mice_eg_with_hipdata.Rmd

Marta Pineda-Moncusi

2022-11-11

rm(list=ls())

Tutorial for multiple imputation using mice package

Introduction text: seven steps

#**0. Load and evaluate the data** We are going to use the hip.RData:

The hip dataset contains information on seven hundred and eight patients receiving primary hip replacement surgery for osteoarthritis (variable id is the unique patient identifier). Prior to the operation, patients completed a pre-operative Oxford Hip Score (OHS) and EQ5D (Euroqol) questionnaire with a follow-up questionnaire being filled in at 6-months post-surgery. The OHS consists of 12 questions asking patients to describe their hip pain and function during the past 4 weeks. An overall score is created by summing the responses to each of the 12 questions, ranging from 0 to 48, where 0 is the worst possible score (severe symptoms) and 48 the best score (excellent joint function). Variable ohs0 is the preoperative score and ohs6 post-operative. The absolute change in OHS between preand post-operative assessments (variable ohsdiff) is negative if patient symptoms have improved and positive for worsening.

The pre-operative EQ5D contains information from 5 questions asking about a patient's health state today, covering mobility, self-care, usual activities, pain, and anxiety. The EQ5D has been converted to a single summary score (variable EQ5D0), anchored at 0 for death and 1 for full health, with some health states being worse than dead (-0.594).

Six months after their operation patients were asked about their overall satisfaction with the outcome of surgery measured on a visual analogue scale from 0 to 100 (variable satisfaction).

Pre-operative information was collected on age at the time of surgery, sex (0 = Male; 1 = Female), height (metres) and weight (kg) (from which body mass index (bmi) is calculated), side of surgery (Left; Right), ethnic group (0 = White; 1 = Non white), whether or not they are retired (0 = Not retired; 1 = Retired). The Index of Multiple Deprivation is a measure of social deprivation, linked to the area a patient lives in (variable imdscore).

OUTCOME OF THE STUDY: improvement in the ohs (i.e., ohsdiff > 0)

```
#load data
library(haven)
```

```
## Warning: package 'haven' was built under R version 4.1.3
```

```
data <- read_dta('C:/Users/martapm/Documents/GitHub/Practical_Imputation/multiple-imputation_with_mice/
#transform/include labels to data
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.1.3
```

```
##
## 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
```

```
data = data %>%
            mutate(sex = factor(sex, levels = c(0,1), labels = c("Male","Female")),
                 ethnic = factor(ethnic, levels = c(0,1), labels = c("White","Non white")),
                retired = factor(retired, levels = c(0,1), labels = c("Not retired","Retired")),
                improve = ifelse(ohsdiff<=0,0,1 ) #"NO","YES"
              )


#data$improve[data$ohsdiff<=0]<-0 # 0 if not improved
#data$improve[data$ohsdiff>0]<-1 # 1 if improved

#Observe data
dim(data)
```

```
## [1] 708  17
```

```
summary(data)
```

```
##        id              sex           age               retired         ohs0
##  Min.   :  1.0   Male  :271   Min.   :31.00   Not retired:109   Min.   : 0.00
##  1st Qu.:177.8   Female:436   1st Qu.:66.00   Retired    :145   1st Qu.:13.00
##  Median :354.5   NA's  :  1   Median :75.00   NA's       :454   Median :20.00
##  Mean   :354.5                Mean   :73.52                     Mean   :20.03
##  3rd Qu.:531.2                3rd Qu.:82.00                     3rd Qu.:26.00
##  Max.   :708.0                Max.   :99.00                     Max.   :46.00
##
##       ohs6          ohsdiff          EQ5D0             height
##  Min.   : 6.00   Min.   :-43.0   Min.   :-0.5940   Min.   :  1.080
##  1st Qu.:33.00   1st Qu.:-26.0   1st Qu.: 0.0550   1st Qu.:  1.610
##  Median :42.00   Median :-19.0   Median : 0.5160   Median :  1.660
##  Mean   :38.63   Mean   :-18.6   Mean   : 0.3804   Mean   :  6.993
##  3rd Qu.:46.00   3rd Qu.:-12.0   3rd Qu.: 0.6910   3rd Qu.:  1.740
##  Max.   :48.00   Max.   : 19.0   Max.   : 1.0000   Max.   :183.000
##                                  NA's   :18        NA's   :359
##      weight       satisfaction         bmi           bmi_imputed
##  Min.   : 44.0   Min.   :  0.00   Min.   : 0.0022   Min.   : 0.00225
##  1st Qu.: 65.0   1st Qu.: 90.00   1st Qu.:23.4509   1st Qu.:23.26808
##  Median : 74.0   Median :100.00   Median :26.3465   Median :26.68621
##  Mean   : 76.1   Mean   : 89.22   Mean   :26.4614   Mean   :26.74399
```

```
##  3rd Qu.: 86.0   3rd Qu.:100.00   3rd Qu.:30.3692   3rd Qu.:30.82529
##  Max.   :186.0   Max.   :100.00   Max.   :63.4431   Max.   :63.44307
##  NA's   :359     NA's   :52       NA's   :359
##        ethnic         side              imdscore        improve
##  White    :300   Length:708         Min.   : 0.92   Min.   :0.00000
##  Non white: 53   Class :character   1st Qu.: 7.07   1st Qu.:0.00000
##  NA's     :355   Mode  :character   Median :11.07   Median :0.00000
##                                      Mean   :13.82   Mean   :0.04379
##                                      3rd Qu.:18.86   3rd Qu.:0.00000
##                                      Max.   :48.05   Max.   :1.00000
##                                      NA's   :11
```

```
#str(data)
```

Check the structure of the data: - How many cases (individuals) are in the data? 708 - How many variables do we have? 17 - Which are numeric? age, oh0, oh6, ohsdiff, EQ5D0, height, weight, satisfaction, bmi, imdscore - Which are factor? improve (but hasn't been formatted on propose), sex, retired, ethnic, side - Is there any variable with missing? EQ5D0, height, weight, bmi, satisfaction, ethnic, imdscore

