model4

Liancheng

2023-11-25

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

```
##          used (Mb) gc trigger (Mb) max used (Mb)
## Ncells 472218 25.3    1018764 54.5   660860 35.3
## Vcells 896957  6.9    8388608 64.0  1800812 13.8
```

```
set.seed(123)
library(car)
library(ggplot2)
library(olsrr)
library(lmtest)
################ (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"
```

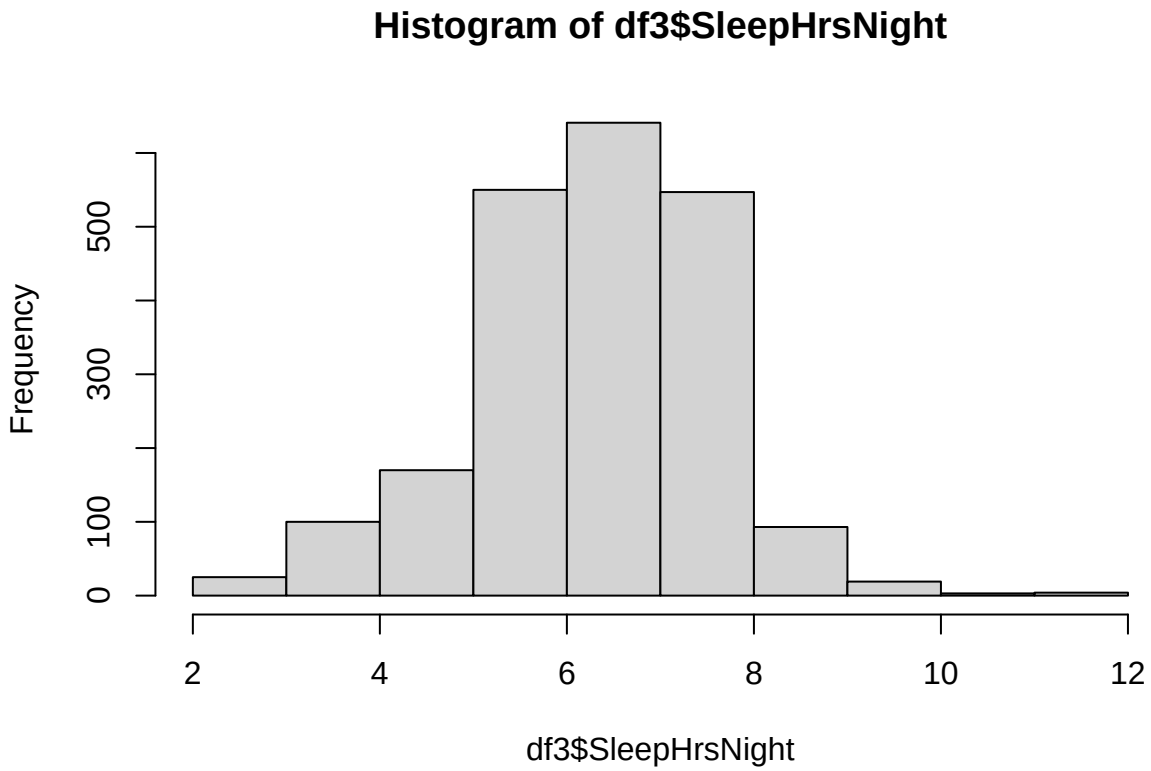
```
# df$BPSysAve
library(dplyr)

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
```

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




```r
# colSums(is.na(df2)) / nrow(df2)
fit0 <-
  lm(SleepHrsNight ~ .,
     data = df3)
#data type
df3$Gender <- ifelse(df3$Gender == "male", 0, 1)
```

```r
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 <- df3 %>%
  mutate(
    HealthGen = case_when(
      HealthGen == 'Poor' ~ 1,
      HealthGen == 'Fair' ~ 2,
      HealthGen == 'Good' ~ 3,
      HealthGen == 'Vgood' ~ 4,
      HealthGen == 'Excellent' ~ 5,
      TRUE ~ NA_integer_  # Default value if none of the conditions are met
    )
  )

## model_4 add additional risk factors ##
m_full = lm(
  BMI ~ SleepHrsNight + Age + Gender + factor(Race1)  + Poverty + TotChol + BPDiaAve + BPSysAve + Alcoho
    DaysPhysHlthBad + factor(HealthGen) + PhysActive + SleepHrsNight*Age + SleepHrsNight*Gender,
  df3
)
summary(m_full)
```

```
##
## Call:
## lm(formula = BMI ~ SleepHrsNight + Age + Gender + factor(Race1) +
##     Poverty + TotChol + BPDiaAve + BPSysAve + AlcoholYear + Smoke100 +
##     UrineFlow1 + DaysMentHlthBad + DaysPhysHlthBad + factor(HealthGen) +
##     PhysActive + SleepHrsNight * Age + SleepHrsNight * Gender,
##     data = df3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.011  -4.048  -0.582   3.184  35.959
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          29.497094   3.183685   9.265  < 2e-16 ***
## SleepHrsNight        -0.568047   0.378050  -1.503 0.133098
## Age                  -0.109637   0.062677  -1.749 0.080395 .
## Gender                3.598603   1.433671   2.510 0.012145 *
## factor(Race1)2       -1.999695   0.640341  -3.123 0.001815 **
## factor(Race1)3       -1.208467   0.561320  -2.153 0.031439 *
## factor(Race1)4       -1.490610   0.420779  -3.543 0.000405 ***
```

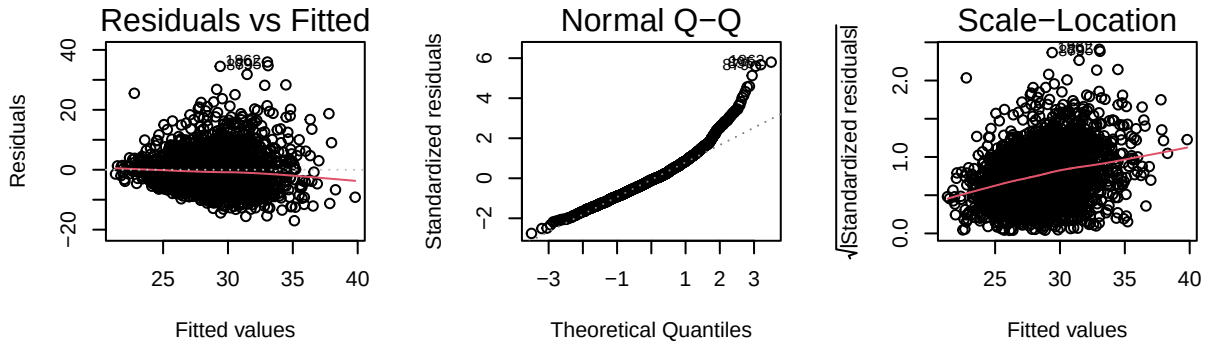
```
## factor(Race1)5       -3.291829   0.630346  -5.222 1.94e-07 ***
## Poverty               0.055441   0.091718   0.604 0.545594
## TotChol               0.012610   0.135822   0.093 0.926038
## BPDiaAve              0.058604   0.013689   4.281 1.94e-05 ***
## BPSysAve              0.051654   0.011799   4.378 1.26e-05 ***
## AlcoholYear          -0.008707   0.001513  -5.753 1.00e-08 ***
## Smoke100             -0.860788   0.287629  -2.993 0.002797 **
## UrineFlow1           -0.096878   0.142283  -0.681 0.496020
## DaysMentHlthBad      -0.032321   0.018003  -1.795 0.072745 .
## DaysPhysHlthBad       0.014228   0.020914   0.680 0.496377
## factor(HealthGen)2   -2.349775   1.001810  -2.346 0.019091 *
## factor(HealthGen)3   -4.101774   0.991935  -4.135 3.69e-05 ***
## factor(HealthGen)4   -5.805697   1.018118  -5.702 1.35e-08 ***
## factor(HealthGen)5   -7.663419   1.075785  -7.124 1.43e-12 ***
## PhysActive           -0.864154   0.294685  -2.932 0.003399 **
## SleepHrsNight:Age     0.017362   0.009010   1.927 0.054117 .
## SleepHrsNight:Gender -0.459043   0.206678  -2.221 0.026452 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.228 on 2128 degrees of freedom
## Multiple R-squared:  0.1587, Adjusted R-squared:  0.1496
## F-statistic: 17.46 on 23 and 2128 DF,  p-value: < 2.2e-16
```

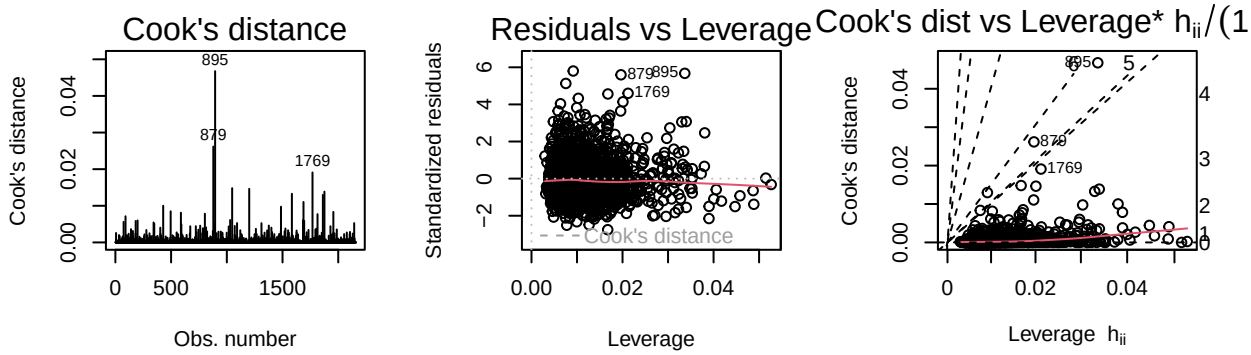
```
car::Anova(m_full, type = "III")
```

```
## Anova Table (Type III tests)
##
## Response: BMI
##                       Sum Sq   Df F value    Pr(>F)
## (Intercept)             3329    1 85.8417 < 2.2e-16 ***
## SleepHrsNight             88    1  2.2577  0.133098
## Age                      119    1  3.0598  0.080395 .
## Gender                   244    1  6.3004  0.012145 *
## factor(Race1)           1144    4  7.3727 6.787e-06 ***
## Poverty                   14    1  0.3654  0.545594
## TotChol                    0    1  0.0086  0.926038
## BPDiaAve                 711    1 18.3273 1.943e-05 ***
## BPSysAve                 743    1 19.1637 1.258e-05 ***
## AlcoholYear             1284    1 33.0990 1.002e-08 ***
## Smoke100                 347    1  8.9563  0.002797 **
## UrineFlow1                18    1  0.4636  0.496020
## DaysMentHlthBad          125    1  3.2232  0.072745 .
## DaysPhysHlthBad           18    1  0.4628  0.496377
## factor(HealthGen)       4608    4 29.7014 < 2.2e-16 ***
## PhysActive               334    1  8.5994  0.003399 **
## SleepHrsNight:Age        144    1  3.7132  0.054117 .
## SleepHrsNight:Gender     191    1  4.9331  0.026452 *
## Residuals              82537 2128
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

