STATS 101A - Final Project Report: Predicting Sleep Duration from a person's daily information

Introduction

This research aims to develop a predictive model for sleep duration and explore the influences of different lifestyle factors on sleep duration. The source of the dataset is from Kaggle ,with 200 observations and six major variables, which are Age, Sleep Duration, Quality of Sleep, Stress Level, Physical Activity Level, Heart Rate, and Daily Steps. We deleted 6 variables: Person ID, Gender, Occupation, BMI Category, Blood Pressure and Sleep Disorder because they are categories or text descriptions. Through this analysis, we are looking for the relationships that influence sleep duration.

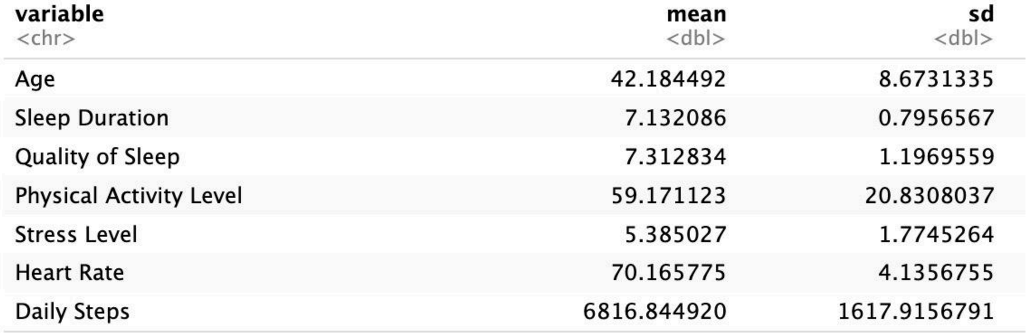
Age: The age of the person in years.
Sleep Duration (hours): The number of hours the person sleeps per day.
Quality of Sleep (scale: 1-10): A subjective rating of the quality of sleep, ranging from 1 to 10.
Physical Activity Level (minutes/day): The number of minutes the person engages in physical activity daily.
Stress Level (scale: 1-10): A subjective rating of the stress level experienced by the person, ranging from 1 to 10.
Heart Rate (bpm): The resting heart rate of the person in beats per minute.
Daily Steps: The number of steps the person takes per day.

We focus on researching how various lifestyle factors influence sleep duration which is the response variable, and the predictor variables include Age, Quality of Sleep, Physical Activity Level, Stress Level, Heart Rate, and Daily Steps. All analyses were conducted by R, and employed multiple linear regression to model the relationship between the predictor and response variables. We used this approach because an initial scatterplot matrix which suggested a linear relationship among the variables

We begin with analyzing the general summary statistics of all variables to understand their distribution and relationships. A generalized full model was initially fitted using all untransformed predictor variables, including Age, Quality of Sleep, Physical Activity Level, Stress Level, Heart Rate, and Daily Steps, to examine their impact on sleep duration. To optimize the model, power transformations were applied to certain variables to improve normality and linearity, and it was followed by stepwise variable selection to retain only the most significant predictors. Through this process, every model was evaluated for assumption satisfaction by residual diagnostics and variance inflation factor analysis to check multicollinearity. After conducting the series of investigations, a final optimized model was identified and it was concluded that Quality of Sleep, Physical Activity Level, Age, and Daily Steps were the most significant factors that influence the response variable with an Adjusted R-squared of 0.8307, and the output indicates strong predictive power.

