project

Liancheng

2023-11-21

(1) Data cleaning

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

```
##          used (Mb) gc trigger (Mb) max used (Mb)
## Ncells 469578 25.1    1011221 54.1   660860 35.3
## Vcells 877810  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)]
df <- df[!duplicated(df), ]
# 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
```

```
library(car)
```

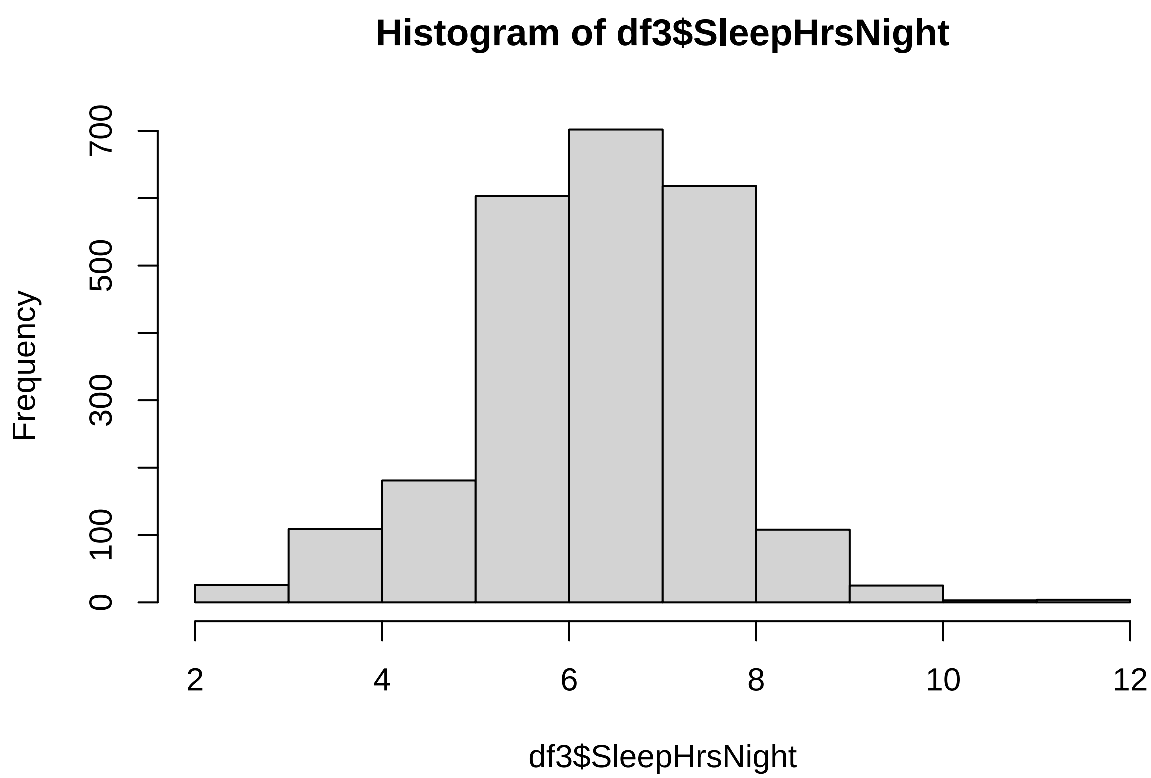
```
## Loading required package: carData

##
## Attaching package: 'car'

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

```
df2 <- df %>% select(
  SleepHrsNight,
  BMI,
```

```r
  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 2379   6.81  1.31   7.00    6.87  1.48  2.00   12.00
## BMI                2 2379  28.78  6.79  27.52   28.10  5.95 15.02   69.00
## DirectChol         3 2379   1.34  0.41   1.27    1.30  0.39  0.39    3.83
## Age                4 2379  38.65 11.58  39.00   38.60 14.83 18.00   59.00
## Gender*            5 2379   1.54  0.50   2.00    1.55  0.00  1.00    2.00
## Race1*             6 2379   3.43  1.16   4.00    3.56  0.00  1.00    5.00
## TotChol            7 2379   5.06  1.05   4.99    5.00  1.04  1.53   13.65
## BPDiaAve           8 2379  71.25 11.63  71.00   71.33 10.38  0.00  116.00
## BPSysAve           9 2379 117.55 14.40 116.00  116.59 13.34 78.00  226.00
## AlcoholYear       10 2379  68.91 93.00  24.00   49.33 35.58  0.00  364.00
## Poverty           11 2379   2.79  1.69   2.71    2.83  2.42  0.00    5.00
## HomeRooms         12 2379   6.02  2.24   6.00    5.88  1.48  1.00   13.00
## SexNumPartnLife   13 2379  15.97 63.17   6.00    8.52  5.93  0.00 2000.00
## SexNumPartYear    14 2379   1.37  2.76   1.00    1.00  0.00  0.00   69.00
## DaysMentHlthBad   15 2379   4.42  7.98   0.00    2.35  0.00  0.00   30.00
##                   range  skew kurtosis   se
## SleepHrsNight     10.00 -0.30     0.64 0.03
## BMI               53.98  1.23     2.65 0.14
## DirectChol         3.44  1.14     2.57 0.01
## Age               41.00  0.02    -1.16 0.24
## Gender*            1.00 -0.15    -1.98 0.01
## Race1*             4.00 -1.10     0.06 0.02
## TotChol           12.12  0.88     3.24 0.02
## BPDiaAve         116.00 -0.36     3.11 0.24
## BPSysAve         148.00  1.12     3.93 0.30
## AlcoholYear      364.00  1.70     2.16 1.91
## Poverty            5.00  0.03    -1.48 0.03
## HomeRooms         12.00  0.64     0.43 0.05
## SexNumPartnLife 2000.00 19.57   497.14 1.30
## SexNumPartYear    69.00 13.35   250.92 0.06
## DaysMentHlthBad   30.00  2.19     3.88 0.16
```

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




```r
# 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      2379  Observations
## ---------------------------------------------------------------------------------
## SleepHrsNight
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2379        0       11     0.94    6.807    1.417        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    23   109   181   603   702   618   108    25     3     4
## Proportion 0.001 0.010 0.046 0.076 0.253 0.295 0.260 0.045 0.011 0.001 0.002
## ---------------------------------------------------------------------------------
## BMI
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2379        0     1139        1    28.78    7.297    20.15    21.39
##      .25      .50      .75      .90      .95
##    23.97    27.52    32.20    37.47    41.46
##
## 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
##     2379        0       99    0.999     1.34   0.4424     0.80     0.88
##      .25      .50      .75      .90      .95
##     1.06     1.27     1.55     1.86     2.07
##
## 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
##     2379        0       42    0.999    38.65    13.37       21       23
##      .25      .50      .75      .90      .95
##       29       39       48       55       57
##
## lowest : 18 19 20 21 22, highest: 55 56 57 58 59
## ---------------------------------------------------------------------------------
## Gender
##        n  missing distinct     Info      Sum     Mean      Gmd
##     2379        0        2    0.746     1102   0.4632   0.4975
##
## ---------------------------------------------------------------------------------
## Race1
##        n  missing distinct     Info     Mean      Gmd
##     2379        0        5     0.77    3.427    1.127
##
```

