Missing Data Analysis with Binary and Ordinal Outcomes

Missing Data Concepts

*MAR and MCAR***.** A distinction of the type of missing data was made by Rubin (1976; Little, 1995), who classified missing values as missing at random (MAR), missing completely at random (MCAR), or neither. Both MAR and MCAR require that the true values of the variable with missing values be unrelated to whether or not a person has missing values on that variable. For example, if those with lower incomes are more likely to have missing values on an income question, 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 or not someone has missing values on a particular variable. For example, are older people more likely to refuse to respond to an income question? The term “missing at random” is confusing because values are not really missing at random—missingness seems to depend on some of the variables in the data set.

Determining whether missing values are MAR or MCAR. Although univariate or multivariate tests (Dixon, 1988; Little, 1988) can be conducted to investigate whether any variables in the data set are related to the probability of missingness on a particular variable, such tests cannot be definitive for a variety of reasons (Enders, 2022), including insufficient power, and could lead one to a false sense of security.[1] And for modern missing data approaches, meeting the MAR rather than the MCAR assumption is what is crucial. 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). One may have to provide a theoretical argument that missingness is not associated with the variable or rely on information in the literature.

In longitudinal studies, it may be useful to distinguish between attrition and intermittent missingness patterns in addition to missingness that occurs within a particular time point (Little, 1995; Newsom, 2024). An attrition or dropout pattern (sometimes called "monotone") occurs when an individual discontinues participation in the study after a certain time point. An intermittent missing data pattern (or nonmonontonic) in which values are missing any particular time point but are present at least once again (including missing values for just particular variables).[2] The missing data pattern does not necessarily imply anything about whether the MAR (or MCAR) assumption has been met or not. Attrition patterns, however, deserve greater suspicion that the variable of interest may be related to the probability of missingness, and therefore not MAR, because health and motivational factors are known to be a factor in tendency to drop out of a study. With longitudinal data, however, analyses can be used to explore this suspicion by examining 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 (Little, 1995).

General Missing Data Remedies

There are a variety of missing data imputation approaches, but most of them are older approaches that produce poor estimates (e.g., mean imputation; Enders, 2022). I highlight listwise deletion, because it is the most common and the default for nearly all analysis procedures in nearly all statistical packages.

---

[1] The Little test is provided in the SPSS missing data module and Mplus, and Craig Ender's has a SAS macro http://www.appliedmissingdata.com/macro-programs.html.

[2] If a participant did not complete the last wave of the study, it may be impossible to classify that individual as belonging to an intermittent or attrition pattern.

Listwise Deletion. Listwise deletion means that complete data on each case is required, and any individual who has missing information on any variable is eliminated. For example,

	j	Y_{ij}	X_{1ij}	X_{2ij}
1	1	10	8	8
2	1	.	9	.
3	1	1	5	5
4	2	3	.	5
5	2	7	8	8
6	2	10	8	.

With listwise deletion, 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. In the example above, only cases 1, 3, and 5 are used in the analysis with listwise deletion. In repeated measures (and growth curve) analysis, each time point (rather that case) must have complete data. Listwise deletion reduces the sample size, adversely impacting significance tests, and will lead to biases in estimates unless data are MCAR (e.g., Enders & Bandalos, 2004; Kim & Curry, 1977).

Other conventional approaches. There are a number of other approaches to data analysis with incomplete data shown to produced biased estimated or significance tests. *Mean imputation* use the average from the sample (or group mean in multilevel analysis) to replace missing values on a variable. Mean substitution generally reduces the variance of variables and therefore leads to underestimate of standard errors (Enders & Bandalos, 2004; Schafer & Schenker, 2000). *Pairwise deletion* is a method of handling data sometimes an option available with OLS regression procedures (or multilevel procedures). 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. The attempt is to maximize sample size by not requiring complete data on all variables in the model. This approach can lead to serious problems and assumes data are MCAR (Little, 1992). Last observation carried forward uses the most recent value obtained for a participant in a longitudinal study. Although sometimes thought to be a conservative approach, last observation can lead to biases in either direction (Molenberghs & Kenward, 2007). Hot-deck imputation replaces values with values from similar other cases, which can lead to substantial biases in regression analysis (Schafer & Graham, 2002).

Modern Missing Data Methods. Modern approaches, in particular multiple imputation (MI; Rubin, 1987) and full maximum likelihood (Dempster, Laird, & Rubin, 1977), which uses a structural modeling approach), produce superior estimates compared with listwise deletion and the other conventional methods mentioned above as long as data are at least MAR (Enders, 2022; Schafer & Graham, 2002). Although these missing data approaches have been shown repeatedly to be less biased and more powerful, they often may not be a dramatic improvement over the default analysis approach using listwise deletion when the amount of missing data is small (perhaps less than 10% of the sample missing if listwise was used; see results of Arbuckle, 1996, for instance). The ease with which they can now be employed, however, suggests there is little cost in likely gain in accuracy by using them more routinely.

The standard multiple imputation approach requires an initial step in which multiple data sets are imputed with some degree of uncertainty built into the imputed estimates. Common recommendations are for approximately 10 to 20 imputed data sets (Graham, Olchowski & Gilreath, 2007; 20 seems to be the most commonly suggested number currently), but Enders (2022) argues that more (e.g., 100) is not too computationally intensive and will not hurt. The

second step combines (or "pools") the analyses from separate data sets and uses variability across the multiple imputations to better estimate standard errors.
Full maximum likelihood generally refers to missing data estimation that is part of testing a structural equation model in software such as Mplus and the lavaan R package. As part of these models, the missing data estimation is employed seamlessly in a single step when specifying a model.

