Model2

Liancheng, He Zhang, Zhengrui Huang, Zibo Yu

2023-11-25

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

```
##          used (Mb) gc trigger (Mb) max used (Mb)
## Ncells 469887 25.1    1012104 54.1   660860 35.3
## Vcells 879834  6.8    8388608 64.0  1800812 13.8
```

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

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

```
library(ggplot2)
library(olsrr)
```

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

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

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

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

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

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

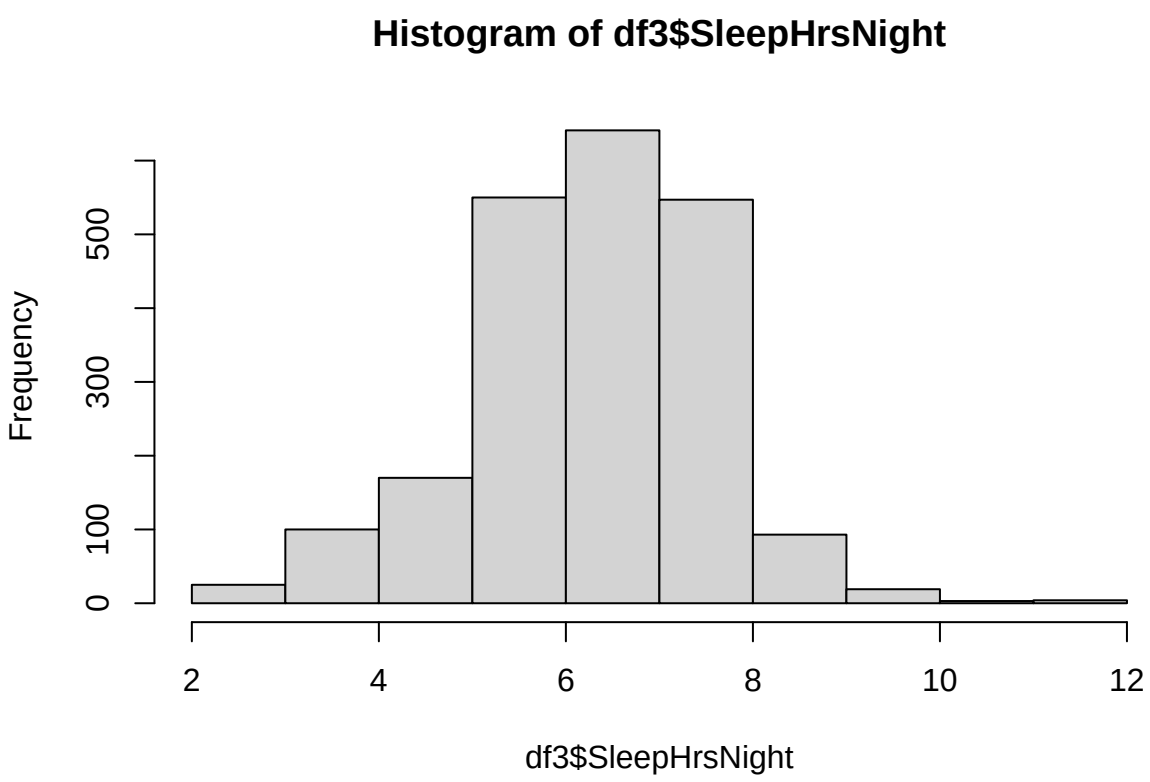
```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
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
    )
  )

## model_2 add known risk factors ##
m_2 = lm(
  BMI ~ SleepHrsNight + Age + Gender + Race1 + TotChol + BPDiaAve + BPSysAve + AlcoholYear + Smoke100 +
    DaysPhysHlthBad + PhysActive,
  df3
)
```

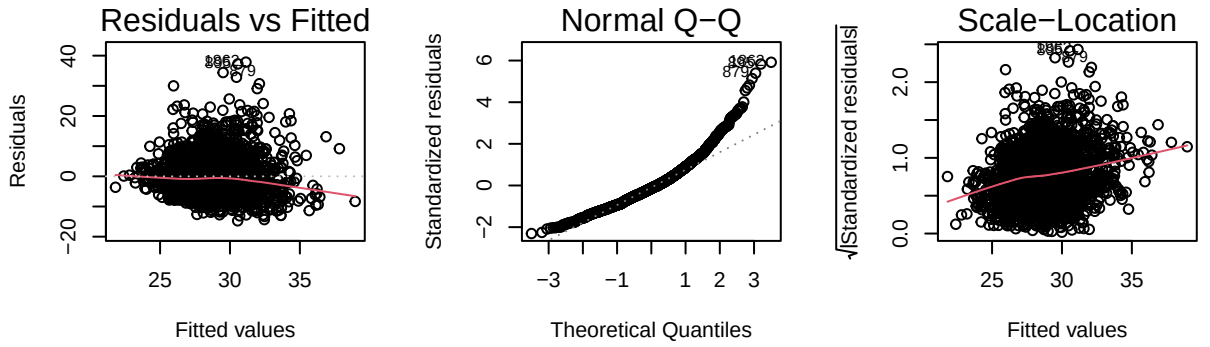
```r
summary(m_2)
```

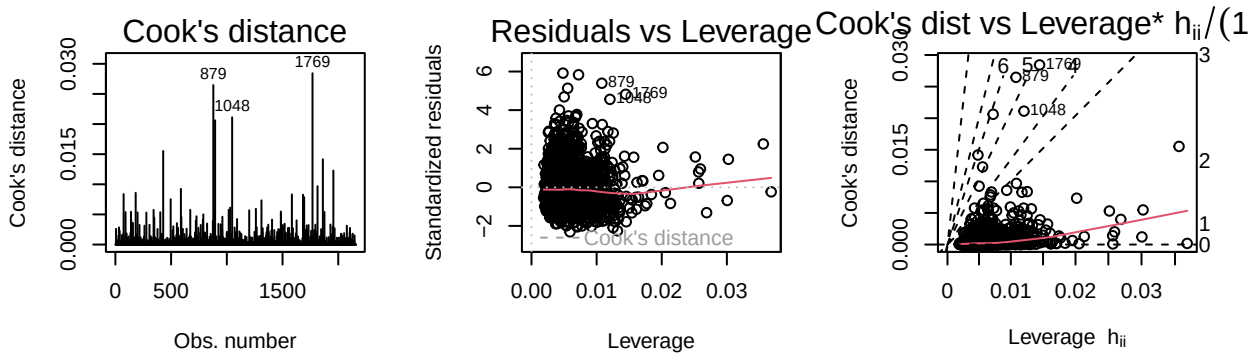
```
##
## Call:
## lm(formula = BMI ~ SleepHrsNight + Age + Gender + Race1 + TotChol +
##     BPDiaAve + BPSysAve + AlcoholYear + Smoke100 + DaysPhysHlthBad +
##     PhysActive, data = df3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.752  -4.236  -0.849   3.055  37.857
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     21.023150   1.610401  13.055  < 2e-16 ***
## SleepHrsNight   -0.212193   0.107400  -1.976 0.048314 *
## Age              0.012839   0.013495   0.951 0.341528
## Gender           0.514621   0.291331   1.766 0.077463 .
## Race1           -0.622971   0.122615  -5.081 4.09e-07 ***
## TotChol          0.076572   0.139325   0.550 0.582658
## BPDiaAve         0.054500   0.014049   3.879 0.000108 ***
## BPSysAve         0.066004   0.012027   5.488 4.55e-08 ***
## AlcoholYear     -0.009762   0.001533  -6.368 2.34e-10 ***
## Smoke100        -0.507830   0.287921  -1.764 0.077911 .
## DaysPhysHlthBad  0.066309   0.019785   3.352 0.000818 ***
## PhysActive      -1.260928   0.292769  -4.307 1.73e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.413 on 2140 degrees of freedom
## Multiple R-squared:  0.1029, Adjusted R-squared:  0.09826
## F-statistic: 22.31 on 11 and 2140 DF,  p-value: < 2.2e-16
```

