DATA-390B: Final Paper (1st draft)

Maha Mapara

29th November, 2020

1 Introduction

Obesity is one of the consequences of the development of behavioral risk factors like an improper diet, physical inactivity, smoking and excessive alcohol consumption. According to the CDC, the age-adjusted prevalence of obesity among American adults was 42.4% in 2017–18.[1] Research shows that excess body fat increases your risk for several cancers, including colorectal, post-menopausal breast, uterine, esophageal, kidney and pancreatic cancers. Moreover, approximately 50% of diabetic patients are obese. And spikes in bad cholesterol and triglyceride levels, high blood pressure (common in obese people) are also a common cause of heart attacks. The World Health Organization has anticipated that 30% of global death will be caused by lifestyle diseases by 2030 and it can be prevented with the appropriate identification of associated risk factors and behavioral intervention plans [15]. Health behavior change should be given priority to avoid life-threatening damages.

My research question is centered around predicting obesity rates for some states and counties in the US, to be used by public health groups to assess the need for a public health intervention. Initially the research focus was in seeing how modifiable risk factors (like sedentary lifestyle, smoking, consuming non-nutritious foods) can be used to predict cancer, heart disease and diabetes linked to these factors. But as many of the modifiable risk factors/lifestyle choices lead to obesity and obesity increases the likelihood of these diseases, I am using obesity as an indicator for the diseases that will become prevalent or are prevalent in a population. Therefore, the premise of this research is that if obesity can be predicted, we can work towards preventing it and prevent heart diseases, cancers and diabetes caused by lifestyle choices, as a consequence.

Obesity can also be caused by certain genetic factors and hormonal conditions. For this research, I am only looking at the behavioral aspect of obesity, by focusing on behaviors like physical inactivity, dietary patterns, medication use, and other exposures. Many of these behaviors are related to an individual's environment. For example, a person may not walk or bike to the store or to work because of a lack of sidewalks or safe bike trails. Similarly, a person may have an unhealthy diet due to lack of food access. This project is predicting diabetes using data on food access, prevalence of quick service restaurants, food purchase, and nutrition. By doing so, the idea is to see a pattern of behavior that can be recognised earlier when a population is at risk for obesity instead of already being obese; and facing the health issues that accompany obesity.

1.1 Synopsis for my research project

1. The data set being used for obesity prediction is data from CDC's County Health Rankings website. An ensemble model will be used for the best prediction result. [10]

2. To understand the relation of obesity to food access and food deserts, data from CDC's Food Access Research Atlas is being used. [14]

3. To assess the relationship of food access and food deserts with quick service

restaurant's [8] (like McDonald's) and presence of super stores like Walmart (with limited fresh food options), I am using data from County Business Patterns [9] and Walmart store openings [6]. This will get a sense of fast food restaurant and superstore density in an area.

2 Related Work

Dunstan et al predicted obesity prevalence at the country-level using national sales data from 79 countries. The data came from 2 sources: 1. food and beverage sales data in 48 categories for 79 countries from the Euromonitor data set; 2. the percentage of obese adult population in these countries during 2008 estimated by Ng et al. On this data, they used 3 machine learning methods: support vector machines, random forests and extreme gradient boosting. The exploratory analysis was done first using principal component analysis and through visual exploration. For obesity prediction, random forest model performed the best, closely followed by extreme gradient tree boosting. For feature selection, the variable importance list was used; through this they found that baked goods/flours was the best predictor of obesity prevalence, followed by cheese and later by carbonated drinks with less than half the importance of the best predictor.

Singh et al. [13] evaluated different multivariate regression methods and multilayer perceptron (MLP) feed-forward neural network models on the data set obtained from a millennium cohort study (MCS) with over 90% accuracy to predict teenager BMI from previous BMI values. Twenty neurons in the hidden layer resulted in the lowest mean absolute error (MAE), with a mean training time of 1.63 s and a regularization factor of 0.9.

Bassam et al. [3] performed a study on data obtained from the Kuwait Health

Network (KHN) to build prognostic models to predict the future risk of diabetes (type II) using machine learning algorithms (logistic regression, k-nearest neighbor (KNN), support vector machine (SVM)) with a five-fold cross-validation technique. The study included age, sex, body mass index (BMI), pre-existing hypertension, family history of hypertension, and diabetes (type II) as baseline non-invasive parameters. As a result, KNN outperformed the other models, with area under the ROC (receiver operating characteristic) curve (AUC)values of 0.83, 0.82, and 0.79 for 3-, 5-, and 7-year prediction limits.

Meghana et al. [11] used 'auto-sklearn', an automatic machine learning (AutoML) library for developing classifiers of CVDs. They experimented on both the heart UCI dataset and a cardiovascular disease dataset consisting of 70,000 records of patients and, as a result, AutoML outperformed traditional machine learning classifiers.

Selya et al.[12] studied how to classify obesity from dietary and physical activity patterns using machine learning classification algorithms and, as a result, support vector machine (SVM) outperformed other classifiers.

Jindal et al. [7] performed ensemble machine learning approaches for obesity prediction based on the key determinants—age, height, weight, and BMI. The ensemble model utilized Random Forest (RF), generalized linear model, and partial least square, with a prediction accuracy of 89.68%.

Grabner at al. [5] performed a study on National Health and Nutrition Examination Survey (NHANES), National Health Interview Survey (NHIS), and "Behavioral Risk Factor Surveillance System (BRFSS)" data sets from the 1970s to 2008 to analyze the trend of BMI in the USA over time and across race, gender, socioeconomic background, and status (SES). It was observed that SES–BMI gradients were steadily more significant for women than for men.

