CHAPTER 3
Introduction to regression

eBook*plus*

DIGITAL DOC
doc-9418
10 Quick Questions

CHAPTER CONTENTS

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.

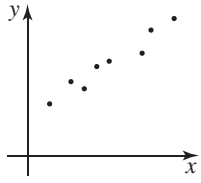
studyon

Units:	3 & 4
AOS:	DA
Topic:	7
Concept:	1

Concept summary
Read a summary of this concept.

3A 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.

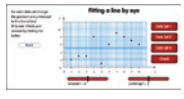


studyon

Units:	3 & 4
AOS:	DA
Topic:	7
Concept:	2

Concept summary
Read a summary of this concept.

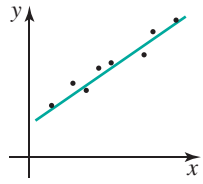
Do more
Interact with lines of best fit.



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.

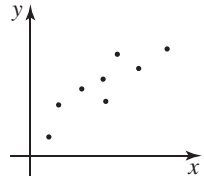
Regression analysis is concerned with finding these straight lines using various methods so that the number of points above and below the lines are 'balanced'.

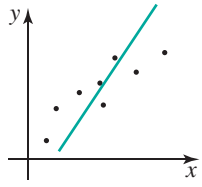
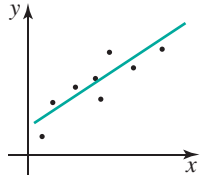
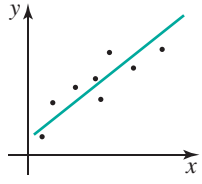
Method of fitting lines by eye

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.



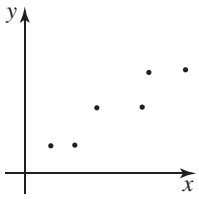
THINK	DRAW
1 Note that the number of points (n) is 8.	
2 Fit a line where 4 points are above the line. Using a clear plastic ruler, try to fit the best line.	
3 The first attempt has only 3 points below the line where there should be 4. Make refinements.	
4 The second attempt is an improvement, but the 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.	

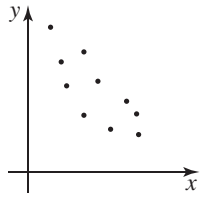
Exercise 3A Fitting a straight line by eye

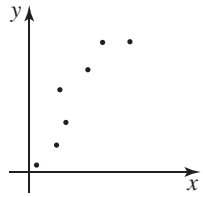
The questions in this exercise represent data collected by groups of students conducting different environmental projects. The students have to fit a straight line to their data sets.

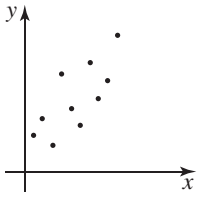
Note: For many of these questions your answers may differ somewhat from those at the end of the chapter. The answers are provided as a guide but there are likely to be individual differences when fitting straight lines by eye.

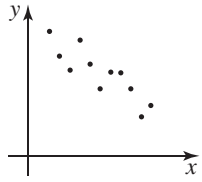
1 WE1 Fit a straight line to the data in the scatterplots using the equal-number-of-points method.

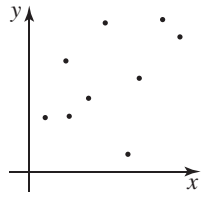
a


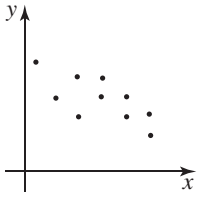
b


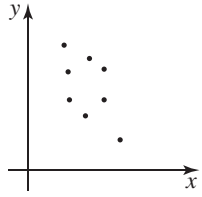
c


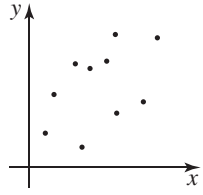
d


e


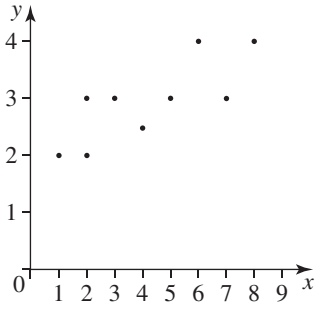
f


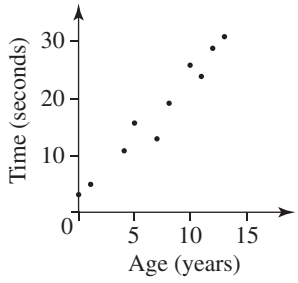
g


h


i


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

a


b


3B 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** on page 98). 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.

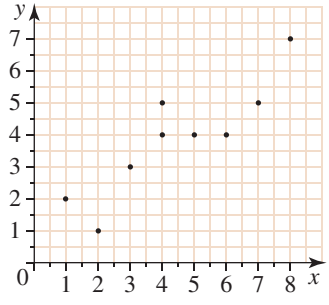


Figure **1**

studyon

Units:	3 & 4
AOS:	DA
Topic:	7
Concept:	3

Concept summary
Read a summary of this concept.

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.

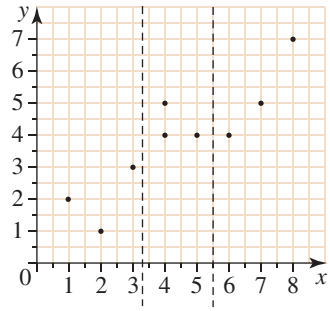


Figure **2**

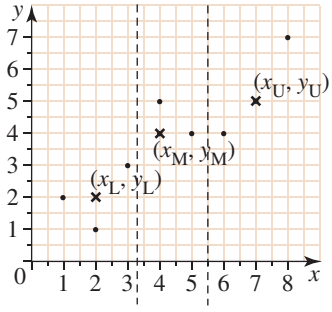


Figure **3**

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 $y = mx + c$).
First, use the coordinates of the lower and upper medians to find the gradient: $m = \dfrac{y_U - y_L}{x_U - x_L}$.
Next, find the y-intercept. If the scale on the axes begins at zero, you can read off the y-intercept of the line. Otherwise, substitute the coordinates of any point on the line into equation and solve for c.

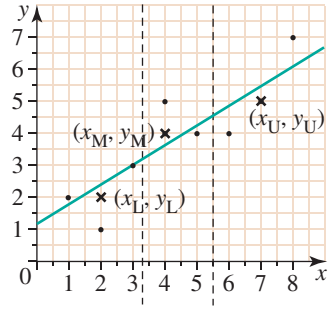


Figure **4**

Arithmetic approach

Using the arithmetic approach, you will proceed as follows.

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

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 $y = mx + c$.

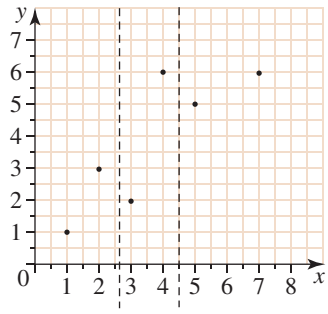
Using a CAS calculator

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

WORKED EXAMPLE 2

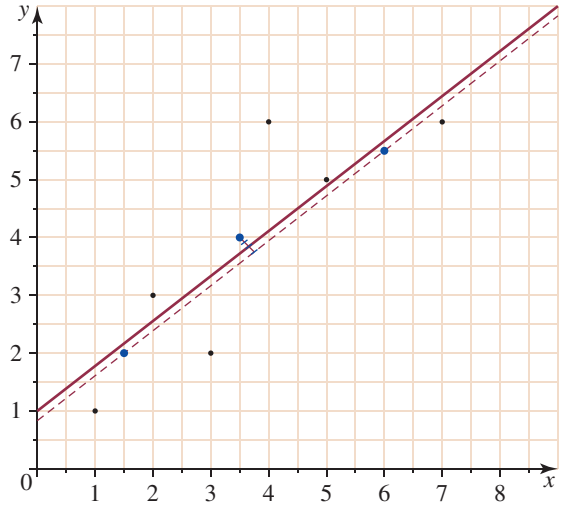
Find the equation of the regression line for the data in the table below using the 3-median method. Give coefficients correct to 2 decimal places.

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

THINK	WRITE/DRAW
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.

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 Use the coordinates of the lower and upper medians to calculate the gradient (m). The points are: (1.5, 2) and (6.5, 5).	$\text{Gradient } (m) = \dfrac{y_U - y_L}{x_U - x_L}$ $= \dfrac{5.5-2}{6-1.5}$ $= \dfrac{3.5}{4.5}$ $= \dfrac{7}{9}$ ≈ 0.78
6 Write the equation of the 3-median regression line.	$y = mx + c$ $= 0.78x + 1$

Method 2: Using the arithmetic approach

3 Find the gradient using the formula, and the upper and lower medians found previously. (This is the same as the graphical approach.)	$m = \dfrac{y_U - y_L}{x_U - x_L}$ $= \dfrac{5.5-2}{6-1.5}$ $= \dfrac{3.5}{4.5}$ $= \frac{7}{9}$ ≈ 0.78

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

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

5 State the regression equation.

$$y = 0.78x + 0.98$$

Note: There are slight variations in the values of the y-intercept of the line between the graphical and the arithmetic approaches. This is because the arithmetic method gives precise values for the y-intercept, whereas the graphical method gives approximate values.

eBook*plus*

DIGITAL DOCS
doc-9419
SkillSHEET 3.1
Finding the median
doc-9420
SkillSHEET 3.2
Gradient
doc-9398
SkillSHEET 3.3
The equation of a straight line

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

1 WE2 Find the regression line for the data in the table below using the 3-median method.

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

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.

