General Maths Unit 3

Core Topic: Statistics
Chapter 4
Data Transformation

Area of Study 1 – Data Analysis

- 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?

The Squared Transformation – Ex 4a

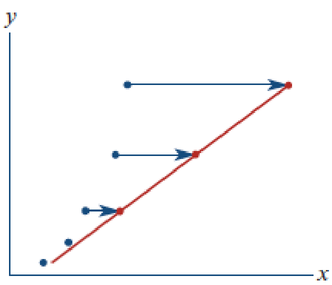
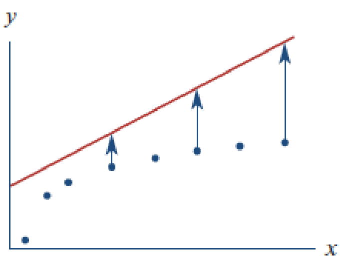
Scatterplots are drawn to check whether 2 variables are related or not and if the relationship is linear or not.

If the **residual plot shows a pattern (i.e. the relationship is non linear),** we can transform the data to find a more accurate equation. Fore example, if the residual plot shows a parabola pattern, it indicates that the original data may have had a quadratic relationship rather than a linear relationship. Therefore, it is more appropriate to fit a quadratic equation.

The squared Transformation

- A stretching transformation
- Stretches out the upper end of the scale of either the x or y axis
- Can apply x^2 or y^2

Graphically…

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.	

CAS Skills

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

a $y = 7 + 8x^2$ when $x = 1.25$

b $y = 7 + 3x^2$ when $x = 1.25$

c $y = 24.56 - 0.47x^2$ when $x = 1.23$

d $y = -4.75 + 5.95x^2$ when $x = 4.7$

Example 1 – the squared transformation

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 Linearise the scatterplot 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.

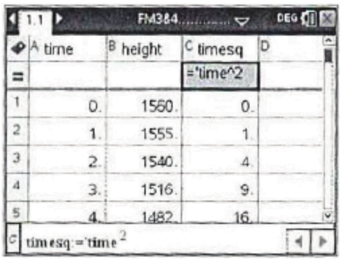
Steps

1 Start a new document by pressing ctrl + N .

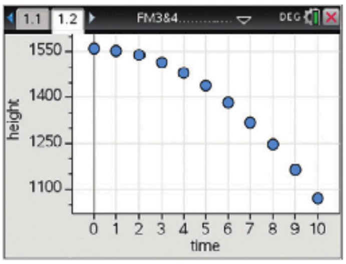
2 Select **Add Lists & Spreadsheet**.
Enter the data into lists named **time** and **height**, as shown.

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

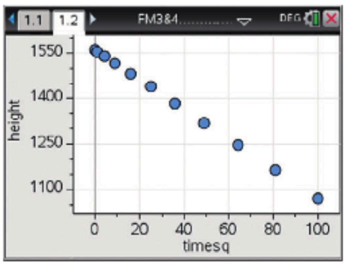
4 Move the cursor to the grey cell below **timesq**. Enter the expression = **time^2** by pressing = , then typing **time^2**. Pressing enter calculates and displays the values of **timesq**.



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.

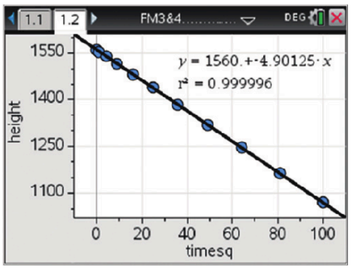


6 Press ctrl + I and select **Add Data & Statistics**.
Construct a scatterplot of *height* against *time*^{2}.
The plot is now linear.



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 $time^2$.



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

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

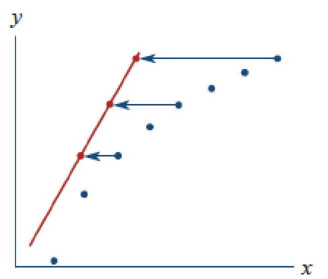
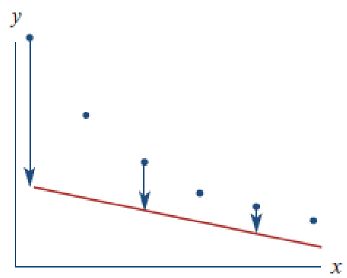
9 Substitute 3.4 for *time* in the equation to find the height after 3.4 seconds.

