Statistical Learning Project

Filippo Santin, Gurjeet Singh, Francesca Zen

18/5/2021

Introduction

In the following report we present the analysis computed on a particular disease, stroke, and its correlation to other factors such as smoking, glucose level, bmi and so on.

"Stroke" is the medical term for damage to brain tissue or the death of a portion of it, due to insufficient blood supply to an area of the brain.

Our aim is to see if and how the variables we are dealing with are related to each other, in order to predict which individual is more probable to get a stroke.

The symptoms of stroke vary from patient to patient, depending on the severity of the condition, the affected brain area, causes, type of stroke, etc.

Stroke is characterized by sudden onset and for this reason it involves the need for immediate therapeutic intervention and adapted to the needs of the patient. In this sense, looking for relation between features may help to prevent it.

In order to have a guide for the interpretation of the data we underline the following information:

- The normal values of glucose level are between 60 and 110 mg/dl and with a value greater than 126 mg/dl a person is considered diabetic;
- a body mass index (BMI) between 18.5-24.9 indicates a normal/healthy weight, below 18.5 indicates underweight, 25.0-29.9 indicates overweight and above 30.0 indicates obese person

Exploiting the Dataset

The dataset we used was provided by kaggle [1] and it is composed of 5,110 entries with a total of 12 columns: `id`, `gender`, `age`, `hypertension`, `heart_disease`, `ever_married`, `work_type`, `Residence_type`, `avg_glucose_level`, `bmi`, `smoking_status`, `stroke`.

```
stroke_data <- read.csv('healthcare-dataset-stroke-data.csv')
attach(stroke_data)
```

The preliminary part of the analysis focuses on the study of the dataset: we looked at the `id` column and verified that all the data collected was referring to different people, so no recidivist status were involved.

```
stroke_data<-stroke_data[,-1]
```

In order to use the variables through the analysis we then transformed the categorical variables into factors:

```
stroke_data$gender<- as.factor(gender)
stroke_data$ever_married<-as.factor(ever_married)
stroke_data$work_type<-as.factor(work_type)
```

[1] https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

```
stroke_data$Residence_type<-as.factor(Residence_type)
stroke_data$smoking_status<-as.factor(smoking_status)
```

What's more, the variable `bmi` was not numeric because of the presence of a string "N/A" which identifies the lack of the value, and so we transformed it using the command `stroke_data$bmi <- as.numeric(bmi)`; and then removed those NA values:`stroke_data<- na.omit(stroke_data)`.

```
## Warning: NA introdotti per coercizione
```

We ended up having 4,909 entries and 11 total columns. Here we give a quick overview of the main information about the dataset:

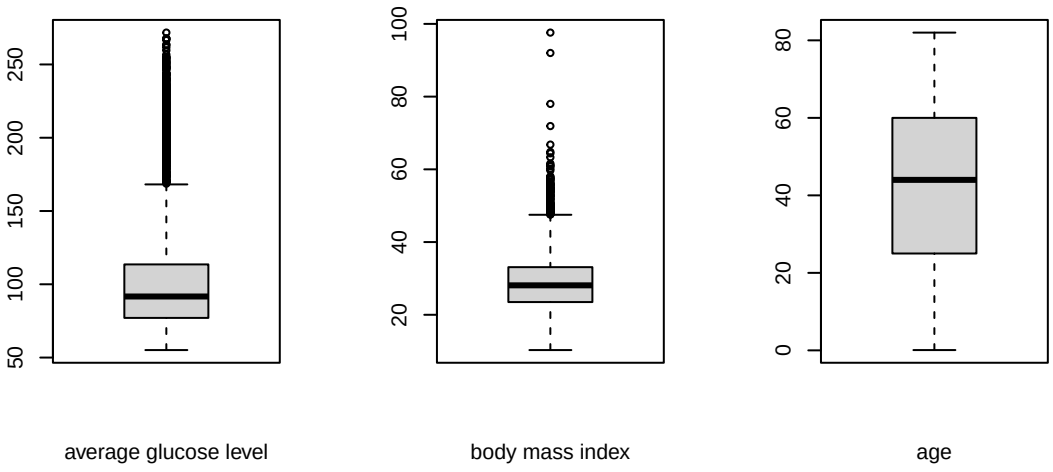
```
summary(stroke_data)
```

```
##     gender          age         hypertension     heart_disease    ever_married
##  Female:2897   Min.   : 0.08   Min.   :0.00000   Min.   :0.0000   No :1705
##  Male  :2011   1st Qu.:25.00   1st Qu.:0.00000   1st Qu.:0.0000   Yes:3204
##  Other :   1   Median :44.00   Median :0.00000   Median :0.0000
##                Mean   :42.87   Mean   :0.09187   Mean   :0.0495
##                3rd Qu.:60.00   3rd Qu.:0.00000   3rd Qu.:0.0000
##                Max.   :82.00   Max.   :1.00000   Max.   :1.0000
##          work_type    Residence_type avg_glucose_level      bmi
##  children     : 671   Rural:2419     Min.   : 55.12    Min.   :10.30
##  Govt_job     : 630   Urban:2490     1st Qu.: 77.07    1st Qu.:23.50
##  Never_worked :  22                  Median : 91.68    Median :28.10
##  Private      :2811                  Mean   :105.31    Mean   :28.89
##  Self-employed: 775                  3rd Qu.:113.57    3rd Qu.:33.10
##                                      Max.   :271.74    Max.   :97.60
##          smoking_status     stroke
##  formerly smoked: 837   Min.   :0.00000
##  never smoked   :1852   1st Qu.:0.00000
##  smokes         : 737   Median :0.00000
##  Unknown        :1483   Mean   :0.04257
##                         3rd Qu.:0.00000
##                         Max.   :1.00000
```

```
attach(stroke_data)
```

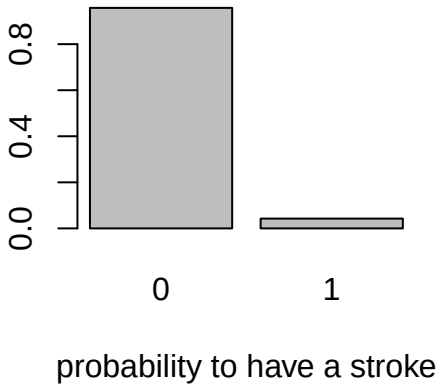
A visual transformation of these values is provided in the following boxplots:

```
par(mfrow=c(1,3))
boxplot(avg_glucose_level, xlab= 'average glucose level' )
boxplot(bmi, xlab = 'body mass index')
boxplot(age, xlab = 'age',pch=20)
```



Through the analysis on the Stroke dataset, we discovered that it was strongly bias, in the sense that 209 people on a total of 4909 get a stroke:

```
barplot(table(stroke)/dim(stroke_data)[1],
        xlab='probability to have a stroke')
```



This value is representative of the real situation in which there are not many stroke cases compared with the whole population. In Italy, for example, we have 200,000 cases over 59.226.539 people, i.e. 0.33%.

Searching for relationships

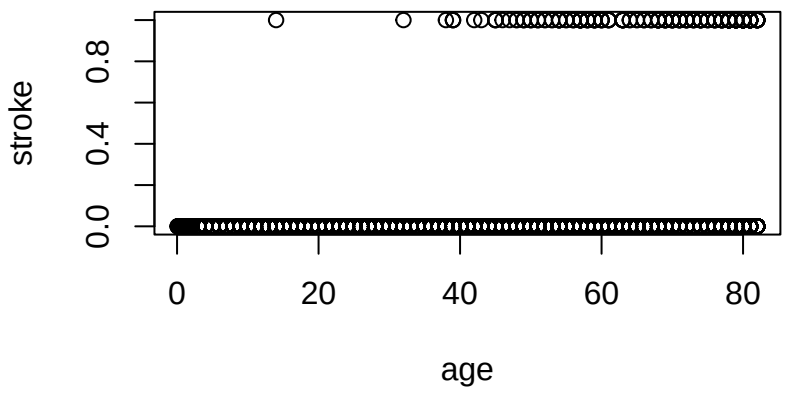
At this point we can ask some questions:

- Is it possible to prevent ictus?
- Which factors are the most related to it?
- How strong are the relations between the features?

