Missing data (from CRAN Task View: Multivariate Statistics)
Packages in R that deal with missing values: add Amelia, missPCA,

Maintainer: Paul Hewson

- mitools provides tools for multiple imputation; mice provides (among many other functions) multivariate imputation by chained equations; mvnmle provides ML estimation for multivariate normal data with missing values, mix provides multiple imputation for mixed categorical and continuous data. pan provides multiple imputation for missing panel data. VIM provides methods for the visualisation as well as imputation of missing data. `aregImpute()` and `transcan()` from Hmisc provide further imputation methods. monomvn deals with estimation models where the missing data pattern is monotone.

Visualising multivariate data

- *Graphical Procedures:* A range of base graphics (e.g. `pairs()` and `coplot()`) and lattice functions (e.g. `xyplot()` and `splom()`) are useful for visualising pairwise arrays of 2-dimensional scatterplots, clouds and 3-dimensional densities. `scatterplot.matrix` in the car provides usefully enhanced pairwise scatterplots. The cwhmisc package provides `plotSplomT()` which displays correlation values and adds histograms on the diagonal of scatterplot matrices. Beyond this, scatterplot3d provides 3 dimensional scatterplots, aplpack provides bagplots and `spin3R()`, a function for rotating 3d clouds. misc3d, dependent upon rgl, provides animated functions within R useful for visualising densities. YaleToolkit provides a range of useful visualisation techniques for multivariate data. More specialised multivariate plots include the following: `faces()` in aplpack provides Chernoff's faces; `parcoord()` from MASS provides parallel coordinate plots; `stars()` in graphics provides a choice of star, radar and cobweb plots respectively. `mstree()` in ade4 and `spantree()` in vegan provide minimum spanning tree functionality. calibrate supports biplot and scatterplot axis labelling, chplot provides convex hull plots. geometry, which provides an interface to the qhull library, gives indices to the relevant points via `convexhulln()`. ellipse draws ellipses for two parameters, and provides `plotcorr()`, visual display of a correlation matrix. denpro provides level set trees for multivariate visualisation. Mosaic plots are available via `mosaicplot()` in graphics and `mosaic()` in vcd that also contains other visualization techniques for multivariate categorical data. gclus provides a number of cluster specific graphical enhancements for scatterplots and parallel coordinate plots See the links for a reference to GGobi; rggobi interfaces with GGobi, DescribeDisplay provides an interface to

GGobi plugins yielding publication quality graphs. xgobi interfaces to the XGobi and XGvis programs which allow linked, dynamic multivariate plots as well as projection pursuit. Finally, iplots allows particularly powerful dynamic interactive graphics, of which interactive parallel co-ordinate plots and mosaic plots may be of great interest. Seriation methods are provided by seriation which can reorder matrices and dendrograms.

- *Data Preprocessing:* `summarize()` and `summary.formula()` in Hmisc assist with descriptive functions; from the same package `varclus()` offers variable clustering while `dataRep()` and `find.matches()` assist in exploring a given dataset in terms of representativeness and finding matches. Whilst `dist()` in base and `daisy()` in cluster provide a wide range of distance measures, proxy provides a framework for more distance measures, including measures between matrices. simba provides functions for dealing with presence / absence data including similarity matrices and reshaping.

Missing Data and Missing Data Estimation – by Craig Enders ('07) {Searching on 'missing data Enders' will be worthwhile; he also has a new book on md}

Listwise Deletion

Until recently, listwise deletion has been the most common way of dealing with missing data in SEM. That is, complete data were required on all variables in the analysis—any cases with missing data on one or more of the variables was eliminated from the analysis. In the last few years, however, researchers have begun to use data estimation techniques when there are missing data among the variables in a structural model. And simulation studies convincingly show that when there are a lot of missing data, listwise deletion will have biased parameters and standard errors (see Enders, 2001, for an illustration).

MAR and MCAR

A distinction of the type of missing data was made by Rubin (1976), who classified missing data as missing at random (MAR), missing completely at random (MCAR), or neither. Both MAR and MCAR require that the variable with missing data be unrelated to whether or not a person has missing data on that variable. For example, if those with lower incomes are more likely to have missing data on the income variable, the data cannot be MAR or MCAR. When data are not MAR or MCAR, missingness is sometimes said to be "nonignorable". The difference between MAR and MCAR is whether or not other variables in the data set are associated with whether someone has missing data on a particular variable. For example, are older people more likely to refuse to respond to the income variable? The term MAR is 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 variable with missing data, as long as that relationship can be accounted for by other variables in the data set. When missing data are not at least MAR, missingness is said to be nonignorable.

Determining If Missing Data is 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 a recent discussion of missing data estimation, Schafer and Graham (2002) state: "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).

We may not be completely in the dark in all situations, however. 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. With missing data on indicators of latent vairables, an approximate approach might be to attempt to show that missingness on particular items is unrelated to scale scores for that measure. 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.

(Little (1988) has a test for MCAR, however, and Enders offers a macro to conduct the test, https://webapp4.asu.edu/directory/person/839490. Newsom, USP 655 SEM, Winter 2010)

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. FIML, sometimes called "direct maximum likelihood," "raw maximum likelihood" or just "ML," is currently available in Amos, Mplus, Mx, and Lisrel. FIML requires that data 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 in is derived from a summation across fit functions for individual cases, and, thus, model fit information is based on all 389 cases. Rather than the traditional approach to calculating chi-square, FIML estimates two models, the H0 model and the H1 model. The H0 model is the "unrestricted" model, meaning that all variables are correlated. The H1 model is the specified model. The difference between the two loglikelihoods is used to derive the chi-square. This approach allows one to use all the available information in the variables.

Recent work illustrates that using modern missing data estimation approaches may be reasonable even if missingness is nonignorable (i.e., MAR assumptions have not been met) provided correlates of missingness (auxiliary variables) are included in the model. Inclusion of auxiliary variables has the most impact when their

association with missingness is high (e.g., > .4) and when the amount of missing data 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.

Other Missing Data Approaches

Multigroup SEM Approach. Another 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 data—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 data, 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 data. 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 non-positive definite matrices and to standardized values over 1. There are other potential problems with the approach and I do not recommend it.

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 Expectation Maximization (EM; which is a maximum likelihood-based approach; 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 is a large amount of missing data and data are at least 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 missing data is sometimes discussed based on the percentage missing for a certain variable. It makes more sense to me to examine the percentage of cases missing if listwise deletion were to be used. With this method, data sets (i.e., the set of variables in the model) in which more than roughly 20% of the cases are excluded by listwise deletion seem to lead to substantial bias in estimates (e.g., Arbuckle, 1996). With fewer than this much missing data, missing

data may not be as consequential. Simulation results now also suggest that even if data are not at least MAR, modern missing data estimation will be preferable to listwise deletion if auxiliary variable as included in the model.

Given that FIML is now easy to implement in the packages where it is available, 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.

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.

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, & 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. (2006). Analyzing structural equation models with missing data. In G.R. Hancock & R.O. Mueller (Eds.), Structural equation modeling: A second course (pp. 313-342). . Greenwich, CT:

Information Age Publishing.

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., & 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.

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.A. & Rubin, D.B. (1987). Statistical analysis with missing data. New York: Wiley.

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.

Schafer, J.L., & Graham, J.W. (2002). Missing data: Our view of the state of the art. Psychological Methods, 7, 147-177.

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.