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

Ex 4a

The log Transformation – Ex 4b

- Log transformation to x compresses the higher x-values thereby straightening out the line
- Log to y compresses larger y values

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.	

CAS 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 $y = 196.1 - 23.2 \log 303$

$\text{Log}_{10}(x)$ Transformation

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:

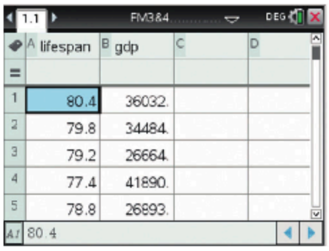
- linearise the data, and fit a regression line to the transformed data (*GDP* is the EV)
- write its equation in terms of the variables *lifespan* and *GDP* correct to three significant figures.
- use the equation of the regression line to predict the lifespan in a country with a GDP of \$20 000, correct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

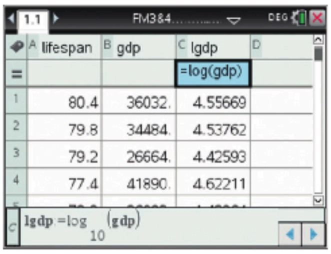
1 Start a new document by pressing ctrl + N .

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

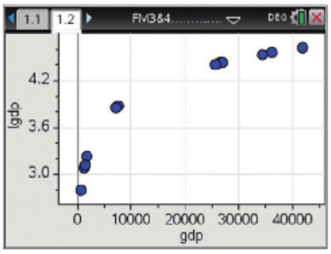


3 Name column C as **lgdp** (short for log (*GDP*)).
Now calculate the values of log (*GDP*) and store them in the list named **lgdp**.

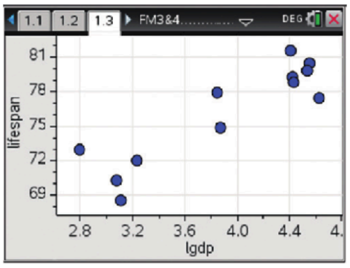
4 Move the cursor to the grey cell below the **lgdp** heading.
We need to enter the expression = **log(gdp)**.
To do this, press = then type in **log(gdp)**. Pressing enter calculates and displays the values of **lgdp**.



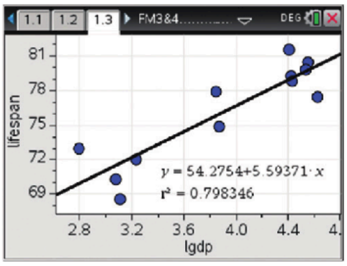
5 Press ctrl + I and select **Add Data & Statistics**.
Construct a scatterplot of *lifespan* against *GDP*. Let *GDP* be the explanatory variable and *lifespan* the response variable. The plot is clearly non-linear.



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).



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

lifespan = 54.3 + 5.59 × log (GDP)

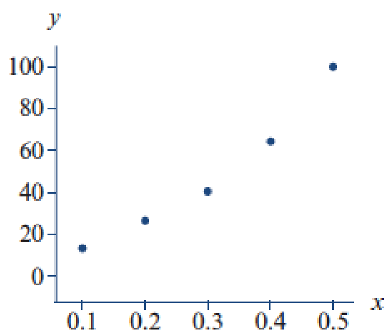
9 Substitute 20 000 for *GDP* in the equation to find the lifespan of people in a country with GDP of $20 000.

lifespan = 54.3 + 5.59 × log 20 000
= 78.3 years

Log y transformation

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.

