Exploring Influences on Sleep Quality and Quantity

Math448 Final Project Report, Spring 2017

Paul Previde

Abstract

The ability to get a full night of high-quality sleep on a daily basis is crucial to a person’s health and well-being. Good sleep hygiene requires both that person sleep long enough, but also that the sleep be of high quality (*i.e.*, uninterrupted, continuous, and in an environment conducive to falling and staying asleep). However, there are many physiological, environmental, and behavioral factors that can interfere with the ability of many people to get both high-quality and sufficient sleep, and the deficiency causes a loss of well-being and health for the afflicted individuals and imposes tremendous economic costs on the entire community.

This project uses a vast repository of health-related data, maintained by the Centers for Disease Control in the National Health and Nutrition Information Survey, to explore the quantity and quality of patients’ sleep. Specifically, the effects of a number of physiological, demographic, and behavioral factors on patients’ sleep were examined through regression and classification analyses. These studies provided evidence of the significance of advancing age and obesity as possible risk factors in the incidence of sleep difficulties. Even recognizing the fact that sleep hygiene is subject to complicated psychological, physiological, and environmental factors, the techniques described in this report were still useful for highlighting some potential areas for further study. In addition, this project provided a valuable opportunity to apply the statistical learning techniques covered in this course to a real, and widely-cited, data set, and as such provided a valuable and realistic data analysis experience.

I. Introduction

The ability to get a full, high-quality night of sleep on a daily basis is crucial to a person's well-being as well as to a person's ability to be productive and effective at work and in all other endeavors. The absence of this ability can be debilitating to a person in a variety of ways. Indeed, the Centers for Disease Control and Prevention has cited insufficient or poor-quality sleep as a public health concern [1]. Unfortunately, sleep problems are very common, and an estimated 50-70 million Americans report trouble sleeping each year [2], and this tally doesn't take into account those individuals who have sleep problems but never report them, choosing instead to simply endure them. Sleep problems are also costly: the total costs associated with insomnia alone (one of many potential sleep disorders that can affect someone) is estimated to exceed $100 billion in terms of the expenses associated with medical involvement and lost workplace productivity [3]. Simply put, sleep problems represent an important and expensive issue that deserves careful study and substantial effort at mitigation and/or management.

This project explores some health- and lifestyle-related factors that may be associated with reduction in sleep quantity or quality. This project recognizes that sleep deficits can manifest as subjective (*i.e.*, poor quality of sleep) and/or quantitative (*i.e.*, not enough hours of sleep). In terms of sleep quantity, getting less than seven hours of sleep consistently has been considered inadequate [5, 6]. In particular, this work explores the effects, if any, of various demographic, physiological, and behavioral factors on how much and how well a person sleeps, in an effort to better understand some of the basic influences on sleep. As such, this study represents an exercise in supervised learning, with a focus in inference rather than prediction

II. Data and Methods

1. Dataset

This project utilizes the extensive National Health and Nutrition Examination Survey ("NHANES") dataset, which is maintained and overseen by the Centers for Disease Control and Prevention headquartered in Atlanta, Georgia. The 2013-14 NHANES data set was obtained from the Kaggle web site [8] and consists of information, measurements, and questionnaire and interview responses from thousands of patients randomly chosen from across the country every year.[1]

The NHANES data set consists of six separate data tables, each in a comma-separated values file. Each data table has thousands of patient observations, with each observation containing information on between 13 and 952 measurements or responses from the patient. Significantly, each patient observation includes a unique patient identifier, called "SEQN" for "sequence number", that allows a patient's observations from across multiple data tables to be combined for analysis. As such, the NHANES dataset offers a valuable look at a wide variety of properties, measurements, and questionnaire responses for the same patient. Table 1 below presents the number of patients and the number of predictors (excluding the patient identifier) in each data table.

[1] The participation of each patient is voluntary, but patients are paid for their time and are not required to transport themselves to a health facility – NHANES personnel can go to the patient. Thus, many people in rural areas get the benefit of health care access because of the NHANES program, yet another positive benefit of the program.

data table	number of predictors*	number of patients
Questionnaire	952	10,175
Examinations	223	9,813
Demographics	46	10,175
Dietary	167	9,813
Laboratory	423	9,813
Medications	13	20,194
Total	1,824	10,175 distinct patients

Table 1: The data tables of the NHANES data set.

