STAT 306 Project

Andy Liang, Eric Tang, Shaloo Menon, Yunpeng Gao

Studying Factors Associated with Deep Sleep Percentage through Linear Regression

Introduction

We chose to use the sleep efficiency dataset that was collected as part of a study conducted by a team of researchers from the University of Oxfordshire. The researchers used participants from the local community and conducted an observational study via surveys, actigraphy, and a sleep monitoring technique called polysomnography. From the dataset, we picked a subset including the following variables associated with the subjects studied: age in years, gender between male and female, bedtime and wakeup time as timestamps (year-month-day-hours-minutes-seconds), sleep duration in hours, sleep efficiency (proportion of time asleep while in bed), deep, light, and REM sleep as a percentages, number of awakenings at night, caffeine (mg) and alcohol (oz) consumption 24 hours before bedtime, smokes status as a boolean, the number of times spent exercising in a week.

Deep sleep is associated with changes in the body. During periods of deep sleep, our bodies replace cells, build muscle tissues and heal wounds. We intend, through this project, to investigate whether we could accurately predict deep sleep percentage (DSP), our response variable, based on the other factors.

Preliminary Data Analysis

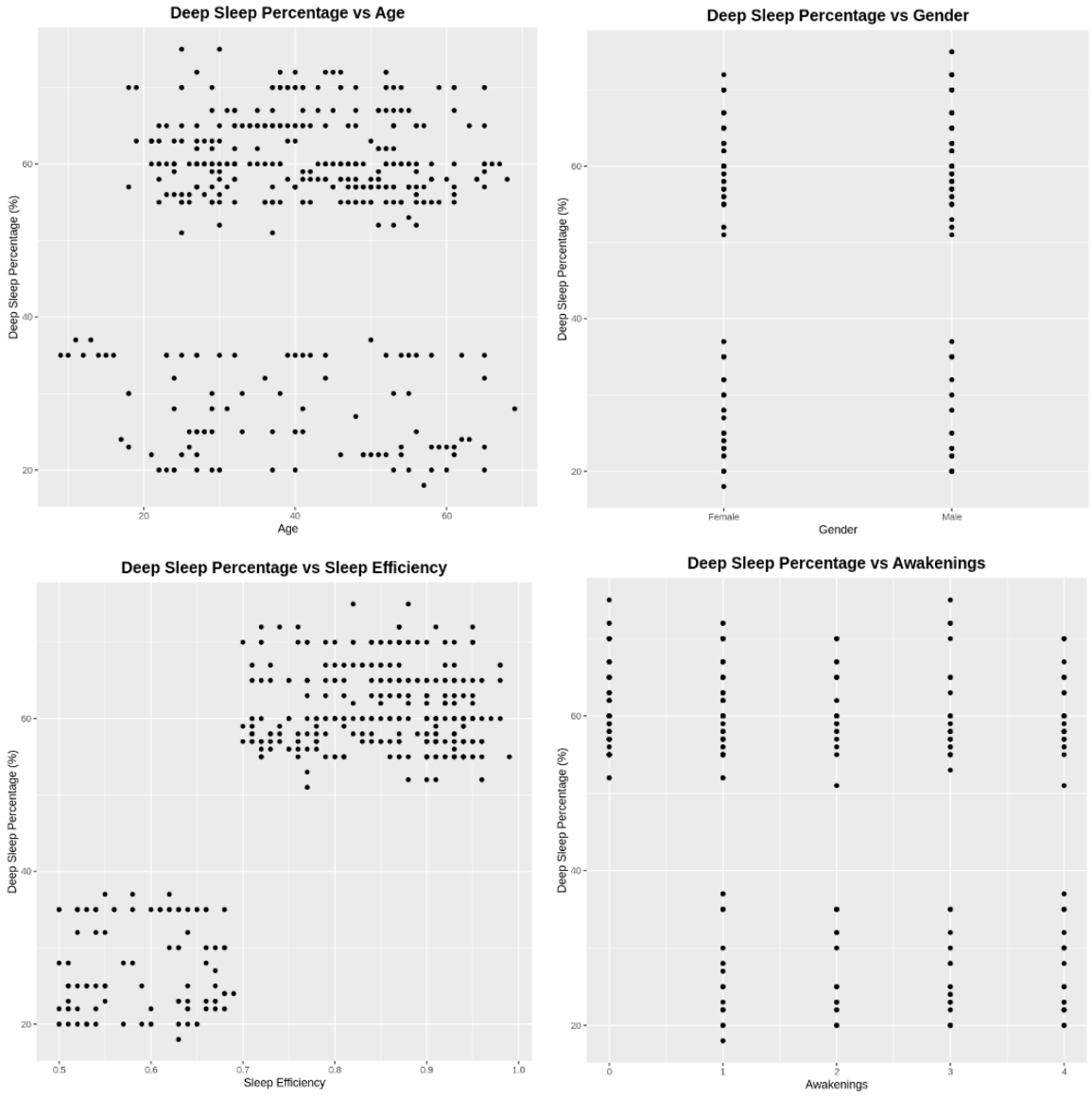
Data Wrangling and Cleaning

Immediately after reading the data of 452 observations, rows containing null values were filtered out, removing 64 unsuitable observations and leaving 388. Analysis involving the year, month, and day of Bedtime (BT) and Wake-up Time (WT) variables were not considered in favour of considering hours, minutes, and seconds. The earliest and latest BT (9pm and 2:30am) and WT (3am and 12:30pm) were found, allowing for the variables to be bisected and categorised by BT before 0am (0) and after 0am (1) and WT before 8am (0) and after 8am (1). The decision to categorise BT and WT stems from the data's observations being limited to certain BT and WT hours. This means that exploring interaction with BT or WT would be difficult. Furthermore, light and REM sleep percentages were removed as they directly influence DSP since the sum of deep, light, and REM percentages is equal to 1.

