Missing Data and Missing Data Estimation in SEM

Listwise Deletion

For many analyses, listwise deletion is the most common way of dealing with missing data. That is, complete data are required on all variables in the analysis—any cases with missing values on one or more of the variables was eliminated from the analysis. Simulation studies, however, convincingly show that when there are a lot of missing values, listwise deletion will have biased parameters and standard errors (see Enders, 2001, for an illustration). With SEM software, estimation that uses all cases often has been integrated into the analyses by default. They type of model estimation is an extension of maximum likelihood for complete cases, called full information maximum likelihood (FIML). If certain assumptions are met, FIML consistently outperforms listwise deletion on parameter, with unbiased or less biased parameter estimates, more efficient standard errors, accurate Type I error rates, and greater statistical power (e.g., Enders & Bandalos, 2001).

Missing Data Assumptions: MCAR, MAR, and MNAR

A distinction of the type of missing data was made by Rubin (1976), who classified missing values as missing at random (MAR), missing completely at random (MCAR), or neither. These classifications are referred to as "missing data mechanisms." Both MAR and MCAR require that the variable with missing values be unrelated to whether or not a person has a missing value for that variable. For example, if those with lower incomes are more likely to have missing values on the income variable, the data cannot be MAR or MCAR. The difference between MAR and MCAR is whether or not other variables in the data set are associated with whether someone has missing values on a particular variable. For example, are older people more likely to refuse to respond to the income variable? The term MAR can be confusing because data are not really missing at random—missingness seems to depend on some of the variables in the data set. In fact, missingness can even be related to the real values of the variable with missing values as long as that relationship can be accounted for by other variables in the data set. When missing values are not at least MAR, missingness is said to be *not missing at random* (NMAR or MNAR) or *nonignorable* (there is a distinction between MNAR and nonignorable, but they are often treated as synonymous in practice). Several sources provide more comprehensive overviews of missing data mechanisms (Enders, 2020; Graham, 2012; Little & Rubin, 2020).

Determining Whether Missing Values are MAR or MCAR

Modern missing data analysis approaches assume that the data are at least MAR. But, practically speaking, it is not really possible to know for sure that your data are MAR, because you do not have information about the value of the variable that is missing. In the words of Schafer and Graham (2002): "When missingness is beyond the researcher's control, its distribution is unknown and MAR is only an assumption. In general, there is no way to test whether MAR holds in a data set, except by obtaining follow-up data from nonrespondents or by imposing an unverifiable model." (p. 152).[1] We may not be completely in the dark in all situations. Some circumstances allow for a little bit of information about the likelihood that values are MAR, although none provide certainty. With longitudinal data and data missing due to attrition, one could explore whether missingness is associated with the value of the variable by examining whether the variable at Time 1 (i.e., with complete data) is associated with the missingness for that variable at Time 2. The passage of time, however, makes the Time 1 values an imperfect proxy for the Time 2 values that are missing. For latent variable estimation with missing values on some indicators, an approximate approach might be to attempt to show that missingness on particular items is unrelated to scale scores for that measure (this can be also be approached within SEM by estimating the association between the latent variable and a missingness indicator, Falcaro, Pendleton, & Pickles, 2013). In other circumstances, one may want to attempt a theoretical argument that missingness is not associated with the variable or rely on information in the literature.

[1] Little (1988) has a test for MCAR, however, and Enders offers a macro to conduct the test, http://www.appliedmissingdata.com/macro-programs.html.

FIML

Probably the most pragmatic missing data estimation approach for structural equation modeling is full information maximum likelihood (FIML), which has been shown to produce unbiased parameter estimates and standard errors under MAR and MCAR (Enders & Bandalos, 2001). FIML, sometimes called "direct maximum likelihood," "raw maximum likelihood" or just "ML," is currently available in all major SEM packages. FIML requires that missing values to be at least MAR (i.e., either MAR or MCAR are ok). The process works by estimating a likelihood function for each individual based on the variables that are present so that all the available data are used. For example, there may be some variables with data for all 389 cases but some variables may have data for only 320 of the cases. Model fit information is derived from a summation across fit functions for individual cases, and, thus, model fit information is based on all 389 cases.

With missing data, the FIML fit function (Arbuckle, 1996) is computed for each set of cases with the same unique pattern of missing values—a casewise likelihood. So, an i subscript is used in the equation below to show that the fit function is for each particular case:

$$\log L_1 = -\frac{k_i}{2}\log(2\pi) - \frac{1}{2}\log\left|\boldsymbol{\Sigma}_i\right| - \frac{1}{2}\left(\mathbf{Y}_i - \boldsymbol{\mu}_i\right)^T \boldsymbol{\Sigma}_i^{-1}\left(\mathbf{Y}_i - \boldsymbol{\mu}_i\right)$$

$\log$ is the natural logarithm (with base e), π is the mathematical constant, $\boldsymbol{\Sigma}_i$ is the population covariance matrix, $\mathbf{Y}$ is vector of observed variables, $\boldsymbol{\mu}_i$ is the vector of their means, and the superscript T is the matrix transpose function. The subscripts i denote individual cases in which the $\mathbf{Y}_i$ vectors differ in length depending on the number of present values. Similarly, matrices $\boldsymbol{\Sigma}_i$ and $\boldsymbol{\mu}_i$ vary by deleting rows and columns for missing variables. The sum of the individual log likelihoods is then computed. Rather than the traditional approach to calculating chi-square, FIML estimates two models, the H_0 model and the H_1 model. The H_1 model is the "unrestricted" model, meaning that all variables are correlated. The H_0 model is the specified model. The difference between the two log-likelihoods is used to derive the chi-square (Muthén, 1998-2004, Appendix 6). This process approach allows one to use all the available information in the variables.

Mplus and lavaan allow the user to specify the type of information matrix used in the FIML estimation. The information matrix (or Fisher's information matrix—the inverse of the Hessian matrix) is used for the standard error estimates. The information matrix using the observed matrix (by default in Mplus and lavaan) produces better standard errors (Savalei, 2010) for most typical models with missing data. Expected values may be used instead,[2] but this is not usually recommended because the standard errors may be underestimated if values are only MAR (Enders, 2010; Kenward & Molenberghs, 1998).

Auxiliary Variables

There is good evidence to suggest that using modern missing data estimation approaches may be advantageous even if missingness is initially nonignorable (i.e., MAR assumptions have not been met) provided correlates of missingness, referred to as *auxiliary variables*, are included in the model. Auxiliary variables are expected to be correlated with missingness on the key variables in the model but are not variables that would have been otherwise included in the mdoel. Inclusion of auxiliary variables has the most impact when their association with missingness is high (e.g., > .4) and when the amount of missing values is large (e.g., > 25%; Collins, Schafer, & Cam, 2001; Graham, 2003). Graham shows that two methods of modeling these auxiliary variables (either as dependent variables or correlated variables) are equally effective in reducing parameter biases, but including auxiliary variables as correlates has a greater impact on reducing biases in model fit.

[2] In the ANALYSIS: command, one can specify INFORMATION = EXPECTED; to change the default. In R, using information.expected with lavTech() changes the matrix from the observed.

Other Missing Data Approaches

Multigroup SEM Approach. An older approach to missing data analysis uses a multigroup structural model approach, suggested by Muthen, Kaplan, and Hollis (1987). The same model is estimated in different groups. The groups are based on different patterns of missing values—one group for each pattern. A few hand calculations must be done. This is a fairly impractical approach if there are many patterns of missing values, but might be especially useful if data are missing by design. This approach has been superceded in some cases by a latent class approach to missing data (Muthen & Muthen, 2002).

Pairwise Deletion. Pairwise deletion is sometimes used to estimate models when there are missing values. With pairwise deletion, a covariance (or correlation) matrix is computed where each element is based on the full number of cases with complete data for each pair of variables. This approach may lead to nonpositive definite matrices and to standardized values over 1. There are other potential problems with the approach and I do not recommend it (see Enders, 2010, for more information).

Other imputation methods. There are several other estimation approaches in which the data are imputed. That is, a full data set is created based on the imputation method that fills in data based on information from existing data. Older methods, such as mean imputation (the average scores is filled in), regression-based methods (a regression is used to predict a score), and resemblance-based "hot-deck imputation" (which imputes new values from similar cases) do not perform as well as other methods, and some may produce highly biased coefficients and/or standard errors (Gold & Bentler, 2000). Two newer methods, multiple imputation (MI; see Graham & Hofer, 2000) and a separate maximum likelihood estimation step using an expectation maximization algorithm (EM; see Enders & Peugh, 2004) provide estimates on par with those obtained with FIML, but tend to be less convenient because separate steps are usually required.

