**Abstract**: This study looks at the causes of stress in engineering students, focusing on factors such as sleep quality, study load, headaches, academic performance, and extracurricular activities. Engineering programs are known for being very tough, with heavy workloads and not enough sleep being big causes of stress. Many students think feeling stressed is just part of their education, and since they don’t use the help available, the stress gets worse (Jensen, 2023).

The results show that poor sleep and heavy study loads are the biggest causes of stress. Other factors, such as frequent headaches and extracurricular activities, also contribute to stress, especially when combined. A linear model was used to understand how these things affect stress. Helping students sleep better and giving them less work could help lower stress and improve their well-being.

Exploratory data analysis showed some patterns, such as the difference in sleep quality and the fact that many students have heavy workloads. The results suggest that helping students manage their sleep, use their time better, and reduce their study load could help lower stress. These changes could also lead to better grades, better health, and a better quality of life.

**Introduction and Motivation:** Engineering students face high stress due to long study hours, heavy coursework, and tough exams. Many believe stress is just part of their education, making

it is harder to seek help (Jensen, 2023). The University of Michigan reports that engineering students face more challenges than others due to their heavy workloads and limited use of support services.

This study examines the causes of stress in engineering students, including sleep quality, study load, headaches, academic performance, and extracurricular activities. These factors influence both mental and physical health, with poor sleep and heavy workloads increasing stress. A linear regression model is used to understand how these factors work together and how lack of support affects students. The goal is to provide suggestions for reducing stress and improving mental health for better academic performance.

**Data Description**: The dataset includes responses from 520 engineering students, who rated different aspects of their academic and personal lives on a scale from 1 to 5. The variables are as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Description | Type | Scale/Range |
| Sleep Quality | Rate the quality of sleep, from 1 (poor) to 5 (excellent). | Discrete | 1–5 |
| Headaches | How many times a week do you suffer headaches, rated from 1 (rarely) to 5 (very frequent). | Continuous | 1-5 |
| Academic Performance | Rate academic performance, from 1 (poor) to 5 (excellent). | Discrete | 1-5 |
| Study Load | Rate study load from 1 (very light) to 5 (very heavy)? | Discrete | 1-5 |
| Extracurricular Activities | How many times a week do you practice extracurricular activities, rated from 1 to 5. | Continuous | 1-5 |
| Stress Levels | Rate stress levels from 1 (very low) to 5 (very high)? | Discrete | 1-5 |

Each variable represents a different aspect of student life. Variables like Sleep Quality, Academic Performance, and Study Load are treated as separate categories for the model, even though they are scored from 1 to 5. This helps understand how these factors are related to stress. Other variables like Headaches and Extracurricular Activities give more detailed information about student experiences. Together, these variables help us analyze and address stress in engineering students.

**Exploratory Data Analysis (EDA):** The exploratory data analysis focuses on understanding the distributions, correlations and patterns within the dataset to determine how each variable relates to stress levels.

**1. Useful Variables**: The variables analyzed in this study include sleep quality, headaches, academic performance, study load, and extracurricular activities. These variables represent different aspects of student life that may impact their stress levels. Based on initial observations:

• Sleep Quality and Study Load seem to be strong predictors of stress, meaning they likely have a big impact on how stressed students feel.

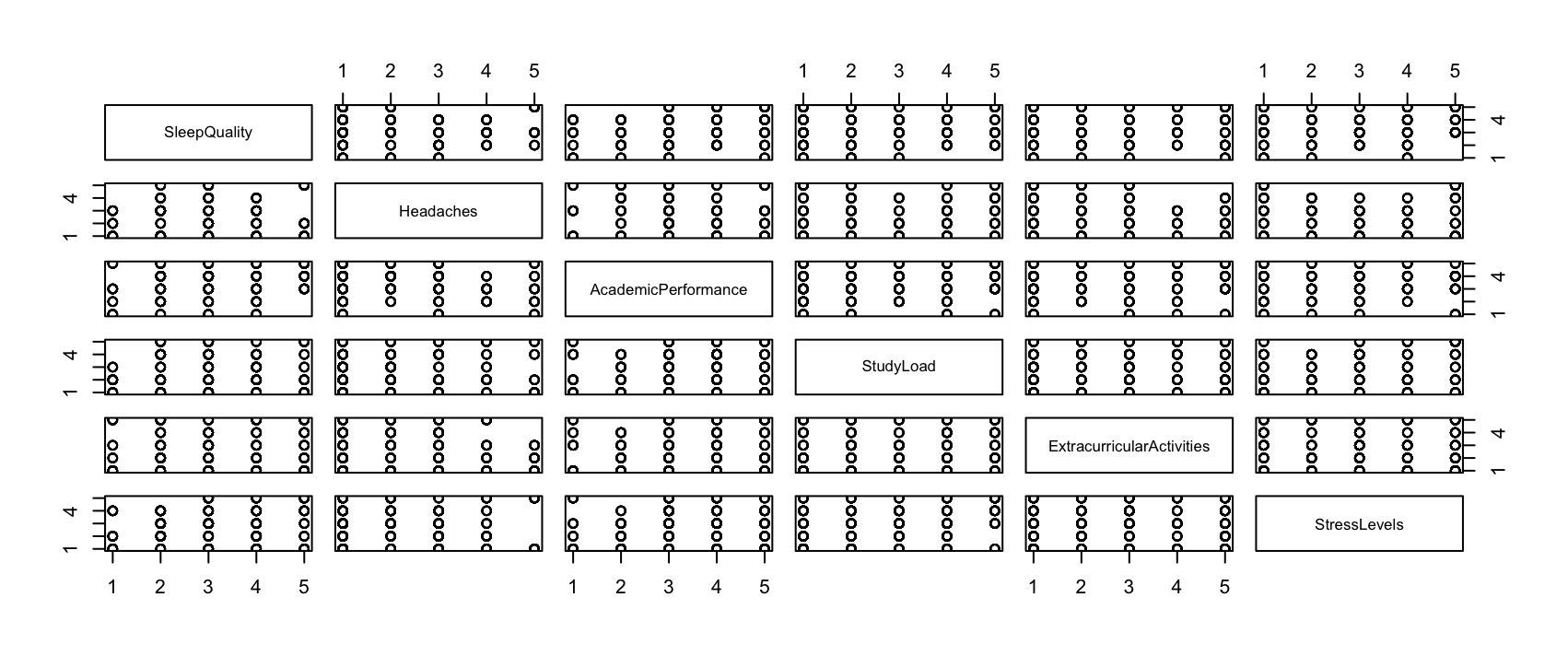
• Headaches and Extracurricular Activities also seem to play a role but may not be as strong.

R code:summary(stress\_data)

pairs(stress\_data)

**Summary of Variables**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Min | median | mean | Max |
| Sleep Quality | 1 | 3 | 3.13 | 5 |
| Headaches | 1 | 2 | 2.18 | 5 |
| Academic Performance | 1 | 3 | 3.33 | 5 |
| Study Load | 1 | 2.5 | 2.75 | 5 |
| Extracurricular Activities | 1 | 3 | 2.68 | 5 |
| Stress Levels | 1 | 3 | 2.88 | 5 |

****

**Interpretation:** Summary statistics show the range and variability of each variable, helping to understand the dataset.

• Most variables range from 1 to 5 with averages close to 3, indicating mid-level responses

for sleep quality, academic performance, study load, and stress levels.

• Headaches and extracurricular activities vary more but their averages are closer to the lower range (around 2).

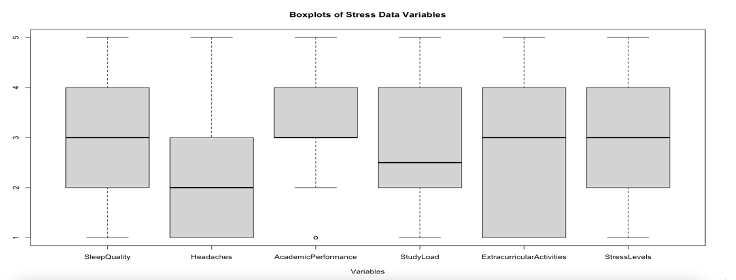
• Scatterplots indicate that stress levels increase as sleep quality decreases or study load increases, suggesting a positive correlation.

**2. High-order Terms or Transformations:** At this stage, no transformations or higher-order terms have been applied, as the data is on a 1–5 scale. However, the scale alone does not guarantee that the relationships between variables and stress levels are linear. If exploratory data analysis (EDA) reveals patterns where stress levels increase or decrease in non-linear ways, transformations (e.g., square, log) or higher-order terms may be used to better model these effects and improve the accuracy of the analysis.

**3. Outliers**: Boxplots were used to find outliers or data points that are very different from the others. These could be rare but valid cases, or they might be mistakes in the data.

R code:

boxplot(stress\_data)



**Findings:**

• Academic Performance: One outlier was found with a much lower score than the others. After checking, it was confirmed as valid data, representing a student with very low academic performance.

• Other Variables: Sleep Quality, Headaches, Study Load, Extracurricular Activities, and StressLevels showed no significant outliers. The data for these variables looks normal and consistent.

**Interpretation:** The outlier in Academic Performance was kept because it is real data and adds variety to the dataset. Since no major outliers were found in other variables, the data is considered reliable and good for analysis.

**4. Assumptions**: For this analysis, we assume:

1. Each variable is independent and does not influence the others.

2. The relationship between the variables and stress levels is linear.

3. The residuals (errors) of the model follow a normal distribution.

If these assumptions are violated, we may need to adjust the model by using transformations or adding interaction terms.

