model1log.R

zhang alice

2023-11-25

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

```
##          used (Mb) gc trigger (Mb) max used (Mb)
## Ncells 469974 25.1    1020662 54.6   644240 34.5
## Vcells 856635  6.6    8388608 64.0  1634810 12.5
```

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

```
## Warning: package 'car' was built under R version 4.2.3
```

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

```
## Warning: package 'carData' was built under R version 4.2.3
```

```r
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.2.3
```

```r
############### (1) Data cleaning ########################################
## select variables
library(NHANES)
```

```
## Warning: package 'NHANES' was built under R version 4.2.3
```

```r
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"
```

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

```
## Warning: package 'dplyr' was built under R version 4.2.3
```

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

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

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

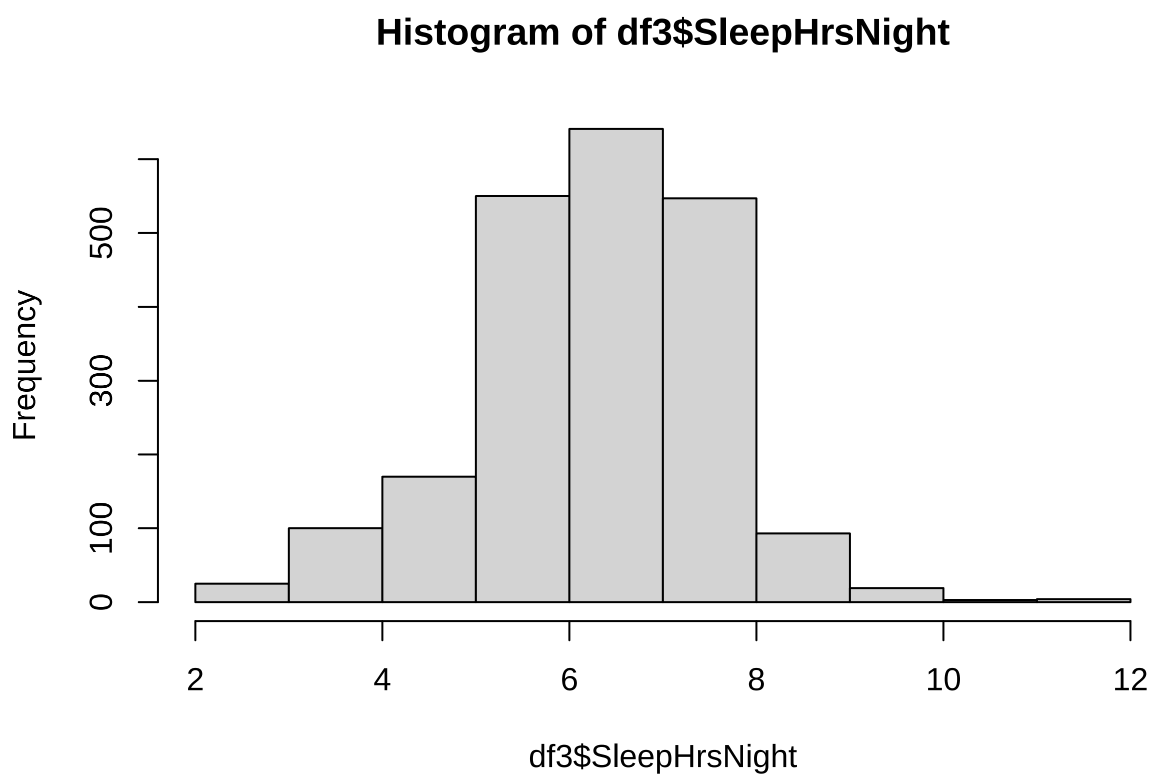
```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
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$logBMI = log(df3$BMI+1)



### multiple linear regression###
```

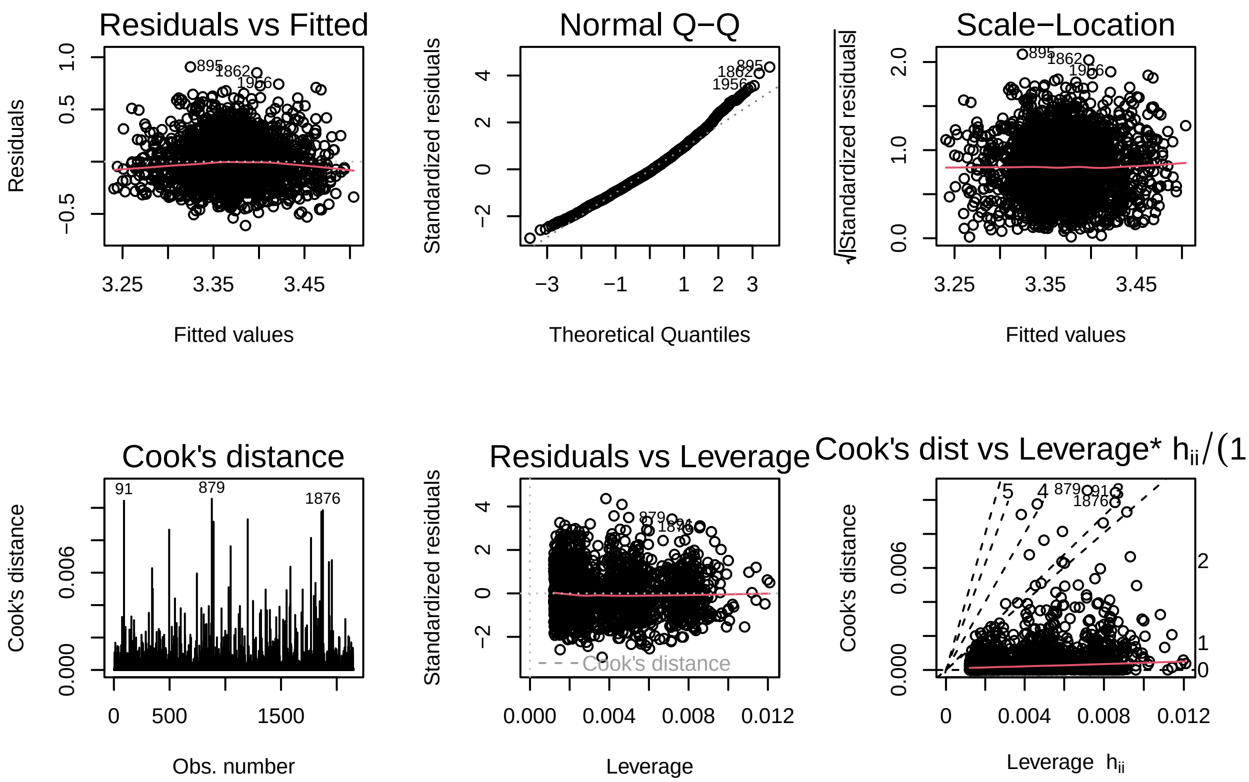
```r
# model_1 add demographic
m_1.log= lm(logBMI ~ SleepHrsNight + Age + Gender + factor(Race1), df3)
summary(m_1.log)
```

```
##
## Call:
## lm(formula = logBMI ~ SleepHrsNight + Age + Gender + factor(Race1),
##     data = df3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.61132 -0.14248 -0.02134  0.12467  0.90697
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.419148   0.030668 111.491  < 2e-16 ***
## SleepHrsNight   -0.009753   0.003460  -2.819  0.00486 **
## Age              0.001969   0.000402   4.898 1.04e-06 ***
## Gender          -0.002541   0.009063  -0.280  0.77919
## factor(Race1)2 -0.057106   0.021235  -2.689  0.00722 **
## factor(Race1)3 -0.015746   0.018549  -0.849  0.39603
## factor(Race1)4 -0.071198   0.013607  -5.232 1.84e-07 ***
## factor(Race1)5 -0.127712   0.020786  -6.144 9.56e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2083 on 2144 degrees of freedom
## Multiple R-squared:  0.03912,    Adjusted R-squared:  0.03598
## F-statistic: 12.47 on 7 and 2144 DF,  p-value: 9.611e-16
```

