model2.R

zhang alice

2023-11-25

```
rm(list = ls())
gc()
```

```
##          used (Mb) gc trigger (Mb) max used (Mb)
## Ncells 469971 25.1    1020654 54.6   644240 34.5
## Vcells 855986  6.6    8388608 64.0  1634809 12.5
```

```
set.seed(123)
library(car)
```

```
## Warning: package 'car' was built under R version 4.2.3
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.2.3
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.2.3
```

```
library(olsrr)
```

```
## Warning: package 'olsrr' was built under R version 4.2.3
```

```
##
## Attaching package: 'olsrr'
```

```
## The following object is masked from 'package:datasets':
##
##     rivers
```

```
############### (1) Data cleaning ########################################
## select variables
library(NHANES)
```

```
## Warning: package 'NHANES' was built under R version 4.2.3
```

```
df0 <- NHANES
df <- NHANES[NHANES$Age >= 18 & NHANES$Age < 60,]
# colSums(is.na(df)) / nrow(df)
df <- df[, which(colSums(is.na(df)) / nrow(df) < 0.3)]
# exclude duplication
df <- df[!duplicated(df),]
names(df)
```

```
##  [1] "ID"              "SurveyYr"        "Gender"          "Age"
##  [5] "AgeDecade"       "Race1"           "Education"       "MaritalStatus"
##  [9] "HHIncome"        "HHIncomeMid"     "Poverty"         "HomeRooms"
```

```
## [13] "HomeOwn"         "Work"            "Weight"          "Height"
## [17] "BMI"             "BMI_WHO"         "Pulse"           "BPSysAve"
## [21] "BPDiaAve"        "BPSys1"          "BPDia1"          "BPSys2"
## [25] "BPDia2"          "BPSys3"          "BPDia3"          "DirectChol"
## [29] "TotChol"         "UrineVol1"       "UrineFlow1"      "Diabetes"
## [33] "HealthGen"       "DaysPhysHlthBad" "DaysMentHlthBad" "LittleInterest"
## [37] "Depressed"       "SleepHrsNight"   "SleepTrouble"    "PhysActive"
## [41] "Alcohol12PlusYr" "AlcoholYear"     "Smoke100"        "Smoke100n"
## [45] "Marijuana"       "RegularMarij"    "HardDrugs"       "SexEver"
## [49] "SexAge"          "SexNumPartnLife" "SexNumPartYear"  "SameSex"
## [53] "SexOrientation"
```

```
# df$BPSysAve
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.2.3
```

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

```
## The following object is masked from 'package:car':
##
##     recode
```

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

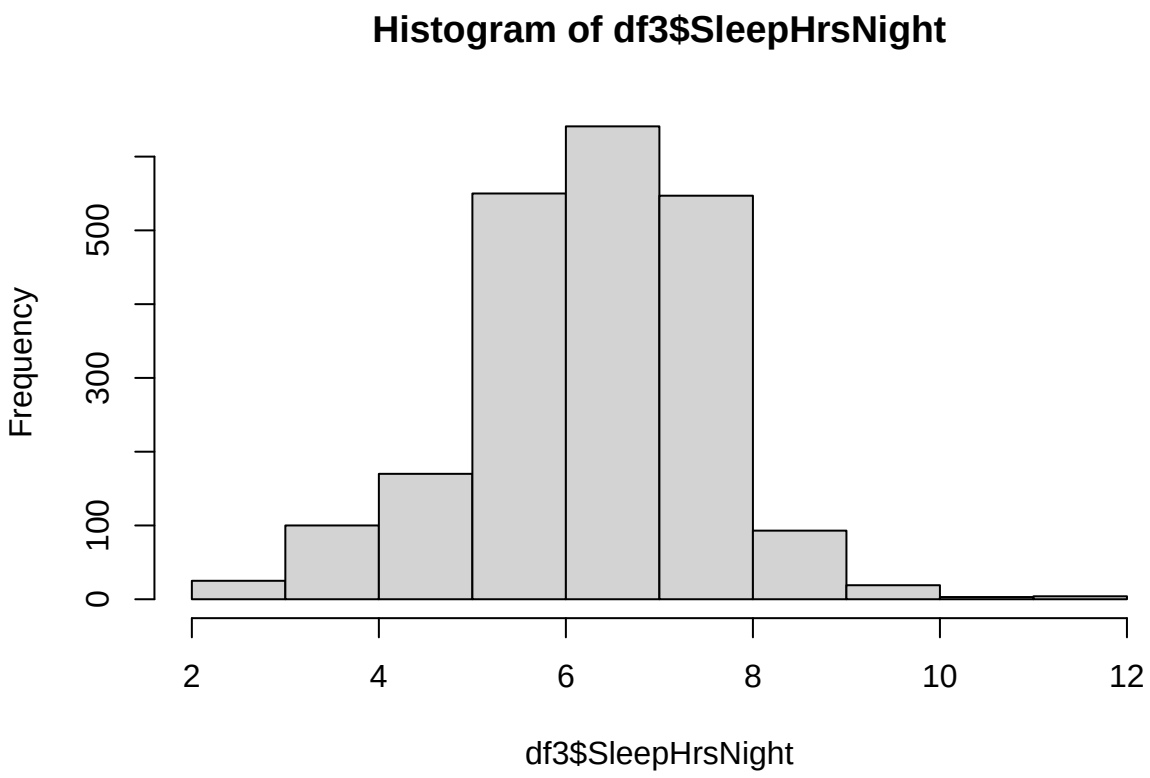
```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
df2 <- df %>% select(
  SleepHrsNight,
  BMI,
  DirectChol,
  Age,
  Gender,
  Race1,
  TotChol,
  BPDiaAve,
  BPSysAve,
  AlcoholYear,
  Poverty,
  SexNumPartnLife,
  SexNumPartYear,
  DaysMentHlthBad,
  UrineFlow1,
  PhysActive,
  DaysPhysHlthBad,
  Smoke100,
  Depressed,
  HealthGen,
  SexAge
)

df3 <- na.omit(df2)
```

```
#df3$SleepHrsNight <- df3$SleepHrsNight * 60
#df3 <- df3[, -which(names(df3) %in% "SleepHrsNight")]
# cor(df3$BPSysAve,df3$BPDiaAve)
psych::describe(df3)
```

```
##                  vars    n   mean    sd median trimmed   mad   min     max
## SleepHrsNight       1 2152   6.78  1.31   7.00    6.85  1.48  2.00   12.00
## BMI                 2 2152  28.77  6.75  27.60   28.09  5.78 15.02   69.00
## DirectChol          3 2152   1.35  0.41   1.29    1.31  0.39  0.39    3.83
## Age                 4 2152  39.18 11.33  39.00   39.15 14.83 20.00   59.00
## Gender*             5 2152   1.53  0.50   2.00    1.54  0.00  1.00    2.00
## Race1*              6 2152   3.43  1.15   4.00    3.57  0.00  1.00    5.00
## TotChol             7 2152   5.07  1.05   4.99    5.01  1.04  1.53   13.65
## BPDiaAve            8 2152  71.19 11.84  71.00   71.28 10.38  0.00  116.00
## BPSysAve            9 2152 117.43 14.28 116.00  116.50 13.34 78.00  209.00
## AlcoholYear        10 2152  70.59 94.22  24.00   50.94 35.58  0.00  364.00
## Poverty            11 2152   2.84  1.69   2.78    2.89  2.49  0.00    5.00
## SexNumPartnLife    12 2152  16.73 66.13   7.00    8.91  5.93  0.00 2000.00
## SexNumPartYear     13 2152   1.38  2.59   1.00    1.04  0.00  0.00   69.00
## DaysMentHlthBad    14 2152   4.47  8.02   0.00    2.40  0.00  0.00   30.00
## UrineFlow1         15 2152   1.07  0.97   0.81    0.91  0.60  0.00   10.14
## PhysActive*        16 2152   1.58  0.49   2.00    1.60  0.00  1.00    2.00
## DaysPhysHlthBad    17 2152   3.16  7.19   0.00    1.12  0.00  0.00   30.00
## Smoke100*          18 2152   1.46  0.50   1.00    1.45  0.00  1.00    2.00
## Depressed*         19 2152   1.30  0.58   1.00    1.16  0.00  1.00    3.00
## HealthGen*         20 2152   2.64  0.94   3.00    2.65  1.48  1.00    5.00
## SexAge             21 2152  17.10  3.39  17.00   16.80  2.97  9.00   44.00
##                   range  skew kurtosis   se
## SleepHrsNight     10.00 -0.30     0.69 0.03
## BMI               53.98  1.28     2.96 0.15
## DirectChol         3.44  1.09     2.27 0.01
## Age               39.00  0.02    -1.15 0.24
## Gender*            1.00 -0.12    -1.99 0.01
## Race1*             4.00 -1.13     0.08 0.02
## TotChol           12.12  0.92     3.47 0.02
## BPDiaAve         116.00 -0.39     3.13 0.26
## BPSysAve         131.00  1.00     2.94 0.31
## AlcoholYear      364.00  1.66     1.98 2.03
## Poverty            5.00 -0.01    -1.47 0.04
## SexNumPartnLife 2000.00 18.82   456.62 1.43
## SexNumPartYear    69.00 14.07   293.16 0.06
## DaysMentHlthBad   30.00  2.16     3.76 0.17
## UrineFlow1        10.14  2.89    14.06 0.02
## PhysActive*        1.00 -0.32    -1.90 0.01
## DaysPhysHlthBad   30.00  2.80     7.06 0.15
## Smoke100*          1.00  0.15    -1.98 0.01
## Depressed*         2.00  1.83     2.21 0.01
## HealthGen*         4.00  0.11    -0.33 0.02
## SexAge            35.00  1.51     5.56 0.07
```

```
# psych::pairs.panels(df3)
hist(df3$SleepHrsNight)
```




```
# colSums(is.na(df2)) / nrow(df2)
fit0 <-
  lm(SleepHrsNight ~ .,
     data = df3)
#data type
df3$Gender <- ifelse(df3$Gender == "male", 0, 1)
df3$Smoke100 <- ifelse(df3$Smoke100 == "No", 0, 1)
df3$PhysActive <- ifelse(df3$PhysActive == "No", 0, 1)
df3 <- df3 %>%
  mutate(
    Race1 = case_when(
      Race1 == 'Black' ~ 1,
      Race1 == 'Hispanic' ~ 2,
      Race1 == 'Mexican' ~ 3,
      Race1 == 'White' ~ 4,
      Race1 == 'Other' ~ 5,
      TRUE ~ NA_integer_  # Default value if none of the conditions are met
    )
  )
df3$logBMI = log(df3$BMI+1)
## model_2 add known risk factors ##
m_2 = lm(
  logBMI ~ SleepHrsNight + Age + Gender + factor(Race1) + TotChol + BPDiaAve + BPSysAve + AlcoholYear +
    DaysPhysHlthBad + PhysActive,
  df3
)
```

```
summary(m_2)
```