The combined data set contains a total of 10,175 distinct patients and 1,824 distinct predictors, excluding the patient identifier column found in each data table. This study required information from only three of the data tables: Demographics, Examination, and Questionnaire. To facilitate analysis and the expeditious handling of missing values, these three tables were combined into a single data frame using the R merge() function.

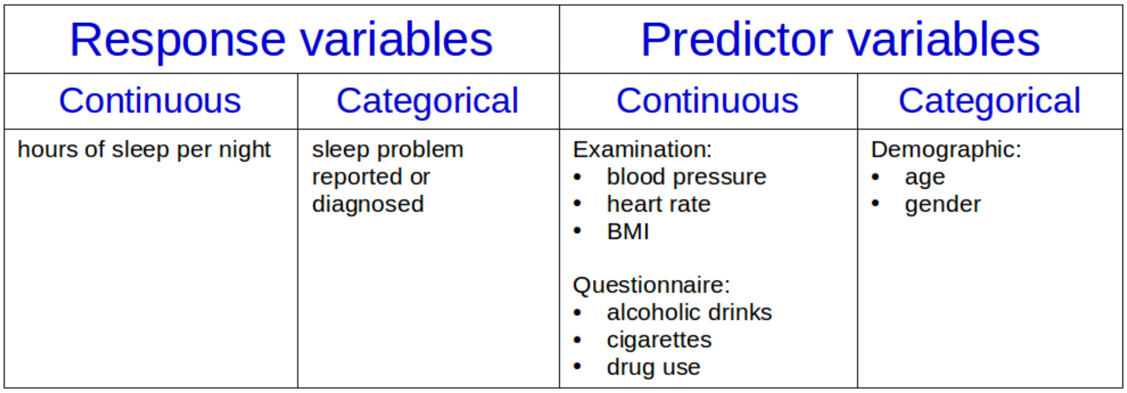
2. Missing Values

The combined dataset suffered from large proportion of missing values. This problem is not surprising, given the sheer scope of the data being collected. In fact, not for a single patient is there information for every predictor; in other words, there are no complete records anywhere in the full dataset. Table 2 presents some information about the missing values problem.

number of complete patient records in the full dataset	0
total number of cells in the dataset	12,433,850
number of cells with missing values	8,593,662 (69%)
mean number of missing values per patient	844 ± 126 out of 1,824

Table 2: Description of the missing values problem in the NHANES dataset

Various strategies were used to handle the missing values in the dataset. Most importantly, because the existing data is valuable, rows with missing values were not removed until absolutely necessary. For example, some functions could not be performed with missing values, but only when those functions were being performed were the rows with missing values removed. As such, as much data as possible was used for regression, classification, etc. Likewise, no imputation of missing values was performed; it was not desired to try to estimate what a patient’s measurement or questionnaire response would be, out of concern that the resulting estimates would essentially be incorrect guesswork rather than accurate information. Lastly, wherever possible (*e.g.*, when performing numerical calculations within columns, such as calculating the mean or the maximum value), the option “na.rm = TRUE” was used rather than removing the row. The combined effect of these steps was to ensure that all available patient data was used to study the influences on sleep quantity and quality, with none of that data removed just as an expediency in dealing with missing values. After removing unneeded variables, the data frame had three response variables and eight predictors (Figure 1).



Response variables		Predictor variables	
Continuous	Categorical	Continuous	Categorical
hours of sleep per night	sleep problem reported or diagnosed	Examination: • blood pressure • heart rate • BMI Questionnaire: • alcoholic drinks • cigarettes • drug use	Demographic: • age • gender

Figure 1: The predictors and response variables after data pre-processing.

The first response variable is quantitative and indicates the average number of hours of sleep per night reported by the respondent. The second and third are categorical and indicate whether the patient is having any sleeping trouble or whether the patient has sought medical help for a sleeping problem, respectively. Figure 1 also shows the predictor variables and the data tables from which they were collected.