#**1. Check that variables with missing data are associated with variables that are complete** To do so, crate a dummy variable were missing values are 1 and complete values are 0, and run a logistic regression.

```
#list the variables you want to test
test_var_list = c("EQ5D0", "height", "weight", "bmi", "satisfaction", "ethnic", "imdscore")
complete_var  = c("age", "sex", "retired", "ohs0","ohs6", "side") #, "improve") #Q: Since the outcome (

for (v in 1:length(test_var_list)){
 test_var = ifelse(is.na(data[[test_var_list[v]]]), 1,0 )
 test = glm(test_var ~ ., data=data[complete_var] )
  print(paste("Tested variable: ", test_var_list[v], sep=""))
  print(summary(test))   #Check p_value
}
```

```
## [1] "Tested variable: EQ5D0"
##
## Call:
## glm(formula = test_var ~ ., data = data[complete_var])
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.07917  -0.04793  -0.03600  -0.01848   0.99077
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.069143   0.219033  -0.316    0.753
## age             0.000399   0.001335   0.299    0.765
## sexFemale       0.003063   0.025045   0.122    0.903
## retiredRetired  0.025519   0.027099   0.942    0.347
## ohs0           -0.001337   0.001420  -0.942    0.347
## ohs6            0.000557   0.001413   0.394    0.694
## sideLeft        0.066388   0.188608   0.352    0.725
## sideRight       0.058065   0.188724   0.308    0.759
##
```

```
## (Dispersion parameter for gaussian family taken to be 0.03488539)
##
##     Null deviance: 8.6811  on 253  degrees of freedom
## Residual deviance: 8.5818  on 246  degrees of freedom
##   (454 observations deleted due to missingness)
## AIC: -121.65
##
## Number of Fisher Scoring iterations: 2
##
## [1] "Tested variable: height"
##
## Call:
## glm(formula = test_var ~ ., data = data[complete_var])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8914  -0.5306   0.3080   0.4250   0.6764
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.545667   0.575730   2.685  0.00775 **
## age            -0.007817   0.003509  -2.228  0.02679 *
## sexFemale      -0.081315   0.065832  -1.235  0.21794
## retiredRetired  0.136713   0.071229   1.919  0.05610 .
## ohs0           -0.005346   0.003733  -1.432  0.15339
## ohs6            0.002202   0.003714   0.593  0.55375
## sideLeft       -0.401278   0.495755  -0.809  0.41905
## sideRight      -0.379217   0.496061  -0.764  0.44533
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.2410242)
##
##     Null deviance: 62.224  on 253  degrees of freedom
## Residual deviance: 59.292  on 246  degrees of freedom
##   (454 observations deleted due to missingness)
## AIC: 369.29
##
## Number of Fisher Scoring iterations: 2
##
## [1] "Tested variable: weight"
##
## Call:
## glm(formula = test_var ~ ., data = data[complete_var])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8914  -0.5306   0.3080   0.4250   0.6764
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.545667   0.575730   2.685  0.00775 **
## age            -0.007817   0.003509  -2.228  0.02679 *
## sexFemale      -0.081315   0.065832  -1.235  0.21794
```

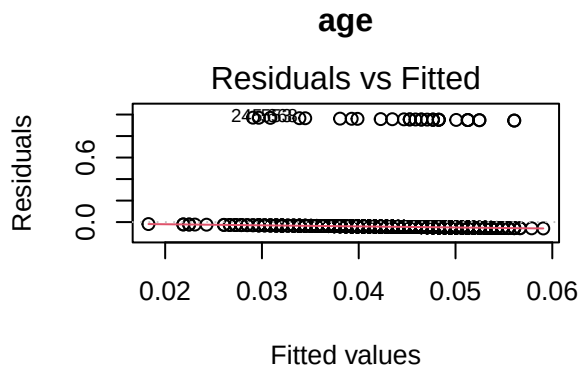
```
## retiredRetired  0.136713   0.071229   1.919  0.05610 .
## ohs0           -0.005346   0.003733  -1.432  0.15339
## ohs6            0.002202   0.003714   0.593  0.55375
## sideLeft       -0.401278   0.495755  -0.809  0.41905
## sideRight      -0.379217   0.496061  -0.764  0.44533
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.2410242)
##
##     Null deviance: 62.224  on 253  degrees of freedom
## Residual deviance: 59.292  on 246  degrees of freedom
##   (454 observations deleted due to missingness)
## AIC: 369.29
##
## Number of Fisher Scoring iterations: 2
##
## [1] "Tested variable: bmi"
##
## Call:
## glm(formula = test_var ~ ., data = data[complete_var])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8914  -0.5306   0.3080   0.4250   0.6764
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.545667   0.575730   2.685  0.00775 **
## age            -0.007817   0.003509  -2.228  0.02679 *
## sexFemale      -0.081315   0.065832  -1.235  0.21794
## retiredRetired  0.136713   0.071229   1.919  0.05610 .
## ohs0           -0.005346   0.003733  -1.432  0.15339
## ohs6            0.002202   0.003714   0.593  0.55375
## sideLeft       -0.401278   0.495755  -0.809  0.41905
## sideRight      -0.379217   0.496061  -0.764  0.44533
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.2410242)
##
##     Null deviance: 62.224  on 253  degrees of freedom
## Residual deviance: 59.292  on 246  degrees of freedom
##   (454 observations deleted due to missingness)
## AIC: 369.29
##
## Number of Fisher Scoring iterations: 2
##
## [1] "Tested variable: satisfaction"
##
## Call:
## glm(formula = test_var ~ ., data = data[complete_var])
##
## Deviance Residuals:
```

