Imputation accuracy

Rianne Margaretha Schouten[1,2]

[1]University of Utrecht, Department of Methodology and Statistics, Sjoerd Groenman building, Padualaan 14, 3584 CH Utrecht, The Netherlands
[2]DPA Professionals, Business Unit DPA IT, Team Data Science, Gatwickstraat 11, 1043 GL Amsterdam, The Netherlands

Simulation setup

1. Generate multivariate normal data: y, x_1 and x_2, n = 1000

2. Generate missingness with two patterns: in x_1 and in x_2

3. Four mechanisms: MCAR, MAR based on other x, MAR based on y, MNAR

4. Four missing data methods: drop (listwise deletion), mean imputation, regression imputation, stochastic regression imputation. Nsim = 100

5. Evaluate imputation accuracy: how close are imputed values to the original values. Calculated with the following formula (based on Tang & Ishwaran, 2016):

$$\frac{1}{J}\sum_{j=1}^{J=2}\sqrt{\frac{\sum_{i=1}^{n_j}(X^*_{i,j}-X_{i,j})^2/n_j}{\sum_{i=1}^{n_j}(X_{i,j}-\bar{X}_j)^2/n_j}} \tag{1}$$

   In words: numerator: mean of the squared distance between the imputed values (X^*) and the original values ($X_{i,j}$). denominator: mean of the squared distance between the original values (only of the amputed values) and the mean of those values. Take the sqrt of this value, do this for x_1 and x_2, and sum those values.

   Implementation in R:

```r
R <- as.data.frame(1 * is.na(ampdata))

num <- mean((imps1 - data[R$x1 == 1, "x1"])^2)
denom <- mean((data[R$x1 == 1, "x1"] - mean(data[R$x1 == 1, "x2"]))^2)
mse.x1 <- sqrt(num / denom)

num <- mean((imps2 - data[R$x2 == 1, "x2"])^2)
denom <- mean((data[R$x2 == 1, "x2"] - mean(data[R$x2 == 1, "x2"]))^2)
mse.x2 <- sqrt(num / denom)

srmse <- (mse.x1 + mse.x2) / 2
```

Interpretation SRMSE:

- SRMSE = 1.0 then num = denom then var imps = var data
- SRMSE > 1.0 then num > denom then var imps > var data
- SRMSE < 1.0 then num < denom then var imps < var data

Results

See Figures 1 to 4, note that the drop method is not visible since it is not possible to calculate imputation accuracy with this missing data method.

