Chapter 3 Linear Models for Numerical Data

LINE OF BEST FIT OR REGRESSION LINE

A scatterplot establishes if a relationship exists between two variables and can be used to make an estimate of one of the variables for a given value of the other. The question arises how confident can we be in the accuracy of the estimate? When estimating a data point using the scattergram we assume that the estimate is probably close to the actual point but it may equally be far away. Hence it can be stated that the accuracy of predictions using data points are very questionable.

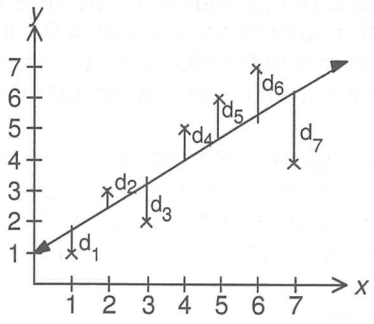
In order to use an established linear relationship between two variables for prediction purposes a **line of best fit** or **regression line** is drawn on the scatter diagram.
The line of best fit or regression line may be drawn on a scatter diagram by examining the given plotted points and then drawing a line which "fits in the best". This method of fitting a line is not very precise and hence using it for prediction purposes will be rather limited.

To obtain a mathematically based line of best fit we use the "method of least squares". This method results in a line which minimises the sum of the squares of the deviations of each point from the line. Now any point not in a line deviates from the line both vertically and horizontally. Thus when we determine a regression line we must specify which deviation has been minimised.
Note (i) We need to square the deviations as their sum will always be zero.
(ii) These deviations from the line of best fit are called the linear regression **residuals**.

Least squares regression line of y on x.

In order to find the least squares regression line for predicting a value for y given a value of x we need to minimise the sum of the squares of the vertical deviations or distances of all points from the regression line.
That is, the fitted least squares regression line is such that
$(d_1)^2 + (d_2)^2 + (d_3)^2 + (d_4)^2 + (d_5)^2 + (d_6)^2 + (d_7)^2$ is a minimum.

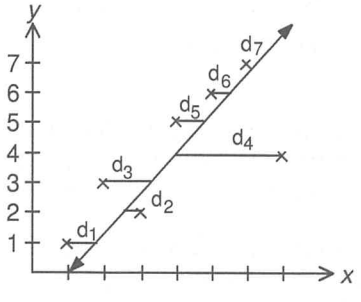
The equation of the least squares regression line for predicting a value of y for a given value of x is commonly called the ***y on x regression line*** and is given by y = ax + b where a is the gradient of the regression line and b is the y-intercept.



Least squares regression line of x on y.

In order to find the least squares regression line for predicting a value of x given a value of y we need to minimise the sum of the squares of the horizontal deviations or distances of all points from the regression line.
That is, the fitted least squares regression line is such that
$(d_1)^2 + (d_2)^2 + (d_3)^2 + (d_4)^2 + (d_5)^2 + (d_6)^2 + (d_7)^2$ is a minimum.

The equation of the least squares regression line for predicting a value of x for a given value of y is commonly called the ***x on y regression line*** and is given by x = ay + b where a is the gradient of the regression line and b is the y-intercept.



Note that the least squares regression line is that line that minimises the sum of the squares of the residuals of each point in the distribution from the line of best fit.

INTERCEPTS AND SLOPE OF REGRESSION LINES

All lines with two variables may be written in the form $y = mx + c$ which enables us to identify the slope and y intercept by inspection. Remember that vertical and horizontal lines only have one variable as the other is constant (i.e. it does not vary).
The y on x regression line is usually written in the form $\hat{y} = ax + b$ where $\hat{y}$ (read as "y hat") is the predicted value of y for some given value x, **a** indicates the slope of the line and **b** the y intercept.

The **slope** of a regression line represents the average change in the predicted value of y the response variable, for each increase of one unit in x the explanatory variable, using the concept that slope equals rise over run or a change in the value of y over a change in the value of x.

When dealing with regression lines and their **intercepts** it must be noted that the intercepts may or may not have meaning this will depend on the situation being discussed. In order to determine if an intercept has a meaning the following need to be considered:

(i) Can the value of x be zero and be meaningful? If the x value is zero and meaningful is the y value meaningful?
(ii) Can the value of y be zero and be meaningful? If the y value is zero and meaningful is the x value meaningful?
(iii) if the x intercept is negative then it has no practical meaning as negative x values are not possible in practical situations. Although a negative x intercept has no practical application it still exists as it is where the regression line will intersect with the x axis if it is extended.
(iv) if the y intercept is negative then it has no practical meaning as negative y values are not possible in practical situations. Although a negative y intercept has no practical application it still exists as it is where the regression line will intersect with the y axis if it is extended.

The concepts presented here for the **y on x regression line** may be applied to the **x on y regression line** of the form $\hat{x} = ay + b$.

Situation 1

A weight-height comparative study of a sample of senior rugby players was found to be modelled by the linear regression equation $\hat{h} = 0.6w + 118$ where $\hat{h}$ represents the predicted height of the players in centimetres and w the weight of the players in kilograms.

The slope of this equation is positive and is given by 0.6 which tells us that every increase of 1 in w (1 kg in weight) results in an increase of 0.6 in h (0.6 cm in height).
Relating this to the situation being considered we can state that on average for every kilogram increase in weight we would expect an increase of 0.6 centimetres in height.

The vertical axis intercept of 118 means that when $w = 0$ then the value of $\hat{h} = 118$. Relating this to the situation implies that for a rugby player with a weight of 0 kilograms the expected height is 118 centimetres. For this example the value of zero would not be in the domain of w and hence has no meaning. For our situation the vertical axis intercept is not relevant as a rugby player cannot have a weight of 0 kilograms.

Situation 2

A study to investigate the relationship between the number of hours per week spent engaged in part time employment x and performance factor in examinations y by full time high school students was conducted over a period of time.
Data collected and processed gave the equation of the regression line as $\hat{y} = -0.2x + 7.2$ where $\hat{y}$ represents the predicted performance factor and x the number of hours per week spent engaged in part time employment.

The slope of the regression line is negative and is given by -0.2 which tells us that for every increase of 1 in x there is a corresponding decrease of 0.2 in y. This tell us that for the situation being considered there is on average a drop of 0.2 in the performance factor for an increase of 1 hour spent in part time employment.

The y intercept is 7.2 which tells us that when $x = 0$ the expected value of y is 7.2.
In this situation the y intercept is relevant as it is known that many full time students are not engaged in part time employment.
The x intercept in this case is calculated to be $x = 36$ which tells us that when $y = 0$ the expected value of x is 36. The x intercept is not relevant in the context of the situation. To estimate a value of x for a given value of y that is $y = 0$ is not appropriate as the given regression line is for predicting a value of y for a given value of x, to estimate a value of x we need to use the x on y regression line. Furthermore for high school students to work 36 hours part time per week and attend school full time during the same week would be extremely rare if at all possible.

PROPERTIES OF THE REGRESSION LINE

1. A regression line in the form $y = ax + b$, is a mathematical model for a given bivariate data set.
2. The regression line will always pass through the **centroid of the distribution**, that is the point $(\bar{x}, \bar{y})$ as it is used to determine the regression line.
3. The gradient of the regression line **a**, is closely related to the correlation coefficient as it may be written in the following form $\mathbf{a} = \frac{s_{xy}}{s^2_x} = \frac{s_{xy}}{s_x s_y} \times \frac{s_y}{s_x} = r\frac{s_y}{s_x}$. The value of **a** informs us that a change of one standard deviation in x gives a change of r standard deviations in y.
4. The vertical axis intercept **b** is the predicted value of y when $x = 0$ and can only be considered if x can take on values close to zero and therefore in many bivariate distributions it has no meaning.
5. The closer the value of r to ± 1 the smaller the difference between the two regression lines that is, the y on x regression line and the x on y regression line. For $r = \pm 1$ the two regression lines overlap and are in fact the one and the same line indicating that there is perfect linear relationship between the two variables.
6. **Correlation** is the degree of **association** between the two variables.
   Regression is using one of the variables to **predict** the other.
7. The better the fit of the line of regression then the stronger is the correlation (association) between the two variables.
8. The least squares regression line minimises the sum of the squares of the deviations, hence the effect of outliers is very significant.
9. The interpretation of the gradient or slope of the least squares regression line equation is the average rate of change in the response variable for each increase of one unit of the explanatory variable.
10. The interpretation of the y-intercept of the least squares regression line equation is the predicted value of the response variable when the explanatory is zero. In many situations you will find that the interpretation of the y-intercept makes no sense when put into context. This is because real life data does not involve values of zero for the explanatory variable.

EXERCISE 3A

1. To study the relationship between a father's height in cm and his son's height in cm, data was collected from a large number of father-son pairs and the following least squares regression line equation determined.

   Son's height = 0.54 x Father's height + 89.58

   (a) Identify and interpret the slope of this regression line.

   (b) Identify and interpret the y intercept.

2. Shoe size and height data was collected from a large group of 16 – 19 year old students to check the claim that you can tell the size of a person's shoe by considering their height. The least squares regression line for shoe size and height was calculated to be:

   Predicted shoe size = 0.25 x height – 33.4.

   (a) State the explanatory variable and the response variable.

   (b) Identify and interpret the slope of this regression line.

   (c) Identify and interpret the y intercept.

3. The relationship between the number of hours of sunshine (per year) and the number of rainy days (per year) for a number of cities is given by $\hat{y} = 2920 - 7.4x$, where x is the number of rainy days and y the number of hours of sunshine.

   (a) Identify the explanatory variable and the response variable.

(b) Identify and interpret the slope of this regression line.

(c) Identify and interpret the y intercept.

4. The least squares regression line for the heart weight in mg and the body weight in grams of a small marsupial species is given by: Heart weight = 2.74 x body weight + 45.28

(a) Identify and interpret the slope of this regression line.

(b) Identify and interpret the y intercept.

5. In a study of the effectiveness of a pain relief drug the response time, in minutes, was measured for different drug doses in milligrams. From the study the following least squares regression line was determined. Response time = -8.54 x drug dose + 45.55

(a) Identify and interpret the slope of this regression line.

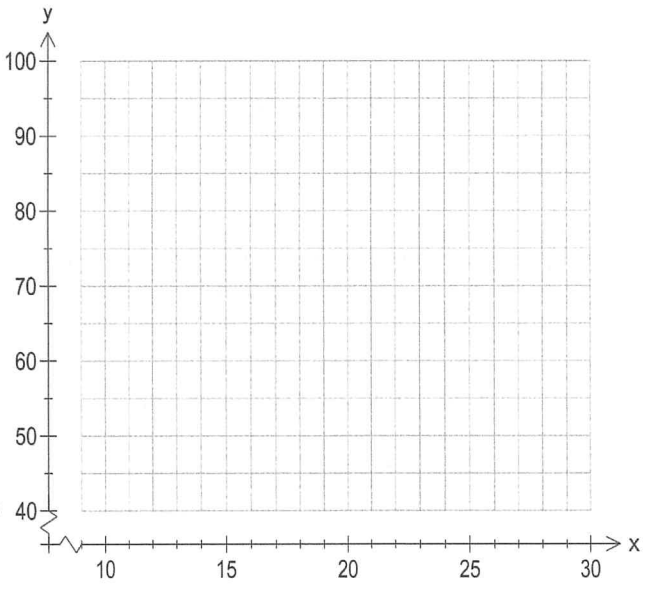
(b) Identify and interpret the y intercept.

6. Consider the following bivariate data set.

x	25	12	20	26	15	10	22	28	27	24
y	90	55	80	84	58	48	68	85	95	77

(a) Construct a scatter plot for the given data set.

(b) Calculate the value of the correlation coefficient. What does it tell you about the relationship between x and y?



(c) Determine the least squares regression line y on x. Draw your regression line on your scatter plot.

(d) Identify and interpret the slope of your regression line.

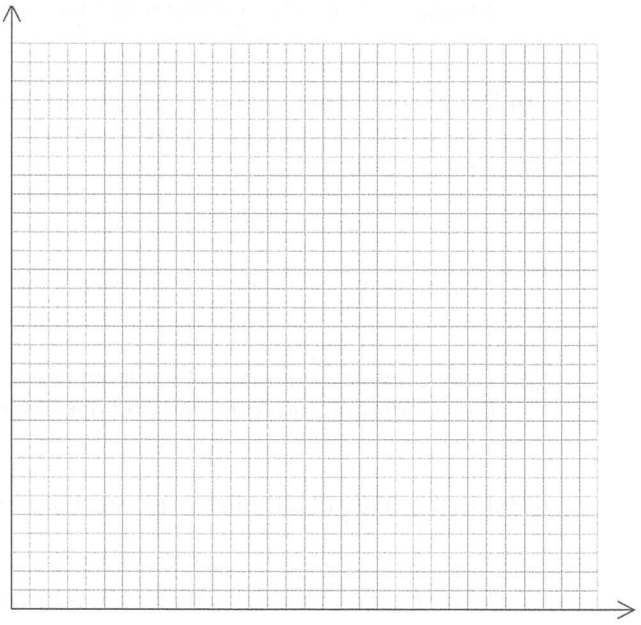
(e) Identify and interpret the y intercept.

7. A survey was conducted by the Australian Chicken Meat Federation (ACMF) comparing the number of weeks a chicken has been on a special diet and the weight gain during that period. The data collected by the ACMF is tabled below.

Number of weeks on the diet (n)	5	15	18	9	20	7	20	30	15	25
Weight gain in grams (w)	80	200	220	150	260	100	240	300	180	280

(a) Construct a scatter plot for the given data set.

(b) Calculate the value of the correlation coefficient. What does it tell you about the relationship between diet time and weight gain?



(c) Determine the equation of the least squares line that models the relationship between the time in weeks and the weight gain for the data provided. Draw this line on your scatter plot

(d) Identify and interpret the gradient of your regression line.

(e) Identify and interpret the vertical axis intercept.

8. For the financial year 2016, the owner of an ice cream vending van kept a record of the profit on sales for each month, rounded to the nearest $100, and the average monthly maximum temperature, to the nearest degree.

Month	July	Aug	Sept	Oct	Nov	Dec	Jan	Feb	Mar	April	May	June
Av. temp (°C)	15	14	17	22	30	32	36	35	32	28	22	18
Profit	15	16	19	25	29	35	44	42	38	28	20	16

(a) Identify the explanatory and response variables.

(b) Does there appear to be a relationship between the monthly maximum average temperature and profit? Refer to an appropriate graph on your calculator. There is no need to draw the graph.

(c) Calculate the correlation coefficient, r. What does r tell you about this relationship?

(d) Determine the equation of the least squares regression line that models the relationship between the monthly maximum average temperatures and profit.

(e) Identify and interpret the gradient of the least squares regression line in this context.

(f) Identify and interpret the vertical axis intercept in this context.

9. The quantity of potatoes, p kilograms, sold each day at Potato Shed and the corresponding price, d dollars per kilogram, were recorded over a period of ten days.

Day	1	2	3	4	5	6	7	8	9	10
Quantity (p kg)	220	350	210	240	280	230	180	200	290	330
Price ($d)	1.75	1.00	1.85	1.55	1.20	1.50	2.00	1.80	1.05	1.10

(a) Identify the explanatory and response variables.

(b) Does there appear to be a relationship between the quantity of potatoes sold and the kilogram price of these potatoes? Refer to an appropriate graph on your calculator. There is no need to draw the graph.

(c) Calculate the correlation coefficient, r. What does r tell you about this relationship?

(d) Determine the equation of the least squares regression line that models the relationship between the quantity of potatoes sold and price per kilogram.

(e) Identify and interpret the gradient of the least squares regression line in this context.

(f) Identify and interpret the vertical axis intercept in this context.

(g) Identify and interpret the horizontal axis intercept in this context.

10. The table give the ages of a particular model of car together with the selling price taken form the classified ads of a weekend newspaper.

Age (years)	4	1	2	4	6	9	10	12	8	2
Price ($'000)	10.3	12.5	11.7	9.9	8	7.5	6.0	5.5	8.5	12.8

(a) Identify the explanatory and response variables.

(b) Does there appear to be a relationship between the selling price of these cars and their age? Refer to an appropriate graph on your calculator. There is no need to draw the graph.

(c) Calculate the correlation coefficient, r. What does r tell you about this relationship?

(d) Determine the equation of the least squares regression line that models the relationship between the selling price of these cars and their age.

(e) Identify and interpret the slope of the least squares regression line in this context.

(f) Identify and interpret the vertical axis intercept in this context.

(g) Identify and interpret the horizontal axis intercept in this context.

