IB Math AI HL — Topic 4

Correlation & Regression

Scatter Diagrams · PMCC · Spearman's r_s · Coefficient of Determination · Linear Regression

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

Name:______________________________ Date:_______________

Score:__________/120

Scan me for help

Instructions: Answer all questions. Show all working. Give answers correct to 3 significant figures unless otherwise stated. A GDC is permitted throughout.

Key Formulae

Pearson's PMCC: $r = \dfrac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$, where $S_{xy} = \sum x_i y_i - n\bar{x}\bar{y}$, $S_{xx} = \sum x_i^2 - n\bar{x}^2$, $S_{yy} = \sum y_i^2 - n\bar{y}^2$

Regression line y on x: $\hat{y} = ax + b$, $a = \dfrac{S_{xy}}{S_{xx}}$, $b = \bar{y} - a\bar{x}$

Spearman's rank: $r_s = 1 - \dfrac{6\sum d_i^2}{n(n^2-1)}$, where d_i = difference in ranks

Coefficient of determination: $R^2 = r^2$ (for linear model); $0 \le R^2 \le 1$

Linearising: if $y = ab^x$ plot ln y vs x; if $y = ax^n$ plot ln y vs ln x; if $y = a \ln x + b$ plot y vs ln x

GDC Tips

- **PMCC & regression:** `STAT` $\to$ `CALC` $\to$ `8:LinReg(a+bx)` — outputs r, a, b simultaneously.
- **Diagnostic plot:** Enable `DiagnosticOn` (`CATALOG`) so the GDC displays r and r^2 values.
- **Spearman's r_s:** Enter the ranks as two new lists, then run `LinReg` on the ranked lists.
- **Non-linear fit:** Try `ExpReg`, `PwrReg`, `QuadReg` — compare R^2 values to select the best model.
- **Tied ranks:** Average the ranks of tied values (e.g. two values tied for 3rd and 4th both get rank 3.5).

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

Scan me for help

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

A fitness researcher records the daily step count x (thousands) and the resting heart rate y (beats per minute) of seven volunteers. The data is shown in the table below.

Steps, x (thousands)	4.0	5.5	6.0	7.5	8.0	9.0	10.5
Heart rate, y (bpm)	78	74	72	68	66	63	60

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

(b) Write down the equation of the regression line of y on x, giving coefficients correct to 3 significant figures. [1]

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

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

A marine biologist studies the relationship between the water temperature t (°C) and the number of fish n observed in a reef zone per hour. For eight observations, she calculates:

$$r = -0.923, \qquad \hat{n} = -3.14\,t + 87.2$$

(a) Describe the correlation between water temperature and fish count. [1]

(b) Use the regression equation to estimate the number of fish observed when $t = 18\,$°C. Give your answer correct to the nearest whole number. [1]

(c) State whether this estimate involves interpolation or extrapolation. The water temperatures recorded ranged from $14\,$°C to $26\,$°C. [1]

3. EASY GDC **[3 marks]**

Six student athletes are ranked by their coach and by a performance algorithm. The rankings are shown below.

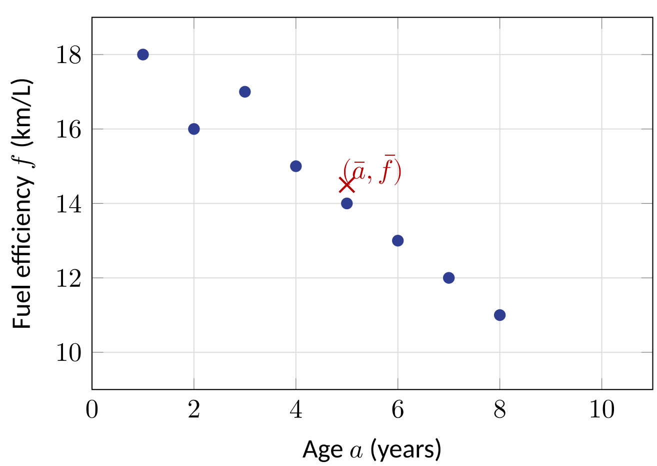
Athlete	A	B	C	D	E	F
Coach rank	1	2	3	4	5	6
Algorithm rank	2	1	4	3	6	5

(a) Calculate the value of Spearman's rank correlation coefficient r_s. [2]

(b) Interpret the value of r_s in context. [1]

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

The scatter diagram below shows the age a (years) and the fuel efficiency f (km per litre) of eight cars. The mean point $(\bar{a}, \bar{f}) = (5, 14.2)$ is also plotted.



The data points are: $(1, 18)$, $(2, 16)$, $(3, 17)$, $(4, 15)$, $(5, 14)$, $(6, 13)$, $(7, 12)$, $(8, 11)$.

(a) Find the value of r. [1]

(b) Describe the type and strength of the correlation. [1]

(c) One data point appears inconsistent with the general trend. Write down its coordinates. [1]

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

A café owner records daily high temperature T (°C) and the number of iced drinks D sold. For ten days she finds:

$$\bar{T} = 24.0, \quad \bar{D} = 86, \quad S_{TT} = 180, \quad S_{TD} = 1260$$

(a) Find the gradient a of the regression line $\hat{D} = aT + b$. [1]

(b) Find the y-intercept b. [1]

(c) Use the regression equation to predict the number of iced drinks sold when $T = 28\,^{\circ}\mathrm{C}$. [1]

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

Five cities are ranked by air quality index (AQI) and by public transport score. The table gives the raw data.

City	P	Q	R	S	T
AQI rank	1	2	3	4	5
Transport rank	3	1	2	5	4

(a) Calculate r_s using the formula $r_s = 1 - \dfrac{6\sum d_i^2}{n(n^2-1)}$. [2]

(b) Comment on whether the AQI rank and transport rank appear to be associated. [1]

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

A linear regression model is fitted to data on advertising spend x (hundreds of dollars) and weekly revenue y (thousands of dollars). The Pearson correlation coefficient is found to be $r = 0.874$.

