STAT463 Project: Sleep Health and Lifestyle

Lisa Lu 31088272

```
# Import libraries
library(dplyr)
library(ggplot2)
library(stringr)
library(gridExtra)
```

Data Exploration and Preprocessing

```
# Read in dataset
dataset <- read.table("Sleep_health_and_lifestyle_dataset.csv",
                      header = TRUE, sep = ',', na.strings = "na")

# Print the first few rows of data frame
head(dataset)
```

```
##   Person.ID Gender Age          Occupation Sleep.Duration Quality.of.Sleep
## 1         1   Male  27    Software Engineer            6.1                6
## 2         2   Male  28              Doctor            6.2                6
## 3         3   Male  28              Doctor            6.2                6
## 4         4   Male  28 Sales Representative            5.9                4
## 5         5   Male  28 Sales Representative            5.9                4
## 6         6   Male  28    Software Engineer            5.9                4
##   Physical.Activity.Level Stress.Level BMI.Category Blood.Pressure Heart.Rate
## 1                      42            6   Overweight         126/83         77
## 2                      60            8       Normal         125/80         75
## 3                      60            8       Normal         125/80         75
## 4                      30            8        Obese         140/90         85
## 5                      30            8        Obese         140/90         85
## 6                      30            8        Obese         140/90         85
##   Daily.Steps Sleep.Disorder
## 1        4200           None
## 2       10000           None
## 3       10000           None
## 4        3000    Sleep Apnea
## 5        3000    Sleep Apnea
## 6        3000       Insomnia
```

```
# Explore the structure of the dataset
str(dataset)
```

```
## 'data.frame':    374 obs. of  13 variables:
```

```
##  $ Person.ID              : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Gender                 : chr  "Male" "Male" "Male" "Male" ...
##  $ Age                    : int  27 28 28 28 28 28 29 29 29 29 ...
##  $ Occupation             : chr  "Software Engineer" "Doctor" "Doctor" "Sales Representative" ...
##  $ Sleep.Duration         : num  6.1 6.2 6.2 5.9 5.9 5.9 6.3 7.8 7.8 7.8 ...
##  $ Quality.of.Sleep       : int  6 6 6 4 4 4 6 7 7 7 ...
##  $ Physical.Activity.Level: int  42 60 60 30 30 30 40 75 75 75 ...
##  $ Stress.Level           : int  6 8 8 8 8 8 7 6 6 6 ...
##  $ BMI.Category           : chr  "Overweight" "Normal" "Normal" "Obese" ...
##  $ Blood.Pressure         : chr  "126/83" "125/80" "125/80" "140/90" ...
##  $ Heart.Rate             : int  77 75 75 85 85 85 82 70 70 70 ...
##  $ Daily.Steps            : int  4200 10000 10000 3000 3000 3000 3500 8000 8000 8000 ...
##  $ Sleep.Disorder         : chr  "None" "None" "None" "Sleep Apnea" ...
```

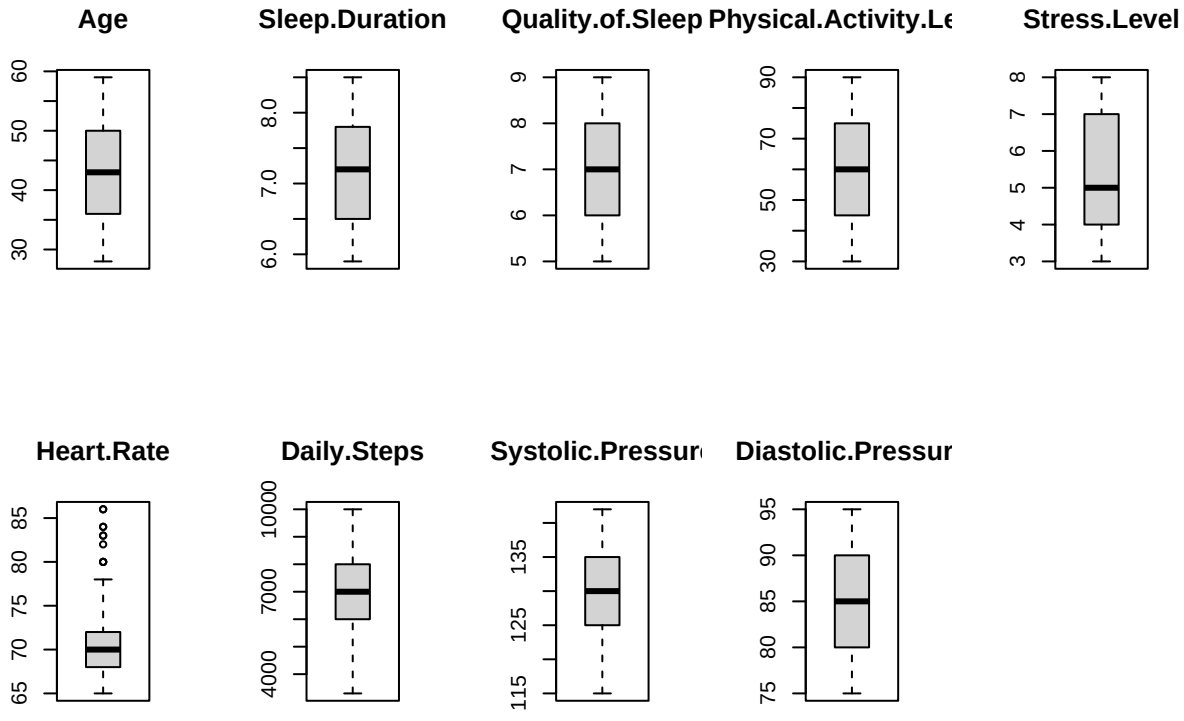
```r
# Get a descriptive statistics
summary(dataset)
```

```
##    Person.ID         Gender               Age         Occupation
##  Min.   :  1.00   Length:374         Min.   :27.00   Length:374
##  1st Qu.: 94.25   Class :character   1st Qu.:35.25   Class :character
##  Median :187.50   Mode  :character   Median :43.00   Mode  :character
##  Mean   :187.50                      Mean   :42.18
##  3rd Qu.:280.75                      3rd Qu.:50.00
##  Max.   :374.00                      Max.   :59.00
##  Sleep.Duration  Quality.of.Sleep Physical.Activity.Level  Stress.Level
##  Min.   :5.800   Min.   :4.000    Min.   :30.00           Min.   :3.000
##  1st Qu.:6.400   1st Qu.:6.000    1st Qu.:45.00           1st Qu.:4.000
##  Median :7.200   Median :7.000    Median :60.00           Median :5.000
##  Mean   :7.132   Mean   :7.313    Mean   :59.17           Mean   :5.385
##  3rd Qu.:7.800   3rd Qu.:8.000    3rd Qu.:75.00           3rd Qu.:7.000
##  Max.   :8.500   Max.   :9.000    Max.   :90.00           Max.   :8.000
##  BMI.Category       Blood.Pressure       Heart.Rate     Daily.Steps
##  Length:374         Length:374         Min.   :65.00   Min.   : 3000
##  Class :character   Class :character   1st Qu.:68.00   1st Qu.: 5600
##  Mode  :character   Mode  :character   Median :70.00   Median : 7000
##                                        Mean   :70.17   Mean   : 6817
##                                        3rd Qu.:72.00   3rd Qu.: 8000
##                                        Max.   :86.00   Max.   :10000
##  Sleep.Disorder
##  Length:374
##  Class :character
##  Mode  :character
##
##
##
```

```r
# Data preprocessing
# Split Blood Pressure column into systolic and diastolic pressure as numeric data
dataset[c('Systolic.Pressure', 'Diastolic.Pressure')] <- as.numeric(str_split_fixed(dataset$Blood.Pressu

# Combine "Normal" and "Normal Weight" values in BMI.Category
dataset$BMI.Category[dataset$BMI.Category == "Normal Weight"] <- "Normal"
# Change "Obese" into "Overweight"
```

