

model4

Liancheng

2023-11-25

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

##           used (Mb) gc trigger (Mb) max used (Mb)
## Ncells  472187 25.3    1018675 54.5   660860 35.3
## Vcells  896974  6.9     8388608 64.0  1800812 13.8

set.seed(123)
library(car)
library(ggplot2)
library(olsrr)

##### (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

```

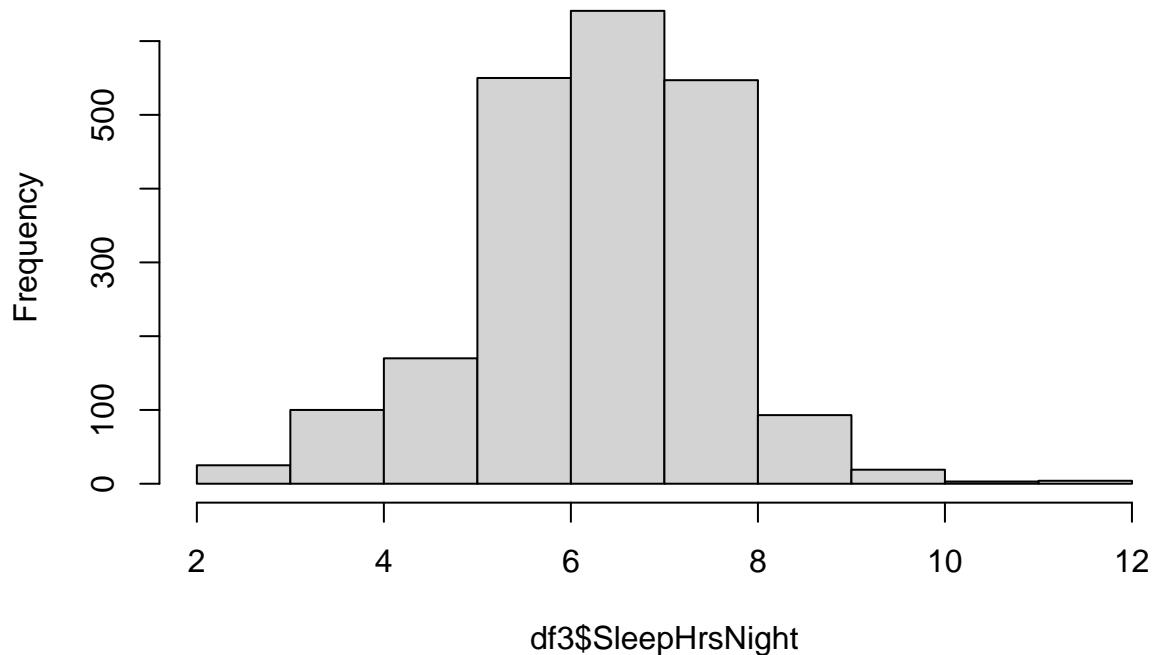
```

## 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

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

```

Histogram of df3\$SleepHrsNight



```

# colSums(is.na(df2)) / nrow(df2)
fit0 <-
  lm(SleepHrsNight ~ .,
    data = df3)
#data type
df3$Gender <- ifelse(df3$Gender == "male", 0, 1)
df3$Smoke100 <- ifelse(df3$Smoke100 == "No", 0, 1)

```

```

df3$PhysActive <- ifelse(df3$PhysActive == "No", 0, 1)
df3 <- df3 %>%
  mutate(
    Race1 = case_when(
      Race1 == 'Black' ~ 1,
      Race1 == 'Hispanic' ~ 2,
      Race1 == 'Mexican' ~ 3,
      Race1 == 'White' ~ 4,
      Race1 == 'Other' ~ 5,
      TRUE ~ NA_integer_ # Default value if none of the conditions are met
    )
  )
df3 <- 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 + AlcoholYear +
  DaysPhysHlthBad + factor(HealthGen) + PhysActive + SleepHrsNight*Age + SleepHrsNight*Gender + SleepHrsNight*factor(Race1)
  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 +
##     SleepHrsNight * factor(Race1), data = df3)
##
## Residuals:
##       Min     1Q   Median     3Q    Max 
## -16.958 -4.088 -0.576  3.191 36.357 
##
## Coefficients:
## (Intercept)           Estimate Std. Error t value Pr(>|t|)    
## SleepHrsNight          29.960505  3.507260   8.542 < 2e-16 ***
## Age                     -0.672481  0.440017  -1.528  0.12658  
## Gender                  -0.080205  0.063471  -1.264  0.20649  
## factor(Race1)2         3.956938  1.441705   2.745  0.00611 ** 
## factor(Race1)3         -0.876564  3.071088  -0.285  0.77535  
## factor(Race1)4         -3.407567  2.778060  -1.227  0.22011  
## factor(Race1)5         -5.099534  1.932348  -2.639  0.00838 ** 
## Poverty                 1.067095  3.249451   0.328  0.74265  
## factor(Race1)1          0.054070  0.091689   0.590  0.55544 

```

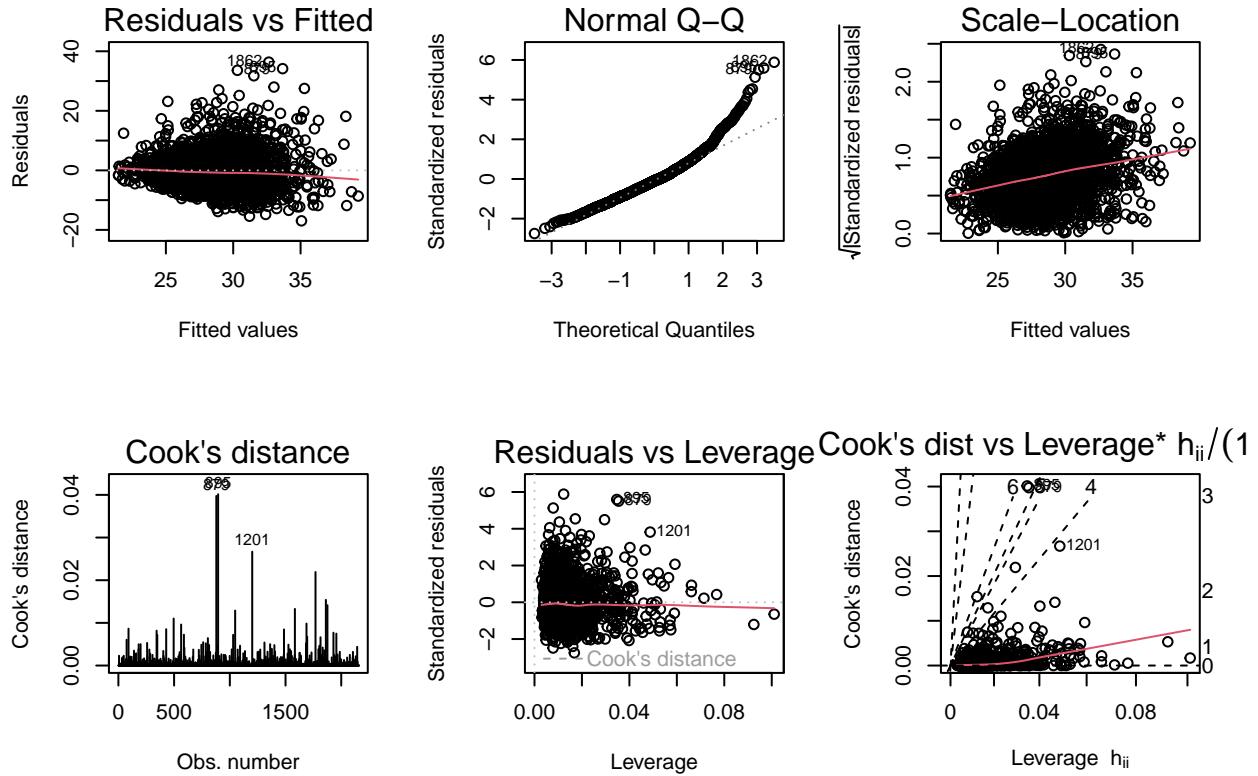
```

