The log transformation

Thursday, 23 January 2020 8:17 pm

★ By the end of the lesson I would hope that you have an understanding and be able to apply to questions the following concepts:

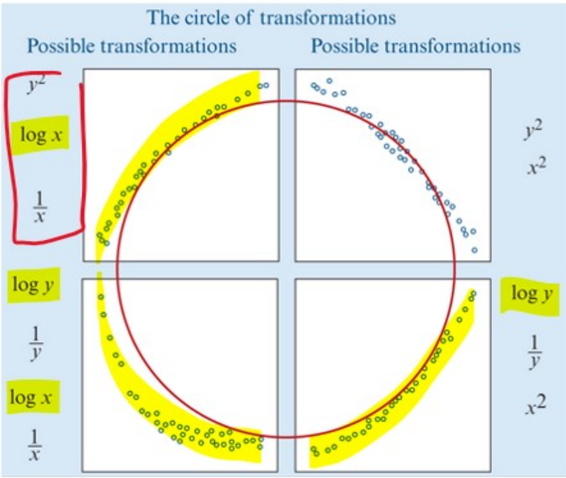
- How to apply the log transformation
- Why we need to apply the log transform

RECAP:

This lesson builds on the work we need to turn curves into lines.
In the last lesson we looked at how to use the squared transformation to turn data from a curve into more linear. This was then used to find the least squares line to allow us to predict values.

This lesson is going to look at another transform, the log transform.

Remember, the circle of transformations will help us choose the most appropriate transform.



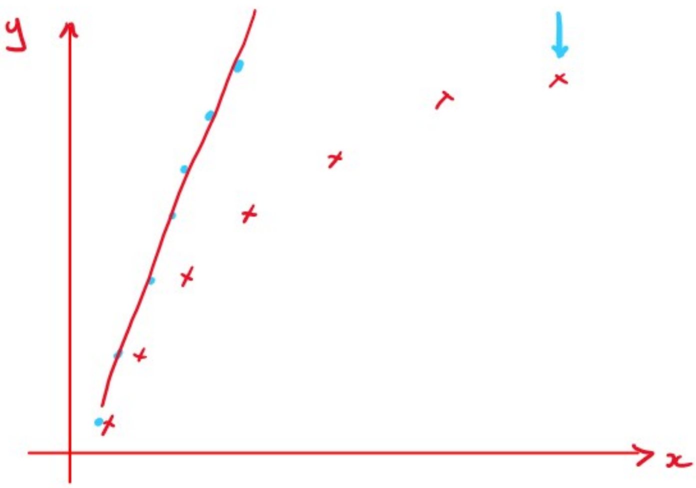
$\log_{10} 10 = 1$

$\log_{10} 100 = 2$

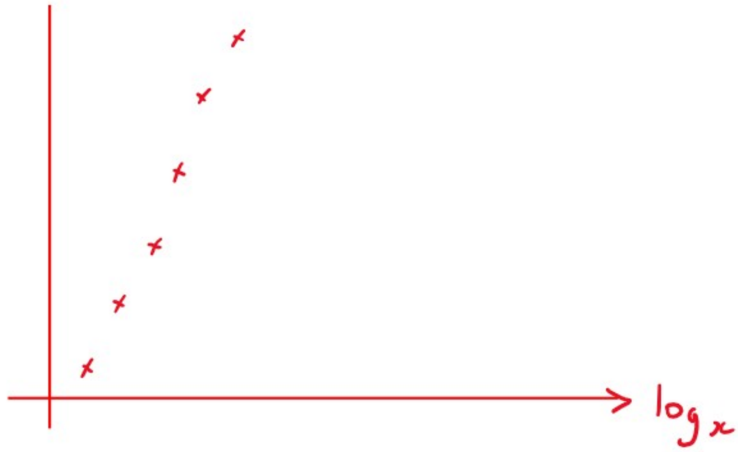
$\log_{10} 1000 = 3$

The log transform

It can be that our data looks like the following:



It would be great then if we could keep the y values the same but compress the larger x values to turn the data linear. Like so:



As it turns out, there is a way and it's called using the log transform.
This is a function on your calculator and, once you know where it is, and how to use it, the questions turn into the same as the previous lesson.

