model4

Liancheng

2023-11-25

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

```
##          used (Mb) gc trigger (Mb) max used (Mb)
## Ncells 471666 25.2    1017187 54.4   660860 35.3
## Vcells 892801  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"
```

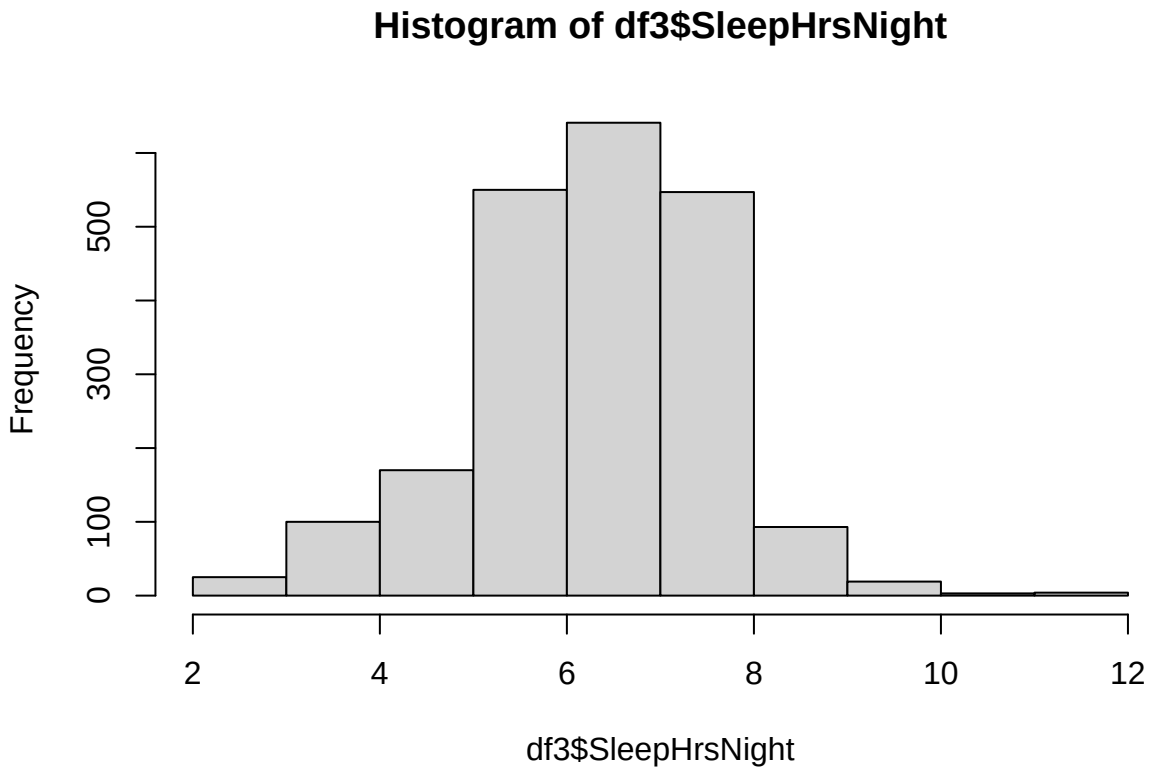
```
# df$BPSysAve
library(dplyr)

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 ##
df3$logBMI = log(df3$BMI + 1)
m_full = lm(
  logBMI ~ SleepHrsNight + Age + Gender + factor(Race1)  + Poverty + TotChol + BPDiaAve + BPSysAve + Al
    DaysPhysHlthBad + factor(HealthGen) + PhysActive + SleepHrsNight*Age + SleepHrsNight*Gender,
  df3
)
summary(m_full)
```

```
##
## Call:
## lm(formula = logBMI ~ 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
## -0.62950 -0.12824 -0.00334  0.12019  0.80003
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           3.354e+00  9.989e-02  33.573  < 2e-16 ***
## SleepHrsNight        -2.008e-02  1.186e-02  -1.693 0.090664 .
## Age                  -3.452e-03  1.967e-03  -1.755 0.079362 .
## Gender                1.099e-01  4.498e-02   2.443 0.014639 *
## factor(Race1)2       -4.911e-02  2.009e-02  -2.444 0.014603 *
## factor(Race1)3       -2.132e-02  1.761e-02  -1.210 0.226302
```

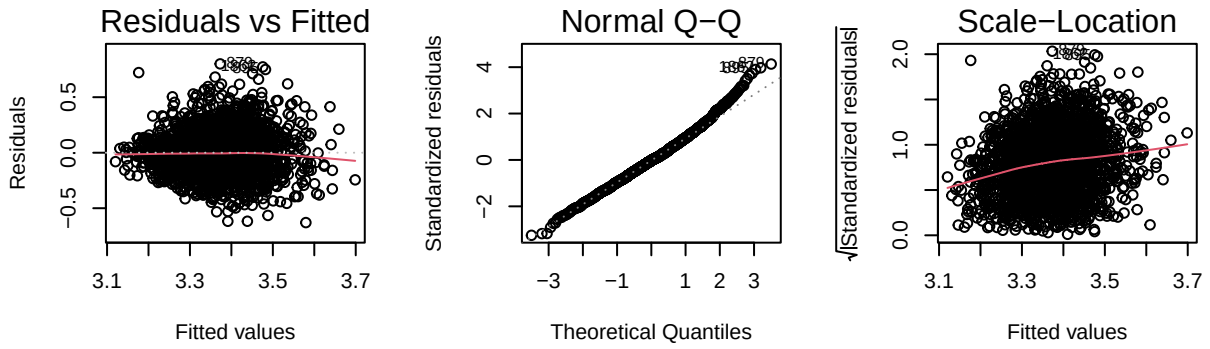
```
## factor(Race1)4       -4.146e-02  1.320e-02  -3.140 0.001710 **
## factor(Race1)5       -1.023e-01  1.978e-02  -5.175 2.50e-07 ***
## Poverty               2.889e-03  2.878e-03   1.004 0.315630
## TotChol               4.030e-03  4.262e-03   0.946 0.344403
## BPDiaAve              1.868e-03  4.295e-04   4.350 1.43e-05 ***
## BPSysAve              1.684e-03  3.702e-04   4.548 5.71e-06 ***
## AlcoholYear          -2.890e-04  4.749e-05  -6.087 1.36e-09 ***
## Smoke100             -2.766e-02  9.025e-03  -3.065 0.002203 **
## UrineFlow1           -3.458e-03  4.464e-03  -0.775 0.438618
## DaysMentHlthBad      -1.077e-03  5.649e-04  -1.907 0.056592 .
## DaysPhysHlthBad       4.421e-04  6.562e-04   0.674 0.500608
## factor(HealthGen)2   -6.900e-02  3.143e-02  -2.195 0.028263 *
## factor(HealthGen)3   -1.151e-01  3.112e-02  -3.698 0.000223 ***
## factor(HealthGen)4   -1.670e-01  3.195e-02  -5.227 1.89e-07 ***
## factor(HealthGen)5   -2.316e-01  3.375e-02  -6.860 9.00e-12 ***
## PhysActive           -2.500e-02  9.246e-03  -2.704 0.006907 **
## SleepHrsNight:Age     5.891e-04  2.827e-04   2.084 0.037310 *
## SleepHrsNight:Gender -1.554e-02  6.485e-03  -2.397 0.016623 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1954 on 2128 degrees of freedom
## Multiple R-squared:  0.1611, Adjusted R-squared:  0.152
## F-statistic: 17.76 on 23 and 2128 DF,  p-value: < 2.2e-16
```

