Statistical Analysis of a Good Night's Sleep

STAT 306 Group Project Report

Group A2

Frederick Wang, Horton Lai, Suliat Yakubu

…

Table of Contents

Introduction

A myriad of lifestyle factors have been hypothesized to influence the quality of sleep. As part of the natural cycle in human behaviour, a good night's sleep is paramount to one's productivity during the day, since insufficient sleep negatively impacts one's focus and creativity. In addition, a lack of sleep may weaken one's immune system, leading to greater susceptibility to short-term diseases as well as an increased risk of developing long-term illnesses. Sleep disorders and changes in sleep are associated with an increase in all-cause mortality and cardiovascular disease [1], as well as an increase in sudden cardiac death [2]. Despite these detrimental consequences, the fast-paced lifestyle in modern society incentivizes people to sleep less so as to work more. Students are motivated to reduce their sleep quantity to allow more time for studying. Workers are incentivized to work overtime, as a result pushing back their bedtime. Hence, sleeping has now become a balancing act. One has to decide whether he should sleep more so as to increase productivity during the day or sleep less in order to get more work done. This poses a very interesting optimization problem, which is the centre of this observational study.

There are many lifestyle factors observed in this study with known effects on sleep duration and quality, such as alcohol use, caffeine consumption, smoking status, and exercise frequency. The effects of alcohol on sleep quality are well-studied. In the early phase of the night, when blood alcohol content is higher, alcohol acts as a sedative, decreasing latency to sleep, while also increasing slow-wave sleep, and suppressing REM sleep. In the later stages of the night, there is increased wakefulness/light sleep [3].

Caffeine is a widely consumed stimulant, and has many sources in the modern diet, including unexpected foods such as sodas, medicines, and chocolate. Caffeine is an adenosine receptor antagonist, which is involved in brain functions related to sleep, arousal and cognition. Studies report a positive effect of caffeine consumption on physical and cognitive performance, but a detrimental effect on sleep quality [4], leading to a vicious cycle wherein individuals consume caffeine for its beneficial effects, which results in decreased sleep quality, leading individuals to consume more caffeine to counter the effects of a poor night of sleep. In addition to worsening sleep quality, caffeine can also attenuate sleep pressure, potentially resulting in a longer time until the onset of sleep after going to bed [5].

Nicotine is a stimulant acting on nicotinic acetylcholine receptors, which indirectly affect pathways involved in dopamine and serotonin release, neurotransmitters that promote wakefulness and inhibit REM sleep [6], [7]. Cigarette smokers are more likely to experience effects such as insomnia and poor sleep quality as marked by disturbances like shorter sleep duration, increased sleep latency, and daytime sleepiness [8].

Exercise has generally been recommended as a sleep aid, however, the mechanism through which exercise affects sleep quality and vice versa is not yet understood [9]. In adults, acute and regular exercise frequency has been shown to have positive effects on sleep quality, with acute exercise sessions showing small effects on factors such as total sleep time and time in slow wave sleep, and regular exercise showing improvements on multiple factors of sleep quality, as well as increasing total sleep time and efficiency [10].

It is understood that these lifestyle choices all have an effect on sleep quality, and eliminating all choices with a detrimental effect on sleep while upregulating all choices with a beneficial effect would be the obvious solution for optimizing one's quality of sleep. However, whether an individual partakes in these behaviours is dependent on a multitude of factors, including social climate and external demands.

Therefore, it is worth examining how these various factors correlate with a good night's sleep, to potentially optimize sleep in a manner that more closely reflects the demands of the modern lifestyle.

Our statistical analysis is based on the dataset available at https://www.kaggle.com/datasets/equilibriumm/sleep-efficiency. The data is collected by a group of artificial intelligence engineering students in High National School for Computer Science and Systems Analysis (ENSIAS). Participants were recruited from local communities in Morocco and various metrics were monitored using self-reported surveys, actigraphy, and polysomnography over the span of a few months. The dataset covers multiple aspects of the participants' daily life. Physical attributes include age and gender (male or female). Behavioral factors include bedtime (in YYYY-MM-DD HH:MM:SS), caffeine (mg in 24 hours prior to bedtime) and alcohol (oz in 24 hours prior to bedtime) consumption, smoking habits (yes or no), and exercise frequency. Biological measurements of how well the participants slept include wake-up time (YYYY-MM-DD HH:MM:SS), sleep duration (in hours), number of awakenings, the percentage of REM, deep, and light sleep, and sleep efficiency (%). It is important to note that the goodness of sleep is evaluated not only by its quantity but also by its quality. The data provided contains qualitative information such as the percentage of REM and deep sleep. However, it is unclear how different sleep stages precisely affect sleep quality. Therefore, our study will focus mainly on sleep efficiency, which is the proportion of time spent in bed when the participant is actually asleep.

Our primary research question is to determine which factors appear to be correlated with efficiency. We will also produce an optimal prediction model for sleep efficiency with these factors. With regards to factors which are lifestyle choices, this will allow readers to identify the degree to which poor lifestyle decisions can negatively impact their sleep, as well as which positive lifestyle factors can improve their sleep.

Analysis

Initial Exploration

We start off the analysis by standardizing the bedtime and wakeup time variables into the number of hours after a specified time, 9 pm for bedtime and 3 am for wakeup time. To explore the data, we plotted our response variable, sleep efficiency, against other explanatory variables in the dataset. Individually, each explanatory variable does not seem to exhibit any particular trends with our response variable (Figures 1, 2, 3). Apart from deep sleep percentage and light sleep percentage where the data seems to be separated into 2 groups, the data in other plots are scattered throughout. Based on these plots, we have no reason to perform any transformation of our data, and hence, we decided to start with a model with only linear explanatory variables.

Then we began our analysis by training a full model on all of the explanatory variables in the dataset minus the "ID" column. From the residual plot of the full model, we can see that it seems like a random scatter of points centered around 0 with no apparent changes in variation (Figure 4). The lack of pattern seems to support the assumption of homoscedasticity, suggesting that no transformations are required from that assumption alone. There do seem to be two clusters of residual points, though, perhaps suggesting some other explanatory categorical variable that is missing from the model. There is also a pattern in the residuals where in the rightmost part of the right cluster, the points form sloped lines, which also suggest some missing explanatory categorical variable. These issues, unfortunately, don't seem to be able to be addressed with the given dataset as this model is the full model fitted with all of the given explanatory variables. This is a potential issue in observatory studies such as this one, as variables unaccounted for in the model, from uncollected data, can confound the relationships observed.

Model Selection

First, we calculated the partial correlation for each explanatory variable to determine if any of them are highly correlated, indicating close to collinearity. We identified Bedtime, Wakeup time, Sleep duration, REM sleep percentage, deep sleep percentage, and light sleep percentage as the variables causing collinearity in our full model because their partial R^2 values are 1 and variance inflation factors (VIF) are clearly greater than the conventional threshold of 10 (Table 1). This drew our attention to these variables in the dataset. Upon careful inspection, we realized that the sleep duration is the difference in bedtime and wake-up time. Their collinearity is consistent with our expectations. Hence, any 2 of these variables would be sufficient in predicting the 3rd one. We kept this in mind as we perform our model selection, that in the final model, these 3 variables should not coexist together. Also, REM sleep percentage, deep sleep percentage, and light sleep percentage add up to 1, which presents the same issue. Similarly, a maximum of 2 out of these 3 variables can exist in our final model.