(a) Write down the coefficient of determination R^2. [1]

(b) Interpret the value of R^2 in context. [1]

(c) State one reason why $r = 0.874$ alone does not confirm that increased advertising *causes* increased revenue. [1]

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

A botanist models the height h (cm) of a plant seedling using the number of days d after planting. The regression line of h on d is:

$$\hat{h} = 1.35d + 4.20$$

This model was built using data for d in the range $5 \leq d \leq 30$.

(a) Use the model to estimate the height of the seedling when $d = 20$ days. [1]

(b) Use the model to estimate the height when $d = 45$ days, and comment on the reliability of this estimate. [2]

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

A psychologist collects data on reaction time x (ms) and a memory-recall score y from nine participants. The scatter diagram shows one extreme outlier.

(a) State which correlation coefficient — Pearson's r or Spearman's r_s — would be more appropriate to use, and give **one** reason for your choice. [2]

(b) The nine participants are ranked. For one pair of participants the reaction times are equal. Explain how this tied rank should be handled when computing r_s. [1]

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

Two regression models are fitted to the same data set relating distance x (km) and delivery time y (minutes):

- **Model A** (linear): $R^2 = 0.762$
- **Model B** (quadratic): $R^2 = 0.891$

(a) Write down which model gives the better fit to the data. [1]

(b) For Model B, interpret the value $R^2 = 0.891$ in the context of the question. [1]

(c) Give **one** reason why selecting the model with the highest R^2 value is not always the best strategy. [1]

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

Scan me for help

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

A nutritionist records the daily sugar intake s (grams) and blood glucose level g (mmol/L) for eight patients. The data is given in the table.

Sugar, s (g)	30	45	50	60	70	80	95	110
Glucose, g (mmol/L)	4.2	4.8	5.1	5.5	5.8	6.2	6.9	7.4

(a) Find the value of Pearson's correlation coefficient r. [1]

(b) Find the equation of the regression line of g on s. [2]

(c) A patient has a daily sugar intake of 75 g. Use the regression line to estimate their glucose level. Comment on the reliability of your estimate. [1]

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

Seven students sit a theoretical exam (score out of 100) and a practical skills test (score out of 50). Their results are shown below.

Student	A	B	C	D	E	F	G
Theory score	88	73	91	64	79	55	83
Practical score	42	31	44	27	38	22	36

(a) Rank the students for each assessment (rank 1 = highest score) and calculate r_s. [3]

(b) The Pearson coefficient for this data is $r = 0.996$. State which coefficient is more appropriate here and give a reason. [1]

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

A physicist records the pressure P (kPa) of a gas and its volume V (litres) at constant temperature. The data is given below.

Volume V (L)	1.0	1.5	2.0	2.5	3.0	4.0
Pressure P (kPa)	300	200	150	120	100	75

(a) Using your GDC, fit a linear model and a power model ($P = aV^n$) to the data. Write down the value of R^2 for each model. [2]

(b) Write down the equation of the power model. [1]

(c) Hence, estimate the pressure when $V = 3.5$ L. Give your answer to 3 significant figures. [1]

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

A sociologist believes that the number of social media followers F of a celebrity grows exponentially with the number of years t since their first post, so that $F = ab^t$.

The following table shows the natural logarithm of F for four selected years.

Years since first post, t	1	2	3	4
ln F	10.82	11.51	12.21	12.90

(a) Find the equation of the regression line of ln F on t. Write the gradient and intercept correct to 3 significant figures. [2]

(b) Hence find the values of a and b in the model $F = ab^t$. [2]

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

An economist records the unemployment rate u (%) and average weekly consumer spending c (USD) across nine regions of a country.

$$r = -0.887, \qquad \bar{u} = 6.2, \qquad \bar{c} = 412$$

The regression line of c on u passes through $(\bar{u}, \bar{c})$ and has gradient $a = -18.3$.

(a) Write down the regression equation $\hat{c} = au + b$. [2]

(b) Estimate the consumer spending in a region where the unemployment rate is 4.5%. [1]

(c) State one limitation of using this regression line to predict consumer spending for a region with unemployment rate of 20%. [1]

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

A sports scientist records the training hours per week h and competition ranking k of eight athletes. Computing both coefficients gives:

$$r = 0.631, \qquad r_s = 0.857$$

(a) Explain why the values of r and r_s differ significantly in this data set. [2]

(b) Justify which coefficient is more reliable here, and state what it tells you about the relationship. [2]

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

A wildlife researcher believes the body mass M (kg) of a species of bird is related to its wingspan w (cm) by a power law $M = a\,w^n$. To verify this, she plots ln M against ln w for six birds.

ln w	3.00	3.30	3.50	3.70	3.91	4.09
ln M	−0.92	0.07	0.69	1.28	1.89	2.44

(a) Find the equation of the regression line of ln M on ln w. [2]

(b) Hence write down the values of a and n in the model $M = a\,w^n$. Give a correct to 3 significant figures. [2]

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

Data collected from ten city parks relates the area A (hectares) and the number of bird species B observed. Summary statistics give:

$$\bar{A} = 8.4, \quad \bar{B} = 23.6, \quad S_{AA} = 142, \quad S_{BB} = 410, \quad S_{AB} = 228$$

(a) Find the value of Pearson's r. [2]

(b) Find the equation of the regression line of B on A. [2]

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

Six wines are assessed by a sommelier (score out of 20) and by a chemical sensor (score out of 100). The scores are given below.

Wine	A	B	C	D	E	F
Sommelier score	16	14	18	12	14	11
Sensor score	82	71	90	58	69	52

(a) Note that wines B and E have equal sommelier scores. Assign appropriate ranks and compute r_s. [3]

(b) Interpret the value of r_s in context. [1]

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

A technology analyst records the market share y (%) of a new app and the number of months x since its launch. She fits two models using GDC regression:

- **Linear:** $\hat{y} = 1.84x + 2.10, \quad R^2 = 0.813$
- **Exponential:** $\hat{y} = 1.95\,e^{0.291x}, \quad R^2 = 0.964$

The data covers months $1 \leq x \leq 8$.

