The Stata Journal (2004)
4, Number 3, pp. 227–241

Multiple imputation of missing values

Patrick Royston
Cancer Division
MRC Clinical Trials Unit
222 Euston Road
London NW1 2DA
UK

Abstract. Following the seminal publications of Rubin about thirty years ago, statisticians have become increasingly aware of the inadequacy of "complete-case" analysis of datasets with missing observations. In medicine, for example, observations may be missing in a sporadic way for different covariates, and a complete-case analysis may omit as many as half of the available cases. Hotdeck imputation was implemented in Stata in 1999 by Mander and Clayton. However, this technique may perform poorly when many rows of data have at least one missing value. This article describes an implementation for Stata of the MICE method of multiple multivariate imputation described by van Buuren, Boshuizen, and Knook (1999). MICE stands for multivariate imputation by chained equations. The basic idea of data analysis with multiple imputation is to create a small number (e.g., 5–10) of copies of the data, each of which has the missing values suitably imputed, and analyze each complete dataset independently. Estimates of parameters of interest are averaged across the copies to give a single estimate. Standard errors are computed according to the "Rubin rules", devised to allow for the between- and within-imputation components of variation in the parameter estimates. This article describes five ado-files. `mvis` creates multiple multivariate imputations. `uvis` imputes missing values for a single variable as a function of several covariates, each with complete data. `micombine` fits a wide variety of regression models to a multiply imputed dataset, combining the estimates using Rubin's rules, and supports survival analysis models (`stcox` and `streg`), categorical data models, generalized linear models, and more. Finally, `misplit` and `mijoin` are utilities to interconvert datasets created by `mvis` and by the `miset` program from John Carlin and colleagues. The use of the routines is illustrated with an example of prognostic modeling in breast cancer.

Keywords: st0067, mvis, uvis, micombine, mijoin, misplit, missing data, missing at random, multiple imputation, multivariate imputation, regression modeling