STATS 415 Group Project
Group #5
Members: Nicky Li Alyssa Yang Kiley Price Amanda Li

In our report, we plan on using the NHANES 2013-2014 dataset. Specifically, we will be using subsets of data from the questionnaire, laboratory, and examination surveys to answer the following questions.

Question 1 (Classification):

- Is the structure of variability the same between the two classes of those who self-identify as depressed and those who do not?
- We will be classifying the response of “feeling down/depressed/hopeless” (DPQ020) by building a model with the following predictors:
    - Quantitative:
        - Average # of drinks/day (ALQ130)
        - Vitamin D level (LBXVIDMS)
        - Amount of sleep per night in hours (SLD010H)
        - Monthly family income (IND235)
        - BMI (BMXBMI)
    - Qualitative:
        - Worried about food (FSD032A)
- We will first split the dataset into different folds. We will create and analyze both LDA and QDA models to determine which one provides a better fit using their respective cross-validated MSEs and the corresponding confidence intervals of the MSEs.
- Tools Used: LDA/QDA, Cross-Validation

Question 2 (Regression):

- What is the effect of each of the predictors on Body Mass Index? I.e. which ones are the most influential when predicting BMI?
- We will be regressing the response of Body Mass Index (BMXBMI) by building a model with the following predictors:
    - Quantitative:
        - Hours spent watching TV or videos over past 30 days (PAQ710)
        - Minutes spent outdoors 9am - 5pm not work day (DED125)
        - Direct LDL Cholesterol levels (mg/dL) (LBDHDD)
        - Direct HDL Cholesterol levels (mg/dL) (LBDLDL)
        - # of times in past year you had a sunburn (DEQ034D)
        - # of meals not prepared at home (DBD895)
        - Systolic blood pressure (BPXSY1)
- We will split our dataset into training and testing sets. We will then be building two regression models using both lasso and ridge constraints (using cross-validation to select lambda) and determining the best model through comparison of the test MSE. Then, we will interpret the coefficients of our chosen model to answer the question of how each predictor influences the response variable.
- Tools used: lasso and ridge regression, cross-validation

Kaggle Report

Members: Nicky Li Alyssa Yang Kiley Price Amanda Li
Name of Kaggle account: Amanda Li

1. Data Cleaning

After downloading all the files, we combined X_train and y_train on their SEQN columns to obtain a training set on each student and their respective data. We converted the categorical variables self_eval, teacher_eval, and district into factors and then removed the column SEQN because the student's identifying number should not affect their performance on a standardized test.

2. OLS Model with All Predictors

To set an initial baseline for performance, we first fit the model to an Ordinary Least Squares Model that indicated all predictors were significant (using a significance level of $p<0.05$), and it produced an R-Squared value of 0.6354.

3. Lasso and Ridge Regression

We then decided to try using lasso and ridge regression because they offer a way to control overfitting and handle multicollinearity by introducing a penalty term on the coefficients. Due to the large number of SRP terms, we thought that these models would be useful in reducing the impact of some that might not be as influential or important and thus improve our model.

After splitting the data into a train (70%) and test set (30%), we used cross-validation with grid search to select the best $\boldsymbol{\lambda}$ to use in our model according to which one gave us the lowest test MSE. Thus, we produced our predictions on X_test and found that both lasso and ridge regression did not perform well, both giving R-squared values of 0.44.

Because they didn't perform well, we came to the conclusion that the data might not be best fit using a linear regression model—both lasso and ridge are regularization techniques for preventing overfitting in linear regression models, and all three linear models thus far haven't produced high R-squared values. Not only this, but lasso and ridge performed even worse than OLS which suggests that most, if not all, of the predictors are important and should be included. Thus, we scrapped these ideas and tried using random forests instead.

4. Feature Engineering

Through the results of the OLS, lasso, and ridge regressions, we learned that we should include most, if not all, of the SRP values, but thought that including all of them separately could potentially exacerbate outliers, increase noise, and thus learn relationships that aren't generalizable to new and unseen data. We thought that it didn't make much sense to look at each SRP value on their own because they can fluctuate sporadically from day to day and are incredibly specific. Due to this, we decided to create a new variable that is the average of the SRP values because it would reduce the dimensionality of the model and help mitigate noise in the potential outliers of specific SRP values; summarizing them all into one feature lets the model capture the overall trend better and thus would lead to better generalization.

