Data Analysis Exercises for Chapter 4: *Applied Regression Analysis, Generalized Linear Models, and Related Methods, Third Edition* (Sage, 2016)

John Fox

Last modified: 2015-01-29

Exercise D4.1 Return to the distributions that you examined in Exercise D3.1. If a distribution is not symmetric, attempt to make it symmetric by an appropriate transformation on the ladder of powers and roots. If a variable is a proportion or a percent, consider using a logit transformation or another transformation that is appropriate for proportions.

Exercise D4.2 Return to the scatterplots that you constructed in Exercise D3.3 to examine relationships between a quantitative response variable and quantitative explanatory variables. If the relationship in a scatterplot is nonlinear, attempt to make the relationship linear by transforming one or both variables, or explain why a transformation to linearity is not an appropriate strategy.

Exercise D4.3 Return to the parallel boxplots that you constructed in Exercise D3.3 to examine relationships between a quantitative response variable and categorical explanatory variables. If the boxplots manifest non-constant spread, try to equalize the spreads by a suitable transformation of the response, or explain why transforming the response is not an appropriate strategy.

Exercise D4.4 *Return to the scatterplot matrix that you constucted in Exercise D3.5. Using the method of maximum likelihood, estimate the Box-Cox transformations that make the joint distribution of the variables as close to mulivariate-normal as possible. How successful is this endeavor? Was it sensible to try to normalize the joint distribution of the variables using transformations in the Box-Cox family?

Exercise D4.5 Examine the distributions of prestige, income, and education in Duncan's occupational prestige data (in `Duncan.txt`). All of these variables are percentages (of raters classifying an occupation as good or better in prestige; of occupational incumbents earning $3500 or more; of high-school graduates). Transform the data using logits (and, possibly, other folded-power transformations), and re-examine the distributions. Does the transformation make the distributions of these variables more nearly normal?