Chapter 4

Data transformation

Chapter questions

- ▶ What is a squared transformation and when is it used?
- ▶ What is a log transformation and when is it used?
- ▶ What is a reciprocal transformation and when is it used?
- ▶ How do I interpret a least squares line fitted to transformed data?
- ▶ How do I use a least squares line fitted to transformed data for prediction?
- ▶ How do I use a residual plot to assess the effectiveness of a data transformation?
- ▶ How do I use the coefficient of determination to assess the effectiveness of a data transformation?

You may recall from your study of Variation in General Mathematics 12 that a non-linear association could be transformed into a linear association using **data transformation**. The transformations introduced were the squared, log and reciprocal transformations. In this chapter we consider the effect of each of these three transformations when applied to one axis only (either x or y, but not both), using them to linearise scatterplots. This is the first step towards solving problems involving non-linear associations.

ISBN 978-1-009-11041-9

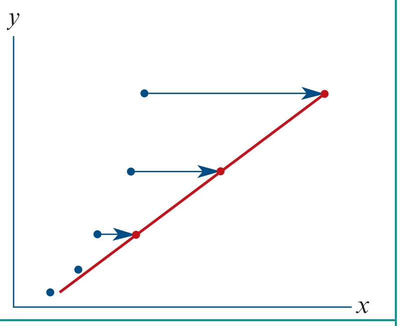
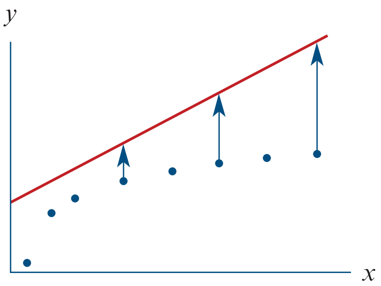
Cambridge University Press

4A The squared transformation

Learning intentions

- To be able to apply a squared transformation to either x or y.
- To be able to fit a least squares regression line to the transformed data.
- To be able to use the least squares regression line fitted to the transformed data for prediction.

The **squared transformation** is a *stretching* transformation. It works by *stretching* out the upper end of the scale on either the x- or y-axis. The effects of applying the x^2 and y^2 transformations (separately) to a scatterplot are illustrated graphically below.

Transformation	*Outcome*	*Graph*
x^2	Spreads out the high x-values relative to the lower x-values, leaving the y-values unchanged. This has the effect of straightening out curves like the one shown opposite.	
y^2	Spreads out the high y-values relative to the lower y-values, leaving the x-values unchanged. This has the effect of straightening out curves like the one shown opposite.	

The following example shows how the x-squared transformation works in practice.

Example 1 Applying the x-squared transformation

A base jumper leaps from the top of a cliff, 1560 metres above the valley floor. The scatterplot below shows the height (in metres) of the base jumper above the valley floor every second, for the first 10 seconds of the jump.

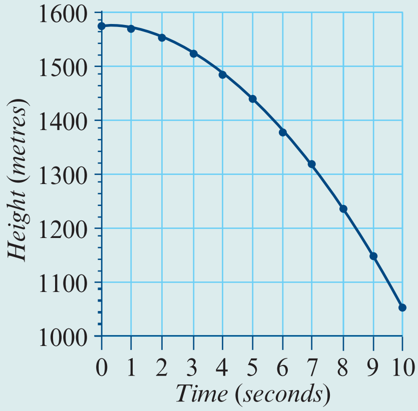
ISBN 978-1-009-11041-9

Cambridge University Press

A scatterplot shows that there is a strong negative association between the *height* of the base jumper above the ground and *time*.

a Apply a squared transformation to the variable *time*, and determine the least squares regression line for the transformed data.

b Use the least squares equation to predict to the nearest metre the height of the base jumper after 3.4 seconds.



Solution

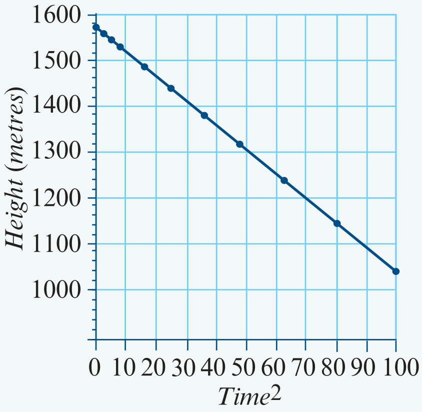
a Applying the squared transformation involves changing the scale on the time axis to *time*^{2}.

From the plot opposite we see that the association between *height* and *time*^{2} is now linear.

Now that we have a linearised scatterplot, we can use a least squares line to model the association between *height* and *time*^{2}.

The equation of this line is:

$$height = 1560 - 4.90 \times time^2$$



b Like any regression line, we can use its equation to make predictions. After 3.4 seconds, we predict that the height of the base jumper is:

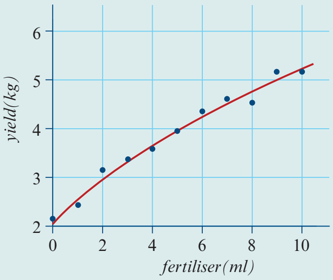
$$height = 1560 - 4.90 \times 3.4^2 = 1503 \text{ m (to nearest m)}$$

The next example shows how the *y*-squared transformation works in practice.

Example 2 Applying the *y*-squared transformation

In a study of the effectiveness of fertiliser on the yield of strawberry plants, differing amounts of liquid fertiliser (in mL) were given to groups of plants, and their average yield (in kg) measured.

A scatterplot shows that there is a strong positive association between the *fertiliser* and *yield*.



a Apply a squared transformation to the variable *yield*, and determine the least squares regression line for the transformed data.

b Use the least squares equation to predict the *yield* of a plant given 6.5 mL of fertiliser, giving your answer to 1 decimal place.

Solution

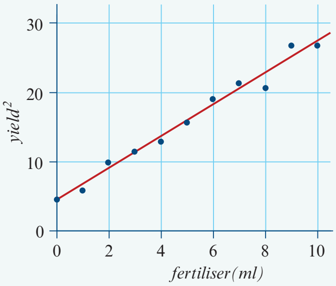
a Applying the *y*-squared transformation involves changing the scale on the *y*-axis to *yield*^{2}.

From the plot opposite we see that the association between *yield*^{2} and *fertiliser* is now linear.

Now that we have a linearised scatterplot, we can use a least squares line to model the association between *yield*^{2} and *fertiliser*.

The equation of this line is:

$$yield^2 = 4.45 + 2.29 \times fertiliser$$



b Using this equation, when we predict that:

$$yield^2 = 4.45 + 2.29 \times 6.5 = 19.34$$

and $yield = \sqrt{19.34} = \pm 4.4$

Looking at the scatterplot we can see that only the positive value of the square root makes sense, so our prediction is 4.4 kg.

Performing a data transformation is quite computationally intensive, but your CAS calculator is well suited to the task.

CAS 1: Using the TI-Nspire CAS to perform a squared transformation

The table shows the height (in m) of a base jumper for the first 10 seconds of her jump.

Time	0	1	2	3	4	5	6	7	8	9	10
Height	1560	1555	1540	1516	1482	1438	1383	1320	1246	1163	1070

a Construct a scatterplot displaying *height* (the RV) against *time* (the EV).

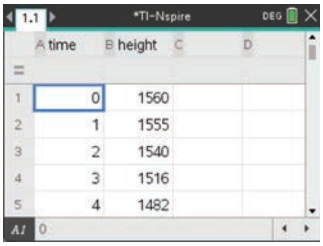
b Apply an *x*-squared transformation and fit a least squares line to the transformed data.

c Use the regression line to predict the height of the base jumper after 3.4 seconds.

Steps

1 Start a new document by pressing ctrl + N .

2 Select **Add Lists & Spreadsheet**.
Enter the data into lists named *time* and *height*, as shown.

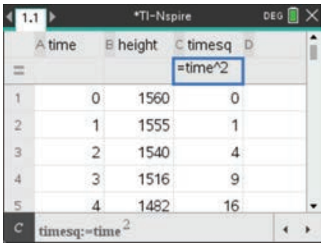


	A time	B height	C	D
=				
1	0	1560		
2	1	1555		
3	2	1540		
4	3	1516		
5	4	1482		

A1 0

3 Name column C as *timesq* (short for ‘time squared’).

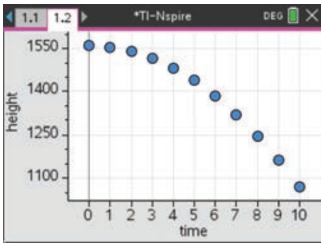
4 Move the cursor to the formula cell below *timesq*.
Enter the expression = *time*^2 by pressing (=) , then typing **time∧2**. Pressing enter calculates and displays the values of *timesq*.



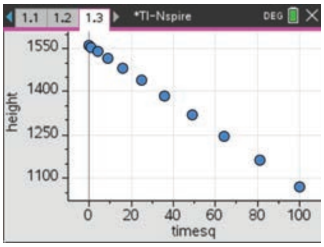
	A time	B height	C timesq	D
=			=time^2	
1	0	1560	0	
2	1	1555	1	
3	2	1540	4	
4	3	1516	9	
5	4	1482	16	

C timesq:=time^2

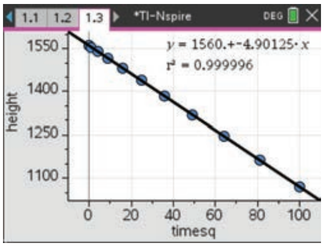
5 Press ctrl + I and select **Add Data & Statistics**.
Construct a scatterplot of *height* against *time*. Let *time* be the explanatory variable and *height* the response variable. The plot is clearly non-linear.



6 Press ctrl + I and select **Add Data & Statistics**.
Construct a scatterplot of *height* against *time*^{2}.
The plot is now linear.



7 Press menu**>Analyze>Regression>Show Linear (a** + **bx)** to plot the line on the scatterplot with its equation.
Note: The x in the equation on the screen corresponds to the transformed variable *time*^{2}.



$y = 1560.+-4.90125\cdot x$
$r^2 = 0.999996$

8 Write down the regression equation in terms of the variables *height* and *time*^{2}.

$height = 1560 - 4.90 \times time^2$

9 Substitute 3.4 for *time* in the equation to find the height after 3.4 seconds.

$height = 1560 - 4.90 \times 3.4^2 = 1503$ m

ISBN 978-1-009-11041-9

Cambridge University Press

CAS 1: Using the CASIO Classpad to perform a squared transformation

The table shows the height (in m) of a base jumper for the first 10 seconds of her jump.

Time	0	1	2	3	4	5	6	7	8	9	10
Height	1560	1555	1540	1516	1482	1438	1383	1320	1246	1163	1070

a Construct a scatterplot displaying *height* (the RV) against *time* (the EV).

b Apply an *x*-squared transformation and fit a least squares line to the transformed data.

c Use the regression line to predict the height of the base jumper after 3.4 seconds.

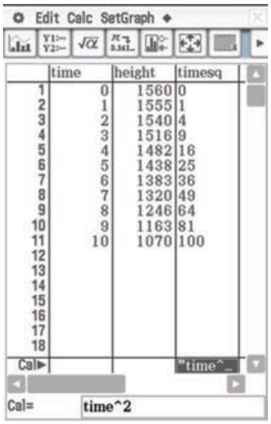
Steps

1 In the Statistics application enter the data into lists named *time* and *height*.

2 Name the third list *timesq* (short for *time* squared).

3 Place the cursor in the calculation cell at the bottom of the third column and type *time*^2. This will calculate the values of *time*^{2}.

Let *time* be the explanatory variable (x) and *height* the response variable (y).

