Dealing with Missing Data

(Tabachnick & Fidell, 2013)

The first step in dealing with missing data is to observe their pattern to try to determine whether data are randomly missing. Deletion of cases is a reasonable choice if the pattern appears random and if only a very few cases have missing data and those cases are missing data on different variables. However, if there is evidence of nonrandomness in the pattern of missing data, methods that pre- serve all cases for further analysis are preferred.

Deletion of a variable with a lot of missing data is also acceptable as long as that variable is not critical to the analysis. Or, if the variable is important, use a dummy variable that codes the fact that the scores are missing coupled with mean substitution to preserve the variable and make it possible to analyze all cases and variables.

It is best to avoid mean substitution unless the proportion of missing values is *very* small and there are no other options available to you. Using prior knowledge requires a great deal of confidence on the part of the researcher about the research area and expected results. Regression methods may be implemented (with some difficulty) without specialized software but are less desirable than EM methods.

EM methods sometimes offer the simplest and most reasonable approach to imputation of missing data, as long as your preliminary analysis provides evidence that scores are missing randomly (MCAR or MAR). Use of an EM covariance matrix, if the technique permits it as input, provides a less biased analysis a data set with imputed values. However, unless the EM program provides appropriate standard errors, the strategy should be limited to data sets in which there is not a great deal of missing data, and inferential results (e.g., *p* values) are interpreted with caution. EM is especially appropriate for techniques that do not rely on inferential statistics, such as exploratory factor analysis. Better yet is to incorporate EM methods into multiple imputation.

Multiple imputation is currently considered the most respectable method of dealing with missing data. It has the advantage of not requiring MCAR (and perhaps not even MAR) and can be used for any form of GLM analysis, such as regression, ANOVA, and logistic regression. The problem is that it is more difficult to implement and does not provide the full richness of output that is typical with other methods.

Using a missing data correlation matrix is tempting if your software offers it as an option for your analysis because it requires no extra steps. It makes most sense to use when missing data are scattered over variables, and there are no variables with a lot of missing values. The vagaries of missing data correlation matrices should be minimized as long as the data set is large and missing values are few.

Repeating analyses with and without missing data is highly recommended whenever any imputation method or a missing data correlation matrix is used and the proportion of missing values is high—especially if the data set is small.

Reference:

Tabachnick, B. G., & Fidell, L. S. (2013). *Using Multivariate Statistics*. Pearson Education. https://books.google.com/books?id=ucj1ygAACAAJ