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

```
## Anova Table (Type III tests)
##
## Response: BMI
##                 Sum Sq   Df  F value    Pr(>F)
## (Intercept)       7009    1 170.4228 < 2.2e-16 ***
## SleepHrsNight      161    1   3.9035 0.0483145 *
## Age                 37    1   0.9051 0.3415284
## Gender             128    1   3.1203 0.0774631 .
## Race1             1062    1  25.8136 4.086e-07 ***
## TotChol             12    1   0.3020 0.5826579
## BPDiaAve           619    1  15.0491 0.0001079 ***
## BPSysAve          1239    1  30.1167 4.550e-08 ***
## AlcoholYear       1668    1  40.5487 2.340e-10 ***
## Smoke100           128    1   3.1109 0.0779110 .
## DaysPhysHlthBad    462    1  11.2328 0.0008176 ***
## PhysActive         763    1  18.5494 1.730e-05 ***
## Residuals        88016 2140
## ---
## 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)
```

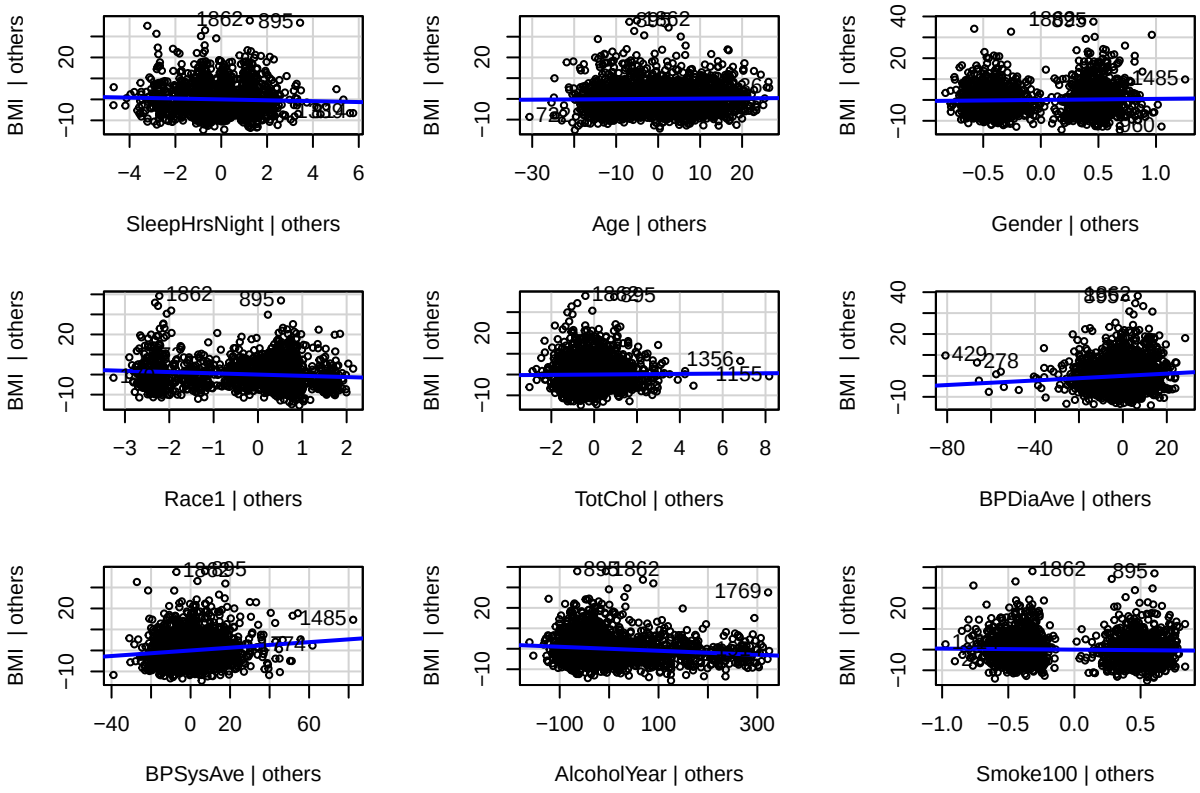
```
## m_2.h
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2152        0     2152        1 0.005576 0.002829 0.002806 0.003052
##      .25      .50      .75      .90      .95
## 0.003715 0.004755 0.006420 0.009242 0.010979
##
## lowest : 0.001859911 0.001924278 0.001978480 0.001997045 0.002020609
```

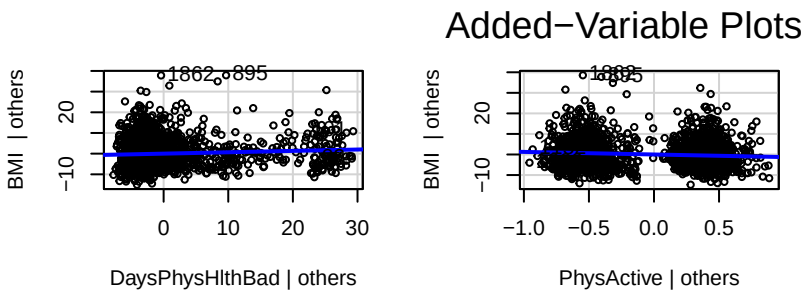
```
## highest: 0.026887311 0.030059375 0.030231906 0.035596339 0.036802689
```