Recent work illustrates that including potential causes or correlates of the variables with missing values (known as “auxiliary” variables) as part of the analysis has important advantages when data are only MAR, particularly when the association of those with the variable with missing values is high (e.g., > .4) and when the amount of missing data is large (e.g., > 25%; Collins, Schafer, & Cam, 2001; Graham, 2003). Both multiple imputation and full information maximum likelihood can incorporate auxiliary variables. Because inclusion of auxiliary variables in the analyses increases the likelihood of meeting the MAR assumption and can reduce the bias when data are MNAR, it is likely preferable to use modern missing data methods with auxiliary variables over default listwise deletion even if there is no way to know whether the MAR assumption is valid or not.

Missing Data with Categorical Data Analysis
For most analyses that we have covered in the present class, the primary option for handling missing data is multiple imputation. Multiple imputation with categorical variables is possible in a number of different software packages, such as `mice` (van Buuren & Groothuis-Oudshoorn, 2011) `mix` (Schafer & Ripley, 2024) R packages or the free standalone software Blimp (Enders & Keller, 2023).[3] The multiple imputation process has two general steps. In the first, multiple data sets are generated, each with different values filling in for the missing values. The replaced values are generated in various ways depending on the program and researcher’s choices, but mostly based on a regression-based process that predicts values using other variables and adds random variation. One common method for this is “factored regression,” as it is called with continuous variables (Lipsitz, & Ibrahim, 1996) or “sequentially specification,” as it is called with categorical variables (Lüdtke et al., 2020), that estimates a multivariate distribution with a product of several conditional univariate distributions, each using decreasing subsets of the conditional univariate variable distributions. In the second step, the planned analysis (e.g., logistic regression) is conducted in each data set and then all of the results are combined, a step that requires special features.

With missing categorical values, sequential probit-based regressions are used for the estimation of missing values at each imputation with values chosen through a Bayesian Markov chain Monte Carlo (MCMC) process (see Enders, 2022 for a more detailed description). The probit regressions enable random draws using the continuous latent y^* distribution. The sequential specification, because used separately for each missing value, can be applied when there is a mix of binary and continuous variables that need imputation.

Some models, such as IRT psychometric analyses or latent class models can be tested in software that incorporates missing data as part of estimation process. With continuous variables, this missing data estimation is known as *full information maximum likelihood* (FIML). The FIML process does not impute any values but instead fits the model and derives the parameter estimates (e.g., loadings) using information for each case. With categorical variables, different estimators are possible, with diagonal weighted least squares (the WLSMV estimation option in Mplus and the lavaan R package). This approach is not entirely full information and therefore requires a stricter missing data assumption of MCAR, whereas the full information methods only require MAR. Another option is

[3] SPSS has a separate module for missing data analysis that will do multiple imputation, but it is a separate add-on for an additional price.

marginal maximum likelihood which is a full information estimator. Bayesian estimation within structural equation modeling or latent class modeling software is expected to perform more like a full information method or multiple imputation, assuming the less strict MAR mechanism.

References

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling* (pp. 243–277). Mahwah, NJ: Erlbaum

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods, 6*, 330_351.

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

Enders, C.K. (2022). *Applied missing data analysis, second edition.* New York: Guilford Press.

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: A Multidisciplinary Journal, 8*, 430–457

Graham, J.W., Olchowski, A.E. and Gilreath, T.D. (2007) How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science, 8*, 206-213.

Keller, B. T., & Enders, C. K. (2023). *Blimp user's guide* (Version 3). Retrieved from www.appliedmissingdata.com/blimp

Kim, J., & Curry, J. (1977). The treatment of missing data in multivariate analyses. *Sociological Methods and Research, 6*, 215–240.

Little, R.J.A. (1988). Missing-data adjustments in large surveys. *Journal of Business and Economic Statistics, 6*, 287–296.

Little, R.J.A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association , 83*, 1198-1202).

Little, R. J. (1992). Regression with missing X's: a review. *Journal of the American Statistical Association, 87*, 1227-1237.

Little, R. J. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association, 90*, 1112–1121.

Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, 2nd edition, New York: John Wiley.

Lüdtke, O., Robitzsch, A., & West, S. G. (2020). Regression models involving nonlinear effects with missing data: A sequential modeling approach using Bayesian estimation. *Psychological Methods, 25*(2), 157.

Molenberghs, G., & Kenward, M. G. (2007). Missing data in clinical studies. Chichester, UK: John Wiley & Sons, Ltd.

Newsom, J.T. (2024). *Longitudinal Structural Equation Modeling: A Comprehensive Introduction*, second edition. New York: Routledge. Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: Wiley.

Schafer JL, Ripley B (2024). *mix: Estimation/Multiple Imputation for Mixed Categorical and Continuous Data. R package version 1.0-13*, https://CRAN.R-project.org/package=mix.

Schafer, J. L., & Schenker, N. (2000). Inference with imputed conditional means. Journal of the American Statistical Association, 449, 144– 154.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*, 1-67.