IB Math AI SL — Topic 4

Correlation & Regression

Scatter Diagrams · Pearson's r & Spearman's r_s · Linear Regression

Difficulty	EASY	MEDIUM	HARD
Questions	10	10	10
Total Marks	30	40	50

Scan me for help

Name:________________________ Date:______________

Score:__________/120

Instructions: Answer all questions. Show all working. Give answers correct to 3 significant figures unless stated otherwise. A graphic display calculator is required throughout; support GDC answers with the values entered and the feature used.

Key Formulae

- Pearson's product-moment correlation coefficient (by hand): $r = \dfrac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$, $\quad S_{xy} = \sum x_i y_i - \dfrac{(\sum x_i)(\sum y_i)}{n}$, $\quad S_{xx} = \sum x_i^2 - \dfrac{(\sum x_i)^2}{n}$, $\quad S_{yy} = \sum y_i^2 - \dfrac{(\sum y_i)^2}{n}$
- Range of Pearson's r: $-1 \le r \le 1$; $r = \pm 1$ perfect linear, $r = 0$ no linear association.
- Regression line of y on x: $\hat{y} = ax + b$, $\quad a = \dfrac{S_{xy}}{S_{xx}}$, $\quad b = \bar{y} - a\bar{x}$.
- Mean point property: The regression line of y on x always passes through $(\bar{x}, \bar{y})$.
- Spearman's rank correlation coefficient: $r_s = 1 - \dfrac{6\sum d_i^2}{n(n^2-1)}$, $\quad$ where $d_i = \text{rank}(x_i) - \text{rank}(y_i)$
- Tied ranks: assign the mean of the positions that would have been occupied, then apply the formula (or use GDC on the rank data).
- Scaling invariance: Pearson's r is unchanged by any linear transformation $x \mapsto cx + d$ or $y \mapsto ey + f$ $(c, e \neq 0)$; units do not affect r.

GDC Tips

- **Pearson's r and regression line y on x:** Enter data into two lists. STAT $\rightarrow$ CALC $\rightarrow$ 8: LinReg(a+bx), specify Xlist and Ylist. The output gives a (gradient), b (intercept), and r directly. Ensure DiagnosticOn is active (MODE or CATALOG) or r will not display.
- **Spearman's r_s:** Manually convert raw data to ranks (smallest = 1) in two new lists, then run LinReg(a+bx) on those rank lists; the r value shown is r_s.
- **Checking the mean point:** After finding a and b, substitute $\bar{x}$ (from STAT $\rightarrow$ CALC $\rightarrow$ 1-Var Stats) and verify $a\bar{x} + b = \bar{y}$; use this to recover a missing data value algebraically.
- **Predicting from the regression line:** Store the equation in Y1 and use TABLE or direct substitution. Always state whether the x-value lies inside the data range (interpolation, reliable) or outside (extrapolation, unreliable).
- **Scatter diagram:** Enter (x, y) pairs in L1, L2, then STAT PLOT $\rightarrow$ Plot1 ON, type = scatter, Xlist: L1, Ylist: L2. Press ZOOM $\rightarrow$ 9: ZoomStat for an automatic window.

Common Mistakes to Avoid

1. **Predicting x from y using the y-on-x line.** The regression line of y on x minimises vertical residuals and must only be used to predict y given x. Using it to find x from a known y gives a biased answer; you would need the separate regression line of x on y.

2. **Forgetting DiagnosticOn and missing r.** If the diagnostic setting is off, the GDC silently omits r from the LinReg output. Always confirm DiagnosticOn at the start of any correlation question; a blank where r should appear costs every accuracy mark that depends on it.

3. **Treating correlation as causation.** A strong r (or r_s) shows association, not cause and effect. A conclusion that "x causes y" scores zero; you must name a plausible confounding variable or state that both variables may be driven by a third factor.

4. **Extrapolating without critique.** Substituting an x-value outside the observed data range into the regression equation produces a prediction, but full marks require an explicit statement that this is extrapolation and the result may be unreliable because the linear trend may not continue beyond the data.

5. **Mishandling the gradient's units and sign in context.** The gradient a must be interpreted as "for each additional one [unit of x], y increases/decreases by $|a|$ [units of y]". Omitting the context nouns, the units, or the direction (sign) loses the interpretation mark every time.

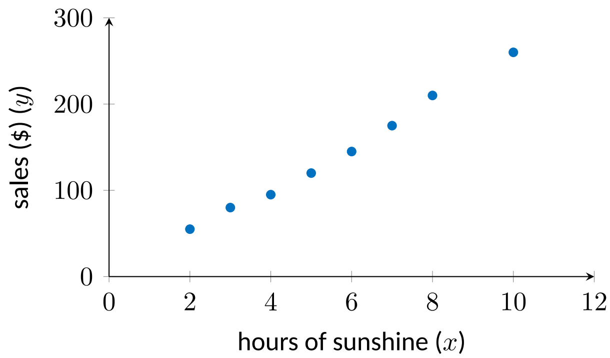
6. **Confusing r and r_s when one outlier is present.** Pearson's r is sensitive to outliers because it uses raw values; Spearman's r_s is not, because it uses only ranks. When r and r_s differ substantially, the correct conclusion is that an outlier is distorting r, and r_s gives a more reliable measure of monotonic association. Swapping this explanation, or claiming both measure the same thing, loses the comparison (R) mark.

Section A — Easy EASY (10 questions × 3 marks = 30 marks)

Scan me for help

1. EASY Calculator **[3 marks]**

The scatter diagram below shows the relationship between the number of hours of sunshine per day, x, and the daily ice-cream sales in dollars, y, at a beach kiosk over eight days.



Describe the correlation between hours of sunshine and daily ice-cream sales shown in the scatter diagram.

2. EASY Calculator **[3 marks]**

The table below shows the number of weekly training sessions, x, and the 5 km race time in minutes, y, for seven athletes.

x (sessions)	1	2	3	4	5	6	7
y (minutes)	34	31	28	26	24	22	20

Using your GDC, find Pearson's product-moment correlation coefficient, r, for this data. State what this value indicates about the relationship between training sessions and race time.

3. **EASY** **Calculator** **[3 marks]**

A researcher records the average daily temperature in °C, x, and the number of visitors to an indoor swimming pool, y, over six days. The regression line of y on x is found to be

$$y = -3.2x + 180.$$

Write down the gradient of this regression line and interpret its meaning in context.

4. **EASY** **Calculator** **[3 marks]**

The table below shows the scores of six students in a mathematics test, x, and a science test, y.

x (maths)	45	52	60	67	74	83
y (science)	48	55	62	65	76	85

Using your GDC, find the equation of the regression line of y on x in the form $y = ax + b$.