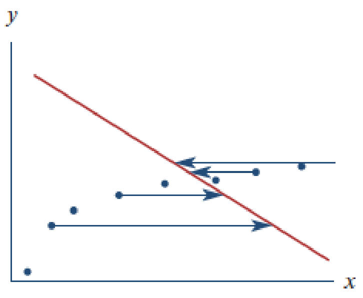
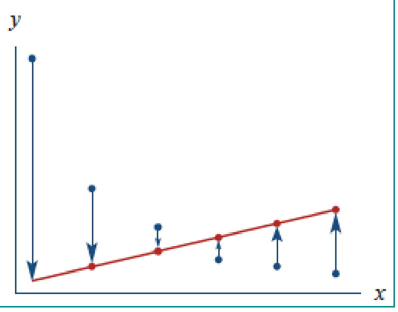
a Linearise the scatterplot by applying a log y transformation and fit a least squares line to the transformed data.

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.

Ex 4b

The Reciprocal Transformation – Ex 4c

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.	

Example Using CAS

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:

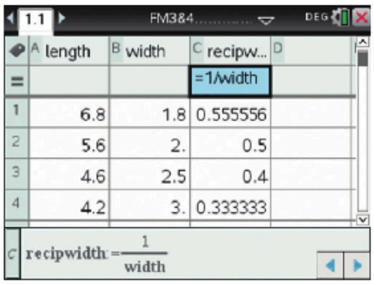
- linearise the data, and fit a regression line to the transformed data (*length* is the EV)
- write its equation in terms of the variables *length* and *width*
- use the equation to predict the width of a sticky label with a length of 5 cm.

Steps

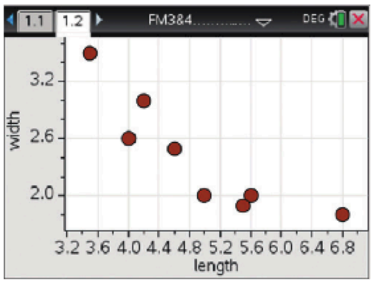
1 Start a new document by pressing [ctrl] + [N].

2 Select **Add Lists & Spreadsheet**.
Enter the data into lists named **length** and **width**.

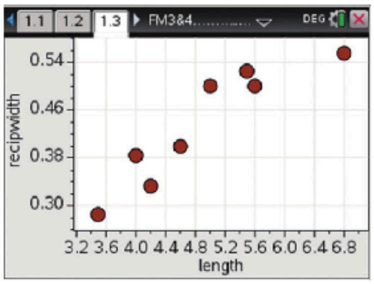
3 Name column C as **recipwidth** (short for 1/width).
Calculate the values of **recipwidth**.
Move the cursor to the grey cell below the **recipwidth** heading. Type in **=1/width**. Press [enter] to calculate the values of **recipwidth**.



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.

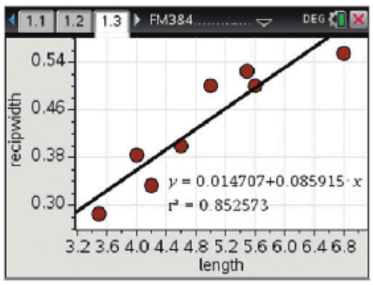


5 Press [ctrl] + [I] and select **Add Data & Statistics**.
Construct a scatterplot of **recipwidth** ($1/width$) against **length**. The plot is now clearly linear.



