Remedies for Assumption Violations and Multicollinearity

Outliers

- If the outlier is due to a data entry error, just correct the value. This is a good reason why raw data should be retained for many years after it is collected as some professional associations recommend. With direct computer entry during interviews, entry errors cannot always be determined for certain.

- If the outlier is due to an invalid case because the protocol was not followed or the inclusion criteria was incorrectly applied for that case, the researcher may be able to just eliminate that case. I recommend at least reporting the number of cases excluded and the reason. Consider analyzing the data and/or reporting analyses with and without the deleted case(s).

- Transformation of the data might be considered (see other handout). Square root transformations, for instance, bring outlying cases closer to the others. Transformations, however, can make results difficult to interpret sometimes.

- Analyze the data with and without the outlier(s). If the implications of the results do not differ, one could simply state that the analyses were conducted with and without the outlier(s) and that the results do not differ substantially. If they do differ, both results could be presented and a discussion about the potential causes of the outlier(s) could be discussed (e.g., different subgroup of cases, potential moderator effect).

- An alternative estimation method could be used, such as *least absolute residuals*, *weighted least squares*, *bootstrapping*, or *jackknifing*. See discussion of these below.

Serial Dependency

- Use time as a covariate (e.g., by transforming the date into the number of days since the start date).

- Transform the X and Y variables using the following formulas: $X^* = X - r_{lag1}X_{t-1}$ and $Y^* = Y - r_{lag1}Y_{t-1}$. In the formula for X*, X_{t-1} is the value of X at the previous timepoint. Similarly, in the formula for Y*, Y_{t-1} is the value of Y at the previous time point. r_{lag1} is the estimation of the correlation between Y and Y_{t-1}, called the *autocorrelation*. This approach may require some difficult reconfiguring of the data and a particular study design.

- Use a different analysis technique such as time series analysis, hierarchical linear modeling (HLM), or structural equation modeling. The ability to use these approaches may depend on particular features of the study design.

Clustering

- Use hierarchical linear modeling (see Raudenbush & Bryk, 2002).

- Generalized Estimating Equations (GEE) may offer an analysis alternative in some circumstances.

- Use complex sampling design adjustments (e.g., see Lee & Forthofer, 2006).

Heteroscedasticity (Nonconstant Variance)

- Some heteroscedasticity problems may be due to the presence of an outlier or group of outliers. In this case, one could follow the remedies presented above.

- Alternative analysis techniques, such as *least absolute residuals*, *weighted least squares*, *bootstrapping*, or *jackknifing*, are also designed to be used for heteroscedasticity problems. (see below)

- Use robust standard errors, also referred to as Huber-White standard errors, or "sandwich estimator." Regression estimates are the same as OLS, and robust standard errors will be equal to OLS standard errors under homoscedasticity. Estimates may be best with large samples.

- Some data transformations of X or Y may be useful.

Multicollinearity

- Check for errors or problematic computations of predictor variables.

- Eliminate one of the redundant variables.

- Average the redundant variables and reconceptualize the meaning of the predictor.

Alternative Regression Estimation

- Historically, alternative estimation procedures have not faired very well in Monte Carlo studies, although newer methods using robust standard errors or bootstrapping do a much better job.
- One general type of approach is referred to as "robust regression", which adjusts standard errors (e.g., White, 1980; sometimes referred to as "sandwich" estimator, or, Eicker, Huber-White), is gaining increasing popularity in some circles.[1] Regression and ANOVA are fairly robust to normality assumption violations, but in more serious cases, this approach may be helpful. They also appear to be useful for heteroscedasticity problems as well for sufficiently large sample sizes (Hayes & Cai, 2007).
- *Least absolute residuals* or *least absolute deviation.* Minimizes |e| instead of e^2. This reduces impact of large residuals.
- *Least trimmed squares*. Standard errors are recomputed by eliminating the most extreme positive or negative residuals. This method may have problems when there is a cluster of outliers but can work well for individual outliers.
- *Weighted least squares (WLS).* WLS does not require the assumption that there will be equal variances. With WLS, cases are reweighted based on the size of X relative to the variance of X, so that more extreme cases are given less weight in the analysis. Generally, there have been a number of criticisms of WLS estimates. There are several newer varieties of weighted least squares (e.g., M-estimates, bounded influence estimates) which offer some improvements over regular WLS. A version called iterated WLS uses many iterations to arrive at the best weighting scheme and shows promise for improved estimates. I recommend reporting both OLS and the WLS estimates if a weighted least squares approach is chosen.
- *Bootstrapping*. Many random samples are drawn from the full sample and estimates of the standard errors are computed for each. An average of all the bootstrap samples is then computed. This reduces the impact of any particular case and can help with heterscedasticity. Boostrapping may be a good alternative estimator, but is still relatively new. This approach is receiving a lot of attention as a data analytic remedy in this realm and others (e.g., missing data analysis).
- *Jackknifing*. Similar to bootstrapping, except that the standard error is recomputed many times, each time eliminating one case. An average of the standard errors is then used.

[1] See GENLIN procedure in SPSS and rlm function from the MASS package.