3. Statistical Learning Techniques Applied to the Data

Because the response variables include both categorical and quantitative variables, both the regression and the classification techniques covered in the course could be brought to bear on the data set. These techniques included linear and polynomial regression as well as LDA, QDA, bagging and random forest classification. This study also included some basic visualization techniques, such as histograms and bar charts, that provided a surprisingly large number of insights about the data.

III. Results and Discussion

Before turning to the regression and classification techniques that were the main subjects of the course, the NHANES data was first subjected to some basic visualization techniques such as histograms.

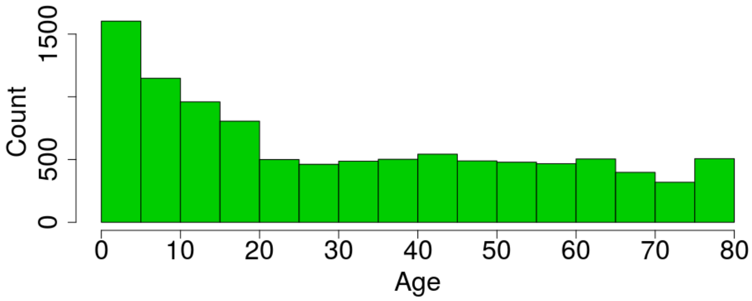


Figure 2: Histogram of the ages of respondents in the NHANES data set.

The histogram of the ages of the respondents for the NHANES dataset (Figure 2) demonstrates that the respondents' ages range from young children to seniors, with a fairly balanced age distribution except that there is a larger proportion of respondents under age 20. The gender ratio of the respondents was fairly balanced as well, with 5,003 males (49.2%) and 5,172 females (50.8%).

A histogram of the hours of sleep per night reported by the respondent pool (Figure 3) provides evidence that insufficient sleep is a common problem. A large proportion of the respondent pool reported less than seven hours of sleep, with a significant portion getting even less than six hours.

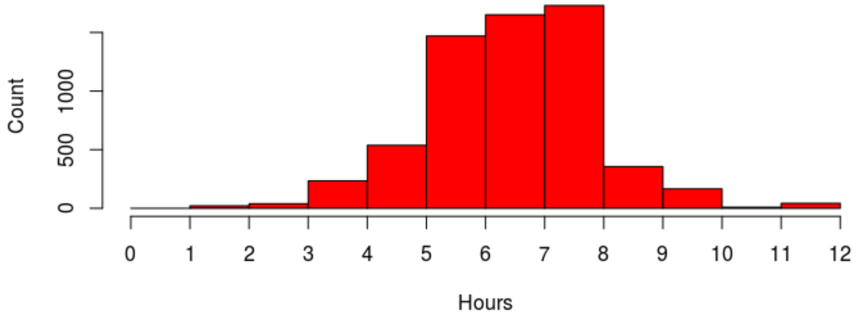


Figure 3: Histogram of the hours per night of sleep, on average, reported by respondents.

Likewise, in terms of sleep quality, 1,548 (24.7%) out of the 6,264 respondents for whom data on sleep quality was available reported sleep problems.

However, despite the prevalence of sleep issues, most cases go unreported and undiagnosed: many people who experience difficulty sleeping never seek a diagnosis, and instead, simply choose to live with insufficient or poor-quality sleep. Figure 4 demonstrates this phenomenon.

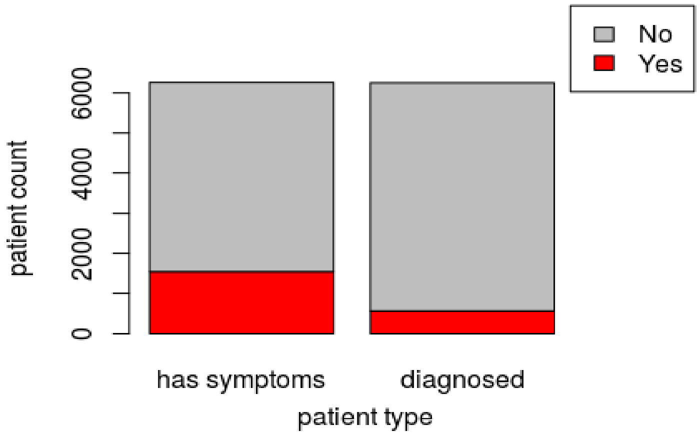


Figure 4: Bar chart showing the difference between the proportion of respondents who indicated sleep problems (left bar) and the proportion of respondents who sought a diagnosis for their condition (right bar).