Total number of points (n)	Lower group	Middle group	Upper group
10	3	4	3
11			
12			
13			
14			
26			
43			
58			
698			

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

x	5	10	20	20	30	40	50	55	60	65	75
y	80	60	50	70	40	55	40	30	10	25	15

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

Day	1	2	3	4	5	6	7	8	9	10	11	12
Temperature (°C)	14	13	15	17	16	18	19	17	22	20	21	24

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.

Month (x)	1	2	3	4	5	6	7	8	9	10
Sales (y)	85	77	81	73	68	72	64	57	53	49

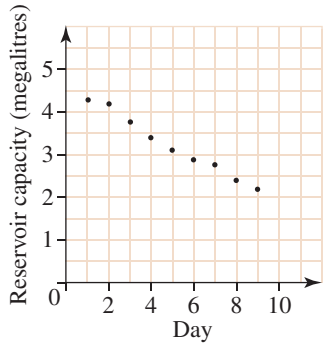
Find the equation of the 3-median regression line.

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

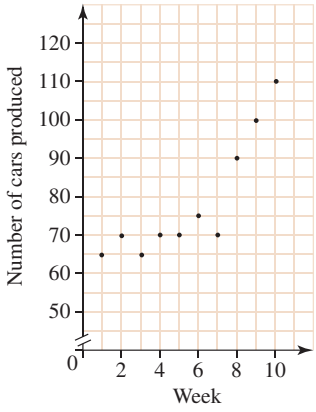
x	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
y	3	5	6	9	11	16	15	13	19	22	26	24	28	31	30	32	36	29	39	40	44

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

8 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.

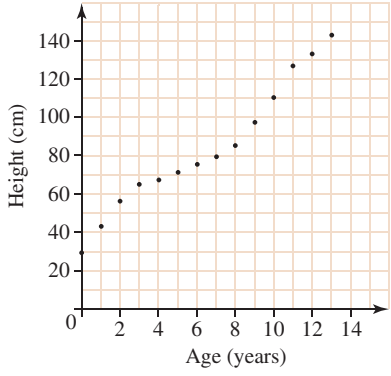


9 Since management instituted new policies, the productivity at DMH car plant has been improving. The scatterplot below shows the number of cars produced each week over a 10-week period.



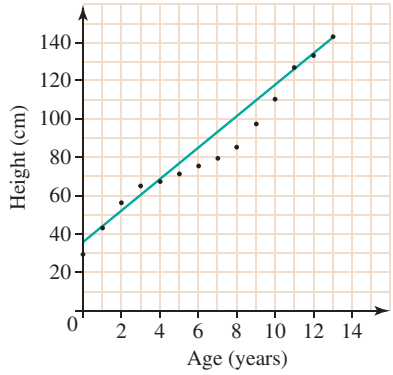
a What are the coordinates of the points used to find the gradient? Use them to find the gradient.
b What are the coordinates of the median of the middle group?
c Using the graph and medians found, estimate the y-intercept.
d State the relationship between cars produced and week as an equation.
e Check your answers using a CAS calculator.

10 **MC** 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?

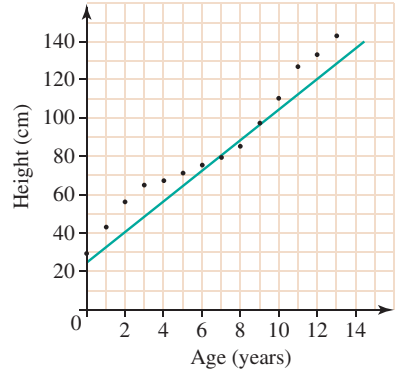


A

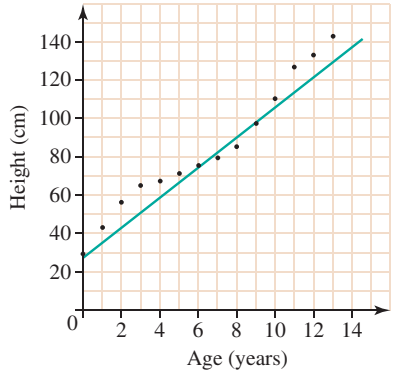
B



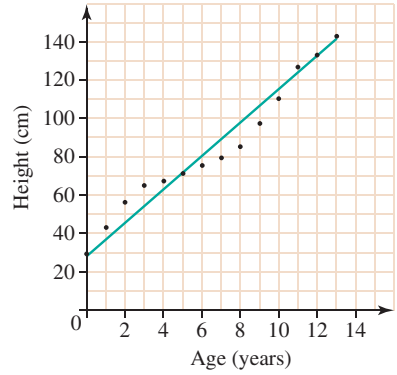
C



D



E



11 **MC** When using the 3-median method for fitting a straight line, which of the following statements is false?

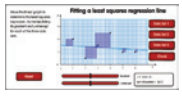
A The straight line is not affected by outliers.
B The two outside medians are used for the gradient of the line.
C For the y-intercept move the line one-third of the way towards the middle median.
D The gradient changes when moving the line towards the middle median.
E The number of points in each group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

study on

Units:	3 & 4
AOS:	DA
Topic:	7
Concept:	4

Concept summary
Read a summary of this concept.

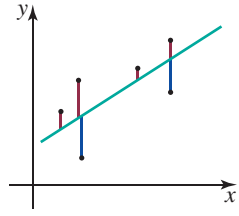
Do more
Interact with least squares regression.



3C 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.

WORKED EXAMPLE 3

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.

Number of minutes (x)	1	3	4	7	10	12	14	15
Number of calls (y)	11	9	10	6	8	4	3	1

THINK	WRITE
1 Enter the data into your calculator and use it to find the equation of least squares regression line.	$y = {}^{-}0.634\,271x + 11.7327$
2 Write the equation with coefficients expressed to 2 decimal places.	$y = {}^{-}0.63x + 11.73$
3 Write the equation in terms of the variable names. Replace x with number of minutes and y with number of calls.	Number of calls = ⁻0.63 × no. of minutes + 11.73

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 $a = \bar{y} - b\bar{x}$.

WORKED EXAMPLE 4

eBook*plus*
TUTORIAL
eles-1264
Worked example 4

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 × 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).

THINK	WRITE
a Recall that the dependent variable is the subject of the equation in $y = mx + c$ form; that is, y.	a The dependent variable is the height of the wife.

b 1	The value of m is the gradient of the regression line. Write the formula and state the required values.	**b** $m = r\frac{s_y}{s_x}$ $\quad r = 0.85$, $s_y = 4.8$ and $s_x = 5.3$
2	Substitute the values into the formula and evaluate m.	$= 0.85 \times \frac{4.8}{5.3}$ $= 0.7698$ ≈ 0.77
c 1	The value of c is the y-intercept of the regression line. Write the formula and state the required values.	**c** $c = \bar{y} - m\bar{x}$ $\bar{y} = 169$, $\bar{x} = 180$ and $m = 0.7698$ (from part **b**)
2	Substitute the values into the formula and evaluate c.	$= 169 - 0.7698 \times 180$ $= 30.436$ ≈ 30.44
d 1	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.	**d** $y = 0.77x + 30.44$ or height of wife = 0.77 × height of husband + 30.44
2	The height of the husband is 195 cm, so substitute $x = 195$ into the equation and evaluate.	$= 0.77 \times 195 + 30.44$ $= 180.59$
3	Write a statement, rounding your answer to the nearest cm.	Using the equation of the regression line found, the wife's height would be 181 cm.

eBook*plus*

DIGITAL DOC
doc-9421
Spreadsheet
Least-squares regression

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.

x	4	6	7	9	10	12	15	17
y	10	8	13	15	14	18	19	23

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

x	1	2	3	4	5	6	7	8	9
y	35	28	22	16	19	14	9	7	2

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

x	$^-4$	$^-2$	$^-1$	0	1	2	4	5	5	7
y	6	7	3	10	16	9	12	16	11	21

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 × 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.

a $\bar{x} = 5.6$ $s_x = 1.2$ $\bar{y} = 110.4$ $s_y = 5.7$ $r = 0.7$
b $\bar{x} = 110.4$ $s_x = 5.7$ $\bar{y} = 5.6$ $s_y = 1.2$ $r = {}^{-}0.7$
c $\bar{x} = 25$ $s_x = 4.2$ $\bar{y} = 10\,200$ $s_y = 250$ $r = 0.88$
d $\bar{x} = 10$ $s_x = 1$ $\bar{y} = 20$ $s_y = 2$ $r = {}^{-}0.5$

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.