## TotChol          0.012933  0.135840  0.095  0.92416
## BPDiaAve        0.057750  0.013676  4.223  2.52e-05 ***
## BPSysAve         0.052227  0.011793  4.429  9.96e-06 ***
## AlcoholYear      -0.009047 0.001517  -5.966  2.84e-09 ***
## Smoke100         -0.847770 0.287236  -2.951  0.00320 **
## UrineFlow1        -0.088739 0.142102  -0.624  0.53238
## DaysMentHlthBad -0.032621 0.017991  -1.813  0.06993 .
## DaysPhysHlthBad  0.014998 0.020905  0.717  0.47319
## factor(HealthGen)2 -2.171783 1.002132  -2.167  0.03033 *
## factor(HealthGen)3 -3.905240 0.993873  -3.929  8.79e-05 ***
## factor(HealthGen)4 -5.635919 1.018732  -5.532  3.55e-08 ***
## factor(HealthGen)5 -7.518320 1.075750  -6.989  3.69e-12 ***
## PhysActive        -0.891431 0.294530  -3.027  0.00250 **
## SleepHrsNight:Age 0.012971 0.009134  1.420  0.15574
## SleepHrsNight:Gender -0.508514 0.207897  -2.446  0.01453 *
## SleepHrsNight:factor(Race1)2 -0.160437 0.452870  -0.354  0.72317
## SleepHrsNight:factor(Race1)3 0.334059 0.403929  0.827  0.40832
## SleepHrsNight:factor(Race1)4 0.544607 0.287394  1.895  0.05823 .
## SleepHrsNight:factor(Race1)5 -0.629379 0.475731  -1.323  0.18599
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.217 on 2124 degrees of freedom
## Multiple R-squared:  0.1631, Adjusted R-squared:  0.1525
## F-statistic: 15.33 on 27 and 2124 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) 2821   1 72.9730 < 2.2e-16 ***
## SleepHrsNight 90    1  2.3357  0.126585
## Age          62    1  1.5968  0.206493
## Gender        291   1  7.5330  0.006109 **
## factor(Race1) 409   4  2.6474  0.031870 *
## Poverty       13    1  0.3478  0.555444
## TotChol        0    1  0.0091  0.924157
## BPDiaAve      689   1 17.8308  2.516e-05 ***
## BPSysAve       758   1 19.6127  9.965e-06 ***
## AlcoholYear    1376   1 35.5910  2.844e-09 ***
## Smoke100       337   1  8.7112  0.003197 **
## UrineFlow1     15    1  0.3900  0.532385
## DaysMentHlthBad 127   1  3.2879  0.069935 .
## DaysPhysHlthBad 20    1  0.5147  0.473188
## factor(HealthGen) 4622   4 29.8917 < 2.2e-16 ***
## PhysActive     354   1  9.1605  0.002503 **
## SleepHrsNight:Age 78    1  2.0165  0.155745
## SleepHrsNight:Gender 231   1  5.9828  0.014526 *
## SleepHrsNight:factor(Race1) 430   4  2.7841  0.025329 *
## Residuals      82107 2124
## ---
## 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  0.01301 0.008551 0.004865 0.005528
##    .25          .50     .75       .90       .95
##  0.007242 0.010523 0.015295 0.023125 0.031581
##
## lowest : 0.002972138 0.003170460 0.003312764 0.003418926 0.003468613
```

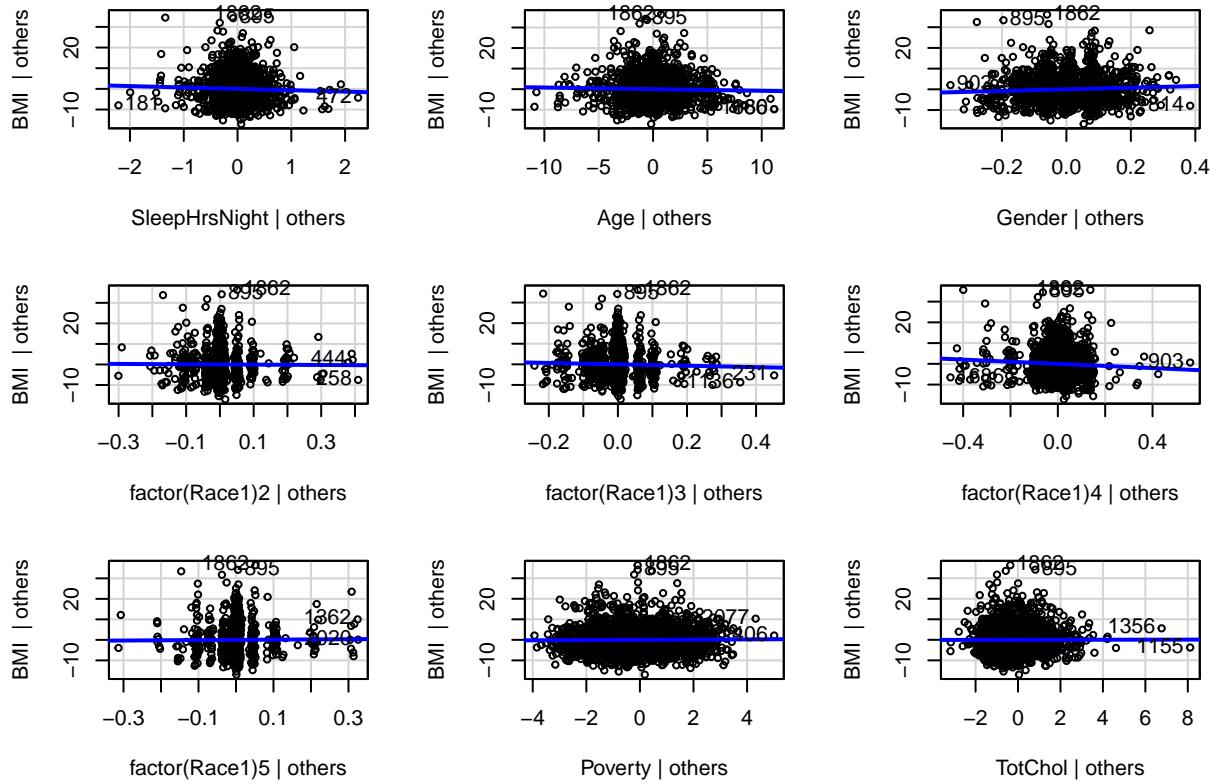
```

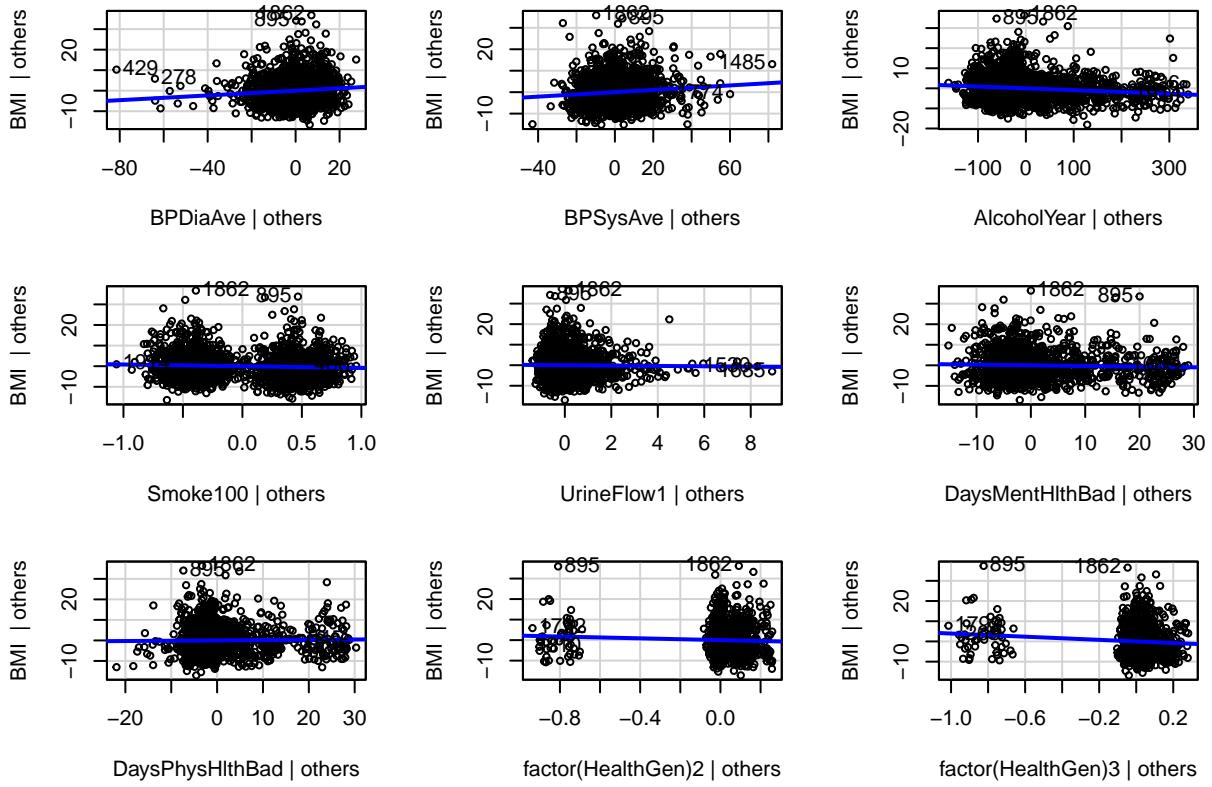
## highest: 0.066174863 0.071440848 0.076776919 0.092465730 0.101033382
m_full.h[which.max(m_full.h)]

