Model1

Liancheng

2023-11-25

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

```
##          used (Mb) gc trigger (Mb) max used (Mb)
## Ncells 472528 25.3    1019650 54.5   660860 35.3
## Vcells 898767  6.9    8388608 64.0  1800812 13.8
```

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

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

```
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'
```

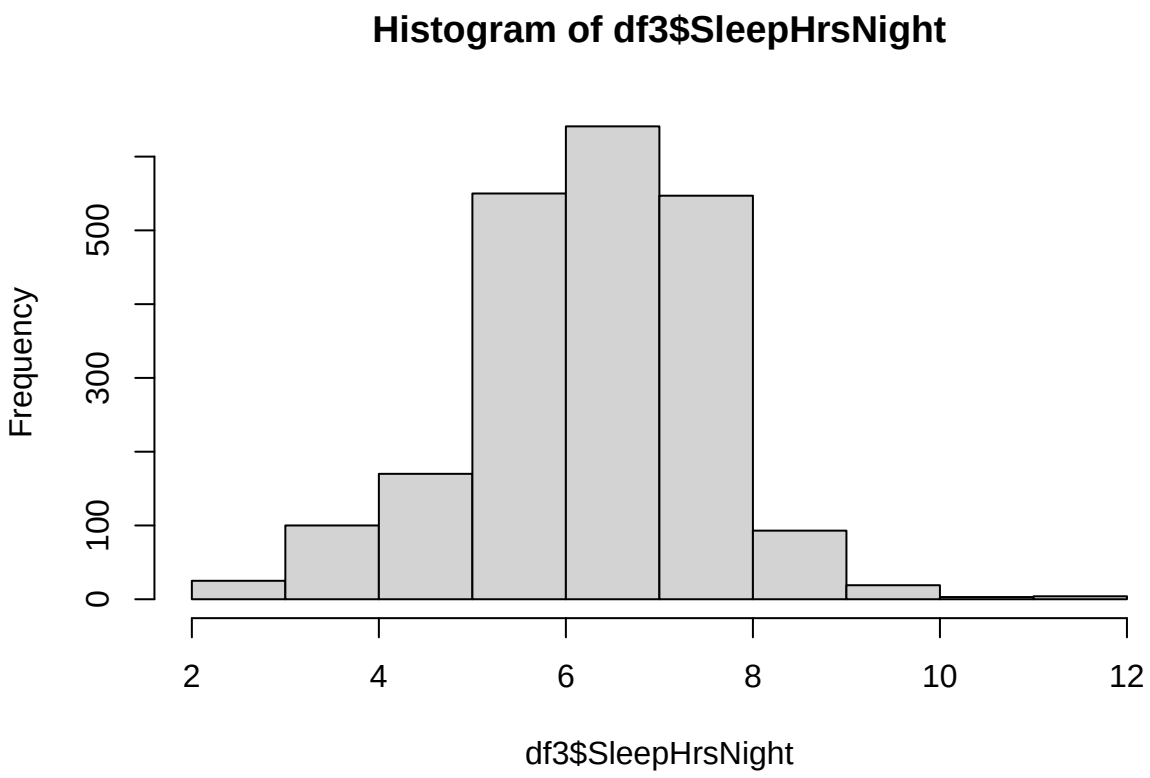
```
## The following object is masked from 'package:car':
##
##     recode

## 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
    )
  )

##variables diagnosis? normality

##descriptive statisctis covariates (like modulek p55)
#psych::pairs.panels(df3)

############### (2) Baseline characteristics ##########################################
```

```
## categorize sleeptime, show the distribution of variables(classified into short sleep (<7 h/day), reg
Hmisc::describe(df3)
```

```
## df3
##
##  21  Variables      2152  Observations
## ---------------------------------------------------------------------------
## SleepHrsNight
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2152        0       11     0.94    6.781    1.415        4        5
##      .25      .50      .75      .90      .95
##        6        7        8        8        9
##
## lowest :  2  3  4  5  6, highest:  8  9 10 11 12
##
## Value          2     3     4     5     6     7     8     9    10    11    12
## Frequency      3    22   100   170   550   641   547    93    19     3     4
## Proportion 0.001 0.010 0.046 0.079 0.256 0.298 0.254 0.043 0.009 0.001 0.002
## ---------------------------------------------------------------------------
## BMI
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2152        0     1072        1    28.77    7.223    20.18    21.50
##      .25      .50      .75      .90      .95
##    24.00    27.60    32.00    37.36    41.22
##
## lowest : 15.02 15.80 15.98 16.51 16.70, highest: 62.80 63.30 63.91 67.83 69.00
## ---------------------------------------------------------------------------
## DirectChol
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2152        0       98    0.999    1.346   0.4446     0.80     0.91
##      .25      .50      .75      .90      .95
##     1.06     1.29     1.58     1.89     2.09
##
## lowest : 0.39 0.41 0.52 0.54 0.57, highest: 3.13 3.41 3.44 3.59 3.83
## ---------------------------------------------------------------------------
## Age
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2152        0       40    0.999    39.18    13.08       21       23
##      .25      .50      .75      .90      .95
##       30       39       49       55       57
##
## lowest : 20 21 22 23 24, highest: 55 56 57 58 59
## ---------------------------------------------------------------------------
## Gender
##        n  missing distinct     Info      Sum     Mean      Gmd
##     2152        0        2    0.747     1011   0.4698   0.4984
##
## ---------------------------------------------------------------------------
## Race1
##        n  missing distinct     Info     Mean      Gmd
##     2152        0        5    0.758    3.428    1.115
##
## lowest : 1 2 3 4 5, highest: 1 2 3 4 5
##
```

```
## Value          1     2     3     4     5
## Frequency    289   145   230  1333   155
## Proportion 0.134 0.067 0.107 0.619 0.072
## ---------------------------------------------------------------------------
## TotChol
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2152        0      208        1    5.069    1.151     3.57     3.85
##      .25      .50      .75      .90      .95
##     4.32     4.99     5.69     6.36     6.83
##
## lowest :  1.53  2.69  2.74  2.79  2.82, highest:  9.31  9.34  9.90 12.28 13.65
## ---------------------------------------------------------------------------
## BPDiaAve
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2152        0       84    0.999    71.19    12.83       53       57
##      .25      .50      .75      .90      .95
##       64       71       78       85       89
##
## lowest :   0  20  21  22  25, highest: 108 109 110 114 116
## ---------------------------------------------------------------------------
## BPSysAve
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2152        0       98    0.999    117.4    15.44       97      101
##      .25      .50      .75      .90      .95
##      108      116      125      134      142
##
## lowest :  78  83  84  85  86, highest: 182 184 191 202 209
## ---------------------------------------------------------------------------
## AlcoholYear
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2152        0       56    0.993    70.59     91.9        0        0
##      .25      .50      .75      .90      .95
##        4       24      104      208      260
##
## lowest :   0   1   2   3   4, highest: 260 300 312 360 364
## ---------------------------------------------------------------------------
## Poverty
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2152        0      393    0.988    2.841    1.931    0.340    0.660
##      .25      .50      .75      .90      .95
##    1.277    2.780    4.817    5.000    5.000
##
## lowest : 0.00 0.02 0.03 0.04 0.05, highest: 4.95 4.96 4.97 4.99 5.00
## ---------------------------------------------------------------------------
## SexNumPartnLife
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2152        0       81    0.995    16.73    22.47        1        1
##      .25      .50      .75      .90      .95
##        3        7       15       30       50
##
## lowest :    0    1    2    3    4, highest:  600  800  999 1000 2000
## ---------------------------------------------------------------------------
## SexNumPartYear
##        n  missing distinct     Info     Mean      Gmd      .05      .10
```