```
plot(m_full, which = 2)
plot(m_full, which = 3)
plot(m_full, which = 4)
plot(m_full, which = 5)
plot(m_full, which = 6)
```





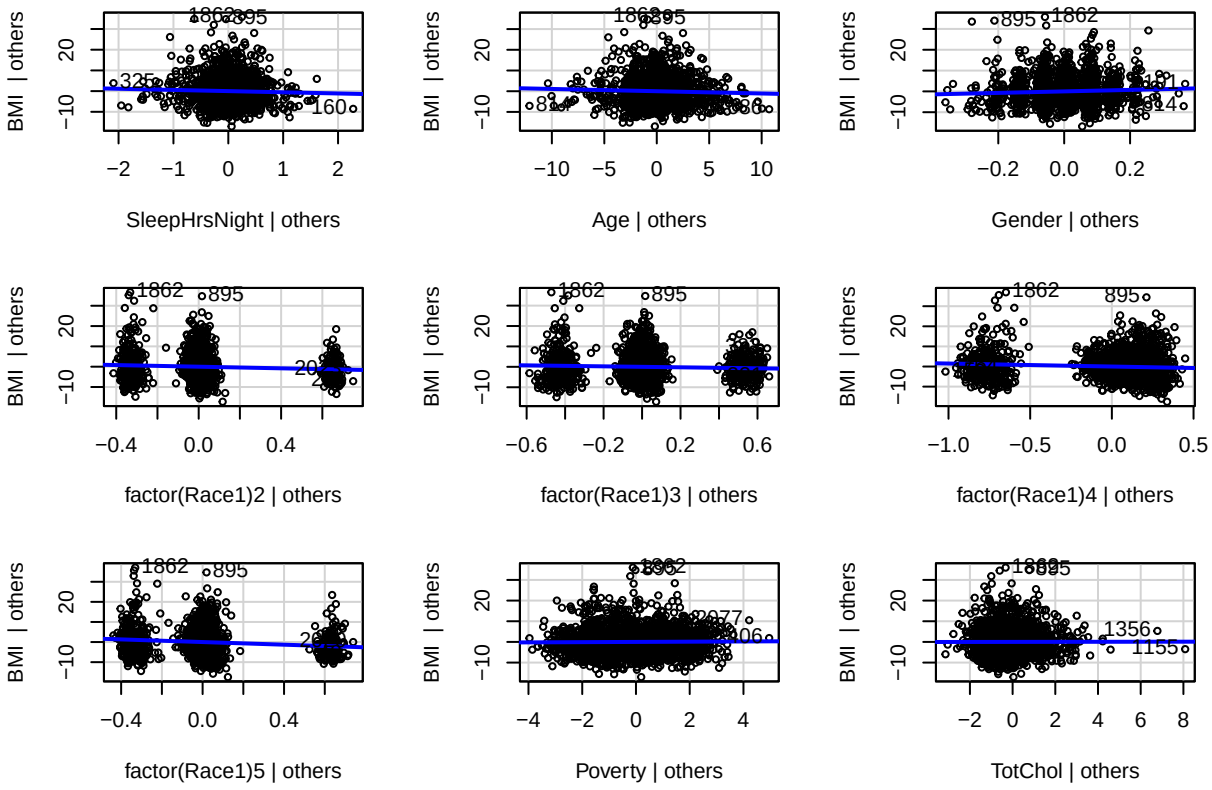

```
par(mfrow = c(1, 1)) # reset

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

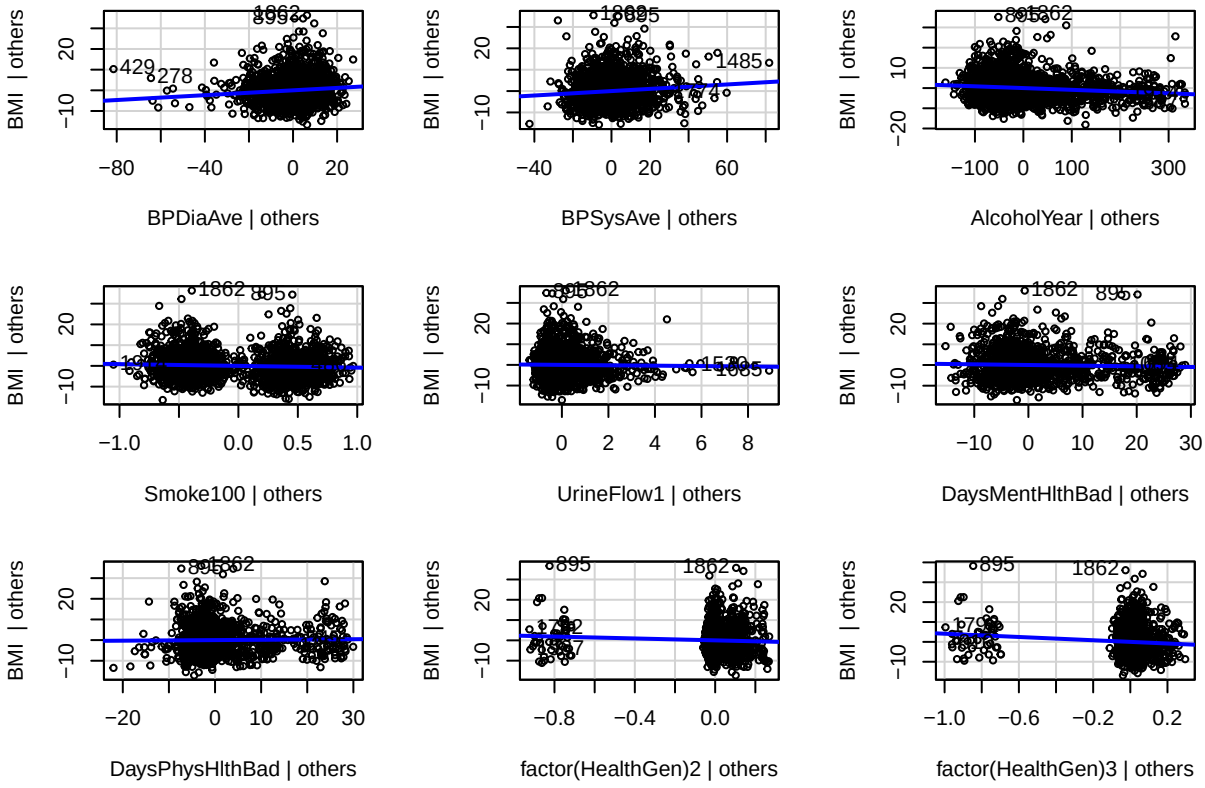
```
## m_full.h
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2152        0     2152        1  0.01115 0.006051 0.004809 0.005406
##      .25      .50      .75      .90      .95
## 0.007016 0.009823 0.013468 0.017982 0.022203
##
## lowest : 0.002961683 0.003144042 0.003307734 0.003408092 0.003459008
## highest: 0.045798704 0.048693974 0.048841649 0.051487655 0.052656174
```

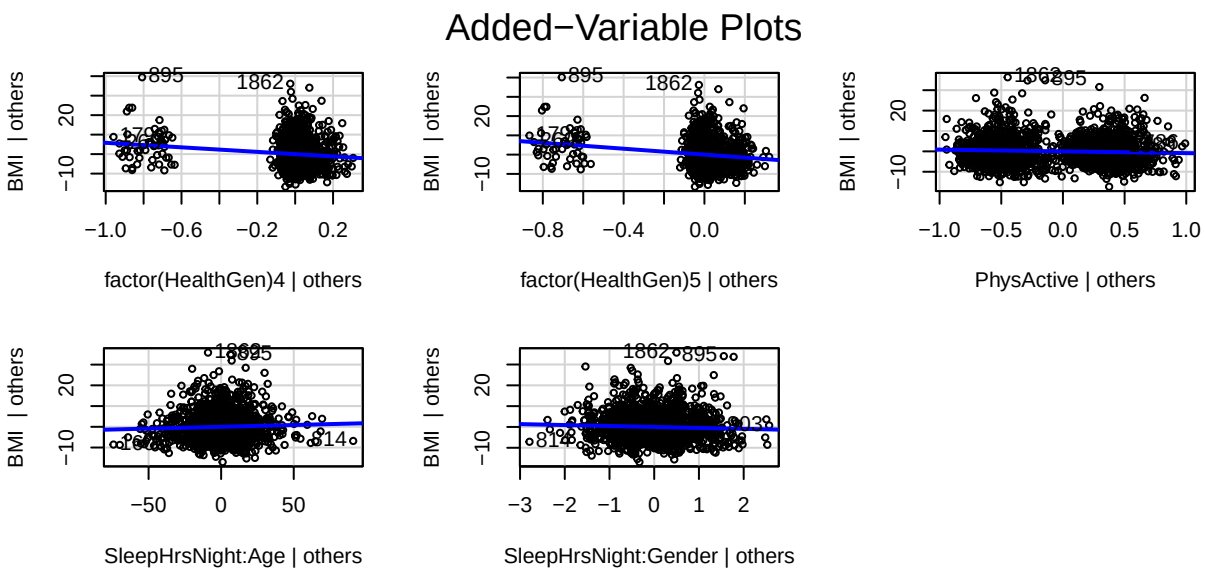
```
m_full.h[which.max(m_full.h)]
```

```
##        1685
## 0.05265617
```

