# Examining the Association between Sleep and Lifestyle: The Roles of Habits, Occupation, and Sleep Disorder

BIOS 611 Fall 2023

Qinghua Li (730305170)

## Introduction

Sleep is essential for maintaining overall health and well-being. Despite its critical importance, it often remains undervalued. According to CDC, 1/3 of US adults report that they do not get the recommended amount of sleep[1]. Based on the recommendation of CDC, adults on average need at least 7 hours of sleep per night and sleep efficient should be at least 85%[2]. Sleep time and sleep could be affected by various factors, including exercise level, smoking status, caffeine consumption, alcohol consumption, occupation, etc[3,4,5,6].

In this study, we aim to explore the association of sleep efficiency (SE) with life habits, including smoking status, exercise frequency and level, stress level, and alcohol and caffeine consumption. Also, the association of total sleep time (TST) occupation, sleep disorder, BMI, and blood pressure are investigated

## Dataset Description

It is hard to find data including all the interested variables, therefore, I downloaded two datasets from Kaggle.

Dataset 1 includes 15 columns:

- ID: participant ID
- Age (y): in years, numerical variable
- Gender: male and female, categorical variable
- Bedtime: MM/DD/YYYY HOUR:MINUTE, the time to go to bed
- Wakeup time: MM/DD/YYYY HOUR:MINUTE, the time to wakeup
- Sleep duration (in hours: numerical variable
- Sleep efficiency (SE): SE = TST/total time in bed, in decimals, numerical variable. It was used as outcome in the analysis for this dataset. In marginal association analysis, I created a binary category for SE: ≥ 85% (normal SE) and < 85% (abnormal SE). In the linear model, numerical SE was used.
- REM sleep percentage (Rapid eye movement sleep percentage): %, numerical variable
- Deep sleep percentage: %, numerical variable
- Light sleep percentage: %, numerical variable
- Awakenings: number of awakenings during sleep, integer, numerical variable

- Caffeine consumption: the amount of caffeine consumed in the 24 hours prior to bedtime (in mg). I created a categorical variable with 0, 25, 50, ≥ 75 since the original data contains 0, 25, 50, 75, 100, 200 and very few observations in 75, 100, and 200 groups.
- Alcohol consumption: the amount of alcohol consumed in the 24 hours prior to bedtime (in oz), numerical variable
- Smoking status: yes or no, categorical variable
- Exercise frequency: the number of times subject exercise per week, ranging 0-5. I categorized it into 0, 1, 2, and ≥ 3 due to few cases with exercise frequency 4 and 5.

Dataset 1 can be found and downloaded from the following link on Kaggle:
https://www.kaggle.com/datasets/equilibriumm/sleep-efficiency/data

Dataset 2 includes 13 columns:

- ID: participant ID
- Gender: male or female, categorical variable
- Age (y): age of the participant
- Occupation: the occupation or profession of the participant, I categorized the occupation into Medical Field and non-Medical Field
- Sleep Duration (hours): the number of hours the person sleeps per day. It was used as the outcome in the analysis for this dataset. I created a binary category for marginal association analysis: ≥ 7 (normal sleep time for adults) and < 7 (abnormal sleep time).
- Quality of Sleep (scale 1-10): a subjective rating of the quality of sleep, ranging 0-10
- Physical Activity Level (minutes/day): the number of minutes the person engages in physical activity daily.
- Stress Level (scale 1-10): a subjective rating of the stress level experienced by the person, ranging 1-10.
- BMI Category: underweight, normal, overweight
- Blood Pressure (systolic/diastolic): blood pressure of the person, indicated as systolic pressure/diastolic pressure. I divided it into systolic blood pressure and diastolic blood pressure for the analysis.
- Heart Rate (bpm): the resting heart rate of the person in beats per minutes.
- Daily Steps: number of steps taken daily.
- Sleep Disorder: presence or absence of a sleep disorder, including None, Insomnia, and Sleep Apnea.