##      231
## 0.1010334
##### Assumption:LINE #####
#(1)Linear: 2 approaches

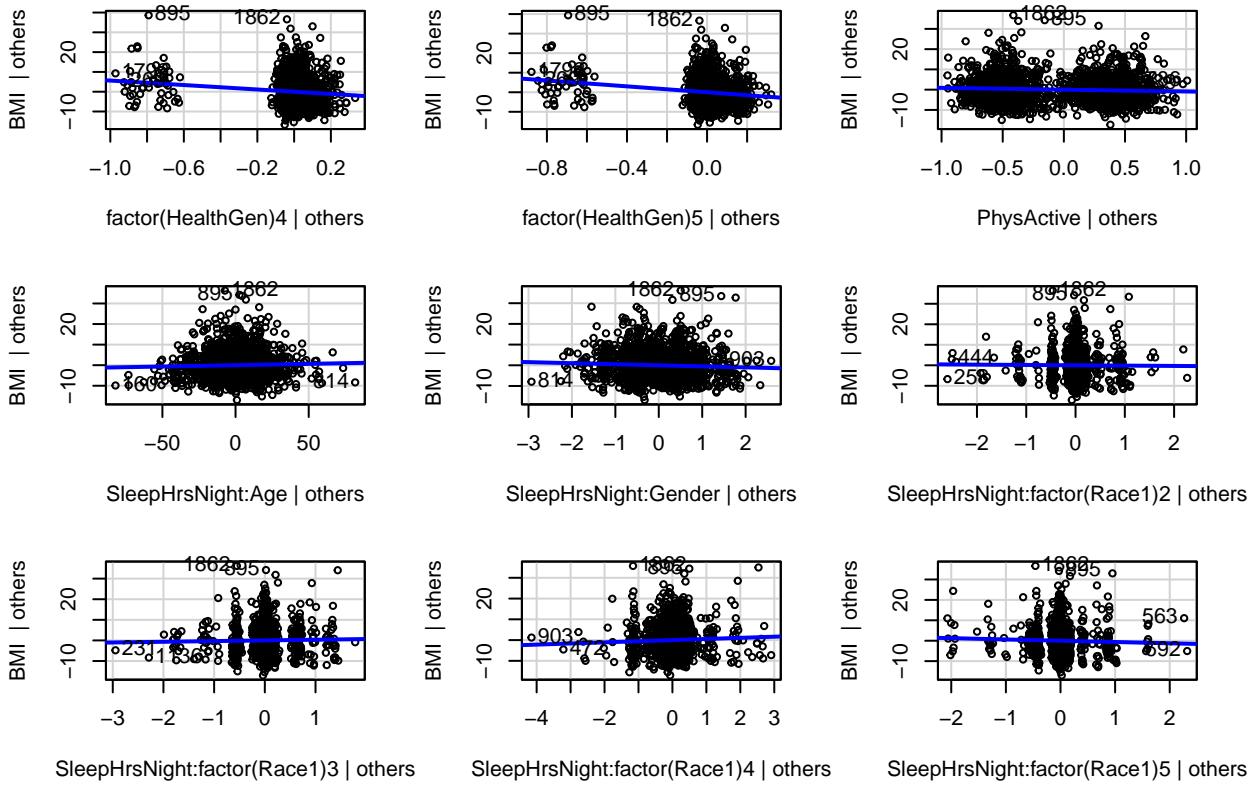
# partial regression plots
car::avPlots(m_full)

```





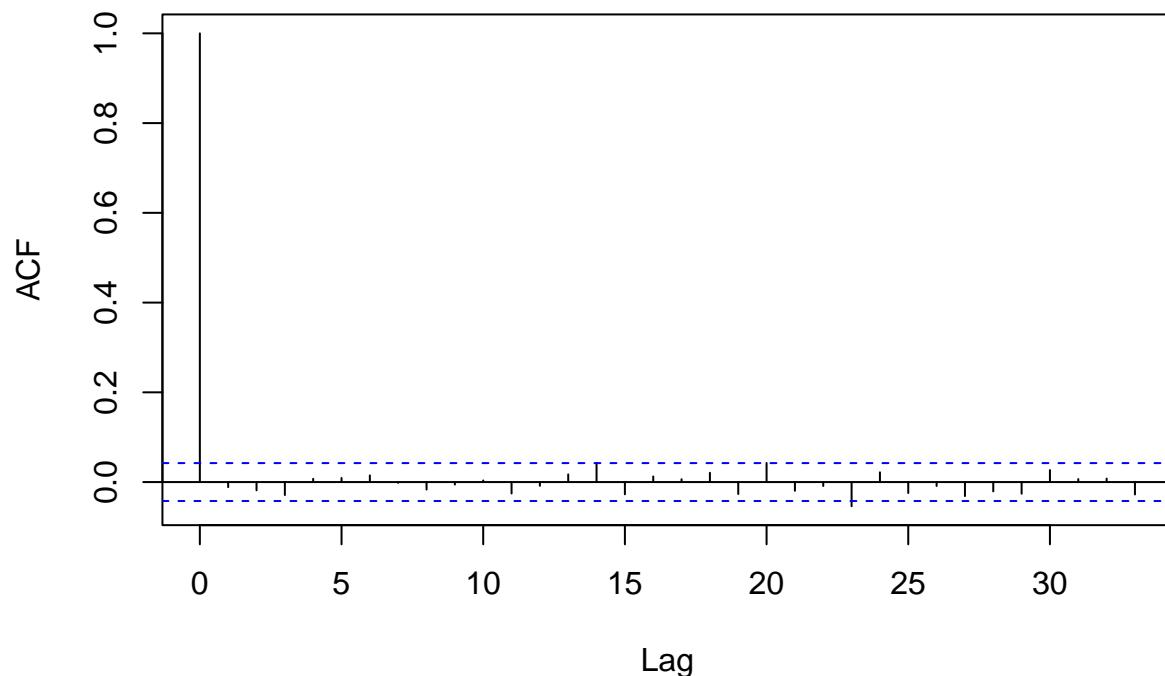
Added-Variable Plots



#(2) Independence:

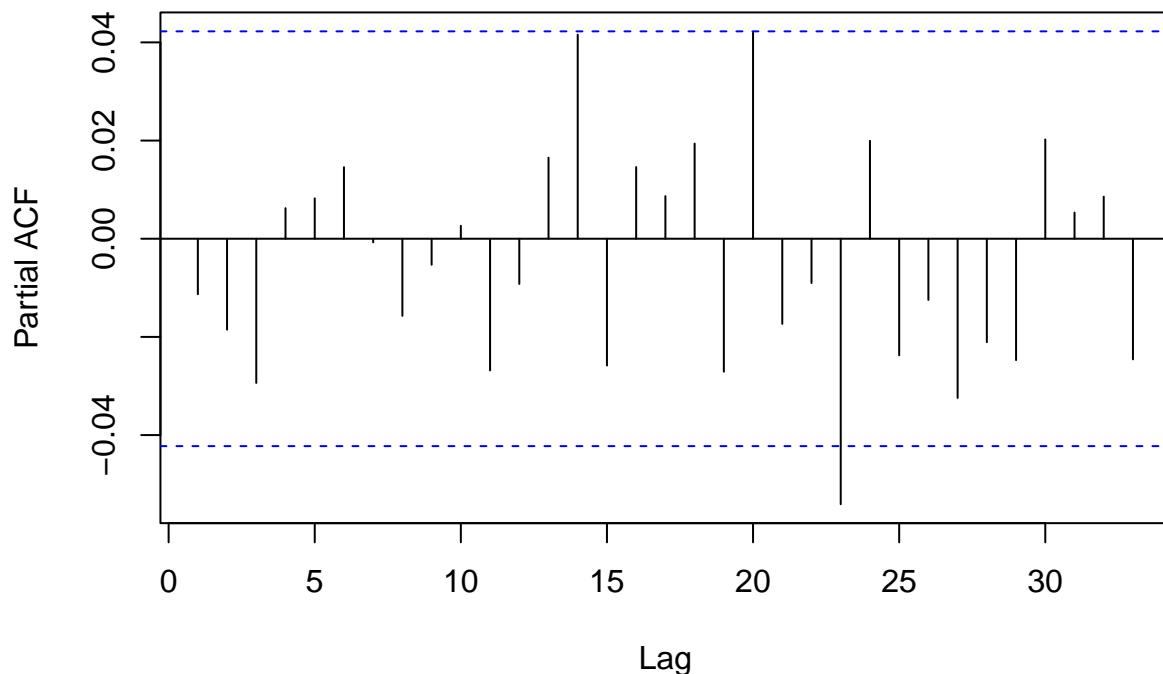
```
residuals <- resid(m_full)
acf(residuals, main = "Autocorrelation Function of Residuals")
```

Autocorrelation Function of Residuals

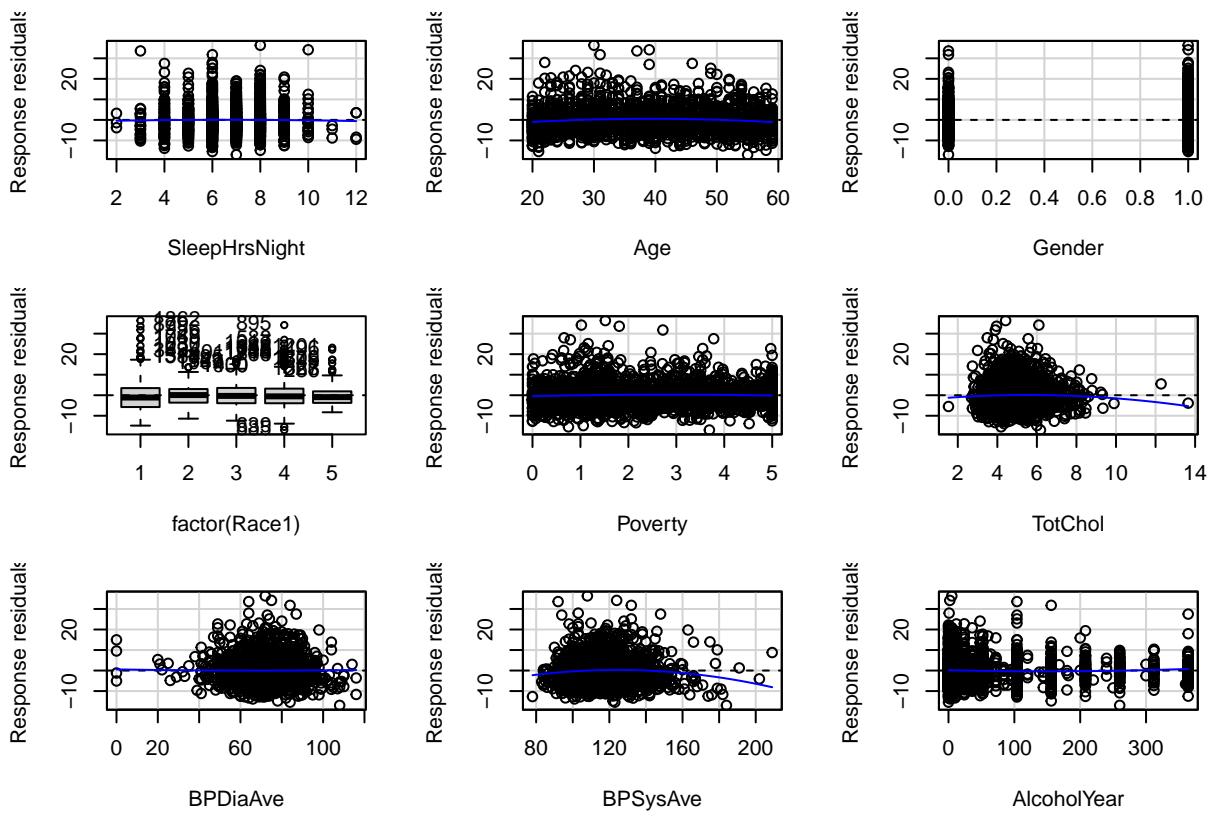


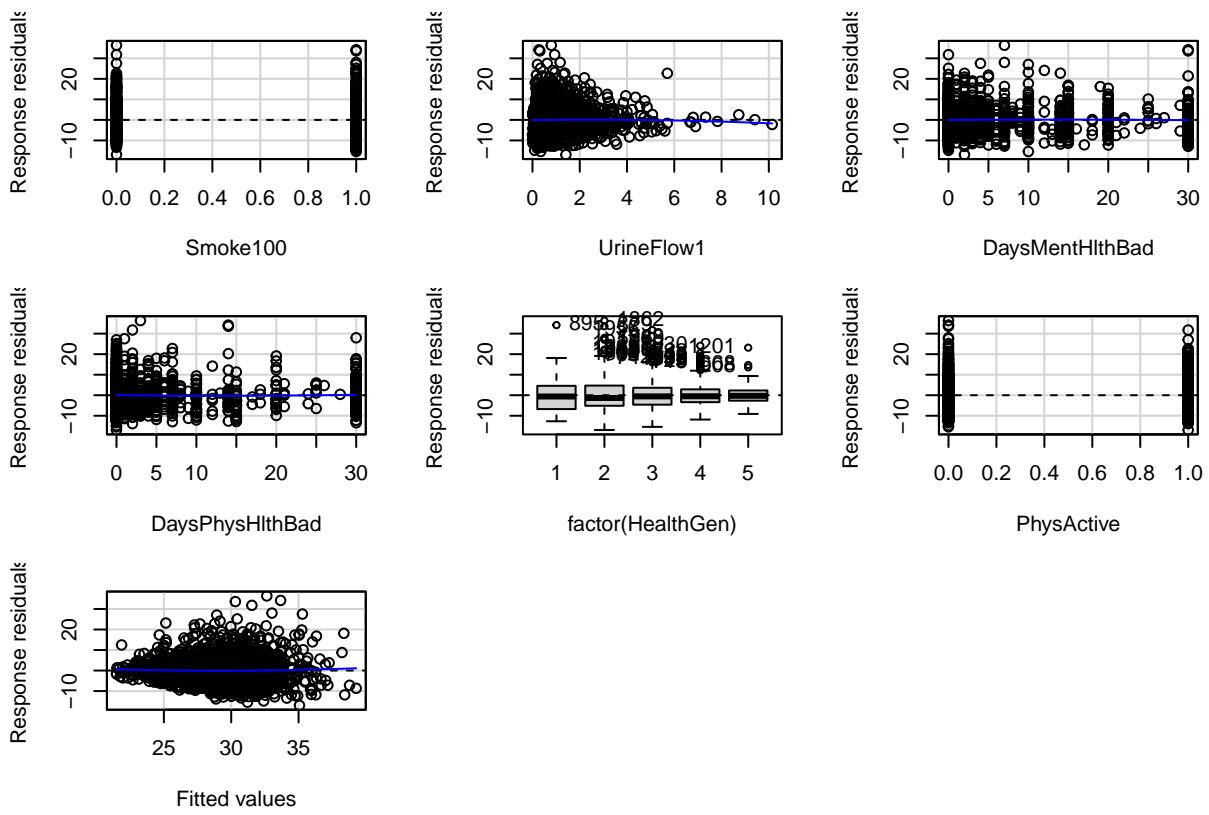
```
pacf(residuals, main = "Partial Autocorrelation Function of Residuals")
```

Partial Autocorrelation Function of Residuals



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

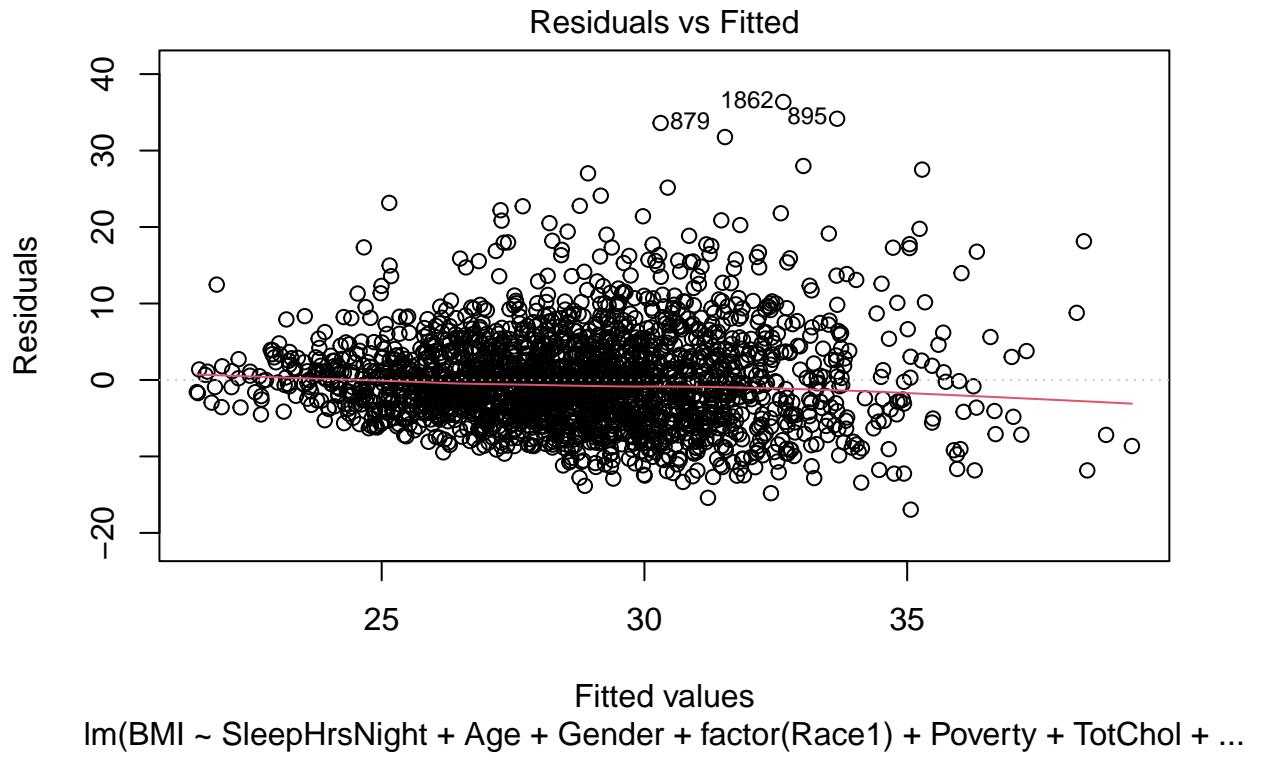




