Statistical Learning Project

Filippo Santin, Gurjeet Singh, Francesca Zen

18/5/2021

1. Introduction

In the following report we present an analysis computed on stroke diseases, and we try to explain from statistical analysis some statistics and correlation factors of the given features/predictors by constructing predictive statistical models in order to assess possible linear and non-linear relationships of features (predictors) to predict a stroke disease in a person (predicted variable).

"Stroke" is the medical term which causes damage to brain tissue or death of its portions, due to insufficient blood supply in vessels to an area of the brain.

Our aim is to see if and how the variables we are dealing with are related, in order to predict which individual is more probable to have 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 or assess 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

2. Exploiting the Dataset

The dataset we used is 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 and its pre-processing: we looked at the `id` column and verified that all the data collected was referring to different people, so no recidivist status were involved. After this check we removed the column from the dataset as it does not hold useful information for our study

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

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

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

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

In addition, the variable `bmi` was not numeric because of the presence of "N/A" string values which identify missing information, and so we transformed it into numeric values and then removed the NA values.

```
## Warning: NAs introduced by coercion
```

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)
```

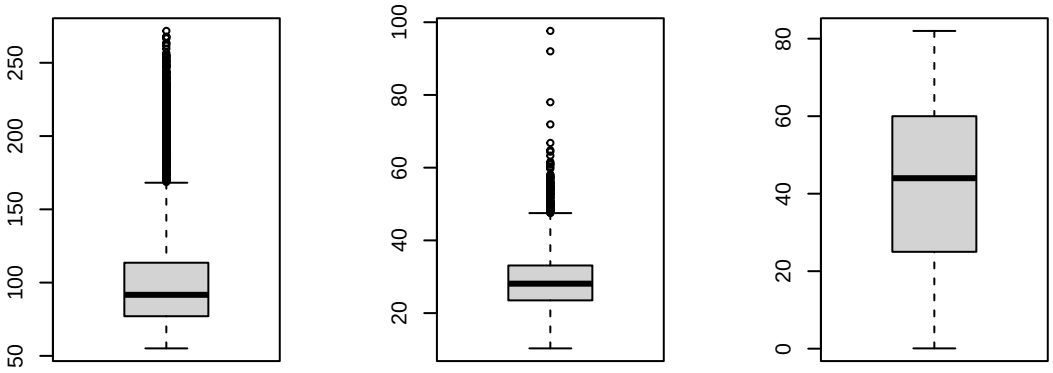
```
##     gender          age         hypertension     heart_disease    ever_married
##  Female:2897   Min.   : 0.08   Min.   :0.00000   Min.   :0.0000   No :1705
##  Male  :2011   1st Qu.:25.00   1st Qu.:0.00000   1st Qu.:0.0000   Yes:3204
##  Other :   1   Median :44.00   Median :0.00000   Median :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
```

But in order to highlight and study better the data we used some plot to study their statistics and data distribution.