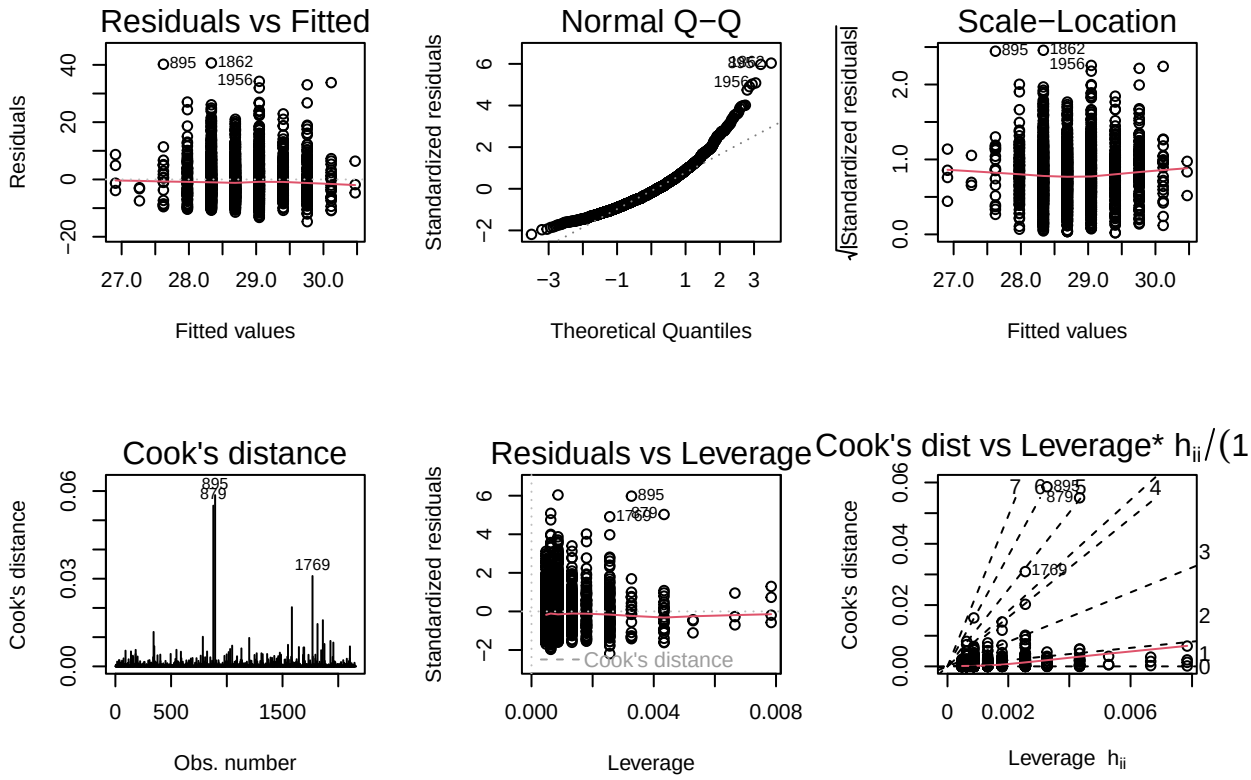
```
##     2152        0       21    0.645    1.381     1.18        0        0
##      .25      .50      .75      .90      .95
##        1        1        1        2        3
##
## lowest :  0  1  2  3  4, highest: 19 20 30 50 69
## ------------------------------------------------------------------------------
## DaysMentHlthBad
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2152        0       28    0.844    4.475    6.894        0        0
##      .25      .50      .75      .90      .95
##        0        0        5       15       30
##
## lowest :  0  1  2  3  4, highest: 25 26 27 29 30
## ------------------------------------------------------------------------------
## UrineFlow1
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2152        0     1337        1    1.074   0.9061   0.1960   0.2775
##      .25      .50      .75      .90      .95
##   0.4580   0.8100   1.3618   2.1929   2.7780
##
## lowest :  0.000  0.006  0.011  0.014  0.016, highest:  7.325  7.826  8.730  9.410 10.143
## ------------------------------------------------------------------------------
## PhysActive
##        n  missing distinct     Info      Sum     Mean      Gmd
##     2152        0        2    0.731     1246    0.579   0.4877
##
## ------------------------------------------------------------------------------
## DaysPhysHlthBad
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2152        0       24    0.708    3.165    5.318     0.00     0.00
##      .25      .50      .75      .90      .95
##     0.00     0.00     2.00    10.00    24.45
##
## lowest :  0  1  2  3  4, highest: 24 25 26 28 30
## ------------------------------------------------------------------------------
## Smoke100
##        n  missing distinct     Info      Sum     Mean      Gmd
##     2152        0        2    0.746      997   0.4633   0.4975
##
## ------------------------------------------------------------------------------
## Depressed
##        n  missing distinct
##     2152        0        3
##
## Value         None Several    Most
## Frequency     1657     355     140
## Proportion   0.770   0.165   0.065
## ------------------------------------------------------------------------------
## HealthGen
##        n  missing distinct
##     2152        0        5
##
## lowest : Excellent Vgood     Good      Fair      Poor
## highest: Excellent Vgood     Good      Fair      Poor
```

```
##
## Value      Excellent     Vgood      Good      Fair      Poor
## Frequency        240       697       854       313        48
## Proportion     0.112     0.324     0.397     0.145     0.022
## ------------------------------------------------------------------------------
## SexAge
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2152        0       28    0.985     17.1    3.463    13.00    14.00
##      .25      .50      .75      .90      .95
##    15.00    17.00    18.00    21.00    23.45
##
## lowest :  9 10 11 12 13, highest: 32 34 35 37 44
## ------------------------------------------------------------------------------
```

```r
################ (3) linear regression model ##########################################
##simple linear regression##
model1 = lm(df3$BMI ~ df3$SleepHrsNight, data = df3)
summary(model1)
```

```
##
## Call:
## lm(formula = df3$BMI ~ df3$SleepHrsNight, data = df3)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -14.74  -4.65  -1.14   3.31  40.67
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        31.1848     0.7660  40.712  < 2e-16 ***
## df3$SleepHrsNight  -0.3563     0.1109  -3.213  0.00133 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.739 on 2150 degrees of freedom
## Multiple R-squared:  0.004778,   Adjusted R-squared:  0.004315
## F-statistic: 10.32 on 1 and 2150 DF,  p-value: 0.001334
```

```r
par(mfrow = c(2, 3)) #read more from ?plot.lm
plot(model1, which = 1)
plot(model1, which = 2)
plot(model1, which = 3)
plot(model1, which = 4)
plot(model1, which = 5)
plot(model1, which = 6)
```




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


age_quant = quantile(df3$Age)
df3$AgeC = 0
df3$AgeC[df3$Age > age_quant[2] & df3$Age <= age_quant[3]] = 1
df3$AgeC[df3$Age > age_quant[3] & df3$Age <= age_quant[4]] = 2
df3$AgeC[df3$Age > age_quant[4]] = 3


### multiple linear regression###
# model_1 add demographic
m_1= lm(BMI ~ SleepHrsNight + Age + Gender + factor(Race1), df3)
summary(m_1)
```

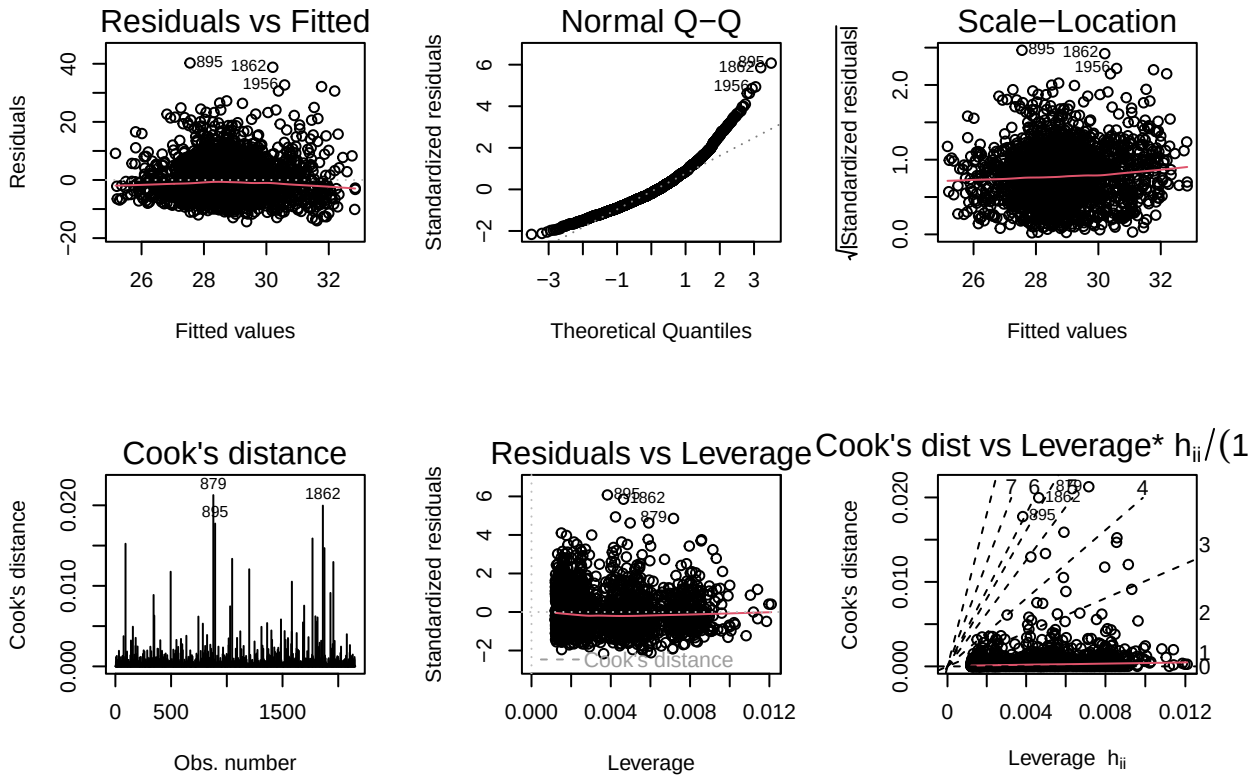
```
##
## Call:
## lm(formula = BMI ~ SleepHrsNight + Age + Gender + factor(Race1),
##     data = df3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.347  -4.497  -1.201   3.190  40.277
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)    30.78080    0.97780  31.480  < 2e-16 ***
## SleepHrsNight  -0.29383    0.11031  -2.664 0.007785 **
## Age             0.05055    0.01282   3.944 8.26e-05 ***
## Gender          0.25869    0.28895   0.895 0.370740
## factor(Race1)2 -2.28054    0.67704  -3.368 0.000769 ***
## factor(Race1)3 -1.02309    0.59140  -1.730 0.083782 .
## factor(Race1)4 -2.51942    0.43385  -5.807 7.30e-09 ***
## factor(Race1)5 -4.14341    0.66274  -6.252 4.88e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.643 on 2144 degrees of freedom
## Multiple R-squared:  0.03564,    Adjusted R-squared:  0.03249
## F-statistic: 11.32 on 7 and 2144 DF,  p-value: 3.698e-14
```

```
car::Anova(m_1,type="III")
```

```
## Anova Table (Type III tests)
##
## Response: BMI
##               Sum Sq   Df  F value    Pr(>F)
## (Intercept)    43731    1 990.9782 < 2.2e-16 ***
## SleepHrsNight    313    1   7.0957  0.007785 **
## Age              687    1  15.5578 8.259e-05 ***
## Gender            35    1   0.8015  0.370740
## factor(Race1)   2413    4  13.6678 5.235e-11 ***
## Residuals      94612 2144
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# age centered + quadratic term of age
df3$Age.c=Age=median(df3$Age,na.rm=T)
m_1.2= lm(BMI ~ SleepHrsNight + Age.c+I(Age.c^2) + Gender + factor(Race1), df3)



########### model 1 diagnosis ###########

par(mfrow = c(2, 3)) #read more from ?plot.lm
plot(m_1, which = 1)
plot(m_1, which = 2)
plot(m_1, which = 3)
plot(m_1, which = 4)
plot(m_1, which = 5)
plot(m_1, which = 6)
```




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



m_1.yhat=m_1$fitted.values
m_1.res=m_1$residuals
m_1.h=hatvalues(m_1)
m_1.r=rstandard(m_1)
m_1.rr=rstudent(m_1)
#which subject is most outlying with respect to the x space
Hmisc::describe(m_1.h)
```

```
## m_1.h
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2152        0     1023        1 0.003717 0.002568 0.001335 0.001433
##      .25      .50      .75      .90      .95
## 0.001727 0.002591 0.005301 0.007614 0.008238
##
## lowest : 0.001205683 0.001206922 0.001211889 0.001215604 0.001225540
## highest: 0.011404275 0.011408686 0.011854168 0.011975653 0.012091402
```

```
m_1.h[which.max(m_1.h)]
```

```
##       325
## 0.0120914
```

```
length(df3$Age)
```

```
## [1] 2152
```