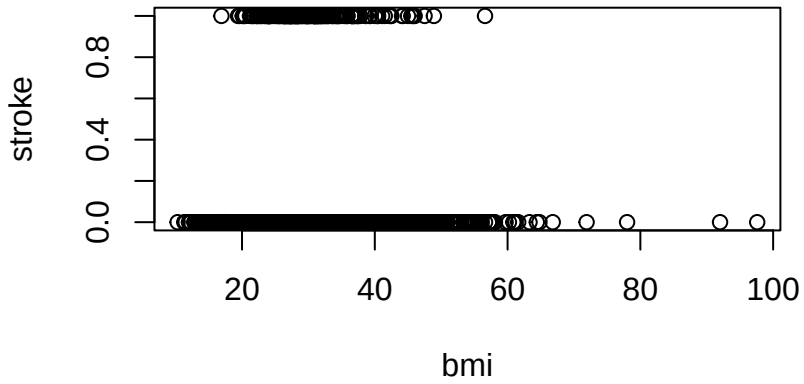
We will explore the data trying to answer them.

First of all we look at some intuitive relation between `stroke` and `age`,`bmi` and `avg_glucose_level`:

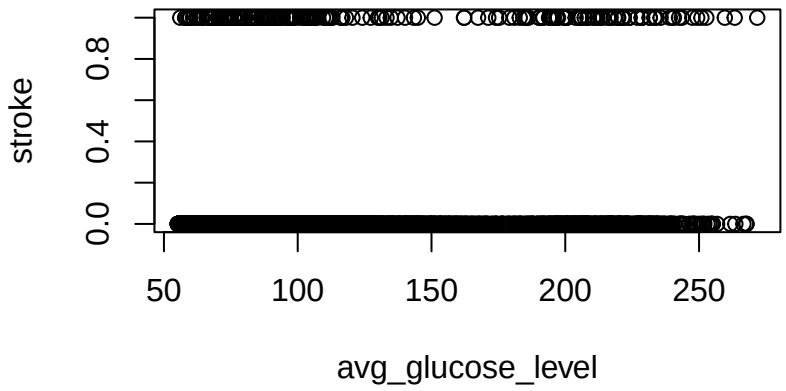
```
plot(stroke~age)
```




```
plot(stroke~bmi)
```



