Introduction to regression

3

VCE coverage
Area of study
Units 3 & 4 • Data analysis

In this chapter
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
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.

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. 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’.

Methods of fitting lines by eye

There are many different methods of fitting a straight line by eye. They may appear logical or even obvious but fitting by eye involves a considerable margin of error.

There are two basic methods of fitting a line by eye:

1. balancing the number of points
2. balancing errors.
Method 1: Balancing the number of points

The first method is to fit a line so that there is 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.

**WORKED Example 1**

Fit a straight line to the data in the figure 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.

**Method 2: Balancing errors**

The second method, balancing the errors, is more suitable when the points are more scattered and fitting a straight line does not appear as obvious. Errors are the *vertical distances* between the points and the fitted straight line. The errors can be marked on a scatterplot using *error bars* (as shown in worked example 2).

The aim is to balance the sum of the error bars above the line with the sum of the error bars below the line. In effect, this means that if all the error bars above the line were joined end to end they would be of equal length to the joined error bars from below the line.

Try to balance the sum of errors of the points above the line and the points below the line.
A hospital administrator wants to know if the time that patients spend in the operating theatre is related to the time they take to recover before they are allowed to be discharged. He sketches the graph at right, and decides to fit a straight line to the data using the equal-errors method.

Note that in this example, the points are somewhat scattered. There is no obvious best straight line; there may even be no linear relationship at all.

**THINK**

1. Fit a straight line which seems to best fit the data and mark in the error bars.

2. It appears that the first attempt involves greater error above the line (red) than below the line (blue). If the red error bars were placed end to end they would exceed the combined lengths of the blue bars. To rectify this the line will need to be raised.

*Note:* The raised line has only 2 points below the line, yet there are 4 above it. This line is still preferable because the errors are more balanced.

**DRAW**

**remember**

To fit a straight line to bivariate data by eye there are two methods.

Method 1: Balance by ensuring an equal number of points above/below the fitted line.

Method 2: Balance by ensuring an equal sum of errors above/below the fitted line.
Fitting a straight line by eye

The questions below 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.

   ![Scatterplot a](image1)
   ![Scatterplot b](image2)
   ![Scatterplot c](image3)

2. Fit a straight line to this data using the equal-errors method.

   ![Scatterplot a](image4)
   ![Scatterplot b](image5)
   ![Scatterplot c](image6)

3. Fit a straight line to the data in the scatterplots below using the most appropriate of the two methods: equal-number-of-points or equal-errors.

   ![Scatterplot a](image7)
   ![Scatterplot b](image8)
   ![Scatterplot c](image9)
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 \( y = mx + b \) where \( m \) is the gradient and \( b \) 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 and 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 scatter diagram. 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, two different approaches may be taken from here: graphical and arithmetic.
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 \( y = mx + b \)). Draw on your knowledge of finding equations of lines to find the equation of the line drawn on the scatterplot. There are two general methods.
(a) Method A: Choose two points which lie on the line and use these 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 the gradient of the line. 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 \((b)\) of the line.
Use the rule: \( b = \frac{1}{2} \left( (y_L + y_M + y_U) - m(x_L + x_M + x_U) \right) \)

Thus, the equation of the regression line is \( y = mx + b \).

WORKED Example 3

Find the equation of the regression line for the data in the table at right using the 3-median method.

<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.

We now have the option of following either the graphical or the arithmetic approach. In this worked example, we shall also employ a method using the graphics calculator.

Continued over page
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.

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.

2. Using the arithmetic approach

3. Find the gradient using the formula.

\[
m = \frac{y_U - y_L}{x_U - x_L}
\]

\[
= \frac{5.5 - 2}{6 - 1.5}
\]

\[
= \frac{3.5}{4.5}
\]

\[
= \frac{7}{9}
\]

\[
\approx 0.78
\]

4. Find the $y$-intercept using the formula.

\[
b = \frac{1}{3}[(y_L + y_M + y_U) - m(x_L + x_M + x_U)]
\]

\[
= \frac{1}{3}[(2 + 4 + 5.5) - \frac{7}{9} (1.5 + 3.5 + 6)]
\]

\[
= \frac{1}{3}[11.5 - \frac{7}{9} (11)]
\]

\[
= \frac{1}{3}[11.5 - 8.555]
\]

\[
b \approx 0.98
\]

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.

A third method of obtaining the equation of the line using the 3-median method is available using a graphics calculator.
THINK

WRITE/DISPLAY

3. Using a graphics calculator

1. Turn off any existing statistical plots.
   Press \( \text{2nd} \) [STAT PLOT].
   Select \( 4: \text{PlotsOff} \).
   Press \( \text{ENTER} \).

