Introduction to regression

3A Fitting a straight line by eye
3B Fitting a straight line — the 3-median method
3C Fitting a straight line — least-squares regression
3D Interpretation, interpolation and extrapolation
3E Residual analysis
3F Transforming to linearity

**Areas of Study**

- Independent and dependent variables, fitting lines to bivariate numerical data, by eye, the 3-median line (as a graphical technique) and the least-squares method, interpretation of slope and intercepts, and use of lines to make predictions; extrapolation and interpolation; residual analysis to check quality of fit (residual value defined as actual value – predicted value)
- Estimation of the equation of an appropriate line of best fit from a scatterplot, use of the formulas $b = r \frac{S_y}{S_x}$ and $a = \bar{y} - b \bar{x}$; and use of technology with bivariate statistics to determine the coefficients of the corresponding equation, $y = a + bx$, of the least-squares regression line
- Transformation of some forms of non-linear data to linearity by transforming one of the axes scales using a square, log or reciprocal transformation.

**Fitting straight lines to bivariate data**

The process of ‘fitting’ straight lines to bivariate data enables us to analyse relationships between the data and possibly make predictions based on the given data set.

We will consider the three most common techniques for fitting a straight line and determining its equation, namely:

1. Line fit by eye
2. 3-median method
3. Least squares.

The linear relationship expressed as an equation is often referred to as the linear regression equation or line.

Recall from the previous chapter that when we display bivariate data as a scatterplot, the independent variable is placed on the horizontal axis and the dependent variable is placed on the vertical axis. When the relationship between two variables ($x$ and $y$) is described in equation form, such as $y = mx + c$, the subject, $y$, is the dependent variable and $x$ is the independent variable.
Fitting a straight line by eye

Consider the set of bivariate data points shown at right. In this case the $x$-values could be heights of married women, while $y$-values could be the heights of their husbands. We wish to determine a linear relationship between these two random variables.

Of course, there is no single straight line which would go through all the points, so we can only estimate such a line.

Furthermore, the more closely the points appear to be on or near a straight line, the more confident we are that such a linear relationship may exist and the more accurate our fitted line should be.

Consider the estimate, drawn ‘by eye’ in the figure below right. It is clear that most of the points are on or very close to this straight line. This line was easily drawn since the points are very much part of an apparent linear relationship.

However, note that some points are below the line and some are above it. Furthermore, if $x$ is the height of wives and $y$ is the height of husbands, it seems that husbands are generally taller than their wives.

Regression analysis is concerned with finding these straight lines using various methods so that the number of points above and below the lines are ‘balanced’.
Method of fitting lines by eye

There should be an equal number of points above and below the line. For example, if there are 12 points in the data set, 6 should be above the line and 6 below it. This may appear logical or even obvious, but fitting by eye involves a considerable margin of error.

**WORKED EXAMPLE 1**

Fit a straight line to the data in the figure at right using the equal-number-of-points method.

**THINK**

1. Note that the number of points (n) is 8.

2. Fit a line where 4 points are above the line. Using a clear plastic ruler, try to fit the best line.

3. The first attempt has only 3 points below the line where there should be 4. Make refinements.

4. The second attempt is an improvement, but the line is too close to the points above it. Improve the position of the line until a better ‘balance’ between upper and lower points is achieved.

**REMEMBER**

To fit a straight line by eye, when using bivariate data, make sure there are an equal number of points above and below the fitted line.

**EXERCISE**

3A Fitting a straight line by eye

The questions in this exercise represent data collected by groups of students conducting different environmental projects. The students have to fit a straight line to their data sets. **Note:** For many of these questions your answers may differ somewhat from those in the back of the book. The answers are provided as a guide but there are likely to be individual differences when fitting straight lines by eye.
1. **Fit a straight line to the data in the scatterplots using the equal-number-of-points method.**

   - **Diagram a**
   - **Diagram b**
   - **Diagram c**
   - **Diagram d**
   - **Diagram e**
   - **Diagram f**
   - **Diagram g**
   - **Diagram h**
   - **Diagram i**

2. **For the following scatterplots, fit a line of best fit by eye and determine the equation of the line.**

   - **Diagram a**
   - **Diagram b**

**Exam Tip:**
To draw a straight line, students are expected to bring a straight edge (for example, a ruler) into the examination. [Assessment report 2 2007]
Fitting a straight line — the 3-median method

Fitting lines by eye is useful but it is not the most accurate of methods. Greater accuracy is achieved through closer analysis of the data. Upon closer analysis, it is possible to find the equation of a line of best fit of the form \( y = mx + c \) where \( m \) is the gradient and \( c \) is the \( y \)-intercept. Several mathematical methods provide a line with a more accurate fit.

One of these methods is called the 3-median method. It involves the division of the data set into 3 groups and the use of the 3 medians in these groups to determine a line of best fit.

It is used when data show a linear relationship. It can even be used when the data contain outliers.

The 3-median method is best described as a step-by-step method.

Step 1. Plot the points on a scatterplot. This is shown in figure 1.

Step 2. Divide the points into 3 groups using vertical divisions (see figure 2). The number of points in a data set will not always be exactly divisible by 3. Thus, there will be three alternatives, as follows.

(a) If the number of points is divisible by 3, divide them into 3 equal groups, for example, 3, 3, 3 or 7, 7, 7.

(b) If there is 1 extra point, put the extra point in the middle group, for example, 3, 4, 3 or 7, 8, 7.

(c) If there are 2 extra points, put 1 extra point in each of the outer groups, for example, 4, 3, 4 or 8, 7, 8.

Step 3. Find the median point of each of the 3 groups and mark each median on the scatterplot (see figure 3). Recall that the median is the ‘middle’ value. So, the median point of each group has an \( x \)-coordinate which is the median of the \( x \)-values in the group and a \( y \)-coordinate which is the median of the \( y \)-values in the group.

(a) The left group is the lower group and its median is denoted by \((x_L, y_L)\).

(b) The median of the middle group is denoted by \((x_M, y_M)\).

(c) The right group is the upper group and its median is denoted by \((x_U, y_U)\).

Note: Although the \( x \)-values are already in ascending order on the scatterplot, the \( y \)-values within each group may need re-ordering before you can find the median.

To complete steps 4 and 5, three different approaches may be taken from here: graphical, arithmetic or you can use a CAS calculator.
**Graphical approach**

The graphical approach is fast and, therefore, usually the preferred method (see figure 4).

**Step 4.** Draw in the line of best fit. Place your ruler so that it passes through the lower and upper medians. Move the ruler a third of the way toward the middle group median while maintaining the slope. Hold the ruler there and draw the line.

**Step 5.** Find the equation of the line (general form $y = mx + c$). There are two general methods.

(a) Method A: Choose two points which lie on the line. Use their coordinates to find the gradient of the line and then the equation of the line.

(b) Method B: If the scale on the axes begins at zero, you can read off the $y$-intercept of the line and calculate its gradient. This will enable you to find the equation of the line.

**Arithmetic approach**

Using the arithmetic approach, you will proceed as follows.

**Step 4.** Calculate the gradient ($m$) of the line. Use the rule: $m = \frac{y_U - y_L}{x_U - x_L}$

**Step 5.** Calculate the $y$-intercept ($c$) of the line.

Use the rule: $c = \frac{1}{3}[(y_L + y_M + y_U) - m(x_L + x_M + x_U)]$

Thus, the equation of the regression line is $y = mx + c$.

**Using a CAS calculator**

Most CAS calculators have an inbuilt function, such as Median–Median, for fitting a line using the 3-median method. This function will be presented in the following worked example and can be used in most of the exercise questions.

**WORKED EXAMPLE 2**

Find the equation of the regression line for the data in the table at right using the 3-median method. Give coefficients correct to 2 decimal places.

<table>
<thead>
<tr>
<th>$x$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>6</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>

**THINK**

1. Plot the points on a scatterplot and divide the data into 3 groups. Note there are 6 points, so the division will be 2, 2, 2.

2. Find the median point of each group. Since each group has only 2 points, medians are found by averaging them.

$(x_L, y_L) = (1.5, 2)$

$(x_M, y_M) = (3.5, 4)$

$(x_U, y_U) = (6, 5.5)$
We now have the option of following either the graphical approach or the arithmetic approach. If using a CAS calculator, the previous 2 steps can be skipped.

**Method 1: Using the graphical approach**

3 Mark in the medians and place a ruler on the outer 2 medians. Maintaining the same slope on the ruler, move it one-third of the way towards the middle median. Draw the line.

4 Read off the $y$-intercept from the graph. $y$-intercept = 1

5 Choose 2 points to calculate the gradient ($m$). Select (0, 1) and (5, 5) as they are on a line parallel to the final solid line.

6 Write the equation of the 3-median regression line.

$$y = 0.8x + 1$$

**Method 2: Using the arithmetic approach**

3 Find the gradient using the formula, and the upper and lower medians found previously.

$$m = \frac{y_U - y_L}{x_U - x_L} = \frac{5.5 - 2}{6 - 1.5} = \frac{3.5}{4.5} = \frac{7}{9} \approx 0.78$$

4 Find the $y$-intercept by using the coordinates of all 3 medians in the formula.

$$c = \frac{1}{3}[(y_L + y_M + y_U) - m(x_L + x_M + x_U)] = \frac{1}{3}[2 + 4 + 5.5 - \frac{7}{9}(1.5 + 3.5 + 6)] = \frac{1}{3}[11.5 - \frac{7}{9}(11)] = \frac{1}{3}[11.5 - 8.555] = 0.98$$

5 State the regression equation.

$$y = 0.78x + 0.98$$

*Note:* There are slight variations in the values of the gradient and the $y$-intercept of the line between the graphical and the arithmetic approaches. This is because the arithmetic method gives more precise values for the gradient and the $y$-intercept, whereas the graphical method gives approximate values.
Method 3: Using a CAS calculator

1. On the Statistics screen, label list1 as ‘x’ and list2 as ‘y’. Enter the data and then tap:
   - SetGraph
   - Settings
     Set:
     Type: Scatter  XList: main\x
     YList: main\y  Freq: 1
     Mark: square
   Tap:
   - Set
   - Calc

2. The graph should appear as shown. To fit the regression line, tap:
   - Calc
   - MedMed Line
     Set:
     XList: main\x
     YList: main\y
     Freq: 1
     Copy Formula: y1
   Then tap OK.

   The values for $a$ and $b$ should appear as shown. These are the values for the equation of the regression line, in the form $y = ax + b$. 
3 To fit the regression line, tap OK.

4 Write your answer correct to 2 decimal places.

The equation of the 3-median regression line is $y = 0.78x + 0.98$.

**REMEMBER**

1. Assuming data points are in order of increasing $x$-values:
   - Step 1. Divide the data points into 3 groups
   - Step 2. Adjust for ‘unequal’ groups; if there is 1 extra point, put it in the middle; if there are 2 extra points, put one in each of the outer groups.
   - Step 3. Calculate the medians for the 3 groups $(x_1, y_1), (x_M, y_M), (x_U, y_U)$.