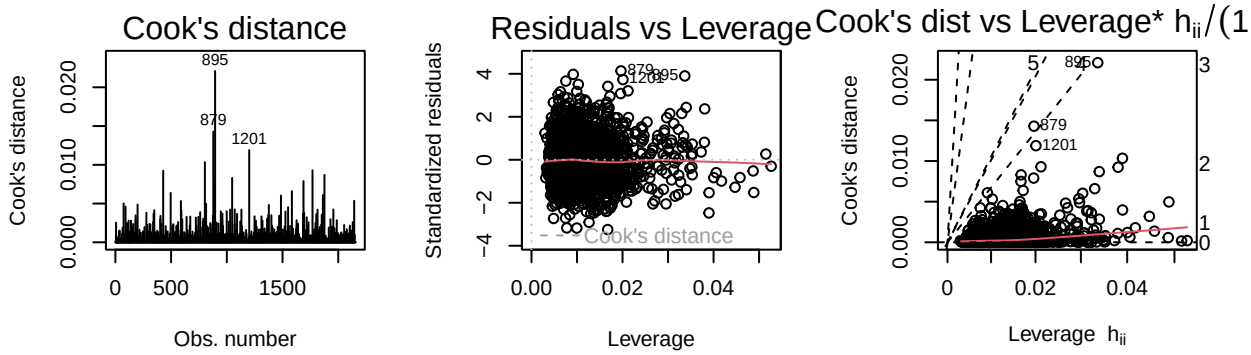
```
car::Anova(m_full, type = "III")
```

```
## Anova Table (Type III tests)
##
## Response: logBMI
##                       Sum Sq   Df   F value    Pr(>F)
## (Intercept)           43.042    1 1127.1743 < 2.2e-16 ***
## SleepHrsNight          0.109    1    2.8651  0.090664 .
## Age                    0.118    1    3.0809  0.079362 .
## Gender                 0.228    1    5.9692  0.014639 *
## factor(Race1)          1.105    4    7.2344 8.762e-06 ***
## Poverty                0.038    1    1.0074  0.315630
## TotChol                0.034    1    0.8944  0.344403
## BPDiaAve               0.722    1   18.9202 1.427e-05 ***
## BPSysAve               0.790    1   20.6887 5.707e-06 ***
## AlcoholYear            1.415    1   37.0507 1.362e-09 ***
## Smoke100               0.359    1    9.3955  0.002203 **
## UrineFlow1             0.023    1    0.6001  0.438618
## DaysMentHlthBad        0.139    1    3.6385  0.056592 .
## DaysPhysHlthBad        0.017    1    0.4538  0.500608
## factor(HealthGen)      4.271    4   27.9621 < 2.2e-16 ***
## PhysActive             0.279    1    7.3112  0.006907 **
## SleepHrsNight:Age      0.166    1    4.3416  0.037310 *
## SleepHrsNight:Gender   0.219    1    5.7449  0.016623 *
## Residuals             81.259 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
```

```r
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)
```





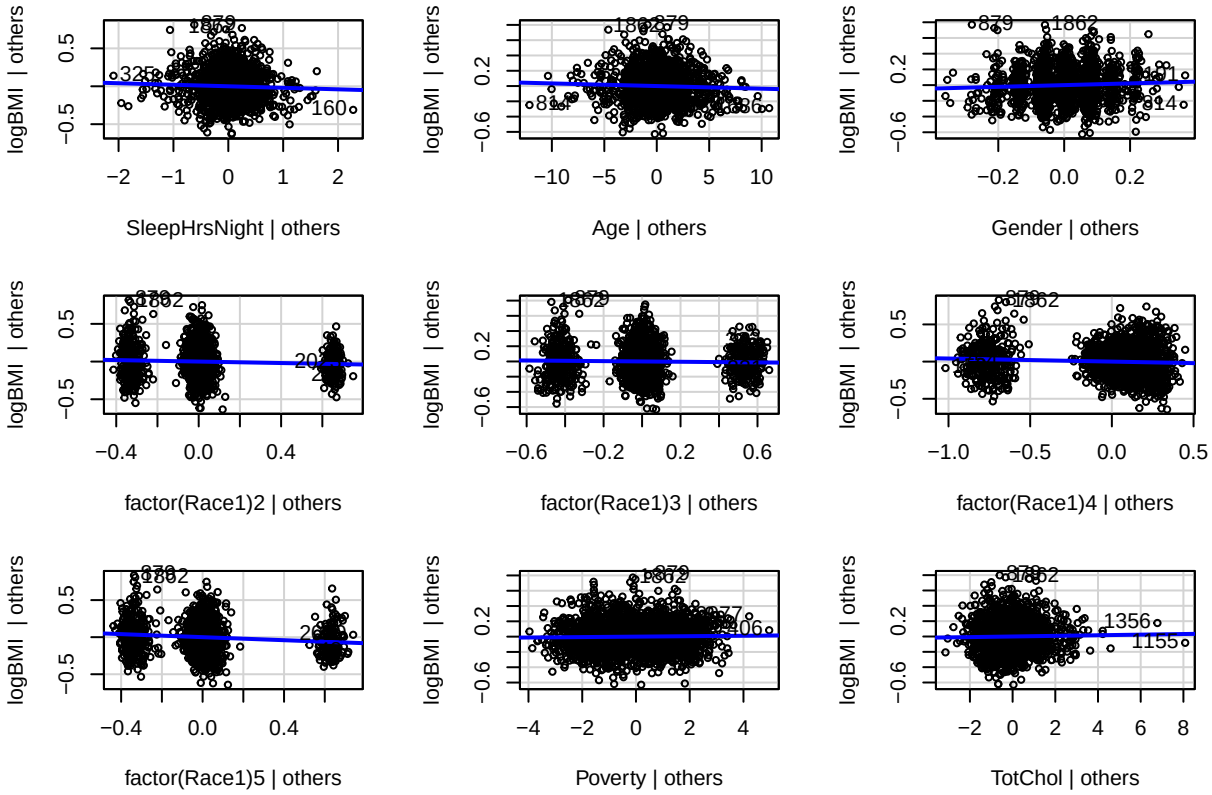

```r
par(mfrow = c(1, 1)) # reset

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)
```

