Regressão Exemplo

Ana Cristina Braga

2023-12-07

Importação de dados

Importação direta de bases de dados existentes do R

```
#install.packages("faraway")
library(faraway)
data(pima, package = "faraway")
View(pima)
```

#Análise exploratória das variáveis

pregnant - variável quantitativa discreta glucose - variável quantitativa proporcional (contínua) diastolica - variável quantitativa proporcional (contínua) triceps - variável quantitativa proporcional (contínua) insulina - variável quantitativa proporcional (contínua) bmi - variável quantitativa proporcional (contínua) diabetes - variável quantitativa proporcional (contínua) age - variável quantitativa proporcional (contínua) test - variável qualitativa nominal

```
summary(pima)
```

```
##     pregnant         glucose        diastolic         triceps
##  Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00
##  1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00
##  Median : 3.000   Median :117.0   Median : 72.00   Median :23.00
##  Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54
##  3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00
##  Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00
##     insulin           bmi           diabetes           age
##  Min.   :  0.0   Min.   : 0.00   Min.   :0.0780   Min.   :21.00
##  1st Qu.:  0.0   1st Qu.:27.30   1st Qu.:0.2437   1st Qu.:24.00
##  Median : 30.5   Median :32.00   Median :0.3725   Median :29.00
##  Mean   : 79.8   Mean   :31.99   Mean   :0.4719   Mean   :33.24
##  3rd Qu.:127.2   3rd Qu.:36.60   3rd Qu.:0.6262   3rd Qu.:41.00
##  Max.   :846.0   Max.   :67.10   Max.   :2.4200   Max.   :81.00
##       test
##  Min.   :0.000
##  1st Qu.:0.000
##  Median :0.000
##  Mean   :0.349
##  3rd Qu.:1.000
##  Max.   :1.000
```

Tratamento inicial

```
pima$diastolic[pima$diastolic == 0] <- NA
pima$glucose[pima$glucose == 0] <- NA
```

```r
pima$triceps[pima$triceps == 0] <- NA
pima$insulin[pima$insulin == 0] <- NA
pima$bmi[pima$bmi == 0] <- NA
pima$test <- factor(pima$test)
summary(pima$test)
```

```
##   0   1
## 500 268
```

```r
levels(pima$test) <- c("negative", "positive")
summary(pima)
```

```
##     pregnant         glucose        diastolic         triceps
##  Min.   : 0.000   Min.   : 44.0   Min.   : 24.00   Min.   : 7.00
##  1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 64.00   1st Qu.:22.00
##  Median : 3.000   Median :117.0   Median : 72.00   Median :29.00
##  Mean   : 3.845   Mean   :121.7   Mean   : 72.41   Mean   :29.15
##  3rd Qu.: 6.000   3rd Qu.:141.0   3rd Qu.: 80.00   3rd Qu.:36.00
##  Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00
##                   NA's   :5       NA's   :35       NA's   :227
##     insulin            bmi           diabetes           age
##  Min.   : 14.00   Min.   :18.20   Min.   :0.0780   Min.   :21.00
##  1st Qu.: 76.25   1st Qu.:27.50   1st Qu.:0.2437   1st Qu.:24.00
##  Median :125.00   Median :32.30   Median :0.3725   Median :29.00
##  Mean   :155.55   Mean   :32.46   Mean   :0.4719   Mean   :33.24
##  3rd Qu.:190.00   3rd Qu.:36.60   3rd Qu.:0.6262   3rd Qu.:41.00
##  Max.   :846.00   Max.   :67.10   Max.   :2.4200   Max.   :81.00
##  NA's   :374      NA's   :11
##        test
##  negative:500
##  positive:268
##
##
##
##
##
```

```r
#install.packages("Amelia")

library(Amelia)
```

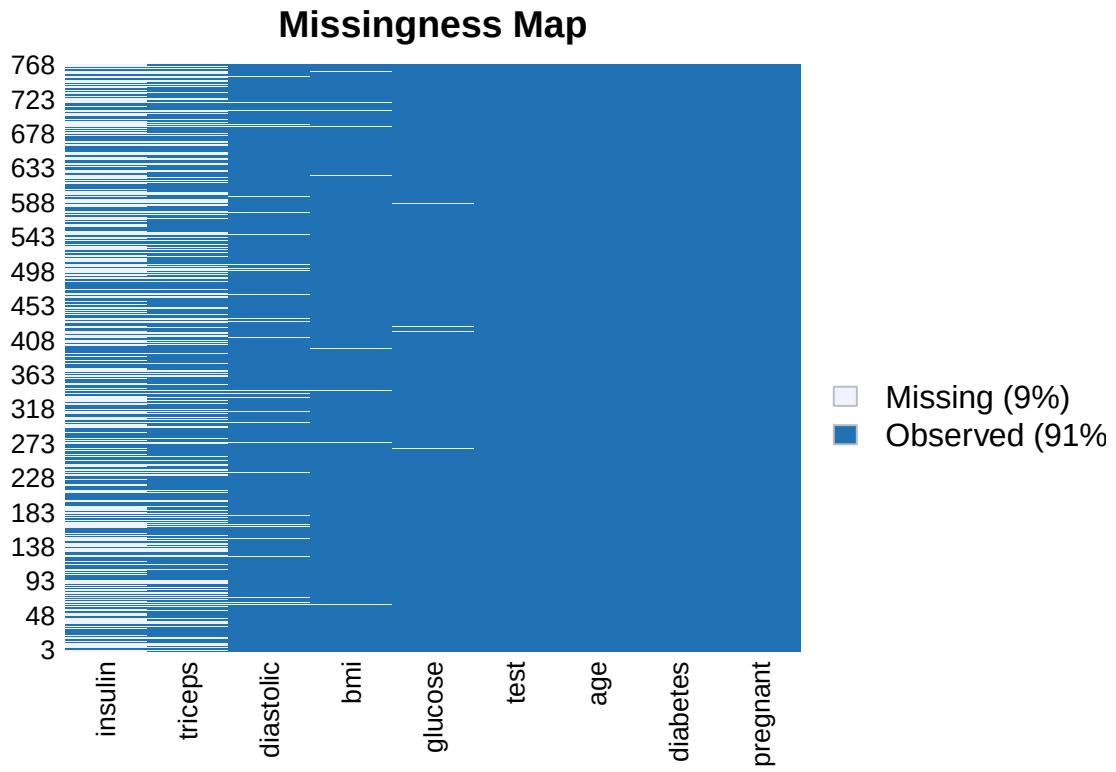
```
## Loading required package: Rcpp
```

```
## ##
## ## Amelia II: Multiple Imputation
## ## (Version 1.8.1, built: 2022-11-18)
## ## Copyright (C) 2005-2023 James Honaker, Gary King and Matthew Blackwell
## ## Refer to http://gking.harvard.edu/amelia/ for more information
## ##
```

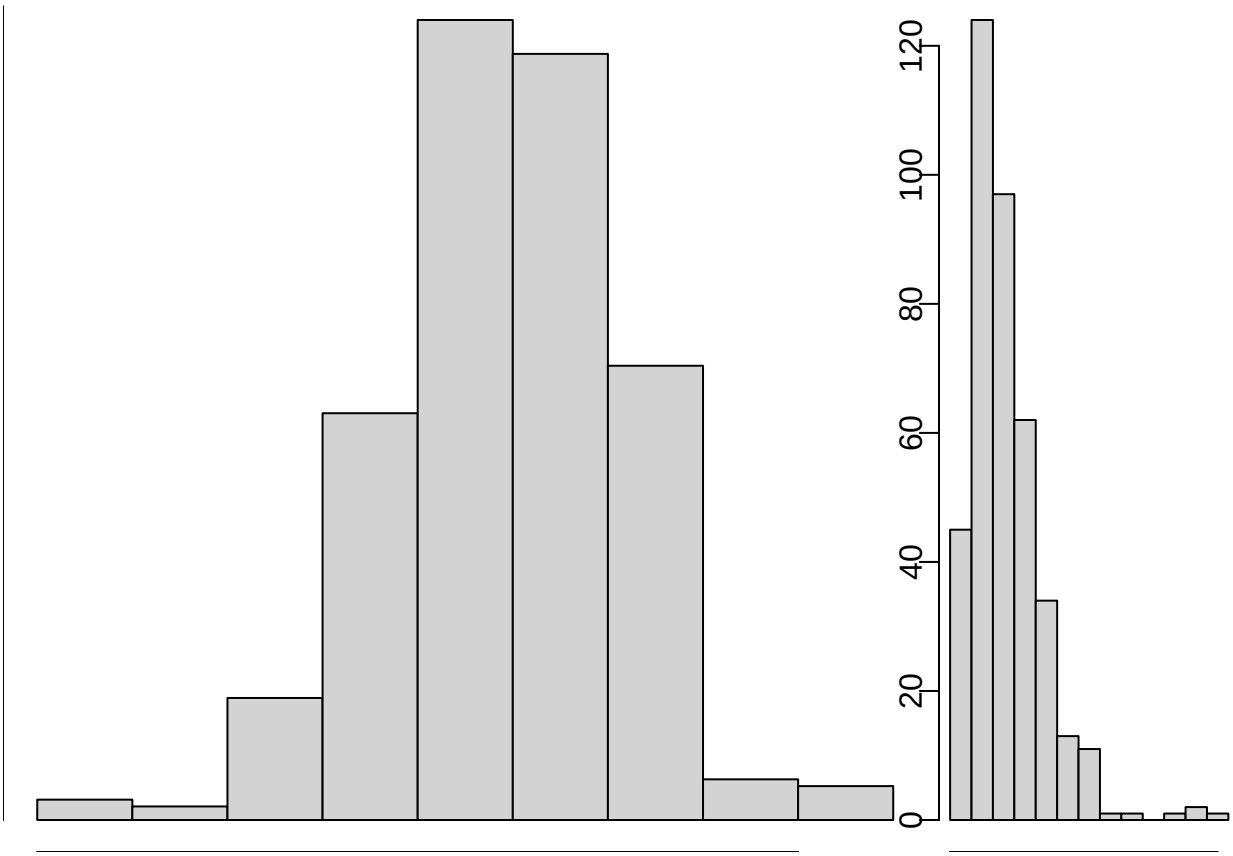
```
##
## Attaching package: 'Amelia'
```

```
## The following object is masked from 'package:faraway':
##
##     africa
```