Figure 5 below indicates the breakdown of sleeping problems by age, and reveals that at the largest proportion of people who reported sleeping problems was for the groups of people over 40.

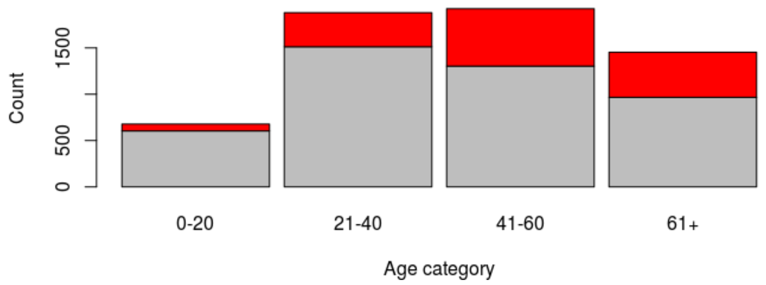


Figure 5: Proportions of people in different age ranges who reported sleeping problems.

Examination of the breakdown of sleeping problems by gender (Figure 6) indicates that females have a somewhat higher proportion of respondents who reported sleeping problems than males.

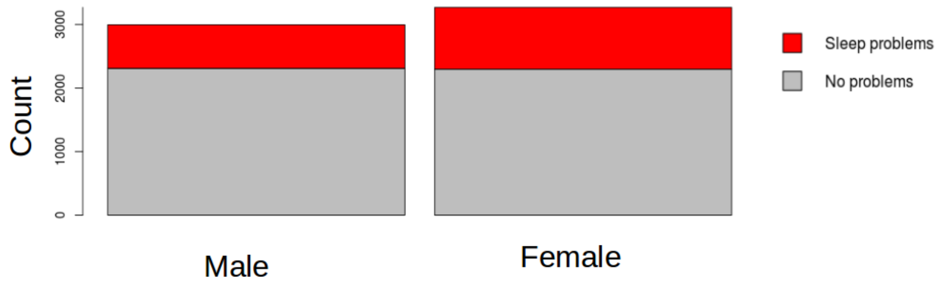


Figure 6: Proportions of males and females who reported sleeping problems.

Figure 7 presents the breakdown of sleep problems by body mass index category. People with obesity have a higher proportion of individuals who reported trouble sleeping compared with other categories.

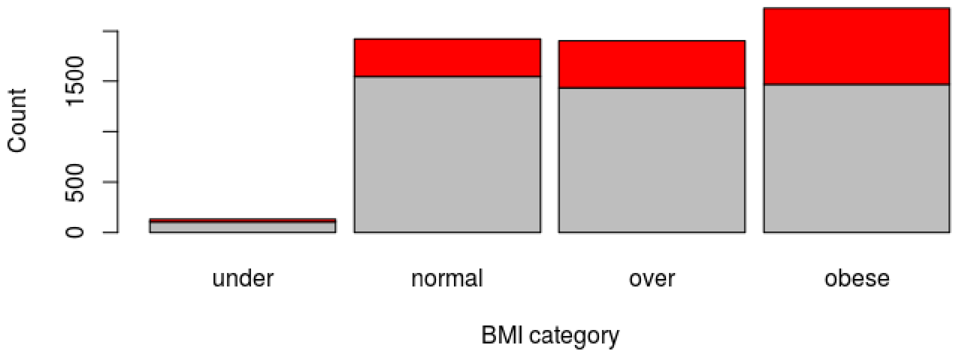
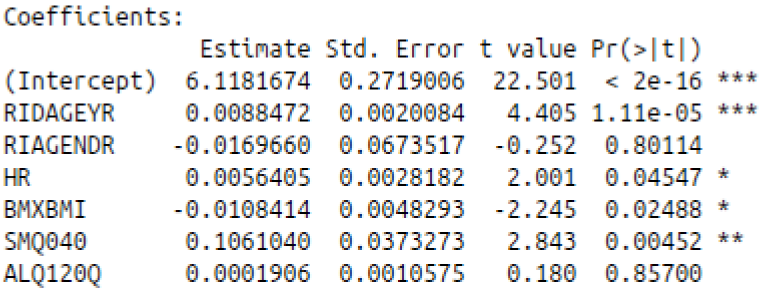