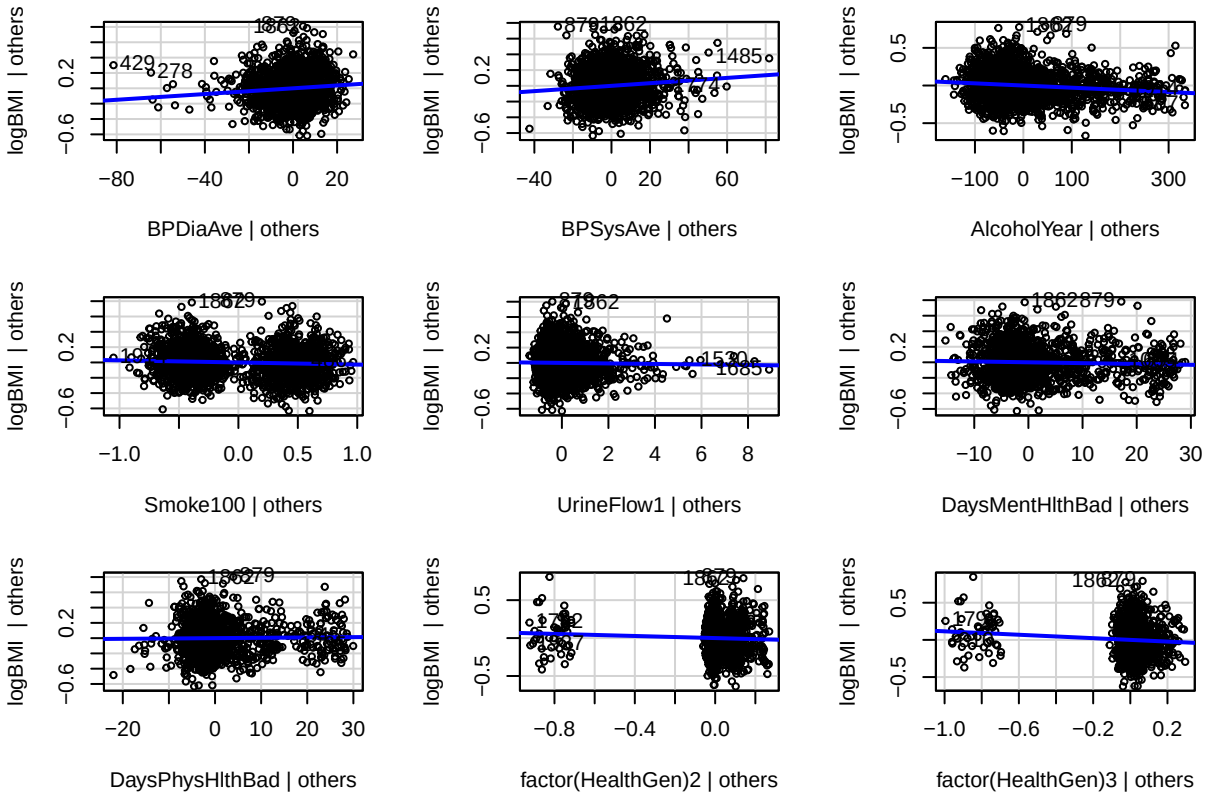
```
## m_full.h
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2152        0     2152        1  0.01115 0.006051 0.004809 0.005406
##      .25      .50      .75      .90      .95
## 0.007016 0.009823 0.013468 0.017982 0.022203
##
## lowest : 0.002961683 0.003144042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

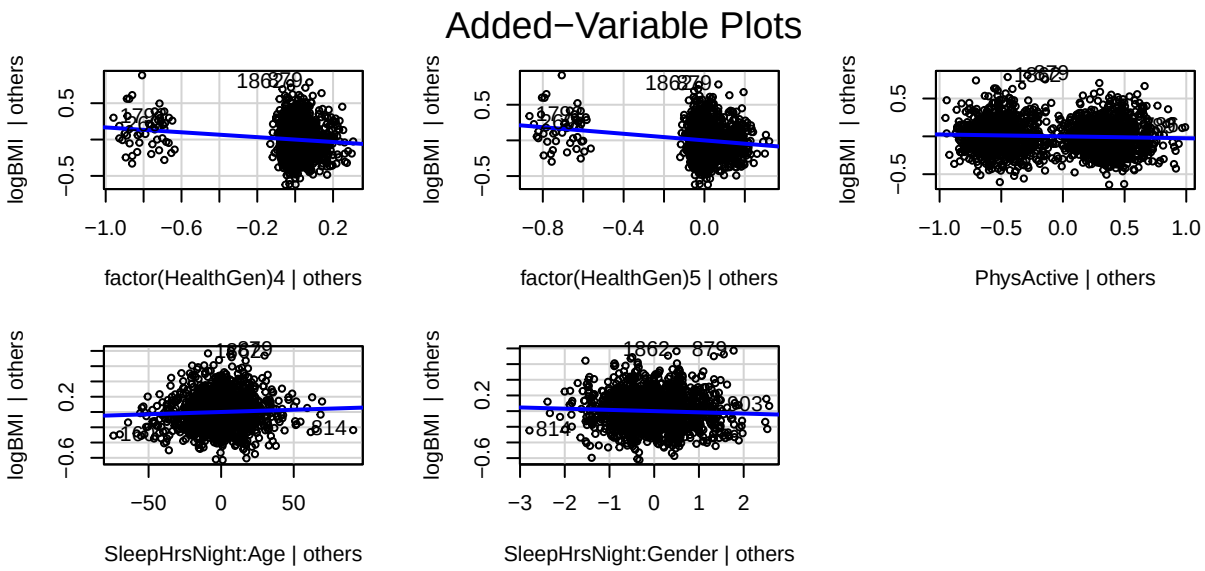
```
m_full.h[which.max(m_full.h)]
```

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

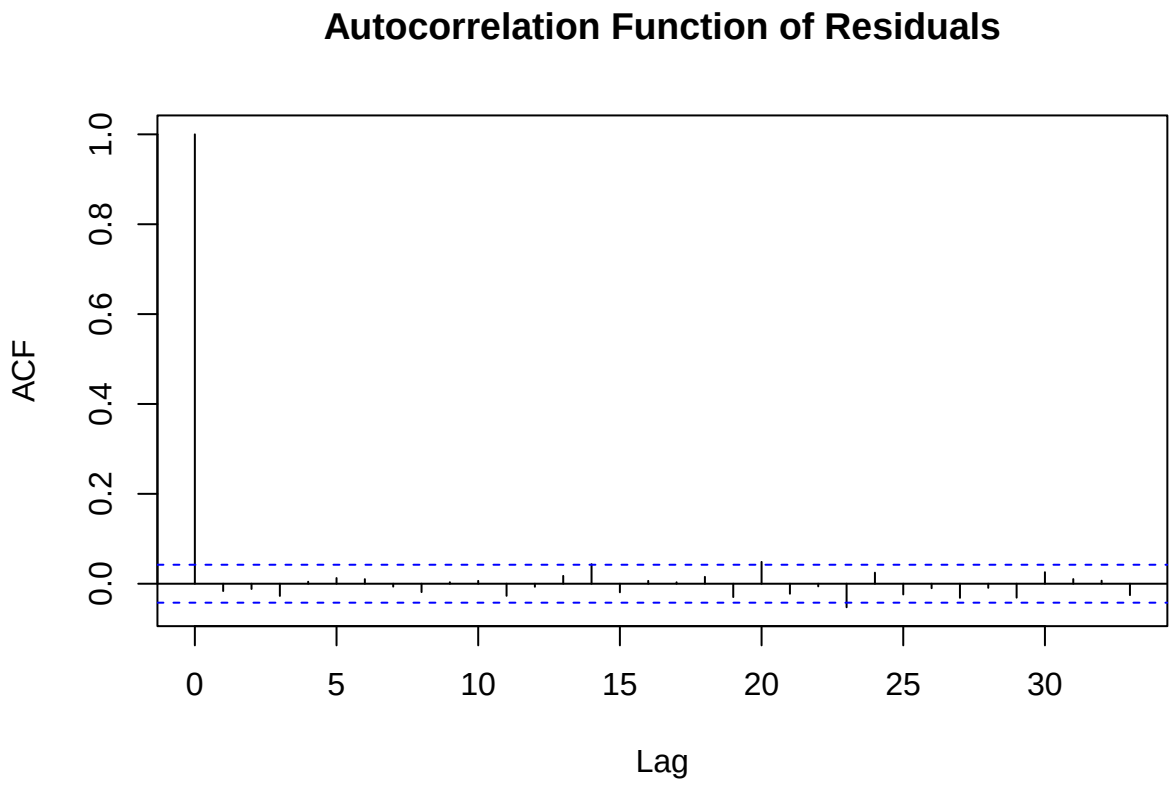
```
###################### Assumption:LINE ##############################