Dataset 2 can be found and downloaded from the following link on Kaggle:
https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset

## Methods, Results & Discussion

Dockerfile and R were used. Mean and standard deviation (SD) were used for numerical variables and n (%) for categorical variables.

In dataset 1, SE was used as outcome. In dataset 2, sleep duration was used as outcome. In marginal association analysis, SE was divided into 2 categories: ≥ 85% (normal SE) and < 85% (abnormal SE), and TST was divided into 2 categories: ≥ 7 (normal sleep time for adults) and < 7 (abnormal sleep time). Student t-test was used for numerical variables and Chi-square for categorical variables. In linear regression model, numerical SE and TST were used. Coefficients and P-values were explored for the linear regression model and residual plots were used for model performance evaluation.

### Results using Dataset 1

Table 1 showed the characteristics and marginal association between the variables and categorical SE using dataset 1. Deep sleep percentage, light sleep percentage, number of awakenings during sleep, alcohol consumption, and weekly exercise frequency are significantly associated with SE (Table 1). The results show that participants with SE ≥ 85% generally had lower alcohol consumption, more deep sleep, less light sleep, and exercise more.

**Table 1: Participants' Characteristics and Their Marginal Association with Sleep Efficiency**

| | Sleep Efficiency < 85% (N=251) | Sleep Efficiency > 85% (N=201) | P-value |
|---|---|---|---|
| Age | 39.7 (14.1) | 41.0 (11.9) | 0.277 |
| Male | 131 (52.2%) | 97 (48.3%) | 0.462 |
| * REM Sleep % | 22.4 (3.55) | 22.8 (3.50) | 0.233 |
| Deep Sleep % | 46.2 (17.9) | 61.1 (5.18) | < 0.001 |
| Light sleep % | 31.4 (17.5) | 16.0 (3.45) | < 0.001 |
| Number of Awakenings | 2.54 (1.12) | 0.503 (0.501) | < 0.001 |
| Caffeine consumption (mg) | | | 0.038 |
| *0* | 122 (48.6%) | 89 (44.3%) | |
| *25* | 45 (17.9%) | 34 (16.9%) | |
| *50* | 60 (23.9%) | 47 (23.4%) | |
| *≥ 75* | 9 (3.6%) | 21 (10.4%) | |
| Alcohol consumption (oz) | 1.53 (1.75) | 0.723 (1.32) | < 0.001 |
| ^ Smoking status - Yes | 93 (37.1%) | 61 (30.3%) | 0.163 |
| Weekly exercise frequency | | | |
| *0* | 82 (32.7%) | 34 (16.9%) | < 0.001 |
| *1* | 59 (23.5%) | 38 (18.9%) | |
| *2* | 29 (11.6%) | 25 (12.4%) | |
| *≥ 3* | 76 (30.3%) | 103 (51.2%) | |

* REM sleep: rapid eye movement sleep
^ There are two smoking statuses in this data: Yes and No.
Mean (SD) and t-test was used for numerical variables, and n (%) and Chi-square for categorical variables.

Figure 1 showed participants in SE < 85% group had higher alcohol consumption and more awakenings at night than SE ≥ 85% group. Combine the results in Table 1, alcohol consumption and number of awakenings at night have negative association with SE.

Figure 2 showed that deep sleep percentage and light sleep percentage were highly negatively correlated. Light sleep percentage was not included in the linear regression model.

