Sleep Efficiency Statistical Inference and Prediction Final Report

Group A5: Daniel Choi, Clare Pan, Thê Quach, Qi Xu

STAT 306
August 11, 2023

Introduction

Sleep is a crucial aspect of your everyday routine, and getting quality sleep is just as essential to survival as food and water (NINDS, 2023). It is a physiological process that allows your brain and body to recover from the stresses of the day (NIH, 2022). Being sleep deprived can stop you from establishing or maintaining the neural pathways in your brain that allow you to learn and create new memories, as well as causing difficulties to concentrate and respond quickly (NINDS, 2023).

We will be using the Sleep Efficiency Dataset from Kaggle (Equilibriumm, 2023) and explore the quality of sleep among individuals and the potential factors influencing sleep. The following are the variables that are measured in the dataset:

- ID: a unique identifier for each test subject
- Age: age of the test subject (in years)
- Gender: male or female
- Bedtime: the time the test subject does to bed each night (in YYYY-MM-DD HH:mm:ss)
- Wakeup time: the time the test subject (in YYYY-MM-DD HH:mm:ss)
- Sleep duration: the total amount of time the test subject slept (in hours)
- Sleep efficiency: a measure of the proportion of time in bed spent asleep (total sleep time/time in bed*100)
- REM sleep percentage: the percentage of total sleep time spent in REM sleep
- Deep sleep percentage: the percentage of total sleep time spent in deep sleep
- Light sleep percentage: the percentage of total sleep time spent in light sleep
- Awakenings: the number of times the test subject wakes up during the night
- Caffeine consumption: the amount of caffeine consumed in the 24 hours prior to bedtime (in mg)
- Alcohol consumption: the amount of alcohol consumed in the 24 hours prior to bedtime (in oz)
- Smoking status: whether or not the test subject smokes
- Exercise frequency: the number of times the test subject exercises each week

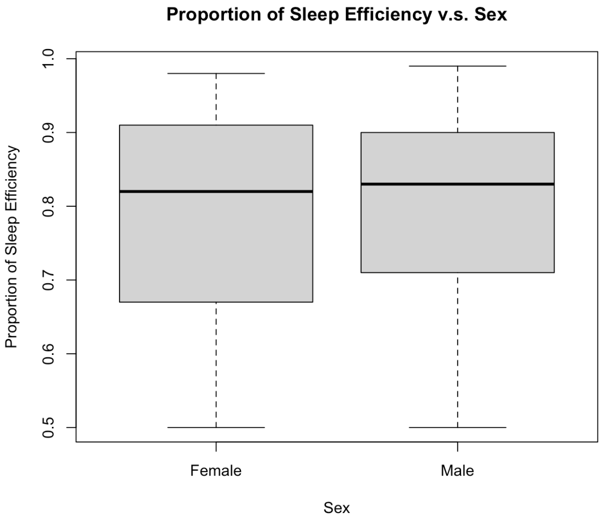
Our group's motivation is that we want to identify factors related to sleep quality and improve the quality of people's sleep so that they can live a healthier life. We plan on doing so by fitting a linear model by the Sleep Efficiency Dataset to find critical factors and patterns that will allow us to find a convincing solution.

Analysis

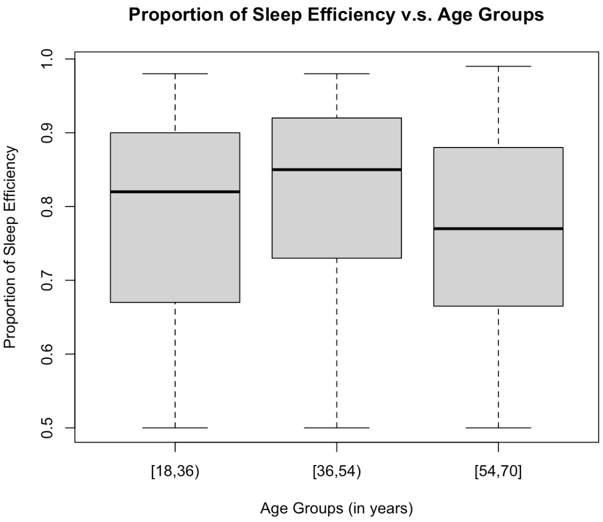
After cleaning the data we removed the id, bedtime, wake time, awakenings variables as they were unrelated to fitting our model. When choosing a linear model to fit, we decided that some of the variables may be correlated with sleep efficiency such as percent of time spent in REM sleep, percent of time spent in light sleep, percent of time spent in deep sleep, and sleep duration. So we did not include these variables in our linear model.

