Final Report

Introduction

Myocardial infarction (MI) is a serious disease with high incidence rate across the world and especially the developed world. South Asians represent approximately 25 percent of the world's population but account for 60 percent of the world's heart disease patients. There is a genetic predisposition towards cardiovascular disease for the population making this topic/dataset is of high interest to me.

Data

a.Data Source

Citation: Golovenkin, S.E., Bac, J., Chervov, A., Mirkes, E.M., Orlova, Y.V., Barillot, E., Gorban, A.N. and Zinovyev, A., 2020. Trajectories, bifurcations, and pseudo-time in large clinical datasets: applications to myocardial infarction and diabetes data. GigaScience, 9(11), p.giaa128., DOI: 10.1093/gigascience/giaa128

Link: The dataset from taken from UCI Machine Learning Repository https://archive.ics.uci.edu/ml/datasets/Myocardial+infarction+complications#

b. Data Collection

Database was collected in the Krasnoyarsk Interdistrict Clinical Hospital No20 named after I. S. Berzon (Russia) in 1992-1995. Database contains 1700 records (patients), 111 input features and 12 complications. Database contains 7.6% of missing values.

c. Units of observations: The rows represent patient diagnostic points at certain intervals during the first three days of hospital stay after MI and the development of complications after that.

d.Variables: variables describing the end of the first day (24 hours after admission to the hospital): all input columns (2- 112) except 94, 95, 101, 102, 104, 105 can be studies and used for prediction.

```
data_raw <-read.csv("MICDatabase.csv", header=TRUE)
```

Data has more than 1700 rows and 124 variables. The number of columns are large in number and have been condensed into groups according for better explation as below. Only the column numbers are noted:

- Column -> 1: ID Numeric
- Column -> 2: Age Numeric
- Column -> 3: Sex Categorical.
- Column -> 4 - 6: Deal with History and Class of MI according to with Medical History (Anamsesis) in Ordinal values. INF_ANAM is the quantity of MI, STENOK_AN is last time the patient had Exertional angina pectoris; FK_STENOK is Functional class (FC) of Angina Pectoris in the last year.
- Column -> 7: Coronary heart disease (CHD) in recent weeks, days before admission to hospital (IBS_POST) in Ordinal Values.
- Column -> 8-11: Conditions from Medical History (Anamsesis) in Ordinal values dealing with presense and duration of variations of hypertension.

- Column -> 12-34: Presence of various medical conditions recorded from Patient Medical History (Anamsesis) in Binary 0, 1 representing No, Yes.
- Column -> 35-38: Blood pressure recordings by various teams in the process of admission in numeric values (mmHg)
- Column -> 39-44: Patient Condition at admission to ICU in Binary 0, 1 representing No, Yes.
- Column -> 45-49: Different Types of MI Conditions deduced from ECG in categorical values.
- Column -> 50-75: ECG Rhythm recordings at time of admission to Hospital in Binary 0, 1 representing No, Yes.
- Column -> 75-82: Administration of various Fibrinolytic Therapies by in Binary 0, 1 representing No, Yes.
- Column -> 83: Hypokalemia in Binary 0, 1 representing No, Yes.
- Column -> 84-91: Blood tests in Numeric concentration volume terms.
- Column -> 92: Time elapsed from the beginning of the attack of CHD to the hospital (TIME_B_S) in Ordinal values.

- Column -> 93-95: Relapse of pain for different days of Hospitalization in ordinal values.
- Column -> 96-112: Use of Opioid Drugs over the period of hospitalization from Emergency to ICU in ordinal values.
- Column -> 113-124: Output columns concerning various complications and Outcomes of MI.

e. Type of study: This was an observational study. f. Data clean-up Categorical variables are loaded as numerical variables from the data set. And some quantitative variables are loaded as integer columns. Converting the column types to factor(categorical) and numerical columns for better analysis.

```
# Data imported as Integer Columns -> Converted to Numeric Columns
data_raw[] <- lapply(data_raw, function(x) if(is.integer(x)) as.numeric(x) else x)

# Storing the Column Names
col_names <- colnames(data_raw)

# Storing the Numeric Column Names
numeric_cols <- c('AGE', 'S_AD_KBRIG', 'D_AD_KBRIG', 'S_AD_ORIT', 'D_AD_ORIT',
                  'K_BLOOD', 'NA_BLOOD', 'ALT_BLOOD', 'AST_BLOOD', 'KFK_BLOOD',
                  'L_BLOOD', 'ROE')

# Separating the Categorical Columns
categorical_cols <- col_names[! col_names %in% c('ID', numeric_cols)]

# The Categorical Columns are imported as Numeric Columns
# -> Converting them to categorical (factor) columns
data_raw[categorical_cols] <- lapply(data_raw[categorical_cols], factor)
```

Imputing Missing Values with missForest Library

```
diagnose(data_raw)
```

```
## # A tibble: 124 x 6
##    variables  types   missing_count missing_percent unique_count unique_rate
##    <chr>      <chr>           <int>           <dbl>        <int>       <dbl>
##  1 ID         numeric             0           0             1700     1
##  2 AGE        numeric             8           0.471           63     0.0371
##  3 SEX        factor              0           0                2     0.00118
##  4 INF_ANAM   factor              4           0.235            5     0.00294
```

```
##  5 STENOK_AN  factor            106          6.24            8     0.00471
##  6 FK_STENOK  factor             73          4.29            6     0.00353
##  7 IBS_POST   factor             51          3               4     0.00235
##  8 IBS_NASL   factor           1628         95.8             3     0.00176
##  9 GB         factor              9          0.529           5     0.00294
## 10 SIM_GIPERT factor              8          0.471           3     0.00176
## # ... with 114 more rows
```

From the diagnosis, columns like KFK_BLOOD, S_AD_KBRIG, D_AD_KBRIG, IBS_NASL have missing values more than 60% (up to ~95%) can are removed from the analysis. The ID column can also be removed as it doesn't add to the analysis.

```
data_cols = subset(data_raw, select = -c(ID, KFK_BLOOD, S_AD_KBRIG, D_AD_KBRIG, IBS_NASL))
```

This data set has a heavy mix of continuous and categorical variables. For each variable `missForest` fits a random forest on the observed part and then predicts the missing part. missForest is a nonparametric imputation method for basically any kind of data. It can cope with mixed-type of variables, nonlinear relations, complex interactions and high dimensionality (p » n). It only requires the observation (i.e. the rows of the data frame supplied to the function) to be pairwise independent.

```
data_fill <- missForest(data_cols)
```

Using data from cached File:

```
data_fill$OOBerror
```

```
##      NRMSE        PFC
## 0.17176160 0.08014556
```

NRMSE is normalized mean squared error. It is used to represent error derived from imputing continuous values. PFC (proportion of falsely classified) is used to represent error derived from imputing categorical values.

This suggests that categorical variables are imputed with 8% error and continuous variables are imputed with ~17% error.

Exploratory Data Analysis

Descriptive statistics and visualization of key variables

The data-set has 112 independent variables and 12 dependent variables.

- The analysis will focus on the particular binary dependent variable `REC_IM` -> Relapse of the Myocardial Infarction (MI) after hospitalization. Its a complication arising from the initial MI and I would like to analyse the factors that could be used to predict the it. Remove other depended variables except the Relapse of the myocardial infarction (REC_IM):

```
df = subset(data_fill$ximp, select = -c(FIBR_PREDS, PREDS_TAH, JELUD_TAH, FIBR_JELUD, A_V_BLOK, OTEK_LA

freq(select(df, REC_IM))
```

```
## Frequencies
## REC_IM
## Type: Factor
##
##                Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- ------ --------- -------------- --------- --------------
##           0   1541     90.65          90.65     90.65          90.65
##           1    159      9.35         100.00      9.35         100.00
##        <NA>      0                                0.00         100.00
##       Total   1700    100.00         100.00    100.00         100.00
```

The Frequency of 'No Relapse of MI', i.e, value 0 is 90.65%. This is an imbalanced data-set which could affect the performance of the logistic regression and PCA that would be performed later.