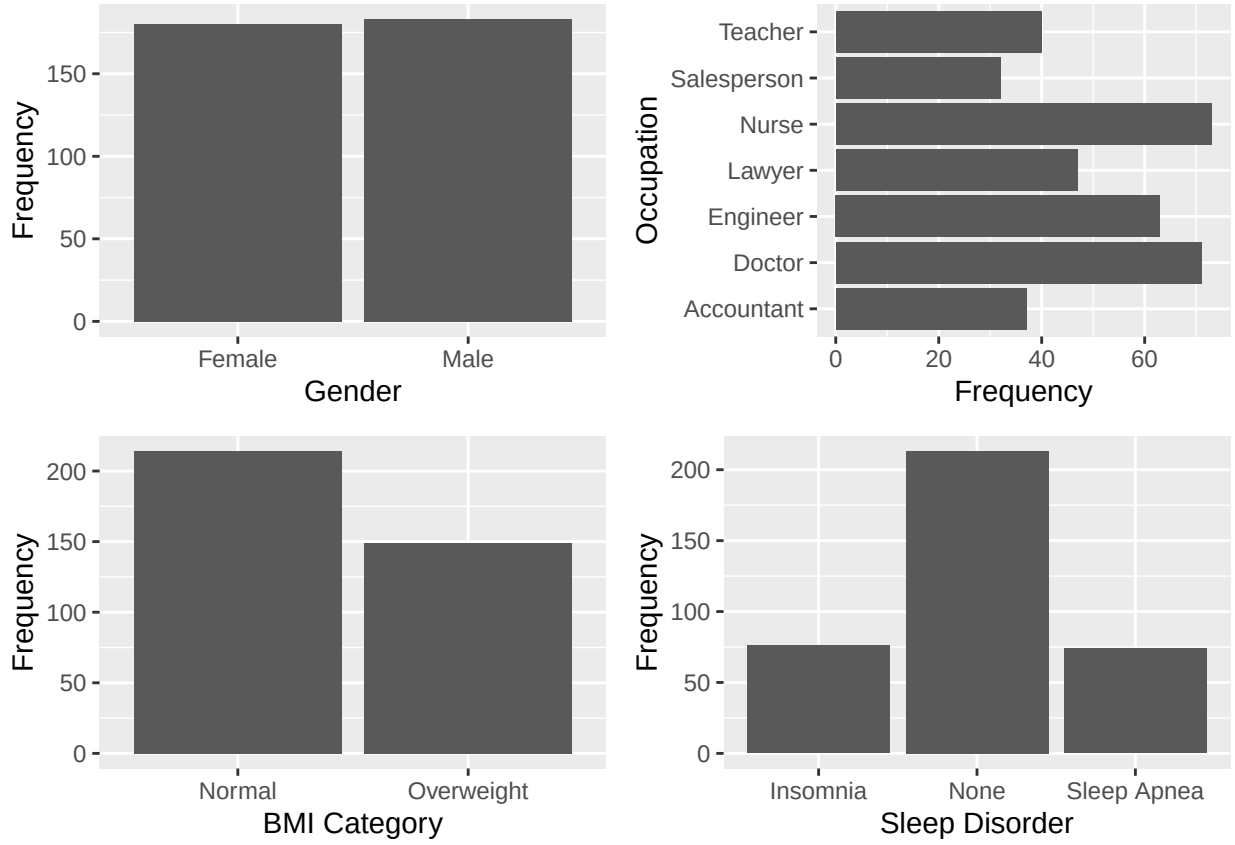
```r
dataset$BMI.Category[dataset$BMI.Category == "Obese"] <- "Overweight"

# Remove Occupations that has count under 10 (namely, Software Engineer, Scientist, Sales Representativ
dataset <- dataset[!(dataset$Occupation =="Software Engineer" | dataset$Occupation =="Scientist" |
          dataset$Occupation =="Sales Representative" |
          dataset$Occupation =="Manager"),]
```

```r
# Get the numeric data
numeric.data <- dataset[,c(3,5,6,7,8,11,12,14,15)]

# Plot the data -- Boxplot for numeric data and Histogram for categorical data
# Boxplots

par(mfrow=c(2,5))
for (i in 1:length(numeric.data)) {
        boxplot(numeric.data[,i], main=names(numeric.data[i]), type="l")
    }
```



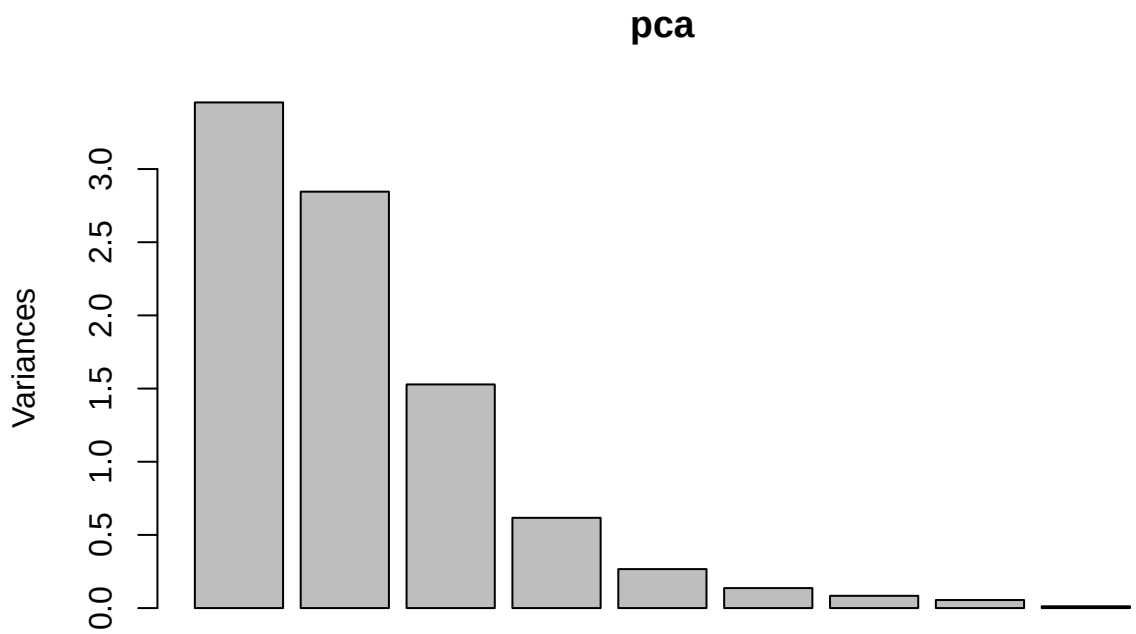

```r
# Frequency charts of the categorical data
gender <- ggplot(data = dataset, aes(x = Gender)) +
geom_bar() +
labs(y = "Frequency", x = "Gender")

occupation <- ggplot(data = dataset, aes(y = Occupation)) +
geom_bar() +
```

```
labs(y = "Occupation", x = "Frequency")

bmi <- ggplot(data = dataset, aes(x = BMI.Category)) +
geom_bar() +
labs(y = "Frequency", x = "BMI Category")

sleep_disorder <- ggplot(data = dataset, aes(x = Sleep.Disorder)) +
geom_bar() +
labs(y = "Frequency", x = "Sleep Disorder")

grid.arrange(gender,occupation, bmi, sleep_disorder, ncol = 2, nrow = 2)
```



Explore the factors affecting quality of sleep

```
# Use PCA analysis
pca <- prcomp(scale(numeric.data))
screeplot(pca)
```




```
summary(pca)
```

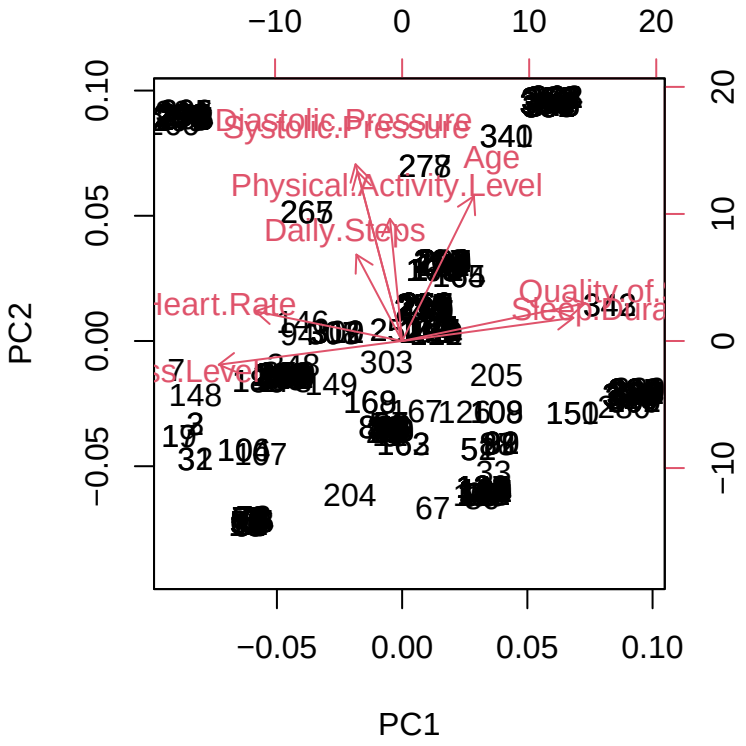
```
## Importance of components:
##                           PC1    PC2    PC3     PC4    PC5     PC6     PC7
## Standard deviation     1.8588 1.6869 1.2362 0.78539 0.5161 0.36972 0.28959
## Proportion of Variance 0.3839 0.3162 0.1698 0.06854 0.0296 0.01519 0.00932
## Cumulative Proportion  0.3839 0.7001 0.8699 0.93843 0.9680 0.98322 0.99254
##                           PC8     PC9
## Standard deviation     0.23447 0.11029
## Proportion of Variance 0.00611 0.00135
## Cumulative Proportion  0.99865 1.00000
```

```
pca$rotation
```

```
##                                  PC1         PC2          PC3         PC4
## Age                        0.19928711  0.44360323 -0.298540468  0.08771766
## Sleep.Duration             0.47580701  0.07097410  0.147664001 -0.39223004
## Quality.of.Sleep           0.50839857  0.11555311  0.088271016 -0.12864909
## Physical.Activity.Level   -0.03396907  0.37410114  0.583481169 -0.22866523
## Stress.Level              -0.50935587 -0.07257793  0.044612374  0.05249403
## Heart.Rate                -0.40800403  0.09097020 -0.006752184 -0.77875457
## Daily.Steps               -0.12716351  0.26462392  0.634042308  0.39106322
## Systolic.Pressure         -0.12416495  0.51817687 -0.307746206  0.03352334
## Diastolic.Pressure        -0.12950346  0.54110853 -0.205118630  0.07971363
```

```
##                                PC5         PC6         PC7         PC8
## Age                    -0.759419284  0.12742274  0.25945473 -0.08005358
## Sleep.Duration         -0.002161696 -0.72099302  0.08530658 -0.25294162
## Quality.of.Sleep       -0.061329680  0.14508373 -0.50037733  0.62508863
## Physical.Activity.Level  0.162440451  0.23941747  0.56829562  0.24389998
## Stress.Level           -0.308697422 -0.55162110  0.08243336  0.53688313
## Heart.Rate             -0.207320541  0.18357928 -0.33597242 -0.16950227
## Daily.Steps            -0.223331870 -0.08818022 -0.45430985 -0.27923957
## Systolic.Pressure       0.299336157 -0.16856011 -0.12913019  0.22599882
## Diastolic.Pressure      0.339667903 -0.10671010 -0.09656019 -0.18265880
##                                PC9
## Age                     0.0006974879
## Sleep.Duration          0.0427669874
## Quality.of.Sleep       -0.1947103477
## Physical.Activity.Level -0.0012798921
## Stress.Level           -0.1895374969
## Heart.Rate              0.0214759011
## Daily.Steps             0.1297631409
## Systolic.Pressure       0.6591506141
## Diastolic.Pressure     -0.6874249478
```

```
biplot(pca)
```

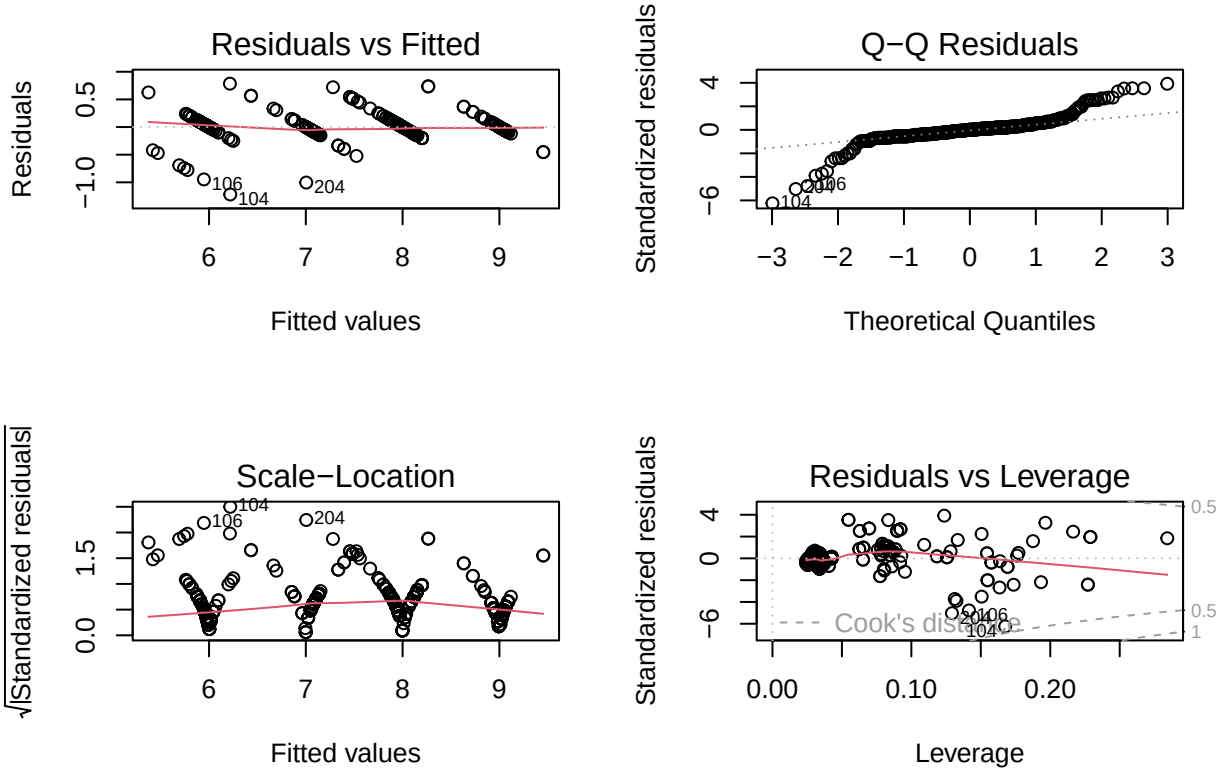



```
# Use linear regression model
sleep <- dataset[, -c(1, 10)]
```

```r
# Get the full multiple linear regression model
lr_full <- lm(Quality.of.Sleep ~ ., data = sleep)
summary(lr_full)
```

```
##
## Call:
## lm(formula = Quality.of.Sleep ~ ., data = sleep)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.21920 -0.08037  0.00079  0.05887  0.78430
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 5.981e+00  8.365e-01   7.150 5.25e-12 ***
## GenderMale                  5.439e-01  6.814e-02   7.982 2.17e-14 ***
## Age                         5.301e-02  4.990e-03  10.623  < 2e-16 ***
## OccupationDoctor           -5.502e-01  7.836e-02  -7.022 1.18e-11 ***
## OccupationEngineer         -7.224e-01  7.763e-02  -9.306  < 2e-16 ***
## OccupationLawyer           -4.678e-01  9.166e-02  -5.104 5.51e-07 ***
## OccupationNurse            -4.195e-01  9.050e-02  -4.635 5.07e-06 ***
## OccupationSalesperson      -9.431e-01  8.827e-02 -10.685  < 2e-16 ***
## OccupationTeacher          -5.804e-01  7.276e-02  -7.977 2.25e-14 ***
## Sleep.Duration              2.816e-01  4.575e-02   6.154 2.10e-09 ***
## Physical.Activity.Level    -2.122e-03  1.441e-03  -1.473 0.141747
## Stress.Level               -3.957e-01  2.041e-02 -19.388  < 2e-16 ***
## BMI.CategoryOverweight     -4.135e-01  8.594e-02  -4.811 2.25e-06 ***
## Heart.Rate                 -1.156e-02  6.433e-03  -1.797 0.073205 .
## Daily.Steps                 5.140e-05  1.988e-05   2.585 0.010149 *
## Sleep.DisorderNone          1.979e-01  5.148e-02   3.843 0.000145 ***
## Sleep.DisorderSleep Apnea   2.723e-01  5.684e-02   4.791 2.47e-06 ***
## Systolic.Pressure           2.544e-02  1.483e-02   1.716 0.087141 .
## Diastolic.Pressure         -3.784e-02  2.003e-02  -1.889 0.059678 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2139 on 344 degrees of freedom
## Multiple R-squared:  0.9667, Adjusted R-squared:  0.9649
## F-statistic: 554.4 on 18 and 344 DF,  p-value: < 2.2e-16
```

```r
par(mfrow = c(2,2))
plot(lr_full)
```




```
cor(numeric.data, numeric.data$Quality.of.Sleep)
```

```
##                                [,1]
## Age                       0.44999752
## Sleep.Duration            0.88356596
## Quality.of.Sleep          1.00000000
## Physical.Activity.Level   0.14682864
## Stress.Level             -0.90722043
## Heart.Rate               -0.61066265
## Daily.Steps              -0.07080688
## Systolic.Pressure        -0.08851489
## Diastolic.Pressure       -0.09182610
```

```
library(lme4)
```

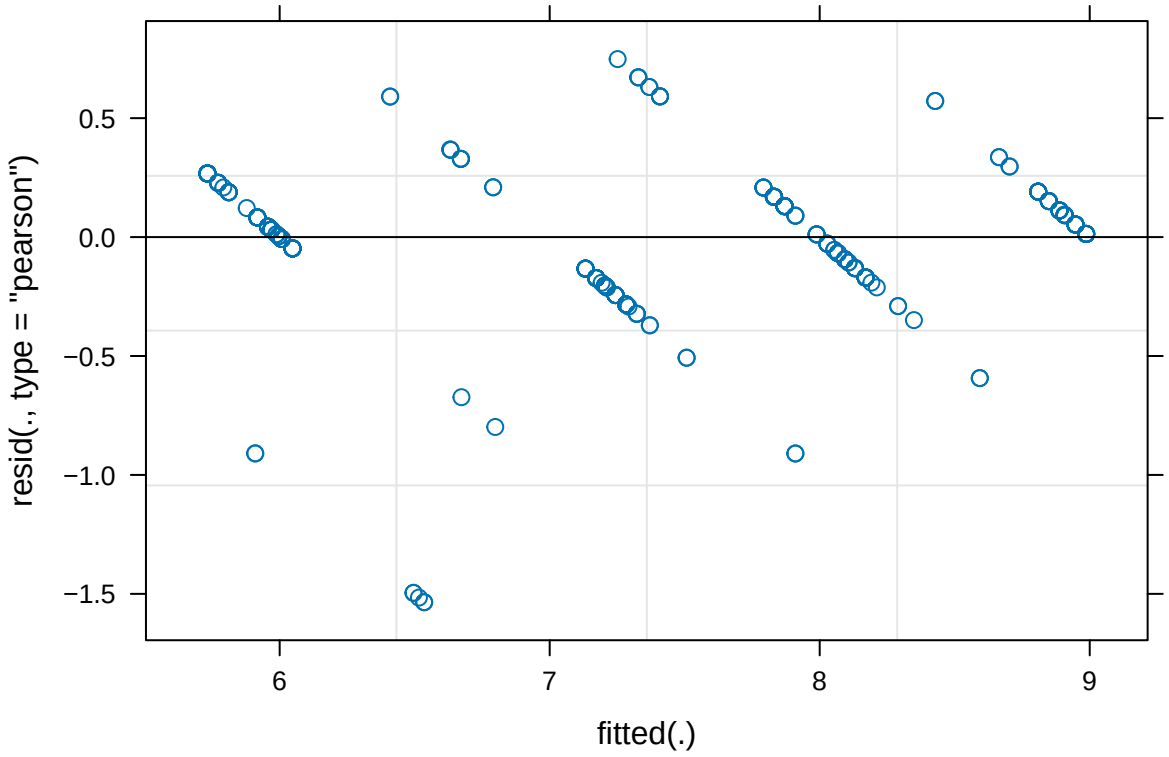
```
## Loading required package: Matrix
```

```
m1 <- lmer(Quality.of.Sleep ~ Sleep.Duration + Stress.Level + (1|Occupation), data = sleep)
summary(m1)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Quality.of.Sleep ~ Sleep.Duration + Stress.Level + (1 | Occupation)
##    Data: sleep
##
```

```
## REML criterion at convergence: 199.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.0854 -0.4395  0.1433  0.5595  2.4792
##
## Random effects:
##  Groups     Name        Variance Std.Dev.
##  Occupation (Intercept) 0.10052  0.3170
##  Residual               0.09112  0.3019
## Number of obs: 363, groups:  Occupation, 7
##
## Fixed effects:
##                 Estimate Std. Error t value
## (Intercept)     6.76664    0.47721  14.179
## Sleep.Duration  0.39383    0.04984   7.903
## Stress.Level   -0.42085    0.02251 -18.699
##
## Correlation of Fixed Effects:
##             (Intr) Slp.Dr
## Sleep.Durtn -0.959
## Stress.Levl -0.895  0.869
```

```
plot(m1)
```



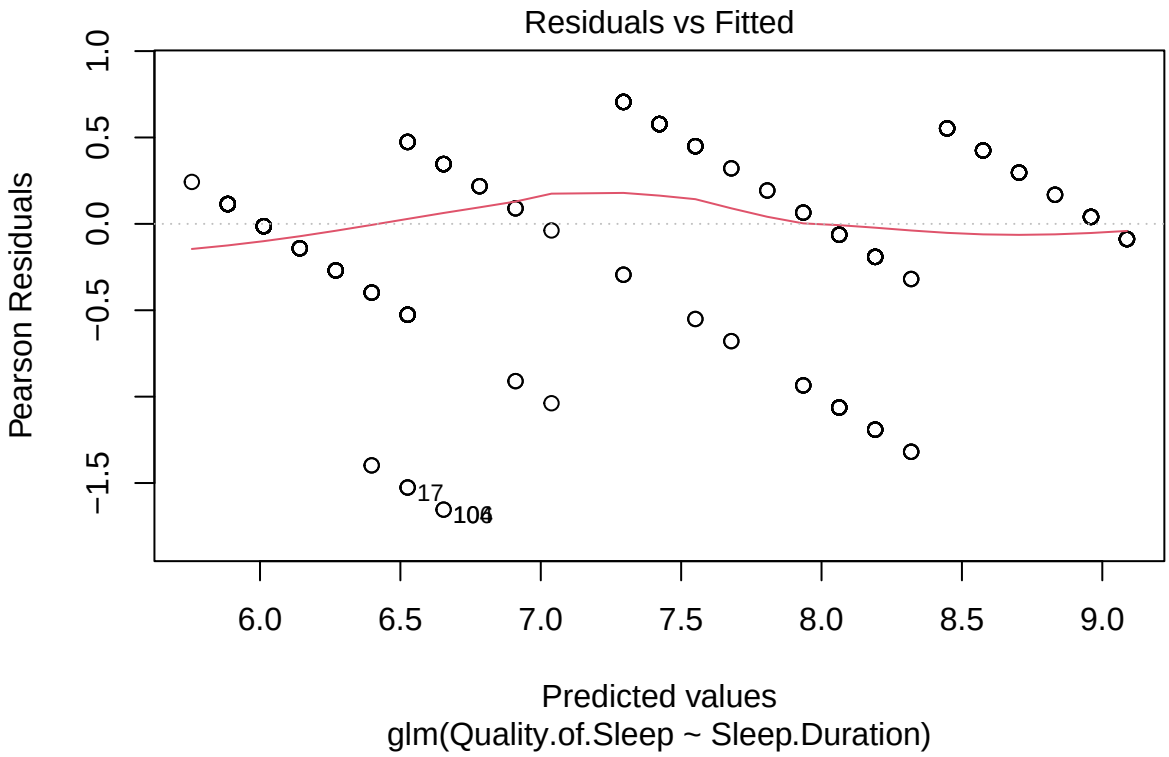
```r
m2 <- lmer(Quality.of.Sleep ~ Sleep.Duration + (1|Occupation), data=sleep)
summary(m2)
```

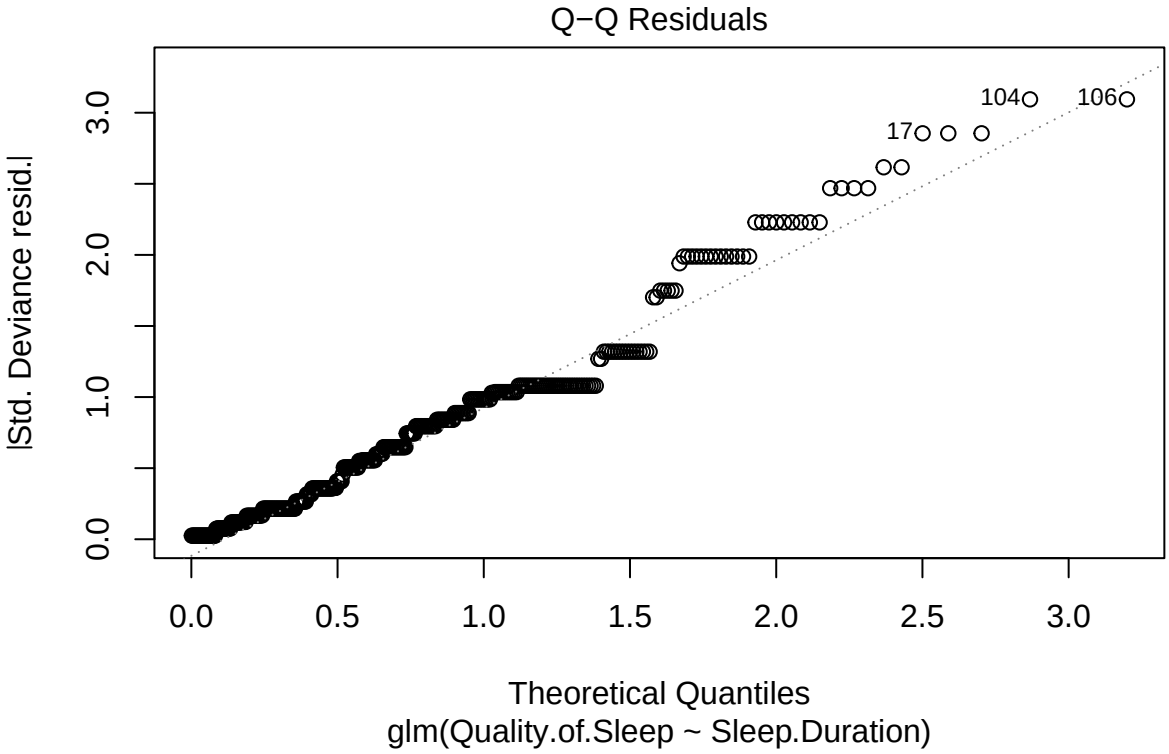
```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Quality.of.Sleep ~ Sleep.Duration + (1 | Occupation)
##    Data: sleep
##
## REML criterion at convergence: 438.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.3894 -0.4401  0.0525  0.6185  2.3162
##
## Random effects:
##  Groups     Name        Variance Std.Dev.
##  Occupation (Intercept) 0.1418   0.3766
##  Residual               0.1800   0.4242
## Number of obs: 363, groups:  Occupation, 7
##
## Fixed effects:
##                Estimate Std. Error t value
## (Intercept)    -1.22745    0.28487  -4.309
## Sleep.Duration  1.20434    0.03463  34.776
##
## Correlation of Fixed Effects:
##             (Intr)
## Sleep.Durtn -0.862
```