```
## lowest : 1 2 3 4 5, highest: 1 2 3 4 5
##
## Value          1     2     3     4     5
## Frequency    318   160   271  1447   183
## Proportion 0.134 0.067 0.114 0.608 0.077
## ---------------------------------------------------------------------------
## TotChol
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2379        0      212        1    5.057     1.15     3.54     3.83
##      .25      .50      .75      .90      .95
##     4.32     4.99     5.69     6.36     6.83
##
## 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
##     2379        0       84    0.999    71.25    12.62       53       58
##      .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
##     2379        0       99    0.999    117.6    15.47     98.0    101.8
##      .25      .50      .75      .90      .95
##    108.0    116.0    125.0    134.0    142.0
##
## lowest :  78  83  84  85  86, highest: 184 191 202 209 226
## ---------------------------------------------------------------------------
## AlcoholYear
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2379        0       56    0.993    68.91    90.19        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
##     2379        0      398    0.989    2.794    1.932    0.329    0.658
##      .25      .50      .75      .90      .95
##    1.225    2.710    4.710    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
##     2379        0       13    0.978    6.024    2.459        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     25    34   168   408   441   438   331   213   134    93    42
## Proportion 0.011 0.014 0.071 0.172 0.185 0.184 0.139 0.090 0.056 0.039 0.018
##
## Value         12    13
## Frequency     26    26
## Proportion 0.011 0.011
## ---------------------------------------------------------------------------
## SexNumPartnLife
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2379        0       81    0.996    15.97    21.68        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
##     2379        0       22    0.683    1.374    1.258        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
##     2379        0       28    0.842    4.422    6.829        0        0
##      .25      .50      .75      .90      .95
##        0        0        5       15       30
##
## lowest :  0  1  2  3  4, highest: 25 26 27 29 30
## ---------------------------------------------------------------------------
```

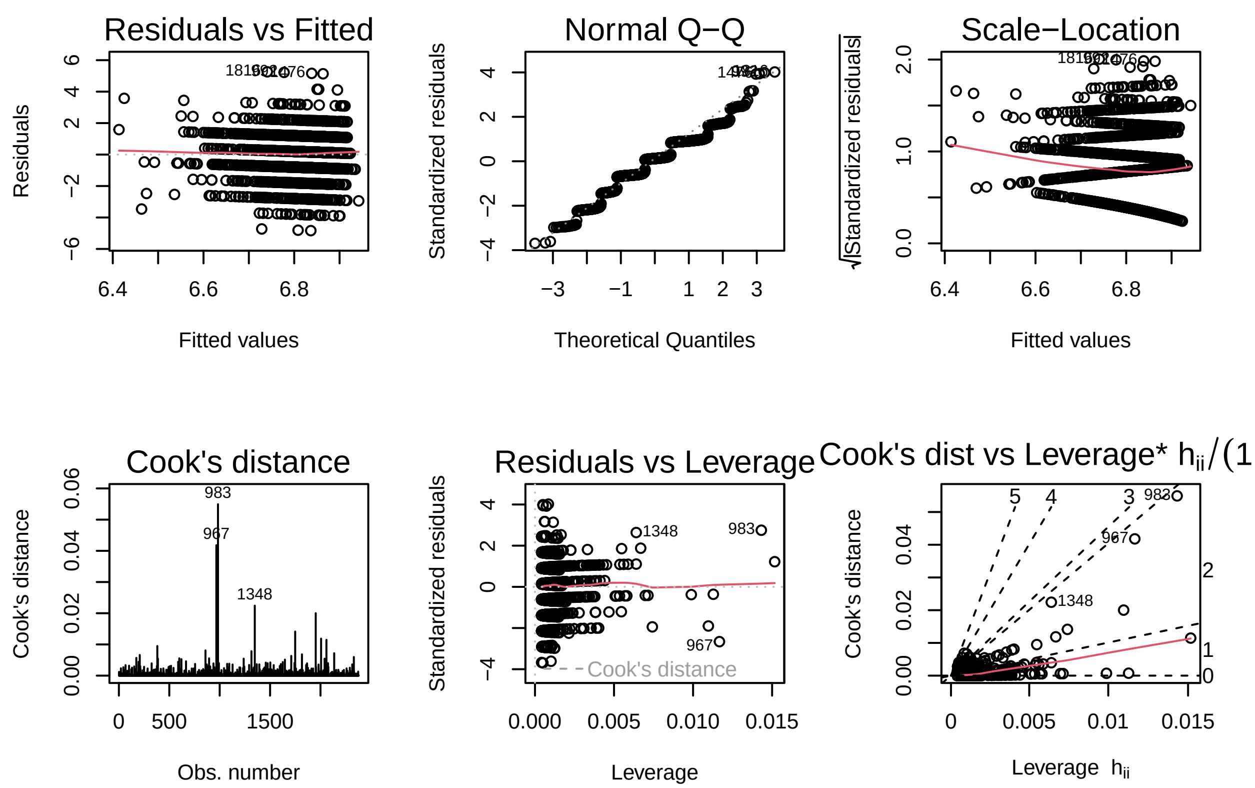
(3) linear regression model

```
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
## -4.8366 -0.8209  0.1606  1.1457  5.2593
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.089190   0.116862  60.663   <2e-16 ***
## df3$BMI     -0.009790   0.003953  -2.477   0.0133 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1.309 on 2377 degrees of freedom
## Multiple R-squared:  0.002574,   Adjusted R-squared:  0.002155
## F-statistic: 6.135 on 1 and 2377 DF,  p-value: 0.01332
```

```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)
```




```r
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
## -4.9314 -0.8178  0.1258  1.0506  5.3685
```

```
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     6.926204   0.165453  41.862  < 2e-16 ***
## BMI            -0.006542   0.003986  -1.641  0.10086
## Age            -0.008653   0.002331  -3.712  0.00021 ***
## Gender          0.190406   0.053607   3.552  0.00039 ***
## factor(Race1)2  0.219456   0.126253   1.738  0.08230 .
## factor(Race1)3  0.463251   0.107923   4.292 1.84e-05 ***
## factor(Race1)4  0.364311   0.081381   4.477 7.94e-06 ***
## factor(Race1)5  0.346075   0.121784   2.842  0.00453 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.298 on 2371 degrees of freedom
## Multiple R-squared:  0.02277,    Adjusted R-squared:  0.01988
## F-statistic: 7.892 on 7 and 2371 DF,  p-value: 1.734e-09
```

```
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
## -4.9890 -0.8396  0.0721  0.9860  5.3001
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      7.0320813  0.2778091  25.313  < 2e-16 ***
## BMI             -0.0044747  0.0040584  -1.103 0.270326
## Age             -0.0094332  0.0024860  -3.794 0.000152 ***
## Gender           0.2375920  0.0557852   4.259 2.13e-05 ***
## factor(Race1)2   0.2194679  0.1252138   1.753 0.079775 .
## factor(Race1)3   0.4198626  0.1074580   3.907 9.60e-05 ***
## factor(Race1)4   0.3457651  0.0809536   4.271 2.02e-05 ***
## factor(Race1)5   0.3059503  0.1211322   2.526 0.011610 *
## TotChol          0.0047497  0.0266203   0.178 0.858404
## BPDiaAve         0.0005694  0.0027009   0.211 0.833058
## BPSysAve        -0.0010233  0.0022623  -0.452 0.651072
## AlcoholYear      0.0004989  0.0002951   1.691 0.091045 .
## DaysMentHlthBad -0.0263313  0.0033210  -7.929 3.38e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.281 on 2366 degrees of freedom
## Multiple R-squared:  0.0491, Adjusted R-squared:  0.04427
```