#(1)Linear: 2 approaches

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



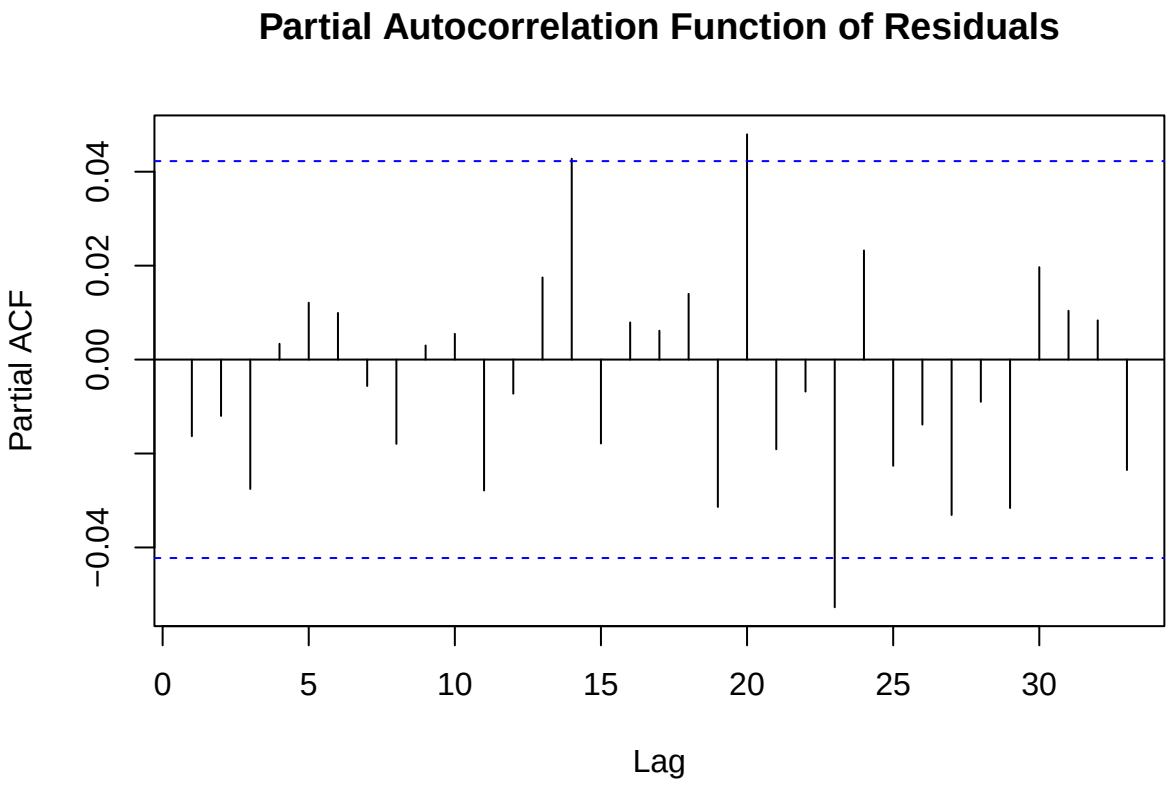

logBMI | others
BPDiaAve | others
BPSysAve | others
AlcoholYear | others
Smoke100 | others
UrineFlow1 | others
DaysMentHlthBad | others
DaysPhysHlthBad | others
factor(HealthGen)2 | others
factor(HealthGen)3 | others




```
#(2)Independence:

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



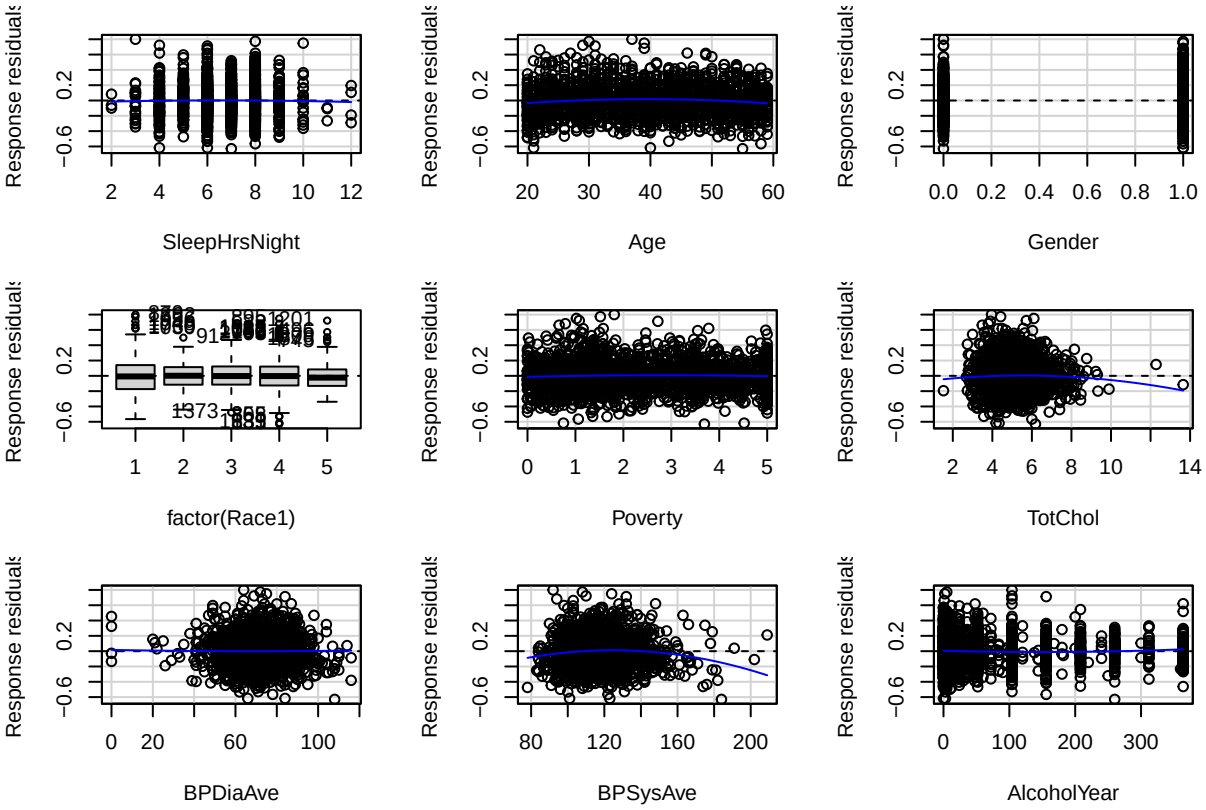

```
pacf(residuals, main = "Partial Autocorrelation Function of Residuals")
```

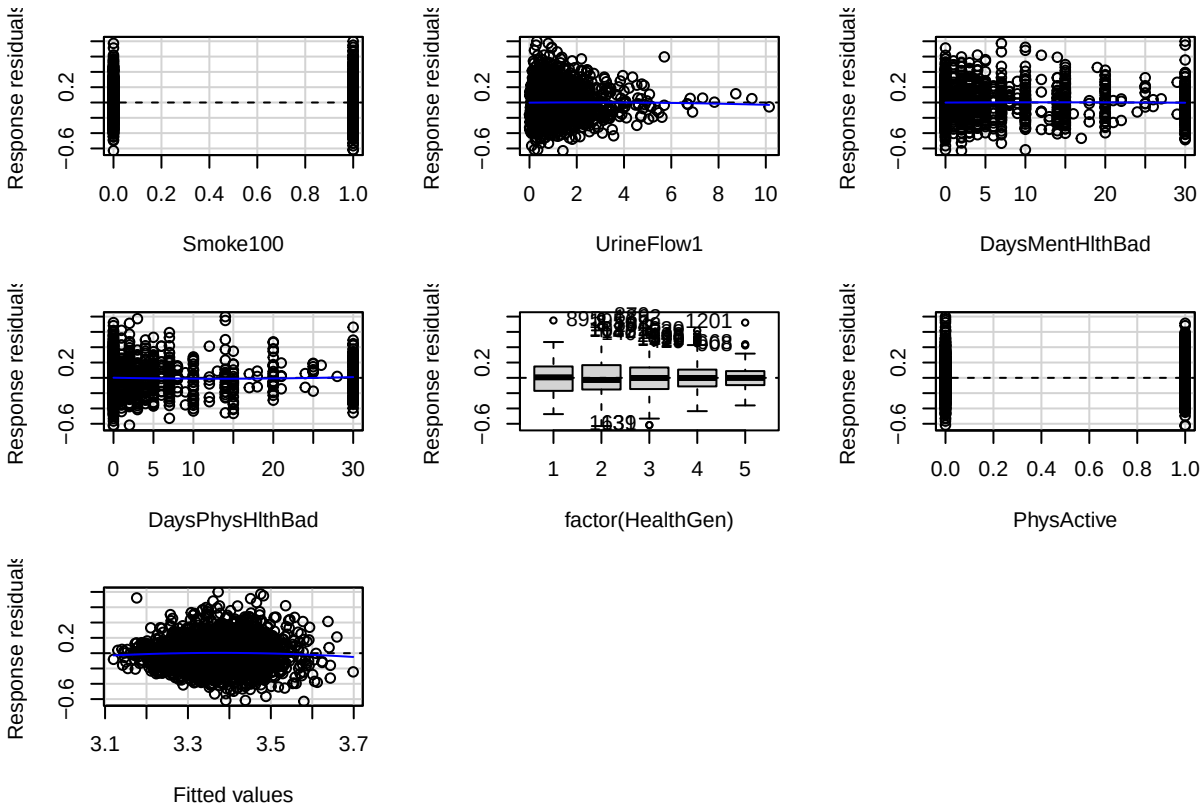



