Investigating Factors Affecting Sleep Efficiency

1 Introduction

Motivation

Understanding the quality of our sleep goes beyond mere hours spent in bed. Sleep efficiency, quantified as the percentage of time spent asleep while in bed, is calculated by dividing the amount of time spent asleep by the total amount of time in bed (measured in minutes). As suggested by professionals, normal sleep efficiency is considered to be 85% or higher (Sleep Efficiency, n.d.). Although some individuals are naturally better at falling asleep than others, numerous factors potentially affect the efficiency of sleep. Diet, exercise, stress, anxiety, age, lifestyle and other factors have emerged from previous studies as apparent contributors to the ability to fall asleep (Rangtell, 2020). In our fast-paced modern landscape, more and more people start paying attention to sleep efficiency because of its significance in maintaining overall health and well-being. Therefore, understanding the factors influencing sleep efficiency is crucial in addressing the concerns about sleep disorders. In this report, we want to explore the relationships between sleep efficiency and factors that may have effects on it.

About the Dataset

We have chosen to use the Sleep Efficiency Dataset sourced from Kaggle (Sleep Efficiency Dataset, n.d.), which contains a total of 452 rows of data, with 65 missing values. This data was obtained from a study conducted by a group of artificial intelligence engineering students from ENSIAS in Moraco. There are 15 variables in this dataset:

1. **ID**: the unique identifier for each test subject;
2. **Age**: the age of the test subjects, measured in years, continuous variable, ranged from 6 to 69;
3. **Gender**: the gender of the test subjects, a categorical variable with two categories: male and female;
4. **Bedtime**: the time the test subject goes to bed each night, date variable, measured exact time (format: year-month-day hour-minute-second);
5. **Wakeup time**: the time the test subject wakes up each morning, date variable, measured exact time (format: year-month-day hour-minute-second);
6. **Sleep duration**: the total amount of time the test subject slept, measured in hours, continuous variable, ranged from 5 to 10;
7. **Sleep efficiency**: a measure of the proportion of time in bed spent asleep, a continuous variable, ranging from 0.5 to 0.99, serves as the response variable in this study;
8. **REM sleep percentage**: the percentage of total sleep time spent in REM sleep, a continuous variable, ranging from 15% to 30%;
9. **Deep sleep percentage**: the percentage of total sleep time spent in deep sleep, a continuous variable, ranging from 18% to 75%;
10. **Light sleep percentage**: the percentage of total sleep time spent in light sleep, a continuous variable, ranging from 7% to 63%;
11. **Awakenings**: the number of times the test subjects wake up during the night, a continuous variable, ranging from 0 to 4;
12. **Caffeine consumption**: the amount of caffeine consumed in the 24 hours prior to bedtime, measured in mg, a continuous variable, ranging from 0 to 200;
13. **Alcohol consumption**: the amount of alcohol consumed in the 24 hours prior to bedtime, measured in oz, a continuous variable, ranging from 0 to 5;

14. **Smoking status**: whether the test subject smokes, a categorical variable with two categories: true and false;

15. **Exercise frequency**: the number of times the test subject exercises each week, a continuous variable, ranging from 0 to 5.

Research Question

We want to investigate, to what extent sleep efficiency relates to a set of variables including age, gender, sleep duration, REM sleep percentage, deep sleep percentage, awakenings, caffeine consumption, alcohol consumption, smoking status, and exercise frequency. In addition, we want to explore and select the linear model that can best explain the relationships between sleep efficiency and the factors.

By generating and exploring multiple linear regression (MLR) models based on these variables, we can offer some valuable insights to guide mental decisions and the allocation of healthcare resources.

