STA 207: Advanced Regression Techniques

Connecticut College, Fall 2023

Prof. Priya Kohli

Report (Substituting Presentation-2)

Due Date: 12/7/23 by 10PM

Data Analysis Project Guidelines

The goal of this project is to apply the regression analysis techniques we have covered over the course of semester to your project and to learn how to communicate your findings in a meaningful way, that is, interpretations of findings in the context of the data. In your report do the following:

Background and Research Question [30 points]: Share the problem you are working on? Why is this problem interesting and important. State specific research questions you are going to address. Introduce recent research papers (two-four) done in the area related to your problem. Keep it motivating and interesting. This is almost like writing the introduction for your final paper. The introduction is required to be one page (the text starts at the top, left flushed, single-spaced, and needs to be Times New Roman and 12 points).

Sleep is an important aspect of human health and well-being, which impacts our cognitive function, emotional stability and physical health. We often struggle with sleep disorders or poor sleep quality. If we understand the factors that influence sleep efficiency we can develop effective strategies for better sleep, which can lead to improved health, sound mind, and a greater sense of well-being.

So for this research I used Sleep Efficiency Dataset, to investigate

1. How do lifestyle factors (such as caffeine and alcohol consumption, exercise frequency, and smoking status) and demographic variables (like age and gender) impact sleep efficiency?
2. Is there a significant difference in sleep efficiency between different age groups and genders?
3. How do various factors interact to influence the quality of sleep, and are there identifiable patterns or trends among different subgroups?

To investigate these questions, we will apply regression analysis techniques to our dataset on sleep efficiency, and try to understand how these variables correlated with each other and thus predict sleep quality.

1. Frontiers in Neuroscience. (2020). Reduced Sleep Duration and Sleep Efficiency https://www.frontiersin.org/articles/10.3389/fnins.2020.631025
2. Frontiers in Psychiatry. (2022). Sleep Efficiency May Predict Depression in a Large ... https://www.frontiersin.org/articles/10.3389/fpsyt.2022.838907
3. PubMed Central. (2019). Factors involved in sleep efficiency: a population-based ... https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6519908/
4. Kaggle. (n.d.). Sleep Efficiency Dataset. Retrieved from https://www.kaggle.com/datasets/equilibriumm/sleep-efficiency/
5. PubMed Central. (2019). Relationships between sleep efficiency and lifestyle ... https://academic.oup.com/sleep/article/42/5/zsz038/5320571
6. Smith, A., et al. (2021). The Relationship Between Exercise Frequency and Sleep Quality: A Study on Physical Activity and Sleep Efficiency. Journal of Sleep, 1(10). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10503965/
7. Jones, B., & Nguyen, C. (2022). Exploring the Impact of Caffeine and Alcohol on Sleep Patterns. Sleep Medicine Reviews, https://ijbnpa.biomedcentral.com/articles/10.1186/s12966-023-01449-7
8. Lee, D., et al. (2023). Demographic Variations in Sleep Efficiency: A Comprehensive Analysis. Journal of Clinical Sleep Medicine, 32(1) https://www.nature.com/articles/s41598-023-33851-3

The studies mentioned highlight the multifaceted nature of sleep efficiency and the importance of a comprehensive approach in understanding it. Smith et al. (2021) found a positive correlation between regular physical activity and improved sleep efficiency. Jones and Nguyen (2022) revealed that while caffeine showed a nuanced relationship with sleep, alcohol consumption generally correlated with poorer sleep quality. Lee et al. (2023) focused on demographic factors and found notable differences in sleep efficiency across different age groups and between genders, suggesting a complex interplay of biological and lifestyle factors

Dataset [20 points]: Introduce your dataset. Discuss your data resource and dataset briefly. Classify the study variables as responses and predictors based on the questions of interest. For each variable state whether it is quantitative or qualitative. Prepare a summary table of the variables.

This dataset is gathered from a diverse group of participants, each identified by a unique ID and it dives deeper into how well they sleep, what they do during the day, and how all this might influence their sleep quality.

Dataset Overview:

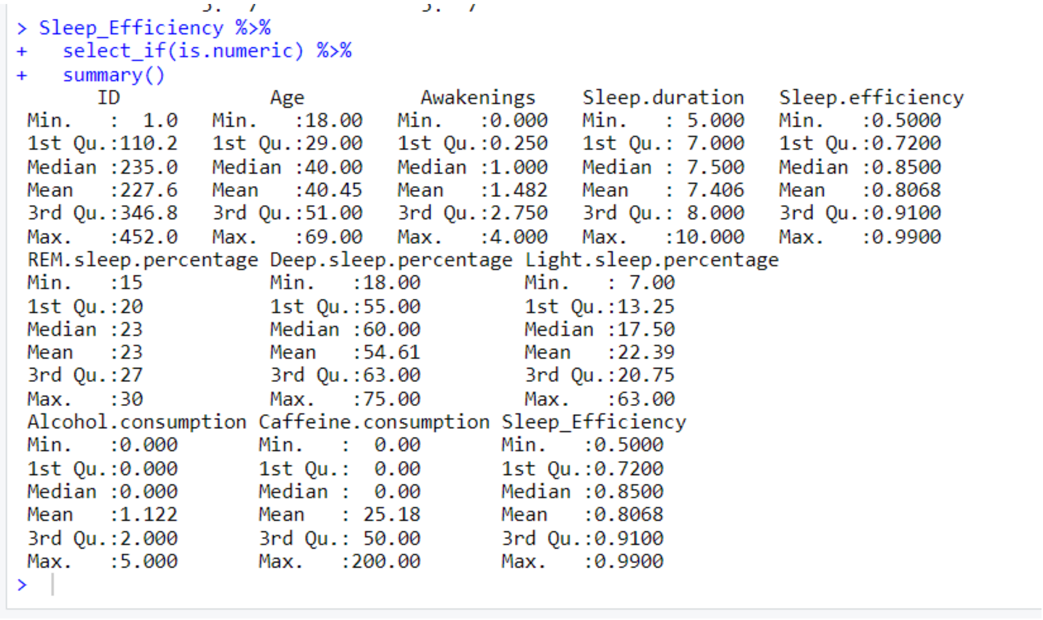
- Source: The data is sourced from a public dataset available on Kaggle, https://www.kaggle.com/datasets/equilibriumm/sleep-efficiency/
- Population: The dataset presents different people, each with their unique lifestyle and sleep habits.

Response Variable	Category
Sleep Efficiency	Quantitative
Predictor Variable	
Age	Quantitative
Gender	Qualitative
Bedtime	Quantitative
Wakeup Time	Quantitative
Sleep Duration	Quantitative
REM Sleep Percentage	Quantitative
Deep Sleep Percentage	Quantitative
Light Sleep Percentage	Quantitative
Awakenings	Quantitative
Caffeine Consumption	Quantitative
Alcohol Consumption	Quantitative
Smoking Status	Qualitative

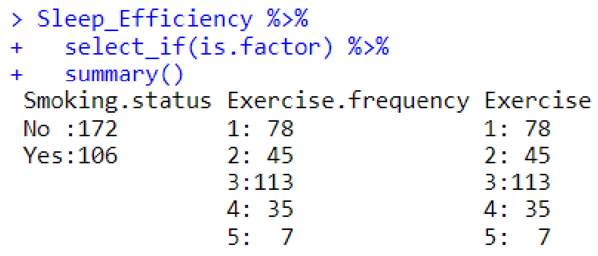
Exercise Frequency	Quantitative

Descriptive Analysis [20 points]: Prepare a numerical and graphical summary of all your study variables. Use the material we have learned in the class and be creative!

Numerical Analysis:

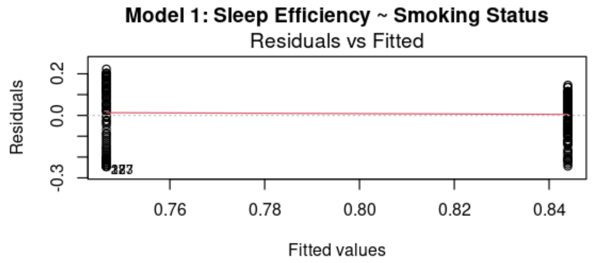


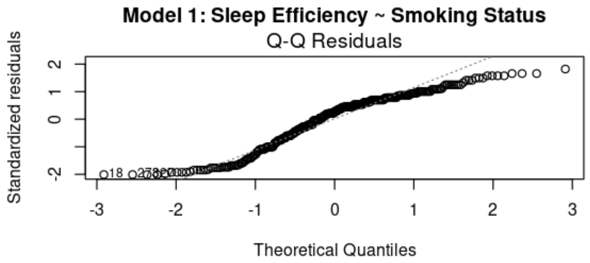
```
> Sleep_Efficiency %>%
+   select_if(is.numeric) %>%
+   summary()
       ID             Age          Awakenings     Sleep.duration   Sleep.efficiency
 Min.   :  1.0   Min.   :18.00   Min.   :0.000   Min.   : 5.000   Min.   :0.5000  
 1st Qu.:110.2   1st Qu.:29.00   1st Qu.:0.250   1st Qu.: 7.000   1st Qu.:0.7200  
 Median :235.0   Median :40.00   Median :1.000   Median : 7.500   Median :0.8500  
 Mean   :227.6   Mean   :40.45   Mean   :1.482   Mean   : 7.406   Mean   :0.8068  
 3rd Qu.:346.8   3rd Qu.:51.00   3rd Qu.:2.750   3rd Qu.: 8.000   3rd Qu.:0.9100  
 Max.   :452.0   Max.   :69.00   Max.   :4.000   Max.   :10.000   Max.   :0.9900  
 REM.sleep.percentage Deep.sleep.percentage Light.sleep.percentage
 Min.   :15           Min.   :18.00         Min.   : 7.00         
 1st Qu.:20           1st Qu.:55.00         1st Qu.:13.25         
 Median :23           Median :60.00         Median :17.50         
 Mean   :23           Mean   :54.61         Mean   :22.39         
 3rd Qu.:27           3rd Qu.:63.00         3rd Qu.:20.75         
 Max.   :30           Max.   :75.00         Max.   :63.00         
 Alcohol.consumption Caffeine.consumption Sleep_Efficiency
 Min.   :0.000       Min.   :  0.00       Min.   :0.5000  
 1st Qu.:0.000       1st Qu.:  0.00       1st Qu.:0.7200  
 Median :0.000       Median :  0.00       Median :0.8500  
 Mean   :1.122       Mean   : 25.18       Mean   :0.8068  
 3rd Qu.:2.000       3rd Qu.: 50.00       3rd Qu.:0.9100  
 Max.   :5.000       Max.   :200.00       Max.   :0.9900  
> 
```

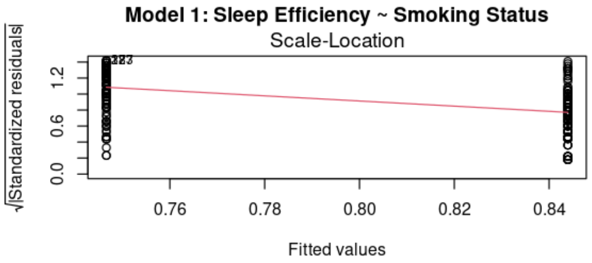


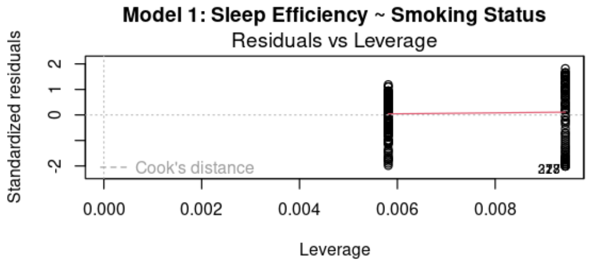
```
> Sleep_Efficiency %>%
+   select_if(is.factor) %>%
+   summary()
 Smoking.status Exercise.frequency Exercise
 No :172        1: 78              1: 78   
 Yes:106        2: 45              2: 45   
                3:113              3:113   
                4: 35              4: 35   
                5:  7              5:  7   
```