As we have insignificant variables in the full model from looking at the t-values and the respective p-values, we used backward selection to remove the insignificant explanatory variables. In order, we removed: gender, bedtime, wakeup time, then sleep duration. At this point, we still have an issue with the REM, deep, and light sleep percentages being related to each other as they add up to 100%. To resolve this problem, we utilized the best subset selection algorithm and selected the best set of variables using adjusted R^2 and Mallows' C_p. Based on the values calculated, we determined that the best model would be the one with 6 explanatory variables since the adjusted R^2 is 0.7984 which is close to the maximum among the subsets which is 0.8008 (Figure 5). For Mallows' C_p, the 6-variable model is 6.3671 which we deemed acceptable (Figure 5). However, the difference between a 6-variable model and a 7-variable model is the exclusion of caffeine consumption, which, based on previous literature, is expected to affect sleep quality and quantity. Hence, we decided that a model including caffeine consumption is more consistent with existing knowledge since we know that caffeine consumption should affect sleep. In the end, we chose to only include the light sleep percentage in the model out of the three sleep percentages. The reasoning behind this decision was that the three criteria suggested that a model with 7, 8, or 9 explanatory variables is similar in quality and therefore, the smallest model of size 7 was selected (Figure 5). The intermediate linear model with no interaction terms we chose includes the following variables: age, light sleep percentage, awakenings, caffeine consumption, alcohol consumption, smoking status, and exercise frequency. Again, we recalculated the VIF of each variable in this model and confirmed that the collinearity seen previously had been dealt with properly (Table 2).

Next, among the terms in this intermediate model, we tested for interaction terms based on interactions reported in the literature. Specifically, we tried adding interactions between 1) awakenings and alcohol consumption, as alcohol consumption results in increased awakenings in the later stages of sleep [3], 2) age and light sleep percentage, as deep sleep percentage decreases with age [11], 3) caffeine consumption and smoking status, as nicotine use increases the speed of caffeine metabolism by as much as 50% [4], and 4) age and exercise frequency, as research points strongly to the ability of exercise to improve sleep quality in older adults [9]. In the model which includes all 4 of these interaction terms, we found that their p-values are 0.0037, 0.0020, 0.3327, and 0.5733 respectively. Clearly, only interactions 1 and 2 are significant based on a 0.05 threshold. Therefore, we sequentially removed the interaction terms with the largest insignificant p-value, until we have achieved a model with only significant interaction terms. In addition to the 7 explanatory variables, our final optimized model includes the interaction terms between awakenings and alcohol consumption, and between age and light sleep percentage, both of which have a p-value of less than 0.003. To evaluate the quality of this model from another angle, we generated the residual and normal quantile plots. Similarly, the residuals do not show any pattern except the 2 groups which we already saw in the full model (Figure 6). At the same time, the q-q plot of the residuals for the optimized model looks to be reasonably in line with the theoretical quantiles of the residuals, validating the assumption that the residuals are normally distributed in the final optimized model (Figure 7).

Leverage and Outliers

After selecting the model, we searched for outliers by checking the standardized residuals. Using the rule that a standardized residual greater than 3 as a criteria for evaluating outliers, our data does not contain any outliers. However, the 3 highest values of our absolute standardized residuals are 2.8159, 2.5157, and 2.5096 respectively. Despite being less than 3, these values are approaching 3, suggesting that these data points may potentially warrant investigation. To evaluate the effect of these data points on our model, we evaluated their leverage and influence using the diagonal entries of the hat matrix as well as Cook's distance. The leverages of these data points are sufficiently low so we have confirmed that these data points are not outliers.

When determining the leverage earlier, we noticed 3 other data points with a leverage value of approximately 0.11. Using 2k/n = 0.0515 as an evaluation criterion, these data points have unusually high leverage, which may potentially exert an unusually high influence on our model. To further investigate this to make sure our model is not compromised, we looked at the influence that each data point has on the model. The maximum Cook's distance between our data points is 0.051005. Since it is less than 1, we concluded that no data points are particularly influential on the fitting of our model.

Statistical Inference

In the end, we have a multiple linear regression model for the prediction of sleep efficiency with the following terms and respective coefficients:

Terms	**Coefficients**
Intercept	0.9432837
Age	0.0009649
Light Sleep Percentage	-0.0055693
Awakenings	-0.0318042
Caffeine Consumption	0.0002489
Alcohol Consumption	-0.0062626
Smoking Status	-0.0454178
Exercise Frequency	0.0065921
Awakenings:Alcohol Consumption	0.00413
Age:Light Sleep Percentage	0.00004492

The intercept represents the sleep efficiency when the individual's age is 0, light sleep percentage is 0, with 0 awakenings, no caffeine and alcohol consumed in the 24 hours prior to bedtime, who does not smoke and does not exercise. For the continuous explanatory variables, the coefficients of the non-interaction terms above can be interpreted as the change in sleep efficiency per one unit increase in its respective term, with all other terms held constant. The coefficients for the interaction terms can be interpreted as the change in sleep efficiency when the interacting product increases by one unit, with all other non-interacting terms held constant. This change in sleep efficiency from interactions is on top of the changes from the interactant terms' individual coefficients. Meanwhile, for the categorical explanatory variable, smoking status, the coefficient of the non-interaction terms represents the difference in the intercept between the different categories (yes vs no in smoking status), but our model does not allow a change in the slope between regression of different smoking statuses since smoking status is not involved in any interaction terms.

Using a multiple linear regression model, we identified age, light sleep percentage, awakenings, caffeine consumption, alcohol consumption, smoking, and exercise frequency as factors that appear to have linear relationships with sleep efficiency.

Discussion

Unsurprisingly, our model identified that negative lifestyle choices such as alcohol consumption and smoking correlate with decreased sleep efficiency. Our model predicts a 0.62% decrease in sleep efficiency per ounce of alcohol consumed 24 hours prior to bedtime. Unlike alcohol which contained the quantity consumed in the last 24 hours, smoking was a categorical variable, as such our model predicts that being a smoker leads to a 4.54% decrease in sleep efficiency.

Interestingly enough, our model predicts an increase in sleep efficiency when caffeine is consumed, specifically an increase of 0.024% per milligram. Obviously, this contrasts with our intuitive understanding of the effects of caffeine on sleep, as well as with research on the matter suggesting that sleep timing can be affected by caffeine [12]. We note that caffeine consumption over the past 24 hours is too big of a window since a person could drink a cup of coffee in the morning and have a large amount of it eliminated from their system by the time they go to bed. The half-life of caffeine in females studied was 17.63 hours [13], which is approximately the amount of time between a morning cup of coffee and bedtime. This may be the reason for the disagreement between our model coefficient and the literature. If the dataset instead recorded caffeine consumption in the past 12 hours, it would likely more accurately record caffeine consumption that would affect sleep, which may have led to a more accurate coefficient for caffeine consumption.

Consistent exercise is also positively correlated with sleep. Our model predicts an increase of 0.65% in sleep efficiency per day of weekly exercise. That being said, the results of our model are consistent with exercising's positive effect on sleep that has been observed in empirical studies [9].

The reason that some of our coefficients do not agree with the results from literature in sleep studies may also suggest that some of the factors may be missing from our model due to unassessed explanatory variables. For example, as mentioned before, the residuals are partitioned into roughly 2 groups, which we cannot adjust for. Therefore, the predictions from our model may be slightly weaker in predictive power when it is applied to individuals in the population, and more research would be required to determine the missing explanatory variables to resolve the inconsistencies between our model and the literature.

