**RAFFLES INSTITUTION**
**H2 Mathematics 9758**
**2025 Year 6**

## Tutorial S6: Correlation and Regression

**1** The amounts, $x$ grams, of catalyst used in a chemical reaction and the resulting times, $y$ hours, taken to complete the reaction were recorded. The results are given in the following table.

| $x$ | 2.0 | 2.5 | 3.0 | 3.5 | 4.0 | 4.5 | 5.0 |
|---|---|---|---|---|---|---|---|
| $y$ | 62.1 | 51.2 | 44.1 | 39.2 | 35.0 | 37.3 | 33.0 |

**(i)** Calculate the value of the product moment correlation coefficient for the data.
**(ii)** State what your value indicates about the relation between $x$ and $y$, and what you expect about the scatter diagram from this sample.
**(iii)** Draw a scatter diagram of the data and comment on your answer to part **(ii)**.

**[(i)** $-0.925$**]**

**2** The variables $x$ and $y$ are believed to be linearly related and a set of values of $x$ and $y$ are obtained experimentally. Both variables are subject to experimental error. State circumstances under which the regression line of $x$ on $y$ should be used, rather than the regression line of $y$ on $x$, when the value of one variable is to be estimated from a given value of the other.

The heart rate ($x$) and diastolic blood pressure ($y$), both in suitable units, were measured for each of 10 hospital patients after being given a certain drug. The results were as follows.

| $x$ | 49 | 51 | 54 | 58 | 63 | 64 | 68 | 70 | 75 | 78 |
|---|---|---|---|---|---|---|---|---|---|---|
| $y$ | 90 | 88 | 85 | 91 | 82 | 85 | 76 | 77 | 70 | 71 |

**(i)** By referring to the scatter diagram of the data obtained using your graphing calculator, state what the scatter diagram indicates about the correlation between $x$ and $y$.

**(ii)** Calculate the equation of the estimated least square regression line of $y$ on $x$ in the form $y = a + bx$, giving the values of $a$ and $b$ correct to 3 significant figures.

**(iii)** Use a suitable regression line to estimate the heart rate when the diastolic blood pressure is 80.
Give a reason why it might be unwise to use either of the regression lines to estimate the diastolic blood pressure when the heart rate is 90.

**(iv)** Find the product moment correlation coefficient for the data and state, giving a reason, whether its value confirms your statement in part **(i)**.

**[ (ii)** $y = 126 - 0.704x$ **(iii)** 64.8 **(iv)** $-0.919$ **]**

**3** The following data is collected from a trial of a weedkiller.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $v$ | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 |
| $x$ | 46 | 32 | 31 | 27 | 18 | 15 | 14 | 12 | 11 |

Trial field areas, each of area 1 hectare, were treated with different volumes of weedkiller. The volume of weedkiller applied is $v$ litres, and $x$ is the number of weeds found in the corresponding hectare after the weedkiller is applied.

**(i)** Calculate the equation of the regression line of $x$ on $v$ in the form $x = a + bv$.

**(ii)** Use your answer to estimate the number of weeds that would be found in an area of 1 hectare treated with 35 litres of weedkiller.

**(iii)** Interpret the coefficient $b$ in the context of volume of weedkiller and the number of weeds.

**(iv)** Give a reason why $a$ may not necessarily give the expected number of weeds per hectare when no weedkiller is used.

**(v)** Explain why in this context a linear model would probably not be appropriate for large values of $v$.
**[(i)** $x = 43.4 - 0.410v$; **(ii)** 29**]**

**4** With the aid of a suitable diagram, describe the difference between the regression line of $Y$ on $X$ and that of $X$ on $Y$. [3]
Under what circumstances would the above two lines be coincident? [1]

The following summarizes the data from 10 sets of lengths($x$) and breadths($y$) in mm:

$$\sum x = 1782,\ \sum y = 1483,\ \sum x^2 = 318086,\ \sum y^2 = 220257,\ \sum xy = 264582$$