Data Visualization

DSP was plotted against each explanatory variable to attempt to discern linear relationships. Some plots are shown below and plots that were not shown were similarly

distributed. Specifically DSP plotted against continuous variables aside from sleep efficiency and age tended to look like DSP plotted against awakenings, and DSP plotted against categorical variables tended to look like DSP plotted against gender.



While there may be a positive linear relationship between DSP and sleep efficiency, there is likely no obvious linear relationship between DSP and any other explanatory variables, however all plots seem to indicate two groups of DSP values. From this it could be hypothesised that the original dataset omitted an important categorical variable or linear term.

Model Selection

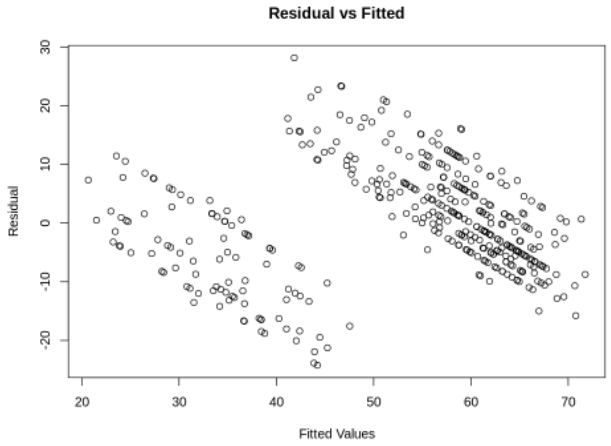
Choosing an Additive Linear Model

First, a full additive model was fitted using all explanatory variables and the output is shown below. From the summary of the model, coefficients for gender, sleep duration, smoking status, exercise frequency, bedtime and wakeup time were insignificant, suggesting that their corresponding variables may not have explanatory effects on DSP and corroborates the conclusions drawn from the plots in the preliminary data analysis. The full additive model had a rounded adjusted R-squared value of 0.6533.

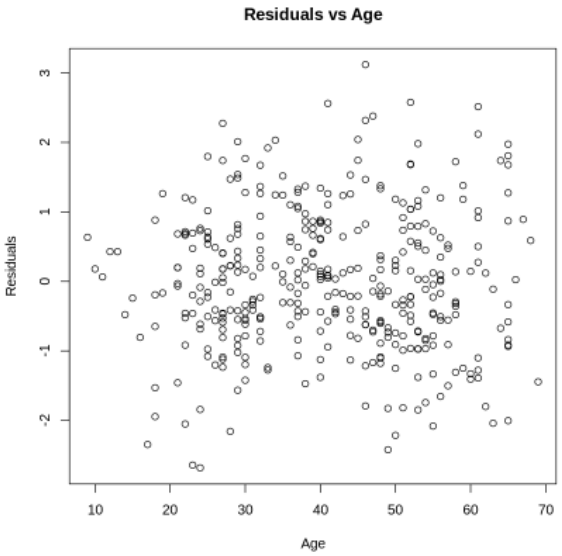
Coefficient	Estimate	Pr(>|t|)	Coefficient	Estimate	Pr(>|t|)
(Intercept)	-24.07787	0.000243	Caffeine.consumption	-0.04385	0.010482
Age	-0.08939	0.018092	Alcohol.consumption	-0.60328	0.061545
GenderMale	1.28077	0.227460	Smoking.statusYes	1.70244	0.120704
Sleep.duration	-0.56599	0.350957	Exercise.frequency	-0.33543	0.378953
Sleep.efficiency	104.13659	< 2e-16	After0am1	0.35687	0.817851
Awakenings	2.05131	3.74e-06	After8am1	0.28827	0.859873

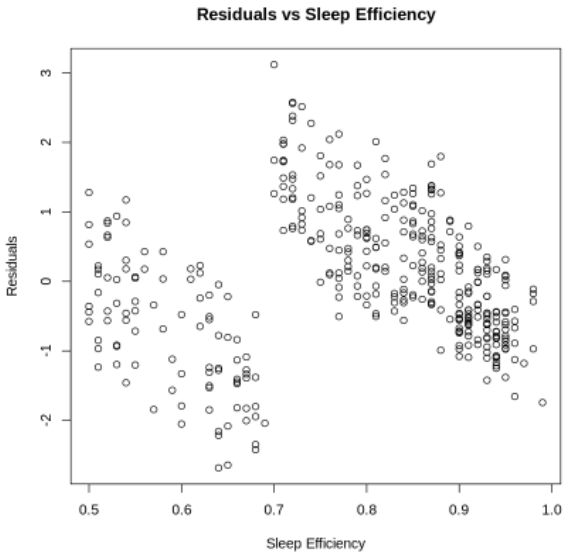
Another model was created by using the model with the highest adjusted R-squared from regsubsets, causing the removal of BT, WT, and sleep duration as explanatory variables. The resulting model included explanatory variables: age, gender, sleep efficiency, awakenings, caffeine consumption, alcohol consumption, smoking status and exercise frequency. This model proved to have a greater adjusted R-squared value of 0.655 and will be referred to as model2.

The residuals and fitted values of model2 were then plotted against each other. A clear pattern could be observed from the plot, showing two separate groups with negative linear correlation.

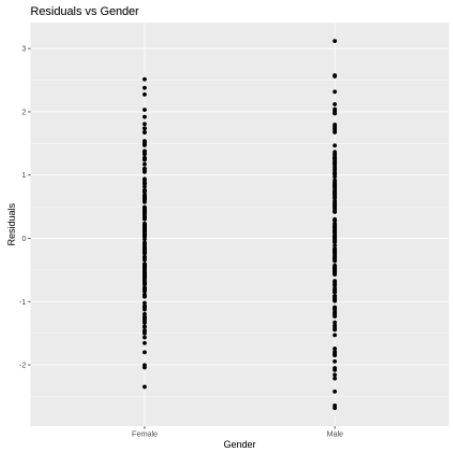


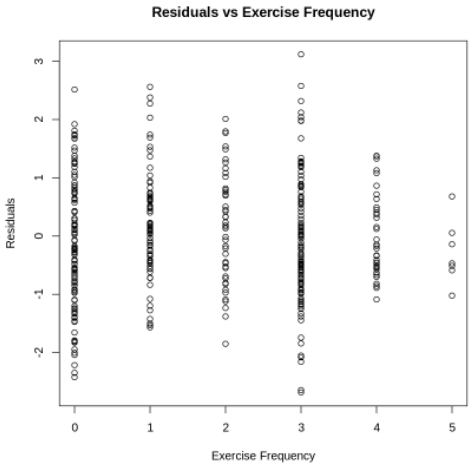
Next, the standardised residuals and explanatory variables of model2 were plotted against each other. The residuals against age plot had no pattern, while the residuals against sleep efficiency plot had a pattern that reflected the residuals versus fitted values plot. Hence, we tried adding higher order terms for sleep efficiency.





Most other residual plots against a continuous variable resembled the right graph below, and most other residual plots against a categorical variable resembled the left graph below.

