Transforming the Data

We are focusing on simple linear regression—however, not all bivariate relationships are linear. Some are curved…we will now look at how to straighten out two large families of curves.

The Exponential Transformation

Exponential functions are seen quite a bit out in the world—well, actually they only do a good job of modeling a part of what we see; in reality, nothing can follow an exactly exponential function. But I digress…

An exponential function is of the form $y = a^x$. The variable a is called the base, and is typically restricted to be a positive real number. Notice that x is the exponent. Below are a few examples of exponential function graphs.

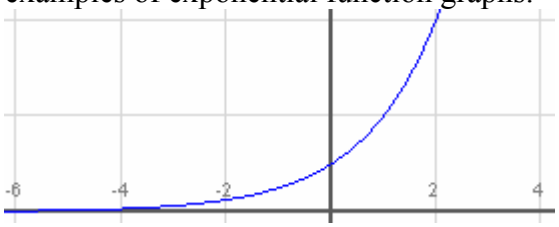

Figure 1 - Exponential Example 1

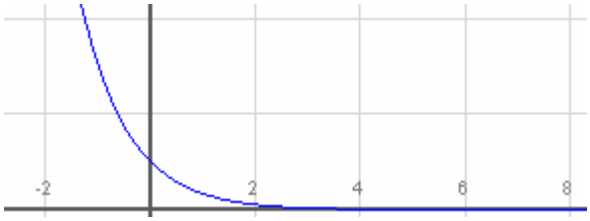

Figure 2 - Exponential Example 2

The essential feature of an exponential graph is that it rises quite sharply on one end, and flattens out completely on the other.

Every point on an exponential function is of the form (x, a^x). The simplest line would have points of the form (x, x). What could we do to change (x, a^x) into (x, x)? How do we get the x out of the exponent?

Logarithms! Specifically, take the logarithm of the y-coordinate (the response variable).

Now, when I say *logarithm*, I mean *natural logarithm*, since that's the most important kind. It is probably the case that you have been taught to use *common logarithms*—which will work, but they're just so…*common*…

If we take the logarithm of the response variable, then plot the new y-coordinates against the original x-coordinates, then the points will be of the form $(x, x \cdot \ln(a))$. This is akin to the graph with coordinates (x, ax), which is a line (with slope a).

If this new plot—using the transformed response variable—looks linear, then we can perform linear regression on it. The resulting regression equation will be $\ln(\hat{y}) = a + bx$, which takes in a value of the explanatory variable and spits out the log of the predicted response.

For example, perhaps you are told that a regression of $\sqrt{y}$ vs. x was performed, and the least squares line is $\sqrt{\hat{y}} = 5.218 - 0.197x$. What is the prediction when $x = 10$?

Plugging in 10 for x produces a value of 3.428, but this isn't the predicted response—it's the square root of the predicted response! You've got to square that to get the actual prediction of 10.550.

An Example

[1.] A classic example—the length of a year for a planet, based on its distance from the sun. Here are the data:

Table 1 - Orbital Data

Distance (millions of miles)	Year (# of Earth-years)
36	0.24
67	0.61
93	1
142	1.88
484	11.86
887	29.46
1784	84.07
2796	164.82
3666	247.68

First of all, let's see that the data have a curved relationship.