```
dw_test <- dwtest(m_full)
print(dw_test)
```

```
##
##  Durbin-Watson test
##
## data:  m_full
## DW = 2.0326, p-value = 0.7748
## alternative hypothesis: true autocorrelation is greater than 0
```

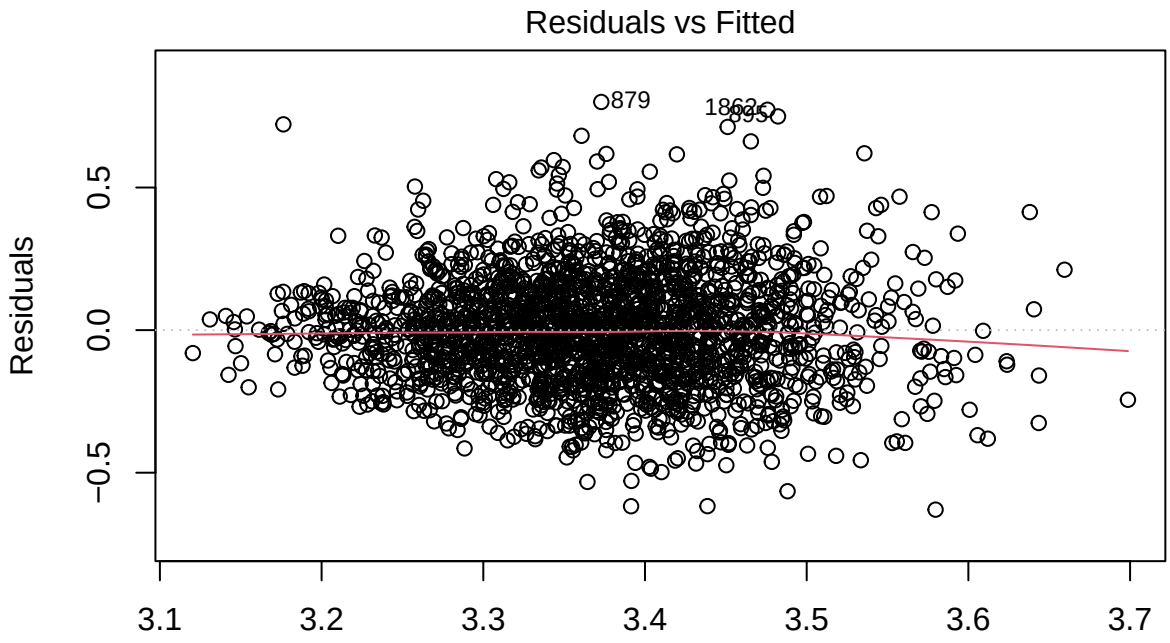
```
#(3)E: constant var: residuals-fitted values; transform for variance-stable...(total: 4 solutions)

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


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

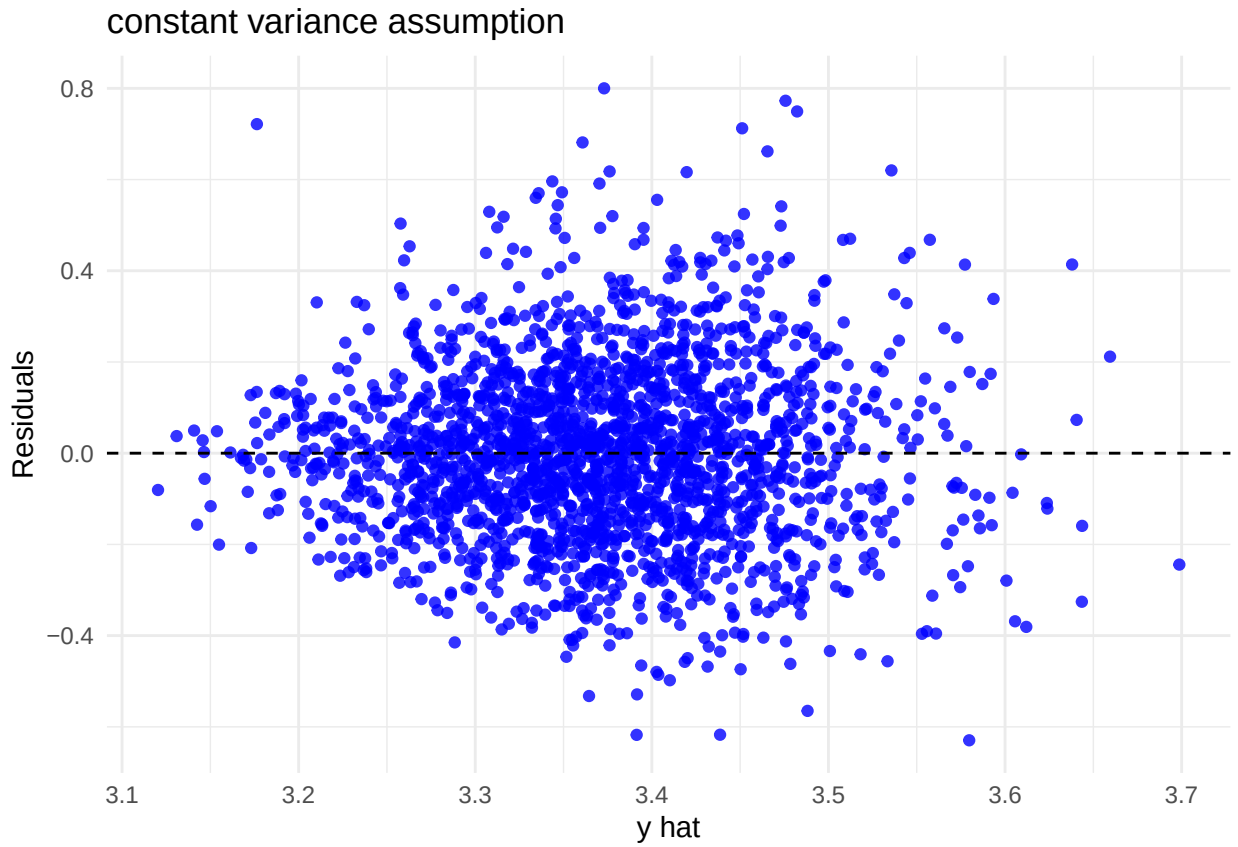



```
##                    Test stat Pr(>|Test stat|)
## SleepHrsNight        -0.5006          0.61667
## Age                  -3.9577        7.816e-05 ***
## Gender                0.2788          0.78041
## factor(Race1)
## Poverty              -1.2289          0.21924
## TotChol              -1.7024          0.08883 .
## BPDiaAve              0.2919          0.77039
## BPSysAve             -4.5027        7.072e-06 ***
## AlcoholYear           1.8682          0.06187 .
## Smoke100             -0.2153          0.82956
## UrineFlow1           -0.3487          0.72734
## DaysMentHlthBad      -0.3170          0.75124
## DaysPhysHlthBad       1.0275          0.30431
## factor(HealthGen)
## PhysActive           -0.6370          0.52419
## Tukey test           -1.5375          0.12418
## ---
## 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()
```