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

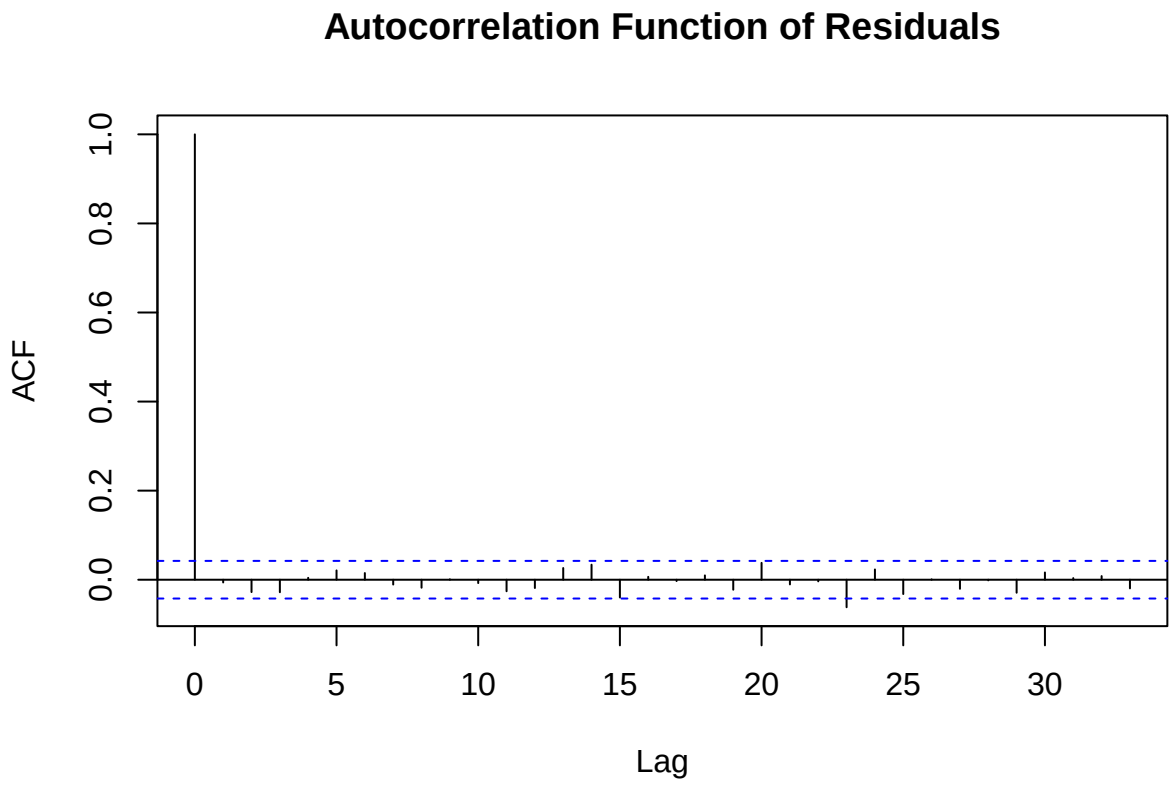
```
##       1155
## 0.03680269
```

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

#(1)Linear: 2 approaches

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

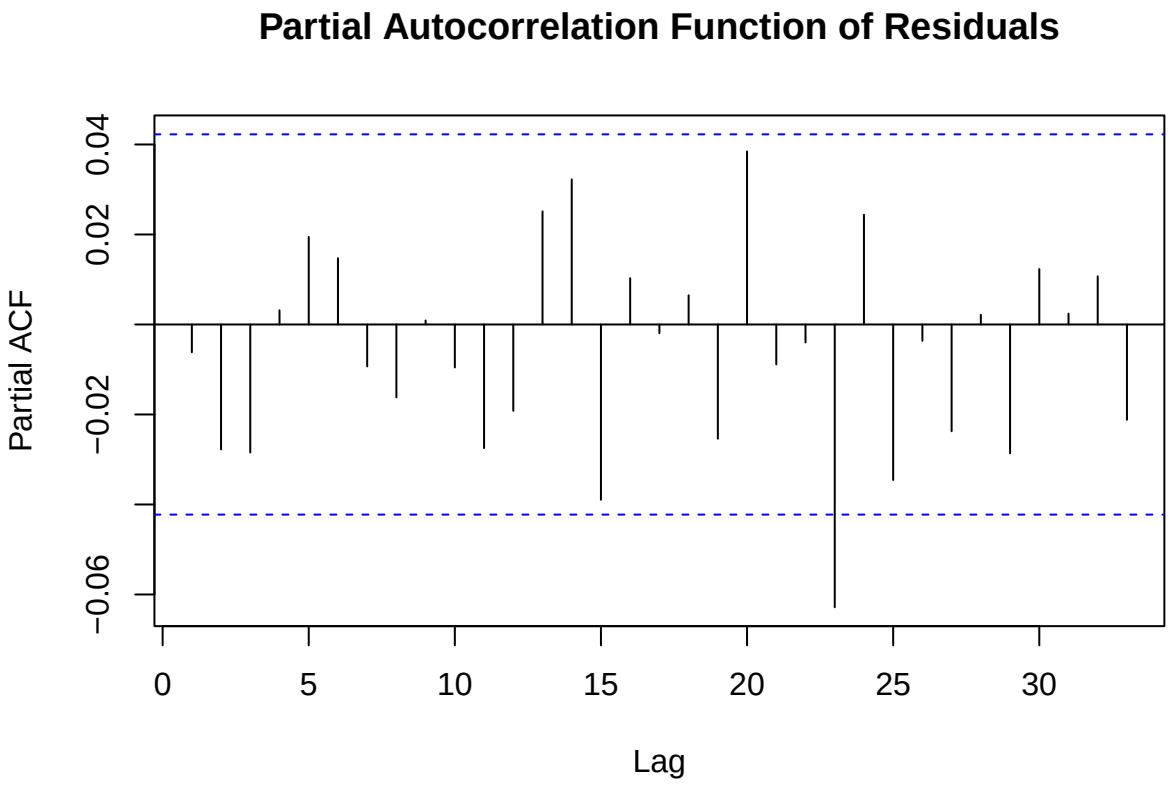




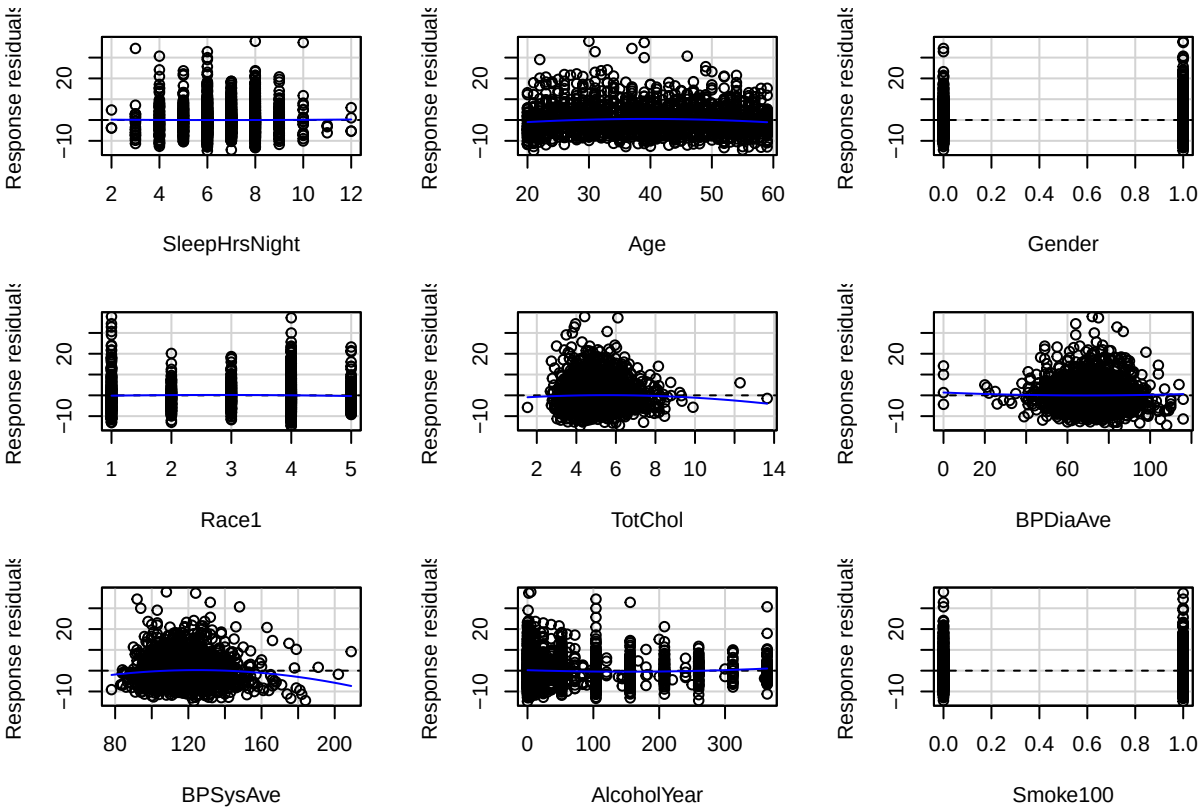

```
#(2)Independence:

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

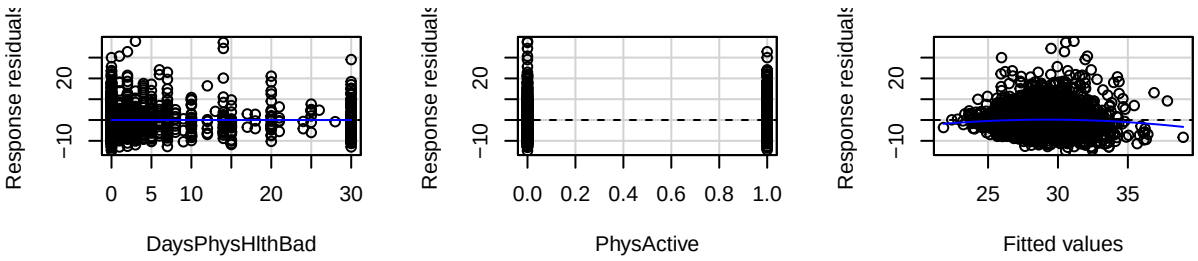