Conclusion

From the final optimized model trained on the given dataset, we were able to determine that age, light sleep percentage, awakenings, caffeine consumption, alcohol consumption, smoking status, and exercise frequency have a statistically significant linear relationship with one's sleep efficiency. We also found that the interaction between alcohol and awakenings, and age and light sleep was linearly related to sleep efficiency. Our findings match up well with the findings of other research papers, such as the negative effects of caffeine on sleep [4], wakefulness caused by nicotine [7], [8], and the benefits exercise has on sleep [9].

With this model, we can say that there are correlations between these factors and sleep efficiency and that if one would like to improve their sleep efficiency, changing lifestyle choices, including, caffeine and alcohol consumption, smoking, and exercise are good places to start. Of course, as this is only a predictive linear model, we cannot say for certain that changing these factors will have a certain effect on sleep efficiency; further research experiments will need to be conducted to demonstrate causation.

Figures and Tables

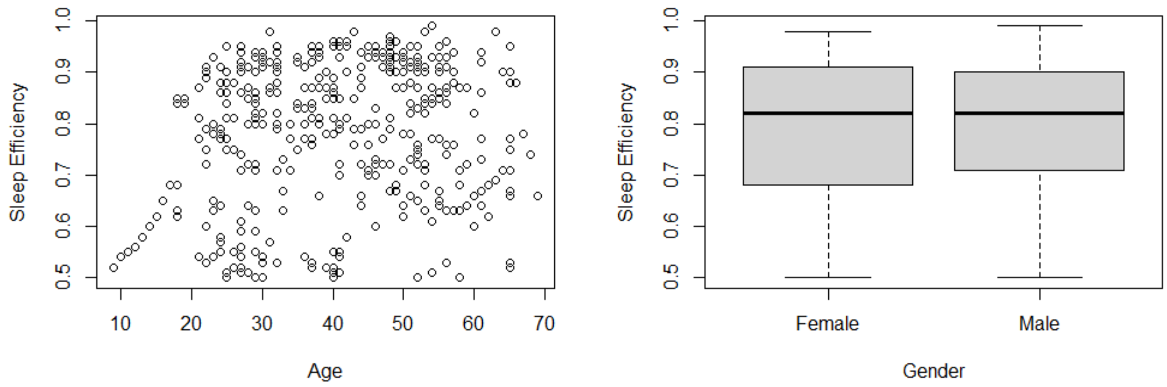


Figure 1: Sleep efficiency vs physical attributes such as age and gender.

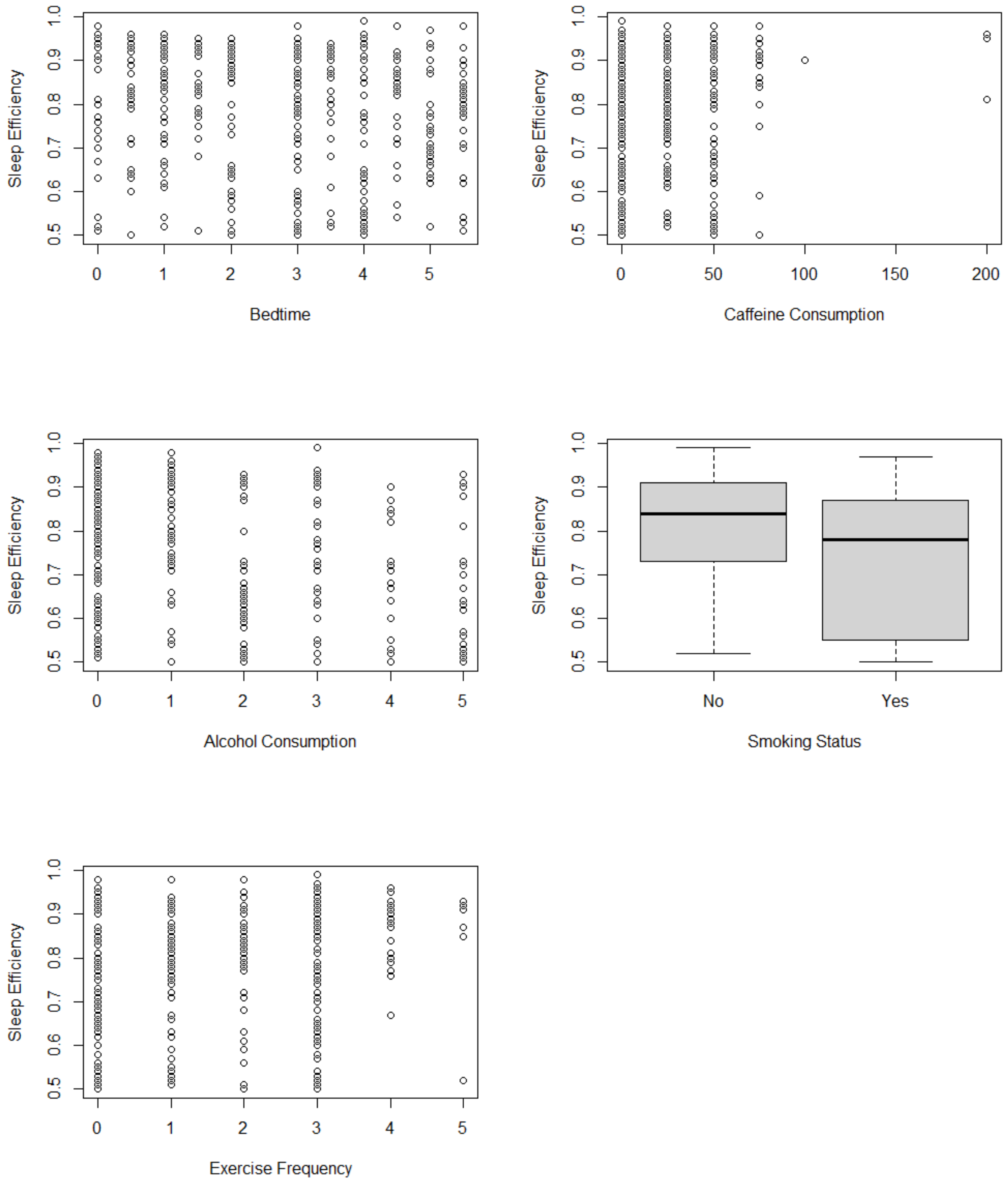


Figure 2: Sleep efficiency vs behavioral factors such as bedtime, caffeine consumption, alcohol consumption, smoking status, and exercise frequency.