Again, you need to remember that the equation of a straight line (or the least squares line) must be written correctly.

Label of the x-axis

Label of the y-axis

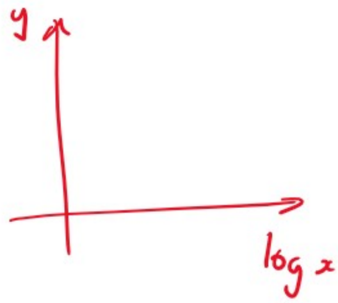
$$y = a + bx$$

y-axis intercept

Gradient of the line

$y = mx + c$

$y = a + bx$



It's **vitally important** that you realise the labels of the x- (and or y-) axes will now change.

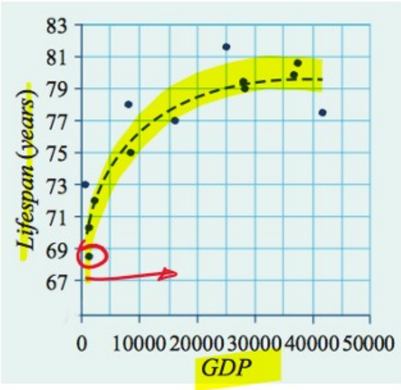
Hence, for the example above, we would have an equation in the form:

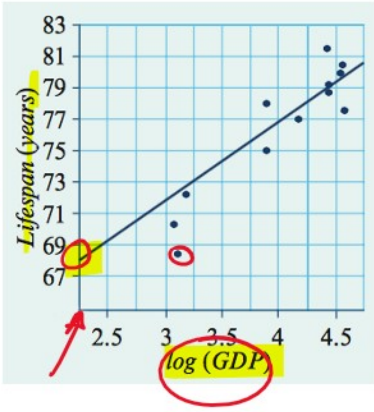
$y = a + bx$

$$y = a + b \times (\log x)$$

The x-axis is now showing all the values of $\log x$

For a real world example:





y = a + bx

0

+ve

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

Notice how this is the log of the GDP.

An example using the CAS

Remember, we can use the CAS in this whole course.
We are used to using the Statistics screen to enter values.
We will enter the raw data items we have and then get the calculator to change the x-values and then plot the appropriate graph.

The following example has been extracted with permission from the Cambridge Further Maths Units 3 and 4 textbook.

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:

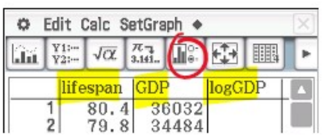
- *linearise the data, and fit a regression line to the transformed data (GDP is the EV)*
- *write its equation in terms of the variables lifespan and GDP correct to three significant figures.*
- *use the equation to predict the lifespan in a country with a GDP of $20000 correct to one decimal place.*

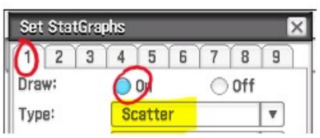
Lifespan	*GDP*
80.4	36 032
79.8	34 484
79.2	26 664
77.4	41 890
78.8	26 893
81.5	25 592
74.9	7 454
72.0	1 713
77.9	7 073
70.3	1 192
73.0	631
68.6	1 302

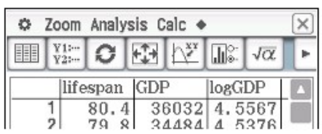
Firstly, notice they have told you which transform to use!

