Model3

Liancheng

2023-11-25

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

```
##          used (Mb) gc trigger (Mb) max used (Mb)
## Ncells 471519 25.2    1016767 54.4   660860 35.3
## Vcells 891425  6.9    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)
############### (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
```

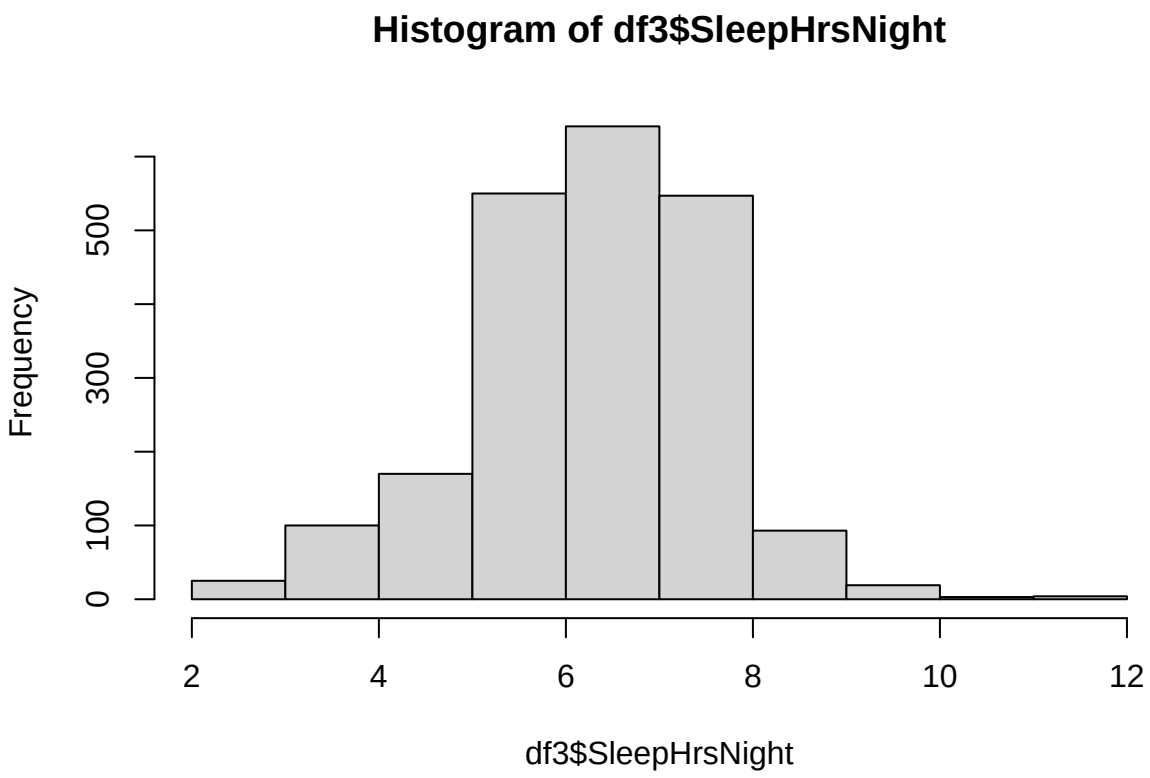
```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
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,
```

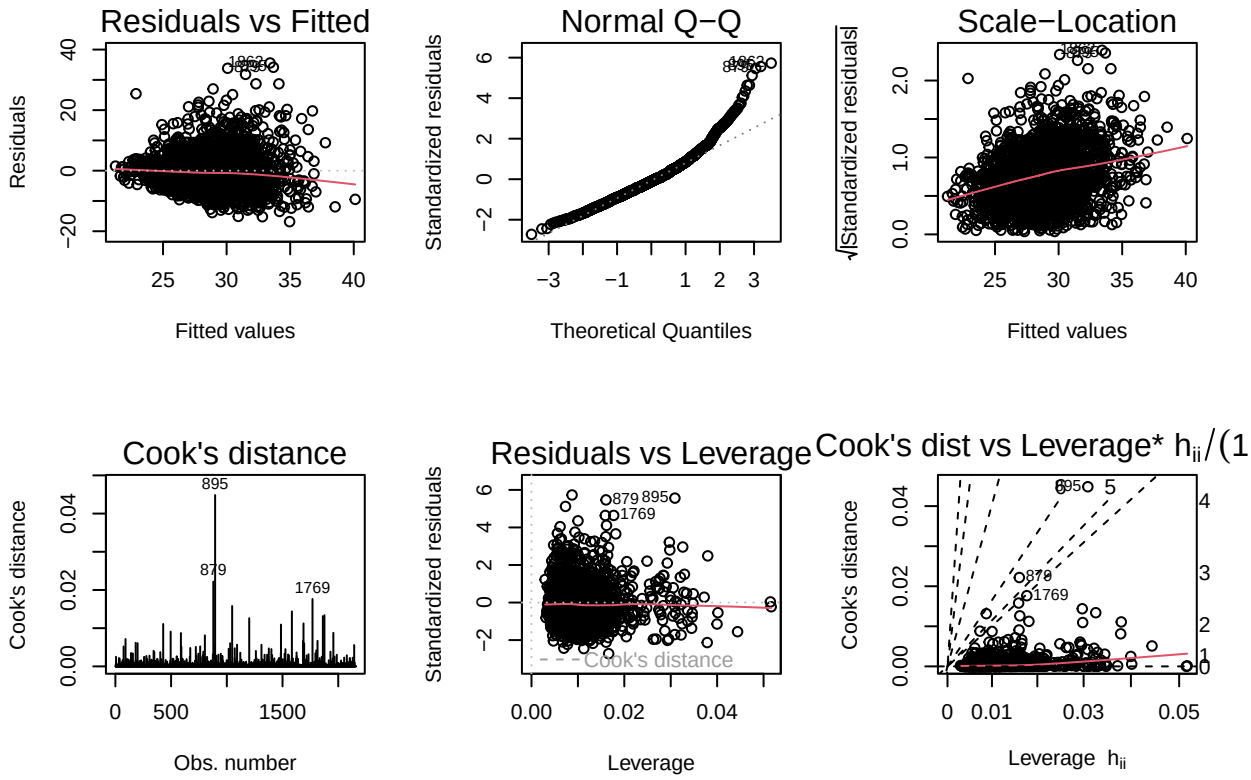
```
      HealthGen == 'Vgood' ~ 4,
      HealthGen == 'Excellent' ~ 5,
      TRUE ~ NA_integer_  # Default value if none of the conditions are met
    )
  )
## model_3 add additional risk factors ##
m_3 = lm(
  BMI ~ SleepHrsNight +Age + Gender + factor(Race1)  + Poverty + TotChol+ BPDiaAve + BPSysAve + Alcohol
)
summary(m_3)
```

```
##
## Call:
## lm(formula = BMI ~ 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
## -16.846  -4.036  -0.638   3.219  35.582
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        26.386667   1.890991  13.954  < 2e-16 ***
## SleepHrsNight      -0.127697   0.106301  -1.201 0.229780
## Age                 0.008073   0.013731   0.588 0.556655
## Gender              0.485216   0.287603   1.687 0.091730 .
## factor(Race1)2     -1.950677   0.641044  -3.043 0.002371 **
## factor(Race1)3     -1.185845   0.562009  -2.110 0.034974 *
## factor(Race1)4     -1.455221   0.421226  -3.455 0.000562 ***
## factor(Race1)5     -3.317989   0.631208  -5.257 1.61e-07 ***
## Poverty             0.056394   0.091716   0.615 0.538701
## TotChol             0.027346   0.135922   0.201 0.840570
## BPDiaAve            0.058320   0.013708   4.254 2.19e-05 ***
## BPSysAve            0.050981   0.011813   4.316 1.66e-05 ***
## AlcoholYear        -0.008650   0.001515  -5.709 1.30e-08 ***
## Smoke100           -0.866408   0.288024  -3.008 0.002660 **
## UrineFlow1         -0.108771   0.142376  -0.764 0.444969
## DaysMentHlthBad    -0.032496   0.018025  -1.803 0.071559 .
## DaysPhysHlthBad     0.013749   0.020944   0.656 0.511596
## factor(HealthGen)2 -2.270473   1.001099  -2.268 0.023430 *
## factor(HealthGen)3 -4.001479   0.992277  -4.033 5.71e-05 ***
## factor(HealthGen)4 -5.723244   1.018409  -5.620 2.16e-08 ***
## factor(HealthGen)5 -7.573168   1.076429  -7.035 2.67e-12 ***
## PhysActive         -0.843726   0.294095  -2.869 0.004160 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.237 on 2130 degrees of freedom
## Multiple R-squared:  0.1554, Adjusted R-squared:  0.1471
## F-statistic: 18.66 on 21 and 2130 DF,  p-value: < 2.2e-16
```

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

```
## Anova Table (Type III tests)
##
## Response: BMI
##                    Sum Sq   Df  F value    Pr(>F)
## (Intercept)          7575    1 194.7109 < 2.2e-16 ***
## SleepHrsNight          56    1   1.4430   0.22978
## Age                    13    1   0.3456   0.55666
## Gender                111    1   2.8463   0.09173 .
## factor(Race1)        1149    4   7.3829 6.660e-06 ***
## Poverty                15    1   0.3781   0.53870
## TotChol                 2    1   0.0405   0.84057
## BPDiaAve              704    1  18.0997 2.187e-05 ***
## BPSysAve              725    1  18.6262 1.663e-05 ***
## AlcoholYear          1268    1  32.5872 1.299e-08 ***
## Smoke100              352    1   9.0488   0.00266 **
## UrineFlow1             23    1   0.5837   0.44497
## DaysMentHlthBad       126    1   3.2501   0.07156 .
## DaysPhysHlthBad        17    1   0.4309   0.51160
## factor(HealthGen)    4589    4  29.4921 < 2.2e-16 ***
## PhysActive            320    1   8.2305   0.00416 **
## Residuals           82864 2130
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
########### 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)
```