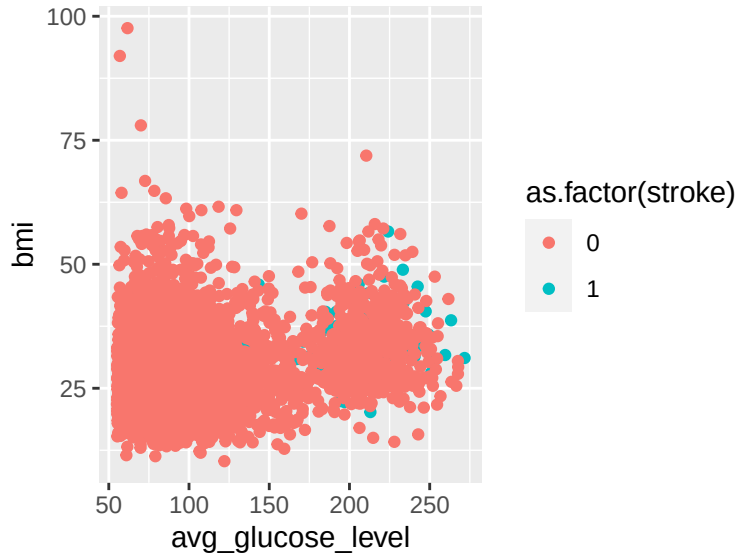

```
plot(stroke~avg_glucose_level)
```



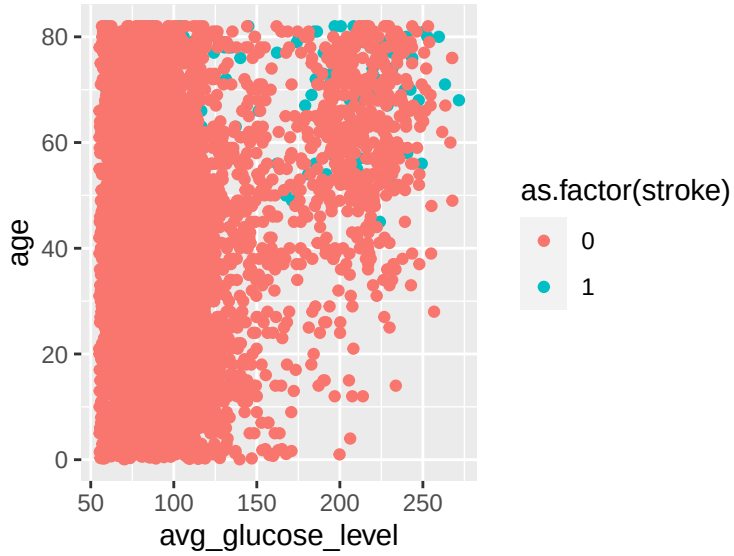
Looking to these plots we can see that the incidence of the virus increases progressively with age and that if we sum also the information about `avg_glucose_level` we may wonder if diabetic people are more probable to get a stroke or not. In addition there is no apparent relation of `stroke` with `bmi`.

We now highlight other visual relationship thanks to the scatter plots:

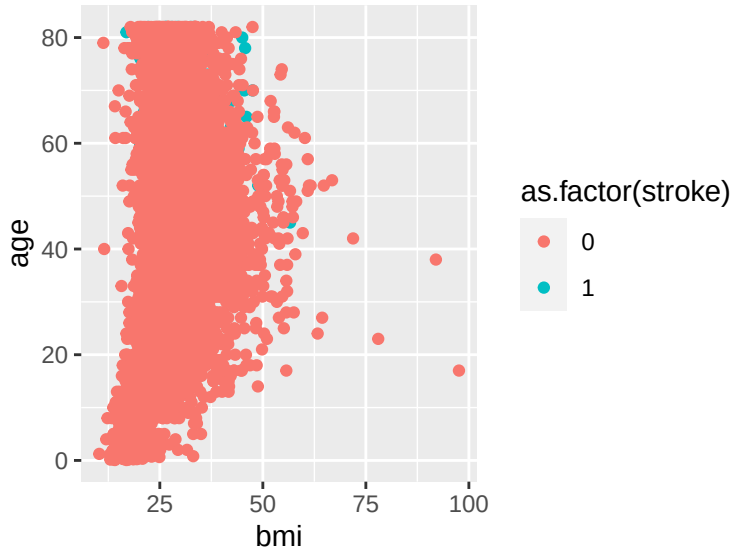
```
library(ggplot2)
par(mfrow=c(1,3))
ggplot(stroke_data, aes(x = avg_glucose_level, y = bmi,
                        col = as.factor(stroke))) + geom_point()
```




```
ggplot(stroke_data, aes(x = avg_glucose_level, y = age,
                        col = as.factor(stroke))) + geom_point()
```




```
ggplot(stroke_data, aes(x = bmi, y = age,
                        col = as.factor(stroke))) + geom_point()
```




```
par(mfrow=c(1,1))
```

discussione

Analysis on models

Full and Reduced Models

```
mod.full <- glm(stroke~., data=stroke_data, family = binomial)
summary(mod.full)
```