```
attach(stroke_data)
```

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

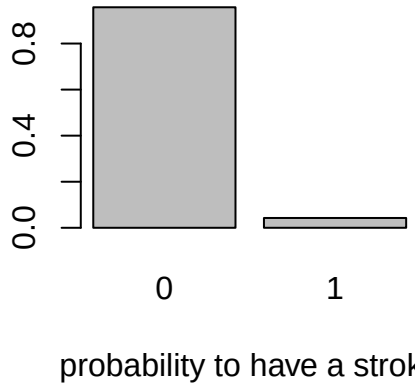
```
par(mfrow=c(1,3))
boxplot(avg_glucose_level, xlab= 'average glucose level' )
boxplot(bmi, xlab = 'body mass index')
boxplot(age, xlab = 'age',pch=20)
```



From the above boxplots we can see, TODO: "Outliers", but they represent reality and possible interesting cases for stroke prediction

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%.

3. Searching for Relationships

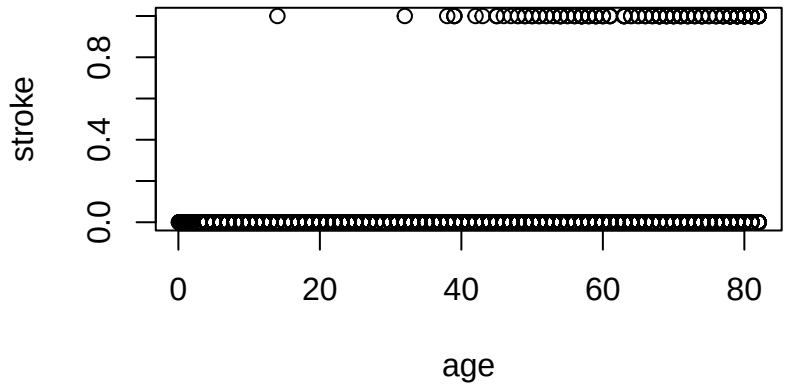
At this point we can ask some questions:

- Is it possible to prevent ictus?
- Which factors are the most related to it?
- How strong are the relations between the features?
- Are the given variables enough to predict a good accuracy of some possible person affected by ictus?

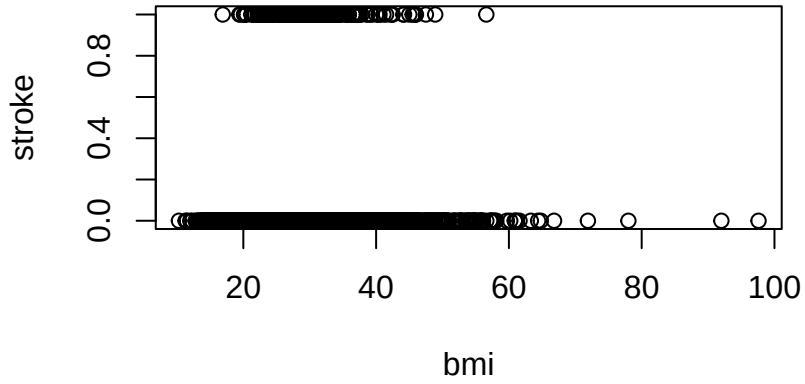
We will explore the data trying to answer them.

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

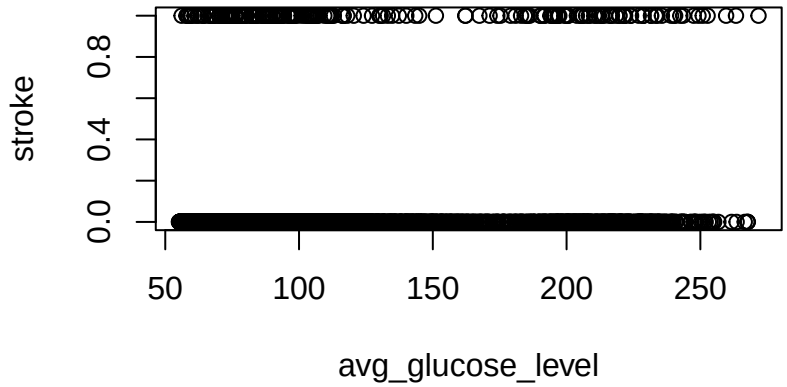
```
plot(stroke~age)
```




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



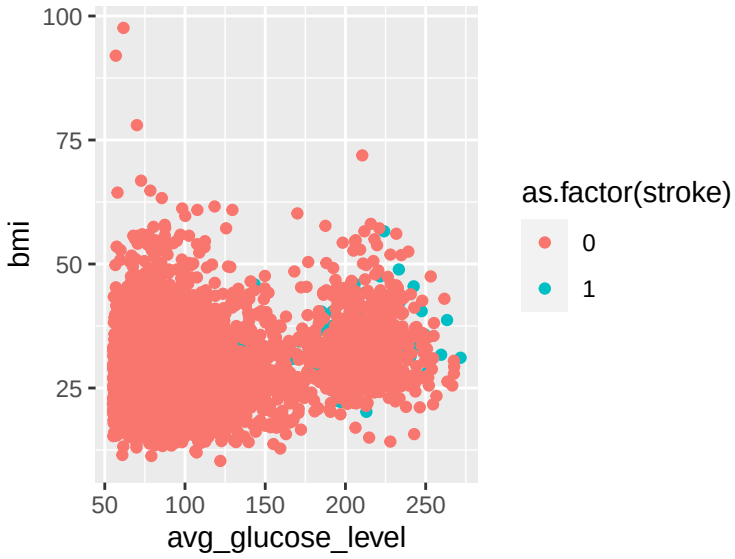

```
plot(stroke~avg_glucose_level)
```



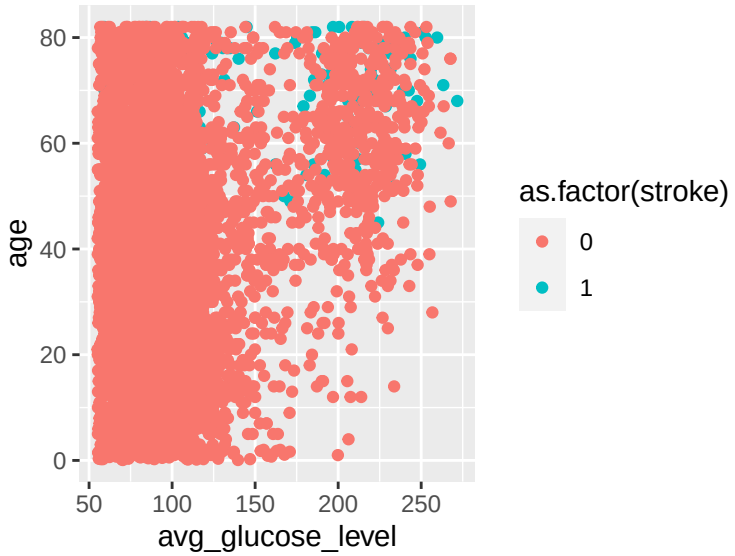
Looking to these plots we can see that the incidence of the virus increases progressively with age and that if we sum also the information about `avg_glucose_level` we may wonder if diabetic people are more probable to get a stroke or not. In addition there is no apparent relation of `stroke` with `bmi`.

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

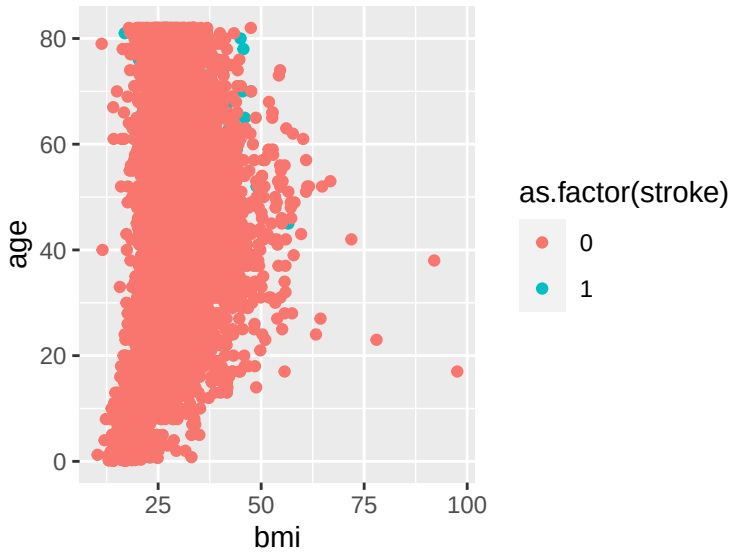
```
library(ggplot2)
par(mfrow=c(1,3))
ggplot(stroke_data, aes(x = avg_glucose_level, y = bmi,
                        col = as.factor(stroke))) + geom_point()
```




