

project

Liancheng

2023-11-21

(1) Data cleaning

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

##           used (Mb) gc trigger (Mb) max used (Mb)
## Ncells  469562 25.1    1011175 54.1    660860 35.3
## Vcells  877833  6.7     8388608 64.0   1800812 13.8

set.seed(123)
##### (1) Data cleaning #####
library(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)]
# colSums(is.na(df)) / nrow(df)
# df$BPSysAve
library(dplyr)

##
## Attaching package: 'dplyr'
## 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,
  HomeRooms,
```

```

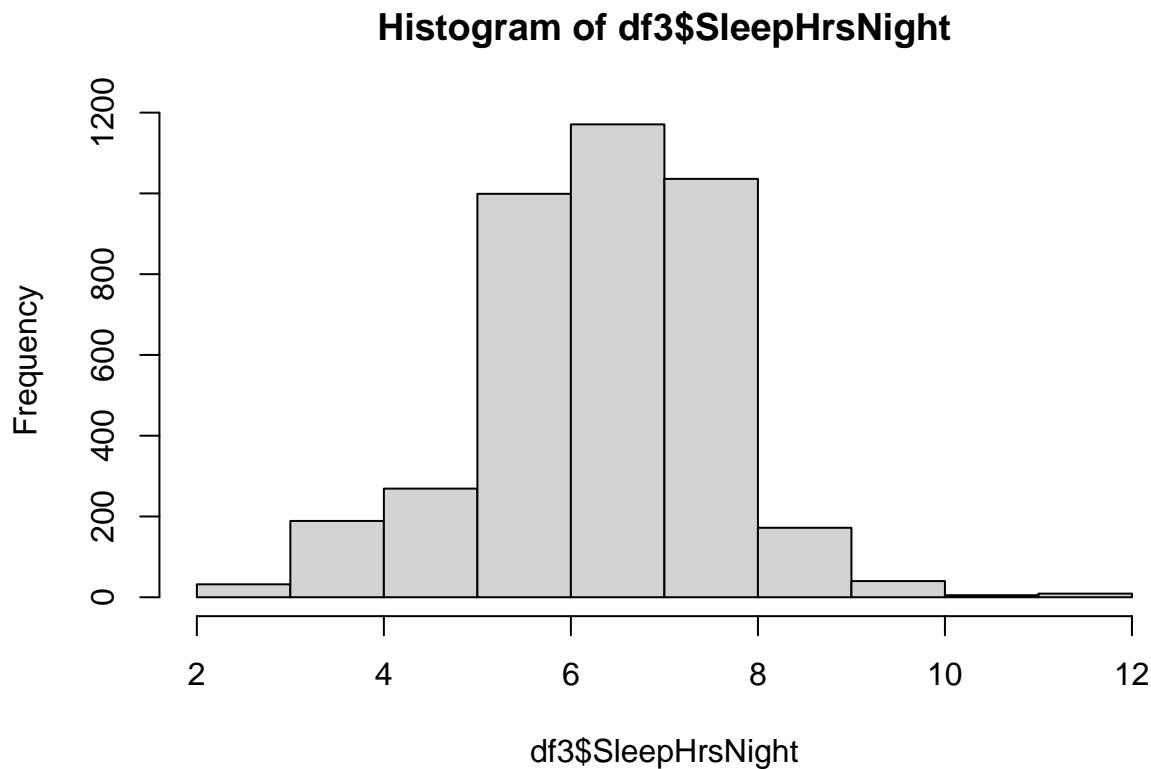
SexNumPartnLife,
SexNumPartYear,
DaysMentHlthBad
)

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 3922  6.83  1.30    7.00    6.90  1.48  2.00 12.00
## BMI                      2 3922 28.64  6.59   27.50   28.05  6.08 15.02 69.00
## DirectChol              3 3922  1.35  0.42    1.29    1.31  0.39  0.39  3.83
## Age                      4 3922 39.34 11.63   40.00   39.41 14.83 18.00 59.00
## Gender*                  5 3922  1.54  0.50    2.00    1.55  0.00  1.00  2.00
## Race1*                   6 3922  3.57  1.04    4.00    3.76  0.00  1.00  5.00
## TotChol                  7 3922  5.08  1.06    5.02    5.03  1.04  1.53 13.65
## BPDiaAve                 8 3922 71.51 11.40   72.00   71.62 10.38  0.00 116.00
## BPSysAve                 9 3922 117.72 14.28 116.00  116.85 13.34 78.00 226.00
## AlcoholYear              10 3922 71.91 95.14   24.00   52.26 35.58  0.00 364.00
## Poverty                  11 3922  3.01  1.66    3.15    3.08  2.65  0.00  5.00
## HomeRooms                12 3922  6.14  2.29    6.00    6.02  1.48  1.00 13.00
## SexNumPartnLife           13 3922 16.21 61.34   6.00    8.64  5.93  0.00 2000.00
## SexNumPartYear             14 3922  1.38  3.04    1.00    0.99  0.00  0.00  69.00
## DaysMentHlthBad            15 3922  4.41  7.99    0.00    2.34  0.00  0.00 30.00
##                                range   skew kurtosis    se
## SleepHrsNight          10.00 -0.25     0.72 0.02
## BMI                     53.98  1.10     2.20 0.11
## DirectChol              3.44  1.15     2.49 0.01
## Age                     41.00 -0.06    -1.18 0.19
## Gender*                  1.00 -0.17    -1.97 0.01
## Race1*                   4.00 -1.48     1.25 0.02
## TotChol                  12.12  0.76     2.22 0.02
## BPDiaAve                 116.00 -0.30    2.51 0.18
## BPSysAve                 148.00  1.08     3.90 0.23
## AlcoholYear              364.00  1.62     1.82 1.52
## Poverty                  5.00 -0.15    -1.43 0.03
## HomeRooms                12.00  0.53     0.27 0.04
## SexNumPartnLife          2000.00 17.33   399.45 0.98
## SexNumPartYear            69.00 12.99   222.05 0.05
## DaysMentHlthBad            30.00  2.19     3.89 0.13

# 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 <- 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
    )
  )
```

(2) Baseline characteristics

```
Hmisc::describe(df3)
```

```
## df3
##
```

```

## 15 Variables      3922 Observations
## -----
## SleepHrsNight
##      n    missing   distinct     Info      Mean      Gmd      .05      .10
##      3922        0        11    0.938    6.829    1.398        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     29    189    269    999   1171   1036    172     40      5      9
## Proportion 0.001 0.007 0.048 0.069 0.255 0.299 0.264 0.044 0.010 0.001 0.002
## -----
## BMI
##      n    missing   distinct     Info      Mean      Gmd      .05      .10
##      3922        0       1139        1    28.64    7.162    19.99    21.20
##      .25       .50       .75       .90       .95
##      23.90    27.50    32.25    37.20    40.74
##
## 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
##      3922        0        99        1    1.351    0.4553    0.80     0.88
##      .25       .50       .75       .90       .95
##      1.06     1.29     1.58     1.91     2.12
##
## lowest : 0.39 0.41 0.52 0.54 0.57, highest: 3.41 3.44 3.59 3.72 3.83
## -----
## Age
##      n    missing   distinct     Info      Mean      Gmd      .05      .10
##      3922        0        42    0.999    39.34    13.42      21      23
##      .25       .50       .75       .90       .95
##      29        40        49        55        57
##
## lowest : 18 19 20 21 22, highest: 55 56 57 58 59
## -----
## Gender
##      n    missing   distinct     Info      Sum      Mean      Gmd
##      3922        0        2    0.745    1795    0.4577    0.4965
##
## -----
## Race1
##      n    missing   distinct     Info      Mean      Gmd
##      3922        0        5    0.668    3.574    0.9307
##
## lowest : 1 2 3 4 5, highest: 1 2 3 4 5
##
## Value      1      2      3      4      5
## Frequency 387    215    345   2709    266
## Proportion 0.099 0.055 0.088 0.691 0.068
## -----
## TotChol

```

```

##          n  missing distinct      Info      Mean      Gmd      .05      .10
##  3922        0       212         1    5.083    1.169    3.540    3.830
##  .25        .50       .75         .90       .95
##  4.320     5.020     5.720       6.390     6.877
##
## lowest :  1.53  2.43  2.59  2.69  2.74, highest:  9.31  9.34  9.90 12.28 13.65
## -----
## BPDiaAve
##          n  missing distinct      Info      Mean      Gmd      .05      .10
##  3922        0       84         0.999   71.51    12.42      54       58
##  .25        .50       .75         .90       .95
##  65         72       78         85       89
##
## lowest :  0  20  21  22  25, highest: 108 109 110 114 116
## -----
## BPSysAve
##          n  missing distinct      Info      Mean      Gmd      .05      .10
##  3922        0       99         0.999  117.7    15.36      98      102
##  .25        .50       .75         .90       .95
##  108       116      125        134      142
##
## lowest :  78  83  84  85  86, highest: 184 191 202 209 226
## -----
## AlcoholYear
##          n  missing distinct      Info      Mean      Gmd      .05      .10
##  3922        0       56         0.993  71.91    93.2       0       0
##  .25        .50       .75         .90       .95
##  4          24      104        208      300
##
## lowest :  0  1  2  3  4, highest: 260 300 312 360 364
## -----
## Poverty
##          n  missing distinct      Info      Mean      Gmd      .05      .10
##  3922        0       398        0.984  3.009   1.892    0.450    0.751
##  .25        .50       .75         .90       .95
##  1.380     3.150     5.000      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
## -----
## HomeRooms
##          n  missing distinct      Info      Mean      Gmd      .05      .10
##  3922        0       13        0.98   6.141   2.528       3       4
##  .25        .50       .75         .90       .95
##  4          6        7          9       10
##
## lowest :  1  2  3  4  5, highest:  9 10 11 12 13
## -----
## Value      1      2      3      4      5      6      7      8      9      10      11
## Frequency  59     46     258    622    692    703    575    371    263    165     79
## Proportion 0.015 0.012 0.066 0.159 0.176 0.179 0.147 0.095 0.067 0.042 0.020
##
## Value      12     13
## Frequency  42     47
## Proportion 0.011 0.012

```