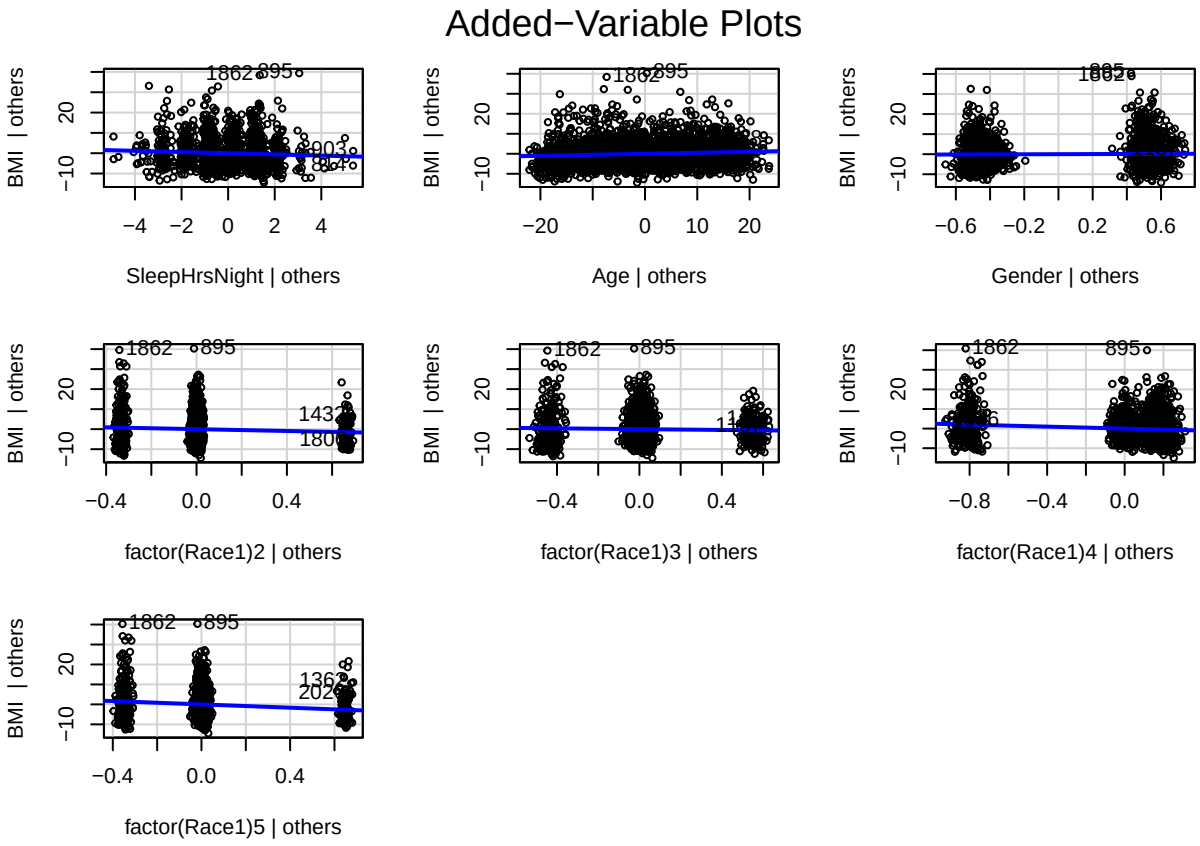
```
length(df3$BMI)
```

```
## [1] 2152
```

```
length(m_1.yhat)# why the length of yhat is diff with y
```

```
## [1] 2152
```

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

#(1)Linear: 2 approaches

# partial regression plots
car::avPlots(m_1)
```



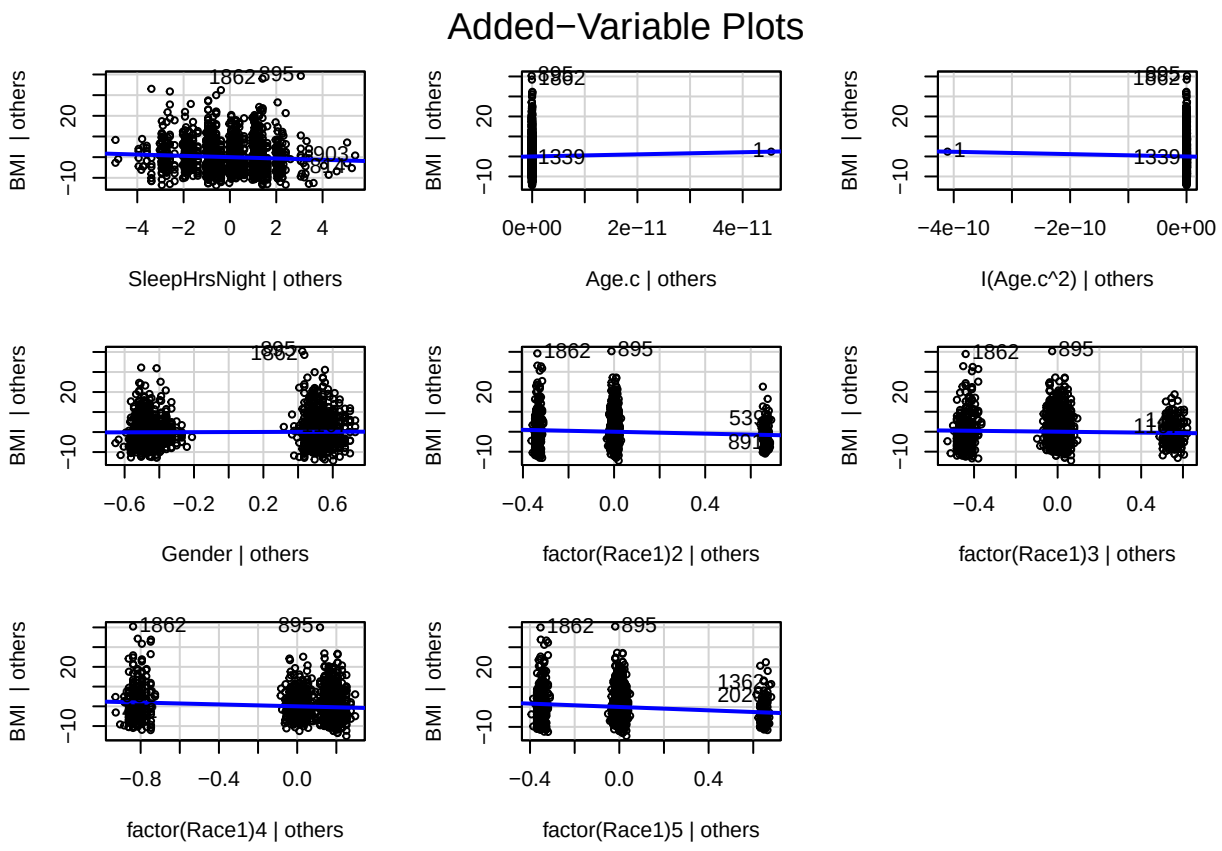

```
#age a set of quartiles
car::avPlots(m_1.2)
```

```
## Warning in lsfit(mod.mat[, -var], cbind(mod.mat[, var], response), wt = wt, :
## 'X' matrix was collinear
```

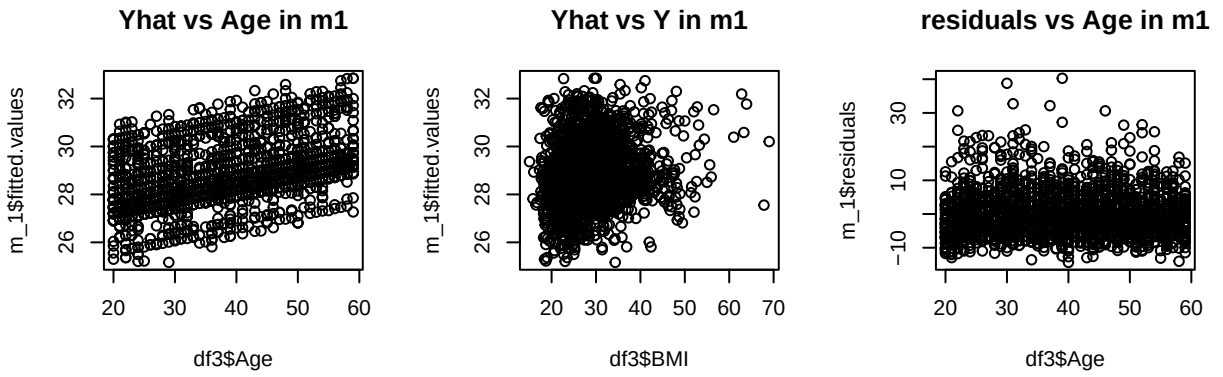
```
## Warning in lsfit(mod.mat[, -var], cbind(mod.mat[, var], response), wt = wt, :
## 'X' matrix was collinear
```

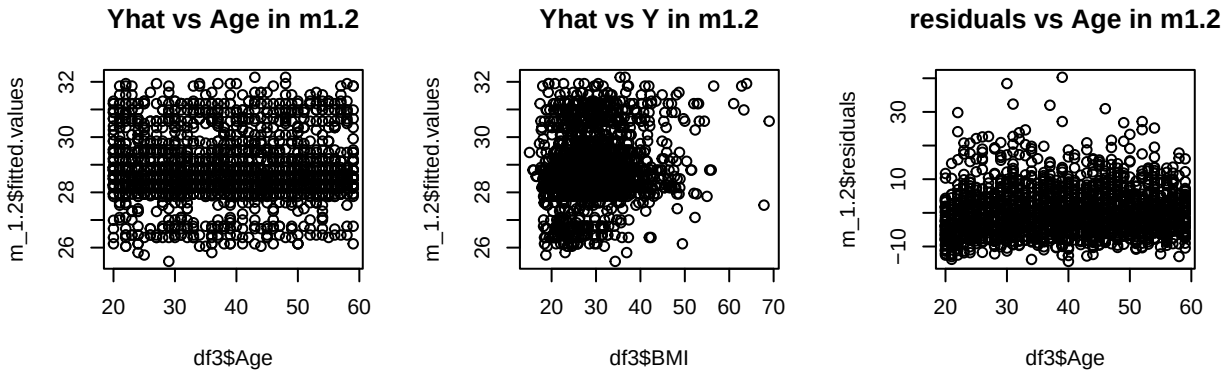
```
## Warning in lsfit(mod.mat[, -var], cbind(mod.mat[, var], response), wt = wt, :
## 'X' matrix was collinear
```

```
## Warning in lsfit(mod.mat[, -var], cbind(mod.mat[, var], response), wt = wt, :
## 'X' matrix was collinear