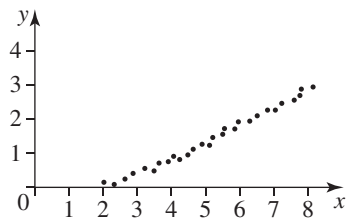
x	1	2	3	4	5	6	7	8	9	10
y	20	18	16	14	12	10	8	6	4	2

a Fit a least-squares regression line.
b Comment on any interesting features of this line.
c Now fit the 'opposite regression line', namely:

x	20	18	16	14	12	10	8	6	4	2
y	1	2	3	4	5	6	7	8	9	10

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

8 **MC** 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 **MC** 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 a certain country over the last 220 years has been recorded.

Year	1780	1800	1820	1840	1860	1880	1900	1920	1940	1960	1980	2000
Life span (years)	51.2	52.4	51.7	53.2	53.1	54.7	59.9	62.7	63.2	66.8	72.7	79.2

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.

Cost of call ($)	1.25	1.85	2.25	2.50	3.25	3.70	4.30	4.90	5.80
Duration of call (seconds)	30	110	250	260	300	350	420	500	600

Cost of call ($)	7.50	8.00	9.25	10.00	12.00	13.00	14.00	16.00	18.00
Duration of call (seconds)	840	1000	1140	1200	1500	1860	2400	3600	7200

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 MC 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) = $^{-}13.6 + 0.54 \times$ hours of daylight
B height of plant (cm) = $^{-}8.5 + 0.34 \times$ hours of daylight
C height of plant (cm) = $2.1 + 0.18 \times$ hours of daylight
D height of plant (cm) = $28.0 + 1.50 \times$ hours of daylight
E height of plant (cm) = $35.68 + 0.54 \times$ 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.

x	1	2	3	4	5	6
y	12	16	17	21	25	29

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.

eBook*plus*

DIGITAL DOC
doc-9422
WorkSHEET 3.1

study on

Units:	3 & 4
AOS:	DA
Topic:	7
Concept:	6

Concept summary
Read a summary of this concept.

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 change in dependent variable as independent variable increases by 1 unit.

The y-intercept indicates the value of the dependent variable when independent variable = 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:

a the equation, describing the relationship between the two variables
b the rate at which bacteria are growing
c the number of bacteria at the start of the experiment.

Day of experiment	1	4	5	9	11
Number of bacteria	500	1000	1100	2100	2500

THINK

a 1 Find the equation of the least-squares regression line using your calculator.

2 Replace x and y with the variables in question.

b The rate at which bacteria are growing is given by the gradient of the least-squares regression.

WRITE

a

Number of bacteria = 202.5 + 206.25 × day of experiment

b m is 206.25, hence on average, the number of bacteria increases by approximately 206 per day.

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

c The y-intercept is 202.5, hence the initial number of bacteria present was approximately 203.

Interpolation and extrapolation

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

$$y = mx + c,$$

where m is the gradient and c is the y-intercept.

This equation can be used to 'predict' the y-value 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 within the range of data in a set, that is, the values that are *in between the 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 a high r value.

Extrapolation

Extrapolation is the use of the regression line to predict values outside the range of data in a set, that is, values that are *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.

WORKED EXAMPLE 6

Using interpolation and the following data set, predict the height of an 8-year-old girl.

Age (years)	1	3	5	7	9	11
Height (cm)	60	76	115	126	141	148

THINK	WRITE
1 Find equation of the least-squares regression line using your calculator. (Age is the independent variable and height is the dependent one.)	$y = 9.23x + 55.63$
2 Replace x and y with the variables in question.	Height $= 9.23 \times$ age $+ 55.63$
3 Substitute 8 for age into equation and evaluate.	When age = 8, Height $= 9.23 \times 8 + 55.63$ $= 129.5$ (cm)
4 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	WRITE
1 Use the regression equation to calculate the girl's height at age 15.	Height = 9.23 × age + 55.63 = 9.23 × 15 + 55.63 =194.08 cm
2 Analyse the result.	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.

eBook*plus*

DIGITAL DOC
doc-9423
Spreadsheet
Interpolation/ Extrapolation

Exercise 3D Interpretation, interpolation and extrapolation

1 **WE5** 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.

Day of experiment	4	5	6	7	8	9
Red blood cell count	210	240	230	260	260	290

Find:

a the equation, describing the relationship between the variables in the form $y = a + bx$

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.

Number of animals	6	4	8	5	7	6
Number of visitors	311	220	413	280	379	334

Find:

a the rate of increase of visitors as the number of live animals is increased by 1

b the predicted number of visitors if there are no live animals.

eBook*plus*

DIGITAL DOC
doc-9424
SkillSHEET 3.4
Using the regression line to make predictions

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

Selling price ($)	60	80	100	120	140	160	200	220	240	260
Sales volume (× 1000)	400	300	275	250	210	190	150	100	50	0

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.

Weekly income ($)	100	200	300	400	500	600	700	800	900	1000
Number of restaurant visits per year	5.8	2.6	1.4	1.2	6	4.8	11.6	4.4	12.2	9

Use the data to predict:

a **WE6** the number of visits per year by a person on a weekly income of $680

b **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.

x	0	1	2	4	5	6	8	10
y	2	3	7	12	17	21	27	35

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.

Distance from Melbourne (km)	10	20	30	40	50	60	70	80
Factory 1 cost ($)	70	70	90	100	110	120	150	180
Factory 2 cost ($)	70	75	80	100	100	115	125	135

a Find the least-squares regression equation for each factory.

b Which factory is likely to have the lowest cost to ship to a shop in Melbourne?

c Which factory is likely to have the lowest cost to ship to Mytown, 115 kilometres from Melbourne?

d 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.

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.

IQ	80		92	102	105		107	111	115	121
Test result (%)	56	60	68	65		74	71	73		92

9 The least-squares regression line for a starting salary (s) as a function of number of years of schooling (n) is given by the rule: $s = 18\,500 + 900n$.

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.

3E 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.

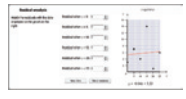
x	2	3	5	8	9	9
y	3	7	12	10	12	16

study on

Units:	3 & 4
AOS:	DA
Topic:	7
Concept:	7

Concept summary
Read a summary of this concept.

Do more
Interact with residual analysis.



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**.

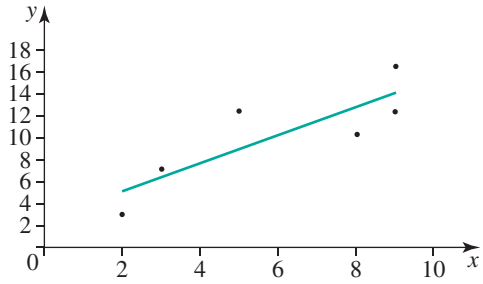


Figure **1**

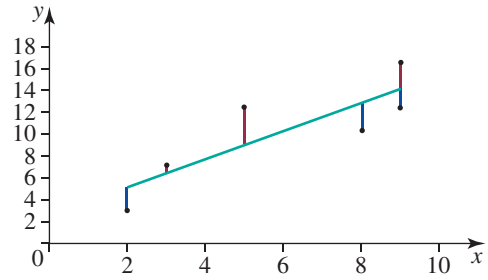


Figure **2**

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 equation of the least-squares regression line and calculate the residuals.

x	1	2	3	4	5	6	7	8	9	10
y	5	6	8	15	24	47	77	112	187	309

THINK | **WRITE**

1. Find the equation of a least-squares regression line using a calculator.

   $y = 28.7x - 78.7$

2. Use the equation of the least-squares regression line to calculate the predicted y-values (these are labelled as y_{pred}) for every x-value in the table. That is, substitute each x-value into the equation and evaluate record results in the table.

x-values	1	2	3	4	5
y-values	5.0	6.0	8.0	15.0	24.0
Predicted y-values	⁻50.05	⁻21.38	7.3	35.98	64.66
Residuals ($y - y_{pred}$)	**55.05**	**27.38**	**0.7**	**⁻20.98**	**⁻40.66**

3. Calculate residuals for each point by subtracting predicted y-values from the actual y-value. (That is, residual = observed y-value – predicted y-value). Record results in the table.

x-values	6	7	8	9	10
y-values	47.0	77.0	112.0	187.0	309.0
Predicted y-values	93.34	122.02	150.7	179.38	208.06
Residuals ($y - y_{pred}$)	**⁻46.34**	**⁻45.02**	**⁻38.7**	**7.62**	**100.94**

Notes

1. The residuals may be determined by ($y - y_{pred}$); 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

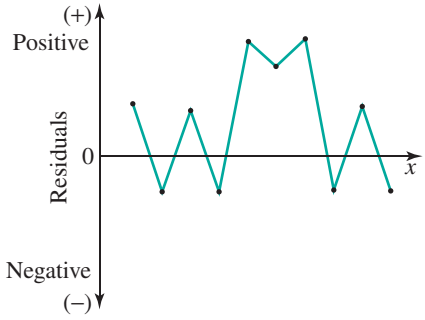
As we observed in the previous worked example, there is not really a good fit between the data and the least-squares regression line; however, there seems to be a pattern in the residuals. How can we observe this pattern in more detail?

