RESEARCH ANALYSIS

Investigating the impact of lifestyle factors such as Sleep Duration, Physical Activity Level, Stress Level, Heart Rate, Blood pressure (Systolic/Diastolic), and BMI Category on Sleep Quality

Ziyi Zhang 1005282720

INTRODUCTION

Sleep quality plays a vital role in maintaining individuals' overall health and well-being. Research have found that adequate and restful sleep is essential for physical, cognitive, and emotional functioning[1]. Numerous lifestyle factors such as sleep duration, physical activity, stress[2], diet, body mass index, and occupation[3] all could contribute a significant influence in sleep quality, which could potentially lead to sleep-related disorders like insomnia and sleep apnea[4]. In order to develop effective methods and promote healthy sleep habits, it is crucial for individuals to understand the impact of these lifestyle factors on sleep quality.

The purpose of this study aims to investigate the impact of lifestyle factors[4] on sleep quality, and to examine the associations among sleep duration, physical activity level, stress, heart rate, blood pressure, and BMI category. The findings will contribute to the existing knowledge and research on sleep health and aims to provide comprehensive analysis to shed light on the complex relationship between lifestyle factors and sleep quality.

METHODS

Study population

The Sleep Health Lifestyle Dataset contains information about a group of test subjects and their sleep patterns. This analysis was conducted using data collected by surveying individuals in their late twenties to late fifties. Like any form of self-report measures, we need to keep in mind that these data collected are susceptible to various type of biases[2]. The author of this dataset did not disclose the survey instruments used to conduct the data collecting. The limitations, bias and potential uncertainties associated with the dataset is acknowledged.

Measures of Interest

Below are all the variables of interest of this study. Sleep quality, sleep duration, physical activity level, stress level was all collected through self-assessment surveys. Blood Pressure, Heart Rate, and BMI Category variables were measured and recorded down by respondents.

Sleep Quality

The Primary exposures of interest in this study was Sleep Quality. Respondents were asked to give a subjective rating of the quality of sleep, ranging from 1 to 10. This method relies on individuals' self-perception and self-reporting of their own sleep quality.

Sleep Duration

Sleep duration was calculated by the number of hours the person sleeps per day, which was self-reported by the respondents is another common approach to assess sleep habits and patterns in

research studies. Various methods that researchers could use to collect the data are questionnaires, interviews, or sleep diaries[1].

Physical Activity Level
Physical Activity Level was reported daily by the respondents. Individuals have to self-report the number of minutes that they engaged in physical activity every day. It wasn't made clear if definitions and examples of what counts as physical activity were provided to the respondents.

Stress Level
The data for Stress Level variable is collected through a survey where respondents gave a subjective rating of the stress level, they had experienced, ranging from 1 to 10. It was not disclosed how the stress level assessment was tested to measure the individual's perceived stress levels.

Blood Pressure (Systolic/Diastolic)
In the dataset Blood Pressure was presented as a categorical character variable like Systolic/Diastolic. To use this variable, two new numerical columns: Systolic and Diastolic were created. The numbers indicating systolic pressure were added into Systolic column, and the numbers indicating diastolic pressure were added into the Diastolic column.

Heart Rate (bpm)
Individuals were also asked to report their rest heart rate in beats per minute. The methods of measuring bpm were not disclosed in this dataset. Some common techniques used include electrocardiography (ECG), photoplethysmography (PPG), or wearable devices[3].

BMI Category
Body Mass Index (BMI) are usually calculated using measured individuals' height and weight. It is then classified into body weight status (Normal, Normal weight, Obese, and Overweight). Not enough data found under "Obese" and "Normal weight" thus it was added into "Overweight" and "Normal" respectively. To use a categorical data in regression analysis, BMI Category variable was transformed to numerical data, with "Normal" = 1, and "Overweight" = 2.

Variable Selection
Data from a total of 374 individuals were to be used in analysis. No apparent outliers can be detected in the data. All parameters seem intuitively related to sleep quality. Lifestyle and health factors such as sleep duration, physical activity, stress level, heart rate, blood pressure, and BMI Category were used as independent variables. Simple Linear Regression (SLR) and Multiple Linear Regression (MLR) were used to model the relationship between these lifestyle factors (independent variables) and sleep quality (dependent variable), providing insights and associations of how different aspects of lifestyle could influence sleep quality. Analysis of Variance (ANOVA) tables were used to assess the significance of the regression models and was also used to compare the fit of different models. Statistical tools and techniques like Statistical Criteria, Information Criterion and Residual Analysis were used to help in selecting the most appropriate predictors and refining the model.