THE COEFFICIENT OF DETERMINATION

The coefficient of determination is another measure that helps us to get some information about the correlation between two variables. The coefficient of determination is used to inform us how well our regression line actually represents the data set under consideration.

Consider the following data set comparing the variables student height and student weight.

Height (cm)	160	172	173	180	166	187	172	191	158	164	186	156
Weight (kg)	52.3	54.8	53.4	58.2	56.8	62.6	53.4	62.8	45.8	50.1	69.2	49.4

Using a calculator the following statistics may be displayed:

For the given data set the correlation coefficient, r, is 0.8874051.

For the given data set the coefficient of determination, r^2, is 0.7874874.

The coefficient of determination is simply the square of the correlation coefficient. That is $r^2 = (0.8874051)^2 = 0.7875$ rounded to 4 d.p.

Linear Reg

$y = ax + b$
$a = 0.4948204$
$b = -29.41701$
$r = 0.8874051$

$r^2 = 0.7874878$

The coefficient of determination gives us the variation in the response or dependent variable that is directly related to or explained by the variation in the explanatory or the independent variable.

For our example the coefficient of determination gives the proportion of the variation in weight that is directly related to the variation in the height of these students.

If we change the value of r^2 into a percentage, that is $r^2 = 0.7875 \times 100\% = 78.75\%$, then this tells us that approximately 79% of the variation in the weight of these students can be accounted for or explained by the variation of the height of these students. This indicates that there is a strong association or link between the weight and height of these students.

Now this infers that the percentage of unexplained variation is given by $100\% - 78.75\% = 21.25\%$. This tells us that approximated 21% of the variation in weight is unexplained by the variation in height.

NOTE: $1 - r^2$ is known as the **coefficient of non-determination**. The coefficient of non-determination measures the percentage of the variation in the response variable which is explained by **chance** and **other factors** or the percentage of variation in the response variable NOT explained by variation in the explanatory variable.

Correlation Coefficient and Coefficient of Determination

The **correlation coefficient**, r or r_{xy}, is also called Pearson's Correlation Coefficient.

The correlation coefficient is a measure of linear association between two variables and it gives us a measure of the **strength** and **direction** of a linear relationship.
The value of the correlation coefficient for a data set always lies in the interval $-1 \leq r \leq 1$.

The **coefficient of determination**, r^2, the square of the correlation coefficient, is a measure of how much of the variation in the **response** or **dependent variable** can be explained by the variation in the **explanatory** or **independent variable**.
The value of the coefficient of determination for a data set always lies in the interval $0 \leq r^2 \leq 1$, because squaring any negative number results in a positive result.

For a correlation coefficient of $r = 0.5$ the coefficient of determination is $r^2 = (0.5)^2 = 0.25$. Hence a correlation of 0.5 implies that 25% of the variation in the response variable is accounted for, or explained by the variation in the explanatory variable.

The coefficient of determination is a more accurate measure of the strength of relationship than the correlation coefficient, especially if we are making comparisons.
For a data set with $r = 0.8$ we would think that the strength of this relationship would be twice as strong as that of a data set with $r = 0.4$. However this is not so because an r value of 0.8 indicates a relationship that is four times as strong as one with an r value of 0.4. This is because examination of r^2 for an r value of 0.8 is 0.64 and the r^2 for an r value of 0.4 is only 0.16. Stating the obvious, 0.64 is four times 0.16.

By definition, the coefficient of determination, r^2 can be found from the correlation coefficient by squaring the correlation coefficient. For example a data set with $r = 0.7$ has a coefficient of determination (r^2) of $(0.7)^2 = 0.49$ or 49%

The correlation coefficient, r, can be determined give the coefficient of determination.

For example if $r^2 = 0.81$, then $r = \pm\sqrt{0.81} = \pm 0.9$. Now to determine whether the r value is positive or negative we need to examine either the scatterplot of the data set or the equation of the regression line. The scatterplot will identify whether the direction of the data set is positive or negative.
Identifying the gradient of the regression line will tell us whether the correlation coefficient, r, is positive or negative. If the gradient of the regression line is positive so also is the value of r, and if the gradient of the regression line is negative then the value of r is also negative.

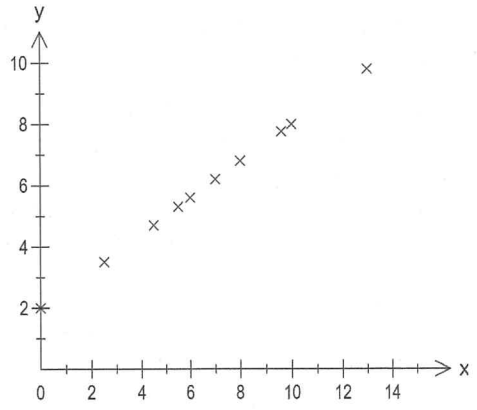
Properties of the Coefficient of Determination

- because the coefficient of determination, r^2, is the result of squaring the correlation coefficient, r, the coefficient of determination can never be negative, thus $0 \leq r^2 \leq 1$,
- $r^2 = 1$ indicates that the regression line perfectly fits the data, that is all the data points line on the regression line.
  When $r^2 = 1$ then this means that 100% of the variation in the response or dependent variable can be explained by the variation in the explanatory or independent variable.
- $r^2 = 0$ indicates that the regression line does not fit the data at all. This may be because the data is not linear, or the data is random. In such situations the regression line is a horizontal line,(i.e. slope is zero) and the y intercept is the mean of the values of the response or dependent variable.
  When $r^2 = 0$ then this means that none or 0% of the variation in the response or dependent variable can be explained by the variation in the explanatory or independent variable.
- $0 < r^2 < 1$ indicates that the regression line fits the data to some degree, the closer r^2 is to 1 the better the goodness of fit or in other words the more the variation in the response or dependent variable can be explained by the variation in the explanatory or independent variable.
  As r^2 gets closer to 1 the values of the response or dependent variable get closer to the regression line, and as r^2 gets closer to 0 the values of the response or dependent variable get further from the regression line.
  If $r^2 = 0.8$, then this means that 80% of the variation in the response or dependent variable can be explained by the variation in the explanatory or independent variable.
  The remaining 20% of the total variation in the response or dependent variable remains unexplained.
- A high value of r^2 does not imply causality. When we examine the value of r^2 we may conclude that the two variables are related but the value of r^2 does not tell us that changes in one variable cause the changes in the other.

Example 1
Consider the following bivariate data set.

x	6	10	7	13	4.5	8	9.6	2.5	0	5.5
y	5.6	8	6.2	9.8	4.7	6.8	7.76	3.5	2	5.3

(a) Construct a scatter plot and comment on the relationship between x and y.
(b) Determine the value of the correlation coefficient.
(c) Describe the data using your value of r.
(d) Determine the value of the coefficient of determination.
(e) Interpret the coefficient of determination.
(f) Determine the equation of the least squares regression line that models the relationship between x and y.
(g) Identify and interpret the slope of the least squares regression line in this context.



Solution

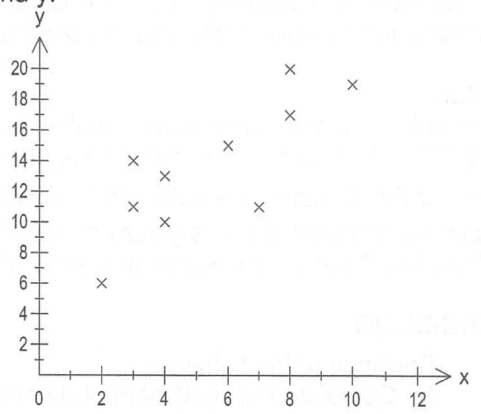
(a) The scatter graph shows a perfect positive linear relationship between x and y.
(b) Using a calculator, $r = 1$
(c) There is a perfect positive linear relationship between x and y.
(d) Now $r = 1$ hence $r^2 = (1)^2 = 1$. Coefficient of determination is 1.
(e) 100% of the variation in y can be explained by the variation in x.
(f) The equation of the regression line is $\hat{y} = 0.6x + 2$
(g) The slope is 0.6. The slope of the regression line predicts that on average we can expect an increase of 0.6 units in y for a one unit increase in x.

Example 2

Consider the following bivariate data set.

x	8	4	3	6	7	3	2	4	8	10
y	17	13	14	15	11	11	6	10	20	19

(a) Construct a scatter plot to show the relationship between x and y.
(b) Describe the relationship shown in the scatter plot.
(c) Determine the value of the correlation coefficient. Round your answer to 4 decimal places.
(d) Describe the data using your value of r.
(e) Determine the value of the coefficient of determination.
(f) Interpret the coefficient of determination.
(g) Determine the equation of the least squares regression line that models the relationship between x and y. Give the regression line parameters rounded to two decimal places.
(h) Identify and interpret the gradient of the least squares regression line in this context.



Solution

(a) See graph.
(b) The relationship between x and y can be described as a moderate to strong positive linear relationship.
(c) Using a calculator, $r = 0.8060$ (4 d.p.)
(d) There is a strong positive linear relationship between x and y.
(e) Now $r = 0.8060$ hence $r^2 = (0.8060)^2 = 0.6496$ (4 d.p.)
Coefficient of determination is 0.6496
(f) 64.96% of the variation in y can be explained by the variation in x.
(g) The equation of the regression line is $\hat{y} = 1.30x + 6.44$.
(h) The gradient is 1.30. The gradient of the regression line predicts that on average we can expect an increase of 1.30 units in y for a unit increase in x.

Example 3

Data was collect by a maths teacher to investigate the relationship between student test marks and the number of hours spent studying for the test. The collected data is displayed in the table below where the test marks have been recorded as percentages.

Number of hours of study	8	4	2	6	6	4	5	5	7	8	3	9
Test mark (%)	88	62	28	62	70	55	62	70	80	79	44	96

(a) Identify The explanatory and response variables.
(b) View, on your calculator or computer, the scatter plot of this data and describe the relationship between the variables.
(c) Determine the value of the coefficient of determination. Give your answer rounded to 3 decimal places.
(d) Interpret the coefficient of determination.
(e) Give the equation of the regression line that models this data. Round parameters to two decimal places.
(f) Is the linear regression model found in (e) appropriate for this data set? Justify your response.

Solution

(a) Explanatory variable is the number of hours of study and the response variable is the test mark given as a percentage.
(b) The scatter plot shows a linear relationship between these variables which is positive and strong.
(c) From the calculator $r^2 = 0.909$.
(d) 90.9% of the variation in the test mark percentage can be explained by the variation in the number of hours of study.
(e) From the calculator $y = 8.32x + 19.88$. Using the correct variable names that model this data the required equation is given by: Test mark = 8.32 x number of hours of study + 19.88
(f) Yes the model is appropriate. The regression model is appropriate because the relationship between the variables is linear and strong with a high value of r^2. The value of r^2 informs us that only 9.1% (100 – 90.9) of the variation between the variables is unexplained by the model.

Example 4

The correlation coefficient for a bivariate data set was found to be -0.75. Find and interpret the coefficient of determination.

Solution

$r^2 = (-0.75)^2 = 0.5625$. Coefficient of determination is 0.5625.
This coefficient of determination tells us that 56.25% of the variation in the response variable can be explained by the variation in the explanatory variable.

Example 5

The least squares regression line y = 11.4 – 0.35x was calculated using latitude measured in degrees and average January temperature in degrees Celsius for 25 cities. The value of the coefficient of determination was given as 0.7257 correct to 4 decimal places.

(a) Identify the explanatory and response variables.
(b) What does the coefficient of determination tell you about these variables.
(c) Calculate the value of Pearson's correlation coefficient.

Solution

(a) Latitude (x) is the explanatory variable average January temperature (y) is the response variable.
(b) 72.57% of the variation in average January temperature is explained by the variation in latitude.
(c) $r^2 = 0.7257$, hence $r = \pm\sqrt{0.7257} = \pm 0.8519$ (4 d.p.)
Now the gradient of the regression line is negative hence the r value is also negative.
Therefore Pearson's correlation coefficient is -0.8519

EXERCISE 3B

1. For each of the following:
   (i) Calculate the coefficient of determination for each of the following values of r giving your answer correct to 3 decimal places.
   (ii) Interpret the coefficient of determination.

   (a) r = 0.58

   (b) r = - 0.709

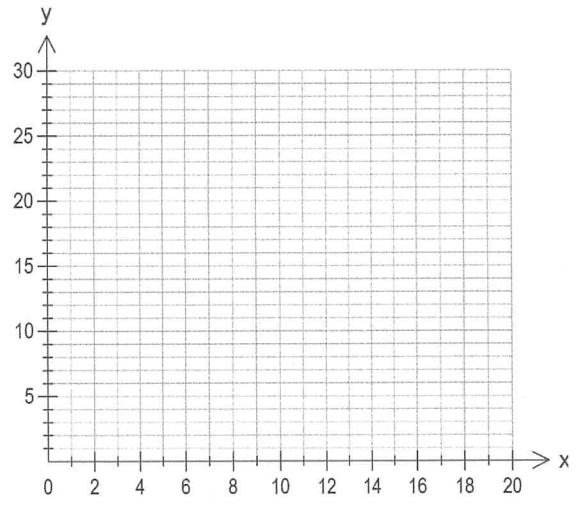
   (c) r = 0.421

   (d) r = -0.8999

2. Consider the following bivariate data set.

x	6	13	10	8	18	4	17	5	2	15	17	9
y	16	20	25	14	28	9	29	8	11	27	24	20

   (a) Construct a scatter plot and comment on the relationship between x and y.



   (b) Determine the value of the correlation coefficient.

   (c) Describe the data using your value of r.

   (d) Determine the value of the coefficient of determination.

   (e) Interpret the coefficient of determination.

   (f) Determine the equation of the least squares regression line that models the relationship between x and y.

   (g) Identify and interpret the slope of the least squares regression line.

   (h) Is the linear regression model appropriate? Justify your answer.

3. The following table gives the age of cars, in years and the cost of repairs in hundreds of dollars of 15 cars of a particular model.

Age of car	12	9	11	8	10	4	3	9
Cost of repairs ($'00)	34.2	23.8	21.3	19.3	30.2	18.2	11.6	29.3

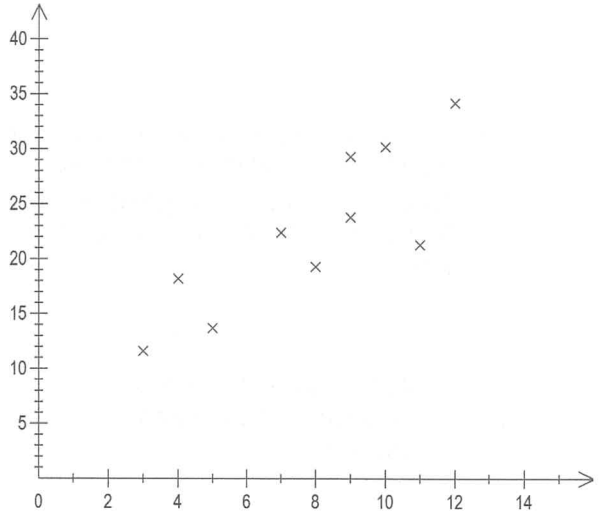
Age of car	7	5	**14**	**11**	**6**	**10**	**12**
Cost of repairs ($'00)	22.4	13.7	**30.4**	**29.9**	**19.0**	**28.0**	**27.1**

(a) Identify the explanatory and response variables.

(b) Complete the scatter plot of the given data by labelling the axes and adding the five points tabled in **bold**.

(c) Determine, correct to 4 decimal places the coefficient of determination.

(d) Interpret the coefficient of determination.



(e) Find the equation of the least squares regression line that models this data and add it to the scatter plot.

(f) Is the model found in (e) appropriate? Justify your response.

4. The table shows the number of widgets purchased at one time and the cost per widget.

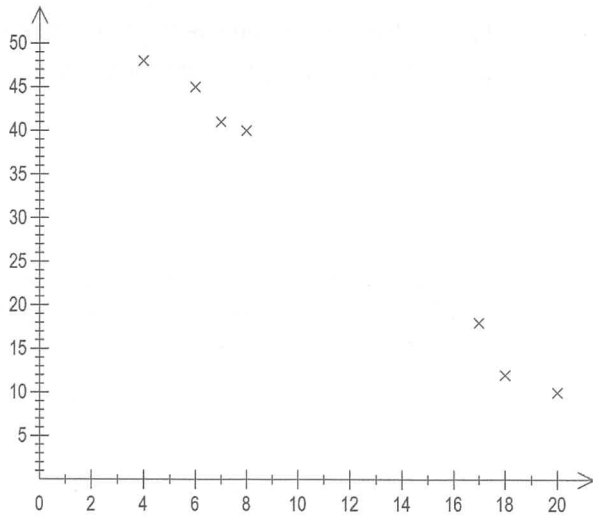
Number purchased	4	6	7	8	**9**	**10**	**12**	**14**	**15**	17	18	20
Cost per widget (cents)	48	45	41	40	**38**	**34**	**32**	**22**	**20**	18	12	10

(a) Identify the explanatory and response variables.

