Analysis-COPD

Lina

2024-05-29

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
#################################################################
# Housekeeping Use for All Analyses #
#################################################################
date() # Current system time and date.
```

```
## [1] "Wed Jun 26 12:06:54 2024"
```

```
Sys.time() # Current system time and date (redundant).
```

```
## [1] "2024-06-26 12:06:54 +07"
```

```
R.version.string # R version and version release date.
```

```
## [1] "R version 4.2.3 (2023-03-15 ucrt)"
```

```
options(digits=6) # Confirm default digits.
options(scipen=999)# Suppress scientific notation.
options(width=60) # Confirm output width.
ls() # List all objects in the working # directory.
```

```
## character(0)
```

```
rm(list = ls()) # CAUTION: Remove all files in the #working directory. If this action is not desired, u
ls.str() # List all objects with finite detail.
getwd() # Identify the current working directory
```

```
## [1] "C:/Users/linan/Documents/GitHub/R-COPD-regression-modelling"
```

```r
setwd("C:/Users/linan/Documents/GitHub/R-COPD-regression-modelling") # Set to a new working directory.
getwd()# Confirm the working directory.
```

```
## [1] "C:/Users/linan/Documents/GitHub/R-COPD-regression-modelling"
```

```r
list.files()# List files at the PC directory
```

```
## [1] "~$alysis-report.docx"
## [2] "10.1177_1479972317694622.pdf"
## [3] "analysis-report.docx"
## [4] "copd-12-467.pdf"
## [5] "copd-multivariate-modelling.Rmd"
## [6] "copd-multivariate-modelling_files"
## [7] "COPD_student_dataset.csv"
```

```r
.libPaths()# Library pathname
```

```
## [1] "C:/Users/linan/AppData/Local/R/win-library/4.2"
## [2] "C:/Program Files/R/R-4.2.3/library"
```

```r
.Library# Library pathname.
```

```
## [1] "C:/PROGRA~1/R/R-42~1.3/library"
```

```r
sessionInfo() # R version, locale, and packages.
```

```
## R version 4.2.3 (2023-03-15 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 22631)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_Indonesia.utf8
## [2] LC_CTYPE=English_Indonesia.utf8
## [3] LC_MONETARY=English_Indonesia.utf8
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_Indonesia.utf8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets
## [6] methods   base
##
## loaded via a namespace (and not attached):
##  [1] compiler_4.2.3    fastmap_1.1.1     cli_3.6.1
##  [4] tools_4.2.3       htmltools_0.5.8   rstudioapi_0.16.0
##  [7] yaml_2.3.8        rmarkdown_2.26    knitr_1.45
## [10] xfun_0.40         digest_0.6.31     rlang_1.1.1
## [13] evaluate_0.23
```

```r
search()# Attached packages and objects.
```

```
## [1] ".GlobalEnv"        "package:stats"
## [3] "package:graphics"  "package:grDevices"
## [5] "package:utils"     "package:datasets"
## [7] "package:methods"   "Autoloads"
## [9] "package:base"
```

```r
searchpaths() # Attached packages and objects.
```

```
## [1] ".GlobalEnv"
## [2] "C:/Program Files/R/R-4.2.3/library/stats"
## [3] "C:/Program Files/R/R-4.2.3/library/graphics"
## [4] "C:/Program Files/R/R-4.2.3/library/grDevices"
## [5] "C:/Program Files/R/R-4.2.3/library/utils"
## [6] "C:/Program Files/R/R-4.2.3/library/datasets"
## [7] "C:/Program Files/R/R-4.2.3/library/methods"
## [8] "Autoloads"
## [9] "C:/PROGRA~1/R/R-42~1.3/library/base"
```

```r
######################################################################
```

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(Hmisc)
```

```
##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:dplyr':
##
##     src, summarize

## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```r
library(gmodels)
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ---- tidyverse 2.0.0 --
## 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     1.3.1
## v readr     2.1.5

## -- Conflicts ---------------------- tidyverse_conflicts() --
## x dplyr::filter()    masks stats::filter()
## x dplyr::lag()       masks stats::lag()
## x Hmisc::src()       masks dplyr::src()
## x Hmisc::summarize() masks dplyr::summarize()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(mctest)
```

MULTIVARIATE LINEAR REGRESSION MODEL

The aim of this project is to predict the association of disease severity to quality of life in patients with COPD.

Data consists of 101 observations with 24 variables measured.

```r
copd <- read.table(file="COPD_student_dataset.csv", header=TRUE, dec=".", sep = ",")
```

```r
str(copd)
```

```
## 'data.frame':    101 obs. of  24 variables:
##  $ X           : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ ID          : int  58 57 62 145 136 84 93 27 114 152 ...
##  $ AGE         : int  77 79 80 56 65 67 67 83 72 75 ...
##  $ PackHistory : num  60 50 11 60 68 26 50 90 50 6 ...
##  $ COPDSEVERITY: chr  "SEVERE" "MODERATE" "MODERATE" "VERY SEVERE" ...
##  $ MWT1        : int  120 165 201 210 204 216 214 214 231 226 ...
##  $ MWT2        : int  120 176 180 210 210 180 237 237 237 240 ...
##  $ MWT1Best    : int  120 176 201 210 210 216 237 237 237 240 ...
##  $ FEV1        : num  1.21 1.09 1.52 0.47 1.07 1.09 0.69 0.68 2.13 1.06 ...
##  $ FEV1PRED    : num  36 56 68 14 42 50 35 32 63 46 ...
##  $ FVC         : num  2.4 1.64 2.3 1.14 2.91 1.99 1.31 2.23 4.38 2.06 ...
##  $ FVCPRED     : int  98 65 86 27 98 60 48 77 80 75 ...
##  $ CAT         : int  25 12 22 28 32 29 29 22 25 31 ...
##  $ HAD         : num  8 21 18 26 18 21 30 2 6 20 ...
##  $ SGRQ        : num  69.5 44.2 44.1 62 75.6 ...
##  $ AGEquartiles: int  4 4 4 1 1 2 2 4 3 3 ...
##  $ copd        : int  3 2 2 4 3 2 3 3 2 3 ...
##  $ gender      : int  1 0 0 1 1 0 0 1 1 0 ...
##  $ smoking     : int  2 2 2 2 2 1 1 2 1 2 ...
##  $ Diabetes    : int  1 1 1 0 0 1 1 1 1 0 ...
##  $ muscular    : int  0 0 0 0 1 0 0 0 0 1 ...
##  $ hypertension: int  0 0 0 1 1 0 0 0 0 0 ...
##  $ AtrialFib   : int  1 1 1 1 0 1 1 1 1 0 ...
##  $ IHD         : int  0 1 0 0 0 0 0 0 0 0 ...
```

Variables

Characters : Age, Gender, Pack History, Smoking Disease : COPDSeverity, CAT Walking ability : MWT1, MWT2, MWT1Best Lung function : FEV1, FEV1PRED, FVC, FVCPRED Anxiety&Depression : HAD QOL : SGRQ Comorbidities : Diabetes, Muscular, Hypertension, AtrialFib, IHD

numeric : Age, PackHistory, FEV, FEV1PRED, FVC, FVCPRED, CAT, HAD, MWT1, MWT2, MWT1Best, SGRQ factor : Gender, COPDseverity, copd, smoking, Diabetes, Muscular, Hypertension, AtrialFib, IHD

Change variable type :

```
#Numeric data
copd$AGE <- as.numeric(copd$AGE)
copd$MWT1 <- as.numeric(copd$MWT1)
copd$MWT2 <- as.numeric(copd$MWT2)
copd$MWT1Best<-as.numeric(copd$MWT1Best)
copd$FEV1PRED <- as.numeric(copd$FEV1PRED)
copd$FVCPRED <- as.numeric(copd$FVCPRED)
copd$CAT <- as.numeric(copd$CAT)
```

```
#Categorical cata
copd$AGEquartiles <- as.factor(copd$AGEquartiles)
copd$copd <- as.factor(copd$copd)
copd$gender <- as.factor(copd$gender)
copd$Diabetes <- as.factor(copd$Diabetes)
copd$smoking <- as.factor(copd$smoking)
copd$muscular <- as.factor(copd$muscular)
copd$hypertension <- as.factor(copd$hypertension)
copd$AtrialFib <- as.factor(copd$AtrialFib)
copd$IHD <- as.factor(copd$IHD)
```

```
#Check each data types
str(copd)
```

```
## 'data.frame':    101 obs. of  24 variables:
##  $ X           : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ ID          : int  58 57 62 145 136 84 93 27 114 152 ...
##  $ AGE         : num  77 79 80 56 65 67 67 83 72 75 ...
##  $ PackHistory : num  60 50 11 60 68 26 50 90 50 6 ...
##  $ COPDSEVERITY: chr  "SEVERE" "MODERATE" "MODERATE" "VERY SEVERE" ...
##  $ MWT1        : num  120 165 201 210 204 216 214 214 231 226 ...
##  $ MWT2        : num  120 176 180 210 210 180 237 237 237 240 ...
##  $ MWT1Best    : num  120 176 201 210 210 216 237 237 237 240 ...
##  $ FEV1        : num  1.21 1.09 1.52 0.47 1.07 1.09 0.69 0.68 2.13 1.06 ...
##  $ FEV1PRED    : num  36 56 68 14 42 50 35 32 63 46 ...
##  $ FVC         : num  2.4 1.64 2.3 1.14 2.91 1.99 1.31 2.23 4.38 2.06 ...
##  $ FVCPRED     : num  98 65 86 27 98 60 48 77 80 75 ...
##  $ CAT         : num  25 12 22 28 32 29 29 22 25 31 ...
##  $ HAD         : num  8 21 18 26 18 21 30 2 6 20 ...
##  $ SGRQ        : num  69.5 44.2 44.1 62 75.6 ...
##  $ AGEquartiles: Factor w/ 4 levels "1","2","3","4": 4 4 4 1 1 2 2 4 3 3 ...
##  $ copd        : Factor w/ 4 levels "1","2","3","4": 3 2 2 4 3 2 3 3 2 3 ...
##  $ gender      : Factor w/ 2 levels "0","1": 2 1 1 2 2 1 1 2 2 1 ...
```

```
##  $ smoking     : Factor w/ 2 levels "1","2": 2 2 2 2 2 1 1 2 1 2 ...
##  $ Diabetes    : Factor w/ 2 levels "0","1": 2 2 2 1 1 2 2 2 2 1 ...
##  $ muscular    : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 1 1 1 2 ...
##  $ hypertension: Factor w/ 2 levels "0","1": 1 1 1 2 2 1 1 1 1 1 ...
##  $ AtrialFib   : Factor w/ 2 levels "0","1": 2 2 2 2 1 2 2 2 2 1 ...
##  $ IHD         : Factor w/ 2 levels "0","1": 1 2 1 1 1 1 1 1 1 1 ...
```

```
describe(copd)
```

```
## copd
##
##  24  Variables      101  Observations
## ---------------------------------------------------------
## X
##        n  missing distinct     Info     Mean      Gmd
##      101        0      101        1       51       34
##      .05      .10      .25      .50      .75      .90
##        6       11       26       51       76       91
##      .95
##       96
##
## lowest :   1   2   3   4   5, highest:  97  98  99 100 101
## ---------------------------------------------------------
## ID
##        n  missing distinct     Info     Mean      Gmd
##      101        0       97        1    91.41    59.56
##      .05      .10      .25      .50      .75      .90
##       10       18       49       87      143      159
##      .95
##      164
##
## lowest :   1   2   3   6   8, highest: 165 166 167 168 169
## ---------------------------------------------------------
## AGE
##        n  missing distinct     Info     Mean      Gmd
##      101        0       33    0.998     70.1     8.73
##      .05      .10      .25      .50      .75      .90
##       55       60       65       71       75       79
##      .95
##       81
##
## lowest : 44 49 52 53 54, highest: 80 81 82 83 88
## ---------------------------------------------------------
## PackHistory
##        n  missing distinct     Info     Mean      Gmd
##      101        0       48    0.998     39.7    27.35
##      .05      .10      .25      .50      .75      .90
##        6       10       20       36       54       75
##      .95
##       90
##
## lowest :   1   3   5   6   8, highest:  90 100 103 105 109
## ---------------------------------------------------------
## COPDSEVERITY
```

```
##        n  missing distinct
##      101        0        4
##
## Value             MILD    MODERATE      SEVERE VERY SEVERE
## Frequency           23          43          27           8
## Proportion       0.228       0.426       0.267       0.079
## ------------------------------------------------------------
## MWT1
##        n  missing distinct     Info     Mean      Gmd
##       99        2       69        1    385.9    117.6
##      .05      .10      .25      .50      .75      .90
##    212.7    226.0    300.0    419.0    460.5    495.2
##      .95
##    510.1
##
## lowest : 120 165 201 204 210, highest: 511 522 558 576 688
## ------------------------------------------------------------
## MWT2
##        n  missing distinct     Info     Mean      Gmd
##      100        1       72        1    390.3    121.7
##      .05      .10      .25      .50      .75      .90
##    210.0    237.0    303.8    399.0    459.0    518.7
##      .95
##    541.1
##
## lowest : 120 176 180 210 230, highest: 563 575 577 582 699
## ------------------------------------------------------------
## MWT1Best
##        n  missing distinct     Info     Mean      Gmd
##      100        1       71        1    399.1    119.7
##      .05      .10      .25      .50      .75      .90
##    215.7    240.0    303.8    420.0    465.2    518.7
##      .95
##    540.9
##
## lowest : 120 176 201 210 216, highest: 558 575 577 582 699
## ------------------------------------------------------------
## FEV1
##        n  missing distinct     Info     Mean      Gmd
##      101        0       85        1    1.604   0.7645
##      .05      .10      .25      .50      .75      .90
##     0.68     0.73     1.10     1.60     1.96     2.70
##      .95
##     2.90
##
## lowest : 0.45 0.47 0.51 0.6  0.65, highest: 2.93 2.97 3.02 3.06 3.18
## ------------------------------------------------------------
## FEV1PRED
##        n  missing distinct     Info     Mean      Gmd
##      101        0       51    0.999    58.53    25.56
##      .05      .10      .25      .50      .75      .90
##       24       30       42       60       75       90
##      .95
##       93
```

```
##
## lowest : 3.29 3.39 14   17   24  , highest: 92   93   95   98   102
## ---------------------------------------------------------------
## FVC
##        n  missing distinct     Info     Mean      Gmd
##      101        0       80        1    2.955    1.108
##      .05      .10      .25      .50      .75      .90
##     1.56     1.89     2.27     2.77     3.63     4.39
##      .95
##     4.70
##
## lowest : 1.14 1.31 1.47 1.52 1.56, highest: 4.72 4.9  5.15 5.23 5.37
## ---------------------------------------------------------------
## FVCPRED
##        n  missing distinct     Info     Mean      Gmd
##      101        0       57    0.999    86.44    24.92
##      .05      .10      .25      .50      .75      .90
##       53       60       71       84      103      118
##      .95
##      122
##
## lowest :  27  45  48  51  53, highest: 121 122 123 125 132
## ---------------------------------------------------------------
## CAT
##        n  missing distinct     Info     Mean      Gmd
##      101        0       30    0.997    19.34    12.28
##      .05      .10      .25      .50      .75      .90
##        5        5       12       18       24       29
##      .95
##       30
##
## lowest :   3   4   5   6   7, highest:  29  30  31  32 188
## ---------------------------------------------------------------
## HAD
##        n  missing distinct     Info     Mean      Gmd
##      101        0       28    0.997    11.18    8.984
##      .05      .10      .25      .50      .75      .90
##        1        2        6       10       15       22
##      .95
##       26
##
## lowest : 0    1    2    3    4   , highest: 23   26   29   30   56.2
## ---------------------------------------------------------------
## SGRQ
##        n  missing distinct     Info     Mean      Gmd
##      101        0       89        1    40.19    20.88
##      .05      .10      .25      .50      .75      .90
##    10.92    16.29    28.41    38.21    55.23    67.56
##      .95
##    72.24
##
## lowest : 2     8.12  8.25  10.01 10.92
## highest: 72.56 73.82 75.56 76.5  77.44
## ---------------------------------------------------------------
```

```
## AGEquartiles
##        n  missing distinct
##      101        0        4
##
## Value          1     2     3     4
## Frequency     26    24    28    23
## Proportion 0.257 0.238 0.277 0.228
## ------------------------------------------------------------
## copd
##        n  missing distinct
##      101        0        4
##
## Value          1     2     3     4
## Frequency     23    43    27     8
## Proportion 0.228 0.426 0.267 0.079
## ------------------------------------------------------------
## gender
##        n  missing distinct
##      101        0        2
##
## Value          0     1
## Frequency     36    65
## Proportion 0.356 0.644
## ------------------------------------------------------------
## smoking
##        n  missing distinct
##      101        0        2
##
## Value          1     2
## Frequency     16    85
## Proportion 0.158 0.842
## ------------------------------------------------------------
## Diabetes
##        n  missing distinct
##      101        0        2
##
## Value          0     1
## Frequency     80    21
## Proportion 0.792 0.208
## ------------------------------------------------------------
## muscular
##        n  missing distinct
##      101        0        2
##
## Value          0     1
## Frequency     82    19
## Proportion 0.812 0.188
## ------------------------------------------------------------
## hypertension
##        n  missing distinct
##      101        0        2
##
## Value          0     1
## Frequency     89    12
```

```
## Proportion 0.881 0.119
## ---------------------------------------------------------
## AtrialFib
##        n  missing distinct
##      101        0        2
##
## Value          0     1
## Frequency     81    20
## Proportion 0.802 0.198
## ---------------------------------------------------------
## IHD
##        n  missing distinct
##      101        0        2
##
## Value          0     1
## Frequency     92     9
## Proportion 0.911 0.089
## ---------------------------------------------------------
```

There is no missing value found in the data

Create variable comorbid

```
comorbid <- length(copd$Diabetes) #create a variable with length similar with Diabetes variable
comorbid[copd$Diabetes ==1 | copd$muscular == 1 | copd$hypertension ==1 | copd$AtrialFib ==1 | copd$IHD
comorbid[is.na(comorbid)] <- 0
comorbid <- factor(comorbid)
```

```
copd$comorbid <- comorbid
```

```
str(copd)
```

```
## 'data.frame':    101 obs. of  25 variables:
##  $ X           : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ ID          : int  58 57 62 145 136 84 93 27 114 152 ...
##  $ AGE         : num  77 79 80 56 65 67 67 83 72 75 ...
##  $ PackHistory : num  60 50 11 60 68 26 50 90 50 6 ...
##  $ COPDSEVERITY: chr  "SEVERE" "MODERATE" "MODERATE" "VERY SEVERE" ...
##  $ MWT1        : num  120 165 201 210 204 216 214 214 231 226 ...
##  $ MWT2        : num  120 176 180 210 210 180 237 237 237 240 ...
##  $ MWT1Best    : num  120 176 201 210 210 216 237 237 237 240 ...
##  $ FEV1        : num  1.21 1.09 1.52 0.47 1.07 1.09 0.69 0.68 2.13 1.06 ...
##  $ FEV1PRED    : num  36 56 68 14 42 50 35 32 63 46 ...
##  $ FVC         : num  2.4 1.64 2.3 1.14 2.91 1.99 1.31 2.23 4.38 2.06 ...
##  $ FVCPRED     : num  98 65 86 27 98 60 48 77 80 75 ...
##  $ CAT         : num  25 12 22 28 32 29 29 22 25 31 ...
##  $ HAD         : num  8 21 18 26 18 21 30 2 6 20 ...
##  $ SGRQ        : num  69.5 44.2 44.1 62 75.6 ...
##  $ AGEquartiles: Factor w/ 4 levels "1","2","3","4": 4 4 4 1 1 2 2 4 3 3 ...
##  $ copd        : Factor w/ 4 levels "1","2","3","4": 3 2 2 4 3 2 3 3 2 3 ...
##  $ gender      : Factor w/ 2 levels "0","1": 2 1 1 2 2 1 1 2 2 1 ...
##  $ smoking     : Factor w/ 2 levels "1","2": 2 2 2 2 2 1 1 2 1 2 ...
##  $ Diabetes    : Factor w/ 2 levels "0","1": 2 2 2 1 1 2 2 2 2 1 ...
```

```
##  $ muscular    : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 1 1 1 2 ...
##  $ hypertension: Factor w/ 2 levels "0","1": 1 1 1 2 2 1 1 1 1 1 ...
##  $ AtrialFib   : Factor w/ 2 levels "0","1": 2 2 2 2 1 2 2 2 2 1 ...
##  $ IHD         : Factor w/ 2 levels "0","1": 1 2 1 1 1 1 1 1 1 1 ...
##  $ comorbid    : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
```

Check categorical variables using crosstable

```
# Assuming 'cat_vars' contains the names of categorical variables
cat_vars <- c("gender", "COPDSEVERITY", "copd", "smoking", "Diabetes", "muscular", "hypertension", "Atri

# Create repeated CrossTable for each categorical variable
for(var in cat_vars) {
  cat(sprintf("Variable: %s\n", var))
  cat_table <- CrossTable(copd[, var], prop.chisq = FALSE)
  print(cat_table)
}
```

```
## Variable: gender
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  101
##
##
##           |         0 |         1 |
##           |-----------|-----------|
##           |        36 |        65 |
##           |     0.356 |     0.644 |
##           |-----------|-----------|
##
##
##
##
## $t
##       0  1
## [1,] 36 65
##
## $prop.row
##             0        1
## [1,] 0.356436 0.643564
##
## $prop.col
##      0 1
## [1,] 1 1
##
## $prop.tbl
##             0        1
```

```
## [1,] 0.356436 0.643564
##
## Variable: COPDSEVERITY
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  101
##
##
##              |        MILD |    MODERATE |      SEVERE | VERY SEVERE |
##              |-------------|-------------|-------------|-------------|
##              |          23 |          43 |          27 |           8 |
##              |       0.228 |       0.426 |       0.267 |       0.079 |
##              |-------------|-------------|-------------|-------------|
##
##
##
##
## $t
##      MILD MODERATE SEVERE VERY SEVERE
## [1,]   23       43     27           8
##
## $prop.row
##          MILD MODERATE   SEVERE VERY SEVERE
## [1,] 0.227723 0.425743 0.267327   0.0792079
##
## $prop.col
##      MILD MODERATE SEVERE VERY SEVERE
## [1,]    1        1      1           1
##
## $prop.tbl
##          MILD MODERATE   SEVERE VERY SEVERE
## [1,] 0.227723 0.425743 0.267327   0.0792079
##
## Variable: copd
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  101
##
##
##           |         1 |         2 |         3 |         4 |
```

```
##           |-----------|-----------|-----------|-----------|
##           |        23 |        43 |        27 |         8 |
##           |     0.228 |     0.426 |     0.267 |     0.079 |
##           |-----------|-----------|-----------|-----------|
##
##
##
##
## $t
##       1  2  3 4
## [1,] 23 43 27 8
##
## $prop.row
##             1        2        3         4
## [1,] 0.227723 0.425743 0.267327 0.0792079
##
## $prop.col
##      1 2 3 4
## [1,] 1 1 1 1
##
## $prop.tbl
##             1        2        3         4
## [1,] 0.227723 0.425743 0.267327 0.0792079
##
## Variable: smoking
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  101
##
##
##           |         1 |         2 |
##           |-----------|-----------|
##           |        16 |        85 |
##           |     0.158 |     0.842 |
##           |-----------|-----------|
##
##
##
##
## $t
##       1  2
## [1,] 16 85
##
## $prop.row
##             1        2
## [1,] 0.158416 0.841584
##
```

```
## $prop.col
##      1 2
## [1,] 1 1
##
## $prop.tbl
##             1        2
## [1,] 0.158416 0.841584
##
## Variable: Diabetes
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  101
##
##
##            |         0 |         1 |
##            |-----------|-----------|
##            |        80 |        21 |
##            |     0.792 |     0.208 |
##            |-----------|-----------|
##
##
##
##
## $t
##       0  1
## [1,] 80 21
##
## $prop.row
##             0        1
## [1,] 0.792079 0.207921
##
## $prop.col
##      0 1
## [1,] 1 1
##
## $prop.tbl
##             0        1
## [1,] 0.792079 0.207921
##
## Variable: muscular
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
```

```
##
##
## Total Observations in Table:  101
##
##
##           |         0 |         1 |
##           |-----------|-----------|
##           |        82 |        19 |
##           |     0.812 |     0.188 |
##           |-----------|-----------|
##
##
##
##
## $t
##       0  1
## [1,] 82 19
##
## $prop.row
##             0        1
## [1,] 0.811881 0.188119
##
## $prop.col
##      0 1
## [1,] 1 1
##
## $prop.tbl
##             0        1
## [1,] 0.811881 0.188119
##
## Variable: hypertension
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  101
##
##
##           |         0 |         1 |
##           |-----------|-----------|
##           |        89 |        12 |
##           |     0.881 |     0.119 |
##           |-----------|-----------|
##
##
##
##
## $t
##       0  1
```

```
## [1,] 89 12
##
## $prop.row
##             0        1
## [1,] 0.881188 0.118812
##
## $prop.col
##      0 1
## [1,] 1 1
##
## $prop.tbl
##             0        1
## [1,] 0.881188 0.118812
##
## Variable: AtrialFib
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  101
##
##
##            |         0 |         1 |
##            |-----------|-----------|
##            |        81 |        20 |
##            |     0.802 |     0.198 |
##            |-----------|-----------|
##
##
##
##
## $t
##       0  1
## [1,] 81 20
##
## $prop.row
##            0       1
## [1,] 0.80198 0.19802
##
## $prop.col
##      0 1
## [1,] 1 1
##
## $prop.tbl
##            0       1
## [1,] 0.80198 0.19802
##
## Variable: IHD
##
```

```
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  101
##
##
##            |         0 |         1 |
##            |-----------|-----------|
##            |        92 |         9 |
##            |     0.911 |     0.089 |
##            |-----------|-----------|
##
##
##
##
## $t
##       0 1
## [1,] 92 9
##
## $prop.row
##             0         1
## [1,] 0.910891 0.0891089
##
## $prop.col
##      0 1
## [1,] 1 1
##
## $prop.tbl
##             0         1
## [1,] 0.910891 0.0891089
##
## Variable: comorbid
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  101
##
##
##            |         0 |         1 |
##            |-----------|-----------|
##            |        46 |        55 |
##            |     0.455 |     0.545 |
##            |-----------|-----------|
```

```
##
##
##
##
## $t
##       0  1
## [1,] 46 55
##
## $prop.row
##             0        1
## [1,] 0.455446 0.544554
##
## $prop.col
##      0 1
## [1,] 1 1
##
## $prop.tbl
##             0        1
## [1,] 0.455446 0.544554
```

Summary and histogram for numerical data

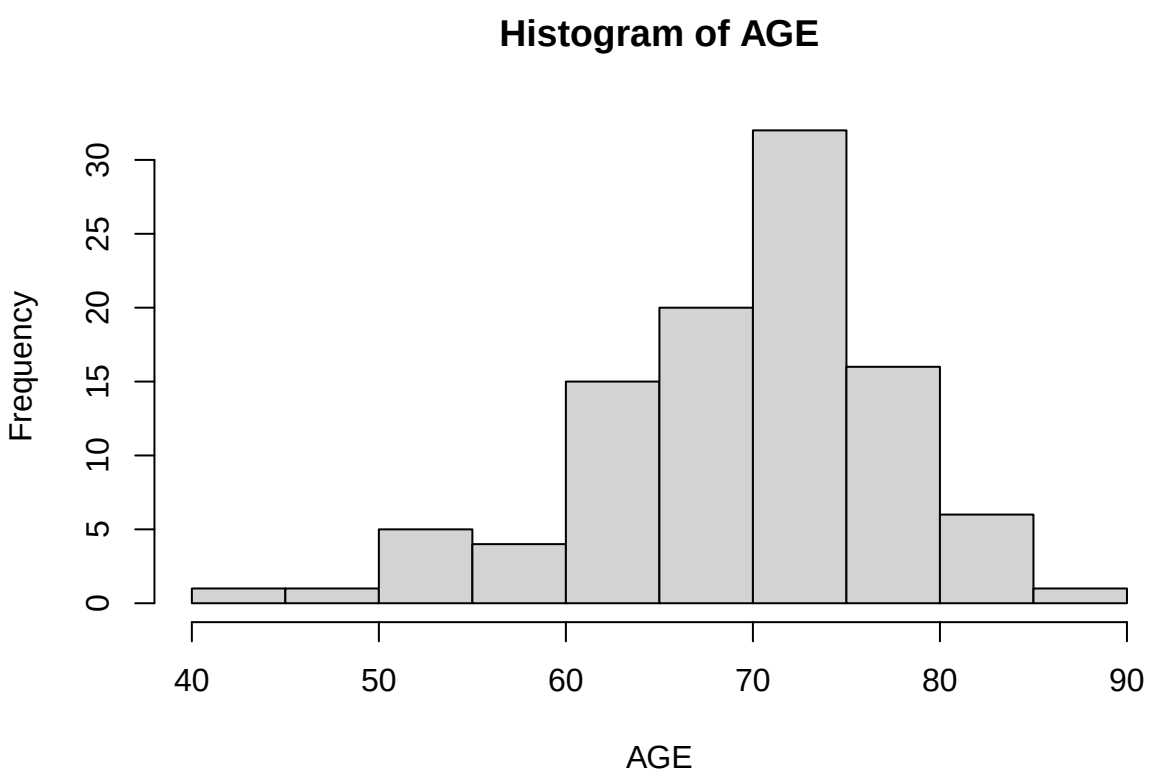
```
num_var <- c("AGE", "PackHistory", "MWT1", "MWT2", "MWT1Best", "FEV1", "FEV1PRED", "FVC", "FVCPRED","CA
```

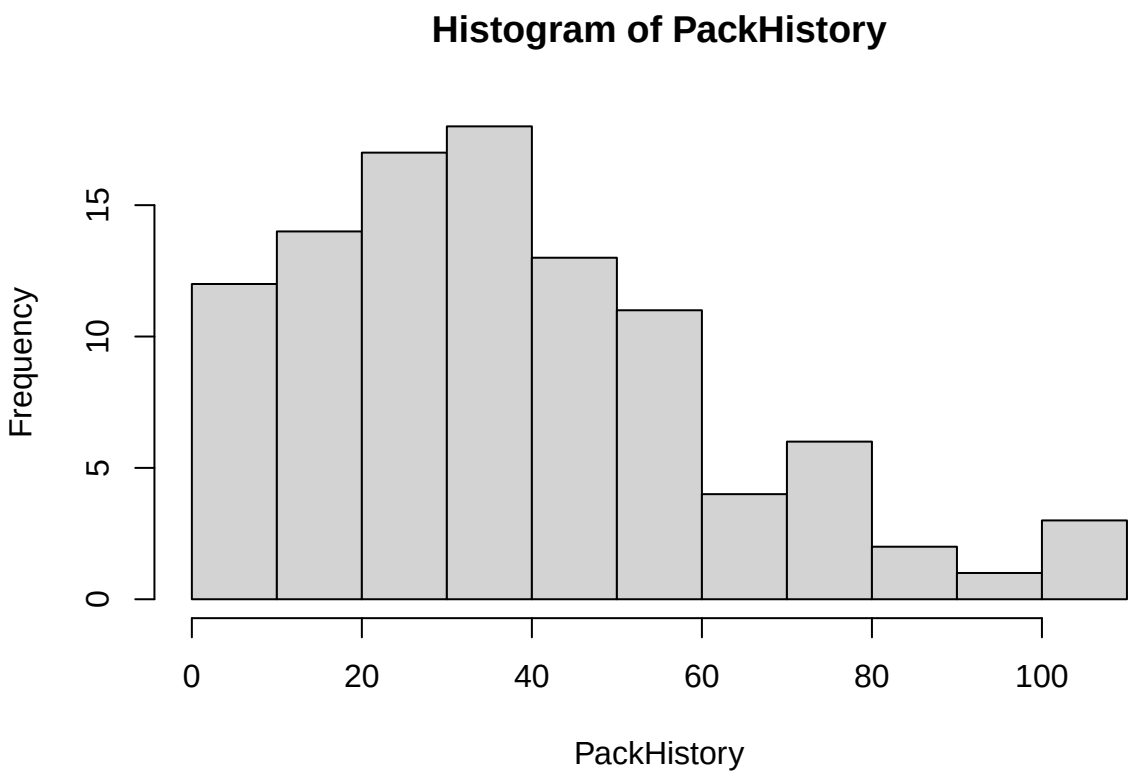
```
summary(copd)
```

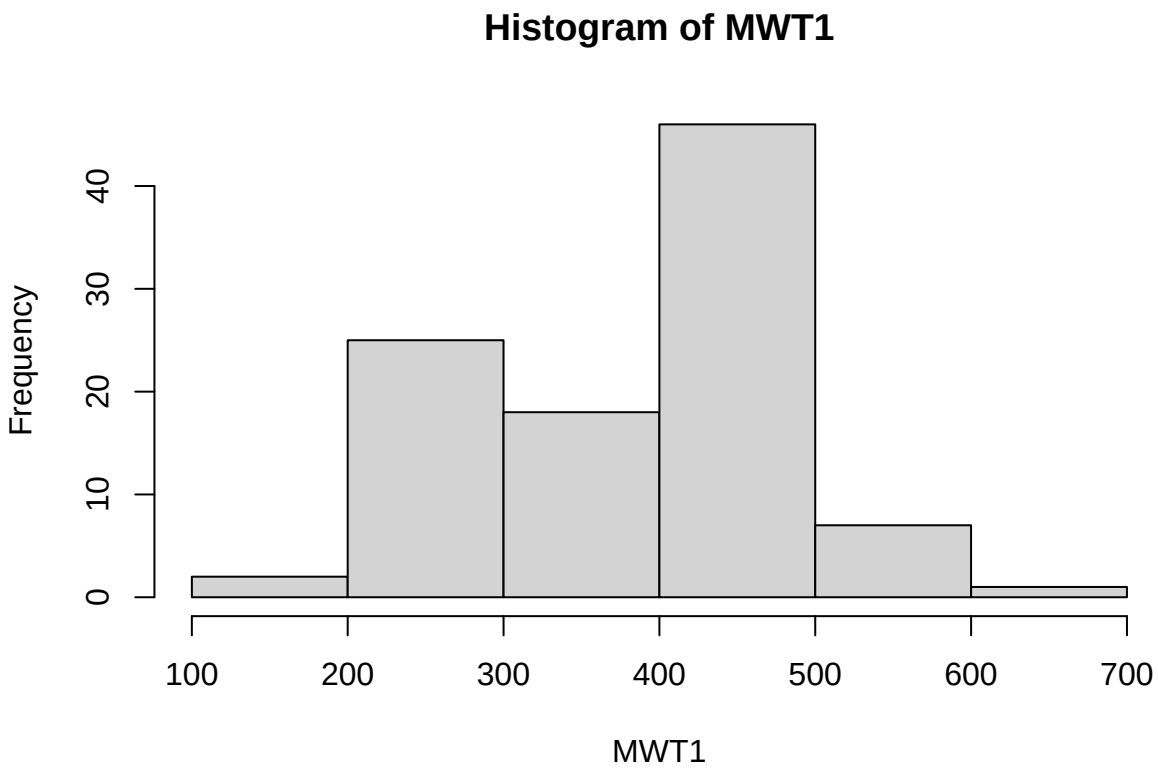
```
##        X              ID              AGE
##  Min.   :  1   Min.   :  1.0   Min.   :44.0
##  1st Qu.: 26   1st Qu.: 49.0   1st Qu.:65.0
##  Median : 51   Median : 87.0   Median :71.0
##  Mean   : 51   Mean   : 91.4   Mean   :70.1
##  3rd Qu.: 76   3rd Qu.:143.0   3rd Qu.:75.0
##  Max.   :101   Max.   :169.0   Max.   :88.0
##
##   PackHistory    COPDSEVERITY            MWT1
##  Min.   :  1.0   Length:101         Min.   :120
##  1st Qu.: 20.0   Class :character   1st Qu.:300
##  Median : 36.0   Mode  :character   Median :419
##  Mean   : 39.7                      Mean   :386
##  3rd Qu.: 54.0                      3rd Qu.:460
##  Max.   :109.0                      Max.   :688
##                                     NA's   :2
##       MWT2        MWT1Best        FEV1
##  Min.   :120   Min.   :120   Min.   :0.45
##  1st Qu.:304   1st Qu.:304   1st Qu.:1.10
##  Median :399   Median :420   Median :1.60
##  Mean   :390   Mean   :399   Mean   :1.60
##  3rd Qu.:459   3rd Qu.:465   3rd Qu.:1.96
##  Max.   :699   Max.   :699   Max.   :3.18
##  NA's   :1     NA's   :1
##     FEV1PRED           FVC          FVCPRED
##  Min.   :  3.29   Min.   :1.14   Min.   : 27.0
##  1st Qu.: 42.00   1st Qu.:2.27   1st Qu.: 71.0
```

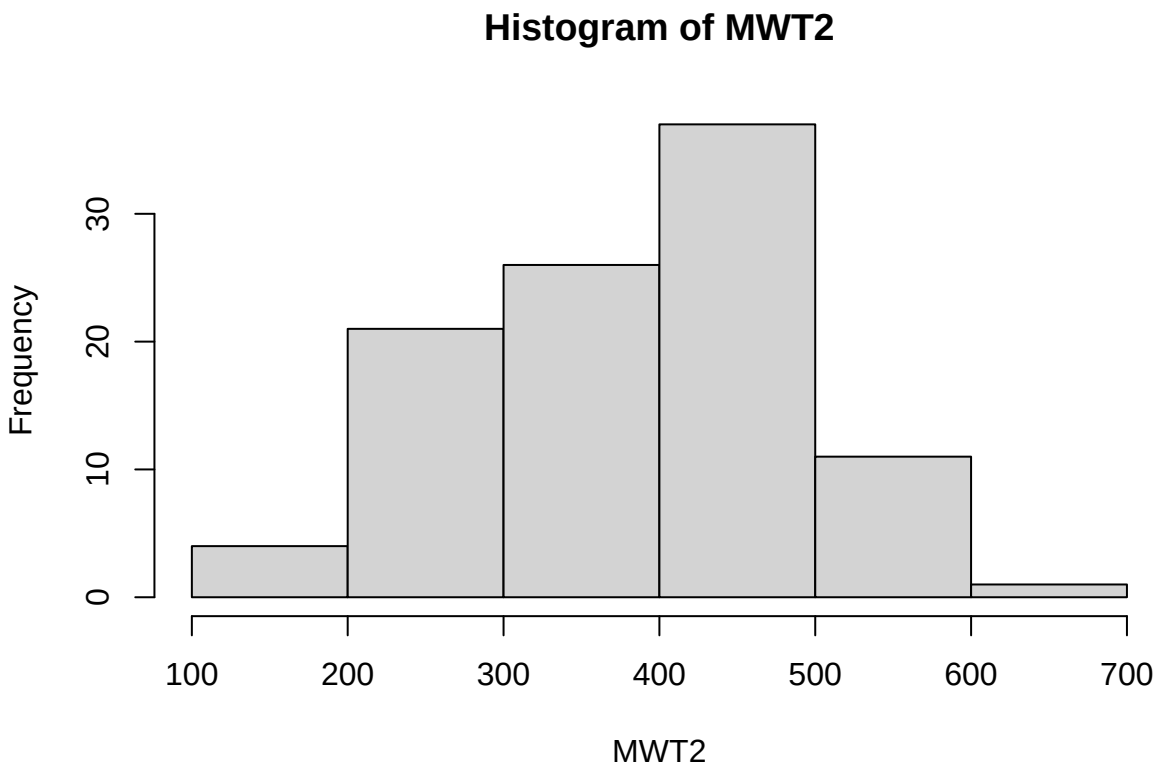
```
##  Median : 60.00   Median :2.77   Median : 84.0
##  Mean   : 58.53   Mean   :2.95   Mean   : 86.4
##  3rd Qu.: 75.00   3rd Qu.:3.63   3rd Qu.:103.0
##  Max.   :102.00   Max.   :5.37   Max.   :132.0
##
##       CAT             HAD            SGRQ      AGEquartiles
##  Min.   :  3.0   Min.   : 0.0   Min.   : 2.0   1:26
##  1st Qu.: 12.0   1st Qu.: 6.0   1st Qu.:28.4   2:24
##  Median : 18.0   Median :10.0   Median :38.2   3:28
##  Mean   : 19.3   Mean   :11.2   Mean   :40.2   4:23
##  3rd Qu.: 24.0   3rd Qu.:15.0   3rd Qu.:55.2
##  Max.   :188.0   Max.   :56.2   Max.   :77.4
##
##  copd   gender smoking Diabetes muscular hypertension
##  1:23   0:36   1:16    0:80     0:82     0:89
##  2:43   1:65   2:85    1:21     1:19     1:12
##  3:27
##  4: 8
##
##
##
##  AtrialFib IHD    comorbid
##  0:81      0:92   0:46
##  1:20      1: 9   1:55
##
##
##
##
##
```

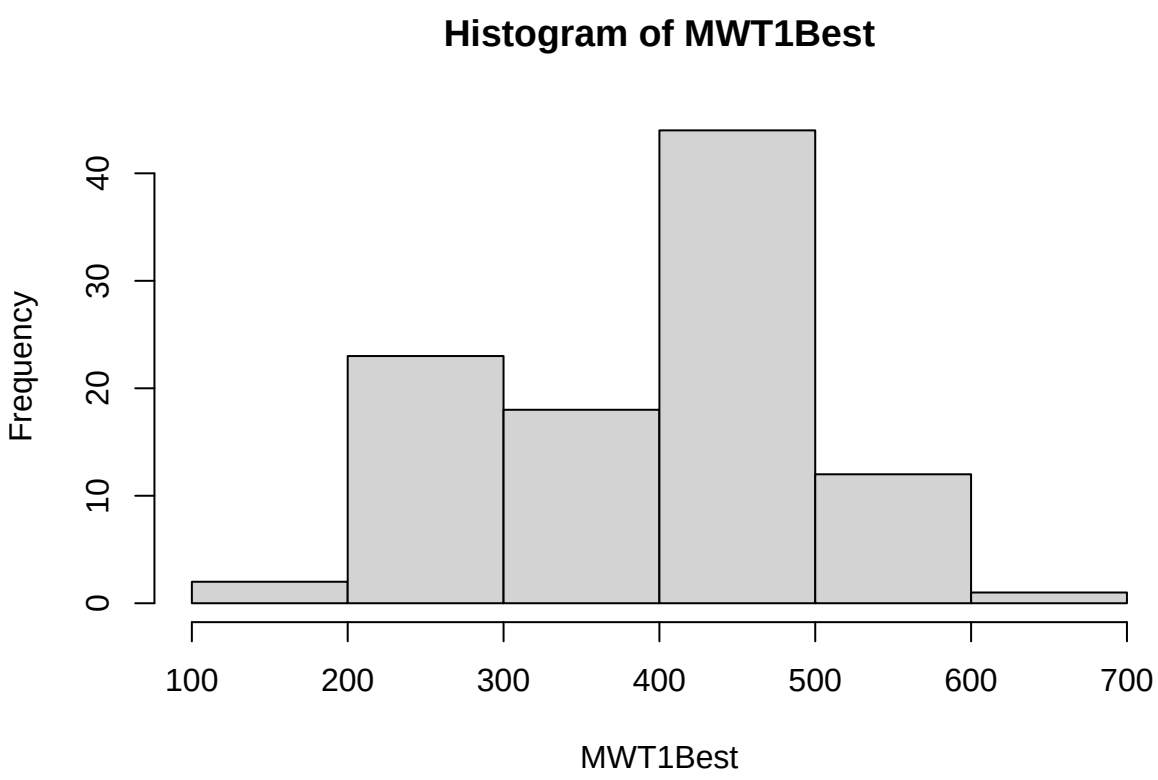
```r
# Create repeated bar plots for each categorical variable
for(var in num_var) {
  hist(copd[[var]], main = paste("Histogram of", var), xlab = var)
}
```