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

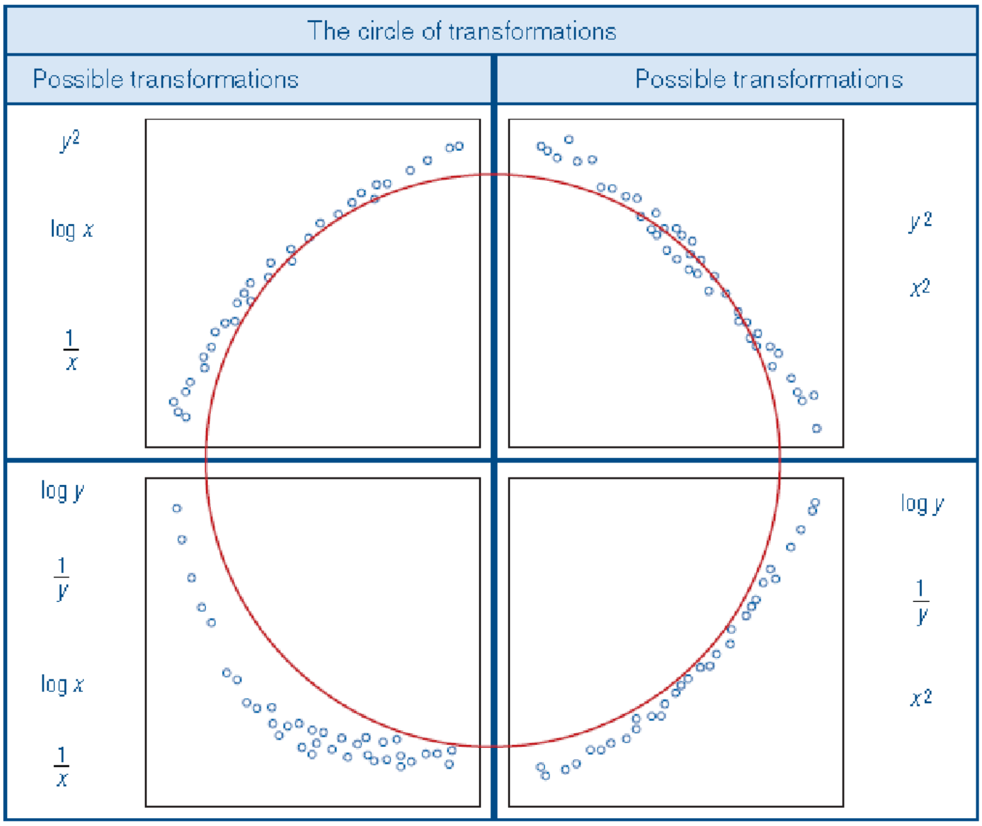
Note: The y in the equation on the screen corresponds to the transformed variable $1/width$.



7	Write down the regression equation in terms of the variables *width* and *length*.	1/width = 0.015 + 0.086 × length
8	Substitute 5 cm for *length* in the equation.	1/width = 0.015 + 0.086 × 5 = 0.445 or width = 1/0.445 = 2.25 cm (to 2 d.p.)

Ex 4c

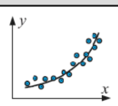
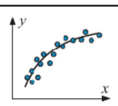
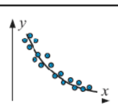
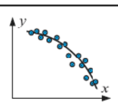
Choosing the Appropriate Transformation – Ex 4d

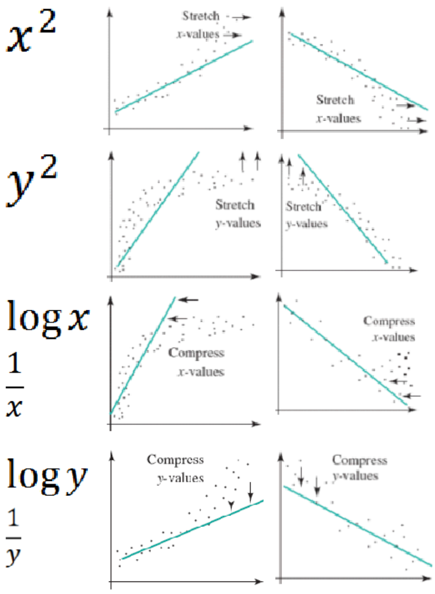


Another helpful tool….

Transformations

The following tables shows which transformations can be used on certain scatterplot shapes, and the effect of certain transformations.

Curve	*Transformations to test*
	x^2, $\log y$, or $\frac{1}{y}$
	y^2, $\log x$, or $\frac{1}{x}$
	$\log x$, $\log y$, $\frac{1}{x}$, or $\frac{1}{y}$
	x^2 or y^2



When a transformation is needed, use the values of r and r^2 to help determine which is best.

