Predicting REM Sleep Duration Based on Lifestyle Habits

Glynn Smith

KU

DEPARTMENT OF
BIOSTATISTICS & DATA SCIENCE
The University of Kansas Medical Center

Table of Contents

List of Figures

List of Tables

Title

Analyzing the association between exercise frequency and the duration of REM sleep and predicting an individual’s REM sleep duration using variables of their lifestyle habits.

Summary

In the increasingly health-conscious climate focusing on well-being and preventive measures, grasping how lifestyle habits influence REM sleep becomes integral to overarching wellness initiatives. This understanding becomes particularly crucial for athletes, where REM sleep plays a vital role in muscle recovery.[1] It can guide the development of strategies to enhance training effectiveness, reduce injury risks, and foster enduring athletic well-being by aligning exercise routines with the body's inherent sleep patterns.

To explore the association between exercise frequency and the duration of REM sleep, along with attempting to accurately predict the duration of REM sleep based on lifestyle habits, a final model was built on ambient predictor variables. The final model was developed through the preliminary model, containing all possible ambient variables. Model selection through best subset regression and stepwise regression was conducted, followed by validation via data splitting. The resulting final model consisted of the exercise frequency, age, awakenings, alcohol consumption, and smoking status variables. Exercise frequency had very minimal positive association with the duration of REM sleep.

Introduction

The data for this analysis was collected from Kaggle.[2] The data is a result of a study conducted by AI Engineering students at the High National School for Computer Science located in Morocco. The students gathered data on 452 local participants through the following data-collection methods: actigraphy, self-reported surveys, and polysomnography. Each individual had a single observation with the following features: a unique ID, age (age of the individual), gender (gender of the individual), bedtime (date and time the individual goes to bed), wake-up time (date and time the individual wakes up), duration of sleep (total amount of hours the individual slept), efficiency of sleep (percentage of time spent in bed asleep), REM sleep percentage (percentage of time slept spent in REM sleep), deep sleep percentage (percentage of time slept spent in deep sleep), light sleep percentage (percentage of time slept spent in light sleep), number of awakenings (number of times the individual wakes up at night), caffeine consumption (number of milligrams an individual consumes in 24 hours prior to bedtime), alcohol consumption (number of fluid ounces an induvial consumes 24 hours prior to bedtime), smoker status (indicates if the individual is a smoker), and exercise frequency (number of times the individual exercises throughout the week.

Primary Analysis Aim

The aim of this analysis is to identify the relationship (if any) that exercise frequency has on REM sleep duration and to build a model to best predict REM sleep duration based on exercise frequency and lifestyle habits. Allowing an individual to observe how certain lifestyle habits impact the duration of REM sleep they experience. Potentially improving sleep quality and muscle recovery as this is the sleep stage that is most associated with feeling well rested[3] and restorative sleep.[1]

Methods

Data Source

The dataset consists of 14 total variables: ID, ranging from 1 to 452; Age, ranging from 9 to 69; Gender, with values 1 or 0; Bedtime, ranging from 2021-01-03 00:30:00 to 2021-12-31 21:00:00; Wakeup.time, ranging from 2021-01-03 08:30:00 to 2021-12-31 06:30:00; Sleep.duration, ranging from 5 to 10; Sleep.efficiency, ranging from 0.50 to 0.89; REM.sleep.percentage, ranging from 15 to 30; Deep.sleep.percentage, ranging from 18 to 75; Light.sleep.percentage, ranging from 7 to 63; Awakenings, ranging from 0 to 4; Caffeine.consumption, ranging from 0 to 200; Alcohol.consumption, ranging from 0 to 5; Smoking.status, with values 1 or 0; Exercise.frequency, ranging from 0 to 5.

Statistical Analysis

The data was downloaded and stored in .xlsx (excel) format. Data analysis was then conducted using the statistical software R and focused on multiple linear regression. Each predictor variable was individually investigated and screened during the preliminary investigation. There were 64 observations identified as having missing values and were dropped from the dataset, resulting in 388 observations. The dependent variable (REM.sleep.time) was created by multiplying the product of Sleep.duration and Sleep.efficiency by REM.sleep.percentage, then dividing by 60 to convert the time to hours. It was found that a log transformation of REM.sleep.time was needed, creating the variable LogREM.sleep.time. Identification number (idnum), Sleep.duration, Sleep.efficiency, REM.sleep.percentage, Deep.sleep.percentage, Light.sleep.percentage, and REM.sleep.time were dropped from the dataset as they are not part of the variables of interest. The variable Gender was transformed to numerical values: 1 (for male) and 0 (for female). The variable Smoking.status was transformed to numerical values: 1 (to indicate an active smoker) and 0 (to indicate a non-smoker).

The final model was determined by using best subset regression and stepwise regression in combination with criterion-based statistics and validated through data splitting. The model assumptions are evaluated and confirmed before suggesting the final model.

Model Assumption

Significance value $\alpha = 0.05$. The variables of interest are continuous and qualitative. The model assumptions: linearity, homoscedasticity, normality of errors, independence of errors, and constant variances are checked before suggesting the final model.

Preliminary Analysis

A multiple linear regression model with all possible predictors is considered. Further investigating the relationship between the dependent variable and independent variables, identifying potential outlying/influential observations, detecting skewness, and the possibility of independent and dependent variable transformations.

A multiple linear regression model with all ambient predictor variables is considered.

$$\log{(Y_i)} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \varepsilon_i$$

- Y_i is the amount of time (hours) spent in REM sleep for the i^{th}individual
- X_{i1} is the age (years) for the i^{th} individual
- X_{i2} is the gender (1 if male, 0 if female) for the i^{th} individual
- X_{i3} is the number of awakenings that occurred at night for the i^{th} individual
- X_{i4} is the amount of caffeine (in mg) consumed 24 hours prior to the bedtime of the i^{th} individual
- X_{i5} is the amount of alcohol (in fl oz.) consumed 24 hours prior to the bedtime of the i^{th} observation
- X_{i6} is the smoking status (1 if active smoker, 0 if non-smoker) of the i^{th} individual
- X_{i7} is the weekly exercise frequency of the i^{th} individual
- ε_i is the error term: $\varepsilon_i \sim iidN(0, \sigma^2)$
- $i = 1,2,3, \dots 388$
- β_0, β_1, β_2, β_3, β_4, β_5 and σ^2 are the unknown parameters to be estimated.

The full dataset was split into training and validation datasets for final model validation. The training dataset contains the first 195 of the total 388 observations, while the validation dataset contains the remaining 193 observations. The preliminary model was built using the training dataset. The log transformation of the dependent variable was found to be necessary while analyzing the distribution of the model.

Final Model

Model selection was conducted through the analysis of criterion-based statistics produced by the best subsets regression and stepwise regression selection methods;

resulting in the gender and caffeine consumption variables being dropped from the preliminary model. Several potential outlying observations and a large amount of potential influential observations were identified. However, the final model was found to be robust to these observations. No issues with non-constant variance, normality, heteroskedasticity, and independence were discovered. A more in-depth analysis is shown in the appendix.

The final models can be expressed as:

$$\log(Y_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon_i$$

- Y_i the amount of time (hours) spent in REM sleep
- X_1 is the weekly exercise frequency of the individual
- X_2 is the age (years) of the individual
- X_3 is the number of awakenings that occurred at night for the individual
- X_4 is the amount of alcohol (in fl oz.) consumed 24 hours prior to the bedtime of the individual
- X_5 is the smoking status of the individual: 1 if active smoker, 0 if non-smoker
- ε_i is the error term: $\varepsilon_i \sim iidN(0, \sigma^2)$
- $i = 1,2,3, \dots 388$
- β_0, β_1, β_2, β_3, β_4, β_5 and σ^2 are the unknown parameters to be estimated.

Results

The final model is shown above. In this model, exercise frequency explains about 0.94% of the variance in log(REM.sleep.time), as compared to 1.04% for age, 10.34% for awakenings, 6.98% for alcohol consumption, and 4.08% for smoking status. Based on the final model, a one-unit increase in exercise frequency is associated with an expected increase in the number of hours an individual is in REM sleep by 0.0164%, assuming the other variables remain constant. While exercise frequency shows a marginal association, age, awakenings, alcohol consumption, and smoking status exhibit more significant relationships with the dependent variable.