```

## -----
## SexNumPartnLife
##      n   missing  distinct      Info      Mean      Gmd      .05      .10
##    3922        0       81    0.996    16.21    21.98      1       1
##    .25       .50       .75      .90      .95
##    3         6       15      30      50
##
## lowest :  0   1   2   3   4, highest:  600  800  999 1000 2000
## -----
## SexNumPartYear
##      n   missing  distinct      Info      Mean      Gmd      .05      .10
##    3922        0       22    0.673    1.381    1.286      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
##    3922        0       28    0.841    4.412    6.821      0       0
##    .25       .50       .75      .90      .95
##    0         0       5       15      30
##
## lowest :  0   1   2   3   4, highest: 25  26  27  29  30
## -----

```

(3) linear regression model

```

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

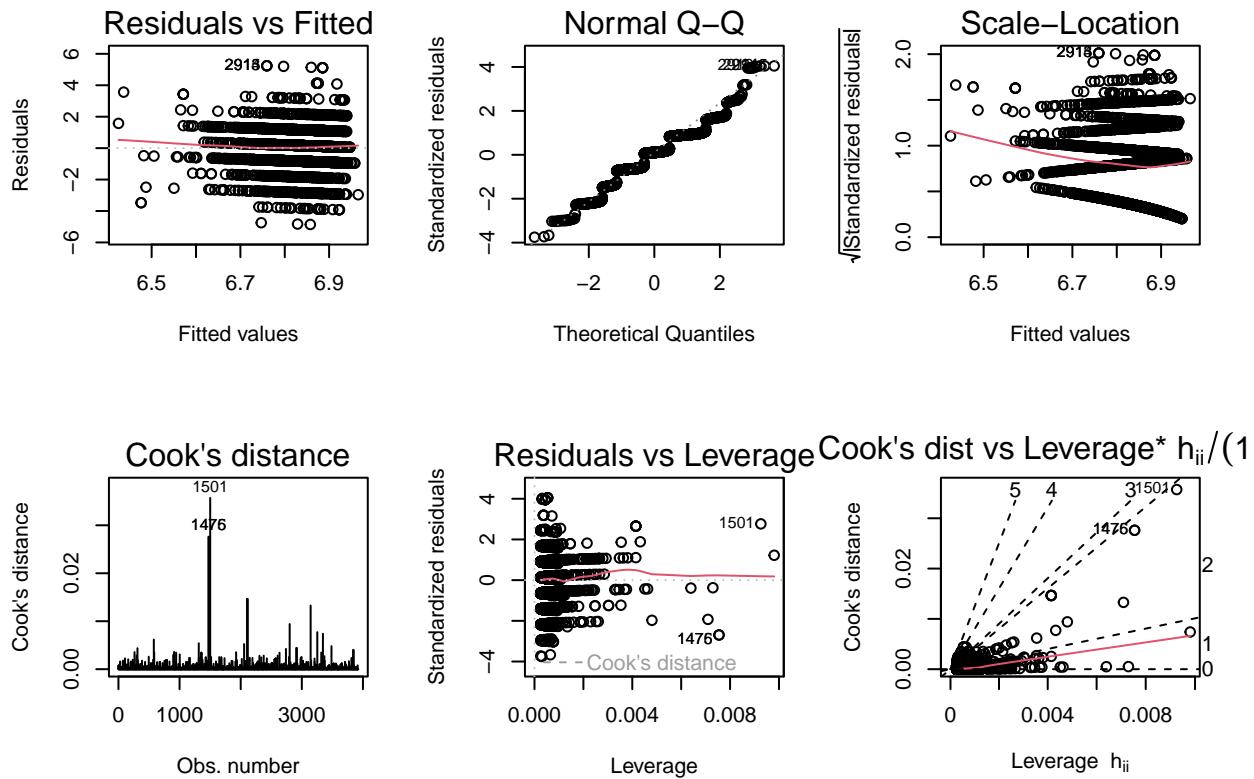
##
## Call:
## lm(formula = df3$SleepHrsNight ~ df3$BMI, data = df3)
##
## Residuals:
##      Min       1Q     Median       3Q      Max
## -4.8576 -0.8355  0.1395  1.1249  5.2404
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.115837  0.092166 77.207 < 2e-16 ***
## df3$BMI     -0.010008  0.003136 -3.192  0.00143 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.295 on 3920 degrees of freedom
## Multiple R-squared:  0.002592,  Adjusted R-squared:  0.002338
## F-statistic: 10.19 on 1 and 3920 DF,  p-value: 0.001426

```

```

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

## multiple linear regression##
m_initial = lm(SleepHrsNight ~ BMI + Age + Gender + factor(Race1), df3)
summary(m_initial)

##
## Call:
## lm(formula = SleepHrsNight ~ BMI + Age + Gender + factor(Race1),
##      data = df3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -4.9640 -0.8046  0.1101  1.0409  5.3517 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 6.910657   0.133893 51.613 < 2e-16 ***
## 
```

```

## BMI          -0.006587  0.003157  -2.086 0.037016 *
## Age          -0.007866  0.001800  -4.370 1.28e-05 ***
## Gender       0.198904  0.041324   4.813 1.54e-06 ***
## factor(Race1)2 0.174457  0.109626   1.591 0.111603
## factor(Race1)3 0.409723  0.095539   4.289 1.84e-05 ***
## factor(Race1)4 0.370461  0.070687   5.241 1.68e-07 ***
## factor(Race1)5 0.355100  0.103282   3.438 0.000592 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.285 on 3914 degrees of freedom
## Multiple R-squared:  0.01997,    Adjusted R-squared:  0.01822
## F-statistic:  11.4 on 7 and 3914 DF,  p-value: 2.187e-14

m_knrisk = lm(
  SleepHrsNight ~ BMI + Age + Gender + factor(Race1) + TotChol + BPDiaAve +
  BPSysAve + AlcoholYear + DaysMentHlthBad,
  df3
)
summary(m_knrisk)

##
## Call:
## lm(formula = SleepHrsNight ~ BMI + Age + Gender + factor(Race1) +
##     TotChol + BPDiaAve + BPSysAve + AlcoholYear + DaysMentHlthBad,
##     data = df3)
##
## Residuals:
##      Min        1Q        Median        3Q        Max 
## -5.0151 -0.8371  0.0538  0.9651  5.3364 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 7.1462829  0.2154155 33.174 < 2e-16 ***
## BMI         -0.0032441  0.0032012 -1.013 0.310923  
## Age         -0.0087863  0.0019017 -4.620 3.96e-06 ***
## Gender       0.2421933  0.0423999  5.712 1.20e-08 ***
## factor(Race1)2 0.1615075  0.1080191  1.495 0.134949  
## factor(Race1)3 0.3670216  0.0943591  3.890 0.000102 ***
## factor(Race1)4 0.3361684  0.0697583  4.819 1.50e-06 ***
## factor(Race1)5 0.3107938  0.1019396  3.049 0.002313 ** 
## TotChol      0.0027643  0.0202261  0.137 0.891300  
## BPDiaAve    0.0020128  0.0021165  0.951 0.341646  
## BPSysAve    -0.0030312  0.0017413 -1.741 0.081793 .  
## AlcoholYear  0.0006543  0.0002219  2.949 0.003209 ** 
## DaysMentHlthBad -0.0299239  0.0025406 -11.778 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.262 on 3909 degrees of freedom
## Multiple R-squared:  0.05591,    Adjusted R-squared:  0.05302
## F-statistic: 19.29 on 12 and 3909 DF,  p-value: < 2.2e-16

m_full = lm(
  SleepHrsNight ~ BMI + Age + Gender + factor(Race1) + TotChol + BPDiaAve +

```

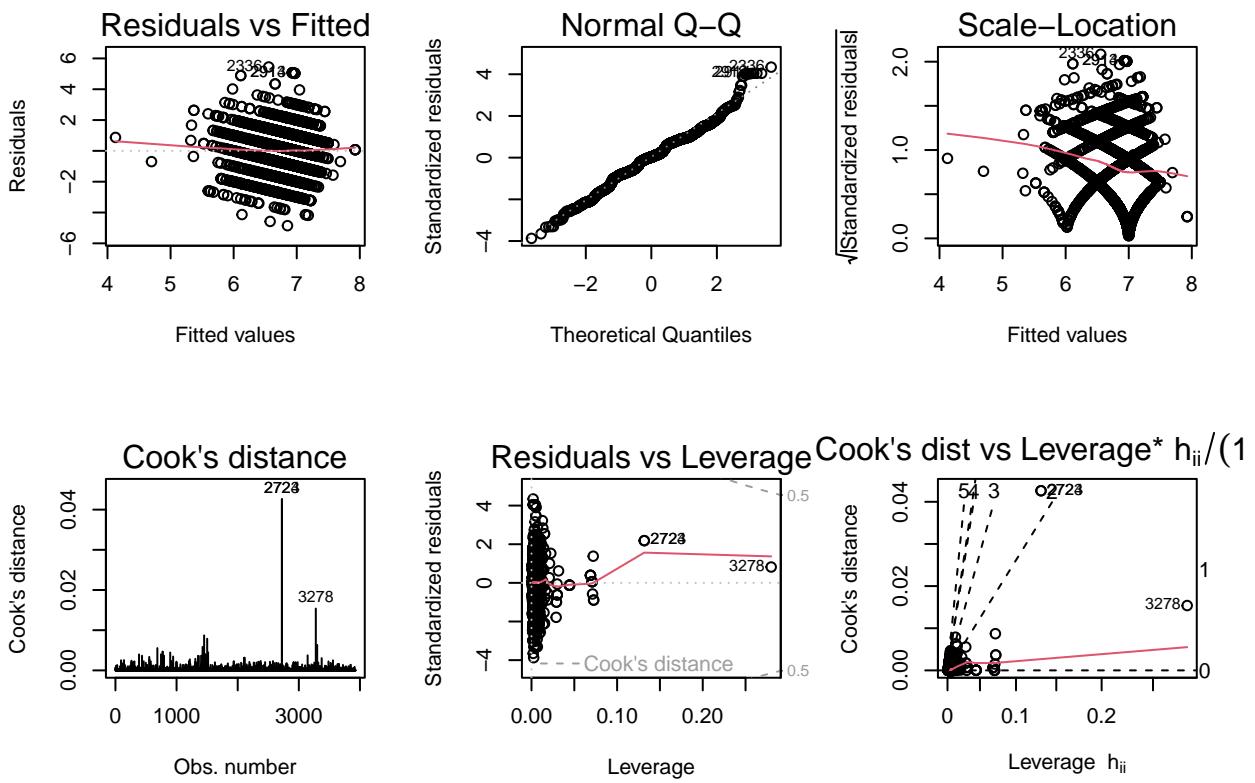
```

