

A MULTIPLE IMPUTATION STRATEGY FOR MISSING DATA IN LONGITUDINAL STUDIES

Fiona O'Callaghan *

Abstract

This article describes a strategy to deal with missing data using the SOLAS for Missing Data Analysis software package to perform multiple imputation. Short examples of some single imputation techniques are presented in order to illustrate the disadvantages of these methods, and the concept of multiple imputation is introduced. Finally, an example outlining how multiple imputation is applied in SOLAS is described.

Keywords: Missing-data uncertainty, Imputation, Multiple Imputation, Propensity score, Covariate.

Missing data is a problem that confronts every data analyst. Missing values lead to less efficient estimates because of the reduced size of the database, and standard complete-data methods of analysis no longer apply. For example, analyses such as multiple regression use only cases that have complete data, so including a variable with numerous missing values would severely reduce the sample size and potentially result in biased estimates depending on the missing data mechanism.

Until recently, the only missing-data methods available to most data analysts have been relatively ad-hoc practices such as listwise deletion. These ad-hoc methods, though simple to implement, have serious drawbacks, which have been well documented.

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

* Customer Support Manager, Statistical Solutions Ltd., 8 South Bank, Crosse's Green, Cork, Ireland. eMail: support@statsol.ie

for missing-data uncertainty. 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.

Single Imputation - Means

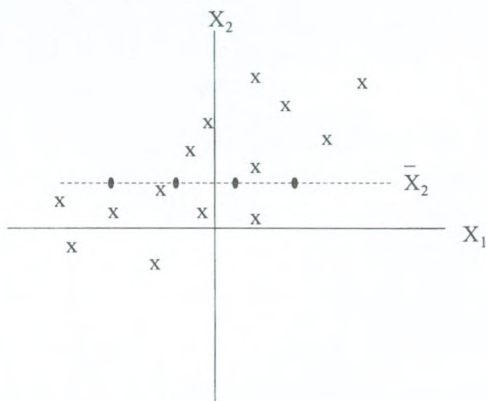


Figure 1. Mean Imputation

As Figure 1 shows, imputing the mean will systematically underestimate the variability of the resulting imputed distribution because there will be no residual variance. Also, any correlation that exists between the variables is not being kept.

Single Imputation - Regression

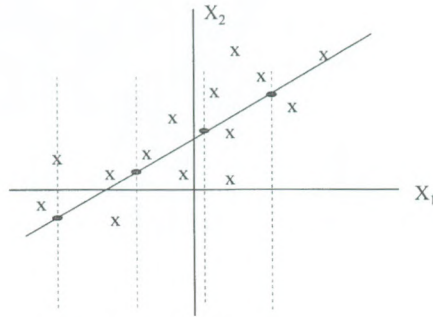


Figure 2. Regression Imputation

Similarly, *Figure 2* shows that when the imputed values all lie on the best fit regression line, the residuals for all of these values will be zero.

An Example of Bias Introduced by Last Value Carried Forward (LVCF)

LVCF is a technique that is frequently used in the pharmaceutical industry. The last observed value for a patient is used to fill in missing values at a later point in the study.

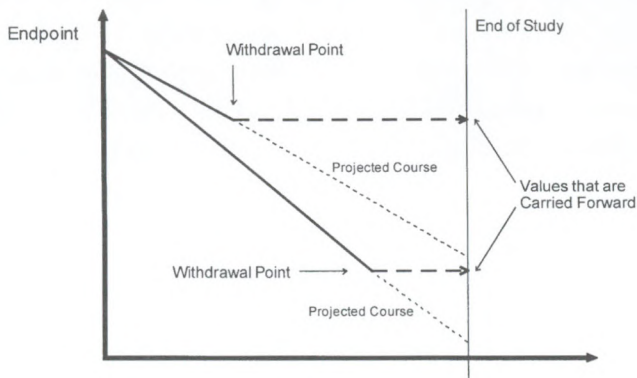


Figure 3. Example of Last Value Carried Forward

An obvious example where LVCF will result in biased results would be in the case of degenerative diseases. Using the last observed value to impute for missing data at a later time point in the study means that a high observation will be carried forward, resulting in an overestimation of the true end-of-study measurement.

