Unit 4: Bivariate Data & Linear Correlation
IB Math AA SL

Answer all 15 questions. Show all working. For Paper 1 questions, use analytical methods. For Paper 2 questions, use your graphic display calculator (GDC) efficiently.

1. **[Paper 2 Style, Calculator Required, Easy, 4 marks]**
The following table shows the mean weight, y kg, of primary school children who are x years old.

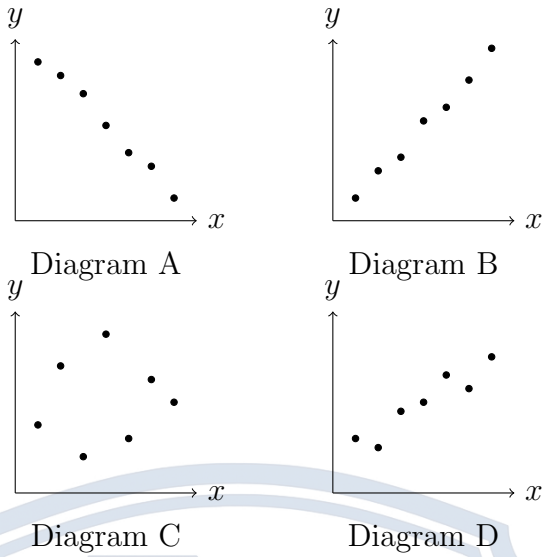
Age (x years)	6.25	7.35	8.50	9.25	10.75
Mean Weight (y kg)	21.5	23.1	26.9	28.6	32.0

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

(a) Find the value of a and the value of b.

(b) Write down the value of Pearson's product-moment correlation coefficient, r.

2. **[Paper 1 Style, Non-Calculator, Easy, 4 marks]**
Four scatter diagrams are shown below, labelled A, B, C, and D.



Match each of the following values of Pearson's product-moment correlation coefficient, r, to the correct scatter diagram:

$$r = 0.98, \quad r = -0.99, \quad r = 0.05, \quad r = 0.82$$

3. **[Paper 2 Style, Calculator Required, Easy, 5 marks]**
Jennifer sells cups of tea at her shop and notices she sells more tea on cooler days. Over five days, she records the maximum daily temperature, T in °C, and the number of cups of tea sold, C.

Temperature, T (°C)	3	5	8	9	12
Cups of tea sold, C	37	34	33	26	21

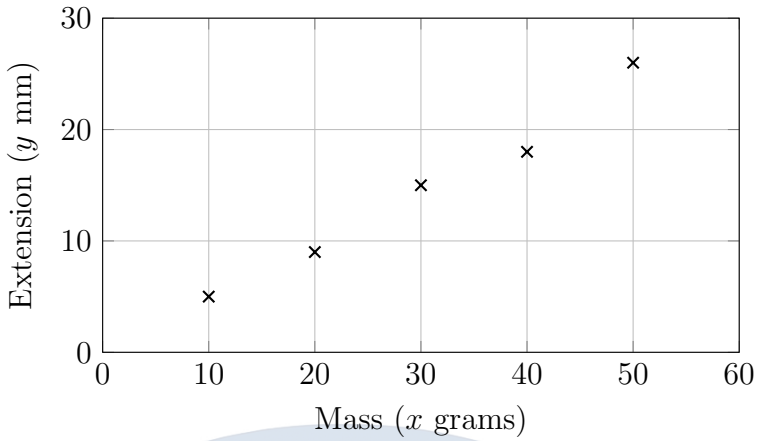
The relationship is modelled by the regression line $C = aT + b$.

(a) Find the value of a and of b.

(b) Use your regression equation to estimate the number of teas Jennifer will sell on a day when the maximum temperature is 11°C.

(c) State the type of correlation shown by this data.

4. **[Paper 2 Style, Calculator Required, Easy, 5 marks]**
The data from a physics experiment measuring the extension of a spring, y mm, when a mass of x grams is attached, is shown in the scatter diagram below.



The mean mass is $\bar{x} = 30$ g and the mean extension is $\bar{y} = 14.6$ mm.

(a) On the scatter diagram, carefully plot and label the mean point $M(\bar{x}, \bar{y})$.

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

(c) Use your line of best fit to estimate the extension when a mass of 35 g is attached.

5. **[Paper 1 Style, Non-Calculator, Medium, 4 marks]**
The equation of a regression line linking the number of hours a student spends studying for a math exam (x) and their final score out of 100 (y) is given by $y = 5.2x + 34.5$.

(a) Interpret the meaning of the parameter 5.2 in the context of the question.

(b) Interpret the meaning of the parameter 34.5 in the context of the question.

(c) A student concludes that "studying for exactly 10 hours *causes* your score to increase by 52 marks." Explain why this statement is statistically flawed.

6. **[Paper 2 Style, Calculator Required, Medium, 5 marks]**
A study measures the average number of hours of sleep per night (x) and the score on a short-term memory test (y) for 6 participants.

Participant	A	B	C	D	E	F
Hours of sleep (x)	6.8	7.2	8.1	9.4	5.9	7.5
Test score (y)	72	70	82	79	62	80

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

(b) Use the regression equation to estimate the memory test score of a participant who sleeps for 10.5 hours per night.

(c) Comment on the reliability of your estimate in part (b), giving a clear statistical reason.

7. **[Paper 2 Style, Calculator Required, Medium, 6 marks]**
A coach investigates the relationship between a rugby player's weight (x kg) and their maximum sprint velocity (y ms^{-1}).