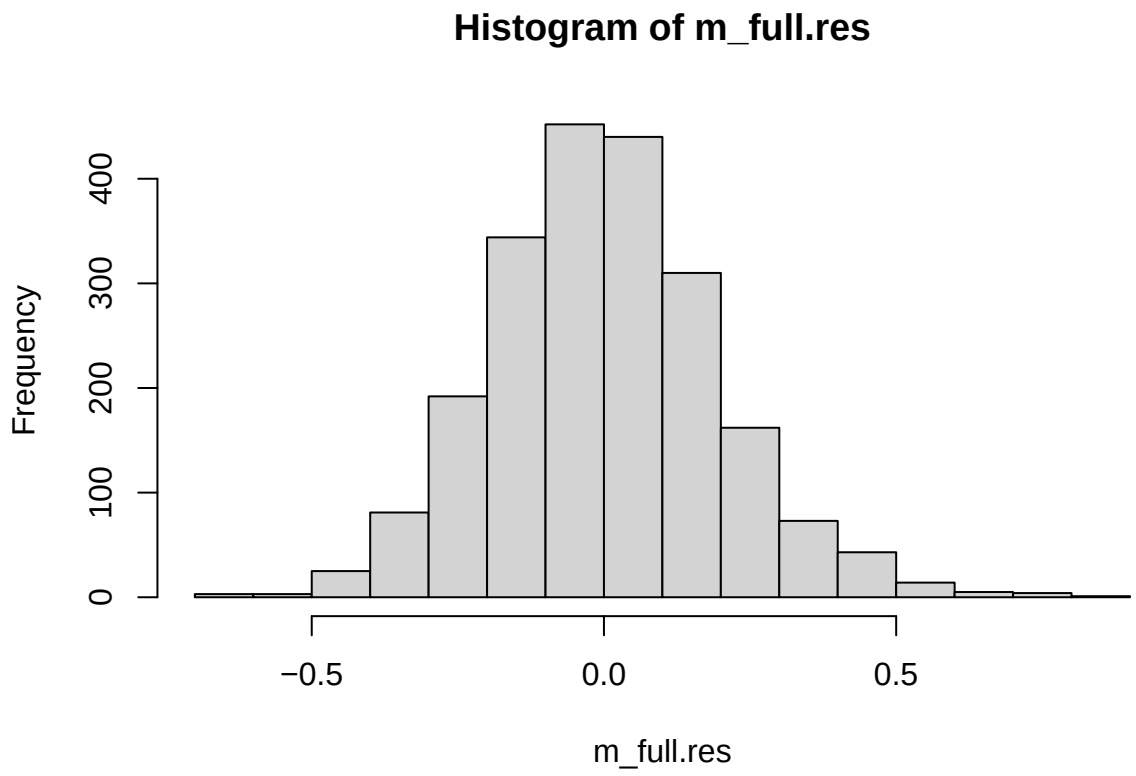

```
#conclusion: the constant variance assumption is basically not violated. The spread of the residuals ap


#(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 6.617e-19    0.2166 -0.304335 -0.242536
##       .25       .50       .75       .90       .95
## -0.128236 -0.003338  0.120187  0.242476  0.326940
##
## lowest : -0.6295009 -0.6175743 -0.6172554 -0.5650262 -0.5323952
## highest:  0.7124199  0.7215555  0.7493782  0.7727731  0.8000316
```

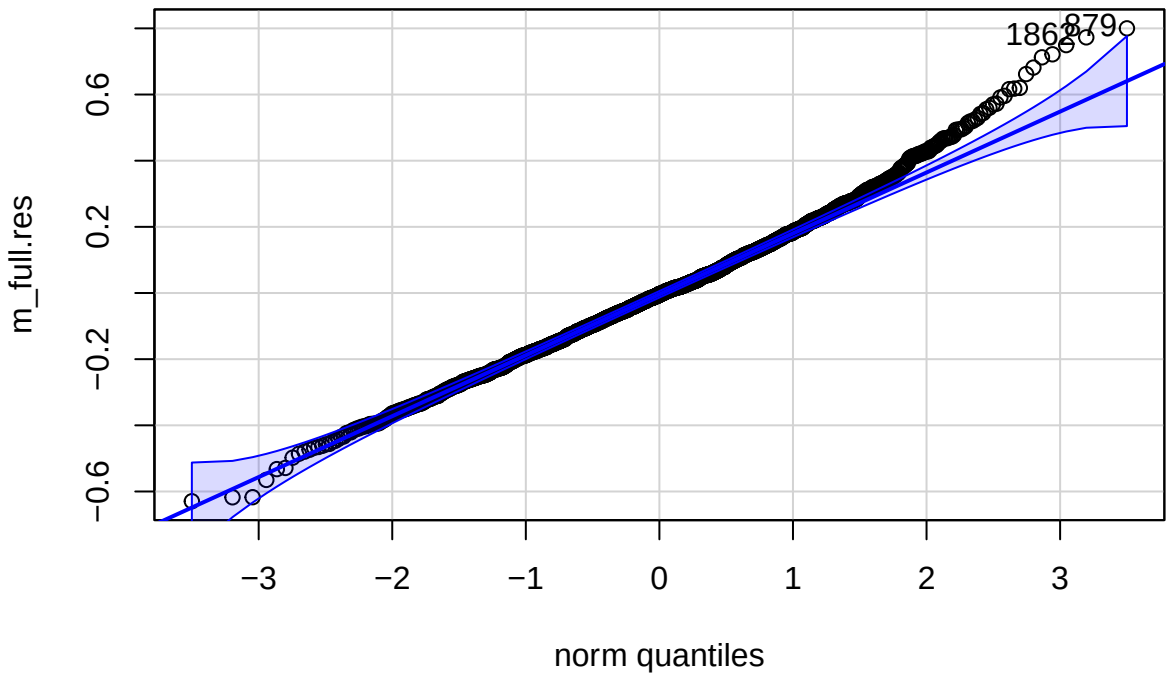
```
Hmisc::describe(m_full.res)$counts[c(".25", ".50", ".75")] #not symmetric
```

```
##         .25         .50         .75
## "-0.128236" "-0.003338" " 0.120187"
```

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



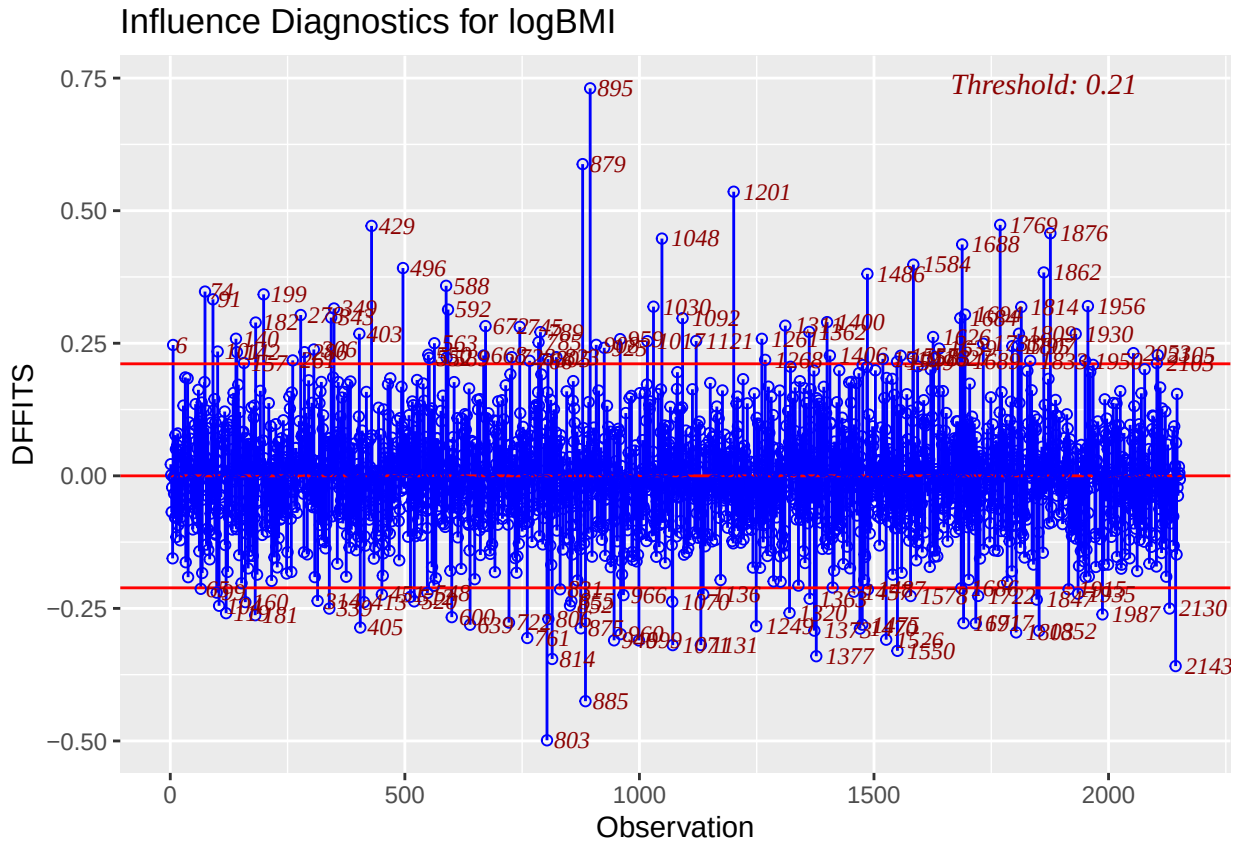

```
# Q-Q plot
qq.m_full.res = car::qqPlot(m_full.res)
```