```
pacf(residuals, main = "Partial Autocorrelation Function of Residuals")
```



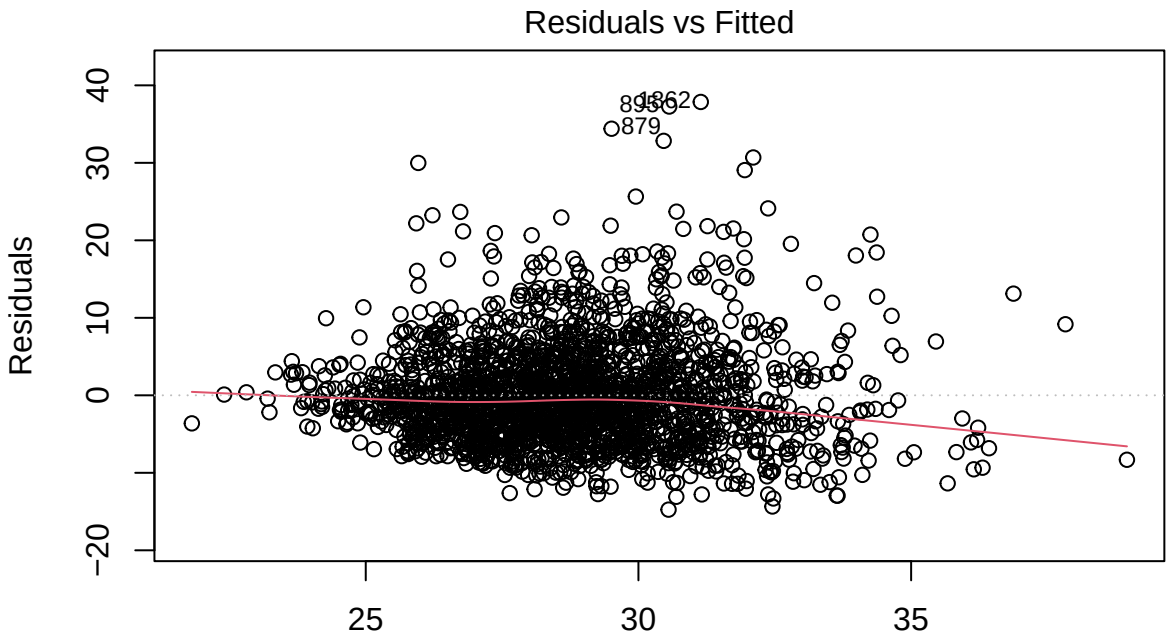

```
#(3)E: constant var: residuals-fitted values; transform for variance-stable...(total: 4 solutions)

car::residualPlots(m_2, type = "response")
```


Response residuals
SleepHrsNight
Age
Gender
Race1
TotChol
BPDiaAve
BPSysAve
AlcoholYear
Smoke100

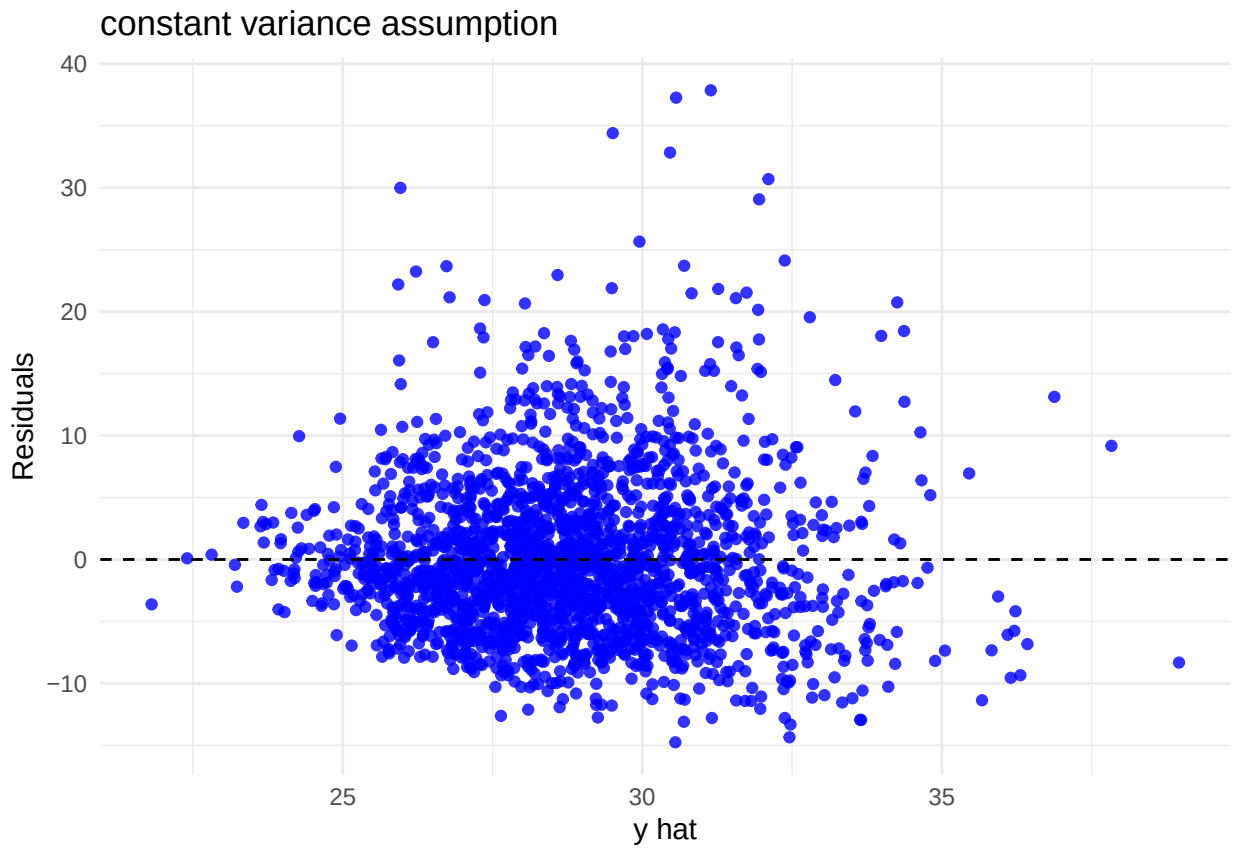



```
##                  Test stat Pr(>|Test stat|)
## SleepHrsNight       0.1824        0.8552520
## Age                -3.5792        0.0003523 ***
## Gender              0.1199        0.9045983
## Race1              -0.9232        0.3560240
## TotChol            -1.0692        0.2851166
## BPDiaAve            0.7104        0.4775572
## BPSysAve           -3.1904        0.0014413 **
## AlcoholYear         2.3916        0.0168588 *
## Smoke100           -0.9301        0.3524433
## DaysPhysHlthBad    -0.0482        0.9615513
## PhysActive          0.4249        0.6709259
## Tukey test         -2.0974        0.0359585 *
## ---
## 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()
```



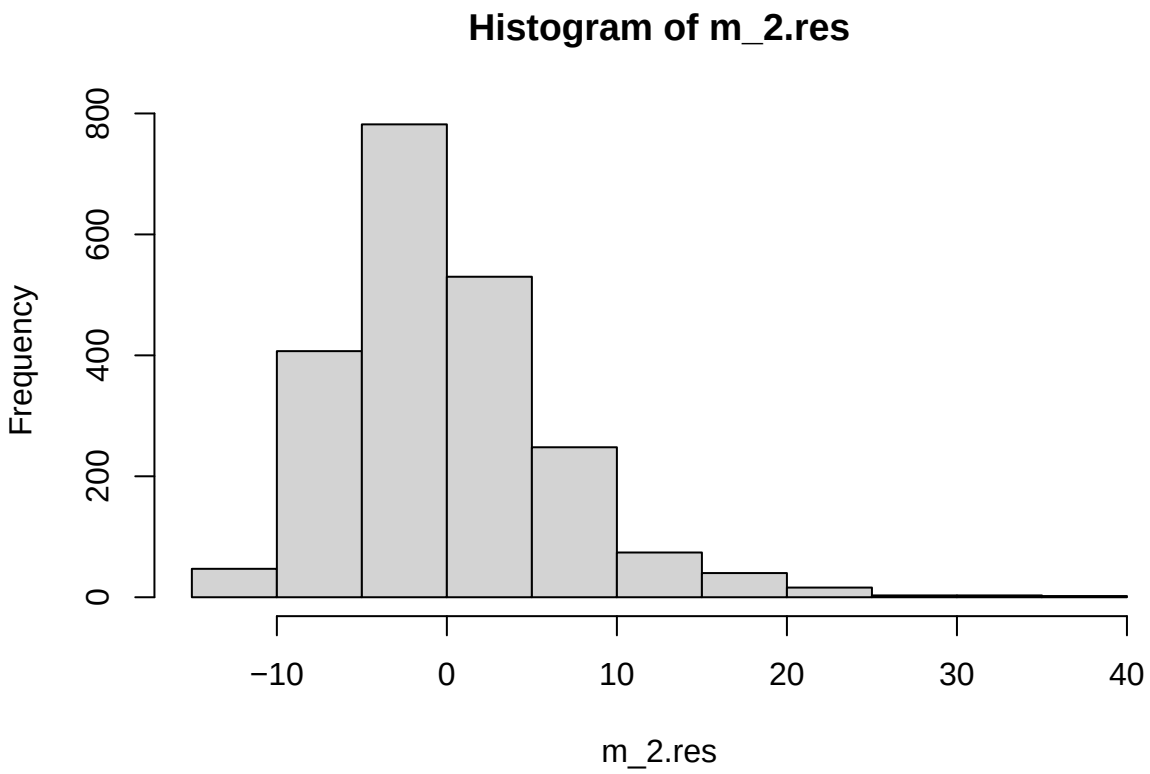