(a) State which model better fits the data, and justify your choice. [2]

(b) Use the better model to predict the market share at $x = 6$ months. [1]

(c) State one reason why predicting the market share at $x = 24$ months using this model would be unreliable. [1]

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

Scan me for help

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

A data set of six pairs (x_i, y_i) has the following summary statistics:

$$n = 6, \quad \sum x = 42, \quad \sum y = 78, \quad \sum x^2 = 364, \quad \sum y^2 = 1150, \quad \sum xy = 607$$

(a) Find the values of $\bar{x}$, $\bar{y}$, S_{xx}, S_{yy}, and S_{xy}. [2]

(b) Hence find the value of Pearson's correlation coefficient r. [1]

(c) A seventh data point $(10, k)$ is added to the data set. Given that the mean $\bar{x}$ of all seven values is now $\frac{52}{7}$, verify that this is consistent, and find the new $\bar{y}$ in terms of k. Then show that the regression line of y on x for the original six points passes through $(7, 13)$. [2]

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

Eight job applicants are assessed on a written test (rank 1 = best) and an interview (rank 1 = best). The rankings are given, but one interview rank is unknown.

Applicant	A	B	C	D	E	F	G	H
Test rank	1	2	3	4	5	6	7	8
Interview rank	3	1	4	2	6	p	8	7

The interview ranks are a permutation of $1, 2, \ldots, 8$.

(a) Write down the value of p. [1]

(b) Calculate r_s. [2]

(c) The critical value of r_s at the 5% significance level (two-tailed) for $n = 8$ is 0.738. Determine whether there is significant evidence of association between test rank and interview rank. [2]

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

A chemist investigates how the reaction rate R (μmol/s) depends on the concentration c (mol/L) of a reactant. She believes $R = k\,c^m$ for constants $k > 0$ and $m > 0$. The data below has been transformed by taking natural logarithms.

ln c	−2.30	−1.61	−0.92	−0.51	0.00
ln R	−0.51	0.26	1.04	1.48	2.12

(a) Find the equation of the regression line of ln R on ln c. [2]

(b) Hence write down the values of k and m, giving k correct to 3 significant figures. [2]

(c) Suggest one assumption required for the power law model $R = kc^m$ to be valid in this context. [1]

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

A data set of five bivariate values has:

$$\sum x = 25, \quad \sum y = 40, \quad \sum x^2 = 145, \quad \sum xy = 215, \quad \sum y^2 = Q$$

The Pearson correlation coefficient is $r = 0.900$ (correct to 3 significant figures).

(a) Find $\bar{x}$, $\bar{y}$, S_{xx}, and S_{xy}. [2]

(b) Show that $S_{yy} = \left(\dfrac{S_{xy}}{r}\right)^2 \div S_{xx}$, and hence find the exact value of S_{yy}. [2]

(c) Determine the value of $Q = \sum y^2$. [1]

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

Nine consumers rate a new food product on two attributes: *flavour* (1 = worst, 9 = best, all scores distinct) and *texture* (1 = worst, 9 = best, all scores distinct). Their rankings (not raw scores) are:

Consumer	1	2	3	4	5	6	7	8	9
Flavour rank	1	2	3	4	5	6	7	8	9
Texture rank	2	4	1	6	3	9	5	7	8

(a) Calculate r_s for this data. [2]

(b) The Pearson coefficient for the original (non-rank) scores is $r = 0.761$. Explain why r_s and r may differ even when computed on the same underlying data. [2]

(c) A tenth consumer gives a flavour rank of 5 and a texture rank of 5. Without recalculating r_s, explain what effect, if any, this addition would have on the ranks of the existing nine consumers. [1]

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

Six data points (x_i, y_i) lie on the regression line $y = 2.4x + 1.7$. The first five points are:

$$(1,\ 4.1),\ (2,\ 6.5),\ (3,\ 8.9),\ (4,\ 11.3),\ (5,\ 13.7)$$

The sixth point is $(6,\ y_6)$ where y_6 is unknown.

(a) Verify that the five known points lie exactly on the regression line. [1]

(b) The mean of all six y-values is $\bar{y} = 9.37$. Find y_6. [2]

(c) Compute S_{xy} and S_{xx} for all six pairs, and verify that the gradient of the regression line of y on x is 2.4. [2]

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

An environmental scientist models the concentration C (μg/m^3) of a pollutant at distance d (km) from a source. The following data is recorded.

Distance d (km)	0.5	1.0	1.5	2.0	3.0	4.0
Concentration C (μg/m^3)	480	240	160	120	80	60

Two models are proposed: $C = ad^n$ (power) and $C = \alpha e^{-\beta d}$ (exponential decay).

(a) Use your GDC to fit both models. Write down R^2 for each. [2]

(b) Write down the power model equation. [1]

(c) Use the power model to estimate C at $d = 5$ km. Give your answer to 3 significant figures. [1]

(d) Suggest one physical reason why the power model may be more appropriate than the exponential model for describing pollutant dispersion with distance. [1]

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

A researcher is investigating whether there is a linear relationship between hours of sunlight per day x and crop yield y (tonnes/hectare) for a sample of $n = 10$ farms. She calculates $r = 0.672$.

The critical value of r for a one-tailed test at the 5% significance level with $n - 2 = 8$ degrees of freedom is $r_{\text{crit}} = 0.5494$.

(a) State the null and alternative hypotheses for this test, using correct notation. [2]

(b) Perform the test and state the conclusion in context. [2]

(c) State one assumption required for the validity of this test. [1]

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

A geographer believes that the population density ρ (persons/km^2) of settlements decreases logarithmically with distance d (km) from the city centre, modelled by $\rho = a \ln d + b$.

The following data was collected:

Distance d (km)	1	2	5	10	20
Population density ρ (persons/km^2)	8200	7300	5800	4600	3400

(a) Explain which two quantities should be plotted in order to linearise this model. [1]

(b) Find the regression equation of ρ on $\ln d$, giving coefficients to 3 significant figures. [2]

(c) Estimate the population density at $d = 8$ km. [1]

(d) Give **one** reason why the model $\rho = a \ln d + b$ may eventually give unrealistic predictions. [1]

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

A professor computes $r = 0.481$ and $r_s = 0.845$ for the same set of eight bivariate data points relating exam mark x and project mark y. Investigating further, she discovers there is one extreme outlier in the data.