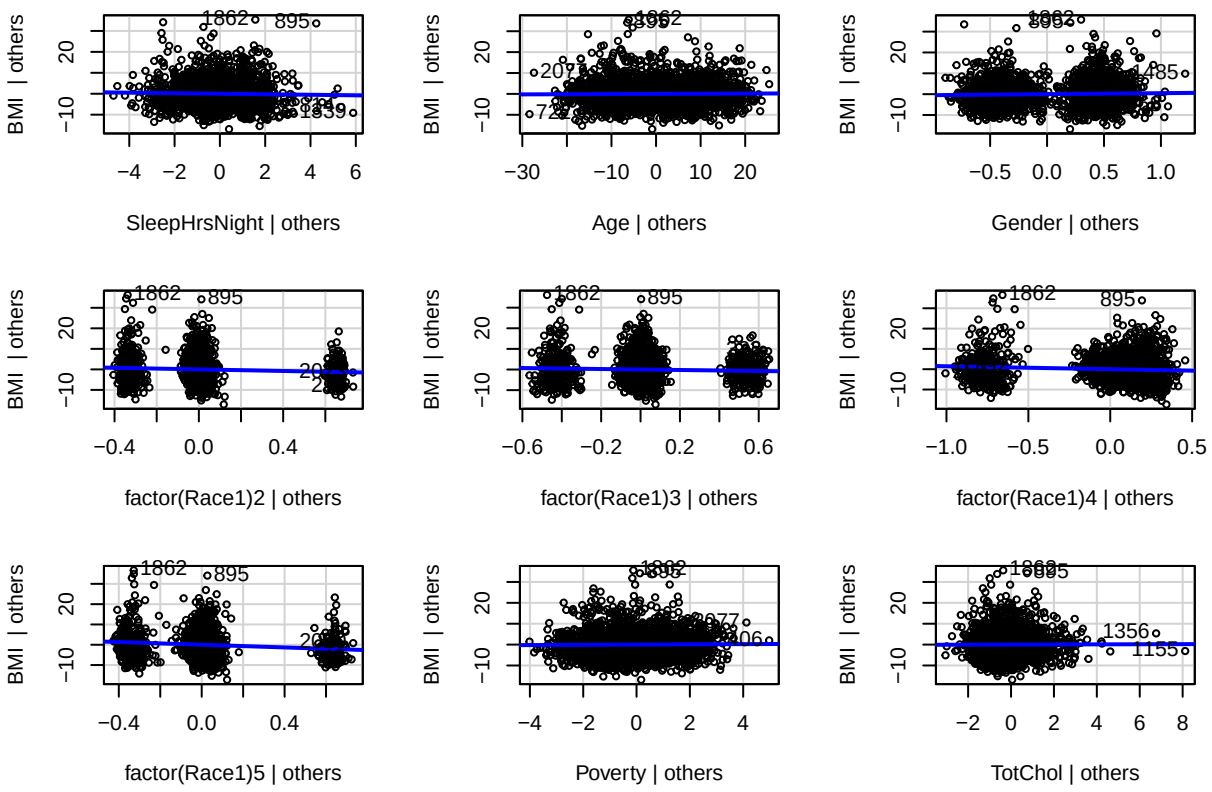
```
## m_3.h
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2152        0     2152        1  0.01022 0.005352 0.004645 0.005142
##      .25      .50      .75      .90      .95
## 0.006490 0.009118 0.012287 0.015996 0.019196
##
## lowest : 0.002934252 0.003128726 0.003292556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

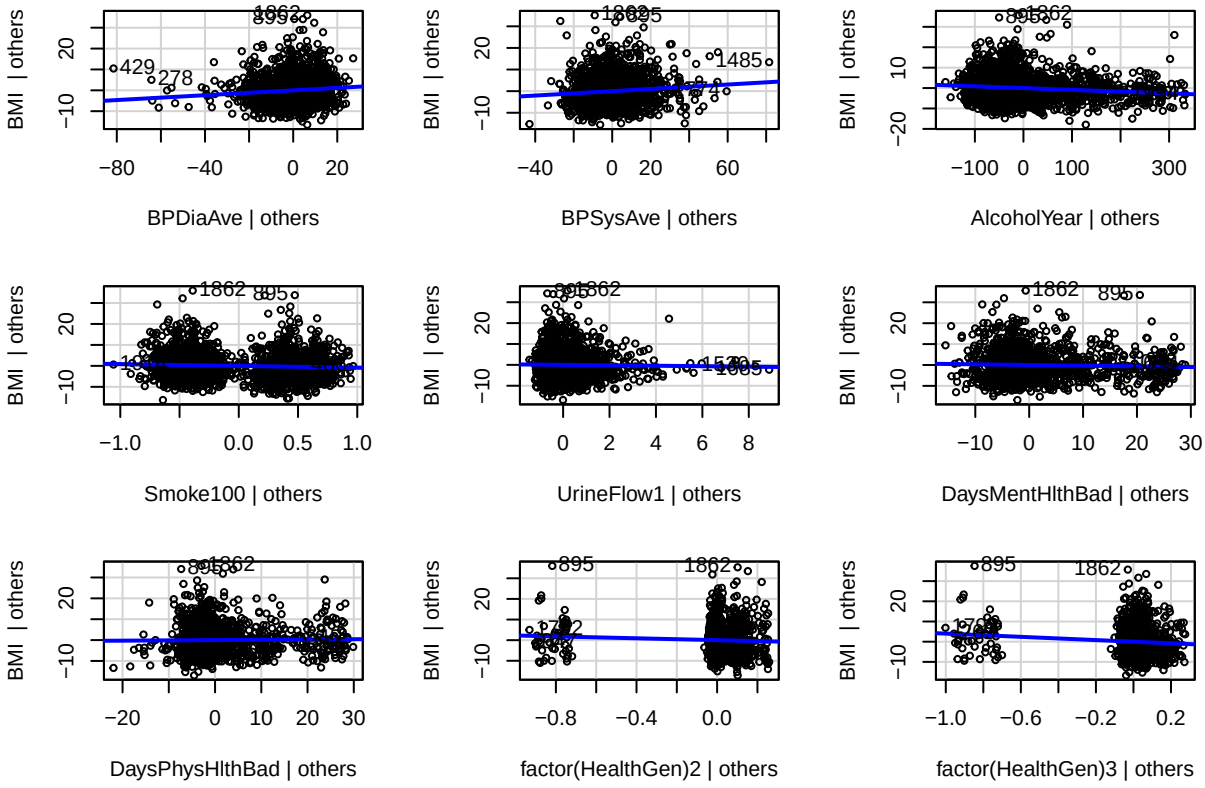
```
m_3.h[which.max(m_3.h)]
```

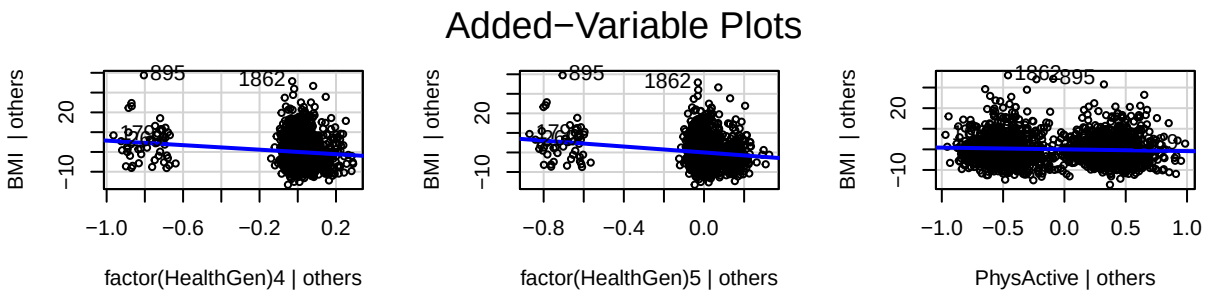
```
##       1685
## 0.05164063
```

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

#(1)Linear: 2 approaches
```