```
#conclusion: the constant variance assumption is basically not violated. The spread of the residuals ap


#(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 -9.645e-17      6.864    -8.3268
##        .10        .25        .50        .75        .90        .95
##    -7.0969    -4.2358    -0.8494     3.0548     8.0950    11.7305
##
## lowest : -14.75183 -14.34698 -13.31334 -13.09414 -12.95173
## highest:  30.69456  32.83699  34.40252  37.26529  37.85659
```

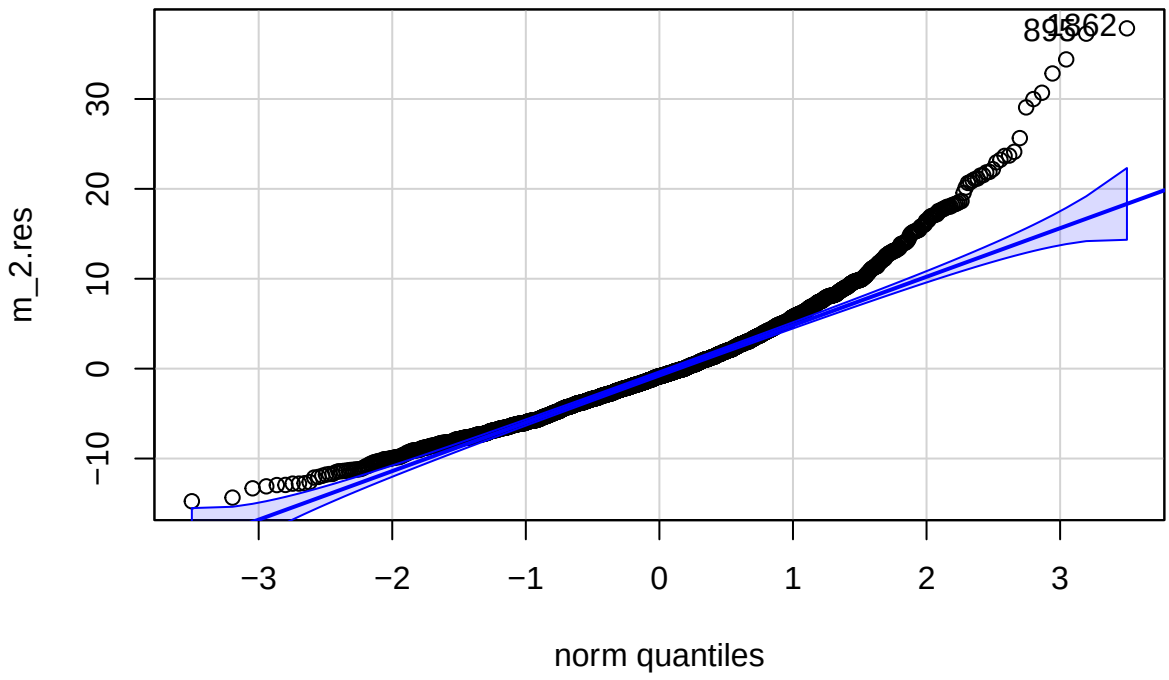
```
Hmisc::describe(m_2.res)$counts[c(".25", ".50", ".75")] #not symmetric
```

```
##       .25       .50       .75
## "-4.2358" "-0.8494" " 3.0548"
```

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



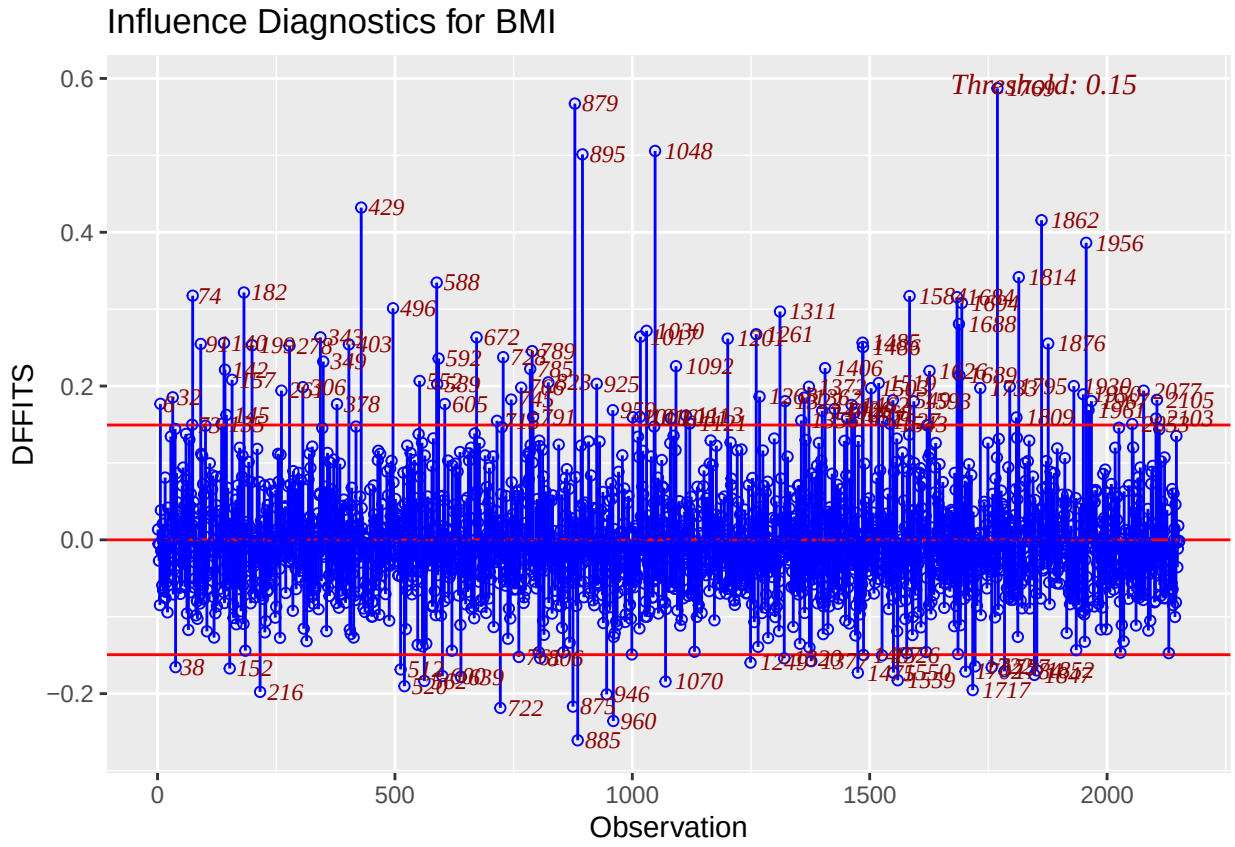

```
# Q-Q plot
qq.m_2.res = car::qqPlot(m_2.res)
```




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

