CSCU Cornell Statistical Consulting Unit

Use of Transformations in Statistical Models

Francoise Vermeylen

Transformations of the dependent or independent variables in statistical models can be useful for improving interpretability, model fit, or adherence to assumptions. This newsletter clarifies some uses for transformations, draws attention to an often-overlooked family of transformations that is now readily available, and highlights some pitfalls that can accompany the use of transformations.

A transformation of the dependent variable is often applied to meet the assumptions required for certain models that the residuals be normally distributed with constant variance. Such a transformation will affect the distribution of the residuals. For example, if a scatter plot shows an increase of the spread of the residuals with larger values of the dependent variable, a square-root transformation of the dependent variable may stabilize the variance. This might at the same time help normalize the residuals if need be.

Although regression models usually have no assumptions for the distributions of independent variables, an independent variable can be transformed to more suitable units or to make its distribution more symmetric. For example, one might transform an income variable to make its distribution more symmetric and alleviate the excessive influence of an outlying value.

Transformations can also be applied to the dependent and/or independent variables to convert a non-linear to a linear regression model. For example, a log transformation of the dependent variable would change an exponential growth model into a linear regression, thereby simplifying its estimation procedure.

Because transformations have multiple effects, their use requires care. For example, when transforming a variable to meet the assumption of normality and constant variance of residuals, to address skew, one may overlook the fact that this transformation might alter the relations that existed among the original variables. Scatter and residual plots should be used extensively to assess the suitability of a specific transformation.

If the most common transformations do not yield the intended results, one might turn to the Box-Cox procedure to find a more appropriate one. This procedure will help choose a transformation of the form Y to the L power, the family of power transformations. The log and square-root transformations are special cases of this family, with L equal to 0 and to 0.5 respectively. A Box-Cox procedure is now available in STATA and in the latest SAS version 8.2 in the Transreg procedure. These procedures will find the optimum L using the method of maximum likelihood, estimating L in addition to the usual regression parameters.

Transformations affect the interpretation of the obtained model coefficients. If desired, for some transformations, an inverse function can be applied to the obtained model coefficients that allow interpretation in the original units. The log transformation is popular because such a back-transformation can be applied. With a back-transformation of a model coefficient, the standard error will no longer be symmetric, so a better way to represent variability in the original units is obtained by back-transforming the end points of confidence intervals.

A good introduction to transformations, including the Box-Cox transformation can be found in :
Neter J., Kutner, M.H., Nachtsheim C.J., Wasserman W. (1996) Applied Linear Regression Models, Irwin, Third Ed., pp. 126-134.

Created October 2002. Last updated April 2022.