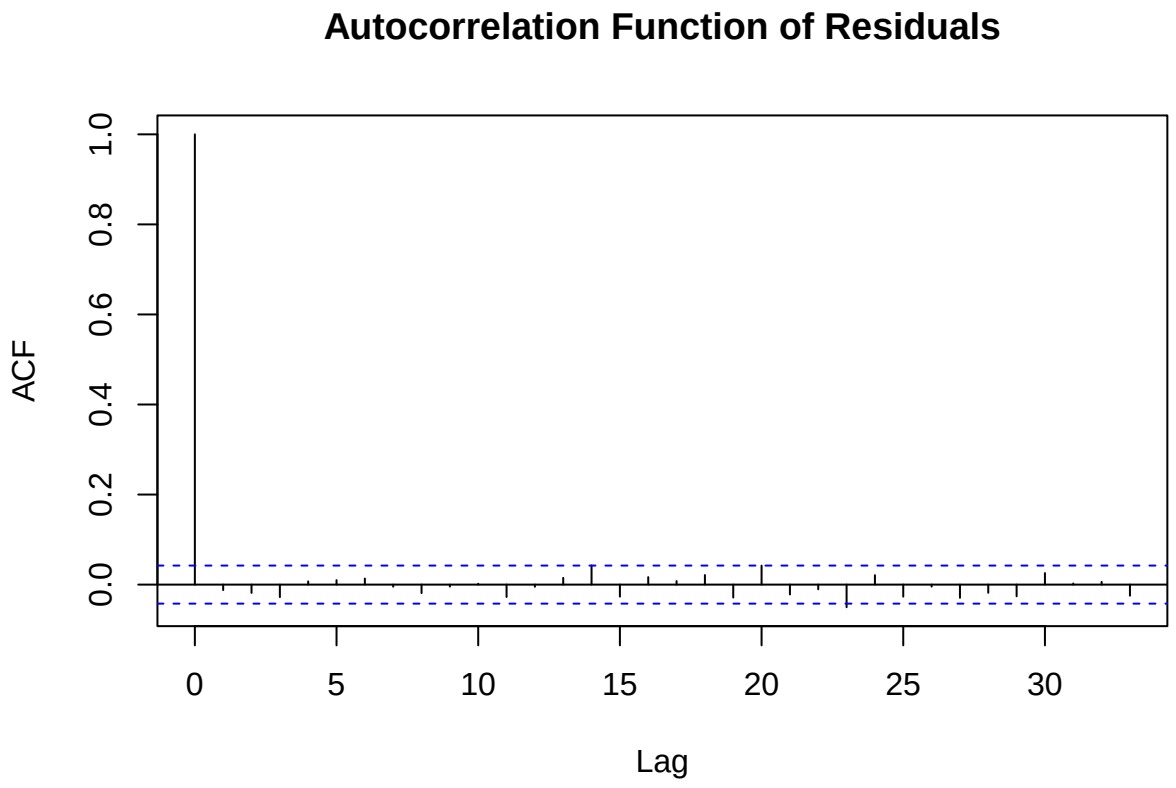
```
###################### Assumption:LINE ##############################

#(1)Linear: 2 approaches

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



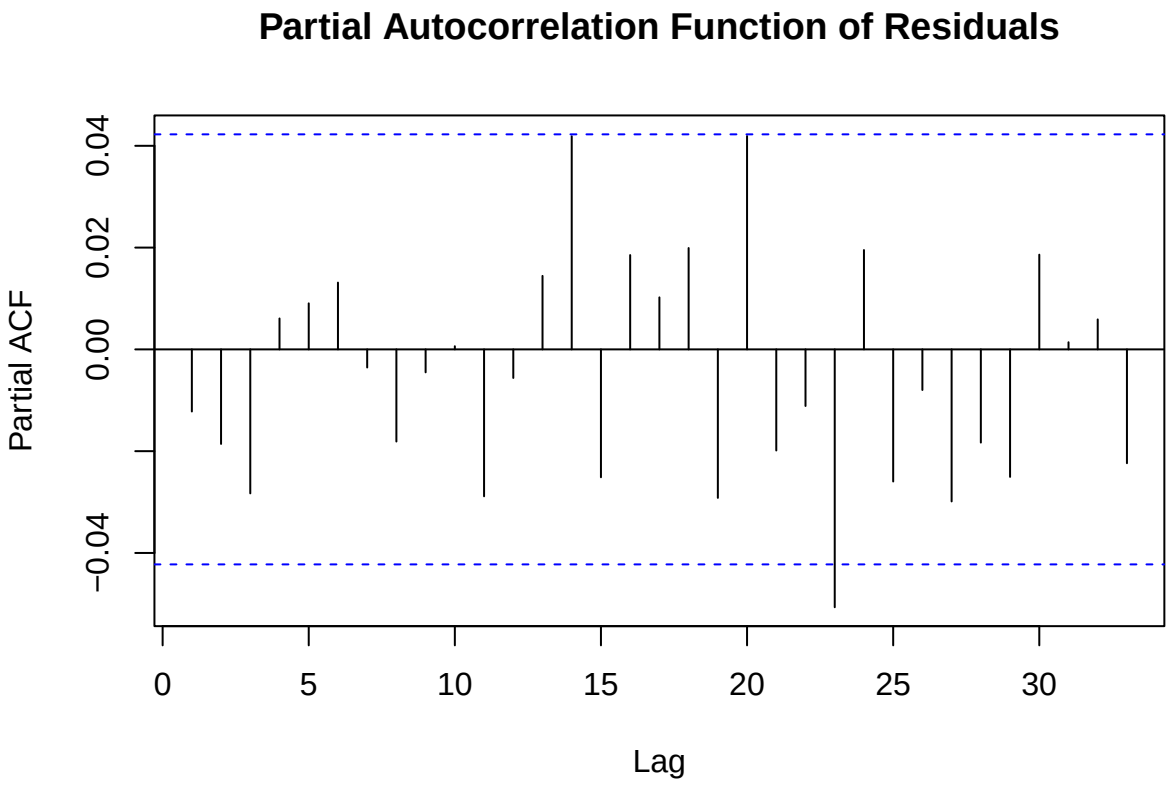

BMI | others
BPDiaAve | others
BPSysAve | others
AlcoholYear | others
Smoke100 | others
UrineFlow1 | others
DaysMentHlthBad | others
DaysPhysHlthBad | others
factor(HealthGen)2 | others
factor(HealthGen)3 | others




```
#(2)Independence:

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



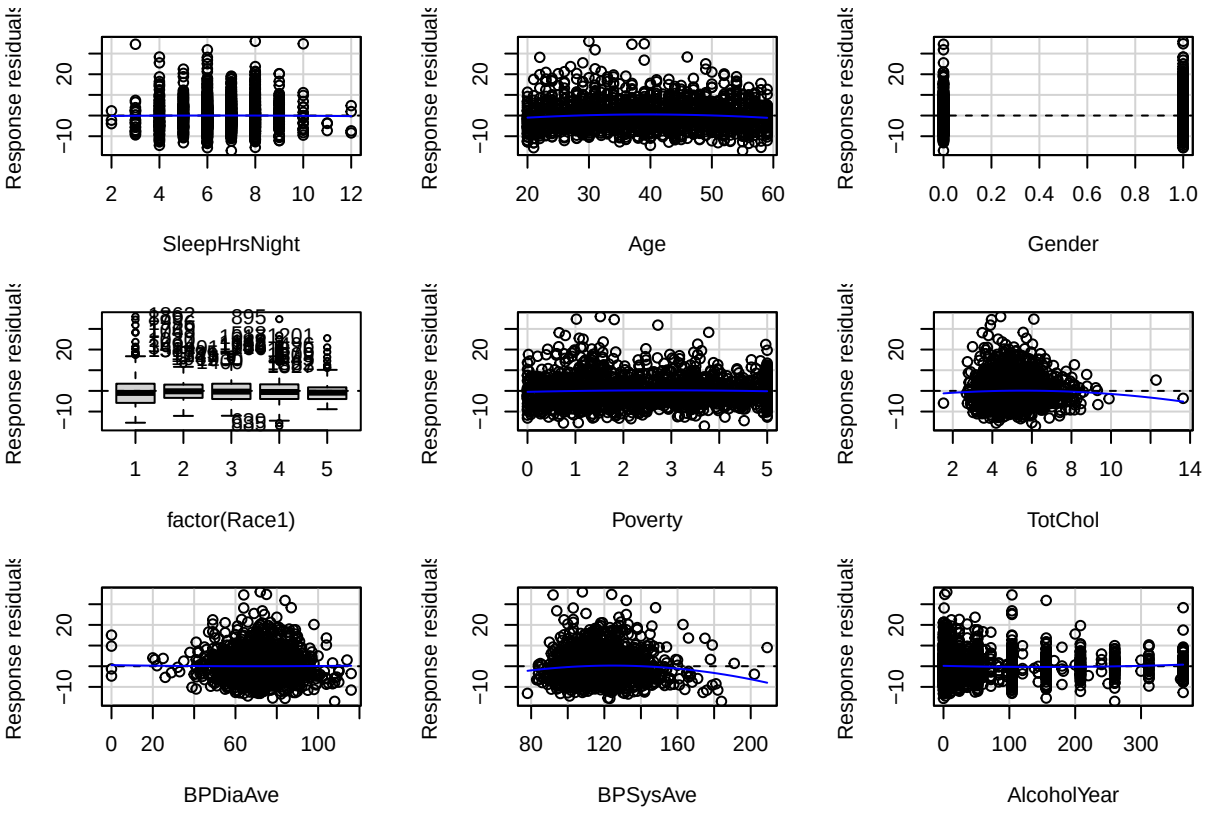

```
pacf(residuals, main = "Partial Autocorrelation Function of Residuals")
```

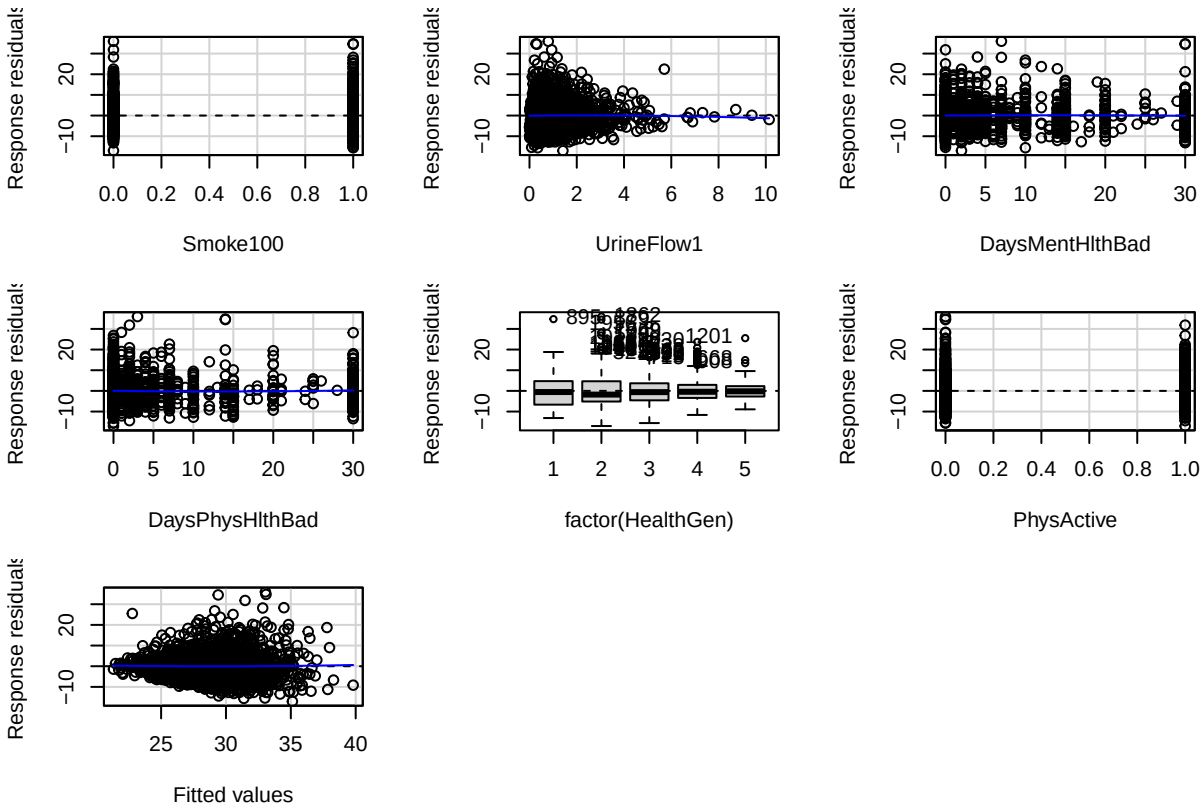



```
dw_test <- dwtest(m_full)
print(dw_test)
```

```
##
##  Durbin-Watson test
##
## data:  m_full
## DW = 2.0244, p-value = 0.7139
## alternative hypothesis: true autocorrelation is greater than 0
```