2. For a graphical approach:
   - Step 4. Place a ruler through the two ‘outer’ medians and move the ruler one-third of the way towards the middle median.
   - Step 5. Calculate the $y$-intercept and the gradient and use these to find the equation of the regression line.

3. For an arithmetic approach:
   - Step 4. Calculate the gradient using the formula:
     \[ m = \frac{y_U - y_L}{x_U - x_L} \]
   - Step 5. Calculate the $y$-intercept using the formula:
     \[ c = \frac{1}{3}[(y_L + y_M + y_U) - m(x_L + x_M + x_U)] \]

**EXERCISE 3B**

Fitting a straight line — the 3-median method

1. Find the regression line for the data in the table below using the 3-median method.

<table>
<thead>
<tr>
<th>$x$</th>
<th>2</th>
<th>3</th>
<th>3</th>
<th>4</th>
<th>4</th>
<th>5</th>
<th>5</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>3</td>
<td>6</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
</tbody>
</table>

2. Copy and complete the following table for the division of data points into three groups in the 3-median regression line method. The first row of the table has been completed for you.
### Data Analysis

<table>
<thead>
<tr>
<th>Total number of points (n)</th>
<th>Lower group</th>
<th>Middle group</th>
<th>Upper group</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>3</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>43</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>58</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>698</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3. Using the data in the table below, find the regression line using the 3-median method on your CAS calculator.

<table>
<thead>
<tr>
<th>$x$</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>55</th>
<th>60</th>
<th>65</th>
<th>75</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>80</td>
<td>60</td>
<td>50</td>
<td>70</td>
<td>40</td>
<td>55</td>
<td>40</td>
<td>30</td>
<td>10</td>
<td>25</td>
<td>15</td>
</tr>
</tbody>
</table>

Questions 4 and 5 refer to the data in the table below.

<table>
<thead>
<tr>
<th>Day</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temperature ($^\circ$C)</td>
<td>14</td>
<td>13</td>
<td>15</td>
<td>17</td>
<td>16</td>
<td>18</td>
<td>19</td>
<td>17</td>
<td>22</td>
<td>20</td>
<td>21</td>
<td>24</td>
</tr>
</tbody>
</table>

4. **MC** The gradient of the 3-median regression line for the above data set is:
   - A 0.56
   - B 0.75
   - C 1
   - D 0.88
   - E 0.5

5. **MC** The $y$-intercept of the 3-median regression line for the data set above is:
   - A 12.00
   - B 12.15
   - C 17.83
   - D 23.52
   - E 36.44

6. The sales figures (in thousands) for a company over a 10-month period were recorded as follows.

<table>
<thead>
<tr>
<th>Month ($x$)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sales ($y$)</td>
<td>85</td>
<td>77</td>
<td>81</td>
<td>73</td>
<td>68</td>
<td>72</td>
<td>64</td>
<td>57</td>
<td>53</td>
<td>49</td>
</tr>
</tbody>
</table>

Find the equation of the 3-median regression line.

7. The unemployment rate (as a percentage) is measured in 9 towns. The data are summarised in the table below. Note that town size is measured in thousands.

<table>
<thead>
<tr>
<th>Town size ($\times 1000$)</th>
<th>13</th>
<th>34</th>
<th>67</th>
<th>90</th>
<th>102</th>
<th>146</th>
<th>167</th>
<th>189</th>
<th>203</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unemployed (%)</td>
<td>12.3</td>
<td>11.6</td>
<td>9.4</td>
<td>9.6</td>
<td>8.1</td>
<td>8.2</td>
<td>6.2</td>
<td>5.4</td>
<td>4.5</td>
</tr>
</tbody>
</table>

**a** State the independent variable.

**b** Use your calculator to draw a scatterplot and show the 3-median regression line.
c State the regression lines in terms of the variable names *unemployed* and *town size*.
d Predict the unemployment rate for a town size of 300 000.

8 During an experiment, a research worker gathers the following data set:

<table>
<thead>
<tr>
<th>x</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>3</td>
<td>5</td>
<td>6</td>
<td>9</td>
<td>11</td>
<td>16</td>
<td>15</td>
<td>13</td>
<td>19</td>
<td>22</td>
<td>26</td>
<td>24</td>
<td>28</td>
<td>31</td>
<td>30</td>
<td>32</td>
<td>29</td>
<td>39</td>
<td>40</td>
<td>44</td>
<td></td>
</tr>
</tbody>
</table>

a Plot the data as a scatterplot.
b Find the equation of the 3-median regression line from the graph.

9 The graph at right shows the daily water level in a reservoir during a drought. From the graph (you may use the formulas or your calculator to check your answers):
a find the coordinates of the points used to find the gradient. Use these to find the gradient.
b find the coordinates of the median of the middle group
c estimate the y-intercept (use the graph and medians)
d state the relationship between water level and day as an equation.

10 Since management instituted new policies, the productivity at DMH car plant has been improving. The scatterplot at right shows the number of cars produced each week over a 10-week period.
a What are the coordinates of the points used to find the gradient? Use them to find the gradient.
b What are the coordinates of the median of the middle group?
c Using the graph and medians found, estimate the y-intercept.
d State the relationship between cars produced and week as an equation.
e Check your answers using a CAS calculator.

11 The graph at right represents the height of Louise, measured each year. Which graph best shows the line of best fit using the 3-median method?
When using the 3-median method for fitting a straight line, which of the following statements is false?

A The straight line is not affected by outliers.
B The two outside medians are used for the gradient of the line.
C For the y-intercept move the line one-third of the way towards the middle median.
D The gradient changes when moving the line towards the middle median.
E The number of points in each group must be balanced.

Fitting a straight line — least-squares regression

Another method for finding the equation of a straight line which is fitted to data is known as the method of least-squares regression. It is used when data show a linear relationship and have no obvious outliers.

To understand the underlying theory behind least-squares, consider the regression line shown below.

We wish to minimise the total of the vertical lines, or ‘errors’ in some way. For example, balancing the errors above and below the line. This is reasonable, but for sophisticated mathematical reasons it is preferable to minimise the sum of the squares of each of these errors. This is the essential mathematics of least-squares regression.

The calculation of the equation of a least-squares regression line is simple using a CAS calculator.
A study shows the more calls a teenager makes on their mobile phone, the less time they spend on each call. Find the equation of the linear regression line for the number of calls made plotted against call time in minutes using the least-squares method on a CAS calculator. Express coefficients correct to 2 decimal places.

<table>
<thead>
<tr>
<th>Number of minutes (x)</th>
<th>1</th>
<th>3</th>
<th>4</th>
<th>7</th>
<th>10</th>
<th>12</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of calls (y)</td>
<td>11</td>
<td>9</td>
<td>10</td>
<td>6</td>
<td>8</td>
<td>4</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

**Worked Example 3**

To draw a scatterplot, on the Statistics screen, label list1 as ‘minutes’ and list2 as ‘calls’. Enter the data appropriately, then follow step 1 of Method 3 from Worked example 2 to generate the scatterplot.

To find the linear regression model, tap:

- Calc
- Linear Reg

Set:

XList: main\minutes
YList: main\calls
Freq: 1

Then tap OK.

The equation with the coefficients should appear.

To fit the least-squares regression line, tap OK.
Write the equation with coefficients expressed to 2 decimal places.

Write the equation in terms of the variable names. Replace \( x \) with number of minutes and \( y \) with number of calls.

\[
y = -0.63 \times \text{number of minutes} + 11.73
\]

\[
\text{Number of calls} = -0.63 \times \text{number of minutes} + 11.73
\]

**Calculating the least-squares regression line using a CAS calculator**

The least-squares regression can be displayed in two different forms. Care must be taken when completing a statistical analysis that requires specifically one of the formats. The following uses the data from Worked example 3.

To fit the least squares regression line in the form of \( y = ax + b \), complete the following steps. Tap:

- **Calc**
- **Linear Reg**

Set: XList: main\minutes

YList: main\calls

Freq: 1

Copy formula: y1

Then tap OK twice.

Here the equation is written as \( y = -0.6343x + 11.733 \), where \( a \) is the gradient and \( b \) is the \( y \)-intercept.

The equation could also be written in the form \( y = a + bx \). That is, \( y = 11.733 - 0.6343x \).

Different pronumerals are used in both instances; however, the equation is still the same. It is worth noting that \( x \) and \( y \) usually will not change; it is only the coefficient of \( x \) (the gradient of the regression line) and the \( y \)-intercept (\( a \), \( b \) or \( c \), depending on the form) that may be written with a different pronumeral.
We recall from our work in chapter 2 that $r$ gives us a numerical value indicating how close to linear a set of data is. In this example, $r = -0.93$, which indicates a strong, negative linear relationship. Also, $r^2$ indicates the extent to which one variable can be predicted from another variable given that the two variables are linearly related. In this example, $r^2 = 0.87$. This means that about 87% of the variation in the values of $y$ can be accounted for by the variation in the $x$-values.

As you can see, finding the least-squares equation using a CAS calculator is an efficient method.

**Calculating the least-squares regression line by hand**

The least-squares regression equation minimises the average deviation of the points in the data set from the line of best fit. This can be shown using the following summary data and formulas to arithmetically determine the least-squares regression equation.

**Summary data needed:**

- $\bar{x}$ the mean of the independent variable ($x$-variable)
- $\bar{y}$ the mean of the dependent variable ($y$-variable)
- $s_x$ the standard deviation of the independent variable
- $s_y$ the standard deviation of the dependent variable
- $r$ Pearson’s product–moment correlation coefficient.

**Formula to use:**

The general form of the least-squares regression line is

$$y = mx + c$$
where:
the slope of the regression line is \( m = r \frac{s_y}{s_x} \),
the y-intercept of the regression line is \( c = \bar{y} - m\bar{x} \).

Alternatively, if the general form is given as \( y = a + bx \), then \( b = r \frac{s_y}{s_x} \) and \( a = \bar{y} - b\bar{x} \).

A study to find a relationship between the height of husbands and the height of their wives revealed the following details.
Mean height of the husbands: 180 cm
Mean height of the wives: 169 cm
Standard deviation of the height of the husbands: 5.3 cm
Standard deviation of the height of the wives: 4.8 cm
Correlation coefficient, \( r = 0.85 \)
The form of the least-squares regression line is to be:
Height of wife = \( m \times \) height of husband + \( c \)

a Which variable is the dependent variable?
b Calculate the value of \( m \) for the regression line (to 2 decimal places).
c Calculate the value of \( c \) for the regression line (to 2 decimal places).
d Use the equation of the regression line to predict the height of a wife whose husband is 195 cm tall (to the nearest cm).