```

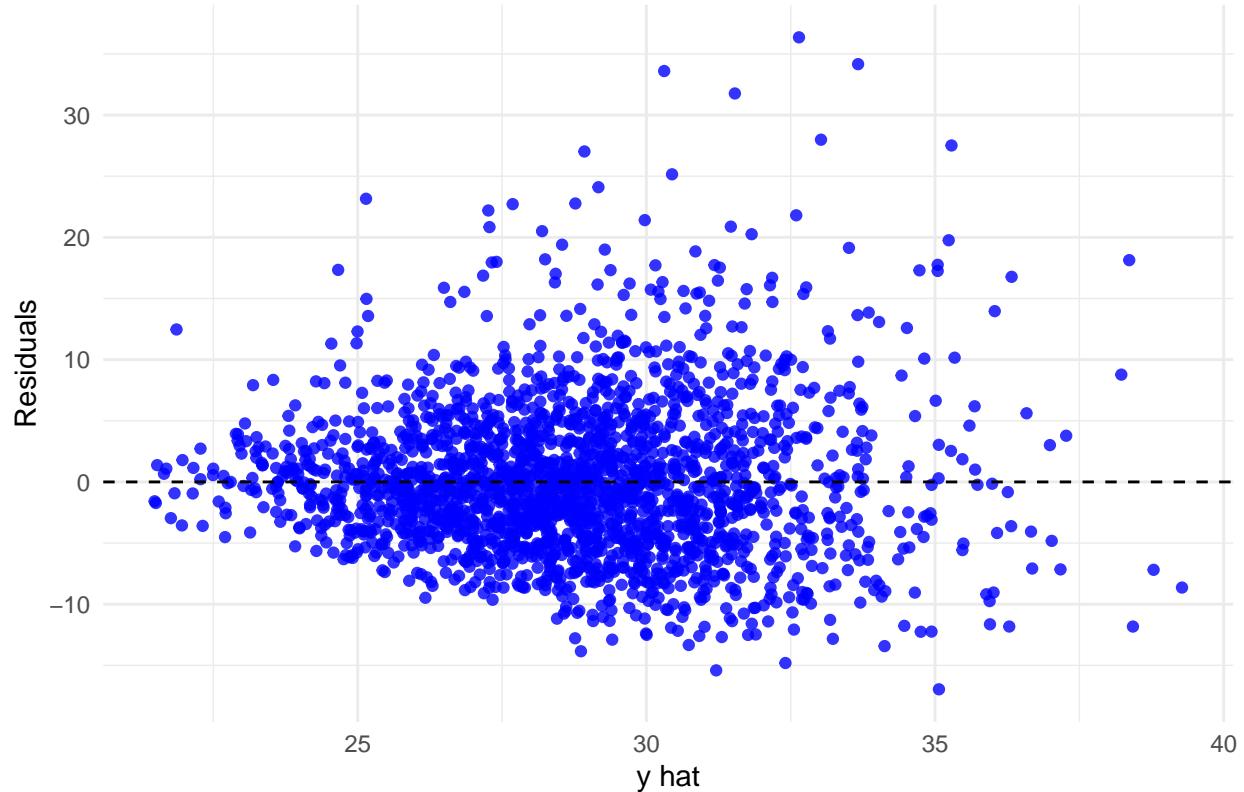
##              Test stat Pr(>|Test stat|)
## SleepHrsNight      -0.4613    0.6446186
## Age                 -3.7447   0.0001854 ***
## Gender                0.3845   0.7006805
## factor(Race1)
## Poverty             -1.4419   0.1494722
## TotChol              -1.4869   0.1371842
## BPDiaAve              0.3082   0.7579352
## BPSysAve             -3.6531   0.0002654 ***
## AlcoholYear            1.9135   0.0558163 .
## Smoke100               0.0927   0.9261142
## UrineFlow1              -0.6018  0.5473911
## DaysMentHlthBad        -0.4951   0.6205792
## DaysPhysHlthBad         0.7946   0.4269417
## factor(HealthGen)
## PhysActive             -0.4795   0.6315971
## Tukey test              1.2073   0.2273166
## ---
## 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()
```

constant variance assumption



```
#conclusion: the constant variance assumption is basically not violated. The spread of the residuals appears roughly constant across the range of y-hat.

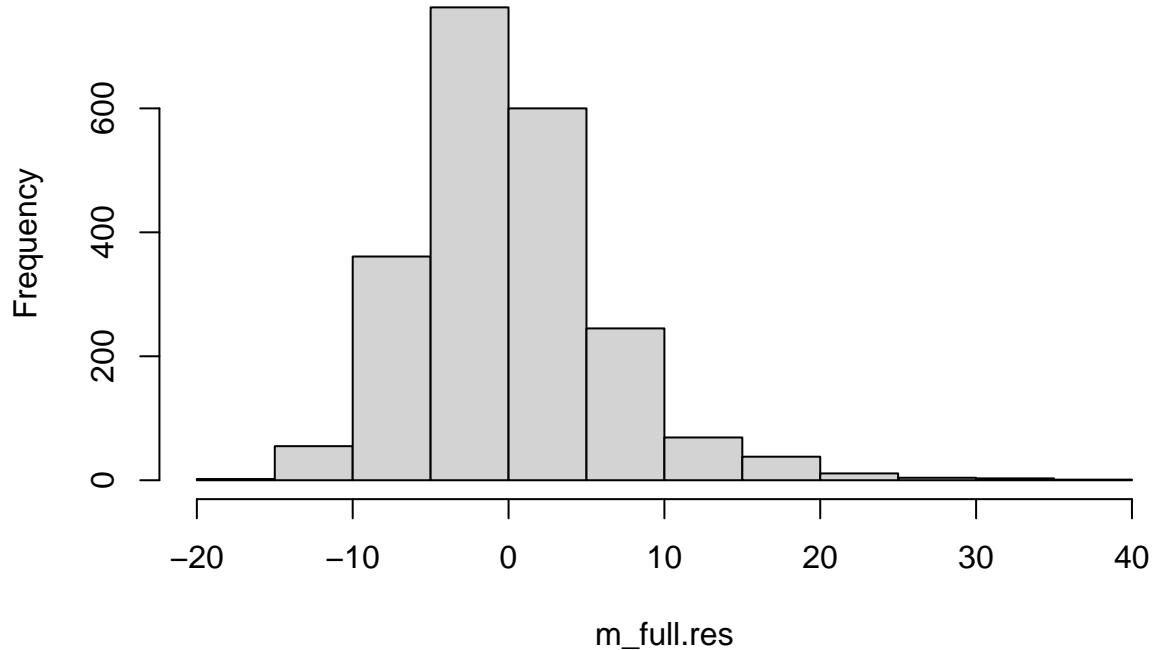
#(4)Normality: residuals freq - residuals (4 plots: his, box, Q-Q, shapiro); transform
#exam quartiles of the residuals
Hmisc::describe(m_full.res)

## m_full.res
##      n    missing   distinct      Info      Mean       Gmd      .05
##     2152        0     2152      1 -2.195e-17     6.655 -8.6296
##     .10        .25     .50      .75      .90      .95
##    -7.0359   -4.0879   -0.5759     3.1910     7.5243    10.4454
##
## lowest : -16.95814 -15.40920 -14.80810 -13.84646 -13.42587
## highest:  27.98564  31.76705  33.60166  34.16306  36.35676
Hmisc::describe(m_full.res)$counts[c(".25", ".50", ".75")] #not symmetric