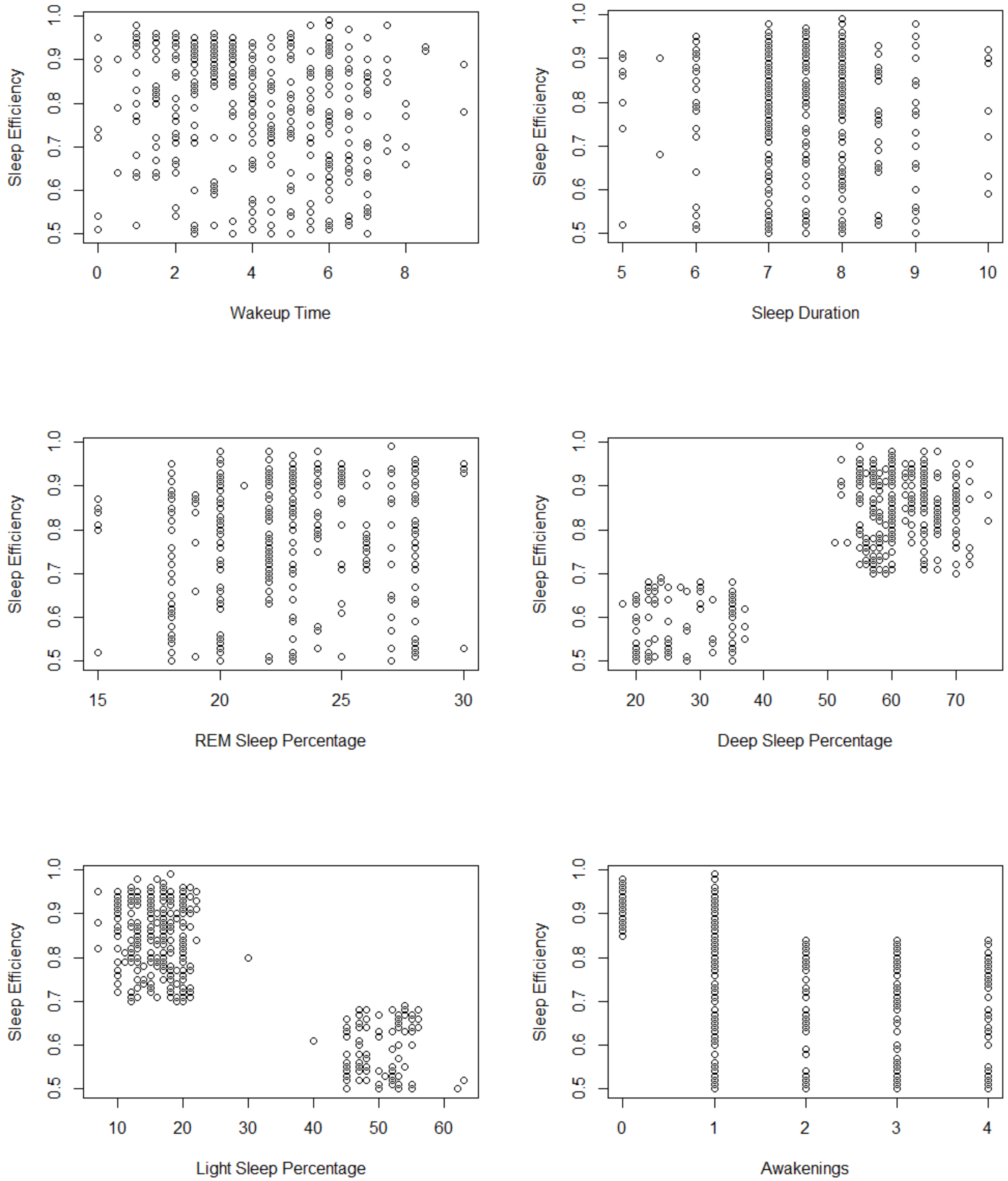


Figure 3: Sleep efficiency vs biological measurements such as wake-up time, sleep duration, REM sleep percentage, deep sleep percentage, light sleep percentage, and awakenings.

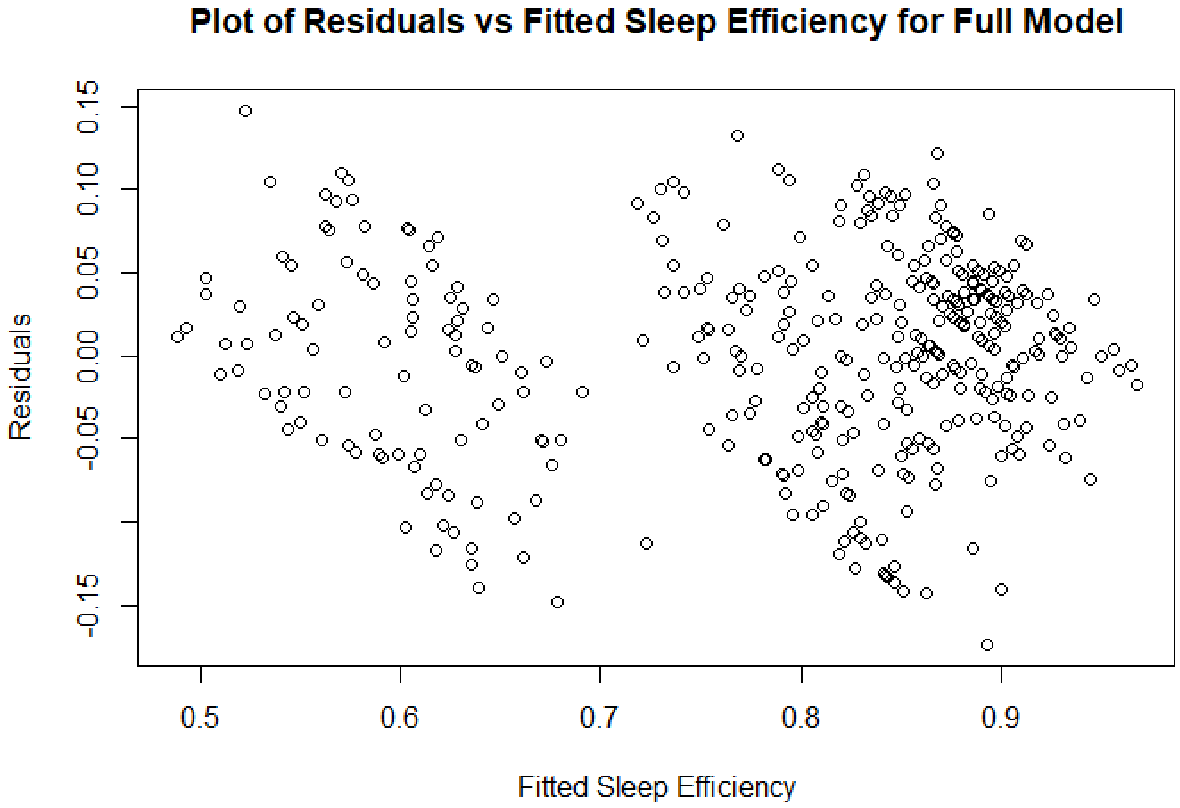


Figure 4: As we have discussed in the analysis, it seems that in the residual plot for the full model, there are two groupings of data, one with high sleep efficiency and one with low sleep efficiency. In the cluster of high sleep efficiency data, there also seem to be some patterns of sloped lines of data. Both of these characteristics suggest missing explanatory variables, but unfortunately, there are no further variables given in the dataset that could be used to resolve this issue. Otherwise, the residual plot looks reasonably random centered at 0 and with no large deviations in variance.

Explanatory Variable in Full Model	Partial R^2	Variance Inflation Factor (VIF)
Age	0.1304	1.1500
Bedtime	1	∞
Wakeup time	1	∞
Sleep duration	1	∞
REM sleep percentage	1	∞
Deep sleep percentage	1	∞
Light sleep percentage	1	∞
Awakenings	0.1903	1.2350
Caffeine consumption	0.1188	1.1348
Alcohol consumption	0.1835	1.2247
Exercise frequency	0.2183	1.2792

Table 2: Partial R^2 and Variance Inflation Factor (VIF) of continuous explanatory variables in the full model.

