Data 602 Final Project

Statistically Examining the Topic of Sleep Efficiency

Presented By: Maria Delgado, Safa Hadi

February 14, 2024

Part I

Starting with the Barry Bonds case, we used linear regression modeling to answer the guiding question, which asks whether Barry Bonds was on steroids for the 2001 season. We started by visualizing how the season affects the number of home-runs for Barry Bonds (See Figure AA). There is a clear positive relationship between the two variables, as it's clear to see that the further the season gets, the higher the number of home-runs. The season coefficient (Beta 1) is equal to 0.0040442, which indicates a weak positive relationship between the season and the number of home-runs. Essentially, suggesting that the number of home-runs increases by 0.0040442, as the season increases by 1. First, we tested the assumptions of normality and independence for this linear model. Beginning with the assumption of normality, this condition is valid for this dataset, as the points were aligned to the Q-Q line (see Figure BB). Next, the assumption of independence is invalid, as there seems to be a pattern on the plot, suggesting that there is an association between the two variables (see Figure CC).

We conducted a t-test to further investigate if there is a positive association between the season and number of home-runs. The t-test resulted in a p-value of 0.0006222045, which is less than alpha of 0.05. Therefore, we can be 95% confident in saying that we have strong evidence against the null hypothesis that states that there is no relationship between the season and the home-runs for Barry Bonds. Meaning, we would be accepting the alternative hypothesis. Therefore, concluding that there is a weak positive correlation between the season and the number of home-runs that Barry Bonds scored.

Lastly, we predicted the number of home-runs for the season of 2001. It equalled 0.09988334. The resulting 95% prediction interval was (0.08353537 , 0.1162313). Therefore, we are 95% confident in saying that the predicted number of homeruns for Barry Bonds for the season of 2001 was somewhere between (0.08353537 , 0.1162313). The actual number of homeruns that was recorded on the dataset was equal to 0.1534. The difference between the actual and predicted is equal to 0.0371687. There is a significant difference in the actual versus the predicted, but it's hard to say whether Barry Bonds was on steroids or not. However, statistically speaking, we can conclude that there was a significant gap between his actual performance, and the predicted performance for the 2001 season.

Part II

Sleep efficiency is a very important topic, as it highlights the important variables that contribute to our overall sleep quality. We chose this dataset on the basis of trying to understand how certain variables, such as caffeine consumption and smoking status, would affect an individual’s quality of sleep. This dataset was developed as a result of trying to investigate how a person’s lifestyle would affect their overall sleep efficiency, it took into consideration their daily habits, such as smoking status, exercise frequency, alcohol consumption, and caffeine consumption. As well as, their gender and age. To explore how these habits would affect their sleep, the researchers further explored their sleep efficiency by measuring their sleep duration, number of awakenings that they incurred per night, and their deep sleep percentage. This would result in a holistic look into how their daily habits would reflect on their sleep quality. Examining the correlation between the different variables and what kind of an effect each would have on one another, would allow us to make an informed conclusion regarding the habits that actually affect our sleep quality.

The first question that we wanted to answer was, if there is a significant difference in the mean of the sleep efficiency of those who smoke versus those who don’t smoke. First we visualized the mean sleep efficiency for non-smokers (see figure A). As well as, the mean sleep efficiency for smokers (see figure B). Our hypothesis test is a two-tailed test, with the null-hypothesis being: the mean sleep efficiency for smokers is equal to the mean sleep efficiency for non-smokers. The alternative hypothesis would be: the mean sleep efficiency for smokers is not equal to the mean sleep efficiency of non-smokers.

(H_0): **μ_*smokers* = *μ*_nonsmokers**
(H_a): **μ_smokers ≠ *μ*_nonsmokers**

The mean difference in sleep efficiency for smokers and non-smokers, was -0.083. This indicates that there is, in fact, an inverse relationship between sleep efficiency and smoking. This is because the mean sleep efficiency for those who smoke is less than the mean sleep efficiency for those who do not smoke. We used an alpha of 0.05, 0.025 for the two-tailed test, to compare the p-value against. The resulting p-value for this hypothesis was equal to 3.603e-08. This p-value is significantly less than the chosen alpha of 0.025. Therefore, we have strong evidence against the null hypothesis that states that the mean sleep efficiency of smokers is equal to the mean efficiency of non-smokers. We, therefore, would reject the null hypothesis and accept the alternative hypothesis. We then built a 95% confidence interval to get a closer look at the interval for the mean difference. The resulting interval was (-0.11423816, -0.05303355). Therefore, we are 95% confident in saying that the difference in mean sleep efficiency for smokers versus non-smokers falls within -0.114 to -0.053. This would further reaffirm our conclusion from the observation of the p-value, leading us to the conclusion that there is a significant difference between the mean sleep efficiency for those who smoke versus those who do not smoke. Lastly, the t-value is equal to -5.705, which translates to the fact that, on average, non-smokers have higher sleep efficiency than smokers.