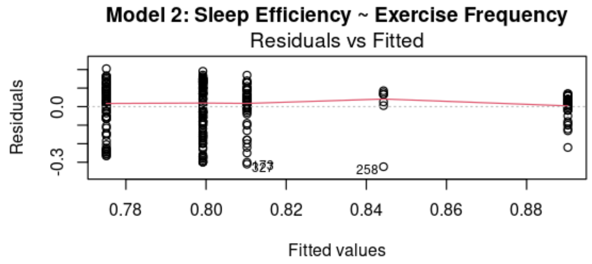
Graphical Analysis:

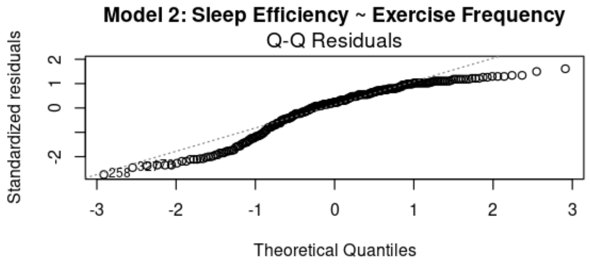

Model 1: Sleep Efficiency ~ Smoking Status
Residuals vs Fitted
Residuals
0.2
0.0
-0.3
0.76
0.78
0.80
0.82
0.84
Fitted values

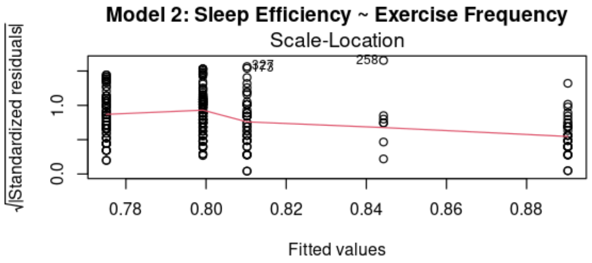

Model 1: Sleep Efficiency ~ Smoking Status
Q-Q Residuals
Standardized residuals
2
1
0
-2
18
-3
-2
-1
0
1
2
3
Theoretical Quantiles

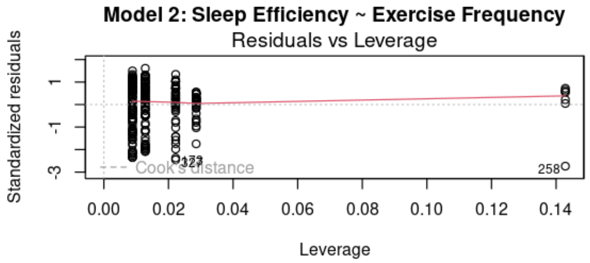

Model 1: Sleep Efficiency ~ Smoking Status
Scale-Location
√|Standardized residuals|
1.2
0.6
0.0
0.76
0.78
0.80
0.82
0.84
Fitted values

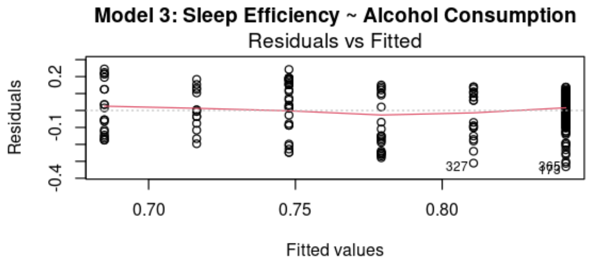

Model 1: Sleep Efficiency ~ Smoking Status
Residuals vs Leverage
Standardized residuals
2
1
0
-2
Cook's distance
0.000
0.002
0.004
0.006
0.008
Leverage

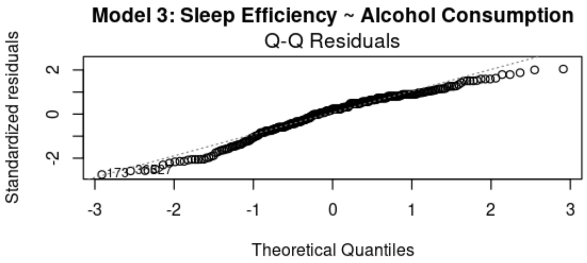

Model 2: Sleep Efficiency ~ Exercise Frequency
Residuals vs Fitted
Residuals
0.0
-0.3
327
258
0.78
0.80
0.82
0.84
0.86
0.88
Fitted values

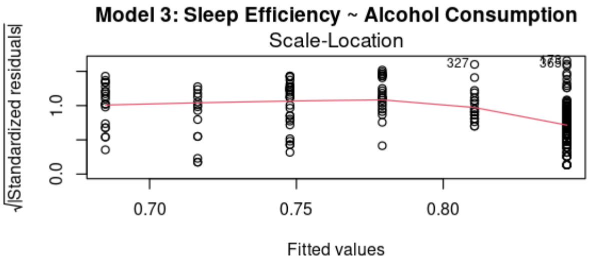

Model 2: Sleep Efficiency ~ Exercise Frequency
Q-Q Residuals
Standardized residuals
2
1
0
-2
258
-3
-2
-1
0
1
2
3
Theoretical Quantiles

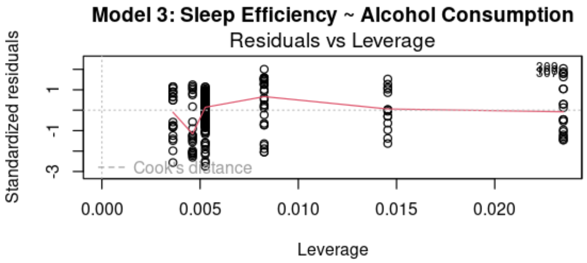

Model 2: Sleep Efficiency ~ Exercise Frequency
Scale-Location
√|Standardized residuals|
1.0
0.0
258
0.78
0.80
0.82
0.84
0.86
0.88
Fitted values

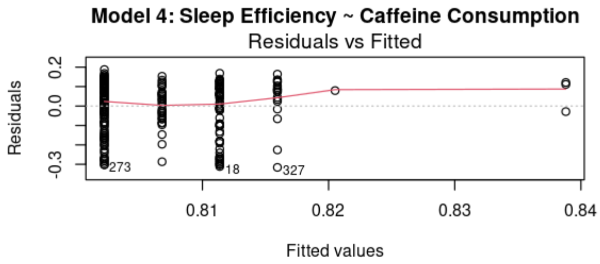

Model 2: Sleep Efficiency ~ Exercise Frequency
Residuals vs Leverage
Standardized residuals
1
-1
-3
Cook's distance
327
258
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Leverage

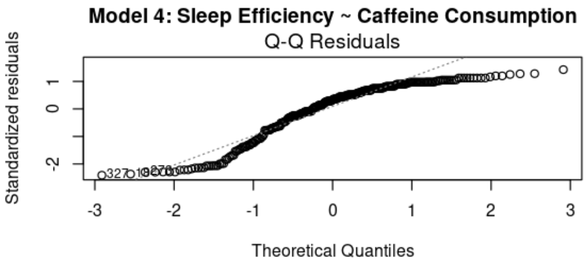

Model 3: Sleep Efficiency ~ Alcohol Consumption
Residuals vs Fitted
Residuals
0.2
-0.1
-0.4
0.70
0.75
0.80
Fitted values
327
365
173

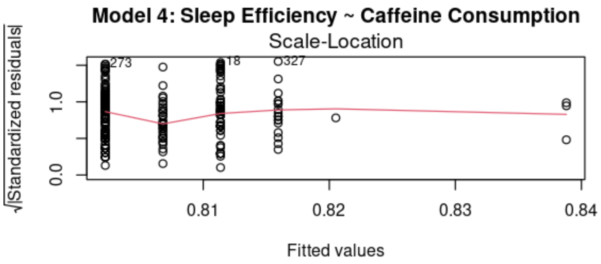

Model 3: Sleep Efficiency ~ Alcohol Consumption
Q-Q Residuals
Standardized residuals
2
0
-2
-3
-2
-1
0
1
2
3
Theoretical Quantiles
173
365
327

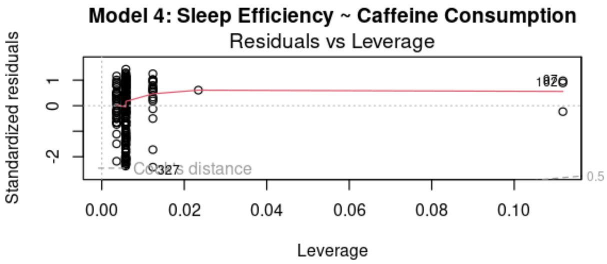

Model 3: Sleep Efficiency ~ Alcohol Consumption
Scale-Location
√|Standardized residuals|
1.0
0.0
0.70
0.75
0.80
Fitted values
327
173
365

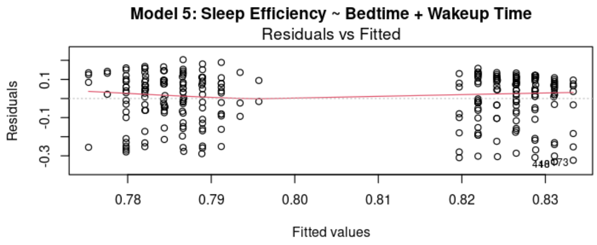

Model 3: Sleep Efficiency ~ Alcohol Consumption
Residuals vs Leverage
Standardized residuals
1
-1
-3
0.000
0.005
0.010
0.015
0.020
Leverage
Cook's distance
389
367

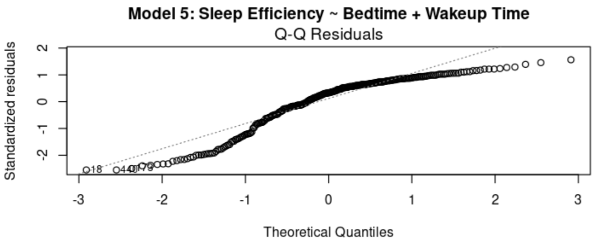

Model 4: Sleep Efficiency ~ Caffeine Consumption
Residuals vs Fitted
Residuals
0.2
0.0
-0.3
0.81
0.82
0.83
0.84
Fitted values
273
18
327

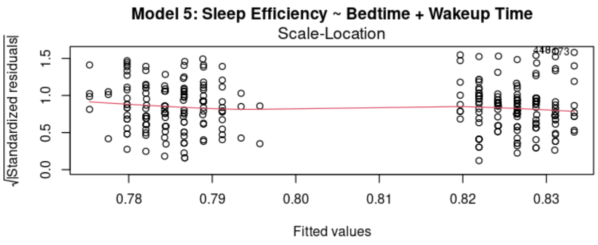

Model 4: Sleep Efficiency ~ Caffeine Consumption
Q-Q Residuals
Standardized residuals
1
0
-2
-3
-2
-1
0
1
2
3
Theoretical Quantiles
327
18
273

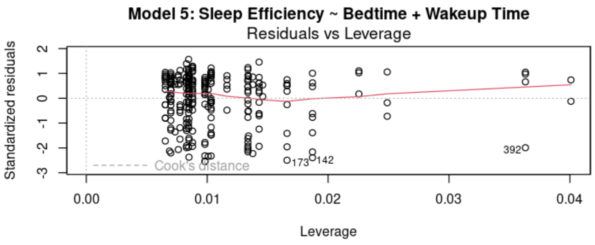

Model 4: Sleep Efficiency ~ Caffeine Consumption
Scale-Location
√|Standardized residuals|
1.0
0.0
0.81
0.82
0.83
0.84
Fitted values
273
18
327

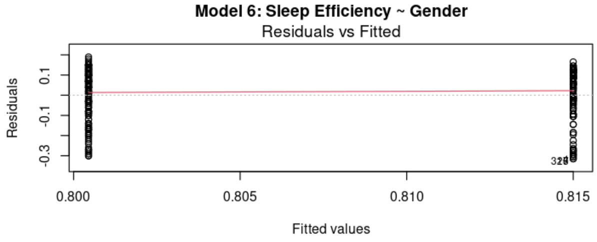

Model 4: Sleep Efficiency ~ Caffeine Consumption
Residuals vs Leverage
Standardized residuals
1
0
-2
0.00
0.02
0.04
0.06
0.08
0.10
Leverage
Cook's distance
327
182
0.5

