

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Statistical Data Editing  
(Rome, Italy, 2-4 June 1999)

Topic (ii): Generalized software packages for statistical data editing, their evaluation

**MULTIPLE IMPUTATION FOR MISSING DATA USING THE  
“SOLAS FOR MISSING DATA ANALYSIS” SOFTWARE APPLICATION**

Submitted by Statistical Solutions Ltd, Cork, Ireland<sup>1</sup>

**Contributed paper**

**I. Introduction**

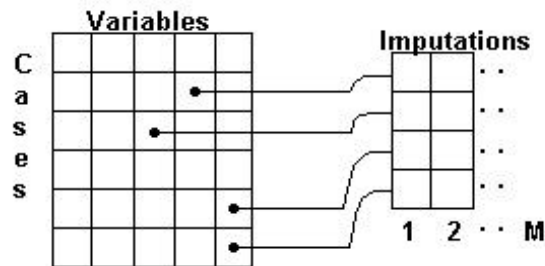
1. Analyses of multivariate data are frequently hampered by missing values. Until recently, the only missing-data methods available to most data analysts have been relatively adhoc practices such as list-wise deletion. These ad-hoc methods, though simple to implement, have serious drawbacks that have been well documented.
2. *Single Imputation* refers to any method whereby each missing value in a dataset is filled in with one value, yielding one complete dataset. The imputed dataset will fail to provide accurate measures of variability because subsequent analyses would fail to account for missing-data uncertainty.
3. Regardless of the imputation method, imputed values are only estimates of the unknown true values. *Any analysis which ignores the uncertainty of missing data prediction will lead to standard errors that are too small, p-values that are artificially low, and rates of Type I error that are higher than nominal levels.*
4. The Multiple Imputation approach, first developed by Dr. Donald B. Rubin in the 1970s, is designed to address the inherent weaknesses of single imputation methods. Multiple Imputation has been implemented by Statistical Solutions as part of the SOLAS for Missing Data Analysis software application. This presentation will focus on the use of Multiple Imputation in SOLAS as an imputation method of choice.

---

<sup>1</sup> Prepared by Aidan McDonnell.

## II. Multiple Imputation

5. Multiple Imputation is a technique that replaces each missing datum with a set of  $m > 1$  plausible values.



6. The  $m$  versions of the complete data are analysed by standard complete-data methods, and the results are combined using simple rules to yield estimates, standard errors, and p-values that formally incorporate missing data uncertainty. The variation among the  $m$  imputations reflects the uncertainty with which the missing values can be predicted from the observed ones.

7. Once MI's have been created, the datasets may be analysed by any method that would be appropriate if the data were complete. For example, one could perform linear or logistic regression procedures using any standard statistical package. Any model would have to be fitted  $m$  times, once for each imputed dataset, and the results across these datasets will vary as a reflection of missing data uncertainty.

8. **SOLAS for Missing Data Analysis** is the only commercially available software package that performs Multiple Imputation. SOLAS applies an implicit model approach based on Propensity Scores and an Approximate Bayesian Bootstrap to generate the imputations. The multiple imputations are independent repetitions from a Posterior Predictive Distribution for the missing data, given the observed data.

## III. Basic Steps in MULTIPLE IMPUTATION using Propensity Scores

9. For each time period/variable and each group:

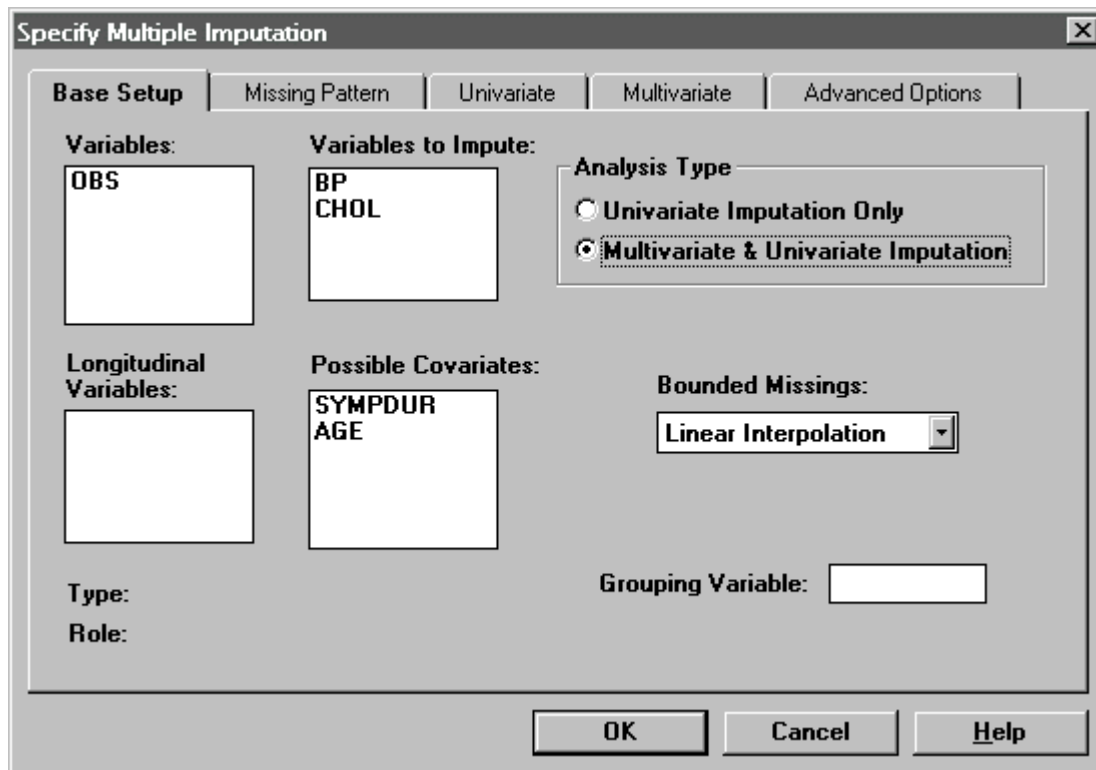
- Via logistic regression, model the missingness using only data that had been observed prior to the missing value.
- Based on results of the logistic regression, calculate the propensity that a subject would have a missing value at that period.
- Group subjects based on quintiles of the propensity score.

