Dealing with Missing Data: Resources

Andrew Miles
Last updated: 12/1/14

I originally prepared this document several months ago for Duke's Social Science Research Institute, as suggested resources for learning about appropriately managing missing data. I have modified that document slightly and am posting it in case it might be helpful to others.

Concepts

Book length treatment:
Craig Enders. *Applied Missing Data Analysis*.

I own this book, and think it is great. It provides conceptual overviews of both major approaches to handling missing data (multiple imputation and maximum likelihood approaches), has lots of examples, and is fairly accessible.

For a general overview of missing data methods that is article length, I suggest:

Schafer, Joseph L., and John W. Graham. 2002. "Missing Data: Our View of the State of the Art." *Psychological Methods* 7(2):147–77.

For a fairly recent resource that addresses many practical questions in multiple imputation, I suggest:

White, Ian R., Patrick Royston, and Angela M. Wood. 2011. "Multiple Imputation Using Chained Equations: Issues and Guidance for Practice." *Statistics in Medicine* 30(4):377–99.

This reference addresses the important practical question of which of the two major imputation methods to use:

Lee, Katherine J., and John B. Carlin. 2010. "Multiple Imputation for Missing Data: Fully Conditional Specification Versus Multivariate Normal Imputation." *American Journal of Epidemiology* 171(5):624–32.

This provides a very useful idea for imputing large data sets with many variables:

http://crmda.dept.ku.edu/resources/kuantguides/11_ImputationWithLargeDataSets.pdf

Applied Guides (software-specific)

STATA
http://www.ats.ucla.edu/stat/stata/seminars/missing_data/mi_in_stata_pt1.htm
http://www.ats.ucla.edu/stat/stata/seminars/missing_data/mi_in_stata_pt2.htm

SAS
http://www.ats.ucla.edu/stat/sas/seminars/missing_data/part1.htm
http://www.ats.ucla.edu/stat/sas/seminars/missing_data/part2.htm

Practical Guides for different R packages:

Packages using multiple imputation with chained equations

Su, Yu-Sung, Andrew Gelman, Jennifer Hill, and Masanao Yajima. 2011. "Multiple Imputation with Diagnostics (mi) in R: Opening Windows into the Black Box." *Journal Of Statistical Software* 45(2).

Van Buuren, Stef, and Karin Groothuis-Oudshoorn. 2011. "mice: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software* 45(3).

Packages using multiple imputation that assumes multivariate normality of data

http://r.iq.harvard.edu/docs/amelia/amelia.pdf