Multiple imputation of missing data by MICE: a comparison of Random Forest and parametric imputation models

Anoop Shah, Clinical Epidemiology Group, University College London

5th February 2013

Contents

1 Multivariate normal model for continuous variables

We simulated multivariate normal datasets with 2000 observations of three predictor variables (x_1, x_2, x_3), one dependent variable (y) and an auxiliary variable (x_4) which was associated with the other variables but was not part of the analysis of interest. The x variables had the variance-covariance matrix:

$$\begin{pmatrix} 1 & 0.2 & 0.1 & -0.7 \\ 0.2 & 1 & 0.3 & 0.1 \\ 0.1 & 0.3 & 1 & 0.2 \\ -0.7 & 0.1 & 0.2 & 1 \end{pmatrix}$$

The data generating model for y was:

$$y = x_1 + x_2 + x_3 + e \text{ where } e \sim N(0,1)$$

The analysis of interest was $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ where the true value of β_0 was 0 and the true value of the other β coefficients was 1.

Some values of variables x_1 and x_2 were made missing according to the following mechanisms:

1. Missing Completely at Random (MCAR), where a missingness indicator was generated for each value of x_1 and x_2 as a random draw from a Bernoulli distribution with probability 0.2.

2. Missing at Random (MAR), in which missingness was dependent on the outcome (y) and x_3, both of which were fully observed. The probability that x_1 was missing was the logistic of $x_3 + y + c$, where c was a constant chosen such that the overall proportion of x_1 missing was 0.2. Missingness was independently introduced into x_2 using the same model and with the same probability as for x_1.

MICE (Multivariate Imputation by Chained Equations) (1) was used to impute missing values of x_1 and x_2, with x_3, y and the auxiliary variable x_4 in the imputation model. Imputations were drawn after 10 iterations, and 10 imputed datasets were created. Coefficients of the linear regression of y on x_1, x_2, and x_3 were estimated for each imputed dataset and the estimates combined using Rubin's rules. We compared MICE with normal-based Bayesian linear regression to MICE with the following novel Random Forest-based imputation functions:

1. Fit a Random Forest with 1 tree to the predictors in the imputation model and the observed values of variable to be imputed. The tree is fitted to a bootstrap sample of the data (sampling with replacement, with the sample size equal to the total number of observations). The observations not included in the bootstrap sample ('out-of-bag' observations) are used to estimate the mean square prediction error (2). Each missing value is imputed as a random draw from a normal distribution with the mean defined by the prediction of the tree and variance estimated by the out of bag mean square error.

2. Fit a Random Forest with 10 or 100 trees to a bootstrap sample of the data. Impute each missing value as a random draw from a normal distribution with the mean defined by the prediction of the Random Forest and variance estimated by the out of bag mean square error.

3. Fit a Random Forest as above, but apply the experimental bias correction ('biascorr') option in Random Forest before drawing imputed values. This attempts to correct for the bias in Random Forest predictions; predictions for extreme values are on average based on values of the outcome variable closer to the center of the data and are biased away from the extremes (3). Hence predictions for high values are underestimates and predictions for low values are overestimates. The bias correction option uses the Random Forest to predict the outcome values ($\hat{z}$) and replaces the original z values with the predictions from a linear regression of z on $\hat{z}$.

Table 1: Analysis of simulated multivariate normal datasets with predictor variables Missing Completely at Random (MCAR) using MICE and normal-based linear regression or Random Forest imputation methods