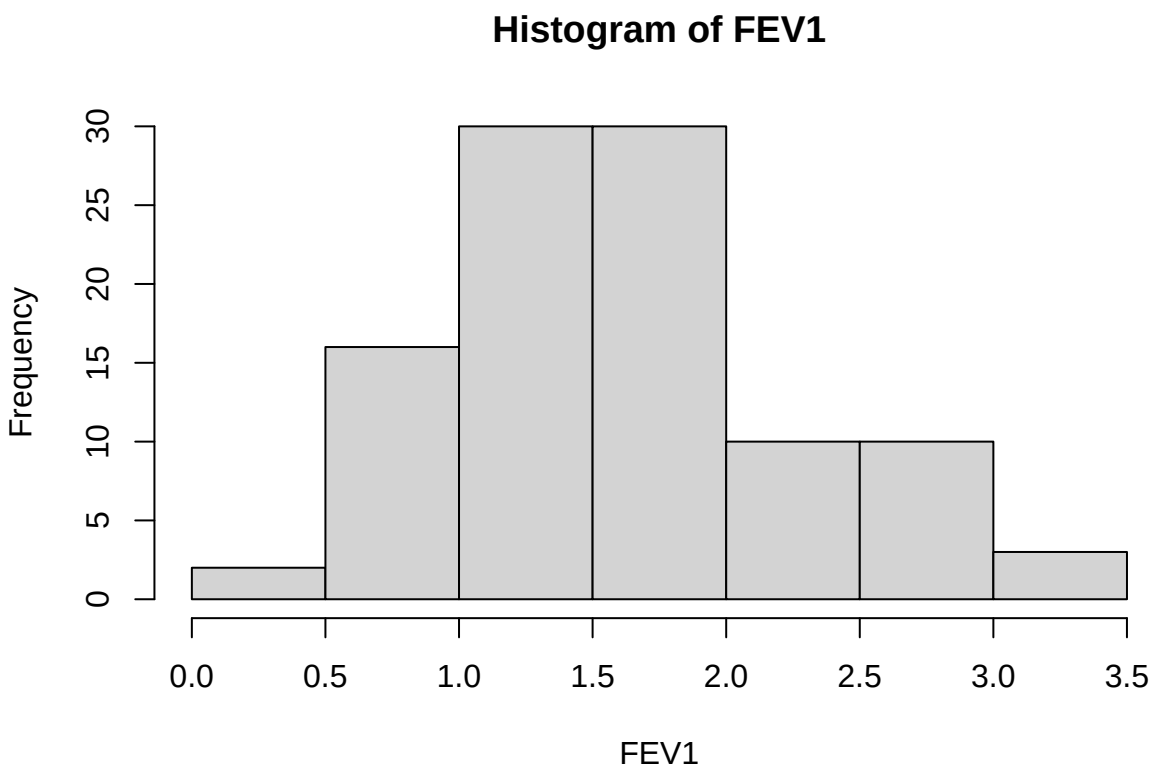

Histogram of AGE
Frequency
0
5
10
15
20
25
30
40
50
60
70
80
90
AGE

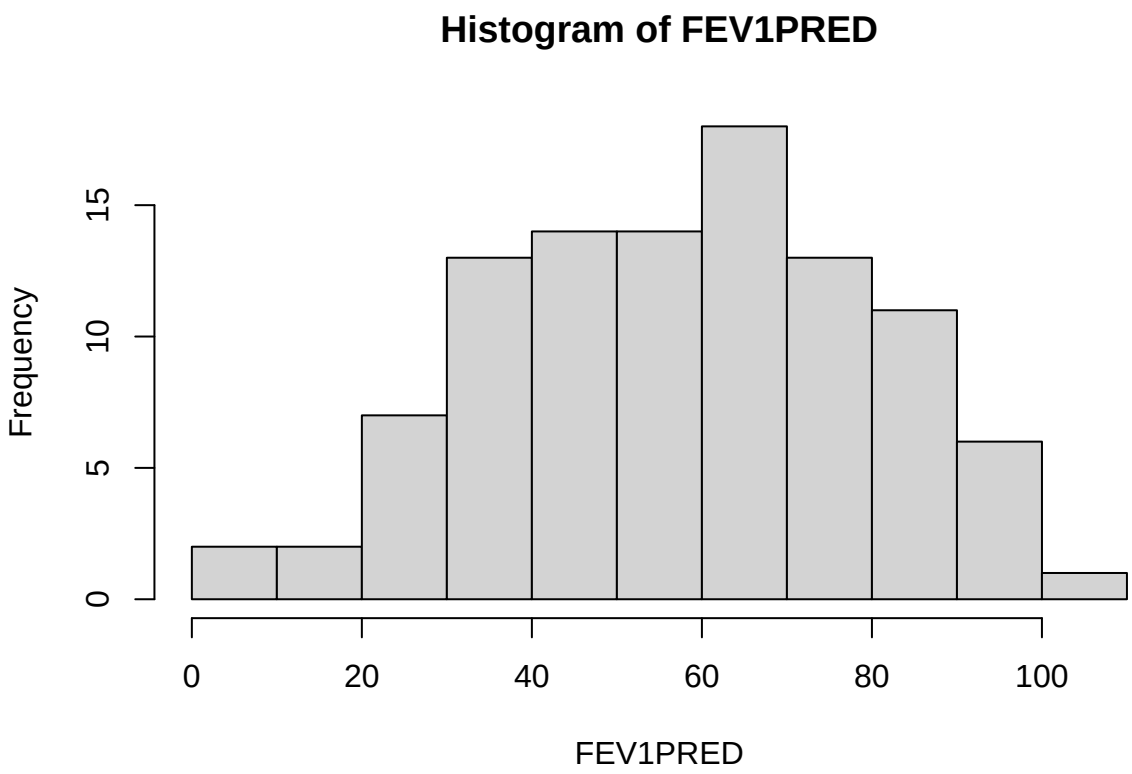

Histogram of PackHistory
Frequency
0
5
10
15
0
20
40
60
80
100
PackHistory

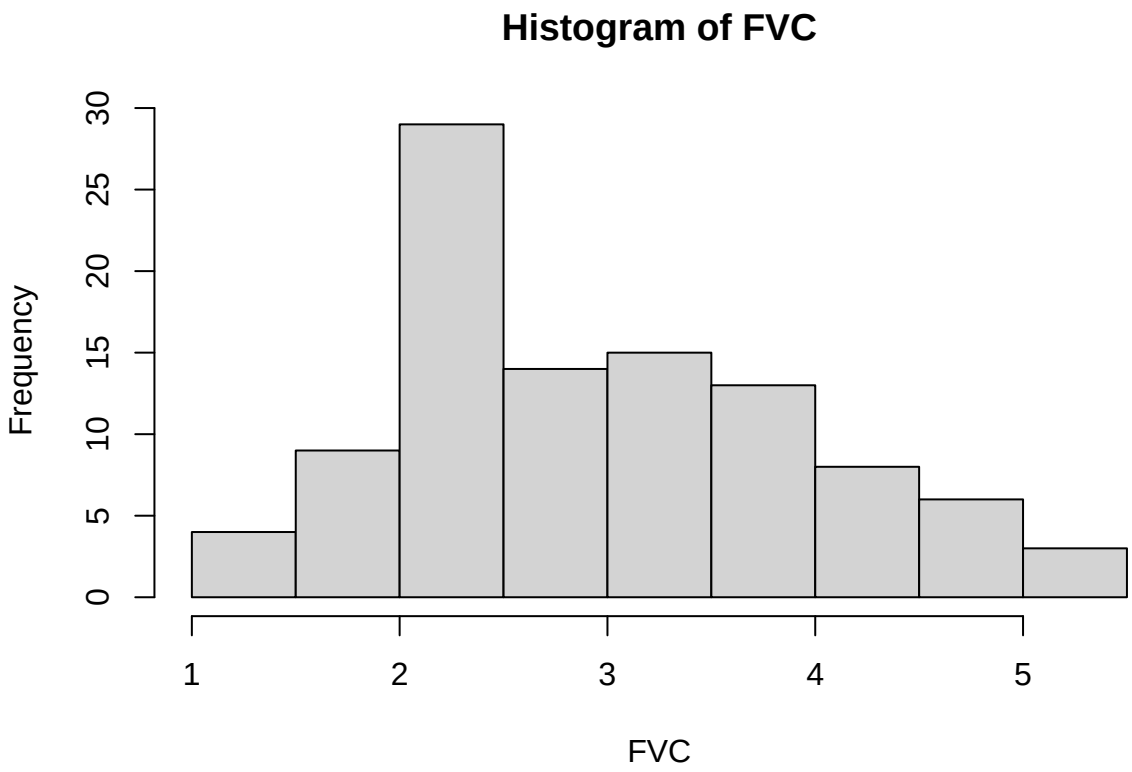

Histogram of MWT1
Frequency
0
10
20
30
40
100
200
300
400
500
600
700
MWT1

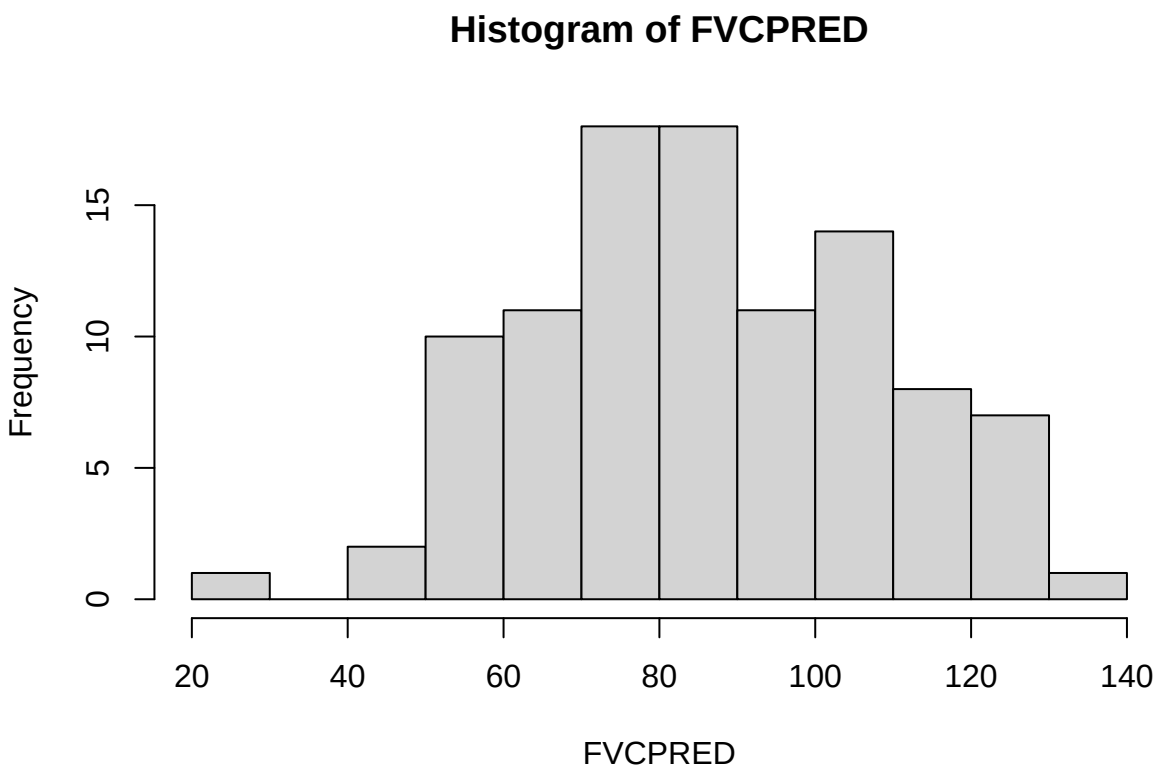

Histogram of MWT2
Frequency
0
10
20
30
100
200
300
400
500
600
700
MWT2

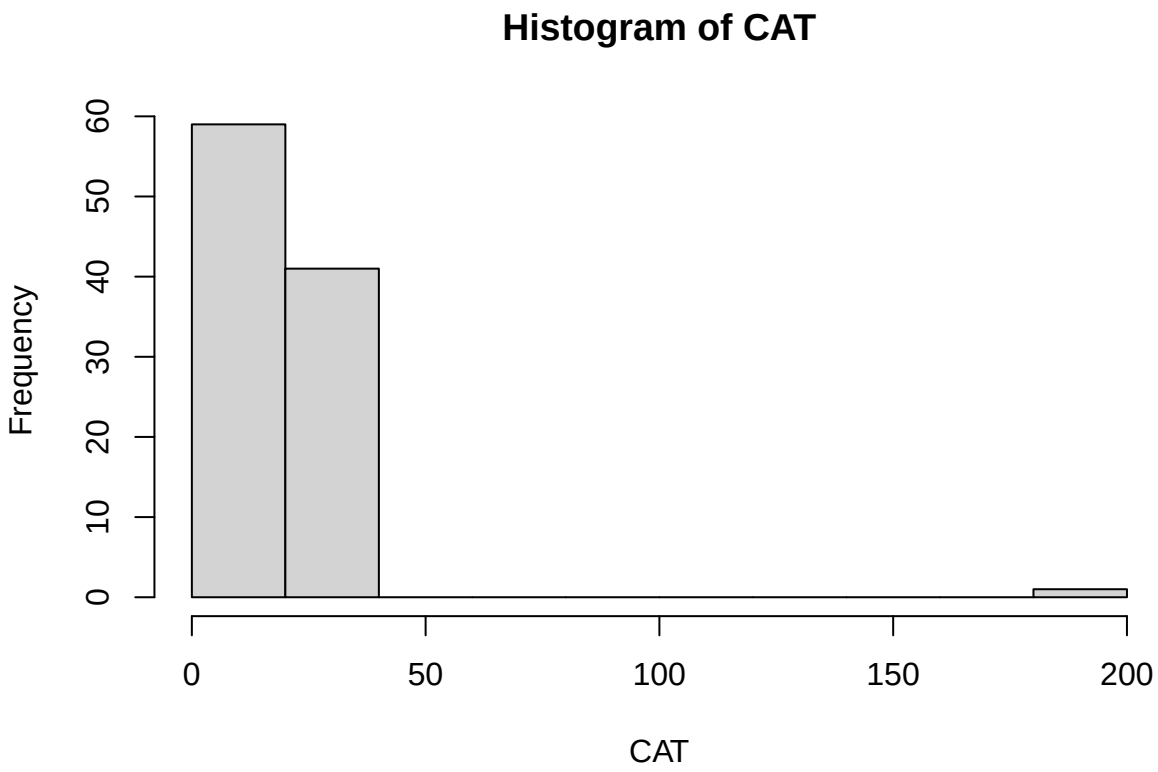

Histogram of MWT1Best
Frequency
0
10
20
30
40
100
200
300
400
500
600
700
MWT1Best

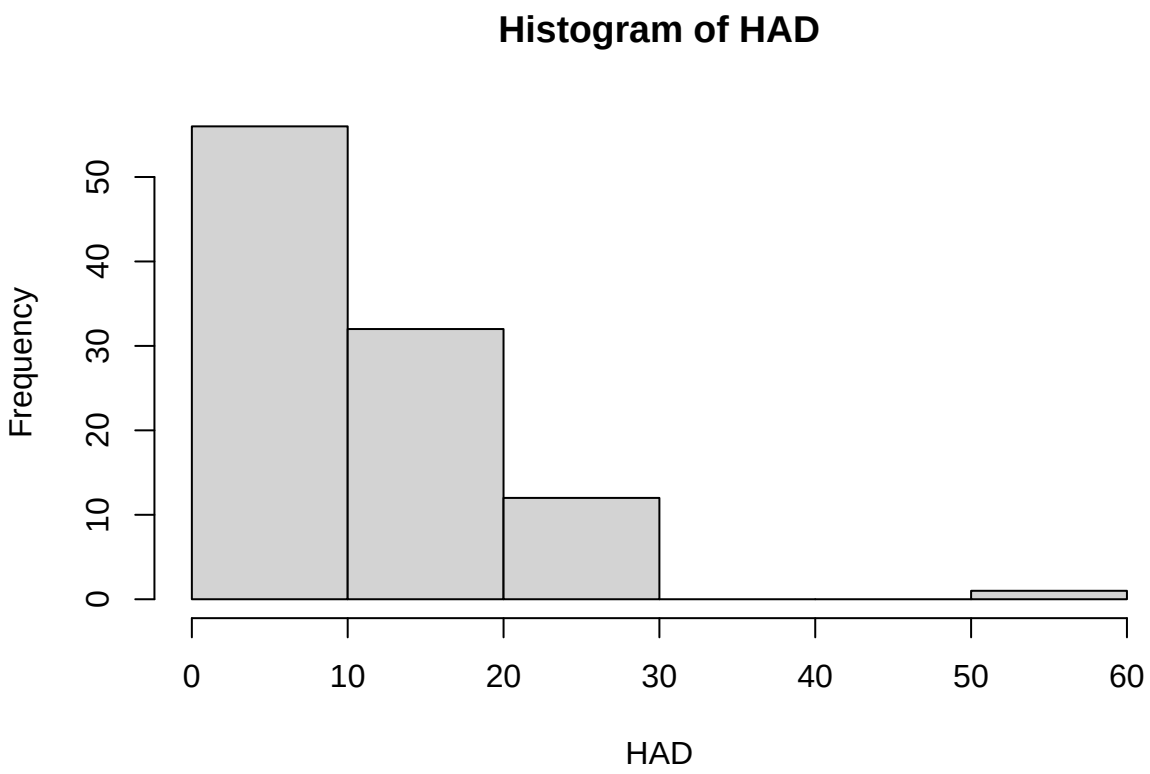

Histogram of FEV1
Frequency
0
5
10
15
20
25
30
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
FEV1

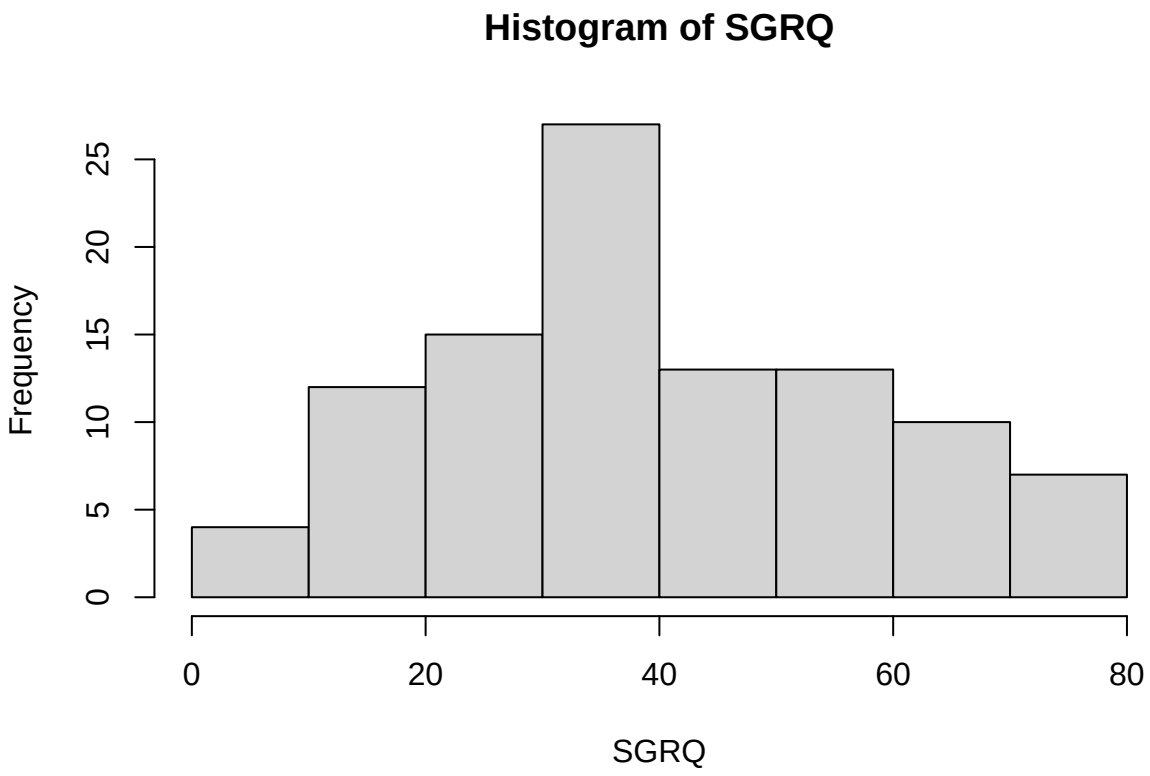

Histogram of FEV1PRED
Frequency
0
5
10
15
0
20
40
60
80
100
FEV1PRED


Histogram of FVC
Frequency
0
5
10
15
20
25
30
1
2
3
4
5
FVC

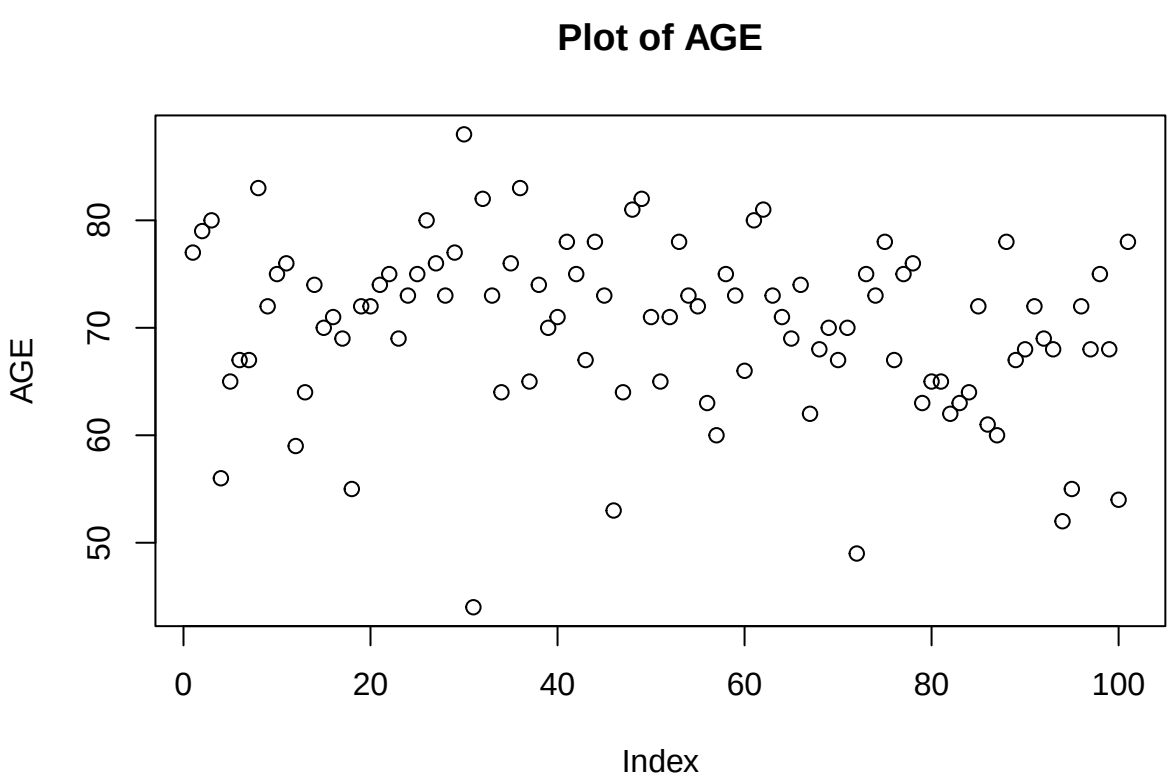

Histogram of FVCPRED
Frequency
0
5
10
15
20
40
60
80
100
120
140
FVCPRED

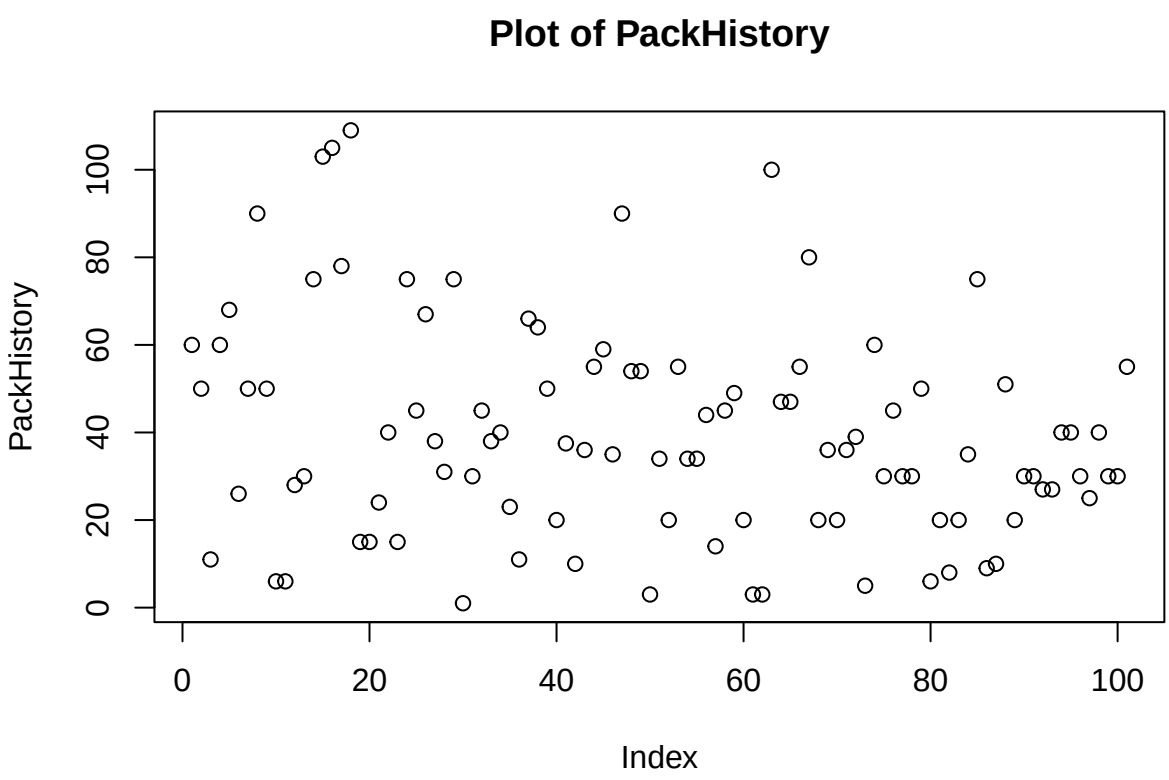

Histogram of CAT
Frequency
0
10
20
30
40
50
60
0
50
100
150
200
CAT

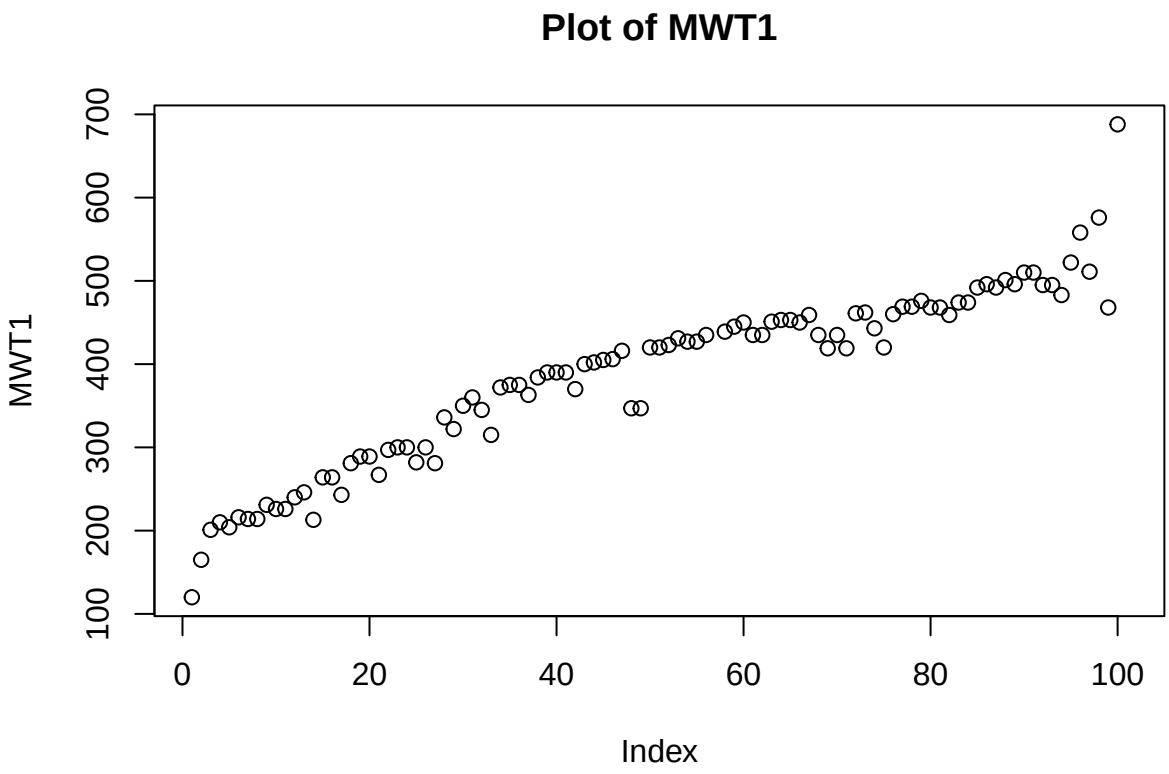

Histogram of HAD
Frequency
0
10
20
30
40
50
0
10
20
30
40
50
60
HAD

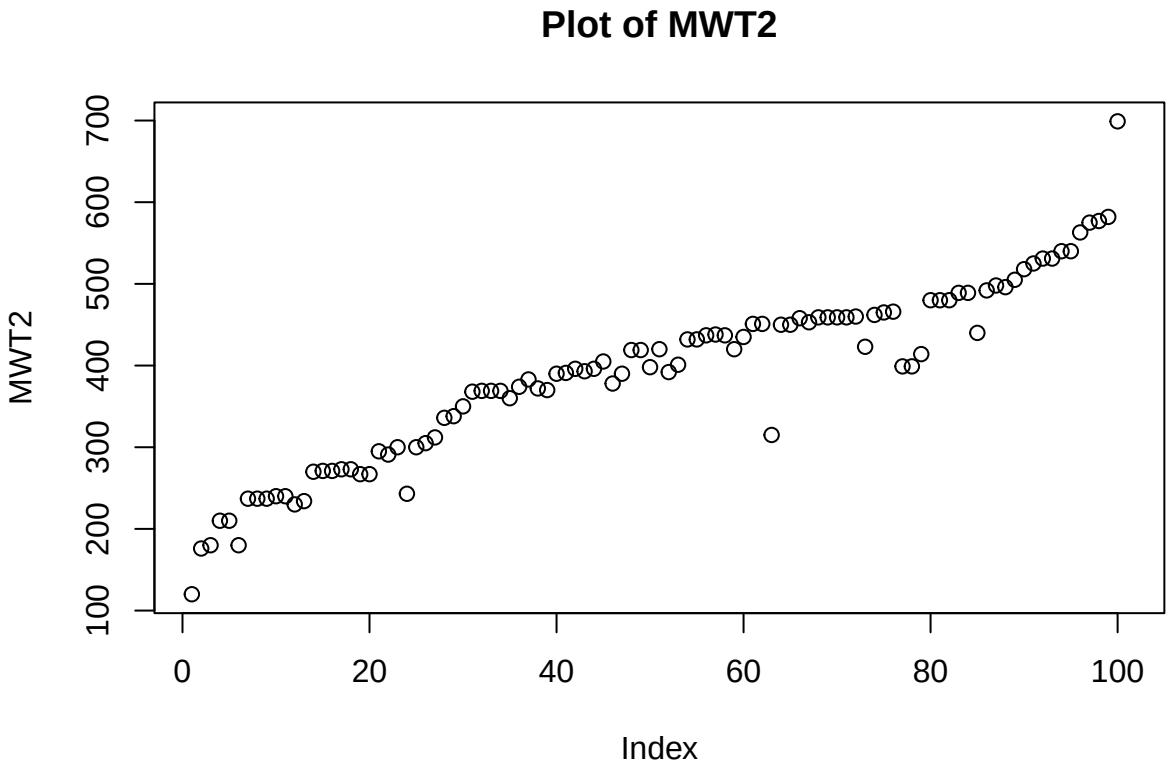


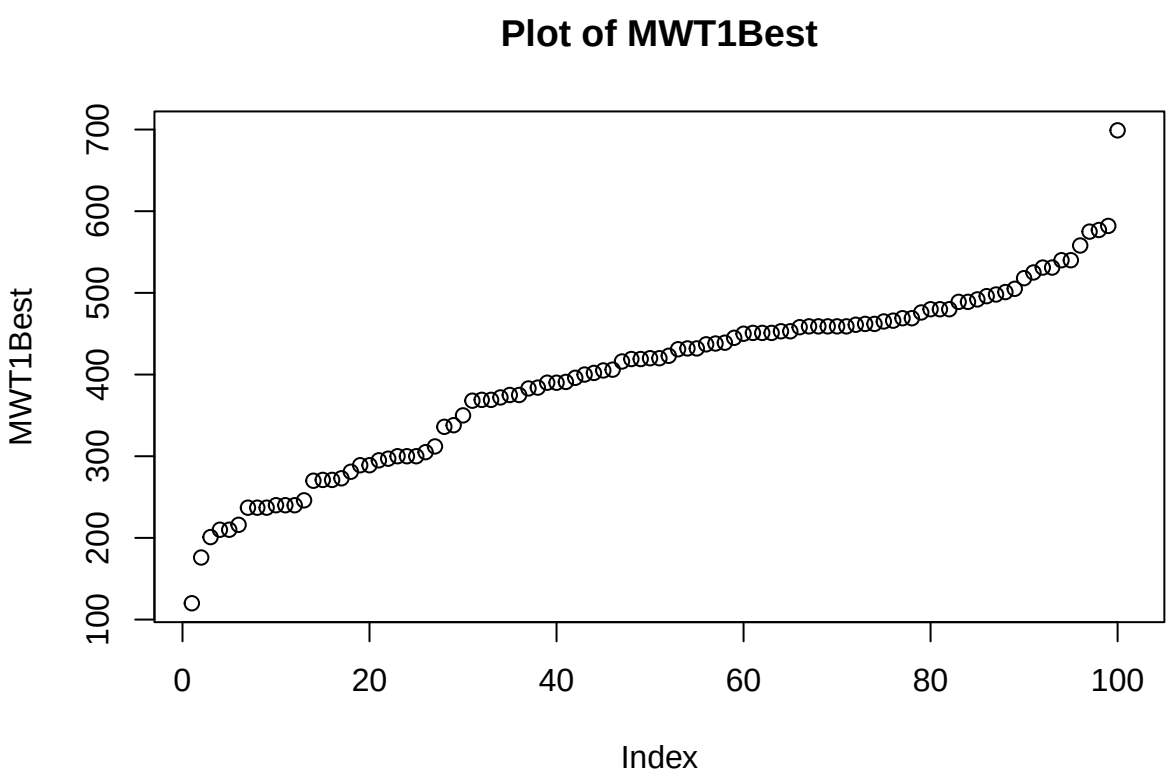

```
# Create repeated plots for each numerical variable
for(var in num_var) {
  plot(copd[[var]], main = paste("Plot of", var), xlab = "Index", ylab = var)
}
```

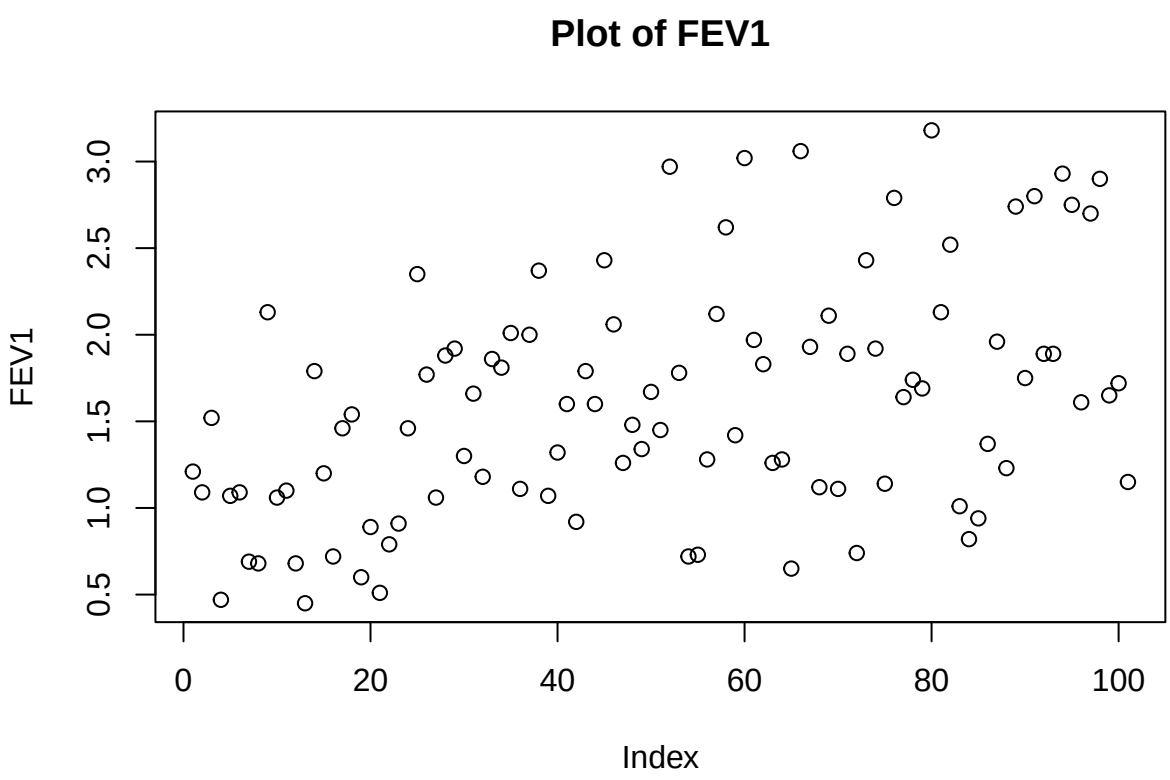

Plot of AGE
AGE
50
60
70
80
0
20
40
60
80
100
Index

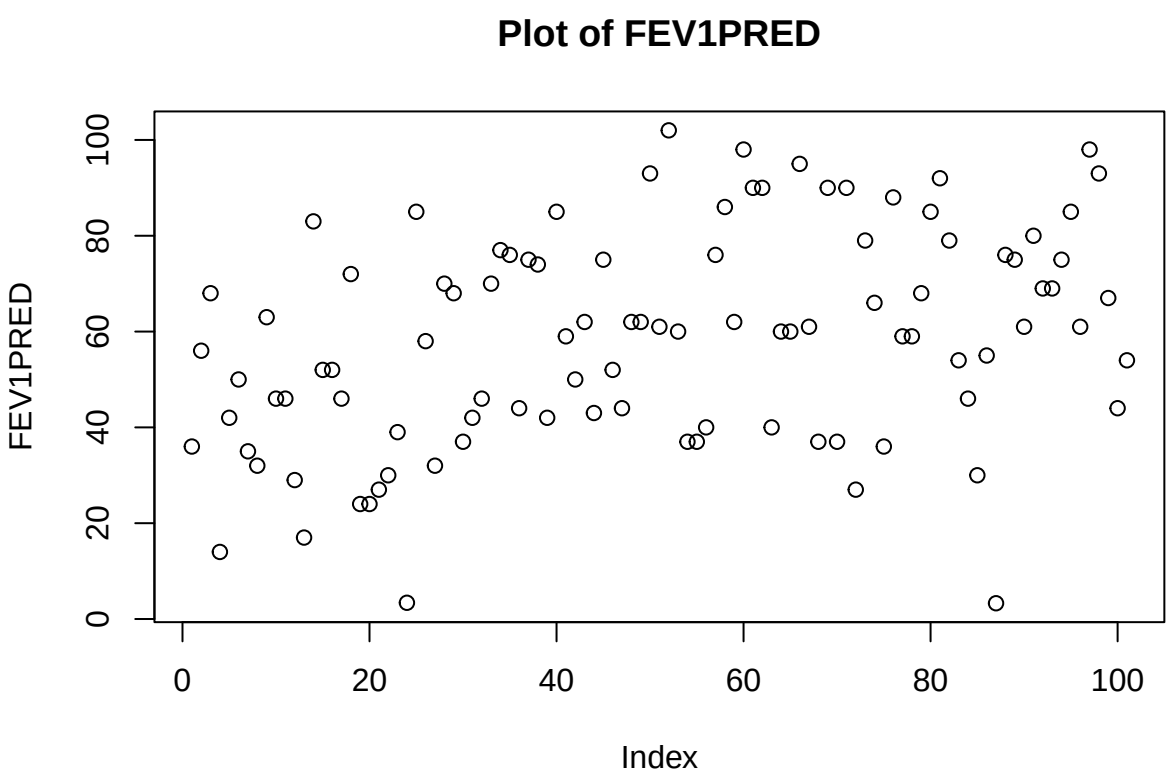

Plot of PackHistory
PackHistory
0
20
40
60
80
100
Index
0
20
40
60
80
100

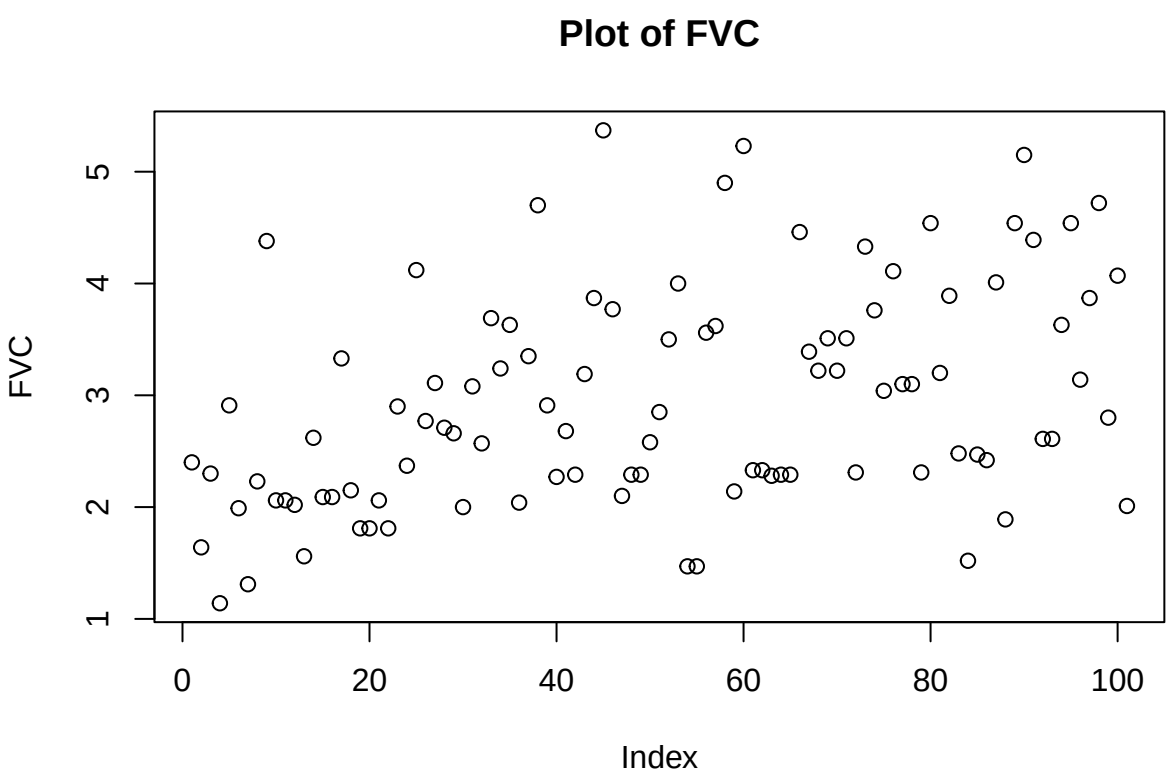

Plot of MWT1
MWT1
100
200
300
400
500
600
700
0
20
40
60
80
100
Index

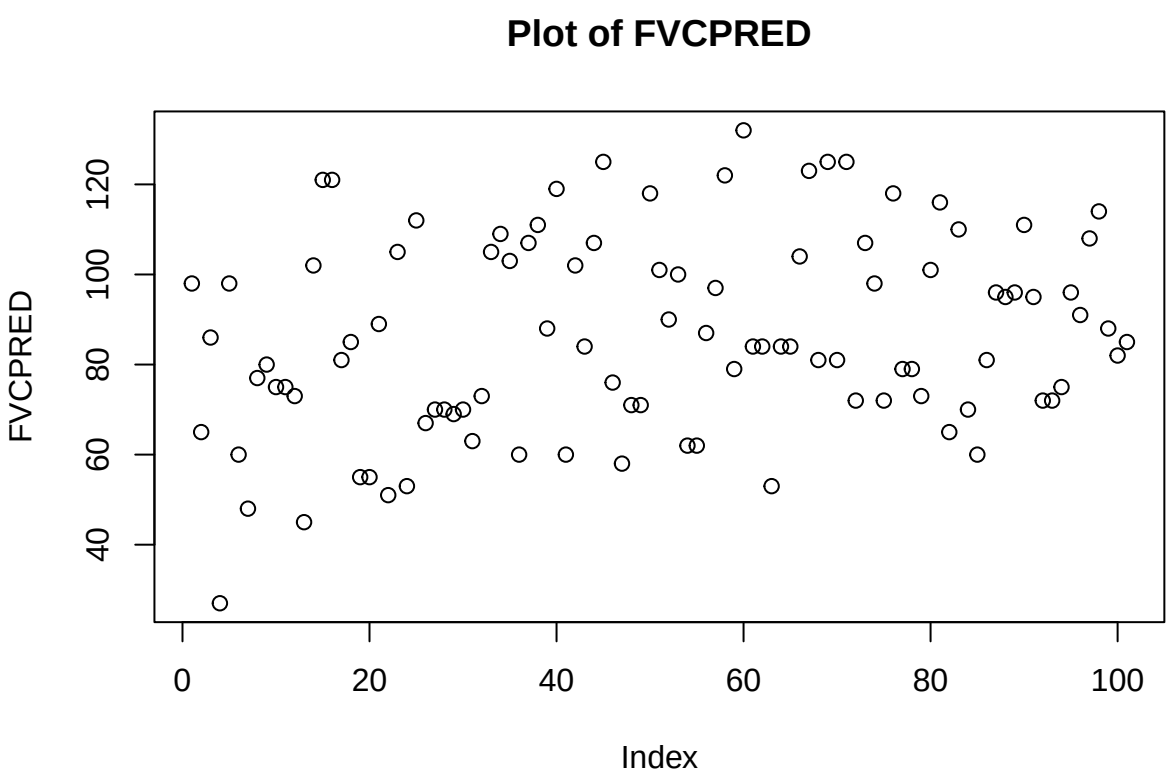

Plot of MWT2
MWT2
100
200
300
400
500
600
700
0
20
40
60
80
100
Index

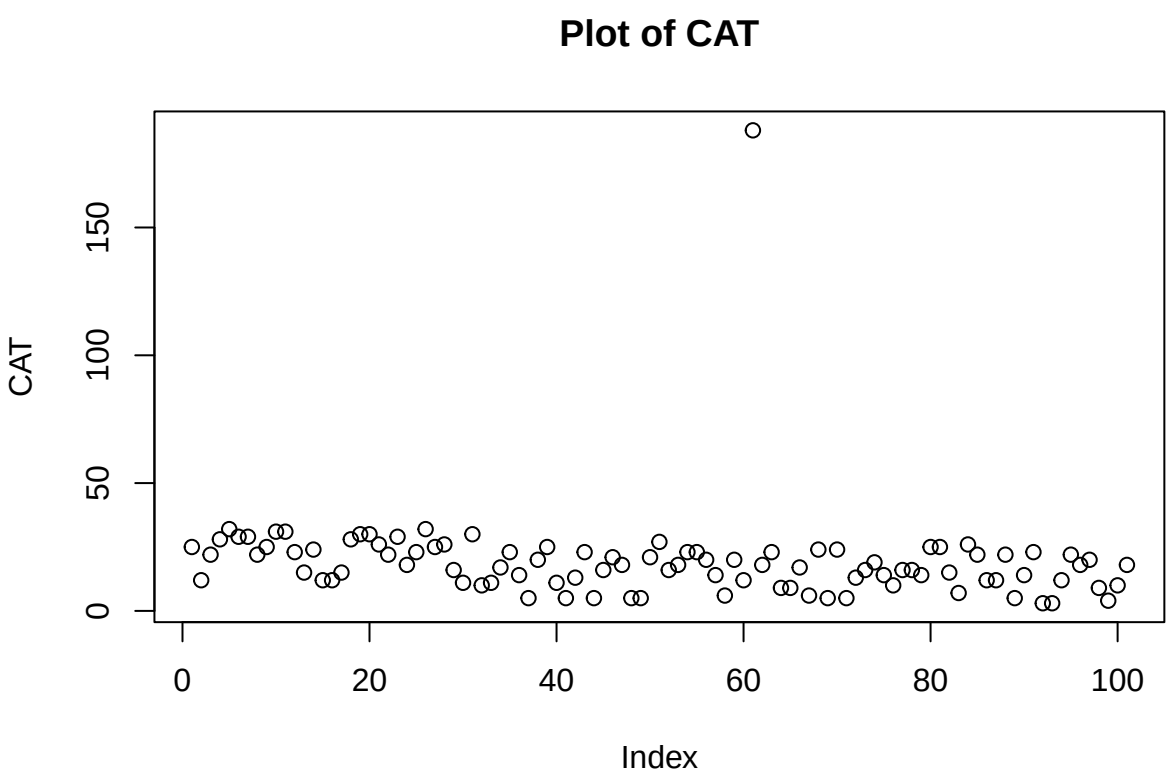

Plot of MWT1Best
MWT1Best
100
200
300
400
500
600
700
0
20
40
60
80
100
Index


Plot of FEV1
FEV1
0.5
1.0
1.5
2.0
2.5
3.0
0
20
40
60
80
100
Index

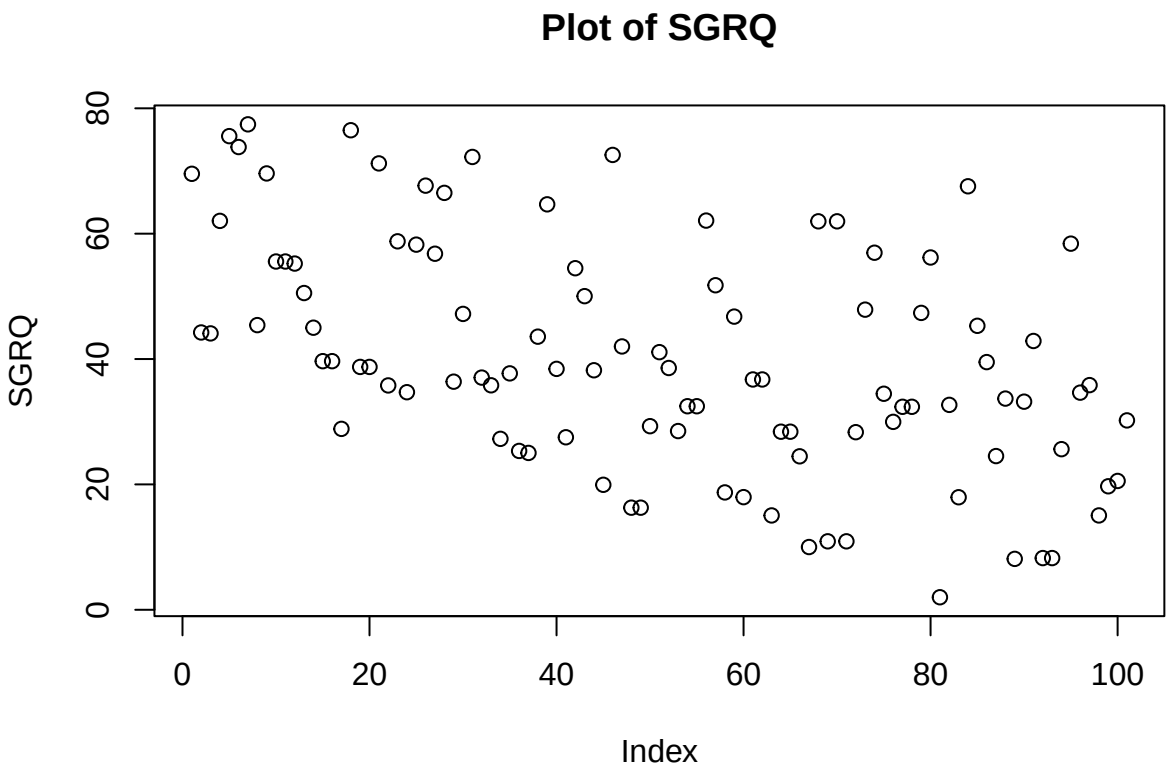

Plot of FEV1PRED
FEV1PRED
0
20
40
60
80
100
Index


Plot of FVC
FVC
1
2
3
4
5
0
20
40
60
80
100
Index


Plot of FVCPRED
FVCPRED
40
60
80
100
120
0
20
40
60
80
100
Index


Plot of CAT
CAT
0
50
100
150
0
20
40
60
80
100
Index


Plot of HAD
HAD
0
10
20
30
40
50
0
20
40
60
80
100
Index



We find a value that might be outlier in MWT1BEst with value= 699 and <150, previously CAT (>40), and HAD (>50)

```
copd$CAT[copd$CAT>40] <- NA
```

```
summary(copd$MWT1Best)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     120     304     420     399     465     699       1
```

```
subset(copd, MWT1Best>650)
```

```
##       X  ID AGE PackHistory COPDSEVERITY MWT1 MWT2 MWT1Best
## 100 100 108  54          30       SEVERE  688  699      699
##     FEV1 FEV1PRED  FVC FVCPRED CAT HAD  SGRQ AGEquartiles
## 100 1.72       44 4.07      82  10  10 20.55            1
##     copd gender smoking Diabetes muscular hypertension
## 100    3      1       2        0        0            0
##     AtrialFib IHD comorbid
## 100         0   0        0
```

```
subset(copd, MWT1Best<150)
```

```
##   X ID AGE PackHistory COPDSEVERITY MWT1 MWT2 MWT1Best FEV1
```

```
## 1 1 58  77          60       SEVERE  120  120      120 1.21
##   FEV1PRED FVC FVCPRED CAT HAD  SGRQ AGEquartiles copd
## 1       36 2.4      98  25   8 69.55            4    3
##   gender smoking Diabetes muscular hypertension AtrialFib
## 1      1       2        1        0            0         1
##   IHD comorbid
## 1   0        1
```

```
subset(copd, HAD>50)
```

```
##     X ID AGE PackHistory COPDSEVERITY MWT1 MWT2 MWT1Best
## 81 81 18  65          20         MILD  468  480      480
##    FEV1 FEV1PRED FVC FVCPRED CAT  HAD SGRQ AGEquartiles
## 81 2.13       92 3.2     116  25 56.2    2            1
##    copd gender smoking Diabetes muscular hypertension
## 81    1      0       2        0        0            1
##    AtrialFib IHD comorbid
## 81         0   0        1
```

```
# Subset the rows based on the condition
subset_copd <- copd[copd$MWT1Best < 650 & copd$MWT1Best > 150, ]
subset_copd <- subset_copd[subset_copd$CAT<40,]
subset_copd <- subset_copd[subset_copd$HAD<50,]

# Check the dimensions of the subsetted data
dim(subset_copd)
```

```
## [1] 98 25
```

```
summary(subset_copd)
```

```
##        X              ID             AGE
##  Min.   : 2.0   Min.   :  1.0   Min.   :44.0
##  1st Qu.:25.8   1st Qu.: 48.8   1st Qu.:65.8
##  Median :49.5   Median : 88.5   Median :71.0
##  Mean   :50.1   Mean   : 91.7   Mean   :70.1
##  3rd Qu.:74.2   3rd Qu.:143.2   3rd Qu.:75.0
##  Max.   :99.0   Max.   :169.0   Max.   :88.0
##  NA's   :2      NA's   :2       NA's   :2
##   PackHistory    COPDSEVERITY           MWT1
##  Min.   :  1.0   Length:98          Min.   :165
##  1st Qu.: 22.2   Class :character   1st Qu.:300
##  Median : 36.0   Mode  :character   Median :416
##  Mean   : 40.0                      Mean   :384
##  3rd Qu.: 51.8                      3rd Qu.:460
##  Max.   :109.0                      Max.   :576
##  NA's   :2                          NA's   :3
##       MWT2        MWT1Best        FEV1
##  Min.   :176   Min.   :176   Min.   :0.45
##  1st Qu.:304   1st Qu.:304   1st Qu.:1.09
##  Median :398   Median :420   Median :1.57
##  Mean   :388   Mean   :398   Mean   :1.60
```

```
##  3rd Qu.:459   3rd Qu.:463   3rd Qu.:1.94
##  Max.   :582   Max.   :582   Max.   :3.18
##  NA's   :2     NA's   :2     NA's   :2
##     FEV1PRED          FVC          FVCPRED
##  Min.   :  3.29   Min.   :1.14   Min.   : 27.0
##  1st Qu.: 42.00   1st Qu.:2.26   1st Qu.: 70.8
##  Median : 60.00   Median :2.79   Median : 84.0
##  Mean   : 58.29   Mean   :2.96   Mean   : 86.1
##  3rd Qu.: 75.00   3rd Qu.:3.63   3rd Qu.:103.2
##  Max.   :102.00   Max.   :5.37   Max.   :132.0
##  NA's   :2        NA's   :2      NA's   :2
##       CAT            HAD            SGRQ       AGEquartiles
##  Min.   : 3.0   Min.   : 0.0   Min.   : 8.12   1   :24
##  1st Qu.:12.0   1st Qu.: 6.0   1st Qu.:28.41   2   :24
##  Median :18.0   Median :10.0   Median :38.50   3   :28
##  Mean   :17.6   Mean   :10.9   Mean   :40.62   4   :20
##  3rd Qu.:23.0   3rd Qu.:15.2   3rd Qu.:55.31   NA's: 2
##  Max.   :32.0   Max.   :30.0   Max.   :77.44
##  NA's   :2      NA's   :2      NA's   :2
##    copd     gender   smoking   Diabetes  muscular
##  1   :21   0   :35   1   :16   0   :76   0   :78
##  2   :42   1   :61   2   :80   1   :20   1   :18
##  3   :25   NA's: 2   NA's: 2   NA's: 2   NA's: 2
##  4   : 8
##  NA's: 2
##
##
##  hypertension AtrialFib   IHD     comorbid
##  0   :85      0   :77   0   :87   0   :44
##  1   :11      1   :19   1   : 9   1   :52
##  NA's: 2      NA's: 2   NA's: 2   NA's: 2
##
##
##
##
```

Pairwise correlation

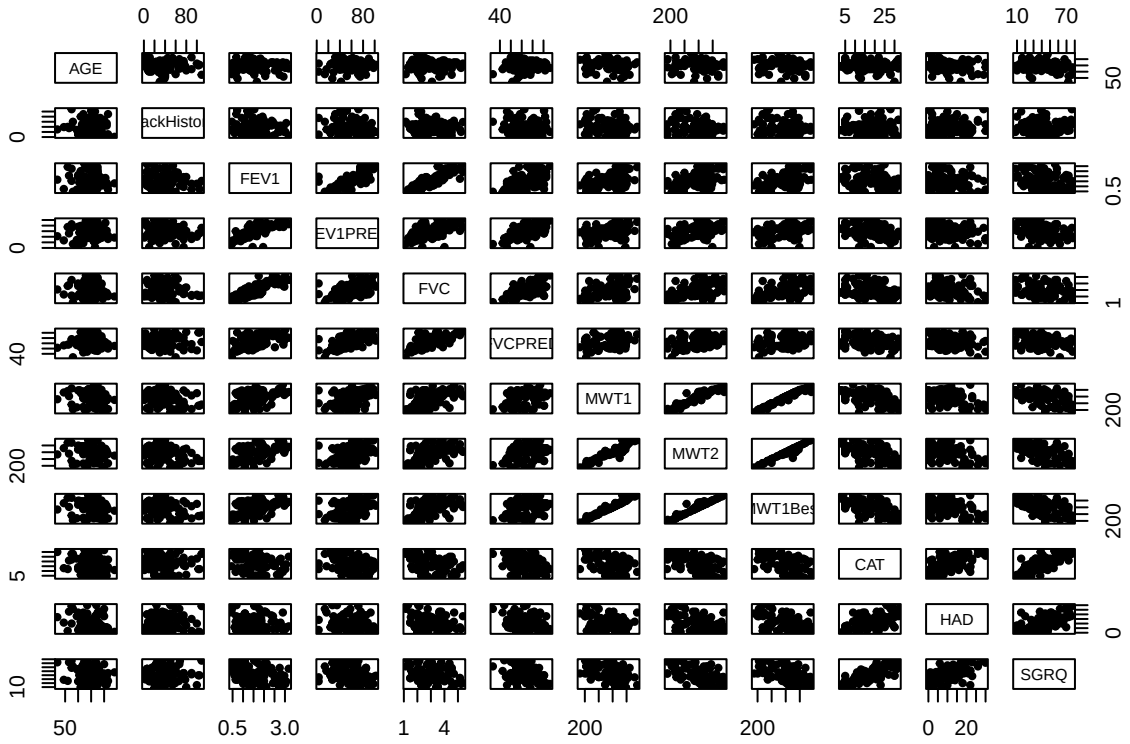
```
my_data<-subset_copd[,c("AGE","PackHistory","FEV1","FEV1PRED","FVC","FVCPRED","MWT1","MWT2","MWT1Best",
cor_matrix <- cor(my_data, use = "complete.obs")
```

```
round(cor_matrix,4)
```

```
##                 AGE PackHistory    FEV1 FEV1PRED     FVC
## AGE          1.0000     -0.0249 -0.0836   0.0730 -0.0952
## PackHistory -0.0249      1.0000 -0.1030  -0.0876 -0.0819
## FEV1        -0.0836     -0.1030  1.0000   0.7784  0.8288
## FEV1PRED     0.0730     -0.0876  0.7784   1.0000  0.5484
## FVC         -0.0952     -0.0819  0.8288   0.5484  1.0000
## FVCPRED      0.0173      0.0057  0.5173   0.6350  0.6388
## MWT1        -0.1863     -0.2280  0.4716   0.3802  0.4441
## MWT2        -0.1742     -0.2581  0.4833   0.4270  0.4518
## MWT1Best    -0.1659     -0.2253  0.4814   0.4064  0.4386
```

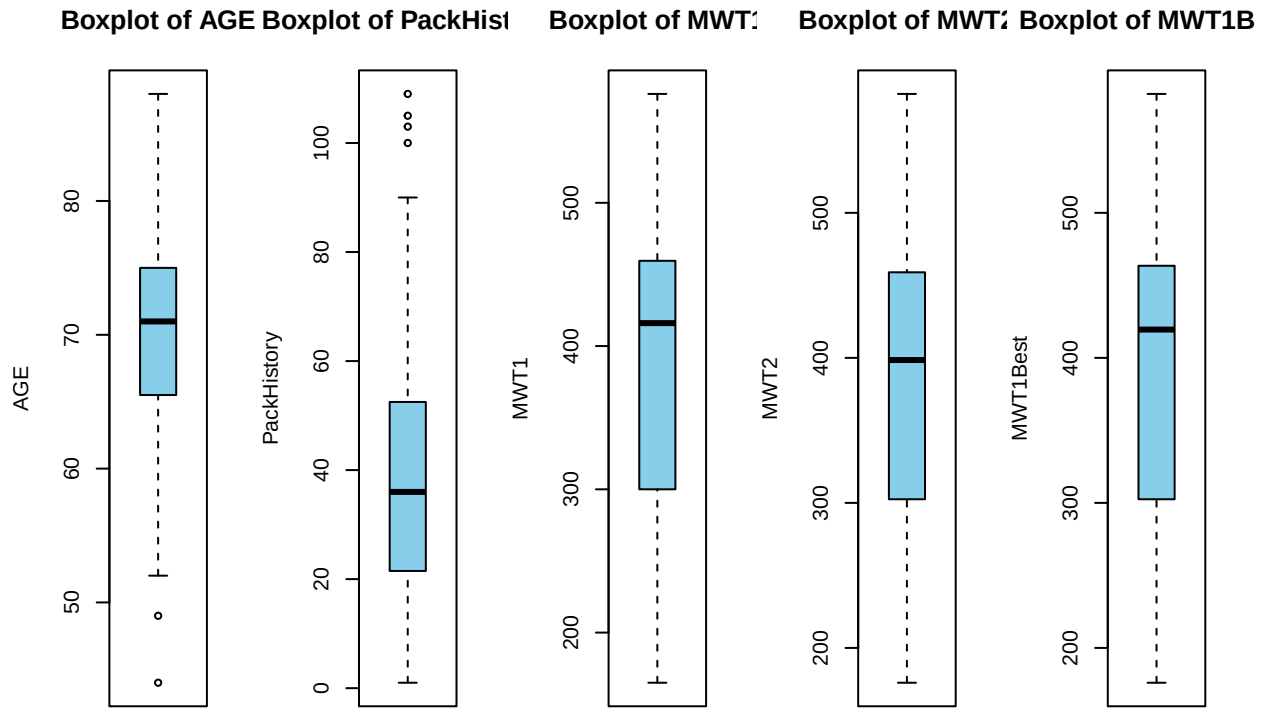
```
## CAT         -0.1091     -0.0268 -0.2763  -0.3107 -0.2290
## HAD         -0.2061      0.0993 -0.2435  -0.2423 -0.1954
## SGRQ        -0.1907      0.0050 -0.3001  -0.3186 -0.2178
##             FVCPRED    MWT1    MWT2 MWT1Best     CAT
## AGE          0.0173 -0.1863 -0.1742  -0.1659 -0.1091
## PackHistory  0.0057 -0.2280 -0.2581  -0.2253 -0.0268
## FEV1         0.5173  0.4716  0.4833   0.4814 -0.2763
## FEV1PRED     0.6350  0.3802  0.4270   0.4064 -0.3107
## FVC          0.6388  0.4441  0.4518   0.4386 -0.2290
## FVCPRED      1.0000  0.2936  0.3393   0.2951 -0.3334
## MWT1         0.2936  1.0000  0.9459   0.9791 -0.4209
## MWT2         0.3393  0.9459  1.0000   0.9791 -0.4922
## MWT1Best     0.2951  0.9791  0.9791   1.0000 -0.4735
## CAT         -0.3334 -0.4209 -0.4922  -0.4735  1.0000
## HAD         -0.2791 -0.3976 -0.4396  -0.4376  0.5774
## SGRQ        -0.2915 -0.4746 -0.4912  -0.5086  0.7700
##                 HAD    SGRQ
## AGE         -0.2061 -0.1907
## PackHistory  0.0993  0.0050
## FEV1        -0.2435 -0.3001
## FEV1PRED    -0.2423 -0.3186
## FVC         -0.1954 -0.2178
## FVCPRED     -0.2791 -0.2915
## MWT1        -0.3976 -0.4746
## MWT2        -0.4396 -0.4912
## MWT1Best    -0.4376 -0.5086
## CAT          0.5774  0.7700
## HAD          1.0000  0.6253
## SGRQ         0.6253  1.0000
```

```
pairs(~AGE+PackHistory+FEV1+FEV1PRED+FVC+FVCPRED+MWT1+MWT2+MWT1Best+CAT+HAD+SGRQ, data=subset_copd, pch=
```



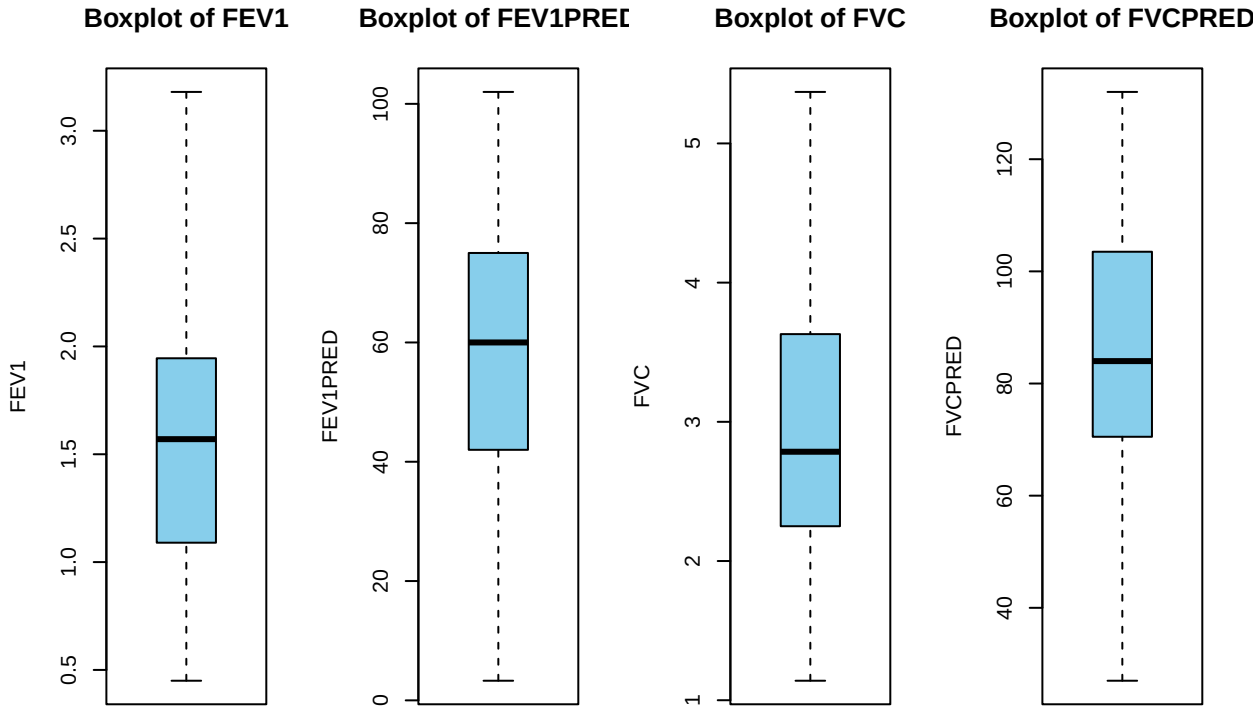

```
num_vars1 <- c("AGE", "PackHistory", "MWT1", "MWT2", "MWT1Best")
num_vars2 <- c("FEV1", "FEV1PRED", "FVC", "FVCPRED")
num_vars3 <- c("CAT","HAD","SGRQ")
```

```
# Create boxplots for each numerical variable
par(mfrow = c(1, length(num_vars1)), mar = c(5, 4, 4, 2))  # Adjusting margins

for(i in 1:length(num_vars1)) {
  boxplot(subset_copd[[num_vars1[i]]], main = paste("Boxplot of", num_vars1[i]),
          ylab = num_vars1[i], col = "skyblue", border = "black")
}
```



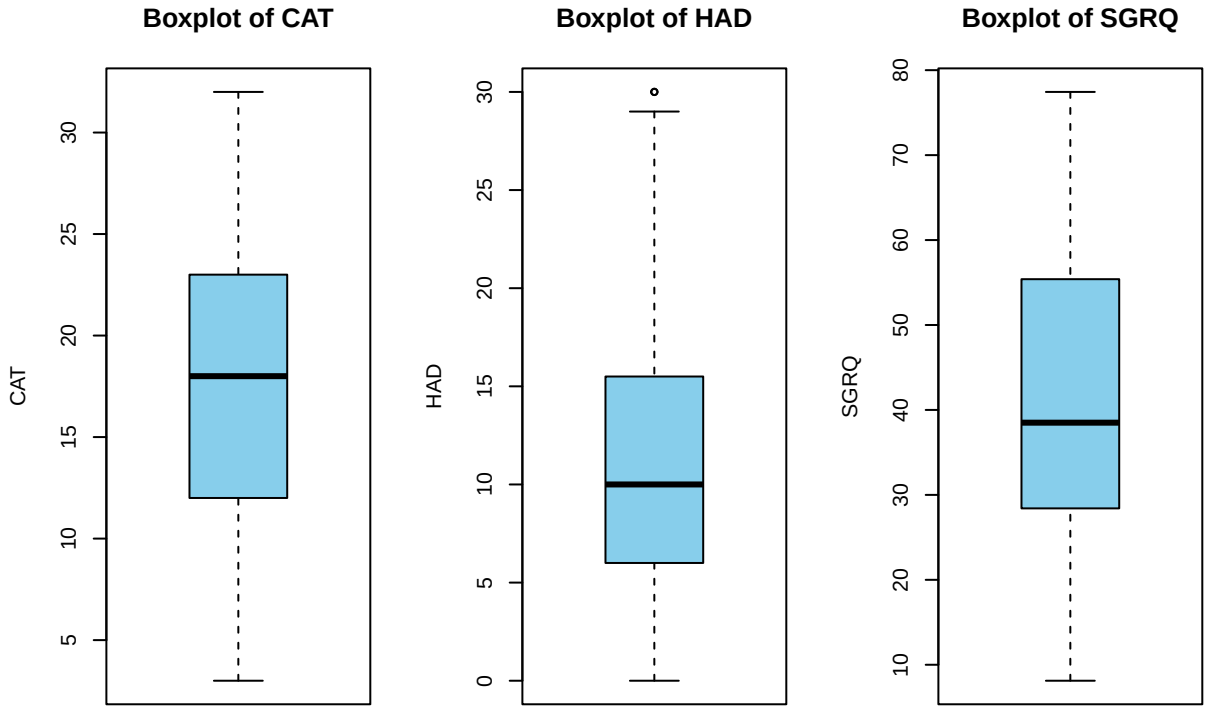

```r
par(mfrow = c(1, 1))  # Reset to single-panel plot
```

```r
# Create boxplots for each numerical variable
par(mfrow = c(1, length(num_vars2)), mar = c(5, 4, 4, 2))  # Adjusting margins

for(i in 1:length(num_vars2)) {
  boxplot(subset_copd[[num_vars2[i]]], main = paste("Boxplot of", num_vars2[i]),
          ylab = num_vars2[i], col = "skyblue", border = "black")
}
```




```
par(mfrow = c(1, 1))  # Reset to single-panel plot
```

```
# Create boxplots for each numerical variable
par(mfrow = c(1, length(num_vars3)), mar = c(5, 4, 4, 2))  # Adjusting margins

for(i in 1:length(num_vars3)) {
  boxplot(subset_copd[[num_vars3[i]]], main = paste("Boxplot of", num_vars3[i]),
          ylab = num_vars3[i], col = "skyblue", border = "black")
}
```

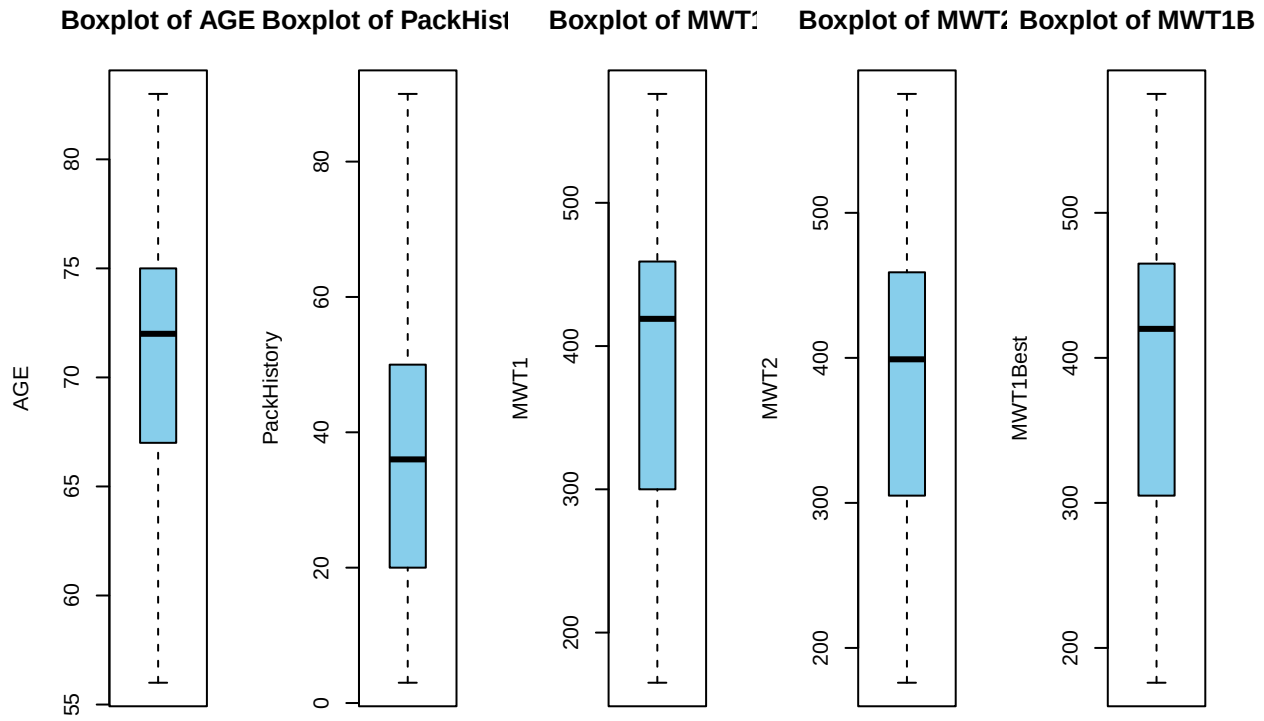



```r
par(mfrow = c(1, 1))  # Reset to single-panel plot
```

Impute more outliers

```r
# Subset the rows based on the condition
subset_copd <- subset_copd[subset_copd$AGE>55 & subset_copd$AGE<85,]
subset_copd <- subset_copd[subset_copd$PackHistory<95,]
subset_copd <- na.omit(subset_copd)
```

```r
# Create boxplots for each numerical variable
par(mfrow = c(1, length(num_vars1)), mar = c(5, 4, 4, 2))  # Adjusting margins
for(i in 1:length(num_vars1)) {
  boxplot(subset_copd[[num_vars1[i]]], main = paste("Boxplot of", num_vars1[i]),
          ylab = num_vars1[i], col = "skyblue", border = "black")
}
```




```
par(mfrow = c(1, 1))  # Reset to single-panel plot
```

```
# Check the dimensions of the subsetted data
dim(subset_copd)
```

```
## [1] 85 25
```

```
describe(subset_copd)
```

```
## subset_copd
##
##  25  Variables      85  Observations
## ---------------------------------------------------------
## X
##        n  missing distinct     Info     Mean      Gmd
##       85        0       85        1    50.24    33.16
##      .05      .10      .25      .50      .75      .90
##      6.2     10.4     26.0     50.0     75.0     88.6
##      .95
##     92.8
##
## lowest :  2  3  4  5  6, highest: 93 96 97 98 99
## ---------------------------------------------------------
## ID
```

```
##        n  missing distinct     Info     Mean      Gmd
##       85        0       83        1    90.12    60.05
##      .05      .10      .25      .50      .75      .90
##     10.2     19.2     48.0     81.0    144.0    158.0
##      .95
##    164.8
##
## lowest :   2   3   6   8  10, highest: 165 166 167 168 169
## ------------------------------------------------------------
## AGE
##        n  missing distinct     Info     Mean      Gmd
##       85        0       26    0.997    71.25    6.835
##      .05      .10      .25      .50      .75      .90
##     62.0     63.4     67.0     72.0     75.0     78.6
##      .95
##     81.0
##
## lowest : 56 59 60 61 62, highest: 79 80 81 82 83
## ------------------------------------------------------------
## PackHistory
##        n  missing distinct     Info     Mean      Gmd
##       85        0       41    0.998    37.95    24.43
##      .05      .10      .25      .50      .75      .90
##      6.0     10.0     20.0     36.0     50.0     67.6
##      .95
##     75.0
##
## lowest :  3  5  6  8  9, highest: 68 75 78 80 90
## ------------------------------------------------------------
## COPDSEVERITY
##        n  missing distinct
##       85        0        4
##
## Value             MILD    MODERATE      SEVERE VERY SEVERE
## Frequency           18          38          22           7
## Proportion       0.212       0.447       0.259       0.082
## ------------------------------------------------------------
## MWT1
##        n  missing distinct     Info     Mean      Gmd
##       85        0       59        1    384.1    112.1
##      .05      .10      .25      .50      .75      .90
##    213.2    226.0    300.0    419.0    459.0    495.0
##      .95
##    508.2
##
## lowest : 165 201 204 210 213, highest: 501 510 511 558 576
## ------------------------------------------------------------
## MWT2
##        n  missing distinct     Info     Mean      Gmd
##       85        0       62        1    389.1    115.4
##      .05      .10      .25      .50      .75      .90
##    214.0    237.0    305.0    399.0    459.0    502.2
##      .95
##    531.0
```

```
##
## lowest : 176 180 210 230 234, highest: 531 563 575 577 582
## ------------------------------------------------------------
## MWT1Best
##        n  missing distinct     Info     Mean      Gmd
##       85        0       61        1    397.5    113.9
##      .05      .10      .25      .50      .75      .90
##    220.2    240.0    305.0    420.0    465.0    503.4
##      .95
##    531.0
##
## lowest : 176 201 210 216 237, highest: 531 558 575 577 582
## ------------------------------------------------------------
## FEV1
##        n  missing distinct     Info     Mean      Gmd
##       85        0       72        1    1.594   0.7745
##      .05      .10      .25      .50      .75      .90
##    0.656    0.724    1.090    1.600    1.920    2.668
##      .95
##    2.880
##
## lowest : 0.45 0.47 0.51 0.6  0.65, highest: 2.9  2.97 3.02 3.06 3.18
## ------------------------------------------------------------
## FEV1PRED
##        n  missing distinct     Info     Mean      Gmd
##       85        0       48    0.999    58.66    25.95
##      .05      .10      .25      .50      .75      .90
##     24.0     30.0     42.0     61.0     75.0     89.2
##      .95
##     93.0
##
## lowest : 3.29 3.39 14   17   24  , highest: 90   93   95   98   102
## ------------------------------------------------------------
## FVC
##        n  missing distinct     Info     Mean      Gmd
##       85        0       69        1    2.975    1.141
##      .05      .10      .25      .50      .75      .90
##    1.528    1.810    2.290    2.800    3.630    4.432
##      .95
##    4.716
##
## lowest : 1.14 1.31 1.47 1.52 1.56, highest: 4.72 4.9  5.15 5.23 5.37
## ------------------------------------------------------------
## FVCPRED
##        n  missing distinct     Info     Mean      Gmd
##       85        0       50    0.999    86.31    25.44
##      .05      .10      .25      .50      .75      .90
##     53.4     60.0     71.0     84.0    104.0    116.4
##      .95
##    122.8
##
## lowest :  27  45  48  51  53, highest: 119 122 123 125 132
## ------------------------------------------------------------
## CAT
```

```
##        n  missing distinct     Info     Mean      Gmd
##       85        0       29    0.997    17.52    9.323
##      .05      .10      .25      .50      .75      .90
##      5.0      5.0     12.0     18.0     23.0     28.6
##      .95
##     30.0
##
## lowest :  3  4  5  6  7, highest: 28 29 30 31 32
## ------------------------------------------------------------
## HAD
##        n  missing distinct     Info     Mean      Gmd
##       85        0       26    0.997    10.76    8.034
##      .05      .10      .25      .50      .75      .90
##      1.0      2.0      6.0      9.0     15.0     20.6
##      .95
##     22.8
##
## lowest :  0  1  2  3  4, highest: 21 22 23 26 30
## ------------------------------------------------------------
## SGRQ
##        n  missing distinct     Info     Mean      Gmd
##       85        0       76        1    39.68    19.96
##      .05      .10      .25      .50      .75      .90
##    10.92    16.95    28.41    37.71    54.49    63.64
##      .95
##    69.22
##
## lowest : 8.12  8.25  10.01 10.92 15.05
## highest: 69.61 71.21 73.82 75.56 77.44
## ------------------------------------------------------------
## AGEquartiles
##        n  missing distinct
##       85        0        4
##
## Value          1     2     3     4
## Frequency     17    22    27    19
## Proportion 0.200 0.259 0.318 0.224
## ------------------------------------------------------------
## copd
##        n  missing distinct
##       85        0        4
##
## Value          1     2     3     4
## Frequency     18    38    22     7
## Proportion 0.212 0.447 0.259 0.082
## ------------------------------------------------------------
## gender
##        n  missing distinct
##       85        0        2
##
## Value          0     1
## Frequency     31    54
## Proportion 0.365 0.635
## ------------------------------------------------------------
```

```
## smoking
##        n  missing distinct
##       85        0        2
##
## Value          1     2
## Frequency     14    71
## Proportion 0.165 0.835
## ------------------------------------------------------------
## Diabetes
##        n  missing distinct
##       85        0        2
##
## Value        0   1
## Frequency   68  17
## Proportion 0.8 0.2
## ------------------------------------------------------------
## muscular
##        n  missing distinct
##       85        0        2
##
## Value          0     1
## Frequency     69    16
## Proportion 0.812 0.188
## ------------------------------------------------------------
## hypertension
##        n  missing distinct
##       85        0        2
##
## Value          0     1
## Frequency     76     9
## Proportion 0.894 0.106
## ------------------------------------------------------------
## AtrialFib
##        n  missing distinct
##       85        0        2
##
## Value          0     1
## Frequency     66    19
## Proportion 0.776 0.224
## ------------------------------------------------------------
## IHD
##        n  missing distinct
##       85        0        2
##
## Value          0     1
## Frequency     78     7
## Proportion 0.918 0.082
## ------------------------------------------------------------
## comorbid
##        n  missing distinct
##       85        0        2
##
## Value          0     1
## Frequency     38    47
```

```
## Proportion 0.447 0.553
## ------------------------------------------------------------
```

```
summary(subset_copd)
```

```
##        X              ID              AGE
##  Min.   : 2.0   Min.   :  2.0   Min.   :56.0
##  1st Qu.:26.0   1st Qu.: 48.0   1st Qu.:67.0
##  Median :50.0   Median : 81.0   Median :72.0
##  Mean   :50.2   Mean   : 90.1   Mean   :71.2
##  3rd Qu.:75.0   3rd Qu.:144.0   3rd Qu.:75.0
##  Max.   :99.0   Max.   :169.0   Max.   :83.0
##   PackHistory   COPDSEVERITY            MWT1
##  Min.   : 3.0   Length:85          Min.   :165
##  1st Qu.:20.0   Class :character   1st Qu.:300
##  Median :36.0   Mode  :character   Median :419
##  Mean   :37.9                      Mean   :384
##  3rd Qu.:50.0                      3rd Qu.:459
##  Max.   :90.0                      Max.   :576
##       MWT2        MWT1Best        FEV1
##  Min.   :176   Min.   :176   Min.   :0.45
##  1st Qu.:305   1st Qu.:305   1st Qu.:1.09
##  Median :399   Median :420   Median :1.60
##  Mean   :389   Mean   :397   Mean   :1.59
##  3rd Qu.:459   3rd Qu.:465   3rd Qu.:1.92
##  Max.   :582   Max.   :582   Max.   :3.18
##     FEV1PRED           FVC          FVCPRED
##  Min.   :  3.29   Min.   :1.14   Min.   : 27.0
##  1st Qu.: 42.00   1st Qu.:2.29   1st Qu.: 71.0
##  Median : 61.00   Median :2.80   Median : 84.0
##  Mean   : 58.66   Mean   :2.98   Mean   : 86.3
##  3rd Qu.: 75.00   3rd Qu.:3.63   3rd Qu.:104.0
##  Max.   :102.00   Max.   :5.37   Max.   :132.0
##       CAT            HAD            SGRQ       AGEquartiles
##  Min.   : 3.0   Min.   : 0.0   Min.   : 8.12   1:17
##  1st Qu.:12.0   1st Qu.: 6.0   1st Qu.:28.41   2:22
##  Median :18.0   Median : 9.0   Median :37.71   3:27
##  Mean   :17.5   Mean   :10.8   Mean   :39.68   4:19
##  3rd Qu.:23.0   3rd Qu.:15.0   3rd Qu.:54.49
##  Max.   :32.0   Max.   :30.0   Max.   :77.44
##  copd   gender smoking Diabetes muscular hypertension
##  1:18   0:31   1:14    0:68     0:69     0:76
##  2:38   1:54   2:71    1:17     1:16     1: 9
##  3:22
##  4: 7
##
##
##  AtrialFib IHD    comorbid
##  0:66      0:78   0:38
##  1:19      1: 7   1:47
##
##
##
##
```

```r
# Initialize vectors to store results
means <- numeric()
stds <- numeric()

# Loop through each column in subset_copd
for (col in names(subset_copd)) {
  # Check if the column is numeric
  if (is.numeric(subset_copd[[col]])) {
    # Calculate mean and standard deviation
    mean_value <- mean(subset_copd[[col]])
    sd_value <- sd(subset_copd[[col]])

    # Print or store results
    cat("Column:", col,"\n")
    cat("Mean:", mean_value, "\n")
    cat("Standard deviation:", sd_value, "\n\n")

    # Store results in vectors
    means <- c(means, mean_value)
    stds <- c(stds, sd_value)
  }
}
```

```
## Column: X
## Mean: 50.2353
## Standard deviation: 28.5601
##
## Column: ID
## Mean: 90.1176
## Standard deviation: 51.9541
##
## Column: AGE
## Mean: 71.2471
## Standard deviation: 5.97596
##
## Column: PackHistory
## Mean: 37.9471
## Standard deviation: 21.4343
##
## Column: MWT1
## Mean: 384.082
## Standard deviation: 98.9755
##
## Column: MWT2
## Mean: 389.106
## Standard deviation: 101.604
##
## Column: MWT1Best
## Mean: 397.459
## Standard deviation: 100.516
##
## Column: FEV1
## Mean: 1.59435
```

```
## Standard deviation: 0.681496
##
## Column: FEV1PRED
## Mean: 58.6551
## Standard deviation: 22.6763
##
## Column: FVC
## Mean: 2.97506
## Standard deviation: 1.00217
##
## Column: FVCPRED
## Mean: 86.3059
## Standard deviation: 22.1591
##
## Column: CAT
## Mean: 17.5176
## Standard deviation: 8.07791
##
## Column: HAD
## Mean: 10.7647
## Standard deviation: 7.1325
##
## Column: SGRQ
## Mean: 39.6791
## Standard deviation: 17.4346
```

```
# Print overall means and standard deviations
cat("Overall Means: \n")
```

```
## Overall Means:
```

```
print(means)
```

```
##  [1]  50.23529  90.11765  71.24706  37.94706 384.08235
##  [6] 389.10588 397.45882   1.59435  58.65506   2.97506
## [11]  86.30588  17.51765  10.76471  39.67906
```

```
cat("Overall standard deviation: \n")
```

```
## Overall standard deviation:
```

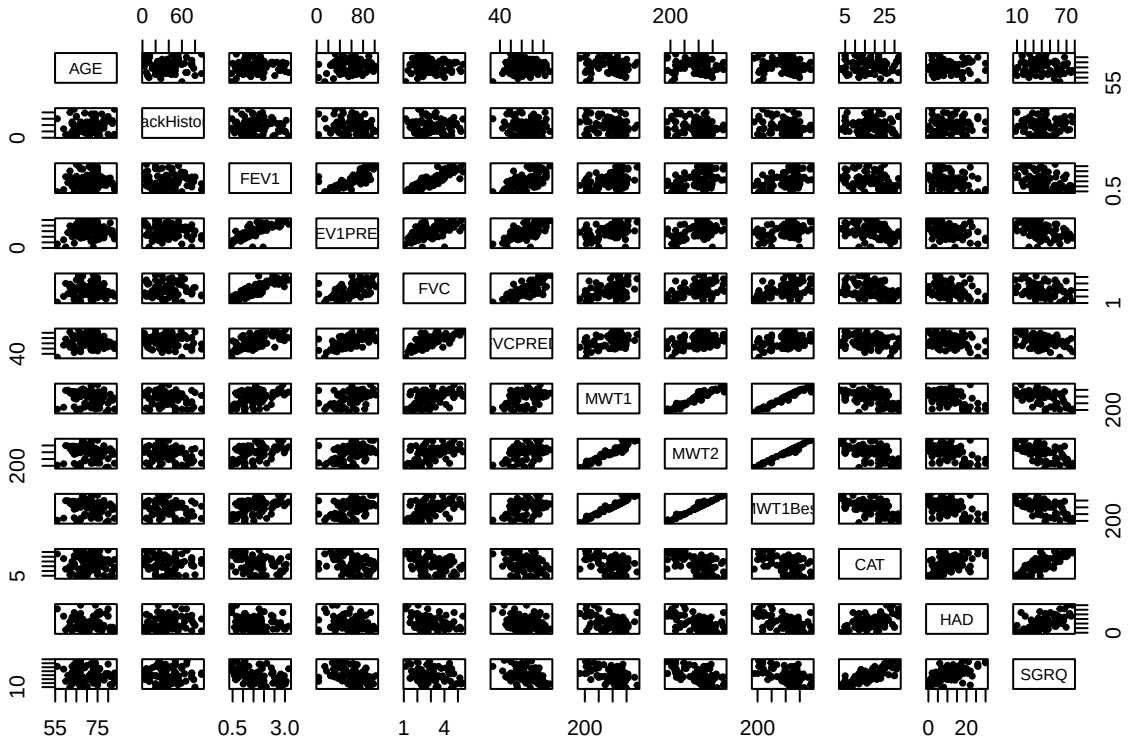
```
print(stds)
```

```
##  [1]  28.560148  51.954057   5.975956  21.434288  98.975491
##  [6] 101.603926 100.516470   0.681496  22.676284   1.002175
## [11]  22.159070   8.077912   7.132499  17.434577
```

```
cor_matrix <- cor(my_data, use = "complete.obs")
round(cor_matrix,4)
```

```
##                 AGE PackHistory    FEV1 FEV1PRED     FVC
## AGE          1.0000     -0.0249 -0.0836   0.0730 -0.0952
## PackHistory -0.0249      1.0000 -0.1030  -0.0876 -0.0819
## FEV1        -0.0836     -0.1030  1.0000   0.7784  0.8288
## FEV1PRED     0.0730     -0.0876  0.7784   1.0000  0.5484
## FVC         -0.0952     -0.0819  0.8288   0.5484  1.0000
## FVCPRED      0.0173      0.0057  0.5173   0.6350  0.6388
## MWT1        -0.1863     -0.2280  0.4716   0.3802  0.4441
## MWT2        -0.1742     -0.2581  0.4833   0.4270  0.4518
## MWT1Best    -0.1659     -0.2253  0.4814   0.4064  0.4386
## CAT         -0.1091     -0.0268 -0.2763  -0.3107 -0.2290
## HAD         -0.2061      0.0993 -0.2435  -0.2423 -0.1954
## SGRQ        -0.1907      0.0050 -0.3001  -0.3186 -0.2178
##             FVCPRED    MWT1    MWT2 MWT1Best     CAT
## AGE          0.0173 -0.1863 -0.1742  -0.1659 -0.1091
## PackHistory  0.0057 -0.2280 -0.2581  -0.2253 -0.0268
## FEV1         0.5173  0.4716  0.4833   0.4814 -0.2763
## FEV1PRED     0.6350  0.3802  0.4270   0.4064 -0.3107
## FVC          0.6388  0.4441  0.4518   0.4386 -0.2290
## FVCPRED      1.0000  0.2936  0.3393   0.2951 -0.3334
## MWT1         0.2936  1.0000  0.9459   0.9791 -0.4209
## MWT2         0.3393  0.9459  1.0000   0.9791 -0.4922
## MWT1Best     0.2951  0.9791  0.9791   1.0000 -0.4735
## CAT         -0.3334 -0.4209 -0.4922  -0.4735  1.0000
## HAD         -0.2791 -0.3976 -0.4396  -0.4376  0.5774
## SGRQ        -0.2915 -0.4746 -0.4912  -0.5086  0.7700
##                 HAD    SGRQ
## AGE         -0.2061 -0.1907
## PackHistory  0.0993  0.0050
## FEV1        -0.2435 -0.3001
## FEV1PRED    -0.2423 -0.3186
## FVC         -0.1954 -0.2178
## FVCPRED     -0.2791 -0.2915
## MWT1        -0.3976 -0.4746
## MWT2        -0.4396 -0.4912
## MWT1Best    -0.4376 -0.5086
## CAT          0.5774  0.7700
## HAD          1.0000  0.6253
## SGRQ         0.6253  1.0000
```

```
pairs(~AGE+PackHistory+FEV1+FEV1PRED+FVC+FVCPRED+MWT1+MWT2+MWT1Best+CAT+HAD+SGRQ, data=subset_copd, pch=
```



Dependent variable : SGRQ / Quality of Life Independet variable : AGE,PackHistory,FEV1,FVC,MWT1Best, CAT, HAD

Linear regression model

```
# List of independent variables (replace with your actual variable names)
independent_vars <- c("AGE", "PackHistory", "FEV1", "FVC", "MWT1Best","CAT","HAD","gender","COPDSEVERIT

# Dependent variable
dependent_var <- "SGRQ"  # Replace with your dependent variable name

# Perform linear regression for each independent variable
for(var in independent_vars) {
  formula <- paste(dependent_var, "~", var)
  model <- lm(formula, data = subset_copd)
  cat("Linear Regression Summary for", var, ":\n")
  print(summary(model))
  cat("95% CI", var, ":\n")
  print(confint(model))
  cat("\n")
}
```

```
## Linear Regression Summary for AGE :
##
## Call:
## lm(formula = formula, data = subset_copd)
```

```
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -32.68 -11.23  -1.17  12.31  36.64
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   58.542     22.800    2.57    0.012 *
## AGE           -0.265      0.319   -0.83    0.409
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.5 on 83 degrees of freedom
## Multiple R-squared:  0.00824,    Adjusted R-squared:  -0.00371
## F-statistic: 0.689 on 1 and 83 DF,  p-value: 0.409
##
## 95% CI AGE :
##                 2.5 %     97.5 %
## (Intercept) 13.192812 103.890627
## AGE         -0.899051   0.369551
##
## Linear Regression Summary for PackHistory :
##
## Call:
## lm(formula = formula, data = subset_copd)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -31.59 -11.26  -1.99  14.77  37.78
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept) 39.73759    3.88555   10.23 0.00000000000000023
## PackHistory -0.00154    0.08928   -0.02                0.99
##
## (Intercept) ***
## PackHistory
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.5 on 83 degrees of freedom
## Multiple R-squared:  3.6e-06,    Adjusted R-squared:  -0.012
## F-statistic: 0.000298 on 1 and 83 DF,  p-value: 0.986
##
## 95% CI PackHistory :
##                2.5 %    97.5 %
## (Intercept) 32.00938 47.465797
## PackHistory -0.17912  0.176035
##
## Linear Regression Summary for FEV1 :
##
## Call:
```

```
## lm(formula = formula, data = subset_copd)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -28.54 -10.94  -0.94  11.47  35.17
##
## Coefficients:
##             Estimate Std. Error t value
## (Intercept)    55.26       4.50   12.29
## FEV1           -9.77       2.60   -3.77
##                          Pr(>|t|)
## (Intercept) < 0.0000000000000002 ***
## FEV1                      0.00031 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.2 on 83 degrees of freedom
## Multiple R-squared:  0.146,  Adjusted R-squared:  0.136
## F-statistic: 14.2 on 1 and 83 DF,  p-value: 0.000309
##
## 95% CI FEV1 :
##                2.5 %   97.5 %
## (Intercept)  46.3191 64.20204
## FEV1        -14.9345 -4.61137
##
## Linear Regression Summary for FVC :
##
## Call:
## lm(formula = formula, data = subset_copd)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -33.17 -11.15  -2.63  11.52  36.62
##
## Coefficients:
##             Estimate Std. Error t value          Pr(>|t|)
## (Intercept)    53.85       5.76    9.35 0.000000000000013
## FVC            -4.76       1.84   -2.59             0.011
##
## (Intercept) ***
## FVC         *
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.9 on 83 degrees of freedom
## Multiple R-squared:  0.075,  Adjusted R-squared:  0.0638
## F-statistic: 6.73 on 1 and 83 DF,  p-value: 0.0112
##
## 95% CI FVC :
##                2.5 %   97.5 %
## (Intercept) 42.38800 65.30944
## FVC         -8.41571 -1.10993
```

```
##
## Linear Regression Summary for MWT1Best :
##
## Call:
## lm(formula = formula, data = subset_copd)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -23.91 -11.09  -1.04  10.59  36.45
##
## Coefficients:
##             Estimate Std. Error t value
## (Intercept)  76.8661     6.5700   11.70
## MWT1Best     -0.0936     0.0160   -5.84
##                          Pr(>|t|)
## (Intercept) < 0.0000000000000002 ***
## MWT1Best               0.0000001 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.8 on 83 degrees of freedom
## Multiple R-squared:  0.291,  Adjusted R-squared:  0.282
## F-statistic: 34.1 on 1 and 83 DF,  p-value: 0.0000001
##
## 95% CI MWT1Best :
##                  2.5 %     97.5 %
## (Intercept) 63.798595 89.9336523
## MWT1Best    -0.125448 -0.0616766
##
## Linear Regression Summary for CAT :
##
## Call:
## lm(formula = formula, data = subset_copd)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -22.40  -7.24  -0.68   9.32  22.58
##
## Coefficients:
##             Estimate Std. Error t value
## (Intercept)    9.566      2.761    3.47
## CAT            1.719      0.143   12.00
##                          Pr(>|t|)
## (Intercept)               0.00084 ***
## CAT         < 0.0000000000000002 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.6 on 83 degrees of freedom
## Multiple R-squared:  0.634,  Adjusted R-squared:  0.63
## F-statistic:  144 on 1 and 83 DF,  p-value: <0.0000000000000002
##
```

```
## 95% CI CAT :
##                2.5 %   97.5 %
## (Intercept) 4.07551 15.05688
## CAT         1.43407  2.00393
##
## Linear Regression Summary for HAD :
##
## Call:
## lm(formula = formula, data = subset_copd)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -41.68  -8.86  -0.77   8.56  37.16
##
## Coefficients:
##             Estimate Std. Error t value         Pr(>|t|)
## (Intercept)    23.34       2.71    8.61 0.00000000000039
## HAD             1.52       0.21    7.22 0.00000000022818
##
## (Intercept) ***
## HAD         ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.7 on 83 degrees of freedom
## Multiple R-squared:  0.386,  Adjusted R-squared:  0.378
## F-statistic: 52.1 on 1 and 83 DF,  p-value: 0.000000000228
##
## 95% CI HAD :
##                2.5 %   97.5 %
## (Intercept) 17.94617 28.72904
## HAD          1.09979  1.93632
##
## Linear Regression Summary for gender :
##
## Call:
## lm(formula = formula, data = subset_copd)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -31.24 -11.96  -1.74  13.24  36.79
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)    41.25       3.14   13.13 <0.0000000000000002
## gender1        -2.48       3.94   -0.63                0.53
##
## (Intercept) ***
## gender1
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 17.5 on 83 degrees of freedom
## Multiple R-squared:  0.00474,    Adjusted R-squared:  -0.00725
## F-statistic: 0.395 on 1 and 83 DF,  p-value: 0.531
##
## 95% CI gender :
##                2.5 %   97.5 %
## (Intercept)  35.0035 47.50484
## gender1     -10.3216  5.36282
##
## Linear Regression Summary for COPDSEVERITY :
##
## Call:
## lm(formula = formula, data = subset_copd)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -28.03 -11.97  -0.71  11.21  37.67
##
## Coefficients:
##                           Estimate Std. Error t value
## (Intercept)                  30.69       3.76    8.17
## COPDSEVERITYMODERATE          5.46       4.56    1.20
## COPDSEVERITYSEVERE           19.05       5.07    3.76
## COPDSEVERITYVERY SEVERE      19.63       7.10    2.76
##                                  Pr(>|t|)
## (Intercept)             0.0000000000035 ***
## COPDSEVERITYMODERATE            0.23487
## COPDSEVERITYSEVERE              0.00032 ***
## COPDSEVERITYVERY SEVERE         0.00705 **
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.9 on 81 degrees of freedom
## Multiple R-squared:  0.194,  Adjusted R-squared:  0.164
## F-statistic: 6.49 on 3 and 81 DF,  p-value: 0.000545
##
## 95% CI COPDSEVERITY :
##                              2.5 %  97.5 %
## (Intercept)              23.21521 38.1681
## COPDSEVERITYMODERATE    -3.61669 14.5355
## COPDSEVERITYSEVERE       8.96572 29.1282
## COPDSEVERITYVERY SEVERE  5.50488 33.7632
##
## Linear Regression Summary for comorbid :
##
## Call:
## lm(formula = formula, data = subset_copd)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -31.77  -9.31  -3.21  13.45  35.66
##
## Coefficients:
```

```
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)    37.08       2.82   13.15 <0.0000000000000002
## comorbid1       4.71       3.79    1.24                0.22
##
## (Intercept) ***
## comorbid1
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.4 on 83 degrees of freedom
## Multiple R-squared:  0.0182, Adjusted R-squared:  0.00639
## F-statistic: 1.54 on 1 and 83 DF,  p-value: 0.218
##
## 95% CI comorbid :
##                2.5 %  97.5 %
## (Intercept) 31.46982 42.6844
## comorbid1   -2.83507 12.2464
##
## Linear Regression Summary for Diabetes :
##
## Call:
## lm(formula = formula, data = subset_copd)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -30.1  -10.9   -1.6   11.8   37.4
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)    38.18       2.09    18.2 <0.0000000000000002
## Diabetes1       7.51       4.68     1.6                0.11
##
## (Intercept) ***
## Diabetes1
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.3 on 83 degrees of freedom
## Multiple R-squared:  0.03,   Adjusted R-squared:  0.0183
## F-statistic: 2.57 on 1 and 83 DF,  p-value: 0.113
##
## 95% CI Diabetes :
##                2.5 %  97.5 %
## (Intercept) 34.0108 42.3436
## Diabetes1   -1.8071 16.8256
##
## Linear Regression Summary for hypertension :
##
## Call:
## lm(formula = formula, data = subset_copd)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max
## -31.06 -10.77  -2.44  15.31  38.26
##
## Coefficients:
##               Estimate Std. Error t value
## (Intercept)      39.18       2.00   19.54
## hypertension1     4.75       6.16    0.77
##                          Pr(>|t|)
## (Intercept)   <0.0000000000000002 ***
## hypertension1                0.44
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.5 on 83 degrees of freedom
## Multiple R-squared:  0.00712,    Adjusted R-squared:  -0.00485
## F-statistic: 0.595 on 1 and 83 DF,  p-value: 0.443
##
## 95% CI hypertension :
##                   2.5 %  97.5 %
## (Intercept)   35.18861 43.1632
## hypertension1 -7.50188 17.0056
##
## Linear Regression Summary for muscular :
##
## Call:
## lm(formula = formula, data = subset_copd)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -31.14 -10.85  -1.55  14.06  38.18
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)    39.26       2.11   18.62 <0.0000000000000002
## muscular1       2.24       4.86    0.46                0.65
##
## (Intercept) ***
## muscular1
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.5 on 83 degrees of freedom
## Multiple R-squared:  0.00256,    Adjusted R-squared:  -0.00945
## F-statistic: 0.213 on 1 and 83 DF,  p-value: 0.645
##
## 95% CI muscular :
##                  2.5 %  97.5 %
## (Intercept) 35.06225 43.4508
## muscular1   -7.42258 11.9120
##
## Linear Regression Summary for AtrialFib :
##
```

```
## Call:
## lm(formula = formula, data = subset_copd)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -35.27  -9.66  -1.19  10.28  37.49
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)    38.07       2.13   17.90 <0.0000000000000002
## AtrialFib1      7.22       4.50    1.61                0.11
##
## (Intercept) ***
## AtrialFib1
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.3 on 83 degrees of freedom
## Multiple R-squared:  0.0301, Adjusted R-squared:  0.0184
## F-statistic: 2.58 on 1 and 83 DF,  p-value: 0.112
##
## 95% CI AtrialFib :
##                2.5 %  97.5 %
## (Intercept) 33.83655 42.2944
## AtrialFib1  -1.72583 16.1633
##
## Linear Regression Summary for IHD :
##
## Call:
## lm(formula = formula, data = subset_copd)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -31.65 -11.36  -2.06  14.72  37.67
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)    39.77       1.99   20.03 <0.0000000000000002
## IHD1           -1.14       6.92   -0.16                0.87
##
## (Intercept) ***
## IHD1
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.5 on 83 degrees of freedom
## Multiple R-squared:  0.000328,   Adjusted R-squared:  -0.0117
## F-statistic: 0.0272 on 1 and 83 DF,  p-value: 0.869
##
## 95% CI IHD :
##                2.5 %  97.5 %
## (Intercept)  35.8238 43.7224
```

```
## IHD1        -14.9036 12.6203
##
## Linear Regression Summary for smoking :
##
## Call:
## lm(formula = formula, data = subset_copd)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -29.8  -12.9   -1.9   12.6   37.6
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)    48.67       4.56   10.67 <0.0000000000000002
## smoking2      -10.76       4.99   -2.16               0.034
##
## (Intercept) ***
## smoking2    *
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.1 on 83 degrees of freedom
## Multiple R-squared:  0.053,  Adjusted R-squared:  0.0416
## F-statistic: 4.65 on 1 and 83 DF,  p-value: 0.034
##
## 95% CI smoking :
##                2.5 %    97.5 %
## (Intercept)  39.5943 57.739959
## smoking2    -20.6875 -0.833282
```

Variables which show significant correlation with SGRQ : FEV1, FVC, MWT1Best, CAT, HAD, COPD-Severity

MULTIPLE LINEAR REGRESSION

Include all variables

```
sgrq_model <- lm(SGRQ~AGE+gender+FEV1+FVC+CAT+MWT1Best+COPDSEVERITY+HAD+comorbid+smoking+Diabetes+hyper
```

```
summary(sgrq_model)
```

```
##
## Call:
## lm(formula = SGRQ ~ AGE + gender + FEV1 + FVC + CAT + MWT1Best +
##     COPDSEVERITY + HAD + comorbid + smoking + Diabetes + hypertension +
##     muscular + AtrialFib + IHD, data = subset_copd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.870  -6.913  -0.291   6.283  19.413
##
## Coefficients:
##                         Estimate Std. Error t value
```

```
## (Intercept)              37.8685    25.1568    1.51
## AGE                      -0.2046     0.2380   -0.86
## gender1                   2.0044     3.1970    0.63
## FEV1                     -6.4394     4.2714   -1.51
## FVC                       2.0352     2.2613    0.90
## CAT                       1.3734     0.2025    6.78
## MWT1Best                 -0.0185     0.0201   -0.92
## COPDSEVERITYMODERATE      0.8246     3.6229    0.23
## COPDSEVERITYSEVERE       -0.8901     5.7670   -0.15
## COPDSEVERITYVERY SEVERE -10.6587     8.4018   -1.27
## HAD                       0.4702     0.2066    2.28
## comorbid1                 0.3530     4.5884    0.08
## smoking2                 -3.1801     3.2354   -0.98
## Diabetes1                 0.9140     3.9499    0.23
## hypertension1             5.5435     4.8509    1.14
## muscular1                -1.9994     4.0464   -0.49
## AtrialFib1                1.2913     4.3486    0.30
## IHD1                      0.9149     4.9366    0.19
##                            Pr(>|t|)
## (Intercept)                   0.137
## AGE                           0.393
## gender1                       0.533
## FEV1                          0.136
## FVC                           0.371
## CAT                  0.0000000037 ***
## MWT1Best                      0.362
## COPDSEVERITYMODERATE          0.821
## COPDSEVERITYSEVERE            0.878
## COPDSEVERITYVERY SEVERE       0.209
## HAD                           0.026 *
## comorbid1                     0.939
## smoking2                      0.329
## Diabetes1                     0.818
## hypertension1                 0.257
## muscular1                     0.623
## AtrialFib1                    0.767
## IHD1                          0.854
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.97 on 67 degrees of freedom
## Multiple R-squared:  0.739,  Adjusted R-squared:  0.673
## F-statistic: 11.2 on 17 and 67 DF,  p-value: 0.000000000000149
```

```
confint(sgrq_model)
```

```
##                               2.5 %     97.5 %
## (Intercept)             -12.3447206 88.0816907
## AGE                      -0.6797869  0.2705043
## gender1                  -4.3769043  8.3856755
## FEV1                    -14.9651335  2.0862725
## FVC                      -2.4784655  6.5488411
## CAT                       0.9692270  1.7776588
```

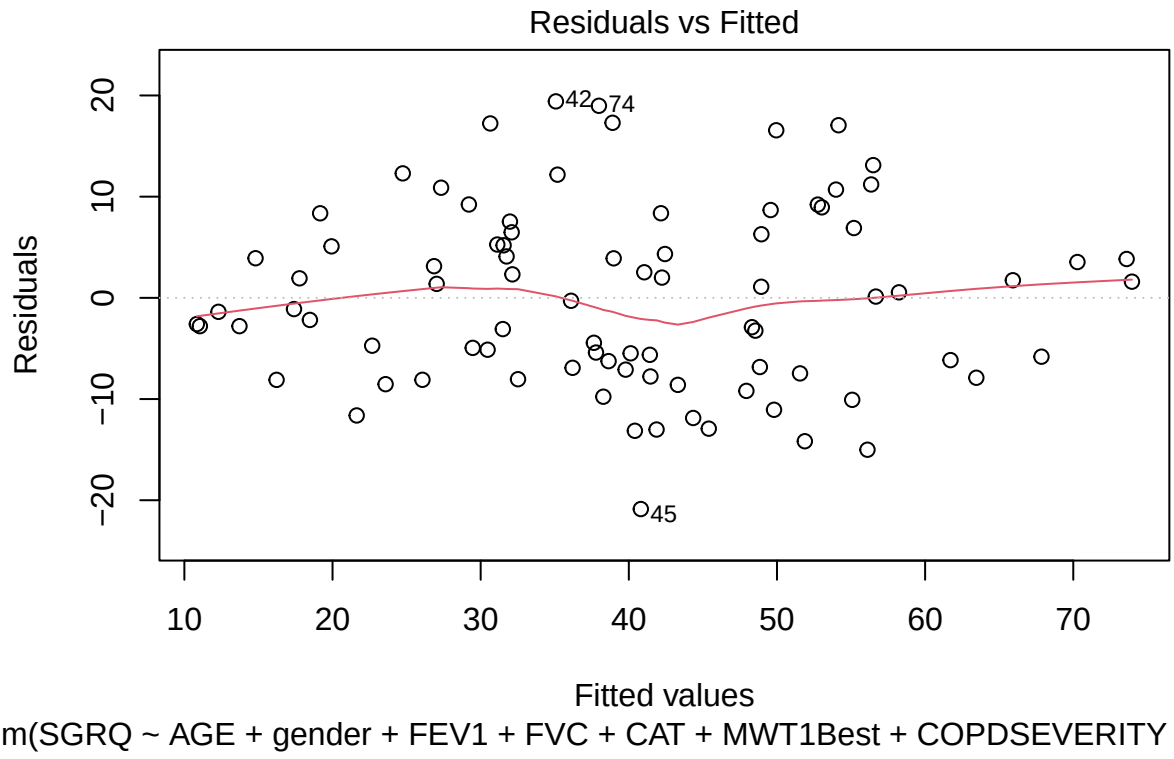
```
## MWT1Best                  -0.0586784  0.0217023
## COPDSEVERITYMODERATE      -6.4067067  8.0558696
## COPDSEVERITYSEVERE       -12.4009811 10.6208790
## COPDSEVERITYVERY SEVERE  -27.4288332  6.1114260
## HAD                        0.0577520  0.8825875
## comorbid1                 -8.8054117  9.5113601
## smoking2                  -9.6380098  3.2777810
## Diabetes1                 -6.9700148  8.7980734
## hypertension1             -4.1389448 15.2258667
## muscular1                -10.0759414  6.0771701
## AtrialFib1                -7.3884744  9.9710142
## IHD1                      -8.9385769 10.7683700
```

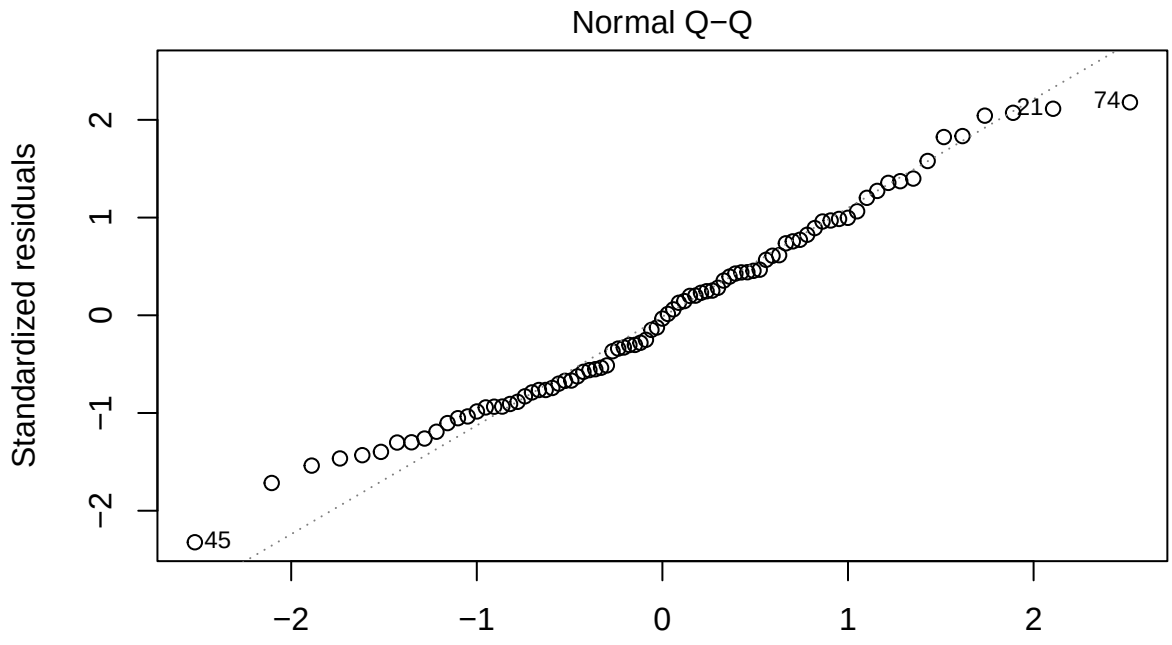
Fit the model :

```
predictedsgrqmodel1 <- predict(sgrq_model)
residualsgrqmodel1 <- residuals(sgrq_model)
```

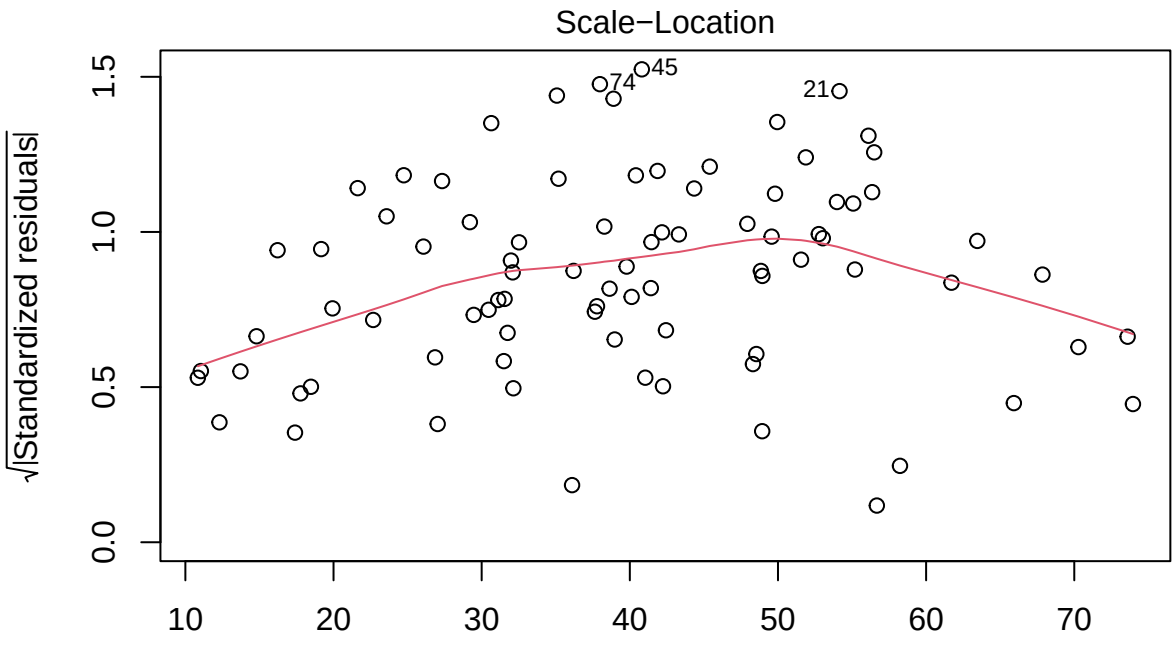
Check using plots :

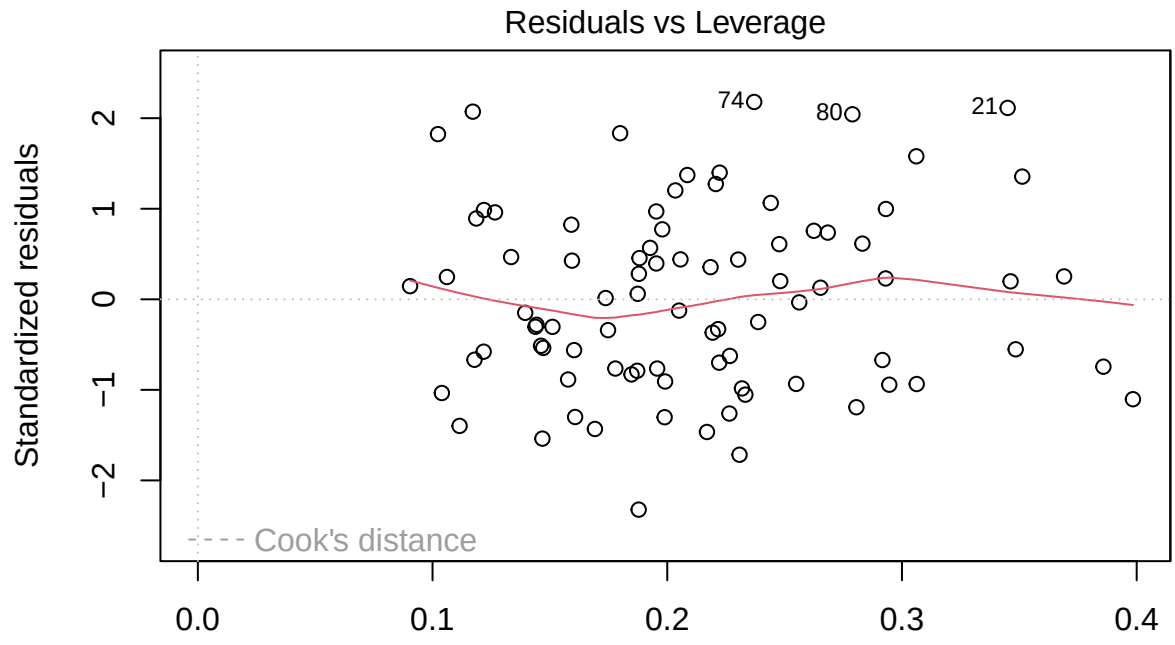
```
plot(sgrq_model)
```





m(SGRQ ~ AGE + gender + FEV1 + FVC + CAT + MWT1Best + COPDSEVERITY + HA


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
10
20
30
40
50
60
70
74
45
21
Fitted values
m(SGRQ ~ AGE + gender + FEV1 + FVC + CAT + MWT1Best + COPDSEVERITY + HA



m(SGRQ ~ AGE + gender + FEV1 + FVC + CAT + MWT1Best + COPDSEVERITY + HA

```
imcdiag(sgrq_model)
```

```
##
## Call:
## imcdiag(mod = sgrq_model)
##
##
## All Individual Multicollinearity Diagnostics Result
##
##                              VIF   TOL     Wi     Fi Leamer
## AGE                        1.710 0.585  3.017  3.265  0.765
## gender1                    2.025 0.494  4.355  4.714  0.703
## FEV1                       7.159 0.140 26.178 28.333  0.374
## FVC                        4.339 0.230 14.193 15.361  0.480
## CAT                        2.261 0.442  5.360  5.801  0.665
## MWT1Best                   3.461 0.289 10.460 11.321  0.538
## COPDSEVERITYMODERATE       2.774 0.360  7.540  8.160  0.600
## COPDSEVERITYSEVERE         5.455 0.183 18.933 20.492  0.428
## COPDSEVERITYVERY SEVERE    4.561 0.219 15.134 16.381  0.468
## HAD                        1.835 0.545  3.549  3.841  0.738
## comorbid1                  4.450 0.225 14.660 15.868  0.474
## smoking2                   1.231 0.812  0.983  1.064  0.901
## Diabetes1                  2.134 0.469  4.821  5.218  0.685
## hypertension1              1.905 0.525  3.845  4.161  0.725
## muscular1                  2.139 0.468  4.841  5.240  0.684
```

```
## AtrialFib1              2.806 0.356  7.676  8.308  0.597
## IHD1                    1.575 0.635  2.442  2.643  0.797
##                           CVIF Klein  IND1  IND2
## AGE                     -0.512     0 0.138 0.704
## gender1                 -0.606     0 0.116 0.858
## FEV1                    -2.144     1 0.033 1.459
## FVC                     -1.299     1 0.054 1.305
## CAT                     -0.677     0 0.104 0.946
## MWT1Best                -1.036     0 0.068 1.206
## COPDSEVERITYMODERATE    -0.831     0 0.085 1.085
## COPDSEVERITYSEVERE      -1.633     1 0.043 1.385
## COPDSEVERITYVERY SEVERE -1.366     1 0.052 1.324
## HAD                     -0.549     0 0.128 0.772
## comorbid1               -1.332     1 0.053 1.315
## smoking2                -0.369     0 0.191 0.319
## Diabetes1               -0.639     0 0.110 0.901
## hypertension1           -0.570     0 0.124 0.806
## muscular1               -0.640     0 0.110 0.903
## AtrialFib1              -0.840     0 0.084 1.092
## IHD1                    -0.471     0 0.149 0.619
##
## 1 --> COLLINEARITY is detected by the test
## 0 --> COLLINEARITY is not detected by the test
##
## AGE , gender1 , FEV1 , FVC , MWT1Best , COPDSEVERITYMODERATE , COPDSEVERITYSEVERE , COPDSEVERITYVERY
##
## R-square of y on all x: 0.739
##
## * use method argument to check which regressors may be the reason of collinearity
## ===================================
```

```
imcdiag(sgrq_model, method = "VIF")
```

```
##
## Call:
## imcdiag(mod = sgrq_model, method = "VIF")
##
##
##  VIF Multicollinearity Diagnostics
##
##                           VIF detection
## AGE                     1.710         0
## gender1                 2.025         0
## FEV1                    7.159         0
## FVC                     4.339         0
## CAT                     2.261         0
## MWT1Best                3.461         0
## COPDSEVERITYMODERATE    2.774         0
## COPDSEVERITYSEVERE      5.455         0
## COPDSEVERITYVERY SEVERE 4.561         0
## HAD                     1.835         0
## comorbid1               4.450         0
## smoking2                1.231         0
## Diabetes1               2.134         0
```

```
## hypertension1           1.905         0
## muscular1               2.139         0
## AtrialFib1              2.806         0
## IHD1                    1.575         0
##
## NOTE:  VIF Method Failed to detect multicollinearity
##
##
## 0 --> COLLINEARITY is not detected by the test
##
## ===================================
```

1. FEV1, FVC, CAT According to literature, these thre variables are assessment measured in patient with COPD. So I start the model with these variables.

```
sgrq_model_1 <- lm(SGRQ~FEV1+FVC+CAT, data=subset_copd)
```

```
summary(sgrq_model_1)
```

```
##
## Call:
## lm(formula = SGRQ ~ FEV1 + FVC + CAT, data = subset_copd)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -24.71  -6.99  -1.44   8.65  19.52
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)   14.391      5.018    2.87              0.0053
## FEV1          -6.178      3.038   -2.03              0.0453
## FVC            2.255      2.026    1.11              0.2689
## CAT            1.623      0.149   10.89 <0.0000000000000002
##
## (Intercept) **
## FEV1        *
## FVC
## CAT         ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.4 on 81 degrees of freedom
## Multiple R-squared:  0.656,  Adjusted R-squared:  0.643
## F-statistic: 51.4 on 3 and 81 DF,  p-value: <0.0000000000000002
```

Rsquared = 0.668 F(3,92) = 40.3 p-value = <0.0001

```
confint(sgrq_model_1)
```

```
##                 2.5 %    97.5 %
## (Intercept)   4.40645 24.376090
```

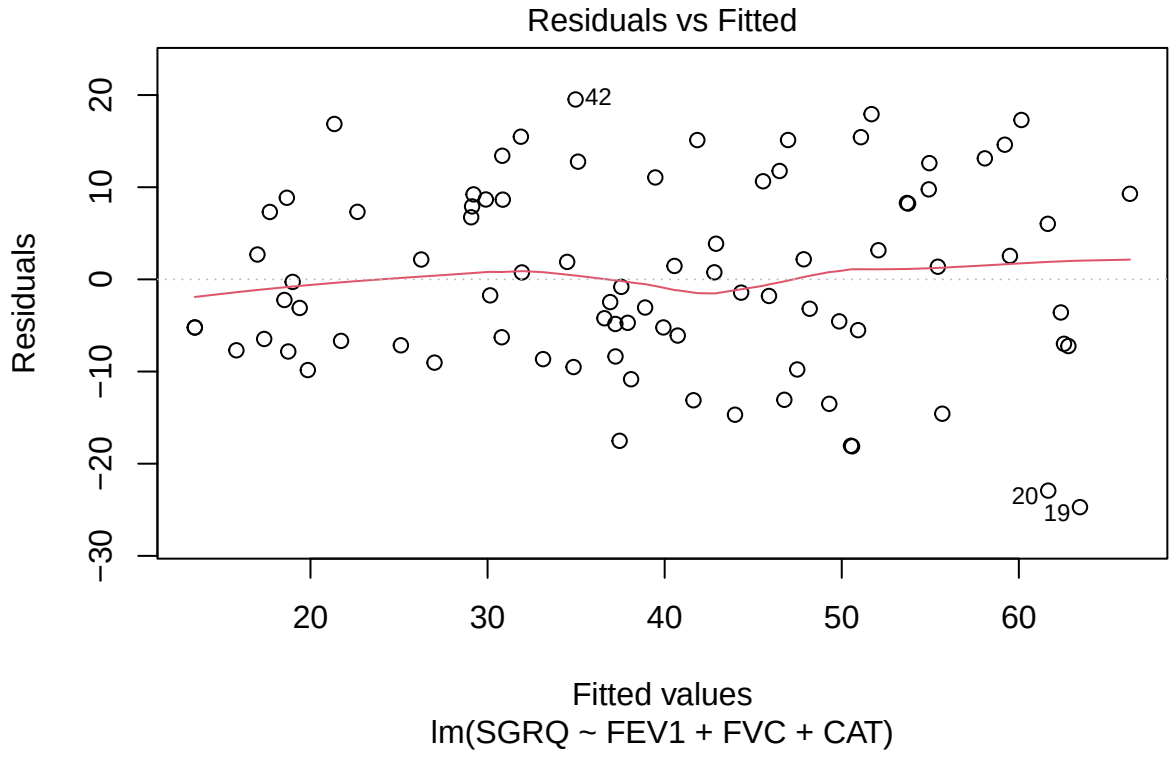
```
## FEV1        -12.22146 -0.133797
## FVC          -1.77508  6.285394
## CAT           1.32621  1.919421
```

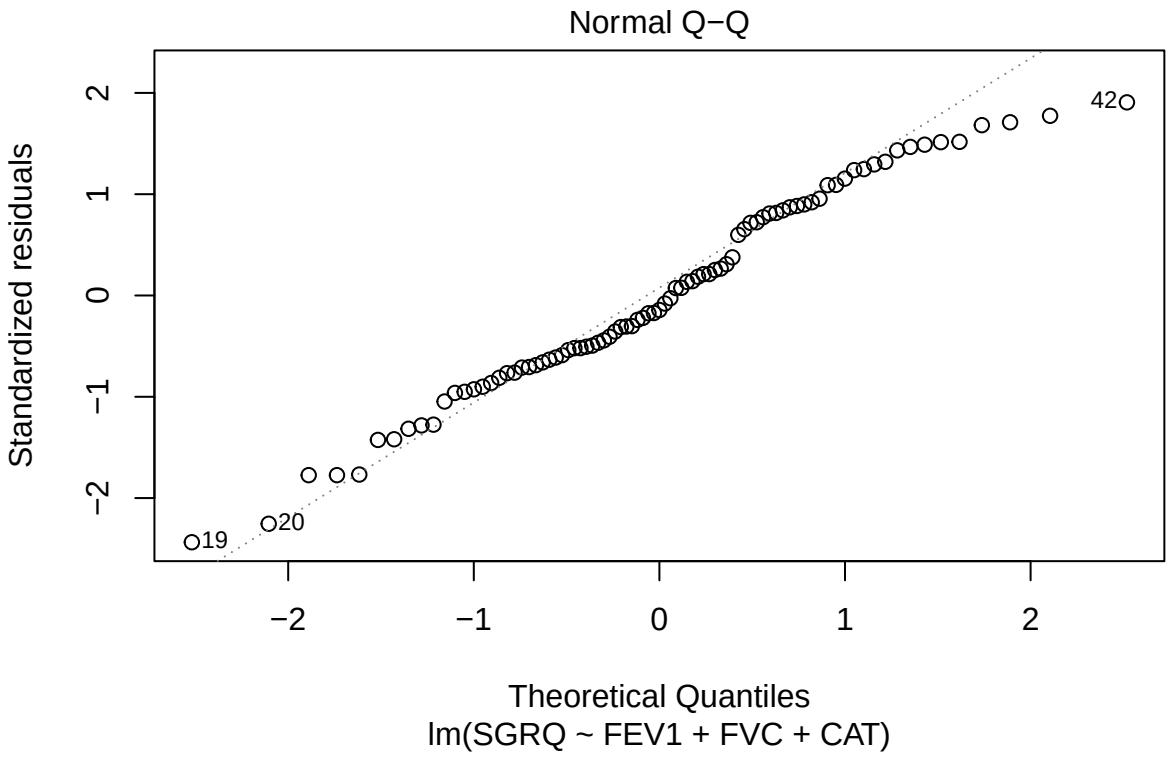
Fit the model :

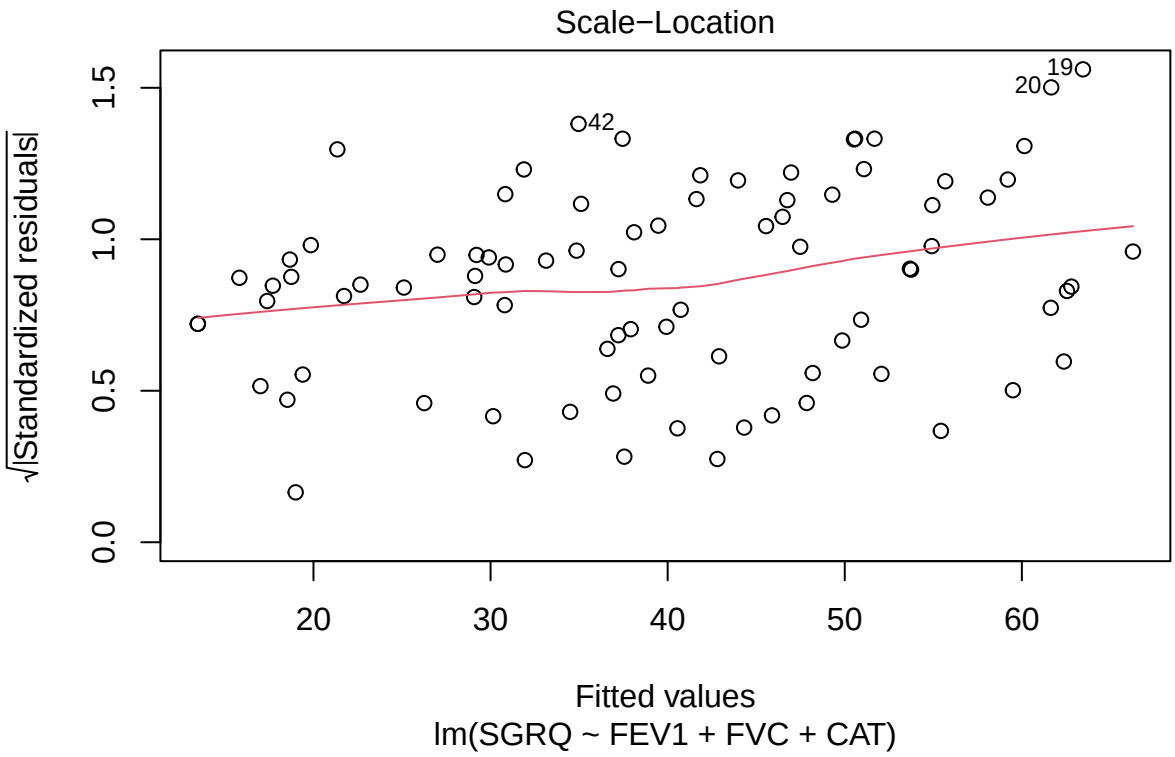
```
predictedsgrqmodel1 <- predict(sgrq_model_1)
residualsgrqmodel1 <- residuals(sgrq_model_1)
```

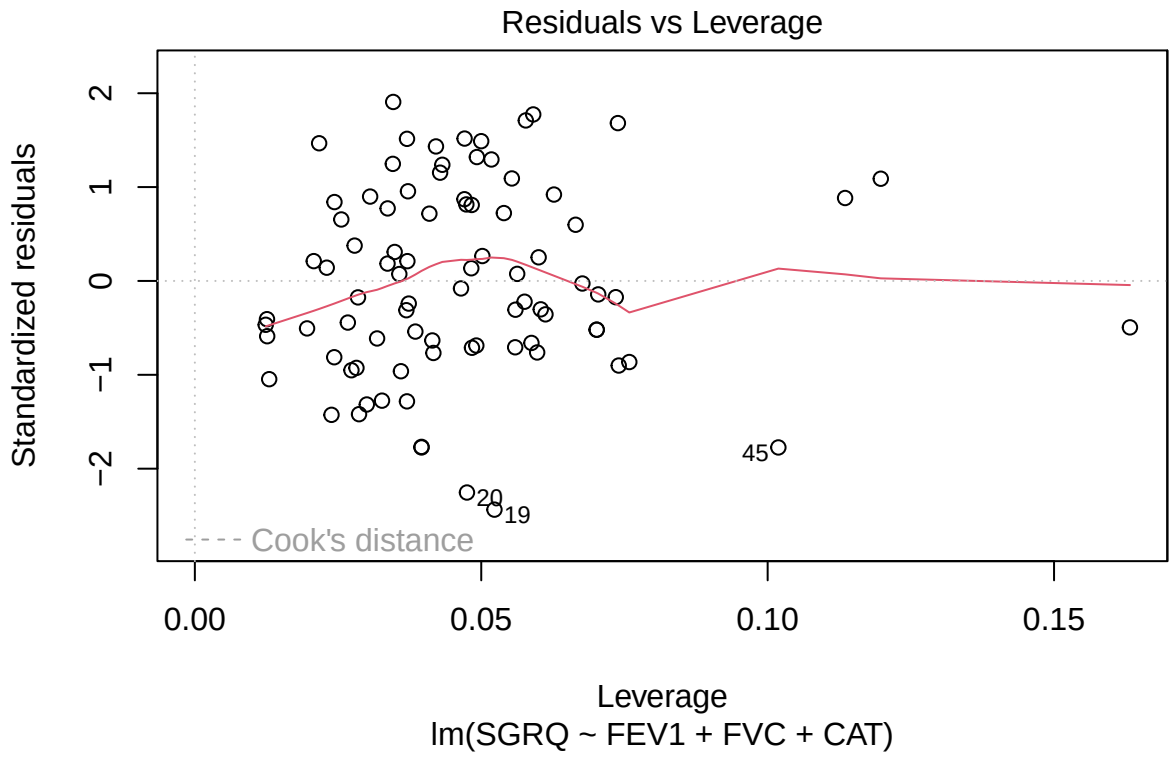
Check using plots :

```
plot(sgrq_model_1)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(SGRQ ~ FEV1 + FVC + CAT)
19
20
42
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
20
30
40
50
60
42
20
19
Fitted values
lm(SGRQ ~ FEV1 + FVC + CAT)



2. Let's include AGE, gender,comorbid and COPD Severity

```
sgrq_model_2 <- lm(SGRQ~AGE+gender+COPDSEVERITY+FEV1+FVC+CAT+MWT1Best+comorbid, data=subset_copd)
```

```
summary(sgrq_model_2)
```

```
##
## Call:
## lm(formula = SGRQ ~ AGE + gender + COPDSEVERITY + FEV1 + FVC +
##     CAT + MWT1Best + comorbid, data = subset_copd)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -17.57  -6.16  -1.20   7.50  23.77
##
## Coefficients:
##                            Estimate Std. Error t value
## (Intercept)                 60.6286    21.9473    2.76
## AGE                         -0.4299     0.2161   -1.99
## gender1                      2.2468     3.1425    0.71
## COPDSEVERITYMODERATE         1.4294     3.6178    0.40
## COPDSEVERITYSEVERE          -0.6210     5.6100   -0.11
## COPDSEVERITYVERY SEVERE    -11.0387     8.1567   -1.35
## FEV1                        -6.7766     4.1923   -1.62
## FVC                          2.0568     2.2272    0.92
## CAT                          1.5052     0.1708    8.81
```

```
## MWT1Best                  -0.0347     0.0155   -2.24
## comorbid1                  1.4695     2.4579    0.60
##                                  Pr(>|t|)
## (Intercept)                        0.0072 **
## AGE                                0.0504 .
## gender1                            0.4769
## COPDSEVERITYMODERATE               0.6939
## COPDSEVERITYSEVERE                 0.9122
## COPDSEVERITYVERY SEVERE            0.1801
## FEV1                               0.1103
## FVC                                0.3588
## CAT                      0.00000000000038 ***
## MWT1Best                           0.0284 *
## comorbid1                          0.5518
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.1 on 74 degrees of freedom
## Multiple R-squared:  0.702,  Adjusted R-squared:  0.662
## F-statistic: 17.4 on 10 and 74 DF,  p-value: 0.000000000000000906
```

Rsquared = 0.702 F(3,92) = 17.4 p-value = <0.00001

```
confint(sgrq_model_2)
```

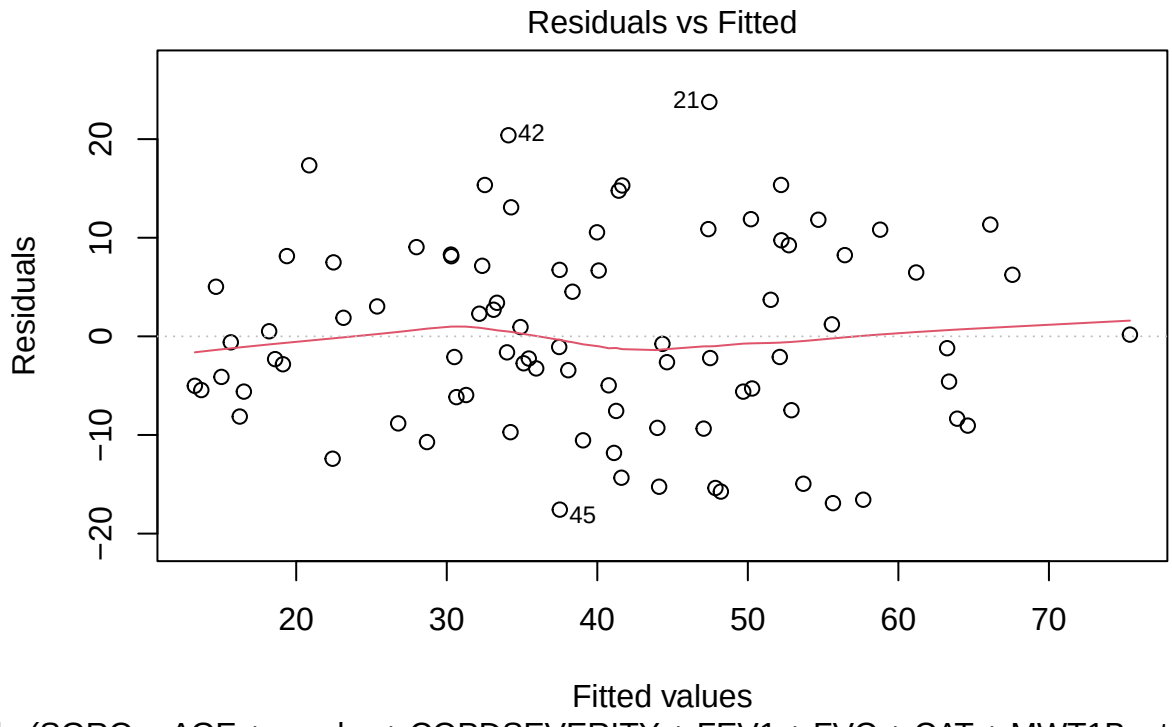
```
##                                2.5 %        97.5 %
## (Intercept)              16.8976250 104.359636486
## AGE                      -0.8605597   0.000699642
## gender1                  -4.0146393   8.508324620
## COPDSEVERITYMODERATE     -5.7791890   8.637930567
## COPDSEVERITYSEVERE      -11.7992583  10.557229631
## COPDSEVERITYVERY SEVERE -27.2913059   5.213957357
## FEV1                    -15.1299014   1.576778256
## FVC                      -2.3811110   6.494638601
## CAT                       1.1648738   1.845536109
## MWT1Best                 -0.0657196  -0.003771360
## comorbid1                -3.4280760   6.367043255
```

Fit the model :

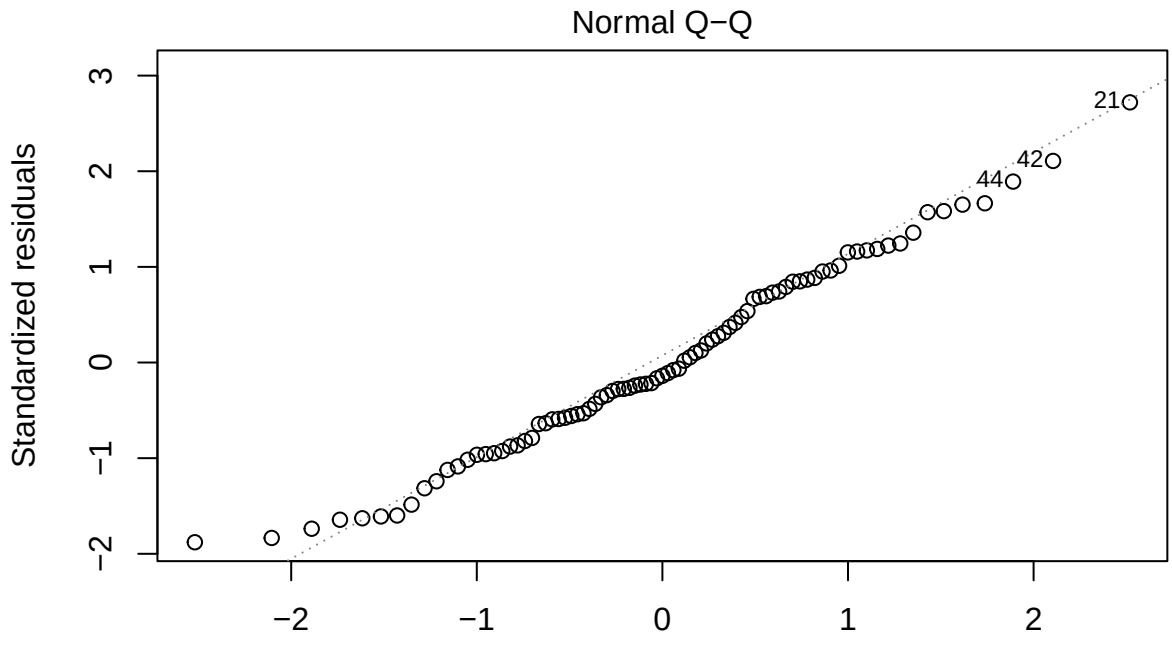
```
predictedsgrqmodel2 <- predict(sgrq_model_2)
residualsgrqmodel2 <- residuals(sgrq_model_2)
```

Check using plots :

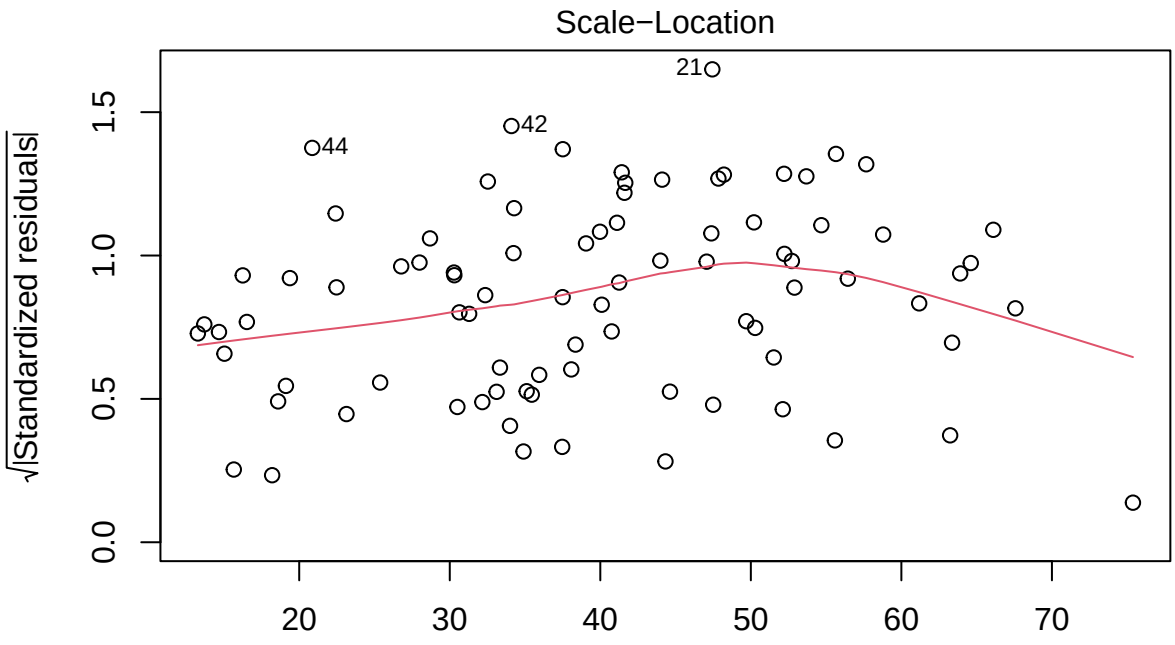
```
plot(sgrq_model_2)
```



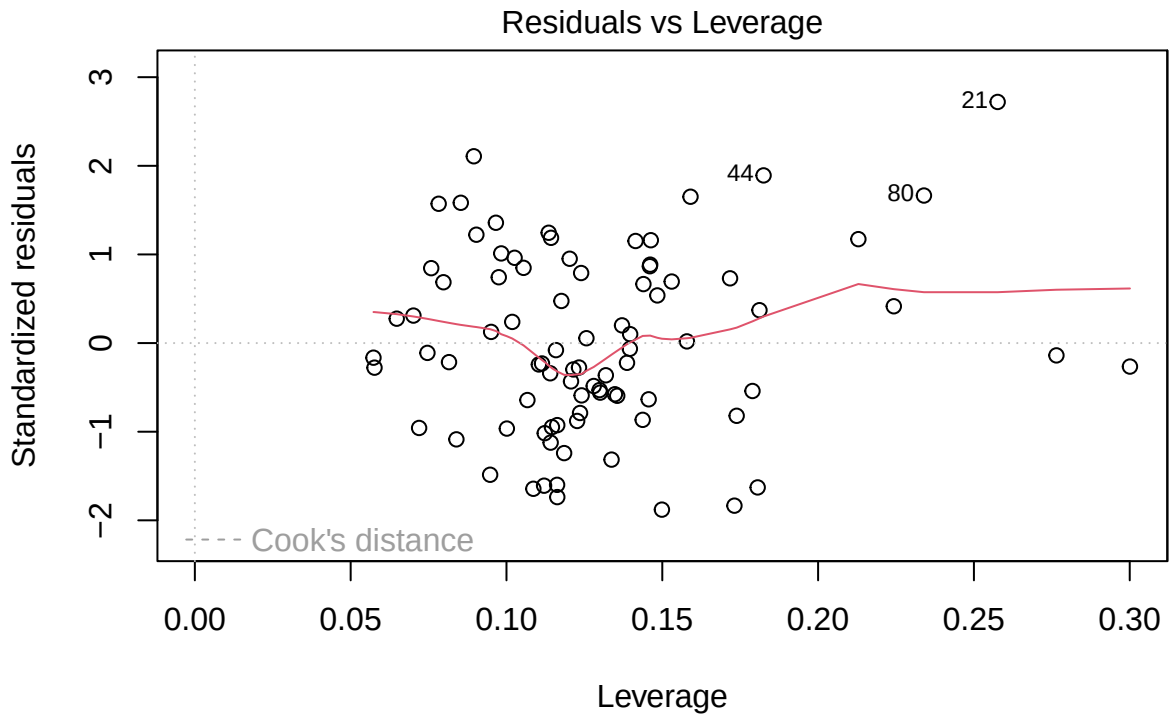
m(SGRQ ~ AGE + gender + COPDSEVERITY + FEV1 + FVC + CAT + MWT1Best + cor



m(SGRQ ~ AGE + gender + COPDSEVERITY + FEV1 + FVC + CAT + MWT1Best + cor



m(SGRQ ~ AGE + gender + COPDSEVERITY + FEV1 + FVC + CAT + MWT1Best + cor



Checking if there's any collinearity in the model

```
imcdiag(sgrq_model_2)
```

```
##
## Call:
## imcdiag(mod = sgrq_model_2)
##
##
## All Individual Multicollinearity Diagnostics Result
##
##                              VIF   TOL     Wi     Fi Leamer
## AGE                        1.363 0.734  3.022  3.445  0.857
## gender1                    1.891 0.529  7.428  8.468  0.727
## COPDSEVERITYMODERATE       2.675 0.374 13.954 15.908  0.611
## COPDSEVERITYSEVERE         4.991 0.200 33.257 37.913  0.448
## COPDSEVERITYVERY SEVERE    4.156 0.241 26.302 29.985  0.491
## FEV1                       6.668 0.150 47.235 53.848  0.387
## FVC                        4.070 0.246 25.584 29.165  0.496
## CAT                        1.555 0.643  4.626  5.274  0.802
## MWT1Best                   1.995 0.501  8.288  9.448  0.708
## comorbid1                  1.235 0.810  1.955  2.228  0.900
##                             CVIF Klein  IND1  IND2
## AGE                       -1.121     0 0.088 0.478
## gender1                   -1.555     0 0.063 0.846
## COPDSEVERITYMODERATE      -2.199     0 0.045 1.124
```

```
## COPDSEVERITYSEVERE      -4.104     1 0.024 1.435
## COPDSEVERITYVERY SEVERE -3.418     1 0.029 1.363
## FEV1                    -5.483     1 0.018 1.525
## FVC                     -3.347     1 0.029 1.354
## CAT                     -1.279     0 0.077 0.641
## MWT1Best                -1.640     0 0.060 0.895
## comorbid1               -1.015     0 0.097 0.341
##
## 1 --> COLLINEARITY is detected by the test
## 0 --> COLLINEARITY is not detected by the test
##
## AGE , gender1 , COPDSEVERITYMODERATE , COPDSEVERITYSEVERE , COPDSEVERITYVERY SEVERE , FEV1 , FVC , c
##
## R-square of y on all x: 0.702
##
## * use method argument to check which regressors may be the reason of collinearity
## ===================================
```

```
imcdiag(sgrq_model_2, method="VIF")
```

```
##
## Call:
## imcdiag(mod = sgrq_model_2, method = "VIF")
##
##
##  VIF Multicollinearity Diagnostics
##
##                           VIF detection
## AGE                     1.363         0
## gender1                 1.891         0
## COPDSEVERITYMODERATE    2.675         0
## COPDSEVERITYSEVERE      4.991         0
## COPDSEVERITYVERY SEVERE 4.156         0
## FEV1                    6.668         0
## FVC                     4.070         0
## CAT                     1.555         0
## MWT1Best                1.995         0
## comorbid1               1.235         0
##
## NOTE:  VIF Method Failed to detect multicollinearity
##
##
## 0 --> COLLINEARITY is not detected by the test
##
## ===================================
```

```
imcdiag(sgrq_model_2, method="TOL")
```

```
##
## Call:
## imcdiag(mod = sgrq_model_2, method = "TOL")
##
##
```

```
##  TOL Multicollinearity Diagnostics
##
##                            TOL detection
## AGE                      0.734         0
## gender1                  0.529         0
## COPDSEVERITYMODERATE     0.374         0
## COPDSEVERITYSEVERE       0.200         0
## COPDSEVERITYVERY SEVERE 0.241         0
## FEV1                     0.150         0
## FVC                      0.246         0
## CAT                      0.643         0
## MWT1Best                 0.501         0
## comorbid1                0.810         0
##
## NOTE:  TOL Method Failed to detect multicollinearity
##
##
## 0 --> COLLINEARITY is not detected by the test
##
## ===================================
```

```
imcdiag(sgrq_model_2, method="Wi")
```

```
##
## Call:
## imcdiag(mod = sgrq_model_2, method = "Wi")
##
##
##  Wi Multicollinearity Diagnostics
##
##                               Wi detection
## AGE                        3.022         1
## gender1                    7.428         1
## COPDSEVERITYMODERATE      13.954         1
## COPDSEVERITYSEVERE        33.257         1
## COPDSEVERITYVERY SEVERE 26.302         1
## FEV1                      47.235         1
## FVC                       25.584         1
## CAT                        4.626         1
## MWT1Best                   8.288         1
## comorbid1                  1.955         0
##
## Multicollinearity may be due to AGE gender1 COPDSEVERITYMODERATE COPDSEVERITYSEVERE COPDSEVERITYVERY
##
## 1 --> COLLINEARITY is detected by the test
## 0 --> COLLINEARITY is not detected by the test
##
## ===================================
```

```
imcdiag(sgrq_model_2, method="Leamer")
```

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

```
## imcdiag(mod = sgrq_model_2, method = "Leamer")
##
##
##  Leamer Multicollinearity Diagnostics
##
##                          Leamer detection
## AGE                       0.857         0
## gender1                   0.727         0
## COPDSEVERITYMODERATE      0.611         0
## COPDSEVERITYSEVERE        0.448         0
## COPDSEVERITYVERY SEVERE   0.491         0
## FEV1                      0.387         0
## FVC                       0.496         0
## CAT                       0.802         0
## MWT1Best                  0.708         0
## comorbid1                 0.900         0
##
## NOTE:  Leamer Method Failed to detect multicollinearity
##
##
## 0 --> COLLINEARITY is not detected by the test
##
## ===================================
```

```
imcdiag(sgrq_model_2, method="CVIF")
```

```
##
## Call:
## imcdiag(mod = sgrq_model_2, method = "CVIF")
##
##
##  CVIF Multicollinearity Diagnostics
##
##                            CVIF detection
## AGE                      -1.121         0
## gender1                  -1.555         0
## COPDSEVERITYMODERATE     -2.199         0
## COPDSEVERITYSEVERE       -4.104         0
## COPDSEVERITYVERY SEVERE -3.418         0
## FEV1                     -5.483         0
## FVC                      -3.347         0
## CAT                      -1.279         0
## MWT1Best                 -1.640         0
## comorbid1                -1.015         0
##
## NOTE:  CVIF Method Failed to detect multicollinearity
##
##
## 0 --> COLLINEARITY is not detected by the test
##
## ===================================
```

```r
imcdiag(sgrq_model_2, method="Klein")
```

```
##
## Call:
## imcdiag(mod = sgrq_model_2, method = "Klein")
##
##
##  Klein Multicollinearity Diagnostics
##
##                            R2j R2(overall) Difference
## AGE                      0.266       0.702     -0.436
## gender1                  0.471       0.702     -0.231
## COPDSEVERITYMODERATE     0.626       0.702     -0.076
## COPDSEVERITYSEVERE       0.800       0.702      0.098
## COPDSEVERITYVERY SEVERE 0.759       0.702      0.057
## FEV1                     0.850       0.702      0.148
## FVC                      0.754       0.702      0.052
## CAT                      0.357       0.702     -0.345
## MWT1Best                 0.499       0.702     -0.203
## comorbid1                0.190       0.702     -0.512
##                          detection
## AGE                              0
## gender1                          0
## COPDSEVERITYMODERATE             0
## COPDSEVERITYSEVERE               1
## COPDSEVERITYVERY SEVERE          1
## FEV1                             1
## FVC                              1
## CAT                              0
## MWT1Best                         0
## comorbid1                        0
##
## Multicollinearity may be due to COPDSEVERITYSEVERE COPDSEVERITYVERY SEVERE FEV1 FVC regressors
##
##
## 1 --> COLLINEARITY is detected by the test
## 0 --> COLLINEARITY is not detected by the test
##
## ===================================
```

```r
imcdiag(sgrq_model_2, method="IND1")
```

```
##
## Call:
## imcdiag(mod = sgrq_model_2, method = "IND1")
##
##
##  IND1 Multicollinearity Diagnostics
##
##                           IND1 detection
## AGE                      0.088         0
## gender1                  0.063         0
```

```
## COPDSEVERITYMODERATE    0.045         0
## COPDSEVERITYSEVERE      0.024         0
## COPDSEVERITYVERY SEVERE 0.029         0
## FEV1                    0.018         1
## FVC                     0.029         0
## CAT                     0.077         0
## MWT1Best                0.060         0
## comorbid1               0.097         0
##
## Multicollinearity may be due to FEV1 regressors
##
## 1 --> COLLINEARITY is detected by the test
## 0 --> COLLINEARITY is not detected by the test
##
## ====================================
```

```
imcdiag(sgrq_model_2, method="IND2")
```

```
##
## Call:
## imcdiag(mod = sgrq_model_2, method = "IND2")
##
##
##  IND2 Multicollinearity Diagnostics
##
##                          IND2 detection
## AGE                     0.478         0
## gender1                 0.846         0
## COPDSEVERITYMODERATE    1.124         0
## COPDSEVERITYSEVERE      1.435         0
## COPDSEVERITYVERY SEVERE 1.363         0
## FEV1                    1.525         0
## FVC                     1.354         0
## CAT                     0.641         0
## MWT1Best                0.895         0
## comorbid1               0.341         0
##
## NOTE:  IND2 Method Failed to detect multicollinearity
##
##
## 0 --> COLLINEARITY is not detected by the test
##
## ====================================
```

Multicollinearity is detected in COPDSEVERITY, FEV1 and FVC variables

3. Check if one of those variables are removed

```
sgrq_model_3 <- lm(SGRQ~AGE+gender+FEV1+FVC+CAT+comorbid, data=subset_copd)
```

```
summary(sgrq_model_3)
```

```
##
## Call:
## lm(formula = SGRQ ~ AGE + gender + FEV1 + FVC + CAT + comorbid,
##     data = subset_copd)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -22.40  -6.83  -1.35   8.19  21.89
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)   25.765     15.430    1.67               0.099
## AGE           -0.188      0.199   -0.94               0.348
## gender1       -0.117      2.699   -0.04               0.965
## FEV1          -6.856      3.103   -2.21               0.030
## FVC            2.784      2.210    1.26               0.211
## CAT            1.602      0.150   10.72 <0.0000000000000002
## comorbid1      3.491      2.334    1.50               0.139
##
## (Intercept) .
## AGE
## gender1
## FEV1        *
## FVC
## CAT         ***
## comorbid1
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.4 on 78 degrees of freedom
## Multiple R-squared:  0.669,  Adjusted R-squared:  0.644
## F-statistic: 26.3 on 6 and 78 DF,  p-value: <0.0000000000000002
```

```
confint(sgrq_model_3)
```

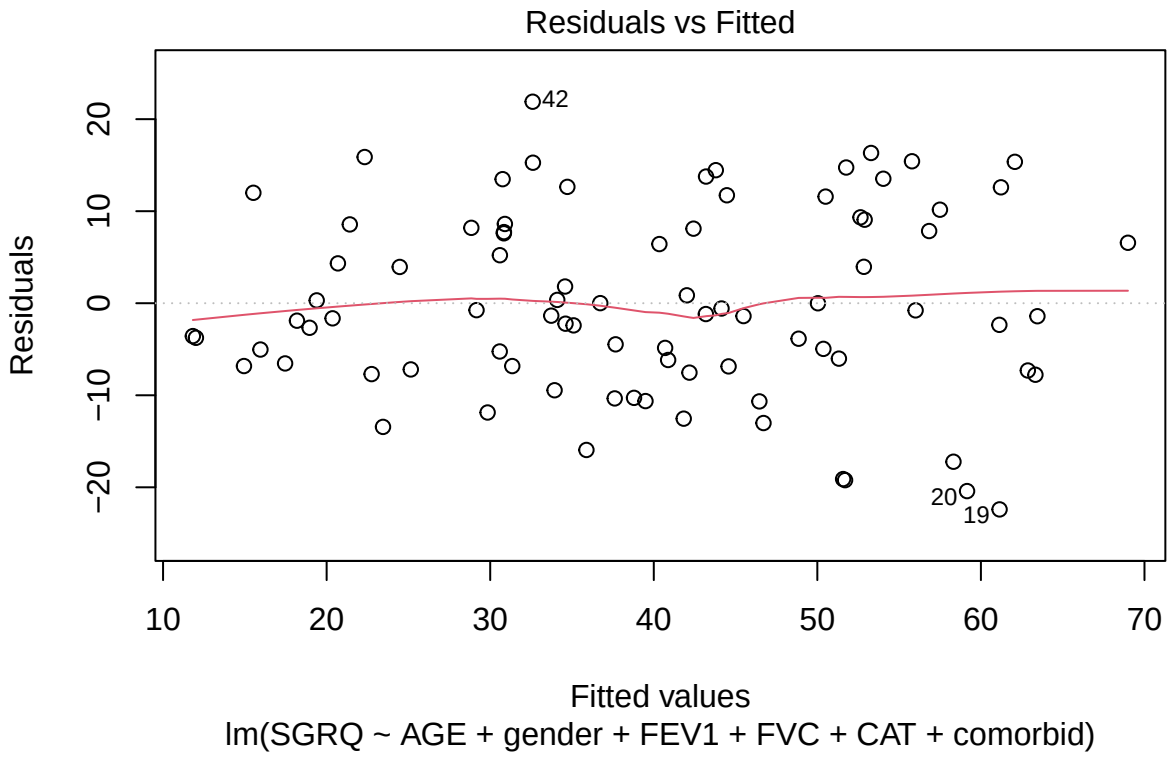
```
##                  2.5 %    97.5 %
## (Intercept)  -4.954850 56.484107
## AGE          -0.582922  0.207803
## gender1      -5.490303  5.255569
## FEV1        -13.033099 -0.679793
## FVC          -1.615397  7.183283
## CAT           1.304783  1.900099
## comorbid1    -1.155104  8.136942
```

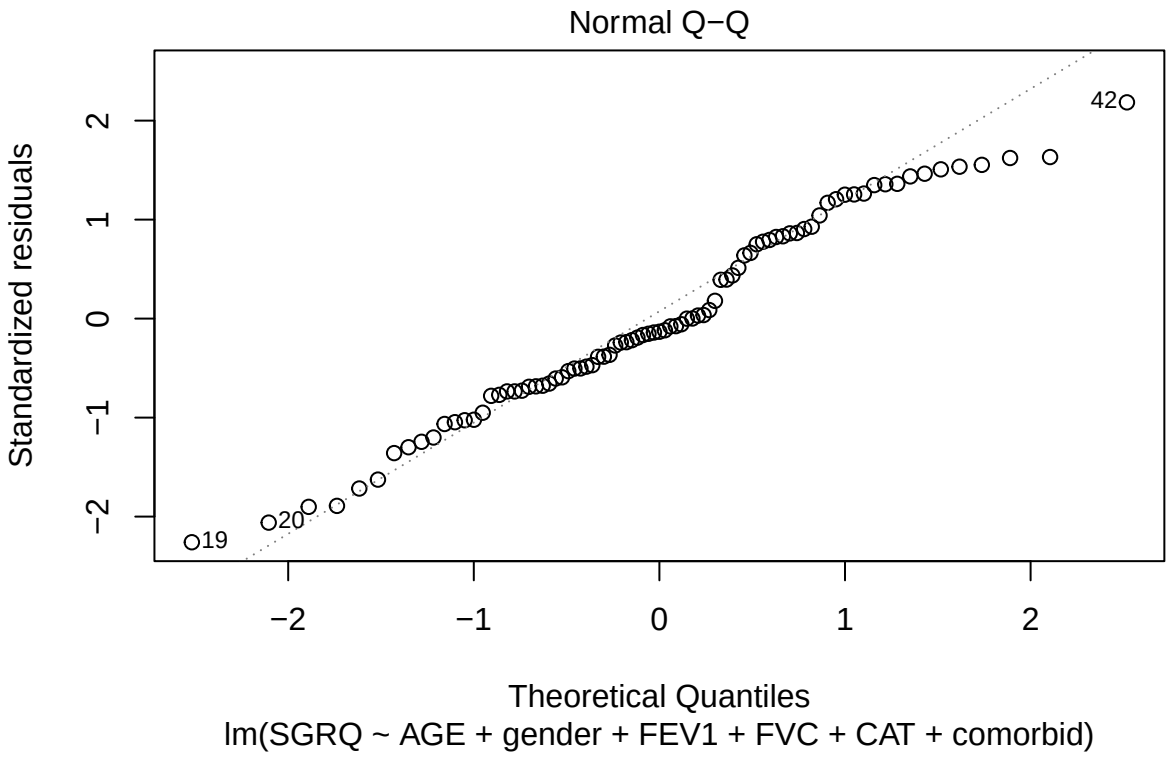
Fit the model :

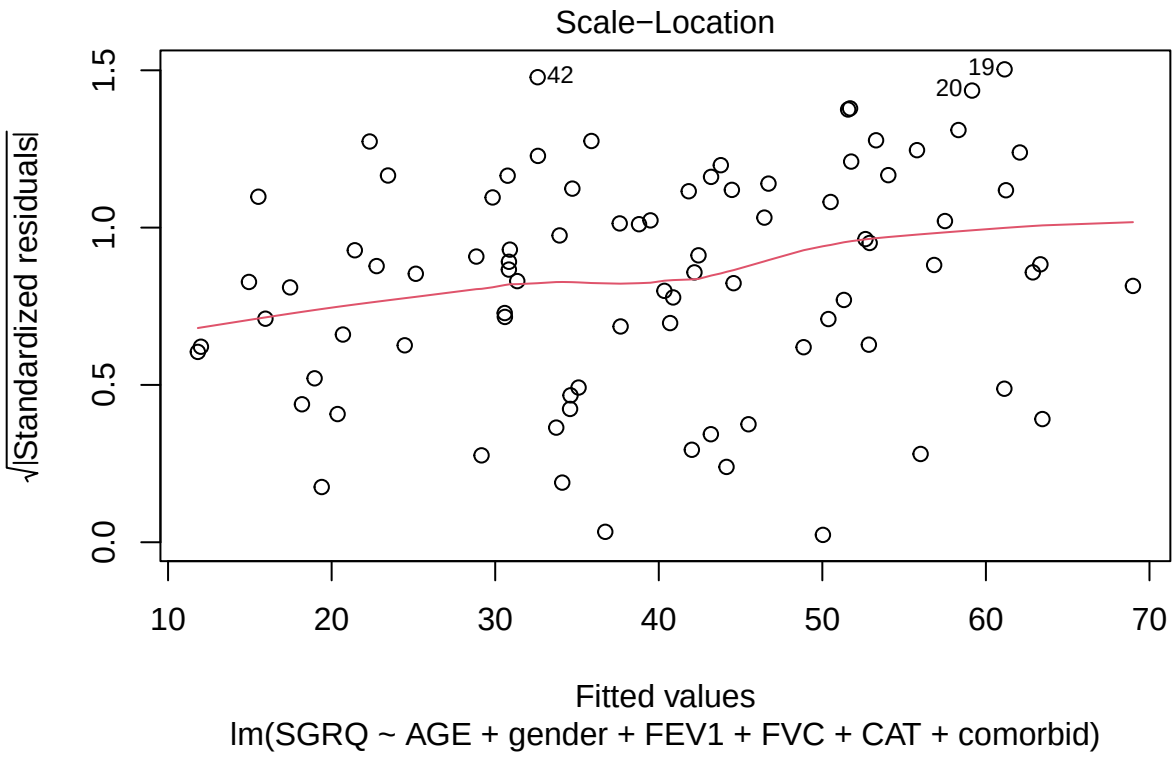
```
predictedsgrqmodel3 <- predict(sgrq_model_3)
residualsgrqmodel3 <- residuals(sgrq_model_3)
```

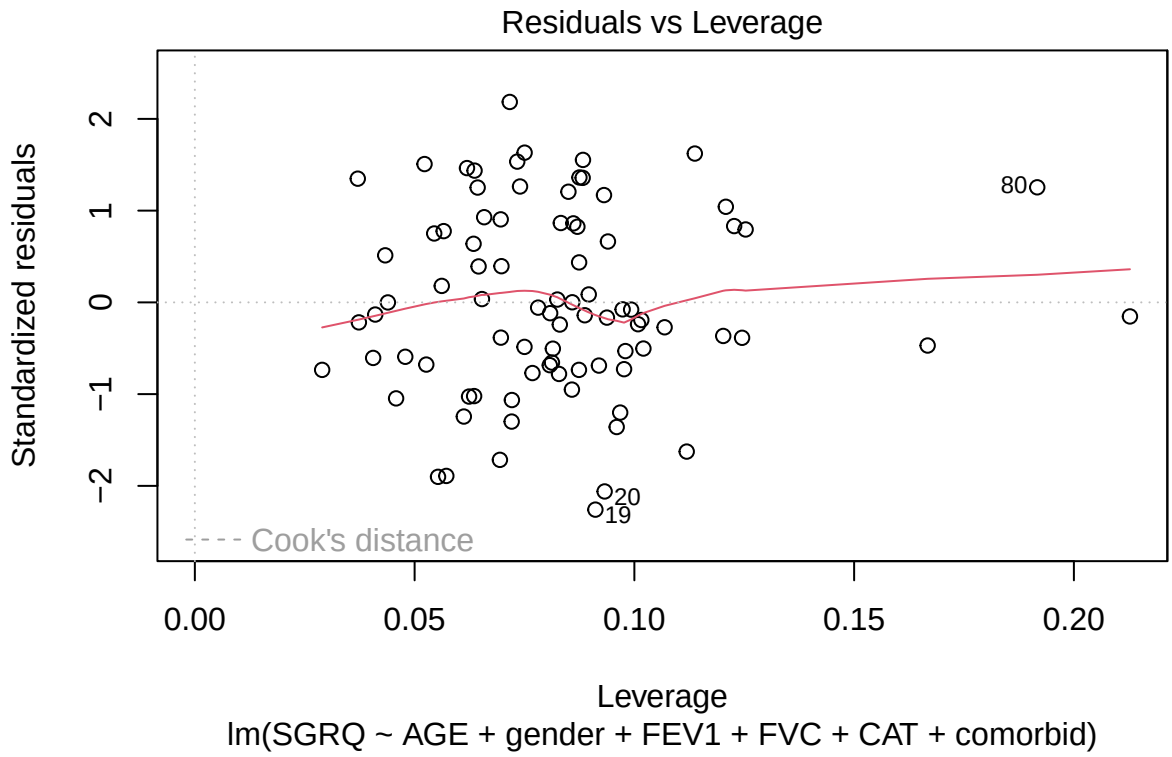
Check using plots :

```
plot(sgrq_model_3)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(SGRQ ~ AGE + gender + FEV1 + FVC + CAT + comorbid)
19
20
42


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
42
19
20
10
20
30
40
50
60
70
Fitted values
lm(SGRQ ~ AGE + gender + FEV1 + FVC + CAT + comorbid)




```
sgrq_model_4 <- lm(SGRQ~AGE+gender+COPDSEVERITY+FVC+CAT+comorbid, data=subset_copd)
```

```
summary(sgrq_model_4)
```

```
##
## Call:
## lm(formula = SGRQ ~ AGE + gender + COPDSEVERITY + FVC + CAT +
##     comorbid, data = subset_copd)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -20.43  -7.50  -1.16   8.34  23.35
##
## Coefficients:
##                           Estimate Std. Error t value
## (Intercept)                 23.576     18.003    1.31
## AGE                         -0.251      0.213   -1.18
## gender1                     -0.418      3.109   -0.13
## COPDSEVERITYMODERATE         5.247      3.341    1.57
## COPDSEVERITYSEVERE           7.288      4.291    1.70
## COPDSEVERITYVERY SEVERE      2.247      6.652    0.34
## FVC                         -0.112      1.811   -0.06
## CAT                          1.643      0.162   10.16
## comorbid1                    2.446      2.384    1.03
##                                    Pr(>|t|)
```

```
## (Intercept)                             0.194
## AGE                                     0.244
## gender1                                 0.894
## COPDSEVERITYMODERATE                    0.120
## COPDSEVERITYSEVERE                      0.094 .
## COPDSEVERITYVERY SEVERE                 0.736
## FVC                                     0.951
## CAT                      0.00000000000000084 ***
## comorbid1                               0.308
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.6 on 76 degrees of freedom
## Multiple R-squared:  0.668,  Adjusted R-squared:  0.633
## F-statistic: 19.1 on 8 and 76 DF,  p-value: 0.00000000000000203
```

```
confint(sgrq_model_4)
```

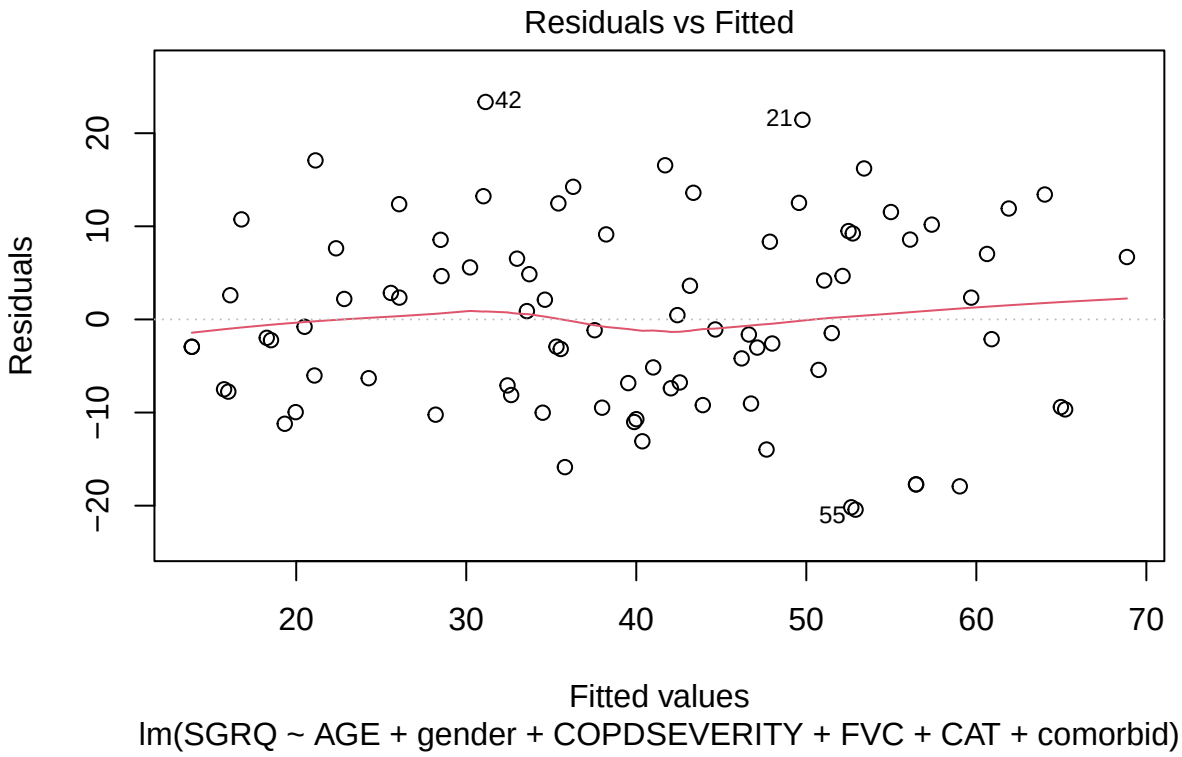
```
##                                2.5 %    97.5 %
## (Intercept)              -12.279520 59.430724
## AGE                       -0.675111  0.174056
## gender1                   -6.609519  5.774346
## COPDSEVERITYMODERATE      -1.406327 11.900048
## COPDSEVERITYSEVERE        -1.258009 15.834631
## COPDSEVERITYVERY SEVERE -11.001494 15.496188
## FVC                       -3.719664  3.495481
## CAT                        1.320827  1.965233
## comorbid1                 -2.301307  7.193099
```

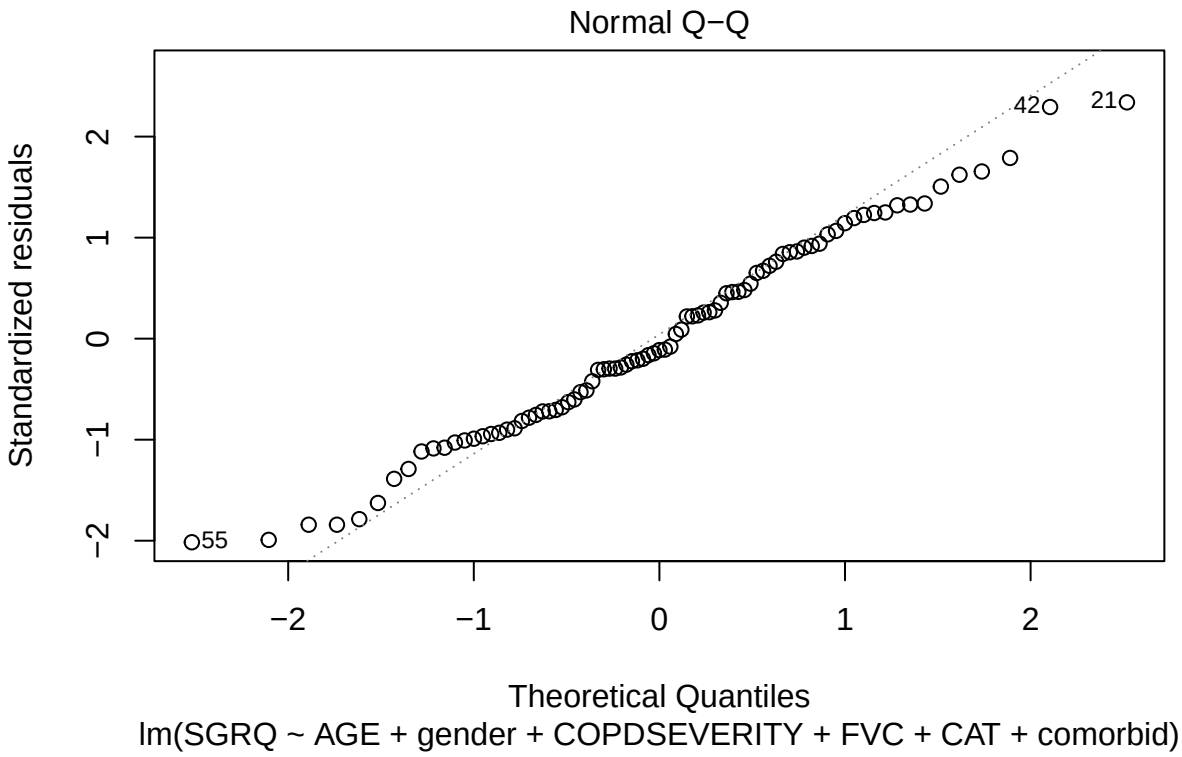
Fit the model :

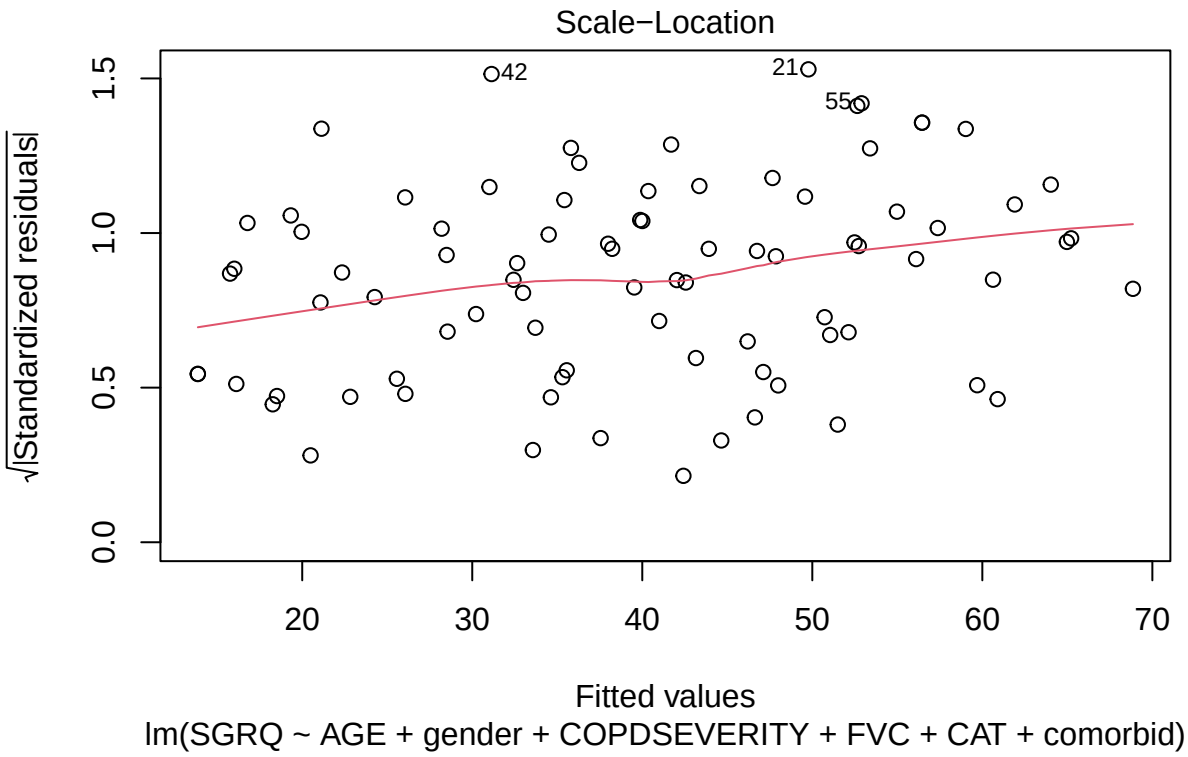
```
predictedsgrqmodel4 <- predict(sgrq_model_4)
residualsgrqmodel4 <- residuals(sgrq_model_4)
```

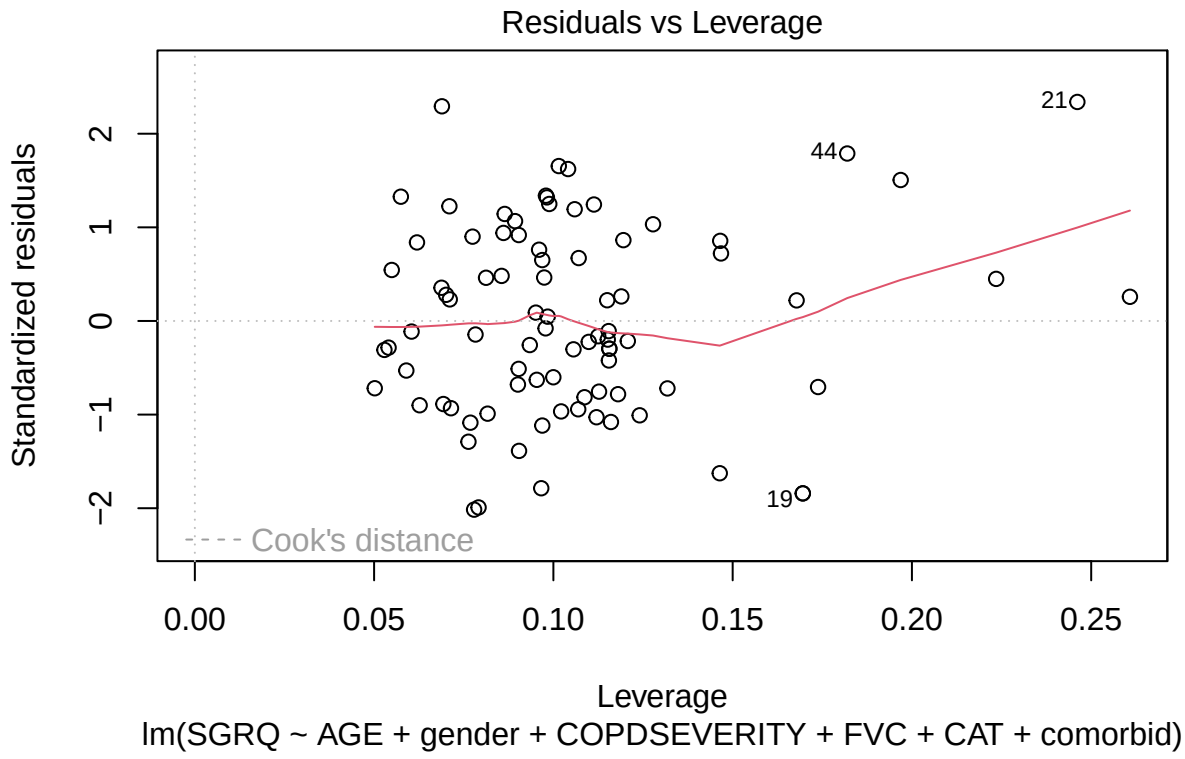
Check using plots :

```
plot(sgrq_model_4)
```


Residuals vs Fitted
Residuals
Fitted values
lm(SGRQ ~ AGE + gender + COPDSEVERITY + FVC + CAT + comorbid)
42
21
55
20
10
0
−10
−20
20
30
40
50
60
70


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(SGRQ ~ AGE + gender + COPDSEVERITY + FVC + CAT + comorbid)
55
42
21
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
42
21
55
20
30
40
50
60
70
Fitted values
lm(SGRQ ~ AGE + gender + COPDSEVERITY + FVC + CAT + comorbid)




```
sgrq_model_5 <- lm(SGRQ~AGE+gender+COPDSEVERITY+FEV1+CAT+comorbid, data=subset_copd)
```

```
summary(sgrq_model_5)
```

```
##
## Call:
## lm(formula = SGRQ ~ AGE + gender + COPDSEVERITY + FEV1 + CAT +
##     comorbid, data = subset_copd)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -20.30  -7.53  -1.20   8.23  23.13
##
## Coefficients:
##                           Estimate Std. Error t value
## (Intercept)                 37.850     17.709    2.14
## AGE                         -0.322      0.207   -1.55
## gender1                      2.095      3.050    0.69
## COPDSEVERITYMODERATE         2.206      3.696    0.60
## COPDSEVERITYSEVERE           0.961      5.716    0.17
## COPDSEVERITYVERY SEVERE     -6.568      8.059   -0.82
## FEV1                        -4.828      3.358   -1.44
## CAT                          1.657      0.160   10.36
## comorbid1                    2.559      2.308    1.11
##                                    Pr(>|t|)
```

```
## (Intercept)                           0.036 *
## AGE                                   0.124
## gender1                               0.494
## COPDSEVERITYMODERATE                  0.552
## COPDSEVERITYSEVERE                    0.867
## COPDSEVERITYVERY SEVERE               0.418
## FEV1                                  0.155
## CAT                    0.00000000000000034 ***
## comorbid1                             0.271
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.4 on 76 degrees of freedom
## Multiple R-squared:  0.677,  Adjusted R-squared:  0.643
## F-statistic: 19.9 on 8 and 76 DF,  p-value: 0.00000000000000076
```

```
confint(sgrq_model_5)
```

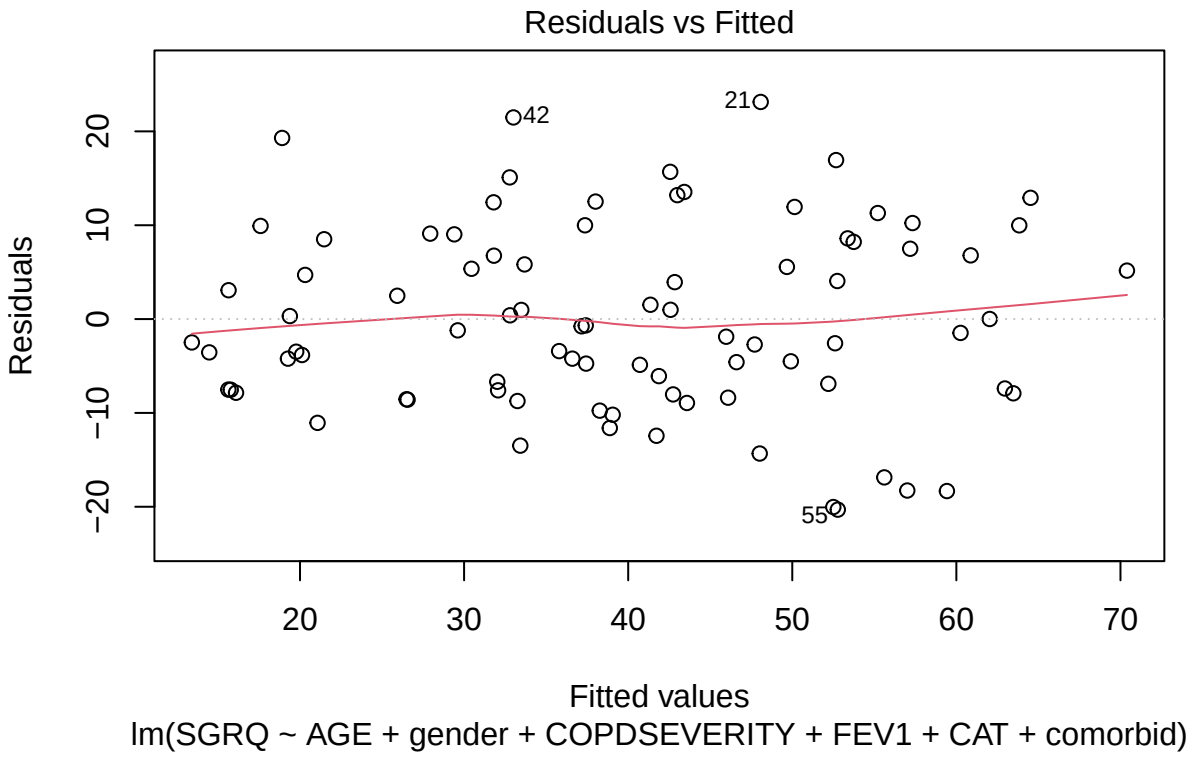
```
##                               2.5 %     97.5 %
## (Intercept)                2.578477 73.1212162
## AGE                       -0.734602  0.0906223
## gender1                   -3.978810  8.1695487
## COPDSEVERITYMODERATE      -5.155513  9.5670693
## COPDSEVERITYSEVERE       -10.423908 12.3463575
## COPDSEVERITYVERY SEVERE  -22.619328  9.4828549
## FEV1                     -11.516758  1.8601978
## CAT                        1.338591  1.9755612
## comorbid1                 -2.037708  7.1549682
```

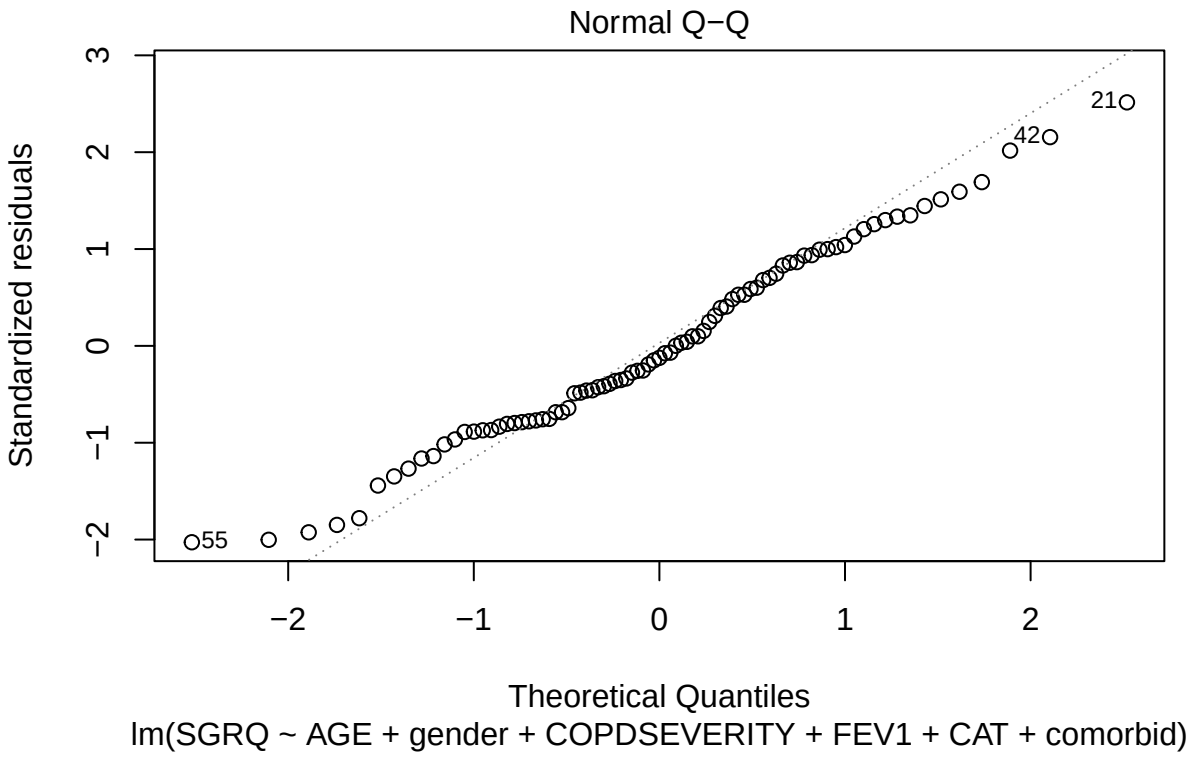
Fit the model :

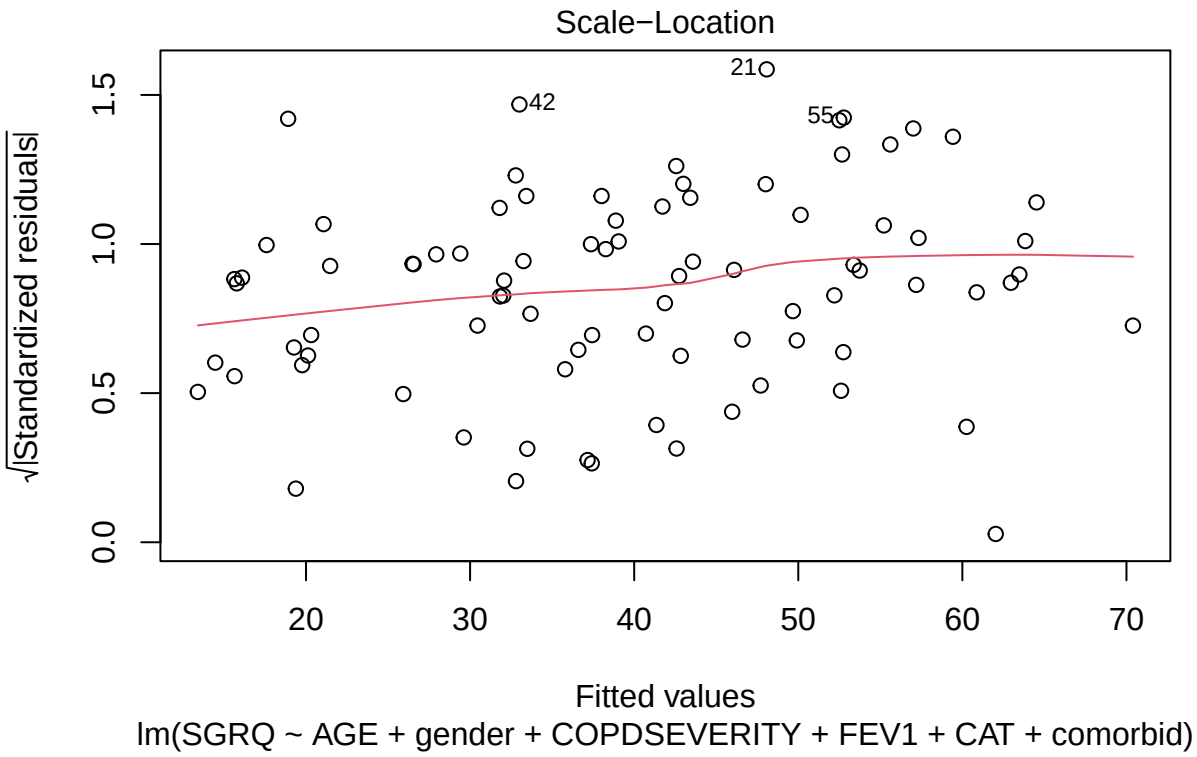
```
predictedsgrqmodel5 <- predict(sgrq_model_5)
residualsgrqmodel5 <- residuals(sgrq_model_5)
```

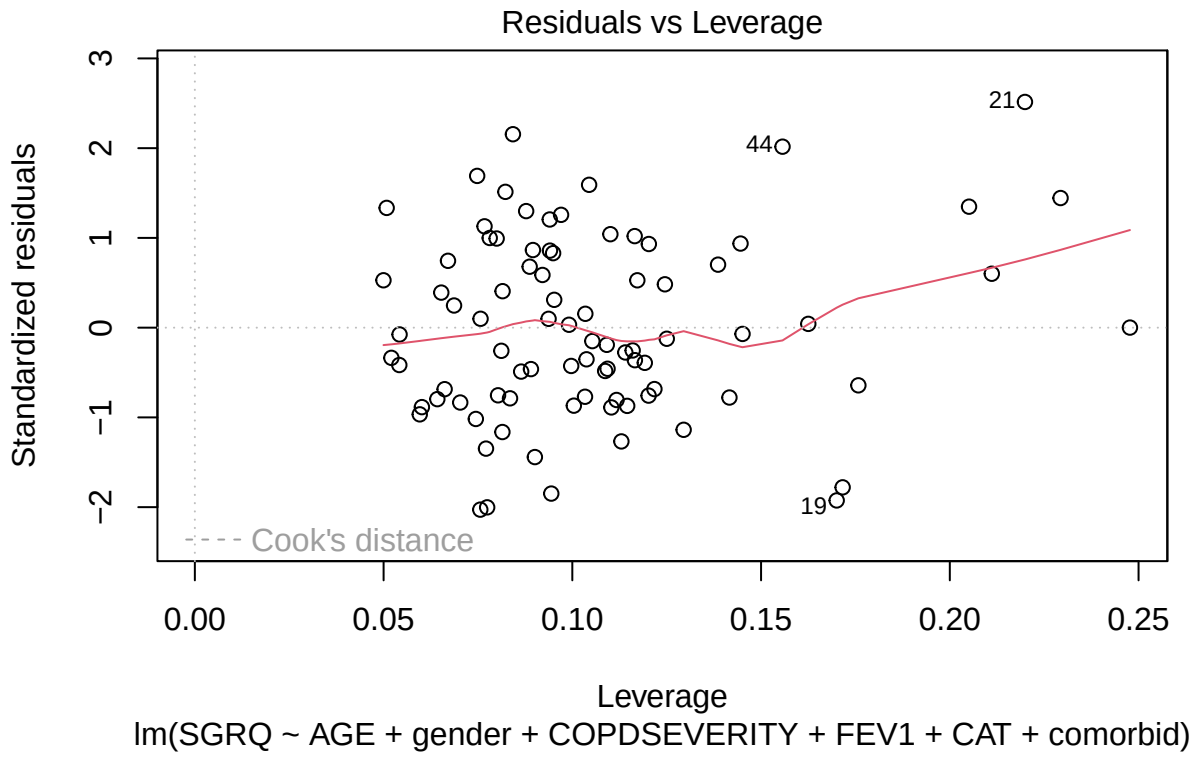
Check using plots :

```
plot(sgrq_model_5)
```


Residuals vs Fitted
Residuals
20
10
0
−10
−20
42
21
55
20
30
40
50
60
70
Fitted values
lm(SGRQ ~ AGE + gender + COPDSEVERITY + FEV1 + CAT + comorbid)


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(SGRQ ~ AGE + gender + COPDSEVERITY + FEV1 + CAT + comorbid)
55
42
21
−2
−1
0
1
2
3


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
20
30
40
50
60
70
21
42
55
Fitted values
lm(SGRQ ~ AGE + gender + COPDSEVERITY + FEV1 + CAT + comorbid)




```
sgrq_model_6 <- lm(SGRQ~AGE+gender+COPDSEVERITY+CAT+comorbid, data=subset_copd)
```

```
summary(sgrq_model_6)
```

```
##
## Call:
## lm(formula = SGRQ ~ AGE + gender + COPDSEVERITY + CAT + comorbid,
##     data = subset_copd)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -20.39  -7.39  -1.24   8.23  23.37
##
## Coefficients:
##                           Estimate Std. Error t value
## (Intercept)                 22.919     14.444    1.59
## AGE                         -0.247      0.202   -1.22
## gender1                     -0.534      2.457   -0.22
## COPDSEVERITYMODERATE         5.334      3.008    1.77
## COPDSEVERITYSEVERE           7.436      3.545    2.10
## COPDSEVERITYVERY SEVERE      2.514      5.039    0.50
## CAT                          1.643      0.161   10.22
## comorbid1                    2.475      2.323    1.07
##                                    Pr(>|t|)
## (Intercept)                           0.117
```

```
## AGE                                  0.226
## gender1                              0.829
## COPDSEVERITYMODERATE                 0.080 .
## COPDSEVERITYSEVERE                   0.039 *
## COPDSEVERITYVERY SEVERE              0.619
## CAT                     0.00000000000000054 ***
## comorbid1                            0.290
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.5 on 77 degrees of freedom
## Multiple R-squared:  0.668,  Adjusted R-squared:  0.638
## F-statistic: 22.2 on 7 and 77 DF,  p-value: 0.000000000000000412
```

```
confint(sgrq_model_6)
```

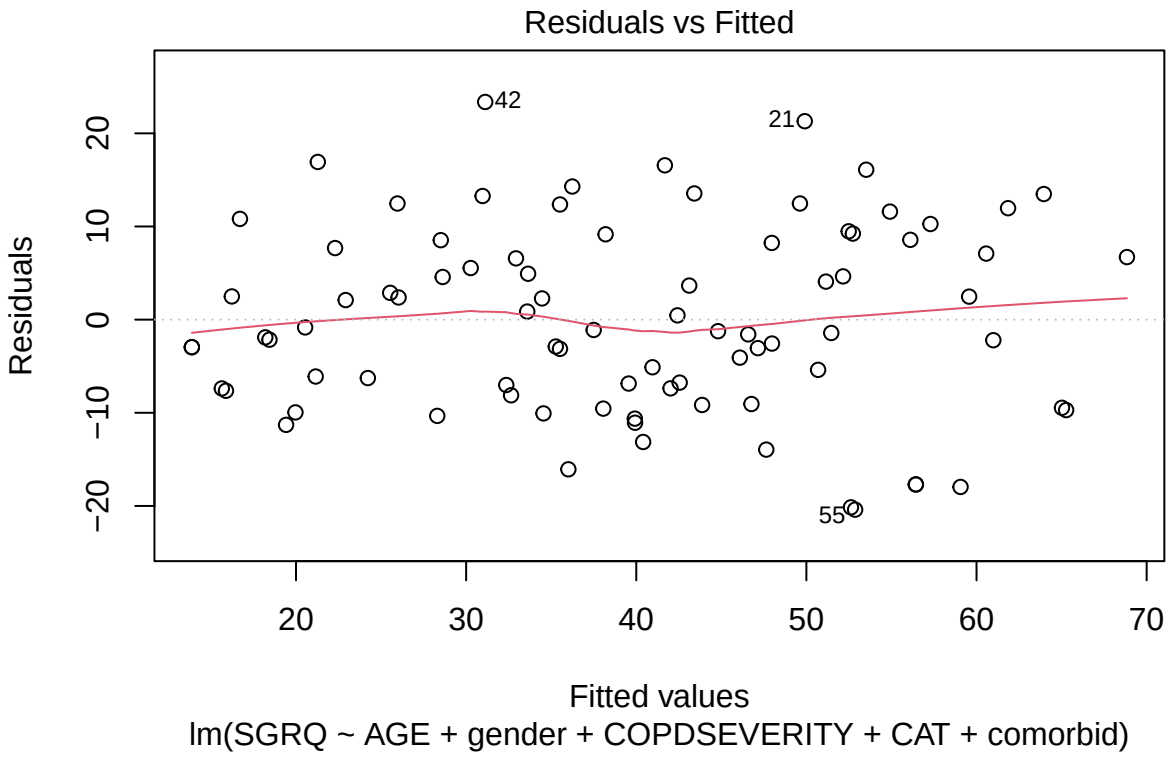
```
##                              2.5 %    97.5 %
## (Intercept)             -5.842719 51.679817
## AGE                     -0.648341  0.155299
## gender1                 -5.427465  4.359185
## COPDSEVERITYMODERATE    -0.656120 11.324457
## COPDSEVERITYSEVERE       0.376137 14.495426
## COPDSEVERITYVERY SEVERE -7.520884 12.548252
## CAT                      1.322866  1.962874
## comorbid1               -2.151078  7.100163
```

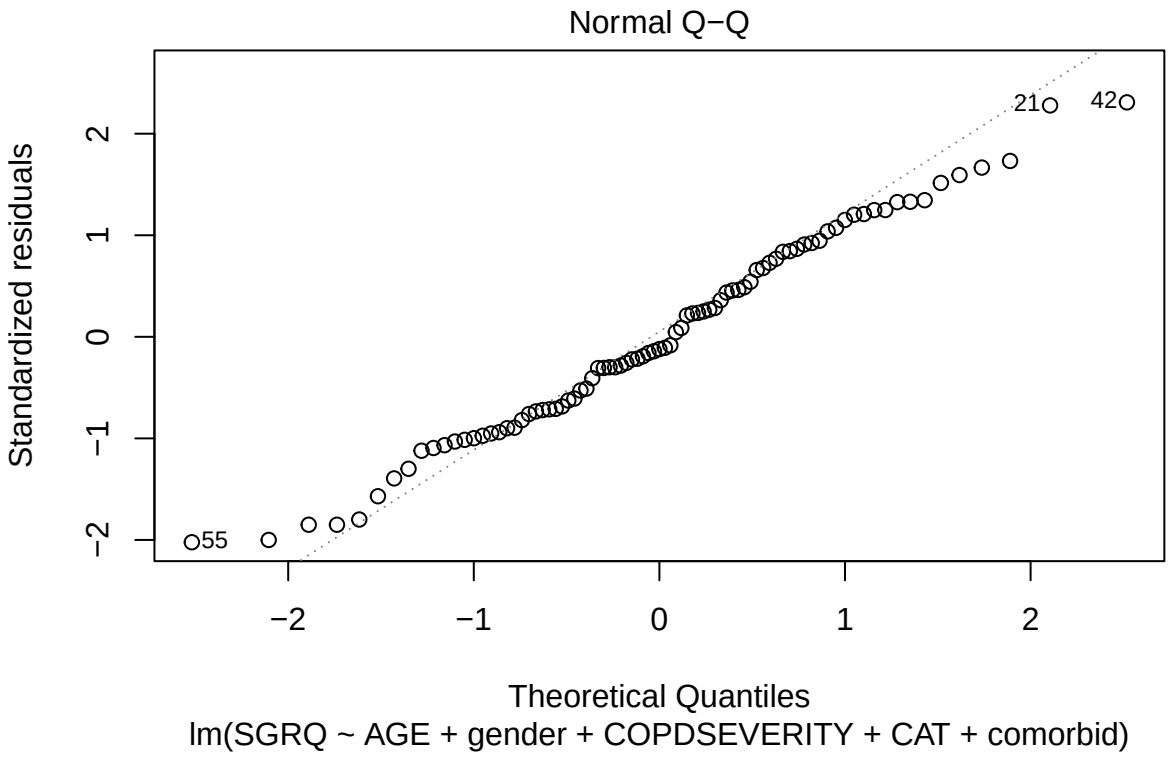
Fit the model :

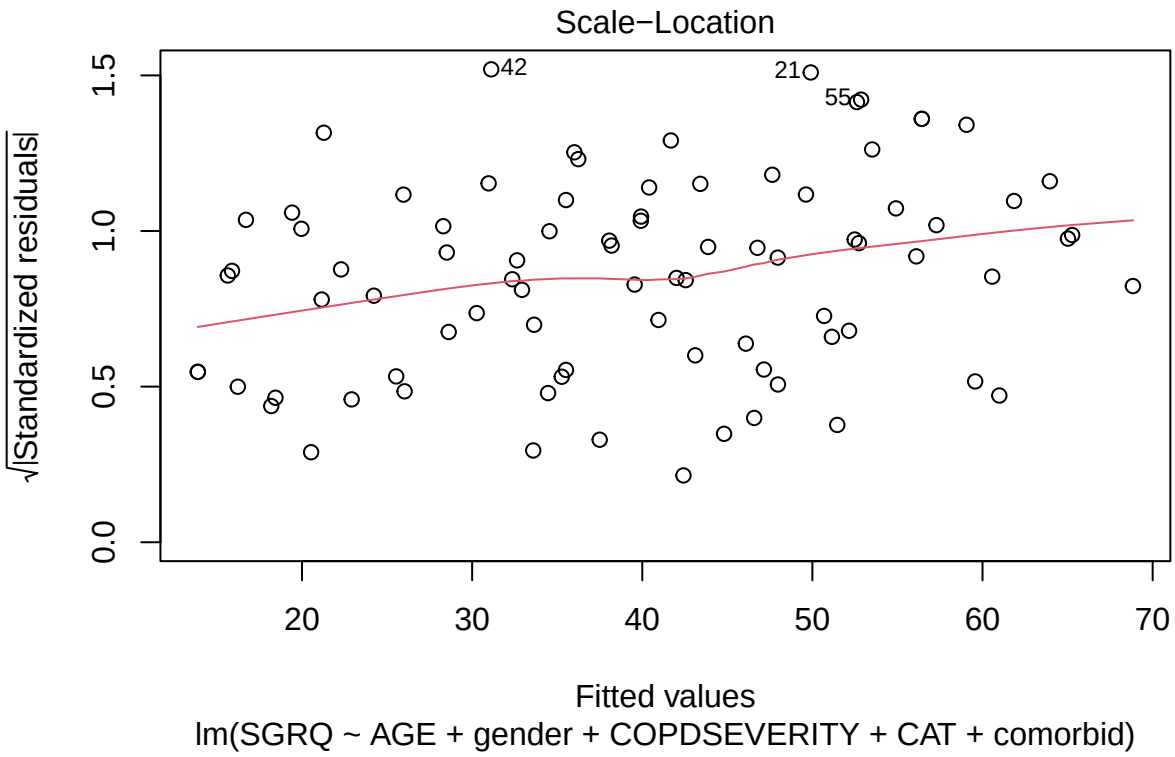
```
predictedsgrqmodel6 <- predict(sgrq_model_6)
residualsgrqmodel6 <- residuals(sgrq_model_6)
```

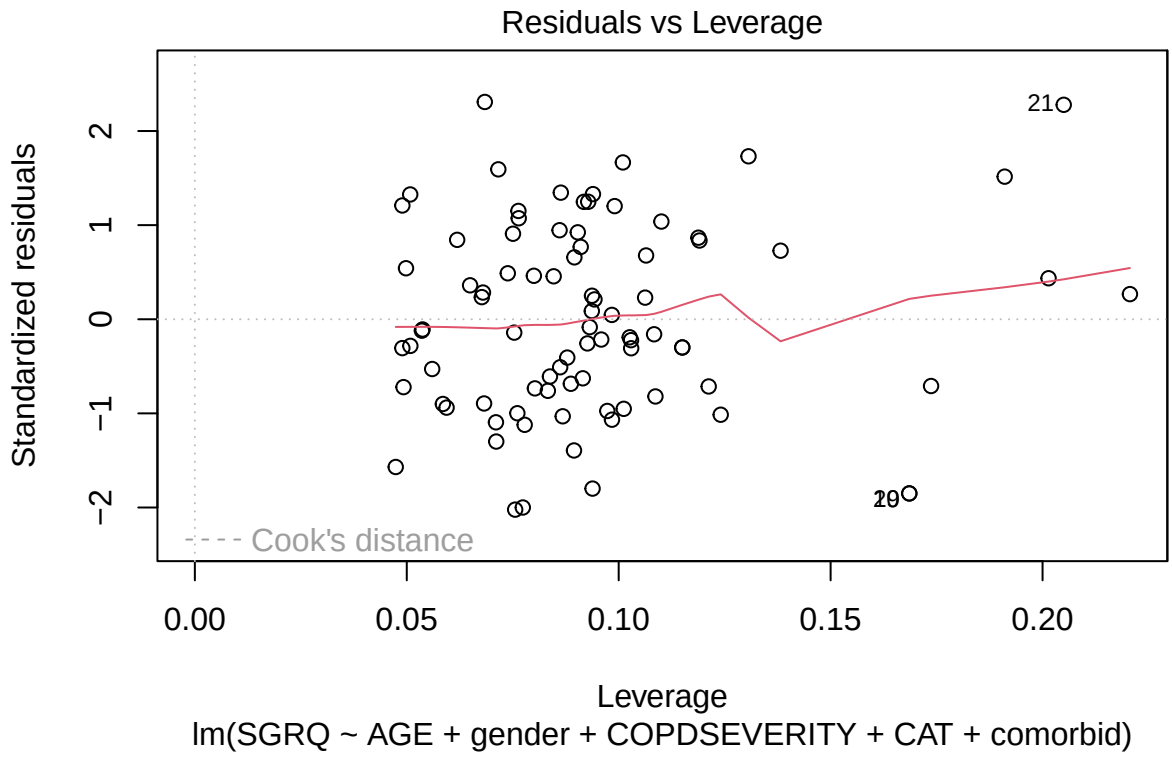
Check using plots :

```
plot(sgrq_model_6)
```


Residuals vs Fitted
Residuals
Fitted values
lm(SGRQ ~ AGE + gender + COPDSEVERITY + CAT + comorbid)
42
21
55
20
10
0
−10
−20
20
30
40
50
60
70


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(SGRQ ~ AGE + gender + COPDSEVERITY + CAT + comorbid)
55
21
42
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
42
21
55
20
30
40
50
60
70
Fitted values
lm(SGRQ ~ AGE + gender + COPDSEVERITY + CAT + comorbid)




```
sgrq_model_7 <- lm(SGRQ~AGE+gender+FEV1+CAT+comorbid, data=subset_copd)
```

```
summary(sgrq_model_7)
```

```
##
## Call:
## lm(formula = SGRQ ~ AGE + gender + FEV1 + CAT + comorbid, data = subset_copd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.282  -6.923  -0.329   7.736  22.638
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)   31.588     14.776    2.14               0.036
## AGE           -0.228      0.197   -1.16               0.249
## gender1        1.062      2.541    0.42               0.677
## FEV1          -3.721      1.860   -2.00               0.049
## CAT            1.602      0.150   10.67 <0.0000000000000002
## comorbid1      2.835      2.283    1.24               0.218
##
## (Intercept) *
## AGE
## gender1
## FEV1        *
```

```
## CAT         ***
## comorbid1
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.4 on 79 degrees of freedom
## Multiple R-squared:  0.663,  Adjusted R-squared:  0.641
## F-statistic: 31.1 on 5 and 79 DF,  p-value: <0.0000000000000002
```

```
confint(sgrq_model_7)
```

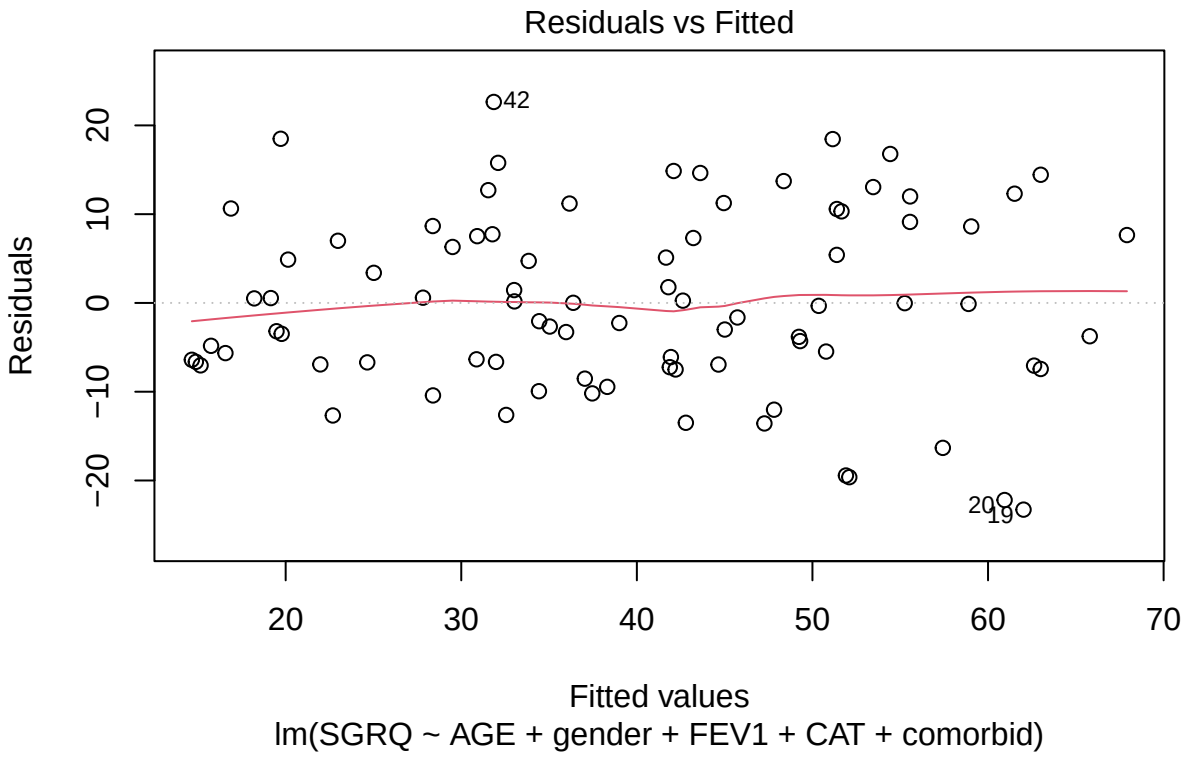
```
##                2.5 %     97.5 %
## (Intercept)  2.17593 60.9994249
## AGE         -0.61990  0.1629042
## gender1     -3.99489  6.1190742
## FEV1        -7.42245 -0.0199073
## CAT          1.30320  1.9006025
## comorbid1   -1.70960  7.3802865
```

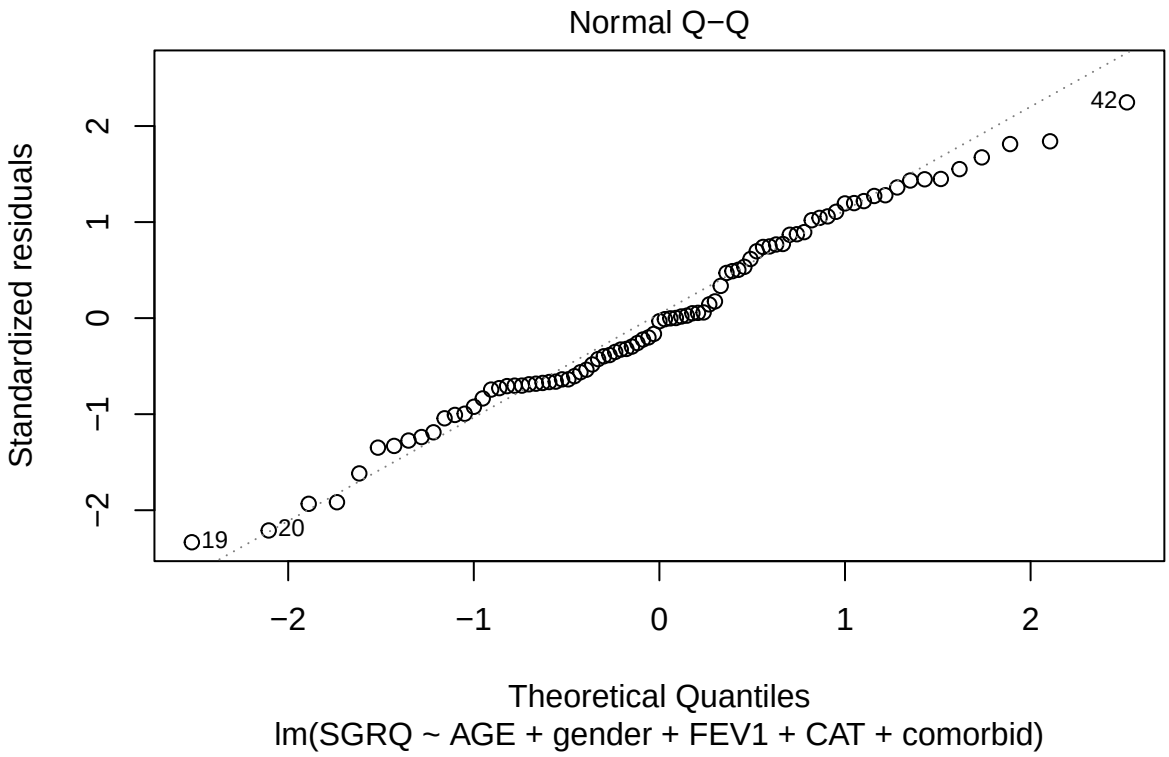
Fit the model :

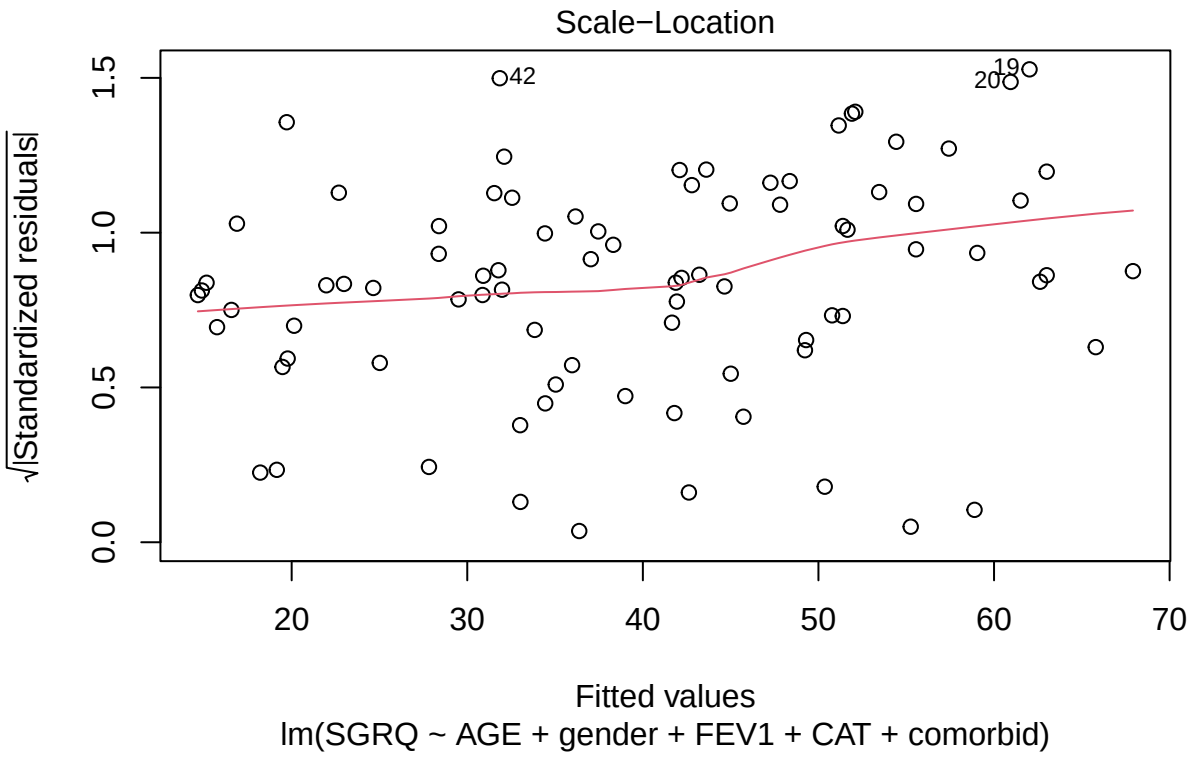
```
predictedsgrqmodel7 <- predict(sgrq_model_7)
residualsgrqmodel7 <- residuals(sgrq_model_7)
```

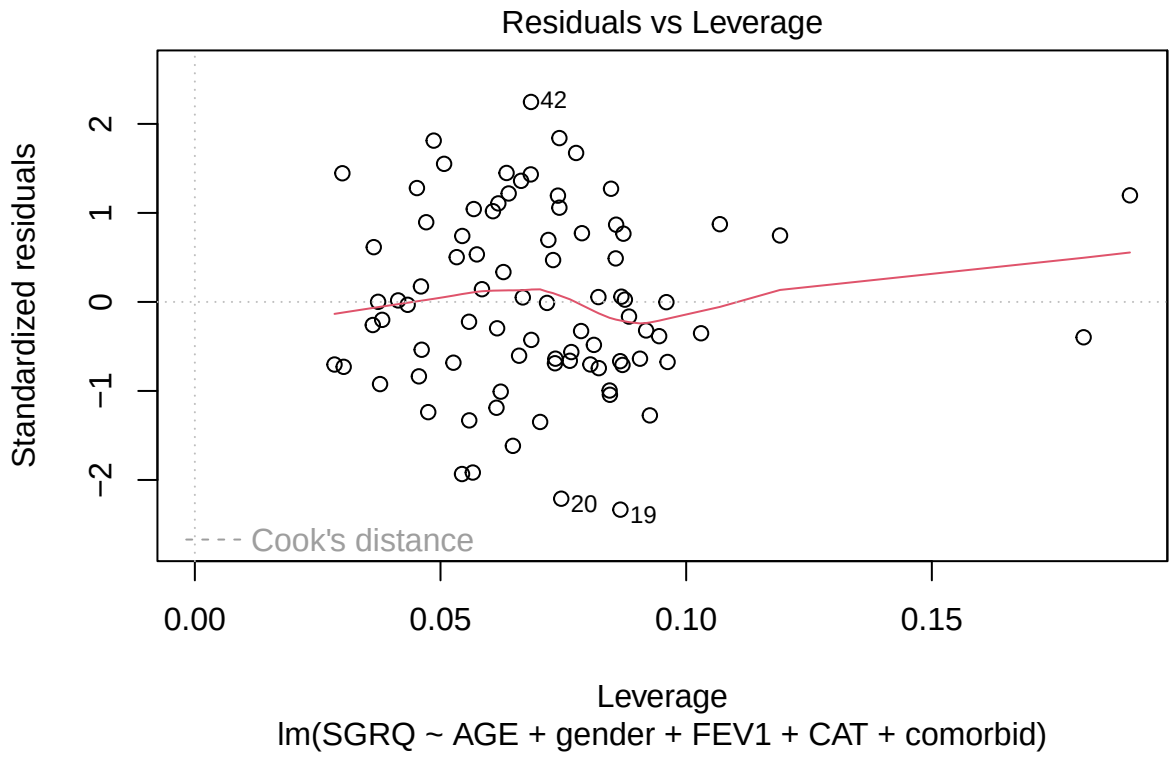
Check using plots :

```
plot(sgrq_model_7)
```


Residuals vs Fitted
Residuals
Fitted values
lm(SGRQ ~ AGE + gender + FEV1 + CAT + comorbid)
42
20
19
20
30
40
50
60
70
−20
−10
0
10
20


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(SGRQ ~ AGE + gender + FEV1 + CAT + comorbid)
19
20
42
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
42
19
20
20
30
40
50
60
70
Fitted values
lm(SGRQ ~ AGE + gender + FEV1 + CAT + comorbid)




```
sgrq_model_8 <- lm(SGRQ~AGE+gender+FVC+CAT+comorbid, data=subset_copd)
```

```
summary(sgrq_model_8)
```

```
##
## Call:
## lm(formula = SGRQ ~ AGE + gender + FVC + CAT + comorbid, data = subset_copd)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -21.69  -7.38  -1.30   7.96  24.14
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)   27.921     15.773    1.77               0.081
## AGE           -0.222      0.203   -1.09               0.277
## gender1        0.523      2.748    0.19               0.849
## FVC           -1.133      1.352   -0.84               0.404
## CAT            1.667      0.150   11.10 <0.0000000000000002
## comorbid1      2.542      2.350    1.08               0.283
##
## (Intercept) .
## AGE
## gender1
## FVC
```

```
## CAT         ***
## comorbid1
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.7 on 79 degrees of freedom
## Multiple R-squared:  0.649,  Adjusted R-squared:  0.627
## F-statistic: 29.2 on 5 and 79 DF,  p-value: <0.0000000000000002
```

```
confint(sgrq_model_8)
```

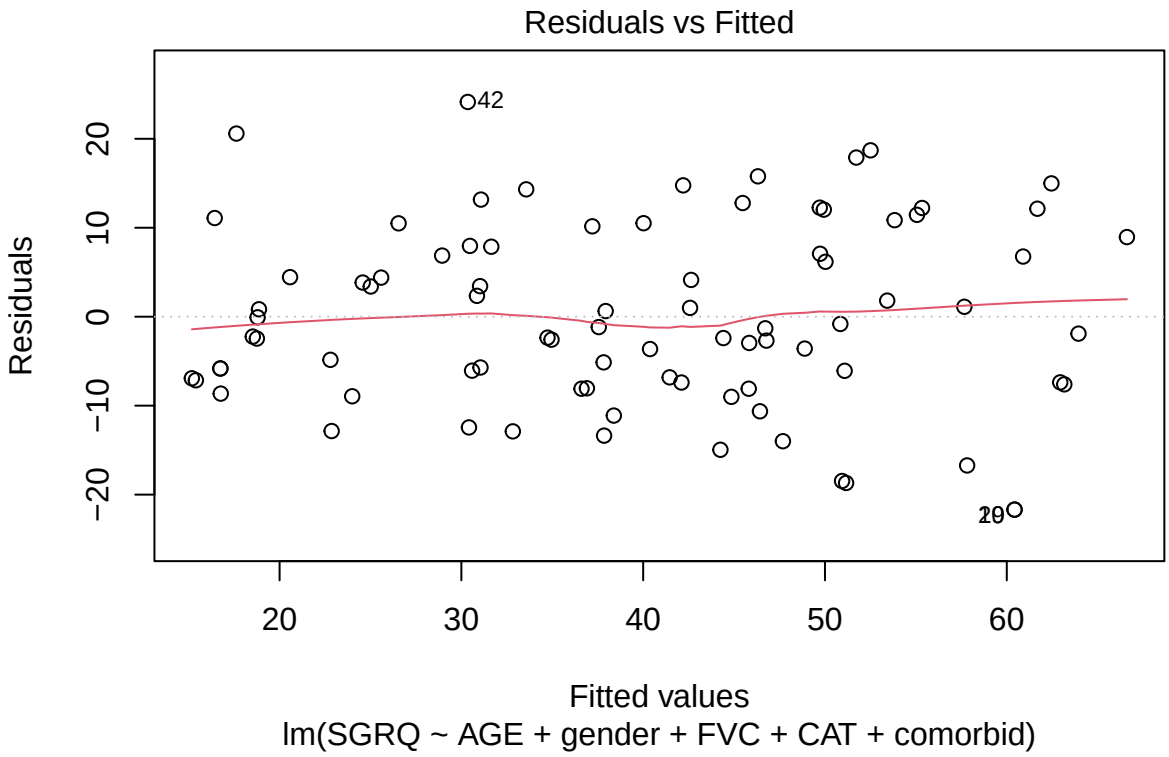
```
##                 2.5 %    97.5 %
## (Intercept) -3.475136 59.317486
## AGE         -0.625582  0.181696
## gender1     -4.947245  5.993752
## FVC         -3.823601  1.556978
## CAT          1.368197  1.966053
## comorbid1   -2.134336  7.219172
```

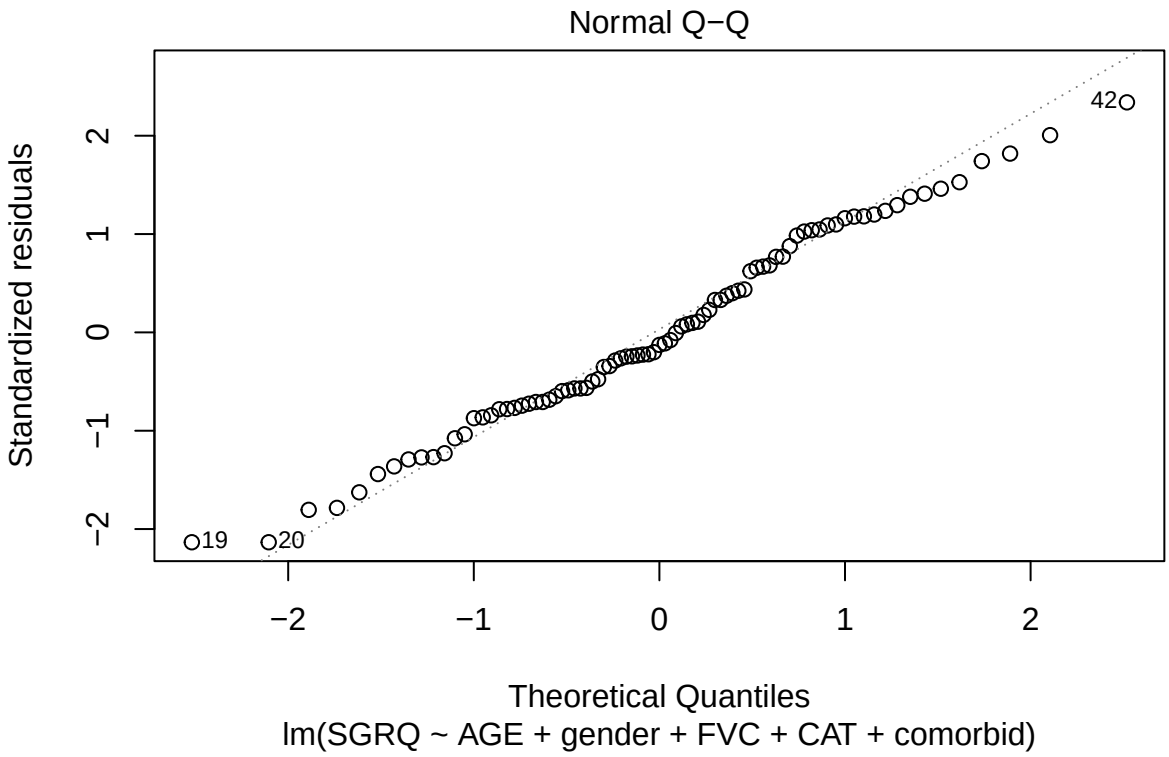
Fit the model :

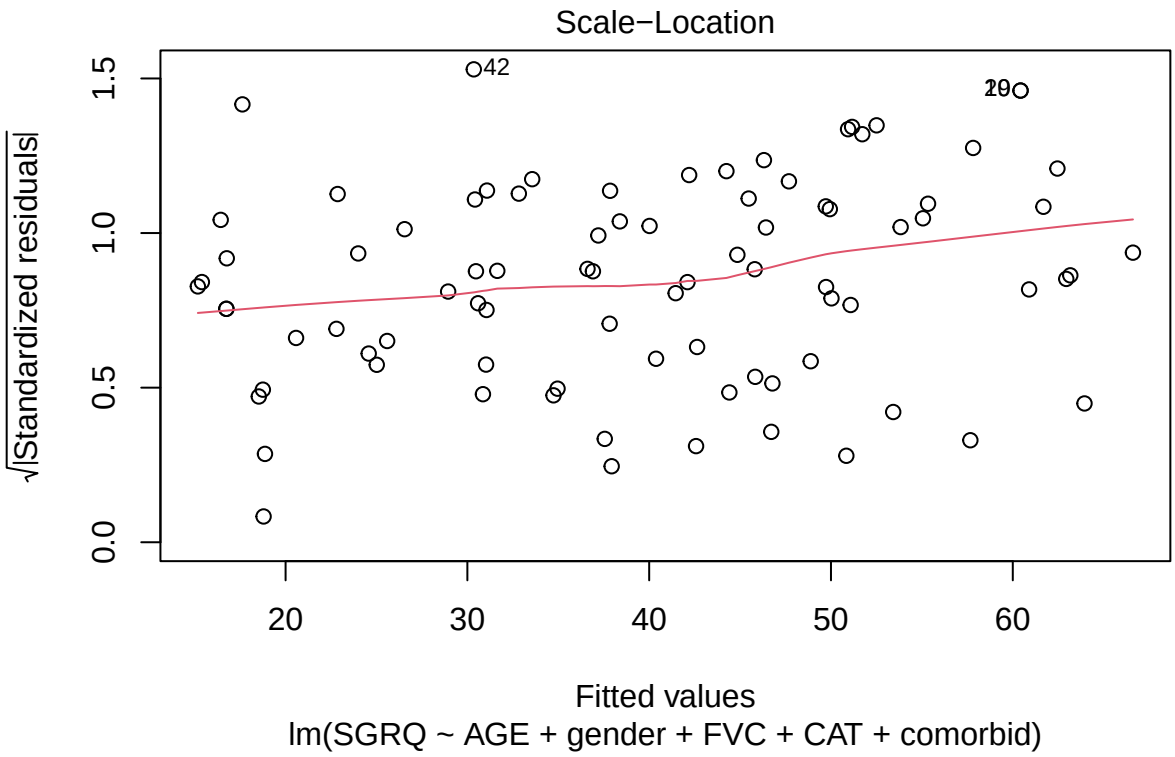
```
predictedsgrqmodel8 <- predict(sgrq_model_8)
residualsgrqmodel8 <- residuals(sgrq_model_8)
```

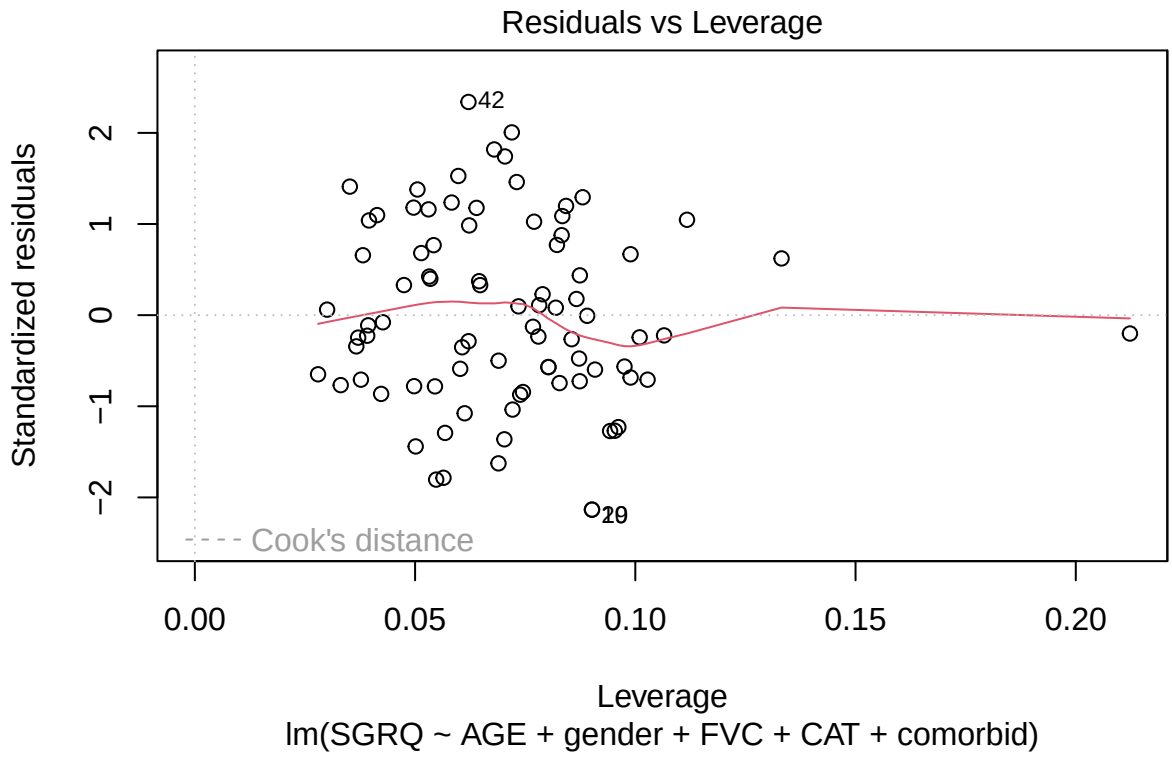
Check using plots :

```
plot(sgrq_model_8)
```

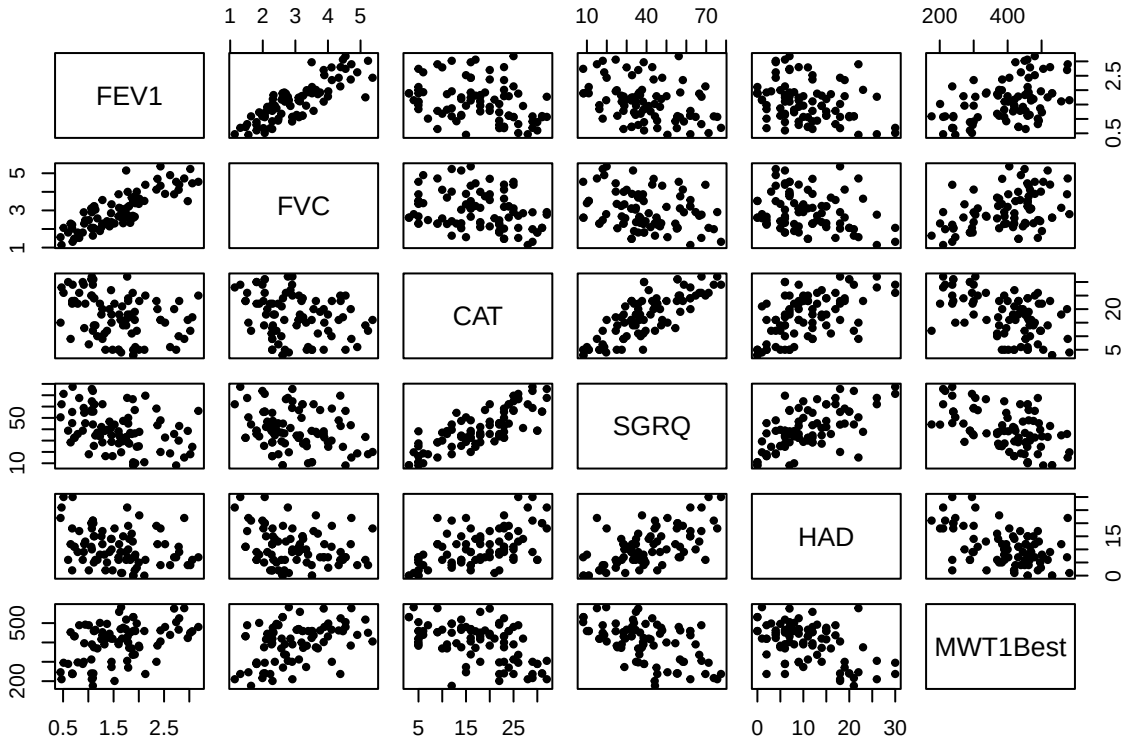

Residuals vs Fitted
Residuals
20
10
0
−10
−20
42
20
20
30
40
50
60
Fitted values
lm(SGRQ ~ AGE + gender + FVC + CAT + comorbid)


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(SGRQ ~ AGE + gender + FVC + CAT + comorbid)
19
20
42
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
42
20
20
30
40
50
60
Fitted values
lm(SGRQ ~ AGE + gender + FVC + CAT + comorbid)




```
pairs(~FEV1+FVC+CAT+SGRQ+HAD+MWT1Best, data=subset_copd, pch=20,cex=1)
```



According to correlation matrix above, it is found that FEV1 and FVC has quite high correlation with each other while their correlations with SGRQ are quite spurious. These are the explanation that previous models have collinearity with FEV1, FVC, and COPDSEVERITY. CAT and HAD are two variables which has better correlation with SGRQ. So,removing variables FEV1, FVC and COPDSEVERITY data and use CAT as the only predictor of lung function in COPD.

```
sgrq_model_9 <- lm(SGRQ~CAT+HAD, data=subset_copd)
```

```
summary(sgrq_model_9)
```

```
##
## Call:
## lm(formula = SGRQ ~ CAT + HAD, data = subset_copd)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -22.19  -6.79  -0.39   7.26  22.41
##
## Coefficients:
##             Estimate Std. Error t value          Pr(>|t|)
## (Intercept)    8.169      2.625    3.11           0.00256
## CAT            1.409      0.162    8.72 0.00000000000026
## HAD            0.635      0.183    3.47           0.00084
##
## (Intercept) **
## CAT         ***
```

```
## HAD         ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.96 on 82 degrees of freedom
## Multiple R-squared:  0.681,  Adjusted R-squared:  0.673
## F-statistic: 87.6 on 2 and 82 DF,  p-value: <0.0000000000000002
```

```
confint(sgrq_model_9)
```

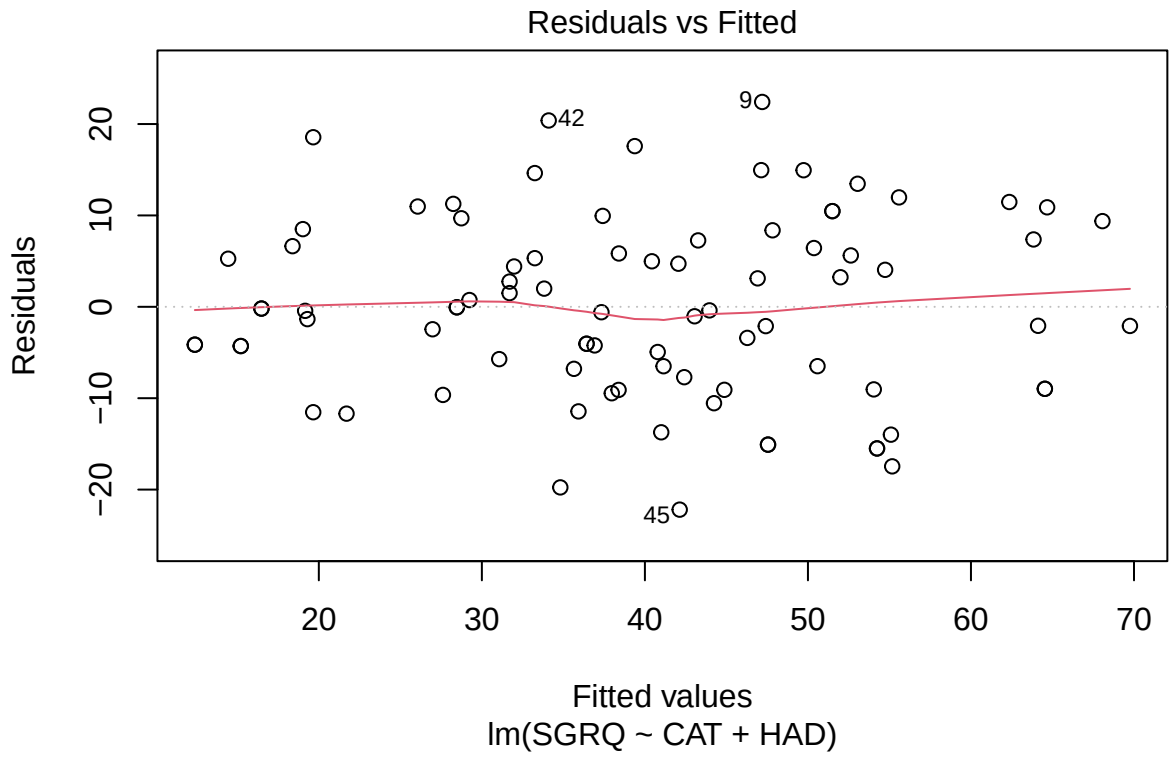
```
##                2.5 %    97.5 %
## (Intercept) 2.947057 13.390064
## CAT         1.087180  1.730203
## HAD         0.270682  0.998938
```

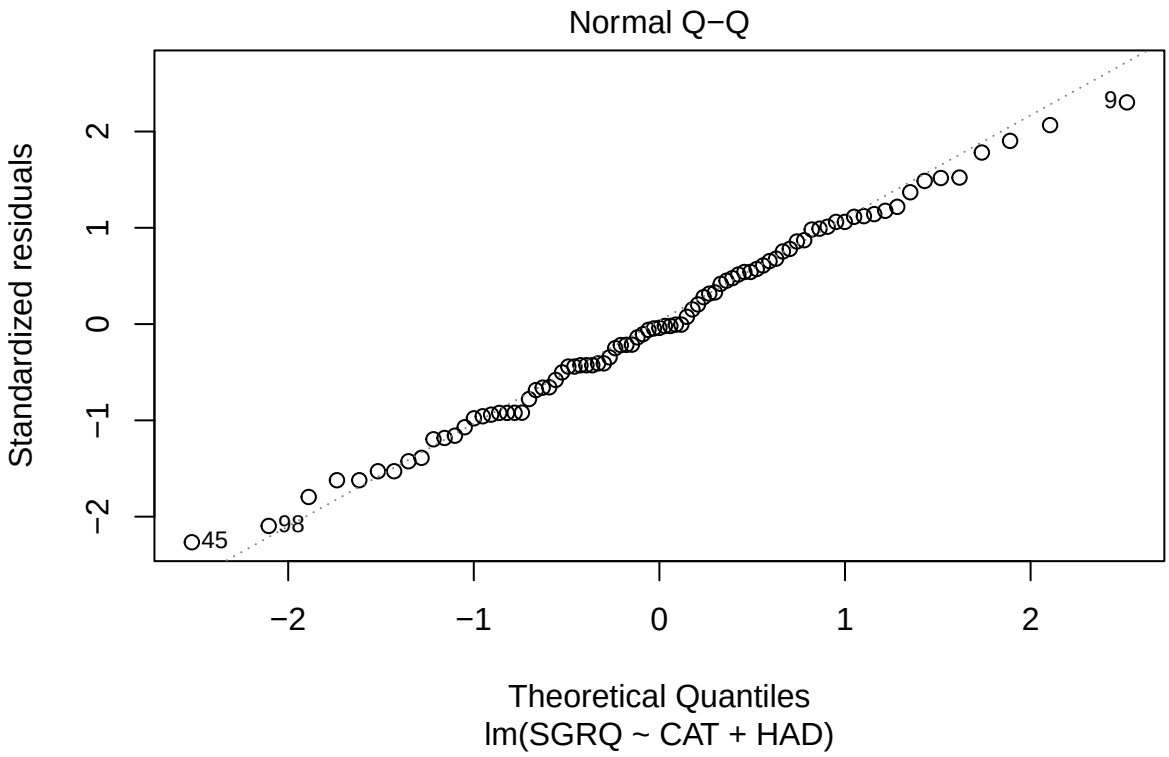
Fit the model :

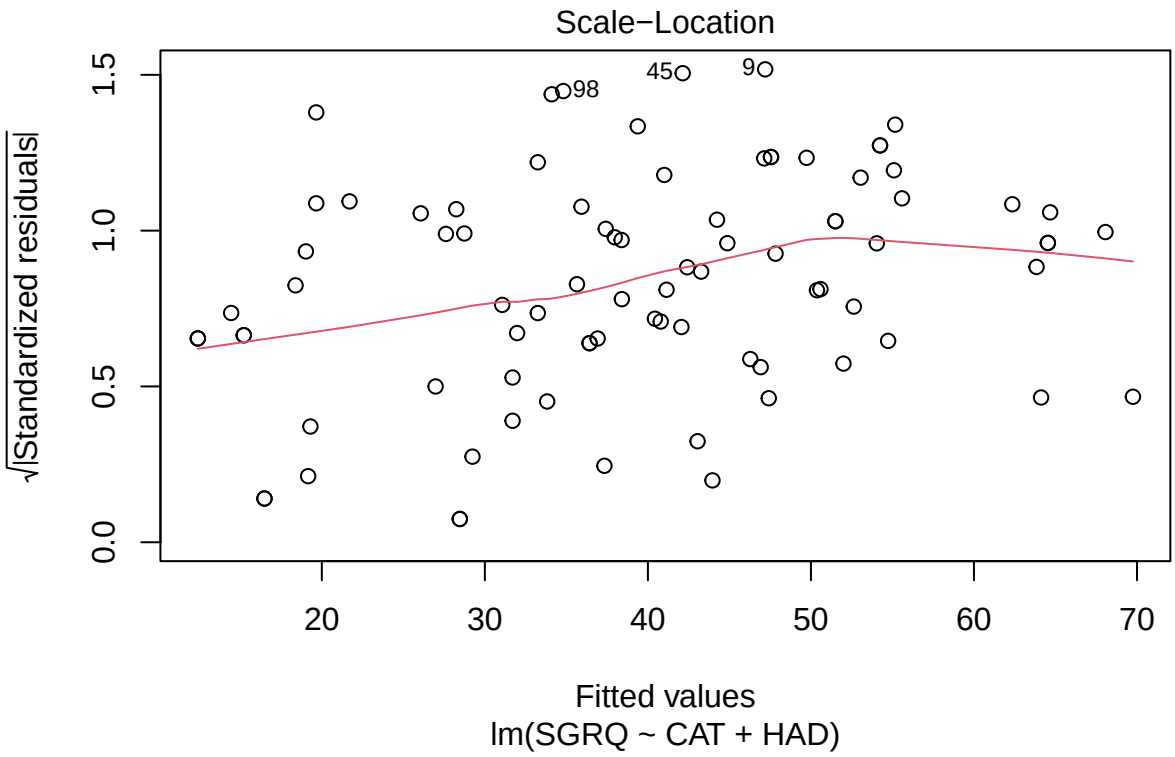
```
predictedsgrqmodel9 <- predict(sgrq_model_9)
residualsgrqmodel9 <- residuals(sgrq_model_9)
```

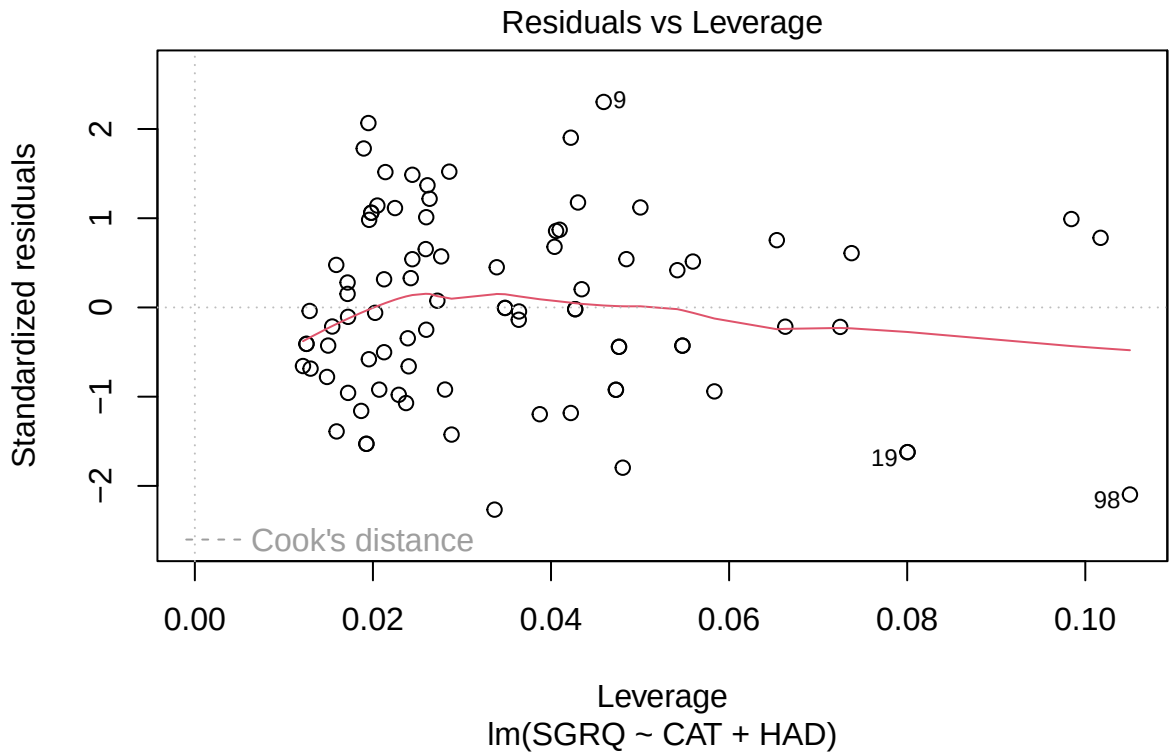
Check using plots :

```
plot(sgrq_model_9)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(SGRQ ~ CAT + HAD)
45
98
9
−2
−1
0
1
2


Scale−Location
45
98
9
√|Standardized residuals|
0.0
0.5
1.0
1.5
20
30
40
50
60
70
Fitted values
lm(SGRQ ~ CAT + HAD)



After removing FEV1, FVC, and COPD Severity predictors, it is found that the multiple Rsquared is improving to 0.681 with significancy <0.00001. Moreover, significancy of CAT and HAD retained with value <0.05.

```
sgrq_model_10 <- lm(SGRQ~CAT+HAD+MWT1Best, data=subset_copd)
```

```
summary(sgrq_model_10)
```

```
##
## Call:
## lm(formula = SGRQ ~ CAT + HAD + MWT1Best, data = subset_copd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.619  -9.097  -0.331   6.792  20.094
##
## Coefficients:
##             Estimate Std. Error t value       Pr(>|t|)
## (Intercept)  17.6980     7.3378    2.41         0.0181 *
## CAT           1.3277     0.1709    7.77 0.000000000022 ***
## HAD           0.5574     0.1903    2.93         0.0044 **
## MWT1Best     -0.0183     0.0132   -1.39         0.1685
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 9.91 on 81 degrees of freedom
## Multiple R-squared:  0.689,  Adjusted R-squared:  0.677
## F-statistic: 59.7 on 3 and 81 DF,  p-value: <0.0000000000000002
```

```
confint(sgrq_model_10)
```

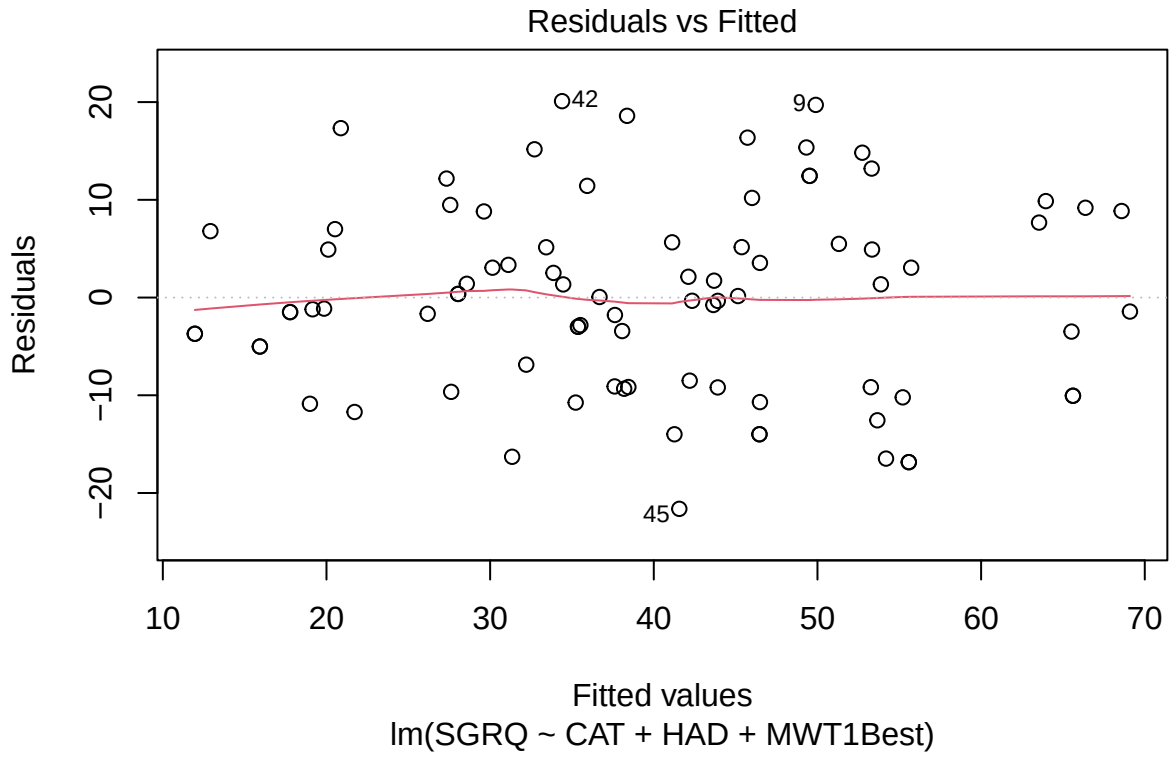
```
##                 2.5 %      97.5 %
## (Intercept)  3.0981565 32.29788181
## CAT          0.9876326  1.66786668
## HAD          0.1787215  0.93614884
## MWT1Best    -0.0445349  0.00790904
```

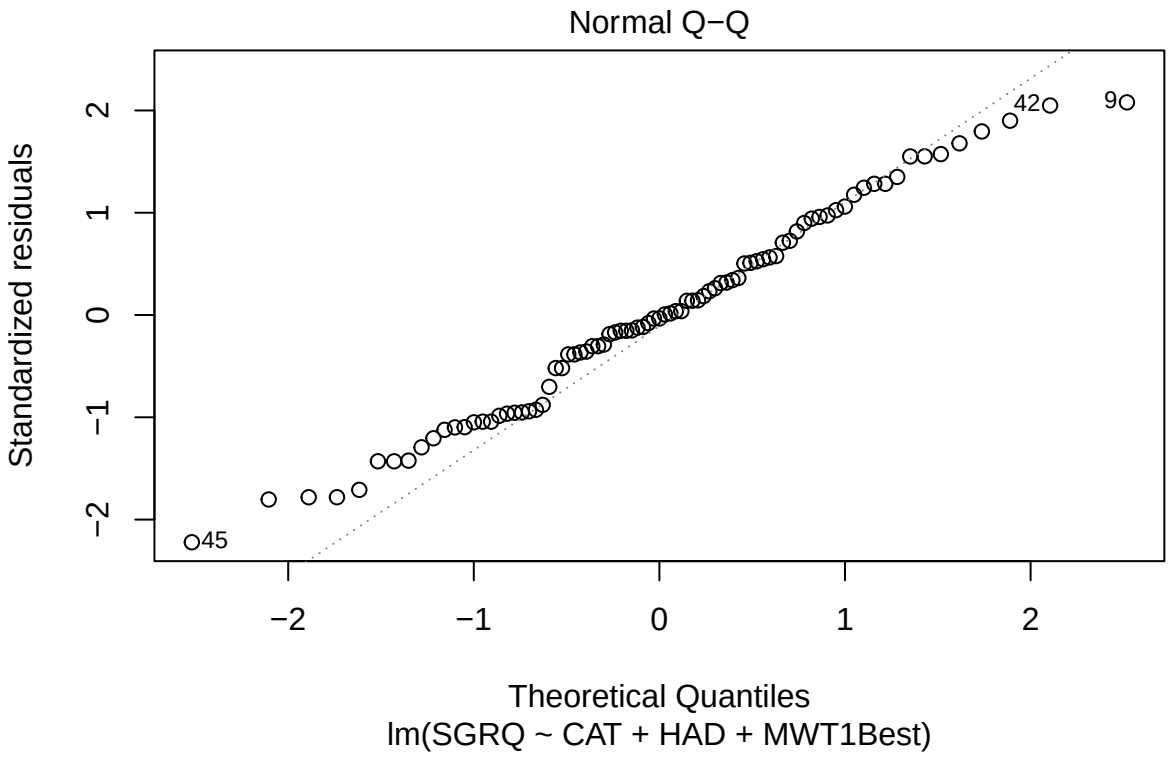
Fit the model :

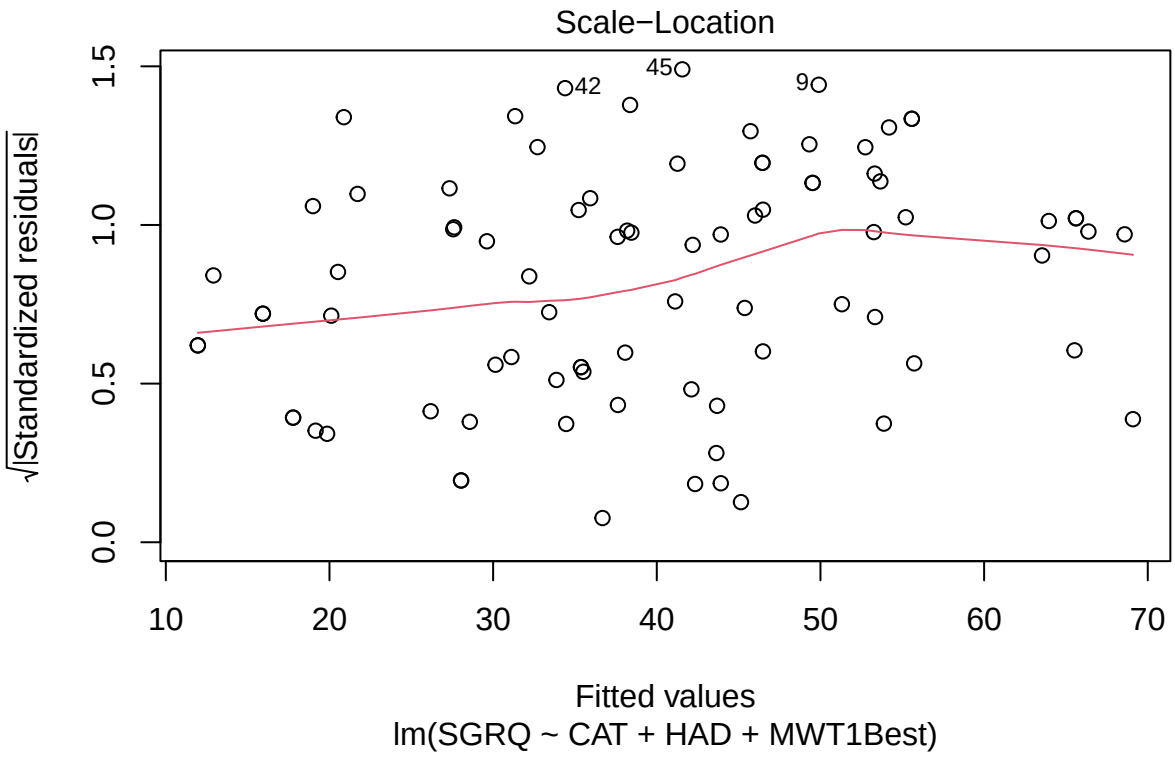
```
predictedsgrqmodel10 <- predict(sgrq_model_10)
residualsgrqmodel10 <- residuals(sgrq_model_10)
```

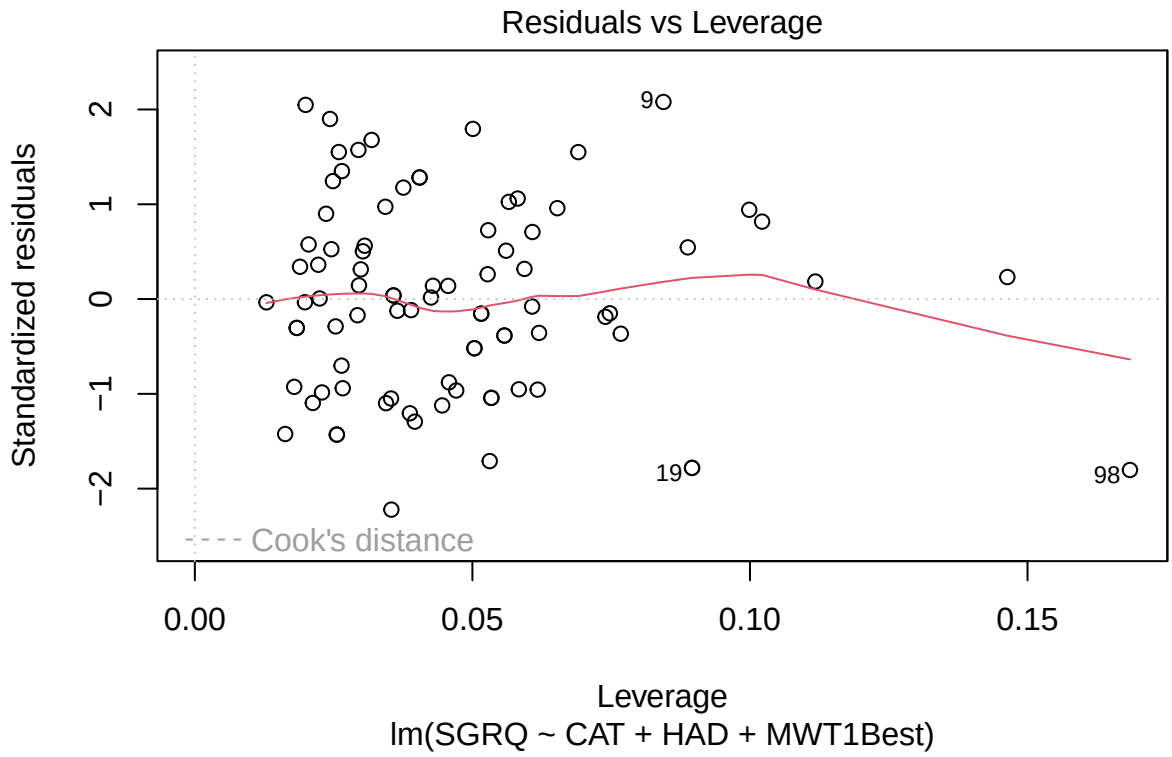
Check using plots :

```
plot(sgrq_model_10)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(SGRQ ~ CAT + HAD + MWT1Best)
45
42
9
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
10
20
30
40
50
60
70
42
45
9
Fitted values
lm(SGRQ ~ CAT + HAD + MWT1Best)



Exploring the effect of categorical variables

gender,comorbid, Diabetes, IHD, AtrialFib, hypertension

```
mlr1 <- lm(SGRQ~CAT+HAD+AGE+gender, data=subset_copd)
```

```
summary(mlr1)
```

```
##
## Call:
## lm(formula = SGRQ ~ CAT + HAD + AGE + gender, data = subset_copd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.066  -7.260   0.026   7.296  22.244
##
## Coefficients:
##             Estimate Std. Error t value         Pr(>|t|)
## (Intercept)   17.105     13.638    1.25            0.2134
## CAT            1.413      0.163    8.65 0.00000000000044
## HAD            0.624      0.188    3.32            0.0014
## AGE           -0.129      0.190   -0.68            0.4972
## gender1        0.506      2.349    0.22            0.8301
##
## (Intercept)
## CAT         ***
```

```
## HAD         **
## AGE
## gender1
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.1 on 80 degrees of freedom
## Multiple R-squared:  0.683,  Adjusted R-squared:  0.667
## F-statistic: 43.1 on 4 and 80 DF,  p-value: <0.0000000000000002
```

```
confint(mlr1)
```

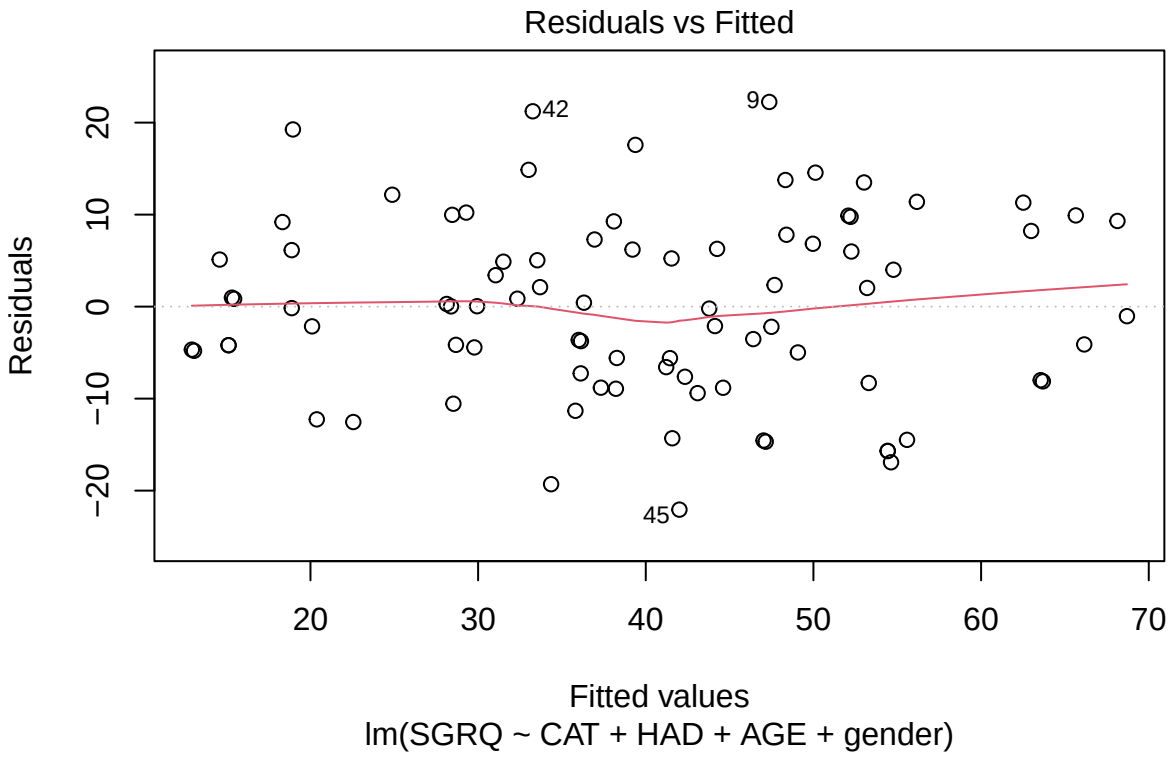
```
##                  2.5 %    97.5 %
## (Intercept) -10.035866 44.246703
## CAT           1.087744  1.738205
## HAD           0.250113  0.997676
## AGE          -0.506779  0.248083
## gender1      -4.169048  5.180428
```

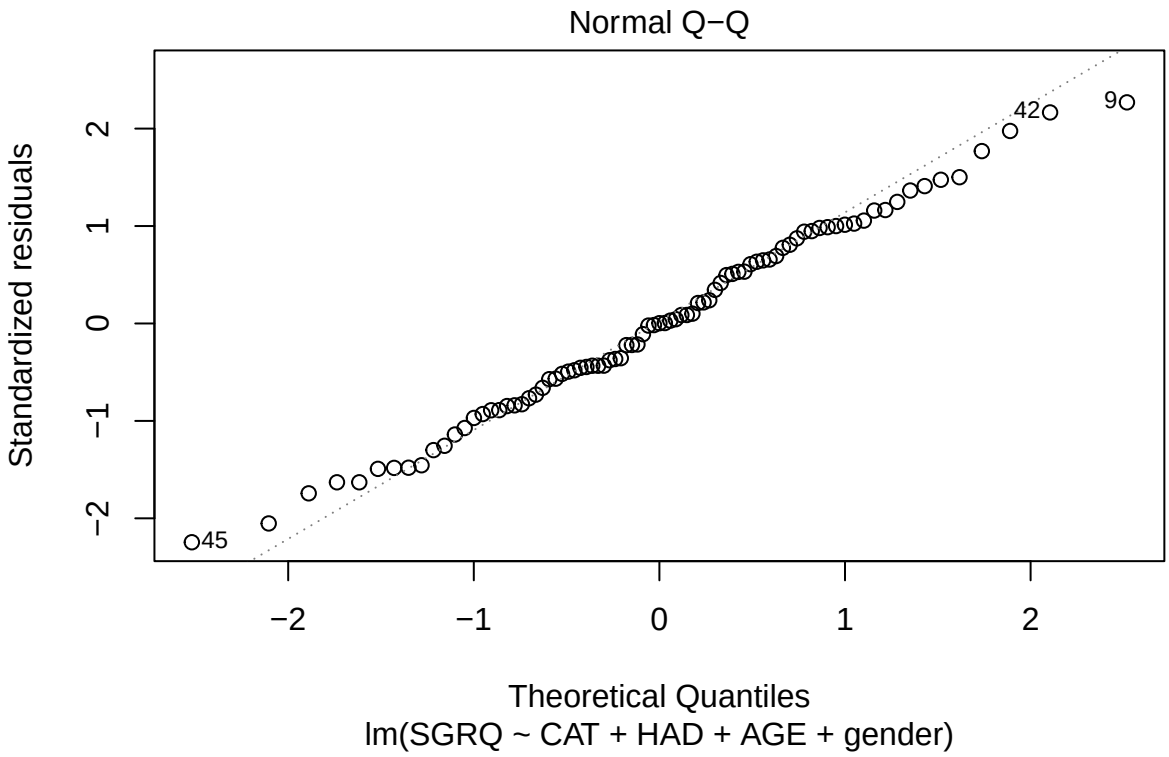
fit the model :

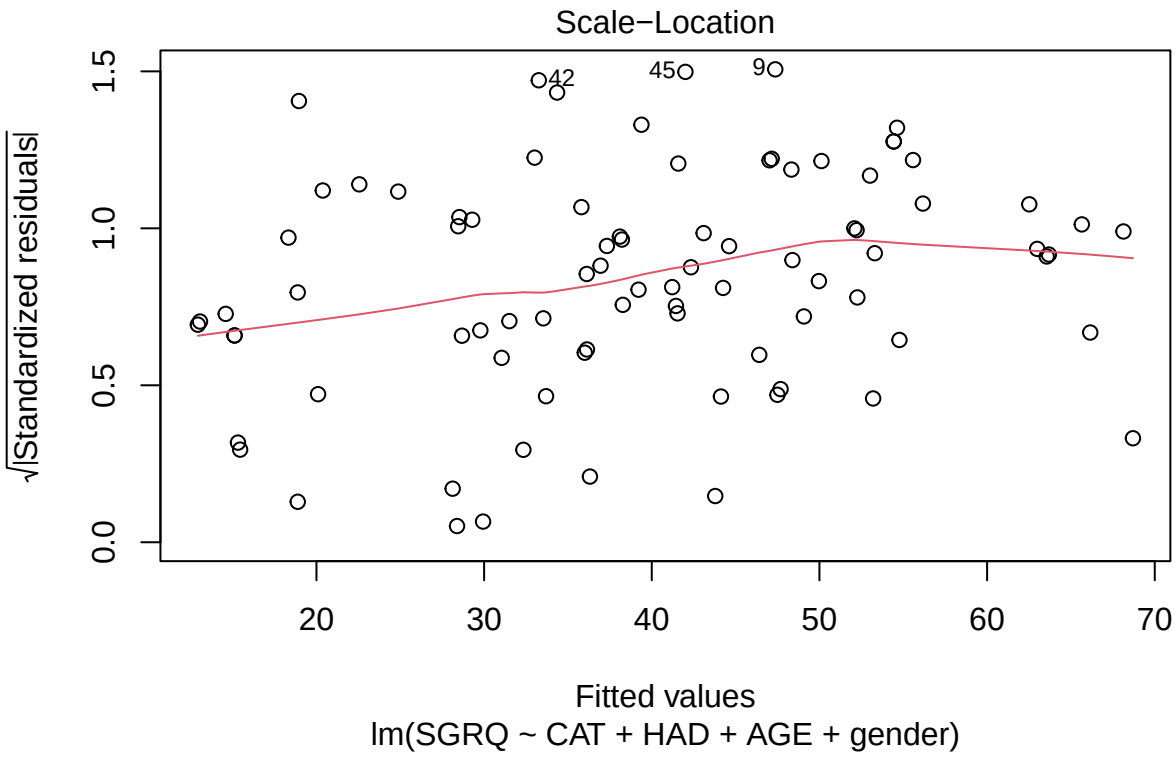
```
predictedsgrqmodel9 <- predict(mlr1)
residualsgrqmodel9 <- residuals(mlr1)
```

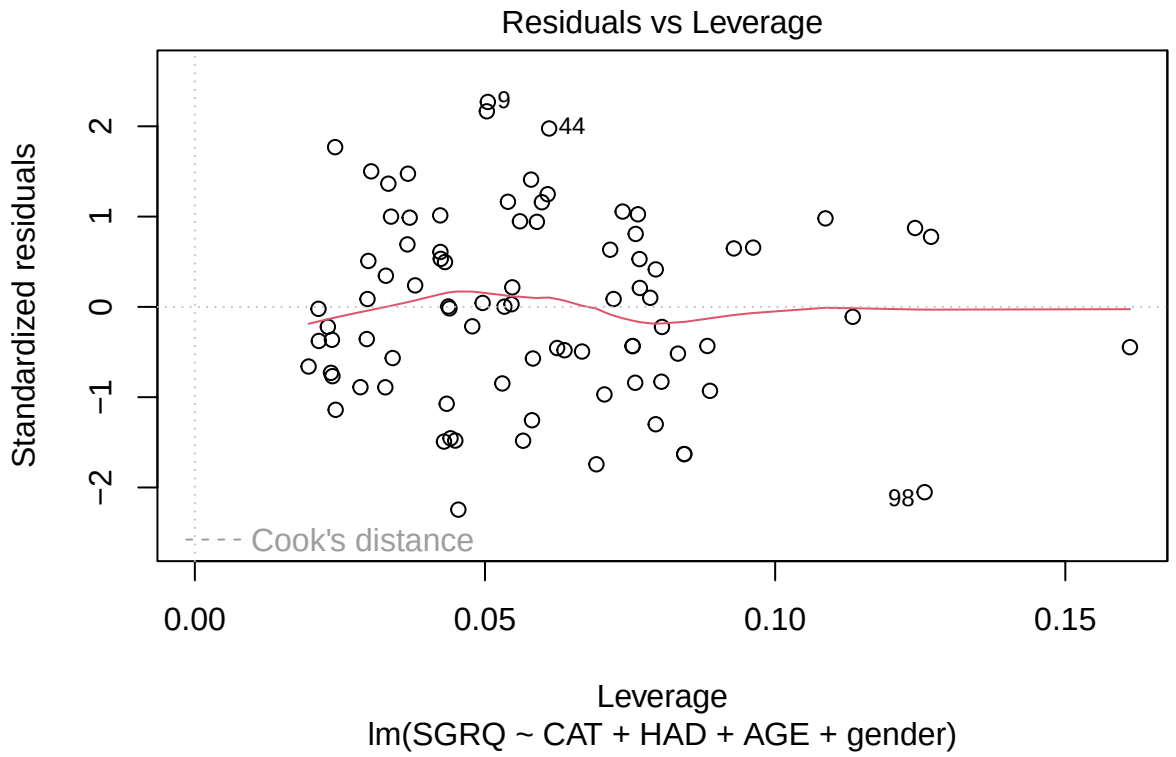
Check using plots :

```
plot(mlr1)
```


Residuals vs Fitted
Residuals
Fitted values
lm(SGRQ ~ CAT + HAD + AGE + gender)
42
9
45
20
10
0
−10
−20
20
30
40
50
60
70


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(SGRQ ~ CAT + HAD + AGE + gender)
45
42
9
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
42
45
9
20
30
40
50
60
70
Fitted values
lm(SGRQ ~ CAT + HAD + AGE + gender)



Adding gender in the model doesn't change much in value of multiple R-squared and p-value. Gender is not significant predictor to SGRQ.

```
mlr2 <- lm(SGRQ~CAT+HAD+AGE+gender+comorbid, data=subset_copd)
```

```
summary(mlr2)
```

```
##
## Call:
## lm(formula = SGRQ ~ CAT + HAD + AGE + gender + comorbid, data = subset_copd)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -20.93  -7.90  -0.03   7.79  22.19
##
## Coefficients:
##             Estimate Std. Error t value         Pr(>|t|)
## (Intercept)   16.553     13.693    1.21           0.2303
## CAT            1.416      0.164    8.64 0.00000000000049
## HAD            0.601      0.191    3.15           0.0023
## AGE           -0.132      0.190   -0.69           0.4903
## gender1        0.453      2.356    0.19           0.8481
## comorbid1      1.705      2.233    0.76           0.4472
##
## (Intercept)
```

```
## CAT         ***
## HAD         **
## AGE
## gender1
## comorbid1
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.1 on 79 degrees of freedom
## Multiple R-squared:  0.685,  Adjusted R-squared:  0.665
## F-statistic: 34.4 on 5 and 79 DF,  p-value: <0.0000000000000002
```

```
confint(mlr2)
```

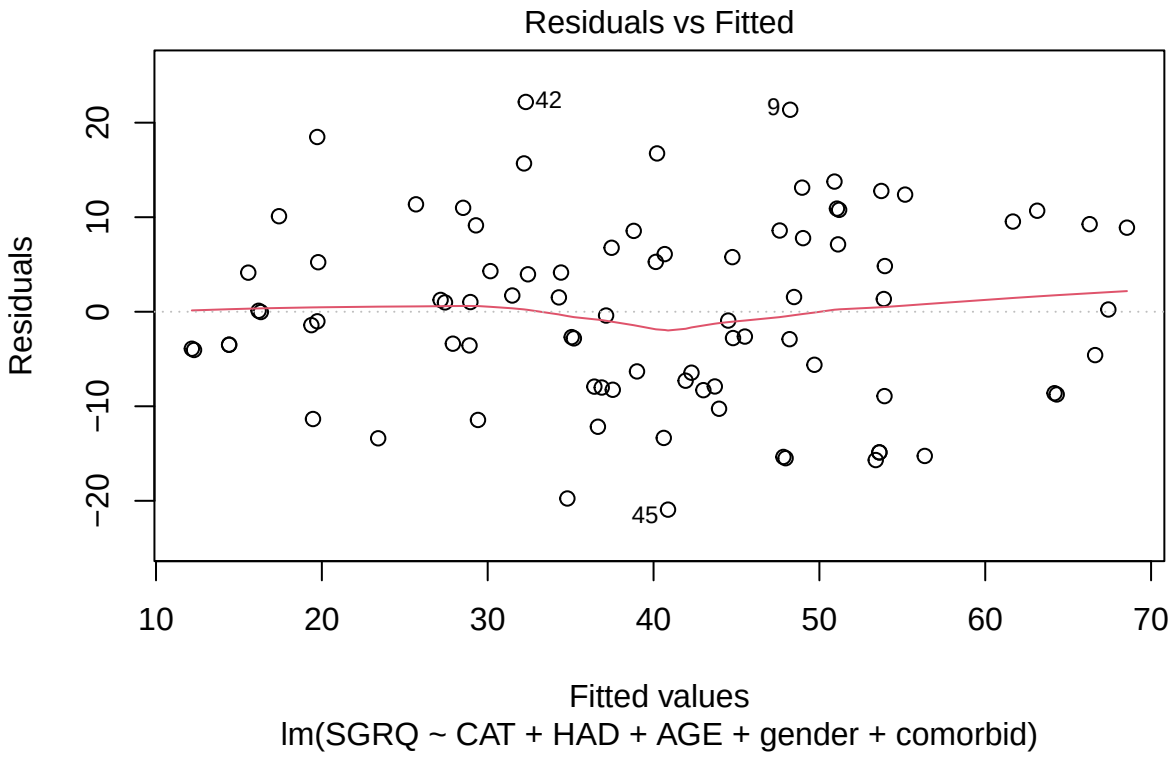
```
##                  2.5 %    97.5 %
## (Intercept) -10.702143 43.808849
## CAT           1.090147  1.742682
## HAD           0.221950  0.980759
## AGE          -0.510347  0.246743
## gender1      -4.237236  5.142568
## comorbid1    -2.738404  6.149102
```

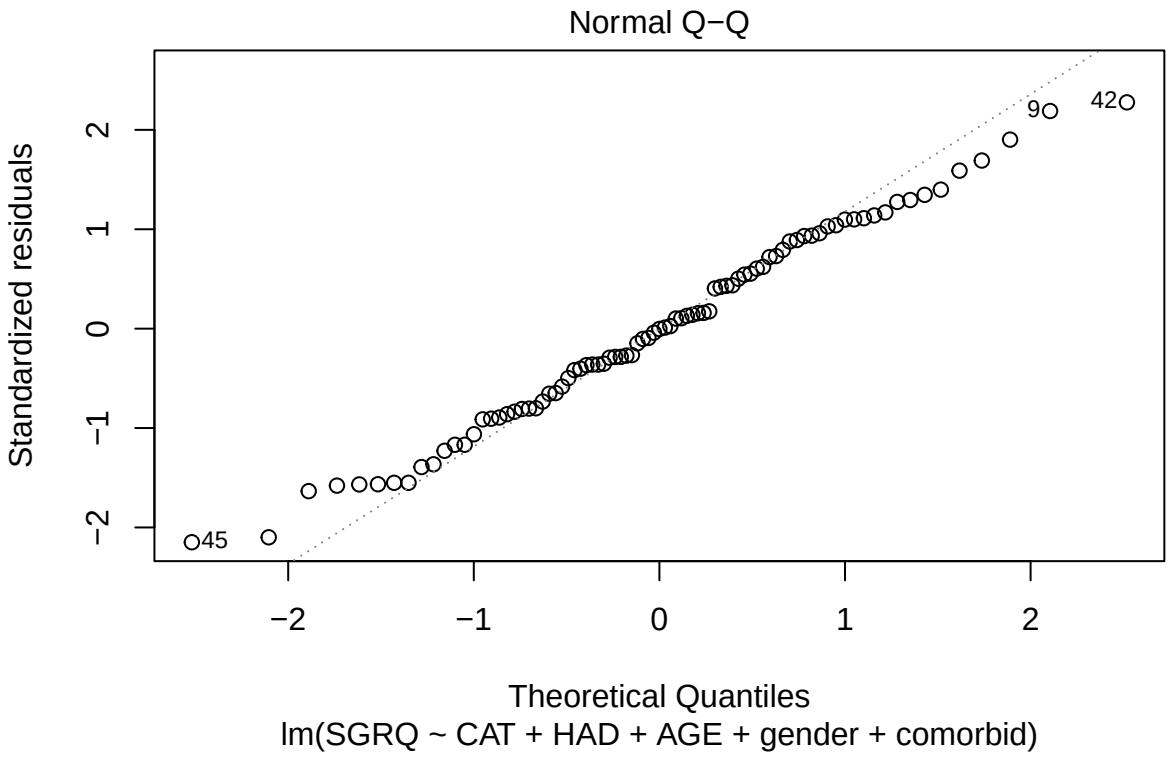
it the model :

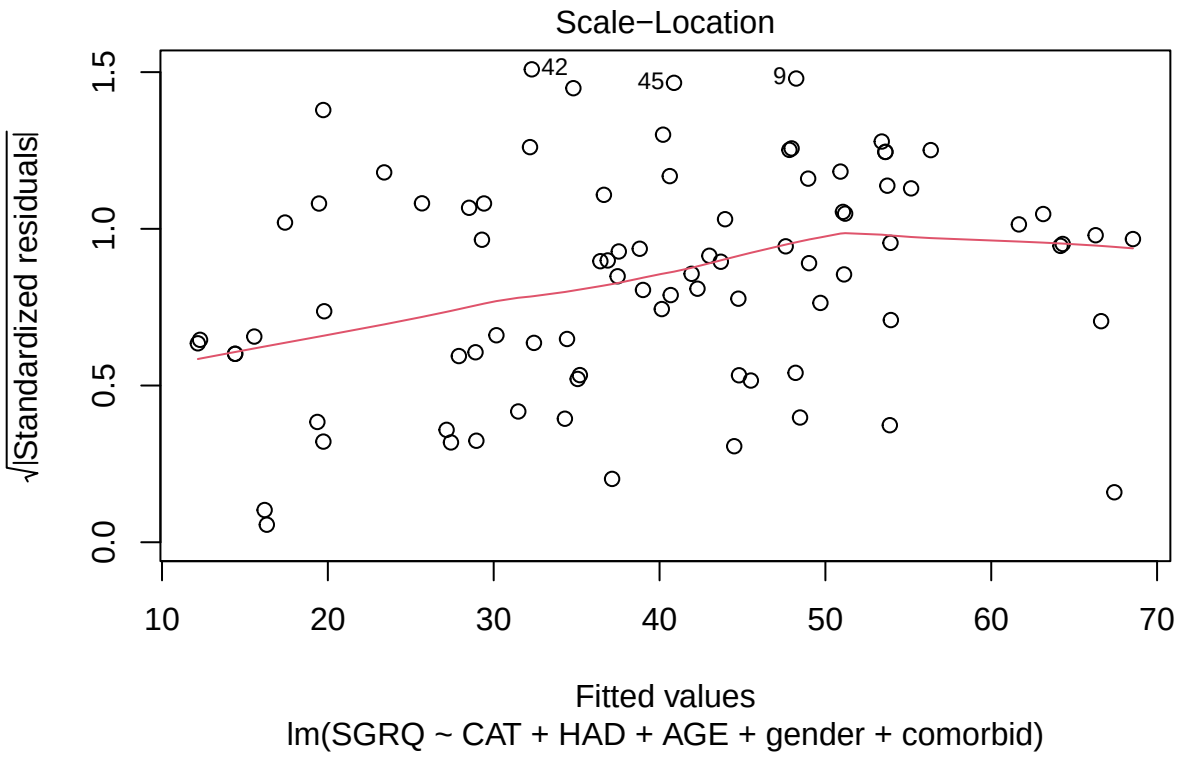
```
predictedsgrqmodel9 <- predict(mlr2)
residualsgrqmodel9 <- residuals(mlr2)
```