Scatter plots and residual plots were used to test the assumptions:

• Scatter plots check if the relationship between each variable, such as sleep quality and study load, and stress levels is linear.

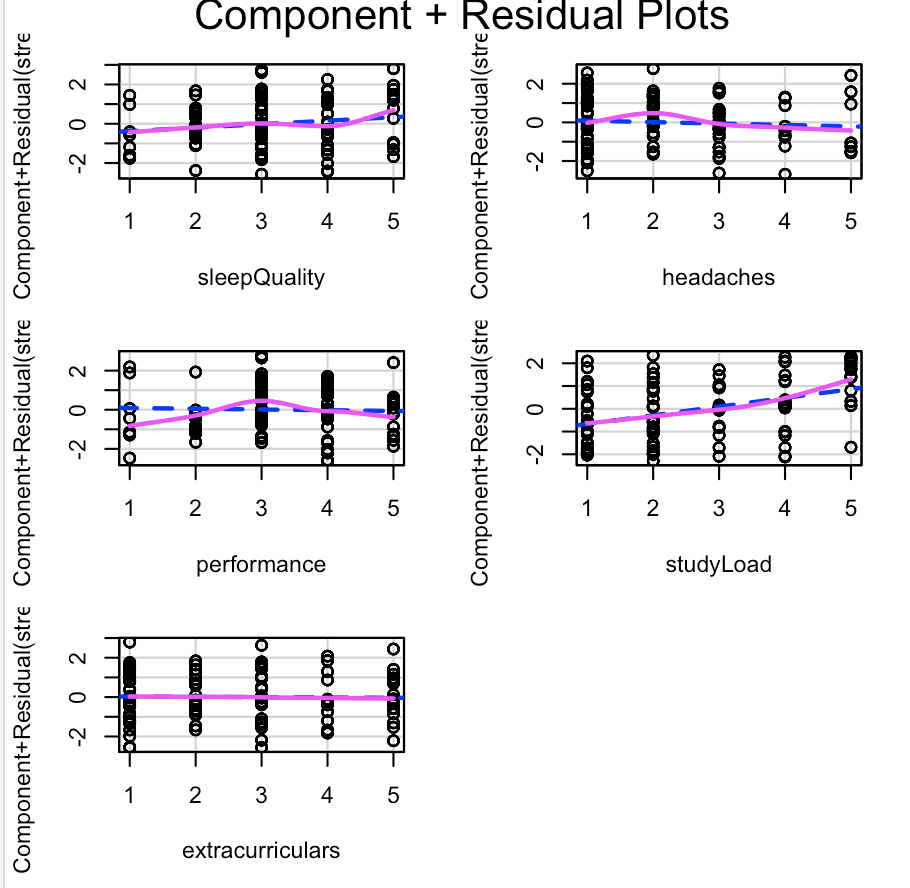
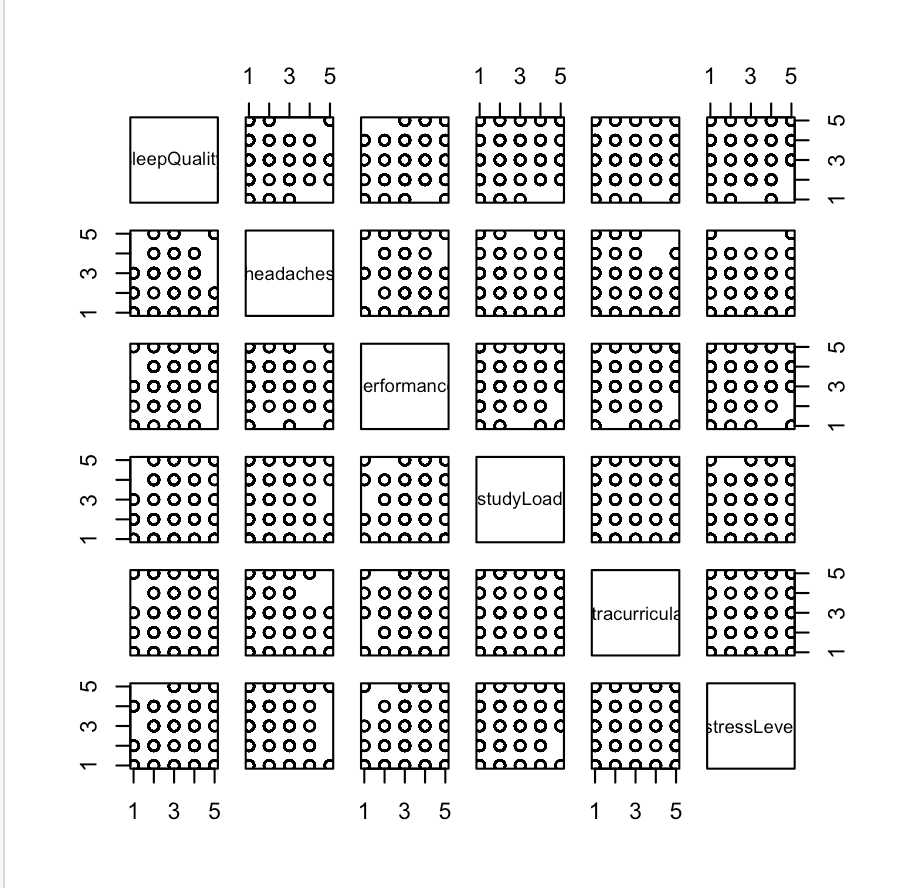
• Residual plots check if the model’s errors are normally distributed, which is necessary for the model to be valid.

R Code:

pairs(stress\_data)

full\_model <- lm(stressLevel ~ ., data = stress\_data)

crPlots(full\_model)



**Findings:**

• Linearity: Sleep quality, study load, and extracurricular activities show a straight-line relationship with stress levels. Headaches and academic performance show slight curves, indicating a possible non-linear relationship.

• Normality of Residuals: The residuals appear close to normal, so the assumption holds.

**Interpretation:**

Most variables meet the assumptions, but headaches and academic performance may need further checks for non-linearity. Overall, the assumptions are fine for using the linear regression model.

**5. Collinearity**: Collinearity happens when variables are highly related, making it hard to separate their effects. A correlation matrix was used to check for collinearity.

R Code:

# Calculate the correlation matrix

cor(stress\_data)

Correlation Matrix Output:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable | SleepQuality | Headaches | Academic  Performance | StudyLoad | Extracurricular Activities | Stress  Levels |
| SleepQuality | 1.00 | -0.06 | 0.25 | 0.07 | 0.00 | 0.17 |
| Headaches | -0.06 | 1.00 | -0.21 | -0.01 | -0.07 | -0.07 |
| Academic  Performance | 0.25 | - 0.21 | 1.00 | 0.10 | 0.15 | 0.06 |
| StudyLoad | 0.07 | -0.01 | 0.10 | 1.00 | 0.18 | 0.39 |
| ExtracurricularActivities | 0.00 | -0.07 | 0.15 | 0.18 | 1.00 | 0.05 |
| StressLevels | 0.17 | -0.07 | 0.06 | 0.39 | 0.05 | 1.00 |

**Findings:**

• Most variables show weak to moderate correlations, indicating that collinearity is not a significant concern.

• The highest correlation is between **Study Load** and **Stress Levels** (0.39), suggesting a moderate positive relationship that may influence stress levels.

• The weakest correlation is between **Sleep Quality** and **Extracurricular Activities** (0.00), confirming no direct relationship.

**Interpretation:** The correlation matrix shows minimal collinearity, meaning the regression model should work well. Study Load and Stress Levels have a moderate relationship but it’s not a problem. If needed, further checks like the Variance Inflation Factor (VIF) can help identify collinearity issues.

**6. Interaction Terms**

Interaction terms were added to see if combining variables affects stress more than individual factors. For example, combining **Study Load** with **Sleep Quality** or **Headaches** with **Extracurricular Activities** could reveal stronger effects on stress.

R code:

# Model with multiple interaction terms

model\_multiple\_interactions <- lm(StressLevels ~ StudyLoad \* SleepQuality + Headaches \* ExtracurricularActivities, data=stress\_data)

summary(model\_multiple\_interactions)

**Summary model**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Term** | **Estimate** | **Std. Error** | **t-value** | **p-value** | **Significance** |
| (Intercept) | 2.16624 | 0.41394 | 5.233 | 2.43e-07 | \*\*\* |
| StudyLoad | 0.29739 | 0.12159 | 2.446 | 0.014785 | \* |
| SleepQuality | 0.17607 | 0.10145 | 1.736 | 0.083235 | . |
| Headaches | -0.36292 | 0.09100 | -3.988 | 7.63e-05 | \*\*\* |
| ExtracurricularActivities | -0.27585 | 0.07939 | -3.475 | 0.000555 |  |
| StudyLoad:SleepQuality | 0.01743 | 0.03551 | 0.491 | 0.623811 | \*\*\* |
| Headaches:ExtracurricularActivities | 0.12608 | 0.03426 | 3.681 | 0.000257 | \*\*\* |

**Findings:**

• The interaction between Headaches and Extracurricular Activities is significant (p < 0.001), showing a strong combined effect on stress levels.

• The interaction between Study Load and Sleep Quality is not significant (p = 0.62), suggesting minimal combined influence.

• Individual variables like Headaches (p < 0.001), Extracurricular Activities (p < 0.001) and Study Load (p < 0.05) significantly affect stress levels.

• The model explains about 19.8% of the variation in stress levels (R² = 0.1983).

**Interpretation**: Adding interaction terms shows how variables can work together to impact stress. The results suggest that students with frequent headaches and a lot of extracurricular activities tend to have higher stress levels. However, the combination of study load and sleep quality doesn’t seem to have a big effect on stress.

**7. Training/Testing Split** : The dataset was divided into training and testing sets to check the model’s performance and avoid overfitting. Split Ratio: 70% for training, 30% for testing.

• Purpose: The training set was used to build the model, while the testing set checks how well the model predicts stress levels on new data.

• Method: The split was random to make sure both sets represent the data fairly.

• Evaluation: The testing set will be used to calculate metrics like Mean Squared Error (MSE) and R-squared (R²) to measure how well the model performs.

R code:

set.seed(123) # For reproducibility

train\_index <- sample(1:nrow(stress\_data), 0.7 \* nrow(stress\_data))

train\_data <- stress\_data[train\_index, ]

test\_data <- stress\_data[-train\_index, ]

**Next Steps :** After splitting the data, the model will be built using the training set and tested on the testing set to check how well it works. This helps make sure the results are reliable and not overly focused on just the training data.

**Model Diagnostics:** Model diagnostics were done to check if the model followed the main linear regression assumptions, such as linearity, normality of residuals, independence, and no multicollinearity. Outliers were also checked to make sure they didn’t influence the results too much.

Steps and Methods:

**1. Linearity Assumption:**

The residual vs. fitted plot was used to check if the relationship between the predictors (like sleep quality and study load) and stress levels is linear. The residuals should be randomly scattered around zero.

R code:

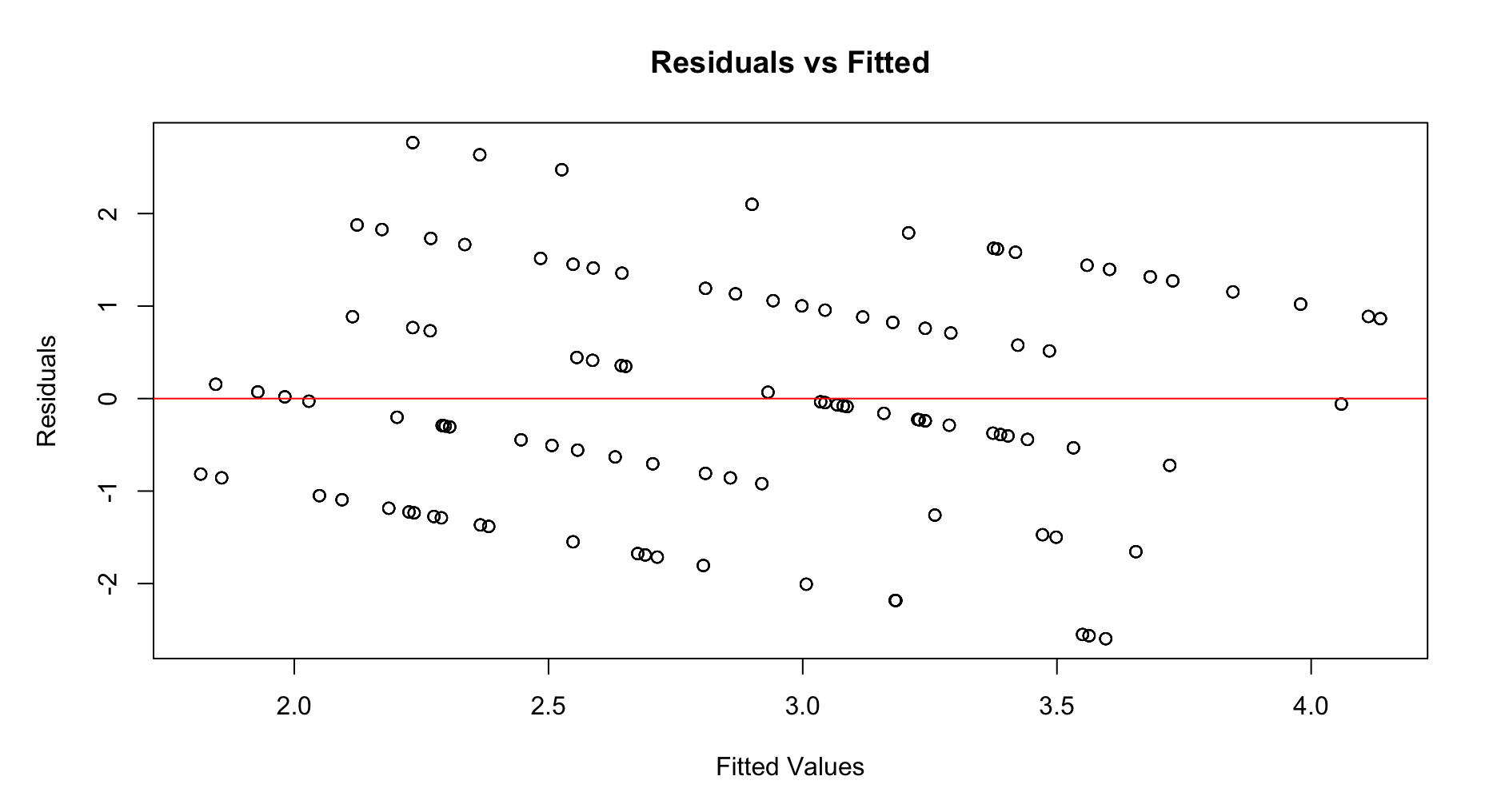
plot(full\_model$residuals ~ full\_model$fitted.values,

main = "Residuals vs Fitted",

xlab = "Fitted Values",

ylab = "Residuals")

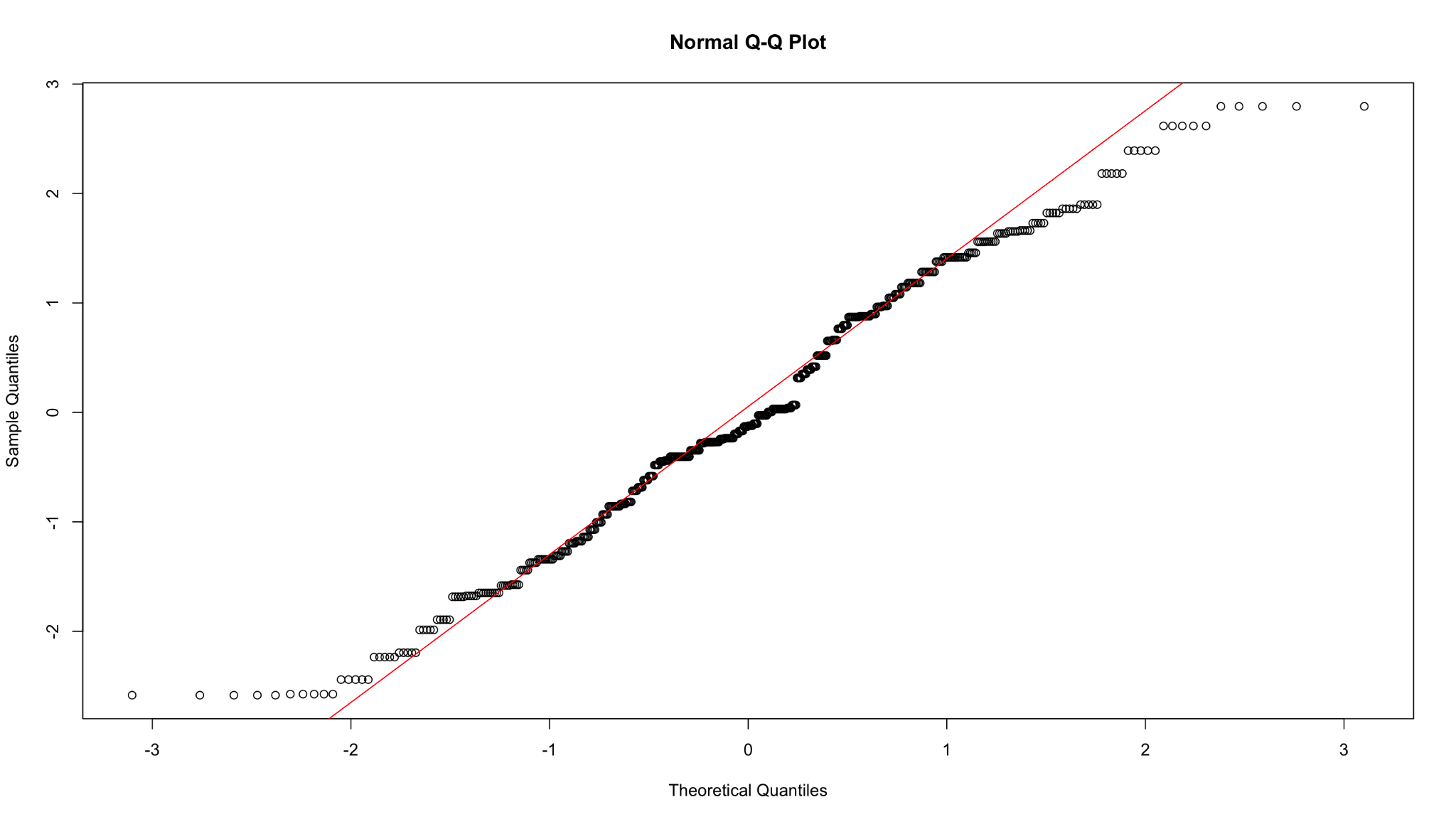
abline(h = 0, col = "red")



**2. Normality Check** : Q-Q plot was used to confirm that residuals followed a normal distribution.

R code: qqnorm(full\_model$residuals)

qqline(full\_model$residuals, col = "red")



• Q-Q Plot: Most points follow the normal line, with small deviations at the ends, showing slight non-normality.