Imputation Method	Coefficient	Number of Trees	Estimate: Mean	Estimate: SD	Bias of Estimate: Mean	Bias of Estimate: 95% CI	95% Confidence Interval: Length	95% Confidence Interval: Coverage, %
Full data	β_1	N/A	1.001	0.023	0	-0.001, 0.001	0.090	94.7
Full data	β_2		1.000	0.024	-0.001	-0.002, 0.000	0.094	94.9
MICE normal	β_1	N/A	1.001	0.026	0	-0.001, 0.001	0.101	95.0
MICE normal	β_2		1.000	0.026	-0.001	-0.002, 0.000	0.104	95.5
MICE RF	β_1	1	0.893	0.025	-0.107	-0.108, -0.106	0.142	8.9
MICE RF	β_2	1	0.809	0.026	-0.192	-0.193, -0.191	0.173	0.1
MICE RF	β_1	10	1.001	0.026	0.000	-0.001, 0.001	0.107	96.2
MICE RF	β_2	10	0.985	0.026	-0.016	-0.017, -0.014	0.113	93.9
MICE RF biascorr	β_1	10	1.003	0.026	0.003	0.002, 0.004	0.107	96.4
MICE RF biascorr	β_2	10	0.986	0.026	-0.015	-0.016, -0.014	0.114	94.3
MICE RF	β_1	100	1.018	0.026	0.017	0.016, 0.018	0.101	88.3
MICE RF	β_2	100	1.017	0.026	0.017	0.015, 0.018	0.107	90.8
MICE RF biascorr	β_1	100	1.008	0.026	0.007	0.006, 0.008	0.100	93.7
MICE RF biascorr	β_2	100	1.016	0.026	0.016	0.015, 0.017	0.104	91.0
MICE RF choose	β_1	10	0.985	0.026	-0.016	-0.017, -0.015	0.111	93.8
MICE RF choose	β_2	10	0.966	0.025	-0.034	-0.035, -0.033	0.119	84.5

4. Fit 10 Random Forests each with 1 tree to a bootstrap sample of the data. Choose a tree at random for each missing value, and impute the value as the prediction of the randomly chosen tree. This method is denoted 'choose', and does not use the normal distribution at all. It is also possible to use it for categorical variables.

We simulated and analysed 2000 datasets using these methods. We saved the random seeds to enable the analysis to be repeated if necessary. The estimates of β_1 and β_2 were compared between methods, assuming that the empirical mean from the full data analyses was the 'correct' result.

Results

As expected, the full data analyses and parametric MICE produced unbiased parameter estimates with correct coverage of confidence intervals, both under MCAR (Table 1) and MAR (Table 2). All the Random Forest methods produced biased parameter estimates, but the bias was less than 2% with 10

Table 2: Analysis of simulated multivariate normal datasets with predictor variables Missing at Random (MAR) using MICE and normal-based linear regression or Random Forest imputation methods

Imputation Method	Coefficient	Number of Trees	Estimate: Mean	Estimate: SD	Bias of Estimate: Mean	Bias of Estimate: 95% CI	95% Confidence Interval: Length	95% Confidence Interval: Coverage, %
Full data	β_1	N/A	1.001	0.023	0.000	-0.001, 0.001	0.090	94.7
Full data	β_2		1.000	0.024	-0.001	-0.002, 0.000	0.094	94.9
MICE normal	β_1	N/A	1.000	0.026	0.000	-0.001, 0.001	0.102	94.8
MICE normal	β_2		0.999	0.026	-0.002	-0.003, 0.000	0.105	95.8
MICE RF	β_1	1	0.888	0.029	-0.112	-0.114, -0.111	0.202	38.4
MICE RF	β_2	1	0.783	0.031	-0.218	-0.219, -0.216	0.253	1.3
MICE RF	β_1	10	1.016	0.029	0.016	0.014, 0.017	0.131	94.9
MICE RF	β_2	10	0.972	0.029	-0.028	-0.03, -0.027	0.140	92.1
MICE RF biascorr	β_1	10	1.021	0.029	0.020	0.019, 0.021	0.131	93.1
MICE RF biascorr	β_2	10	0.971	0.030	-0.030	-0.031, -0.029	0.142	92.3
MICE RF	β_1	100	1.037	0.029	0.036	0.035, 0.038	0.117	77.4
MICE RF	β_2	100	1.007	0.03	0.006	0.005, 0.008	0.125	96.2
MICE RF biascorr	β_1	100	1.025	0.029	0.024	0.023, 0.025	0.117	86.8
MICE RF biascorr	β_2	100	1.01	0.029	0.009	0.008, 0.011	0.121	95.3
MICE RF choose	β_1	10	0.986	0.029	-0.015	-0.016, -0.013	0.142	97.3
MICE RF choose	β_2	10	0.938	0.029	-0.062	-0.064, -0.061	0.156	70.4

