MEASUREMENT AND ASSESSMENT

Missing Data by Design: The Good News about Missing Data

—by Jennie G. Noll

Few aspects of doing research are as frustrating as the loss of data. Data are lost or missed in various ways and for many reasons: questions are accidentally omitted, subjects skip pages or items, hard-drives crash, and questionnaire items are inapplicable. What is the solution to this problem, which is so endemic to the study of human behavior?

Several widely used techniques for dealing with "missingness" are less than optimal. At the same time, technology has the capacity to deal with missingness in statistically sound ways. This discussion outlines current, widely used missing data techniques and their shortcomings and presents research designs, statistical techniques, and the conceptual means to better deal with the reality of missing values.

This discussion outlines current, widely used missing data techniques and their shortcomings and presents research designs, statistical techniques, and the conceptual means to better deal with the reality of missing values.

Perhaps the most maddening piece of missing data is the one questionnaire item that a subject passed over. In such a case, a researcher must face the possibility of having to drop the subject from final analysis because a linear composite score will not be computed if even one question is skipped.

Common options for addressing missing data

At this point, several commonly used options for dealing with these missing data are widely available:

- *Listwise deletion of cases.* This is perhaps the most obvious method for dealing with incomplete data. The computer program discards all cases with missing values. For most multivariate algorithms, this is usually the default.

- *Mean substitution.* This is the most commonly used and (being the default in many statistical software packages) most widely distributed technique to deal with missing values. The logic behind the procedure stipulates that a missing value can be "replaced" with the mean value of that item. The mean value is calculated first for the subjects who completed the item, and everyone who did not answer the item receives this mean value. While this technique does result in an increase of statistical power—the total sample size goes up—the variance of the variable is truncated. Thus, this procedure can lead to misleading results because of its tendency to attenuate variance and covariances.

- *Pairwise deletion of cases.* This method, also called the piecemeal method, employs all available pairs of values in the computation of covariances. As long as the integrity of the covariance matrix is maintained, this procedure has been shown to be preferable to listwise deletion or mean substitution (Raymond & Roberts, 1987). But the pairwise procedure does not help a researcher who is concerned with calculating means, variances, linear composites, or other item-level statistics.

Regression-based procedures

These three procedures offer little for the frustrated researcher in our scenario. Other procedures, however, may appease this researcher—procedures that use relevant, present information to estimate missingness at the item level. These procedures are regression based and use information present in the data to estimate the missing values for the variable of interest. The item for which there is at least one missing case is the dependent variable, and items that are present in the data are used as independent, or predictor, variables, in the equation. Some of these procedures include multiple regression algorithms (e.g., Beale & Little, 1979) as well as iterative principal components procedures (e.g., Gleason & Staelin, 1975). One popular regression-based program is the AM procedure included in BMDP. Other programs are available, however, which allow greater flexibility and minimal estimation bias.

EMCOV23.EXE

One such program is the EMCOV 23.EXE program (Graham & Hofer, 1995).

continued on next page

The Good News about Missing Data

Continued from page 22

This iterative estimation program employs the EM (expectation-maximization) algorithm (Dempster, Laird, & Rubin, 1977). The algorithm begins with the E-step, involving the collection of sums and sums of squares and cross-products as the data are read into the program. For sums, if the data value is present, the value is added to the overall variable sum. If the data value is missing, the best estimate for the value is added to the overall variable sum. The best estimate is based on a regression equation, with all other variables in the solution as predictors. For sums of squares and cross-products, if either data value is present, the square or cross-product is based on the actual value for the present variable and the best estimate of the other variable. If both values are missing, the square or cross-product is based on the best estimate of the values plus a penalty term.

The worst case scenario is when missingness is nonrandom and is nonignorable.

The M-step of EMCOV23 involves simply calculating the covariance matrix elements based on the sums and sums of squares and cross-products obtained from the E-step. The regression weights from this part of the procedure are used in the next E-step to obtain the best estimates for missing values. These steps are repeated until the change in the estimates of the covariance matrix reaches some minimal convergence criterion. The program is user-friendly, widely available in the public domain, and very flexible. However, minimizing the residual component of the estimation is not the only concern for a researcher who wishes to exert maximum control over estimation bias. A number of other factors must be considered, such as the percentage of missingness, the extent to which the missingness occurs at random, and the relevance of the information used to estimate the missing values.