We examined the validity of our test, by looking at the normality of our variables. Starting with the mean for non-smokers, we visualize the data using a histogram (see Figure A). As portrayed by the histogram, the data of the mean for non-smokers, seems to not be following a normal distribution. The same result was also depicted in the histogram of the mean sleep efficiency for smokers (see Figure B). The histogram is not representative of normally distributed data. Lastly, for the normality test, we performed a Shapiro Wilk test for normality. It resulted in a p-value of 1.879e-10. In this case, the null hypothesis states that the data comes from a normally distributed population and the alternative states that the data does not come from a normally distributed population. Being that the p-value is equal to 1.879e-10, we would reject the null hypothesis and accept the alternative hypothesis that states that the data collected does not come from a normally distributed population. Despite the invalidity of normality, we continued to use the t-test, given that our sample size is quite large. The t-test is known to be robust to violations of normality, particularly with larger sample sizes.

Next, we looked at another hypothesis that entails whether the consumption of caffeine had an effect on an individual’s sleep efficiency. The null hypothesis states that there is no significant impact of caffeine consumption on sleep efficiency among individuals. The alternative hypothesis states that there is a significant impact of caffeine consumption on sleep efficiency. This was an important measure for us, as a great amount of individuals are avid coffee drinkers. We wanted to investigate and see if the consumption of coffee had an effect on sleep quality.

(H_0): **Beta_1=** ***0***

(H_a): **Beta_1>** ***0***

We built a linear regression model, where x is the caffeine consumption and y is the sleep efficiency. We chose x to be the caffeine consumption, as x is the independent variable that either affects or does not affect sleep efficiency, therefore, sleep efficiency being the dependent variable in this case. As per the linear regression plot, there seems to be evidence of an extremely weak relationship between the two variables (see Figure C). Meaning that, just by observing the plot, it can be assumed that there is a slight relationship between caffeine consumption and sleep efficiency. With regards to the linear regression model, it is important to look at the coefficients of the model, as they give us a clearer sense of the relationship at hand. Starting with the intercept coefficient, which equals approximately 0.7817. This suggests that the estimated sleep efficiency, when all independent variables, including caffeine consumption, are set to a value of zero, is equal to 0.7817. Next let's look at the coefficient for caffeine consumption, also known as the slope. The slope coefficient is equal to 0.0003314. This indicates that for each additional unit increase in caffeine consumption, the model predicts a negligible increase in sleep efficiency, approximately 0.0003314 units, holding all other factors constant. Therefore, the relationship can be seen to have a very weak positive relationship.

We wanted to build a hypothesis test, to further test whether there was a relationship between both variables. The null hypothesis indicates that the slope coefficient is equal to zero, meaning there is no relationship between caffeine consumption and sleep efficiency. The alternative hypothesis states that there is a positive relationship between caffeine consumption and sleep efficiency. We performed a t-test, and set an alpha equal to 0.05 to test the p-value against. The resulting p-value was 0.08974389. Given that the p-value is greater than the alpha of 0.05, we therefore, would fail to reject the fact that there is no relationship between the consumption of caffeine and sleep efficiency (null hypothesis). We built a confidence interval to test the range that our slope would fall into. We can be 95% confident in saying that the slope of this linear regression model would fall within (-0.0001343157, 0.0007159287). Again, the interval provides an extremely small range for the slope, which indicates a very weak relationship between caffeine consumption and sleep efficiency.

In conclusion to this linear model, while there is a very slight positive association between caffeine consumption and sleep efficiency according to the model, the effect size is extremely minimal and may not be practically significant. It would be fair to say that caffeine consumption is not an essential factor in determining the sleep efficiency of individuals. This might be contrary to what certain individuals might believe about the significance of caffeine consumption and sleep quality.

When using a linear regression model, it's important to assess whether this model is the best fit to calculate the relationship between these variables. To test the fitability of these variables, there are two specific tests, one that deals with normality of residuals and another that tests for homoscedasticity. As shown on the plots, it's evident to see that the residuals are not normally distributed (see Figure D). Next we tested for homoscedasticity, which essentially means that the spread of the data points around the regression line remains constant throughout the range of values of the independent variables. As seen on the scatter plot, this assumption is violated (see Figure E). Lastly, we conducted a Shapiro-Wilk normality test on the residuals. The resulting p-value was 2.231e-13. This suggests strong evidence against the null hypothesis that the residuals are normally distributed, as it is smaller than a significance level of 0.05. After all the conducted tests, we can conclude that the residuals deviate significantly from a normal distribution, suggesting that the linear regression model may not be suitable for these variables.