5. **EASY** **Calculator** **[3 marks]**

For a set of eight data pairs, the mean value of x is $\bar{x} = 5$ and the mean value of y is $\bar{y} = 34$. The regression line of y on x passes through the mean point $(\bar{x}, \bar{y})$ and has gradient $a = 4.5$.
Find the y-intercept of the regression line, giving your answer in the form $y = ax + b$.

6. **EASY** **Calculator** **[3 marks]**

Seven cities are ranked by their annual rainfall, R, and the number of rainy days per year, D. The ranks are given in the table below.

City	A	B	C	D	E	F	G
Rank R	1	2	3	4	5	6	7
Rank D	2	1	4	3	6	5	7

Using your GDC, find Spearman's rank correlation coefficient, r_s, for this data.

7. EASY Calculator **[3 marks]**

A study records the number of hours of sleep, x, and the reaction time in milliseconds, y, for ten participants. The regression line of y on x is

$$y = -12.4x + 310.$$

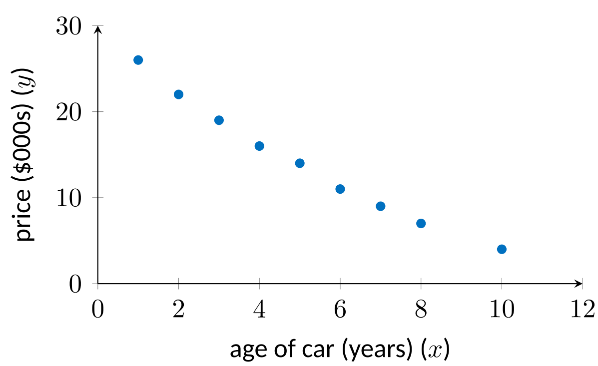
Use this equation to estimate the reaction time for a participant who sleeps for 7 hours.

8. EASY Calculator **[3 marks]**

For a dataset of paired values, Pearson's product-moment correlation coefficient is calculated as $r = 0.94$. Write down what this value of r tells you about the linear relationship between the two variables.

9. EASY Calculator **[3 marks]**

The scatter diagram below shows the relationship between the age of a car in years, x, and its second-hand selling price in thousands of dollars, y, for nine cars.



Describe the correlation between the age of a car and its selling price shown in the scatter diagram.

10. **EASY** **Calculator** **[3 marks]**

A sports scientist records the daily protein intake in grams, x, and the muscle mass gained in kilograms, y, for eight athletes over one month. Using a GDC, the regression line of y on x is found to be

$$y = 0.045x + 0.3.$$

(a) Write down the value of the gradient and interpret its meaning in context. [2]

(b) Write down the y-intercept and state what it represents in context. [1]

Section B — Medium MEDIUM (10 questions × 4 marks = 40 marks)

Scan me for help

11. MEDIUM Calculator **[4 marks]**

A marine biologist measures the shell length, x mm, and the mass, y g, of eight clams. The data are shown in the table below.

Shell length x (mm)	22	27	31	35	38	42	46	51
Mass y (g)	14	19	23	28	31	37	42	49

(a) Find Pearson's product-moment correlation coefficient, r. [1]

(b) Find the equation of the regression line of y on x, giving your answer in the form $y = ax + b$. [2]

(c) Interpret the value of a in context. [1]

12. MEDIUM Calculator **[4 marks]**

A sports scientist records the weekly training hours, x, and the reaction time, y ms, of nine athletes. The Pearson correlation coefficient is found to be $r = -0.94$.

(a) Describe the correlation between training hours and reaction time. [2]

(b) The scientist converts training hours to training minutes by multiplying each x-value by 60. State the value of the new Pearson correlation coefficient. [1]

(c) Explain why the correlation coefficient does not change. [1]

13. MEDIUM Calculator **[4 marks]**

Seven students each sit a written test and a practical exam. Their scores out of 50 are given in the table below.

Student	A	B	C	D	E	F	G
Written score x	12	18	24	29	33	38	45
Practical score y	16	21	20	30	28	40	43

Complete the rank table below and hence find Spearman's rank correlation coefficient, r_s.

Student	A	B	C	D	E	F	G
Rank of x	1	2	3	4	5	6	7
Rank of y							

14. MEDIUM Calculator **[4 marks]**

A nutritionist collects data on the daily sugar intake, x g, and blood glucose level, y mmol/L, of ten participants. The regression line of y on x is found to be $y = 0.032x + 4.1$.

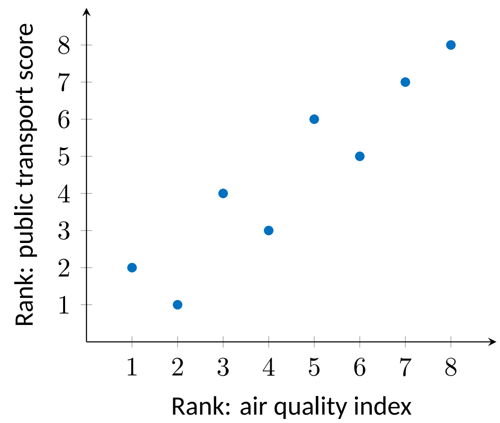
(a) The mean daily sugar intake is $\bar{x} = 85$ g. Find the mean blood glucose level, $\bar{y}$. [2]

(b) Estimate the blood glucose level of a participant with a daily sugar intake of 70 g. [1]

(c) A participant reports a daily sugar intake of 350 g. State, with a reason, whether it is appropriate to use the regression line to estimate their blood glucose level. [1]

15. MEDIUM Calculator **[4 marks]**

Eight cities are ranked by air quality index (rank 1 = best) and by public transport score (rank 1 = best). The scatter diagram below shows these ranks.



(a) Write down the ranks from the scatter diagram and find Spearman's rank correlation coefficient, r_s, using your GDC. [2]

(b) Describe what the value of r_s suggests about the relationship between air quality index and public transport score for these cities. [1]

(c) State one advantage of using r_s rather than Pearson's r for ranked data. [1]

16. MEDIUM Calculator **[4 marks]**

The table below shows the altitude, x m, and mean annual temperature, y °C, at seven weather stations.