Figure 7: Proportions of people from different BMI categories who reported sleeping problems.

2. Quantity of Sleep

The response variable corresponding to quantity of sleep per night was explored using linear, polynomial, ridge, and lasso regression. Linear regression of the average hours of sleep per night onto predictor variables for age, gender, heart rate, body mass index, smoking frequency, and number of alcoholic drinks per week was performed. The results are shown in Figure 8.

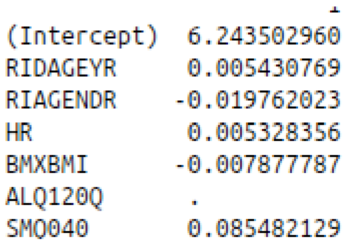


```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  6.1181674  0.2719006  22.501  < 2e-16 ***
RIDAGEYR     0.0088472  0.0020084   4.405 1.11e-05 ***
RIAGENDR    -0.0169660  0.0673517  -0.252  0.80114    
HR           0.0056405  0.0028182   2.001  0.04547 *  
BMXBMI      -0.0108414  0.0048293  -2.245  0.02488 *  
SMQ040       0.1061040  0.0373273   2.843  0.00452 ** 
ALQ120Q      0.0001906  0.0010575   0.180  0.85700    
```

Figure 8: Linear regression of average hours of sleep per night onto the predictors.

Linear regression reveals evidence of an association between sleep quantity per night and the predictors associated with body mass index, heart rate, age, and smoking frequency. On the other hand, the regression did not yield evidence of any association between sleep quantity and gender or alcohol use.

Ridge and lasso regression were then performed, and lasso regression was of particular interest to see if any of the coefficients would be reduced to zero (a capability which ridge regression does not offer). The selection of λ, the penalty parameter for the number of predictors in each regression, was determined by using a grid of 10,000 evenly-spaced candidate values from 10^{10} to 10^{-4}, in cross-validation. Lasso regression reduced one of the model coefficients (alcohol consumption) to zero at the chosen value of λ, as shown in Figure 9 below.

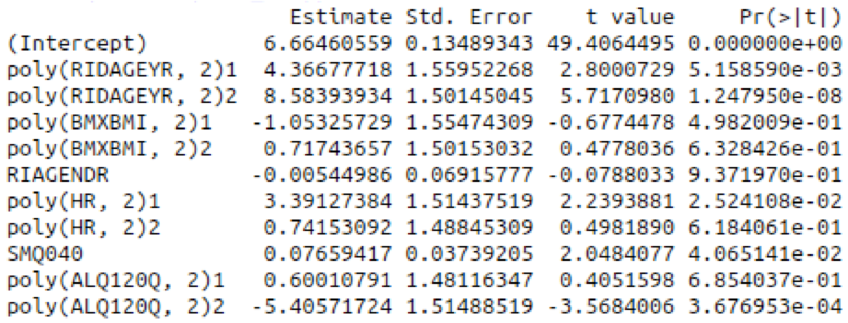


```
                       1
(Intercept)  6.243502960
RIDAGEYR     0.005430769
RIAGENDR    -0.019762023
HR           0.005328356
BMXBMI      -0.007877787
ALQ120Q      .          
SMQ040       0.085482129
```

Figure 9: Reduction to zero of one of the model coefficients by lasso.

Reduction of the alcohol coefficient to zero is consistent with the finding of linear regression that there was no evidence of an association between alcohol and the response variable sleep quantity.