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.

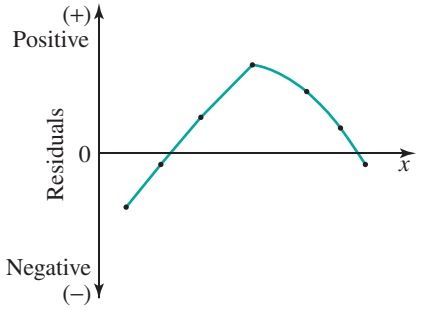
Types of residual plots

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

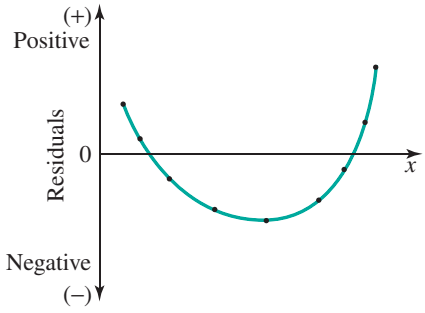
Note: The points are joined together to see the patterns more clearly.



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 (∩), 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 (∪), 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.

WORKED EXAMPLE 9

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

THINK

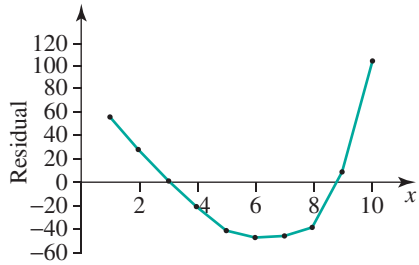
1 Generate a table of values of residuals against *x*.

WRITE/DRAW

***x*-values**	1	2	3	4	5
Residuals $(y - y_{pred})$	55.05	27.38	0.7	⁻20.98	⁻40.66

***x*-values**	6	7	8	9	10
Residuals $(y - y_{pred})$	⁻46.34	⁻45.02	⁻38.7	7.62	100.94

2 Plot the residuals against x. To see the pattern clearer, join the consecutive points with straight line segments.



3 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. 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.

Exercise 3E Residual analysis

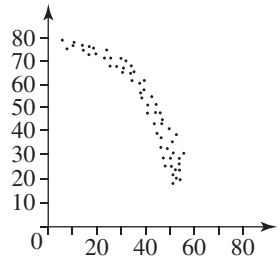
1 WE8 Find the residuals for the following data.

x	1	2	3	4	5	6
y	1	9.7	12.7	13.7	14.4	14.5

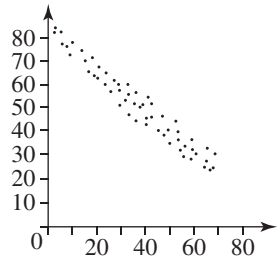
2 WE9 For the results of question **1**, plot the residuals and discuss whether the relationship between x and y is linear.

3 MC Which of the following scatterplots shows linear relationship between the variables?

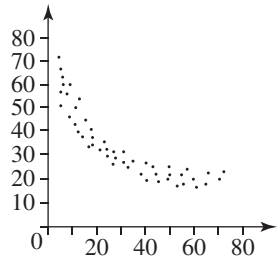
i



ii



iii



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.

Day	2	4	5	7	8	9	10	11
Defective rate (%)	15	10	12	4	9	7	3	4

a The results of least-squares regression were: $m = {}^{-}1.19$, $c = 16.34$, $r = {}^{-}0.87$. Use the above information to calculate the predicted defective rates (y_{pred}).
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.)

Day	Mon.	Tues.	Wed.	Thurs.	Fri.	Sat.	Sun.
Bookings in hotel	158	124	74	56	31	35	22

The results of least-squares regression were:

$m = {}^{-}22.5$, $c = 161.3$, $r = {}^{-}0.94$.

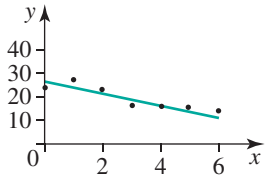
a Find the predicted hotel bookings (y_{pred}) for each day of the week.

b Find the residuals ($y - y_{pred}$).

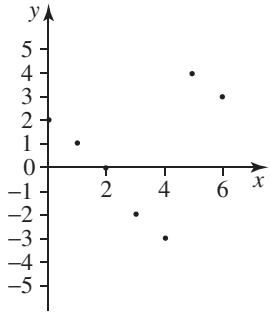
c Plot the residuals and comment on the likely linearity of the data.

d Would this regression line be a typical one for this hotel?

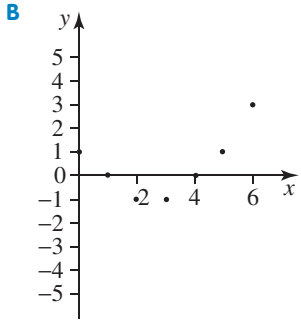
6 **MC** A least-squares regression is fitted to the points shown in the scatterplot at right. Which of the following looks most similar to the residual plot for the data?



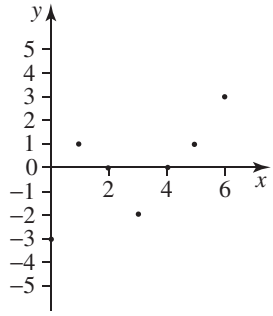
A



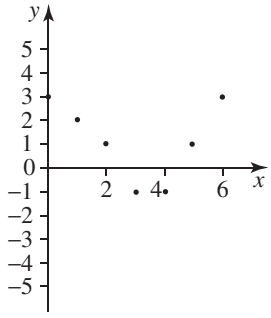
B



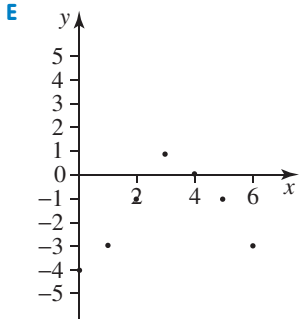
C



D



E



7 From each table of residuals, decide whether or not the relationship between the variables is likely to be linear.

a

x	y	Residuals
1	2	⁻1.34
2	4	⁻0.3
3	7	⁻0.1
4	11	0.2
5	21	0.97
6	20	2.3
7	19	1.2
8	15	⁻0.15
9	12	⁻0.9
10	6	⁻2.8

b

x	y	Residuals
23	56	0.12
21	50	⁻0.56
19	43	1.30
16	41	0.20
14	37	⁻1.45
11	31	2.16
9	28	⁻0.22
6	22	⁻3.56
4	19	2.19
3	17	⁻1.05

c

x	y	Residuals
1.2	23	0.045
1.6	25	0.003
1.8	24	⁻0.023
2.0	26	⁻0.089
2.2	28	⁻0.15
2.6	29	⁻0.98
2.7	34	⁻0.34
2.9	42	⁻0.01
3.0	56	0.45
3.1	64	1.23

8 Consider the following data set.

x	0	1	2	3	4	5	6	7	8	9	10
y	1	4	15	33	60	94	134	180	240	300	390

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

b Find the correlation coefficient and interprete its value.

c Calculate the coefficient of determination and explain its meaning.

d Find the residuals.

e Construct the residual plot and use it to comment on the appropriateness of the assumption that the relationship between the variables is linear.

eBook*plus*

DIGITAL DOC
doc-9425
WorkSHEET 3.2

eBook*plus*

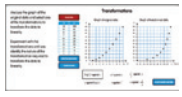
INTERACTIVITY
int-0184
Transforming to linearity
eLESSON
eles-0050
Which way to stretch?

study on

Units:	3 & 4
AOS:	DA
Topic:	7
Concept:	9

Concept summary
Read a summary of this concept.

Do more
Interact with transforming data.



3F 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:

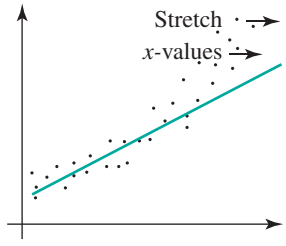
Logarithmic transformations:	y versus $\log_{10}(x)$	$\log_{10}(y)$ versus x
Quadratic transformations:	y versus x^2	y^2 versus x
Reciprocal transformations:	y versus $\frac{1}{x}$	$\frac{1}{y}$ versus x

Choosing the correct transformations

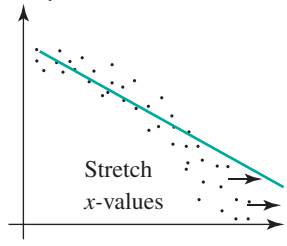
To decide on an appropriate transformation, examine the points on a scatterplot with high values of x and/or y (that is, away from the origin) and decide for each axis whether it needs to be stretched or compressed to make the points line up. The best way to see which of the transformations to use is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

Quadratic transformations

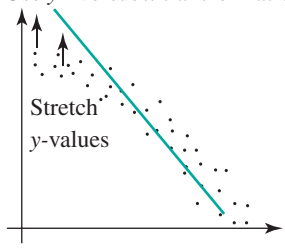
1. Use y versus x^2 transformation.