After fitting a linear model, predictor variables were added or removed based on Statistical Criteria like p-values, F-statistic, and standard errors in order to test for variable significance. This will

help to determine whether the model provides a significant fit to the data. Then models were compared using Information Criterion such as Adjusted R^2, Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC) – higher values of R^2_{adj}, and lower values of AIC or BIC indicate better-fitting models. These criterions provide a quantitative measure of model fit that balances the goodness of fit with model complexity. After modelling, various evaluation metrics like R^2, mean squared error (MSE), and root mean squared error (RMSE) are used to assess the goodness of fit of the linear regression model. Assumptions of whether a linear model will hold needs to be checked by Residual Analysis. Analyzing residuals of the model can help identify any patterns or violations of assumptions. Assumptions of linear regression need to be evaluated and validated through techniques like plotting residuals against predicted values, checking for heteroscedasticity, and assessing normality – which can all guide the refinement of the final model. To correct any violated assumptions such as uncorrelated error, or non-constant variance, the final selected model was weighed and transformed to reduce the biasness in dataset.

Model Validation

The model will be validated using a train-test split, which involves dividing the original dataset into two separate subsets. The split ratio 60/40 was used, 60% for the training data set, and 40% for the testing data set. The training set was used to fit the linear regression model in R, which involves estimating the coefficients (parameters) of the model based on the training data. Once the model is trained, it is used to make predictions on the testing set. When the predicted values for the response variable is obtained, it is compared with the actual values of the response variable in the testing set. Evaluation metrics such as MSE, RMSE, and R^2 were all used to assess the performance of the model.

Model Violations and Diagnostics

Model violations that occurred in analysis were linearity assumptions, and constant variances. Preliminary informal assessment of assumptions for linearity, constant variance, and normality were conducted by analyzing scatter plots and histograms before modelling. After verifying condition 1: linearity, and condition 2: constant variance, Residual plots vs Fitted values, and Residual plots vs Predictor values for the model were built to check for assumptions. The residual vs fitted graph helps assess the assumption of linearity and homoscedasticity. Looking for patterns such as fanning and clusters in the Residual vs Predictor graphs could indicate some non-constant variance. To try and correct the violation of linearity and non-constant variance, method of weighted least squares and transformation of the model were considered.

RESULTS

Description of Data

Characteristics of the surveyed individuals are shown in Table 1. It was reported that individuals slept an average of 7.13 hours on a typical night (median 7.20 hours, range 5.8-8.5 hours). It's noted that 59.6% of individuals did not have sleeping disorder, while 20.6% had insomnia and 20.9% had sleep apnea. An interesting factor, the BMI Category, which is good indicator of a person's health – 57.8% of individuals were normal weight, while 42.2% were overweight.