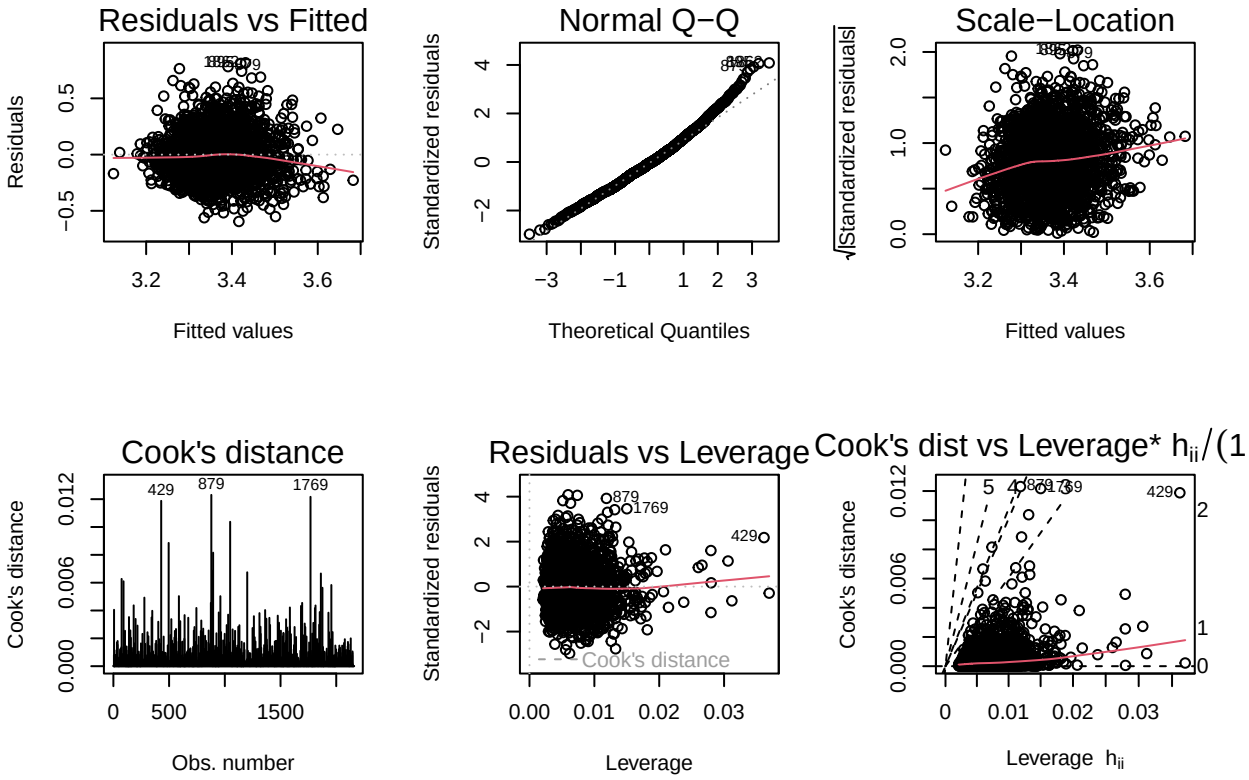
```
##
## Call:
## lm(formula = logBMI ~ SleepHrsNight + Age + Gender + factor(Race1) +
##     TotChol + BPDiaAve + BPSysAve + AlcoholYear + Smoke100 +
##     DaysPhysHlthBad + PhysActive, data = df3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.59481 -0.13429 -0.01056  0.11809  0.81672
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.092e+00  5.055e-02  61.162  < 2e-16 ***
## SleepHrsNight   -7.668e-03  3.365e-03  -2.279  0.02279 *
## Age              6.691e-04  4.245e-04   1.576  0.11519
## Gender           4.868e-03  9.178e-03   0.530  0.59591
## factor(Race1)2  -5.235e-02  2.058e-02  -2.544  0.01102 *
## factor(Race1)3  -1.511e-02  1.798e-02  -0.840  0.40078
## factor(Race1)4  -5.292e-02  1.330e-02  -3.980 7.11e-05 ***
## factor(Race1)5  -1.077e-01  2.019e-02  -5.336 1.05e-07 ***
## TotChol          5.920e-03  4.365e-03   1.356  0.17518
## BPDiaAve         1.764e-03  4.402e-04   4.007 6.37e-05 ***
## BPSysAve         2.029e-03  3.772e-04   5.381 8.23e-08 ***
## AlcoholYear     -3.261e-04  4.817e-05  -6.770 1.66e-11 ***
## Smoke100        -1.821e-02  9.029e-03  -2.017  0.04385 *
## DaysPhysHlthBad  1.883e-03  6.190e-04   3.042  0.00238 **
## PhysActive      -3.612e-02  9.211e-03  -3.921 9.10e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2006 on 2137 degrees of freedom
## Multiple R-squared:  0.1122, Adjusted R-squared:  0.1064
## F-statistic: 19.29 on 14 and 2137 DF,  p-value: < 2.2e-16
```

```
car::Anova(m_2, type = "III")
```

```
## Anova Table (Type III tests)
##
## Response: logBMI
##                   Sum Sq   Df   F value    Pr(>F)
## (Intercept)     150.530    1 3740.7560 < 2.2e-16 ***
## SleepHrsNight     0.209    1    5.1921  0.022788 *
## Age               0.100    1    2.4836  0.115186
## Gender            0.011    1    0.2813  0.595907
## factor(Race1)     1.444    4    8.9685 3.509e-07 ***
## TotChol           0.074    1    1.8393  0.175178
## BPDiaAve          0.646    1   16.0540 6.367e-05 ***
## BPSysAve          1.165    1   28.9518 8.232e-08 ***
## AlcoholYear       1.844    1   45.8302 1.660e-11 ***
## Smoke100          0.164    1    4.0672  0.043848 *
## DaysPhysHlthBad   0.372    1    9.2535  0.002379 **
## PhysActive        0.619    1   15.3743 9.096e-05 ***
```

```
## Residuals        85.994 2137
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
############ model 2 diagnosis ############
par(mfrow = c(2, 3)) #read more from ?plot.lm
plot(m_2, which = 1)
plot(m_2, which = 2)
plot(m_2, which = 3)
plot(m_2, which = 4)
plot(m_2, which = 5)
plot(m_2, which = 6)
```