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

```
## Anova Table (Type III tests)
##
## Response: logBMI
##                Sum Sq   Df    F value    Pr(>F)
## (Intercept)    539.59    1 12430.1801 < 2.2e-16 ***
## SleepHrsNight    0.34    1     7.9471  0.004861 **
## Age              1.04    1    23.9941 1.039e-06 ***
## Gender           0.00    1     0.0786  0.779189
## factor(Race1)    2.35    4    13.5564 6.456e-11 ***
## Residuals       93.07 2144
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
########### model 1.log diagnosis ###########

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




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



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)
#which subject is most outlying with respect to the x space
Hmisc::describe(m_1.log.h)
```

```
## m_1.log.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 : 0.00120568 0.00120692 0.00121189 0.0012156  0.00122554
## highest: 0.0114043  0.0114087  0.0118542  0.0119757  0.0120914
```

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

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

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

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

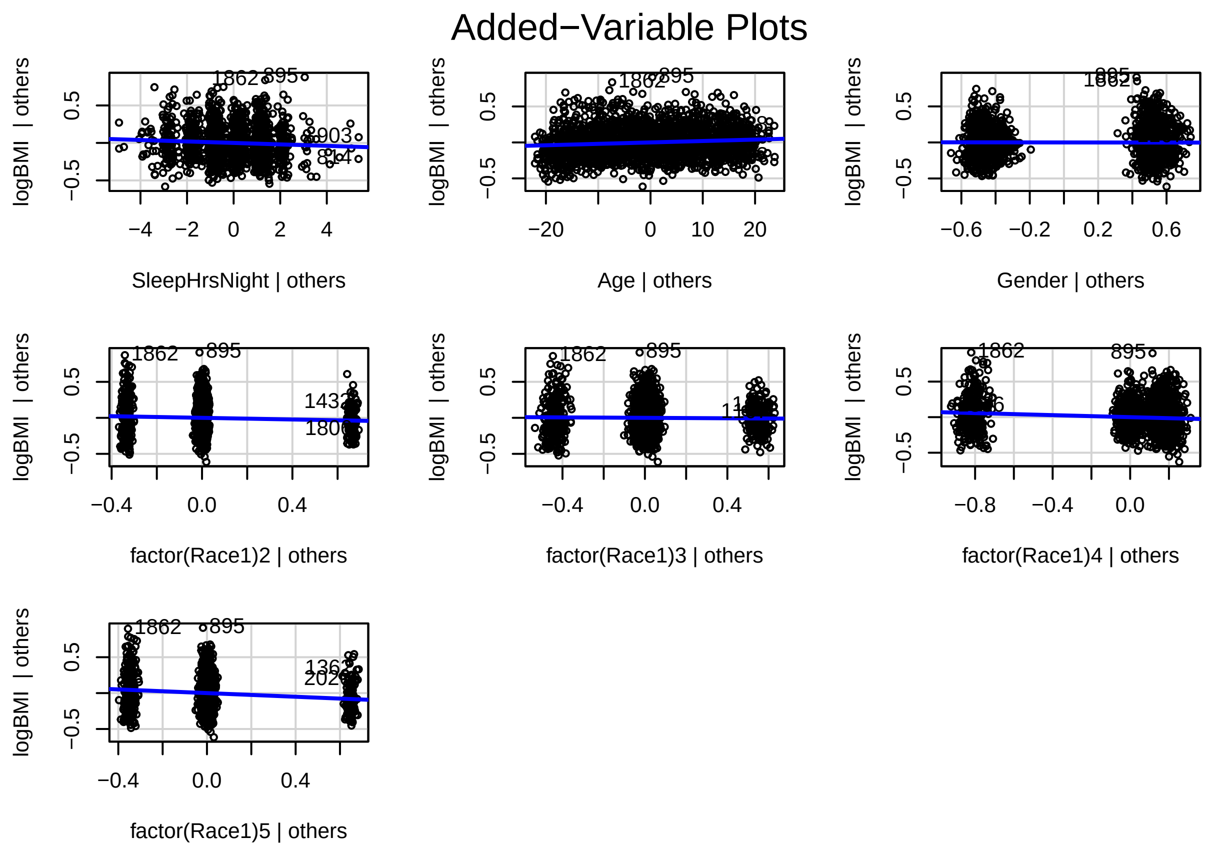
```
length(df3$logBMI)
```

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

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

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

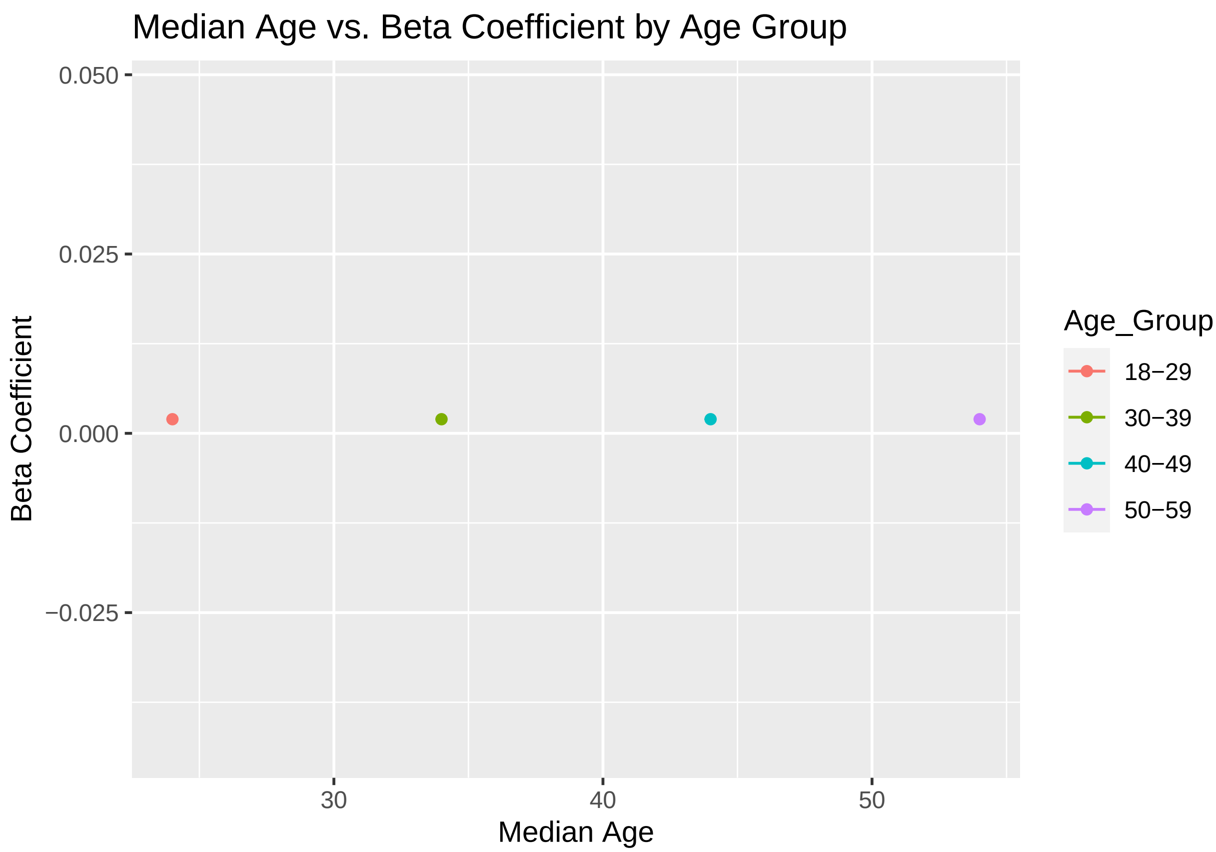
```
###################### Assumption:LINE ##############################

#(1)Linear: 2 approaches

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



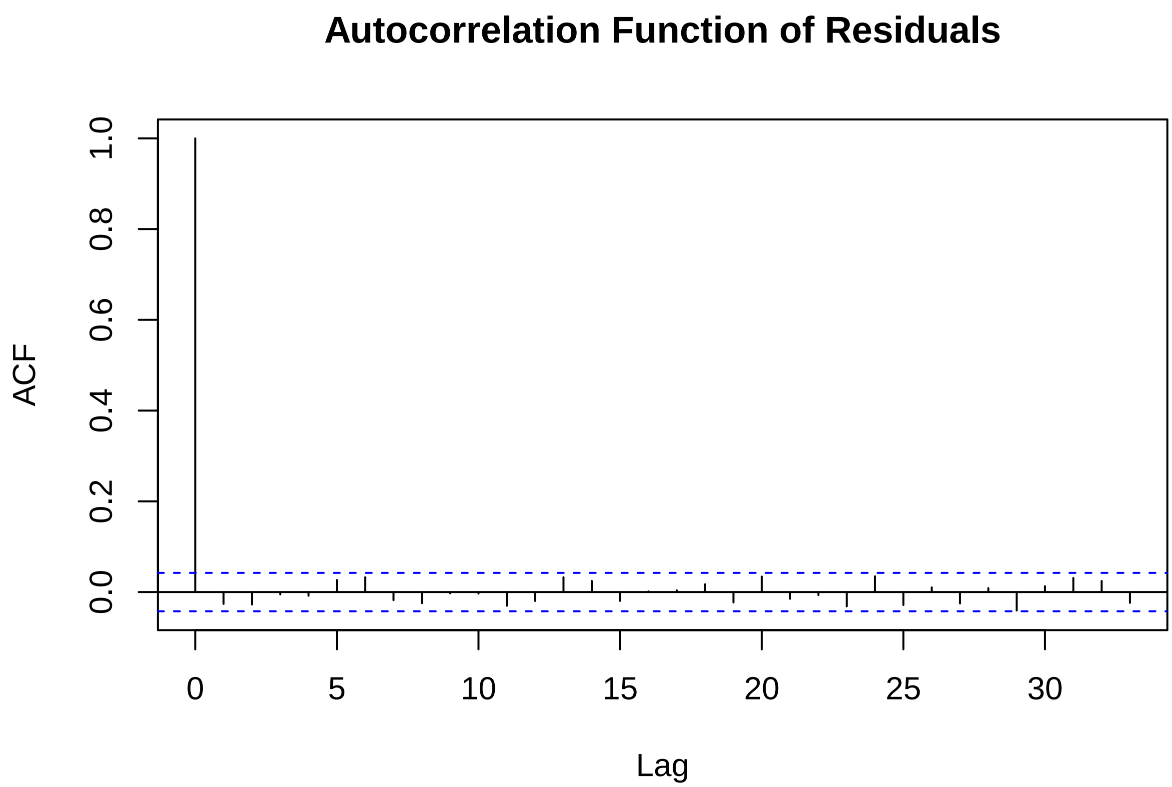

```
#categoraize age ---beta plot
df3 <- df3 %>%
  mutate(Age_Group = cut(Age, breaks = c(18, 29, 39, 49, 59), labels = c("18-29", "30-39", "40-49", "50-

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

ggplot(summary_stats, aes(x = Median_Age, y = Beta_Coefficient, group = Age_Group, color = Age_Group)) +
  geom_line() +
  geom_point() +
```