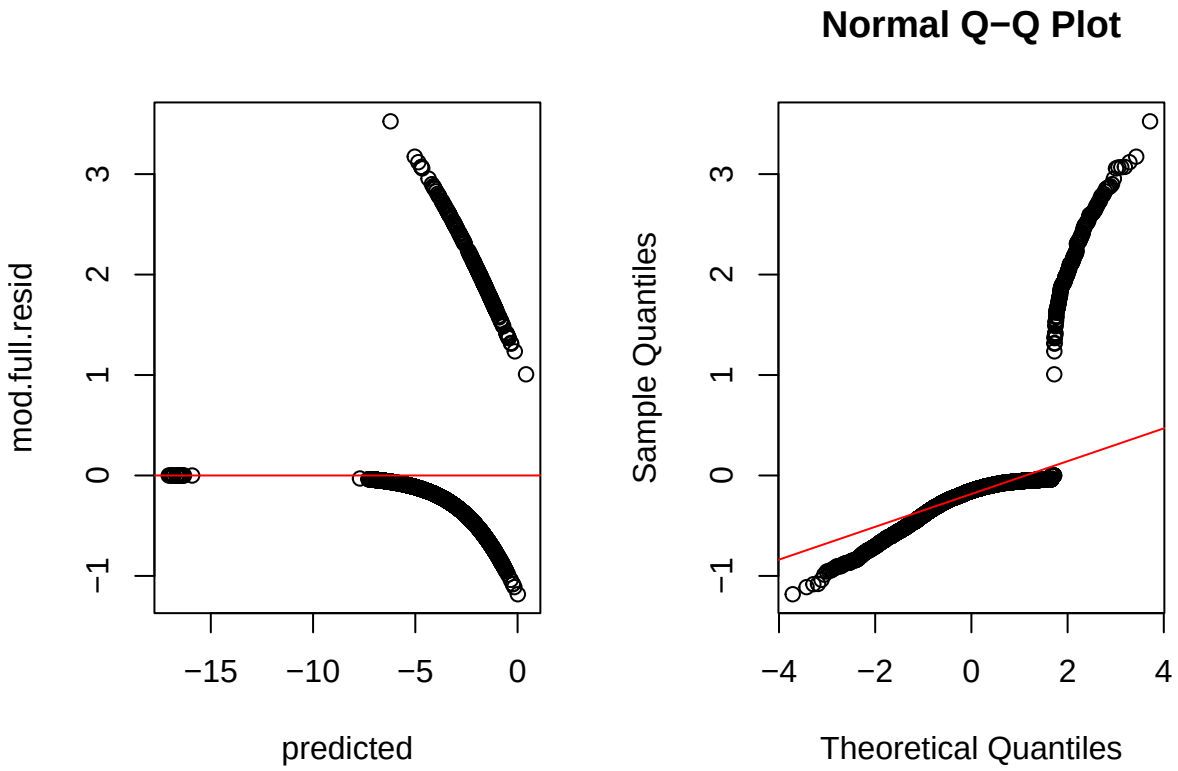
```
##
## Call:
```

```
## glm(formula = stroke ~ ., family = binomial, data = stroke_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1823  -0.2947  -0.1524  -0.0744   3.5251
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -7.360e+00  1.067e+00  -6.895 5.37e-12 ***
## genderMale                  -1.463e-02  1.544e-01  -0.095 0.924525
## genderOther                 -1.135e+01  2.400e+03  -0.005 0.996225
## age                          7.348e-02  6.347e-03  11.578  < 2e-16 ***
## hypertension                 5.249e-01  1.750e-01   2.999 0.002711 **
## heart_disease                3.488e-01  2.072e-01   1.683 0.092381 .
## ever_marriedYes             -1.152e-01  2.473e-01  -0.466 0.641394
## work_typeGovt_job           -6.817e-01  1.114e+00  -0.612 0.540660
## work_typeNever_worked       -1.082e+01  5.090e+02  -0.021 0.983036
## work_typePrivate            -5.208e-01  1.100e+00  -0.473 0.635943
## work_typeSelf-employed      -9.459e-01  1.119e+00  -0.845 0.397906
## Residence_typeUrban          4.514e-03  1.500e-01   0.030 0.975990
## avg_glucose_level            4.652e-03  1.294e-03   3.595 0.000324 ***
## bmi                          4.062e-03  1.188e-02   0.342 0.732387
## smoking_statusnever smoked -6.722e-02  1.886e-01  -0.356 0.721556
## smoking_statussmokes         3.139e-01  2.295e-01   1.368 0.171310
## smoking_statusUnknown       -2.753e-01  2.471e-01  -1.114 0.265193
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1728.4  on 4908  degrees of freedom
## Residual deviance: 1363.2  on 4892  degrees of freedom
## AIC: 1397.2
##
## Number of Fisher Scoring iterations: 15
```

Here we see that `age` and `hypertension` are the variables most related to `stroke`.