Zheng et al. [16] used binary logistic regression, improved decision tree (IDT), weighted k-nearest neighbor (KNN), and artificial neural network (ANN) on nine health-related behaviors from the 2015 Youth Risk Behavior Surveillance System (YRBSS) for the state of Tennessee in their study to predict obesity in high school students by focusing on both risk and protective factors. The result showed that the IDT model achieved an 80.23% accuracy and 90.74% specificity, the weighted KNN model achieved an 88.82% accuracy and 93.44% specificity, and the ANN model achieved an 84.22% accuracy and 99.46% specificity in the classification problem.

DeGregory et al. [2] suggested in their literature review of "machine learning in obesity" that smart wearable wireless sensors, electronic medical health records, smartphone apps, and insurance data are rich sources of obesity-related data and are quite promising to treat and prevent obesity/overweight. Machine learning algorithms do have the potential to describe, classify, and predict obesity-related risks and consequences. They reviewed various machine learning methods, such as linear and logistic regression, artificial neural networks, deep learning, decision tree analysis, cluster analysis, principal component analysis (PCA), network science, and topological data analysis with the strengths and limitations of each method on the National Health and Nutrition Examination Survey to demonstrate the methodology, utility, and outcomes.

Golino et al. [4] used a machine learning technique, namely, a classification tree, to investigate the prediction of increased blood pressure by body mass index (BMI), waist (WC) and hip circumference (HC), and waist–hip ratio (WHR) on 400 college students from 16-63 years of age (56.3% women). The model outperformed the traditional logistic regression model in terms of predictive power. The model presented a sensitivity of 80.86% and specificity of 81.22% in the training set and, respectively, 45.65% and 65.15% in the test sample for women and a sensitivity of 72% and specificity of 86.25% in the training set and, respectively, 58.38% and 69.70% in the test set for men.

3 Louisiana County Health Rankings- Adult Obesity Prediction

This data was collected from the Louisiana County Health Rankings website. The data was available from 2010 to 2020, but as individual data sets. I had to manually merge the data sets and pick appropriate variables, due to the formatting of the data sets. After merging the data, it was pre-processed and explored as explained below.

3.1 Data Pre-processing and Exploration

1. The data originally had 715 rows and 65 columns. After removing rows for overall state values for each variable (which occurred 11 times), the total rows were 704.

Figure 1 shows some of the data types in the data set.

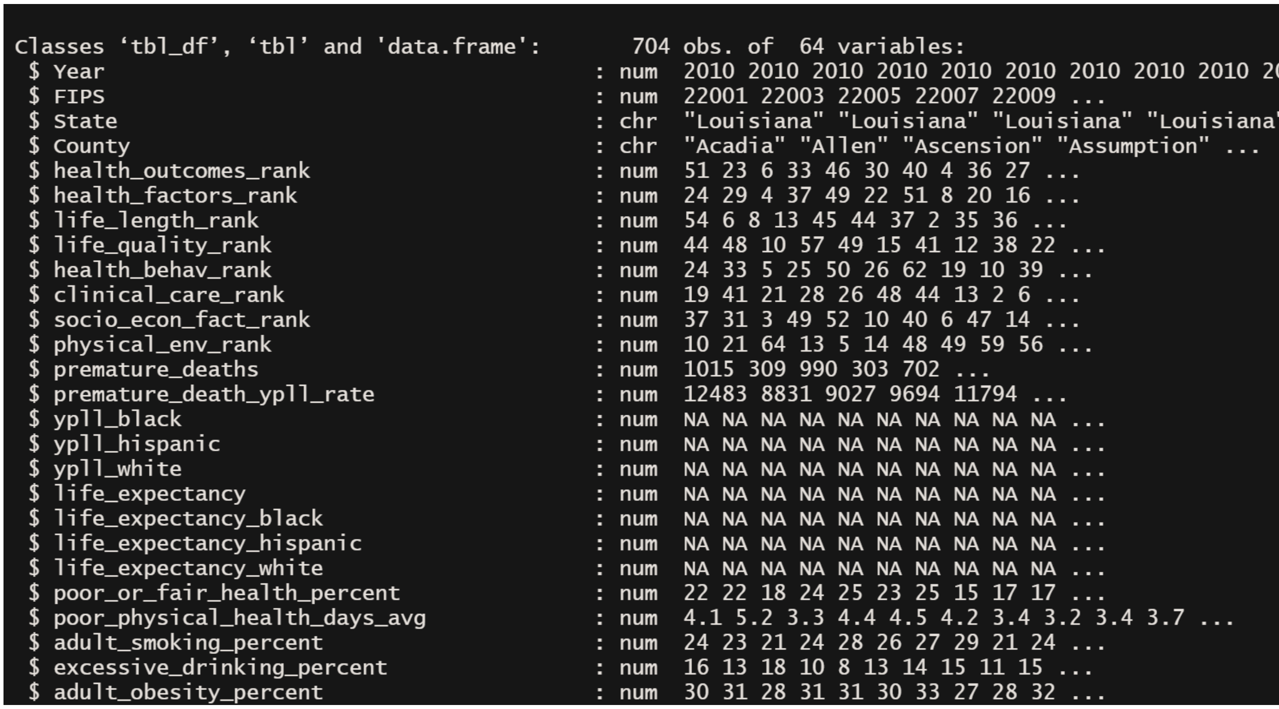
2. Dummy coding

As county was a categorical variable, it had to be changed from a character/string to a numeric factor variable. Figure 2 shows a part of the summary of the data set, reflecting this change.

The county variable was then coded into dummy variables, i.e, 64 columns of 0s and 1s were created to reflect which county's data a particular row was about.

