A Level H2 Math

Correlation and Linear Regression Test 2

Q1

The table below shows the petrol mileage, y km/L and the weight, x kg in thousands for various car models in the year 1995.

x	3.5	3	2.75	2.5	2.25	2	1.75	1.5	1.25
y	7.5	8.0	8.5	8.7	10.0	k	13.5	16.8	18.0

(i) The equation of the regression line of y on x is $y = 22.51355 - 4.908387x$. Show that $k = 11.0$. [2]

(ii) Draw a scatter diagram to illustrate the data. [1]

(iii) With reference to the scatter diagram and context of the question, explain why model (C) below is the most appropriate for modelling the data as compared to the other 2 models.

(A) $y = a + bx$, where a is positive and b is negative,

(B) $y = a + b\ln x$, where a is positive and b is negative,

(C) $y = a + \frac{b}{x}$, where a and b are positive. [1]

(iv) Calculate the least squares estimates of a and b for model (C). [1]

(v) Predict the weight of the car if the petrol mileage is 12 km/L. Comment on the reliability of your prediction. [2]

(vi) Suppose there was an error in recording the y values and all the y values must be increased by a constant M km/L, state any change you would expect in the values of

(a) $\overline{y}$, [1]

(b) standard deviation of y and [1]

(c) the correlation coefficient. [1]

Q2

In the fishery sciences, researchers often need to determine the length of a fish as a function of its age. The table below shows the average length, L inches, at age, t years, of a kind of fish called the North Sea Sole.

t	1	2	3	4	5	6	7	8
L	3.6	7.5	10.1	11.7	12.7	13.4	14.0	14.4

(i) Draw a scatter diagram of these data, and explain how you know from your diagram that the relationship between L and t should not be modelled by an equation of the form $L = at + b$. [3]

(ii) Which of the formulae $L = a\sqrt{t} + b$ and $L = c \ln t + d$, where a, b, c and d are constants, is the better model for the relationship between L and t? Explain fully how you decided, and find the constants for the better formula. [3]

(iii) Use the formula you chose from part **(ii)** to estimate the average length of a six-month old Sole. Explain whether your estimate is reliable. [2]

A popular approach to determine the average length of a fish as a function of its age is the von Bertalanffy model. The model shows the relationship between the average length that is yet to be grown, G inches, at age, t years. The maximum average length attained by the Sole is 14.8 inches.

(iv) The product moment correlation between L and t is given as r_1 while that between G and t is given as r_2. State the relationship between r_1 and r_2. [1]

Q3

A retail manager of a large electrical appliances store wants to investigate the relationship between the monthly advertising expenditure, x hundred dollars, and the monthly sales of their refrigerators, y thousand dollars. The table below shows the results of the investigation.

x	5	8	12	16	18	20	23
y	12.5	12.9	13.6	14.8	17.0	19.3	25.1

(i) The manager concludes that an increase in monthly advertising expenditure will result in an increase in the monthly sales of refrigerators. State, with a reason, whether you agree with his conclusion. [1]

(ii) Draw a scatter diagram to illustrate the above data. Explain why a linear model is not likely to be appropriate. [2]

It is thought that the monthly sales y thousand dollars can be modelled by one of the formulae

$$y = a + b\mathrm{e}^{\sqrt{x}} \quad \text{or} \quad y = a + bx^2$$

where a and b are constants.

(iii) Find, correct to 4 decimal places, the value of the product moment correlation coefficient between

(A) $\mathrm{e}^{\sqrt{x}}$ and y,

(B) x^2 and y.