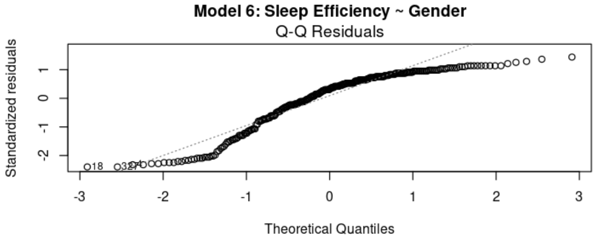

Model 5: Sleep Efficiency ~ Bedtime + Wakeup Time
Residuals vs Fitted
Residuals
0.1
-0.1
-0.3
0.78
0.79
0.80
0.81
0.82
0.83
Fitted values
448
173

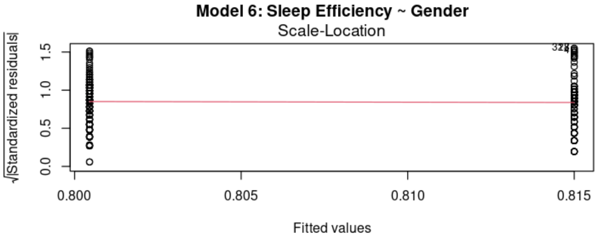

Model 5: Sleep Efficiency ~ Bedtime + Wakeup Time
Q-Q Residuals
Standardized residuals
2
1
0
-1
-2
-3
-2
-1
0
1
2
3
Theoretical Quantiles
18
448
173

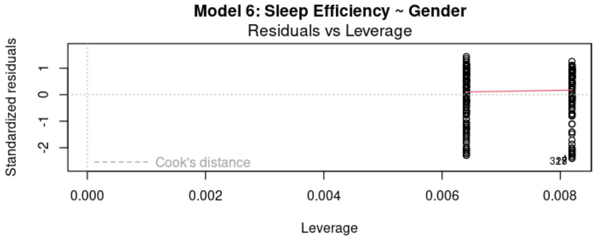

Model 5: Sleep Efficiency ~ Bedtime + Wakeup Time
Scale-Location
√|Standardized residuals|
1.5
1.0
0.5
0.0
0.78
0.79
0.80
0.81
0.82
0.83
Fitted values
448
173

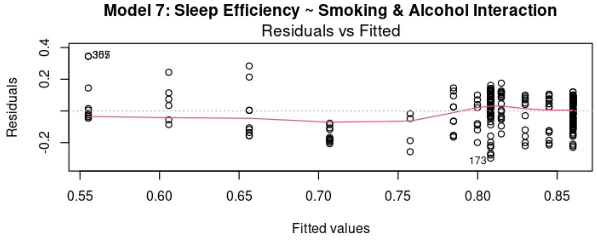

Model 5: Sleep Efficiency ~ Bedtime + Wakeup Time
Residuals vs Leverage
Standardized residuals
2
1
0
-1
-2
-3
0.00
0.01
0.02
0.03
0.04
Leverage
Cook's distance
173
142
392

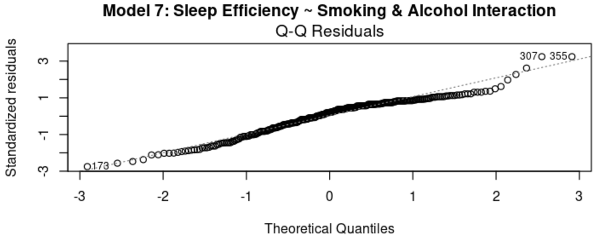

Model 6: Sleep Efficiency ~ Gender
Residuals vs Fitted
Residuals
0.1
-0.1
-0.3
0.800
0.805
0.810
0.815
Fitted values

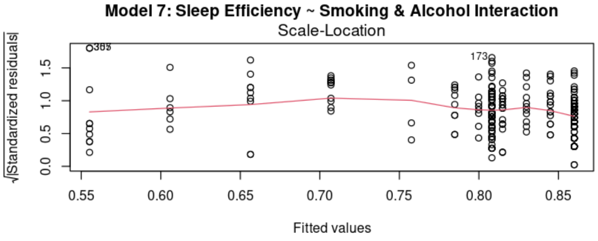

Model 6: Sleep Efficiency ~ Gender
Q-Q Residuals
Standardized residuals
1
0
-1
-2
-3
-2
-1
0
1
2
3
Theoretical Quantiles
18

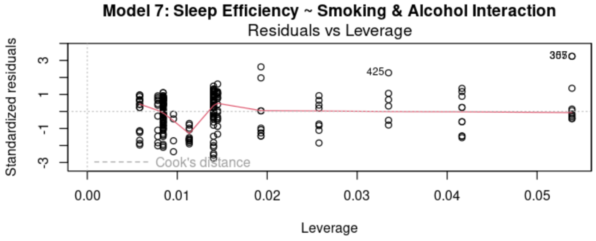

Model 6: Sleep Efficiency ~ Gender
Scale-Location
√|Standardized residuals|
1.5
1.0
0.5
0.0
0.800
0.805
0.810
0.815
Fitted values


Model 6: Sleep Efficiency ~ Gender
Residuals vs Leverage
Standardized residuals
1
0
-1
-2
Cook's distance
0.000
0.002
0.004
0.006
0.008
Leverage


Model 7: Sleep Efficiency ~ Smoking & Alcohol Interaction
Residuals vs Fitted
Residuals
0.4
0.2
-0.2
0.55
0.60
0.65
0.70
0.75
0.80
0.85
Fitted values
173

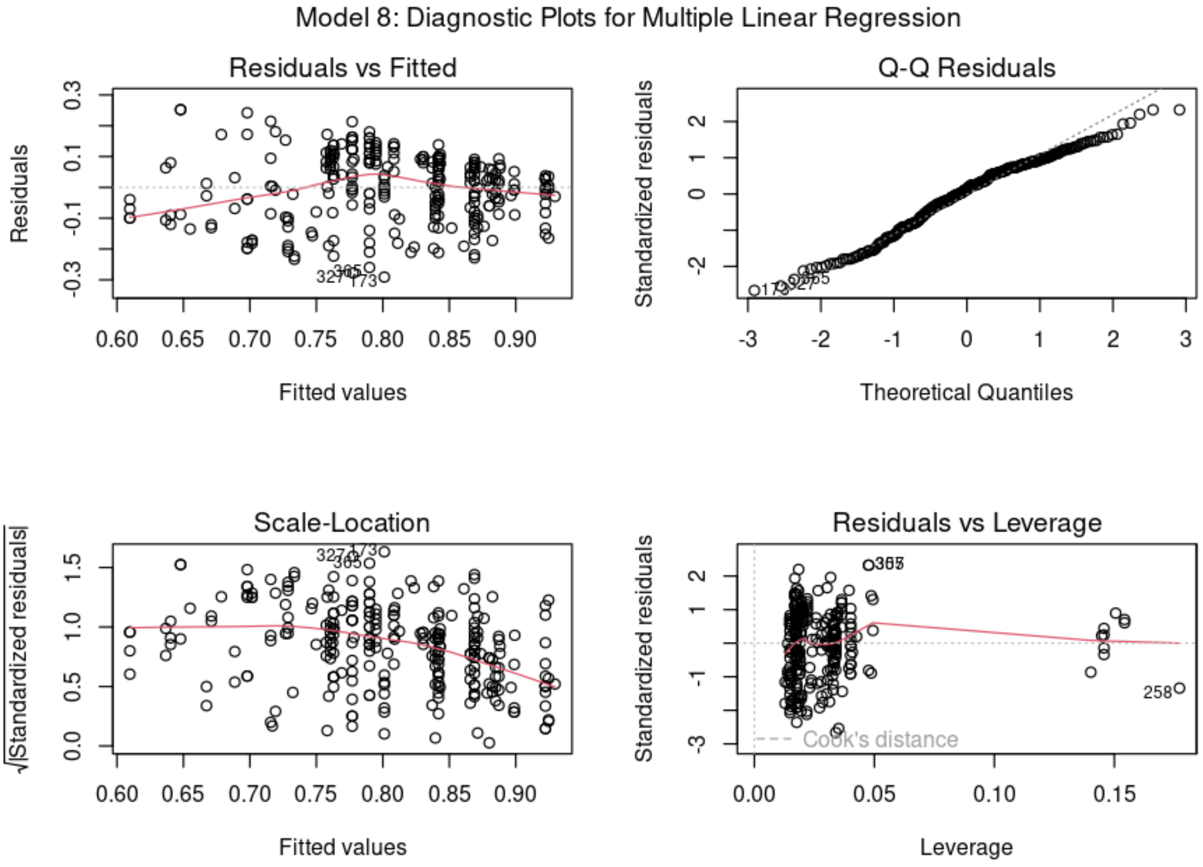

Model 7: Sleep Efficiency ~ Smoking & Alcohol Interaction
Q-Q Residuals
Standardized residuals
3
1
-1
-3
-3
-2
-1
0
1
2
3
Theoretical Quantiles
307
355
173


Model 7: Sleep Efficiency ~ Smoking & Alcohol Interaction
Scale-Location
√|Standardized residuals|
1.5
1.0
0.5
0.0
0.55
0.60
0.65
0.70
0.75
0.80
0.85
Fitted values
173


Model 7: Sleep Efficiency ~ Smoking & Alcohol Interaction
Residuals vs Leverage
Standardized residuals
3
1
-1
-3
Cook's distance
0.00
0.01
0.02
0.03
0.04
0.05
Leverage
425

Modeling:

I) Exploratory Analysis [30 points]



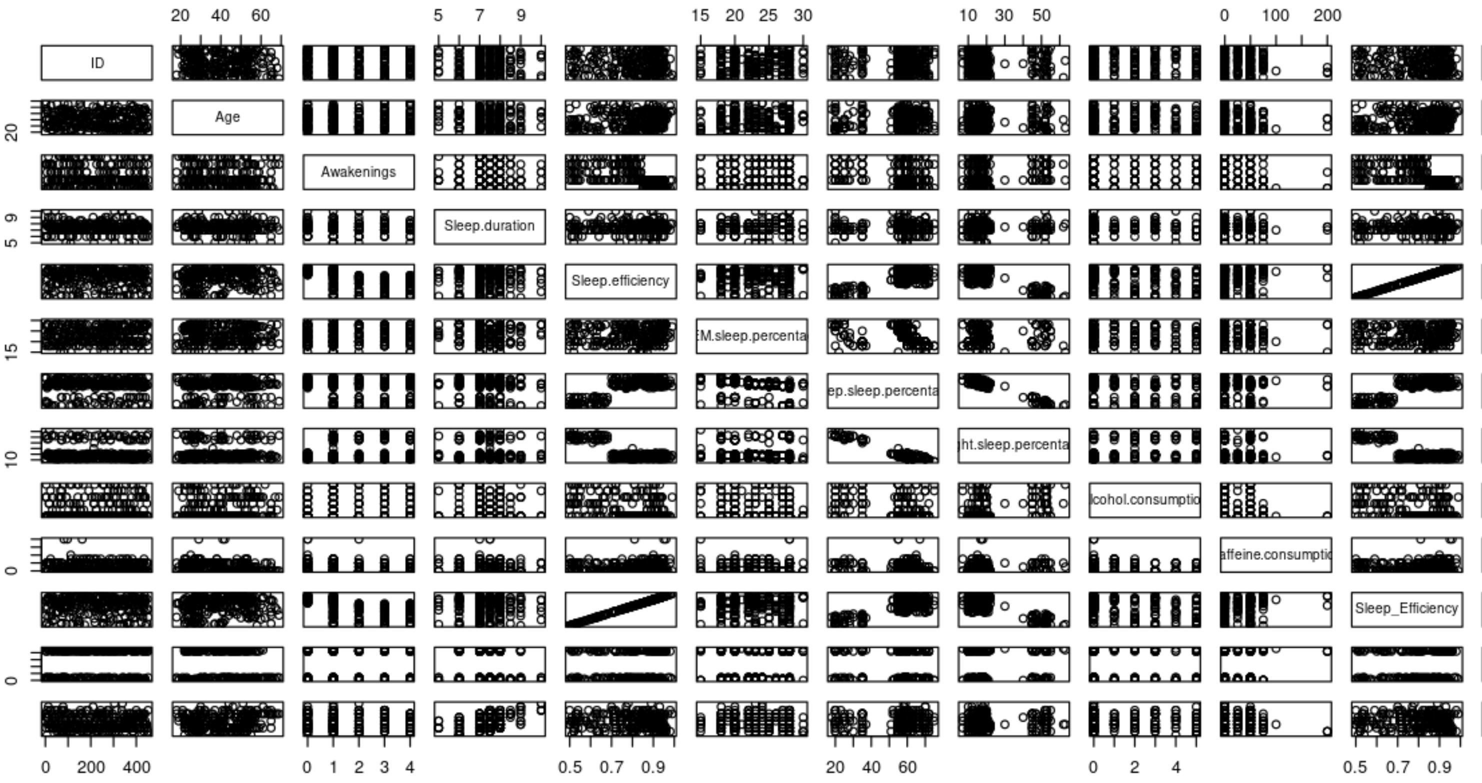
Model 8: Diagnostic Plots for Multiple Linear Regression