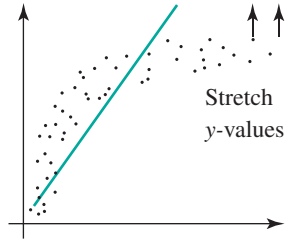
2. Use y versus x^2 transformation.



3. Use y^2 versus x transformation.

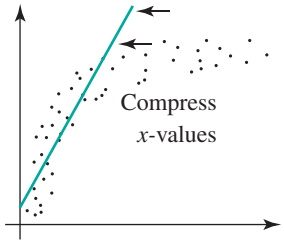


4. Use y^2 versus x transformation.

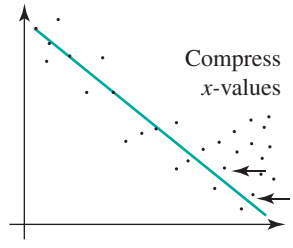


Logarithmic and reciprocal transformations

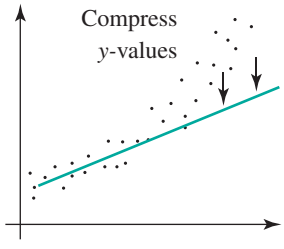
1. Use y versus $\log_{10}(x)$ or y versus $\frac{1}{x}$ transformation.



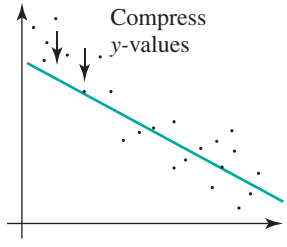
2. Use y versus $\log_{10}(x)$ or y versus $\frac{1}{x}$ transformation.



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.



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

eBook*plus*

TUTORIAL
eles-1265
Worked example 10

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

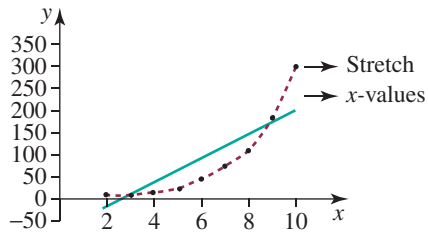
$$y = 28.7x - 78.7 \text{ with } r = 0.87.$$

x	1	2	3	4	5	6	7	8	9	10
y	5	6	8	15	24	47	77	112	187	309

THINK

WRITE/DRAW

1. Plot the data and the regression line to check that a quadratic 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.

x^2	1	4	9	16	25	36	49	64	81	100
y	5	6	8	15	24	47	77	112	187	309

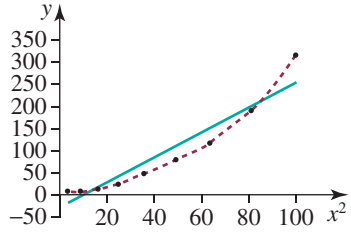
3. Find the equation of the least-squares regression line for the transformed data.
   Using a calculator or spreadsheet:
   (a) gradient (m) = 2.78
   (b) y-intercept (c) = ⁻28.0
   (c) correlation (r) = 0.95.

   $y = 2.78x_T - 28$ where $x_T = x^2$; that is,
   $y = 2.78x^2 - 28$

4. Plot the new transformed data and regression line.

   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.

WORKED EXAMPLE 11

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

Heart rate = 93.2 – 6.97 × time, with $r = {}^{-}0.90$.

Time after operation (h)	x	1	2	3	4	5	6	7	8
Heart rate (beats/min)	y	100	80	65	55	50	51	48	46

THINK

1 Transform the y data by calculating the log of y-values or, in this problem, the log of heart rate.

WRITE

Time	x	1	2	3	4	5	6	7	8
log (heart rate)	$\log y$	2	1.903	1.813	1.740	1.694	1.708	1.681	1.663

2 Use a calculator to find the equation of least-squares regression line for x and $\log y$.

$\log_{10}(y) = 1.98 - 0.05x$

3 Rewrite the equation in terms of the variables in question.

$\log_{10}(\text{heart rate}) = 1.98 - 0.05 \times \text{time}$ (i.e. time = number of hours after the operation.)

4 State the value of r and comment on the result.

$r = {}^{-}0.93$
There is a slight improvement of the correlation coefficient that resulted from applying logarithmic transformation.

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

Once the appropriate model has been established and the equation of least-squares regression line has been found, the equation can be used for predictions.

WORKED EXAMPLE 12

a Using a 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.

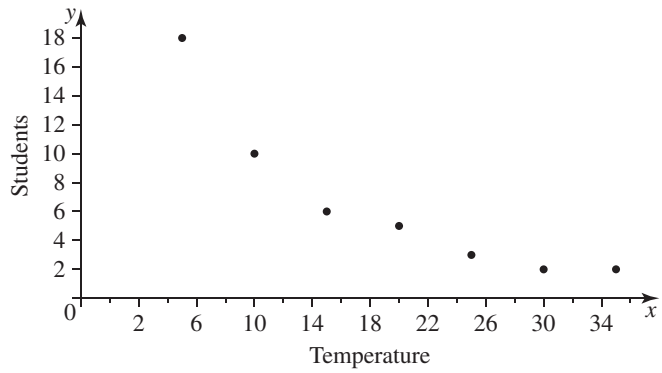
Temperature (°C)	x	5	10	15	20	25	30	35
Number of students in a class wearing jumpers	y	18	10	6	5	3	2	2

THINK	WRITE/DRAW

a 1 Construct the scatterplot. Temperature is the independent variable, while the number of students wearing jumpers is the dependant one.

Therefore, put *temperature* on the horizontal axis and *students* on the vertical axis.

a



2 The x-values should be compressed, so it may be appropriate to transform the x-data by calculating the reciprocal of temperature). Reciprocate each x-value (that is, find $\frac{1}{x}$).

$\frac{1}{\text{Temperature}}$	$\frac{1}{x}$	$\frac{1}{5}$	$\frac{1}{10}$	$\frac{1}{15}$	$\frac{1}{20}$	$\frac{1}{25}$	$\frac{1}{30}$	$\frac{1}{35}$
Number of students wearing jumpers	y	18	10	6	5	3	2	2

3 Use a CAS calculator to find the equation of least-squares regression line for $\frac{1}{x}$ and y.

$y = 94.583x_T - 0.4354$, where $x_T = \frac{1}{x}$

or $y = \frac{94.583}{x} - 0.4354$

4 Replace x and y with the variables in question.

The number of students in class wearing jumpers

$= \frac{94.583}{\text{Temperature}} - 0.4354.$

b 1 Substitute 12 for x into equation of regression line and evaluate.

b Number of students wearing jumpers

$= \frac{94.583}{\text{Temperature}} - 0.4354$

$= \frac{94.583}{12} - 0.4354$

$= 7.447$

2 Write your answer to the nearest whole number.

7 students are predicted to wear jumpers when the temperature is 12 °C.

Note: If the residual plot exhibits a clear pattern, the relationship between the variables is probably not linear. To find an appropriate model, a logarithmic, quadratic or receiprocal transformation can be attempted.

Exercise 3F Transforming to linearity

eBook*plus*

DIGITAL DOC
doc-9482
Spreadsheet
Transforming data

1 **WE 10** Apply a quadratic (x^2) transformation to the following data set. The regression line has been determined as $y = {}^{-}27.7x + 186$ with $r = {}^{-}0.91$.

x	2	3	4	5	7	9
y	96	95	92	90	14	⁻100

2 **WE 11** The *average* heights of 50 girls of various ages were measured as follows.

Age group (years)	9	10	11	12	13	14	15	16	17	18
Average height (cm)	128	144	148	154	158	161	165	164	166	167

The original linear regression yielded:

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

a Plot the original data and regression line.
b Apply $\log_{10}(x)$ 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 Which of the predictions in part **a** were obtained by interpolating?

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

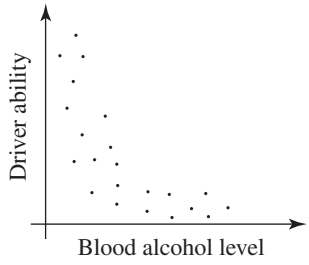
5 **a** **WE12** Apply a reciprocal transformation to the following data obtained by a physics student studying light intensity.

Distance from light source (metres)	1	2	3	4	5	10
Intensity (candlepower)	90	60	28	22	20	12

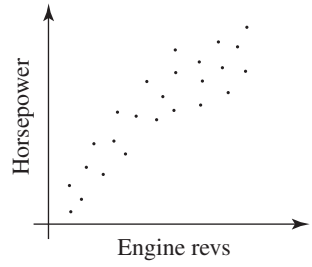
b Use the transformed regression equation to predict the intensity at a distance of 20 metres.

6 For each of the following scatterplots suggest an appropriate transformation(s).

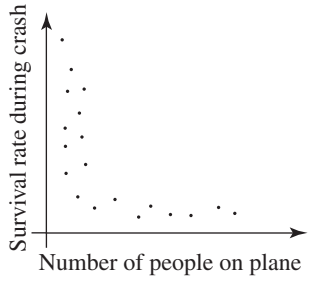
a



b



c



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.

