Solutions for Missing Data in Structural Equation Modeling

Rufus Lynn Carter
Marymount University

Abstract

Many times in both educational and social science research it is impossible to collect data that is complete. When administering a survey, for example, people may answer some questions and not others. This missing data causes a problem for researchers using structural equation modeling (SEM) techniques for data analyses. Because SEM and multivariate methods require complete data, several methods have been proposed for dealing with these missing data. What follows is a review of several methods currently used, a description of strengths and weaknesses of each method, and a proposal for future research.

Methods for dealing with missing data

Listwise deletion

Listwise deletion is an ad hoc method of dealing with missing data in that it deals with the missing data before any substantive analyses are done. It is considered the easiest and simplest method of dealing with missing data (Brown, 1983). It involves removing incomplete cases (record with missing data on any variable) from the dataset. This means the researcher removes all the records that have missing data on any variable. Depending on the sample size and number of variables this can result in a great reduction in the sample size available for data analysis. Listwise deletion assumes that the data are missing completely at random (MCAR). Data are missing completely at random when the probability of obtaining a particular pattern of missing data is not dependant on the values that are missing and when the probability of obtaining the missing data pattern in the sample is not dependant on the observed data (Rubin, 1976). An advantage in using listwise deletion is that all analyses are calculated with the same set of cases.

Pairwise Deletion

Another ad hoc method of dealing with missing data, pairwise deletion (PD), uses all available data. This means for each pair of variables PD calculates the covariance estimates from all cases with complete observations on both variables (Wothke, 1998). Pairwise deletion assumes that the data are missing completely at random (MCAR). Cases are removed when they have missing data on the variables involved in that particular computation (Kline, 1998). This can be problematic in that each element of the covariance matrix could be based on different groups of subjects. For example, if 300 subjects had complete scores for variables X1 and X2 then the effective sample size for the covariance between X1 and X2 is 300. Likewise, if 200

subjects had complete scores on X1 and X3 then the sample size for this covariance would be only 200. Kline (1998) points out that it would be impossible to derive some of these covariances if they were calculated using data from all subjects as in listwise deletion.

Imputation

The method of imputation involves placing estimated scores into the data set in the location of the missing data. Kline (1998) discusses three basic types of imputation.
In each of these three types of imputations, the data are assumed to be MCAR. Mean imputation involves substituting missing cases with the overall sample average for each particular variable with missing data. While simple to execute this method does not take into consideration subjects patterns of scores across all the other variables. Regression imputation takes this into consideration by predicting a score for each subject by using multiple regression based on their non missing scores for other variables. For this method to work, Kline (1998) states that the variable with missing data must co-vary at least moderately with the other variables.

Pattern matching is the third from of imputation Kline (1998) describes. In this method the missing score is replaced with a score from another subject who has a similar profile of scores across the other variables. This method is not widely available on software packages but is available via PRELIS2 (Joreskog & Sorbom, 1996b), which performs pattern matching and can be used with LISREL. Kline (1998) notes these methods seem to work best with the proportion of missing data is low and scattered across different variables.

Another imputation method is that of multiple imputations with the Expectation-Maximization (EM) algorithm. Dempster, Laird, and Rubin (1977) presented an algorithm for computing maximum likelihood estimates from missing data sets. Each iteration of their algorithm consists of an expectation step followed by a maximization step. They assume a family of sampling densities $f(x|\varphi)$ depending on parameters φ and they then derive their corresponding family of sampling densities $g(y|\varphi)$. The EM algorithm attempts to find a value of φ which maximizes $g(y|\varphi)$ given an observed y, but it does this by making use of the related family $f(x|\varphi)$. Schafer and Olsen (1998) state that with the development of the EM algorithm, statisticians have stopped viewing missing data as a "nuisance" and have reevaluated it as a source of variability to be averaged over. Schafer and Olsen (1998) describe a technique developed by Rubin (1987) where each value is replaced with a set of $m > 1$ plausible values which allows the variances reported above to be averaged by simulation. After performing multiple imputations, each of these *m* data sets can be analyzed by SEM techniques intended for complete data. Then through a series of complex rules the estimates and standard errors are combined to provide overall estimates and standard errors that reflect missing data uncertainty. These rules properly applied are thought to provide unbiased estimates.