Explain which of $y = a + b\mathrm{e}^{\sqrt{x}}$ or $y = a + bx^2$ is the better model. [2]

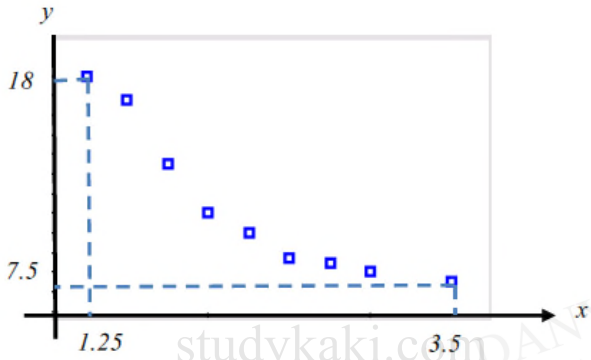
Assume that the better model in part **(iii)** holds for part **(iv)**.

(iv) The manager forgot to record the monthly advertising expenditure when the monthly sales of refrigerators was \$11300. Combining this with the above data set, it is found that $a = 10.876$ and $b = 0.09906$ for the model. Find the monthly advertising expenditure that the manager forgot to record, leaving your answer to the nearest hundred. [3]

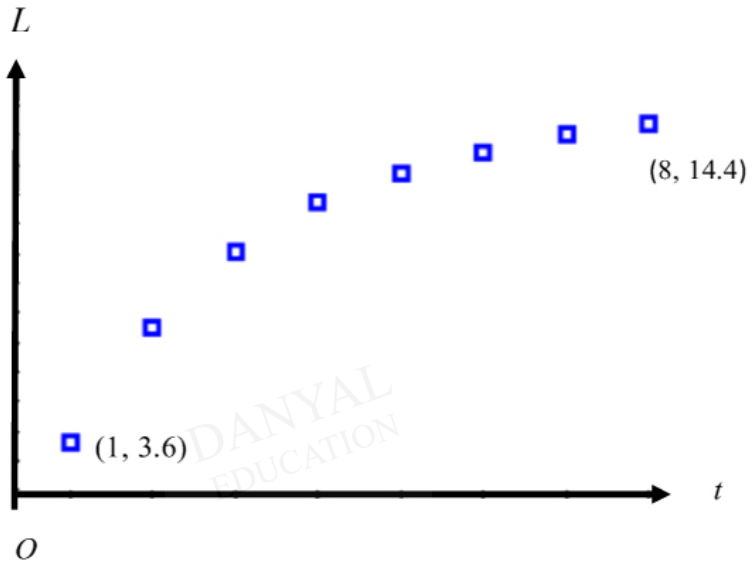
Answers

Correlation and Linear Regression Test 2

Q1

(i)	$\bar{y} = 22.51355 - 4.908387\bar{x}$ $\left(\frac{91+k}{9}\right) = 22.51355 - 4.908387\left(\frac{20.5}{9}\right)$ $k = 11.0$	Poorly attempted. Many students simply substituted $x = 2$ into the equation of the regression line and hoped that the resulting y, i.e. k will be 2. They failed to understand that the point $x = 2$ may not pass through the regression line. Students must understand the concept that $(\bar{x}, \bar{y})$ lies on the regression line.
(ii)		Most students handled this part accurately. There were students who carelessly wrote the x-intercepts as y-intercepts and vice versa. Others thought that $1.25 > 3.5$ and $7.5 > 18$. All these could have been avoided if students made an effort to check their scatter diagram before proceeding.
(iii)	As x increases, y decreases at a decreasing rate and tends towards a limit.	Poorly attempted. Students merely says that the graph of model (C) is similar to the graph in the scatter diagram. This warrants no marks. Students are reminded that they need to describe the shape of the graph. Students are advised against describing the gradient of the scatter diagram as it is prone to careless mistakes. In this question, as x increases, the gradient actually increases because it becomes less negative.
(iv)	$a = 0.257$ $b = 22.8$	Many students failed to leave their final answer in 3 s.f.
(v)	$y = 0.25681 + \frac{22.837}{x}$ $12 = 0.25681 + \frac{22.837}{x}$ $x = 1.94$ (3 s.f.) The weight of the car is 1940kg. The prediction is reliable as $y = 12$ is within the data range of y and the $\lvert r \rvert$-value is close to 1.	Many students left their answer as $x = 1.94$. They did not conclude that the weight of the car is 1940kg or 1.94 kg in thousands. Many students failed to mention that the $\lvert r \rvert$-value is close to 1 when stating that the prediction is reliable.
(vi)	(a) $\bar{y}$ will be increased by a. (b) Standard deviation of y remain unchanged. (c) Correlation coefficient remain unchanged.	Well-attempted by students.