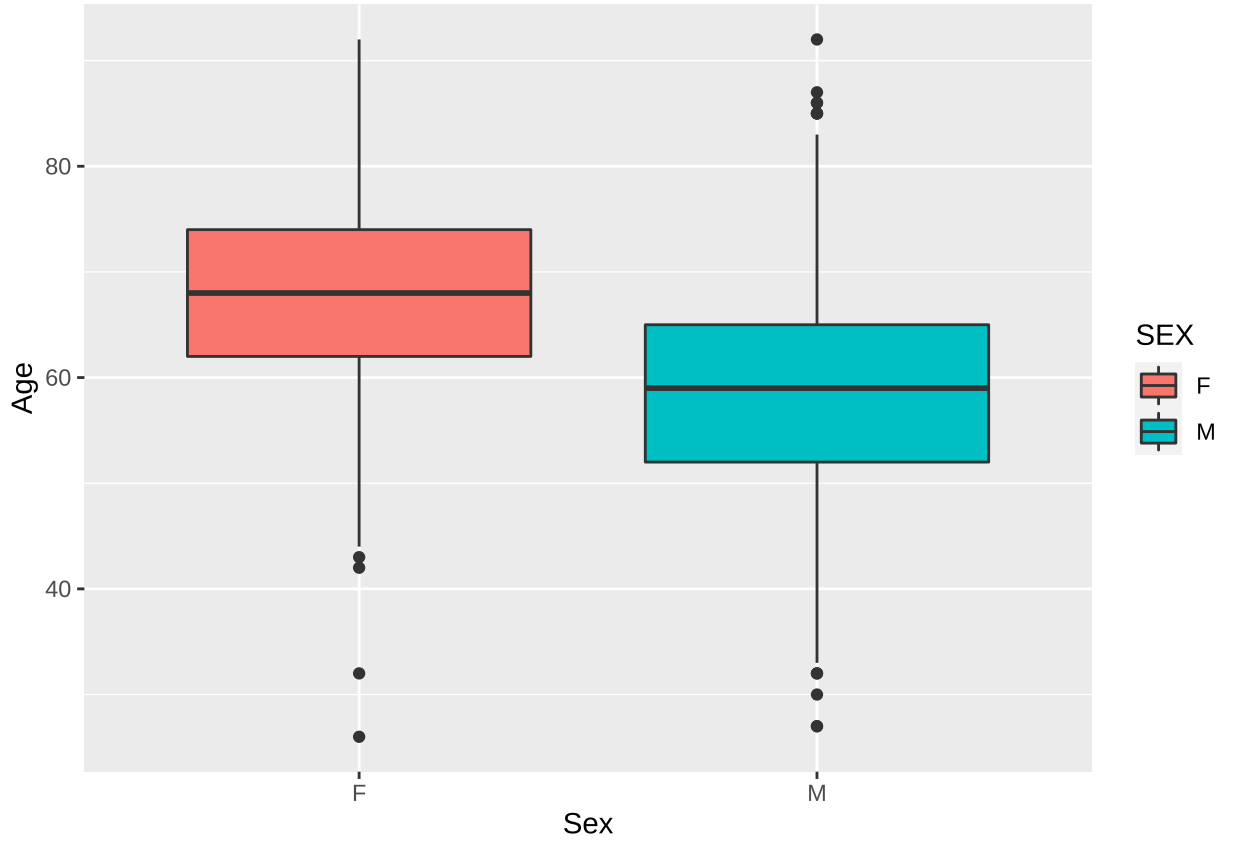
The depended variables of interest are analyses below

1. AGE - The age of the patient; numerical variable.
2. SEX - Gender of the patient; Binary Variable.
3. STENOK_AN - Exertional angina pectoris in the Patient Medical History from the last few years; Ordinal variable.
4. IBS_POST - Coronary heart disease (CHD) in recent weeks, days before admission to hospital; Categorical variable.
5. S_AD_ORIT - Systolic blood pressure according to ICU in mmHg; Numerical variable;
6. D_AD_ORIT - Diastolic blood pressure according to ICU in mmHg; Numerical variable;

Descriptive statistics and visualization of key variables

Age distribution between male and female patients:

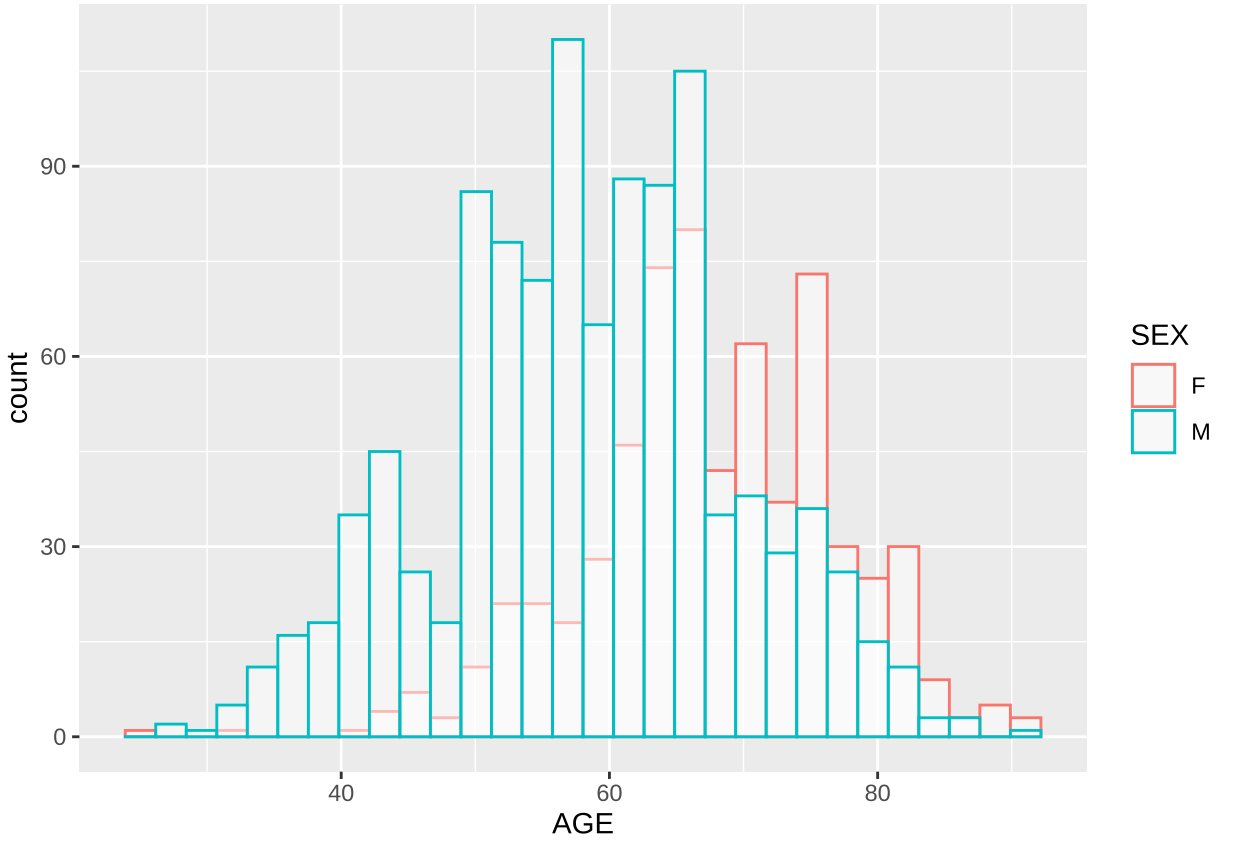
```
df$SEX <-recode(df$SEX, '0'="F", '1'= "M")

ggplot(df, aes(x = as.factor(SEX),y = AGE,fill= SEX))+
  geom_boxplot() +
  labs(x="Sex", y = "Age") +
  theme(plot.caption = element_text(hjust = 0.5))
```



On average, the age of admission for female patients is higher than male. It is in line with the generally known result that their is a higher incidence of heart related ailments in male than female.

```
ggplot(df, aes(x=AGE, color=SEX)) +
  geom_histogram(fill="white", alpha=0.5, position="identity")
```

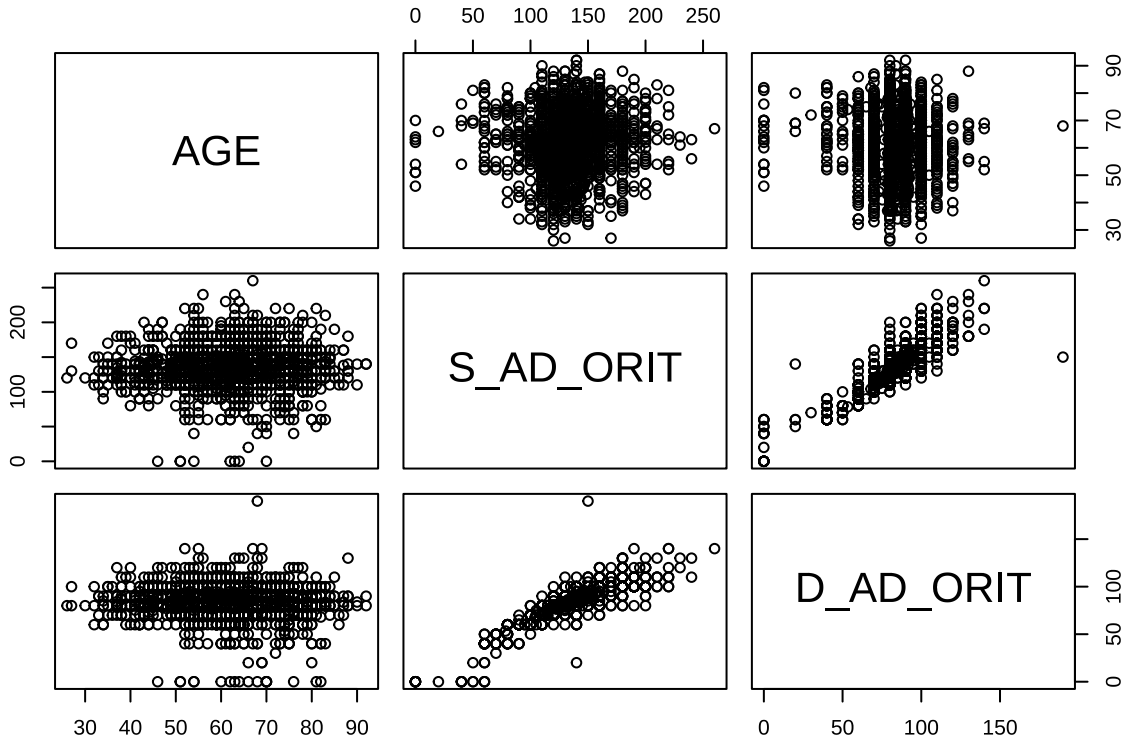
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Its clear from the graph that male patients admitted to hospital are more in the data-set and younger than female patients.

