

A Level H2 Math

Correlation and Linear Regression Test 5

Q1

Six cities in a certain country are linked by rail to city O . The rail company provides the information about the distance of each city to city O and the rail fare from that city to city O on its website. Charles copied the table below from the website, but he had copied one of the rail fares wrongly.

City	A	B	C	D	E	F
Distance, x km	100	270	120	56	289	347
Rail fare, \$ y	11.1	17.1	6.44	7.62	17.9	18.8

- (i) Give a sketch of the scatter diagram for the data as shown on your calculator. On your diagram, circle the point that Charles has copied wrongly. [2]

For parts (ii), (iii) and (iv) of this question you should **exclude** the point for which Charles has copied the rail fare value wrongly.

- (ii) Find, correct to 4 decimal places, the product moment correlation coefficient between
(a) $\ln x$ and y ,
(b) x^2 and y . [2]
- (iii) Using parts (i) and (ii), explain which of the cases in part (ii) is more appropriate for modelling the data. [2]
- (iv) By using the equation of a suitable regression line, estimate the rail fare when the distance is 210 km. Explain if your estimate is reliable. [3]

Q2

An experiment to determine the effect of a fertilizer on crop yield was carried out. A field was divided into eight plots of equal area and eight different amounts of fertilizer, one for each plot, were used. The table below shows the amount of fertilizer, x grams, and the crop yield, y grams, for each plot.

Amount of fertilizer (x)	15	22	37	55	62	69	78	90
Yield (y)	101	123	137	150	150	154	158	160

- (i) Draw the scatter diagram for these values, labelling the axes. [1]

It is thought that the yield of a crop, y grams, can be modelled by one of the formulae

$$y = a + bx \quad \text{or} \quad y = c + d \ln x$$

where a , b , c and d are constants.

- (ii) Find the value of the product moment correlation coefficient between
(a) x and y ,
(b) $\ln x$ and y . [2]
- (iii) Use your answers to parts (i) and (ii) to explain which of $y = a + bx$ or $y = c + d \ln x$ is the better model. [2]
- (iv) For a plot of land, the yield of the crop was 144 grams. Using a suitable regression line estimate the amount of fertilizer used, giving your answer to the nearest gram. [2]
- (v) Comment on the reliability of the model in part (iv) in predicting the value of y when $x = 110$. [1]

Q3

- (a) Comment briefly on the following statements:
- (i) Flowers in a garden are watered and the product moment correlation coefficient between petal size and the amount of water given is 0.073, so it follows that there is no relation between petal size and quantity of water given to the flower. [1]
- (ii) The product moment correlation coefficient between the risk of heart disease and amount of red wine intake is found to be approximately -1 . Therefore we conclude that red wine intake causes the risk of heart disease to decrease. [1]
- (b) The median age of residents in Singapore across the years are given in the table.

Year (x)	1984	1988	1992	1996	2000	2004	2008	2012	2016
Median age (y)	26.7	28.8	27.0	32.3	34.0	35.4	36.7	38.4	40.0

It is thought that the median age of residents in year x can be modelled by one of the formulae

$$y = \frac{a}{x} + b, \quad y = c \ln x + d,$$

where a , b , c and d are constants.

- (i) Plot a scatter diagram on graph paper for these values, labelling the axis, using a scale of 2cm to represent 5 years on the y -axis and an appropriate scale for the x -axis. One of the values of y was quoted wrongly. Indicate this point as P on your diagram. [2]

For parts (ii), (iii), (iv) of this question, you should **exclude** the point P .