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

m_2.yhat = m_2$fitted.values
m_2.res = m_2$residuals
m_2.h = hatvalues(m_2)
m_2.r = rstandard(m_2)
m_2.rr = rstudent(m_2)
#which subject is most outlying with respect to the x space
Hmisc::describe(m_2.h)
```

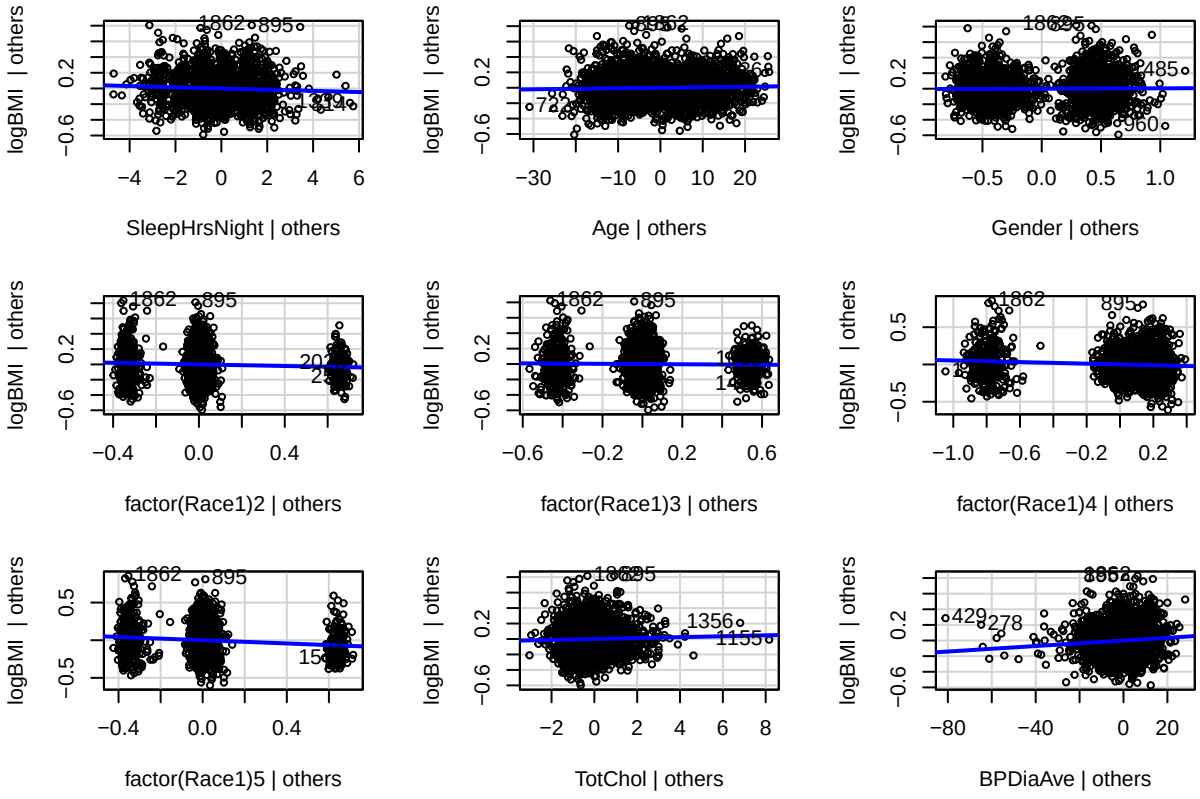
```
## m_2.h
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2152        0     2152        1  0.00697 0.003726 0.003065 0.003360
##      .25      .50      .75      .90      .95
```

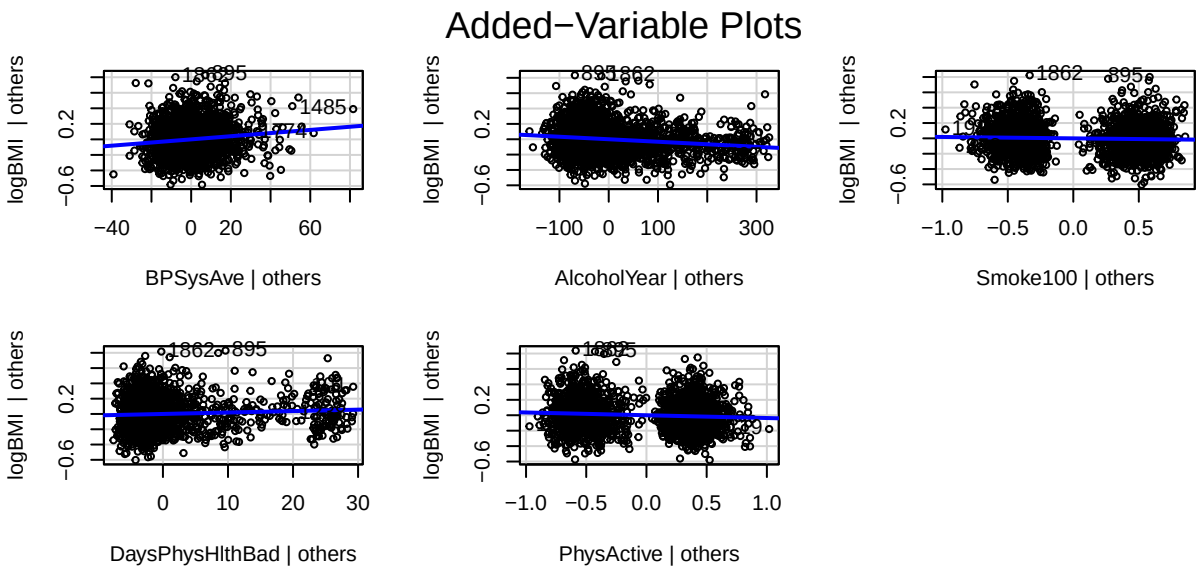
```
## 0.004188 0.006211 0.008970 0.011331 0.012996
##
## lowest : 0.00209248 0.0021336  0.00216141 0.00216865 0.00224259
## highest: 0.0280094  0.0306139  0.0312045  0.0361597  0.0369534
```

```
m_2.h[which.max(m_2.h)]
```

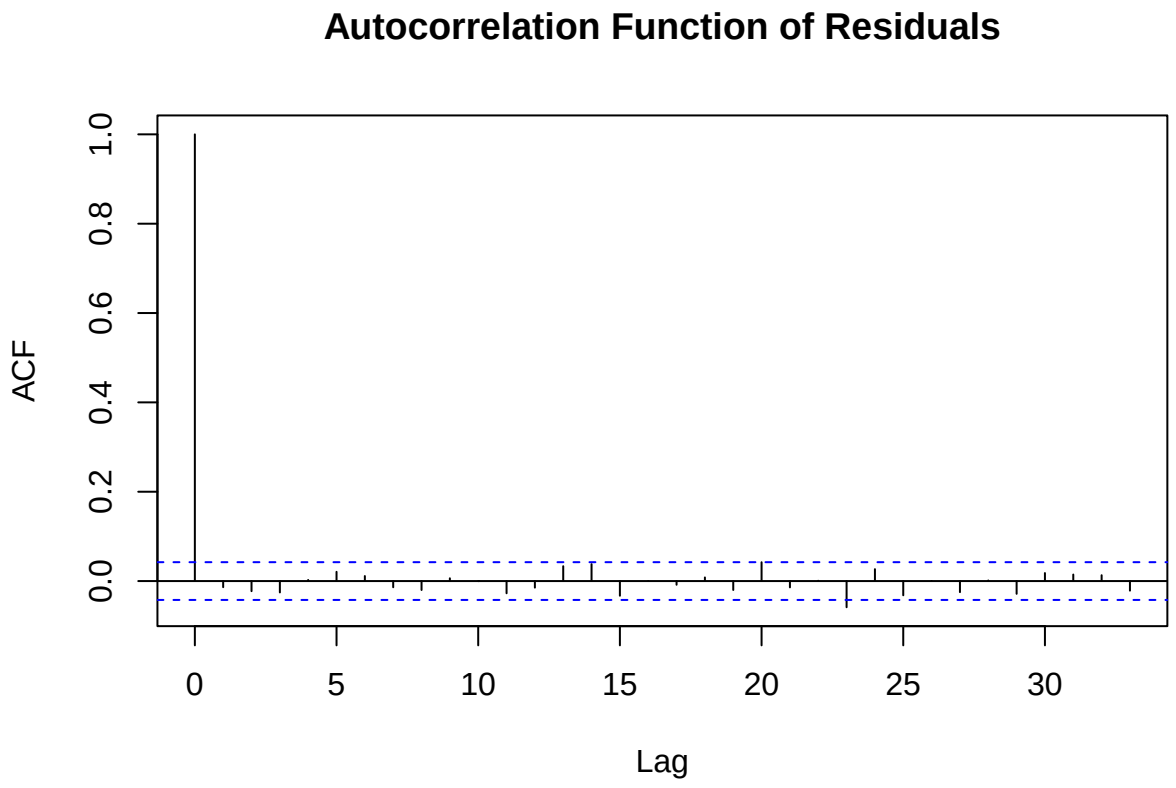
```
##       1155
## 0.03695339
```