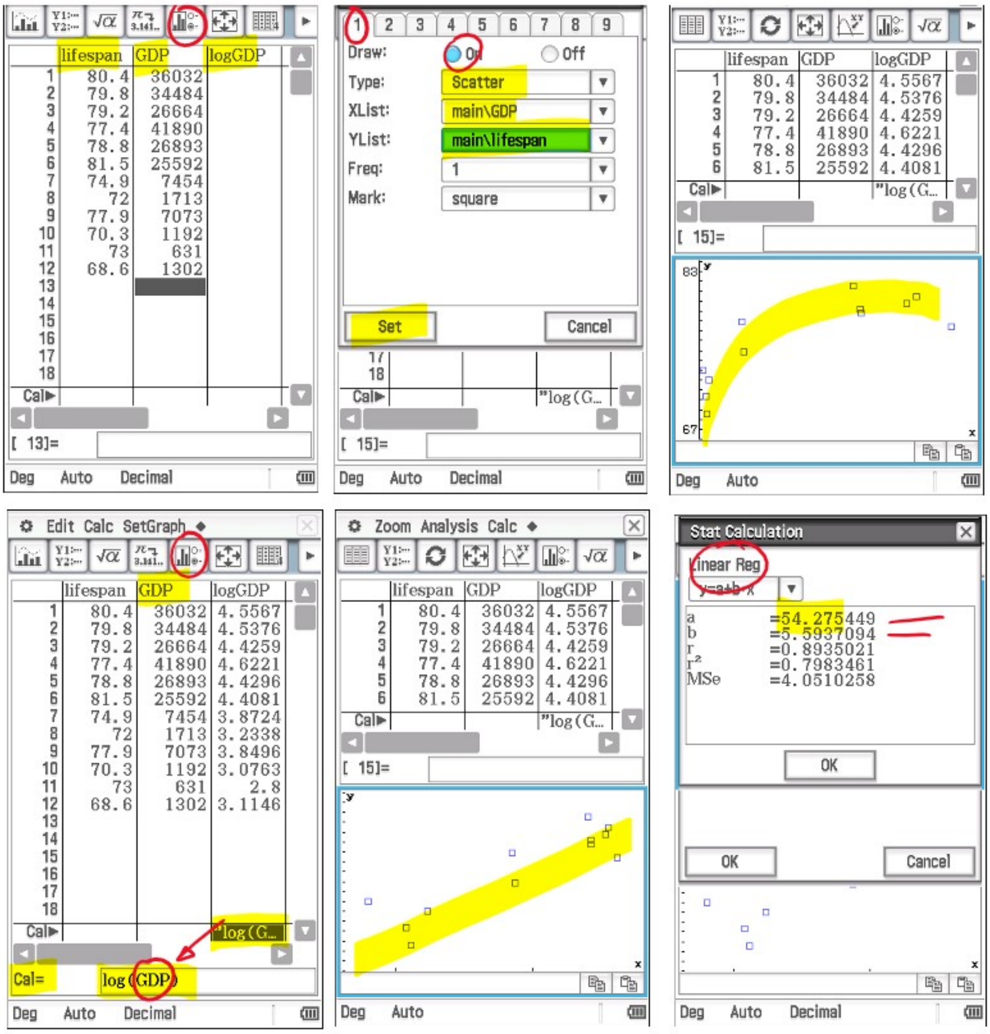
So, the rest is a case of:

- entering the data into the calculator
- Adding a column called log(GDP)
- Having the calculator do the calculations to find the least squares regression line.







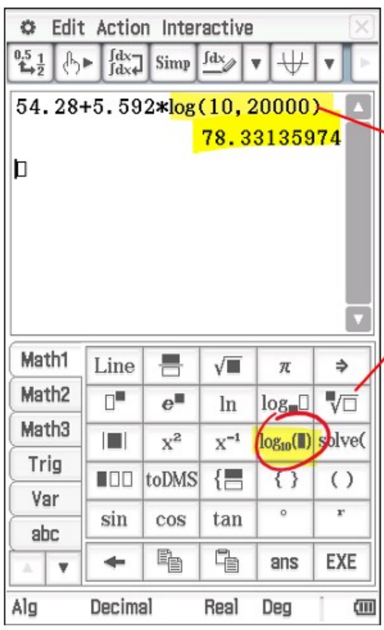


Hence, using the information above, we can see that the relationship can be expressed as:

$= 54\cdot3 + 5\cdot59 \times \log(GDP)$

$$lifespan = 54.28 + 5.593\ \times \log(GDP)$$

Answering the last part of the question with a GDP of $20,000 ...