(a) Explain why r and r_s differ so substantially, referencing the effect of the outlier on each coefficient. [2]

(b) She removes the outlier and recomputes r for the remaining seven points, obtaining $r = 0.913$. The critical value for r (one-tailed, 5% level, $n - 2 = 5$ df) is $r_{\text{crit}} = 0.6694$. Determine whether there is significant evidence of positive linear correlation after removing the outlier, and state your conclusion. [2]

(c) Suggest a statistical reason why removing a data point must be justified carefully before a test is performed. [1]

Worked Solutions — Official Mark Scheme

Notation: **M1** = method mark (earned for correct method even with wrong numbers); **A1** = accuracy mark (dependent on preceding M); **R1** = reasoning mark; **(M1)/(A1)** = implied by subsequent correct work; **FT** = follow-through from earlier answer.

Section A — Easy EASY

Q1 — PMCC and Regression (r find, equation, gradient interpret) [EASY

] **(a)** Using GDC with data as listed: **(M1)**
$r = -0.992$ (3 sf) **A1**

Note: Accept -0.991 to -0.993. Full value: $r = -0.9919\ldots$

(b) Regression line y on x: **(M1)**
$\hat{y} = -2.78x + 89.4$ **A1**

Note: Accept $a = -2.78$ (full: $-2.7846\ldots$), $b = 89.4$ (full: $89.424\ldots$).

(c) The gradient -2.78 means that for each additional thousand steps walked per day, the resting heart rate decreases by approximately 2.78 beats per minute, on average. **R1**

Note: Must refer to context (steps and heart rate). Award R1 for any correct contextual interpretation referencing the units.

Q2 — Describe Correlation, Predict, Interpolation Comment [EASY

] **(a)** The correlation is strong and negative: as water temperature increases, the number of fish observed per hour decreases. **R1**

(b) $\hat{n} = -3.14(18) + 87.2 = -56.52 + 87.2 = 30.68\ldots \approx 31$ fish **A1**

(c) Interpolation, since $t = 18\,^\circ$C lies within the observed range $[14, 26]\,^\circ$C. **A1**
Note: Must state "interpolation" and a correct reason (within observed range).

Q3 — Spearman's r_s: Direct Ranks [EASY

] **(a)** d values (Coach − Algorithm): $1-2, 2-1, 3-4, 4-3, 5-6, 6-5 = -1, 1, -1, 1, -1, 1$ **M1**

$\sum d_i^2 = 1 + 1 + 1 + 1 + 1 + 1 = 6$

$r_s = 1 - \dfrac{6 \times 6}{6(36-1)} = 1 - \dfrac{36}{210} = 1 - 0.1714\ldots = 0.829$ (3 sf) **A1**

(b) The value $r_s = 0.829$ indicates a strong positive association between the coach's rankings and the algorithm's rankings: the two systems broadly agree on the relative ordering of athletes. **R1**

Q4 — Scatter Diagram Reading + r + Identify Outlier [EASY

] **(a)** Enter data $(1,18),(2,16),(3,17),(4,15),(5,14),(6,13),(7,12),(8,11)$ in GDC: **(M1)**

$r = -0.972$ (3 sf) **A1**

Note: Full value $r = -0.9722\ldots$

(b) Strong negative (linear) correlation. **A1**

(c) The point $(3,17)$ appears above the general trend (all other points follow the decreasing pattern, but at $a = 3$ the efficiency is 17 rather than the expected ≈ 15.7). **A1**

Q5 — Regression from Summary Statistics [EASY

] **(a)** $a = \dfrac{S_{TD}}{S_{TT}} = \dfrac{1260}{180} = 7.00$ **M1 A1**

(b) The regression line passes through $(\bar{T}, \bar{D}) = (24.0,\ 86)$:

$b = \bar{D} - a\bar{T} = 86 - 7.00 \times 24.0 = 86 - 168 = -82.0$ **A1**

(c) $\hat{D} = 7.00(28) + (-82.0) = 196 - 82 = 114$ iced drinks **A1**

Note: FT from their a and b in (a) and (b).

Q6 — Spearman's r_s from Given Ranks [EASY

] **(a)** d values: $1–3 = -2,\ 2–1 = 1,\ 3–2 = 1,\ 4–5 = -1,\ 5–4 = 1$ **M1**

$\sum d_i^2 = 4 + 1 + 1 + 1 + 1 = 8$

$r_s = 1 - \dfrac{6 \times 8}{5(25-1)} = 1 - \dfrac{48}{120} = 1 - 0.4 = 0.600$ **A1**

(b) There is a moderate positive association between AQI rank and transport rank; cities with better air quality tend to also have better public transport, though the relationship is not strong. **R1**

Q7 — Coefficient of Determination: Write Down + Interpret + Causation [EASY

] **(a)** $R^2 = (0.874)^2 = 0.764$ (3 sf) **A1**

Note: Accept 0.763 to 0.764.

(b) Approximately 76.4% of the variation in weekly revenue can be explained by the linear relationship with advertising spend. **R1**

(c) Correlation does not imply causation: a third variable (e.g. seasonal demand, economic conditions) could be responsible for both higher advertising and higher revenue. **R1**

Note: Accept any valid confounding factor. Do not accept merely "could be a coincidence".

Q8 — Regression Predict + Extrapolation Comment [EASY

] **(a)** $\hat{h} = 1.35(20) + 4.20 = 27.0 + 4.20 = 31.2$ cm **A1**

(b) $\hat{h} = 1.35(45) + 4.20 = 60.75 + 4.20 = 64.95 \approx 65.0$ cm **A1**

This estimate is obtained by extrapolation ($d = 45$ is outside the range $5 \leq d \leq 30$ used to build the model), so it is unreliable; the linear growth may not continue beyond the observed range. **R1**

Note: Must use the word "extrapolation" or equivalent, and state $d = 45$ is outside observed range.