2. Enter the data.
   Press \( \text{STAT} \).
   Select \( 1: \text{Edit} \).
   Press \( \text{ENTER} \).
   Enter the \( x \)-values in \( L1 \).
   Enter the \( y \)-values in \( L2 \).

3. Set up a scatterplot as follows.
   Set \( \text{WINDOW} \) settings as shown.
   Press \( \text{2nd} \) [STAT PLOT].
   Select \( 1: \text{Plot 1} \).
   Press \( \text{ENTER} \).
   Select \( \text{On} \), and the scatterplot in the \( \text{Type:} \) menu.
   Enter \( L1 \) for the \( Xlist: \) and \( L2 \) for the \( Ylist: \).
   Select the square in \( \text{Mark:} \).

4. Find the gradient and \( y \)-intercept of the regression line.
   Press \( \text{STAT} \).
   Select \( \text{CALC} \) and \( 3: \text{Med-Med} \).
   Press \( \text{ENTER} \).
   Or, after \( \text{Med-Med} \) appears, type \( L1, L2, Y1 \).
   (To type \( Y1 \), press \( \text{2nd} \) \( \text{VARS} \) and select \( \text{Y-VARS} \) and \( 1: \text{Function} \) then \( 1:Y1 \).) Press \( \text{ENTER} \). This allows the graph of the line to be seen as it enters the regression equation into \( Y1 \).
   Hence, the 3-median regression line equation is:

5. To view the data and regression line, press \( \text{GRAPH} \).
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 them in the outer groups.
   Step 3: Calculate the medians for the 3 groups $(x_L, y_L), (x_M, y_M), (x_U, y_U)$.

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:
   $$b = \frac{1}{2}[(y_L + y_M + y_U) - m(x_L + x_M + x_U)]$$

**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. a Using the data in the table below, find the regression line using the 3-median method.
   b Using this regression line determine:
   i the value of $y$ when $x = 25$
   ii the value of $x$ when $y = 65$
   iii the $x$-intercept of the regression line.

<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>
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.

<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>

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 (°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>

**multiple choice**

4. 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

**multiple choice**

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

<table>
<thead>
<tr>
<th>x</th>
<th>23</th>
<th>25</th>
<th>27</th>
<th>31</th>
<th>35</th>
<th>37</th>
<th>42</th>
<th>48</th>
<th>51</th>
<th>55</th>
<th>56</th>
<th>61</th>
<th>64</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>32</td>
<td>34</td>
<td>38</td>
<td>43</td>
<td>48</td>
<td>44</td>
<td>47</td>
<td>51</td>
<td>53</td>
<td>21</td>
<td>22</td>
<td>50</td>
<td>53</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Month</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</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>

a. Find the equation of the 3-median regression line.

b. According to this regression, when will the company have no sales? Discuss whether this prediction is reasonable.
8. 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 ('000)</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. Find the 3-median regression line.

b. Predict the employment rate for a town of size 300 (that is, 300 000).

c. According to this regression, what sized town will have no unemployment?

d. Discuss the merit of producing such a regression.

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

| x  | 0  | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| y  | 3  | 5  | 6  | 9  | 11 | 16 | 15 | 13 | 19 | 22 | 26 | 24 | 28 | 31 | 30 | 32 | 36 | 29 | 39 | 40 | 44 |

a. Plot the data as a scatter diagram.

b. Find the equation of the 3-median regression line.
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 from an earlier section, reproduced here.

We wish to minimise the total of the vertical lines, or ‘errors’ in some way. For example, in the earlier section we minimised the ‘sum’, 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 graphics calculator. The arithmetic background to its calculation is shown here for interest.

The least-squares formulas

Like other regression methods, we assume we have a set of \((x, y)\) values.

The number of such values is \(n\).

Let \(\bar{x}, \bar{y}\) be the averages (means) of the \(x-\) and \(y\)-values.

Recall from an earlier chapter the formulas for variances \((s_x^2)\) and covariance \((s_{xy})\) of bivariate data:

\[
s_x^2 = \frac{\sum (x - \bar{x})^2}{n - 1}
\]

or

\[
s_x^2 = \frac{\sum xy - n\bar{x}\bar{y}}{n - 1}
\]

\[
s_{xy} = \frac{\sum (x - \bar{x})(y - \bar{y})}{n - 1}
\]

or

\[
s_{xy} = \frac{\sum x^2 - n\bar{x}^2}{n - 1}
\]

Note that the summations (\(\sum\)) are over all points in the data set. It is beyond the scope of Further Mathematics to derive the formulas for the slope and intercept of the least-squares regression line, so they are simply stated.