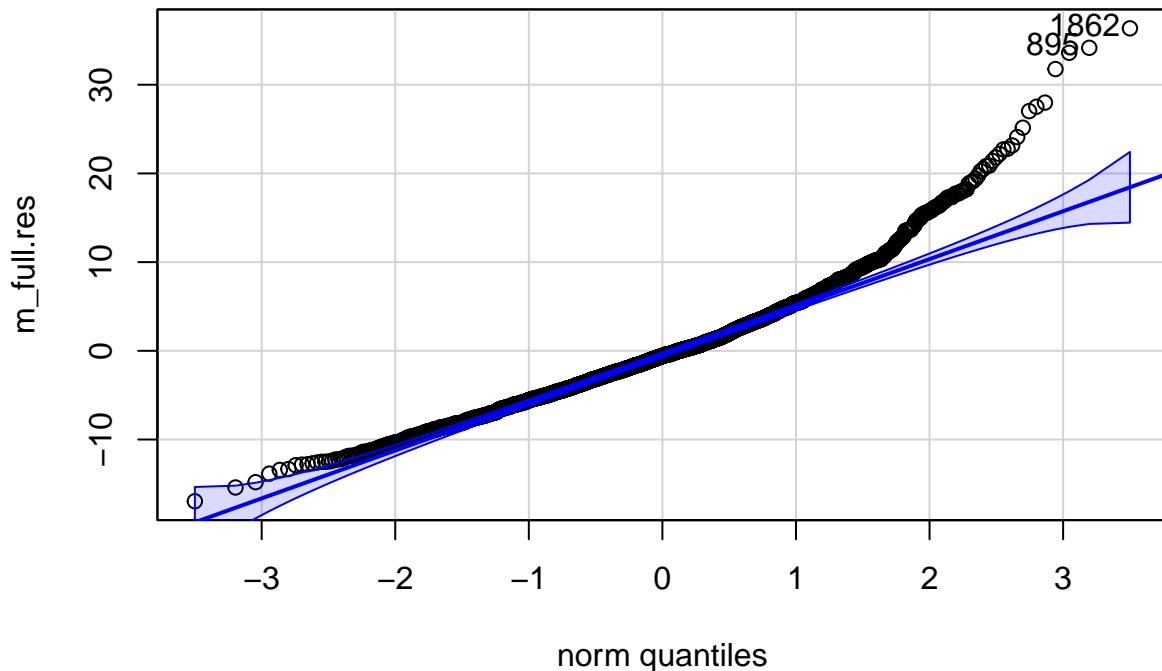
##      .25      .50      .75
## "-4.0879" "-0.5759" " 3.1910"

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

Histogram of m_full.res



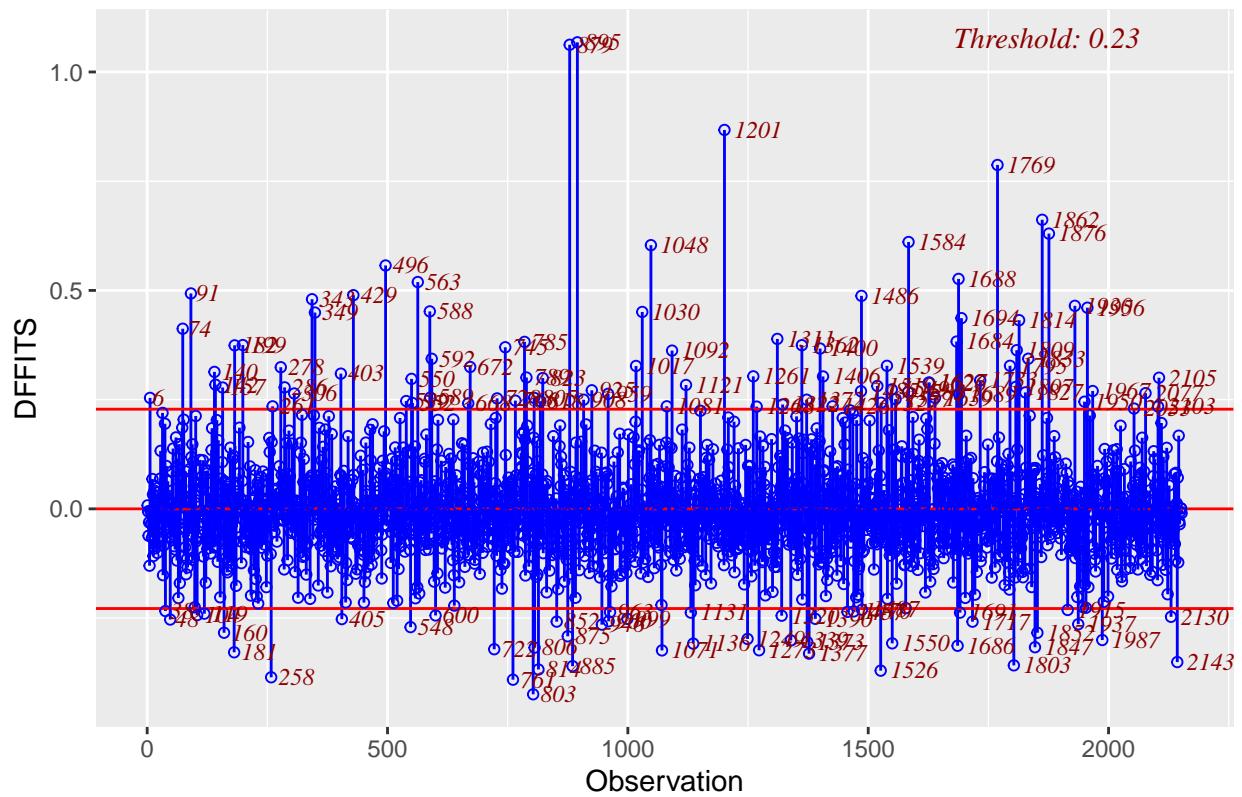
```
# Q-Q plot
qq.m_full.res = car::qqPlot(m_full.res)
```



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

##      1862      895
## 36.35676 34.16306
##### 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)
```

Influence Diagnostics for BMI



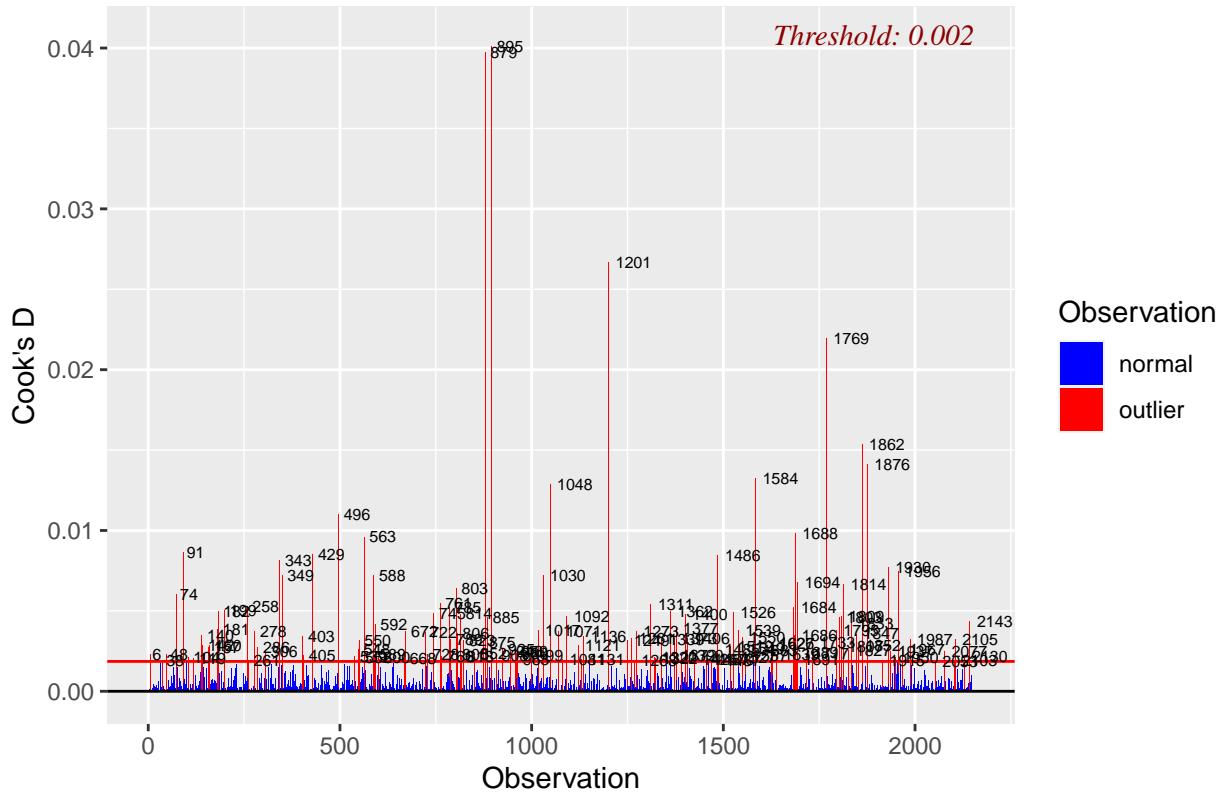
```
influence4[order(abs(influence4$DFFITS), decreasing = T), ] %>% head()
```

```
##      Residual Rstudent   HatDiagH CovRatio     DFFITS COOKsDistance
## 895  34.16306 5.632875 0.03468934 0.6928884 1.0678114 0.04014143
## 879  33.60166 5.541168 0.03544669 0.7027303 1.0622481 0.03974315
## 1201 23.15612 3.830897 0.04875209 0.8783372 0.8672618 0.02669040
## 1769 27.51498 4.512694 0.02952588 0.7991973 0.7871278 0.02192759
## 1862 36.35676 5.930967 0.01229464 0.6475374 0.6617130 0.01539036
## 1876 17.25396 2.846975 0.04668488 0.9553282 0.6300187 0.01412858
```

#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)
```