Multiple Imputation

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

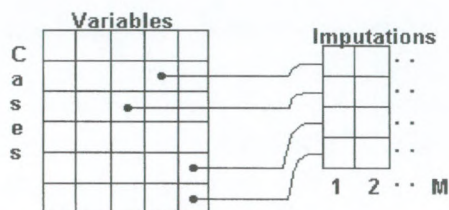


Figure 4. Multiple imputation

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.

Once Multiple Imputations (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.

An overall set of estimated coefficients and standard errors can be obtained by combining the results using the rules given by Rubin (1987).

Rubins rules are as follows:

For each of the M complete datasets, let, $\hat{\theta}_m, m = 1, \dots, M$, be M complete-data estimates for a parameter θ , and $U_m, m = 1, \dots, M$, be their associated variances. The MI estimate, or overall estimate, is given by

$$\bar{\theta} = \sum \hat{\theta}_m / M$$

The total variance for the estimate has two components that take into account variability within each dataset and across datasets. The within-imputation variance is

$$\bar{U} = \sum U_m / M$$

The between imputation variance is

$$B = \sum (\hat{\theta}_m - \bar{\theta})^2 / (M - 1)$$

The total variance, T , is the sum of the two components with an additional correction factor to account for the simulation error is

$$T = \bar{U} + (1 - M^{-1})B$$

Basic Steps in Multiple Imputation using Propensity Scores

The aim of the Multiple Imputation procedure in SOLAS is to create m multiply imputed data sets in such a way that they can be combined to produce valid inferences. The propensity score method, which SOLAS uses, is a nonparametric method, in the sense that you do not have to make explicit assumptions about the model that your data follow.*

* SOLAS 2.0 will also have a parametric (regression) model for multiple imputation.

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

As an illustration of the two random samplings with replacement, say that in quintile number 3 there are 10 cases in total – 7 cases have observed values and 3 cases have missing values. *Figure 5* below shows how the sampling is performed.

* SOLAS 2.0 allows the user to set the number of imputation subgroups, and to match on covariates in addition to the propensity score.

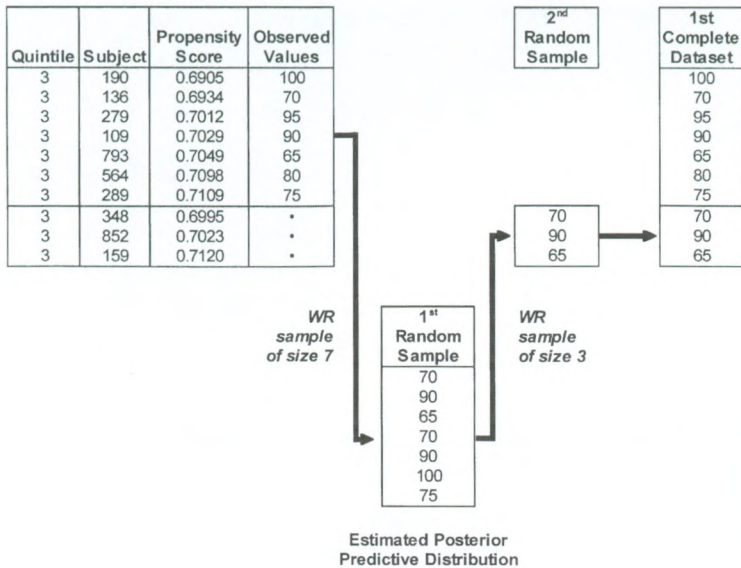


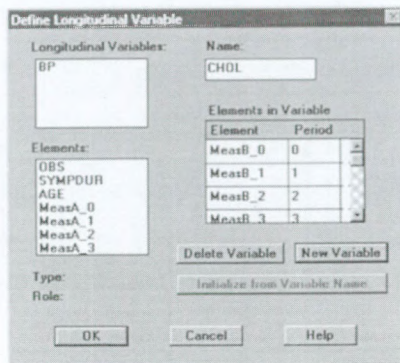
Figure 5. An illustration of the 2 random samplings with replacement