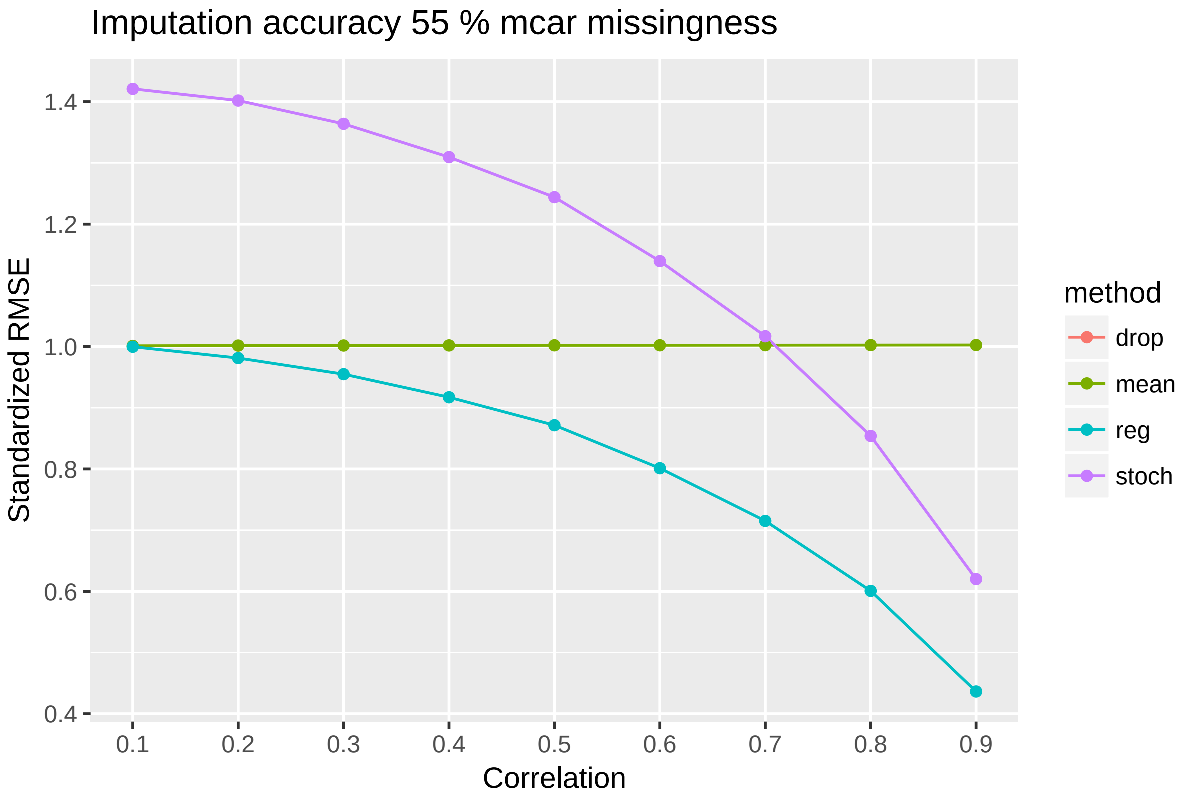
1. MCAR: mean imputation gives SRMSE of 1.0 and around 1.0 for MARRIGHT Y and MNAR. Only for MARRIGHT X, SRMSE changes when correlation changes.

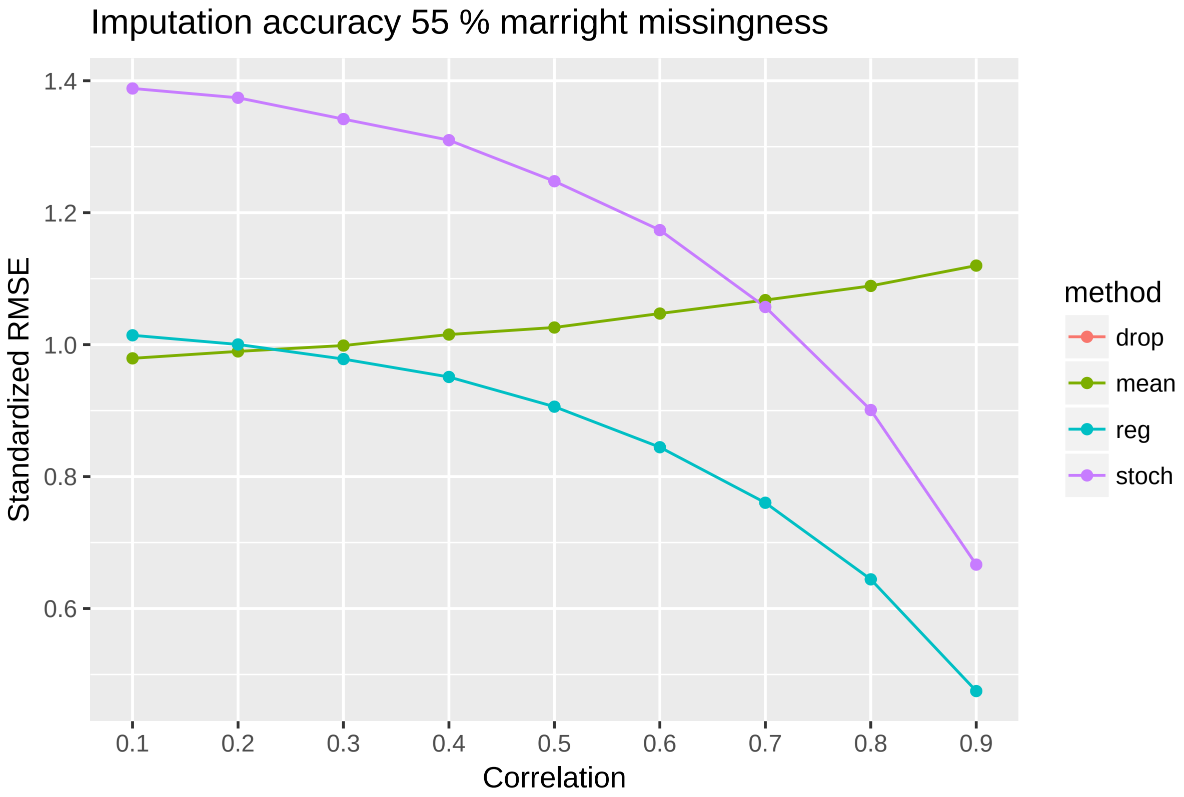
2. Stochastic regression imputation always results in a larger SRMSE than normal regression imputation, since the variance in the imputations is larger.

