1

Transformations and Non-Linear Modeling

This document considers transformations in the analysis of RET data. Transformations have been used in several ways. First, they often are used as a tool to address assumption violations of statistical methods. The statistical workhorse for many RETs is some form of linear regression, so the assumption violations typically addressed are those dealing with normality, variance homogeneity of disturbance terms, and linearity. Also important is the assumption of independent disturbances and, potentially, interval level metrics. A second use of transformations is to formally model non-linear relationships between variables, such as between a continuous mediator and a continuous outcome. I address primarily the first use of transformations in this document. For a discussion of using transformations for non-linear modeling, see Chapter 15 (and, in particular, the section on log regression).

TRANSFORMATIONS TO ADDRESS ASSUMPTION VIOLATIONS

Transformations often are used to help make data conform to the assumptions of a statistical test. This strategy was important in earlier days of statistical analysis when the analytic tools we had were limited and assumption laden. In modern statistics, a large number of robust statistical procedures have become available that do not require the assumptions of the older methods. These newer approaches include such strategies as the invocation of sandwich estimators for standard errors, bootstrapping, and a host of newer robust regression methods described in Wilcox (2022). With the advent of these newer techniques, the need for using transformations to deal with assumption violations has diminished.

Common transformation strategies include log transformations, which are thought to reduce outliers and to deal with data that are right or positively skewed. Log transformations also are thought by some to address matters of non-normality and disturbance variance heterogeneity. Square root transformations often are invoked for positively skewed data where disturbance variance heterogeneity is proportional to the mean. Reciprocal transformations often are used when the disturbance variance is proportional to the fourth power of the mean. Box-Cox transformations are a family of power transformations that can stabilize heterogeneous disturbance variances and improve disturbance non-normality. Sometimes these strategies are effective at reducing assumption violations, sometimes not. Current wisdom is that it usually is better to use more modern analytic methods that do not

make assumptions that are problematic in the first place.

Sometimes a transformation will remedy one assumption violation but unwittingly create another violation (see Budescu & Appelbaum, 1981). As well, transformations have been found to be suboptimal when working with certain forms of heavy-tailed distributions (Rasmussen, 1989; Doksum &Wong, 1983). Log transformations to deal with outliers may or may not suffice. Wilcox (2003) describes how log transformations sometimes increase rather than decrease the number of outliers in data. There are a wide range of outlier resistant regression methods that handle outliers in more elegant ways than transformations and that allow one to maintain original variable metrics in the process (see Wilcox, 2008, 2010, 2017). I describe some of these regression methods in Chapter 6 and provide programs on my website to implement them.

Transformed variables often are difficult to interpret, which is another reason not to use them. For example, most of us have a good sense of income and what different values of income imply, but few of us have a good sense of the log of income and what different values of log income imply. How well off is someone who earns a log income of 10.23 and how much better off is that individual than someone who earns a log income of 9.33? This difference is not intuitive. When transformations make non-arbitrary metrics arbitrary, the transformations can be counterproductive interpretationally.

Doubts also can be raised about the meaningfulness of transformed scores if, for example, the transformations make constructs that are inherently skewed be normally distributed. The distribution of depression among the general population is inherently skewed with most people not suffering from depression. What exactly are we measuring when these depression scores are transformed to have a normal distribution? Do these transformed scores validly reflect depression? In some sense they do, but the non-linear function between the scores and the underlying latent construct of depression might make score interpretation difficult.

Yet another issue with a transformation strategy is that transformations in multivariate models can sometimes lead us to unwittingly work with models we never would have posited because of their conceptual inappropriateness. For example, suppose we want to work with the model $Y = X + Z$ but because of non-linearities, transform the variables using log transforms such that $\ln(Y) = \ln(X) + \ln(Z)$. It turns out that this latter model implies that X and Z in their original metrics combine multiplicatively to impact Y because the log of the equation $Y = XZ$ is $\ln(Y) = \ln(X) + \ln(Z)$. Is a multiplicative model what we intended when we model the log transformed predictors?

Transformations of a variable affects not just its association with one predictor in a regression equation; it also potentially changes the relationship between the transformed variable and *all* the other variables in the regression analysis. This can result in model

misspecification and biased coefficients and significance tests. One must be careful when using transformations in multivariate systems because the transformation can affect the entire system, not just one pair of variables.

