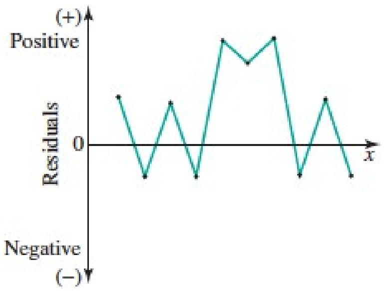
Transformations

Transformations

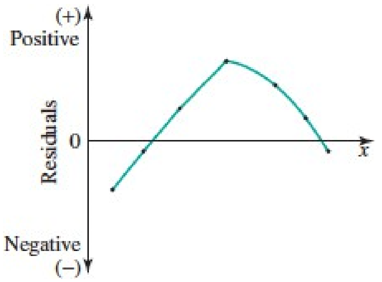
Several data sets may appear to be quite linear based upon a high Pearson product-moment correlation coefficient (r). However, when examined more closely, they may actually be better explained by a non-linear model such as a reciprocal, logarithmic or squared relationship.

There are two main measures of a data sets linearity:

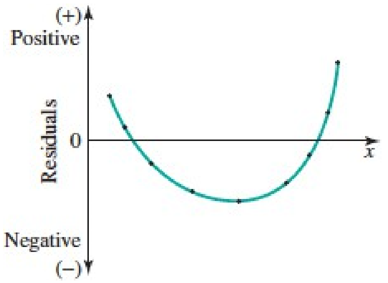
1. Pearson product-moment correlation (r). Whereby a high value represents a strong association between the two variables
2. The residual plot



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



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

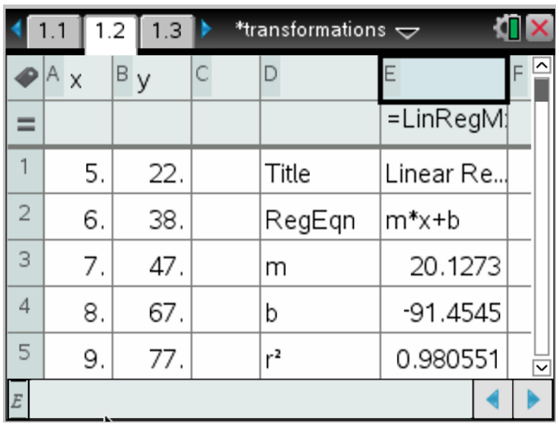


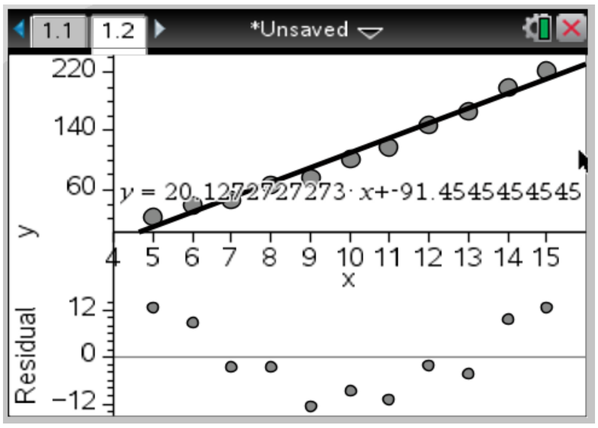
The points of the residuals show a curved pattern (∪), with a series of positive, then negative and back to positive residuals along the *x*-axis. The original data probably have a *non-linear* relationship. Transformation of the data may be required.

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

Entered into the TI-Nspire CAS calculator the linear regression provides the following results:





Summary:
Least squares regression line: Y = 20.13X - 91.45
r = 0.99

Residual plot:
The points of the residuals show a curved pattern (ᴗ) 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.***