Q9 — Choosing Between r and r_s [EASY

] **(a)** Spearman's r_s is more appropriate. **A1**

Reason: Pearson's r is heavily influenced by extreme outliers, so with one extreme outlier present, r would give a distorted picture of the overall association; Spearman's r_s, based on ranks, is more resistant to the effect of the outlier. **R1**

(b) The two participants with equal reaction times share ranks. Their ranks are averaged: if they are tied for ranks k and $k + 1$, each is assigned rank $k + 0.5$. **R1**

Q10 — Model Comparison via R^2 [EASY

] **(a)** Model B (quadratic), since $R^2 = 0.891 > 0.762$. **A1**

(b) 89.1% of the variation in delivery time is explained by the quadratic relationship with distance. **R1**

(c) A higher R^2 may be due to over-fitting: adding more parameters to a model always increases R^2 even if the extra complexity does not reflect a genuine relationship in the data; the model may not generalise well to new data. **R1**

Note: Accept references to over-fitting, model complexity, parsimony, or lack of predictive validity for new observations.

Section B — Medium MEDIUM

Q11 — PMCC + Regression + Predict + Reliability Comment [MED

] **(a)** GDC with $s = 30, 45, 50, 60, 70, 80, 95, 110$ and $g = 4.2, 4.8, 5.1, 5.5, 5.8, 6.2, 6.9, 7.4$: **(M1)**
$r = 0.999$ (3 sf) **A1**

Note: Full value $r = 0.99924\ldots$

(b) GDC gives: $a = 0.0381$ (full: $0.038128\ldots$), $b = 3.07$ (full: $3.0706\ldots$) **M1**
$\hat{g} = 0.0381s + 3.07$ **A1**

(c) $\hat{g} = 0.0381(75) + 3.07 = 2.858 + 3.07 = 5.93$ mmol/L **A1**

This is an interpolation ($s = 75$ is within $[30, 110]$) and r is very close to 1, so the estimate is reliable. **R1**

Note: Award A1FT for correct substitution using their equation. Award R1 for correct interpolation statement with reference to strength of correlation.

Q12 — Spearman's r_s Full Calculation + Comparison [MED

] **(a)** Ranking theory (rank 1 = highest): C(91)=1, A(88)=2, G(83)=3, F(79)... wait:

Theory ranks: C=1, A=2, G=3, E=4, B=5, D=6, F=7

Practical ranks: C=1, A=2, G=3, E=4, B=5, D=6, F=7 **M1 M1**

(Re-ordered theory scores: C=91, A=88, G=83, E=79, B=73, D=64, F=55. Practical: C=44, A=42, E=38, G=36, B=31, D=27, F=22.)

Differences d_i: all $d_i = 0$ for all 7 students.

$\sum d_i^2 = 0$

$r_s = 1 - \dfrac{6 \times 0}{7(49-1)} = 1 - 0 = 1.00$ **A1**

(b) Both $r = 0.996$ and $r_s = 1.00$ indicate near-perfect positive association. Pearson's r is appropriate here because the data are measured on numeric (interval) scales with no extreme outliers and the relationship appears linear; both coefficients are essentially equivalent in this case. **R1**

Note: Accept Spearman's is more appropriate if student correctly notes that the data are effectively rank-equivalent; award R1 for a valid justification.

Q13 — Non-linear Regression: Power vs Linear, Choose, Predict [MED

] **(a)** Linear model ($P = aV + b$): GDC gives $R^2 = 0.843$ **A1**

Power model ($P = aV^n$): GDC gives $R^2 = 1.000$ (or $R^2 = 0.9999\ldots$) **A1**

Note: For linear model accept R^2 in range $[0.84, 0.85]$. For power model R^2 must be ≥ 0.999.

(b) Power model: $P = 300\,V^{-1.00}$ (i.e. $P = \frac{300}{V}$) **A1**

Note: Accept $P = 300V^{-1}$. GDC output: $a = 300$, $n = -1.00$.

(c) $P = \dfrac{300}{3.5} = 85.7$ kPa (3 sf) **A1**

Q14 — Exponential Model via Linearisation [MED

] **(a)** GDC with $x = t = 1, 2, 3, 4$ and $y = \ln F = 10.82, 11.51, 12.21, 12.90$: **M1**

Gradient $= 0.693$ (3 sf), Intercept $= 10.1$ (3 sf) **A1**

$\ln F = 0.693\,t + 10.1$

Note: Full values: gradient $= 0.6933\ldots$, intercept $= 10.127\ldots$

(b) $\ln F = \ln a + t \ln b$, so comparing: **M1**

ln $b = 0.693 \Rightarrow b = e^{0.693} = 2.00$ (3 sf) **A1**

ln $a = 10.1 \Rightarrow a = e^{10.1} = 24\,300$ (3 sf)

Note: Full value $a = e^{10.127\ldots} = 25\,000$ (to 3 sf using full intercept). Accept a in range $[24000, 25200]$, b in range $[1.99, 2.01]$.

Q15 — Regression Equation from Summary + Predict + Limitation [MED

] **(a)** The line passes through $(\bar{u}, \bar{c}) = (6.2, 412)$ with gradient $a = -18.3$: **M1**

$b = 412 - (-18.3)(6.2) = 412 + 113.46 = 525.46\ldots \approx 525$

$\hat{c} = -18.3u + 525$ (3 sf) **A1**

(b) $\hat{c} = -18.3(4.5) + 525 = -82.35 + 525 = 442.65 \approx 443$ USD **A1**

(c) An unemployment rate of 20% is well outside the observed range of the data, so this would be extrapolation; the linear model may not remain valid at such extreme values, and the prediction is unreliable.**R1**

Note: Accept reference to the linear model breaking down, or to 20% being unrealistic within the data range. Do not accept bare "extrapolation" without explanation.

