# Multiple Linear Regression on Sleep Data

Ethan Newcomb Kyubin Im Matthew Murnane

# Introduction

## Motivation

Sleep is an important contributor to our well-being and is deeply connected to our physical health, and cognitive function. However we often find that many individuals struggle with getting long hours sleep, often finding themselves tossing and turning throughout the night. Consequently, there is now a growing interest in understanding the lifestyle and bio metric factors that influence sleep patterns and duration. General knowledge and intuition suggest factors such as physical activity and occupational stress may influence sleep duration. Therefore, our goal is to use linear regression techniques to quantify exactly how much these factors reduce or extend sleep time.

Before, analyzing the data set, we expect a number of our indicators to contribute significantly to sleep duration. In general, we would assume that occupation will play an important role in analyzing our data as a more stressful work environment may hinder sleep patterns and disturb the body's natural sleep cycle. Additionally, we may expect that higher levels of physical activity will work in tandem with deeper and longer duration of sleep.

Also, we seek to filter through and find the most important variables in our analysis. We expect that variables like Daily Steps and Physical Activity Level will be a contribute to sleep duration in a similar manner. In our data set, we have numerous combinations of variables that are alike in this way and therefore we want to reduce our factors to give the most precise and simple model possible.

## About the Data

Table 1: Variables From Our Data

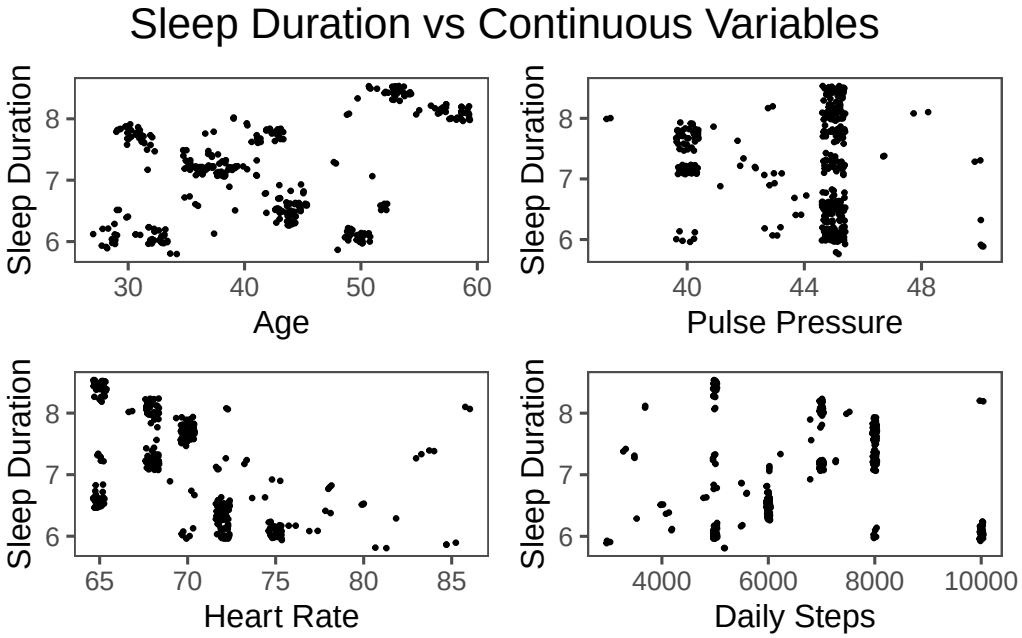
| | | |
|---|---|---|
| Gender | Age | Occupation |
| Sleep Duration (hours) | Quality of Sleep (scale: 1–10) | Physical Activity Level (minutes/day) |
| Stress Level (scale: 1–10) | BMI Category | PP (Pulse Pressure) |
| Heart Rate (bpm) | Daily Steps | Sleep Disorder |

Sleep Duration is a continuous variable and will be our response variable. The rest will be our covariates. Gender, Occupation, Quality of Sleep, Stress Level, BMI Category and Sleep Disorder are categorical variables. Physical Activity, PP, agem and Daily steps are continuous variables. We have a total of 374 observations.