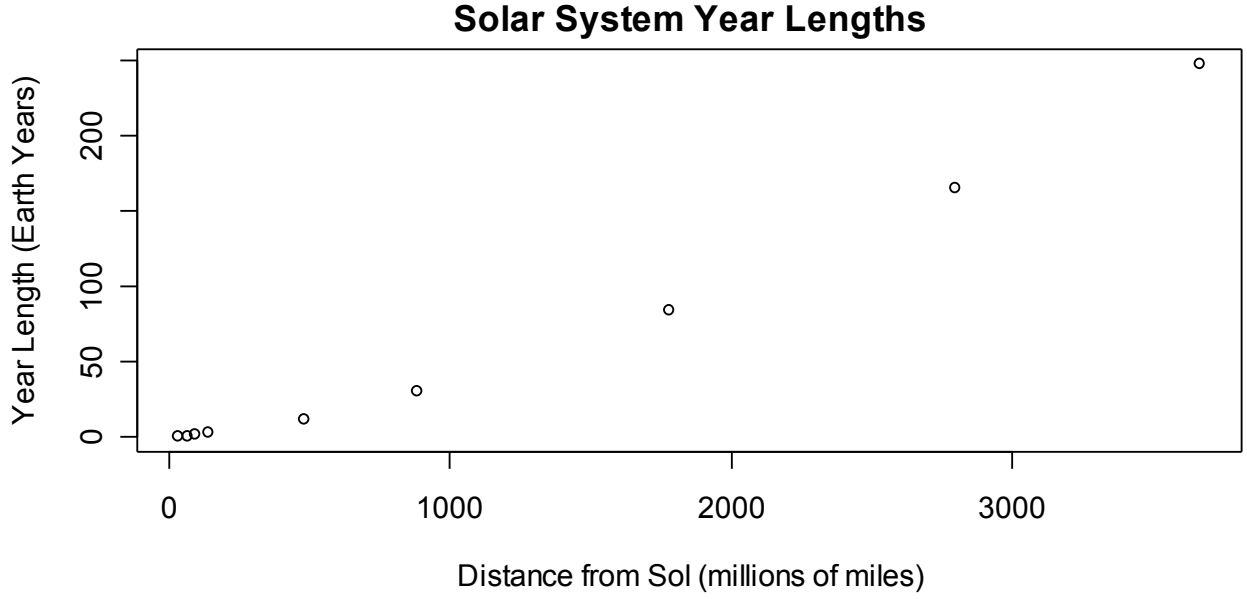

Figure 5 - Scatterplot of Orbital Data

Definitely curved. Don't believe me? Look at the residuals.

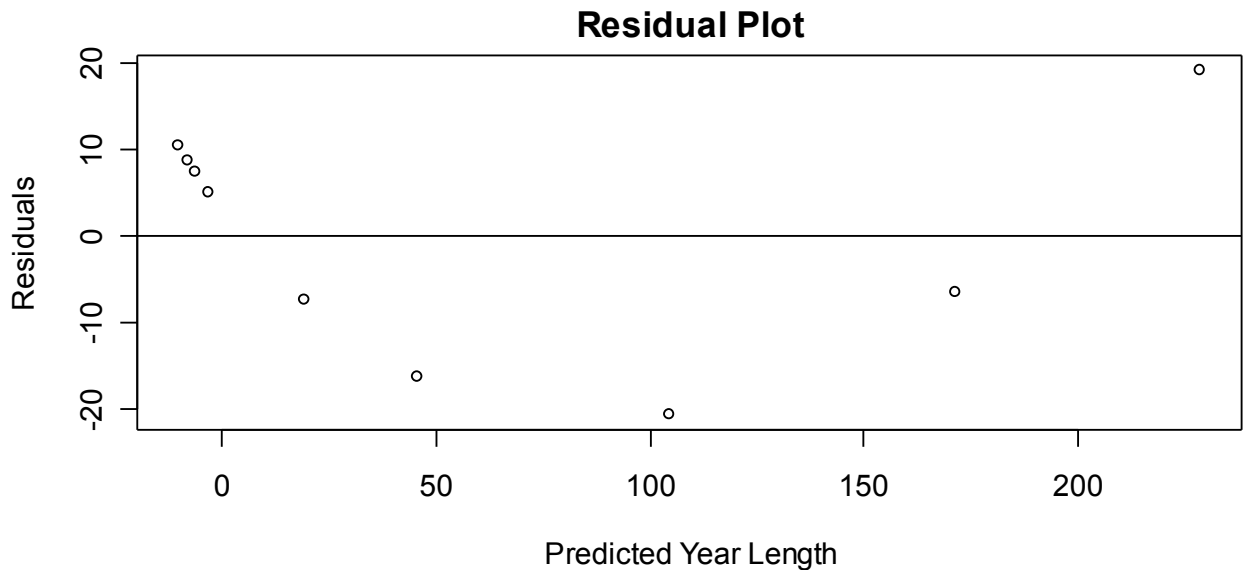

Figure 6 - Residual Plot of Orbital Data

Yeah—that's bad.

But what transformation should be applied?