(b) Complete the scatter plot of the given data by labelling the axes and adding the five points tabled in **bold**.

(c) Determine, correct to 4 decimal places the coefficient of determination.

(d) Interpret the coefficient of determination.



(e) Find the equation of the least squares regression line that models this data and add it to the scatter plot.

(f) Identify and interpret the slope of the least squares regression line.

(g) Identify and interpret the y intercept.

(h) Is the model found in (e) appropriate? Justify your response.

5. The number of hours students spent playing video games and their History test marks (%) were recorded for a group of Year 12 students. The recorded results are displayed in the table below.

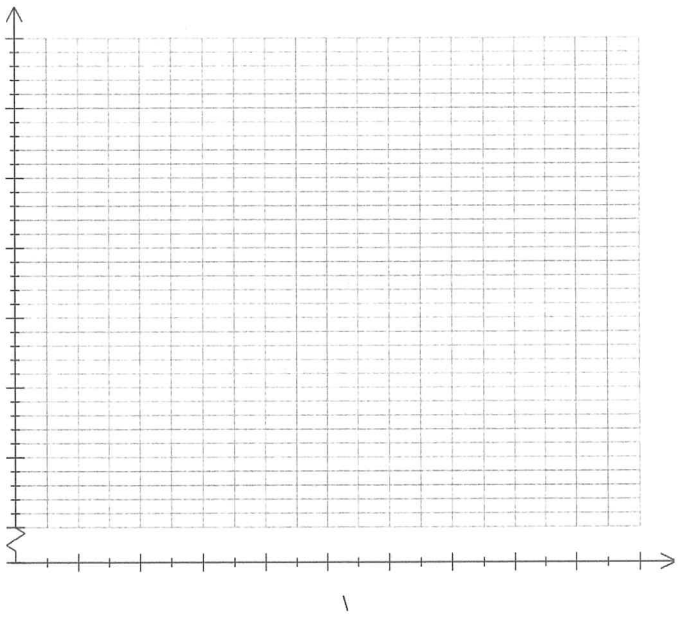
History(%)	72	53	94	36	91	90	68	43	86	72	69	85
Video game hours	6	9	1	17	3	0	5	18	7	5	10	5

(a) Identify the explanatory and response variables.

(b) On the given axes construct a scatter plot to represent these variables.

(c) Describe the relationship between these variables shown in the scatter plot.

(d) State the value of Pearson's correlation coefficient and interpret it.



(e) State the coefficient of determination and interpret it.

(f) Find the equation of the regression line that models the data.

(g) Identify and interpret the gradient of the least squares regression line.

(h) Identify and interpret the y-intercept of the least squares regression line.

(i) Add the regression line to your scatter plot.

(i) Is this model appropriate for the given data set? Justify your answer.

6. Mike is conducting research on monthly expenses for medical care. He collected data from 578 families noting the monthly expense, in dollars, and the number of people in each of these families. On entering the data into his calculator the following screen was displayed

(a) Identify the explanatory and response variables.

(b) Write the equation of the least squares regression line in the context of Mike's research.

(c) What percentage of the variation in medical expenses is explained by the size of the family?

(d) Is the linear model that Mike is considering based on the calculator display appropriate? Justify.

Linear Reg

$y = ax + b$

a = 16.831330
b = 110.472056
r = 0.695421
r^2 = 0.483610

7. The regression model for a data set is given by $\hat{y} = 467.5 - 12.8x$. If the coefficient of determination for the data set is 0.687, what is the value of the correlation coefficient? Give your answer rounded to four decimal places.

8. The value of the correlation coefficient is reported by a researcher to be $r_{ab} = -0.5$. Determine the coefficient of determination for this data and explain what it means.

9. Medical research data on length of the left forearm (cm) and height (cm) of a group of students is displayed in the table below.

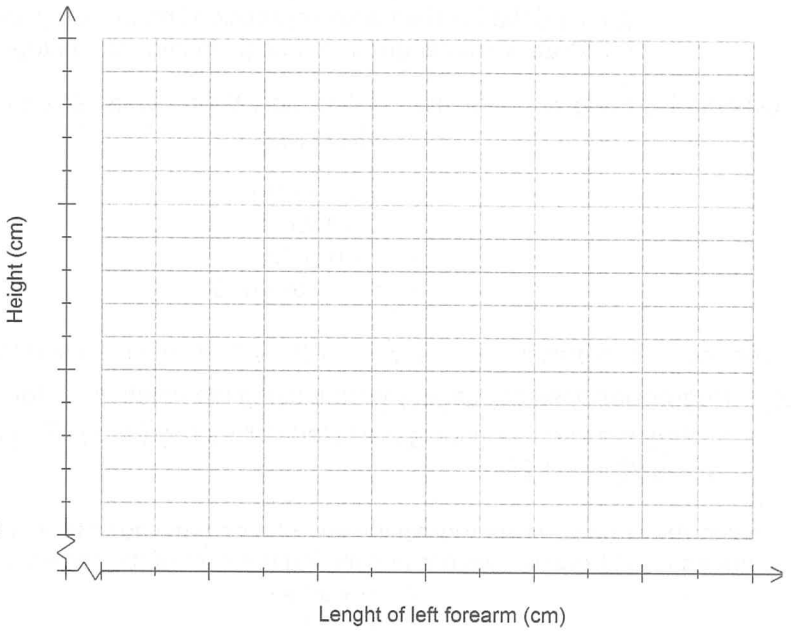
Left forearm length (cm)	22.8	29.1	24.3	20.4	21.8	27.2	25.4	28.4	30.6	22.1	20.6	30.7
Height (cm)	169	178	173	167	174	176	176	180	184	169	171	182

(a) Identify the response and explanatory variables.

(b) Construct a scatter plot to represent the data.

(c) Determine the value of the correlation coefficient and interpret it.

(d) Determine the value of the coefficient of determination and interpret it.



(e) Find the equation of the least squares regression line that models the data.

(f) Find the gradient of the regression line and explain the gradient in context of the variables.

(g) Find the y-intercept of the regression line and explain the y-intercept in context of the variables.

(h) Add the regression line to your scatter plot.
(i) Is the linear model appropriate for this data. Justify your response.

10. If the coefficient of correlation is 0.90, what is the percentage of the variation in the response variable explained by the variation in the explanatory variable?

11. The least squares regression line is given by $\hat{y} = 1.4x - 55.3$. If the coefficient of determination is 0.49, what is the coefficient of correlation?

POPULATION AND SAMPLE REGRESSION LINES

When calculating a population regression line all the data from the population under consideration must be used and for any given population the regression line will have fixed parameters for the slope and intercept. Sample regression lines are calculated by taking subsets of data points from the population, these subsets must be **large** and the data points **randomly** selected such that these subsets are representative of the population.

Different random samples from the same population will have different means, standard deviations, covariance, correlation coefficient and obviously different regression lines. Also these sample regression lines will have different parameters to those of the population regression line.

If a sample is large, randomly selected and represents the population from which it is drawn then its regression line should not be significantly different to that of the population regression line (or another sample from the same population) and can be used with a high degree of confidence.

Example 6 Consider the given bivariate data below:

x	5	1	7	3	5	4	2	6	3	6
y	4	1	6	4	6	5	4	7	2	4

(a) Find the least squares regression line of y on x and state how it is used.
(b) Find the least squares regression line of x on y and state how it is used.
(c) Draw a scatter graph of the given information together with both of the regression lines.

(a) On entering the given data into the calculator the following screen may be obtained.

Linear Reg
y = ax + b
a = 0.7261904
b = 1.25
r = 0.76725
r^2 = 0.5886726

The required regression line is $\hat{y} = 0.726x + 1.25$ and it is used to predict a value for y given a value of x.

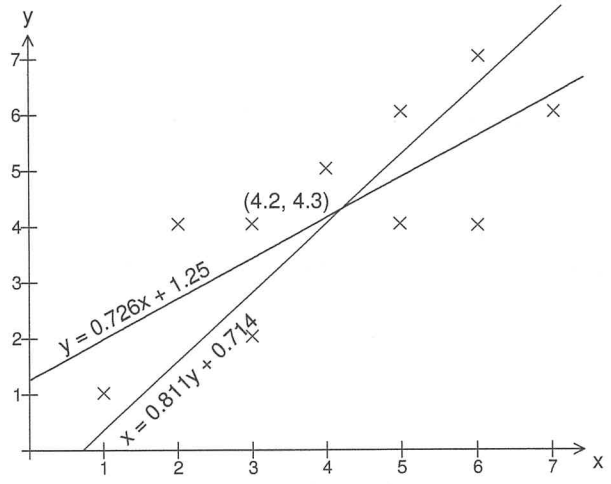
Note: Regression lines are usually written using the notation $\hat{y}$ (or $\hat{x}$) which is read as "y hat" (or "x hat") as they are used for finding predicted values of y (or x). Thus the regression line may be written as $\hat{y} = 0.726x + 1.25$.

(b) To find the least squares regression line of x on y we need to instruct the calculator to read the ordered pairs as (y, x) because we are minimising the sum of the squares of the horizontal deviations.

Linear Reg
y = ax + b
a = 0.8106312
b = 0.7142857
r = 0.76725
r^2 = 0.5886726

The required regression line is $\hat{x} = 0.811y + 0.714$ and it is used to predict a value for x given a value of y.

(c)



NOTE: The point of intersection of the pair of regression lines is (4.2, 4.3) the respective means of the variables x and y. The mean of the x values is 4.2 and the mean of the y values is 4.3.

MAKING PREDICTIONS, INTERPOLATION AND EXTRAPOLATION

Interpolation is predicting a value within the range of the given data. Interpolation is usually fairly reliable for strong correlation between the variables.

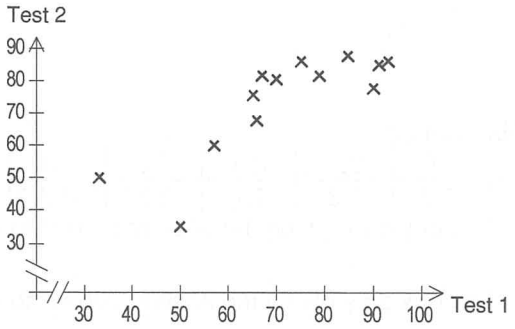
Extrapolation is predicting a value outside the range of the given data. Extrapolation, very close to the range of the data is fairly reliable if it is associated with a very strong correlation between the variables. Extrapolations any distance from the given data should be considered with caution and are usually unreliable predictions.

Example 7 The test results for an Unit 3 Mathematics class are shown below.

Student	1	2	3	4	5	6	7	8	9	10	11	12	13
Test 1(x)	90	67	50	75	70	65	85	91	57	33	66	93	79
Test 2(y)	78	82	35	86	81	76	88	85	60	50	68	86	82

(a) Draw a scatter diagram of the test results to determine if a linear relationship exists between them.
(b) Find the regression line for predicting a test 2 mark using a 2 decimal place degree of accuracy.
(c) Predict a test 2 mark for a student that scored 83 in test 1.
(d) Predict a test 1 mark for a student that scored 83 in test 2.
(e) Comment in the reliability of your predictions in (c) and (d).

(a) The scatter diagram and the correlation coefficient of approximately 0.82 indicate that there is a strong positive linear relationship between test 1 and test 2 marks.



(b) From the calculator we obtain a = 0.75365833 and b = 20.2215904
Hence the equation of the regression line for predicting a test 2 mark is given by:
Predicted test 2 mark = 0.75 x (test 1 mark) + 20.22

(c) When test 1 mark = 83
Predicted test 2 mark = 0.75(83) + 20.22
= 82.47
Predicted test 2 mark is 82.

(d) To predict a test 1 mark we need to find the linear regression line of x on y
From the calculator we obtain a = 0.88627759 and b = 5.60248783
Hence: Predicted test 1 mark = 0.89 x (test 2 mark) + 5.60
When test 2 mark = 83
Predicted test 1 mark = 0.89(83) = 0.89(83) + 5.60 = 79.47
Predicted test 1 mark is 79.

(e) The predictions in (c) and (d) may be considered to be reasonably reliable as both are interpolations and the linear association between the tests is strong as indicated by the correlation coefficient value of 0.82. The size of the data set is small and this will reduce the reliability of the predictions

EXERCISE 3C

1. Consider the following bivariate distribution.

x	10	8	8	7	7	7	6	5
y	13	15	16	12	8	6	4	3

(a) Find the least squares regression line of y on x giving the answer correct to 4 decimal places.

(b) Find the least squares regression line of x on y giving the answer correct to 4 decimal places.

(c) Predict a value for y when x = 9.

(d) Predict a value for x when y = 10.

2. Consider the following bivariate distribution.

x	1	3	5	7	8	9	11	14
y	9	8	11	7	14	13	15	16

(a) Find the least squares regression line of y on x giving the answer correct to 4 decimal places.

(b) Find the least squares regression line of x on y giving the answer correct to 4 decimal places.

(c) Predict a value for y when x = 6.

(d) Predict a value for x when y = 12.

3. Consider the following bivariate distribution.

x	46	38	34	57	46	47	38	42
y	48	48	33	54	46	47	44	43

(a) Find the least squares regression line of y on x giving the answer correct to 4 decimal places.

(b) Find the least squares regression line of x on y giving the answer correct to 4 decimal places.

(c) Predict a value for y when x = 50.

(d) Predict a value for x when y = 50.

4. Consider the following bivariate distribution.

x	2.8	2.5	1.1	1.5	2.7	2.4	2.2	1.4	1.9	2.6
y	1.1	1.3	1.1	1.2	1.3	1.7	1.5	2.3	1.1	1.4

(a) Find the least squares regression line of y on x giving the answer correct to 4 decimal places.

(b) Find the least squares regression line of x on y giving the answer correct to 4 decimal places.

(c) Predict a value for y when x = 2.1.

(d) Predict a value for x when y = 1.6.

5. Shown below in the table are the values of two related variables x and y.

x	4	8	2	7	9	6	4	3	8	7
y	3	7	1	5	6	7	4	3	6	4

(a) Find the equation of the regression line of y on x.

(b) Estimate the y value for x = 5.

(c) Estimate the y value for x = 13.

(d) Which estimate is more reliable? Why?

6. Shown below in the table are the values of two related variables x and y.

x	4	8	2	7	9	6	4	3	8	7
y	3	7	1	5	6	7	4	3	6	4

(a) Find the equation of the regression line of x on y.

(b) Estimate the x value for y = 2.

(c) Estimate the x value for y = 13.

(d) Which estimate is more reliable? Why?

7. Consider the bivariate data set shown in the table below.

p	50	25	35	45	75	20	80	25	70	40
q	25	65	40	20	25	70	20	60	10	35

(a) Determine the correlation coefficient r_{pq} giving your answer correct to 2 d.p.

(b) Determine the equation of the regression line of p on q.

(c) Determine the equation of the regression line of q on p.

(d) Using the appropriate regression line predict the value of p for q = 30.

(e) Using the appropriate regression line predict the value of q for p = 30.

8. The table shows the relationship between the explanatory variable m and the response variable n.

m	3	13	7	4	10	5	11	20	1	16	9
n	26	46	34	28	40	30	42	60	22	52	38

(a) Determine the equation of the regression line for predicting the n value given the value for m.

(b) Determine the equation of the regression line for predicting the m value given the value for n.

(c) Determine the correlation coefficient r_{mn} for this bivariate data set. What does the value of the correlation coefficient tell you about the equations of the regression lines found in (a) and (b)?

(d) Verify your answer to (c) by expressing both regression line equations in terms of the explanatory variable.

(e) For what other value(s) of r does the above hold?

(f) (i) Predict n for m = 15. (ii) Predict m for n = 50.

(g) The predicted value for n is some value p when m = q. What will be the prediction for m when n is p? Why?

9. The table below gives the height and weight of 10 football players belonging to an AFL team.

Height (cm)	198	203	184	205	194	190	207	188	201	199
Weight (kg)	89	98	80	101	83	91	103	85	97	88

(a) Identify the explanatory and response variables.

(b) Determine the coefficient of correlation between the heights(h) and weights(w) of these football players. Interpret this value of the correlation coefficient.

(c) Is there a linear relationship between h and w? Discuss.