Circle	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
Number of seeds	3	5	8	13	21	34	55	89	144	233

a Plot the data and fit in a least-squares regression line.
b Find the correlation coefficient and interprete its value.
c Using the equation of the regression line, predict the number of seeds in the 11th circle.
d Find the residuals.
e Construct the residual on plot. Is the relation between the number of the circle and the number of seeds 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 Find the equation of the least-squares regression line for the transformed data.
d Calculate the coefficient of determination and interprete its value.
e Using the equation of 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: $m = \dfrac{y_U - y_L}{x_U - x_L}$.
  Step 5. Calculate the y-intercept using the formula:

$$c = \tfrac{1}{3}[(y_L + y_M + y_U) - m(x_L + x_M + x_U)].$$

  Step 6. Substitute m and c into the equation $y = mx + c$.

Fitting a straight line — least-squares regression

- Use a calculator to find the equation of the least-squares regression line. The equation can be obtained in one of these forms:

$$y = mx + b$$
or
$$y = a + bx.$$

- 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) $m = r\dfrac{s_y}{s_x}$ (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 $b = r\dfrac{s_y}{s_x}$ and $a = \bar{y} - b\bar{x}$.

Interpretation, interpolation and extrapolation

- The slope (m) of the regression line $y = mx + c$ indicates the change in the dependent variable as independent variable increases by 1.
- The y-intercept, c, indicates the value of the dependent variable when independent variable = 0.
- Interpolation is the use of the regression line to predict values 'between' the values already in the data set (predicting within the range of 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 (predicting outside the data set).

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).
- Construct the residual plot.
- If the residual plot shows points randomly scattered around zero (i.e. there is no clear pattern), the relationship between the variables in question is probably linear.
- If the residual plot shows a clear pattern, the relationship between variables is probably not linear.

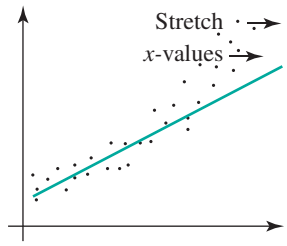
Transforming to linearity

- Transform non-linear data to linearity by using one or more of the following possible transformations.

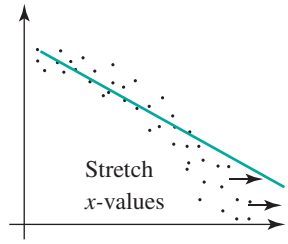
Compressing axis:	y versus $\log_{10}(x)$	y versus $\frac{1}{x}$
	$\log_{10}(y)$ versus x	$\frac{1}{y}$ versus x
Stretching axis:	y versus x^2	y^2 versus x

Quadratic transformations

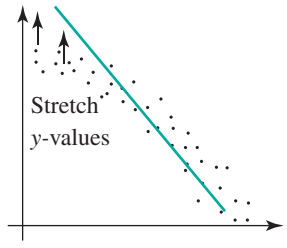
1. Use y versus x^2 transformation.



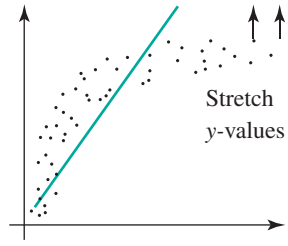
2. Use y versus x^2 transformation.



3. Use y^2 versus x transformation.

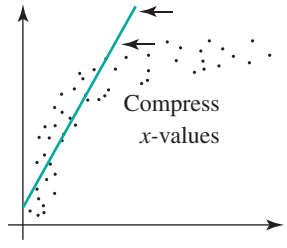


4. Use y^2 versus x transformation.

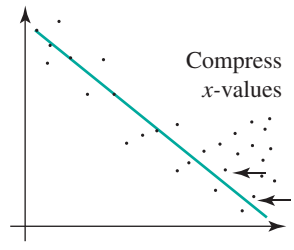


Logarithmic and reciprocal transformations

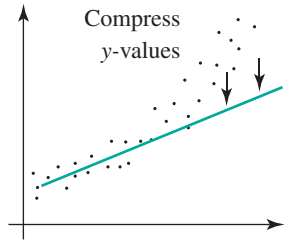
1. Use y versus $\log_{10}(x)$ or y versus $\frac{1}{x}$ transformation.



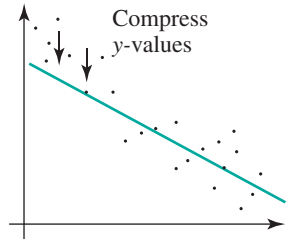
2. Use y versus $\log_{10}(x)$ or y versus $\frac{1}{x}$ transformation.



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



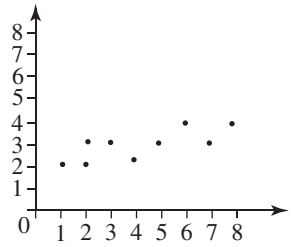
4. Use $\log_{10}(y)$ versus x or $\frac{1}{y}$ versus x transformation.



Chapter review

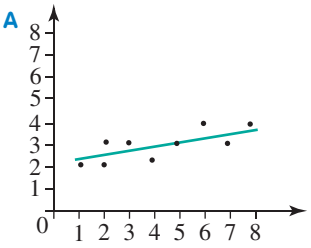
MULTIPLE CHOICE

Use the figure at right to answer questions **1** and **2**.

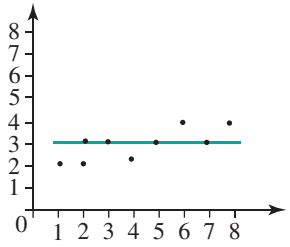


1 The most appropriate line of best fit for the figure is:

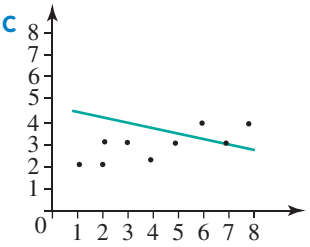
A



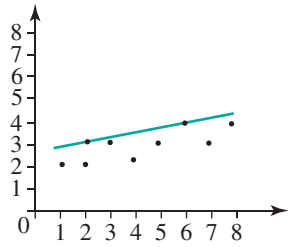
B



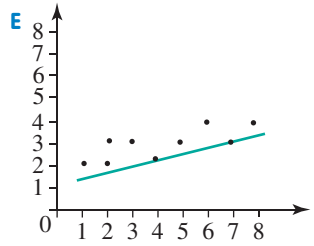
C



D



E



2 The gradient of the 3-median regression line is:

A $\frac{5}{2}$ **B** $\frac{1}{2}$ **C** $\frac{1}{5}$ **D** $\frac{3}{5}$ **E** $\frac{2}{5}$

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

x	25	36	45	78	89	99	110
y	78	153	267	456	891	1020	1410

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 statistics

$\bar{x} = 154.4$ $s_x = 5.8$ $\bar{y} = 172.5$ $s_y = 7.4$ $r = 0.9$

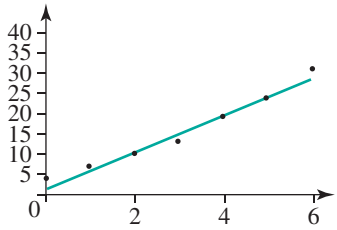
the values of m and c, respectively, for the equation of the regression line $y = mx + c$ are:

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:

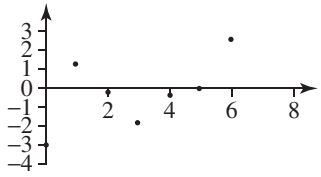
A 21.32 **B** 18.92 **C** 16.52
D 1.58 **E** $^{-}2.4$

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

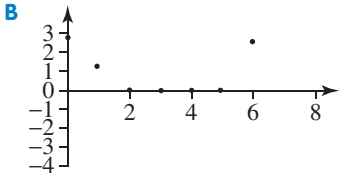


The residual plot would look most similar to:

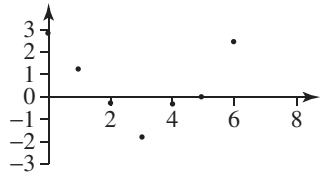
A



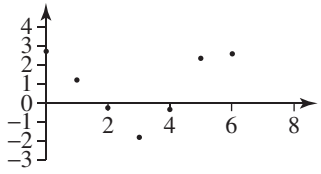
B



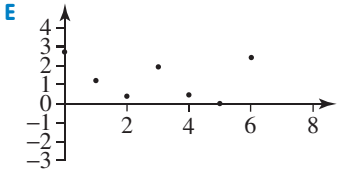
C



D



E



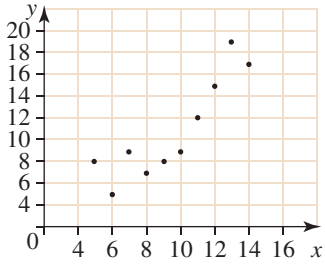
10 After a transformation, a relationship was found to be $y = 0.4x^2 + 12.1$. The predicted value for y given that $x = 2.5$ is:

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.



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.

x-values	1	2	4	8	9	10	11	15
y-values	23	21	20	14	16	9	12	5

4 Use the data from question **2** to fit a least-squares regression line. Express the equation in the form $y = a + bx$.