## Warning in lsfit(mod.mat[, -var], cbind(mod.mat[, var], response), wt = wt, :
## 'X' matrix was collinear

## Warning in lsfit(mod.mat[, -var], cbind(mod.mat[, var], response), wt = wt, :
## 'X' matrix was collinear

## Warning in lsfit(mod.mat[, -var], cbind(mod.mat[, var], response), wt = wt, :
## 'X' matrix was collinear

## Warning in lsfit(mod.mat[, -var], cbind(mod.mat[, var], response), wt = wt, :
## 'X' matrix was collinear
```



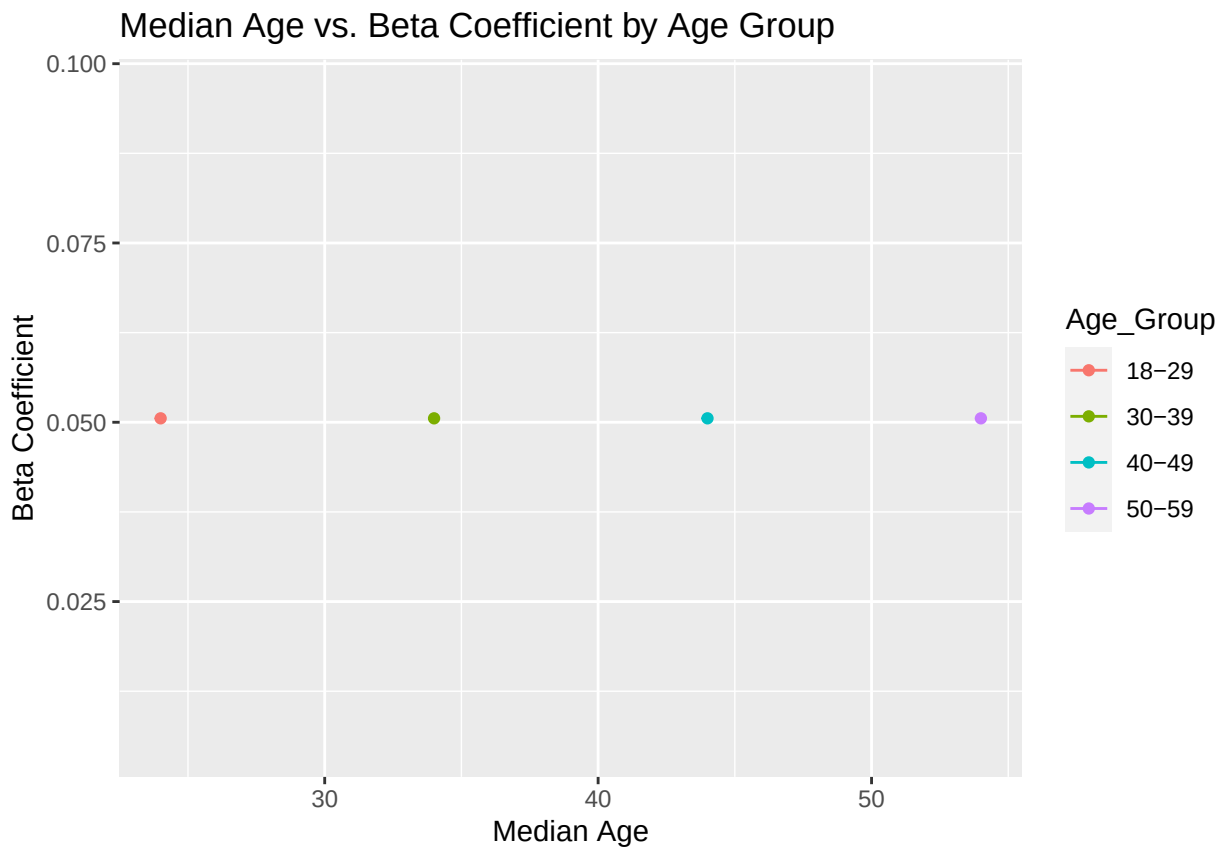

```
#contain quadratic term of age
par(mfrow=c(2,3))
plot(x=df3$Age,y=m_1$fitted.values,main="Yhat vs Age in m1")
plot(x=df3$BMI,y=m_1$fitted.values,main="Yhat vs Y in m1")
plot(x=df3$Age,y=m_1$residuals,main="residuals vs Age in m1")
plot(x=df3$Age,y=m_1.2$fitted.values,main="Yhat vs Age in m1.2")
plot(x=df3$BMI,y=m_1.2$fitted.values,main="Yhat vs Y in m1.2")
plot(x=df3$Age,y=m_1.2$residuals,main="residuals vs Age in m1.2")
```



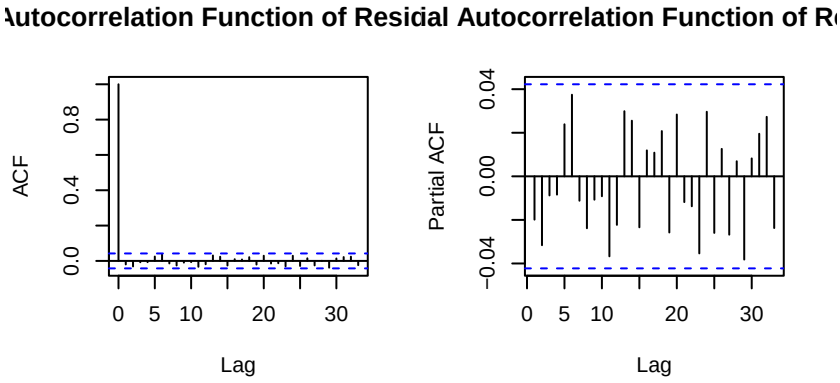


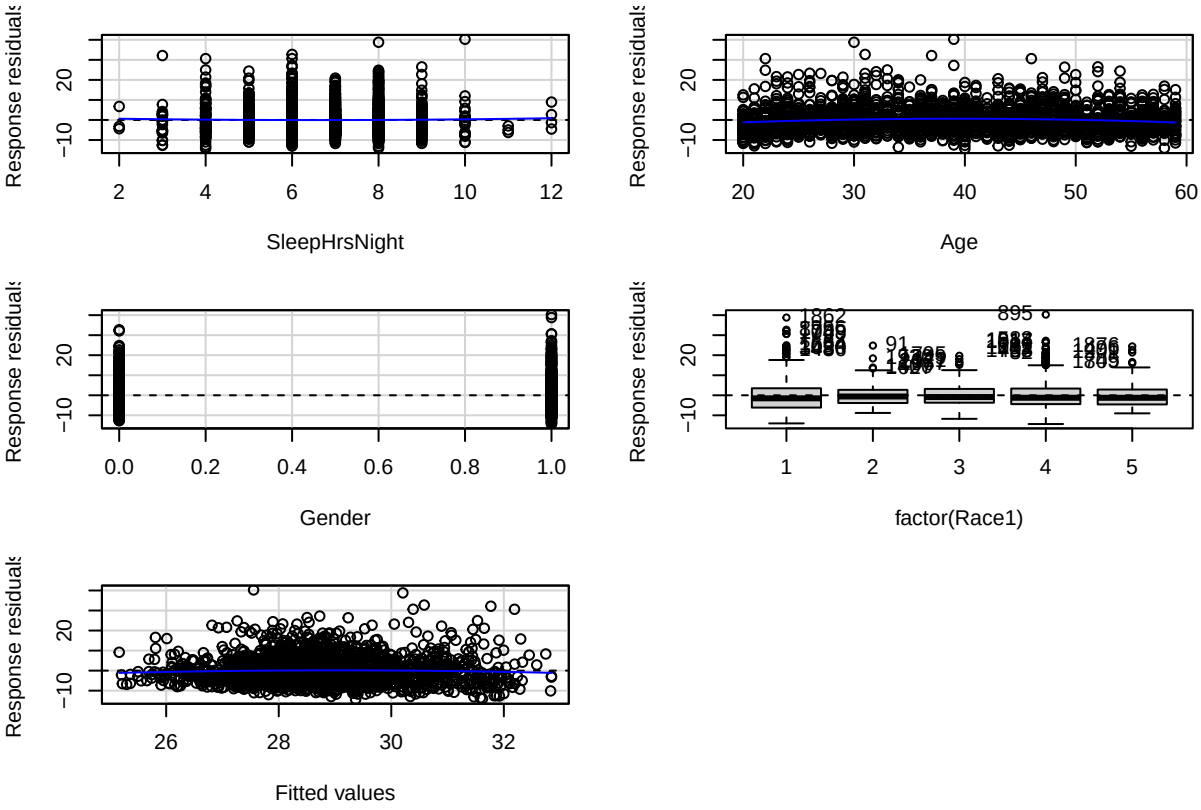

```
#categoraize age ---beta plot
df3 <- df3 %>%
  mutate(Age_Group = cut(Age, breaks = c(18, 29, 39, 49, 59), labels = c("18-29", "30-39", "40-49", "50-

summary_stats <- df3 %>%
  group_by(Age_Group) %>%
  summarise(Median_Age = median(Age), Beta_Coefficient = coef(m_1)['Age'])

ggplot(summary_stats, aes(x = Median_Age, y = Beta_Coefficient, group = Age_Group, color = Age_Group)) +
  geom_line() +
  geom_point() +
  labs(title = "Median Age vs. Beta Coefficient by Age Group",
       x = "Median Age",
       y = "Beta Coefficient")
```

```
## `geom_line()`: Each group consists of only one observation.
## i Do you need to adjust the group aesthetic?
```




```
#(2)Independence:

residuals <- resid(m_1)
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_1,type="response")
```


Autocorrelation Function of Resid
Autocorrelation Function of R
ACF
0.0
0.4
0.8
0
5
10
20
30
Lag
Partial ACF
−0.04
0.00
0.04
0
5
10
20
30
Lag




```
##                Test stat Pr(>|Test stat|)
## SleepHrsNight    0.6210          0.53467
## Age             -4.0097        6.287e-05 ***
## Gender          -0.0409          0.96735
## factor(Race1)
## Tukey test      -2.1001          0.03572 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(m_1, which = 1)
#or
ggplot(m_1, aes(x = m_1.yhat, y = m_1.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_1.res)
```

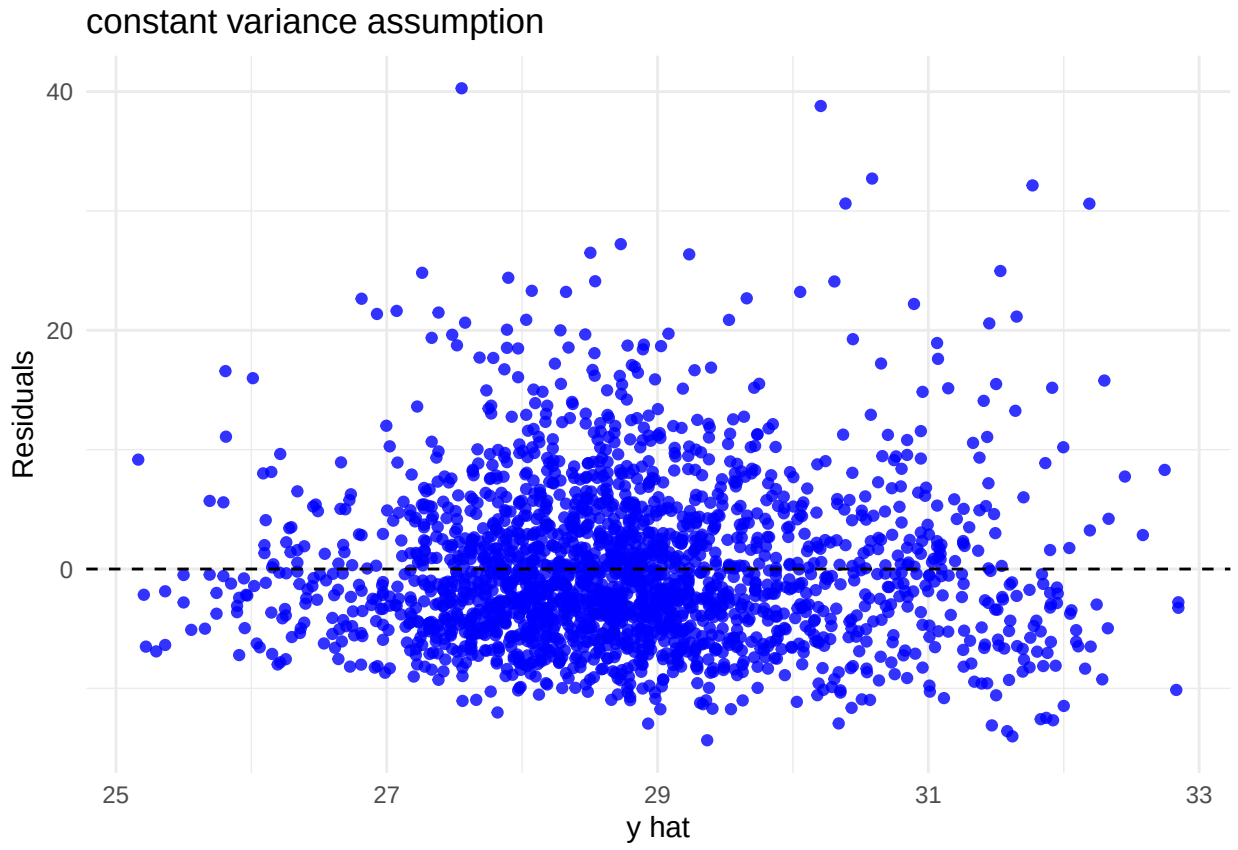
```
## m_1.res
```