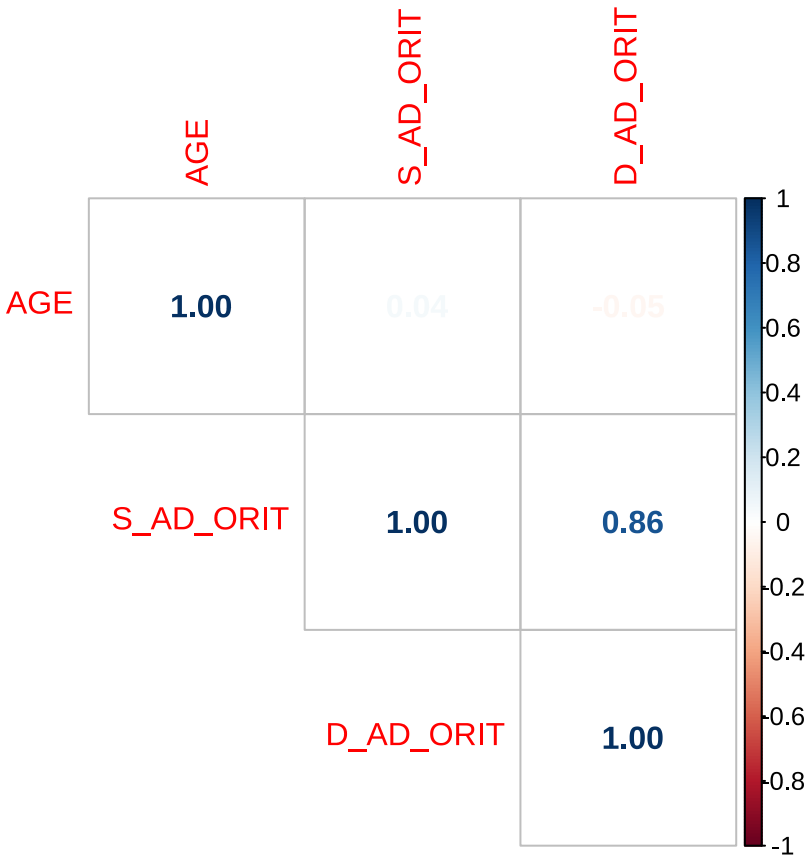
Correlation between continuous variables:

```
pairs(df[, c("AGE", "S_AD_ORIT", "D_AD_ORIT")])
```



There seems to be positive correlation between the two BP variables.

```
corrplot(cor(df[, c("AGE", "S_AD_ORIT", "D_AD_ORIT")]),
         method = "number",
         type = "upper")
```



From the correlation plot, the Blood Pressure variables (S_AD_ORIT, D_AD_ORIT) are naturally related but surprisingly there seems to be little relationship with age. Admission to hospital for CHD/MI related ailments generally have higher BP recordings at the time of admission. Being relatively positively related, they are not independent of each other.

```
(test <- cor.test(df$S_AD_ORIT, df$D_AD_ORIT) )
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$S_AD_ORIT and df$D_AD_ORIT
## t = 70.199, df = 1698, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8496947 0.8741028
## sample estimates:
##       cor
## 0.8623991
```

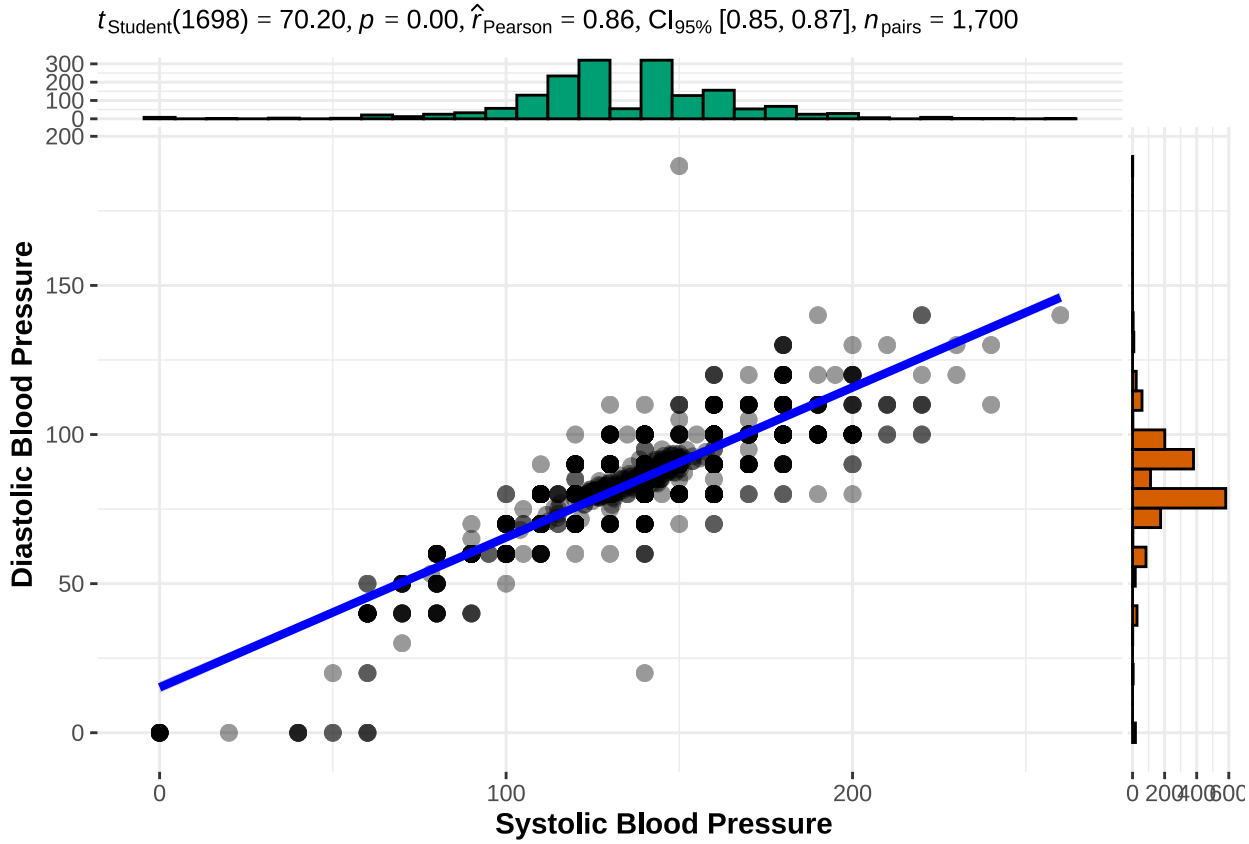
Just to confirm, they are correlated.

```
ggscatterstats(
  data = df,
  x = S_AD_ORIT,
  y = D_AD_ORIT,
  xlab = 'Systolic Blood Pressure',
```

```
  ylab = 'Diastolic Blood Pressure',
  bf.message = FALSE
)
```

```
## Registered S3 method overwritten by 'ggside':
##   method from
##   +.gg   ggplot2
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Inference: PCA

1. Approach

The data-set consists of a mix of both categorical variables and continuous. PCA mostly does well with only continuous variables. So in our case, the Factor analysis of mixed data (FAMD), a principal component method dedicated to analyze a data set containing both quantitative and qualitative variables should be used. It makes it possible to analyze the similarity between individuals by taking into account a mixed types of variables. Additionally, one can explore the association between all variables, both quantitative and qualitative variables.

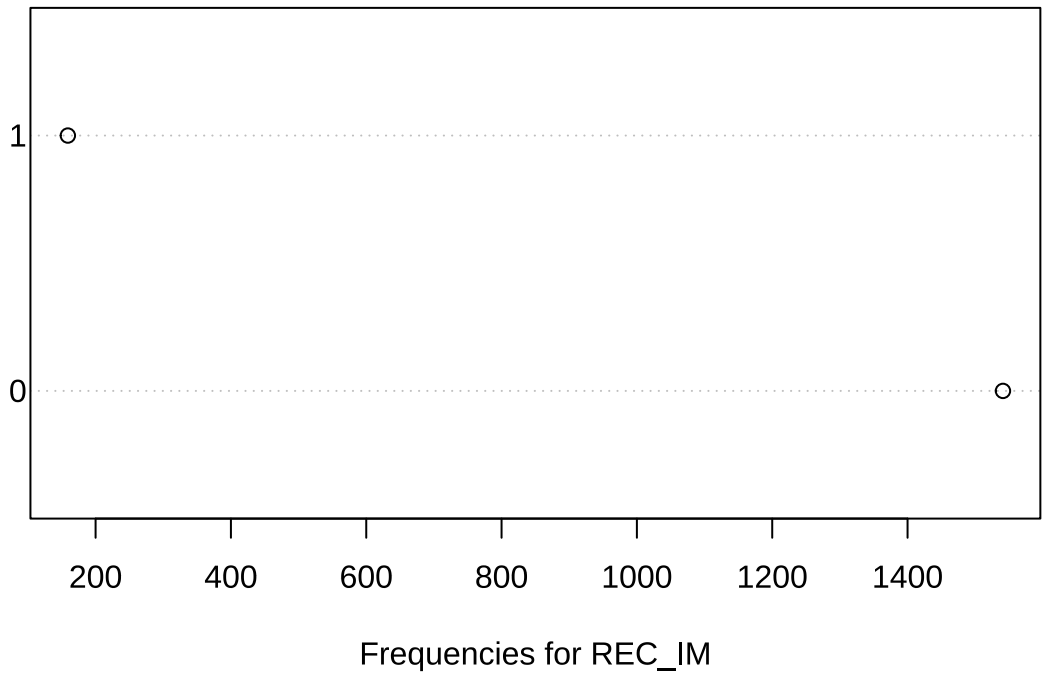
Roughly speaking, the FAMD algorithm can be seen as a mixed between PCA for quantitative variables and multiple correspondence analysis (MCA) for qualitative variables. Quantitative and qualitative variables are normalized during the analysis in order to balance the influence of each set of variables.

2. Formulation