```
## F-statistic: 10.18 on 12 and 2366 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)
```

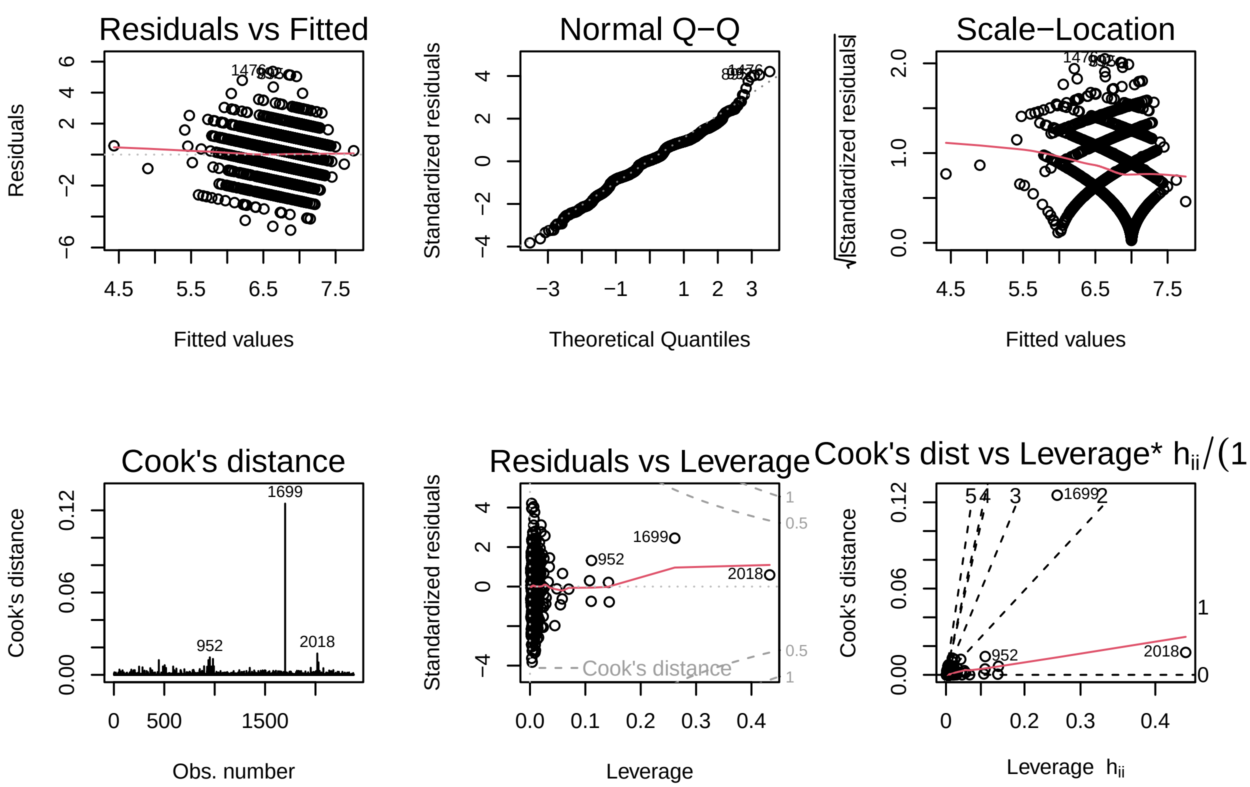
```
##
## Call:
## lm(formula = SleepHrsNight ~ BMI + Age + Gender + factor(Race1) +
##     TotChol + BPDiaAve + BPSysAve + AlcoholYear + DaysMentHlthBad +
##     HomeRooms + SexNumPartnLife + SexNumPartYear + Poverty, data = df3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8773 -0.8413  0.0550  0.9634  5.3687
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      6.7900913  0.2880350  23.574  < 2e-16 ***
## BMI             -0.0040314  0.0040533  -0.995 0.320044
## Age             -0.0108102  0.0025767  -4.195 2.82e-05 ***
## Gender           0.2295409  0.0559260   4.104 4.19e-05 ***
## factor(Race1)2   0.2329616  0.1249975   1.864 0.062483 .
## factor(Race1)3   0.4443994  0.1077464   4.124 3.84e-05 ***
## factor(Race1)4   0.3135085  0.0815958   3.842 0.000125 ***
## factor(Race1)5   0.2899386  0.1211140   2.394 0.016746 *
## TotChol          0.0053548  0.0265885   0.201 0.840408
## BPDiaAve         0.0006679  0.0026998   0.247 0.804627
## BPSysAve        -0.0005452  0.0022628  -0.241 0.809634
## AlcoholYear      0.0003952  0.0002967   1.332 0.182986
## DaysMentHlthBad -0.0247441  0.0033503  -7.386 2.09e-13 ***
## HomeRooms        0.0213263  0.0127462   1.673 0.094431 .
## SexNumPartnLife -0.0009946  0.0004243  -2.344 0.019168 *
## SexNumPartYear   0.0149274  0.0097588   1.530 0.126243
## Poverty          0.0372512  0.0173927   2.142 0.032314 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.278 on 2362 degrees of freedom
## Multiple R-squared:  0.05601,    Adjusted R-squared:  0.04962
## F-statistic: 8.759 on 16 and 2362 DF,  p-value: < 2.2e-16
```

```
vif(m_full)
```

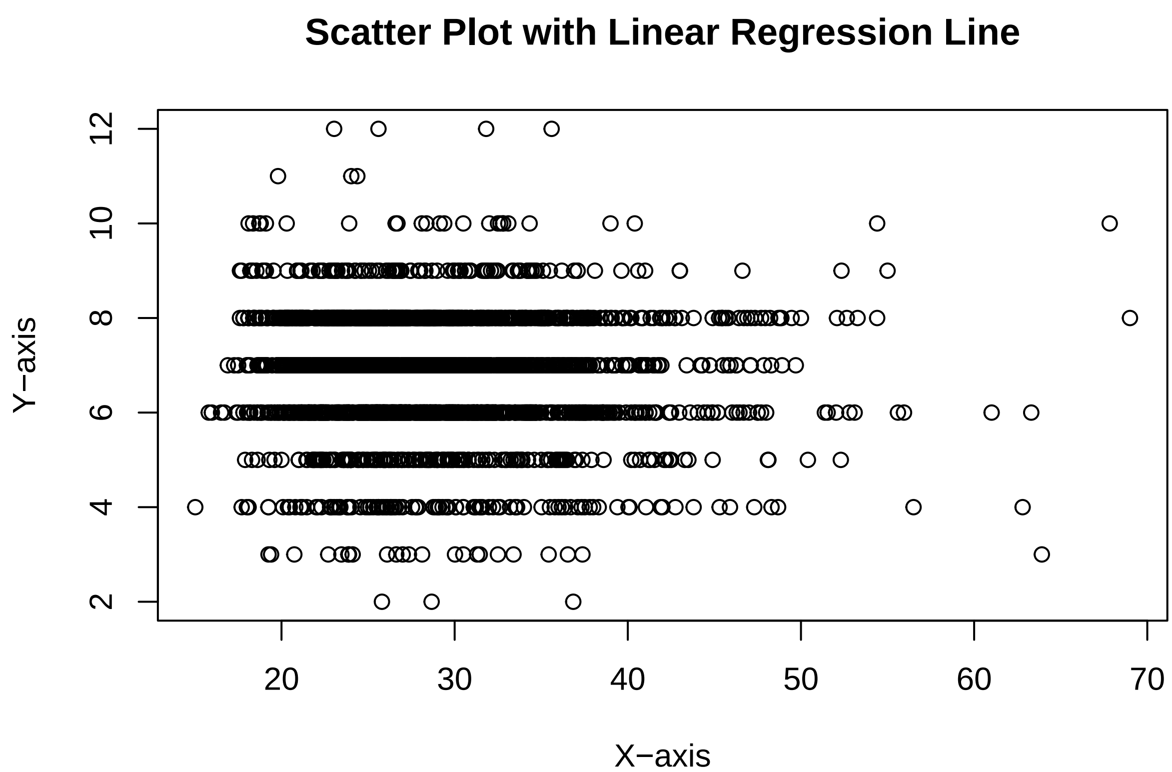
```
##                     GVIF Df GVIF^(1/(2*Df))
## BMI             1.104174  1        1.050797
## Age             1.296882  1        1.138807
## Gender          1.133022  1        1.064435
## factor(Race1)   1.208657  4        1.023972
## TotChol         1.133282  1        1.064557
## BPDiaAve        1.436892  1        1.198704
```

```
## BPSysAve         1.545631  1        1.243234
## AlcoholYear      1.108683  1        1.052940
## DaysMentHlthBad 1.041358  1        1.020469
## HomeRooms        1.182949  1        1.087635
## SexNumPartnLife 1.046533  1        1.023002
## SexNumPartYear   1.053685  1        1.026491
## Poverty          1.261580  1        1.123201
```