The slope of the regression line \(y = mx + b\) is \(m = \frac{s_{xy}}{s_x^2}\)

\[
m = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}
\]

or

\[
m = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2}
\]

The \(y\)-intercept of the regression line is: \(b = \bar{y} - m\bar{x}\)
Calculating the least-squares regression line by hand

Let us calculate an equation using the least-squares method ‘by hand’, to see how it can be set out.

Consider the following data set.

<table>
<thead>
<tr>
<th>x</th>
<th>5</th>
<th>8</th>
<th>9</th>
<th>11</th>
<th>12</th>
<th>15</th>
<th>17</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>6</td>
</tr>
</tbody>
</table>

First, calculate the means, $\bar{x}$ and $\bar{y}$:

$$\bar{x} = \frac{77}{7} = 11$$
$$\bar{y} = \frac{21}{7} = 3$$

These means are used to calculate ‘deviations’ from the mean (see italic numbers in columns 3 and 5 below).

Given the number of calculations to be done, a table (as set out below) is useful.

Columns 4 and 6 contain the squares of columns 3 and 5 (the deviations from the means) respectively.

Column 7 contains the product of the numbers in columns 3 and 5.

The sum of each column is in bold.

Note that columns 3 and 5 must add up to 0. This acts as a check of your preliminary calculations.

**Calculation of regression line from table**

Recall that the gradient is given by the formula:

$$m = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

The numerator is the total in column 7 of the previous table.

The denominator is the total in column 4, so

$$m = \frac{34}{102}$$
$$= \frac{1}{3}$$

The y-intercept is calculated from the formula:

$$b = \bar{y} - m\bar{x}$$
Substituting previous calculations of $x$, $y$ and $m$:

$$b = 3 - \frac{1}{3} \times 11$$

$$= \frac{2}{3}$$

Thus the equation of the regression line is: $y = \frac{1}{3}x - \frac{2}{3}$

**WORKED Example 4**

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

<table>
<thead>
<tr>
<th>$x$</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>$y$</td>
<td>11</td>
<td>9</td>
<td>10</td>
<td>8</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Using a graphics calculator

**THINK**

1. Turn off any existing statistical plots.
   - Press **2nd** [STAT PLOT].
   - Select 4:PlotsOff.
   - Press **ENTER**.

2. Enter the data.
   - Press **STAT**.
   - Select 1:Edit.
   - Press **ENTER**.
   - Enter the $x$-values in L1 and the $y$-values in L2.

3. Set up a stat plot as follows.
   - Set WINDOW as shown opposite.
   - Select 2nd [STAT PLOT].
   - Select 1:Plot 1.
   - Press **ENTER**.
   - Select On, and the scatterplot in the Type: menu.

Enter L1 for the Xlist: and L2 for the Ylist.
Select the square in Mark:. 

Continued over page
4 Find the gradient and $y$-intercept of the least-squares regression line.

Press (STAT).
Select CALC and 4:LinReg(ax+b).
After LinReg(ax+b) appears, type L1, L2, Y1.
(To type Y1, press VARS and select Y-VARS and 1:Function then 1:Y1. Press ENTER.)
Write the least-squares regression.

Using arithmetic

1 Calculate the means.

\[ \bar{x} = \frac{66}{8} \quad \bar{y} = \frac{52}{8} \]
\[ = 8.25 \quad = 6.5 \]

2 Draw up the required table and complete all calculations.

<table>
<thead>
<tr>
<th>$x$</th>
<th>$y$</th>
<th>$(x - \bar{x})(y - \bar{y})$</th>
<th>$(x - \bar{x})^2$</th>
<th>$(y - \bar{y})^2$</th>
<th>$(x - \bar{x})(y - \bar{y})^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11</td>
<td>-7.25</td>
<td>56.625</td>
<td>4.5</td>
<td>-32.625</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>-5.25</td>
<td>27.5625</td>
<td>2.5</td>
<td>-13.235</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>-4.25</td>
<td>18.0625</td>
<td>3.5</td>
<td>-14.875</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>-1.25</td>
<td>1.5625</td>
<td>-0.5</td>
<td>0.25</td>
</tr>
<tr>
<td>10</td>
<td>8</td>
<td>1.75</td>
<td>3.0625</td>
<td>1.5</td>
<td>2.625</td>
</tr>
<tr>
<td>12</td>
<td>4</td>
<td>3.75</td>
<td>14.0625</td>
<td>-2.5</td>
<td>9.375</td>
</tr>
<tr>
<td>14</td>
<td>3</td>
<td>5.75</td>
<td>33.0625</td>
<td>-3.5</td>
<td>-20.125</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>6.75</td>
<td>45.5625</td>
<td>-5.5</td>
<td>-37.125</td>
</tr>
<tr>
<td>66</td>
<td>52</td>
<td>0</td>
<td>195.5</td>
<td>0</td>
<td>-124</td>
</tr>
</tbody>
</table>

Total each column.

