Missing Data in Multilevel Regression

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.

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,

---

[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.

i	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. This approach is sometimes thought to be a conservative approach but can lead to biases in either direction (Molenberghs & Kenward, 2007). Hot-deck imputuation 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). 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. Current recommendations are for approximately 10 to 20 imputed data sets (Graham, Olchowski & Gilreath, 2007; 20 seems to be the most commonly suggested number currently). The second step combines (or "pools") the separate data sets and uses variability across the multiple imputations to better estimate standard errors. Structural equation modeling packages, such as Mplus and AMOS, use full information maximum likelihood that is employed seamlessly in a single step when specifying a model. 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 20% of the sample missing if listwise was used; see results of Arbuckle, 1996, for instance).

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). The T1 values of the dependent variable can be included in the longitudinal attrition case, and this could often serve as a key auxiliary variable. 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 may be preferable to use modern missing data methods with auxiliary variables over listwise deletion even if there is no way to know whether the MAR assumption is valid or not.

Missing Data with Multilevel Models

Like other analysis procedures, multilevel regression procedures by default do not allow missing data on any of the predictors or the dependent variable for any given case. One exception is that all multilevel analyses are estimated allowing different group sizes, or, in the case of longitudinal data, differing number of time points. Both of these circumstances can be considered missing data problems.

Unequal group sizes as a missing data problem. A term commonly applied to unequal group sizes is *unbalanced* n_j or *unbalanced design*. Consider the following small, hypothetical example with 5 cases, i, and two groups, j, and only one predictor.

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

Because there are a different number of cases in the two groups, one can consider case 6 to have missing data in a design that is supposed to have 6 cases. If averages for groups 1 and 2 are computed and used at level 2 (as when intercepts are estimated), the data from group two have estimated the mean using information only from the two remaining cases. Thus, the estimation of the intercept from group 2 has essentially "imputed" data from case 6. As discussed earlier in class, information is borrowed from the full sample to estimated intercepts (or slopes) for a particular group, and these empirical Bayes estimates are "shrunken" toward the grand mean values. Smaller groups are shrunken more toward the grand mean value.

This same type of missing data issue occurs with longitudinal data too. If the data were in an original wide data (repeated measures) format, it is easier to see that in the same data as in the table above (if the j units were considered individuals and i units were considered time points) is an instance of missing data, where all of the values are complete for the first individual but the value for Y_3 is missing for the second individual.

i	Y_1	Y_2	Y_3
1	10	4	1
2	7	10	.

In multilevel regression, the data are not actually imputed but the model is estimated making use of the incomplete data in a way that does not bias estimates under certain conditions.[3] The estimation in multilevel software (REML) is related to the missing data estimation for more general applications using full maximum likelihood (Little & Rubin, 2002; Raudenbush & Bryk, 2002). The process is comparable to the full maximum likelihood approach in structural equation modeling. And, under certain restrictive conditions, when values are MAR or MCAR, missing data estimation in this sense produces unbiased estimation of the complete data and better estimation than if conventional missing data estimation approaches would have been used.

Other Ways Data May Be Missing in Multilevel Models. van Buuren (2011) provides a nice summary of the ways in which data can be missing with multilevel regression models. Data may be missing on the dependent variable, Y, the level-1 predictors, X, or the level-2 predictors, Z, all of which will lead to loss of

[3] This is not necessarily true about the standard errors, as shrinkage of the estimated variance from this "imputation" leads to standard errors that are too small (Raudenbush & Bryk, 2002, p. 47).

information because of the default missing data handling of multilevel procedures.[4] Exclusion of cases because values are missing in one or more of these ways (listwise deletion), like single-level analyses, can potentially lead to loss of power, biased estimates, and biased standard errors.

Multilevel Missing Data Remedies. As with single-level analyses, full maximum likelihood and multiple imputation approaches offer superior estimates to those that would be obtained with listwise deletion. There are complications, however, in that simple approaches that do not take the nested structure into account can be problematic, and additional precautions are needed because values are missing from nested data are not independent (Dreschler, 2015; van Buuren, 2011). Although there are some preferable approaches that appear to be emerging, there is no clearly ideal solution yet and the area is still under development (see Enders, 2022; Enders & Hayes, 2022; Enders, Keller, & Levy, 2018; Hox, van Buuren, & Jolani, 2015; Grund, Lüdtke, & Robitzsch, 2016, for recent reviews).

Full maximum likelihood (FIML) for missing data is available through some structural equation modeling software, such as Mplus and the `xxM` package (Mehta, 2013) that are also capable of estimating multilevel models. These packages handle missing data within the specified model including auxiliary variables that are included.

There are several possible approaches that have been discussed for multiple imputation with multilevel analysis. Flat file imputation uses does not use the nested structure on the imputation step. Separate classes (or clusters) models takes the groups into account by using the intercepts in the imputation model. A full multilevel approach can use the variables in the model and an iterative process that takes the within and between variance into account (van Buuren, 2011). A full multilevel approach could take several forms. A Bayesian estimation strategy can be used to find distributional prior through a Markov Chain Monte Carlo (MCMC) approach, which can be computationally intense or even impossible for larger models. A fully conditional specification (FCS; sometimes chained equations or sequential regression), which is often used for single level multiple imputation on the initial step, regresses each variable on all other predictors for initial imputations. The multilevel version, however, uses the multilevel model in the imputation step taking into account the multilevel variances. A joint modeling approach (Schafer, 2007) is similar to the FCS for single-level data in which the distributions are assumed to be normally distributed, so does not take the multilevel error structure as fully into account or allow for error heterogeneity (Enders et al.,2018).