Altitude x (m)	120	250	410	580	730	890	1050
Temperature y (°C)	18.4	16.1	13.8	11.2	9.3	7.0	4.8

(a) Find the equation of the regression line of y on x in the form $y = ax + b$. [2]

(b) Interpret the value of b in context, and explain whether this interpretation is meaningful. [1]

(c) Use the regression line to estimate the mean annual temperature at an altitude of 500 m. [1]

17. MEDIUM Calculator **[4 marks]**

A researcher studying social media use records the daily screen time, x hours, and the self-reported wellbeing score, y (out of 100), for eleven teenagers. She calculates $r = -0.81$ and $r_s = -0.76$.

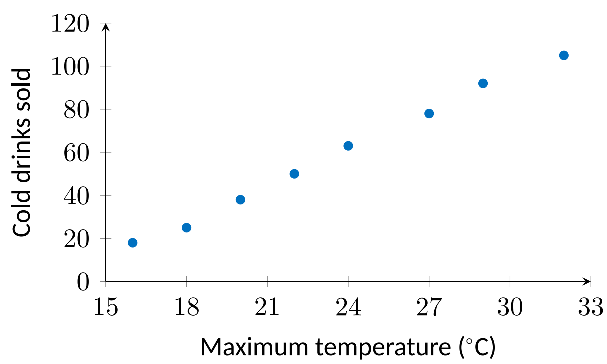
(a) Describe the correlation shown by r. [1]

(b) One teenager's screen time is an outlier. Explain, with reference to the values of r and r_s, which coefficient is less affected by this outlier. [2]

(c) State why a strong correlation between screen time and wellbeing does not imply that screen time causes lower wellbeing. [1]

18. MEDIUM Calculator **[4 marks]**

A café owner records the daily maximum temperature, x °C, and the number of cold drinks sold, y, over eight days. The data are shown in the scatter diagram below.



The regression line of y on x passes through the mean point $(\bar{x}, \bar{y})$. Given that the equation is $y = 5.3x - 67.4$:

(a) Find the mean number of cold drinks sold, $\bar{y}$, given that $\bar{x} = 23.5$ °C. [2]

(b) Estimate the number of cold drinks sold on a day when the maximum temperature is 25 °C. [1]

(c) The owner claims the regression line can be used to predict sales when the temperature reaches 45 °C. Give one reason why this claim is not valid. [1]

19. MEDIUM Calculator **[4 marks]**

The table shows the price, x (thousands of USD), and the age, y years, of eight second-hand cars of the same model.

Price x (thousands USD)	18.5	16.2	14.0	11.8	9.4	7.6	5.3	3.1
Age y (years)	1	2	3	4	5	6	7	8

(a) Find Pearson's product-moment correlation coefficient, r. [1]

(b) Find the regression line of y on x in the form $y = ax + b$. [2]

(c) A car of this model is advertised at a price of USD 10 000. Estimate its age. [1]

20. MEDIUM Calculator **[4 marks]**

A geologist records the depth below ground, x m, and the soil temperature, y °C, at six locations. She finds Pearson's $r = 0.97$ and the regression line of y on x is $y = 0.045x + 8.2$. The data cover depths from 10 m to 80 m.

(a) Interpret the gradient 0.045 in context. [1]

(b) Estimate the soil temperature at a depth of 50 m. [1]

(c) The geologist uses the regression line to estimate the soil temperature at a depth of 500 m. Determine whether this estimate is reliable. Justify your answer. [2]

Section C — Hard **HARD** (10 questions × 5 marks = 50 marks)

Scan me for help

21. **HARD** **Calculator** **[5 marks]**

A researcher records the average daily temperature, x (°C), and the number of hot drinks sold, y, at a mountain café over eight randomly selected days.

x (°C)	-4	0	3	7	10	14	18	22
y (drinks)	98	91	84	73	65	54	41	29

(a) Find Pearson's product-moment correlation coefficient r for this data. [1]

(b) Find the equation of the regression line of y on x in the form $y = ax + b$. [2]

(c) The researcher claims that for every 1°C rise in temperature, approximately 3.5 fewer hot drinks are sold. Determine whether this claim is supported by the regression equation. Justify your answer. [1]

(d) A temperature of 30°C is forecast. State, giving a reason, whether it would be appropriate to use the regression line to predict the number of hot drinks sold. [1]

22. **HARD** **Calculator** **[5 marks]**

Eight athletes compete in a sprint event. Their training hours per week, h, and their finishing rank (1 = fastest) are recorded. A coach calculates both Pearson's r and Spearman's r_s for the data.

Athlete	A	B	C	D	E	F	G	H
h (hours/week)	5	8	10	12	15	18	20	24
Finishing rank	7	6	5	8	4	2	3	1

(a) Find Spearman's rank correlation coefficient r_s. [2]

(b) Find Pearson's correlation coefficient r between h and finishing rank. [1]

(c) The coach notices that Athlete D trained 12 hours yet finished last. Compare r and r_s and explain which coefficient better describes the relationship between training hours and finishing rank for this data. [2]

23. **HARD** **Calculator** **[5 marks]**

The table below shows the number of years of experience, x, and the monthly salary, y (in thousands of dollars), for nine employees at a consultancy firm.

x (years)	1	2	4	5	6	8	10	13	15
y ($'000)	3.2	3.8	4.5	5.1	5.6	6.8	8.0	9.5	11.2

(a) Show that the mean experience is $\bar{x} = \frac{64}{9} \approx 7.11$ years and the mean salary is $\bar{y} \approx 6.41$ thousand dollars. [1]

(b) Find the equation of the regression line of y on x. [2]

(c) A new employee has 7 years of experience. Use the regression line to estimate their monthly salary. [1]

(d) An employee earning $12 500 per month asks a colleague to use the regression line to find their expected years of experience. State why this is not a valid use of the regression line of y on x, and determine their expected experience using an appropriate method. [1]

24. HARD Calculator **[5 marks]**

A marine biologist measures the shell diameter, d (mm), and the mass, m (g), of ten sea urchins collected from a reef. The regression line of m on d is found to be $m = 0.84d - 11.2$.
It is also known that $\bar{d} = 52.0$ mm and Pearson's $r = 0.973$.