Note: Sum of columns 3 and 5 is 0.

3 Use the formula to calculate the gradient by substituting values derived from the table.

\[ m = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2} \]
\[ = \frac{-124}{195.5} \]
\[ = -0.6343. \]

4 Use the formula to calculate the $y$-intercept.

\[ b = \bar{y} - m\bar{x} \]
\[ = 6.5 - (-0.6343)8.25 \]
\[ = 11.733 \]

5 Write the equation of the least-squares linear regression line.

\[ y = -0.634x + 11.733 \]
The TI-83 may be set up to display \( r \) and \( r^2 \) with regression line information as follows.

Choose **CATALOG** and **DiagnosticOn**, and press **ENTER**.

The result, using the previous example, is shown below.

The value of \( r \), Pearson's correlation coefficient, can also be found using a graphics calculator by pressing **VARS** and selecting **5:Statistics**, **EQ** and \( r \). 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 graphics calculator is a far more efficient method.

---

**Remember**

1. Use a graphics calculator to find the equation of the least-squares regression line.
2. To find the least-squares regression line 'by hand':
   (a) set up columns of \( x \)- and \( y \)-values
   (b) calculate means (\( \bar{x}, \bar{y} \))
   (c) calculate deviations (\( x - \bar{x} \), \( y - \bar{y} \))
   (d) calculate the square of these deviations (\( (x - \bar{x})^2 \), \( (y - \bar{y})^2 \), and the product, \( (x - \bar{x})(y - \bar{y}) \))
   (e) add up the numbers in each of these columns
   (f) use the formula to calculate the gradient, \( m = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2} \)
   (g) calculate the \( y \)-intercept from the formula, \( b = \bar{y} - m\bar{x} \).
1. Find the equation of the linear regression line for the following data set using the least-squares method.

<table>
<thead>
<tr>
<th>x</th>
<th>4</th>
<th>6</th>
<th>7</th>
<th>9</th>
<th>10</th>
<th>12</th>
<th>15</th>
<th>17</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>10</td>
<td>8</td>
<td>13</td>
<td>15</td>
<td>14</td>
<td>18</td>
<td>19</td>
<td>23</td>
</tr>
</tbody>
</table>

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

<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>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>35</td>
<td>28</td>
<td>22</td>
<td>16</td>
<td>19</td>
<td>14</td>
<td>9</td>
<td>7</td>
<td>2</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>x</th>
<th>-4</th>
<th>-2</th>
<th>-1</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>5</th>
<th>5</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>6</td>
<td>7</td>
<td>3</td>
<td>10</td>
<td>16</td>
<td>9</td>
<td>12</td>
<td>16</td>
<td>11</td>
<td>21</td>
</tr>
</tbody>
</table>

4. A software company has researched the volume of sales of a new operating system as a function of the selling price.
   a. Find the equation of the linear regression line for the following data set using the least-squares method.

<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 ('000)</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>

   b. What would be the volume of sales if the selling price is $95?
   c. Discuss how reasonable this result is, given the data in the table.

5. 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. Can you find 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?
Chapter 3 Introduction to regression

6. The best estimate of the least-squares regression line for the scatterplot below is:

\[ y = 2x \]  \hspace{1cm}  \text{A} \hspace{1cm}  y = \frac{1}{2}x + 2 \]  \hspace{1cm}  \text{C} \hspace{1cm}  y = \frac{1}{2}x - 2 \]  \hspace{1cm}  \text{D} \hspace{1cm}  y = \frac{1}{2}x - 1 \]  \hspace{1cm}  \text{E}

7. 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>Lifespan (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>

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

8. Least-squares regression is a popular technique in medical research. Consider the example of the testing of a new drug to reduce blood pressure for people with abnormally high (systolic) pressure. The dosage is varied and the blood pressure is recorded as demonstrated in the following data set.

<table>
<thead>
<tr>
<th>Drug dosage</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blood pressure</td>
<td>200</td>
<td>190</td>
<td>185</td>
<td>175</td>
<td>160</td>
<td>155</td>
</tr>
</tbody>
</table>

\( \text{a} \) Fit a least-squares regression line to the data.
\( \text{b} \) Comment on the reliability of this line to predict blood pressure given the dosage.