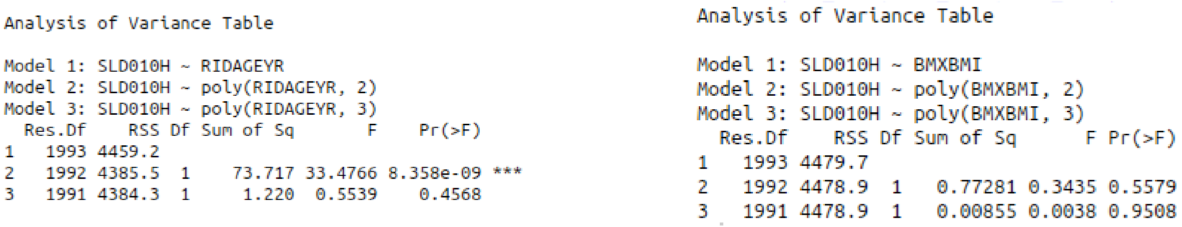
Polynomial regression was then performed with the same response and predictor variables as was used for the other regression types, in order to explore the possibility that the relationship between sleep quantity and the predictors is not simply linear. The polynomial regression checked up to the quadratic term associated with each predictor (*i.e*, each term and its square) were variables in the polynomial regression model, except for the categorical variables gender and smoking. Figure 10 below shows the terms, their coefficients, and their p-values in the polynomial regression model.



```
                     Estimate Std. Error    t value     Pr(>|t|)
(Intercept)        6.66460559 0.13489343 49.4064495 0.000000e+00
poly(RIDAGEYR, 2)1  4.36677718 1.55952268  2.8000729 5.158590e-03
poly(RIDAGEYR, 2)2  8.58393934 1.50145045  5.7170980 1.247950e-08
poly(BMXBMI, 2)1   -1.05325729 1.55474309 -0.6774478 4.982009e-01
poly(BMXBMI, 2)2    0.71743657 1.50153032  0.4778036 6.328426e-01
RIAGENDR           -0.00544986 0.06915777 -0.0788033 9.371970e-01
poly(HR, 2)1        3.39127384 1.51437519  2.2393881 2.524108e-02
poly(HR, 2)2        0.74153092 1.48845309  0.4981890 6.184061e-01
SMQ040              0.07659417 0.03739205  2.0484077 4.065141e-02
poly(ALQ120Q, 2)1   0.60010791 1.48116347  0.4051598 6.854037e-01
poly(ALQ120Q, 2)2  -5.40571724 1.51488519 -3.5684006 3.676953e-04
```

Figure 10: Polynomial regression of sleep quantity onto the predictors and their quadratic terms.

To further explore the importance of the age and gender terms in polynomial regression, analysis of variance was performed as follows. First, a polynomial regression was conducted of sleep quantity onto the age predictor, its square term, then its cubic term. Likewise, polynomial regression onto the predictor BMI, its square term, and its cubic term was performed as well. Analysis of variance was used to see if the higher-order regression models were justified. The results are shown in Figures 11(a) and 11(b) below.



```
Analysis of Variance Table

Model 1: SLD010H ~ RIDAGEYR
Model 2: SLD010H ~ poly(RIDAGEYR, 2)
Model 3: SLD010H ~ poly(RIDAGEYR, 3)
  Res.Df    RSS Df Sum of Sq       F    Pr(>F)
1   1993 4459.2
2   1992 4385.5  1    73.717 33.4766 8.358e-09 ***
3   1991 4384.3  1     1.220  0.5539    0.4568
```

(a)

```
Analysis of Variance Table

Model 1: SLD010H ~ BMXBMI
Model 2: SLD010H ~ poly(BMXBMI, 2)
Model 3: SLD010H ~ poly(BMXBMI, 3)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1   1993 4479.7
2   1992 4478.9  1   0.77281 0.3435 0.5579
3   1991 4478.9  1   0.00855 0.0038 0.9508
```

(b)

Figure 11. (a) ANOVA for regression of sleep quantity onto age. (b) ANOVA for regression of sleep quantity onto BMI.

The ANOVA results of Figure 11(a) show that the quadratic regression model of age is justified, as shown by the very low p-value associated with the second-degree polynomial. On the other hand, Figure 11(b) shows that there is insufficient evidence to justify a quadratic or cubic polynomial as a model for the relationship between sleep quantity and BMI. These findings suggest a more complicated relationship between sleep quantity and age than was suggested by linear regression.