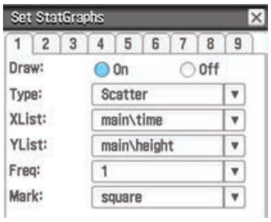


	time	height	timesq
1	0	1560	0
2	1	1555	1
3	2	1540	4
4	3	1516	9
5	4	1482	16
6	5	1438	25
7	6	1383	36
8	7	1320	49
9	8	1246	64
10	9	1163	81
11	10	1070	100
12			
13			
14			
15			
16			
17			
18			
Cal▸			"time^...

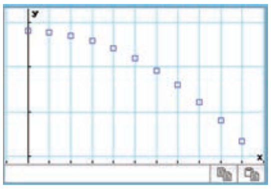
Cal= time^2

4 Construct a scatterplot of *height* against *time*.

- Tap 📊 and complete the **Set StatGraphs** dialog box as shown.
- Tap 📊 to view the scatterplot. The plot is clearly non-linear.

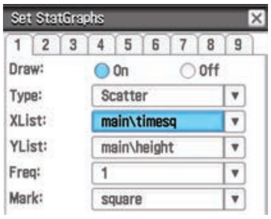


Draw: ● On ○ Off
Type: Scatter
XList: main\time
YList: main\height
Freq: 1
Mark: square

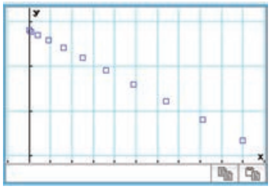


5 Construct a scatterplot of *height* against *time*^{2}.

- Tap 📊 and complete the **Set StatGraphs** dialog box as shown.
- Tap 📊 to view the scatterplot. The plot is now clearly linear.



Draw: ● On ○ Off
Type: Scatter
XList: main\timesq
YList: main\height
Freq: 1
Mark: square

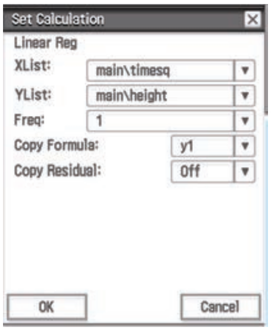


6 Fit a regression line to the transformed data.

- Go to **Calc, Regression, Linear Reg**.
- Complete the **Set Calculation** dialog box as shown and tap **OK**.

Note: The '*x*' in the linear equation corresponds to the transformed variable *time*^{2}.

- Tap **OK** a second time to plot and display the regression line on the scatterplot.



Set Calculation

Linear Reg

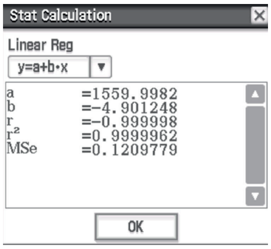
XList: main\timesq

YList: main\height

Freq: 1

Copy Formula: y1

Copy Residual: Off

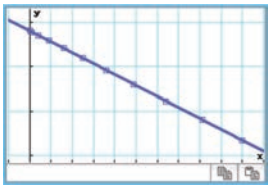


Stat Calculation

Linear Reg

y=a+b·x

```
a    =1559.9982
b    =-4.901248
r    =-0.999998
r²   =0.9999962
MSe  =0.1209779
```



7 Write down the equation in terms of *height* and *time*^{2}. $\quad$ $height = 1560 - 4.90 \times time^2$.

8 Substitute 3.4 for *time* in the equation. $\quad$ $height = 1560 - 4.90 \times 3.4^2 = 1503$ m

Skill-sheet

Exercise 4A

The *x*-squared transformation: some prerequisite skills

1 Evaluate *y* in the following expression, rounded to one decimal place.

a $y = 7 + 8x^2$ when $x = 1.25$ $\quad$ **b** $y = 7 + 3x^2$ when $x = 1.25$

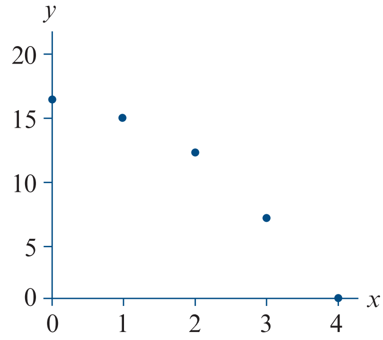
c $y = 24.56 - 0.47x^2$ when $x = 1.23$ $\quad$ **d** $y = -4.75 + 5.95x^2$ when $x = 4.7$

The *x*-squared transformation: calculator exercises

Example 1

2 The scatterplot opposite was constructed from the data in the table below.

x	0	1	2	3	4
y	16	15	12	7	0



a Linearise the scatterplot by applying an *x*-squared transformation and fit a least squares line to the transformed data.

b Give its equation.

c Use the equation to predict the value of *y* when $x = -2$.

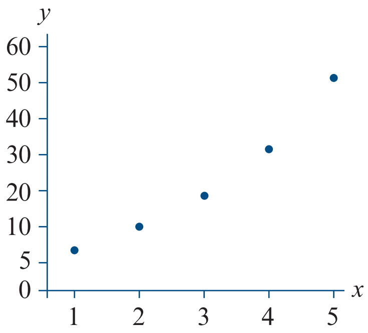
ISBN 978-1-009-11041-9

Cambridge University Press

3 The scatterplot opposite was constructed from the data in the table below.

x	1	2	3	4	5
y	3	9	19	33	51

From the scatterplot, the association between y and x is non-linear.



a Linearise the scatterplot by applying an x-squared transformation and fit a least squares line to the transformed data.

b Give its equation.

c Use the equation to predict the value of y when $x = 6$.

The *y*-squared transformation: some prerequisite skills

4 Evaluate y in the following expression. Give the answers rounded to one decimal place.

a $y^2 = 16 + 4x$ when $x = 1.57$

b $y^2 = 1.7 – 3.4x$ when $x = 0.03$

c $y^2 =16 + 2x$ when $x = 10$ $(y > 0)$

d $y^2 = 58 + 2x$ when $x = 3$ $(y < 0)$

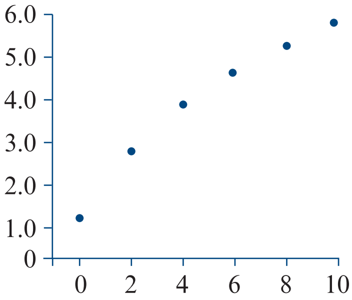
The *y*-squared transformation: calculator exercises

Example 2

5 The scatterplot opposite was constructed from the data in the table below.

x	0	2	4	6	8	10
y	1.2	2.8	3.7	4.5	5.1	5.7

From the scatterplot, the association between y and x is non-linear.



a Linearise the scatterplot by applying a y-squared transformation and fit a least squares line to the transformed data.

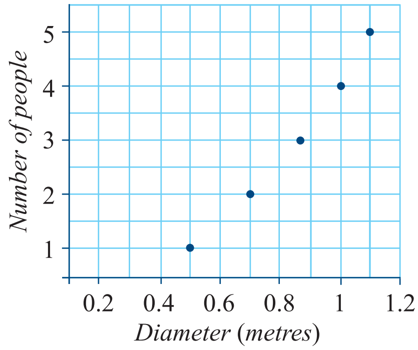
b Give its equation. Write the coefficient, rounded to two significant figures.

c Use the equation to predict the value of y when $x = 9$. Give the answer rounded to one decimal place.

Applications of the squared transformation

6 The table gives the *diameter* (in m) of five different umbrellas and the *number of people* each umbrella is designed to keep dry. A scatter plot is also shown.

Diameter	*Number*
0.50	1
0.70	2
0.85	3
1.00	4
1.10	5



a Apply the squared transformation to the variable *diameter* and determine the least squares regression line for the transformed data. *Diameter* is the EV.

Write the slope and intercept of this line, rounded to one decimal place, in the spaces provided.

$$number = \boxed{} + \boxed{} \times diameter^2$$

b Use the equation to predict the number of people who can be sheltered by an umbrella of diameter 1.3 m. Give your answer rounded to the nearest person.

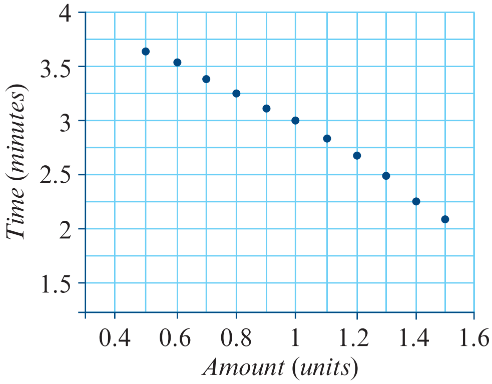
7 The time (in minutes) taken for a local anaesthetic to take effect is associated with to the amount administered (in units). To investigate this association a researcher collected the following data.

Amount	0.5	0.6	0.7	0.8	0.9	1.0	1.1	1.2	1.3	1.4	1.5
Time	3.7	3.6	3.4	3.3	3.2	3.0	2.9	2.7	2.5	2.3	2.1

The association between the variables *amount* and *time* is non-linear as can be seen from the scatterplot below. A squared transformation applied to the variable *time* will linearise the scatterplot.

a Apply the squared transformation to the variable *time* and fit a least squares regression line to the transformed data. *Amount* is the EV. Write the equation of this line with the slope and intercept rounded to two significant figures.

b Use the equation to predict the time for the anaesthetic to take effect when the dose is 0.4 units. Give the answer rounded to one decimal place.



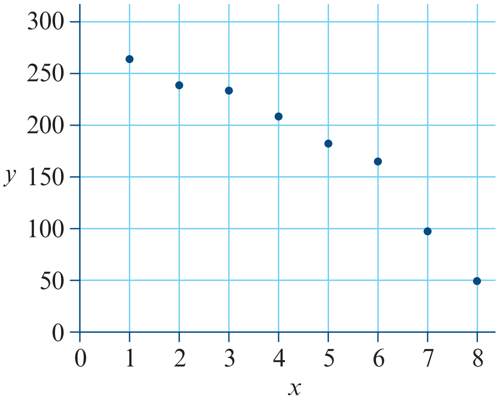
ISBN 978-1-009-11041-9 Cambridge University Press

Exam 1 style questions

The following information relates to Questions 8 and 9

A student uses the data in the table below to construct the scatterplot shown:

x	y
1	264
2	239
3	234
4	208
5	182
6	164
7	98
8	49



8 A squared transformation is applied to x to linearise the association. A least squares line is fitted to the transformed data, with x^2 as the explanatory variable.

The equation of this least squares line is closest to

A $y = 310 - 29.1x^2$

B $y = 10.2 - 0.032x^2$

C $y = 1.80 - 0.106x^2$

D $y = 263 - 3.26x^2$

E $y = 79.9 - 0.303x^2$

9 A y^2 transformation could also be used to linearise this association. A least squares line is fitted to the transformed data, with y^2 as the response variable, and the equation of the least squares line is

$$y^2 = 79973 - 9533.4x$$

Using this equation, the value of y when $x = 4$ is closest to:

A 205 **B** 208 **C** 247 **D** 531 **E** 41839

10 The association between the *cost* of a certain precious stone (in $) and its *weight* (in mg) is non-linear. A squared transformation was applied to the explanatory variable *weight*, and a least squares line fitted to the transformed data. The equation of the least squares line is:

$$\textit{cost} = 2370 + 0.238 \times \textit{weight}^2$$

Using this equation, the cost of a precious stone weighing 75mg is closest to:

A $2389 **B** $3709 **C** $2689 **D** $7995 **E** $177,768

ISBN 978-1-009-11041-9

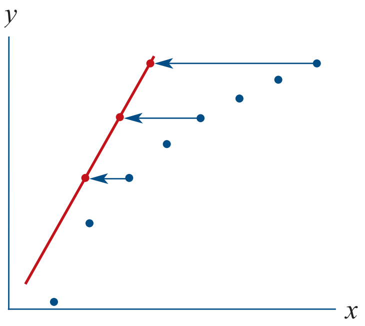
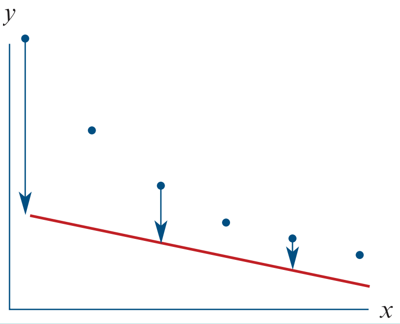
Cambridge University Press

4B The log transformation

Learning intentions

- To be able to apply a $\log_{10}$ transformation to either x or y.
- To be able to fit a least squares regression line to the transformed data.
- To be able to use the least squares regression line fitted to the transformed data for prediction.