Within each quintile:

- (i) From the observed data in this quintile, create a Posterior Predictive Distribution (PPD) of observed data by taking a random sample WR equivalent in size to the number of observed data. (ii) From the PPD, randomly sample WR to choose an imputed value for each missing value.
- Repeat the entire procedure to create M complete datasets.

10. SOLAS imputes variables from left to right through the dataset, so that values which are imputed for one variable can be used in the prediction model for missing values occurring in variables to the right of it.

11. To set up a Multiple Imputation in SOLAS is very easy. The Base Setup dialog box is where you define which variables you want to impute, and which variables you want to use as possible covariates for the logistic regression.



12. If you want to run an imputation using all of the system defaults, then you can select OK at this point. However, if you want to make some changes to the logistic regressions that are performed, you can do so in the **Univariate**, **Multivariate** and **Advanced Options** tabs.

13. The **Univariate** tab affords the user complete control over the logistic regression for all univariately missing values. (A value is considered univariately missing if it is missing in a certain period of one longitudinal variable, but observed in the same period of the other longitudinal variable). Variables can be added to or removed from the covariate pool, and terms can be forced into the regression model. If the logistic regression does not converge, you have the option to select a variable, the values of which will be used as a propensity score.

14. Similarly, in the **Multivariate** tab, the user can customise the regressions for all of the multivariately missing values. (A value is considered multivariately missing if it is missing for a certain period in one longitudinal variable, and is also missing in that same period of the other longitudinal variable.)

15. The **Advanced Options** tab allows you to further control the logistic regressions. Here you can set parameters such as the model tolerance, the convergence criterion and the number of iterations to convergence.

16. Once the multiple imputation has run, the imputed datapages appear with the imputed values appearing in blue. The default for the number of imputations is 5, but this can be set to between 2 and 10 imputations. Each of these datapages can be saved for later analysis or exported to any other statistics package.

	MeasA_0	MeasA_1	MeasA_2	MeasA_3
1	177	174	111	57
2	165	150	78	33
3	270	240	255	261
4	276	276	297	291
5	306	294	297	285
6	198	228	162	150
7	147	186	177	381
8	321	321	336	318
9	213	213	201	270
10	276	216	252	273

#### IV. Other features of the SOLAS system

17. In addition to Multiple Imputation, SOLAS allows the following imputation approaches:

- 'Hot-Decking'
- Group Means Imputation
- Last Value Carried Forward

#### V. Future Releases

18. Version 2.0 of SOLAS will include several enhancements to the propensity score approach. One of these, for example is that, rather than automatically dividing the dataset into quintiles based on propensity, the user will have the option to specify the number of subgroups as a function of the number of cases in the dataset. It is also planned to add a regression or model-based approach for creating multiple imputations with a monotone missing data pattern.

#### References

1. Rubin, D. Multiple Imputation for Nonresponse in Surveys, Wiley, New York, 1987.
  2. Rubin, D. and Schenker, N. "Multiple imputation in health-care data bases: an overview and some applications", *Statistics in Medicine*, 10, 585-598 (1991).
  3. Lavori, P. , Dawson, R. and Shera, D. "A multiple imputation strategy for clinical trials with truncation of patient data", *Statistics in Medicine*, 14, 1913-1925 (1995).
  4. Rubin, D. "Multiple imputation after 18+ years", *Journal of the American Statistical Association*, 91, 473-489 (1996).
- Reference 1 provided the theoretical background for multiple imputation.
  - References 2 and 3 served as our primary references.
  - Reference 4 is a good overview of multiple imputation and has a very thorough list of literature references.