\textbf{THINK}

\begin{enumerate}
\item Recall that the dependent variable is the subject of the equation in \( y = mx + c \) form; that is, \( y \).
\item The value of \( m \) is the gradient of the regression line. Write the formula and state the required values.
\item Substitute the values into the formula and evaluate \( m \).
\item The value of \( c \) is the y-intercept of the regression line. Write the formula and state the required values.
\item Substitute the values into the formula and evaluate \( c \).
\item State the equation of the regression line, using the values calculated from parts b and c. In this equation, \( y \) represents the height of the wife and \( x \) represents the height of the husband.
\item The height of the husband is 195 cm, so substitute \( x = 195 \) into the equation and evaluate.
\item Write a statement, rounding your answer to the nearest cm.
\end{enumerate}

\textbf{WRITE}

\begin{enumerate}
\item The dependent variable is the height of the wife.
\item \( m = r \frac{s_y}{s_x} \quad r = 0.85, \quad s_y = 4.8 \) and \( s_x = 5.3 \)
\item \( = 0.85 \times \frac{4.8}{5.3} \)
\item \( = 0.7698 \)
\item \( \approx 0.77 \)
\item \( c = \bar{y} - m\bar{x} \)
\item \( \bar{y} = 169, \quad \bar{x} = 180 \) and \( m = 0.7698 \) (from part b)
\item \( = 169 - 0.7698 \times 180 \)
\item \( = 30.436 \)
\item \( \approx 30.44 \)
\item \( y = 0.77x + 30.44 \) or height of wife = \( 0.77 \times \) height of husband + 30.44
\item \( = 0.77 \times 195 + 30.44 \)
\item \( = 180.59 \)
\end{enumerate}

Using the equation of the regression line found, the wife’s height would be 181 cm.
1. Use a CAS calculator to find the equation of the least-squares regression line.
   (a) Enter the data in the Statistics screen.
   (b) Draw the scatterplot by tapping **SetGraph** and completing the fields.
   (c) To fit the regression line \( y = ax + b \), tap:
   - **Calc**
   - **Linear Reg**

   Complete the relevant fields and then tap **OK**.

   Record the values of \( a \), \( b \) and \( r \) for the linear regression equation.

   Tap **OK** again to view the graph.

2. To find the equation of the least-squares regression line ‘by hand’:
   (a) the summary data needed are:
      (i) \( \bar{x} \) and \( s_x \) the mean and standard deviation of the independent variable
      (ii) \( \bar{y} \) and \( s_y \) the mean and standard deviation of the dependent variable
      (iii) \( r \) Pearson’s product–moment correlation coefficient.

   (b) the formulas to use are:
      (i) \( m = r \frac{s_y}{s_x} \)
      (ii) \( c = \bar{y} - mx \)

   where \( m \) is the slope of the regression line and \( c \) is the \( y \)-intercept.

   Alternatively, if the general form of the regression line is given as \( y = a + bx \), then
   \( b = r \frac{s_y}{s_x} \) and \( a = \bar{y} - b\bar{x} \).

---

**EXERCISE 3C**

**Fitting a straight line — least-squares regression**

1. **WE3** Find the equation of the linear regression line for the following data set using the least-squares method.

   \[
   \begin{array}{c|ccccccccc}
   x & 4 & 6 & 7 & 9 & 10 & 12 & 15 & 17 \\
   y & 10 & 8 & 13 & 15 & 14 & 18 & 19 & 23 \\
   \end{array}
   \]

2. **WE3** Find the equation of the linear regression line for the following data set using the least-squares method.

   \[
   \begin{array}{c|cccccccc}
   x & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\
   y & 35 & 28 & 22 & 16 & 19 & 14 & 9 & 7 & 2 \\
   \end{array}
   \]

3. **WE3** Find the equation of the linear regression line for the following data set using the least-squares method.

   \[
   \begin{array}{c|cccccccccccc}
   x & -4 & -2 & -1 & 0 & 1 & 2 & 4 & 5 & 5 & 7 \\
   y & 6 & 7 & 3 & 10 & 16 & 9 & 12 & 16 & 11 & 21 \\
   \end{array}
   \]

4. **WE4** The following summary details were calculated from a study to find a relationship between mathematics exam marks and English exam marks from the results of 120 Year 12 students.

   - Mean mathematics exam mark = 64%
   - Mean English exam mark = 74%
   - Standard deviation of mathematics exam mark = 14.5%
   - Standard deviation of English exam mark = 9.8%
   - Correlation coefficient, \( r = 0.64 \)
The form of the least-squares regression line is to be:
Mathematics exam mark = \( m \times \) English exam mark + \( c \)

a Which variable is the dependent variable (y-variable)?

b Calculate the value of \( m \) for the least-squares regression line (correct to 2 decimal places).

c Calculate the value of \( c \) for the least-squares regression line (correct to 2 decimal places).

d Use the regression line to predict the expected mathematics exam mark if a student scores 85% in an English exam (to the nearest percentage).

5 Find the least-squares regression equation, given the following summary data.

\[ \bar{x} = 5.6 \quad s_x = 1.2 \quad \bar{y} = 110.4 \quad s_y = 5.7 \quad r = 0.7 \]

\[ \bar{x} = 110.4 \quad s_x = 5.7 \quad \bar{y} = 5.6 \quad s_y = 1.2 \quad r = -0.7 \]

\[ \bar{x} = 25 \quad s_x = 4.2 \quad \bar{y} = 10200 \quad s_y = 250 \quad r = 0.88 \]

\[ \bar{x} = 10 \quad s_x = 1 \quad \bar{y} = 20 \quad s_y = 2 \quad r = -0.5 \]

6 Repeat questions 1, 2 and 3, collecting the values for \( \bar{x}, s_x, \bar{y}, s_y \) and \( r \) from the calculator. Use these data to find the least-squares regression equation.

Compare your answers to the ones obtained earlier from questions 1, 2 and 3. What do you notice?

7 A mathematician is interested in the behaviour patterns of her kitten, and collects the following data on two variables. Help her manipulate the data.

<table>
<thead>
<tr>
<th>x</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>20</td>
<td>18</td>
<td>16</td>
<td>14</td>
<td>12</td>
<td>10</td>
<td>8</td>
<td>6</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>

a Fit a least-squares regression line.

b Comment on any interesting features of this line.

c Now fit the ‘opposite regression line’, namely:

<table>
<thead>
<tr>
<th>x</th>
<th>20</th>
<th>18</th>
<th>16</th>
<th>14</th>
<th>12</th>
<th>10</th>
<th>8</th>
<th>6</th>
<th>4</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
</tbody>
</table>

d In comparing the regression line from part a with that from part c, what other interesting features do you find?

8 The best estimate of the least-squares regression line for the scatterplot at right is:

A \( y = 2x \) B \( y = \frac{1}{2}x \) C \( y = \frac{1}{2}x + 2 \)

D \( y = \frac{1}{2}x - 2 \) E \( y = \frac{1}{2}x - 1 \)

9 Given the summary details

\[ \bar{x} = 5.4 \quad s_x = 1.8 \quad \bar{y} = 12.5 \quad s_y = 1.4 \quad r = -0.57 \]

the values of \( m \) and \( c \) for the equation of the regression line \( y = mx + c \) are

A -0.44 and 14.9 B -0.73 and 14.6 C 0.44 and 10.1 D 0.44 and 14.9 E -1.32 and 3.8

10 The life span of adult males in the country of Upper Slobovia over the last 220 years has been recorded.

<table>
<thead>
<tr>
<th>Year</th>
<th>1780</th>
<th>1800</th>
<th>1820</th>
<th>1840</th>
<th>1860</th>
<th>1880</th>
<th>1900</th>
<th>1920</th>
<th>1940</th>
<th>1960</th>
<th>1980</th>
<th>2000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Life span (years)</td>
<td>51.2</td>
<td>52.4</td>
<td>51.7</td>
<td>53.2</td>
<td>53.1</td>
<td>54.7</td>
<td>59.9</td>
<td>62.7</td>
<td>63.2</td>
<td>66.8</td>
<td>72.7</td>
<td>79.2</td>
</tr>
</tbody>
</table>

a Fit a least-squares regression line to these data.
b Plot the data and the regression line on a scatterplot.
c Do the data really look linear? Discuss.

11 The price of a long distance telephone call changes as the duration of the call increases. The cost of a sample of calls from Melbourne to Slovenia are summarised in the table below.

<table>
<thead>
<tr>
<th>Cost of call ($)</th>
<th>1.25</th>
<th>1.85</th>
<th>2.25</th>
<th>2.50</th>
<th>3.25</th>
<th>3.70</th>
<th>4.30</th>
<th>4.90</th>
<th>5.80</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration of call (seconds)</td>
<td>30</td>
<td>110</td>
<td>250</td>
<td>260</td>
<td>300</td>
<td>350</td>
<td>420</td>
<td>500</td>
<td>600</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cost of call ($)</th>
<th>7.50</th>
<th>8.00</th>
<th>9.25</th>
<th>10.00</th>
<th>12.00</th>
<th>13.00</th>
<th>14.00</th>
<th>16.00</th>
<th>18.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration of call (seconds)</td>
<td>840</td>
<td>1000</td>
<td>1140</td>
<td>1200</td>
<td>1500</td>
<td>1860</td>
<td>2400</td>
<td>3600</td>
<td>7200</td>
</tr>
</tbody>
</table>

a What is the independent variable likely to be?
b Fit a least-squares regression line to the data.
c View the data on a scatterplot and comment on the reliability of the regression line in predicting the cost of telephone calls. (That is, consider whether the regression line you found proves that costs of calls and duration of calls are related.)

12 In a study to find a relationship between the height of plants and the hours of daylight they were exposed to, the following summary details were obtained.

Mean height of plants = 40 cm
Mean hours of daylight = 8 hours
Standard deviation of plant height = 5 cm
Standard deviation of daylight hours = 3 hours
Pearson’s correlation coefficient = 0.9

The most appropriate regression equation is:
A height of plant (cm) = −13.6 + 0.54 × hours of daylight
B height of plant (cm) = −8.5 + 0.34 × hours of daylight
C height of plant (cm) = 2.1 + 0.18 × hours of daylight
D height of plant (cm) = 28.0 + 1.50 × hours of daylight
E height of plant (cm) = 35.68 + 0.54 × hours of daylight

13 You saw with the 3-median method that at least six points were needed to perform meaningful analysis and generate a linear equation. Is the same true of least-squares linear regression?

Consider the following data set.

<table>
<thead>
<tr>
<th>x</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>12</td>
<td>16</td>
<td>17</td>
<td>21</td>
<td>25</td>
<td>29</td>
</tr>
</tbody>
</table>

a Perform a least-squares regression on the first two points only.
b Now add the 3rd point and repeat.
c Repeat for the 4th, 5th and 6th points.
d Comment on your results.

3D Interpretation, interpolation and extrapolation

Interpreting slope and intercept (m and c)

Once you have a linear regression line, the slope and intercept can give important information about the data set.

The slope \( m \) indicates the rate at which the data are increasing or decreasing.
The \( y \)-intercept indicates the approximate value of the data when \( x = 0 \).
In the study of the growth of a species of bacterium, it is assumed that the growth is linear. However, it is very expensive to measure the number of bacteria in a sample. Given the data listed below, find:

a. the relationship on a CAS calculator
b. the rate at which bacteria are growing
c. the number of bacteria at the start of the experiment.