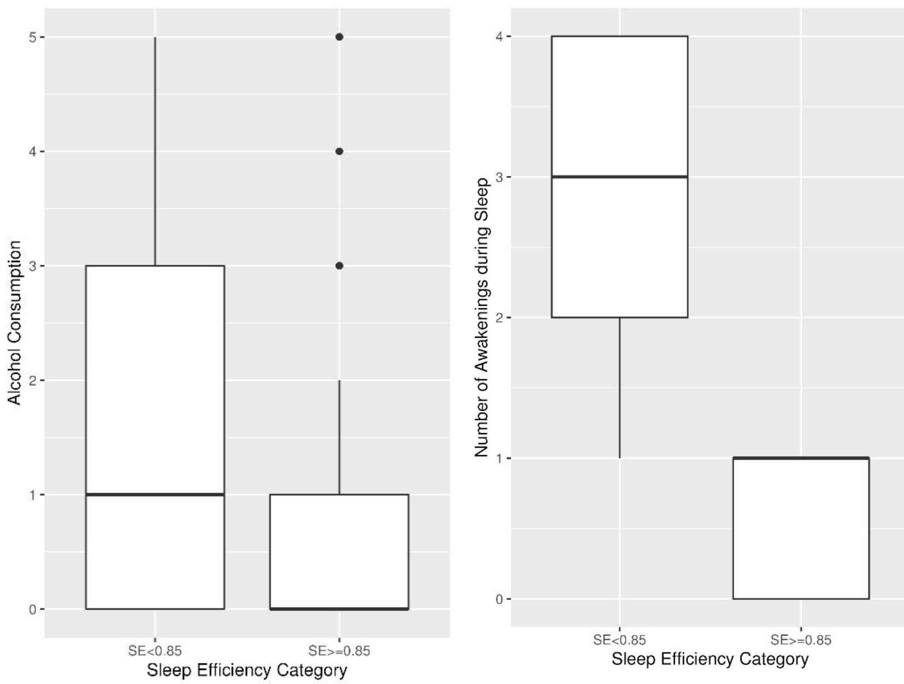


Figure 1: Alcohol consumption (left) and number of awakenings (right) in different SE groups: participants in SE ≥ 85% group had less alcohol consumption and number of awakenings compared with SE < 85% group.

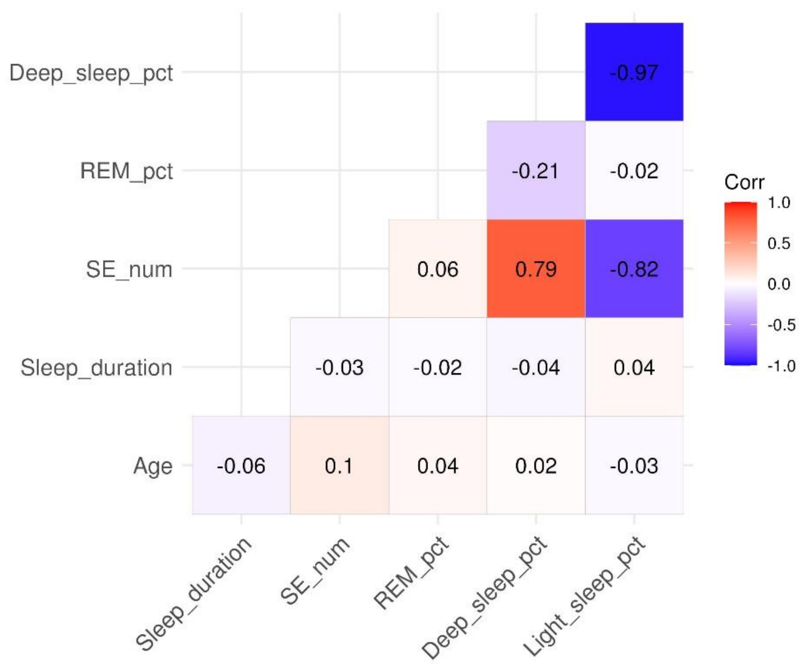


Figure 2: Correlation using dataset 1: Deep sleep percentage and light sleep percentage are highly negatively correlated.

Linear model: formula = Numeric SE ~ Gender + Age + REM Sleep % + Deep sleep % + Awake nings + Smoking status + Exercise frequency + Caffeine consumption + Alcohol consumption.

Age, REM sleep %, Deep sleep %, Awakenings at night, Smoking Status, and Caffeine and Alco hol consumption are significantly associated with SE (Table 2). Compared with non-smokers, par ticipants who smoke have worse sleep efficiency indicated by negative coefficient. Alcohol cons umption is negatively correlated with SE. Age, REM sleep %, and Deep sleep % are positively c orrelated with sleep efficiency. Table 2 showed that more caffeine consumption seemed to be ben eficial to sleep. However, Figure 3 showed that as individuals age, there tends to be a decrease in caffeine consumption. Age is positively correlated with SE. This suggests that the effect of caffei ne consumption on sleep may be moderated by age.

**Table 2: Linear model summary for Dataset 1**

| | Coefficient | Std. Error | t value | P-value |
|---|---|---|---|---|
| Intercept | 0.355 | 0.030 | 11.788 | < 0.0001 |
| Gender: Male [a] | -0.002 | 0.007 | -0.258 | 0.7964 |
| Age | 0.001 | 0.0002 | 3.725 | < 0.0001 |
| REM Sleep % | 0.007 | 0.001 | 7.436 | < 0.0001 |
| Deep Sleep % | 0.006 | 0.0002 | 23.110 | < 0.0001 |
| Awakenings | -0.032 | 0.0025 | -12.877 | < 0.0001 |
| Smoking Status: Yes [b] | -0.045 | 0.0069 | -6.529 | < 0.0001 |
| Exercise Frequency [c] | | | | |
| 1 | -0.0140 | 0.001 | -1.402 | 0.1618 |
| 2 | 0.0223 | 0.0117 | 1.907 | 0.0572 |
| ≥3 | 0.0149 | 0.0082 | 1.816 | 0.0702 |
| Caffeine Consumption | | | | |
| 25 | 0.00001 | 0.001 | 0 | 0.9996 |
| 50 | 0.003 | 0.001 | 3.003 | 0.0029 |
| ≥ 75 | 0.0015 | 0.0015 | 1.012 | 0.3124 |
| Alcohol Consumption | -0.0062 | 0.0021 | -2.927 | 0.0036 |

Adjusted $R^2$: 0.8024.
a female is the reference group;
b No smoking is the reference group
c Exercise frequency 0 is the reference group

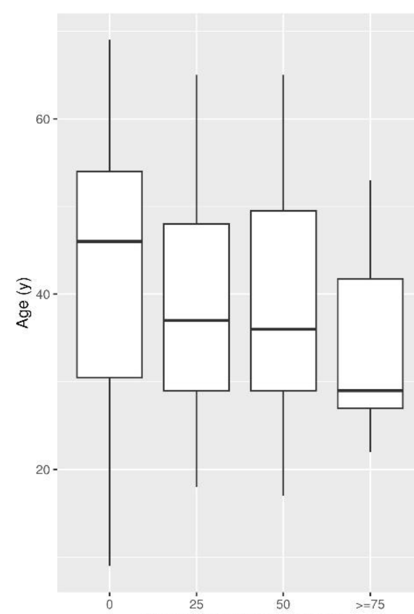


Figure 3: Age distribution among different caffeine consumption group: as individuals age, there tends to be a decrease in caffeine consumption.

**Results using Dataset 2**

Table 3 showed the characteristics and their marginal association with binary TST using dataset 2. Physical activity level, stress level, BMI category, blood pressure, and sleep disorder are strongly associated with TST. Age, Occupation, and Daily steps are significantly associated with TST. Take Sleep Disorder for example, participants with TST ≥ 7 hours per night had higher percentage of None sleep disorder and much lower percentage of sleep disorder compared with those who did not sleep enough.

**Table 3: Participants' Characteristics and Their Marginal Association with Total Sleep Time (TST) in hours**

| | TST < 7 (N=155) | TST ≥ 7 (N=219) | P-value |
|---|---|---|---|
| Male | 77 (49.7%) | 112 (51.1%) | 0.862 |
| Age | 41.3 (7.47) | 42.8 (9.40) | 0.077 |
| Occupation | | | 0.0341 |
| *Medical Field* | 70 (45.2%) | 74 (33.8%) | |
| *Non-Medical Field* | 85 (54.8%) | 145 (66.2%) | |
| Physical Activity Level | 50.5 (21.3) | 65.3 (18.2) | < 0.001 |
| Stress Level | 6.88 (1.42) | 4.32 (1.11) | < 0.001 |
| BMI Category | | | < 0.001 |
| *Normal* | 39 (25.2%) | 177 (80.8%) | |
| *Obese* | 4 (2.6%) | 6 (2.7%) | |
| *Overweight* | 112 (72.3%) | 36 (16.4%) | |
| Systolic Blood Pressure (mmHg) | 132 (5.89) | 126 (8.12) | < 0.001 |
| Diastolic Blood Pressure (mmHg) | 86.9 (5.31) | 83.0 (6.22) | < 0.001 |
| Heart Rate (bpm) | 72.3 (4.34) | 68.7 (3.24) | < 0.001 |
| Daily Steps | 6570 (1990) | 6990 (1270) | 0.0236 |
| Sleep Disorder | | | < 0.001 |
| *Insomnia* | 69 (44.5%) | 8 (3.7%) | |
| *None* | 46 (29.7%) | 173 (79.0%) | |
| *Sleep Apnea* | 40 (25.8%) | 38 (17.4%) | |

Mean (SD) and t-test was used for numerical variables, and n (%) and Chi-square for categorical variables.

Similarly, Pearson correlation was explored for numeric variables in Dataset 2 (Figure 4). Quality of sleep is highly negatively correlated with stress level (-0.9), and systolic blood pressure is highly positively correlated with diastolic blood pressure (0.97). Therefore, quality of sleep and diastolic blood pressure were not included in the linear regression model.

Linear model: formula = Numeric TST ~ Gender + Age + Occupation + Physical Activity level + Stress level + BMI category + SBP + HR + Daily steps + Sleep disorder.

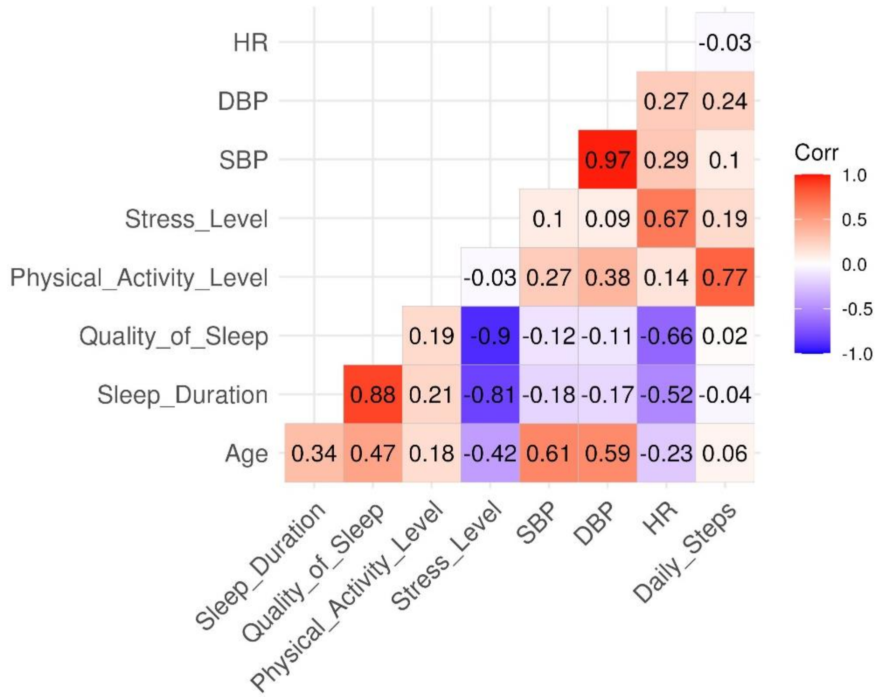


Figure 4: Correlation plot for numeric variables in Dataset 2.

