

## **Missing Data Imputation Software: A sampling of available programs**

**Compiled by David R. Johnson: Department of Sociology**

### **SPSS: MVA: Missing Values Analysis (Missing Data Module)**

An optional module for SPSS. Is included in the site license at Penn State.

Estimates a variety of single imputation methods, although the only really useful one is an implementation of the EM algorithm.

The EM estimated variance and covariance matrices output by the program are unbiased, but the actual values imputed are biased because they do not include an error component. See von Hippel, Paul T. 2004. Biases in SPSS 12.0 Missing Values Analysis. *The American Statistician*. 58: 160-164.

### **Stata: ICE: Imputation with Chained Equations: (user supplied ado to Stata)**

This can be found by running typing `findit ice` in the Stata command window. The procedure can then be installed.

This is an upgrade to the Stata program MVIS which is an implementation for Stata of the program MICE (originally implemented in R). Provides single and multiple imputations using a model based approach based on chained regression equations. The user can chose a regression model based on the level of measurement of the variables (regression, logit, ordered logit, multinomial regression...). A supplied program MICOMBINE allows rolling up the estimates from multiple imputed data sets for many types of regression-based models.

See

Royston, Patrick. 2004. Multiple imputation of missing values. *The Stata Journal*. 4: 227-241.

Royston, Patrick. 2005. Multiple imputation of missing values:Update. *The Stata Journal*. 5: 1-14.

### **SAS: MI: Multiple Imputation**

An “experimental” procedure in SAS, but bundled with the SAS available at Penn State.

This is an quite complete (and can be complex) implementation of a variety of model based imputation procedures based on Rubin and Shaefer’s approaches which are generally based on a multivariate normal model. There are a number of options to regarding the specifics of the imputation process and the program can produce lots of diagnostics information to evaluate the adequacy of the imputations for the data. It can relatively easily with the program defaults.

Choice of the specific options for different starting values and imputation models can be challenging. Studies have shown that even though the distributions are assumed to be normal, that the imputations are quite robust with respect to violations of this assumption. Also includes the option to generate an EM single-imputed data set. This implementation differs from the SPSS EM in that it does properly include an error component in the imputed values.

See

Yuan, Jang C. Multiple imputation for missing data: Concepts and new development. SAS paper P267-25. SAS Institute

SAS OnlineDoc: Chapter 9: The MI Procedure. (available through the online documentation on the PRI website).

### **SAS: IVEware: Imputation and Variance Estimation Software.**

This is a free package from the University of Michigan ISR and can be downloaded over the internet ( <http://www.isr.umich.edu/src/smp/ive/> ). It is a program that runs in the SAS environment. The program can be used for both imputation and variance estimation in survey data (clustered, stratified, and weighted data). It has been used to generate single imputations (and multiple) for several large government studies (e.g., the most recent NSFG). It uses a “sequence of regression models” to yield the imputations. The regression models can differ depending on the level of measurement of the variable. Bounds can be imposed and cases can be identified in where imputations will not be created for a given variable.

See

Raghunathan, T. E., Lepkowski, J. W., Van Hoewyk, J., & Solenberger, P. 2001. A multivariate technique for multiply imputing missing values using a sequence of regression models.

Raghunathan, T. E. , Solenberger, P, & Van Hoewyk, J. 2002. IVEware: Imputation and Variance Estimation Software Users Guide. University of Michigan: Survey Research Center, Institute for Social Research.

### **NORM (and related programs) (We are.... Penn State!!!)**

This is a set of programs by Shaefer and his group at Penn State for model based multiple imputation of missing data. Many of the imputation models used are similar to those in PROC MI in SAS. The programs can be downloaded and are available at <http://www.stat.psu.edu/~jls/misoftwa.html> . They are written for use in S-Plus, but there is a version of NORM that can be run stand-alone in Windows. Shaefer has a readable and useful FAQ on missing data at <http://www.stat.psu.edu/~jls/mifaq.html> . John Graham has developed a set of utilities to facilitate the use of NORM by SAS and SPSS users <http://mcgee.hhdev.psu.edu/missing/index.html> .