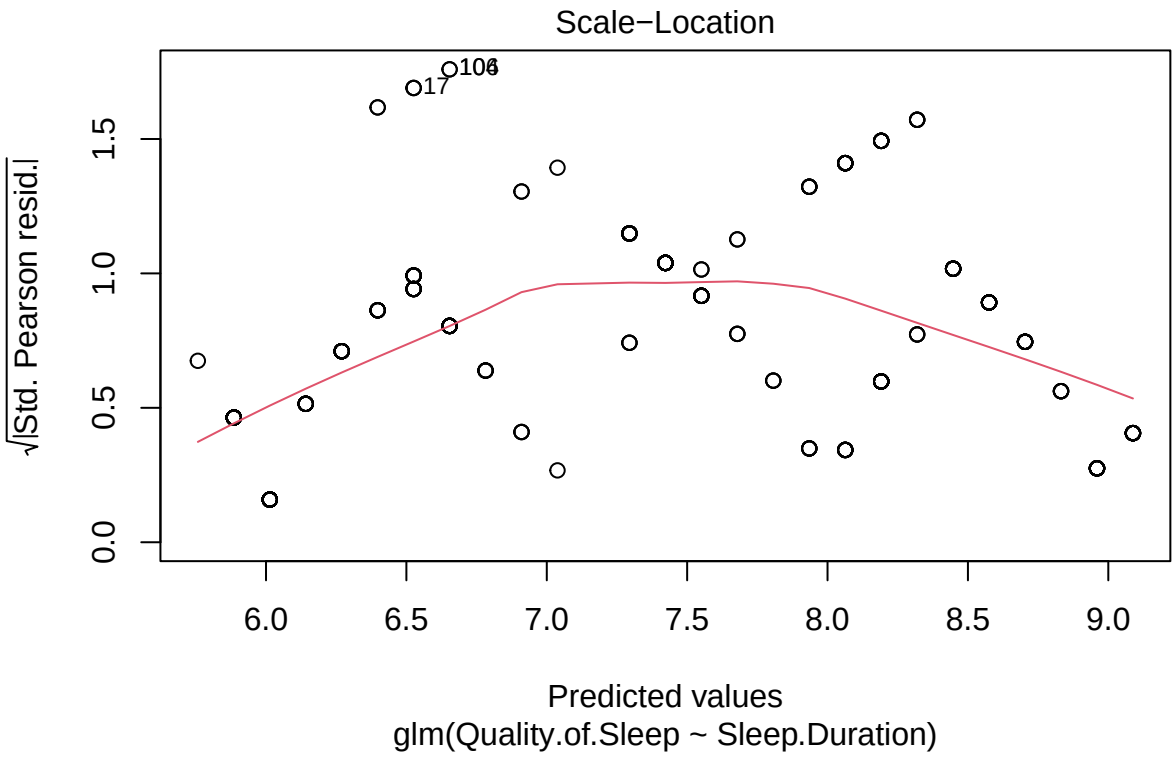
```r
lr2 <- glm(Quality.of.Sleep ~ Sleep.Duration, data = sleep)
summary(lr2)
```

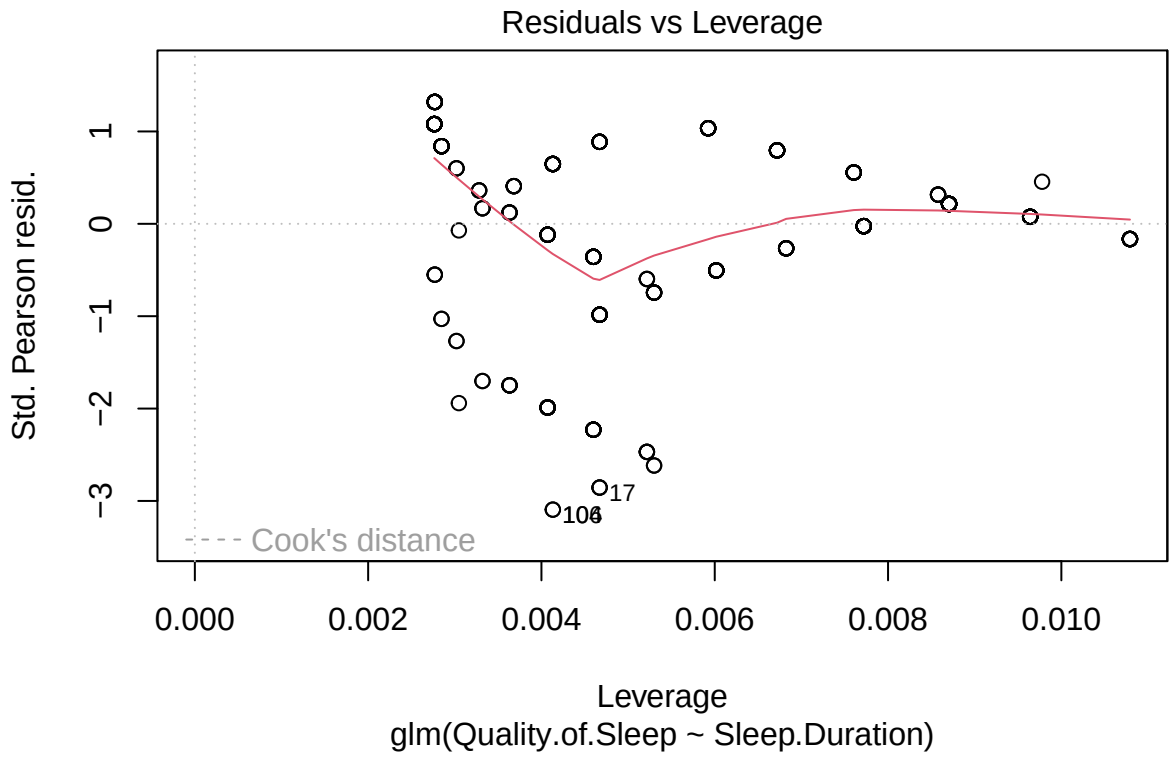
```
##
## Call:
## glm(formula = Quality.of.Sleep ~ Sleep.Duration, data = sleep)
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1.80048    0.25726  -6.999 1.26e-11 ***
## Sleep.Duration  1.28097    0.03573  35.848  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.2869088)
##
##     Null deviance: 472.27  on 362  degrees of freedom
## Residual deviance: 103.57  on 361  degrees of freedom
## AIC: 580.91
##
## Number of Fisher Scoring iterations: 2
```

```r
plot(lr2)
```