This led us to run a GLS model, which offers a more flexible modeling framework to address violations of homoscedasticity and normality of residuals that are often encountered in linear regression. GLS adjusts for varying levels of variance across observations by modeling the variance-covariance structure of the errors, thus yielding more accurate parameter estimates. Unlike linear regression, GLS does not assume that the residuals follow a normal distribution. In this GLS model, the estimated intercept is approximately 0.7818, and the estimated coefficient for Caffeine consumption is approximately 0.0003274. However, similar to the linear regression model, it appears that in the context of this data and model, caffeine does not have a significant impact on sleep efficiency.

Lastly, we wanted to see if there is a significant relationship between exercise frequency and sleep efficiency. If a person increases their exercise frequency, are they likely to have better sleep? To answer this question, we built a regression model, with x being the exercise frequency and y being the sleep efficiency. This linear regression model will help us in determining if there is a relationship between the two variables, and whether such relationship is inverse or direct (See Figure F).

(H_0): **Beta_1= *0***
(H_a): **Beta_1> *0***

If the exercise frequency coefficient (Beta 1) is equal to zero, this would be indicative of no relationship. If Beta 1 is greater than zero, this yields a positive relationship. To test this, we used the bootstrapping technique to generate our own sample with 1000 iterations. This resulted in a confidence interval of (0.01578825, 0.03402403). The results indicate that the relationship between exercise frequency and sleep efficiency has a positive slope. This means that sleep efficiency, on average, will fall within the interval of 0.016 to 0.03, when exercise frequency is equal to 1. Essentially, indicating that an increase in exercise frequency will, in turn, slightly increase sleep efficiency.

When working with the sleep efficiency dataset, there were many variables of interest that we wanted to further investigate and explore. Statistical analysis, such as hypothesis testing and linear regression modeling allowed us to dig deeper into the relationship between the different variables. As well as, gain a greater perspective into the habits that have an actual effect on sleep efficiency. The first question that we wanted to answer was whether the smoking status of an individual would affect their sleep efficiency. This question might sound unusual, as normally we don’t think to associate smoking status with sleep efficiency. After conducting the hypothesis testing, we came to the conclusion that smoking status does, in fact, influence sleep efficiency. Non-smokers are more likely, on average, to have higher sleep efficiency than those who smoke. Next, we wanted to analyze whether there was a relationship between caffeine consumption and sleep efficiency. Again, intuitively, one might think that caffeine consumption does have a significant impact on sleep quality. After conducting a linear regression test and a GLS test, it turns out that there is no significant association between caffeine consumption and sleep quality. However, as it is widely recognized that caffeine disrupts sleep patterns, further exploration may be warranted. Particularly by considering higher levels of caffeine intake and examining the timing of caffeine consumption relative to sleep.Lastly, we examined exercise frequency, and its effect on sleep efficiency. We conducted a bootstrapping technique to solve this question. The results yielded a weak positive relationship between exercise frequency and sleep quality. This project was very enlightening for us, as it allowed us to dive deeper, and put aside our preconceived notions regarding sleep efficiency, to focus on examining the daily habits that actually influence sleep efficiency.

Figure AA

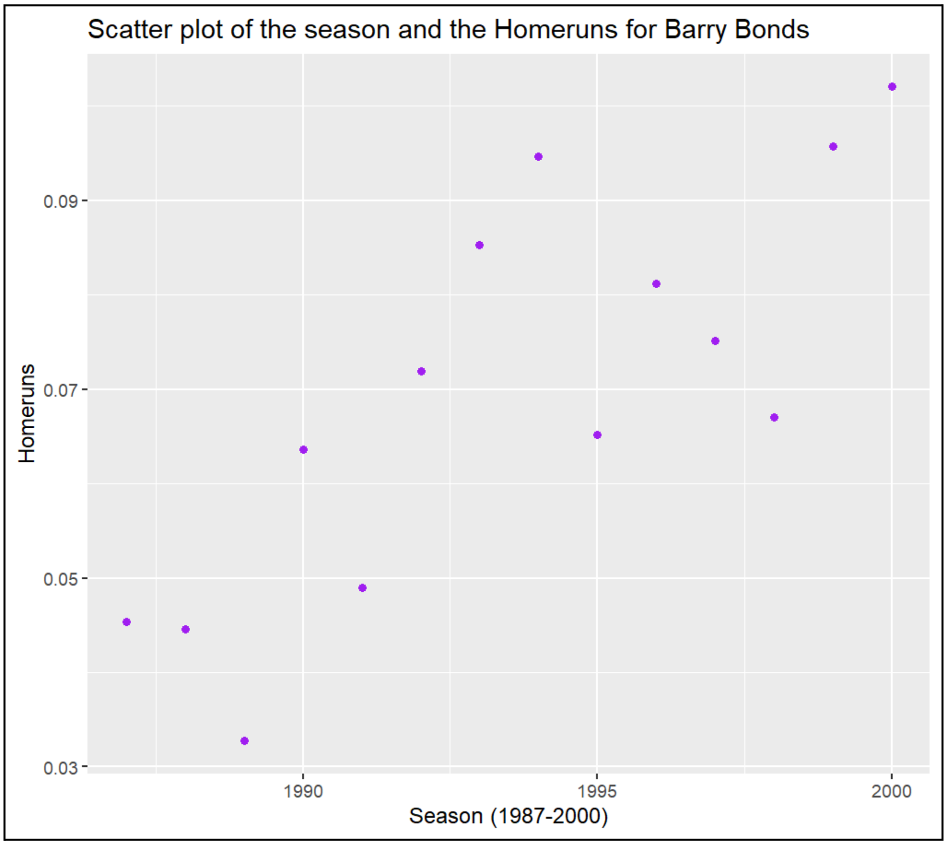


Figure BB

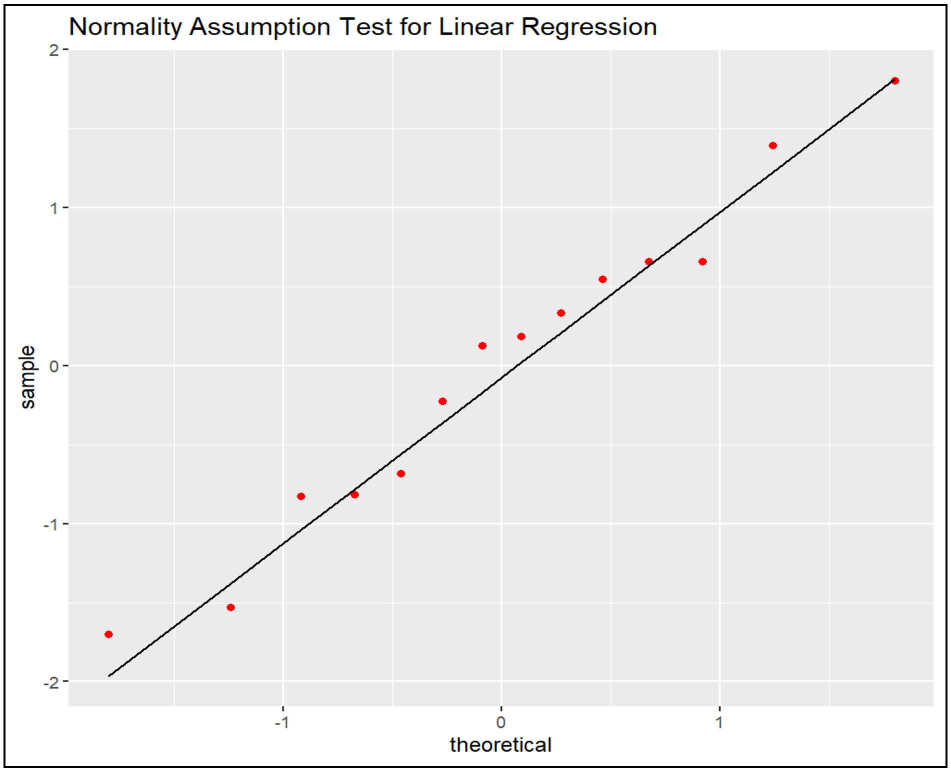


Figure CC

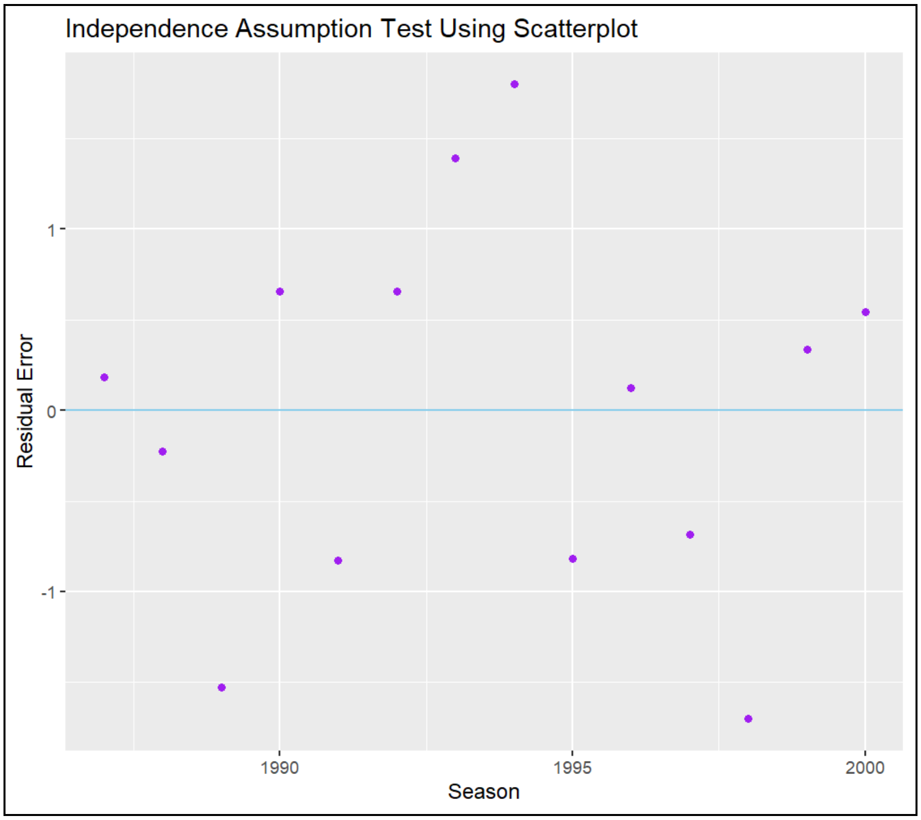