9. The price of eggs in China is compared with the price of petrol in Queensland. The following table contains the data for this comparison, recorded in 10 cities in China and 10 in Queensland.

<table>
<thead>
<tr>
<th>Eggs ($/dozen)</th>
<th>1.09</th>
<th>1.22</th>
<th>1.23</th>
<th>1.31</th>
<th>1.31</th>
<th>1.32</th>
<th>1.35</th>
<th>1.36</th>
<th>1.36</th>
<th>1.38</th>
</tr>
</thead>
<tbody>
<tr>
<td>Petrol ($/litre)</td>
<td>0.73</td>
<td>0.78</td>
<td>0.77</td>
<td>0.79</td>
<td>0.78</td>
<td>0.81</td>
<td>0.80</td>
<td>0.80</td>
<td>0.85</td>
<td>0.89</td>
</tr>
</tbody>
</table>

\( \text{a} \) Fit a least-squares regression line to the data.
\( \text{b} \) Which is likely to be the independent variable, eggs or petrol? Why?
\( \text{c} \) Comment on the reliability of your regression in predicting prices of eggs or petrol. (That is, consider whether the regression line you determined proves that egg and petrol prices are directly related.)
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.

**History of mathematics**

**CARL FRIEDRICH GAUSS (30.04.1777 – 23.02.1855)**

During his lifetime...
Captain Cook is killed in Hawaii.
Mozart and Haydn write symphonies.
Laughing gas is developed.
The Rosetta Stone is found.

Carl Gauss was a German mathematician, astronomer and physicist. He was born in Brunswick and was the son of a bricklayer. At the age of three Gauss taught himself how to work with numbers and while still a young child was able to correct his father’s calculations regarding his workers’ wages. At school, when a teacher asked the class to add the numbers 1 to 100 (the result is 5050), Gauss completed the sum before the teacher had even explained how to do it.

His parents had trouble paying for his university fees but fortunately the Duke of Brunswick heard about this gifted young man and paid for his education at the University of Gottingen. When he was nineteen years old Gauss wrote a paper stating that every algebraic equation had a solution. He developed this idea as part of his doctorate. He also worked on elliptic functions, and statistics and probability. From 1807 until his death he was the director of Gottingen Observatory. He was also in charge of the geodetic survey of Hanover from 1818 to 1825. Much of his work involved large amounts of routine calculation and he became interested in the theory of errors of observation. As a result he developed the method of least squares.

Gauss was also good at managing money. He read the financial sections of the newspapers and played the stock market. By the time he died he was a wealthy man.

**Questions**

1. What did Gauss invent with Weber?
2. What did his work on lenses help to correct?
3. What is a heliotrope used for?
4. Was Gauss a wealthy person?
5. What mathematical area did he work on which is related to regression?
Interpretation, interpolation and extrapolation

Interpreting slope and intercept ($m$ and $b$)
Once you have a least-squares 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 $x = 0$.

**WORKED Example 5**

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:

- the rate at which bacteria are growing
- 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>
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

\[ y = mx + b \]

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) and that there were actual data on that day.

**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 therefore, 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.

---

**WORKED Example 6**

**Interpolation**

Use the following data set to 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. Use a graphics calculator to obtain the regression line as described in worked example 4.

2. Press **GRAPH** to view the data and the regression line.

**WRITE/DISPLAY**

```
LinReg
y=ax+b
a=9.228571429
b=55.62857143
r=.9489181097
r= .9700000000
```

$y = 9.23x + 55.63$
THINK
3 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.

Alternatively, get the graphics calculator to do the work by calculating \( Y_1(8) \).

Notes:
1. \( Y_1 \) is under VARS, Y-VARS, 1:Function, 1:Y1
2. Since there was a good fit (\( r = 0.97 \)), then one can be confident of an accurate prediction.

WRITE/DISPLAY
Height = 9.23 \times \text{age} + 55.63
= 9.23 \times 8 + 55.63
= 129.5 \text{ cm}

WORKED Example 7
Extrapolation
Use 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 graphics calculator to find \( Y_1(15) \).
2 Analyse the result.

WRITE
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**

1. 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 were 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 rate at which the red blood cell count was changing

b. 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 ('000)</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-square 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. the number of visits per year by a person on a weekly income of $680

b. 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:
- the regression equation
- \( y \) when \( x = 3 \)
- \( y \) when \( x = 12 \)
- \( x \) when \( y = 7 \)
- \( x \) when \( y = 25 \).

