Exploration into Factors Affecting Sleep Quality

Frank Cui
Steve He
Zihan Wang
Zelalem Araya

Introduction

Sleep has been proven to affect many aspects of our well-being, having impacts on physical health, mental health, and general personal growth. (Hamilton, 2006). Actions while awake can have negative or positive impacts on our sleep, and thus have negative or positive impacts on our lives as a whole. This dataset includes variables that document the quality of sleep of their subjects and the various actions that may impact their sleep, including exercise per week and whether a subject is a smoker.

This research explores the relationship between a person's sleep quality and their lifestyle, notably habits such as exercise, alcohol intake, and smoking. The adverse effects of many of these habits on sleep are well-documented: cigarette smokers display worse sleep quality than non-smokers, and the same is seen in avid drinkers. (Ledger et. al, 2022). In contrast, exercise has been widely reported to improve sleep quality (Driver & Taylor, 2000). However, there is a noticeable lack of research on the interaction between these variables. Is an individual who is a frequent smoker, but exercises more than average, able to offset the negative impacts of smoking on their sleep quality? Do those who drink and smoke experience worse sleep quality than those who only do one or the other?

Our report uses a series of statistical analyses to explore the relationships between sleep quality, measured by sleep efficiency, and other lifestyle factors to determine what factors are most correlated to sleep efficiency. We will be defining "most correlated" as factors whose relationships to sleep efficiency has a statistical significance of below 0.05.

The dataset used was collected as part of a study conducted by ENIAS in Morocco, and contains information on 452 individuals, with 15 variables on each individual's sleep patterns and lifestyle, sourced from Kaggle. The variables are described as follows:

1. **ID:** The identifier for each subject in this study, no specific meaning
2. **Age:** the age of the subject, in years
3. **Gender:** male or female
4. **Bedtime:** the time that the subject goes to bed
5. **Wakeup time:** the time that the subject wakes up
6. **Sleep duration:** the amount of time (in hours) that the subject spends sleeping
7. **Sleep efficiency:** a measure of the proportion of time in bed spent asleep (the range is from 0 to 1, with 1 being the highest sleep efficiency)
8. **REM sleep percentage:** the percentage of time the subject spent in REM (rapid eye movement) sleep
9. **Deep sleep percentage:** the percentage of time the subject spent in deep sleep
10. **Light sleep percentage:** the percentage of time the subject spent in light sleep. The sum of the previous three values for one subject should be 100.
11. **Awakenings:** the number of times the subject wakes up during the night
12. **Caffeine consumption:** the amount of caffeine consumed in the 24 hours prior to bedtime (in mg)
13. **Alcohol consumption:** the amount of alcohol consumed in the 24 hours prior to bedtime (in oz)
14. **Smoking status:** whether or not the subject smokes
15. **Exercise frequency:** the number of times the subject exercises each week

We removed Bedtime and Wakeup time from our analysis due to their impacts being summarized in Sleep duration, along with ID since it provides no information to our analysis. The rest of the variables being used will be determined through our preliminary analysis.

Exploratory Data Analysis

We will first investigate the statistics for some categorical variables. There are 452 subjects in this study. Among all subjects, 228 of them are male and 224 of them are female. 154 out of the 452 subjects smoke and 206 subjects consume alcohol in the 24 hours prior to bedtime.

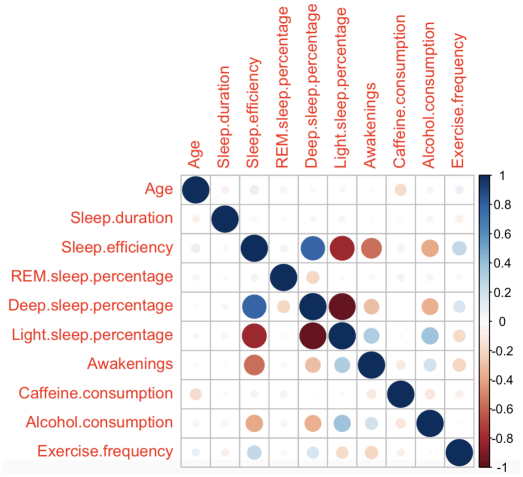

Fig 1: Correlation Matrix for continuous variables

We start the analysis by creating the correlation matrix for all the continuous variables in the dataset so that the correlation between different variables can be visualized. A strong negative correlation between deep sleep percentage and light sleep percentage is expected since deep sleep percentage is calculated using light sleep percentage and REM sleep percentage. There is a positive correlation between deep sleep percentage and sleep efficiency and a negative correlation between light sleep percentage and sleep efficiency. The potential reason why there is no strong correlation between REM sleep percentage and sleep efficiency is that their correlation might be masked by deep sleep percentage and light sleep percentage due to collinearity.

