Transformation of Data

Introduction

We are often in the situation where the original scatterplot of *y* vs. *x* produces a plot that has a definite non-linear trend to the data. Researchers then often work with a function of the original data, since the function can greatly simplify the statistical analysis of the data. Sometimes, however, the scientist will take the logarithm, square root, or reciprocal of a quantitative variable. Such processes mean the scientist has **transformed** the data. Researchers transform either *x* or *y* and sometimes both *x* and *y*. the effect of the transformation can be amazing in simplifying the data. Since the transformation is one to one it is easy to go back to the original units.

Linear Transformations

Transforming the data can change the scale of the data. For example, you could measure temperature in Fahrenheit or Celsius, or you could measure distance in miles or kilometers, etc. These changes are called **linear transformations**.

A linear transformation changes the original variable (*x)* into a new variable (*y)* given by

$$y = a + bx$$

Adding a constant a shifts all *x* values up or down by the same amount, thus changing the zero point. Multiplying by *b* (a positive constant) changes the size of the unit of measurement. In terms of the equation of a line, which we have previously studied, *a* is the y-intercept and *b* is the slope of the line.

Examples

1) If we were to measure distance x in kilometers, this distance could be changed to miles using the following equation:

$$y = .62\,x$$

Thus, a 10 K run could also be a 6.2 mile run. Keep in mind that zero kilometers equals zero miles.

2) If we were to measure temperature in Fahrenheit (*x*), we could convert to Celsius (*y*) using the following equation:

$$y = \frac{5}{9}(x-32) = -\frac{160}{9} + \frac{5}{9}x$$

Thus, 95^o F could also be 35^o C. Keep in mind that the freezing point of water is 32^o F but 0^o C.

Linear transformations do not change the shape of a distribution. If the original data is skewed or symmetric then the transformed variable will also be skewed or symmetric. At the same time, the center and spread of the variable will also change.

Properties of a Linear Transformation

1) Multiplying each of observation by positive *b* multiplies both the measures of center (mean and median) and the measures of spread (interquartile range and standard deviation) by *b*.

2) Adding the same number, *a*, to each observation adds *a* to the measures of center and other percentiles, but *does not* change the spread.

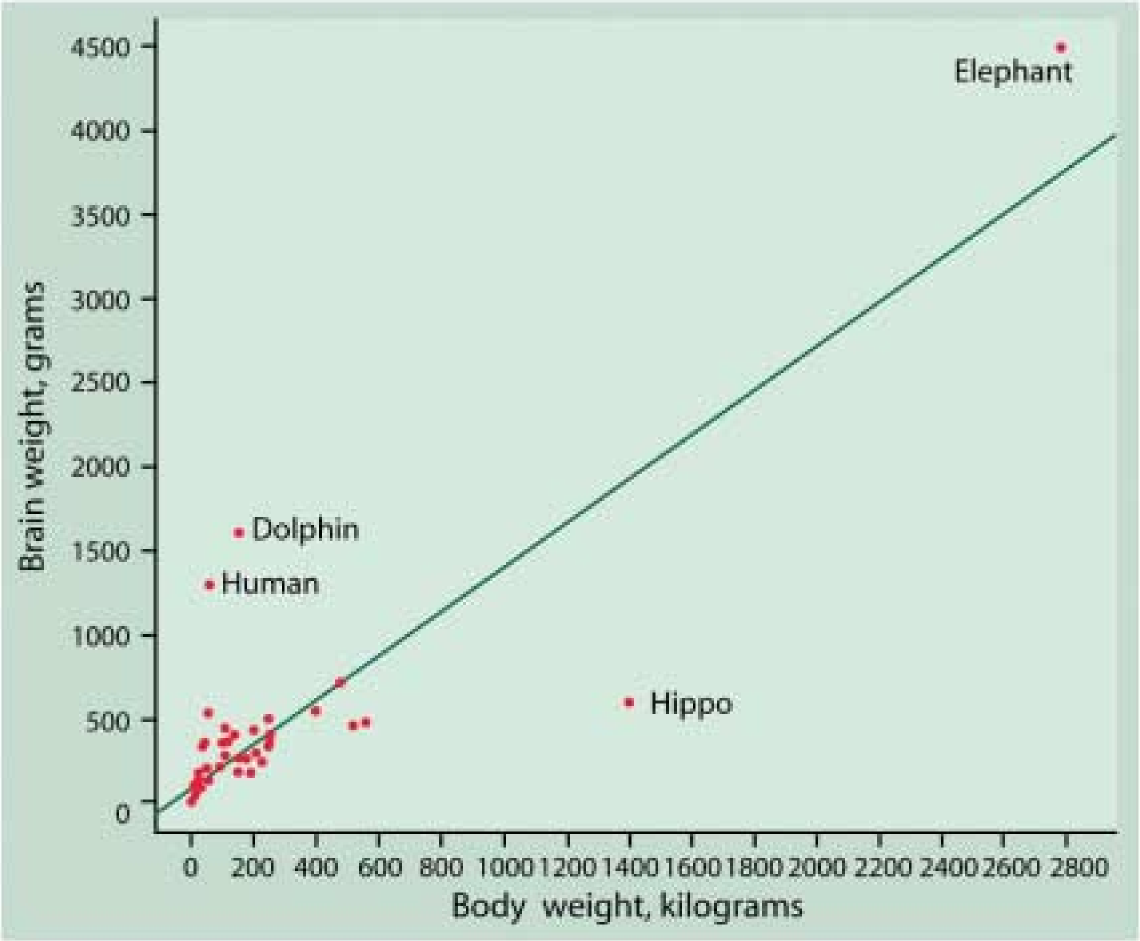
Example

1) Jenny and Becky investigate the cost of products to consumers. Jenny measures the cost in dollars of 5 products whereas Becky measures the cost in cents.