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

```
##       879      1862
## 0.8000316 0.7727731
```

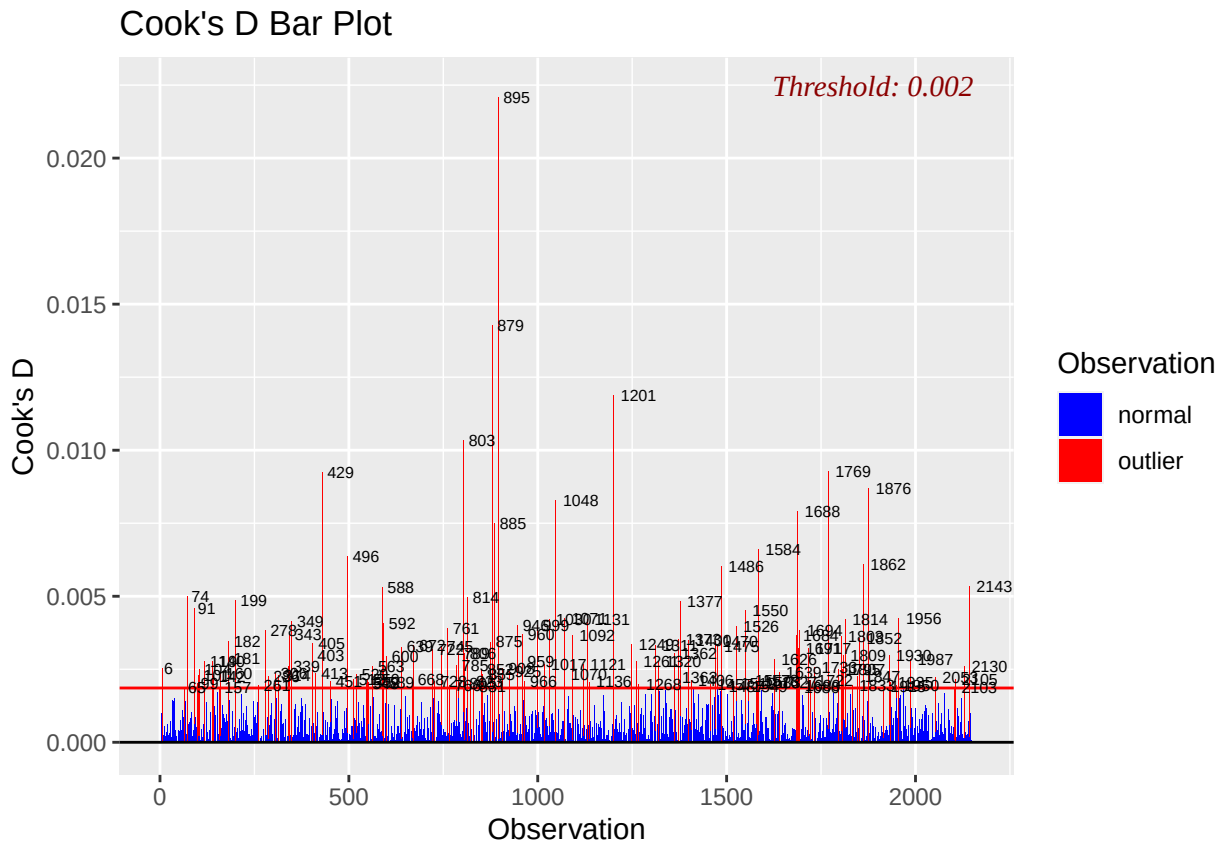
```
################ influential observations  ##################

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   0.7493782  3.914236 0.03367642 0.8809802  0.7307160   0.022099023
## 879   0.8000316  4.150680 0.01966016 0.8500383  0.5877930   0.014286906
## 1201  0.7215555  3.741538 0.02009154 0.8817344  0.5357522   0.011886988
## 803  -0.4737505 -2.476052 0.03899166 0.9821368 -0.4987490   0.010339678
## 1769  0.6200737  3.214408 0.02121657 0.9198195  0.4732550   0.009291347
## 429   0.4536477  2.369539 0.03805140 0.9868916  0.4712738   0.009234098
```

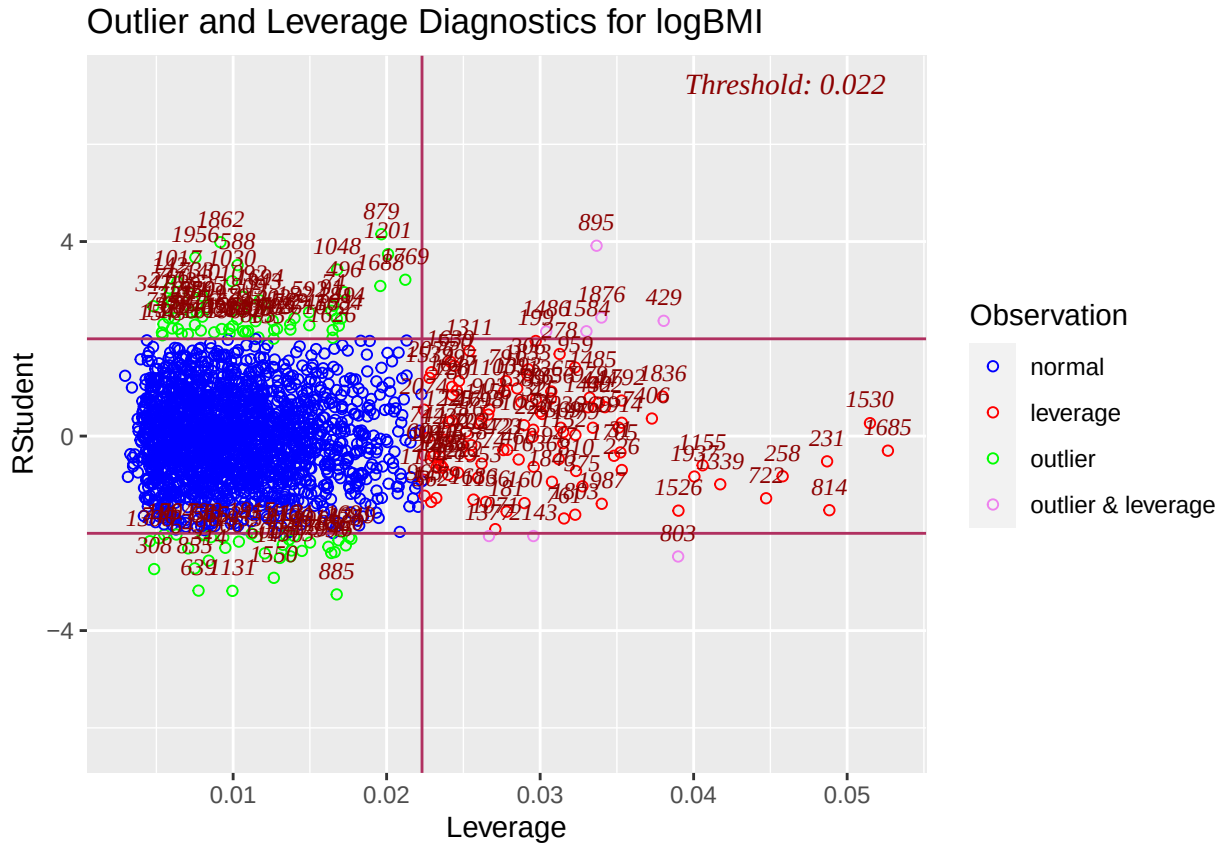
```
#From the plot above, we can see 2 observations with the largest (magnitude) of DFFITS, observation 879

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




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

