Prediction 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 prediction accuracy: perform 5-fold crossvalidation, for every fold: apply imputation method on trainingdata (80%), calculate parameters of linear regression model ($y \sim x_1 + x_2$), apply same imputation method on validationdata (20%), calculate y predictions using training model, calculate difference between predictions and true y values as RMSE. Average over the 5 folds.

For instance, one fold for mean imputation:

```
R.train <- as.data.frame(1 * is.na(training))
training[R.train$x1 == 1, "x1"] <- mean(training[R.train$x1 == 0, "x1"])
training[R.train$x2 == 1, "x2"] <- mean(training[R.train$x2 == 0, "x2"])
fit <- lm(y ~ x1 + x2, training)
coefs <- unname(fit$coefficients)

R.val <- as.data.frame(1 * is.na(validatie))
validatie[R.val$x1 == 1, "x1"] <- mean(training[R.train$x1 == 0, "x1"])
validatie[R.val$x2 == 1, "x2"] <- mean(training[R.train$x2 == 0, "x2"])

predictions <- coefs[1] + coefs[2]*validatie[, "x1"] + coefs[3]*validatie[,
rmse <- sqrt(mean((predictions - validatie[, "y"])^2))
```

Interpretation RMSE: the smaller the RMSE, the better the performance of the imputation and linear regression model in predicting y from x_1 and x_2 with missing values in these x features.

Results

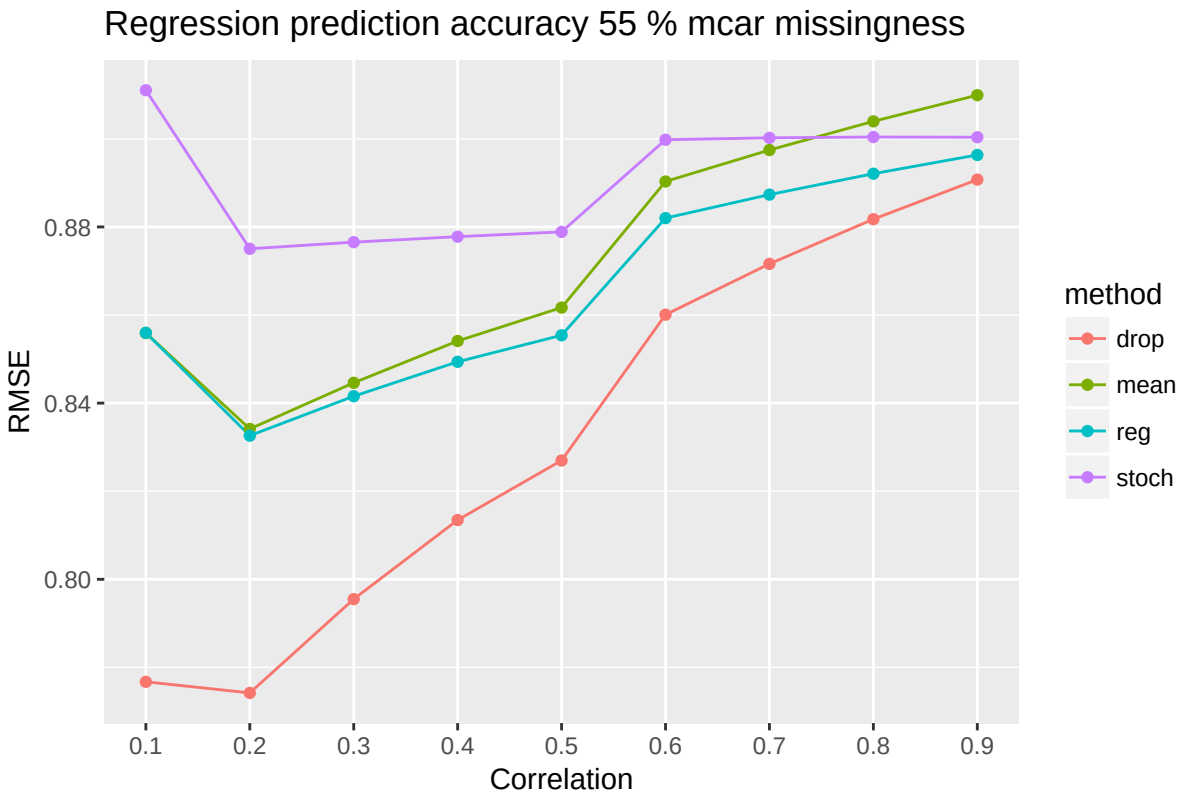
1. Drop gives lowest accuracy for every mechanism. With MARRIGHT on Y, drop is extremely lower than the other methods.