**(i)** Find the value of the product moment correlation coefficient. [1]
**(ii)** Find the equation of the regression line of $y$ on $x$ and that of $x$ on $y$. [2]
**(iii)** State a data point which can be included such that $r$ remains the same for the new set of data. [1]
**(iv)** Predict the breadth when the length is 185 mm. Comment on the reliability of your prediction. [2]
**[(i)** 0.744; **(ii)** $y = 0.584x + 44.3$, $x = 0.949y + 37.4$; **(iii)** $(178.2, 148.3)$; **(iv)** 152**]**

**5** **(a)** Draw separate scatter diagrams, each with 8 points, all in the first quadrant, which represents the situation where the product moment correlation coefficent between variable $x$ and $y$ is
**(i)** $-1$,
**(ii)** 0,
**(iii)** between 0.5 and 0.9. [3]

**(b)** An investigation into the effect of a fertiliser on yields of corn found that the amount of fertiliser applied, $x$, resulted in the average yields of corn, $y$, given below, where $x$ and $y$ are measured in suitable units.

| $x$ | 0 | 40 | 80 | 120 | 160 | 200 |
|---|---|---|---|---|---|---|
| $y$ | 70 | 104 | 118 | 119 | 126 | 129 |

**(i)** Draw a scatter diagram for these values. State which of the following equations, where $a$ and $b$ are positive constants, provides the most accurate model of the relationship between $x$ and $y$.

(A) $y = ax^2 + b$ (B) $y = \dfrac{a}{x^2} + b$

(C) $y = a\ln 2x + b$ (D) $y = a\sqrt{x} + b$ [2]

**(ii)** Using the model you chose in part **(i)**, write down the equation for the relationship between $x$ and $y$, giving the numerical values of the coefficients. State the product moment correlation coefficient for this model. [3]

**(iii)** Give two reasons why it would be reasonable to use your model to estimate the value of $y$ when $x = 189$. [2]

**6** The radiation intensity $I$ at time $t$, from a radioactive source, is given by the formula $I = I_0 e^{-kt}$, where $I_0$ and $k$ are constants. Show that the relation between $\ln I$ and $t$ is linear.

The following data were obtained from a particular source. The values of $t$ may be considered to be exact, while the values of $I$ are subject to experimental error.

| $t$ | 0.2 | 0.4 | 0.6 | 0.8 | 1.0 |
|---|---|---|---|---|---|
| $I$ | 3.22 | 1.63 | 0.89 | 0.41 | 0.36 |

**(i)** Find the equation of the estimated regression line of $\ln I$ on $t$ and hence give estimates for $I_0$ and $k$.

**(ii)** Calculate the radiation intensity that would be expected at time $t = 0.5$.

**(iii)** Calculate the product moment correlation coefficient between $t$ and $\ln I$.

**(iv)** It is required to estimate the value of $t$ for which $I = 1.5$. Explain why neither the regression equation of $t$ on $I$ nor the regression line of $t$ on $\ln I$ should be used.

**(v)** Use a regression line to give the best estimate that you can of the time when the radiation intensity is 1.5.

**[** $\ln I = -2.88t + 1.65$, $I_0 = 5.23$, $k = 2.88$, **(ii)** 1.24, **(iii)** $-0.984$, **(v)** 0.433 **]**

**7** Amy is revising for a mathematics examination and takes a different practice paper each week. Her marks, $y$% in week $x$, are as follows.

| Week, $x$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Percentage mark $y$ | 38 | 63 | 67 | 75 | 71 | 82 |

**(i)** Draw a scatter diagram showing these marks. [1]

**(ii)** Suggest a possible reason why one of the marks does not seem to follow the trend. [1]

**(iii)** It is desired to predict Amy's marks on future papers. Explain why, in this context, neither a linear nor a quadratic model is likely to be appropriate. [2]

It is decided to fit a model of the form $\ln(L - y) = a + bx$, where $L$ is a suitable constant. The product moment correlation coefficient between $x$ and $\ln(L - y)$ is denoted by $r$. The following table gives values of $r$ for some possible values of $L$.