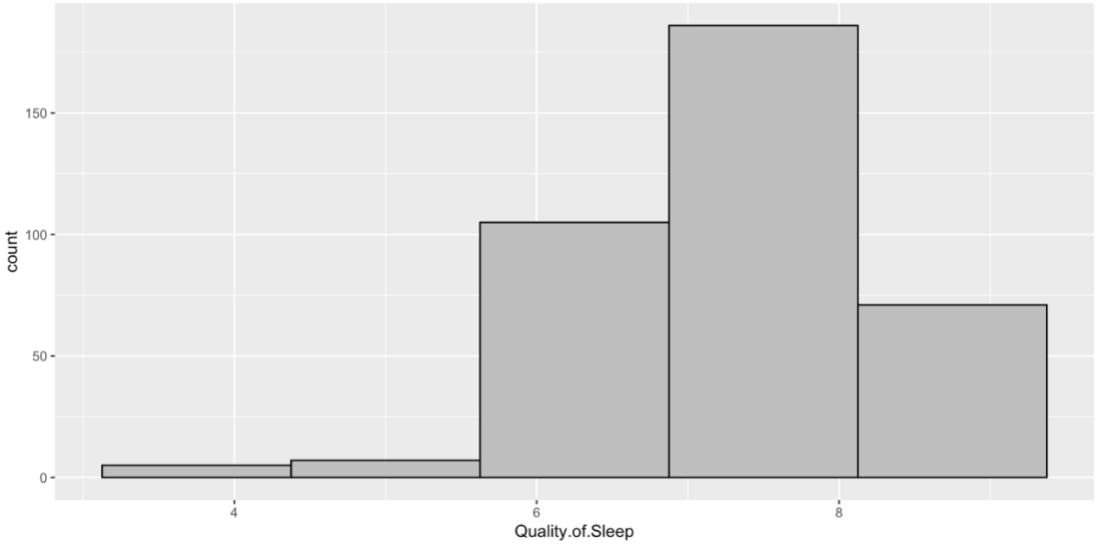
Table 1: Characteristics of individuals

	Overall (N=374)
Gender	
Female	185 (49.5%)
Male	189 (50.5%)
Age	
Mean (SD)	42.2 (8.67)
Median [Min, Max]	43.0 [27.0, 59.0]
Occupation	
Accountant	37 (9.9%)
Doctor	71 (19.0%)
Engineer	67 (17.9%)
Lawyer	47 (12.6%)
Nurse	73 (19.5%)
Others	5 (1.3%)
Salesperson	34 (9.1%)
Teacher	40 (10.7%)
Sleep.Duration	
Mean (SD)	7.13 (0.796)
Median [Min, Max]	7.20 [5.80, 8.50]
Quality.of.Sleep	
Mean (SD)	7.31 (1.20)
Median [Min, Max]	7.00 [4.00, 9.00]
Physical.Activity.Level	
Mean (SD)	59.2 (20.8)
Median [Min, Max]	60.0 [30.0, 90.0]

Table 1: Characteristics of individuals

	Overall (N=374)
Stress.Level	
Mean (SD)	5.39 (1.77)
Median [Min, Max]	5.00 [3.00, 8.00]
BMI.Category	
Normal	216 (57.8%)
Overweight	158 (42.2%)
Systolic	
Mean (SD)	129 (7.75)
Median [Min, Max]	130 [115, 142]
Diastolic	
Mean (SD)	84.6 (6.16)
Median [Min, Max]	85.0 [75.0, 95.0]
Heart.Rate	
Mean (SD)	70.2 (4.14)
Median [Min, Max]	70.0 [65.0, 86.0]
Daily.Steps	
Mean (SD)	6820 (1620)
Median [Min, Max]	7000 [3000, 10000]
Sleep.Disorder	
Insomnia	77 (20.6%)
None	219 (58.6%)
Sleep Apnea	78 (20.9%)

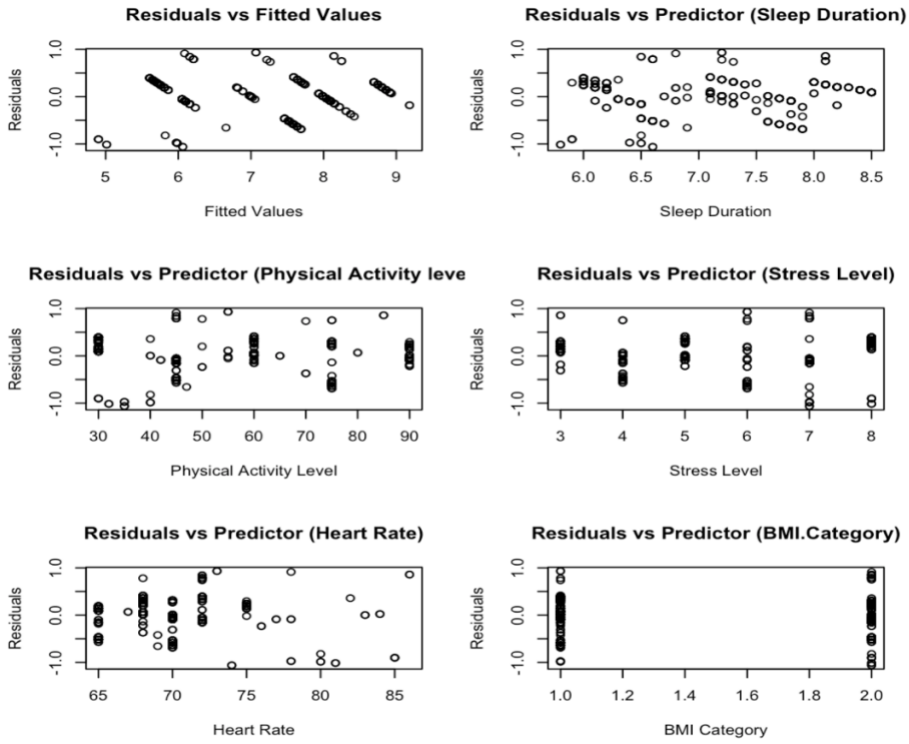
Quality of Sleep Distribution [Figure 1]



The response variable quality of sleep was distributed throughout the study population [Figure 1]. The graph of the response seems a little skewed to the right, highlighting there could be potential for a Normality violation. The histogram distributions of the variables shows that there are peculiarities with the quality of sleep, stress level, and physical activity level where they seem to be spaced out – possibly due to issues with the questions asked. The same sporadic pattern is observed in blood pressure and steps, which needs further explanation, it will likely provide problem confirming any distributions.