Cook's D Bar Plot



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

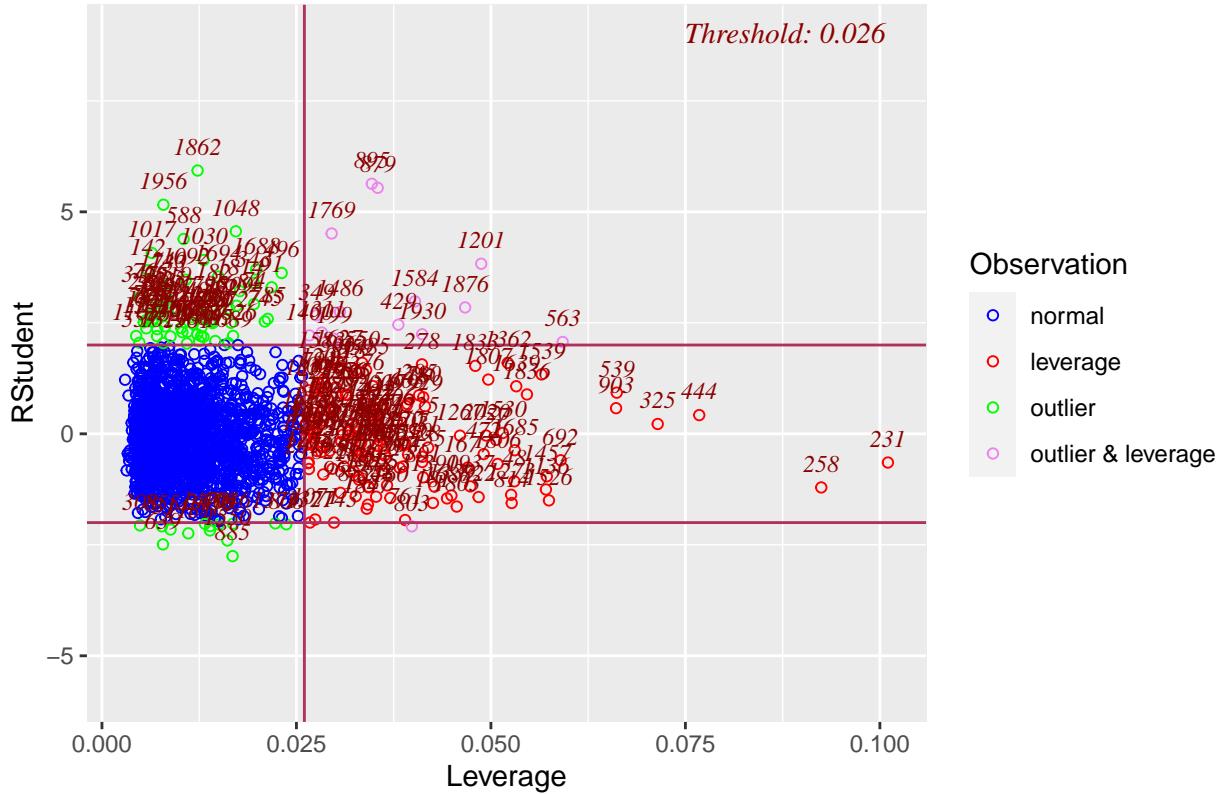
```
##      Residual Rstudent   HatDiagH CovRatio    DFFITS COOKsDistance
## 895  34.16306 5.632875 0.03468934 0.6928884 1.0678114  0.04014143
## 879  33.60166 5.541168 0.03544669 0.7027303 1.0622481  0.03974315
## 1201 23.15612 3.830897 0.04875209 0.8783372 0.8672618  0.02669040
## 1769 27.51498 4.512694 0.02952588 0.7991973 0.7871278  0.02192759
## 1862 36.35676 5.930967 0.01229464 0.6475374 0.6617130  0.01539036
## 1876 17.25396 2.846975 0.04668488 0.9553282 0.6300187  0.01412858
```

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

#leverage

```
ols_plot_resid_lev(m_full)
```

Outlier and Leverage Diagnostics for BMI



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

```
##      Residual   Rstudent   HatDiagH CovRatio      DFFITS COOKsDistance
## 231 -3.813195 -0.6467635 0.10103338 1.120952 -0.21682342 0.0016794740
## 258 -7.153198 -1.2078235 0.09246573 1.095243 -0.38553371 0.0053072906
## 444  2.520223  0.4217833 0.07677692 1.094967  0.12163298 0.0005285825
## 325  1.326059  0.2212833 0.07144085 1.090527  0.06137863 0.0001346080
## 539  5.570046  0.9270421 0.06617486 1.072851  0.24678184 0.0021751896
## 903  3.470564  0.5775193 0.06608684 1.080214  0.15362814 0.0008431791
```

#high studentized residual

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

```
##      Residual   Rstudent   HatDiagH CovRatio      DFFITS COOKsDistance
## 1862 36.35676 5.930967 0.012294639 0.6475374 0.6617130 0.015390364
## 895  34.16306 5.632875 0.034689340 0.6928884 1.0678114 0.040141432
## 879  33.60166 5.541168 0.035446686 0.7027303 1.0622481 0.039743153
## 1956 31.76705 5.160492 0.007897548 0.7204138 0.4604251 0.007480843
## 1048 27.98564 4.561499 0.017207050 0.7846276 0.6035730 0.012890516
## 1769 27.51498 4.512694 0.029525875 0.7991973 0.7871278 0.021927594
```

#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 e

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 + AlcoholYear +
    DaysPhysHlthBad + factor(HealthGen) + PhysActive + SleepHrsNight*Age + SleepHrsNight*Gender + SleepHrsNight*factor(Race1)
  rm4.df3
)
## Before removing these observations, the estimated coefficients are:
summary(m_full)$coef

##                                     Estimate Std. Error   t value Pr(>|t|) 
## (Intercept)                29.960505073 3.507260454 8.54242377 2.466760e-17
## SleepHrsNight              -0.672481427 0.440016672 -1.52830897 1.265847e-01
## Age                         -0.080204768 0.063470508 -1.26365410 2.064929e-01
## Gender                       3.956938092 1.441705429 2.74462315 6.109021e-03
## factor(Race1)2             -0.876564287 3.071087624 -0.28542471 7.753466e-01
## factor(Race1)3             -3.407566549 2.778060078 -1.22659930 2.201092e-01
## factor(Race1)4             -5.099534349 1.932348034 -2.63903513 8.375006e-03
## factor(Race1)5              1.067094548 3.249451033 0.32839225 7.426475e-01
## Poverty                      0.054070144 0.091688515 0.58971556 5.554441e-01
## TotChol                      0.012933319 0.135840431 0.09520965 9.241573e-01
## BPDiaAve                     0.057749909 0.013676197 4.22265846 2.516211e-05
## BPSysAve                      0.052226688 0.011792973 4.42862767 9.964814e-06
## AlcoholYear                   -0.009047472 0.001516551 -5.96582184 2.844035e-09
## Smoke100                      -0.847770435 0.287235608 -2.95148099 3.197186e-03
## UrineFlow1                     -0.088738883 0.142102264 -0.62447199 5.323847e-01
## DaysMentHlthBad               -0.032621346 0.017990576 -1.81324636 6.993492e-02
## DaysPhysHlthBad                0.014998194 0.020905479 0.71742883 4.731884e-01
## factor(HealthGen)2            -2.171782863 1.002132374 -2.16716166 3.033334e-02
## factor(HealthGen)3            -3.905239626 0.993873382 -3.92931303 8.791217e-05
## factor(HealthGen)4             -5.635919387 1.018731903 -5.53228908 3.550503e-08
## factor(HealthGen)5             -7.518320187 1.075750186 -6.98890903 3.694385e-12
## PhysActive                     -0.891431344 0.294529536 -3.02662801 2.502685e-03
## SleepHrsNight:Age              0.012970832 0.009134171 1.42003383 1.557446e-01
## SleepHrsNight:Gender            -0.508514014 0.207897348 -2.44598605 1.452629e-02
## SleepHrsNight:factor(Race1)2   -0.160436519 0.452869868 -0.35426627 7.231745e-01
## SleepHrsNight:factor(Race1)3   0.334059261 0.403929165 0.82702436 4.083162e-01
## SleepHrsNight:factor(Race1)4   0.544606530 0.287394425 1.89497946 5.823080e-02
## SleepHrsNight:factor(Race1)5   -0.629379219 0.475730981 -1.32297295 1.859868e-01

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

##                                     Estimate Std. Error   t value Pr(>|t|) 
## (Intercept)                28.268008351 3.459409877 8.17133828 5.189771e-16
## SleepHrsNight              -0.424021324 0.433321364 -0.97853778 3.279203e-01
## Age                         -0.079264749 0.062184319 -1.27467423 2.025646e-01
## Gender                       3.918842197 1.416789399 2.76600192 5.724139e-03
## factor(Race1)2              1.368019475 3.022848671 0.45255970 6.509122e-01
## factor(Race1)3              -0.913434900 2.739731420 -0.33340308 7.388631e-01
## factor(Race1)4              -2.879953436 1.919328994 -1.50050015 1.336342e-01
## factor(Race1)5              3.200399722 3.194996265 1.00169122 3.166073e-01
## Poverty                      0.055168101 0.089769348 0.61455387 5.389155e-01

```

```

## TotChol          0.030174864 0.135186651 0.22320891 8.233945e-01
## BPDiaAve        0.054810943 0.013401685 4.08985453 4.478056e-05
## BPSysAve         0.053259584 0.011621651 4.58278995 4.8555550e-06
## AlcoholYear      -0.009863362 0.001492621 -6.60808052 4.912196e-11
## Smoke100          -0.906976594 0.281453978 -3.22246856 1.290233e-03
## UrineFlow1        -0.103228010 0.139065343 -0.74229861 4.579889e-01
## DaysMentHlthBad   -0.027067099 0.017667463 -1.53203084 1.256644e-01
## DaysPhysHlthBad   -0.006917921 0.020728014 -0.33374742 7.386033e-01
## factor(HealthGen)2 -2.860916730 0.983991462 -2.90746093 3.681707e-03
## factor(HealthGen)3 -4.310290167 0.973430101 -4.42794009 9.998425e-06
## factor(HealthGen)4 -6.058304209 0.998133224 -6.06963486 1.515217e-09
## factor(HealthGen)5 -7.927510866 1.053831429 -7.52256068 7.893859e-14
## PhysActive         -0.802153028 0.288476385 -2.78065404 5.473008e-03
## SleepHrsNight:Age  0.013395759 0.008948908 1.49691548 1.345645e-01
## SleepHrsNight:Gender -0.522213533 0.204157956 -2.55788970 1.060010e-02
## SleepHrsNight:factor(Race1)2 -0.423357311 0.445647906 -0.94998160 3.422301e-01
## SleepHrsNight:factor(Race1)3  0.036331910 0.398266905 0.09122503 9.273224e-01
## SleepHrsNight:factor(Race1)4  0.283662710 0.285240570 0.99446832 3.201087e-01
## SleepHrsNight:factor(Race1)5 -0.880859176 0.467706683 -1.88335811 5.978896e-02

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

##             (Intercept)           SleepHrsNight
##               5.649093            36.946761
##                 Age              Gender
##               1.172024            0.962762
##   factor(Race1)2    factor(Race1)3
##             256.066075            73.193923
##   factor(Race1)4    factor(Race1)5
##             43.525168            199.917165
##                 Poverty            TotChol
##             2.030615            133.311061
##                 BPDiaAve           BPSysAve
##             5.089126            1.977715
##   AlcoholYear          Smoke100
##             9.017876            6.983749
##                 UrineFlow1          DaysMentHlthBad
##             16.327823            17.026419
##   DaysPhysHlthBad    factor(HealthGen)2
##             146.125029            31.731251
##   factor(HealthGen)3    factor(HealthGen)4
##             10.371977            7.494515
##   factor(HealthGen)5          PhysActive
##             5.442581            10.015165
##   SleepHrsNight:Age     SleepHrsNight:Gender
##             3.276022            2.694030
## SleepHrsNight:factor(Race1)2 SleepHrsNight:factor(Race1)3
##             163.878395            89.124112
## SleepHrsNight:factor(Race1)4 SleepHrsNight:factor(Race1)5
##             47.914192            39.956826

