STATS 415

DATA MINING AND STATISTICAL LEARNING

Obese or Not Report

Relating eating and health habits with Obesity rate as classified by BMI.



**By Christian Burns, Redmond Xia, Jiayu Zhang**

**Data**

The dataset is 2014-2016 Eating & Health Module Data sponsored by the U.S. Department of Agriculture’s Economic Research Service. The dataset started with 37 variables and 10210 observation. We took out 24 variables because many of the variables weren’t useful or relevant. Specifically, some variables we took out included who performed the survey, height and weight, variables with no variation, and how respondents received their groceries. We emphasized height and weight because they are a multicollinear covariate of BMI. i.e. we can find BMI through a linear combination of height and weight.

The dataset has 13 variables including household income information, eating habits as well as activity habits. There are in total 325 valid observations after we excluded the respondents with blank responses in the questionnaire. Setting the seed to 12345, we splitted the data 80% into training and 20% in test set. We kept the 12 variables (ERBMI represents BMI, EEINCOME1 represents income compared with 185% of poverty threshold, ERTPREAT represents total amount of time spent in primary eating and drinking (in minutes) and ERTSEAT represents secondary eating and drinking, EUDIETSODA represents soft drink consumption (diet or regular or both), EUEAT represents whether a respondent ate or drank while doing other things, EUEXFREQ represents exercise frequency, EUFASTFDFRQ represents times the respondents purchased prepared food, EUFDSIT represents whether the respondent had enough food to eat, EUSNAP represents whether the respondent received SNAP benefits, EUMEAT represents meat consumption, EUWIC whether the respondent recieved WIC benefits) in most of our analysis because in linear regression model, we only found one variable to be significant under the 10% significance level and only two were under 20% level.

**Abstract**

Our goal was to use the various machine learning algorithms we learned this semester to classify if the respondent has obesity or not. This meant BMI (Body Mass Index) values over 30 indicated people with obesity. Different degrees of obesity were not considered. We used response variables based on health habits, as mentioned above to classify if respondents were obese or not. This is based on the assumption that people who have worse health habits were more likely be obese. So people who exercise less and eat more junk food, on average, are more obese.

We used the majority of the regressions and classification techniques we learned in class to predict whether the person has obesity or not. The methods were Logistic, KNN, LDA, QDA, Trees, SVM, Lasso, Ridge, AIC, BIC, PLS, PCR, GAM, and multilinear regression. We did not know exactly how classification boundaries as we could not plot in a high dimensional plot. Instead, we compared the performances of these methods and analyzed how the data set behaved based on performance. Overall, the results high in accuracy obtained around a mid-high 70% success in prediction. However, in the regression analysis, many of the variables were not significant enough and the MSE was quite small. We used all 12 of the variables because we believed after downsizing the initial 37, the leftover covariates had some relations with weight gains. Using the AIC criterion, we would just have a single regression or with two covariates, which is unrealistic in the real world model.

Table of contents

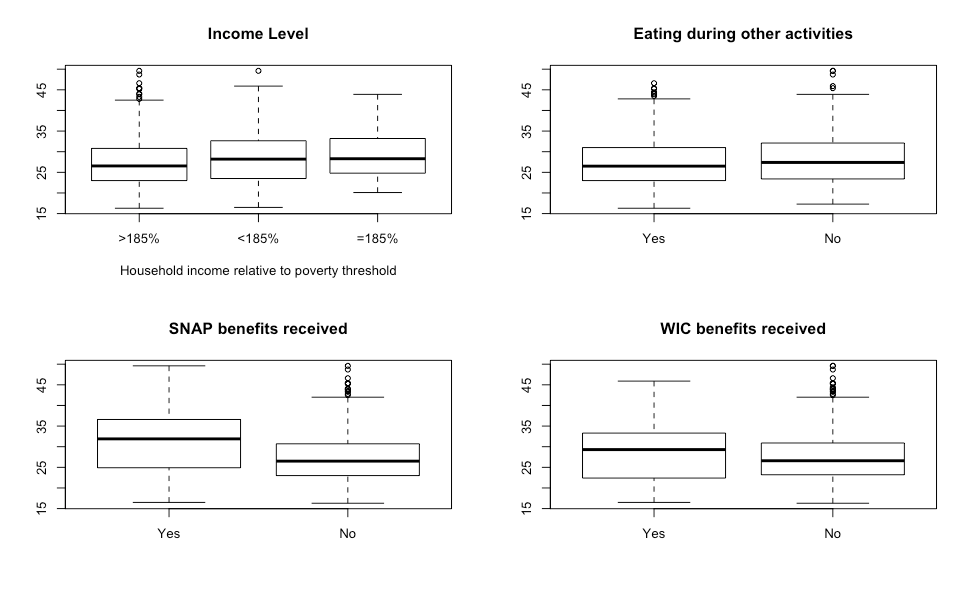
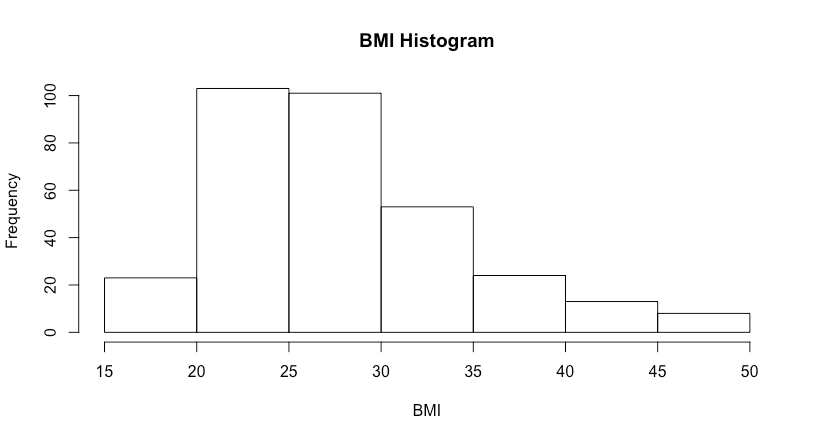
1. Visualization
2. Regression
3. Classification
4. Limitations
5. Conclusion