We had no missing data in the data set but we did have to make a mutation using `dplyr`. In the original data set there was a blood pressure variable that was encoded as a character but was a ratio, e.g. "$\frac{120}{80}$". We separated the blood pressure variable into two: Systolic and Diastolic. These two we suspected to be highly correlated but did not want to just drop one or the other. We had options of combining the two into something meaningful. One was Pulse Pressure (PP), which is calculated as $\text{PP} = \text{systolic} - \text{diastolic}$. PP is the force of the hearts contraction on the arteries. The other is called Mean Arterial Pressure (MAP), which is calculated as $\text{MAP} = \frac{\text{systolic} \cdot \text{diastolic}^2}{3}$. MAP is the average blood pressure throughout a cardiac cycle. We decided on PP because it was simpler to calculate and understand.

For transparency it should be noted that the data is synthetic. Even though it is synthetic we still treat the analysis and report as if it was real data.

**Exploratory Data Analysis**




Visual relationship can be seen for `Sleep Duration` with `Age` and `sleep_duration`. Not so much for `PP` and `daily_steps`. The correlation between our numeric covariates is most notable between `sleep_duration` and `heart_rate`. `age` and `hearts_pressure`

Table 2: Correlation Matrix of Continuous Varaibles

| | sleep_duration | age | PP | heart_rate | daily_steps |
|---|---|---|---|---|---|
| sleep_duration | 1.00 | 0.34 | -0.16 | -0.52 | -0.04 |
| age | 0.34 | 1.00 | 0.46 | -0.23 | 0.06 |
| PP | -0.16 | 0.46 | 1.00 | 0.27 | -0.31 |
| heart_rate | -0.52 | -0.23 | 0.27 | 1.00 | -0.03 |
| daily_steps | -0.04 | 0.06 | -0.31 | -0.03 | 1.00 |

## Finding Our Model

**Full Model**

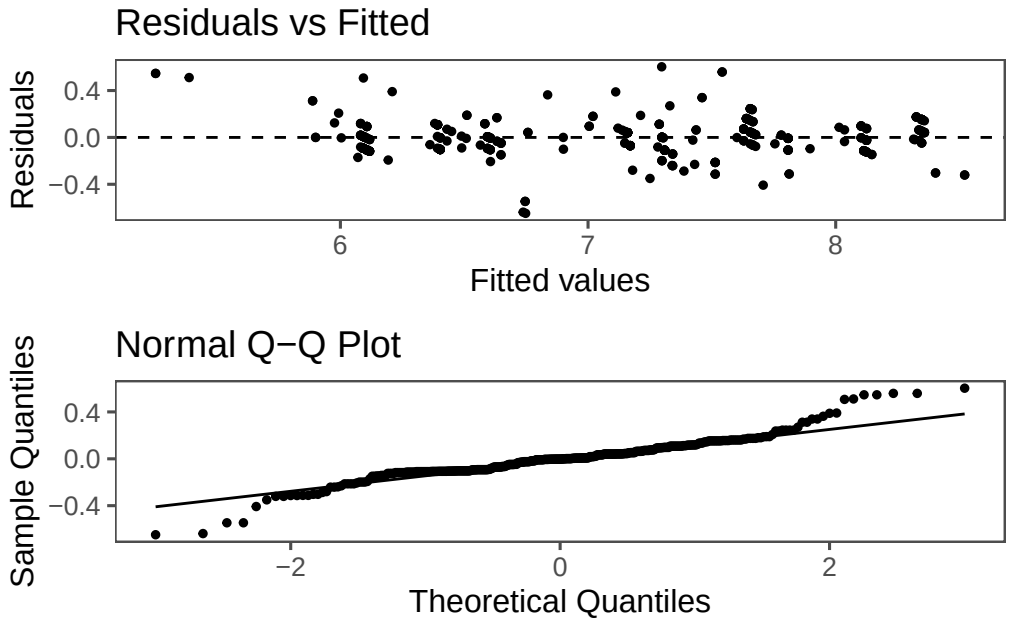
Running the full model we can see that our omnibus hypothesis test has an F-stat of 426.6 which is greater than the 99th percentile for our F-distribution, $F_{31,342,.99} = 1.74$. So we reject the null that all

coefficients are 0. We also see a high Adjusted $R^2$ and very low Residual Sum of Squares compared to the Total Sum of Squares.

Table 3: Full Linear Model Outcome

| F-statistic | Adjusted R² | Residual Sum of Squares | Total Sum of Squares |
|---|---|---|---|
| 313.34 | 0.958 | 9.28 | 236.13 |

We also saw all coefficients have standard error less than one.




As far as assumptions go homoscedasticity seems to be true and normality might be slightly violated with heavy tails but there is no excessive deviation from normality. We will explore multicollinearity and outliers next.

| | GVIF | Df | GVIF^(1/(2*Df)) |
|---|---|---|---|
| gender | 15.879968 | 1 | 3.984968 |
| age | 26.029394 | 1 | 5.101901 |
| occupation | 9618.696656 | 10 | 1.581815 |
| quality_of_sleep | 30.593110 | 1 | 5.531104 |
| physical_activity_level | 10.667470 | 1 | 3.266109 |
| stress_level | 8788.039407 | 5 | 2.479643 |
| bmi_category | 125.886061 | 3 | 2.238702 |
| heart_rate | 14.888503 | 1 | 3.858562 |
| daily_steps | 10.043217 | 1 | 3.169103 |

| | GVIF | Df | GVIF^(1/(2*Df)) |
|---|---|---|---|
| sleep_disorder | 12.533533 | 2 | 1.881561 |
| PP | 8.672479 | 1 | 2.944907 |

`age` and `quality_of_sleep` have $\text{GVIF}^{2\cdot\text{DF}} > 5$. We may want to explore Lasso to deal with this.

Looking now towards outliers we selected observations that had one of following: leverage $> 2 \cdot \frac{p+1}{n}$, studentized residual > 2, and a cooks threshold > 1. We saw a total of 57 observations that exceeded at least one of these thresholds. Since we can't say that any of these observation are errors we will only remove row 264 because it has a leverage of 1 and `NaN` for the rest. This is clearly a problem observation.

Table 5: Row with Highest Leverage

| | Leverage | Studentized | DFFITS | Cooks_Distance |
|---|---|---|---|---|
| 264 | 1 | NaN | NaN | NaN |

**Updating Model: VIF and Lasso Considerations**

We removed variables `age` and `quality_of_sleep` and ran a MLR model. Our results showed that the reduced VIF model was very similar to the full model. Adjusted R^2 was essentially unchanged. Therefore, we concluded that `age` and `quality_of_sleep` did not contribute much to the regression analysis because they were linear combinations of other independent variables.

Table 6: MLR Outcome for age and quality of sleep removed

| F-statistic | Adjusted $R^2$ | Residual Sum of Squares | Total Sum of Squares |
|---|---|---|---|
| 308.43 | 0.952 | 10.6 | 236.08 |

In order to avoid overfitting issues, we utilized a greedy algorithmn process to select variables that provided the lowest AIC value. We ran 3 model selection process as discussed in class and recorded the variables selected for each method. We found that all 3 methods took all of our variables in our data set. We knew that AIC penalizes the fitted model for each additional variables taken. Despite this, we saw that all variables were significant. Note that the variable selection process was done after our high VIF variables were removed.

Table 7: Variable Selection

| Variable | Backward | Forward | Stepwise |
|---|---|---|---|
| bmi_category | Yes | Yes | Yes |
| daily_steps | Yes | Yes | Yes |
| gender | Yes | Yes | Yes |

| Variable | Backward | Forward | Stepwise |
|---|---|---|---|
| heart_rate | Yes | Yes | Yes |
| occupation | Yes | Yes | Yes |
| physical_activity_level | Yes | Yes | Yes |
| PP | Yes | Yes | Yes |
| sleep_disorder | Yes | Yes | Yes |
| stress_level | Yes | Yes | Yes |

For our suggested regression, we used LASSO regression. We hoped to extract the most import input features to achieve a simple yet powerful model. To find the optimal $\lambda$ or the regularization parameter, we used a cross-validation process by splitting our data set: train = 80%. We saw that `gender`, `occupation`, `bmi_category`, and `sleep_disorder` converged to zero. Optimal lambda value was 0.0009622.