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

6 Use the following summary statistics to find:

a the slope, m, of the least-squares regression line

b the y-intercept, c, of the least-squares regression line

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

$\bar{x} = 15$ $s_x = 5$ $\bar{y} = 10$ $s_y = 2.5$ $r = {}^{-}0.9$

7 Using the least-squares regression line from question **5**, copy and complete the following table of predicted values.

x	3	5	7	9	11	13	15	17	20
y_{pred}									

8 For the least-squares regression line from question **5**, find the residuals.

EXTENDED RESPONSE

Task 1

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

Day	1	2	3	4	5	6	7	8
Units sold	1	2	4	9	20	44	84	124

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 Use a CAS calculator to construct the scatterplot of the data. What kind of relationship between the variables does the scatterplot suggest?

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

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

2 Transform the data from question **1** using a quadratic (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.

Month (1 = January)	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.
Production (tonnes)	3	8	10.8	12	11.6	14	15.5	15	18.1

a Plot the data and fit a line of best fit by eye.

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 Find the equation of the least-squares regression 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.

eBook*plus*

DIGITAL DOC
doc-9426
Test Yourself
Chapter 3

2 Using the data from question **1** above, answer the following questions.

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 Predict the level of production of gold after 12 months using the equation obtained in part **d**. Compare the prediction from question **1e** above with the one obtained using the logarithmic transformation.

study on

Units: 3 & 4
AOS: DA

Practice VCE exam questions

Use StudyON to access all exam questions on this topic since 2002.

eBookplus ICT activities

Chapter opener

DIGITAL DOC

- 10 Quick Questions doc-9418: Warm up with a quick quiz on introduction to regression. *(page 95)*

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

DIGITAL DOCS

- SkillSHEET 3.1 doc-9419: Practise finding the median. *(page 100)*
- SkillSHEET 3.2 doc-9420: Practise calculating the gradient (I). *(page 100)*
- SkillSHEET 3.3 doc-9398: Practise finding the equation of a straight line. *(page 100)*

3C Fitting a straight line — least-squares regression

DIGITAL DOCS

- Spreadsheet doc-9421: Investigate least-squares regression. *(page 104)*
- WorkSHEET 3.1 doc-9422: Fitting a straight line by eye and using the 3-median regression line. *(page 106)*

TUTORIAL

- WE 4 eles-1264: Learn how to find the equation of the least-squares regression line using r, s_x, and s_y. *(page 103)*

3D Interpretation, interpolation and extrapolation

DIGITAL DOCS

- Spreadsheet doc-9423: Investigate interpolation and extrapolation on a scatterplot. *(page 108)*
- SkillSHEET 3.4 doc-9424: Practise using the regression line to make predictions. *(page 108)*

3E Residual analysis

DIGITAL DOC

- WorkSHEET 3.2 doc-9425: 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 113)*

3F Transforming to linearity

DIGITAL DOC

- Spreadsheet doc-9482: Investigate different transformations to linearity. *(page 117)*

TUTORIALS

- WE 10 eles-1265: Watch a tutorial on applying a parabolic transformation to data using a CAS calculator. *(page 115)*

INTERACTIVITY

- Transforming to linearity int-0184: Use the interactivity to consolidate your understanding of applying appropriate transformations to achieve linearity. *(page 114)*

eLESSON

- Which way to stretch? eles-0050: Discover how to use a scatterplot displaying a non-linear relationship to determine how to transform data to achieve linearity. *(page 114)*

Chapter review

DIGITAL DOC

- Test Yourself doc-9426: Take the end-of-chapter test to test your progress. *(page 124)*

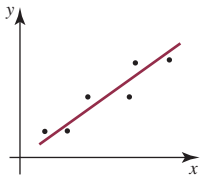
To access eBookPLUS activities, log on to www.jacplus.com.au

INTRODUCTION TO REGRESSION

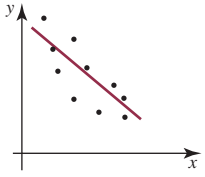
Exercise 3A Fitting a straight line by eye

(*Note:* Best fit lines are indicated as a guide only.)

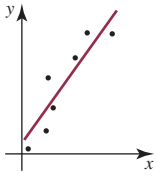
1 a



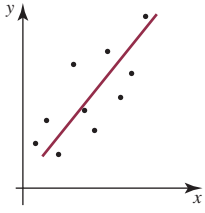
b



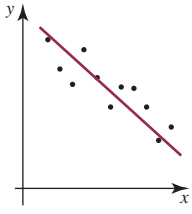
c



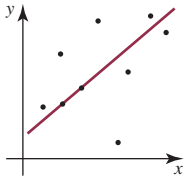
d



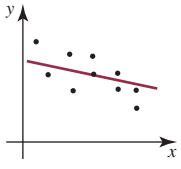
e



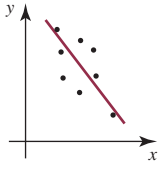
f



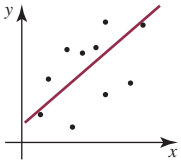
g



h



i



2 a $y = 0.4x + 1.3$
b Time = 2 × age + 2.5

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

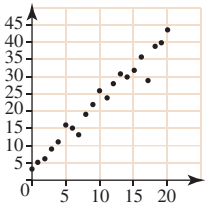
1 $y = 2x - 3.2$

2

n (total)	Lower group	Middle group	Upper group
10	3	4	3
11	4	3	4
12	4	4	4
13	4	5	4
14	5	4	5
26	9	8	9
43	14	15	14
58	19	20	19
698	233	232	233

3 $y = {}^{-}0.95x + 78.8$ **4** D
5 B **6** $y = {}^{-}4x + 90$

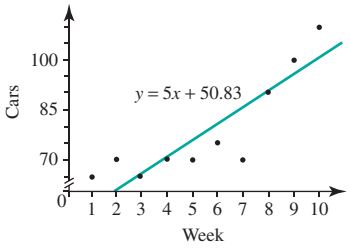
7 a



b $y = 1.93x + 3.71$

8 a (2, 4.2) and (8, 2.4), ⁻0.3
b (5, 3.1)
c Approximately 4.7 (using calculator 4.733)
d Level = ⁻0.3 × day + 4.7

9 a (2, 65) and (9, 100), 5
b (5.5, 70)
c From the graph, approximately 50
d Number of cars = 5 × week + 50
e



10 A **11** D

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

1 $y = 1.04x + 4.57$
2 $y = {}^{-}3.72x + 35.47$
3 $y = 1.2x + 9.06$
4 a The mathematics exam mark
b 0.95 **c** ⁻6.07
d 75%

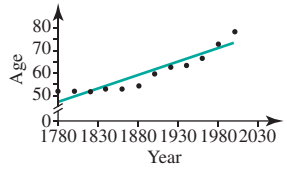
5 a $y = 3.33x + 91.78$
b $y = {}^{-}0.15x + 21.87$
c $y = 52.38x + 8890.48$
d $y = {}^{-}x + 30$

6 The least-squares regression equations are exactly the same as obtained in questions **1**, **2** and **3**.

7 a $y = {}^{-}2x + 22$
b A 'perfect' fit
c $y = {}^{-}0.5x + 11$
d The two lines are inverses of each other.

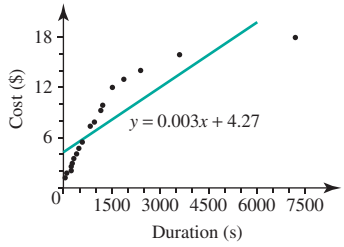
8 E **9** A

10 a $y = 0.119x - 164.7$
b



c The data definitely are not linear; there are big increases from 1880–1920, 1940–2000.

11 a Duration of call
b Cost of call ($) = $0.002 57 × duration of call (sec) + $4.27
c



The line does not fit closely for all data points. The equation is not reliable due to outliers. If you eliminate the last two calls then there is a direct relationship.

12 D

13 a $y = 4x + 8$, perfect fit, but meaningless
b $y = 2.5x + 10$, good fit, but almost meaningless
c $y = 2.8x + 9.5$, $y = 3.1x + 8.9$, $y = 3.3x + 8.4$, good fit.
d Perhaps 'converging' to the correct line?

Exercise 3D Interpretation, interpolation and extrapolation

1 a $y = 157.3 + 14x$ **b** 14 cells per day
c 157

2 a 48.5, or 49 people per extra animal
b 31.8, or 32 visitors

3 $y = {}^{-}1.72x + 464$, $r = {}^{-}0.98$. Gradient shows a drop of 1720 sales for every $1 increase in the price of the item. Clearly, the *y*-intercept is nonsensical in this case since an item is not going to be sold for $0! This is a case where extrapolation of the line makes no sense.

4 a 7 **b** 18

5 a $y = 3.381x + 0.286$ **b** 10.4 **c** 40.9
d 1.99 **e** 7.31 **f** c

6 a Factory1: $y = 1.51x + 43.21$;
Factory 2: $y = 0.96x + 56.61$
b Factory 1 is cheaper at $43.21 (compared to Factory 2 at $56.61).

c Factory 2 is cheaper at \$167.47 (compared to Factory 1 at \$216.86).
d Factory 2 is marginally more linear (Factory 1: $r = 0.97$; Factory 2: $r = 0.99$).