The final model has a coefficient of determination $R^2 = 0.2517$. Indicating that the percentage of explained variation of the model is 25.17%, leaving 74.83% of the variation left unexplained. The final model has an $MSE = 0.0567$, indicating that, on average, the squared difference between predicted and actual values is relatively small. The final model has a residual standard error value of 0.2381, indicating a relatively small average deviation of the residuals.

Limitations

There were 64 observations with missing values identified and removed from the dataset, reducing the size of the original dataset by about 15%. Most of the variables in the dataset rely on a self-reported survey. This survey appears to have collected responses in a manner so that the variables are categorical, only allowing for distinct responses for certain variables even though individuals fluctuate between these levels. Looking at caffeine consumption, there are inputs in intervals of 25mg (25, 50, 75, 100, and 200). This is also observed for alcohol consumption as there are only whole number inputs in the range of 1-5 fl oz.

The dependent variable had to be created from the product of multiple variables. This product could be wrong due to poor variable description. The sleep duration variable (the total hours the test subject slept) seems to be the difference between the bedtime and wake-up time variables. However, bedtime is described as the time the individual goes to bed and not falls asleep. Furthermore, sleep efficiency variable is described as the proportion of time the individual spent in bed asleep. Indicating that there is time between the bedtime and wake-up time when the individual is not asleep, which would affect the sleep duration variable.

Discussion and Conclusion

The estimated regression function from this data analysis is:

$$\log(Y)_i = 0.8451 + 0.0164X_1 + 0.0018X_2 - 0.0624X_3 - 0.0414X_4 - 0.1030X_5 + \varepsilon_i$$

- Y_i the amount of time (hours) spent in REM sleep
- X_1 is the weekly exercise frequency of the individual
- X_2 is the age (years) of the individual
- X_3 is the number of awakenings that occurred at night for the individual
- X_4 is the amount of alcohol (in fl oz.) consumed 24 hours prior to the bedtime of the individual
- X_5 is the smoking status of the individual: 1 if active smoker, 0 if non-smoker
- ε_i is the error term: $\varepsilon_i \sim iidN(0, \sigma^2)$
- $i = 1,2,3, \dots 388$
- $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ and σ^2 are the unknown parameters to be estimated.

The final model's small standard error and *MSE* values are indicators that the model's predictions are close to the actual values. However, there may be room for improvement, especially given that a significant portion of the variance remains unexplained ($R^2 = 0.2517$). A much higher R^2 would be desired.

To improve this analysis, better data is needed. Starting with how the data the data is collected. If a self-reported survey is going to be used, the responses should be unrestricted. Restricted responses in a self-reported survey, can diminish variability and make it harder to identify subtle patterns or trends. This can cause the survey results might to be not entirely accurate, leading to biased estimates. Having multiple observations per person in a study on lifestyle habits and REM sleep would be better than a single observation. More data improves the ability to find smaller effects, understand daily variations, and detect patterns over time. Multiple observations would increase the reliability of measurements and help account for natural differences amongst the test subjects. This approach would be more robust in dealing with missing data, as the dataset would be more comprehensive. A controlled experiment would be ideal, but very difficult to conduct for these variables, as most are lifestyle habits. Though a slight positive association between exercise frequency and REM sleep duration was observed: it could be entirely possible that the duration of REM sleep is just not associated with exercise frequency as I predicted. It may be more beneficial to study the association between the duration or intensity of an individual's workout with REM sleep rather than the frequency of their workouts. Furthermore, the type of workout needs to be accounted for as there is a large difference in the duration, intensity, and physical toll that a weightlifting session has compared to an abdominal workout or yoga session.

References

[1] Suni, Eric. “Stages of Sleep: What Happens in a Sleep Cycle.” Edited by Abhinav Singh, *Sleep Foundation*, https://www.sleepfoundation.org/stages-of-sleep#:~:text=in%20N2%20sleep.,Stage%203,the%20body%20relaxes%20even%20further.

[2] Sleep Efficiency. Kaggle; 2023. https://www.kaggle.com/datasets/equilibriumm/sleep-efficiency/data. Updated January, 2023. Accessed November 12, 2023.

[3] “Sleep: What It Is, Why It’s Important, Stages, REM & NREM.” *Cleveland Clinic*, https://my.clevelandclinic.org/health/body/12148-sleep-basics.

[4] Kutner, M.H., Nachtsheim, C.J., Neter, J., and Li, W. (2005). *Applied Linear Regression Models,* 4^{th} *ed.,* McGraw-Hill Irwin

Appendix

Distribution of Variables

Figure 1 shows the boxplots of the variables that are numerical. Gender and Smoking.status are binary variables; therefore, a bar chart is more appropriate to represent their distribution (figure 2). Figures 1 (a), (b), and (f) appear to be symmetrical. Figures 1 (c), (d), and (e) are right skewed. This skewness indicates a need for a possible logarithmic or square root transformation. However, due to how the data was collected, these variables are categorical. Therefore, no transformations are needed. Figure 1(c) and (f) have an observed outlier that will need to be investigated further. Figure 3 (a) shows the residuals vs. fitted values of the preliminary model. There appears to be an issue with non-constant variance, depicted by the expanding cone shape in the plot. Applying a log transformation to the dependent variable (REM.sleep.time) appears to have fixed the issue of non-constant variance (figure 4). The dependent variable will now be the log of REM.sleep.time (LogREM.sleep.time)