Residuals vs Fitted
Pearson Residuals
1.0
0.5
0.0
−0.5
−1.5
17
106
6.0
6.5
7.0
7.5
8.0
8.5
9.0
Predicted values
glm(Quality.of.Sleep ~ Sleep.Duration)


Q−Q Residuals
|Std. Deviance resid.|
0.0
1.0
2.0
3.0
104
106
17
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Theoretical Quantiles
glm(Quality.of.Sleep ~ Sleep.Duration)


Scale−Location
√|Std. Pearson resid.|
0.0
0.5
1.0
1.5
6.0
6.5
7.0
7.5
8.0
8.5
9.0
Predicted values
glm(Quality.of.Sleep ~ Sleep.Duration)




```
step.model <- step(lr_full, direction = "backward")
```

```
## Start:  AIC=-1101.2
## Quality.of.Sleep ~ Gender + Age + Occupation + Sleep.Duration +
##     Physical.Activity.Level + Stress.Level + BMI.Category + Heart.Rate +
##     Daily.Steps + Sleep.Disorder + Systolic.Pressure + Diastolic.Pressure
##
##                           Df Sum of Sq    RSS      AIC
## <none>                                  15.738 -1101.20
## - Physical.Activity.Level  1    0.0992 15.838 -1100.92
## - Systolic.Pressure        1    0.1347 15.873 -1100.11
## - Heart.Rate               1    0.1477 15.886 -1099.81
## - Diastolic.Pressure       1    0.1633 15.902 -1099.46
## - Daily.Steps              1    0.3057 16.044 -1096.22
## - BMI.Category             1    1.0590 16.797 -1079.56
## - Sleep.Disorder           2    1.1904 16.929 -1078.74
## - Sleep.Duration           1    1.7328 17.471 -1065.29
## - Gender                   1    2.9150 18.653 -1041.52
## - Age                      1    5.1629 20.901 -1000.22
## - Occupation               6    8.7876 24.526  -952.17
## - Stress.Level             1   17.1982 32.937  -835.13
```

```
summary(step.model)
```

```
##
```

```
## Call:
## lm(formula = Quality.of.Sleep ~ Gender + Age + Occupation + Sleep.Duration +
##     Physical.Activity.Level + Stress.Level + BMI.Category + Heart.Rate +
##     Daily.Steps + Sleep.Disorder + Systolic.Pressure + Diastolic.Pressure,
##     data = sleep)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.21920 -0.08037  0.00079  0.05887  0.78430
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 5.981e+00  8.365e-01   7.150 5.25e-12 ***
## GenderMale                  5.439e-01  6.814e-02   7.982 2.17e-14 ***
## Age                         5.301e-02  4.990e-03  10.623  < 2e-16 ***
## OccupationDoctor           -5.502e-01  7.836e-02  -7.022 1.18e-11 ***
## OccupationEngineer         -7.224e-01  7.763e-02  -9.306  < 2e-16 ***
## OccupationLawyer           -4.678e-01  9.166e-02  -5.104 5.51e-07 ***
## OccupationNurse            -4.195e-01  9.050e-02  -4.635 5.07e-06 ***
## OccupationSalesperson      -9.431e-01  8.827e-02 -10.685  < 2e-16 ***
## OccupationTeacher          -5.804e-01  7.276e-02  -7.977 2.25e-14 ***
## Sleep.Duration              2.816e-01  4.575e-02   6.154 2.10e-09 ***
## Physical.Activity.Level    -2.122e-03  1.441e-03  -1.473 0.141747
## Stress.Level               -3.957e-01  2.041e-02 -19.388  < 2e-16 ***
## BMI.CategoryOverweight     -4.135e-01  8.594e-02  -4.811 2.25e-06 ***
## Heart.Rate                 -1.156e-02  6.433e-03  -1.797 0.073205 .
## Daily.Steps                 5.140e-05  1.988e-05   2.585 0.010149 *
## Sleep.DisorderNone          1.979e-01  5.148e-02   3.843 0.000145 ***
## Sleep.DisorderSleep Apnea   2.723e-01  5.684e-02   4.791 2.47e-06 ***
## Systolic.Pressure           2.544e-02  1.483e-02   1.716 0.087141 .
## Diastolic.Pressure         -3.784e-02  2.003e-02  -1.889 0.059678 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2139 on 344 degrees of freedom
## Multiple R-squared:  0.9667, Adjusted R-squared:  0.9649
## F-statistic: 554.4 on 18 and 344 DF,  p-value: < 2.2e-16
```