You will recall from Chapter 1 that the shape of a highly skewed single variable distribution could be changed to become more symmetric by changing the scale from x to $\log_{10}x$. When applied to bivariate data, the effect of the **logarithmic transformation** is to again to *compress* the upper end of the scale on either the x- or the y-axis, potentially linearising a non-linear association. The effect of applying the $\log_{10}x$ and $\log_{10}y$ transformations (separately) to a scatterplot are illustrated graphically below.

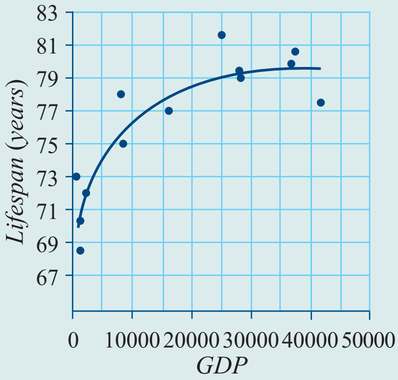
Transformation	*Outcome*	*Graph*
$\log_{10}x$	Compresses the higher x-values relative to the lower x-values, leaving the y-values unchanged. This has the effect of straightening out curves like the one shown.	
$\log_{10}y$	Compresses larger y values relative to the smaller y values This has the effect of straightening out curves like the one shown.	

Following the normal convention, $\log x$ means $\log_{10}x$.

Example 3 Applying the log x transformation

The general wealth of a country, often measured by its Gross Domestic Product (*GDP*), is one of several variables associated with *lifespan* in different countries. However, the association is not linear, as can be seen in the scatterplot below which plots *lifespan* (in years) against *GDP* per person (in dollars) for 13 different countries.

The scatterplot shows that there is a strong positive association between the *lifespan* and *GDP*.



a Apply a log transformation to the variable *GDP*, and determine the least squares regression line for the transformed data.

b Use the least squares equation to predict the *lifespan* of a country with a *GDP* of \$20 000 per person, giving your answer rounded to one decimal place.

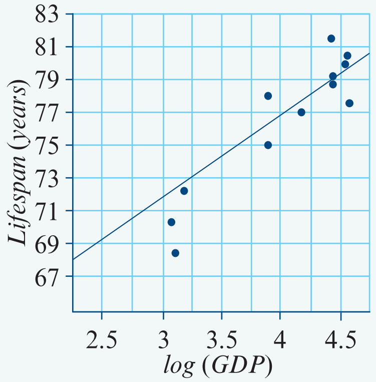
Solution

a Applying the log x transformation involves changing the scale on the x-axis to log(*GDP*).

When we make this change, we see that the association between the variables *lifespan* and log (*GDP*) is linear. See the plot opposite.

Note: On the plot, when log (*GDP*) = 4, the actual GDP is 10^4 or \$10 000.

We can now fit a least squares line to model the association between the variables *lifespan* and log(*GDP*).



The equation of this line is:

$$lifespan = 54.3 + 5.59 \times \log(GDP)$$

b Using this equation, for a country with a GDP of \$20 000, the lifespan is predicted as:

$$lifespan = 54.3 + 5.59 \times \log 20\,000 = 78.3 \text{ years (to one decimal place)}$$

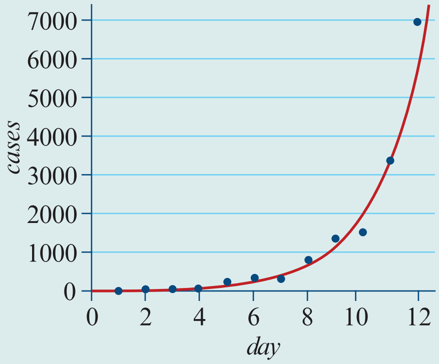
Example 4 Applying the log *y* transformation

The numbers of cases of a very infectious disease were recorded over a 12 day period. The association is not linear, as can be seen in the scatterplot below which plots *cases* against *days*.

The scatterplot shows that there is a strong positive association between the number of *case* and *day*.

a Apply a log transformation to the variable *cases*, and determine the least squares regression line for the transformed data.

b Use the least squares equation to predict the *cases* on day 13.



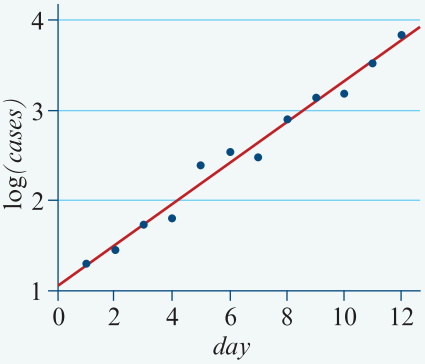
Solution

a Applying the log y transformation involves changing the scale on the y-axis to log(*cases*).

When we make this change, we see from the plot the association between the variables log(*cases*) and *day* is linear.

Note: On the plot, when log (*cases*) =3, the actual number of cases is 10^3 or 1000.

We can now fit a least squares line to model the association between the variables log(*cases*) and *day*.



The equation of this line is:

$$\log(\textit{cases}) = 1.046 + 0.227 \times \textit{day}$$

b Using this equation, on day 13 the number of cases is predicted as:

$$\log(\textit{cases}) = 1.046 + 0.227 \times 13 = 3.997$$

To find the number of cases we use the calculator to evaluate $10^{3.997}$ = 9931 cases (to the nearest whole number).

CAS 2: Using the TI-Nspire CAS to perform a log transformation

The table shows the *lifespan* (in years) and *GDP* (in dollars) of people in 12 countries. The association is non-linear. Using the log *x* transformation:

- linearise the data, and fit a regression line to the transformed data (*GDP* is the EV)
- write its equation in terms of the variables *lifespan* and *GDP* rounded to three significant figures.
- use the equation of the regression line to predict the lifespan in a country with a GDP of $20 000, rounded to one decimal place.

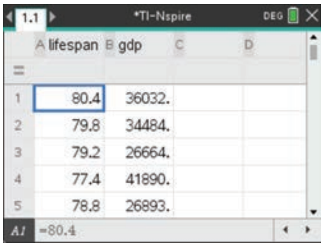
Lifespan	*GDP*
80.4	36 032
79.8	34 484
79.2	26 664
77.4	41 890
78.8	26 893
81.5	25 592
74.9	7 454
72.0	1 713
77.9	7 073
70.3	1 192
73.0	631
68.6	1 302

Steps

1 Start a new document by pressing ctrl + N .

2 Select **Add Lists & Spreadsheet**.
Enter the data into lists named *lifespan* and *gdp*.

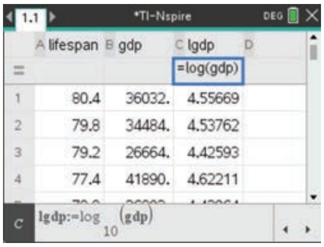
3 Name column C as *lgdp* (short for log (*GDP*)).
Now calculate the values of log (*GDP*) and store them in the list named *lgdp*.



	A lifespan	B gdp	C	D
=				
1	80.4	36032.		
2	79.8	34484.		
3	79.2	26664.		
4	77.4	41890.		
5	78.8	26893.		

A1 =80.4

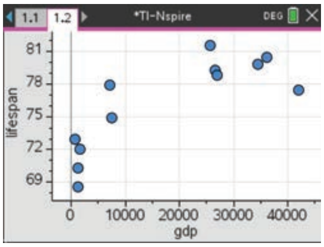
4 Move the cursor to the formula cell below the *lgdp* heading.
We need to enter the expression = **log(gdp)**.
To do this, press = then type in **log(gdp)**. Pressing enter calculates and displays the values of *lgdp*.



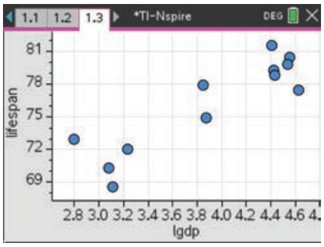
	A lifespan	B gdp	C lgdp	D
=			=log(gdp)	
1	80.4	36032.	4.55669	
2	79.8	34484.	4.53762	
3	79.2	26664.	4.42593	
4	77.4	41890.	4.62211	

C lgdp:=$\log_{10}$(gdp)

5 Press ctrl + I and select **Add Data & Statistics**.
Construct a scatterplot of *lifespan* against *GDP*. Let *GDP* be the explanatory variable and *lifespan* the response variable. The plot is clearly non-linear.

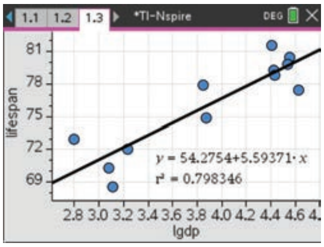


6 Press ctrl + I and select **Add Data & Statistics**. Construct a scatterplot of *lifespan* against log(*GDP*). The plot is now clearly linear.



7 Press menu **>Analyze>Regression>Show Linear (a + bx)** to plot the line on the scatterplot with its equation.

Note: The x in the equation on the screen corresponds to the transformed variable log (*GDP*).



$y = 54.2754+5.59371 \cdot x$
$r^2 = 0.798346$

8 Write the regression equation in terms of the variables *lifespan* and log (*GDP*).

$lifespan = 54.3 + 5.59 \times \log(GDP)$

9 Substitute 20 000 for *GDP* in the equation to find the lifespan of people in a country with GDP of $20 000.

$$\begin{aligned} lifespan &= 54.3 + 5.59 \times \log 20\,000 \\ &= 78.3 \text{ years} \end{aligned}$$

CAS 2: Using the CASIO Classpad to perform a log transformation

The table shows the *lifespan* (in years) and *GDP* (in dollars) of people in 12 countries. The association is non-linear. Using the log x transformation:

- linearise the data, and fit a regression line to the transformed data (*GDP* is the EV)
- write its equation in terms of the variables *lifespan* and *GDP* rounded to three significant figures.
- use the equation to predict the lifespan in a country with a GDP of $20 000 rounded to one decimal place.

Lifespan	*GDP*
80.4	36 032
79.8	34 484
79.2	26 664
77.4	41 890
78.8	26 893
81.5	25 592
74.9	7 454
72.0	1 713
77.9	7 073
70.3	1 192
73.0	631
68.6	1 302

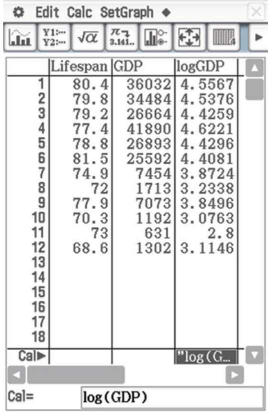
ISBN 978-1-009-11041-9 Cambridge University Press

Steps

1 In the Statistics application enter the data into lists named *Lifespan* and *GDP*.

2 Name the third list *logGDP*.

3 Place the cursor in the calculation cell at the bottom of the third column and type **log (GDP)**.
Let *GDP* be the explanatory variable (x) and *lifespan* the response variable (y).

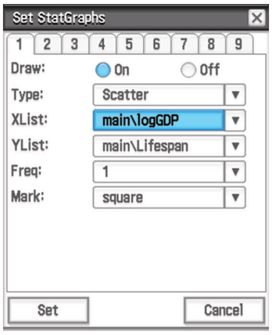


	Lifespan	GDP	logGDP
1	80.4	36032	4.5567
2	79.8	34484	4.5376
3	79.2	26664	4.4259
4	77.4	41890	4.6221
5	78.8	26893	4.4296
6	81.5	25592	4.4081
7	74.9	7454	3.8724
8	72	1713	3.2338
9	77.9	7073	3.8496
10	70.3	1192	3.0763
11	73	631	2.8
12	68.6	1302	3.1146
13			
14			
15			
16			
17			
18			
Cal▶			"log(G...

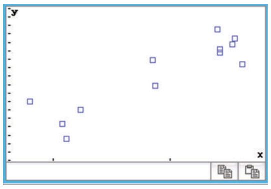
Cal= log(GDP)

4 Construct a scatterplot of *lifespan* against *log (GDP)*.

- Tap and complete the **Set StatGraphs** dialog box as shown.
- Tap to view the scatterplot.
- The plot is linear.

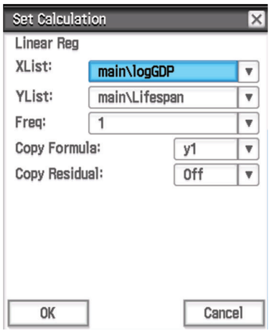


Draw: ● On ○ Off
Type: Scatter
XList: main\logGDP
YList: main\Lifespan
Freq: 1
Mark: square

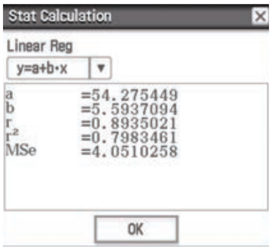


5 To find the least squares regression equation and fit a regression line to the transformed data.

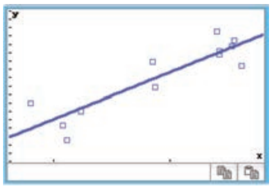
- Go to **Calc, Regression, Linear Reg**.
- Complete the **Set Calculation** dialog box as shown and tap **OK**.
  This generates the regression results.
  Note: The x in the linear equation corresponds to the transformed variable log (GDP).



Linear Reg
XList: main\logGDP
YList: main\Lifespan
Freq: 1
Copy Formula: y1
Copy Residual: Off



Linear Reg
y=a+b·x
a =54.275449
b =5.5937094
r =0.8935021
r² =0.7983461
MSe =4.0510258



- Tap **OK** a second time to plot and display the regression line on the scatterplot.

6 Write the equation in terms of *lifespan* and log (*GDP*).

$$lifespan = 54.3 + 5.59 \times log\ (GDP)$$

7 Substitute 20 000 for *GDP* in the equation.

$$lifespan = 54.3 + 5.59 \times log\ 20\,000$$
$$= 78.3 \text{ years}$$

ISBN 978-1-009-11041-9 Cambridge University Press

Skill-sheet

Exercise 4B

The log *x* transformation: some prerequisite skills

1 Evaluate the following expressions rounded to one decimal place.

a $y = 5.5 + 3.1 \log 2.3$ **b** $y = 0.34 + 5.2 \log 1.4$

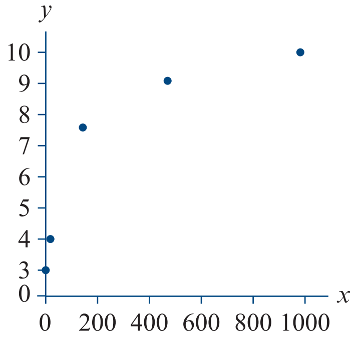
c $y = -8.5 + 4.12 \log 20$ **d** $y = 196.1 - 23.2 \log 303$

The log *x* transformation: calculator exercise

Example 3

2 The scatterplot opposite was constructed from the data in the table below.

x	5	10	150	500	1000
y	3.1	4.0	7.5	9.1	10.0



From the scatterplot, it is clear that the association between y and x is non-linear.

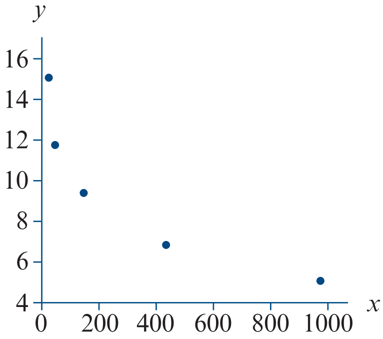
a Linearise the scatterplot by applying a log x transformation and fit a least squares line to the transformed data.

b Write down the equation, with the coefficients rounded to one significant figure.

c Use the equation to predict the value of y when $x = 100$.

3 The scatterplot opposite was constructed from the data in the table below.

x	10	44	132	436	981
y	15.0	11.8	9.4	6.8	5.0



From the scatterplot, it is clear that the association between y and x is non-linear.

a Linearise the scatterplot by applying a log x transformation and fit a least squares line to the transformed data.

b Write down the equation, with the coefficients rounded to one significant figure.

c Use the equation to predict the value of y when $x = 1000$.

The log *y* transformation: some prerequisite skills

4 Find the value of y in the following, rounded to one decimal place if not exact.

a $\log y = 2$ **b** $\log y = 2.34$

c $\log y = 3.5 + 2x$ where $x = 1.25$ **d** $\log y = -0.5 + 0.024x$ where $x = 17.3$

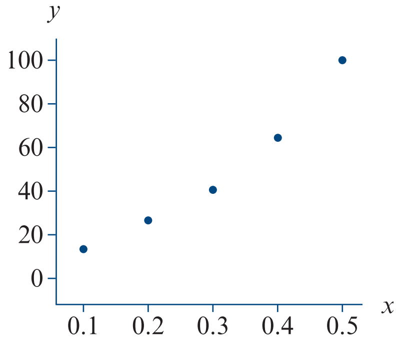
ISBN 978-1-009-11041-9 Cambridge University Press

The log *y* transformation: calculator exercise

Example 4

5 The scatterplot opposite was constructed from the data in the table below.

x	0.1	0.2	0.3	0.4	0.5
y	15.8	25.1	39.8	63.1	100.0



From the scatterplot, it is clear that the association between y and x is non-linear.

a Linearise the scatterplot by applying a log y transformation and fit a least squares line to the transformed data.

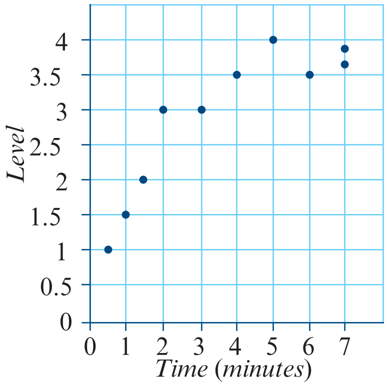
b Write down the equation, with the coefficients rounded to one significant figure.

c Use the equation to predict the value of y when $x = 0.6$, rounded to one decimal place.

Applications of the log transformation

6 The table below shows the level of performance level achieved by 10 people on completion of a task. Also shown is the time spent (in minutes) practising the task. In this situation, *time* is the EV. The association between the *level* and *time* is non-linear as seen in the scatterplot.

Time	*Level*
0.5	1
1	1.5
1.5	2
2	3
3	3
4	3.5
5	4
6	3.5
7	3.9
7	3.6



A log transformation can be applied to the variable *time* to linearise the scatterplot.

a Apply the log transformation to the variable *time* and fit a least squares line to the transformed data. log (*time*) is the EV.

Write the slope and intercept of this line, rounded to two significant figures, in the spaces provided.

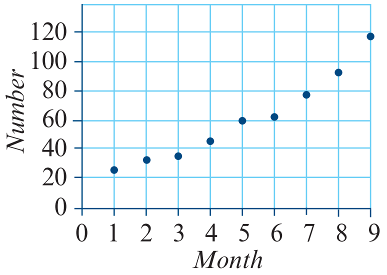
$$level = \boxed{} + \boxed{} \times \log(time)$$

b Use the equation to predict the level of performance (rounded to one decimal place) for a person who spends 2.5 minutes practising the task.

ISBN 978-1-009-11041-9 Cambridge University Press

7 The table below shows the number of internet users signing up with a new internet service provider for each of the first nine months of their first year of operation. A scatterplot of the data is also shown.

Month	*Number*
1	24
2	32
3	35
4	44
5	60
6	61
7	78
8	92
9	118



The association between *number* and *month* is non-linear.

a Apply the log transformation to the variable *number* and fit a least squares line to the transformed data. *Month* is the EV.

Write the slope and intercept of this line, rounded to four significant figures, in the spaces provided.

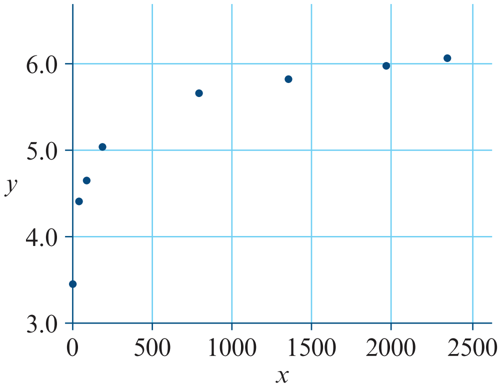
log (*number*) = □ + □ × *month*

b Use the equation to predict the *number* of internet users after 10 months. Give answer to the nearest whole number.

Exam 1 style questions

8 A student uses the data in the table below to construct the scatterplot shown.

x	y
5	3.45
44	4.41
94	4.64
187	5.03
791	5.65
1350	5.81
1960	5.97
2345	6.06



A log transformation is applied to x to linearise the association. A least squares line is fitted to the transformed data, with log x as the explanatory variable.

ISBN 978-1-009-11041-9

Cambridge University Press

The equation of this least squares line is closest to

A $y = 4.43 + 0.001 \log x$

B $y = -3570 + 861 \log x$

C $y = 2.78 + 0.976 \log x$

D $y = -2.85 + 1.03 \log x$

E $y = 0.642 + 0.00724 \log x$

9 The association between the power of a car (in horsepower) and the *time* it takes to accelerate from 0 to 100 km/hr (in seconds) is non-linear. A log transformation was applied to the explanatory variable *horsepower*, and a least squares line fitted to the transformed data. The equation of the least squares line is:

$$time = 42.7 - 13.9 \times \log(horsepower)$$

Using this equation, the time it would take for a car with 180 horsepower to accelerate from 0 to 100km/hr is closest to:

A 29.5 seconds **B** 65 seconds **C** 28.8 seconds **D** 11.3 seconds **E** 11.4 seconds

10 The price of shares in a newly formed technology company *price* has increased non-linearly since the company was formed 12 months ago. A log transformation was applied to the maximum share price each month (*share price*), and a least squares line fitted to the transformed data, with *month* as the explanatory variable. The equation of the least squares line is:

$$\log(shareprice) = 1.39 + 0.050 \times month$$

Using this equation, the maximum monthly share price in month 14 is closest to:

A \$123.03 **B** \$2.09 **C** \$8.08 **D** \$20.16 **E** \$25.25

4C The reciprocal transformation

Learning intentions

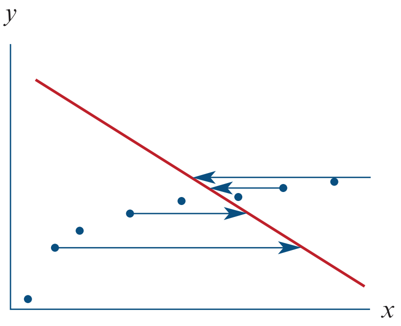
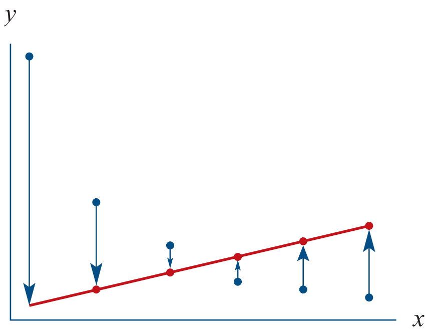
- To be able to apply a reciprocal transformation to either x or y.
- To be able to fit a least squares regression line to the transformed data.
- To be able to use the least squares regression line fitted to the transformed data for prediction.