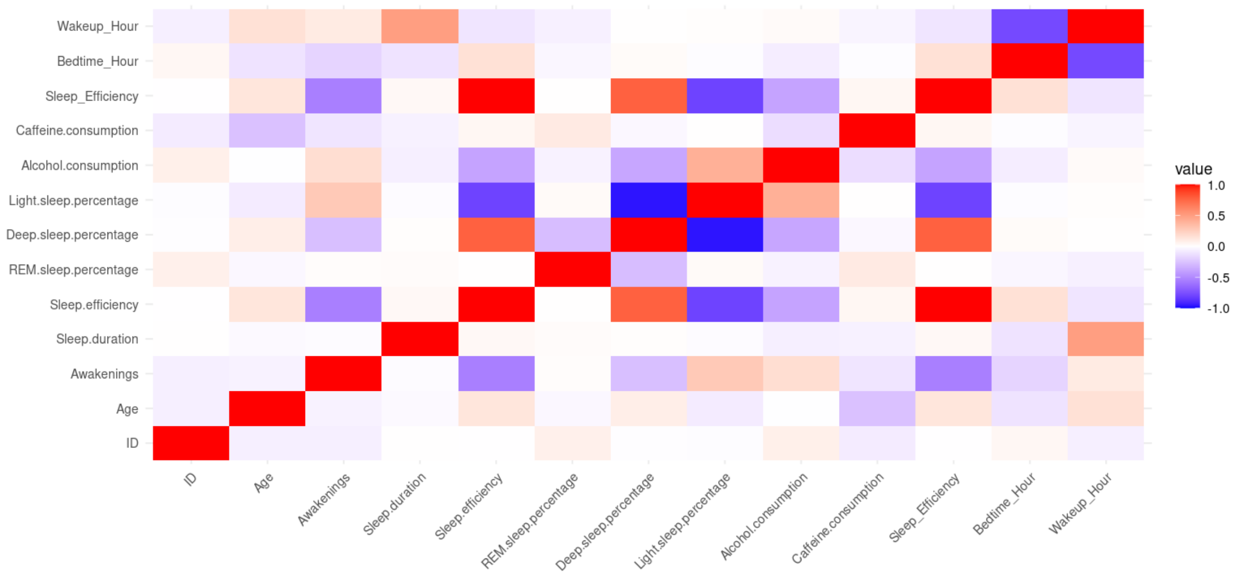
Residuals vs Fitted

Q-Q Residuals

Scale-Location

Residuals vs Leverage





1. State the research goals of the analysis in terms of regression model.

Research Goals in Terms of Regression Model

1. Determine the strength and nature (positive/negative) of the relationships between sleep efficiency and lifestyle factors such as caffeine and alcohol consumption, exercise frequency, and smoking status.

2. Assess the impact of demographic factors like age and possibly infer from bedtime and wake-up times on sleep quality.
3. Explore whether there are interaction effects between these factors that significantly affect sleep efficiency.

2. Study the relationship between all study variables using scatterplot matrix and corrplot. Be as creative as you can be using the R resources. Comment on all findings from these and discuss what would work in favor of your potential model and what might work against it.

Findings from the Scatterplot Matrix and Heatmap

1. **Correlation Observations:** The heatmap shows some variables with strong positive or negative correlations. For example, variables related to the sleep cycle stages (REM, light, and deep sleep percentages) are strongly correlated with each other and with sleep efficiency.

2. **Multicollinearity:** Multicollinearity refers to a situation in which two or more explanatory variables in a multiple regression model are highly linearly correlated.

Smoking.status: VIF is slightly above 1, which suggests a low level of multicollinearity.
Exercise.frequency: VIF is close to 2, which indicates a moderate level of multicollinearity. This could be due to exercise frequency correlating with another variable in the model.
Alcohol.consumption: VIF is slightly above 1, indicating a low level of multicollinearity.
Caffeine.consumption: VIF is close to 1.2, indicating a low level of multicollinearity.

There is some multicollinearity present, but it is not severe enough to distort the regression coefficients excessively.

3. **Outliers and Data Spread:** The scatterplot matrix reveals the presence of outliers in some variables and the distribution of data points, which could influence the regression model. Outliers in the Awakenings or Alcohol.consumption variables need further investigation or data preprocessing steps.

4. **Non-linear Relationships:** Some scatter plots suggest non-linear relationships, such as between age and certain sleep metrics, which means that linear regression may not be the best fit for these variables. A transformation or a different type of model, like polynomial regression, might be required.

II) Multiple Linear Regression [100 points]

- [10 points] Fit a MLR model to your data and report the fitted model. Report the goodness of fit of your model. Do not use any categorical or numeric discrete predictors variables.

$\widehat{Y}$ = 0.2479668+0.0008339(Age)−0.0325232(Awakenings)+0.0078798(Sleep Duration)+ 0.0073938(REM Sleep Percentage)+ 0.0065146(Deep Sleep Percentage)−0.0028165(Alcohol Consumption)+ 0.0001181(Caffeine Consumption)+ 0.0021404(GenderMale)+ 0.0008451(Bedtime)−0.0031928(Wakeup Time)

Goodness of Fit Measures:

1. Coefficient of Determination (R-squared):

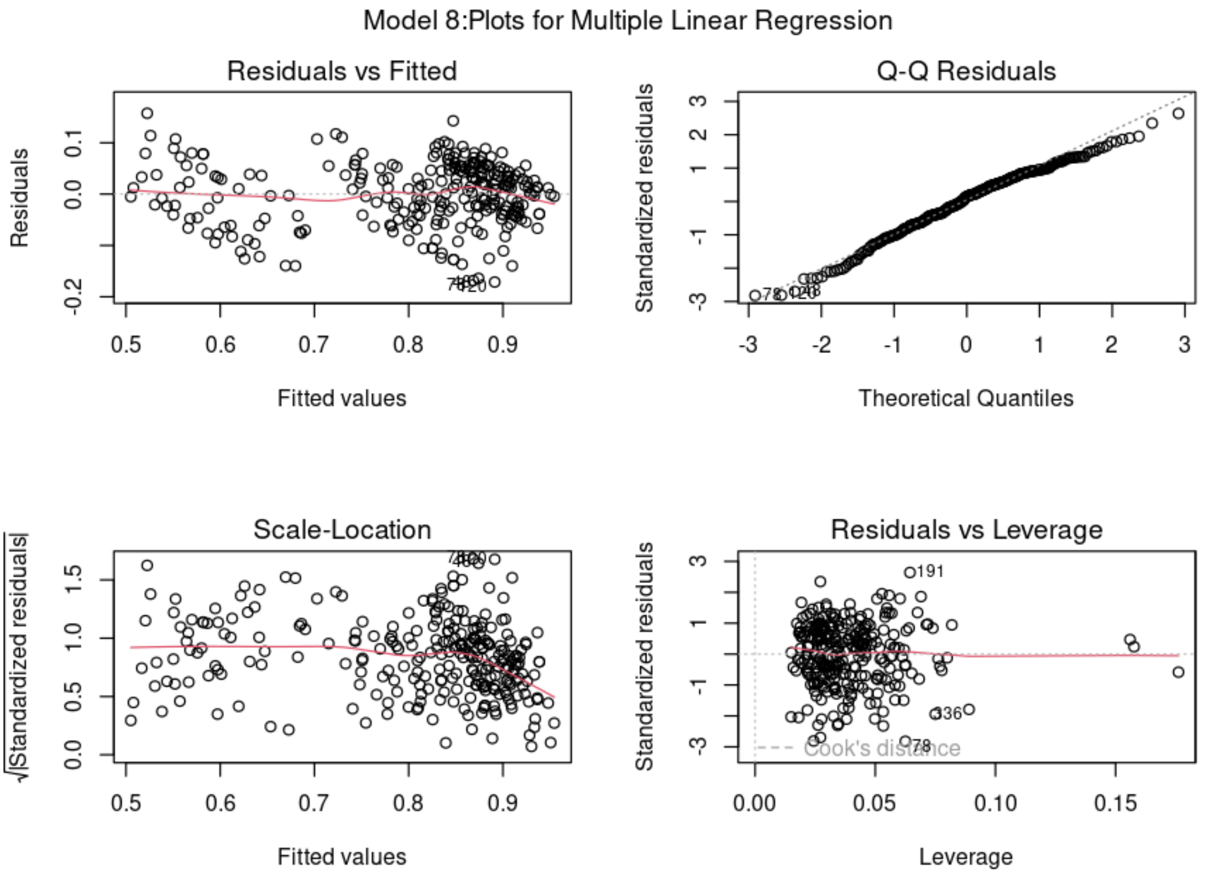
The R-squared value is 0.7891, indicating that approximately 78.91% of the variability in Sleep Efficiency is explained by the model's predictors.

2. Variability in Errors (Residual Standard Error):

The residual standard error (RSE) is 0.06168, which is pretty low, indicating a good predictive accuracy.

3. Residual Plot Analysis:

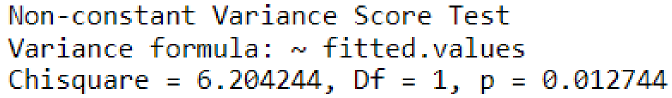
The residual plots suggest that this regression model does not fully meet the assumptions of linear regression. The Residuals vs Fitted plot indicates potential non-linearity and heteroscedasticity, as the residuals increase with the fitted values. The Q-Q plot shows deviations from normality, particularly with extreme values.The Scale-Location plot also suggests heteroscedasticity, and the Residuals vs Leverage plot highlights a few outliers



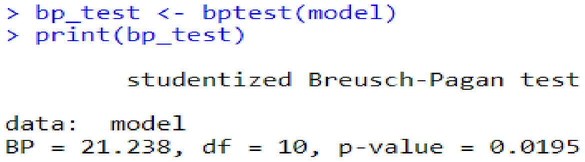
- [10 points] Report the LINE conditions and state whether the assumptions are met or not. Use all techniques we have learned and practiced.

Linearity: The scatter plot of observed vs fitted sleep efficiency (Fig(a)) suggests almost a linear relationship, as the points cluster around a line. However, a slight curvature is visible, indicating potential non-linearity.
Normality of Errors: The histogram of residuals (Fig(b)) and the Q-Q plot (not shown) would be used to assess the normality of errors. If the histogram is bell-shaped and the Q-Q plot follows the line closely, the normality assumption is met.
Equal Variance (Homoscedasticity): The residual plot (Fig(d)) and the residuals index plot (Fig(e)) should display a random spread of residuals to support homoscedasticity. The Breusch-Pagan test result (p-value = 0.0195) indicates potential heteroscedasticity, contradicting the visual assessment.
Independence: The residuals index plot (Fig(e)) should show no obvious patterns or trends if errors are independent. This seems to be the case in the provided plot.
Tukey's Test: A non-constant variance score test with a p-value of 0.012744 suggests that the variance of the residuals may be changing with the fitted values, which violates the assumption of homoscedasticity.