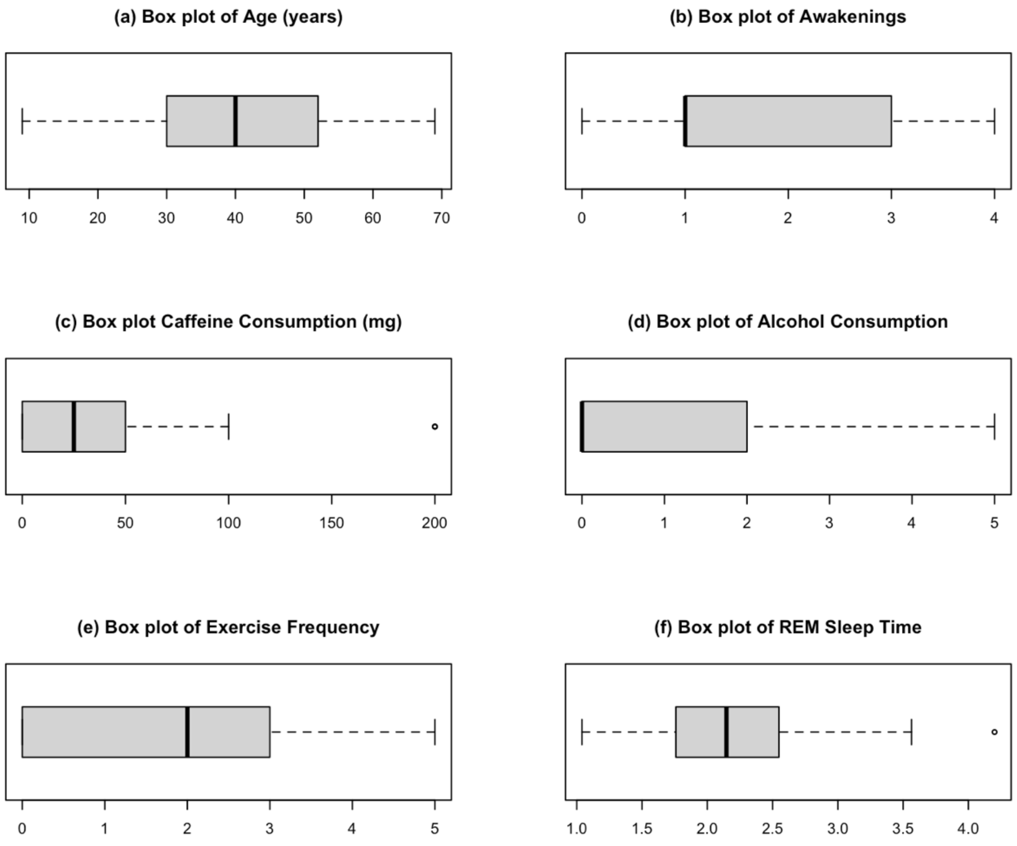

Figure 1: Box Plots of Numerical Variables

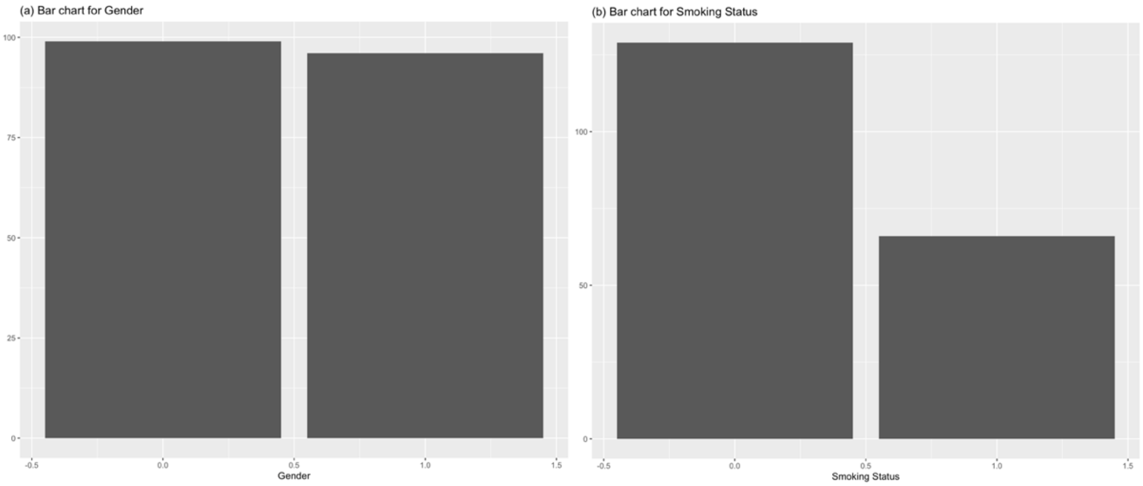

Figure 2: Bar Charts of Binary Variables

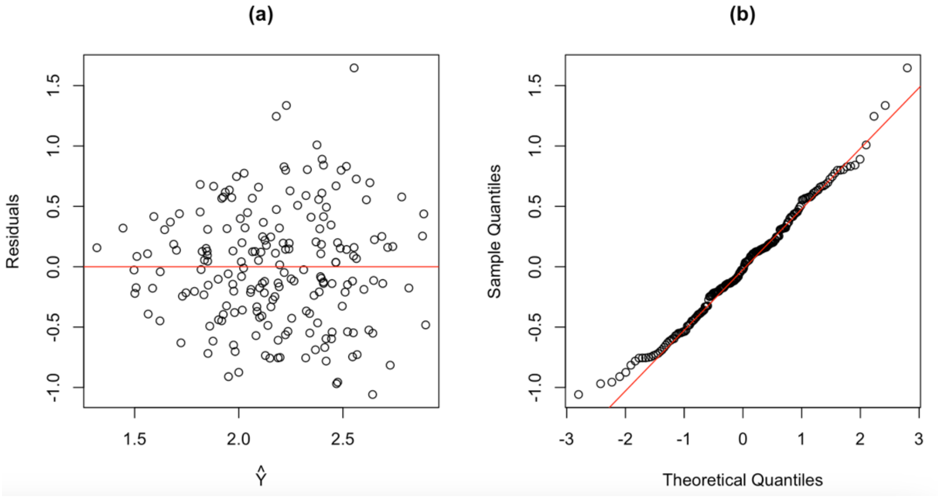

Figure 3: Preliminary Model Residual Diagnostics

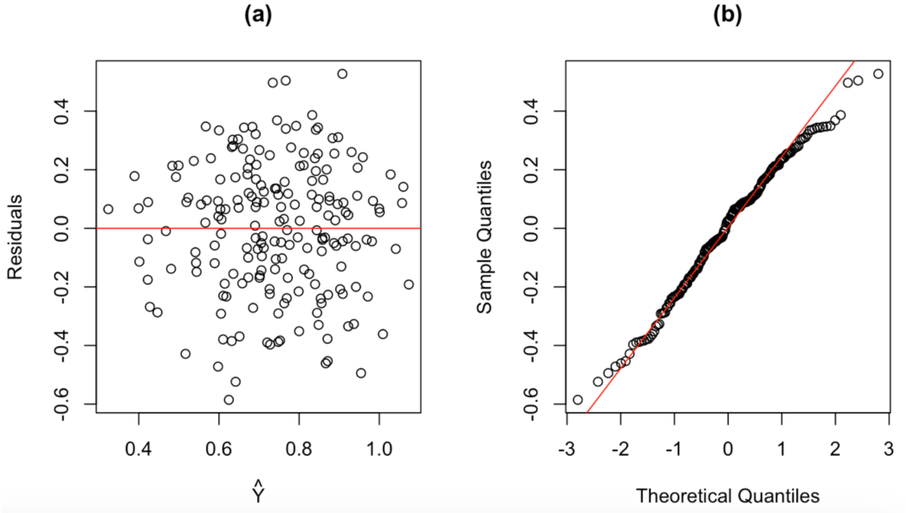

Figure 4: Log Transformed Preliminary Model Residual Diagnostics

Multicollinearity between Predictor Variables

Table 1 shows the Pearson correlation coefficient (r) between all of the variables in the dataset. Figure 5 shows both the r and a scatter plot of the variables. Each predictor variable appears to have some sort of relationship with the dependent variable (LogREM.sleep.time); the strongest being alcohol consumption. However, all predictor variables have weak relationships with the dependent variable, represented by the corresponding small r values. The closer the values of r are to $\pm$ 1, the stronger the correlation is.[4] The variable of interest, exercise frequency, has a r value of 0.2309 between LogREM.sleep time. A positive yet weak relationship. There doesn't appear to be an issue of multicollinearity between the predictor variables in the dataset.

	Log(REM sleep time)	Age	Gender	Awakenings	Caffeine consumption	Alcohol consumption	Smoking status	Exercise frequency
Log(REM sleep time)	1.0000	0.1289	-0.0747	-0.4084	0.1557	-0.3258	-0.1885	0.2309
Age	0.1289	1.0000	0.2577	-0.0400	-0.1069	0.0866	0.0621	0.1128
Gender	-0.0747	0.2577	1.0000	0.1363	-0.1513	0.0438	0.1628	0.2604
Awakenings	-0.4084	-0.0400	0.1363	1.0000	-0.2086	0.2064	0.0295	-0.1999
Caffeine consumption	0.1557	-0.1069	-0.1513	-0.2086	1.0000	-0.0738	-0.0136	0.0100
Alcohol consumption	-0.3258	0.0866	0.0438	0.2064	-0.0738	1.0000	0.0545	0.0427
Smoking status	-0.1885	0.0621	0.1628	0.0295	-0.0136	0.0545	1.0000	-0.0280
Exercise frequency	0.2309	0.1128	0.2604	-0.1999	0.0100	0.0427	-0.0280	1.0000