| $L$ | 91 | 92 | 93 |
|---|---|---|---|
| $r$ | | $-0.929944$ | $-0.929918$ |

**(iv)** Calculate the value of $r$ for $L = 91$, giving your answer correct to 6 decimal places. [1]

**(v)** Use the table and your answer to part **(iv)** to suggest with a reason which of 91, 92 or 93 is the most appropriate value for $L$. [1]

**(vi)** Using this value for $L$, calculate the values of $a$ and $b$, and use them to predict the week in which Amy will obtain her first mark of at least 90%. [4]

**(vii)** Give an interpretation, in context, of the value of $L$. [1]

**[** $-0.929744$ **(vi)** $a = 4.10$ and $b = -0.280$ , week 13**]**

**8** Abi and Bhani find the fuel consumption for a car driven at different constant speeds. The table shows the fuel consumption, $y$ kilometres per litre, for different constant speeds, $x$ kilometres per hour.

| $x$ | 40 | 45 | 50 | 55 | 60 |
|---|---|---|---|---|---|
| $y$ | 22 | 20 | 18 | 17 | 16 |

**(i)** Abi decides to model the data using the line $y = 35 - \frac{1}{3}x$.

**(a)** On the grid on page 5
- draw a scatter diagram of the data,
- draw the line $y = 35 - \frac{1}{3}x$. [2]

**(b)** For a line of fit $y = \mathrm{f}(x)$, the residual for a point $(a. b)$ plotted on the scatter diagram is the vertical distance between $(a, \mathrm{f}(a))$ and $(a. b)$. Mark the residual for each point on your diagram. [1]

**(c)** Calculate the sum of the squares of the residuals for Abi's line. [1]

**(d)** Explain why, in general, the sum of the squares of the residuals rather than the sum of the residuals is used. [1]

Bhani models the same data using a straight line passing through the points (40, 22) and (55, 17). The sum of the squares of the residuals for Bhani's line is 1.

**(ii)** State, with a reason, which of the two models, Abi's or Bhani's, gives a better fit. [1]

**(iii)** State the coordinates of the point that the least squares regression line must pass through. [1]

**(iv)** Use your calculator to find the equation of the least squares regression line of $y$ on $x$. State the value of the product moment correlation coefficient. [3]

**(v)** Use the equation of the regression line to estimate the fuel consumption when the speed is 30 kilometres per hour. Explain whether you would expect this value to be reliable. [2]

**(vi)** Cerie performs a similar experiment on a different car. She finds that the sum of the squares of the residuals for her line is 0. What can you deduce about the data points in Cerie's experiment? [1]

**[ (i)(c)** $\sum_{i=1}^{5}\left(e_i\right)^2=\frac{4}{3}$ **(ii)** Bhani's model is a better fit **(iii)** $(50,\ 18.6)$
**(iv)** $y=-0.3x+33.6$ ; $r=-0.985$ **(v)** $y=24.6$ (kilometres per litre) **]**

**9** A random sample of eight pairs of values of $x$ and $y$ is used to obtain the following equations of the regression lines of $y$ on $x$ and of $x$ on $y$ respectively.

$$y=-\frac{7}{10}x+\frac{151}{10}, \quad x=-\frac{7}{6}y+20.$$

Seven of the pairs of data are given in the table.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $x$ | 10 | 11 | 12 | 11 | 17 | 14 | 19 |
| $y$ | 9 | 8 | 7 | 6 | 5 | 4 | 1 |

**(i)** Find the eighth pair of values of $x$ and $y$. [5]

**(ii)** Let $Y$ be the value obtained by substituting a sample value of $x$ into the equation of the regression line of $y$ on $x$. Evaluate $Y$ for each of the eight values of $x$ and verify that $\sum(y-Y)^2=8.8$. [3]

**(iii)** For each of the eight sample values of $x$, $Y'$ is given by $Y' = a + bx$, where $a$ and $b$ are any constants. What can you say about the value of $\sum(y-Y')^2$ ? [2]

**[(i)** $x=10,\ y=8$ **(iii)** $r=-0.904$ **]**