By looking at the plot, we notice that there is a negative correlation between alcohol consumption and sleep efficiency, as well as alcohol consumption and deep sleep percentage. On the other hand, exercise frequency is positively correlated with sleep efficiency and deep sleep percentage. Hence, alcohol consumption and exercise frequency will be variables that we will investigate later.

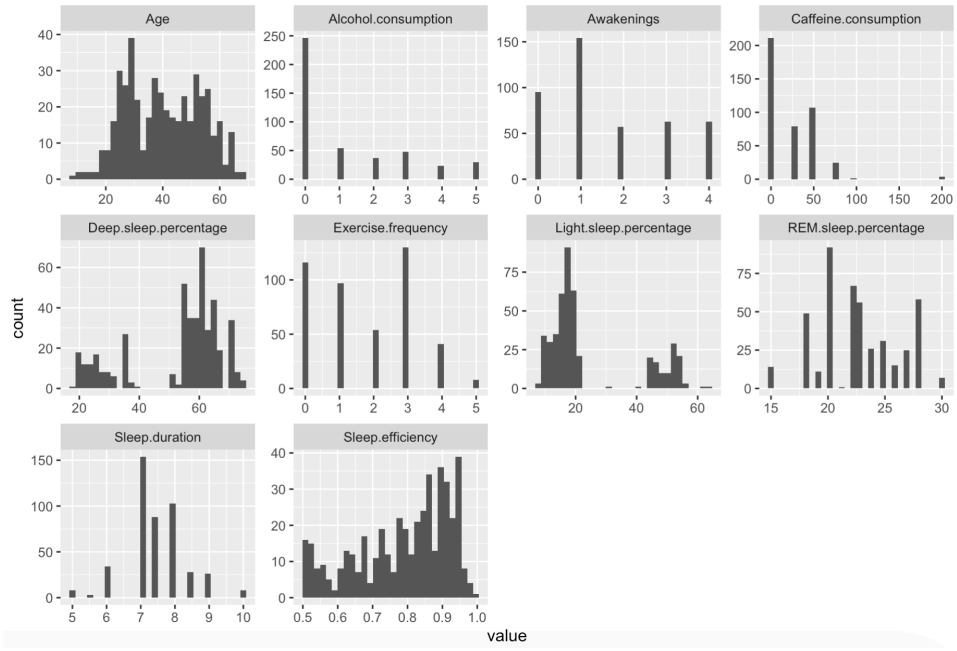

Fig 2: Histograms for Continuous Variables

We then create histograms for the variables and look at the distribution of all variables. The distribution for sleep efficiency is left skewed and the data range from 0.5 to 1.0. One other thing worth mentioning is that there is a clear cluster between the distribution for deep sleep percentage and light sleep percentage.