```
##     1862      895
## 37.85659 37.26529
```

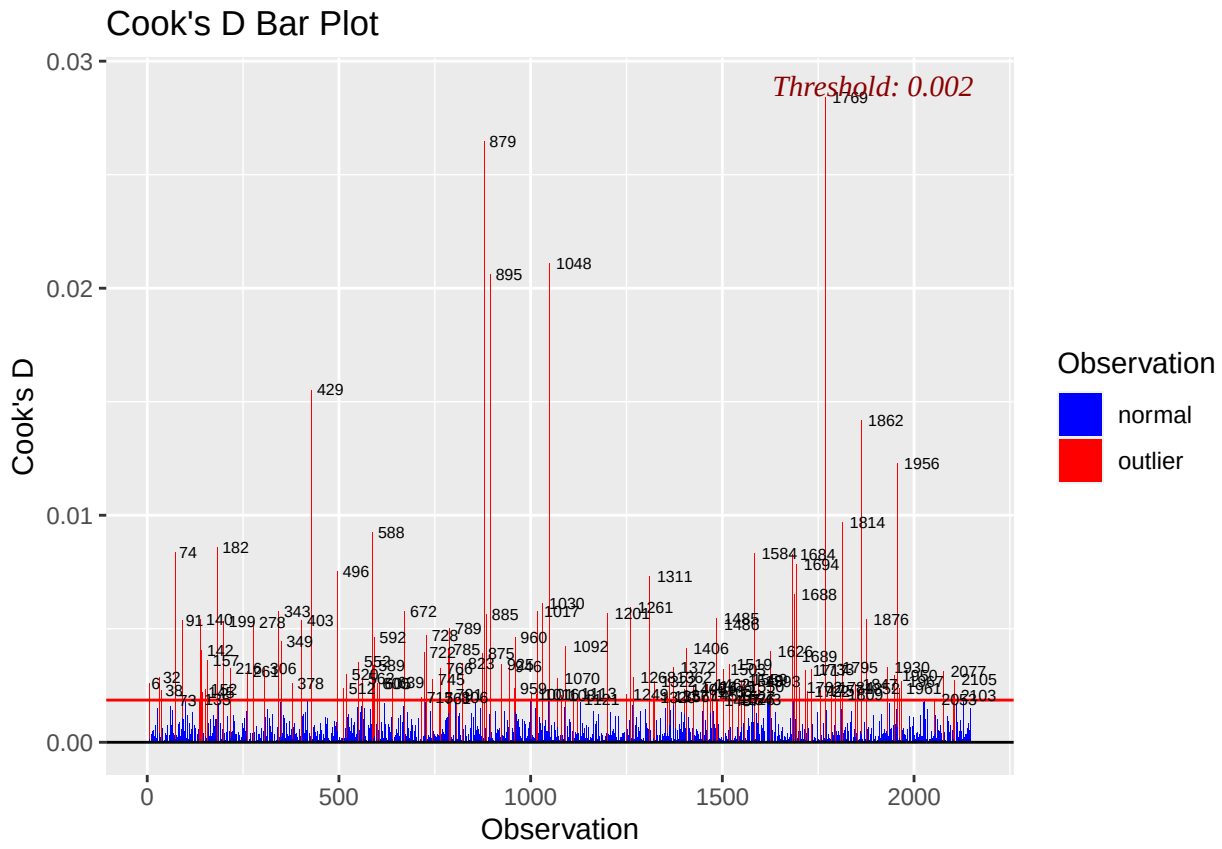
```
################ influential observations  ##################

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
## 1769 30.69456 4.846434 0.014465665 0.8950540 0.5871589    0.02843087
## 879  34.40252 5.429323 0.010800527 0.8626254 0.5673176    0.02646855
## 1048 29.06060 4.580162 0.012045991 0.9053925 0.5057475    0.02111790
## 895  37.26529 5.877363 0.007221976 0.8357826 0.5012847    0.02061736
## 429  14.15086 2.249012 0.035596339 1.0136031 0.4320808    0.01552837
## 1862 37.85659 5.964869 0.004832020 0.8290200 0.4156398    0.01416744
```

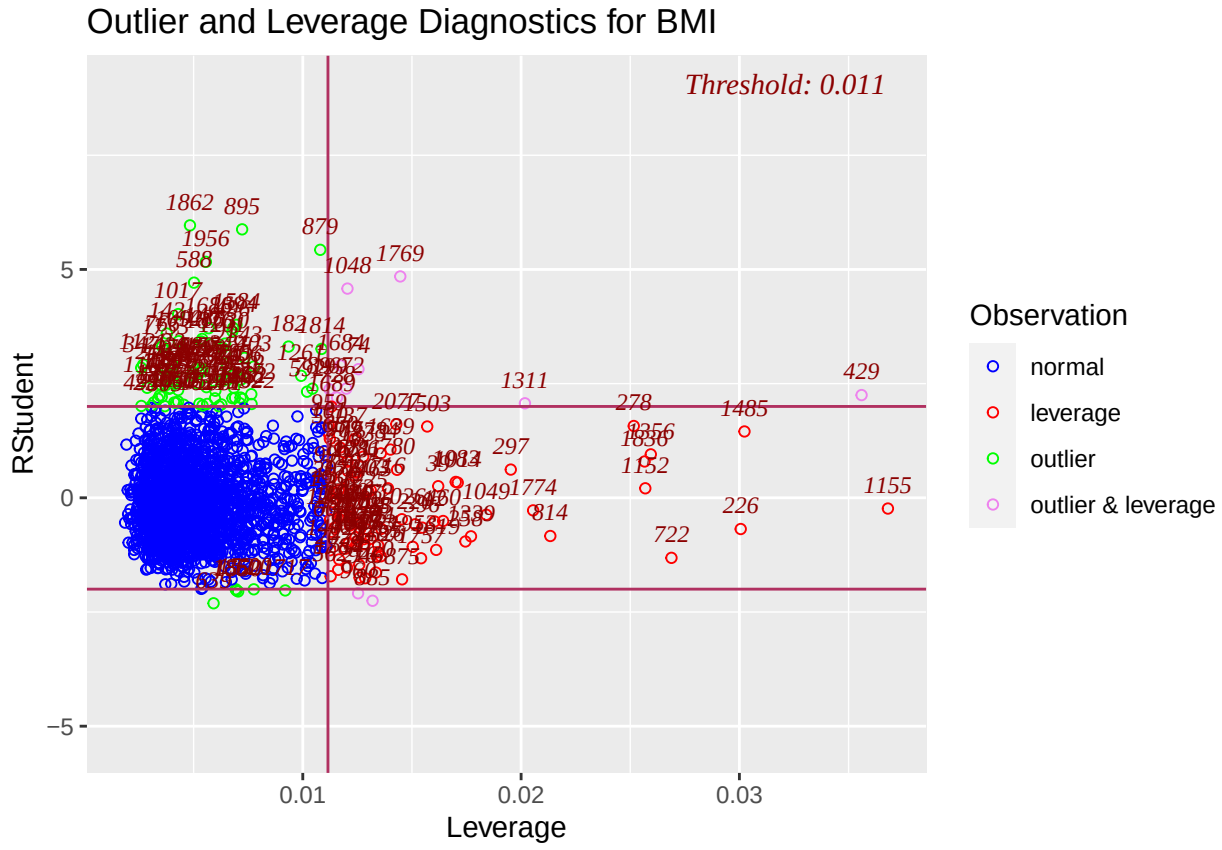
```
#From the plot above, we can see 2 observations with the largest (magnitude) of DFFITS, observation 879

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




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

```
##      Residual Rstudent    HatDiagH  CovRatio    DFFITS COOKsDistance
## 1769 30.69456 4.846434 0.014465665 0.8950540 0.5871589    0.02843087
## 879  34.40252 5.429323 0.010800527 0.8626254 0.5673176    0.02646855
## 1048 29.06060 4.580162 0.012045991 0.9053925 0.5057475    0.02111790
## 895  37.26529 5.877363 0.007221976 0.8357826 0.5012847    0.02061736
## 429  14.15086 2.249012 0.035596339 1.0136031 0.4320808    0.01552837
## 1862 37.85659 5.964869 0.004832020 0.8290200 0.4156398    0.01416744
```

```
#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 -1.479250 -0.2349714 0.03680269 1.043725 -0.04593007  0.0001758753
## 429  14.150865  2.2490119 0.03559634 1.013603  0.43208083  0.0155283738
## 1485  9.170096  1.4523768 0.03023191 1.024781  0.25643549  0.0054770905
## 226  -4.317916 -0.6835564 0.03005937 1.034076 -0.12033494  0.0012070086
## 722  -8.317574 -1.3149702 0.02688731 1.023438 -0.21857880  0.0039800348
## 1356  6.022294  0.9514512 0.02594104 1.027177  0.15527002  0.0020091538
```