3. Looking for outliers

a. In the univariate approach, I looked at histogram and boxplots of the dependent variable, adult_obesity_percent. Figure 3 shows a box plot with 4 suspected



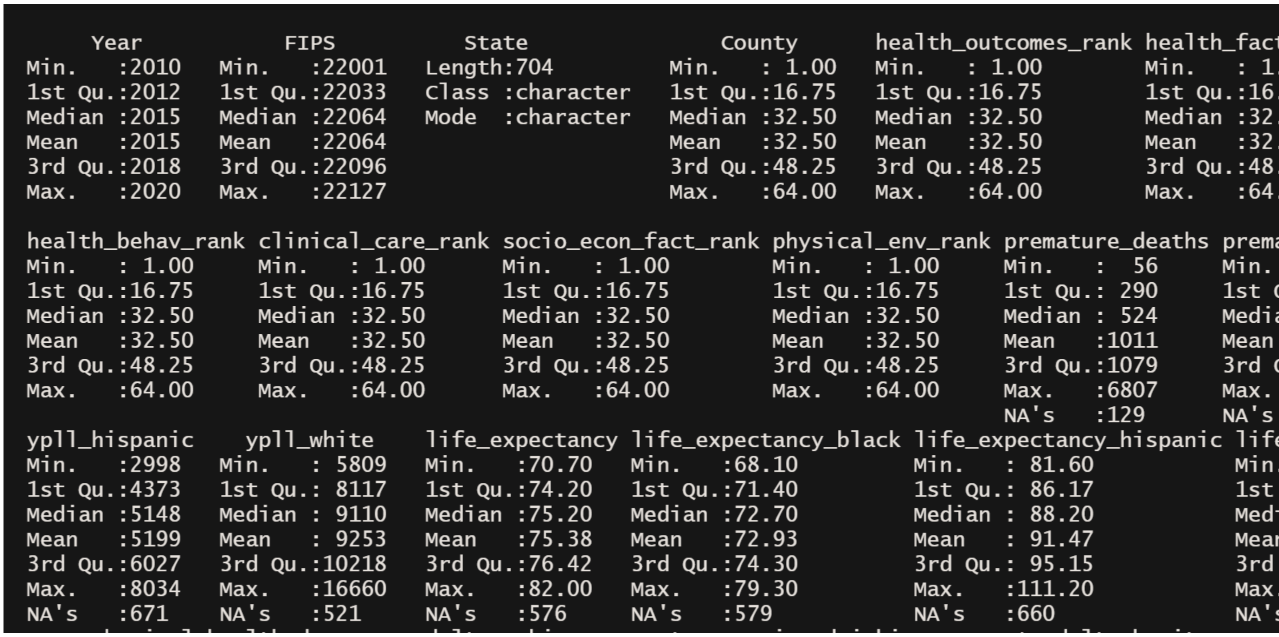
```
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':       704 obs. of  64 variables:
 $ Year                             : num  2010 2010 2010 2010 2010 2010 2010 2010 2010 20
 $ FIPS                             : num  22001 22003 22005 22007 22009 ...
 $ State                            : chr  "Louisiana" "Louisiana" "Louisiana" "Louisiana"
 $ County                           : chr  "Acadia" "Allen" "Ascension" "Assumption" ...
 $ health_outcomes_rank             : num  51 23 6 33 46 30 40 4 36 27 ...
 $ health_factors_rank              : num  24 29 4 37 49 22 51 8 20 16 ...
 $ life_length_rank                 : num  54 6 8 13 45 44 37 2 35 36 ...
 $ life_quality_rank                : num  44 48 10 57 49 15 41 12 38 22 ...
 $ health_behav_rank                : num  24 33 5 25 50 26 62 19 10 39 ...
 $ clinical_care_rank               : num  19 41 21 28 26 48 44 13 2 6 ...
 $ socio_econ_fact_rank             : num  37 31 3 49 52 10 40 6 47 14 ...
 $ physical_env_rank                : num  10 21 64 13 5 14 48 49 59 56 ...
 $ premature_deaths                 : num  1015 309 990 303 702 ...
 $ premature_death_ypll_rate        : num  12483 8831 9027 9694 11794 ...
 $ ypll_black                       : num  NA NA NA NA NA NA NA NA NA NA ...
 $ ypll_hispanic                    : num  NA NA NA NA NA NA NA NA NA NA ...
 $ ypll_white                       : num  NA NA NA NA NA NA NA NA NA NA ...
 $ life_expectancy                  : num  NA NA NA NA NA NA NA NA NA NA ...
 $ life_expectancy_black            : num  NA NA NA NA NA NA NA NA NA NA ...
 $ life_expectancy_hispanic         : num  NA NA NA NA NA NA NA NA NA NA ...
 $ life_expectancy_white            : num  NA NA NA NA NA NA NA NA NA NA ...
 $ poor_or_fair_health_percent      : num  22 22 18 24 25 23 25 15 17 17 ...
 $ poor_physical_health_days_avg    : num  4.1 5.2 3.3 4.4 4.5 4.2 3.4 3.2 3.4 3.7 ...
 $ adult_smoking_percent            : num  24 23 21 24 28 26 27 29 21 24 ...
 $ excessive_drinking_percent       : num  16 13 18 10 8 13 14 15 11 15 ...
 $ adult_obesity_percent            : num  30 31 28 31 31 30 33 27 28 32 ...
```

Figure 1: Data types for the data set

outliers at the very top and bottom of the plot (outside the interquartile range). Based on the interquartile range, the values of these outliers was extracted. After this I decided to use some statistical techniques to detect whether these values were truly outliers. First, I used Grubb's test which detects whether the highest or lowest value in a data set is an outlier. The result showed that both the lowest and highest values of the obesity percent were outliers. Figure 4 shows the Grubb's test result for the lowest value.