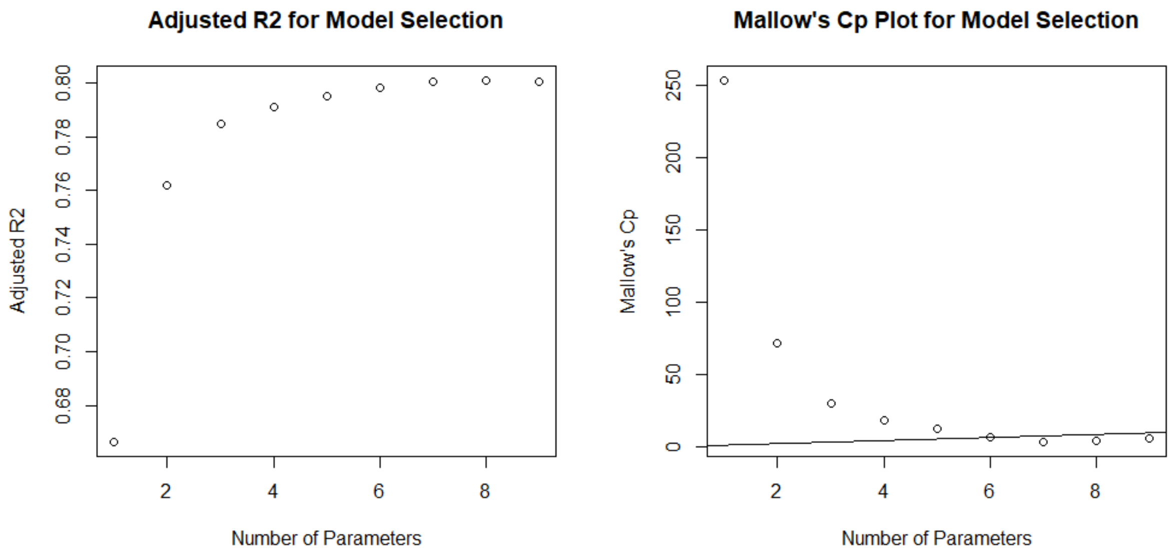


Figure 5: Adjusted R^2 and Mallow's C_p Plot for selecting the best subset of the variables remaining after backward selection.

Explanatory Variable in Intermediate Model	Partial R^2	Variance Inflation Factor (VIF)
Age	0.0424	1.0443
Light sleep percentage	0.2766	1.3824
Awakenings	0.1731	1.2093
Caffeine consumption	0.0571	1.0606
Alcohol consumption	0.1795	1.2188
Exercise frequency	0.0869	1.0952

Table 3: Partial R^2 and VIF of continuous explanatory variables in the intermediate model with no interaction terms. Column 2 contains the partial R^2 between the response variable and each explanatory variable in the optimized model, controlled for the other variables. Column 3 contains the Variance Inflation Factor of each explanatory variable.

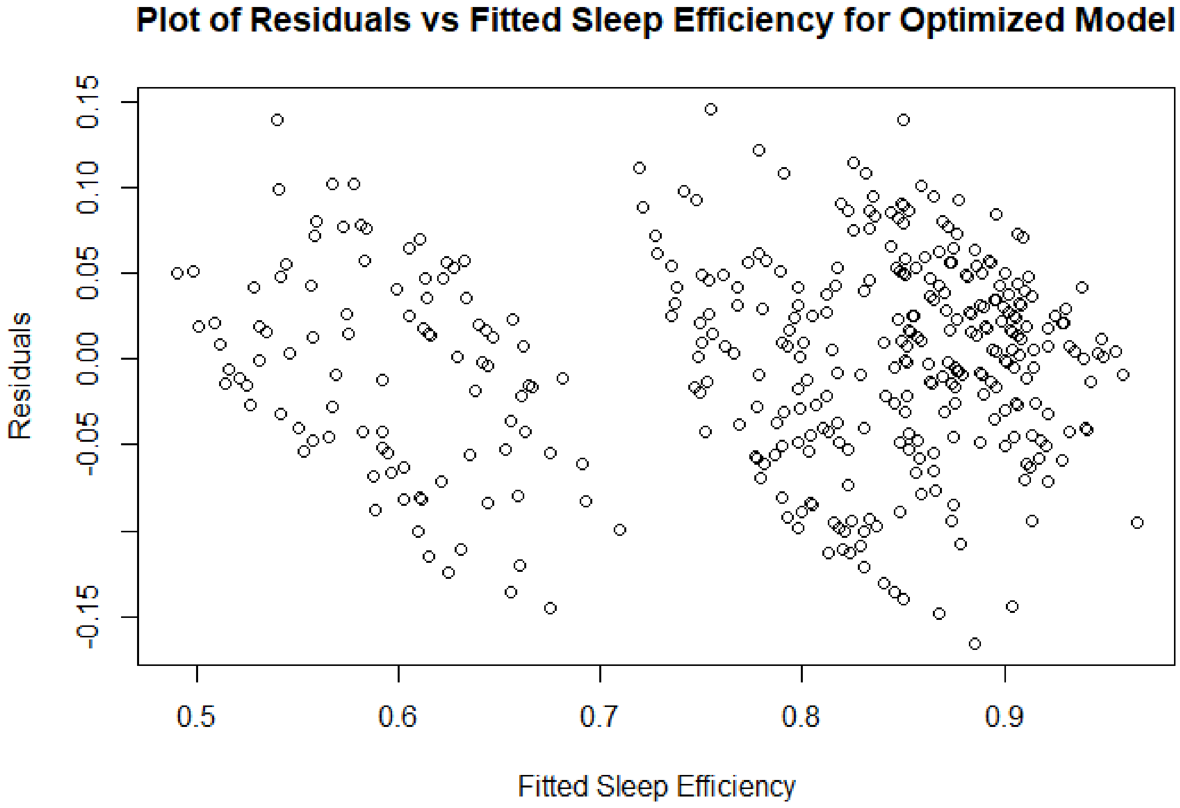


Figure 6: Residuals vs the fitted sleep efficiency for the final optimized model with interaction terms. The same issue with the two clusters of data applies here as that with the plot for the full model.

