Section 3.1.16 Transformations to linearity

VCAA "Dot Points"

Investigating and modelling linear associations, including:

- *interpretation of the slope and intercepts of the least squares line in the context of the situation being modelled, including:*
  - *— use of the rule of the fitted line to make predictions being aware of the limitations of extrapolation*
  - *— use of the coefficient of determination, r2, to assess the strength of the association in terms of explained variation*
  - *— use of residual analysis to check quality of fit*

- *data transformation and its use in transforming some forms of non-linear data to linearity using a square, log or reciprocal transformation (on one axis only)*

- *interpretation and use of the equation of the least squares line fitted to the transformed data to make predictions.*

Checking for Linearity

To determine whether a **linear relationship** exists between two variables, **three "checks"** can be applied.

Check 1 – The correlation coefficient (r)
A high correlation coefficient indicates a strong linear association/relationship between the two variables.

Check 2 – The coefficient of determination (r^2)
A high correlation coefficient indicates that the response variable can be predicted, to a high degree, by the explanatory variable.

NB: Several data sets may appear to be linear based upon a high Pearson product-moment correlation coefficient (r). However, when examined more closely, the relationship may actually be better explained by a non-linear model such as a reciprocal, logarithmic or squared relationship.

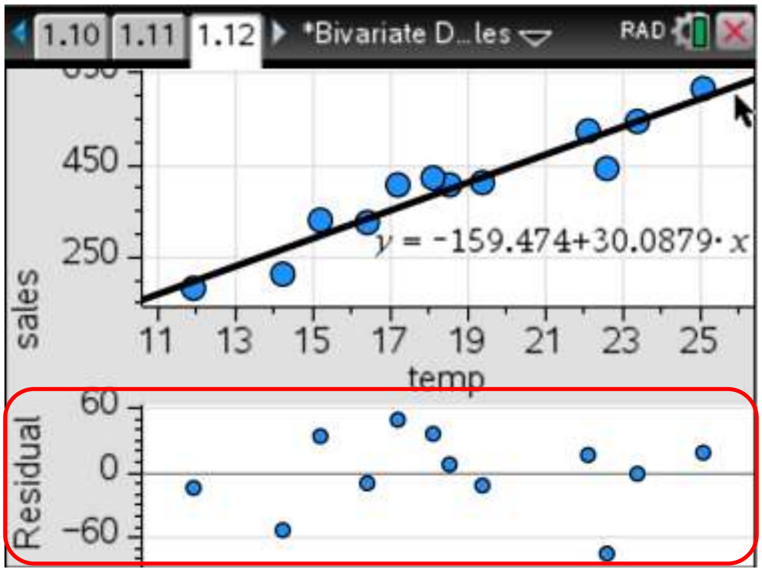
Check 3 – Residual plot
The third and final "check" to determine whether a linear relationship exists between two variables is to analyse the residual plot created by the data.

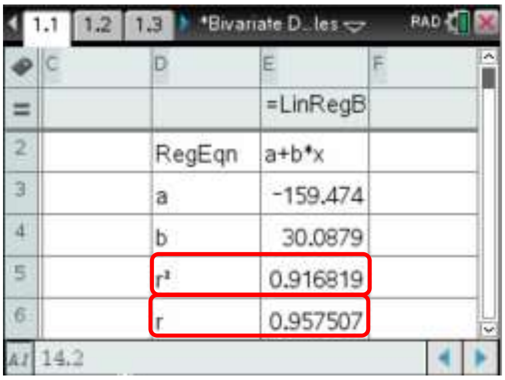
If a linear relationship exists, then the residual plot will display:

- An appropriately **equal number of points** above and below the axis
- A **random scattering** of points above and below the x-axis
- **No clear pattern**

Example 1

Consider the ice cream sales v daily temperature example from Notes 3.1.12. The linear regression and residual plot for this data is as follows:





Linearity Check List

☑ High r (0.958)
☑ High r^2 (0.917)
☑ Evenly scattered residual plot

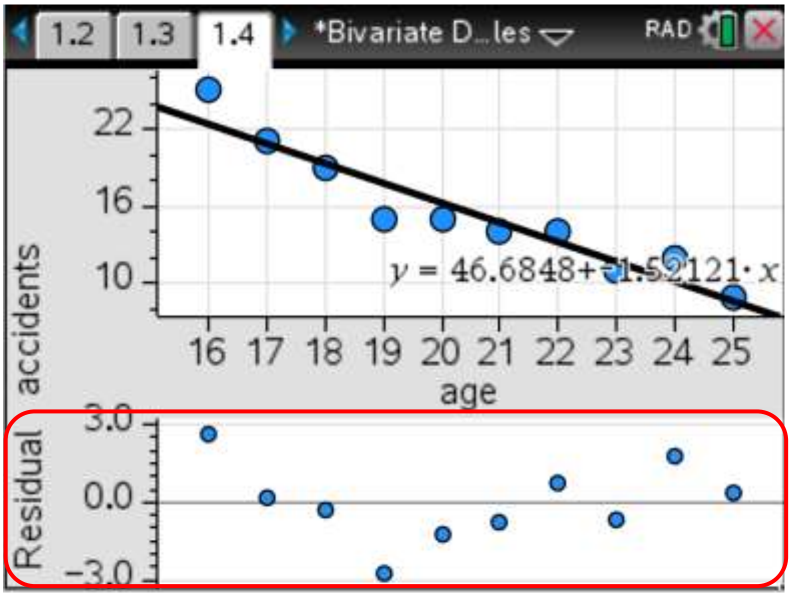
Outcome

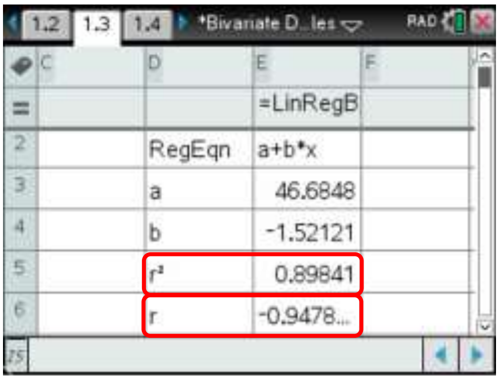
The original data probably have a linear relationship.

$Ice\ cream\ sales\ (\$) = -159.47 + 30.09 \times Temperature\ (°C)$

Example 2

Consider the ice driver age v no. of accidents example from Notes 3.1.12. The linear regression and residual plot for this data is as follows:





Linearity Check List

☑ High r (-0.948)
☑ High r^2 (0.898)
☑ Evenly scattered residual plot

Outcome

The original data probably have a linear relationship.

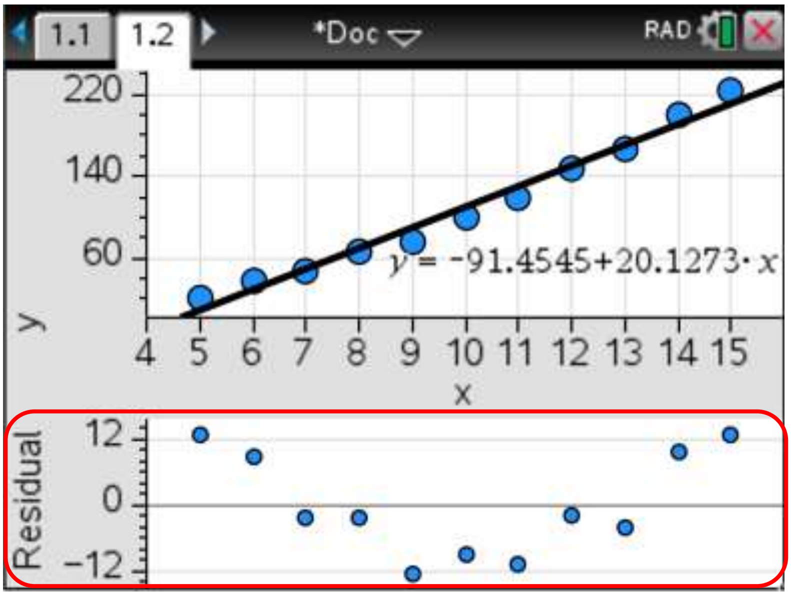
$No.of\ accidents(per\ 100) = 46.68 - 1.52 \times Age(years)$

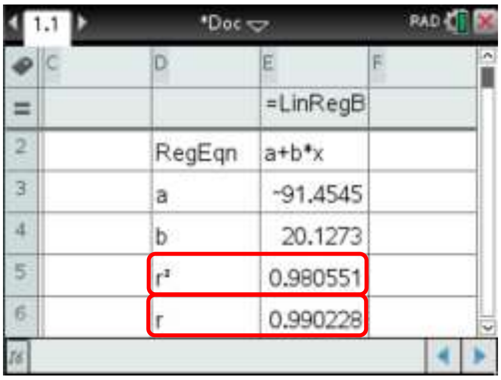
Example 3

Consider the following data for example:

x	5	6	7	8	9	10	11	12	13	14	15
y	22	38	47	67	77	101	119	148	166	200	223

The linear regression and residual plot for this data is as follows:





Linearity Check List
☑ High r (0.990)
☑ High r^2 (0.981)
☒ Unevenly scattered residual plot
There appears to be a curved pattern

Outcome
The original data probably have a non-linear relationship.
Transformation of the data may be required

NB: a very strong correlation coefficient **does not** guarantee a linear relationship.