<table>
<thead>
<tr>
<th>Day of experiment</th>
<th>1</th>
<th>4</th>
<th>5</th>
<th>9</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of bacteria</td>
<td>500</td>
<td>1000</td>
<td>1100</td>
<td>2100</td>
<td>2500</td>
</tr>
</tbody>
</table>

**THINK**

1. On the Statistics screen, label `list1` as ‘day’ and `list2` as ‘bacteria’. Enter the data accordingly. To find the linear regression model, tap:
   - Calc
   - Linear Reg
   Set:
     - XList: main\day
     - YList: main\bacteria
     - Freq: 1

   Then tap OK. Note the pronumerals used by the calculator when writing your answer.

2. To plot the points and draw the linear regression line, tap:
   - SetGraph
   - Settings
   Select the relevant variables and tap:
   - Set
   - \text{list1}

3. Write the equation.

   \[ y = 202.5 + 206.25x \]

   Note: \textit{r} indicates a strong positive linear association and from \textit{r}^2 we see that 99\% of the variation in the number of bacteria can be explained by the number of days over which the experiment ran. This means that this data set provides a very good predictor of the number of bacteria present.
The rate at which bacteria are growing is defined by the gradient of the least-squares regression.

The number of bacteria at the start of the experiment is denoted by the y-intercept of the least-squares regression line.

As can be observed from the graph, the number of bacteria when time = 0 (about 200) can be seen as the y-intercept of the graph, and the daily rate of increase (about 200) is the gradient of the straight line.

Interpolation and extrapolation

As we have already observed, any linear regression method produces a linear equation in the form:

\[ y = (\text{gradient}) \times x + (\text{y-intercept}) \]

or

\[ y = m \times x + c \]

This line can be used to ‘predict’ data values for a given value of \( x \). Of course, these are only approximations, since the regression line itself is only an estimate of the ‘true’ relationship between the bivariate data. However, they can still be used, in some cases, to provide additional information about the data set (that is, make predictions).

There are two types of prediction: interpolation and extrapolation.

Interpolation

Interpolation is the use of the regression line to predict values in between two values already in the data set. If the data are highly linear (\( r \) near +1 or −1) then we can be confident that our interpolated value is quite accurate. If the data are not highly linear (\( r \) near 0) then our confidence is duly reduced. For example, medical information collected from a patient every third day would establish data for day 3, 6, 9, . . . and so on. After performing regression analysis, it is likely that an interpolation for day 4 would be accurate, given good \( r \) values.

Extrapolation

Extrapolation is the use of the regression line to predict values smaller than the smallest value already in the data set or larger than the largest value.

Two problems may arise in attempting to extrapolate from a data set. Firstly, it may not be reasonable to extrapolate too far away from the given data values. For example, suppose there is a weather data set for 5 days. Even if it is highly linear (\( r \) near +1 or −1) a regression line used to predict the same data 15 days in the future is highly risky. Weather has a habit of randomly fluctuating and patterns rarely stay stable for very long.

Secondly, the data may be highly linear in a narrow band of the given data set. For example, there may be data on stopping distances for a train at speeds of between 30 and 60 km/h. Even if they are highly linear in this range, it is unlikely that things are similar at very low speeds (0–15 km/h) or high speeds (over 100 km/h).

Generally, one should feel more confident about the accuracy of a prediction derived from interpolation than one derived from extrapolation. Of course, it still depends upon the correlation coefficient (\( r \)). The closer to linearity the data are, the more confident our predictions in all cases.
Using interpolation and the following data set, predict the height of an 8-year-old girl.

<table>
<thead>
<tr>
<th>Age (years)</th>
<th>1</th>
<th>3</th>
<th>5</th>
<th>7</th>
<th>9</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height (cm)</td>
<td>60</td>
<td>76</td>
<td>115</td>
<td>126</td>
<td>141</td>
<td>148</td>
</tr>
</tbody>
</table>

THINK

1. On the Statistics screen, label list1 as ‘age’, list2 as ‘height’ and enter the data. Then tap:
   - SetGraph
   - Settings
   Set:
   Type: Scatter   XLList: main\age
   YList: main\height   Freq: 1
   Mark: square

   Tap:
   - Set
   - [Graph]

2. To fit the least-squares regression line in the form \( y = ax + b \), tap:
   - Calc
   - Linear Reg
   Set:
   XLList: main\x
   YList: main\y
   Freq: 1
   Copy Formula: y1

   Then tap OK. Since the data was a good fit, \( r = 0.97 \), one can be confident of an accurate prediction.

3. To fit a least-squares regression line, tap OK. Using the regression equation, find the height when the age is 8. Take into account that in \( y = 9.23x + 55.63 \), \( x \) is age in years and \( y \) is height in centimetres.

\[
y = 9.23x + 55.63
\]

Height = \( 9.23 \times 8 + 55.63 \)
= \( 9.23 \times 8 + 55.63 \)
= 129.5 cm
To find the height when the age is 8, tap Main on the icon panel. Complete the entry line as: 
\[ y_1(8) \]
Then press \( \text{EXE} \). 
Round the answer to 1 decimal place.

Write the answer. At age 8, the predicted height is 129.5 cm.

WORKED EXAMPLE 7

Use extrapolation and the data from Worked example 6 to predict the height of the girl when she turns 15. Discuss the reliability of this prediction.

**THINK**

1. Use the regression equation to calculate the girl’s height at age 15. Alternatively, use the CAS calculator to find \( f_1(15) \).

**WRITE**

\[
\text{Height} = 9.23 \times \text{age} + 55.63
\]
\[
= 9.23 \times 15 + 55.63
\]
\[
= 194.08 \text{ cm}
\]

Since we have extrapolated the result (that is, since the greatest age in our data set is 11 and we are predicting outside the data set) we cannot claim that the prediction is reliable.

**REMEMBER**

1. The slope \((m)\) indicates the rate at which the data are increasing or decreasing.
2. The \(y\)-intercept indicates the approximate value of the data when \(x = 0\).
3. Interpolation is the use of the regression line to predict values in between two values already in the data set.
4. Extrapolation is the use of the regression line to predict values smaller than the smallest value already in the data set or larger than the largest value.
5. The reliability of these predictions depends on the value of \(r^2\) and the limits of the data set.

**EXERCISE**

**3D Interpretation, interpolation and extrapolation**

A drug company wishes to test the effectiveness of a drug to increase red blood cell counts in people who have a low count. The following data are collected.

<table>
<thead>
<tr>
<th>Day of experiment</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Red blood cell count</td>
<td>210</td>
<td>240</td>
<td>230</td>
<td>260</td>
<td>260</td>
<td>290</td>
</tr>
</tbody>
</table>
Find:
\(a\) the relationship in the form \(y = a + bx\) on a CAS calculator
\(b\) the rate at which the red blood cell count was changing
\(c\) the red blood cell count at the beginning of the experiment (that is, on day 0).

2 A wildlife exhibition is held over 6 weekends and features still and live displays. The number of live animals that are being exhibited varies each weekend. The number of animals participating, together with the number of visitors to the exhibition each weekend, is shown below.

<table>
<thead>
<tr>
<th>Number of animals</th>
<th>6</th>
<th>4</th>
<th>8</th>
<th>5</th>
<th>7</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of visitors</td>
<td>311</td>
<td>220</td>
<td>413</td>
<td>280</td>
<td>379</td>
<td>334</td>
</tr>
</tbody>
</table>

Find:
\(a\) the rate of increase of visitors as the number of live animals is increased
\(b\) the predicted number of visitors if there are no live animals.

3 An electrical goods warehouse produces the following data showing the selling price of electrical goods to retailers and the volume of those sales.

<table>
<thead>
<tr>
<th>Selling price ($)</th>
<th>60</th>
<th>80</th>
<th>100</th>
<th>120</th>
<th>140</th>
<th>160</th>
<th>200</th>
<th>220</th>
<th>240</th>
<th>260</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sales volume ((\times 1000))</td>
<td>400</td>
<td>300</td>
<td>275</td>
<td>250</td>
<td>210</td>
<td>190</td>
<td>150</td>
<td>100</td>
<td>50</td>
<td>0</td>
</tr>
</tbody>
</table>

Perform a least-squares regression analysis and discuss the meaning of the gradient and y-intercept.

4 A study of the dining-out habits of various income groups in a particular suburb produces the results shown in the table below.

<table>
<thead>
<tr>
<th>Weekly income ($)</th>
<th>100</th>
<th>200</th>
<th>300</th>
<th>400</th>
<th>500</th>
<th>600</th>
<th>700</th>
<th>800</th>
<th>900</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of restaurant visits per year</td>
<td>5.8</td>
<td>2.6</td>
<td>1.4</td>
<td>1.2</td>
<td>6</td>
<td>4.8</td>
<td>11.6</td>
<td>4.4</td>
<td>12.2</td>
<td>9</td>
</tr>
</tbody>
</table>

Use the data to predict:
\(a\) \(\text{WE6}\) the number of visits per year by a person on a weekly income of $680
\(b\) \(\text{WE7}\) the number of visits per year by a person on a weekly income of $2000.

5 Fit a least-squares regression line to the following data.

<table>
<thead>
<tr>
<th>(x)</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>8</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>(y)</td>
<td>2</td>
<td>3</td>
<td>7</td>
<td>12</td>
<td>17</td>
<td>21</td>
<td>27</td>
<td>35</td>
</tr>
</tbody>
</table>

Find:
\(a\) the regression equation
\(b\) \(y\) when \(x = 3\)
\(c\) \(y\) when \(x = 12\)
\(d\) \(x\) when \(y = 7\)
\(e\) \(x\) when \(y = 25\).
\(f\) Which of \(b\) to \(e\) above are extrapolations?

6 The following table represents the costs for shipping a consignment of shoes from Melbourne factories. The cost is given in terms of distance from Melbourne. There are two factories that can be used. The data are summarised below.

<table>
<thead>
<tr>
<th>Distance from Melbourne (km)</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>70</th>
<th>80</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factory 1 cost ($)</td>
<td>70</td>
<td>70</td>
<td>90</td>
<td>100</td>
<td>110</td>
<td>120</td>
<td>150</td>
<td>180</td>
</tr>
<tr>
<td>Factory 2 cost ($)</td>
<td>70</td>
<td>75</td>
<td>80</td>
<td>100</td>
<td>100</td>
<td>115</td>
<td>125</td>
<td>135</td>
</tr>
</tbody>
</table>
7 A factory produces calculators. The least-squares regression line for cost of production \((C)\) as a function of numbers of calculators \((n)\) produced is given by:

\[
C = 600 + 7.76n
\]

Furthermore, this function is deemed accurate when producing between 100 and 1000 calculators.

a Find the cost to produce 200 calculators.

b How many calculators can be produced for $2000?

c Find the cost to produce 10,000 calculators.

d What are the ‘fixed’ costs for this production?

e Which of a to c above is an interpolation?

8 A study of the relationship between IQ and results in a mathematics exam produced the following results. Unfortunately, some of the data were lost. Copy and complete the table by using the least-squares equation with the data that were supplied.

Note: Only use \((x, y)\) pairs if both are in the table.