or 100 trees under MCAR, or with 10 trees under MAR. With 10 trees, coverage of nominal 95% confidence intervals was around 92–94% for both β_1 and β_2 under MAR, and 94–96% under MCAR. Under MCAR and MAR, Random Forest with 1 tree produced estimates which were biased towards the null, with very wide confidence intervals but low coverage. The performance of Random Forest with 100 trees was poor for β_1 (x_1 was strongly negatively correlated with the completely observed auxiliary variable, x_4) but good for β_2 (x_2 was poorly correlated with other variables). The bias correction option in Random Forest did not consistently reduce the bias or improve the coverage of confidence intervals. Random Forest with 100 trees produced narrower confidence intervals than with 10 trees but coverage was worse.

With the 'choose' method (choosing an imputed value from 10 single trees), estimates for β_1 were satisfactory but estimates of β_2 were biased towards the null, with coverage of 95% confidence intervals only 85% (under MCAR) or 70% (under MAR).

2 Mixed linear and binary variables

We simulated datasets as in section 1, but replaced the values of x_2 by random draws from a Bernoulli distribution with probability equal to the logistic of x_2. We made x_1 and x_2 partially missing according to the same missing at random mechanism as in section 1.

We analysed this dataset using MICE with linear and logistic regression, or MICE with the 'choose' Random Forest method (choosing a prediction of a random tree) for x_2, the binary variable, and the same methods as in section 1 for x_1.

Results

The full data analyses and parametric MICE produced unbiased parameter estimates with correct coverage of confidence intervals, both under MCAR (Table 3) and MAR (Table 4). Imputation of the binary variable x_2 by Random Forest resulted in estimates β_2 with 3–5% bias under MCAR and 3–9% bias under MAR, and coverage of 95% confidence intervals was around 90%. Random Forest with 10 trees performed moderately well in estimating β_1 and β_2, with 3–6% bias of estimates and 89–90% coverage of 95% confidence intervals under MAR. Random Forest with 100 trees performed better in estimating β_1 but worse on β_2.

3 Discussion

Parametric MICE had better performance than any of the Random Forest methods used on the simulated data. This is unsurprising as the data were drawn from a normal distribution and the parametric model is completely correct. Random Forest does not assume that linear relations hold, and this uncertainty is manifest in the form of less efficient parameter estimates and wider confidence intervals.

Increasing the number of trees in a Random Forest prediction model should increase the precision of estimates (4). We found that confidence intervals were narrower when using 100 trees for imputation, some the parameter estimates were frequently biased, and the results were worse than Random Forest with 10 trees. A possible reason for this might be that the prediction bias of Random Forest is not improved by increasing the number of trees, as shown in figure 1. Predictions for extreme values are on average based on values of the outcome variable closer to the center of the data and are biased away from the extremes (3), and increasing the number of trees does not reduce this bias but does reduce the out of bag mean square error. In our MAR model, higher values of x_1 and x_2 had a greater probability of being missing, because missingness depended on the outcome y and both x_1 and x_2

Table 3: Analysis of simulated datasets with continuous and binary predictor variables Missing Completely at Random (MCAR) using MICE and parametric or Random Forest imputation methods