Schafer and Olsen (1998) describe their own iterative process, data augmentation (DA), which alternately fills in the missing data and makes inferences about the unknown parameters. The process is similar to the EM algorithm as DA fills in the missing data either randomly or

else based on conjecture. DA performs a random imputation of missing data under assumed values of the parameters and then draws new parameters from a Bayesian posterior distribution based on the observed and imputed data. Schafer and Olsen explain the Bayesian distribution as requiring the researcher to specify a prior distribution for the parameters of the imputated model. Schafer (1997) developed a computer program NORM using the multivariate normal distribution to generate imputations for the missing values.

Schafer and Olsen (1998) note that multiple imputation methods resemble other methods of ad hoc case deletion because it addresses the missing-data issue at the beginning, before substantive analyses are run. They argue that unlike the other ad hoc methods, multiple imputations do not have to be MCAR but instead need only meet the less rigorous assumption that the missing data are missing at random (MAR). Data are missing at random when probability of obtaining a particular pattern of missing data is not dependant on the values that are missing (Rubin, 1987). Schafer and Olsen (1998), also state that while multiple imputation techniques are statistically defensible and incorporate missing-data into all summary statistics, They do suggest that the direct maximum likelihood methods may be more efficient than multiple imputations because they do not rely on simulation.

SEM Methods

One option available by SEM to deal with the problem of missing data is illustrated by Allison (1987). He proposes a maximum likelihood estimation for incomplete data. His model assumes multivariate normality, which as he states implies that the means, variances, and covariances are the sufficient statistics. However he also states that violations of multivariate normality will not seriously compromise the estimates. Allison (1987) discusses a confirmatory factor model where the goal is to estimate the correlation between father's occupational status (FAOC) and father's educational attainment (FAED) for black men in the U.S. He reports previous studies had estimated the correlation to be 0.433. He split a sample of 2,020 taken from Bielby et al. (1977b) into two groups, 348 with complete data and 1,672 with incomplete data. The small complete sample had two indicators of FAOC (y1 and y2) and two indicators of FAED (y3 and y4). The large sub-sample had only y1 and y3. Allison states that this design virtually guarantees that the missing data are missing completely at random. Sample variances and covariances for the complete-data sub-sample were obtained from the correlation matrix and standard deviations in the original study. By calculating sums of squares and cross-products from the reported correlations and standard deviations of sample with the missing data, comparisons can then be made between the re-measurement sample and the full sample. These values are then used to recreate the covariance matrix for the sample with missing data (Allison, 1987). He goes on to state that while his method using LISREL produces non-biased estimates; it is exceedingly complex with the addition of more variables. The relationship of number of variables to number of possible missing data patterns is $2^k - 1$. In these cases Allison (1987) suggests using the previously mentioned listwise and pairwise ad hoc practices to eliminate

minor missing data patterns. His LISREL runs require the sample means and requires that each latent variable in each sub-sample have at least one indicator with a fixed, nonzero λ coefficient. The nonzero λ coefficients for y1 and y3 are fixed at 1.0, which define the metrics for the latent variables (Allison 1987). For the sub-sample with no observations on y2 and y4 he set λ_{21}, λ_{23}, λ_{42}, and λ_{43} equal to 0.0 and constrained variances ε_2 and ε_4 equal to 1.0. All the free parameters were constrained to be equal across sub samples (Allison, p. #,).

Another method of using maximum likelihood to estimate missing data is the Full-Information Maximum Likelihood (FIML) method. "The FIML method uses all of the information of the observed data, including mean and variance for the missing portions of a variable, given the observed portion(s) of other variables" (Wothke, 1998). Muthén, Kaplan, and Hollis (1987) present how the method applies to structural equation modeling. They state that their method using LISREL allows for the latent variable model to include missingness. Their paper examines maximum likelihood estimation of the θ parameters. Wothke(1998) states that FIML assumes multivariate normality, and maximizes the likelihood of the model with the observed data. He also states that two structural equation modeling programs, AMOS (Arbuckle, 1995) and Mx (Neale, 1994), implement this FIML method for dealing with missing data. He critiques other methods for estimation using FIML and states that those approaches are only practical when the data have just a few distinct patterns of missing data. In addition, he and states that using AMOS (Arbuckle, 1995) and Mx do not require the same level of technical expertise as do the methods of presented by Dempster et al. (1977) and Muthén et al. (1987) do. Wothke(1998) suggests that both AMOS and Mx maximize the case-wise likelihood of the observed data, computed by minimizing the function. He further states that both AMOS and Mx are not limited by the number of missing-data patterns, and do not require complex steps to accommodate missing data.

Comparisons of Methods in the Literature

Several of the techniques described earlier have been compared to determine which yields the least biased estimates in SEM. Wothke (1998) examined listwise, pairwise, mean imputation and maximum likelihood methods for growth curve modeling for examples where the data were MCAR and MAR. For the MCAR data estimates of the model parameters were unbiased for FIML, LD and MD, while mean imputation showed no bias in means but exhibited strongly biased variance and covariance estimates. For the MAR data FMIL produced unbiased estimates while PD estimates exhibited a small negative bias. Listwise deletion and mean imputation methods resulted in sampling distributions that did not include the parameter value. Similar results are reported in the literature by Muthén et al. (1987) and Arbuckle (1996). In these and other studies the comparison seems to be that of FIML methods with listwise and pairwise deletion. The results of the comparisons of these methods in the literature indicate that when the data are MCAR there is little difference in the estimation bias for listwise deletion, pairwise deletion and maximum likelihood. Some Other comparisons, were notably absent from

the literature, and are the subject of the research proposal discussed below.

Future Research

In the literature, little attention has been paid to the use of pattern-based imputation in the literature. For MCAR data it would appear to be a viable alternative to listwise and pairwise deletion and perhaps to both multiple imputation methods and maximum likelihood methods. Further investigation into this area is needed.

One suggestion is to generate population values from a complete data set having no missing values. A random number generator like that found in SAS (version 9) software can provide random missing data points for an adequate number of data sets. A single model can be fit to each random sample taken from the original population sample as described above. Model fit can then be examined using FIML, listwise deletion, pairwise deletion, an application of the EM algorithm using NORM (Schafer, 1997) and finally the pattern-matching imputation method. This will enable researchers to make comparisons about estimate bias for missing data in SEM for the MCAR condition.

References

Allison, P. D. (1987). Estimation of linear models with incomplete data. In C. C. Clogg (Ed.), *Sociological Methodology* (pp. 71-103). San Francisco: Jossey-Bass.

Arbuckle, J. L. (1995). *AMOS for Windows Analysis of Moment Structures. Version 3.5.* Chicago: SmallWaters Corp.

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides and R. E. Schumacker (Eds.). *Advanced structural equation modeling* . Mahwah, NJ: Lawrence Erlbaum Publishers.

Brown, C. H. (1983). Asymptotic comparison of missing data procedures for estimating factor loadings. *Psychometrika, 48*(2), 269-292.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39*, 1-38.

Kline, R. B. (1998). *Principles and practices of structural equation modeling.* New York: Guilford.

Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika, 52*, 431-462.

Neale, M. C. (1994) *Mx: Statistical modeling (2nd Edition).* Department of Psychiatry: Medical College of Virginia.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 61*, 581-592.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: J. Wiley & Sons.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data.* London: Chapman.

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research, 33*(4), 545-571.

Wothke, W. (1998). Longitudinal and multi-group modeling with missing data. In T. D. Little,

K. U. Schnabel, and J. Baumert [Eds.]. *Modeling longitudinal and multiple group data: Practical issues, applied approaches and specific examples.*. Mahwah, NJ: Lawrence Erlbaum Publishers.