(a) Find the mean mass $\bar{m}$ of the sea urchins. [2]

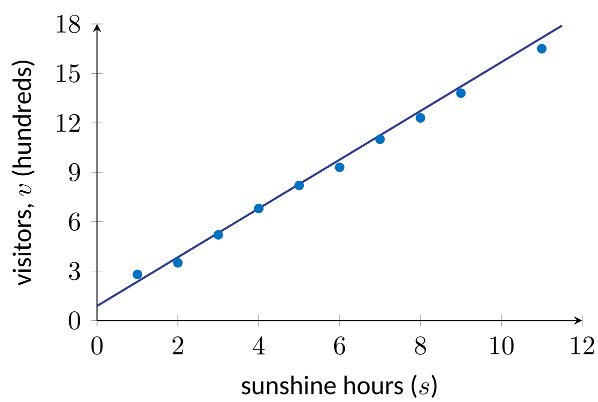
(b) Interpret the gradient of the regression line in context. [1]

(c) The biologist states that the value $r = 0.973$ confirms that larger shell diameter causes greater mass. Explain why this statement is not valid. [1]

(d) A sea urchin has shell diameter $d = 38$ mm. Determine whether using the regression line to predict its mass is appropriate, given that the diameters in the sample range from 40 mm to 68 mm. [1]

25. HARD Calculator **[5 marks]**

The scatter diagram below shows the relationship between daily sunshine hours, s, and the number of visitors, v (hundreds), to a coastal park on ten randomly selected days. The regression line of v on s passes through the point $(4, 6.8)$.



The gradient of the regression line is $a = 1.48$ (3 s.f.).

(a) Write down the y-intercept of the regression line, showing your method clearly. [2]

(b) Interpret the gradient in context. [1]

(c) Pearson's $r = 0.999$. A student argues that sunshine causes more visitors. Identify the statistical error in this reasoning and suggest one possible confounding variable. [1]

(d) Estimate the number of visitors when there are 6 hours of sunshine. Give your answer to the nearest ten. [1]

26. HARD Calculator **[5 marks]**

A tutor records the number of hours spent studying per week, x, and the test score achieved, y (%), for a sample of 10 students. The following summary statistics are known:

$$n = 10, \quad \sum x = 52, \quad \sum y = 706, \quad \sum x^2 = 318, \quad \sum xy = 3942, \quad \sum y^2 = 51\,394$$

(a) Show that $\bar{x} = 5.2$ hours and $\bar{y} = 70.6\%$. [1]

(b) Find Pearson's correlation coefficient r. [2]

(c) Find the gradient a of the regression line of y on x. [1]

(d) Hence write down the full equation of the regression line of y on x. [1]

27. HARD Calculator **[5 marks]**

Seven students sit two examinations: Paper 1 (score out of 100) and Paper 2 (score out of 100). Their scores are shown below.

Student	1	2	3	4	5	6	7
Paper 1, p	45	52	61	68	74	80	93
Paper 2, q	50	48	65	70	71	85	90

(a) Assign ranks to each student for Paper 1 and Paper 2 (rank 1 = highest score). Show the complete rank table. [1]

(b) Hence find Spearman's rank correlation coefficient r_s. [1]

(c) Find Pearson's r for the data. [1]

(d) Student 2 scored unusually low on Paper 2 relative to Paper 1. Explain how this is reflected in the values of r and r_s, and state which coefficient is more appropriate to report. [2]

28. **HARD** **Calculator** **[5 marks]**

A car dealership records the age, t (years), and selling price, p (thousands of euros), of nine second-hand cars of the same model. The regression line of p on t has gradient $a = -2.35$ and passes through $(\bar{t}, \bar{p})$ where $\bar{t} = 5.0$ years.
It is also known that $\bar{p} = 14.8$ thousand euros.

(a) Find the equation of the regression line of p on t. [2]

(b) A car is 3 years old. Use the regression line to estimate its selling price. [1]

(c) Show that a car predicted to sell for exactly €8 000 would be approximately 7.89 years old according to the model. [1]

(d) The dealership wishes to use the model for a 15-year-old car. Determine whether this is appropriate, given that the ages of cars in the sample range from 1 to 10 years. Justify your answer. [1]

29. HARD Calculator **[5 marks]**

A psychologist studies the relationship between hours of sleep, s (hours per night), and a cognitive performance score, c (out of 50), for eight participants.

s (hours)	4.0	5.0	5.5	6.0	7.0	7.5	8.0	9.0
c (score)	18	24	27	30	38	40	44	46

(a) Find Pearson's r and Spearman's r_s for this data. [2]

(b) The regression line of c on s is found. Given that $\bar{s} = 6.5$ hours and $\bar{c} = 33.375$, and the gradient of the regression line is $a = 5.40$ (3 s.f.), find the y-intercept b and hence write down the full regression equation. [2]

(c) A participant sleeps 6.5 hours per night. Write down the predicted cognitive score. Explain why this result could have been obtained without computing b. [1]

30. HARD Calculator **[5 marks]**

A sports physiologist records the number of hours per week spent on structured exercise, x, and the resting heart rate, y (beats per minute), for ten adults.

x (hours)	1	2	3	4	5	6	7	8	9	10
y (bpm)	82	80	78	76	72	70	68	64	62	58

(a) Find the equation of the regression line of y on x in the form $y = ax + b$. [2]

(b) A person exercises for 6.5 hours per week. Estimate their resting heart rate. [1]

(c) Show that a resting heart rate of 60 bpm corresponds to approximately 9.64 hours of exercise per week according to the model. [1]

(d) A physiologist claims the data proves that exercising more directly causes a lower resting heart rate. State one reason why this conclusion is not justified from the regression analysis alone. [1]

Mark Schemes

Q1 — Describing Correlation from a Scatter Diagram EASY

To describe correlation we comment on direction, strength and form of the pattern in context.
The points rise from bottom-left to top-right, lie close to a straight line, with no obvious outliers. **A1**
The correlation is strong and positive. **A1**
This means that as the number of hours of sunshine increases, the daily ice-cream sales also increase. **A1**

Note: All three elements — direction (positive), strength (strong), and context statement — are needed for full marks.

Q2 — Pearson's r by GDC EASY

Enter the x and y values into two lists and use the linear regression function to obtain r.
GDC: Statistics $\to$ Linear Regression (a+bx) on lists $x = \{1, 2, 3, 4, 5, 6, 7\}$ and $y = \{34, 31, 28, 26, 24, 22, 20\}$; read off r. **M1**
$r = -0.995$ (unrounded: $-0.9952\ldots$) **A1**
Since r is very close to -1, this indicates a strong negative linear relationship: as the number of training sessions increases, the race time decreases. **A1**

Note: The sign of r must be negative here; a positive value means the lists were entered in the wrong order or mis-keyed.

Q3 — Interpreting the Gradient of a Regression Line EASY

The gradient is the coefficient of x in the equation $y = -3.2x + 180$. **A1**
Gradient $= -3.2$ **A1**
This means that for each additional degree Celsius increase in average daily temperature, the number of visitors to the indoor swimming pool decreases by 3.2 visitors. **A1**

Note: The interpretation must use the correct context units (visitors and °C) and must state the direction of change (decrease, not increase).