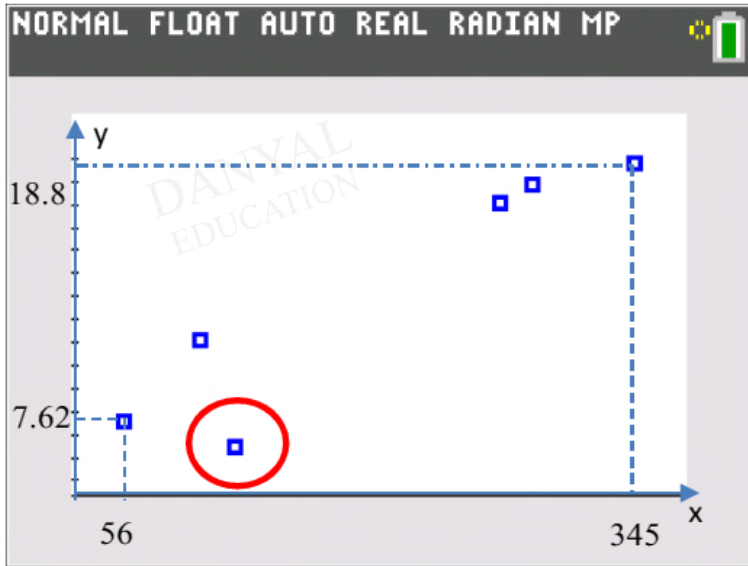
- (ii) Find, correct to 5 decimal places, the value of the product moment correlation coefficient between
(A) x^{-1} and y
(B) $\ln x$ and y . [2]
- (iii) Explain which model is more appropriate to predict the median age of residents in Singapore and find the equation of the least squares regression line for this model, giving your answer to 2 decimal places. [2]
- (iv) Explain why neither the regression line of x^{-1} on y nor the regression line of $\ln x$ on y should be used to estimate the year when the median age is 30. [1]
- (v) Give a possible reason for the rise in the median age. [1]

Answers

Correlation and Linear Regression Test 5

Q1

(i)



(ii) (a)

Product moment correlation coefficient, $r = 0.99959$

(b)

Product moment correlation coefficient, $r = 0.95137$

(iii)

From the scatter diagram, as x increases, the value of y increases at a decreasing rate, that seems to fit model (a) better. Also, the value of $|r|$ for model (a) is closer to 1 as compared to model (b).

(iv)

We use the regression line y on $\ln x$

$$y = 6.1619(\ln x) - 17.223 \approx 6.16 \ln x - 17.2$$

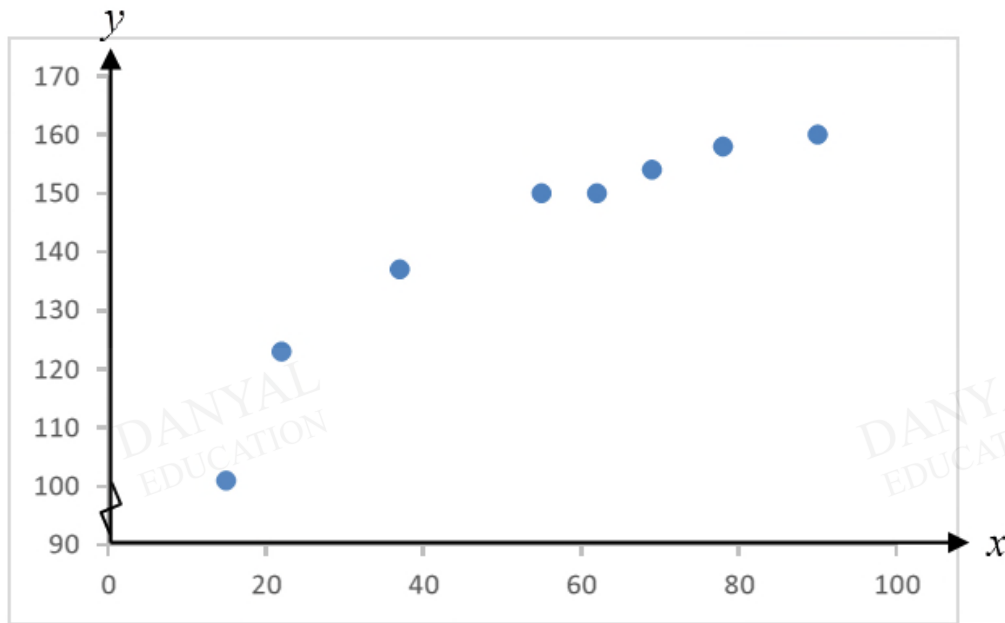
When $x = 210$,

$$y = 6.1619(\ln 210) - 17.223 = 15.725 \approx 15.7$$

As the value of $|r|$ is close to 1 and $x = 210$ is within the given data range, the estimation may be reliable.