Transformation Options

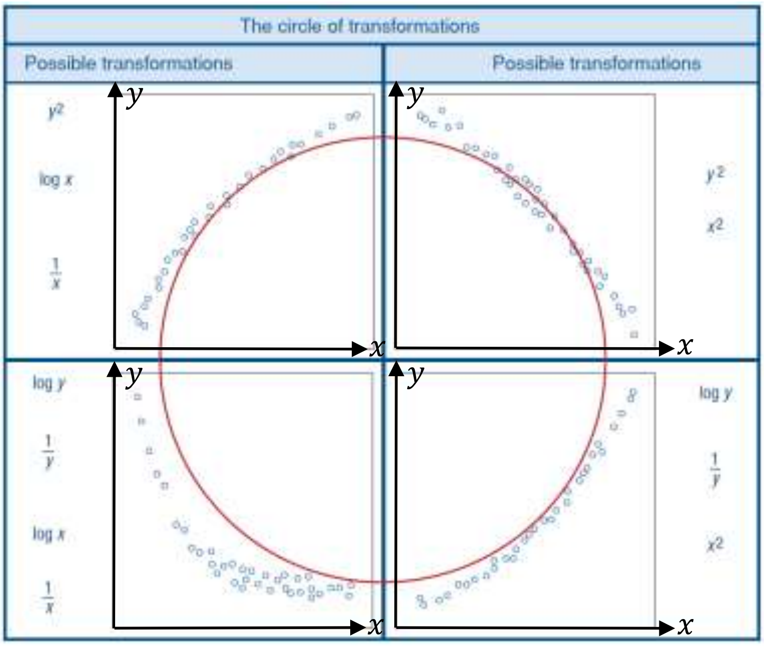
There are six different transformations available.

x transformations	Reciprocal of x: $\frac{1}{x}$	Logarithm of x: $log_{10}(x)$	x squared: x^2
y transformations	Reciprocal of y: $\frac{1}{y}$	Logarithm of y: $log_{10}(y)$	y squared: y^2

The questions is which transformation should we use?

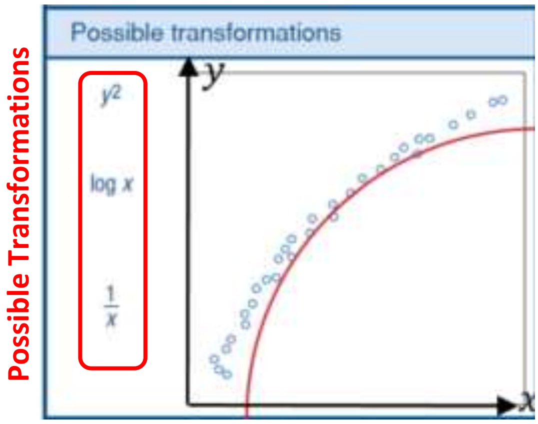
We can examine the shape of the original data and see which quadrant it fits in upon **the circle of transformation** diagram.

The Circle of Transformation

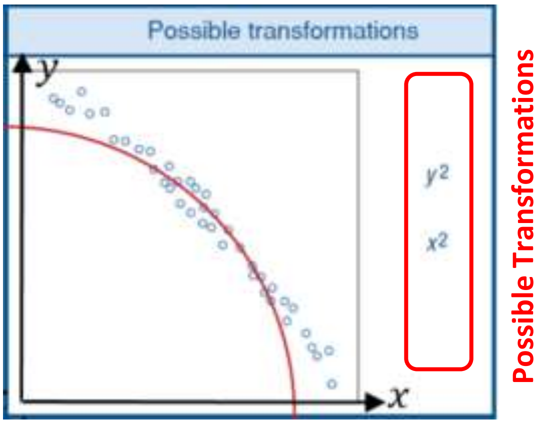


The **circle of transformation** is a visual tool used to select the most **appropriate transformation** for a given set of non-linear data.