Figure A

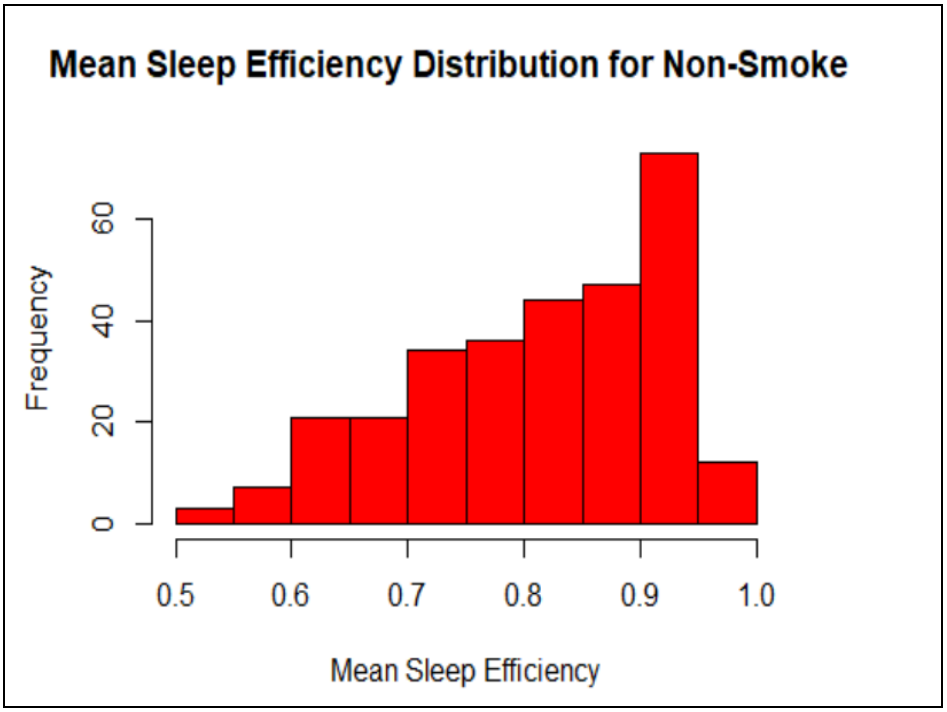


Figure B

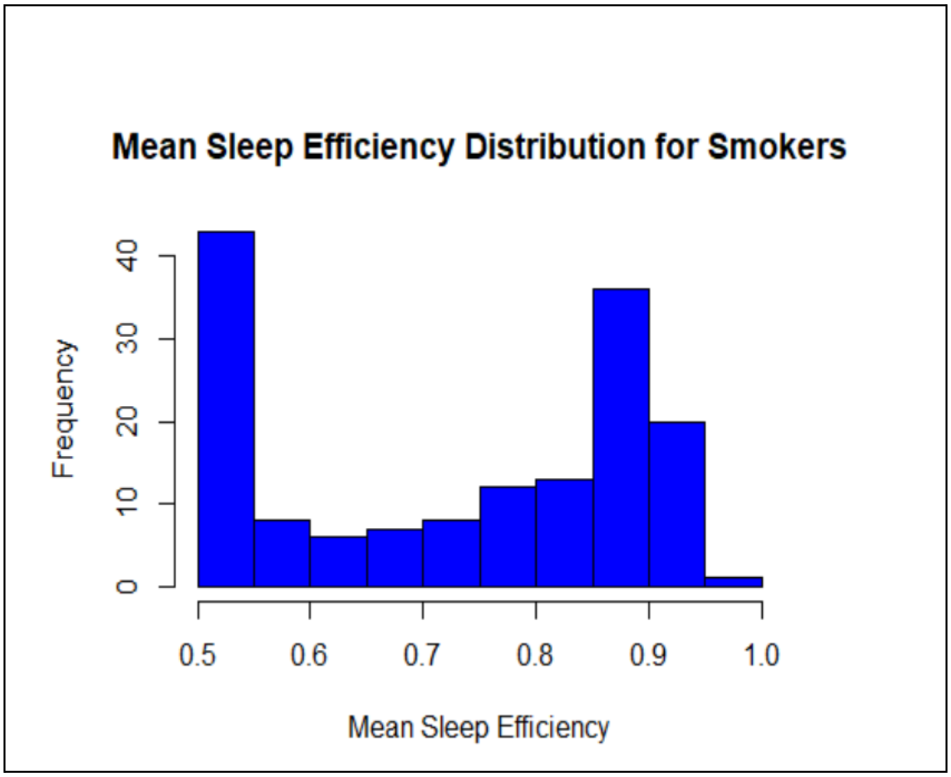


Figure C