    BPSysAve + AlcoholYear + DaysMentHlthBad + HomeRooms + SexNumPartnLife +
    SexNumPartYear + Poverty,
  df3
)
summary(m_full)

##
## Call:
## lm(formula = SleepHrsNight ~ BMI + Age + Gender + factor(Race1) +
##      TotChol + BPDiaAve + BPSysAve + AlcoholYear + DaysMentHlthBad +
##      HomeRooms + SexNumPartnLife + SexNumPartYear + Poverty, data = df3)
##
## Residuals:
##       Min     1Q   Median     3Q    Max 
## -4.8534 -0.8280  0.0354  0.9312  5.4440 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 6.8271794  0.2226486 30.663 < 2e-16 ***
## BMI        -0.0026447  0.0031871 -0.830 0.406694  
## Age        -0.0107341  0.0019748 -5.435 5.80e-08 ***
## Gender      0.2300247  0.0423898  5.426 6.10e-08 ***
## factor(Race1)2 0.1634606  0.1075484  1.520 0.128622  
## factor(Race1)3 0.3982799  0.0942020  4.228 2.41e-05 ***
## factor(Race1)4 0.2862593  0.0702207  4.077 4.66e-05 ***
## factor(Race1)5 0.2854605  0.1016592  2.808 0.005010 ** 
## TotChol      0.0047184  0.0201452  0.234 0.814828  
## BPDiaAve    0.0018866  0.0021093  0.894 0.371149  
## BPSysAve    -0.0022470  0.0017400 -1.291 0.196654  
## AlcoholYear   0.0005280  0.0002223  2.375 0.017598 *  
## DaysMentHlthBad -0.0280171  0.0025566 -10.959 < 2e-16 ***
## HomeRooms    0.0260173  0.0095185  2.733 0.006298 ** 
## SexNumPartnLife -0.0011068  0.0003339 -3.315 0.000925 *** 
## SexNumPartYear  0.0187508  0.0067967  2.759 0.005828 ** 
## Poverty      0.0522337  0.0137235  3.806 0.000143 *** 
## ---        
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.255 on 3905 degrees of freedom
## Multiple R-squared:  0.06694,    Adjusted R-squared:  0.06312 
## F-statistic: 17.51 on 16 and 3905 DF,  p-value: < 2.2e-16

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

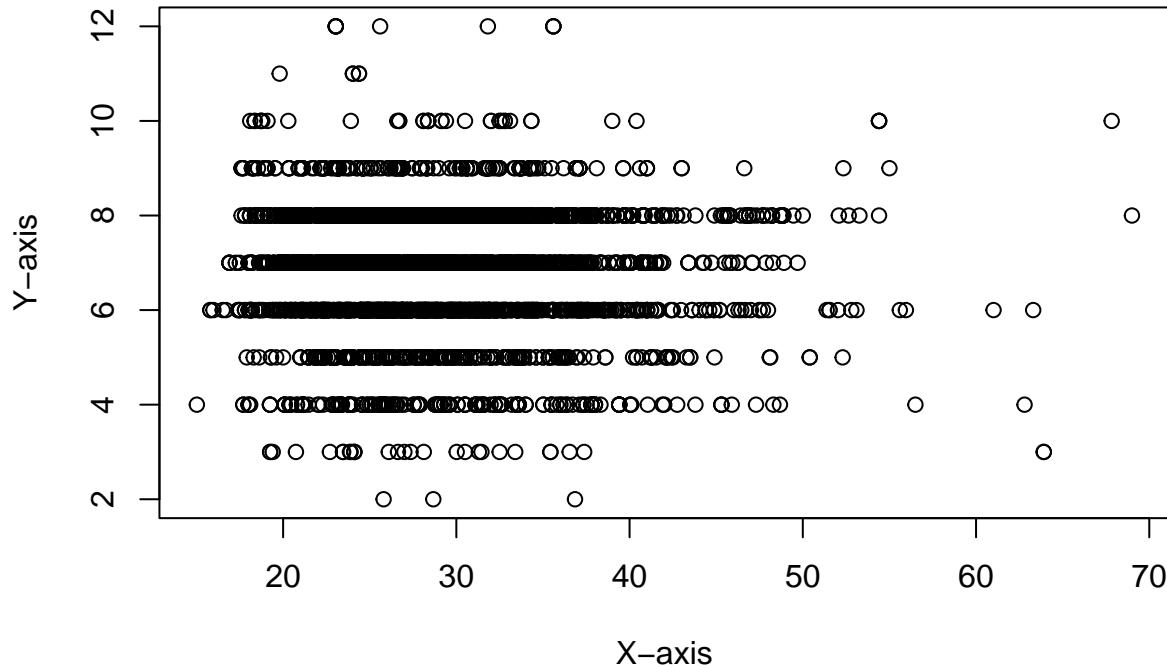
```



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

plot(
  df3$BMI,
  df3$SleepHrsNight,
  main = "Scatter Plot with Linear Regression Line",
  xlab = "X-axis",
  ylab = "Y-axis"
)
```

Scatter Plot with Linear Regression Line



```
#log outcome
df3$logSleepHrsNight = log(df3$SleepHrsNight + 1)
m_logfull_1 = lm(
  logSleepHrsNight ~ BMI + Age + Gender + factor(Race1) + TotChol + BPDiaAve +
  BPSysAve + AlcoholYear + DaysMentHlthBad + HomeRooms + SexNumPartnLife +
  SexNumPartYear + Poverty,
  df3
)
summary(m_logfull_1)

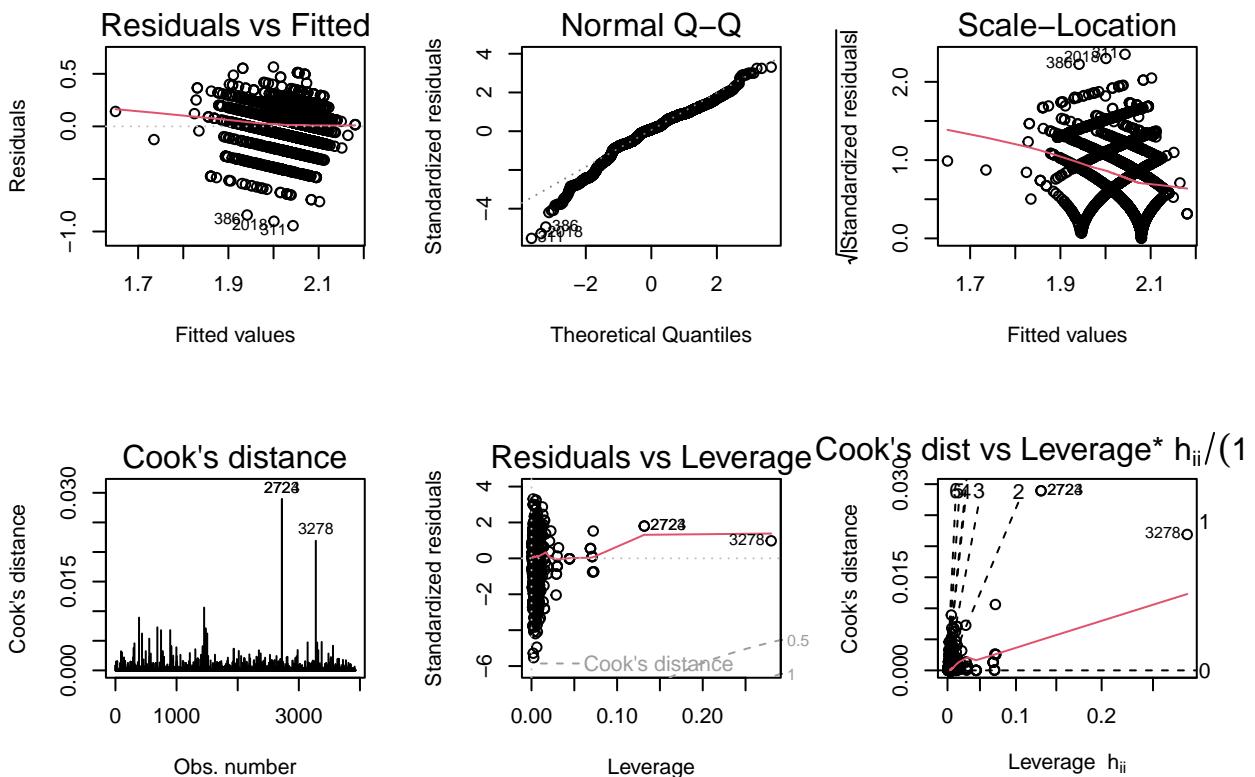
##
## Call:
## lm(formula = logSleepHrsNight ~ BMI + Age + Gender + factor(Race1) +
##     TotChol + BPDiaAve + BPSysAve + AlcoholYear + DaysMentHlthBad +
##     HomeRooms + SexNumPartnLife + SexNumPartYear + Poverty, data = df3)
##
## Residuals:
##      Min        1Q        Median         3Q        Max 
## -0.94458 -0.09816  0.01636  0.12163  0.56510 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 2.033e+00 3.033e-02 67.026 < 2e-16 ***
## BMI        -4.730e-04 4.342e-04 -1.090 0.275981  
## Age        -1.485e-03 2.690e-04 -5.520 3.60e-08 ***
## Gender      2.838e-02 5.774e-03  4.915 9.26e-07 ***
```

```

## factor(Race1)2  2.478e-02  1.465e-02   1.691 0.090882 .
## factor(Race1)3  5.693e-02  1.283e-02   4.437 9.38e-06 ***
## factor(Race1)4  4.259e-02  9.566e-03   4.453 8.72e-06 ***
## factor(Race1)5  4.232e-02  1.385e-02   3.056 0.002260 **
## TotChol         4.578e-04  2.744e-03   0.167 0.867504
## BPDiaAve       3.782e-04  2.873e-04   1.316 0.188144
## BPSysAve        -2.977e-04 2.370e-04  -1.256 0.209220
## AlcoholYear     8.234e-05  3.028e-05   2.719 0.006578 **
## DaysMentHlthBad -4.145e-03 3.483e-04  -11.903 < 2e-16 ***
## HomeRooms        3.765e-03  1.297e-03   2.904 0.003705 **
## SexNumPartnLife -1.623e-04 4.548e-05  -3.569 0.000362 ***
## SexNumPartYear   2.441e-03  9.258e-04   2.637 0.008400 **
## Poverty          8.175e-03  1.869e-03   4.373 1.26e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1709 on 3905 degrees of freedom
## Multiple R-squared:  0.07513,    Adjusted R-squared:  0.07134
## F-statistic: 19.83 on 16 and 3905 DF,  p-value: < 2.2e-16

par(mfrow = c(2, 3)) #read more from ?plot.lm
plot(m_logfull_1, which = 1)
plot(m_logfull_1, which = 2)
plot(m_logfull_1, which = 3)
plot(m_logfull_1, which = 4)
plot(m_logfull_1, which = 5)
plot(m_logfull_1, which = 6)