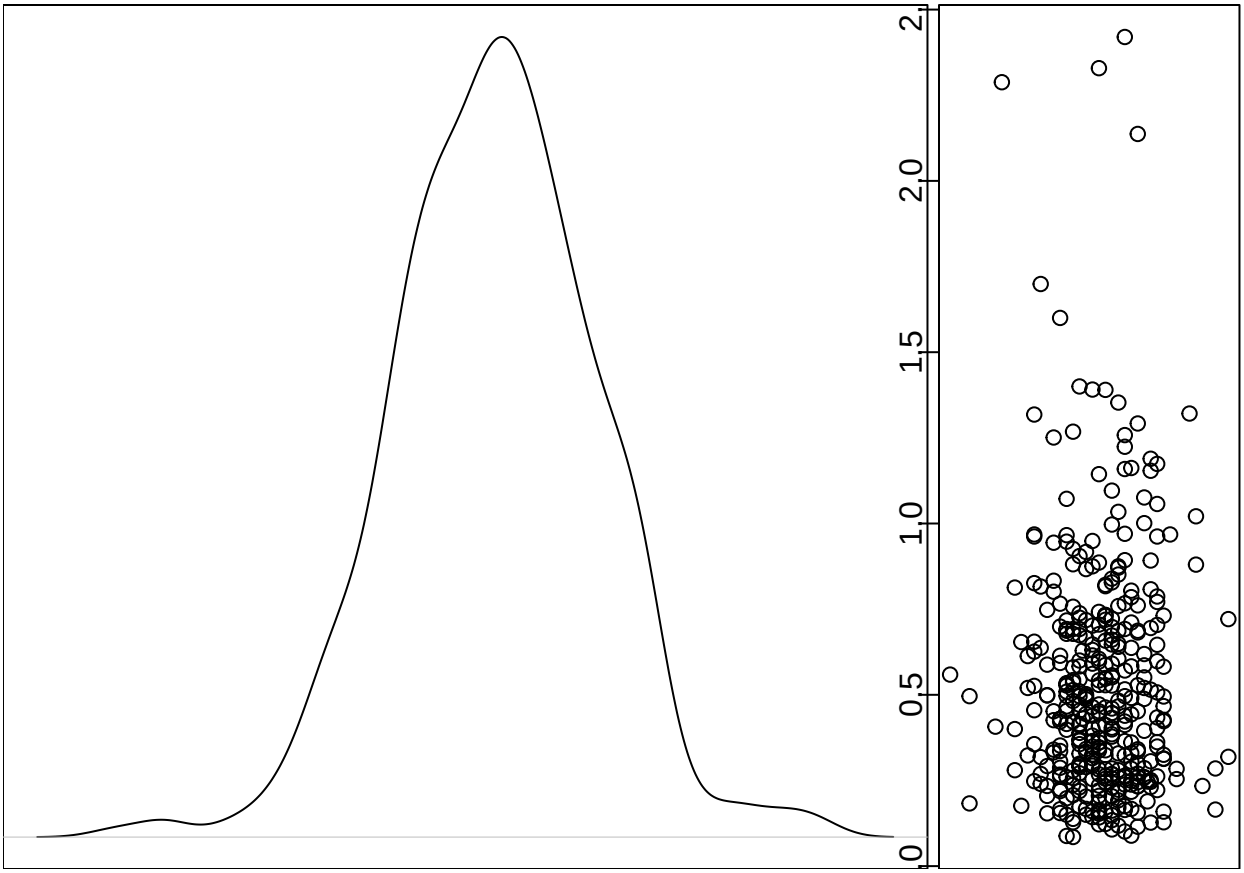
```
missmap(pima)
```



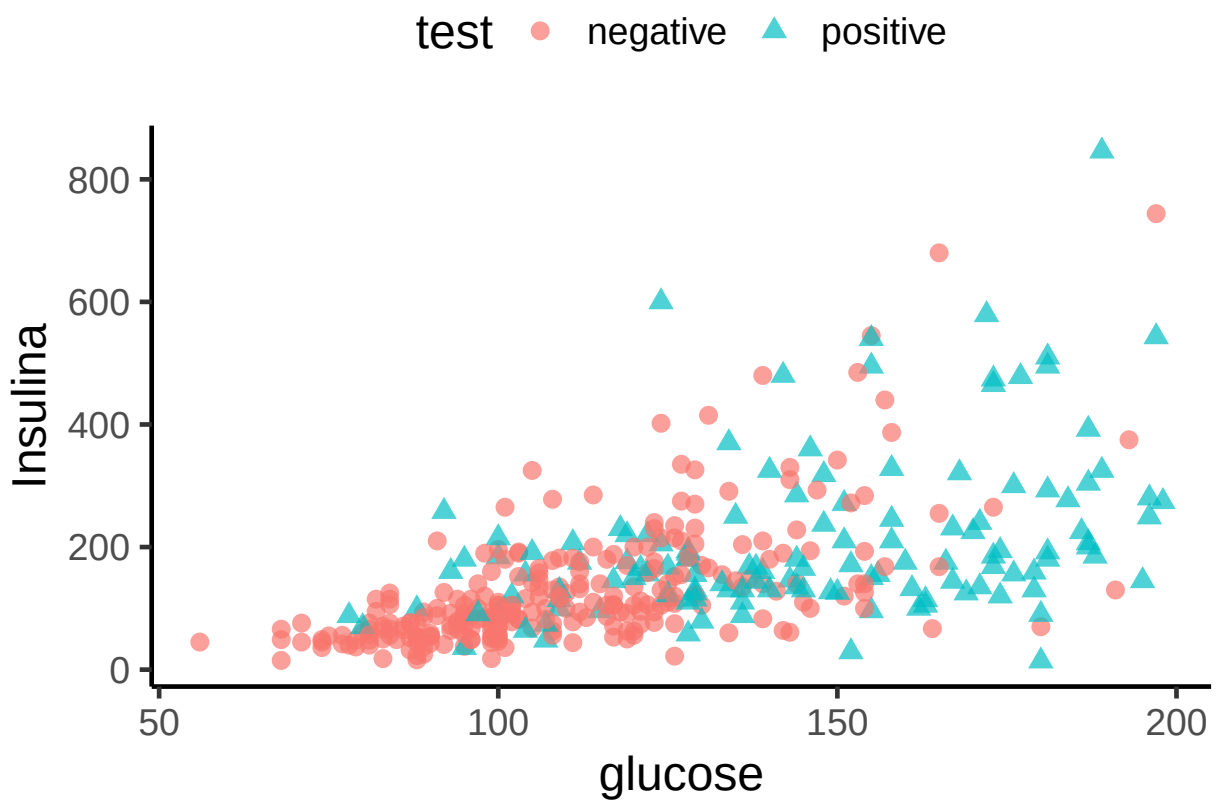

```
pima_NA<-na.omit((pima))

hist(pima_NA$diastolic, xlab = "Diastolic", main = "")


hist(pima_NA$diabetes, xlab = "Diabetes", main = "")
```




```
plot(density(pima_NA$diastolic, na.rm = TRUE), main = "")


plot(diabetes ~ diastolic, pima_NA)
```




```
plot(diabetes ~ test, pima_NA)


library(ggplot2)
ggplot(pima_NA, aes(y = insulin, x = glucose ,
                    shape = test, color = test)) +
  geom_point(size = 3, alpha = .7) +
  theme_classic(base_size = 18) +
  theme(legend.position = "top") +
  xlab("glucose") +
  ylab("Insulina")
```



Matriz de correlação entre variáveis quantitativas

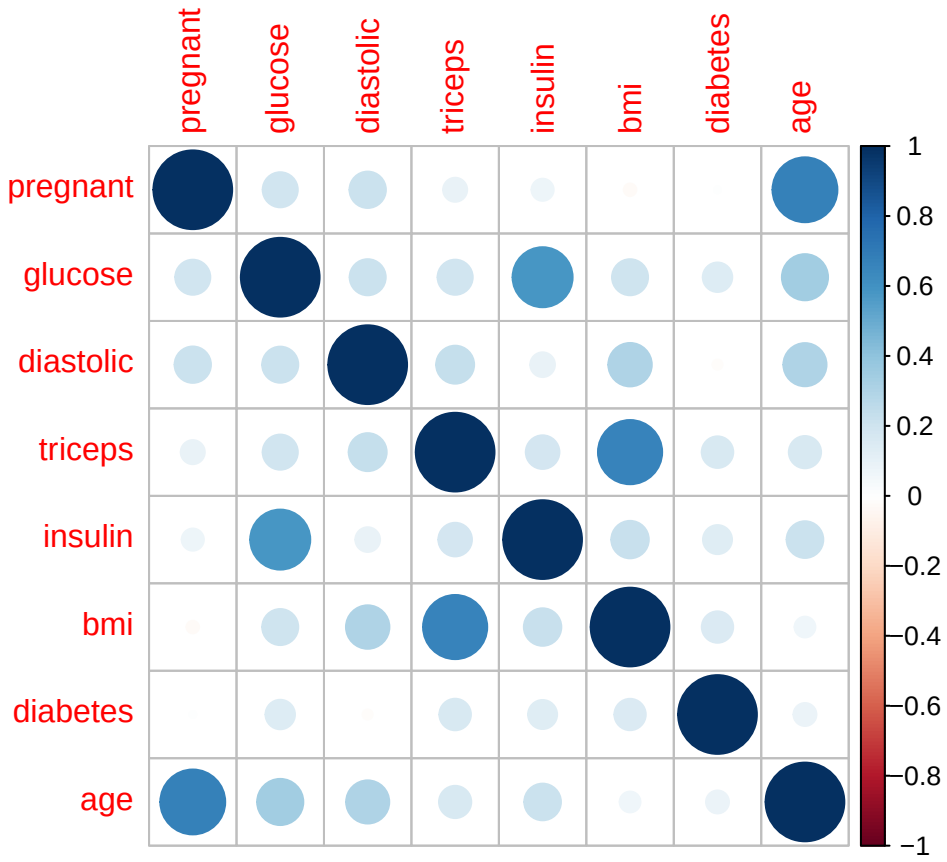
```
#install.packages("corrplot")
library(corrplot)
```

```
## corrplot 0.92 loaded
```

