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 469621 25.1    1011344 54.1   660860 35.3
## Vcells 878082  6.7    8388608 64.0  1800812 13.8
```

```
set.seed(123)
############### (1) Data cleaning #########################################
## select variables
library(NHANES)
df0 <- NHANES
df <- NHANES[NHANES$Age >= 18 & NHANES$Age < 60,]
# colSums(is.na(df)) / nrow(df)
df <- df[, which(colSums(is.na(df)) / nrow(df) < 0.3)]
# exclude duplication
df <- df[!duplicated(df),]
names(df)
```

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

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

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

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

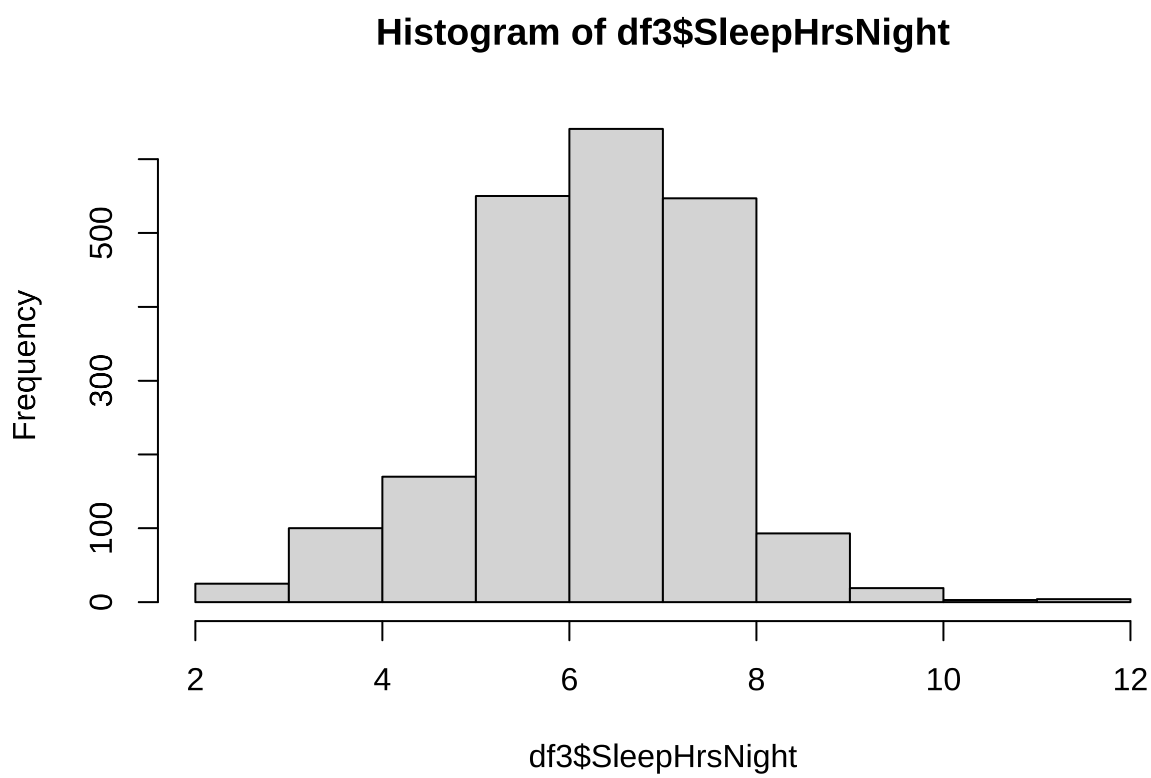
```
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

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

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

```
##                 vars    n   mean    sd median trimmed   mad   min     max
## SleepHrsNight      1 2152   6.78  1.31   7.00    6.85  1.48  2.00   12.00
## BMI                2 2152  28.77  6.75  27.60   28.09  5.78 15.02   69.00
## DirectChol         3 2152   1.35  0.41   1.29    1.31  0.39  0.39    3.83
## Age                4 2152  39.18 11.33  39.00   39.15 14.83 20.00   59.00
## Gender*            5 2152   1.53  0.50   2.00    1.54  0.00  1.00    2.00
## Race1*             6 2152   3.43  1.15   4.00    3.57  0.00  1.00    5.00
## TotChol            7 2152   5.07  1.05   4.99    5.01  1.04  1.53   13.65
## BPDiaAve           8 2152  71.19 11.84  71.00   71.28 10.38  0.00  116.00
## BPSysAve           9 2152 117.43 14.28 116.00  116.50 13.34 78.00  209.00
## AlcoholYear       10 2152  70.59 94.22  24.00   50.94 35.58  0.00  364.00
## Poverty           11 2152   2.84  1.69   2.78    2.89  2.49  0.00    5.00
## SexNumPartnLife   12 2152  16.73 66.13   7.00    8.91  5.93  0.00 2000.00
## SexNumPartYear    13 2152   1.38  2.59   1.00    1.04  0.00  0.00   69.00
## DaysMentHlthBad   14 2152   4.47  8.02   0.00    2.40  0.00  0.00   30.00
## UrineFlow1        15 2152   1.07  0.97   0.81    0.91  0.60  0.00   10.14
## PhysActive*       16 2152   1.58  0.49   2.00    1.60  0.00  1.00    2.00
## DaysPhysHlthBad   17 2152   3.16  7.19   0.00    1.12  0.00  0.00   30.00
```

