TutorialImputingMissingData

Tutorial on Missing Data Imputation

We follow the tutorial on R packages for missing data imputation by MANISH SARASWAT which can be found here: https://www.analyticsvidhya.com/blog/2016/03/tutorial-powerful-packages-imputing-missing-values/

MICE Package

```
library(missForest)
library(mice)
library(VIM)
```

```
data <- iris
```

Generate Missing Data with missForest

Generate 10% missing values at Random using the missForest package

```
iris.mis <- prodNA(iris, noNA = 0.1)
summary(iris.mis)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.200   Min.   :1.000   Min.   :0.10
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.575   1st Qu.:0.30
##  Median :5.800   Median :3.000   Median :4.350   Median :1.35
##  Mean   :5.828   Mean   :3.054   Mean   :3.730   Mean   :1.22
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.80
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.50
##  NA's   :17      NA's   :12      NA's   :14      NA's   :18
##        Species
##  setosa    :47
##  versicolor:42
##  virginica :47
##  NA's      :14
##
##
##
```

Remove categorical variables and focus on continuous variables

```
iris.mis <- subset(iris.mis, select = -c(Species))
summary(iris.mis)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.200   Min.   :1.000   Min.   :0.10
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.575   1st Qu.:0.30
##  Median :5.800   Median :3.000   Median :4.350   Median :1.35
##  Mean   :5.828   Mean   :3.054   Mean   :3.730   Mean   :1.22
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.80  
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.50  
##  NA's   :17      NA's   :12      NA's   :14      NA's   :18    
```

Inspect Missing Pattern with MICE

```
md.pattern(iris.mis)
```

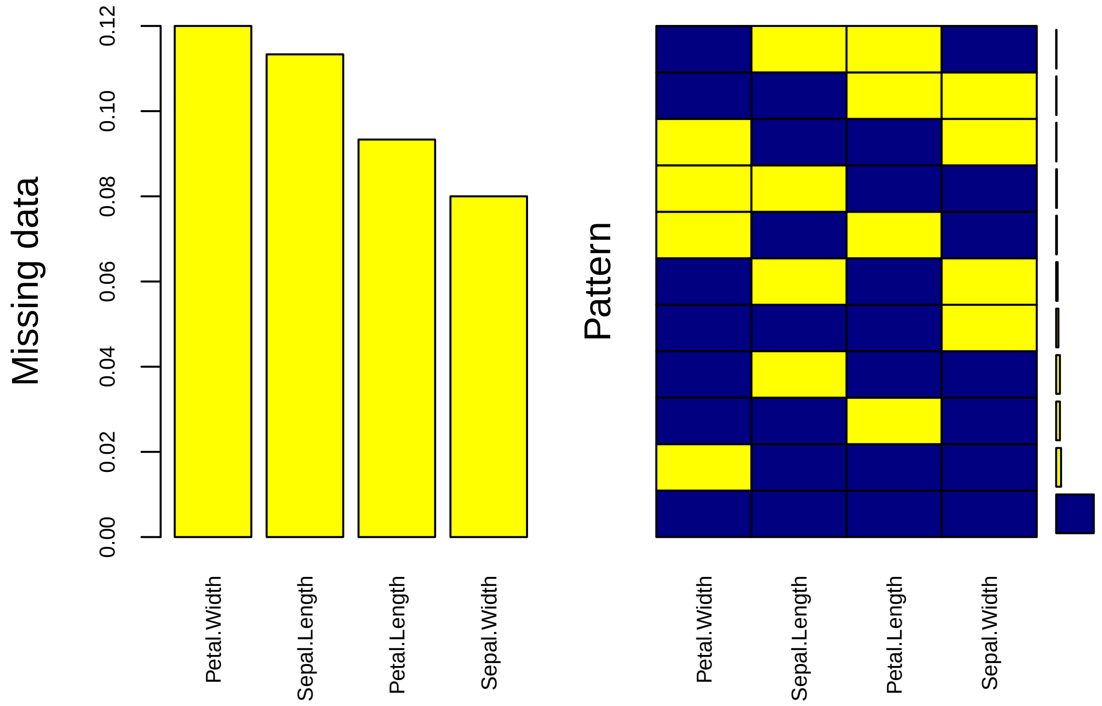
```
##     Sepal.Width Petal.Length Sepal.Length Petal.Width   
## 100           1            1            1           1  0
##  10           1            1            0           1  1
##   6           0            1            1           1  1
##  10           1            0            1           1  1
##  13           1            1            1           0  1
##   4           0            1            0           1  2
##   1           1            0            0           1  2
##   1           0            0            1           1  2
##   2           1            1            0           0  2
##   1           0            1            1           0  2
##   2           1            0            1           0  2
##              12           14           17          18 61
```

```
md.pattern(iris.mis)
```

```
##     Sepal.Width Petal.Length Sepal.Length Petal.Width   
## 100           1            1            1           1  0
##  10           1            1            0           1  1
##   6           0            1            1           1  1
##  10           1            0            1           1  1
##  13           1            1            1           0  1
##   4           0            1            0           1  2
##   1           1            0            0           1  2
##   1           0            0            1           1  2
##   2           1            1            0           0  2
##   1           0            1            1           0  2
##   2           1            0            1           0  2
##              12           14           17          18 61
```