Which of \( b-e \) 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 which 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>

- Find the least-squares regression equation for each factory.
- Which factory is likely to have the lowest cost to ship to a shop in Melbourne?
- Which factory is likely to have the lowest cost to ship to Mytown, 115 kilometres from Melbourne?
- Which factory has the most ‘linear’ shipping rates?

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.

- Find the cost to produce 200 calculators.
- How many calculators can be produced for $2000?
- Find the cost to produce 10 000 calculators.
- What are the ‘fixed’ costs for this production?
- Which of \( a-c \) 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. \( \text{Note: Only use } (x, y) \text{ pairs if both are in the table.} \)

<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: \( s = 18500 + 900n \).

- Find the salary for a woman who completed 10 years of schooling.
- Find the salary for a man who completed 12 years of schooling.
- Find the salary for a woman who completed 15 years of schooling.
- Mary earned $30 400. What was her likely schooling experience?
- 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. These are nothing more than the ‘error’ bars of the figure in worked example 2.

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, below. 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. Find 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. Find the equation of a least-squares regression line using a calculator or spreadsheet. Results are:
   (a) gradient ($m$) = 28.7
   (b) $y$-intercept ($b$) = −78.7
   (c) correlation ($r$) = 0.87.

2. Plot the data and regression line.
   Calculate the predicted $y$-values — these are labelled as $y_{\text{pred}}$.
   **Note:** In this example, despite the relatively good fit ($r = 0.87$) it is clear that the original data are not really linear. There seems to be a ‘pattern’ in the error bars. The first few are above the line, then there is a group below the line, and then a few above the line again.

3. Calculate residuals for each point using $y = 28.7x - 78.7$.

<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>Predicted $y$-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>Residuals ($y - y_{\text{pred}}$)</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>Predicted $y$-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>Residuals ($y - y_{\text{pred}}$)</td>
<td>−46.34</td>
<td>−45.02</td>
<td>−38.7</td>
<td>7.62</td>
<td>100.94</td>
</tr>
</tbody>
</table>

**WRITE**

So $y = 28.7x - 78.7$

### Notes:

1. The residuals may be determined by either $(y - y_{\text{pred}})$ or $(y_{\text{pred}} - y)$, as long as we use the *same* formula for all the points.

2. The *sum* of all the residuals *always* adds to 0 (or very close after rounding), when least-squares regression is used. This can act as a check on 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. 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.

WORKED Example 9

Using the same data as in worked example 8, plot the residuals and discuss the features of the residual plot.

THINK

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; they are not part of the residual plot.)

remember

1. Calculate predicted values ($y_{pred}$) from the regression equation ($y = mx + b$) for all values of x.
2. Calculate residuals ($y - y_{pred}$) for all values of x.
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.
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?
   - A. All of them
   - B. None of them
   - C. i and iii only
   - D. ii only
   - E. ii and iii only

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: \( m = -1.19, b = 16.34, r = -0.87 \).
   - a. Find the predicted defective rate (\( y_{pred} \)) based upon this regression line.
   - b. Find the residuals (\( y - y_{pred} \)).
   - 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>Day</th>
<th>Mon</th>
<th>Tues</th>
<th>Wed</th>
<th>Thurs</th>
<th>Fri</th>
<th>Sat</th>
<th>Sun</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: \( m = -22.5, b = 161.3, r = -0.94 \).
   - a. Find the predicted hotel bookings (\( y_{pred} \)) based upon this regression line.
   - b. Find the residuals (\( y - y_{pred} \)).
   - 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?
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, three transformations are studied:

1. Logarithmic transformation:  \( y \) versus \( \log_{10} x \)

2. Quadratic transformation:  \( y \) versus \( x^2 \)

3. Reciprocal transformation:  \( y \) versus \( \frac{1}{x} \)

Choosing the correct transformation

The best way to see which of the transformations to use is to look at a number of ‘data patterns’.

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

   ![Graph](image1)

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

   ![Graph](image2)

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

   ![Graph](image3)

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

   ![Graph](image4)

If the gradient is increasing in magnitude as \( x \) increases, an \( x^2 \) transformation may be appropriate.

If the gradient is decreasing in magnitude as \( x \) increases, a \( \log_{10} x \) or \( \frac{1}{x} \) transformation may be appropriate.

We shall look at one example of each transformation.
Quadratic transformation

Apply a parabolic transformation to the data from worked example 8, reproduced here. The regression line has been determined as \( y = 28.7x - 78.7 \) 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.

   Note: The data are ‘parabolic’ and the gradient is increasing in magnitude, so we need to apply an \( x^2 \) transformation to the data.

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 \((b) = -28.0\)
   (c) correlation \((r) = 0.95\).