7 a \$2152 **b** 180 **c** \$78 200
d \$600 **e a, b** only

8

IQ	Test result (%)
80	56
87	60
92	68
102	65
105	73
106	74
107	71
111	73
115	80
121	92

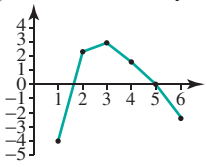
9 a \$27 500 **b** \$29 300
c \$32 000 **d** About 13 years
e Various answers

Exercise 3E Residual analysis

1

x	y	y_{pred}	Residuals
1	1	5.1	⁻4.1
2	9.7	7.46	2.24
3	12.7	9.82	2.88
4	13.7	12.18	1.52
5	14.4	14.54	⁻0.14
6	14.5	16.9	⁻2.4

2 By examining the original scatterplot, and residual plot, data are clearly not linear.

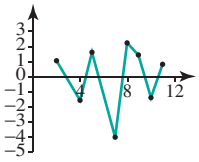


3 D

4 a, b

Day	Defective rate (%)	y_{pred}	Residuals
2	15	13.96	1.04
4	10	11.58	⁻1.58
5	12	10.39	1.61
7	4	8.01	⁻4.01
8	9	6.82	2.18
9	7	5.63	1.37
10	3	4.44	⁻1.44
11	4	3.25	0.75

c



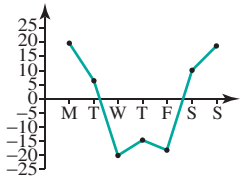
No apparent pattern in the residuals — likely to be linear

d 15.15%
e 13.7 days. Unlikely that extrapolation that far from data points is accurate. Unlikely that there would be 0% defectives.

5 a, b

Day	Bookings in hotel	y_{pred}	Residuals
1	158	138.8	19.2
2	124	116.3	7.7
3	74	93.8	⁻19.8
4	56	71.3	⁻15.3
5	31	48.8	⁻17.8
6	35	26.3	8.7
7	22	3.8	18.2

c



Slight pattern in residuals — may not be linear

d Decline in occupancy likely due to drought — an atypical event.

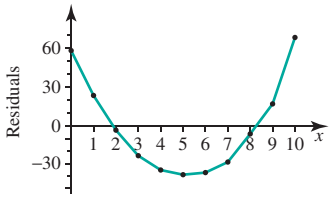
6 C

7 a Non-linear **b** Linear
c Non-linear

8 a $y = 37.93x - 57.73$
b $r = 0.958$. This means that there is a strong positive relationship between variables x and y.
c 0.9177, therefore 91.8% of the variation in y can be explained by the variation in x.
d

x	Residuals
0	58.7
1	23.8
2	⁻3.1
3	⁻23.1
4	⁻34.0
5	⁻37.9
6	⁻35.8
7	⁻27.8
8	⁻5.7
9	16.4
10	68.5

e

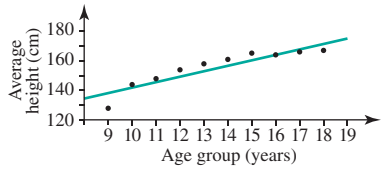


There is a clear pattern; the relationship between the variables is non-linear.

Exercise 3F Transforming to linearity

1 $y = {}^{-}2.62x_T + 128.15$ where $x_T = x^2$, $r = {}^{-}0.97$, which shows some improvement.

2 a



b

log (age group)	Average height (cm)
0.954	128
1	144
1.041	148
1.079	154
1.114	158
1.146	161
1.176	165
1.204	164
1.230	166
1.255	167

c $y = 117.2x_T + 24.26$ where $x_T = \log_{10}(x)$, $r = 0.95$, most non-linearity removed.

3 a i 123.3 cm **ii** 143.9 cm **iii** 176.7 cm
b a ii

4 Normal growth is linear only within given range; eventually the girl stops growing. Thus logarithmic transformation is a big improvement over the original regression.

5 a $y = 90.867x_T + 2.572$ where $x_T = \frac{1}{x}$ $(r = 0.9788)$
b Intensity is 7.1 candlepower.

6 a Compresst he y- or x-values using logs or reciprocals.
b Stretch the y-values using y^2 or compress the x-values using logs or reciprocals.
c Compress the y- or x-values using logs or reciprocals.

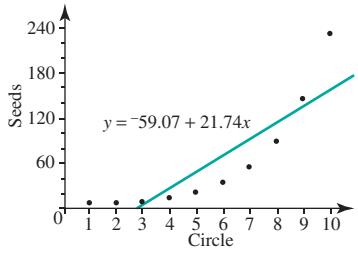
7 a ⁻11.25 **b** ⁻11.25

8 a ⁻24.55
b Cannot take the log of a negative number.
c Cannot take the log of zero.

9 a 3.39 **b** 0.34

10 a ⁻0.08 **b** ⁻0.08

11 a

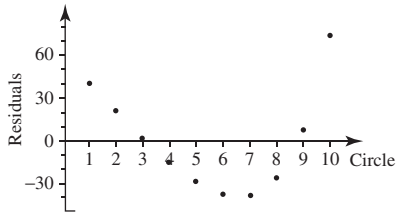


b $r = 0.87$, which means it is a strong and positive relationship.
c 180

d

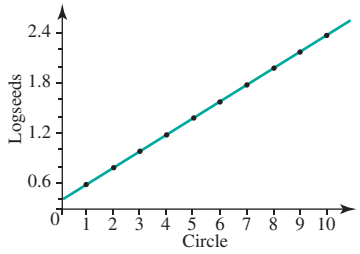
Circle	Seeds	Residual
1	3	40.33
2	5	20.59
3	8	1.85
4	13	⁻14.89
5	21	⁻28.36
6	34	⁻37.37
7	55	⁻38.12
8	89	⁻25.58
9	144	7.41
10	233	74.67

e



f **i** Stretch using x^2
ii Compress either $\log_{10}(y)$ or $\frac{1}{y}$

12 a



b $r = 0.9999$, this is an almost perfect relation.
c $\log_{10}(y) = 0.2094x + 0.2746$
$\log_{10}$ (number of seeds) $= 0.2094 \times$ circle number $+ 0.2746$
d 0.9999, 99.99% (100.0%) of variation in number of seeds is due to number of circles. This is a perfect relation, often found in nature (see the Golden Ratio).
e 378
f This is a much better prediction as it follows the steep upward trend.

CHAPTER REVIEW

MULTIPLE CHOICE

1 A **2** E **3** B **4** E
5 D **6** C **7** B **8** C
9 C **10** C

SHORT ANSWER

1 $y = {}^{-}3.5x + 25$ **2** $y = \frac{9}{7}x - \frac{22}{21}$
3 $y = {}^{-}1.33x + 24.56$ **4** $y = {}^{-}2.055 + 1.364x$
5 $y = {}^{-}1.25x + 24.35$, $r = {}^{-}0.96$
6 a ⁻0.45 **b** 16.75

7

x	y_{pred}
3	20.6
5	18.1
7	15.6
9	13.1
11	10.6
13	8.1
15	5.6
17	3.2
20	⁻0.65

8

x	y	y_{pred}	$y - y_{pred}$
1	23	23.1	⁻0.1
2	21	21.85	⁻0.85
4	20	19.35	0.65
8	14	14.35	⁻0.35
9	16	13.1	2.9
10	9	11.85	⁻2.85
11	12	10.6	1.4
15	5	5.6	⁻0.6

EXTENDED RESPONSE

Task 1

1 a Likely to be a y versus x^2 relationship
b A poor predictor for most values of x
c 128

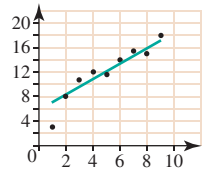
2

Day	1	4	9	16	25	36	49	64
Units sold	1	2	4	9	20	44	84	124

3 $y = 1.96x_T - 13.86$, where $x_T = x^2$
4 182

Task 2

1 a



b Something like $y = 1.26x + 5.77$
c $y = 1.25x + 5.58$
d $y = 1.55x + 4.27$
e 22.87
f Simplicity of eye fitting versus accuracy in this case is quite good. Little difference in the sum of squared errors. Least-squares regression gives quite a different answer from the other 2 methods, with consequent change in errors. (The 3-median method is subject to errors due to outliers, and computational errors.)

2 a Not very linear, logarithmic transformation suggested
b

$\log_{10}$ (month)	Production (tonnes)
0	3
0.301	8
0.477	10.8
0.602	12
0.699	11.6
0.778	14
0.845	15.5
0.903	15
0.954	18.1

c $y = 14.08x_T + 3.30$, where $x_T = \log_{10}(x)$
d Square of residual error is reduced, correlation (0.98) is closer to 1, graph looks more linear
e Prediction of 18.49 against 22.87 using untransformed data. Given the nature of the data, likely to be more accurate.