Table 1: Correlation Matrix for the Dataset

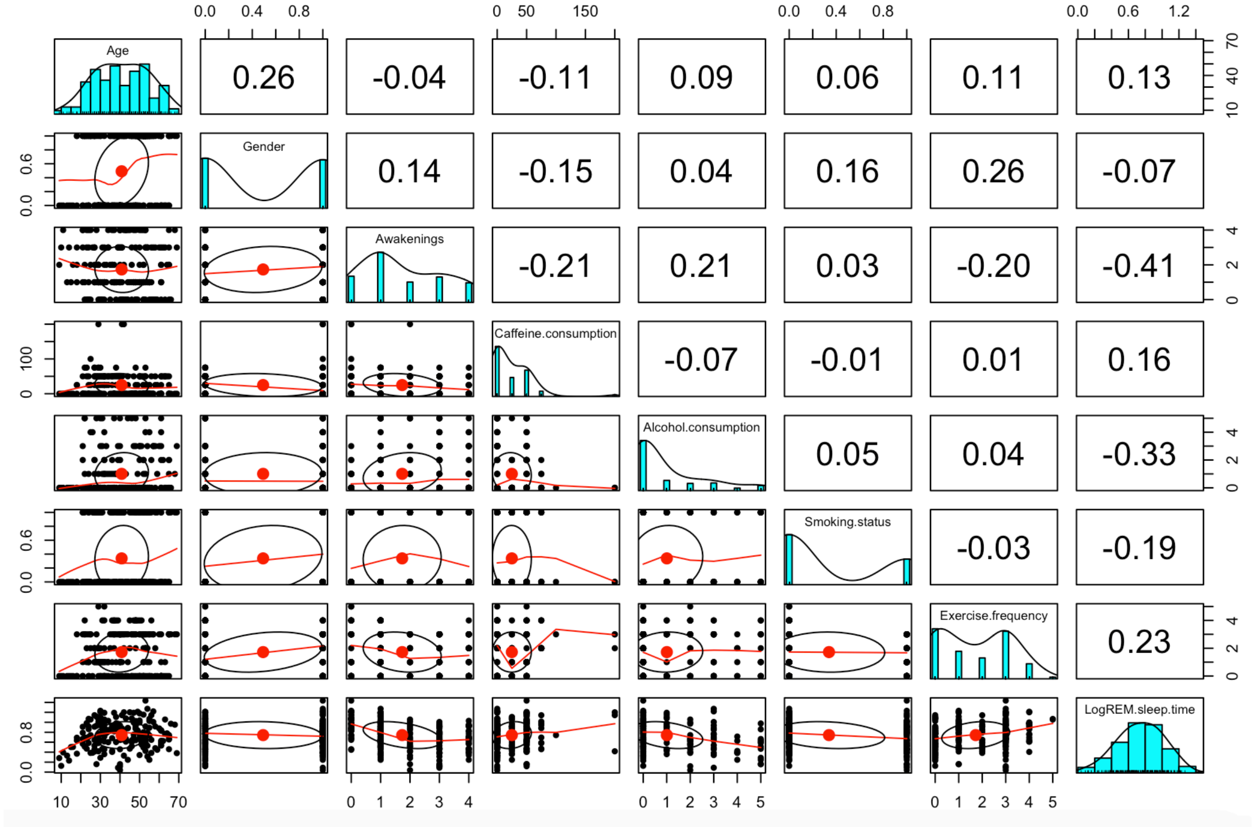

Figure 5: Scatter Plot Matrix for the Dataset

In examining the final model, the variance inflation factors *(VIF)* were calculated. The *VIF* measures how much the variances of the estimated regression coefficients are inflated as compared to when the predictor variables aren't related. When the

VIF value exceeds 10, it is an indication that multicollinearity is present.[4] Table 2 shows the *VIF* values of the final model, all being less than 10; confirming that multicollinearity is not an issue in the final model.

Exercise frequency	Age	Awakenings	Alcohol consumption	Smoking status
1.066	1.0101	1.1097	1.0615	1.0080

Table 2: VIF Values for the Final Model

Partial Residual Plots

Partial residual plots illustrate the significance of a variable within the context of other variables in the model. They depict the nature of the relationship between the variable and the dependent variable, also offering insights into whether a transformation may be necessary. Figure 6 displays the partial residual plots for the full model. Taking into consideration how the data may have been collected, it doesn't appear that any of the variables need a transformation as there are linear relationships apparent for each variable. However, none of the predictors display a strong linear relationship with the dependent variable, which is explained by their small r values shown in table 1.

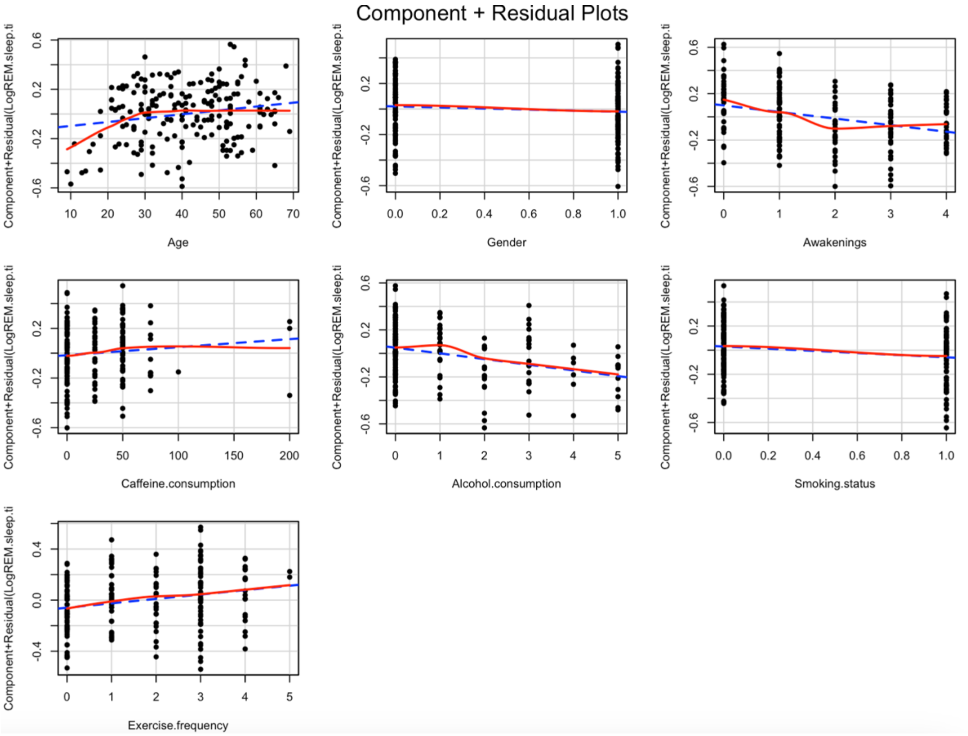

Figure 6: Partial Residual Plots for the Full Model

Model Selection

Best Subsets Regression

As a starting point in eliminating redundant variables, the best subsets regression method was used. This method identifies the several 'good' subsets for each possible number of predictor variables in the model. The variable of interest (exercise frequency) is forced into each model.

The *Mallow's CP*, Schwarz Bayesian criterion *(BIC),* residual sum of squares *(RSS),* and adjusted R_a^2 values of the suggested models are examined further to determine possible candidate models from the selection results. For adjusted R_a^2, models with the highest value are considered to be the best. Models with the lowest *BIC* and *RSS* are considered to be the best.[4] Interpreting *Mallow's CP* is subjective as there are two decision rules for this statistic:

1. Having a low *CP* value
2. *CP* = *P*, where *P* = the number of total parameters in the model

The further the *CP* value is from *P*, the more biased the model is. Thus, we want a model with a low *CP* and low bias.[4]

Table 3 shows the criterion-based statistics resulting from the best subsets regression method. There does not appear to be a clear 'best' model. Therefore, three candidate models from the best subsets regression method are selected for further consideration: model 4, model 5, and model 6 (preliminary/full model)

	Dependent variable: LogREM.sleep.time			
Predictors	C_p	R_a^2	*BIC*	*RSS*
Exercise.frequency + Awakenings	31.3359	0.1816	-25.2752	11.5451
Exercise.frequency + Awakenings + Alchohol.consumption	14.7792	0.2466	-37.1581	10.5727
Exercise.frequency + Awakenings + Alchohol.consumption + Smoking.status	9.8272	0.2687	-38.7220	10.2085
Exercise.frequency + Age + Awakenings + Alchohol.consumption + Smoking.status	7.1834	0.2824	-38.1532	9.9652
Exercise.frequency + Age + Awakenings + Caffeine.consumption +Alchohol.consumption + Smoking.status	7.2778	0.2858	-34.8442	9.8653
Exercise.frequency + Age + Gender + Awakenings + Caffeine.consumption + Alchohol.consumption + Smoking.status	8.000	0.2868	-30.8988	9.7983

Table 3: Best Subset Regression Selection Statistics

Stepwise Selection Regression

The stepwise selection regression method is a combination of forward selection and backward elimination, where predictor variables are added or removed sequentially. The variable of interest (exercise frequency) is forced into each model. Similar to best subsets regression, criterion-based statistics can be examined further to determine the best candidate model from the selection results.

Table 4 shows the criterion-based statistics resulting from the stepwise selection regression method. No clear ‘best’ model was identified. Model 5 and model 6 were selected for further consideration. Model 5 was found to be robust to both methods. The final candidate models are model 5, model 6, and model 4 from the best subsets regression method.

