4.3 Further Correlation & Regression

Question Paper

Course	DP IB Maths
Section	4. Statistics & Probability
Topic	4.3 Further Correlation & Regression
Difficulty	Very Hard

Time allowed: 120

Score: /93

Percentage: /100

Question 1a

Effie is carrying out some research into dragon egg incubation patterns. She studies five dragon eggs and records the number of days it takes for the egg to hatch and the length of the wingspan on the newly hatched dragon. Some of Effie's data is displayed in the following table.

Incubation time, t days	22	58	41	13	96	11
Wingspan length, l cm	52.368	x	59.151	y	34.016	16.761

Effie uses a least squares regression curve in the form $l = at^3 + bt^2 + ct - 52$ to model the data and calculates the sum of the squared residuals, SS_{res}, to be zero.

a)
Explain what a value of $SS_{res} = 0$ means about using this least squares regression curve to model the data and hence write down the coefficient of determination.

[1 mark]

Question 1b

b)
Use the data given in the table to find the values of a, b and c and hence, find the values of x and y.

[4 marks]

Question 1c

Effie manages to collect data on one more dragon egg which has an incubation period of 33 days. Using the same model Effie now calculates the value of the sum of the squared residuals to be 0.0441.

c)
Find the two possible wingspan lengths of this newly hatched dragon.

[3 marks]

Question 2a

Gromit is researching the perfect basketball shot to get the basketball into a net of height 3.5 metres. He records himself taking ten shots from the same position and uses tracking software to measure the vertical and horizontal distances from the net at time t seconds after he shoots the ball. The mean distances are recorded in the table below. A positive value means the ball is above or in front of the net and a negative value means the ball is below or behind the net.

Time, t seconds	0.0	0.3	0.6	0.9	1.2	1.5	1.8
Horizontal Distance, x (m)	2.52	2.44	2.07	1.12	0.28	0.01	-0.02
Vertical Distance, y (m)	-2.05	0.56	2.98	4.09	2.23	0.23	-0.08

a)
Find the height from which Gromit throws the ball.

[2 marks]

Question 2b

b)
By first finding the equation of a quadratic, cubic and quartic least squares regression curve, investigate which is best for Gromit to use to model the trajectory of the ball. Give a reason why each equation is either suitable or unsuitable.

[6 marks]

Question 3a

Paul moors his sailboat in a harbour and wants to come up with a model for the daily tidal pattern at a certain time of year. He collects data on the depth of the water at every hour over a 24–hour period and uses it to calculate a least squares regression curve in the form $D = a\sin\left(\frac{\pi}{12}(t-2)\right) + b$, where D is the water depth in metres and is the number of hours after midnight. Some of his data is given in the table below.

Time, *t* hours	3	8	16	20
Depth, *D* metres	x	12.87	7.12	4.93
Residual	0.065	-0.13	0.12	y

a)
Use the information in the table to find the values of a and b.

[4 marks]

Question 3b

b)
Hence, find the values of x and y.

[3 marks]

Question 3c

c)
Find an estimate for the maximum and minimum tidal depths and the times at which they should occur.

[2 marks]

Question 4a

Jusef is modelling the height of some water as it drains out of a conical tank. He believes that the vertical distance of the water in the tank from the top of the tank, H cm over time, t seconds, can be modelled using the equation $H = Ab^t$, where $A, b \in \mathbb{R}$.

Time, t seconds	1	5	7	10	12	14	15
Distance from top, H cm	0.1	1.3	2.2	5.1	9.4	16.2	26.0

a)
Find the equation of the least squares regression curve and hence write down the values of A and b.

[3 marks]

Question 4b

b)
Use mathematical reasoning to comment on the suitability of the model found in part (a).

[2 marks]

Question 4c

Jusef linearises the data and models it using a semi-log graph.

c)
Calculate
(i)
the equation of the least squares regression line for the semi-log graph,
(ii)
the Pearson's product-moment correlation coefficient, r, and comment on the validity of the regression line found in part (c) (i).

[4 marks]

Question 5a

Oakley records the vertical height, h in metres, of a red kite in flight over the course of 20 seconds. He wishes to find an equation that will model the movement of the red kite.

The height measurements of some points during the flight of the red kite are shown in the table below.

Time, *t*	0.5	2.2	4.7	8.2	12.1	14.6	16.3	17.7	19.1
Height, *h*	19.3	19.2	19.1	18.7	18.2	1.9	1.8	1.4	1.5

a)
Suggest a reason why Oakley may choose to use a logistic function to model the flight of the red kite.

[1 mark]

Question 5b

b)
Use the data given to find an appropriate logistic model for the flight of the red kite, giving limits for the domain of this model.

[3 marks]

Question 5c

c)
Use the model to estimate the height of the red kite at $t = 9.8$ seconds.

[1 mark]

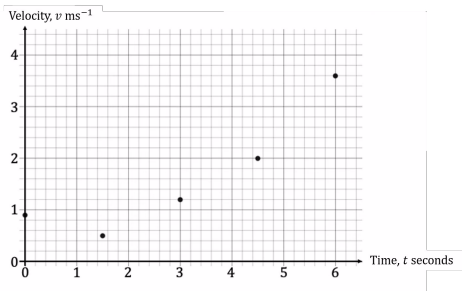
Question 5d

d)
Oakley noted that the red kite was level with his window at a certain point on its flight. If Oakley's window is 11.1 metres above ground level, find an approximation for the time at which the red kite was level with his window.