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

```
##      Residual Rstudent    HatDiagH  CovRatio    DFFITS COOKsDistance
## 1862 37.85659 5.964869 0.004832020 0.8290200 0.4156398    0.01416744
## 895  37.26529 5.877363 0.007221976 0.8357826 0.5012847    0.02061736
## 879  34.40252 5.429323 0.010800527 0.8626254 0.5673176    0.02646855
## 1956 32.83699 5.165249 0.005561149 0.8714776 0.3862642    0.01228591
## 1769 30.69456 4.846434 0.014465665 0.8950540 0.5871589    0.02843087
## 588  29.98670 4.710724 0.005018304 0.8929612 0.3345481    0.00923542
```

```
#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
```

```
#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(
  BMI ~ SleepHrsNight + Age + Gender + Race1 + TotChol + BPDiaAve + BPSysAve + AlcoholYear + Smoke100 +
    DaysPhysHlthBad + PhysActive,
  rm2.df3
)
## Before removing these observations, the estimated coefficients are:
summary(m_2)$coef
```

```
##                     Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)     21.023150172 1.61040076 13.0546077 1.560026e-37
## SleepHrsNight   -0.212193168 0.10740028 -1.9757226 4.831448e-02
## Age              0.012838871 0.01349526  0.9513613 3.415284e-01
## Gender           0.514620931 0.29133077  1.7664489 7.746310e-02
## Race1           -0.622970728 0.12261497 -5.0807068 4.086286e-07
## TotChol          0.076571710 0.13932513  0.5495901 5.826579e-01
## BPDiaAve         0.054499935 0.01404885  3.8793173 1.079145e-04
## BPSysAve         0.066003942 0.01202724  5.4878708 4.549610e-08
## AlcoholYear     -0.009761821 0.00153300 -6.3677900 2.339900e-10
## Smoke100        -0.507829750 0.28792054 -1.7637844 7.791095e-02
## DaysPhysHlthBad  0.066308553 0.01978451  3.3515386 8.176133e-04
## PhysActive      -1.260928313 0.29276914 -4.3069031 1.730378e-05
```

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

```
##                    Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)     20.35681885 1.591402458 12.7917478 3.758113e-36
## SleepHrsNight   -0.16992002 0.105487052 -1.6108140 1.073684e-01
## Age              0.01603432 0.013256274  1.2095650 2.265800e-01
## Gender           0.38567462 0.286107176  1.3480075 1.777992e-01
## Race1           -0.49166686 0.120767754 -4.0711767 4.848398e-05
## TotChol          0.09239091 0.138702544  0.6661082 5.054140e-01
## BPDiaAve         0.05080608 0.013773193  3.6887658 2.309889e-04
## BPSysAve         0.06676170 0.011857120  5.6305153 2.034003e-08
## AlcoholYear     -0.01071753 0.001509265 -7.1011608 1.678800e-12
## Smoke100        -0.57220143 0.282360003 -2.0264961 4.283829e-02
## DaysPhysHlthBad  0.04676614 0.019585316  2.3878164 1.703545e-02
## PhysActive      -1.18819428 0.286816208 -4.1427027 3.566379e-05
```

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

```
##     (Intercept)   SleepHrsNight             Age          Gender           Race1
##        3.169512       19.922012       24.888895       25.056561       21.077052
##         TotChol        BPDiaAve        BPSysAve     AlcoholYear        Smoke100
##       20.659324        6.777717        1.148045        9.790278       12.675840
## DaysPhysHlthBad      PhysActive
##       29.471935        5.768293
```

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

###################   multicollinearity   ######################
#Pearson correlations
```

```r
var2 = c(
  "BMI",
  "SleepHrsNight",
  "Age",
  "Gender",
  "Race1",
  "TotChol",
  "BPDiaAve",
  "BPSysAve",
  "AlcoholYear",
  "Smoke100",
  "DaysPhysHlthBad",
  "PhysActive"
)
newData2 = df3[, var2]
library("corrplot")
```

```
## corrplot 0.92 loaded
```

```r
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)
```

```
##   SleepHrsNight             Age          Gender           Race1         TotChol
##        1.035419        1.223319        1.106167        1.045711        1.122357
##        BPDiaAve        BPSysAve     AlcoholYear        Smoke100 DaysPhysHlthBad
##        1.447702        1.542999        1.091195        1.078534        1.057582
##      PhysActive
##        1.093222
```

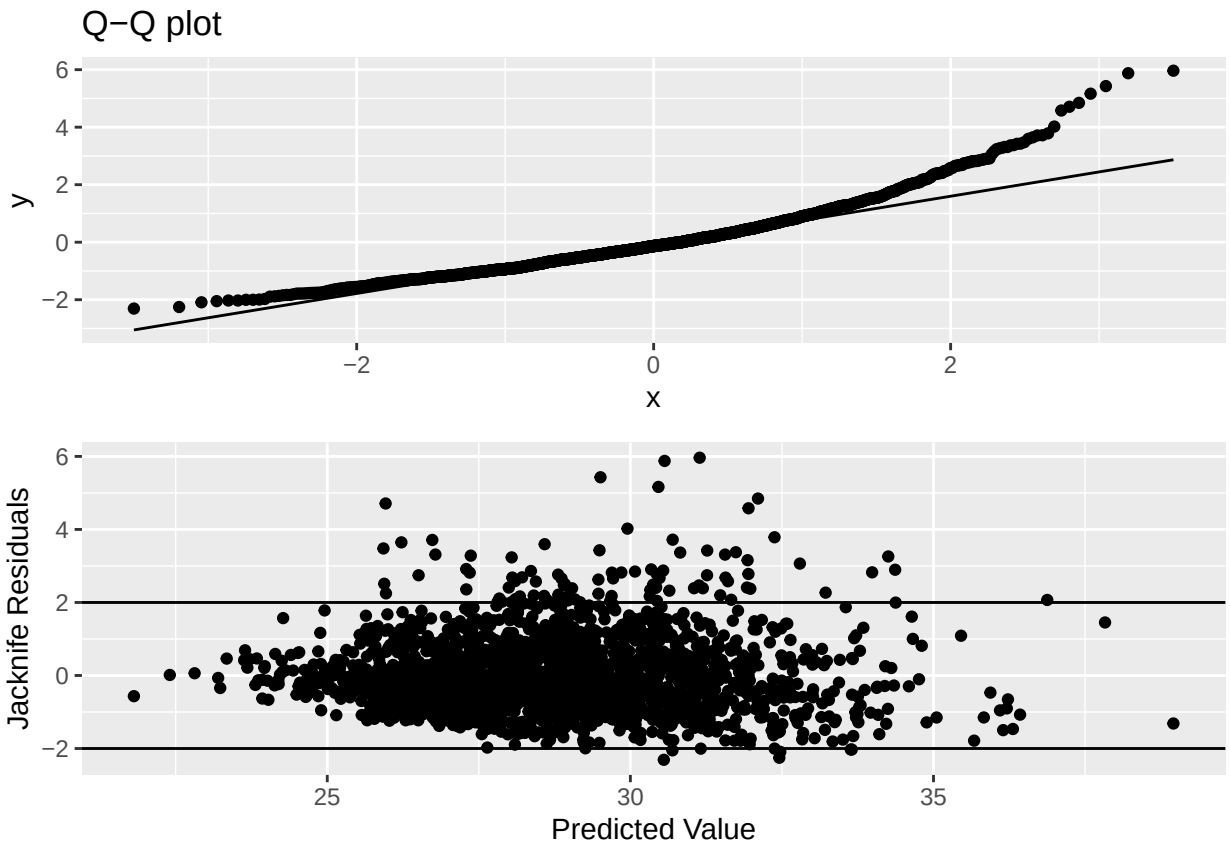
```r
#From the VIF values in the output above, once again we do not observe any potential collinearity issue

################## using log-transformed BMI  ###################
# log BMI
df3$logBMI = log(df3$BMI + 1)
m_2.log = lm(
  logBMI ~ SleepHrsNight + Age + Gender + Race1 + TotChol + BPDiaAve + BPSysAve + AlcoholYear + Smoke10
```