Table 8: Non-Zero Coefficients from LASSO (Selected by Cross-Validation)

| Variable | Coefficient |
|---|---|
| (Intercept) | 10.8659 |
| gender | 0.0000 |
| occupation | 0.0000 |
| physical_activity_level | 0.0133 |
| stress_level | -0.3173 |
| bmi_category | 0.0000 |
| heart_rate | -0.0141 |
| daily_steps | -0.0001 |
| sleep_disorder | 0.0000 |
| PP | -0.0260 |

Table 9: Optimal Lambda from Cross-Validation

| Description | Value |
|---|---|
| Optimal lambda (lambda.min) | 0.0009622 |

Table 10: Final Model Statistics

| F-statistic | Adjusted $R^2$ | Residual Sum of Squares | Total Sum of Squares |
|---|---|---|---|
| 343.56 | 0.88 | 27.61 | 236.08 |

```
Call:
lm(formula = sleep_duration ~ ., data = sleep_test_final)

Residuals:
     Min       1Q   Median       3Q      Max
-1.11141 -0.17912  0.01931  0.16623  1.14412

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             10.5004509  0.4454436  23.573  < 2e-16 ***
physical_activity_level  0.0041665  0.0008093   5.148 4.31e-07 ***
stress_level4           -1.3067982  0.0502288 -26.017  < 2e-16 ***
stress_level5           -0.7804699  0.0506108 -15.421  < 2e-16 ***
stress_level6           -0.9717802  0.0657348 -14.783  < 2e-16 ***
stress_level7           -1.6569968  0.0615906 -26.903  < 2e-16 ***
stress_level8           -2.1389377  0.0580852 -36.824  < 2e-16 ***
heart_rate              -0.0067484  0.0051389  -1.313     0.19
PP                      -0.0458751  0.0087270  -5.257 2.51e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2754 on 364 degrees of freedom
Multiple R-squared:  0.883, Adjusted R-squared:  0.8805
F-statistic: 343.6 on 8 and 364 DF,  p-value: < 2.2e-16
```

## Discussion

Upon comparing the final model to the original full model, we observe a reduction in the Adjusted $R^2$ from 0.958 to 0.880, and an increase in the Residual Sum of Squares (RSS) from 9.28 to 27.61. The F-statistic increased slightly from 313.34 to 343.56, while the Total Sum of Squares remained nearly unchanged at approximately 236. This suggests that, although the simplified model explains slightly less of the variation in sleep duration, it still performs quite well in terms of overall fit and statistical significance.

Despite the modest drop in Adjusted $R^2$, we are confident in selecting the final model. This model is much simpler, yet retains highly significant explanatory power, making it more interpretable and potentially more useful in real-world applications. The use of Lasso regularization, which penalizes more complex models, likely contributed to the decrease in Adjusted $R^2$. This trade-off was intentional. We prioritized a simpler model with a more strict variable selection. This has important practical benefits: a simpler model enables more efficient and cost-effective data collection for future studies. For example, if a healthcare provider wanted to screen individuals for sleep issues, collecting a smaller set of variables would be faster and less resource-demanding. In the final model, several predictors emerged as statistically significant at the 5% level, including: Physical Activity Level, Stress Levels 4 through 8, and Pulse Pressure (PP).

These results align well with our expectations—individuals experiencing higher stress levels or reduced physical activity tend to sleep less. Interestingly, heart rate, while retained in the model, was not statistically significant at the 5% level, suggesting its role may be less direct in predicting sleepduration. Lastly, 10-fold cross-validation showed that both the full and final models performed well, with the final model achieving only a slight decrease in $R^2$ but with a meaningful gain in simplicity. This further supports our choice to recommend the final model as a strong, interpretable, and efficient tool for predicting sleep duration.

Table 11: 10-Fold Cross-Validation Results for Initial and Final Models

| Model | RMSE | Rsquared |
|---|---|---|
| Initial Model | 0.1921801 | 0.9435390 |
| Final Model | 0.2761879 | 0.8782622 |