**Table 4: Linear model summary for Dataset 2**

| | Coefficient | Std. Error | t value | P-value |
|---|---|---|---|---|
| Intercept | 8.08 | 0.80 | 10.160 | < 0.0001 |
| Gender: Male [a] | 0.051 | 0.045 | 11.223 | < 0.0001 |
| Age | 0.044 | 0.0003 | 13.456 | < 0.0001 |
| Occupation: Non-Medical Field [b] | -0.247 | 0.004 | -6.161 | < 0.0001 |
| Physical Activity Level | 0.007 | 0.002 | 4.038 | < 0.0001 |
| Stress Level | -0.350 | 0.023 | -15.349 | < 0.0001 |
| BMI | | | | |
| Overweight | -0.509 | 0.062 | -8.233 | < 0.0001 |
| Obese | -0.638 | 0.173 | -3.691 | 0.0003 |
| Systolic Blood Pressure | -0.022 | 0.004 | -5.627 | < 0.0001 |
| Heart Rate | 0.028 | 0.010 | 2.874 | 0.0043 |
| Daily Steps | -0.0001 | 0.00002 | -2.445 | 0.0149 |
| Sleep Disorder [c] | | | | |
| Sleep Apnea | 0.059 | 0.069 | 0.853 | 0.394 |
| Insomnia | -0.019 | 0.059 | -0.328 | 0.743 |

Adjusted $R^2$: 0.8674.
a female is the reference group
b Medical field is the reference group
c None is the reference group

Table 4summarized the results for the linear regression model using Dataset 2. Age, male (compared with female), physical activity level, and heart rate have positive coefficients with P-value < 0.05, indicating positive association with TST. Take age for example, holding other variables constant, for each additional year of age, TST increases by 0.044 hours. Non-medical field participants have longer TST compared with medical field workers indicated by negative coefficients in the linear model. Upon examining the coefficients, it was observed that participants with sleep apnea had longer TST compared with those without any sleep disorder. However, this finding is challenging to interpret when solely considering the coefficients. Figure 5 showed that participants with sleep disorders (sleep apnea and insomnia) tend to have higher stress level and SBP. This suggests that the impact of sleep disorder on TST may be moderated by stress level and SBP.

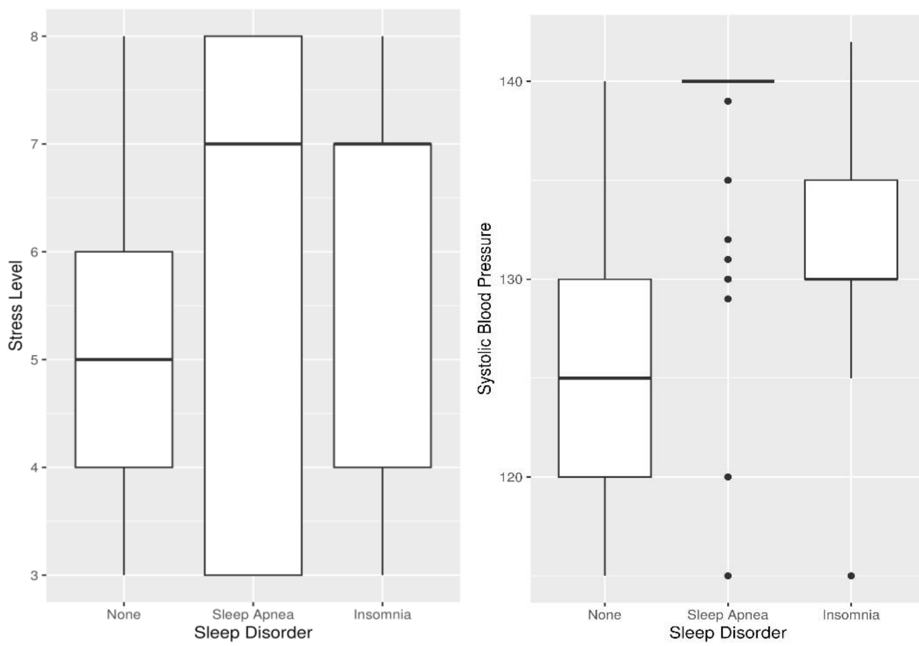


Figure 5: Stress level and systolic blood pressure (SBP) among different sleep disorder groups: participants with sleep disorders tend to have higher stress level and SBP.