Product	1	2	3	4	5
Jenny (dollars)	5	4	1	2	3
Becky (cents)	500	400	100	200	300

Remember the product cost for Jenny and Becky is identical except for the units. Prove to yourself the rules above (the properties of a linear transformation) apply to the product data.

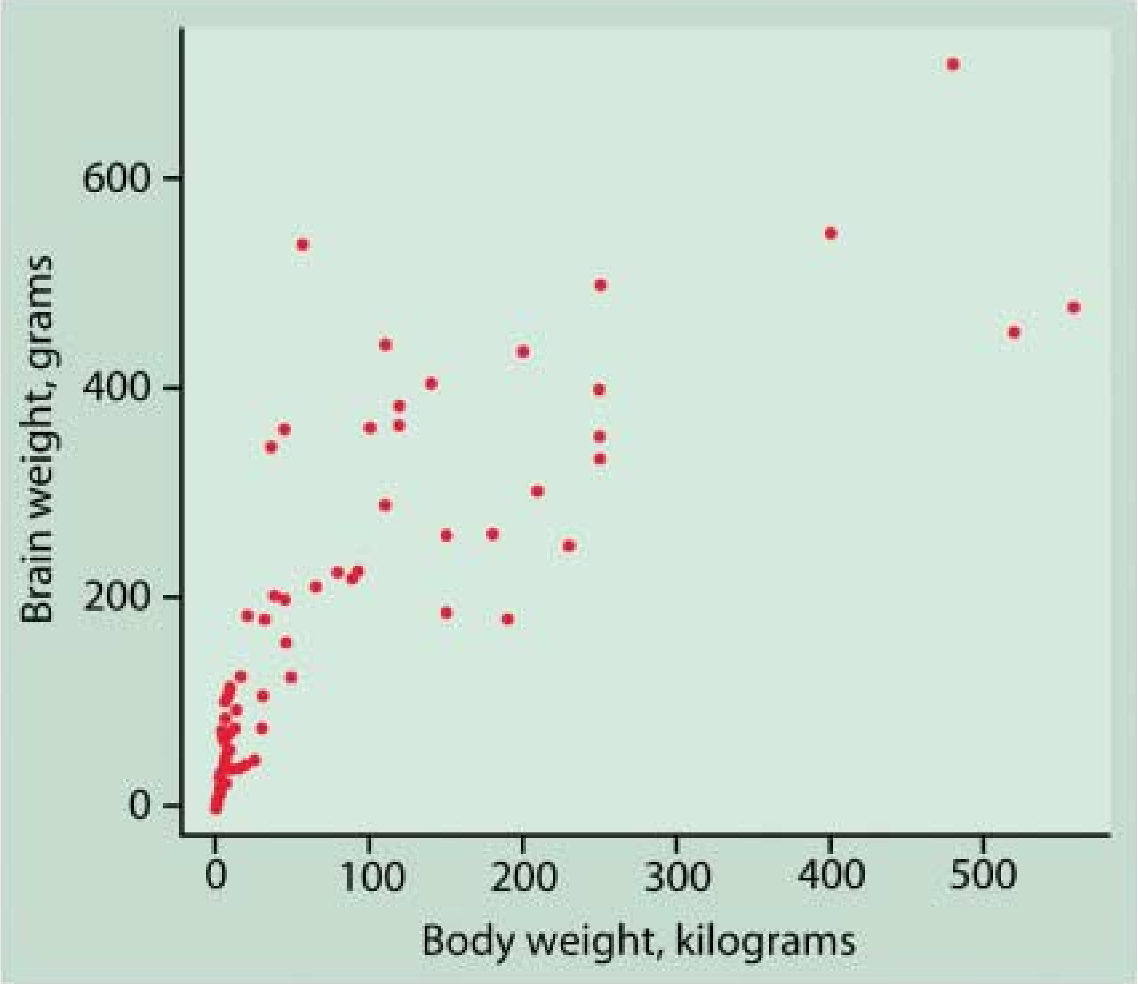
2) How is brain weight related to body weight?
Examine the graph below. Notice that most of the data is clustered in the bottom left corner.