Q4 — Regression Line Equation by GDC EASY

Enter the maths scores as list x and science scores as list y, then apply linear regression.
GDC: Statistics $\to$ Linear Regression ($y = ax + b$) on $x = \{45, 52, 60, 67, 74, 83\}$ and $y = \{48, 55, 62, 65, 76, 85\}$. **M1**
Reading off: $a = 0.9555\ldots \approx 0.956$ and $b = 4.490\ldots \approx 4.49$ **A1**

$$y = 0.956x + 4.49$$

A1

Note: This is the regression line of y on x only — it is used to predict science scores from maths scores, not the other way around.

Q5 — Finding the y-intercept Using the Mean Point EASY

Every regression line of y on x passes through the mean point $(\bar{x}, \bar{y})$, so we substitute these values into $y = ax + b$ to find b. **M1**

Substituting $\bar{x} = 5$, $\bar{y} = 34$, and $a = 4.5$:

$$34 = 4.5(5) + b$$

$$34 = 22.5 + b$$

$$b = 11.5$$

A1

The regression line is $y = 4.5x + 11.5$. **A1**

Note: Substituting $\bar{x}$ and $\bar{y}$ correctly is the key step; using any other point on the supposed line would not give the correct b.

Q6 — Spearman's Rank Correlation Coefficient by GDC EASY

Enter the rank values directly into two lists and apply linear regression to obtain Spearman's r_s from the correlation coefficient of the ranks.

GDC: Statistics $\to$ Linear Regression on rank $R = \{1, 2, 3, 4, 5, 6, 7\}$ and rank $D = \{2, 1, 4, 3, 6, 5, 7\}$; read off r, which equals r_s when computed on ranks. **M1**

$r_s = 0.893$ (unrounded: $0.8929\ldots$) **A1**

This indicates a strong positive monotonic relationship between annual rainfall and number of rainy days. **A1**

Note: Spearman's r_s is found by entering the ranks into the GDC, not the original data values.

Q7 — Using a Regression Line to Estimate a Value EASY

Substitute $x = 7$ into the regression equation to estimate the reaction time.

GDC (or by hand): substitute $x = 7$ into $y = -12.4(7) + 310$. **M1**

$$y = -12.4 \times 7 + 310 = -86.8 + 310 = 223.2$$

A1

The estimated reaction time for a participant who sleeps 7 hours is 223 milliseconds (to 3 s.f.). **A1**

Note: Keep the unrounded value (223.2) visible before stating the rounded answer.

Q8 — Interpreting a Value of Pearson's r EASY

Pearson's r measures the strength and direction of the linear association between two variables; its value lies in $[-1, 1]$. **A1**
Since $r = 0.94$ is close to $+1$, the two variables have a strong positive linear relationship. **A1**
As one variable increases, the other tends to increase in an approximately linear manner. **A1**

Note: Always describe both strength (strong/moderate/weak) and direction (positive/negative) when interpreting r.

Q9 — Describing Correlation from a Scatter Diagram EASY

To describe correlation we comment on direction, strength and form in context.
The points fall from top-left to bottom-right, following a roughly linear pattern closely. **A1**
The correlation is strong and negative. **A1**
This means that as the age of a car increases, its second-hand selling price decreases. **A1**

Note: All three descriptive elements — direction (negative), strength (strong), and a statement linking the two context variables — are required for full credit.

Q10 — Interpreting Gradient and Intercept of a Regression Line EASY

(a) The gradient is the coefficient of x in $y = 0.045x + 0.3$.
Gradient $= 0.045$ **A1**
This means that for each additional gram of daily protein intake, the muscle mass gained increases by 0.045 kilograms. **A1**

(b) The y-intercept is $b = 0.3$. **A1**
This represents the predicted muscle mass gained (0.3 kg) when protein intake is zero grams per day; in practice this value may not be meaningful as zero protein intake is unrealistic in this context.

Note: The gradient interpretation must include the correct context units (grams and kilograms) and state the direction (increases).

Q11 — Clam Shell Length vs Mass and Regression MEDIUM

(a) To find Pearson's r, enter all eight (x, y) pairs into the GDC and read the correlation coefficient directly.
GDC: Statistics $\rightarrow$ Linear Regression (LinReg) on lists $x = \{22, 27, 31, 35, 38, 42, 46, 51\}$, $y = \{14, 19, 23, 28, 31, 37, 42, 49\}$; read off r.
$r = 0.998$ (3 s.f.) **A2**

(b) Using the same GDC output, the regression line of y on x is obtained from the same LinReg screen.
GDC: same regression output gives $a = 1.21$ (unrounded: $1.2125\ldots$) and $b = -13.9$ (unrounded: $-13.880\ldots$).

$$y = 1.21x - 13.9$$

A1

Note: Always write the line as y on x (predicts y from x); reversing to predict x from y requires a different regression line. (M1) implied by correct form

(c) The gradient $a = 1.21$ means that for each additional 1 mm in shell length, the mass of the clam increases by approximately 1.21 g. **A1**

Q12 — Training Hours and Reaction Time and Scaling MEDIUM

(a) The correlation is described using direction, strength, and the variable names in context.
There is a strong, negative correlation between weekly training hours and reaction time: as training hours increase, reaction time decreases. **A1 A1**
(Award first A1 for "strong negative"; second A1 for a context sentence using both variable names.)

(b) Multiplying every x-value by the same positive constant (60) is a linear scaling; Pearson's r is invariant under linear transformations of either variable.
New $r = -0.94$ (unchanged). **A1**
Note: A common error is to multiply r by 60. The correlation coefficient has no units and is unaffected by any positive linear rescaling.

(c) Pearson's r measures the strength of the linear relationship between variables; scaling x by a positive constant does not change the relative spread or direction of the data, so the standardised covariance (i.e. r) is unchanged. **A1**

Q13 — Written and Practical Exam Spearman Ranks MEDIUM

Rank the practical scores y: ordering $16, 20, 21, 28, 30, 40, 43$ gives ranks $1, 3, 2, 5, 4, 6, 7$ for students A–G respectively.

Student	A	B	C	D	E	F	G
Rank of x	1	2	3	4	5	6	7
Rank of y	1	3	2	5	4	6	7

M1 (for correct ranks of y)
Enter the two rank lists into the GDC and apply linear regression / correlation on the ranks.
GDC: Statistics $\rightarrow$ LinReg on rank-x list $\{1, 2, 3, 4, 5, 6, 7\}$ and rank-y list $\{1, 3, 2, 5, 4, 6, 7\}$; read r (which equals r_s when applied to ranks).
$r_s = 0.929$ (unrounded: $0.9286\ldots$) **A1**
The value $r_s \approx 0.93$ indicates a strong positive monotonic association between written and practical exam scores. **A1**

Note: Practical scores must be ranked in ascending order (rank 1 = lowest score). Student C's practical score (20) is lower than Student B's (21), so C receives rank 2 and B receives rank 3.