<table>
<thead>
<tr>
<th>IQ</th>
<th>80</th>
<th>92</th>
<th>102</th>
<th>105</th>
<th>107</th>
<th>111</th>
<th>115</th>
<th>121</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test result (%)</td>
<td>56</td>
<td>60</td>
<td>68</td>
<td>65</td>
<td>74</td>
<td>71</td>
<td>73</td>
<td>92</td>
</tr>
</tbody>
</table>

9 The least-squares regression line for salary \((s)\) as a function of number of years of schooling \((n)\) is given by the rule: \(s = 18500 + 900n\).

a Find the salary for a person who completed 10 years of schooling.

b Find the salary for a person who completed 12 years of schooling.

c Find the salary for a person who completed 15 years of schooling.

d Mary earned $30,400. What was her likely schooling experience?

e Discuss the reasonableness of predicting salary on the basis of years of schooling.

Residual analysis

There are situations where the mere fitting of a regression line to some data is not enough to convince us that the data set is truly linear. Even if the correlation is close to +1 or -1 it still may not be convincing enough.

The next stage is to analyse the residuals, or deviations, of each data point from the straight line.

A residual is the vertical difference between each data point and the regression line.

Calculating residuals

A sociologist gathers data on the heights of brothers and sisters in families from different ethnic backgrounds. He enters his records in the table below.

<table>
<thead>
<tr>
<th>(x)</th>
<th>2</th>
<th>3</th>
<th>5</th>
<th>8</th>
<th>9</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>(y)</td>
<td>3</td>
<td>7</td>
<td>12</td>
<td>10</td>
<td>12</td>
<td>16</td>
</tr>
</tbody>
</table>

He then plots each point, and fits a regression line as shown in figure 1, which follows. He then decides to calculate the residuals.

The residuals are simply the vertical distances from the line to each point. These lines are shown as blue and red bars in figure 2.
Finally, he calculates the residuals for each data point. This is done in two steps.

Step 1. He calculates the predicted value of \( y \) from the regression equation.

Step 2. He calculates the difference between this predicted value and the original value.

**WORKED EXAMPLE 8**

Consider the data set below. Draw a least-squares regression line and calculate the residuals.

<table>
<thead>
<tr>
<th>( x )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y )</td>
<td>5</td>
<td>6</td>
<td>8</td>
<td>15</td>
<td>24</td>
<td>47</td>
<td>77</td>
<td>112</td>
<td>187</td>
<td>309</td>
</tr>
</tbody>
</table>

**THINK**

1. On the Statistics screen, label list 1 as ‘\( x \)’, list 2 as ‘\( y \)’ and enter the data values.

   Complete the steps from Worked example 6 to draw a scatterplot.

   Then tap:
   - Calc
   - Linear Reg

   Set the screen as shown and tap OK twice.

**WRITE/DISPLAY**

So \( y = 28.7x - 78.7 \)

Method 1: Predicting from the equation

2. The equation of the least-squares regression line is generated. The results are:

   - Gradient (\( m \)) = 28.7
   - y-intercept (\( c \)) = -78.7
   - Correlation (\( r \)) = 0.87
3 Calculate residuals for each point by substituting each \( x \)-value into the equation \( y = 28.7x - 78.7 \).

<table>
<thead>
<tr>
<th>( x )-values</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y )-values</td>
<td>5.0</td>
<td>6.0</td>
<td>8.0</td>
<td>15.0</td>
<td>24.0</td>
</tr>
<tr>
<td><strong>Predicted</strong> ( y )-values</td>
<td>(-50.05)</td>
<td>(-21.38)</td>
<td>7.3</td>
<td>35.98</td>
<td>64.66</td>
</tr>
<tr>
<td><strong>Residuals</strong> ((y - y_{\text{pred}}))</td>
<td>55.05</td>
<td>27.38</td>
<td>0.7</td>
<td>(-20.98)</td>
<td>(-40.66)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( x )-values</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y )-values</td>
<td>47.0</td>
<td>77.0</td>
<td>112.0</td>
<td>187.0</td>
<td>309.0</td>
</tr>
<tr>
<td><strong>Predicted</strong> ( y )-values</td>
<td>93.34</td>
<td>122.02</td>
<td>150.7</td>
<td>179.38</td>
<td>208.06</td>
</tr>
<tr>
<td><strong>Residuals</strong> ((y - y_{\text{pred}}))</td>
<td>(-46.34)</td>
<td>(-45.02)</td>
<td>(-38.7)</td>
<td>7.62</td>
<td>100.94</td>
</tr>
</tbody>
</table>

Method 2: Using a CAS calculator

2 The values of the residuals are placed in list3. Write down all the values displayed in the list as your answer.

3 To plot the points and draw the linear regression line, tap:
   - SetGraph
   - Settings
   Select the relevant variables and tap:
   - Set
   - List

Notes
1. The residuals may be determined by \((y - y_{\text{pred}})\); that is, the actual values minus the predicted values.
2. The sum of all the residuals always adds to 0 (or very close to 0 after rounding), when least-squares regression is used. This can act as a check for our calculations.
Introduction to residual analysis

As we observed in the previous worked example, there is not really a good fit between the data and the least-squares regression line; however, there seems to be a pattern in the residuals. How can we observe this pattern in more detail?

The answer is to plot the residuals themselves against the original $x$-values. If there is a pattern, it should become clearer after they are plotted.

Types of residual plots

There are three basic types of residual plots. Each type indicates whether or not a linear relationship exists between the two variables under investigation.

Note: The points are joined together to see the patterns more clearly.

The points of the residuals are randomly scattered above and below the $x$-axis. The original data probably have a linear relationship.

The points of the residuals show a curved pattern ($\cap$), with a series of negative, then positive and back to negative residuals along the $x$-axis. The original data probably have a non-linear relationship. Transformation of the data may be required.

The points of the residuals show a curved pattern ($\cup$), with a series of positive, then negative and back to positive residuals along the $x$-axis. The original data probably have a non-linear relationship. Transformation of the data may be required.

The transformation of data suggested in the last two residual plots will be studied in more detail in the next section.
Using the same data as in Worked example 8, plot the residuals and discuss the features of the residual plot.

**THINK**

**Method 1: Plotting the graph**

1. Generate a table of values of residuals against x.

<table>
<thead>
<tr>
<th>x-values</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residuals ( (y - y_{pred}) )</td>
<td>55.05</td>
<td>27.38</td>
<td>0.7</td>
<td>-20.98</td>
<td>-40.66</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>x-values</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residuals ( (y - y_{pred}) )</td>
<td>-46.34</td>
<td>-45.02</td>
<td>-38.7</td>
<td>7.62</td>
<td>100.94</td>
</tr>
</tbody>
</table>

2. Plot the residuals against x. If the relationship was linear the residuals would be scattered randomly above and below the line. However, in this instance there is a pattern which looks somewhat like a parabola. This should indicate that the data were not really linear, but were more likely to be quadratic. (The line segments between the points are included for clarity.)

**Method 2: Using a CAS calculator**

1. To set up a residual plot, return to the Statistics screen.

2. Then tap:
   - **SetGraph**
   - **Setting**
   Set the screen as shown since the residuals are in list3.
Comment on the residual plot and its relevance.

The residual plot indicates a distinct pattern suggesting that a non-linear model could be more appropriate.

REMEMBER

1. Calculate predicted values ($y_{\text{pred}}$) from the regression equation ($y = mx + c$) for all values of $x$.
2. Calculate residuals ($y - y_{\text{pred}}$) for all values of $x$ (that is, actual values – predicted values).
3. Observe the data and the plot residuals.
4. If a residual plot shows points randomly scattered above and below zero then the original data probably have a linear relationship.
5. Conversely, if a residual plot shows some sort of pattern, then there is probably not a linear relationship between the original data sets.

EXERCISE 3E

Residual analysis

1. Find the residuals for the following data.

<table>
<thead>
<tr>
<th>$x$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>1</td>
<td>9.7</td>
<td>12.7</td>
<td>13.7</td>
<td>14.4</td>
<td>14.5</td>
</tr>
</tbody>
</table>

2. For the results of question 1, plot the residuals and discuss whether the data are really linear.

3. Which of the following data sets are likely to be linear?

   i
   ![Graph i]
   
   A All of them
   D ii only

   ii
   ![Graph ii]
   
   B None of them
   E ii and iii only

   iii
   ![Graph iii]
   
   C i and iii only

EXAM TIP

Students find it difficult to explain why a residual plot might suggest a non-linear relationship may be more appropriate.  
[Assessment report 2 2005]
4 Consider the following table from a survey conducted at a new computer manufacturing factory. It shows the percentage of defective computers produced on 8 different days after the opening of the factory.

<table>
<thead>
<tr>
<th>Day</th>
<th>2</th>
<th>4</th>
<th>5</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Defective rate (%)</td>
<td>15</td>
<td>10</td>
<td>12</td>
<td>4</td>
<td>9</td>
<td>7</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

The results of least-squares regression were: $m = -1.19$, $c = 16.34$, $r = -0.87$.

a Find the predicted defective rate ($y_{\text{pred}}$) based upon this regression line.
b Find the residuals ($y - y_{\text{pred}}$).
c Plot the residuals and comment on the likely linearity of the data.
d Estimate the defective rate after the first day of the factory’s operation.
e Estimate when the defective rate will be at zero. Comment on this result.

5 The following data represent the number of tourists booked into a hotel in central Queensland during the first week of a drought. (Assume Monday = 1.)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Bookings in hotel</td>
<td>158</td>
<td>124</td>
<td>74</td>
<td>56</td>
<td>31</td>
<td>35</td>
<td>22</td>
</tr>
</tbody>
</table>

The results of least-squares regression were: $m = -22.5$, $c = 161.3$, $r = -0.94$.

a Find the predicted hotel bookings ($y_{\text{pred}}$) based upon this regression line.
b Find the residuals ($y - y_{\text{pred}}$).
c Plot the residuals and comment on the likely linearity of the data.
d Would this regression line be a typical one for this hotel?

6 (MC) A least-squares regression is fitted to the points shown in the scatterplot at right. Which of the following looks most similar to the plot of residuals?

- [A]
- [B]
- [C]
From each table of residuals, decide whether or not the original data set is likely to be linear.