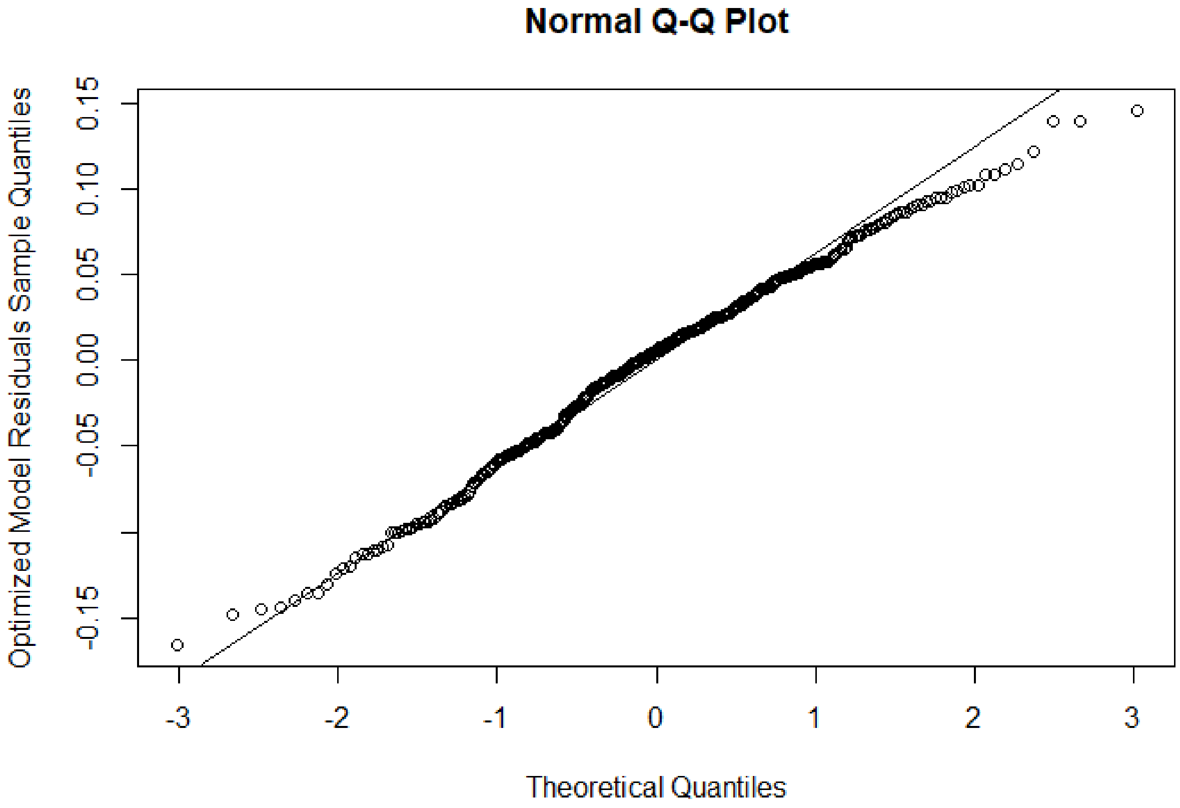
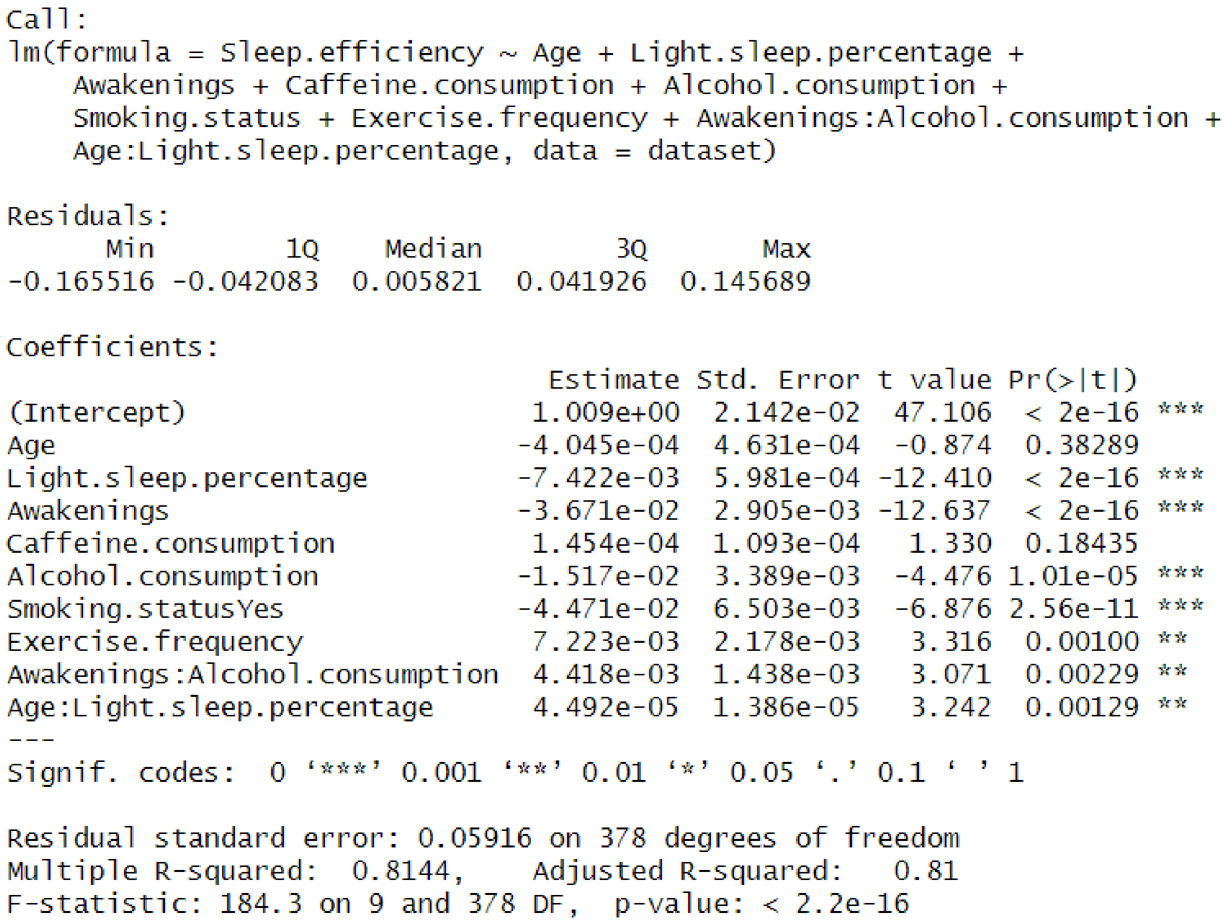


Figure 7: The q-q plot of the residuals for the final optimized model looks to be reasonably in line with the theoretical quantiles of the residuals, suggesting that the assumption that the residuals are normally distributed can be made.



```
Call:
lm(formula = Sleep.efficiency ~ Age + Light.sleep.percentage + 
    Awakenings + Caffeine.consumption + Alcohol.consumption + 
    Smoking.status + Exercise.frequency + Awakenings:Alcohol.consumption + 
    Age:Light.sleep.percentage, data = dataset)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.165516 -0.042083  0.005821  0.041926  0.145689 

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     1.009e+00  2.142e-02  47.106  < 2e-16 ***
Age                            -4.045e-04  4.631e-04  -0.874  0.38289    
Light.sleep.percentage         -7.422e-03  5.981e-04 -12.410  < 2e-16 ***
Awakenings                     -3.671e-02  2.905e-03 -12.637  < 2e-16 ***
Caffeine.consumption            1.454e-04  1.093e-04   1.330  0.18435    
Alcohol.consumption            -1.517e-02  3.389e-03  -4.476 1.01e-05 ***
Smoking.statusYes              -4.471e-02  6.503e-03  -6.876 2.56e-11 ***
Exercise.frequency              7.223e-03  2.178e-03   3.316  0.00100 ** 
Awakenings:Alcohol.consumption  4.418e-03  1.438e-03   3.071  0.00229 ** 
Age:Light.sleep.percentage      4.492e-05  1.386e-05   3.242  0.00129 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05916 on 378 degrees of freedom
Multiple R-squared:  0.8144,	Adjusted R-squared:   0.81 
F-statistic: 184.3 on 9 and 378 DF,  p-value: < 2.2e-16
```

Figure 8: Summary Statistics of the final optimized model with interaction terms.

Citations

[1] T. Kendzerska, T. Mollayeva, A. S. Gershon, R. S. Leung, G. Hawker, and G. Tomlinson, "Untreated obstructive sleep apnea and the risk for serious long-term adverse outcomes: a systematic review," *Sleep Med Rev*, vol. 18, no. 1, pp. 49–59, Feb. 2014, doi: 10.1016/j.smrv.2013.01.003.

[2] A. S. Gami *et al.*, "Obstructive sleep apnea and the risk of sudden cardiac death: a longitudinal study of 10,701 adults," *J Am Coll Cardiol*, vol. 62, no. 7, pp. 610–616, Aug. 2013, doi: 10.1016/j.jacc.2013.04.080.

[3] I. M. Colrain, C. L. Nicholas, and F. C. Baker, "Alcohol and the Sleeping Brain," *Handb Clin Neurol*, vol. 125, pp. 415–431, 2014, doi: 10.1016/B978-0-444-62619-6.00024-0.