(d) Determine the equation of the regression line you would use to predict the height of a football player if his weight was known.

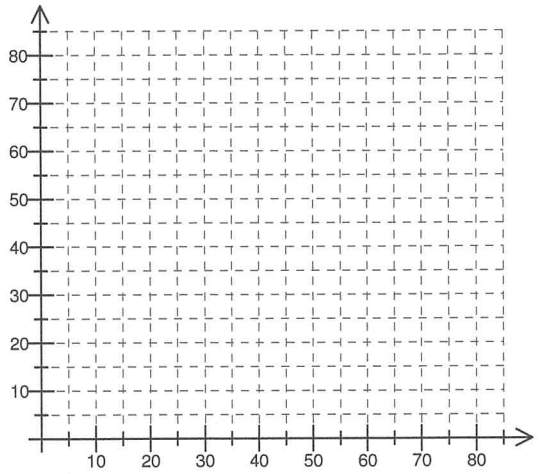
(e) Predict the height (to the nearest cm) of a football player belonging to this team if his weight is 100 kg.

(f) Predict the weight of a football player of height 219 cm. How valid is this prediction?

10. A class of students sat for a Mock ATAR examination in Chemistry in September and the ATAR in November. The marks obtained by the students in both exams were as follows:

Student	1	2	3	4	5	6	7	8	9	10
Mock Atar (m)	38	41	45	49	54	57	60	63	71	73
Atar (t)	41	44	51	52	56	60	65	61	78	80

(a) Draw a scatter diagram to establish whether a linear relationship exists between the variables m and t. Comment on your findings.



(b) Draw a line of best fit 'by eye' to enable you to predict a ATAR score for a given Mock ATAR mark.

(c) A student obtained a mark of 65 in the Mock ATAR and was absent for the ATAR. Show how to use your line of best fit find this student's probable ATAR mark.

(d) Will your answer to (c) above be the same as other class members? Explain.

(e) Using the method of least squares, find the equation of the line of best fit to predict ATAR marks in Chemistry for the class.

(f) Using the calculated regression line determine a ATAR mark for the student that did not sit the exam.

(g) If a student had scored 50 in ATAR Chemistry what would have you expected him to score in the Mock ATAR.

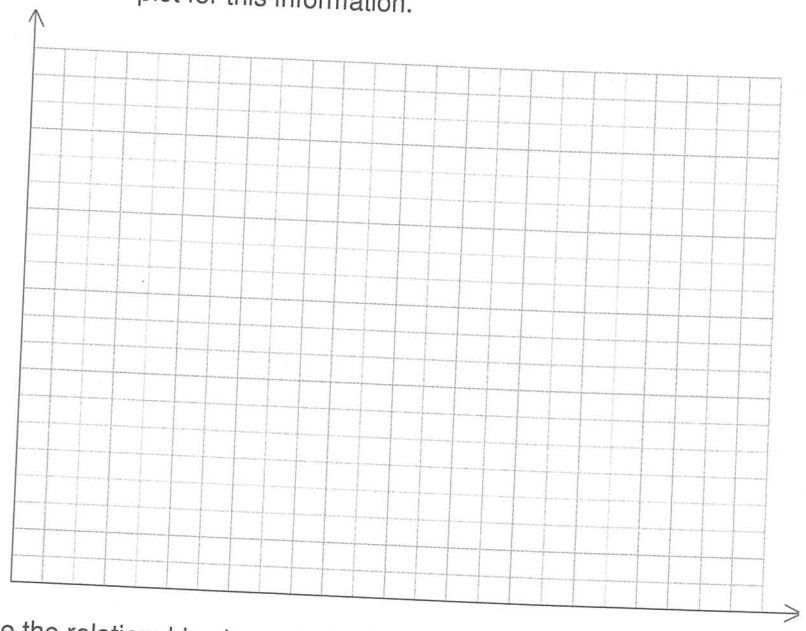
(h) Comment on the reliability of your prediction for (g).

11. The music exam marks (m) and the number of hours of practice (h) by each student are given in the following table.

Student	JM	HJ	RT	BA	PM	GD	VG	LW	BF	AP
Hours of practice (h)	95	50	34	105	55	70	115	85	48	81
Exam mark (m)	87	62	28	90	50	62	94	76	45	72

(a) Identify the response and explanatory variables.

(b) Sketch a scatter plot for this information.



(c) Describe the relationship shown in the scatter plot

(d) Determine the value of the correlation coefficient and describe the data using it value.

(e) Determine the equation of the least squares regression line for predicting exam marks and add it to your scatter plot.

(f) Identify and interpret the slope of your regression line in context.

(g) Identify and interpret the y-intercept of your regression line in context.

(h) Student MO who had practised for 20 hours wasn't able to sit the examination due to illness. Predict an exam mark for MO and comment on the reliability of your prediction.

(i) If both regression lines, m on h and h on m were graphed on the same set of axes, determine their point of intersection. What is significant about this point of intersection?

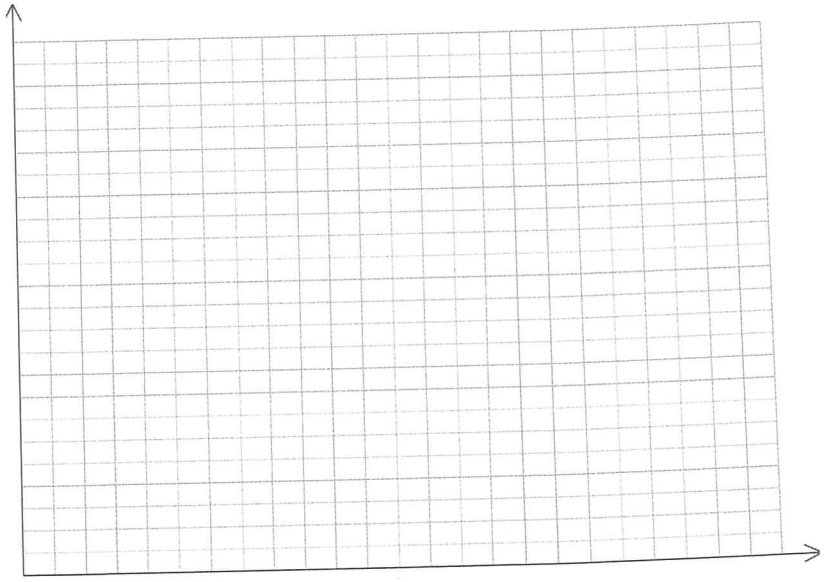
12. Data was collected regarding the weight loss (w) in kilograms and the daily calorie intake (c) for a group of twenty overweight people. The collected data is presented in the table below.

Daily calorie intake (c)	1000	1100	1100	1200	1300	1400	1400	1400	1600	1700
Weight loss (w)	2.1	2.0	1.8	1.8	1.7	1.9	1.6	1.5	1.2	1.6

Daily calorie intake (c)	1700	1700	1800	1900	2000	2000	2100	2200	2400	2500
Weight loss (w)	1.4	1.2	1.3	1.1	1.1	0.9	0.8	0.7	0.7	0.6

(a) State the explanatory and response variables. Justify you choice.

(b) On the given axes construct a scatter plot to represent the data.



(c) Describe the relationship, if any, shown in the scatter plot.

(d) Calculate Pearson's coefficient and interpret its meaning.

(e) Determine the line of regression that could be used to predict the weight loss from the daily calorie intake. Sketch the regression line on your scatter plot constructed in part (b).

(f) Interpret the vertical axis intercept of your regression line in context.

(g) Interpret the horizontal axis intercept of your regression line in context.

(h) Interpret the gradient or slope of your regression line in context.

(i) Use your regression line to predict the weight loss for a person whose intake is 5000 calories. Comment on your prediction.

(j) If a person wishes to lose a kilogram, what advice would you give this person with regards to calorie intake?

(k) Comment on the reliability of your answer to (i).

13. The following table shows the Probability, Matrices and Statistics test results(%) for 18 mathematics students.

Probability(p)	65	85	66	91	57	33	66	92	93	79	90	67	50	90	75	70	93	75
Matrices(m)	76	88	64	85	33	50	68	78	86	82	78	82	35	86	86	81	87	87
Statistics(s)	59	74	47	62	56	31	65	66	78	56	66	64	45	67	63	62	78	63

(a) Find the correlation coefficients r_{pm}, r_{ps} and r_{ms}.

(b) Give the equation of the regression line that will enable you to predict a Matrices mark given a Probability mark.

(c) Give the equation of the regression line that will enable you to predict a Probability mark given a Matrices mark.

(d) Give the equation of the regression line that will enable you to predict a Statistics mark given a probability mark.

(e) Give the equation of the regression line that will enable you to predict a Matrices mark given a Statistics mark.

(f) Which of the above tests is the best indicator of performance in the Matrices test. Justify using statistics.

(g) A student with a score of 88% in Probability test and 76% in the Statistics test did not sit the Matrices test due to illness. Using the better indicator of performance in the Matrices test estimate a Matrices test mark for this student.

14. A researcher discovered a relationship between an ant's speed, S, measured in metres per second, and the temperature of the ant's surroundings, T in degrees Celsius.
His research included the following data.

Temperature (T)	10	14	15	18	20	22	25	28	30	33
Speed (S)	1.1	1.8	2.9	3.4	4.2	4.8	5.3	5.8	6.4	7.2

(a) Determine the equation of the least squares line that models the relationship between the temperature in degrees Celsius and the speed in metres per second for the data provided.

(b) Predict the average speed of ants if the surrounding temperature is 24 degrees Celsius.

(c) Which would be the more accurate prediction: The speed when the surrounding temperature is 24 degrees or when the surrounding temperature is 8 degrees Celsius? Justify your answer.

(d) The data show an increasing trend in the value of the speed of the ants as the temperature increases. What does the equation as determined in part (a), indicate is the rate of increase in speed?

(e) What percentage of the variation in speed can be explained by the variation in temperature.

OUTLIERS AND REGRESSION LINES

An outlier is a point of the data under consideration that differs from the overall pattern as indicated by a scatterplot and it has a large residual value.
Consider the scatterplot shown below on the left. The scatterplot indicates that there is a strong positive linear relationship between the variables x and y if we ignore the point (1, 9). The point (1, 9) differs from the other data points and the graph on the right shows that the residual value of the point (1, 9) is very large in comparison to the residual values of all of the other points. The point (1, 9) is clearly an outlier of this distribution.

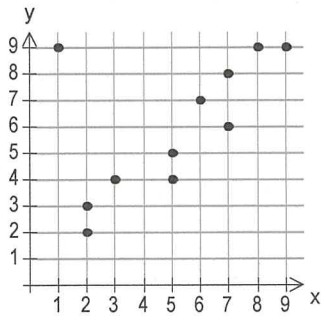
To calculate the residual values shown in the graph on the left we proceed as follows:

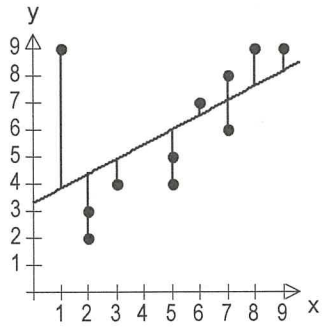
Determine the equation of the least squares regression line.

in this case the equation of the regression line is $\hat{y} = 0.54x + 3.29$

Residual = Observed value – Predicted value

Residual $(y - \hat{y})$ = 9 – (0.54(1) + 3.29) = 5.17

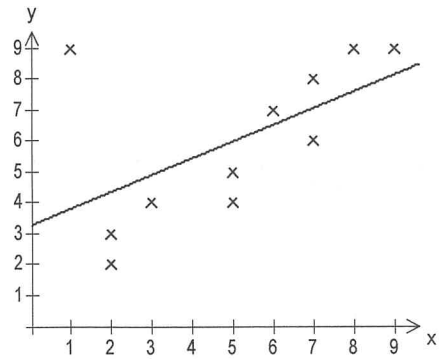




Removal of Outliers

An outlier present in a distribution will **skew the regression line towards it** and will add to the unreliability of any use that is made of the regression line. If an outlier is a consequence of an error in the collection of the data or in the recording of the data then it should be removed before applying any statistical methods.
If an outlier is a real data point then it should not be removed unless there is evidence to suggest that the data point in question has special characteristics which are not present in the rest of the distribution.
Furthermore if there is no evidence to suggest that the outlier should be removed then a different regression model may need to be considered.
Alternatively, both sets of statistics could be presented to interested viewers of the information that is with the outlier present and without the presence of the outlier and allow the user of the information to make their own conclusion.
The set of graphs below compare the distributions with and without the outlier.

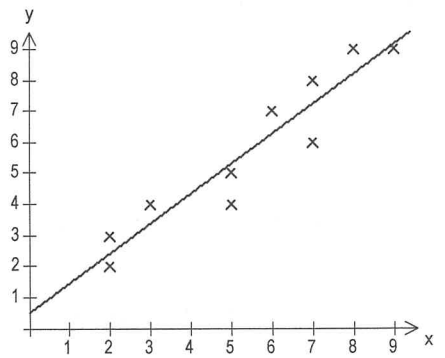
Outlier included



Regression line: $\hat{y} = 0.54x + 3.29$

Correlation coefficient: r = 0.5658

Outlier not included



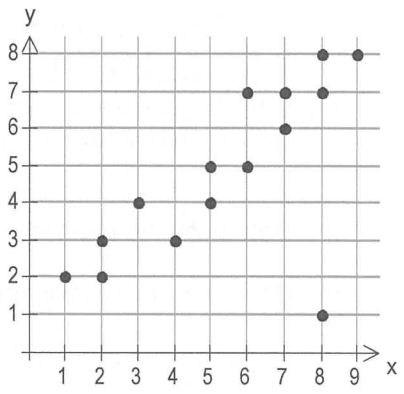
Regression line: $\hat{y} = 0.96x + 0.52$

Correlation coefficient: r = 0.9449

Examination of the graphs above shows that the outlier has a great effect on the regression line and the correlation coefficient. The outlier appears to have "pulled" the line towards itself resulting in very significant differences in the regression line and the correlation coefficient.
Using the outlier included regression line results in a predicted value of 5.45 for y when x = 4 and a significantly different predicted value of 4.36 when x = 4 when using the outlier not included regression model.
When a data point (or points) is intentionally left out in the calculation of statistics we say that the data has been **cropped**. When cropping data it must be made clear to the reader of the cropped statistics that the statistics have been obtained from data that has been cropped and a reason or reasons given for cropping the data.
Cropping data strengthens the linear relationship (or other relationships) between the two variables but it does not present the "true picture" and care must be taken when cropping in making estimates and drawing conclusions about the cropped data.

EXERCISE 3D

1. Consider the bivariate distribution given by the scatter graph shown on the right.

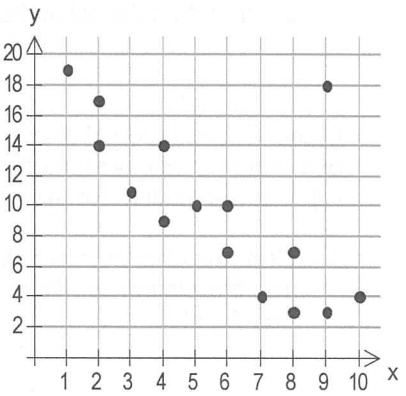


(a) Determine the following statistics:
(i) mean of x (ii) median of y
(iii) standard deviation of x (iv) r_{xy}
(v) y on x regression line (vi) r^2

The given data includes an outlier.
(b) Identify the outlier.

(c) Determine the following statistics after cropping the outlier.
(i) mean of x (ii) median of y (iii) standard deviation of x
(iv) r_{xy} (v) y on x regression line (vi) r^2

2. Consider the bivariate distribution given by the scattergram shown on the right.



(a) Determine the following statistics:
(i) mean of x (ii) mean of y
(iii) standard deviation of x (iv) r_{xy}
(v) y on x regression line (vi) r^2

The given data includes an outlier.
(b) Identify the outlier.

(c) Determine the following statistics after cropping the outlier.
(i) mean of x (ii) mean of y (iii) standard deviation of x
(iv) r_{xy} (v) y on x regression line (vi) r^2

3. Consider the following bivariate distribution.

x	24	35	28	48	67	20	29	44	53	57	61	65	39	22	15	19
y	61	51	56	68	26	63	50	42	42	39	33	35	41	52	60	53

(a) Calculate the value of the correlation coefficient for the given data and comment on the relationship between x and y.

(b) Determine the regression line for predicting a value of y for a given value of x for the given data.

The given data contains an outlier.
(c) Identify the outlier.

(d) Determine the value of the correlation coefficient without the outlier and comment on the relationship between x and y.