```r
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)
```




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

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

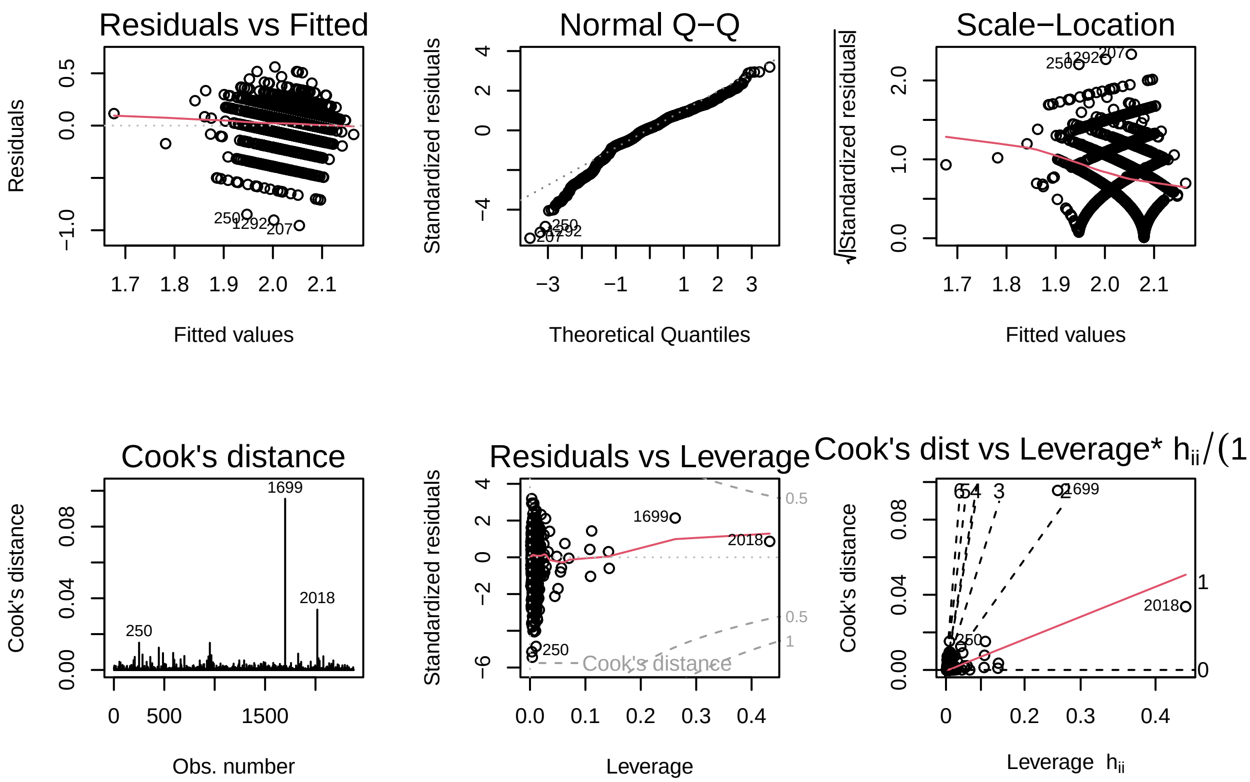



```
#log x
df3$logBMI = log(df3$BMI + 1)
df3$logSleepHrsNight = log(df3$SleepHrsNight + 1)
df3$logDaysMentHlthBad = log(df3$DaysMentHlthBad + 1)
df3$invTotChol = 1 / df3$TotChol
df3$sqrtDaysMentHlthBad = sqrt(df3$DaysMentHlthBad)
df3$sqBMI = (df3$BMI - mean(df3$BMI)) ^ 2
m_logfull_2 = lm(
  logSleepHrsNight ~ Age + Gender + factor(Race1) + logBMI + invTotChol +
    BPDiaAve + BPSysAve + AlcoholYear + sqrtDaysMentHlthBad + HomeRooms + SexNumPartnLife +
    SexNumPartYear + Poverty,
  df3
)
summary(m_logfull_2)
```

```
##
## Call:
## lm(formula = logSleepHrsNight ~ Age + Gender + factor(Race1) +
##     logBMI + invTotChol + BPDiaAve + BPSysAve + AlcoholYear +
##     sqrtDaysMentHlthBad + HomeRooms + SexNumPartnLife + SexNumPartYear +
##     Poverty, data = df3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.95516 -0.09798  0.01973  0.12503  0.56082
##
```

```
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)    
## (Intercept)          2.091e+00  7.094e-02  29.476  < 2e-16 ***
## Age                 -1.567e-03  3.540e-04  -4.427 9.99e-06 ***
## Gender               3.130e-02  7.727e-03   4.050 5.28e-05 ***
## factor(Race1)2       3.462e-02  1.719e-02   2.014  0.04416 *  
## factor(Race1)3       6.283e-02  1.482e-02   4.239 2.33e-05 ***
## factor(Race1)4       4.723e-02  1.121e-02   4.212 2.63e-05 ***
## factor(Race1)5       4.506e-02  1.666e-02   2.705  0.00688 ** 
## logBMI              -2.052e-02  1.776e-02  -1.156  0.24797    
## invTotChol          -1.352e-02  8.866e-02  -0.152  0.87885    
## BPDiaAve             1.927e-04  3.715e-04   0.519  0.60410    
## BPSysAve            -5.826e-05  3.116e-04  -0.187  0.85168    
## AlcoholYear          6.167e-05  4.089e-05   1.508  0.13167    
## sqrtDaysMentHlthBad -1.748e-02  2.215e-03  -7.891 4.55e-15 ***
## HomeRooms            3.031e-03  1.755e-03   1.727  0.08428 .  
## SexNumPartnLife     -1.467e-04  5.844e-05  -2.511  0.01211 *  
## SexNumPartYear       1.880e-03  1.344e-03   1.399  0.16203    
## Poverty              6.249e-03  2.396e-03   2.608  0.00917 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.1759 on 2362 degrees of freedom
## Multiple R-squared:  0.06159,    Adjusted R-squared:  0.05523 
## F-statistic: 9.689 on 16 and 2362 DF,  p-value: < 2.2e-16
```

```r
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)
```




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

(4) Diagnosis: 10-fold CV

```r
library(caret)
```

```
## Loading required package: ggplot2

## Loading required package: lattice
```

```r
splitIndex <-
  createDataPartition(df3$SleepHrsNight, p = 0.7, list = FALSE)
trainData <- df3[splitIndex, ]
testData <- df3[-splitIndex, ]
predictions <- predict(m_logfull_2, newdata = testData)
mse <- mean((testData$SleepHrsNight - predictions) ^ 2)
control <-
  trainControl(method = "cv", number = 10)  # 10-fold cross-validation
cv_model <-
  train(
    SleepHrsNight ~ .,
    data = df3,
    method = "lm",
    trControl = control
  )
```