**a**

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
<th>Residuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>-1.34</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>-0.3</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>-0.1</td>
</tr>
<tr>
<td>4</td>
<td>11</td>
<td>0.2</td>
</tr>
<tr>
<td>5</td>
<td>21</td>
<td>0.97</td>
</tr>
<tr>
<td>6</td>
<td>20</td>
<td>2.3</td>
</tr>
<tr>
<td>7</td>
<td>19</td>
<td>1.2</td>
</tr>
<tr>
<td>8</td>
<td>15</td>
<td>-0.15</td>
</tr>
<tr>
<td>9</td>
<td>12</td>
<td>-0.9</td>
</tr>
<tr>
<td>10</td>
<td>6</td>
<td>-2.8</td>
</tr>
</tbody>
</table>

**b**

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
<th>Residuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td>56</td>
<td>0.12</td>
</tr>
<tr>
<td>21</td>
<td>50</td>
<td>-0.56</td>
</tr>
<tr>
<td>19</td>
<td>43</td>
<td>1.30</td>
</tr>
<tr>
<td>16</td>
<td>41</td>
<td>0.20</td>
</tr>
<tr>
<td>14</td>
<td>37</td>
<td>-1.45</td>
</tr>
<tr>
<td>11</td>
<td>31</td>
<td>2.16</td>
</tr>
<tr>
<td>9</td>
<td>28</td>
<td>-0.22</td>
</tr>
<tr>
<td>6</td>
<td>22</td>
<td>-3.56</td>
</tr>
<tr>
<td>4</td>
<td>19</td>
<td>2.19</td>
</tr>
<tr>
<td>3</td>
<td>17</td>
<td>-1.05</td>
</tr>
</tbody>
</table>

**c**

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
<th>Residuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.2</td>
<td>23</td>
<td>0.045</td>
</tr>
<tr>
<td>1.6</td>
<td>25</td>
<td>0.003</td>
</tr>
<tr>
<td>1.8</td>
<td>24</td>
<td>-0.023</td>
</tr>
<tr>
<td>2.0</td>
<td>26</td>
<td>-0.089</td>
</tr>
<tr>
<td>2.2</td>
<td>28</td>
<td>-0.15</td>
</tr>
<tr>
<td>2.6</td>
<td>29</td>
<td>-0.98</td>
</tr>
<tr>
<td>2.7</td>
<td>34</td>
<td>-0.34</td>
</tr>
<tr>
<td>2.9</td>
<td>42</td>
<td>-0.01</td>
</tr>
<tr>
<td>3.0</td>
<td>56</td>
<td>0.45</td>
</tr>
<tr>
<td>3.1</td>
<td>64</td>
<td>1.23</td>
</tr>
</tbody>
</table>

Consider the following data set.

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>33</td>
</tr>
<tr>
<td>4</td>
<td>60</td>
</tr>
<tr>
<td>5</td>
<td>94</td>
</tr>
<tr>
<td>6</td>
<td>134</td>
</tr>
<tr>
<td>7</td>
<td>180</td>
</tr>
<tr>
<td>8</td>
<td>240</td>
</tr>
<tr>
<td>9</td>
<td>300</td>
</tr>
<tr>
<td>10</td>
<td>390</td>
</tr>
</tbody>
</table>

**a** Plot the data and fit a least-squares regression line.

**b** Find the correlation coefficient. What does this mean?

**c** Calculate the coefficient of determination. How does this explain variation?

**d** Find the residuals.

**e** Plot the residuals on a separate graph. Comment on the residual plot.

---

**Transforming to linearity**

Although linear regression might produce a ‘good’ fit (high r value) to a set of data, the data set may still be non-linear. To remove (as much as is possible) such non-linearity, the data can be transformed.

Either the x-values, y-values, or both may be transformed in some way so that the transformed data are more linear. This enables more accurate predictions (extrapolations and interpolations) from the regression equation. In Further Mathematics, six transformations are studied:

- **Logarithmic transformations:** \( y \) versus \( \log_{10}(x) \)
- **Quadratic transformations:** \( y \) versus \( x^2 \)
- **Reciprocal transformations:** \( y \) versus \( \frac{1}{x} \)

Where \( y \) is the dependent variable and \( x \) is the independent variable.
Choosing the correct transformations

To decide on an appropriate transformation, examine the points on a scatterplot with high values of \( x \) and/or \( y \) (that is, away from the origin) and decide for each axis whether it needs to be stretched or compressed to make the points line up. The best way to see which of the transformations to use is to look at a number of ‘data patterns’.

### Quadratic transformations

1. Use \( y \) versus \( x^2 \) transformation.

![Graph showing quadratic transformation](image)

2. Use \( y \) versus \( x^2 \) transformation.

![Graph showing quadratic transformation](image)

3. Use \( y^2 \) versus \( x \) transformation.

![Graph showing quadratic transformation](image)

4. Use \( y^2 \) versus \( x \) transformation.

![Graph showing quadratic transformation](image)

### Logarithmic and reciprocal transformations

1. Use \( y \) versus \( \log_{10} (x) \) or \( y \) versus \( \frac{1}{x} \) transformation.

![Graph showing logarithmic transformation](image)

2. Use \( y \) versus \( \log_{10} (x) \) or \( y \) versus \( \frac{1}{x} \) transformation.

![Graph showing logarithmic transformation](image)

3. Use \( \log_{10} (y) \) versus \( x \) or \( \frac{1}{y} \) versus \( x \) transformation.

![Graph showing logarithmic transformation](image)

4. Use \( \log_{10} (y) \) versus \( x \) or \( \frac{1}{y} \) versus \( x \) transformation.

![Graph showing logarithmic transformation](image)

### Testing transformations
As there are at least two possible transformations for any given non-linear scatterplot, the decision as to which is the best comes from the coefficient of correlation. The least-squares regression equation that has a Pearson correlation coefficient closest to 1 or -1 should be considered as the most appropriate. However, there may be very little difference so common sense needs to be applied. It is sometimes more useful to use a linear function rather than one of the six non-linear functions.

WORKED EXAMPLE 10

Apply a parabolic transformation to the data from Worked example 8, reproduced here. The regression line has been determined as

\[ y = 28.7x - 78.7 \text{ with } r = 0.87. \]

<table>
<thead>
<tr>
<th>x</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>5</td>
<td>6</td>
<td>8</td>
<td>15</td>
<td>24</td>
<td>47</td>
<td>77</td>
<td>112</td>
<td>187</td>
<td>309</td>
</tr>
</tbody>
</table>

**THINK**

1. Plot the data and the regression line to check that a parabolic transformation is suitable. One option is to stretch the x-axis. This requires an \( x^2 \) transformation.

2. Square the x-values to give a transformed data set.

3. Find the equation of the least-squares regression line of the transformed data. Using a calculator or spreadsheet:
   (a) gradient (\( m \)) = 2.78
   (b) y-intercept (\( c \)) = -28.0
   (c) correlation (\( r \)) = 0.95.

4. Plot the new transformed data and regression line.

**WRITE/DRAW**

\[ y = 2.78x_T - 28 \text{ where } x_T = x^2; \text{ that is, } x_T \text{ represents the transformed data.} \]

**Notes**

1. These data are still not truly linear, but are ‘less’ parabolic. Perhaps another transformation would improve things even further. This could involve transforming the y-values, such as \( \log_{10}(y) \), and applying another linear regression.

2. See Worked examples 11 and 12 for a CAS calculator approach to transforming data.
Apply a logarithmic transformation to the following data which represent a patient’s heart rate as a function of time. The regression line has been determined as

\[
\text{Heart rate} = 93.2 - 6.97 \times \text{time}, \quad \text{with } r = -0.90.
\]

<table>
<thead>
<tr>
<th>Time after operation (h)</th>
<th>x</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heart rate (beats/min)</td>
<td>y</td>
<td>100</td>
<td>80</td>
<td>65</td>
<td>55</td>
<td>50</td>
<td>51</td>
<td>48</td>
<td>46</td>
</tr>
</tbody>
</table>

THINK

1. To view the scatterplot, on the Statistics screen, label list1 as ‘time’, list2 as ‘heart’ and enter the data. Then tap:
   - SetGraph
   - Settings
   Set:
   
   Type: Scatter   XList: main\time
   YList: main\heart   Freq: 1

   Tap:
   - Set
   - OK

2. To show the linear regression equation, tap:
   - Calc
   - Linear Reg
   Set:
   
   XList: main\time
   YList: main\heart
   Freq: 1
   Copy Formula: y1
   Copy Residual: list3

   Then tap OK. This displays the coefficients of the equation.

3. To view the residual plot, on the Statistics screen, tap:
   - SetGraph
   - Settings
   Set:
   
   Type: Scatter   XList: main\time
   YList: list3   Freq: 1
   Mark: square

   Tap:
   - OK
   - OK

Heart rate = 93.2 – 6.9 × time
Transform the $y$ (heart) data by calculating the log of $y$ (heart). Tap the cell where Cal and list3 intersect; this sets list3 as the column where the $\log_{10}$ (heart) values will be stored. Tap the Cal= cell and complete the entry line as:

$\log(\text{main\heart})$

Then press $\text{Ex}$.

To view the equation of the regression line, draw a scatterplot using XList: main\time and YList: list3 [$\log_{10}$ (heart)]. Tap:
- Calc
- Linear Reg
Set:
- XList: main\time
- YList: list3 [$\log_{10}$ (heart)]
Then tap OK.
The equation for the linear regression should appear. The $r$ and $r^2$ values are also displayed.

To view the transformed data and regression line, tap OK.

Analysing the results, there has been an improvement of the correlation coefficient from $r = -0.90$ to a marginally better $r$ value of $-0.93$. State the regression equation, keeping in mind that $y$ has been transformed to logheartrate.

$r$ has improved from $-0.90$ to $-0.93$

$y_T = -0.0456x + 1.9809$

In the context of the question, $\log_{10}$ of the heart rate is equal to $-0.0456 \times$ number of hours after the operation + 1.9809.
Further investigation

Often all appropriate transformations need to be performed to choose the best one. Extend Worked example 11 by compressing the $y$ data using the reciprocals of the $y$ data or even compress the $x$ data. Go back to the steps for transforming the data. Did you get a better $r$ value and thus a more reliable line of best fit? (*Hint:* The best transformation gives $r = -0.98$.)

Using the transformed line for predictions

When predicting $y$-values using either the $x^2$ or $\log_{10}(x)$ transformation, remember to transform the original $x$-value first. For instance, in Worked example 10, if we wish to know the value of $y$ when $x = 2.5$, we must square $x$ first (6.25) and put this value into the transformed linear regression equation.

---

**WORKED EXAMPLE 12**

- **Using a CAS calculator, apply a reciprocal transformation to the following data.**
- **Use the transformed regression equation to predict the number of students wearing a jumper when the temperature is 12°C.**

<table>
<thead>
<tr>
<th>Temperature (°C)</th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
<th>30</th>
<th>35</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number of students in a class wearing jumpers</strong></td>
<td>$y$</td>
<td>18</td>
<td>10</td>
<td>6</td>
<td>5</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

**THINK**

1. On the Statistics screen, label `list1` as ‘temp’, `list2` as ‘students’ and enter the data. Then tap:
   - **SetGraph**
   - **Settings**
   Set Type as Scatter, XList as main\temp, YList as main\students, Freq as 1 and Mark as square. Then tap:
   - **Set**
   - **Calc**
   - **Linear Reg**
   Set the same XList and YList and set the residuals to `list3`. Then tap **OK** twice.

2. A view of the residual plot indicates a transformation is appropriate since there is a clear pattern in the points. To show this on the Statistics screen, tap:
   - **SetGraph**
   - **Settings**
   Set:
     - XList: main\temp
     - YList: list3 (the residuals are stored here)
     - Freq: 1
     - Mark: square
   Then tap **Set**.
The $x$-values should be compressed, so it may be appropriate to transform the $x$-data by calculating the reciprocal of temperature.