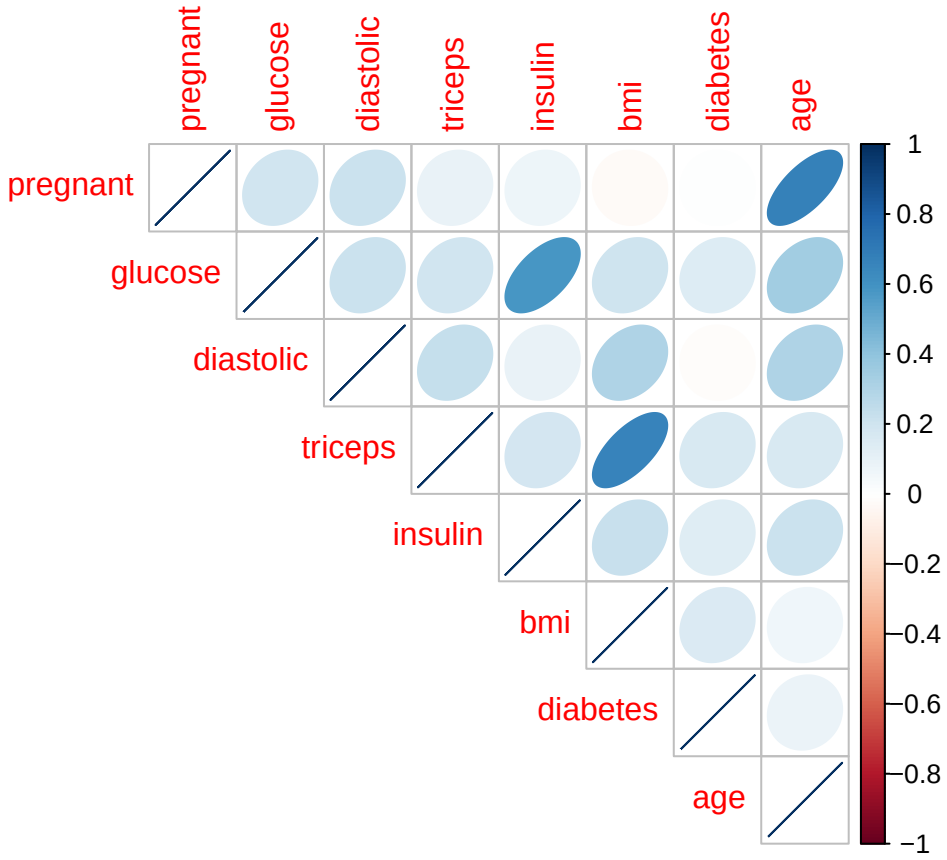
```
pima_q<-pima_NA[,-9]
head(pima_q)
```

```
##    pregnant glucose diastolic triceps insulin  bmi diabetes age
## 4         1      89        66      23      94 28.1    0.167  21
## 5         0     137        40      35     168 43.1    2.288  33
## 7         3      78        50      32      88 31.0    0.248  26
## 9         2     197        70      45     543 30.5    0.158  53
## 14        1     189        60      23     846 30.1    0.398  59
## 15        5     166        72      19     175 25.8    0.587  51
```

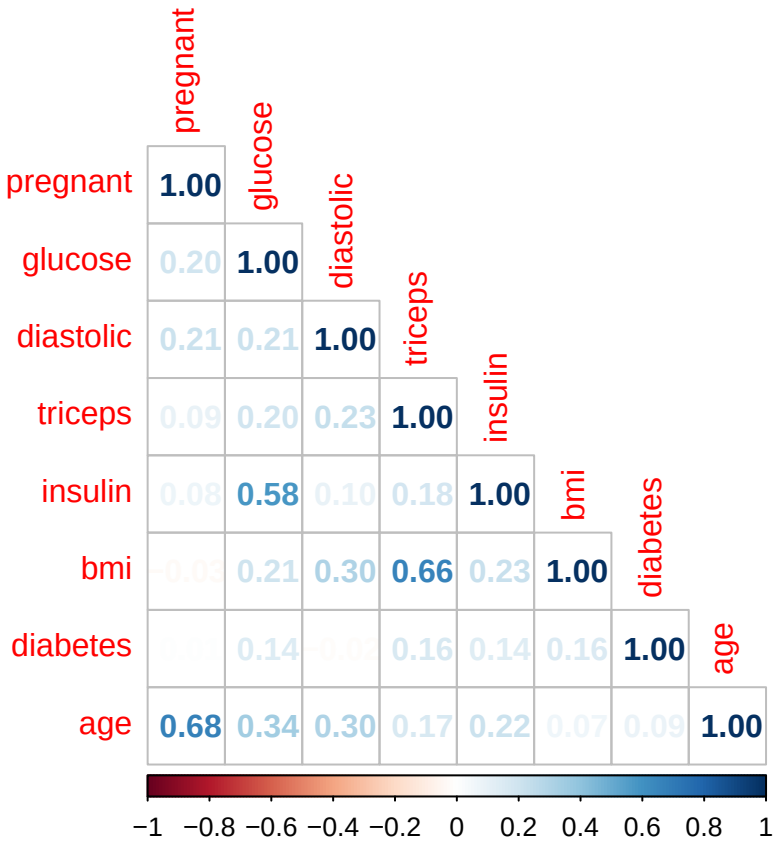
```
M <- cor(pima_q)
corrplot(M, method = "circle")
```



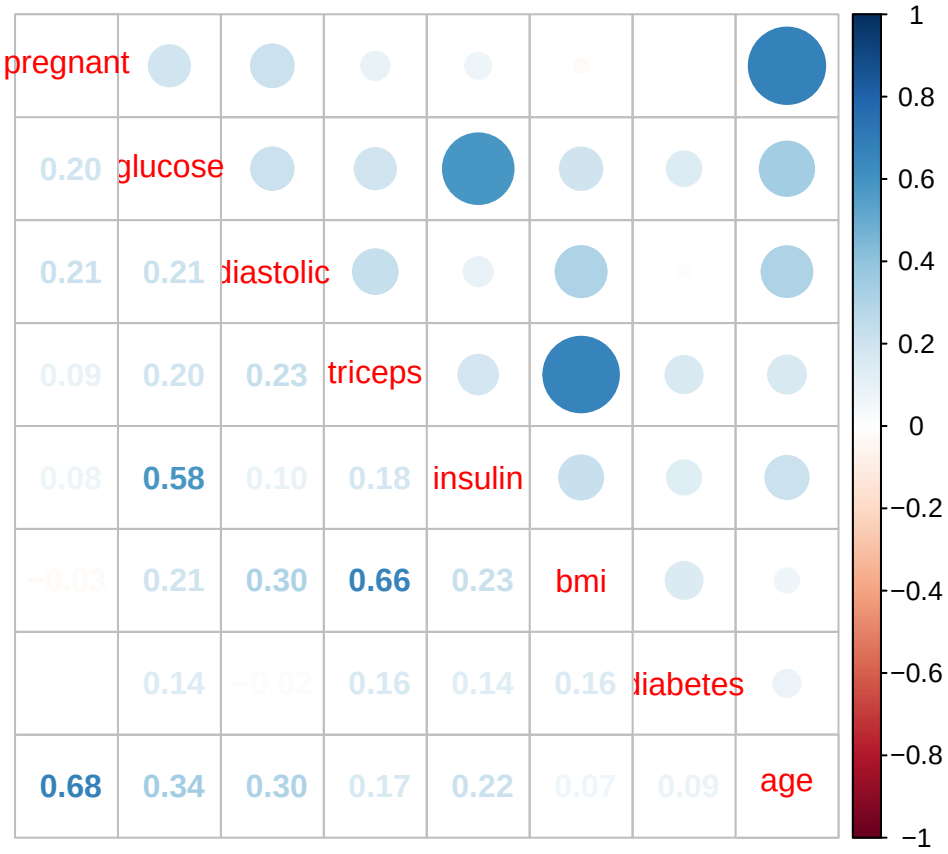

```
corrplot(M, method = "ellipse", type = "upper")
```



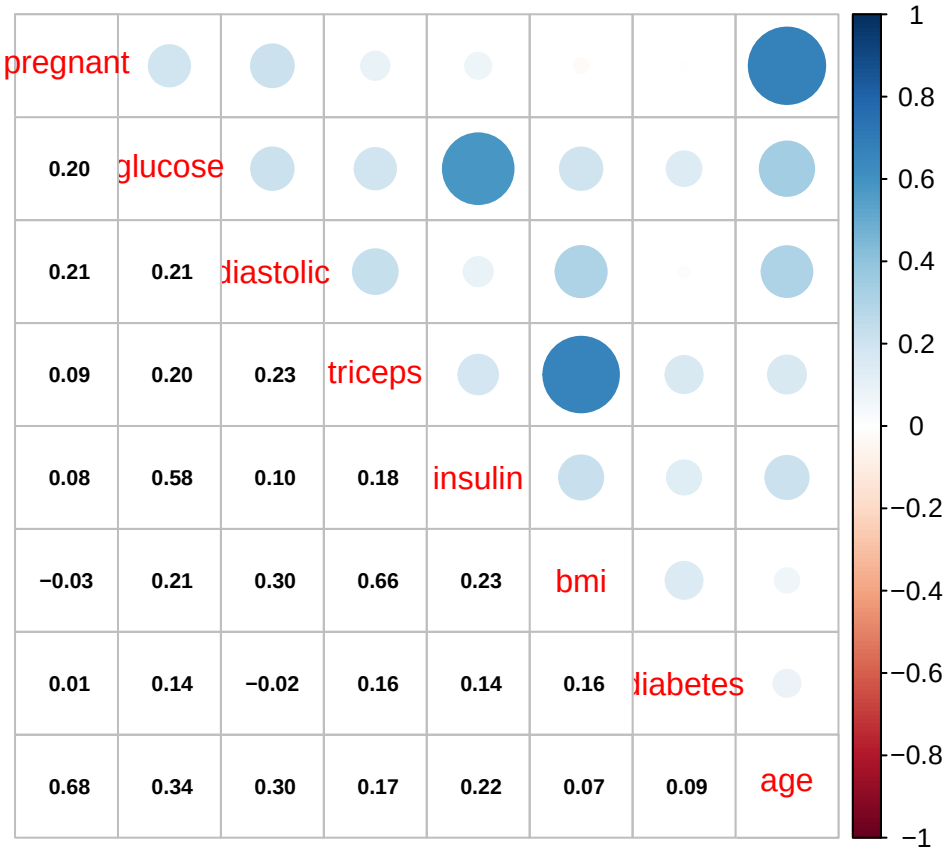

```
corrplot(M, method = "number", type = "lower")
```