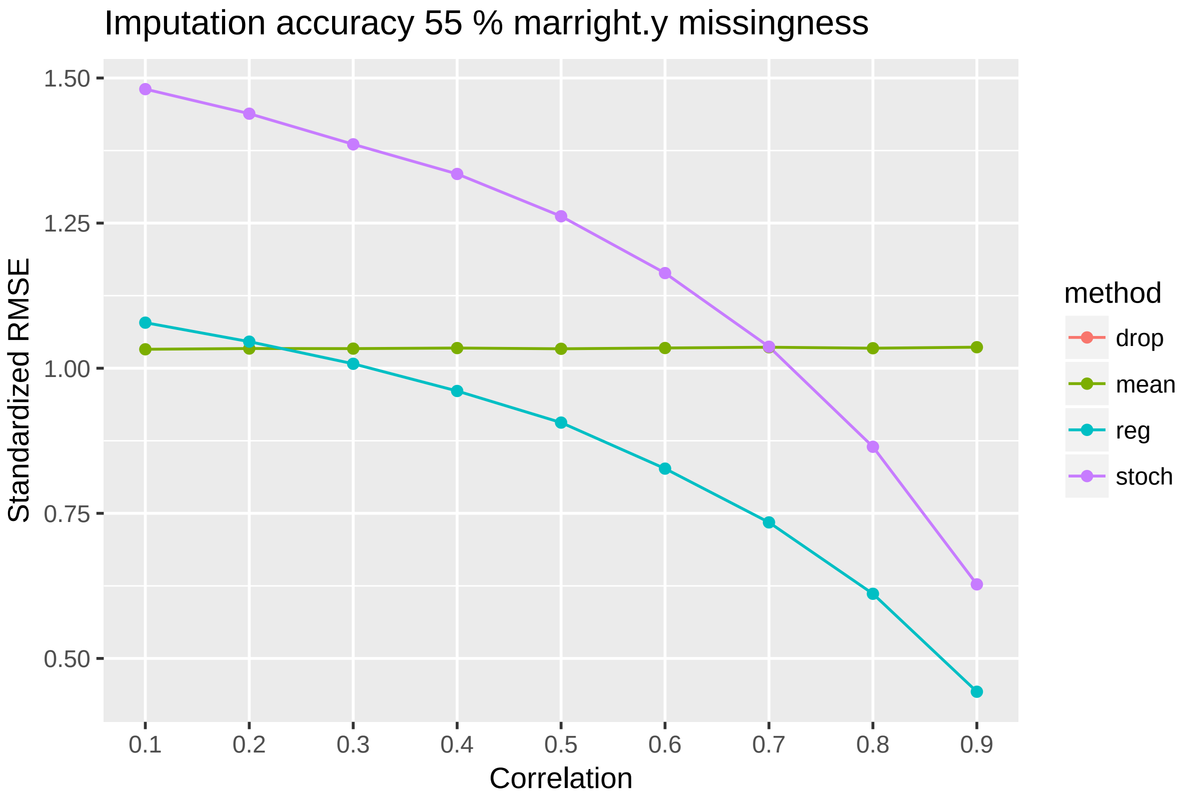
3. For both stochastic and normal regression imputation, SRMSE decreases when correlation in data increases.