```r
cv_model
```

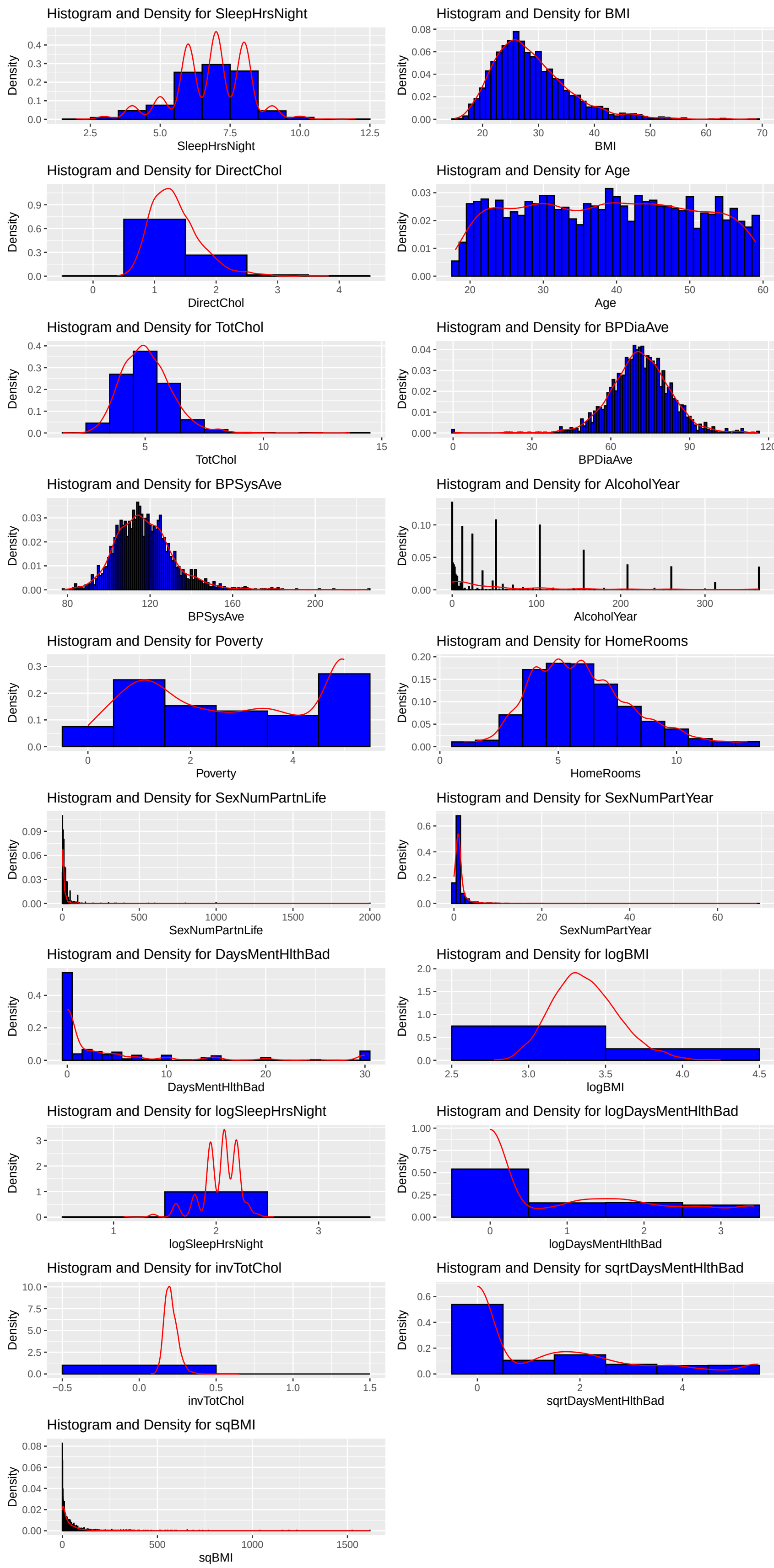
```
## Linear Regression
##
## 2379 samples
##   20 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 2141, 2140, 2141, 2141, 2142, 2142, ...
## Resampling results:
##
##   RMSE       Rsquared  MAE
##   0.1900161  0.979765  0.1266116
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```r
(cv_results <- cv_model$results)
```

```
##   intercept      RMSE Rsquared       MAE     RMSESD  RsquaredSD      MAESD
## 1      TRUE 0.1900161 0.979765 0.1266116 0.03595476 0.005748305 0.00897214
```

(4) Diagnosis: Normality Assumption

```r
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)
```



Histogram and Density for SleepHrsNight
Density
SleepHrsNight
Histogram and Density for BMI
Density
BMI
Histogram and Density for DirectChol
Density
DirectChol
Histogram and Density for Age
Density
Age
Histogram and Density for TotChol
Density
TotChol
Histogram and Density for BPDiaAve
Density
BPDiaAve
Histogram and Density for BPSysAve
Density
BPSysAve
Histogram and Density for AlcoholYear
Density
AlcoholYear
Histogram and Density for Poverty
Density
Poverty
Histogram and Density for HomeRooms
Density
HomeRooms
Histogram and Density for SexNumPartnLife
Density
SexNumPartnLife
Histogram and Density for SexNumPartYear
Density
SexNumPartYear
Histogram and Density for DaysMentHlthBad
Density
DaysMentHlthBad
Histogram and Density for logBMI
Density
logBMI
Histogram and Density for logSleepHrsNight
Density
logSleepHrsNight
Histogram and Density for logDaysMentHlthBad
Density
logDaysMentHlthBad
Histogram and Density for invTotChol
Density
invTotChol
Histogram and Density for sqrtDaysMentHlthBad
Density
sqrtDaysMentHlthBad
Histogram and Density for sqBMI
Density
sqBMI

```r
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 0.9354644 6.065754e-31 1.273808e-29
## BMI                                  BMI 0.9301559 5.692561e-32 1.195438e-30
## DirectChol                    DirectChol 0.9405789 6.876212e-30 1.444005e-28
## Age                                  Age 0.9582706 1.360245e-25 2.856515e-24
## Gender                            Gender 0.6346474 1.545071e-57 3.244648e-56
## Race1                              Race1 0.7427298 1.802728e-51 3.785730e-50
## TotChol                          TotChol 0.9663542 3.785914e-23 7.950419e-22
## BPDiaAve                        BPDiaAve 0.9726214 6.250883e-21 1.312685e-19
## BPSysAve                        BPSysAve 0.9484045 3.946229e-28 8.287082e-27
## AlcoholYear                  AlcoholYear 0.7403964 1.270928e-51 2.668949e-50
## Poverty                          Poverty 0.8951549 1.570942e-37 3.298979e-36
## HomeRooms                      HomeRooms 0.9553923 2.237881e-26 4.699550e-25
## SexNumPartnLife          SexNumPartnLife 0.1509787 1.499112e-73 3.148134e-72
## SexNumPartYear            SexNumPartYear 0.2545353 5.992229e-71 1.258368e-69
## DaysMentHlthBad          DaysMentHlthBad 0.6076574 8.193380e-59 1.720610e-57
## logBMI                            logBMI 0.9877235 1.946304e-13 4.087239e-12
## logSleepHrsNight        logSleepHrsNight 0.8984084 4.408251e-37 9.257327e-36
## logDaysMentHlthBad    logDaysMentHlthBad 0.7729598 2.157265e-49 4.530256e-48
## invTotChol                    invTotChol 0.9572292 7.005059e-26 1.471062e-24
## sqrtDaysMentHlthBad sqrtDaysMentHlthBad 0.7619387 3.557376e-50 7.470490e-49
## sqBMI                              sqBMI 0.4152373 3.203569e-66 6.727494e-65
```

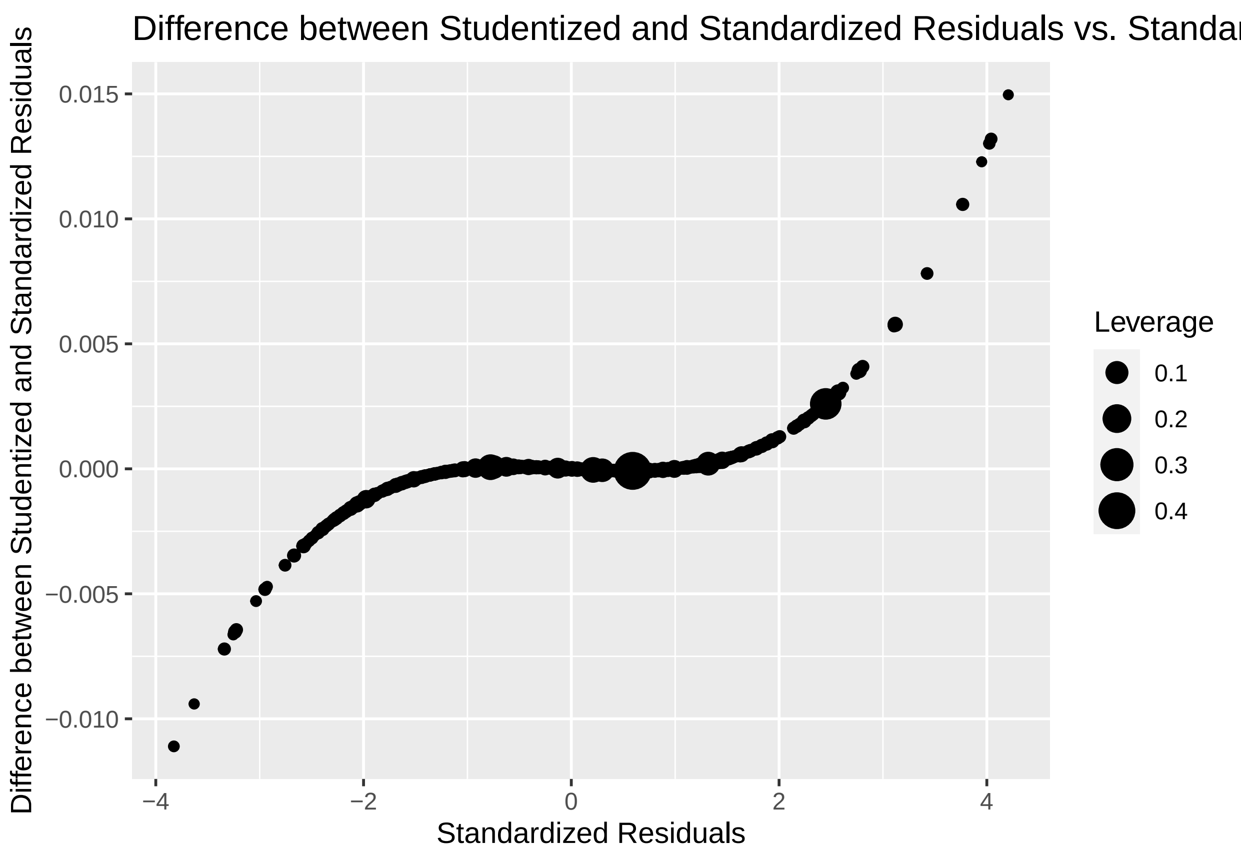
Standardized residuals, Studentized residuals