This can't bottom off—that would force us to accept negative distances. Thus, the power transformation is indicated. Here is the new plot:

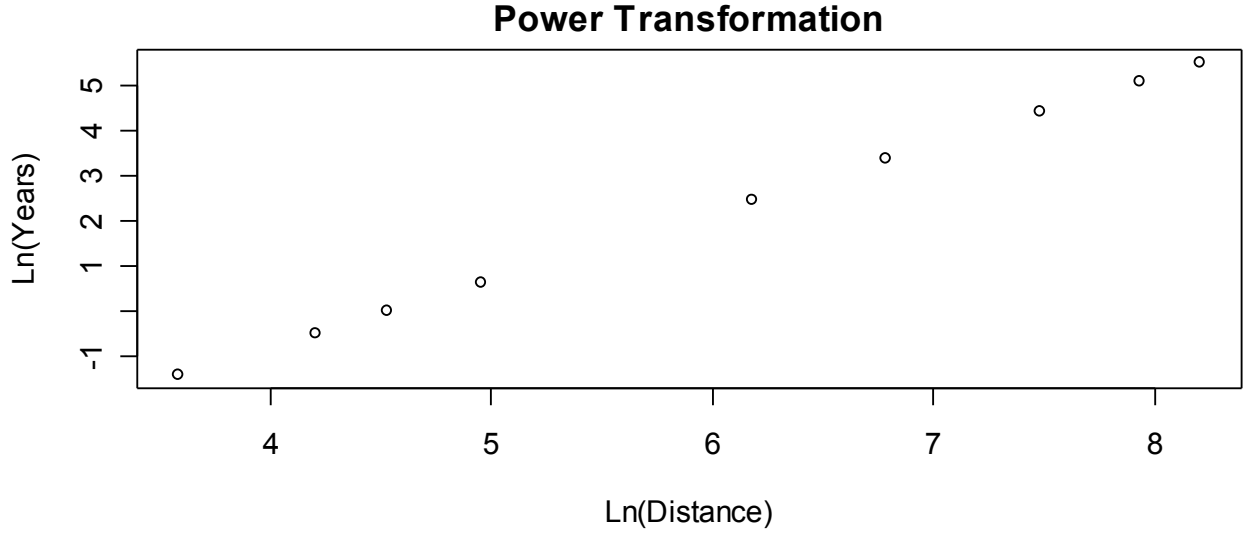

Figure 7 - Power Transformation Scatterplot

Looks pretty good. Linear, with strong positive association. There appears to be a gap between 5 and 6 (the asteroid belt!). The correlation is 1! 100% of the variation in the natural log of year length can be explained by the linear regression with the natural log of distance. Check that fit!