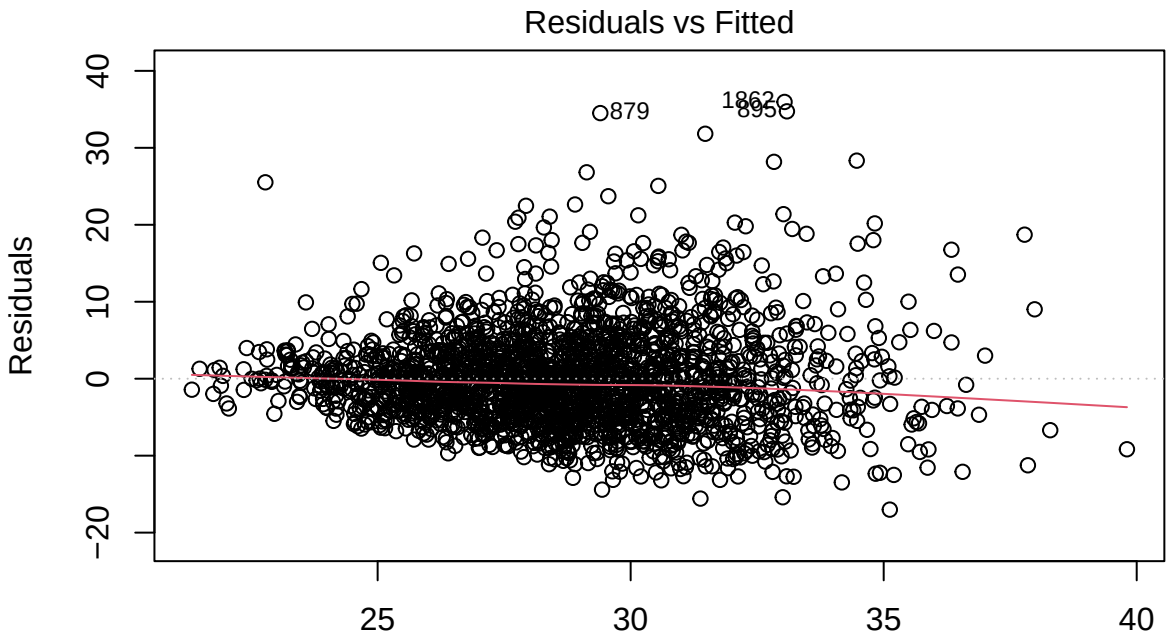
```
#(3)E: constant var: residuals-fitted values; transform for variance-stable...(total: 4 solutions)

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


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

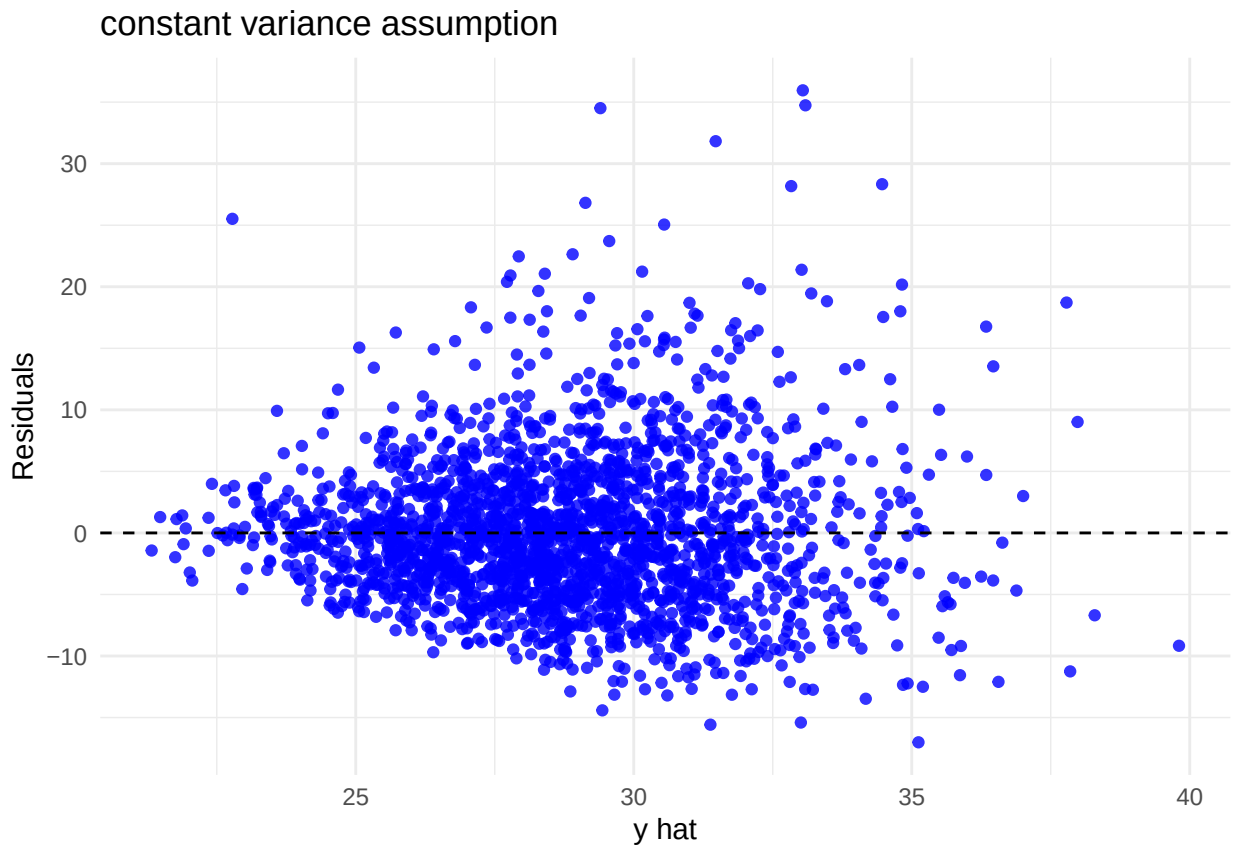



```
##                    Test stat Pr(>|Test stat|)
## SleepHrsNight        -0.2139        0.8306569
## Age                  -3.7721        0.0001663 ***
## Gender                0.4227        0.6725509
## factor(Race1)
## Poverty              -1.4176        0.1564669
## TotChol              -1.4430        0.1491614
## BPDiaAve              0.2951        0.7679152
## BPSysAve             -3.5794        0.0003520 ***
## AlcoholYear           1.8524        0.0641046 .
## Smoke100              0.0552        0.9560070
## UrineFlow1           -0.4682        0.6396812
## DaysMentHlthBad      -0.4875        0.6259662
## DaysPhysHlthBad       0.8219        0.4112378
## factor(HealthGen)
## PhysActive           -0.5312        0.5953404
## Tukey test            0.5422        0.5876498
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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




```
#or
ggplot(m_full, aes(x = m_full.yhat, y = m_full.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()
```



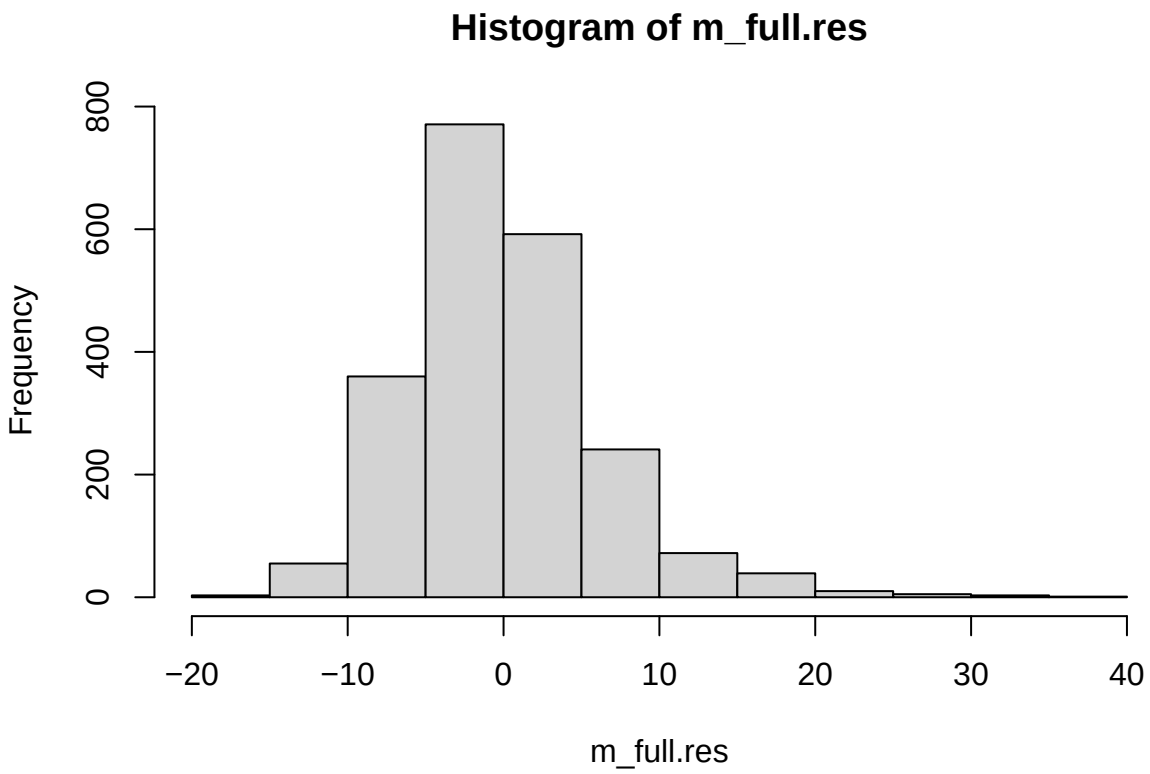

```
#conclusion: the constant variance assumption is basically not violated. The spread of the residuals ap


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

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

```
## m_full.res
##         n   missing  distinct      Info      Mean       Gmd       .05       .10
##      2152         0      2152         1 1.149e-16     6.668   -8.6636   -7.0207
##       .25       .50       .75       .90       .95
##   -4.0480   -0.5821    3.1841    7.4997   10.6326
##
## lowest : -17.01112 -15.57898 -15.40568 -14.41497 -13.47341
## highest:  28.33273  31.82618  34.51006  34.74208  35.95902
```

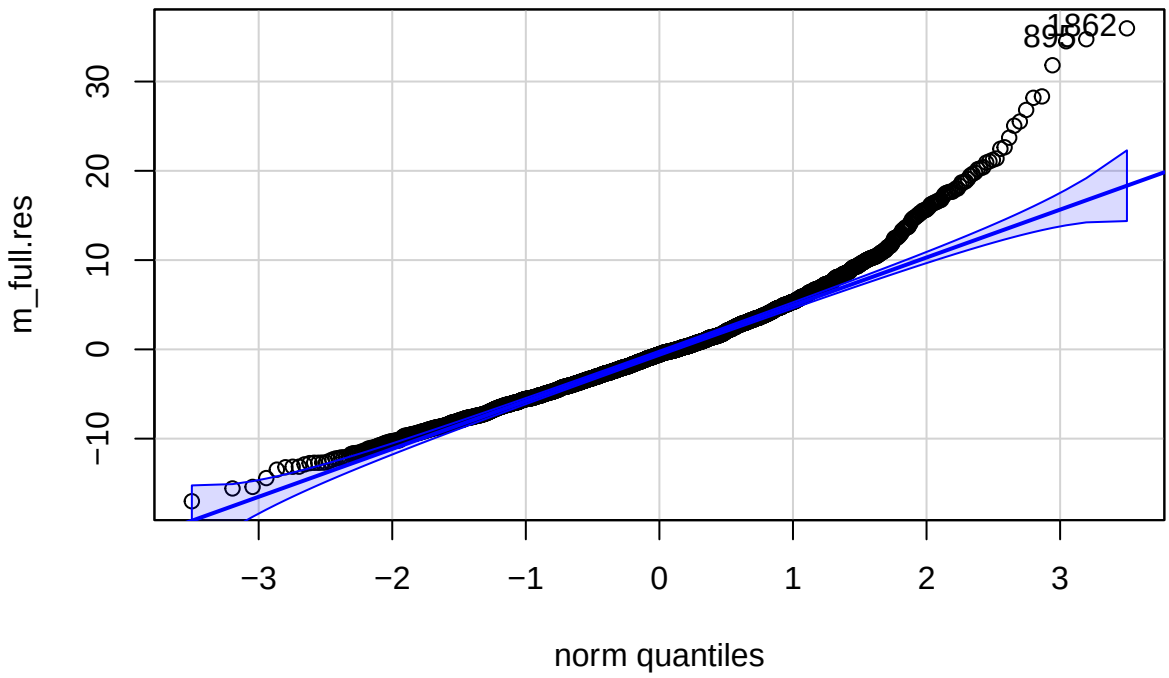
```
Hmisc::describe(m_full.res)$counts[c(".25", ".50", ".75")] #not symmetric
```