```
freq(select(data_cols, REC_IM))
```

```
## Frequencies
## REC_IM
## Type: Factor
##
##                Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- ------ --------- -------------- --------- --------------
##            0   1541     90.65          90.65     90.65          90.65
##            1    159      9.35         100.00      9.35         100.00
##        <NA>      0                                0.00         100.00
##       Total   1700    100.00         100.00    100.00         100.00
```

```
hist.data.frame(data_fill$ximp %>%
                  select(c('REC_IM')))
```



The data-set is imbalanced (90.65% for the negative class) and it will affect the performance of the FADM and logistic regression. One way to solve the problem is use the ROSE algorithm. It is similar to up-sampling and is to create synthetic samples. Adding synthetic samples is also only done after the train-test split, into the training data. So first the data has to be split into training and testing sets but for a imbalanced set, stratified sampling is used in order to make sure that both the sets have the same proportion of minority/majority classes.

```
#set.seed(123)
train.index <- createDataPartition(df$REC_IM, p = .7, list = FALSE)
train <- df[ train.index,]
test  <- df[-train.index,]
```

Now we can see that the proportion remains same in the training set.

```
freq(select(train, REC_IM))
```

```
## Frequencies
## REC_IM
## Type: Factor
##
##               Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- ------ --------- -------------- --------- --------------
##           0   1079     90.60          90.60     90.60          90.60
##           1    112      9.40         100.00      9.40         100.00
##        <NA>      0                                0.00         100.00
##       Total   1191    100.00         100.00    100.00         100.00
```

ROSE method

ROSE (Random Over-Sampling Examples) aids the task of binary classification in the presence of rare classes. It produces a synthetic, possibly balanced, sample of data simulated according to a smoothed-bootstrap approach. ROSE package helps us to generate artificial data based on sampling methods and smoothed bootstrap approach. Before resampling:

```
table(train$REC_IM)
```

```
##
##    0    1
## 1079  112
```

After resampling:

```
data.rose <- ROSE( REC_IM ~ ., data = train, seed = 1)$data
table(data.rose$REC_IM)
```

```
##
##   0   1
## 624 567
```

We can see that the classes are not now balanced better. Only the training set is balanced and NOT the test set. Also to be noted is that the data generation has the SAME number of TOTAL observations as BEFORE.

FAMD

The data-set has many dependent variables. This analysis focuses on REC_IM, that is the Relapse of the Myocardial Infarction after admission to hospital. Selecting the required dependent variable and discarding others:

4. Validation of the approach

The function FAMD() [from FactoMiner] can be used to compute FAMD. Format:
* base : a data frame with n rows (individuals) and p columns (variables). * ncp: the number of dimensions kept in the results (by default 5) * sup.var: a vector indicating the indexes of the supplementary variables. * ind.sup: a vector indicating the indexes of the supplementary individuals. * graph : a logical value. If TRUE a graph is displayed.

The grouping of columns have been done for easy selection during the analysis. Iterative FAMD using different columns have resulted in the progressive deletion of these columns from the analysis and finally using only a fraction of them.

```
#Remove cols_anamnesis, cols_ECG and Target Var
df_dependents = subset(data.rose, select = -c( GB, SIM_GIPERT, DLIT_AG, ZSN_A, nr_11, nr_01, nr_02, nr_

#Remove above PLUS cols_time_to_hosp, cols_relapse_pain
df_dependents = subset(df_dependents, select = -c(TIME_B_S, R_AB_1_n, R_AB_2_n, R_AB_3_n))

# Remove above PLUS cols_hypokalemia, cols_condition_ICU
df_dependents = subset(df_dependents, select = -c(GIPO_K, O_L_POST, K_SH_POST, MP_TP_POST, SVT_POST, GT_

# Remove above PLUS cols_opioids, cols_fibrinolytic_therapy
df_dependents = subset(df_dependents, select = -c(fibr_ter_01, fibr_ter_02, fibr_ter_03, fibr_ter_05, fi

# Remove GIPER_Na
df_dependents = subset(df_dependents, select = -c(GIPER_NA))

# Remove D_AD_ORIT
df_dependents = subset(df_dependents, select = -c(D_AD_ORIT))

#Remove Blood Tests:
df_dependents = subset(df_dependents, select = -c(K_BLOOD, NA_BLOOD, ALT_BLOOD, AST_BLOOD, L_BLOOD, ROE

#Remove Quantity of myocardial infarctions in the anamnesis (INF_ANAM):
df_dependents = subset(df_dependents, select = -c(INF_ANAM))

#Remove Functional class (FC) of angina pectoris in the last year (FK_STENOK)
df_dependents = subset(df_dependents, select = -c(FK_STENOK))
```

These columns were selected in a best-guess manner as discussed later.

```
head(df_dependents)
```

```
##        AGE SEX STENOK_AN IBS_POST S_AD_ORIT
## 1 63.39332   M         6        1  179.9379
## 2 66.71402   M         3        2  139.1958
## 3 76.49845   M         5        2   77.5382
## 4 68.36027   F         4        1  112.9787
## 5 42.06832   M         0        0  125.9443
## 6 63.01472   M         3        2  161.8574
```

Performing the clustering:

```
res.famd <- FAMD(df_dependents, ncp = 5, graph = FALSE)
print(res.famd)
```

```
## *The results are available in the following objects:
##
##   name          description
## 1 "$eig"        "eigenvalues and inertia"
## 2 "$var"        "Results for the variables"
## 3 "$ind"        "results for the individuals"
## 4 "$quali.var"  "Results for the qualitative variables"
## 5 "$quanti.var" "Results for the quantitative variables"
```

get_famd_ind(res.famd): Extract the results for individuals. get_famd_var(res.famd): Extract the results for quantitative and qualitative variables. fviz_famd_ind(res.famd), fviz_famd_var(res.famd): Visualize the results for individuals and variables, respectively.

Eigenvalues / Variances

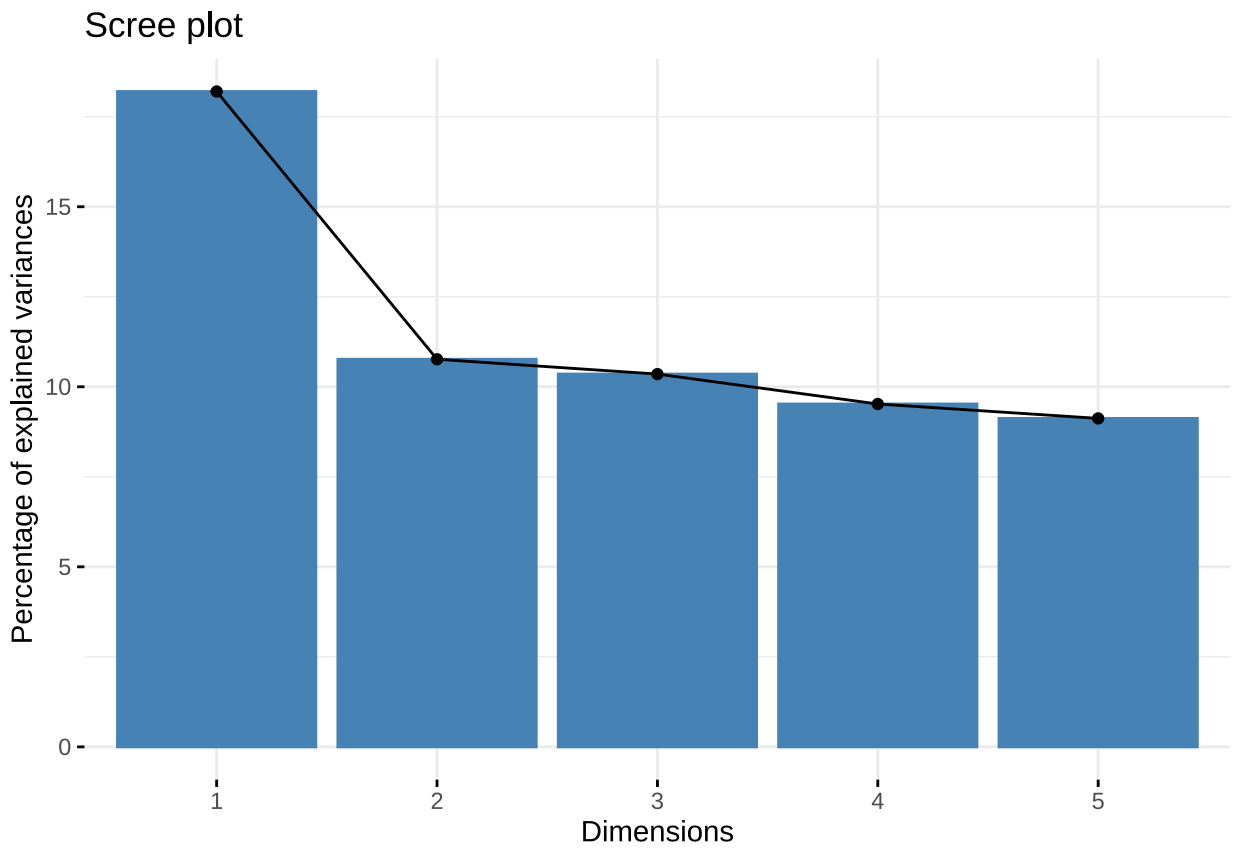
The proportion of variances retained by the different dimensions (axes) using the function get_eigenvalue():

