

Multivariate Imputation by Chained Equations

Stef van Buuren and Karin Oudshoorn
 TNO Prevention and Health, Leiden, The Netherlands
 S.vanBuuren@pg.tno.nl
 K.Oudshoorn@pg.tno.nl



TNO Prevention and Health

Practical problems in multivariate imputation

- Large numbers of predictor variables
- Predictors themselves can be incomplete
- Circularities occur, where Y_1 is imputed given Y_2 , and Y_2 given Y_1
- Order of imputation can be meaningful
- Transformed versions of imputed data might be needed
- Mixed measurement levels
- Nonlinear imputation models and interaction terms
- Survey weights
- Hierarchical data structure
- Uncongenial imputation and complete-data models

MICE

Multivariate Imputation by Chained Equations (MICE) is a flexible and general methodology for generating multiple imputations in multivariate data. S-Plus software is available that assists in performing the steps required in a full multiple imputation analysis: generation of imputations, repeated analyses on the imputed data, and pooling of the results. Figure 1 provides a graphic representation of these steps.

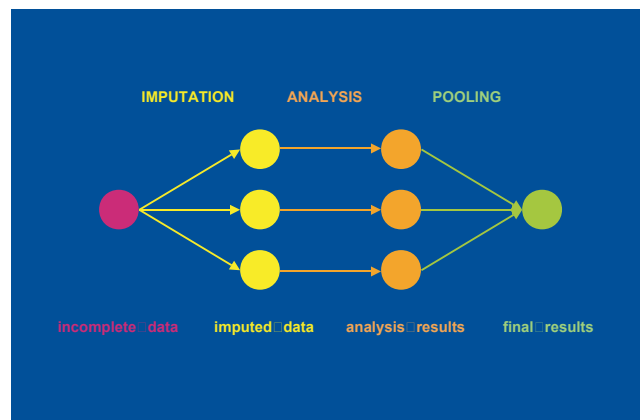


Figure 1

Specific features of the software are

- columnwise specification of the imputation model
- arbitrary patterns of missing data
- transformations and index variables
- subset selection of predictors
- supports standard lm and glm complete-data methods
- automated pooling using the BarnardRubin adjustment
- callable user-written imputation functions
- online help files

MICE imputation method

For each missing variable, a conditional distribution for the missing data given the other data can be specified. The method consists of iterating over these conditional densities by means of Gibbs

sampling. Table 1 contains the elementary imputation method currently available.

Does it work? How many iterations?

It is hard to establish convergence in the general case, but simulation studies suggest that the coverage properties in some important practical cases are quite good. Brand (1999) performed simulations using just 5 iterations. See Table 2.

Points of interest

- Variables Y_1 - Y_4 are of mixed type (Y_5 - Y_7 are not shown)
- Missing data seriously affect the means, proportions and correlations
- Imputation 'restores' the appropriate population values
- Coverage percentages are generally close to the nominal value (95)

Conclusion

MICE consists of a set of flexible tools for creating multiple imputations and for the analysis of multiply imputed data sets. The Gibbs sampling approach provides a solution for most problems mentioned in the introduction. Empirical evidence suggests that the imputations created by MICE are proper with even as low as five main iteration steps.

REFERENCES

- Brand JPL (1999). Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets. Thesis, Erasmus University Rotterdam/TNO Prevention and Health, Leiden.
- Van Buuren S, Boshuizen HC & Knook DL (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, 18, 681-694.
- Van Buuren S and Oudshoorn CGM (2000). *Multivariate Imputation by Chained Equations: MICE V1.0 User's Manual*. TNO Report PG/VGZ/00.038.

Availability

MICE works under S-Plus 4.5 and S-Plus 2000. The software and the documentation can be downloaded from

www.multiple-imputation.com.

Table 1 Built-in elementary imputation methods in MICE..

| Method | Description | Type of target |
|-----------------|---|----------------|
| impute.norm | Bayesian linear regression | Numeric |
| impute.pmm | Predictive mean matching | Numeric |
| impute.mean | Unconditional mean imputation | Numeric |
| impute.logreg | Logistic regression | 2 categories |
| impute.logreg2 | Logistic regression (direct minimization) | 2 categories |
| impute.polyreg | Polytomous regression (generalized logit) | >= 2 cat(nom) |
| impute.llda | Linear discriminant analysis | >= 2 cat(nom) |
| impute.propodds | Proportional odds model | >= 2 cat (ord) |
| impute.sample | Random sample from observed values | Any |

Table 2 Simulation results. Data 600 observations, 7 variables Y_1 - Y_7 (4 are incomplete with 31% missing entries each), 500 replications, 5 Gibbs sampling iterations, 10 multiple imputations. MAR missing data mechanism, non-monotone pattern. Source: Table 5.12 of J.P.L. Brand, (1999)

| | type | statistic | population | complete cases only | multiple imputation (MICE) | 95% coverage |
|------------|-------------|---------------|------------|---------------------|----------------------------|--------------|
| Y_1 | numeric | mean | 12.08 | 11.33 | 12.09 | 92.6 |
| Y_2 | binary | $p(Y_2=0)$ | 0.48 | 0.55 | 0.47 | 96.0 |
| | | $p(Y_2=1)$ | 0.52 | 0.45 | 0.53 | 96.0 |
| Y_3 | categorical | $p(Y_3=0)$ | 0.34 | 0.41 | 0.34 | 95.2 |
| | | $p(Y_3=1)$ | 0.35 | 0.33 | 0.35 | 97.8 |
| | | $p(Y_3=2)$ | 0.32 | 0.26 | 0.31 | 98.0 |
| Y_4 | numerical | mean | 9.67 | 9.01 | 9.67 | 96.2 |
| Y_1, Y_4 | numerical | $r(Y_1, Y_4)$ | 0.56 | 0.21 | 0.56 | 97.8 |