```
##          n    missing   distinct       Info       Mean        Gmd        .05
##       2152          0       2148          1 -1.789e-16      7.089     -8.376
##        .10        .25        .50        .75        .90        .95
##     -7.067     -4.497     -1.201      3.190      8.316     12.176
##
## lowest : -14.34683 -14.02086 -13.58293 -13.09920 -12.95082
## highest:  30.62135  32.14027  32.71507  38.79459  40.27670
```

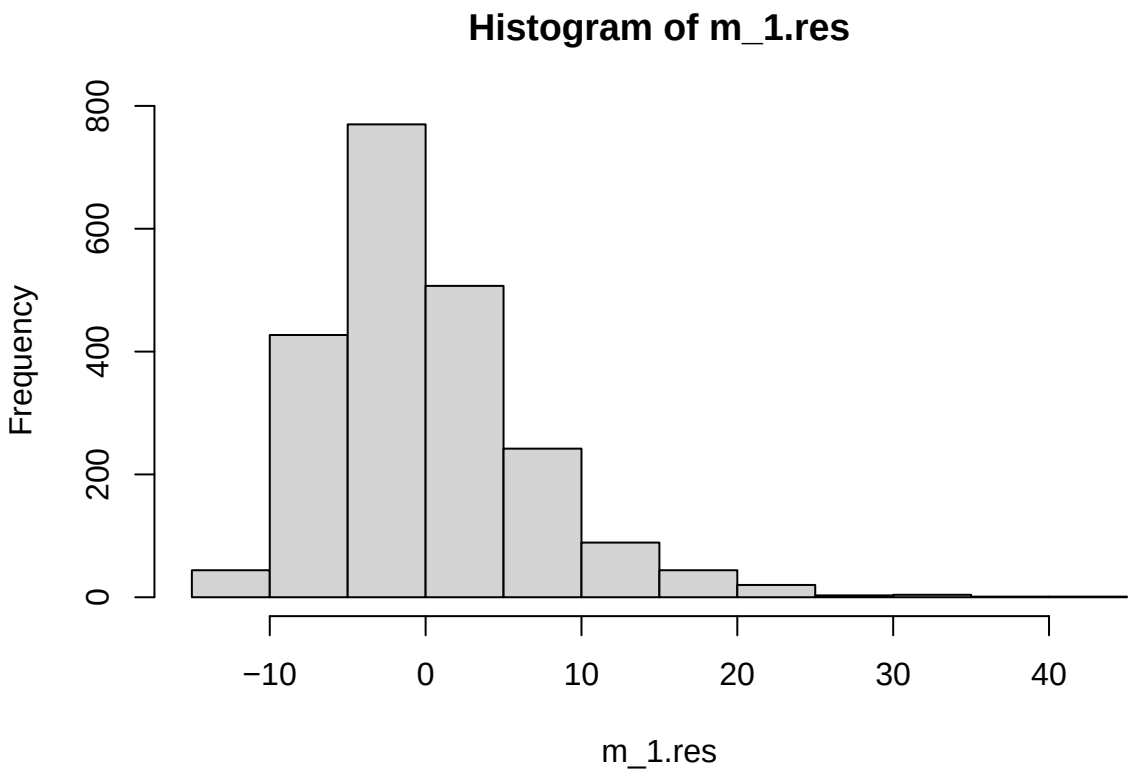
```
Hmisc::describe(m_1.res)$counts[c(".25",".50",".75")] #not symmetric
```

```
##      .25      .50      .75
## "-4.497" "-1.201" " 3.190"
```

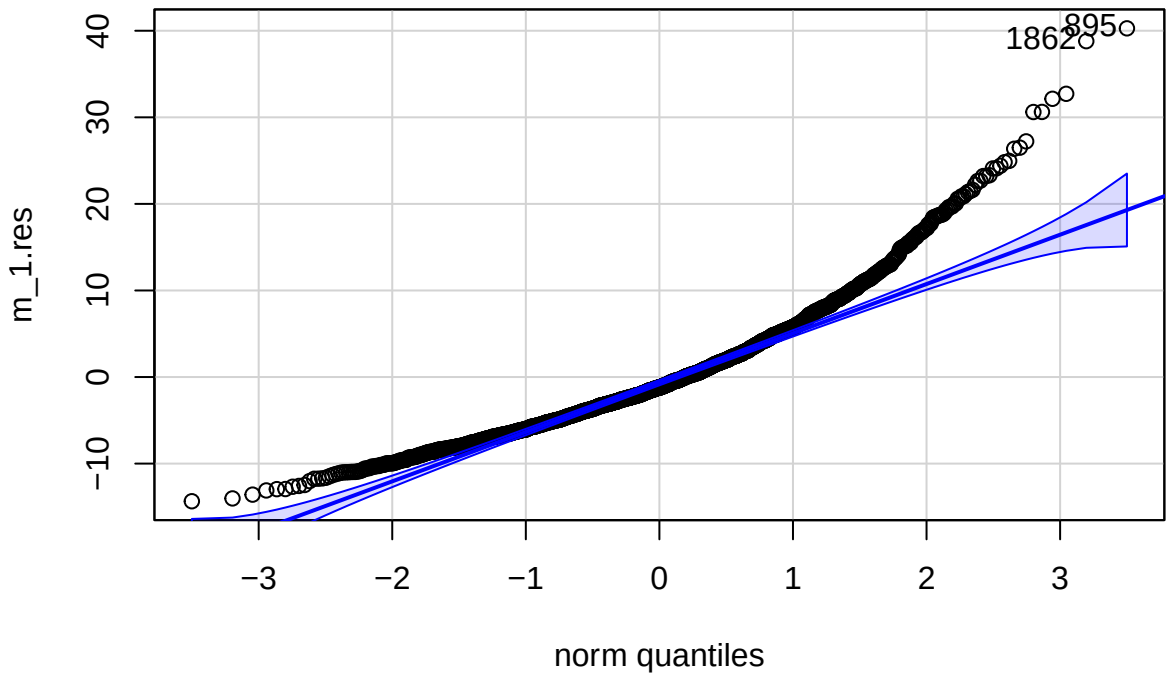
```
#histogram
par(mfrow = c(1, 1))
```




```
hist(m_1.res,breaks = 15)
```




```
# Q-Q plot
qq.m_1.res=car::qqPlot(m_1.res)
```



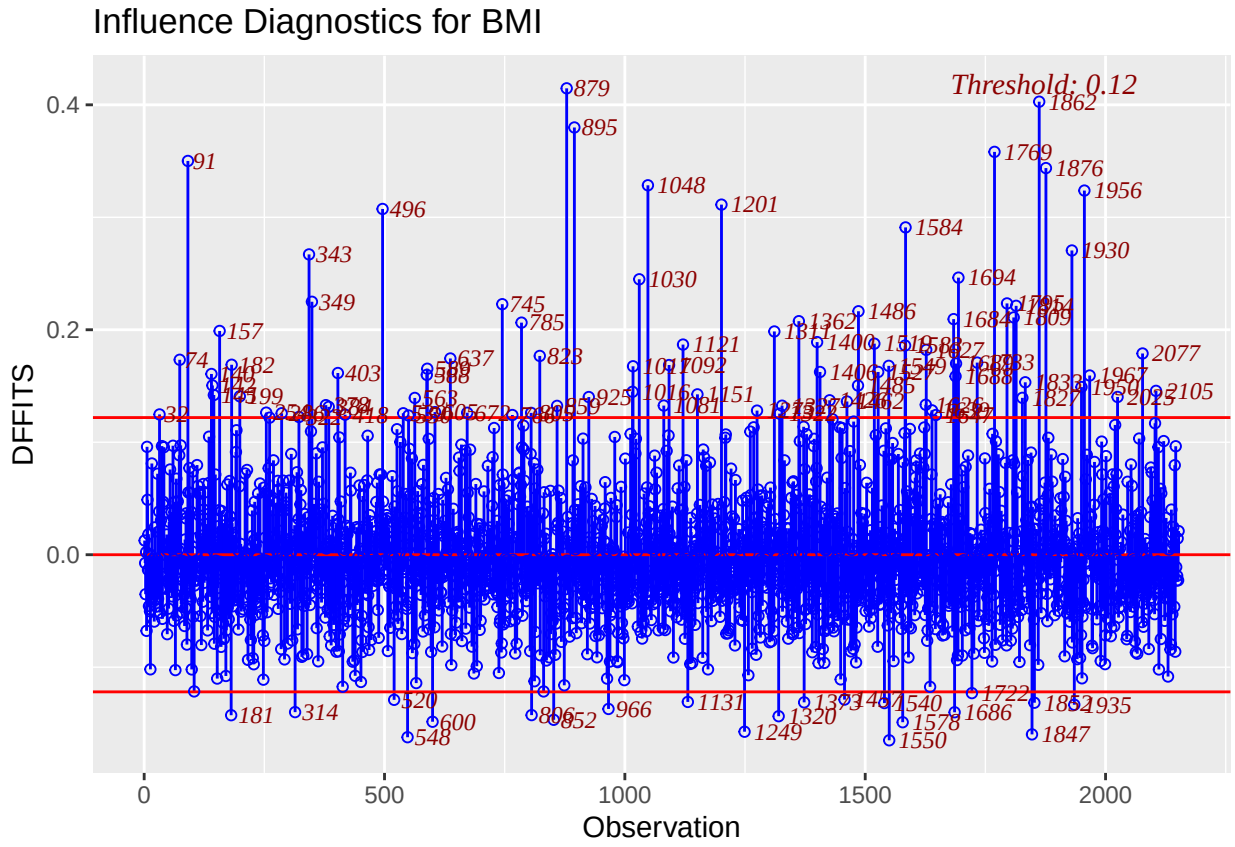