```
#Extract the eigenvalues/variances retained by each dimension (axis).
eig.val <- get_eigenvalue(res.famd)
head(eig.val)
```

```
##       eigenvalue variance.percent cumulative.variance.percent
## Dim.1   2.001965        18.199679                    18.19968
## Dim.2   1.184056        10.764150                    28.96383
## Dim.3   1.138792        10.352650                    39.31648
## Dim.4   1.047398         9.521802                    48.83828
## Dim.5   1.003400         9.121814                    57.96010
```

Plotting Scree:

```
# Scree plot (the percentages of inertia explained by each FAMD dimensions)
fviz_screeplot(res.famd)
```



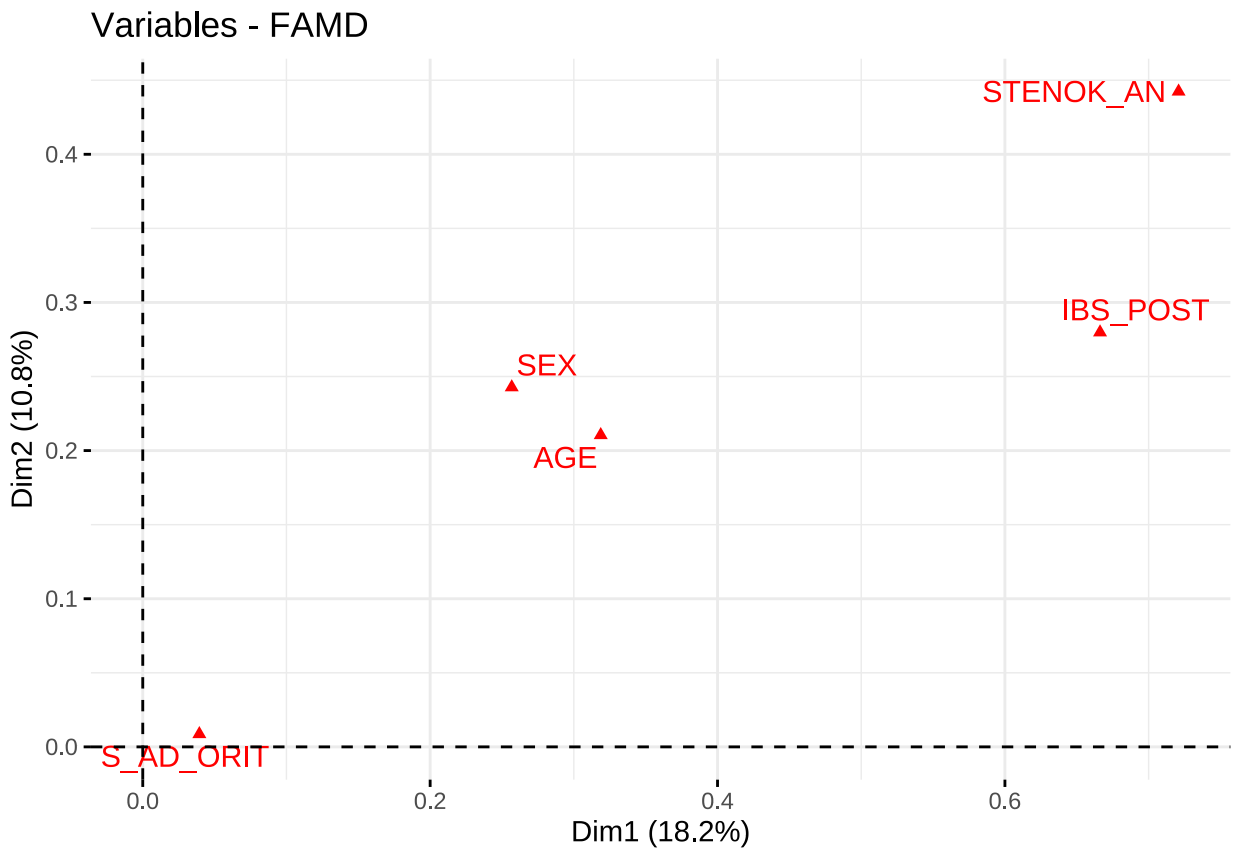
The elbow occurs at 3rd dimension.

Plotting dimensions: The function get_mfa_var() is used to extract the results for variables. By default, this function returns a list containing the coordinates, the cos2 and the contribution of all variables:

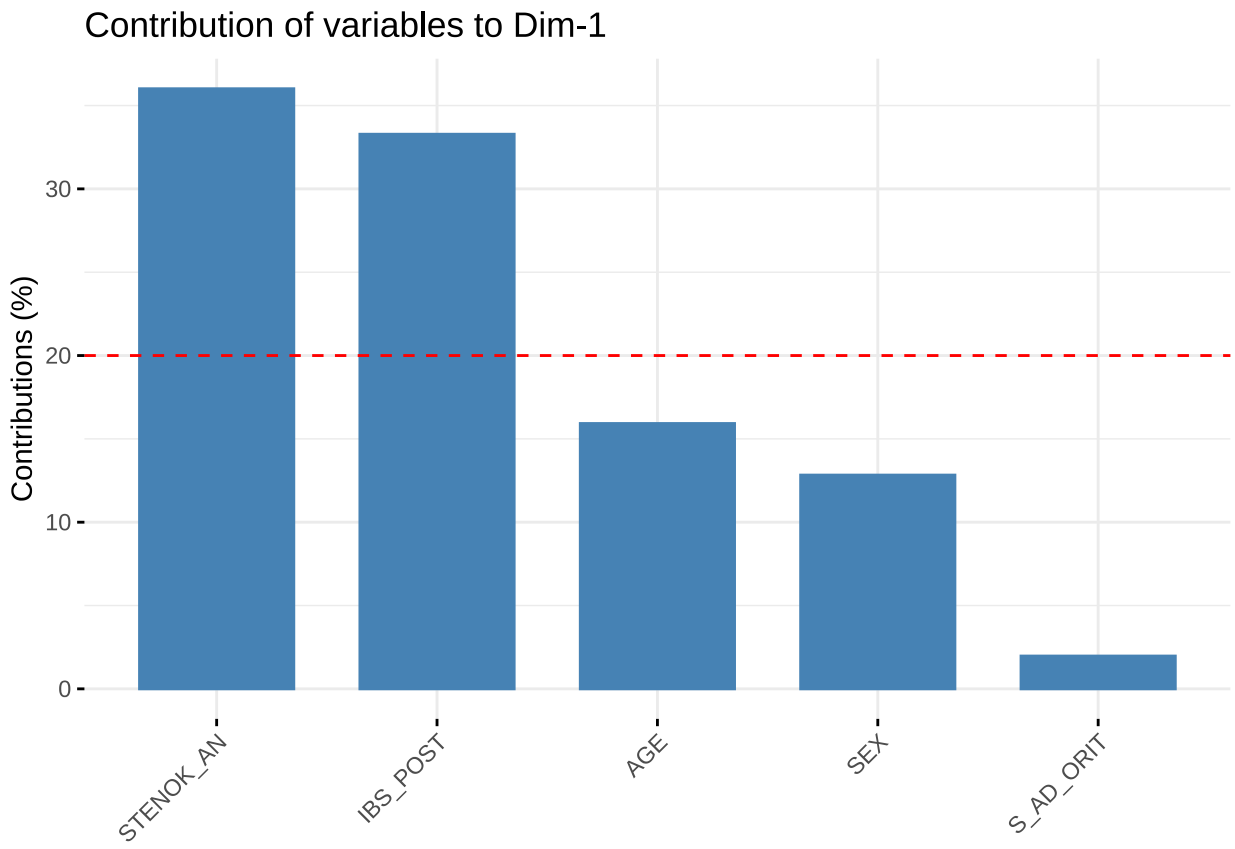
```
var <- get_famd_var(res.famd)
# Access components using head(var$coord/var$cos2/var$contrib)
```

To show the correlation between variables - both quantitative and qualitative variables - and the principal dimensions, as well as, the contribution of variables to the dimensions 1 and 2:

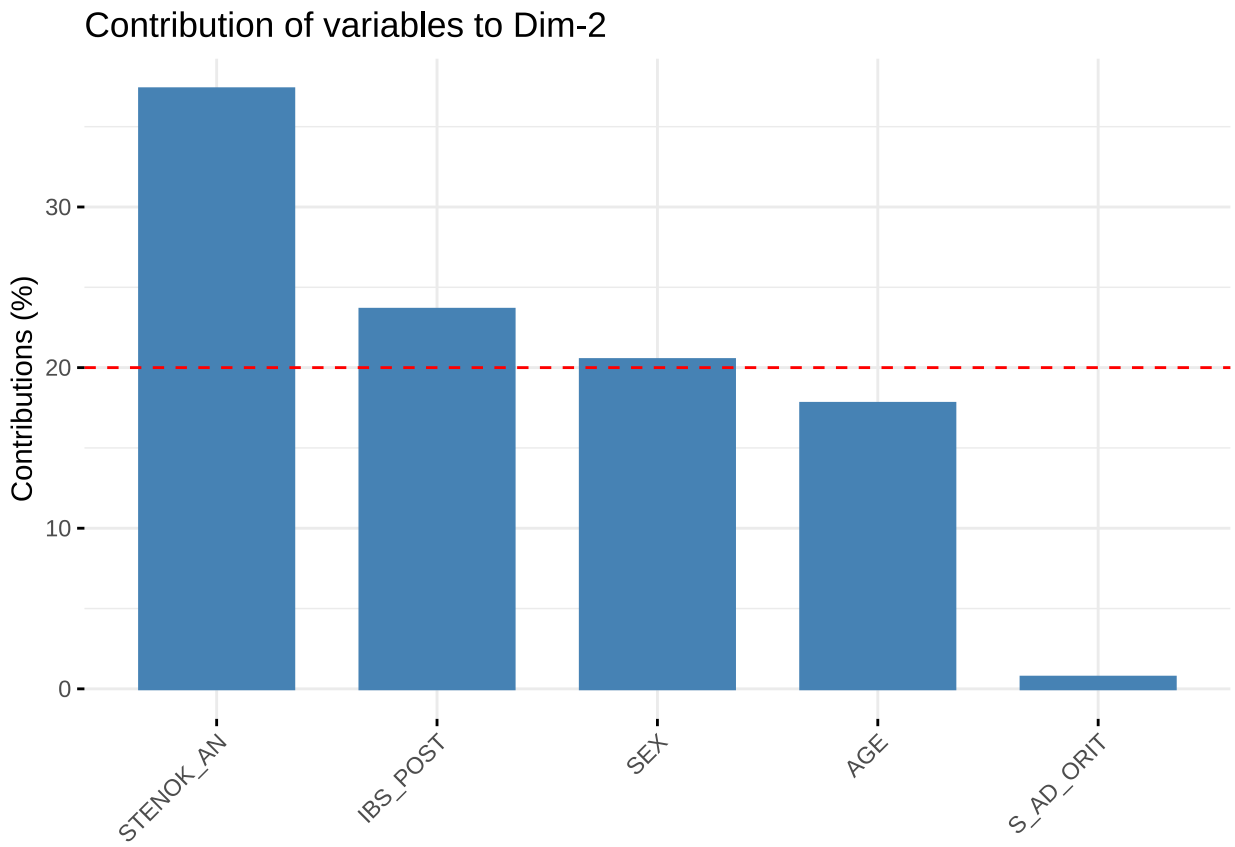
```
# plot both quantitative and qualitative variables
fviz_famd_var(res.famd, repel = TRUE)
```