```



```

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

#log x
df3$logBMI = log(df3$BMI + 1)
m_logfull_2 = lm(
  SleepHrsNight ~ logBMI + Age + Gender + factor(Race1) + TotChol + BPDiaAve +
  BPSysAve + AlcoholYear + DaysMentHlthBad + HomeRooms + SexNumPartnLife +
  SexNumPartYear + Poverty,
  df3
)
summary(m_logfull_2)

##
## Call:
## lm(formula = SleepHrsNight ~ logBMI + Age + Gender + factor(Race1) +
##     TotChol + BPDiaAve + BPSysAve + AlcoholYear + DaysMentHlthBad +
##     HomeRooms + SexNumPartnLife + SexNumPartYear + Poverty, data = df3)
##
## Residuals:
##      Min        1Q    Median        3Q       Max
## -4.8498 -0.8252  0.0376  0.9327  5.4441
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.0697215  0.3707264 19.070 < 2e-16 ***
## logBMI     -0.0968912  0.0998871 -0.970 0.332103

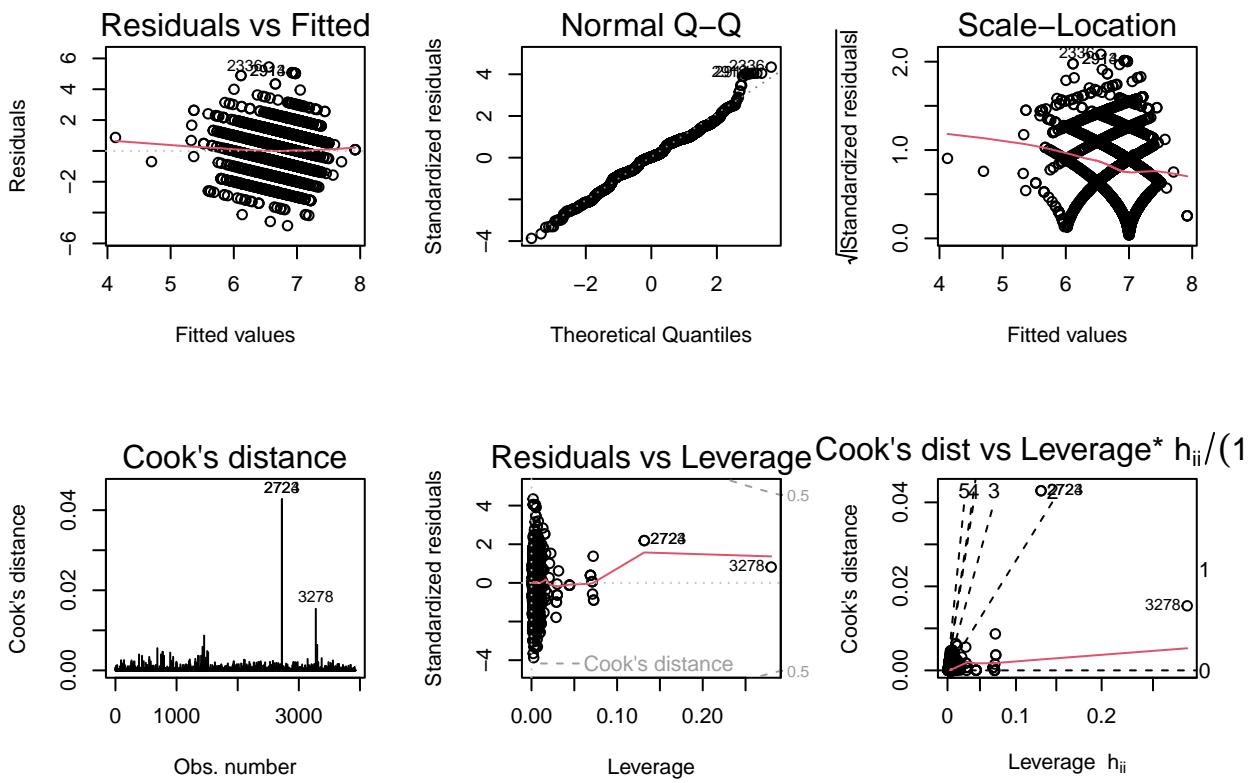
```

```

## Age          -0.0106914  0.0019765  -5.409 6.71e-08 ***
## Gender       0.2289323  0.0424081   5.398 7.13e-08 ***
## factor(Race1)2 0.1638502  0.1074512   1.525 0.127370
## factor(Race1)3 0.3991272  0.0941219   4.241 2.28e-05 ***
## factor(Race1)4 0.2857969  0.0701448   4.074 4.71e-05 ***
## factor(Race1)5 0.2844552  0.1015942   2.800 0.005137 **
## TotChol      0.0050053  0.0201492   0.248 0.803830
## BPDiaAve    0.0019128  0.0021095   0.907 0.364596
## BPSysAve     -0.0022143  0.0017404  -1.272 0.203335
## AlcoholYear  0.0005223  0.0002224   2.348 0.018906 *
## DaysMentHlthBad -0.0280111  0.0025564 -10.957 < 2e-16 ***
## HomeRooms    0.0260275  0.0095182   2.735 0.006276 **
## SexNumPartnLife -0.0011073  0.0003339  -3.316 0.000920 ***
## SexNumPartYear  0.0187471  0.0067960   2.759 0.005833 **
## Poverty      0.0522811  0.0137199   3.811 0.000141 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.255 on 3905 degrees of freedom
## Multiple R-squared:  0.067, Adjusted R-squared:  0.06318
## F-statistic: 17.53 on 16 and 3905 DF, p-value: < 2.2e-16

par(mfrow = c(2, 3)) #read more from ?plot.lm
plot(m_logfull_2, which = 1)
plot(m_logfull_2, which = 2)
plot(m_logfull_2, which = 3)
plot(m_logfull_2, which = 4)
plot(m_logfull_2, which = 5)
plot(m_logfull_2, which = 6)