```
##       .25       .50       .75
## "-4.0480" "-0.5821" " 3.1841"
```

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



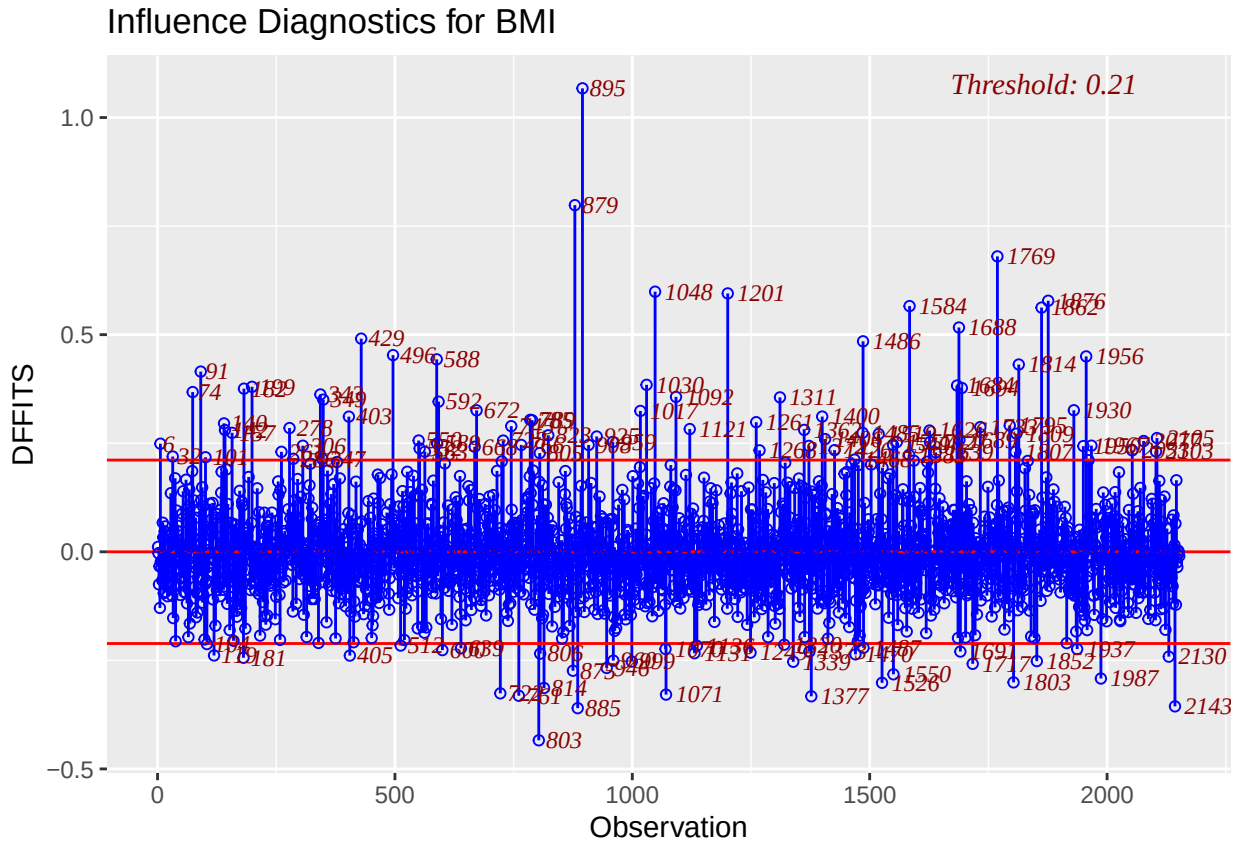

```
# Q-Q plot
qq.m_full.res = car::qqPlot(m_full.res)
```




```
m_full.res[qq.m_full.res]
```

```
##     1862      895
## 35.95902 34.74208
```

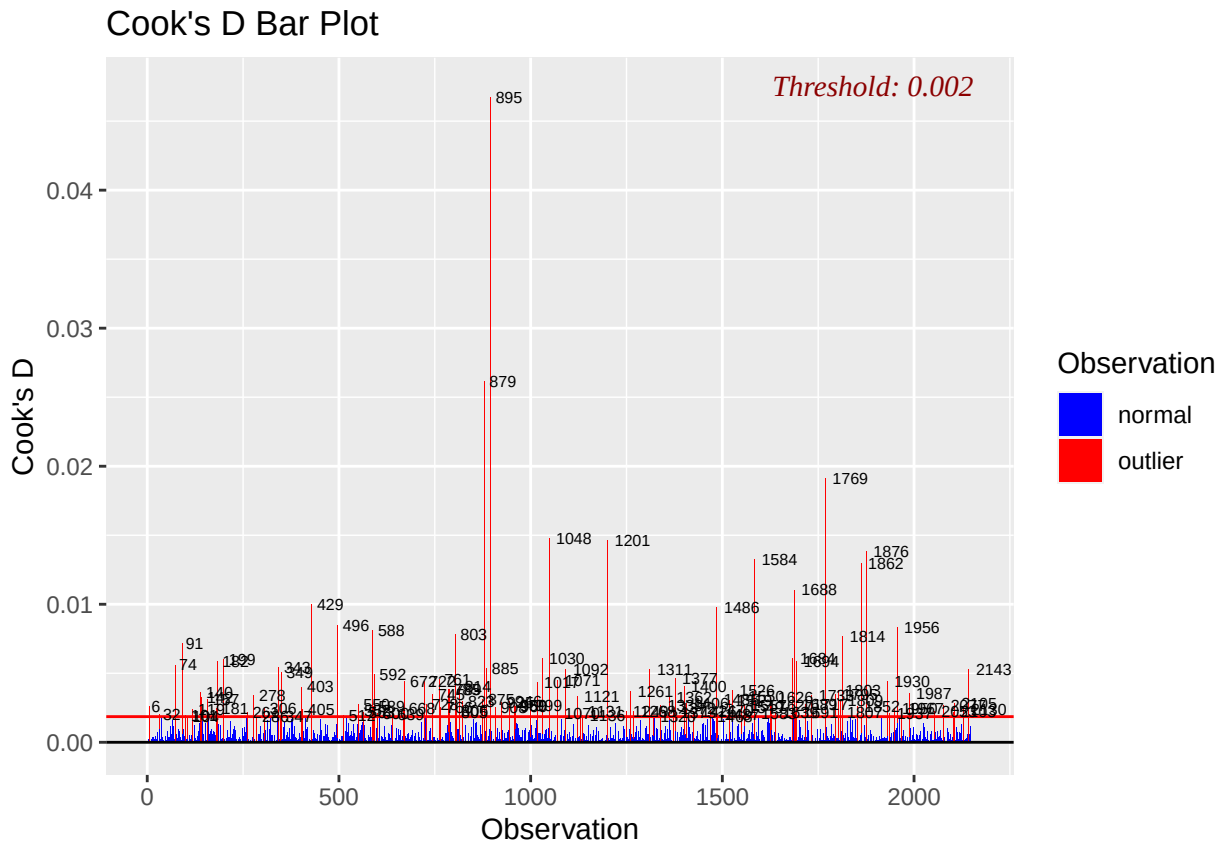
```
################ influential observations  ##################

influence4 = data.frame(
  Residual = resid(m_full),
  Rstudent = rstudent(m_full),
  HatDiagH = hat(model.matrix(m_full)),
  CovRatio = covratio(m_full),
  DFFITS = dffits(m_full),
  COOKsDistance = cooks.distance(m_full)
)
# DFFITS
ols_plot_dffits(m_full)
```




```
influence4[order(abs(influence4$DFFITS), decreasing = T), ] %>% head()
```

```
##      Residual Rstudent   HatDiagH  CovRatio    DFFITS COOKsDistance
## 895  34.74208 5.716954 0.03367642 0.7258276 1.0672504    0.04676306
## 879  34.51006 5.636843 0.01966016 0.7227199 0.7982540    0.02617191
## 1769 28.33273 4.620323 0.02121657 0.8130629 0.6802470    0.01909806
## 1048 28.17758 4.584300 0.01678444 0.8124018 0.5989658    0.01480904
## 1201 25.51917 4.155163 0.02009154 0.8500582 0.5949794    0.01463813
## 1876 18.82482 3.081536 0.03399534 0.9407850 0.5780791    0.01386861
```

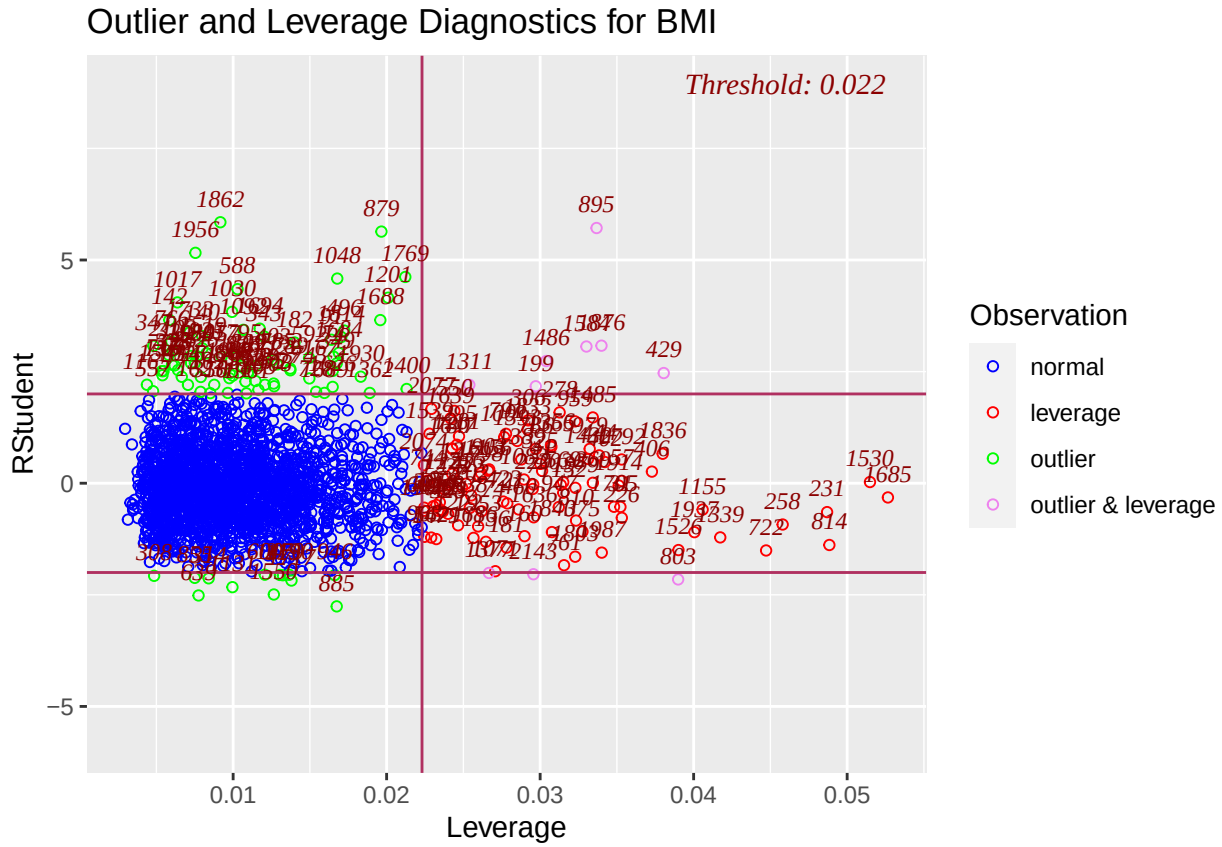
```
#From the plot above, we can see 2 observations with the largest (magnitude) of DFFITS, observation 879