[4] F. O'Callaghan, O. Muurlink, and N. Reid, "Effects of caffeine on sleep quality and daytime functioning," *Risk Manag Healthc Policy*, vol. 11, pp. 263–271, Dec. 2018, doi: 10.2147/RMHP.S156404.

[5] J. Weibel *et al.*, "Regular Caffeine Intake Delays REM Sleep Promotion and Attenuates Sleep Quality in Healthy Men," *J Biol Rhythms*, vol. 36, no. 4, pp. 384–394, Aug. 2021, doi: 10.1177/07487304211013995.

[6] A. Jaehne, B. Loessl, Z. Bárkai, D. Riemann, and M. Hornyak, "Effects of nicotine on sleep during consumption, withdrawal and replacement therapy," *Sleep Medicine Reviews*, vol. 13, no. 5, pp. 363–377, Oct. 2009, doi: 10.1016/j.smrv.2008.12.003.

[7] H. Li *et al.*, "Association of Cigarette Smoking with Sleep Disturbance and Neurotransmitters in Cerebrospinal Fluid," *Nat Sci Sleep*, vol. 12, pp. 801–808, Oct. 2020, doi: 10.2147/NSS.S272883.

[8] H. Purani, S. Friedrichsen, and A. Allen, "Sleep quality in cigarette smokers: Associations with smoking-related outcomes and exercise," *Addict Behav*, vol. 90, pp. 71–76, Mar. 2019, doi: 10.1016/j.addbeh.2018.10.023.

[9] B. A. Dolezal, E. V. Neufeld, D. M. Boland, J. L. Martin, and C. B. Cooper, "Interrelationship between Sleep and Exercise: A Systematic Review," *Adv Prev Med*, vol. 2017, p. 1364387, 2017, doi: 10.1155/2017/1364387.

[10] M. A. Kredlow, M. C. Capozzoli, B. A. Hearon, A. W. Calkins, and M. W. Otto, "The effects of physical activity on sleep: a meta-analytic review," *J Behav Med*, vol. 38, no. 3, pp. 427–449, Jun. 2015, doi: 10.1007/s10865-015-9617-6.

[11] "Sleep in Normal Aging - PMC." https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5841578/ (accessed Aug. 11, 2023).

[12] G. Mathew, D. Reichenberger, L. Master, O. Buxton, A.-M. Chang, and L. Hale, "0184 Too Jittery to Sleep? Temporal Associations of Actigraphic Sleep and Caffeine in Adolescents," *Sleep*, vol. 45, no. Supplement_1, p. A85, Jun. 2022, doi: 10.1093/sleep/zsac079.182.

[13] A. Ali, J. M. O'Donnell, C. Starck, and K. J. Rutherfurd-Markwick, "The Effect of Caffeine Ingestion during Evening Exercise on Subsequent Sleep Quality in Females," *Int J Sports Med*, vol. 36, no. 06, pp. 433–439, Jun. 2015, doi: 10.1055/s-0034-1398580.

Appendix

Dataset Limitations

Each data point represented a single time point collected for a single subject: There is an inherent weakness in the dataset in that there is variation in one's quality of sleep from night to night, even if all variables are held constant- by capturing a single time point, we have nothing to compare the difference in sleep with alcohol vs sleep without alcohol in a single subject, with sleep without alcohol serving as the baseline (for example). Additionally, factors such as alcohol, caffeine and nicotine consumption have differing effects on sleep when considered under acute or chronic circumstances, therefore a person's habitual consumption needs to be considered alongside consumption the day prior.

Gender and sex have been conflated in this dataset: how male/female vs man/woman/other was determined is unclear. Neither gender nor sex form a true binary, therefore there is an inherent limitation in that a misclassification due to limited options could confound the relationships.

The data was collected by a group of artificial intelligence engineers rather than sleep scientists: This is important, as data collected by AI engineers, for the purpose of training the model, may miss confounding variables, or lack underlying assumptions that a group of sleep scientists would be on the lookout for

How each type of data is collected is not reported in the dataset: for example, it is unclear whether the sleep duration was collected using polysomnography or actigraphy. This could be relevant, because if a polysomnograph was used, which is the gold standard for conducting sleep studies, then confounding effects of sleeping in an unfamiliar environment may play into the data of some members of the population, whereas an actigraph collects data at home in a familiar environment, but is not as reliable. Data collection modality could be an unaccounted-for, significant explanatory variable in the model.