```



```

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

# x^2
df3$sqBMI = (df3$BMI - mean(df3$BMI))^2
m_sqfull_1 = lm(
  SleepHrsNight ~ BMI + sqBMI + Age + Gender + factor(Race1) + TotChol +
  BPDiaAve + BPSysAve + AlcoholYear + DaysMentHlthBad + HomeRooms + SexNumPartnLife +
  SexNumPartYear + Poverty,
  df3
)
summary(m_sqfull_1)

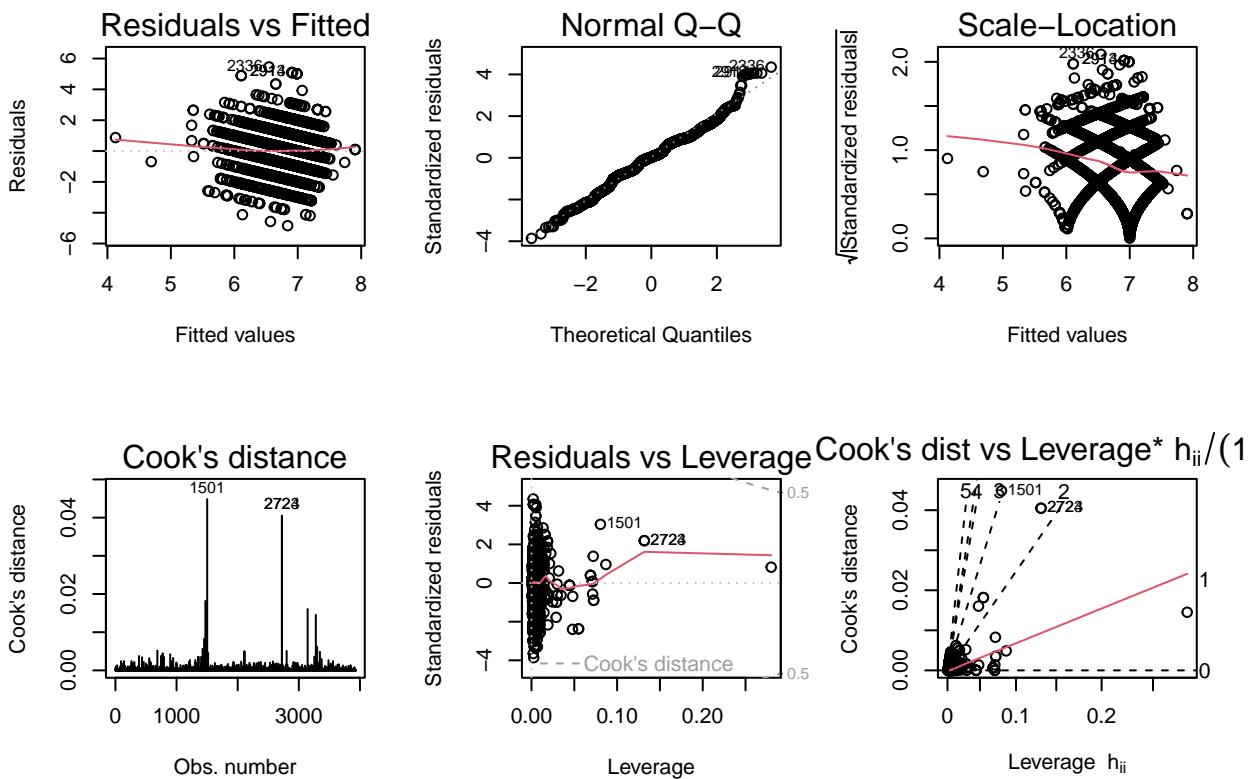
##
## Call:
## lm(formula = SleepHrsNight ~ BMI + sqBMI + Age + Gender + factor(Race1) +
##     TotChol + BPDiaAve + BPSysAve + AlcoholYear + DaysMentHlthBad +
##     HomeRooms + SexNumPartnLife + SexNumPartYear + Poverty, data = df3)
##
## Residuals:
##      Min        1Q    Median        3Q       Max 
## -4.8406 -0.8248  0.0403  0.9355  5.4532 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 6.8508755  0.2236619 30.630 < 2e-16 ***
## BMI         -0.0049446  0.0038007 -1.301 0.193344
## 
```

```

## sqBMI      0.0003061  0.0002756   1.111 0.266797
## Age       -0.0105379  0.0019827  -5.315 1.13e-07 ***
## Gender     0.2230819  0.0428470   5.206 2.02e-07 ***
## factor(Race1)2 0.1730766  0.1078932   1.604 0.108762
## factor(Race1)3 0.4096193  0.0947509   4.323 1.58e-05 ***
## factor(Race1)4 0.2908673  0.0703411   4.135 3.62e-05 ***
## factor(Race1)5 0.2890734  0.1017082   2.842 0.004504 **
## TotChol     0.0061198  0.0201841   0.303 0.761753
## BPDiaAve   0.0019388  0.0021097   0.919 0.358163
## BPSysAve   -0.0021929  0.0017406  -1.260 0.207810
## AlcoholYear 0.0005170  0.0002225   2.323 0.020214 *
## DaysMentHlthBad -0.0280502  0.0025567 -10.971 < 2e-16 ***
## HomeRooms    0.0261720  0.0095193   2.749 0.005998 **
## SexNumPartnLife -0.0011077  0.0003339  -3.318 0.000916 ***
## SexNumPartYear  0.0188640  0.0067973   2.775 0.005542 **
## Poverty      0.0528991  0.0137362   3.851 0.000119 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.255 on 3904 degrees of freedom
## Multiple R-squared:  0.06724,    Adjusted R-squared:  0.06318
## F-statistic: 16.55 on 17 and 3904 DF,  p-value: < 2.2e-16

par(mfrow = c(2, 3)) #read more from ?plot.lm
plot(m_sqfull_1, which = 1)
plot(m_sqfull_1, which = 2)
plot(m_sqfull_1, which = 3)
plot(m_sqfull_1, which = 4)
plot(m_sqfull_1, which = 5)
plot(m_sqfull_1, which = 6)

```



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

(4) Diagnosis: 10-fold CV

```
library(caret)

## Loading required package: ggplot2
## Loading required package: lattice
splitIndex <-
  createDataPartition(df3$SleepHrsNight, p = 0.7, list = FALSE)
trainData <- df3[splitIndex,]
testData <- df3[-splitIndex,]
predictions <- predict(m_sqfull_1, newdata = testData)
mse <- mean((testData$SleepHrsNight - predictions) ^ 2)
control <-
  trainControl(method = "cv", number = 10) # 10-fold cross-validation
cv_model <-
  train(
    SleepHrsNight ~ .,
    data = df3,
    method = "lm",
    trControl = control
  )
```

```

cv_model

## Linear Regression
##
## 3922 samples
##    17 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 3529, 3529, 3529, 3530, 3530, 3530, ...
## Resampling results:
##
##    RMSE      Rsquared     MAE
##    0.1816134  0.9805096  0.1194418
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
(cv_results <- cv_model$results)

##    intercept      RMSE   Rsquared       MAE   RMSESD   RsquaredSD       MAESD
## 1      TRUE 0.1816134 0.9805096 0.1194418 0.0307151 0.005095776 0.007554956

```

(4) Diagnosis: Normality Assumption

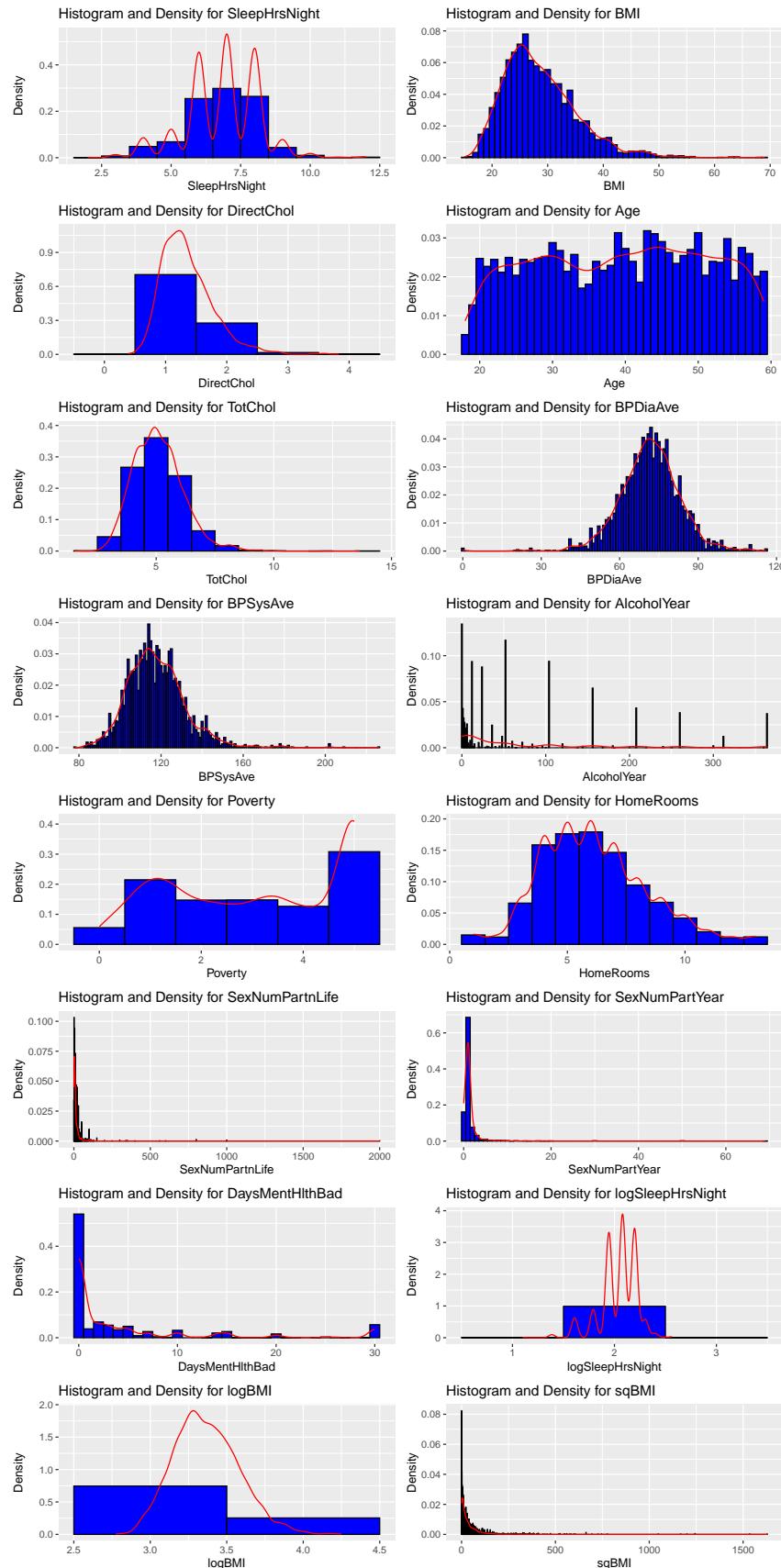
```

library(ggplot2)
library(patchwork)
# Initializes an empty patchwork object
plot_list <- list()

# Draw a histogram for each numeric variable (except Race1 and Gender) and add it to the list
for (var in names(df3)) {
  if (is.numeric(df3[[var]]) && !(var %in% c("Race1", "Gender"))) {
    p <- ggplot(df3, aes(x = .data[[var]])) +
      geom_histogram(
        aes(y = after_stat(density)),
        binwidth = 1,
        fill = "blue",
        color = "black"
      ) +
      geom_density(col = "red") +
      ggtitle(paste("Histogram and Density for", var)) +
      xlab(var) +
      ylab("Density")
    plot_list[[length(plot_list) + 1]] <- p
  }
}

# Use patchwork to put all the charts together
combined_plot <- wrap_plots(plot_list, ncol = 2)
print(combined_plot)