Q14 — Sugar Intake and Blood Glucose Mean Point MEDIUM

(a) The mean point $(\bar{x}, \bar{y})$ always lies on the regression line of y on x. Substituting $\bar{x} = 85$ into the equation gives $\bar{y}$.

$\bar{y} = 0.032(85) + 4.1$ **M1**

$= 2.72 + 4.1 = 6.82$ mmol/L **A1**

(b) Substituting $x = 70$ into the regression equation gives the estimate.

$y = 0.032(70) + 4.1 = 2.24 + 4.1 = 6.34$ mmol/L **A1**

(c) A daily sugar intake of 350 g lies well outside the data range. This is extrapolation, so the estimate is not reliable (the linear relationship may not hold beyond the observed range). **R1**

Q15 — City Air Quality and Transport Ranks Spearman MEDIUM

(a) Read the eight coordinate pairs from the scatter diagram: $(1, 2), (2, 1), (3, 4), (4, 3), (5, 6), (6, 5), (7, 7), (8, 8)$.

Enter rank-x list $\{1, \ldots, 8\}$ and rank-y list $\{2, 1, 4, 3, 6, 5, 7, 8\}$ into the GDC.

GDC: Statistics $\to$ LinReg on the two rank lists; read r as r_s. **M1**

$r_s = 0.929$ (unrounded: $0.9286\ldots$) **A1**

(b) The value $r_s \approx 0.93$ indicates a strong positive monotonic association: cities with better (lower-numbered) air quality ranks tend to also have better public transport scores. **A1**

(c) Spearman's r_s measures monotonic association and is less sensitive to outliers than Pearson's r, making it more appropriate when data are given as ranks or may contain extreme values. **A1**

Note: Do not confuse "monotonic" with "linear".

Q16 — Altitude vs Temperature Intercept Interpretation MEDIUM

(a) Enter the altitude and temperature data into the GDC and perform linear regression of y on x.

GDC: Statistics $\to$ LinReg on $x = \{120, 250, 410, 580, 730, 890, 1050\}$, $y = \{18.4, 16.1, 13.8, 11.2, 9.3, 7.0, 4.8\}$; read a and b. **M1**

$a = -0.014456\ldots \approx -0.0145, \quad b = 19.837\ldots \approx 19.8$

$$y = -0.0145x + 19.8$$

A1

(b) The intercept $b \approx 19.8$ would represent the mean annual temperature at an altitude of 0 m (sea level). However, this interpretation is not fully meaningful because the lowest altitude in the dataset is 120 m; predicting at $x = 0$ is a mild extrapolation beyond the observed data range. **A1**

(c) Substituting $x = 500$ into the regression equation:

$$y = -0.0145(500) + 19.8 = -7.25 + 19.8 = 12.6\,^{\circ}\text{C (3 s.f.)}$$

A1

Note: 500 m lies within the data range (120 m to 1050 m), so this is interpolation and the estimate is valid.

Q17 — Screen Time and Wellbeing r vs rs Comparison MEDIUM

(a) $r = -0.81$ indicates a strong negative correlation between daily screen time and wellbeing score: as screen time increases, wellbeing decreases. **A1**

(b) Spearman's r_s is based only on the ranks of the data values; an outlier changes only one rank, so it has little effect on r_s. Pearson's r, by contrast, uses the actual data values and is pulled more strongly by extreme observations.
Since $|r_s| = 0.76 < |r| = 0.81$, the outlier inflates (strengthens) Pearson's r more than Spearman's; r_s is less affected by the outlier. **R1**
Therefore r_s is the less affected coefficient. **A1**

(c) Correlation measures the strength of association between two variables, but does not establish causation. A plausible confounding variable — for example, underlying mental health status — could independently affect both screen time and wellbeing, producing the observed correlation without any direct causal link. **A1**

Q18 — Temperature and Cold Drinks Mean Point and Extrapolation MEDIUM

(a) The regression line always passes through the mean point $(\bar{x}, \bar{y})$. Substituting $\bar{x} = 23.5$ into $y = 5.3x - 67.4$:
$\bar{y} = 5.3(23.5) - 67.4$ **M1**
$= 124.55 - 67.4 = 57.15 \approx 57.2$ cold drinks **A1**

(b) Substituting $x = 25$ into the regression equation:
$y = 5.3(25) - 67.4 = 132.5 - 67.4 = 65.1 \approx 65$ cold drinks **A1**
(Accept 65 or 65.1; the data range is 16–32 °C so $x = 25$ is valid interpolation.)

(c) A temperature of 45 °C lies well outside the observed data range of 16 °C to 32 °C. Using the regression line here is extrapolation; the linear relationship between temperature and sales may not continue to hold at such extreme temperatures, so the estimate is not reliable. **R1**

Q19 — Second-hand Car Price vs Age Regression Reversal Trap MEDIUM

(a) Enter the eight (x, y) pairs into the GDC and perform linear regression of y on x.
GDC: Statistics → LinReg on $x = \{18.5, 16.2, 14.0, 11.8, 9.4, 7.6, 5.3, 3.1\}$, $y = \{1, \dots, 8\}$; read r.
$r = -1.00$ (3 s.f.; unrounded $-0.9998\dots$) **A2**

(b) From the same GDC output, the regression line of y on x:
$a = -0.4566\dots \approx -0.457, \quad b = 9.4023\dots \approx 9.40$

$$y = -0.457x + 9.40$$

A1
Note: This is the regression line of y (age) on x (price). To estimate age from price, this is the correct line.

(c) A price of USD 10 000 means $x = 10$ (thousands USD). Substituting into the regression line:

$y = -0.457(10) + 9.40 = -4.57 + 9.40 = 4.84 \approx 4.84$ years **A1**
(Accept 4.84 years; this lies within the data range so interpolation is valid.)