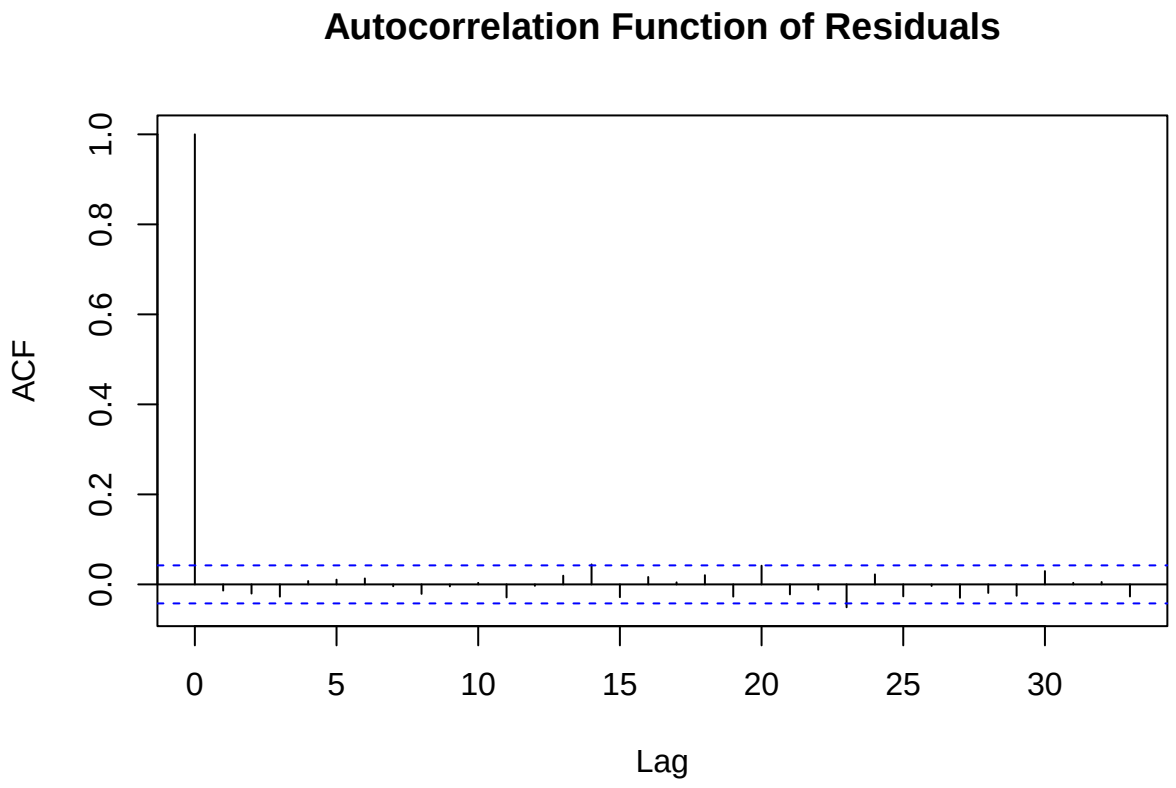
```
# partial regression plots
car::avPlots(m_3)
```



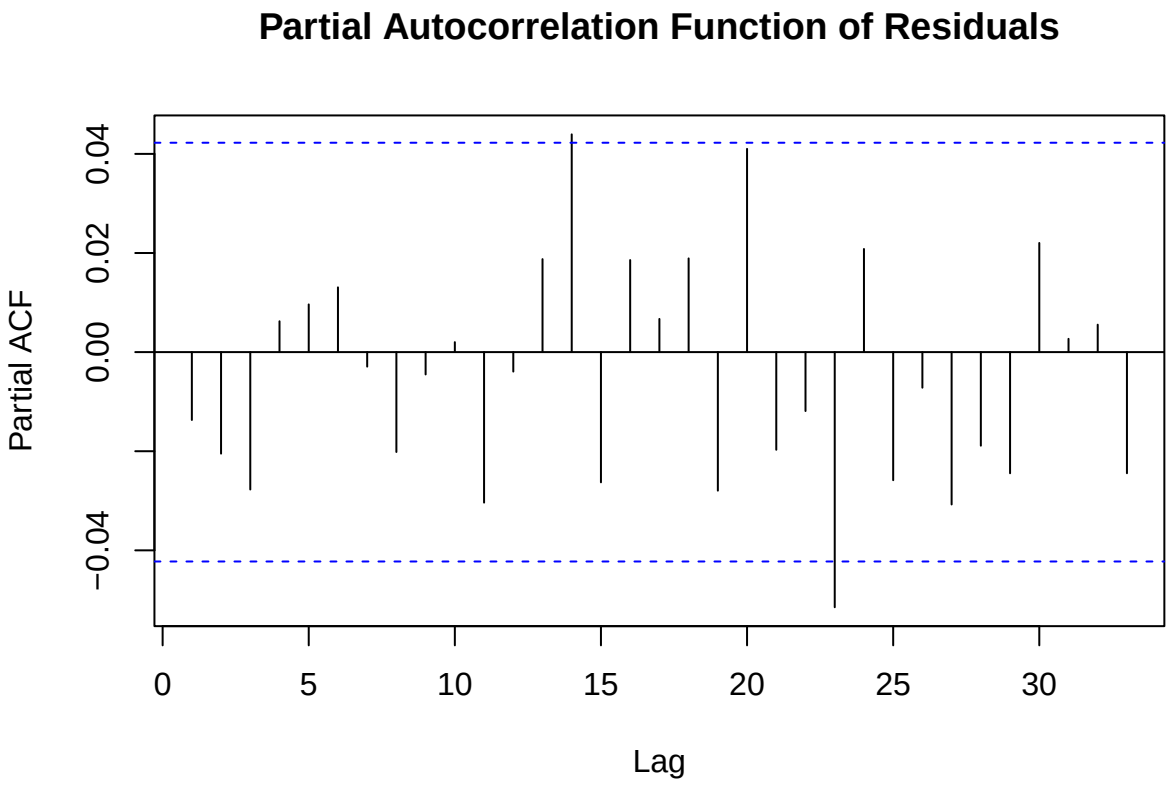

BMI | others
BPDiaAve | others
BPSysAve | others
AlcoholYear | others
Smoke100 | others
UrineFlow1 | others
DaysMentHlthBad | others
DaysPhysHlthBad | others
factor(HealthGen)2 | others
factor(HealthGen)3 | others



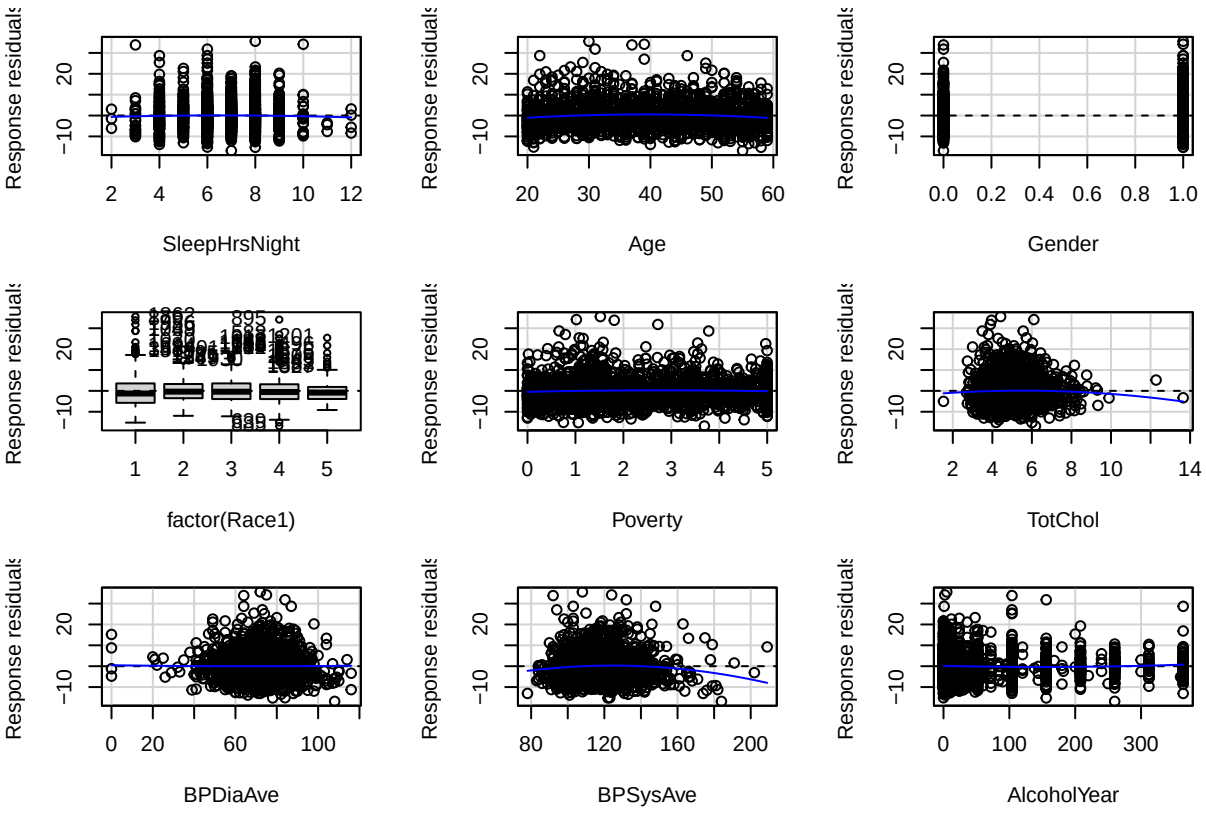

```r
#(2)Independence:

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

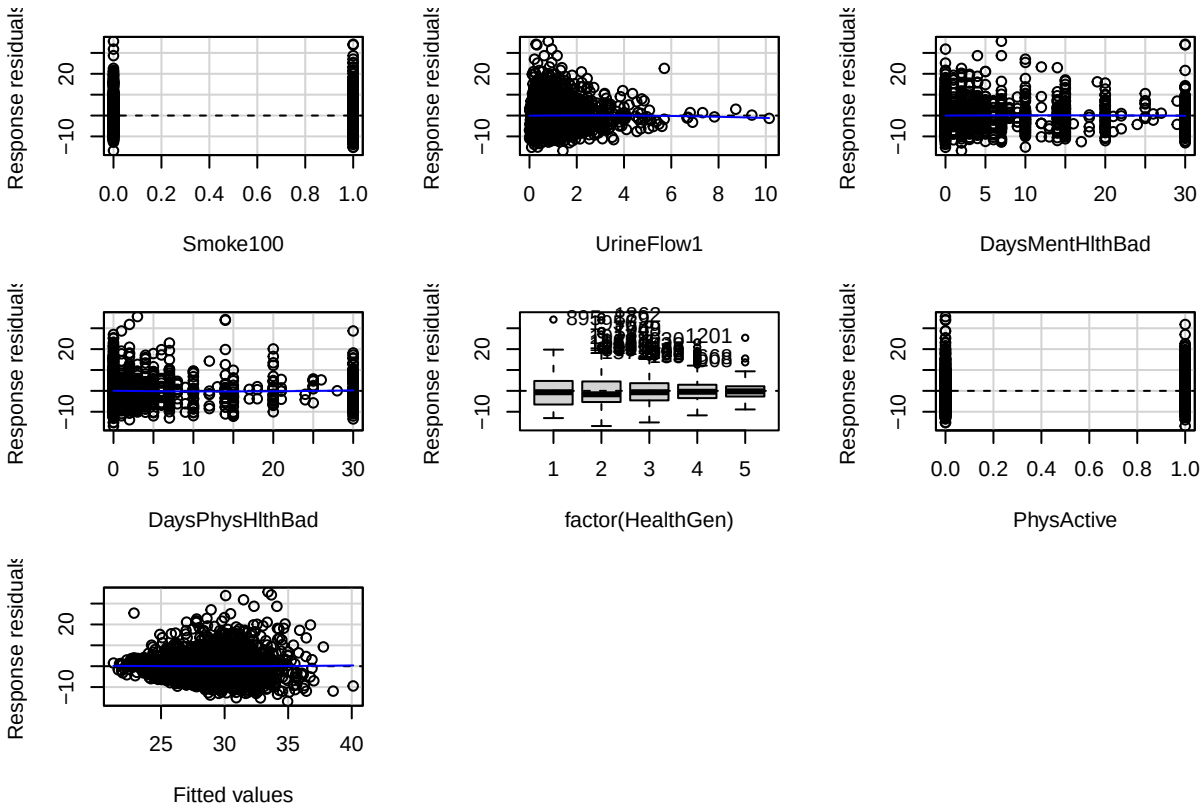