	Dependent variable: LogREM.sleep.time			
Predictors	C_p	R_a^2	*BIC*	*RSS*
Exercise.frequency + Awakenings	31.3359	0.1816	-25.2752	11.5451
Exercise.frequency + Awakenings + Alchohol.consumption	14.7792	0.2466	-37.1581	10.5727
Exercise.frequency + Awakenings + Alchohol.consumption + Smoking.status	9.8272	0.2687	-38.7220	10.2085
Exercise.frequency + Age + Gender +Awakenings + Caffeine.consumption	30.6207	0.1939	-15.4916	11.1932
Exercise.frequency + Age + Awakenings + Caffeine.consumption +Alchohol.consumption + Smoking.status	7.2775	0.2858	-34.8442	9.8653
Exercise.frequency + Age + Gender + Awakenings + Caffeine.consumption + Alchohol.consumption + Smoking.status	8.000	0.2868	-30.8988	9.7983

Table 4: Stepwise Selection Regression Selection Statistics

Model Validation

To further evaluate the performance of these three candidate models, validation through data splitting is conducted. Comparing the prediction sum of squares (*PRESS*), mean squared error (*MSE)*, sum of squared errors (*SSE)*, mean squared prediction error (*MSPR*), estimated regression coefficients, their estimated standard deviations, and R_a^2 for each model fitted to both the training and validation datasets. Tables 5, 6, and 7 show these values for each of the candidate models.

Models with small *PRESS* values are considered good candidate models.[4] A *PRESS* value reasonably close to *SSE*, supports the validity of the fitted regression model and of *MSE* as an indicator of the predictive capability of the model.[4] All three models have *PRESS* values reasonably close to *SSE*. However, the *PRESS* and *SSE* values are quite large, questioning the validity of each model. To examine the predictive capability of the candidate models, *MSPR* and *MSE* are considered. *MSPR* measures the predictive capability of the fitted regression model built on the training data to predict the new data in the validation dataset. If the *MSPR* is reasonably similar to the *MSE* based on the training dataset, then the *MSE* for the model is not seriously biased and gives an appropriate indication of the predictive capability of the model.[4]

The *MSPR* values for each model are 0.0627, 0.0627, and 0.0628. The *MSE* values based on the training dataset are 0.0527, 0.0525, and 0.0523. The *MSPR* will generally be larger than the *MSE* due to the introduction of new data. The *MSPR* and *MSE* values for each model are not substantially different. Indicating that the *MSE* based on the training dataset is a reasonable indicator of the predictive ability of the model.[4] The similar *MSPR* values indicate that the three candidate models perform similarly in predictive capability.

When fitting model 6 to the validation dataset, b_4 was negated. Therefore, model 6 was eliminated from further consideration. The final model selection was based on parsimony. Model 5 and model 4 perform similarly, but model 4 was able to do perform just as well with one less variable. Therefore, model 4 is chosen to be the final model and fit to the full dataset.

	Model 4	
	Training Dataset	**Validation Dataset**
$\boldsymbol{b_0}$	0.7743	0.9261
$\boldsymbol{s(b_0)}$	0.0607	0.0678
$\boldsymbol{b_1}$	0.0302	0.0025
$\boldsymbol{s(b_1)}$	0.0118	0.0127
$\boldsymbol{b_2}$	0.0026	0.0008
$\boldsymbol{s(b_2)}$	0.0012	0.0014
$\boldsymbol{b_3}$	-0.0633	-0.0589
$\boldsymbol{s(b_3)}$	0.0130	0.0137
$\boldsymbol{b_4}$	-0.0481	-0.0392
$\boldsymbol{s(b_4)}$	0.0111	0.0109
$\boldsymbol{b_5}$	-0.0962	-0.1119
$\boldsymbol{s(b_5)}$	0.0349	0.0375
SSE	9.9652	11.2878
PRESS	10.5775	12.0300
MSE	0.0527	0.0604
MSPR	0.0627	–
$\boldsymbol{R_a^2}$	0.2824	0.2023

Table 5: Model 4 Statistics

	Model 5	
	Training Dataset	**Validation Dataset**
$\boldsymbol{b_0}$	0.7416	0.9166
$\boldsymbol{s(b_0)}$	0.0650	0.0774
$\boldsymbol{b_1}$	0.0305	0.0033
$\boldsymbol{s(b_1)}$	0.0117	0.0130
$\boldsymbol{b_2}$	0.0028	0.0009
$\boldsymbol{s(b_2)}$	0.0012	0.0014
$\boldsymbol{b_3}$	-0.0595	-0.0588
$\boldsymbol{s(b_3)}$	0.0132	0.0138
$\boldsymbol{b_4}$	0.0007	0.0002
$\boldsymbol{s(b_4)}$	0.0005	0.0008
$\boldsymbol{b_5}$	-0.0478	-0.0389
$\boldsymbol{s(b_5)}$	0.0111	0.1100
$\boldsymbol{b_6}$	-0.0962	-0.1128
$\boldsymbol{s(b_6)}$	0.0348	0.0377
$\boldsymbol{SSE}$	9.8653	11.2838
$\boldsymbol{PRESS}$	10.6406	12.1527
$\boldsymbol{MSE}$	0.0525	0.0607
$\boldsymbol{MSPR}$	0.0627	$-$
$\boldsymbol{R_a^2}$	0.2858	0.1983

Table 6: Model 5 Statistics

	Model 6	
	Training Dataset	**Validation Dataset**
$\boldsymbol{b_0}$	0.7364	0.9258
$\boldsymbol{s(b_0)}$	0.0652	0.0780
$\boldsymbol{b_1}$	0.0031	0.0012
$\boldsymbol{s(b_1)}$	0.0013	0.0014
$\boldsymbol{b_2}$	-0.0412	-0.0407
$\boldsymbol{s(b_2)}$	0.0365	0.0419
$\boldsymbol{b_3}$	-0.0567	-0.0578
$\boldsymbol{s(b_3)}$	0.0135	0.0138
$\boldsymbol{b_4}$	0.0006	-0.0001
$\boldsymbol{s(b_4)}$	0.0005	0.0009
$\boldsymbol{b_5}$	-0.0483	-0.0394
$\boldsymbol{s(b_5)}$	0.0111	0.0110
$\boldsymbol{b_6}$	-0.0895	-0.1052
$\boldsymbol{s(b_6)}$	0.0353	0.0385
$\boldsymbol{b_7}$	0.0345	0.0049
$\boldsymbol{s(b_7)}$	0.01224	0.1310
SSE	9.7983	11.2266
PRESS	10.6747	12.2107
MSE	0.0523	0.0607
MSPR	0.0628	–
$\boldsymbol{R_a^2}$	0.2868	0.1981

Table 7: Model 6 Statistics

Outliers and Influential Observations

To detect outliers and potential influential observations, the deleted studentized residuals, leverage, DFFITS, DFBETAS, and Cook's distance values are considered. For the deleted studentized residuals, a formal test by means of the Bonferroni test procedure is applied. If the absolute value of the deleted studentized residual ($|t_i|$) is larger than the critical value ($qt\left(1-\frac{\alpha}{2n};n-p-1\right)$), then the observation is considered an outlier regarding its Y-value.[4] Setting $\alpha = 0.05$, n = 388 observations, p = 6 total parameters, the

Bonferroni critical value is 3.868316. No observations were found to be outliers regarding their Y-values (figure 8(a)).

The leverage values (h_{ii}) are taken from the diagonal elements of the hat matrix. If h_{ii} is larger than $\frac{2*p}{n}$, the observation is considered outlying regarding its X-value.[4] The critical value for h_{ii} is 0.0309. Observations 118, 155, 258, 303, and 379 were identified as being outliers regarding their X-values. These observations will need further investigation to see how influential they are.

For *DFFITS*, if the absolute value exceeds $2\sqrt{\frac{p}{n}}$ the observation is considered influential.[4]The critical value for *DFFITS* is 0.2487. There were 15 cases identified based on their *DFFITS* value (figure 8(b)). The high-leverage observations 258 and 303 are included in these influential observations.

For *DFBETAS*, if the absolute value exceeds $\frac{2}{\sqrt{n}}$ the observation is considered influential.[4] The critical value for *DFBETAS* is 0.1015. There were 128 beta values identified as being influential based on their absolute *DFBETAS* value (figure 8(c)).

For Cook's distance, the value is related to the $F(p, n-p)$ distribution. The distribution is then compared to the corresponding percentile value. If the percentile value is near 50%, the observation is considered influential.[4] No influential observations were discovered based on Cook's distance (figure 8(d)).