Weight (x kg)	85	88	95	96	98	112
Velocity (y ms^{-1})	5.8	10.1	8.8	7.5	6.9	6.1

(a) Calculate the Pearson's product-moment correlation coefficient, r, for all 6 players.

(b) The coach realizes that the player weighing 88 kg with a velocity of 10.1 ms^{-1} had an exceptionally strong tailwind during their sprint and removes this data point as an outlier. Calculate the new value of r for the remaining 5 players.

(c) Comment on how the removal of this outlier affected the strength and direction of the linear correlation.

8. **[Paper 2 Style, Calculator Required, Medium, 6 marks]**
A movie cinema records the daily revenue from popcorn (x, in dollars) and the daily revenue from drinks (y, in dollars) over 6 days.

Popcorn revenue (x)	30.80	78.20	102.50	132.40	154.80	200.50
Drinks revenue (y)	60.70	202.10	308.50	198.70	368.20	300.00

(a) Calculate the mean daily revenue from popcorn, $\bar{x}$, and the mean daily revenue from drinks, $\bar{y}$.

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

(c) Show algebraically that the mean point $(\bar{x}, \bar{y})$ lies exactly on your regression line.

9. **[Paper 1 Style, Non-Calculator, Hard, 5 marks]**
For a dataset containing $n = 10$ pairs of bivariate data (x, y), the equation of the regression line of y on x is given by $y = 3.2x - 14.5$. Given that the sum of all the x-values is $\sum x = 60$:

   (a) Find the mean of x, $\bar{x}$.

   (b) Use the properties of the regression line to find $\bar{y}$.

   (c) Hence, find the exact sum of all the y-values, $\sum y$.

10. **[Paper 2 Style, Calculator Required, Hard, 6 marks]**
The table below shows the distance, D in km, to 5 ferry destinations and the corresponding price of the cruise, P in euros (€).

Distance, D	174	620	730	810	933
Price, P (€)	30.50	65.00	45.75	125.00	85.50

   (a) Find the equation of the regression line of P on D.

   (b) Find the equation of the regression line of D on P.

   (c) A customer has exactly €100 to spend on a cruise. Explain which regression line from parts (a) or (b) should be used to estimate the maximum distance they can travel, and use it to find this distance.

11. **[Paper 2 Style, Calculator Required, Hard, 5 marks]**
Ten students are asked to give a score from 1 to 10 on how much they enjoy watching football (x) and cricket (y). The regression line of y on x is calculated to be $y = -0.85x + 9.5$. The mean score for football is $\bar{x} = 5.6$. A new student joins the class and gives a football score of 8. Unfortunately, the survey system fails to record their cricket score.

   (a) Estimate the missing cricket score for this new student using the regression line.

   (b) A researcher claims that if we sum the y-values of all 10 original students, the total must be 47.4. Verify whether this claim is correct, showing your working.

12. **[Paper 2 Style, Calculator Required, Hard, 5 marks]**
Body Mass Index (BMI) is an indicator used to categorize weight. The following table shows the BMI (x) and body fat percentage (y) for 6 male participants.

BMI (x)	18.1	19.8	22.4	25.5	29.8	31.2
Body fat % (y)	15.1	20.1	22.1	24.2	31.1	16.2

(a) Calculate Pearson's r.

(b) Notice the participant with a BMI of 31.2 has a very low body fat percentage (16.2%). They are a professional bodybuilder. State the statistical term for this data point and briefly explain its effect on the correlation coefficient.

(c) Remove the bodybuilder from the dataset and find the new equation of the regression line y on x.

13. **[Paper 1 Style, Non-Calculator, Very Hard, 6 marks]**
A bivariate dataset consists of variables x and y. The regression line of y on x is given by:

$$L_1: \quad 3x - 4y + 12 = 0$$

The regression line of x on y is given by:

$$L_2: \quad 5x - 8y + 16 = 0$$

(a) Find the exact coordinates of the mean point $(\bar{x}, \bar{y})$.

(b) By rearranging both equations into the forms $y = m_1x + c_1$ and $x = m_2y + c_2$, find the exact value of the product m_1m_2.

(c) State what the value of m_1m_2 tells you about the correlation coefficient r.

14. **[Paper 2 Style, Calculator Required, Very Hard, 6 marks]**
Five points are plotted on a scatter diagram. The coordinates of four of the points are:

$$(2, 5), \quad (4, 9), \quad (5, 11), \quad (7, 14)$$

The fifth point is (p, q). The equation of the regression line of y on x for all five points is $y = 1.9x + 1.2$. The mean of the x-values is exactly $\bar{x} = 5$.

(a) Find the value of p.

(b) Find the value of q.

(c) Calculate the correlation coefficient r for the complete set of five points.

15. **[Paper 2 Style, Calculator Required, Very Hard, 7 marks]**
The average weekly temperature x (°C) and the energy consumption y (kWh) of an air conditioning unit were recorded over several weeks. The regression line of y on x was calculated to be $y = 12.5x + 150$. The correlation coefficient was $r = 0.88$. The energy company decides to change the units. They convert the temperature to Fahrenheit (F) using the formula $F = 1.8x + 32$, and they convert the energy to Joules (J) using $J = 3.6 \times 10^6 y$.

(a) State the new correlation coefficient between F and J. Give a reason for your answer.

(b) Find the new equation of the regression line of J on F, giving your answer in the form $J = aF + b$.