```
pacf(residuals, main = "Partial Autocorrelation Function of Residuals")
```



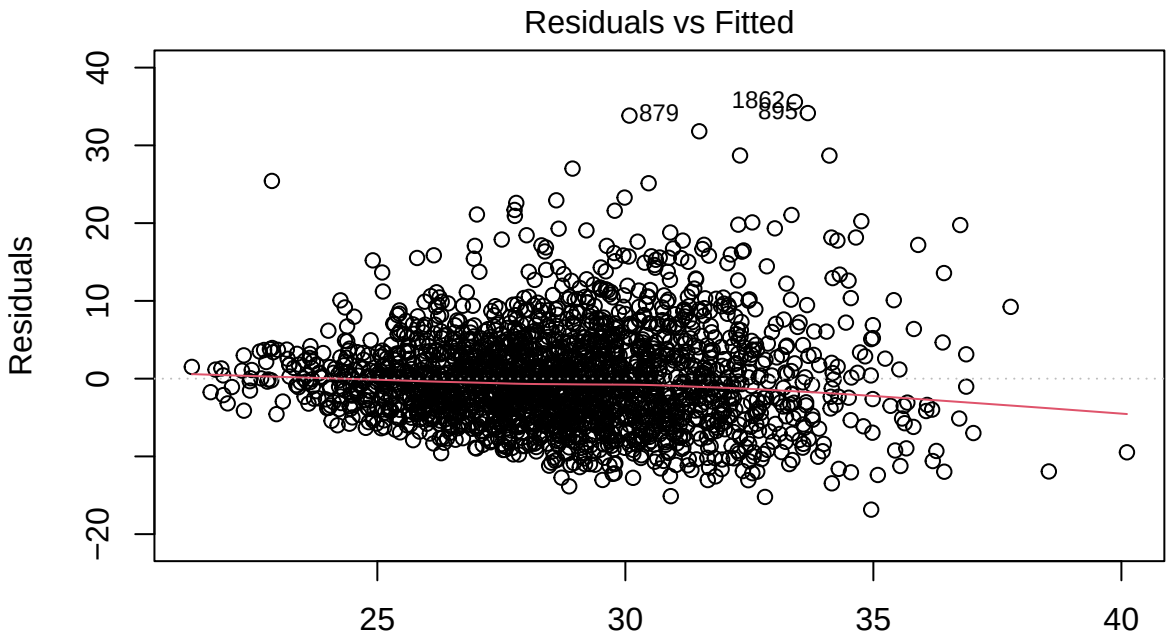

```
#(3)E: constant var: residuals-fitted values; transform for variance-stable...(total: 4 solutions)

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


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

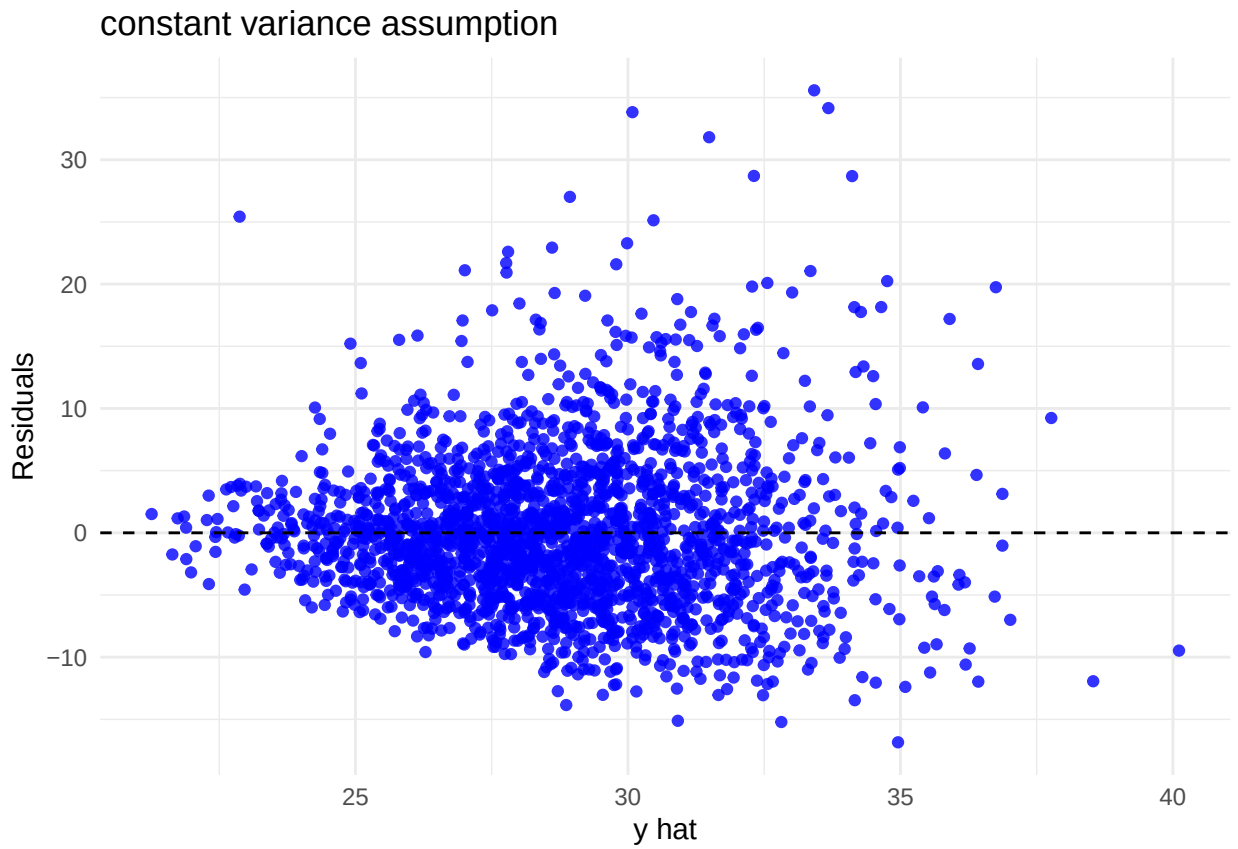



```
##                    Test stat Pr(>|Test stat|)
## SleepHrsNight        -0.6894        0.4906502
## Age                  -3.8160        0.0001395 ***
## Gender                0.4364        0.6625881
## factor(Race1)
## Poverty              -1.3727        0.1699992
## TotChol              -1.4181        0.1563160
## BPDiaAve              0.2591        0.7955996
## BPSysAve             -3.5685        0.0003669 ***
## AlcoholYear           1.8242        0.0682681 .
## Smoke100              0.0065        0.9948406
## UrineFlow1           -0.4240        0.6716176
## DaysMentHlthBad      -0.3811        0.7031405
## DaysPhysHlthBad       0.7522        0.4520044
## factor(HealthGen)
## PhysActive           -0.5408        0.5886962
## Tukey test            0.3156        0.7522713
## ---
## 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()
```



