# Exploring Predictive Factors for Sleep Efficiency: A Comprehensive Analysis and Linear Regression Modeling Approach

**STAT 306 Group A4:**

Minh Vu (#33077769)
Jessie Shang (#82738477)
Carson Lu (#37798238)
Lucy Vincent-Smith (#14810148)

# 1. Introduction

### 1.1 Dataset Source

The dataset referenced for this project is the Sleep Efficiency Dataset from Kaggle, linked here: Kaggle Sleep Efficiency Dataset.

### 1.2 Dataset Description

The dataset contains information about 452 human test subjects and their sleep patterns. Each subject's age, gender, bedtime and wakeup times, sleep duration, number of awakenings, caffeine and alcohol consumption, smoking status, and exercise frequency are recorded. REM, Deep, and Light sleep percentages are also recorded, allowing for calculation of a “Sleep Efficiency” variable for each subject. A subject is identified by a unique "Subject ID".

### 1.3 Motivation and Research Question

Given that various sleep stages significantly influence daily functioning across multiple facets of life, it's crucial to delve into sleep efficiency as a vital parameter. Rapid Eye Movement (REM) sleep aids in the consolidation and processing of fresh information within the brain, fostering enhancements in mental focus and the regulation of mood. Additionally, deep sleep plays a similarly pivotal role in promoting brain health and maintaining hormonal equilibrium. Therefore, our study focuses on investigating sleep efficiency, as it serves as a valuable metric encompassing the effectiveness of these essential sleep stages.

There is evidence to suggest that some factors play a significant role in an individual’s achievement of good sleep efficiency and quality. Our group would like to study **which factors have the most significant influence on an individual’s sleep efficiency, and how can these variables be modeled to predict sleep efficiency of a new test subject?** More specifically, this study aims to explore the relationship between alcohol and caffeine consumption, exercise frequency, smoking status, and gender, and how these factors collectively influence sleep efficiency.

### 1.4 Variables Used

The response variable in this analysis is sleep efficiency, measured as the proportion of time in bed spent asleep.

The explanatory variables examined are alcohol consumption, caffeine consumption, exercise frequency, smoking status, and gender. Alcohol consumption, caffeine consumption, and exercise frequency are quantitative, with units of oz, mg, and frequency, respectively. Smoking status and gender are categorical and are both binary (true/false and male/female, respectively) due to the way in which data was recorded.

The variables age, bedtime, wakeup time, sleep duration, and REM, deep, and light sleep percentages were excluded from models, as the other chosen variables are hypothesized to be better predictors than age, while the sleep-related data will be captured indirectly by the response variable (sleep efficiency).

## 2. Discussion

### 2.1 Data Cleaning and Exploration

#### 2.1.1 Cleaning

Table 1 shows a brief description and missing value count of each variable used in our analysis.

*Table 1. Variables used in analysis.*

| Variable Name | Unit | Description | Categorical or Quantitative | Missing Value Count |
|---|---|---|---|---|
| Gender | male/female | Gender of the test subject | Categorical | 0 |
| Sleep efficiency | proportion | measure of proportion of time in bed spent asleep | Quantitative | 0 |
| Caffeine consumption | mg | amount of caffeine consumed in 24 hours prior to bedtime | Quantitative | 25 |
| Alcohol consumption | oz | amount of alcohol consumed in 24 hours prior to bedtime | Quantitative | 16 |
| Smoking status | true/false | whether or not the test subject smokes | Categorical | 0 |
| Exercise frequency | times/week | number of times test subject exercises each | Quantitative | 6 |

| | | | | |
|---|---|---|---|---|
| | | week | | |

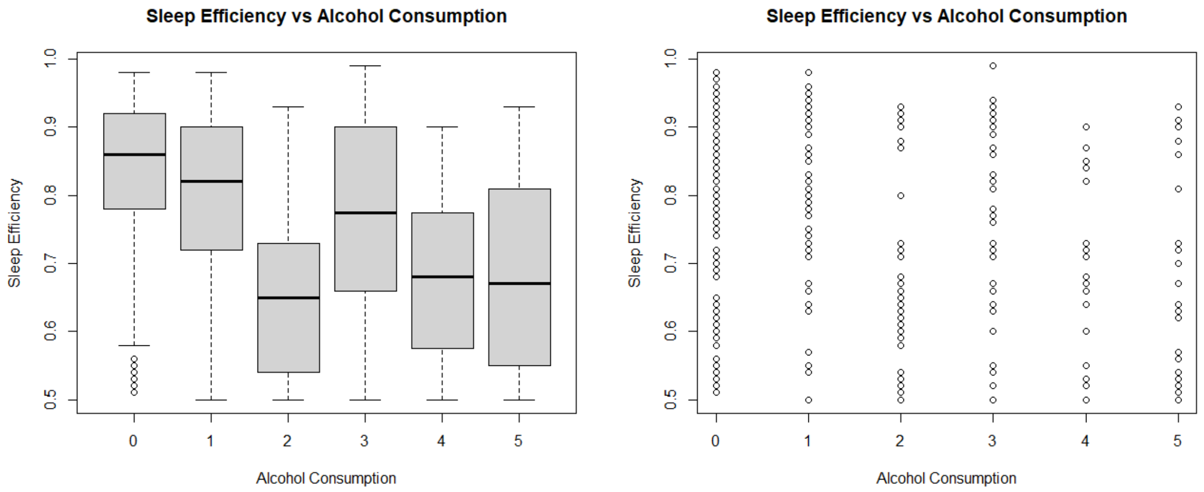
The data was first cleaned by removing extraneous variables and any rows with missing data. After handling missing values, the dataset contains 407 examples. Smoking status and gender were converted to factors to use as categorical variables. These changes are executed in the attached R code.

The variables age, bedtime, wakeup time, sleep duration, and REM, deep, and light sleep percentages were excluded from model analysis moving forward.

### 2.1.2 Visualizations

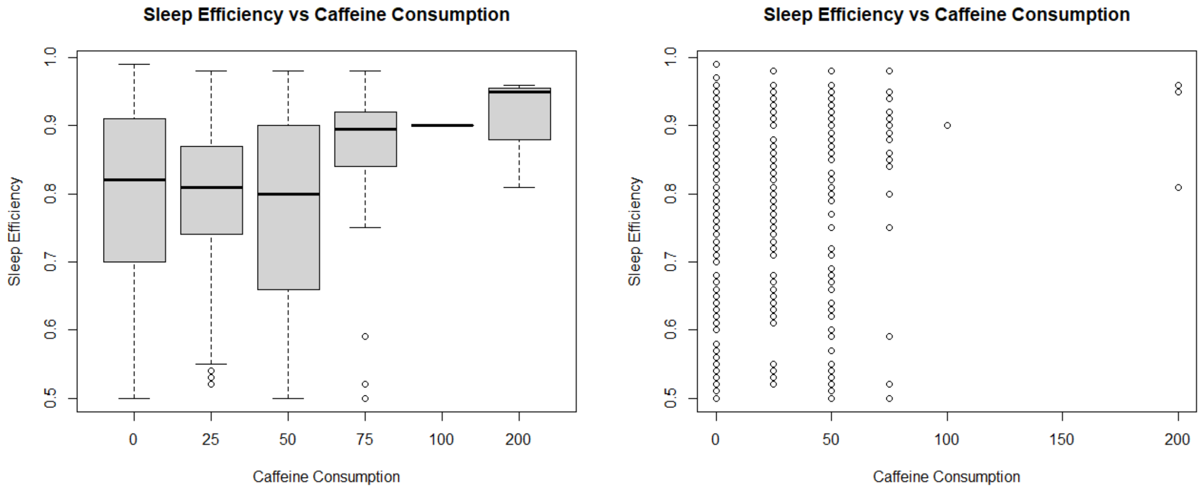
For our visualizations, we plot each of our potential predictors against sleep efficiency to see if there are any noticeable patterns or issues with fitting a linear model. For quantitative plots, integer values made typical scatter plots look odd, making the boxplots easier to interpret. Scatterplots were still included as some levels had very few values, making the boxplots difficult to interpret.

*Figure 1: Plots of sleep efficiency vs. alcohol consumption.*



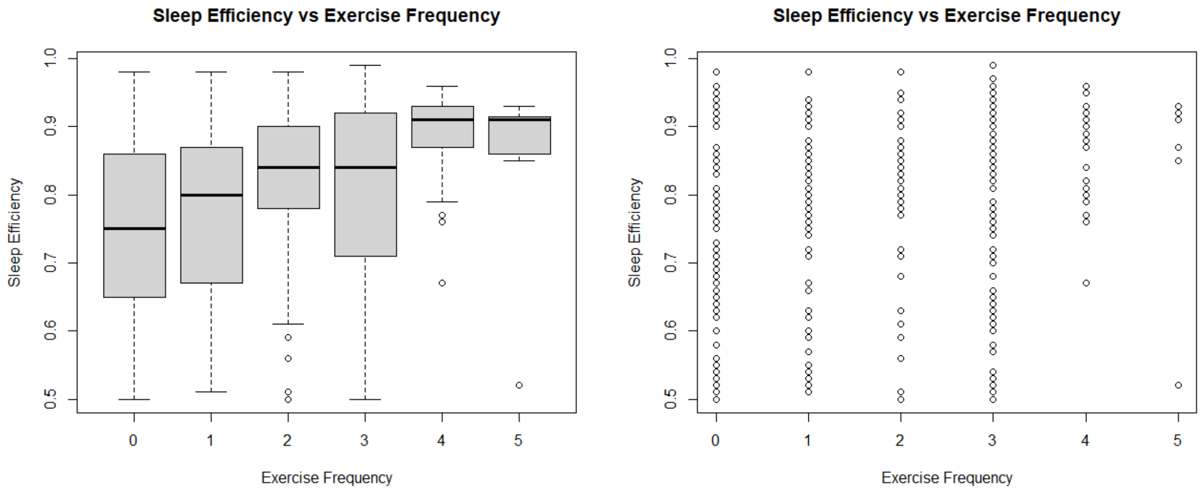
Using the boxplot in figure 1, it can be seen that the density of observations suggests a negative correlation between sleep efficiency and alcohol consumption.

*Figure 2: Plots of sleep efficiency vs. caffeine consumption.*



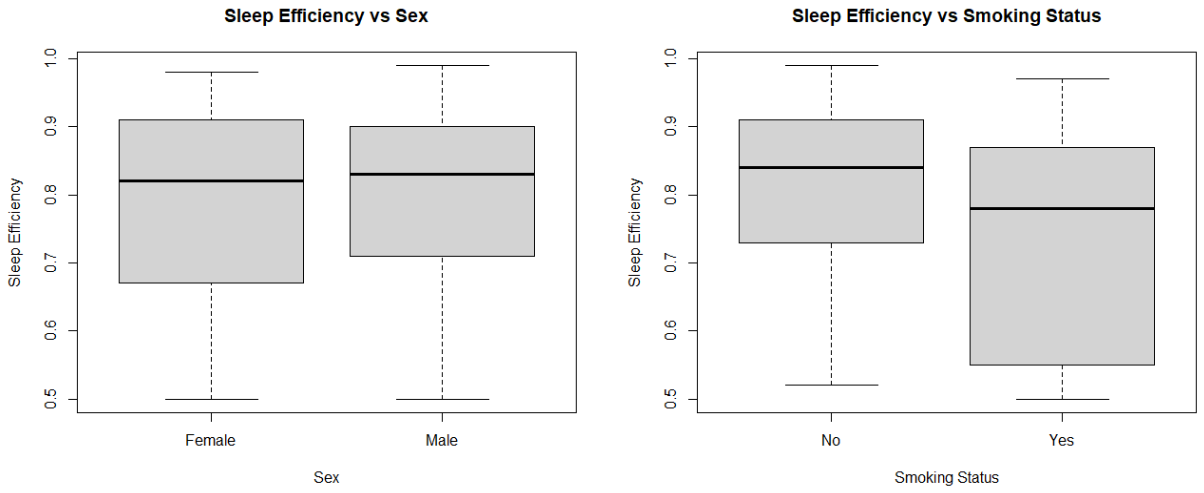
In figure 2, we note that the first few data points in the boxplot show a positive correlation between sleep efficiency and caffeine consumption; from the scatterplot we see a few outliers in the 200mg caffeine consumption range but these still follow the general positive trend.

*Figure 3: Plots of sleep efficiency vs. exercise frequency.*



Similar to caffeine consumption, we see that exercise frequency is positively correlated with sleep efficiency. We see that the spread of values is reduced for larger values of exercise frequency, but this is likely due to a lack of sample points at these values.

*Figure 4: Plots of sleep efficiency vs. sex and sleep efficiency vs. smoking status.*



In figure 4, we see that the distributions of sleep efficiency for sex (male and female) are similar. For smoking, we see that smokers have a much larger spread of sleep efficiency compared to non-smokers.

### 2.2 Model Selection

We perform a best subset selection algorithm using the exhaustive method to determine the predictor variables of the optimal model. This process considers all combinations of explanatory variables and outputs the subsets that best predict the response variable, sleep efficiency. Table 2 below shows the optimal combinations of different numbers of predictors.

| #vars | intercept | Alcohol consumption | Exercise frequency | Caffeine consumption | Smoking status (yes) | Gender (male) |
|---|---|---|---|---|---|---|
| 1 | TRUE | TRUE | FALSE | FALSE | FALSE | FALSE |
| 2 | TRUE | TRUE | TRUE | FALSE | FALSE | FALSE |
| 3 | TRUE | TRUE | TRUE | FALSE | TRUE | FALSE |
| 4 | TRUE | TRUE | TRUE | TRUE | TRUE | FALSE |
| 5 | TRUE | TRUE | TRUE | TRUE | TRUE | TRUE |

*Table 2: Components of the output of regsubsets() with exhaustive method, providing the best subsets of variables of different model sizes.*

We found the adjusted-R$^2$, Mallow’s $C_p$, and residual square of each number of explanatory variables, shown in table 3.

| #variables -> | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| adjusted-R$^2$ | 0.1435820 | 0.2140089 | 0.2763900 | 0.2767982 | 0.2753709 |
| Mallows’ $C_p$ | 75.657679 | 37.210983 | 3.433214 | 4.208173 | 6.000000 |
| Residual square | 0.1456914 | 0.2178807 | 0.2817369 | 0.2839234 | 0.2842949 |

*Table 3: Adjusted-R$^2$, Mallow’s $C_p$, and residual square of models with 1-5 predictor variables (chosen via the exhaustive method)*

We then used a similar approach to the above algorithm, this time making use of the **forward selection method**. This method allows the algorithm to start with one best predictor model, then add on more predictor variables one at a time.

| #vars | intercept | Alcohol consumption | Exercise frequency | Caffeine consumption | Smoking status (yes) | Gender (male) |
|---|---|---|---|---|---|---|
| 1 | TRUE | TRUE | FALSE | FALSE | FALSE | FALSE |
| 2 | TRUE | TRUE | TRUE | FALSE | FALSE | FALSE |
| 3 | TRUE | TRUE | TRUE | FALSE | TRUE | FALSE |
| 4 | TRUE | TRUE | TRUE | TRUE | TRUE | FALSE |
| 5 | TRUE | TRUE | TRUE | TRUE | TRUE | TRUE |

*Table 4: Components of the output of regsubsets() with forward selection method, providing the best subsets of variables of different model sizes*

We found the adjusted-R$^2$, Mallow’s $C_p$, and residual square of each number of explanatory variables (using the results of forward selection), shown in table 5.

| #variables -> | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| adjusted-$R^2$ | 0.1435820 | 0.2140089 | 0.2763900 | 0.2767982 | 0.2753709 |
| Mallows' $C_p$ | 75.657679 | 37.210983 | 3.433214 | 4.208173 | 6.000000 |
| Residual square | 0.1456914 | 0.2178807 | 0.2817369 | 0.2839234 | 0.2842949 |

*Table 5: Adjusted-$R^2$, Mallow's $C_p$, and residual square of models with 1-5 predictor variables (chosen via the forward selection method)*

We observe above that tables 4 and 5 are exactly the same as tables 2 and 3, respectively. This suggests that both selection methods give the same results in this case.

## 2.3 Model Analysis

### 2.3.1 Initial Model Summary

First we fit a linear model involving the response variable, sleep efficiency, and all 5 predictor variables without interaction terms. Beta values are shown in figure 5 below.

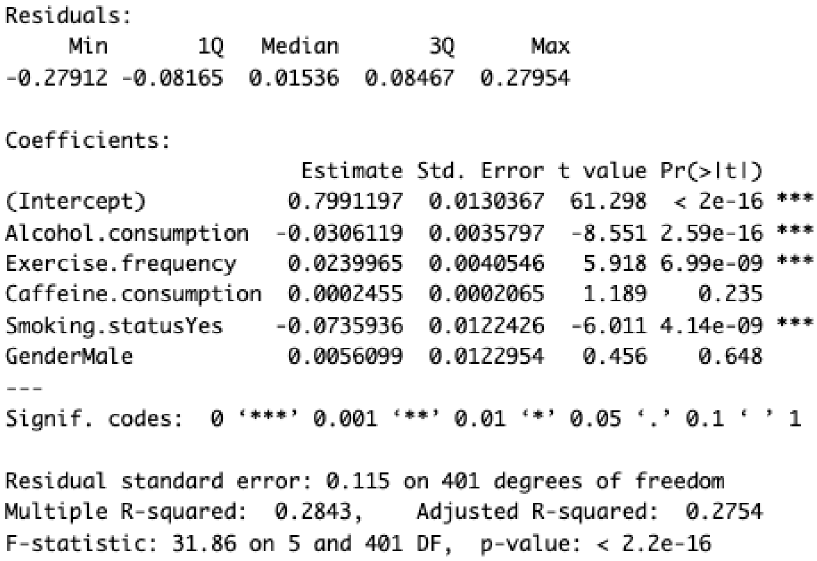
**Model 1: Multiple Regression Without Interaction**

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 z_{1i} + \beta_5 z_{2i} + \varepsilon_i \ , \ \text{i = 1, 2, … 407}$$

*Table 6. Description of variables in the multiple regression model, i = 1, 2, … 407.*

| **Symbol** | **Interpretation** |
|---|---|
| $Y_i$ | Sleep efficiency (proportion) |
| $x_{1i}$ | Alcohol consumption (oz) |
| $x_{2i}$ | Exercise frequency (times/week) |
| $x_{3i}$ | Caffeine consumption (mg) |
| $z_{1i}$ | Dummy variable for smoking status: $z_{1i}$ = 1 if true, $z_{1i}$ = 0 if false |
| $z_{2i}$ | Dummy variable for gender: $z_{2i}$ = 1 if male, $z_{2i}$ = 0 if female |
| $\varepsilon_i$ | Error, follows a normal distribution with mean 0 |

*Figure 5. Summary of the regression model as output by R.*



```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.27912 -0.08165  0.01536  0.08467  0.27954 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)          0.7991197  0.0130367  61.298  < 2e-16 ***
Alcohol.consumption -0.0306119  0.0035797  -8.551 2.59e-16 ***
Exercise.frequency   0.0239965  0.0040546   5.918 6.99e-09 ***
Caffeine.consumption 0.0002455  0.0002065   1.189    0.235    
Smoking.statusYes   -0.0735936  0.0122426  -6.011 4.14e-09 ***
GenderMale           0.0056099  0.0122954   0.456    0.648    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.115 on 401 degrees of freedom
Multiple R-squared:  0.2843,	Adjusted R-squared:  0.2754 
F-statistic: 31.86 on 5 and 401 DF,  p-value: < 2.2e-16
```

Assuming a significance level of 0.05, we can conduct hypothesis tests to evaluate the significance of the coefficient of each individual predictor variable in the model:

$$H_0\colon\ \beta_i = 0$$

$$H_A\colon\ \beta_i \neq 0$$

Based on figure 5, if p-value < 0.05, there exists sufficient evidence to suggest that the corresponding Beta is statistically significant. We can conclude that $\beta_0$, $\beta_1$, $\beta_2$ and $\beta_4$ are significant, and $\beta_3$, $\beta_5$ are not.

The results of this analysis agree with the findings from 2.2 using the best subset selection algorithm. The most significant predictors in a multiple regression model with 3 variables are exercise frequency, alcohol consumption, and smoker status.

### 2.3.2 Investigating Interaction Terms

Investigating interaction terms can provide a deeper understanding of how variables interact to influence the response. By exploring these interactions, we hope to uncover nuanced effects that can inform our analysis and model selection.

There are a few approaches to exploring interaction between the 5 predictor variables. One could include all combinations of predictor variables in the model, or include the subset of interaction terms containing a significant predictor from Model 1 above. However, there is a risk of overfitting with both these approaches. An excessive number of interaction terms can lead to model complexity, hampering its generalization to novel data (Giesselmann & Schmidt-Catran, 2021), and makes the model difficult to interpret.

The significance of smoker status as a predictor warrants further examination, given its established association with disease and organ impairment (West, 2017). A directed investigation into interaction terms involving smoker status may unveil new relationships in the data. In particular, smoking can cause heart disease, lung diseases, diabetes, and problems of the immune system - so, it is of interest to see if smoker status impacts how caffeine consumption, alcohol consumption, and exercise frequency affect sleep efficiency. Furthermore, the interaction of smoking status and gender may give us insight on potential gender-specific impacts of smoking on sleep (Allen et al., 2014).

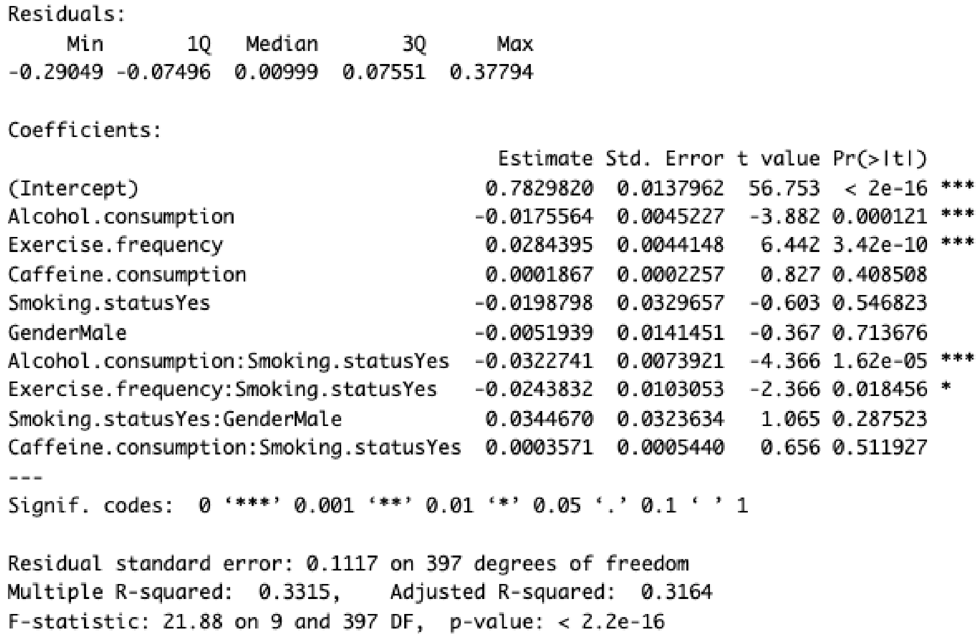
Below we fit a linear model using all 5 predictor variables and interaction terms involving the smoking status predictor. Beta values are shown in figure 6 below.

**Model 2: Multiple Regression With Interaction**

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 z_{1i} + \beta_5 z_{2i} + \beta_6 z_{1i} x_{1i} + \beta_7 z_{1i} x_{2i} + \beta_8 z_{1i} x_{3i} + \beta_9 z_{1i} z_{2i} + \varepsilon_i \ ,$$

i = 1, 2, … 407

*Figure 6. Summary of the regression model as output by R.*



```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.29049 -0.07496  0.00999  0.07551  0.37794 

Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)                          0.7829820  0.0137962  56.753  < 2e-16 ***
Alcohol.consumption                 -0.0175564  0.0045227  -3.882 0.000121 ***
Exercise.frequency                   0.0284395  0.0044148   6.442 3.42e-10 ***
Caffeine.consumption                 0.0001867  0.0002257   0.827 0.408508    
Smoking.statusYes                   -0.0198798  0.0329657  -0.603 0.546823    
GenderMale                          -0.0051939  0.0141451  -0.367 0.713676    
Alcohol.consumption:Smoking.statusYes  -0.0322741  0.0073921  -4.366 1.62e-05 ***
Exercise.frequency:Smoking.statusYes   -0.0243832  0.0103053  -2.366 0.018456 *  
Smoking.statusYes:GenderMale         0.0344670  0.0323634   1.065 0.287523    
Caffeine.consumption:Smoking.statusYes  0.0003571  0.0005440   0.656 0.511927    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1117 on 397 degrees of freedom
Multiple R-squared:  0.3315,	Adjusted R-squared:  0.3164 
F-statistic: 21.88 on 9 and 397 DF,  p-value: < 2.2e-16
```

Again, if p-value < 0.05, there exists sufficient evidence to suggest that the corresponding Beta is statistically significant. From figure 6 we can conclude that only $\beta_0$, $\beta_1$, $\beta_2$, $\beta_6$ and $\beta_7$ are significant.

Among the four interaction terms incorporated, only two interactions stand out as statistically significant at the 95% confidence level. These significant interactions are observed between alcohol consumption and smoking status, as well as between exercise frequency and smoking status.

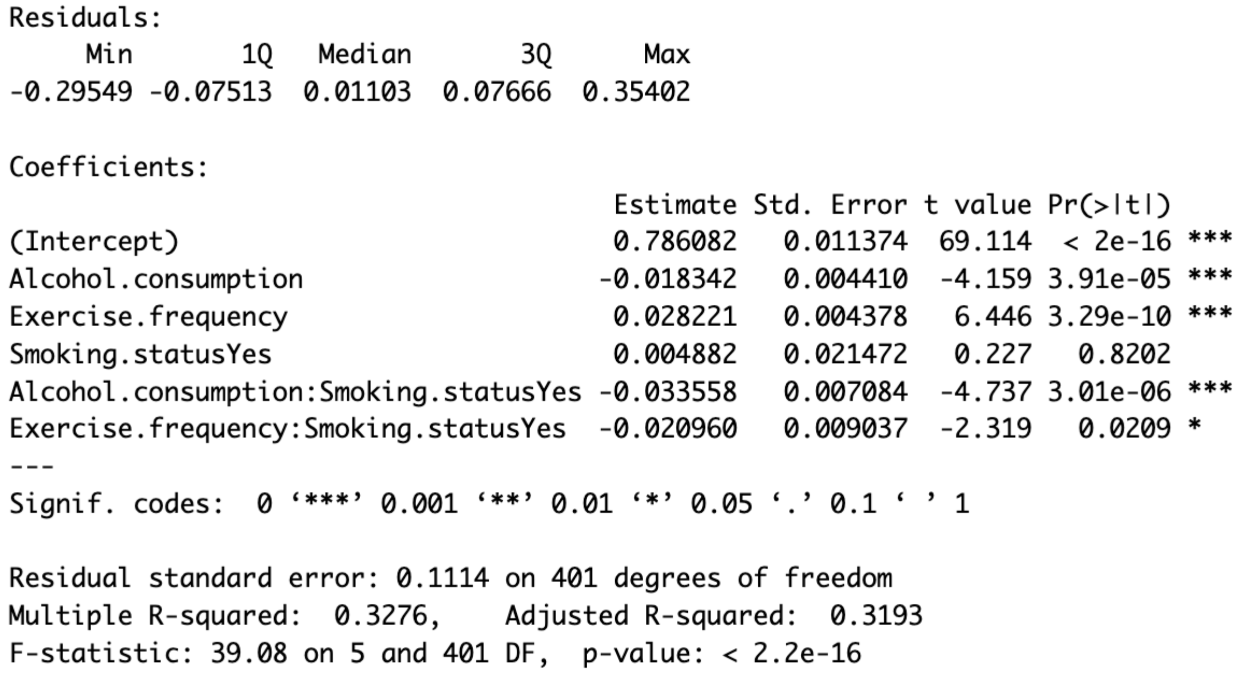
Notably, the previously significant predictor, smoking status, loses its significance within this regression model. However, it is imperative to consider that this might be attributed to the interaction terms diverting significance from its "main effect." Consequently, the lack of significance for smoking status might not accurately reflect its true importance (Giesselmann & Schmidt-Catran, 2021). Therefore, we choose to retain smoking status as a predictor variable in the model, recognizing its potential contribution alone and in conjunction with other variables.

Based on the combined analysis of model 1 and model 2, we construct a third predictive model for sleep efficiency using the predictors alcohol consumption, exercise frequency, and smoking status, as well as the interaction terms between smoking status and alcohol consumption and between smoking status and exercise frequency. The adjusted multiple regression model is denoted as Model 3. Beta values are shown in figure 7.

**Model 3: Adjusted Multiple Regression Model**

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 z_{1i} + \beta_4 z_{1i} x_{1i} + \beta_5 z_{1i} x_{2i} + \varepsilon_i \ , \ \text{i = 1, 2, … 407}$$

*Figure 7. Summary of the regression model as output by R.*



```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.29549 -0.07513  0.01103  0.07666  0.35402 

Coefficients:
                                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)                            0.786082   0.011374  69.114  < 2e-16 ***
Alcohol.consumption                   -0.018342   0.004410  -4.159 3.91e-05 ***
Exercise.frequency                     0.028221   0.004378   6.446 3.29e-10 ***
Smoking.statusYes                      0.004882   0.021472   0.227   0.8202    
Alcohol.consumption:Smoking.statusYes -0.033558   0.007084  -4.737 3.01e-06 ***
Exercise.frequency:Smoking.statusYes  -0.020960   0.009037  -2.319   0.0209 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1114 on 401 degrees of freedom
Multiple R-squared:  0.3276,	Adjusted R-squared:  0.3193 
F-statistic: 39.08 on 5 and 401 DF,  p-value: < 2.2e-16
```

## 3. Discussion

### 3.1 Model Selection

Firstly, recall that our goal is to identify the best set of predictors for a model with sleep efficiency as the response variable.

After performing the best subset selection algorithm we observe that the chosen models are the same using both the exhaustive and forward selection methods. That is, whether we compare all possible combinations of subsets of explanatory variables or add on more variables one by one to compare, the result is the same. This consistency is noteworthy, as it suggests that the choice of selection method does not significantly impact the resulting model.

Referring to Tables 3 and 5, the model with the highest adjusted $R^2$ has 3 or 4 explanatory variables. It is preferable to work with fewer variables when the difference is insignificant, so we choose a model with 3 variables. The adjusted $R^2$ increases when adding a variable to the model more than would be expected by chance. It increases when we add another variable to the model with 2 variables, making 3 a considerably better number of explanatory variables for the model. These models also have a Mallow’s $C_p$ closest to the number of corresponding predictors; preferably we want the model to have a relatively small Mallows $C_p$ close to the number of predictors of the model. Therefore, the model with 3 predictors should be the most suitable.

Looking at Tables 2 and 4, the common variables appearing in the selected models across both methods are alcohol consumption and exercise frequency. These variables consistently contribute to the model's explanatory power regardless of the selection approach. This suggests that both alcohol consumption and exercise frequency play a substantial role in predicting sleep efficiency. Other variables are less significant so not all models included them. Models with 3 variables include alcohol consumption, exercise frequency, and smoking status as predictors.

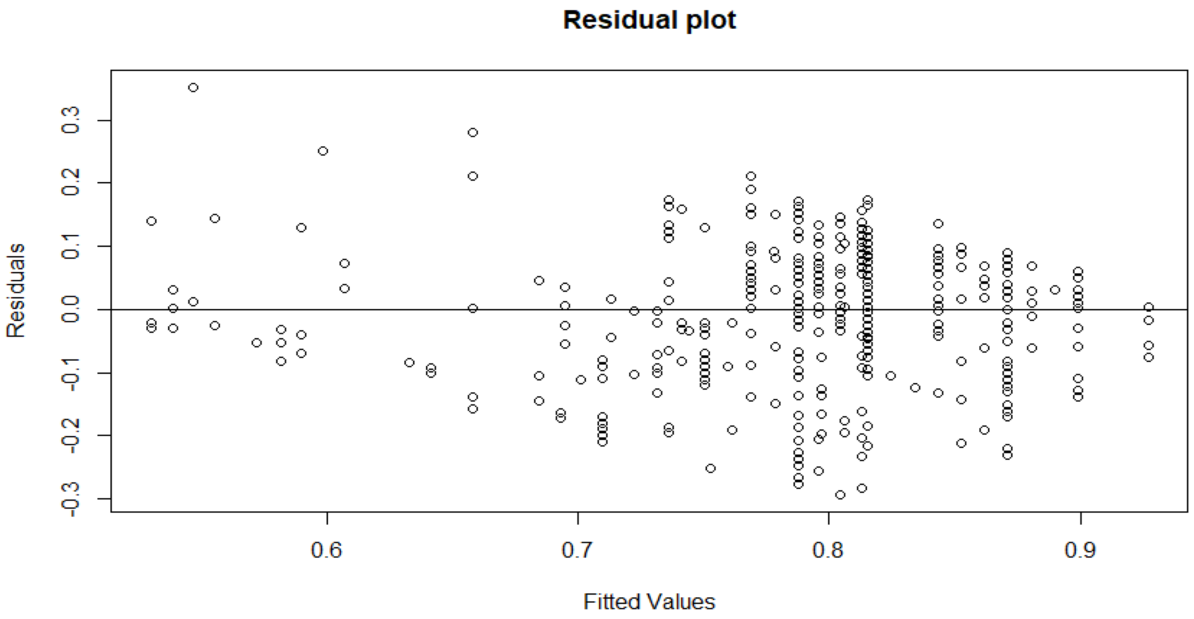
### 3.2 Regression Model Analysis

Our final model is the following

$$x_1 = \text{Alcohol consumption}$$
$$x_2 = \text{Exercise frequency}$$
$$z_1 = \text{Smoking Status (no} = 0\text{, yes} = 1)$$
$$\hat{y} = 0.786082 - 0.018342x_1 + 0.028221x_2 + 0.004882z_1 - 0.033558x_1z_1 - 0.020960x_2z_1$$

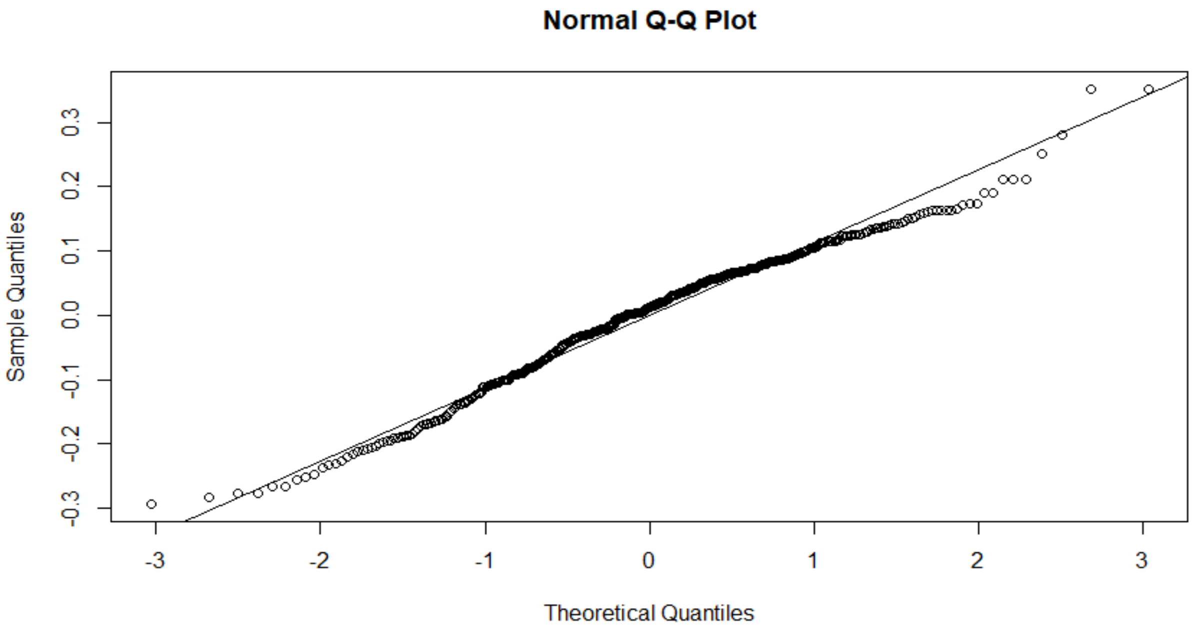
Below we analyze the residuals with our selected model to see if it is a “good” fit.

*Figure 8: Plot of residuals against fitted values.*



From the residual plot of our model, shown in figure 8, we can notice many values stacked on top of each other. However, this is due to the way our data is recorded. All of our quantitative predictors are recorded as integers which results in this equal spacing between values and apparent stacking of values. We can see from the plot that there is no discernible pattern, so it appears that there is no violation of the assumption of constant variance (homoscedasticity) in our model.

*Figure 9: Normal Q-Q Plot.*



From the Normal Q-Q plot in figure 9, we see that the residuals lie along the reference line in an almost straight line. There is a small amount of deviation towards the right end of the plot, but not enough that the normality assumption appears invalid.

## 4. Conclusion

In this report, we used the Kaggle Sleep Efficiency Dataset to examine the relationship between the response variable sleep efficiency and the explanatory variables alcohol consumption, caffeine consumption, exercise frequency, smoking status, and gender.

We cleaned the data by removing extraneous variables and any rows with missing data.Smoking status and gender were converted to factors to use as categorical variables. We created initial plots of sleep efficiency against each individual predictor to visualize potential patterns in the dataset.

Finally, we performed analysis to address our research questions. We performed a best subset selection algorithm using both the exhaustive and forward selection methods to determine the best predictor variables to use in our regression model. We then used analysis of adjusted R-squared, Mallow’s $C_p$, residuals, and plots to determine the best model for predicting sleep efficiency, with the following results.

### 4.1 Synopsis of Findings

Based on the above analysis, the best model found to predict sleep efficiency from patient data includes three predictors: alcohol consumption, exercise frequency, and smoking status. Interaction terms between smoking status and each of the two other predictors are also significant and were thus included in the best model:

$$x_1 = \text{Alcohol consumption}$$
$$x_2 = \text{Exercise frequency}$$
$$z_1 = \text{Smoking Status (no} = 0\text{, yes} = 1)$$
$$\hat{y} = 0.786082 - 0.018342x_1 + 0.028221x_2 + 0.004882z_1 - 0.033558x_1z_1 - 0.020960x_2z_1$$

With this “best” model, the adjusted $R^2$ is 0.3193. This is a value generally not considered in practice to indicate a good fit (Chicco et al., 2021). Based on this observation, it is likely that other predictors are needed to more adequately explain sleep efficiency as a response variable. One hypothesized potential predictor is consistency of bedtime; this variable was not used as most subjects are inconsistent and it is more difficult to work with this data in the given form, but further studies could explore possible impact in this area. Other potential predictors of interest could require gathering more data; e.g., patient routine prior to bedtime.

### 4.2 Comments

The analysis documented in this report suggests that the following patient features contribute to greater sleep efficiencies:

- Lower alcohol consumption
- Higher exercise frequency
- Status of non-smoker

The interaction terms add to the observation regarding smoking status, as both interaction terms have a negative coefficient in the final regression model.

Rapid Eye Movement (REM) sleep and Deep each sleep play an extremely important role in daily performance. Concentration, mood regulation, and processing of new information are just a few of the areas for which sleep quality and efficiency are necessary. The observations above provide real-world use in that individuals aiming to improve sleep efficiency can begin to do so by attempting to improve these aspects of their lives.

## References

1. Allen, A. M., Oncken, C., & Hatsukami, D. (2014). Women and smoking: The effect of gender on the epidemiology, health effects, and cessation of smoking. Current Addiction Reports, 1(1), 53-60. https://doi.org/10.1007/s40429-013-0003-6

2. Giesselmann, M., & Schmidt-Catran, A. (2021). Interactions in Fixed Effects Regression Models. https://doi.org/10.31235/osf.io/m78qf

3. West, R. (2017). Tobacco smoking: Health impact, prevalence, correlates and interventions. Psychology & Health, 32(8), 1018-1036. https://doi.org/10.1080/08870446.2017.1325890

4. Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Computer Science, 7, e623. https://doi.org/10.7717/peerj-cs.623