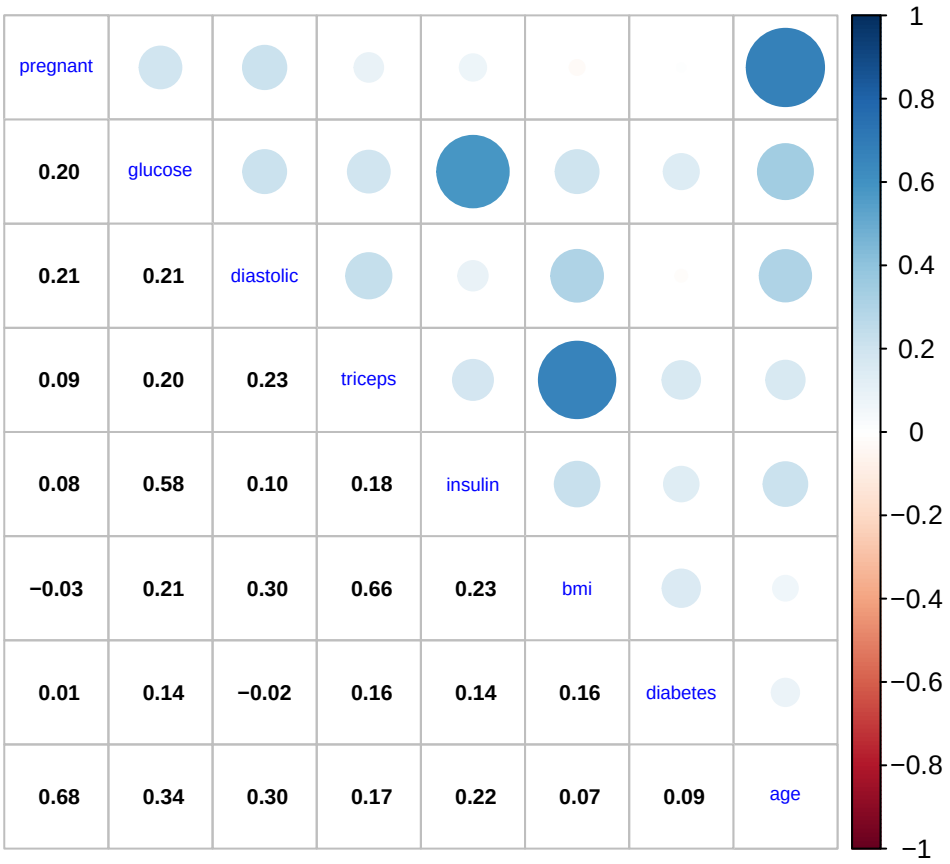

```
corrplot.mixed(M)
```




```
corrplot.mixed(M,lower.col = "black", number.cex = .7)
```




```
corrplot.mixed(M,lower.col = "black", number.cex = .7, tl.col = "blue", tl.cex = .6)
```




```
lmod <- lm(diabetes ~ ., pima_NA)
coef(lmod)
```

```
##   (Intercept)      pregnant       glucose     diastolic       triceps
##  4.016211e-01 -8.290548e-03  4.401971e-05 -2.802941e-03  2.499518e-03
##       insulin           bmi           age  testpositive
##  1.522927e-04  3.677822e-03  2.705798e-03  1.195425e-01
```

```
summary(lmod)
```

```
##
## Call:
## lm(formula = diabetes ~ ., data = pima_NA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.60980 -0.22779 -0.06745  0.15580  1.71843
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.016e-01  1.329e-01   3.022  0.00268 **
## pregnant     -8.291e-03  7.304e-03  -1.135  0.25709
## glucose       4.402e-05  7.667e-04   0.057  0.95424
## diastolic    -2.803e-03  1.502e-03  -1.866  0.06283 .
## triceps       2.500e-03  2.197e-03   1.138  0.25588
## insulin       1.523e-04  1.783e-04   0.854  0.39345
## bmi           3.678e-03  3.422e-03   1.075  0.28312
```

```
## age           2.706e-03  2.441e-03   1.109  0.26829
## testpositive  1.195e-01  4.414e-02   2.708  0.00707 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.336 on 383 degrees of freedom
## Multiple R-squared:  0.07373,    Adjusted R-squared:  0.05439
## F-statistic: 3.811 on 8 and 383 DF,  p-value: 0.0002525
```

```
#install.packages("tidyverse")
#install.packages("caret")
#install.packages("leaps")
library(MASS)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x dplyr::select() masks MASS::select()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(caret)
```

```
## Loading required package: lattice
##
## Attaching package: 'lattice'
##
## The following object is masked from 'package:faraway':
##
##     melanoma
##
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(leaps)


step.model <- stepAIC(lmod, direction = "both",
                      trace = FALSE)
summary(step.model)
```

```
##
## Call:
## lm(formula = diabetes ~ diastolic + bmi + test, data = pima_NA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -0.53252 -0.22906 -0.08082  0.14938  1.82645
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.439474   0.112731   3.898 0.000114 ***
## diastolic    -0.002601   0.001437  -1.810 0.070994 .
## bmi           0.006680   0.002604   2.566 0.010673 *
## testpositive  0.139802   0.037681   3.710 0.000237 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3357 on 388 degrees of freedom
## Multiple R-squared:  0.06301,    Adjusted R-squared:  0.05576
## F-statistic: 8.697 on 3 and 388 DF,  p-value: 1.347e-05
```

```
step.model_back <- stepAIC(lmod, direction = "backward",
                      trace = FALSE)
summary(step.model_back)
```

```
##
## Call:
## lm(formula = diabetes ~ diastolic + bmi + test, data = pima_NA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.53252 -0.22906 -0.08082  0.14938  1.82645
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.439474   0.112731   3.898 0.000114 ***
## diastolic    -0.002601   0.001437  -1.810 0.070994 .
## bmi           0.006680   0.002604   2.566 0.010673 *
## testpositive  0.139802   0.037681   3.710 0.000237 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3357 on 388 degrees of freedom
## Multiple R-squared:  0.06301,    Adjusted R-squared:  0.05576
## F-statistic: 8.697 on 3 and 388 DF,  p-value: 1.347e-05
```

```
step.model_for <- stepAIC(lmod, direction = "forward",
                      trace = FALSE)
summary(step.model_for)
```

```
##
## Call:
## lm(formula = diabetes ~ pregnant + glucose + diastolic + triceps +
##     insulin + bmi + age + test, data = pima_NA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.60980 -0.22779 -0.06745  0.15580  1.71843
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)   4.016e-01  1.329e-01   3.022  0.00268 **
## pregnant     -8.291e-03  7.304e-03  -1.135  0.25709
## glucose       4.402e-05  7.667e-04   0.057  0.95424
## diastolic    -2.803e-03  1.502e-03  -1.866  0.06283 .
## triceps       2.500e-03  2.197e-03   1.138  0.25588
## insulin       1.523e-04  1.783e-04   0.854  0.39345
## bmi           3.678e-03  3.422e-03   1.075  0.28312
## age           2.706e-03  2.441e-03   1.109  0.26829
## testpositive  1.195e-01  4.414e-02   2.708  0.00707 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.336 on 383 degrees of freedom
## Multiple R-squared:  0.07373,    Adjusted R-squared:  0.05439
## F-statistic: 3.811 on 8 and 383 DF,  p-value: 0.0002525
```

```
AIC(step.model)
```

```
## [1] 262.7021
```

```
AIC(step.model_back)
```