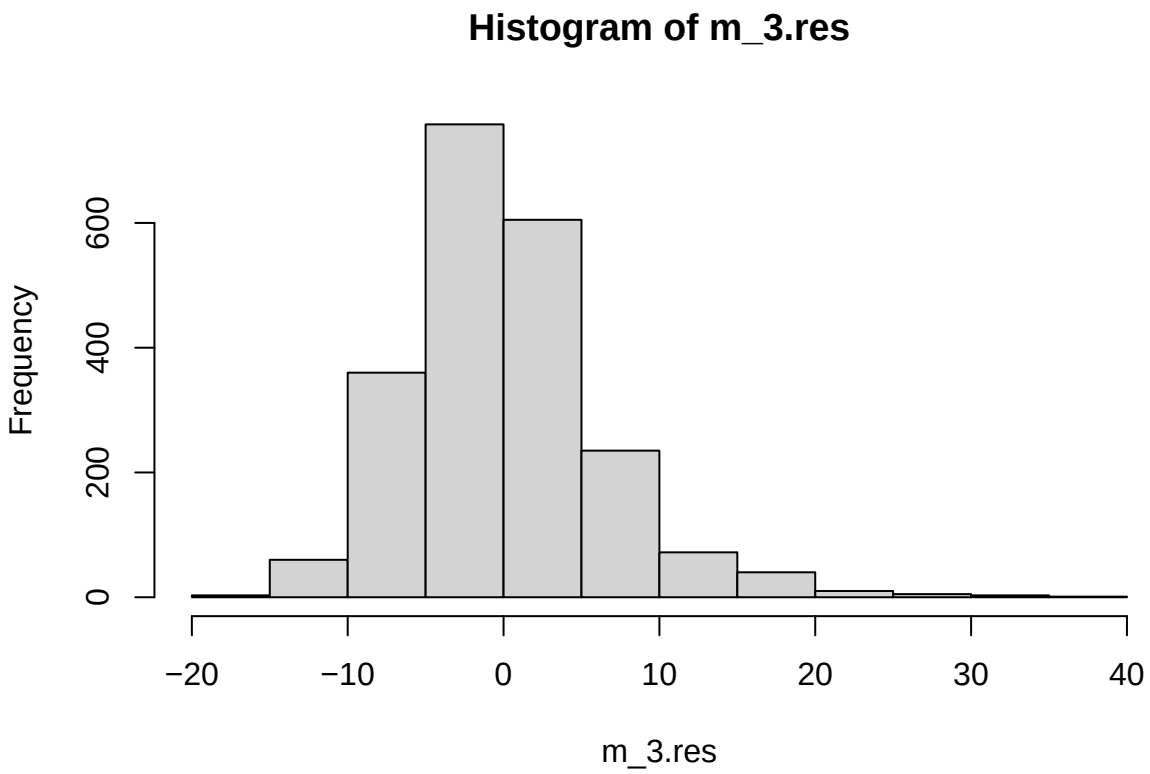

```
#conclusion: the constant variance assumption is basically not violated. The spread of the residuals ap


#(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 -2.624e-17      6.688    -8.6788
##        .10        .25        .50        .75        .90        .95
##    -7.0145    -4.0358    -0.6378     3.2188     7.5294    10.5440
##
## lowest : -16.84565 -15.21486 -15.11587 -13.84800 -13.46278
## highest:  28.69839  31.80996  33.82742  34.15169  35.58233
```

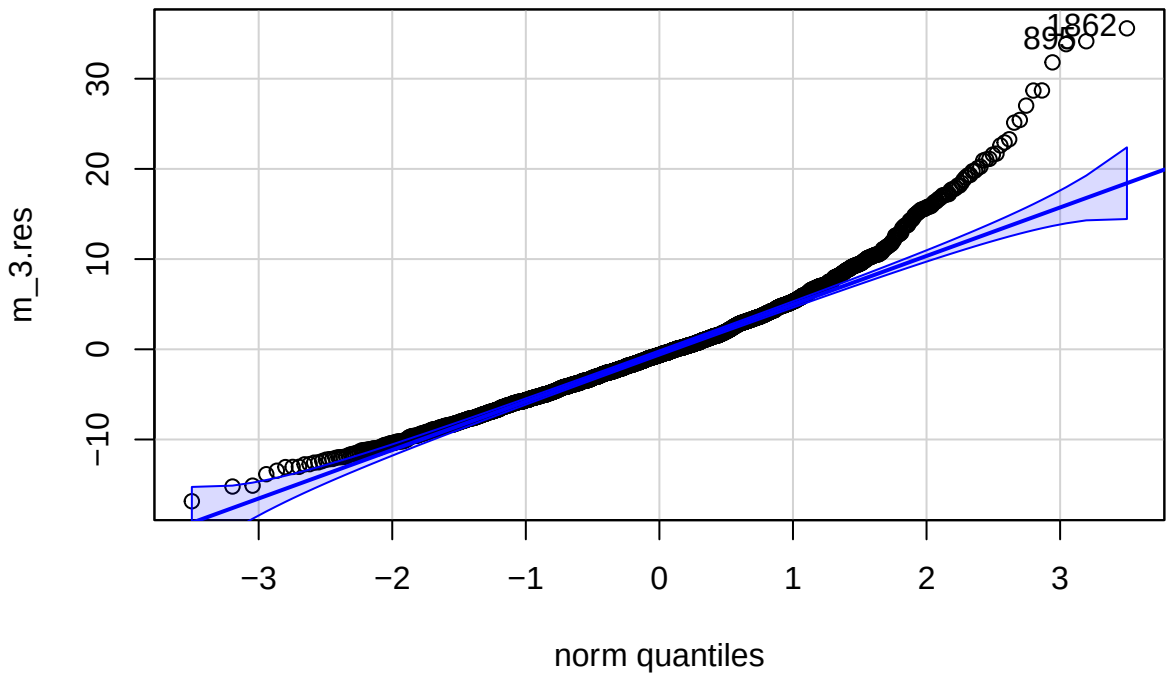
```
Hmisc::describe(m_3.res)$counts[c(".25",".50",".75")] #not symmetric
```

```
##       .25       .50       .75
## "-4.0358" "-0.6378" " 3.2188"
```

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



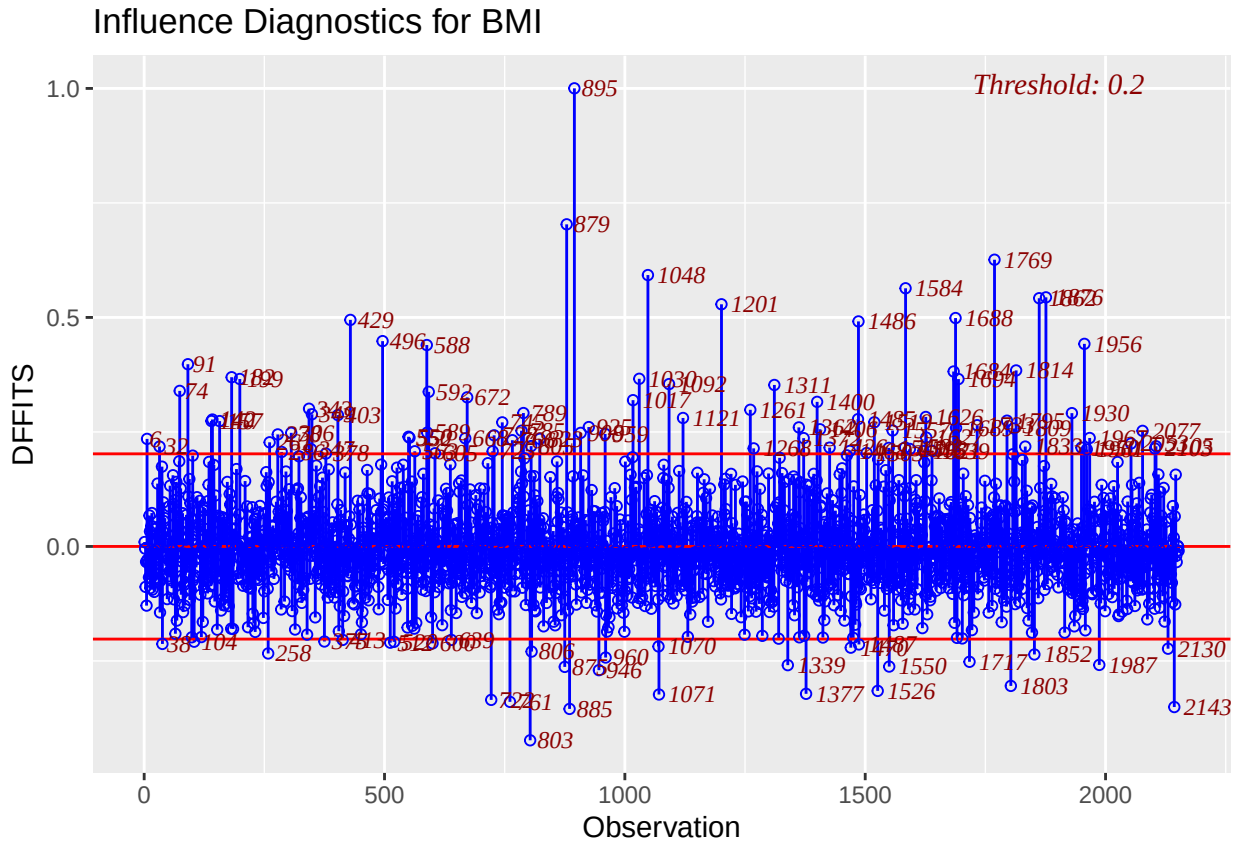

```
# Q-Q plot
qq.m_3.res=car::qqPlot(m_3.res)
```




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

```
##     1862      895
## 35.58233 34.15169
```