Additionally, through EDA, we also observed that the variables self_eval, teacher_eval, and district can be factored into distinct levels. We also observed that observations with an even-numbered district have higher y values than observations with an odd-numbered district. Thus, we also created a dummy variable indicating even/odd district by labeling them with 0 and 1, respectively. Running this back through the OLS model, we found that this dummy variable resulted in a better performance compared to using the original district variable, so this dummy variable replaced the original.

5. Random Forest + Average SRP

We settled on using random forest because it is very robust and is well-suited to handle non-linear relationships between the predictors and the response. It averages the predictions across multiple trees, so the final model would have decreased variance, reduced overfitting, and improved generalization. It is able to manipulate high-dimensional data with many predictors which would be good for this dataset due to the abundance of SRP values (although we ended up mitigating this another way) and also makes no assumptions about the data's distribution due to it being non-parametric—this is beneficial because we don't know the dataset's true distribution. Thus, in our final model, we include only four variables: teacher_eval, self_eval, district_dummy, and avg_SRP.

After using a random forest model with the average SRP value, we obtained an R-squared value of 0.70725 which is better than all of our previous models, suggesting that we were correct in thinking that the data is non-linear and that using the average SRP value instead of individual SRP values helped our model's overall performance.

6. Hyperparameter Tuning

Thus, we decided to stick with random forests and wanted to tune hyperparameters within the model so that it would be able to capture the trends in the data better. We chose to tune the hyperparameter mtry (the number of variables randomly sampled as candidates at each split) because it controls feature selection—a smaller mtry reduces the model's tendency to overfit the data while a larger mtry would be able to capture more complex relationships. Tuning mtry would directly influence our model's complexity, ability to generalize, and overall performance.

We tuned the hyperparameter mtry through grid search (values 1-4) and chose the best mtry according to which one gave us the highest repeated cross-validated (with 3 repeats) R-squared score during the tuning process. This value was mtry = 2. We saved that value and used it in our final random forest model and obtained an R-squared value of 0.877 which indeed improved our performance compared to before tuning mtry.

For our code, we referred to this website which shows how to tune the random forest model through mtry: https://machinelearningmastery.com/tune-machine-learning-algorithms-in-r/

We also decided to tune the number of trees generated by the random forest algorithm by running through a for loop with 6 models starting with 500 trees and ending with 3000 trees (with an increase of 500 trees between each model). We trained each model on a training dataset with 70% of the original data, and measured performance based on the RMSE of the predicted values for the remaining 30% of the data (the validation set). This resulted in us choosing ntree=1500.

7. Contributions

Initially, we each brainstormed and discussed the provided data and our approach to the Kaggle competition; however, we soon realized that iterating on and making minor adjustments to models in R was not compatible with a team of four with busy schedules. We instead decided to delegate tasks and worked independently but cross-validating our work regularly. After optimizing models and testing our code, we shared our findings with the whole group and worked together to write, edit, and finalize the report. After both cleaning the data, Alyssa worked on building the lasso and ridge regression models, the initial random forest model, re-featurized the SRP variables, and tuned mtry, and Amanda worked on the initial OLS model, worked on EDA to help factor any qualitative variables as well as create the dummy variable for district, split the data into training and validation, compared repeated CV and non-repeated CV results, and wrote the iterative search to tune for ntree. Kiley and Nicky helped proofread the report and ensure all ideas were flushed out and coherent.

Project Report

Members: Nicky Li Alyssa Yang Kiley Price Amanda Li

1. Introduction

In our report, we set out to answer two questions:

1) Is the structure of variability the same between the two classes of those who self-identify as depressed and those who do not? i.e. What are the differences between the conditional distribution of predictors given those who self-identify as depressed and those who do not?
   - Answering this question would provide insight as to whether the distribution differs between these two populations, which can then be used to better identify patterns specific to each group, aiding in early detection of depression and effectiveness of interventions.
   - In this case, our goal for this question is model interpretability. We believe that the conditional probabilities of both classes obtained from LDA and QDA models are easier to interpret than the results of more complex models like random forests, kernel svm etc.

2) What is the effect of various lifestyle and health factors on Body Mass Index and which are the most influential when predicting BMI?
   - Answering this question would provide insight into building a predictive model for BMI as well as help improve lifestyle recommendations for those who hope to improve their BMI.