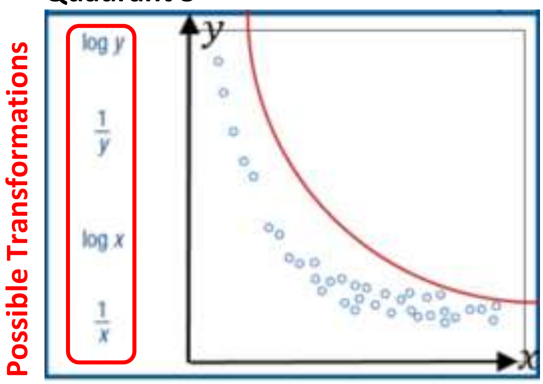
Quadrant 1



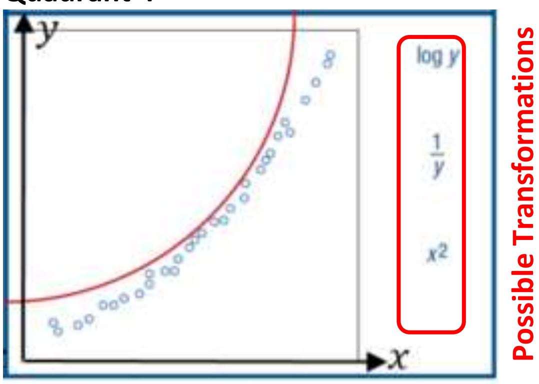
Quadrant 2



Quadrant 3



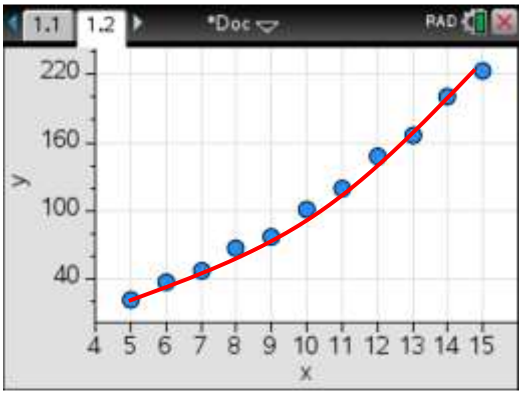
Quadrant 4



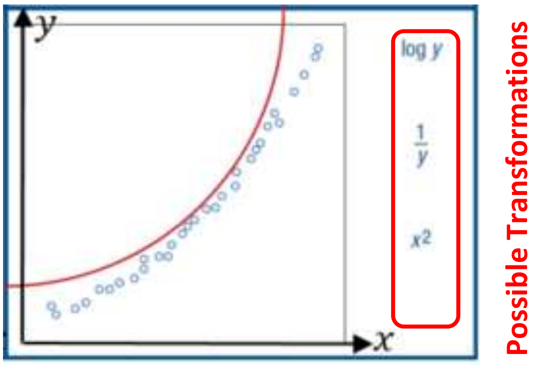
Example 3 (cont)

The original data <u>probably</u> have a non-linear relationship. **Transformation** of the data may be required.

Qn: Which transformation to test?



Quadrant 4

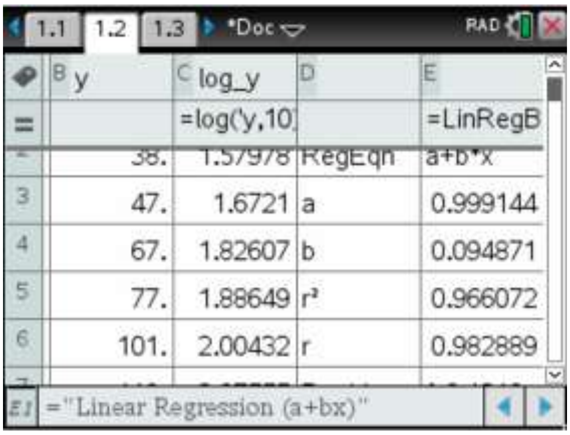


The shape of the original data best matches that of Quadrant 4.
The following transformations should be investigated:

- Logarithm of y: $log_{10}(y)$
- Reciprocal of y: $\dfrac{1}{y}$
- x squared: x^2

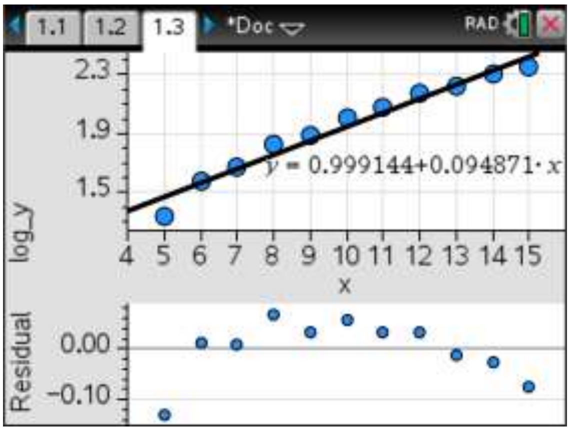
Let's now exam each of the three recommended transformations to determine what relationship actually exists between the two variables.

Transformation 1: Logarithm of y: $log_{10}(y)$



	B y	C log_y	D	E
=		=log('y,10)		=LinRegB
2	38.	1.57978	RegEqn	a+b*x
3	47.	1.6721	a	0.999144
4	67.	1.82607	b	0.094871
5	77.	1.88649	r²	0.966072
6	101.	2.00432	r	0.982889

E1 ="Linear Regression (a+bx)"

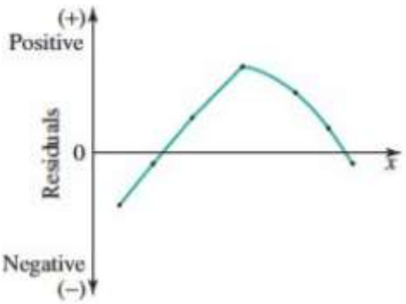


NB: the y-axis now represents $log_{10}(y)$

Summary:
Least squares regression line: $log_{10}(y) = 0.9991 + 0.0949x$
r = 0.983
r^2 = 0.966

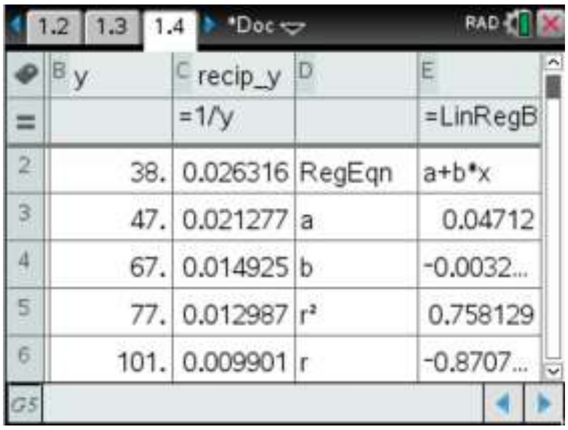
Residual plot:
The points of the residuals show a curved pattern (∩) with a series of negative, then positive and back to negative residuals along the x-axis. The original data probably have a non-linear relationship. An alternative transformation of the data may be required.