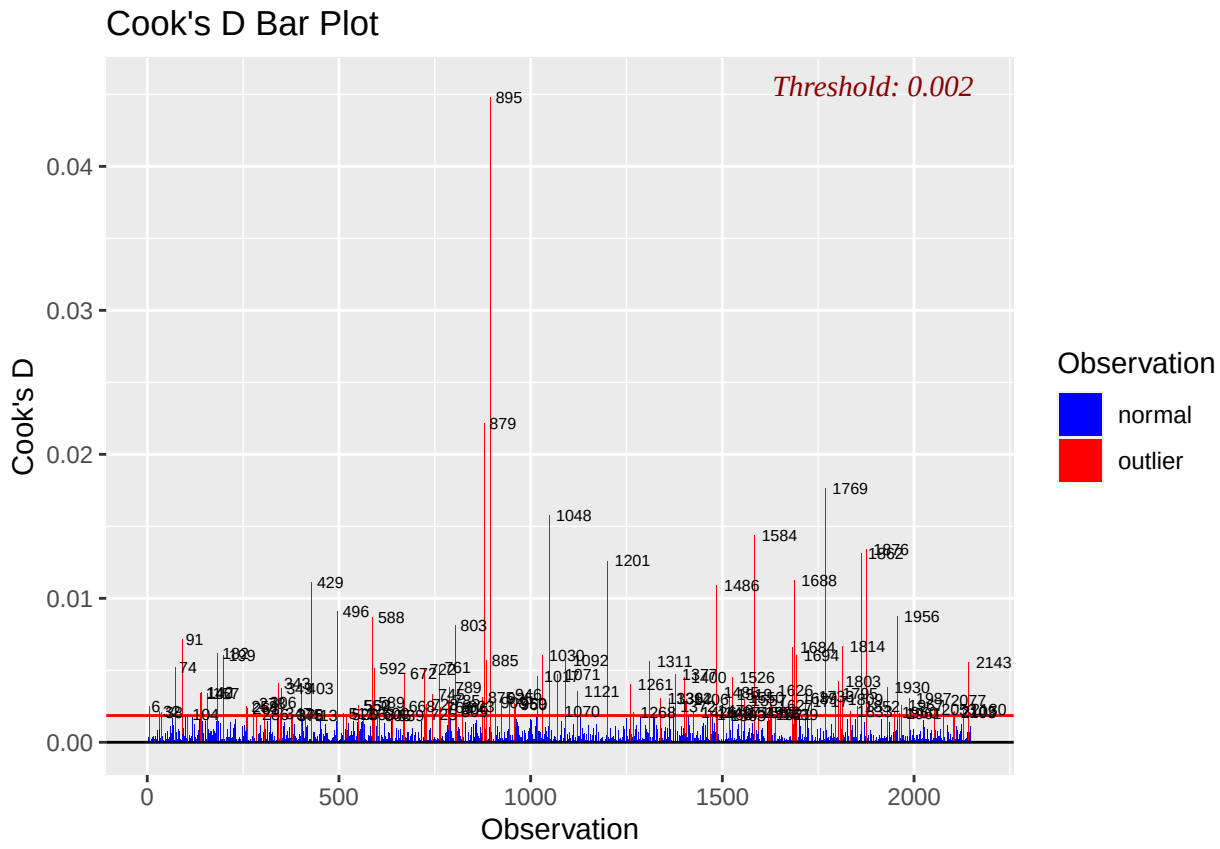
```
################ influential observations  #################

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  34.15169 5.601548 0.03089695 0.7556659 1.0001862    0.04483210
## 879  33.82742 5.504997 0.01605793 0.7524436 0.7032619    0.02217569
## 1769 28.68626 4.662976 0.01770429 0.8225458 0.6260101    0.01764132
## 1048 28.69839 4.660656 0.01590422 0.8212229 0.5924947    0.01580308
## 1584 19.74878 3.221428 0.02970559 0.9356728 0.5636581    0.01437808
## 1876 18.14589 2.963259 0.03258574 0.9540004 0.5438475    0.01339516
```

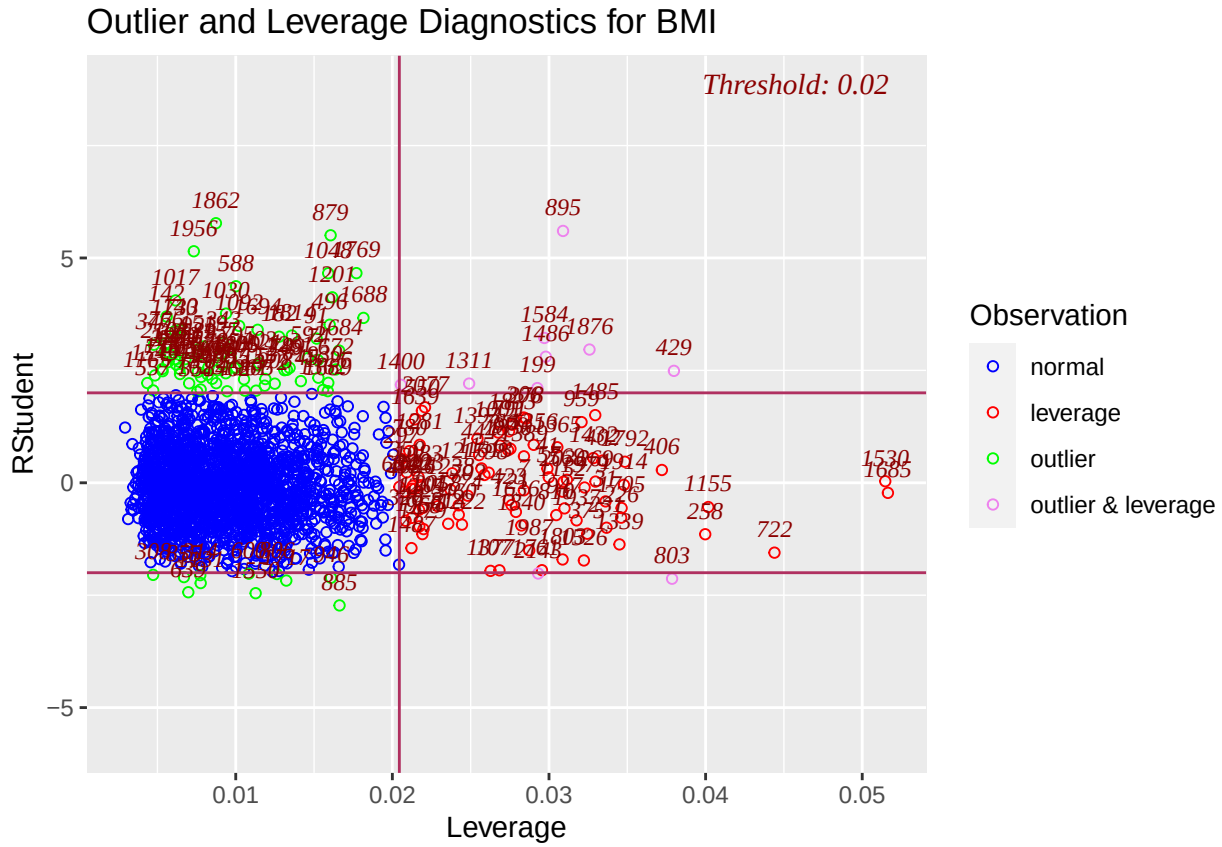
```
#From the plot above, we can see 2 observations with the largest (magnitude) of DFFITS, observation 879

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




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

```
##      Residual Rstudent   HatDiagH  CovRatio    DFFITS COOKsDistance
## 895  34.15169 5.601548 0.03089695 0.7556659 1.0001862    0.04483210
## 879  33.82742 5.504997 0.01605793 0.7524436 0.7032619    0.02217569
## 1769 28.68626 4.662976 0.01770429 0.8225458 0.6260101    0.01764132
## 1048 28.69839 4.660656 0.01590422 0.8212229 0.5924947    0.01580308
## 1584 19.74878 3.221428 0.02970559 0.9356728 0.5636581    0.01437808
## 1876 18.14589 2.963259 0.03258574 0.9540004 0.5438475    0.01339516
```

```
#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 -1.3626686 -0.22429149 0.05164063 1.0648490 -0.052338600  1.245705e-04
## 1530  0.2082322  0.03427099 0.05147106 1.0651993  0.007983308  2.898323e-06
## 722  -9.4718286 -1.55397971 0.04439799 1.0312845 -0.334956562  5.096428e-03
## 1155 -3.2990503 -0.53979083 0.04017058 1.0495065 -0.110428837  5.544812e-04
## 258  -6.9957375 -1.14480045 0.03997521 1.0383041 -0.233605976  2.480173e-03
## 429  15.2117071  2.48954775 0.03797756 0.9852067  0.494642452  1.109434e-02
```

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

```
##      Residual Rstudent    HatDiagH  CovRatio    DFFITS COOKsDistance
## 1862 35.58233 5.773197 0.008734615 0.7242337 0.5419302    0.01314988
## 895  34.15169 5.601548 0.030896951 0.7556659 1.0001862    0.04483210
## 879  33.82742 5.504997 0.016057928 0.7524436 0.7032619    0.02217569
## 1956 31.80996 5.149333 0.007324108 0.7752074 0.4423078    0.00878729
## 1769 28.68626 4.662976 0.017704294 0.8225458 0.6260101    0.01764132
## 1048 28.69839 4.660656 0.015904216 0.8212229 0.5924947    0.01580308
```