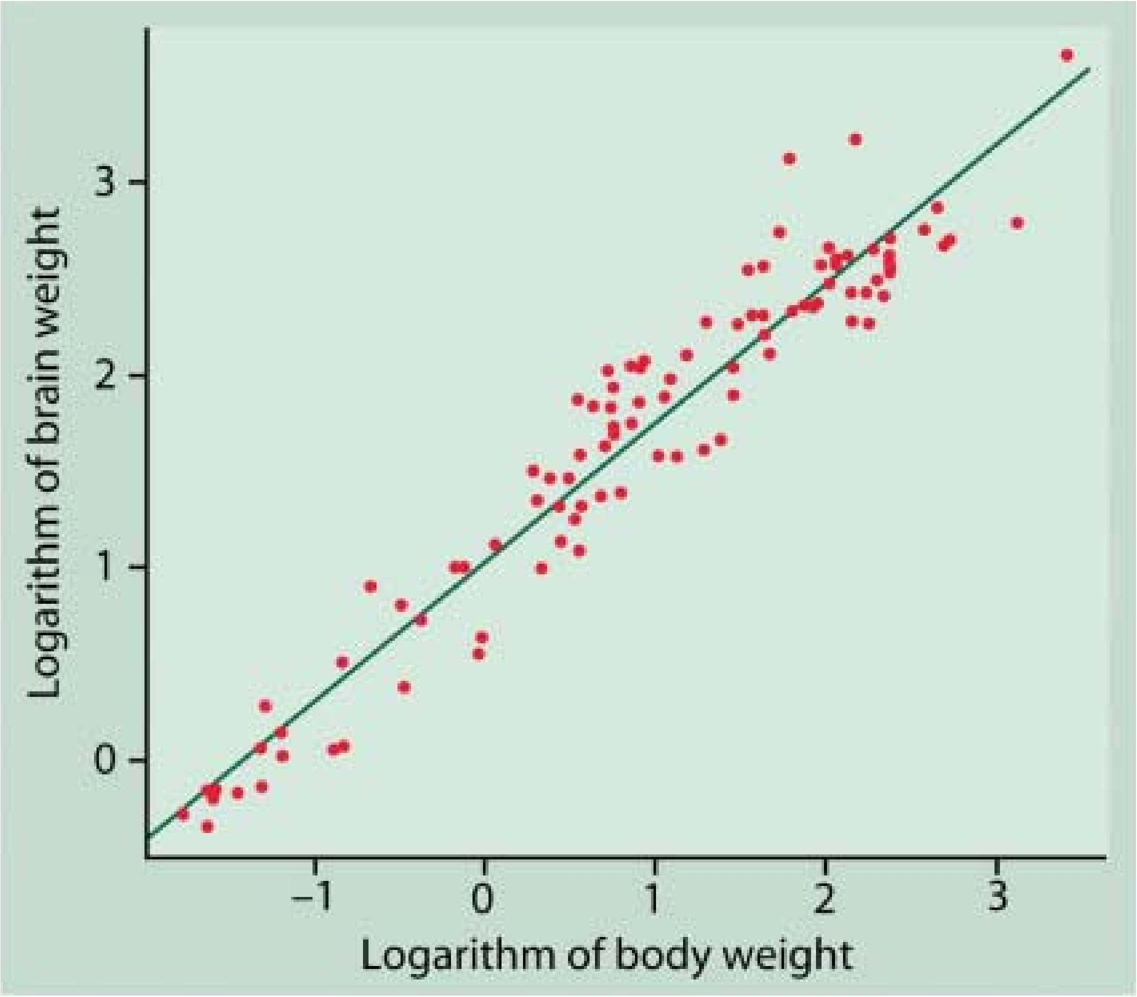
What is the correlation, r, for the above relationship between brain and body weight? Now we want to remove the elephant from our data because it is so large. What is the new correlation (r) between brain and body weight with the elephant removed?

In the graph below, the outliers of the Elephant, Hippo, Dolphin, and Human have been removed. Notice the relationship between brain and body weight has now become non linear.

The graph now bends to the right – it is not linear.



Try taking logarithm of data (include outliers). Take the logarithm of both *x* and *y* and plot the transformed data.



Amazing – things are linear!!

Transformation Notation

As noted previously, applying a function to the data (such as a log, square root, etc.) is called a transformation. We denote a transformation as *t*, which is the transformed variable of *x* or *y* etc.

It was also mentioned earlier that temperature can be measured in F or C, or distance in miles or kilometers. These changes are called linear transformations, where linear means they *cannot* straighten a curved relationship between two variables.

To be able to straighten a curved relationship, we use a nonlinear function, such as the logarithm transformation in the above examples. Some other non-linear transformations will be discussed below.

Non-Linear Functions

We have mentioned the non-linear transformation of the logarithm. Other non-linear transformations are powers (t^1, t^2, t^3) and reciprocals (which are negative powers: t^{-1}).

Example

A car that gets 25 Km to the liter can be represented as:

$$\frac{1}{\text{km per liter}} = \frac{1}{25} = .04 \text{ liters per km}$$

Reciprocals are negative powers, which are represented as:

$$1/t = t^{-1}$$

Monotonic functions

The logarithm, linear, and positive (t^2) and negative powers (t^{-2} – reciprocal) are called monotonic. A linear function is a line. A monotonic increasing function is increasing everywhere (for example, the line of a positive slope). A monotonic decreasing function is decreasing everywhere (for example, the line with negative slope). We can try these transformations on our data to see if we can simplify our scatterplot to a form that is easier to interpret or perform statistical analyses. Remember how the logarithm of brain and body weight simplified the scatterplot of the species.

A monotonic function, denoted as f(*t*), moves in one direction as *t* increases.

A monotonic increasing function preserves the order of data, thus, if a > b, then f(a) > f(b)

$$(a)\quad \left\{\begin{array}{lll} \text{Examples} & & \\ \quad\text{Linear} & a + bt; & \text{slope } b > 0 \\ \quad\text{Square} & t^2 & \\ \quad\text{Logarithm} & \log t & \end{array}\right.$$

A monotonic decreasing function reverses the order of the data. If a>b, then f(a) <f(b)

(b) Examples

Linear	$a + bt$; slope $b < 0$
Reciprocal square root	$1\sqrt{t}$ or $t^{-1/2}$
Reciprocal	$1/t$ or t^{-1}

Recipe for Creating Monotonic Functions

(1) Apply a transformation to make all data positive. To do this, you often just neet to add a constant to the data.