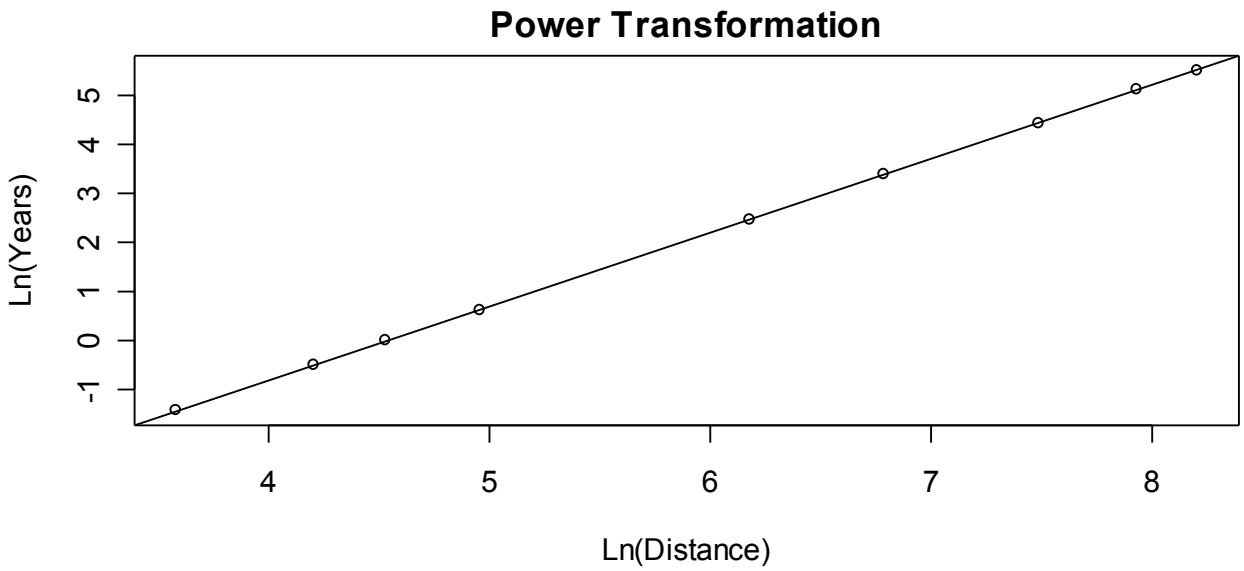

Figure 8 - Power Transformation Least Squares Line

OK, we'd better check the residuals. One residual plot, coming up.

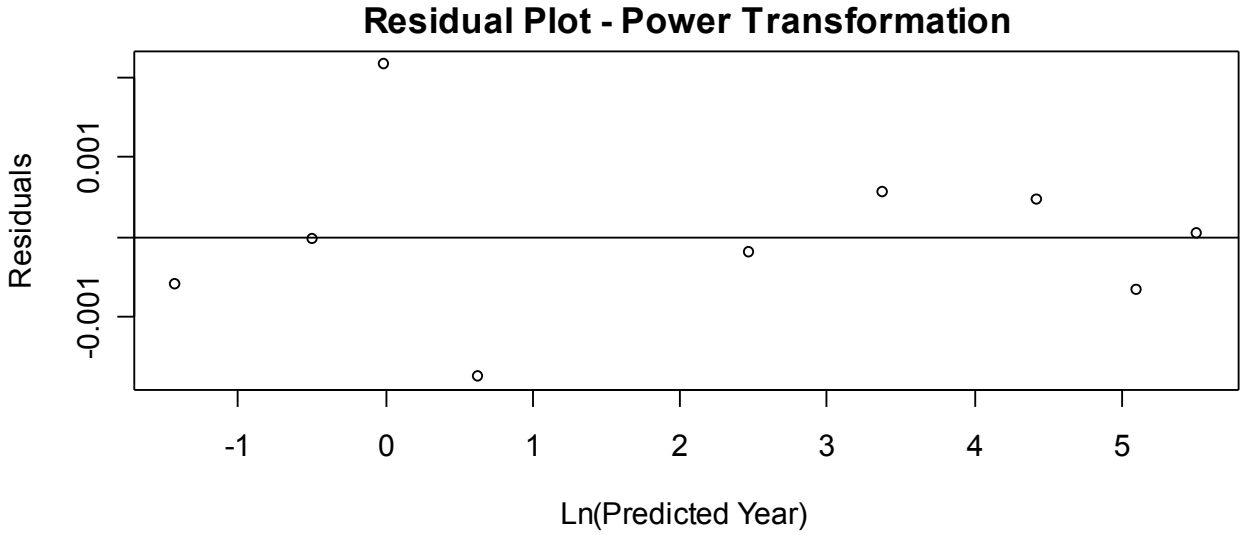

Figure 9 - Power Transformation Residual Plot

Nice and random. Now for a check of normality in the residuals.