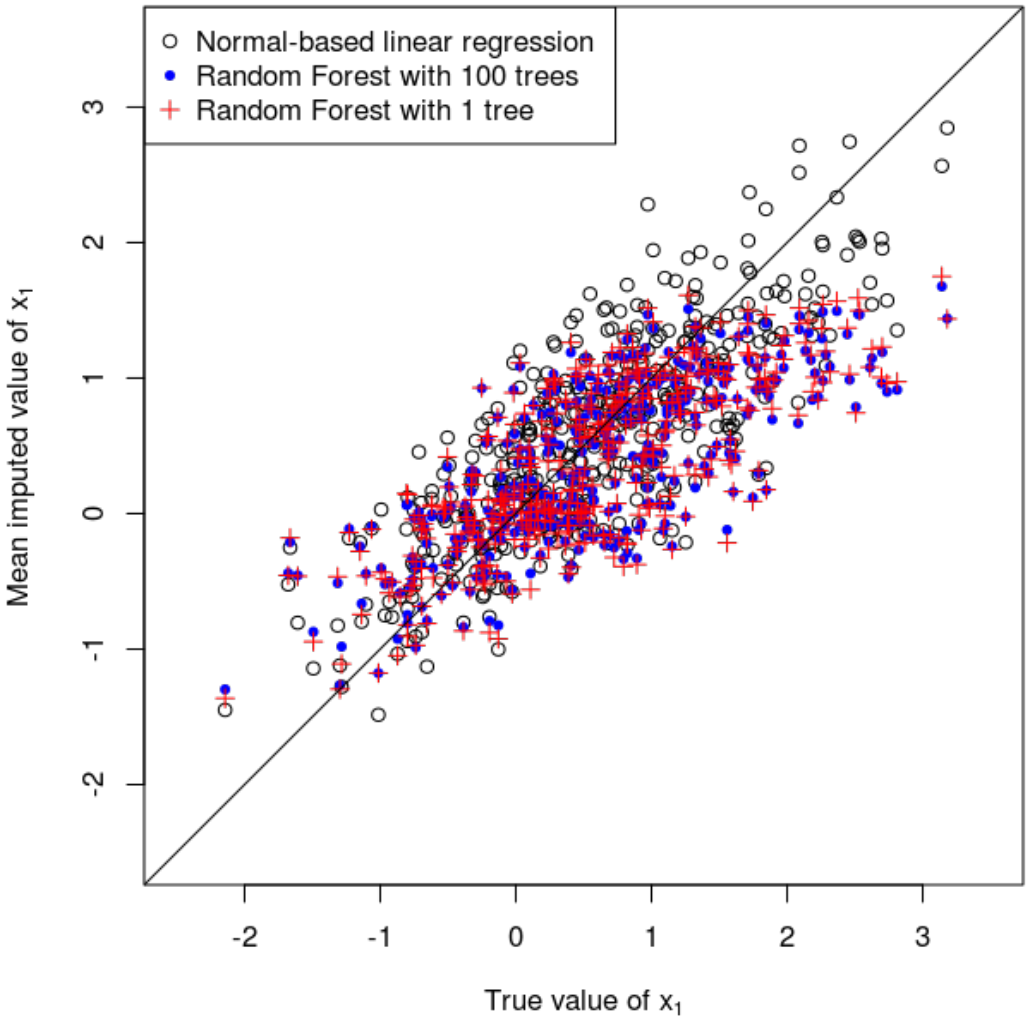
Imputation Method	Coefficient	Number of Trees	Estimate: Mean	Estimate: SD	Bias of Estimate: Mean	Bias of Estimate: 95% CI	95% Confidence Interval: Length	95% Confidence Interval: Coverage, %
Full data	β_1	N/A	1.000	0.024	0.000	-0.001, 0.001	0.088	94.2
Full data	β_2		0.999	0.046	0.000	-0.002, 0.002	0.177	95.0
MICE normal	β_1	N/A	0.999	0.026	0.000	-0.001, 0.001	0.096	93.6
MICE logistic	β_2		0.999	0.052	0.000	-0.003, 0.002	0.202	94.6
MICE RF	β_1	1	0.893	0.026	-0.107	-0.108, -0.106	0.174	23.6
MICE RF choose	β_2	1	0.977	0.052	-0.022	-0.025, -0.020	0.230	95.7
MICE RF	β_1	10	0.982	0.026	-0.018	-0.019, -0.017	0.113	92
MICE RF choose	β_2	10	0.964	0.051	-0.036	-0.038, -0.034	0.216	91.7
MICE RF biascorr	β_1	10	0.982	0.025	-0.018	-0.019, -0.017	0.108	91.1
MICE RF choose	β_2	10	0.957	0.051	-0.043	-0.045, -0.040	0.214	88.7
MICE RF	β_1	100	1.002	0.026	0.002	0.001, 0.003	0.104	95.4
MICE RF choose	β_2	100	0.96	0.051	-0.039	-0.042, -0.037	0.212	89.5
MICE RF biascorr	β_1	100	1.003	0.025	0.003	0.002, 0.004	0.098	94.3
MICE RF choose	β_2	100	0.947	0.051	-0.053	-0.055, -0.051	0.207	82.8
MICE RF choose	β_1	10	1.015	0.026	0.015	0.014, 0.016	0.105	92.7
MICE RF choose	β_2	10	0.959	0.051	-0.041	-0.043, -0.039	0.21	88.9