```
###################### Assumption:LINE ################################

#(1)Linear: 2 approaches

# partial regression plots
car::avPlots(m_2)
```

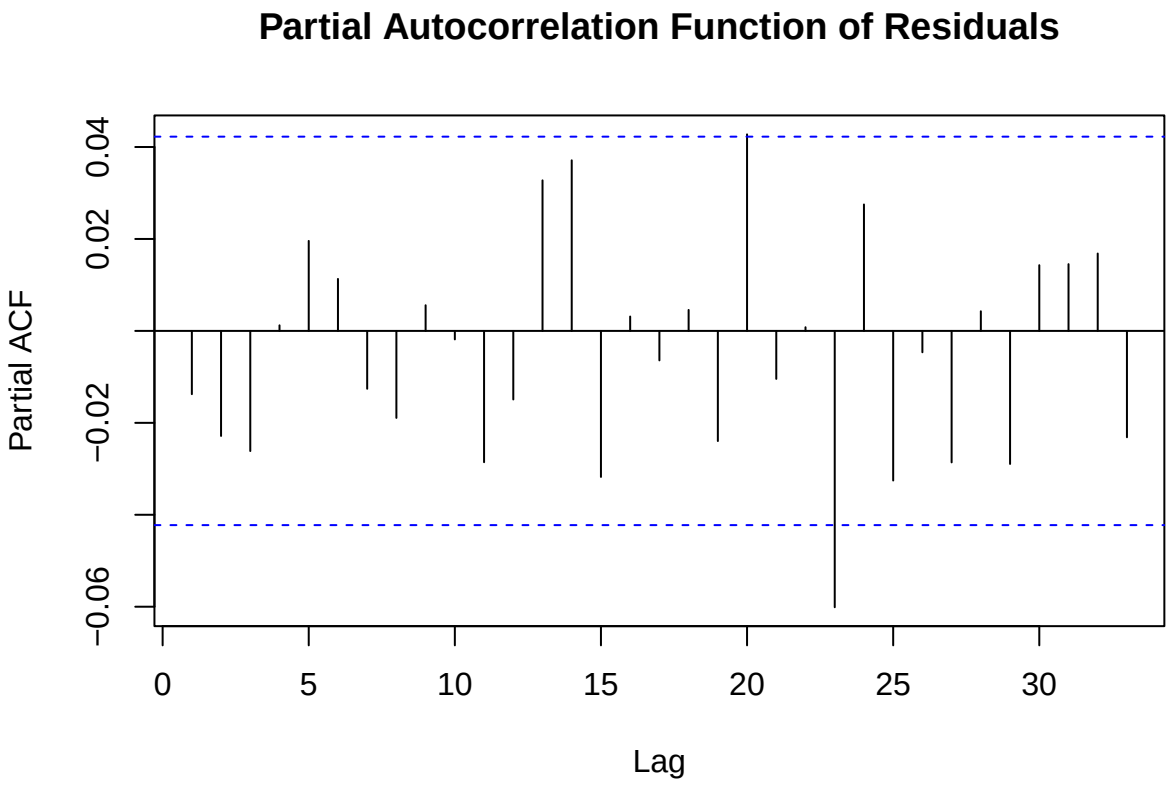





```
#(2)Independence:

residuals <- resid(m_2)
acf(residuals, main = "Autocorrelation Function of Residuals")
```




```
pacf(residuals, main = "Partial Autocorrelation Function of Residuals")
```




```
library(lmtest)
```

```
## Warning: package 'lmtest' was built under R version 4.2.3
```

```
## Loading required package: zoo
```

```
## Warning: package 'zoo' was built under R version 4.2.3
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
# Perform Durbin-Watson test
dw_test_result <- dwtest(m_2, alternative = "two.sided")

# Print the Durbin-Watson test result
print(dw_test_result)
```

```
##
##  Durbin-Watson test
##
## data:  m_2
## DW = 2.0275, p-value = 0.5235
## alternative hypothesis: true autocorrelation is not 0
```

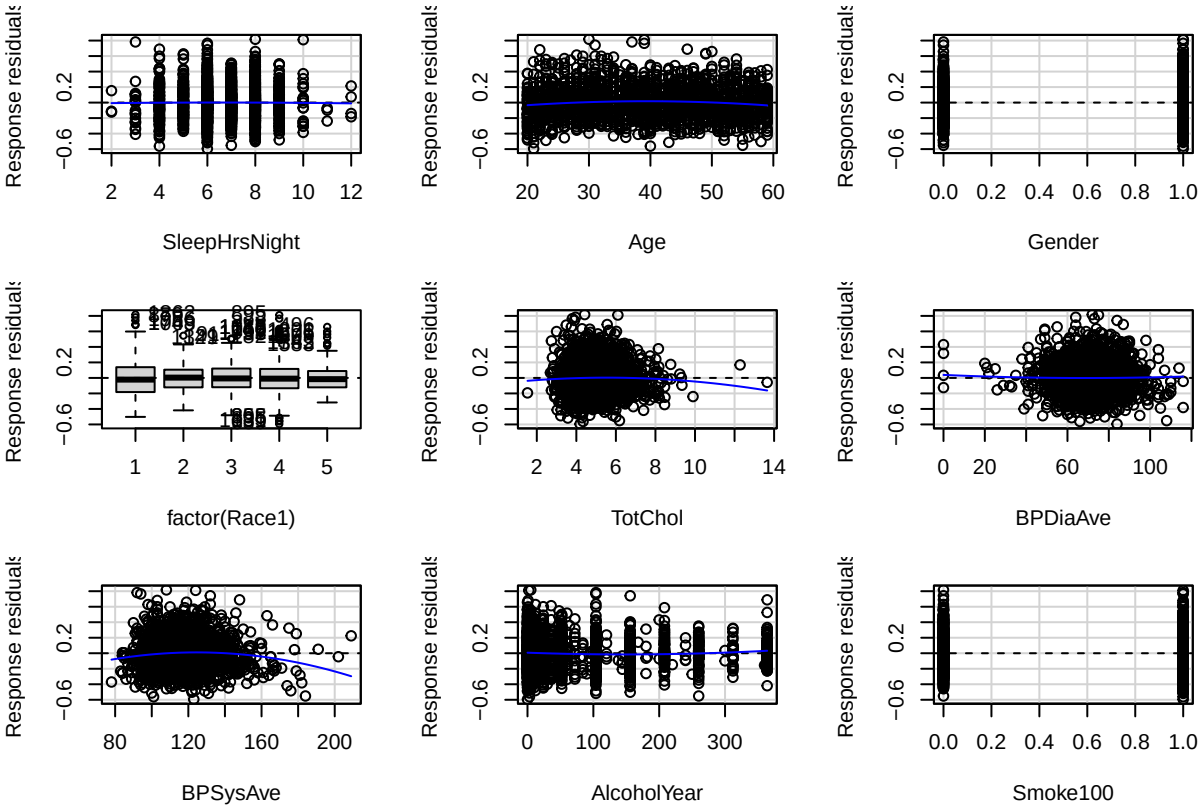
```
library(lmtest)
bptest(m_2)
```

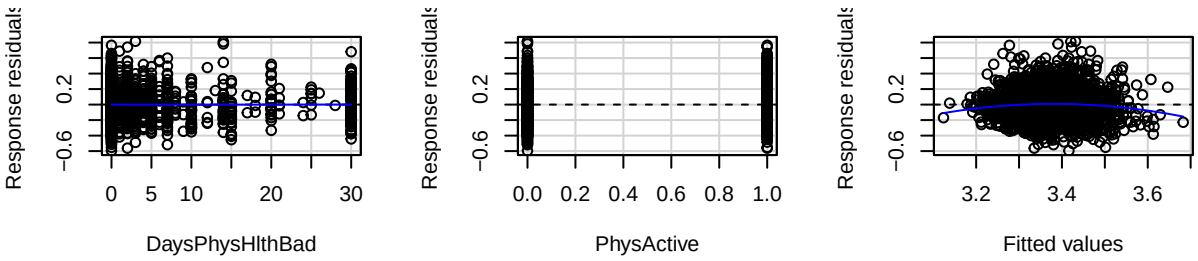
```
##
##  studentized Breusch-Pagan test
##
## data:  m_2
## BP = 134.94, df = 14, p-value < 2.2e-16
```

```
epsilon1<-residuals(m_2)
weights1 <- 1/abs(epsilon1)
m_2.2<-lm(logBMI ~ SleepHrsNight + Age + Gender + factor(Race1), df3,weights = weights1)
summary(m_2.2)
```