(e) Determine the regression line for predicting a value of y for a given value of x for the cropped data.

4. Consider the following bivariate distribution.

x	1.3	3.6	2.3	2.1	1.7	3.8	2.0	0.4	0.9	2.6	2.7	1.2	2.9	3.2	3.1	3.9
y	2.4	3.6	3.3	3.0	2.7	1.2	2.2	1.9	1.5	3.0	3.8	2.0	2.9	3.6	3.4	3.9

(a) Calculate the value of the correlation coefficient for the given data and comment on the relationship between x and y.

(b) Determine the regression line for predicting a value of y for a given value of x for the given data.

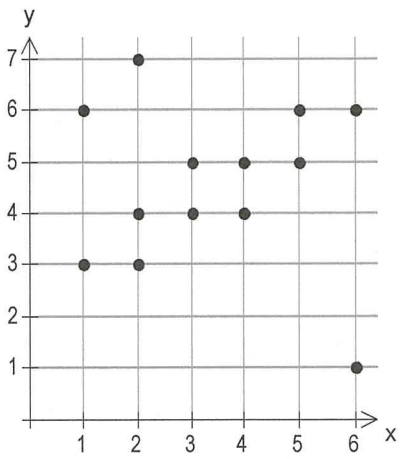
(c) Using your regression model predict y for x = 3.7.

(d) The given data contains an outlier. Identify the outlier.

(e) Eliminate the outlier from the data set and calculate the value of the correlation coefficient for the cropped data and comment on the relationship between x and y.

5. The scattergram shows the relationship between the variables x and y.

(a) Calculate the value of the correlation coefficient and comment on the relationship between x and y.



Closer examination of the scattergram reveals that three data points do not appear to fit the data set.

(b) List the outliers for this data set.

(c) Crop the data set to exclude these outliers and recalculate the correlation coefficient. Comment on the cropped relationship.

(d) If the outliers were real values of the distribution, justify their removal.

6. Consider the following bivariate distribution.

x	37	60	12	71	10	88	12	15	20	63	66	19	69	23	74
y	51	38	20	55	32	67	26	30	25	44	50	31	45	35	53

x	77	30	26	80	85	28	21	33	83	73	86	90	8	79	32
y	51	39	41	60	59	36	37	45	63	49	65	62	22	57	49

(a) Calculate the correlation coefficient for the given data and comment on the relationship between x and y.

(b) Determine the regression line for predicting a value of y for a given value of x for the given data.

(c) Use your regression line to estimate a value of y for x = 50.

(d) Examine the scatterplot of the given data and comment on observed relationship between the two variables.

(e) Comment on the validity of your answers to parts (a), (b) and (c) found above.

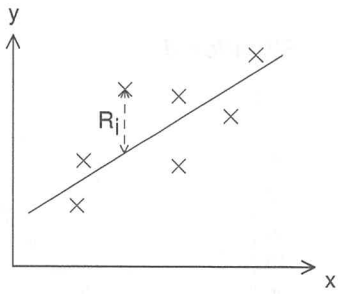
TESTING THE LINEAR REGRESSION MODEL - RESIDUALS

On establishing a linear relationship between two variables on the basis of the correlation coefficient in conjunction with the scatterplot of the data, a line of best fit is fitted to the data to enable predictions to be made. This line of best fit is obtained by the "method of least squares" and we call this line of best fit the linear regression model.

This linear regression model may or may not be suitable for prediction purposes as a high correlation coefficient between the two variables does not necessarily imply that the linear regression model is the most suitable for prediction purposes.

In order to determine the suitability of a linear regression model for prediction purposes we need to examine the **residuals** by constructing a **residual plot**.

The diagram on the right shows the regression line and a residual R_i.



A **residual** is the difference between an observed data point and a fitted or predicted value. It is the distance of a point from a curve, and is positive if above the curve and negative if below the curve. The residual R_i is calculated as follows:

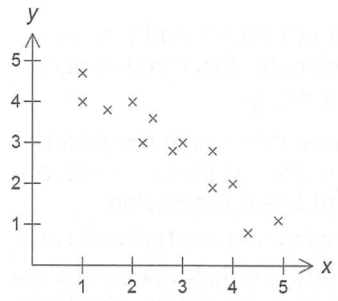
That is R_i = observed value – fitted value.

$= y_i - \hat{y}_i$

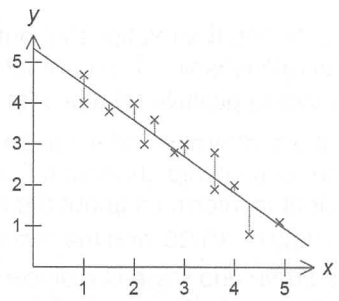
where $\hat{y}_i = ax_i + b$ is the value predicted by the linear regression model.

Consider the following bivariate situations:

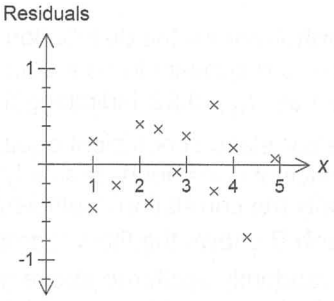
Situation1



Graph A scatterplot



Graph B regression line



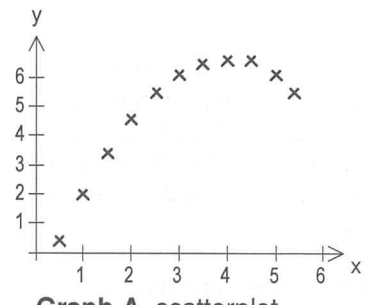
Graph C residual plot

Graph A shows the distribution of bivariate data and the relationship between the variables x and y appears to be a strong negative linear relationship. The correlation coefficient $r_{xy} = -0.94$ supports the scatterplot.

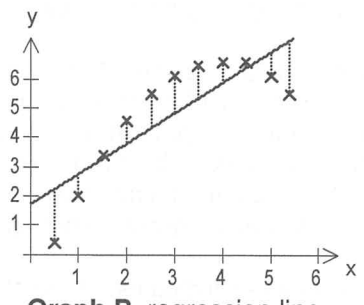
Graph B shows the fitted regression line $\hat{y} = -0.88x + 5.34$ and the residuals, showing that the residuals are randomly scattered and there is no discernible pattern. Hence the linear regression model suits the given data and is appropriate model for regression (making predictions).

Graph C is a residual plot with the x-axis the same as that for the scatterplot and the y-axis showing the possible magnitudes of the residuals. The residual plot gives the same information as graph B but it is much easier to see the random distribution of the residual values.

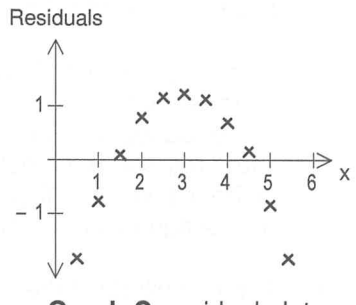
Situation 2



Graph A scatterplot



Graph B regression line



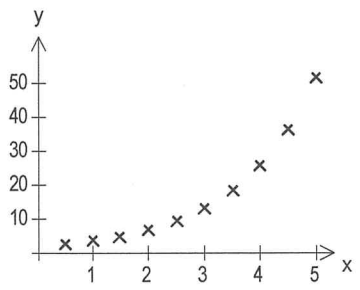
Graph C residual plot

Graph A shows the distribution of bivariate data and the relationship between the variables x and y is not linear and appears to be a strong quadratic relationship. The correlation coefficient for this scatterplot is given as $r_{xy} = 0.83$ indicating that there is a strong positive relationship between x and y.

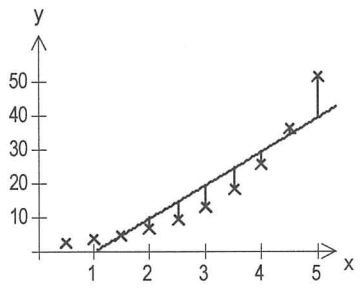
The correlation coefficient does not support the scatterplot and it can be clearly seen that a high correlation coefficient does not necessarily imply that the relationship under consideration is linear. Thus consideration of only the correlation coefficient is not sufficient to inform us about the suitability of linear regression.

Graph B shows the fitted regression line $\hat{y}=1.04x+1.72$ and the residuals, showing that the residuals are not randomly scattered above and below the curve and there is a discernible pattern. For low values of x the residuals are below the regression line that is negative, for middle values of x the residuals are above the regression line that is positive and for high values of x the residuals are negative. Hence using the linear regression model $\hat{y}=1.04x+1.72$ would not be appropriate for regression as the relationship between x and y is clearly not linear. Regression for this data set must be done using a different regression model as linear regression is not appropriate. Perhaps quadratic regression may be appropriate in this case.
Graph C the residual plot gives the same information as graph B but it is much easier to see the non-random nature of the distribution of the residual values and the magnitude of the residuals.

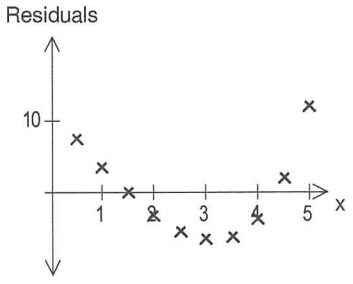
Situation 3



Graph A scatterplot



Graph B regression line



Graph C residual plot

Graph A shows the distribution of bivariate data and the relationship between the variables x and y is not linear and appears to be a strong exponential relationship. The correlation coefficient for this scatterplot is given as $r_{xy}=0.92$ indicating that there is a strong positive relationship between x and y.

The correlation coefficient does not support the scatterplot and it can be clearly seen that a high correlation coefficient does not necessarily imply that the relationship under consideration is linear. Thus consideration of only the correlation coefficient is not sufficient to inform us about the suitability of linear regression.
Graph B shows the fitted regression line $\hat{y}=9.89x-10.26$ and the residuals, showing that the residuals are not randomly scattered above and below the curve and there is a discernible pattern. For low values of x the residuals are positive, for middle values of x the residuals are negative and for high values of x the residuals are positive. Hence using the linear regression model $\hat{y}=9.89x-10.26$ would not be appropriate for regression as the relationship between x and y is clearly not linear. Regression for this data set must be done using a different regression model as linear regression is not appropriate, perhaps exponential regression may be appropriate in this case.
Graph C the residual plot gives the same information as graph B but it is much easier to see the non-random nature of the distribution of the residual values and the magnitude of the residuals.

In summing up:

- The correlation coefficient on its own is a very poor indicator in the validation of a linear regression model and on its own cannot be used to justify a linear relation between two variables. If used with an accompanying scatter diagram its use is strengthened to some degree but it is still not an acceptable justification as the relationship may not be linear.

- A residual plot will allow us to determine the validity of the linear regression model as a predictor. As residuals are like random errors the residual plot should be a random scatter of points in the plane, any evidence of a non-random distribution (or pattern) in the residual plot would indicate that our linear regression model is not the curve of best fit. If a pattern or a non-random distribution is evident in the residual plot then another regression model must be used for prediction purposes, for example, a least squares parabola, that is a quadratic regression model may be appropriate.

- The magnitude of the residuals gives us an indication of the strength of the linear relationship between the two variables. The smaller the magnitude of the residuals the higher the correlation between the two variables resulting in an increase in the reliability of any predictions that are made.

- Any outliers in the residual plot have to be checked out and if found to be real and not as a result of misreporting or other error then either the outlier or the linear regression model must be discarded.

- When making predictions the reliability of the prediction may be questionable if the number of data points being used is small.

Example 8

The Test 1 and Test 2 results of the Unit 3 Applications course for a class are shown below. Consider the suitability of a linear model for predicting a Test 2 score for a given Test 1 score.

Student	1	2	3	4	5	6	7	8	9	10	11	12
Test 1(x)	67	50	75	70	65	85	91	57	33	66	93	79
Test 2(y)	82	35	86	81	76	88	85	60	50	68	86	82

From the calculator we obtain $a = 0.8171986$

$b = 16.658999$

$r = 0.8397605$

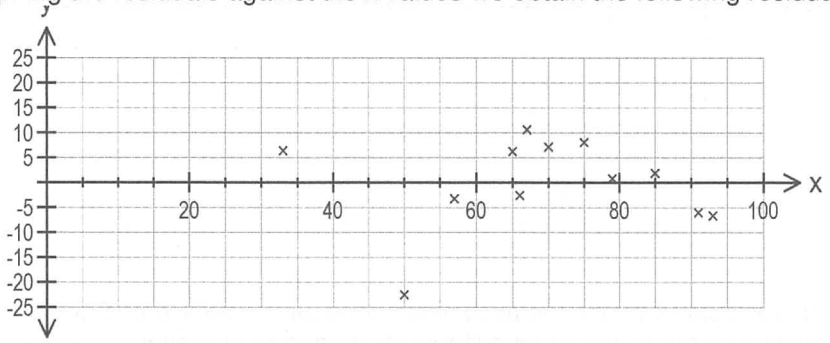
Hence the linear regression model for y on x is given by $\hat{y} = 0.8171986x + 16.658999$

To calculate the residuals we make use of the following table:

Student	1	2	3	4	5	6	7	8	9	10	11	12
Test 1(x)	67	50	75	70	65	85	91	57	33	66	93	79
Test 2(y)	82	35	86	81	76	88	85	60	50	68	86	82
$\hat{y}$	71.4	57.5	77.9	73.9	69.8	86.1	91.0	63.2	43.6	70.6	92.7	81.2
$y - \hat{y}$	10.6	-22.5	8.1	7.1	6.2	1.9	-6.0	-3.2	6.4	-2.6	-6.7	0.8

Where $\hat{y}$ are the predicted values given by $\hat{y} = 0.81719856x + 16.6589991$ and $y - \hat{y}$ are the residuals.

Graphing the residuals against the x values we obtain the following residual plot.

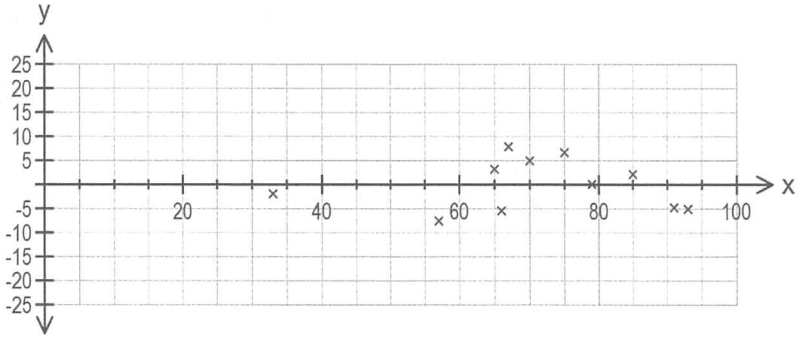


The residual plot does not suggest any definite pattern, thus the linear regression model given by $\hat{y} = 0.81719856x + 16.6589991$ appears to be suitable for this bivariate distribution.

However closer examination of the residuals suggests that the residual for the student numbered 2 with ordered pair (50, 35) is comparatively large and hence should be checked out.
In this case the point is real, that is no errors have been made, hence we would use this model for prediction of a test 2 score for a given test 1 score or consider a different regression model.

However if it was found that student numbered 2 had been absent for an extended period of time prior to test 2 due to illness then in this case there is justification for cropping the data set as there is evidence of a characteristic which is not present in the rest of the distribution.

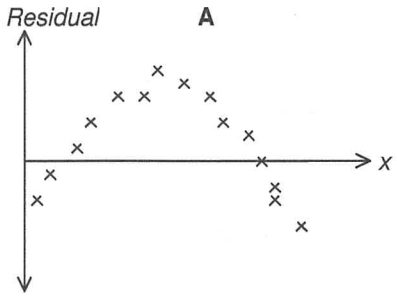
On removal of student 2 the revised linear regression model is given by $y = 0.6528839x + 30.372512$ and the resulting residual plot, shown below does not suggest any definite pattern indicating that the linear regression model is suitable for prediction purposes. Furthermore, with the removal of the outlier, student numbered 2, the value of the correlation coefficient has increased from 0.8397605 to 0.90069847 supporting the suitability of the linear model.

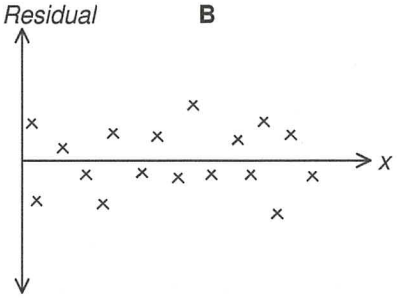


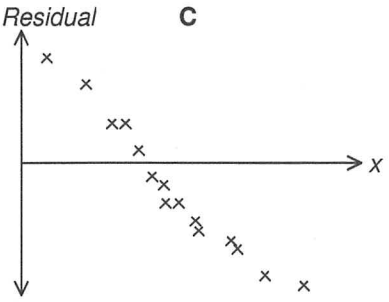
Note that the magnitude of the residuals in the cropped residual plot are much smaller than that of the original data set further supporting the appropriateness of the cropped linear model.

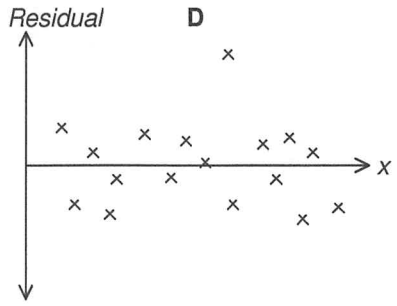
EXERCISE 3E

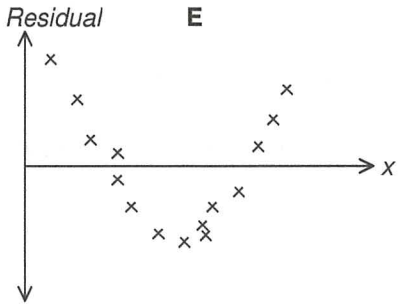
1. Consider the following residual plots and comment on the suitability of the linear regression model associated with the residual plot.

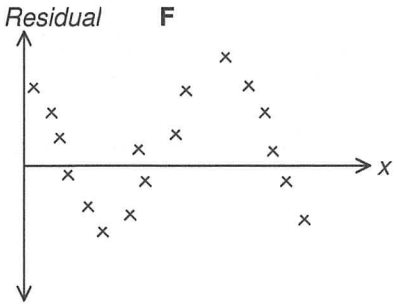




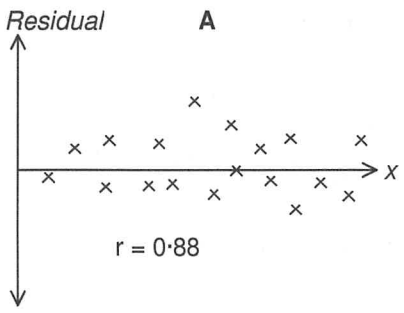


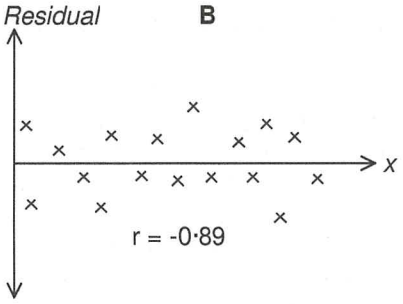


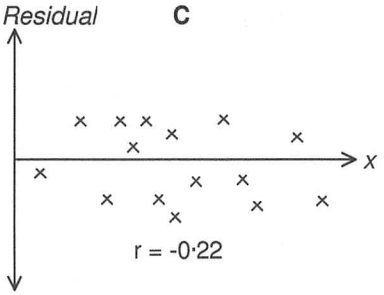


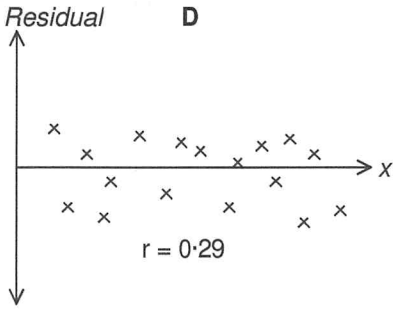


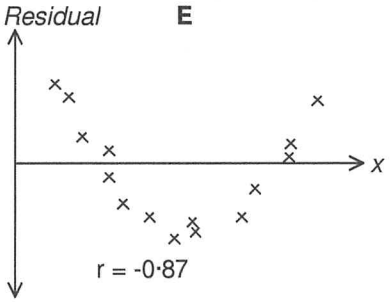
2. Consider the following residual plots and the associated correlation coefficient. For each case state whether the linear regression model would be suitable for prediction purposes.

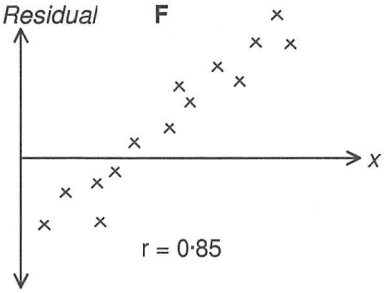












3. The first row of each data set shown below contains the explanatory variable. For each data set shown below:
   (i) Plot the scatter diagram using a graphics calculator and comment on the linear relationship between the variables.
   (ii) Determine the coefficient of correlation and comment on the relationship between the variables.
   (iii) Determine the equation of the line of best fit for predicting the response variable.
   (iv) Complete each table.
   (v) Draw a residual plot using a graphics calculator and comment on the suitability of using the linear regression model for prediction purposes.

(a)

x	1	2	3	4	5	6
y	3.3	5.7	7.5	9.9	11.2	13.8
$\hat{y}$						
$y-\hat{y}$						

(b)

m	20	30	50	70	90	100
n	90	200	410	670	980	1350
$\hat{n}$						
$n-\hat{n}$						

(c)

p	2	3.25	4.21	5	7.1	8	9.95	12	16
q	3.35	3.6	3.7	3.81	4	4.1	4.14	4.24	4.4
$\hat{q}$									
$q-\hat{q}$									

(d)

s	90	80	70	60	50	40	30	20	10
t	10	60	50	100	140	150	190	240	200
$\hat{t}$									
$t-\hat{t}$									

4. Consider the following bivariate distribution.

x	1	1	2	4	4	5	5	6	7	8	8	9	9
y	1	3	3	3	4	3	4	5	5	4	6	5	6

(a) Find the coefficient of correlation correct to four decimal places and comment on the relationship between the variables x and y.

(b) Determine the linear regression line of y on x.

(c) Examine the residual plot and comment on the suitability of the linear regression model.

(d) Predict a value for y when x = 10 and comment on the reliability of your prediction.

(e) Predict a value for y when x = 20 and comment on the reliability of your prediction.

5. Consider the following data.

x	1	2	3	7	12	18
y	9	7	6	5	3	1

(a) Examine the scatter plot of the given data set and comment on the relationship between x and y.

(b) Find r_{xy} and comment on the relationship between x and y.

(c) Find the least squares regression line of y on x.

(d) Construct a residual plot and comment on the suitability of using the least squares regression line of y on x for predicting the value of y for x = 9.

6. Consider the following data.

x	1	2	3	4	5	6	7	8	9	10	11	12
y	360	700	1100	1465	2000	2222	2809	3355	4167	4833	5532	6201

(a) Examine the scatter plot of the given data set and comment on the relationship between x and y.

(b) Find r_{xy} and comment on the linear relationship between x and y.

(c) Find the least squares regression line of y on x

(d) Construct a residual plot and comment on the suitability of using the least squares regression line of y on x for predicting the value of y for x = 4.6

7. Consider the following data.

x	3	4	5	6	7	8	9	10	11	12
y	150	62	32	16	8	4	2	1	0.5	0.2

(a) Examine the scatter plot of the given data set and comment on the relationship between x and y.

(b) Find r_{xy} and comment on the linear relationship between x and y.

(c) Find the least squares regression line of y on x

(d) Construct a residual plot and comment on the suitability of using the least squares regression line of y on x for predicting the value of y for x = 6.5

8. A class of music students sat for their Theory (t) and Practical (p) examinations, their results expressed as a percentage have been listed in the spreadsheet below.

	A	B	C	D	E	F	G	H
1	Student	Theory (t)	Practical (p)	$\hat{p}$	$p-\hat{p}$	$\hat{t}$	$t-\hat{t}$	
2	1	58	56					
3	2	61	63					
4	3	65	66					
5	4	69	70					
6	5	74	71					
7	6	77	75					
8	7	80	83					
9	8	87	80					
10	9	91	91					
11	10	93	87					
12								

(a) Using a graphics calculator or a suitable computer programme examine the scatter diagram of the above data and then comment on the relationship between the practical music marks and the theory music marks.

(b) Find a linear regression model for prediction of practical marks given a music theory mark.

On the spreadsheet $\hat{t}$ represents the predicted theory mark and $\hat{p}$ represents the predicted practical mark.

(c) Give the following spreadsheet entries for each of the following cells.

(i) Cell D2 (ii) Cell E2

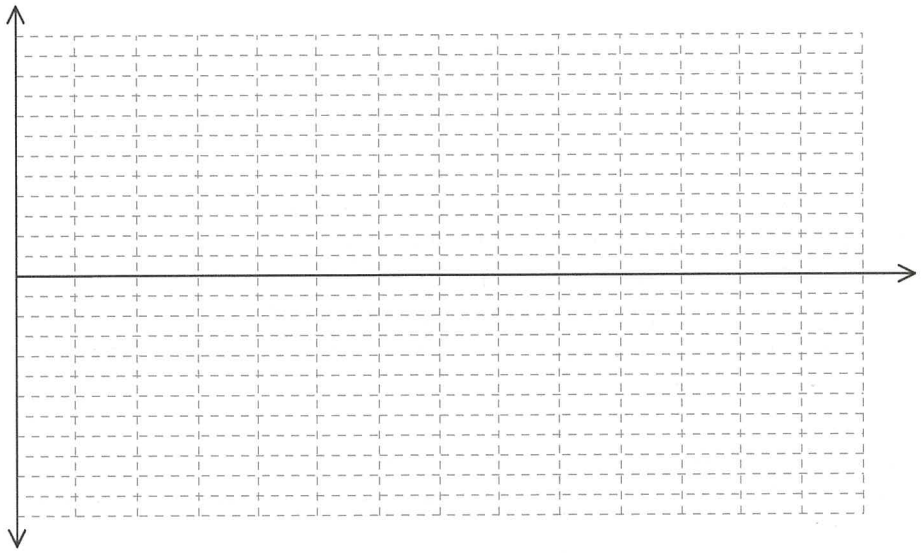
(iii) Cell F2 (ii) Cell G2

(d) Complete the spreadsheet by making entries in all of the empty cells.

(e) What information is contained in column E of the spreadsheet?

(f) What use can be made of the column E entries.

(g) Construct a residual plot on the given axes and comment on the suitability of using the least squares regression line found in (b) for making predictions.



(h) A student scored 90 marks in her theory examination but due to a hand injury she could not sit the practical examination. What practical mark should she be awarded on the basis of her performance in the theory examination?

(i) What theory mark should a student be awarded if the student received a practical mark of 87?

(j) Justify the suitability of your regression model for the prediction of the theory mark.

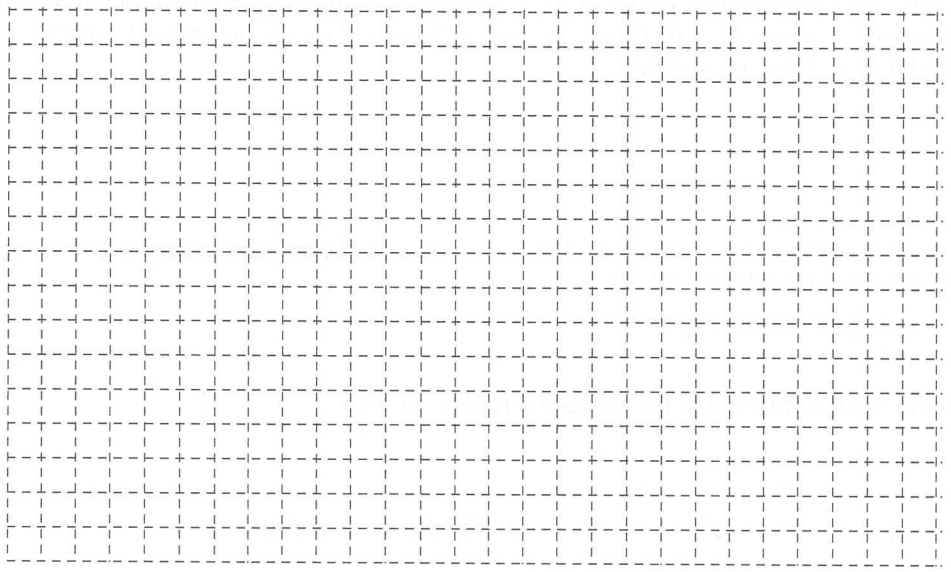
9. A study of teenagers revealed that the linear correlation coefficient between level of obesity and number of hours spent using a computer was 0.97.

(a) According to the result of this study what percentage of the variation in the level of obesity could be accounted for by the variation in the number of hours spent using a computer.

(b) Does this study suggest that by increasing the number of hours a teenager spends using a computer then the teenagers level of obesity will increase. Justify your response.

10. Hot food is put into a refrigerator. The refrigerator is kept at a constant temperature of 3°C. The temperature of the food is checked every 10 minutes and it is recorded to the nearest degree.

Time elapsed (t mins)	10	20	30	40	50	60	70	80	90	100
Temperature (°C)	48	38	30	4	18	14	9	5	4	3



(a) On the grid above draw a scatter plot representing the information in the table.
(b) Describe the relationship shown in the scatter plot.

(c) Calculate and interpret the value of the correlation coefficient between elapsed time and temperature.

From your graph is should be obvious that one data point does not fit the trend of the data.
(d) Identify this data point.

It was found that the data point in part (d) was a recording error and it should have been recorded as data point (40, 24).
(e) Without performing any calculations what will be the effect on the value of the correlation coefficient if the recording error is corrected?

(f) Correct the recording error and then calculate the value of the correlation coefficient.

(g) Find the equation of the least squares regression line for predicting the temperature of the food for elapsed time. Give the parameters of your equation rounded to 3 decimal places.

(h) Identify and interpret the y-intercept of your regression line in context.

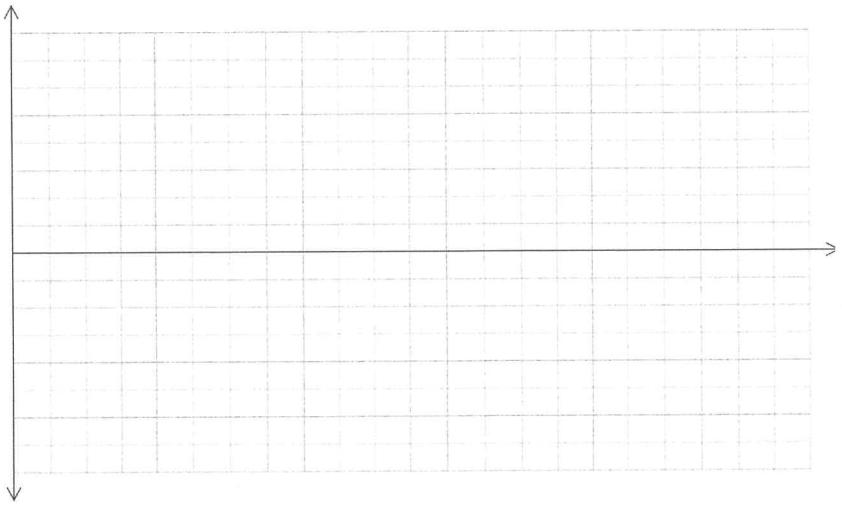
(i) Identify and interpret the gradient of your regression line in context.

(j) Calculate the value of the coefficient of determination and use it together with the correlation coefficient determined in part (f) to determine if the linear regression model in part (g) is appropriate for making predictions.

(k) Use your regression model to predict the temperature of the food, to one decimal place, after being in the refrigerator for 1 hour.

(l) Calculate the residual value for elapsed time of one hour.

(m) On the given axes construct a residual plot for the corrected data set.



(n) Does your residual plot support your answer to part (j) above? Justify your response.

11. The owner of a free-range egg farm noted that there appeared to be a relationship between the number of eggs laid and the average daily temperature. The table below gives the weekly average number of eggs laid (n) by each hen and average weekly temperature (t) in degrees Celsius over a ten-week period.

Week	1	2	3	4	5	6	7	8	9	10
n	5.9	6.6	5.7	6.1	7	7.5	8	7.6	8.1	7.9
t	19.4	19.5	19.3	19.5	19.6	19.8	19.9	19.9	19.8	19.8

(a) Identify the explanatory variable and the response variable.

(b) Find the value of r_{nt} . What does the value of r_{nt} indicate?