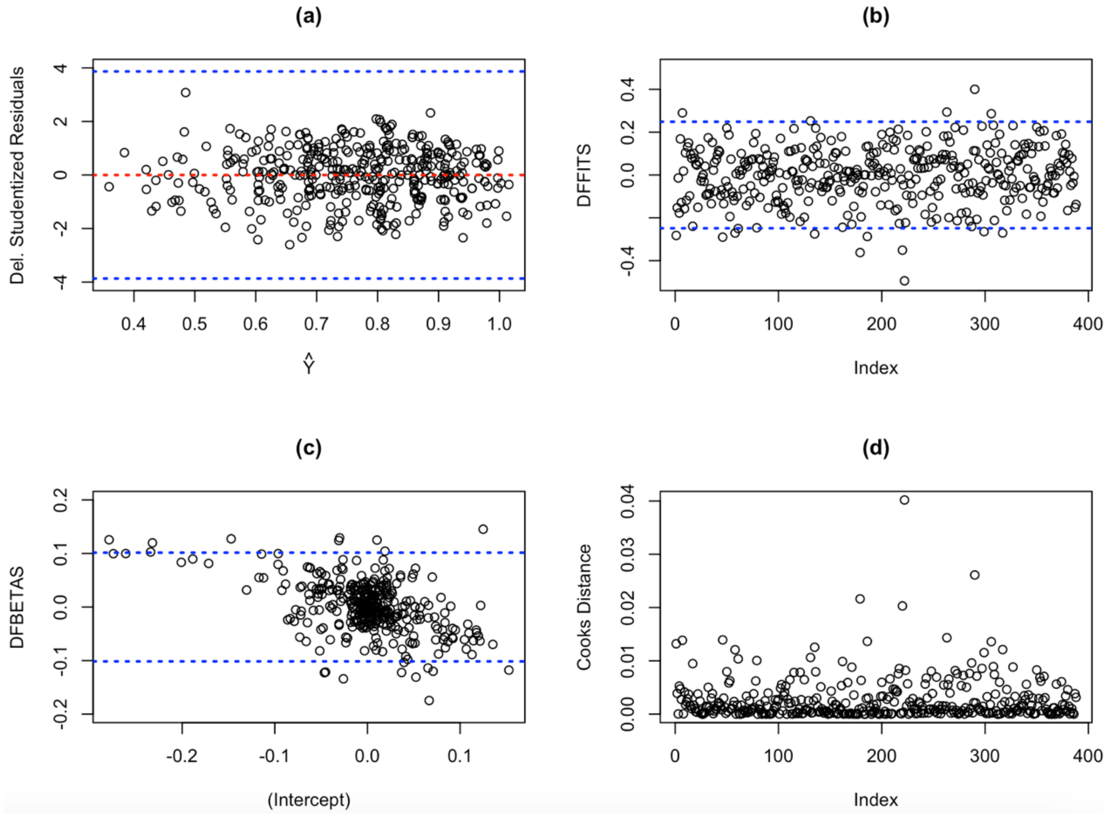

Figure 7: Detecting Outlying and Influential Observations

Observations 258 and 303 stuck out by being both outlying regarding their X-values and having a high influence regarding their DFFITS values. A fit of the final model without them showed no significant changes in the fit of the model. A fit of the final model without the 16 influential observations based on their DFFITS values showed similar results. A fit of the final model without outlying and influential observations based on DFFITS showed similar results. Indicating that the final model is robust to these observations. Table 8 shows the fit of the final model with and without the outlying and influential observations based on DFFITS values.

	Final Model	
	Full Dataset	**Reduced Dataset**
$\boldsymbol{b_0}$	0.8451	0.8727
$\boldsymbol{s(b_0)}$	0.0453	0.0438
$\boldsymbol{b_1}$	0.0164	0.0120
$\boldsymbol{s(b_1)}$	0.0086	0.0082
$\boldsymbol{b_2}$	0.0018	0.0013
$\boldsymbol{s(b_2)}$	0.0009	0.0009
$\boldsymbol{b_3}$	-0.0624	-0.0619
$\boldsymbol{s(b_3)}$	0.0094	0.0089
$\boldsymbol{b_4}$	-0.0414	-0.0436
$\boldsymbol{s(b_4)}$	0.0077	0.0075
$\boldsymbol{b_5}$	-0.1030	-0.1124
$\boldsymbol{s(b_5)}$	0.0256	0.0245
$\boldsymbol{SSE}$	21.6653	17.6834
$\boldsymbol{MSE}$	0.0567	0.0486
$\boldsymbol{R^2}$	0.2517	0.2847
$\boldsymbol{R_a^2}$	0.2419	0.2748

Table 8: Model 4 Fit to Full Dataset and Reduced Dataset

Residual Diagnostics

Figure 8 (a) shows the final model's absolute residuals vs. the fitted values and figure 8 (b) shows the final model's QQ-normal probability plot. Taking into consideration the potential outlying and influential observations that were not removed from the dataset, figure 8 (a) shows no issues with non-constant variance. Figure 8 (b) shows a little departure from linearity with a slight deviation in the upper tail, indicating the residuals follow a normal distribution. To further examine the normality of the residuals, the coefficient of correlation between the ordered residuals and expected values under normality was obtained. The coefficient was found to be 0.9964. While there is no critical value for a dataset this large, a critical value of a dataset with $n = 100$ and setting $\alpha = 0.05$, is 0.987[4]. The coefficient of correlation between the ordered residuals and expected values under normality is larger than the critical value based on a smaller dataset and $\alpha = 0.05$, supporting my original claim of normality.