R Code

```r
#Loading Libraries
library(lubridate)
library(leaps)
library(car)

#Data preprocessing
dataset = read.table('Sleep_Efficiency.csv', TRUE, ',')
dataset = na.omit(dataset)

dataset$Smoking.status = as.factor(dataset$Smoking.status)
dataset$Gender = as.factor(dataset$Gender)
dataset$Bedtime = as.POSIXlt(dataset$Bedtime)
dataset$Wakeup.time = as.POSIXlt(dataset$Wakeup.time)

#Turn bedtime into number of hours after 9:00pm
hours = as.numeric(hour(dataset$Bedtime))
hours = as.numeric(hours + minute(dataset$Bedtime)/60)
hours = ifelse(21-hours <= 0, hours - 21, hours + 3)
dataset$Bedtime = hours
dataset$Bedtime

#Turn wake up time into hours since 3:00am
hours = as.numeric(hour(dataset$Wakeup.time))
hours = as.numeric(hours + minute(dataset$Wakeup.time)/60)
hours = hours - 3
dataset$Wakeup.time = hours
dataset$Wakeup.time



#Main model selection
plot(Sleep.efficiency ~ Age, data = dataset, xlab = 'Age', ylab = 'Sleep
Efficiency')
plot(Sleep.efficiency ~ Gender, data = dataset, xlab = 'Gender', ylab = 'Sleep
Efficiency')
plot(Sleep.efficiency ~ Sleep.duration, data = dataset, xlab = 'Sleep
Duration', ylab = 'Sleep Efficiency')
plot(Sleep.efficiency ~ REM.sleep.percentage, data = dataset, xlab = 'REM
Sleep Percentage', ylab = 'Sleep Efficiency')
plot(Sleep.efficiency ~ Deep.sleep.percentage, data = dataset, xlab = 'Deep
Sleep Percentage', ylab = 'Sleep Efficiency')
```

```r
plot(Sleep.efficiency ~ Light.sleep.percentage, data = dataset, xlab = 'Light
Sleep Percentage', ylab = 'Sleep Efficiency')
plot(Sleep.efficiency ~ Awakenings, data = dataset, xlab = 'Awakenings', ylab
= 'Sleep Efficiency')
plot(Sleep.efficiency ~ Caffeine.consumption, data = dataset, xlab = 'Caffeine
Consumption', ylab = 'Sleep Efficiency')
plot(Sleep.efficiency ~ Alcohol.consumption, data = dataset, xlab = 'Alcohol
Consumption', ylab = 'Sleep Efficiency')
plot(Sleep.efficiency ~ Smoking.status, data = dataset, xlab = 'Smoking
Status', ylab = 'Sleep Efficiency')
plot(Sleep.efficiency ~ Exercise.frequency, data = dataset, xlab = 'Exercise
Frequency', ylab = 'Sleep Efficiency')
plot(Sleep.efficiency ~ Bedtime, data = dataset, xlab = 'Bedtime', ylab =
'Sleep Efficiency')
plot(Sleep.efficiency ~ Wakeup.time, data = dataset, xlab = 'Wakeup Time',
ylab = 'Sleep Efficiency')


full_model = lm(Sleep.efficiency ~ Age + Gender + Sleep.duration +
REM.sleep.percentage + Deep.sleep.percentage + Light.sleep.percentage +
Awakenings + Caffeine.consumption +
                 Alcohol.consumption + Smoking.status + Exercise.frequency +
Bedtime + Wakeup.time, data = dataset)
summary(full_model)

resids = full_model$residuals
fitted = full_model$fitted.values

resids_plot = plot(fitted, resids, xlab = 'Fitted Sleep Efficiency', ylab =
'Residuals')
title('Plot of Residuals vs Fitted Sleep Efficiency for Full Model')

# Partial R squared and VIF values for the initial full model
vif_full_dataset <- dataset[-7][-1]
age_pr2_full <- summary(lm(Age ~ ., data = vif_full_dataset))$r.squared
bedtime_pr2_full <- summary(lm(Bedtime ~ ., data =
vif_full_dataset))$r.squared
wakeuptime_pr2_full <- summary(lm(Wakeup.time ~ ., data =
vif_full_dataset))$r.squared
duration_pr2_full <- summary(lm(Sleep.duration ~ ., data =
vif_full_dataset))$r.squared
REM_pr2_full <- summary(lm(REM.sleep.percentage ~ ., data =
vif_full_dataset))$r.squared
```

```
deep_pr2_full <- summary(lm(Deep.sleep.percentage ~ ., data =
vif_full_dataset))$r.squared
light_pr2_full <- summary(lm(Light.sleep.percentage ~ ., data =
vif_full_dataset))$r.squared
awakening_pr2_full <- summary(lm(Awakenings ~ ., data =
vif_full_dataset))$r.squared
caffeine_pr2_full <- summary(lm(Caffeine.consumption ~ ., data =
vif_full_dataset))$r.squared
alcohol_pr2_full <- summary(lm(Alcohol.consumption ~ ., data =
vif_full_dataset))$r.squared
exercise_pr2_full <- summary(lm(Exercise.frequency ~ ., data =
vif_full_dataset))$r.squared
pr2_full_model <- c(age_pr2_full, bedtime_pr2_full, wakeuptime_pr2_full,
duration_pr2_full, REM_pr2_full,
                    deep_pr2_full, light_pr2_full, awakening_pr2_full,
caffeine_pr2_full, alcohol_pr2_full,
                    exercise_pr2_full)
vif_full_model <- 1 / (1 - pr2_full_model)


#Backwards Selection
ver2 = lm(Sleep.efficiency ~ Age + Sleep.duration + REM.sleep.percentage +
Deep.sleep.percentage + Light.sleep.percentage + Awakenings +
Caffeine.consumption +
            Alcohol.consumption + Smoking.status + Exercise.frequency +
Bedtime + Wakeup.time, data = dataset)
summary(ver2)

ver3 = lm(Sleep.efficiency ~ Age + Sleep.duration + REM.sleep.percentage +
Deep.sleep.percentage + Light.sleep.percentage + Awakenings +
Caffeine.consumption +
            Alcohol.consumption + Smoking.status + Exercise.frequency +
Wakeup.time, data = dataset)
summary(ver3)

ver4 = lm(Sleep.efficiency ~ Age + Sleep.duration + REM.sleep.percentage +
Deep.sleep.percentage + Light.sleep.percentage + Awakenings +
Caffeine.consumption +
            Alcohol.consumption + Smoking.status + Exercise.frequency, data =
dataset)
summary(ver4)

ver5 = lm(Sleep.efficiency ~ Age + REM.sleep.percentage +
Deep.sleep.percentage + Light.sleep.percentage + Awakenings +
Caffeine.consumption +
            Alcohol.consumption + Smoking.status + Exercise.frequency, data =
dataset)
summary(ver5)


#To choose from REM, deep, light sleep percentage
mod_sub = regsubsets(Sleep.efficiency ~., data = dataset, method =
'exhaustive')
ss = summary(mod_sub)
ss$which
ss$adjr2
ss$cp
ss$rsq

plot(x = c(1, 2, 3, 4, 5, 6, 7, 8, 9), ss$adjr2, xlab = 'Number of
Parameters', ylab = "Adjusted R2")
title("Adjusted R2 for Model Selection")

plot(x = c(1, 2, 3, 4, 5, 6, 7, 8, 9), ss$cp, xlab = 'Number of Parameters',
ylab = "Mallow's Cp")
abline(0, 1)
title("Mallow's Cp Plot for Model Selection")


optimized = lm(Sleep.efficiency ~ Age + Light.sleep.percentage + Awakenings +
Caffeine.consumption +
            Alcohol.consumption + Smoking.status + Exercise.frequency, data =
dataset)
summary(optimized)

optimized_resids = optimized$residuals
optimized_fitted = optimized$fitted.values

resids_plot = plot(optimized_fitted, optimized_resids, xlab = 'Fitted Sleep
Efficiency', ylab = 'Residuals')
title('Plot of Residuals vs Fitted Sleep Efficiency for Optimized Model')


qq_plot = qqnorm(optimized_resids, ylab = 'Optimized Model Residuals Sample
Quantiles')
qqline(optimized_resids)
```

```
# Partial R squared and VIF values for the intermediate model without
interaction
vif_int_dataset <- dataset[-9][-8][-7][-6][-5][-4][-3][-1]
age_pr2_int <- summary(lm(Age ~ ., data = vif_int_dataset))$r.squared
light_pr2_int <- summary(lm(Light.sleep.percentage ~ ., data =
vif_int_dataset))$r.squared
awakening_pr2_int <- summary(lm(Awakenings ~ ., data =
vif_int_dataset))$r.squared
caffeine_pr2_int <- summary(lm(Caffeine.consumption ~ ., data =
vif_int_dataset))$r.squared
alcohol_pr2_int <- summary(lm(Alcohol.consumption ~ ., data =
vif_int_dataset))$r.squared
exercise_pr2_int <- summary(lm(Exercise.frequency ~ ., data =
vif_int_dataset))$r.squared
pr2_int_model <- c(age_pr2_int, light_pr2_int, awakening_pr2_int,
caffeine_pr2_int,
                   alcohol_pr2_int, exercise_pr2_int)
vif_int_model <- 1 / (1 - pr2_int_model)



#Sample calculation for VIF of age
mod_justage = lm(Age ~ Light.sleep.percentage + Awakenings +
Caffeine.consumption +
                   Alcohol.consumption + Smoking.status + Exercise.frequency,
data = dataset)
optimized_agevif = 1 / (1 - summary(mod_justage)$r.squared)


## Optimized model including interaction terms
optimized_int = lm(Sleep.efficiency ~ Age + Light.sleep.percentage +
Awakenings + Caffeine.consumption +
                     Alcohol.consumption + Smoking.status + Exercise.frequency
+
                     Awakenings:Alcohol.consumption +
Age:Light.sleep.percentage, data = dataset)
summary(optimized_int)
```

```
optimized_int_resids = optimized_int$residuals
optimized_int_fitted = optimized_int$fitted.values

resids_int_plot = plot(optimized_int_fitted, optimized_int_resids, xlab =
'Fitted Sleep Efficiency', ylab = 'Residuals')
title('Plot of Residuals vs Fitted Sleep Efficiency for Optimized Model with
Interactions')

qq_int_plot = qqnorm(optimized_int_resids, ylab = 'Optimized Model With
Interactions Residuals Sample Quantiles')
qqline(optimized_int_resids)


#Leverage and Outlier Checks
leverages <- as.data.frame(hatvalues(optimized_int))
cookdstes <- as.data.frame(cooks.distance(optimized_int))
standard_res <- rstandard(optimized_int)
```