Levels of randomness

Studies have shown that estimation by regression appears to be most valuable in circumstances in which 10% to 40% of the data are missing (e.g., Little, 1979) and that the advantages of iterative procedures (versus simple regression) become more apparent as the percentage of missing values exceeds 15% to 20% (Raymond & Roberts, 1987). Whenever large proportions of data are missing, however, the appropriateness of using any statistical procedure should be called into question. Why are the items missing? Are they missing at random?

"Randomness" occurs at several levels. The best case scenario—when a researcher knows that estimation bias is considerably controlled—occurs when the data are missing *completely at random* (Rubin, 1987). This term refers to cases in which the cause for missingness is a random process and (by definition) is uncorrelated with other variables. Unfortunately, this is either rare or difficult to confirm.

At the next level, the missingness is nonrandom but is *ignorable* (Little & Rubin, 1987). Ignorable missingness occurs when the cause for missingness can be identified, is unrelated to (not dependent upon) the main dependent variable of the study, and is "accessible" (Graham & Donaldson, 1993). Accessible missing data mechanisms are those causes of missingness that have been measured for all cases and are available for analyses. For example, a packet of questionnaires that was missing a page of items was administered on a day when the youngest subjects turned out for testing. Age is not related to the main dependent variable of the study and has been measured for all subjects. This is a likely scenario.

The worst case scenario is when missingness is nonrandom and is nonignorable (Little & Rubin, 1987). This occurs when the cause for missingness is related to the main dependent variable and is "inaccessible"—has not been measured for everyone or is otherwise unavailable for analysis (Graham & Donaldson, 1993). This would be the case in the scenario described earlier if age were a key hypothesis (directly related to the main dependent variable of the study) or if age were not measured for all subjects.

Most relevant variables

Another way to control estimation bias in missing data imputation is to use only the

continued on next page

The Good News about Missing Data

Continued from page 23

most relevant information to estimate missing values. Using the most relevant information to estimate missingness reduces bias because variables that are irrelevant, or uncorrelated, do not enter into the estimation of the variable of interest. The most relevant variables for the estimation of a missing value are the variables that "load" together with the missing variable in a factor-analytic sense. Variables that are part of the same common factor are the most relevant to one another and should be the only variables used to estimate missingness in one another. For example, if one item of a self-concept scale is missing, only the items that make up the self-concept factor should be used to estimate that missing item. This procedure works best when subscales are reliable (internally consistent) and measurement invariance has been demonstrated for all subjects (or groups of subjects) in the sample (Horn & McArdle, 1992). The EMCOV23 program allows a researcher to include only the most relevant variables for the estimation of missing values.

Missing data do not have to ruin a good research project. In fact, the best research designs of the future will be those that systematically build in missingness.

Second-order missing data imputation

These considerations address missingness only at the item level. What happens, however, when an entire scale is missing (e.g., the depression measure was inadvertently not given), or a piece of data for which there are no apparent predictors is lost (e.g., blood samples were lost due to storage problems)? These problems are more complex, but can be solved using the same methods discussed earlier. The difference is that the missing data imputation is now at the second order: composite scores or factor scores are used to impute other composite or factor scores. Constructs that are highly correlated can be used to predict scores on similar constructs. In these examples, scores on self-esteem and anxiety measures could be used to predict depression scores, and blood hormone levels could be predicted from saliva samples or other physical growth measurements that are present. Second-order missing data imputation should be done on an analysis-by-analysis basis. To avoid linear dependencies and spurious findings, constructs that are used to impute missing values for the construct of interest must not be used in subsequent analyses where the construct of interest is present.

Conclusion

In conclusion, some simple, underused programs are readily available to deal with missing data at the item level or at the second-order (or construct) level with limited estimation bias. The data do not necessarily have to be missing completely at random, but the cause of the missingness has to be at least ignorable under nonrandom conditions. As long as the nonrandom conditions are ignorable, iterative regression-based procedures, such as the EMCOV23 program, can deal with considerably large proportions of missing data (15% to 20%) with little bias.

Researchers must make every attempt to use only the most relevant information when selecting the variables that will be used to predict missing values. Missing data do not have to ruin a good research project. In fact, the best research designs of the future will be those that systematically build in missingness. Several researchers have suggested that missing data is a less than monumental problem, and have advocated research designs with planned missingness (Bell, 1954; McArdle & Hamagami, 1992; McArdle, 1994; Graham, Hofer, & MacKinnon, 1996).

