Topic: Bivariate Data & Linear Regression
IB Math AI SL

Answer all questions. Show all working where appropriate. Total: 82 marks.

1. **[Paper 1 Style, Short Answer, Easy, 4 marks]**
A researcher records the number of hours of sunshine, x, and the number of ice creams sold, y, at a beach kiosk over 5 days.

Sunshine hours (x)	2	4	6	8	10
Ice creams sold (y)	12	20	26	36	46

   (a) Find the Pearson's product-moment correlation coefficient, r.

   (b) Describe the correlation between the number of sunshine hours and the number of ice creams sold.

2. **[Paper 1 Style, Short Answer, Easy, 5 marks]**
Refer to the data in Question 1.

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

   (b) Use your regression equation to estimate the number of ice creams sold on a day with 7 hours of sunshine.

3. **[Paper 1 Style, Short Answer, Easy, 4 marks]**
A set of bivariate data has a regression line y on x with the equation $y = 3.5x + 12$. The mean value of the x-data is $\overline{x} = 8$.

   (a) State the exact coordinates of the mean point, $M(\overline{x}, \overline{y})$, for this data set.

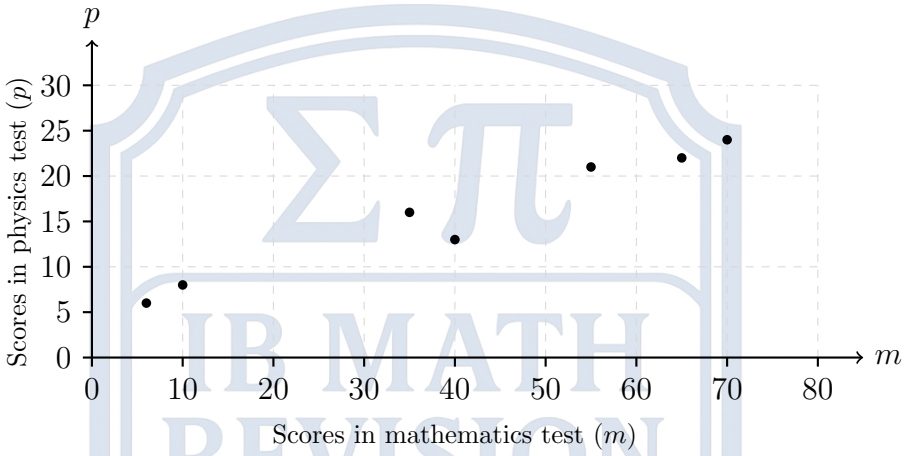
   (b) Explain the mathematical significance of the mean point in relation to the line of best fit.

4. **[Paper 1 Style, Short Answer, Medium, 4 marks]**
A student investigates the relationship between the shoe size of primary school children and their reading speed in words per minute. The student calculates a Pearson's correlation coefficient of $r = 0.89$.

   (a) State the direction and strength of the linear correlation.

   (b) The student concludes that having larger feet causes children to read faster. Explain why this conclusion is invalid, suggesting a possible lurking variable.

5. **[Paper 1 Style, Short Answer, Medium, 6 marks]**
The scatter diagram below shows the test scores of seven students in Mathematics (m) and Physics (p). The mean point, M, for these data is $(40, 16)$.



   (a) Plot and clearly label the mean point $M(\overline{m}, \overline{p})$ on the scatter diagram.

   (b) Draw a line of best fit, by eye, on the scatter diagram.

   (c) Using your line of best fit, estimate the physics test score for a student with a score of 20 in their mathematics test.

6. **[Paper 1 Style, Short Answer, Medium, 5 marks]**
The data below shows the number of hours five students spent playing video games (x) and their scores on a final exam (y). The regression line equation is $y = -2.8x + 85$. The data was collected from students who played video games for between 2 and 12 hours.

   (a) Use the regression equation to estimate the exam score of a student who played video games for 25 hours.

   (b) State whether your estimate in part (a) is reliable. Justify your answer.