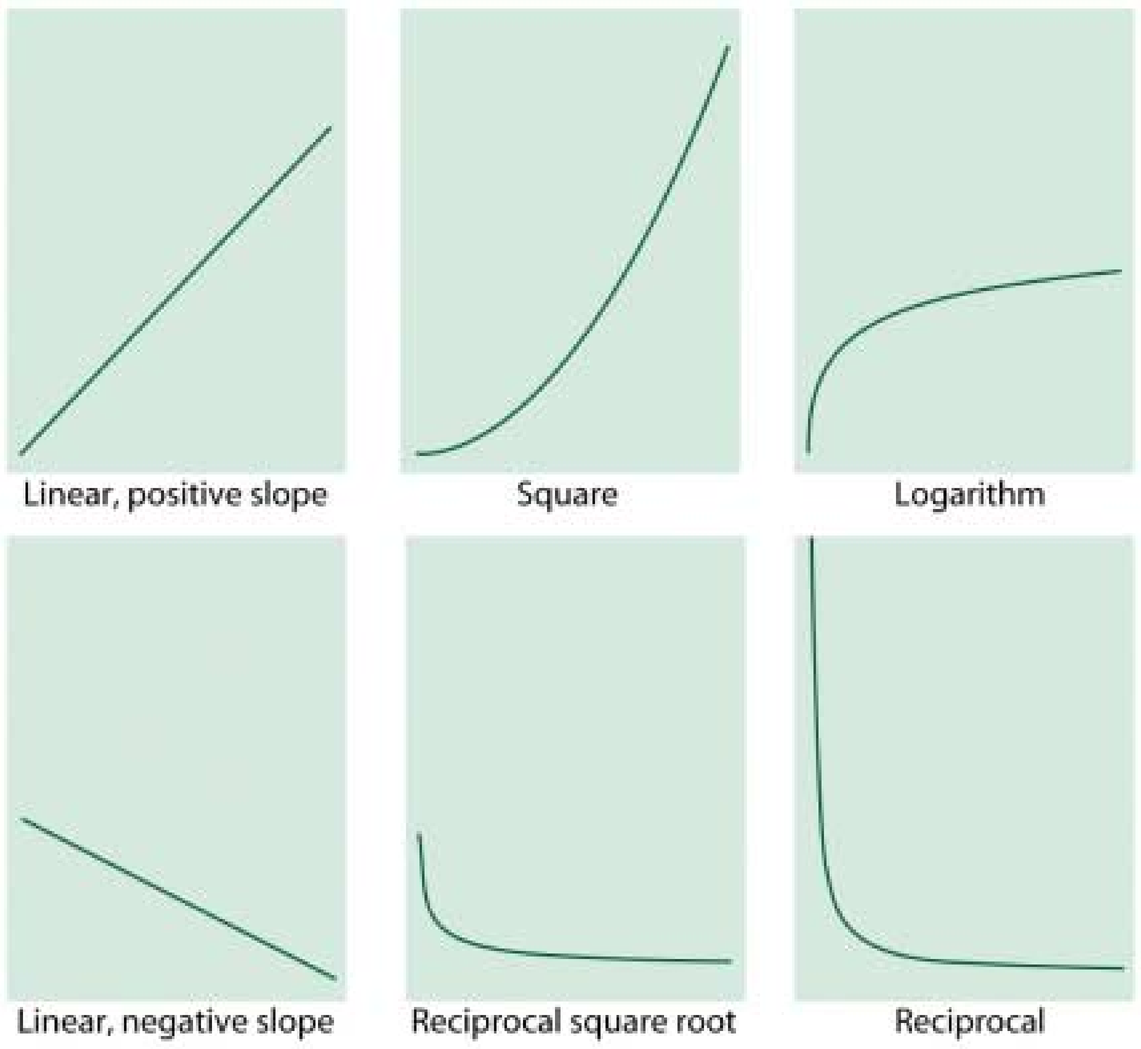
(2) Choose a transformation that simplifies the data, ie. a transformation that straightens the scatterplot.

Samples of Power Transformations

Think of powers such as

$\ldots, t^{-1}, t^{-1/2}, t^{1/2}, t, t^{2}, \ldots$

as possible power transformations for your data. In the graph below what functions are monotonic increasing/decreasing?