Data Description

We first start with examining the summary statistics of the variables. The variables in this data set show many distinct distribution patterns. Age is about normally distributed around 35 to 45 years, and there are a few people at the extremes. Sleep Duration is concentrated around 6.5 to 7.5 hours. Quality of Sleep is a little bit left skewed, which peaks at 6 to 9, with very few low scores. Physical Activity Level is mostly concentrates at 30, 40, 55, 70, and 85, and very rarely distributed in other levels. Stress Level peak is somewhat uniformly distributed. Heart Rate is tightly clustered around 70 to 75 bpm, and it is right skewed with low variability. Daily Steps is a little bit left skewed, with most values between 5000 to 7000 steps.



variable <chr>	mean <dbl>	sd <dbl>
Age	42.184492	8.6731335
Sleep Duration	7.132086	0.7956567
Quality of Sleep	7.312834	1.1969559
Physical Activity Level	59.171123	20.8308037
Stress Level	5.385027	1.7745264
Heart Rate	70.165775	4.1356755
Daily Steps	6816.844920	1617.9156791

Table 1. Variable means and standard deviations

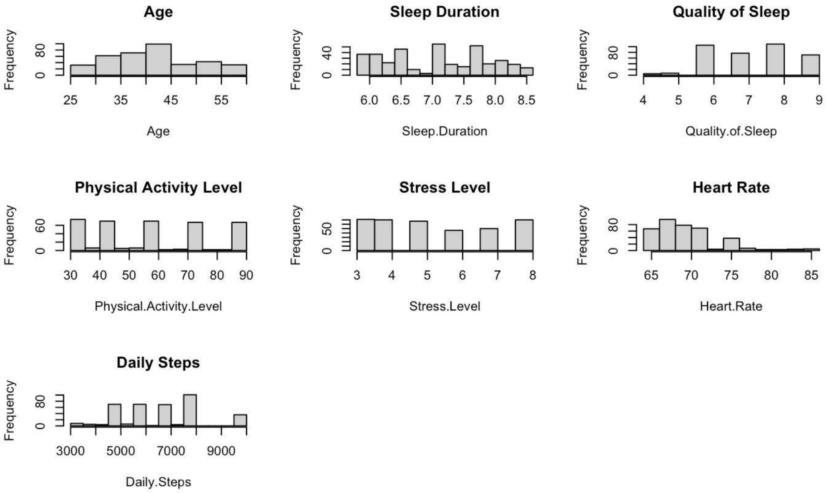


Figure 1. Distribution of variables

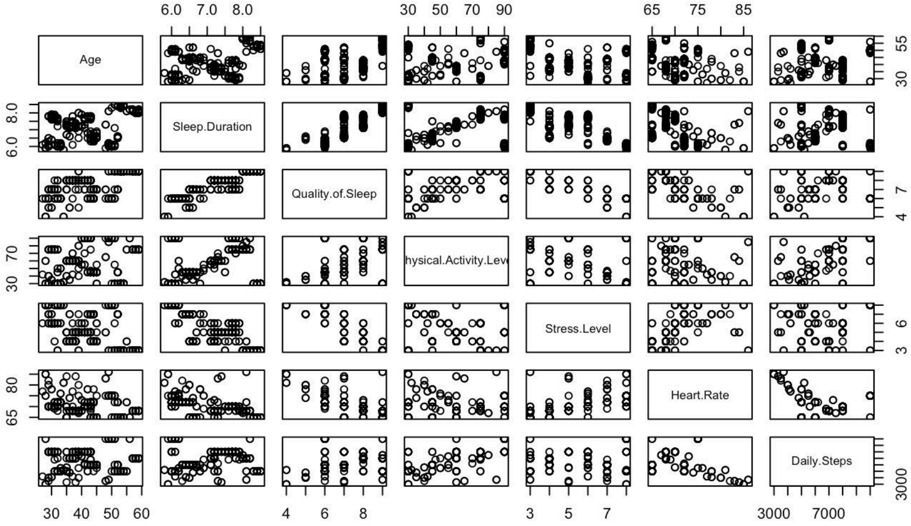


Figure 2. Scatterplot matrix of variables

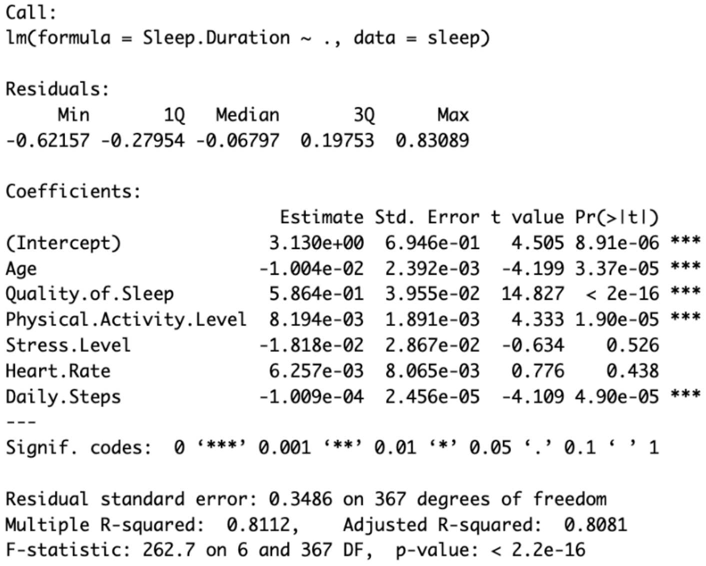
The scatter plot matrix shows that Sleep Duration has a strong positive relationship with Quality of Sleep and a weak positive relationship with Physical Activity Level. However, it has a strong negative relationship with Stress Level and no significant correlation with Age, Heart Rate, and Daily Steps, which suggests there's no multicollinearity. Stress Level has a negative linear relationship with Quality of Sleep and Physical Activity Level, and the Heart Rate shows a positive relationship with Age and a negative relationship with Physical Activity Level. Daily Steps shows a strong positive relationship with Physical Activity Level. The heteroscedasticity is very obvious, with greater variability at higher activity levels. By these patterns and stable predictor independence, a multiple linear regression model is appropriate to explore these relationships further.