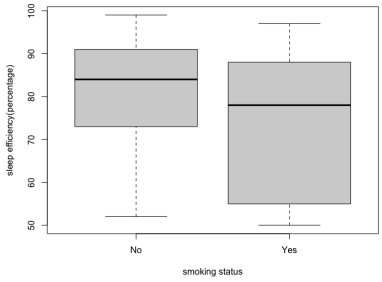

Fig 3: Smoking Status vs Sleep Efficiency

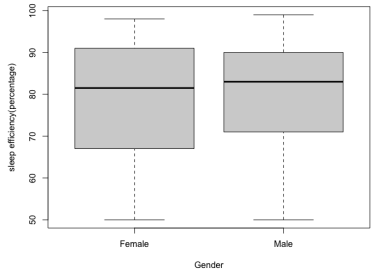

Fig 4: Gender vs Sleep Efficiency

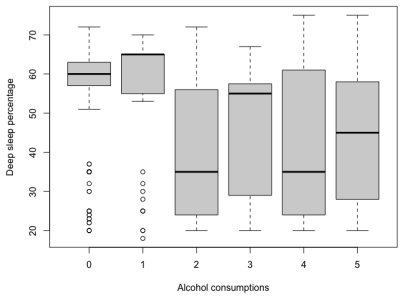

Fig 5: Alcohol Consumption vs Deep Sleep Percentage

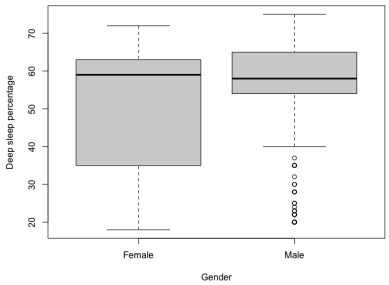

Fig 6: Gender vs Deep Sleep Percentage

We decided to create boxplots for some categorical variables to investigate what is the factor(s) that cause the cluster. By visualizing the boxplots of deep sleep percentage against alcohol consumption, we found that the subjects who consume very little or no alcohol (0-1 oz) generally have a higher deep sleep percentage than those who consume more than or equal to 2 oz of alcohol. Boxplots were also created for sleep efficiency against smoking status and gender. We noticed that subjects who smoke not only have a wider range but also a lower median sleep efficiency, indicating that smoking is likely negatively correlated with sleeping efficiency. There is no obvious difference between the sleep efficiency for males and females since the median and distribution for the two groups are similar.

Model Selection: No Interactions

First, we wanted to find the best linear model without any interactions. Light sleep percentage is dropped since it is just (1 - deep sleep percentage - REM sleep percentage) and highly collinear with the other two variables. In addition, Awakenings is also dropped because it overlaps with our response variable and might mask other associations. The purpose of this model is to predict sleep efficiency based on multiple predictor variables. Through the application of the `regsubsets()` function on the initial model with all the variables, an optimal set of 7 variables should be selected for constructing the model since it has Cp value = 7.675 and including more variables does not increase our adj R squared much. Thus, the new linear model is composed of the following variables: Age, REM sleep percentage, Deep sleep percentage, Caffeine consumption, Alcohol consumption, Smoking status, and Exercise frequency.

Through model analysis, we have identified this as an optimal regression model encompassing all variables exhibiting statistically significant linear correlations with sleep efficiency. Upon careful examination of the intercepts (β = 24.058, *se* =3.322, *p* = 2.28e-12***) of additional indicators, we have established that a unit increase or decrease in these indicators leads to a corresponding increase or decrease in sleep efficiency. Among these explanatory variables, Age (β = 0.087, *se* =0.028, *p* = 0.0017**), REM sleep percentage (β = 0.748, *se* =0.109, *p* = 2.76e-11***), Deep sleep percentage (β = 0.641, *se* =0.027, *p* < 2e-16***), Caffeine consumption (β = 0.037, *se* =0.013, *p* = 0.0043**), and Exercise Frequency (β = 1.102, *se* =0.255, *p* = 1.99e-05***), exhibit a positive correlation with sleep efficiency. Conversely, the remaining two predictors, Alcohol consumption (β = -0.816, *se* =0.245, *p* = 0.0009***) and Smoking status (β = -3.768, *se* =0.776, *p* = 1.75e-06***) , show a negative correlation.

In tandem, to assess the model's overall adequacy, we have observed that the mean of residuals—reflecting the disparities between initial and predicted values—is proximate to zero. This observation underscores the model's commendable ability to capture the variability of the response indicator.

Furthermore, the model's holistic performance is noteworthy, as evidenced by the low p-value of the F-statistic (< 0.0001). The model aptly elucidates a significant proportion of the sleep efficiency variance, substantiated by a multiple R-squared of 0.7174. The adjusted R-squared of 0.7125 accommodates the count of predictors while remaining substantial. The residual standard error, a measure of residuals' variability, approximates 7.243.

In conclusion, our meticulous analysis culminates in the identification of an optimal regression model, amalgamating indicators with statistically significant linear correlations to sleep efficiency. The model's ability to discern meaningful relationships, coupled with the negligible residual mean, underscores its robustness in elucidating the multifaceted factors shaping sleep efficiency.

Model Selection: With Interactions

To investigate whether there is any evidence that some of the explanatory variables might interact with each other in predicting our response variable, we added six interaction terms in our model, namely Smoking status × Exercise Frequency, Smoking status × Alcohol consumption, Gender × Caffeine consumption, Gender × Alcohol consumption, Gender × Smoking status, and Gender × Exercise frequency. These six interactions were selected based on the conjecture that the association between some of the explanatory variables and Sleep Efficiency might vary depending on the subject's gender and smoking status.

The results indicated that out of the six interaction terms, Smoking status x Alcohol consumption is the only significant one in the model and the sign of the coefficient estimate also translates well into real life. Our model predicts that for non-smokers, alcohol consumption is negatively associated with sleep efficiency (β = -0.166, *se* = 0.413, *p* = 0.688) though it's not statistically significant. However, compared with non-smokers which are the baseline, the magnitude of negative association between alcohol consumption and sleep efficiency is significantly bigger in smokers (β = -0.955, *se* = 0.485, *p* = 0.0495*). In this model, age (β = 0.084, *se* = 0.029, *p* = 0.0040**), REM sleep percentage (β = 0.732, *se* = 0.112, *p* = 2.33e-10 ***), deep sleep percentage (β = 0.633, *se* = 0.027, *p* < 2e-16 ***) and exercise frequency (β = 1.400, *se* = 0.352, *p* = 8.39e-05 ***) stayed significant with positive coefficients. In other words, an increase in those explanatory variables are associated with better sleep efficiency.

Model fit is also improved by the addition of interactions (adj R squared increased to 0.7147, compared with 0.7125 from our best model without any interaction term).

Model Diagnostics

Overall, the diagnostic plots showed that our model with interactions met the assumptions for linear regression (Fig X) and fit well with an adj R squared of 0.7147.

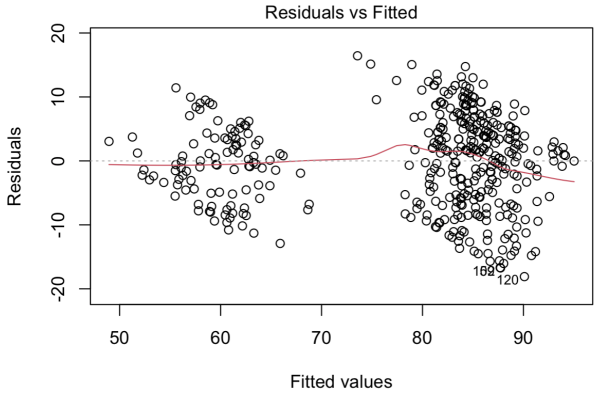

Fig. 7a Fitted values VS Residuals plot. The residual plot shows no obvious pattern and that the residuals are randomly distributed around 0 with constant variance.

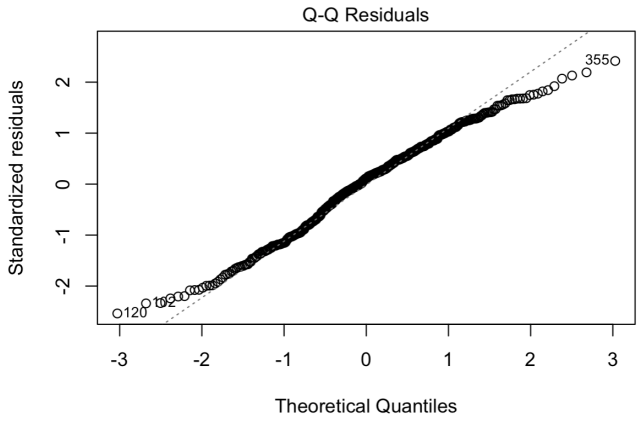

Fig. 7b Q-Q plot. It indicates that standardized residuals mostly agree with the theoretical quantiles with minor lighter tails.

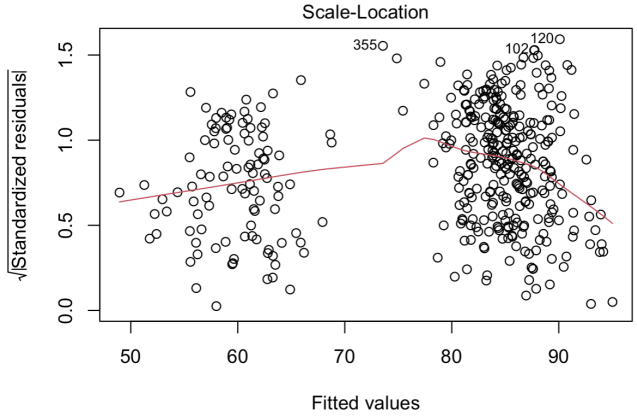

Fig. 7c Fitted values VS square root of standardized residuals plot. It shows no obvious pattern.

