**Missing Values (using R package “MICE”):**

From: <https://datascienceplus.com/handling-missing-data-with-mice-package-a-simple-approach/>

Note, the “long tutorial” is posted as well.

**Handling missing data with MICE package; a simple approach**

This is a quick, short and concise tutorial on how to impute missing data. Previously, we have published [an extensive tutorial on imputing missing values](http://datascienceplus.com/imputing-missing-data-with-r-mice-package/) with MICE package. Current tutorial aim to be simple and user friendly for those who just starting using R.

**Preparing the dataset**

I have created a simulated dataset, which you can load on your R environment by using the following code.

dat <- read.csv(url("<https://goo.gl/4DYzru>"), header=TRUE, sep=",")

Let’s see the header of dataset.

head(dat)

## Age Gender Cholesterol SystolicBP BMI Smoking Education

## 1 67.9 Female 236.4 129.8 26.4 Yes High

## 2 54.8 Female 256.3 133.4 28.4 No Medium

## 3 68.4 Male 198.7 158.5 24.1 Yes High

## 4 67.9 Male 205.0 136.0 19.9 No Low

## 5 60.9 Male 207.7 145.4 26.7 No Medium

## 6 44.9 Female 222.5 130.6 30.6 No Low

Check the data for missing values.

sapply(dat, function(x) sum(is.na(x)))

## Age Gender Cholesterol SystolicBP BMI Smoking

## 0 0 0 0 0 0

## Education

## 0

Since there are no missings, I will add some NA in dataset, but before I will duplicate original dataset to evaluate the accuracy of imputation later.

original <- dat

Now I will add some missings in few variables.

set.seed(10)

dat[sample(1:nrow(dat), 20), "Cholesterol"] <- NA

dat[sample(1:nrow(dat), 20), "Smoking"] <- NA

dat[sample(1:nrow(dat), 20), "Education"] <- NA

dat[sample(1:nrow(dat), 5), "Age"] <- NA

dat[sample(1:nrow(dat), 5), "BMI"] <- NA

Confirm the presence of missings in the dataset.

sapply(dat, function(x) sum(is.na(x)))

## Age Gender Cholesterol SystolicBP BMI Smoking

## 5 0 20 0 5 20

## Education

## 20

Next step is to transform the variables in factors or numeric. For example, smoking and education are categorical variables, whereas cholesterol level is continuous.

library(dplyr)

dat <- dat %>%

mutate(

Smoking = as.factor(Smoking),

Education = as.factor(Education),

Cholesterol = as.numeric(Cholesterol)

)

Look the dataset structure.

str(dat)

## 'data.frame': 250 obs. of 7 variables:

## $ Age : num 67.9 54.8 68.4 67.9 60.9 44.9 49.9 NA 57.5 77.2 ...

## $ Gender : Factor w/ 2 levels "Female","Male": 1 1 2 2 2 1 2 1 2 2 ...

## $ Cholesterol: num 236 256 199 205 208 ...

## $ SystolicBP : num 130 133 158 136 145 ...

## $ BMI : num 26.4 28.4 24.1 19.9 26.7 30.6 27.3 27.5 28.3 29.1 ...

## $ Smoking : Factor w/ 2 levels "No","Yes": 2 1 2 1 1 1 1 1 1 1 ...

## $ Education : Factor w/ 3 levels "High","Low","Medium": 1 3 1 NA NA 2 3 2 1 1 ...

Everything looks OK, so lets proceed with imputation.

**Imputation**

Now that the dataset is ready for imputation, we will call the mice package. The code below is standard and you dont need to change anything besides the dataset name.

library(mice)

init = mice(dat, maxit=0)

meth = init$method

predM = init$predictorMatrix

To impute the missing values, mice package use an algorithm in a such a way that use information from other variables in dataset to predict and impute the missing values. Therefore, you may not want to use certain variable as predictors. For example the ID variable does not have any predictive value.

The code below will remove the variable as predictor but still will be imputed. Just for illustration purposes I select the BMI variable to not be included as predictor during imputation.

predM[, c("BMI")]=0

If you want to skip a variable from imputation use the code below. This variable will be used for prediction.

meth[c("Age")]=""

Now let specify the methods for imputing the missing values. There are specific methods for continues, binary and ordinal variables. I set different methods for each variable. You can add more than one variable in each methods.

meth[c("Cholesterol")]="norm"

meth[c("Smoking")]="logreg"

meth[c("Education")]="polyreg"

Now it is time to run the multiple (m=5) imputation.

set.seed(103)

imputed = mice(dat, method=meth, predictorMatrix=predM, m=5)

## iter imp variable

## 1 1 Cholesterol BMI Smoking Education

## 1 2 Cholesterol BMI Smoking Education

## 1 3 Cholesterol BMI Smoking Education

## 1 4 Cholesterol BMI Smoking Education

## 1 5 Cholesterol BMI Smoking Education

## 2 1 Cholesterol BMI Smoking Education

## 2 2 Cholesterol BMI Smoking Education

...

Create a dataset after imputation.

imputed <- complete(imputed)

Check for missings in the imputed dataset.

sapply(imputed, function(x) sum(is.na(x)))

## Age Gender Cholesterol SystolicBP BMI Smoking

## 5 0 0 0 0 0

## Education

## 0

**Accuracy**

In this example, we know the actual values of missing data, since I added the missings. This indicate that we can check the accuracy of the imputation. However, we should acknowledge that this is an simulated dataset, and therefore, variables have no scientific meanings and are not correlated to each other. Therefore I expect a lower rate of accuracy for this imputation.

# Cholesterol

actual <- original$Cholesterol[is.na(dat$Cholesterol)]

predicted <- imputed$Cholesterol[is.na(dat$Cholesterol)]

# Smoking

actual <- original$Smoking[is.na(dat$Smoking)]

predicted <- imputed$Smoking[is.na(dat$Smoking)]

table(actuals)

table(predicted)

mean(actual)

mean(predicted)

## [1] 231.07

## [1] 231.3564

## actual

## No Yes

## 11 9

## predicted

## No Yes

## 14 6

The mean of actual and predicted for Cholesterol is almost identical, which shows a high accuracy of imputation, whereas for smoking is low.