```r
# Regular residuals
residual_1 <- fit0$residuals

# Standardized residuals
```

```r
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 7.060476e-17 1.273554 -4.878636 5.368822
## 2           Standardized 1.960572e-04 1.000639 -3.826002 4.207084
## 3            Studentized 1.585968e-04 1.001202 -3.837105 4.222048
## 4 Externally Studentized 1.585968e-04 1.001202 -3.837105 4.222048
```

```r
# 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")
```

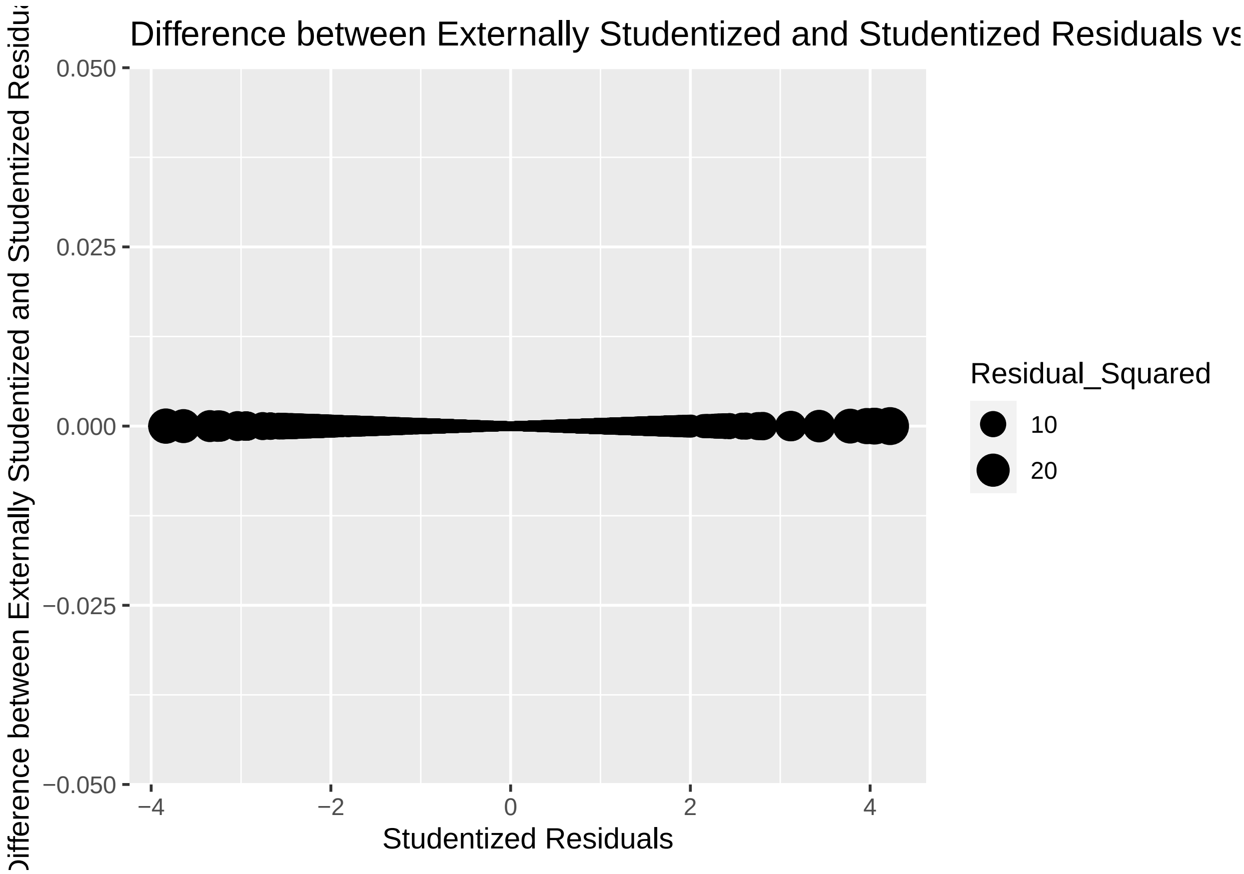



```r
# Display the plot
print(ggplot)
```

```
## function (data = NULL, mapping = aes(), ..., environment = parent.frame())
## {
##     UseMethod("ggplot")
## }
## <bytecode: 0x4e12680>
## <environment: namespace:ggplot2>
```

```r
# 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,
```

```r
  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")
```

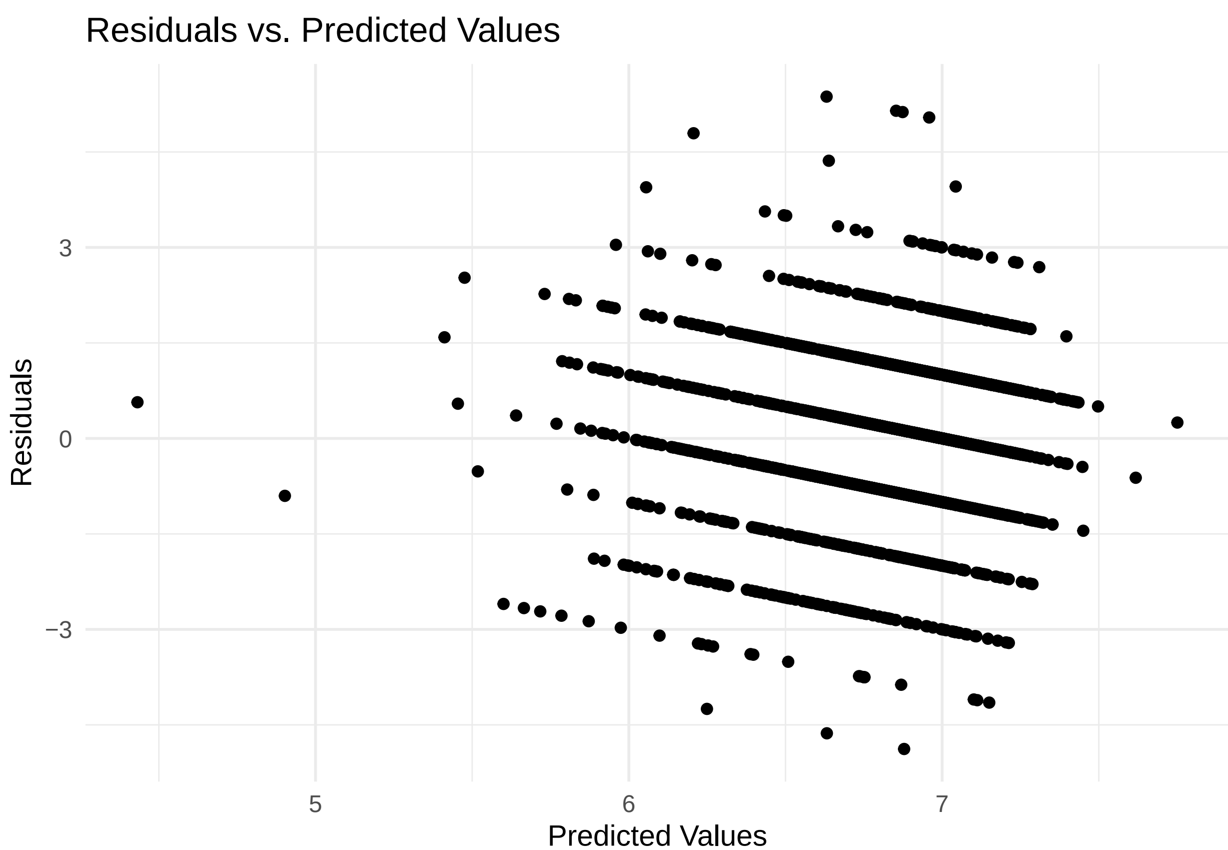