Comments

If there are missing values for a large number of cases and the mechanism is MAR, there are clear advantages to using modern missing data approaches (FIML, EM, or MI) compared with listwise deletion or older imputation methods. What is a large amount of missing data? The percentage of cases with missing values is sometimes discussed based on the percentage missing for a certain variable, which can be confusing when the cases that are missing values differs across variables. It makes most sense to me to consider the percentage of cases missing if listwise deletion were to be used (percentage of cases missing data on one or more variables). With this definition, my reading of the simulation studies suggests data sets (i.e., the set of variables in the model) in which more than roughly 10-20% of the cases are excluded by listwise deletion seem to lead to substantial bias in regression estimates and relative efficiency (standard errors; e.g., Arbuckle, 1996). With fewer than this many missing cases, the impact of missing data may not be as consequential, but, at least within structural equation modeling packages, the potential gain in at least some precision, given its convenience in implementing, make use of FIML instead of listwise deletion the wisest choice. Even if the mechanism is not MCAR or MAR (i.e., data are NMAR), modern missing data estimation will be preferable to listwise deletion if appropriate auxiliary variables are included in the model, because their inclusion reduces the impact of NMAR (Graham, 2009; although see Thoemmes & Rose, 2014, for some exceptions). Given that FIML is now easy to implement in the packages where it is available (often even the default), it is increasingly difficult to argue that one should *not* use it. Missing data estimation with nonnormal is also available in some packages (e.g., EQS, Mplus). Scaled chi-square and robust standard errors obtained with this estimation approach appears to work well (Yuan & Bentler, 2000). In Mplus, *estimator=MLR* is used to obtain the robust estimates with missing data (more on robust estimates later).

References and Further Readings

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: Issues and Techniques* (243-277).. Mahwah, NJ: Lawrence Erlbaum Associates.

Collins, L. M, Schafer, J.L., & Kam, C-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. Structural Equation Modeling, 6, 330-351.

Enders, C.K. (2001). A primer on maximum likelihood algorithms available for use with missing data. Structural Equation Modeling, 8, 128-141.

Enders, C. K (2010). *Applied missing data analysis*. New York: Guilford.
Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural equation modeling*, *8*(3), 430-457.
Enders, C.K, & Peugh, J.L. (2004). Using an EM covariance matrix to estimate structural equation models with missing data: Choosing an adjusted sample size to improve the accuracy of inferences. Structural Equation Modeling, 11, 1-19.
Enders, C.K. (2013). Analyzing structural equation models with missing data. In G.R. Hancock & R.O. Mueller (Eds.), *Structural equation modeling: A second course, second edition* (pp. 493-520). Charlotte, NC: Information Age Publishing.
Falcaro, M., Pendleton, N., & Pickles, A. (2013). Analysing censored longitudinal data with non-ignorable missing values: depression in older age. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176, 415–430.
Gold, M.S., & Bentler, P.M. (2000). Treatments of missing data: A Monte Carlo comparison of RBHDI, iterative stochastic regression imputation, and expectation-maximization. Structural Equation Modeling, 7, 319-355.
Graham, J.W. (2003). Adding missing-data-relevant variables to FIML-based structural equation models. Structural Equation Modeling, 10, 80-100.
Graham, J. W. (2012). *Missing data: Analysis and design*. Springer Science & Business Media.
Graham, J.W., & Hofer, S.M. (2000). Multiple imputation in multivariate research. In T.D. Little, K.U. Schnabel, and J. Baumert (Eds.), *Modeling longitudinal and multilevel data: Practical issues, applied approaches, and specific examples* (pp. 201-218). Mahwah, NJ: Erlbaum.
Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60*, 549-576.
Kenward, M. G., & Molenberghs, G. (1998). Likelihood based frequentists inference when data are missing at random. *Statistical Science, 13,* 236-247.
Little, R.J.A., & Rubin, D.B. (1989). The analysis of social science data with missing values, Sociological Methods and Research, 18, 292-326.
Little, R. J., & Rubin, D. B. (2020). *Statistical analysis with missing data, third edition*. *New York: Wiley.*
Muthén, B.O. (1998-2004). *Mplus Technical Appendices*. Los Angeles, CA: Muthén & Muthén
Muthen, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. Psychometrika, 51,431-462.
Muthén, L.K. and Muthén, B.O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. Structural Equation Modeling, 4, 599-620.
Marsh, H.W. (1998). Pairwise deletion for missing data in structural equation models with missing data: Nonpositive definite matrices, parameter estimates, goodness of fit, and adjusted sample sizes. Structural Equation Modeling, 5, 22-36.
Rubin, D.B. (1976). Inference with missing data. Biometrika, 63, 581-592.
Savalei, V. (2010). Expected versus observed information in SEM with incomplete normal and nonnormal data. *Psychological methods, 15*(4), 352.
Schafer, J.L., & Graham, J.W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147-177.
Thoemmes, F., & Rose, N. (2014). A cautious note on auxiliary variables that can increase bias in missing data problems. *Multivariate Behavioral Research, 49*, 443-459.
Yuan K. H. and P.M. Bentler. 2000. “Three Liklihood-Based Methods for Mean and Covariance Structure Analysis with Non-Normal Missing Data.” Sociological Methodology 2000:165-200. Washington, D.C. American Sociological.