Then I conducted Rosner's test for outliers, which is useful for data sets for more than one outlier and data sets with n greater than 25. The result (shown in figure 5) was that no outliers were detected.

b. In the multivariate approach for outlier detection, I used Cook's distance. Cook's distance checks the influential observations in a predictive model to



```
     Year           FIPS          State               County      health_outcomes_rank health_fact
Min.   :2010   Min.   :22001   Length:704         Min.   : 1.00   Min.   : 1.00        Min.   : 1.
1st Qu.:2012   1st Qu.:22033   Class :character   1st Qu.:16.75   1st Qu.:16.75        1st Qu.:16.
Median :2015   Median :22064   Mode  :character   Median :32.50   Median :32.50        Median :32.
Mean   :2015   Mean   :22064                      Mean   :32.50   Mean   :32.50        Mean   :32.
3rd Qu.:2018   3rd Qu.:22096                      3rd Qu.:48.25   3rd Qu.:48.25        3rd Qu.:48.
Max.   :2020   Max.   :22127                      Max.   :64.00   Max.   :64.00        Max.   :64.

health_behav_rank clinical_care_rank socio_econ_fact_rank physical_env_rank premature_deaths prema
Min.   : 1.00     Min.   : 1.00      Min.   : 1.00        Min.   : 1.00     Min.   :  56     Min.
1st Qu.:16.75     1st Qu.:16.75      1st Qu.:16.75        1st Qu.:16.75     1st Qu.: 290     1st Q
Median :32.50     Median :32.50      Median :32.50        Median :32.50     Median : 524     Media
Mean   :32.50     Mean   :32.50      Mean   :32.50        Mean   :32.50     Mean   :1011     Mean
3rd Qu.:48.25     3rd Qu.:48.25      3rd Qu.:48.25        3rd Qu.:48.25     3rd Qu.:1079     3rd Q
Max.   :64.00     Max.   :64.00      Max.   :64.00        Max.   :64.00     Max.   :6807     Max.
                                                                            NA's   :129      NA's
ypll_hispanic    ypll_white    life_expectancy life_expectancy_black life_expectancy_hispanic life
Min.   :2998   Min.   : 5809   Min.   :70.70   Min.   :68.10         Min.   : 81.60           Min.
1st Qu.:4373   1st Qu.: 8117   1st Qu.:74.20   1st Qu.:71.40         1st Qu.: 86.17           1st
Median :5148   Median : 9110   Median :75.20   Median :72.70         Median : 88.20           Med
Mean   :5199   Mean   : 9253   Mean   :75.38   Mean   :72.93         Mean   : 91.47           Mean
3rd Qu.:6027   3rd Qu.:10218   3rd Qu.:76.42   3rd Qu.:74.30         3rd Qu.: 95.15           3rd
Max.   :8034   Max.   :16660   Max.   :82.00   Max.   :79.30         Max.   :111.20           Max.
NA's   :671    NA's   :521     NA's   :576     NA's   :579           NA's   :660              NA's
```

Figure 2: Summary of the variables in the data set

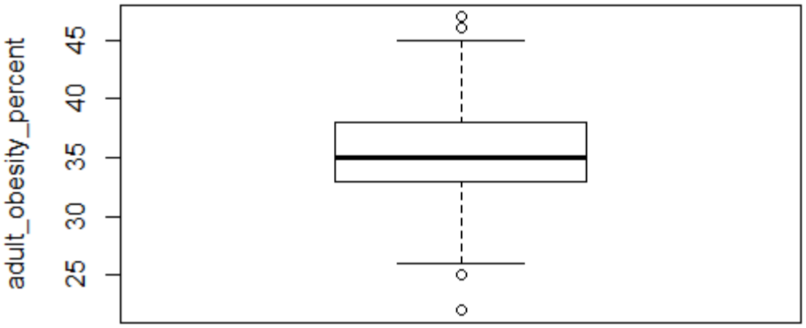
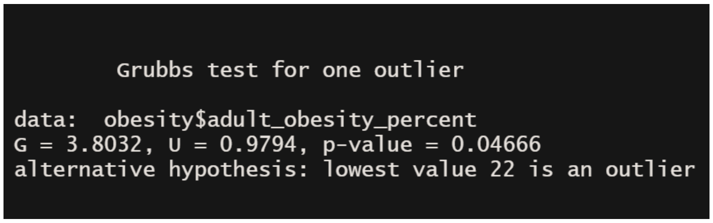


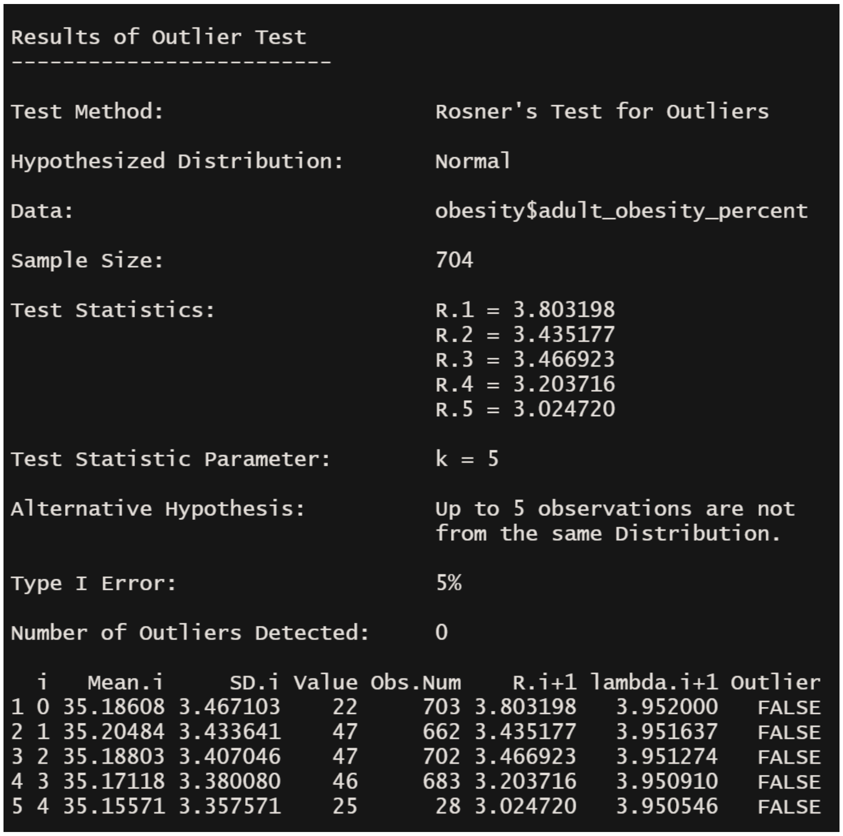
Figure 3: Box plot to detect outliers in the obesity variable