Linear, positive slope

Square

Logarithm

Linear, negative slope

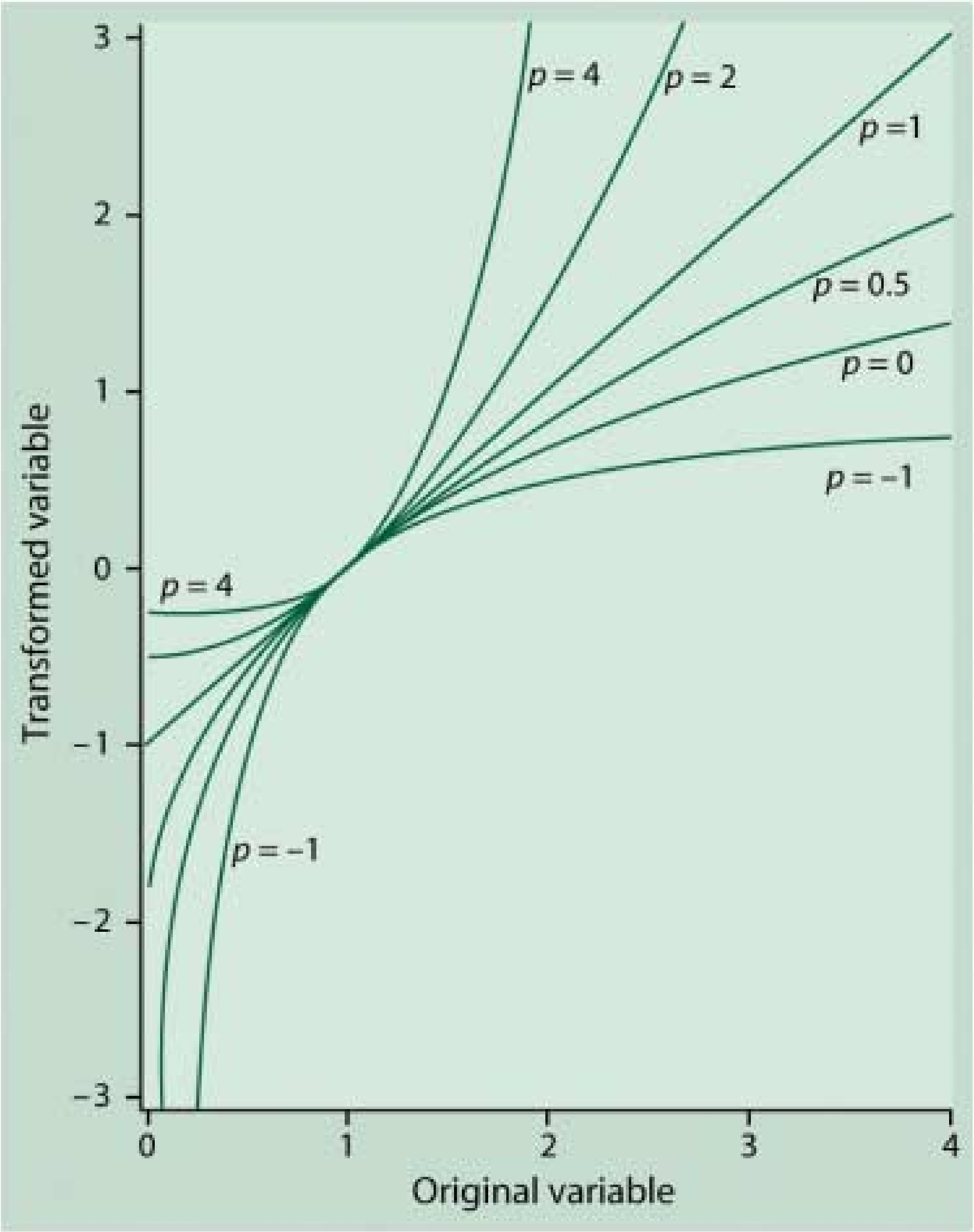
Reciprocal square root

Reciprocal

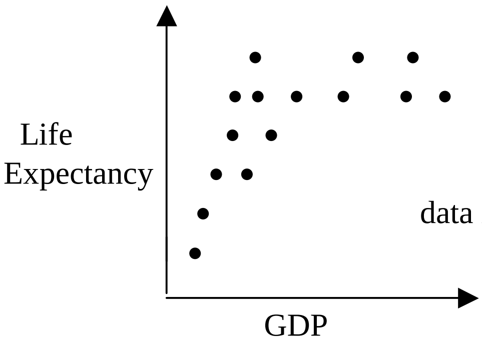
Some facts about these transformations will help us to choose the right one for our data:

Power functions are denoted as t^p for positive powers of *p*. These power functions are monotonic where they are increasing for values of $t > 0$. Keep in mind that power functions preserve the order of data (note: this is the same for logarithmic transformations). Negative powers are denoted as *p* and are monotonic functions where they are decreasing for values of $t > 0$. Power transformations (t^p), for powers of *p* greater than one, are concave up (ie. like the letter ∪) they push out the right tail and suck in the left. For *p* less than one, these power transformations are concave down, like ∩) they suck in the right tail and push out the left.

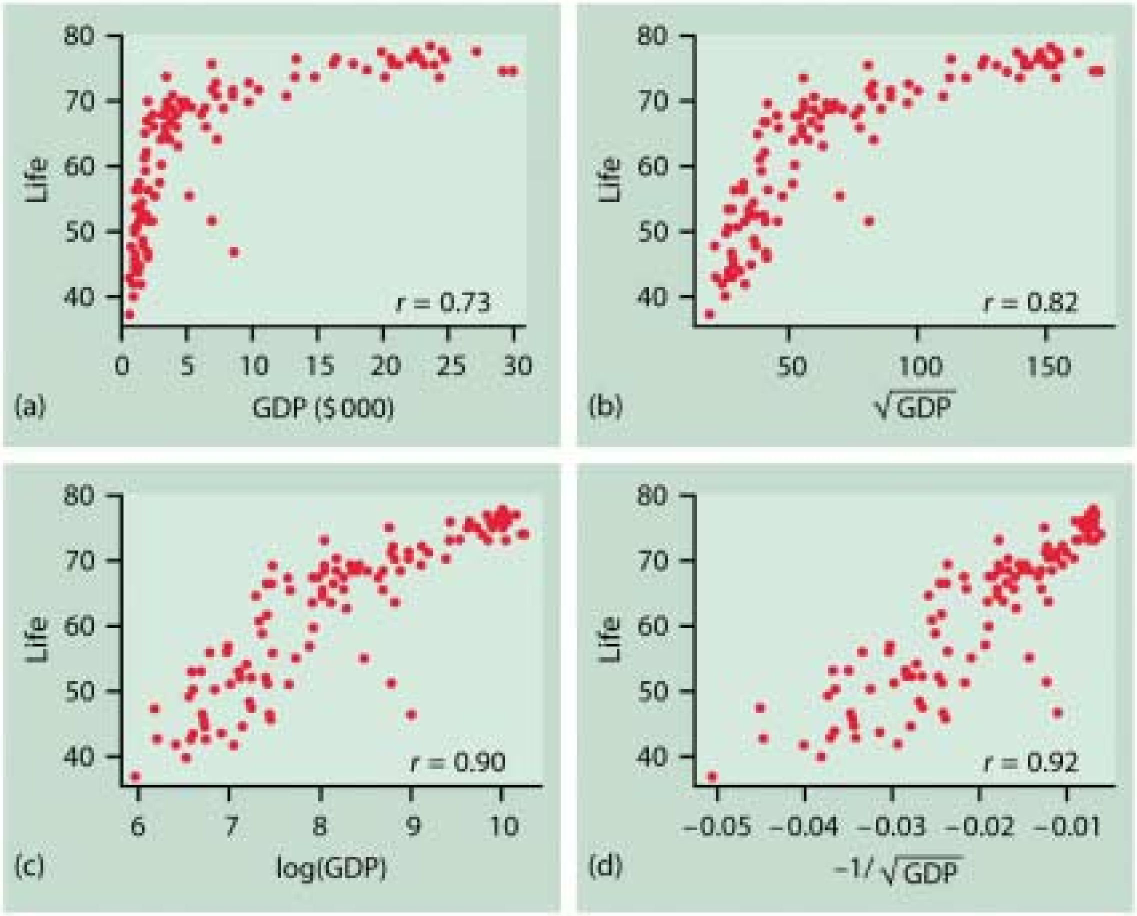
The graph below shows the effect of p, note the logarithm is given by p=0.



