A Level H2 Math

Correlation and Linear Regression Test 3

Q1

A large cohort of students sat for a mathematics examination. Based on selected data of the examination results, the following table shows y, the proportions of students who scored x marks.

x	20	30	40	50	60	70	80	90
y	0.00029	0.00174	0.00663	0.0161	0.0252	0.0252	0.0161	0.00663

(i) Draw a scatter diagram for these values, labelling the axes. [2]

(ii) Explain why, in this context, a linear model is not appropriate. [1]

It is decided to fit a model of the form $\ln y = -a(x-m)^2 + b$, where $a > 0$ and m is a suitable constant, to the data. The product moment correlation coefficient between $(x-m)^2$ and $\ln y$ is denoted by r. The table below gives values of r for some possible values of m.

m	62.5	65	67.5
r	0.9899292		0.9938968

(iii) Calculate the value of r for $m = 65$, giving your answer correct to 7 decimal places. [1]

(iv) Use the table and your answer in part **(iii)** to suggest with a reason which of 62.5, 65 or 67.5 is the most appropriate value for m. [1]

(v) Using the value of m found in part **(iv)**, calculate the values of a and b, and use them to predict the proportion of students who scored 45 marks.
Comment on the reliability of your prediction. [5]

Q2

(a) Traffic engineers are studying the correlation between traffic flow on a busy main road and air pollution at a nearby air quality monitoring station. Traffic flow, x, is recorded automatically by sensors and reported each hour as the average flow in vehicles per hour for the preceding hour. The air quality monitoring station provides, each hour, an overall pollution reading, y, in a suitable unit (higher readings indicate more pollution). Data for a random sample of 8 hours are as follows.

Traffic flow, x	1796	1918	2120	2315	2368	2420	2588
Pollution reading, y	1.0	2.2	3.5	4.2	4.3	4.5	4.9

(i) Draw the scatter diagram for these values, labelling the axes. [2]

It is thought that the pollution y can be modelled by one of the formulae

$$y = a + bx \qquad\qquad y^2 = c + dx$$

where a, b, c and d are constants.

(ii) Find the value of the product moment correlation coefficient between

(a) x and y,

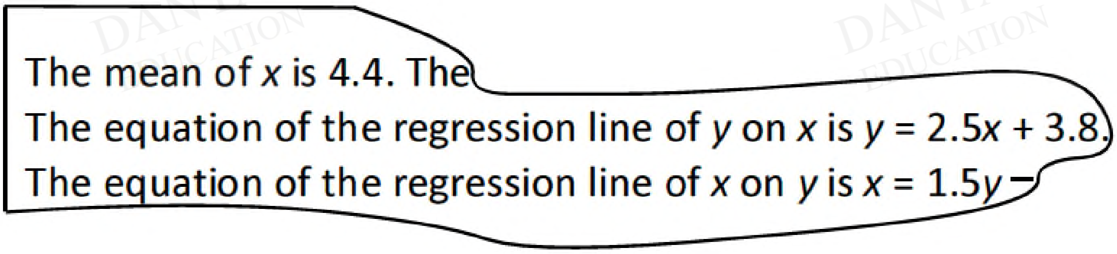
(b) x and y^2. [2]

(iii) Use your answers to parts **(i)** and **(ii)** to explain which of $y = a + bx$ or $y^2 = c + dx$ is the better model. [2]

(iv) It is required to estimate the value of y for which $x = 2000$. Find the equation of a suitable regression line, and use it to find the required estimate. [2]

(v) The local newspaper carries a headline "Heavy traffic causes air pollution". Comment briefly on the validity of this headline in the light of your results. [1]

(b) The diagram below shows an old research paper that has been partially destroyed. The surviving part of the paper contains incomplete information about some bivariate data from an experiment. Calculate the missing constant at the end of the equation of the second regression line. [3]



The mean of x is 4.4. The

The equation of the regression line of y on x is $y = 2.5x + 3.8$

The equation of the regression line of x on y is $x = 1.5y -$

Q3

In an experiment to investigate the decay of organic material over time, a bag of leaf litter was allowed to sit for a 20-week period in a moderately forested area.

The table below shows the weight of the remaining leaf litter (y kg) when x number of weeks have passed.

x	1	2	4	6	8	9	11	15	20
y	60.9	51.8	34.7	26.2	14.0	12.3	8.2	3.1	1.4

(i) Draw a scatter diagram of these data. [1]

(ii) Find the equation of the regression line of y on x and calculate the corresponding estimated value of y when $x = 17$.
Comment on the suitability of the linear model for these data. [3]

The variable W is defined as $W = \ln y$.