1. Visualization

After removing observations that were missing values, we began our initial analysis of the data set. We found the mean of BMI to be 27.94 with a variance of 43.59551. In the classification problem, we had 30.02% of the data classified as obese. Next, we started the analysis of correlations between the numeric variables. As seen in the figure below, no set of variables had a particularly strong correlation.



With the numeric predictors, observations seemed concentrated on the lower end of their scales. With most categorical variables, observations were concentrated in one of two classes. The diet vs. regular soda, and eating during other activities had around a 60/40 split. The distribution of BMI’s within the sample seemed to be relatively normal, as seen below.



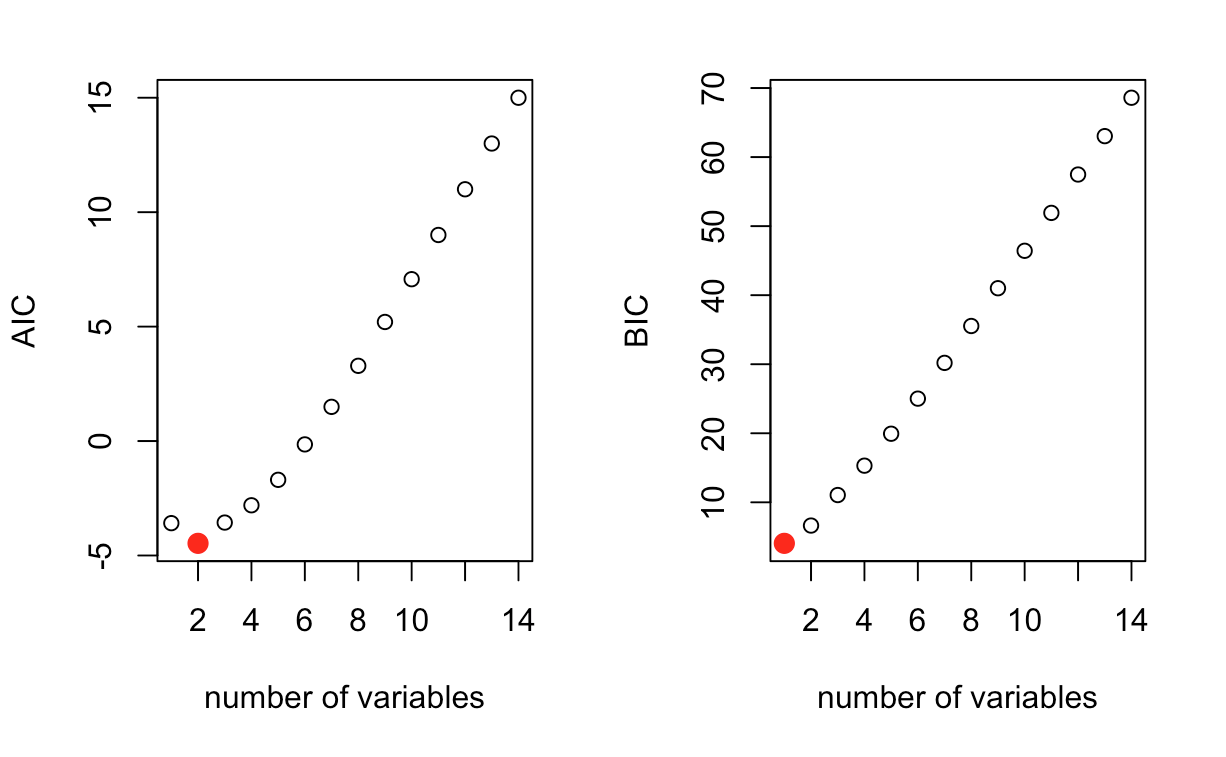
Next we looked at group means of the numeric variables split by categorical variable values. Nearly all were identical except for BMI with respect to income level, whether or not the respondent ate outside of meals, whether or not the respondent received SNAP benefits, and whether or not the respondent received WIC benefits. The relevant plots are pictured below.

2. Regression

We aimed to make best prediction on BMI with the 13 variables we chose in this section. We first divided the dataset into training data (80%) and test data (20%). Specifically, we chose the best subset model with AIC and BIC criterion, and we generated shrinkage models with Lasso and Ridge.

We computed the cross-validation error for lambda values of each Lasso and Ridge model under consideration, and then select the values for which the resulting estimated test error is smallest.

The AIC criterion selected 2 variables as the most critical subset in prediction of BMI: whether the respondent ate or drank while doing other things and whether the respondent received benefits from SNAP. BIC criterion selected only one variable as the most critical subset in prediction of BMI value: whether the respondent received benefits from SNAP. These results suggested that eating and drinking habit play an important role in BMI status. Not eating or drinking while doing other things resulted in a significantly higher BMI value, probably because distraction led to less food or drink consumption. SNAP (Supplemental Nutrition Assistance Program) offers nutrition assistance to low-income individuals and families and provides economic benefits to communities. From our results, respondents who did not receive benefits from SNAP had significantly lower BMI, which indicates that families that are not eligible for SNAP perform better at keeping fit.



Lasso and Ridge shrinkage gave similar results. The 2 most significant variables were also whether the respondent ate or drank while doing other things and whether the respondent received benefits from SNAP in both Lasso and Ridge shrinkage. Lasso shrank other coefficients to 0 while Ridge shrank them to extremely small values.

For both PCR and PLS only one component was used. This is because for both methods a single component produced the lowest CV error on the training data. For PCR, one component was able to explain 60.8% of the variance in X, but only 0.3394% of the variance in BMI. For PLS, one component was able to explain 54.4% of the variance in X, but only 0.66% of the variance in BMI. For both methods, the first (and second) components were linear combinations of time spent eating during meals and time spent outside of meals. Test errors are given below:

For GAM and multilinear regression, we ran the parameters chosen with all 12 variables that we used.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | AIC | BIC | Ridge | Lasso | PCR | PLS | GAM |
| Test error | 24.16274 | 22.93261 | 22.80036 | 22.90809 | 23.0196 | 23.58899 | 21.89058 |

All the errors were within an acceptable range. Since BMI is affected by many complicated factors, we suggest keeping all the 12 variables in prediction. Therefore, we recommend Ridge shrinkage may to model this dataset.

For multi-linear regressions, our resulting equation was

ERBMI = 30.4716 + 0.0415\*EEINCOME12 + 1.79163\*EEINCOME13 + 0.002457\*ERTPREAT + -0.004925\*ERTSEAT + -0.053685\*EUDIETSODA2 + 0.0112\*EUDIETSODA3 + 1.33475\*EUEAT2 + -0.166461EUEXFREQ + 0.2888309\*EUFASTFDFRQ + -16.01812\*EUFDSIT2 + 2.70301\*EUFDSIT3 + -3.951787\*EUSNAP2 + -1.094117\*EUEAT2 + 0.65593\*EUWIC2

This equation had an R^2 of 0.05539 and Adjusted R^2 of 0.00141. Only EUSNAP2 was slightly significant, meanwhile the majority of the data was not close to the 20% significant levels.

3. Classification

The goal of classification, which was the real aim for this project, was to classify if the person is obese or not based on their eating habit. We coded anyone with a BMI over 30 as obese. As mentioned before, we did not care which class of obese the respondents were.