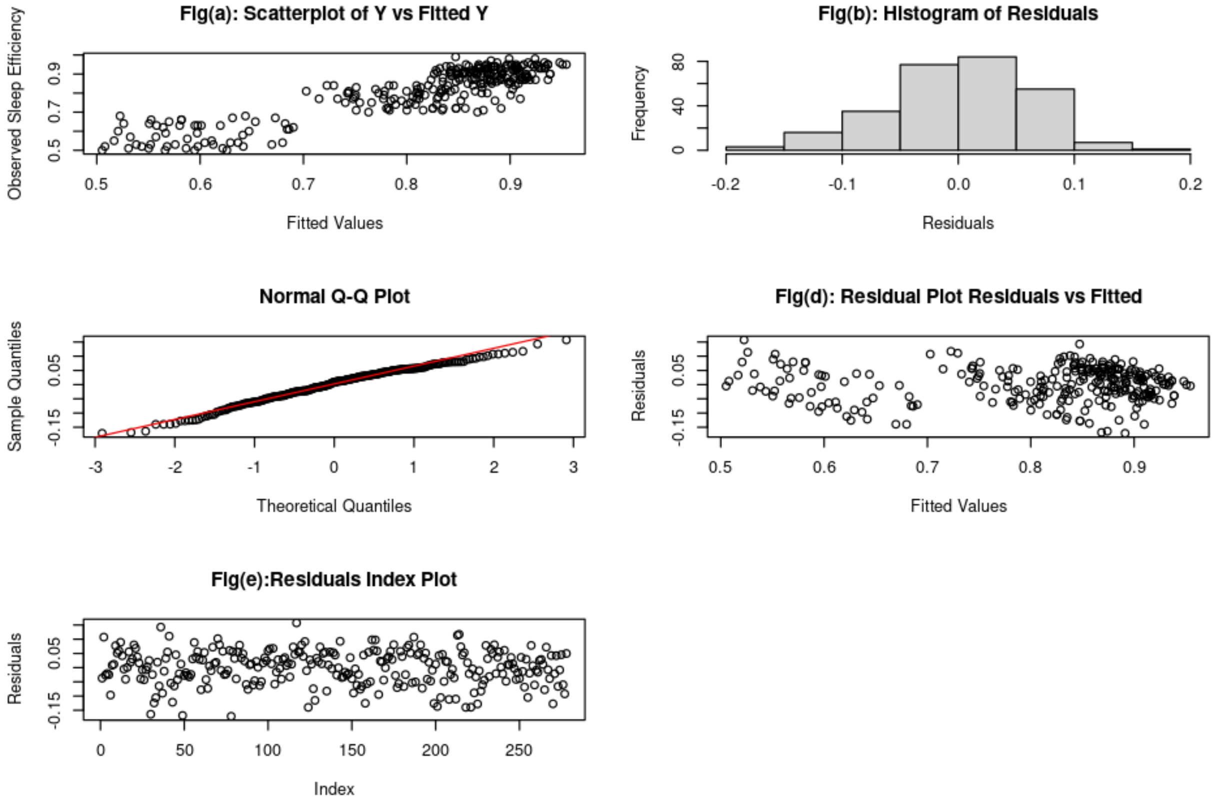
Overall, We see that the linearity and independence assumptions appear to be met, the homoscedasticity assumption fails meaning Breusch-Pagan and Tukey's test result fails .



```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 6.204244, Df = 1, p = 0.012744
```



```
> bp_test <- bptest(model)
> print(bp_test)

        studentized Breusch-Pagan test

data:  model
BP = 21.238, df = 10, p-value = 0.0195
```



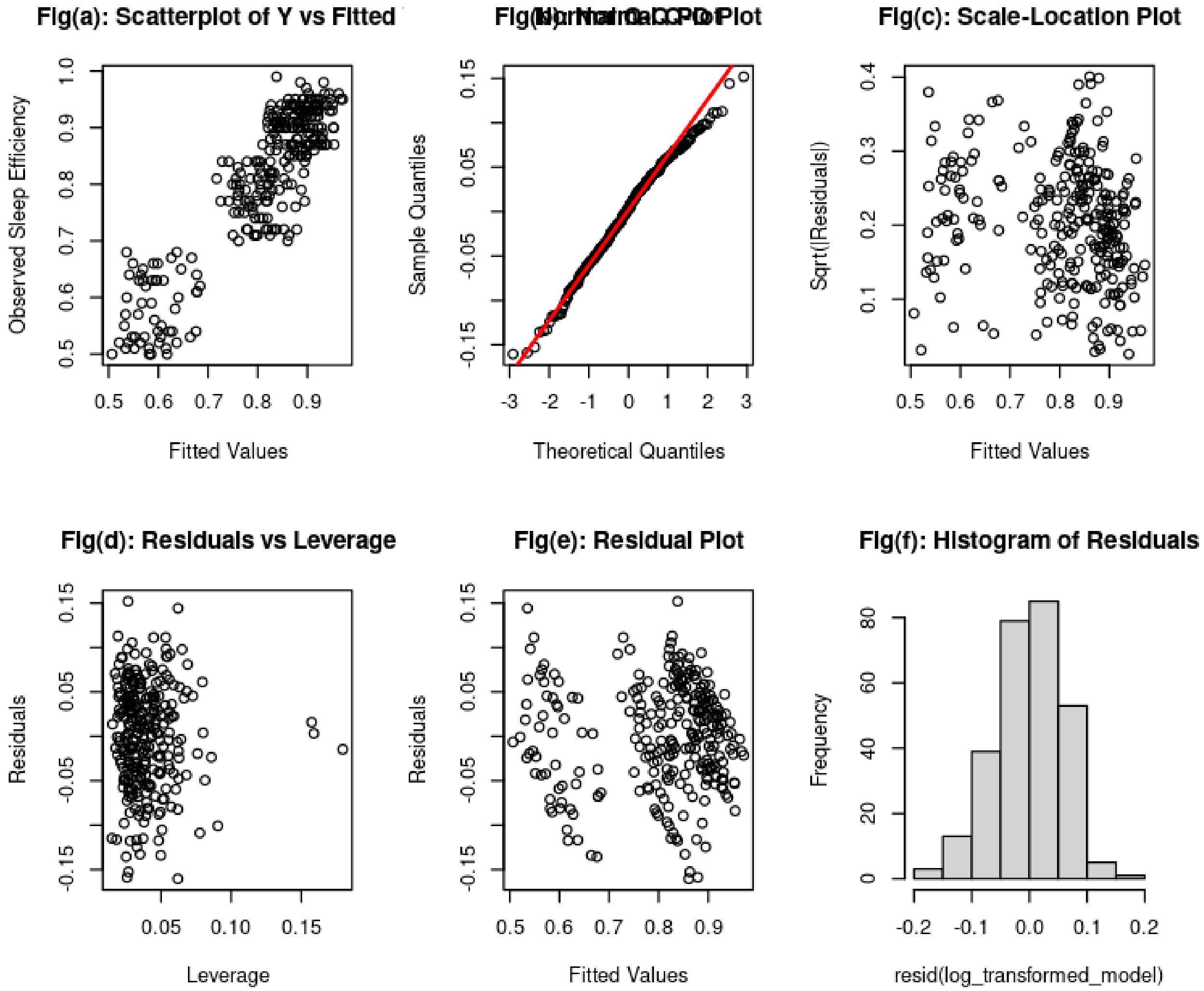
- [10 points] If some of the assumptions were not met, then try transformation. Report any transformations that you did and present the fitted model. Discuss the goodness of
fit of this model.

Model Transformation:

We observed heteroscedasticity in our previous model, as indicated by the Breusch-Pagan test and Tukey’s curve test. We then tried the log and square root transformations, which did not fully satisfy these assumptions. The Box-Cox transformation, which is a power transformation that includes log and square root as special cases, didn't resolve the issues either. The polynomial transformation made a slight improvement, but there were still concerns with the Breusch-Pagan test indicating potential heteroscedasticity.

Since all transformations have not led to satisfactory diagnostics, we tried to address this issue by applying a log transformation to some predictor variables, which are known to be skewed or have a multiplicative relationship with the response variable. This approach resulted in a partially log-transformed model that showed marked improvements in the diagnostic plots, suggesting better compliance with the LINE assumptions. Specifically, the normal Q-Q plot and the residuals vs leverage plot indicated that the normality and independence assumptions were reasonably met. The Breusch-Pagan test still indicates potential heteroscedasticity (p-value = 0.01595), but we can definitely see improvements in the model.

Transformation Type	R-squared	Adjusted R-squared	Residual Standard Error	BP Test p-value	Tukey Test p-value
None (Original)	0.7891	0.7812	0.06168	0.0195	0.012744
Log	0.8004	0.7929	0.08236	6.038e-05	8.9866e-07
Square Root	0.7959	0.7882	0.0354	0.002214	0.00030484
Box-Cox	0.4997	0.481	6.161e-16	0.06715	1
Polynomial	0.7916	0.783	0.06142	0.04712	0.013642
Partial Log	0.8007	0.7932	0.05996	0.01595	0.030518



Fig(a): Scatterplot of Y vs Fitted

Fig(b): Normal Q-Q Plot

Fig(c): Scale-Location Plot

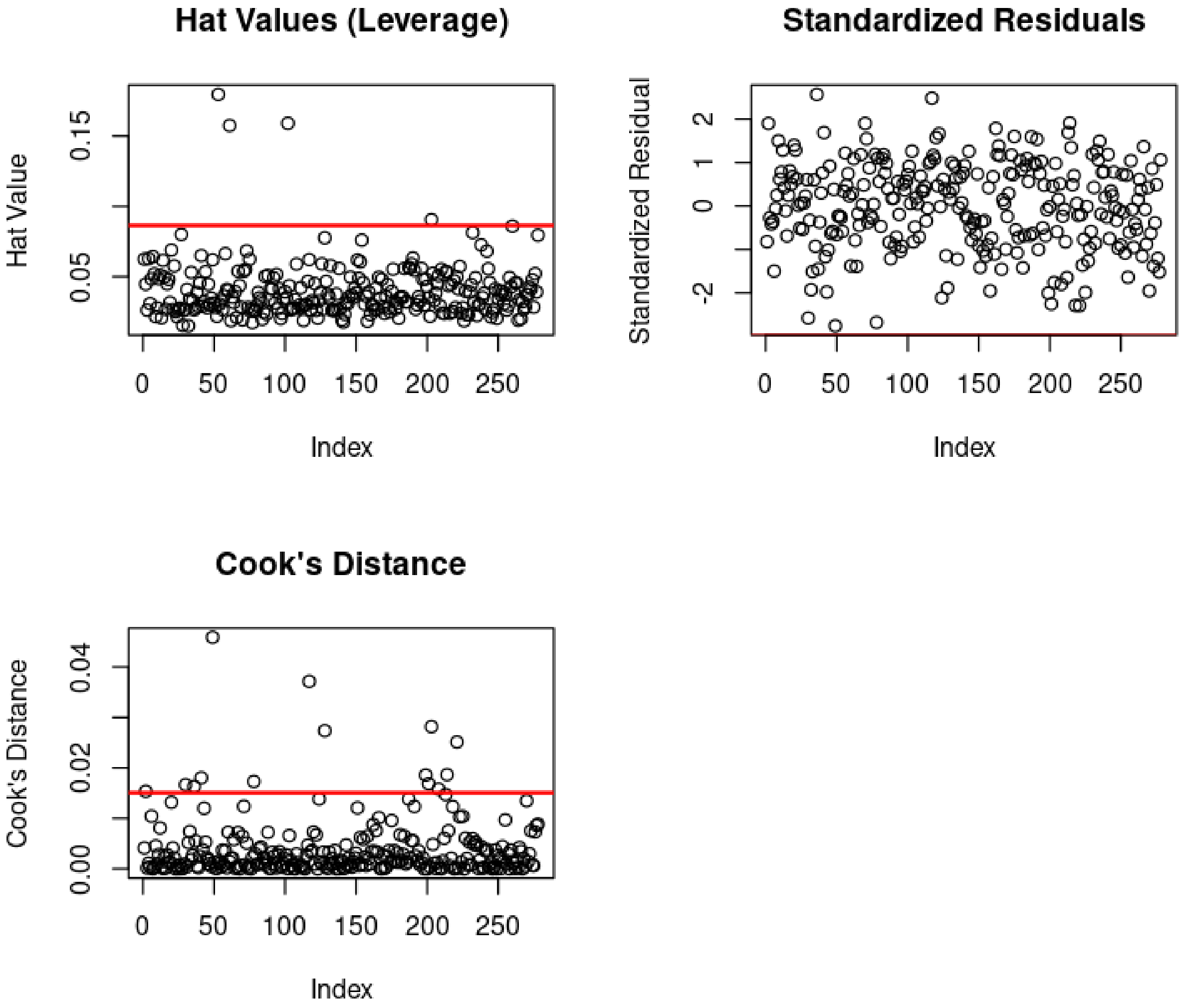
Fig(d): Residuals vs Leverage

Fig(e): Residual Plot

Fig(f): Histogram of Residuals

Note that for the rest of the points, your model is the original MLR model you reported if no transformation was done and it is the transformed model if transformation was done.