Results and Interpretation

We first started our project with the linear regression model that included all predictor variables: Age, Quality of Sleep, Stress Level, Physical Activity Level, Heart Rate, and Daily Steps. The regression equation for the full model is:

$$\text{Sleep.D}\hat{\text{u}}\text{ration} = 3.130 - 1.004 \times 10^{-2} \cdot \text{Age} + 5.864 \times 10^{-1} \cdot \text{Quality.of.Sleep}$$

$$+8.194 \times 10^{-3} \cdot \text{Physical.Activity.Level} - 1.818 \times 10^{-2} \cdot \text{Stress.Level} + 6.257 \times 10^{-3} \cdot \text{Heart.Rate} - 1.009 \times 10^{-4} \cdot \text{Daily.Steps}$$



```
Call:
lm(formula = Sleep.Duration ~ ., data = sleep)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.62157 -0.27954 -0.06797  0.19753  0.83089 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)              3.130e+00  6.946e-01   4.505 8.91e-06 ***
Age                     -1.004e-02  2.392e-03  -4.199 3.37e-05 ***
Quality.of.Sleep         5.864e-01  3.955e-02  14.827  < 2e-16 ***
Physical.Activity.Level  8.194e-03  1.891e-03   4.333 1.90e-05 ***
Stress.Level            -1.818e-02  2.867e-02  -0.634    0.526    
Heart.Rate               6.257e-03  8.065e-03   0.776    0.438    
Daily.Steps             -1.009e-04  2.456e-05  -4.109 4.90e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3486 on 367 degrees of freedom
Multiple R-squared:  0.8112,	Adjusted R-squared:  0.8081 
F-statistic: 262.7 on 6 and 367 DF,  p-value: < 2.2e-16
```

Figure 3. Summary output of the Full model

Based on figure 3, we observe that Age, Quality of Sleep, Physical Activity Level, and Daily Steps have significant p-values (all < 0.001), while Stress Level and Heart Rate are not statistically significant at the 5% level. So we have four significant predictors and 2 non significant predictors in our first try. The model’s R^2 of 0.8112 indicates that these predictors collectively explain 81.12% of the variation in sleep duration. Moreover, the overall F-test yields a p-value < 2.2e-16, suggesting that at least one of the predictors is indeed important for explaining sleep duration.