(iii) Find the product moment correlation coefficient between W and x. [1]

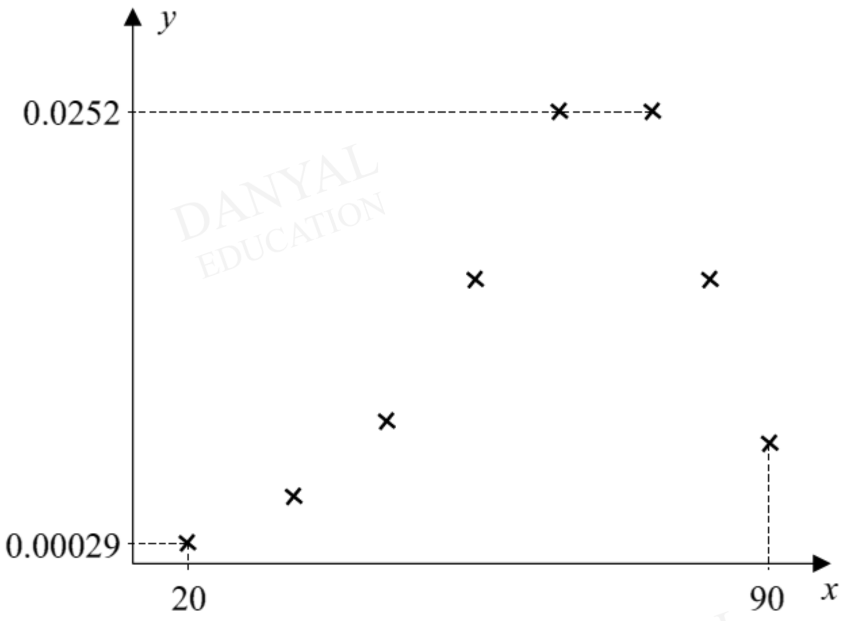
(iv) It is given that the weight of the leaf litter in the bag was 75.0 kg initially. Using an appropriate regression line, estimate how long it takes for the weight of the leaf litter to drop to half its initial value, giving your answer to one decimal place. [3]
Give two reasons why you would expect this estimate to be reliable. [2]

Answers

Correlation and Linear Regression Test 3

Q1

(i)



(ii)
The scatter diagram displays a curvilinear relationship which suggests the presence of a maximum point. Hence a linear model is inappropriate.

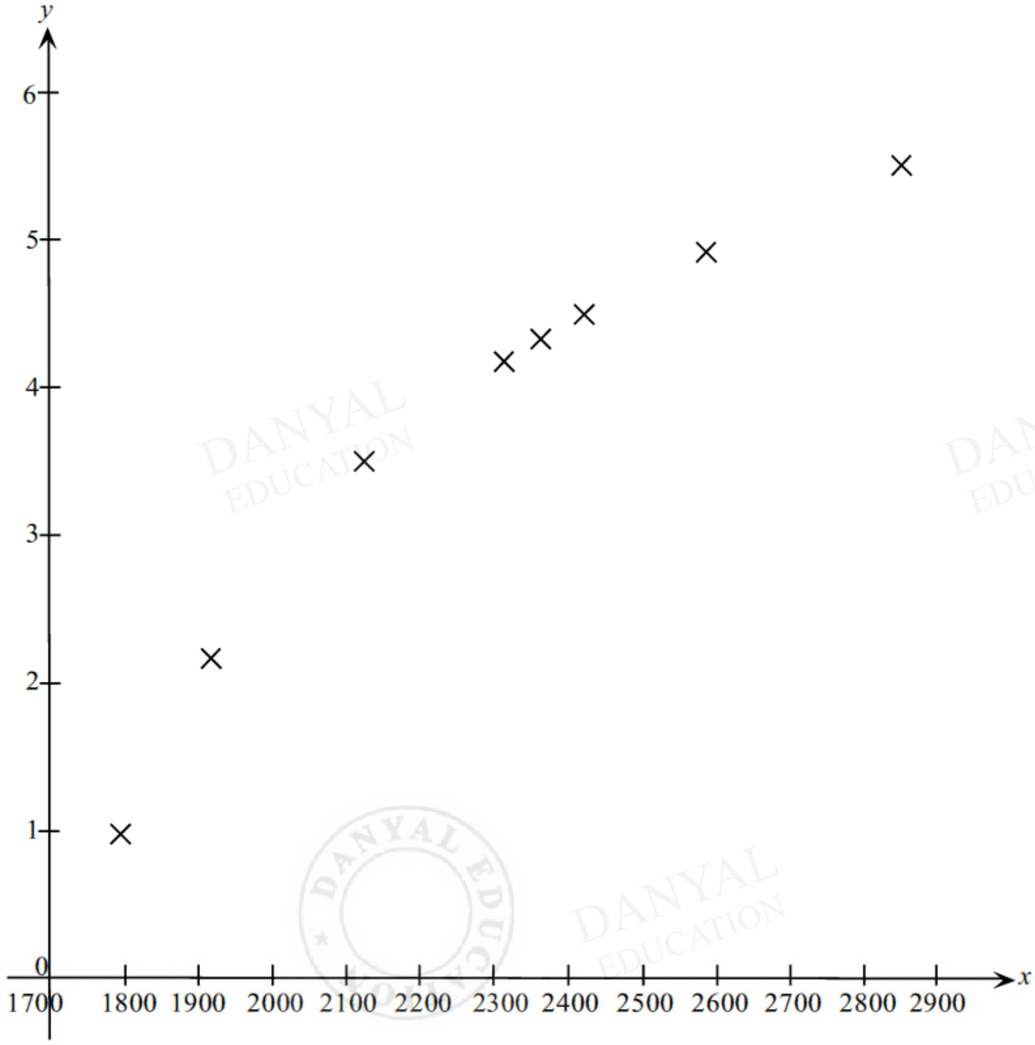
(iii)
$r = -0.9999984$ (7 decimal places)

