Lab – Multiple Imputation

The Mosaic Data set includes the (longitudinal) data of 924 subjects with a diagnosis of Type 2 Diabetes (T2D)

Variable Name	Description	Type
patID	Patient identifier	integer
date	Observation Date	character
time	Time (days) from the first Observation	integer
timeSEQ	Observation Sequence	integer
gender	M=Male F=female	character
smoke	Y=Yes, N=No, Ex=Ex Smoker	character
age	Age at visit (year)	integer
t2d	Years from the T2D diagnosis	numeric
hab1c	Glycated Hemoglobin Values	numeric
bmi	Body Mass Index	numeric
sbp	Systolic Blood Pressure	numeric
colTot	Total Cholesterol	numeric
trigl	Triglycerides	numeric
CreatCL	Creatinine Clearance	numeric
MicroAln	Micro Albuminuria	numeric
hasRET	Retinopathy diagnosis	integer (0-1)
hasNEPH	Nephropathy diagnosis	integer (0-1)
hasNEU	Neuropathy diagnosis	integer (0-1)
hasSTR	Stroke diagnosis	integer (0-1)
hasOCC	Arterial occulsion diagnosis	integer (0-1)
hasCIHD	Chronic ischemic heart disease diagnosis	integer (0-1)
hasPVD	Peripheral vascular disease diagnosis	integer (0-1)
MICRO	Microvascular complication diagnosis	integer (0-1)
MACRO	Macrovascular complication diagnosis	integer (0-1)

1. Prepare the Data Set

1.1. Set up your directory and load the MosaicData.csv data set

1.2. Using the dplyr piping:

1..1. create a novel data set (MyDataBaseline) selecting baseline data (time =0),

1..2. exclude all the complications diagnosis as an outcome (hasXXX, MICRO, MACRO)

1..3. exclude also the columns timeSEQ, date and time.

1..4. Transform gender and smoke in a proper format

2. Impute the Data and train a regression model

2.1. Check the missingness pattern with the md.pattern() function, what do you observe? Are there any variables that it would be better to remove? If so, exclude them and re-check the pattern.

2.2. The aim is to build a regression model that uses all the variable selected at this point to predict hba1c values. Build a provisional model with the original data set using the lm() function. Check the results.

2.3. Impute data with mice() using default methods, number of imputations (m=5) and iterations (maxit=5). Set seed. Check the results plotting the iterations results. What do you observe?

2.4. Check the methods used for imputations and change them on the basis of the variable type. Check if the iteration results improve. Plot the iteration results distributions with stripplot() and densityplot(). What do you observe?

2.5. Try to change the visit sequence, does it affect the results?

2.6. Increase the number of imputation and iteration, does it affect the results?

2.7. Train the regression models on the imputed datasets obtained with the chosen imputation strategy and check the results. What do you observe?

2.8. Create a data set only with complete observations, use the na.omit() function. Exclude Creatinine Clearance and Micro Albuminuria. Add random missing values with the function ampute(). Transform gender and smoke into factors.

2.9. Train a regression model on the complete data (we assume this is the ground truth)

2.10. Impute the data set with random missing values with at least two different approaches (i.e. change methods, number of iterations). Train the regression models.

2.11. Compute the RMSE to compare the results of the different imputation approaches. Chose the best one. What do you observe?