Some researchers argue that transformations are unnecessary for linear regression for handling non-normality or disturbance variance heterogeneity because tests of coefficients for regression models based on OLS are robust to such violations. Broad assertions about the robustness of OLS regression are questionable because the fact is that the matter of robustness is nuanced. For example, Wilcox (2022) argues that with certain types of heteroscedasticity, Type I errors in regression models can inflate substantially beyond desired alpha levels. Statistical power also can be adversely affected by violations of normality and heteroscedasticity, especially for heavy tailed distributions (Wilcox, 1998; Wilcox & Rousselet, 2023). Again, when possible, one usually is advised to use analytic methods that do not make these assumptions rather than resort to ad hoc transformations.

Transformations also can obscure effect size estimates and their meaningfulness. Transformations can sometimes help address tests of null hypotheses of zero effects, but they can yield different characterizations of effect size based on transformed versus untransformed data (even after back-transformation).

For some transformations and some analytic situations, it is reasonable to back-transform predicted scores and regression coefficients so that inferences and statements can be made in the original metric of the transformed variables. However, in some scenarios, back transformations can be misleading (Miller, 1984; Dambolena, Eriksen & Kopsco, 2009). Also, what is minimized in the loss function when deriving parameter estimates in the transformed model is not necessarily what is minimized in a non-linear version of that model vis-à-vis traditional non-linear modeling. To illustrate this point, consider an exponential model that appears as follows:

$$Y = (a)(e^{bX}) \quad [1]$$

It turns out, I can "linearize" this model by transforming Y to be the natural log of Y. Taking the log of both sides of Equation 1 produces the model:

$$\ln(Y) = \ln(a) + bX \quad [2]$$

In principle, if I regress ln(Y) onto X using standard linear regression methods, I will obtain estimates of ln(a) and of b (the intercept and slope in the log transformed equation). The exponent of the intercept from the linear model is an estimate of a in Equation 1 and b in Equation 2 is an estimate of b in Equation 1. Note that when I use classic non-linear

regression modeling, I introduce a disturbance term into Equation 1 to reflect random disturbances and I then use the least squares loss function to derive estimates of the adjustable constants in Equation 1. , with the model taking the form

$$Y = (a)(e^{bX}) + d \quad [3]$$

where *d* is the disturbance term. I seek to minimize the squared discrepancies between Y and $\hat{Y}$ in this model where $\hat{Y}$ is $(a)(e^{bX})$. By contrast, when I apply traditional OLS regression per Equation 2, I minimize instead the sum of the squared differences between ln(Y) and $\hat{Y}$ where $\hat{Y}$ is defined as (ln(a) + bX), which, it turns out, yields a different error structure from the traditional non-linear model (see Seber & Wild, 2003). The parameter estimates, thus, can be different as can the results of significance tests in the log transformed linear model compared to the more traditional non-linear approach.

CONCLUDING COMMENTS

For all of the above reasons, I personally think it is better to avoid transformations whenever possible rather than embrace them. To be sure, there are modeling scenarios where transformations can be useful (see, for example, Chapters 12 to 15), but as a go-to strategy to deal with skewed data or assumption violations of variance heterogeneity and non-normality, there generally are better ways of addressing such matters. Using transformations in such scenarios is outdated.

REFERENCES

Budescu, D. V., & Appelbaum, M. I. (1981). Variance stabilizing transformations and the power of the F test. Journal of Educational Statistics, 6, 483-497.

Doksum, K. A., & Wong, C.-W. (1983). Statistical tests based on transformed data. *Journal of the American Statistical Association*, 78, 411–417.

Miller, D. M. (1984). Reducing transformation bias in curve fitting. *The American Statistician*, 38, 124–126.

Rasmussen, J. L. (1989). Data transformation, Type I error rate and power. *British Journal of Mathematical and Statistical Psychology*, 42, 203–211.

Seber, G. & Wild, C. (2003). *Nonlinear regression*. Wiley.

Wilcox, R. (1998). How many discoveries have been lost by ignoring modern statistical methods. American Psychologist, 53, 300-314.

Wilcox, R. R. (2003). *Applying contemporary statistical techniques.* Academic Press.

Wilcox, R. R. (2008). Robust multivariate regression when there is heteroscedasticity. *Communications in Statistics - Simulation and Computation*, 38, 1–13.

Wilcox, R. R. (2010). *Fundamentals of modern statistical methods*. Springer

Wilcox, R. R. (2022). *Introduction to robust estimation and hypothesis testing* (5th ed). Academic Press.

Wilcox, R. R., & Rousselet, G. A. (2023). An updated guide to robust statistical methods in neuroscience. *Current Protocols*, 3, e719.