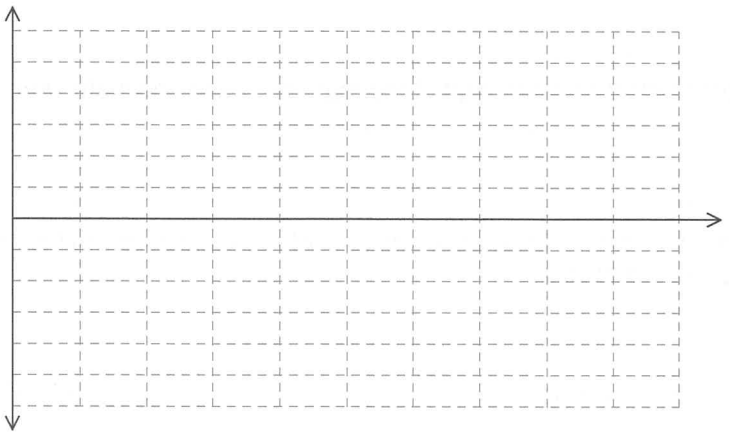
(c) Calculate the line of regression that can be used to predict the average laying rate from the average weekly temperature.

(d) State the gradient of the regression line and explain what it means.

(e) State the vertical axis intercept of the regression line and state what it means.

In order to check on the reliability of the regression model it is necessary to examine the residual plot for the model in question.

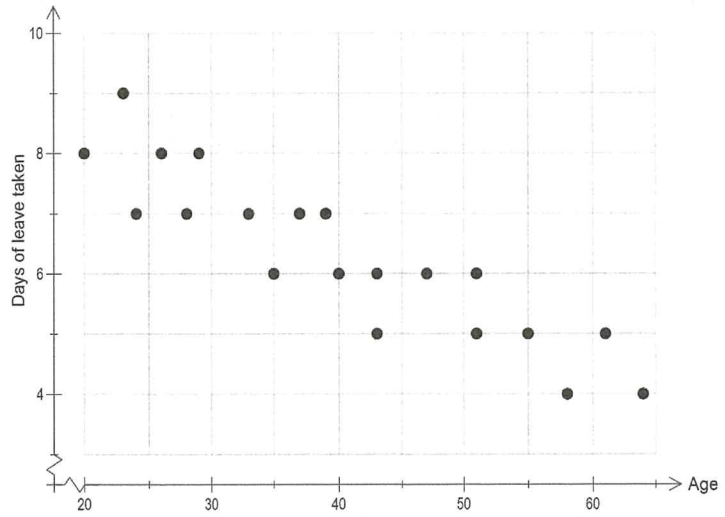
(f) On the grid below draw a residual plot for this model and comment on how it is used to validate the reliability of prediction.



(g) Using your prediction model determine the average number of eggs that can be expected if the average weekly temperature is 20°C.

(h) Discuss the reliability of your prediction.

12. In a study between the age, in years, of all employees in a small company, and the number of days sick leave taken last year by these employees resulted in the graph below.



The correlation coefficient for this data is -0.9060.
The equation of the least squares regression line for predicting the number of days sick leave is:

Number of days leave = 10.093 – 0.094 x age

(a) Graph the regression line on the scatter plot above showing two key points used in drawing the line.

(b) State the gradient of the regression line and explain what it means.

(c) State the vertical axis intercept of the regression line and state what it means.

(d) Explain the relationship between the sign of the correlation coefficient and gradient of the least squares regression line.

(e) Calculate the coefficient of determination and state what it tells you.

(f) Predict how many days a fifty-five year old employee might be expected to be absent.

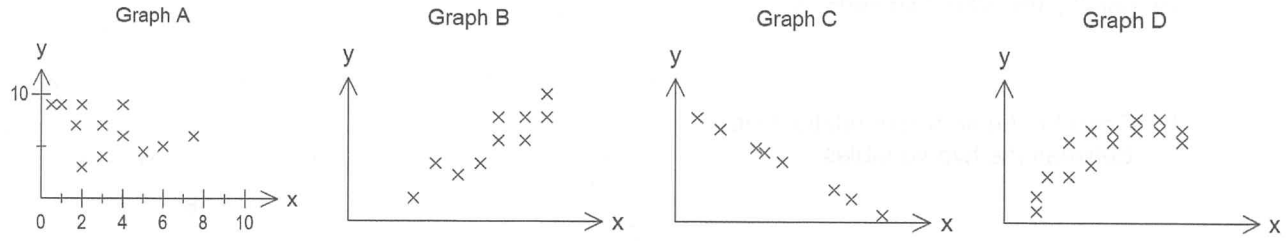
(g) Comment on the reliability of your prediction. Give reasons.

(h) Find the residual when the age of an employee is 55 years.

(i) Explain what your answer to part (f) means.

CHAPTER THREE REVIEW EXERCISE

1. Consider the following scatter diagrams:



Match each of the above scatterplots with the number next to the appropriate value of r^2 given below. Then calculate the value of r_{xy} in each case.

(i). $r^2 = 1$ (ii). $r^2 = 0.81$ (iii) $r^2 = 0.16$ (iv) $r^2 = 0.64$

Scatter plot	Matching number of r^2	Value of r_{xy}
A		
B		
C		
D		

2. The quantity of grapes sold, in kilograms, each day at Spudmart and the corresponding price in dollars per kilogram were recorded for eight days.

Day	1	2	3	4	5	6	7	8
Price ($)	3.15	4.17	2.96	5.42	5.30	5.80	4.45	3.57
Quantity (kg)	350	300	310	230	220	200	250	300

(a) Identify the response variable and the explanatory variable.

(b) Find the equation of least squares line for predicting the quantity of grapes sold based on the given data.

(c) Estimate the change in sales of grapes if the price is increased by 50 cents per kilogram.

(d) Estimate the change in sales of grapes if the price is lowered by 20 cents per kilogram.

(e) Calculate the correlation coefficient.

(f) Calculate the coefficient of determination.

(g) What percentage of the variation in the quantity sold each day is unexplained by the variation in the price?

(h) Estimate the quantity of grapes sold if the price is set at $4.50 per kilogram. Comment on the reliability of your estimate.

(k) Determine, to the nearest percent, the unexplained variation between the ages of these chickens and the average number of eggs laid per month.

5. As the average daily maximum temperature increases, hot chocolate drink sales decrease at a particular sidewalk café. 64% of the variation in hot chocolate drink sales can be attributed to the variation in the average daily maximum temperature. What is the value of the correlation coefficient between average daily maximum temperature and hot chocolate sales?

6. The coefficient of determination resulting from a regression analysis was found to be 0.81. What would be the value of the correlation coefficient?

7. The length, in cm and weight in grams of a particular lizard were measured and recorded as follows:

Length	32.4	32.1	27.4	34.1	31.5	33.8	28.3	29.8	29.3	26.2	30.7	27.1	28.5	33.0	32.6
Weight	390	344	292	421	364	381	269	335	308	271	321	266	322	385	360

(a) Find the equation of the least squares regression line that models the data and enables you to make a prediction for the weight of a lizard given its length. State coefficients rounded to three decimal places.

(b) Examination of the scatter plot (on your calculator or computer) shows that the data has an increasing trend in the value of the weight. What does the equation, as determined in part (a), indicate is the rate of increase in the weight.

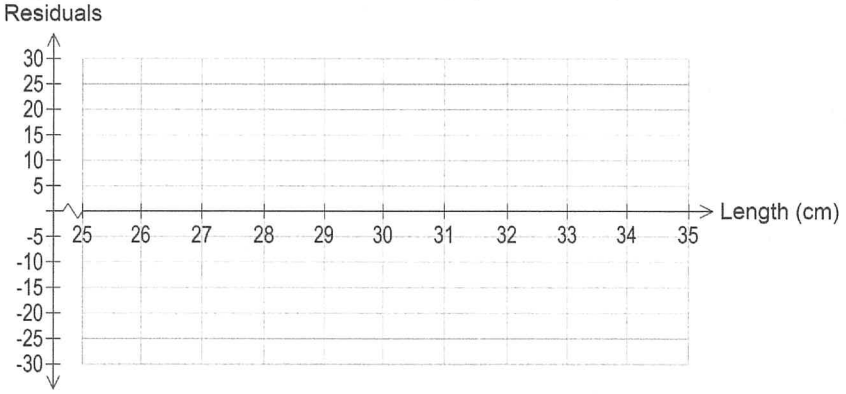
(c) State the values of the correlation coefficient and the coefficient of determination.

(d) Use the values of the correlation coefficient and the coefficient of determination to determine if the regression line found in part (a) is appropriate.

(e) Use your regression model to predict the weight of a lizard of length 31.5 cm. Give your answer correct to nearest gram.

(f) Calculate the residual of a lizard of length 31.5 cm. Round your answer to one decimal place.

(g) Use the given information to construct a residual plot on the axes below.



(h) Does the residual plot support your answer to part (d)? Justify your response.

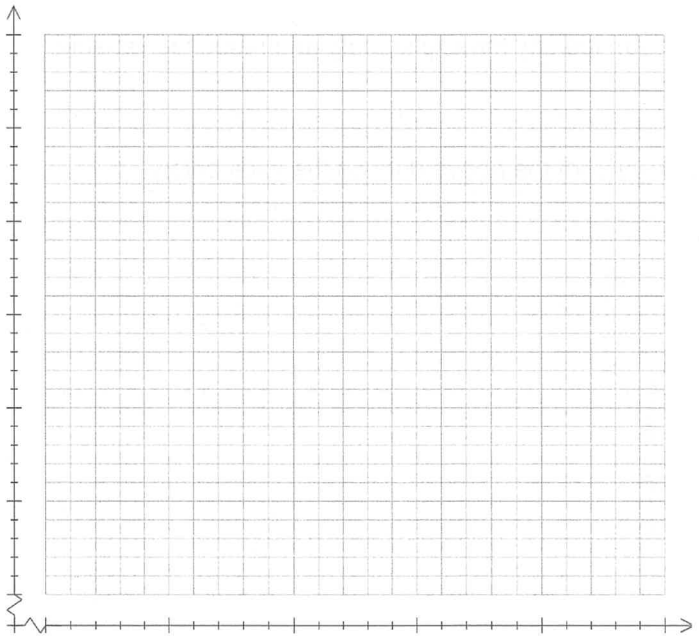
8. An ice-cream vending van owner keeps records of the profit of sales (P) for each month (rounded to the nearest $100) and the average monthly maximum temperature (T) rounded to the nearest degree Celsius for a year.

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Temperature (T)	35	33	28	26	22	21	18	17	21	24	27	29
Profit (P)	34	35	29	28	16	18	17	12	22	24	26	31
Residuals	-2.83	0.69	0.99	2.51	-4.49	-1.19						

(a) Identify the response and explanatory variables.

(b) Plot the above data on the axes right.

(c) Determine the equation of the least squares line that models the relationship between the average monthly maximum temperature and the profit of sales.



(d) Add the least squares line to your graph.

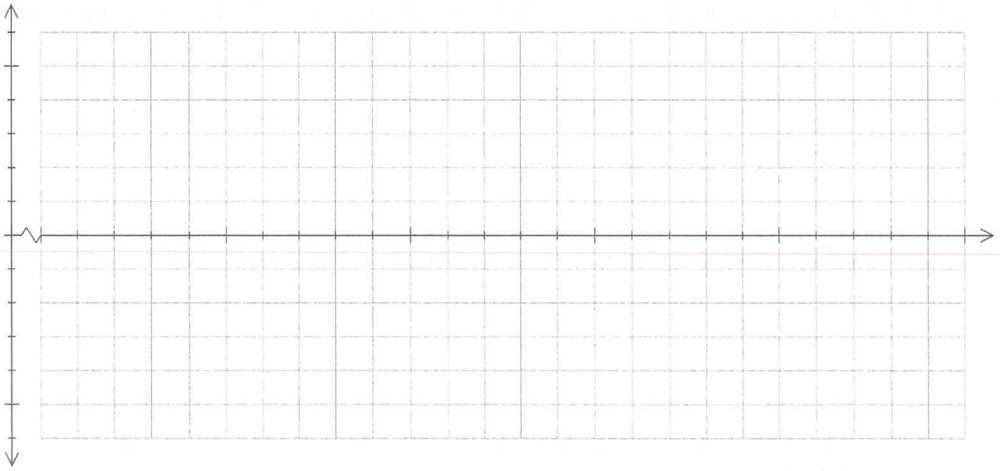
(e) Estimate the value of r the correlation coefficient and explain your estimate.

(f) Determine the correlation coefficient for the given data.

(g) What does the correlation coefficient tell you about the relationship between the variables?

(h) Complete the 4th row of the table above by listing the missing residuals for the response variable.

(i) On the axes below construct a residual plot for the response variable.



(j) Does the least squares line determined in part (c) provide a good model for the given data? Justify your answer using evidence from your residual plot.

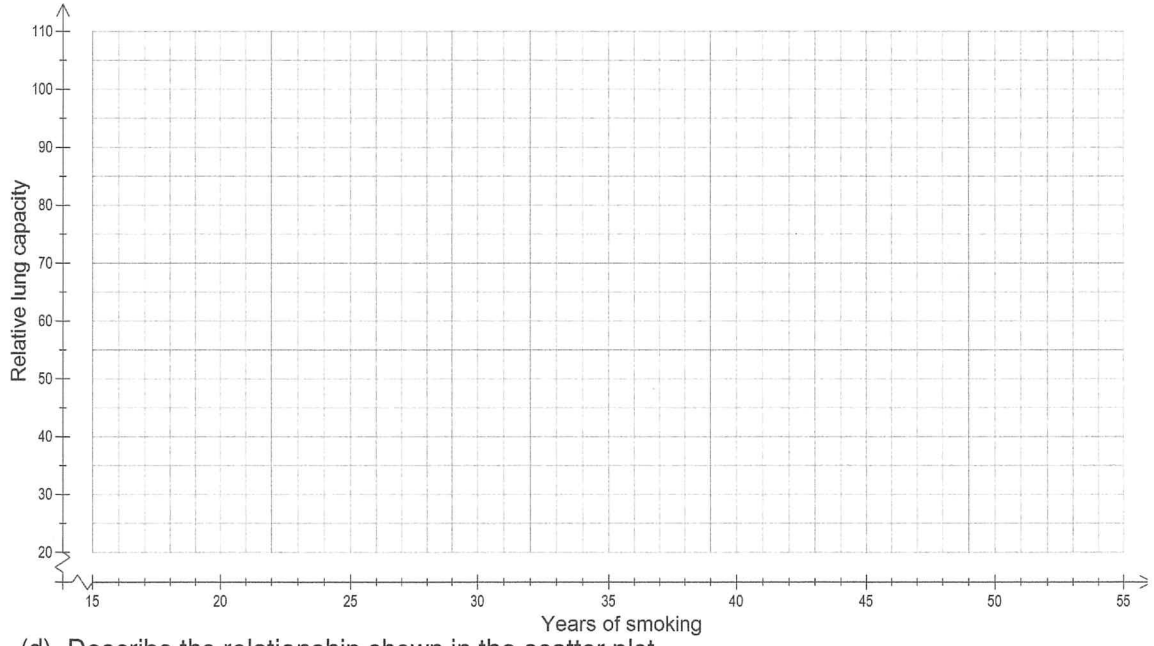
9. A random sample of 15 male patients suffering from emphysema were tested to determine their lung capacity (0 to 150 units), and how long they had been smoking. The test data has been tabled below.

Patient	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Years smoking (Y)	30	16	25	20	50	45	41	24	23	26	37	18	17	34	22
Relative lung capacity (C)	66	99	77	90	28	38	46	78	80	75	54	95	98	58	84

(a) What assumption must be made before any valid statistical investigation is carried out using the data.

(b) Identify the explanatory and response variable.

(c) Plot the above data on the axes below.



(d) Describe the relationship shown in the scatter plot.

(e) Determine the equation of the least squares line that models the relationship between years of smoking and relative lung capacity. Add the least squares line to you scatter plot.

(f) Identify and interpret the y-intercept of your least squares regression line in context.

(g) Identify and interpret the slope of your least squares regression line in context.