In the following report, we applied various statistical tools which include but are not limited to: Linear and Quadratic Discriminant Analysis (LDA/QDA), cross-validation, bootstrap, and Lasso and Ridge regression.

2. Data

In order to answer the two questions above, we cleaned and joined subsets of data from the questionnaire, laboratory, and examination from the 2013-2014 National Health and Nutrition Examination Survey (NHANES).

To determine if the structure of variability is the same between two classes of individuals who self-identify as depressed and those who do not, we built a classification model with one response variable and six predictors (see Figure 2.1). In order to answer the binary question at hand, we transformed the response variable (feeling down/depressed/hopeless), which originally included four levels of frequency, into a two-level factor variable by categorizing all who responded “not at all” as one group, and the others as another. The six predictors we chose to accomplish this task were 1) average number of drinks per day, 2) vitamin D levels in nmol/L, 3) amount of sleep per night in hours, 4) monthly family income level, 5) body mass index (BMI) in kg/m^2, and 6) how frequently a respondent worries about running out of food.

To determine the effect of various lifestyle and health factors on Body Mass Index and which are the most influential when predicting BMI, we built a regression model with one response variable (body mass index (BMI) in kg/m^2) and eight predictors (see Figure 2.2). The eight predictors we chose to accomplish this task were 1) number of hours spent watching TV or videos over the past 30 days, 2) minutes spent outdoors between 9am and 5pm on a non-work day, 3) direct LDL cholesterol levels in mg/dl, 4) direct HDL cholesterol levels in mg/dl, 5) how frequently a respondent applies sunscreen, 6) number of meals a respondent consumes that were not prepared at home, 7) systolic blood pressure from the first reading in mm Hg, and 8) systolic blood pressure from the second reading in mm Hg. We selected this assortment of predictors because, from our perspective, some variables such as time spent watching TV, high levels of LDL cholesterol, and high blood pressure are notoriously associated with unhealthy lifestyles (and therefore BMI) while others such as minutes spent outdoors and how frequently a respondent applies sunscreen are not as directly associated. We intentionally chose predictors that might have a direct impact on BMI and those that may not, along with variables that may be highly correlated (such as blood pressure from the first and second readings) to test the performance of the Lasso and Ridge methods.

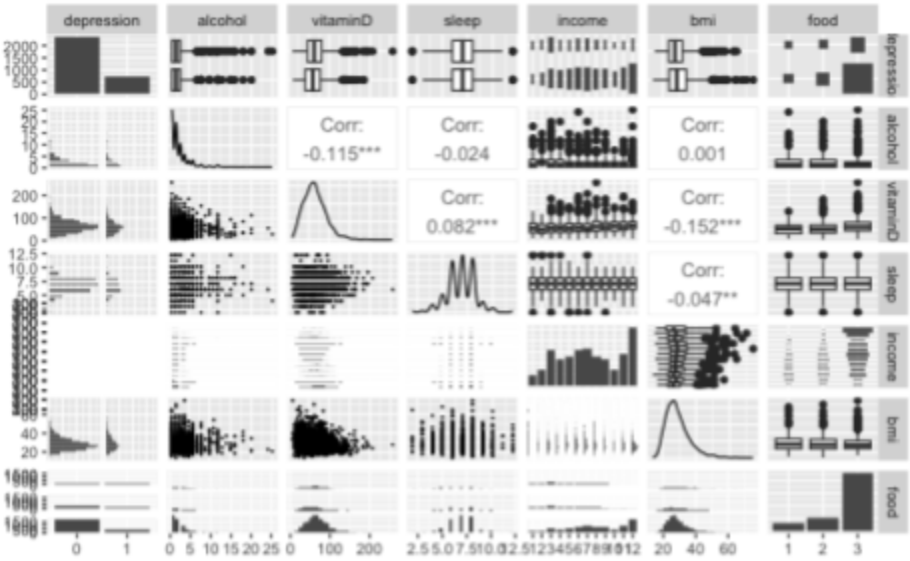

Figure 2.1 Correlation matrix and graphical representations of the relationships between features of depression dataset.

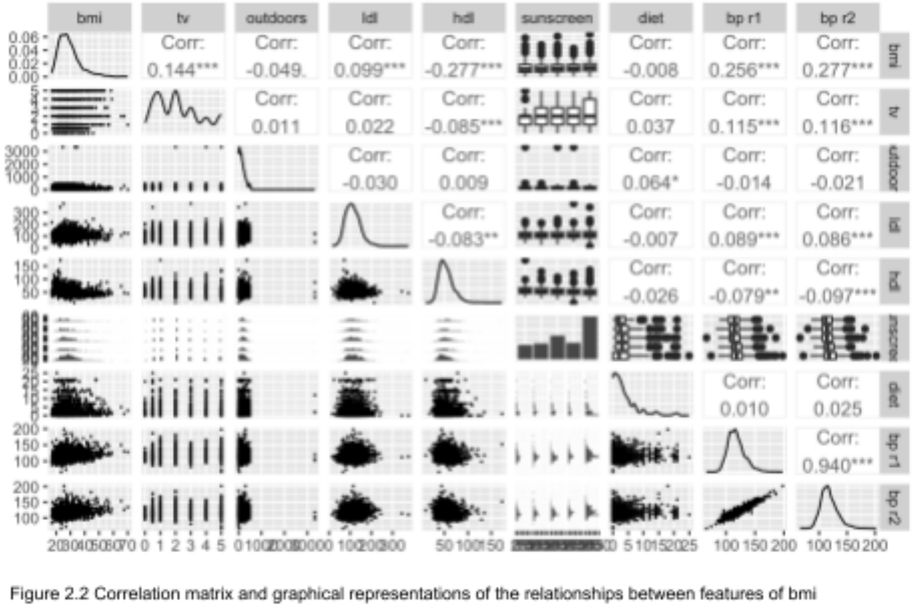

Figure 2.2 Correlation matrix and graphical representations of the relationships between features of bmi dataset.

3. Methods

Methods which lend themselves well to comparing the structure of variability between two classes are Linear and Quadratic Discriminant analysis (LDA/QDA). LDA with multiple predictors assumes that $X = (X_1, X_2, ..., X_p)$ is drawn from a multivariate Gaussian distribution with a class-specific mean vector and a common covariance matrix whereas QDA assumes that the observations from each class are drawn from a Gaussian distribution and each class has its own covariance. Another valuable statistical tool is cross-validation; although we had to remove approximately 2,300 records from our sample set due to incomplete or missing data, implementing k-folds cross-validation instead of the traditional validation set approach allows us to build k different models, utilize all remaining 3,107 observations for both testing and training, and continue to evaluate our model on data it has never seen before. For this project, we chose to execute LDA and QDA using 10-fold cross-validation and compare their respective cross-validated mean squared errors (MSEs) to determine which method is more appropriate.

To analyze the effect of various lifestyle and health factors on BMI and which are the most influential when predicting BMI, shrinkage methods such as Lasso and Ridge regression seemed appropriate as they are able to handle both high dimensionality and collinearity concerns. In this instance, our data contains ten features and a sample size of 1509, thus high dimensionality should not be of concern; however, we may run the risk of collinearity. In our model, we are considering systolic blood pressure from the first reading and from the second reading. Assuming that both readings were conducted properly, these two values for any individual should not differ drastically. Looking at figure (2.2), these two variables have a high correlation coefficient of approximately 0.94, thus we have reason to believe that collinearity is present among our data within these two variables. Since Ridge and Lasso regression account for this phenomenon, we do not have to worry about an imprecise estimate of $\boldsymbol{\beta}$, increased sensitivity to measurement errors, or numerical instability in our results.

In general, Ridge and Lasso regression aim to lower the variance of the model by introducing the penalty parameter $\boldsymbol{\lambda}$. When $\boldsymbol{\lambda}$ equals 0, the model becomes ordinary least squares. When $\boldsymbol{\lambda}$ approaches infinity, the model becomes less flexible and the coefficients shrink towards 0. Note that Lasso regression can be seen as a variable selection method since it can assign 0 to any variable that's not useful in predicting the response variable, whereas Ridge regression would have nonzero coefficients for all predictor variables. When performing Ridge and Lasso regression on the data, we use cross validation to find the best penalty parameter $\boldsymbol{\lambda}$ that would minimize the mean squared error between the prediction and the actual value. Running 100 samples of bootstrap ensures that the coefficient we get for each variable is a good approximation of its effect on the response variable BMI.