Example: a graph of life expectancy and GDP may look like the following:



data is skewed to the right



Note the effect of the various transformations on the original data in a. The data is skewed to the right so we wish to bring in the right tail consequently p is < 1.

Exponential Growth

A variable grows linearly if over time it adds a fixed increment in each time period. Thus, exponential growth occurs when a variable is multiplied by a fixed number in each time period.

Linear growth

Increases by a fixed amount in each equal time period.

Exponential growth

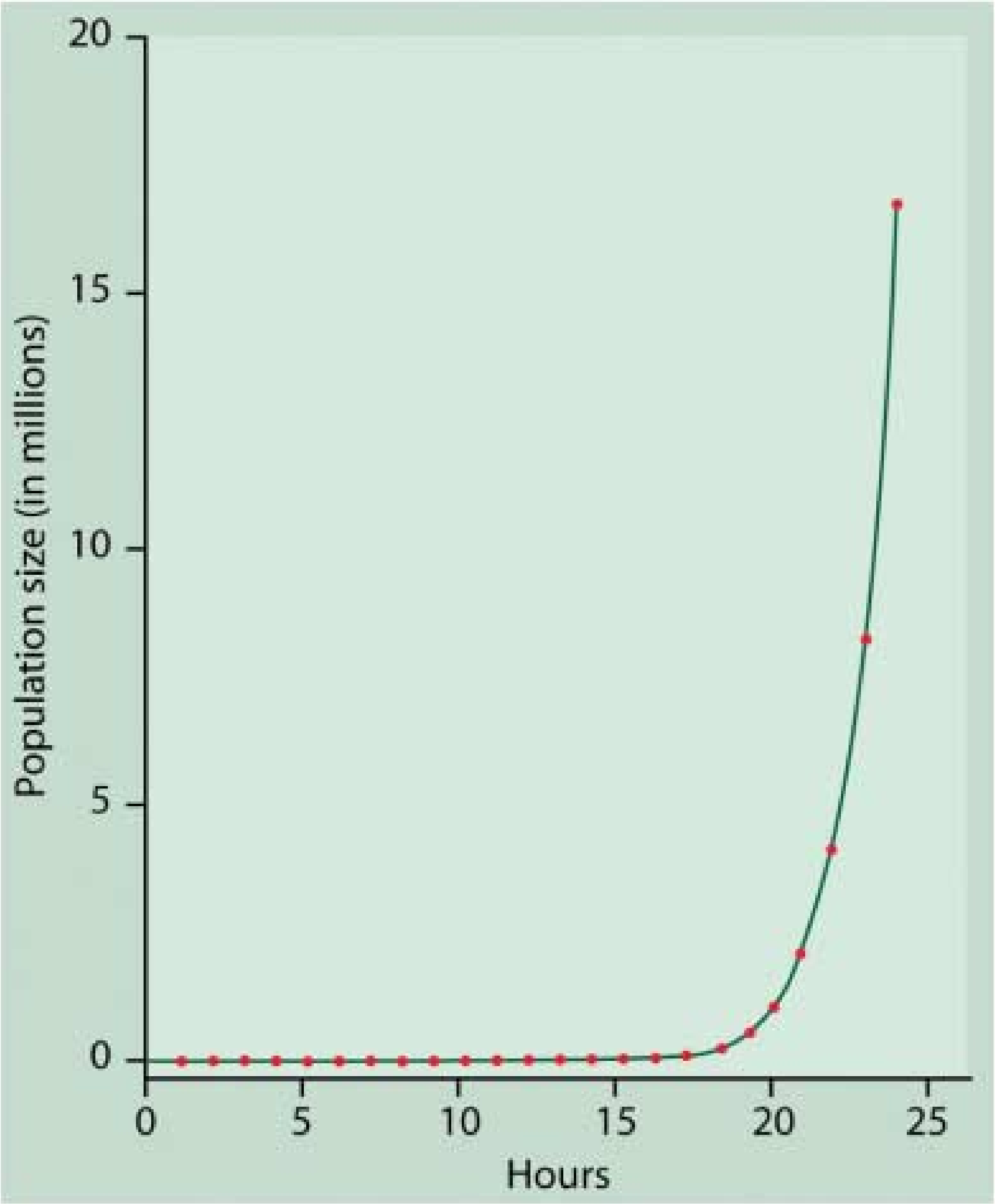
Increases by a fixed percentage of the previous total.