The test MSE of linear, ridge, lasso, and polynomial regression are shown in Table 3 below.

Regression	Test MSE
Linear	2.247
Ridge	2.343
Lasso	2.266
Polynomial	2.194

Table 3: Comparison of the test MSE results of all regression techniques of sleep quantity.

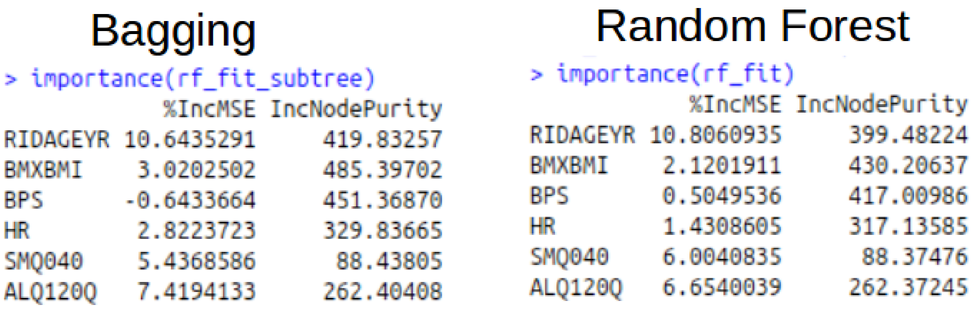
Table 3 reveals the interesting finding that the test MSE was lowest for the polynomial regression model, suggesting that the simpler models associated with linear, ridge, and lasso techniques are insufficient to accurately describe the relationship between sleep quantity and those predictors. Table 3 can also be viewed in light of the bias-variance trade-off: while less flexible models like linear regression can reduce variance, they may also suffer from more bias resulting from the error introduced by attempting to describe a complex reality with an overly simple model.

Lastly, bagging and random forest techniques were used to develop predictive models for sleep quantity based on the predictors associated with age, body mass index, blood pressure, heart rate, smoking, and alcohol use. The test MSE results are shown in Table 4 below.

Technique	Test MSE	Most important variables
Bagging	2.399	age, alcohol
RF	2.343	age, alcohol

Table 4: Test MSE of bagging and random forest prediction models for sleep quantity.

The test MSE was somewhat worse for bagging and random forest techniques than for the regression models. Both bagging and random forest determined that the most influential predictors on the prediction models were age and alcohol, so these techniques provided another line of evidence in support of an association between age and sleep quantity (Figure 12).



Bagging

```
> importance(rf_fit_subtree)
            %IncMSE IncNodePurity
RIDAGEYR 10.6435291     419.83257
BMXBMI    3.0202502     485.39702
BPS      -0.6433664     451.36870
HR        2.8223723     329.83665
SMQ040    5.4368586      88.43805
ALQ120Q   7.4194133     262.40408
```

Random Forest

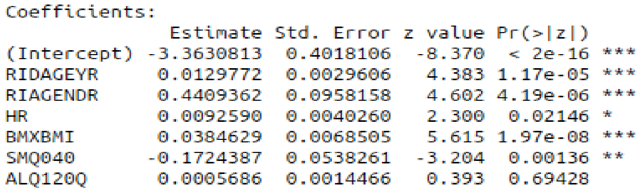
```
> importance(rf_fit)
            %IncMSE IncNodePurity
RIDAGEYR 10.8060935     399.48224
BMXBMI    2.1201911     430.20637
BPS       0.5049536     417.00986
HR        1.4308605     317.13585
SMQ040    6.0040835      88.37476
ALQ120Q   6.6540039     262.37245
```

Figure 12: The most influential predictors in the bagging and random forest models.

3. Classification Techniques

The categorical response variable associated with sleep quality (as reflected in whether a patient indicated sleep problems or sought professional help for sleep problems) was studied using classification techniques.