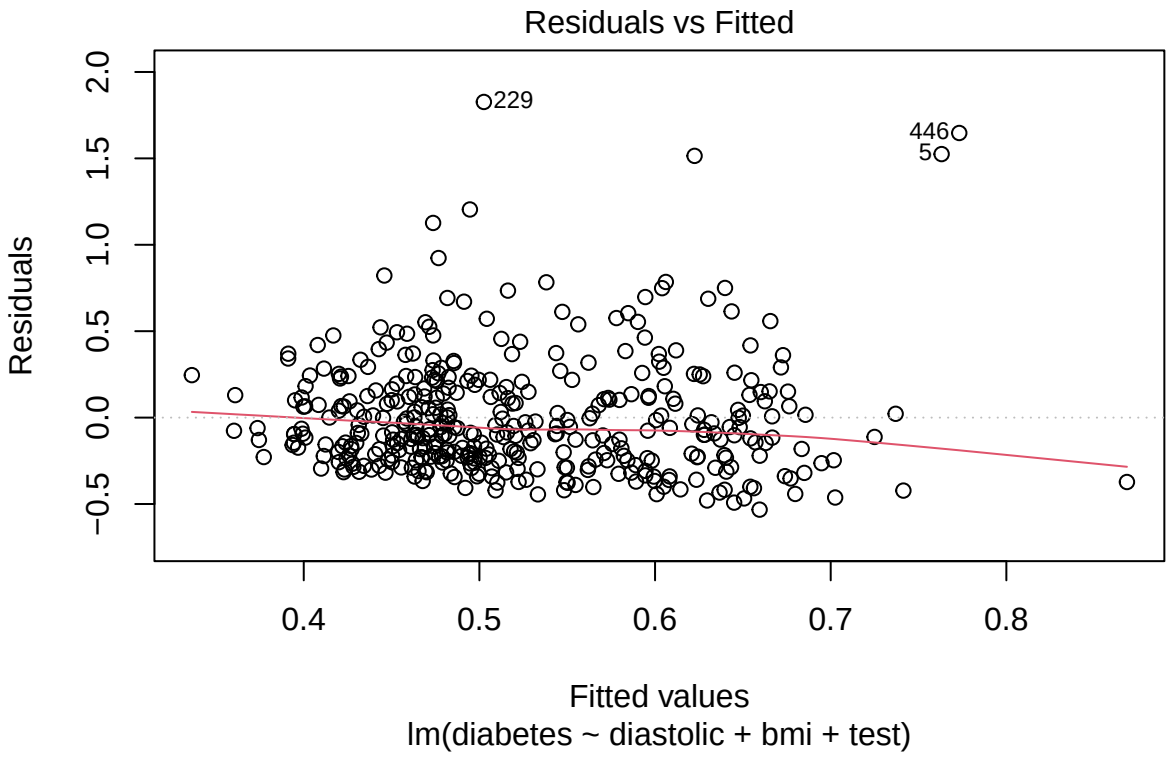
```
## [1] 262.7021
```

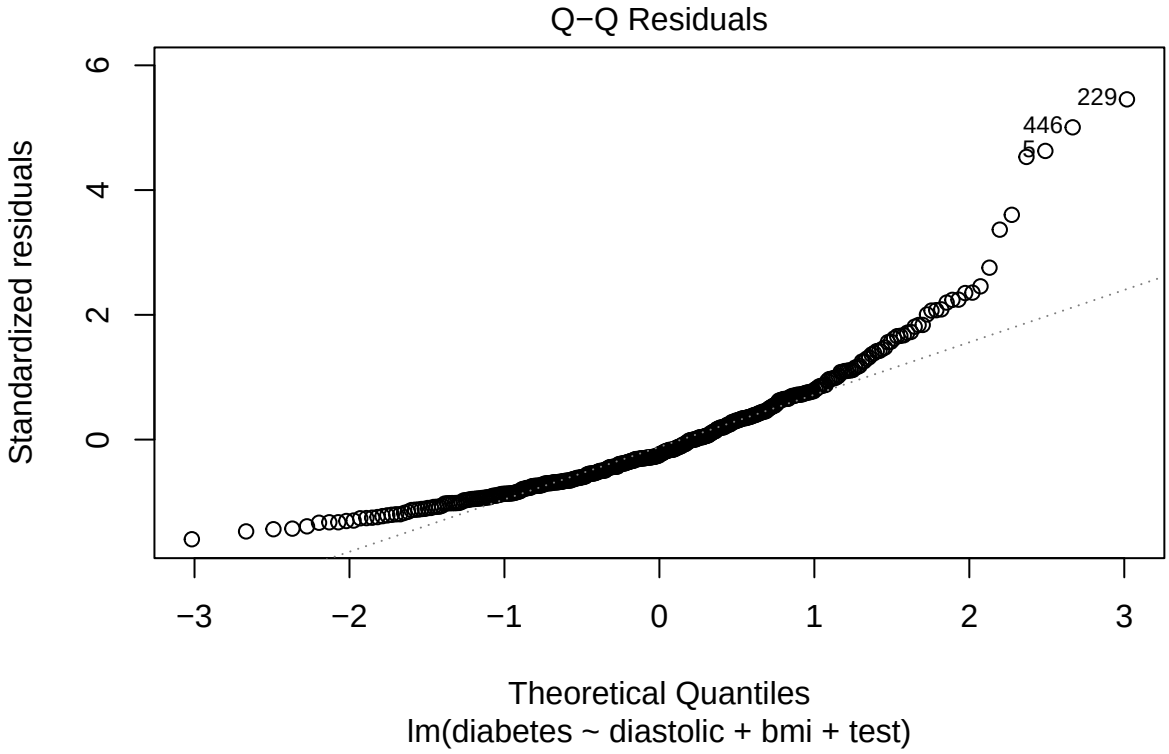
```
AIC(step.model_for)
```

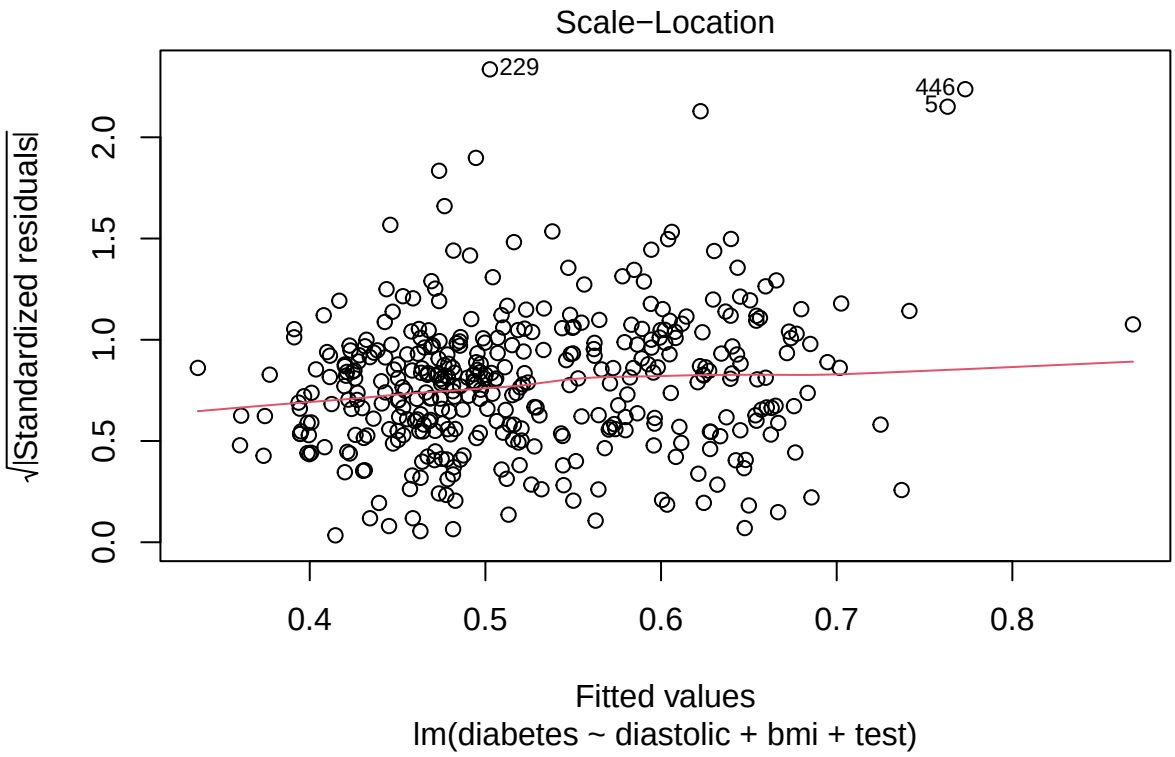
```
## [1] 268.1891
```

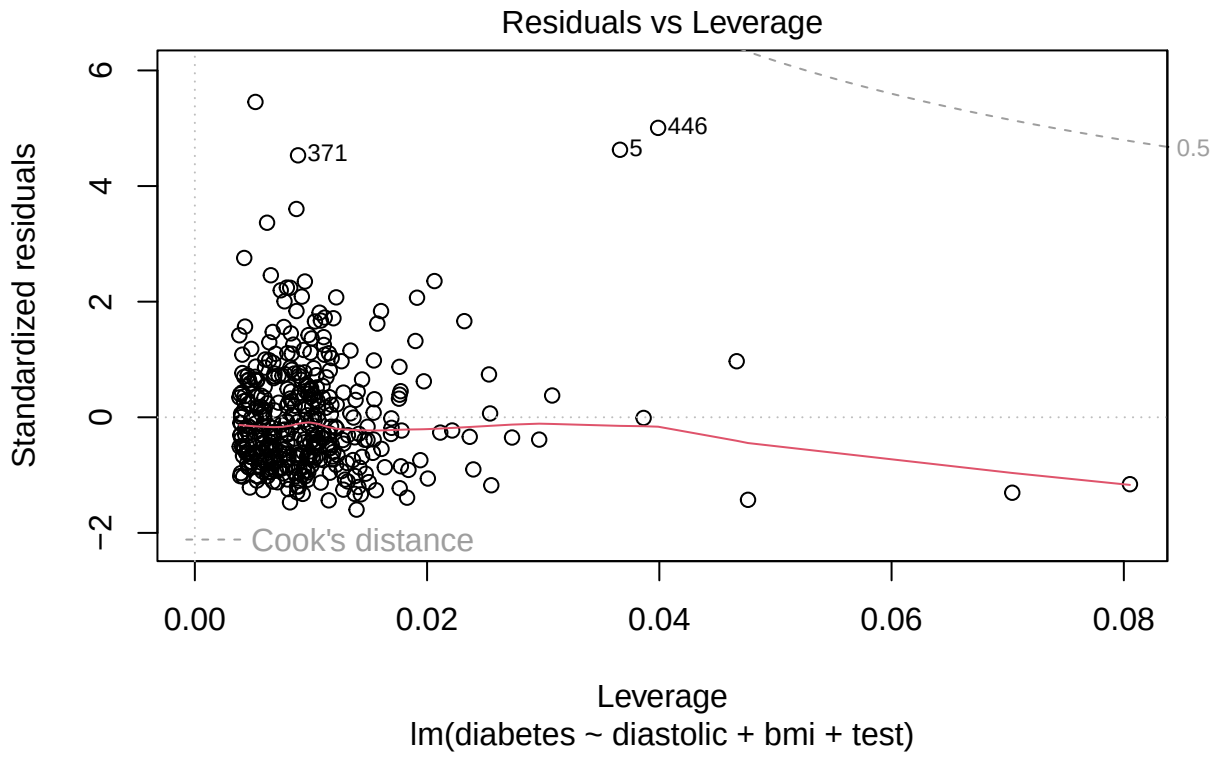
Análise de resíduos:

```
plot(step.model)
```


Residuals vs Fitted
Residuals
Fitted values
lm(diabetes ~ diastolic + bmi + test)
229
446
5
2.0
1.5
1.0
0.5
0.0
−0.5
0.4
0.5
0.6
0.7
0.8


Q−Q Residuals
Standardized residuals
0
2
4
6
229
446
5
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(diabetes ~ diastolic + bmi + test)