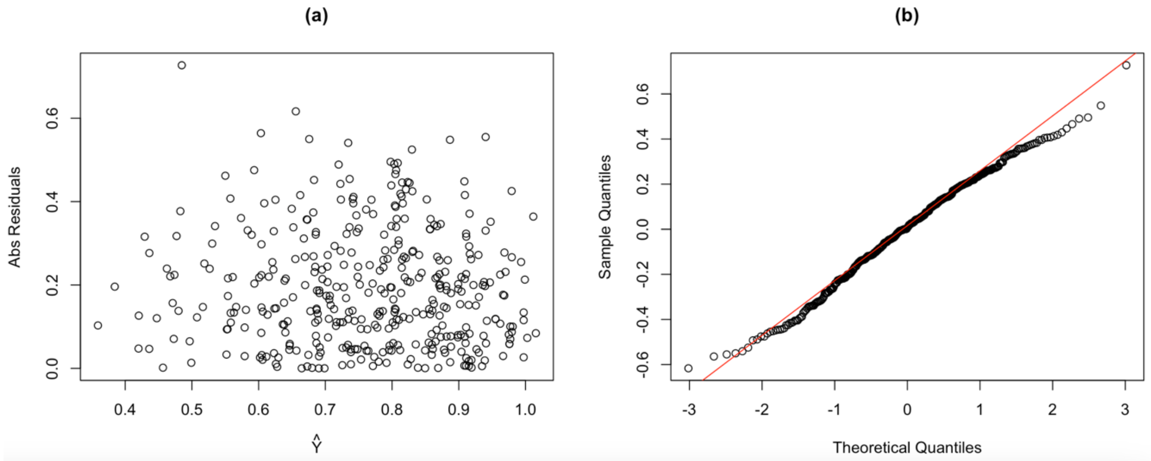

Figure 8: Final Model Residual Diagnostics

R-Code

```
#loading packages
library(readxl)
library(tidyverse)
library(caret)
library(asbio)
library(olsrr)
library(xtable)
library(shiny)
library(knitr)
library(DT)
require(scatterplot3d)
require(Hmisc)
require(rgl)
require(faraway)
library(car)
library(readxl)
library(leaps)
library(onewaytests)
library(dplyr)
library(psych)
library(ggplot2)
library(car)
#load in data
setwd("~/Documents/Masters/STAT 840/Final Proj")
data <- read.csv("Sleep_Efficiency.csv", header = TRUE)
###########################################################
```

```
#Exploring the data
nrow(data)# 452 rows
#Checking/removing rows with N/A values
SD<-na.omit(data)
nrow(SD)#388 rows
#################


#Creating variable for REM sleep time in hours
SD$REM.sleep.time <- ((SD$Sleep.duration*SD$Sleep.efficiency)*SD$REM.sleep.percentage)/60
#omitting uneeded variables:
SD <- subset(SD, select = -c(ID, Bedtime, Wakeup.time,REM.sleep.percentage,
                             Deep.sleep.percentage, Light.sleep.percentage,
                             Sleep.duration, Sleep.efficiency))
#transforming Gender: 1 for Fale, 0 for Female
SD$Gender <- ifelse(SD$Gender == "Male",1,0)
#transforming Smoking Status: 1 for Yes, 0 for No
SD$Smoking.status <- ifelse(SD$Smoking.status == 'Yes', 1,0)
###########################################################


#splitting the data for model training/validation
#first 195 observations will be used for model-building/training
training <- SD[1:195,]
#last 193 observations will be used for validation
validation <- SD[196:388,]
#applying Dependent Log transformation used in training dataset
validation$LogREM.sleep.time <- log(validation$REM.sleep.time)
validation <- subset(validation, select = -c(REM.sleep.time))
#############################################################


#Box plots of the variables
par(mfrow = c(3,2))
#Age
boxplot(training$Age, horizontal =TRUE,
        main = '(a) Box plot of Age (years)')
#Awakenings
boxplot(training$Awakenings, horizontal = TRUE,
        main='(b) Box plot of Awakenings')
#Caffeine Consumption
boxplot(training$Caffeine.consumption, horizontal = TRUE,
        main = '(c) Box plot Caffeine Consumption (mg)')
#Alcohol Consumption
boxplot(training$Alcohol.consumption, horizontal = TRUE,
        main = '(d) Box plot of Alcohol Consumption')
#Exercise Frequency
boxplot(training$Exercise.frequency, horizontal = TRUE,
        main = '(e) Box plot of Exercise Frequency')
#REM.sleep.time
```

```
boxplot(training$REM.sleep.time, horizontal = TRUE,
        main = '(f) Box plot of REM Sleep Time')

#Barcharts for Gender and Smoking.status as they're categorical
#Gender
training %>%
  ggplot(aes(x=Gender)) +
  geom_bar() +
  ggtitle('(a) Bar chart for Gender') +
  ylab('')

#Smoking
training %>%
  ggplot(aes(x=Smoking.status)) +
  geom_bar() +
  ggtitle('(b) Bar chart for Smoking Status') +
  xlab('Smoking Status')+
  ylab('')

##########


#preliminary model/full model
m1 <- lm(REM.sleep.time ~., data = training)
summary(m1)
#preliminary model: residuals vs. fitted values
par(mfrow = c(1,2))
plot(residuals(m1)~fitted(m1), main = '(a)', ylab = 'Residuals',
     xlab = expression(hat(Y)))
abline(h=0, col = 'red')
#issue of non-constant varaiance, consider a log transformation
qqnorm(residuals(m1), main = '(b)')
qqline(residuals(m1), col = 'red')

#Log Transformation of REM Sleep Time
training$LogREM.sleep.time <- log(training$REM.sleep.time)
training <- subset(training, select = -c(REM.sleep.time))
m1.t <- lm(LogREM.sleep.time~., data = training)
#transformed preliminary model: residuals vs. fitted values
par(mfrow = c(1,2))
plot(residuals(m1.t)~fitted(m1.t), main = '(a)', ylab = 'Residuals',
     xlab = expression(hat(Y)))
abline(h=0, col = 'red')

qqnorm(residuals(m1.t), main = '(b)')
qqline(residuals(m1.t), col = 'red')

#Partial Residual Plots of the transformed preliminary model
crPlots(m1.t,
        pch = 16, col.lines = c('blue', 'red'))
```

```r
#scatter-plot matrix and correlation matrix of the preliminary model
pairs.panels(training)

cor(training)
#VIF of the preliminary model
vif(m1.t)
#########


#model selection using automatic selection methods: best subsets and stepwiseforce in exercise frequency as it is the main variable of interest


#method 1: best subsets
summary(m1)
best.subs <- regsubsets(LogREM.sleep.time~., force.in = 7, data = training)
(best.subs.sum <- summary(best.subs))


#Screening further to select the best model based on: adjusted R^2, CP, BIC, and RSS values


#adjusted R^2: bigger is better
#fullmodel/model 6: 0.2868291
#model 5: 0.2857763
#model 4: 0.2823637
best.subs.sum$adjr2


#Mallow's CP: low CP value and CP=P is best:
#model 5: CP = 7.277507, p = 7, bias = 0.277507
#full model/model 6: Cp = 8, p = 8, bias = 0
best.subs.sum$cp


#BIC: smaller is better:
#model 2: --37.15807
#model 3: -38.72200
#model 4: -38.15318
best.subs.sum$bic


#RSS: smaller is better:
#model 4: 9.965151
#model 5: 9.865288
#fullmodel/model 6: 9.798349
best.subs.sum$rss
#Three candidate models from best subsets: model 4, model 5, and fullmodel/model 6


#Method 2: Automatic selection method - stepwise
step.subs <- regsubsets(LogREM.sleep.time~., force.in = 7, data = training, method = "seqrep")
(step.subs.sum <- summary(step.subs))
```

```
#Screening further to select the best model based on: adjusted R^2, CP, BIC, and RSS values


#Adjusted R^2: bigger is better:
#full model/model 6: 0.2868291
#model 5: 0.2857763
step.subs.sum$adjr2


#Mallow's CP: low CP value and CP=P is best:
#model 5: CP = 7.277507, p = 7, bias = 0.277507
#fullmodel/model 6: Cp = 8, p = 8, bias = 0
step.subs.sum$cp


#BIC: smaller is better:
#model 3: -38.72200
step.subs.sum$bic


#RSS: smaller is better:
#model 5: 9.865288
#fullmodel/model 6: 9.798349
step.subs.sum$rss
#Two candidate models from stepwise method: model 5 and full model/model 6.

#Model 5 and  was robust to both methods. Therefore, will be moving with the full model/model 6, model
5, and model 4 from best subsets to the next step of validation via data-splitting


fullmodel <- lm(LogREM.sleep.time~., data = training)
#model 5 from best subsets
m5 <- lm(LogREM.sleep.time~Exercise.frequency + Age + Awakenings +
            Caffeine.consumption + Alcohol.consumption +
            Smoking.status, data = training)
#model 4 from best subsets
m4 <- lm(LogREM.sleep.time~Exercise.frequency + Age + Awakenings +
          + Alcohol.consumption + Smoking.status, data = training)
#######################################################################


#Validation of the fullmodel/model 6
anova(fullmodel)$'Sum Sq'[8] #SSE: 9.798349
anova(fullmodel)$'Mean Sq'[8]#MSE: 0.05239759
press(fullmodel)#PRESS: 10.67473

fullmodel.valid <-lm(LogREM.sleep.time~., data = validation)
summary(fullmodel)
summary(fullmodel.valid)
anova(fullmodel.valid)$'Sum Sq'[8]#SSE.v: 11.22664
anova(fullmodel.valid)$'Mean Sq'[8]#MSE.v: 0.06068455
press(fullmodel.valid)#PRESS.v: 12.21069


#calculate MSPR and compare to MSE for training
#MSPR
#fit <- predict(traininig model, newdata = testing/validation data)
#pe <- testing/validation data$dependent variable - fit
#(MSPR <- (t(pe)%*%pe)/length(pe))
fit <- predict(fullmodel, newdata = validation)
pe <- validation$LogREM.sleep.time-fit
(MSPR <- (t(pe)%*%pe)/length(pe)) #0.06276787
#MSE for training set
anova(fullmodel)$'Mean Sq'[8]#0.05239759
#MSPR/MSE difference:0.01037028