2 Exploratory Data Analysis

Data Cleansing and Wrangling

First, since there are 65 observations containing missing values, we remove these rows in this study. Bedtime and Wakeup time are two timestamp variables, in the format of year-month-day hour-minute-second, which are difficult to work with, so we remove these two variables from the dataset. Subsequent cleaning and data wrangling includes the change in the data type of categorical variables from "character" to "factor" as factors have predefined levels which is much easier to work with than characters (this can be done by the "as_factor" function in R), thus, we convert the data type of "gender" and "smoking status" from "characters" to "factors". We also remove the "ID" column which is just an index of each test subject and is redundant in the exploration of our dataset. The clean data now has 388 rows with 12 variables. The clean data is shown below:

Age	Gender	Sleep.duration	Sleep.efficiency	REM.sleep.percentage	Deep.sleep.percentage
65	Female	6	0.88	18	70
69	Male	7	0.66	19	28
40	Female	8	0.89	20	70
40	Female	6	0.51	23	25
57	Male	8	0.76	27	55
27	Female	6	0.54	28	25

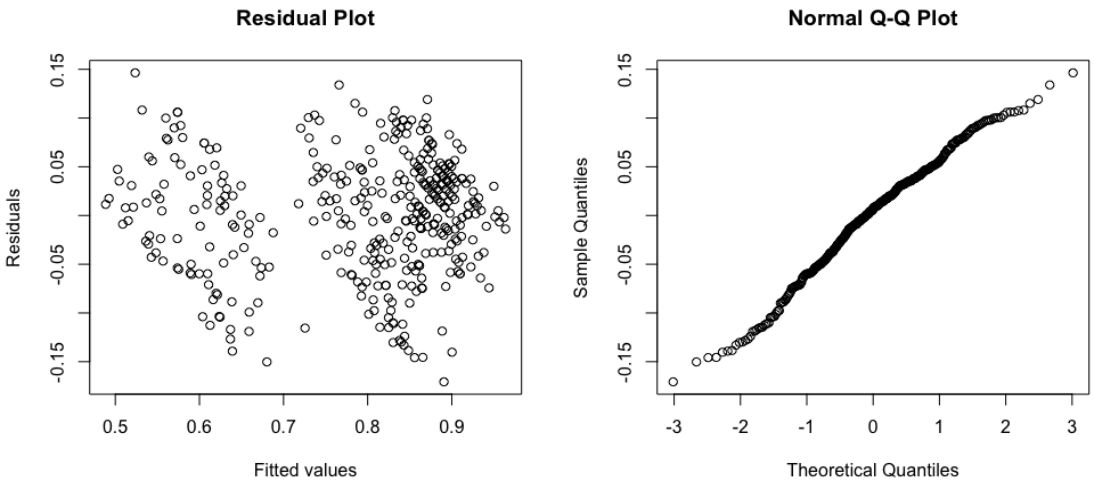
Light.sleep.percentage	Awakenings	Caffeine.consumption	Alcohol.consumption	Smoking.status	Exercise.frequency
12	0	0	0	Yes	3
53	3	0	3	Yes	3
10	1	0	0	No	3
52	3	50	5	Yes	1
18	3	0	3	No	3
47	2	50	0	Yes	1

Initial Model Fitting

To start, we constructed an initial linear model using all of the above variables as explanatory variables, and sleep efficiency as the response variable. The linear model equation is as follows:

$$\begin{aligned} sleep\ efficiency = \beta_0 + \beta_1 age + \beta_2 gender_{male} + \beta_3 sleep\,duration + \beta_4 REM\ sleep\,pct + \\ \beta_5 Deep\,sleep\,pct + \beta_6 Light\,sleep\,pct + \beta_7 Awakenings + \beta_8 Caffeine\ consumption \\ + \beta_9 Alcohol\,consumption + \beta_{10} Smoking_{yes} + \beta_{11} Exercise\ efficiency \end{aligned}$$

We obtained the residual plots:



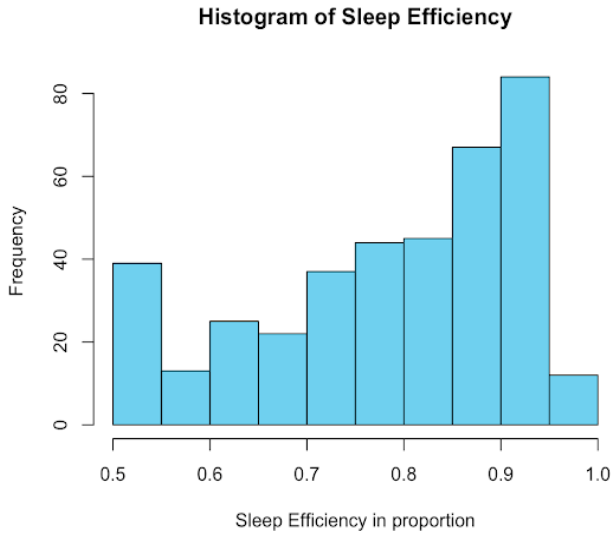
From the above residual plot and normal Q-Q plot, the residuals are normally distributed which does not violate our assumption of fitting the linear regression model. Then, we want to check the summary of the initial model:

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	0.3489326880	0.0413857549	8.4312268	7.348402e-16
Age	0.0009514933	0.0002427368	3.9198566	1.052276e-04
GenderMale	0.0014379397	0.0069492648	0.2069197	8.361843e-01
Sleep.duration	0.0017481986	0.0035195519	0.4967106	6.196827e-01
REM.sleep.percentage	0.0066732949	0.0009385640	7.1101121	5.855199e-12
Deep.sleep.percentage	0.0055670615	0.0002376502	23.4254414	1.522018e-75
Light.sleep.percentage	NA	NA	NA	NA
Awakenings	-0.0318783398	0.0025248580	-12.6257951	1.010137e-30
Caffeine.consumption	0.0002411676	0.0001131318	2.1317400	3.367431e-02
Alcohol.consumption	-0.0061302708	0.0021173665	-2.8952336	4.009340e-03
Smoking.statusYes	-0.0460091746	0.0067857893	-6.7802244	4.646933e-11
Exercise.frequency	0.0063931820	0.0022972803	2.7829350	5.658047e-03

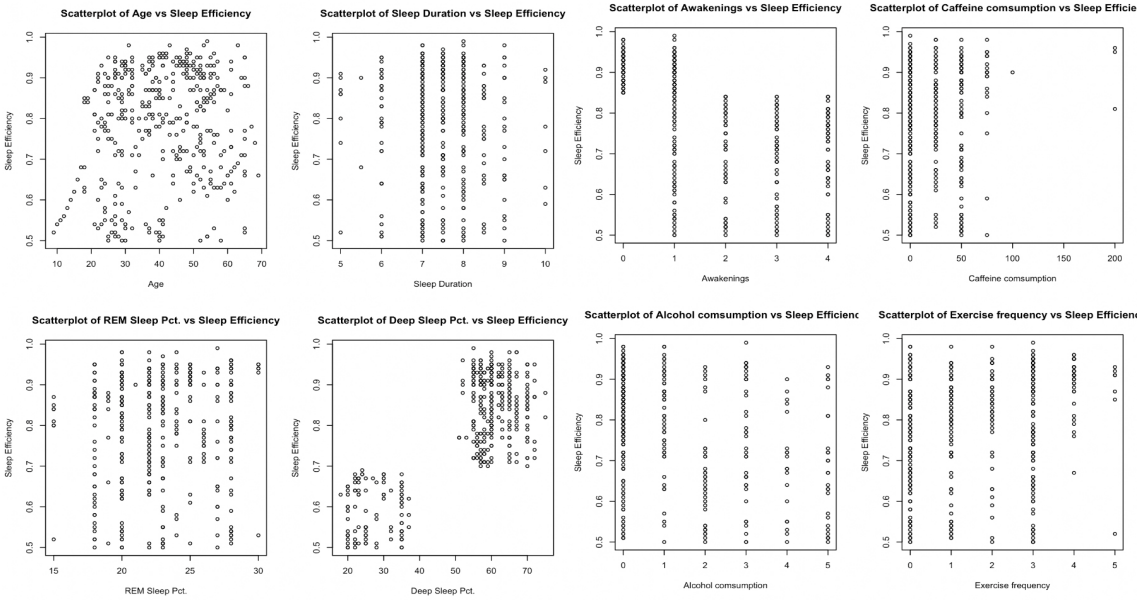
We notice that there's one coefficient (for light sleep percentage) is not defined due to singularities. This occurs due to perfect multicollinearity or when the predictor variables are not orthogonal, resulting in a singular or nearly singular matrix. By checking the correlation between the explanatory variables, we found that light sleep percentage is highly correlated with deep sleep percentage, which is -0.975. As such, we can conclude that perfect multicollinearity exists. Thus, we will remove this variable from subsequent analysis. Now, we have a clean dataset consisting of 11 variables (1 response variable and 10 explanatory variables).