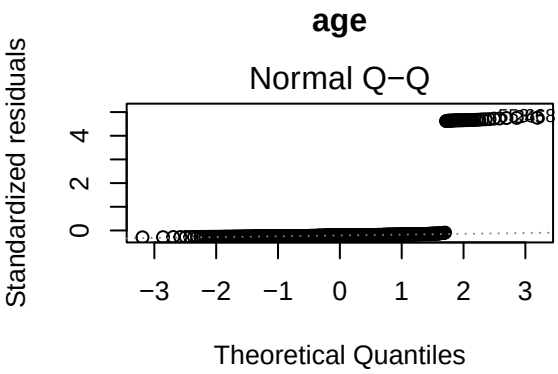
```
##      Min        1Q    Median        3Q       Max
## -0.23047  -0.11064  -0.07911  -0.05401   0.96391
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.0072187  0.3386165  -0.021   0.9830
## age             0.0021651  0.0020636   1.049   0.2951
## sexFemale      -0.0120803  0.0387191  -0.312   0.7553
## retiredRetired -0.0253641  0.0418934  -0.605   0.5454
## ohs0           -0.0006963  0.0021954  -0.317   0.7514
## ohs6           -0.0036805  0.0021842  -1.685   0.0932 .
## sideLeft        0.1133223  0.2915795   0.389   0.6979
## sideRight       0.1136654  0.2917591   0.390   0.6972
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.08337578)
##
##     Null deviance: 20.917  on 253  degrees of freedom
## Residual deviance: 20.510  on 246  degrees of freedom
##   (454 observations deleted due to missingness)
## AIC: 99.655
##
## Number of Fisher Scoring iterations: 2
##
## [1] "Tested variable: ethnic"
##
## Call:
## glm(formula = test_var ~ ., data = data[complete_var])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8816  -0.5207   0.2977   0.4223   0.6827
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.555678   0.575312   2.704  0.00733 **
## age            -0.007799   0.003506  -2.224  0.02703 *
## sexFemale      -0.087448   0.065784  -1.329  0.18497
## retiredRetired  0.145204   0.071177   2.040  0.04242 *
## ohs0           -0.005762   0.003730  -1.545  0.12365
## ohs6            0.001893   0.003711   0.510  0.61038
## sideLeft       -0.401472   0.495396  -0.810  0.41849
## sideRight      -0.371248   0.495701  -0.749  0.45461
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.2406749)
##
##     Null deviance: 62.362  on 253  degrees of freedom
## Residual deviance: 59.206  on 246  degrees of freedom
##   (454 observations deleted due to missingness)
## AIC: 368.92
##
```

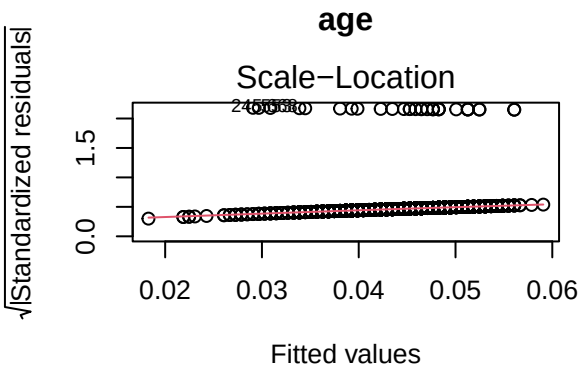
```
## Number of Fisher Scoring iterations: 2
##
## [1] "Tested variable: imdscore"
##
## Call:
## glm(formula = test_var ~ ., data = data[complete_var])
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.07577  -0.04249  -0.02553  -0.00379   0.97603
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.0689639  0.1782698   0.387   0.6992
## age            -0.0015618  0.0010864  -1.438   0.1518
## sexFemale      -0.0093973  0.0203842  -0.461   0.6452
## retiredRetired  0.0566976  0.0220554   2.571   0.0107 *
## ohs0           -0.0002142  0.0011558  -0.185   0.8532
## ohs6            0.0003515  0.0011499   0.306   0.7601
## sideLeft        0.0376007  0.1535064   0.245   0.8067
## sideRight       0.0376510  0.1536010   0.245   0.8066
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.0231089)
##
##     Null deviance: 5.8583  on 253  degrees of freedom
## Residual deviance: 5.6848  on 246  degrees of freedom
##   (454 observations deleted due to missingness)
## AIC: -226.26
##
## Number of Fisher Scoring iterations: 2
```

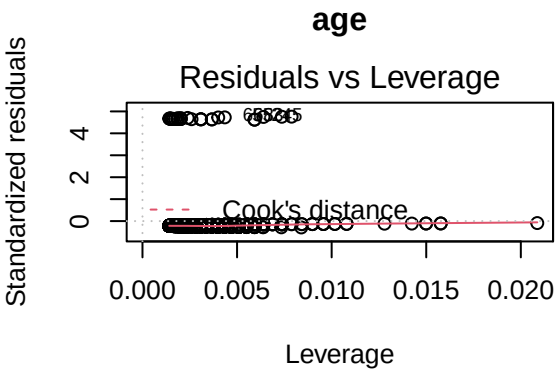
2. Check that your numeric variables are linear with your outcome

