Analyzing Behavioural Risk Factors to Predict Obesity

Prinsa Gandhi

December 22, 2020

Abstract

This paper will be analyzing survey data collected in 2013 by the Behavioral Risk Factor Surveillance System (BRFSS) project in the United States. The focus of this analysis will be on *Obesity*. This paper includes a comprehensive analysis of the BRFSS dataset, and will focus on the predicting the response variable of whether a person is obese, based on several key predictors. After cleaning and preparing the BRFSS data, exploratory data analysis was performed. Next, numerous variable selection methods were used. Finally, the best model for prediction of obesity was chosen. There are very interesting results obtained during this analysis, which will be discussed throughout the paper. This analysis will be very important for public health workers, medical field workers, and government officials to make decisions to decrease the risk of obesity, as well as reduce healthcare spending which can be due to obesity. In addition, insurance companies and the general public will benefit from this analysis, and use the data to make important decisions.

The code and data supporting this analysis is avaliable at:

github https://github.com/username1-p/Analyzing-Behavioural-Risk-Factors-to-predict-Obesity

Keywords

Obesity, Regression, Prediction, Behavioral Risk Factors, Survey, Center for Disease Control and Prevention, Health, Generalized Linear Model

Introduction

Obesity has become a growing health concern in recent years. In Canada, approximately one in four Canadian adults are obese (PHAC, CIHI). For children and young adults in the ages of six to 17, 8.6% are obese (PHAC, CIHI). There are numerous costs that are associated with obesity, as well as common diseases that can be linked to Obesity. This can decrease Canadian life expectancy, affect the health of Canadians, create costs for the healthcare systems, as well as increase costs incurred by life and health insurers. The World Health Organization states that Obesity has tripled since 1975.

It is important to know that Obesity is preventable (WHO). This analysis will be using a survey dataset that has collected data on behavioral factors that affect various preventable diseases. The goal of this analysis is to identify various behavioral factors that are important predictors of Obesity, and interpret these predictors. This will aid in making important government decisions, and help insurers make important decisions regarding actions they may want to take to reduce obesity and reduce the harmful economic costs and health care challenges that arise because of Obesity. The analysis will be using a regression analysis to identify important variables, as well as interpret them. This will aid in identifying the most important behavioral factors.

This analysis will be based on a survey that collects data from people across the United States, thus, it is using survey data. The survey will be used to determine which individuals are obese based on the data collected. The interpretation of the results can be extremely important to aid in important decision making to reduce Obesity, and help pre-obese patients take steps to mitigate the risk of Obesity and help physicians take steps to direct patients to prevent obesity.

This analysis will describe the dataset used, and methodologies used to collect the survey data in the "Data section". The regression model used for statistical analysis will be described in the "Methods section". In the "Results section", the results of the regression analysis will be shown, which will include selected predictors, coefficients, and the final model selected. In the "Discussion section", there will be a description of the results of modeling, as well as any additional observations regarding the predictors and outcome. There will be important interpretation discussed, as well as recommendations for decision makers in the government that would be interested in these key findings. There will be a discussion of weaknesses and further recommendations that would have led to a better study and analysis.

Data

This information is shown in greater detail, on the CDC website, which is mentioned in the references. While the original data is from the CDC website, the data version I have used is obtained from Kaggle, and it is a bit more usable format of the data, since it is a bit more cleaned than the original data. However, it is originally obtained from the CDC website, thus, I will mention all information about the data from the CDC website.

The data used is collected by the Behavioral Risk Factor Surveillance System (BRFSS). The BRFSS project is used to collect behavioral risk factors for non-institutionalized adults living in the United States. The definition of adults is those who are 18 years of age or older. The very complex survey project is a collaboration states and territories in the United States, and Centers for Disease Control and Prevention (CDC). The BRFSS is administered by the CDC's population health surveillance branch. This survey collected data from all 50 states, and the District of Columbia, Puerto Rico, Guam, and US Virgin Islands. According to CDC, it is stated that "The BRFSS objective is to collect uniform, state-specific data on preventive health practices and risk behaviors that are linked to chronic diseases, injuries, and preventable infectious diseases that affect the adult population" (CDC). There is a wide range of data collected in the survey, which include factors such as exercise, sleep, health outlook, smoking and alcohol use, healthy eating, and important health conditions. This is a very rich dataset to analyze, however, in my analysis, I decided to clean the original data and select key predictors that seemed to be most important to predict whether an individual is obese.

In terms of collecting the data, the BRFSS conducted surveys on both landline and mobile phones. There are interviewers who collect data from adults in a household, whom are randomly selected to take part in this survey. The adults who are called on mobile phones are those who live in college or private residence. The survey development process is a collaboration between CDC and state health departments. Then, the state health departments conduct the surveys themselves, and the CDC collects all state data to analyze and process.

The data collection process was on a telephone-based survey. The BRFSS uses a raking method to weight the data. They adjust underrepresented groups to be more accurately represented in the final dataset. This is also used to adjust and weight the mobile phone surveys into the survey. The raking method was used in age and race by gender, education group, and additional categories.

Survey Design

The survey developed has 3 components. There is a core component which a common set of questions all states use, a set of optional questions by the BRFSS, and some state-specific questions. Additionally, some questions are taken from other national surveys such as the National Health Interview Survey. The BRFSS states they are used because data can be compared with those surveys, and the questions have been tested

before. The team managing this program is a group of state health officials, along with CDC's BRFSS management team.