Data Visualization

Since sleep efficiency serves as the response variable in this study, we first want to visualize its distribution using a histogram.



As observed, the histogram shows an asymmetric distribution and a bit left-skewed. The sleep efficiency ranges from 0.5 to 1.0 with most individuals having sleep efficiency of 0.8 0.9. Next, we want to visualize the relationship between sleep efficiency and each continuous explanatory variable using scatterplots and a correlation table.



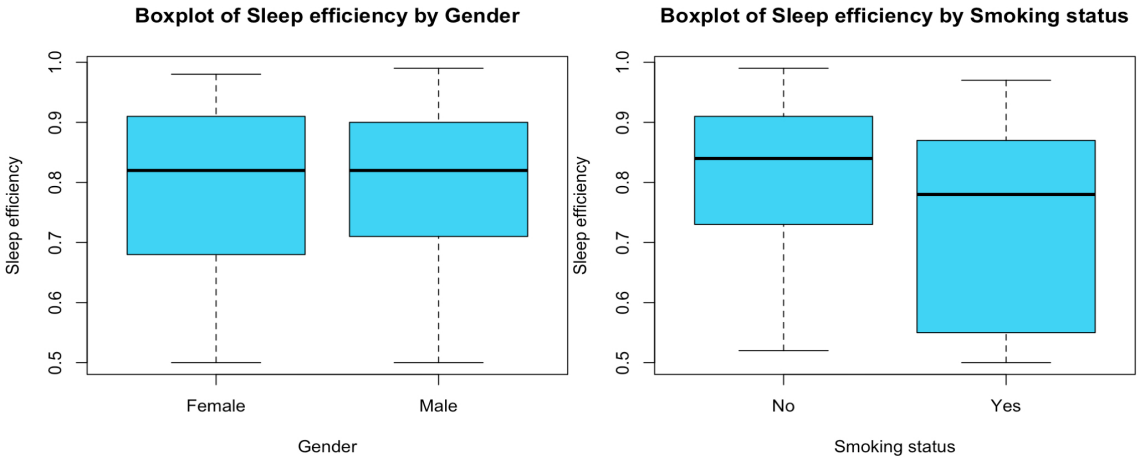
	Age	Sleep.duration	Sleep.efficiency	REM.sleep.percentage	Deep.sleep.percentage	Awakenings	Caffeine.consumption	Alcohol.consumption	Exercise.frequency
Age	1.000	-0.066	0.124	0.015	0.058	-0.004	-0.169	0.069	0.071
Sleep.duration	-0.066	1.000	-0.019	-0.015	-0.035	-0.009	-0.031	-0.048	-0.047
Sleep.efficiency	0.124	-0.019	1.000	0.064	0.789	-0.568	0.071	-0.397	0.266
REM.sleep.percentage	0.015	-0.015	0.064	1.000	-0.186	-0.025	0.114	-0.040	0.044
Deep.sleep.percentage	0.058	-0.035	0.789	-0.186	1.000	-0.327	-0.025	-0.375	0.172
Awakenings	-0.004	-0.009	-0.568	-0.025	-0.327	1.000	-0.113	0.210	-0.231
Caffeine.consumption	-0.169	-0.031	0.071	0.114	-0.025	-0.113	1.000	-0.098	-0.083
Alcohol.consumption	0.069	-0.048	-0.397	-0.040	-0.375	0.210	-0.098	1.000	0.004
Exercise.frequency	0.071	-0.047	0.266	0.044	0.172	-0.231	-0.083	0.004	1.000

From the above scatterplots, we observe a slight positive linear association between the explanatory variable Age and the response variable Sleep Efficiency, further supported by the correlation coefficient of 0.124. The scatterplot with Sleep Duration appears to be largely concentrated nearer the centre, and the correlation coefficient is low at -0.019. There also seems to be a moderate negative linear association between Awakenings and the response variable, with a correla-

tion coefficient of -0.568. We can also deduce a slight linear association in the caffeine consumption plot, which has a correlation coefficient of 0.071. In the second row of scatter plots, REM Sleep Percentage looks to be similarly concentrated nearer the centre of the plot. It has a correlation coefficient of 0.064. In the deep sleep percentage, we can detect two distinct groups of data points, illustrating a strong positive association, further supported by a correlation coefficient of 0.789. The Alcohol Consumption plot has points more concentrated towards the left suggesting a negative association with the response variable, supported by a correlation coefficient of -0.397. Exercise frequency appears to have a slight positive linear association, with a correlation coefficient of 0.266.

It is also worth noting that in the correlation matrix, none of the pairs of explanatory variables exhibit a high correlation (coefficient), suggesting that perfect multicollinearity is unlikely to be a significant issue.

We also want to visualize the relationships between sleep efficiency and categorical variables, gender and smoking status, using a boxplot.



From the left boxplot, we can see that female sleep efficiency has a slightly wider range than male sleep efficiency but their median are almost the same. This observation indicates gender may not have a significant effect on sleep efficiency. Observing the right boxplot, the smoking group generally has lower sleep efficiency than the non-smoking group. The median sleep efficiency of the smoking group is lower than the non-smoking group and its range is wider than that of the non-smoking group. This finding implies that smoking status might be a significant factor that affects sleep efficiency. We will explore and confirm our thoughts later in this report.

Data Summary

We summarise the statistics including mean, median, standard deviation, variance, min and max of all continuous variables as shown below:

	name	mean	sd	median	variance	max	min
1	Age	40.8298969	13.4031865	41.00	179.64540877	69.00	9.0
2	Alcohol.consumption	1.1469072	1.6127931	0.00	2.60110152	5.00	0.0
3	Awakenings	1.6185567	1.3559577	1.00	1.83862117	4.00	0.0
4	Caffeine.consumption	22.6804124	28.9975661	0.00	840.85884014	200.00	0.0
5	Deep.sleep.percentage	52.8221649	15.5715179	58.00	242.47216894	75.00	18.0
6	Exercise.frequency	1.7577320	1.4478236	2.00	2.09619329	5.00	0.0
7	REM.sleep.percentage	22.6804124	3.4305103	22.00	11.76840086	30.00	15.0
8	Sleep.duration	7.4510309	0.8834826	7.50	0.78054157	10.00	5.0
9	Sleep.efficiency	0.7892526	0.1357064	0.82	0.01841624	0.99	0.5

We also summarised the proportion of each level in the categorical variables below:

	Female	Male
gender	194.0	194.0
genderprop	0.5	0.5

	No	Yes
smoke	255.000	133.000
smokeprop	0.657	0.343

From the above, it is worth noting that there is an equal number of female and male participants present in this dataset. Additionally, there is a larger proportion of the participants who don't smoke than those who do.

3 Methods & Plan

After excluding the variable 'light sleep percentage' due to its multicollinearity issues, 10 explanatory variables remain. We prefer a model with fewer variables to prevent overfitting and maintain model simplicity. To produce a more interpretable linear model with fewer covariates, forward selection is applied as our model selection method. It begins with the empty model. A new variable with the most statistical significance is added to the current model at each step. This step is repeated until no variables in the rest set are significant.

Several information criteria are used for variable selection, including the adjusted R^2, Mallows' Cp statistic, and the Bayesian Information Criterion (BIC). Each of these criteria evaluates the goodness of fit of a model. In short, the adjusted R^2 measures the proportion of the variation explained by the regression model. The Cp estimates the size of the bias of the model while the BIC estimates the likelihood of a model to predict. Therefore, we generally prefer models with a higher adjusted R^2 and lower values of Mallows' Cp and BIC.

Following our plan, we implement the forward selection by "regsubsets" in R. The results of the forward selection based on 10 explanatory variables are as follows:

	(Intercept)	Age	GenderMale	Sleep.duration	REM.sleep.percentage	Deep.sleep.percentage	Awakenings	Caffeine.consumption	Alcohol.consumption	Smoking.statusYes	Exercise.frequency
1	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
2	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE
3	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE
4	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	FALSE
5	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	FALSE
6	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	FALSE
7	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE
8	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
9	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
10	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE

Now, we have the best variable combinations for each size of the model (from size 2 to 11, size equals to number of parameters in the model). Next, we will compute the adjusted R^2, Cp and BIC to compare those different-sized models and choose the one we believe can best explain our data.

n_input_variables	ADJ.R2	Cp	BIC
<int>	<dbl>	<dbl>	<dbl>
1	0.6216810	345.751259	-366.2247
2	0.7286273	140.103176	-490.1810
3	0.7624349	75.872469	-536.8530
4	0.7855301	32.482334	-571.5856
5	0.7919887	21.081537	-578.5028
6	0.7956678	15.037225	-580.4828
7	0.7989668	9.751538	-581.8570
8	0.8007956	7.283339	-580.4641
9	0.8003959	9.042816	-574.7505
10	0.7998892	11.000000	-568.8336

The table shows the criteria (adjusted R^2, Cp, BIC) for each model. We can see that the model with 8 variables has the largest value of adjusted R^2 and a relatively low BIC compared to the other models. By checking the Akaike Information Criterion (AIC) of this model, we observe a relatively low value, approximately -1064. This indicates that the model experiences minimal information loss.

Since β_0 is counted in the number of parameters, the model with 8 variables has 9 parameters, the Cp statistic for this model is around 7.28, which is the smallest and closely aligns with the number of parameters in the model.

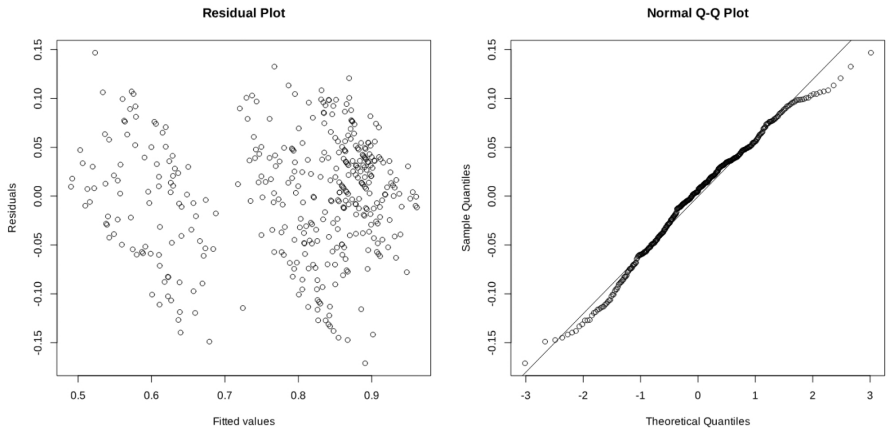
We can observe that the Cp statistic for the model with 9 variables is closer to its number of parameters. However, this model exhibits a higher AIC and lower adjusted R^2. Additionally, the additional variable (sleep duration) introduced in the 9-variable model is statistically insignificant, as indicated by a p-value of 0.62. Thus, we can conclude that the linear regression model with 8 variables outperforms the one with 9 variables.

Based on these observations, we have decided to select the variables in the 8-variable model as our explanatory variables. Having decided to choose the model with 8 variables, we have identified the following variables that are included in this model: **Age, REM.sleep.percentage, Deep.sleep.percentage, Awakenings, Caffeine.consumption, Alcohol.consumption, Smoking.statusYes and Exercise.frequency**.

Below is the summary of the inference model with the 8 selected explanatory variables. This model exhibits a high adjusted R^2 of approximately 0.801, indicating a strong fit. While most parameters are statistically significant, Caffeine.consumption and Exercise.frequency stand out with p-values of 0.03477 and 0.00396, respectively. Despite their relatively higher p-values, they are still significant at 5% significance level, so these parameters will be retained in the inference model. Their inclusion, as determined by the forward selection, is crucial for maintaining the overall significance of other parameters and preserving the high adjusted R^2 value.

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	0.3634104710	0.0298043725	12.193193	4.450181e-29
Age	0.0009561706	0.0002348634	4.071177	5.696450e-05
REM.sleep.percentage	0.0066437927	0.0009314321	7.132879	5.021654e-12
Deep.sleep.percentage	0.0055625145	0.0002365262	23.517537	4.635769e-76
Awakenings	-0.0318493010	0.0024973038	-12.753475	3.042330e-31
Caffeine.consumption	0.0002333085	0.0001101231	2.118615	3.477310e-02
Alcohol.consumption	-0.0062024223	0.0021081747	-2.942082	3.460505e-03
Smoking.statusYes	-0.0457656360	0.0066572544	-6.874551	2.571124e-11
Exercise.frequency	0.0064591314	0.0022282587	2.898735	3.964649e-03

The following graphs are the residual plot and Normal Q-Q plot of the inference model. In the Residual Plot, the scatter of residuals appears to be random without an obvious pattern. The variance of the residuals seems to be constant across the full range of fitted values, indicating homoscedasticity. Regarding the Normal Q-Q plot, it shows that the residuals from the inference model closely follow a normal distribution, as most points lie along the line. In short, these two graphs suggest that the inference model satisfies the assumption of linear regression.



4 Conclusion

From the previous section, the model with the best fit has 8 variables:

$$\hat{sleep\ efficiency} = 0.363 + 0.001 * age + 0.007 * REM\ sleep\ pct + 0.006 * deep\ sleep\ pct - 0.032 * awakenings + 0.0002 * caffeine\ consumption - 0.006 * alcohol\ consumption - 0.046 * smoking_{yes} + 0.006 * exercise\ efficiency$$

To address our research question, we can conclude that the following variables are significant factors affecting sleep efficiency based on the data: Age, REM sleep percentage, Deep sleep percentage, Awakenings, Caffeine consumption, Alcohol consumption, Smoking and Exercise frequency. More specifically, being male, having a higher percentage of REM sleep time and Deep sleep time, and greater Caffeine intake and Exercise frequency are associated with increased sleep efficiency, whereas frequent awakenings, higher alcohol intake, and smoking are associated with decreased sleep efficiency.

With a clearer understanding of the factors significantly influencing sleep efficiency and the intricate relationships among these variables, we are equipped with valuable insights to navigate the complexities of addressing sleep disorders. Armed with this knowledge, healthcare professionals, researchers, and individuals alike can devise more precise and effective approaches, ranging from lifestyle modifications to personalized treatments, fostering better sleep habits and, consequently, enhanced health outcomes. Understanding these connections creates opportunities for more research and the development of comprehensive solutions that emphasize the importance of sleep for overall health and well-being.

5 References

1. Sleep Efficiency. (n.d.). Hypersomnia Foundation. https://www.hypersomniafoundation.org/
2. Rangtell, Frida. (2020, December 29). *What determines our sleep efficiency?* Sleep cycle.https://www.sleepcycle.com/sleep-habits-and-health/what-determines-sleep-efficiency/
3. Sleep Efficiency Dataset. (n.d.). Kaggle. https://www.kaggle.com/datasets/equilibriumm/sleep-efficiency/data

6 Appendix

```
data = read.csv("Sleep_Efficiency.csv",header=T)
sum(is.na(data))
data$Exercise.frequency = as.integer(data$Exercise.frequency)

# remove bedtime and wake up time
dat <- data[, -c(4, 5)]

# remove NA
dat <- na.omit(dat)

# convert to factor
dat$Gender = as.factor(dat$Gender)
dat$Smoking.status=as.factor(dat$Smoking.status)

# remove ID
dat <- dat[, -c(1)]

numdat = dat[,-c(2,11)]
cor.table=round(cor(numdat),3)

# initial model fitting
ini.model = lm(Sleep.efficiency~., data=dat)
residuals <- residuals(ini.model)

# Create a residual plot
par(mfrow = c(1, 2))
plot(predict(ini.model), residuals,
     xlab = "Fitted values", ylab = "Residuals",
     main = "Residual Plot")

# create QQ normal plot
qqnorm(residuals)
qqline(residuals)

summary(ini.model)
par(mfrow = c(1, 1))

# remove light sleep pct column
sleep = dat[, -c(7)]

# histogram of response
hist(sleep$Sleep.efficiency,
     main = "Histogram of Sleep Efficiency",
     xlab = "Sleep Efficiency in proportion", ylab = "Frequency",
```

```
     col = "skyblue", border = "black")

# create ggpairs for continuous variables
par(mfrow = c(2, 4))

plot(x = sleep$Age, y = sleep$Sleep.efficiency,
     main = "Scatterplot of Age vs Sleep Efficiency",
     xlab = "Age", ylab = "Sleep Efficiency",cex=0.8)

plot(x = sleep$Sleep.duration, y = sleep$Sleep.efficiency,
     main = "Scatterplot of Sleep Duration vs Sleep Efficiency",
     xlab = "Sleep Duration", ylab = "Sleep Efficiency",cex=0.8)

plot(x = sleep$REM.sleep.percentage, y = sleep$Sleep.efficiency,
     main = "Scatterplot of REM Sleep Pct. vs Sleep Efficiency",
     xlab = "REM Sleep Pct.", ylab = "Sleep Efficiency",cex=0.8)

plot(x = sleep$Deep.sleep.percentage, y = sleep$Sleep.efficiency,
     main = "Scatterplot of Deep Sleep Pct. vs Sleep Efficiency",
     xlab = "Deep Sleep Pct.", ylab = "Sleep Efficiency",cex=0.8)

plot(x = sleep$Awakenings, y = sleep$Sleep.efficiency,
     main = "Scatterplot of Awakenings vs Sleep Efficiency",
     xlab = "Awakenings", ylab = "Sleep Efficiency",cex=0.8)

plot(x = sleep$Caffeine.consumption, y = sleep$Sleep.efficiency,
     main = "Scatterplot of Caffeine comsumption vs Sleep Efficiency",
     xlab = "Caffeine comsumption", ylab = "Sleep Efficiency",cex=0.8)

plot(x = sleep$Alcohol.consumption, y = sleep$Sleep.efficiency,
     main = "Scatterplot of Alcohol comsumption vs Sleep Efficiency",
     xlab = "Alcohol comsumption", ylab = "Sleep Efficiency",cex=0.8)

plot(x = sleep$Exercise.frequency, y = sleep$Sleep.efficiency,
     main = "Scatterplot of Exercise frequency vs Sleep Efficiency",
     xlab = "Exercise frequency", ylab = "Sleep Efficiency",cex=0.8)

par(mfrow = c(1, 1))

# create boxplot for categorical variables
boxplot(Sleep.efficiency ~ Gender, data = sleep,
        main = "Boxplot of Sleep efficiency by Gender",
        xlab = "Gender", ylab = "Sleep efficiency",
        col = "skyblue", border = "black")

boxplot(Sleep.efficiency ~ Smoking.status, data = sleep,
        main = "Boxplot of Sleep efficiency by Smoking status",
        xlab = "Smoking status", ylab = "Sleep efficiency",
        col = "skyblue", border = "black")

# data summary for continuous variable
library(tidyr)
sleep_summary <- sleep %>% select(c(-Gender,-Smoking.status)) %>%
  pivot_longer(cols = everything()) %>% group_by(name) %>%
  summarise(
    mean = mean(value,na.rm = T), sd = sd(value,na.rm = T),
```

```
    median = median(value, na.rm = T), variance = var(value, na.rm = T),
    max = max(value, na.rm = T), min = min(value, na.rm = T))

# data summary for categorical variable
gender=table(sleep$Gender)
genderprop=prop.table(gender)

smoke=table(sleep$Smoking.status)
smokeprop=prop.table(smoke)

# model selection
library(leaps)
s <- regsubsets(Sleep.efficiency ~ ., data=sleep, nvmax=10, method="forward")
ss=summary(s)

# cp, bic, r^2
forward_summary_df <- tibble(
  n_input_variables = 1:10,
  adj.r2=ss$adjr2, Cp = ss$cp, BIC = ss$bic)
forward_summary_df

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

model_with_9var <- lm( Sleep.efficiency~Age+REM.sleep.percentage+Deep.sleep.percentage
    +Awakenings+Caffeine.consumption+ Alcohol.consumption+ Smoking.status
    + Exercise.frequency+Sleep.duration ,data=sleep)
summary(model_with_9var)

AIC(inference_model) #-1063.89288498968
AIC(model_with_9var) #-1062.14031891154

residuals <- residuals(inference_model)

# Create a residual plot
plot(predict(inference_model), residuals,
     xlab = "Fitted values", ylab = "Residuals", main = "Residual Plot")

# create QQ normal plot
qqnorm(residuals)
qqline(residuals)
```