```r
  labs(title = "Median Age vs. Beta Coefficient by Age Group",
       x = "Median Age",
       y = "Beta Coefficient")
```

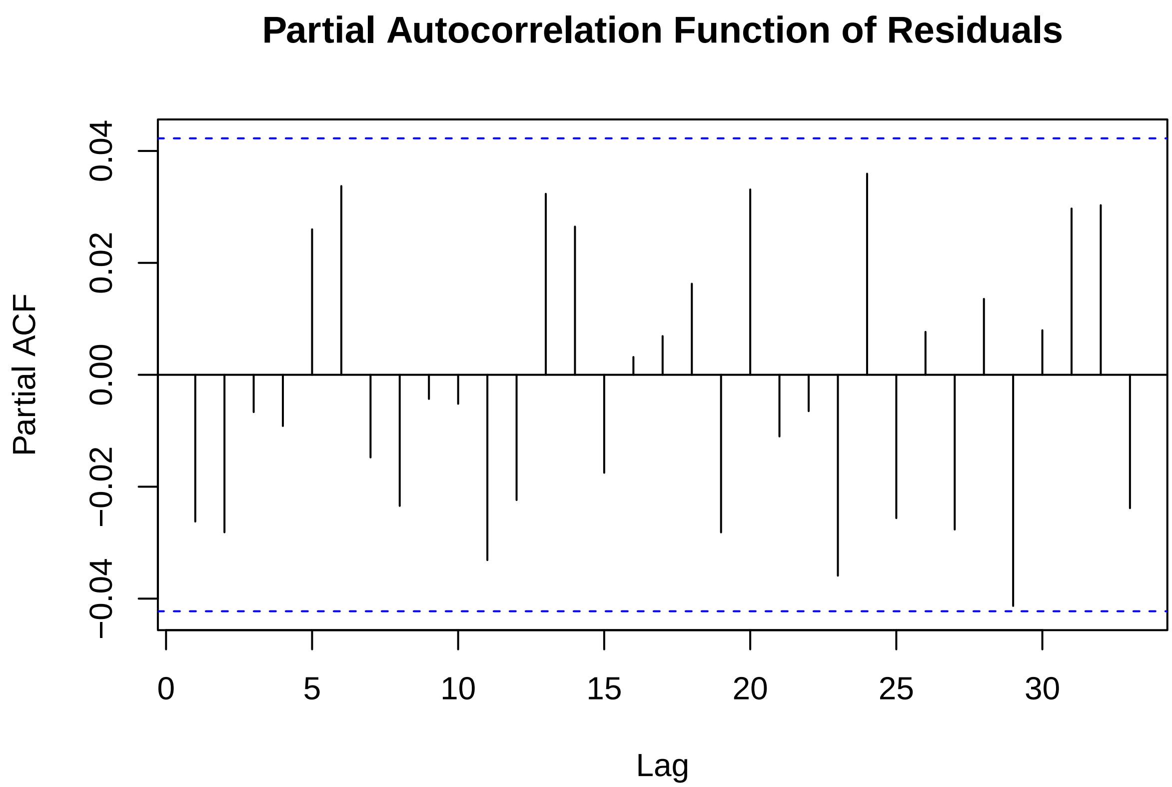
```
## `geom_line()`: Each group consists of only one observation.
## i Do you need to adjust the group aesthetic?
```




```r
#(2)Independence:

residuals <- resid(m_1.log)
acf(residuals, main = "Autocorrelation Function of Residuals")
```




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




```r
# Assuming m_1.log is your linear regression model
# Assuming df3 is your data frame

library(lmtest)
```

```
## Warning: package 'lmtest' was built under R version 4.2.3
```

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

```
## Warning: package 'zoo' was built under R version 4.2.3
```

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

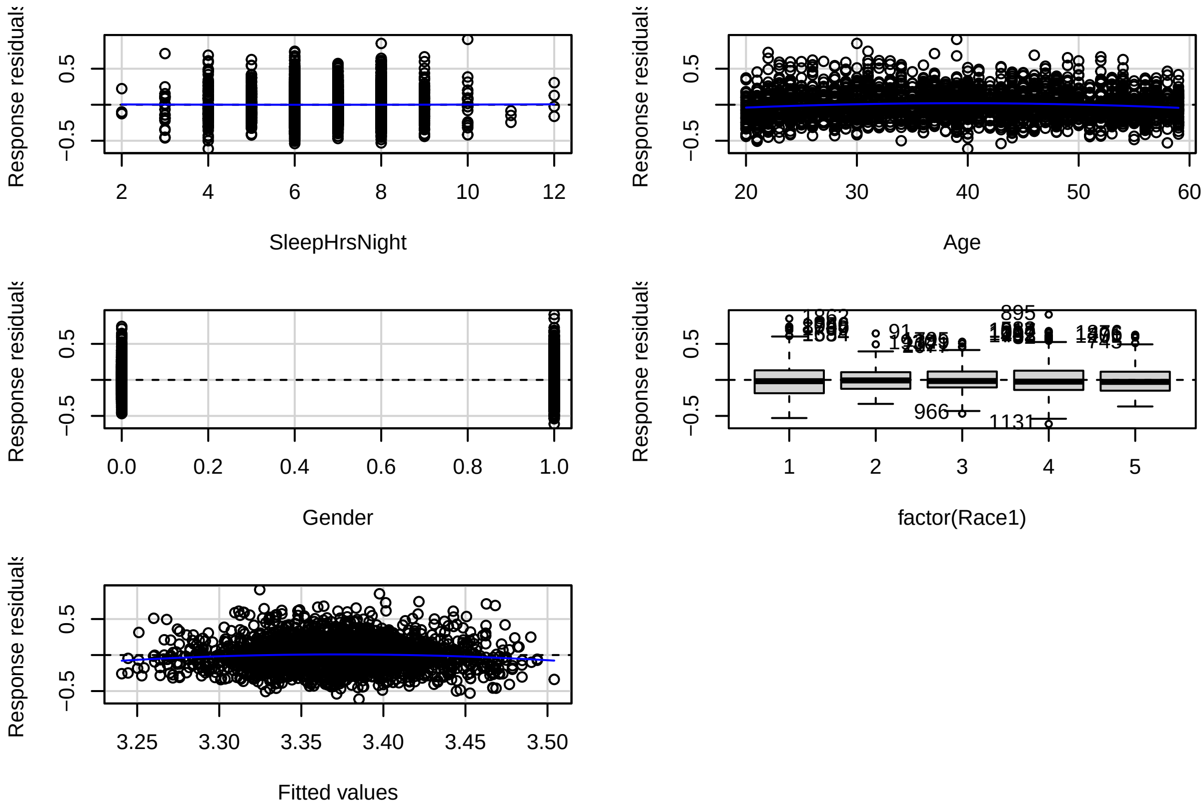
```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
# Perform Durbin-Watson test
dw_test_result <- dwtest(m_1.log, alternative = "two.sided")