We used boxplots to check the relationships between sleep efficiency and all the other variables. Females and males in the sample have a similar proportion of sleep efficiency at around 0.825 according to the boxplot (Figure 1). We decided to split the people into age groups of young adults (18-36 years old), adults (36- 54 years old), and seniors (54+ years old) when creating the boxplot. Adults have a higher mean proportion of sleep efficiency, whereas seniors have a lower mean (Figure 2). In the Proportion of Sleep Efficiency v.s. Caffeine Consumed in 24 Hours boxplot, it shows that consuming 75mg to 200mg can improve the mean proportion of sleep efficiency from ~0.82 to ~0.9 or higher, compared to the proportion of sleep efficiency if 0mg to 50mg caffeine is consumed (Figure 3). In terms of alcohol consumption 24 hours before sleep, having 0 to 1 and 3 ounces of alcohol result in an average proportion of sleep efficiency ranges from 0.78 to 0.86, whereas 2, 4 to 5 ounces of alcohol can significantly decrease sleep efficiency to a mean proportion of 0.6 to 0.7 (Figure 4). People in the sample who do not smoke get a higher sleep efficiency proportion compared to the people who do smoke, and the middle 50% of the IQR of people who smokes is a lot wider and lean towards low sleep efficiency, suggesting that it has a larger standard deviation (Figure 5). The boxplot that compares the proportion of sleep efficiency and exercise sessions each week shows that an increase in exercise sessions generally leads to a higher proportion of sleep efficiency (Figure 6).

(Figure 1)

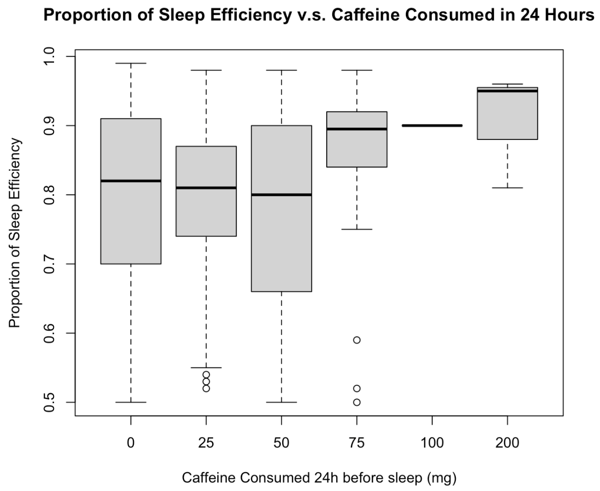


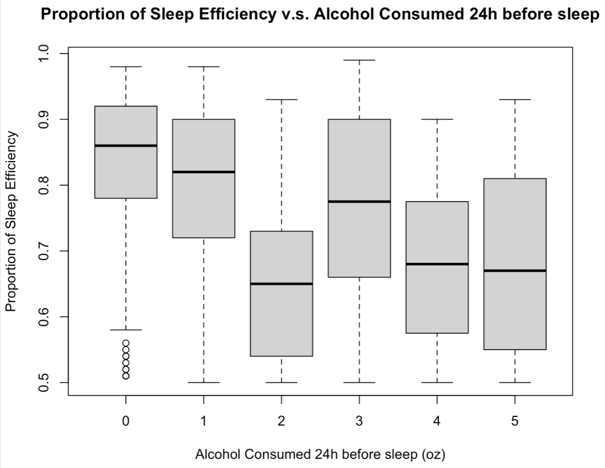
(Figure 2)