Q16 — r vs r_s Discrepancy: Outlier Effect [MED

] **(a)** Pearson's r is based on the actual data values and is sensitive to extreme outliers: a single outlier can pull r significantly toward zero or change its sign. Spearman's r_s is based only on ranks, so a single outlier receives an extreme rank but has limited leverage on the sum $\sum d_i^2$. **R1 R1**

(b) Spearman's $r_s = 0.857$ is more reliable here because it is more robust to the outlier. It indicates a strong positive monotonic relationship between training hours and ranking: athletes who train more tend to have better (lower) competition rankings. **R1 R1**

Note: Award R1 for correctly identifying r_s as more reliable with a valid reason. Award R1 for a correct contextual interpretation of $r_s = 0.857$.

Q17 — Power Law via Log-Log Linearisation [MED

] **(a)** GDC with x = ln w = $3.00, 3.30, 3.50, 3.70, 3.91, 4.09$ and y = ln M = $-0.92, 0.07, 0.69, 1.28, 1.89, 2.44$: **M1**

ln $M = 2.96$ ln $w - 9.80$ (3 sf) **A1**

Note: Full gradient $= 2.9634\ldots$, intercept $= -9.803\ldots$ Accept answers in range gradient $\in [2.96, 2.97]$, intercept $\in [-9.81, -9.80]$.

(b) Comparing ln $M = n$ ln w + ln a: **M1**

$n = 2.96$ (the gradient) **A1**

ln $a = -9.80 \Rightarrow a = e^{-9.80} = 5.52 \times 10^{-5}$ (3 sf)

Note: Accept a in range $[5.4 \times 10^{-5},\ 5.6 \times 10^{-5}]$.

Q18 — PMCC and Regression from Summary Statistics [MED

] **(a)** $r = \dfrac{S_{AB}}{\sqrt{S_{AA} \cdot S_{BB}}} = \dfrac{228}{\sqrt{142 \times 410}}$ **M1**

$= \dfrac{228}{\sqrt{58220}} = \dfrac{228}{241.29\ldots} = 0.945$ (3 sf) **A1**

(b) Gradient: $a = \dfrac{S_{AB}}{S_{AA}} = \dfrac{228}{142} = 1.606\ldots \approx 1.61$ (3 sf) **M1**

Intercept: $b = \bar{B} - a\bar{A} = 23.6 - 1.606\ldots \times 8.4 = 23.6 - 13.49\ldots = 10.1$ (3 sf) **A1**

$\hat{B} = 1.61A + 10.1$

Note: FT from their gradient. Accept b in range $[10.0, 10.2]$.

Q19 — Spearman with Tied Ranks [MED

] **(a)** Sommelier ranking (rank 1 = highest: 18, 16, 14, 14, 12, 11): **M1**

Wine C = rank 1, Wine A = rank 2, Wines B & E tie for 3rd and 4th $\rightarrow$ both get rank 3.5, Wine D = rank 5, Wine F = rank 6.

Sensor ranking: 90=1, 82=2, 71=3, 69=4, 58=5, 52=6
Wine C=1, A=2, B=3, E=4, D=5, F=6 **M1**

d values: C: $1-1=0$; A: $2-2=0$; B: $3.5-3=0.5$; E: $3.5-4=-0.5$; D: $5-5=0$; F: $6-6=0$

$\sum d_i^2 = 0+0+0.25+0.25+0+0 = 0.5$

$r_s = 1 - \dfrac{6 \times 0.5}{6(36-1)} = 1 - \dfrac{3}{210} = 1 - 0.01429\ldots = 0.986$ (3 sf) **A1**

(b) $r_s = 0.986$ indicates a very strong positive association between the sommelier scores and the sensor scores; the two assessment methods rank the wines in almost the same order. **R1**

Q20 — Model Comparison + Predict + Extrapolation Warning [MED

] **(a)** The exponential model is a better fit. **A1**

$R^2 = 0.964 > 0.813$, so the exponential model explains a greater proportion of the variation in market share. **R1**

(b) $\hat{y} = 1.95\,e^{0.291\times 6} = 1.95\,e^{1.746} = 1.95 \times 5.733\ldots = 11.2\%$ (3 sf) **A1**

(c) $x = 24$ is far outside the observed range $[1, 8]$; this is extrapolation, and the exponential model would predict an unrealistically large market share (exponential growth cannot continue indefinitely; market saturation limits growth). **R1**

Section C — Hard **HARD**

Q21 — Summary Statistics + r + Regression Line Verification [HARD

] **(a)** $\bar{x} = \frac{42}{6} = 7, \quad \bar{y} = \frac{78}{6} = 13$ **A1**

$S_{xx} = 364 - 6(7)^2 = 364 - 294 = 70$

$S_{yy} = 1150 - 6(13)^2 = 1150 - 1014 = 136$

$S_{xy} = 607 - 6(7)(13) = 607 - 546 = 61$ **A1**

(b) $r = \frac{61}{\sqrt{70 \times 136}} = \frac{61}{\sqrt{9520}} = \frac{61}{97.57\ldots} = 0.625$ (3 sf) **M1 A1**

Note: Full value $r = 0.6252\ldots$

(c) Gradient of regression line: $a = \frac{S_{xy}}{S_{xx}} = \frac{61}{70} = 0.871$ (3 sf) **M1**

Intercept: $b = \bar{y} - a\bar{x} = 13 - 0.871 \times 7 = 13 - 6.10 = 6.90$

At $x = 7$: $\hat{y} = 0.871(7) + 6.90 = 6.10 + 6.90 = 13.0 = \bar{y}$? **AG**

Since the regression line always passes through $(\bar{x}, \bar{y}) = (7, 13)$, the line passes through $(7, 13)$. **A1**

Q22 — Unknown Rank + r_s + Significance Test [HARD

] **(a)** The interview ranks must be a permutation of $1, \ldots, 8$. The given interview ranks are $3, 1, 4, 2, 6, p, 8, 7$. The missing value is $p = 5$. **A1**

(b) d_i = Test rank − Interview rank: **M1**
A: $1 - 3 = -2$; B: $2 - 1 = 1$; C: $3 - 4 = -1$; D: $4 - 2 = 2$; E: $5 - 6 = -1$; F: $6 - 5 = 1$; G: $7 - 8 = -1$; H: $8 - 7 = 1$