The diagnostic plots(***Figure 4***) reveal key issues with the initial regression model. The Residuals vs. Fitted plot shows a curve in the mean of the variance, indicating a potential violation of the linearity assumption. The Scale-Location plot suggests heteroscedasticity, as the spread of residuals is not uniform across fitted values. The Q-Q plot displays deviations from the 45-degree line, particularly in the tails, suggesting that the residuals are not perfectly normally distributed. Lastly, the Residuals vs. Leverage plot shows potential high-leverage points, though none exhibit extreme influence requiring removal. These findings justify the need for transformations to improve model assumptions.

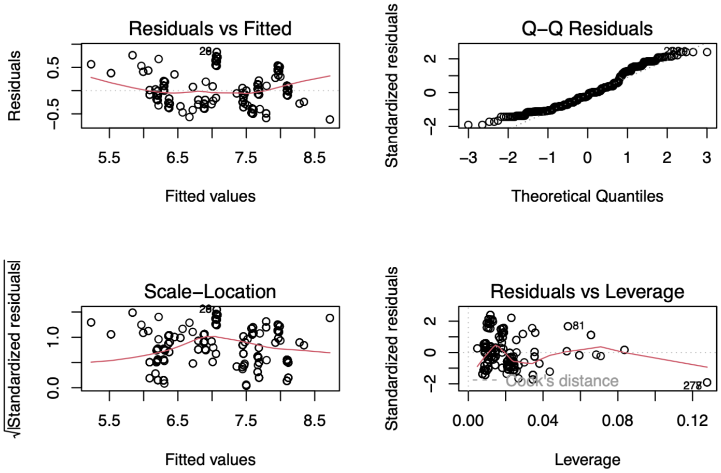


Figure 4. Diagnostic plots of the Full Model

We used the Box-Cox method to the predictors to improve the diagnostic plots and determine appropriate transformations for the predictors. The likelihood ratio test for an all-log transformation resulted in a p-value < 2.22e-16, indicating that a uniform log transformation was not ideal. Similarly, the test for no transformation was also given p < 2.22e-16, confirming that transformations were necessary. Based on these results, we applied a squared transformation to Quality of Sleep (λ=2), square root transformations to Physical Activity Level and Stress Level (λ=0.5), and an inverse transformation to Heart Rate (λ= -5.91), while Age and Daily Steps remained largely unchanged. For the response variable Sleep Duration, we used the inverse response plot to transform and got the result of log transform. So the transformed multiple linear regression model is:

$$\log(\text{Sleep Duration}) = 1.608 - 0.0021(\text{Age}) + 0.006149(\text{Quality of Sleep})^2 + 0.02107(\text{Physical Activity Level})^{0.5} - 1.502e - 05(\text{Daily Steps})$$

For this refined model, Age, Quality of Sleep, Physical Activity Level, and Daily Steps remained significant, while Heart Rate and Stress Level were not, suggesting they may not be strong predictors of Sleep Duration. The new model achieved an Adjusted R² of 0.8282, an improvement over the initial model, confirming that applying these transformations improved model fit and met regression assumptions more effectively.

This Diagnostic Plots after Transformation shows that the full model mostly satisfies the assumptions for multiple linear regression, but some violations remain. The Residuals vs. Fitted and Scale-Location plots exhibit slight curvature and a non-uniform spread, indicating non-linearity and heteroscedasticity. The Normal Q-Q plot suggests left skewness and heavy tails, signaling deviations from normality. Additionally, the Residuals vs. Leverage plot highlights potential outliers, notably observations 87 and 278, but none were influential enough to justify removal.