detect ouliers. I created a separate data set that included only variables that I intended to use in my multiple linear regression model later, ran the model and



```
        Grubbs test for one outlier

data:  obesity$adult_obesity_percent
G = 3.8032, U = 0.9794, p-value = 0.04666
alternative hypothesis: lowest value 22 is an outlier
```

Figure 4: Grubb's test for outlier detection



```
Results of Outlier Test
-------------------------

Test Method:                     Rosner's Test for Outliers

Hypothesized Distribution:       Normal

Data:                            obesity$adult_obesity_percent

Sample Size:                     704

Test Statistics:                 R.1 = 3.803198
                                 R.2 = 3.435177
                                 R.3 = 3.466923
                                 R.4 = 3.203716
                                 R.5 = 3.024720

Test Statistic Parameter:        k = 5

Alternative Hypothesis:          Up to 5 observations are not
                                 from the same Distribution.

Type I Error:                    5%

Number of Outliers Detected:     0

  i   Mean.i     SD.i Value Obs.Num    R.i+1 lambda.i+1 Outlier
1 0 35.18608 3.467103    22     703 3.803198   3.952000   FALSE
2 1 35.20484 3.433641    47     662 3.435177   3.951637   FALSE
3 2 35.18803 3.407046    47     702 3.466923   3.951274   FALSE
4 3 35.17118 3.380080    46     683 3.203716   3.950910   FALSE
5 4 35.15571 3.357571    25      28 3.024720   3.950546   FALSE
```

Figure 5: Rosner's test for outlier detection

calculated Cook's distance and plotted it. Figure 6 shows this.

Then I extracted these particular rows from the data set to see why they might have been seen as outliers.

In the end, I decided to not remove any of the values from both the univariate

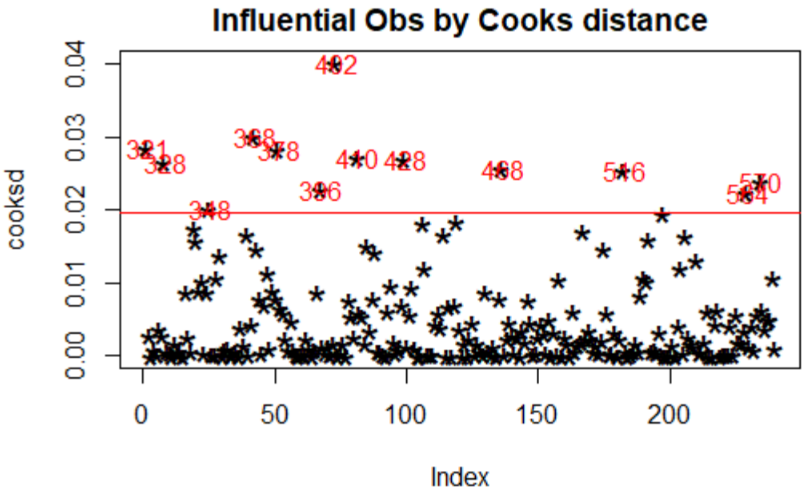


Figure 6: Plot of influential points by Cook's distance

and multivariate approach because each row in the data set is the value for a county in a particular year; and variation between values is expected and what I am trying to understand.

4. Standardizing the features

Some of my features were had values between 1 and 10, other were percentages between 1 and 100 and some were in the 1000s like median household income. To make sure that a model does not assign more importance to one feature over another because of different scaling, I standardized the variables I was using in my model. Variables that were county ranks like for clinical care, were standardized using a rank scale standardization method. Non rank variables were Z score standardized. County dummy variables were not standardized.

The different standardization methods led to different data frames. I had to merge them and add back the county dummy variables.

5. Missing data

The summary in figure 2 also shows many missing values in the form of NAs.The data frame with all the standardized variables then also had those missing values. I replaced all the NAs with 0, as this made those values meaningless but could still be put through a model without having the missing values be an issue.

6. Feature selection

One of the feature selection methods I used was finding variables with a high correlation. I did this using a correlation matrix with a cutoff at 0.5 and 0.75. I removed the highly correlated feature on ranks of health outcomes from my data set.

7. Data split

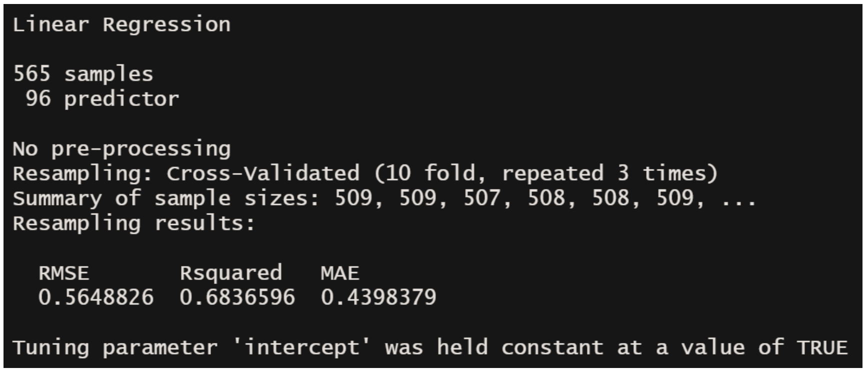
The data was split using the data partition function from the caret package with 80% of the data for the train set and 20% for the test set.