7. **[Paper 2 Style, Longer Question, Medium, 5 marks]**
The table below shows the length of a spring (y, in cm) when different masses (x, in grams) are attached to it.

Mass (x)	10	20	30	40
Length (y)	15.5	18.0	20.5	23.0

   (a) Find the equation of the regression line y on x.

   (b) State the physical meaning of the y-intercept in the context of this problem.

   (c) State the physical meaning of the gradient (slope) in the context of this problem.

8. **[Paper 2 Style, Longer Question, Medium, 6 marks]**
An environmentalist models the relationship between the distance from a highway (d, in metres) and the concentration of a pollutant (C, in parts per million). The equation of the regression line is $C = -0.04d + 8.5$.

   (a) Estimate the concentration of the pollutant at a distance of 50 metres from the highway.

   (b) The actual concentration measured at 50 metres was 7.1 parts per million. Calculate the percentage error of the estimate.

9. **[Paper 1 Style, Short Answer, Hard, 4 marks]**
The weights of 10 athletes and the weights of their corresponding bicycles are recorded. The Pearson's correlation coefficient is found to be $r = 0.76$. The researcher realises that the bicycle weights were recorded in kilograms, but they should have been recorded in grams (by multiplying every bicycle weight by 1000).

   (a) State the new value of r after the bicycle weights are converted to grams.

   (b) Give a mathematical reason for your answer to part (a).

10. **[Paper 2 Style, Longer Question, Hard, 6 marks]**
The equation of the regression line of y on x for a dataset is $y = 1.4x - 5.2$. The equation of the regression line of x on y for the exact same dataset is $x = 0.6y + 4.8$.

    (a) Explain how you can use the two regression lines to find the exact coordinates of the mean point, $(\overline{x}, \overline{y})$.

    (b) Hence, calculate the exact values of $\overline{x}$ and $\overline{y}$.

11. **[Paper 1 Style, Short Answer, Hard, 4 marks]**
A study measures the population size of a town (x, in thousands) and the number of hospitals (y). The data contains a significant outlier where one town has a population of 500,000 but only 1 hospital.

(a) State whether Pearson's product-moment correlation coefficient (r) or Spearman's rank correlation coefficient (r_s) is more appropriate to measure the correlation of this data.

(b) Give a reason for your choice in part (a).

12. **[Paper 1 Style, Short Answer, Hard, 5 marks]**
The table shows the number of hours of sunshine (x) and ice cream sales in thousands of dollars (y) for 6 days.

Sunshine (hrs), x	2	5	3	8	6	4
Sales (\$1000), y	4	7	5	10	8	6

(a) Use your graphic display calculator to calculate Pearson's correlation coefficient, r.

(b) Interpret this value in the context of the problem, referencing both direction and strength.

13. **[Paper 2 Style, Longer Question, Very Hard, 6 marks]**
A group of 7 adult men wanted to see if there was a relationship between their waist size (x, in cm) and their Body Mass Index (y). The data is shown in the table below.

Waist (x cm)	58	63	75	82	93	98	105
BMI (y)	19	20	22	23	25	24	26

The relationship between x and y can be modelled by the regression equation $y = ax + b$.

(a) Write down the value of a and of b.

(b) Find the Pearson's correlation coefficient, r.

(c) Use the regression equation to estimate the BMI of an adult man whose waist size is 95 cm.

14. **[Paper 1 Style, Short Answer, Very Hard, 4 marks]**
The coefficient of determination, R^2, measures the proportion of the variance in the dependent variable that is predictable from the independent variable.

(a) A set of bivariate data is found to have a Pearson's correlation coefficient of $r = -0.80$. Calculate the coefficient of determination, R^2.

(b) State the percentage of the variance that is explained by this linear model.

15. **[Paper 1 Style, Short Answer, Very Hard, 4 marks]**
Pearson's product-moment correlation coefficient (r) is used to measure the strength of linear relationships.

(a) State the range of possible values for r.

(b) An analysis of two variables yields $r = 0.02$. State what this implies about the relationship between the two variables.