Let's use the residual plots to get more information about this model:

```
mod.full.resid <- residuals(mod.full, type="deviance") # because we have a binary response
predicted <- predict(mod.full, type = "link")
par(mfrow=c(1,2))
plot(mod.full.resid~predicted)
abline(h=0, col='red')
qqnorm(mod.full.resid)
qqline(mod.full.resid, col='red')
```



Comparison between Reduced Models

```
# Reduced model 1 : We remove at least all the features that have collinearity between
# each other (work_type, ever_married) and the residence type
mod.red1 <- glm(stroke ~ age + bmi + avg_glucose_level + hypertension + smoking_status +
                  gender, family=binomial)
summary(mod.red1)
```

```
##
## Call:
## glm(formula = stroke ~ age + bmi + avg_glucose_level + hypertension +
##     smoking_status + gender, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0138  -0.2991  -0.1569  -0.0715   3.6908
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -7.915289   0.587035 -13.484  < 2e-16 ***
## age                          0.070996   0.005764  12.318  < 2e-16 ***
## bmi                          0.002695   0.011699   0.230 0.817837
## avg_glucose_level            0.004915   0.001282   3.834 0.000126 ***
## hypertension                 0.527714   0.173926   3.034 0.002412 **
## smoking_statusnever smoked  -0.070206   0.187586  -0.374 0.708209
## smoking_statussmokes         0.343917   0.227595   1.511 0.130766
## smoking_statusUnknown       -0.256575   0.244806  -1.048 0.294603
```

```
## genderMale                    0.017558   0.152828   0.115 0.908533
## genderOther                  -7.261569 324.743861  -0.022 0.982160
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1728.4  on 4908  degrees of freedom
## Residual deviance: 1372.5  on 4899  degrees of freedom
## AIC: 1392.5
##
## Number of Fisher Scoring iterations: 11
```

```
# Reduced model 2: remove gender
mod.red2 <- glm(stroke ~ age + bmi + avg_glucose_level + hypertension + smoking_status, family=binomial)
summary(mod.red2)
```

```
##
## Call:
## glm(formula = stroke ~ age + bmi + avg_glucose_level + hypertension +
##     smoking_status, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0108  -0.2983  -0.1571  -0.0716   3.6890
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -7.906885   0.580996 -13.609  < 2e-16 ***
## age                           0.070991   0.005760  12.325  < 2e-16 ***
## bmi                           0.002663   0.011686   0.228 0.819754
## avg_glucose_level             0.004928   0.001277   3.859 0.000114 ***
## hypertension                  0.527864   0.173922   3.035 0.002405 **
## smoking_statusnever smoked -0.073031   0.185814  -0.393 0.694294
## smoking_statussmokes          0.343591   0.227529   1.510 0.131019
## smoking_statusUnknown        -0.258084   0.244363  -1.056 0.290900
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1728.4  on 4908  degrees of freedom
## Residual deviance: 1372.5  on 4901  degrees of freedom
## AIC: 1388.5
##
## Number of Fisher Scoring iterations: 7
```

```
# Final Reduced Model
mod.red <- glm(stroke~age + bmi + avg_glucose_level+ hypertension, data=stroke_data, family = binomial)
summary(mod.red)
```