3.2 Modeling approach 1: Multiple Linear Regression

I used multiple linear regression as a feature selection method - to see which variables are the best predictors of obesity in Louisiana.

Each model was trained on the training data set using a thrice repeated 10 fold cross validation.

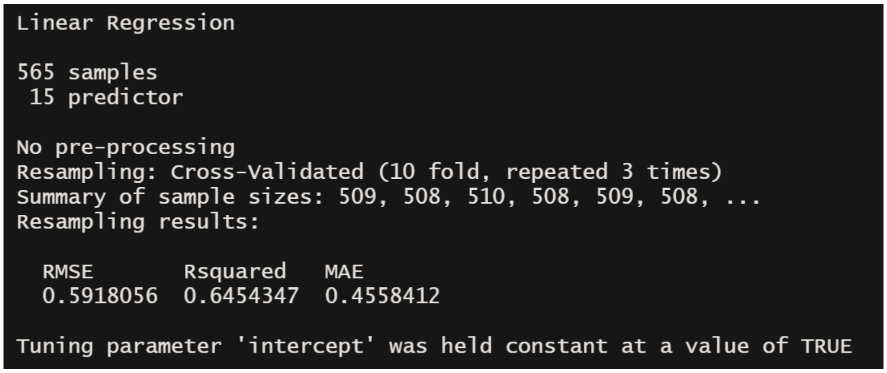
1. The first model had all the variables from my subsetted data set. This had features like life length and clinical care rank, % of adults who smoke, pose secondary education completion rate, % adults that are food insecure etc. Figure 7 shows the root mean squared of error (RMSE), the R-squared and the mean absolute errors for this model.
2. In the second model, I removed all the features for which the p-value was greater than 0.05. The 57th county was not significant in this model and so



```
Linear Regression

565 samples
 96 predictor

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 509, 509, 507, 508, 508, 509, ...
Resampling results:

  RMSE       Rsquared   MAE
  0.5648826  0.6836596  0.4398379

Tuning parameter 'intercept' was held constant at a value of TRUE
```

Figure 7: RMSE, R-squared, MAE for the model with all features

that was removed. Figure 8 shows the the root mean squared of error (RMSE), the R-squared and the mean absolute errors for the model with all statistically significant features. Figure 9 shows the model summary.

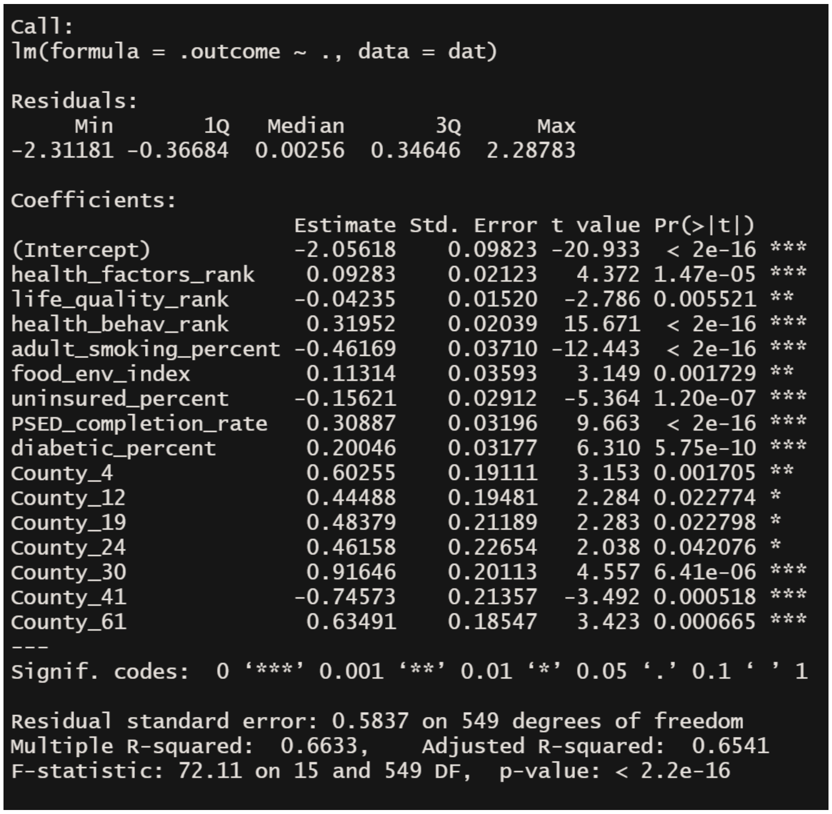


```
Linear Regression

565 samples
 15 predictor

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 509, 508, 510, 508, 509, 508, ...
Resampling results:

  RMSE       Rsquared   MAE
  0.5918056  0.6454347  0.4558412

Tuning parameter 'intercept' was held constant at a value of TRUE
```

Figure 8: RMSE, R-squared, MAE for model with all statistically significant features

