Multiple Linear Regression on Sleep Data

Ethan Newcomb Kyubin Im Matthew Murnane

Motivation

Sleep is important and people are always wondering how they can improve their sleep. Everyone knows how annoying it is to be tossing and turning all night long. Lifestyle and general health seem to be important indicators. In this report we explore a sleep data set and see what lifestyle and health variables have a significant relationship with sleep duration.

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

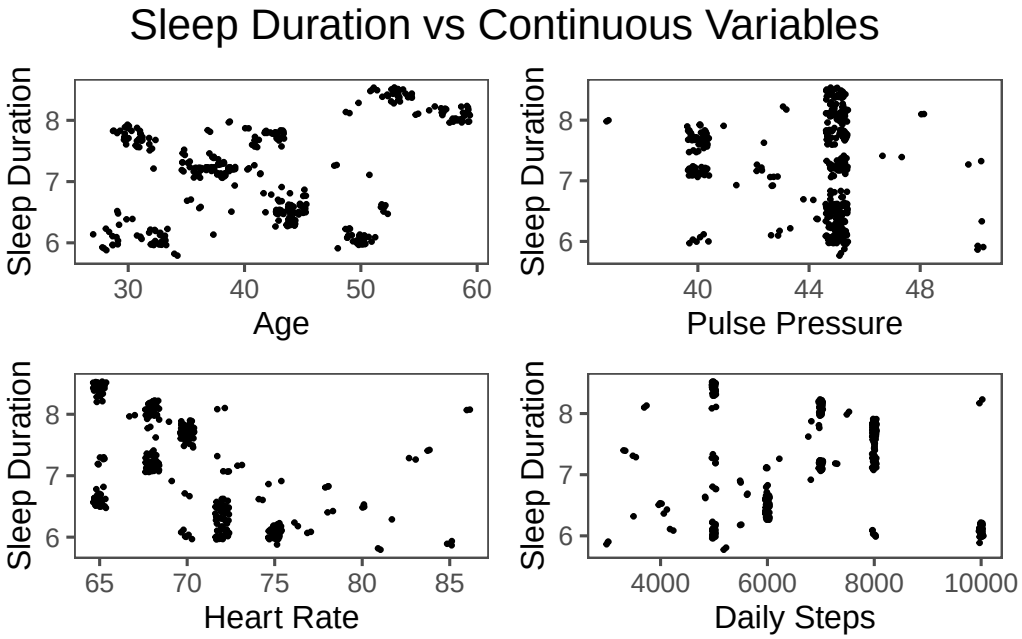


Table 2: Correlation Matrix of Continuous Varaibles

	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

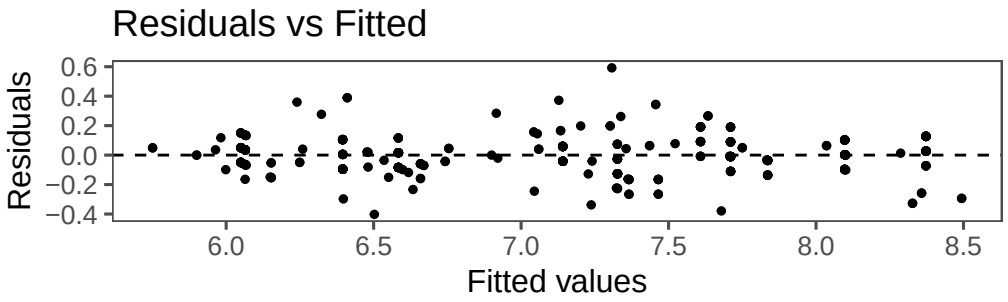
Full Model

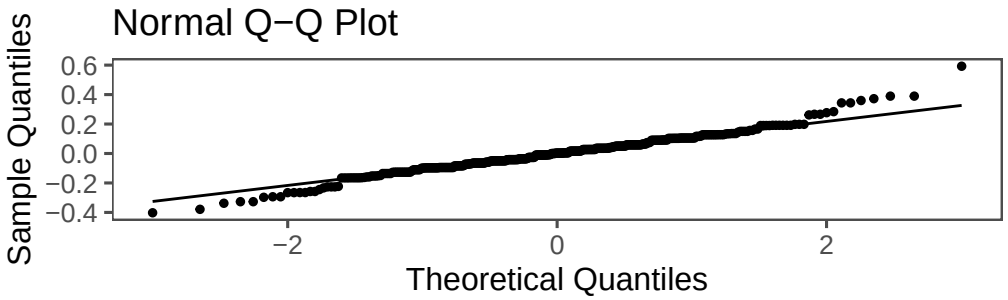
Running the full model we can see that our omnibus hypothesis test has an F-stat of 426.6 which is greater than the 99th percentile for our F-distribution, $F_{31,342,.99} = 1.74$. So we reject the null that all coefficients are 0. We also see a high Adjusted R^2 and very low Residual Sum of Squares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

Table 3: Full Linear Model Outcome

F-statistic	Adjusted R²	Residual Sum of Squares	Total Sum of Squares
426.6	0.973	5.95	236.13

As far as assumptions go heteroskadcity seems to be true and normality might be slightly violated with heavy tails but their is no excessive deviation from normality.


Residuals vs Fitted
Residuals
0.6
0.4
0.2
0.0
−0.2
−0.4
6.0
6.5
7.0
7.5
8.0
8.5
Fitted values


Normal Q−Q Plot
Sample Quantiles
0.6
0.4
0.2
0.0
−0.2
−0.4
−2
0
2
Theoretical Quantiles