A Level H2 Math

Correlation and Linear Regression Test 1

Q1

A swim school takes in both male and female primary school students for competitive swimming lessons. The school assesses its students' progress each year by recording the time, t seconds, each student takes to swim a 50-metre lap in breaststroke, and the number of months, m, that he or she has been at the school. The records for 8 randomly chosen students are shown in the following table.

m	6	7	10	12	15	19	21	24
t	92.32	87.11	66.12	59.41	53.94	43.82	42.07	41.45

(i) Labelling the axes clearly, draw a scatter diagram for the data and explain, in context, why a linear model would not be suitable to predict the time taken by a student to swim a lap of breaststroke given the number of months that he or she has been at the school. [2]

It is desired to fit a model of the form $\ln(t-C)=a+bm$, where C is a suitable constant. The product moment correlation coefficient r between m and $\ln(t-C)$ for some possible values of C are shown in the table below.

C	36	37	38	39
r	–0.992114		–0.992681	–0.992192

(ii) Calculate the value of r for $C=37$, giving your answer correct to 6 decimal places. [1]

(iii) Use the table and your answer to **(ii)** to choose the most appropriate value for C. Explain your choice. [2]

For the remainder of this question, use the value of C that you have chosen in **(iii)**.

(iv) Find the equation of the least squares regression line of $\ln(t-C)$ on m. Give an interpretation of C in the context of the question. [2]

(v) Another student who has been swimming at the school for 9 months clocked a time of 60.33 seconds for a lap of breaststroke. Using your regression line, comment on the student's swimming ability. [2]

(vi) Suggest an improvement to the data collection process so that the results could provide a fairer gauge of the expected outcome for the students in the first 2 years of lessons. [1]

Q2

A pilot records the take-off distance, S metres, for his private aircraft on runways at various altitudes of h metres. The data are shown in the table below.

h	0	300	600	900	1200	1500	1800
S	635	690	750	840	950	1080	1250

(i) Plot a scatter diagram on graph paper for these values, labelling the axes, using a scale of 2 cm to represent a take-off distance of 100 metres on the *y*-axis and an appropriate scale for the *x*-axis. [2]

It is thought that the take-off distance S can be modelled by one of the formulae

$$S = ah + b \qquad \text{or} \qquad S = ch^2 + d,$$

where a, b, c and d are constants.

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

(a) h and S,

(b) h^2 and S. [2]

(iii) Use your answers to parts **(i)** and **(ii)** to explain which of $S = ah + b$ or $S = ch^2 + d$ is the better model. [2]

(iv) Find the equation of the least-square regression line for the model you have chosen in part **(iii)**. [1]

(v) Use the equation of your regression line to estimate the take-off distance for altitude of 2200 metres. Comment on the reliability of your estimate when $h = 2200$. [2]

Q3

In the study of how the population of a harmful bacteria varies with temperature, scientists conducted an experiment to collect the following set of data:

Temperature (x °C)	10	12	14	16	18	20	22	24	26	28
Population (y millions)	25.4	25.1	24.4	22.9	20.8	18.3	15.4	12.2	8.8	5.3

(i) Draw a scatter diagram for the above data, labelling the axes clearly. [2]

(ii) Calculate the value of the product moment correlation coefficient. Explain why a linear model is not appropriate. [2]

It is suggested the relationship between x and y can be modelled by one of the following formulae:

$$y = a + \frac{b}{x} \quad \text{or} \quad y = a - bx^2$$

where a and b are positive constants.

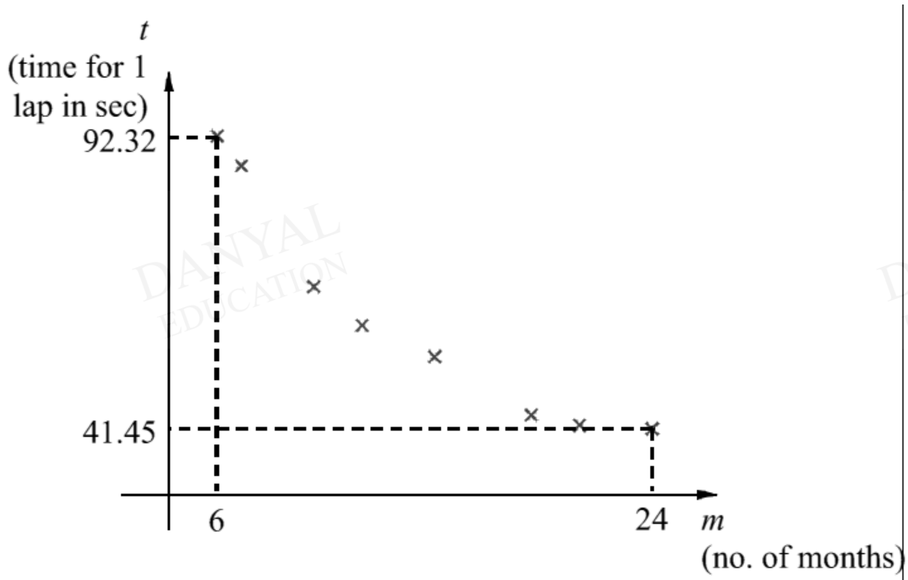
(iii) Explain which of the above two models is the better model and calculate the values of a and b for the chosen model. [3]

(iv) It is required to estimate the temperature when the population of the bacteria is 10 millions. By using an appropriate regression line, find an estimate of the value of x and comment on the reliability of your answer. [2]

Answers

Correlation and Linear Regression Test 1

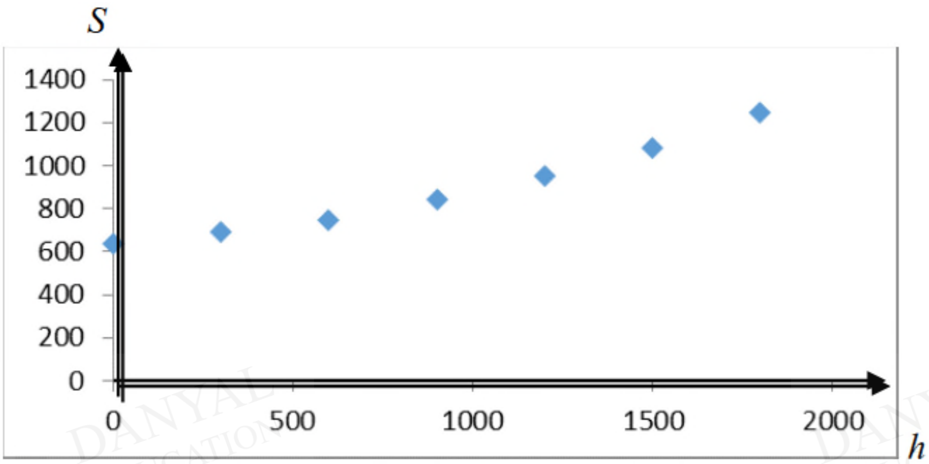
Q1

(i)		3 important points to note for scatter diagram: 1) axes t and m labelled 2) extreme values labelled 3) 8 points in total
	A linear model would imply that in the long run, the time taken to swim a lap would be negative, which is unrealistic. *(Note: Extrapolation is not accepted as a reason, as the question isn't looking for a reason based on the data obtained.)*	Acceptable answers include: - negative time - zero time
(ii)	Using GC, for $C = 37$, $r = -0.992555$	**R**: 6 d.p.
(iii)	The most appropriate value for C is 38, as the magnitude of its corresponding value of r is closest to 1.	Acceptable answers include: - $\lvert r \rvert \approx 1$ - $r \approx -1$ Quite a number of scripts had “closet” instead of “closest”!

(iv)	From GC, least squares regression line of $\ln(t-38)$ on m is $\ln(t-38)=5.01236-0.16349m$ $\Rightarrow$ $\underline{\ln(t-38)=5.01-0.163m}$ (3 s.f.) $C=38$ is the fastest time that a student can expect to complete a lap of breaststroke after spending a long time at the swim school. (Making t the subject in the equation of the regression line gives us $t=38+\mathrm{e}^{5.01-0.163m}$, so as $m\to\infty$, $t\to 38$.)	**R**: use C =38 **R**: $\ln(t-38)$ on m 3 s.f. for final answer Please note that C is NOT the gradient; C is NOT the y-intercept Acceptable answers include: - fastest time after a long period - shortest time after a long period
(v)	When $m=9$, $t\ =38+\mathrm{e}^{5.01236-0.16349(9)}$ $=72.50$ (2 d.p.) A timing of 60.33 seconds is well below the expected timing of 72.50 seconds. Therefore, we can say that the student is exceptionally strong in his/her swimming ability.	Acceptable answers include: - very strong - very talented - way above average
(vi)	The 8 randomly selected students might have been of different genders and ages. To make the results fairer, data could be collected separately based on genders and age ranges.	The following may not give **fairer results**: - increase sample size - increase frequency - group by ability (beginner, intermediate, advanced) is subjective

Q2

(i)



(ii)

(a) $r = 0.980867 \approx 0.9809$ (4 d.p.)

(b) $r = 0.996039 \approx 0.9960$ (4 d.p.)

(iii)

The scatter diagram shows that <u>S increases at an increasing rate as h increases</u>, and for $S = ch^2 + d$, <u>$r \approx 0.9960$ which is closer to 1</u>, so the model $S = ch^2 + d$ is a better model.

(iv)

The equation of regression line is

$$S = 0.0001822853073h^2 + 671.7261905$$

i.e. $S = 0.000182h^2 + 672$ (3 s.f.)

(v)

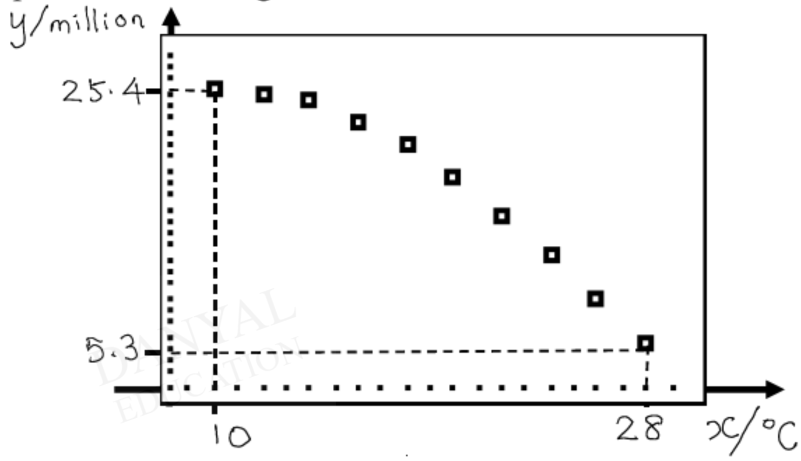
$S = 0.00018229(2200)^2 + 671.73$
$= 1554.0136$
$= 1550$ metres (3 s.f.)

Estimate for when $h = 2200$ metres is <u>not</u> reliable since $h = 2200$ metres is outside the range of the given data and <u>extrapolation</u> is not a good practice.

Q3

(i)

The required scatter diagram is as shown below:



(ii)

From GC, the correlation coefficient $r = -0.973$.
Although the value of r is close to -1 and suggests a strong negative linear relationship between x and y, the scatter diagram shows a curvilinear relationship between x and y. Thus, the a linear relationship between x and y is not appropriate.

(iii)

The scatter diagram shows that when x increases, y decreases at increasing rate. Thus, the model with $y = a - bx^2$ where a, b are positive constants is more appropriate.

Using GC, we found that $a = 29.98560169 = 30.0$ (3 s.f.)

and $b = 0.0307756388 = 0.0308$ (3 s.f.)

(For $a, b > 0$, $y = a + \frac{b}{x}$ decreases at a decreasing rate when x increases)

(iv)

As x is the independent variable and y is the dependent variable, we will still use the regression line $y = 30.0 - 0.0308x^2$ to estimate the value of x.
Thus, when $y = 10$, $x = 25.5$ °C (3 s.f.)
The anwer is reliable for the following reasons:

i) correlation coefficient $r = -0.995$ has absolute value close to 1
ii) the y value of 10 is within data range of the available y values.