```
##
## Call:
## lm(formula = logBMI ~ SleepHrsNight + Age + Gender + factor(Race1),
##     data = df3, weights = weights1)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7529 -0.4483 -0.1021  0.3992  3.3720
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.4264129  0.0106454 321.867  < 2e-16 ***
## SleepHrsNight  -0.0089435  0.0014453  -6.188 7.28e-10 ***
## Age             0.0019187  0.0001629  11.780  < 2e-16 ***
## Gender          0.0124312  0.0034090   3.647 0.000272 ***
## factor(Race1)2 -0.0839735  0.0082767 -10.146  < 2e-16 ***
## factor(Race1)3 -0.0543745  0.0074941  -7.256 5.56e-13 ***
## factor(Race1)4 -0.0832679  0.0043963 -18.940  < 2e-16 ***
## factor(Race1)5 -0.1484716  0.0076906 -19.305  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5058 on 2144 degrees of freedom
## Multiple R-squared:  0.2876, Adjusted R-squared:  0.2853
## F-statistic: 123.7 on 7 and 2144 DF,  p-value: < 2.2e-16
```

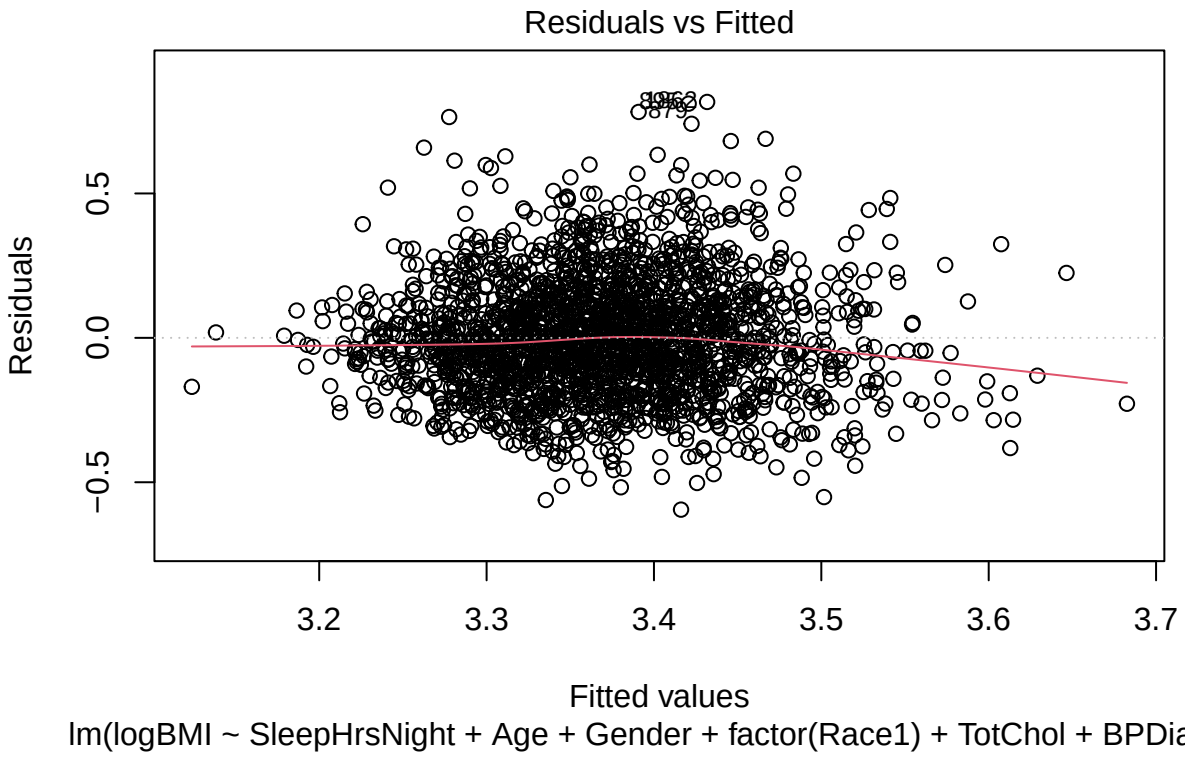
```
#(3)E: constant var: residuals-fitted values; transform for variance-stable...(total: 4 solutions)
library(ggplot2)
car::residualPlots(m_2, type = "response")
```


Response residuals
SleepHrsNight
Age
Gender
factor(Race1)
TotChol
BPDiaAve
BPSysAve
AlcoholYear
Smoke100

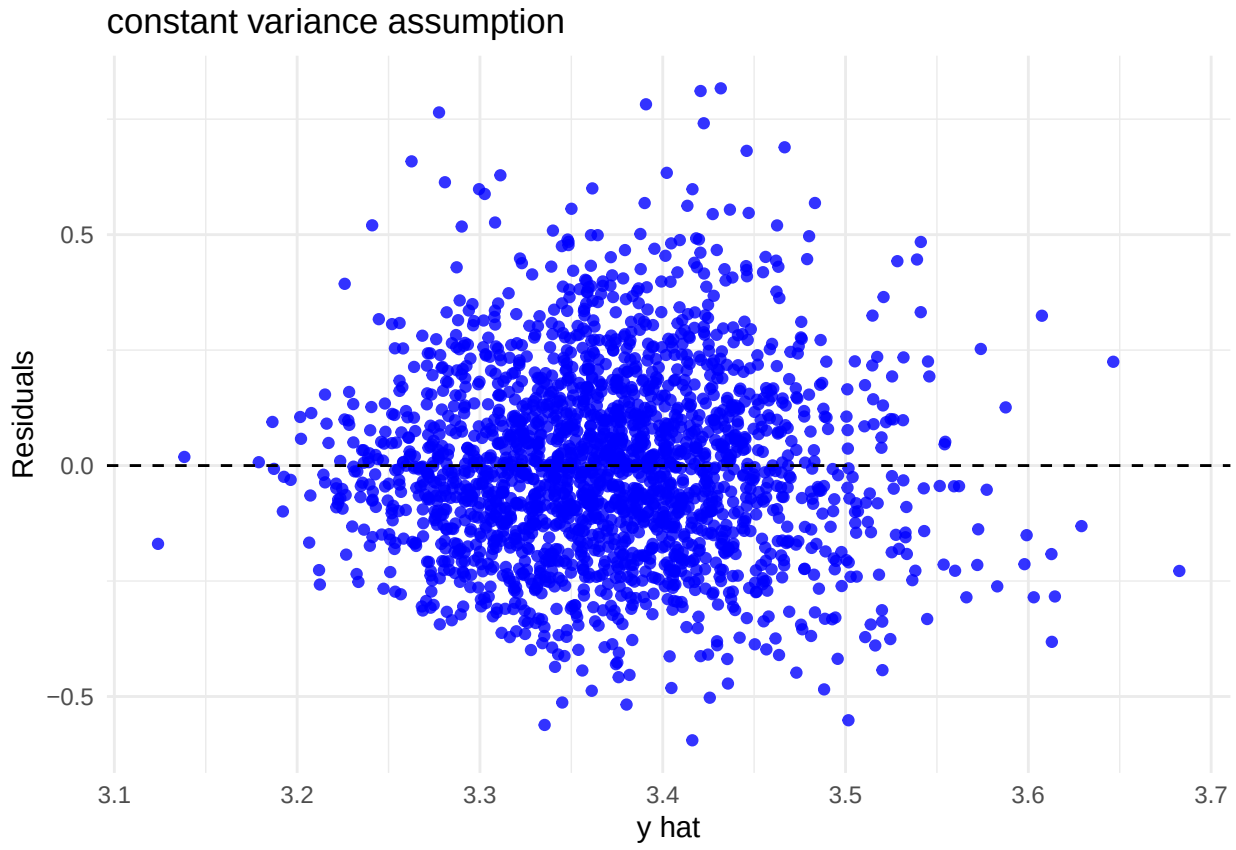



```
##                  Test stat Pr(>|Test stat|)
## SleepHrsNight      -0.2403        0.8101492
## Age                -3.8493        0.0001219 ***
## Gender              0.0374        0.9701789
## factor(Race1)
## TotChol            -1.3940        0.1634675
## BPDiaAve            0.6512        0.5149995
## BPSysAve           -4.1202         3.93e-05 ***
## AlcoholYear         2.3712        0.0178174 *
## Smoke100           -1.1689        0.2425822
## DaysPhysHlthBad     0.0651        0.9480842
## PhysActive         -0.0841        0.9330026
## Tukey test         -3.5414        0.0003980 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(m_2, which = 1)
```