Label list3 as ‘rectemp’. To transform the values found in list1 (temp), tap the cell where Cal and ‘rectemp’ intersect; this sets ‘rectemp’ as the column where the reciprocal values will be stored. Tap the Cal= cell and complete the entry line as:

\[
\frac{1}{\text{temp}}
\]

Then press \(=\).

To view the transformed data and regression line \(y = a + \frac{b}{x}\), tap:

- **SetGraph**
- **Setting**
  
  Set: Type: Scatter
  
  XList: main/rectemp
  
  YList: main/students

Tap:

- **Set**
- **Calc**
- **Linear Reg**

Set the same XList and YList. Then tap **OK** twice.

Write your answer.

b 1 Transform the $x$-value involved in the prediction.

2 Use the transformed value, 0.083 33, in the transformed regression equation to find \(y\) (the number of students wearing jumpers).

3 Write your answer to the nearest whole number.

If temperature \((x) = 12 \, ^\circ\text{C}\),

\[
x_T = \frac{1}{x} = \frac{1}{12} = 0.083 33.
\]

\[
y = 94.583x_T - 0.4354
\]

Number of students wearing jumpers

\[
= \frac{94.583}{\text{temperature}} - 0.4354
\]

\[
= \frac{94.583}{12} - 0.4354
\]

\[
= 7.447
\]

7 students are predicted to wear jumpers.
To transform to linearity:
Step 1. Plot the original data and least-squares regression line. Examine the residuals or do a residual plot to examine if a pattern suggests the data are non-linear.
Step 2. Examine the high values of x and/or y and decide if the data need to be compressed or stretched to make them linear (see diagrams on page 137).
Step 3. Transform the data by:
   (a) compressing x- or y-values using the reciprocal ($\frac{1}{x}$ or $\frac{1}{y}$) or logarithmic $[\log_{10}(x)$ or $\log_{10}(y)]$ functions
   (b) stretching x- or y-values using the square function ($x^2$ or $y^2$).
Step 4. Plot the transformed data and its least-squares regression line. Examine the residuals or correlation coefficient, $r$, to see if it is a better fit.
Step 5. Repeat steps 2 to 4 for all appropriate transformations.

**EXERCISE 3F**

**Transforming to linearity**

1. **WE10** Apply a parabolic ($x^2$) transformation to the following data set. The regression line has been determined as $y = -27.7x + 186$ with $r = -0.91$.

<table>
<thead>
<tr>
<th>x</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>7</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>96</td>
<td>95</td>
<td>92</td>
<td>90</td>
<td>14</td>
<td>-100</td>
</tr>
</tbody>
</table>

2. **WE11** The average heights of 50 girls of various ages were measured as follows.

<table>
<thead>
<tr>
<th>Age group (years)</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average height (cm)</td>
<td>128</td>
<td>144</td>
<td>148</td>
<td>154</td>
<td>158</td>
<td>161</td>
<td>165</td>
<td>164</td>
<td>166</td>
<td>167</td>
</tr>
</tbody>
</table>

The original linear regression yielded:

Height = 3.76 $\times$ age $+$ 104.7, with $r = 0.92$.

a) Plot the original data and regression line.
b) Transform using the $\log_{10}(x)$ transformation.
c) Perform regression analysis on the transformed data and comment on your results.

3. a) Use the transformed data from question 2 to predict the heights of girls of the following ages:
   i) 7 years old
   ii) 10.5 years old
   iii) 20 years old.
b) Identify the interpolations from these predictions.

4. Comment on the suitability of transforming the data of question 2 in order to improve predictions for girls under 8 or over 18.

5. a) **WE12** Apply a reciprocal transformation to the following data obtained by a physics student studying light intensity.

<table>
<thead>
<tr>
<th>Distance from light source (metres)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intensity (candlepower)</td>
<td>90</td>
<td>60</td>
<td>28</td>
<td>22</td>
<td>20</td>
<td>12</td>
</tr>
</tbody>
</table>

b) Use the transformed regression equation to predict the intensity at a distance of 20 metres.
6 For each of the following scatterplots suggest an appropriate transformation(s).

7 Use the equation \( y = 0.2x^2 - 12.5 \), found after transformation, to predict values of \( y \) for the given \( x \)-value (correct to 2 decimal places):
   - \( a \) \( x = 2.5 \)
   - \( b \) \( x = 2.5 \)

8 Use the equation \( y = 1.12 \log_{10}(x) - 25 \), found after transformation, to predict values of \( y \) for the given \( x \)-value (correct to 2 decimal places):
   - \( a \) \( x = 2.5 \)
   - \( b \) \( x = 2.5 \)
   - \( c \) \( x = 0 \)

9 Use the equation \( \log_{10}(y) = 0.2x + 0.03 \), found after transformation, to predict values of \( y \) for the given \( x \)-value (correct to 2 decimal places):
   - \( a \) \( x = 2.5 \)
   - \( b \) \( x = 2.5 \)

10 Use the equation \( \frac{1}{y} = 0.2x - 12.5 \), found after transformation, to predict values of \( y \) for the given \( x \)-value (correct to 2 decimal places):
   - \( a \) \( x = 2.5 \)
   - \( b \) \( x = 2.5 \)

11 The seeds in the sunflower are arranged in spirals for a compact head. Counting the number of seeds in the successive circles starting from the centre and moving outwards, the following number of seeds were counted.

<table>
<thead>
<tr>
<th>Circle</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
<th>5th</th>
<th>6th</th>
<th>7th</th>
<th>8th</th>
<th>9th</th>
<th>10th</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of seeds</td>
<td>3</td>
<td>5</td>
<td>8</td>
<td>13</td>
<td>21</td>
<td>34</td>
<td>55</td>
<td>89</td>
<td>144</td>
<td>233</td>
</tr>
</tbody>
</table>

- a Fit a least-squares regression line and plot the data.
- b Find the correlation coefficient. What does this mean?
- c Using the regression line for the original data, predict the number of seeds in the 11th circle.
- d Find the residuals.
- e Plot the residuals on a separate graph. Are the data linear?
- f What type of transformation could be applied to:
  - i the \( x \)-values? Explain why.
  - ii the \( y \)-values? Explain why.

12 Apply the \( \log_{10}(y) \) transformation to the data used in question 11.
   - a Fit a least-squares regression line to the transformed data and plot it with the data.
   - b Find the correlation coefficient. Is there an improvement? Why?
   - c Complete the least-squares regression for the transformation.
   - d Calculate the coefficient of determination. How does this explain variation?
   - e Using the regression line for the transformed data, predict the number of seeds for the 11th circle.
   - f How does this compare with the prediction from question 11?
**SUMMARY**

### Fitting a straight line by eye
- Make sure there are an equal number of points above/below the fitted line.

### Fitting a straight line — the 3-median method
- Assuming data points are in order of increasing $x$-values:
  - Step 1: Divide data points into 3 groups.
  - Step 2: Adjust for ‘unequal’ groups: if there is 1 extra point, put it in the middle; if there are 2 extra points, put them in the end groups.
  - Step 3: Calculate the medians for the 3 groups $(x_L, y_L)$, $(x_M, y_M)$, $(x_U, y_U)$.
- For a graphical approach:
  - Step 4: Place a ruler through the two ‘outer’ medians and move the ruler one-third of the way towards the middle median.
  - Step 5: Calculate the $y$-intercept and the gradient and use these to find the equation of the regression line.
- For an arithmetic approach:
  - Step 4: Calculate the gradient using the formula: $m = \frac{y_U - y_L}{x_U - x_L}$.
  - Step 5: Calculate the $y$-intercept using the formula:
    $$c = \frac{1}{3}(y_L + y_M + y_U) - m(x_L + x_M + x_U)$$
  - Step 6: Substitute $m$ and $c$ into the equation $y = mx + c$.

### Fitting a straight line — least-squares regression
- Use a CAS calculator to find the equation of the least-squares regression line.
  - (a) Enter the data in the Statistics screen.
  - (b) Draw the scatterplot by tapping **SetGraph** and completing the fields.
  - (c) To fit the regression line ($y = ax + b$), tap:
    - **Calc**
    - **Linear Reg**
  - Record the values of $a$, $b$ and $r$ for the linear regression equation.
- To find the equation of the least-squares regression line ‘by hand’:
  - (a) The summary data needed are:
    - (i) $\bar{x}$ and $s_x$ the mean and standard deviation of the independent variable
    - (ii) $\bar{y}$ and $s_y$ the mean and standard deviation of the dependent variable
    - (iii) $r$ Pearson’s product–moment correlation coefficient.
  - (b) The formulas to use are:
    - (i) $m = r \frac{s_y}{s_x}$
    - (ii) $c = \bar{y} - m\bar{x}$
  - where $m$ is the slope of the regression line and $c$ is the $y$-intercept.
  - Alternatively, if the general form of the regression line is given as $y = a + bx$, then $b = r \frac{s_y}{s_x}$ and $a = \bar{y} - b\bar{x}$.

### Interpretation, interpolation and extrapolation
- The slope ($m$) of the regression line $y = mx + c$ indicates the rate at which the data are increasing or decreasing.
- The $y$-intercept, $c$, indicates the approximate value of the data when $x = 0$.
- Interpolation is the use of the regression line to predict values ‘between’ two values already in the data set.
- Extrapolation is the use of the regression line to predict values smaller than the smallest value already in the data set or larger than the largest value.
Residual analysis

- Calculate predicted values \((y_{pred})\) from the regression equation \((y = mx + c)\) for all values of \(x\).
- Calculate residuals \((y - y_{pred})\) for all values of \(x\) (actual values – predicted values).
- Observe the data and plot the residuals.

Transforming to linearity

- Transform non-linear data to linearity by using one or more of the following possible transformations.
  Compressing axis: \(y\) versus \(\log_{10}(x)\) \(y\) versus \(\frac{1}{x}\) \(\log_{10}(y)\) versus \(x\) \(\frac{1}{y}\) versus \(x\)
  Stretching axis: \(y\) versus \(x^2\) \(y^2\) versus \(x\)

Quadratic transformations

1. Use \(y\) versus \(x^2\) transformation.
2. Use \(y\) versus \(x^2\) transformation.
3. Use \(y^2\) versus \(x\) transformation.
4. Use \(y^2\) versus \(x\) transformation.

Logarithmic and reciprocal transformations

1. Use \(y\) versus \(\log_{10}(x)\) or \(y\) versus \(\frac{1}{x}\) transformation.
2. Use \(y\) versus \(\log_{10}(x)\) or \(y\) versus \(\frac{1}{x}\) transformation.
3. Use \(\log_{10}(y)\) versus \(x\) or \(\frac{1}{y}\) versus \(x\) transformation.
4. Use \(\log_{10}(y)\) versus \(x\) or \(\frac{1}{y}\) versus \(x\) transformation.
**MULTIPLE CHOICE**

Use the figure below to answer questions 1 and 2.