Table 4: Analysis of simulated datasets with continuous and binary predictor variables Missing at Random (MAR) using MICE and parametric or Random Forest imputation methods

Imputation Method	Coeffic-ient	Number of Trees	Estimate: Mean	SD	Bias of Estimate: Mean	95% CI	95% Confidence Interval: Length	Coverage, %
Full data	β_1	N/A	1.000	0.024	0.000	-0.001, 0.001	0.088	94.2
Full data	β_2		0.999	0.046	0.000	-0.002, 0.002	0.177	95.0
MICE normal	β_1	N/A	1.000	0.025	0.000	-0.001, 0.001	0.096	94
MICE logistic	β_2		0.999	0.051	-0.001	-0.003, 0.002	0.203	95.4
MICE RF	β_1	1	0.872	0.028	-0.127	-0.128, -0.126	0.219	29.8
MICE RF choose	β_2	1	0.948	0.060	-0.051	-0.054, -0.049	0.320	94.7
MICE RF	β_1	10	0.970	0.027	-0.030	-0.031, -0.028	0.130	89.2
MICE RF choose	β_2	10	0.939	0.058	-0.061	-0.063, -0.058	0.276	90.3
MICE RF biascorr	β_1	10	0.975	0.027	-0.025	-0.026, -0.024	0.123	90.2
MICE RF choose	β_2	10	0.931	0.057	-0.069	-0.071, -0.066	0.272	86.7
MICE RF	β_1	100	0.991	0.027	-0.009	-0.01, -0.007	0.115	95.6
MICE RF choose	β_2	100	0.938	0.058	-0.062	-0.064, -0.059	0.268	88.1
MICE RF biascorr	β_1	100	1.002	0.027	0.002	0.001, 0.003	0.108	95.6
MICE RF choose	β_2	100	0.918	0.056	-0.081	-0.084, -0.079	0.256	79.6
MICE RF choose	β_1	10	0.988	0.027	-0.012	-0.013, -0.011	0.123	95.8
MICE RF choose	β_2	10	0.941	0.058	-0.059	-0.062, -0.057	0.272	89.4

were positively correlated with y. This means that the imputation models tended to contain lower values of x_1 and x_2 than the missing values they were trying to predict, and in Random Forest this may have introduced bias. Web Figure 1 shows that imputed values using Random Forest were biased downwards for large values of x_1, compared to values imputed by MICE with linear regression.

Figure 1: Comparison of true versus mean imputed values for linear regression and Random Forest imputation functions, based on 1000 imputations of a single partially observed dataset.



It is probable that Random Forest with 10 trees was less biased than 100 trees in this simulation because the extra variance in the forest accommodated the prediction bias. However, using a single tree was also worse than 10 trees. Parameter estimates were biased towards the null with 1 tree, probably because the mean was predicted with a large degree of error, and confidence intervals were wide because of the variability in the imputations.

In this simulation study we used data which conformed perfectly to a specific parametric distribution. However, in real datasets such as electronic health record datasets used in epidemiological studies, values of patient measurements are frequently not normally distributed and may have interactions and non-linear associations. Such datasets have a large number of potential explanatory variables which can be used in imputation models, so it would be worth testing Random Forest imputation in such datasets.

References

[1] van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. J Stat Softw. 2011 12;45(3):1–67. Available from: http://www.jstatsoft.org/v45/i03.

[2] Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2002;2(3):18–22. Available from: http://CRAN.R-project.org/doc/Rnews/.

[3] Mendez G, Lohr S. Estimating residual variance in random forest regression. Computational Statistics & Data Analysis. 2011;55(11):2937–2950. Available from: http://www.sciencedirect.com/science/article/pii/S0167947311001514.

[4] Breiman L, Cutler A. Manual on setting up, using, and understanding Random Forests V3.1; 2002. Available from: http://oz.berkeley.edu/users/breiman/Using_random_forests_V3.1.pdf.