Analysis Process and Results

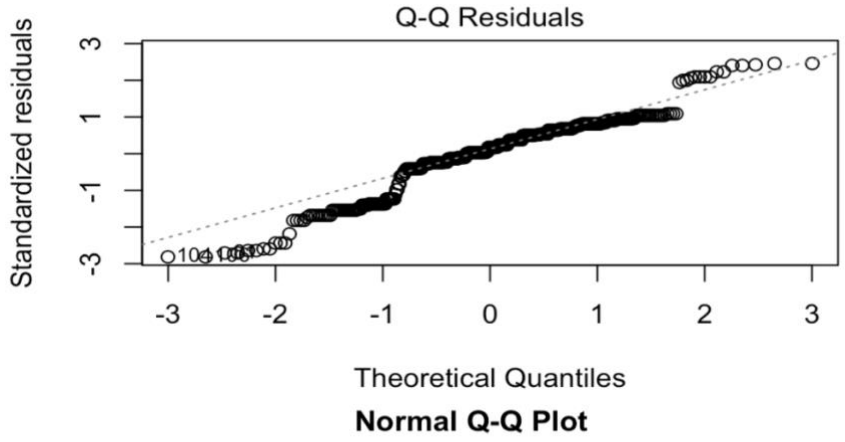
We fit an initial linear model (model_full) for the response variable Quality of Sleep that included predictors Sleep Duration, Physical Activity Level, BMI Category (R^2 = 0.78). With only the predictor Sleep Duration being significantly related to the response. So, we conducted a partial F test to compare the simple linear model (model_1) involving Sleep Duration to the initial model (model_full). The test rejected the null that all predictors in the initial model except Sleep Duration were not necessary (p-value = 0.827), thus we don't need to remove Physical Activity Level and BMI Category. So instead, we added Systolic and Diastolic predictors (model_2) to initial model (model_full) to test whether these two predictors were necessary using a partial F-test and found that they were not (p-value = 0.1721). Lastly, we added Stress Level and Heart Rate to the next model (model_3). The remaining model (model_3) with the predictors Sleep Duration, Physical Activity Level, BMI Category, Stress Level, and Heart Rate had no non-significant predictor. Thus, we ended up with a model involving only 5 predictors (R^2 = 0.9004), indicating that little information was lost by this new modelling. Model_3 has the highest adjusted R-square (R^2_{adj} = 0.899), lowest AIC (AIC= 345.3196, and lowest BIC (BIC = 373.7893).



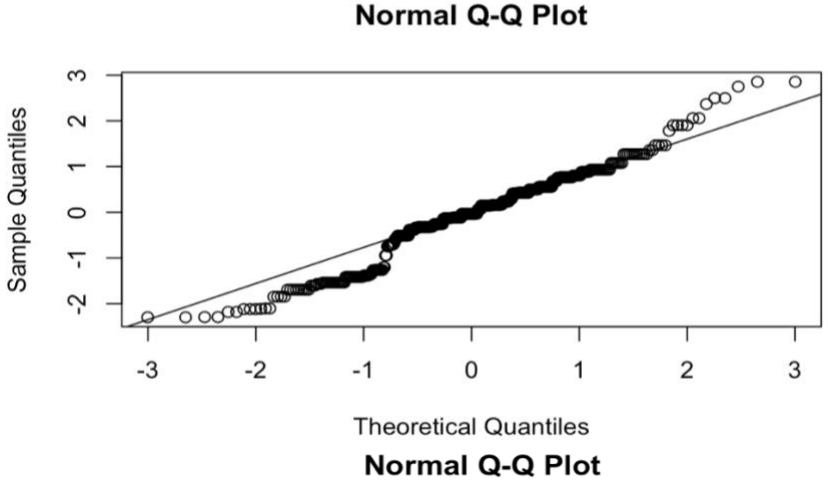
[Figure 2]