1. The most appropriate line of best fit for the figure is:
   - A
   - B
   - C
   - D
   - E

2. The gradient of the 3-median regression line is:
   - A
   - B
   - C
   - D
   - E

3. In using the 3-median method for 34 points, the number of points placed in each group is:
   - A 10, 14, 10
   - B 11, 12, 11
   - C 12, 10, 12
   - D 10, 12, 14
   - E dependent on the decision of the person doing the calculations

4. The correlation between two variables $x$ and $y$ is $-0.88$. Which of the following statements is true?
   - A As $y$ increases it causes $x$ to increase.
   - B As $y$ increases it causes $x$ to decrease.
   - C There is a poor fit between $x$ and $y$.
   - D As $x$ increases, $y$ tends to increase.
   - E As $x$ increases, $y$ tends to decrease.

5. When calculating a least-squares regression line, a correlation coefficient of $-1$ indicates that:
   - A the $y$-axis variable depends linearly on the $x$-axis variable
   - B the $y$-axis variable increases as the $x$-axis variable decreases
   - C the $y$-axis variable decreases as the $x$-axis variable decreases
   - D all the data lie on the same straight line
   - E the two variables depend upon each other

6. For the following data set
   
<table>
<thead>
<tr>
<th>$x$</th>
<th>25</th>
<th>36</th>
<th>45</th>
<th>78</th>
<th>89</th>
<th>99</th>
<th>110</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>78</td>
<td>153</td>
<td>267</td>
<td>456</td>
<td>891</td>
<td>1020</td>
<td>1410</td>
</tr>
</tbody>
</table>

   the coefficient of determination (to 2 decimal places) is closest to:
   - A 14.14
   - B $-381.97$
   - C 0.91
   - D 0.95
   - E 0.94

7. Given the following summary details
   
   $\bar{x} = 154.4$  
   $s_x = 5.8$  
   $\bar{y} = 172.5$  
   $s_y = 7.4$  
   $r = 0.9$

   the values of $m$ and $c$, respectively, for the equation of the regression line $y = mx + c$ are:
   - A 0.71 and 32.72
   - B 1.15 and $-4.79$
   - C 0.44 and 10.1
   - D 0.04 and $-0.16$
   - E $-1.32$ and 3.8
8 A 3-median regression fit yielded the equation $y = 4.3x - 2.4$. The value of $y$ when $x = 4.4$ is:
A 21.32  
B 18.92  
C 16.52  
D 1.58  
E -2.4

9 A least-squares regression is fitted to the 7 points as shown.

The plot of residuals would look most similar to:
A
B
C
D
E

10 The lengths and diameters (in mm) of a sample of jellyfish selected from another location were recorded and displayed in the following scatterplot. The least-squares regression line for these data is shown.

The equation of the least-squares regression line is $length = 3.5 + 0.87 \times diameter$.

The correlation coefficient is $r = 0.9034$.

From the equation of the least-squares regression line, it can be concluded that for these jellyfish, on average:
A there is a 3.5 mm increase in diameter for each 1 mm increase in length  
B there is a 3.5 mm increase in length for each 1 mm increase in diameter  
C there is a 0.87 mm increase in diameter for each 1 mm increase in length  
D there is a 0.87 mm increase in length for each 1 mm increase in diameter  
E there is a 4.37 mm increase in diameter for each 1 mm increase in length

[VCAA 2007]

11 A student uses the following data to construct the scatterplot shown below.

To make the scatterplot linear, she applies a log ($y$) transformation; that is, a log transformation is applied to the $y$-axis scale. She then fits a least-squares regression line to the transformed data.
With \( x \) as the independent variable, the equation of this least-squares regression line is closest to:

A \( \log (y) = 0.217 + 88.0x \)
B \( \log (y) = 3.8 + 4.4x \)
C \( \log (y) = 3.1 + 0.008x \)
D \( \log (y) = 0.88 + 0.23x \)
E \( \log (y) = 1.58 + 0.002x \)  

[VA 2006]

12 After a transformation, a relationship was found to be \( y = 0.4x^2 + 12.1 \). The predicted value for \( y \) given that \( x = 2.5 \) is:

A 6.25  B 2.5  C 14.6  D 13.1  E 12.5

**SHORT ANSWER**

1 Find the equation of the line passing through the point (5, 7.5) with a gradient of \(-3.5\).

2 Fit a 3-median line to the following data.

<table>
<thead>
<tr>
<th>( x )</th>
<th>( y )</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>16</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
</tr>
<tr>
<td>7</td>
<td>18</td>
</tr>
<tr>
<td>9</td>
<td>15</td>
</tr>
<tr>
<td>11</td>
<td>13</td>
</tr>
<tr>
<td>13</td>
<td>9</td>
</tr>
<tr>
<td>15</td>
<td>5</td>
</tr>
<tr>
<td>17</td>
<td>5</td>
</tr>
<tr>
<td>20</td>
<td>15</td>
</tr>
</tbody>
</table>

Express the equation with exact values of \( m \) and \( c \).

3 Find the equation of the 3-median regression line for the following data set.

<table>
<thead>
<tr>
<th>( x )-values</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y )-values</td>
<td>23</td>
<td>21</td>
<td>20</td>
<td>14</td>
<td>16</td>
<td>9</td>
<td>12</td>
<td>5</td>
</tr>
</tbody>
</table>

**EXTENDED RESPONSE**

**Task 1**

1 Consider this data set which measures the sales figures for a new salesperson.

<table>
<thead>
<tr>
<th>Day</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Units sold</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>9</td>
<td>20</td>
<td>44</td>
<td>84</td>
<td>124</td>
</tr>
</tbody>
</table>

The least-squares regression yielded the following equation:

\[
\text{Units sold} = 16.7 \times \text{day} - 39.1
\]

The correlation coefficient was 0.90.

a Do the data exhibit any pattern, and if so what pattern?

b Comment on using the regression line to predict for small values of the independent variable.

c Use the regression line to predict the sales figures for the 10th day.

2 Transform the data from question 1 using a parabolic (\( x^2 \)) transformation.

3 Perform a least-squares regression on the transformed data from question 2.

4 Use the regression line for the transformed data to predict the sales figures for the 10th day. Is this a better prediction than the one found in 1c?
Task 2

1 A mining company wishes to predict its gold production output. It collected the following data over a 9-month period.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Production (tonnes)</td>
<td>3</td>
<td>8</td>
<td>10.8</td>
<td>12</td>
<td>11.6</td>
<td>14</td>
<td>15.5</td>
<td>15</td>
<td>18.1</td>
</tr>
</tbody>
</table>

a Plot the data and fit a line, by eye, using the ‘equal number of points’ method.
b State the equation of this line.
c Fit a straight line to the original data using the 3-median method, stating the equation of this line.
d Now, use the least-squares regression to find another equation for this line.
e Using the line from part d, predict the production after 12 months.
f Comment on the accuracy, usefulness and simplicity of the methods.

2 Let us pursue further the data from question 1 above.
a Looking at the original data set, discuss whether linearity is a reasonable assertion.
b Research into goldmines has indicated that after about 10 months, production tends not to increase as rapidly as in earlier months. Given this information, a logarithmic transformation is suggested. Transform the original data using this method.
c Fit a straight line to this transformed data using least-squares regression.
d Discuss whether or not this transformation has removed any non-linearity.
e Compare the prediction from question 1e above with the one obtained using the logarithmic transformation.

Task 3

1 The heights, in cm, and ages, in months, of a random sample of 15 boys have been plotted in the scatterplot below. The least-squares regression line has been fitted to the data.

The equation of the least-squares regression line is \( height = 75.4 + 0.53 \times age \).
The correlation coefficient is \( r = 0.7541 \).
a Complete the following sentence:
On average, the height of a boy increases by __________ cm for each one-month increase in age.
b i Evaluate the coefficient of determination. Write your answer, as a percentage, correct to 1 decimal place.

ii Interpret the coefficient of determination in terms of the variables height and age.

[VCAA 2006]

2 The heights, in cm, and ages, in months, of the 15 boys are shown in the scatterplot below.

\[\begin{array}{c}
\text{Age (months)} \\
0 & 15 & 20 & 25 & 30 & 35 & 40 \\
\text{Height (cm)} \\
82 & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\
84 & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\
86 & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\
88 & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\
90 & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\
92 & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\
94 & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\
96 & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\
\end{array}\]

a Fit a 3-median line to the scatterplot. Circle the three points you used to determine this 3-median line.

b Determine the equation of the 3-median line. Write the equation in terms of the variables height and age and give the slope and intercept correct to 1 decimal place.

c Explain why the 3-median line might model the relationship between height and age better than the least-squares regression line.

[VCAA 2006]

[EXAM TIP] Many students did not convert the decimal into a percentage as required and gave an answer of 0.6. Some who did the conversion rounded the decimal first before converting to an incorrect answer of 60%.

[Assessment report 2 2006]

[EXAM TIP] Common incorrect interpretations referred to causation. A common incorrect answer was ‘56.9% of the variation in height is due to the variation in age.’

[Assessment report 2 2006]

[EXAM TIP] Many students did not find the required three points. If they had drawn a suitable line through their two end points and then done a one-third shift toward their middle point, a method mark may have been gained.

[Assessment report 2 2006]
**Chapter opener**

Digital doc
- 10 Quick Questions: Warm up with a quick quiz on introduction to regression. *(page 105)*

**3B Fitting a straight line — the 3-median method**

Digital docs
- SkillSHEET 3.1: Practise finding the median. *(page 113)*
- SkillSHEET 3.2: Practise calculating the gradient (I). *(page 113)*
- SkillSHEET 3.3: Practise finding the equation of a straight line. *(page 113)*

**3C Fitting a straight line — least-squares regression**

Digital docs
- Spreadsheet 063: Investigate least-squares regression. *(page 121)*
- WorkSHEET 3.1: Fitting a straight line by eye and using the 3-median regression line. *(page 123)*

Tutorial
- **WE4** int-0436: Learn how to find the equation of the least-squares regression line using $r$, $s_x$, and $s_y$. *(page 120)*

**3D Interpretation, interpolation and extrapolation**

Digital docs
- Spreadsheet 057: Investigate interpolation and extrapolation on a scatterplot. *(page 127)*
- SkillSHEET 3.4: Practise using the regression line to make predictions. *(page 128)*

**3E Residual analysis**

Digital doc
- WorkSHEET 3.2: Fitting a line by using the equal-number-of-points method, the 3-median method, calculate $r$, calculate residuals and make predictions using interpolation and extrapolation. *(page 136)*

Tutorial
- **WE8** int-0842: Learn how to find residuals using a CAS calculator. *(page 130)*

**3F Transforming to linearity**

Digital doc
- Spreadsheet 133: Investigate different transformations to linearity. *(page 143)*

Tutorials
- **WE10** int-0438: Watch a tutorial on applying a parabolic transformation to data using a CAS calculator. *(page 138)*
- **WE11** int-0843: Watch a worked example on applying a logarithmic transformation to data using a CAS calculator. *(page 139)*
- **WE12** int-0844: Watch a worked example on applying a reciprocal transformation to data using a CAS calculator. *(page 141)*

Interactivity int-0184
- Transforming to linearity: Use the interactivity to consolidate your understanding of applying appropriate transformations to achieve linearity. *(page 136)*

**Chapter review**

Digital doc
- Test Yourself: Take the end-of-chapter test to test your progress. *(page 151)*

To access eBookPLUS activities, log on to