# Print the Durbin-Watson test result
print(dw_test_result)
```

```
##
##  Durbin-Watson test
##
## data:  m_1.log
## DW = 2.0523, p-value = 0.2245
```

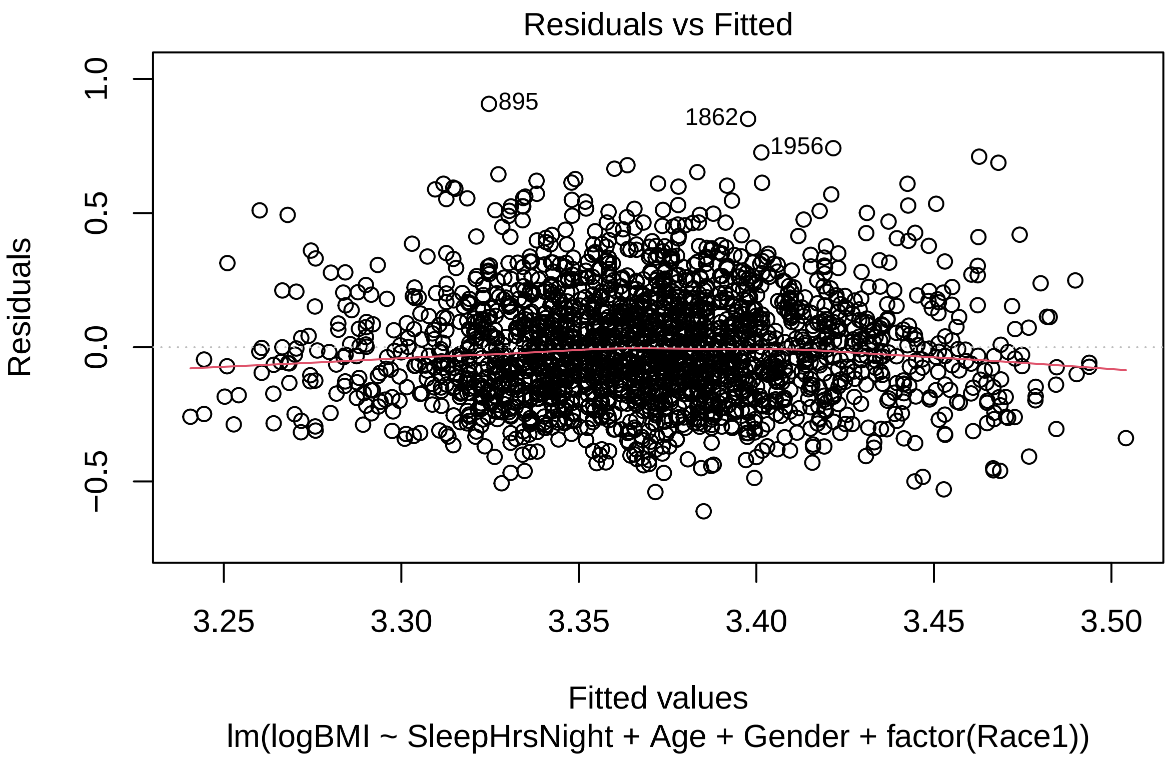
```
## alternative hypothesis: true autocorrelation is not 0
```

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

car::residualPlots(m_1.log,type="response")
```

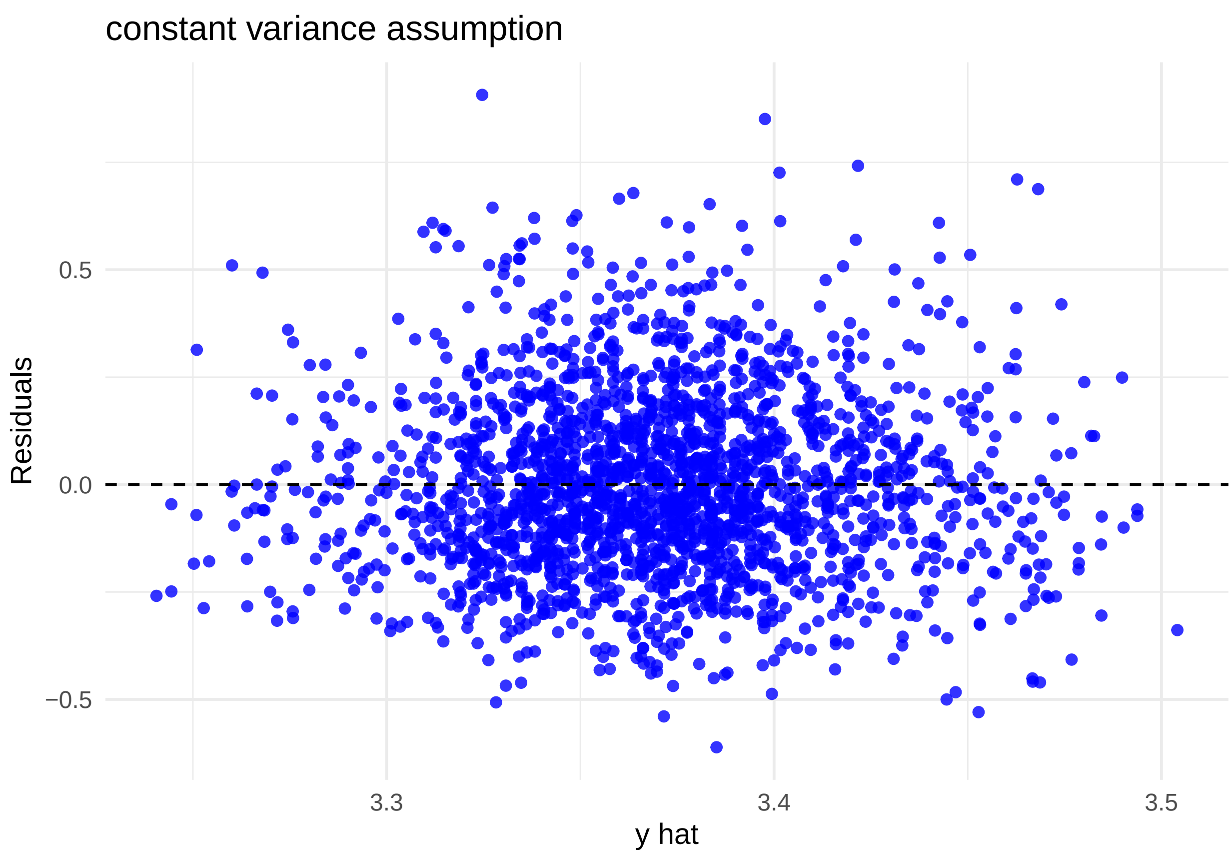



```
##                Test stat Pr(>|Test stat|)
## SleepHrsNight     0.1369         0.891137
## Age              -4.3105        1.702e-05 ***
## Gender           -0.1347         0.892867
## factor(Race1)
## Tukey test       -3.8099         0.000139 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
plot(m_1.log, which = 1)
```




```r
#or
ggplot(m_1.log, aes(x = m_1.log.yhat, y = m_1.log.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()
```