The **reciprocal transformation** is a stretching transformation that compresses the upper end of the scale on either the x- or y-axis.

ISBN 978-1-009-11041-9

Cambridge University Press

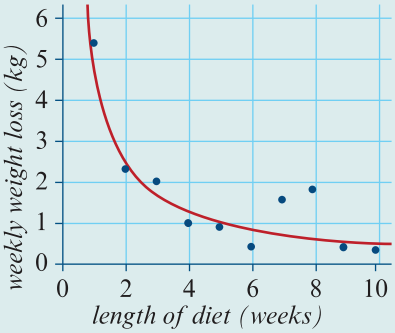
The effect of applying a reciprocal y transformation to a scatterplot is as follows:

Transformation	*Outcome*	*Graph*
$\frac{1}{x}$	Compresses larger x values relative to the smaller data values, but to a greater extent than log x. This has the effect of straightening out curves like the one shown opposite. Note that values of x less than one become greater than 1, and values of x greater than 1 become less than 1, so that the order of the data values is reversed.	
$\frac{1}{y}$	Compresses larger values of y relative to lower values of y. This has the effect of straightening out curves like the one shown opposite. Again, the order of the data is reversed.	

The following example shows how the $1/x$ transformation works in practice.

Example 5 Applying the reciprocal $(1/x)$ transformation

After embarking on a new healthy eating and exercise plan, Ben recorded his weekly weight loss over a 10 week. The association is not linear, as can be seen in the scatterplot below which plots *weekly weight loss* in kg against *length of diet* in weeks.



The scatterplot shows that there is a strong negative association between *weekly weight loss* and *length of diet*.

a Apply a reciprocal transformation to the variable *length of diet*, and determine the least squares regression line for the transformed data.

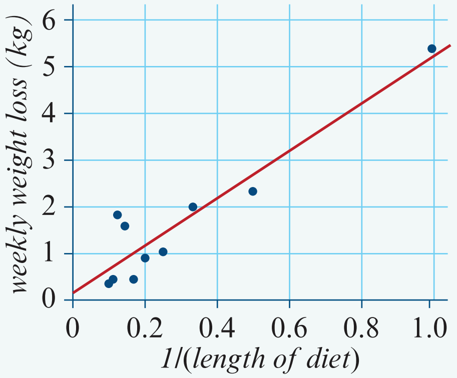
b Use the least squares equation to predict the *weekly weight loss* in week 11, giving your answer to one decimal place.

Solution

a Applying the $1/x$ transformation involves changing the scale on the x-axis to 1/(*length of diet*).

When we make this change, we see from the plot the association between the variables *weekly weight loss* and 1/(*length of diet*) is linear.

We can now fit a least squares line to model the association between the variables *weekly weight loss* and 1/(*length of diet*).



The equation of this line is:

$$\textit{weekly weight loss} = 0.13 + \frac{5.09}{\textit{length of diet}}$$

b Using this equation, in week 11 the weight loss predicted is:

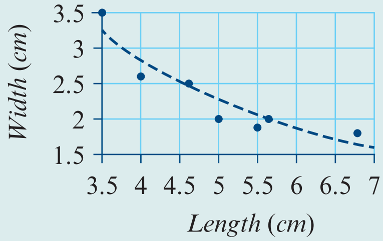
$$\textit{weekly weight loss} = 0.13 + \frac{5.09}{11} = 0.6 \text{ kg}$$

The following example shows how the $1/y$ transformation works in practice.

Example 6 **Applying the reciprocal ($1/y$) transformation**

A homeware company makes rectangular sticky labels with a variety of lengths and widths.

The scatterplot opposite displays the *width* (in cm) and *length* (in cm) of eight of the sticky labels.



The scatterplot shows that there is a strong negative association between the width of the sticky labels and their lengths, but it is clearly non-linear.

a Apply a reciprocal transformation to the variable *width*, and determine the least squares regression line for the transformed data.

b Use the least squares equation to predict the *width* of a sticky label which is 5 cm long, giving your answer to two decimal places.

Solution

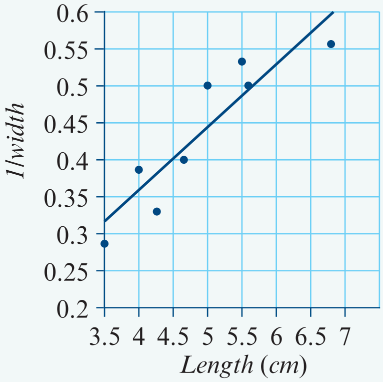
a Applying the $1/y$ transformation involves changing the scale on the y-axis from *width* to $1/(width)$.

When we make this change, we see from the scatterplot that the association between 1/*width* and *length* is linear.

We can now fit a least squares line to model the association between 1/*width* and *length*.

The equation of this line is:

$$1/width = 0.015 + 0.086 \times \ length$$



b For a sticky label of length 5 cm, we would predict that:

$$1/width = 0.015 + 0.086 \times 5 = 0.445$$

$$\text{or } width = \frac{1}{0.445} = 2.25 \text{ cm}$$

CAS 3: Using the TI-Nspire CAS to perform a reciprocal transformation

The table shows the length (in cm) and width (in cm) of eight sizes of sticky labels.

Length	6.8	5.6	4.6	4.2	3.5	4.0	5.0	5.5
Width	1.8	2.0	2.5	3.0	3.5	2.6	2.0	1.9

Using the $1/y$ transformation:

- linearise the data, and fit a regression line to the transformed data (*length* is the EV)
- write its equation in terms of the variables *length* and *width*
- use the equation to predict the width of a sticky label with a length of 5 cm.

Steps

1 Start a new document by pressing ctrl + N.

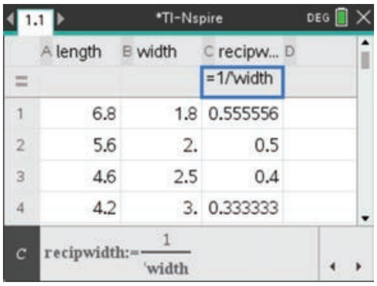
2 Select **Add Lists & Spreadsheet**.

Enter the data into lists named *length* and *width*.

3 Name column C as *recipwidth* (short for 1/width).

Calculate the values of *recipwidth*.

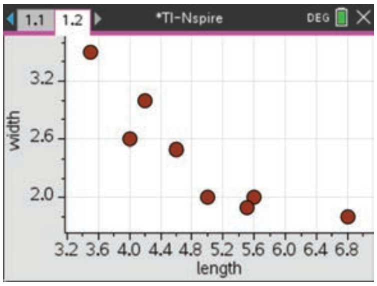
Move the cursor to the formula cell below the *recipwidth* heading. Type in =**1/width**. Press enter to calculate the values of *recipwidth*.



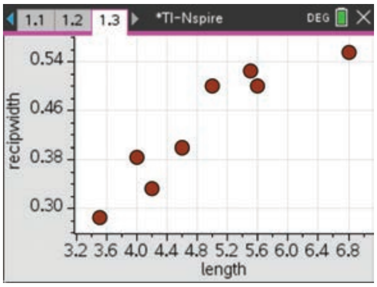
	A length	B width	C recipw...	D
=			=1/width	
1	6.8	1.8	0.555556	
2	5.6	2.	0.5	
3	4.6	2.5	0.4	
4	4.2	3.	0.333333	

C $recipwidth := \frac{1}{width}$

4 Press ctrl + I and select **Add Data & Statistics**. Construct a scatterplot of *width* against *length*. Let *length* be the explanatory variable and *width* the response variable. The plot is clearly non-linear.

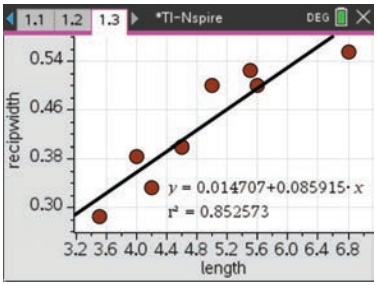


5 Press ctrl + I and select **Add Data & Statistics**. Construct a scatterplot of *recipwidth* ($1/width$) against *length*. The plot is now clearly linear.



6 Press menu**>Analyze>Regression>Show Linear (a + bx)** to plot the line on the scatterplot with its equation.

Note: The y in the equation on the screen corresponds to the transformed variable $1/width$.



$y = 0.014707 + 0.085915 \cdot x$

$r^2 = 0.852573$

7 Write down the regression equation in terms of the variables *width* and *length*.

$1/width = 0.015 + 0.086 \times length$

8 Substitute 5 cm for *length* in the equation.

$1/width = 0.015 + 0.086 \times 5 = 0.445$

Thus width $= 1/0.445 = 2.25$ cm (to 2 d.p.)

CAS 3: Using the CASIO Classpad to perform a reciprocal transformation

The table shows the length (in cm) and width (in cm) of eight sizes sticky labels.

Length	6.8	5.6	4.6	4.2	3.5	4.0	5.0	5.5
Width	1.8	2.0	2.5	3.0	3.5	2.6	2.0	1.9

Using the $1/y$ transformation:

- linearise the data, and fit a regression line to the transformed data. *Length* is the EV.
- write its equation in terms of the variables *length* and *width*.
- use the equation to predict the width of a sticky label with length of 5 cm.

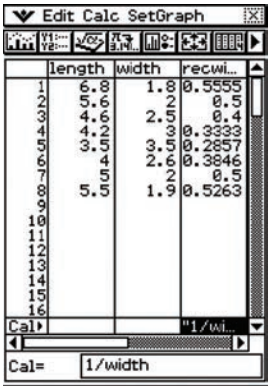
Steps

1 Open the Statistics application and enter the data into lists named *length* and *width*.

2 Name the third list *recwidth* (short for reciprocal width).

3 Place the cursor in the calculation cell at the bottom of the third column and type *1/width*. This will calculate all the reciprocal values of the width.

Let *length* be the explanatory variable (x) and *width* the response variable (y).



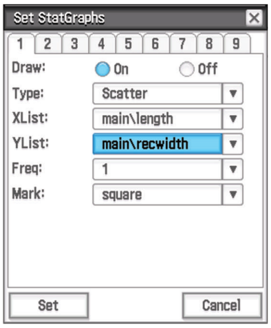
	length	width	recwi...
1	6.8	1.8	0.5555
2	5.6	2	0.5
3	4.6	2.5	0.4
4	4.2	3	0.3333
5	3.5	3.5	0.2857
6	4	2.6	0.3846
7	5	2	0.5
8	5.5	1.9	0.5263
9			
10			
11			
12			
13			
14			
15			
16			
Cal▸			"1/wi...

Cal= 1/width

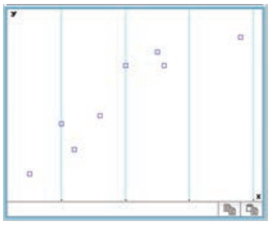
4 Construct a scatterplot of *1/width* against *length*.