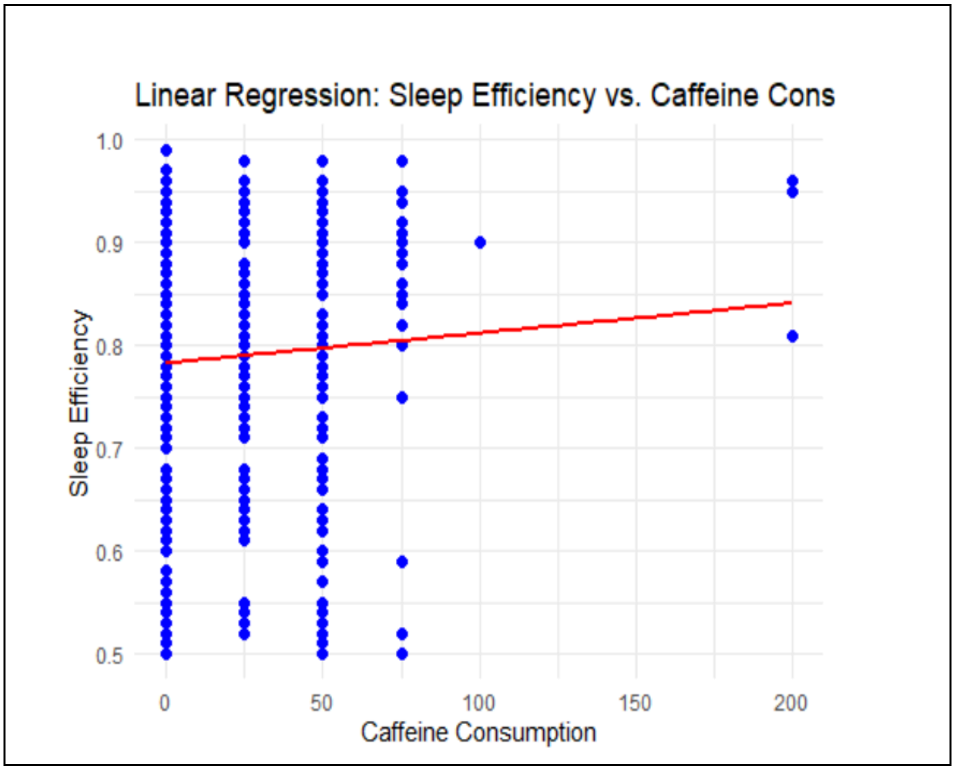


Figure D

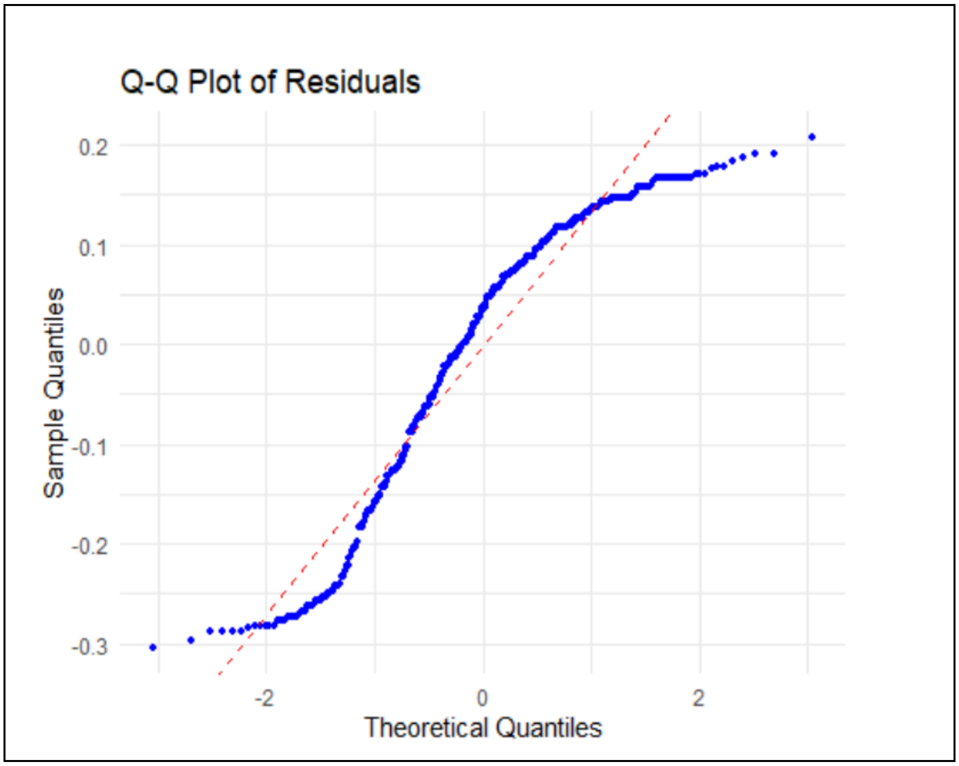


Figure E

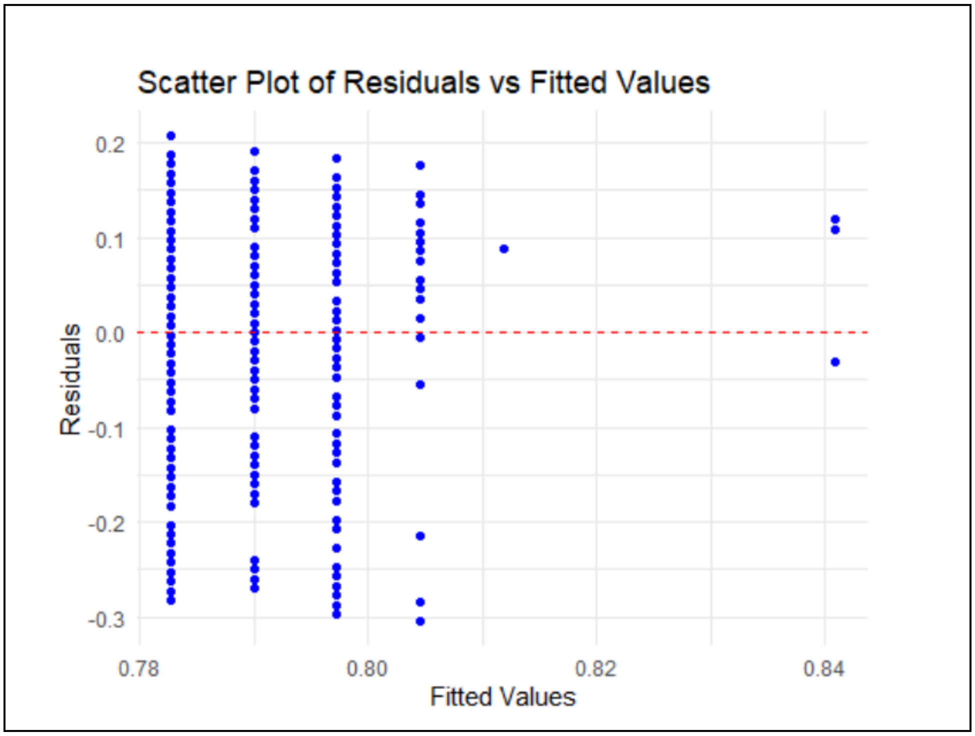