$\sum d_i^2 = 4 + 1 + 1 + 4 + 1 + 1 + 1 + 1 = 14$

$r_s = 1 - \frac{6 \times 14}{8(64 - 1)} = 1 - \frac{84}{504} = 1 - \frac{1}{6} = 0.833$ (3 sf) **A1**

(c) $H_0 : \rho_s = 0$ (no association); $H_1 : \rho_s \neq 0$ (association exists) **M1**

Since $r_s = 0.833 > 0.738 = r_{\text{crit}}$, we reject H_0. **R1**

There is significant evidence (at the 5% level) of a positive association between written test rank and interview rank for these applicants. **A1**

Q23 — Power Law via ln–ln: Regression + Back-transform + Assumption [HARD

] **(a)** GDC with $x = \ln c = -2.30, -1.61, -0.92, -0.51, 0.00$ and $y = \ln R = -0.51, 0.26, 1.04, 1.48, 2.12$: **M1**

Gradient $= 1.12$ (3 sf), Intercept $= 2.11$ (3 sf) **A1**

$\ln R = 1.12 \ln c + 2.11$

Note: Full values: gradient $= 1.121\ldots$, *intercept* $= 2.113\ldots$. *Accept gradient in* $[1.12, 1.13]$, *intercept in* $[2.11, 2.12]$.

(b) Comparing $\ln R = m \ln c + \ln k$: **M1**

$m = 1.12$ (reaction order) **A1**

$\ln k = 2.11 \Rightarrow k = e^{2.11} = 8.24$ μmol/s (3 sf)

Note: Accept k *in range* $[8.2, 8.3]$.

(c) The temperature is constant (isothermal conditions) throughout the experiment, so the rate constant k does not vary with the reaction conditions. **R1**

Q24 — Unknown $\sum y^2$ from Constrained r [HARD

] **(a)** $n = 5$: $\bar{x} = \dfrac{25}{5} = 5$, $\bar{y} = \dfrac{40}{5} = 8$ **A1**

$S_{xx} = 145 - 5(5)^2 = 145 - 125 = 20$

$S_{xy} = 215 - 5(5)(8) = 215 - 200 = 15$ **A1**

(b) From $r = \dfrac{S_{xy}}{\sqrt{S_{xx}\,S_{yy}}}$: **M1**

$$r^2 = \frac{S_{xy}^2}{S_{xx}\,S_{yy}} \implies S_{yy} = \frac{S_{xy}^2}{r^2\,S_{xx}} = \frac{15^2}{(0.9)^2 \times 20} = \frac{225}{0.81 \times 20} = \frac{225}{16.2} = \frac{125}{9}$$ **AG A1**

Note: The "Show that" identity $S_{yy} = (S_{xy}/r)^2/S_{xx}$ *follows immediately from* $r = S_{xy}/\sqrt{S_{xx}S_{yy}}$.

(c) $S_{yy} = \sum y^2 - n\bar{y}^2 \implies Q = S_{yy} + n\bar{y}^2 = \dfrac{125}{9} + 5(64) = \dfrac{125}{9} + 320 = \dfrac{125 + 2880}{9} = \dfrac{3005}{9} \approx 334$
A1

Note: Exact value $Q = \frac{3005}{9} = 333.8\overline{8}$. *Accept 334 (3 sf).*

Q25 — r_s Nine Points + Compare with r + Tied Rank Argument [HARD

] **(a)** Differences d_i (Flavour rank − Texture rank): **M1**

$(1-2), (2-4), (3-1), (4-6), (5-3), (6-9), (7-5), (8-7), (9-8)$
$= -1, -2, 2, -2, 2, -3, 2, 1, 1$

$\sum d_i^2 = 1 + 4 + 4 + 4 + 4 + 9 + 4 + 1 + 1 = 32$ **M1**

$r_s = 1 - \dfrac{6 \times 32}{9(81-1)} = 1 - \dfrac{192}{720} = 1 - \dfrac{4}{15} = \dfrac{11}{15} = 0.733$ (3 sf) **A1**

(b) Pearson's r is based on the actual numeric scores and measures the strength of the *linear* relationship; r_s is based on ranks and measures the strength of any *monotonic* relationship. Even when the underlying data is the same, r and r_s can differ if the relationship is monotonic but non-linear, or if the score scale is compressed or expanded unevenly (e.g. bunching of scores at one end). **R1 R1**

(c) Adding a tenth consumer does not change the existing nine ranks if the new scores are distinct from all current scores. However, the tenth consumer's flavour score of 5 duplicates the existing rank 5 position; both consumers would need new tied ranks of 5.5, changing all consumers originally ranked above 5 in flavour to be re-evaluated. **R1**

Note: Award R1 for recognising that the existing ranks would be disrupted because a score of rank 5 in flavour now ties with the new entrant.

Q26 — Verify Points on Line, Find y_6, Verify Gradient [HARD

] **(a)** Check: $y = 2.4(1)+1.7 = 4.1$✓; $y = 2.4(2)+1.7 = 6.5$✓; $2.4(3)+1.7 = 8.9$✓; $2.4(4)+1.7 = 11.3$✓; $2.4(5)+1.7 = 13.7$✓. All five known points lie on the line. **A1**

(b) $\bar{y} = \dfrac{4.1 + 6.5 + 8.9 + 11.3 + 13.7 + y_6}{6} = 9.37$ **M1**

$44.5 + y_6 = 56.22 \implies y_6 = 11.72 \approx 11.7$ (3 sf) **A1**

Note: $6 \times 9.37 = 56.22$; $56.22 - 44.5 = 11.72$.