**Shapiro-Wilk Test**:

R code: shapiro.test(full\_model$residuals)

The p-value from the Shapiro-Wilk test is 0.00002465. As the p-value < α = 0.05, we can reject the null hypothesis and conclude that normality is violated.

**Interpretation**: Since normality is violated, a transformation may be required to improve the model’s accuracy.

**3.** **Homoscedasticity :** The scale-location was used to check if the residuals are spread out evenly for all predicted values.

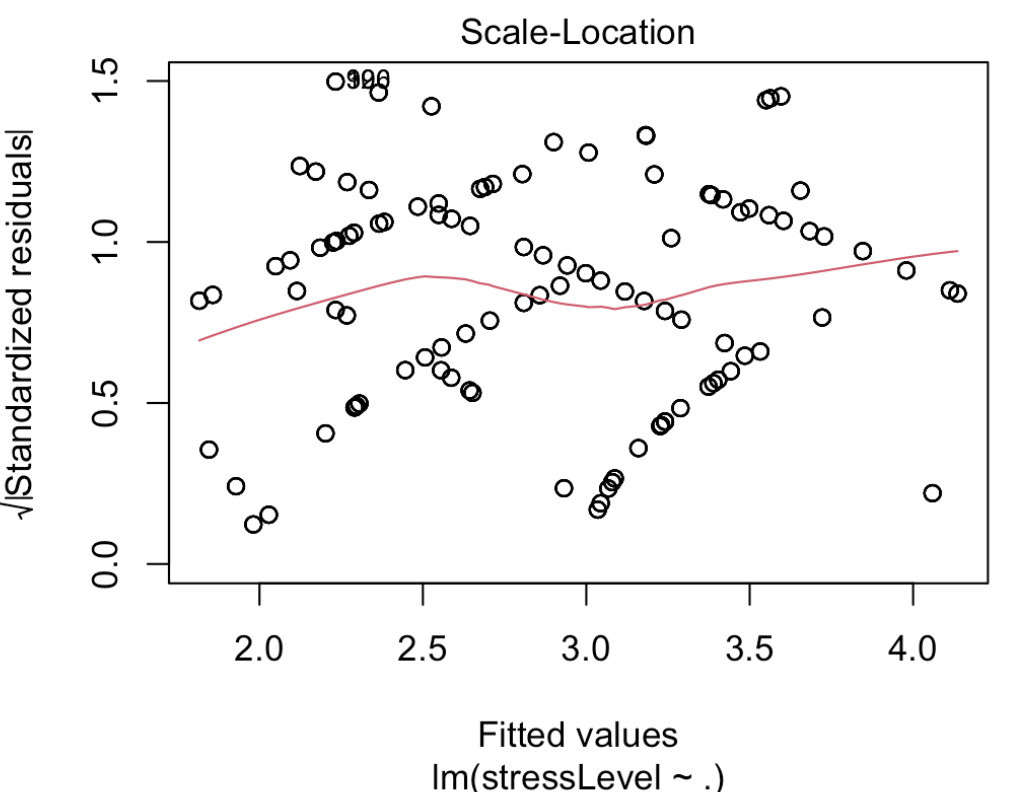
R code:

plot(full\_model$fitted.values, sqrt(abs(residuals(full\_model))), main = "Scale-Location",

xlab = "Fitted Values",

ylab = "Square Root of |Residuals|",

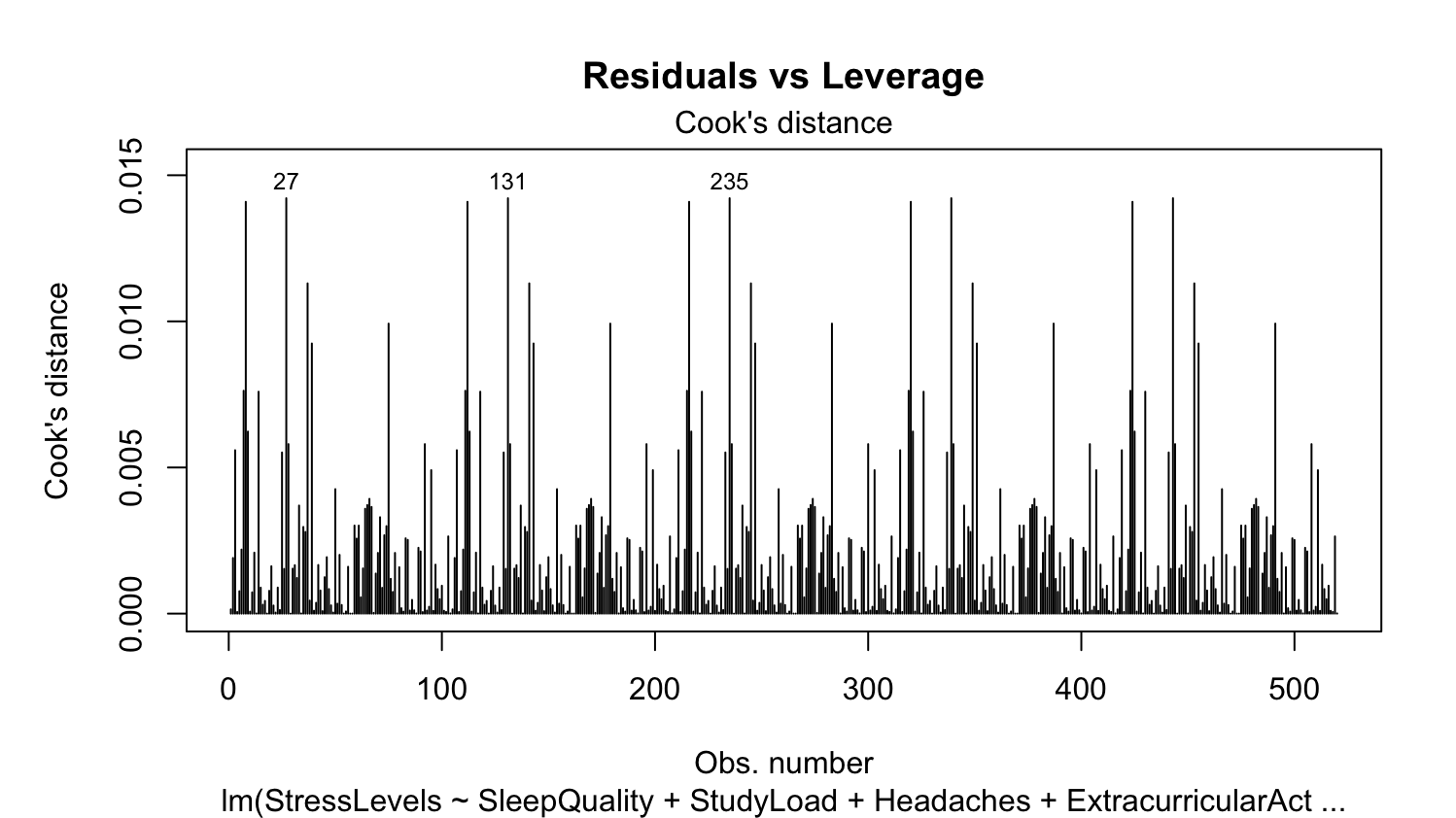
col = "black")

abline(h = 0, col = "red") # Add a horizontal line at 0

**interpretation** : if the points scattered randomly around the red line, it means the assumption is okay and the variance is consistent.

**4. Outlier Detection:** Outliers were checked using residual plots and Cook’s distance. Points with high leverage or large Cook’s distance may unduly influence the model.

R code : plot(full\_model, which = 4, main = "Residuals vs Leverage")



**Findings:**

**•** The plot showed that observations 27, 131, and 235 had higher Cook’s distances, meaning they might have more impact on the model.

• After checking their values, it was found that all three points had the same responses for variables like sleep quality, headaches and stress levels.

• A check confirmed these points are unique and not duplicates.

**Conclusion:**

These points were kept in the dataset because they represent students with high stress levels. Removing them would mean losing important information about these cases. Their influence on the results will be considered.

**Leverage** : Leverage measures how much influence a data point has on the model

The threshold for high leverage was calculated with 5 predictors and 520 observations

R code :

leverage\_values <- hatvalues(full\_model)

top\_leverage <- tail(sort(leverage\_values), n = 21)

print(top\_leverage)

**Findings:** The threshold value for high leverage was t = 0.0231 . Observations 109, 213, and 511 had leverage values higher than this threshold, meaning they had a strong influence on the model. However, these points were kept in the model because their influence wasn’t large enough to remove them.

**Jackknife Residuals**: Jackknife residuals help check if any data points are influencing the model too much. A big residual means that one data point has a large effect on the results.

The largest Jackknife residual for this model was 1.96

R code:

# Calculate Jackknife residuals

jackknife\_residuals <- studres(full\_model)

# Find the largest Jackknife residual

max(jackknife\_residuals)

**Interpretation**: The biggest residual is 1.96, which isn’t very big. This shows that no data points are messing with the model too much

**5. Multicollinearity Check:** Variance Inflation Factor (VIF) was calculated to detect multicollinearity.

R Code:> vif(full\_model)

The VIF values output :

• VIF for SleepQuality = 1.073457

• VIF for StudyLoad = 1.040015

• VIF for Headaches = 1.046513

• VIF for ExtracurricularActivities = 1.056736

• VIF for AcademicPerformance = 1.140660

**Findings:**

• The Variance Inflation Factor (VIF) was checked for the predictors in the model. All values were very close to 1, showing no serious multicollinearity.

• This means the predictors are independent and do not interfere with each other in explaining stress levels.