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

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




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

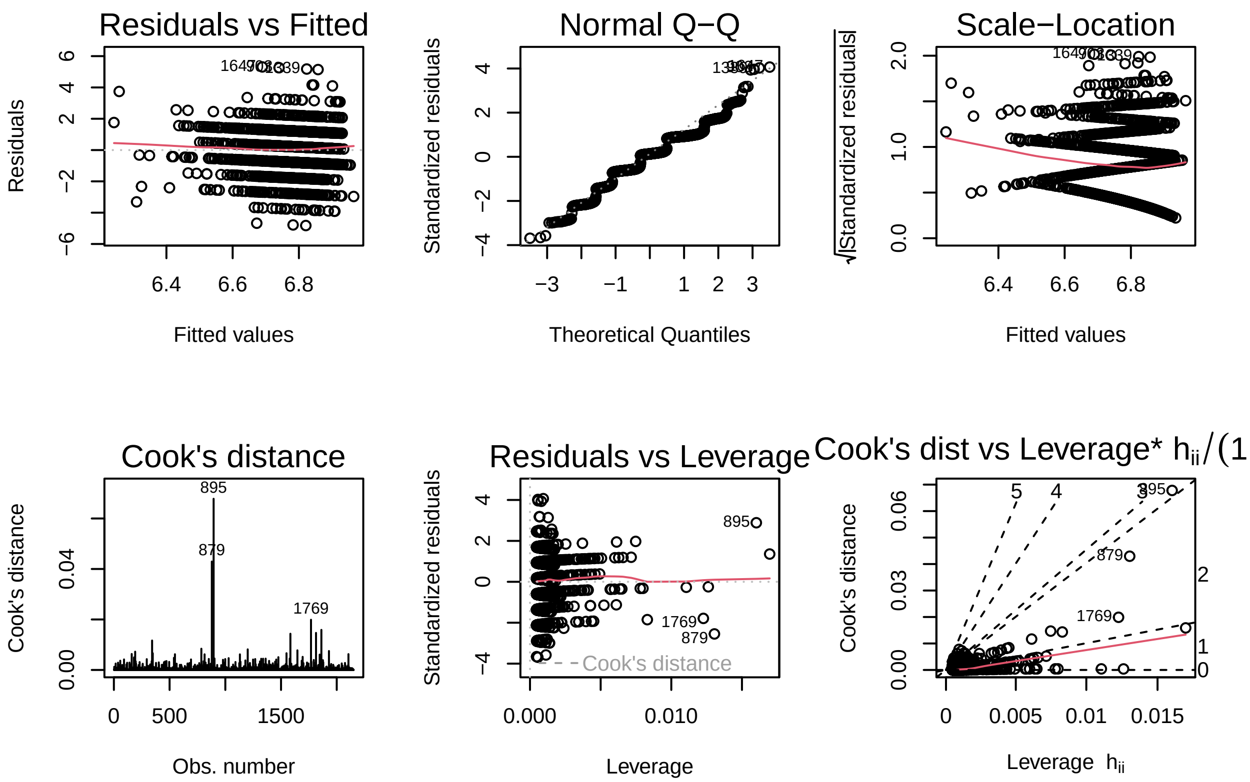
```
## highest: Excellent Vgood     Good      Fair      Poor
##
## Value      Excellent     Vgood      Good      Fair      Poor
## Frequency        240       697       854       313        48
## Proportion     0.112     0.324     0.397     0.145     0.022
## ---------------------------------------------------------------------------
## 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
```

```r
## 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









## 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
```

```
# 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 + SleepHrsNight * factor(Race1),
  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 +
##     SleepHrsNight * factor(Race1), data = df3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.958  -4.088  -0.576   3.191  36.357
##
## Coefficients: (1 not defined because of singularities)
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 22.175411   3.663849   6.052 1.68e-09 ***
## SleepHrsNight               -0.672481   0.440017  -1.528  0.12658
## Age                         -0.080205   0.063471  -1.264  0.20649
## Gender                       3.956938   1.441705   2.745  0.00611 **
## Race1                        0.266774   0.812363   0.328  0.74265
## Poverty                      0.054070   0.091689   0.590  0.55544
## TotChol                      0.012933   0.135840   0.095  0.92416
## BPDiaAve                     0.057750   0.013676   4.223 2.52e-05 ***
## BPSysAve                     0.052227   0.011793   4.429 9.96e-06 ***
## AlcoholYear                 -0.009047   0.001517  -5.966 2.84e-09 ***
## Smoke100Yes                 -0.847770   0.287236  -2.951  0.00320 **
## UrineFlow1                  -0.088739   0.142102  -0.624  0.53238
## DaysMentHlthBad             -0.032621   0.017991  -1.813  0.06993 .
## DaysPhysHlthBad              0.014998   0.020905   0.717  0.47319
## HealthGenVgood               1.882401   0.469175   4.012 6.23e-05 ***
## HealthGenGood                3.613081   0.467141   7.734 1.59e-14 ***
## HealthGenFair                5.346537   0.574641   9.304  < 2e-16 ***
## HealthGenPoor                7.518320   1.075750   6.989 3.69e-12 ***
## PhysActiveYes               -0.891431   0.294530  -3.027  0.00250 **
## factor(Race1)2              -1.143338   2.946825  -0.388  0.69806
```

```
## factor(Race1)3              -3.941114   2.751175  -1.433  0.15214
## factor(Race1)4              -5.899855   2.349469  -2.511  0.01211 *
## factor(Race1)5                     NA         NA      NA       NA
## SleepHrsNight:Age            0.012971   0.009134   1.420  0.15574
## SleepHrsNight:Gender        -0.508514   0.207897  -2.446  0.01453 *
## SleepHrsNight:factor(Race1)2 -0.160437   0.452870  -0.354  0.72317
## SleepHrsNight:factor(Race1)3  0.334059   0.403929   0.827  0.40832
## SleepHrsNight:factor(Race1)4  0.544607   0.287394   1.895  0.05823 .
## SleepHrsNight:factor(Race1)5 -0.629379   0.475731  -1.323  0.18599
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.217 on 2124 degrees of freedom
## Multiple R-squared:  0.1631, Adjusted R-squared:  0.1525
## F-statistic: 15.33 on 27 and 2124 DF,  p-value: < 2.2e-16
```

(4) Diagnosis: 10-fold CV

```
library(caret)
```

```
## Loading required package: ggplot2
```

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

```
splitIndex <-
  createDataPartition(df3$SleepHrsNight, p = 0.7, list = FALSE)
trainData <- df3[splitIndex, ]
testData <- df3[-splitIndex, ]
predictions <- predict(m_3, newdata = testData)
mse <- mean((testData$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
##
## 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
```

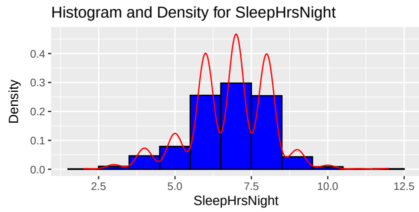
```
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

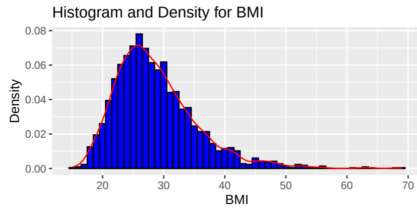
```
(cv_results <- cv_model$results)
```

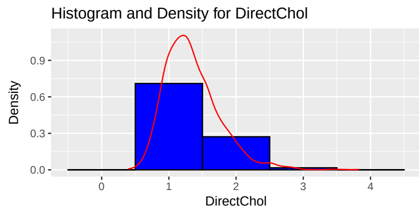
```
##   intercept     RMSE   Rsquared       MAE     RMSESD RsquaredSD      MAESD
## 1      TRUE 1.280209 0.05043061 0.9931499 0.04543809 0.02732622 0.02794626
```

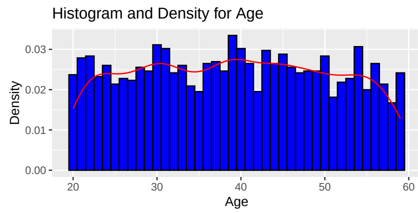
(4) Diagnosis: Normality Assumption

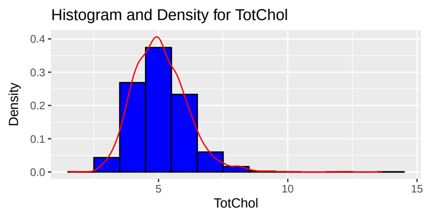
```
library(ggplot2)
library(patchwork)
# Initializes an empty patchwork object
plot_list <- list()

# 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)
```

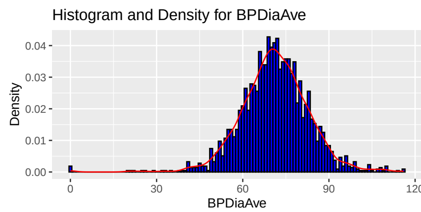

Histogram and Density for SleepHrsNight
Density
SleepHrsNight

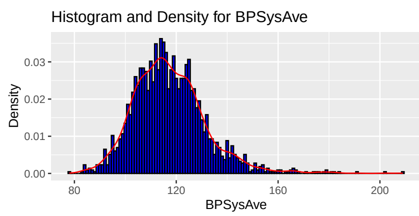

Histogram and Density for BMI
Density
BMI

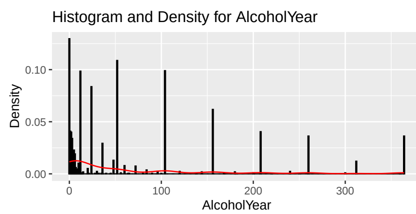

Histogram and Density for DirectChol
Density
DirectChol

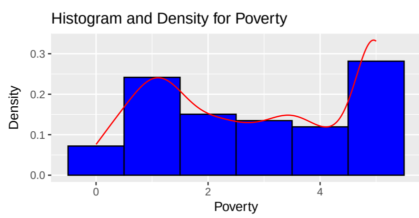

Histogram and Density for Age
Density
Age

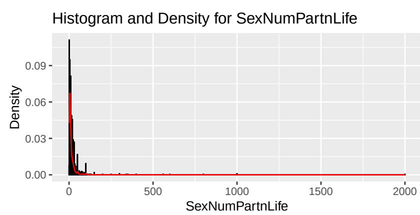

Histogram and Density for TotChol
Density
TotChol

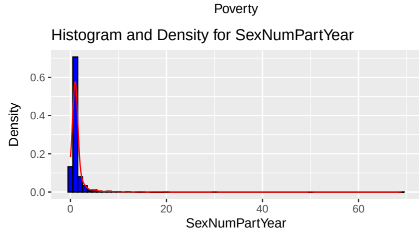

Histogram and Density for BPDiaAve
Density
BPDiaAve

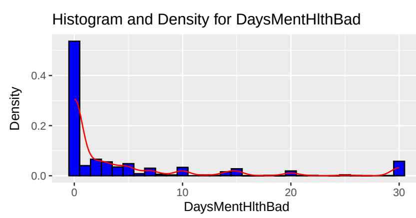

Histogram and Density for BPSysAve
Density
BPSysAve

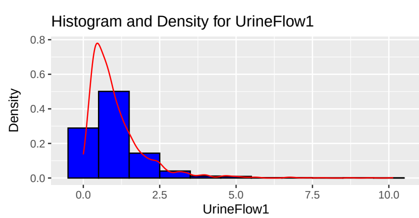

Histogram and Density for AlcoholYear
Density
AlcoholYear

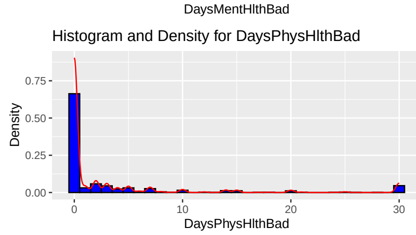

Histogram and Density for Poverty
Density
Poverty

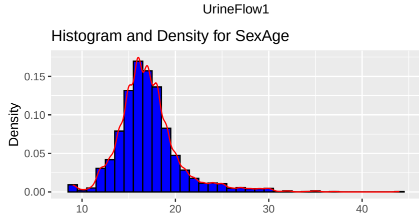

Histogram and Density for SexNumPartnLife
Density
SexNumPartnLife

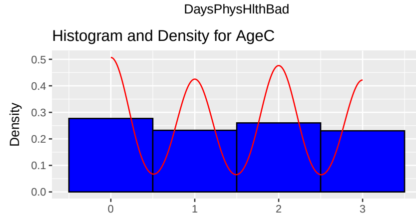

Histogram and Density for SexNumPartYear
Density
SexNumPartYear


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
```

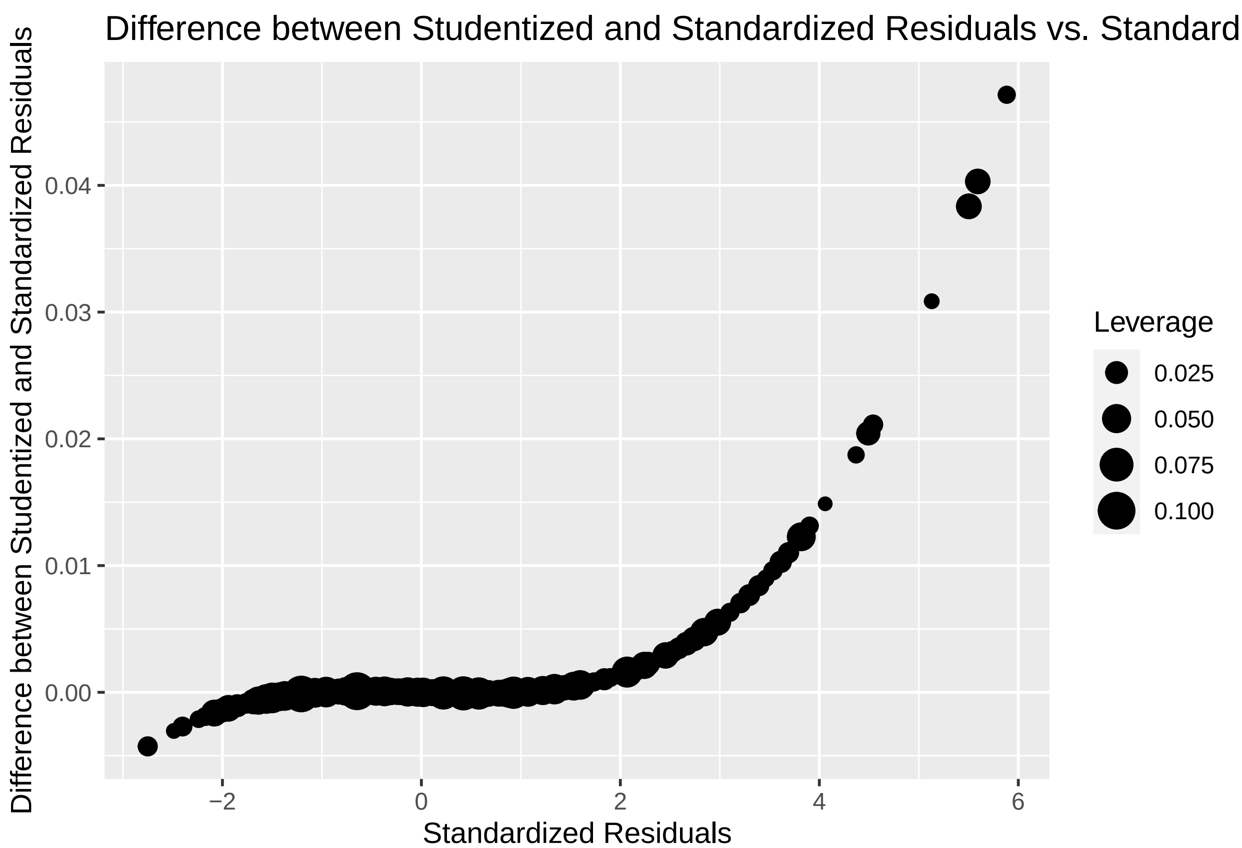
Standardized residuals, Studentized residuals

```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 1.457066e-16 6.178303 -16.958137 36.356759
## 2           Standardized 2.759984e-05 1.001216  -2.750704  5.883818
## 3            Studentized 2.756080e-04 1.002353  -2.754968  5.930967
## 4 Externally Studentized 2.756080e-04 1.002353  -2.754968  5.930967
```

```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")
```

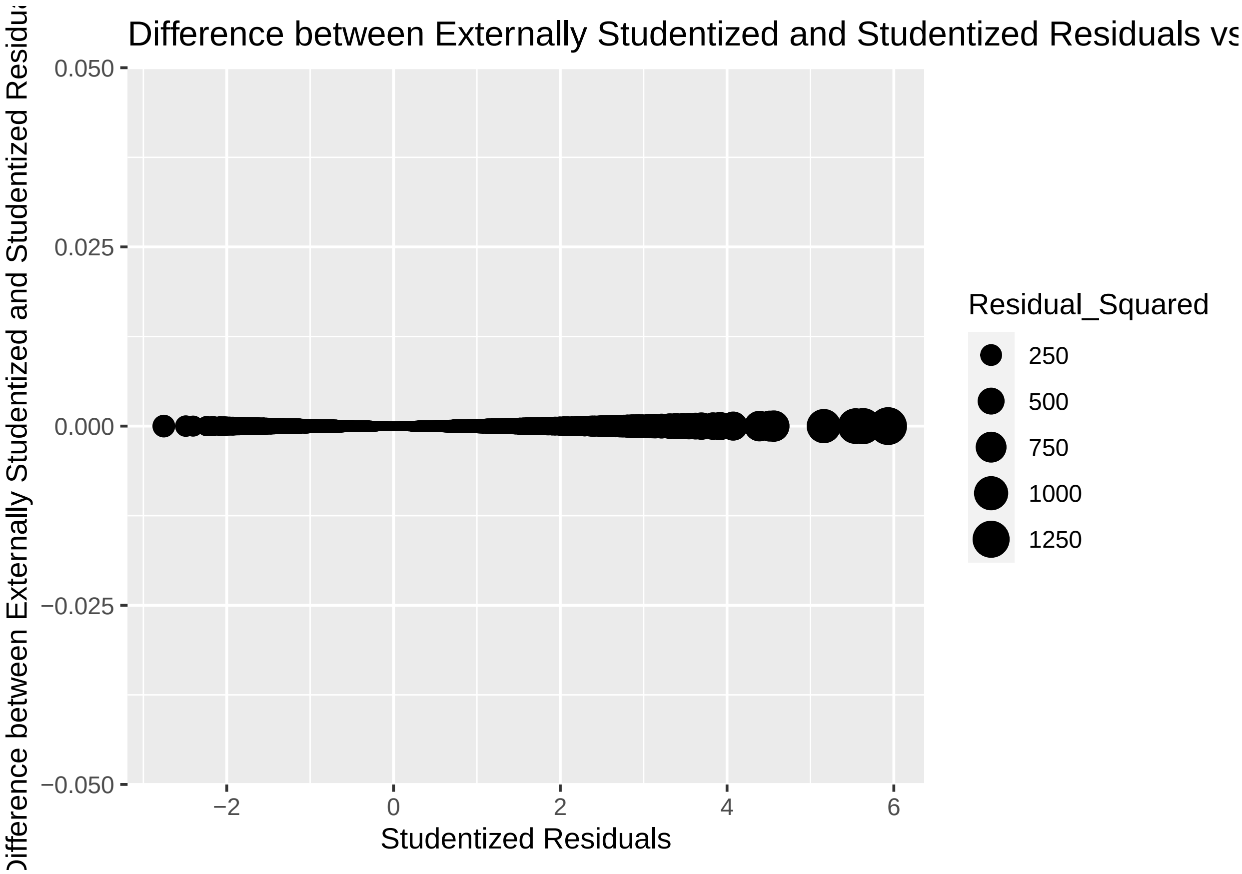



```r
# Display the plot
print(ggplot)
```

```
## function (data = NULL, mapping = aes(), ..., environment = parent.frame())
## {
##     UseMethod("ggplot")
## }
## <bytecode: 0x4c54e60>
## <environment: namespace:ggplot2>
```

```r
# 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,
```

```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")
```

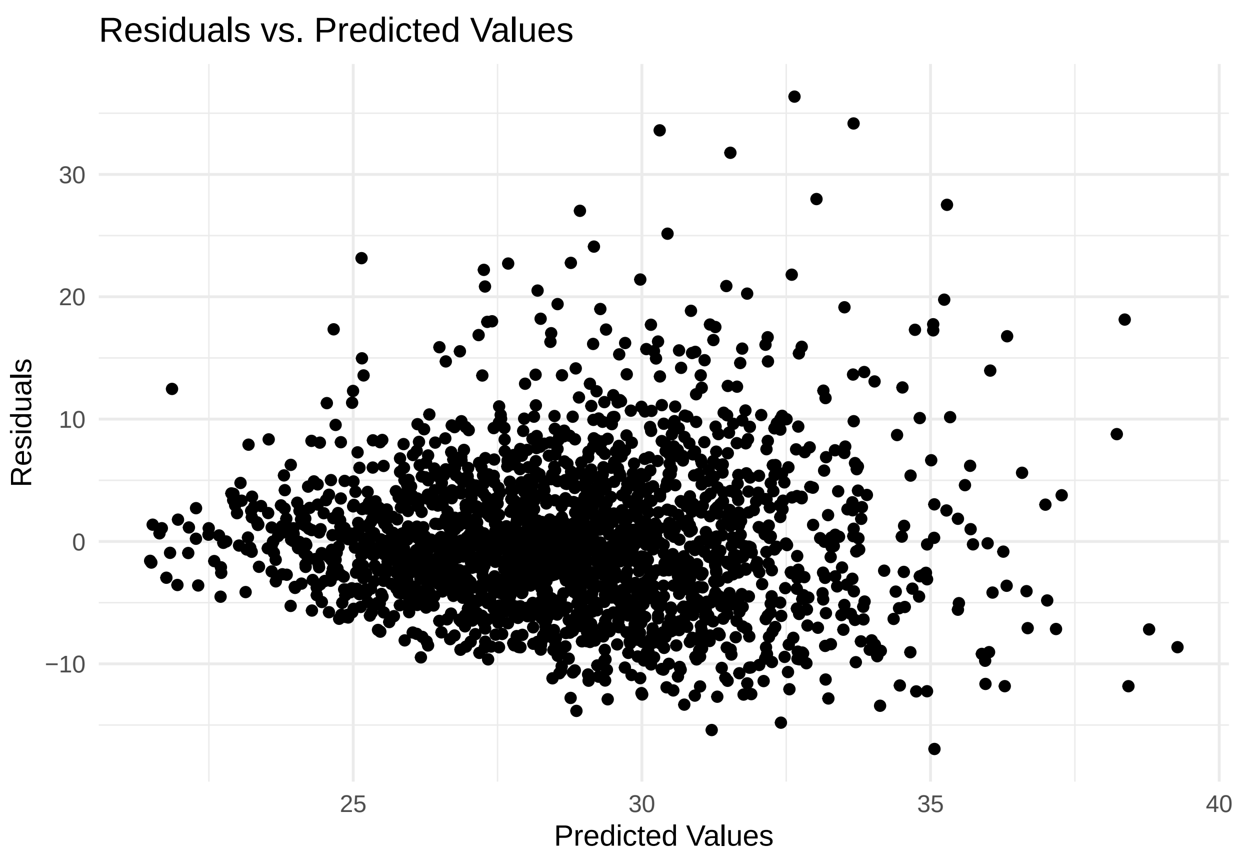



```r
# Display the plot
print(ggplot)
```

```
## function (data = NULL, mapping = aes(), ..., environment = parent.frame())
## {
##     UseMethod("ggplot")
## }
## <bytecode: 0x4c54e60>
## <environment: namespace:ggplot2>
```

```r
# 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()
```

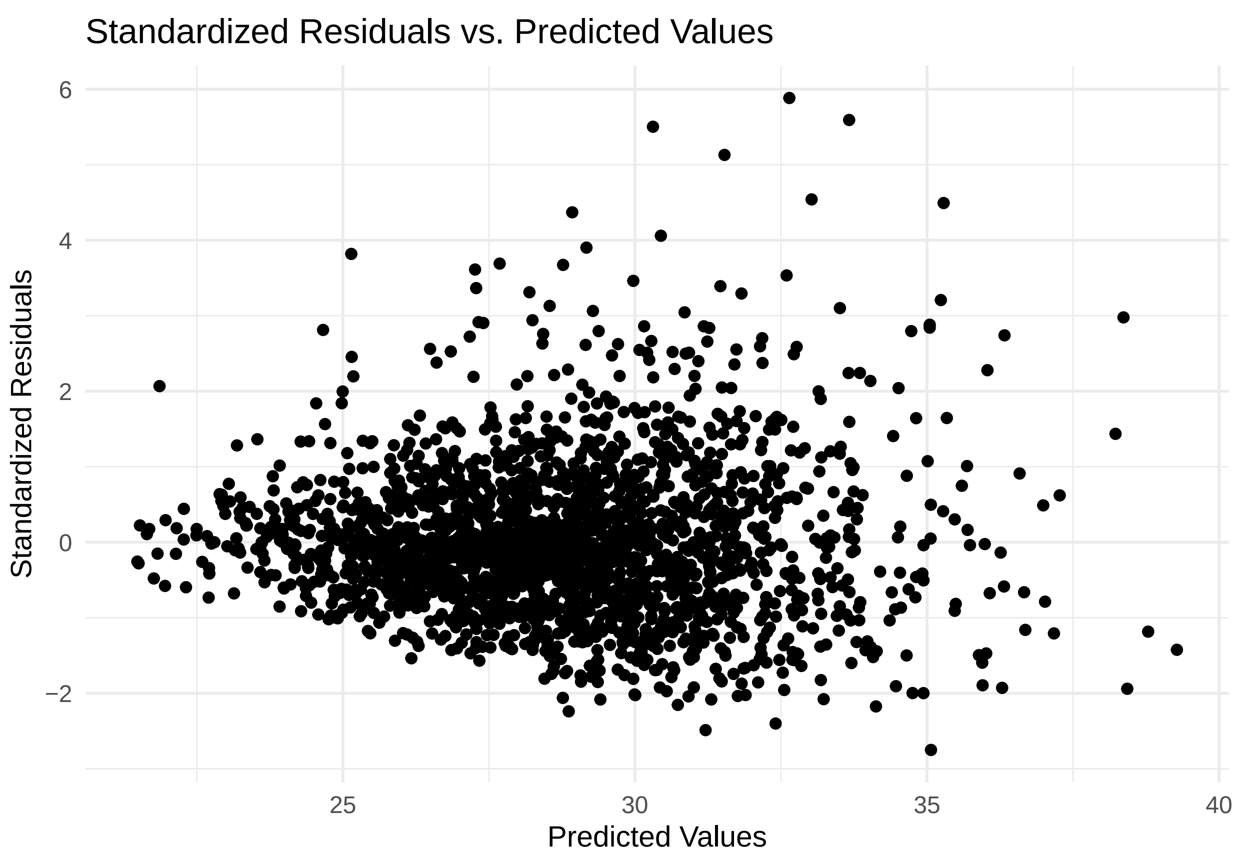



```r
# Display the plot
print(ggplot)
```

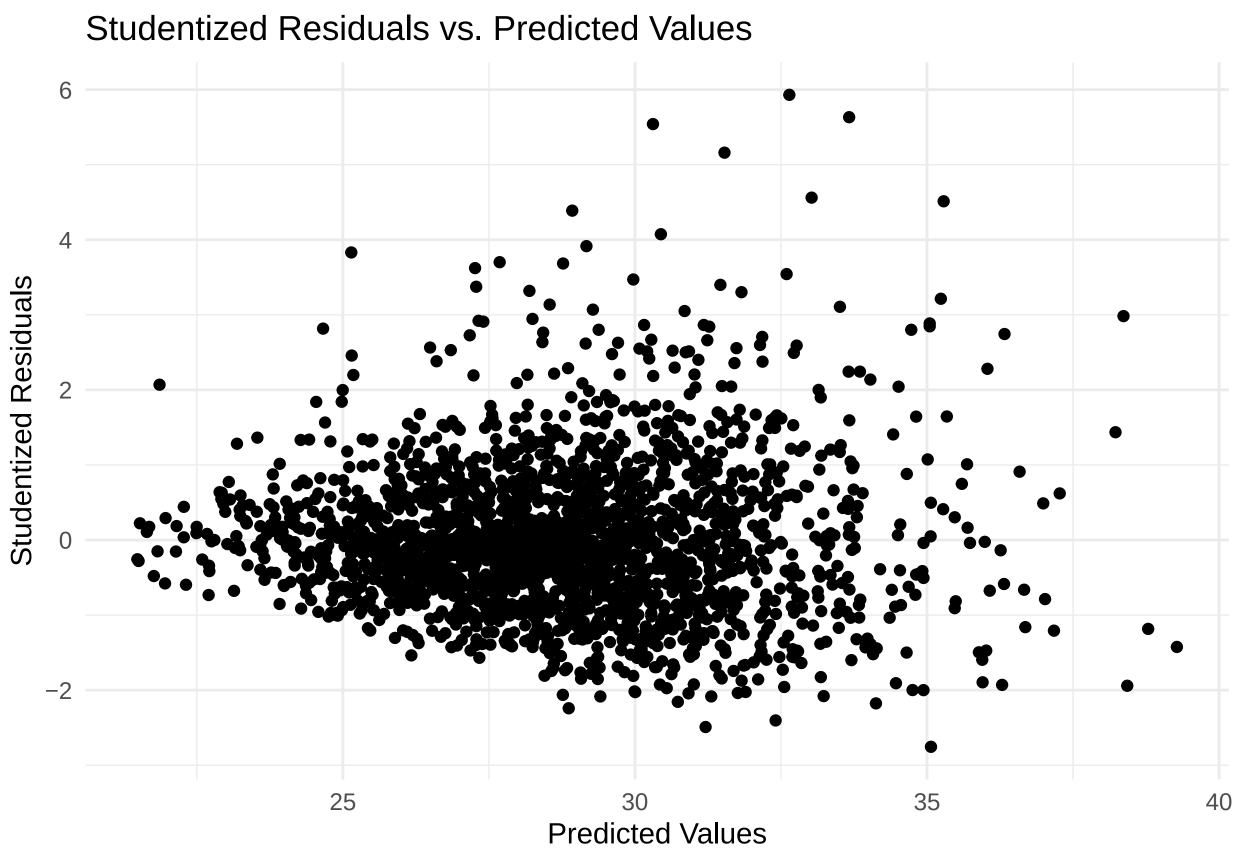
```
## function (data = NULL, mapping = aes(), ..., environment = parent.frame())
## {
##     UseMethod("ggplot")
## }
## <bytecode: 0x4c54e60>
## <environment: namespace:ggplot2>
```

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

# Assuming m_full is your linear model
```

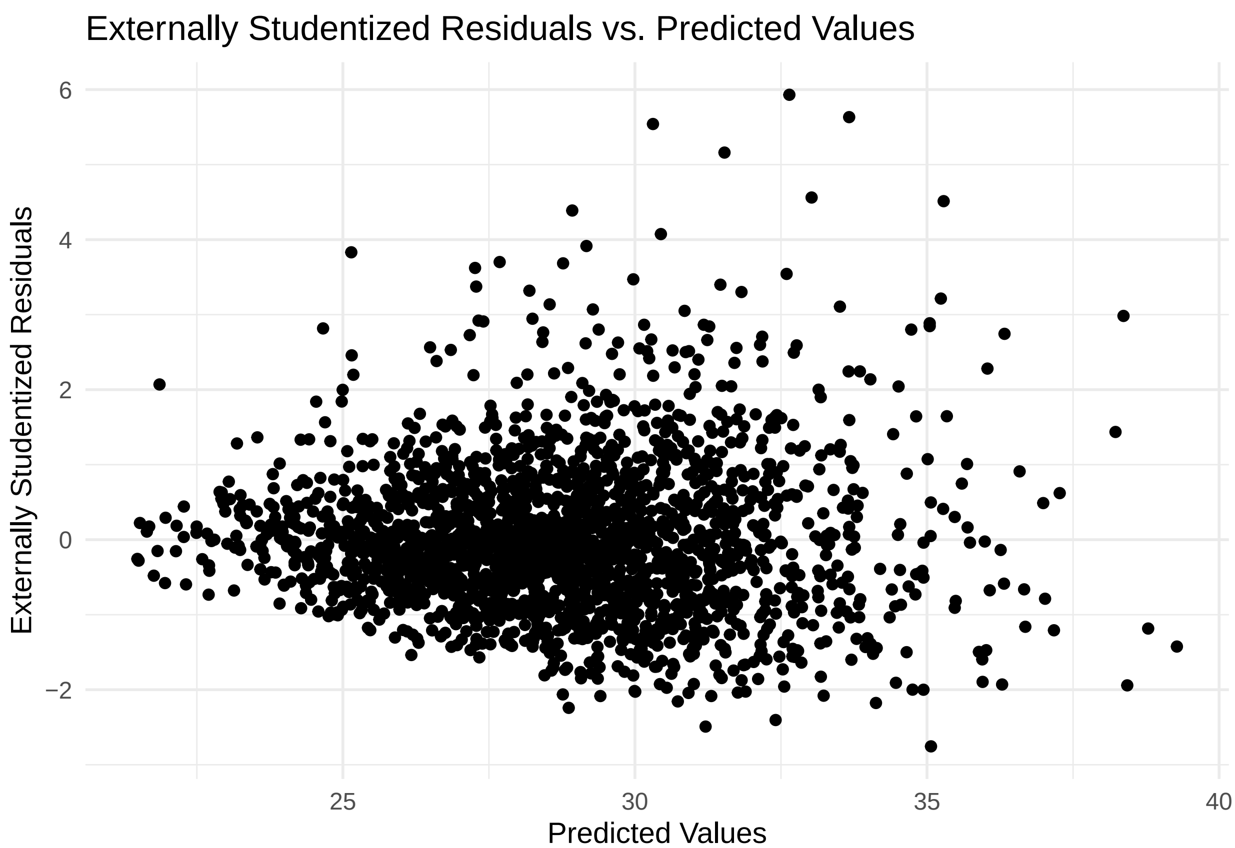
```r
# m_full <- lm(SleepMinNight ~ ., data = df3)

# 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()
```




```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(m_full)
```

```
## Start:  AIC=7892.77
## BMI ~ SleepHrsNight + Age + Gender + Race1 + Poverty + TotChol +
##     BPDiaAve + BPSysAve + AlcoholYear + Smoke100 + UrineFlow1 +
##     DaysMentHlthBad + DaysPhysHlthBad + HealthGen + PhysActive +
##     SleepHrsNight * Age + SleepHrsNight * Gender + SleepHrsNight *
##     factor(Race1)
##
##
## Step:  AIC=7892.77
## BMI ~ SleepHrsNight + Age + Gender + Poverty + TotChol + BPDiaAve +
##     BPSysAve + AlcoholYear + Smoke100 + UrineFlow1 + DaysMentHlthBad +
##     DaysPhysHlthBad + HealthGen + PhysActive + factor(Race1) +
##     SleepHrsNight:Age + SleepHrsNight:Gender + SleepHrsNight:factor(Race1)
##
##                                  Df Sum of Sq   RSS    AIC
## - TotChol                         1       0.4 82107 7890.8
## - Poverty                         1      13.4 82120 7891.1
## - UrineFlow1                      1      15.1 82122 7891.2
## - DaysPhysHlthBad                 1      19.9 82127 7891.3
## <none>                                        82107 7892.8
```

```
## - SleepHrsNight:Age            1      78.0 82185 7892.8
## - DaysMentHlthBad              1     127.1 82234 7894.1
## - SleepHrsNight:factor(Race1)  4     430.5 82537 7896.0
## - SleepHrsNight:Gender         1     231.3 82338 7896.8
## - Smoke100                     1     336.7 82444 7899.6
## - PhysActive                   1     354.1 82461 7900.0
## - BPDiaAve                     1     689.3 82796 7908.8
## - BPSysAve                     1     758.2 82865 7910.6
## - AlcoholYear                  1    1375.8 83483 7926.5
## - HealthGen                    4    4622.1 86729 8002.6
##
## Step:  AIC=7890.78
## BMI ~ SleepHrsNight + Age + Gender + Poverty + BPDiaAve + BPSysAve +
##     AlcoholYear + Smoke100 + UrineFlow1 + DaysMentHlthBad + DaysPhysHlthBad +
##     HealthGen + PhysActive + factor(Race1) + SleepHrsNight:Age +
##     SleepHrsNight:Gender + SleepHrsNight:factor(Race1)
##
##                               Df Sum of Sq   RSS    AIC
## - Poverty                      1      13.4 82120 7889.1
## - UrineFlow1                   1      15.0 82122 7889.2
## - DaysPhysHlthBad              1      19.9 82127 7889.3
## <none>                                     82107 7890.8
## - SleepHrsNight:Age            1      78.5 82186 7890.8
## - DaysMentHlthBad              1     127.4 82235 7892.1
## - SleepHrsNight:factor(Race1)  4     430.5 82538 7894.0
## - SleepHrsNight:Gender         1     231.6 82339 7894.8
## - Smoke100                     1     338.1 82445 7897.6
## - PhysActive                   1     354.6 82462 7898.1
## - BPDiaAve                     1     698.7 82806 7907.0
## - BPSysAve                     1     759.9 82867 7908.6
## - AlcoholYear                  1    1377.1 83484 7924.6
## - HealthGen                    4    4628.8 86736 8000.8
##
## Step:  AIC=7889.13
## BMI ~ SleepHrsNight + Age + Gender + BPDiaAve + BPSysAve + AlcoholYear +
##     Smoke100 + UrineFlow1 + DaysMentHlthBad + DaysPhysHlthBad +
##     HealthGen + PhysActive + factor(Race1) + SleepHrsNight:Age +
##     SleepHrsNight:Gender + SleepHrsNight:factor(Race1)
##
##                               Df Sum of Sq   RSS    AIC
## - UrineFlow1                   1      13.2 82134 7887.5
## - DaysPhysHlthBad              1      19.3 82140 7887.6
## <none>                                     82120 7889.1
## - SleepHrsNight:Age            1      81.0 82201 7889.3
## - DaysMentHlthBad              1     133.0 82253 7890.6
## - SleepHrsNight:factor(Race1)  4     431.2 82552 7892.4
## - SleepHrsNight:Gender         1     228.7 82349 7893.1
## - PhysActive                   1     342.9 82463 7896.1
## - Smoke100                     1     373.7 82494 7896.9
## - BPDiaAve                     1     702.9 82823 7905.5
## - BPSysAve                     1     751.3 82872 7906.7
## - AlcoholYear                  1    1363.8 83484 7922.6
## - HealthGen                    4    4694.1 86815 8000.8
##
```

```
## Step:  AIC=7887.48
## BMI ~ SleepHrsNight + Age + Gender + BPDiaAve + BPSysAve + AlcoholYear +
##     Smoke100 + DaysMentHlthBad + DaysPhysHlthBad + HealthGen +
##     PhysActive + factor(Race1) + SleepHrsNight:Age + SleepHrsNight:Gender +
##     SleepHrsNight:factor(Race1)
##
##                               Df Sum of Sq   RSS    AIC
## - DaysPhysHlthBad              1      19.3 82153 7886.0
## <none>                                     82134 7887.5
## - SleepHrsNight:Age            1      83.3 82217 7887.7
## - DaysMentHlthBad              1     134.4 82268 7889.0
## - SleepHrsNight:factor(Race1)  4     433.9 82568 7890.8
## - SleepHrsNight:Gender         1     229.6 82363 7891.5
## - PhysActive                   1     352.7 82486 7894.7
## - Smoke100                     1     372.5 82506 7895.2
## - BPDiaAve                     1     705.1 82839 7903.9
## - BPSysAve                     1     748.9 82883 7905.0
## - AlcoholYear                  1    1388.6 83522 7921.6
## - HealthGen                    4    4725.3 86859 7999.9
##
## Step:  AIC=7885.98
## BMI ~ SleepHrsNight + Age + Gender + BPDiaAve + BPSysAve + AlcoholYear +
##     Smoke100 + DaysMentHlthBad + HealthGen + PhysActive + factor(Race1) +
##     SleepHrsNight:Age + SleepHrsNight:Gender + SleepHrsNight:factor(Race1)
##
##                               Df Sum of Sq   RSS    AIC
## <none>                                     82153 7886.0
## - SleepHrsNight:Age            1      82.2 82235 7886.1
## - DaysMentHlthBad              1     120.8 82274 7887.1
## - SleepHrsNight:factor(Race1)  4     432.0 82585 7889.3
## - SleepHrsNight:Gender         1     230.0 82383 7890.0
## - PhysActive                   1     363.3 82516 7893.5
## - Smoke100                     1     366.2 82519 7893.6
## - BPDiaAve                     1     696.3 82849 7902.1
## - BPSysAve                     1     750.4 82903 7903.5
## - AlcoholYear                  1    1403.8 83557 7920.4
## - HealthGen                    4    5179.9 87333 8009.6
```

```
##
## Call:
## lm(formula = BMI ~ SleepHrsNight + Age + Gender + BPDiaAve +
##     BPSysAve + AlcoholYear + Smoke100 + DaysMentHlthBad + HealthGen +
##     PhysActive + factor(Race1) + SleepHrsNight:Age + SleepHrsNight:Gender +
##     SleepHrsNight:factor(Race1), data = df3)
##
## Coefficients:
##                 (Intercept)                SleepHrsNight
##                   22.630757                    -0.683663
##                         Age                       Gender
##                   -0.079622                     3.952743
##                    BPDiaAve                     BPSysAve
##                    0.057703                     0.051854
##                 AlcoholYear                  Smoke100Yes
##                   -0.009057                    -0.868816
```

```
##              DaysMentHlthBad                HealthGenVgood
##                    -0.031240                      1.870626
##                HealthGenGood                 HealthGenFair
##                     3.611653                      5.361278
##                HealthGenPoor                 PhysActiveYes
##                     7.674279                     -0.892385
##               factor(Race1)2                factor(Race1)3
##                    -0.830286                     -3.287026
##               factor(Race1)4                factor(Race1)5
##                    -5.098631                      1.043670
##            SleepHrsNight:Age          SleepHrsNight:Gender
##                     0.013291                     -0.506819
## SleepHrsNight:factor(Race1)2  SleepHrsNight:factor(Race1)3
##                    -0.169941                      0.312389
## SleepHrsNight:factor(Race1)4  SleepHrsNight:factor(Race1)5
##                     0.545229                     -0.627600
```

```
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)
```

```
## Note: model has aliased coefficients
##       sums of squares computed by model comparison

## Note: model has aliased coefficients
##       sums of squares computed by model comparison
## Note: model has aliased coefficients
##       sums of squares computed by model comparison
## Note: model has aliased coefficients
##       sums of squares computed by model comparison
## Note: model has aliased coefficients
##       sums of squares computed by model comparison
## Note: model has aliased coefficients
##       sums of squares computed by model comparison
## Note: model has aliased coefficients
##       sums of squares computed by model comparison
## Note: model has aliased coefficients
##       sums of squares computed by model comparison
## Note: model has aliased coefficients
##       sums of squares computed by model comparison

## Warning in b * sx: longer object length is not a multiple of shorter object
## length

## Note: model has aliased coefficients
##       sums of squares computed by model comparison
## Note: model has aliased coefficients
##       sums of squares computed by model comparison
## Note: model has aliased coefficients
##       sums of squares computed by model comparison
```

```
## Note: model has aliased coefficients
##       sums of squares computed by model comparison
## Note: model has aliased coefficients
##       sums of squares computed by model comparison
## Note: model has aliased coefficients
##       sums of squares computed by model comparison

## Warning in b * sx: longer object length is not a multiple of shorter object
## length

## Note: model has aliased coefficients
##       sums of squares computed by model comparison
## Note: model has aliased coefficients
##       sums of squares computed by model comparison
## Note: model has aliased coefficients
##       sums of squares computed by model comparison
## Note: model has aliased coefficients
##       sums of squares computed by model comparison
## Note: model has aliased coefficients
##       sums of squares computed by model comparison

## Warning in b * sx: longer object length is not a multiple of shorter object
## length

## Note: model has aliased coefficients
##       sums of squares computed by model comparison
## Note: model has aliased coefficients
##       sums of squares computed by model comparison
## Note: model has aliased coefficients
##       sums of squares computed by model comparison
## Note: model has aliased coefficients
##       sums of squares computed by model comparison

## Warning in b * sx: longer object length is not a multiple of shorter object
## length

## Note: model has aliased coefficients
##       sums of squares computed by model comparison
## Note: model has aliased coefficients
##       sums of squares computed by model comparison
## Note: model has aliased coefficients
##       sums of squares computed by model comparison

## Warning in b * sx: longer object length is not a multiple of shorter object
## length

## Note: model has aliased coefficients
##       sums of squares computed by model comparison
## Note: model has aliased coefficients
##       sums of squares computed by model comparison

## Warning in b * sx: longer object length is not a multiple of shorter object
## length

## Note: model has aliased coefficients
##       sums of squares computed by model comparison

##
##                                      Selection Summary
```

```
## -----------------------------------------------------------------------------------------------
##        Variable                                       Adj.
## Step              Entered                R-Square    R-Square     C(p)          AIC          RMSE
## -----------------------------------------------------------------------------------------------
##    1    HealthGen                           0.0826      0.0809    180.3262    14153.5414    6.4747
##    2    BPDiaAve                            0.1066      0.1045    121.3681    14098.4490    6.3908
##    3    AlcoholYear                         0.1226      0.1201     82.8326    14061.6280    6.3349
##    4    factor(Race1)                       0.1381      0.1341     45.3890    14031.1683    6.2844
##    5    BPSysAve                            0.1450      0.1406     29.8339    14015.8275    6.2606
##    6    Smoke100                            0.1483      0.1436     23.4714    14009.5178    6.2500
##    7    PhysActive                          0.1519      0.1467     16.5415    14002.6086    6.2386
##    8    SleepHrsNight:factor(Race1)         0.1567      0.1496      6.1651    14000.1995    6.2279
##    9    SleepHrsNight                       0.1567      0.1496      8.1651    14000.1995    6.2279
##   10    Gender                              0.1578      0.1503      7.3501    13999.3671    6.2252
##   11    SleepHrsNight:Gender                0.1602      0.1523      3.3336    13995.3009    6.2179
##   12    Poverty                             0.1606      0.1523      4.4769    13996.4356    6.2181
##   13    Race1                               0.1606      0.1523      4.4769    13998.4356    6.2181
##   14    DaysPhysHlthBad                     0.1606      0.1520      6.2801    14000.2368    6.2193
##   15    Age                                 0.1608      0.1518      7.7645    14001.7160    6.2200
##   16    TotChol                             0.1608      0.1514      9.7324    14003.6835    6.2214
##   17    UrineFlow1                          0.1610      0.1512     11.2234    14005.1691    6.2222
##   18    DaysMentHlthBad                     0.1623      0.1521     10.0165    14003.9255    6.2189
## -----------------------------------------------------------------------------------------------
```

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

```
## Note: model has aliased coefficients
##       sums of squares computed by model comparison
## Note: model has aliased coefficients
##       sums of squares computed by model comparison
## Note: model has aliased coefficients
##       sums of squares computed by model comparison
## Note: model has aliased coefficients
##       sums of squares computed by model comparison
## Note: model has aliased coefficients
##       sums of squares computed by model comparison
## Note: model has aliased coefficients
##       sums of squares computed by model comparison
## Note: model has aliased coefficients
##       sums of squares computed by model comparison
## Note: model has aliased coefficients
##       sums of squares computed by model comparison
## Note: model has aliased coefficients
##       sums of squares computed by model comparison

## Warning in b * sx: longer object length is not a multiple of shorter object
## length

## Note: model has aliased coefficients
##       sums of squares computed by model comparison
## Note: model has aliased coefficients
##       sums of squares computed by model comparison
## Note: model has aliased coefficients
##       sums of squares computed by model comparison
## Note: model has aliased coefficients
##       sums of squares computed by model comparison
```

```
## Note: model has aliased coefficients
##       sums of squares computed by model comparison
## Note: model has aliased coefficients
##       sums of squares computed by model comparison

## Warning in b * sx: longer object length is not a multiple of shorter object
## length

## Note: model has aliased coefficients
##       sums of squares computed by model comparison
## Note: model has aliased coefficients
##       sums of squares computed by model comparison
## Note: model has aliased coefficients
##       sums of squares computed by model comparison
## Note: model has aliased coefficients
##       sums of squares computed by model comparison
## Note: model has aliased coefficients
##       sums of squares computed by model comparison

## Warning in b * sx: longer object length is not a multiple of shorter object
## length

## Note: model has aliased coefficients
##       sums of squares computed by model comparison
## Note: model has aliased coefficients
##       sums of squares computed by model comparison
## Note: model has aliased coefficients
##       sums of squares computed by model comparison
## Note: model has aliased coefficients
##       sums of squares computed by model comparison

## Warning in b * sx: longer object length is not a multiple of shorter object
## length

## Note: model has aliased coefficients
##       sums of squares computed by model comparison
## Note: model has aliased coefficients
##       sums of squares computed by model comparison
## Note: model has aliased coefficients
##       sums of squares computed by model comparison

## Warning in b * sx: longer object length is not a multiple of shorter object
## length

## Note: model has aliased coefficients
##       sums of squares computed by model comparison
## Note: model has aliased coefficients
##       sums of squares computed by model comparison

## Warning in b * sx: longer object length is not a multiple of shorter object
## length

## Note: model has aliased coefficients
##       sums of squares computed by model comparison

##
##                                     Selection Summary
## -----------------------------------------------------------------------------------------
##         Variable                                        Adj.
```

```
## Step                Entered                   R-Square     R-Square       C(p)          AIC          RMSE
## ----------------------------------------------------------------------------------------------------
##    1     HealthGen                               0.0826       0.0809     180.3262    14153.5414     6.4747
##    2     BPDiaAve                                0.1066       0.1045     121.3681    14098.4490     6.3908
##    3     AlcoholYear                             0.1226       0.1201      82.8326    14061.6280     6.3349
##    4     factor(Race1)                           0.1381       0.1341      45.3890    14031.1683     6.2844
##    5     BPSysAve                                0.1450       0.1406      29.8339    14015.8275     6.2606
##    6     Smoke100                                0.1483       0.1436      23.4714    14009.5178     6.2500
##    7     PhysActive                              0.1519       0.1467      16.5415    14002.6086     6.2386
##    8     SleepHrsNight:factor(Race1)             0.1567       0.1496       6.1651    14000.1995     6.2279
##    9     SleepHrsNight                           0.1567       0.1496       8.1651    14000.1995     6.2279
##   10     Gender                                  0.1578       0.1503       7.3501    13999.3671     6.2252
##   11     SleepHrsNight:Gender                    0.1602       0.1523       3.3336    13995.3009     6.2179
##   12     Poverty                                 0.1606       0.1523       4.4769    13996.4356     6.2181
##   13     Race1                                   0.1606       0.1523       4.4769    13998.4356     6.2181
##   14     DaysPhysHlthBad                         0.1606       0.1520       6.2801    14000.2368     6.2193
##   15     Age                                     0.1608       0.1518       7.7645    14001.7160     6.2200
##   16     TotChol                                 0.1608       0.1514       9.7324    14003.6835     6.2214
##   17     UrineFlow1                              0.1610       0.1512      11.2234    14005.1691     6.2222
##   18     DaysMentHlthBad                         0.1623       0.1521      10.0165    14003.9255     6.2189
## ----------------------------------------------------------------------------------------------------
```

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

```
## [1] 36.32895
```