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. For direct computer entry or online data collection, 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 overhead on transformations). 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 reconfiguring of the data and a particular study design.
- Use a different analysis technique such as time series analysis, hierarchical linear modeling (HLM; also known as multilevel modeling), or structural equation modeling. The ability to use these approaches may depend on particular features of the study design.

Clustering

- Use hierarchical linear (multilevel) 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 (or Huber-White-Eiker) 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, the older alternative estimation procedures have not always faired very well in simulation studies, but newer methods, such as those using robust standard errors or bootstrapping, have shown much more

promise. Unfortunately, there is an overwhelming number of particular approaches for each one of the alternative methods described below, and there is a dizzying array of specific situations (distribution shapes, number of outliers, extremity of outliers, sample sizes, number of predictors, overall fit of the model, and effect sizes) that can have an important effect on whether these methods perform better and in what ways compared with OLS regression. There are still quite a few authors who contend that OLS works well in most general circumstances (e.g., Cohen, Cohen, West, & Aiken, 2003; Lewis-Beck & Lewis-Beck, 2015; Lumley, 2002). Others have been stronger proponents of alternative or robust regression methods (e.g., Fox, 2015; Fox & Weisberg, 2011).

- One general type of approach, often simply 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 superior. The robust estimation approach appears to be useful for heteroscedasticity problems as well, provided the sample size is sufficiently large (Hayes & Cai, 2007).
- *Least absolute residuals, least absolute deviation,* or *least absolute values.* Minimizes $|e|$ instead of e^2. This reduces impact of large residuals. There is a wide variety of specific methods and tests used (Dielman, 2005), and the methods may be preferable when there are distribution issues, outliers, or heteroscedasticity depending on the method and the circumstances.
- *Least trimmed squares (*or more generally*, bounded estimates)*. Standard errors are recomputed by eliminating the most extreme positive or negative residuals. There are a variety of related methods (Nguyen & Welsch, 2010), some of which 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 (often based on leverage), so that more extreme cases are given less weight in the analysis. In its basic form, WLS can be problematic unless the choice of weights is accurate. There are several newer varieties in combination with other techniques (e.g., M-estimates, bounded influence estimates, bisquare, minimax) that offer some important improvements over regular WLS (Strutz, 2016). For any of these WLS-related methods, I recommend reporting both OLS and the WLS estimates, as it is difficult to know which one is the optimal approach for any particular set of circumstances.
- *Bootstrapping*. Many random samples are drawn from the full sample with replacement. Expected values and the variance of this mini sampling distribution can be used for addressing parameter bias or improving confidence interval (or standard error) estimation. Bootstrapping can reduce the impact of particular outlier cases and improve estimates when heteroscedasticity is present. It is a promising alternative method receiving a lot of attention (Chernick, González-Manteiga, Crujeiras, & Barrios, 2014; Davison & Hinkley, 2003) as a data analytic remedy in this realm and others (e.g., non-normality correction, mediation analysis, and missing data analysis).
- *Jackknifing*. Similar to bootstrapping but less commonly employed. Resampling uses the sample size minus one case, using the multiple estimates to correct bias and estimate standard errors.

References

Chernick, M. R., González-Manteiga, W., Crujeiras, R. M., & Barrios, E. B. (2011). Bootstrap methods. In M. Lovric (Ed.), *International encyclopedia of statistical science* (pp. 169-174). Heidelberg, Germany: Springer.
Davison, A. C., Hinkley, D. V., & Young, G. A. (2003). Recent developments in bootstrap methodology. *Statistical Science, 18*, 141-157.
Dielman, T. E. (2005). Least absolute value regression: recent contributions. *Journal of Statistical Computation and Simulation*, *75*(4), 263-286.
Fox, J. (2015). *Applied regression analysis and generalized linear models.* Thousand Oaks, CA: Sage Publications.
Fox, J., & Weisberg, S. (2011). *An R companion to applied regression. T*housand Oaks, CA: Sage Publications.
Hayes, A. F., & Cai, L. (2007). Using heteroskedasticity-consistent standard error estimators in OLS regression: An introduction and software implementation. *Behavior research methods, 39*(4), 709-722.
Huynh, H. (1982). A comparison of four approaches to robust regression. *Psychological Bulletin, 92*(2), 505.
Lee, E. S., & Forthofer, R. N. (2006). *Analyzing complex survey data, second edition* (Vol. 71). Thousand Oaks, CA: Sage Publications.
Nguyen, T. D., & Welsch, R. (2010). Outlier detection and least trimmed squares approximation using semi-definite programming. *Computational Statistics & Data Analysis, 54*(12), 3212-3226.
Lewis-Beck, C., & Lewis-Beck, M. (2015). *Applied regression: An introduction (Vol. 22)*. Thousand Oaks, CA: Sage publications.
Raudenbush, S. W., & Bryk, A. S. (2002). H*ierarchical linear models: Applications and data analysis methods* (Vol. 1). Thousand Oaks, CA: Sage.
Strutz, T. (2016). *Data Fitting and Uncertainty (A practical introduction to weighted least squares and beyond)*. Weisbaden, Germany: Vieweg + Teubner Fachmedien Wiesbaden GmbH Springer.
White, H. (1980). "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica, 48*, 817–38.
Wilcox, R. R. (1996). A review of some recent developments in robust regression. *British Journal of Mathematical and Statistical Psychology, 49*(2), 253-274.
Yu, C., & Yao, W. (2017). Robust linear regression: A review and comparison. *Communications in Statistics-Simulation and Computation, 46*(8), 6261-6282.

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