4. Plot the new transformed data and regression line.

   Note:
   
   1. These data are still not truly linear, but are ‘less’ parabolic. Perhaps another \( x^2 \) transformation would improve things even further. This would involve squaring for a second time the \( x \)-values and applying another linear regression.

   2. See worked example 12 for a graphics calculator approach to transforming data.
Logarithmic transformation

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 $y = -6.97x + 93.2$, 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. Plot the data and the regression line to check that a logarithmic transformation is suitable.
   *Note:* Although the fit appears good ($r = -0.9$), the data clearly are of the same form as in worked example 10, with a gradient which is decreasing in magnitude; so a logarithmic transformation is appropriate.

2. Transform by taking logarithms of the $x$ data (rounded to 3 decimal places).

3. Find the equation of least-squares regression line using the transformed data. Using a calculator or spreadsheet:
   (a) gradient ($m$) = $-61.26$
   (b) $y$-intercept ($b$) = 97.14
   (c) correlation ($r$) = $-0.98$.

4. Plot the transformed data.
   *Notes:*
   1. Almost all the non-linearity has been removed from the data set and the correlation coefficient went from $-0.9$ to $-0.98$.
   2. See worked example 12 for a graphics calculator approach to transforming data.
   3. Logarithmic transformation cannot be applied if any of the $x$-values are negative. This is because the logarithmic function is defined for positive numbers only.

$y = -61.26x_T + 97.14$

where $x_T = \log_{10}x$. 
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**

Reciprocal transformation

\( a \) Using a graphics calculator, apply a reciprocal transformation to the following data.

\( b \) 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 (°)</th>
<th>( x )</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>Number of students in a class wearing jumpers</td>
<td>( y )</td>
<td>18</td>
<td>10</td>
<td>6</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

**THINK**

1. Enter the \( x \)-data in L1 and \( y \)-data in L2. Set up and view a STAT PLOT of the raw data.

2. Transform by calculating \( \frac{1}{x} \).

   Press (STAT), select EDIT and enter \( 1/L1 \) in the heading for L3.

3. Find the regression equation for the transformed data.

   To calculate the regression equation, use (STAT), CALC, 4:LinReg, and L3,L2,Y1.

   \[
   y = 94.583x - 0.4354, \text{ where } x = \frac{1}{x}.
   \]
THINK

4 Set up a STAT PLOT for the transformed data that uses L3 and L2 for the Xlist and Ylist respectively.

WRITE/DISPLAY

5 Plot the transformed data and the new regression equation by setting a suitable WINDOW and pressing GRAPH.

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

2 Use the transformed value, 0.08333, in the transformed regression equation to find y (the number of students wearing jumpers). Alternatively, use Y1(0.08333).

b If temperature \( x = 12 \)°C,
\[
x_r = \frac{1}{x} = \frac{1}{12} = 0.08333.
\]
\[
y = 94.583x_r - 0.4354 = 94.583 \times 0.08333 - 0.4354 = 7.447
\]
So 7 students are predicted to wear jumpers.

remember

To transform to linearity:
1. Calculate predicted values \( y_{pred} \) and residuals \( y - y_{pred} \) from the regression equation.
2. Observe data and plot residuals.
3. Transform either with \( y \) versus \( x^2 \) or \( y \) versus \( \log_{10} x \) depending on the ‘shape’ of the data set.
4. If the gradient is increasing in magnitude as \( x \) increases, an \( x^2 \) transformation is appropriate.
5. If the gradient is decreasing in magnitude as \( x \) increases, a \( \log_{10} x \) transformation is appropriate.
6. When predicting \( y \)-values using either the \( x^2 \) or \( \log_{10} x \) transformation, transform the original \( x \)-value first.
Transforming to linearity

1. 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. The average heights of 50 girls of various ages were measured as follows.

<table>
<thead>
<tr>
<th>Age group (years)</th>
<th>Average height (cm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>128</td>
</tr>
<tr>
<td>10</td>
<td>144</td>
</tr>
<tr>
<td>11</td>
<td>148</td>
</tr>
<tr>
<td>12</td>
<td>154</td>
</tr>
<tr>
<td>13</td>
<td>158</td>
</tr>
<tr>
<td>14</td>
<td>161</td>
</tr>
<tr>
<td>15</td>
<td>165</td>
</tr>
<tr>
<td>16</td>
<td>164</td>
</tr>
<tr>
<td>17</td>
<td>166</td>
</tr>
<tr>
<td>18</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} 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. 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.
Fitting straight lines by eye
• Method 1: Balance by having equal number of points above/below the fitted line.
• Method 2: Balance by having equal errors above/below the fitted line.