First, logistic regression was performed of sleep quality onto the predictors associated with age, gender, heart rate, body mass index, smoking, and alcohol use (Figure 13). Logistic regression provided evidence of an association between sleep quality and the predictors associated with age, gender, heart rate, body mass index, and heart rate. Logistic regression therefore provided corroboration for many of the findings in linear regression: specifically, both techniques provided evidence of an association between body mass index, age, heart rate, and smoking.



```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.3630813  0.4018106  -8.370  < 2e-16 ***
RIDAGEYR     0.0129772  0.0029606   4.383 1.17e-05 ***
RIAGENDR     0.4409362  0.0958158   4.602 4.19e-06 ***
HR           0.0092590  0.0040260   2.300  0.02146 *
BMXBMI       0.0384629  0.0068505   5.615 1.97e-08 ***
SMQ040      -0.1724387  0.0538261  -3.204  0.00136 **
ALQ120Q      0.0005686  0.0014466   0.393  0.69428
```

Figure 13: Logistic regression of sleep quality onto the predictors.

Next, linear and quadratic discriminant analysis (LDA and QDA respectively) were performed to discern how well a classifier could be developed to predict whether a patient would have sleep problems based on the predictors age, gender, blood pressure, heart rate, body mass index, smoking, and alcohol use. The best misclassification rates were determined via cross validation, and were determined to be 35.8% for LDA and 35.6% for QDA. Thus, neither was significantly better than the other.The classification success rates were determined via cross validation.

The effectiveness of LDA and QDA were compared to that of bagging and random forest classifiers for the same response and predictor variables, again by cross validation. Bagging and random forest had somewhat higher misclassification rates than LDA or QDA, but revealed that BMI, blood pressure, and age are the top three most influential predictors (Table 5).

Technique	Misclassification rate	Most important variables
Bagging	36.7%	BMI, blood pressure, age
RF	36.8%	BMI, age, blood pressure

Table 5: Bagging and random forest classifier performance and most important variables.

V. Conclusion

This project offered a valuable opportunity to practice many of the techniques covered in this course, and also provided some valuable lessons regarding statistical learning and data analysis. First and foremost, the project demonstrated that there is no “one size fits all” technique that clearly outperforms all others for a certain type of problem. Patience and a methodical approach to evaluating the performance of each technique are crucial. Some techniques might perform better than others in some circumstances, but not in others. Each technique’s results should be viewed as a source of evidence, not as a definite answer to the research question.

This project provided an example of these considerations. Several techniques provided corroborating evidence of the proposition that age and body mass index have an effect on sleep quantity and quality, while alcohol use was indicated as an important factor in only one technique. From this, one cannot conclude that alcohol use is irrelevant or that age and body mass index are clearly important to sleep; rather, all the techniques and results described in this report are simply sources of evidence that must be weighed accordingly.

Overall, the project was a great learning experience and provided a valuable opportunity to learn how R can be a tremendous and versatile tool for data analysis.

References

1. Centers for Disease Control and Prevention web site (<https://www.cdc.gov/features/dssleep>), accessed May 6, 2017.
2. *Sleep Disorders and Sleep Deprivation: an Unmet Public Health Problem*, Institute of Medicine, National Academy of Sciences, Washington, D.C., 2006.
3. E. Wickwire, F. Shaya, S. Scharf, “Health economics of insomnia treatments: the return on investment for a good night’s sleep,” Sleep Medicine Reviews, vol 30, 72-82, Dec. 2016.
4. G. James, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical Learning, ISBN 978-1-4614-7138-7, Springer Science and Business Media, 2013.
5. M. Hirshkowitz et al., “National Sleep Foundation’s sleep time duration recommendations: methodology and results summary,” *Journal of the National Sleep Foundation*, vol. 1, issue 1, 40-43 (2015).
6. Mayo Clinic website (<http://www.mayoclinic.org/healthy-lifestyle/adult-health/expert-answers/how-many-hours-of-sleep-are-enough/faq-20057898>), accessed May 15, 2017.
7. Google Scholar web site (<https://www.google.com/#q=scholar>), accessed May 6, 2017.
8. Kaggle web site (<https://www.kaggle.com/cdc/national-health-and-nutrition-examination-survey>), accessed March-May 2017.