Example

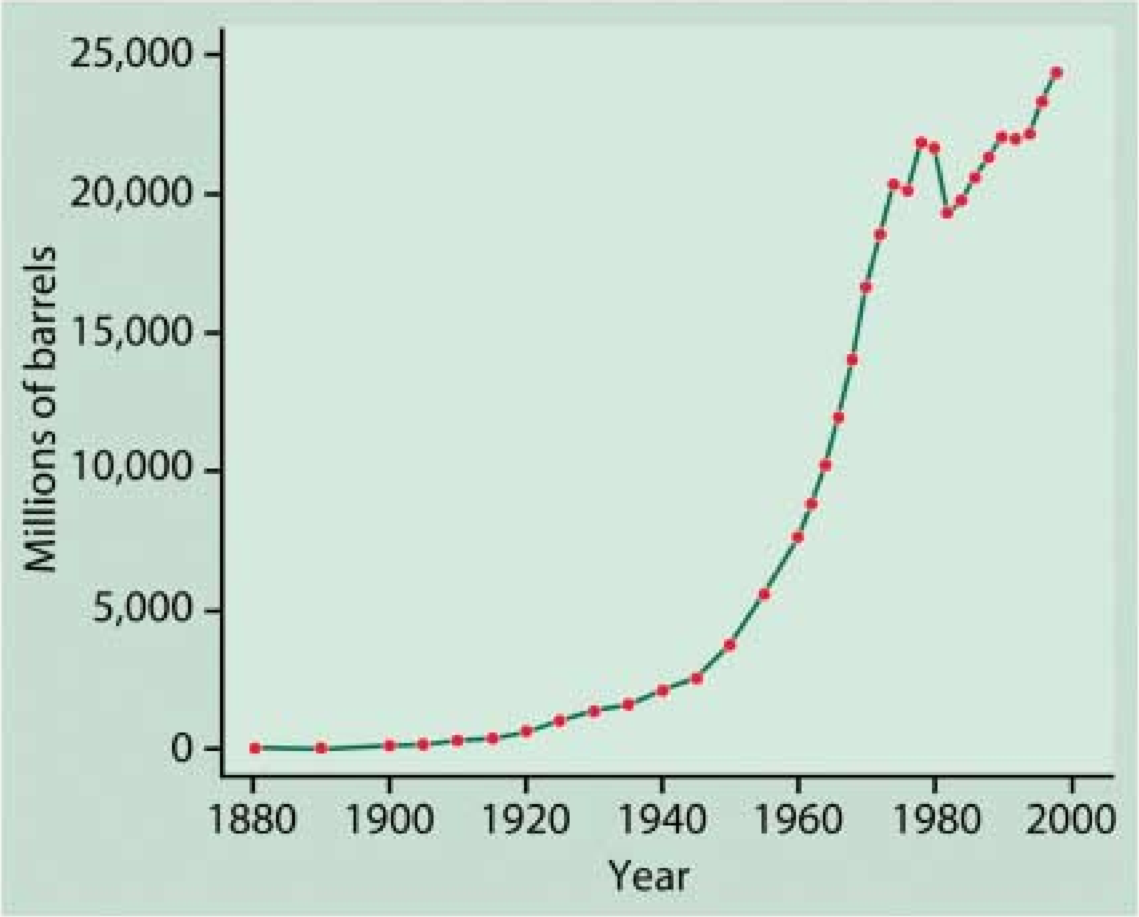
Invest a dollar at the annual rate of 6% interest and in one year you will have \$1.06, whereby you have earned six cents (1 x .06 = \$1.06). Thus, any amount deposited earns 6% annually where the amount is multiplied by 1.06. For the second year, we have 1.06 x 1.06 or $(1.06)^2$ = \$1.12. For the third year, we would multiply 1.12 x 1.06, and on and on for each subsequent year. After *x* number of years, one dollar becomes 1.06^x dollars.

The value of \$5 invested for *x* years at 6% interest is calculated by 5 x 1.06^x. This is an example of exponential growth; $y = a \text{ x } b^x$ for the different constants, *a* and *b*. The response, *y* is multiplied by *b* in each time period.

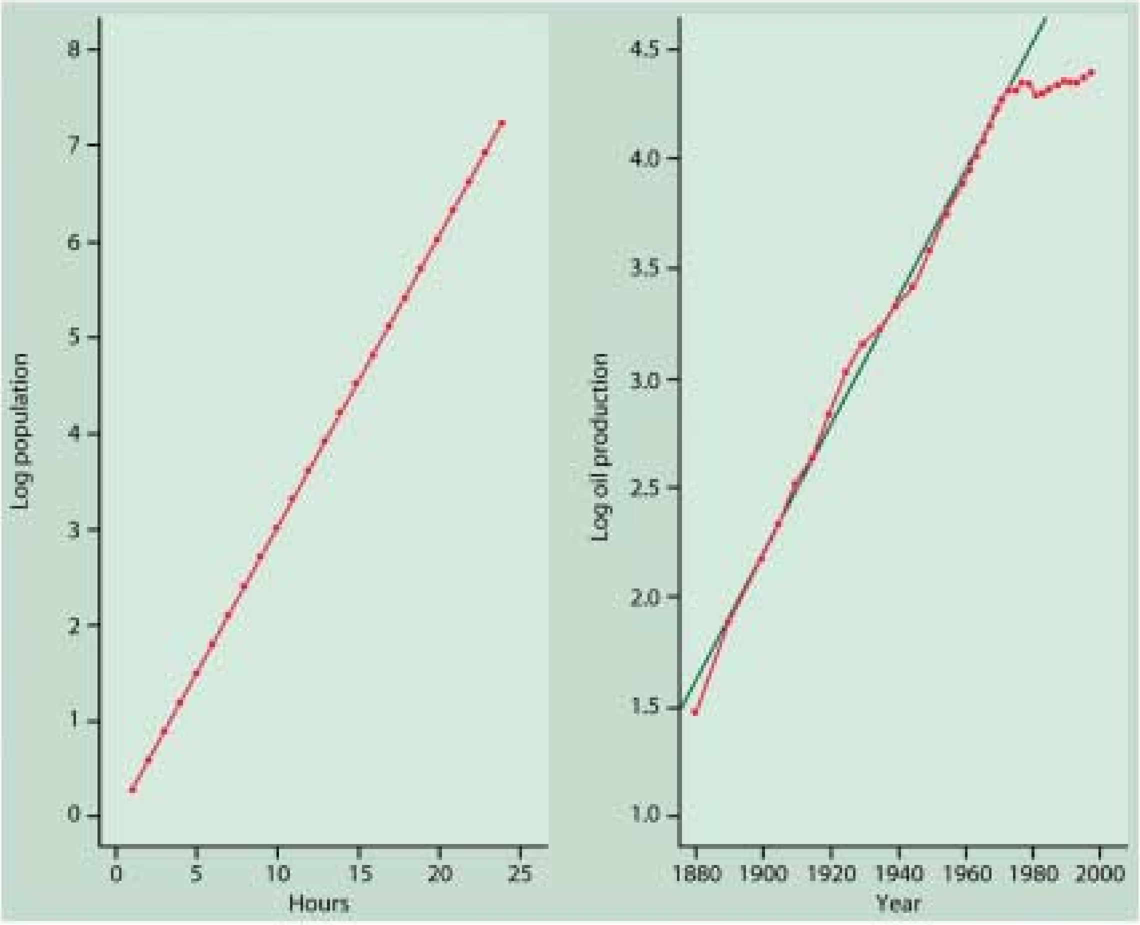
If a variable grows exponentially, its logarithm grows linearly. So we transform one variable by taking its logarithm and thus 'straighten' the relationship.Below we have graphs of bacteria growth and world oil production



Bacteria that double each hour,2^x where x is in hours.