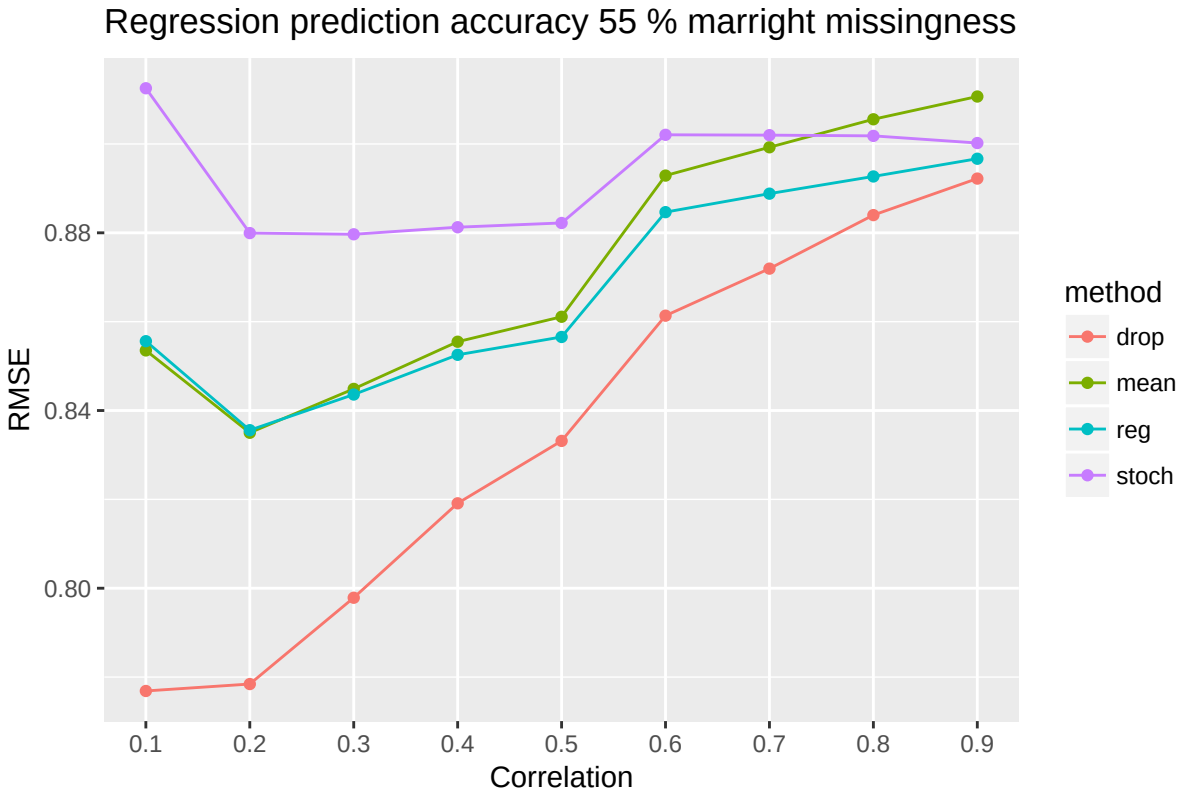
2. Mean and regression imputation overlap for low correlations for MCAR and MARRIGHT on X.

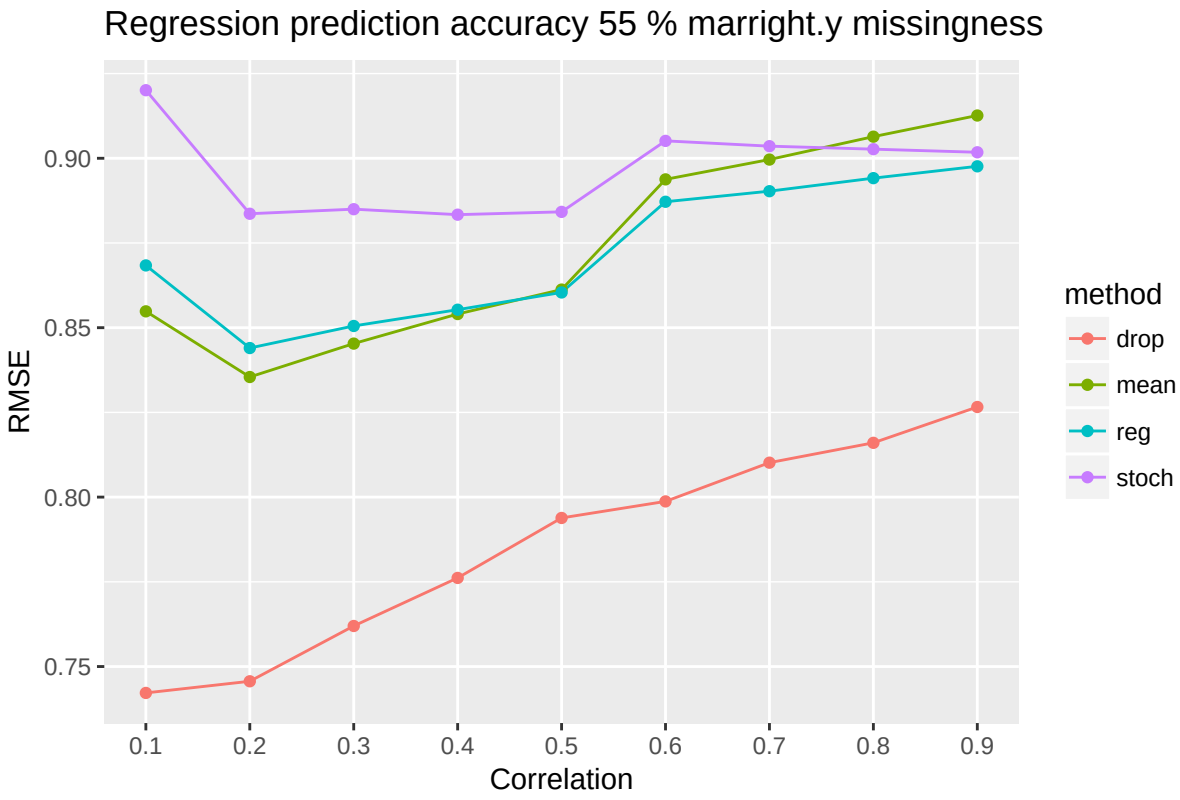
3. Mean gives lower RMSE than regression imputation for low correlations and MARRIGHT on Y and MNAR missingness.