Fitting straight lines by 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:
  $b = \frac{1}{3}[(y_L + y_M + y_U) - m(x_L + x_M + x_U)]$

Fitting lines by least-squares regression
• Use a graphics calculator to find the equation of the least-squares regression line.
  To find the least-squares regression line by hand:
  1. set up columns of $x$- and $y$-values
  2. calculate means ($\bar{x}$, $\bar{y}$)
  3. calculate deviations $(x - \bar{x})$, $(y - \bar{y})$
  4. calculate the square of these deviations $(x - \bar{x})^2$, $(y - \bar{y})^2$, and the product,
     $(x - \bar{x})(y - \bar{y})$
  5. add up numbers in each of these columns
  6. use the formula to calculate the gradient, $m = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2}$
  7. calculate the $y$-intercept using the formula, $b = \bar{y} - m\bar{x}$.

Interpreting slope and intercepts and use of interpolation and extrapolation
• The slope ($m$) of the regression line $y = mx + b$ indicates the rate at which the
  data are increasing or decreasing
• The $y$-intercept, $b$, 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 ($y_{\text{pred}}$) from the regression equation ($y = mx + b$) for all values of $x$.
- Calculate residuals ($y - y_{\text{pred}}$) for all values of $x$.
- Observe the data and plot the residuals.

Transforming to linearity
- Transform with $y$ versus $x^2$ (quadratic), $y$ versus $\log_{10}x$ (logarithmic) or $y$ versus $\frac{1}{x}$ (reciprocal), depending on the shape of the data plot.

Try quadratic transformation.

Try logarithmic or reciprocal transformation.
Multiple choice

1. 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.

2. A least-squares regression is fitted to the 7 points as shown at right. The plot of residuals would look most similar to:
   A
   B
   C
   D
   E

3. In question 2, assume that the regression line goes through the point (2, 10). The best estimate of the regression equation is:
   A \( y = 4x + 2 \)
   B \( y = 4x - 2 \)
   C \( y = 0.25x + 9.5 \)
   D \( y = 8x + 2 \)
   E \( y = 2x + 2 \)

4. 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

5. In the least-squares regression below, the line goes through the point (6, 12). The best estimate of the regression equation is:
   A \( y = 2x + 6 \)
   B \( y = 2x + 4 \)
   C \( y = 2x + 2 \)
   D \( y = 2x \)
   E \( y = 3x - 6 \)

6. The best estimate for the y-intercept for the equation of question 5 is:
   A 6
   B 4
   C 2
   D 0
   E -6
7 The correlation between two variables \( x \) and \( y \) is \(-0.88\). Which of the following statements is true?
   A As \( x \) increases it causes \( y \) to increase.
   B As \( x \) increases it causes \( y \) 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.

8 For a data set, least-squares regression was performed and it was found that \( s_x^2 = 16 \), \( s_y^2 = 100 \), \( s_{xy} = 4 \). The correlation coefficient is:
   A 0.64   B 0.1   C 0.0025   D 0.16   E 0.4

9 The gradient of the 3-median regression line for the figure at right is:

\[
\begin{align*}
A &\quad \frac{5}{2} \\
B &\quad \frac{1}{2} \\
C &\quad \frac{1}{5} \\
D &\quad \frac{3}{5} \\
E &\quad \frac{2}{5}
\end{align*}
\]

10 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

**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.

Express the equation with exact values of \( m \) and \( b \).
Copy and complete this table of calculations for a least squares regression \((n = 8)\).

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

Find the least-squares regression line and the correlation coefficient for the data in question 3. Express your answers to 2 decimal places.

Using the least-squares regression line from question 3, copy and complete the following table of predicted values.

<table>
<thead>
<tr>
<th>(x)</th>
<th>3</th>
<th>5</th>
<th>7</th>
<th>9</th>
<th>11</th>
<th>13</th>
<th>15</th>
<th>17</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>(y_{\text{pred}})</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For the least-squares regression line from question 3, find the residuals.

**Analysis**

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.

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

Perform least-squares regression on the transformed data from question 2.
4 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 Compute the predicted values, then residuals. Square the residuals and add them up.
- d Fit a straight line to the original data using the 3-median method, stating the equation of this line.
- e Again, compute predicted values, and the residuals and square them.
- f Now, use least-squares regression to find another equation for this line.
- g Using the line from part f, predict the production after 12 months.
- h Comment on the accuracy, usefulness and simplicity of the methods.

5 Let us pursue further the data from question 4.

- a Looking at the original data set, discuss whether linearity is a reasonable assertion.
- b Research into gold mines 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 4g with the one obtained using the logarithmic transformation.