Note below the remarkable impact of transforming the data by taking logarithms.



<u>Power Law Models</u>

Say pizzas come in 10, 12 and 14 diameters. But what you eat is the area. The area of a circle is given by π r^2 where r is the radius. So a round pizza with diameter x is

$$\text{Area} = \pi r^2 = \pi (x/2)^2 = \pi (x^2/4) = (\pi/4) x^2$$

This is a power law model of the form

$$y = ax^p$$

where a=(π/4) and p=2

Power law models become linear when the log transformation is applied to both variables, consider

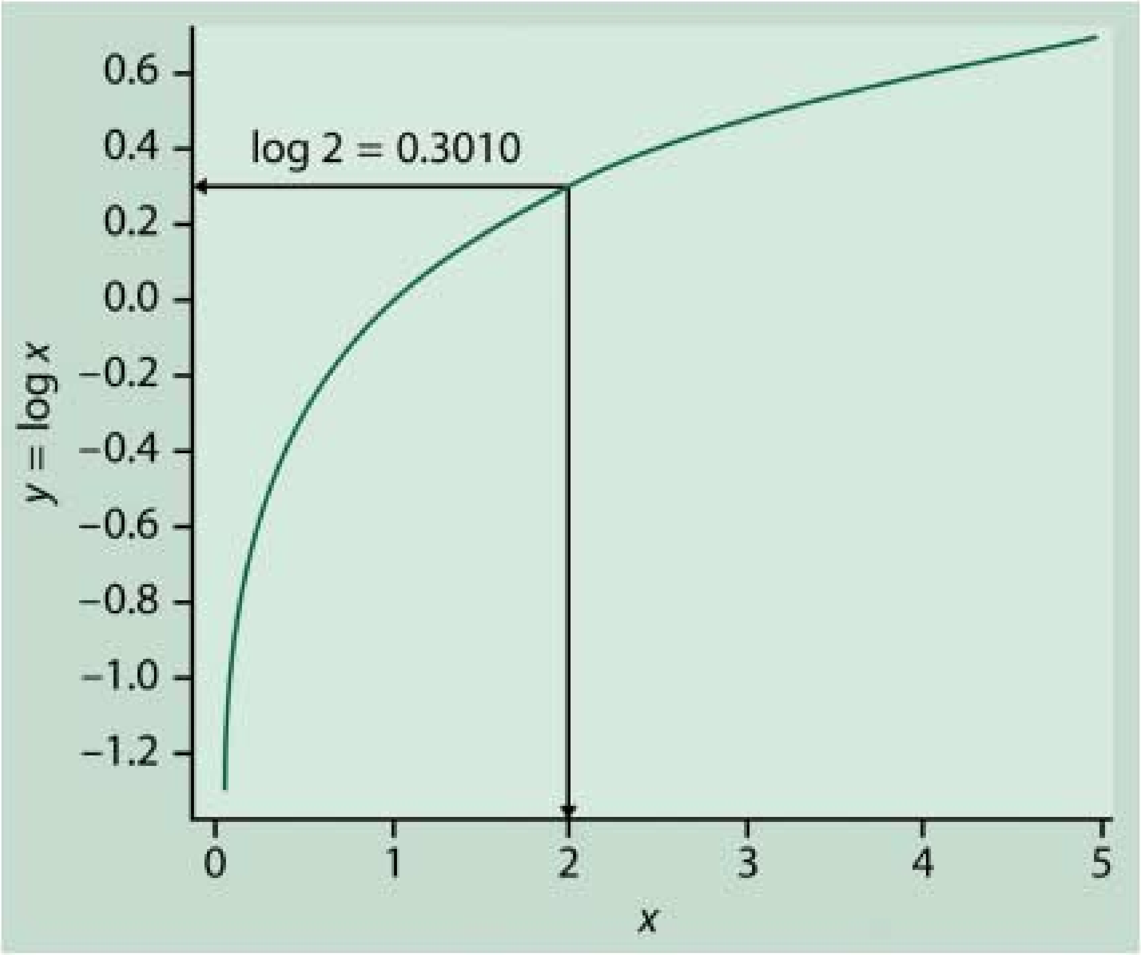
$$y = ax^p$$

now,take the logarithm of both sides

$$\log y = \log a + p \log x$$

by taking the logarithm of both x and y we straighten the scatterplot of y vs x. We see the form of a linear equation with corresponding intercept and slope in the above expression. The power p in the power law becomes the slope of the straight line that links logy to logx. When considering power models both y and x are transformed as is seen above whereas in the case of exponential models only y is transformed.

Note(see below) if x=2 and we take the logarithm of x we obtain logx=.3010. we also can go in reverse, if given the logx=.3010 we can recover 2 since the transformation is one to one or unique. Any number x can be obtained from its common base 10 logarithm logx by $x=10^{logx}$.
For example if logx=.3010 then $x=10^{logx}=10^{.3010} =2.0$.Now if the log of y is predicted from x we have an equation say logy=-52.7+.0289x. Substitute into the equation x in original units and obtain a prediction for log y(4.4353). Now take logy back to y, $10^{4.4353} =27246$.
If logy =1+.2logx we can reverse the transformation,$y=10^{1+.2logx} =10^{1}\times 10^{.2logx}=10 \times (10^{logx})^{.2}=10 \times x^{.2}$. So if a prediction is required sub in x in the above equation to obtain y in original units.


y = log x
log 2 = 0.3010
x