(c) $\bar{x} = \dfrac{1+2+3+4+5+6}{6} = 3.5$ **M1**

$S_{xx} = (1^2 + \cdots + 6^2) - 6(3.5)^2 = 91 - 6(12.25) = 91 - 73.5 = 17.5$

$S_{xy} = (1 \cdot 4.1 + 2 \cdot 6.5 + 3 \cdot 8.9 + 4 \cdot 11.3 + 5 \cdot 13.7 + 6 \cdot 11.72) - 6(3.5)(9.37)$
$= (4.1 + 13 + 26.7 + 45.2 + 68.5 + 70.32) - 196.77 = 227.82 - 196.77 = 41.05\ldots$

Gradient $= \dfrac{S_{xy}}{S_{xx}} = \dfrac{41.05\ldots}{17.5} = 2.346\ldots$ **A1**

Note: The gradient ≈ 2.35, not exactly 2.4, because $y_6 = 11.72$ does not lie exactly on the line $y = 2.4x + 1.7$ (which would give $y_6 = 16.1$). Students should observe the discrepancy — the sixth point $(6, 11.72)$ lies off the trend, so it shifts the regression gradient. Award A1 for correct computation of gradient from their y_6.

Q27 — Non-linear Model Comparison + Predict + Physical Reasoning [HARD

] **(a)** Using GDC with $d = 0.5, 1, 1.5, 2, 3, 4$ and $C = 480, 240, 160, 120, 80, 60$: **M1**

Power model ($C = ad^n$): $R^2 = 0.999$ (accept ≥ 0.998) **A1**

Exponential model ($C = \alpha e^{-\beta d}$): $R^2 = 0.921$ (accept range $[0.91, 0.93]$) **A1**

Note: Award first A1 for both R^2 values stated; exact values depend on GDC. Power: $a = 240, n = -1.00$; Exp: $\alpha \approx 403, \beta \approx 0.386$.

(b) $C = 240\, d^{-1.00} = \dfrac{240}{d}$ **A1**

(c) $C = \dfrac{240}{5} = 48.0\ \mu\text{g/m}^3$ (3 sf) **A1**

(d) Pollutant concentration in the open air decreases inversely with distance due to spherical (or cylindrical) dilution of the emitted mass across an expanding area, which naturally produces an inverse power law rather than exponential decay. **R1**

Q28 — Hypothesis Test for PMCC using Critical Value [HARD

] **(a)** Let ρ denote the population correlation coefficient. **M1**

$H_0: \rho = 0$ (no linear correlation between sunlight hours and crop yield) **A1**

$H_1: \rho > 0$ (positive linear correlation) **A1**

(b) Since $r = 0.672 > 0.5494 = r_{\text{crit}}$, we reject H_0. **R1**

There is significant evidence at the 5% level that there is a positive linear correlation between daily hours of sunlight and crop yield. **A1**

Note: Do not award final A1 if conclusion is assertive ("proves") or not in context.

(c) The population of (x, y) pairs is bivariate normal (i.e. the data for both variables are approximately normally distributed). **R1**

Note: Accept "the relationship between x and y is linear in the population", or "the observations are independent".

Q29 — Logarithmic Model: Linearise, Regress, Predict, Critique [HARD

] **(a)** Plot ρ (population density) on the y-axis against ln d (natural log of distance) on the x-axis, to obtain a linear scatter. **R1**

(b) Compute ln d for $d = 1, 2, 5, 10, 20$: ln $d = 0, 0.693, 1.609, 2.303, 2.996$ **M1**

GDC linear regression of ρ on ln d:

Gradient $a = -1580$ (3 sf); Intercept $b = 8210$ (3 sf) **A1**

$\hat{\rho} = -1580 \ln d + 8210$

Note: Full values: $a = -1578\ldots$, $b = 8210\ldots$. Accept $a \in [-1590, -1570]$, $b \in [8200, 8220]$.

(c) $\hat{\rho} = -1580 \ln(8) + 8210 = -1580(2.0794\ldots) + 8210 = -3285\ldots + 8210 = 4920$ persons/km^2 (3 sf) **A1**

(d) As $d \to \infty$, ln $d \to \infty$, so $\hat{\rho} \to -\infty$; the model predicts negative population density at large distances, which is physically impossible. **R1**

Q30 — Outlier Effect on r vs r_s + Significance Test After Removal [HARD

] **(a)** Pearson's r uses actual data values: the outlier's extreme coordinates inflate the residuals and distort the covariance S_{xy}, pulling r toward zero. Spearman's r_s uses ranks, so the outlier simply occupies one extreme rank position (rank 1 or rank 8), giving it limited influence on $\sum d_i^2$, which is why $r_s = 0.845$

remains high while $r = 0.481$ is substantially reduced. **R1 R1**

(b) $H_0: \rho = 0$; $H_1: \rho > 0$. **M1**

Since $r = 0.913 > 0.6694 = r_{\text{crit}}$, we reject H_0. **R1**

There is significant evidence at the 5% level of a positive linear correlation between exam mark and project mark after removing the outlier. **A1**

(c) Removing a data point changes the sample composition; it must be justified on substantive grounds (e.g. data entry error, a participant who did not follow the assessment protocol) rather than simply because the point reduces correlation. Removing points selectively to improve correlation inflates the Type I error rate and invalidates the stated significance level. **R1**

Common Mistakes to Avoid

1. **Using y on x vs x on y.** Always use the regression line of the *dependent* variable on the *independent* variable for predictions. The two lines are different (unless $r = \pm 1$).

2. **Tied ranks.** When two or more values are equal, assign each the *average* of the ranks they would occupy (e.g. tied for 3rd and 4th $\Rightarrow$ both get rank 3.5). Forgetting this changes r_s.

3. **R^2 vs r.** For a linear model, $R^2 = r^2$, so R^2 is always non-negative. Quoting $R^2 = -0.87$ is nonsensical. Similarly, r can be negative but R^2 cannot.

4. **Interpolation vs Extrapolation.** Interpolation means the input is *within* the range of observed data — extrapolation means it is *outside*. Extrapolation is always less reliable; state this explicitly.

5. **Causation from correlation.** A strong r does not prove that x causes y. There may be confounding variables, reverse causation, or coincidence. Always phrase conclusions carefully.

6. **Back-transforming logarithmic models.** If $\ln y = at + b$, then $y = e^b \cdot (e^a)^t$, so $a = \ln(\text{base})$ and $b = \ln(\text{multiplier})$. A common error is writing $y = at + b$ directly.

7. **Hypothesis test conclusion wording.** Never write "Accept H_0" or "Prove H_1". Write: "There is/is not significant evidence of …" followed by the contextual interpretation.