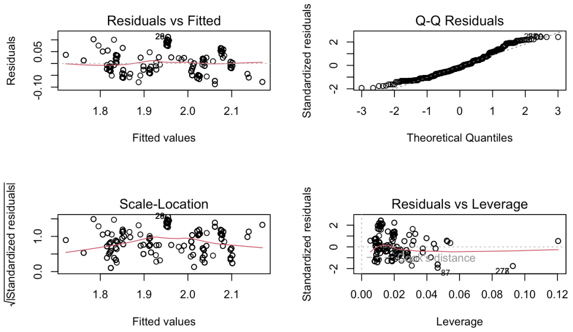


Figure 5. Diagnostic plots of the Transformed Model

After the transformation of the model, we decided to check the VIF of the model and also the added variable plot to see if the variable selection is needed

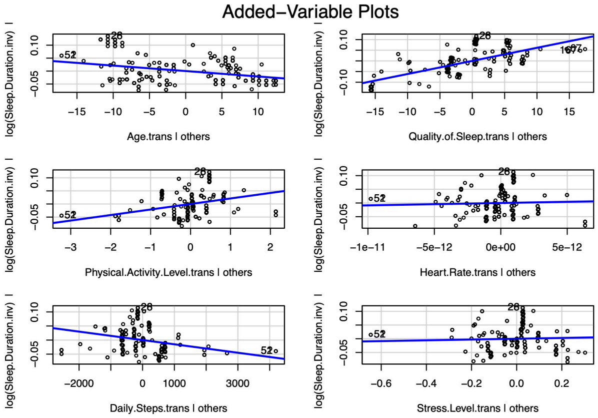
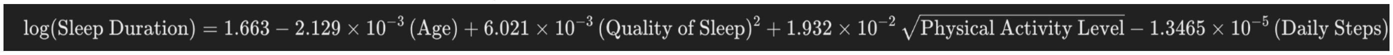


Figure 6. Added variable plot

The VIF result shows the quality of sleep. trans have a VIF of 7, and the stress level has a VIF of 11.3. Which indicates that multicollinearity happened in these two variables. Moreover, based on the added variable plot, we are able to see that the heart rate.trans and the stress level.trans have none linear relationship with the response variable sleep duration. Variable selection is needed.

For the variable selection, our group performed two ways to find the best model, the forward stepwise regression and the backward stepwise regression. Both methods suggested that removing the variables of heart rate and the stress level will give us the lowest AIC of -2291.2. I also performed a partial F test between the reduced and full models, and the resulting p-value is 0.7486 > a=0.05, so we are able to conclude that we fail to reject the null hypothesis and the reduced model is a better fit for our data.

Therefore, our final model, after being transformed and reduced, becomes:



$$\log(\text{Sleep Duration}) = 1.663 - 2.129 \times 10^{-3}\,(\text{Age}) + 6.021 \times 10^{-3}\,(\text{Quality of Sleep})^2 + 1.932 \times 10^{-2}\,\sqrt{\text{Physical Activity Level}} - 1.3465 \times 10^{-5}\,(\text{Daily Steps})$$

For our final model, we have all 4 of the predictors to be statistically significant with the p-values that are all less than 0.05. And an adjusted R^2 of 0.829.

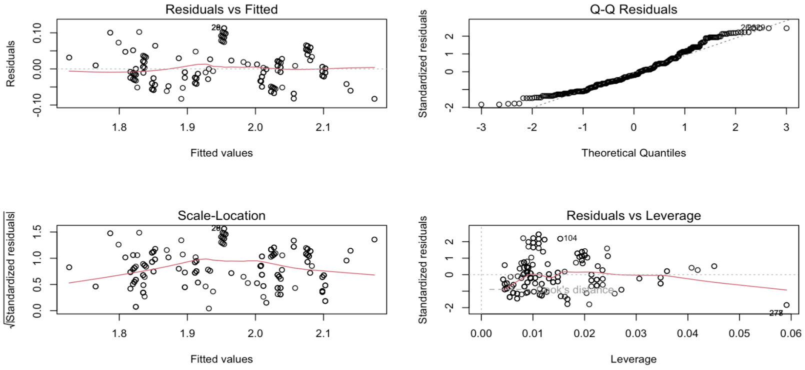
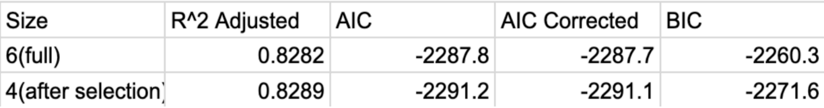