## Conclusion

Various factors are associated with sleep efficiency and total sleep time, including alcohol and caffeine consumption, exercise frequency and activity level, blood pressure and heart rate, sleep disorders and stress level.

Engaging in regular exercise and moderating alcohol and caffeine consumption could potentially improve sleep quality.

## Limitation and Future Study

There are some limitations in this study. First of all, the interested variables with TST and SE are from different data. It would be great if these interested variables were from the same dataset. Second, it would be interesting to explore the association of screen time, education degree, and personality with SE and TST. This could be included in the future study. Third, the residual scatter plot and QQ plot show that the residuals deviate from a normal distribution (Figure 5), and hence, the assumption that residuals follow N $(0, \sigma^2)$ is not met. Consequently, alternative modeling approaches, such as logistic regression using binary categorical SE and TST and classification tree models, could be considered.

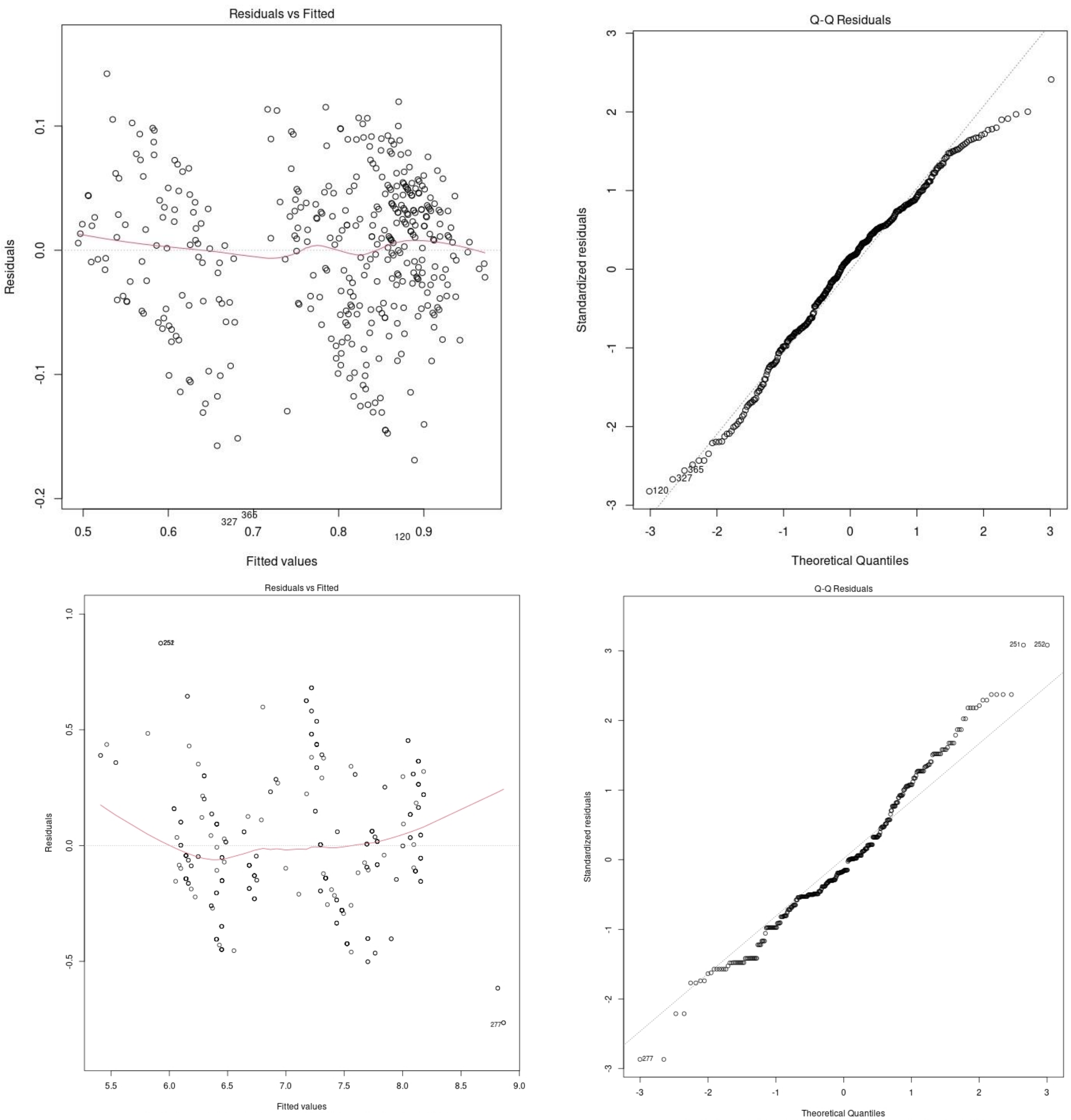


Figure 6: Residual scatter plot and residual QQ plot for the linear models using dataset 1 (top panel) and dataset 2 (bottom panel).

## References


1. https://www.cdc.gov/sleep/index.html

2. https://www.cdc.gov/sleep/features/getting-enough-sleep.html#:~:text=Habits%20to%20Improve%20Your%20Sleep,and%20phones%20from%20the%20bedroom.

3. https://www.sleepfoundation.org/nutrition/alcohol-and-sleep

4. https://www.sleepfoundation.org/nutrition/caffeine-and-sleep#:~:text=Caffeine%20consumption%20can%20make%20you,feeling%20refreshed%20the%20next%20day.