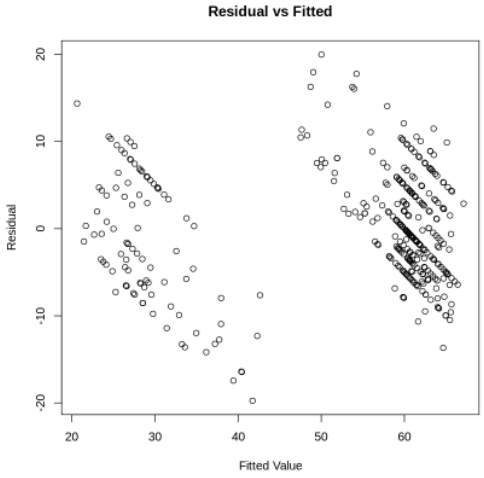




Curve Fitting

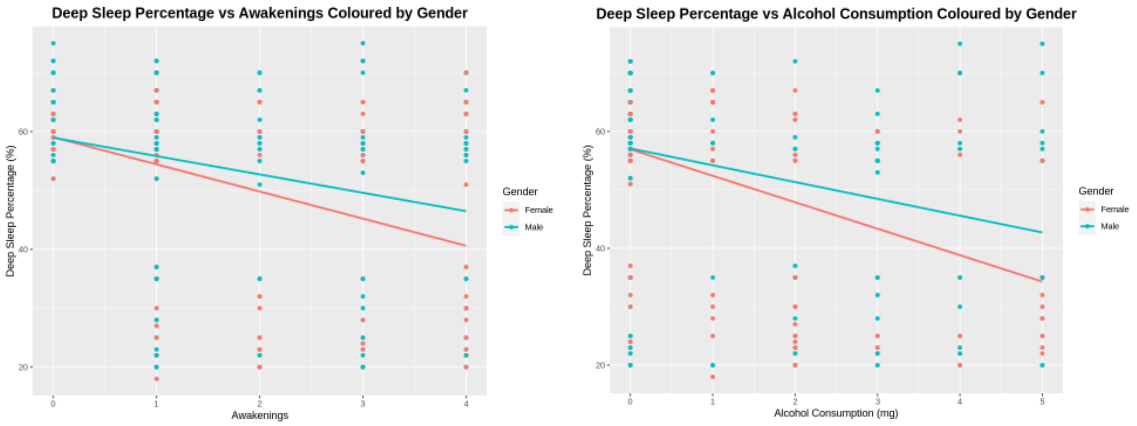
The residual versus fitted values plot of model2 had obvious patterns. Furthermore, the residual versus sleep efficiency plot also showed a pattern. Hence, the next step was to add higher order terms of sleep efficiency. The initial regsubset output also suggested that sleep efficiency is important to modelling DSP since it is present in the one variable model. Higher order terms of sleep efficiency were added according to the algorithm discussed from Activity 12. The new model with the higher order sleep efficiency terms will be referred to as

model4, containing the explanatory variables: age, gender, awakenings, caffeine consumption, alcohol consumption, smoking status, exercise frequency, sleep efficiency terms from the first to the sixth power. Model4 had an adjusted R-squared value of 0.8236 The range in the residual versus fitted plot also decreased.



Interactions

DSP was then plotted against each explanatory variable, indexed by each categorical variable in separate plots. Some of the plots showed possible interaction terms, for instance, the following plots showed possible interactions with gender.



Each model including the addition of an interaction term yielded an adjusted R-squared value within 0.001 of the original value. Therefore, adding interaction terms was necessary.

Testing Model for Improvement

2-fold cross validation was applied to compare model4 with the full additive model. Comparing the 2-fold cross validation errors revealed strong evidence that the new model yielded higher accuracy in prediction than the original additive model. After fitting the models to the testing set, the 2-fold cross validation error value from model4, 46.5, was much less than the 2-fold cross validation error from the full model, 84.6.

Conclusion

Based on our analysis, we have concluded that model4 is relatively suitable in terms of predicting DSP. Model4's adjusted R-squared value of 0.8236 indicates a relatively well-fitted model. In splitting the data into training and testing sets, we were able to train model4 and check the accuracy of its predictions on the test set. When compared to the additive model, model4 yielded a lower cross-validation error which supports our conclusion that this model yields relatively accurate predictions in comparison to other models.

For future studies, analysts may wish to consider the objective of prediction by categorising DSP into "upper" and "lower" using 0.45 as a threshold and using the categorised DSP as a response variable. This would accurately reflect the groupings seen in the preliminary data analysis of this project as it appears that across all variables, DSP has distinct upper and lower groups where 0.45 is safely between both groups. A new prediction goal could be to predict which DSP group a new observation belongs to. For a future explanatory study, seeking only to explain the trends in the given dataset, it would be intuitive to immediately split the data into two separate sets based on the same "upper" and "lower" split. This allows for a regression analysis to be performed under the assumption that the population from which the sample is drawn has individuals that can be characterised by having higher or lower deep sleep percentages. In this way, analysis would provide separate explanations for the two groups.