Multiple Linear Regression on Sleep Data

Ethan Newcomb Kyubin Im Matthew Murnane

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

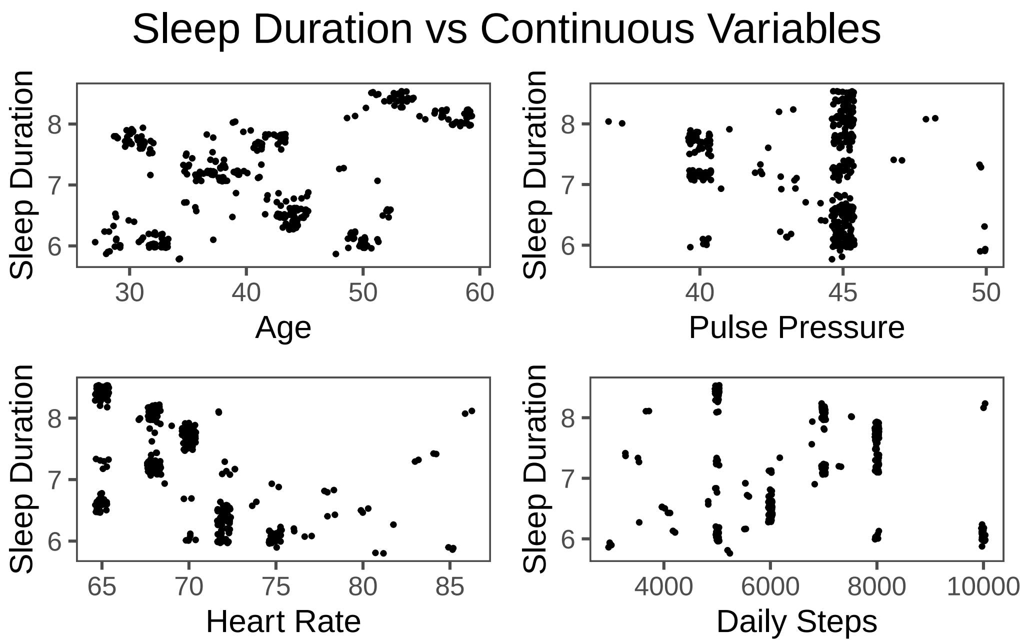
Gender	Age	Occupation
Sleep Duration (hours)	Quality of Sleep (scale: 1–10)	Physical Activity Level (minutes/day)
Stress Level (scale: 1–10)	BMI Category	PP (Pulse Pressure)
Heart Rate (bpm)	Daily Steps	Sleep Disorder

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

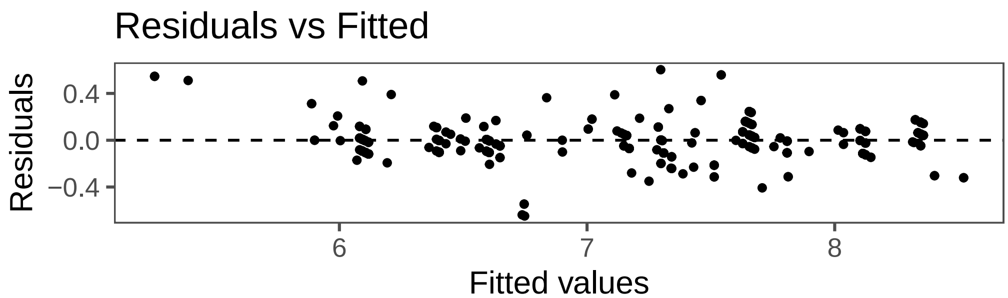
Full Model

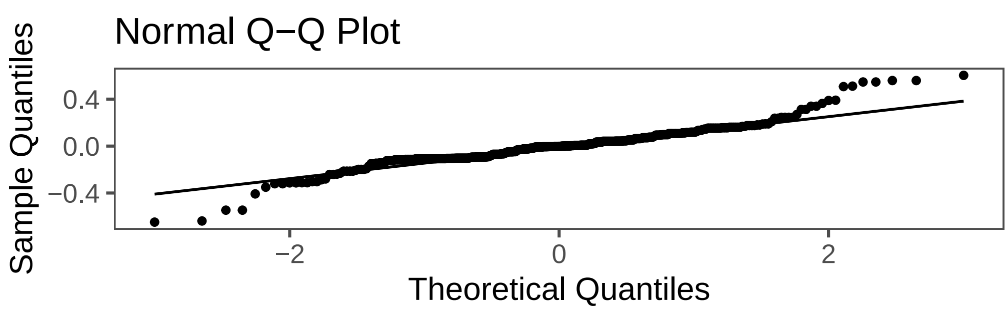
Running the full model we can see that our omnibus hypothesis test has an F-stat of 426.6 which is greater than the 99th percentile for our F-distribution, $F_{31,342,.99} = 1.74$. So we reject the null that all coefficients are 0. We also see a high Adjusted R^2 and very low Residual Sum of Squares compared to the Total Sum of Squares.

Table 3: Full Linear Model Outcome

F-statistic	Adjusted R^2	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 heteroskadcity 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
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

Updated Model