```
# Contribution to the first dimension
fviz_contrib(res.famd, "var", axes = 1)
```




```
# Contribution to the second dimension
fviz_contrib(res.famd, "var", axes = 2)
```



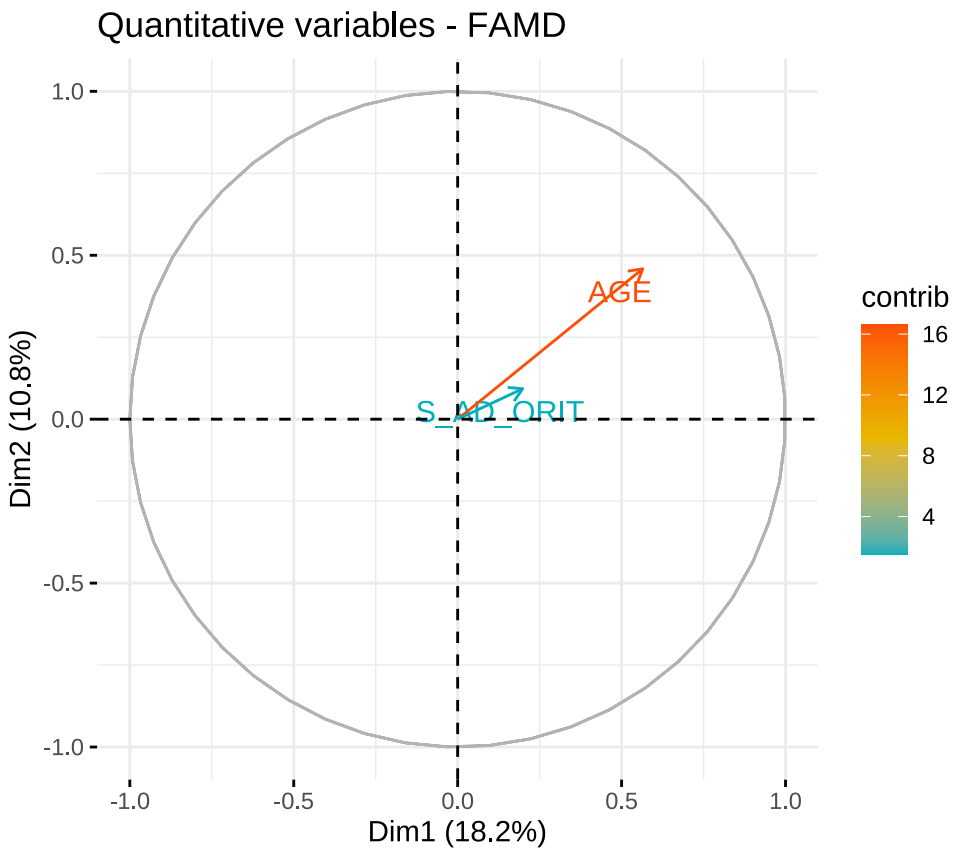
5. Interpretation of the results

An eigenvalue >1 indicates that the PD accounts for more variance than one of the original variables in standardized data. This is commonly used as a cutoff point for which PDs are retained to be used in further analysis. The 5 dimensions chosen together they account for only 58.6% of the total variance in the data set. This suggests that this data-set is quite complex, potentially due to:
* 1) the relationships between the variables being non-linear, and/or * 2) some factors (variables) that can account for variance in this data-set are not included in this analysis.

Contributions to individual dimensions and their graphs:

a) Quantitative Variables Contribution Graph

```
quanti.var <- get_famd_var(res.famd, "quanti.var")
fviz_famd_var(res.famd, "quanti.var", col.var = "contrib",
              gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
              repel = TRUE)
```



Observations:

```
head(var$contrib)
```

```
##               Dim.1      Dim.2     Dim.3      Dim.4       Dim.5
## AGE       15.920548 17.7752886  0.221603  3.1227670  0.50141405
## S_AD_ORIT  1.967747  0.7295267 25.147910 10.3819464  1.38734809
## SEX       12.826674 20.5015326  6.647340  0.3000141  0.31759502
## STENOK_AN 36.007668 37.3614033 52.400008 64.7310061 97.74158957
## IBS_POST  33.277362 23.6322488 15.583139 21.4642664  0.05205326
```

Contributions across categorical variables:

```
quali.var <- get_famd_var(res.famd, "quali.var")
head(quali.var$contrib, 15)
```