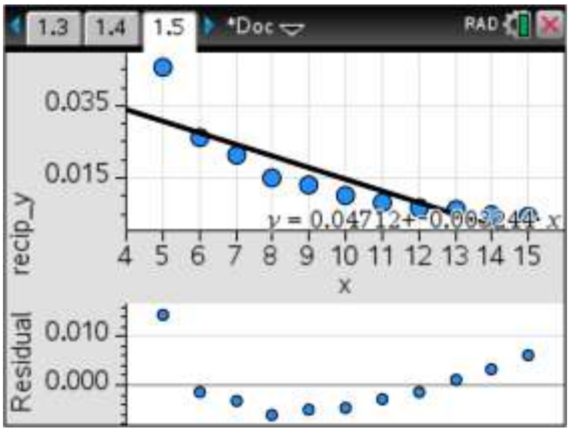
Outcome:
The $log_{10}(y)$ transformation has not improved the original r or r^2 values, nor does the residual plot indicate any improved linearity. Proceed to the next recommended transformation.



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

Transformation 2: Reciprocal of y: $\frac{1}{y}$





NB: the y-axis now represents $\frac{1}{y}$

Summary:

Least squares regression line: $\frac{1}{y} = 0.0471 + 0.0032x$

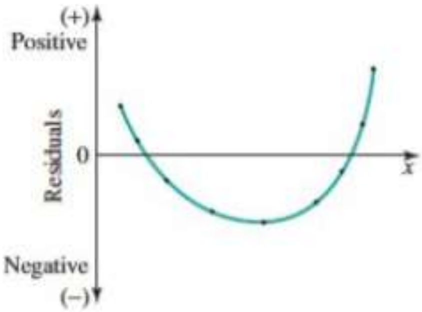
r = -0.8707
r^2 = 0.7581

Residual plot:

The points of the residuals show a curved pattern (◡) with a series of positive, then negative and back to positive residuals along the x-axis. The original data probably have a non-linear relationship. An <u>alternative</u> transformation of the data may be required.

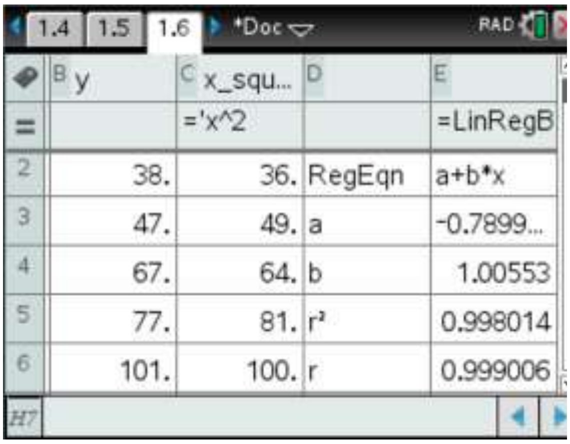
Outcome:

The $\frac{1}{y}$ transformation has not improved the original r or r^2 values, nor does the residual plot indicate any improved linearity. Proceed to the next recommended transformation.

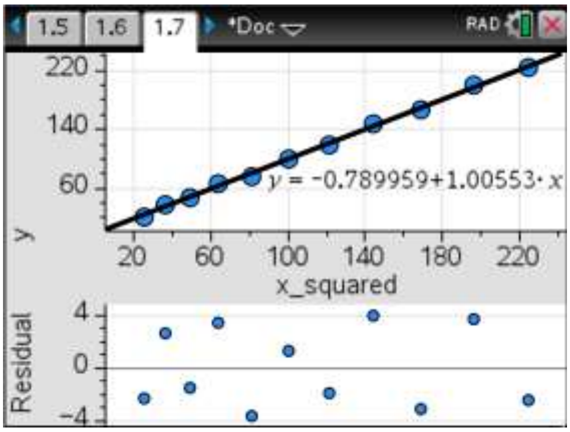


The points of the residuals show a curved pattern (∪), with a series of positive, then negative and back to positive residuals along the *x*-axis. The original data probably have a *non-linear* relationship. Transformation of the data may be required.