```
m_1.res[qq.m_1.res]
```

```
##      895     1862
## 40.27670 38.79459
```

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

influence = data.frame(Residual = resid(m_1), Rstudent = rstudent(m_1),
                       HatDiagH = hat(model.matrix(m_1)),
                       CovRatio = covratio(m_1), DFFITS = dffits(m_1),
                       COOKsDistance = cooks.distance(m_1))
# DFFITS
library(olsrr)
```

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

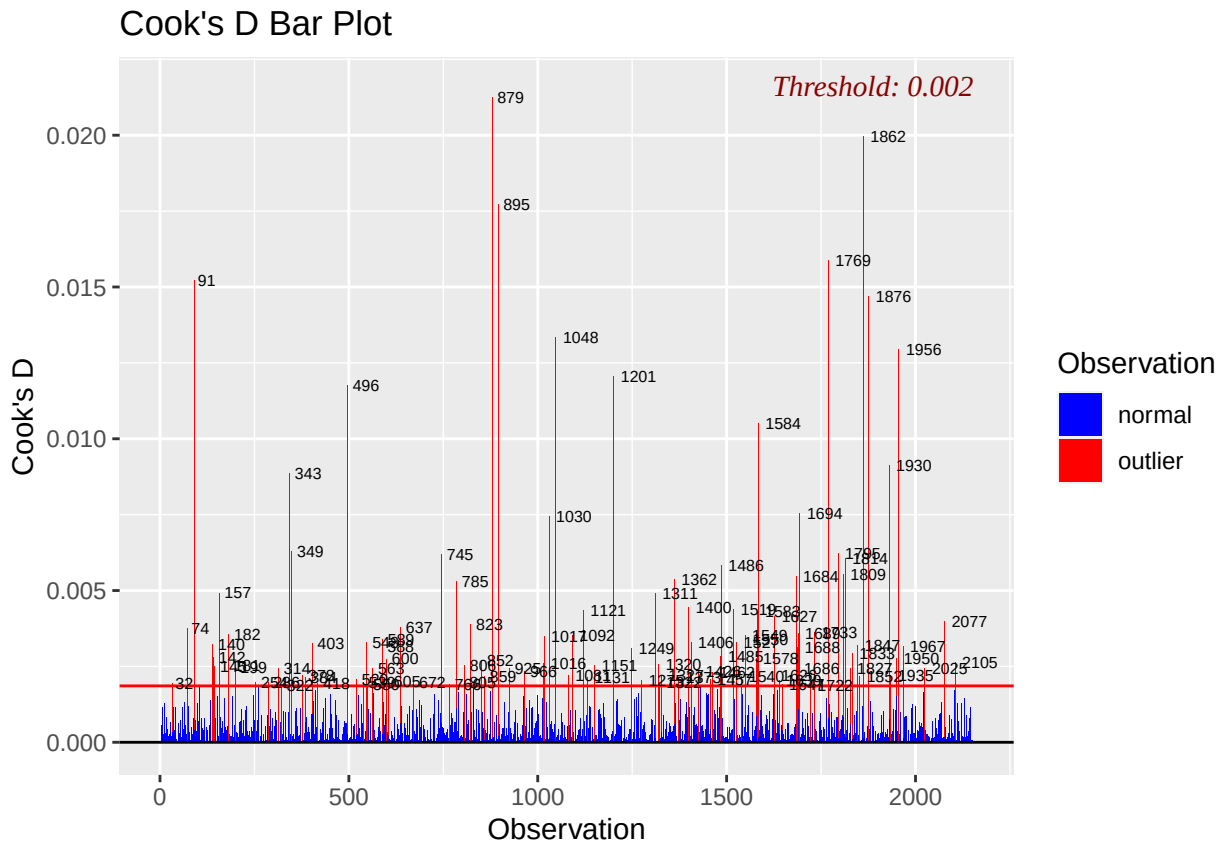
```
ols_plot_dffits(m_1)
```




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

```
##       Residual Rstudent    HatDiagH  CovRatio    DFFITS COOKsDistance
## 879   32.14027 4.881458 0.007163982 0.9253887 0.4146561    0.02126602
## 1862  38.79459 5.899523 0.004639616 0.8864760 0.4027798    0.01996418
## 895   40.27670 6.126250 0.003829692 0.8769334 0.3798480    0.01773341
## 1769  30.61045 4.643754 0.005913164 0.9319868 0.3581517    0.01588175
## 91    24.81823 3.763655 0.008578651 0.9604459 0.3500985    0.01522761
## 1876  24.40196 3.700081 0.008557096 0.9621168 0.3437483    0.01468345
```

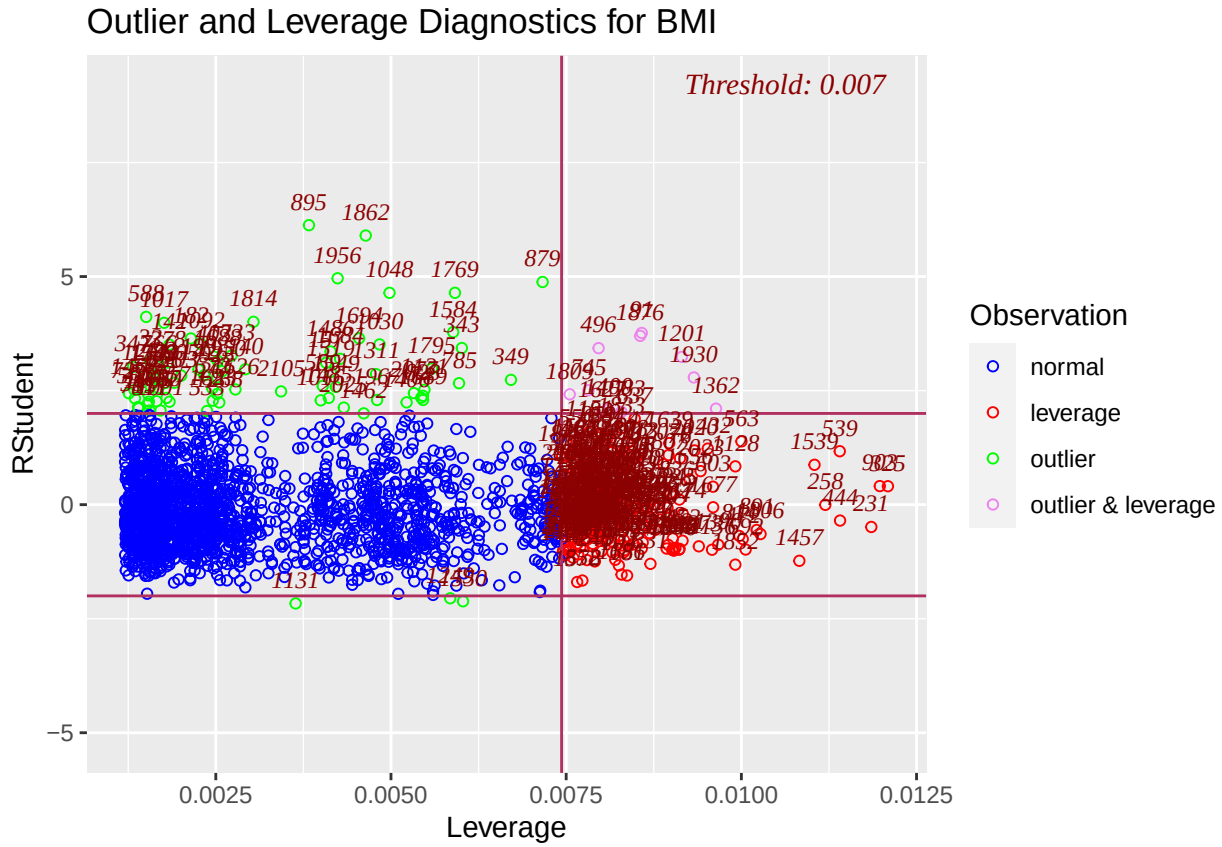
```
#From the plot above, we can see 2 observations with the largest (magnitude) of DFFITS, observation 879

# Cook's D
ols_plot_cooksd_bar(m_1)
```




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

```
##      Residual Rstudent    HatDiagH  CovRatio    DFFITS COOKsDistance
## 879  32.14027 4.881458 0.007163982 0.9253887 0.4146561    0.02126602
## 1862 38.79459 5.899523 0.004639616 0.8864760 0.4027798    0.01996418
## 895  40.27670 6.126250 0.003829692 0.8769334 0.3798480    0.01773341
## 1769 30.61045 4.643754 0.005913164 0.9319868 0.3581517    0.01588175
## 91   24.81823 3.763655 0.008578651 0.9604459 0.3500985    0.01522761
## 1876 24.40196 3.700081 0.008557096 0.9621168 0.3437483    0.01468345
```

```
#From the plot above, we can see that the observation 879 and 1862 also have the largest Cook's Distanc

#leverage
ols_plot_resid_lev(m_1)
```




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

```
##       Residual     Rstudent   HatDiagH CovRatio        DFFITS COOKsDistance
## 325  2.6701881  0.404331534 0.01209140 1.015404  0.0447319220  2.502157e-04
## 903  2.6888017  0.407126452 0.01197565 1.015277  0.0448223986  2.512287e-04
## 231 -3.2298449 -0.489027128 0.01185417 1.014874 -0.0535621462  3.587402e-04
## 444 -2.3031224 -0.348624992 0.01140869 1.014862 -0.0374513890  1.753977e-04
## 539  7.7332215  1.170920877 0.01140428 1.010136  0.1257627521  1.976692e-03
## 258 -0.0252608 -0.003823227 0.01119827 1.015107 -0.0004068656  2.070211e-08
```