5. https://www.sleepfoundation.org/physical-activity/exercise-and-sleep

6. Wetter DW, Young TB. The relation between cigarette smoking and sleep disturbance. Prev Med. 1994 May;23(3):328-34. doi: 10.1006/pmed.1994.1046. PMID: 8078854.


```
Call:
lm(formula = Sleep_Duration ~ Gender + Age + Occupation + Physical_Activity_Level +
    Stress_Level + BMI_Cat + SBP + DBP + HR + Daily_Steps + Sleep_Disorder,
    data = slp_life_df)

Residuals:
```

```
     Min       1Q   Median       3Q      Max 
-0.70052 -0.17500 -0.04997  0.16827  0.73945 

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                8.687e+00  7.737e-01  11.229  < 2e-16 ***
GenderMale                 5.651e-01  4.484e-02  12.603  < 2e-16 ***
Age                        5.586e-02  3.807e-03  14.673  < 2e-16 ***
OccupationNon-Medical Field -9.826e-02  4.727e-02  -2.079 0.038362 *  
Physical_Activity_Level    5.922e-03  1.662e-03   3.563 0.000416 ***
Stress_Level              -2.904e-01  2.450e-02 -11.852  < 2e-16 ***
BMI_CatObese              -6.056e-01  1.665e-01  -3.638 0.000315 ***
BMI_CatOverweight         -7.683e-01  7.623e-02 -10.079  < 2e-16 ***
SBP                       -1.134e-01  1.719e-02  -6.598 1.49e-10 ***
DBP                        1.308e-01  2.403e-02   5.443 9.69e-08 ***
HR                         2.408e-02  9.432e-03   2.554 0.011076 *  
Daily_Steps               -1.131e-04  2.392e-05  -4.729 3.25e-06 ***
Sleep_DisorderSleep Apnea -4.133e-03  6.738e-02  -0.061 0.951124    
Sleep_DisorderInsomnia    -5.467e-02  5.749e-02  -0.951 0.342275    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2789 on 360 degrees of freedom
Multiple R-squared:  0.8814,	Adjusted R-squared:  0.8772 
F-statistic: 205.9 on 13 and 360 DF,  p-value: < 2.2e-16
```