(iv)
$m = 65$. Of the 3 negative r values, the r value corresponding to $m = 65$ is closest to -1 .

(v)
Using GC with $m = 65$,
$a \approx 0.0022230 \approx 0.00222$ (3 s.f.)
$b \approx -3.6269 \approx -3.63$ (3 s.f.)
$\therefore \ \ln y = -0.002223(x-65)^2 - 3.6269$
or $\ln y = -0.00222(x-65)^2 - 3.63$
When $x = 45$,
$y \approx 0.0109$ (3 s.f.)

Since $x = 45$ is within data range and $r = -0.9999984$ is very close to -1 , the prediction is reliable.

Q2



(ii) **(a)** Between x and y: $r = \underline{0.959}$

(b) Between x and y^2: $r = \underline{0.995}$

(iii) From **(i)**, since <u>as x increases, y increases at a decreasing rate</u>, the points on the scatter diagram take the shape of the graph of $y^2 = c + dx$.

Or: From **(i)**, the points on the scatter diagram seem to lie on a <u>concave downward curve</u>.

From **(ii)**, the product moment correlation coefficient between x and y^2 is <u>closer to 1</u>, as compared to that between x and y,

$\therefore$ the model $\underline{\underline{y^2 = c + dx}}$ is the <u>better</u> model.

(iv) From GC, the regression line of y^2 on x is

$y^2 = 0.027897x - 47.985$

$\underline{\underline{y^2 = 0.0279x - 48.0}}$ (3 sf)

When $x = 2000$,

$$\begin{aligned} y^2 &= 0.027897(2000)-47.985 \\ &= 7.809 \\ \therefore\ y &= \underline{\underline{2.79}} \text{ (3 sf) or } \underline{\underline{2.8}} \text{ (1 dp, as shown in the table of values)} \end{aligned}$$

(v) May not be valid as correlation does not necessarily imply causation.

Or: May not be valid as there could be other factors relating traffic flow and air pollution.

(b)

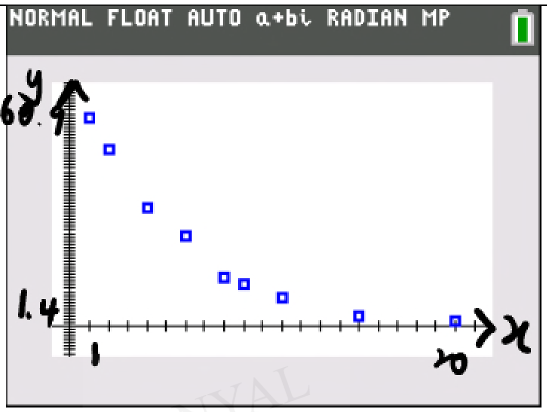
$$\begin{aligned} y &= 2.5x+3.8 \\ \overline{y} &= 2.5\overline{x}+3.8 \\ &= 2.5(4.4)+3.8 \\ &= 14.8 \end{aligned}$$

Let

$$\begin{aligned} x &= 1.5y-k \\ \overline{x} &= 1.5\overline{y}-k \\ 4.4 &= 1.5(14.8)-k \\ k &= 22.2-4.4 \\ &= \underline{\underline{17.8}} \end{aligned}$$

Q3

(i)



(ii) Regression line of y on x is $y = 49.7 - 3.09x$

When $x = 17$, $y = -2.8466\ldots = -2.85$

The linear model is not suitable since

1) the negative value of y is impossible or
2) the scatter diagram shows a curved relationship between the two variables.

(iii) Product moment correlation coefficient between W and x
$= -0.997837\ldots = -0.998$

(iv) Since x is the controlled variable, we use the regression line of $\ln y$ on x :

$\ln y = 4.3549 - 0.20532x$ [from GC]

When $y = \frac{1}{2}(75)$,

we have $\ln\frac{75}{2} = 4.3549 - 0.20532x$

$\Rightarrow x = 3.5581\ldots = 3.6$

The weight will drop to half its original value in 3.6 weeks.

The estimate is reliable since

1) The product moment correlation coefficient between $\ln y$ and x is -0.998 which is very close to -1, showing a strong negative linear correlation between $\ln y$ and x.
2) The estimate is an interpolation, because $y = \frac{1}{2}(75)$ is in the data range $1.4 \le y \le 60.9$.