Writing Exercise on Transformations

Here is my first draft.

Transformations are a common tool in applied statistics, and in regression analysis in particular. They are used to make the data better comply with the assumptions underlying the models we wish to use to analyze the data. There are three general reasons for transforming data:

- *To reduce skewness*, or equivalently enhance symmetry, in the distribution of individual variables. This is done because many basic statistical procedures, up through multiple regression, assume that the underlying variables follow a normal distribution, and the procedures still work reasonably well if the variables are at least symmetrically distributed with no severe outliers (extreme values).

- *To correct nonconstant spread* in one variable when plotted against another variable. This is done again because many basic statistical procedures (and those based on the assumption that the variables jointly follow the normal distribution, in particular) assume that the spread in one variable does not depend on the value of another variable in the problem.

- *To promote a linear relationship* between two or more variables. Again, this is because basic regression analysis assumes a linear relationship between the variables.

The most common transformations for continuous, unbounded variables are power transformations, $y = x^p$. For example, $p < 1$ corrects positive skew (long right tails), and $p > 1$ corrects negative skew (long left tails). The log and exponential functions can be viewed as the special cases: $\log(X)$ corresponds to $p = 0$ and e^x corresponds to $p = \infty$. Generally a transformation should be chosen both to improve the appropriateness of the assumptions of the statistical model, and to facilitate—or at least not hinder—an interpretation of the variables. Thus, powers such as 2, 1/2, and log and exponential functions are to be preferred over powers such as 0.324534 or 4.35254. For reporting to a nonstatistical audience, plots should probably be made in terms of the *untransformed* variables, so that any curvilinear relationships are apparent in the plots.

Writing Exercise on Transformations

Here is my second draft. *The first parag and the bullets were heavily edited, the second parag hardly at all.*

Transformations are a common tool in applied statistics, and in regression analysis in particular, because they can help the data to better comply with the assumptions underlying statistical models. Transformations are used: (a) to reduce skewness and enhance symmetry in the distribution of a single variable, since many basic statistical models assume that the variables are normally distributed, or at least symmetric; (b) to correct nonconstant spread in one variable as a function of another, since the multivariate normal distribution on which many basic procedures are based assumes this; and (c) to promote linearity in the relationship between variables, since basic regression models assume this. For reporting to a nonstatistical audience, plots should probably be made in terms of the *untransformed* variables. For examining the regression the transformed variables may be easier to work with.

The most common transformations for continuous, unbounded variables are power transformations, $y = x^p$. For example, $p < 1$ corrects positive skew (long right tails), and $p > 1$ corrects negative skew (long left tails). The log and exponential functions can be viewed as the special cases: $\log(X)$ corresponds to $p = 0$ and e^x corresponds to $p = \infty$. Generally a transformation should be chosen both to improve the appropriateness of the assumptions of the statistical model, and to facilitate—or at least not hinder—an interpretation of the variables. Powers such as 2, 1/2—or log and exponential functions—are better than 0.324534 or 4.35254, because simple integer and rational powers often relate to simple physical and social laws. For reporting to a nonstatistical audience, plots should probably be made in terms of the *untransformed* variables, so that any curvilinear relationships are apparent in the plots.