```



```

df3 <- data.frame(df3)
library(dplyr)
# Shapiro-Wilk normality test is performed for each numerical variable in df3
results <- sapply(df3, function(x) {
  if (is.numeric(x)) {
    shapiro_test <- shapiro.test(x)
    return(c(shapiro_test$statistic, shapiro_test$p.value))
  } else {
    return(c(NA, NA))
  }
})
# Convert the result to a data box and name the column
results_df <- as.data.frame(t(results))
names(results_df) <- c("W", "p.value")
# Add a variable name as a new column
results_df$Variable <- rownames(results_df)
# Rearrange the order of columns
results_df <- results_df[, c("Variable", "W", "p.value")]
# Calculate the corrected P-value (for example, using Bonferroni correction)
results_df$p.adjusted <-
  p.adjust(results_df$p.value, method = "bonferroni")
print(results_df)

```

	Variable	W	p.value	p.adjusted
## SleepHrsNight	SleepHrsNight	0.9324408	6.174763e-39	1.111457e-37
## BMI	BMI	0.9420252	1.043365e-36	1.878057e-35
## DirectChol	DirectChol	0.9389239	1.850577e-37	3.331039e-36
## Age	Age	0.9565820	1.100461e-32	1.980830e-31
## Gender	Gender	0.6340133	4.238105e-68	7.628589e-67
## Race1	Race1	0.6732812	7.054979e-66	1.269896e-64
## TotChol	TotChol	0.9724090	7.211614e-27	1.298090e-25
## BPDiaAve	BPDiaAve	0.9787402	8.519951e-24	1.533591e-22
## BPSysAve	BPSysAve	0.9505758	1.857649e-34	3.343769e-33
## AlcoholYear	AlcoholYear	0.7494486	7.869506e-61	1.416511e-59
## Poverty	Poverty	0.8916507	3.020524e-46	5.436943e-45
## HomeRooms	HomeRooms	0.9631989	1.707583e-30	3.073650e-29
## SexNumPartnLife	SexNumPartnLife	0.1633647	2.016343e-85	3.629418e-84
## SexNumPartYear	SexNumPartYear	0.2272038	1.134070e-83	2.041325e-82
## DaysMentHlthBad	DaysMentHlthBad	0.6061789	1.487607e-69	2.677692e-68
## logSleepHrsNight	logSleepHrsNight	0.8994157	4.724481e-45	8.504065e-44
## logBMI	logBMI	0.9909264	3.861958e-15	6.951525e-14
## sqBMI	sqBMI	0.4296243	3.806098e-77	6.850976e-76

Standardized residuals, Studentized residuals

```

# Regular residuals
residual_1 <- fit0$residuals

# Standardized residuals
residual_2 <- rstandard(fit0)

# Studentized residuals

```

```

residual_3 <- rstudent(fit0)

# Externally studentized residuals
# Note: Externally studentized residuals are the same as studentized residuals in most cases
residual_4 <- rstudent(fit0)

# Creating a data frame to summarize these residuals
residual_summary <- data.frame(
  Residuals = c("Regular", "Standardized", "Studentized", "Externally Studentized"),
  Mean = c(mean(residual_1), mean(residual_2), mean(residual_3), mean(residual_4)),
  SD = c(sd(residual_1), sd(residual_2), sd(residual_3), sd(residual_4)),
  Min = c(min(residual_1), min(residual_2), min(residual_3), min(residual_4)),
  Max = c(max(residual_1), max(residual_2), max(residual_3), max(residual_4))
)

# Display the summary
print(residual_summary)

##             Residuals      Mean       SD      Min      Max
## 1          Regular -2.268127e-16 1.251851 -4.894975 5.444620
## 2 Standardized  9.976361e-05 1.000389 -3.907567 4.343986
## 3   Studentized  8.874780e-05 1.000738 -3.914730 4.353965
## 4 Externally Studentized  8.874780e-05 1.000738 -3.914730 4.353965

# Load necessary library
library(ggplot2)

# Assuming fit0 is your linear model
# fit0 <- lm(SleepMinNight ~ ., data = df3)

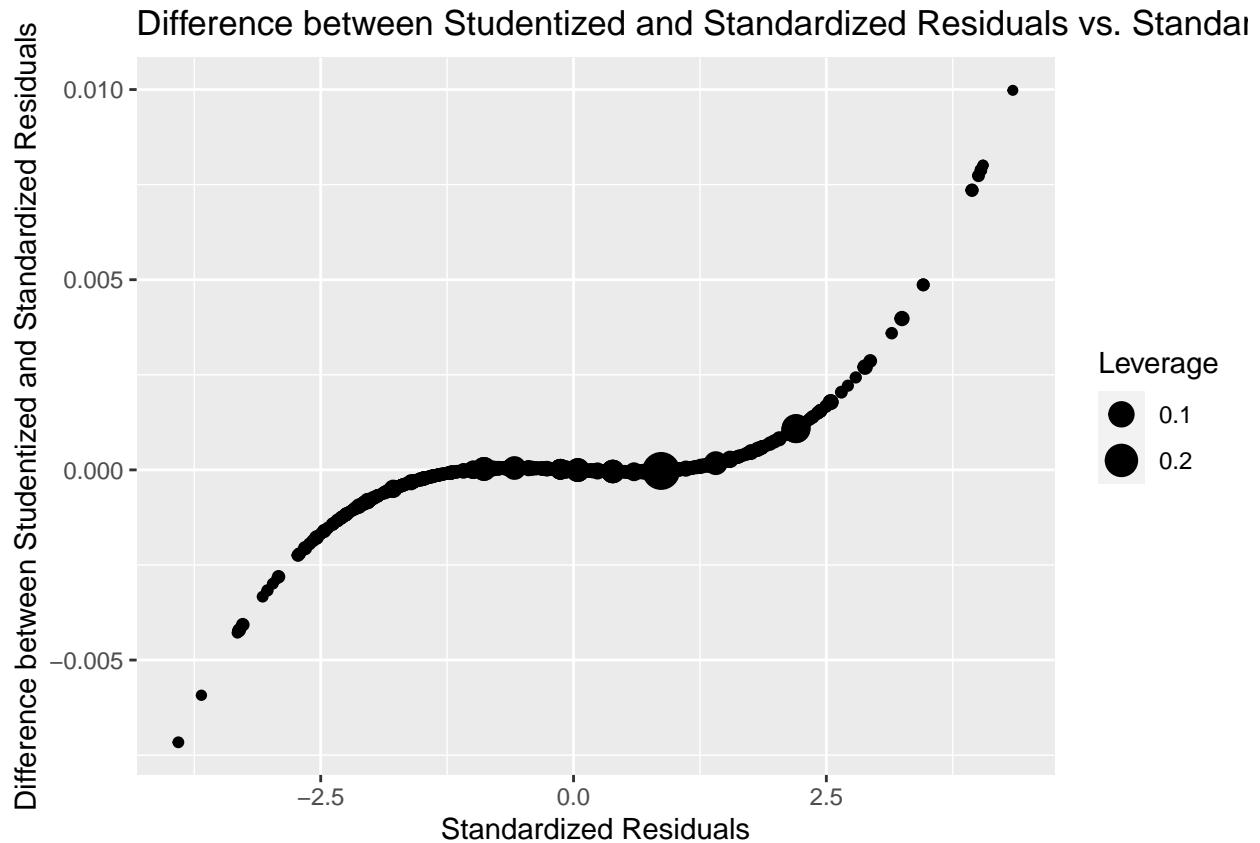
# Calculate standardized and studentized residuals
residual_2 <- rstandard(fit0)
residual_3 <- rstudent(fit0)

# Calculate leverage values
leverage_values <- hatvalues(fit0)

# Create a data frame for plotting
plot_data <- data.frame(
  Standardized_Residuals = residual_2,
  Difference = residual_3 - residual_2,
  Leverage = leverage_values
)

# Create the plot
ggplot(plot_data, aes(x = Standardized_Residuals, y = Difference)) +
  geom_point(aes(size = Leverage)) +
  ggtitle("Difference between Studentized and Standardized Residuals vs. Standardized Residuals") +
  xlab("Standardized Residuals") +
  ylab("Difference between Studentized and Standardized Residuals")

```



```

# Display the plot
print(ggplot)

## function (data = NULL, mapping = aes(), ..., environment = parent.frame())
## {
##   UseMethod("ggplot")
## }
## <bytecode: 0x531ad20>
## <environment: namespace:ggplot2>

# Load necessary library
library(ggplot2)

# Assuming fit0 is your linear model
# fit0 <- lm(SleepMinNight ~ ., data = df3)

# Calculate studentized and externally studentized residuals
residual_3 <- rstudent(fit0)
residual_4 <- rstudent(fit0) # Externally studentized residuals are typically the same as studentized

# Regular residuals
residual_1 <- fit0$residuals

# Create a data frame for plotting
plot_data <- data.frame(
  Studentized_Residuals = residual_3,
  Difference = residual_4 - residual_3,
  Leverage = leverage(fit0))
  
```

```

    Residual_Squared = residual_1^2
}

# Create the plot
ggplot(plot_data, aes(x = Studentized_Residuals, y = Difference)) +
  geom_point(aes(size = Residual_Squared)) +
  ggtitle("Difference between Externally Studentized and Studentized Residuals vs. Studentized Residuals") +
  xlab("Studentized Residuals") +
  ylab("Difference between Externally Studentized and Studentized Residuals")

```

Difference between Externally Studentized and Studentized Residuals

Residual_Squared

- 10
- 20

```

# Display the plot
print(ggplot)

## function (data = NULL, mapping = aes(), ..., environment = parent.frame())
## {
##   UseMethod("ggplot")
## }
## <bytecode: 0x531ad20>
## <environment: namespace:ggplot2>
# Load necessary library
library(ggplot2)

# Assuming fit0 is your linear model
# fit0 <- lm(SleepMinNight ~ ., data = df3)

# Calculate regular residuals