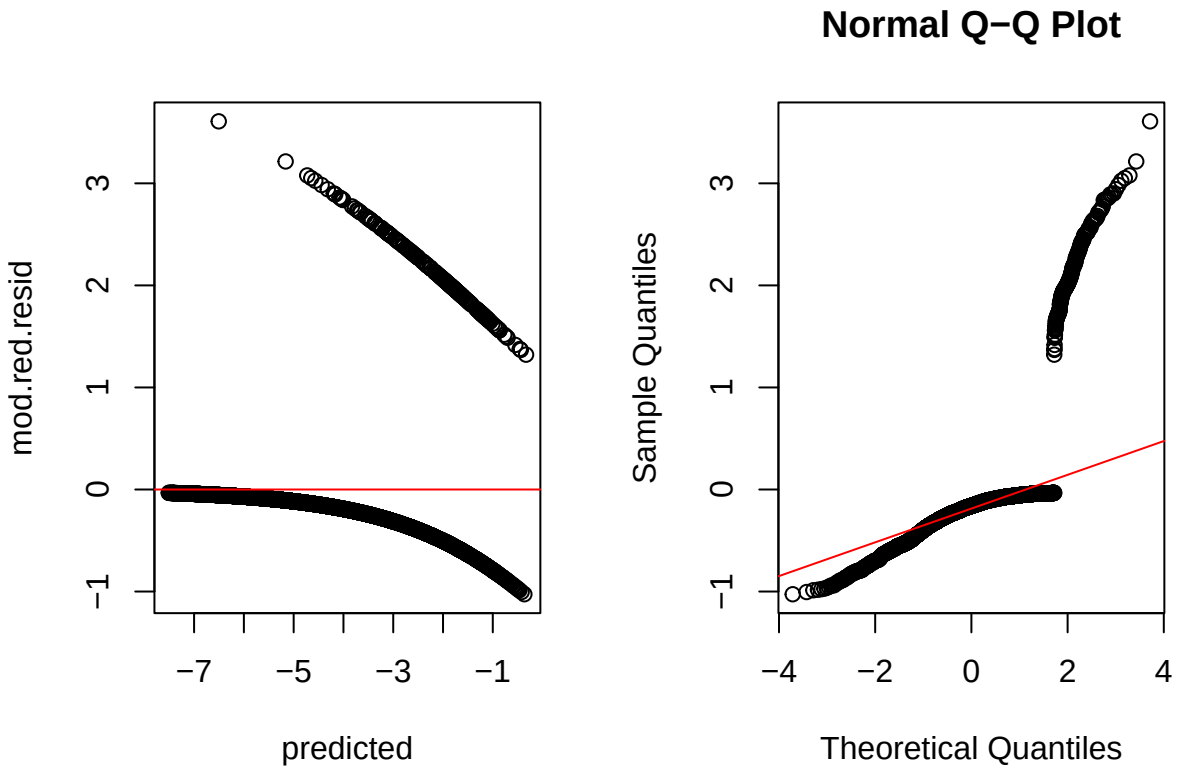
```
##
## Call:
## glm(formula = stroke ~ age + bmi + avg_glucose_level + hypertension,
##     family = binomial, data = stroke_data)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0265  -0.2986  -0.1600  -0.0755   3.6075
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -7.852291   0.541579 -14.499  < 2e-16 ***
## age                0.069793   0.005593  12.479  < 2e-16 ***
## bmi                0.002621   0.011598   0.226  0.82121
## avg_glucose_level  0.004984   0.001276   3.905 9.41e-05 ***
## hypertension       0.543399   0.173304   3.136  0.00172 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1728.4  on 4908  degrees of freedom
## Residual deviance: 1378.4  on 4904  degrees of freedom
## AIC: 1388.4
##
## Number of Fisher Scoring iterations: 7
```

The `mod.red` has `age`, `hypertension`, `avg_glucose_level`, `bmi` as the variables with the highest level of significance.

```
mod.red.resid <- residuals(mod.red, type="deviance")
predicted <- predict(mod.red, type = "link")

par(mfrow=c(1,2))
plot(mod.red.resid~predicted)
abline(h=0, col='red')
qqnorm(mod.red.resid)
qqline(mod.red.resid, col='red')
```



Residual plot:

```
#par(mfrow=c(2,2))
#plot(mod.red)
#par(mfrow=c(1,1))
```

We can see from the residual vs predicted values the presence of high non-linearity in the dataset. In the qqplot instead we see that residuals do no follow a normal distribution.

Instead in the standard deviance vs predicted we can see that homoscedasticity does not hold since the line of the standard residual is not flat, hence even by standardizing the residual we end up having high variance among residuals.

In the end by looking at the leverage plot, we see the presence of some sample with high leverage values (bottom right), which could influence the prediction of the model.

