Tim Vigers

5/5/19

Missing Data

Virtually all real-world studies involve some sort of missing data, and it’s a particularly common issue in clinical research due to participant dropout, loss to follow up, confidentiality concerns, etc. [1] There are many causes of missing data, but they can generally be classified into three groups: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). If the data are MAR (but not completely at random), then using only complete cases can potentially bias the analysis, but statistical techniques like multiple imputation can be used to limit this bias. [2] MAR assumes that “the missingness depends only on observed quantities, which may include outcomes and predictors,” so including a large number of predictors in the model can make the MAR assumption more reasonable. [1]

There are a few ad-hoc approaches to handling missing data, none of which are recommended, but are worth mentioning here as a guide to what not to do. These methods “include replacing missing values with values imputed from the observed data (for example, the mean of the observed values), using a missing category indicator, and replacing missing values with the last measured value (last value carried forward).” [2] Some more generally acceptable approaches include multiple imputation, weighting, and Bayesian methods, although these methods naturally have their own drawbacks. [1]

Multiple imputation is available in several common statistical packages and is based on creating several imputed data sets and combining the results based on all of them. In each dataset copy, missing data are replaced with values randomly sampled from “their predictive distribution based on the observed data—thus multiple imputation is based on a bayesian approach.” This method must also account for uncertainty in predicting missing values by adding appropriate variability in the imputed data. Then the model of interest is applied to each copy of the data set, and the results combined. Obviously this method requires very careful modeling of the distribution of each variable, and can be extremely error-prone. [2]

Weighting methods involve modeling the probability of missingness for a given variable, and then using these probabilities as weights in a complete case analysis. This can be done with any software that can fit weighted models, but “becomes considerably less tractable with multiple missing variables, particularly when they are nonmonotone.” [1]

Although multiple imputation is based on Bayesian methods, there are many other more general approaches that involve “estimation with a prior distribution on the covariates.” [1] Unfortunately there isn’t room to go into detail in this short summary, but Ibrahim et al. (2005) provide much more detail on using prior distributions in GLMs. [3]

Missing data is one of the biggest topics in statistics and one of the most vitally important. There is no perfect way to deal with missing data, and in many cases the optimal approach will depend on the nature of the study. However, multiple imputation appears to be the most generally applicable and robust method, and although it is complex and requires lots of additional time and effort, this work is usually worthwhile for the decrease in bias. [1]

References

1. Horton, N.J. and K.P. Kleinman, *Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models.* Am Stat, 2007. **61**(1): p. 79-90.

2. Sterne, J.A.C., et al., *Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls.* BMJ, 2009. **338**: p. b2393.

3. Ibrahim, J.G., et al., *Missing-Data Methods for Generalized Linear Models.* Journal of the American Statistical Association, 2005. **100**(469): p. 332-346.