Q20 — Soil Depth and Temperature Extrapolation Judgement MEDIUM

(a) The gradient 0.045 means that for each additional 1 m of depth below ground, the soil temperature increases by 0.045 °C. **A1**

(b) Substituting $x = 50$ into the regression equation (50 m lies within the data range of 10 m to 80 m, so this is valid interpolation):
$y = 0.045(50) + 8.2 = 2.25 + 8.2 = 10.45 \approx 10.5\,°$C **A1**

(c) A depth of 500 m lies far outside the data range of 10 m to 80 m; using the regression line here is extrapolation. **R1**
Therefore the estimate is not reliable, as the linear relationship observed between 10 m and 80 m may not continue to hold at 500 m. **A1**

Q21 — Temperature vs Hot Drinks: Regression and Extrapolation HARD

(a) To find r, enter the paired data into the GDC as two lists and request linear regression statistics.
GDC: Statistics → Linear Reg ($ax + b$), List1 $= x$, List2 $= y$; read off r.
$r = -0.997$ (3 s.f.) **A2**

(b) From the same GDC output, record a and b.
GDC: same screen; $a = -2.7025\ldots$, $b = 90.522\ldots$

$$y = -2.70x + 90.5 \text{ (3 s.f.)}$$

M1 A1

(c) The gradient of the regression line is $a = -2.70$ (3 s.f.), which means each 1°C rise is associated with approximately 2.70 fewer drinks sold.
Since $2.70 \neq 3.5$, the claim is **not** supported by the regression equation — the actual rate is noticeably smaller than the researcher's estimate. **R1**

(d) The data range is $-4 \leq x \leq 22\,°$C. Since 30°C lies outside this range, using the regression line would be extrapolation and is therefore not appropriate — the linear relationship may not hold beyond the observed data. **A1**

Q22 — Training Hours vs Finishing Rank: Pearson vs Spearman HARD

(a) Rank the training hours h (rank 1 = smallest): the hours are already in ascending order, so ranks for h are $1, 2, 3, 4, 5, 6, 7, 8$ respectively. The finishing ranks are given directly as $7, 6, 5, 8, 4, 2, 3, 1$.
Enter rank-of-h in List1 and finishing rank in List2.

GDC: Statistics $\to$ Linear Reg, read off r computed on the ranks $= r_s$.
$r_s = -0.833$ (3 s.f.) **M1 A1**

(b) Enter the raw h values in List1 and finishing rank in List2; request linear regression.
GDC: Statistics $\to$ Linear Reg ($ax + b$); read off r.
$r = -0.861$ (3 s.f.) **A1**

(c) Athlete D trained in the 12–15 hour range but achieved the worst finishing rank (8), creating an anomalous point that disrupts the otherwise consistent pattern of "more training hours $\to$ better (lower-numbered) finishing rank".
$|r| = 0.861$ and $|r_s| = 0.833$: both coefficients agree the underlying relationship is strong and negative, and neither is dramatically larger than the other here, because Athlete D's training hours (12) are not an extreme x-value in themselves — only their combination with an unexpectedly poor rank is unusual. **R1**
Even so, r_s is the more appropriate coefficient to report: *finishing rank* is already an ordinal variable, not a continuous measurement, and Spearman's r_s is the statistically correct tool for relating an ordinal variable to a continuous one, regardless of the small numerical gap between the two coefficients here. **A1**

Q23 — Experience vs Salary: Regression Reversal HARD

(a) $\bar{x} = \dfrac{1+2+4+5+6+8+10+13+15}{9} = \dfrac{64}{9} \approx 7.11$ years. **AG**
$\bar{y} = \dfrac{3.2+3.8+4.5+5.1+5.6+6.8+8.0+9.5+11.2}{9} = \dfrac{57.7}{9} \approx 6.41$ thousand dollars. **AG**
(Both values must be shown with their sums to earn the mark.) **M1**

(b) Enter x in List1 and y in List2; request linear regression.
GDC: Statistics $\to$ Linear Reg ($ax + b$); read $a = 0.5597\ldots$, $b = 2.4308\ldots$

$$y = 0.560x + 2.43 \text{ (3 s.f.)}$$

M1 A1

(c) Substitute $x = 7$: $y = 0.560(7) + 2.43 = 3.92 + 2.43 = 6.35$ thousand dollars (3 s.f.), i.e. approximately \$6 350. **A1**

(d) The regression line of y on x predicts y (salary) from x (experience); it cannot validly be rearranged to predict x from a known y — the x on y regression line is required for that.
Using GDC with x and y swapped: regression of x on y gives $x = 1.77y - 4.27$ (unrounded: $a = 1.7748\ldots$, $b = -4.2676\ldots$); substituting $y = 12.5$ into the unrounded line:
$x = 1.7748(12.5) - 4.2676 = 17.9$ years (3 s.f.) **A1**
Note: The most common trap here is simply rearranging $y = ax + b$ to get x; this gives a different (incorrect) line and will lose the mark.

Q24 — Sea Urchins Regression Line Mean Point and Causation HARD

(a) Since the regression line of m on d passes through $(\bar{d}, \bar{m})$, substitute $d = \bar{d} = 52.0$ into $m = 0.84d - 11.2$.
This uses the non-negotiable property that the mean point always lies on the regression line. **M1**
$\bar{m} = 0.84(52.0) - 11.2 = 43.68 - 11.2 = 32.5$ g (3 s.f.) **A1**

(b) The gradient 0.84 means that for each additional 1 mm increase in shell diameter, the mass of the sea urchin increases by 0.84 g on average. **A1**

(c) Correlation does not imply causation; there may be a confounding variable (for example, age of the sea urchin) that causes both larger diameter and greater mass simultaneously. **R1**

(d) The sample diameters range from 40 mm to 68 mm. Since $d = 38$ mm lies below the minimum observed value, using the regression line here would be extrapolation, so it is not appropriate. **A1**

Q25 — Sunshine Hours vs Park Visitors: Intercept from Mean Point HARD

(a) Since the regression line passes through $(\bar{s}, \bar{v}) = (4, 6.8)$ and has gradient $a = 1.48$, substitute into $v = as + b$:
$6.8 = 1.48(4) + b$ **M1**
$b = 6.8 - 5.92 = 0.880$ (3 s.f.) **A1**

(b) The gradient 1.48 means that for each additional hour of sunshine, the number of visitors increases by approximately 148 people (since v is in hundreds). **A1**

(c) The statistical error is concluding causation from correlation alone; even with $r = 0.999$, correlation does not establish a causal relationship. A plausible confounding variable is air temperature — warmer temperatures produce more sunshine hours and independently attract more visitors. **R1**

(d) Substitute $s = 6$ into $v = 1.48(6) + 0.880 = 8.88 + 0.880 = 9.76$ hundreds $= 976$ visitors.
Rounded to the nearest ten: 980 visitors. **A1**

Q26 — Study Hours vs Test Score Regression from Summary Statistics HARD

(a) $\bar{x} = \frac{\sum x}{n} = \frac{52}{10} = 5.2$ hours; $\bar{y} = \frac{\sum y}{n} = \frac{706}{10} = 70.6\%$. Both shown explicitly. **AG**