Transformation 3: x squared: x^2



	B y	C x_squ...	D	E
=		='x^2		=LinRegB
2	38.	36.	RegEqn	a+b*x
3	47.	49.	a	-0.7899...
4	67.	64.	b	1.00553
5	77.	81.	r²	0.998014
6	101.	100.	r	0.999006

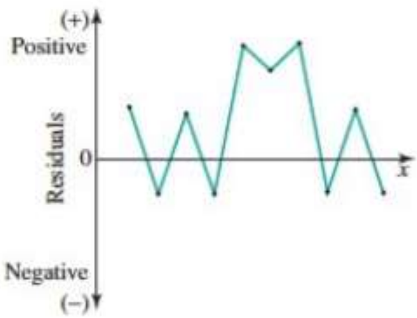


NB: the x-axis now represents x^2

Summary:
Least squares regression line: $y = -0.789959 + 1.00533x^2$
r = 0.999
r^2 = 0.998

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

Outcome:
The x^2 transformation has improved the original r and r^2 values and the residual plot also indicate an improved linearity. <u>This is the correct transformation</u> for this particular set of data. Therefore $y = -0.789959 + 1.00533x^2$ is the best relationship available to these two variables.



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

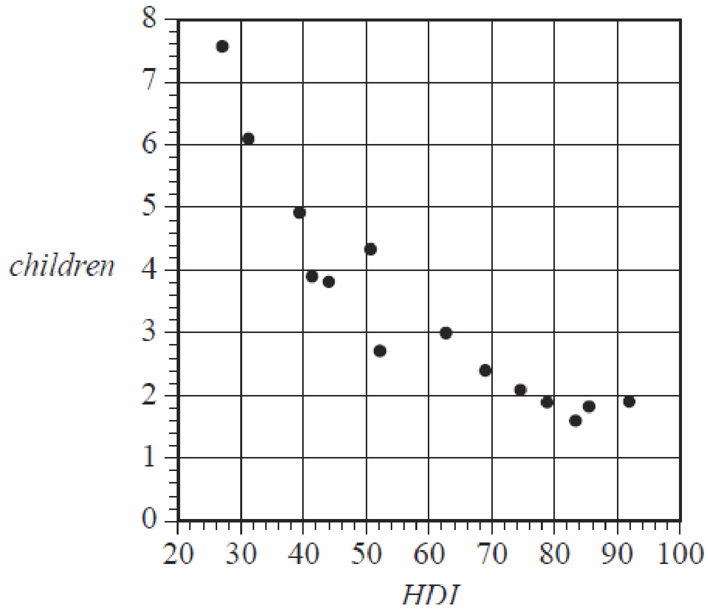
Exam Styled Questions– Multiple Choice

Question 1

(2016 Exam 1 Section A – Qn 11)

The table below gives the Human Development Index (*HDI*) and the mean number of children per woman (*children*) for 14 countries in 2007.
A scatterplot of the data is also shown.

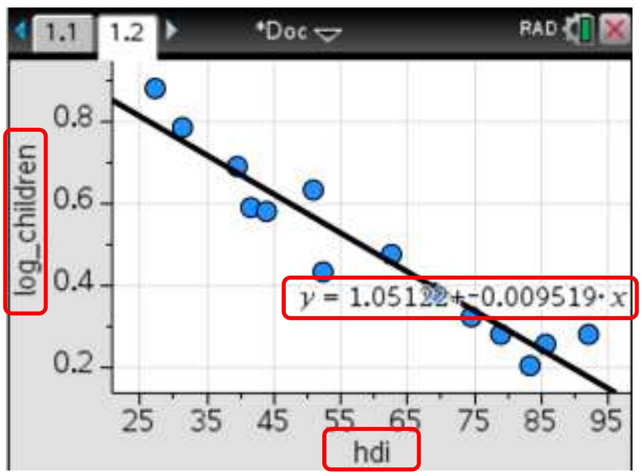
HDI	*Children*
27.3	7.6
31.3	6.1
39.5	4.9
41.6	3.9
44.0	3.8
50.8	4.3
52.3	2.7
62.5	3.0
69.1	2.4
74.6	2.1
78.9	1.9
85.6	1.8
92.0	1.9
83.4	1.6



	A hdi	B children	C log_ch...	D
=			=log(childr	
1	27.3	7.6	0.880814	
2	31.3	6.1	0.78533	
3	39.5	4.9	0.690196	
4	41.6	3.9	0.591065	
5	44	3.8	0.579784	

The scatterplot is non-linear.
A log transformation applied to the variable *children* can be used to linearise the scatterplot.
With *HDI* as the explanatory variable, the equation of the least squares line fitted to the linearised data is closest to

A. $\boxed{log(children) = 1.1 - 0.0095 \times HDI}$
B. $children = 1.1 - 0.0095 \times log(HDI)$
C. $log(children) = 8.0 - 0.77 \times HDI$
D. $children = 8.0 - 0.77 \times log(HDI)$
E. $log(children) = 21 - 10 \times HDI$



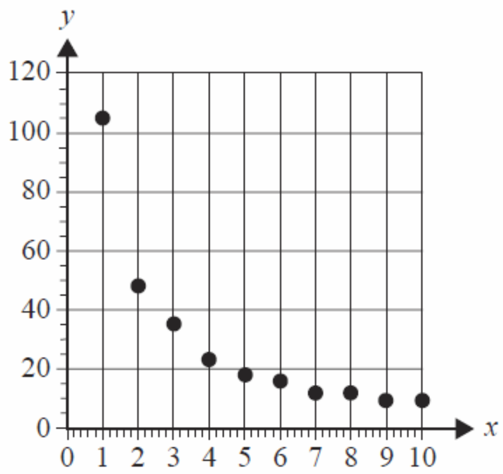
A

Therefore Option A

Question 2
(2018 Exam 1 Section A – Qn 11)