################################


#Validation of the model 5
anova(m5)$'Sum Sq'[7]#SSE; 9.865288
anova(m5)$'Mean Sq'[7]#MSE: 0.05247494
press(m5)#PRESS: 10.64064

m5.valid <-lm(LogREM.sleep.time~Exercise.frequency + Age + Awakenings +
                  Caffeine.consumption + Alcohol.consumption +
                  Smoking.status, data = validation)
summary(m5)
summary(m5.valid)
anova(m5.valid)$'Sum Sq'[7]#SSE.v: 53.82344
anova(m5.valid)$'Mean Sq'[7]#MSE.v: 0.2886186
press(m5.valid)#PRESS.v: 12.15274
#calculate MSPR and compare to MSE for training
#MSPR
fit <- predict(m5, newdata = validation)
pe <- validation$LogREM.sleep.time-fit
(MSPR <- (t(pe)%*%pe)/length(pe))#0.06274656
#MSE for training set
anova(m5)$'Mean Sq'[7]#0.05247494
#MSPR/MSE difference: 0.01027162
#################################


#Validation of the stepwise model 4
anova(m4)$'Sum Sq'[6]#SSE: 9.965151
anova(m4)$'Mean Sq'[6]#MSE: 0.05272567
press(m4)#PRESS: 10.57748

m4.valid <- lm(LogREM.sleep.time~Exercise.frequency + Age + Awakenings +
      + Alcohol.consumption + Smoking.status, data = validation)
summary(m4)
summary(m4.valid)
```

```
anova(m4.valid)$'Sum Sq'[6]#SSE.v: 11.28781
anova(m4.valid)$'Mean Sq'[6]#MSE.v: 0.06036259
press(m4.valid)#PRESS.v: 12.03004
#MSPR
fit <- predict(m4, newdata = validation)
pe <- validation$LogREM.sleep.time-fit
(MSPR <- (t(pe)%*%pe)/length(pe))#0.06274979
#MSE for training set
anova(m4)$'Mean Sq'[6]#0.05272567
#MSPR/MSE difference: 0.01002412
################################


#Fitting the final model to the full data
#adding transformed dependent variable to full data
SD$LogREM.sleep.time <- log(SD$REM.sleep.time)
SD <- subset(SD, select = -c(REM.sleep.time))

m.final <- lm(LogREM.sleep.time~Exercise.frequency + Age + Awakenings +
        + Alcohol.consumption + Smoking.status, data = SD)
summary(m.final)
################


#Detecting outliers and influential observations in the final model


#outlying Y observations
#Bonferroni Critical Value for outlying Y-observations
n<-nrow(SD) #388 observations
p <- 6 #6betas (B0, B1, B2, B3, B4, B5)
alpha <- 0.05
#Bonferroni simultaneous test procedure: qt(1-(alpha/2n), n-p-1)
qt(1-alpha/(2*n), n-p-1)
#if |t-value| >= qt(1-(alpha/2*n), n-p-1), then the observation is an outlier,otherwise, not an outlier
which(abs(rstandard(m.final))>qt(1-alpha/(2*n), n-p-1))


#deleted studentized residuals vs fitted
par(mfrow = c(2,2))
plot(rstandard(m.final)~fitted(m.final),
   ylab = 'Del. Studentized Residuals',
   xlab = expression(hat(Y)),
   main = "(a)",
   ylim = c(-4,4))
abline(h=0, lty = 3, lwd = 2, col = 'red')
abline(h=3.868316, lty = 3, lwd = 2, col = 'blue') #values taken from BF test procedure
abline(h=-3.868316, lty = 3, lwd = 2, col = 'blue')


#Detecting outlying X observations using hat matrix
#Rule of thumb: leverage value greater then (2*p)/n are considered large
```

```
which(hatvalues(m.final)>((2*p)/n))
hatvalues(m.final)[103]#observation 118, different number due to ommitting NA values
hatvalues(m.final)[131]#observation 155
hatvalues(m.final)[222]#observation 258
hatvalues(m.final)[263]#observation 303
hatvalues(m.final)[328]#observation 379
#4 observations w/ high leverage were identified. Need to investigate how influential these observations are in the fitting of the regression function

#Detecting influential cases using DFFITS,DFBETAS, Cooks distance

DFFITS
#Rule of thumb: DFFITS > 2*sqrt(p/n) for large datasets
CT <- 2*sqrt(p/n)
plot(dffits(m.final), ylim = c(-.5,.5),
     ylab = 'DFFITS',
     main = '(b)')
abline(h=CT, lty = 3, lwd = 2, col = 'blue')
abline(h=-CT, lty = 3, lwd = 2, col = 'blue')
which(abs(dffits(m.final))>CT)
#16 influential cases identified (high leverage observations 258 and 303 are influential based on DFFITS and outlying in regards to its X value)

#DFBETAS
#rule of thumb: DFBETASS > 2/sqrt(n)
CV <- 2/sqrt(n)
plot(dfbetas(m.final), ylim = c(-.2,.2),
     ylab = 'DFBETAS',
     main = '(c)')
abline(h=CV, lty = 3, lwd = 2, col = 'blue')
abline(h=-CV, lty = 3, lwd = 2, col = 'blue')
which(abs(dfbetas(m.final))>CV)
nrow(which(abs(dfbetas(m.final))>CV))
#143 betas are considered influential

#Cooks Distance
#rule: if the percentile is less than about 10-20%, there
#is little apparent influence. If the percentile value is near
#50%, the observation is considered influential
plot(cooks.distance(m.final),
     ylab = 'Cooks Distance',
     main = '(d)')

#compare Cook's Distance to F(p, n-p) pg.403
summary(m.final)
q <- pf(cooks.distance(m.final),6,388-6)
which(q>.1)
#no influential observations based on cook's distance are observed

#removing observations 258 and 303
reduced.data <- SD[-c(222,263),]
```

```
nrow(reduced.data)
summary(lm(LogREM.sleep.time~Exercise.frequency + Age + Awakenings +
    + Alcohol.consumption + Smoking.status, data = reduced.data))
#no significant changes in the betas or the fit of the model


#removing all observations with high leverage: 118,155,258,303,379
reduced.data <- SD[-c(103,131,222,263,328),]
nrow(reduced.data)
summary(lm(LogREM.sleep.time~Exercise.frequency + Age + Awakenings +
          + Alcohol.consumption + Smoking.status, data = reduced.data))
summary(m.final)
#no significant changes in the betas or the fit of the model
#removing values based on DFFITS
reduced.data <- SD[-c(1,7,50,136,179,186,216,220,222,263,271,290,297,308,357),]
nrow(reduced.data)
summary(lm(LogREM.sleep.time~Exercise.frequency + Age + Awakenings +
          + Alcohol.consumption + Smoking.status, data = reduced.data))
summary(m.final)
#no significant changes in the betas or the fit of the model


#removing values based on high leverage and DFFITS
reduced.data <- SD[-c(1,7,50,103,131,222,263,328,136,179,186,216,220,222,263,271,290,297,308,357),]
nrow(reduced.data)
m.r<-lm(LogREM.sleep.time~Exercise.frequency + Age + Awakenings +
    + Alcohol.consumption + Smoking.status, data = reduced.data)
summary(m.r)
summary(m.final)

#MSE/SSE of the final model w/full data
anova(m.final)$'Mean Sq'[6]#MSE: 0.05671533
anova(m.final)$'Sum Sq'[6]#SSE: 21.66526
#MSE/SSE of the final model w/reduced datta
anova(m.r)$'Mean Sq'[6]#MSE: 0.04858081
anova(m.r)$'Sum Sq'[6]#SSE: 17.68342

#Final model residual diagnostics
par(mfrow = c(1,2))
#residuals vs. fitted
plot(abs(residuals(m.final))~predict(m.final),
    xlab = expression(hat(Y)),ylab = "Abs Residuals",
    main = '(a)')
#normality plot
qqnorm(residuals(m.final), main = '(b)')
qqline(residuals(m.final), col= 'red')

#VIF of the final model
vif(m.final)
```

```
#coefficient of correlation between ordered resid. vs. expected values under normality of the final model
ols_test_correlation(m.final)\


#Percent of variation in the final model explained by each variable

#Variable of interest: Exercise Frequency
round(partial.R2(lm(LogREM.sleep.time~Age + Awakenings + Alcohol.consumption + Smoking.status, data = SD),m.final)*100,2)
#Age
round(partial.R2(lm(LogREM.sleep.time~Exercise.frequency + Awakenings + Alcohol.consumption + Smoking.status, data = SD),m.final)*100,2)
#Awakenings
round(partial.R2(lm(LogREM.sleep.time~Exercise.frequency + Age + Alcohol.consumption + Smoking.status, data = SD),m.final)*100,2)
#Alcohol consumption
round(partial.R2(lm(LogREM.sleep.time~Exercise.frequency + Age+ Awakenings + Smoking.status, data = SD),m.final)*100,2)
#Smoking Status
round(partial.R2(lm(LogREM.sleep.time~Exercise.frequency + Age+ Awakenings + Alcohol.consumption, data = SD),m.final)*100,2)
```