Q2

6i



ii a From GC, $r = 0.93639 = 0.936$ (3 s.f)

ii b From GC, $r = 0.98775 = 0.988$ (3 s.f)

iii Since

- 1) the points on the scatter diagram seem to lie close to an increasing curve with decreasing gradient (or close to a curve in which y increases by decreasing amounts as x increases), and
- 2) the product moment correlation coefficient between $\ln x$ and y of 0.988 is closer to 1 than the product moment correlation coefficient between x and y of 0.936,

hence $y = c + d \ln x$ is the better model.

iv From (iii), we should use the regression line of y on $\ln x$.
 From GC, the equation of the regression line of y on $\ln x$ is

$$y = 20.8496 + 31.539 \ln x$$

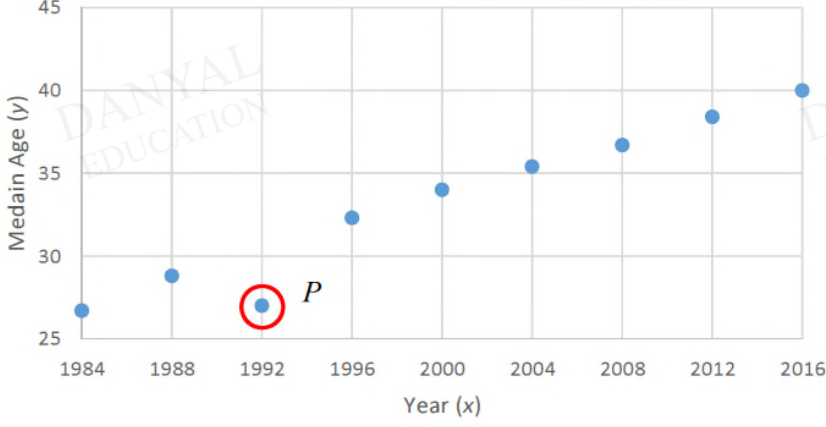
$$y = 20.8 + 31.5 \ln x \quad (3 \text{ s.f})$$

When $y = 144$, $144 = 20.8496 + 31.539 \ln x$

$$\therefore x = 49.635 = 50 \quad (\text{nearest gram})$$

iv Since $x = 110$ is outside the range of data values ($15 \leq x \leq 90$), hence the estimated value of y may not be reliable.

Q3

8(a) (i)	The value of 0.073 indicates that there is a weak linear correlation between petal size and the amount of water but there could be some non-linear relation.	Quite a number of students state that there is still some weak linear correlation.
8(a) (ii)	The approximate value of -1 indicates that there is a strong negative linear correlation between the risk of heart disease and amount of red wine intake. It does not mean that red wine intake decreases the risk of heart disease.	
8(b) (i)		Wrong scale is used by a handful of students.
8(b) (ii)	For Model A, $r = -0.9985438 = 0.99854$ For Model B, $r = 0.9984431 = 0.99844$	Quite a number of students chose model B instead of A simply because r is positive.
8(b) (iii)	Model A as the $ r $ value is closer to 1. The suitable regression line is $y = 848.24 - \frac{1629165.57}{x}$	
8(b) (iv)	This is because age (x) is the controlled variable	Badly answered. Students are not able to identify controlled variable.
8(b) (v)	The rise in the median age is due to the drop in the growth of the population.	