Goodness of the Final Model

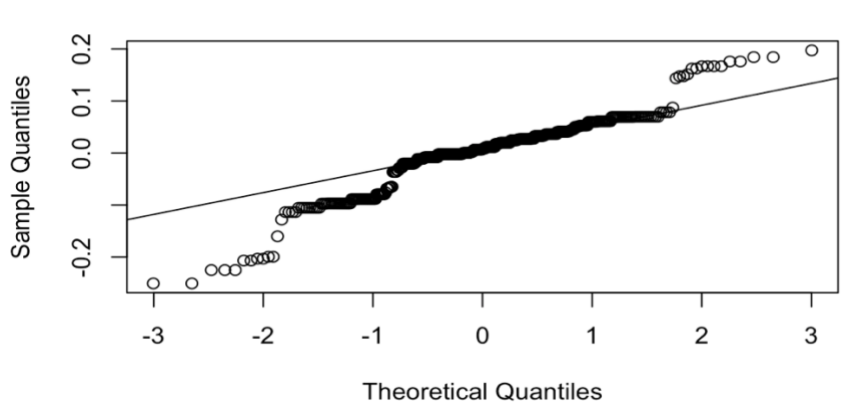
Assumptions for the model_3 were accessed and checked by building residual plots [figure 2]. In the Residuals vs Fitted values, there seems to be a discernible diagonal pattern, which could indicate violation of linearity. Not a lot of clumping noted in the interaction term plots, thus indicating that the assumption of uncorrelated errors has not been violated. Although not strong but in the Residual vs Predictor graphs there seems to be a small pattern in the residuals, which could indicate some non-constant variance. As an attempt to correct linearity and variance, the method weighted least squares and transformation were conducted separately on the model. Then the weighted least squares model [figure 4] and the transformation model [figure 5] were compared to model_3 [figure 3] to check for improvement of linearity and constant variance assumptions, and the fit of the model.



Q-Q plot model_3 [Figure 3]



Q-Q plot methodweights [Figure 4]



Q-Q plot methodtrans [Figure 5]

To validate the final model, sample data were randomly into testing and training data sets, with 60% of observations used for training, and 40% of observations used for testing. The models: model_1, model_2, model_3 was fit to the training data, and predictions for training data and testing data were made using the training models. The RMSE for both testing and training data were generated for each model [figure 6]. It's shown that model_3 had the lowest RMSE for both the train data (0.3733820), and test data (0.3889738), which indicates that model_3 is the best model performance. The trained model 3 (mod3_train) have very similar R^2_{adj} to model_3, all the same predictors appear as significant [figure 7], and the predictions seem to be consistent.

[Figure 6] RMSE for testing and training data

Model <chr>	RMSE_train <dbl>	RMSE_test <dbl>
model_full	0.5581154	0.5693441
model_1	0.5601919	0.5640943
model_2	0.5503277	0.5789503
model_3	0.3733820	0.3889738

[Figure 7] Summary of model_3 and trained model_3 for comparison

Model_3 Summary

```
Call:
lm(formula = Quality.of.Sleep ~ Sleep.Duration + Physical.Activity.Level +
    BMI.Category + Stress.Level + Heart.Rate, data = data2)

Residuals:
    Min      1Q  Median      3Q     Max
-1.0620 -0.1553  0.0647  0.2549  0.9322

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           7.924533   0.573804  13.811  < 2e-16 ***
Sleep.Duration        0.519903   0.051152  10.164  < 2e-16 ***
Physical.Activity.Level  0.007145   0.001036   6.894 2.37e-11 ***
BMI.Category         -0.165351   0.048685  -3.396 0.000757 ***
Stress.Level         -0.348086   0.024420 -14.254  < 2e-16 ***
Heart.Rate           -0.037523   0.006976  -5.379 1.34e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3804 on 368 degrees of freedom
Multiple R-squared:  0.9004,    Adjusted R-squared:  0.899
F-statistic: 665.1 on 5 and 368 DF,  p-value: < 2.2e-16
```

