Modelling Uncertainty in the Risk of Intensive Care Unit Readmission II: Comparison of Imputation Methods

Ben Cooper

March 31, 2021

1 Introduction

2 Methods

2.1 Data source and processing

2.2 Imputation methods

2.2.1 *k*-nearest neighbours (KNN)

2.2.2 Principal components analysis (PCA)

2.2.3 Random Forest (RF)

2.2.4 Amelia

The 'Amelia' R package (Honaker *et al.*, 2011) is an implementation and extension of the 'Expectation Maximisation' (EM) algorithm developed by Dempster *et al.* (1977). *Amelia* runs the EM algorithm on multiple bootstrapped samples from the complete cases to estimate the parameters of the multivariate normal distribution that the data is assumed to follow. These parameters are then used to draw imputed values for the missing cases. Unlike the imputation methods described thus far, *Amelia* performs multiple imputation, i.e., it generates multiple imputed datasets using the same process. The authors recommend carrying out analysis on each of these datasets separately, then combining the results.

2.2.5 Multiple imputation by chained equations (MICE)

Multiple imputation by chained equations (MICE) uses an iterative process to impute datasets, and is implemented in R package *mice* (**?**). The algorithm has the following steps:

1. Missing data for each variable are initially imputed as the mean of that variable.
2. For a dataset of N variables, N models are generated which predict a given variable using all other variables.
3. These models are used to predict the initially missing values for each variable, and these predictions replace the initial mean imputations.
4. Steps 2 and 3 are then repeated until convergence criteria are met, commonly that the average of the imputed values does not change

Like *amelia*, MICE is run as a multiple imputation method, generating m final imputed datasets. Whilst all implementations follow the steps laid out above, the algorithm is flexible as to the type of predictive models used. As implied by the algorithm, MICE assumes a correlation structure among the full data.

2.2.6 Additional methods

2.3 Hyperparameter profiling

2.4 Comparison

3 Results

References

Dempster, A. P., Laird, N. M., & Rubin, D. B. 1977. Maximum likelihood from incomplete data via theEMAlgorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, **39**(1), 1–22.

Honaker, J., King, G., & Blackwell, M. 2011. AmeliaII: A program for missing data. *Journal of Statistical Software*, **45**(7).