4. Results

Upon performing LDA and QDA to classify those who self-identify as depressed and those who do not based on the six predictors we listed above, we received a mean-squared error approximation of 0.239 for LDA and 0.309 for QDA. Additionally, the area under the ROC curve (a tool used to summarize the overall performance of a classification method) is greater for LDA than for QDA, as seen in figures 4.1 and 4.2. Although the ROC curve is not indicative of or dependent on class distribution, it does plot the tradeoff between sensitivity (the true positive rate) and specificity (1 - the false negative rate) as we vary the threshold value for the posterior probability of “feeling down/depressed/hopeless”. An ideal ROC curve hugs the top left corner of the plot; therefore, the larger the area under the curve (AUC), the better the classifier. Since the test mean-squared error is larger for QDA than LDA and the ROC curve for LDA runs closer to the top left corner, we can confidently suggest that Linear Discriminant Analysis is the superior classification method in this scenario. Similarly, we can analyze and compare the confusion matrices for LDA and QDA where 1 represents “feeling down/depressed/hopeless” and 0 represents otherwise.

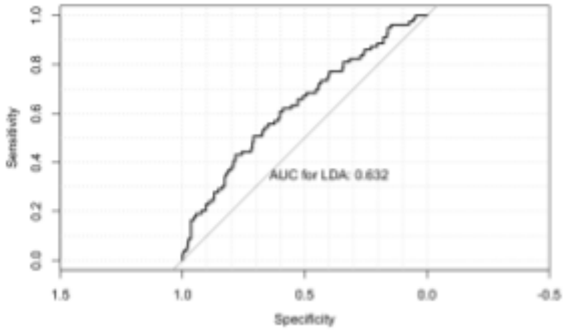

Figure 4.1 ROC curve for Linear Discriminant Analysis

LDA		Actual	
Predicted		0	1
	0	223	70
	1	8	9

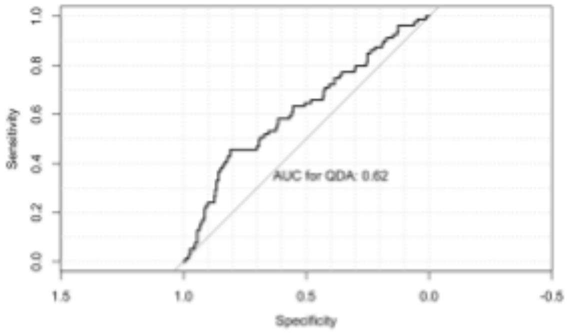

Figure 4.2 ROC curve for Quadratic Discriminant Analysis

QDA		Actual	
Predicted		0	1
	0	168	43
	1	63	36

For the second question, in order to avoid biased coefficient estimates, we scale the variables to a standard normal distribution with a mean 0 and a standard deviation 1 before performing Ridge and Lasso regression. After comparing the predicted values obtained from fitting Ridge regression to the actual values, we found that Ridge regression has a mean squared error (MSE) of 0.8566, as opposed to an MSE of 0.8511 found performing Lasso regression. The difference in MSE might be because Lasso regression removes the dummy variable “no effects on the skin without sunscreen” and the first reading of blood pressure, thus it has two less variables than that of Ridge regression. We already suspected that sunscreen usage would not necessarily be important in predicting BMI and assumed that only one blood pressure reading would be useful, thus it makes logical sense that removing both helped our prediction to be slightly more accurate. After taking 100 bootstrap samples, the fitted Ridge regression model is shown as Equation 4.3 and the fitted lasso regression model is shown as Equation 4.4. We’ve used cross validation in both Ridge and Lasso regression to find $\boldsymbol{\lambda}$ that produces the smallest MSE respectively.

$$\begin{aligned} Y = {} & 0.01978 + 0.08899X_1 - 0.0379X_1 + 0.0517X_3 - 0.2186X_4 - 0.0165X_5 + 0.0632X_6 \\ & + 0.163X_7 - 0.0679X_8 + 0.0265X_9 - 0.0804X_{10} - 0.0112X_{11} \end{aligned} \quad (4.3)$$

$$Cost = \sum(Y - \widehat{Y})^2 + 0.02903 * \sum((0.08899)^2 + (-0.0379)^2 + ... + (-0.0112)^2)$$

$$\begin{aligned} Y = {} & 0.00503 + 0.0797X_1 - 0.0231X_2 + 0.0391X_3 - 0.226X_4 - 0.00396X_5 \\ & + 0.2227X_7 - 0.0156X_8 + 0.00758X_9 - 0.0292X_{10} - 0.000572X_{11} \end{aligned} \quad (4.4)$$

$$Cost = \sum(Y - \widehat{Y})^2 + 0.0123 * \sum(|0.0797| + |- 0.0231| + ... + |- 0.000572|)$$