Figure 7.Diagnostic plots of the Final Model

For our diagnostic plot for this final model, we can see that there is only a good leverage point remaining, which is improved from the transformed model. The Residuals vs Fitted plot seems very good and follows the model assumption of linearity and the nonconstant variance. However, in the Standardized Residuals vs Fitted values plot, we are still able to see some mild heteroscedasticity issues remain unresolved. Also, in the Normal QQ plot, the problem of skew still remains. But based on the result from the power transformation and inverse response transformation, we stick with the transformed model. Given the slightly higher adjusted R Square value of the reduced model and the lower AIC, AIC corrected, BIC values, and the result of the partial F test, we chose the reduced model. The variable selection is needed and statistically significant.

Table 2. Data comparison between Full and Reduced model



Size	R^2 Adjusted	AIC	AIC Corrected	BIC
6(full)	0.8282	-2287.8	-2287.7	-2260.3
4(after selection)	0.8289	-2291.2	-2291.1	-2271.6

Discussion

Our study was to investigate which factors really affect an individual's sleep duration. We started with a dataset that has Age, Quality of Sleep, Physical Activity Level, Stress Level, Heart Rate, and Daily Steps and an outcome variable, Sleep Duration. All four predictors are statistically significant, showing a strong relationship with Sleep Duration. First one Age is negative, telling us that as individuals grow older, their predicted log(Sleep.Duration) decreases. Interpreted in the original scale, this implies a modest decrease in expected Sleep Duration with increasing age. Second one is quality of sleep (squared). This predictor has a strong positive effect on log(Sleep.Duration). Higher reported Quality of Sleep corresponds to higher Sleep Duration, telling us the idea that subjective sleep quality is linked to overall time spent sleeping. The third one is physical activity level (square-root transformed). This is also positive, showing us higher physical activity levels tend to experience slightly longer sleep. While moderate or regular exercise is often associated with improved sleep, the relationship here suggests that increases in activity (especially at lower ranges) may have a clear effect on sleep duration. Our last one is daily steps. Interestingly, the coefficient for Daily Steps is negative, it shows that, holding other factors constant, individuals with more total steps in a day might have slightly lower Sleep Duration. One possibility is that people who are very active or have busy schedules (accumulating many steps) may end up sleeping less. However, the effect size is relatively small, so large changes in daily steps would be required to have a more important impact on total sleep. This result correlates with findings by Dubinina and colleagues at Herzen State Pedagogical University in Saint Petersburg, who says that "having a high physical load six or more times a week is a risk factor for having insomnia symptoms, in particular, sleep-onset difficulties." Their research also tells us that "excessively high physical activity is detrimental for sleep quality," a conclusion supported by a broader study of over 9,000 adults across five countries showing that both very low (fewer than 10 continuous minutes of activity per week) and very high (more than 300 minutes per week) levels of physical activity correlated with an increased risk of insomnia. This evidence shows us the importance of finding a balanced level of exercise while moderate activity can improve sleep, extremes on either end of the activity can cause poorer sleep outcomes.

At the end, this analysis suggests that Age, Quality of Sleep, Physical Activity Level, and Daily Steps each play a meaningful role in determining Sleep Duration, with Quality of Sleep being particularly impactful. Future work and data collection and broader variable sets will refine our understanding and guide us to practice more effective sleep health.

Works Cited:

Travers, Mark. "Too Much Exercise May Be Linked to Sleep Problems." Psychology Today, 27 Oct. 2021