```
##        Dim.1      Dim.2       Dim.3      Dim.4        Dim.5
## F  7.8187785 12.4971559  4.05203087  0.1828801 1.935970e-01
## M  5.0078953  8.0043767  2.59530903  0.1171340 1.239981e-01
## 0 22.5111079  5.3388801  1.17303854  0.1848874 3.003926e-02
## 1  1.0452874 21.5906048  4.66517599 12.0781137 1.302218e+01
## 2  1.7965604  2.9089620  7.69730658 27.9108881 1.610142e+01
## 3  0.8528175  0.9891297  0.01616847 17.0299903 5.242495e+01
```

```
## 4  2.0383203  0.7662235 10.40820727  0.1589640 1.137674e+01
## 5  2.4789049  5.5647922  0.26882721  4.1834830 7.817686e-01
## 6  5.2846701  0.2028110 28.17128378  3.1846797 4.004505e+00
## 0 22.6042882  2.3515026  6.13267277  4.3960907 2.779685e-02
## 1 10.4924013 15.9529997  1.86847901  5.7636057 2.389265e-02
## 2  0.1806726  5.3277465  7.58198716 11.3045700 3.637643e-04
```

- Typically the first dimension has a contribution of more than 30%. But here the top contribution is only 17.5%. When EVERY variable was selected, the contribution was abysmal, it varied between 2-5% for almost every independent variable. After removing many spurious data columns that were grouped as was done above before performing FAMD, we could obtain columns whose contributions are now these. These 5 dimensions/components can be used in regression.
- Dimension 1 (~17.5%): Has most contribution by Exertional Angina Pectoris event in previous years(STENOK_AN, with 7 levels [0, 6]) and CHD event occurring in the previous weeks (IBS_POST with 3 levels [0, 2]). Cross referencing the contribution across levels of categorical variables, this means that the occurrence of STENOK_AN(value = 0) and IBS_POST(value = 0) contribute most, i.e. not having a Angina Pectoris or CHD event previously is the most important factor in relapse of MI event (target variable) after hospital admission.
- Dimension 2(~11.5%): Has most contribution by AGE and Gender(SEX). Cross referencing the contribution across levels of categorical variables, this means that the occurrence of SEX(value = F) contribute most, i.e. being a female patient and their age is the most factor in determining the 2nd dimension in relapse of MI event (target variable) after hospital admission. Probably being Female and younger in age would help in not having a Relapse of MI event after admitting to hospital.

Inference: Logistic Regression

1. Approach

The target variable REC_IM(Relapse of the Myocardial infarction) is a binary and hence we would require to perform Logistic Regression, which also takes care of independent categorical variables.

2. Formulation

Selecting the columns:

```
df_dependents = subset(data.rose, select = c(REC_IM, AGE, S_AD_ORIT, SEX, STENOK_AN, IBS_POST))
```

Fitting the Model

```
# Fit the model
model <- glm( REC_IM ~AGE + S_AD_ORIT + SEX + STENOK_AN + IBS_POST, data = df_dependents, family = bino
```

3. Checks of assumptions

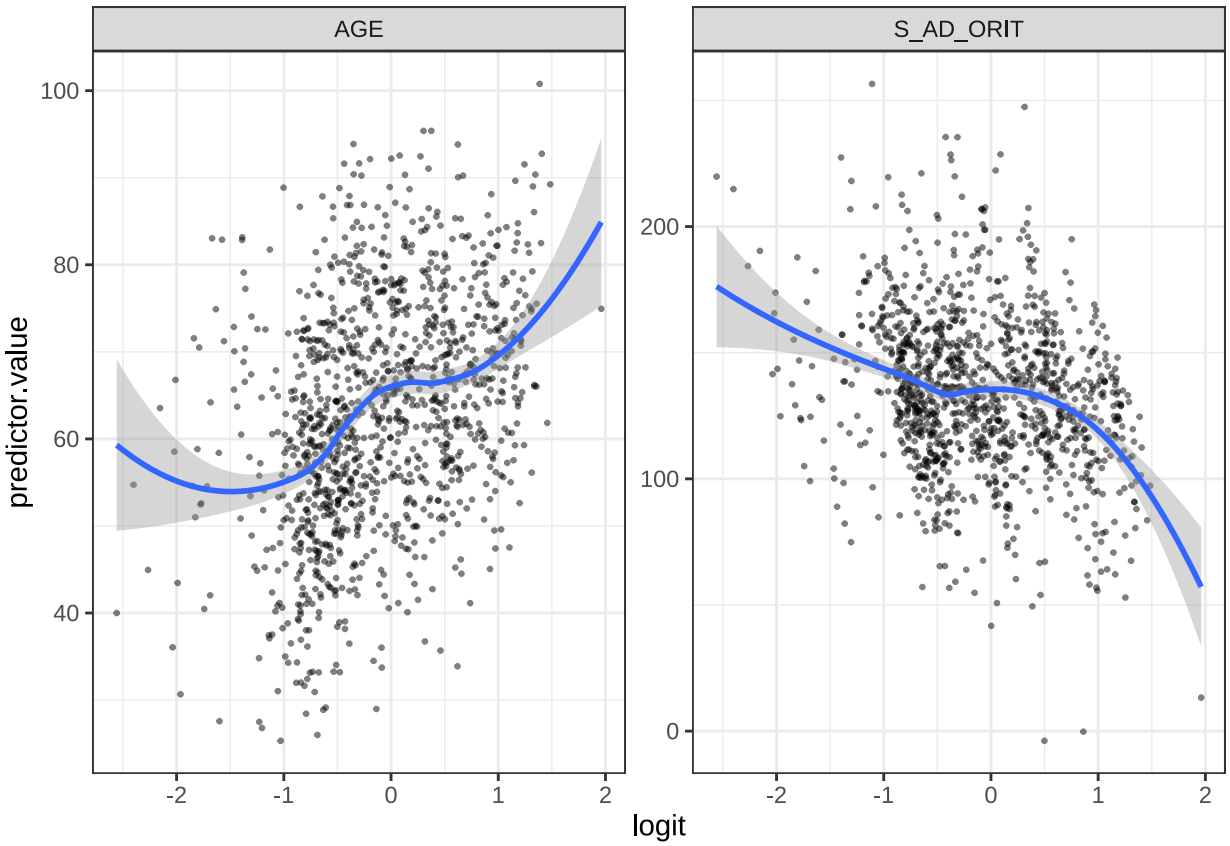
Linearity assumption Linear relationship between continuous predictor variables and the logit of the outcome can be done by visually.

```
probabilities <- predict(model, type = "response")

mydata <- df_dependents %>%
  dplyr::select_if(is.numeric)
predictors <- colnames(mydata)
# Bind the logit and tidying the data for plot
mydata <- mydata %>%
  mutate(logit = log(probabilities/(1-probabilities))) %>%
  gather(key = "predictors", value = "predictor.value", -logit)
```

Scatter plots:

```
ggplot(mydata, aes(logit, predictor.value))+
  geom_point(size = 0.5, alpha = 0.5) +
  geom_smooth(method = "loess") +
  theme_bw() +
  facet_wrap(~predictors, scales = "free_y")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

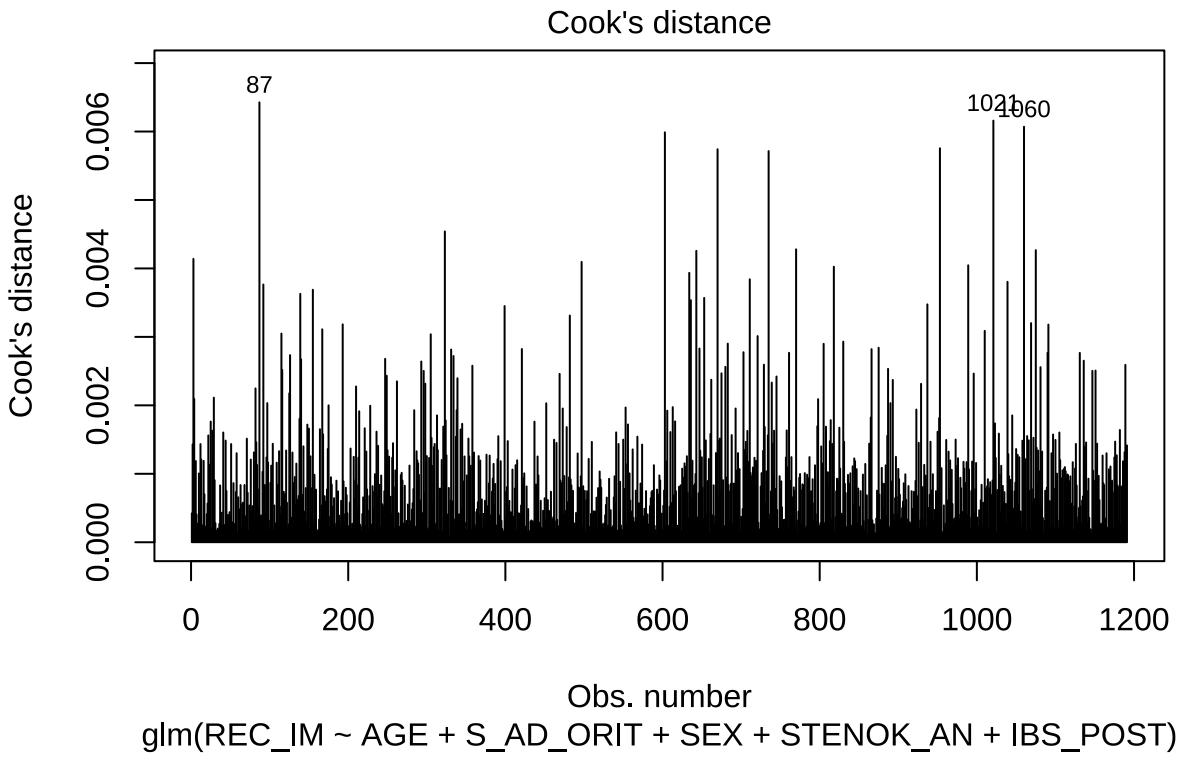


Conclusion: They are reasonably linearly related.