Note: The EMCOV23 program is available via FTP. Use your FTP software to log in to: ftp cac.psu.edu:

login: anonymous

password: your e-mail address

Change to the directory: /pub/people/. DOS files are in the subdirectory "dos," Multiple Imputation files (for DOS) are in the subdirectory "multimp," Windows NT and Windows 95 files are in "NT." The main BINARY file to download is: emcov.exe (in the DOS subdirectory), which is a self-extracting ZIP file Pcs. If you download EMCOV.EXE from this Penn State ftp server, or have trouble doing so, please send an e-mail message to: jwg4@psuvm.psu.edu, telling the author that you have received it.

continued on next page

The Good News about Missing Data

Continued from page 24

References

Beale, E M L., & Little, R J A. (1975). Missing values in multivariate analyses. *Journal of the Royal Statistical Society, Series B*, 37, 129-145.

Bell, R Q. (1954). An experimental test of the accelerated longitudinal approach. *Child Development, 25*, 281-286.

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

Gleason, T C., & Staelin, R. (1975). A proposal for handling missing data. *Psychometrika, 40*, 229-252.

Graham, J W., & Donaldson, S I. (1993). Evaluating interventions with differential attrition: The importance of nonresponse mechanisms and use of follow-up data. *Journal of Applied Psychology, 78*(1), 119-128.

Graham, J W., & Hofer, S M. (1995). *Reference Manual for COVIMP a multiple imputation procedure used in conjunction with EMCOV*. Department of Biobehavioral Health, Penn State University.

Graham, J W., Hofer, S M., & MacKinnon, D P. (1996). Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Research, 31*, 197-218.

Horn, J L., & McArdle, J J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18*, 117-144.

Little, R J A. (1979). Maximum likelihood inferences for multiple regression with missing values: A simulation study. *Journal of the Royal Statistical Society, Series B*, 41, 76-87.

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

McArdle, J J. (1994). Structural factor analysis experiments with incomplete data. *Multivariate Behavioral Research, 29*(4), 409-454.

McArdle, J J., & Hamagami, F. (1992). Modeling incomplete longitudinal and cross-sectional data using latent growth structural models. *Experimental Aging Research, 18*, 145-167.

Raymond, M R., & Roberts, D M. (1987). A comparison of methods for treating incomplete data in selection research. *Educational and Psychological Measurement, 47*, 13-26.

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

Jennie G. Noll, PhD, is at the National Institute of Mental Health, Bethesda, Maryland.

Members Get Members

For every organization, word of mouth is the best form of advertising. Please help strengthen APSAC's voice and achieve APSAC's mission by telling your colleagues and students about APSAC. Urge them to support the organization—first by joining, then by telling yet more colleagues about its mission and benefits. Call 312-554-0166 and ask for Howard Griffin if you would like to receive information about APSAC to distribute to colleagues.

APSAC Benefits of Membership

- The *APSAC Advisor*, the interdisciplinary, hands-on style quarterly newsjournal.
- *Child Maltreatment*, the quarterly, peer-reviewed interdisciplinary journal
- Free copies of APSAC's guidelines for practice, fact sheets, and position papers
- Discounts on APSAC's books, monographs, audiotapes, and other publications.
- Discounts on APSAC's interdisciplinary Colloquium, Institutes, and other conferences nationwide.
- Participation in APSAC's state chapters, committees, task forces, Legislative Network, and Legislative ListServ
- Expert guidance on educating legislators and journalists about child abuse and neglect.
- Support of a national interdisciplinary organization focused on child maltreatment.

APSAC Mission

APSAC's mission is to ensure that everyone affected by child abuse and neglect receives the best possible professional response. APSAC is committed to:

- Providing interdisciplinary professional education which promotes effective, culturally sensitive approaches to the identification, intervention, treatment, and prevention of child abuse and neglect.
- Promoting research and guidelines to inform professional practice.
- Educating the public about child abuse and neglect
- Ensuring that America's public policy concerning child maltreatment is well-informed and constructive.

Every member plays a role in achieving this mission. APSAC's leaders invite members' contributions of time, ideas, energy, and expertise to the wide range of APSAC's activities.