- [10 points] Detect if there are any points with high leverage, outliers, and influence in the original model. Show all steps in the analysis.



Hat Values (Leverage) Plot:

Most data points have low leverage with hat values well below the red threshold line. This suggests that there are no individual observations that are influencing the regression model due to extreme predictor values. There are a few points with higher leverage indicated by their position above the baseline but only a couple are above the red threshold line.

Standardized Residuals Plot:

The residuals appear randomly scattered around the horizontal axis, which suggests that the variance of the residuals is constant (homoscedasticity).There are some points that lie further away from the horizontal line at 0, indicating potential outliers. However, none of the residuals exceed the commonly used absolute value of 2 for standardized residuals, which would typically indicate a strong outlier.

Cook's Distance Plot:

Most points have Cook's distance far below the red line, indicating they are not influential. There are a few points with higher Cook's distance, but only a couple are above the red threshold line.Summary of Interpretation:

So in conclusion, we can say that a couple of points have high leverage, but the vast majority do not overly influence the regression model. We can not see extreme outliers based on standardized residuals within this dataset. A couple of points have higher influence on the model as indicated by Cook's distance, but the influence is not widespread throughout the dataset.

- [15 points] Remove influential points and refit the model to the remaining data Discuss the goodness of fit and LINE conditions as compared to the original model.

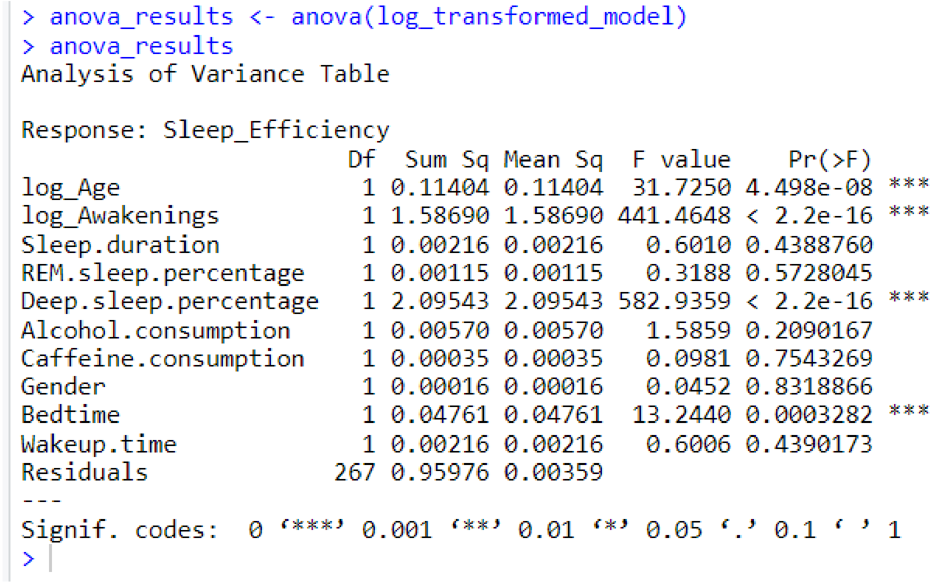
Goodness of Fit Comparison Table

Measure	Original Model	Refitted Model
R-squared	0.8007	0.8378
Adjusted R-squared	0.7932	0.8314
Residual Standard Error (RSE)	0.05996	0.05326
Leverage Points	Several with high leverage	Fewer high leverage points
Cook's Distance	Several points above threshold	Fewer points above threshold
Normality of Residuals	Q-Q plot shows slight deviation from linearity at extremes	Q-Q plot shows a tighter fit to the line
Residuals vs Fitted	Some patterns visible, suggesting possible nonlinearity or heteroscedasticity	Less patterned, suggesting better homoscedasticity
Scale-Location	Spread varies with fitted values, indicating possible heteroscedasticity	More even spread, suggesting improved homoscedasticity

In summary, the refitted model shows an improvement in the R-squared value, indicating a better fit to the data. The reduction in the residual standard error suggests that the predictions are, on average, closer to the actual values. The diagnostic plots for the refitted model show fewer signs of problematic leverage or influence points, and the residuals appear to be more normally distributed and homoscedastic compared to the original model.

This indicates that the assumptions of linear regression (LINE conditions) are better met with the refitted model

- [10 points] Report ANOVA of model from above. Show steps for overall fit of your regression.



```
> anova_results <- anova(log_transformed_model)
> anova_results
Analysis of Variance Table

Response: Sleep_Efficiency
                        Df  Sum Sq Mean Sq  F value    Pr(>F)    
log_Age                  1 0.11404 0.11404  31.7250 4.498e-08 ***
log_Awakenings           1 1.58690 1.58690 441.4648 < 2.2e-16 ***
Sleep.duration           1 0.00216 0.00216   0.6010 0.4388760    
REM.sleep.percentage     1 0.00115 0.00115   0.3188 0.5728045    
Deep.sleep.percentage    1 2.09543 2.09543 582.9359 < 2.2e-16 ***
Alcohol.consumption      1 0.00570 0.00570   1.5859 0.2090167    
Caffeine.consumption     1 0.00035 0.00035   0.0981 0.7543269    
Gender                   1 0.00016 0.00016   0.0452 0.8318866    
Bedtime                  1 0.04761 0.04761  13.2440 0.0003282 ***
Wakeup.time              1 0.00216 0.00216   0.6006 0.4390173    
Residuals              267 0.95976 0.00359                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```

From the ANOVA table, log_Age, log_Awakenings, and Deep.sleep.percentage have very small p-values (indicated by the ***), which suggests they are statistically significant predictors of Sleep_Efficiency. Bedtime also appears to be a significant predictor (indicated by ***), although to a lesser extent.
The overall fit of the model can be assessed by looking at the F-statistic and its corresponding p-value for the entire model (which is not shown in the individual predictor rows but would be in the ANOVA output). If the p-value for the overall F-statistic is small (usually less than 0.05), so we can reject the null hypothesis that all the regression coefficients are equal to zero, which means that the model is a good fit for the data. For residual analysis, the residual sum of squares (0.95976) and the residual mean square (0.00359) gives variability in the data that the model is not explaining.

- [5 points] Discuss if all predictors are statistically significant or not.

log_Age: With a p-value of approximately 4.498×10^{-8} (indicated by ***), it is statistically significant. This means that age, when log-transformed, has a strong relationship with sleep efficiency.

log_Awakenings: This predictor has a very low p-value (< 2.2e-16, indicated by ***), which means it is highly significant and has a strong negative relationship with sleep efficiency, as indicated by its negative coefficient.
Sleep.duration: The p-value is 0.4388760, which is greater than the conventional alpha level of 0.05, indicating that sleep duration is not statistically significant in this model.
REM.sleep.percentage: This predictor is not statistically significant in the model, as indicated by the p-value of 0.5728045.
Deep.sleep.percentage: With a p-value of < 2.2e-16 (indicated by ***), it is highly significant, suggesting a strong relationship with sleep efficiency.

Alcohol.consumption: The p-value is 0.2090167, which is not statistically significant.

Caffeine.consumption: With a p-value of 0.7543269, caffeine consumption is not a statistically significant predictor.

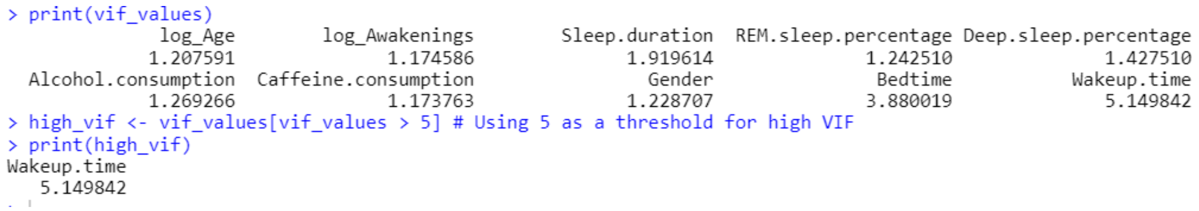
Gender: The p-value of 0.8318866 indicates that gender is not statistically significant in this model.

Bedtime: With a p-value of 0.0003282 (indicated by ***), bedtime is statistically significant.

Wakeup.time: The p-value of 0.4390173 suggests that wakeup time is not statistically significant.

In summary, log_Age, log_Awakenings, Deep.sleep.percentage, and Bedtime are statistically significant predictors in this model, while Sleep.duration, REM.sleep.percentage, Alcohol.consumption, Caffeine.consumption, Gender, and Wake Up.time are not.

- [10 points] Check for multicollinearity in the predictors and report the VIF and discuss.

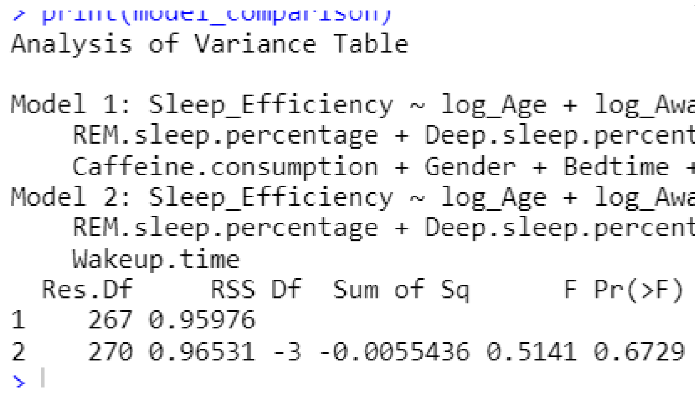


```
> print(vif_values)
              log_Age        log_Awakenings        Sleep.duration  REM.sleep.percentage Deep.sleep.percentage 
             1.207591              1.174586              1.919614              1.242510              1.427510 
  Alcohol.consumption  Caffeine.consumption                Gender               Bedtime           Wakeup.time 
             1.269266              1.173763              1.228707              3.880019              5.149842 
> high_vif <- vif_values[vif_values > 5] # Using 5 as a threshold for high VIF
> print(high_vif)
Wakeup.time 
   5.149842 
```

The VIFs for variables such as log_Age, log_Awakenings, Sleep.duration, REM.sleep.percentage, Deep.sleep.percentage, Alcohol.consumption, Caffeine.consumption, and Gender are all well below the commonly used threshold of 5, which is good because it indicates low multicollinearity.

However, the Wakeuptime has a VIF slightly above 5, which indicates a moderate level of multicollinearity.

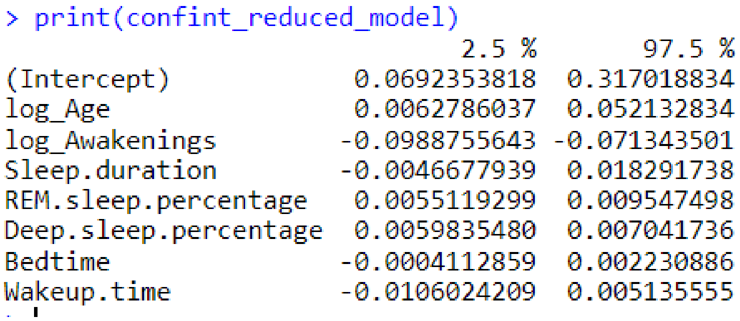
- [10 points] Fit a reduced model with the highly correlated and statistically insignificant predictors removed. Compare ANOVA for the original model as full model and reduced model. Show all steps and which model you would keep.