```
##        Residual  Rstudent   HatDiagH  CovRatio     DFFITS COOKsDistance
## 895   0.7493782  3.914236 0.03367642 0.8809802  0.7307160   0.022099023
## 879   0.8000316  4.150680 0.01966016 0.8500383  0.5877930   0.014286906
## 1201  0.7215555  3.741538 0.02009154 0.8817344  0.5357522   0.011886988
## 803  -0.4737505 -2.476052 0.03899166 0.9821368 -0.4987490   0.010339678
## 1769  0.6200737  3.214408 0.02121657 0.9198195  0.4732550   0.009291347
## 429   0.4536477  2.369539 0.03805140 0.9868916  0.4712738   0.009234098
```

```
#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 -0.05717953 -0.3005693 0.05265617 1.066471 -0.07086233  0.0002093174
## 1530  0.05054162  0.2655116 0.05148766 1.065395  0.06186051  0.0001595165
## 814  -0.29039626 -1.5242293 0.04884165 1.035780 -0.34539739  0.0049677175
## 231  -0.09903365 -0.5195159 0.04869397 1.059879 -0.11753747  0.0005758250
## 258  -0.15784673 -0.8268639 0.04579870 1.051742 -0.18115094  0.0013675226
## 722  -0.24426659 -1.2791298 0.04472129 1.039332 -0.27676219  0.0031906007
```

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

```
##       Residual Rstudent   HatDiagH  CovRatio    DFFITS COOKsDistance
## 879  0.8000316 4.150680 0.01966016 0.8500383 0.5877930   0.014286906
## 1862 0.7727731 3.986736 0.00916931 0.8536949 0.3835183   0.006085999
## 895  0.7493782 3.914236 0.03367642 0.8809802 0.7307160   0.022099023
## 1201 0.7215555 3.741538 0.02009154 0.8817344 0.5357522   0.011886988
## 1956 0.7124199 3.670288 0.00754466 0.8757569 0.3200107   0.004242090
## 588  0.6814039 3.514441 0.01028842 0.8892826 0.3583244   0.005321463
```

```
#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

rm4.df3 = df3[-c(879, 1769, 1155, 1048, 1769, 1684, 74, 72, 1689, 1311), ]
rm.m_full =  lm(
  logBMI ~ SleepHrsNight + Age + Gender + factor(Race1)  + Poverty + TotChol + BPDiaAve + BPSysAve + Al
    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)           3.3537894309 9.989419e-02 33.5734168 1.123611e-198
## SleepHrsNight        -0.0200785445 1.186203e-02 -1.6926741  9.066389e-02
## Age                  -0.0034518684 1.966607e-03 -1.7552406  7.936204e-02
## Gender                0.1099051893 4.498417e-02  2.4431968  1.463857e-02
## factor(Race1)2       -0.0491060825 2.009192e-02 -2.4440717  1.460322e-02
## factor(Race1)3       -0.0213162050 1.761249e-02 -1.2102890  2.263024e-01
## factor(Race1)4       -0.0414615534 1.320273e-02 -3.1403762  1.710483e-03
## factor(Race1)5       -0.1023434373 1.977829e-02 -5.1745331  2.499210e-07
## Poverty               0.0028885308 2.877839e-03  1.0037152  3.156300e-01
## TotChol               0.0040303172 4.261675e-03  0.9457121  3.444028e-01
## BPDiaAve              0.0018683309 4.295272e-04  4.3497388  1.427377e-05
## BPSysAve              0.0016839824 3.702298e-04  4.5484794  5.707379e-06
## AlcoholYear          -0.0002890427 4.748584e-05 -6.0869245  1.361506e-09
## Smoke100             -0.0276631709 9.024910e-03 -3.0652018  2.202553e-03
## UrineFlow1           -0.0034584632 4.464410e-03 -0.7746742  4.386183e-01
## DaysMentHlthBad      -0.0010774854 5.648725e-04 -1.9074842  5.659241e-02
## DaysPhysHlthBad       0.0004420593 6.562161e-04  0.6736489  5.006077e-01
## factor(HealthGen)2   -0.0690003838 3.143369e-02 -2.1951094  2.826341e-02
## factor(HealthGen)3   -0.1150852828 3.112386e-02 -3.6976544  2.231090e-04
## factor(HealthGen)4   -0.1669739921 3.194540e-02 -5.2268550  1.892653e-07
## factor(HealthGen)5   -0.2315533469 3.375480e-02 -6.8598639  9.003994e-12
## PhysActive           -0.0250013367 9.246295e-03 -2.7039304  6.907007e-03
## SleepHrsNight:Age     0.0005890657 2.827069e-04  2.0836625  3.730994e-02
## SleepHrsNight:Gender -0.0155433776 6.484923e-03 -2.3968486  1.662254e-02
```

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

```
##                          Estimate   Std. Error    t value      Pr(>|t|)
## (Intercept)           3.346455e+00 0.0991000876 33.7684321 2.218301e-200
## SleepHrsNight        -1.929861e-02 0.0117433367 -1.6433670  1.004554e-01
## Age                  -3.402999e-03 0.0019464509 -1.7483098  8.055521e-02
## Gender                1.096600e-01 0.0446514120  2.4559137  1.413244e-02
## factor(Race1)2       -3.623456e-02 0.0199279320 -1.8182802  6.916244e-02
## factor(Race1)3       -8.458696e-03 0.0174930186 -0.4835469  6.287574e-01
## factor(Race1)4       -2.887175e-02 0.0131560927 -2.1945537  2.830381e-02
## factor(Race1)5       -9.029518e-02 0.0196073438 -4.6051717  4.365636e-06
## Poverty               2.970943e-03 0.0028462867  1.0437960  2.966988e-01
## TotChol               4.485130e-03 0.0042836257  1.0470407  2.952002e-01
## BPDiaAve              1.783673e-03 0.0004252330  4.1945778  2.846441e-05
## BPSysAve              1.712063e-03 0.0003685889  4.6449115  3.610901e-06
## AlcoholYear          -3.098086e-04 0.0000472204 -6.5609045  6.702052e-11
```

```
## Smoke100             -2.928532e-02 0.0089333001 -3.2782196  1.061536e-03
## UrineFlow1           -3.859864e-03 0.0044135082 -0.8745569  3.819141e-01
## DaysMentHlthBad      -9.465186e-04 0.0005605080 -1.6886800  9.142791e-02
## DaysPhysHlthBad      -8.523803e-05 0.0006572556 -0.1296878  8.968257e-01
## factor(HealthGen)2   -8.463605e-02 0.0311739764 -2.7149584  6.682174e-03
## factor(HealthGen)3   -1.237808e-01 0.0307934490 -4.0197113  6.030688e-05
## factor(HealthGen)4   -1.763913e-01 0.0316175617 -5.5789030  2.730030e-08
## factor(HealthGen)5   -2.407548e-01 0.0334034113 -7.2074911  7.890079e-13
## PhysActive           -2.327114e-02 0.0091467795 -2.5441894  1.102360e-02
## SleepHrsNight:Age     5.954039e-04 0.0002797893  2.1280439  3.344856e-02
## SleepHrsNight:Gender -1.597917e-02 0.0064322249 -2.4842364  1.305968e-02
```

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

```
##          (Intercept)        SleepHrsNight                  Age
##            0.2187035            3.8844074            1.4157305
##               Gender       factor(Race1)2       factor(Race1)3
##            0.2230774           26.2116569           60.3180035
##       factor(Race1)4       factor(Race1)5              Poverty
##           30.3650028           11.7723744            2.8530731
##              TotChol             BPDiaAve             BPSysAve
##           11.2848002            4.5312209            1.6674989
##          AlcoholYear              Smoke100           UrineFlow1
##            7.1843432            5.8639293           11.6063401
##      DaysMentHlthBad      DaysPhysHlthBad   factor(HealthGen)2
##           12.1548589          119.2820354           22.6602580
##   factor(HealthGen)3   factor(HealthGen)4   factor(HealthGen)5
##            7.5556931            5.6399905            3.9737904
##           PhysActive    SleepHrsNight:Age SleepHrsNight:Gender
##            6.9204199            1.0759813            2.8037003
```

```
#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

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
```