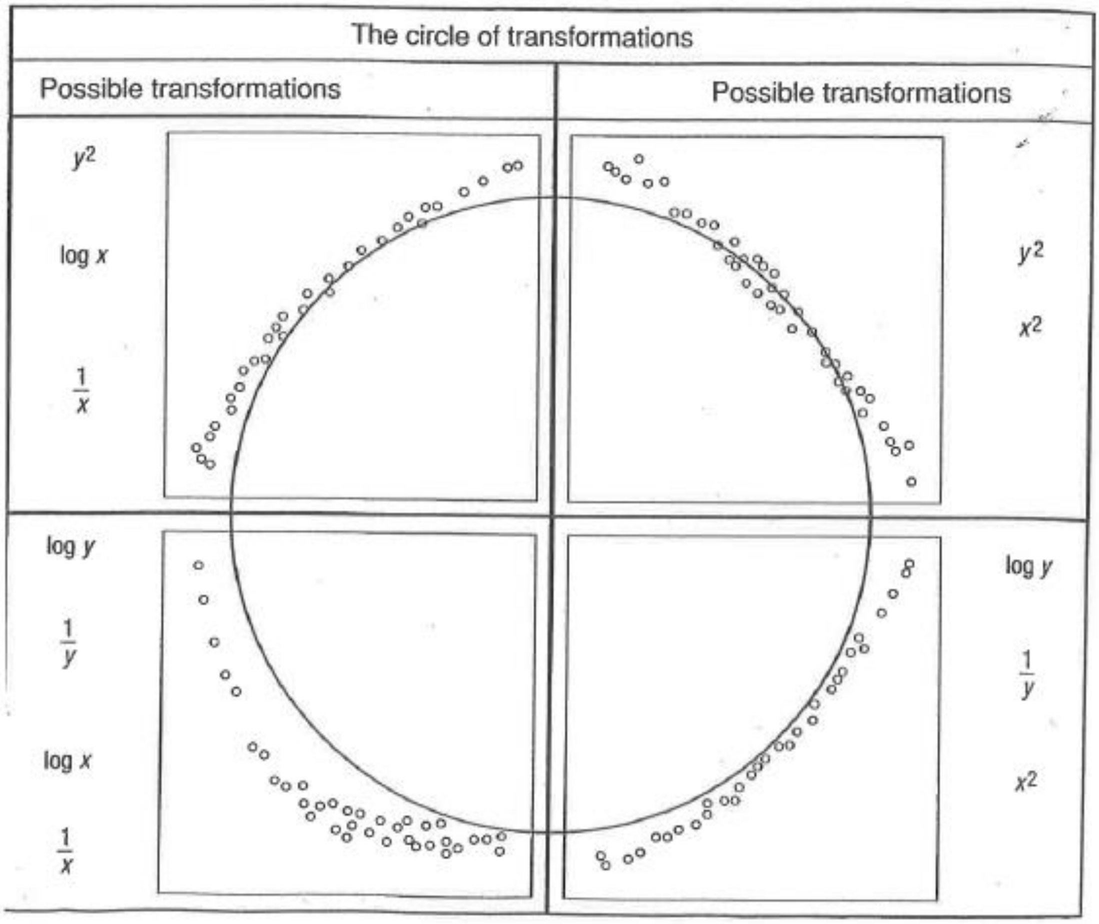
NB: a very strong correlation coefficient doesn't guarantee a linear relationship.

Transformation Options
There are six different transformations available.

X transformations	reciprocal of X: **1/X**	logarithm of X: **$\log_{10}$(X)**	X squared: **X^2**
Y transformations	reciprocal of Y: **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.



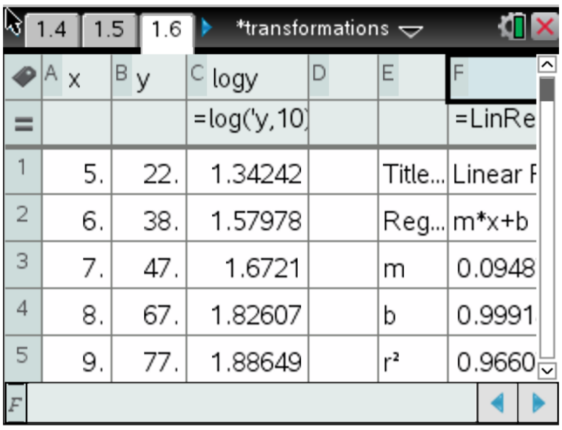
In this particular the original data fits best in bottom right quadrant of the circle of transformations.

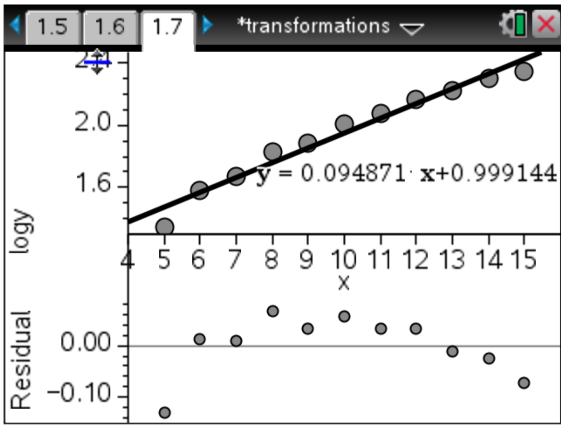
This diagram suggests we should try the following possible transformations:

- logarithm ($\log_{10}Y$), or;
- reciprocal (1/Y), or;
- x squared (X^2)

Let's now exam each of the three recommended transformations:

Transformation.1 Logarithm of Y: $\log_{10}Y$



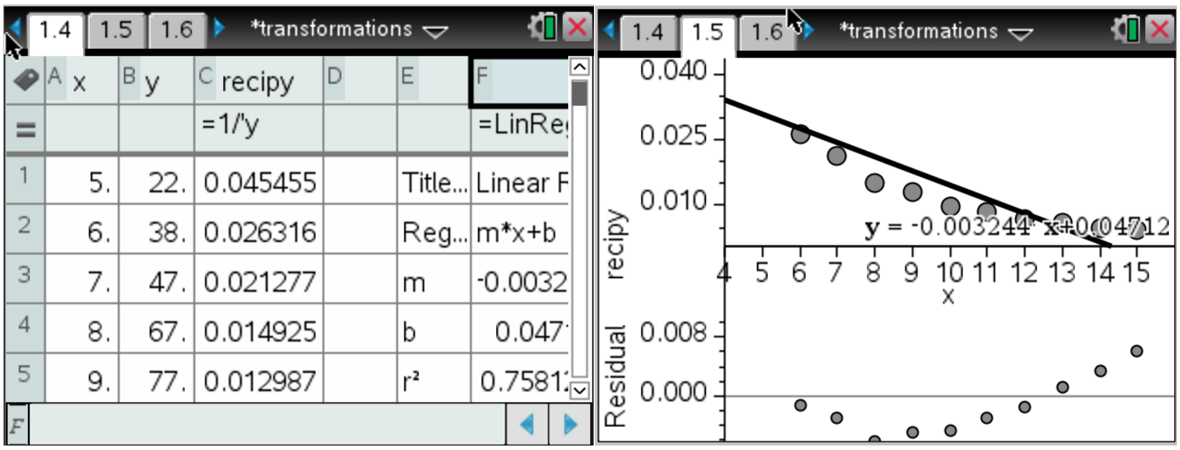


Summary:
Least squares regression line: $\log_{10}(Y) = 0.0949X + 0.9991$
r = 0.98

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. ***Transformation of the data may be required.***

Outcome:
The $\log_{10}(Y)$ transformation hasn't improved the original r value, nor does the residual plot indicate any improved linearity. Proceed to the next recommended transformation.

Transformation.2 Reciprocal of Y: (1/Y)



Summary:

Least squares regression line: 1/Y = -0.0032X + 0.0471

r = -0.8707

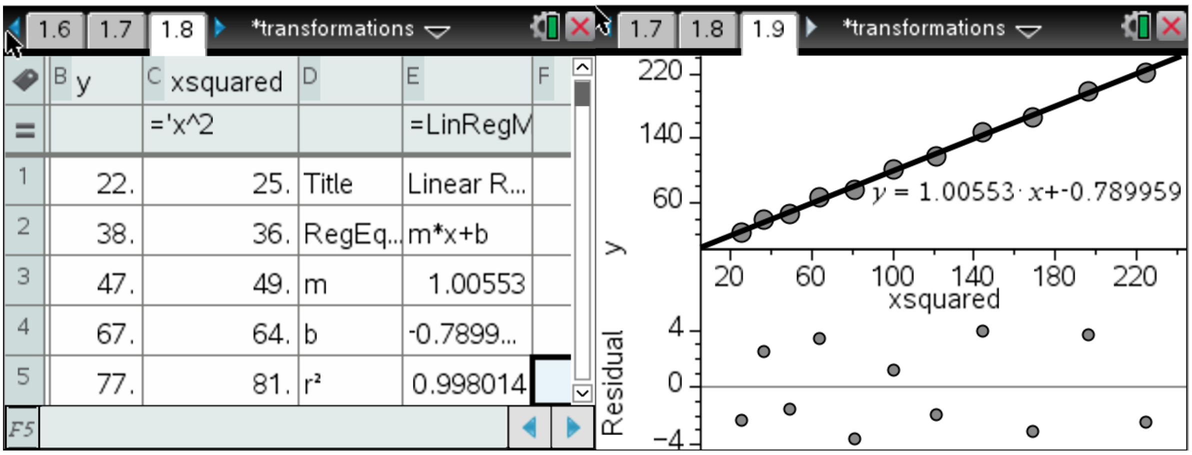
Residual plot:

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. ***Transformation of the data may be required.***

Outcome:

The 1/Y transformation hasn't improved the original r value nor does the residual plot indicate any improved linearity. Proceed to the next recommended transformation.

Transformation.3 X squared: X^2



Summary:

Least squares regression line: $Y = 1.006X^2 - 0.7900$

$r = 0.998$

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 value and the residual plot also indicate an improved linearity. This is the correct transformation for this particular set of data. Therefore $Y = 1.006X^2 - 0.7900$ is the best relationship available to these two variables.