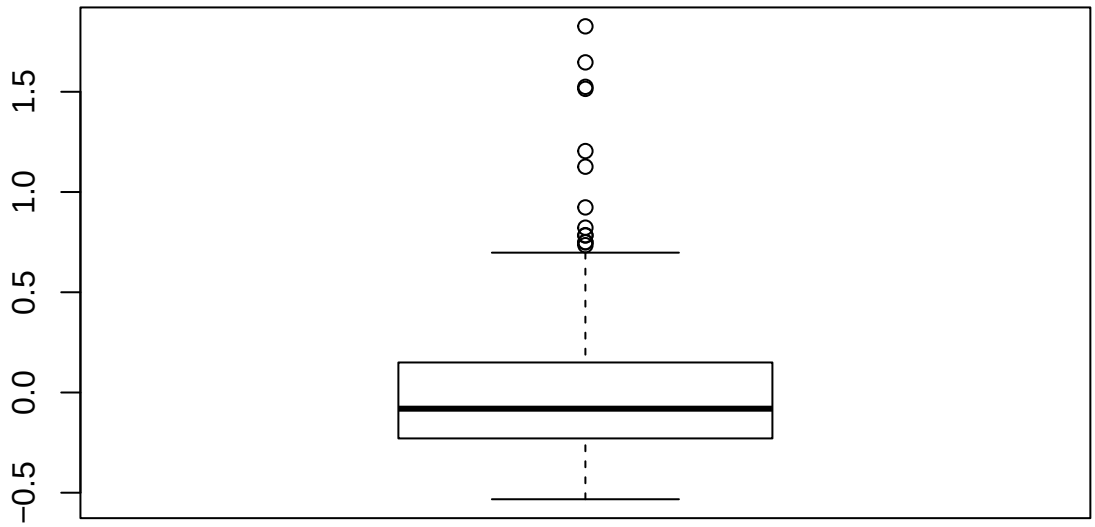

Scale−Location
229
446
5
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
0.4
0.5
0.6
0.7
0.8
Fitted values
lm(diabetes ~ diastolic + bmi + test)



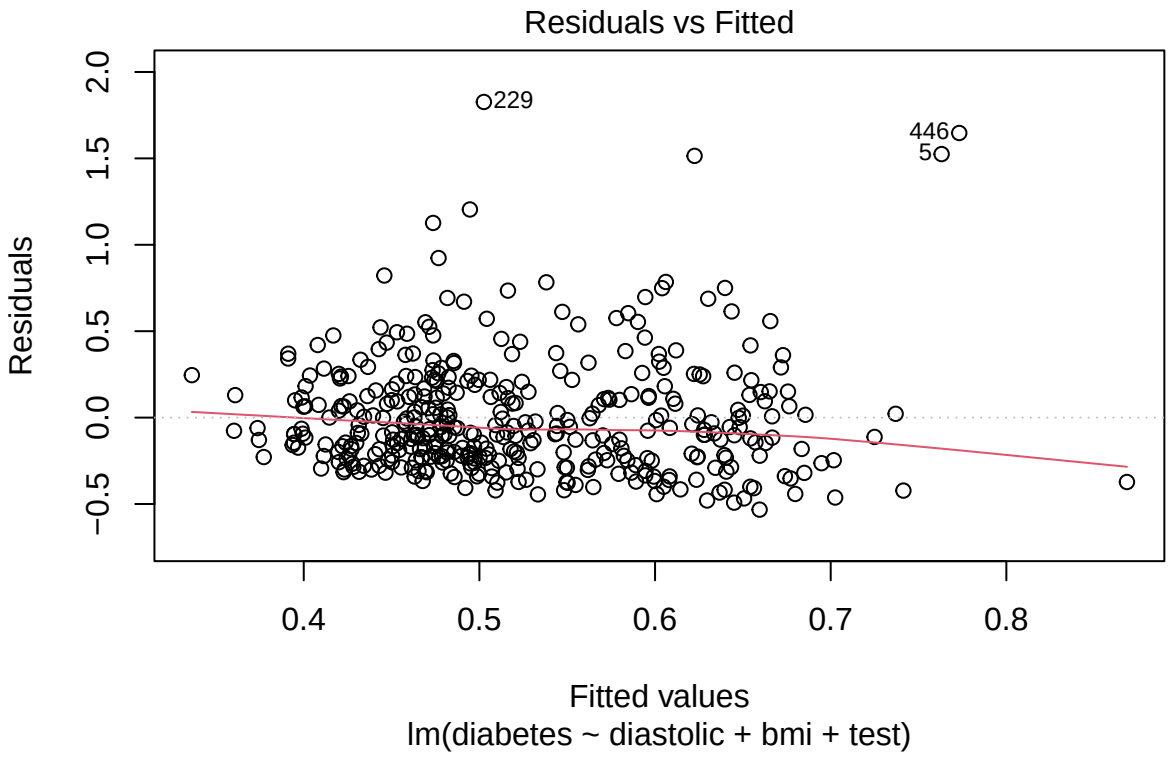

```
shapiro.test(residuals(step.model))
```

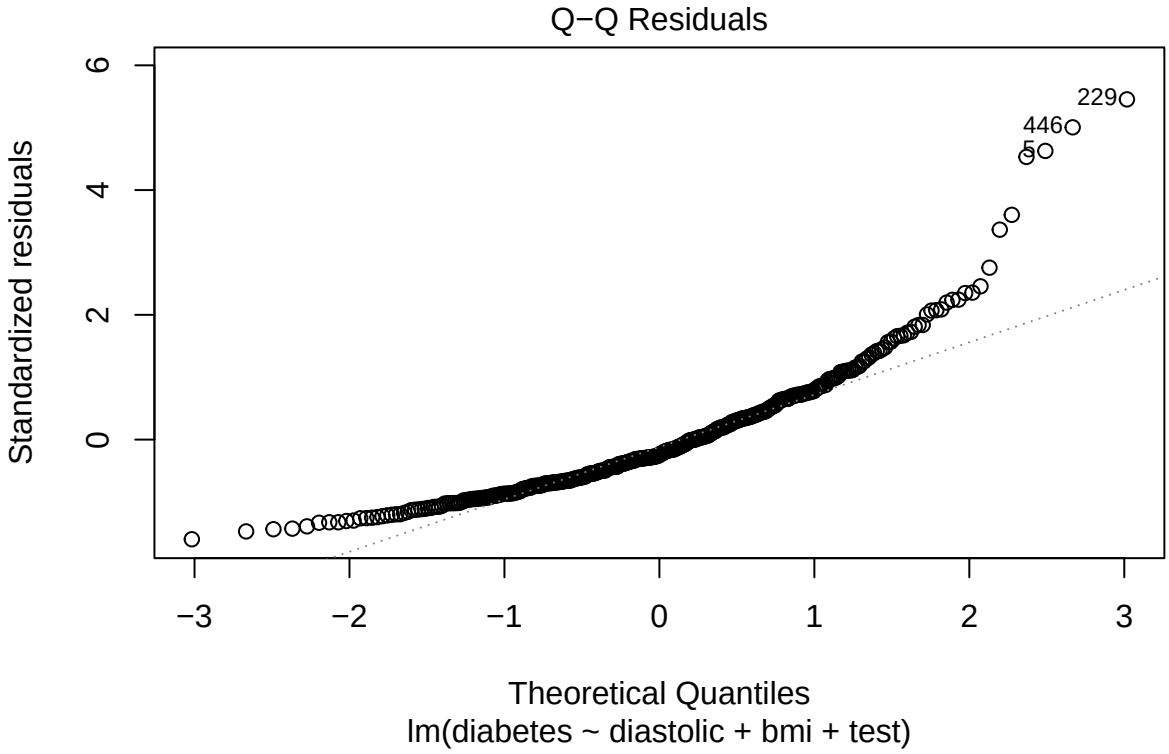
```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(step.model)
## W = 0.87701, p-value < 2.2e-16
```

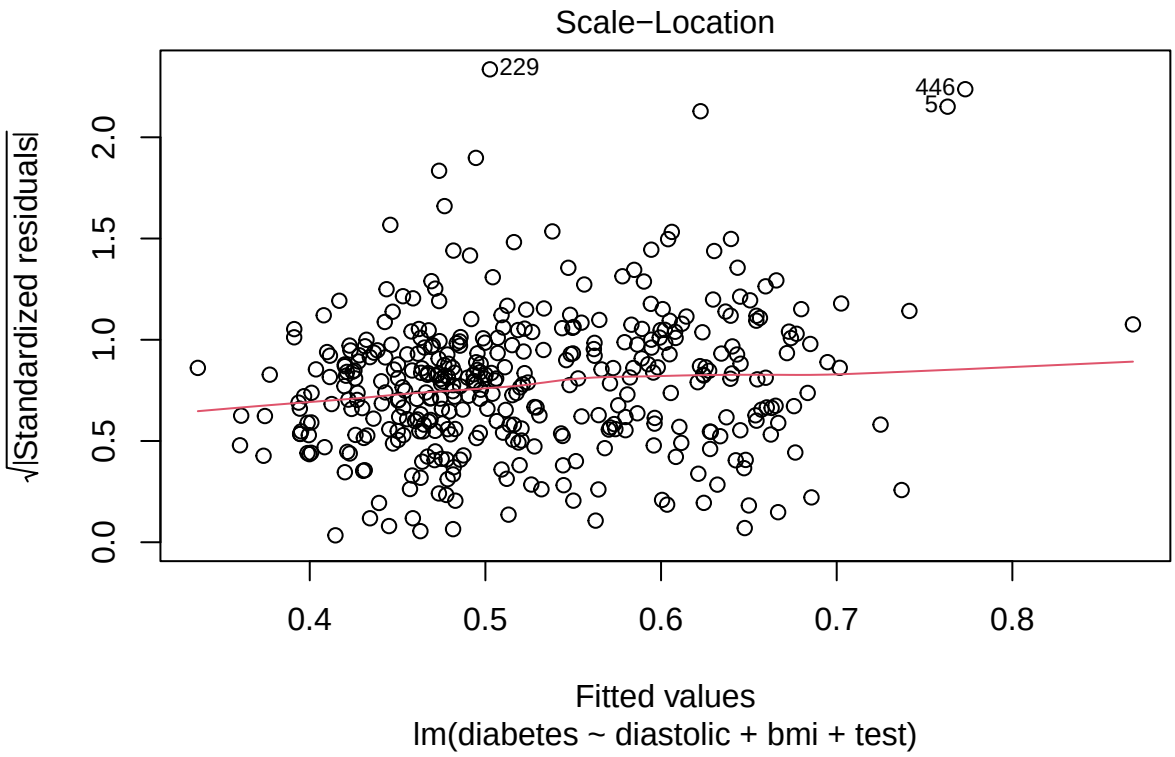
```
boxplot(residuals(step.model),col="white")
```