Where

X_1: Hours spent watching TV or videos over past 30 days

X_2: Minutes spent outdoors 9am - 5pm not work day

X_3: Direct LDL Cholesterol levels (mg/dL)

X_4: Direct HDL Cholesterol levels (mg/dL)

X_5: # of meals not prepared at home

X_6: systolic blood pressure from the first reading in mm Hg

X_7: systolic blood pressure from the second reading in mm Hg

and the categorical variable frequency of using sunscreen is represented as the following dummy variables:

X_8: most of the time

X_9: sometimes

X_{10}: rarely

X_{11}: never

Since Lasso regression has a slightly lower MSE than that of Ridge regression, we will use the model observed from Lasso regression (Equation 4.4) to make our conclusions. From the model, we see that Direct HDL Cholesterol levels and systolic blood pressure from the second reading are the most influential when predicting BMI. In particular, the coefficient for HDL is negative. This makes sense since high-density lipoprotein cholesterol is considered “good” and high levels of HDL are commonly associated with a healthier lifestyle, which could lead to a lower BMI. Compared to HDL, LDL seems to not have much effect in predicting BMI since it only has a coefficient of 0.04079. Additionally, the second systolic blood pressure reading has a positive relationship with BMI, meaning an increase in systolic blood pressure results in an increase in BMI. From the correlation matrix in figure 2.2, we see that there is a high correlation between first systolic blood pressure and second systolic blood pressure. And when this happens, Lasso would exclude one of the two variables that plays a less important role in predicting the response variable. Therefore, in this case, first systolic blood pressure is removed from the model. Comparatively speaking, variables Hours spent watching TV or videos, Minutes spent outdoors 9am - 5pm not work day, and Direct LDL Cholesterol level have less impact in predicting BMI. We know that physical activity can affect BMI level—large hours spent watching TV or videos often suggest a sedentary lifestyle, thus it’s reasonable that there is a positive correlation between number of hours watching TV and BMI levels. In addition, time spent outdoors is more indicative of an active lifestyle, which might explain why minutes spent outdoors on non-work days has a negative correlation with BMI. As the frequency of applying sunscreen decreases, the predicted BMI level first decreases then increases but then decreases again. Since there’s no clear trend between the frequency of applying sunscreen and BMI and the associated coefficients are quite small, we conclude that there’s no relationship between the two. Lastly, we see that the coefficient for the number of meals not prepared at home is small which means this variable plays a minor role in predicting BMI.

5. Conclusions

For our first question, our results lead us to conclude that the structure of variability is the same between the two classes of those who self-identify as depressed and those who do not; thus we may not need to consider different distributions between the two classes when iterating on future models to detect depression and test the effectiveness of interventions. We recognize that LDA and QDA may not be the most sophisticated models and may not lead to the best performance; but these two methods seemed like an appropriate entry point that could inform later decisions to implement more complex performance models such as random forest, kernel svm, lasso logistic and or permutation importance.

Another caveat that may be worth addressing is the false negative rates for these LDA and QDA classification methods. Although LDA has less misclassified points than QDA overall, the amount of observations falsely classified as not “feeling down/depressed/hopeless” is almost double the amount of false negatives for the QDA model (see figures 4.3 and 4.4). If this model were being implemented for diagnoses or prediction in the real world, one might argue that it would be better to err on the side of caution by classifying an individual as depressed and keep a watchful eye on them unnecessarily, than neglecting someone in a bad mental headspace; thus the QDA model might be preferred realistically despite its lower accuracy rate.

The results of the Lasso regression model for BMI (see equation 4.4) lead us to conclude that lifestyle and health factors such as the second reading of blood pressure and HDL cholesterol levels are most influential in predicting BMI since these variables correspond to the largest coefficients. Additionally, since Lasso selected the second of the two collinear blood pressure reading variables, we can infer that a second reading of blood pressure is necessary when performing medical examinations because it may be more accurate and indicative of an individual’s true health condition.