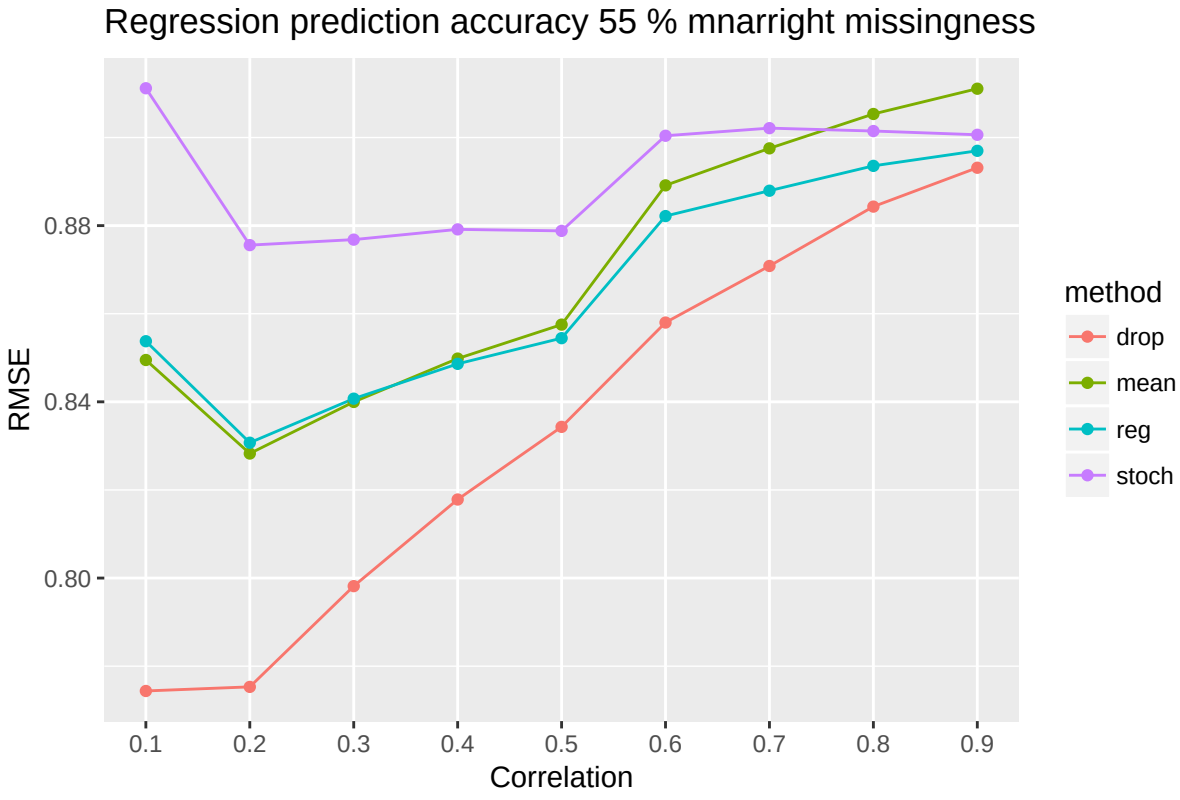
4. For high correlations, regression imputation gives lower RMSE than mean imputation.

5. For very high correlations, stochastic regression imputation gives better results than regression imputation. This probably has to do with the validation procedure: the imputations are based on the trainingdata and with high correlations, the regression line in the trainingset is very different from the one in the validation set. Stochastic regression imputation better accounts for these differences.


Regression prediction accuracy 55 % mcar missingness
RMSE
0.88
0.84
0.80
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Correlation
method
drop
mean
reg
stoch


Regression prediction accuracy 55 % marright missingness
RMSE
0.88
0.84
0.80
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Correlation
method
drop
mean
reg
stoch


Regression prediction accuracy 55 % marright.y missingness
RMSE
0.90
0.85
0.80
0.75
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Correlation
method
drop
mean
reg
stoch


Regression prediction accuracy 55 % mnarright missingness
RMSE
0.88
0.84
0.80
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Correlation
method
drop
mean
reg
stoch