**Conclusion:** The low VIF values confirm that the model’s results are reliable and not affected by multicollinearity.

**6. Transformation :** To make the model better, different transformations were tried on the stress level variable, including log, square root, inverse, Box-Cox, and Y2/3 . The goal was to make the errors (residuals) of the model closer to normal.

R code:

# Log Transformation

train\_data$StressLevels\_log <- log(train\_data$StressLevels)

model\_log <- lm(StressLevels\_log ~ SleepQuality + StudyLoad + Headaches + ExtracurricularActivities + AcademicPerformance, data = train\_data)

# Square Root Transformation

train\_data$StressLevels\_sqrt <- sqrt(train\_data$StressLevels)

model\_sqrt <- lm(StressLevels\_sqrt ~ SleepQuality + StudyLoad + Headaches + ExtracurricularActivities + AcademicPerformance, data = train\_data)

# Box-Cox Transformation

library(MASS)

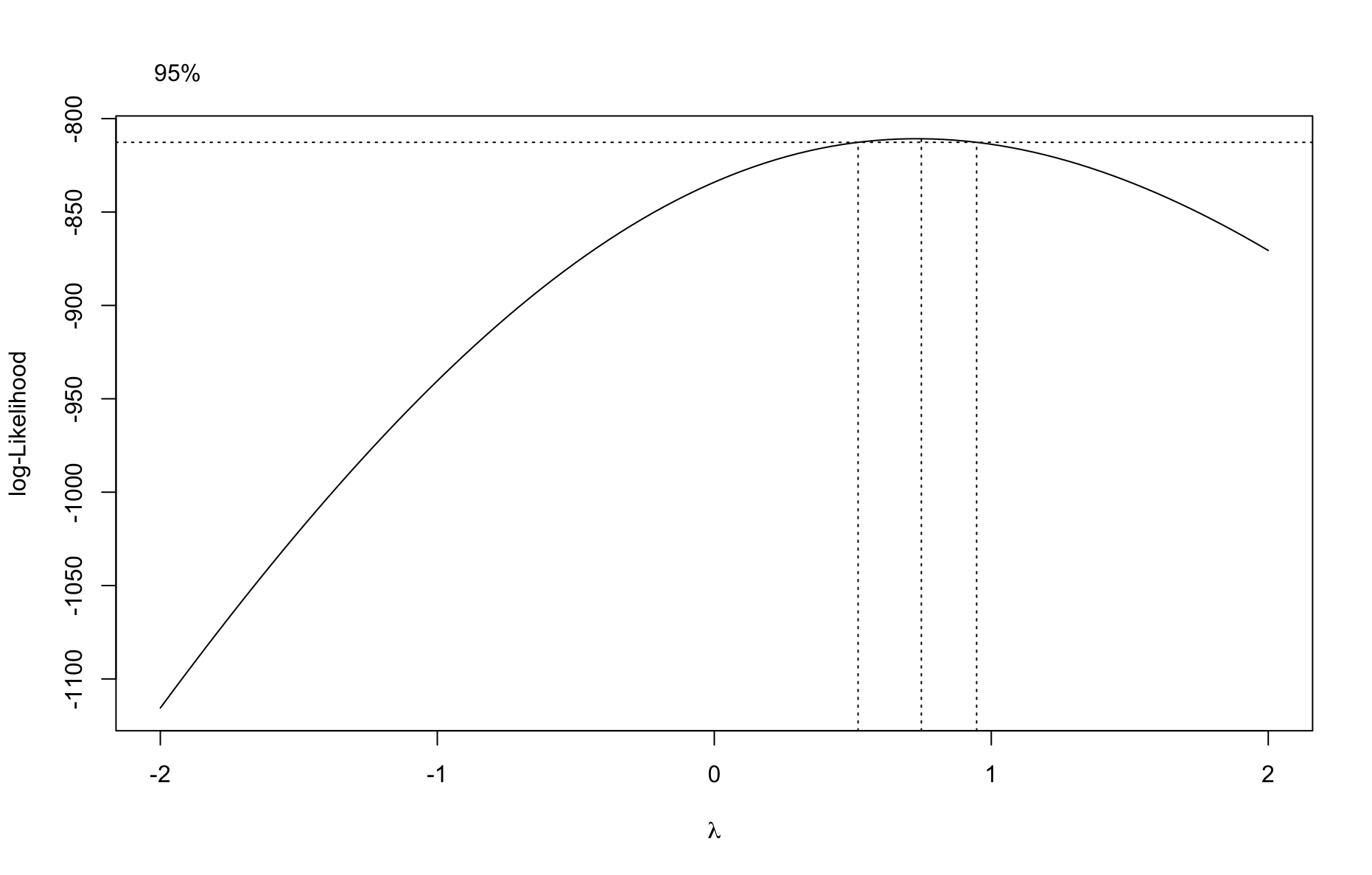
boxcox\_model <- lm(StressLevels ~ SleepQuality + StudyLoad + Headaches + ExtracurricularActivities + AcademicPerformance, data = train\_data)

boxcox(boxcox\_model, lambda = seq(-2, 2, by = 0.1))

# Y^(2/3) Transformation

train\_data$StressLevels\_23 <- (train\_data$StressLevels)^(2/3)

model\_23 <- lm(StressLevels\_23 ~ SleepQuality + StudyLoad + Headaches + ExtracurricularActivities + AcademicPerformance, data = train\_data)



**Findings:**

• Log Transformation: The Shapiro-Wilk test showed a very small p-value (9.79e-10), meaning the errors were still not normal.

• Square Root Transformation: This change also didn’t work well, with a p-value of 8.19e-06 from the Shapiro-Wilk test.

• Box-Cox Transformation: The Box-Cox change made the errors a little more normal (p-value = 9.07e-06), but it still wasn’t perfect.

• Y2/3 Transformation: The Y2/3 transformation gave a p-value of 5.26e-05, which was a small improvement.

**Conclusion**: Even though the Y2/3 transformation helped a little, none of the changes

made the errors perfectly normal. So, it was decided to continue with the original model

because the transformations didn’t make a big difference.

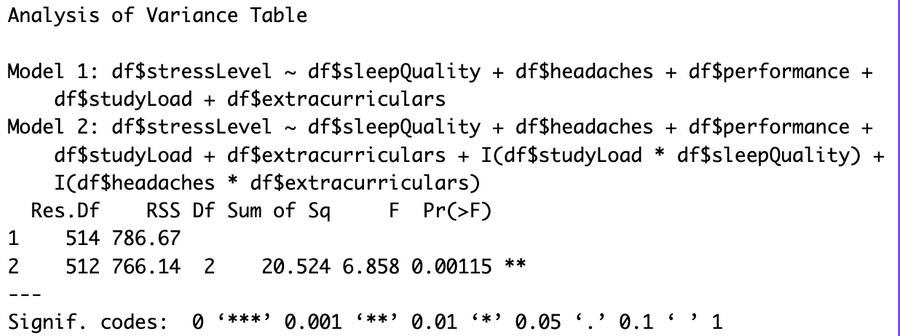
**Model Selection :**  Model Selection includes the process taken to choose the full equation that will be trained upon, which includes new interaction terms and transformations to independent variables. Using the results of the exploratory data analysis and model diagnostics to create the interaction terms and add it to the full model, while also considering dropping variables that may not contribute to the final model.

**ANOVA Table:** For the extended, full model, 2 extra interaction terms were created, and no independent variables are dropped. The 2 new terms added to the model are (StudyLoad \* sleepQuality) and (headaches \* extracurriculars), as they showed possible

interactions during data analysis. To ensure that the interaction variables have a significant effect on the prediction of stress levels, an ANOVA table is created to compare the simple model without interaction terms, to the extended table.

R code:

anova(full\_model, model\_2)



**Results:** At the 𝛼 = 0.05 significance level, there is sufficient evidence to indicate that at least one of the interaction terms is significant for predicting stress levels. As such, the extended model will be used for training and predicting student stress levels.

**Training** **:** During this phase of model selection, different selection models are used to create a regression model to predict our independent variable. Then, different models are compared to find the optimal model.

Before training, the data is split into training and testing sets. Using R, the original data is used and split 50/50 for training and testing.   
R Code:   
train=sample(1:nrow(full),nrow(full)\*0.50)

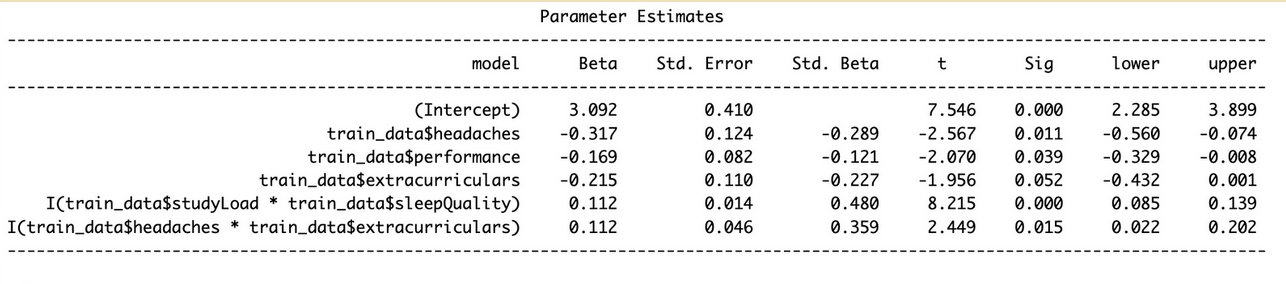
test=(-train)

train\_data <- full[train, ]

test\_data <- full[test, ]