3.3 Results & Interpretation

The second model (figure 9) shows the features important for obesity prediction in Louisiana. These are the health factors rank, life quality rank. health behavior rank, adult smoking %, food environment index, percentage of uninsured people, rate of completion for post-secondary education, percentage of diabetic



```
Call:
lm(formula = .outcome ~ ., data = dat)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.31181 -0.36684  0.00256  0.34646  2.28783 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)           -2.05618    0.09823 -20.933  < 2e-16 ***
health_factors_rank    0.09283    0.02123   4.372 1.47e-05 ***
life_quality_rank     -0.04235    0.01520  -2.786 0.005521 ** 
health_behav_rank      0.31952    0.02039  15.671  < 2e-16 ***
adult_smoking_percent -0.46169    0.03710 -12.443  < 2e-16 ***
food_env_index         0.11314    0.03593   3.149 0.001729 ** 
uninsured_percent     -0.15621    0.02912  -5.364 1.20e-07 ***
PSED_completion_rate   0.30887    0.03196   9.663  < 2e-16 ***
diabetic_percent       0.20046    0.03177   6.310 5.75e-10 ***
County_4               0.60255    0.19111   3.153 0.001705 ** 
County_12              0.44488    0.19481   2.284 0.022774 *  
County_19              0.48379    0.21189   2.283 0.022798 *  
County_24              0.46158    0.22654   2.038 0.042076 *  
County_30              0.91646    0.20113   4.557 6.41e-06 ***
County_41             -0.74573    0.21357  -3.492 0.000518 ***
County_61              0.63491    0.18547   3.423 0.000665 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5837 on 549 degrees of freedom
Multiple R-squared:  0.6633,	Adjusted R-squared:  0.6541 
F-statistic: 72.11 on 15 and 549 DF,  p-value: < 2.2e-16
```

Figure 9: Model summary

people and the Assumption, Cameron, East Faleciana, Iberville, La Salle, Red River, and West Baton Rouge counties.

3.4 Future Plans

First, I will take a closer look at all the data for the counties listed above by exploratory plots of the other significant features from the second model.

Second, I will use random forest to predict obesity and for more optimal feature selection. I expect the results to be better for this than for multiple linear regression.

4 Food Access Research Atlas data

I am using this data to expand on some variables on food access from the county health ranking data. The idea is to magnify correlations between food access, access to transportation, poverty, and family income. This data is a good supplement to the county ranking data as the time frame for the data collection overlaps. This data was collected from 2013 to 2017 and the county data is from 2010 to 2020.

As this data set was not being used for any modeling, I did not have to pre-process it. Figure 10 the poverty rate against the median household income for each county. The points are colored by whether people with low income and low access have vehicular access or not (LILATracts_Vehicle) [Note: LILATracts_Vehicle is a binary variable].

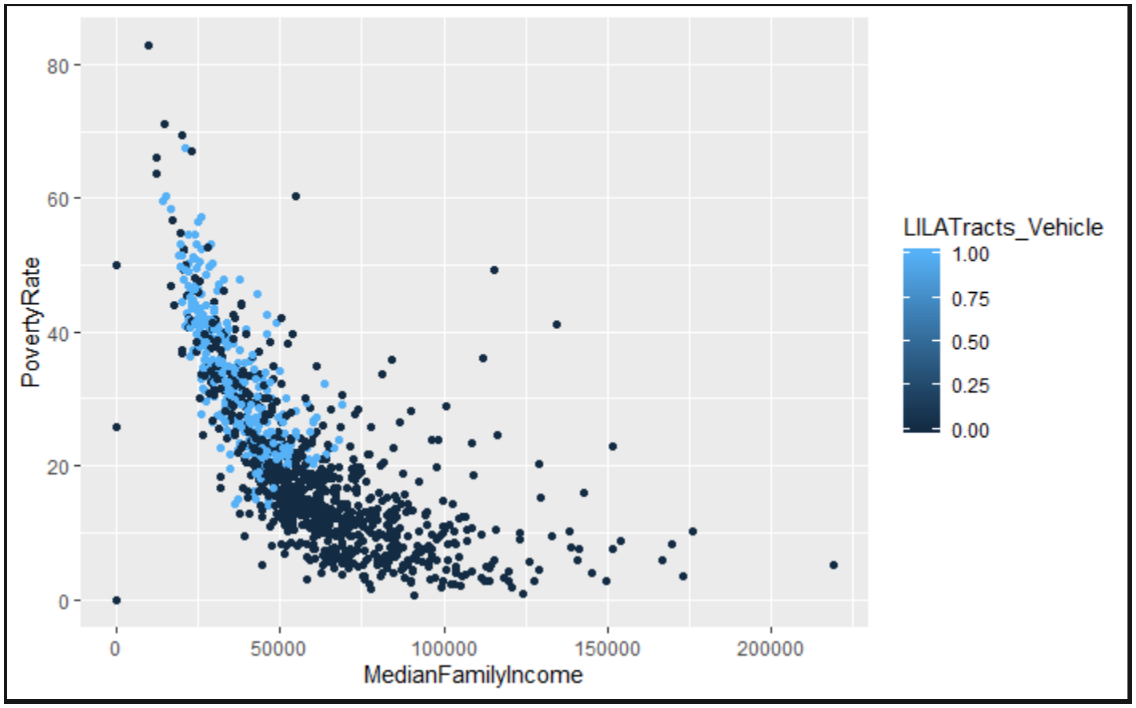


Figure 10: Poverty rate against the median household income for each county

Figure 11 shows the population count beyond 1/2 mile from supermarket against low income population count beyond 1/2 mile from supermarket. The points are colored by LATractsVehicle_20 which is a flag for tract where greater or equal to 100 of households do not have a vehicle, and beyond 1/2 mile from the supermarket; or greater or equal to 500 individuals are beyond 20 miles from the supermarket ; or greater or equal to 33% of individuals are beyond 20 miles from the supermarket.

