Dealing with Missing Data: Practical Use of Multiple Imputation

Jang Myoung-jin PhD

Medical Research Collaborating Center

Seoul National University Hospital, Seoul, Korea

Understanding the reasons why data are missing can help with analyzing data. The missing mechanisms can be classified as MCAR (missing completely at random), MAR (missing at random), and MNAR (missing not at random). The data are MCAR when the probability that a value for a certain variable is missing is unrelated both to the value of other observed variables and to the variable with missing values. An example is when respondents accidentally skip questions. The data are MAR when the probability that a value for a certain variable is missing is related to observed values on other variables, but unrelated to the variable with missing values. An example is that females are less likely to fill in a depression survey than males but this has nothing to do with their level of depression after accounting for sex. Within the group of male and female respondents, the data are MCAR. The data are MNAR when the probability that a value for a certain variable is missing is related to the values of that variable. An example is that respondents with high depression scores intentionally do not respond to the depression question.

When the data are MCAR, deleting missing cases will not introduce any bias into parameter estimates because the cases with complete data is equivalent to a simple random sample from the original sample. If the data are MAR but not MCAR, deleting missing cases can introduce bias because the cases with complete data are not representative of the original sample. In addition, deleting cases with missing values reduces the statistical power of analysis.

When data are MCAR or MAR, imputation method can be used for dealing with missing data. In single imputation method, missing values for any variable are predicted using values from other variables. Then missing values are substituted by predicted values and standard statistical analysis is carried out on the imputed data set, treating imputed data as observed data. Under MCAR and MAR, single imputation gives unbiased estimates. But this method ignoring uncertainty in imputed values leads to standard error estimates that are biased downward. In multiple imputation (MI), this process is repeated several times. MI produces multiple data sets of the original dataset, where missing data are filled in with values that differ slightly between imputed data sets. Thereafter, estimates and standard errors are calculated in each imputation set using a standard statistical method and pooled into one overall estimate and standard error. MI accounts for missing data uncertainty by creating different multiple imputed data sets and reflecting the variability between imputed data sets in the overall inference in the pooling phase. Therefore multiple imputation provides unbiased estimators and standard errors under MCAR and MAR assumption. In this presentation, we review the basic concepts and general methodology for multiple imputation and illustrate the application of this method in SPSS and SAS using example data.