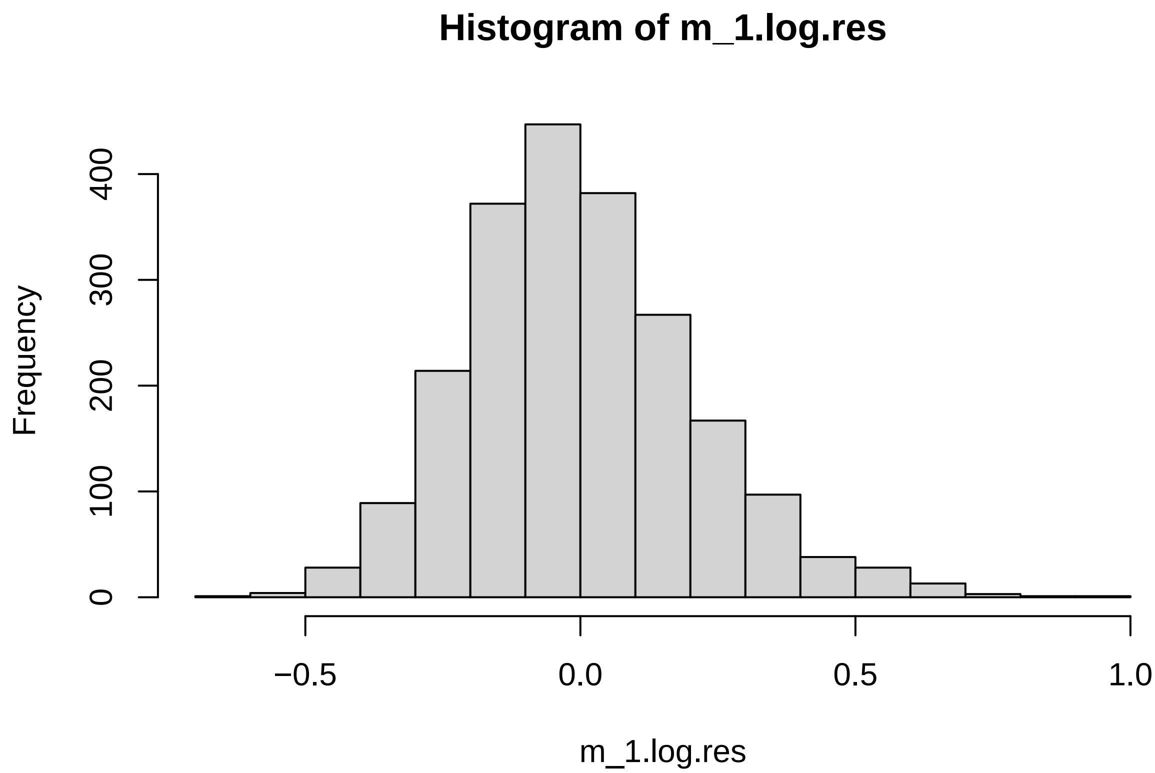

```
#conclusion: the constant variance assumption is basically not violated. The spread of the residuals ap



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

#exam quartiles of the residuals
Hmisc::describe(m_1.log.res)
```

```
## m_1.log.res
##          n    missing   distinct       Info       Mean        Gmd        .05
##       2152          0       2148          1 -3.347e-18     0.2319   -0.30647
##        .10        .25        .50        .75        .90        .95
##   -0.25086   -0.14248   -0.02134    0.12467    0.27089    0.36744
##
## lowest : -0.611319 -0.539522 -0.529624 -0.506861 -0.500102
## highest: 0.710259  0.725888  0.74189   0.850842  0.906969
```

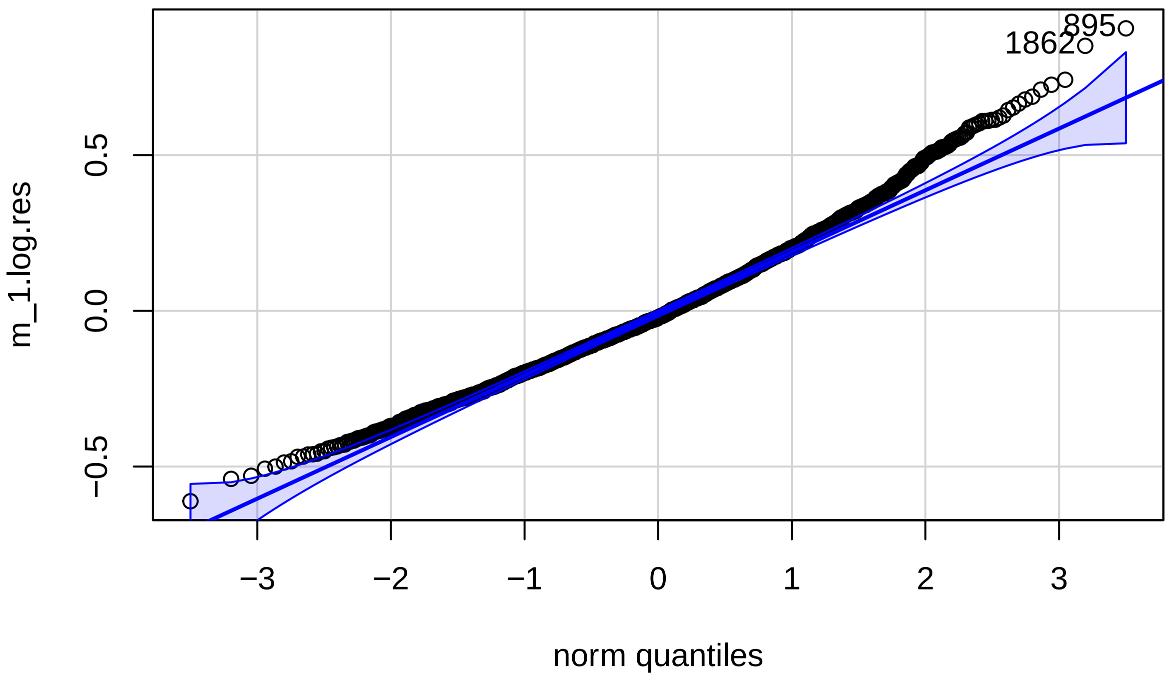
```
Hmisc::describe(m_1.log.res)$counts[c(".25",".50",".75")] #not symmetric
```

```
##        .25         .50         .75
## "-0.14248" "-0.02134" " 0.12467"
```

```
#histogram
par(mfrow = c(1, 1))
hist(m_1.log.res,breaks = 15)
```




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




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

```
##       895      1862
## 0.9069692 0.8508415
```