```
    DaysPhysHlthBad + PhysActive,
  df3
)
p21.log = ols_plot_resid_lev(m_2.log)
p22.log = ols_plot_cooksd_bar(m_2.log)
library(gridExtra)
```

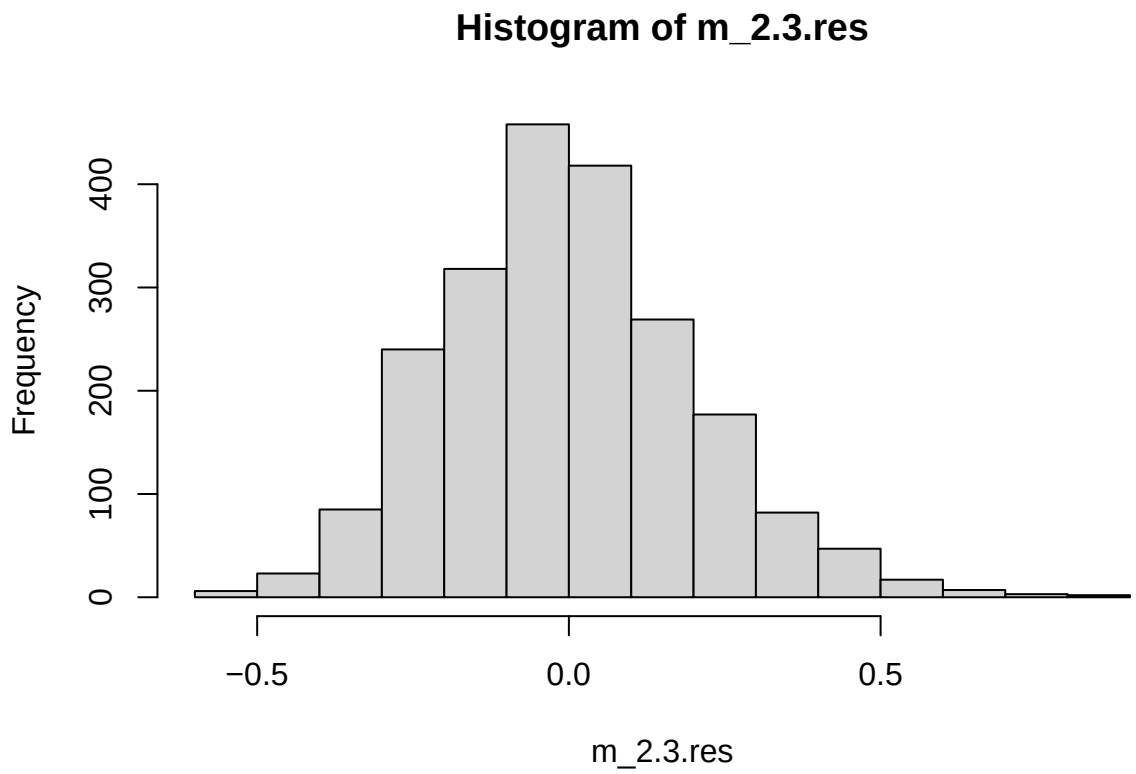
```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

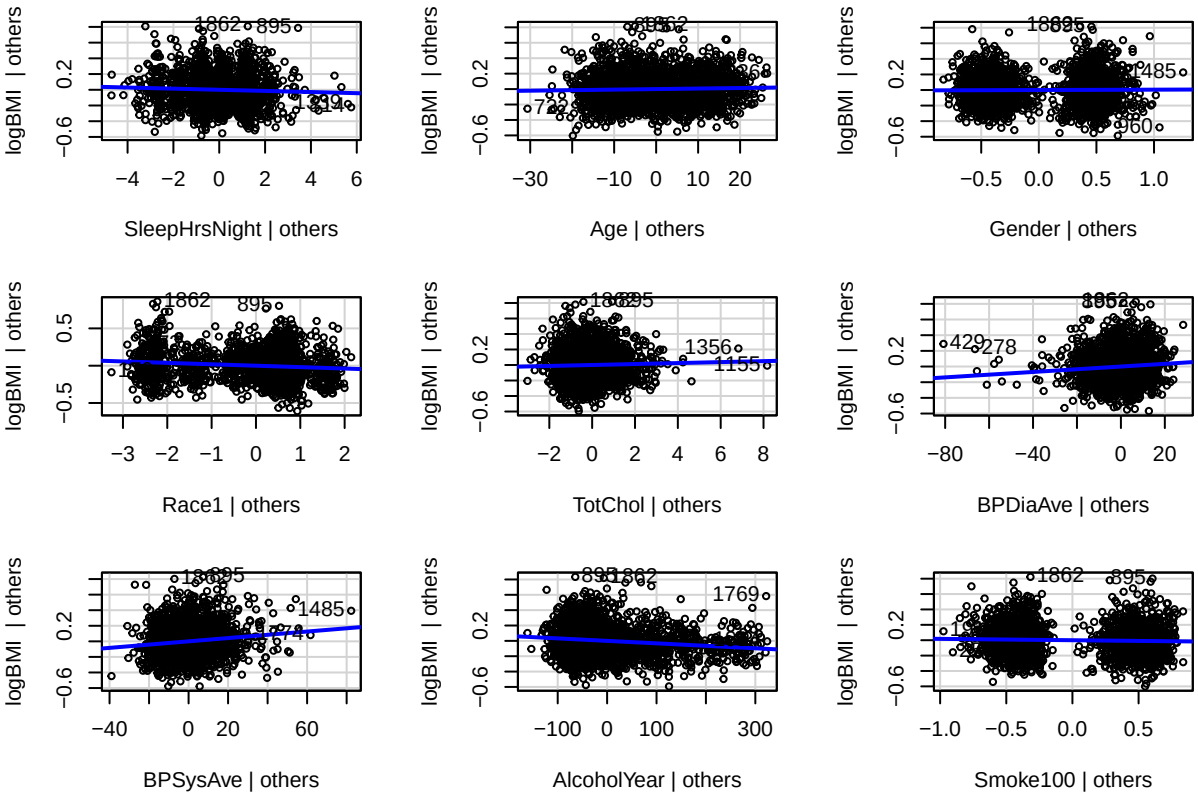
```
p23.log = ggplot(m_2.log, aes(sample = rstudent(m_2.log))) + geom_qq() + stat_qq_line() +
  labs(title = "Q-Q plot")
p24.log = ggplot() + geom_point(aes(y = rstudent(m_2.log), x = m_2.log$fitted.values)) + labs(x = "Predi
  geom_hline(yintercept = c(-2, 2))
grid.arrange(p23.log, p24.log, nrow = 2)


p23 = ggplot(m_2, aes(sample = rstudent(m_2))) + geom_qq() + stat_qq_line() +
  labs(title = "Q-Q plot")
p24 = ggplot() + geom_point(aes(y = rstudent(m_2), x = m_2$fitted.values)) + labs(x = "Predicted Value"
  geom_hline(yintercept = c(-2, 2))
grid.arrange(p23, p24, nrow = 2)


m_2.3.yhat = m_2.log$fitted.values
m_2.3.res = m_2.log$residuals
m_2.3.h = hatvalues(m_2.log)
m_2.3.r = rstandard(m_2.log)
m_2.3.rr = rstudent(m_2.log)

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

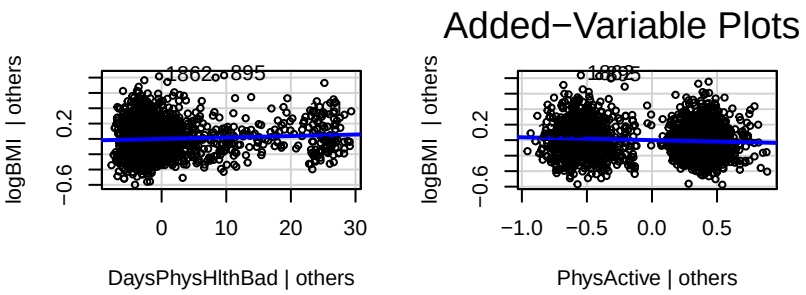



```
hist(m_2.3.res, breaks = 15)
```




```
car::avPlots(m_2.log)
```


logBMI | others
SleepHrsNight | others
Age | others
Gender | others
Race1 | others
TotChol | others
BPDiaAve | others
BPSysAve | others
AlcoholYear | others
Smoke100 | others




```
#After looking at residuals from models using the log-transformed (natural log scale) BMI adjusted for

#collinearity diagnostics

car::vif(m_2.log)
```

```
##   SleepHrsNight             Age          Gender           Race1         TotChol
##        1.035419        1.223319        1.106167        1.045711        1.122357
##        BPDiaAve        BPSysAve     AlcoholYear        Smoke100 DaysPhysHlthBad
##        1.447702        1.542999        1.091195        1.078534        1.057582
##      PhysActive
##        1.093222
```

```
#The VIF from both the models are the same. None of the VIF values are greater than 10. So there are no
```