In comparing several multilevel multiple imputation methods, van Buuren (2011) found that multilevel multiple imputation with FCS performed better than the other two general approaches (flat file and separate classes), but that there were issues for data sets with extreme values of between group variance and small group sizes. Another possible approach is predictive mean matching (PMM: Little, 1988), which uses a similar regression approach with each variable predicted by all others, but then uses a similar case to replace the missing values. Hox and colleagues (2015) compared the same multilevel multiple imputation method that van Buuren used to maximum likelihood as well as the flat file and separate classes multiple imputations and complete case (listwise) analysis. Their finding indicated that MI and FIML outperformed the other approaches and worked equally well to one another. Enders and colleagues have extended the FCS (chained equations or fully conditional) approach of van Buuren to handle random slopes, missing data at level 2, and noncontinuous outcomes (Enders, Kelly, & Levy, 2018). The approach uses the same Bayesian and FCS methodology and partitions the two levels. Their results suggest the method works well, outperforming the joint modeling approach, for fixed effects and for random intercept variance estimates but produces slope variance estimates that are biased low if missing data greater than 15%.

[4] Data missing on the grouping variable (data missing on j), in which information about group membership is missing, is also sometimes discussed. There is little that missing data estimation techniques can do about this situation, but there are some suggested remedies, such as use of latent class analysis (Browne & McNicholas, 2013).

There has been a recent proliferation of software options for implementing FIML and MI with multilevel data. The FCS multiple imputation approach of Enders and colleagues is currently implemented in new software package called BLIMP (Keller & Enders, 2023). The `MICE` (for multivariate imputation by chained equations) R package uses FCS with MCMC for multilevel data (van Buuren & Groothuis- Oudshoorn, 2011). The joint modeling approach is available in Mplus and the `pan` and `mlmm` packages in R. The `pan` package (Schafer & Zhao, 2014; see Grund, Lüdtke, & Robitzsch, 2016 for an introduction) for the imputation step is used together with the `mitml` package which combines the data sets when testing the multilevel model (Grund, Robitzsch, & Lüdtke, 2016). HLM has a facility to pool estimates from the multiple imputations (Other settings → Estimation Settings → Mulitple Imputation). This method could potentially be used to pool data sets from an initial imputation step using one of the above methods. van Ginkel has a macro that can be used for pooling multivariate imputations in SPSS (van Ginkel, 2010, MI-MUL2.SPS macro).

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

Drechsler, J. (2015). Multiple Imputation of Multilevel Missing Data—Rigor Versus Simplicity. *Journal of Educational and Behavioral Statistics*, *40*(1), 69-95.

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

Enders, C.K. & Hayes, T. (2022). Missing data handling for multilevel data. In A.A. O'Connell, D. B., McCoach, & B.A. Bell, Multilevel modeling methods with introductory and advanced applications (pp. 535-566).

Enders, C. K., Keller, B. T., & Levy, R. (2017). A fully conditional specification approach to multilevel imputation of categorical and continuous variables. *Psychological Methods, 23*, 298-317.

Enders, C. K., Mistler, S. A., & Keller, B. T. (2016). Multilevel multiple imputation: A review and evaluation of joint modeling and chained equations imputation. *Psychological Methods, 21*, 222-240.

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.

Grund, S., Lüdtke, O., & Robitzsch, A. (2016). Multiple imputation of multilevel missing data: An introduction to the R package pan. SAGE Open, 6(4), 2158244016668220.

Grund, S., Robitzsch, A., & Lüdtke, O. (2016). *mitml: Tools for multiple imputation in multilevel modeling (Version 0.3-2)* [Computer software].

Hox, J.J., van Buuren, S., & Jolani, S. (2015). Incomplete multilevel data: Problems and solutions. In Harring, J.R., Stapleton, L.M., Beretvas, S.N. & Hancock, G.R. (Eds.) (2015). *Advances in multilevel modeling for educational research: Addressing practical issues found in real-world applications* (pp. 37-58). Charlotte, NC: Information Age Publishing, Inc.

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.

Mehta, P. (2013). xxM: Structural Equation Modeling for Dependent Data (R package version 0.6.0) [Computer program]. Available online at: http://xxm.times.uh.edu/

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

Newsom, J.T. (2015). *Longitudinal Structural Equation Modeling: A Comprehensive Introduction*. New York: Routledge.

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

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

Schafer, J. L., & Zhao, J. H. (2014). *pan: Multiple imputation for multivariate panel or clustered data (Version 0.9*) [Computer software]. Retrieved from http://CRAN.R-project.org/package=pan

van Buuren, S. (2011). Multiple imputation of multilevel data (pp. 173-196).In Hox, J., & Roberts, J. K. (Eds.). (2011). *Handbook of advanced multilevel analysis*. New York: Routledge.

van Ginkel (2010) http://leidenuniv.nl/fsw/pedagogiek/ginkel/MI-mul2.zip which can be retrieved from http://www.universiteitleiden.nl/en/staffmembers/joost-van-ginkel.