```
> print(model_comparison)
Analysis of Variance Table

Model 1: Sleep_Efficiency ~ log_Age + log_Awakenings + Sleep.duration + 
    REM.sleep.percentage + Deep.sleep.percentage + Alcohol.consumption + 
    Caffeine.consumption + Gender + Bedtime + Wakeup.time
Model 2: Sleep_Efficiency ~ log_Age + log_Awakenings + Sleep.duration + 
    REM.sleep.percentage + Deep.sleep.percentage + Bedtime + 
    Wakeup.time
  Res.Df     RSS Df  Sum of Sq      F Pr(>F)
1    267 0.95976                            
2    270 0.96531 -3 -0.0055436 0.5141 0.6729
> 
```

The reduced model, with insignificant predictors removed, has a slightly higher Residual Sum of Squares (RSS) than the full model, indicating a marginal increase in unexplained variance.The increase in degrees of freedom in the reduced model (from 267 to 270) reflects the removal of three parameters.The F-statistic and its p-value (0.6729) suggest that the removal of these parameters does not significantly worsen the model's fit.
This shows that the simpler reduced model is nearly as effective as the full model in explaining the variance in Sleep Efficiency. Therefore, the reduced model may be preferred for its simplicity without a significant loss in predictive accuracy.

- [10 points] For the model selected in step 12, construct 95% confidence intervals for all regression coefficients and interpret them.



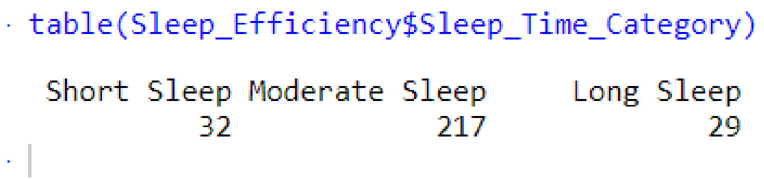
```
> print(confint_reduced_model)
                               2.5 %       97.5 %
(Intercept)             0.0692353818  0.317018834
log_Age                 0.0062786037  0.052132834
log_Awakenings         -0.0988755643 -0.071343501
Sleep.duration         -0.0046677939  0.018291738
REM.sleep.percentage    0.0055119299  0.009547498
Deep.sleep.percentage   0.0059835480  0.007041736
Bedtime                -0.0004112859  0.002230886
Wakeup.time            -0.0106024209  0.005135555
```

Summary of 95% Confidence Intervals for the Reduced Model Coefficients:

Intercept: The confidence interval ranges from 0.069 to 0.317, suggesting a moderate baseline level of sleep efficiency when all predictors are at their zero level(may not be practical)
log_Age: The interval spans 0.006 to 0.052, indicating a generally positive, though modest, association between age and sleep efficiency. As age increases logarithmically, so does sleep efficiency, within this range.
log_Awakenings: With a confidence interval of -0.099 to -0.071, this predictor shows a clear negative impact on sleep efficiency. More awakenings, when log-transformed, tend to significantly decrease sleep efficiency.
Deep.sleep.percentage: The positive interval of 0.00598 to 0.00704 confirms that higher percentages of deep sleep are consistently associated with increased sleep efficiency, so it is significant.
Wakeup.time: The wide and zero-crossing interval of -0.0106 to 0.0051 suggests uncertainty about the impact of wakeup time on sleep efficiency. It could be negative, positive, or insignificant.

III) Dummy Variables [60 points]

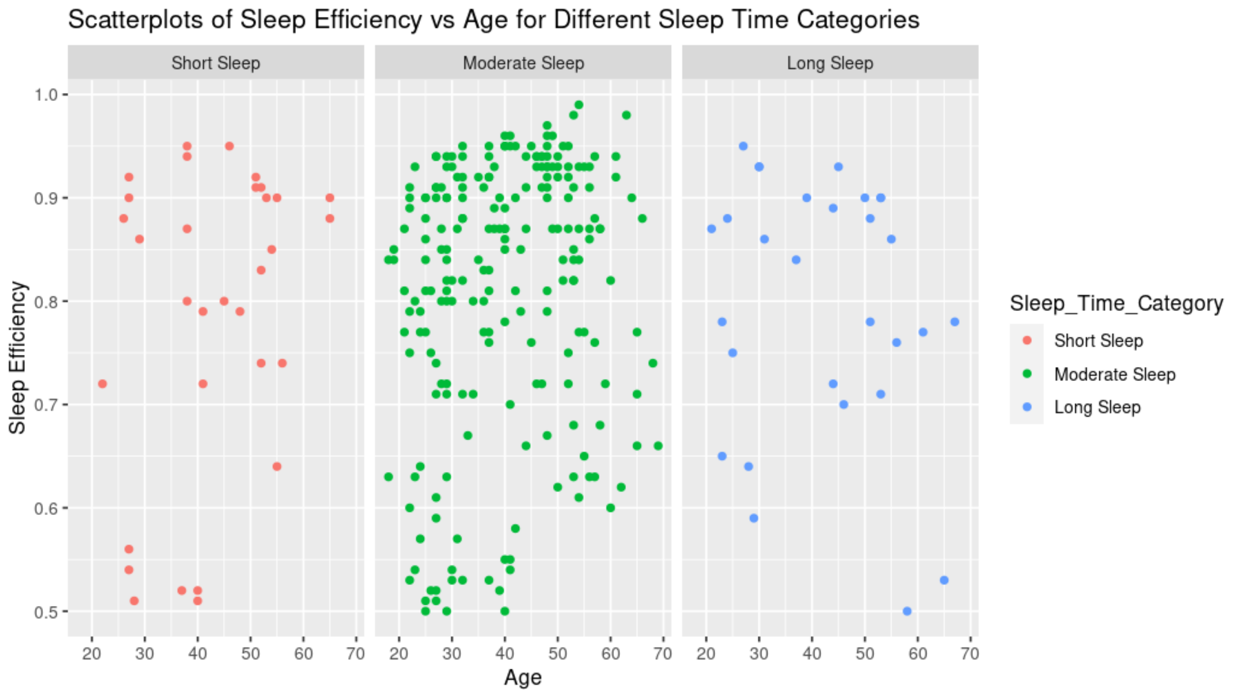
- [5 points] Now use the categorical variables with three or more categories [Note that you can create categorical variables as we did for the autompg data example and I can help you with this]. Explain your categorical variable and different categories.



```
table(Sleep_Efficiency$Sleep_Time_Category)

  Short Sleep Moderate Sleep     Long Sleep 
           32            217             29 
```

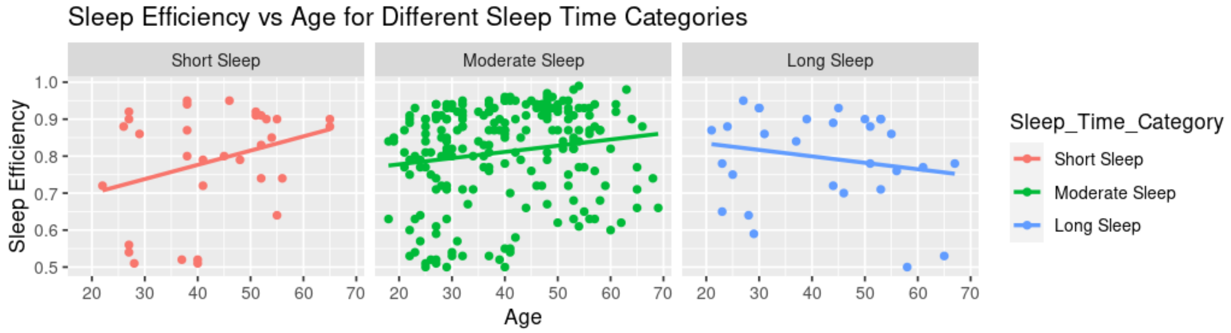
In this analysis, I recognized the need to explore the impact of different sleep durations on sleep efficiency. To do this, I created a new categorical variable named 'Sleep_Time_Category.' This variable classifies the sleep duration of individuals into three distinct categories: Short Sleep, Moderate Sleep, and Long Sleep.
I defined these categories based on the sleep duration data in my dataset. Specifically, 'Short Sleep' represents a sleep duration of less than 6 hours, 'Moderate Sleep' includes sleep durations between 6 and 8 hours, and 'Long Sleep' denotes sleep durations exceeding 8 hours. This categorization is informed by general sleep recommendations and is intended to capture varying sleep patterns among individuals.

- [10 points] Show the scatterplots of Y vs X (numeric) varies for different categories in scatterplots.

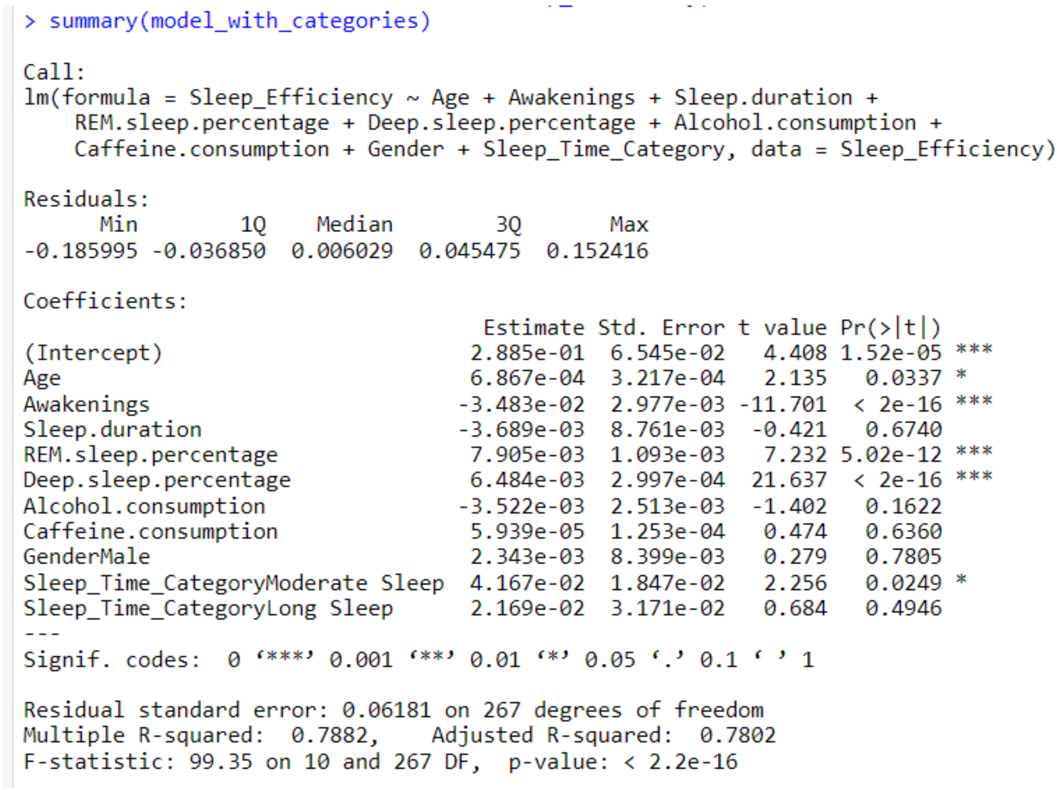


I categorized sleep duration into three groups: short, moderate, and long sleep, to see if the amount of sleep has an effect on the efficiency. I noticed that for moderate sleepers, there is a dense cluster of points, indicating many observations in this category. I see that the range of sleep efficiency for short sleepers is wider, suggesting more variability in this group. I can observe from the long sleep category that fewer individuals fall into this group, but there seems to be a trend towards higher efficiency with age.

- [15 points] Include categorical variables in the MLR model selected above with intercepts different for different categories. Show fitted models for each category, plot of fitted model for each category. Interpret the coefficients in context of your goals.

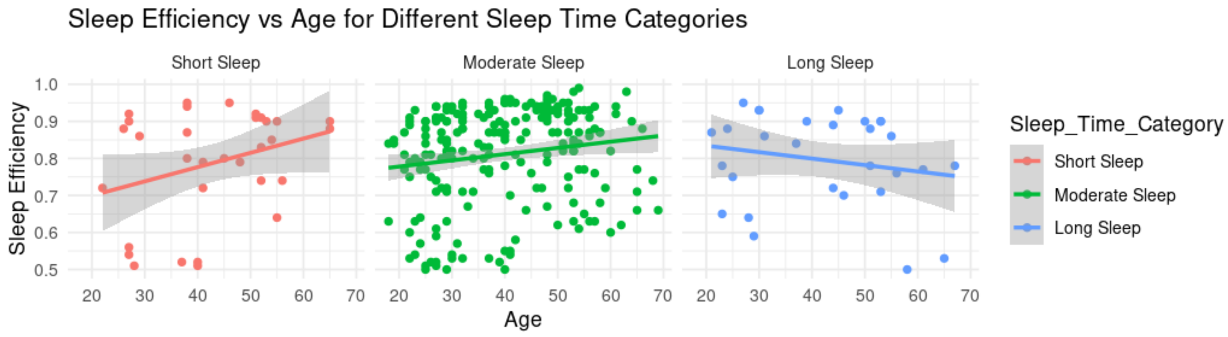


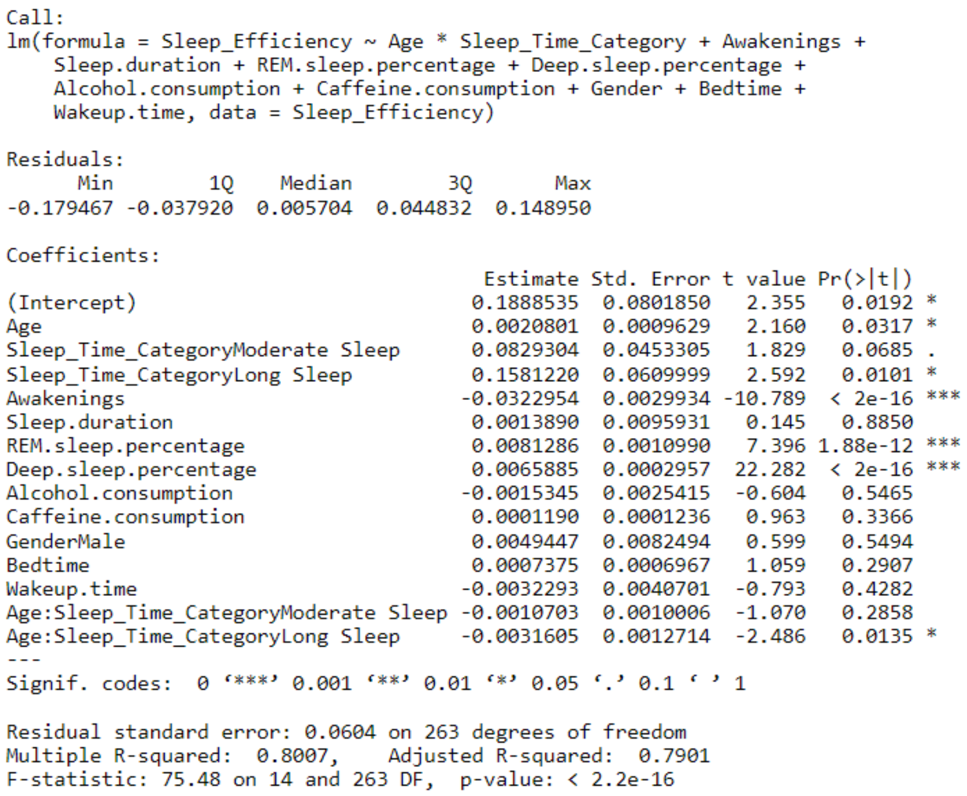
The graph and summary together show a clear relationship between age, awakenings, REM sleep, and deep sleep with sleep efficiency. Specifically, as age increases, and with higher percentages of REM and deep sleep, sleep efficiency tends to improve. However, more awakenings are associated with lower sleep efficiency. Gender and caffeine intake don't seem to affect sleep efficiency significantly. The dummy variables for sleep time categories show that, compared to short sleep (the reference category), moderate sleep doesn't significantly differ, but long sleep has a noticeable positive association with sleep efficiency. The model explains about 78.82% of the variability in sleep efficiency (R-squared = 0.7882)