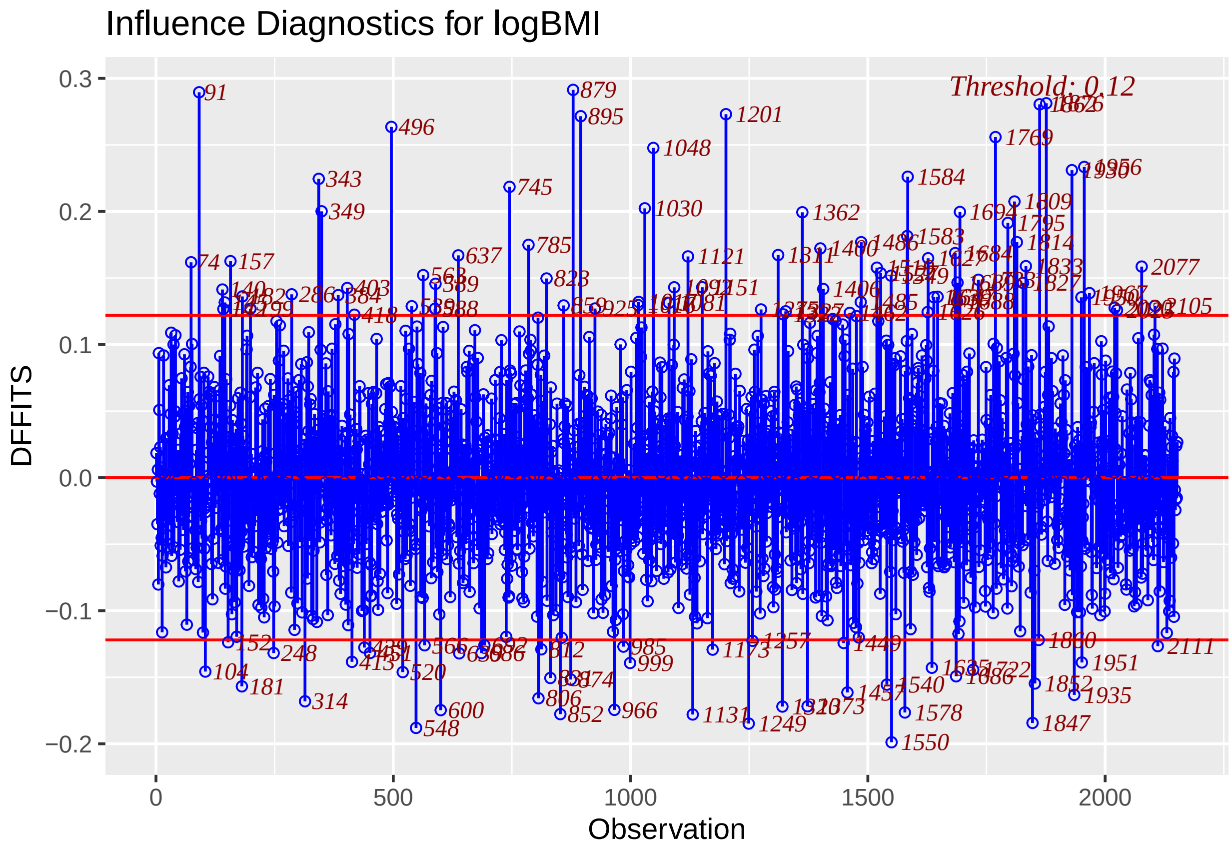
```
############### influential observations  #################

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

```
## Warning: package 'olsrr' was built under R version 4.2.3
```

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

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

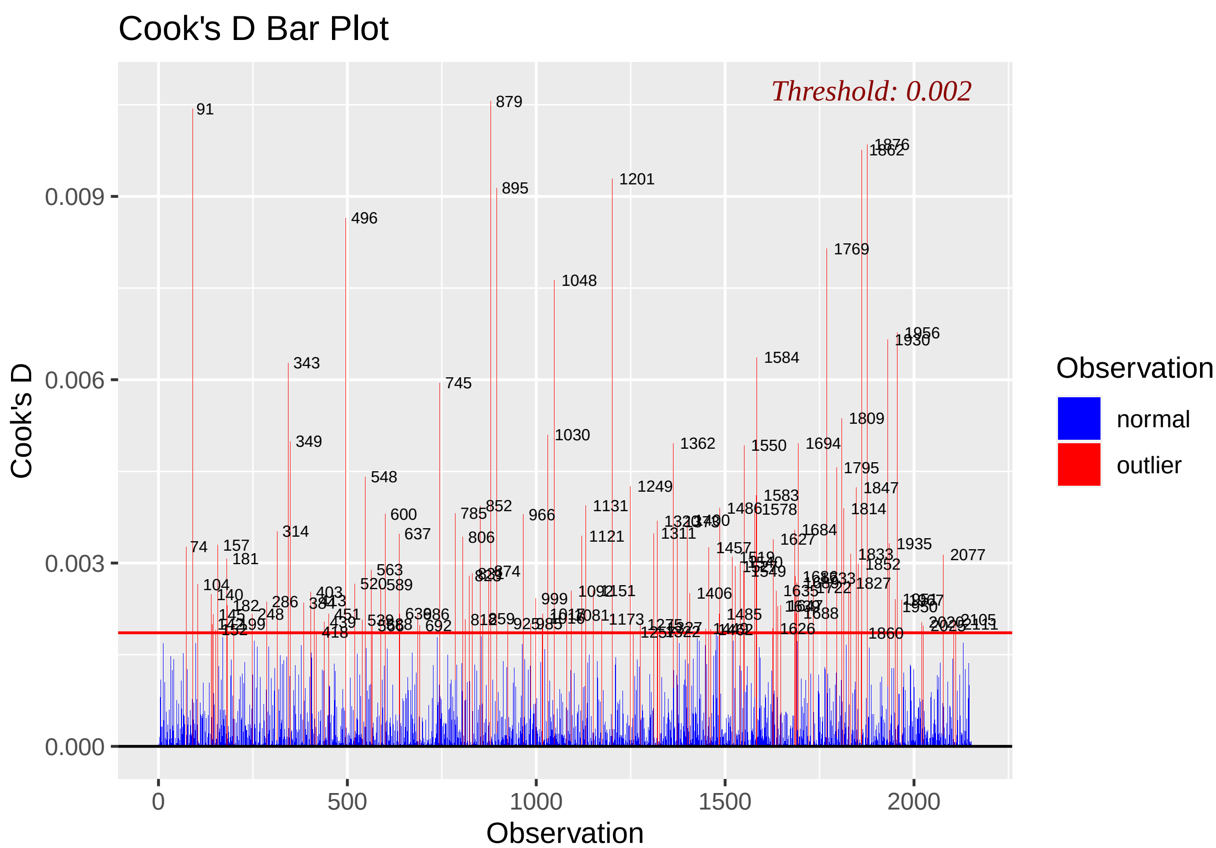
```
ols_plot_dffits(m_1.log)
```




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