```
#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

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

```
##                         Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)        26.386667481 1.890990832 13.9538844 2.066969e-42
## SleepHrsNight      -0.127696522 0.106301357 -1.2012690 2.297804e-01
## Age                 0.008072597 0.013731018  0.5879096 5.566553e-01
## Gender              0.485216014 0.287603245  1.6871020 9.173016e-02
## factor(Race1)2     -1.950676945 0.641044438 -3.0429668 2.371201e-03
## factor(Race1)3     -1.185844891 0.562008984 -2.1100106 3.497379e-02
## factor(Race1)4     -1.455221376 0.421226492 -3.4547242 5.616368e-04
## factor(Race1)5     -3.317988642 0.631208275 -5.2565671 1.614282e-07
## Poverty             0.056394179 0.091715993  0.6148784 5.387007e-01
## TotChol             0.027346036 0.135921718  0.2011896 8.405695e-01
## BPDiaAve            0.058320373 0.013708327  4.2543758 2.187091e-05
## BPSysAve            0.050980958 0.011812618  4.3158051 1.663011e-05
## AlcoholYear        -0.008650015 0.001515282 -5.7085174 1.299002e-08
## Smoke100           -0.866408483 0.288023562 -3.0081167 2.659507e-03
## UrineFlow1         -0.108771300 0.142376192 -0.7639711 4.449691e-01
## DaysMentHlthBad    -0.032495539 0.018024890 -1.8028148 7.155859e-02
## DaysPhysHlthBad     0.013749204 0.020944330  0.6564642 5.115964e-01
## factor(HealthGen)2 -2.270472504 1.001098685 -2.2679807 2.343013e-02
## factor(HealthGen)3 -4.001478912 0.992276845 -4.0326235 5.709988e-05
## factor(HealthGen)4 -5.723244058 1.018408925 -5.6197898 2.162548e-08
## factor(HealthGen)5 -7.573168404 1.076429170 -7.0354545 2.666463e-12
## PhysActive         -0.843725954 0.294095010 -2.8688891 4.159738e-03
```

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

```
##                          Estimate  Std. Error     t value     Pr(>|t|)
## (Intercept)     26.442057879 1.655369868  15.9735044 2.482773e-54
## SleepHrsNight   -0.078469729 0.104172937  -0.7532641 4.513746e-01
## Age              0.014296918 0.013420336   1.0653175 2.868532e-01
## Gender           0.406246753 0.280966372   1.4458910 1.483551e-01
## Race1           -0.366482618 0.119733800  -3.0608117 2.234971e-03
## Poverty          0.070973840 0.088782815   0.7994097 4.241422e-01
## TotChol          0.057202195 0.135190627   0.4231225 6.722487e-01
## BPDiaAve         0.054964501 0.013434024   4.0914398 4.446870e-05
## BPSysAve         0.054662515 0.011621752   4.7034658 2.722010e-06
## AlcoholYear     -0.009268099 0.001490117  -6.2197127 5.979088e-10
## Smoke100        -0.869942857 0.280904254  -3.0969373 1.980833e-03
## UrineFlow1      -0.126487523 0.139418349  -0.9072516 3.643765e-01
## DaysMentHlthBad -0.025740738 0.017648097  -1.4585560 1.448350e-01
## DaysPhysHlthBad -0.003255820 0.020182627  -0.1613180 8.718583e-01
## HealthGen       -1.722854279 0.160250648 -10.7509973 2.739864e-26
## PhysActive      -0.744654891 0.287400481  -2.5910009 9.635138e-03
```

```
#### 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
##       2.099181e-01       3.854983e+01       7.710432e+01       1.627507e+01
##     factor(Race1)2     factor(Race1)3     factor(Race1)4     factor(Race1)5
##       8.121254e+01       1.059851e+02       1.039308e+02       1.016566e+02
##            Poverty            TotChol          BPDiaAve          BPSysAve
##       3.070643e+00       1.338919e+02       1.591662e+03       3.481074e+02
##        AlcoholYear           Smoke100         UrineFlow1    DaysMentHlthBad
##       1.975803e+02       9.962422e+01       1.483924e+03       2.191560e+03
##    DaysPhysHlthBad factor(HealthGen)2 factor(HealthGen)3 factor(HealthGen)4
##       1.922170e+05       9.654390e+01       1.003573e+02       1.070982e+02
## factor(HealthGen)5         PhysActive
##       9.516078e+01       1.084120e+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
```

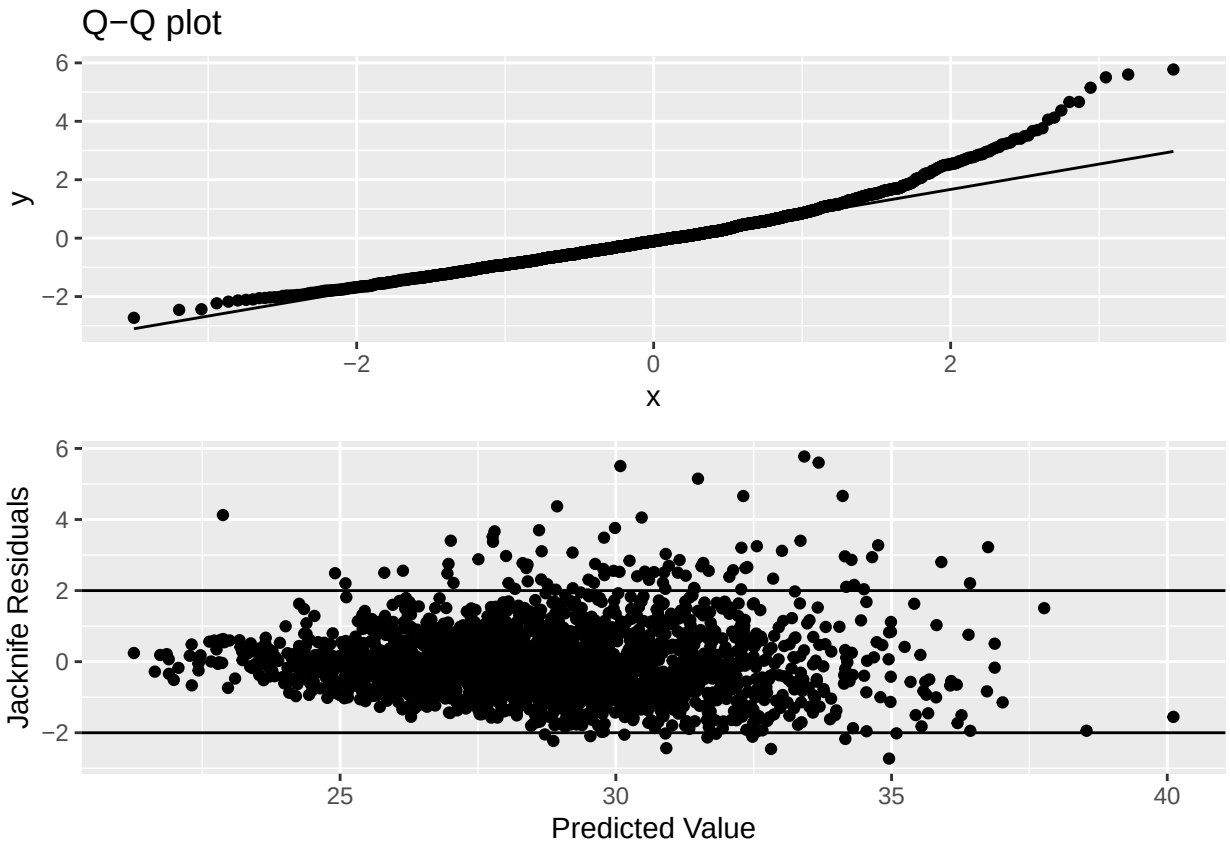
```
#From the VIF values in the output above, once again we do not observe any potential collinearity issue

################## using log-transformed BMI  ###################
# log BMI
df3$logBMI = log(df3$BMI+1)
m_3.log= lm(logBMI ~ SleepHrsNight + Age + Gender + factor(Race1)  + Poverty + TotChol+ BPDiaAve + BPSys
p31.log = ols_plot_resid_lev(m_3.log)
p32.log = ols_plot_cooksd_bar(m_3.log)
library(gridExtra)
```

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

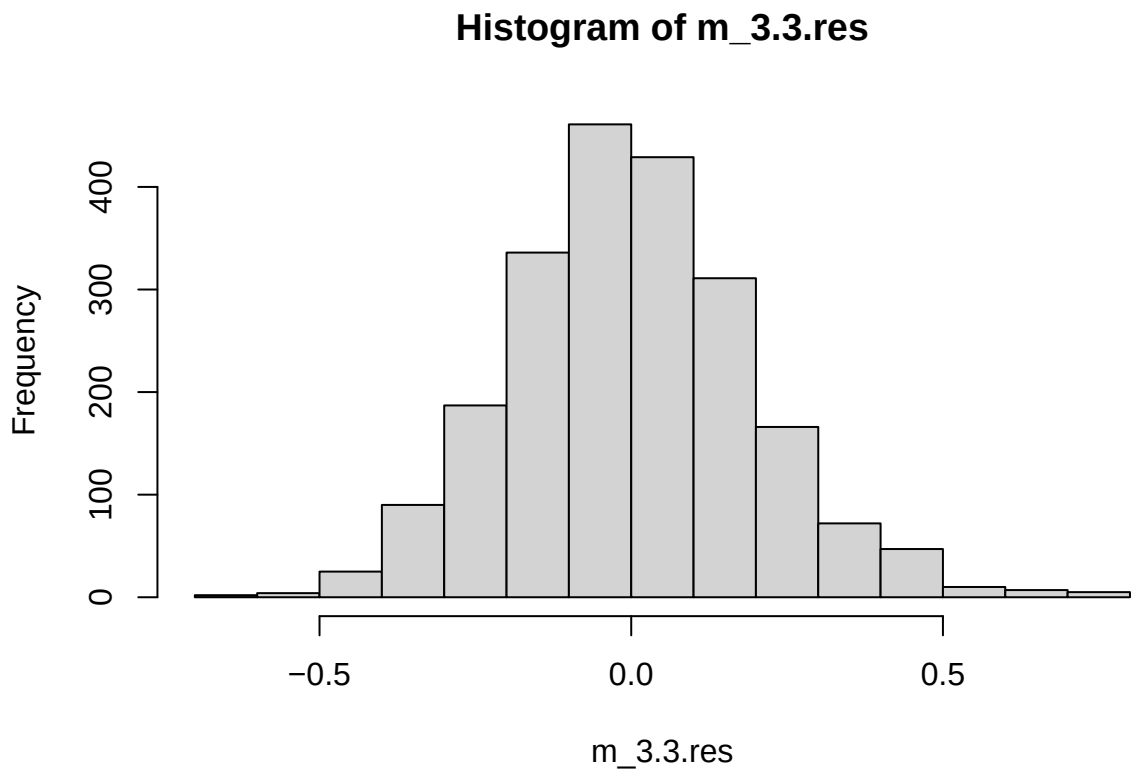
```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
p33.log = ggplot(m_3.log, aes(sample = rstudent(m_3.log))) + geom_qq() + stat_qq_line() +labs(title="Q-
p34.log = ggplot() + geom_point(aes(y = rstudent(m_3.log), x = m_3.log$fitted.values )) + labs(x = "Pre
grid.arrange(p33.log,p34.log, nrow=2)


p33 = ggplot(m_3, aes(sample = rstudent(m_3))) + geom_qq() + stat_qq_line() +labs(title="Q-Q plot")
p34 = ggplot() + geom_point(aes(y = rstudent(m_3), x = m_3$fitted.values )) + labs(x = "Predicted Value
grid.arrange(p33,p34, nrow=2)
```