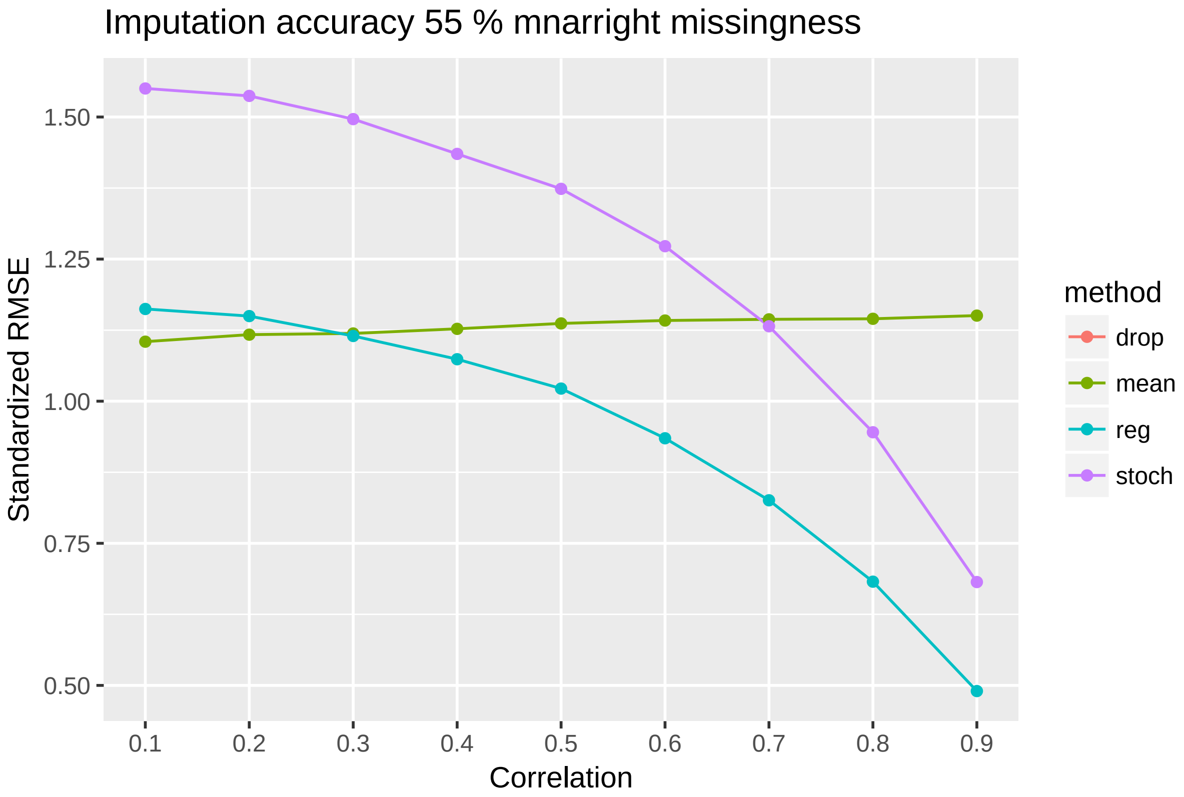
4. Regression imputation SRMSE is almost always < 1.0: this means that variance in imputations is smaller than in the data. For stochastic regression imputation, this occurs when correlation is large.

5. There are only small differences between the missingness mechanisms. Overall SRMSE is smallest for MCAR < MAR < MNAR, but very close!


Imputation accuracy 55 % mcar missingness
Standardized RMSE
1.4
1.2
1.0
0.8
0.6
0.4
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Correlation
method
drop
mean
reg
stoch


Imputation accuracy 55 % marright missingness
Standardized RMSE
1.4
1.2
1.0
0.8
0.6
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Correlation
method
drop
mean
reg
stoch


Imputation accuracy 55 % marright.y missingness
Standardized RMSE
1.50
1.25
1.00
0.75
0.50
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Correlation
method
drop
mean
reg
stoch


Imputation accuracy 55 % mnarright missingness
Standardized RMSE
1.50
1.25
1.00
0.75
0.50
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Correlation
method
drop
mean
reg
stoch