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 most appropriate**. However, there may be very little difference so common sense needs to be applied!

Once the data has been transformed, the form of the equation changes slightly.

- The least squares is in the form $y = a + bx$
- If an x squared transformation occurs the regression is in the form $y = a + bx^2$
- If a logarithmic transformation occurs the regression is in the form $y = a + b \times \log(x)$

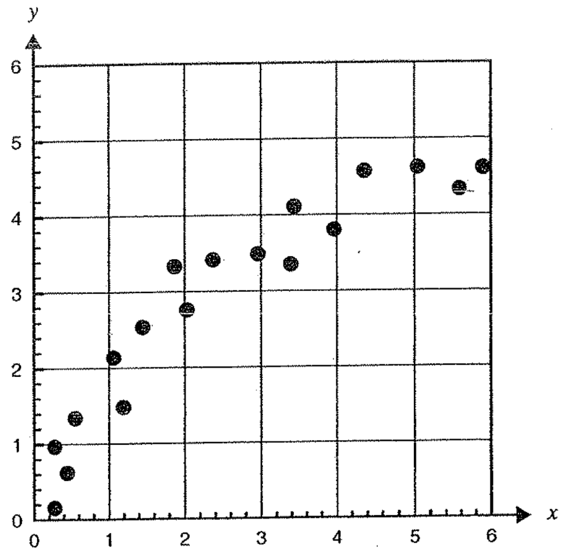
- If a reciprocal transformation occurs the regression is in the form $y = a + \frac{b}{x}$

Ex 4d
Exam Questions

Exam Questions:

1. 2000

The relationship between the two variables y and x as shown in the scatterplot below is clearly non-linear.



In an attempt to transform the data to linearity, a student would be best advised to

A. leave out the first six points.

B. use a y^2 transformation.

C. use a *log y* transformation.

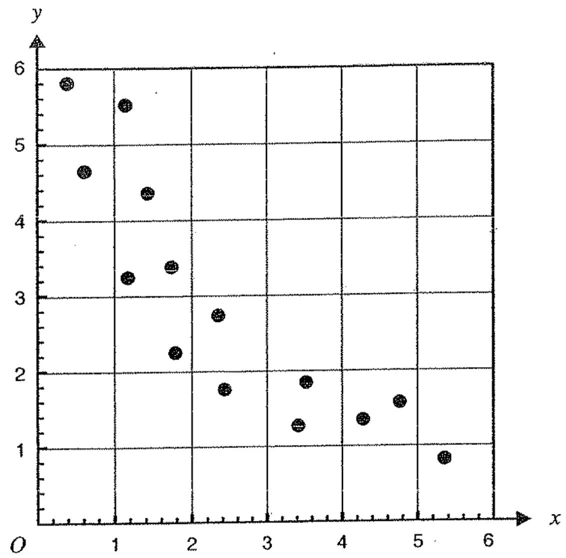
D. use a $\frac{1}{y}$ transformation.

E. use a three median line.

2.

Question 11

The relationship between the two variables y and x, as shown in the scatterplot below, is clearly non-linear.



A student wants to linearise the scatterplot above by transforming the scale on one of the axes only. Which of the following approaches would be the best strategy?

A. Use either a $\frac{1}{x}$ or a $\frac{1}{y}$ transformation.

B. Use either a $\frac{1}{x}$ or a y^2 transformation.

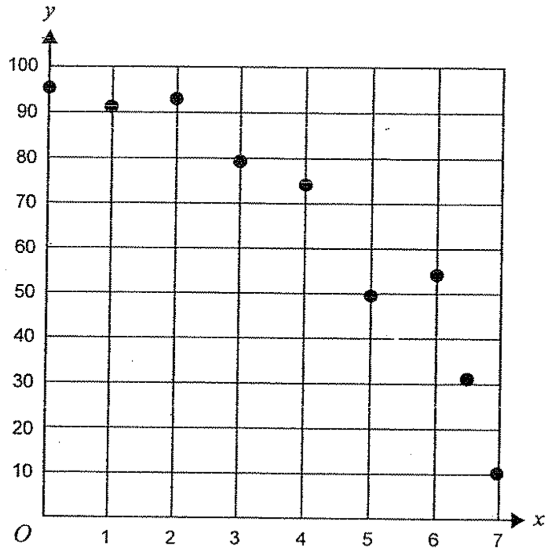
C. Use either a $\log x$ or a y^2 transformation.