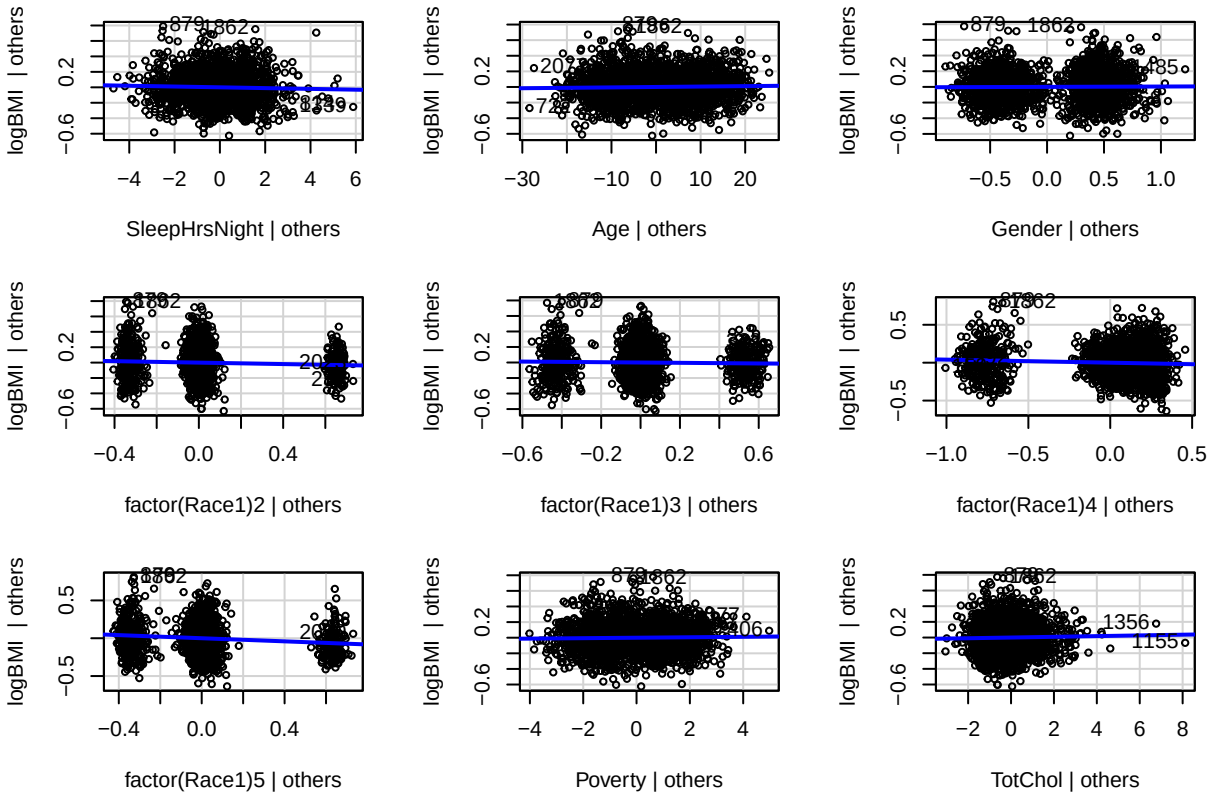
```
m_3.3.yhat=m_3.log$fitted.values
m_3.3.res=m_3.log$residuals
m_3.3.h=hatvalues(m_3.log)
m_3.3.r=rstandard(m_3.log)
m_3.3.rr=rstudent(m_3.log)

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

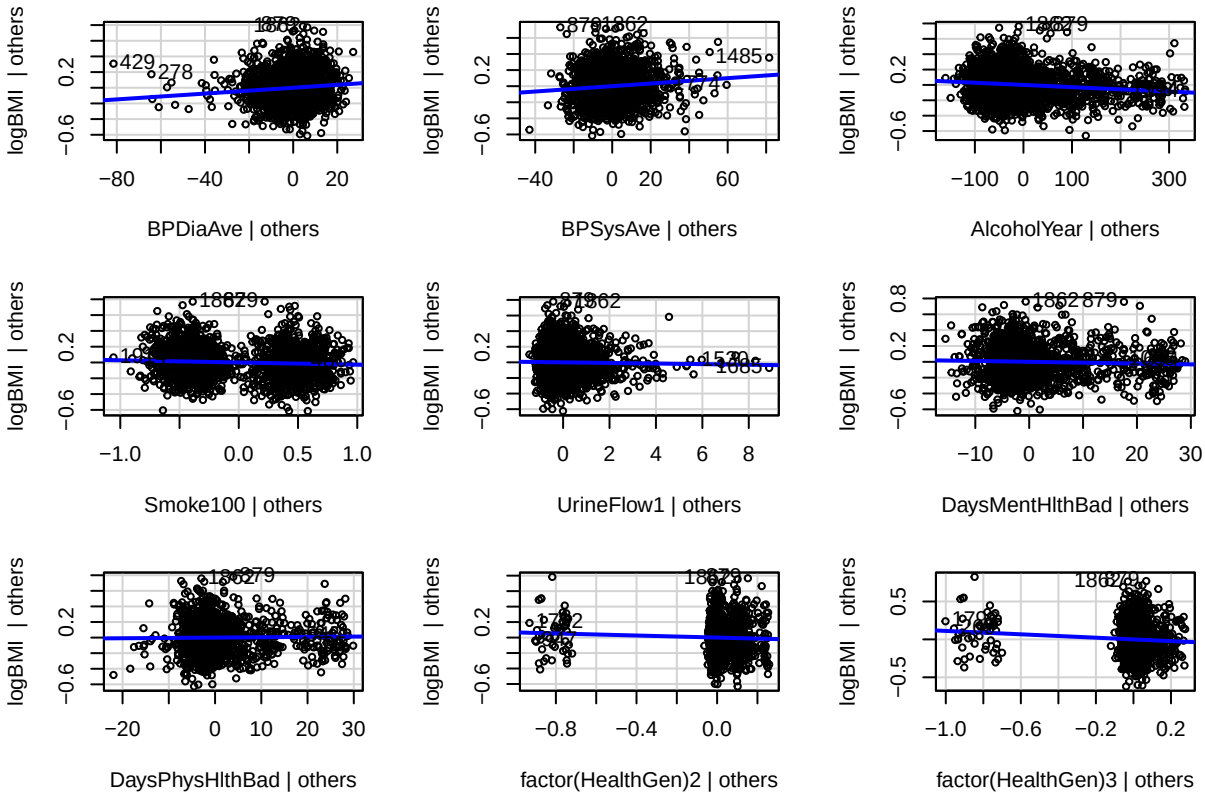


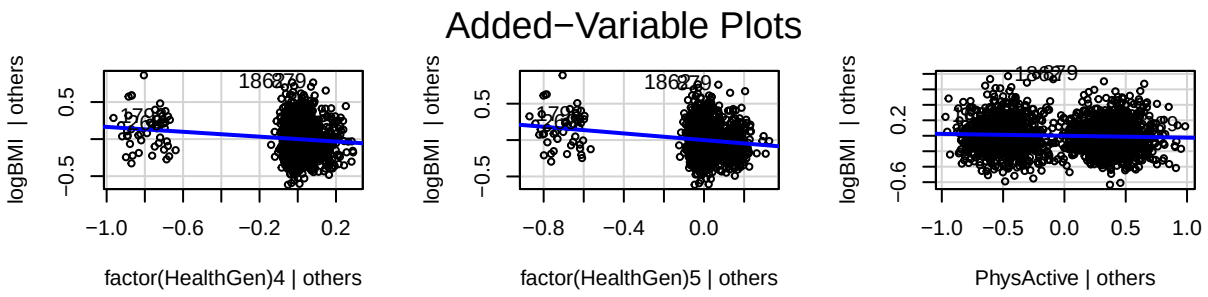

```
hist(m_3.3.res,breaks = 15)
```




```
car::avPlots(m_3.log)
```


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


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




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

#collinearity diagnostics

car::vif(m_3.log)
```

```
##                       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
```

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