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

```
##      Residual Rstudent    HatDiagH  CovRatio    DFFITS COOKsDistance
## 895  40.27670 6.126250 0.003829692 0.8769334 0.3798480    0.01773341
## 1862 38.79459 5.899523 0.004639616 0.8864760 0.4027798    0.01996418
## 1956 32.71507 4.962366 0.004238569 0.9199614 0.3237582    0.01295962
## 879  32.14027 4.881458 0.007163982 0.9253887 0.4146561    0.02126602
## 1769 30.61045 4.643754 0.005913164 0.9319868 0.3581517    0.01588175
## 1048 30.62135 4.643222 0.004979030 0.9311288 0.3284548    0.01335724
```

```
#From the plot above, we can see that the observation 325 has the largest leverage (0.0121). Observatio
```

```
#From the plot above, there is 11 observations(1809,745,496, 1876, 91, 1201, 1930, 1362, 1627, 1583,140
#The thresholds for the externally studentized residual are -2 and 2, i.e. 2 in magnitude. The threshol
```

```
#From (DFFITS), observations 879 and 1862 appear to be influential observations. Observation 325 has ex

rm.df3 = df3[-c(879,1862,325,1809,745,496, 1876, 91, 1201, 1930, 1362, 1627, 1583,1400),]
rm.m_1 = lm(BMI ~ SleepHrsNight + Age + Gender + factor(Race1), rm.df3)
## Before removing these observations, the estimated coefficients are:
summary(m_1)$coef
```

```
##                   Estimate Std. Error    t value      Pr(>|t|)
## (Intercept)    30.78080019 0.97779508 31.4798068 3.939805e-179
## SleepHrsNight  -0.29382901 0.11030542 -2.6637766  7.784737e-03
## Age             0.05055175 0.01281626  3.9443440  8.258992e-05
## Gender          0.25869036 0.28895036  0.8952761  3.707400e-01
## factor(Race1)2 -2.28053615 0.67703708 -3.3684066  7.694110e-04
## factor(Race1)3 -1.02309157 0.59139692 -1.7299575  8.378177e-02
## factor(Race1)4 -2.51941843 0.43384887 -5.8071338  7.302407e-09
## factor(Race1)5 -4.14340723 0.66274194 -6.2519165  4.878127e-10
```

```
## After removing these observations, the estimated coefficients are:
summary(rm.m_1)$coef
```

```
##                   Estimate Std. Error   t value      Pr(>|t|)
## (Intercept)    30.11857029 0.95406699 31.568612 9.144596e-180
## SleepHrsNight  -0.24136738 0.10777490 -2.239551  2.522291e-02
## Age             0.05183874 0.01244376  4.165841  3.225874e-05
## Gender          0.30607721 0.28058739  1.090844  2.754647e-01
## factor(Race1)2 -2.56312594 0.66387036 -3.860883  1.163453e-04
## factor(Race1)3 -0.78603525 0.57340129 -1.370829  1.705726e-01
## factor(Race1)4 -2.29062304 0.42113470 -5.439170  5.966971e-08
## factor(Race1)5 -4.77415992 0.65242607 -7.317549  3.559809e-13
```

```
#### change percent
abs((rm.m_1$coefficients - m_1$coefficients)/(m_1$coefficients) *100)
```

```
##    (Intercept)  SleepHrsNight            Age         Gender factor(Race1)2
##       2.151438      17.854475       2.545893      18.317978      12.391375
## factor(Race1)3 factor(Race1)4 factor(Race1)5
##      23.170586       9.081278      15.223044
```

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



################   multicollinearity   ######################
#Pearson correlations
var= c("BMI","SleepHrsNight","Age","Gender","Race1")
newData = df3[,var]
library("corrplot")
```

```
## corrplot 0.92 loaded
```

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

```
         number.cex = 1,
         diag = FALSE,
         addCoef.col = "black",
         tl.col = "black",
         tl.srt = 90,
         p.mat = p.mat,
         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_1)
```

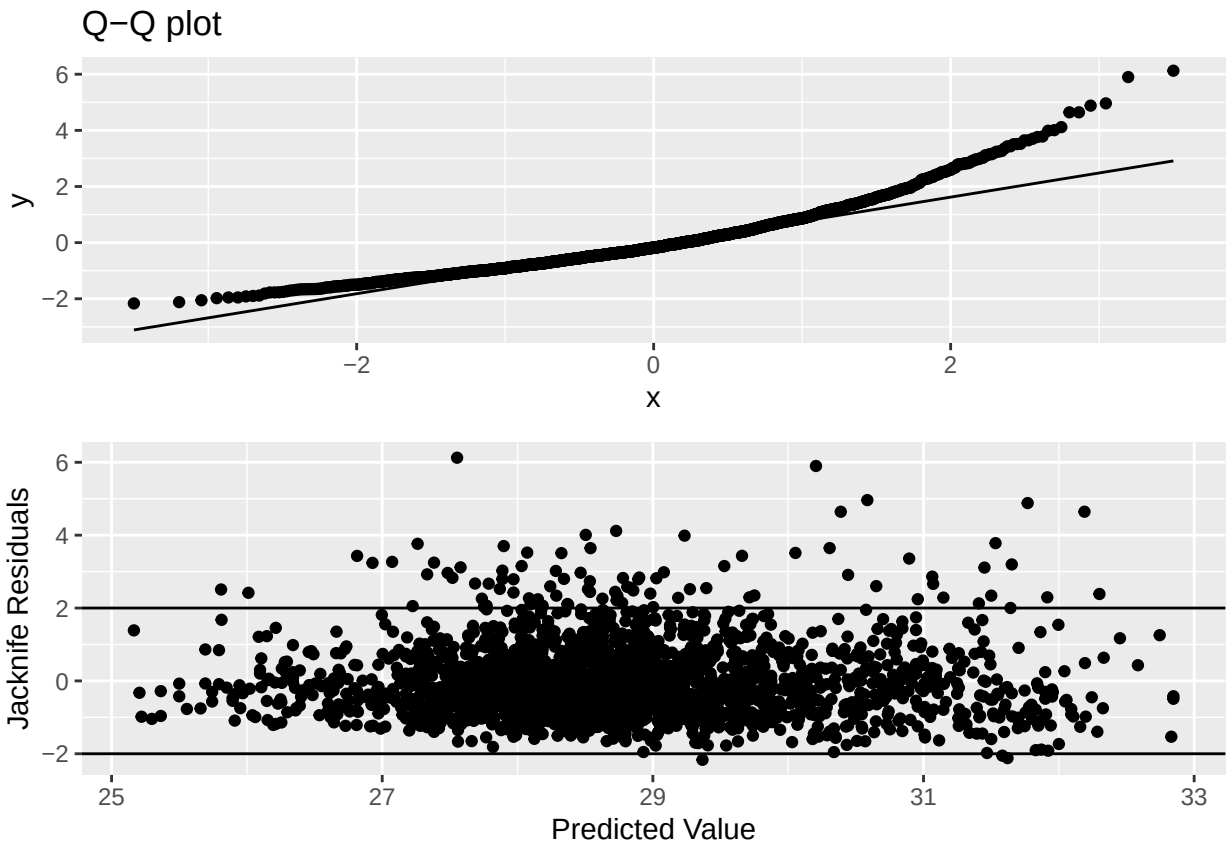
```
##                    GVIF Df GVIF^(1/(2*Df))
## SleepHrsNight 1.017942  1        1.008931
## Age           1.028310  1        1.014056
## Gender        1.014189  1        1.007069
## factor(Race1) 1.042495  4        1.005216
```

```
#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_1.log= lm(logBMI ~ SleepHrsNight + Age + Gender + factor(Race1), df3)
p11.log = ols_plot_resid_lev(m_1.log)
p12.log = ols_plot_cooksd_bar(m_1.log)
library(gridExtra)
```

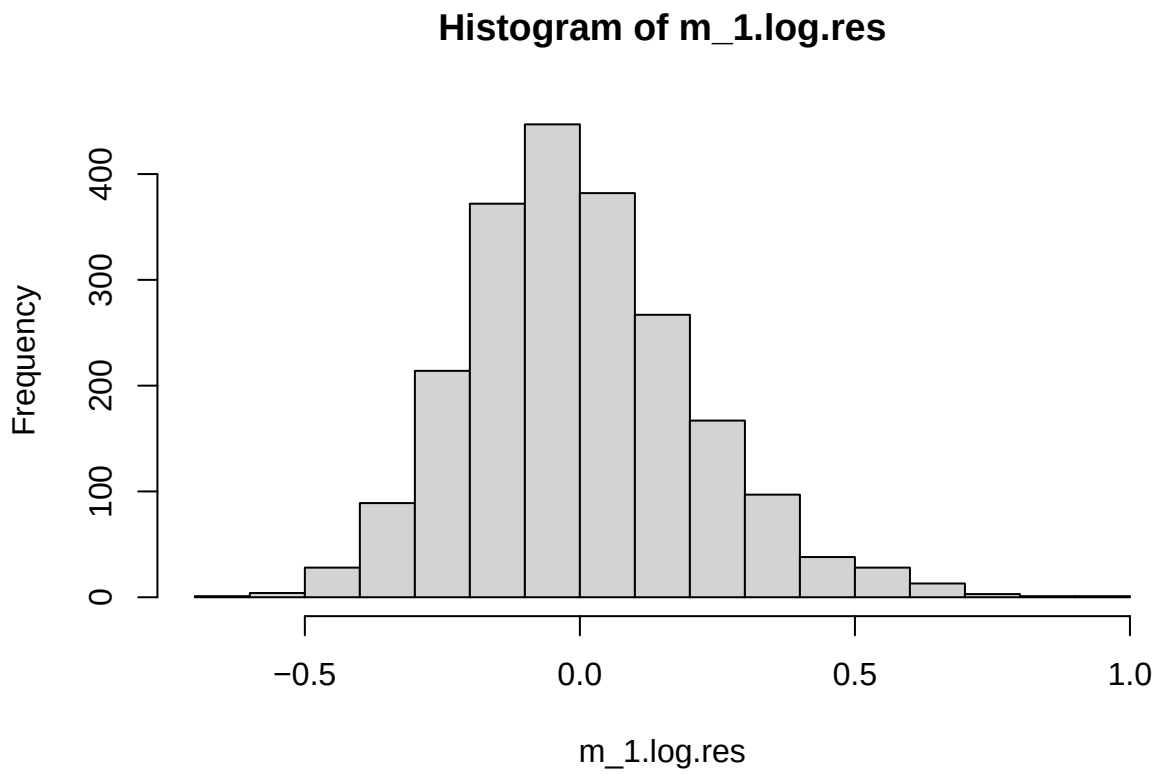
```
##
## Attaching package: 'gridExtra'
```

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

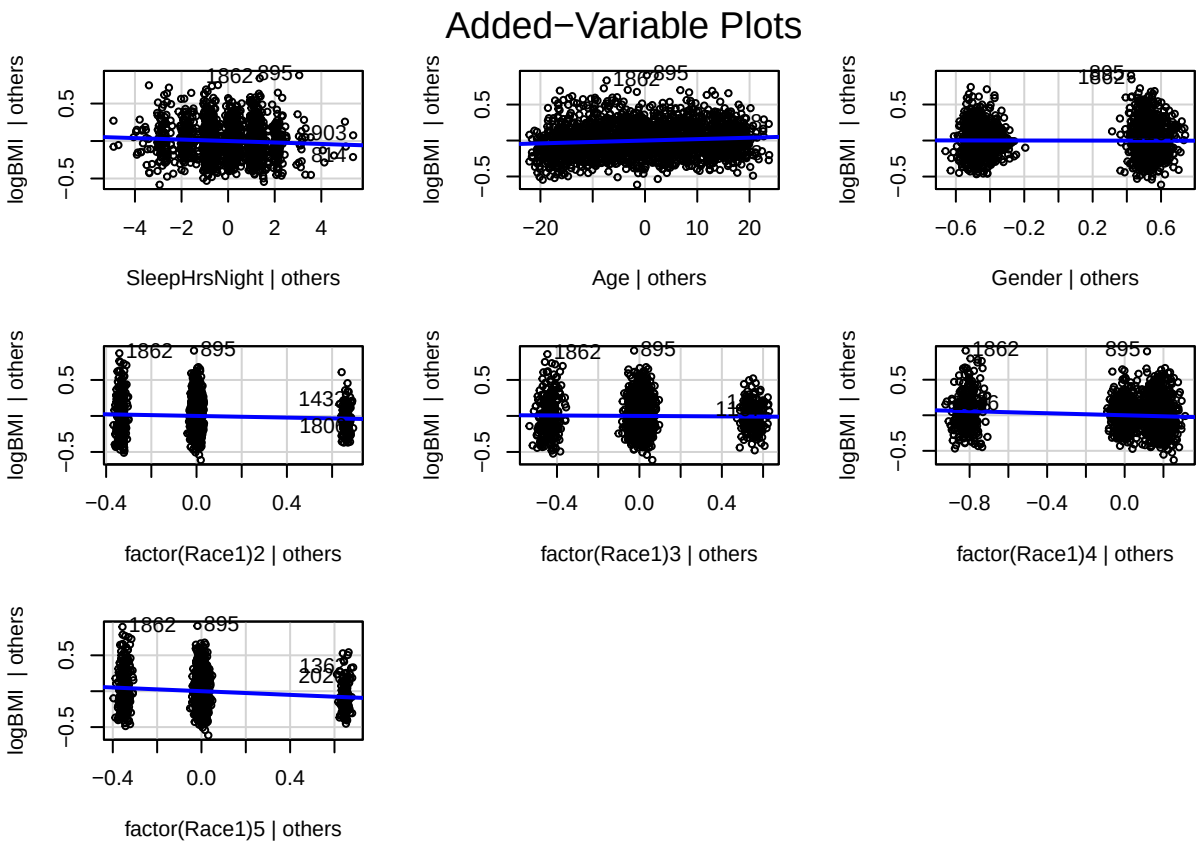
```
p13.log = ggplot(m_1.log, aes(sample = rstudent(m_1.log))) + geom_qq() + stat_qq_line() +labs(title="Q-
p14.log = ggplot() + geom_point(aes(y = rstudent(m_1.log), x = m_1.log$fitted.values )) + labs(x = "Pre
grid.arrange(p13.log,p14.log, nrow=2)