We may also note that the dummy variable “never applying sunscreen” has an average coefficient of -0.000255. But if we look at the 100 coefficients for this variable produced by each bootstrap iteration, we see that it has a coefficient of 0 most of the time, thus it would be more accurate to drop this variable in fitting a lasso regression. Furthermore, since the coefficients for all dummy variables corresponding to frequency of using sunscreen are quite small and unstable, it may be in our best interest to eliminate this variable from our initial model entirely for future investigation. Another potential shortcoming of our model may lie within the shrinkage parameter. We’ve performed bootstrapped Ridge and Lasso regression to find the coefficients used to fit the models with $\boldsymbol{\lambda}$ being the best $\boldsymbol{\lambda}$ obtained via cross validation, respective to each model. However, it’s possible that this $\boldsymbol{\lambda}$ might not be the best when fitting bootstrapped Ridge and Lasso regression.

6. Contributions

Initially, we each brainstormed and discussed our questions and approach to this project; however, we soon realized that executing our proposed ideas in R was not compatible with a team of four with busy schedules. We instead decided to delegate tasks and worked independently but cross-validating our work periodically. After writing and analyzing our code, we shared our findings with the whole group and worked together to write, edit, and finalize the report. Kiley took on the responsibility of cleaning the data and worked on the first question while Nicky primarily worked on answering the second question. Amanda and Alyssa helped edit the report, interpret the results, and ensure all ideas were fully flushed out and coherent.

7. Reproducibility

In order to obtain the results detailed above, one should import data from the appropriate NHANES 2013 - 2014 datasets, extract the sequence number and column(s) that correspond to the desired feature(s), remove any null values or responses that correspond to the aliases

“Refused” and “Don’t Know” and inner join all of the cleaned tables to produce one dataframe. Before proceeding, it is important to also cast any categorical features as factors.

To address the first question, we perform binary classification by categorizing all responses of 1, 2, and 3 to the NHANES survey question “How often have you felt down, depressed, or hopeless over the last 2 weeks?” as 1 since all three responses suggest that the individual has felt “down, depressed, or hopeless” at some point during the given time period. To reproduce analysis we obtained above, create an empty list to hold classification errors and then implement a for loop that repeats 10 times. For every ith iteration, we will initiate k-fold cross-validation but defining the test data as the ith set of 310 (sample size divided by 10) consecutive observations in the original data set and the training set being everything else. Since we are utilizing k-fold cross validation instead of random sampling, we do not need to set a seed. Next, call the lda function, defined in the MASS package, with depression as the response variable and all other features (except for sequence number) on the training data set. Still within the for loop, predict the response produced by the LDA model for the test set by calling the predict function. Then, calculate the test error by computing the mean number of times which LDA misclassifies a test observation and store this in the list that holds classification errors. Outside of the loop, calculate the mean of all 10 classification errors and plot the ROC curve of the test data’s true observations against the posterior probability of “feeling down/depressed/hopeless” according to LDA’s predictions. Repeat this process for QDA and compare results.

To analyze and answer the second question, we first scale all the variables except correspondent ID and sunscreen. Then we create a model matrix that includes all variables except for the intercepts. After that, we randomly allocate 70% of the data into the training data set and the rest to the testing data set. We then fit the model using the training data set and test its accuracy by finding the mean squared error between the predicted BMI and the actual BMI on the test data set. Since a tuning parameter $\boldsymbol{\lambda}$ needs to be selected for a Ridge regression model, we use cross validation to find the $\boldsymbol{\lambda}$ that produces the smallest MSE. Then we fit the Ridge regression model using that ideal $\boldsymbol{\lambda}$. To create 100 bootstrap datasets, we first create a 12*100 matrix. Next, we use a for loop to iterate 100 times. For the ith iteration (where i = 1,2, …,100), a random training and testing dataset is being used. With the same algorithm as above, we find the best $\boldsymbol{\lambda}$ then use it to fit the Ridge regression model. Next we store the coefficients of the model into the ith column of the matrix. After 100 iterations, the 12*100 matrix created before is now filled with coefficients with each row representing different variables. We then find the mean of each row to obtain the coefficient of the bootstrapped Ridge regression model. The same procedure is applied to the Lasso regression model.