```
#or
ggplot(m_2, aes(x = m_2.yhat, y = m_2.res)) +
  geom_point(color = "blue", alpha = 0.8) +
  geom_hline(yintercept = 0,
             linetype = "dashed",
             color = "black") +
  labs(title = "constant variance assumption",
       x = "y hat",
       y = "Residuals") +
  theme_minimal()
```



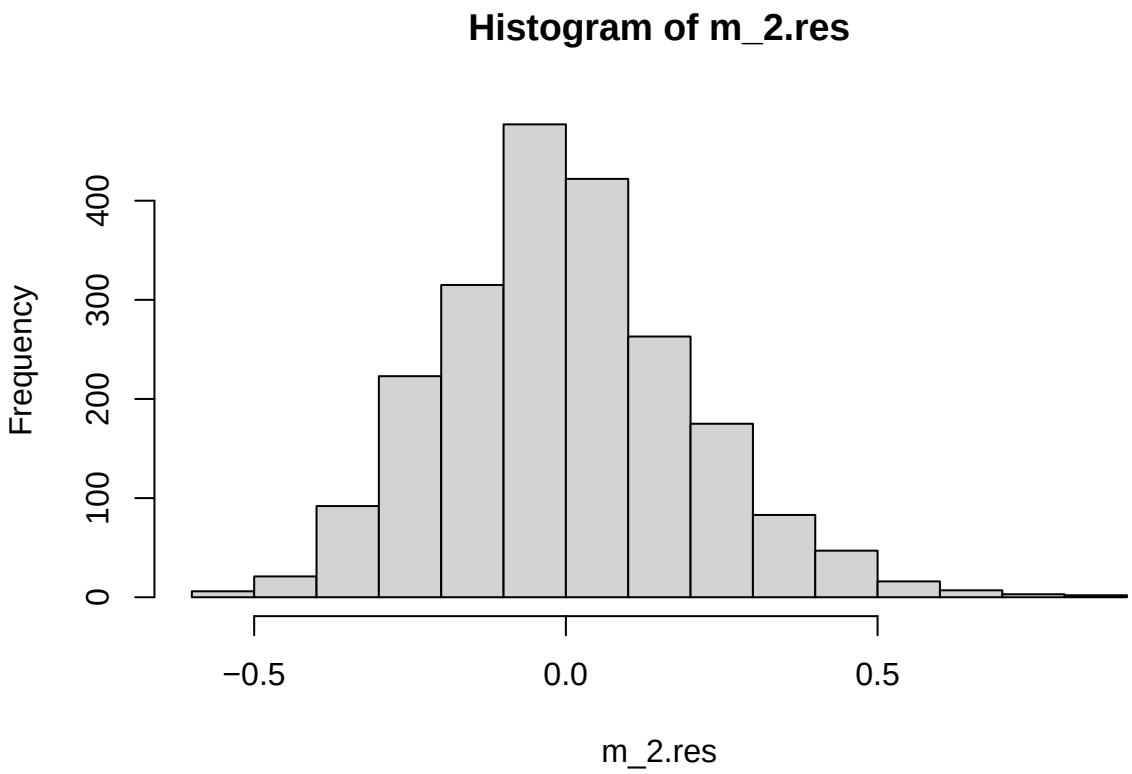

```
#conclusion: the constant variance assumption is basically not violated. The spread of the residuals ap


#(4)Normality: residuals freq - residuals (4 plots: his, box, Q-Q, shapiro); transform

#exam quartiles of the residuals
Hmisc::describe(m_2.res)
```

```
## m_2.res
##          n    missing   distinct       Info       Mean        Gmd        .05
##       2152          0       2152          1 -2.631e-18      0.223   -0.30536
##        .10        .25        .50        .75        .90        .95
##   -0.24203   -0.13429   -0.01056    0.11809    0.26068    0.35099
##
## lowest : -0.594807 -0.561559 -0.55139  -0.517444 -0.513002
## highest: 0.741103  0.764543  0.782158  0.810912  0.816723
```

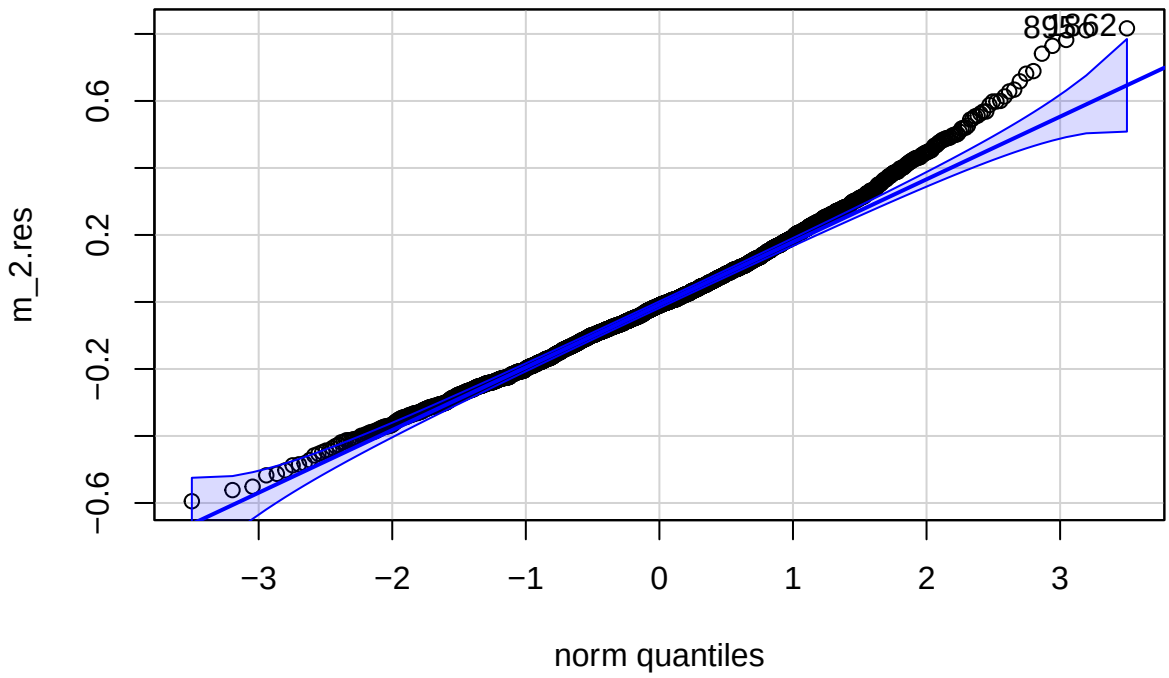
```
Hmisc::describe(m_2.res)$counts[c(".25", ".50", ".75")] #not symmetric
```

```
##        .25        .50        .75
## "-0.13429" "-0.01056" " 0.11809"
```

```
#histogram
par(mfrow = c(1, 1))
hist(m_2.res, breaks = 15)
```



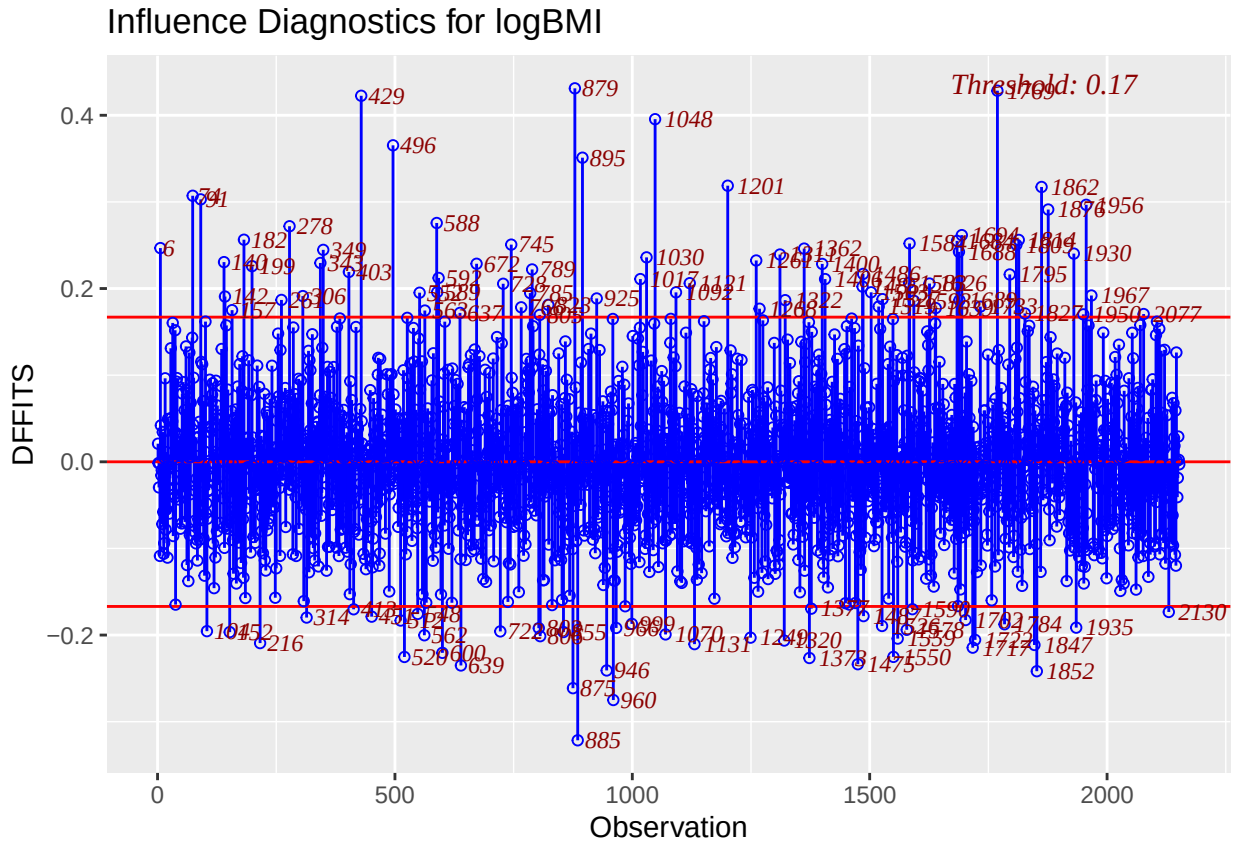