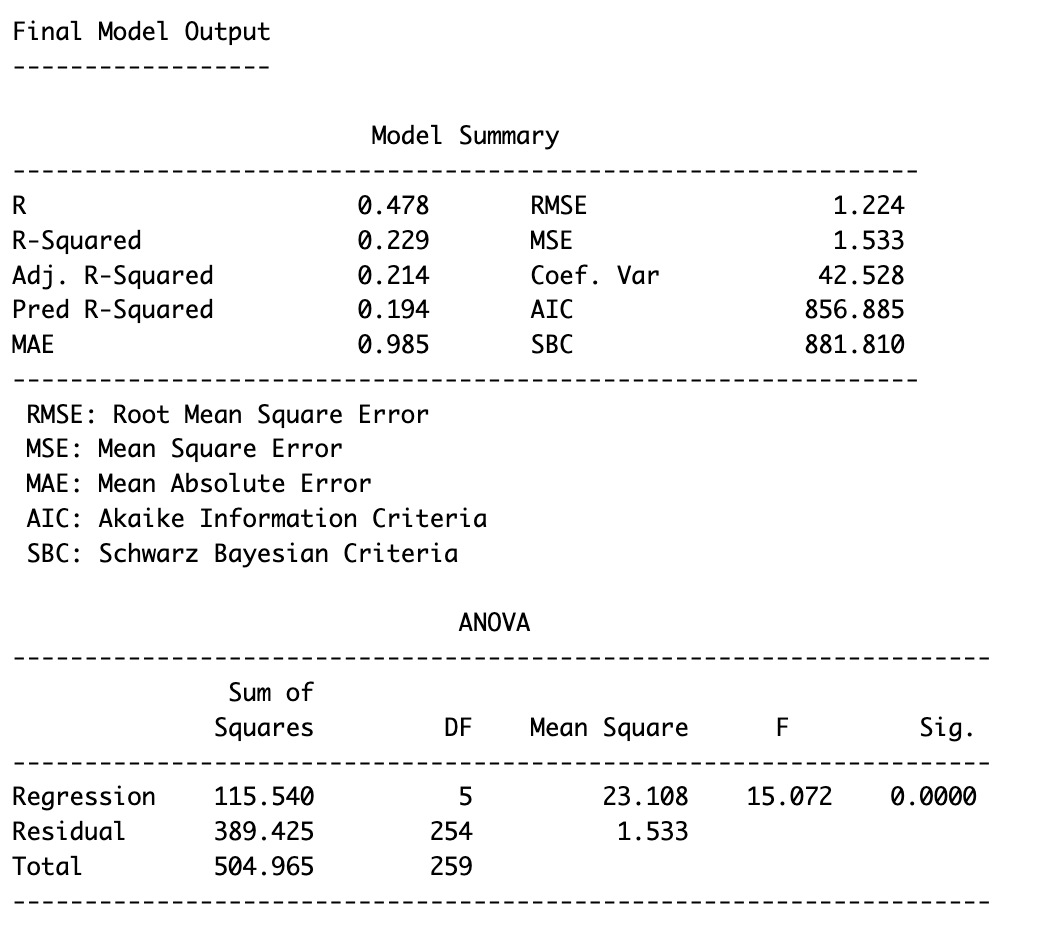
The initial model used for training was backwards-selection using p-values as criteria. This means that as the model is trained, insignificant variables are removed until an optimal value solution is found with the chosen method. However, other models tested include forward selection with p-values, piecewise selection with p-values, and backwards selection with AIC.

R code:

model\_backward <- ols\_step\_backward\_p(final\_model,prem = 0.1 ,details=TRUE)

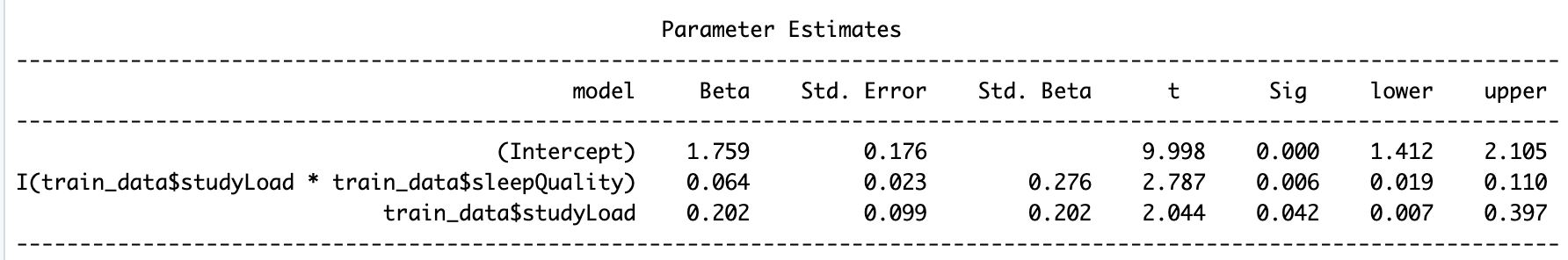
model\_backward <- ols\_step\_backward\_p(final\_model,prem = 0.1 ,details=TRUE)

**Results :** The final model decided on for backwards selection removed 2 variables: sleepQuality, and studyLoad. This means the complete formula was: 

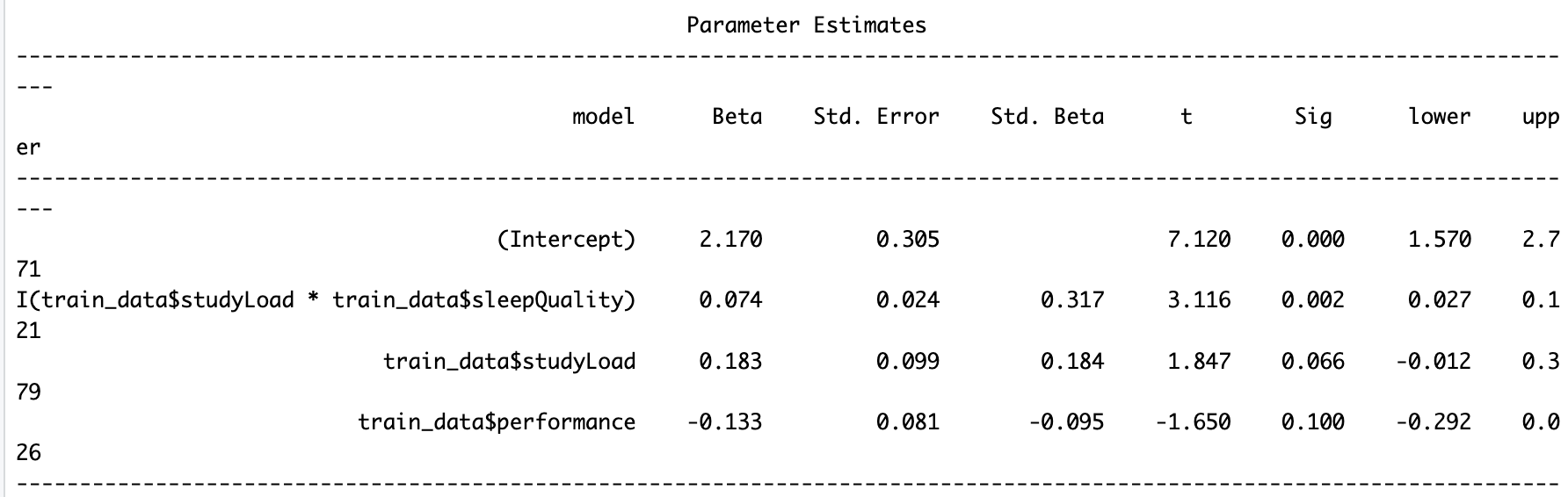


The other models were as follows:

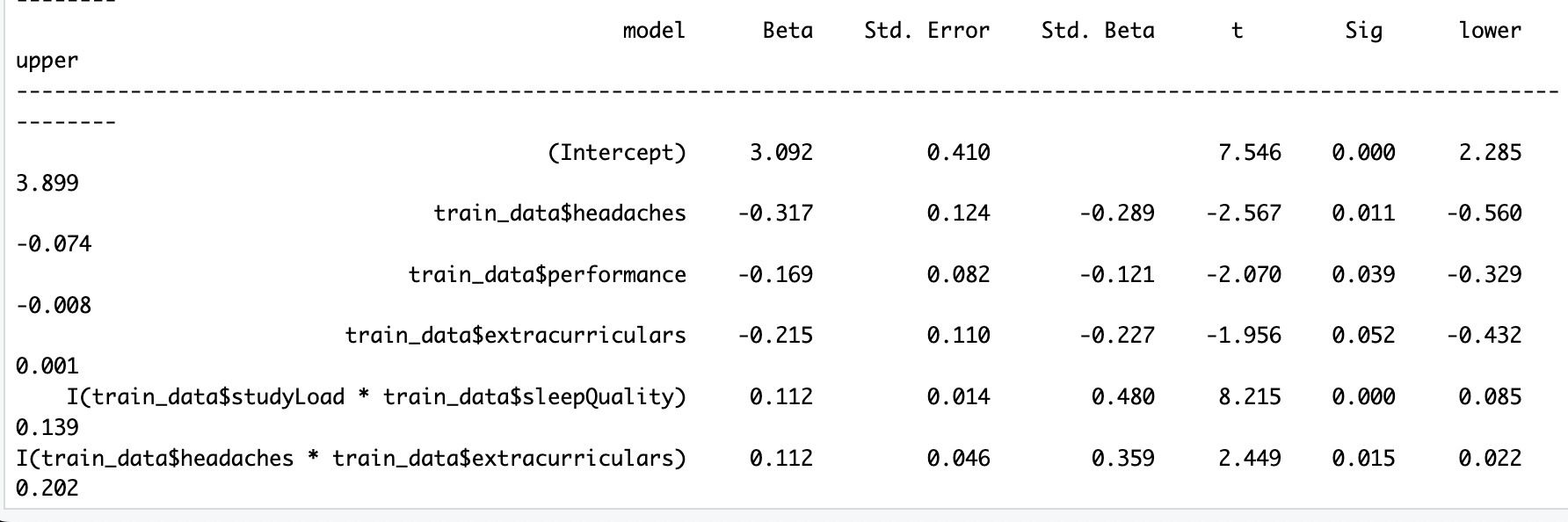
**Piecewise Selection with p-values:**

****

**Forward Selection with p-values:**

****

**Backwards Selection with AIC:**

****

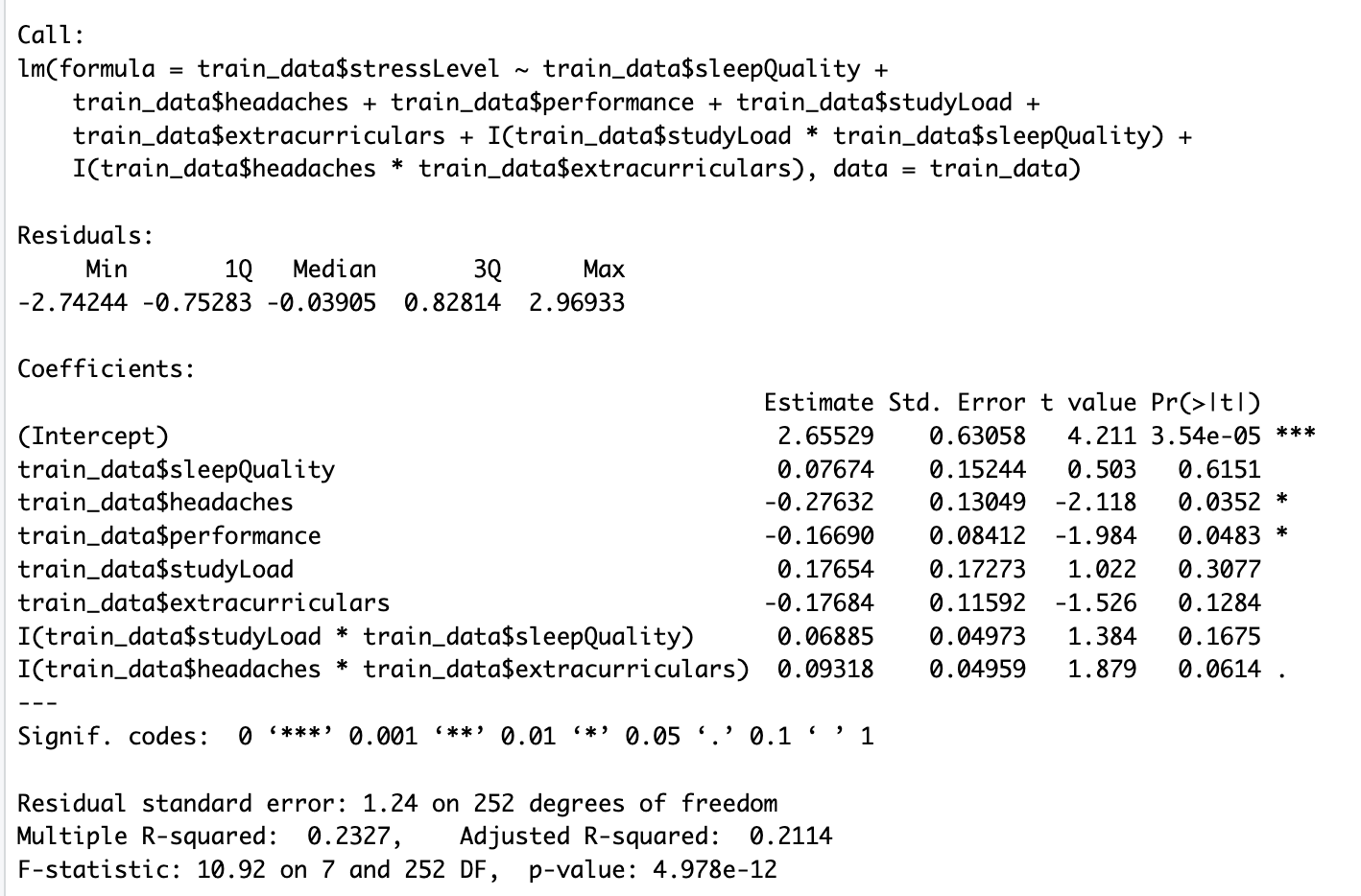
**Model Reliability :** To check model reliability for the backwards selection model, the model is recreated with a new dataset in R, now called df2. This dataframe is specifically used for checking model reliability. Then, the model is analyzed with the summary function. For this example, the backwards selection model was tested for reliability. For this example, backwards selection is used.

R code:

df2 <- train\_data[,-c(2)]

final\_model <- lm(df2$stressLevel ~ df2$headaches + df2$performance + df2$extracurriculars + I(df2$studyLoad \* df2$sleepQuality) + I(df2$headaches \* df2$extracurriculars, data = df2))

summary(final\_model)



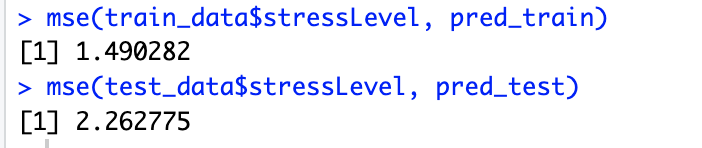
The mean square error of the test and training data is also calculated using R.

R code:

pred\_train <- predict(final\_model, train\_data)

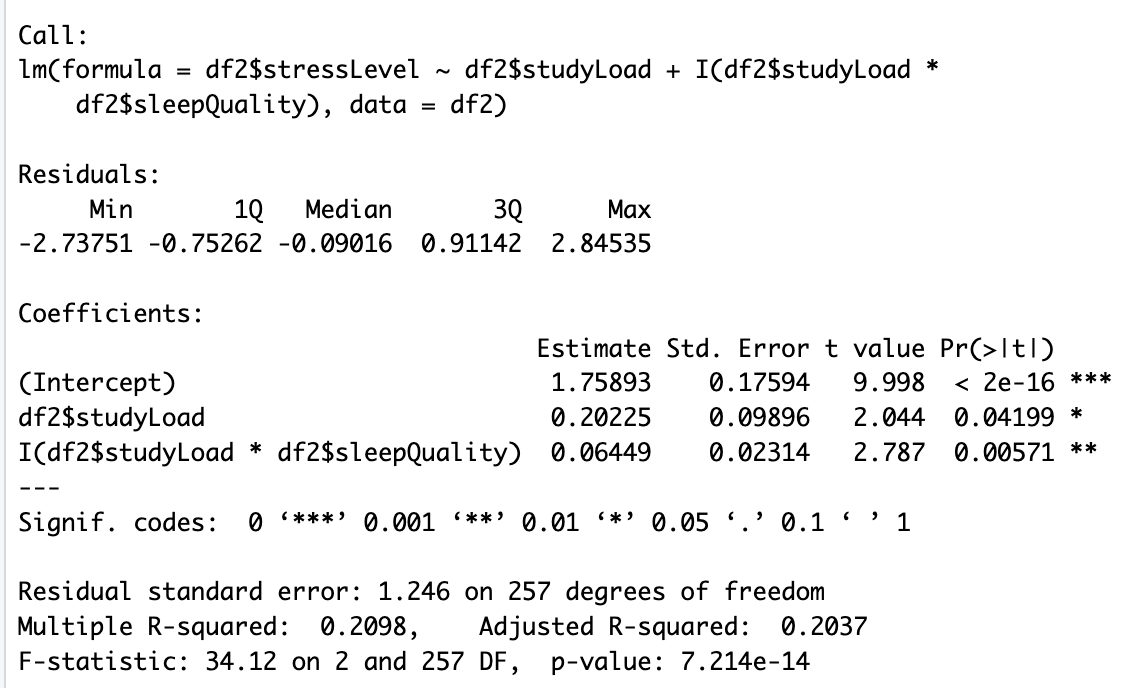
mse(train\_data$stressLevel, pred\_train)

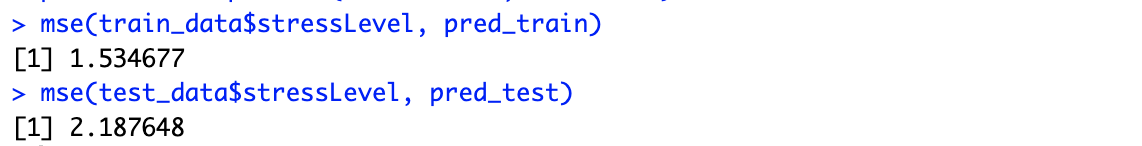
pred\_test <- predict(final\_model, test\_data)

mse(test\_data$stressLevel, pred\_test)