Q2



As L increases at a decreasing rate/concave downwards with respect to t, the linear model $L = at + b$ should not be used.

(ii)

The r value for $L = a\sqrt{t} + b$ is 0.972.

The r value for $L = c\ln t + d$ is 0.996.

Since the value of $|r|$ for $L = c\ln t + d$, is closer to 1, $L = c\ln t + d$ is a better model.

$\therefore c = 5.28248 \approx 5.28$

$\therefore d = 3.92267 \approx 3.92$

(iii)

$$L = 5.28248\ln(0.5) + 3.92267$$
$$= 0.2611$$
$$= 0.261(3\text{ sf})$$

This estimate is not reliable as as the age of the Sole is out of the range of the data.

(iv)

Since

$G = 14.8 - L$

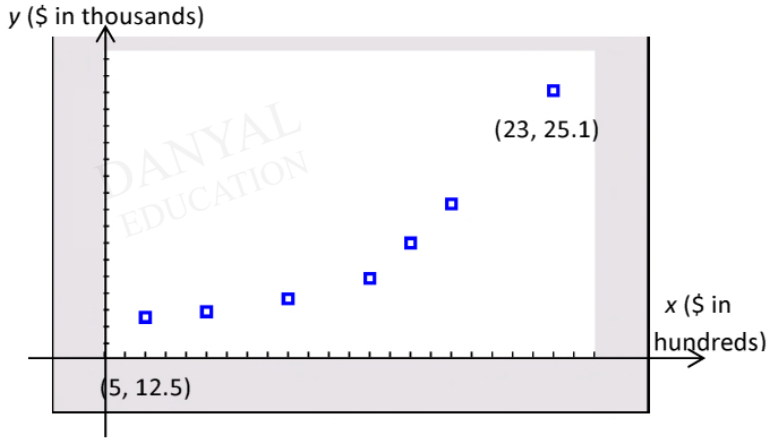
r_1 is positive but r_2 is negative.

$\therefore r_2 = -r_1$

Q3

(i) No, because correlation does not imply causation /
The increase in the monthly sales of refrigerators could be due to other factors such as a rise in the income level.

(ii)



There appears to be a curvilinear/non-linear relationship between x and y, thus a linear model is not likely to be appropriate.

(iii)
(A) $r = 0.9684$ (to 4dp)
(B) $r = 0.9495$ (to 4dp)

Since the r value between $e^{\sqrt{x}}$ and y has an absolute value closer to 1, $y = a + be^{\sqrt{x}}$ is the better model.

(iv) New regression line (8 data points) for y on $e^{(\sqrt{x})}$ is $y = 10.876 + 0.09906e^{(\sqrt{x})}$

$$\overline{y} = \frac{11.3+12.5+12.9+13.6+14.8+17+19.3+25.1}{8} = 15.813$$

Since $\overline{e^{(\sqrt{x})}}$ and $\overline{y}$ lie on the new regression line y on $e^{(\sqrt{x})}$,
and letting $x = m$ when $y = 11.3$,

$$15.813 = 10.876 + 0.09906\left(\frac{\sum_{i=1}^{7} e^{(\sqrt{x_i})} + e^{(\sqrt{m})}}{8}\right)$$

Using GC (1-var stats), $\sum_{i=1}^{7} e^{(\sqrt{x_i})} = 390.96$

$$\therefore 15.813 = 10.876 + 0.09906\left(\frac{390.96 + e^{(\sqrt{m})}}{8}\right)$$

$$\therefore e^{(\sqrt{m})} = 7.7479 \quad \Rightarrow \quad m = 4.19 \approx 4$$

Monthly advertising expenditure = \$400 (nearest hundred)