D. Use either a x^2 or a $\log y$ transformation.

E. Use either an x^2 or a y^2 transformation.

3.

The relationship between the two variables y and x, as shown in the scatterplot below, is nonlinear.



Which one of the following transformations, by itself, is most likely to linearise this data?

A. a $\frac{1}{x}$ transformation

B. a $\frac{1}{y}$ transformation

C. an x^2 transformation

D. a $\log x$ transformation

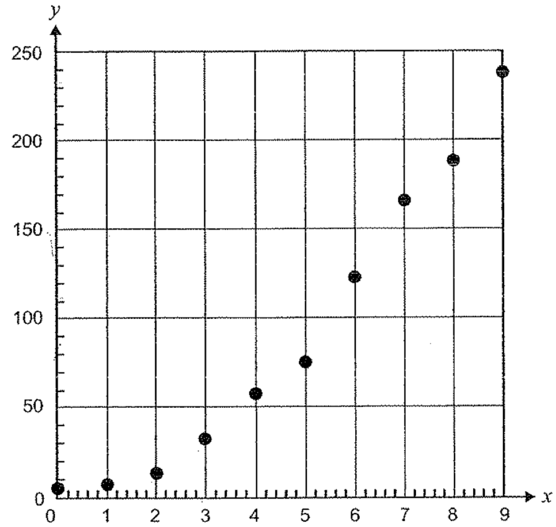
E. a $\log y$ transformation

4. 2009

Question 9

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

x	0	1	2	3	4	5	6	7	8	9
y	5	7	14	33	58	76	124	166	188	238



To linearise the scatterplot, she applies an x-squared transformation.

She then fits a least squares regression line to the **transformed data** with y as the dependent variable.

The equation of this least squares regression line is closest to

A. $y = 7.1 + 2.9x^2$

B. $y = -29.5 + 26.8x^2$

C. $y = 26.8 - 29.5x^2$

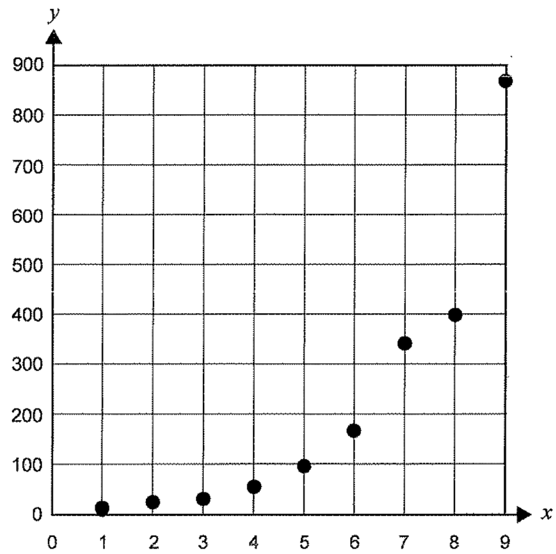
D. $y = 1.3 + 0.04x^2$

E. $y = -2.2 + 0.3x^2$

5. 2007

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

x	1	2	3	4	5	6	7	8	9
y	12	25	33	58	98	168	345	397	869



To linearise the scatterplot, she applies a **log** y transformation; that is, a log transformation is applied to the y-axis scale.

She then fits a least squares regression line to the transformed data.

With x as the independent variable, the equation of this least squares regression line is closest to

A. $\log y = -217 + 88.0\,x$

B. $\log y = -3.8 + 4.4\,x$

C. $\log y = 3.1 + 0.008\,x$

D. $\log y = 0.88 + 0.23\,x$

E. $\log y = 1.58 + 0.002\,x$