STA 4164 – Final Report

Professor: Nathaniel Simone

Spring 2024 | Section 1

Group 18

Brendan B., Jerzy D., & Zachary L.

Motivation and Data Description

This study aims to determine which factors most significantly influence sleep efficiency. Sleep efficiency is the time an individual spends asleep divided by the time an individual dedicates to sleep. The reason sleep efficiency is important is because poor sleep quality results in a variety of health-related issues. Sleep is often measured with sleep efficiency versus hours slept since it is a more consistent way to measure sleep quality across age, gender, and other factors.

Our dataset has 387 unique observations and 11 total variables on numerous different lifestyle factors and sleep quality statistics. The dataset provided was collected from a study conducted in Morocco by a group of artificial engineering students at ENSIAS. The team recruited participants from the local community and collected data over a period of several months using self-reported surveys, actigraphy, and polysomnography. The variables can be split up into two groups: nominal and categorical.

The nominal data variables are age, sleep duration, sleep efficiency, REM sleep percentage, and deep sleep percentage. Age represents the age of the participants. Sleep duration, measured in hours, is the total number of hours participants slept during the study. Sleep efficiency is the overall efficiency of one's sleep, measured as a ratio of the total sleep time to the time in bed – it represents a percentage of the time an individual is sleeping while they are in bed. REM sleep percentage accounts for the total percentage of one's sleep time spent in the REM stage. The last nominal variable is deep sleep percentage, which represents the percentage of one's total sleep time spent in deep sleep.

The remaining gender, awakenings, caffeine consumption, alcohol consumption, smoking status, and exercise frequency variables are all categorical. Gender was one of the dummy variables in the study, with 0 representing females and 1 representing males. Awakenings were categorized by 0, 1 awakening, 2, 3, and +4, representing the total number of times participants woke up from sleep. Caffeine consumption represents the total amount of caffeine consumed (measured in mg) 24 hours before bedtime, in intervals of 0mg, 25mg, 50mg, 75mg, 100mg, and +200mg. Alcohol consumption measures the amount of alcohol (measured in oz) participants consumed 24 hours before bedtime, in intervals of 0oz, 1oz, 2oz, 3oz, 4oz, and +5oz. Smoking status is a dummy variable – 1 represents smokers, and 0 represents non-smokers. The final variable is exercise frequency, measured in the number of days per week participants exercise, with 0, 1 day, 2 days, 3 days, 4 days, and +5 days.

Model Diagnostics:

Age:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
9.00	29.00	41.00	40.79	52.00	69.00

Gender:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0000	0.0000	0.0000	0.4987	1.0000	1.0000

Sleep duration:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
5.000	7.000	7.500	7.451	8.000	10.000

Sleep efficiency:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.500	0.700	0.820	0.789	0.900	0.990

REM Sleep percentage:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
15.00	20.00	22.00	22.69	25.00	30.00

Deep Sleep percentage:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
18.00	51.00	58.00	52.79	63.00	75.00

Awakenings:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.000	1.000	1.000	1.623	3.000	4.000

Caffeine consumption (mg):

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.00	0.00	0.00	22.67	50.00	200.00

Alcohol consumption:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.00	0.00	0.00	1.15	2.00	5.00

Smoking status:

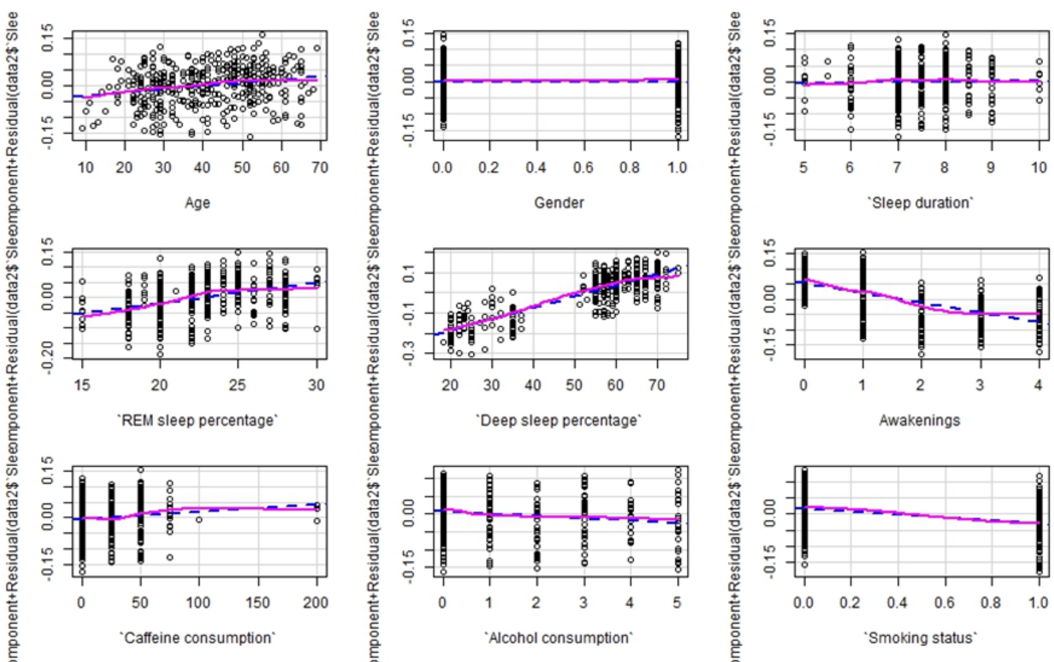
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0000	0.0000	0.0000	0.3411	1.0000	1.0000

Exercise frequency:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.00	0.00	2.00	1.76	3.00	5.00

As we can see from our basic data analysis, we can see that sleep efficiency ranges from a minimum of 0.5 (50%) to 0.99 (99%), our sample tends to consume 22.67 mg of caffeine, get moderate activity levels of 1.76 times a week, and a smoking status of 0.3411 which indicates that only about 30% of our participants are smokers. There were 3 outliers of 200mg for caffeine consumption that were removed from our final data set before training the data set because this would significantly skew our caffeine consumption coefficient, and these individuals drank 10 times the average amount of 22.67mg. After removing the 3 caffeine outliers, we no longer had any obvious outliers in our data set.

In beginning our initial model diagnostics, we obtained the partial plots of the full model.



We disregarded the dummy variables among the plots and focused on the remaining plots. Fortunately, there did not appear to be a need for transformations. Collinearity was not an issue with any variables. The correlation between deep sleep and sleep efficiency was 0.78, suggesting that deep sleep was a good predictor of sleep efficiency. Our final model also validated this since it included an interaction term between deep sleep and awakenings and smoking status. We concluded that deep sleep was a significant predictor in the final model.

Next, we tested for interaction terms. We included all the main effects and a couple of interaction terms that could be used to explain sleep efficiency. We tested two interaction terms: deep sleep and smoking status, along with deep sleep and sleep efficiency. First, we produced an Analysis of Variance (ANOVA) table to determine if the full model with the two interaction terms compared against the reduced model with only the main effects was significant.

```
> anova(reduced_model1, full_model1)
Analysis of Variance Table

Model 1: d$`Sleep efficiency` ~ Age + Gender + `Sleep duration` + `REM sleep percentage` +
    `Deep sleep percentage` + Awakenings + `Caffeine consumption` +
    `Alcohol consumption` + `Smoking status` + `Exercise frequency`
Model 2: d$`Sleep efficiency` ~ Age + Gender + `Sleep duration` + `REM sleep percentage` +
    `Deep sleep percentage` + Awakenings + `Caffeine consumption` +
    `Alcohol consumption` + `Smoking status` + `Exercise frequency` +
    d$`Smoking status`:d$`Deep sleep percentage` + d$Awakenings:d$`Deep sleep percentage`
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    376 1.3893
2    374 1.1855  2   0.20375 32.139 1.317e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Since there was an extremely small p-value of 1.317e-13, we were able to reject the null hypothesis that the variables were insignificant and can conclude that at least one of the two interaction terms are useful in predicting sleep efficiency in the full model. Next, we tested each interaction term individually. We tested deep sleep and smoking status to determine if it was significant in the model compared to the original full model with only the main effects.