SOLAS for Missing Data Analysis

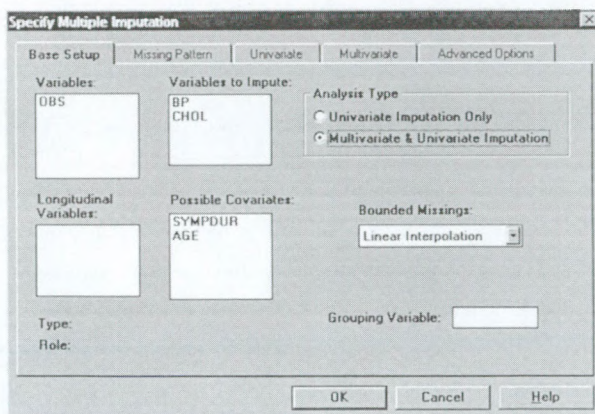
SOLAS for MDA is the only commercially available software package that performs Multiple Imputation. SOLAS applies an implicit model approach based on Propensity Scores and an Approximate Bayesian Bootstrap to generate the imputations. The multiple imputations are independent repetitions from a Posterior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

The data that we will use for this example is from a clinical trial. Two longitudinal variables (blood pressure and cholesterol) were measured for each of 50 patients. Repeated measures were taken at monthly intervals for 4 months.

SOLAS has a feature which allows the user define a longitudinal variable which is made up of a set of repeated measures. So, for example, the longitudinal variable CHOL is made up of the four repeated measures variables MeasB_0, MeasB1, ..., MeasB_3.

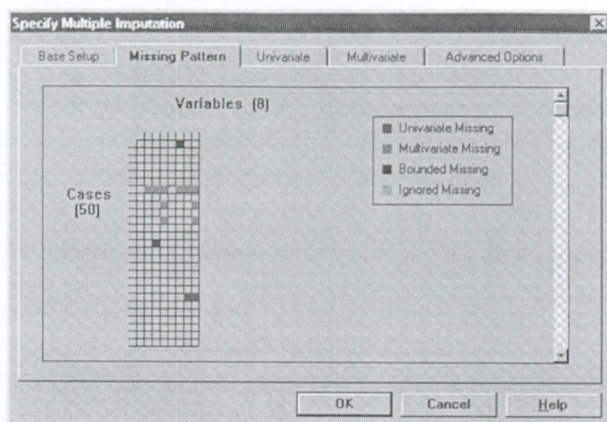


To set up a Multiple Imputation in SOLAS is very easy. The Base Setup dialog box is where you define which variables you want to impute, and which variables you want to use as possible covariates for the logistic regression that is used to predict the response propensity.



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.

Once you have selected which variables you want to impute, you can view the missing pattern in your data and identify which values are bounded missings, which are univariate missings and which are multivariate missings in the *Missing Pattern* tab.



The *Univariate* tab affords the user complete control over the logistic regression for all univariately missing values. (For the purposes of this example, 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.

Specify Multiple Imputation

Base Setup | Missing Pattern | **Univariate** | Multivariate | Advanced Options

Variables:

- OBS
- SYMPDUR
- AGE
- MeasA_0
- MeasA_1
- MeasA_2
- MeasA_3
- MeasB_0
- MeasB_1
- MeasB_2

Variable Name	N	Covariates	Forced	Non-model Propensity
- CHOL	3			
MeasB_2				
		MeasB_0	<input checked="" type="checkbox"/>	<input type="checkbox"/>
		MeasB_1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
		SYMPDUR	<input type="checkbox"/>	<input type="checkbox"/>
		AGE	<input type="checkbox"/>	<input type="checkbox"/>
MeasB_3				
		MeasB_0	<input checked="" type="checkbox"/>	<input type="checkbox"/>

Click on the +/- sign in front of a variable name to expand/contract the list of covariates associated with it.

OK Cancel Help

Similarly, in the *Multivariate* tab, the user can customise the regressions for all of the multivariately missing values. (For the purposes of this example, 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.)