```r
# Display the plot
print(ggplot)
```

```
## function (data = NULL, mapping = aes(), ..., environment = parent.frame())
## {
##     UseMethod("ggplot")
## }
## <bytecode: 0x4e12680>
## <environment: namespace:ggplot2>
```

```r
# Load necessary library
library(ggplot2)

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

# Calculate regular residuals
```

```r
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()
```

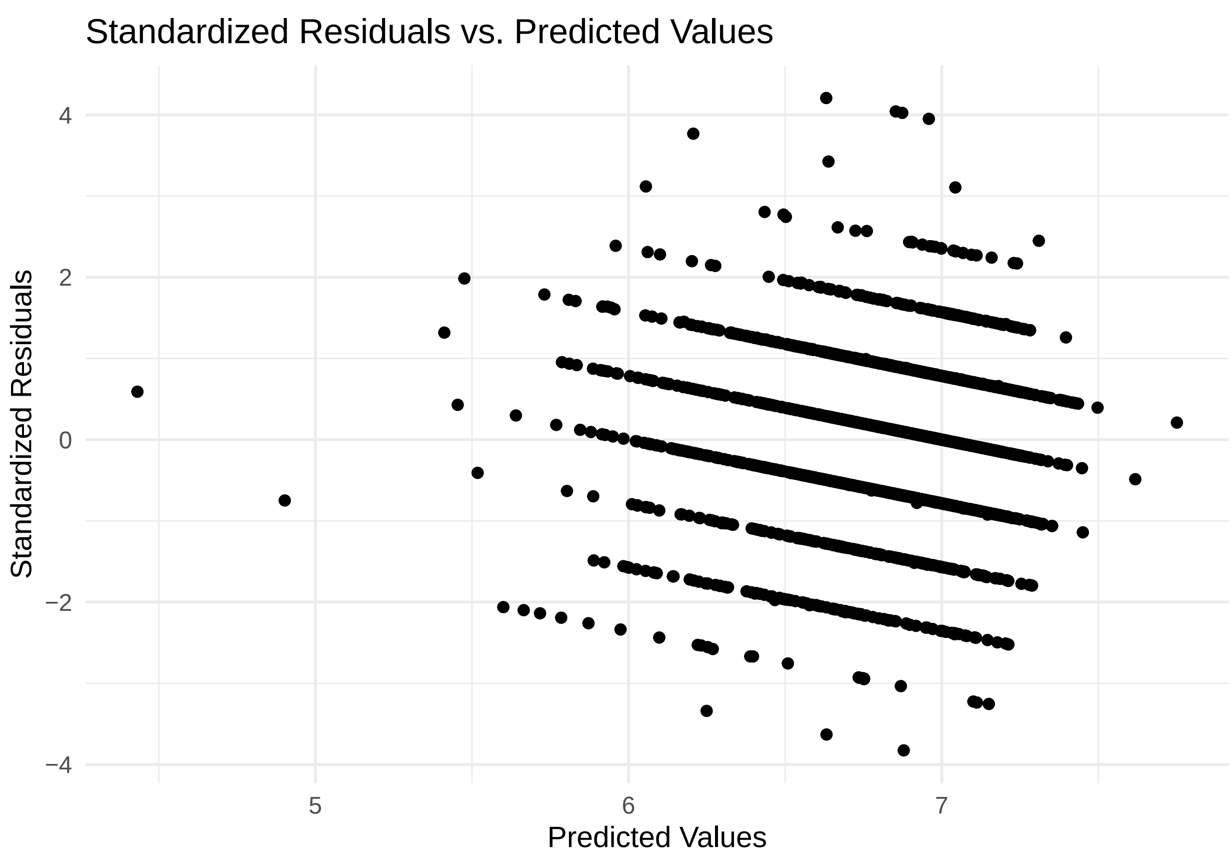



```r
# Display the plot
print(ggplot)
```

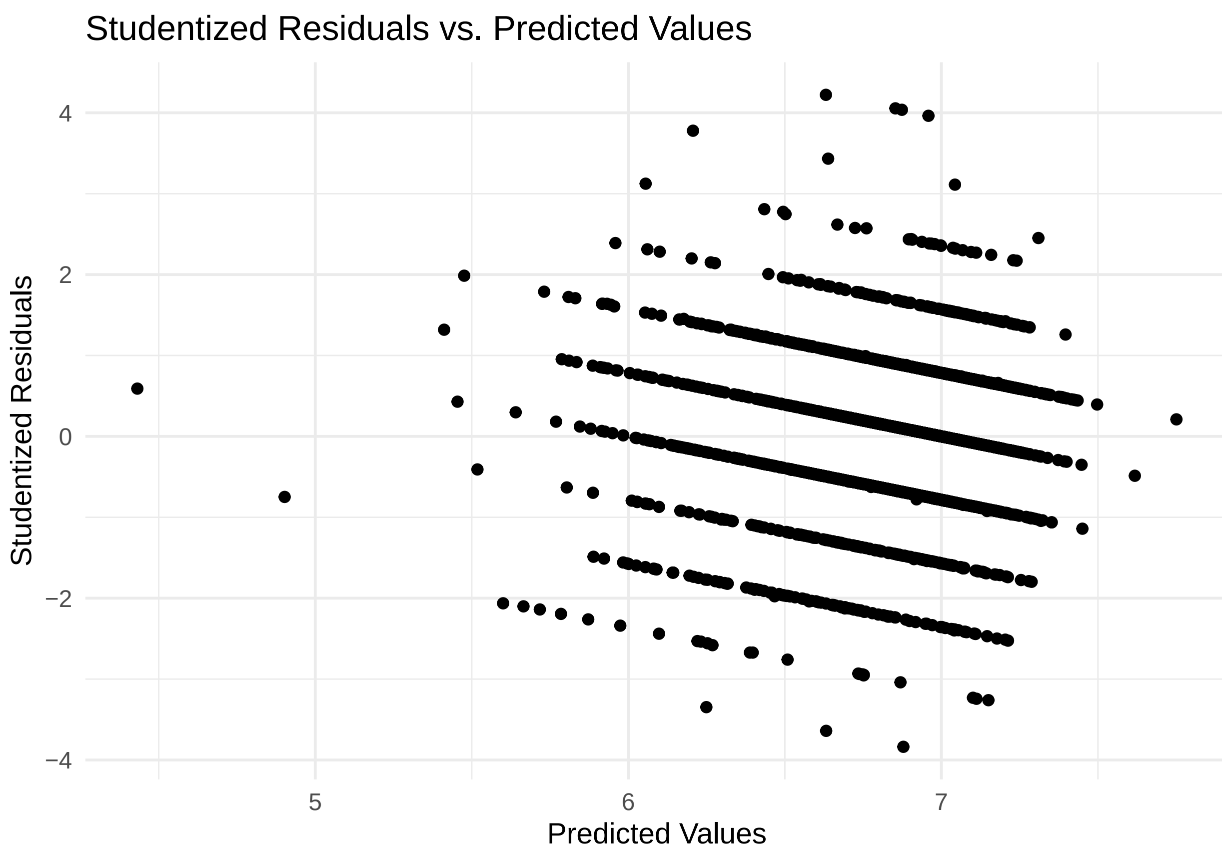
```
## function (data = NULL, mapping = aes(), ..., environment = parent.frame())
## {
##     UseMethod("ggplot")
## }
## <bytecode: 0x4e12680>
## <environment: namespace:ggplot2>
```