```

```

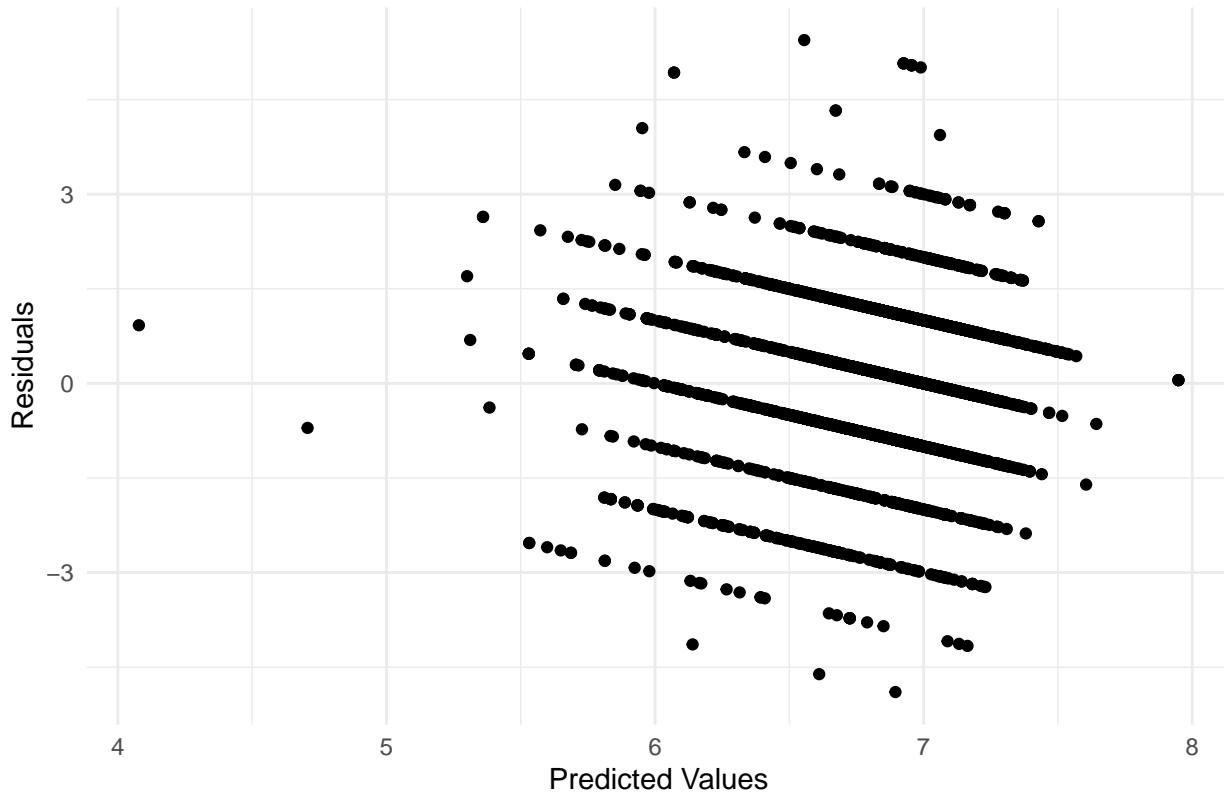
residual_1 <- fit0$residuals

# Get predicted values from the model
predicted_values <- predict(fit0)

# Create the plot
ggplot() +
  geom_point(aes(x = predicted_values, y = residual_1)) +
  ggtitle("Residuals vs. Predicted Values") +
  xlab("Predicted Values") +
  ylab("Residuals") +
  theme_minimal()

```

Residuals vs. Predicted Values



```

# Display the plot
print(ggplot)

## function (data = NULL, mapping = aes(), ..., environment = parent.frame())
## {
##   UseMethod("ggplot")
## }
## <bytecode: 0x531ad20>
## <environment: namespace:ggplot2>

# Load necessary library
library(ggplot2)

# Assuming fit0 is your linear model

```

```

# fit0 <- lm(SleepMinNight ~ ., data = df3)

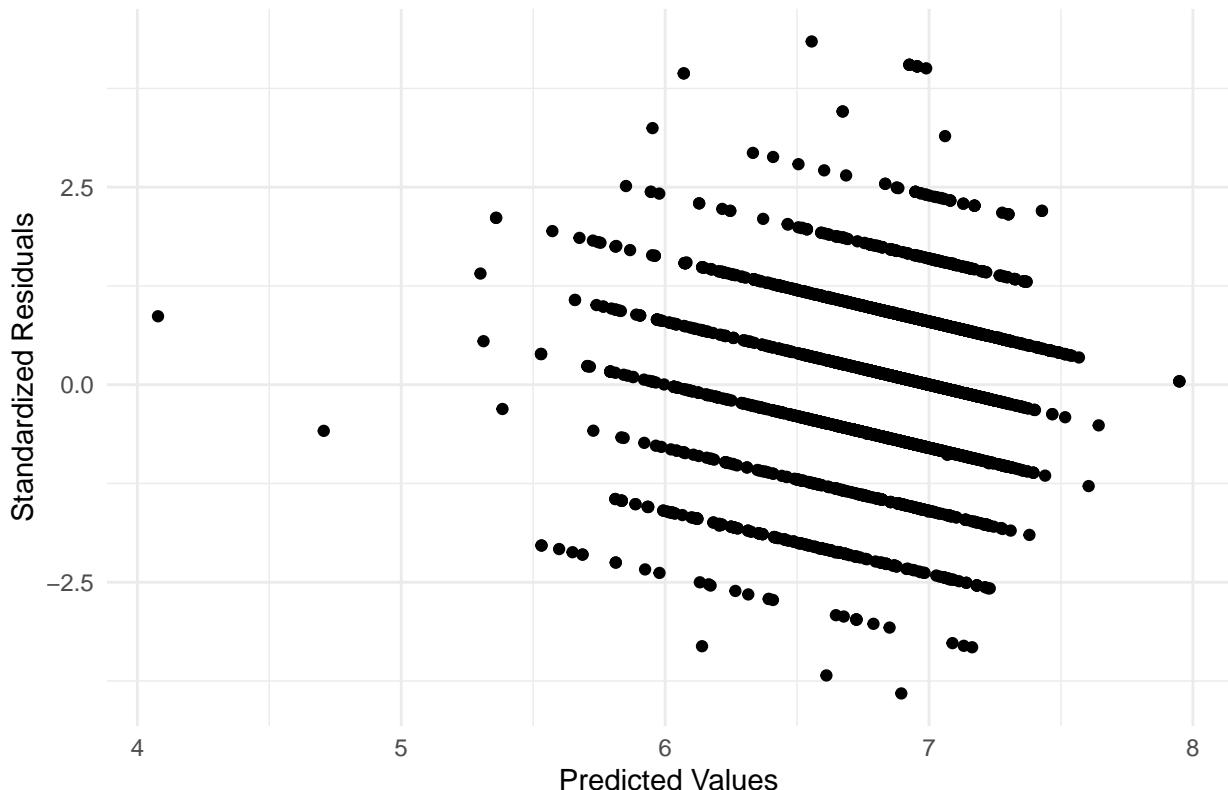
# Calculate different types of residuals
residual_2 <- rstandard(fit0)
residual_3 <- rstUDENT(fit0)
residual_4 <- rstUDENT(fit0) # Externally studentized residuals

# Get predicted values from the model
predicted_values <- predict(fit0)

# Plot for Standardized Residuals
ggplot() +
  geom_point(aes(x = predicted_values, y = residual_2)) +
  ggtitle("Standardized Residuals vs. Predicted Values") +
  xlab("Predicted Values") +
  ylab("Standardized Residuals") +
  theme_minimal()