Trained Model_3 Summary

```
Call:
lm(formula = Quality.of.Sleep ~ Sleep.Duration + Physical.Activity.Level +
    BMI.Category + Stress.Level + Heart.Rate, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-1.08273 -0.15354  0.02363  0.23962  0.91914

Coefficients:
                      Estimate Std. Error t value
(Intercept)           7.766419   0.735350  10.562
Sleep.Duration        0.461980   0.066094   6.990
Physical.Activity.Level  0.006976   0.001339   5.209
BMI.Category         -0.171414   0.064174  -2.671
Stress.Level         -0.388678   0.030392 -12.789
Heart.Rate           -0.025918   0.008335  -3.110
                     Pr(>|t|)
(Intercept)           < 2e-16 ***
Sleep.Duration       3.31e-11 ***
Physical.Activity.Level 4.38e-07 ***
BMI.Category          0.00813 **
Stress.Level          < 2e-16 ***
Heart.Rate            0.00212 **
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3785 on 218 degrees of freedom
Multiple R-squared:  0.8961,    Adjusted R-squared:  0.8937
F-statistic: 375.9 on 5 and 218 DF,  p-value: < 2.2e-16
```

DISCUSSION

Final Model Interpretation and Importance

In the final linear regression model (model_3), we aimed to investigate the impact of lifestyle factors such as sleep duration, physical activity, stress level, heart rate, and BMI category on sleep quality. After validating the model, we obtained the following results:

The coefficient for the predictor variable "Sleep Duration" was found to be statistically significant (p-value < 2e-16), which indicates that it has a significant impact on the response variable Quality of Sleep. Assuming all other variables in the model are held constant, the average of response variable Quality of Sleep increases or decreases by a certain amount determined by the coefficient value when the predictor variable Sleep Duration increase ($\beta_1 = 0.5$).

The findings in the model summary implies that individuals that engage with longer sleep duration tends to have higher or lower quality of sleep depending on the sign and magnitude of the coefficient. This relationship was observed while accounting for the other predictor variables like physical activity, stress level, heart rate, and BMI category. The other predictor variables were all found to be statistically significant, indicating significant impact on our response. Therefore, model_3 provides evidence that there is a statistically significant relationship between the 5 lifestyle factors stated above. This information directly addresses the research question and aids us to understand the specific impact of lifestyle and health choices on quality of sleep.

Limitation of the Analysis

The final model still violates key assumptions such as linearity, constant variance, and normality. These can lead to biased coefficient estimates, inaccurate inference, and unreliable predictions, which could potentially affect the reliability of the model's estimates and predictions. Many of the predictors are skewed, highlighting the potential to see maybe linearity problems or just poorly fitting models. Then constant variance is checked by looking at the scatterplots, which shows small indication that there may be a problem with non-constant variance. A histogram of the response variable was graphed to check for normality. The response is skewed to the right, highlighting the potential to see a Normality violation. The final model's residual graphs show diagonal parallel lines – which is a logical consequence of the fact that the dependent variable only has few possible values. Thus, transforming the variable will not change the pattern of the residual plot nor improve constant variance. These assumptions violations are inherent in the data. The lack of information about the survey instruments used can have an impact on the transparency, reliability, and validity of the dataset. It's important to note that the specific choice of test and measurements may depend on factors such as the intended population[3], purpose of the assessment, and the context in which it will be used. The fundamental problem here could be argued that it's not homoscedasticity at all, but it is the level of measurement of the data that is violating the assumptions of the linear regression analysis, and having to fit a linear model to an outcome variable that is bounded.

References

1. Henrich, Liv C., et al. "Sleep Quality in Students: Associations with Psychological and Lifestyle Factors." *Current Psychology (New Brunswick, N.J.)*, vol. 42, no. 6, 2023, pp. 4601–08, https://doi.org/10.1007/s12144-021-01801-9.

2. Brown, V., and A. Arikawa. "The Relationship Between Caffeine Intake, Lifestyle Factors, Sleep Quality and Perceived Stress in College Students." *Journal of the Academy of Nutrition and Dietetics*, vol. 121, no. 9, 2021, pp. A35–A35, https://doi.org/10.1016/j.jand.2021.06.100.

3. Bukowska, Agnieszka, et al. "Rotating Night Shift Work, Sleep Quality, Selected Lifestyle Factors and Prolactin Concentration in Nurses and Midwives." *Chronobiology International*, vol. 32, no. 3, 2015, pp. 318–26, https://doi.org/10.3109/07420528.2014.975353.

4. Zwart, Tom C., et al. "Long-Term Melatonin Therapy for Adolescents and Young Adults with Chronic Sleep Onset Insomnia and Late Melatonin Onset : Evaluation of Sleep Quality, Chronotype, and Lifestyle Factors Compared to Age-Related Randomly Selected Population Cohorts." *Healthcare (Basel)*, vol. 6, no. 1, 2018, p. 23–, https://doi.org/10.3390/healthcare6010023.