An example of linear regression model after data transformation

We discussed the basics of linear regression in the previous article and noted the usefulness of the linear relationship model for variables x and y in making a prediction within the range of observed values for the explanatory or independent variable x.

However, not every relationship between variables x and y is linear. Some experimental data collected may visually appear to have a non-linear relationship between these two variables. In those situations, the linear regression model is inappropriate. In some cases, however, we may be able to remedy the situation by transforming the response variable in such a way that the set of transformed data shows a linear relationship with the explanatory variable x.

Mathematically we let (x_i, y_i), $i = 1, \ldots, n$ to denote our n pairs of observations and assume that a straight line does not reasonably describe the relationship between x and y.

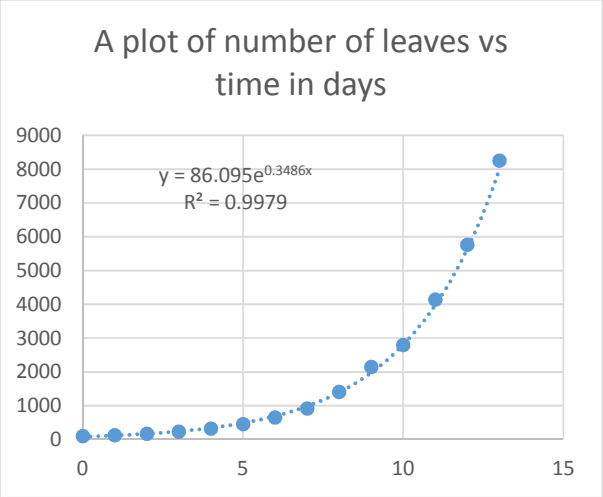
By transformation, we look for a function, f, such that the transformed variables, $z_i = f(y_i)$ can be modelled as a linear function of the x's, i.e.,

$$z = \alpha + \beta x \qquad (1)$$

For example, in a study of the growth of duckweed (*Lemna* or water lens), which are flowering aquatic plants that can float on or just beneath the surface of still water pond, by counting the number of leaves every day over a two-week period, the following data were collected as shown in the table below:

Day	Leaves	Day	Leaves
0	100	7	918
1	127	8	1406
2	171	9	2150
3	233	10	2800
4	323	11	4140
5	452	12	5760
6	654	13	8250

If we were to model the growth of duckweed as a function of day, we have the following plot:



The above plot seems to fit an exponential growth model where the population size at time t is given by the formula:

$$f(t) = c \times \exp(b \times t) \qquad (2)$$

In this example, c = 86.095, b = 0.3486, as given by the MS Excel calculations. Let's see if we can validate these data.

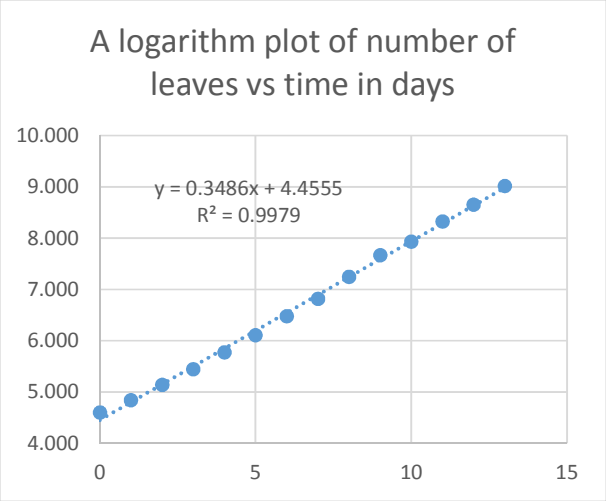
The two parameters, b and c, represent the average population increase per leave per day and the population size at day zero, respectively.

If we take natural logarithms on both sides of equation (2), we get:

$$\log(f(t)) = \underset{\alpha}{\log c} + \underset{\beta}{b} \times t \qquad (3)$$

The equation (3) corresponds to a linear regression model with $\log(f(t))$ as response and t as explanatory or independent variable.

The following figure shows a plot of the logarithm of the number of leaves versus time (days) and we see that a straight line fits the data almost perfectly.



Therefore, we have fitted a linear regression model to the logarithmically transformed leaf count and get estimates of:

$$\hat{\alpha} = 4.4555; \quad \hat{\beta} = 0.3486$$

through the least squares estimation method as described previously. We can now back transform these parameters to the original scale by the anti-logarithm function :

$$\hat{c} = \exp(\hat{\alpha}) = 86.099; \quad \hat{b} = \hat{\beta} = 0.3486$$

The above results indeed confirm the findings of MS Excel® spreadsheet's calculations.

The interpretation of the growth rate, $\hat{\beta} = 0.3486$, is that if we have k leaves in our population, then on average, we will have $\exp(\hat{b}) \times k = 1.417 \times k$ leaves in our population the following day.