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




```
influence4[order(influence4$COOKsDistance, decreasing = T), ] %>% head()
```

```
##      Residual Rstudent   HatDiagH  CovRatio    DFFITS COOKsDistance
## 895  34.74208 5.716954 0.03367642 0.7258276 1.0672504    0.04676306
## 879  34.51006 5.636843 0.01966016 0.7227199 0.7982540    0.02617191
## 1769 28.33273 4.620323 0.02121657 0.8130629 0.6802470    0.01909806
## 1048 28.17758 4.584300 0.01678444 0.8124018 0.5989658    0.01480904
## 1201 25.51917 4.155163 0.02009154 0.8500582 0.5949794    0.01463813
## 1876 18.82482 3.081536 0.03399534 0.9407850 0.5780791    0.01386861
```

```
#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_full)
```




```
#high leverage
influence4[order(influence4$HatDiagH, decreasing = T), ] %>% head()
```

```
##       Residual   Rstudent   HatDiagH CovRatio      DFFITS COOKsDistance
## 1685 -1.931548 -0.3185817 0.05265617 1.066336 -0.07510895  2.351557e-04
## 1530  0.160039  0.0263793 0.05148766 1.066235  0.00614601  1.574633e-06
## 814  -8.403970 -1.3839242 0.04884165 1.040555 -0.31360361  4.096039e-03
## 231  -3.932033 -0.6472298 0.04869397 1.058099 -0.14643200  8.936745e-04
## 258  -5.629922 -0.9253980 0.04579870 1.049696 -0.20273801  1.712728e-03
## 722  -9.167270 -1.5064867 0.04472129 1.031938 -0.32595482  4.424300e-03
```

```
#high studentized residual
influence4[order(influence4$Rstudent, decreasing = T), ] %>% head()
```

```
##      Residual Rstudent   HatDiagH  CovRatio    DFFITS COOKsDistance
## 1862 35.95902 5.845581 0.00916931 0.6962757 0.5623365    0.01297370
## 895  34.74208 5.716954 0.03367642 0.7258276 1.0672504    0.04676306
## 879  34.51006 5.636843 0.01966016 0.7227199 0.7982540    0.02617191
## 1956 31.82618 5.160475 0.00754466 0.7559609 0.4499394    0.00833484
## 1769 28.33273 4.620323 0.02121657 0.8130629 0.6802470    0.01909806
## 1048 28.17758 4.584300 0.01678444 0.8124018 0.5989658    0.01480904
```

```
#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

rm4.df3 = df3[-c(879, 1769, 1155, 1048, 1769, 1684, 74, 72, 1689, 1311), ]
rm.m_full =  lm(
  BMI ~ SleepHrsNight + Age + Gender + factor(Race1)  + Poverty + TotChol + BPDiaAve + BPSysAve + Alcoho
    DaysPhysHlthBad + factor(HealthGen) + PhysActive + SleepHrsNight*Age + SleepHrsNight*Gender,
  rm4.df3
)
## Before removing these observations, the estimated coefficients are:
summary(m_full)$coef
```

```
##                          Estimate  Std. Error     t value     Pr(>|t|)
## (Intercept)          29.497094434 3.183685294  9.26507858 4.619339e-20
## SleepHrsNight        -0.568047125 0.378049577 -1.50257310 1.330975e-01
## Age                  -0.109636600 0.062676892 -1.74923481 8.039460e-02
## Gender                3.598602589 1.433671409  2.51006093 1.214471e-02
## factor(Race1)2       -1.999694994 0.640340871 -3.12286016 1.815123e-03
## factor(Race1)3       -1.208466988 0.561320206 -2.15290128 3.143852e-02
## factor(Race1)4       -1.490610101 0.420778721 -3.54250352 4.048804e-04
## factor(Race1)5       -3.291828521 0.630345573 -5.22226008 1.939627e-07
## Poverty               0.055441226 0.091718379  0.60447237 5.455941e-01
## TotChol               0.012609973 0.135822020  0.09284189 9.260379e-01
## BPDiaAve              0.058604369 0.013689277  4.28104204 1.942795e-05
## BPSysAve              0.051653618 0.011799435  4.37763487 1.257837e-05
## AlcoholYear          -0.008706857 0.001513401 -5.75317255 1.002440e-08
## Smoke100             -0.860788361 0.287629066 -2.99270298 2.796997e-03
## UrineFlow1           -0.096878138 0.142283308 -0.68088196 4.960203e-01
## DaysMentHlthBad      -0.032320815 0.018002812 -1.79532043 7.274450e-02
## DaysPhysHlthBad       0.014228165 0.020913984  0.68031826 4.963770e-01
## factor(HealthGen)2   -2.349774556 1.001809675 -2.34552991 1.909144e-02
## factor(HealthGen)3   -4.101774388 0.991935310 -4.13512287 3.685381e-05
## factor(HealthGen)4   -5.805697020 1.018118341 -5.70237937 1.346102e-08
## factor(HealthGen)5   -7.663418887 1.075784904 -7.12356053 1.433222e-12
## PhysActive           -0.864153767 0.294684715 -2.93246891 3.398698e-03
## SleepHrsNight:Age     0.017361985 0.009010030  1.92696197 5.411737e-02
## SleepHrsNight:Gender -0.459042673 0.206678207 -2.22105019 2.645246e-02
```

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

```
##                          Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)          29.154063884 3.125682556  9.3272632 2.643525e-20
## SleepHrsNight        -0.530063265 0.370392635 -1.4310848 1.525534e-01
## Age                  -0.106814807 0.061392354 -1.7398715 8.202677e-02
## Gender                3.553854215 1.408335178  2.5234435 1.169376e-02
## factor(Race1)2       -1.471032257 0.628540206 -2.3403948 1.935570e-02
## factor(Race1)3       -0.680383502 0.551741419 -1.2331565 2.176542e-01
## factor(Race1)4       -0.975908447 0.414951897 -2.3518592 1.877092e-02
## factor(Race1)5       -2.799397862 0.618428642 -4.5266304 6.325344e-06
## Poverty               0.056473409 0.089773771  0.6290636 5.293752e-01
## TotChol               0.029286646 0.135108400  0.2167641 8.284131e-01
## BPDiaAve              0.055227703 0.013412130  4.1177430 3.972786e-05
## BPSysAve              0.052812605 0.011625539  4.5428092 5.862951e-06
## AlcoholYear          -0.009553623 0.001489363 -6.4145702 1.736495e-10
```

```
## Smoke100             -0.920358696 0.281762217 -3.2664376 1.106478e-03
## UrineFlow1           -0.111498883 0.139204979 -0.8009691 4.232393e-01
## DaysMentHlthBad      -0.027385931 0.017678794 -1.5490837 1.215109e-01
## DaysPhysHlthBad      -0.007174377 0.020730277 -0.3460821 7.293154e-01
## factor(HealthGen)2   -3.009982100 0.983247913 -3.0612647 2.231712e-03
## factor(HealthGen)3   -4.455350481 0.971245827 -4.5872532 4.753425e-06
## factor(HealthGen)4   -6.195074469 0.997238886 -6.2122271 6.269985e-10
## factor(HealthGen)5   -8.043481753 1.053565770 -7.6345322 3.405097e-14
## PhysActive           -0.783362929 0.288495498 -2.7153385 6.674536e-03
## SleepHrsNight:Age     0.017507527 0.008824740  1.9839141 4.739426e-02
## SleepHrsNight:Gender -0.471171487 0.202876643 -2.3224531 2.030285e-02
```

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

```
##          (Intercept)        SleepHrsNight                  Age
##            1.1629300            6.6867444            2.5737691
##               Gender       factor(Race1)2       factor(Race1)3
##            1.2434931           26.4371686           43.6986274
##       factor(Race1)4       factor(Race1)5              Poverty
##           34.5295966           14.9591832            1.8617607
##              TotChol             BPDiaAve             BPSysAve
##          132.2498619            5.7618005            2.2437682
##          AlcoholYear              Smoke100           UrineFlow1
##            9.7252723            6.9204392           15.0918930
##      DaysMentHlthBad      DaysPhysHlthBad   factor(HealthGen)2
##           15.2684392          150.4237691           28.0966335
##   factor(HealthGen)3   factor(HealthGen)4   factor(HealthGen)5
##            8.6200766            6.7068166            4.9594427
##           PhysActive    SleepHrsNight:Age SleepHrsNight:Gender
##            9.3491276            0.8382783            2.6421975
```

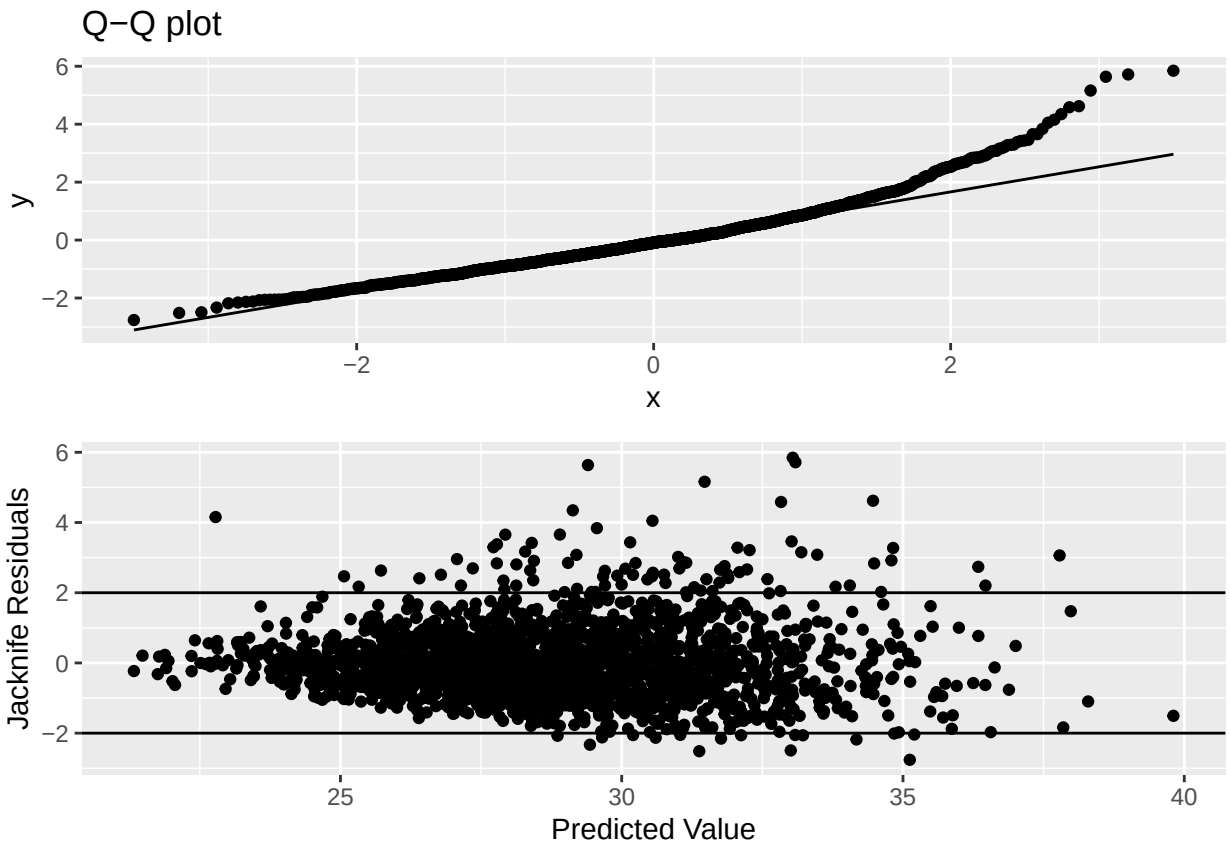
```
#The estimated regression coefficients doesn't change slightly after removing these observations. 5 of

################## multicollinearity ######################
#Pearson correlations
var4 = c(
  "BMI",
  "SleepHrsNight",
  "Age",
  "Gender",
  "Race1",
  "TotChol",
  "BPDiaAve",
  "BPSysAve",
  "AlcoholYear",
  "Smoke100",
  "DaysPhysHlthBad",
  "PhysActive",
  "Poverty",
  "UrineFlow1",
  "DaysMentHlthBad",
  "HealthGen"
)
newData4 = df3[, var4]
```

```r
library("corrplot")
par(mfrow = c(1, 2))
cormat4 = cor(as.matrix(newData4[, -c(1)], method = "pearson"))
p.mat4 = cor.mtest(as.matrix(newData4[, -c(1)]))$p
corrplot(
  cormat4,
  method = "color",
  type = "upper",
  number.cex = 1,
  diag = FALSE,
  addCoef.col = "black",
  tl.col = "black",
  tl.srt = 90,
  p.mat = p.mat4,
  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_full)
```