```r
# Load necessary library
library(ggplot2)

# Assuming fit0 is your linear model
```

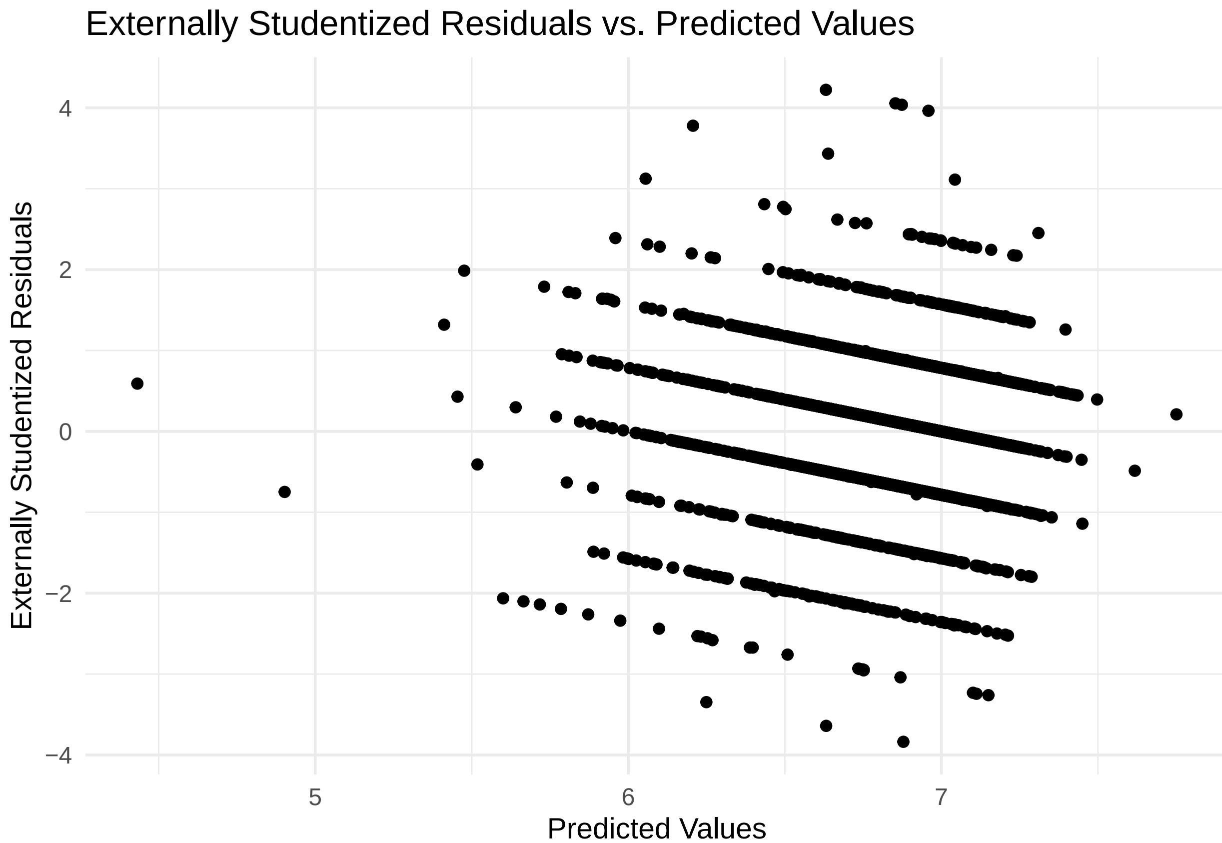
```r
# fit0 <- lm(SleepMinNight ~ ., data = df3)

# 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()
```




```r
# 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()
```




```r
# 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()
```



(5) Model Selection

```
step(fit0)
```

```
## Start:  AIC=1185.54
## SleepHrsNight ~ BMI + DirectChol + Age + Gender + Race1 + TotChol +
##     BPDiaAve + BPSysAve + AlcoholYear + Poverty + HomeRooms +
##     SexNumPartnLife + SexNumPartYear + DaysMentHlthBad
##
##                   Df Sum of Sq    RSS    AIC
## - DirectChol       1     0.003 3857.0 1183.5
## - TotChol          1     0.069 3857.0 1183.6
## - BPSysAve         1     0.093 3857.1 1183.6
## - BPDiaAve         1     0.098 3857.1 1183.6
## - BMI              1     1.435 3858.4 1184.4
## - AlcoholYear      1     2.773 3859.7 1185.2
## <none>                         3857.0 1185.5
## - SexNumPartYear   1     3.823 3860.8 1185.9
## - HomeRooms        1     4.571 3861.5 1186.4
## - Poverty          1     7.466 3864.4 1188.1
## - SexNumPartnLife  1     8.973 3865.9 1189.1
## - Race1            4    32.638 3889.6 1197.6
## - Gender           1    23.929 3880.9 1198.2
## - Age              1    28.718 3885.7 1201.2
```

```
## - DaysMentHlthBad  1    89.039 3946.0 1237.8
##
## Step:  AIC=1187.2
## SleepHrsNight ~ BMI + Age + Gender + Race1 + TotChol + BPDiaAve +
##     BPSysAve + AlcoholYear + Poverty + HomeRooms + SexNumPartnLife +
##     SexNumPartYear + DaysMentHlthBad
##
## Call:
## lm(formula = SleepHrsNight ~ BMI + Age + Gender + Race1 + TotChol +
##     BPDiaAve + BPSysAve + AlcoholYear + Poverty + HomeRooms +
##     SexNumPartnLife + SexNumPartYear + DaysMentHlthBad, data = df3)
##
## Coefficients:
##     (Intercept)              BMI              Age           Gender
##       6.8656069       -0.0040545       -0.0107790        0.2153535
##           Race1          TotChol         BPDiaAve         BPSysAve
##       0.0766796        0.0095890        0.0003656       -0.0007014
##     AlcoholYear          Poverty        HomeRooms  SexNumPartnLife
##       0.0003533        0.0303914        0.0198583       -0.0010323
##  SexNumPartYear  DaysMentHlthBad
##       0.0146137       -0.0253312
```

```r
library(olsrr)
```

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

```r
ols_step_forward_p(m_full, penter = 0.1, details = F)
```

```
##
##                                  Selection Summary
## --------------------------------------------------------------------------------
##        Variable                      Adj.
## Step        Entered        R-Square    R-Square     C(p)         AIC         RMSE
## --------------------------------------------------------------------------------
##    1    DaysMentHlthBad      0.0242      0.0238    66.5525    7985.6561    1.2951
##    2    Gender               0.0311      0.0303    51.4035    7970.8875    1.2908
##    3    Age                  0.0373      0.0361    37.8570    7957.5830    1.2869
##    4    factor(Race1)        0.0471      0.0443    15.3744    7941.2801    1.2815
##    5    Poverty              0.0507      0.0475     8.3826    7934.2917    1.2793
##    6    SexNumPartnLife      0.0525      0.0489     5.7740    7931.6716    1.2783
##    7    HomeRooms            0.0536      0.0496     4.9984    7930.8847    1.2779
## --------------------------------------------------------------------------------
```

```r
ols_step_forward_p(m_full, penter = 0.05, details = F)
```

```
##
##                                  Selection Summary
## --------------------------------------------------------------------------------
##        Variable                      Adj.
## Step        Entered        R-Square    R-Square     C(p)         AIC         RMSE
## --------------------------------------------------------------------------------
```

```
## 1 DaysMentHlthBad 0.0242 0.0238 66.5525 7985.6561 1.2951
## 2 Gender 0.0311 0.0303 51.4035 7970.8875 1.2908
## 3 Age 0.0373 0.0361 37.8570 7957.5830 1.2869
## 4 factor(Race1) 0.0471 0.0443 15.3744 7941.2801 1.2815
## 5 Poverty 0.0507 0.0475 8.3826 7934.2917 1.2793
## 6 SexNumPartnLife 0.0525 0.0489 5.7740 7931.6716 1.2783
## -----------------------------------------------------------------------------------
```

```r
ols_mallows_cp(model = m_logfull_2, fullmodel = m_full)  # Mallows' Cp
```

```
## [1] -2306.233
```