```
# Q-Q plot
qq.m_2.res = car::qqPlot(m_2.res)
```




```
m_2.res[qq.m_2.res]
```

```
##      1862       895
## 0.8167233 0.8109118
```

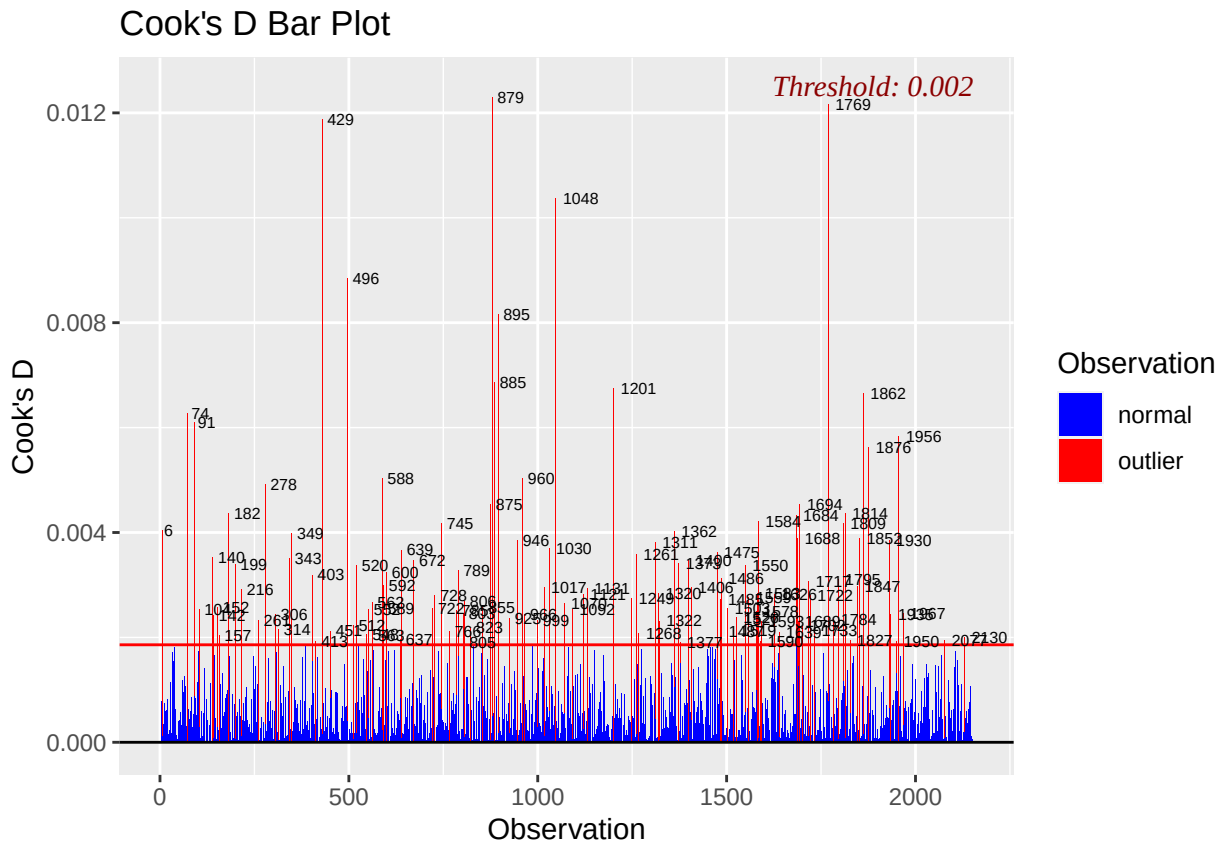
```
################ influential observations  ##################

influence2 = data.frame(
  Residual = resid(m_2),
  Rstudent = rstudent(m_2),
  HatDiagH = hat(model.matrix(m_2)),
  CovRatio = covratio(m_2),
  DFFITS = dffits(m_2),
  COOKsDistance = cooks.distance(m_2)
)
# DFFITS
ols_plot_dffits(m_2)
```




```
influence2[order(abs(influence2$DFFITS), decreasing = T), ] %>% head()
```

```
##       Residual Rstudent    HatDiagH  CovRatio    DFFITS COOKsDistance
## 879  0.7821577 3.935678 0.011856107 0.9144480 0.4311025   0.012306514
## 1769 0.6890643 3.470012 0.015010796 0.9397188 0.4283681   0.012170404
## 429  0.4292211 2.181364 0.036159704 1.0105273 0.4225110   0.011880140
## 1048 0.6813655 3.427753 0.013138069 0.9398468 0.3955005   0.010375852
## 496  0.6585841 3.310678 0.012029452 0.9439799 0.3653153   0.008855742
## 895  0.8109118 4.072181 0.007376848 0.9034121 0.3510510   0.008156314
```

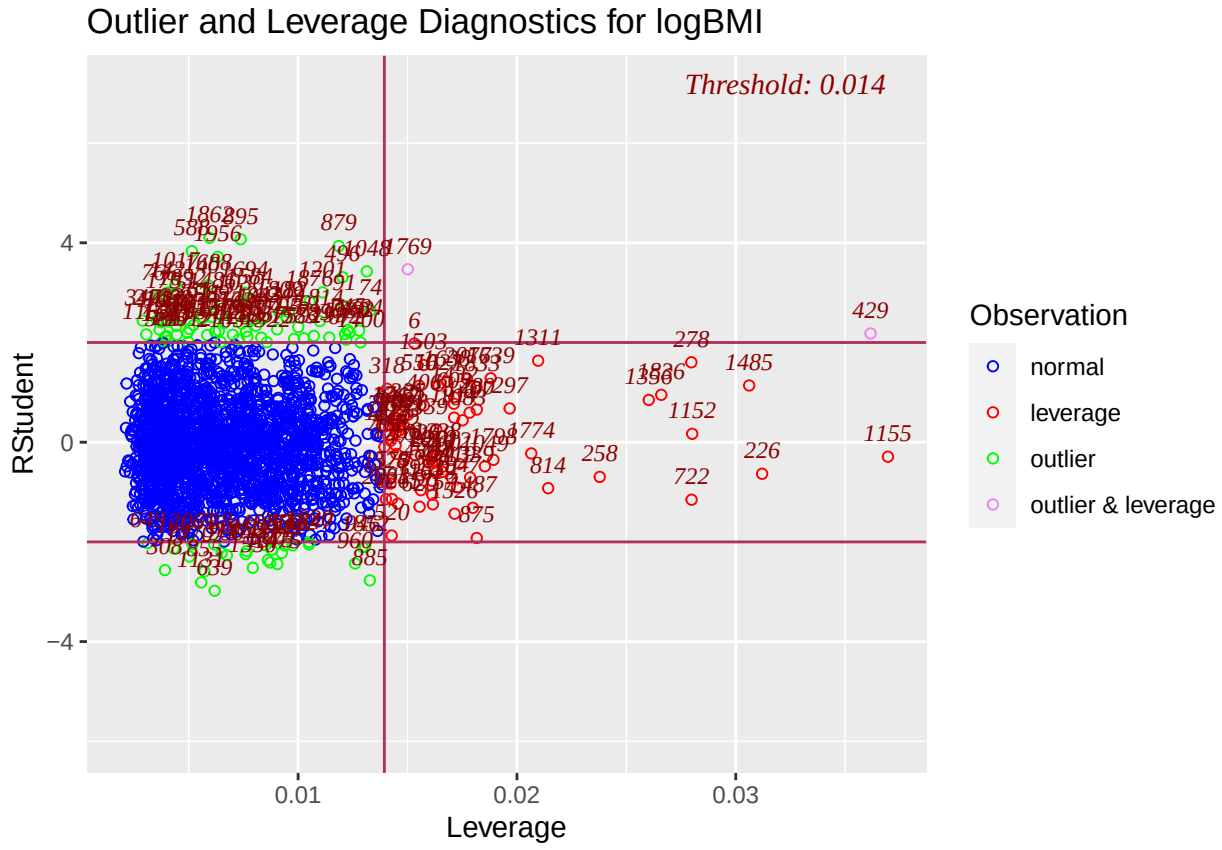
```
#From the plot above, we can see 2 observations with the largest (magnitude) of DFFITS, observation 879