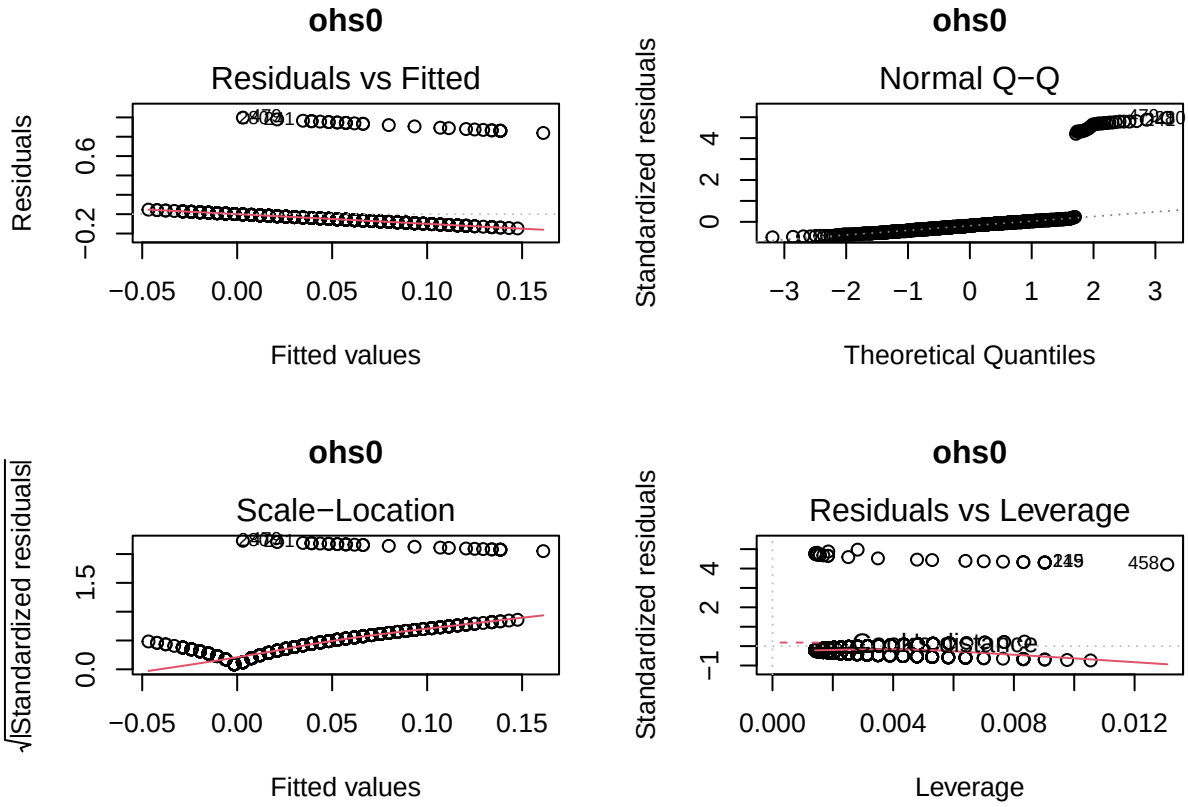
```
#list the variables you want to test
test_var_list = c("age", "ohs0", "ohs6", "ohsdiff", "EQ5D0", "height", "weight", "satisfaction", "bmi",
outcome = "improve"

for (v in 1:length(test_var_list)){
  df = cbind(data[test_var_list[v]],outcome = data[[outcome]])
  #df= as.data.frame(df) #remove tibble
  test = lm(outcome ~ ., data = df)
  par(mfrow = c(2, 2))
  plot(test,main=paste(test_var_list[v],"\n",sep=""))
}
```