```
##        Residual Rstudent    HatDiagH  CovRatio    DFFITS COOKsDistance
## 879  0.7102594 3.429835 0.007163982 0.9676614 0.2913478   0.010557437
## 91   0.6444631 3.112826 0.008578651 0.9765366 0.2895577   0.010438154
## 1876 0.6269385 3.027780 0.008557096 0.9784127 0.2812896   0.009852944
## 1862 0.8508415 4.108358 0.004639616 0.9470723 0.2804911   0.009762111
## 1201 0.5883984 2.841780 0.009149981 0.9829802 0.2730842   0.009291211
## 895  0.9069692 4.379930 0.003829692 0.9382613 0.2715703   0.009141275
```

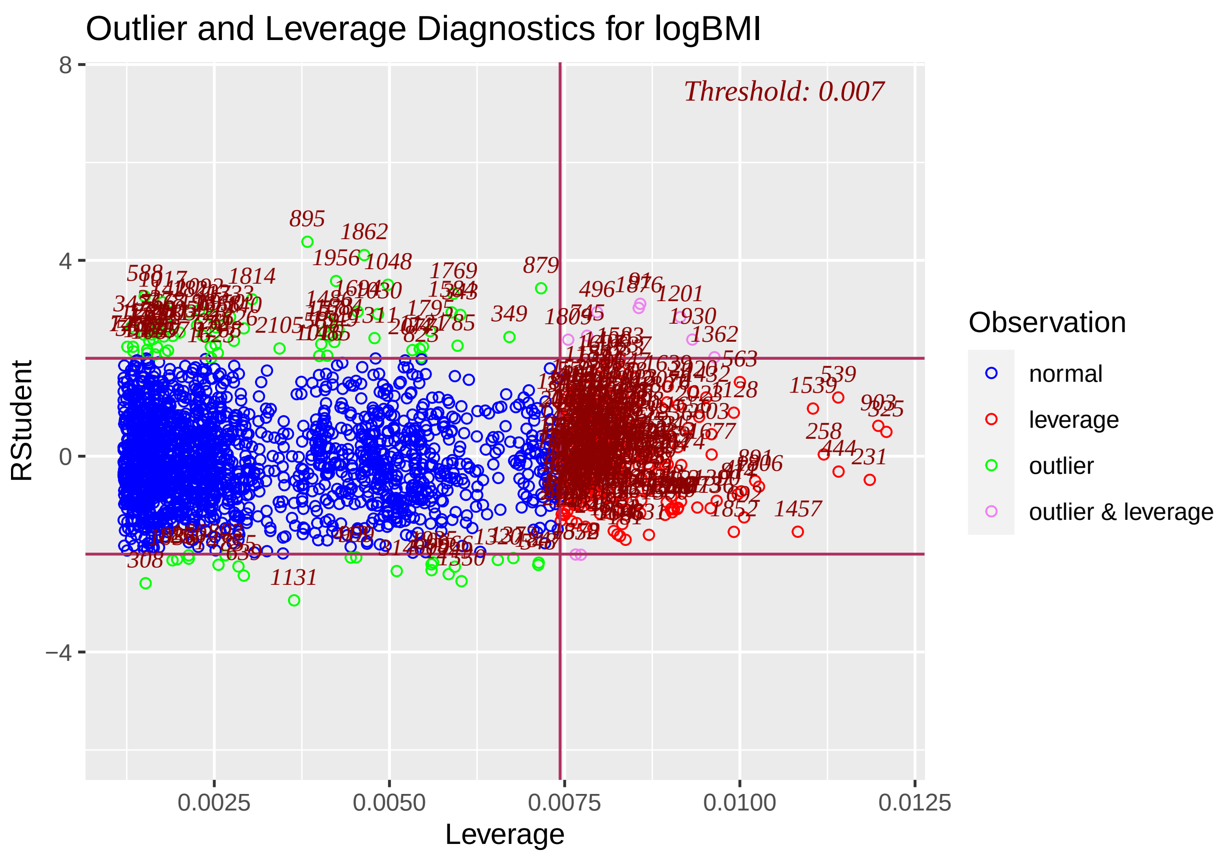
```r
#From the plot above, we can see 2 observations with the largest (magnitude) of DFFITS, observation 879

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




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

```
##        Residual Rstudent    HatDiagH  CovRatio    DFFITS COOKsDistance
## 879  0.7102594 3.429835 0.007163982 0.9676614 0.2913478   0.010557437
## 91   0.6444631 3.112826 0.008578651 0.9765366 0.2895577   0.010438154
## 1876 0.6269385 3.027780 0.008557096 0.9784127 0.2812896   0.009852944
## 1862 0.8508415 4.108358 0.004639616 0.9470723 0.2804911   0.009762111
## 1201 0.5883984 2.841780 0.009149981 0.9829802 0.2730842   0.009291211
## 895  0.9069692 4.379930 0.003829692 0.9382613 0.2715703   0.009141275
```

```
#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.log)
```




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

```
##          Residual    Rstudent   HatDiagH CovRatio      DFFITS COOKsDistance
## 325  0.102903785  0.49682588 0.01209140 1.015089  0.05496474  3.777730e-04
## 903  0.128457873  0.62018622 0.01197565 1.014448  0.06827912  5.829220e-04
## 231 -0.100085644 -0.48316020 0.01185417 1.014896 -0.05291955  3.501851e-04
## 444 -0.065075290 -0.31406819 0.01140869 1.014949 -0.03373909  1.423506e-04
## 539  0.248379354  1.19910720 0.01140428 1.009885  0.12879010  2.072938e-03
## 258  0.007336789  0.03540444 0.01119827 1.015102  0.00376772  1.775292e-06
```

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

```
##       Residual Rstudent    HatDiagH  CovRatio    DFFITS COOKsDistance
## 895  0.9069692 4.379930 0.003829692 0.9382613 0.2715703   0.009141275
## 1862 0.8508415 4.108358 0.004639616 0.9470723 0.2804911   0.009762111
## 1956 0.7418899 3.578175 0.004238569 0.9611028 0.2334498   0.006775053
## 1048 0.7258882 3.501861 0.004979030 0.9637486 0.2477166   0.007630357
## 879  0.7102594 3.429835 0.007163982 0.9676614 0.2913478   0.010557437
## 1769 0.6875841 3.317663 0.005913164 0.9691639 0.2558763   0.008146067
```

```r
#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.log = lm(logBMI ~ SleepHrsNight + Age + Gender + factor(Race1), rm.df3)
## Before removing these observations, the estimated coefficients are:
summary(m_1.log)$coef
```

```
##                     Estimate   Std. Error     t value     Pr(>|t|)
## (Intercept)     3.419148011 0.0306675577 111.4907175 0.000000e+00
## SleepHrsNight  -0.009752870 0.0034596185  -2.8190595 4.860689e-03
## Age             0.001968997 0.0004019691   4.8983792 1.038560e-06
## Gender         -0.002541267 0.0090626370  -0.2804115 7.791889e-01
## factor(Race1)2 -0.057105901 0.0212345861  -2.6892872 7.215952e-03
## factor(Race1)3 -0.015746046 0.0185485688  -0.8489090 3.960267e-01
## factor(Race1)4 -0.071198415 0.0136072328  -5.2323949 1.836209e-07
## factor(Race1)5 -0.127711727 0.0207862332  -6.1440534 9.562198e-10
```

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

```
##                     Estimate   Std. Error     t value     Pr(>|t|)
## (Intercept)     3.402863783 0.0302870425 112.3537824 0.000000e+00
## SleepHrsNight  -0.008353952 0.0034213352  -2.4417227 1.469823e-02
## Age             0.001999779 0.0003950297   5.0623521 4.496670e-07
## Gender         -0.001446304 0.0089073013  -0.1623728 8.710277e-01
## factor(Race1)2 -0.066280209 0.0210746938  -3.1450141 1.683700e-03
## factor(Race1)3 -0.010611356 0.0182027356  -0.5829539 5.599860e-01
## factor(Race1)4 -0.066247643 0.0133690031  -4.9553166 7.790481e-07
## factor(Race1)5 -0.147297721 0.0207113928  -7.1119177 1.556007e-12
```

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

```
##    (Intercept)  SleepHrsNight            Age         Gender factor(Race1)2
##      0.4762657     14.3436616      1.5633338     43.0873099     16.0654282
## factor(Race1)3 factor(Race1)4 factor(Race1)5
##     32.6093955      6.9534859     15.3360965
```

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



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

```
## Warning: package 'corrplot' was built under R version 4.2.3
```

```
## 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.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
```

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