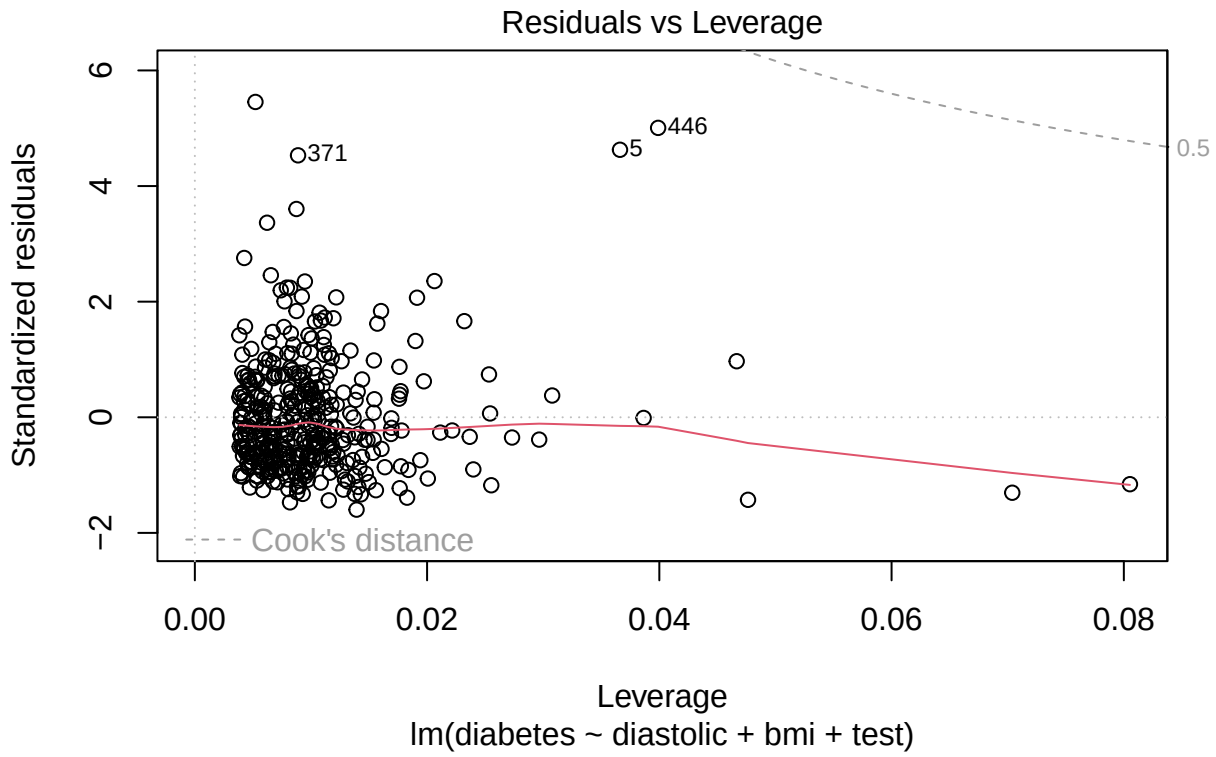



```
plot(step.model_back)
```


Residuals vs Fitted
Residuals
Fitted values
lm(diabetes ~ diastolic + bmi + test)
229
446
5
−0.5
0.0
0.5
1.0
1.5
2.0
0.4
0.5
0.6
0.7
0.8


Q−Q Residuals
Standardized residuals
0
2
4
6
229
446
5
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(diabetes ~ diastolic + bmi + test)

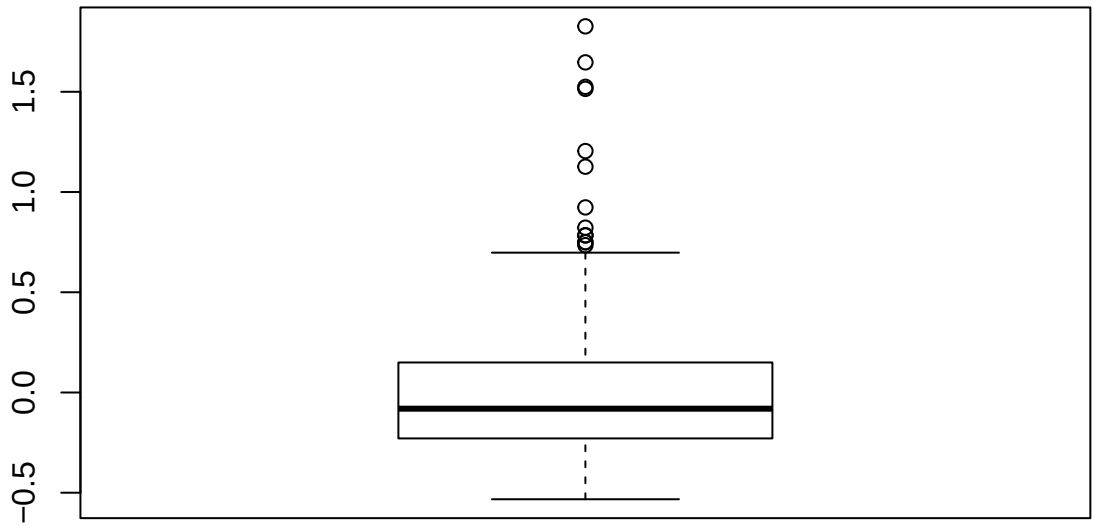

Scale−Location
229
446
5
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
0.4
0.5
0.6
0.7
0.8
Fitted values
lm(diabetes ~ diastolic + bmi + test)



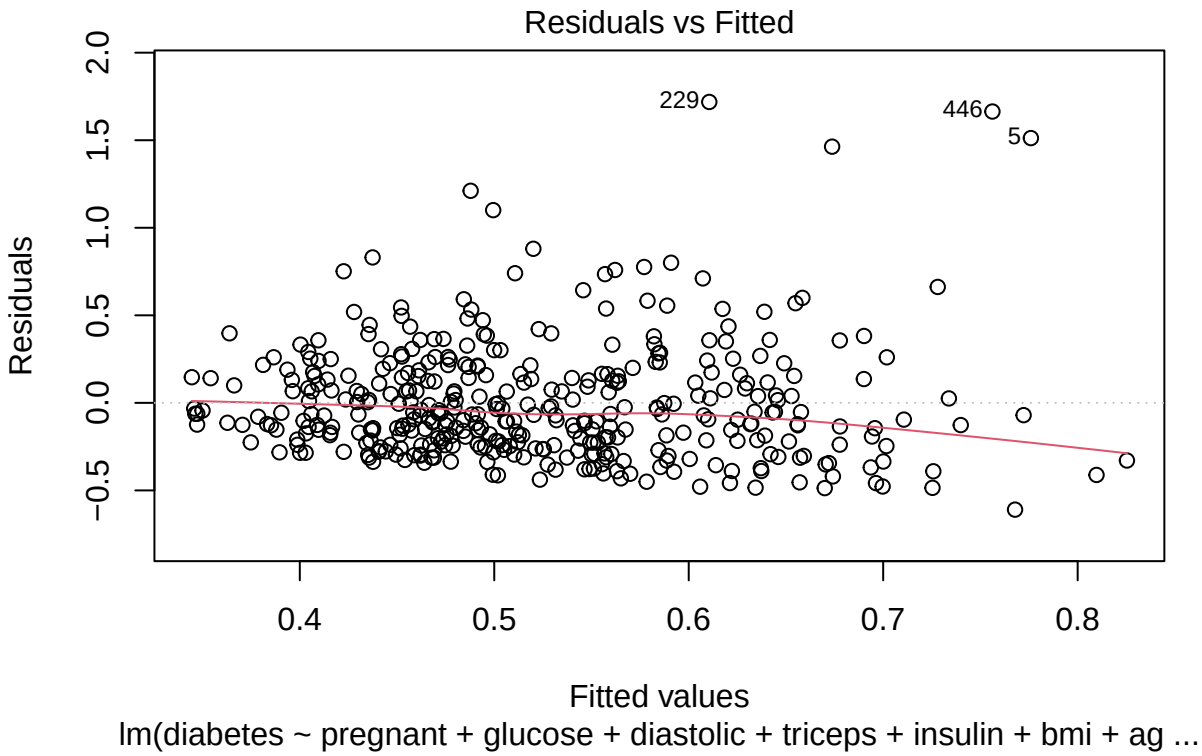

```
shapiro.test(residuals(step.model_back))
```

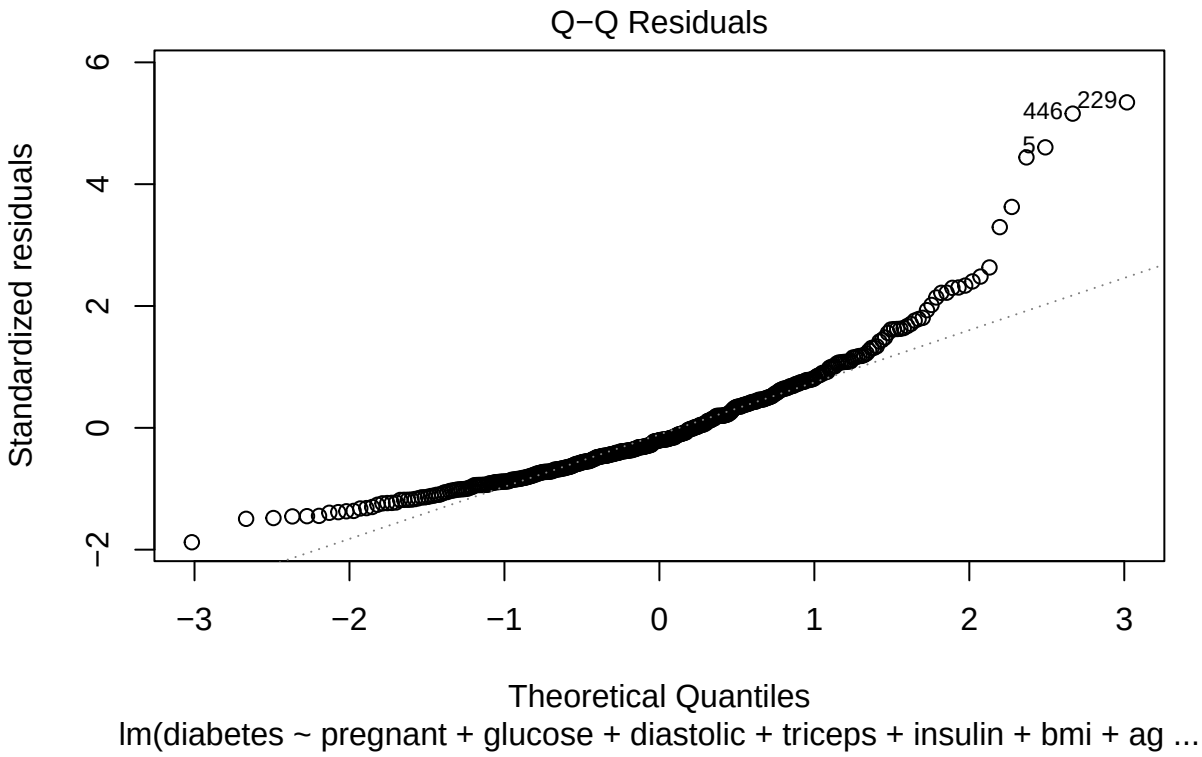
```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(step.model_back)
## W = 0.87701, p-value < 2.2e-16
```