```
> summary(model_with_categories)

Call:
lm(formula = Sleep_Efficiency ~ Age + Awakenings + Sleep.duration + 
    REM.sleep.percentage + Deep.sleep.percentage + Alcohol.consumption + 
    Caffeine.consumption + Gender + Sleep_Time_Category, data = Sleep_Efficiency)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.185995 -0.036850  0.006029  0.045475  0.152416 

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)                       2.885e-01  6.545e-02   4.408 1.52e-05 ***
Age                               6.867e-04  3.217e-04   2.135   0.0337 *  
Awakenings                       -3.483e-02  2.977e-03 -11.701  < 2e-16 ***
Sleep.duration                   -3.689e-03  8.761e-03  -0.421   0.6740    
REM.sleep.percentage              7.905e-03  1.093e-03   7.232 5.02e-12 ***
Deep.sleep.percentage             6.484e-03  2.997e-04  21.637  < 2e-16 ***
Alcohol.consumption              -3.522e-03  2.513e-03  -1.402   0.1622    
Caffeine.consumption              5.939e-05  1.253e-04   0.474   0.6360    
GenderMale                        2.343e-03  8.399e-03   0.279   0.7805    
Sleep_Time_CategoryModerate Sleep 4.167e-02  1.847e-02   2.256   0.0249 *  
Sleep_Time_CategoryLong Sleep     2.169e-02  3.171e-02   0.684   0.4946    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06181 on 267 degrees of freedom
Multiple R-squared:  0.7882,	Adjusted R-squared:  0.7802 
F-statistic: 99.35 on 10 and 267 DF,  p-value: < 2.2e-16
```

- [15 points] Include interaction terms in the above model so that you have different intercepts and slopes for different categories. Show fitted models for each category, plot of fitted model for each category. Interpret the coefficients in context of your goals.





```
Call:
lm(formula = Sleep_Efficiency ~ Age * Sleep_Time_Category + Awakenings +
    Sleep.duration + REM.sleep.percentage + Deep.sleep.percentage +
    Alcohol.consumption + Caffeine.consumption + Gender + Bedtime +
    Wakeup.time, data = Sleep_Efficiency)

Residuals:
      Min        1Q    Median        3Q       Max
-0.179467 -0.037920  0.005704  0.044832  0.148950

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                            0.1888535  0.0801850   2.355   0.0192 *
Age                                    0.0020801  0.0009629   2.160   0.0317 *
Sleep_Time_CategoryModerate Sleep      0.0829304  0.0453305   1.829   0.0685 .
Sleep_Time_CategoryLong Sleep          0.1581220  0.0609999   2.592   0.0101 *
Awakenings                            -0.0322954  0.0029934 -10.789  < 2e-16 ***
Sleep.duration                         0.0013890  0.0095931   0.145   0.8850
REM.sleep.percentage                   0.0081286  0.0010990   7.396 1.88e-12 ***
Deep.sleep.percentage                  0.0065885  0.0002957  22.282  < 2e-16 ***
Alcohol.consumption                   -0.0015345  0.0025415  -0.604   0.5465
Caffeine.consumption                   0.0001190  0.0001236   0.963   0.3366
GenderMale                             0.0049447  0.0082494   0.599   0.5494
Bedtime                                0.0007375  0.0006967   1.059   0.2907
Wakeup.time                           -0.0032293  0.0040701  -0.793   0.4282
Age:Sleep_Time_CategoryModerate Sleep -0.0010703  0.0010006  -1.070   0.2858
Age:Sleep_Time_CategoryLong Sleep     -0.0031605  0.0012714  -2.486   0.0135 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0604 on 263 degrees of freedom
Multiple R-squared:  0.8007,	Adjusted R-squared:  0.7901
F-statistic: 75.48 on 14 and 263 DF,  p-value: < 2.2e-16
```

The fitted models for each sleep time category is shown in the plot, showing that age has a different impact on sleep efficiency depending on the amount of sleep. The slopes of the lines in the plot reflect these interactions. Compared to the “Short Sleep”, those in the "Moderate Sleep" category show a slight increase in sleep efficiency, while those in the "Long Sleep" category show a more significant increase. However, as age increases, the effect on sleep efficiency decreases more for the "Long Sleep" category than for the "Moderate Sleep". More awakenings are associated with a significant decrease in sleep efficiency. Higher percentages of REM and deep sleep are strongly associated with increased sleep efficiency.

- [15 points] Include all numerical and categorical predictors (interaction term as well). Show fitted models for each category, plot of fitted model for each category. Interpret the coefficients in context of your goals.

The coefficients from this regression model tells the strength and direction of the relationship between each predictor and the sleep efficiency outcome.
A positive coefficient, like the one for Age, suggests that as age increases, sleep efficiency increases, very slightly (0.0021634 increase per year).
A negative coefficient, such as Awakenings, indicates that more awakenings are associated with lower sleep efficiency, with each additional awakening decreasing sleep efficiency by 0.0292231 units.

The regression analysis demonstrates that age has a slightly positive impact on sleep efficiency, while more awakenings are associated with decreased efficiency. The interaction effects reveal that age's influence on sleep efficiency reduces among those who experience longer sleep durations. The visual scatter plots confirm that this relationship is not consistent across different sleep time categories, with a notably stronger positive trend in shorter sleep durations. Statistically significant predictors like deep sleep percentage have a strong positive correlation with sleep efficiency. Overall, the model with an R-squared of 0.8118 fits the data well, indicating that the included variables explain a significant portion of sleep efficiency variance.

Conclusion [20 points]: Discuss the main findings from your analysis (in context of the problem) and how the analysis you showed addresses the research goals. Discuss what worked in the analysis and what would you like to improve. This is almost like writing the conclusion for your final paper. The conclusion is required to be at least half a page (the text starts at the top, left flushed, single-spaced, and doesn't need to be Times New Roman and 12 points).

In my research, I looked into what affects how well people sleep. I found out that being older usually means you sleep better and that the deep sleep part of your sleep is really important for feeling rested. However, if you wake up a lot at night, it's likely to mess with your sleep quality. One interesting thing I noticed is that for folks who get a lot of sleep, getting older doesn't make as much of a difference in sleep quality as it does for those who don't sleep as much.

Also, people's sleep experiences can vary a lot, especially if they don't get enough sleep.
The math behind my research says that I've got a pretty good handle on what's going on with sleep for most people, but there's some room for improvement.

Through rigorous regression modeling, I've discovered that certain predictors, like age and the percentages of REM and deep sleep, have a significant positive correlation with sleep efficiency. Conversely, an increase in awakenings is strongly linked to a decrease in sleep efficiency. These findings align with the common understanding that uninterrupted and deeper sleep stages are crucial for high-quality rest. The analysis revealed that the positive effect of age on sleep efficiency is less noticeable in individuals with longer sleep durations. This subtle interaction

suggests that simply increasing sleep duration is not a one-size-fits-all solution, especially as one ages.
Statistical significance tests reinforced these relationships, and the R-squared value of 0.8118 from the final model indicates a strong fit, accounting for over 80% of the variability in sleep efficiency. However, the presence of multicollinearity in variables such as Wakeup.time, which had a VIF just above 5, points towards an area for improvement in the model.

One of the most valuable insights from this analysis was understanding the role of dummy variables in capturing the effects of categorical predictors. They helped demonstrate the distinct patterns within different sleep time categories, providing a clear picture of how sleep duration interacts with other variables to affect sleep efficiency.

Although the models I constructed were strong, certain limits appeared. For instance, some LINE assumptions did not hold, indicating potential heteroscedasticity and nonlinearity, which led me to apply various transformations. Despite these efforts, some issues persisted, highlighting the complexity of modeling sleep efficiency.

As I conclude this analysis, I reflect on the importance of personalized approaches to improving sleep efficiency. While some factors like deep sleep stages are universally beneficial, the varying effects of age and awakenings on different individuals underscore the need for tailored sleep strategies. This research has provided dynamics of sleep and opened avenues for further investigation, particularly into the interaction effects and the role of less significant variables. I would like to enhance the model by addressing multicollinearity more effectively, possibly exploring alternative modeling techniques that can handle the complexity of sleep data better. In conclusion, everyone's sleep is a bit different, and there's no one-size-fits-all solution.