Chapter 10

Re-expressing Data: Get it Strait!

Re-expressing data is important to discover a model that best fits in a certain situation. For example, when racing in a car it is common to measure the given speed in terms of miles per hour (ex. 90 miles per hour). However, if one were to measure the speed of a person running, miles per hour would not be an easy unit to understand; instead most people use the reciprocal of miles per hour by measuring it in saying how long of a mile was run (ex. a 5 minute mile). These examples prove that fact that every individual situation may or may not call for re-expression, and the only way to determine this is to understand the specific case that is being considered. Re-expression of data can accomplish several things. It may change the direction of the relationship, cause a straighter relationship, help satisfy the Strait Enough Condition, and reduce the amount of extrapolation the model predicts.

There are several goals re-expression attempts to accomplish. First, re-expression attempts to make the distribution of a variable more symmetric as seen on a histogram. Next, it tries to make the spread of several groups more alike as seen on side-by-side boxplots. Also, it attempts to make the form of a scatterplot more linear as can be seen on a residual plot. Finally, re-expression tries to make the scatterplot spread out evenly rather than following more of a fan shape.

How does one re-express find the correct way to re-express? The first step is to go through the Ladder of Powers as shown below:

Power	Name
2	The square of the data values, y^2.
1	The raw data-no change at all.
½	The square root of the data values, $\sqrt{y}$.
0	This place is held by logarithms.
-½	The (negative) reciprocal square root $-1/\sqrt{y}$.
-1	The (negative) reciprocal. $-1/y$.

Going through these steps will help find the most linear re-expression to use in the linear model.

What happens if these re-expressions do not provide a linear model in any case? It would then be wise to turn to the "Plan B" of re-expressing known as the Attack of the Logarithms which is shown below:

Model Name	**X-axis**	**Y-axis**
Exponential	X	log(Y)
Logarithmic	log(X)	Y
Power	log(X)	log(Y)

These should only be used if the Ladder of Powers does not produce sufficient re-expression, each of these models are useful in their own individual ways yet one should re-express with each one before making the decision as to which model to use in a linear test.