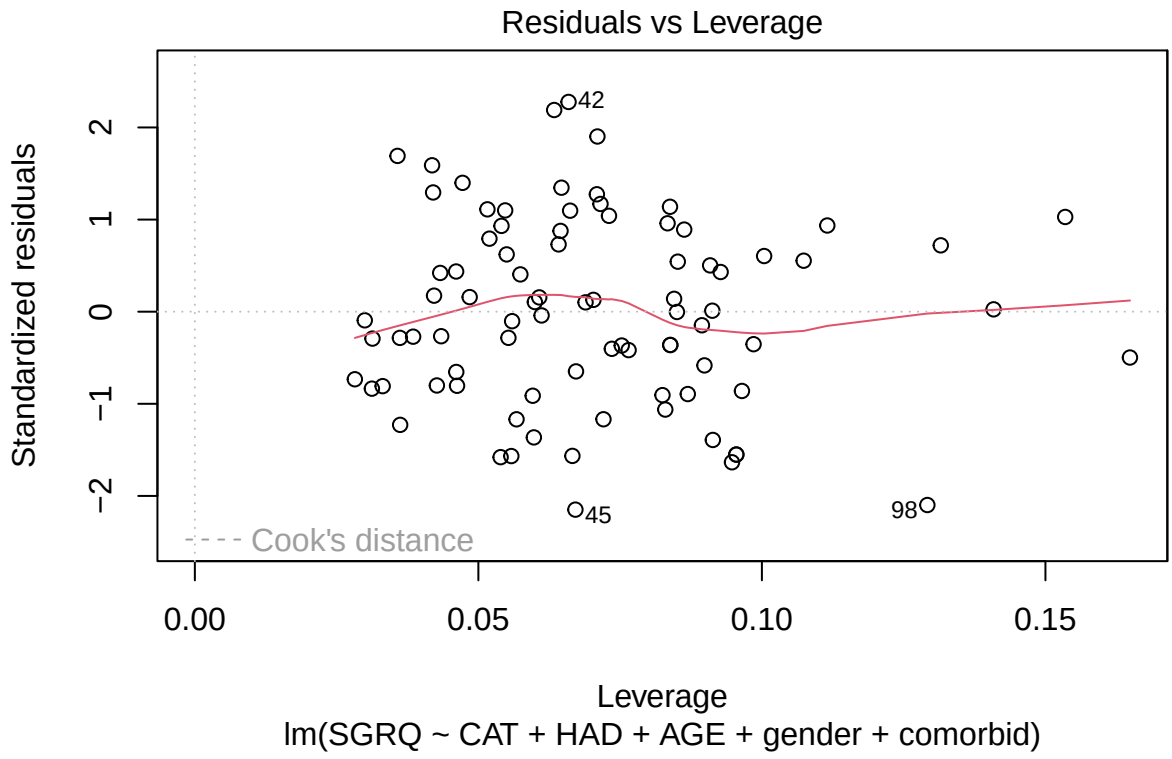
Check using plots :

```
plot(mlr2)
```


Residuals vs Fitted
Residuals
Fitted values
lm(SGRQ ~ CAT + HAD + AGE + gender + comorbid)
42
9
45
20
10
0
−10
−20
10
20
30
40
50
60
70


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(SGRQ ~ CAT + HAD + AGE + gender + comorbid)
45
9
42
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
42
45
9
10
20
30
40
50
60
70
Fitted values
lm(SGRQ ~ CAT + HAD + AGE + gender + comorbid)




```
mlr3 <- lm(SGRQ~CAT+HAD+AGE+MWT1Best+gender+comorbid+Diabetes+hypertension+AtrialFib+IHD, data=subset_c
```

```
summary(mlr3)
```

```
##
## Call:
## lm(formula = SGRQ ~ CAT + HAD + AGE + MWT1Best + gender + comorbid +
##     Diabetes + hypertension + AtrialFib + IHD, data = subset_copd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.114  -8.087   0.266   6.938  22.031
##
## Coefficients:
##                Estimate Std. Error t value      Pr(>|t|)
## (Intercept)    22.0857    20.7796    1.06         0.291
## CAT             1.4048     0.1890    7.43 0.00000000015 ***
## HAD             0.4868     0.2083    2.34         0.022 *
## AGE            -0.1287     0.2165   -0.59         0.554
## MWT1Best       -0.0103     0.0186   -0.55         0.581
## gender1         0.2951     2.4740    0.12         0.905
## comorbid1      -2.3928     3.3695   -0.71         0.480
## Diabetes1       3.3590     3.5847    0.94         0.352
## hypertension1   6.1044     4.5434    1.34         0.183
## AtrialFib1      3.0399     3.9058    0.78         0.439
```

```
## IHD1            1.7620     4.6223    0.38         0.704
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.1 on 74 degrees of freedom
## Multiple R-squared:  0.703,  Adjusted R-squared:  0.663
## F-statistic: 17.5 on 10 and 74 DF,  p-value: 0.000000000000000798
```

```
confint(mlr3)
```

```
##                     2.5 %     97.5 %
## (Intercept)   -19.3185149 63.4898813
## CAT             1.0282178  1.7813421
## HAD             0.0717180  0.9019240
## AGE            -0.5601174  0.3027948
## MWT1Best       -0.0473184  0.0267075
## gender1        -4.6343252  5.2246148
## comorbid1      -9.1066644  4.3211064
## Diabetes1      -3.7836615 10.5015959
## hypertension1  -2.9485815 15.1573845
## AtrialFib1     -4.7426744 10.8224216
## IHD1           -7.4482571 10.9721992
```

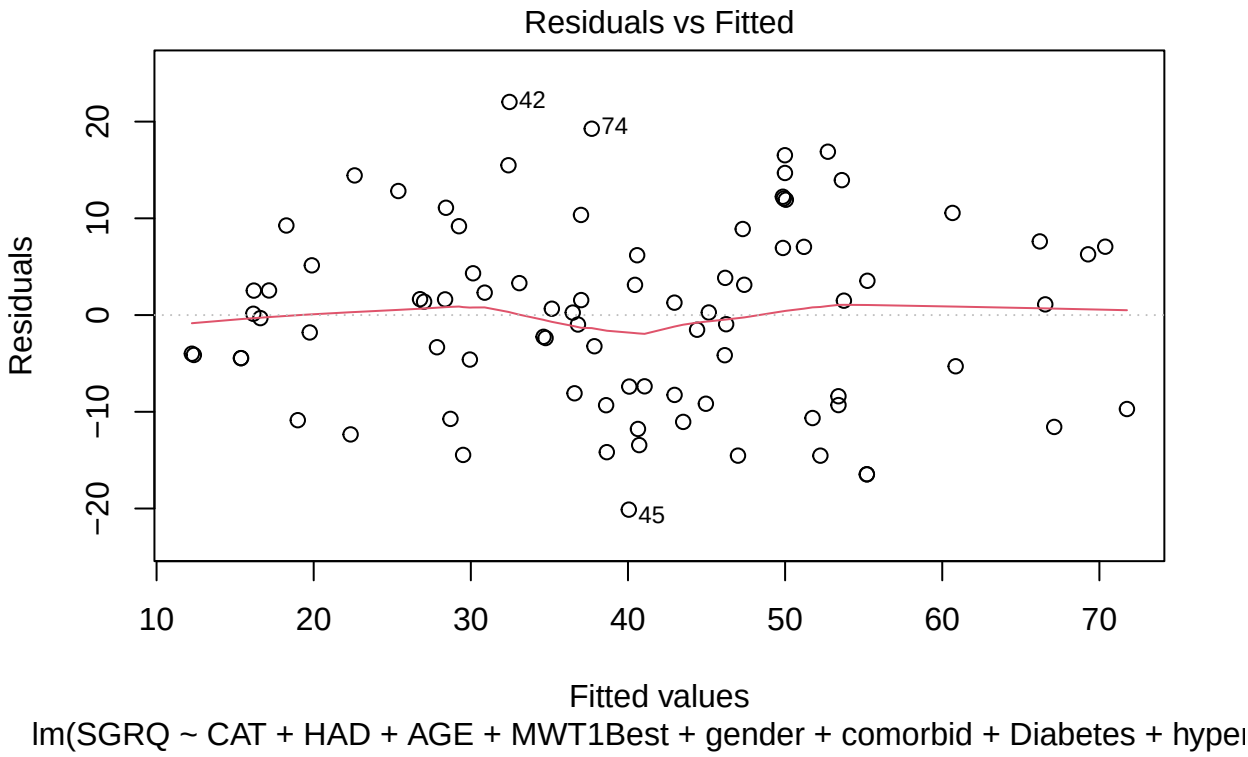
Adding all comorbidities, the Rsquared value increased to 0.703 and Diabetes and hypertension are two predictors that said dignificant.

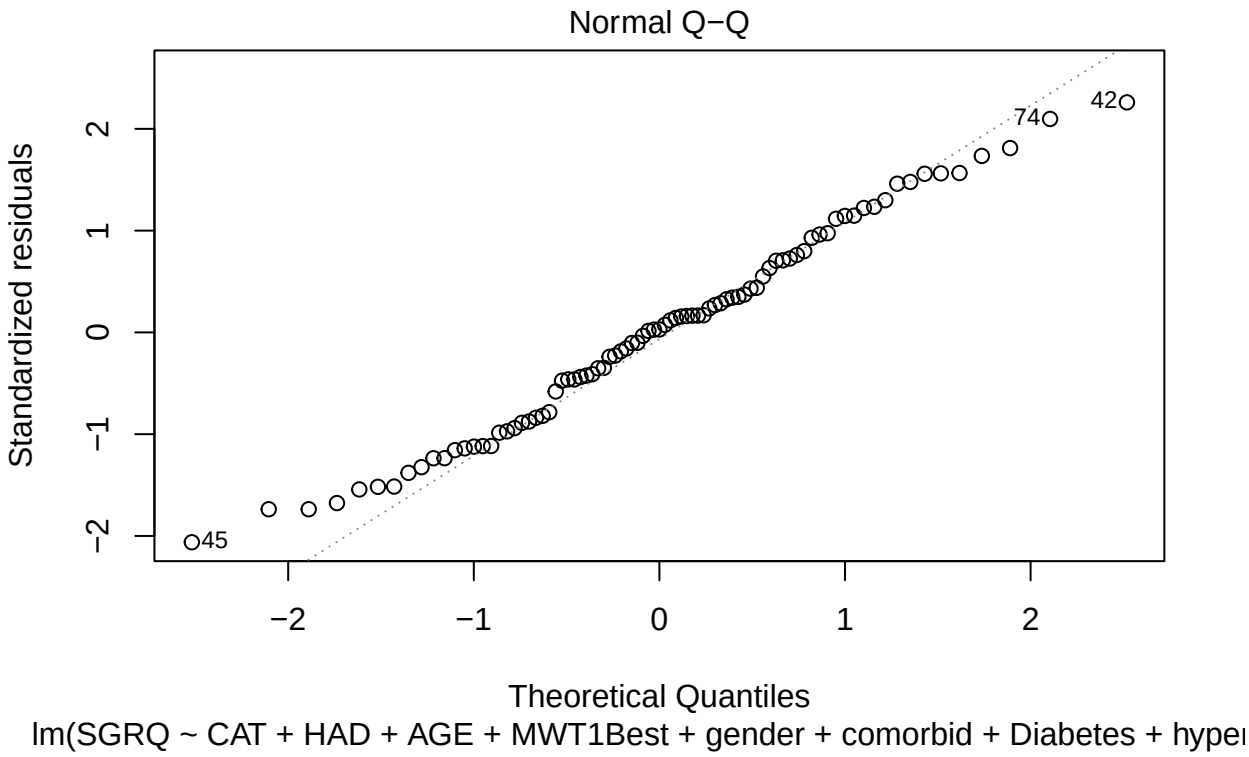
Fit the model :

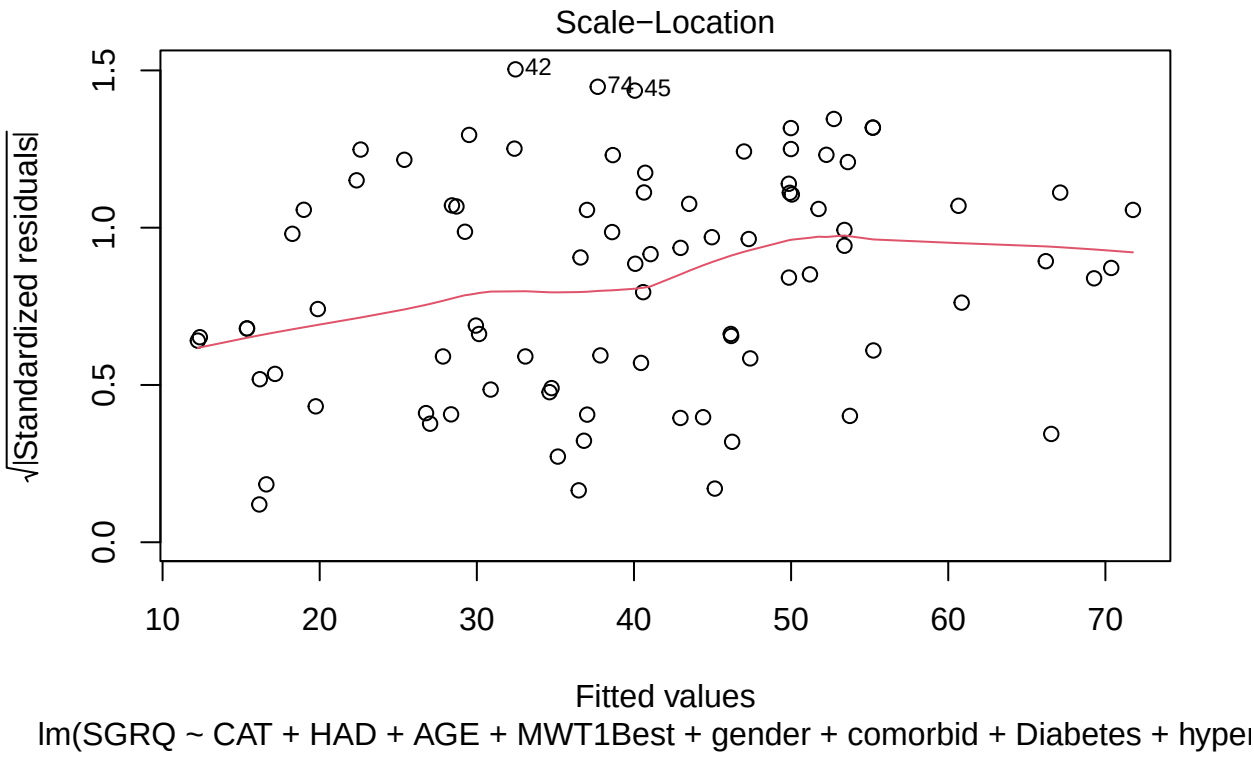
```
predictedmlr3 <- predict(mlr3)
residualsmlr3 <- residuals(mlr3)
```

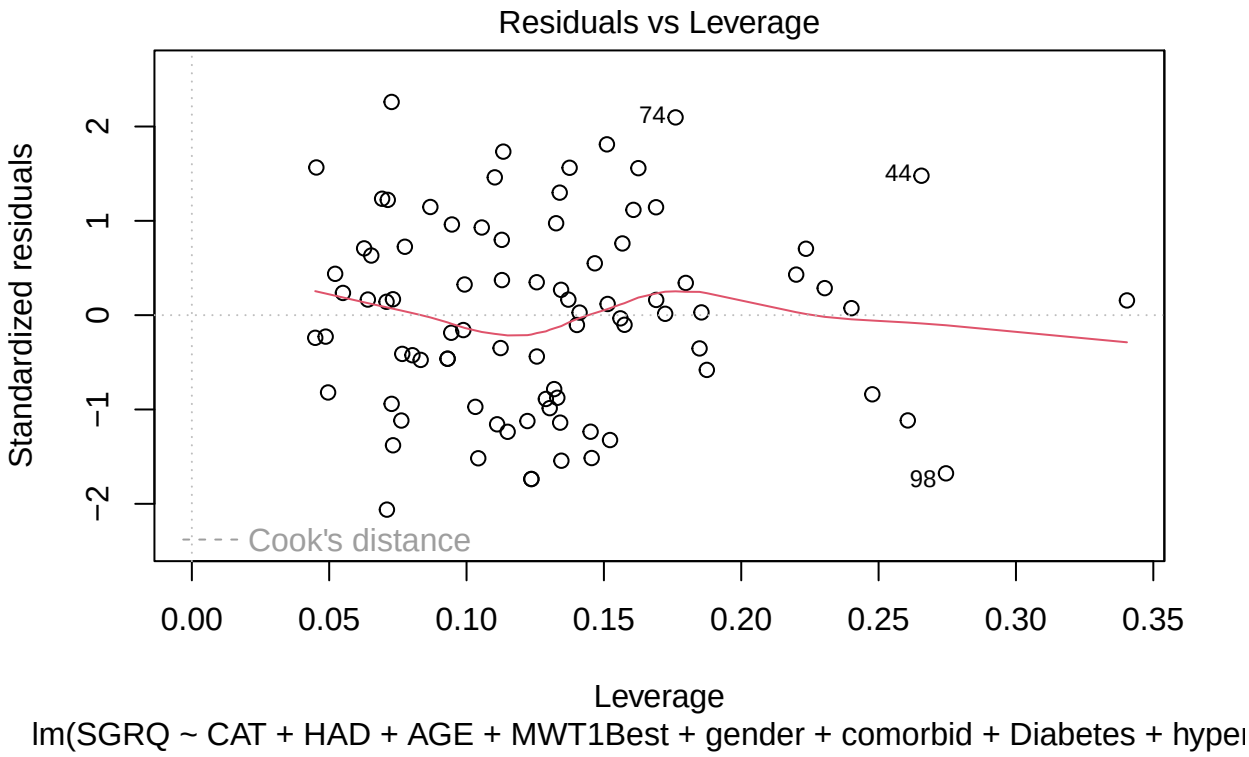
Check using plots :

```
plot(mlr3)
```


Residuals vs Fitted
Residuals
Fitted values
lm(SGRQ ~ CAT + HAD + AGE + MWT1Best + gender + comorbid + Diabetes + hype
42
74
45
20
10
0
−10
−20
10
20
30
40
50
60
70


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(SGRQ ~ CAT + HAD + AGE + MWT1Best + gender + comorbid + Diabetes + hype
45
74
42
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
42
74
45
10
20
30
40
50
60
70
Fitted values
lm(SGRQ ~ CAT + HAD + AGE + MWT1Best + gender + comorbid + Diabetes + hype




```
imcdiag(mlr3)
```

```
##
## Call:
## imcdiag(mod = mlr3)
##
##
## All Individual Multicollinearity Diagnostics Result
##
##                   VIF   TOL     Wi     Fi Leamer   CVIF Klein
## CAT             1.911 0.523  7.589  8.651  0.723 -1.385     0
## HAD             1.810 0.552  6.751  7.696  0.743 -1.312     0
## AGE             1.373 0.728  3.106  3.541  0.853 -0.995     0
## MWT1Best        2.858 0.350 15.485 17.653  0.591 -2.071     0
## gender1         1.176 0.850  1.470  1.676  0.922 -0.852     0
## comorbid1       2.328 0.429 11.069 12.619  0.655 -1.687     0
## Diabetes1       1.706 0.586  5.880  6.704  0.766 -1.236     0
## hypertension1 1.621 0.617  5.177  5.902  0.785 -1.175     0
## AtrialFib1      2.197 0.455  9.972 11.368  0.675 -1.592     0
## IHD1            1.340 0.747  2.829  3.225  0.864 -0.971     0
##                  IND1  IND2
## CAT             0.063 1.145
## HAD             0.066 1.075
## AGE             0.087 0.653
## MWT1Best        0.042 1.562
```

```
## gender1       0.102 0.360
## comorbid1     0.052 1.371
## Diabetes1     0.070 0.994
## hypertension1 0.074 0.921
## AtrialFib1    0.055 1.309
## IHD1          0.090 0.609
##
## 1 --> COLLINEARITY is detected by the test
## 0 --> COLLINEARITY is not detected by the test
##
## AGE , MWT1Best , gender1 , comorbid1 , Diabetes1 , hypertension1 , AtrialFib1 , IHD1 , coefficient(s)
##
## R-square of y on all x: 0.703
##
## * use method argument to check which regressors may be the reason of collinearity
## ===================================
```

Despite increase in multiple R-squared, the residual plot shows overfitting which means the model catch noise in the data.

```
mlr4 <- lm(SGRQ~CAT+HAD+AGE+MWT1Best+gender+Diabetes+hypertension, data=subset_copd)
```

```
summary(mlr4)
```

```
##
## Call:
## lm(formula = SGRQ ~ CAT + HAD + AGE + MWT1Best + gender + Diabetes +
##     hypertension, data = subset_copd)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -20.03  -7.90   0.51   6.87  22.05
##
## Coefficients:
##               Estimate Std. Error t value       Pr(>|t|)
## (Intercept)    27.5568    18.0352    1.53          0.131
## CAT             1.3539     0.1758    7.70 0.000000000038
## HAD             0.4964     0.1977    2.51          0.014
## AGE            -0.1583     0.2057   -0.77          0.444
## MWT1Best       -0.0172     0.0148   -1.17          0.248
## gender1         0.3338     2.4337    0.14          0.891
## Diabetes1       2.4725     3.0690    0.81          0.423
## hypertension1   4.3912     3.8268    1.15          0.255
##
## (Intercept)
## CAT           ***
## HAD           *
## AGE
## MWT1Best
## gender1
## Diabetes1
## hypertension1
## ---
```

```
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.97 on 77 degrees of freedom
## Multiple R-squared:   0.7,   Adjusted R-squared:  0.673
## F-statistic: 25.7 on 7 and 77 DF,  p-value: <0.0000000000000002
```

```
confint(mlr4)
```

```
##                     2.5 %     97.5 %
## (Intercept)   -8.3558647 63.4695086
## CAT            1.0038629  1.7039482
## HAD            0.1027397  0.8900744
## AGE           -0.5678832  0.2513042
## MWT1Best      -0.0465713  0.0121903
## gender1       -4.5123825  5.1799491
## Diabetes1     -3.6387678  8.5836725
## hypertension1 -3.2288449 12.0112906
```

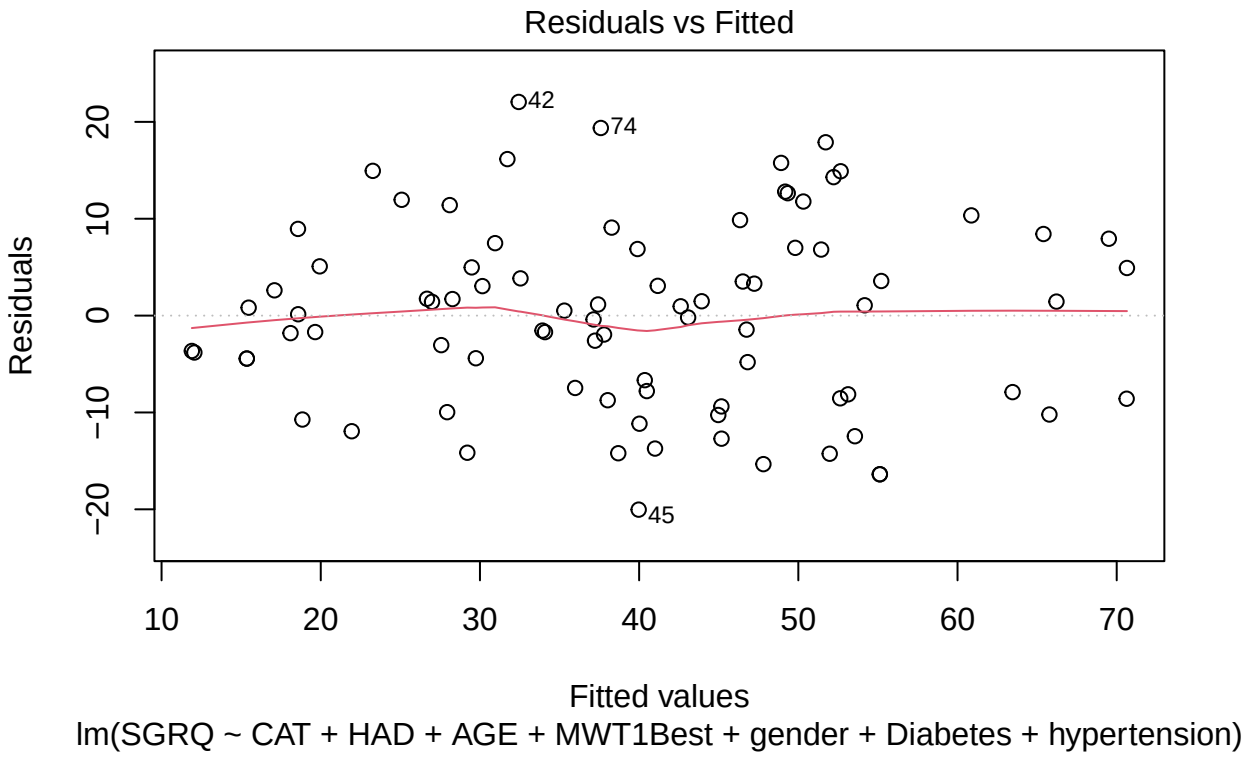
Adding all comorbidities, the Rsquared value increased to 0.7 and no significant categorical predictor.

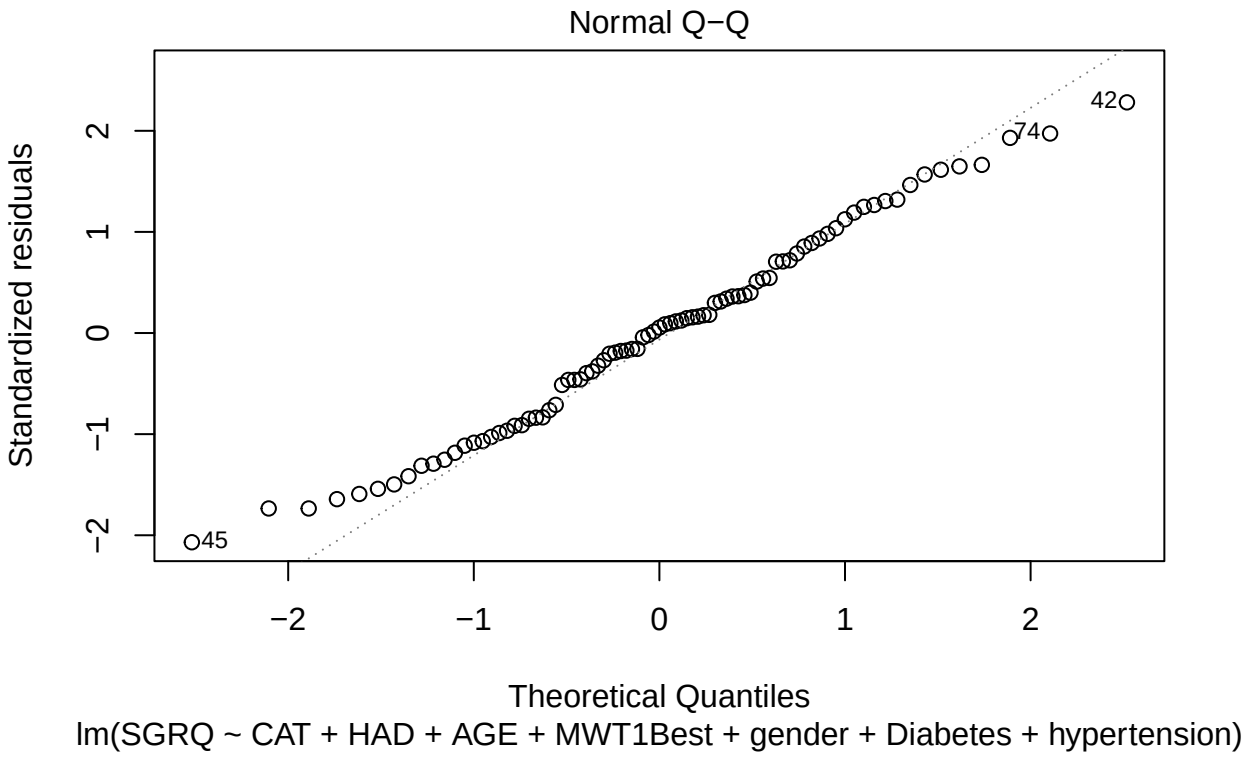
Fit the model :

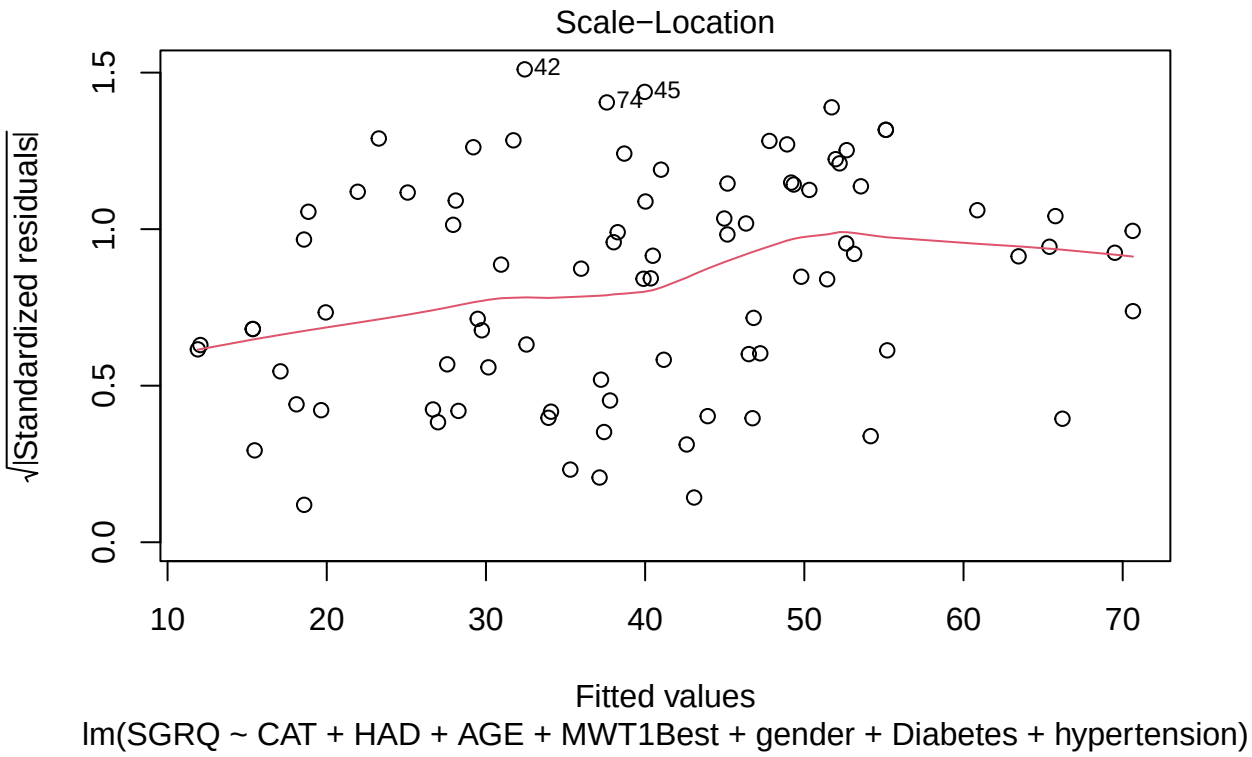
```
predictedmlr4 <- predict(mlr4)
residualsmlr4 <- residuals(mlr4)
```

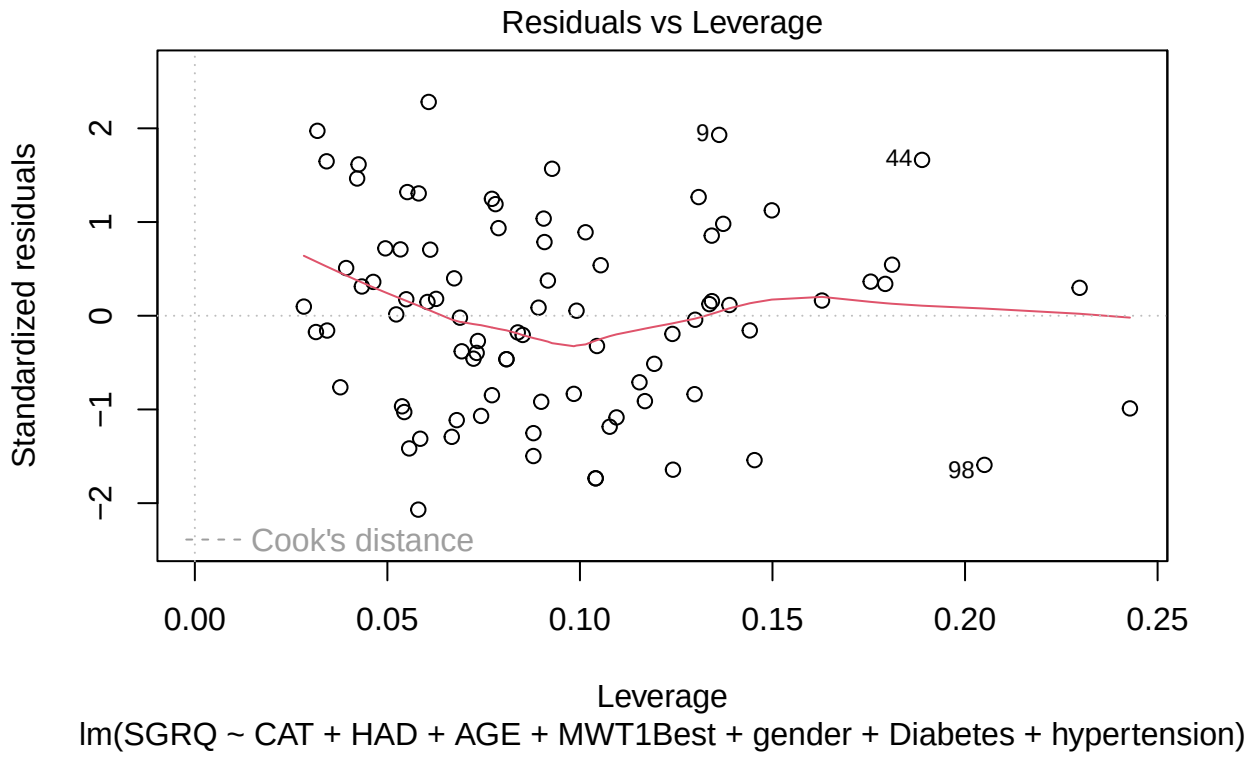
Check using plots :

```
plot(mlr4)
```


Residuals vs Fitted
Residuals
Fitted values
lm(SGRQ ~ CAT + HAD + AGE + MWT1Best + gender + Diabetes + hypertension)
42
74
45
20
10
0
−10
−20
10
20
30
40
50
60
70


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(SGRQ ~ CAT + HAD + AGE + MWT1Best + gender + Diabetes + hypertension)
45
74
42
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
42
74
45
10
20
30
40
50
60
70
Fitted values
lm(SGRQ ~ CAT + HAD + AGE + MWT1Best + gender + Diabetes + hypertension)




```
mlr5 <- lm(SGRQ~CAT+HAD+AGE+gender+Diabetes*hypertension, data=subset_copd)
```

```
summary(mlr5)
```

```
##
## Call:
## lm(formula = SGRQ ~ CAT + HAD + AGE + gender + Diabetes * hypertension,
##     data = subset_copd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.148  -7.529   0.871   7.117  22.304
##
## Coefficients: (1 not defined because of singularities)
##                           Estimate Std. Error t value
## (Intercept)                 14.806     14.367    1.03
## CAT                          1.432      0.163    8.80
## HAD                          0.554      0.192    2.89
## AGE                         -0.105      0.201   -0.52
## gender1                      0.072      2.429    0.03
## Diabetes1                    3.746      2.874    1.30
## hypertension1                4.889      3.812    1.28
## Diabetes1:hypertension1         NA         NA      NA
##                             Pr(>|t|)
## (Intercept)                    0.306
```

```
## CAT                     0.00000000000027 ***
## HAD                                0.005 **
## AGE                                0.603
## gender1                            0.976
## Diabetes1                          0.196
## hypertension1                      0.203
## Diabetes1:hypertension1               NA
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10 on 78 degrees of freedom
## Multiple R-squared:  0.695,  Adjusted R-squared:  0.671
## F-statistic: 29.6 on 6 and 78 DF,  p-value: <0.0000000000000002
```

```
confint(mlr5)
```

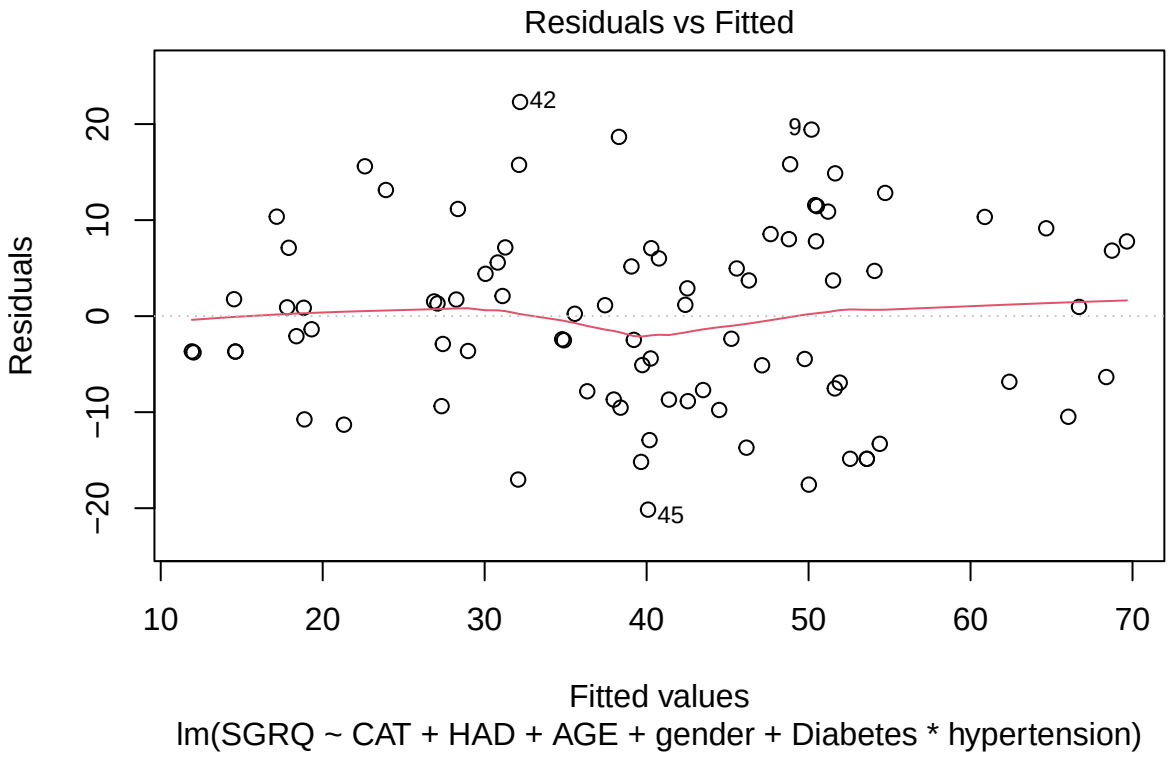
```
##                              2.5 %    97.5 %
## (Intercept)             -13.797939 43.409030
## CAT                       1.108056  1.756327
## HAD                       0.171845  0.935879
## AGE                      -0.505344  0.295095
## gender1                  -4.763556  4.907472
## Diabetes1                -1.977064  9.468189
## hypertension1            -2.698649 12.477575
## Diabetes1:hypertension1         NA        NA
```

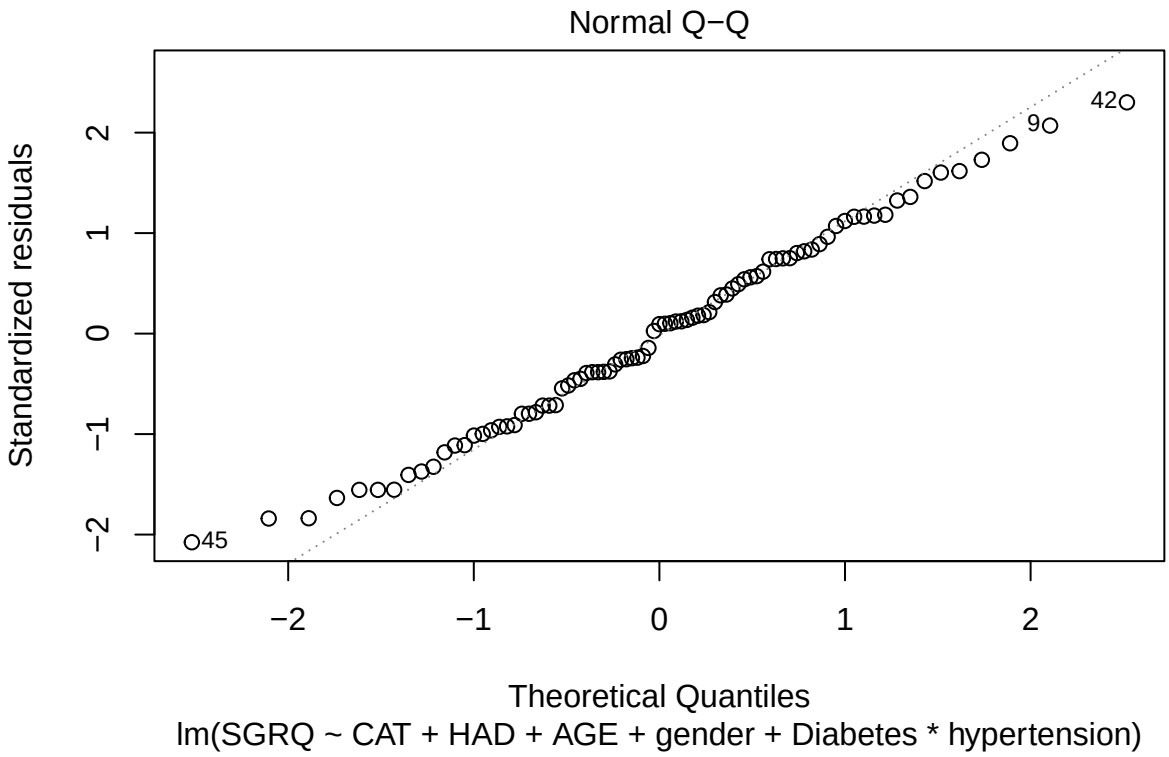
Fit the model :

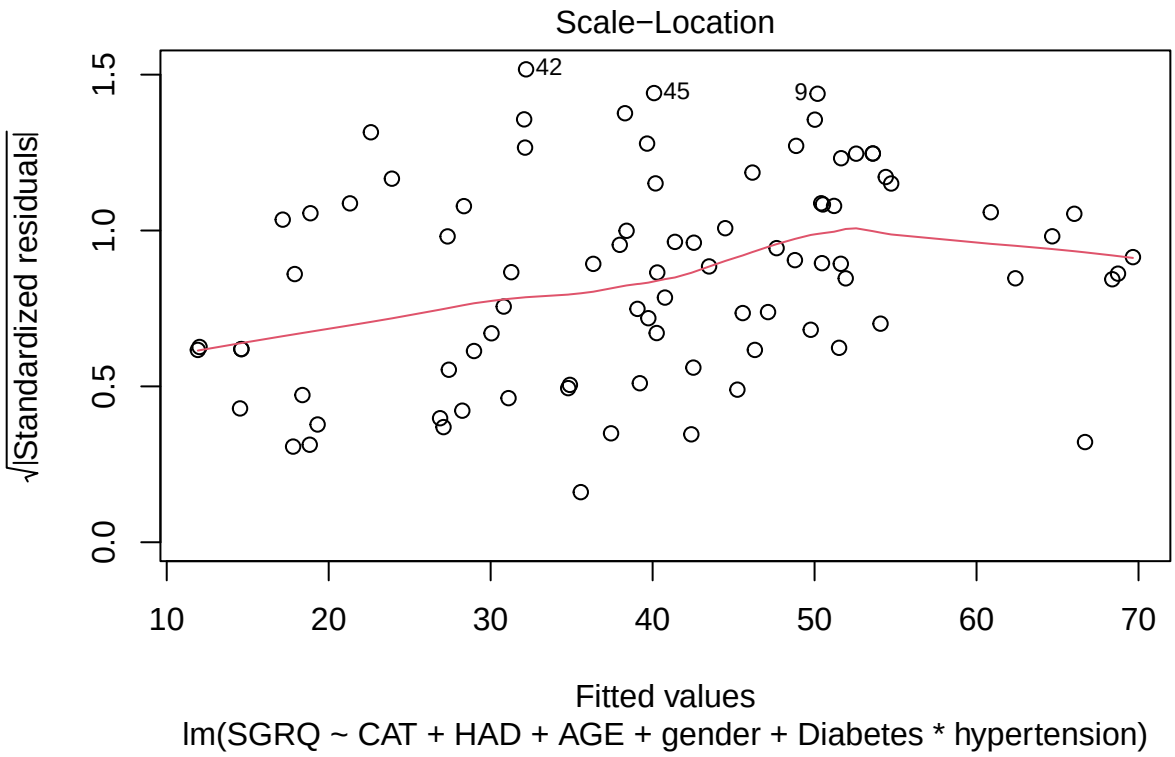
```
predictedmlr4 <- predict(mlr5)
residualsmlr4 <- residuals(mlr5)
```

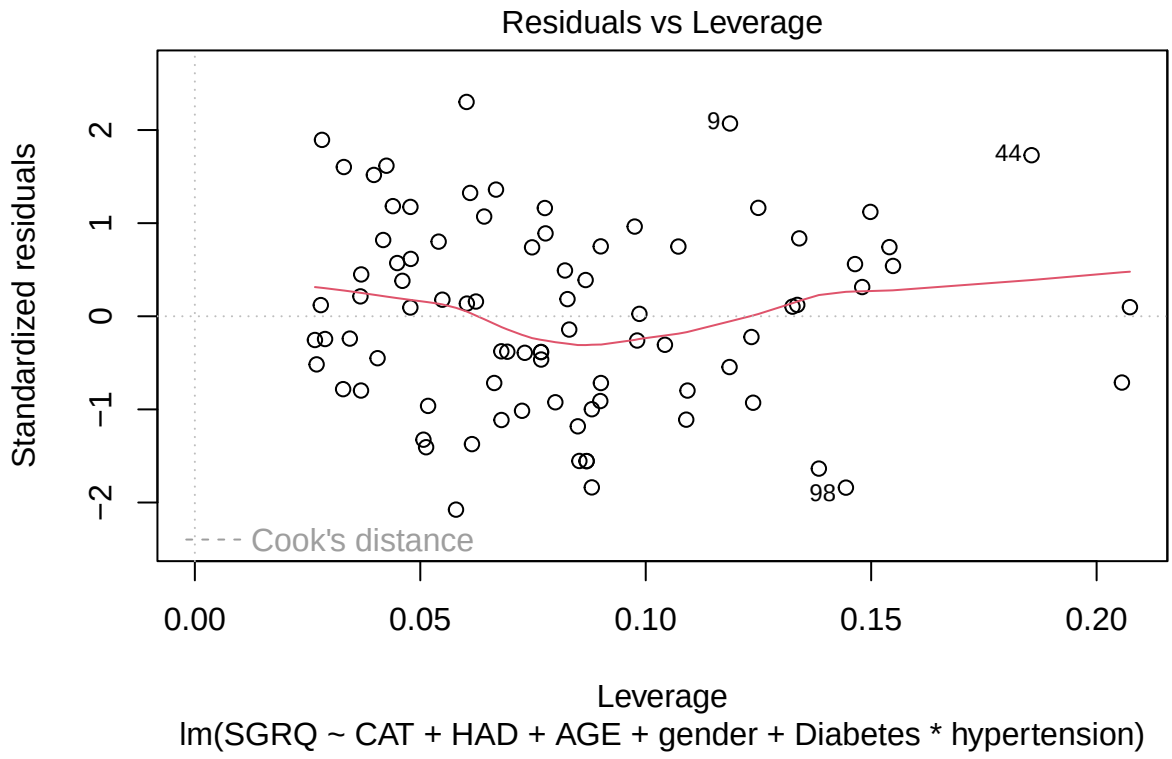
Check using plots :

```
plot(mlr5)
```


Residuals vs Fitted
Residuals
Fitted values
lm(SGRQ ~ CAT + HAD + AGE + gender + Diabetes * hypertension)
42
9
45
20
10
0
−10
−20
10
20
30
40
50
60
70


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(SGRQ ~ CAT + HAD + AGE + gender + Diabetes * hypertension)
45
9
42
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
42
45
9
10
20
30
40
50
60
70
Fitted values
lm(SGRQ ~ CAT + HAD + AGE + gender + Diabetes * hypertension)




```
imcdiag(mlr5)
```

```
##
## Call:
## imcdiag(mod = mlr5)
##
##
## All Individual Multicollinearity Diagnostics Result
##
##                             VIF   TOL    Wi    Fi Leamer
## CAT                       1.454 0.688 5.898 7.168  0.829
## HAD                       1.574 0.635 7.465 9.073  0.797
## AGE                       1.213 0.824 2.768 3.364  0.908
## gender1                   1.162 0.860 2.111 2.566  0.928
## Diabetes1                 1.124 0.889 1.615 1.963  0.943
## hypertension1             1.170 0.855 2.205 2.680  0.925
## Diabetes1:hypertension1     NaN   NaN   NaN   NaN     NA
##                            CVIF Klein  IND1 IND2
## CAT                      -6.326     0 0.044  NaN
## HAD                      -6.850     0 0.040  NaN
## AGE                      -5.278     0 0.052  NaN
## gender1                  -5.058     0 0.054  NaN
## Diabetes1                -4.892     0 0.056  NaN
## hypertension1            -5.090     0 0.054  NaN
## Diabetes1:hypertension1    NaN    NA   NaN  NaN
```

```
##
## 1 --> COLLINEARITY is detected by the test
## 0 --> COLLINEARITY is not detected by the test
##
## AGE , gender1 , Diabetes1 , hypertension1 , coefficient(s) are non-significant may be due to multicol
##
## R-square of y on all x: 0.695
##
## * use method argument to check which regressors may be the reason of collinearity
## ===================================
```