(Figure 3)

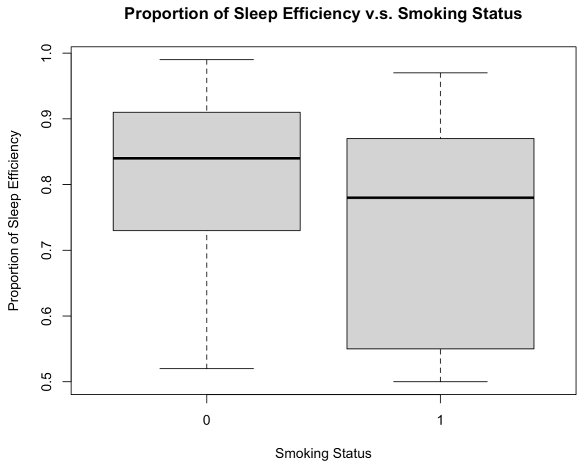
(Figure 4)

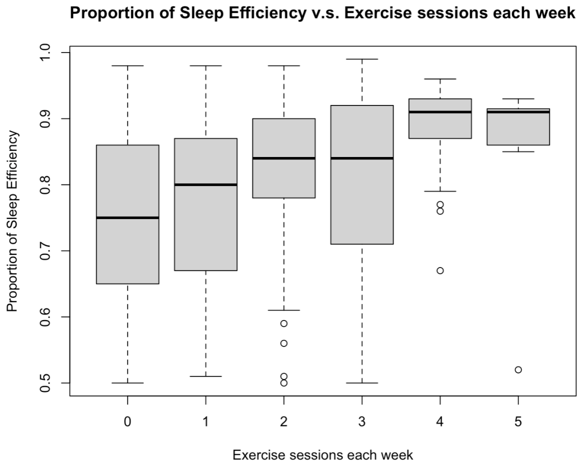




(Figure 5)

(Figure 6)

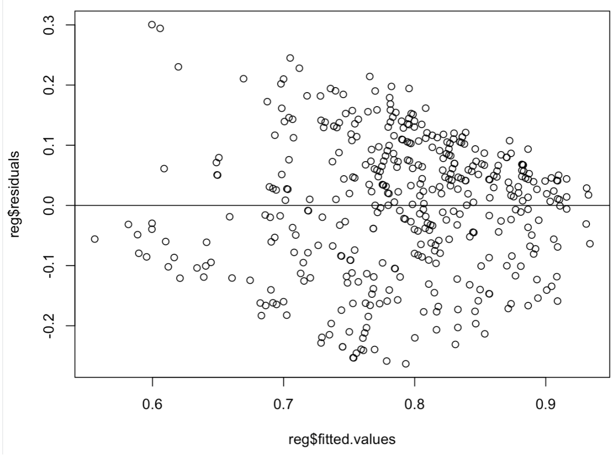




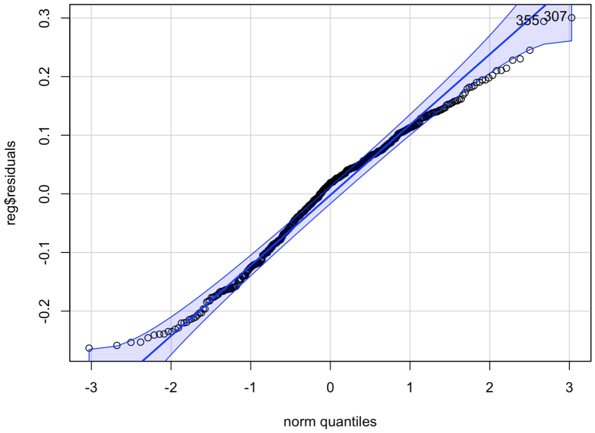
For our analysis, we used sleep efficiency as our response variable, and age, sex, alcohol consumption, caffeine consumption, smoker status, and exercise frequency as our predictor variable to fit our models.

We then created a plot for the fitted values v.s. residual for the linear model to check whether the regression assumptions are met (Figure 7). Figure 7 shows a diamond-shaped pattern between the fitted and the residual values, which suggests that heteroscedasticity exists in the model. The Q-Q plot (Figure 8) of the residual has a heavy-tailed and skewed to the right characteristics. The said characteristics about the model violates the assumptions for simple linear regression, because the residuals do not follow a normal distribution. We also tried fitting the model with the log of response variable, but it did not make any meaningful changes to the plot. Further studies should be made to investigate the model, but for the purpose of the project goal, we will continue to use this model and make analysis.

(Figure 7)



(Figure 8)



When testing at 0.05 significance level, we conclude that the Sex and Caffeine consumed variables are significant to our model, as the p-value > 0.05. Therefore we cannot reject the null hypothesis for these variables that $\beta_{Caffeine} = 0$ and $\beta_{Sex\,(Male)} = 0$.

Our linear model is:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon ,$$

where $x_1$is age, x_2 is exercise frequency, x_3 is alcohol consumption, and x_4 is smoker status (1 for smoker, 0 for non-smoker).

Through the best subset selection algorithm, we arrived at the same model with the same variables; age, alcohol consumption, smoking status, and exercise frequency. We determined 4 variables was most appropriate model as adding a variable past 4 does not dramatically change the $adjR^2$ value.

When fitting this final model in R, we get the following values for each β_i:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4$$

b_0 = 0.759, b_1 = 0.00125, b_2 = 0.0233, b_3 = -0.0318, b_4 = -0.0725

We can interpret these values as follows:

- For every year increase in age, holding all other variables equal, the predicted sleep efficiency proportion increases by 0.00125.
- For every workout day a person participates in, holding all other variables equal, the predicted sleep efficiency proportion increases by 0.0233.
- For every additional ounce of alcohol consumed 24 hours before, holding other variables equal, the predicted sleep efficiency decreases by 0.0318.
- For smokers the intercept of the line is 0.0725 less than for non-smokers.

The R^2 for our model is 0.3012, which suggests that our model explains ~30% of the data, giving us the goodness of fit.

95% confidence intervals for each coefficient estimate:

b_1: (0.000410, 00209)

b_2: (0.0156, 0.0310)

b_3: (-0.0387, -0.0248)

b_4: (-0.0960, -0.0490)

Conclusion

Based on our model we determined that age, exercise frequency, alcohol consumption, and smoking status have an influence on sleep efficiency. Our model is appropriate because the p-values of our coefficient of predictor variables are very small, and the predictors explain ~30% of the variance in the response variable. Based on the residual plot and the Q-Q plot however, it is harder to judge if our model is appropriate since the residuals does not follow a normal distribution and the Q-Q plot shows that heteroscedasticity exists in the model.

It is possible that there are interactions between the variables that affect sleep efficiency, adding these interaction terms could help to improve the pattern in the residual plots which will improve the model's reliability. Alternatively, transforming the data in some way could help to alleviate the heteroscedasticity.

References

Equilibriumm. (2023). *Sleep efficiency dataset*. Kaggle. https://www.kaggle.com/datasets/equilibriumm/sleep-efficiency

U.S. Department of Health and Human Services (NINDS). (2023). *Brain basics: Understanding sleep*. National Institute of Neurological Disorders and Stroke. https://www.ninds.nih.gov/health-information/public-education/brain-basics/brain-basics-understanding-sleep

U.S. National Library of Medicine (NIH). (2022). *Sleep and your health: Medlineplus medical encyclopedia*. MedlinePlus. https://medlineplus.gov/ency/patientinstructions/000871.htm