- Tap and complete the **Set StatGraphs** dialog box as shown.
- Tap to view the scatterplot.

  The plot is now clearly linear.



Draw: ● On ○ Off
Type: Scatter
XList: main\length
YList: main\recwidth
Freq: 1
Mark: square

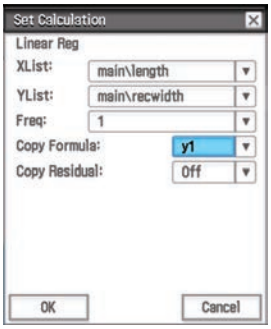


5 Fit a regression line to the transformed data.

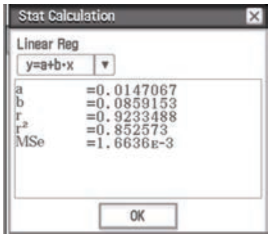
- Go to **Calc, Regression, Linear Reg**.
- Complete the **Set Calculation** dialog box as shown and tap **OK**.

  This generates the regression results.

  Note: The y in the linear equation corresponds to the transformed variable $1/width$; that is $1/y$.

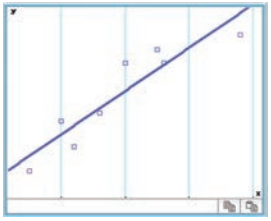


Linear Reg
XList: main\length
YList: main\recwidth
Freq: 1
Copy Formula: y1
Copy Residual: Off



Linear Reg
y=a+b·x

```
a      =0.0147067
b      =0.0859153
r      =0.9233488
r²     =0.852573
MSe    =1.6636E-3
```



- Tap **OK** a second time to plot and display the line on the scatterplot.

6 Write down the equation in terms of the variables *width* and *length*.

$1/width = 0.015 + 0.086 \times length$

7 Substitute 5 cm for *length* in the equation.

$1/width = 0.015 + 0.086 \times 5 = 0.445$

Thus width = 1/0.445 = 2.25 cm (to 2 d.p.)

Skill-sheet

Exercise 4C

The reciprocal (1/*x*) transformation: some prerequisite skills

1 Evaluate the following expressions rounded to one decimal place.

a $y = 6 + \frac{22}{x}$ when $x = 3$

b $y = 4.9 - \frac{2.3}{x}$ when $x = 1.1$

c $y = 8.97 - \frac{7.95}{x}$ when $x = 1.97$

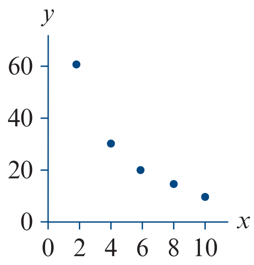
d $y = 102.6 + \frac{223.5}{x}$ when $x = 1.08$

The reciprocal (1/*x*) transformation: calculator exercise

Example 5

2 The scatterplot opposite was constructed from the data in the table below.

x	2	4	6	8	10
y	60	30	20	15	12



From the scatterplot, it is clear that the association between y and x is non-linear.

a Linearise the scatterplot by applying a $1/x$ transformation and fit a least squares line to the transformed data.

b Write down its equation.

c Use the equation to predict the value of y when $x = 5$.

The reciprocal (1/*y*) transformation: some prerequisite skills

3 Find the value of y in the following, rounded to two decimal places.

a $\frac{1}{y} = 3x$ when $x = 2$

b $\frac{1}{y} = 6 + 2x$ when $x = 4$

c $\frac{1}{y} = -4.5 + 2.4x$ when $x = 4.5$

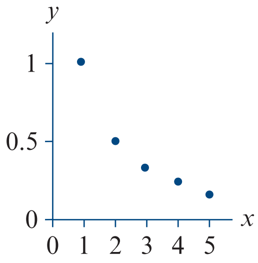
d $\frac{1}{y} = 14.7 + 0.23x$ when $x = 4.5$

The reciprocal (1/*y*) transformation: calculator exercise

Example 6

4 The scatterplot opposite was constructed from the data in the table below.

x	1	2	3	4	5
y	1	0.5	0.33	0.25	0.20



From the scatterplot, it is clear that the association between y and x is non-linear.

a Linearise the scatterplot by applying a $1/y$ transformation and fit a least squares line to the transformed data.

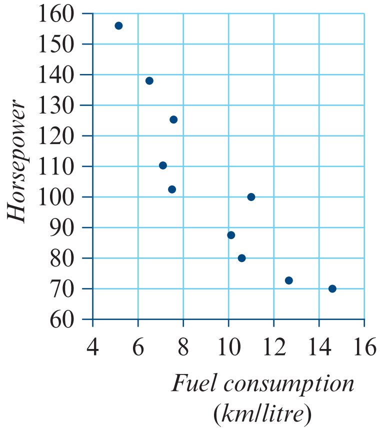
b Write down its equation.

c Use the equation to predict the value of y when $x = 0.25$.

Applications of the reciprocal transformation

5 The table shows the *horsepower* of 10 cars and their *fuel consumption*. From the scatterplot, it is clear that the association between *horsepower* and *fuel consumption* is non-linear.

Fuel consumption	*Horsepower*
5.2	155
7.3	125
12.6	75
7.1	110
6.3	138
10.1	88
10.5	80
14.6	70
10.9	100
7.7	103



a Apply the reciprocal transformation to the variable *fuel consumption* and fit a least squares line to the transformed data. *Horsepower* is the RV.

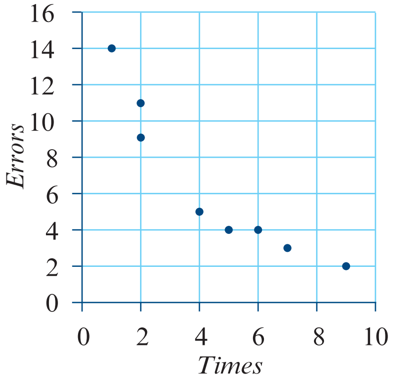
Write the intercept and slope of this line in the boxes provided, rounded to three significant figures.

$$horsepower = \boxed{} + \boxed{} \times \frac{1}{fuel\,consumption}$$

b Use the equation to predict the horsepower of a car with a fuel consumption of 9 km/litre.

6 Ten students were given an opportunity to practise a complex matching task as often as they liked before they were assessed. The number of *times* they practised the task and the number of *errors* they made when assessed are given in the table.

Times	*Errors*
1	14
2	9
2	11
4	5
5	4
6	4
7	3
7	3
9	2



ISBN 978-1-009-11041-9

Cambridge University Press

a Apply the reciprocal transformation to the variable *errors* and determine the least squares regression with the number of times the task was practiced as the EV. Write the intercept and slope of this line in the boxes provided, rounded to two significant figures.

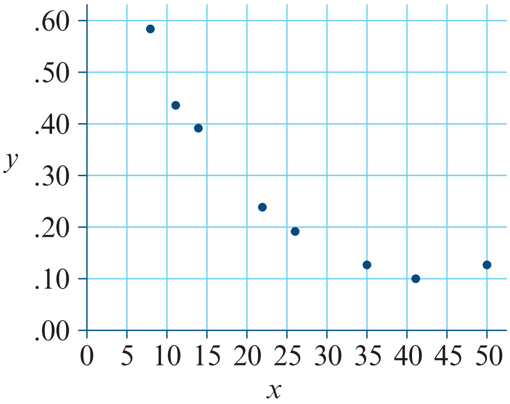
$$\frac{1}{errors} = \boxed{} + \boxed{} \times times$$

b Use the equation to predict the number of errors made when the task is practised six times.

Exam 1 style questions

7 A student used the data in the table below to construct the scatterplot shown

x	y
8	0.58
11	0.43
14	0.39
22	0.24
26	0.19
35	0.13
41	0.10
50	0.12



A reciprocal transformation is applied to y to linearise the association. A least squares line is fitted to the transformed data, with $1/y$ as the response variable. The equation of this least squares line is closest to

A $\frac{1}{y} = 0.198 + 0.196x$

B $\frac{1}{y} = -0.546 - 0.011x$

C $y = 0.013 + \frac{4.665}{x}$

D $\frac{1}{y} = 0.013 + 4.665x$

E $y = 0.546 - 0.011x$

8 The association between score on a problem solving test (*score*) and the number of attempts a person has at the test (*attempts*) is non-linear. A reciprocal transformation was applied to the explanatory variable *attempts*, and a least squares line fitted to the transformed data. The equation of the least squares line is:

$$score = 50 - 22.8 \times \frac{1}{attempts}$$

ISBN 978-1-009-11041-9 Cambridge University Press

Using this equation, the score that a person achieves on their fourth attempt is closest to:

A 6.8 **B** 27.2 **C** 55.7 **D** 41.2 **E** 44.3

9 The price of shares in a newly formed technology company *price* has increased non-linearly since the company was formed 12 months ago. A reciprocal transformation was applied to the maximum share price each month (*share price*), and a least squares line fitted to the transformed data, with *month* as the explanatory variable. The equation of the least squares line is:

$$\frac{1}{share\,price} = 0.0349 - 0.00215 \times month$$

Using this equation, the maximum monthly share price in month 14 is closest to:

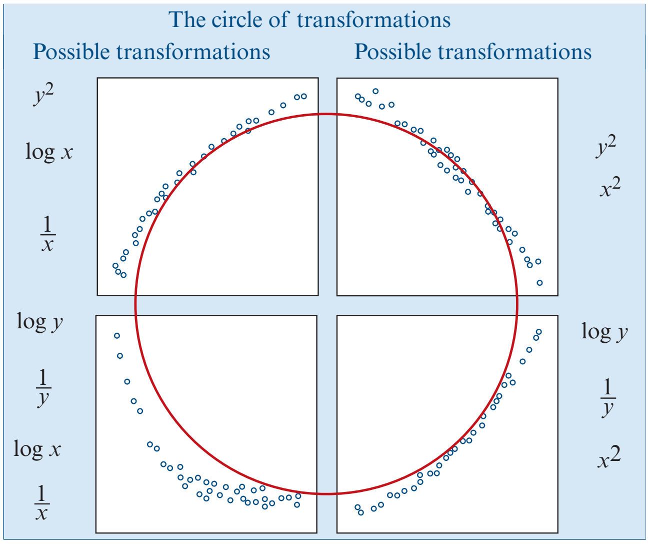
A \$2.18 **B** \$28.78 **C** \$208.33 **D** 0.48 cents **E** 48 cents

4D Choosing and applying the appropriate transformation

Learning intentions

- ▶ To be able to use the circle of transformations to determine which transformations may help linearise a non-linear association.
- ▶ To be able to use a residual plot to assess the effectiveness of a data transformation.
- ▶ To be able to use the coefficient of determination to assess the effectiveness of a data transformation.

The types of scatterplots that can be transformed by the squared, log or reciprocal transformations can be fitted together into what we call the **circle of transformations**.



The purpose of the circle of transformations is to guide us in our choice of transformation to linearise a given scatterplot.

There are two things to note when using the circle of transformations:

1. In each case, there is more than one type of transformation that might work.
2. These transformations only apply to scatterplots with a consistently increasing or decreasing trend.

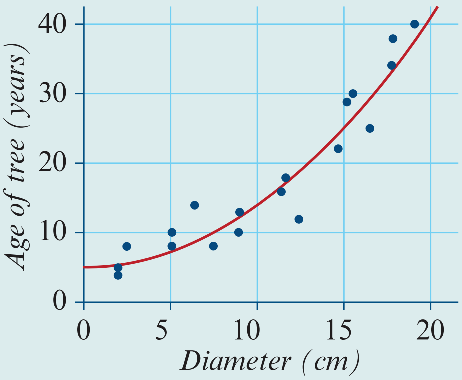
The advantage of having alternatives is that in practice, we can always try each of them to see which gives us the best result. How do we decide which transformation is the best? The best transformation is the one which results in the best linear model. To choose the best linear model we will consider for each transformation applied:

- The residual plot, in order to evaluate the linearity of the transformed association.
- The value of the coefficient of determination, r^2.