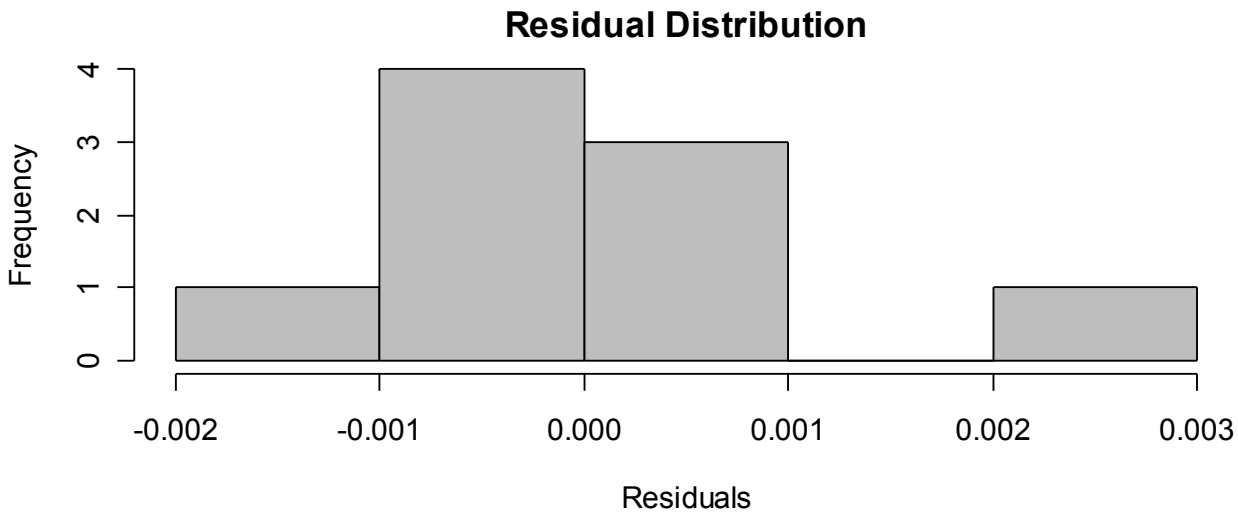

Figure 10 - Histogram of Residuals for Power Transformation

Hmmm…perhaps I'll look at the normal probability plot, too.

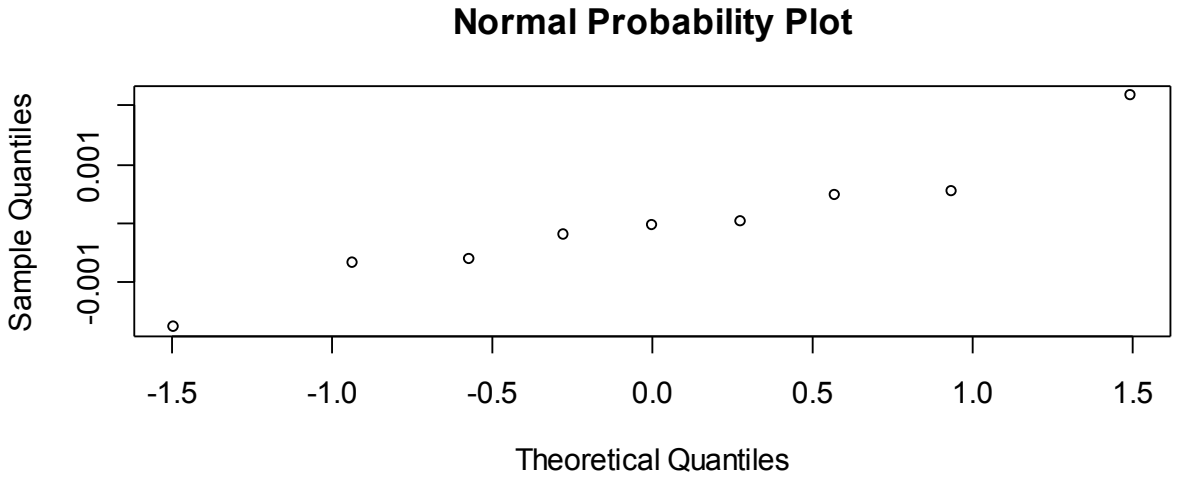

Figure 11 - Normal Probability Plot for Power Transformation

Not bad. Looks like we've got a model!

$\ln(\hat{y}) = -6.8046 + 1.5008 \cdot \ln(x)$, where $\hat{y}$ is the predicted year length, and x is the distance from Sol.

So—let's use this model to predict the year length of a planet that doesn't exist. The halfway point between Mars and Jupiter is around 313 million miles from Sol. What will this model predict for a year length if a planet occupied this position?

Letting $x = 313$, we get

$\ln(\hat{y}) = -6.8046 + 1.5008 \cdot \ln(313)$

$\ln(\hat{y}) = -6.8046 + 1.5008 \cdot 5.7462$

$\ln(\hat{y}) = -6.8046 + 8.6238$

$\ln(\hat{y}) = 1.8192$

$\hat{y} = e^{1.8192} = 6.167$

So our model predict a year length of 6.167 years.

(in fact, the asteroid Ceres orbits at a distance of about 418 million miles, and has a period of 4.6 years…oops? Well, Ceres isn't a *planet*).