```
> # Reject H0, Now Conduction Partial F-test to see which interaction is not 0.
> full_model2 <- lm(d$`Sleep efficiency`~ .+ d$`Smoking status`:d$`Deep sleep percentage` ,data=d)
> anova(full_model2, full_model1)
Analysis of Variance Table

Model 1: d$`Sleep efficiency` ~ Age + Gender + `Sleep duration` + `REM sleep percentage` +
    `Deep sleep percentage` + Awakenings + `Caffeine consumption` +
    `Alcohol consumption` + `Smoking status` + `Exercise frequency` +
    d$`Smoking status`:d$`Deep sleep percentage`
Model 2: d$`Sleep efficiency` ~ Age + Gender + `Sleep duration` + `REM sleep percentage` +
    `Deep sleep percentage` + Awakenings + `Caffeine consumption` +
    `Alcohol consumption` + `Smoking status` + `Exercise frequency` +
    d$`Smoking status`:d$`Deep sleep percentage` + d$Awakenings:d$`Deep sleep percentage`
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    375 1.2852
2    374 1.1855  1  0.099649 31.437 4.009e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Since the p-value = 4.009e-08 is less than alpha = .05, we rejected the null hypothesis that the interaction term was insignificant and included the interaction term of deep sleep and smoking status in our model. We then tested our other interaction terms, deep sleep and awakenings, to determine their significance.

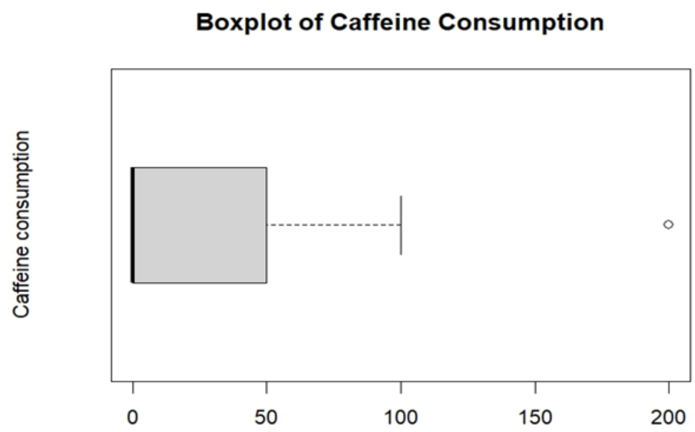
```
> full_model3 <- lm(d$`Sleep efficiency`~.+ d$Awakenings:d$`Deep sleep percentage`,data=d)
> anova(full_model3, full_model1)
Analysis of Variance Table

Model 1: d$`Sleep efficiency` ~ Age + Gender + `Sleep duration` + `REM sleep percentage` +
    `Deep sleep percentage` + Awakenings + `Caffeine consumption` +
    `Alcohol consumption` + `Smoking status` + `Exercise frequency` +
    d$Awakenings:d$`Deep sleep percentage`
Model 2: d$`Sleep efficiency` ~ Age + Gender + `Sleep duration` + `REM sleep percentage` +
    `Deep sleep percentage` + Awakenings + `Caffeine consumption` +
    `Alcohol consumption` + `Smoking status` + `Exercise frequency` +
    d$`Smoking status`:d$`Deep sleep percentage` + d$Awakenings:d$`Deep sleep percentage`
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    375 1.2651
2    374 1.1855  1  0.079597 25.111 8.367e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

We rejected the null hypothesis that the interaction term was insignificant, and since the p-value = 8.367e-07 is less than alpha = .05, we include the Deep Sleep and Awakenings interaction term in our final model.

Before running the model selection, we looked at potential outliers from the dataset using Cook's distance, Leverage, and Jackknife to obtain the following:

```
#Cook's Distance: No Outliers > 1
print(sort(cooks.distance(full_model1)))

#Jackknife |x|>1.966344
qt(0.025,387-12-2, lower.tail=FALSE)
print(sort(studres(full_model1)))
#Outliers: 215,154,155,174,218,225,255,24,284,115,212,37,351,27,81,379,295,287,286,329,183,28