#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]
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-wise

```

# collinearity diagnostics (VIF)
car::vif(m_full)

##                                     GVIF Df GVIF^(1/(2*Df))
## SleepHrsNight           1.849122e+01  1      4.300141
## Age                      2.879008e+01  1      5.365639
## Gender                   2.882194e+01  1      5.368607
## factor(Race1)            5.340628e+05  4      5.199350
## Poverty                  1.338183e+00  1      1.156799
## TotChol                  1.135148e+00  1      1.065433
## BPDiaAve                 1.459656e+00  1      1.208162
## BPSysAve                 1.578347e+00  1      1.256323
## AlcoholYear               1.136198e+00  1      1.065926
## Smoke100                 1.142055e+00  1      1.068670

```

```

## UrineFlow1          1.049202e+00  1      1.024306
## DaysMentHlthBad   1.158938e+00  1      1.076540
## DaysPhysHlthBad   1.256335e+00  1      1.120864
## factor(HealthGen) 1.495098e+00  4      1.051559
## PhysActive         1.177165e+00  1      1.084972
## SleepHrsNight:Age  3.875954e+01  1      6.225716
## SleepHrsNight:Gender 3.046426e+01  1      5.519444
## SleepHrsNight:factor(Race1) 6.995071e+05  4      5.377730

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

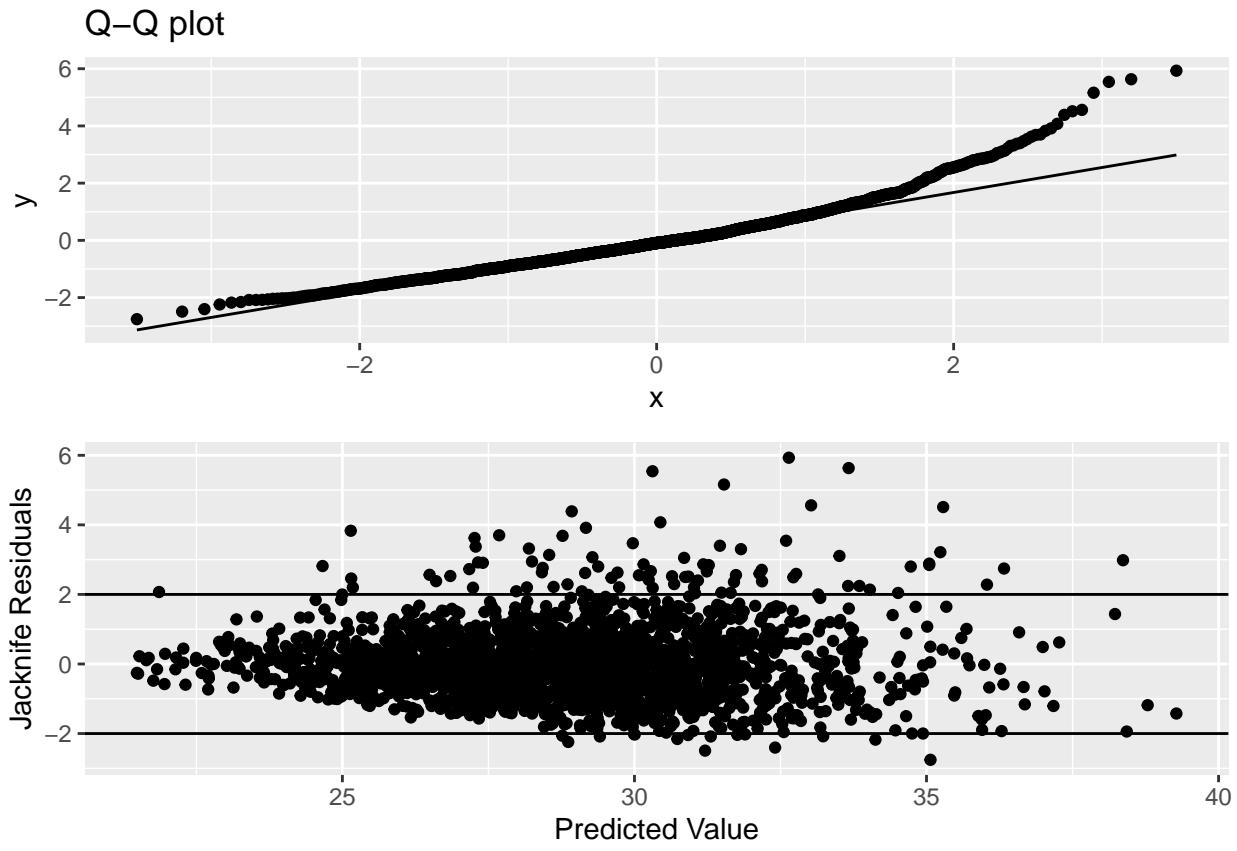
##### using log-transformed BMI #####
# log BMI
df3$logBMI = log(df3$BMI + 1)
m_full.log = lm(
  logBMI ~ SleepHrsNight + Age + Gender + factor(Race1) + Poverty + TotChol + BPDiaAve + BPSysAve + Alc +
    SleepHrsNight*Age+SleepHrsNight*Gender+SleepHrsNight*factor(Race1),
  df3
)
p41.log = ols_plot_resid_lev(m_full.log)
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 Y")
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)

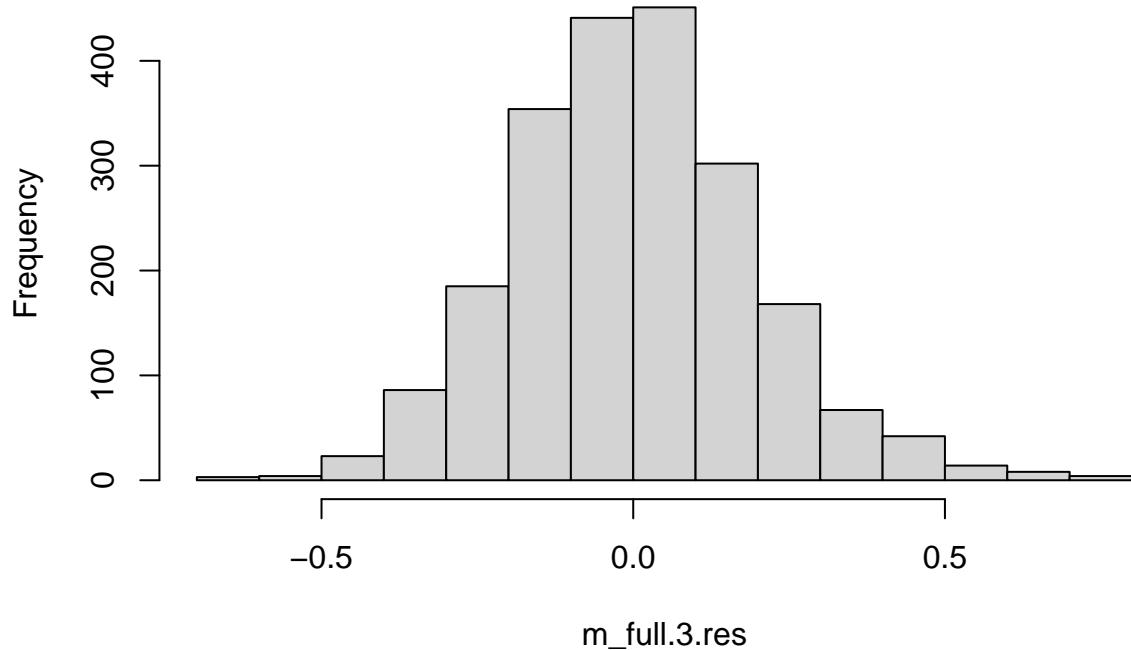
par(mfrow = c(1, 1))

```

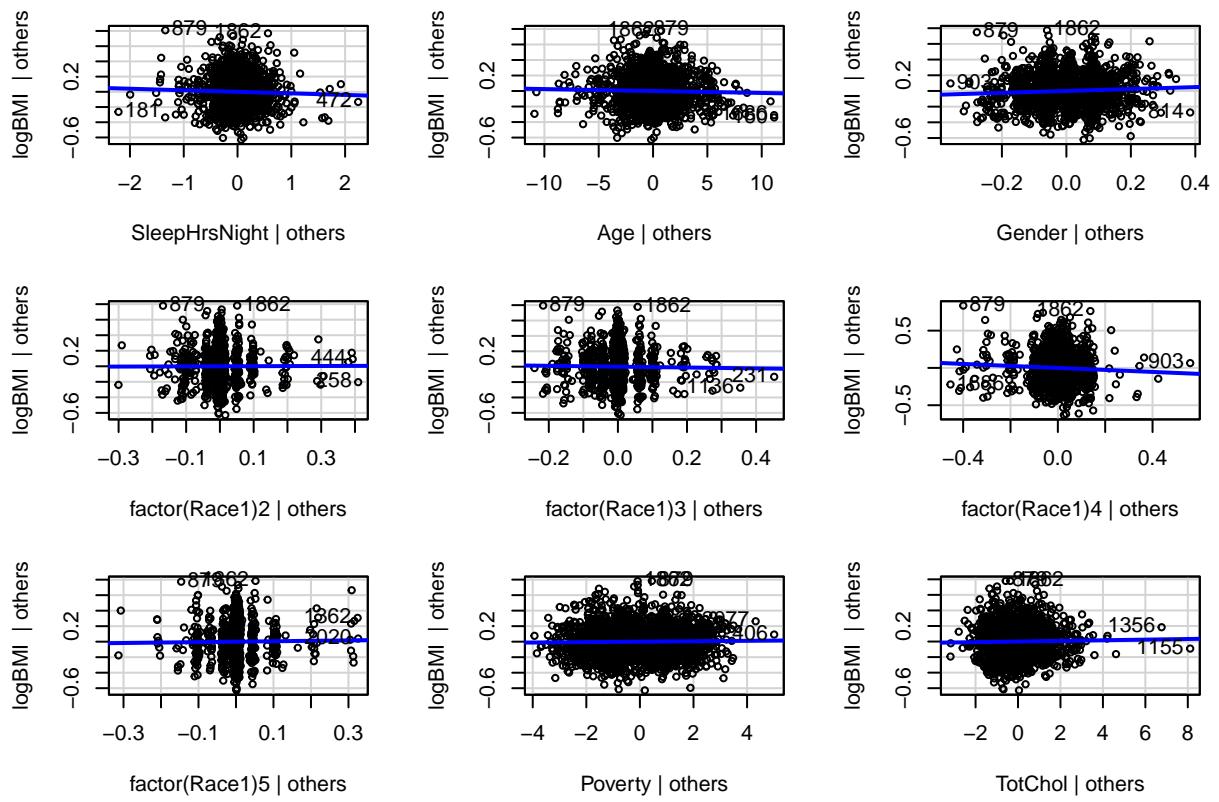


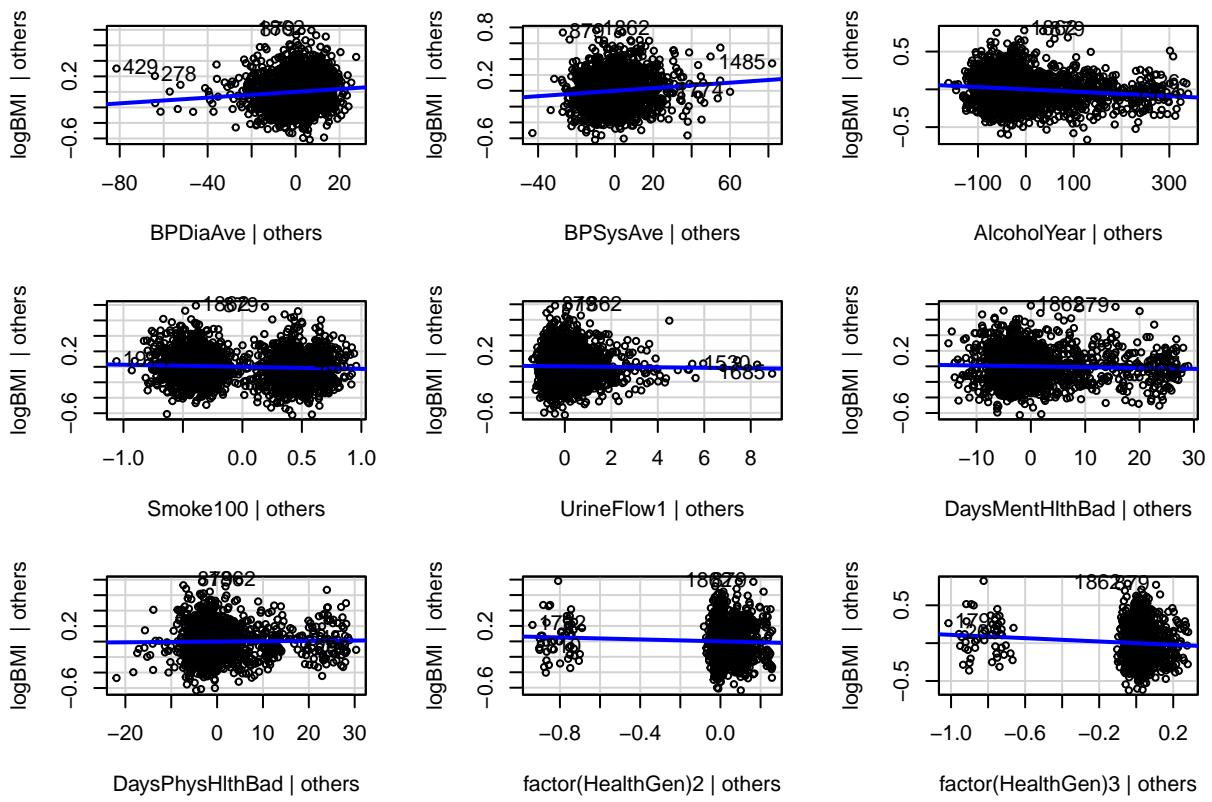
```
hist(m_full.3.res, breaks = 15)
```

Histogram of m_full.3.res

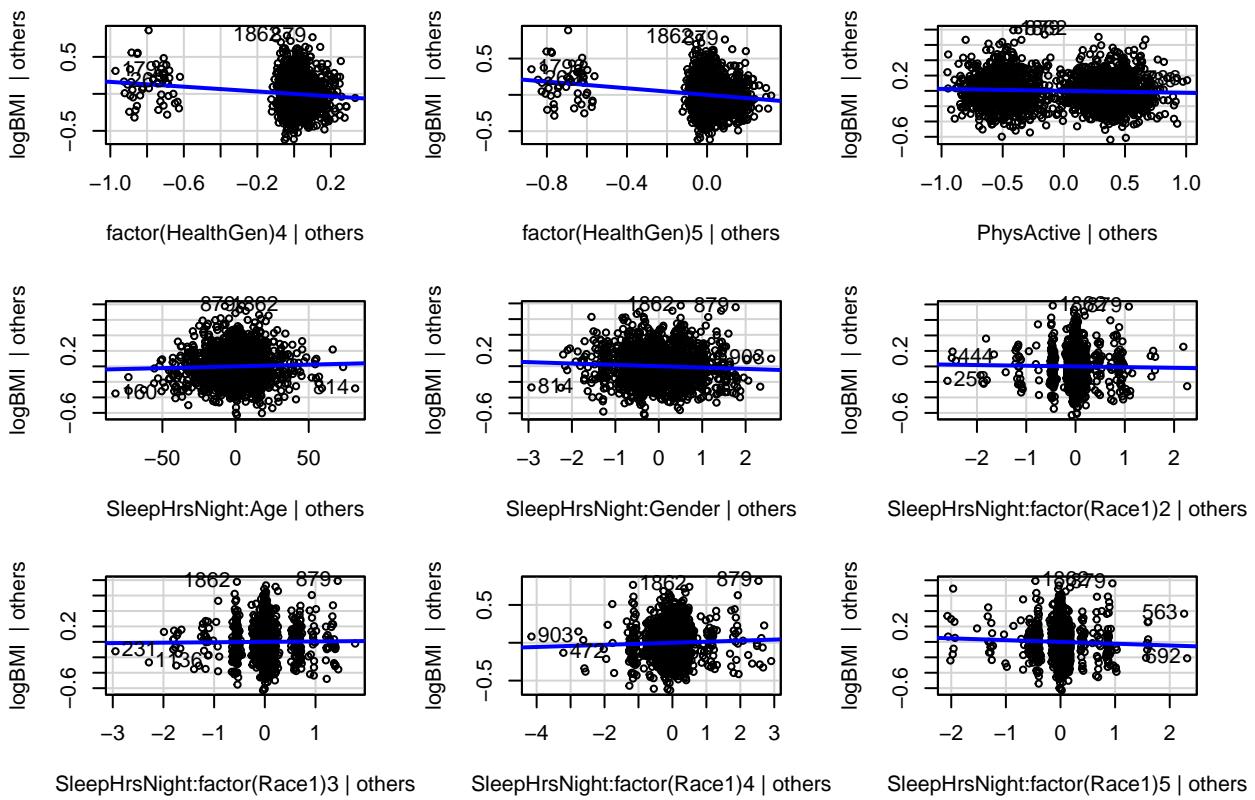


```
car::avPlots(m_full.log)
```





Added-Variable Plots



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

#collinearity diagnostics

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

	GVIF	Df	GVIF^(1/(2*Df))
## SleepHrsNight	1.849122e+01	1	4.300141
## Age	2.879008e+01	1	5.365639
## Gender	2.882194e+01	1	5.368607
## factor(Race1)	5.340628e+05	4	5.199350
## Poverty	1.338183e+00	1	1.156799
## TotChol	1.135148e+00	1	1.065433
## BPDiaAve	1.459656e+00	1	1.208162
## BPSysAve	1.578347e+00	1	1.256323
## AlcoholYear	1.136198e+00	1	1.065926
## Smoke100	1.142055e+00	1	1.068670
## UrineFlow1	1.049202e+00	1	1.024306
## DaysMentHlthBad	1.158938e+00	1	1.076540
## DaysPhysHlthBad	1.256335e+00	1	1.120864
## factor(HealthGen)	1.495098e+00	4	1.051559
## PhysActive	1.177165e+00	1	1.084972
## SleepHrsNight:Age	3.875954e+01	1	6.225716
## SleepHrsNight:Gender	3.046426e+01	1	5.519444
## SleepHrsNight:factor(Race1)	6.995071e+05	4	5.377730

```

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

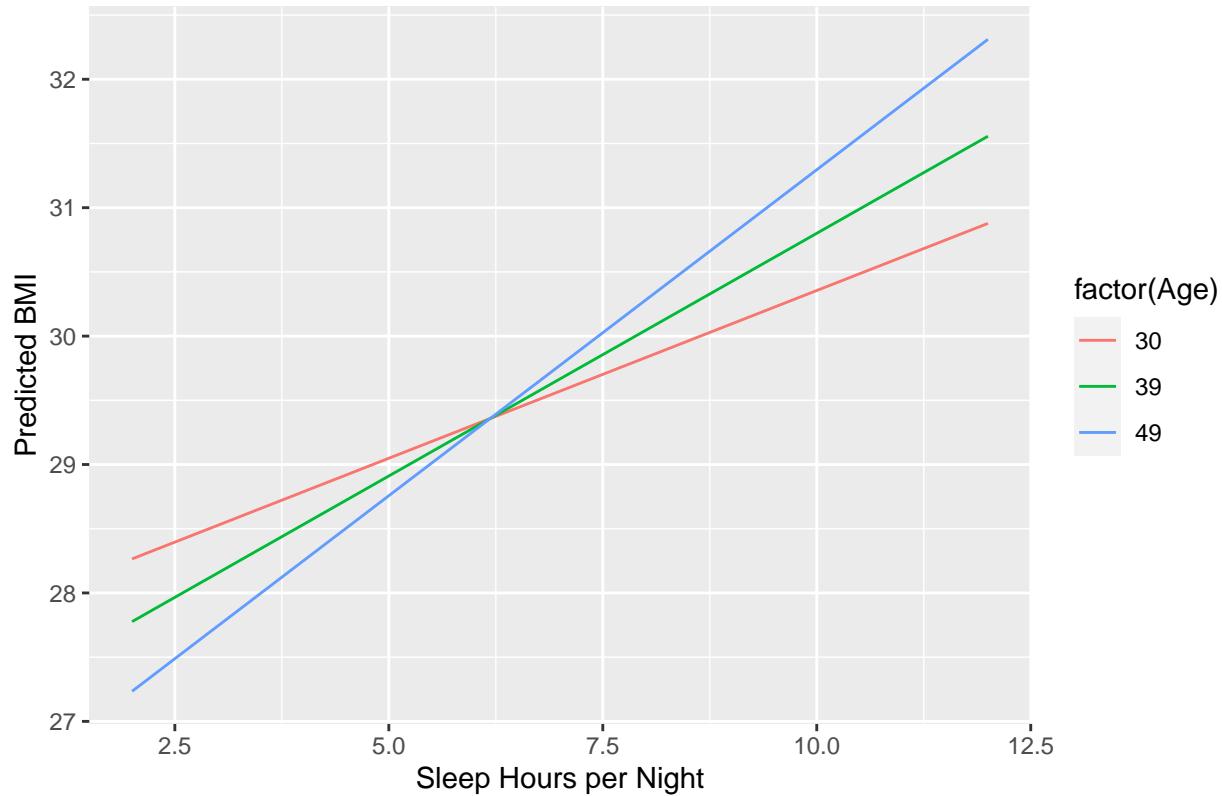
getMode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}

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")

```

Interaction between Sleep Hours and Age on 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
```