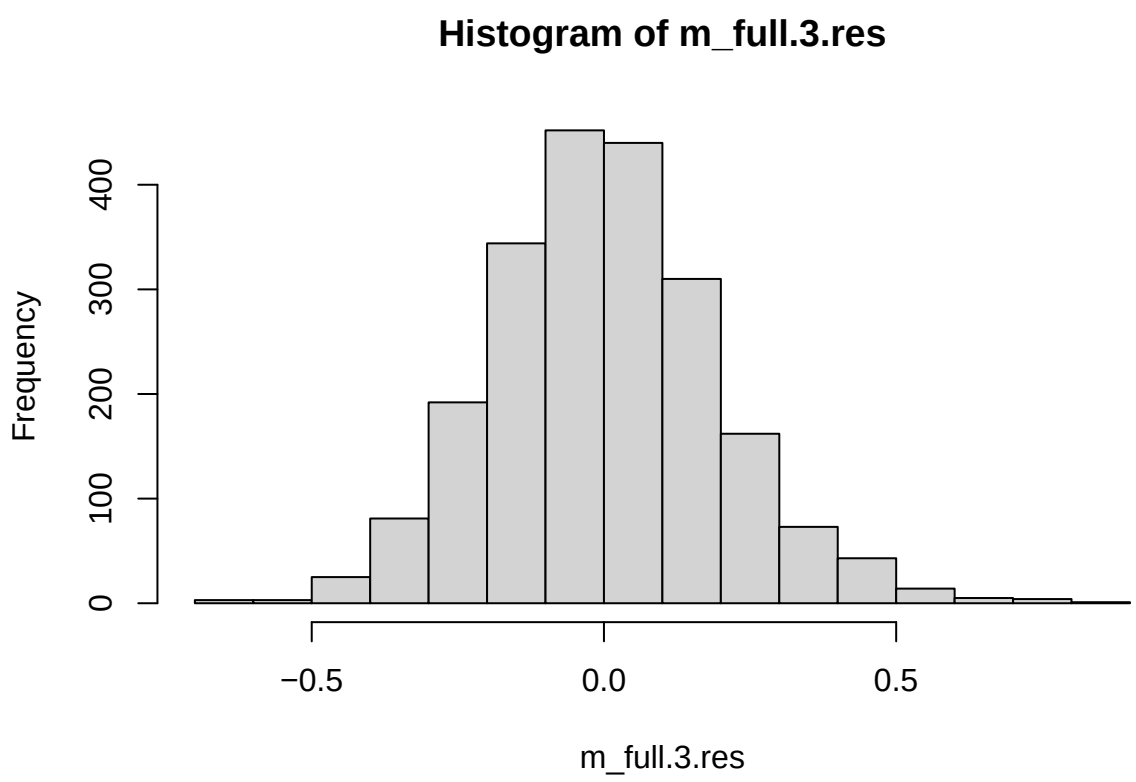
```
##                            GVIF Df GVIF^(1/(2*Df))
## SleepHrsNight         13.604129  1        3.688378
## Age                   27.980786  1        5.289687
## Gender                28.406350  1        5.329761
## factor(Race1)          1.244918  4        1.027762
## Poverty                1.334579  1        1.155240
## TotChol                1.131047  1        1.063507
## BPDiaAve               1.457561  1        1.207295
## BPSysAve               1.574796  1        1.254909
## AlcoholYear            1.127701  1        1.061933
## Smoke100               1.141358  1        1.068344
## UrineFlow1             1.048362  1        1.023895
## DaysMentHlthBad        1.156637  1        1.075470
## DaysPhysHlthBad        1.253155  1        1.119444
## factor(HealthGen)      1.474279  4        1.049718
## PhysActive             1.174467  1        1.083728
## SleepHrsNight:Age     37.587100  1        6.130832
## SleepHrsNight:Gender  30.007381  1        5.477899
```

```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_full.log = lm(
  logBMI ~ SleepHrsNight + Age + Gender + factor(Race1)  + Poverty + TotChol + BPDiaAve + BPSysAve + Al
    SleepHrsNight*Age+SleepHrsNight*Gender,
  df3
)
p41.log = ols_plot_resid_lev(m_full.log)
p42.log = ols_plot_cooksd_bar(m_full.log)
```

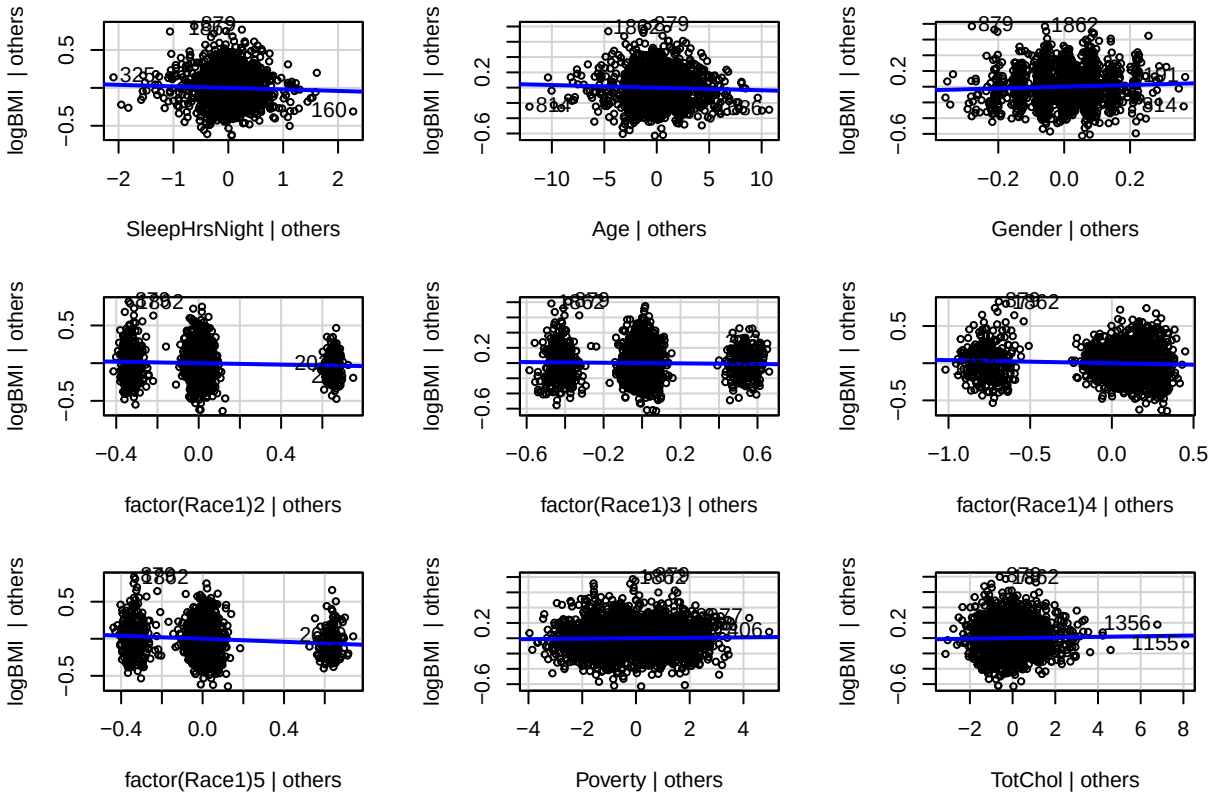
```
library(gridExtra)
p43.log = ggplot(m_full.log, aes(sample = rstudent(m_full.log))) + geom_qq() + stat_qq_line() +
  labs(title = "Q-Q plot")
p44.log = ggplot() + geom_point(aes(y = rstudent(m_full.log), x = m_full.log$fitted.values)) + labs(x =
  geom_hline(yintercept = c(-2, 2))
grid.arrange(p43.log, p44.log, nrow = 2)


p43 = ggplot(m_full, aes(sample = rstudent(m_full))) + geom_qq() + stat_qq_line() +
  labs(title = "Q-Q plot")
p44 = ggplot() + geom_point(aes(y = rstudent(m_full), x = m_full$fitted.values)) + labs(x = "Predicted
  geom_hline(yintercept = c(-2, 2))
grid.arrange(p43, p44, nrow = 2)


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

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

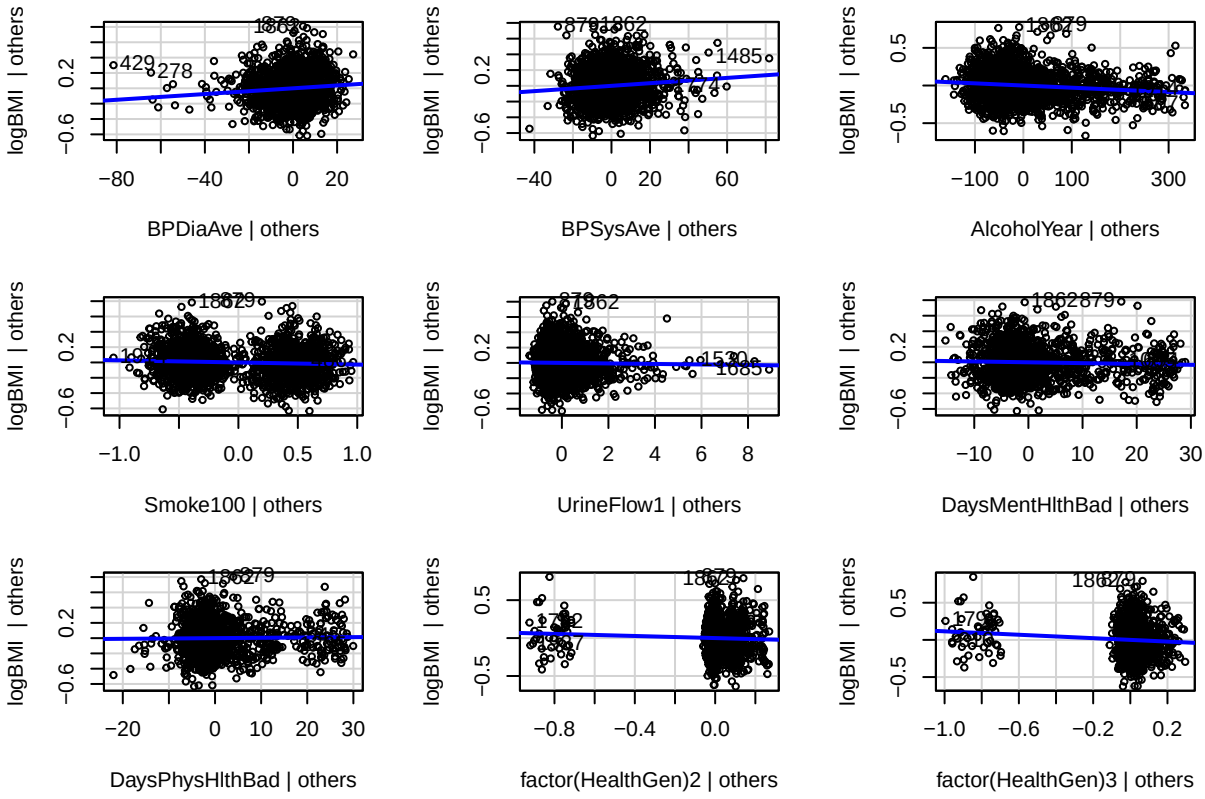


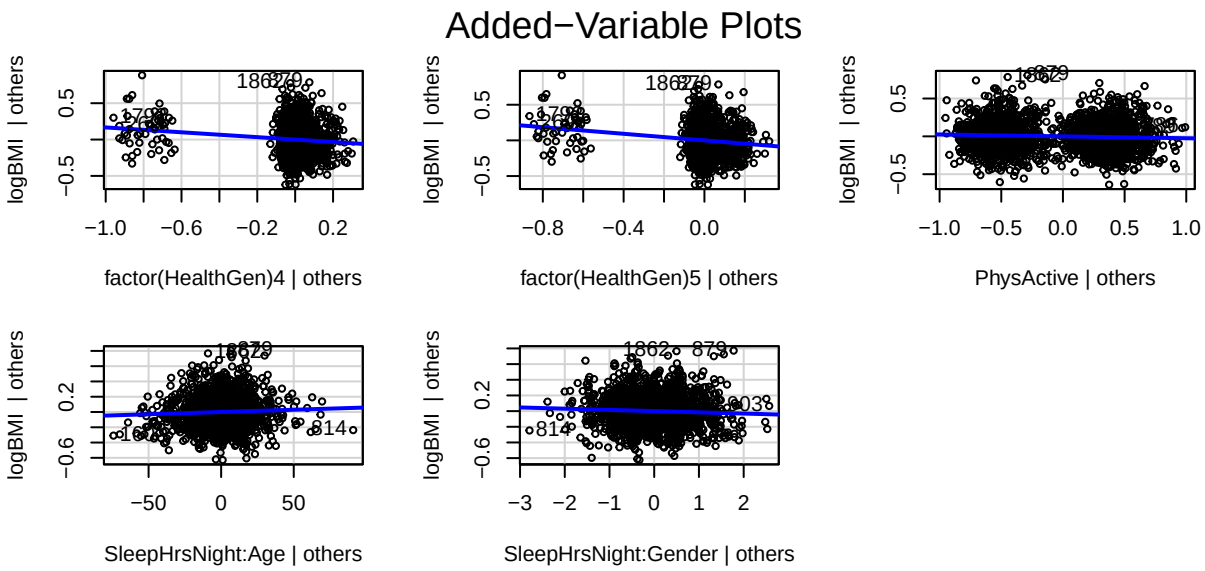