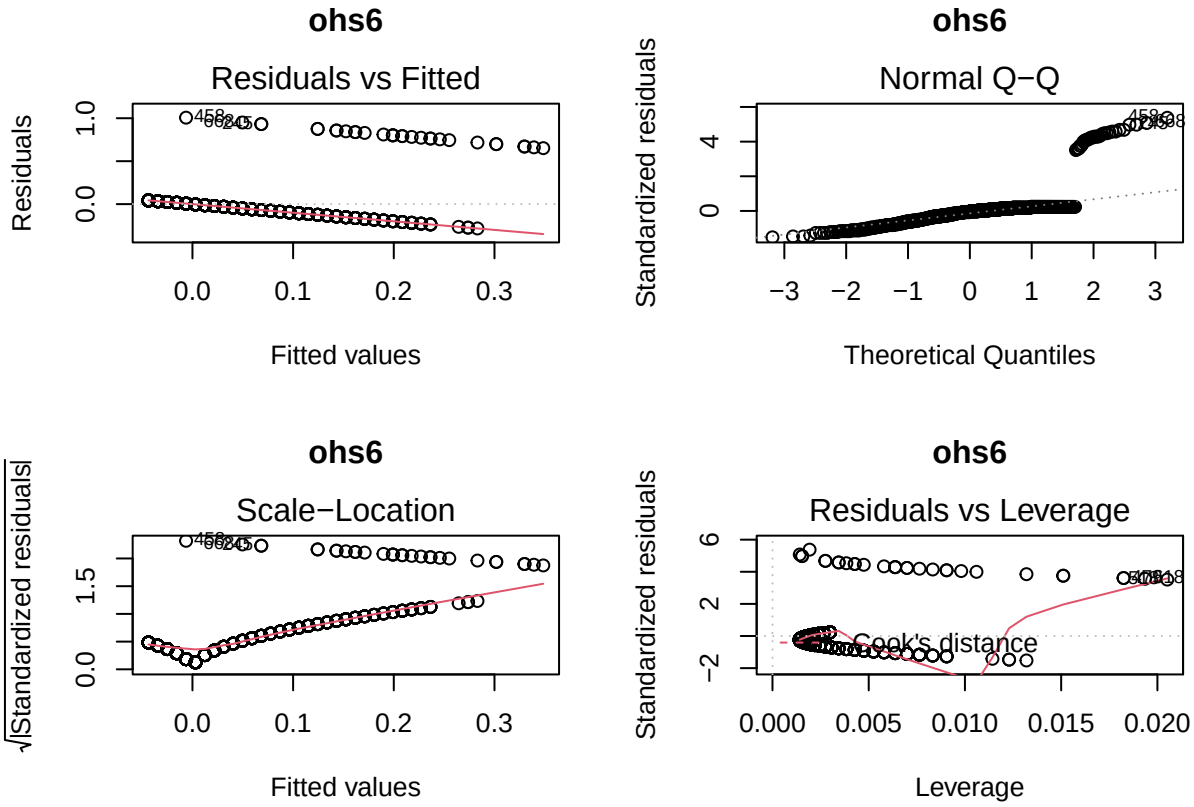

age
Residuals vs Fitted
Residuals
Fitted values

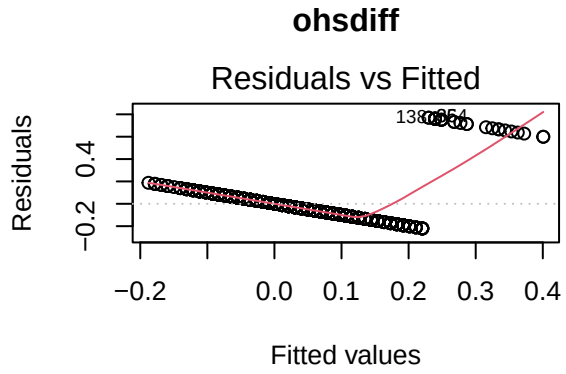

age
Normal Q−Q
Standardized residuals
Theoretical Quantiles

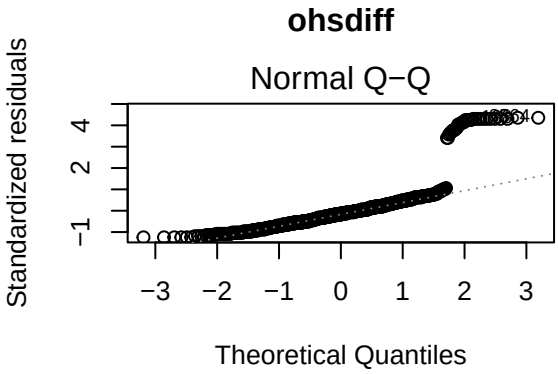

age
Scale−Location
√|Standardized residuals|
Fitted values

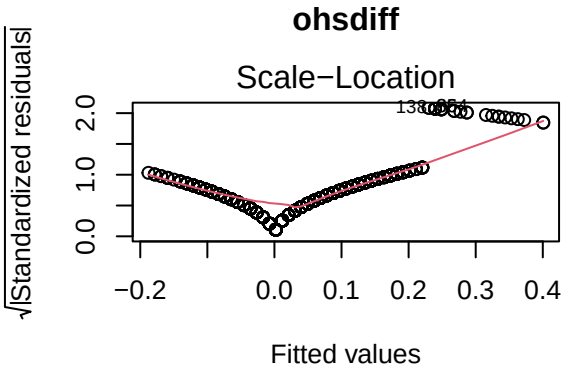

age
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage

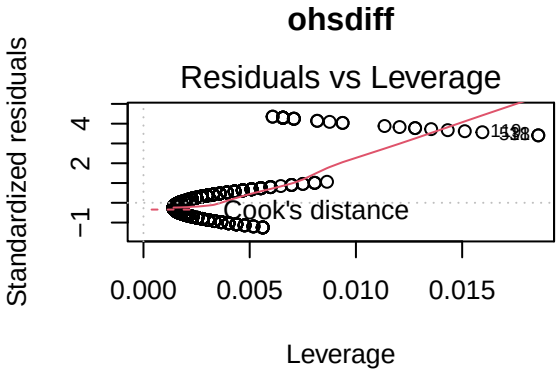

ohs0
Residuals vs Fitted
Residuals
Fitted values
ohs0
Normal Q−Q
Standardized residuals
Theoretical Quantiles
ohs0
Scale−Location
√|Standardized residuals|
Fitted values
ohs0
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage

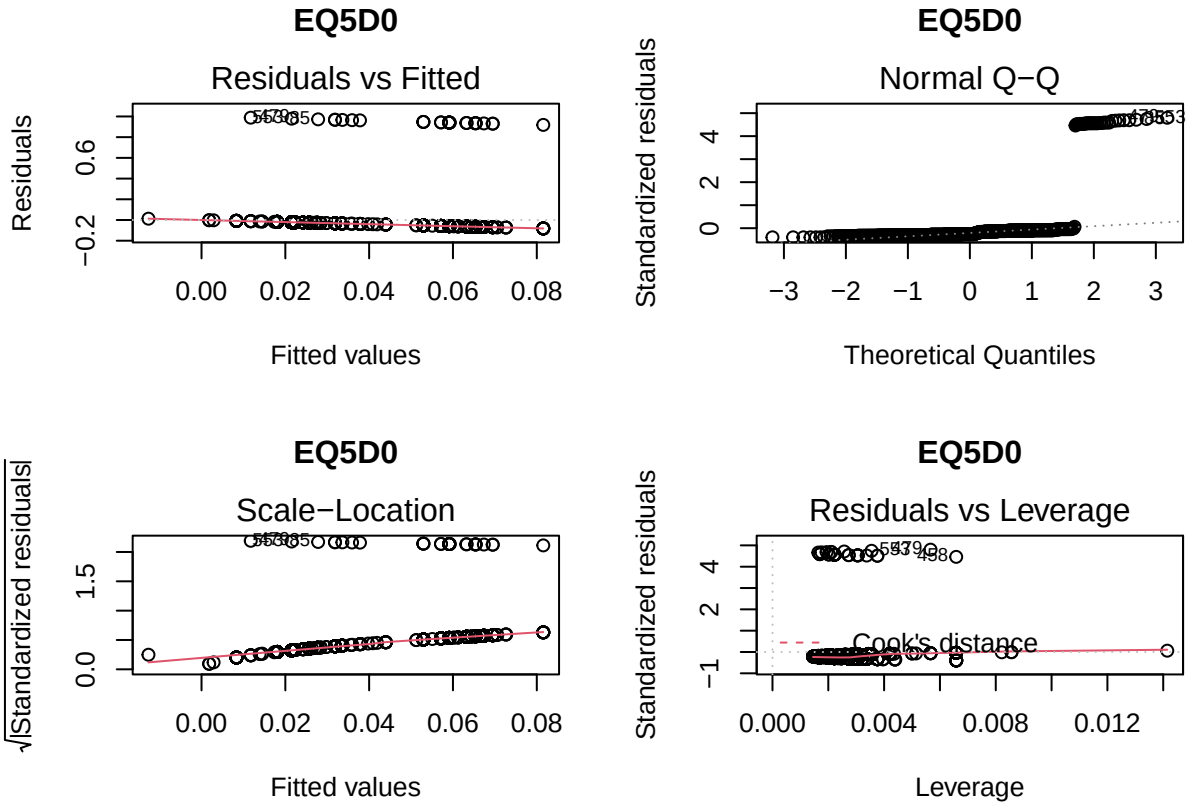

ohs6
Residuals vs Fitted
Residuals
Fitted values
ohs6
Normal Q−Q
Standardized residuals
Theoretical Quantiles
ohs6
Scale−Location
√|Standardized residuals|
Fitted values
ohs6
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage

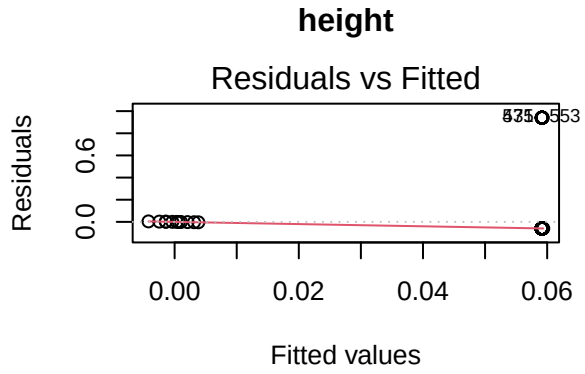

ohsdiff
Residuals vs Fitted
Residuals
Fitted values

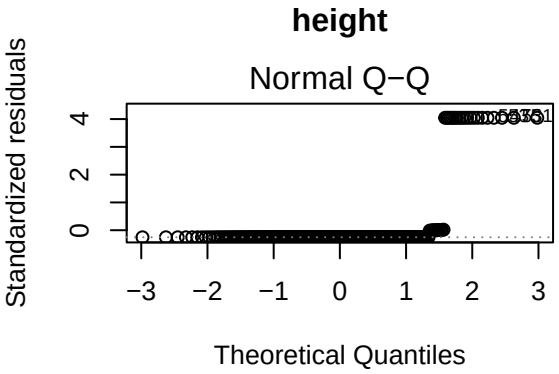

ohsdiff
Normal Q−Q
Standardized residuals
Theoretical Quantiles

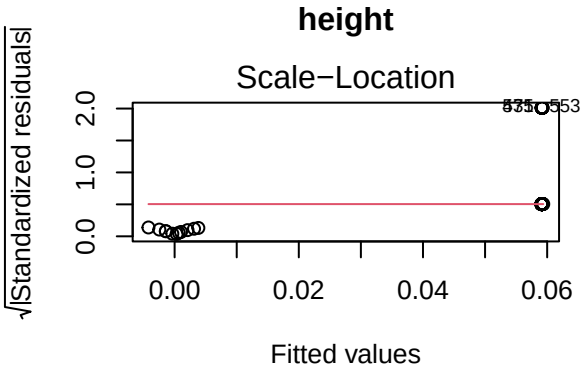

ohsdiff
Scale−Location
√|Standardized residuals|
Fitted values

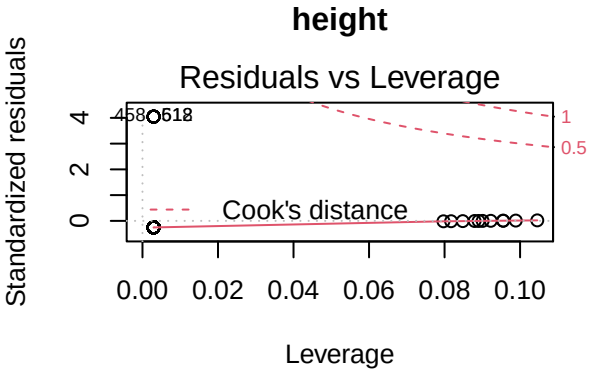

ohsdiff
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage

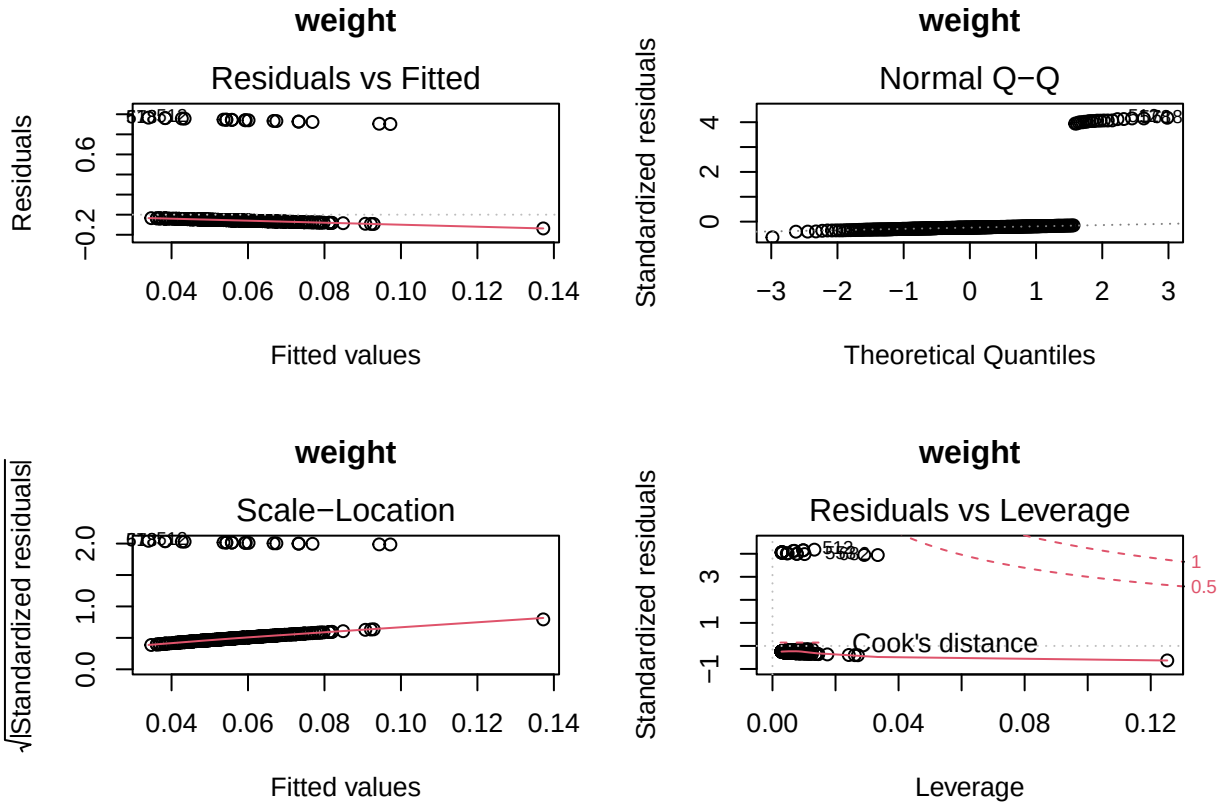

EQ5D0
Residuals vs Fitted
Residuals
Fitted values
EQ5D0
Normal Q−Q
Standardized residuals
Theoretical Quantiles
EQ5D0
Scale−Location
√|Standardized residuals|
Fitted values
EQ5D0
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage

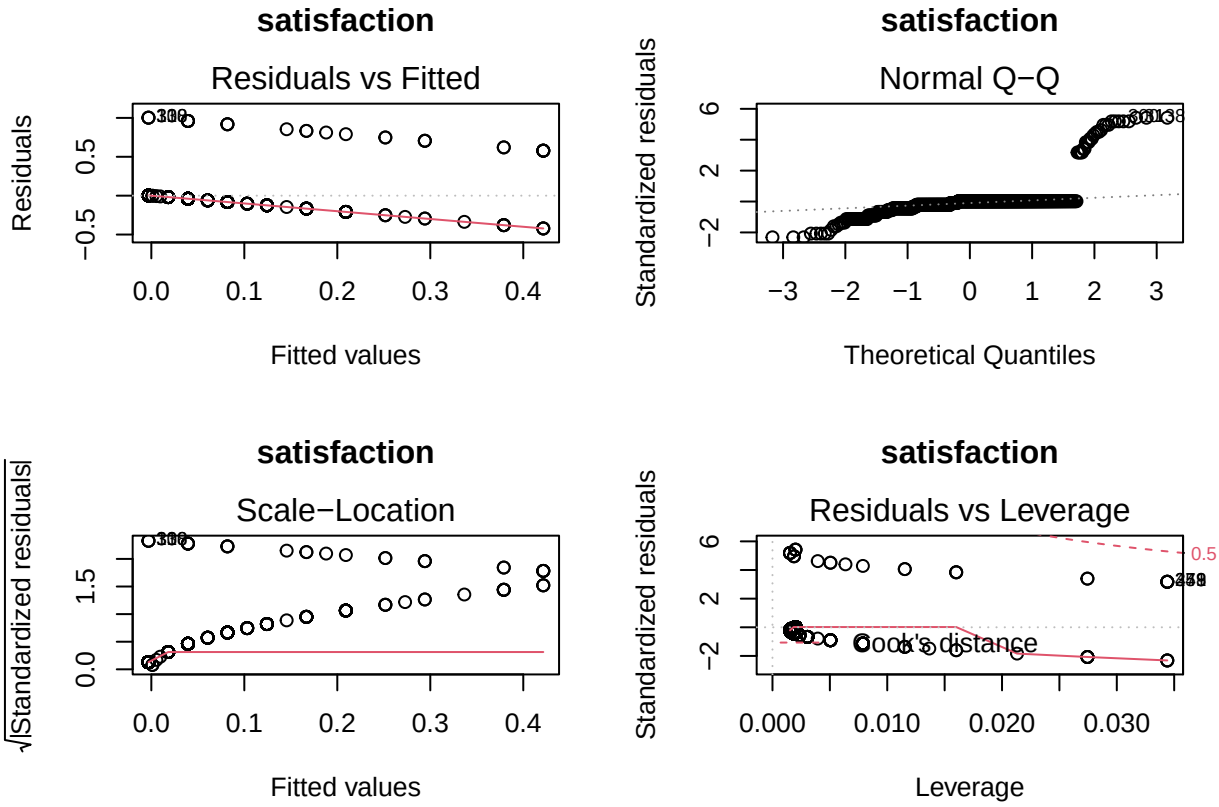

height
Residuals vs Fitted
Residuals
Fitted values
0.0
0.6
0.00
0.02
0.04
0.06

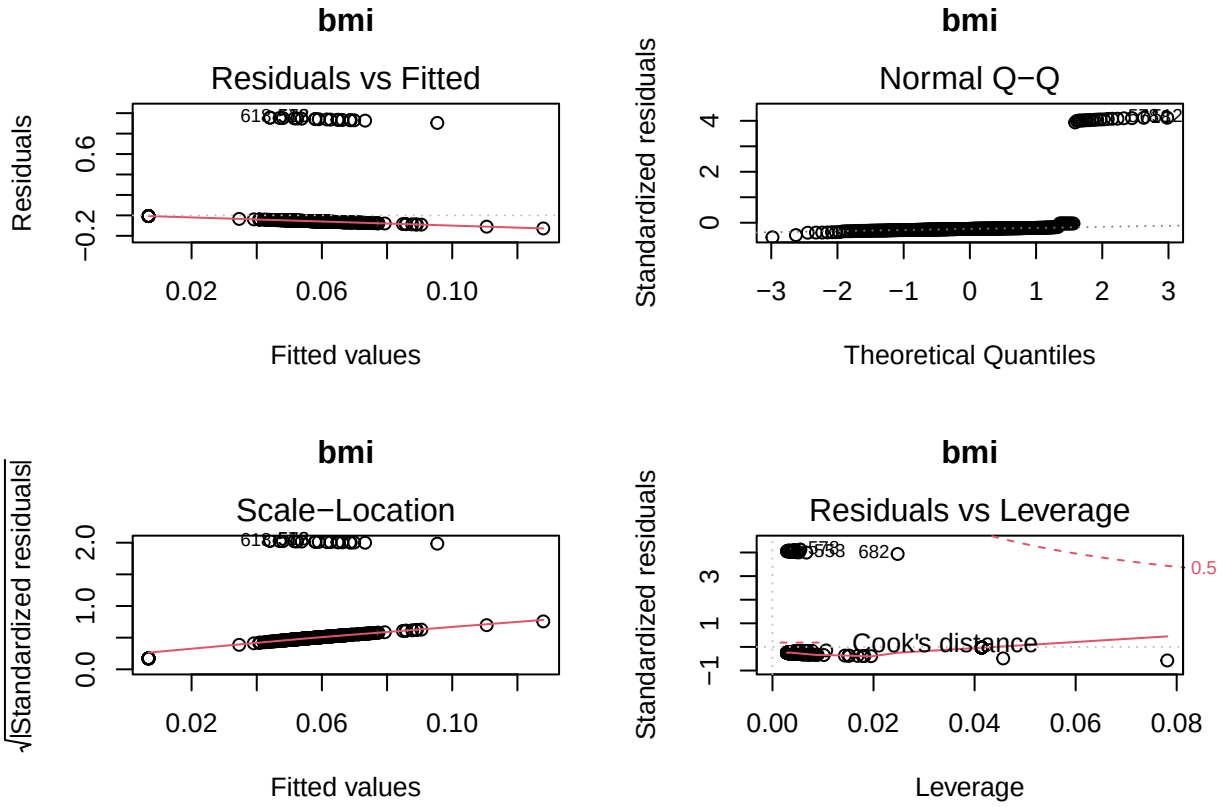

height
Normal Q−Q
Standardized residuals
Theoretical Quantiles
0
2
4
−3
−2
−1
0
1
2
3

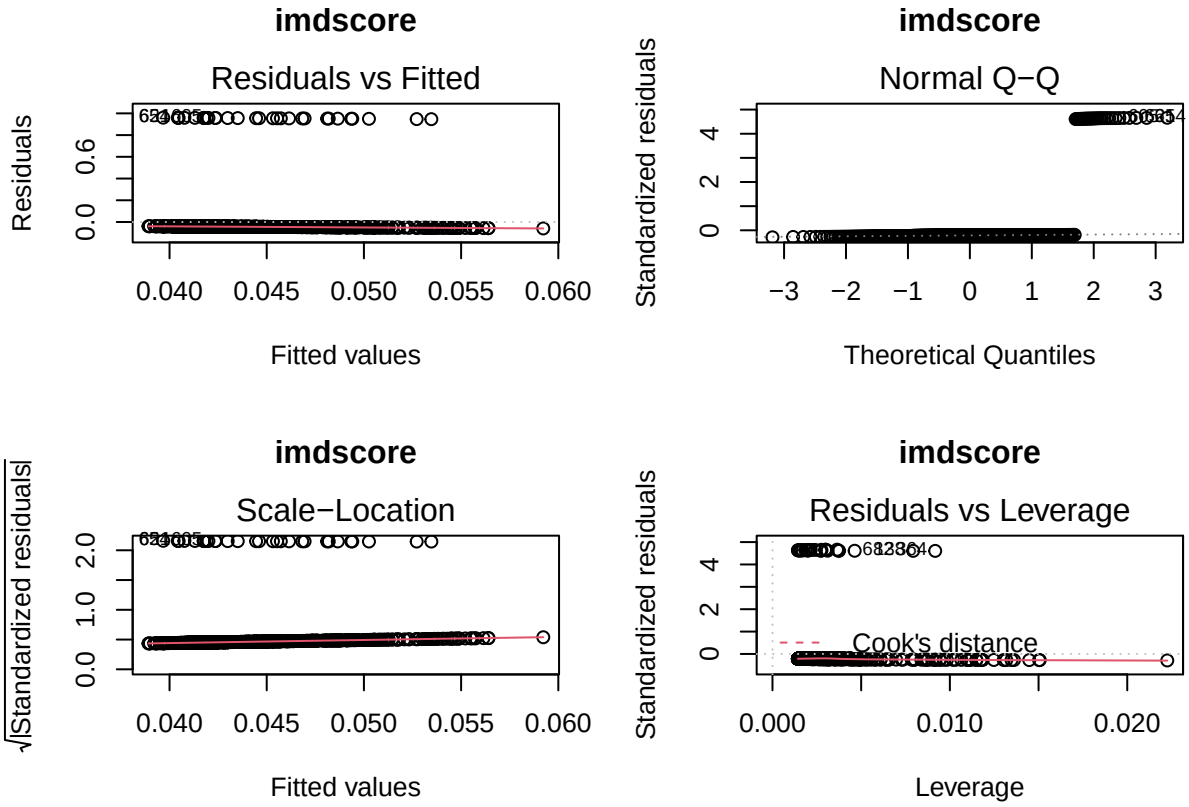

height
Scale−Location
√|Standardized residuals|
Fitted values
0.0
1.0
2.0
0.00
0.02
0.04
0.06


height
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
1
0.5
0
2
4
0.00
0.02
0.04
0.06
0.08
0.10


weight
Residuals vs Fitted
Residuals
Fitted values
weight
Normal Q−Q
Standardized residuals
Theoretical Quantiles
weight
Scale−Location
√|Standardized residuals|
Fitted values
weight
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage


satisfaction
Residuals vs Fitted
Residuals
Fitted values
satisfaction
Normal Q−Q
Standardized residuals
Theoretical Quantiles
satisfaction
Scale−Location
√|Standardized residuals|
Fitted values
satisfaction
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage


bmi
Residuals vs Fitted
Residuals
Fitted values
bmi
Normal Q−Q
Standardized residuals
Theoretical Quantiles
bmi
Scale−Location
√|Standardized residuals|
Fitted values
bmi
Residuals vs Leverage
Standardized residuals
Cook's distance
0.5
Leverage


imdscore
Residuals vs Fitted
Residuals
Fitted values
imdscore
Normal Q−Q
Standardized residuals
Theoretical Quantiles
imdscore
Scale−Location
√|Standardized residuals|
Fitted values
imdscore
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage