Model3

Liancheng

2023-11-25

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

```
##          used (Mb) gc trigger (Mb) max used (Mb)
## Ncells 471047 25.2    1015418 54.3   660860 35.3
## Vcells 887670  6.8    8388608 64.0  1800812 13.8
```

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

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

```
library(olsrr)
```

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

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

```
library(ggplot2)
library(lmtest)
```

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

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

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

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

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

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

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

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

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

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

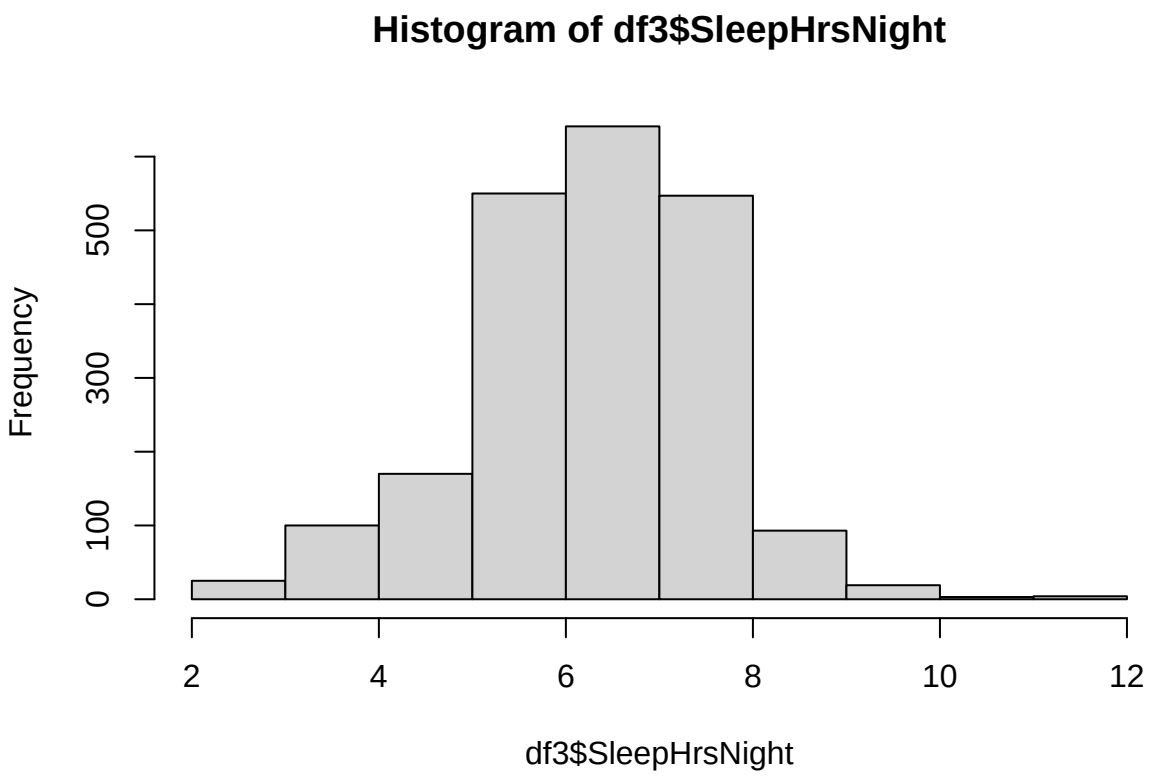
```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

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

df3 <- na.omit(df2)
#df3$SleepHrsNight <- df3$SleepHrsNight * 60
#df3 <- df3[, -which(names(df3) %in% "SleepHrsNight")]
```

```
# cor(df3$BPSysAve,df3$BPDiaAve)
psych::describe(df3)
```

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

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




```
# colSums(is.na(df2)) / nrow(df2)
fit0 <-
  lm(SleepHrsNight ~ .,
     data = df3)
#data type
df3$Gender <- ifelse(df3$Gender == "male", 0, 1)
df3$Smoke100 <- ifelse(df3$Smoke100 == "No", 0, 1)
df3$PhysActive <- ifelse(df3$PhysActive == "No", 0, 1)
df3 <- df3 %>%
  mutate(
    Race1 = case_when(
      Race1 == 'Black' ~ 1,
      Race1 == 'Hispanic' ~ 2,
      Race1 == 'Mexican' ~ 3,
      Race1 == 'White' ~ 4,
      Race1 == 'Other' ~ 5,
      TRUE ~ NA_integer_  # Default value if none of the conditions are met
    )
  )

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_3 add additional risk factors ##
df3$logBMI = log(df3$BMI + 1)
m_3 = lm(
  logBMI ~ SleepHrsNight + Age + Gender + factor(Race1)  + Poverty + TotChol + BPDiaAve + BPSysAve + Al
    DaysPhysHlthBad + factor(HealthGen) + PhysActive,
  df3
)
summary(m_3)
```

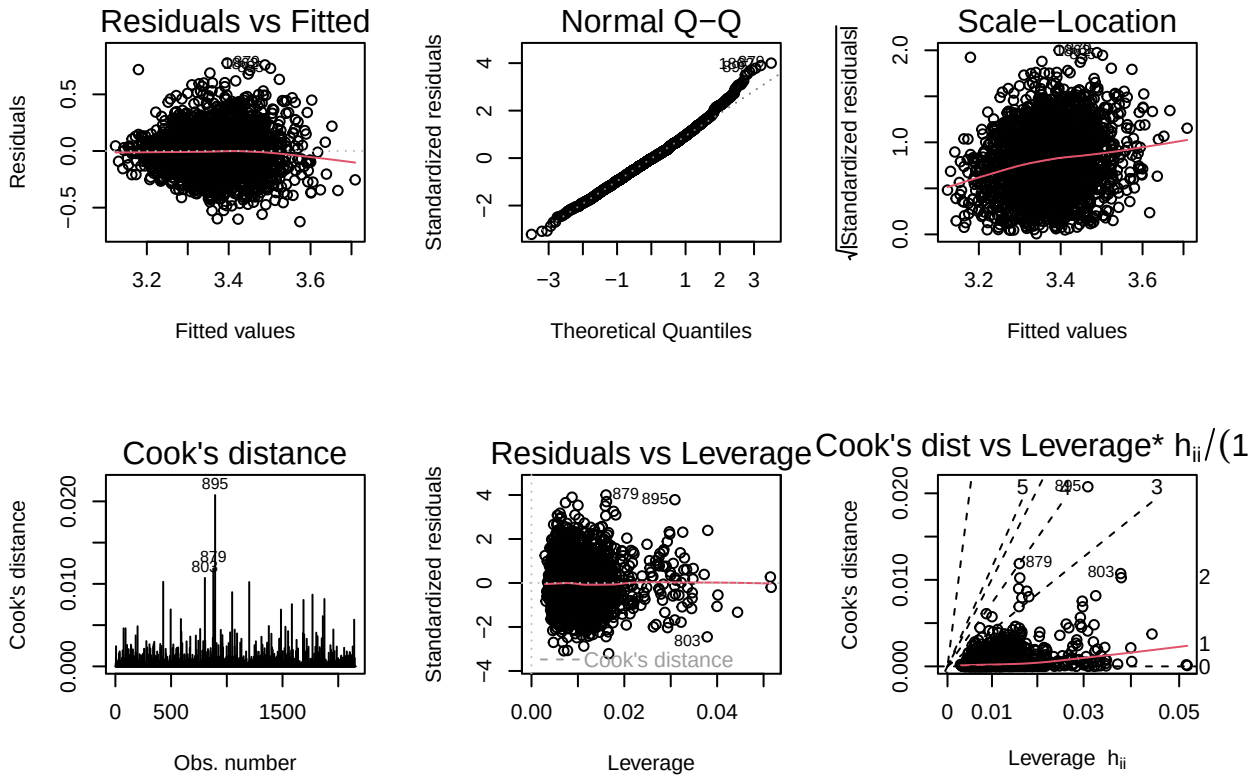
```
##
## Call:
## lm(formula = logBMI ~ SleepHrsNight + Age + Gender + factor(Race1) +
##     Poverty + TotChol + BPDiaAve + BPSysAve + AlcoholYear + Smoke100 +
##     UrineFlow1 + DaysMentHlthBad + DaysPhysHlthBad + factor(HealthGen) +
##     PhysActive, data = df3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.62390 -0.12801 -0.00572  0.12171  0.77693
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          3.248e+00  5.935e-02  54.726  < 2e-16 ***
## SleepHrsNight       -5.122e-03  3.337e-03  -1.535 0.124883
## Age                  5.418e-04  4.310e-04   1.257 0.208812
## Gender               4.485e-03  9.027e-03   0.497 0.619349
## factor(Race1)2      -4.744e-02  2.012e-02  -2.358 0.018464 *
## factor(Race1)3      -2.055e-02  1.764e-02  -1.165 0.244146
## factor(Race1)4      -4.026e-02  1.322e-02  -3.045 0.002353 **
## factor(Race1)5      -1.032e-01  1.981e-02  -5.210 2.07e-07 ***
## Poverty              2.921e-03  2.879e-03   1.015 0.310313
## TotChol              4.530e-03  4.266e-03   1.062 0.288444
## BPDiaAve             1.859e-03  4.303e-04   4.320 1.63e-05 ***
## BPSysAve             1.661e-03  3.708e-04   4.480 7.84e-06 ***
## AlcoholYear         -2.871e-04  4.756e-05  -6.037 1.85e-09 ***
## Smoke100            -2.785e-02  9.040e-03  -3.081 0.002089 **
## UrineFlow1          -3.862e-03  4.469e-03  -0.864 0.387574
## DaysMentHlthBad     -1.083e-03  5.657e-04  -1.915 0.055637 .
## DaysPhysHlthBad      4.258e-04  6.574e-04   0.648 0.517225
## factor(HealthGen)2 -6.631e-02  3.142e-02  -2.110 0.034956 *
## factor(HealthGen)3 -1.117e-01  3.114e-02  -3.586 0.000343 ***
## factor(HealthGen)4 -1.642e-01  3.197e-02  -5.136 3.06e-07 ***
## factor(HealthGen)5 -2.285e-01  3.379e-02  -6.763 1.74e-11 ***
## PhysActive          -2.431e-02  9.231e-03  -2.633 0.008519 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1958 on 2130 degrees of freedom
## Multiple R-squared:  0.1572, Adjusted R-squared:  0.1489
```

```
## F-statistic: 18.92 on 21 and 2130 DF,  p-value: < 2.2e-16
```

```
car::Anova(m_3, type = "III")
```

```
## Anova Table (Type III tests)
##
## Response: logBMI
##                      Sum Sq   Df   F value    Pr(>F)
## (Intercept)       114.784    1 2994.9268 < 2.2e-16 ***
## SleepHrsNight       0.090    1    2.3568  0.124883
## Age                 0.061    1    1.5806  0.208812
## Gender              0.009    1    0.2469  0.619349
## factor(Race1)       1.117    4    7.2873 7.945e-06 ***
## Poverty             0.039    1    1.0298  0.310313
## TotChol             0.043    1    1.1274  0.288444
## BPDiaAve            0.715    1   18.6611 1.633e-05 ***
## BPSysAve            0.769    1   20.0748 7.842e-06 ***
## AlcoholYear         1.397    1   36.4430 1.849e-09 ***
## Smoke100            0.364    1    9.4927  0.002089 **
## UrineFlow1          0.029    1    0.7468  0.387574
## DaysMentHlthBad     0.141    1    3.6669  0.055637 .
## DaysPhysHlthBad     0.016    1    0.4196  0.517225
## factor(HealthGen)   4.255    4   27.7539 < 2.2e-16 ***
## PhysActive          0.266    1    6.9338  0.008519 **
## Residuals          81.634 2130
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
############ model 3 diagnosis ############
par(mfrow = c(2, 3)) #read more from ?plot.lm
plot(m_3, which = 1)
plot(m_3, which = 2)
plot(m_3, which = 3)
plot(m_3, which = 4)
plot(m_3, which = 5)
plot(m_3, which = 6)
```




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

m_3.yhat = m_3$fitted.values
m_3.res = m_3$residuals
m_3.h = hatvalues(m_3)
m_3.r = rstandard(m_3)
m_3.rr = rstudent(m_3)
#which subject is most outlying with respect to the x space
Hmisc::describe(m_3.h)
```

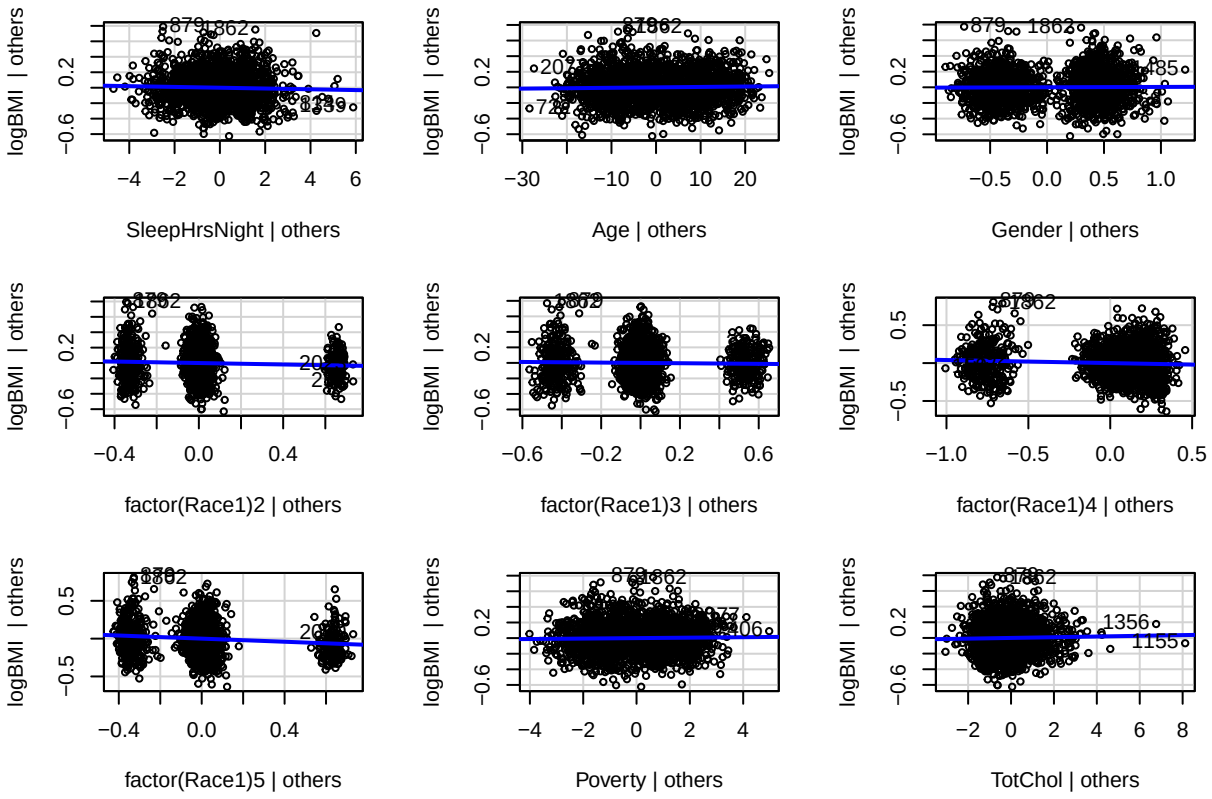
```
## m_3.h
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2152        0     2152        1  0.01022 0.005352 0.004645 0.005142
##      .25      .50      .75      .90      .95
## 0.006490 0.009118 0.012287 0.015996 0.019196
##
## lowest : 0.002934252 0.003128726 0.003292556 0.003347562 0.003402047
## highest: 0.039975211 0.040170575 0.044397994 0.051471058 0.051640631
```

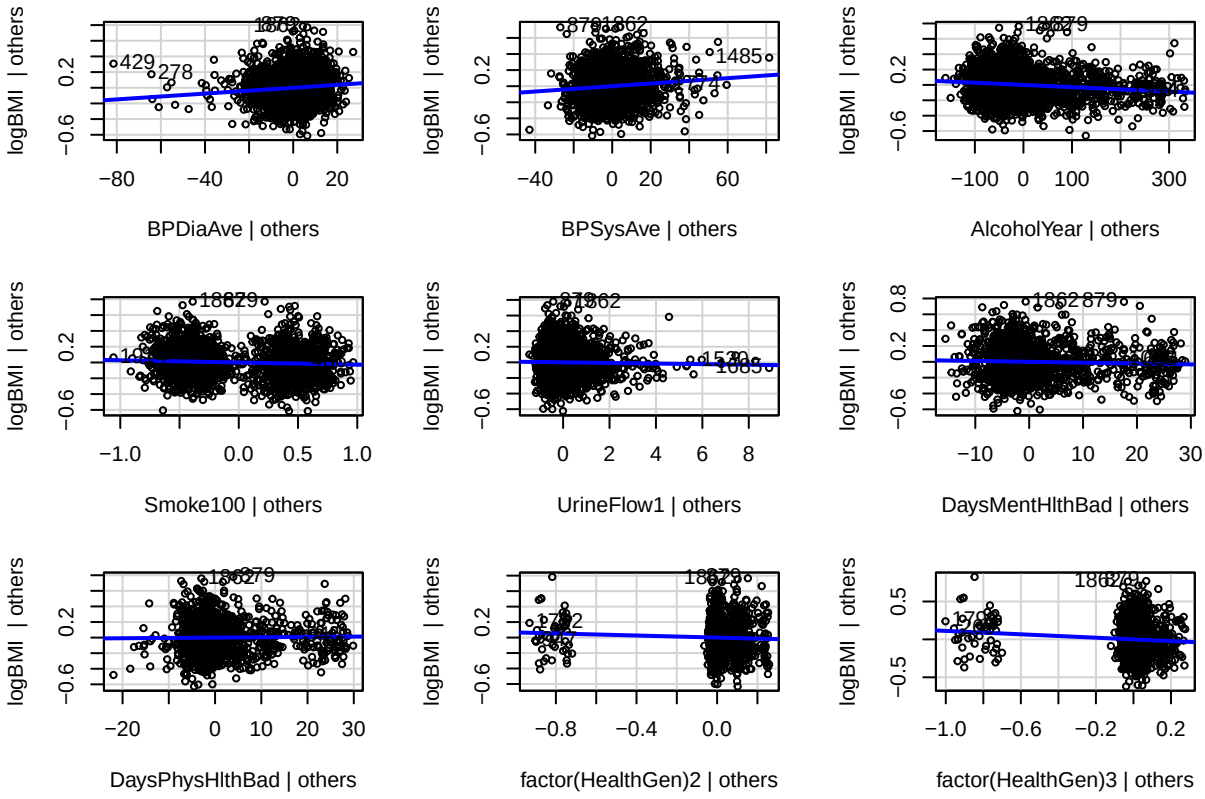
```
m_3.h[which.max(m_3.h)]
```

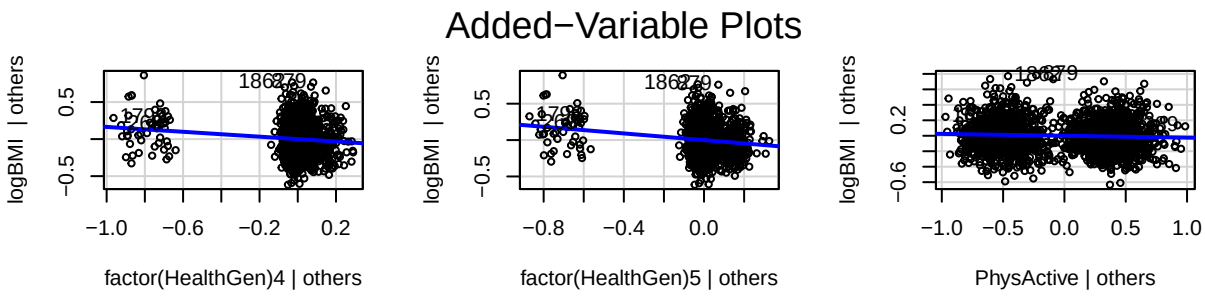
```
##       1685
## 0.05164063
```

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

#(1)Linear: 2 approaches
```

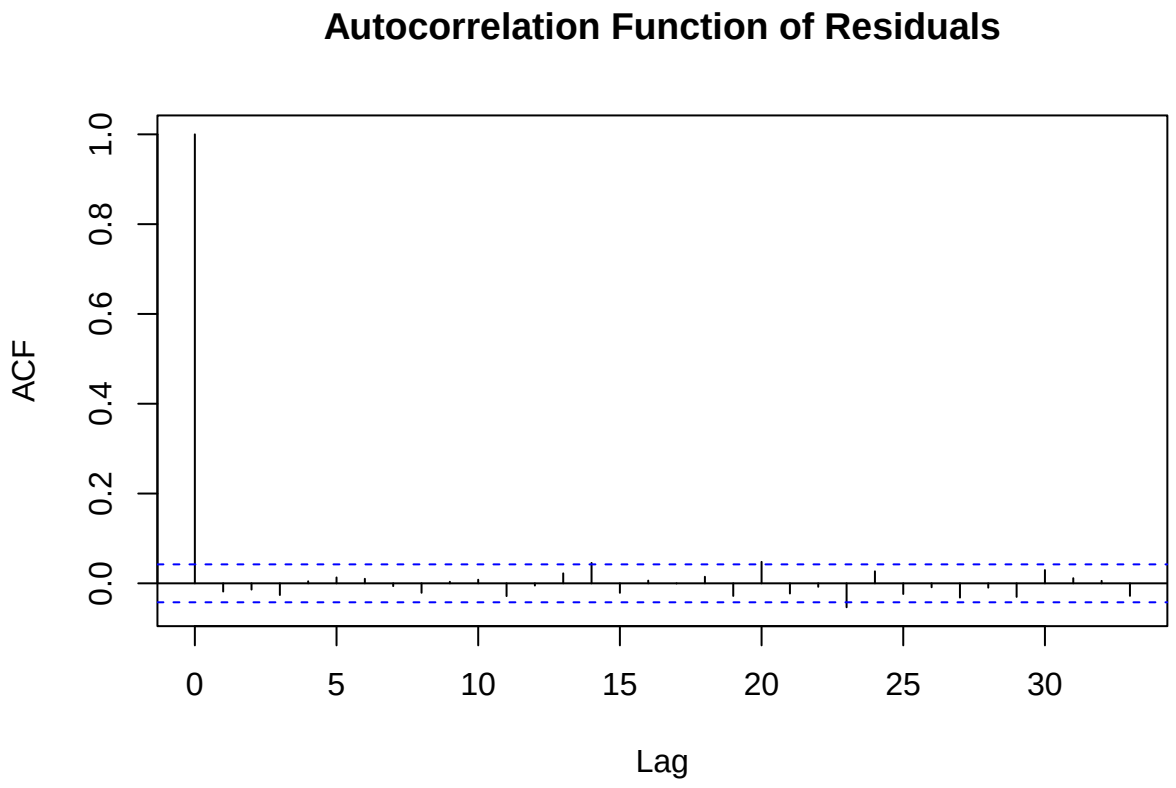
```
# partial regression plots
car::avPlots(m_3)
```



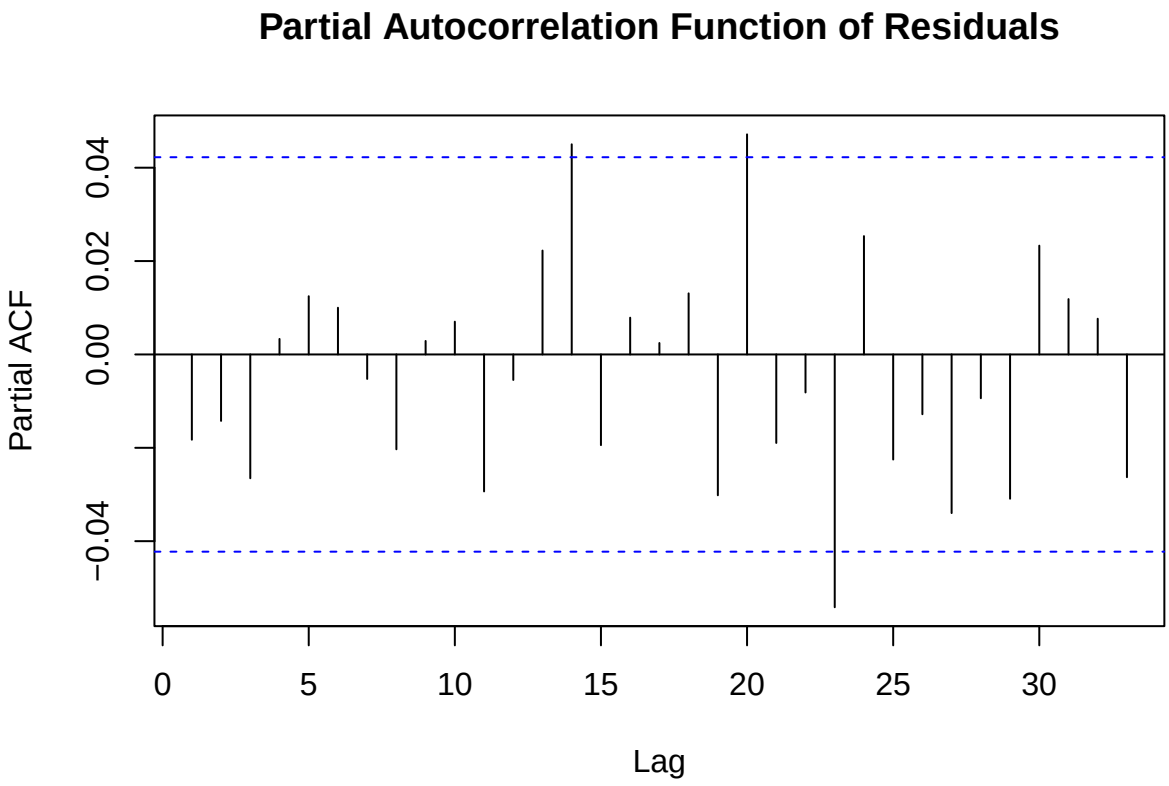

logBMI | others
BPDiaAve | others
BPSysAve | others
AlcoholYear | others
Smoke100 | others
UrineFlow1 | others
DaysMentHlthBad | others
DaysPhysHlthBad | others
factor(HealthGen)2 | others
factor(HealthGen)3 | others




```r
#(2)Independence:

residuals <- resid(m_3)
acf(residuals, main = "Autocorrelation Function of Residuals")
```



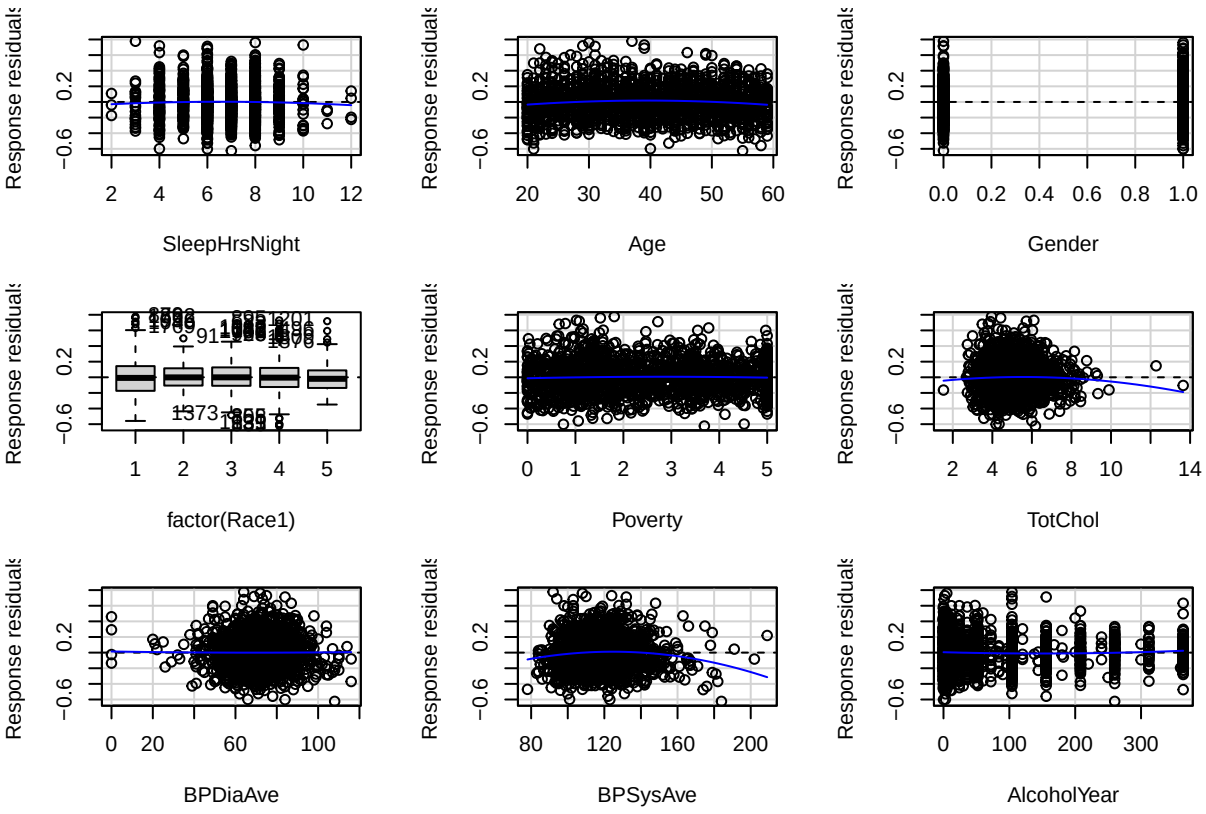

```
pacf(residuals, main = "Partial Autocorrelation Function of Residuals")
```

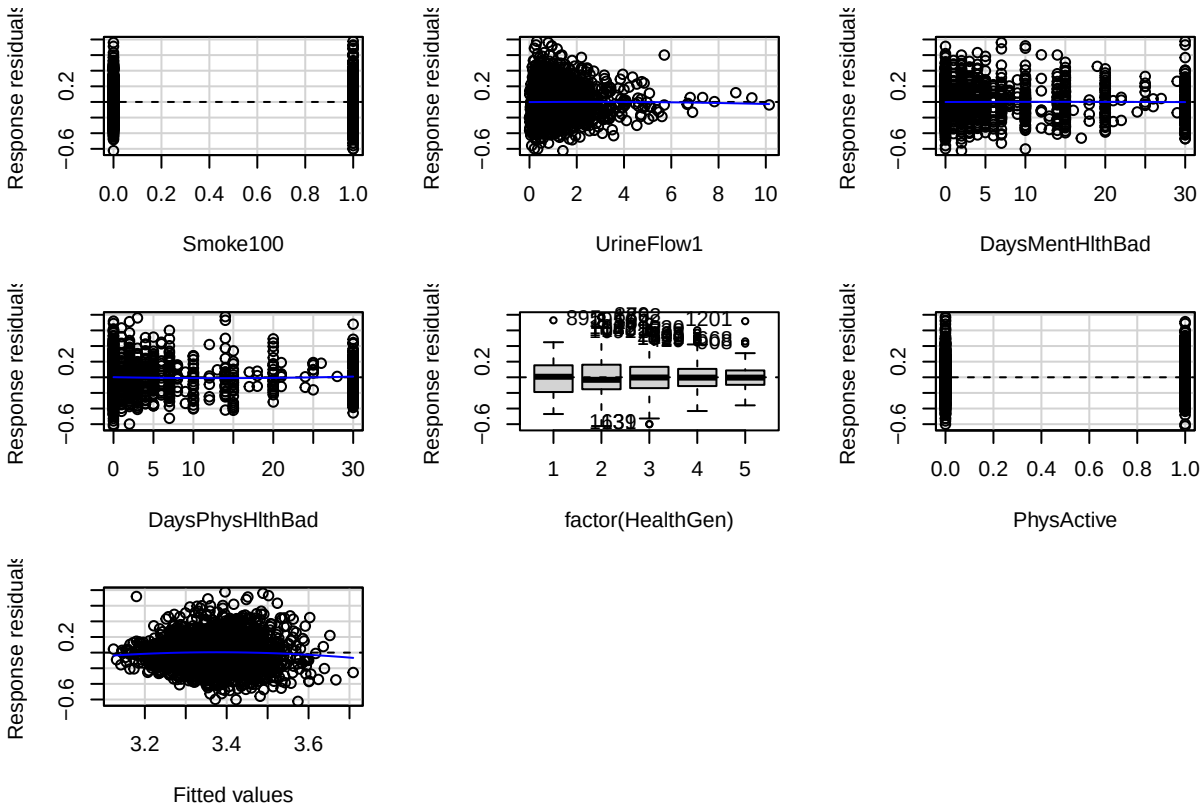



```
dw_test <- dwtest(m_3)
print(dw_test)
```

```
##
##  Durbin-Watson test
##
## data:  m_3
## DW = 2.0365, p-value = 0.8013
## alternative hypothesis: true autocorrelation is greater than 0
```

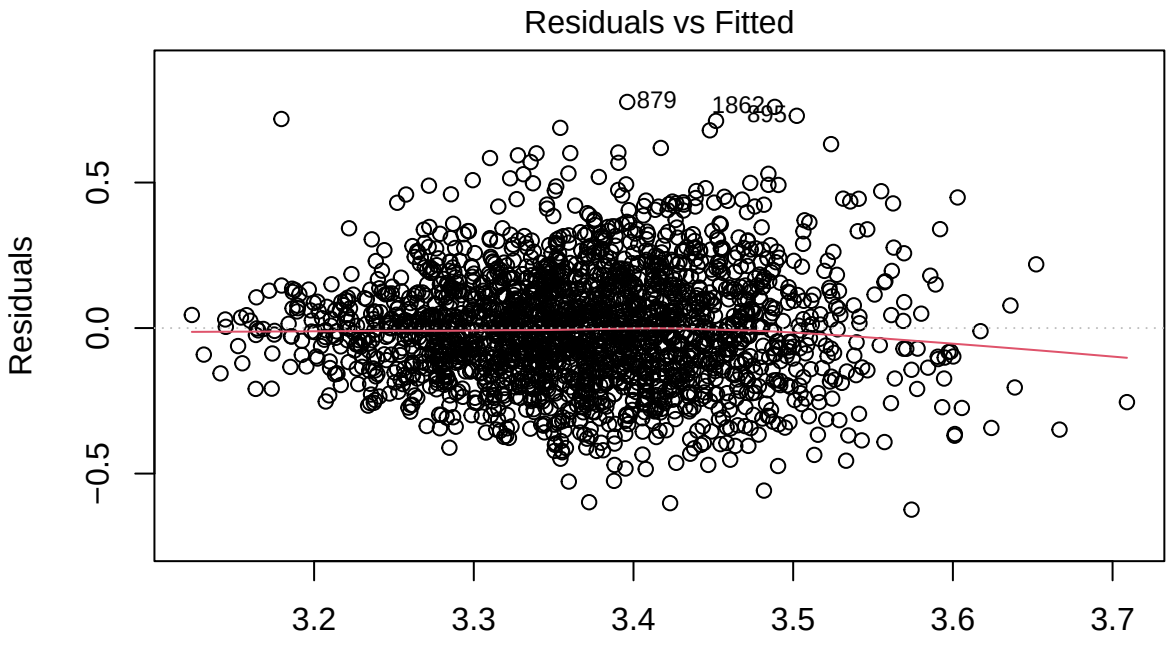
```
#(3)E: constant var: residuals-fitted values; transform for variance-stable...(total: 4 solutions)

car::residualPlots(m_3, type = "response")
```


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

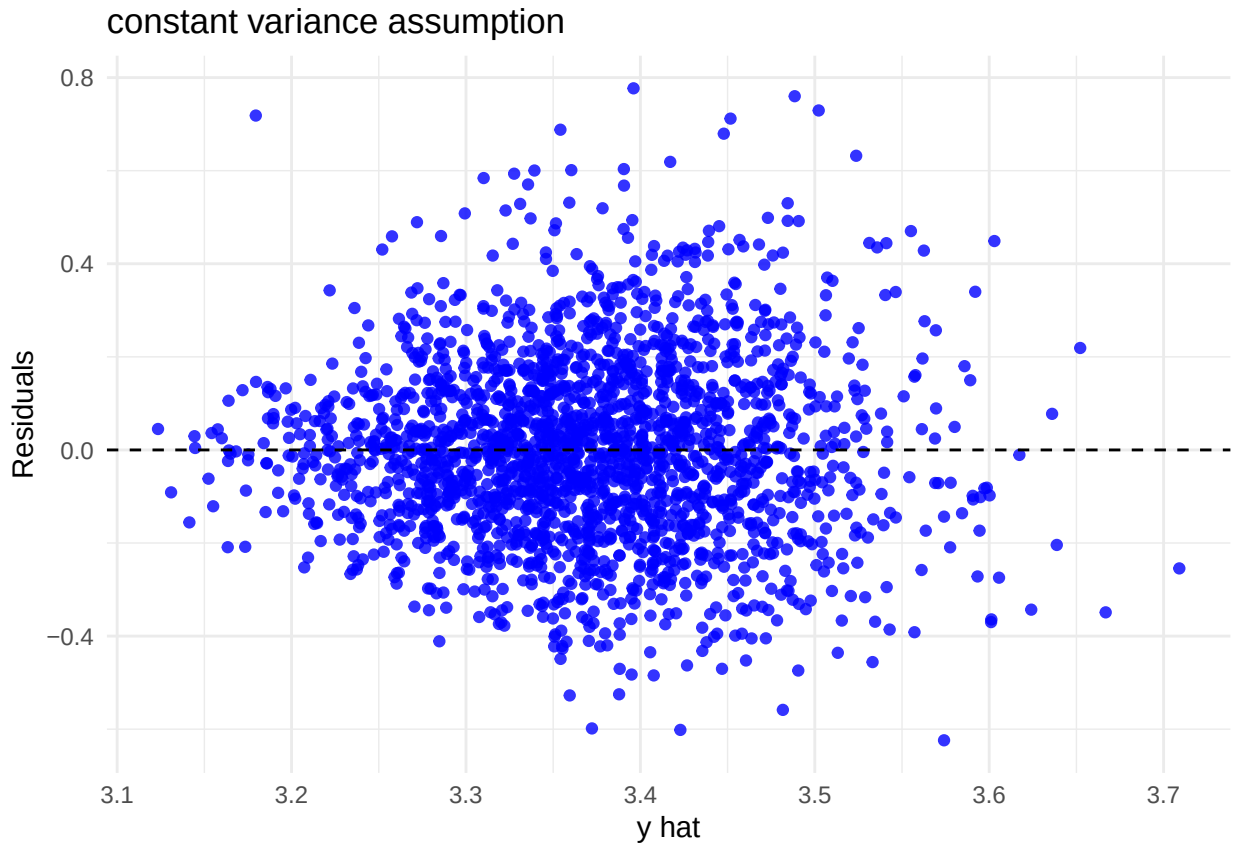



```
##                    Test stat Pr(>|Test stat|)
## SleepHrsNight        -1.0099          0.31265
## Age                  -4.0041        6.439e-05 ***
## Gender                0.2938          0.76896
## factor(Race1)
## Poverty              -1.1806          0.23791
## TotChol              -1.6747          0.09415 .
## BPDiaAve              0.2531          0.80025
## BPSysAve             -4.4884        7.561e-06 ***
## AlcoholYear           1.8373          0.06630 .
## Smoke100             -0.2674          0.78921
## UrineFlow1           -0.3011          0.76337
## DaysMentHlthBad      -0.2026          0.83950
## DaysPhysHlthBad       0.9517          0.34138
## factor(HealthGen)
## PhysActive           -0.6470          0.51773
## Tukey test           -1.9392          0.05248 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(m_3, which = 1)
```




```
#or
ggplot(m_3, aes(x = m_3.yhat, y = m_3.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()
```



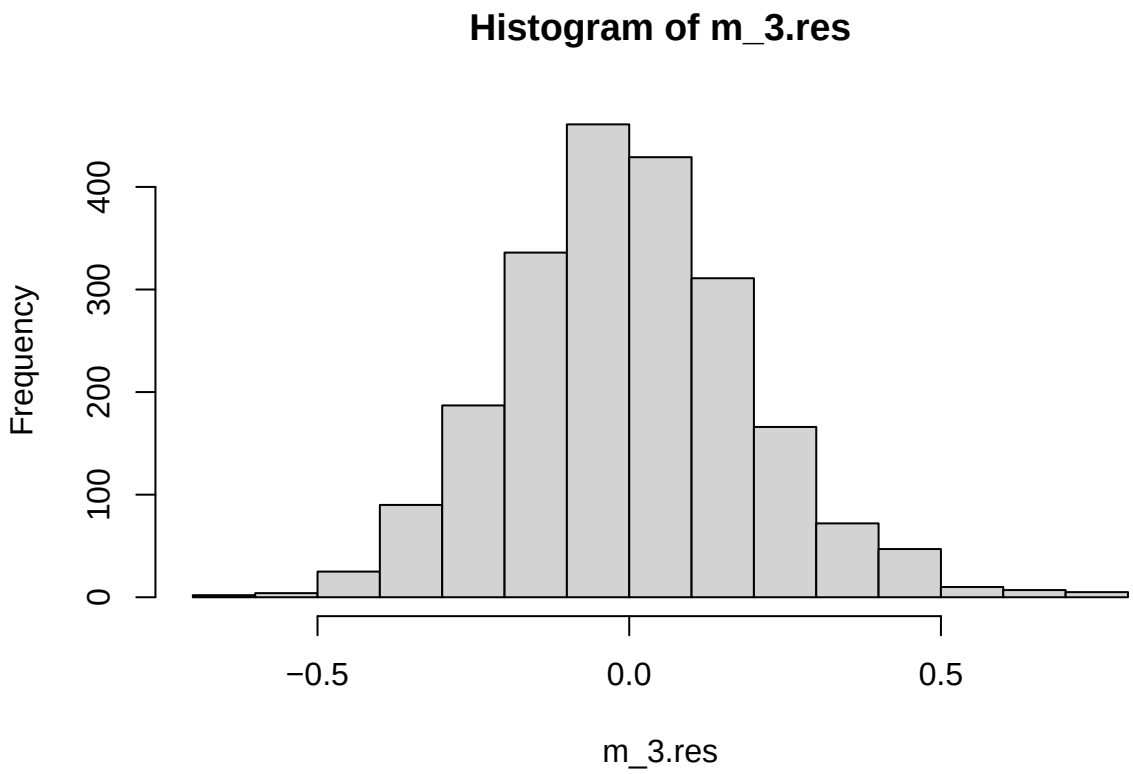

```
#conclusion: the constant variance assumption is basically not violated. The spread of the residuals ap


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

#exam quartiles of the residuals
Hmisc::describe(m_3.res)
```

```
## m_3.res
##          n    missing   distinct       Info       Mean        Gmd        .05
##       2152          0       2152          1 -1.894e-20     0.2173  -0.310238
##        .10        .25        .50        .75        .90        .95
##  -0.240933  -0.128010  -0.005718   0.121707   0.244771   0.328438
##
## lowest : -0.6238971 -0.6015537 -0.5983818 -0.5585373 -0.5274531
## highest:  0.7118792  0.7184194  0.7293958  0.7600078  0.7769265
```

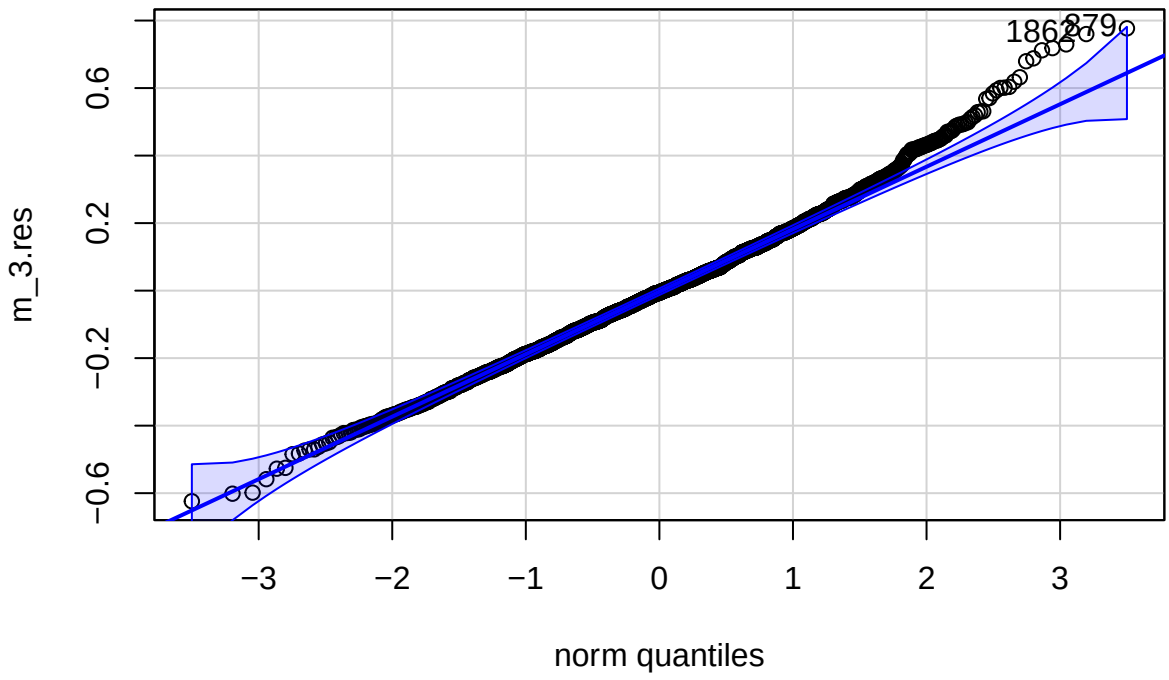
```
Hmisc::describe(m_3.res)$counts[c(".25", ".50", ".75")] #not symmetric
```

```
##         .25         .50         .75
## "-0.128010" "-0.005718" " 0.121707"
```

```
#histogram
par(mfrow = c(1, 1))
hist(m_3.res, breaks = 15)
```



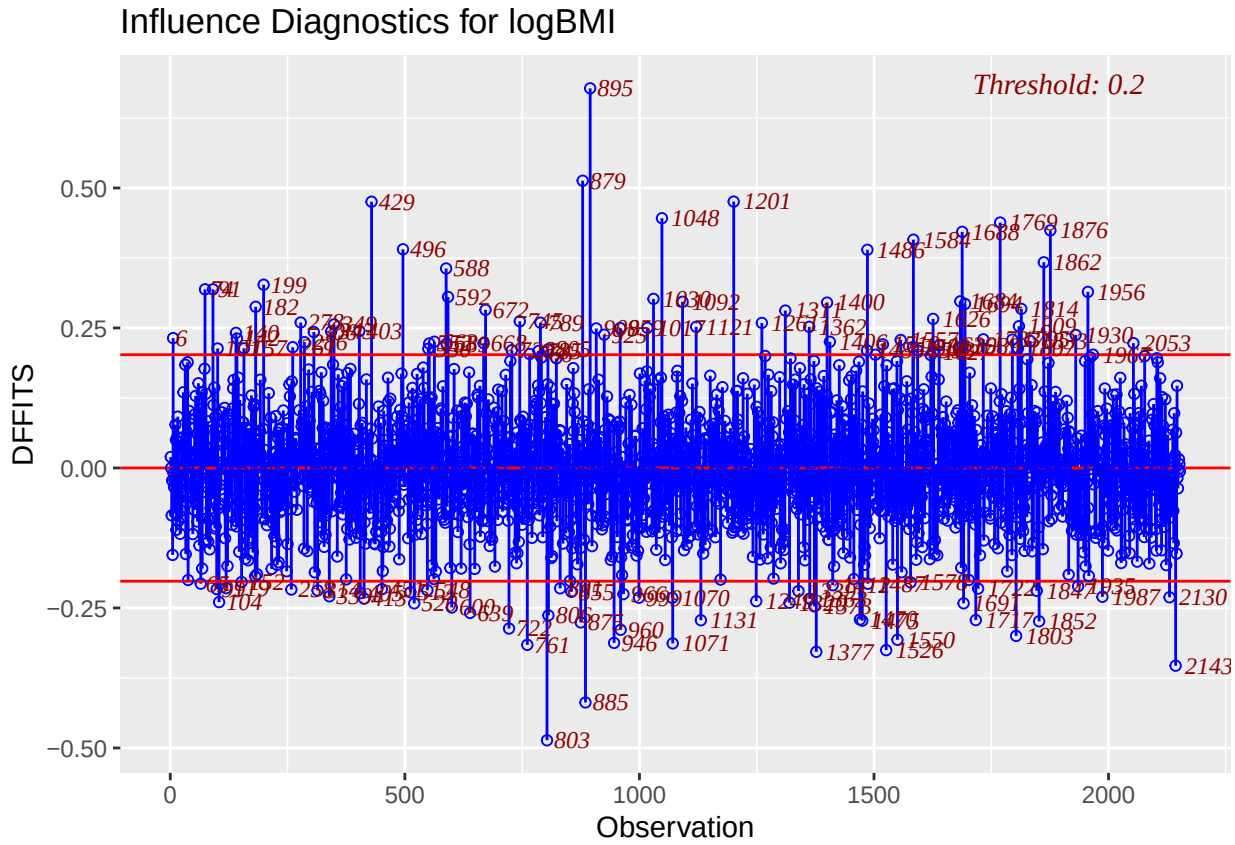

```
# Q-Q plot
qq.m_3.res = car::qqPlot(m_3.res)
```




```
m_3.res[qq.m_3.res]
```

```
##       879      1862
## 0.7769265 0.7600078
```

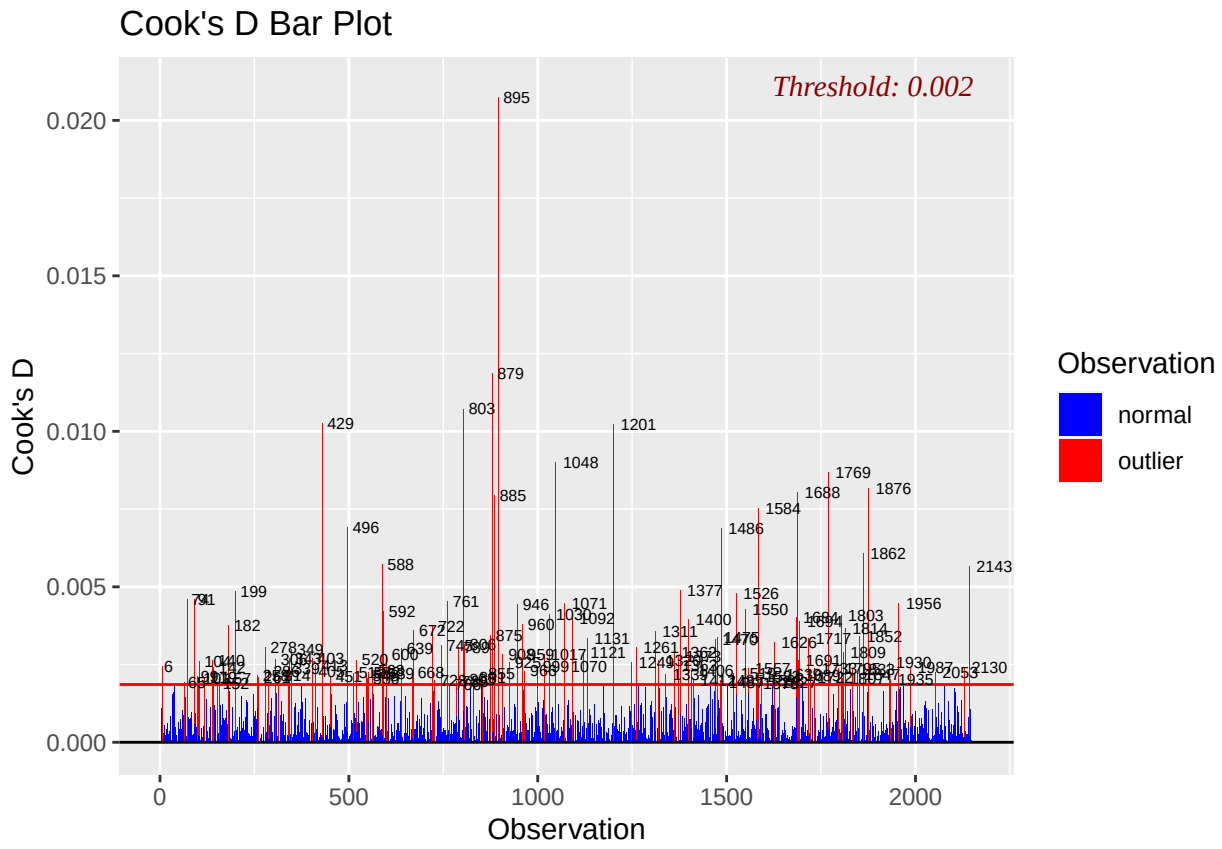
```
################ influential observations  ##################

influence3 = data.frame(
  Residual = resid(m_3),
  Rstudent = rstudent(m_3),
  HatDiagH = hat(model.matrix(m_3)),
  CovRatio = covratio(m_3),
  DFFITS = dffits(m_3),
  COOKsDistance = cooks.distance(m_3)
)
# DFFITS
ols_plot_dffits(m_3)
```




```
influence3[order(abs(influence3$DFFITS), decreasing = T), ] %>% head()
```

```
##        Residual  Rstudent   HatDiagH  CovRatio     DFFITS COOKsDistance
## 895   0.7293958  3.796597 0.03089695 0.8987673  0.6779027   0.020758000
## 879   0.7769265  4.014987 0.01605793 0.8698536  0.5129135   0.011873905
## 803  -0.4702638 -2.451802 0.03785850 0.9869798 -0.4863485   0.010726351
## 1201  0.7184194  3.710796 0.01616157 0.8911795  0.4756056   0.010220573
## 429   0.4589148  2.392619 0.03797756 0.9900227  0.4753839   0.010249532
## 1048  0.6794591  3.507899 0.01590422 0.9044501  0.4459484   0.008991817
```

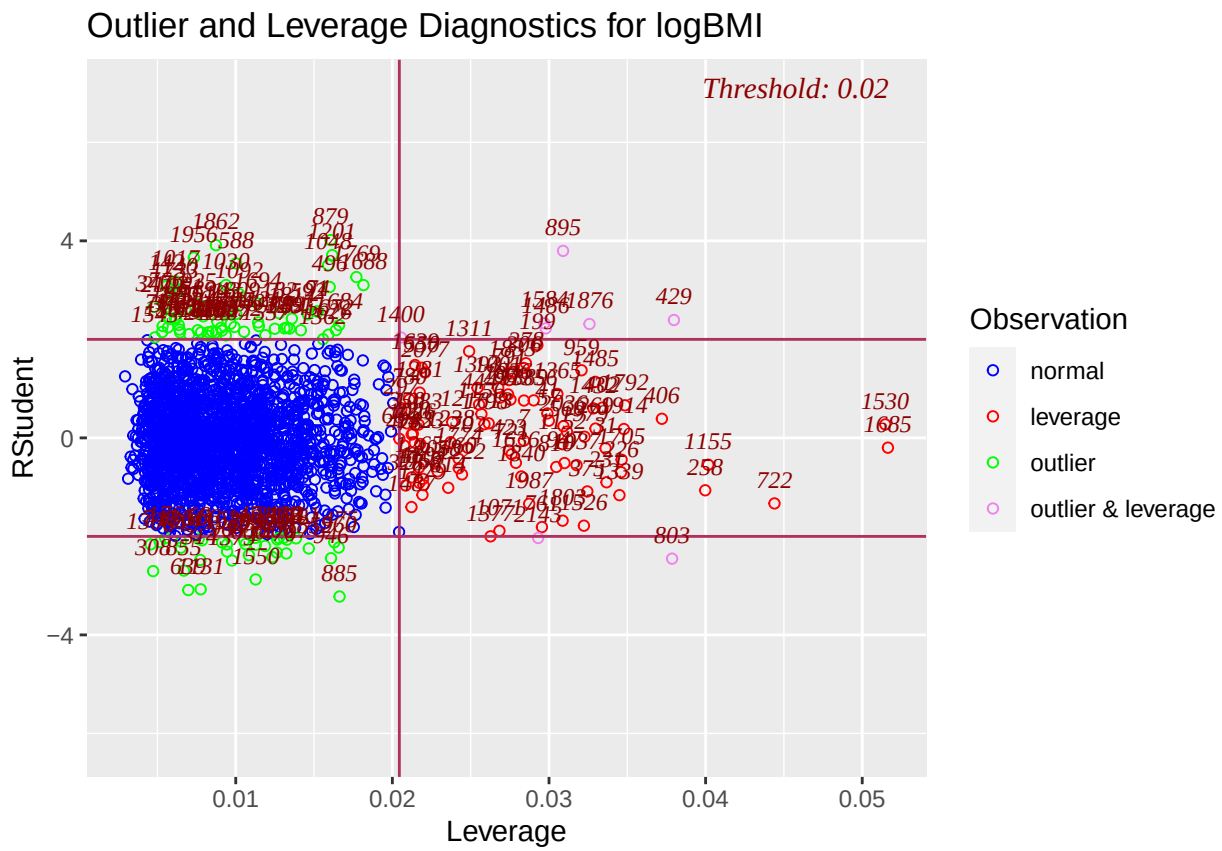
```
#From the plot above, we can see 2 observations with the largest (magnitude) of DFFITS, observation 879

# Cook's D
ols_plot_cooksd_bar(m_3)
```




```
influence3[order(influence3$COOKsDistance, decreasing = T), ] %>% head()
```

```
##        Residual  Rstudent   HatDiagH  CovRatio     DFFITS COOKsDistance
## 895   0.7293958  3.796597 0.03089695 0.8987673  0.6779027   0.020758000
## 879   0.7769265  4.014987 0.01605793 0.8698536  0.5129135   0.011873905
## 803  -0.4702638 -2.451802 0.03785850 0.9869798 -0.4863485   0.010726351
## 429   0.4589148  2.392619 0.03797756 0.9900227  0.4753839   0.010249532
## 1201  0.7184194  3.710796 0.01616157 0.8911795  0.4756056   0.010220573
## 1048  0.6794591  3.507899 0.01590422 0.9044501  0.4459484   0.008991817
```

```
#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_3)
```




```
#high leverage
influence3[order(influence3$HatDiagH, decreasing = T), ] %>% head()
```

```
##         Residual   Rstudent   HatDiagH  CovRatio      DFFITS COOKsDistance
## 1685 -0.03789727 -0.1987360 0.05164063 1.0649679 -0.04637520  9.780132e-05
## 1530  0.05217679  0.2735966 0.05147106 1.0643886  0.06373338  1.847140e-04
## 722  -0.25458400 -1.3305282 0.04439799 1.0381694 -0.28679213  3.737272e-03
## 1155 -0.10460200 -0.5452859 0.04017058 1.0494418 -0.11155301  5.658264e-04
## 258  -0.20412667 -1.0642043 0.03997521 1.0402149 -0.21715967  2.143427e-03
## 429   0.45891480  2.3926191 0.03797756 0.9900227  0.47538393  1.024953e-02
```

```
#high studentized residual
influence3[order(influence3$Rstudent, decreasing = T), ] %>% head()
```

```
##       Residual Rstudent    HatDiagH  CovRatio    DFFITS COOKsDistance
## 879  0.7769265 4.014987 0.016057928 0.8698536 0.5129135   0.011873905
## 1862 0.7600078 3.912277 0.008734615 0.8706688 0.3672456   0.006089523
## 895  0.7293958 3.796597 0.030896951 0.8987673 0.6779027   0.020758000
## 1201 0.7184194 3.710796 0.016161565 0.8911795 0.4756056   0.010220573
## 1956 0.7118792 3.660293 0.007324108 0.8866282 0.3144050   0.004467203
## 588  0.6880235 3.541735 0.010016754 0.8968708 0.3562588   0.005738008
```

```
#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

rm3.df3 = df3[-c(879, 1769, 1155, 1048, 1769, 1684, 74, 72, 1689, 1311), ]
rm.m_3 =  lm(
  logBMI ~ SleepHrsNight + Age + Gender + Race1  + Poverty + TotChol + BPDiaAve + BPSysAve + AlcoholYear
    DaysPhysHlthBad + HealthGen + PhysActive,
  rm3.df3
)
## Before removing these observations, the estimated coefficients are:
summary(m_3)$coef
```

```
##                         Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept)         3.2481448996 0.0593529475 54.7259241 0.000000e+00
## SleepHrsNight      -0.0051222011 0.0033365042 -1.5351999 1.248832e-01
## Age                 0.0005418351 0.0004309785  1.2572208 2.088115e-01
## Gender              0.0044850562 0.0090270667  0.4968454 6.193494e-01
## factor(Race1)2     -0.0474441662 0.0201206036 -2.3579892 1.846416e-02
## factor(Race1)3     -0.0205506376 0.0176399003 -1.1650087 2.441459e-01
## factor(Race1)4     -0.0402625452 0.0132211291 -3.0453182 2.352819e-03
## factor(Race1)5     -0.1032294376 0.0198118738 -5.2104833 2.065048e-07
## Poverty             0.0029213342 0.0028787102  1.0148066 3.103133e-01
## TotChol             0.0045298833 0.0042662050  1.0618063 2.884440e-01
## BPDiaAve            0.0018586878 0.0004302663  4.3198546 1.633047e-05
## BPSysAve            0.0016612099 0.0003707653  4.4804897 7.841655e-06
## AlcoholYear        -0.0002871135 0.0000475605 -6.0368055 1.849337e-09
## Smoke100           -0.0278531968 0.0090402593 -3.0810175 2.089327e-03
## UrineFlow1         -0.0038619406 0.0044687930 -0.8642022 3.875742e-01
## DaysMentHlthBad    -0.0010833643 0.0005657512 -1.9149128 5.563743e-02
## DaysPhysHlthBad     0.0004258120 0.0006573843  0.6477368 5.172250e-01
## factor(HealthGen)2 -0.0663067752 0.0314217059 -2.1102220 3.495556e-02
## factor(HealthGen)3 -0.1116829970 0.0311448128 -3.5859261 3.434475e-04
## factor(HealthGen)4 -0.1641755328 0.0319650262 -5.1360988 3.060189e-07
## factor(HealthGen)5 -0.2284914176 0.0337861205 -6.7628782 1.741484e-11
## PhysActive         -0.0243066776 0.0092308251 -2.6332075 8.519429e-03
```

```
## After removing these observations, the estimated coefficients are:
summary(rm.m_3)$coef
```

```
##                          Estimate   Std. Error      t value     Pr(>|t|)
## (Intercept)       3.271586e+00 5.251846e-02  62.29403017 0.000000e+00
## SleepHrsNight    -3.667352e-03 3.305003e-03  -1.10963653 2.672811e-01
## Age               6.764573e-04 4.257752e-04   1.58876640 1.122618e-01
## Gender            2.198340e-03 8.913972e-03   0.24661728 8.052282e-01
## Race1            -1.217641e-02 3.798689e-03  -3.20542485 1.368623e-03
## Poverty           3.219538e-03 2.816734e-03   1.14300397 2.531656e-01
## TotChol           5.565808e-03 4.289074e-03   1.29767130 1.945410e-01
## BPDiaAve          1.763934e-03 4.262094e-04   4.13865575 3.629609e-05
## BPSysAve          1.754616e-03 3.687131e-04   4.75875836 2.078818e-06
## AlcoholYear      -3.025731e-04 4.727562e-05  -6.40019185 1.903297e-10
## Smoke100         -2.779701e-02 8.912001e-03  -3.11905315 1.838642e-03
## UrineFlow1       -4.357312e-03 4.423203e-03  -0.98510328 3.246854e-01
## DaysMentHlthBad  -9.328717e-04 5.599056e-04  -1.66612312 9.583614e-02
## DaysPhysHlthBad  -5.421605e-05 6.403164e-04  -0.08467072 9.325311e-01
## HealthGen        -5.270135e-02 5.084131e-03 -10.36585202 1.350836e-24
```

```
## PhysActive      -2.235083e-02 9.118101e-03  -2.45125954 1.431568e-02
```

```
#### change percent
abs((rm.m_3$coefficients - m_3$coefficients) / (m_3$coefficients) * 100)
```

```
## Warning in rm.m_3$coefficients - m_3$coefficients: longer object length is not
## a multiple of shorter object length
```

```
##        (Intercept)      SleepHrsNight                Age             Gender
##       7.216908e-01       2.840281e+01       2.484559e+01       5.098524e+01
##     factor(Race1)2     factor(Race1)3     factor(Race1)4     factor(Race1)5
##       7.433528e+01       1.156664e+02       1.138238e+02       1.017088e+02
##            Poverty            TotChol          BPDiaAve          BPSysAve
##       3.993784e+01       1.066795e+02       1.595518e+03       3.622975e+02
##        AlcoholYear           Smoke100         UrineFlow1    DaysMentHlthBad
##       2.249139e+02       9.980535e+01       1.264634e+03       1.963095e+03
##    DaysPhysHlthBad factor(HealthGen)2 factor(HealthGen)3 factor(HealthGen)4
##       7.682170e+05       9.446911e+01       1.006057e+02       1.013390e+02
## factor(HealthGen)5         PhysActive
##       9.467095e+01       1.132455e+02
```

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

##################   multicollinearity   ######################
#Pearson correlations
var3 = c(
  "BMI",
  "SleepHrsNight",
  "Age",
  "Gender",
  "Race1",
  "TotChol",
  "BPDiaAve",
  "BPSysAve",
  "AlcoholYear",
  "Smoke100",
  "DaysPhysHlthBad",
  "PhysActive",
  "Poverty",
  "UrineFlow1",
  "DaysMentHlthBad",
  "HealthGen"
)

newData3 = df3[, var3]
library("corrplot")
```

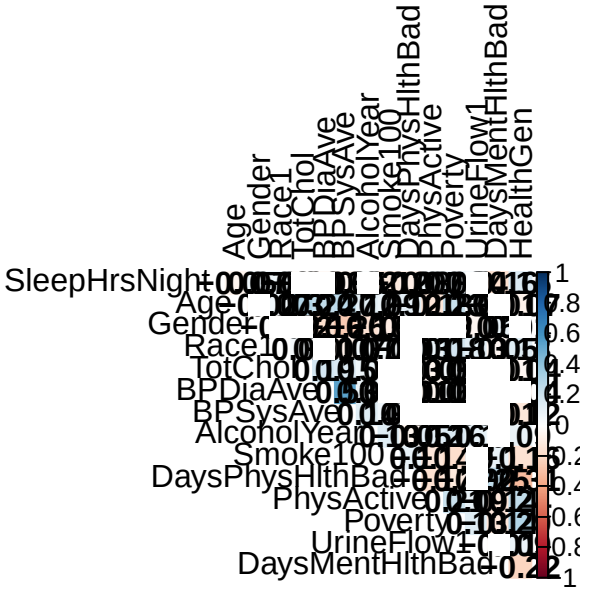
```
## corrplot 0.92 loaded
```

```
par(mfrow = c(1, 2))
cormat3 = cor(as.matrix(newData3[, -c(1)], method = "pearson"))
p.mat3 = cor.mtest(as.matrix(newData3[, -c(1)]))$p
corrplot(
  cormat3,
  method = "color",
  type = "upper",
```

```
  number.cex = 1,
  diag = FALSE,
  addCoef.col = "black",
  tl.col = "black",
  tl.srt = 90,
  p.mat = p.mat3,
  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_3)
```

```
##                       GVIF Df GVIF^(1/(2*Df))
## SleepHrsNight     1.072361  1        1.035549
## Age               1.338877  1        1.157098
## Gender            1.139709  1        1.067571
## factor(Race1)     1.240624  4        1.027318
## Poverty           1.330491  1        1.153469
## TotChol           1.129297  1        1.062684
## BPDiaAve          1.457219  1        1.207153
## BPSysAve          1.573563  1        1.254417
## AlcoholYear       1.127102  1        1.061651
## Smoke100          1.141045  1        1.068197
## UrineFlow1        1.046569  1        1.023020
## DaysMentHlthBad   1.155983  1        1.075167
## DaysPhysHlthBad   1.253009  1        1.119379
## factor(HealthGen) 1.462320  4        1.048649
## PhysActive        1.166248  1        1.079930
```

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


SleepHrsNight
Age
Gender
Race1
TotChol
BPDiaAve
BPSysAve
AlcoholYear
Smoke100
DaysPhysHlthBad
PhysActive
Poverty
UrineFlow1
DaysMentHlthBad
HealthGen
1
0.8
0.6
0.4
0.2
0
0.2
0.4
0.6
0.8
1