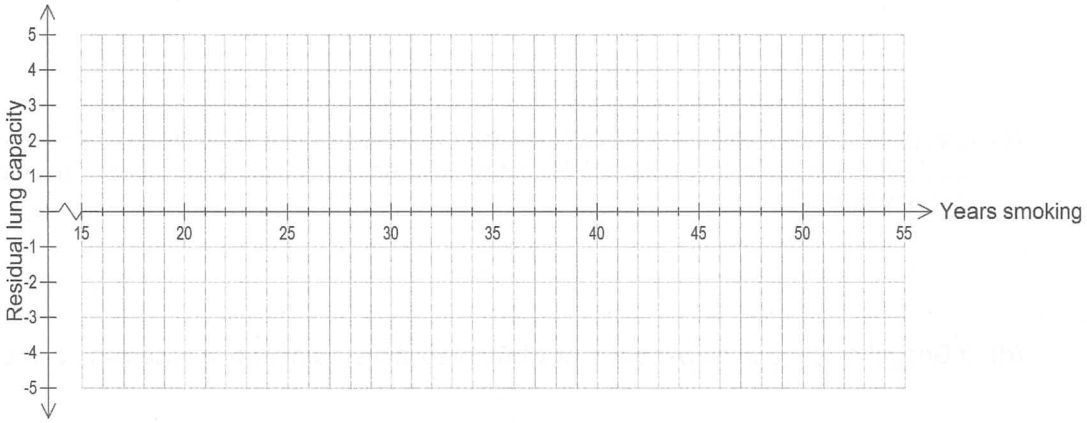
(h) Use your equation to determine the relative lung capacity for a male with emphysema who has been smoking 21 years.

(i) Determine the correlation coefficient for the sample of 15 patients, and use it to decide if there is evidence of linear relationship between relative lung capacity and years of smoking.

(j) Comment on the reliability of your prediction based on the correlation coefficient.

(k) Determine the coefficient of determination. Does the value of the coefficient of determination support your answer to (j)?

(l) On the axes below construct a residual plot and justify that the least squares line determined in part (e) provides a good model for the data.



10. An investigation concerning the relationship between life expectancy in years and birth rates greater than 30 births per one thousand people for a number of countries was found to be linear. The equation of the least squares regression line was determined and is given below.

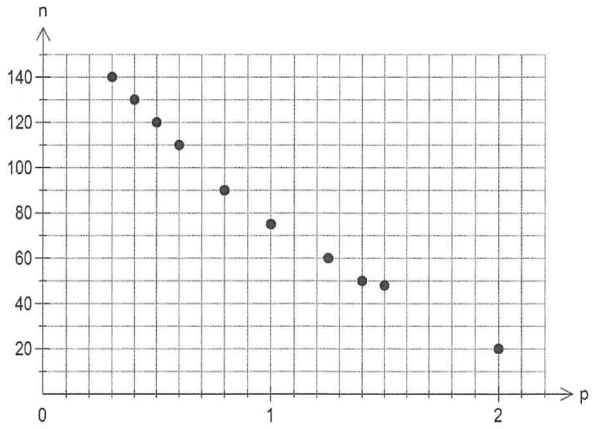
Life expectancy = 98.6 – 1.4 x birth rate.

(a) Identify and interpret the slope of this regression line.

(b) Identify and interpret the y intercept.

11. Bridget is raising money by selling home-made lemonade. She noticed that the number of glasses she sells (n) is related to the price (p) which she charges. She has summarised her results in the table below and the scatter plot of the data on the right.
Bridget has also calculated the equation of the trend line to be: $\hat{n} = 154.7118 - 69.6039p$.

Price ($p)	No. of glasses sold per day (n)	Residuals
0.30	140	7.40
0.40	130	4.56
0.50	120	1.71
0.60	110	-1.13
0.80	(i)	-6.82
1.00	(ii)	-7.51
1.25	60	-4.62
1.40	50	-3.89
1.50	48	(iii)
2.00	20	(iv)



(a) Identify the response and explanatory variable.

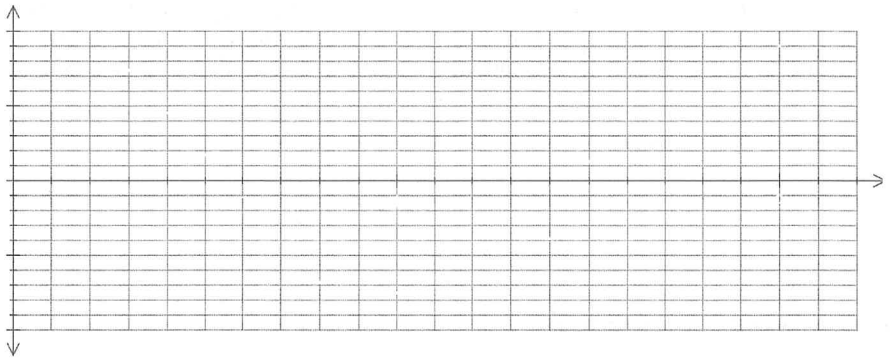
(b) Complete the table by finding entries labelled (i) to (iv).

(c) Calculate the correlation coefficient between these variables. Use the value of the correlation coefficient to comment on the relationship between price and the number of glasses of lemonade per day sold.

(d) If Bridget sets the price per glass at $1.80 predict the number of glasses she can expect to sell.

(e) Comment on the reliability of your prediction in part (d).

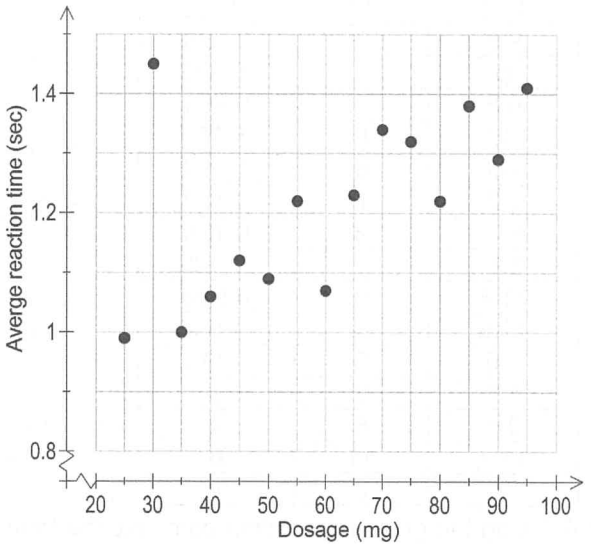
(f) Using the grid below construct a residual plot for the data.



(h) Comment on the appropriateness of using a linear model for the data, making reference to the graph from (f). Hence comment on whether your estimate from (d) is too high or too low?

12. A drug manufacturer tested a new pain-killing drug to see if it had any effect on people's reaction times. If using the drug affected reaction times significantly it would change the drug formula. The following table shows the reaction times, in seconds, of groups of people given dosages of this drug in multiples of 5 milligrams.

Dosage (mg)	Average reaction time (sec)
25	0.99
30	1.45
35	1.00
40	1.06
45	1.12
50	1.09
55	1.22
60	1.07
65	1.23
70	1.34
75	1.32
80	1.22
85	1.38
90	1.29
95	1.41



(a) State the response variable.

(b) Using the given information describe the relationship shown between the drug dosage and average reaction times,

(c) Calculate the correlation coefficient. What does it tell us about the relationship between drug dosage and average reaction times?

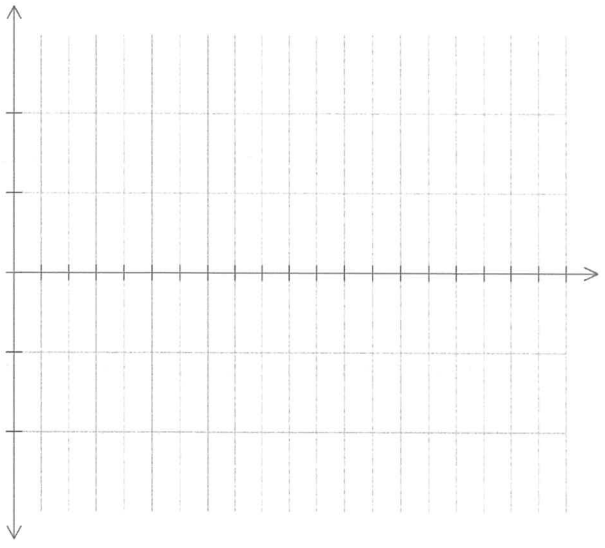
(d) State the equation of the regression line for the given data.

(e) Predict the reaction time if the drug dosage is 43 mg. Comment on the reliability of your prediction.

(f) Which of the data points most clearly appears to be an outlier?

(g) Remove from the data set the outlier you have identified in part (f). What effect does this have on the correlation coefficient?

(h) Use a suitable regression line based on the data after the removal of the outlier identified in part (f) to predict the average reaction times of people taking 43 mg of the drug.

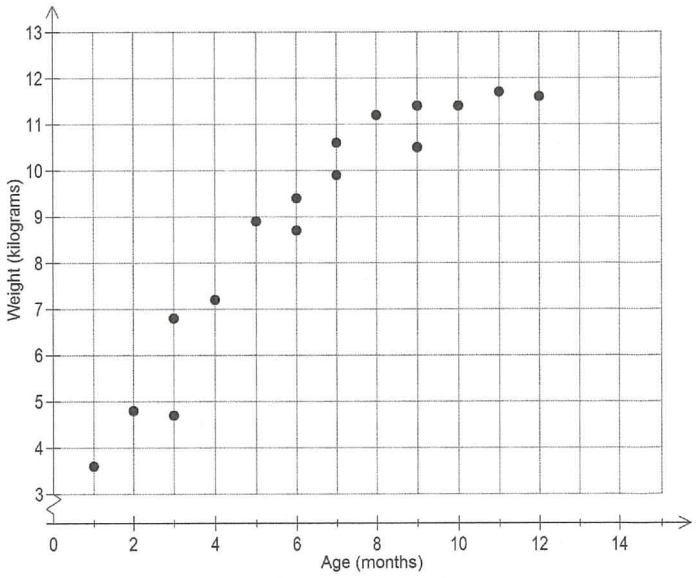


(i) On the given axes construct a residual plot based on the data after the removal of the outlier.

(j) Comment on the reliability of your prediction in part (h) making use of your residual plot.

13. The weights, in kilograms, and the ages, in months, of boys recorded at an Early Childhood Health Centre are recorded and graphed as shown below.

Weight (kg)	Age (months)	Residuals
11.4	10	-0.250
9.9	7	0.566
p	9	-0.378
6.8	3	0.553
11.2	8	1.094
10.6	**q**	1.266
7.2	**r**	0.182
8.7	6	0.138
3.6	1	-1.103
11.4	9	0.522
11.6	12	-1.594
8.9	5	1.110
4.7	3	**s**
9.4	6	**t**
11.7	11	**u**
4.8	2	**v**



(a) Using the given information complete the table by finding the missing entries p, q and r.

(b) Identify the response variable.

(c) Calculate the correlation coefficient and interpret this value.

(d) Calculate the coefficient of determination.

(e) Interpret the coefficient of determination.

(f) State the equation for the least squares line that models these data.

(g) Identify and interpret the slope of the least squares regression line in this context.

(h) Identify and interpret the vertical axis intercept in this context.

(i) Predict the weight of a boy who is 3 months old.

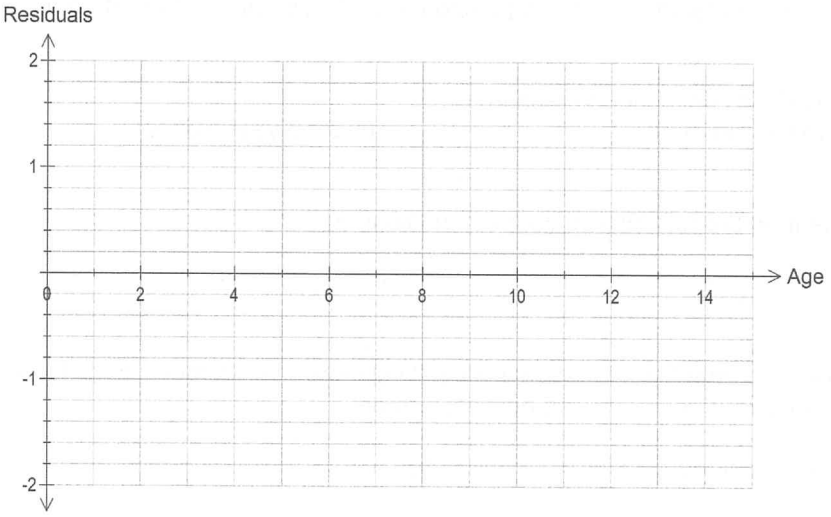
(j) Comment on the reliability of your prediction in part (i). Give a supporting reason.

(k) Calculate the residual **s** for the boy who is 3 months old giving your answer rounded to 3 decimal places.

(l) Explain what your answer to part (i) means.

(m) Complete the residual column of the table by finding values for t u and v.

(n) Graph the residuals on the axes below.



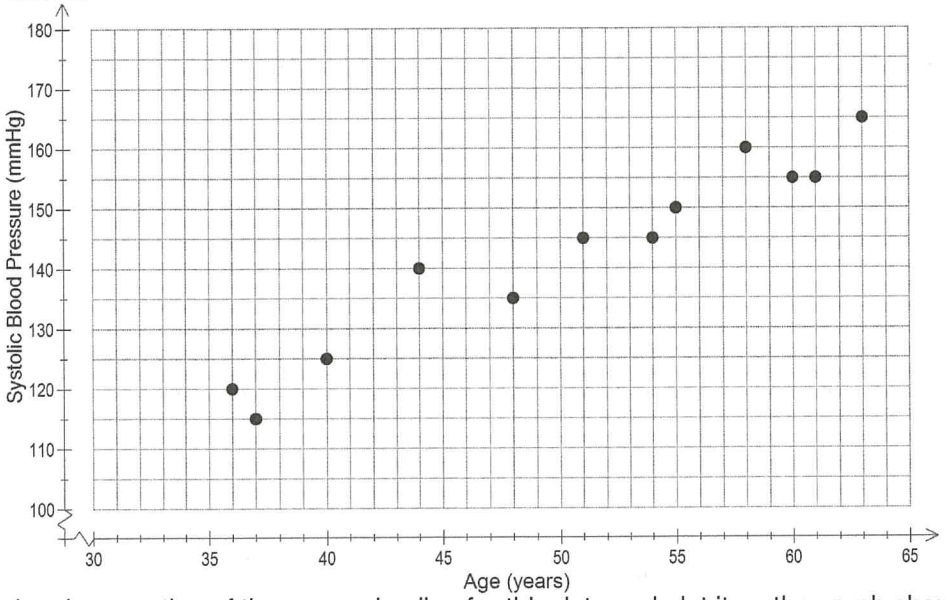
(o) Comment on the graph in terms of making predictions from the linear model in part (f). Give reasons.

(p) A researcher states: 'The high value of r proves that a baby's weight is caused by its age." Comment on this statement giving reasons.

14. The table shows the ages in years and systolic blood pressure in mmHg of 15 men.

Age	37	61	58	36	63	48	60	51	54	44	40	55	**35**	**45**	**57**
Blood Pressure	115	155	160	120	165	135	155	145	145	140	125	150	**125**	**135**	**145**

(a) The scatterplot below is incomplete. Complete the scatterplot by plotting the bold data values tabled above.



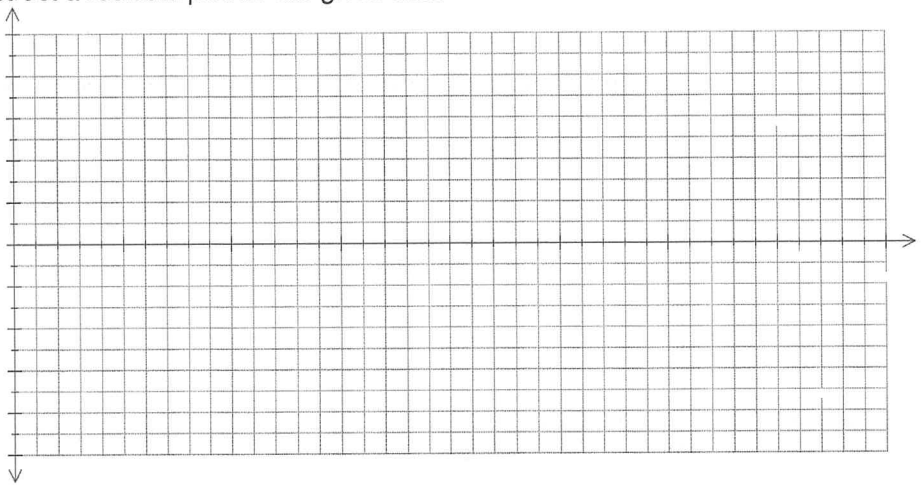
(b) Determine the equation of the regression line for this data and plot it on the graph above.

(c) Estimate the systolic blood pressure of:
(i) a 50 year old man. (ii) a 70 year old man.

(d) Comment on the reliability of your estimates in (c).

(e) What is the percentage variation is systolic blood pressure that is explained by the variation in age? Give your answer to the nearest percent.

(f) Construct a residual plot for the given data.



(g) What purpose/test does the residual plot serve?

(h) Apply the test to the given data.