Specify Multiple Imputation

Base Setup | Missing Pattern | Univariate | **Multivariate** | Advanced Options

Variables:

- OBS
- SYMPDUR
- AGE
- MeasA_0
- MeasA_1
- MeasA_2
- MeasA_3
- MeasB_0
- MeasB_1
- MeasB_2

Variable Name	N	Covariates	Forced	Non-model Propensity
+ Period__	3			
Period__1				
		Period_0	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
		SYMPDUR	<input type="checkbox"/>	<input type="checkbox"/>
		AGE	<input type="checkbox"/>	<input type="checkbox"/>
Period__2				
		Period_0	<input checked="" type="checkbox"/>	<input type="checkbox"/>
		Period_1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

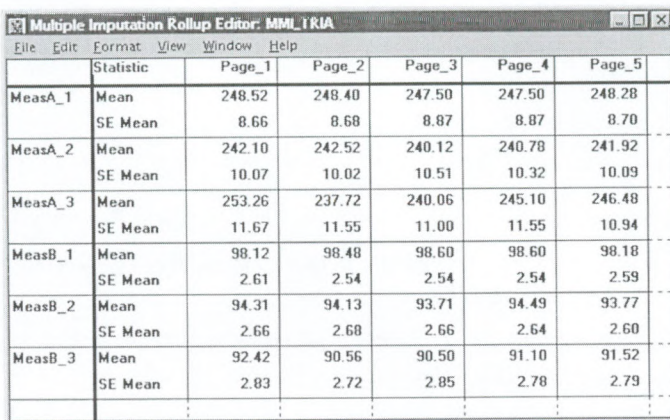
Click on the +/- sign in front of a variable name to expand/contract the list of covariates associated with it.

OK Cancel Help

The *Advanced Options* tab contains controls related to the logistic regressions. Here you can set parameters such as the model tolerance, the convergence criterion and the number of iterations to convergence.

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 stats package.

The multiple imputation Rollup Editor is a table which gives the mean and the standard error of the mean for every imputed variable for each of the 5 multiple imputations.



	Statistic	Page_1	Page_2	Page_3	Page_4	Page_5
MeasA_1	Mean	248.52	248.40	247.50	247.50	248.28
	SE Mean	8.66	8.68	8.87	8.87	8.70
MeasA_2	Mean	242.10	242.52	240.12	240.78	241.92
	SE Mean	10.07	10.02	10.51	10.32	10.09
MeasA_3	Mean	253.26	237.72	240.06	245.10	246.48
	SE Mean	11.67	11.55	11.00	11.55	10.94
MeasB_1	Mean	98.12	98.48	98.60	98.60	98.18
	SE Mean	2.61	2.54	2.54	2.54	2.59
MeasB_2	Mean	94.31	94.13	93.71	94.49	93.77
	SE Mean	2.66	2.68	2.66	2.64	2.60
MeasB_3	Mean	92.42	90.56	90.50	91.10	91.52
	SE Mean	2.83	2.72	2.85	2.78	2.79

SOLAS also provides a Rollup feature which automatically combines these means and standard errors from each of the imputed datasets into one repeated imputation inference, using the rules described earlier.*

* The SOLAS 2.0 Rollup editor will automatically combine and present the results of any analysis performed on multiply imputed datasets.

Multiple Imputation Rollup Statistics : MMI_TRIA					
File Edit View Format Window Help					
Courier New 10 B I U					
Statistic		Mean	Within_Var	Bet_Var	Total_Var
MeasA_1	Mean	248.0400	76.66163	0.2501998	76.96187
MeasA_2	Mean	241.4880	104.1055	0.9997274	105.3051
MeasA_3	Mean	244.5240	128.6684	36.67425	172.6775
MeasB_1	Mean	98.39600	6.566869	0.05327942	6.630804
MeasB_2	Mean	94.08200	7.018794	0.1141199	7.155738
MeasB_3	Mean	91.22000	7.797900	0.6245993	8.547419
Bet_Ratio		D.F.		Gamma	
MeasA_1	0.00391642	262830.0	0.00390114		
MeasA_2	0.01152363	30820.06	0.01139235		
MeasA_3	0.3420352	61.58093	0.2548630		
MeasB_1	0.00973604	43024.01	0.00964216		
MeasB_2	0.01951103	10921.53	0.01913763		
MeasB_3	0.09611808	520.1930	0.08768953		
NUM INS Col: 0 Line: 19 Page: 1					

References

- Rubin, D. *Multiple Imputation for Nonresponse in Surveys*, Wiley, New York, (1987)
- Rubin, D. and Schenker, N. "Multiple imputation in health-care data bases: an overview and some applications", *Statistics in Medicine*, 10, 585-598 (1991)
- 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)
- 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

Ontvangen: 05-01-1999

Geaccepteerd: 23-07-1999