Visual Inspection of Missing Patern with VIM

```
mice_plot <- aggr(iris.mis, col=c('navyblue','yellow'),
                    numbers=TRUE, sortVars=TRUE,
                    labels=names(iris.mis), cex.axis=.7,
                    gap=3, ylab=c("Missing data","Pattern"))
```



```
##
##  Variables sorted by number of missings:
##      Variable      Count
##   Petal.Width 0.12000000
##  Sepal.Length 0.11333333
##  Petal.Length 0.09333333
##   Sepal.Width 0.08000000
```

Imputing the missing data with MICE

```
imputed_Data <- mice(iris.mis, m=5, maxit = 50, method = 'pmm', seed = 500)
```

```
summary(imputed_Data)
```

```
## Multiply imputed data set
## Call:
## mice(data = iris.mis, m = 5, method = "pmm", maxit = 50, seed = 500)
## Number of multiple imputations:  5
## Missing cells per column:
## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##           17           12           14           18
## Imputation methods:
## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##        "pmm"        "pmm"        "pmm"        "pmm"
## VisitSequence:
## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##            1            2            3            4
## PredictorMatrix:
```

```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length            0           1            1           1
## Sepal.Width             1           0            1           1
## Petal.Length            1           1            0           1
## Petal.Width             1           1            1           0
## Random generator seed value:  500
```

```
#check imputed values
imputed_Data$imp$Sepal.Width
```

```
##       1   2   3   4   5
## 12  2.9 3.3 2.8 3.0 2.8
## 33  3.6 3.7 3.0 3.6 3.0
## 36  3.1 3.0 3.5 3.4 3.3
## 38  3.1 3.0 3.4 3.1 3.1
## 49  3.5 3.5 3.8 3.1 3.2
## 52  3.0 3.2 2.6 3.0 2.8
## 60  2.8 3.0 2.5 2.8 2.8
## 61  2.8 2.4 2.7 3.3 2.5
## 69  2.8 2.7 2.8 3.4 3.0
## 71  2.7 2.6 2.5 2.5 2.7
## 112 2.9 2.5 3.4 3.4 2.5
## 144 3.0 3.0 3.0 2.8 3.0
```

```
#get complete data ( 2nd out of 5)
completeData <- complete(imputed_Data,2)
```

Build a model using the imputed data

```
#build predictive model
#Caveat I deviate from the Tutorial by using imputed_Data instead of iris.mis, because it otherwise thr
fit <- with(data = imputed_Data, exp = lm(Sepal.Width ~ Sepal.Length + Petal.Width))

#combine results of all 5 models
combine <- pool(fit)
summary(combine)
```

```
##                     est         se         t       df     Pr(>|t|)
## (Intercept)   1.9116909 0.32765022  5.834548 87.17461 8.996957e-08
## Sepal.Length  0.2884695 0.06752128  4.272275 84.05754 5.071101e-05
## Petal.Width  -0.4536662 0.07110477 -6.380250 97.69335 5.922347e-09
##                   lo 95      hi 95 nmis       fmi    lambda
## (Intercept)   1.2604690  2.5629128   NA 0.1400332 0.1205272
## Sepal.Length  0.1541973  0.4227416   17 0.1466066 0.1265404
## Petal.Width  -0.5947768 -0.3125555   18 0.1191318 0.1012812
```

Build a model without imputation to compare

```
raw.data <- iris
poor_fit <- fit <- with(data = raw.data, exp = lm(Sepal.Width ~ Sepal.Length + Petal.Width))
summary(poor_fit)
```

```
##
```

```
## Call:
## lm(formula = Sepal.Width ~ Sepal.Length + Petal.Width)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.99563 -0.24690 -0.00503  0.23354  1.01131
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.92632    0.32094   6.002 1.45e-08 ***
## Sepal.Length  0.28929    0.06605   4.380 2.24e-05 ***
## Petal.Width  -0.46641    0.07175  -6.501 1.17e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3841 on 147 degrees of freedom
## Multiple R-squared:  0.234,  Adjusted R-squared:  0.2236
## F-statistic: 22.46 on 2 and 147 DF,  p-value: 3.091e-09
```

The point estimates of the poor_fit regression summary (without imputation) differ from the regression coefficients based on the imputed data; the latter also have wider confidence bands expressing the increased uncertainty due to imputation.

AMELIA package