Figure G

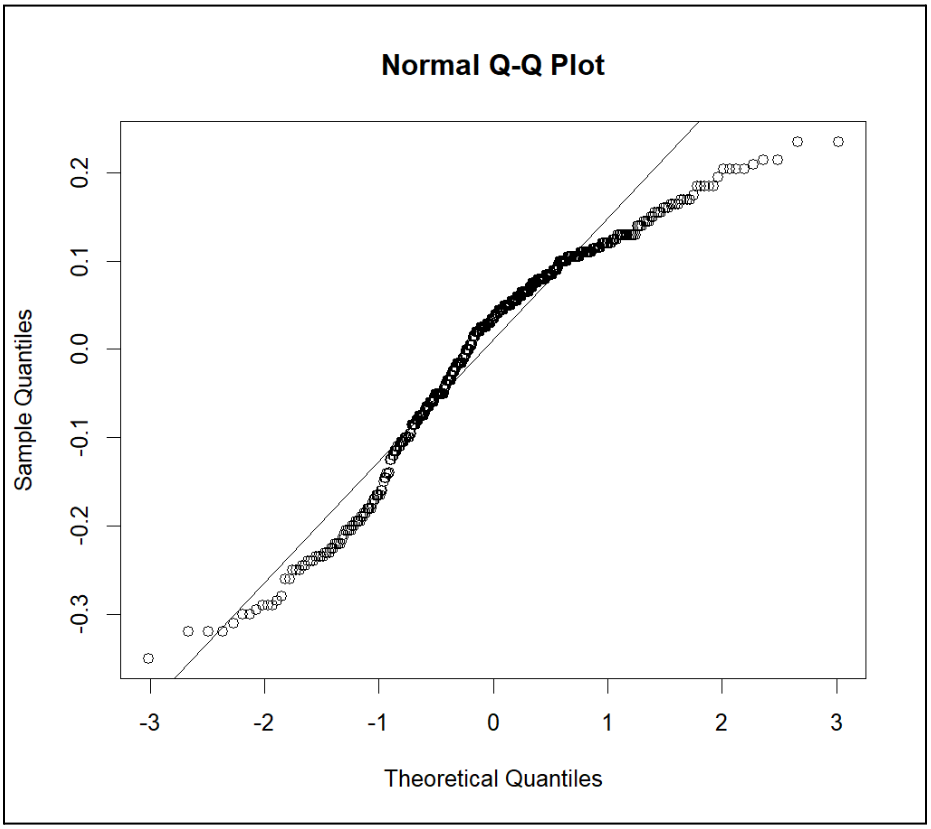
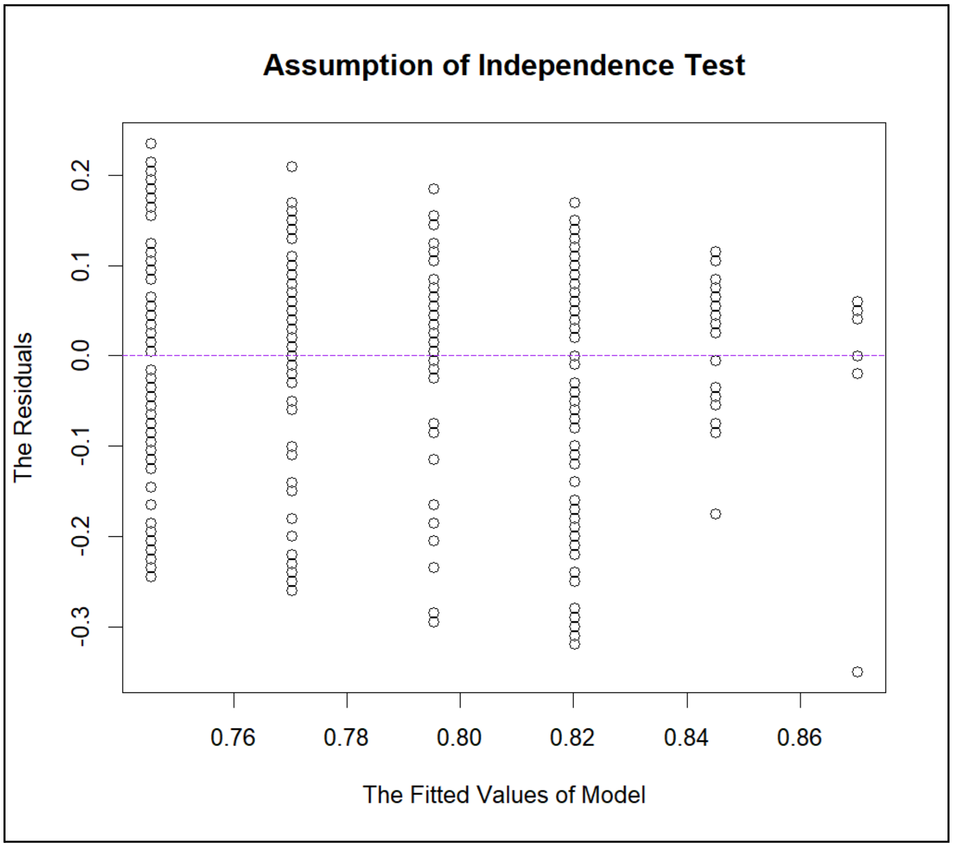


Figure H



Bibliography

Equilibrium *'Sleep Efficiency Dataset '* , Kaggle, Available at: https://www.kaggle.com/datasets/equilibriumm/sleep-efficiency/data (Accessed: 22 January 2024).