```
hist(m_full.3.res, breaks = 15)
```




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


logBMI | others
SleepHrsNight | others
Age | others
Gender | others
factor(Race1)2 | others
factor(Race1)3 | others
factor(Race1)4 | others
factor(Race1)5 | others
Poverty | others
TotChol | others


logBMI | others
BPDiaAve | others
BPSysAve | others
AlcoholYear | others
Smoke100 | others
UrineFlow1 | others
DaysMentHlthBad | others
DaysPhysHlthBad | others
factor(HealthGen)2 | others
factor(HealthGen)3 | others

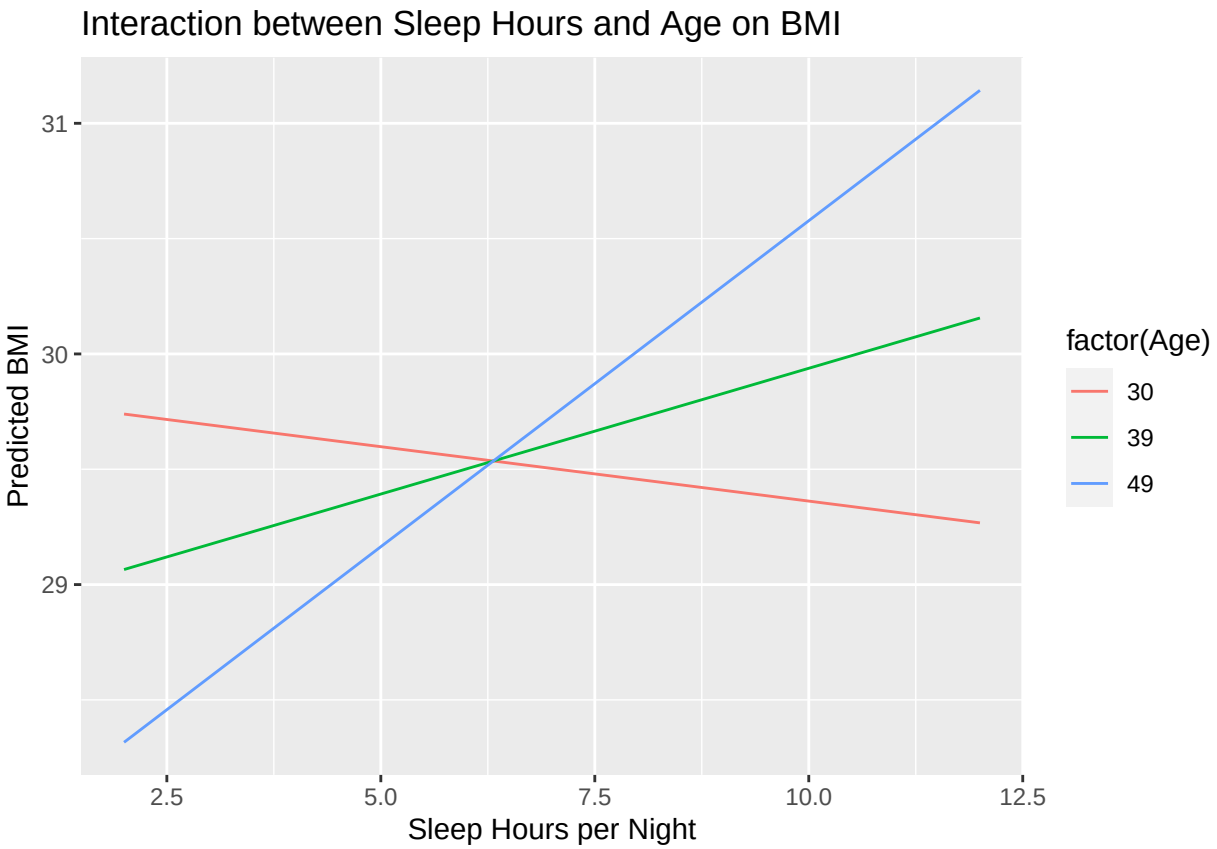



```
#After looking at residuals from models using the log-transformed (natural log scale) BMI adjusted for

#collinearity diagnostics

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

```
##                            GVIF Df GVIF^(1/(2*Df))
## SleepHrsNight         13.604129  1        3.688378
## Age                   27.980786  1        5.289687
## Gender                28.406350  1        5.329761
## factor(Race1)          1.244918  4        1.027762
## Poverty                1.334579  1        1.155240
## TotChol                1.131047  1        1.063507
## BPDiaAve               1.457561  1        1.207295
## BPSysAve               1.574796  1        1.254909
## AlcoholYear            1.127701  1        1.061933
## Smoke100               1.141358  1        1.068344
## UrineFlow1             1.048362  1        1.023895
## DaysMentHlthBad        1.156637  1        1.075470
## DaysPhysHlthBad        1.253155  1        1.119444
## factor(HealthGen)      1.474279  4        1.049718
## PhysActive             1.174467  1        1.083728
## SleepHrsNight:Age     37.587100  1        6.130832
## SleepHrsNight:Gender  30.007381  1        5.477899
```

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

```r
getMode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}

new_data <- expand.grid(SleepHrsNight = seq(min(df3$SleepHrsNight), max(df3$SleepHrsNight), length.out =
                        Age = quantile(df3$Age, probs = c(0.25, 0.5, 0.75)),
                        Gender = median(df3$Gender, na.rm = TRUE),
                        Race1 = median(df3$Race1, na.rm = TRUE),
                        Poverty = median(df3$Poverty, na.rm = TRUE),
                        TotChol = median(df3$TotChol, na.rm = TRUE),
                        BPDiaAve = median(df3$BPDiaAve, na.rm = TRUE),
                        BPSysAve = median(df3$BPSysAve, na.rm = TRUE),
                        AlcoholYear = median(df3$AlcoholYear, na.rm = TRUE),
                        Smoke100 = getMode(df3$Smoke100),
                        UrineFlow1 = median(df3$UrineFlow1, na.rm = TRUE),
                        DaysMentHlthBad = median(df3$DaysMentHlthBad, na.rm = TRUE),
                        DaysPhysHlthBad = median(df3$DaysPhysHlthBad, na.rm = TRUE),
                        HealthGen = getMode(df3$HealthGen),
                        PhysActive = getMode(df3$PhysActive)
)

# predict
new_data$predicted_BMI <- predict(m_full, newdata = new_data)
# interaction
library(ggplot2)
ggplot(new_data, aes(x = SleepHrsNight, y = predicted_BMI, group = factor(Age))) +
  geom_line(aes(color = factor(Age))) +
  labs(title = "Interaction between Sleep Hours and Age on BMI",
       x = "Sleep Hours per Night",
       y = "Predicted BMI")
```




```
# cross validation
library(caret)
splitIndex <-
  createDataPartition(df3$SleepHrsNight, p = 0.7, list = FALSE)
trainData <- df3[splitIndex, ]
testData <- df3[-splitIndex, ]
predictions <- predict(m_full, newdata = testData)
mse <- mean((testData$SleepHrsNight - predictions) ^ 2)
control <-
  trainControl(method = "cv", number = 10)  # 10-fold cross-validation
cv_model <-
  train(
    SleepHrsNight ~ .,
    data = df3,
    method = "lm",
    trControl = control
  )
cv_model
```

```
## Linear Regression
##
## 2152 samples
##   21 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
```

```
## Summary of sample sizes: 1937, 1938, 1936, 1937, 1937, 1937, ...
## Resampling results:
##
##   RMSE      Rsquared    MAE
##   1.280489  0.04937206  0.9968378
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
(cv_results <- cv_model$results)
```

```
##   intercept     RMSE   Rsquared       MAE    RMSESD RsquaredSD      MAESD
## 1      TRUE 1.280489 0.04937206 0.9968378 0.0466957 0.02498676 0.03037066
```