Freya uses the following data to generate the scatterplot below.

x	1	2	3	4	5	6	7	8	9	10
y	105	48	35	23	18	16	12	12	9	9

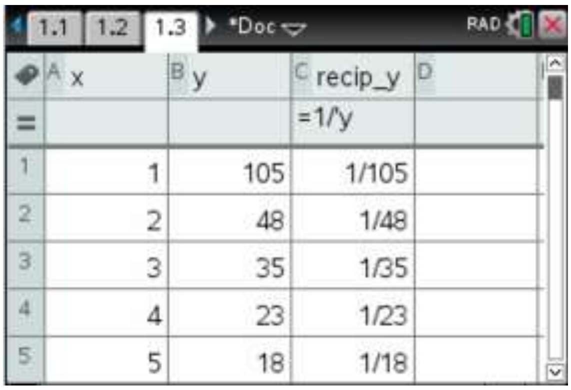


The scatterplot shows that the data is non-linear.
To linearise the data, Freya applies a reciprocal transformation to the variable y.
She then fits a least squares line to the transformed data.
With x as the explanatory variable, the equation of this least squares line is closest to

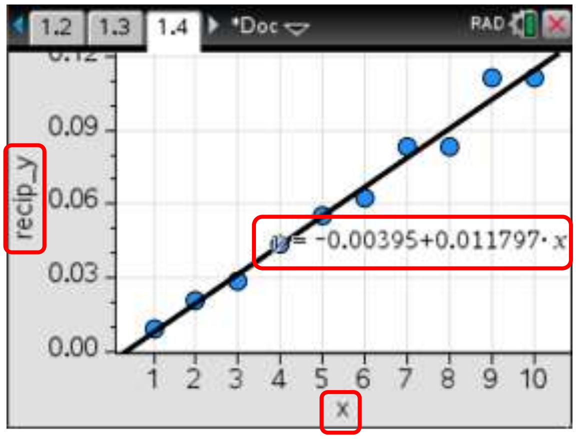
A. $\frac{1}{y} = -0.0039 + 0.012x$
B. $\frac{1}{y} = -0.025 + 1.1x$
C. $\frac{1}{y} = 7.8 - 0.082x$
D. $y = 45.3 + 59.7 \times \frac{1}{x}$
E. $y = 59.7 + 45.3 \times \frac{1}{x}$

A

Therefore Option A



	x	y	recip_y	
1	1	105	1/105	
2	2	48	1/48	
3	3	35	1/35	
4	4	23	1/23	
5	5	18	1/18	



Question 3
(2018 Exam 1 Section A – Qn 12)

A $log_{10}(y)$ transformation was used to linearise a set of non-linear bivariate data.
A least squares line was then fitted to the transformed data.
The equation of this least squares line is

$$log_{10}(y) = 3.1 - 2.3x$$

This equation is used to predict the value of y when $x = 1.1$
The value of y is closest to

A. –0.24
B. 0.57
C. 0.91
D. 1.6
E. 3.7

E

Substitute (place) $x = 1.1$ into the equation of least squares

$log_{10}(y) = 3.1 - 2.3x$
$log_{10}(y) = 3.1 - 2.3 \times 1.1$
$log_{10}(y) = 0.57$

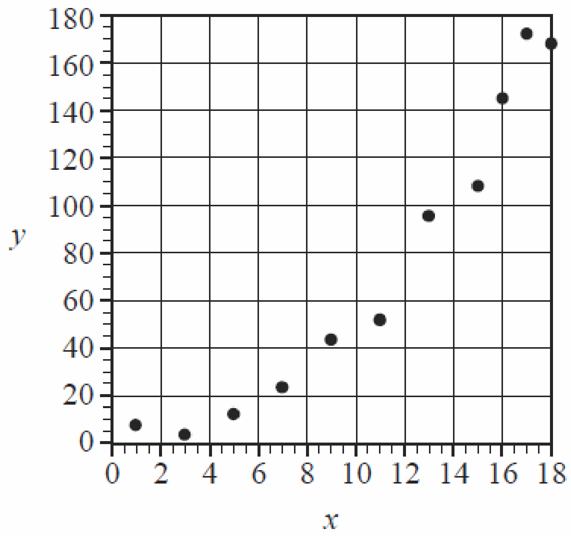
$\text{solve}\left(\log_{10}(y)=0.57, y\right)$ $y=3.71535$

Therefore Option E

Question 4
(2019 Exam 1 Section A – Qn 12)

The table below shows the values of two variables x and y.
The associated scatterplot is also shown.
The explanatory variable is x.

