Data Transformations for Graphics and Modeling

For graphics we often want to have units we understand. Often these are the unit of measure that come with the data.

1. Winsorizing outliers

This is used in graphics to provide greater resolution in the body of the data. The procedure sets a threshold value and recoded all values beyond the threshold as the threshold value. Winsorized values can be show using a different symbol shape or color. The data units data the same.

2. Transformation to unitless variables

Some common data transformation produce unitless variables. Some it is helpful to think in terms variables that have a mean of 0 and a standard deviation of 1. The R scale() is designed subtract the mean and divide by the standard deviation for the columns of a matrix. Dividing by the standard deviation make the variables unitless. For positive variables would could divided by the geometric mean to them unitless. Other common transformation such as percent of total or percent change remove the original units of measure.

3. Power transformations of positive variables

We sometimes want to make variable with thick tails somewhat similar to a normal density. Cleveland (1993) discussed using a power transformation to reduce tail thickness of positive variable. Some other writers use xnew = (x**power - 1)/power because this that smoothly approaches the log transformation as the power approach 0. After making a power transformation the data units seem foreign too many people. There is some hope of clients understanding a square or a log transformation. R. Dennis Cook 1994 commented on historical guidance for selecting the log transformation of a positive variable. The guide to consider the log transformation when the ratio of the max(x) to min(x) is at least 10 preferably when 100 or larger. In term of choosing the base to use for the long Cleveland suggests using 2 as the base when value are not to much beyond 1024 since many people know the low power of 2. For really big numbers I like log because the integer part tells me the power of 10.

There are less frequently used transformation to bring in the both tails of a variable that has both positive and negative values. (Of course it is thinkable to thicken a very thin tail.)

Why bring in thick tails? Kernel density estimates and smoothes and other similar local procedures borrow information from their neighbors. The closest neighbors are the most important for such local methods. Closeness is defined by the gap between the data values. Tail thickness reducing transformations reduces that gaps between extreme points and make them more neighborly. For linear regression transforming the dependent variable can help make the residual distribution loser to a normal distribution. For a explanatory variable the transformation may reduce influence of the extreme value cases and let more of the other case have a say in the model.

4). Normal scores

One can simply replace the variable values by the corresponding quantiles from a standard normal distribution or one that has same mean and standard deviation as the variable. These are sometimes called normal scores.