```

Standardized Residuals vs. Predicted Values

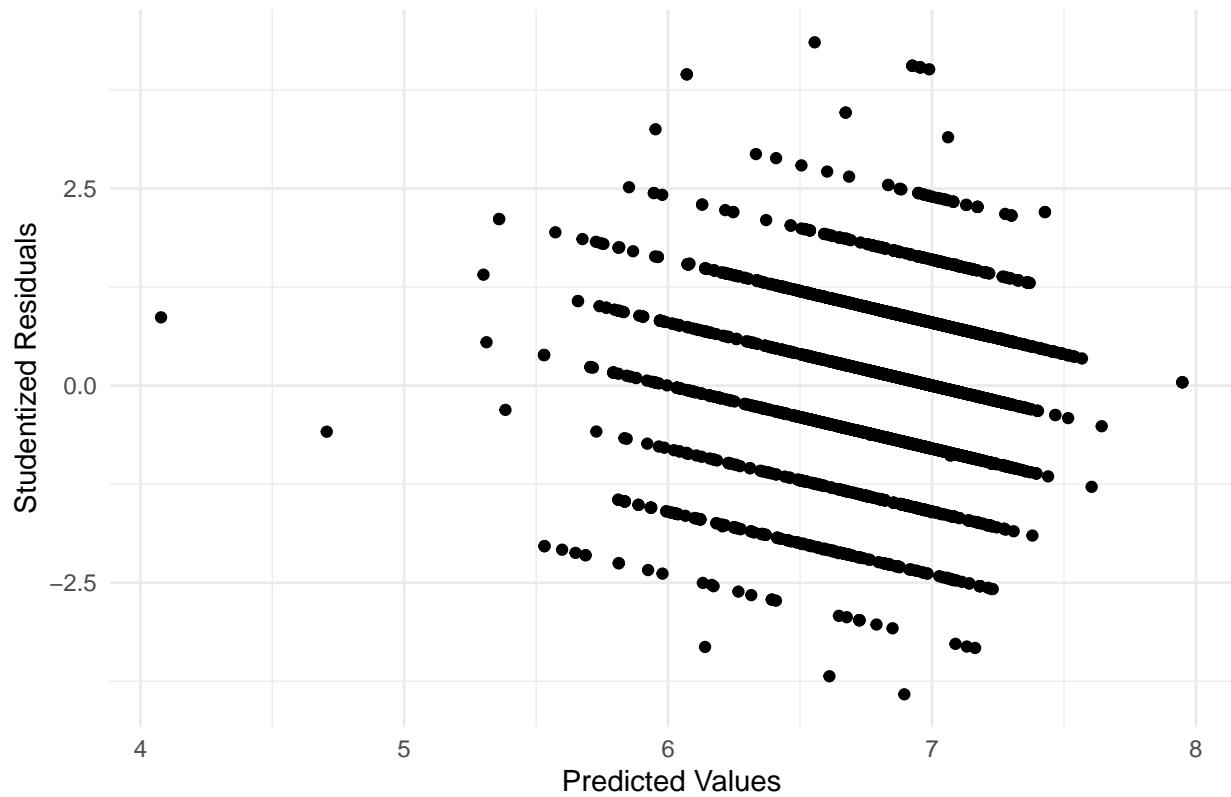


```

# Plot for Studentized Residuals
ggplot() +
  geom_point(aes(x = predicted_values, y = residual_3)) +
  ggtitle("Studentized Residuals vs. Predicted Values") +
  xlab("Predicted Values") +
  ylab("Studentized Residuals") +
  theme_minimal()

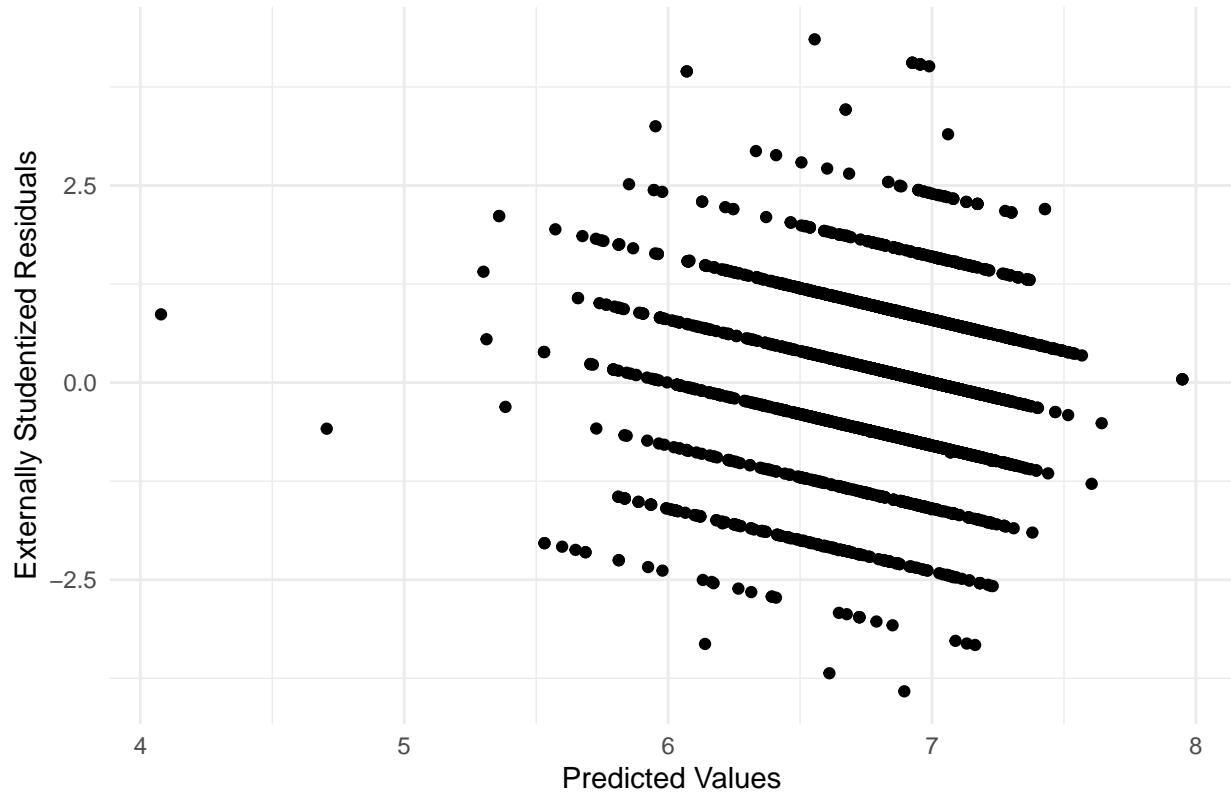
```

Studentized Residuals vs. Predicted Values



```
# Plot for Externally Studentized Residuals
ggplot() +
  geom_point(aes(x = predicted_values, y = residual_4)) +
  ggtitle("Externally Studentized Residuals vs. Predicted Values") +
  xlab("Predicted Values") +
  ylab("Externally Studentized Residuals") +
  theme_minimal()
```

Externally Studentized Residuals vs. Predicted Values



(5) Model Selection

```

step(fit0)

## Start: AIC=1796.94
## SleepHrsNight ~ BMI + DirectChol + Age + Gender + Race1 + TotChol +
##      BPDiaAve + BPSysAve + AlcoholYear + Poverty + HomeRooms +
##      SexNumPartnLife + SexNumPartYear + DaysMentHlthBad
##
##          Df Sum of Sq    RSS    AIC
## - TotChol     1   0.387 6145.1 1795.2
## - BPDiaAve   1   1.147 6145.9 1795.7
## - BPSysAve   1   2.270 6147.0 1796.4
## - BMI        1   2.872 6147.6 1796.8
## <none>           6144.7 1796.9
## - DirectChol  1   3.700 6148.4 1797.3
## - AlcoholYear 1  11.211 6155.9 1802.1
## - SexNumPartYear 1  12.425 6157.1 1802.9
## - HomeRooms   1  12.586 6157.3 1803.0
## - SexNumPartnLife 1  17.677 6162.4 1806.2
## - Poverty     1  24.042 6168.8 1810.3
## - Race1       4  33.645 6178.4 1810.4
## - Age         1  46.071 6190.8 1824.2
## - Gender      1  49.306 6194.0 1826.3

```

```

## - DaysMentHlthBad 1 187.449 6332.2 1912.8
##
## Step: AIC=1798.4
## SleepHrsNight ~ BMI + DirectChol + Age + Gender + Race1 + BPDiaAve +
##      BPSysAve + AlcoholYear + Poverty + HomeRooms + SexNumPartnLife +
##      SexNumPartYear + DaysMentHlthBad
##
## Call:
## lm(formula = SleepHrsNight ~ BMI + DirectChol + Age + Gender +
##      Race1 + BPDiaAve + BPSysAve + AlcoholYear + Poverty + HomeRooms +
##      SexNumPartnLife + SexNumPartYear + DaysMentHlthBad, data = df3)
##
## Coefficients:
## (Intercept)          BMI        DirectChol         Age
## 7.0066602       -0.0045456      -0.0868565     -0.0103630
## Gender            Race1        BPDiaAve        BPSysAve
## 0.2492065       0.0709518       0.0017997     -0.0021224
## AlcoholYear       Poverty       HomeRooms    SexNumPartnLife
## 0.0005835       0.0485063       0.0262798     -0.0011571
## SexNumPartYear  DaysMentHlthBad
## 0.0185115      -0.0283090
library(olsrr)

##
## Attaching package: 'olsrr'
## The following object is masked from 'package:datasets':
##
##   rivers
ols_step_forward_p(fit0,penter=0.1,details=F)

##
##                                     Selection Summary
## -----
##   Variable           Entered      R-Square      Adj.
## Step
## -----
## 1 DaysMentHlthBad 0.0319 0.0316 135.1234 13044.1166 1.2757
## 2 Gender 0.0401 0.0396 102.7354 13012.6991 1.2704
## 3 Race1 0.0454 0.0446 82.6655 12993.1011 1.2671
## 4 Age 0.0512 0.0502 60.3184 12971.1357 1.2634
## 5 Poverty 0.0570 0.0558 38.1552 12949.2058 1.2597
## 6 SexNumPartnLife 0.0591 0.0577 31.1607 12942.2607 1.2584
## 7 SexNumPartYear 0.0610 0.0593 25.1899 12936.3166 1.2573
## 8 HomeRooms 0.0628 0.0609 19.7213 12930.8583 1.2563
## 9 AlcoholYear 0.0641 0.0619 16.4559 12927.5917 1.2556
## -----
ols_step_forward_p(fit0,penter=0.05,details=F)

```

```

##
##                                     Selection Summary
## -----
##   Variable           Entered      R-Square      Adj.
## Step
## -----
## 1 DaysMentHlthBad 0.0319 0.0316 135.1234 13044.1166 1.2757
## 2 Gender 0.0401 0.0396 102.7354 13012.6991 1.2704
## 3 Race1 0.0454 0.0446 82.6655 12993.1011 1.2671
## 4 Age 0.0512 0.0502 60.3184 12971.1357 1.2634
## 5 Poverty 0.0570 0.0558 38.1552 12949.2058 1.2597
## 6 SexNumPartnLife 0.0591 0.0577 31.1607 12942.2607 1.2584
## 7 SexNumPartYear 0.0610 0.0593 25.1899 12936.3166 1.2573
## 8 HomeRooms 0.0628 0.0609 19.7213 12930.8583 1.2563
## 9 AlcoholYear 0.0641 0.0619 16.4559 12927.5917 1.2556
## -----

```

## Step	Entered	R-Square	R-Square	C(p)	AIC	RMSE
## 1	DaysMentHlthBad	0.0319	0.0316	135.1234	13044.1166	1.2757
## 2	Gender	0.0401	0.0396	102.7354	13012.6991	1.2704
## 3	Race1	0.0454	0.0446	82.6655	12993.1011	1.2671
## 4	Age	0.0512	0.0502	60.3184	12971.1357	1.2634
## 5	Poverty	0.0570	0.0558	38.1552	12949.2058	1.2597
## 6	SexNumPartnLife	0.0591	0.0577	31.1607	12942.2607	1.2584
## 7	SexNumPartYear	0.0610	0.0593	25.1899	12936.3166	1.2573
## 8	HomeRooms	0.0628	0.0609	19.7213	12930.8583	1.2563
## 9	AlcoholYear	0.0641	0.0619	16.4559	12927.5917	1.2556
##						
ols_mallows_cp(model =m_logfull_1, fullmodel =m_full) # Mallows' Cp						
## [1] -3821.538						
ols_mallows_cp(model =m_logfull_2, fullmodel =m_full) # Mallows' Cp						
## [1] 10.74774						
ols_mallows_cp(model =m_sqfull_1, fullmodel =m_full) # Mallows' Cp						
## [1] 11.76658						