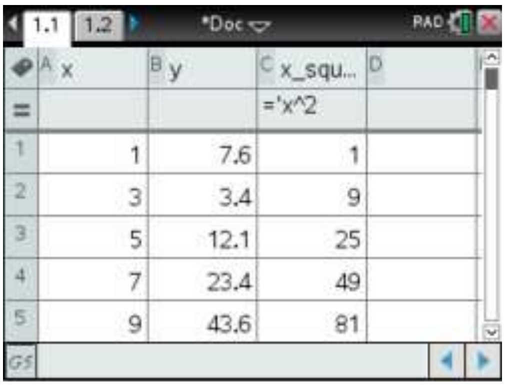
x	y
1	7.6
3	3.4
5	12.1
7	23.4
9	43.6
11	51.8
13	95.4
15	108
16	145
17	172
18	168



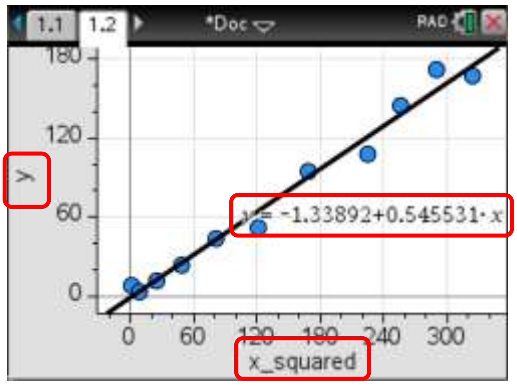
The scatterplot is non-linear.
A squared transformation applied to the variable x can be used to linearise the scatterplot.
The equation of the least squares line fitted to the linearised data is closest to

A. $y = -1.34 + 0.546x$
B. $y = -1.34 + 0.546x^2$
C. $y = 3.93 - 0.00864x^2$
D. $y = 34.6 - 10.5x$
E. $y = 34.6 - 10.5x^2$

B



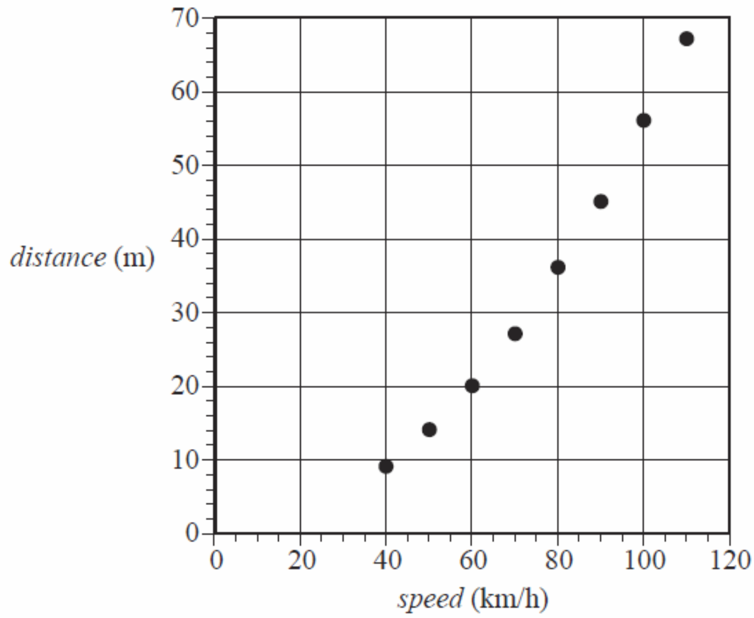
Therefore Option B



Question 4
(2017 NHT Exam 1 Section A – Qn 12)

The table below shows the *speed*, in kilometres per hour, and the braking *distance*, in metres, of a car travelling at eight different speeds. A scatterplot has been constructed from this data.

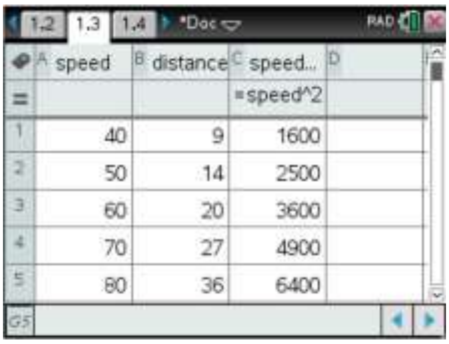
Speed (km/h)	*Distance* (m)
40	9
50	14
60	20
70	27
80	36
90	45
100	56
110	67



The scatterplot shows that the association between $distance$ and $speed$ is non-linear.
A squared transformation is applied to the variable $speed$ to linearise the data.
A least squares line is then fitted to the transformed data with $distance$ as the response variable.
The equation of this least squares line is closest to

A. $distance = -15.6 + 180 \times speed^2$
B. $distance = 0.0056 + 0.092 \times speed^2$
C. $distance = 0.092 + 0.0056 \times speed^2$
D. $speed^2 = 180 - 15.6 \times distance$
E. $speed^2 = 0.0056 + 0.092 \times distance^2$

C



A speed	B distance	C speed...	D
		=speed^2	
40	9	1600	
50	14	2500	
60	20	3600	
70	27	4900	
80	36	6400	

Therefore Option C