Buuut I don't know in which range of levarage value is considered to change a lot the prediction of the model. Furthermore R does not show the index of the sample with high leverage, *I guess because a lot of values could change the prediction.* Some outliers with high variance are: idx: 183, 246, 163

Mixed Approach for Models

In order to see which variables were relevant on our research we tested various models usign the mixed approach: we started by the null model and then added variables and removed those with maximum p-value (which indicates low relationship with the `stroke` variable).

```
mod1 <- glm(stroke~age, family=binomial)
summary(mod1)
```

```
##
## Call:
## glm(formula = stroke ~ age, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7162  -0.3073  -0.1639  -0.0776   3.5579
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.377231   0.362383  -20.36   <2e-16 ***
## age          0.074969   0.005318   14.10   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1728.4  on 4908  degrees of freedom
## Residual deviance: 1407.7  on 4907  degrees of freedom
## AIC: 1411.7
##
## Number of Fisher Scoring iterations: 7
```

```r
mod2 <- glm(stroke~age+avg_glucose_level+age*avg_glucose_level, family = binomial)
summary(mod2) # -> age no correlated

# remove age*avg_glucose_level and
# add age
mod3 <-glm(stroke~avg_glucose_level+hypertension + age,
           family = binomial)
summary(mod3)

# add hypertension*avg_glucose_level+hypertension*age
mod4 <- glm(stroke~avg_glucose_level+age+hypertension+
              hypertension*avg_glucose_level+hypertension*age,
            family = binomial)
summary(mod4)

# remove hypertension*avg_glucose_level+hypertension*age
# add heart_disease
mod5 <- glm(stroke~avg_glucose_level+age*avg_glucose_level+hypertension+age+
           heart_disease, family = binomial)
summary(mod5)

# remove glucose_level*age
# add hypertension*age + heart_disease*avg_glucose_level + heart_disease*hypertension
mod6 <- glm(stroke~age+avg_glucose_level +age*avg_glucose_level+hypertension+
              heart_disease+hypertension*age + heart_disease*avg_glucose_level+
              heart_disease*hypertension, family = binomial)
summary(mod6)


# Da confrontare con R adjusted e validation test
# Question: cosa cambia tra il 6 e il 7, no additive term of heart disease, why so big difference then?
```

```r
mod7 <- glm(stroke~avg_glucose_level+age*avg_glucose_level+hypertension+
              hypertension*age + heart_disease*avg_glucose_level+
              heart_disease*hypertension, family = binomial)
summary(mod7)

# remove age*avg_glucose_level
# add bmi
mod8 <- glm(stroke~avg_glucose_level+hypertension+
              hypertension*age + heart_disease*avg_glucose_level+
              heart_disease*hypertension + bmi, family = binomial  )
summary(mod8)
```

```r
# remove bmi
mod9 <- glm(stroke~avg_glucose_level+hypertension+age+
              hypertension*age + heart_disease*avg_glucose_level+
              heart_disease*hypertension + bmi*age+bmi*avg_glucose_level, family = binomial  )
summary(mod9)
```

```
##
## Call:
## glm(formula = stroke ~ avg_glucose_level + hypertension + age +
##     hypertension * age + heart_disease * avg_glucose_level +
##     heart_disease * hypertension + bmi * age + bmi * avg_glucose_level,
##     family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1184  -0.2946  -0.1549  -0.0698   3.6045
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -8.3319425  1.4703553  -5.667 1.46e-08 ***
## avg_glucose_level               -0.0088875  0.0056733  -1.567   0.1172
## hypertension                     1.6435129  1.0185165   1.614   0.1066
## age                              0.1064769  0.0208413   5.109 3.24e-07 ***
## heart_disease                   -0.3361740  0.5318449  -0.632   0.5273
## bmi                              0.0220495  0.0464285   0.475   0.6348
## hypertension:age                -0.0142223  0.0148358  -0.959   0.3377
## avg_glucose_level:heart_disease  0.0064928  0.0032311   2.009   0.0445 *
## hypertension:heart_disease      -0.6525380  0.4620173  -1.412   0.1578
## age:bmi                         -0.0012261  0.0006809  -1.801   0.0718 .
## avg_glucose_level:bmi            0.0003969  0.0001716   2.312   0.0208 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1728.4  on 4908  degrees of freedom
## Residual deviance: 1360.6  on 4898  degrees of freedom
## AIC: 1382.6
##
## Number of Fisher Scoring iterations: 8
```

Let's see some relevant information, such as outliers, leverage point and collinearity through some plots:

```r
par(mfrow=c(2,2))
#plot(mod9)
par(mfrow=c(1,1))
```

Si stima che la percentuale di persone che possono avere un ictus andrà via via crescendo dal momento che l'età media della popolazione è in costante crescita.