# Cook's D
ols_plot_cooksd_bar(m_2)
```




```
influence2[order(influence2$COOKsDistance, decreasing = T), ] %>% head()
```

```
##       Residual Rstudent    HatDiagH  CovRatio    DFFITS COOKsDistance
## 879  0.7821577 3.935678 0.011856107 0.9144480 0.4311025   0.012306514
## 1769 0.6890643 3.470012 0.015010796 0.9397188 0.4283681   0.012170404
## 429  0.4292211 2.181364 0.036159704 1.0105273 0.4225110   0.011880140
## 1048 0.6813655 3.427753 0.013138069 0.9398468 0.3955005   0.010375852
## 496  0.6585841 3.310678 0.012029452 0.9439799 0.3653153   0.008855742
## 895  0.8109118 4.072181 0.007376848 0.9034121 0.3510510   0.008156314
```

```
#From the plot above, we can see that the observation 879 and 1769 also have the largest Cook's Distanc

#leverage
ols_plot_resid_lev(m_2)
```




```
#high leverage
influence2[order(influence2$HatDiagH, decreasing = T), ] %>% head()
```

```
##         Residual   Rstudent   HatDiagH CovRatio      DFFITS COOKsDistance
## 1155 -0.05733155 -0.2911689 0.03695339 1.045065 -0.05703593  2.169661e-04
## 429   0.42922110  2.1813637 0.03615970 1.010527  0.42251096  1.188014e-02
## 226  -0.12498932 -0.6329420 0.03120446 1.036562 -0.11359421  8.604844e-04
## 1485  0.22472322  1.1378827 0.03061387 1.029449  0.20221258  2.725619e-03
## 1152  0.03325452  0.1681082 0.02800938 1.035859  0.02853708  5.431569e-05
## 722  -0.22818255 -1.1538431 0.02798115 1.026397 -0.19576823  2.554617e-03
```

```
#high studentized residual
influence2[order(influence2$Rstudent, decreasing = T), ] %>% head()
```

```
##       Residual Rstudent    HatDiagH  CovRatio    DFFITS COOKsDistance
## 1862 0.8167233 4.098637 0.005955274 0.9007623 0.3172395   0.006660157
## 895  0.8109118 4.072181 0.007376848 0.9034121 0.3510510   0.008156314
## 879  0.7821577 3.935678 0.011856107 0.9144480 0.4311025   0.012306514
## 588  0.7645426 3.833335 0.005144678 0.9133288 0.2756614   0.005033689
## 1956 0.7411025 3.717264 0.006328375 0.9200244 0.2966525   0.005831869
## 1769 0.6890643 3.470012 0.015010796 0.9397188 0.4283681   0.012170404
```

```
#From the plot above, we can see that the observation 1155 has the largest leverage (0.0368). Observati


#From the plot above, there is 7 observations(1048,1769,1684, 74, 72, 1689, 1311) located in the inters
#The thresholds for the externally studentized residual are -2 and 2, i.e. 2 in magnitude. The threshol
```

```r
#From (DFFITS), observations 879 and 1769 appear to be influential observations. Observation 1155 has es

rm2.df3 = df3[-c(879, 1769, 1155, 1048, 1769, 1684, 74, 72, 1689, 1311), ]
rm.m_2 =  lm(
  logBMI ~ SleepHrsNight + Age + Gender + Race1 + TotChol + BPDiaAve + BPSysAve + AlcoholYear + Smoke10
    DaysPhysHlthBad + PhysActive,
  rm2.df3
)
## Before removing these observations, the estimated coefficients are:
summary(m_2)$coef
```

```
##                      Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept)      3.0918224538 5.055159e-02 61.1617204 0.000000e+00
## SleepHrsNight   -0.0076684090 3.365370e-03 -2.2786229 2.278776e-02
## Age              0.0006690605 4.245454e-04  1.5759458 1.151863e-01
## Gender           0.0048679852 9.178382e-03  0.5303751 5.959070e-01
## factor(Race1)2  -0.0523531993 2.057699e-02 -2.5442596 1.102080e-02
## factor(Race1)3  -0.0151117273 1.798154e-02 -0.8404021 4.007770e-01
## factor(Race1)4  -0.0529198531 1.329526e-02 -3.9803545 7.110965e-05
## factor(Race1)5  -0.1077142022 2.018685e-02 -5.3358603 1.051487e-07
## TotChol          0.0059204590 4.365469e-03  1.3562024 1.751780e-01
## BPDiaAve         0.0017638655 4.402241e-04  4.0067443 6.366926e-05
## BPSysAve         0.0020293687 3.771578e-04  5.3806885 8.231914e-08
## AlcoholYear     -0.0003261218 4.817301e-05 -6.7698025 1.660454e-11
## Smoke100        -0.0182083387 9.028629e-03 -2.0167335 4.384819e-02
## DaysPhysHlthBad  0.0018830175 6.190159e-04  3.0419533 2.379070e-03
## PhysActive      -0.0361179715 9.211392e-03 -3.9210114 9.095701e-05
```

```r
## After removing these observations, the estimated coefficients are:
summary(rm.m_2)$coef
```

```
##                      Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept)      3.0848094788 5.039006e-02 61.2186090 0.000000e+00
## SleepHrsNight   -0.0062973071 3.340135e-03 -1.8853451 5.951920e-02
## Age              0.0007772666 4.197458e-04  1.8517555 6.419917e-02
## Gender           0.0013225431 9.059279e-03  0.1459877 8.839449e-01
## Race1           -0.0158186521 3.823982e-03 -4.1366961 3.660220e-05
## TotChol          0.0066206313 4.391868e-03  1.5074750 1.318372e-01
## BPDiaAve         0.0016386369 4.361135e-04  3.7573635 1.763255e-04
## BPSysAve         0.0021150806 3.754431e-04  5.6335587 1.998934e-08
## AlcoholYear     -0.0003455809 4.778925e-05 -7.2313515 6.636253e-13
## Smoke100        -0.0194374892 8.940628e-03 -2.1740630 2.981013e-02
## DaysPhysHlthBad  0.0014176365 6.201482e-04  2.2859642 2.235410e-02
## PhysActive      -0.0351429330 9.081729e-03 -3.8696301 1.122761e-04
```

```r
#### change percent
abs((rm.m_2$coefficients - m_2$coefficients) / (m_2$coefficients) * 100)
```

```
## Warning in rm.m_2$coefficients - m_2$coefficients: longer object length is not
## a multiple of shorter object length
```

```
##     (Intercept)   SleepHrsNight             Age          Gender  factor(Race1)2
##    2.268233e-01    1.787987e+01    1.617283e+01    7.283182e+01    6.978475e+01
##  factor(Race1)3  factor(Race1)4  factor(Race1)5         TotChol        BPDiaAve
##    1.438112e+02    1.030965e+02    1.019636e+02    1.058371e+02    1.201983e+03
##        BPSysAve     AlcoholYear        Smoke100 DaysPhysHlthBad      PhysActive
```

```
##    3.014396e+01    1.067602e+04    1.704174e+04    4.344264e+02    1.021520e+02
```

```
#The estimated regression coefficients doesn't change slightly after removing these observations. 5 of

##################   multicollinearity   ######################
#Pearson correlations
var2 = c(
  "logBMI",
  "SleepHrsNight",
  "Age",
  "Gender",
  "Race1",
  "TotChol",
  "BPDiaAve",
  "BPSysAve",
  "AlcoholYear",
  "Smoke100",
  "DaysPhysHlthBad",
  "PhysActive"
)
newData2 = df3[, var2]
library("corrplot")
```

```
## Warning: package 'corrplot' was built under R version 4.2.3
```

```
## corrplot 0.92 loaded
```

```
par(mfrow = c(1, 2))
cormat2 = cor(as.matrix(newData2[, -c(1)], method = "pearson"))
p.mat2 = cor.mtest(as.matrix(newData2[, -c(1)]))$p
corrplot(
  cormat2,
  method = "color",
  type = "upper",
  number.cex = 1,
  diag = FALSE,
  addCoef.col = "black",
  tl.col = "black",
  tl.srt = 90,
  p.mat = p.mat2,
  sig.level = 0.05,
  insig = "blank",
)

#None of the covariates seem strongly correlated.There is no evidence of collinearity from the pair-wis

# collinearity diagnostics (VIF)
car::vif(m_2)
```

```
##                     GVIF Df GVIF^(1/(2*Df))
## SleepHrsNight   1.039090  1        1.019358
## Age             1.237393  1        1.112382
## Gender          1.122181  1        1.059330
## factor(Race1)   1.124710  4        1.014799
## TotChol         1.126202  1        1.061227
## BPDiaAve        1.452874  1        1.205352
```

```
## BPSysAve         1.550823  1        1.245321
## AlcoholYear      1.101306  1        1.049431
## Smoke100         1.083963  1        1.041135
## DaysPhysHlthBad 1.058155  1        1.028666
## PhysActive       1.106090  1        1.051708
```

```
#From the VIF values in the output above, once again we do not observe any potential collinearity issue
```