Pearson Edexcel AS Further Mathematics 8FM0

FurStats 2 − 1 Linear Regression

Time allowed: 45 minutes

School: www.CasperYC.club

Name:

Teacher:

Question	Points	Score
1	11	
2	11	
3	14	
4	17	
Total:	53	

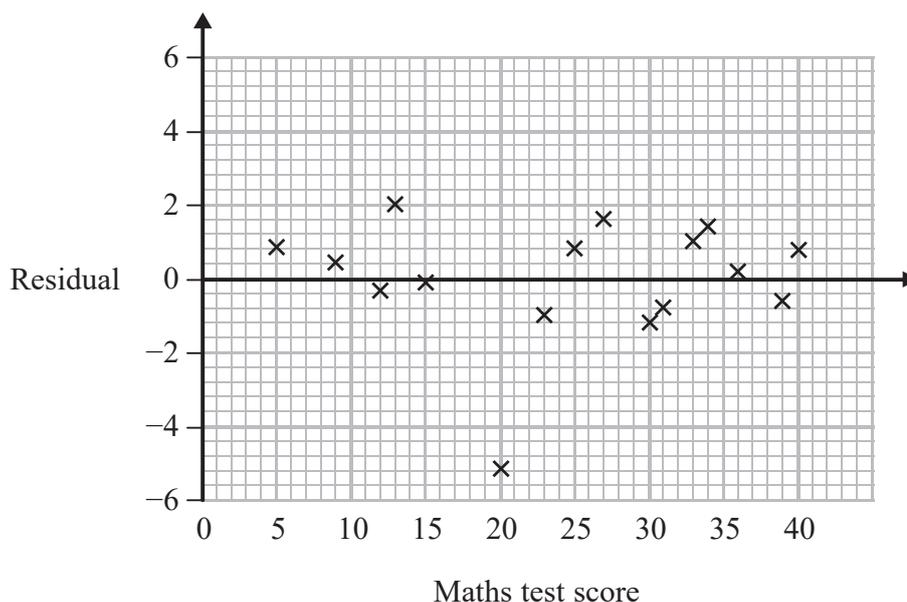**How I can achieve better:**

-
-
-

1. The scores achieved on a maths test, m, and the scores achieved on a physics test, p, by 16 students are summarised below.

$$\sum m = 392 \qquad \sum p = 254 \qquad \sum p^2 = 4748 \qquad S_{mm} = 1846 \qquad S_{mp} = 1115$$

(a) Find the product moment correlation coefficient between m and p. [2]

(b) Find the equation of the linear regression line of p on m. [3]

Figure below shows a plot of the residuals.

Maths test score

(c) Calculate the residual sum of squares (RSS). [2]

For the person who scored 30 marks on the maths test,

(d) find the score on the physics test. [2]

The data for the person who scored 20 on the maths test is removed from the data set.

(e) Suggest a reason why. [1]

The product moment correlation coefficient between m and p is now recalculated for the remaining 15 students.

(f) Without carrying out any further calculations, suggest how you would expect this recalculated value to compare with your answer to part (a). Give a reason for your answer. [1]

Total: 11

(Q1 continued ...)

2. Two students, Jim and Dora, collected data on the mean annual rainfall, w cm, and the annual yield of leeks, l tonnes per hectare, for 10 years.

Jim summarised the data as follows

$$S_{wl} = 42.786 \qquad S_{ww} = 9936.9 \qquad \sum l^2 = 26.2326 \qquad \sum l = 16.06$$

(a) Find the product moment correlation coefficient between l and w. [2]

Dora decided to code the data first using $s = w - 6$ and $t = l - 20$.

(b) Write down the value of the product moment correlation coefficient between s and t. Give a justification for your answer. [1]
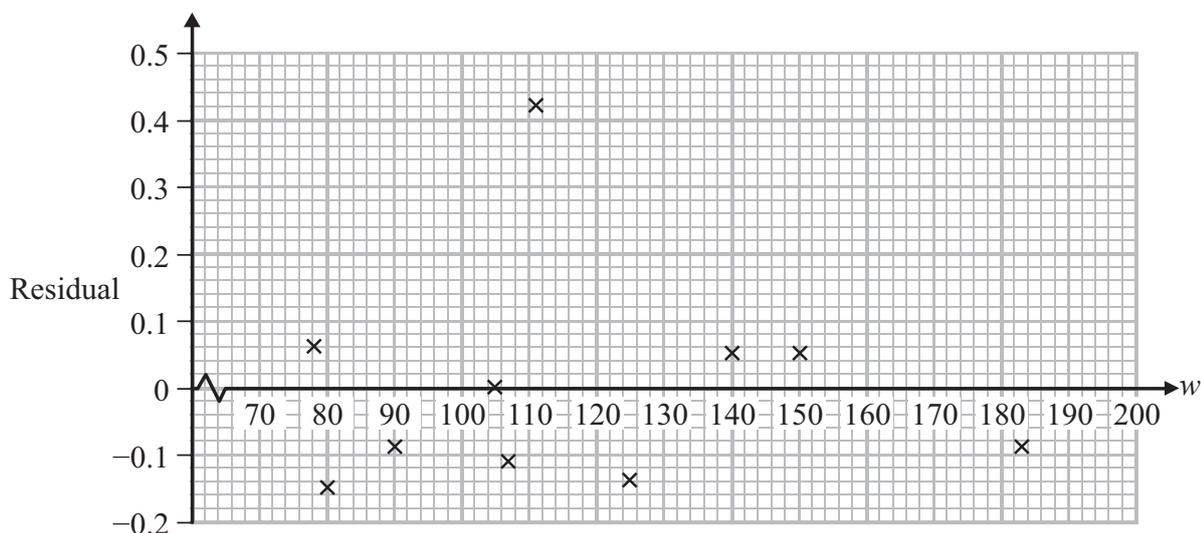
Dora calculates the equation of the regression line of t on s to be $\quad t = 0.00431s - 18.87$.

(c) Find the equation of the regression line of l on w in the form $l = a + bw$, giving the values of a and b to 3 significant figures. [3]

(d) Use your equation to estimate the yield of leeks when w is 100 cm . [1]

(e) Calculate the residual sum of squares. [2]

The graph shows the residual for each value of l.

(f)   i. State whether this graph suggests that the use of a linear regression model is suitable for these data. Give a reason for your answer. [2]

  ii. Other than collecting more data, suggest how to improve the fit of the model in part (c) to the data.

Total: 11

(Q2 continued . . .)
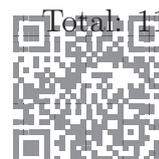
3. Some students are investigating the strength of wire by suspending a weight at the end of the wire. They measure the diameter of the wire, d mm, and the weight, w grams, when the wire fails. Their results are given in the following table.

	These 14 points are plotted on next page														Not yet plotted			
d	0.5	0.6	0.7	0.8	0.9	1.1	1.3	1.6	2	2.4	2.8	3.3	3.5	3.9	4.5	4.6	4.8	5.4
w	1.2	1.7	2.3	3.0	3.8	5.6	7.7	11.6	18	25.9	34.9	47.4	52.7	63.9	81	83.6	89.9	109.4

The first 14 points are plotted on the next page.

(a) On the axes next page, complete the scatter diagram for these data. [1]

(b) Use your calculator to write down the equation of the regression line of w on d. [2]

(c) With reference to the scatter diagram, comment on the appropriateness of using this linear regression model to make predictions for w for different values of d between 0.5 and 5.4. [1]

The product moment correlation coefficient for these data is $r = 0.987$ (to 3 significant figures).

(d) Calculate the residual sum of squares (RSS) for this model. [2]

Robert, one of the students, suggests that the model could be improved and intends to find the equation of the line of regression of w on u, where $u = d^2$.

He finds the following statistics

$$\text{S}_{wu} = 5721.625 \qquad \text{S}_{uu} = 1482.619 \qquad \sum u = 157.57$$

(e) By considering the physical nature of the problem, give a reason to support Robert's suggestion. [1]

(f) Find the equation of the regression line of w on u. [3]

(g) Find the residual sum of squares (RSS) for Robert's model. [2]

(h) State, giving a reason based on these calculations, which of these models better describes these data. [1]

(i) Hence estimate the weight at which a piece of wire with diameter 3 mm will fail. [1]

Total: 14

(Q3 continued . . .)

(Q3 continued . . .)
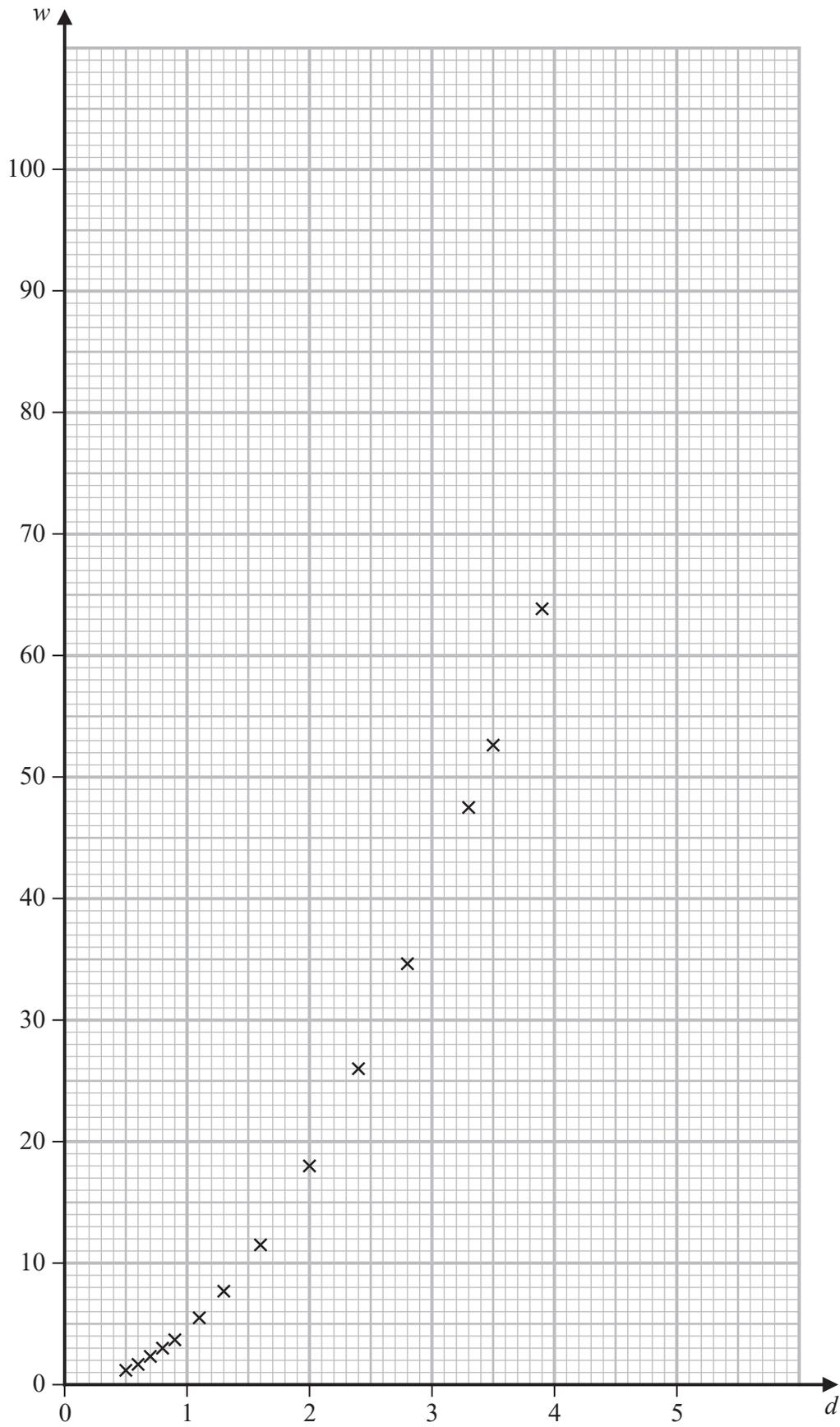
4. Naasir is studying the hearing function of workers in his factory. He takes a random sample of 10 workers from his factory. He records their hearing function as a percentage, h, where 100% indicates perfect hearing. He also records the number of years they have worked in the factory, w.

The results are summarised below.

$$\bar{h} = 86.5 \qquad \bar{w} = 10.5 \qquad S_{hh} = 198.5 \qquad S_{ww} = 470.5 \qquad r = -0.620$$

(a) Stating your hypotheses clearly and using a 5% level of significance, test whether or not there is evidence of a negative correlation between hearing function and the number of years worked in Naasir's factory. [3]

(b) State why your test in part (a) is consistent with the use of a linear regression model to describe the relationship between hearing function and the number of years worked in Naasir's factory. [1]

Naasir wishes to estimate the effect that the number of years spent working in his factory has on the hearing function of his workers. He decides to create a linear regression model.

(c) Using the above data find Naasir's linear regression model. [6]

(d) Use this model to estimate the effect on hearing function of working in Naasir's factory for 5 years. [1]

(e) Calculate the residual sum of squares for this model. [2]

Erica was one of the workers included in the sample.

She is 58 years old, has worked in the factory for 10 years and her hearing function is 80%.

(f) Calculate the residual for Erica's data. [2]

A friend visited Naasir's factory and suggested that all new workers should have an assessment of their hearing function when they start working in his factory.

(g)   i. State, giving your reasons, a limitation of Naasir's current model. [2]

  ii. Explain how he could use this extra information to design an improved model.

Total: 17

(Q4 continued . . .)