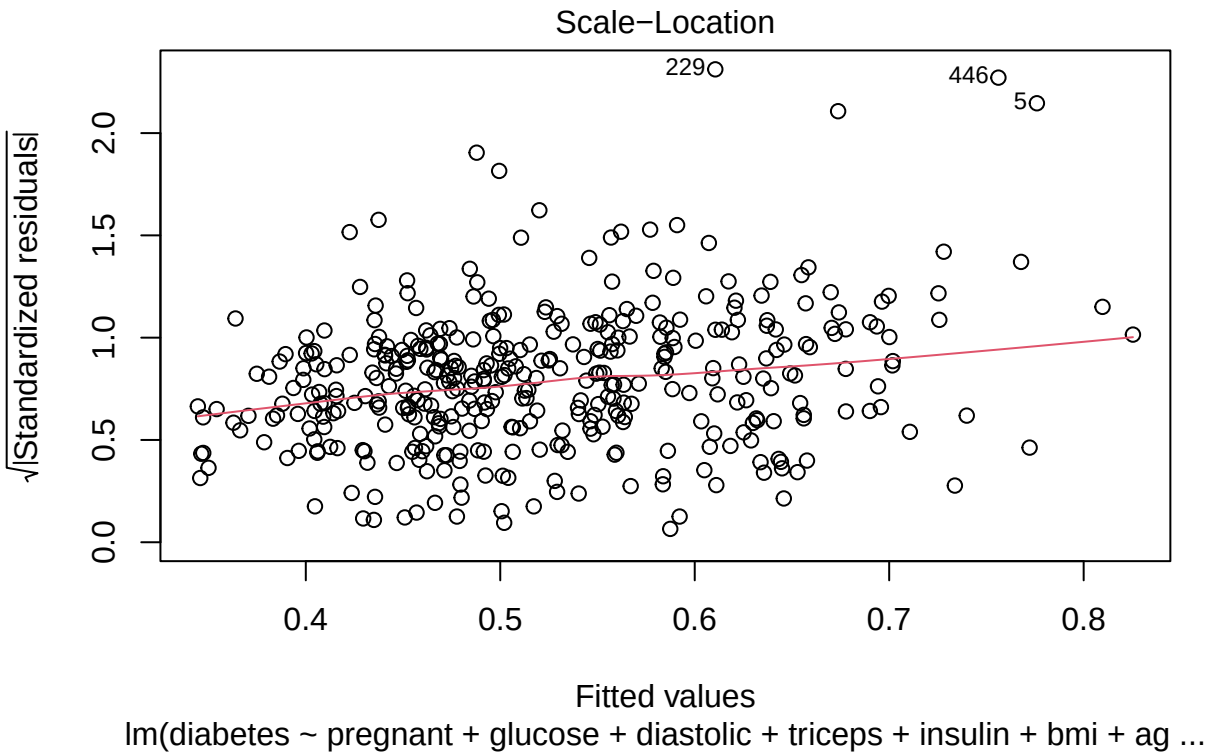
```
boxplot(residuals(step.model_back),col="white")
```

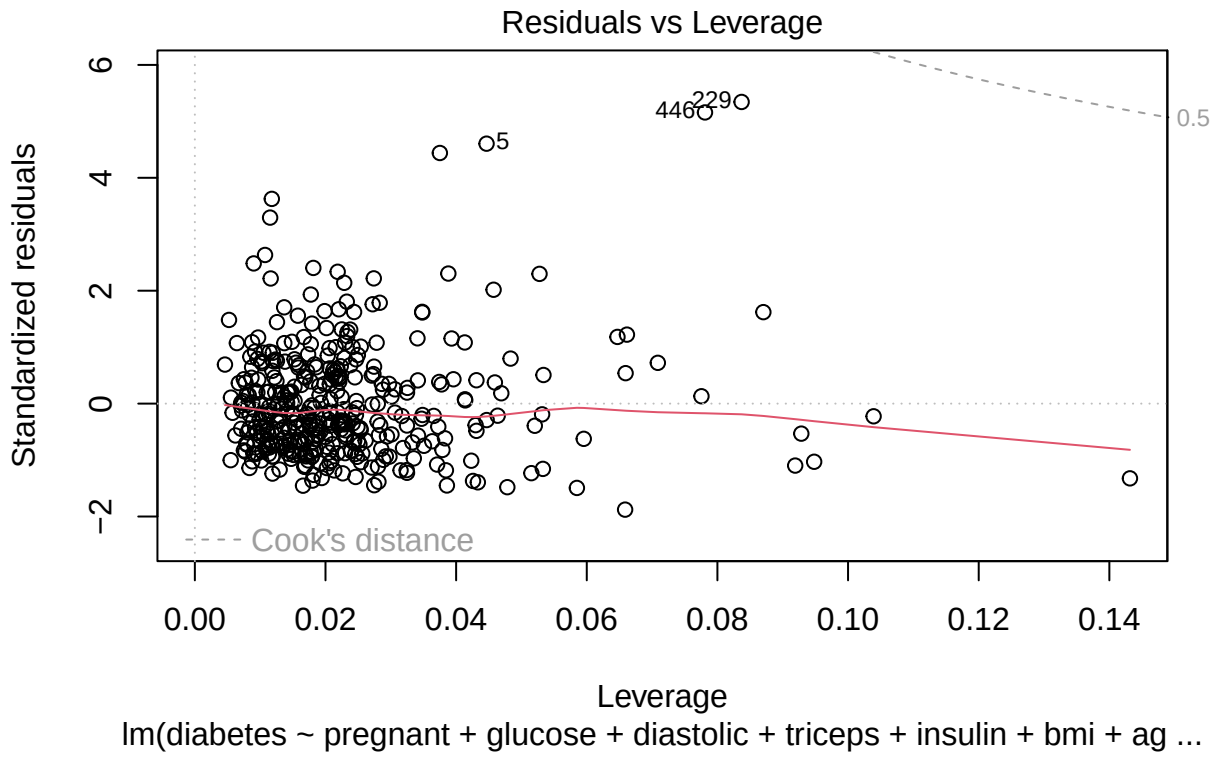



```
plot(step.model_for)
```


Residuals vs Fitted
Residuals
Fitted values
lm(diabetes ~ pregnant + glucose + diastolic + triceps + insulin + bmi + ag ...
229
446
5
2.0
1.5
1.0
0.5
0.0
−0.5
0.4
0.5
0.6
0.7
0.8


Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(diabetes ~ pregnant + glucose + diastolic + triceps + insulin + bmi + ag ...
446
229
5
−2
0
2
4
6
−3
−2
−1
0
1
2
3

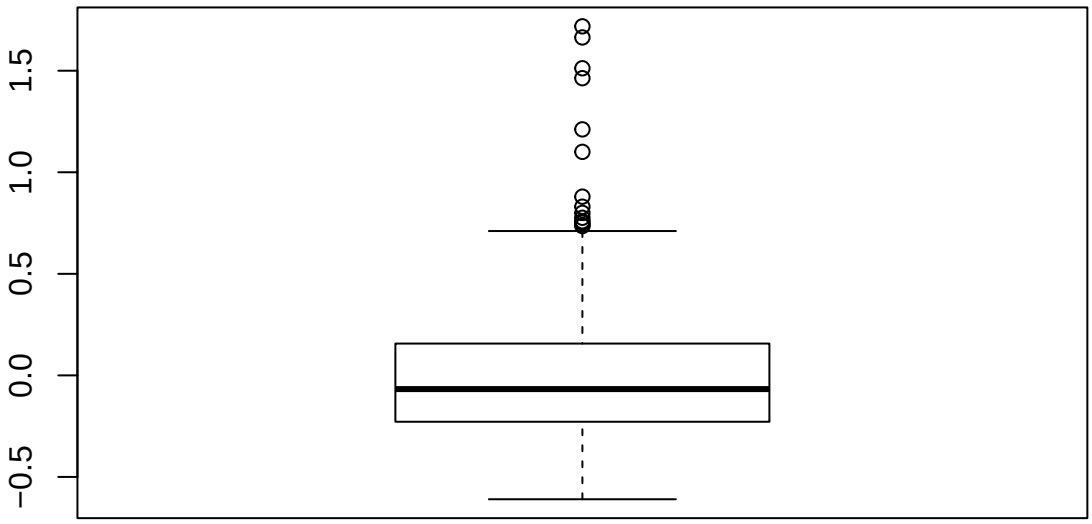

Scale−Location
229
446
5
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
0.4
0.5
0.6
0.7
0.8
Fitted values
lm(diabetes ~ pregnant + glucose + diastolic + triceps + insulin + bmi + ag ...




```
shapiro.test(residuals(step.model_for))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(step.model_for)
## W = 0.8882, p-value = 2.673e-16
```

```
boxplot(residuals(step.model_for),col="white")
```


1.5
1.0
0.5
0.0
−0.5