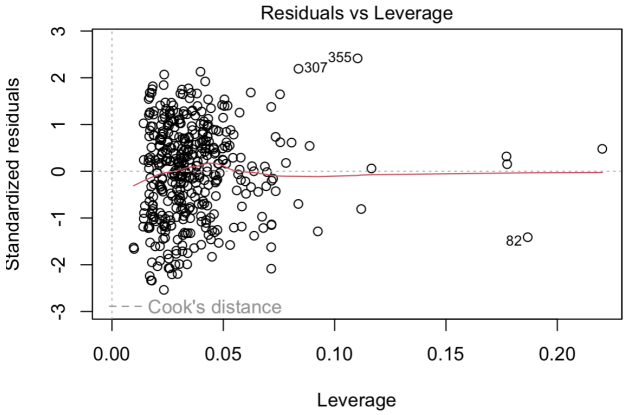

Fig. 7d Leverage VS Standardized residuals. There is no influential point with Cook's distance bigger than one.

Conclusion

From our analysis, we explored the relationships between sleep quality as measured by sleep efficiency, and 9 other variables. From our additive model, Age, REM Sleep Percentage, Deep Sleep Percentage, Alcohol Consumption, Smoking Status, and Exercise Frequency were shown to be linearly correlated with sleep quality, with statistical significance. Specifically, there is a positive relationship between Sleep Efficiency and Age, along with REM Sleep Percentage, Deep Sleep Percentage, and Exercise Frequency. In contrast, the model shows a negative relationship between Sleep Efficiency and being a smoker and between Sleep Efficiency and Alcohol Consumption. These results are expected, as many of these relationships are well documented in other research.

The most interesting of these results is the relationship between aging and sleep quality, as the elderly tend to have the worst quality of sleep compared to all other age groups. However, this is explained through our dataset, where the grand majority of participants are under the age of 65, and since research shows that those between the ages of 45 to 65 experience the best sleep quality, this is likely why we found a positive relationship between Sleep Efficiency and Age (Li et. al., 2018).

In our interaction model, we explored interactions between Gender and other lifestyle factors, along with Smoking and some lifestyle factors. We found only one of our 6 tested interactions to be statistically significant: the Smoking Status $\times$ Alcohol Consumption. This showed that in comparison to non-smokers, the magnitude of the negative associations between Alcohol Consumption and Sleep Efficiency is significantly bigger in smokers. This could have interesting impacts on those who are both heavy drinkers and heavy smokers, who may experience an increase in their sleep quality just by attempting to quit one of their habits. This model also found Age, REM Sleep Percentage, Deep Sleep Percentage, and Exercise Frequency to be statistically significant.

Overall, our findings suggest that Age, REM Sleep Percentage, Deep Sleep Percentage, Exercise Frequency, Smoking & Alcohol Consumption are the factors most correlated with Sleep Efficiency. In the future, it would be beneficial to look at these variables in more detail, for example, a wider range of ages or the number of cigarettes smoked per day among smokers.

Citations

Driver, H. S., & Taylor, S. R. (2000). Exercise and sleep. *Sleep Medicine Reviews*, *4*(4), 387–402. https://doi.org/10.1053/smrv.2000.0110

Hamilton, N. A., Nelson, C. A., Stevens, N., & Kitzman, H. (2006). Sleep and psychological well-being. *Social Indicators Research*, *82*(1), 147–163. https://doi.org/10.1007/s11205-006-9030-1

Leger, D., Andler, R., Richard, J., Nguyen‐Thanh, V., Collin, O., Chennaoui, M., & Metlaine, A. (2022). Sleep, substance misuse and addictions: A nationwide observational survey on smoking, alcohol, cannabis and sleep in 12,637 adults. *Journal of Sleep Research*, *31*(5). https://doi.org/10.1111/jsr.13553

Li, J., Vitiello, M. V., & Gooneratne, N. S. (2018). Sleep in normal aging. *Sleep Medicine Clinics*, *13*(1), 1–11. https://doi.org/10.1016/j.jsmc.2017.09.001