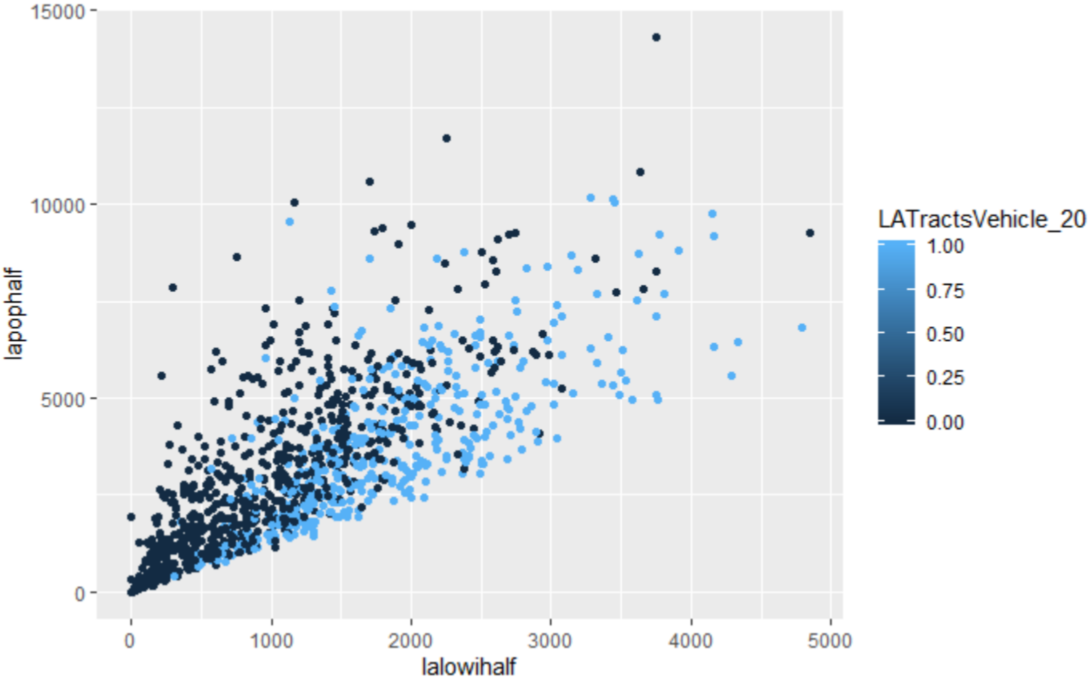


Figure 11: Population count beyond 1/2 mile from supermarket vs. low income population count beyond 1/2 mile from supermarket

4.1 Results & Interpretation

Figure 10 shows a negative correlation between poverty rate and median family income, i.e., as poverty rate increases, median family income decreases. The color separation is very clear showing that people without vehicular access have lower median family incomes and higher poverty rates. I would expect these people to have lower food access as well.

Figure 11 shows a positive correlation between population count beyond 1/2 mile from the supermarket and the low income population count beyond 1/2 mile from the supermarket. This is expected because as more people are beyond 1/2 mile from the supermarket, we would expect more low income people to be affected by that. The color separation is again clear showing that low income people beyond 1/2 mile from the supermarket also do not have vehicular access, and so have low food access.

4.2 Future Plans

I want to further see how different distances from the supermarket affect food access. Using the information on counties that were important for obesity prediction from the multiple linear regression model earlier, I want to see if those counties have lower food access than others and at what distances.

5 Other Future plans

5.1 Walmart store opening data

This data set has information on Walmart openings from 1970 to 2006.[6] The data set does not have any useful variables except for store opening dates and where they opened. I will map the zip codes in the data set for Louisiana and see where they have opened and if that relates to food access.

5.2 County Business Patterns data

The Economic Census County Business Patterns data for numbers of Limited Service Restaurants establishments is also available by county in Louisiana [9].

The Bureau of Labor Statistics Consumer Expenditures Survey also has data on food eaten outside the home. It is not broken down by restaurant or county, but a state overview will be useful to understand these eating patterns in relation to obesity.

Importantly, this will help understand the relationship between food deserts and quick service restaurants. With information from the Food Access Atlas data, I will be able to understand how these restaurants fill a gap where access to fresh foods does not exist.

References

[1] CDC. Adult obesity in the united states. 2020.

[2] Kuiper P. DeSilvio T. Pleuss J.D. Miller R. Roginski J.W. Fisher C.B. Harness D. Viswanath S. Heymsfield S.B. et al. DeGregory, K.W. A review of machine learning in obesity. 2018.

[3] AlWotayan R. Alkandari H. Al-Abdulrazzaq D. Channanath A. Thangavel A.T. Farran, B. Use of non-invasive parameters and machine-learning algorithms for predicting future risk of type 2 diabetes: A retrospective cohort study of health data from kuwait. 2019.

[4] Amaral L.S.D.B. Duarte S.F.P. Gomes-C.M.A. Soares T.D.J. Reis L.A.D. Santos J. Golino, H.F. Predicting increased blood pressure using machine learning. 2014.

[5] M. Grabner. Bmi trends, socioeconomic status, and the choice of dataset. 2012.

[6] Thomas J. Holmes. The diffusion of wal-mart and economies of density. 2011.

[7] Baliyan N. Rana P.S. Jindal, K. Obesity prediction using ensemble machine learning approaches. 2018.

[8] S. Lock. Number of quick service restaurant (qsr) franchise establishments in the united states from 2007 to 2020. 2020.

[9] United States Census of Bureau. County business patterns. 2020.

[10] LA Department of Health. Louisiana county health ranking data. 2020.

[11] Yuan P. Chada G. van Nguyen H. Padmanabhan, M. Physician-friendly machine learning: A case study with cardiovascular disease risk prediction. 2019.

[12] Anshutz D. Selya, A.S. Machine learning for the classification of obesity from dietary and physical activity patterns. 2018.

[13] Tawfik H. Singh, B. A machine learning approach for predicting weight gain risks in young adults. 2019.

[14] USDA. Food access research atlas. 2017.

[15] WHO. Obesity and overweight fact sheet. 2020.

[16] Ruggiero K. Zheng, Z. Using machine learning to predict obesity in high school students. 2017.