(b) $S_{xy} = \sum xy - n\bar{x}\bar{y} = 3942 - 10(5.2)(70.6) = 3942 - 3671.2 = 270.8$
$S_{xx} = \sum x^2 - n\bar{x}^2 = 318 - 10(5.2)^2 = 318 - 270.4 = 47.6$
$S_{yy} = \sum y^2 - n\bar{y}^2 = 51\,394 - 10(70.6)^2 = 51\,394 - 49\,843.6 = 1550.4$ **M1**
$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{270.8}{\sqrt{47.6 \times 1550.4}} = \frac{270.8}{271.66\ldots} = 0.997$ (3 s.f.) **A1**
GDC check: entering the reconstructed data (or using summary-statistics regression mode) confirms $r = 0.997$. **A1**

(c) Gradient: $a = \dfrac{S_{xy}}{S_{xx}} = \dfrac{270.8}{47.6} = 5.69$ (3 s.f.) **M1**

(d) Since $(\bar{x}, \bar{y})$ lies on the regression line: $b = \bar{y} - a\bar{x} = 70.6 - 5.69(5.2) = 70.6 - 29.6 = 41.0$ (3 s.f.)

$$y = 5.69x + 41.0$$

A1

Note: All five summary statistics (n, $\sum x$, $\sum y$, $\sum x^2$, $\sum xy$, $\sum y^2$) are needed and are all given here — unlike a shortcut using separately-quoted standard deviations, working directly from the sums cannot produce a contradiction.

Q27 — Two Exam Papers Spearman vs Pearson and Outlier Effect HARD

(a) Rank with rank 1 = highest score. Paper 1 ranks (highest score = 1):

Student	1	2	3	4	5	6	7
Paper 1 rank	7	6	5	4	3	2	1
Paper 2 rank	6	7	5	4	3	2	1

M1

(b) Enter Paper 1 ranks in List1 and Paper 2 ranks in List2; request linear regression to obtain r on the ranks $= r_s$.
GDC: Statistics $\to$ Linear Reg; read $r = r_s = 0.964$ (3 s.f.) **A1**

(c) Enter raw Paper 1 scores in List1 and raw Paper 2 scores in List2.
GDC: Statistics $\to$ Linear Reg; read $r = 0.969$ (3 s.f.) **A1**

(d) Student 2 scored 48 on Paper 2 despite scoring 52 on Paper 1, a drop that disrupts the otherwise strong linear trend. This anomalous point pulls Pearson's r downward more than Spearman's r_s, since Pearson's uses raw values (sensitive to the magnitude of the deviation) whereas Spearman's uses ranks (Student 2's rank simply moves from 6 to 7, a small change).
Since $r_s = 0.964$ is less affected by this single outlier, r_s is more appropriate to report as a measure of association. **R1 A1**

Q28 — Car Age vs Price: Regression Equation and Extrapolation HARD

(a) The regression line passes through $(\bar{t}, \bar{p}) = (5.0, 14.8)$ with gradient $a = -2.35$.
Substitute to find intercept: $14.8 = -2.35(5.0) + b$ **M1**
$b = 14.8 + 11.75 = 26.55 \approx 26.6$ (3 s.f.)

$$p = -2.35t + 26.6$$

A1

(b) Substitute $t = 3$: $p = -2.35(3) + 26.55 = -7.05 + 26.55 = 19.5$ thousand euros (3 s.f.), i.e. €19 500.

A1

(c) Set $p = 8$ (thousands) and solve for t:
$8 = -2.35t + 26.55$
$2.35t = 26.55 - 8 = 18.55$
$t = \dfrac{18.55}{2.35} = 7.8936\ldots \approx 7.89$ years. **AG**
(Unrounded value $7.8936\ldots$ must be shown before stating the printed target.) **M1**

(d) The sample ages range from 1 to 10 years. Since $t = 15$ years lies outside this range, using the regression line would involve extrapolation; the linear relationship may not continue to hold for such old vehicles, so the prediction is not appropriate. **R1**

Q29 — Sleep vs Cognition Two-technique Fusion and Mean Point HARD

(a) Enter s in List1 and c in List2; request linear regression statistics.
GDC: Statistics → Linear Reg $(ax + b)$; read $r = 0.992$ (3 s.f.). **A1**
To find r_s, first rank both variables (rank 1 = smallest value; all values are distinct so no ties):

s rank	1	2	3	4	5	6	7	8
c rank	1	2	3	4	5	6	7	8

GDC: Linear Reg on ranks; $r_s = 1.00$ (exact, since ranks are identical). **A1**

(b) The regression line passes through $(\bar{s}, \bar{c}) = (6.5, 33.375)$ with gradient $a = 5.40$.
$33.375 = 5.40(6.5) + b$ **M1**
$b = 33.375 - 35.10 = -1.725 \approx -1.73$ (3 s.f.)

$$c = 5.40s - 1.73 \text{ (3 s.f.)}$$

A1

(c) Substitute $s = 6.5$: $c = 5.40(6.5) - 1.725 = 35.10 - 1.725 = 33.375 \approx 33.4$.
This result equals $\bar{c}$ exactly, because every regression line passes through $(\bar{x}, \bar{y})$ — so substituting $s = \bar{s}$ always returns $c = \bar{c}$ regardless of b. **A1**

Q30 — Exercise Hours vs Resting Heart Rate: Regression and Causation HARD

(a) Enter x in List1 and y in List2; request linear regression.
GDC: Statistics → Linear Reg $(ax + b)$; $a = -2.6545\ldots$, $b = 85.6$

$$y = -2.65x + 85.6 \text{ (3 s.f.)}$$

M1 A1

(b) Substitute $x = 6.5$: $y = -2.6545(6.5) + 85.6 = -17.25 + 85.6 = 68.3$ (3 s.f.) bpm. **A1**

(c) Set $y = 60$ and solve using the unrounded regression line:

$60 = -2.6545x + 85.6$
$2.6545x = 25.6$
$x = \dfrac{25.6}{2.6545} = 9.6438\ldots \approx 9.64$ hours. **AG**
(Back-substitution: $-2.6545(9.64) + 85.6 = 60.0$ ✓; this value also lies inside the observed data range $[1, 10]$, so it is interpolation.) **M1**

(d) Correlation (and regression) does not establish causation; a confounding variable — such as overall diet and lifestyle quality — could independently influence both how much a person exercises and their resting heart rate, so the data alone cannot prove that exercise directly causes the lower heart rate. **R1**