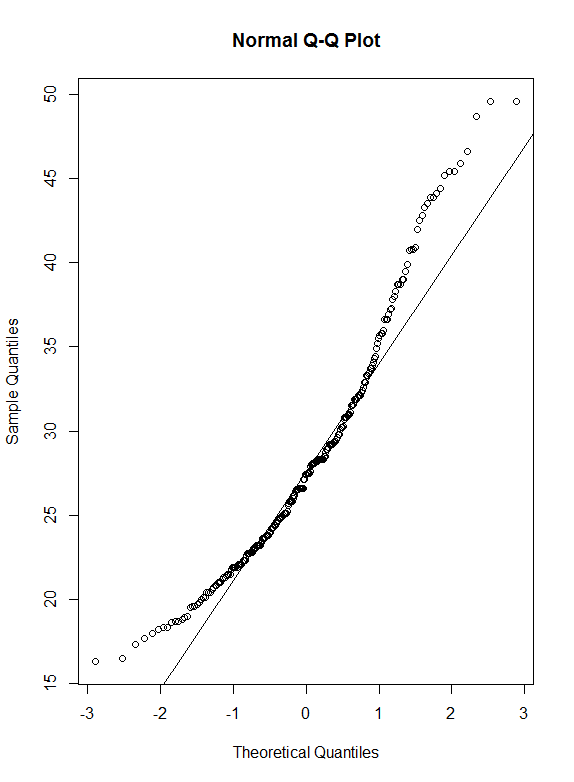
We used 6 techniques: Logistics, KNN (K-Nearest Neighbors), LDA (Linear Discriminant Analysis), QDA (Quadratic Discriminant Analysis), Trees, and SVM (Support Vector Machine). Using supervised learning on the training set, we predicted it on the test set and produced the errors. The different methods’ errors varied, with KNN, k = 7, performing the best. We did not did report the cross-validated error with KNN because it did not outperform the regular KNN. In addition, logistics, lda, and qda were not cross validated. Logistic, LDA, and QDA were used with the specified 12 parameters, KNN used a wide set of K (1:15, 50, 100, 150, 200, 250, 312) and found 7 to give the lowest test error.

The optimal single tree length was chosen through cross validation. This tree had 2 splits, one on whether or not the respondent received SNAP benefits, and the other on how long they spent eating during meals. The optimal tree size for random forest was chosen through lowest classification error. The best forest was comprised of trees with one split.

SVM parameters were chosen through cross-validation across several values. The linear SVM looks at costs from 1 to 10, and chose 5 as the best. The radial SVM looked at the same range of costs and gamma values of: 0.01, 0.1, 0.25, 0.5, 0.75, 1:5. It chose a cost value of 1 and gamma value of 2.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Logistic | 7-KNN | LDA | QDA | Pruned Single Tree | Random Forest | SVM | Radial SVM |
| Test error in % | 23.08 | 18.46 | 23.08 | 29.23 | 23.00 | 21.5 | 23.00 | 26.00 |

This means that the non-parametric method performed the best, and we should not make a big assumption on the function of people’s BMI, i.e. linearity or normality. Error rate or 18.46% is really small. It is also interesting to note that the Logistics and LDA performed the same. As you can see from the Normal Q-Q Plot (found below). The distribution does not follow a normal distribution. Which says people’s BMI does not follow normal. In addition, the right tail is heavier than the left tail, which is really light. This implies that people are more obese than they are underweight, which is supported by the right skewed histogram. With this in mind, we see that logistics performed the same without the normality assumption as the LDA does.



This confirms that it is hard to use parametric methods to model people’s BMI and we should use non-parametric method instead.

4. Limitations

Many of the classification techniques require continuous variables, meanwhile our data used mostly categorical data. The cut-offs for many categories might not be reasonable (i.e. 185% of poverty threshold).

Another shortcoming is that we took out majority of the data, 96.82% of the data to be precise. Taking out most of the data will result in a different analysis than the ERS intended with this survey.

Lastly, it could be possible that the respondents were already obese before taking the survey and most of the variables asked typical or recent habits. Thus, it could be that a very obese person could have healthy habits, but would need a very long time to reduce BMI to non-obese class.

5. Conclusion

In an effort to classify if people are likely to be obese in eating habit, the AIC gave two variables meanwhile linear regression only produced one slightly significant variable. One of the variable described if they multi-tasked and ate or if someone in the family household uses food stamps, with this stamps variable being common to both the linear regression and the AIC and BIC. Using all the variables in classification, we got a good prediction rate, however, many of the variables used seemed useless. In short, it is very hard to draw a conclusion based on the variables used, i.e. eating habits.

**Individual Contribution**

Collaboratively, all three of us chose the topic, looked for the dataset and cleaned up the data. Jiayu mainly performed the linear model selection and regularization part (including AIC, BIC, Lasso and Ridge). Redmond conducted GAM, KNN, logistic regression, LDA and QDA while Christian did trees, random forest, SVM, PLS and PCR. Everyone visualized and commented on their own part and together, we drew our conclusion from those comments. We all contributed to the limitation analysis and report writing.