p13 = ggplot(m_1, aes(sample = rstudent(m_1))) + geom_qq() + stat_qq_line() +labs(title="Q-Q plot")
p14 = ggplot() + geom_point(aes(y = rstudent(m_1), x = m_1$fitted.values )) + labs(x = "Predicted Value
grid.arrange(p13,p14, nrow=2)


m_1.log.yhat=m_1.log$fitted.values
m_1.log.res=m_1.log$residuals
m_1.log.h=hatvalues(m_1.log)
m_1.log.r=rstandard(m_1.log)
m_1.log.rr=rstudent(m_1.log)

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



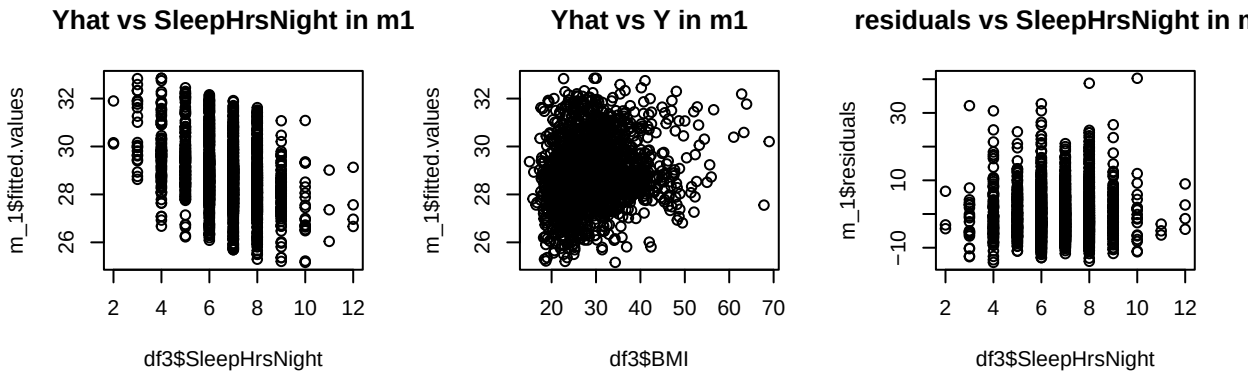

```
hist(m_1.log.res,breaks = 15)
```

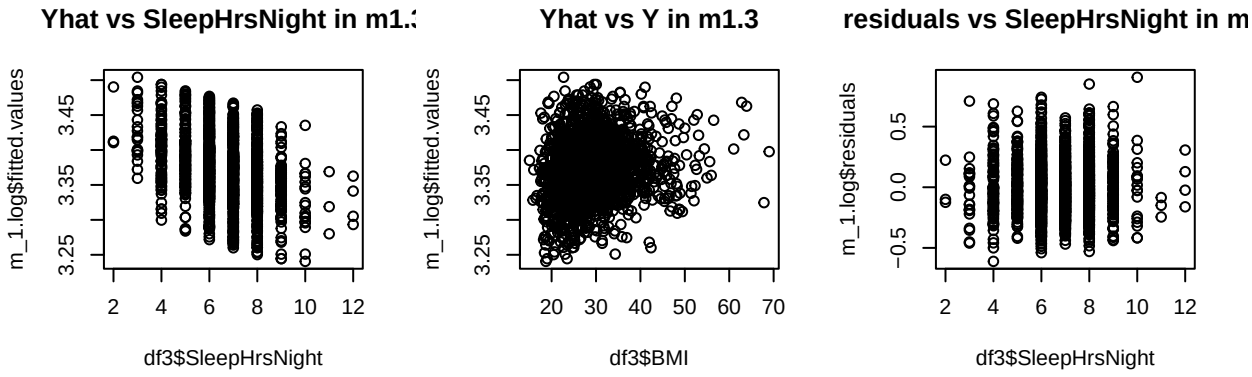



```
car::avPlots(m_1.log)
```




```
par(mfrow=c(2,3))
plot(x=df3$SleepHrsNight,y=m_1$fitted.values,main="Yhat vs SleepHrsNight in m1")
plot(x=df3$BMI,y=m_1$fitted.values,main="Yhat vs Y in m1")
plot(x=df3$SleepHrsNight,y=m_1$residuals,main="residuals vs SleepHrsNight in m1")
plot(x=df3$SleepHrsNight,y=m_1.log$fitted.values,main="Yhat vs SleepHrsNight in m1.3")
plot(x=df3$BMI,y=m_1.log$fitted.values,main="Yhat vs Y in m1.3")
plot(x=df3$SleepHrsNight,y=m_1.log$residuals,main="residuals vs SleepHrsNight in m1.3")
```






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

#collinearity diagnostics
library(Hmisc)
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:dplyr':
##
##     src, summarize
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
var= c("BMI","SleepHrsNight","Age","Gender","Race1","logBMI")
newData.log = df3[,var]
par(mfrow = c(1, 2))
cor = rcorr(as.matrix(newData.log[,-1]),type = "pearson")
corrplot(cor$r,
         method = "color",
         type = "upper",
         number.cex = 0.5,
         diag = FALSE,
         addCoef.col = "black",
         tl.col = "black",
         tl.srt = 90,
```

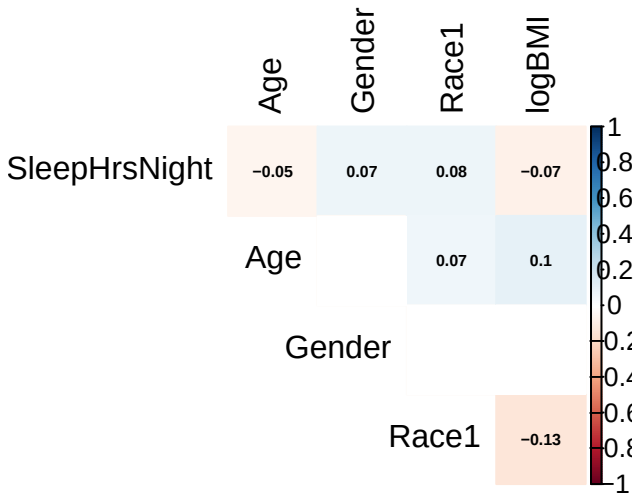
```
         p.mat = cor$P,
         sig.level = 0.05,
         insig = "blank")

car::vif(m_1.log)
```

```
##                    GVIF Df GVIF^(1/(2*Df))
## SleepHrsNight 1.017942  1        1.008931
## Age           1.028310  1        1.014056
## Gender        1.014189  1        1.007069
## factor(Race1) 1.042495  4        1.005216
```

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

```
## model_2 add known risk factors ##
m_2 = lm(
  BMI ~ SleepHrsNight +Age + Gender + Race1 + TotChol+ BPDiaAve + BPSysAve + AlcoholYear+ Smoke100 +Days
)
summary(m_2)
```

```
##
## Call:
## lm(formula = BMI ~ SleepHrsNight + Age + Gender + Race1 + TotChol +
##     BPDiaAve + BPSysAve + AlcoholYear + Smoke100 + DaysPhysHlthBad +
##     PhysActive, data = df3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.752  -4.236  -0.849   3.055  37.857
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     21.023150   1.610401  13.055  < 2e-16 ***
## SleepHrsNight   -0.212193   0.107400  -1.976 0.048314 *
## Age              0.012839   0.013495   0.951 0.341528
## Gender           0.514621   0.291331   1.766 0.077463 .
## Race1           -0.622971   0.122615  -5.081 4.09e-07 ***
## TotChol          0.076572   0.139325   0.550 0.582658
## BPDiaAve         0.054500   0.014049   3.879 0.000108 ***
## BPSysAve         0.066004   0.012027   5.488 4.55e-08 ***
## AlcoholYear     -0.009762   0.001533  -6.368 2.34e-10 ***
## Smoke100        -0.507830   0.287921  -1.764 0.077911 .
## DaysPhysHlthBad  0.066309   0.019785   3.352 0.000818 ***
## PhysActive      -1.260928   0.292769  -4.307 1.73e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.413 on 2140 degrees of freedom
## Multiple R-squared:  0.1029, Adjusted R-squared:  0.09826
## F-statistic: 22.31 on 11 and 2140 DF,  p-value: < 2.2e-16
```


Age
Gender
Race1
logBMI
SleepHrsNight
−0.05
0.07
0.08
−0.07
Age
0.07
0.1
Gender
Race1
−0.13
1
0.8
0.6
0.4
0.2
0
1