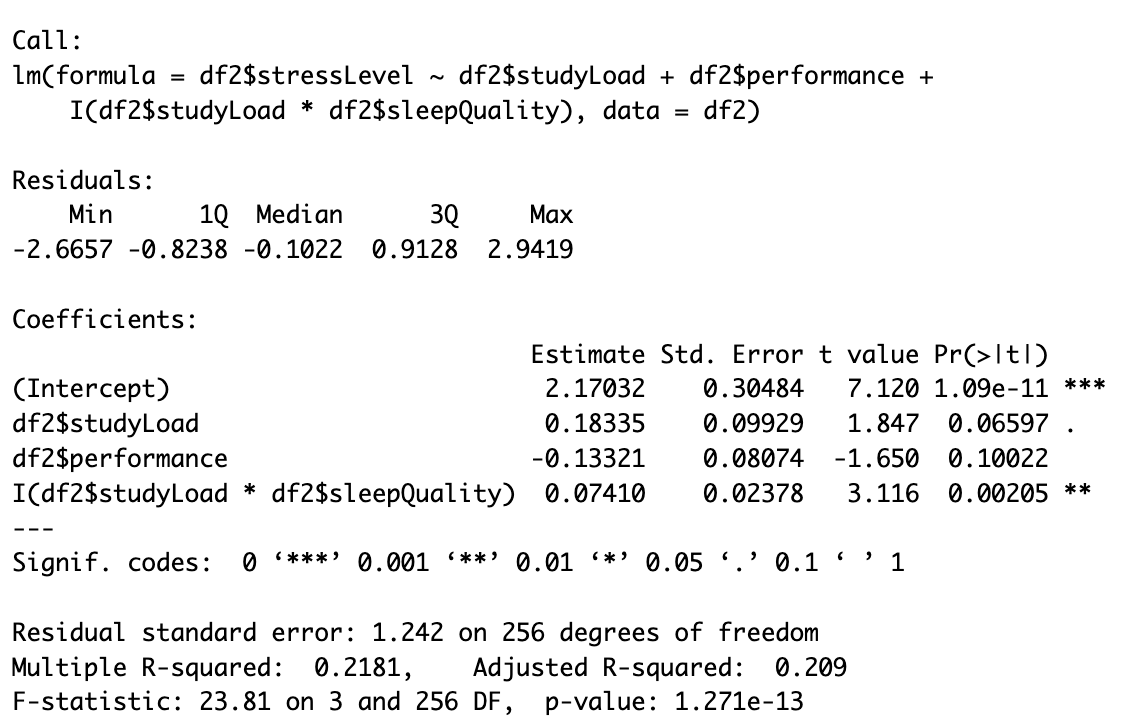
**Interpretation :** As the R2 **=** 0.2066 , 20.66% of the variability of stress levels in engineering students can be explained by the model, and the R2 adjusted = 0.191, approximately 19% of the variability of stress levels in engineering students can be explained by the model, when adjusted for the number of predictors in the model. It is also noticed that the training MSE (1.490282) is less than the testing MSE (2.262775), which shows overfitting is an issue with the backwards selection model. As these results are unsatisfactory, other methods must be used to find the optimal model. The results from the alternative models shown above are as follows:

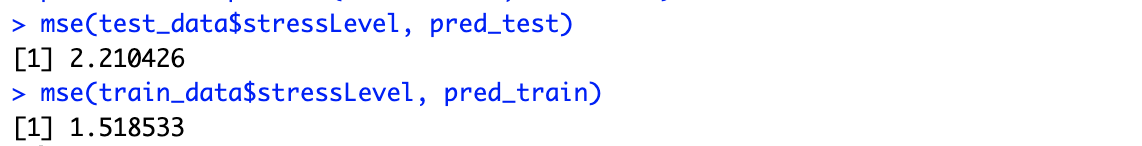
**Piecewise Selection with p-values:**

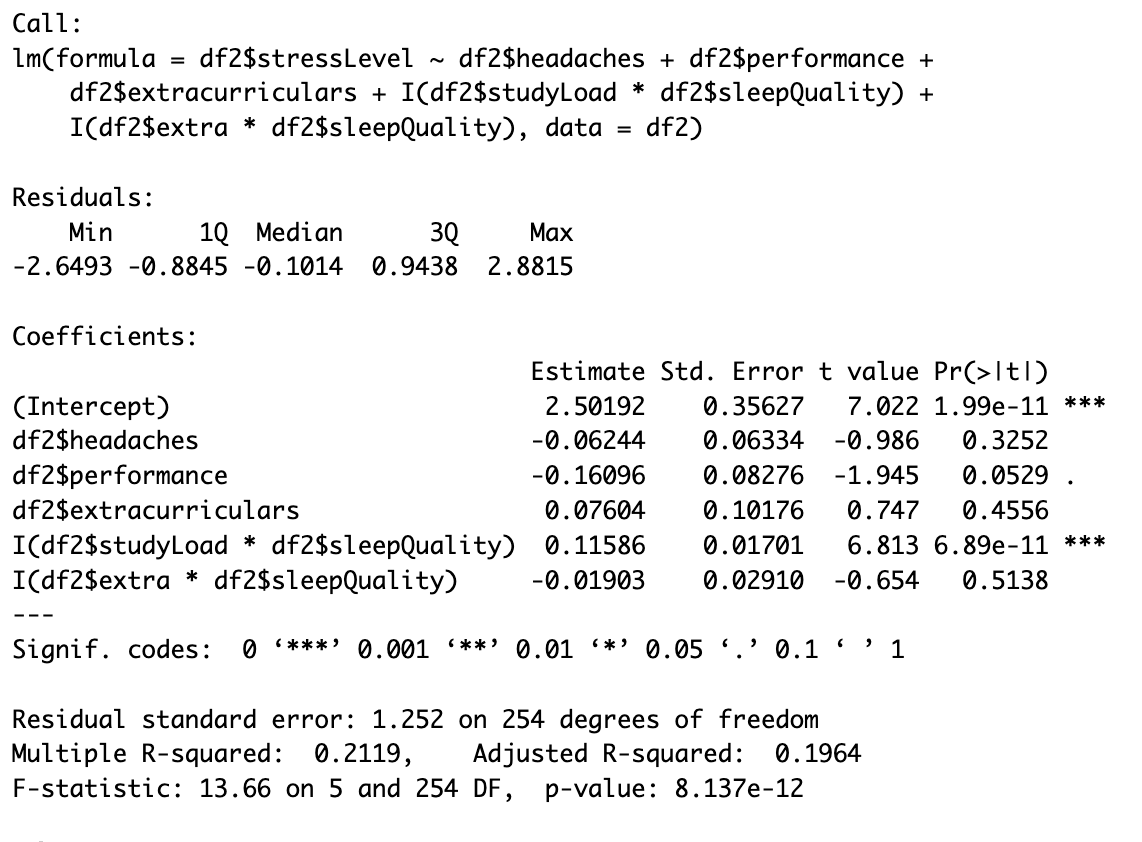
****

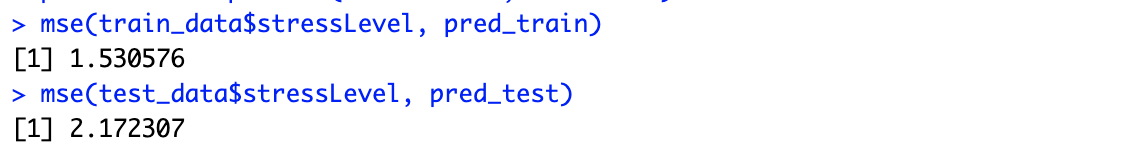
****

**Forward Selection with P-values:**

****

****

**Backwards Selection with AIC:   
**

****

**Results, Summary and Interpretations :** After testing other model types, and comparing the reliability tests, there were a complete total of 3 extra models alongside the preliminary backwards selection with p-values model. Though all models seemed prone to overfitting, the model with optimal results was Piecewise Selection with p-values.

The final results are R2 = 0.2096, Adjusted R2 = 0.2037, with a training MSE of 1.534677 and test MSE of 2.187648. This means that approximately 20.96% of the variability of stress levels in engineering students can be explained by the model, and approximately 20.37% of the variability of stress levels in engineering students can be explained by the model, when adjusted for the number of predictors in the model. The training and testing MSE are also the most similar compared to the rest of the models, concluding our final regression model is:

It can be assumed by this model that increases in study load, and its interaction with a student’s sleep quality, might correlate to heightened stress levels for engineering students.

**Conclusion and Limitations :** The study found that poor sleep and heavy study loads cause the most stress of engineering students. Other factors such as headaches and extracurricular activities,also add to stress.Improving sleep and managing workloads could help reduce stress.

The study only looked at a small group of students, so it might not represent everyone. The data was based on what the student said, which could be biased. Also, the study only considered a few factors, leaving out other possible causes of stress.

**Works Cited :** Jensen, Karin. “Engineering Students’ Perceptions of Stress and Mental Health.” BioMed Central, 2023, [https://blogs.biomedcentral.com/on-society/2023/05/04/engineering-students-perceptions-of-stress-and-mental-health/.](https://blogs.biomedcentral.com/on-society/2023/05/04/engineering-students-perceptions-of-stress-and-mental-health/)

University of Michigan. (2022, August 1). *The silent mental health crisis of engineering graduate students.* Retrieved from <https://ioe.engin.umich.edu/2022/08/01/the-silent-mental-health-crisis-of-engineering-graduate-students/>

International Journal of STEM Education. (2023). High stress culture in engineering students. Retrieved from [https://stemeducationjournal.springeropen.com](https://stemeducationjournal.springeropen.com/articles/10.1186/s40594-023-00419-6).

**Project Member Effort :**

**We, the project teams members, certify that below is an accurate account of the percentage**

**of effort contributed by each team member in the project and report.**

|  |  |
| --- | --- |
| Project Team Member | Percentage of Total Effort |
| Munchootsorn Wangsriviroj | 50% |
| Alexandra Ramlogan | 50% |