BIOSTAT650_Final_Project

Liancheng, He Zhang, Zhengrui Huang, Zibo Yu

2023-11-21

(1) Data cleaning

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

```
##          used (Mb) gc trigger (Mb) max used (Mb)
## Ncells 469679 25.1    1011510 54.1   660860 35.3
## Vcells 878642  6.8    8388608 64.0  1800812 13.8
```

```
set.seed(123)
############### (1) Data cleaning ########################################
## select variables
library(NHANES)
library(car)
```

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

```
df0 <- NHANES
df <- NHANES[NHANES$Age >= 18 & NHANES$Age < 60,]
# colSums(is.na(df)) / nrow(df)
df <- df[, which(colSums(is.na(df)) / nrow(df) < 0.3)]
# exclude duplication
df <- df[!duplicated(df),]
names(df)
```

```
##  [1] "ID"              "SurveyYr"        "Gender"          "Age"
##  [5] "AgeDecade"       "Race1"           "Education"       "MaritalStatus"
##  [9] "HHIncome"        "HHIncomeMid"     "Poverty"         "HomeRooms"
## [13] "HomeOwn"         "Work"            "Weight"          "Height"
## [17] "BMI"             "BMI_WHO"         "Pulse"           "BPSysAve"
## [21] "BPDiaAve"        "BPSys1"          "BPDia1"          "BPSys2"
## [25] "BPDia2"          "BPSys3"          "BPDia3"          "DirectChol"
## [29] "TotChol"         "UrineVol1"       "UrineFlow1"      "Diabetes"
## [33] "HealthGen"       "DaysPhysHlthBad" "DaysMentHlthBad" "LittleInterest"
## [37] "Depressed"       "SleepHrsNight"   "SleepTrouble"    "PhysActive"
## [41] "Alcohol12PlusYr" "AlcoholYear"     "Smoke100"        "Smoke100n"
## [45] "Marijuana"       "RegularMarij"    "HardDrugs"       "SexEver"
## [49] "SexAge"          "SexNumPartnLife" "SexNumPartYear"  "SameSex"
## [53] "SexOrientation"
```

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

```
##
```

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

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

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

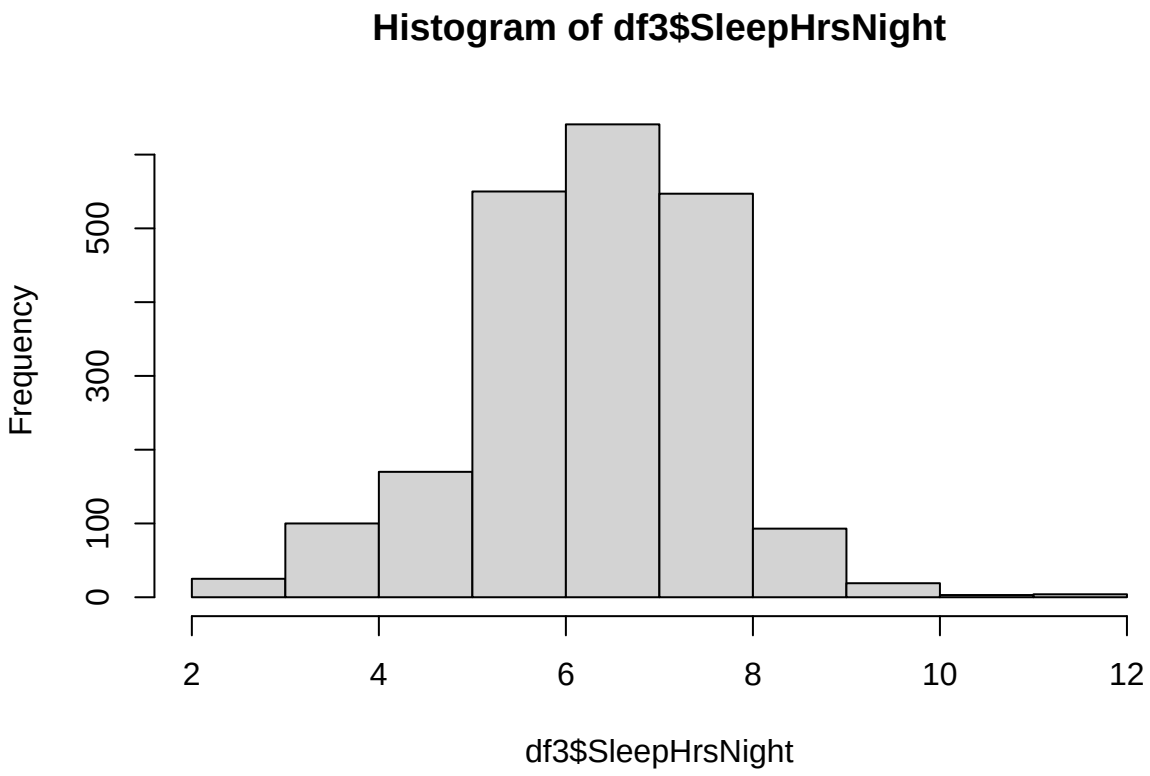
```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
df2 <- df %>% select(
  SleepHrsNight,
  BMI,
  DirectChol,
  Age,
  Gender,
  Race1,
  TotChol,
  BPDiaAve,
  BPSysAve,
  AlcoholYear,
  Poverty,
  SexNumPartnLife,
  SexNumPartYear,
  DaysMentHlthBad,
  UrineFlow1,
  PhysActive,
  DaysPhysHlthBad,
  Smoke100,
  Depressed,
  HealthGen,
  SexAge
)

df3 <- na.omit(df2)
#df3$SleepHrsNight <- df3$SleepHrsNight * 60
#df3 <- df3[, -which(names(df3) %in% "SleepHrsNight")]
# cor(df3$BPSysAve,df3$BPDiaAve)
psych::describe(df3)
```

```
##                 vars    n   mean    sd median trimmed   mad   min     max
## SleepHrsNight      1 2152   6.78  1.31   7.00    6.85  1.48  2.00   12.00
## BMI                2 2152  28.77  6.75  27.60   28.09  5.78 15.02   69.00
## DirectChol         3 2152   1.35  0.41   1.29    1.31  0.39  0.39    3.83
## Age                4 2152  39.18 11.33  39.00   39.15 14.83 20.00   59.00
## Gender*            5 2152   1.53  0.50   2.00    1.54  0.00  1.00    2.00
## Race1*             6 2152   3.43  1.15   4.00    3.57  0.00  1.00    5.00
## TotChol            7 2152   5.07  1.05   4.99    5.01  1.04  1.53   13.65
## BPDiaAve           8 2152  71.19 11.84  71.00   71.28 10.38  0.00  116.00
## BPSysAve           9 2152 117.43 14.28 116.00  116.50 13.34 78.00  209.00
## AlcoholYear       10 2152  70.59 94.22  24.00   50.94 35.58  0.00  364.00
## Poverty           11 2152   2.84  1.69   2.78    2.89  2.49  0.00    5.00
```

```
## SexNumPartnLife   12 2152  16.73 66.13   7.00    8.91  5.93  0.00 2000.00
## SexNumPartYear    13 2152   1.38  2.59   1.00    1.04  0.00  0.00   69.00
## DaysMentHlthBad   14 2152   4.47  8.02   0.00    2.40  0.00  0.00   30.00
## UrineFlow1        15 2152   1.07  0.97   0.81    0.91  0.60  0.00   10.14
## PhysActive*       16 2152   1.58  0.49   2.00    1.60  0.00  1.00    2.00
## DaysPhysHlthBad   17 2152   3.16  7.19   0.00    1.12  0.00  0.00   30.00
## Smoke100*         18 2152   1.46  0.50   1.00    1.45  0.00  1.00    2.00
## Depressed*        19 2152   1.30  0.58   1.00    1.16  0.00  1.00    3.00
## HealthGen*        20 2152   2.64  0.94   3.00    2.65  1.48  1.00    5.00
## SexAge            21 2152  17.10  3.39  17.00   16.80  2.97  9.00   44.00
##                    range  skew kurtosis   se
## SleepHrsNight     10.00 -0.30     0.69 0.03
## BMI               53.98  1.28     2.96 0.15
## DirectChol         3.44  1.09     2.27 0.01
## Age               39.00  0.02    -1.15 0.24
## Gender*            1.00 -0.12    -1.99 0.01
## Race1*             4.00 -1.13     0.08 0.02
## TotChol           12.12  0.92     3.47 0.02
## BPDiaAve         116.00 -0.39     3.13 0.26
## BPSysAve         131.00  1.00     2.94 0.31
## AlcoholYear      364.00  1.66     1.98 2.03
## Poverty            5.00 -0.01    -1.47 0.04
## SexNumPartnLife 2000.00 18.82   456.62 1.43
## SexNumPartYear    69.00 14.07   293.16 0.06
## DaysMentHlthBad   30.00  2.16     3.76 0.17
## UrineFlow1        10.14  2.89    14.06 0.02
## PhysActive*        1.00 -0.32    -1.90 0.01
## DaysPhysHlthBad   30.00  2.80     7.06 0.15
## Smoke100*          1.00  0.15    -1.98 0.01
## Depressed*         2.00  1.83     2.21 0.01
## HealthGen*         4.00  0.11    -0.33 0.02
## SexAge            35.00  1.51     5.56 0.07
```

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




```
# colSums(is.na(df2)) / nrow(df2)
fit0 <-
  lm(SleepHrsNight ~ .,
     data = df3)
#data type
df3$Gender <- ifelse(df3$Gender == "male", 0, 1)
df3 <- 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
##
```

```
##  21  Variables      2152  Observations
## ---------------------------------------------------------------------------
## SleepHrsNight
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2152        0       11     0.94    6.781    1.415        4        5
##      .25      .50      .75      .90      .95
##        6        7        8        8        9
##
## lowest :  2  3  4  5  6, highest:  8  9 10 11 12
##
## Value          2     3     4     5     6     7     8     9    10    11    12
## Frequency      3    22   100   170   550   641   547    93    19     3     4
## Proportion 0.001 0.010 0.046 0.079 0.256 0.298 0.254 0.043 0.009 0.001 0.002
## ---------------------------------------------------------------------------
## BMI
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2152        0     1072        1    28.77    7.223    20.18    21.50
##      .25      .50      .75      .90      .95
##    24.00    27.60    32.00    37.36    41.22
##
## lowest : 15.02 15.80 15.98 16.51 16.70, highest: 62.80 63.30 63.91 67.83 69.00
## ---------------------------------------------------------------------------
## DirectChol
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2152        0       98    0.999    1.346   0.4446     0.80     0.91
##      .25      .50      .75      .90      .95
##     1.06     1.29     1.58     1.89     2.09
##
## lowest : 0.39 0.41 0.52 0.54 0.57, highest: 3.13 3.41 3.44 3.59 3.83
## ---------------------------------------------------------------------------
## Age
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2152        0       40    0.999    39.18    13.08       21       23
##      .25      .50      .75      .90      .95
##       30       39       49       55       57
##
## lowest : 20 21 22 23 24, highest: 55 56 57 58 59
## ---------------------------------------------------------------------------
## Gender
##        n  missing distinct     Info      Sum     Mean      Gmd
##     2152        0        2    0.747     1011   0.4698   0.4984
##
## ---------------------------------------------------------------------------
## Race1
##        n  missing distinct     Info     Mean      Gmd
##     2152        0        5    0.758    3.428    1.115
##
## lowest : 1 2 3 4 5, highest: 1 2 3 4 5
##
## Value          1     2     3     4     5
## Frequency    289   145   230  1333   155
## Proportion 0.134 0.067 0.107 0.619 0.072
## ---------------------------------------------------------------------------
## TotChol
```

```
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2152        0      208        1    5.069    1.151     3.57     3.85
##      .25      .50      .75      .90      .95
##     4.32     4.99     5.69     6.36     6.83
##
## lowest :  1.53  2.69  2.74  2.79  2.82, highest:  9.31  9.34  9.90 12.28 13.65
## ---------------------------------------------------------------------------
## BPDiaAve
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2152        0       84    0.999    71.19    12.83       53       57
##      .25      .50      .75      .90      .95
##       64       71       78       85       89
##
## lowest :   0  20  21  22  25, highest: 108 109 110 114 116
## ---------------------------------------------------------------------------
## BPSysAve
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2152        0       98    0.999    117.4    15.44       97      101
##      .25      .50      .75      .90      .95
##      108      116      125      134      142
##
## lowest :  78  83  84  85  86, highest: 182 184 191 202 209
## ---------------------------------------------------------------------------
## AlcoholYear
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2152        0       56    0.993    70.59     91.9        0        0
##      .25      .50      .75      .90      .95
##        4       24      104      208      260
##
## lowest :   0   1   2   3   4, highest: 260 300 312 360 364
## ---------------------------------------------------------------------------
## Poverty
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2152        0      393    0.988    2.841    1.931    0.340    0.660
##      .25      .50      .75      .90      .95
##    1.277    2.780    4.817    5.000    5.000
##
## lowest : 0.00 0.02 0.03 0.04 0.05, highest: 4.95 4.96 4.97 4.99 5.00
## ---------------------------------------------------------------------------
## SexNumPartnLife
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2152        0       81    0.995    16.73    22.47        1        1
##      .25      .50      .75      .90      .95
##        3        7       15       30       50
##
## lowest :    0    1    2    3    4, highest:  600  800  999 1000 2000
## ---------------------------------------------------------------------------
## SexNumPartYear
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2152        0       21    0.645    1.381     1.18        0        0
##      .25      .50      .75      .90      .95
##        1        1        1        2        3
##
## lowest :  0  1  2  3  4, highest: 19 20 30 50 69
```

```
## ---------------------------------------------------------------------------
## DaysMentHlthBad
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2152        0       28    0.844    4.475    6.894        0        0
##      .25      .50      .75      .90      .95
##        0        0        5       15       30
##
## lowest :  0  1  2  3  4, highest: 25 26 27 29 30
## ---------------------------------------------------------------------------
## UrineFlow1
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2152        0     1337        1    1.074   0.9061   0.1960   0.2775
##      .25      .50      .75      .90      .95
##   0.4580   0.8100   1.3618   2.1929   2.7780
##
## lowest :  0.000  0.006  0.011  0.014  0.016, highest:  7.325  7.826  8.730  9.410 10.143
## ---------------------------------------------------------------------------
## PhysActive
##        n  missing distinct
##     2152        0        2
##
## Value         No   Yes
## Frequency    906  1246
## Proportion 0.421 0.579
## ---------------------------------------------------------------------------
## DaysPhysHlthBad
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2152        0       24    0.708    3.165    5.318     0.00     0.00
##      .25      .50      .75      .90      .95
##     0.00     0.00     2.00    10.00    24.45
##
## lowest :  0  1  2  3  4, highest: 24 25 26 28 30
## ---------------------------------------------------------------------------
## Smoke100
##        n  missing distinct
##     2152        0        2
##
## Value         No   Yes
## Frequency   1155   997
## Proportion 0.537 0.463
## ---------------------------------------------------------------------------
## Depressed
##        n  missing distinct
##     2152        0        3
##
## Value         None Several    Most
## Frequency     1657     355     140
## Proportion   0.770   0.165   0.065
## ---------------------------------------------------------------------------
## HealthGen
##        n  missing distinct
##     2152        0        5
##
## lowest : Excellent Vgood     Good      Fair      Poor
```

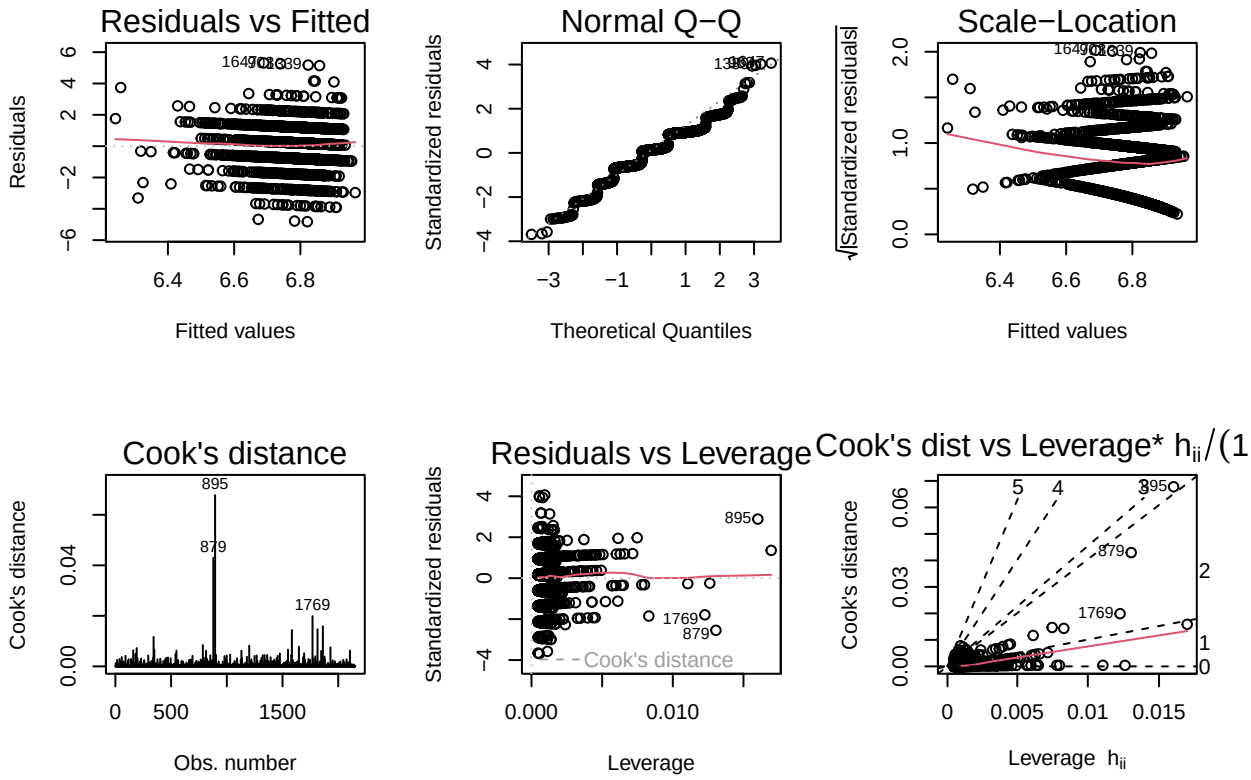
```
## highest: Excellent Vgood     Good      Fair      Poor
##
## Value      Excellent     Vgood      Good      Fair      Poor
## Frequency        240       697       854       313        48
## Proportion     0.112     0.324     0.397     0.145     0.022
## ------------------------------------------------------------------------------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

(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
## -4.8209 -0.8022  0.1710  1.1494  5.3105
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.166900   0.123331  58.111  < 2e-16 ***
## df3$BMI     -0.013409   0.004174  -3.213  0.00133 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.307 on 2150 degrees of freedom
## Multiple R-squared:  0.004778,   Adjusted R-squared:  0.004315
## F-statistic: 10.32 on 1 and 2150 DF,  p-value: 0.001334
```

```
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

dummy_b = 1 * (df3$Race1 == "Black")
dummy_h = 1 * (df3$Race1 == "Hispanic")
dummy_m = 1 * (df3$Race1 == "Mexican")
dummy_w = 1 * (df3$Race1 == "White")
dummy_o = 1 * (df3$Race1 == "Other")

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

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

```
##
## Call:
## lm(formula = BMI ~ SleepHrsNight + Age + Gender + factor(Race1),
##     data = df3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -14.347  -4.497  -1.201   3.190  40.277
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    30.78080    0.97780  31.480  < 2e-16 ***
## SleepHrsNight  -0.29383    0.11031  -2.664 0.007785 **
## Age             0.05055    0.01282   3.944 8.26e-05 ***
## Gender          0.25869    0.28895   0.895 0.370740
## factor(Race1)2 -2.28054    0.67704  -3.368 0.000769 ***
## factor(Race1)3 -1.02309    0.59140  -1.730 0.083782 .
## factor(Race1)4 -2.51942    0.43385  -5.807 7.30e-09 ***
## factor(Race1)5 -4.14341    0.66274  -6.252 4.88e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.643 on 2144 degrees of freedom
## Multiple R-squared:  0.03564,    Adjusted R-squared:  0.03249
## F-statistic: 11.32 on 7 and 2144 DF,  p-value: 3.698e-14
```

```
## model_2 add known risk factors
m_2 = lm(
  BMI ~ SleepHrsNight + Age + Gender + Race1 + TotChol + BPDiaAve + BPSysAve + AlcoholYear + Smoke100 +
    DaysPhysHlthBad + PhysActive,
  df3
)
summary(m_2)
```

```
##
## Call:
## lm(formula = BMI ~ SleepHrsNight + Age + Gender + Race1 + TotChol +
##     BPDiaAve + BPSysAve + AlcoholYear + Smoke100 + DaysPhysHlthBad +
##     PhysActive, data = df3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.752  -4.236  -0.849   3.055  37.857
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     21.023150   1.610401  13.055  < 2e-16 ***
## SleepHrsNight   -0.212193   0.107400  -1.976 0.048314 *
## Age              0.012839   0.013495   0.951 0.341528
## Gender           0.514621   0.291331   1.766 0.077463 .
## Race1           -0.622971   0.122615  -5.081 4.09e-07 ***
## TotChol          0.076572   0.139325   0.550 0.582658
## BPDiaAve         0.054500   0.014049   3.879 0.000108 ***
## BPSysAve         0.066004   0.012027   5.488 4.55e-08 ***
## AlcoholYear     -0.009762   0.001533  -6.368 2.34e-10 ***
## Smoke100Yes     -0.507830   0.287921  -1.764 0.077911 .
## DaysPhysHlthBad  0.066309   0.019785   3.352 0.000818 ***
## PhysActiveYes   -1.260928   0.292769  -4.307 1.73e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.413 on 2140 degrees of freedom
```

```
## Multiple R-squared:  0.1029, Adjusted R-squared:  0.09826
## F-statistic: 22.31 on 11 and 2140 DF,  p-value: < 2.2e-16
```

```
#LINE

#influential observations

#multicollinearity

vif(m_1)
```

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

```
vif(m_2)
```

```
##   SleepHrsNight             Age          Gender           Race1         TotChol
##        1.035419        1.223319        1.106167        1.045711        1.122357
##        BPDiaAve        BPSysAve     AlcoholYear        Smoke100 DaysPhysHlthBad
##        1.447702        1.542999        1.091195        1.078534        1.057582
##      PhysActive
##        1.093222
```

```
## model_3 add additional risk factors

m_3 = lm(
  BMI ~ SleepHrsNight + Age + Gender + Race1  + Poverty + TotChol + BPDiaAve + BPSysAve + AlcoholYear +
    DaysPhysHlthBad + HealthGen + PhysActive,
  df3
)
summary(m_3)
```

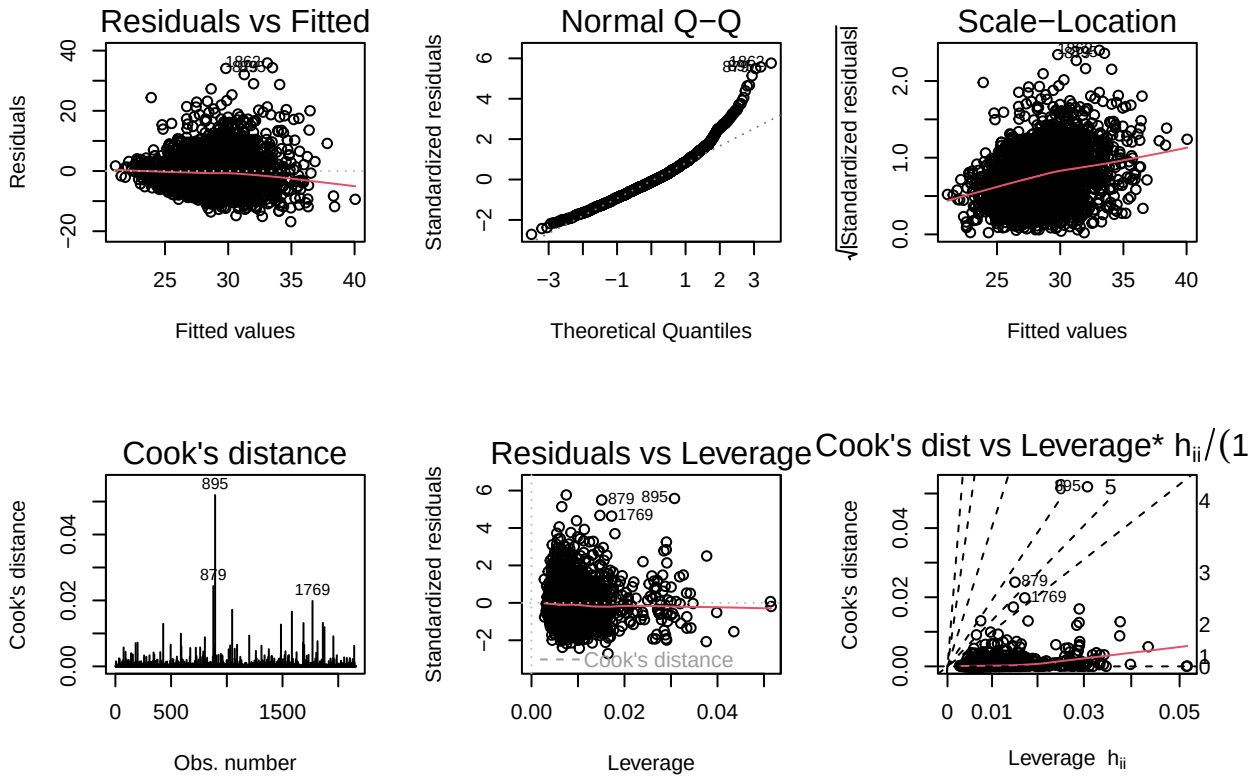
```
##
## Call:
## lm(formula = BMI ~ SleepHrsNight + Age + Gender + Race1 + Poverty +
##     TotChol + BPDiaAve + BPSysAve + AlcoholYear + Smoke100 +
##     UrineFlow1 + DaysMentHlthBad + DaysPhysHlthBad + HealthGen +
##     PhysActive, data = df3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.838  -4.054  -0.646   3.203  35.902
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     18.471020   1.621565  11.391  < 2e-16 ***
## SleepHrsNight   -0.121393   0.106352  -1.141  0.25382
## Age              0.010806   0.013725   0.787  0.43118
## Gender           0.532917   0.286537   1.860  0.06304 .
## Race1           -0.500763   0.122151  -4.100 4.29e-05 ***
## Poverty          0.073370   0.090958   0.807  0.41997
## TotChol          0.030653   0.136000   0.225  0.82170
## BPDiaAve         0.058458   0.013721   4.260 2.13e-05 ***
```

```
## BPSysAve         0.053724   0.011806   4.550 5.65e-06 ***
## AlcoholYear     -0.008337   0.001515  -5.503 4.18e-08 ***
## Smoke100Yes     -0.807332   0.287264  -2.810  0.00499 **
## UrineFlow1      -0.113369   0.142545  -0.795  0.42652
## DaysMentHlthBad -0.030360   0.017984  -1.688  0.09153 .
## DaysPhysHlthBad  0.014779   0.020974   0.705  0.48112
## HealthGenVgood   1.922013   0.470923   4.081 4.64e-05 ***
## HealthGenGood    3.569501   0.468730   7.615 3.93e-14 ***
## HealthGenFair    5.283476   0.575334   9.183  < 2e-16 ***
## HealthGenPoor    7.546146   1.078147   6.999 3.43e-12 ***
## PhysActiveYes   -0.818408   0.294015  -2.784  0.00542 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.251 on 2133 degrees of freedom
## Multiple R-squared:  0.1504, Adjusted R-squared:  0.1432
## F-statistic: 20.97 on 18 and 2133 DF,  p-value: < 2.2e-16
```

```
vif(m_3)
```

```
##                     GVIF Df GVIF^(1/(2*Df))
## SleepHrsNight   1.068552  1        1.033708
## Age             1.331598  1        1.153949
## Gender          1.126176  1        1.061214
## Race1           1.092236  1        1.045101
## Poverty         1.302699  1        1.141358
## TotChol         1.125511  1        1.060901
## BPDiaAve        1.453387  1        1.205565
## BPSysAve        1.564805  1        1.250922
## AlcoholYear     1.121584  1        1.059049
## Smoke100        1.129923  1        1.062979
## UrineFlow1      1.044330  1        1.021925
## DaysMentHlthBad 1.145584  1        1.070320
## DaysPhysHlthBad 1.250957  1        1.118462
## HealthGen       1.435741  4        1.046248
## PhysActive      1.160363  1        1.077202
```

```
par(mfrow = c(2, 3)) #read more from ?plot.lm
plot(m_3, which = 1)
plot(m_3, which = 2)
plot(m_3, which = 3)
plot(m_3, which = 4)
plot(m_3, which = 5)
plot(m_3, which = 6)
```




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

# model_4 add additional risk factors
m_full = lm(
  BMI ~ SleepHrsNight + Age + Gender + Race1  + Poverty + TotChol + BPDiaAve + BPSysAve + AlcoholYear +
    DaysPhysHlthBad + HealthGen + PhysActive + SleepHrsNight * Age + SleepHrsNight *
    Gender,
  df3
)
summary(m_full)
```

```
##
## Call:
## lm(formula = BMI ~ SleepHrsNight + Age + Gender + Race1 + Poverty +
##     TotChol + BPDiaAve + BPSysAve + AlcoholYear + Smoke100 +
##     UrineFlow1 + DaysMentHlthBad + DaysPhysHlthBad + HealthGen +
##     PhysActive + SleepHrsNight * Age + SleepHrsNight * Gender,
##     data = df3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.019  -4.059  -0.648   3.165  36.301
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          21.349305   2.996916   7.124 1.43e-12 ***
```

```
## SleepHrsNight        -0.542616   0.378594  -1.433  0.15194
## Age                  -0.105136   0.062783  -1.675  0.09416 .
## Gender                3.768696   1.435161   2.626  0.00870 **
## Race1                -0.503222   0.121964  -4.126 3.83e-05 ***
## Poverty               0.072729   0.090968   0.800  0.42409
## TotChol               0.014773   0.135905   0.109  0.91345
## BPDiaAve              0.058709   0.013701   4.285 1.91e-05 ***
## BPSysAve              0.054450   0.011792   4.617 4.12e-06 ***
## AlcoholYear          -0.008396   0.001513  -5.549 3.23e-08 ***
## Smoke100Yes          -0.802999   0.286852  -2.799  0.00517 **
## UrineFlow1           -0.102218   0.142435  -0.718  0.47305
## DaysMentHlthBad      -0.030250   0.017962  -1.684  0.09230 .
## DaysPhysHlthBad       0.015142   0.020943   0.723  0.46975
## HealthGenVgood        1.928283   0.470249   4.101 4.28e-05 ***
## HealthGenGood         3.559316   0.468010   7.605 4.24e-14 ***
## HealthGenFair         5.299570   0.575060   9.216  < 2e-16 ***
## HealthGenPoor         7.640142   1.077494   7.091 1.81e-12 ***
## PhysActiveYes        -0.837418   0.294615  -2.842  0.00452 **
## SleepHrsNight:Age     0.017092   0.009024   1.894  0.05837 .
## SleepHrsNight:Gender -0.477032   0.206903  -2.306  0.02123 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.242 on 2131 degrees of freedom
## Multiple R-squared:  0.1538, Adjusted R-squared:  0.1459
## F-statistic: 19.37 on 20 and 2131 DF,  p-value: < 2.2e-16
```

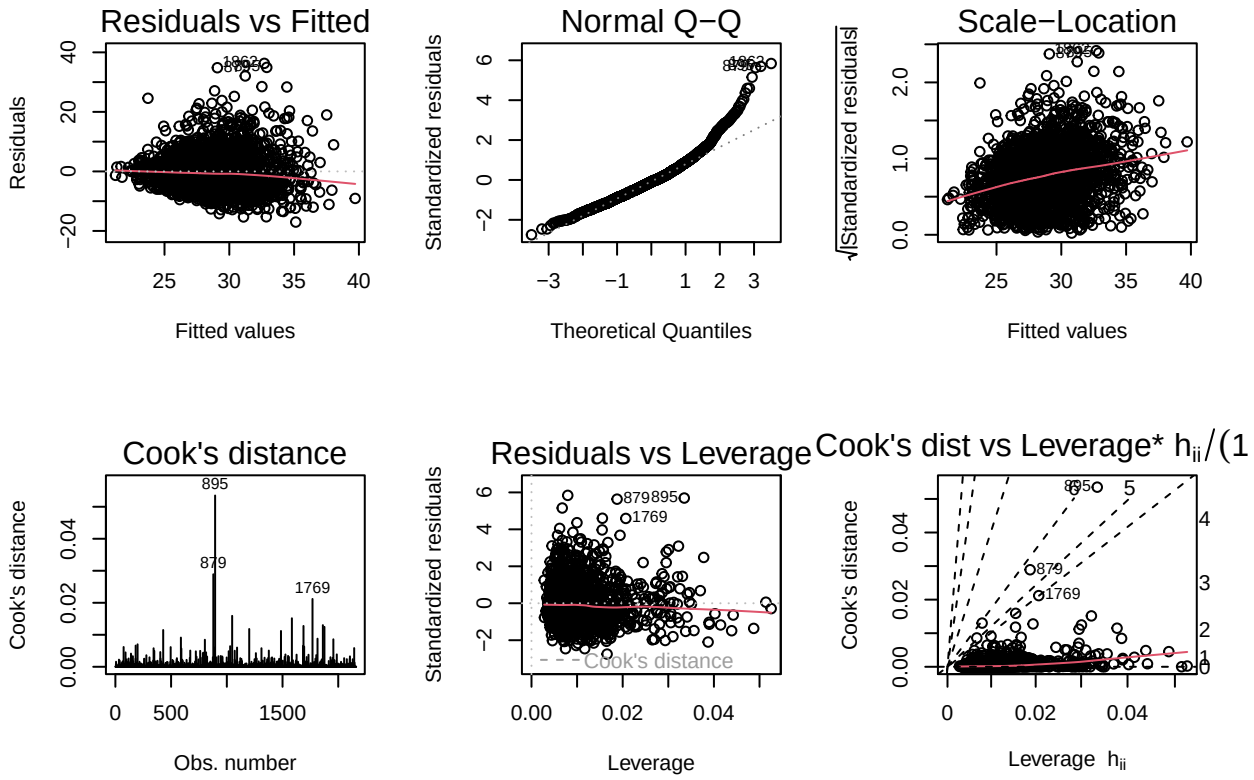
```
vif(m_full)
```

```
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif
```

```
##                          GVIF Df GVIF^(1/(2*Df))
## SleepHrsNight       13.583457  1        3.685574
## Age                 27.952318  1        5.286995
## Gender              28.340488  1        5.323579
## Race1                1.092306  1        1.045135
## Poverty              1.307058  1        1.143266
## TotChol              1.127465  1        1.061822
## BPDiaAve             1.453678  1        1.205686
## BPSysAve             1.565996  1        1.251398
## AlcoholYear          1.122185  1        1.059332
## Smoke100             1.130221  1        1.063119
## UrineFlow1           1.045986  1        1.022734
## DaysMentHlthBad      1.146283  1        1.070646
## DaysPhysHlthBad      1.251072  1        1.118513
## HealthGen            1.447335  4        1.047300
## PhysActive           1.168763  1        1.081093
## SleepHrsNight:Age   37.541993  1        6.127152
## SleepHrsNight:Gender 29.940850  1        5.471823
```

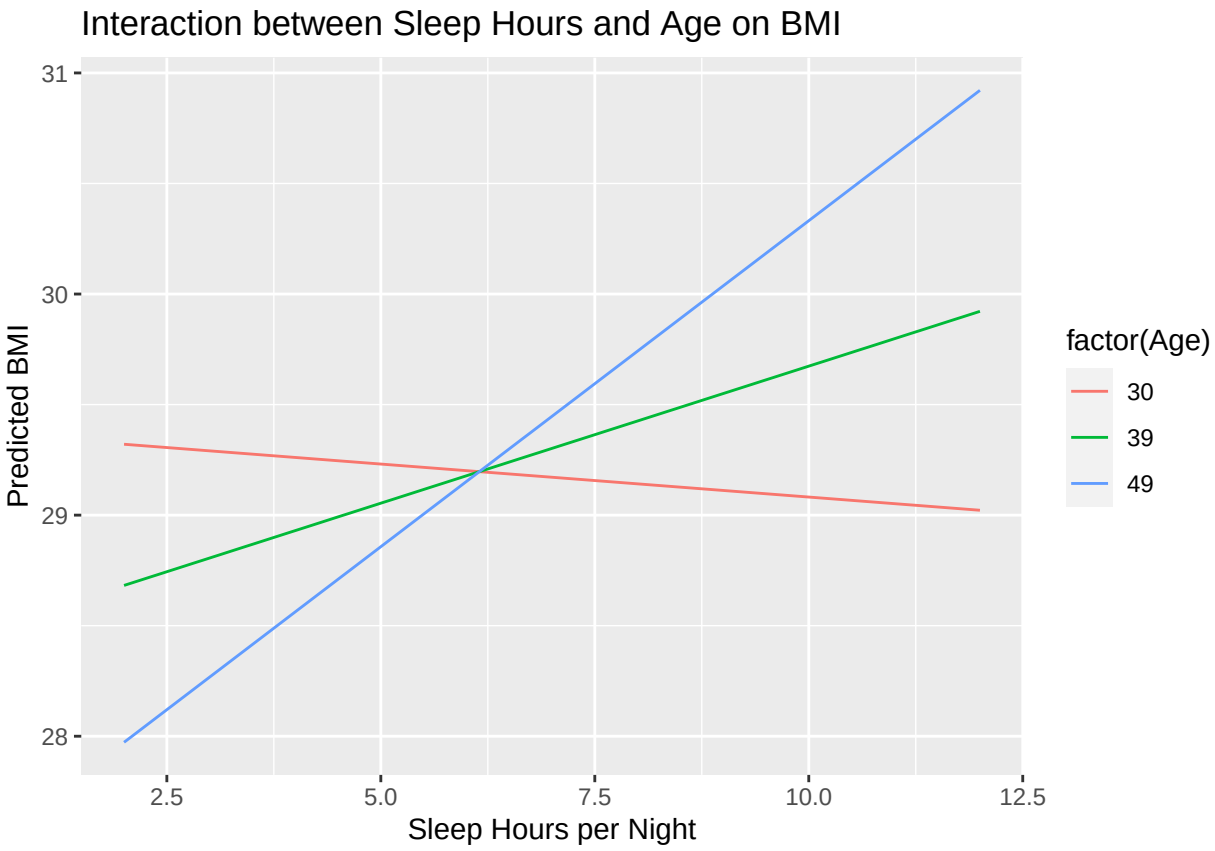
```
par(mfrow = c(2, 3)) #read more from ?plot.lm
plot(m_full, which = 1)
plot(m_full, which = 2)
plot(m_full, which = 3)
plot(m_full, which = 4)
```

```
plot(m_full, which = 5)
plot(m_full, which = 6)
```




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

getMode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}

new_data <- expand.grid(SleepHrsNight = seq(min(df3$SleepHrsNight), max(df3$SleepHrsNight), length.out =
                        Age = quantile(df3$Age, probs = c(0.25, 0.5, 0.75)),
                        Gender = median(df3$Gender, na.rm = TRUE),
                        Race1 = median(df3$Race1, na.rm = TRUE),
                        Poverty = median(df3$Poverty, na.rm = TRUE),
                        TotChol = median(df3$TotChol, na.rm = TRUE),
                        BPDiaAve = median(df3$BPDiaAve, na.rm = TRUE),
                        BPSysAve = median(df3$BPSysAve, na.rm = TRUE),
                        AlcoholYear = median(df3$AlcoholYear, na.rm = TRUE),
                        Smoke100 = getMode(df3$Smoke100),
                        UrineFlow1 = median(df3$UrineFlow1, na.rm = TRUE),
                        DaysMentHlthBad = median(df3$DaysMentHlthBad, na.rm = TRUE),
                        DaysPhysHlthBad = median(df3$DaysPhysHlthBad, na.rm = TRUE),
                        HealthGen = getMode(df3$HealthGen),
                        PhysActive = getMode(df3$PhysActive)
)
```

```r
# predict
new_data$predicted_BMI <- predict(m_full, newdata = new_data)
# interaction
library(ggplot2)
ggplot(new_data, aes(x = SleepHrsNight, y = predicted_BMI, group = factor(Age))) +
  geom_line(aes(color = factor(Age))) +
  labs(title = "Interaction between Sleep Hours and Age on BMI",
       x = "Sleep Hours per Night",
       y = "Predicted BMI")
```



(4) Diagnosis: 10-fold CV

```r
library(caret)
```

```
## Loading required package: lattice
```

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

```r
cv_model <-
  train(
    SleepHrsNight ~ .,
    data = df3,
    method = "lm",
    trControl = control
  )
cv_model
```

```
## Linear Regression
##
## 2152 samples
##   21 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 1937, 1938, 1936, 1937, 1937, 1937, ...
## Resampling results:
##
##   RMSE      Rsquared    MAE
##   1.280209  0.05043061  0.9931499
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

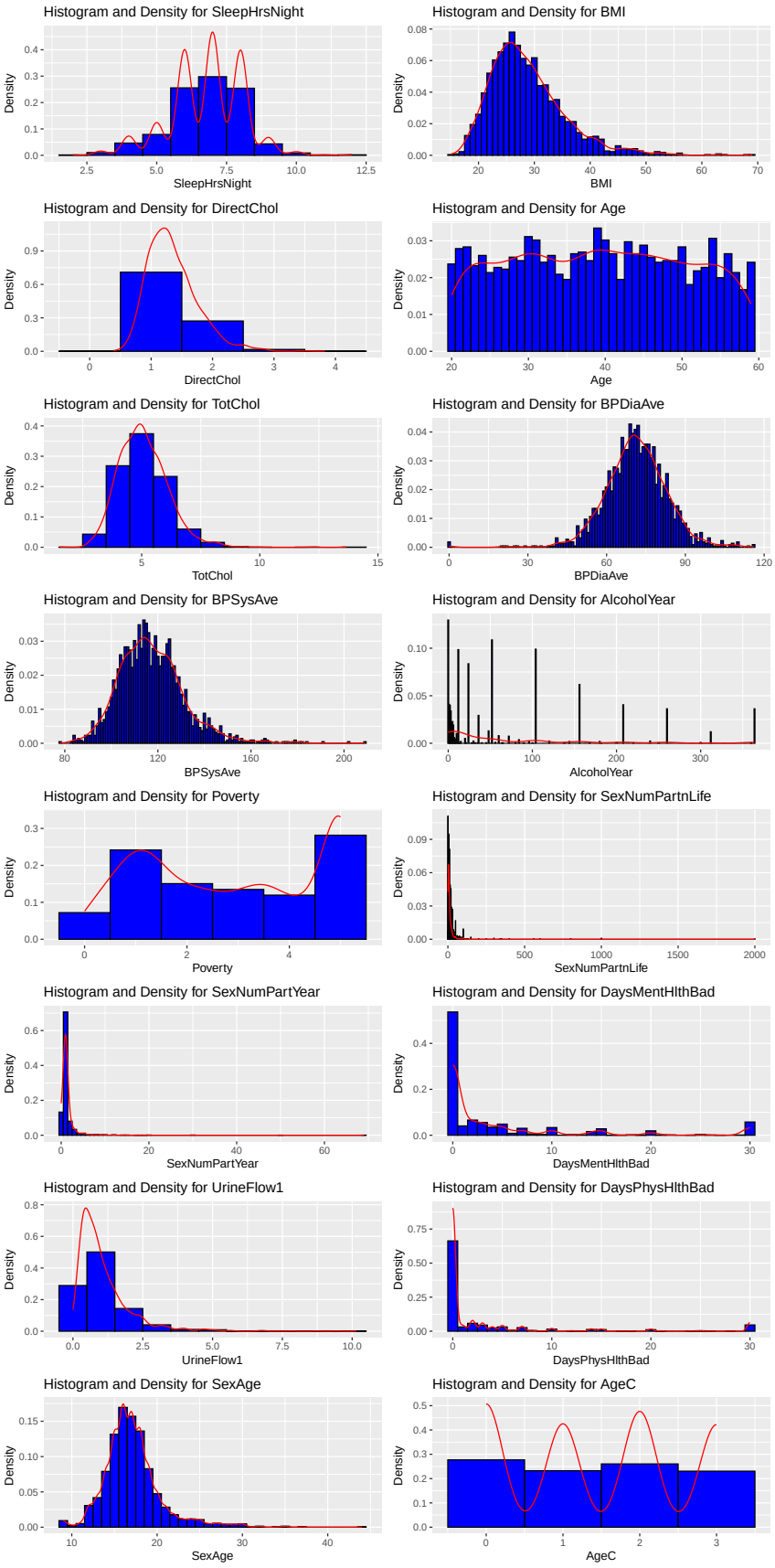
```r
(cv_results <- cv_model$results)
```

```
##   intercept     RMSE   Rsquared       MAE     RMSESD RsquaredSD      MAESD
## 1      TRUE 1.280209 0.05043061 0.9931499 0.04543809 0.02732622 0.02794626
```

(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 SexNumPartnLife
Density
SexNumPartnLife
Histogram and Density for SexNumPartYear
Density
SexNumPartYear
Histogram and Density for DaysMentHlthBad
Density
DaysMentHlthBad
Histogram and Density for UrineFlow1
Density
UrineFlow1
Histogram and Density for DaysPhysHlthBad
Density
DaysPhysHlthBad
Histogram and Density for SexAge
Density
SexAge
Histogram and Density for AgeC
Density
AgeC

```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 0.9347691 1.022342e-29 1.840215e-28
## BMI                         BMI 0.9263898 2.950926e-31 5.311666e-30
## DirectChol           DirectChol 0.9439221 7.552977e-28 1.359536e-26
## Age                         Age 0.9579654 1.832383e-24 3.298290e-23
## Gender                   Gender 0.6352876 1.636740e-55 2.946133e-54
## Race1                     Race1 0.7327797 3.104346e-50 5.587823e-49
## TotChol                 TotChol 0.9642744 1.175111e-22 2.115200e-21
## BPDiaAve               BPDiaAve 0.9718079 3.709893e-20 6.677808e-19
## BPSysAve               BPSysAve 0.9554033 3.865527e-25 6.957949e-24
## AlcoholYear         AlcoholYear 0.7454040 1.944127e-49 3.499428e-48
## Poverty                 Poverty 0.8942742 4.092136e-36 7.365845e-35
## SexNumPartnLife SexNumPartnLife 0.1496531 2.951432e-71 5.312577e-70
## SexNumPartYear   SexNumPartYear 0.2562318 1.244353e-68 2.239836e-67
## DaysMentHlthBad DaysMentHlthBad 0.6112779 1.254550e-56 2.258190e-55
## UrineFlow1           UrineFlow1 0.7555438 8.969094e-49 1.614437e-47
## PhysActive           PhysActive        NA           NA           NA
## DaysPhysHlthBad DaysPhysHlthBad 0.4968273 2.926552e-61 5.267794e-60
## Smoke100               Smoke100        NA           NA           NA
## Depressed             Depressed        NA           NA           NA
## HealthGen             HealthGen        NA           NA           NA
## SexAge                   SexAge 0.8954434 5.842918e-36 1.051725e-34
## AgeC                       AgeC 0.8533480 8.034125e-41 1.446143e-39
```

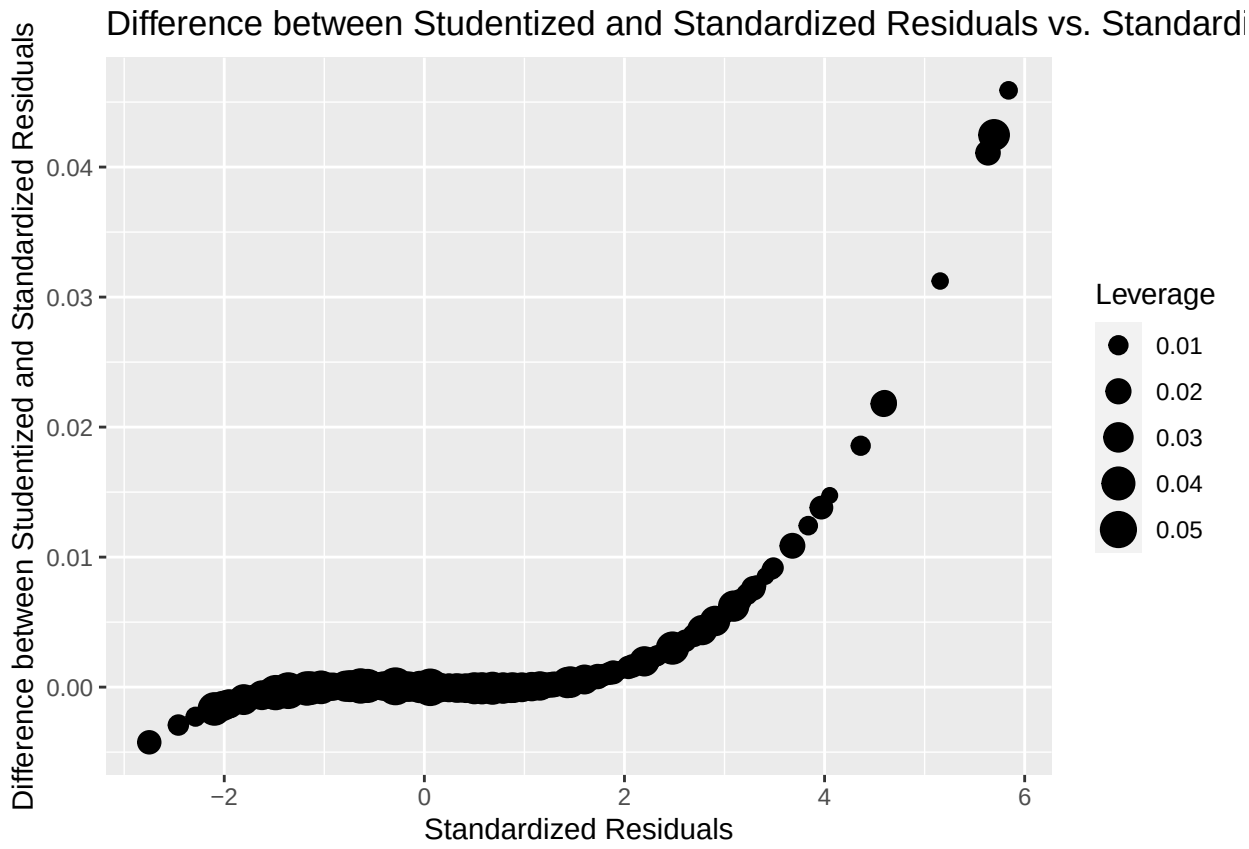
Standardized residuals, Studentized residuals

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

```r
# Standardized residuals
residual_2 <- rstandard(m_full)

# Studentized residuals
residual_3 <- rstudent(m_full)

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

# 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 -1.448790e-16 6.212489 -17.019074 36.300973
## 2           Standardized -2.232345e-05 1.000939  -2.749618  5.839262
## 3            Studentized  2.310529e-04 1.002091  -2.753862  5.885164
## 4 Externally Studentized  2.310529e-04 1.002091  -2.753862  5.885164
```

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

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

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

# Calculate leverage values
leverage_values <- hatvalues(m_full)

# 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")
```

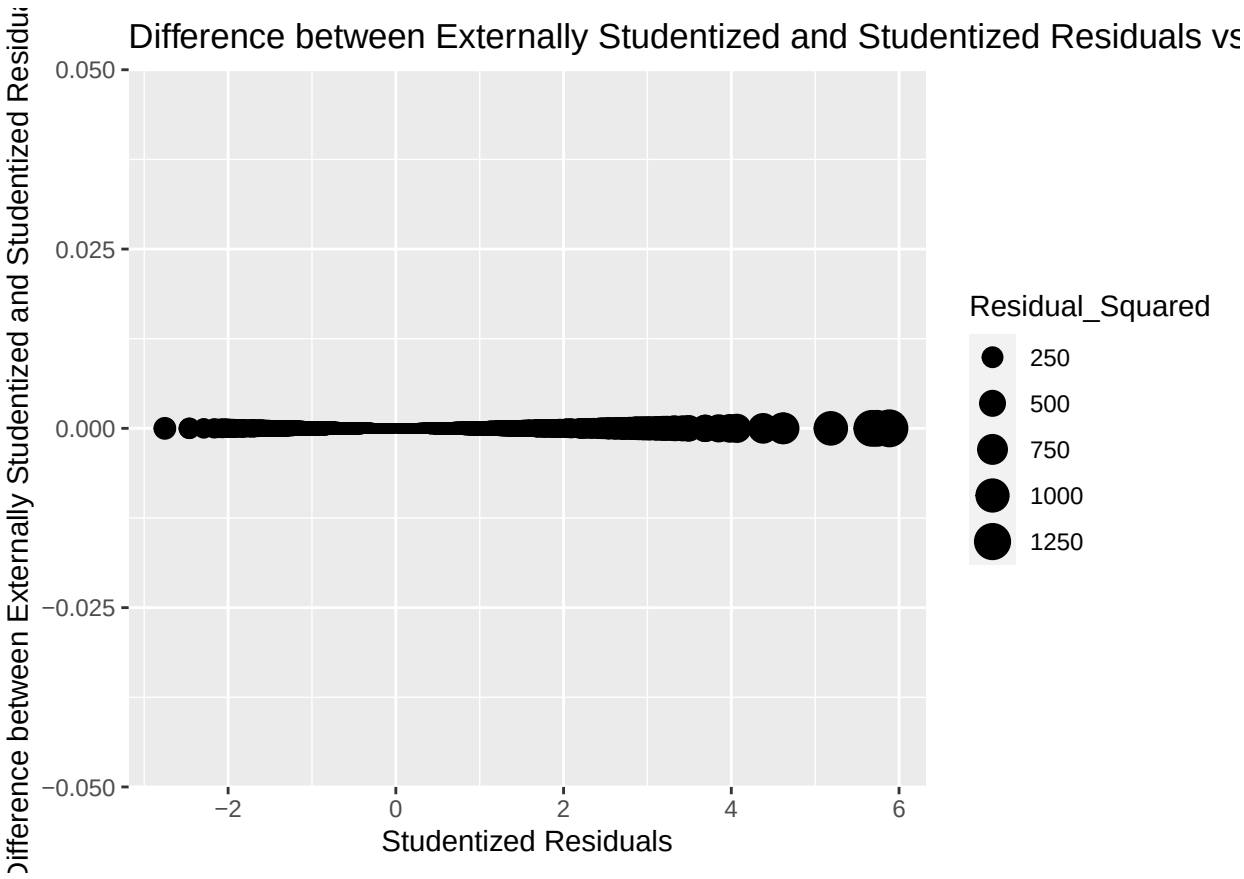



```
# Display the plot
print(ggplot)
```

```
## function (data = NULL, mapping = aes(), ..., environment = parent.frame())
## {
##     UseMethod("ggplot")
## }
## <bytecode: 0x39eced8>
## <environment: namespace:ggplot2>
```

```
# Load necessary library
library(ggplot2)

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

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

# Regular residuals
residual_1 <- m_full$residuals

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

```
  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")
```

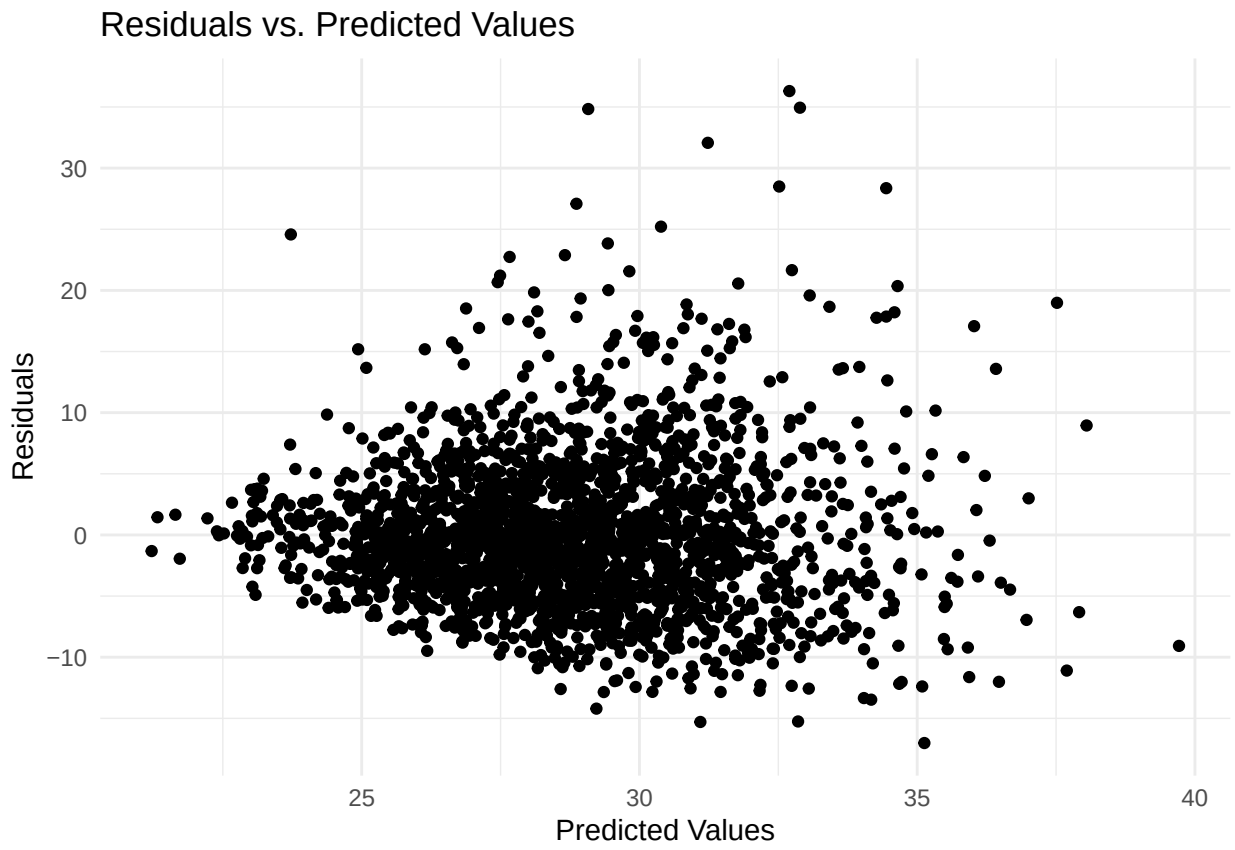



```
# Display the plot
print(ggplot)
```

```
## function (data = NULL, mapping = aes(), ..., environment = parent.frame())
## {
##     UseMethod("ggplot")
## }
## <bytecode: 0x39eced8>
## <environment: namespace:ggplot2>
```

```
# Load necessary library
library(ggplot2)

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

# Calculate regular residuals
```

```r
residual_1 <- m_full$residuals

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

# 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()
```

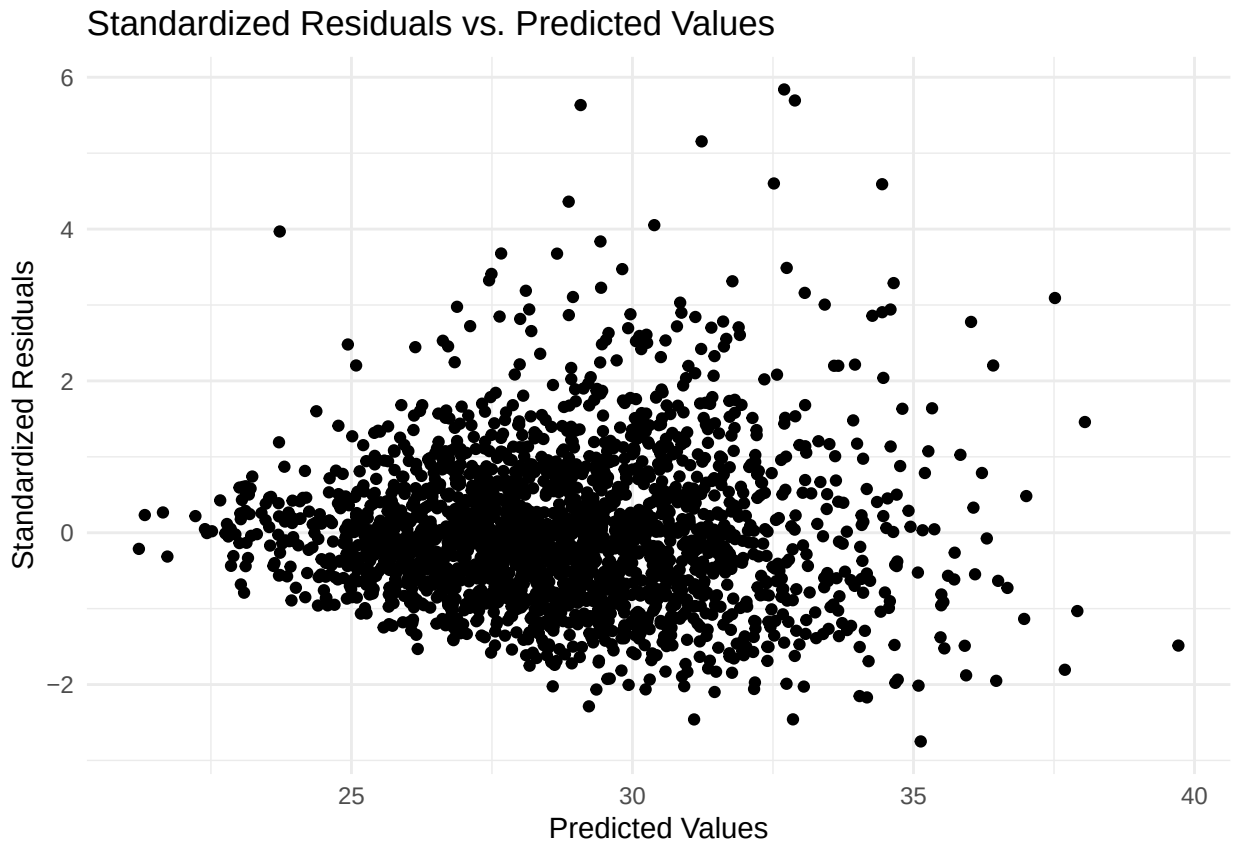



```r
# Display the plot
print(ggplot)
```

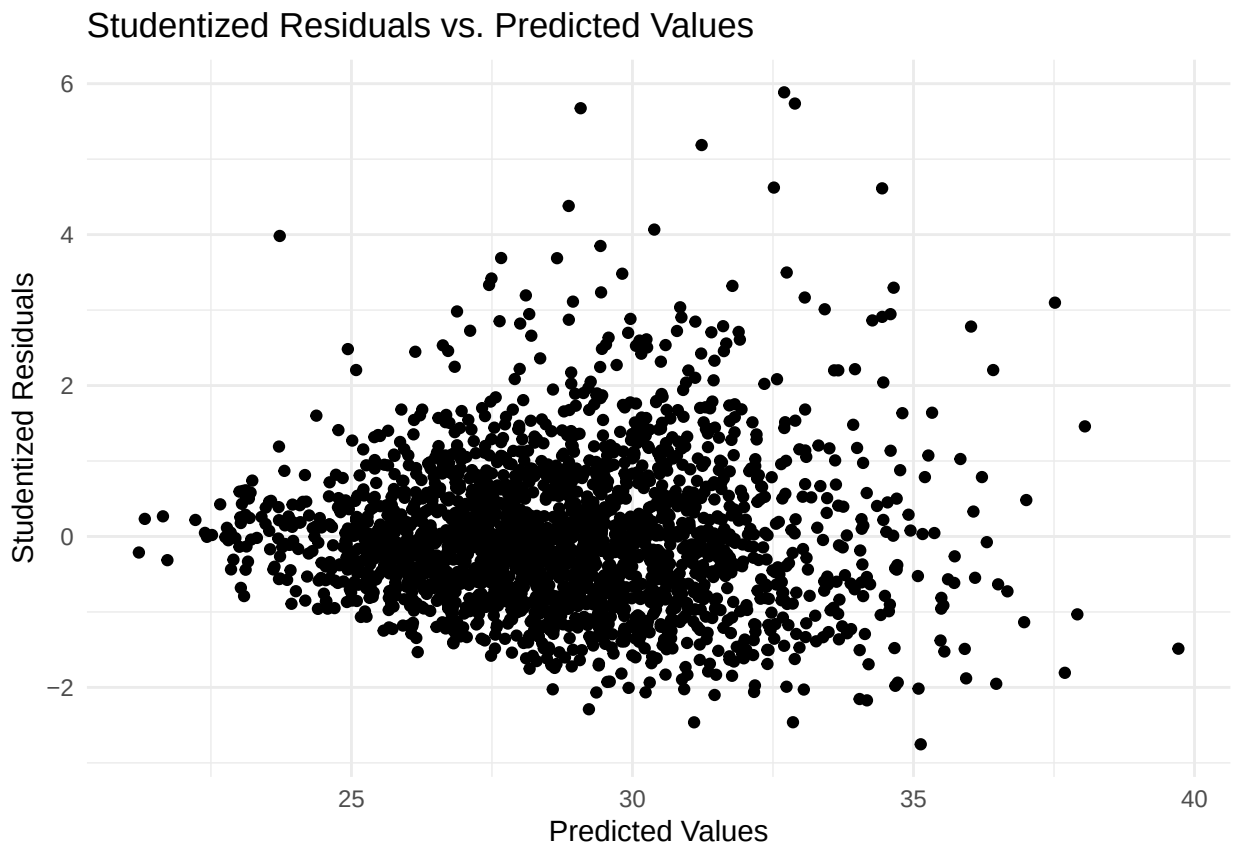
```
## function (data = NULL, mapping = aes(), ..., environment = parent.frame())
## {
##     UseMethod("ggplot")
## }
## <bytecode: 0x39eced8>
## <environment: namespace:ggplot2>
```

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

# Assuming m_full is your linear model
```

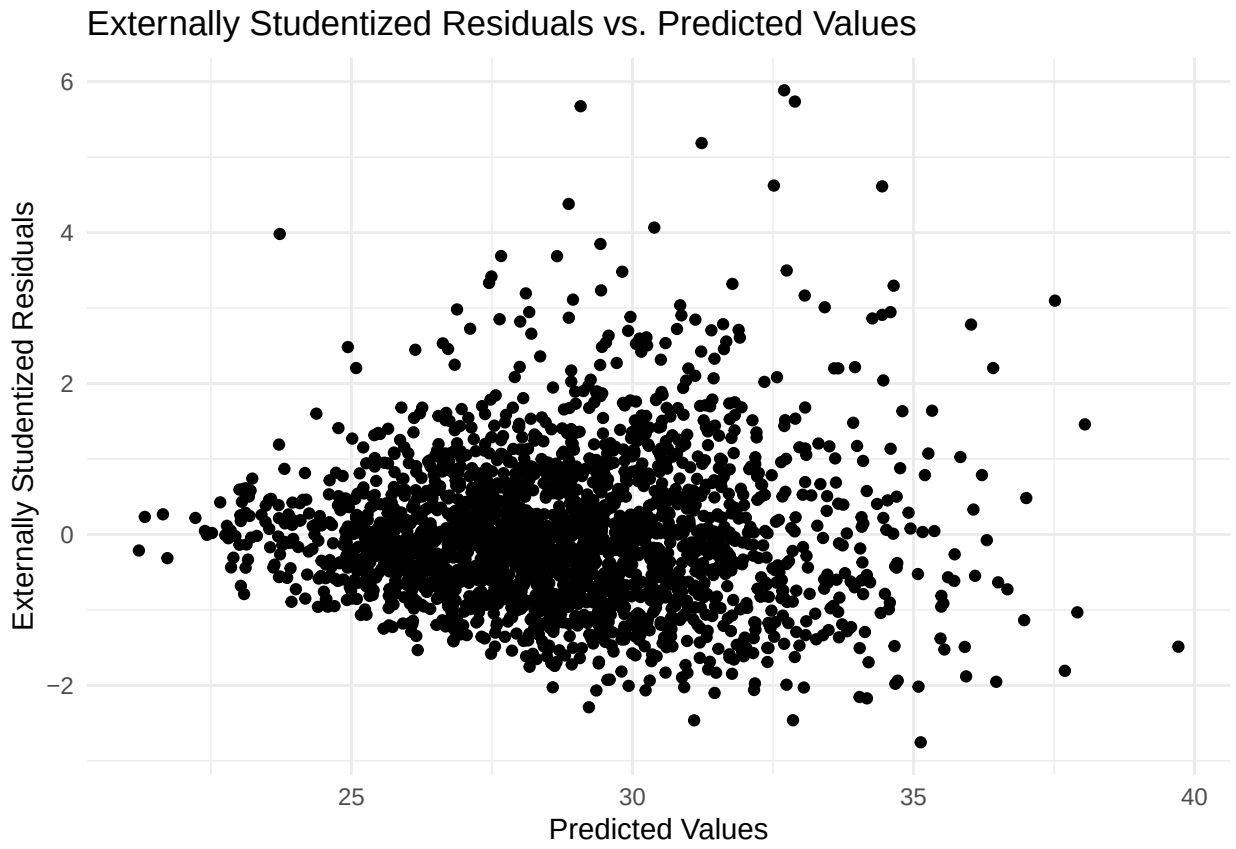
```r
# m_full <- lm(SleepMinNight ~ ., data = df3)

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

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

# 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()
```




```
# 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(m_full)
```

```
## Start:  AIC=7902.52
## BMI ~ SleepHrsNight + Age + Gender + Race1 + Poverty + TotChol +
##     BPDiaAve + BPSysAve + AlcoholYear + Smoke100 + UrineFlow1 +
##     DaysMentHlthBad + DaysPhysHlthBad + HealthGen + PhysActive +
##     SleepHrsNight * Age + SleepHrsNight * Gender
##
##                         Df Sum of Sq   RSS    AIC
## - TotChol                1       0.5 83018 7900.5
## - UrineFlow1             1      20.1 83038 7901.0
## - DaysPhysHlthBad        1      20.4 83038 7901.0
## - Poverty                1      24.9 83043 7901.2
## <none>                                83018 7902.5
## - DaysMentHlthBad        1     110.5 83128 7903.4
## - SleepHrsNight:Age      1     139.7 83158 7904.1
## - SleepHrsNight:Gender   1     207.1 83225 7905.9
## - Smoke100               1     305.3 83323 7908.4
## - PhysActive             1     314.7 83333 7908.7
## - Race1                  1     663.2 83681 7917.6
## - BPDiaAve               1     715.3 83733 7919.0
## - BPSysAve               1     830.6 83848 7921.9
```

```
## - AlcoholYear           1    1199.6 84217 7931.4
## - HealthGen             4    4547.8 87566 8009.3
##
## Step:  AIC=7900.53
## BMI ~ SleepHrsNight + Age + Gender + Race1 + Poverty + BPDiaAve +
##     BPSysAve + AlcoholYear + Smoke100 + UrineFlow1 + DaysMentHlthBad +
##     DaysPhysHlthBad + HealthGen + PhysActive + SleepHrsNight:Age +
##     SleepHrsNight:Gender
##
##                         Df Sum of Sq   RSS    AIC
## - UrineFlow1             1      20.0 83038 7899.1
## - DaysPhysHlthBad        1      20.3 83039 7899.1
## - Poverty                1      24.8 83043 7899.2
## <none>                               83018 7900.5
## - DaysMentHlthBad        1     110.9 83129 7901.4
## - SleepHrsNight:Age      1     140.5 83159 7902.2
## - SleepHrsNight:Gender   1     207.6 83226 7903.9
## - Smoke100               1     306.6 83325 7906.5
## - PhysActive             1     315.3 83334 7906.7
## - Race1                  1     662.9 83681 7915.6
## - BPDiaAve               1     725.3 83744 7917.3
## - BPSysAve               1     832.7 83851 7920.0
## - AlcoholYear            1    1200.1 84218 7929.4
## - HealthGen              4    4554.4 87573 8007.5
##
## Step:  AIC=7899.05
## BMI ~ SleepHrsNight + Age + Gender + Race1 + Poverty + BPDiaAve +
##     BPSysAve + AlcoholYear + Smoke100 + DaysMentHlthBad + DaysPhysHlthBad +
##     HealthGen + PhysActive + SleepHrsNight:Age + SleepHrsNight:Gender
##
##                         Df Sum of Sq   RSS    AIC
## - DaysPhysHlthBad        1      20.2 83058 7897.6
## - Poverty                1      21.7 83060 7897.6
## <none>                               83038 7899.1
## - DaysMentHlthBad        1     113.2 83152 7900.0
## - SleepHrsNight:Age      1     144.9 83183 7900.8
## - SleepHrsNight:Gender   1     207.9 83246 7902.4
## - Smoke100               1     307.7 83346 7905.0
## - PhysActive             1     325.0 83363 7905.5
## - Race1                  1     690.4 83729 7914.9
## - BPDiaAve               1     728.0 83766 7915.8
## - BPSysAve               1     828.9 83867 7918.4
## - AlcoholYear            1    1223.1 84261 7928.5
## - HealthGen              4    4580.6 87619 8006.6
##
## Step:  AIC=7897.57
## BMI ~ SleepHrsNight + Age + Gender + Race1 + Poverty + BPDiaAve +
##     BPSysAve + AlcoholYear + Smoke100 + DaysMentHlthBad + HealthGen +
##     PhysActive + SleepHrsNight:Age + SleepHrsNight:Gender
##
##                         Df Sum of Sq   RSS    AIC
## - Poverty                1      21.3 83080 7896.1
## <none>                               83058 7897.6
## - DaysMentHlthBad        1     100.0 83158 7898.2
```

```
## - SleepHrsNight:Age     1     143.8 83202 7899.3
## - SleepHrsNight:Gender  1     207.9 83266 7901.0
## - Smoke100              1     301.9 83360 7903.4
## - PhysActive            1     334.8 83393 7904.2
## - Race1                 1     688.1 83747 7913.3
## - BPDiaAve              1     719.1 83778 7914.1
## - BPSysAve              1     829.9 83888 7917.0
## - AlcoholYear           1    1235.8 84294 7927.4
## - HealthGen             4    5008.6 88067 8015.6
##
## Step:  AIC=7896.12
## BMI ~ SleepHrsNight + Age + Gender + Race1 + BPDiaAve + BPSysAve +
##     AlcoholYear + Smoke100 + DaysMentHlthBad + HealthGen + PhysActive +
##     SleepHrsNight:Age + SleepHrsNight:Gender
##
##                         Df Sum of Sq   RSS    AIC
## <none>                               83080 7896.1
## - DaysMentHlthBad       1     105.0 83185 7896.8
## - SleepHrsNight:Age     1     148.8 83229 7898.0
## - SleepHrsNight:Gender  1     203.5 83283 7899.4
## - PhysActive            1     317.7 83397 7902.3
## - Smoke100              1     338.7 83418 7902.9
## - Race1                 1     668.7 83748 7911.4
## - BPDiaAve              1     726.3 83806 7912.9
## - BPSysAve              1     818.4 83898 7915.2
## - AlcoholYear           1    1214.9 84295 7925.4
## - HealthGen             4    5098.7 88178 8016.3
##
## Call:
## lm(formula = BMI ~ SleepHrsNight + Age + Gender + Race1 + BPDiaAve +
##     BPSysAve + AlcoholYear + Smoke100 + DaysMentHlthBad + HealthGen +
##     PhysActive + SleepHrsNight:Age + SleepHrsNight:Gender, data = df3)
##
## Coefficients:
##          (Intercept)         SleepHrsNight                   Age
##            21.620969             -0.565231             -0.104961
##               Gender                 Race1              BPDiaAve
##             3.749303             -0.498750              0.058815
##             BPSysAve           AlcoholYear           Smoke100Yes
##             0.053941             -0.008364             -0.832608
##      DaysMentHlthBad        HealthGenVgood         HealthGenGood
##            -0.028955              1.913856              3.548945
##        HealthGenFair         HealthGenPoor         PhysActiveYes
##             5.292807              7.773117             -0.830109
##    SleepHrsNight:Age  SleepHrsNight:Gender
##             0.017598             -0.472517
```

```
library(olsrr)
```

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

```
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    HealthGen                 0.0826      0.0809    162.3618    14153.5414    6.4747
##    2    BPDiaAve                  0.1066      0.1045    103.8740    14098.4490    6.3908
##    3    AlcoholYear               0.1226      0.1201     65.6513    14061.6280    6.3349
##    4    Race1                     0.1323      0.1295     43.2247    14039.7062    6.3013
##    5    BPSysAve                  0.1401      0.1369     25.5844    14022.2771    6.2744
##    6    Smoke100                  0.1430      0.1394     20.2357    14016.9620    6.2652
##    7    PhysActive                0.1463      0.1424     13.8134    14010.5477    6.2544
##    8    Gender                    0.1476      0.1432     12.6813    14009.4102    6.2513
##    9    SleepHrsNight:Gender      0.1497      0.1449      9.4392    14006.1486    6.2451
##   10    DaysMentHlthBad           0.1509      0.1458      8.3036    14004.9953    6.2420
##   11    Poverty                   0.1514      0.1458      9.1090    14005.7926    6.2417
##   12    DaysPhysHlthBad           0.1516      0.1457     10.4990    14007.1783    6.2423
##   13    TotChol                   0.1517      0.1453     12.3665    14009.0448    6.2436
##   14    UrineFlow1                0.1519      0.1452     13.7880    14010.4620    6.2442
##   15    Age                       0.1522      0.1450     15.1916    14011.8610    6.2448
##   16    SleepHrsNight             0.1524      0.1448     16.5871    14013.2516    6.2454
##   17    SleepHrsNight:Age         0.1538      0.1459     15.0000    14011.6323    6.2416
## ------------------------------------------------------------------------------------------
```

```
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    HealthGen         0.0826      0.0809    162.3618    14153.5414    6.4747
##    2    BPDiaAve          0.1066      0.1045    103.8740    14098.4490    6.3908
##    3    AlcoholYear       0.1226      0.1201     65.6513    14061.6280    6.3349
##    4    Race1             0.1323      0.1295     43.2247    14039.7062    6.3013
##    5    BPSysAve          0.1401      0.1369     25.5844    14022.2771    6.2744
##    6    Smoke100          0.1430      0.1394     20.2357    14016.9620    6.2652
##    7    PhysActive        0.1463      0.1424     13.8134    14010.5477    6.2544
## ------------------------------------------------------------------------------
```

```
ols_mallows_cp(model = m_3, fullmodel = m_full)  # Mallows' Cp
```

```
## [1] 19.69163
```