#Leverage >2(k+1)/n = 2(13)/387 = 0.06718346
print(sort(hatvalues(full_model1)))
#Outliers: 193,327,387,149,175,10,200,344,320,113,240,273,229
```

Although there were no outliers in Cook's distance and no outliers in both Jackknife and Leverage, we decided to remove a few caffeine outliers. These were points: 240, 273, and 229 from the leverage test. They had consumed 10 times the average caffeine consumption. Including these could potentially skew the data, moving forward. This was also observed in a box plot:



Model Selection:

Our full model was fit only the full data set without the 3 caffeine outliers, to validate whether our model can accurately predict sleep efficiency of an individual based on the data set. We used forward, stepwise, and backward selection algorithms in R to eliminate the insignificant predictors from the full model using a p-value of 0.1. Here are the 3 models that we ended up with:

```
forward_model <- lm(Sleep.efficiency ~ Deep.sleep.percentage +
Smoking.status + REM.sleep.percentage + Age + Sleep.duration +
Gender + Caffeine.consumption + Alcohol.consumption +
Exercise.frequency + Awakenings + Deep.sleep.percentage :
Awakenings + Deep.sleep.percentage : Smoking.status, data=d)

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

stepwise_model <- lm(Sleep.efficiency ~ Deep.sleep.percentage +
Caffeine.consumption + Alcohol.consumption + Awakenings +
```

```
Exercise.frequency + Deep.sleep.percentage : Smoking.status,
data=d)
```

After conducting the tests, we were left with the following AIC values for each test.

Forwards:

Model Summary

R	0.913	RMSE	0.055
R-Squared	0.834	MSE	0.003
Adj. R-Squared	0.828	Coef. Var	7.135
Pred R-Squared	0.822	AIC	-1113.790
MAE	0.044	SBC	-1058.372

Backwards:

Model Summary

R	0.913	RMSE	0.055
R-Squared	0.833	MSE	0.003
Adj. R-Squared	0.829	Coef. Var	7.117
Pred R-Squared	0.824	AIC	-1117.726
MAE	0.044	SBC	-1070.225

Stepwise:

Model Summary

R	0.870	RMSE	0.067
R-Squared	0.757	MSE	0.005
Adj. R-Squared	0.753	Coef. Var	8.560
Pred R-Squared	0.747	AIC	-978.763
MAE	0.055	SBC	-947.095

From this, we concluded that the Backwards model had the lowest AIC value of -1117.726, which means that it was the best model out of the three. This model included all of the original variables, except for sleep duration and gender, along with both of the interaction terms outlined previously.

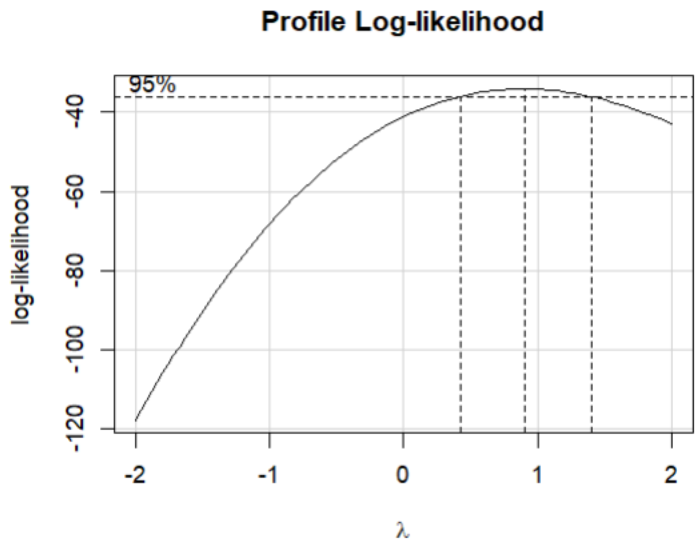
We then conducted an additional Multiple Partial F-Test to determine if each of the variables in the final model were significant. At a p-value of 0.066907, which is greater than alpha = 0.05, we determined that Awakenings is not a significant predictor of sleep efficiency and should be removed from the final model. This, however, did not affect the interaction term that included awakenings (deep sleep and awakenings) as it had a p-value of 4.79e-08, which is less than alpha. That interaction term stayed in the model.

```
Residuals:
      Min        1Q    Median        3Q       Max 
-0.161430 -0.035340  0.002922  0.037314  0.144663 

Coefficients:
                                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)                           0.2900964  0.0370490   7.830 5.11e-14 ***
Deep.sleep.percentage                 0.0065677  0.0004585  14.323  < 2e-16 ***
Smoking.status                       -0.1483413  0.0208604  -7.111 5.91e-12 ***
REM.sleep.percentage                  0.0072076  0.0008847   8.147 5.70e-15 ***
Age                                   0.0008732  0.0002202   3.966 8.77e-05 ***
Caffeine.consumption                  0.0002604  0.0001231   2.116 0.034971 *  
Alcohol.consumption                  -0.0053629  0.0019736  -2.717 0.006888 ** 
Exercise.frequency                    0.0082457  0.0021114   3.905 0.000112 ***
Awakenings                            0.0161053  0.0087640   1.838 0.066907 .  
Deep.sleep.percentage:Awakenings     -0.0008711  0.0001563  -5.574 4.79e-08 ***
Deep.sleep.percentage:Smoking.status  0.0019568  0.0003876   5.049 6.97e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05635 on 373 degrees of freedom
Multiple R-squared:  0.8323,	Adjusted R-squared:  0.8279 
F-statistic: 185.2 on 10 and 373 DF,  p-value: < 2.2e-16
```

After determining the model, we were going to move forward with, we conducted a Shapiro Test to determine if normality was an issue. Since the p-value of the Shapiro-Wilk test was 0.009305, which is less than 0.05, we determined that normality was violated.

```
	Shapiro-Wilk normality test

data:  backward_model_sig_only$residuals
W = 0.98985, p-value = 0.009305
```

To fix this, we conducted a Box-Cox test to determine what we should transform sleep efficiency by. Based off the test, we determined that it should be transformed by 0.9.



To ensure normality was fixed, we conducted an additional Shapiro-Wilk normality test, but transformed sleep efficiency to the power of 0.9. However, this resulted in a p-value of 0.009105, which is still less than 0.05 and confirms that normality is still violated after the transformation. After some guessing and checking, we determined that the transformation to the power of 3 was the closest value that did not result in a violation of normality.

```
        Shapiro-Wilk normality test

data:  backward_model_sig_only$residuals
W = 0.98982, p-value = 0.009105
```

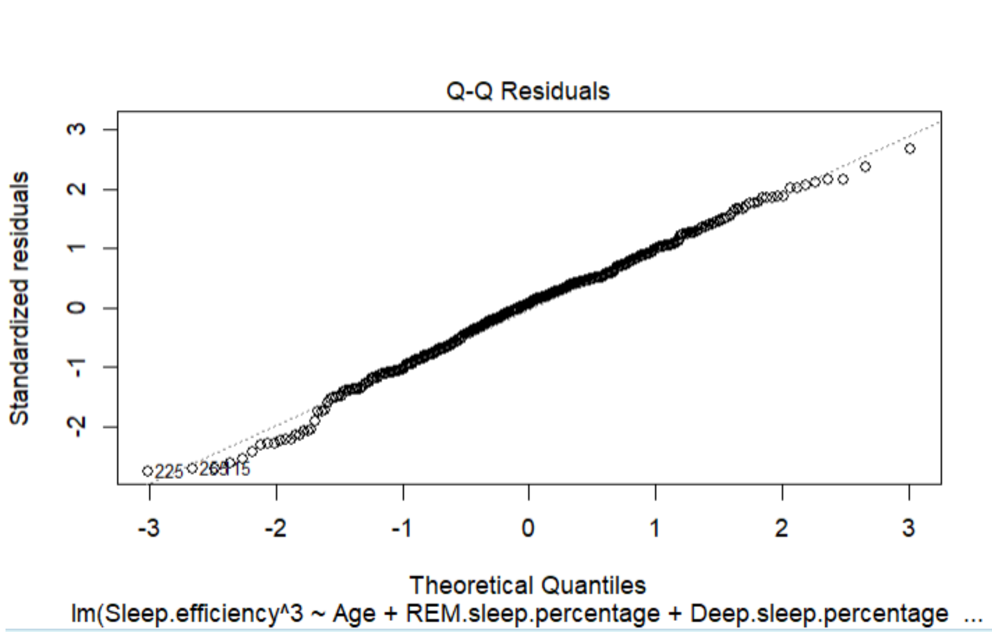
```
        Shapiro-Wilk normality test

data:  backward_model_sig_only$residuals
W = 0.99274, p-value = 0.05968
```

Finally, we conducted one last set of Multiple Partial F-Tests to ensure that each variable was significant. Caffeine consumption had a p-value of 0.081882, which is less than alpha = 0.05 and thus, is not significant to the final model.

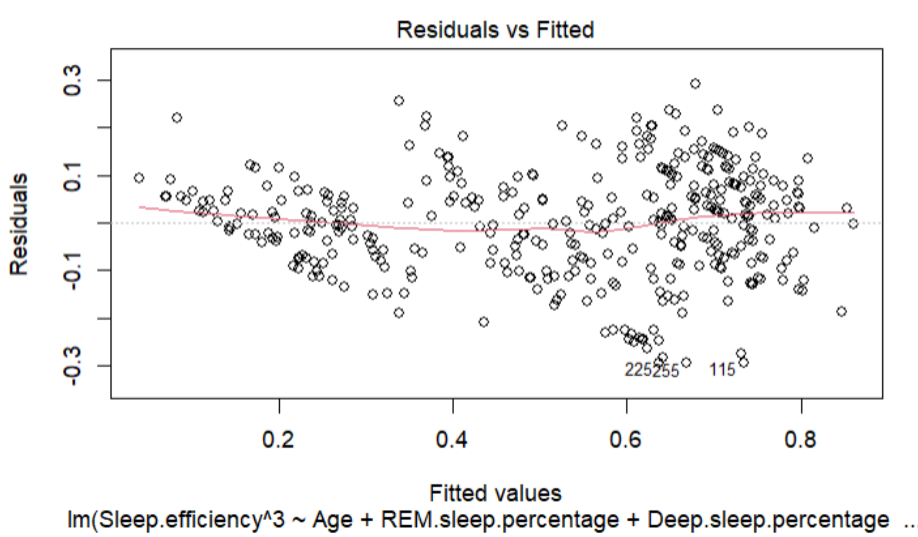
```
Call:
lm(formula = Sleep.efficiency^3 ~ Age + REM.sleep.percentage + 
    Deep.sleep.percentage + Caffeine.consumption + Alcohol.consumption + 
    Smoking.status + Exercise.frequency + Deep.sleep.percentage:Smoking.status + 
    Deep.sleep.percentage:Awakenings, data = d)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.295919 -0.072859  0.007502  0.069415  0.297502 

Coefficients:
                                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)                          -0.2371479  0.0550661  -4.307 2.12e-05 ***
Age                                   0.0013753  0.0004291   3.205 0.001468 ** 
REM.sleep.percentage                  0.0118268  0.0017095   6.918 1.99e-11 ***
Deep.sleep.percentage                 0.0103866  0.0005514  18.835  < 2e-16 ***
Caffeine.consumption                  0.0004185  0.0002399   1.745 0.081882 .  
Alcohol.consumption                  -0.0099700  0.0038421  -2.595 0.009834 ** 
Smoking.status                       -0.1587198  0.0405194  -3.917 0.000107 ***
Exercise.frequency                    0.0150584  0.0041110   3.663 0.000285 ***
Deep.sleep.percentage:Smoking.status  0.0017026  0.0007520   2.264 0.024148 *  
Deep.sleep.percentage:Awakenings     -0.0012759  0.0000812 -15.712  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1098 on 374 degrees of freedom
Multiple R-squared:  0.7859,	Adjusted R-squared:  0.7808 
F-statistic: 152.6 on 9 and 374 DF,  p-value: < 2.2e-16
```

After removing it, we conducted an additional Shapiro test to ensure that this new model would not violate normality. And at a p-value of 0.05754, the Shapiro-Wilk test showed that normality was not violated with transforming Y by the power of 3.

```
	Shapiro-Wilk normality test

data:  backward_model_sig_only$residuals
W = 0.99269, p-value = 0.05754
```



Homoscedasticity, linearity, and normality are all not violated based on the residual vs. fitted plot below. All of the assumptions are satisfied.



Results summary:

Our final model is the following with Y (sleep efficiency) transformed to the power of 3:

```
backward_model <- lm(I(Sleep.efficiency^3) ~ Age +
REM.sleep.percentage + Deep.sleep.percentage +
Alcohol.consumption + Smoking.status + Exercise.frequency +
Deep.sleep.percentage : Smoking.status + Deep.sleep.percentage :
Awakenings, data=d)
```

$Y = \beta0+\beta1x1+\beta2x2+\beta3x3+\beta4x4+\beta5x5+\beta6x6+\beta7x3x5+\beta6x1x7$

$Y =$

$-2.284e-01+1.244e-03x1+1.218e-02x2+1.040e-02x3-1.031e-02x4-1.572e-01x5+1.402e-02x6+1.696e-03x3x5-1.294e-03x3x7$

Our R-squared of 0.7842 shows that 78.42% of the variability in sleep efficiency is explained by our multilinear regression model, while our R of -0.88555 shows that there is a strong negative correlation between the data and our multilinear regression model. Meaning as our indicators and variables decrease, so does sleep efficiency. If someone gets less exercise or has a lower percentage of their total sleep time in the REM stage, their sleep efficiency is going to decrease. These two coefficients are important because they tell us how much variation can be explained by the multilinear regression model and how strong our association is.

Conclusions and limitations:

The most meaningful takeaway from our analysis is that deep sleep percentage and the number of times an individual wakes up throughout the night are the most significant predictors of sleep efficiency for our study. This is not surprising because there are studies such as "REM Sleep: An Unknown Indicator of sleep quality" that show that deep sleep (stages N2-N3), as measured by polysomnography, is highly correlated with the sleep quality of an individual.

Some limitations of our study include a limited sample size for different age groups, especially younger groups. We only had 57 individuals younger than 25, so there is a limited sample, and different age groups have different "normal" sleep efficiency ratios.

A more comprehensive study would group individuals by age and gender to account for any differences between those groups and have different sets of models to predict sleep efficiency for different age groups and genders. Our data also wasn't labeled very well, for example, we don't know if individuals who consumed 20mg of caffeine consumed it 24 hours before sleeping or right before bedtime, which would influence their ability to sleep.

As for programming, we learned a lot about how to use R, which is very useful in data science. The data frames, data cleaning, data analysis, model selection, and diagnostics have given us a good theoretical framework to create different types of multilinear regression models from different data sets that we can transfer to other programming languages and libraries like Python Scikit Learn or Numpy. As for our group, we learned that working as a team can be much more effective by working on the diagnostics together when one of us gets stuck on how to analyze something. 90% of the effort was dedicated to simply cleaning the data and getting background information. Then, model selection was a lot easier once we ran the different algorithms and did the multiple partial F test on the main effects to see what was the most important.

Works Cited

1. Barbato G. (2021). REM Sleep: An Unknown Indicator of Sleep Quality. International journal of environmental research and public health, 18(24), 12976. https://doi.org/10.3390/ijerph182412976
2. Institute of Medicine (US) Committee on Sleep Medicine and Research. (2006). Sleep Disorders and Sleep Deprivation: An Unmet Public Health Problem. In H.R. Colten & B.M. Altevogt (Eds.), Sleep Physiology (2). Washington, DC: National Academies Press (US). https://www.ncbi.nlm.nih.gov/books/NBK19956/
3. Patel, A.K., Reddy, V., Shumway, K.R., et al. (2024). Physiology, Sleep Stages. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK526132/
4. Kaggle. (n.d.). Sleep Efficiency Dataset. Retrieved from Kaggle.