```
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

4. Analysis on models

While going on with the classification using logistic regression we could meet the following problems: non-linearity of the data, correlation of errorterms, heteroschedasticity, outliers, leverage point and collinearity.

4.1 Full and Reduced Models

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

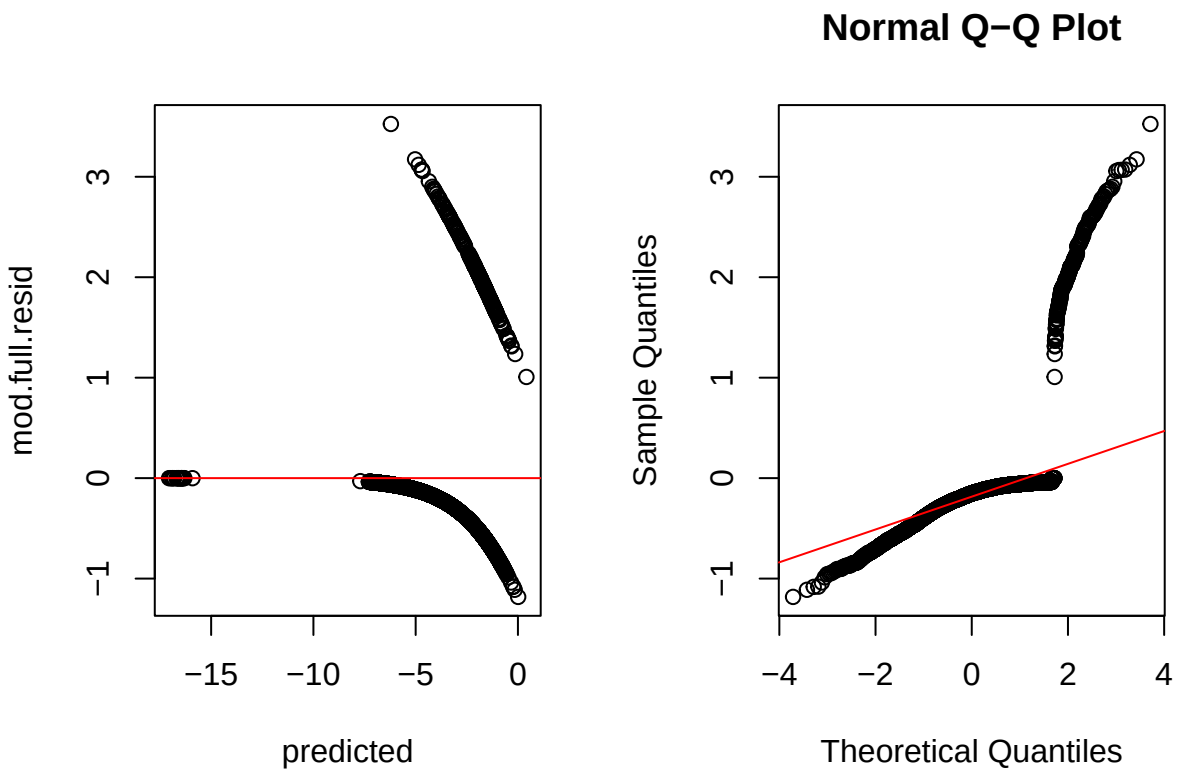
```
##
## Call:
## glm(formula = stroke ~ ., family = binomial, data = stroke_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1823  -0.2947  -0.1524  -0.0744   3.5251
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -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')
```



The residual plots aare not satisfactory. From the right plot we can see that the data are not normal.

4.2 Comparison between Reduced Models

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 + heart_
summary(mod.red1)
```

```
##
## Call:
## glm(formula = stroke ~ age + bmi + avg_glucose_level + hypertension +
##     smoking_status + gender + heart_disease, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0751  -0.2957  -0.1572  -0.0734   3.6706
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -7.823946   0.588445 -13.296  < 2e-16 ***
## age                          0.069041   0.005847  11.808  < 2e-16 ***
## bmi                          0.003458   0.011745   0.294 0.768441
## avg_glucose_level            0.004697   0.001289   3.644 0.000269 ***
## hypertension                 0.517649   0.174438   2.968 0.003002 **
## smoking_statusnever smoked  -0.057792   0.187972  -0.307 0.758502
## smoking_statussmokes         0.321264   0.228512   1.406 0.159754
## smoking_statusUnknown       -0.256978   0.245259  -1.048 0.294740
## genderMale                  -0.011195   0.154011  -0.073 0.942055
## genderOther                 -7.287812 324.743860  -0.022 0.982096
## heart_disease                0.372836   0.206072   1.809 0.070411 .
## ---
## 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: 1369.4  on 4898  degrees of freedom
## AIC: 1391.4
##
## Number of Fisher Scoring iterations: 11
```

```
# Reduced model 2, Remove gender
mod.red2 <- glm(stroke ~ age + bmi + avg_glucose_level + hypertension + smoking_status  + heart_disease
summary(mod.red2)
```

```
##
## Call:
## glm(formula = stroke ~ age + bmi + avg_glucose_level + hypertension +
##     smoking_status + heart_disease, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0766  -0.2964  -0.1572  -0.0733   3.6720
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)                -7.830942   0.582317 -13.448  < 2e-16 ***
## age                         0.069065   0.005842  11.822  < 2e-16 ***
## bmi                         0.003487   0.011744   0.297 0.766535
## avg_glucose_level           0.004689   0.001285   3.648 0.000264 ***
## hypertension                0.517599   0.174427   2.967 0.003003 **
## smoking_statusnever smoked -0.055884   0.186343  -0.300 0.764254
## smoking_statussmokes        0.321811   0.228437   1.409 0.158907
## smoking_statusUnknown      -0.255822   0.244827  -1.045 0.296066
## heart_disease               0.371145   0.204739   1.813 0.069867 .
## ---
## 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: 1369.4  on 4900  degrees of freedom
## AIC: 1387.4
##
## Number of Fisher Scoring iterations: 7
```

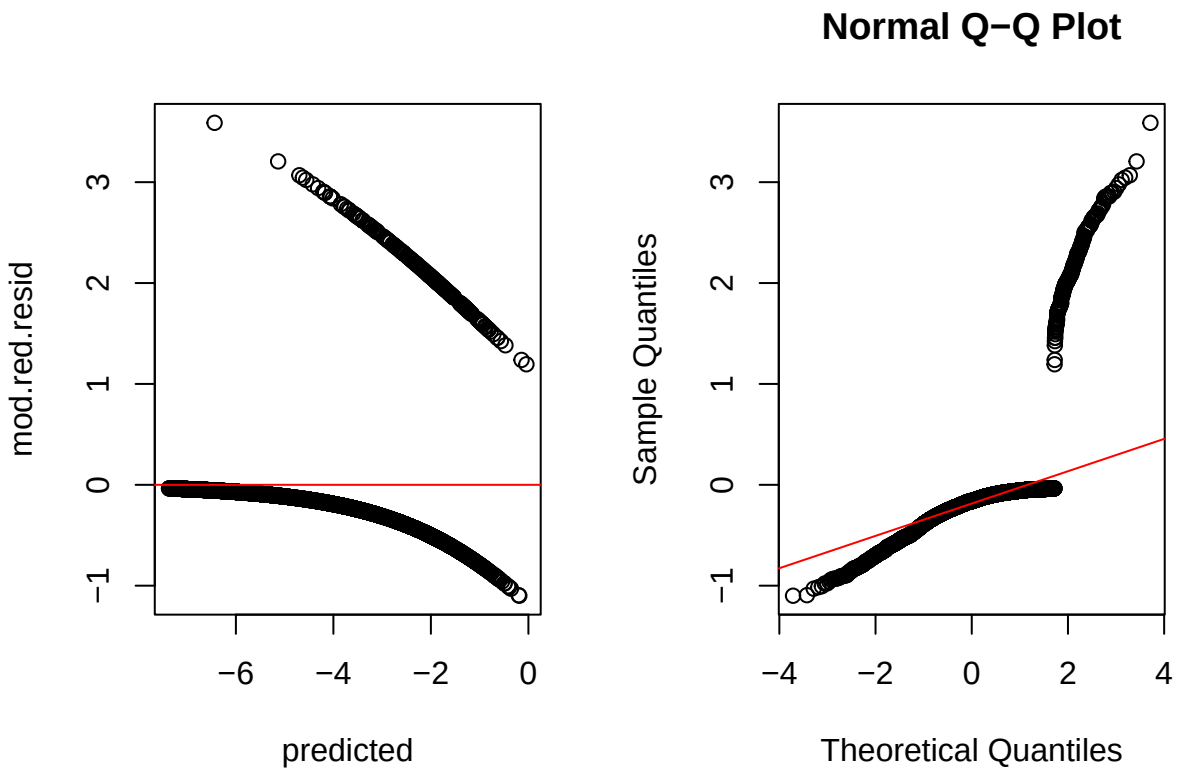
```
#####
mod.red <- glm(stroke~age + heart_disease + avg_glucose_level+ hypertension, data=stroke_data, family =
summary(mod.red)
```

```
##
## Call:
## glm(formula = stroke ~ age + heart_disease + avg_glucose_level +
##     hypertension, family = binomial, data = stroke_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0995  -0.2940  -0.1599  -0.0778   3.5885
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -7.660740   0.387152 -19.787  < 2e-16 ***
## age                0.067547   0.005571  12.124  < 2e-16 ***
## heart_disease      0.404298   0.203447   1.987 0.046895 *
## avg_glucose_level  0.004802   0.001255   3.828 0.000129 ***
## hypertension       0.539613   0.173055   3.118 0.001820 **
## ---
## 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: 1374.6  on 4904  degrees of freedom
## AIC: 1384.6
##
## Number of Fisher Scoring iterations: 7
```

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

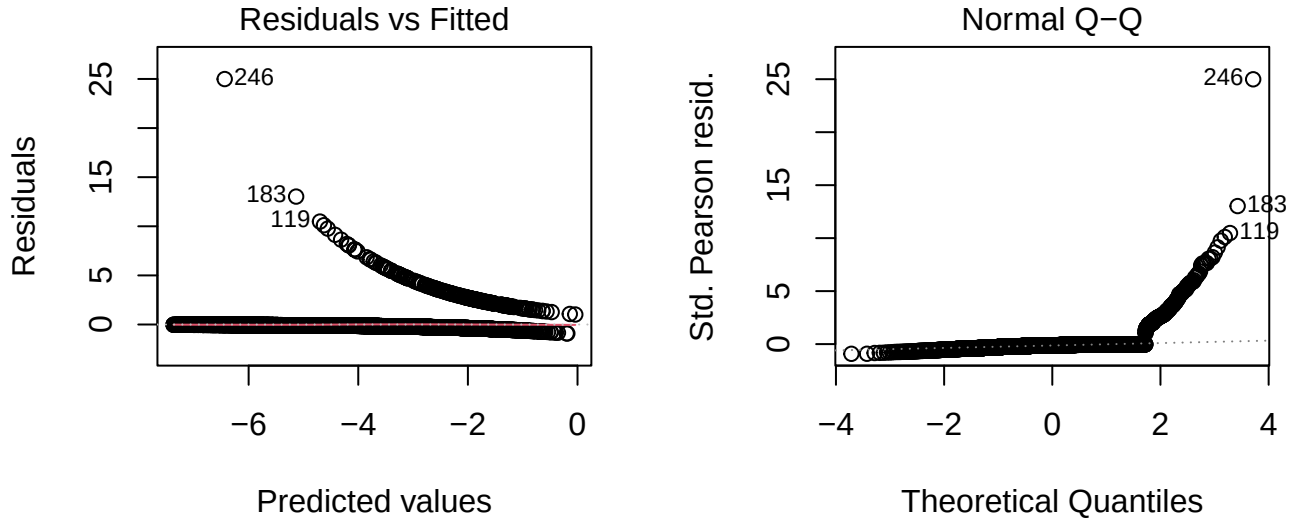
```
mod.red.resid <- residuals(mod.red, type="deviance")
predicted <- predict(mod.red, type = "link")
```

```
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')
```

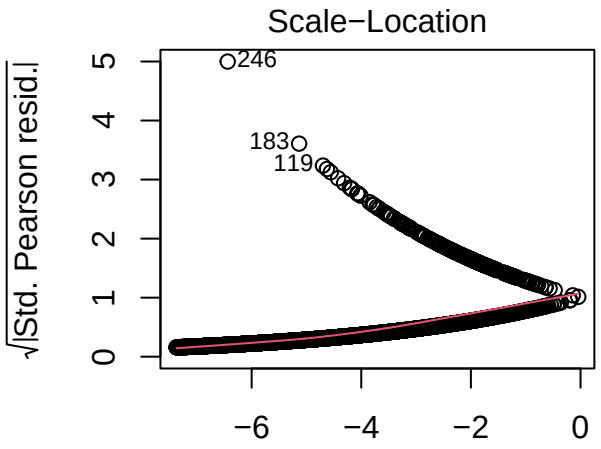


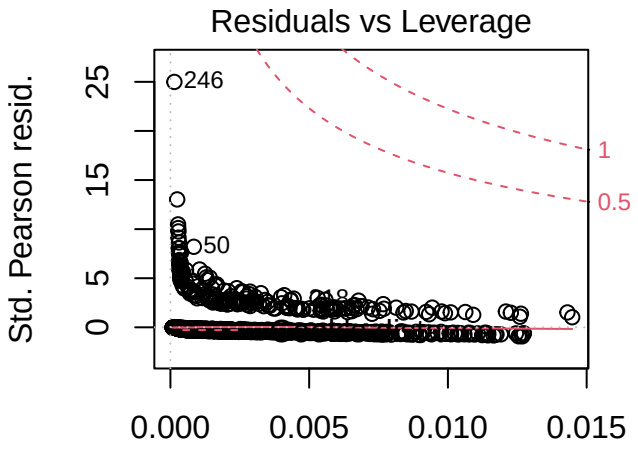
Residual plot:

```
plot(mod.red)
```



:e ~ age + heart_disease + avg_glucose_level + :e ~ age + heart_disease + avg_glucose_level +





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: 119, 183, 246.

Anova computation:

```
anova(mod.full, mod.red, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: stroke ~ gender + age + hypertension + heart_disease + ever_married +
##     work_type + Residence_type + avg_glucose_level + bmi + smoking_status
## Model 2: stroke ~ age + heart_disease + avg_glucose_level + hypertension
##   Resid. Df Resid. Dev  Df Deviance Pr(>Chi)
## 1      4892     1363.2
## 2      4904     1374.7 -12   -11.42   0.4933
```

As expected from the anova test rejects that the complex model is more significant than the reduced one, since the p-value is not less than 5%. Hence the full model does not help with our prediction.

Outliers:

```
library(knitr)
kable(stroke_data[c("119","183","246"), ])
```

	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
119	Female	38	0	0	No	Self-employed	Urban	82.28	24.0	formerly smoked	1
183	Female	32	0	0	Yes	Private	Rural	76.13	29.9	smokes	1
246	Female	14	0	0	No	children	Rural	57.93	30.9	Unknown	1

4.3 Mixed Approach for Models

In order to see which variables were relevant on our research we tested various models using the mixed approach: we started by the reduced model `mod.red` and then tested some interaction between the explanatory variables for improving the performance of the model.

We started with `mod.red`, which had `stroke` as response and `age`, `avg_glucose_level`,`hypertension` and `heart_disease` as predictors. We recall that its AIC was 1384.6.

We started by considering the interaction of `age` with the other numerical features, i.e. `avg_glucose_level`,`hypertension` and `heart_disease`: we find out that only `age*heart_disease` was relevant between the ones tested, with an AIC = 1384, which was lower than the one of the reduced model `mod.red`.

```
mod1 <- glm(stroke~age + avg_glucose_level+ heart_disease+ hypertension +
              age*heart_disease, family=binomial)
summary(mod1)
```

```
##
## Call:
## glm(formula = stroke ~ age + avg_glucose_level + heart_disease +
##     hypertension + age * heart_disease, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9897  -0.2980  -0.1557  -0.0737   3.6232
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -7.816578   0.407175 -19.197  < 2e-16 ***
## age                0.070133   0.005889  11.908  < 2e-16 ***
## avg_glucose_level  0.004702   0.001253   3.752 0.000176 ***
## heart_disease      2.765299   1.396557   1.980 0.047694 *
## hypertension       0.536550   0.172602   3.109 0.001880 **
## age:heart_disease -0.032872   0.019486  -1.687 0.091604 .
## ---
## 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: 1372.0  on 4903  degrees of freedom
## AIC: 1384
##
## Number of Fisher Scoring iterations: 7
```

We went on considering all the interaction of `avg_glucose_level` with the remaining predictors and find out that `avg_glucose_level*hypertension` was the best of the possible interaction but cannot improve the previous model, it had an AIC of 1385.9.

```
mod2 <- glm(stroke~age + avg_glucose_level+ heart_disease+hypertension +
              avg_glucose_level*hypertension, family=binomial)
summary(mod2)
```

In the end it was left the interaction `heart_disease*hypertension` which return an AIC 1384.5 for the model.

```
mod3 <- glm(stroke~age + avg_glucose_level+ heart_disease+hypertension +
             heart_disease*hypertension, family=binomial)
```

```
summary(mod3)
```

We then tried to see the results of the `mod.red` without `heart_disease` which was the explanatory variables with higher p-value, and then we added some interaction terms, starting with the one with `age`:

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

The AIC of this model was 1386.2, higher that the ones seen on the previous tested models. We go forther testing also the interaction add `avg_glucose_level*hypertension`

```
##
## Call:
## glm(formula = stroke ~ age + avg_glucose_level + hypertension +
##     avg_glucose_level * hypertension, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9682  -0.3000  -0.1591  -0.0759   3.6156
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -7.839547   0.393628 -19.916  < 2e-16 ***
## age                              0.069259   0.005505  12.582  < 2e-16 ***
## avg_glucose_level                0.005785   0.001473   3.928 8.58e-05 ***
## hypertension                     0.894651   0.413438   2.164   0.0305 *
## avg_glucose_level:hypertension -0.002479   0.002707  -0.916   0.3598
## ---
## 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: 1377.6  on 4904  degrees of freedom
## AIC: 1387.6
##
## Number of Fisher Scoring iterations: 7
```

It had an AIC = 1387.6. In the end we return on our base model, `mod.red` and add the teo best iteractions, i.e. the two terms that reduced the AIC term:

```
mod6 <- glm(stroke~age + avg_glucose_level+ heart_disease+ hypertension +
              age*heart_disease + heart_disease*hypertension, family=binomial)
summary(mod6)
```

It has AIC = 1384.2 but the interactions had a p-value greater than 0.1 and it didn't seem to represent our data properly. At the end of all we promote `mod1` as the model which better fits our data.

Let's now see some relevant information, such as outliers on the `mod1`:

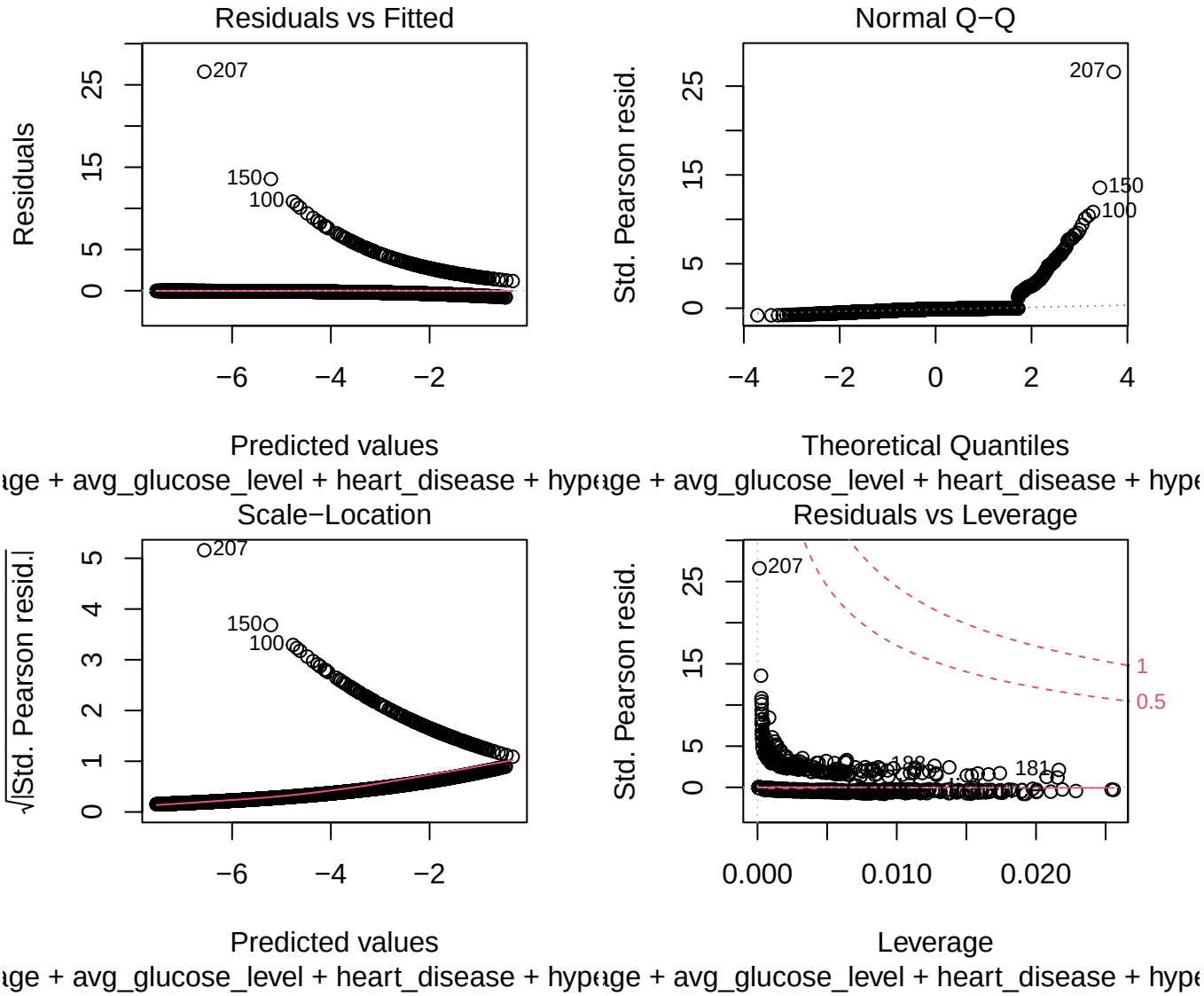
```
kable(stroke_data[c("207","150","100"), ])
```

	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
207	Female	81	0	0	Yes	Private	Rural	80.13	23.4	never smoked	1

	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
150	Female	70	0	1	Yes	Private	Rural	239.07	26.1	never smoked	1
100	Female	69	0	0	Yes	Govt_job	Urban	82.81	28.0	never smoked	1

but also leverage point and collinearity:

```
plot(mod1)
```



4.4 Polynomial models

We tried to use a polynomial model starting from the reduced model `mod.red` with also `bmi` as predictors. Then we contribute with the square of `bmi`, `avg_glucose_level` and then both for the `mod.poly1`,`mod.poly2` and `mod.poly3` respectively.

```
mod.poly1 <- glm(stroke~age + heart_disease + avg_glucose_level+ hypertension+bmi+
                        I(bmi^2), family = binomial)
```

```
summary(mod.poly1)
```

```
##
## Call:
## glm(formula = stroke ~ age + heart_disease + avg_glucose_level +
##     hypertension + bmi + I(bmi^2), family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1025  -0.2962  -0.1603  -0.0767   3.5773
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -8.4496839  1.2063580  -7.004 2.48e-12 ***
## age                0.0675329  0.0057033  11.841  < 2e-16 ***
## heart_disease      0.3987505  0.2038850   1.956 0.050493 .
## avg_glucose_level  0.0047035  0.0012850   3.660 0.000252 ***
## hypertension       0.5363360  0.1737148   3.087 0.002019 **
## bmi                0.0468572  0.0701687   0.668 0.504275
## I(bmi^2)          -0.0006419  0.0010336  -0.621 0.534570
## ---
## 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: 1374.1  on 4902  degrees of freedom
## AIC: 1388.1
##
## Number of Fisher Scoring iterations: 7
```

```
mod.poly2 <- glm(stroke~age + heart_disease + avg_glucose_level+ hypertension+bmi+
                   + I(avg_glucose_level^2), family = binomial)
summary(mod.poly2)
```

```
##
## Call:
## glm(formula = stroke ~ age + heart_disease + avg_glucose_level +
##     hypertension + bmi + +I(avg_glucose_level^2), family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0984  -0.2943  -0.1600  -0.0772   3.5821
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -7.664e+00  7.831e-01  -9.787  < 2e-16 ***
## age                    6.776e-02  5.680e-03  11.929  < 2e-16 ***
## heart_disease          4.066e-01  2.036e-01   1.998  0.04576 *
## avg_glucose_level      2.982e-03  8.865e-03   0.336  0.73661
## hypertension           5.347e-01  1.738e-01   3.077  0.00209 **
## bmi                    3.554e-03  1.166e-02   0.305  0.76056
## I(avg_glucose_level^2) 5.778e-06  2.921e-05   0.198  0.84321
## ---
```

```
## 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: 1374.5  on 4902  degrees of freedom
## AIC: 1388.5
##
## Number of Fisher Scoring iterations: 7
```

```
mod.poly3 <- glm(stroke~age + heart_disease + avg_glucose_level+ hypertension+bmi+
                     I(bmi^2) + I(avg_glucose_level^2), family = binomial)
summary(mod.poly3)
```

```
##
## Call:
## glm(formula = stroke ~ age + heart_disease + avg_glucose_level +
##     hypertension + bmi + I(bmi^2) + I(avg_glucose_level^2), family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1043  -0.2961  -0.1600  -0.0769   3.5687
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -8.336e+00  1.327e+00  -6.281 3.36e-10 ***
## age                     6.746e-02  5.714e-03  11.806  < 2e-16 ***
## heart_disease           3.986e-01  2.039e-01   1.955  0.05063 .
## avg_glucose_level       2.901e-03  8.862e-03   0.327  0.74338
## hypertension            5.365e-01  1.737e-01   3.088  0.00201 **
## bmi                     4.699e-02  7.019e-02   0.670  0.50317
## I(bmi^2)               -6.441e-04  1.034e-03  -0.623  0.53332
## I(avg_glucose_level^2)  6.002e-06  2.920e-05   0.206  0.83715
## ---
## 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: 1374.1  on 4901  degrees of freedom
## AIC: 1390.1
##
## Number of Fisher Scoring iterations: 7
```

At the end of the tests nothing interesting appeared with polynomial models. There were no improvement in the results.

5. LDA

Assumption: samples are normally distributed and have same variance in every class => strong assumption.

```
library(MASS)
lda.fit <- lda(stroke~age+bmi+avg_glucose_level+hypertension+work_type+gender
               +smoking_status+ever_married+Residence_type + heart_disease)
lda.pred <- predict(lda.fit)
```

```
table(lda.pred$class, stroke)
```

```
##    stroke
##       0    1
##   0 4646  186
##   1   54   23
```

```
lda.pred.stroke <- lda.pred$posterior[, 2]
```

6. QDA

Assumption: sample are normally distributed BUT NOT SAME variance among classes.

```
qda.fit <- qda(stroke~age+bmi+avg_glucose_level+hypertension+heart_disease+smoking_status, data = stroke
# ERROR rank deficiency, i.e. some variables
# are collinear and one or more covariance matrices cannot be inverted to obtain the estimates in group
qda.pred <- predict(qda.fit, stroke_data)
qda.pred.stroke <- qda.pred$posterior[, 2]

table(qda.pred$class, stroke)
```

```
##    stroke
##       0    1
##   0 4260  123
##   1  440   86
```

7. ROC and RECALL-PRECISION CURVES

```
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
library(ROCR)
get.roc.recall.values <- function(pred_models, true_value) {
  result <- data.frame(Threshold1=double(), Specificity=double(), Sensitivity=double(),
                       Threshold2=double(), Recall=double(), Precision=double())
  n_models = length(list(mod.red.probs,lda.pred.stroke, qda.pred.stroke))
  par(mfrow=c(n_models, 2))
  for (pred in pred_models) {
    roc.res <- roc(true_value, pred, levels=c("0", "1"))
    plot(roc.res, print.auc=TRUE, legacy.axes=TRUE, xlab="False positive rate", ylab="True positive rat

    tmp.res  <- coords(roc.res, "best")
    pred.rec = prediction(mod.red.probs, true_value)
    perf = performance(pred.rec, "prec", "rec")
    plot(perf)
    pr_cutoffs <- data.frame(cutrecall=perf@alpha.values[[1]], recall=perf@x.values[[1]],
                             precision=perf@y.values[[1]])
```

```
    best_recall <- pr_cutoffs[which.min(pr_cutoffs$recall + pr_cutoffs$precision), ]

    result[nrow(result) + 1,] = c(tmp.res[1, 1], tmp.res[1, 2], tmp.res[1, 3],
                                  best_recall[1, 1], best_recall[1, 2], best_recall[1, 3])
  }
  par(mfrow=c(1, 1))
  return(result)
}


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

```
##
## Call:
## glm(formula = stroke ~ age + avg_glucose_level + hypertension +
##     bmi, 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 ***
## avg_glucose_level  0.004984   0.001276   3.905 9.41e-05 ***
## hypertension       0.543399   0.173304   3.136  0.00172 **
## bmi                0.002621   0.011598   0.226  0.82121
## ---
## 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
```

```
mod.red.probs <- predict(mod.red,type="response")
```

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.