$\log_{10}\square = \log(10,\square)$

$\log(10, 20000)$

This is the button to use for Logs.

Hence, the lifespan will be 78.3 years.

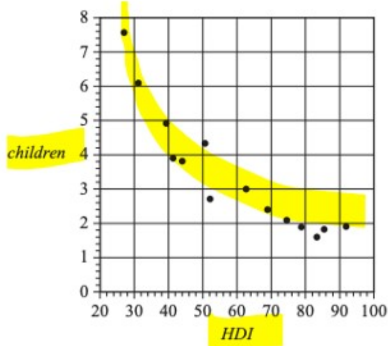
VCAA Exam Question on this concept
2016 Paper 1

Question 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



Data: Gapminder

1.051

log y!

-9.519E-3

.009519

-0.009519

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. log(*children*) = 1.1 – 0.0095 × *HDI*

B. ~~*children* = 1.1 – 0.0095 × log(*HDI*)~~

C. ~~log(*children*) = 8.0 – 0.77 × *HDI*~~

D. ~~*children* = 8.0 – 0.77 × log(*HDI*)~~

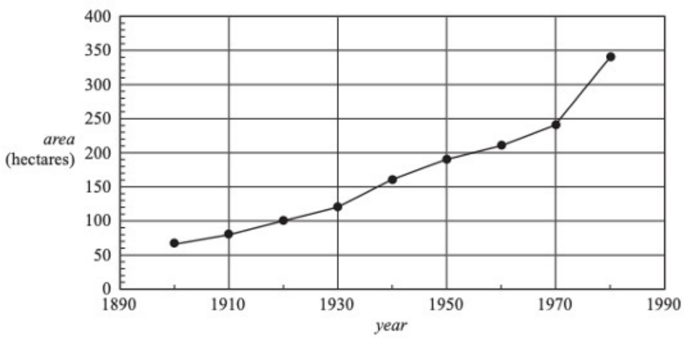
E. ~~log(*children*) = 21 – 10 × *HDI*~~

VCAA Exam Question on this concept
2017 Paper 2

Question 4 (5 marks)

The eggs laid by the female moths hatch and become caterpillars.

The following time series plot shows the total *area*, in hectares, of forest eaten by the caterpillars in a rural area during the period 1900 to 1980. The data used to generate this plot is also given.



Year	1900	1910	1920	1930	1940	1950	1960	1970	1980
***Area* (hectares)**	66	80	100	120	160	190	210	240	340

Area **(hectares)**	66	80	100	120	160	190	210	240	340

The association between ***area*** of forest eaten by the caterpillars and *year* is non-linear.

A $\log_{10}$ transformation can be applied to the variable *area* to linearise the data.

a. When the equation of the least squares line that can be used to predict $\log_{10}$ (*area*) from *year* is determined, the slope of this line is approximately 0.0085385

Round this value to three significant figures. 1 mark

b. Perform the $\log_{10}$ transformation to the variable *area* and determine the equation of the least squares line that can be used to predict $\log_{10}$ (*area*) from *year*.

Write the values of the intercept and slope of this least squares line in the appropriate boxes provided below.

Round your answers to three significant figures. 2 marks

$\log_{10}$ (*area*) = $\boxed{\quad}$ + $\boxed{\quad}$ × *year*

c. **i.** The least squares line predicts that the $\log_{10}$ (*area*) of forest eaten by the caterpillars by the year 2020 will be approximately 2.85

Using this value of 2.85, calculate the expected area of forest that will be eaten by the caterpillars by the year 2020.

Round your answer to the nearest hectare. 1 mark

ii. Give a reason why this prediction may have limited reliability. 1 mark