Missing Data and Regression

Missing data is a common problem in applied research. Missing values may occur because of non-response, errors in the data collection, or dropout. With regression analysis, the default in all programs is to eliminate any cases with missing data on any of the variables (i.e., listwise deletion). As the amount of data that is missing increases, there can be a substantial reduction of sample size and a resulting loss of power. As important, there is a potential for biases in the regression estimates and their standard errors (and therefore the significance tests), depending on which values are missing. If the values observed were simply a random sample of the possible values observed, then the only biases would be due to the loss of sample size, as the observed subsample would be just a random sample form the larger possible values if there had been no missing values. Both the loss of sample size and the biases can be addressed in some cases (for comprehensive treatments, see Enders, 2022; Little & Rubin, 2002; Schafer, 2012).

Mechanisms

MAR and MCAR. A distinction about the nature 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 (say Y). For example, are older people more likely to refuse to respond to an income question? If these other variables are related to missingness on Y but the values of Y are not, then the missing values are MAR. If no other variables are related to missingness, then missing values are MCAR. The term “missing at random” is confusing because values are not really missing at random—for MAR, missingness seems to depend on some of the variables in the data set. MCAR is more what we think of when we think values are missing at random. For MCAR, it is as if we took a completely random selection of cases, and deleted their values for a variable.

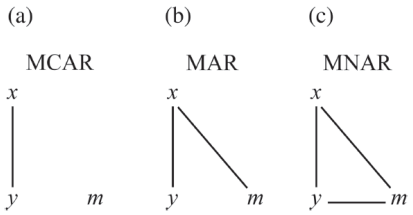

Figure 14.1 Analogue Representation of Missing Data Mechanisms.

Sources: Adapted with permission from Schafer and Graham, 2002, Figure 2, p. 152.

from Newsom (2024), p. 421

Determining Whether Missing Values Are MAR

Researchers can investigate whether any variables in the data set are related to missingness on a variable by computing a new variable that indicates (0, 1) whether data are missing or present and then using correlations or group comparisons. Little (1988) developed a simultaneous test along these lines.[1] If none of the variables in the data set are related to missingness, then the data are observed to be missing completely at random, although this does not guarantee that the values for the missing variable are not related to missingness for that variable (Allison, 2001). Practically speaking it is not possible ever to determine whether the true values of a variable are related to the probability of missingness on that variable, because we do not have the missing information. As Schafer and Graham (2002) state: "When missingness is beyond the researcher's control, its distribution is unknown and MAR is only an

[1] Little's test for MCAR, [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] can be conducted in SPSS with the missing data module (must be separately purchased), in SAS a macro https://communities.sas.com/kntur85557/attachments/kntur85557/sas_iml/4752/1/Little%20Code.docx, with R, https://search.r-project.org/CRAN/refmans/naniar/html/mcar_test.html and through other specialty packages such as Mplus.

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). With attrition over time, it may be possible to test whether missingness is associated with the value of the variable that has present values at an earlier time point (i.e., usually all cases have mostly complete data at the first time point). For example, in a pretest-posttest design, we could investigate whether the variable at Time 1 (i.e., with complete data) is associated with the missingness for that variable at Time 2 (Little, 1995), which provides some information but is nearly always an imperfect proxy. In other circumstances, one may have to provide a theoretical argument that missingness is not associated with the variable or rely on information in the literature. Simulation work illustrates that modeling potential causes or correlates of the variables with missing values has important advantages when values are only MAR, particularly when the association of those “auxiliary” variables 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). So, to the extent that we can incorporate some of the variables or proxies for the variables that may be causally related to the probability of missingness, we may be closer to meeting the MAR assumption. For this reason, there is an argument for always using modern missing data techniques, such as multiple imputation or full maximum likelihood estimation, because there are few if any cases in which listwise deletion would provide better statistical tests.

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,

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 most traditional repeated measures analyses such as ANOVA or regression, each time point (for each 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* uses the average from the sample 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 and sometimes is an option available with OLS regression 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 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 the full maximum likelihood (FIML; Dempster, Laird, & Rubin, 1977) approach used in structural equation modeling, 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 (the I Step) in which multiple data sets are imputed with some degree of uncertainty built into the imputed estimates. There are a number of different methods for doing this (see Enders, 2022, for a nice summary). Current recommendations are for approximately 10 to 20 imputed data sets (Graham, Olchowski & Gilreath, 2007; 20 seems to be the most commonly suggested number lately). In the second step analyses (the P step), the multiple, imputed data sets are analyzed and results are combined (or "pooled") using variability across the multiple imputations to better estimate standard errors in the analysis. The process can be described as a Bayesian process using Monte Carlo simulation to make “draws” from a posterior distribution. Special software or special procedures within existing software are needed for multiple imputation, including SPSS Missing Values (which is an add-on with additional cost), several packages, such as `mice` and `mitml` in R, and free software Blimp (Enders, Keller, & Levy, 2018), which also handles multilevel data sets.

Structural equation modeling packages, such as Mplus, the R lavaan package, and AMOS, use FIML that is employed seamlessly in a single step when specifying a model (Mplus also can be used with MI). Regression models can be specified within these packages conveniently by simply requesting FIML estimation (often it is the default). These packages also easily extend regression models to mediational models. And with Mplus, continuous and categorical variables can be analyzed.

Auxiliary variables. Simulation studies illustrate that including potential causes or correlates of the variables with missing values (known as “auxiliary” variables) as part of the analysis (either with MI or FIML) 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). Auxiliary variables are included in the imputation or estimation (in the case of FIML) without necessarily being included in the model. 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 is likely 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.

References

Allison, P. D. (2001). *Missing data* (Vol. 136). Sage publications.
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
Enders, C.K., Keller, B.T., & Levy, R. (2018). A fully conditional specification approach to multilevel imputation of categorical and continuous variables. *Psychological Methods, 23,* 298-317. http://dx.doi.org/10.1037/met0000148.
Graham, J. W. (2003). Adding missing-data-relevant variables to FIML-based structural equation models. *Structural Equation Modeling*, *10*(1), 80-100.
Graham, J. W. (2012). *Missing data: Analysis and design*. Springer Science & Business Media.
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.
Kim, J., & Curry, J. (1977). The treatment of missing data in multivariate analyses. *Sociological Methods and Research, 6*, 215–240.
Little, R. J. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the american statistical association*, *90*(431), 1112-1121.
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.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, 2nd edition, New York: John Wiley.
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. (1976). Inference and missing data. *Biometrika*, *63*(3), 581-592.
Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: Wiley.
Schafer, J.L., & Graham, J.W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*, 147-177.
Schafer, J. L., & Schenker, N. (2000). Inference with imputed conditional means. Journal of the American Statistical Association, 449, 144–154.