This procedure is illustrated in the following example.

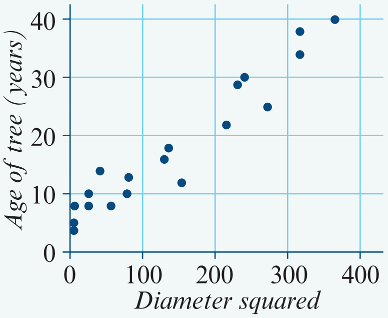
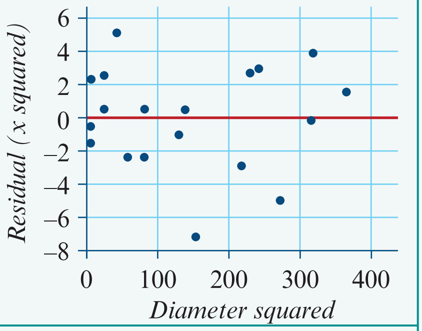
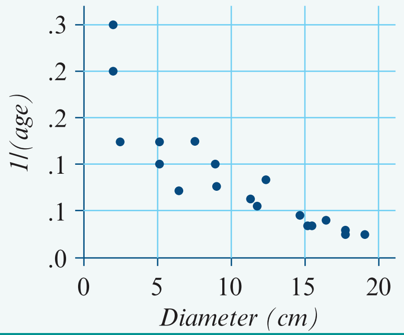
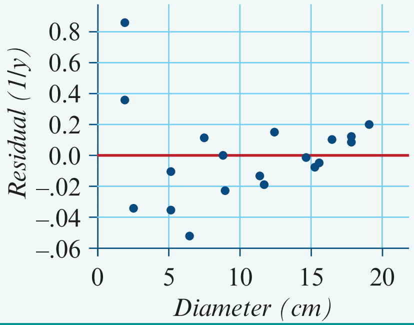
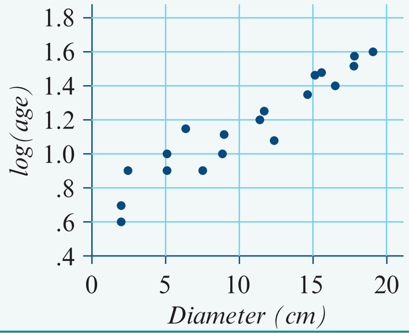
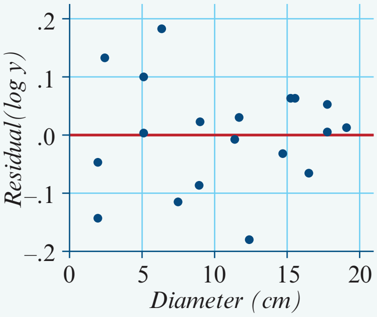
Example 7 Choosing the best transformation

The scatterplot shows the age (in years) and diameter at a height of 1.5 metre (in cm) for a sample of 19 trees of the same species. Use an appropriate transformation to find a regression model which allows the age of this species of tree to be predicted from its diameter.



Solution

The scatterplot has a consistently increasing trend so the circle of transformations applies. Comparing the scatterplot to those in the circle of transformations we see that the x^2, $1/y$ and $\log x$ transformations all have the potential to linearise this scatterplot. All of these transformations have been applied in turn, and the resulting scatterplots and residual plots are shown in the following table.

Transformation	*Scatterplot*	*Residual plot*
x^2		
$\frac{1}{y}$		
log y		

Applying each of these transformations in turn we can see from the residual plots that both the x^2 and the log y transformations have been quite effective in linearising the association between the age of the tree and its diameter. There still seems to be a curve in the residual plot after the the $1/y$ transformation so that has been less effective.

To further help to choose the best transformation we can compare the values of r^2, the coefficient of determination.

- For the x^2 transformation, $r^2 = 92.7\%$
- For the $1/y$ transformation, $r^2 = 75.7\%$
- For the log y transformation, $r^2 = 90.2\%$

Both the x^2 and log y transformations have a very high explanatory power, and either would seem to be acceptable. When more than one transformation is doing a reasonable job of linearising the association, and they have similar value of r^2 then the transformation which is easier to interpret in terms of the variables is preferred. In this case *diameter*^{2} makes more sense in that it tells us that the *age* of the tree relates to the cross sectional area of the tree. The log transformation does not have an equivalent meaningful interpretation.

We can now fit a least squares line to model the association between *age* and *diameter*^{2}

The equation of this line is:

$$age = 5.098 + 0.091 \times diameter^2$$

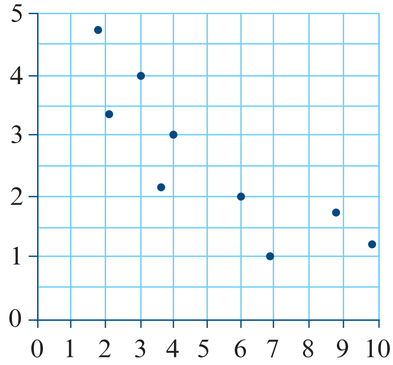
At this stage you might find it helpful to use the interactive 'Data transformation' (accessible through the Interactive Textbook) to see how these different transformations can be used to linearise scatterplots.

Exercise 4D

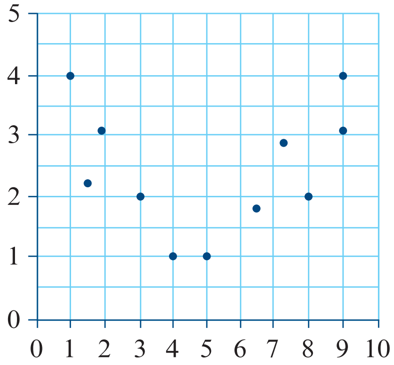
Example 7

1 The scatterplots below are non-linear. For each, identify the transformations x^2, log x, $1/x$, y^2, log y, $1/y$ or none that might be used to linearise the plot.

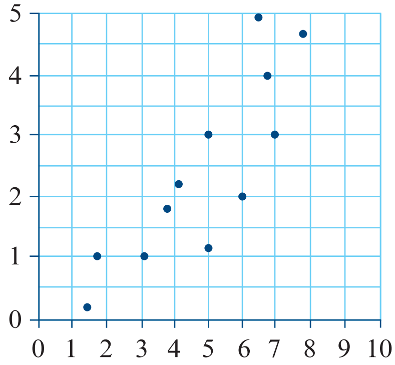
a



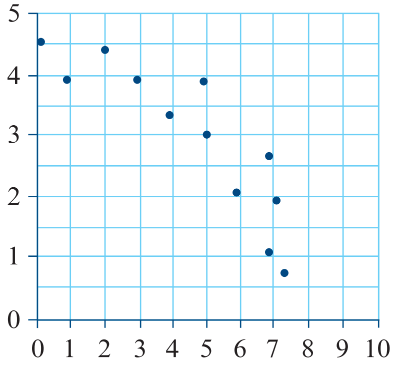
b



c



d



ISBN 978-1-009-11041-9

Cambridge University Press

2 The data below gives the yield in kilograms and length in metres of 12 commercial potato plots.

yield(kg)	346	1798	152	86	436	968
length(m)	12.1	27.4	8.3	5.5	15.7	21.5
yield(kg)	686	257	2435	287	1850	1320
length(m)	19.5	9.0	34.2	14.7	31.9	25.3

a Construct a scatterplot showing the association between *yield* in kilograms (the RV) and *length* of the plot in metres (the EV).

b Fit a least squares regression line to the data. Write down the equation in terms of the variables in the question, giving the values of the intercept and slope rounded to four significant figures.

c Construct a residual plot, and comment on whether the linearity assumption has been met.

d Use the circle of transformations to select which transformations could be considered in order to linearise the association.

e Using an appropriate transformation, recommend a regression model for the association between *yield* and *length* of the plot. Write down the equation in terms of the transformed variables, giving the values of the intercept and slope rounded to four significant figures.

f What is the value of r^2 for the recommended model? Give your answer as a percentage rounded to one decimal place.

3 In order to investigate the association between the average number of cigarettes per day per smoker (*smoking*) and the cost of cigarettes in \$ per cigarette (*cost*) for a group of countries the following data was collected.

cost (\$)	0.67	0.75	0.80	0.92	1.00	1.08	1.17	1.25	1.30	1.40
smoking	16.7	15.5	14.8	13.4	12.5	12.0	11.1	10.9	10.3	9.5

a Construct a scatterplot showing the association between *smoking* (in cigarettes/day) (the RV) and *cost* (\$/cigarette) (the EV).

b Fit a least squares regression line to the data. Write down the equation in terms of the variables in the question, giving the values of the intercept and slope rounded to four significant figures.

c Construct a residual plot, and comment on whether the linearity assumption has been met.

d Use the circle of transformations to select which transformations could be considered in order to linearise the association.

e Using an appropriate transformation, recommend a regression model for the association between *smoking* and *cost*. Write down the equation in terms of the

ISBN 978-1-009-11041-9

Cambridge University Press

transformed variables, giving the values of the intercept and slope rounded to four significant figures.

f What is the value of r^2 for the recommended model? Give your answer as a percentage rounded to one decimal place.

4 The following data shows the population density in people per hectare (*density*) and the distance from the centre of the city in km (*distance*) for a large city.

density	307.58	294.67	283.93	270.82	234.93	175.08	101.56	49.80
distance	0	2	4	6	8	10	12	14

a Construct a scatterplot showing the association between the population *density* in people per hectare (the RV) and *distance* from the centre of the city in km (the EV).

b Fit a least squares regression line to the data. Write down the equation in terms of the variables in the question, giving the values of the intercept and slope rounded to four significant figures.

c Construct a residual plot, and comment on the whether linearity assumption has been met.

d Use the circle of transformations to select which transformations could be considered in order to linearise the association.

e Using an appropriate transformation, recommend a regression model for the association between *density* and *distance*. Write down the equation in terms of the transformed variables, giving the values of the intercept and slope rounded to four significant figures.

f What is the value of r^2 for the recommended model? Give your answer as a percentage rounded to one decimal place.

ISBN 978-1-009-11041-9

Cambridge University Press

Key ideas and chapter summary

Assignment

Data transformation In regression analysis, **data transformation** involves changing the scale on either the *x*- or *y* axis in order to linearise an association prior to fitting a least squares line.

Squared transformation The **squared transformation** *stretches out* the upper end of the scale on an axis.

Logarithmic transformation The **logarithmic transformation** *compresses* the upper end of the scale on an axis.

Reciprocal transformation The **reciprocal transformation** *compresses* the upper end of the scale on an axis but to a greater extent than the log transformation.

Residual plots Residual plots are used to assess the effectiveness of a data transformation.

Coefficient of determination The transformation that results in a linear association (as assessed by the residual plot) and which has the highest value of the coefficient of determination is generally the preferred transformation.

The circle of transformations The **circle of transformations** provides guidance in choosing the transformations that can be used to linearise various types of scatterplots.

Skills checklist

Checklist

Download this checklist from the Interactive Textbook, then print it and fill it out to check your skills. ☑

4A **1** I can apply the x^2 transformation. □

See Example 1, and Exercise 4A Question 2

4A **2** I can apply the y^2 transformation. □

See Example 2, and Exercise 4A Question 5

4B **3** I can apply the log *x* transformation. □

See Example 3, and Exercise 4B Question 2

4B **4** I can apply the log *y* transformation. □

See Example 4, and Exercise 4B Question 5

4C **5** I can apply the reciprocal ($1/x$) transformation. □

See Example 5, and Exercise 4C Question 2

4C **6** I can apply the reciprocal ($1/y$) transformation. □

See Example 6, and Exercise 4C Question 4

4D **7** I can choose the best transformation to apply. □

See Example 7, and Exercise 4D Question 1

Multiple-choice questions

1 Select the statement that correctly completes the sentence:
'The effect of a squared transformation is to . . . '

A stretch the high values in the data **B** maintain the distance between values
C stretch the low values in the data **D** compress the high values in the data
E reverse the order of the data values

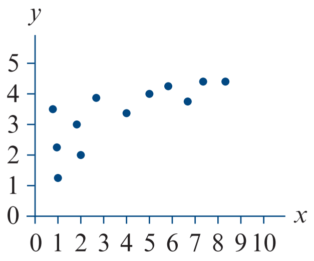
2 Select the statement that correctly completes the sentence:
'The effect of a log transformation is to . . .'

A stretch the high values in the data **B** maintain the distance between values
C stretch the low values in the data **D** compress the high values in the data
E maintain the order of the values in the data

3 Select the statement that correctly completes the sentence:
'The effect of a reciprocal transformation is to . . .'

A stretch the high values in the data **B** maintain the distance between values
C stretch the low values in the data **D** compress the high values in the data
E reverse the order of the values in the data

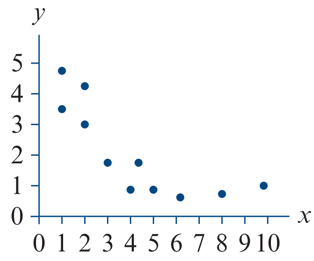
4 The association between two variables y and x, as shown in the scatterplot, is non-linear. In an attempt to transform the association to linearity, a student would be advised to:



A leave out the first four points
B use a y^2 transformation
C use a log y transformation
D use a $1/y$ transformation
E use a least squares regression line

ISBN 978-1-009-11041-9 Cambridge University Press

5 The association between two variables y and x, as shown in the scatterplot, is non-linear. Which of the following sets of transformations could possibly linearise this association?

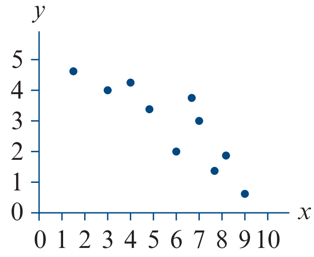


A $\log y$, $1/y$, $\log x$, $1/x$ **B** y^2, x^2

C y^2, $\log x$, $1/x$ **D** $\log y$, $1/y$, x^2

E $ax + b$

6 The association between two variables y and x, as shown in the scatterplot, is non-linear. Which of the following transformations is most likely to linearise the association?



A a $1/x$ transformation **B** a y^2 transformation

C a $\log y$ transformation **D** a $1/y$ transformation

E a $\log x$ transformation

7 The following data were collected for two related variables x and y.

x	1	2	3	4	5	6	7	8	9	10	11
y	7	8.6	8.9	8.8	9.9	9.7	10.4	10.5	10.7	11.2	11.1

A scatterplot indicates a non-linear association. The data is linearised using a log x transformation and a least squares line is then fitted. The equation of this line is closest to:

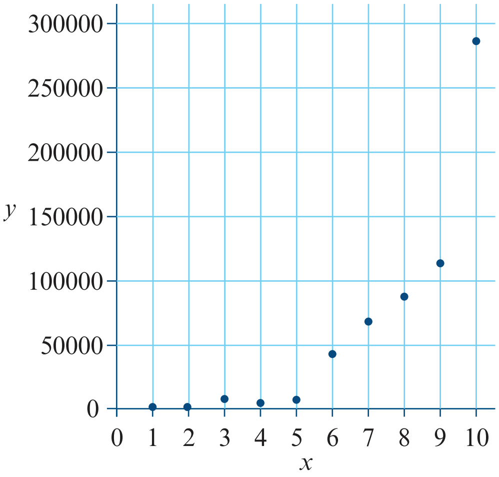
A $y = 7.52 + 0.37 \log x$ **B** $y = 0.37 + 7.52 \log x$

C $y = -1.71 + 0.25 \log x$ **D** $y = 3.86 + 7.04 \log x$

E $y = 7.04 + 3.86 \log x$

8 A student uses the data in the table below to construct the scatterplot shown

x	y
1	2030
2	1265
3	8265
4	5654
5	6893
6	43265
7	67890
8	87803
9	113062
10	286370



ISBN 978-1-009-11041-9 Cambridge University Press

A log transformation is applied to y to linearise the association. A least squares line is fitted to the transformed data, with $\log y$ as the response variable.

The equation of this least squares line is closest to

A $y = 2.88 + 0.256 \log x$

B $\log y = -10.1 + 3.64 \log x$

C $\log y = -69800 + 24000x$

D $\log y = 2.88 + 0.256x$

E $\log y = -4.84 + 4.38x$

9 The association between the total *weight* of produce picked from a vegetable garden and its *width* is non-linear. An x^2 transformation is used to linearise the data.

When a least squares line is fitted to the data, its y-intercept is 10 and its slope is 5.

Assuming that *weight* is the response variable, the equation of this line is:

A $(weight)^2 = 10 + 5 \times\ width$ **B** $width = 10 + 5 \times\ (weight)^2$

C $width = 5 + 10 \times\ (weight)^2$ **D** $weight = 10 + 5 \times\ (width)^2$

E $(weight)^2 = 5 + 10 \times\ weight$

10 A model that describes the association between the hours spent studying for an exam and the mark achieved is:

$mark = 20 + 40 \times \log(hours)$

From this model, we would predict that a student who studies for 20 hours would score a mark (to the nearest whole number) of:

A 80 **B** 78 **C** 180 **D** 72 **E** 140

11 A $1/y$ transformation is used to linearise a scatterplot.

The equation of a least squares line fitted to this data is:

$1/y = 0.14 + 0.045x$

This regression line predicts that, when $x = 6$, y is closest to:

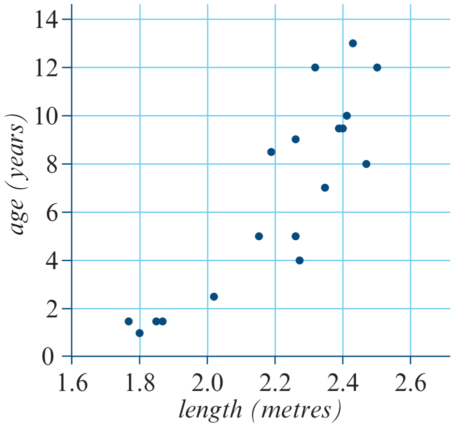
A 0.17 **B** 0.27 **C** 0.41 **D** 2.4 **E** 3.7

ISBN 978-1-009-11041-9 © Peter Jones et al 2023 Cambridge University Press

Written response questions

1 The table below shows the age in years (*age*) and the length in metres (*length*), for a group of 18 dugongs. A scatterplot of the data is also shown.

age	*length*	*age*	*length*
1.0	1.80	8.0	2.47
1.5	1.85	8.5	2.19
1.5	1.87	9.0	2.26
1.5	1.77	9.5	2.40
2.5	2.02	9.5	2.39
4.0	2.27	10.0	2.41
5.0	2.15	12.0	2.50
5.0	2.26	12.0	2.32
7.0	2.35	13.0	2.43



The scatterplot shows that the association is clearly non-linear. A reciprocal $\left(\frac{1}{y}\right)$ transformation can be applied to the variable *age* to linearise the association.

a Apply the reciprocal $\left(\frac{1}{y}\right)$ transformation to the data and use the transformed data to determine the equation of a least squares line that enables $\frac{1}{age}$ to be predicted from *length*. Write the values of the intercept and slope in the the appropriate boxes provided. Round to four significant figures.

$$\frac{1}{age} = \boxed{} - \boxed{} \times length$$

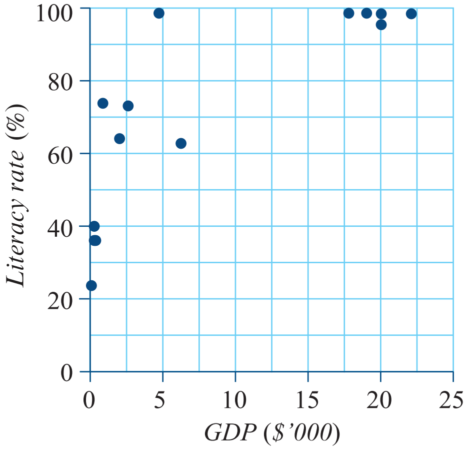
b The association can also be linearised by applying a log transformation to the variable *age*. When this is done, and a least squares line fitted to the transformed data, the resulting equation is:

$$\log(age) = -2.443 + 1.429 \times length$$

Use this equation to predict the age of a dugong with a length of 2.00 metres. Round the answer to one decimal place.

2 The table below shows the percentage of people who can read (*literacy rate*) and the gross domestic product (*GDP*), in dollars/person, for a selection of 14 countries. A scatterplot of the data is also shown.
The scatterplot can be linearised by using a log *x* transformation.

GDP	*literacy rate*
2677	72
260	35
19 904	97
122	24
18 944	99
4 500	99
17 539	99
1 030	73
19 860	99
409	40
406	35
6651	62
22 384	99
2 436	64



a Apply the log transformation to the variable *GDP*, and fit a least squares line to the transformed data. Write down its equation terms of the variables *literacy rate* and log (*GDP*). Give the slope and intercept rounded to three significant figures.

b Verify that the log transformation has linearised the association by constructing a residual plot.

c Use the regression equation to predict the literacy rate of a country with a GDP of $10 000 to the nearest percent.

d Find the value of the residual when the regression equation is used to predict the literacy rate when the GDP is equal to $19 860. Give your answer rounded to two significant figures.

3 Measurements of the *distance* travelled (metres) and *time* taken (seconds) were made on a falling body. The data are given in the table below.

time	0	1	2	3	4	5	6
distance	0	5.2	18.0	42.0	79.0	128.0	168.0

a Construct a scatterplot of the data and comment on its form.

b Determine the values of *time*^{2} and complete the table.

c Construct a scatterplot of *distance* against *time*^{2}.

d Fit a least squares line to the transformed data, with *distance* as the RV.

e Use the regression equation to predict the distance travelled in 7 seconds.

f Obtain a residual plot and comment on whether the assumption of linearity is reasonable.

4 Is the infant mortality rate in a country associated with the number of doctors in that country? The data below gives infant mortality rate in deaths per 1000 births (*mortality*) and the number of doctors per 100 000 of population (*doctors*) for 14 countries.

mortality	12	13	12	10	10	7	111
doctors	192	222	154	182	179	204	61
mortality	15	10	20	54	75	121	71
doctors	270	271	357	79	59	27	52

a Construct a scatterplot of *mortality* against *doctors* and use it to comment on the association between infant mortality rate and doctor numbers.

b Construct a scatterplot of *mortality* against $\dfrac{1}{doctors}$.

c Determine the equation of the least squares regression line which would enable *mortality* to be predicted from $\dfrac{1}{doctors}$. Give the values of the intercept and slope rounded to four significant figures.

d Obtain a residual plot for the model fitted in part c and comment on the linearity.

e Determine the value of coefficient of determination for the model fitted in part c. Give your answer as a percentage rounded to one decimal place.

f Use the regression equation to predict the infant mortality rate in a country where there are 100 doctors per 100,000 people. Give your answer rounded to the nearest whole number.

g Comment on the reliability of the prediction made in part f.