[1 mark]

Question 6a

The velocity, $v\ \text{ms}^{-1}$ of a vehicle was recorded over 6 seconds and the results given in the table and plotted in the velocity time graph below.



a)
Use the trapezoidal rule to find an estimate for the distance travelled by the vehicle.

[3 marks]

Question 6b

b)
Explain whether the answer found in part (a) will be an over or underestimate of the actual distance travelled.

[2 marks]

Question 6c

c)
Use the points on the graph to find the equation of the least squares quadratic regression curve.

[3 marks]

Question 6d

d)
Write down the R^2 value and interpret what this tells you about the model.

[2 marks]

Question 6e

e)
Use the answer found in part (c) to find a better estimate for the distance travelled by the vehicle.

[2 marks]

Question 7a

Each year during the dry season in a particular country a man-made trapezoidal reservoir empties completely and then begins to refill when the rains reappear. In 2021 Bea collected information about the height of the water level, a certain number of days after the beginning of the rainy season. Bea believes thatthe vertical height of the water from the bottom of the reservoir, *H* metres, can be modelled using the equation $H = At^b$, where $A, b \in \mathbb{R}$ and *t* is the number of days since the beginning of the rainy season.

Time, t days	1	2	3	5	10	20	40
Height, H m	1.5	3.8	5.7	8.3	14.5	24.9	30.0

a)
Find the equation of the least squares regression curve and hence write down the values of A and b.

[3 marks]

Question 7b

b)
Use mathematical reasoning to comment on the suitability of the model found in part (a).

[2 marks]

Question 7c

Bea wants to linearise the data using logarithms.

c)
(i)
Suggest, with a reason, whether Bea should use a semi-log graph or a log-log graph to linearise her data.

(ii) For this graph, find the equation of the regression line.

[3 marks]

Question 8a

A researcher in a particular fitness centre has been collecting data on the number of sleeveless T-shirts sold per week, T, and the number of new gym memberships per week, G. The data is shown in the table below.

T	119	54	92	25	442	340	9	261
G	50	15	25	12	129	22	8	21

The researcher suspects that T and G are related in one of two ways:

$$T = aG^m \text{ or } T = bp^G$$

where a, b, m and p are constants.

a)
By finding the values of a, b, m, p and the coefficient of determination for the two different models, decide which better represents the relationship between T and G.

[4 marks]

Question 8b

b)
Without carrying out any further calculations,
(i)
Write down which model has a higher value of the product moment correlation coefficient and give a reason for your answer,
(ii)
explain whether it would be best to represent the data on a semi-log graph or a log-log graph.

[3 marks]

Question 8c

c)
Hence, calculate the value of the product moment correlation coefficient for the chosen model in part (a).

[1 mark]

Question 9a

When brewing beer the temperature that the beer is stored at during fermentation, $T°C$, changes the alcohol content, A %, at the end of the fermentation process. A group of brewers collect data on T and A for their casks of beer. They suspect the data follows a model of the form $A = bp^T$ where b and p are unknown constants. They plot the regression line of $y = \ln A$ on $x = T$ and find that the line has a gradient of 0.0392 and passes through the point $(0, 0.811)$.

a)
Using the line of regression, calculate the values of b and p.

[5 marks]

Question 9b

In the data collected by the brewers, the range of values for T was 15 and the range of values for A was 4. The minimum alcohol content occurred when the temperature was at its minimum and the maximum alcohol content occurred when the temperature was at its maximum.

b)
Find estimates for the minimum values of T and A to 2 significant figures.

[6 marks]

Question 9c

c)
Hence explain why it would not be appropriate to use the model to predict the alcohol content of beer when the temperature during fermentation is $50°C$.

[1 mark]

Question 10a

Nora is building a skate ramp for a project at school. She models the first part of the ramp using the coordinates $A(0, 5)$, $B(1, 8)$, $C(5, 15)$ and $D(10, 18)$. These four points fit a cubic model in the form $y = ax^3 + bx^2 + cx + d$.

a)
Find the model for this section of Nora's skate ramp, giving a, b, c and d as exact values. Hence, write down the value of the sum of the squares residual.

[3 marks]

Question 10b

Nora models the next part of her skate ramp from the point D using a least squares regression curve with equation

$$y = a\cos 0.1(x - 10) + b$$

b)
Given that the skate ramp goes through the point $E(23,7)$ with a residual of -0.0124824 to six significant figures, find the values of a and b.

[4 marks]

Question 10c

The final section of Nora's skate ramp goes through the coordinates $F(30, 2)$, $G(35, 0)$, $H(40, 2)$ and $I(60, 15)$. Nora models this using a quadratic model of $y = 0.028x^2 - 2.1x + 40$.

c)
Find the equation of the least squares trigonometric curve for points F, G, H and I.

[2 marks]

Question 10d

d)
By examining the sum of the squares of the residuals, show that the model found in part (c) is a better fit for the data than Nora's quadratic model.

[4 marks]