Influential values

Influential values are extreme individual data points that can alter the quality of the logistic regression model and can be visualized using the Cook's distance values. Labeling the top 3 largest values:

```
plot(model, which = 4, id.n = 3)
```



Note that, not all outliers are influential observations. But most don't seem to be outliers.

Multicollinearity

Multicollinearity corresponds to a situation where the data contain highly correlated predictor variables. It can be assessed using the R function vif() [car package], which computes the variance inflation factors:

```
car::vif(model)
```

```
##                GVIF Df GVIF^(1/(2*Df))
## AGE        1.181572  1        1.087001
## S_AD_ORIT  1.035294  1        1.017494
## SEX        1.158341  1        1.076263
## STENOK_AN  2.292831  6        1.071596
## IBS_POST   2.142338  2        1.209823
```

A VIF value that exceeds 5 or 10 indicates a problematic amount of collinearity. In our example, there is no collinearity: all variables have a value of VIF well below 5.

4. Validation of the approach

```
# Make predictions
probabilities <- model %>% predict(test, type = "response")
predicted.classes <- ifelse(probabilities > 0.5, "1", "0")
# Model accuracy
mean(predicted.classes == test$REC_IM)
```

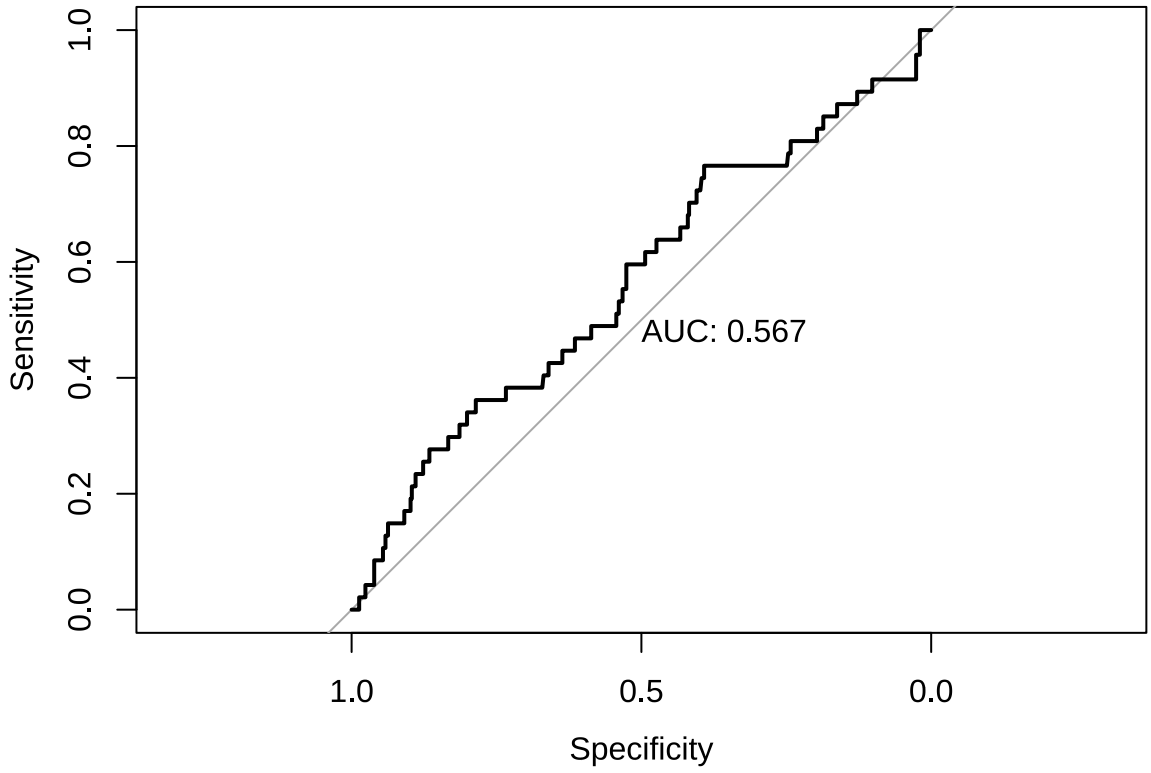
```
## [1] 0.6561886
```

The model has a mean accuracy of 66.79%. But being imbalanced, other measures of accuracy are needed:

```
test_roc = roc(test$REC_IM ~ probabilities, plot = TRUE, print.auc = TRUE)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```




```
as.numeric(test_roc$auc)
```

```
## [1] 0.5667081
```

The performance value of AUC is 0.612, which tells that our model can distinguish between patients with and without a heart disease with a probability of 61.2%. From the curve plotted, the value of 0.6 is good enough to determine true positive rate, being the topmost/leftmost point on the curve. Using that AUC value to determine new accuracy:

```
accuracy_probs <- predict(model, test, type = "response")
predicted.classes <- ifelse(accuracy_probs > 0.6, "1", "0")
mean(predicted.classes == test$REC_IM)
```

```
## [1] 0.7937132
```

Hence, the model has an accuracy of 79.8%.

5. Interpretation of the results

```
summary(model)
```

```
##
## Call:
## glm(formula = REC_IM ~ AGE + S_AD_ORIT + SEX + STENOK_AN + IBS_POST,
##     family = binomial, data = df_dependents)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0456  -1.0348  -0.6678   1.0909   1.9999
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.080055   0.445585   0.180 0.857418
## AGE          0.008305   0.005107   1.626 0.103925
## S_AD_ORIT   -0.007084   0.002031  -3.488 0.000486 ***
## SEXM        -0.369604   0.135130  -2.735 0.006235 **
## STENOK_AN1   0.959624   0.268115   3.579 0.000345 ***
## STENOK_AN2  -0.442824   0.336971  -1.314 0.188802
## STENOK_AN3   0.519544   0.297453   1.747 0.080700 .
## STENOK_AN4   0.946719   0.319224   2.966 0.003020 **
## STENOK_AN5   1.329881   0.277098   4.799 1.59e-06 ***
## STENOK_AN6   1.226229   0.203448   6.027 1.67e-09 ***
## IBS_POST1   -0.601539   0.252796  -2.380 0.017334 *
## IBS_POST2    0.126048   0.201528   0.625 0.531667
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1648.3  on 1190  degrees of freedom
## Residual deviance: 1519.2  on 1179  degrees of freedom
## AIC: 1543.2
##
## Number of Fisher Scoring iterations: 4
```

It can be seen that only 5 out of the 11 predictors are significantly associated to the outcome. These include: Age, Exertional angina pectoris event more than 4 years ago (STENOK_AN5, STENOK_AN6), CHD event of exertional angina pectoris (IBS_POST1), Systolic blood pressure according to intensive care unit (S_AD_ORIT) and Age

The coefficient estimate of the variable Age = 0.014748, which is positive. This means that an increase in Age is associated with increase in the probability of having a relapse of MI after admission. Surprisingly, CHD event of exertional angina pectoris (IBS_POST1) is negative, meaning the chances of relapse decreases if the event happened in the last few weeks, surprisingly.

Conclusion

- The variables needed to predict the relapse in MI conditions after admitting to hospital are relatively small in number like age, gender, the Blood Pressure at entry and CHD event previously etc.
- The use of extensive notes at the time of admission concerning different conditions could not be very helpful given the formatting of the indicators was mostly categorical. Of the 112 variables only 12 continuous out of which 2 were mostly empty. But, despite that the logistical model built on top of the variables having the highest contribution had a decent accuracy.

- I learnt that PCA was only applicable to continuous variables and if I had to extend other kinds of variables, I would have to use techniques like FAMD which involve hot-coding the categorical variables.
- I also learnt that, despite cleaning the data and controlling for imbalance in data-set, the extended-PCA may not work effectively and the interpretations have to be carefully extended.

Limitations of your study and ideas for possible future research.

- The primary limitation is the structure of data-input; the presence of huge number of categorical variables made it particularly harder to study. The results from PCA (FAMD) were mixed. Possible reasons could be that data is not following the Multi-normial distribution, on which PCA works best on or covariate/variable with a significant effect was excluded from the variable & noise reduction procedure.
- Gathering more data and working on larger number of variables could lead to better results.

References

Articles from: http://www.sthda.com were used for the analysis of FAMD and Logistic Regression (with assumptions.)