The sampling frame used is a list of telephone numbers of the United States, which are randomly selected to dial. The sampled population is the selected telephone numbers who picked up the call, and completed the survey. The target population would be the adults in the United States. The BRFSS standard is that in all US states involved, the sampled population should be representative of all households with telephones in said state (CDC). The CDC states 51 states used a disproportionate stratified sample. Guam and Puerto Rico used simple-random sampling. In the disproportionate stratified sample, the BRFSS divided telephone numbers (sampling frame) into 2 strata. Then they were separately sampled with simple-random sampling. A high-density and medium-density stratum were created. The strata that telephone numbers were split into depended on the number of households' telephone numbers in a specific 100 blocks, and then they were split into high-density or medium-density strata. The 100 blocks areas were determined by factors such as area codes. If there were one or more listed telephone numbers of a household, they were high-density. This created a sample with equal probability of representing households across the state.

For mobile phone numbers, in 2013, the sampling frame was the Telecordia database of telephone exchanges. There were numbers sorted by area codes within states. The BRFSS used the population of the sampling frames' telephone numbers divided by desired sample size, to create K. This formed an interval. Then, the desired sample size intervals were created of the size K. Then, in each interval, there was one random selection of a telephone number. The target population for mobile numbers were adults in private residence or college residence, and received larger than 90% of their calls on their mobile number. The sampling frame, and sampled population selection is discussed earlier for mobile phones. Interview Process

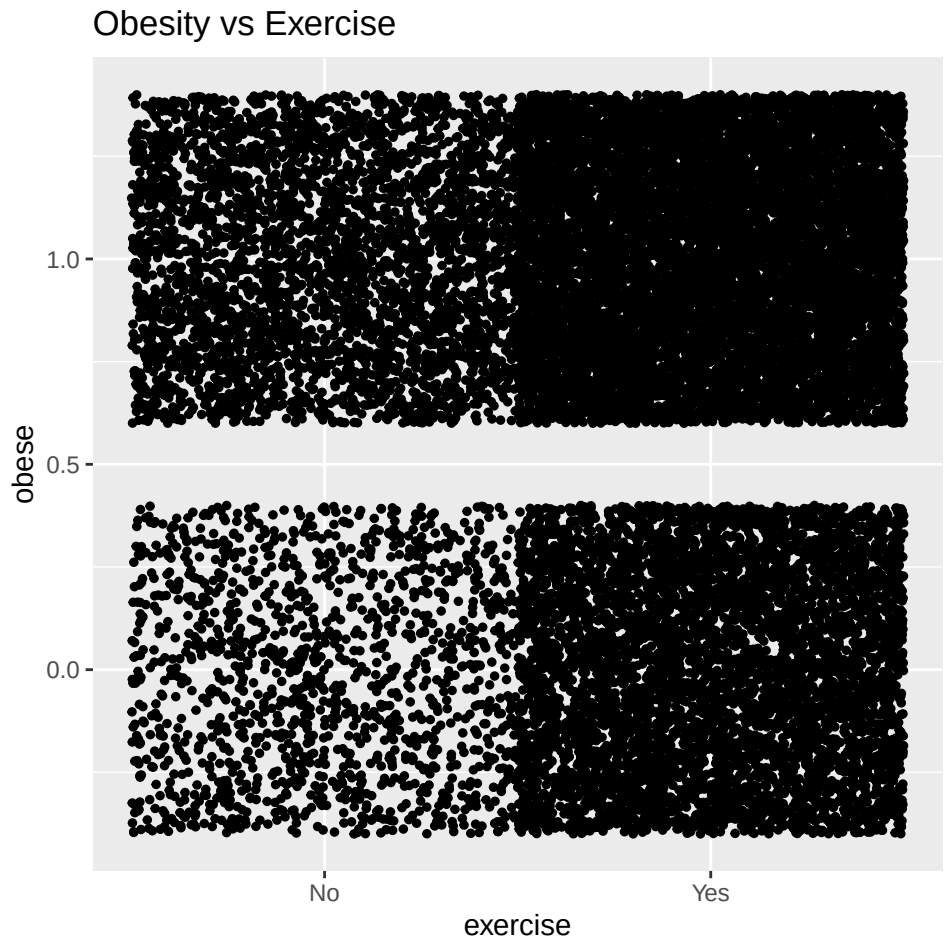
The process of interview used a Computer-Assisted Telephone Interview system. The time taken was about 18 minutes for core questions (CDC). The additional state specific and optional questions took about 5-10 minutes (CDC). Interviewers were given training. Data is submitted to the CDC monthly by states, and CDC continuously works, monitors and edits to prepare the final year-end data. The CDC also continuously monitors and works to improve the CATI system used. The respondents who refused to answer some questions are shown in the dataset as the refused answer option.In the data, most questions had an option for non-response. Then, data weighting is performed, so that assumptions can be met, and that populations can be reflected by the data. They follow design weight and raking methods.

The survey is very complex and well developed. There is some types of improvement discussed in the discussion section, but it is very rich data that can lead to numerous results. It is very representative of America's population,and can be further used to be representative of similar first world country populations. The variables used for this analysis will be shown in the methods and results section, which are obtained using model selection methods. There were a total of 330 original variables.

Exploratory Data Analysis

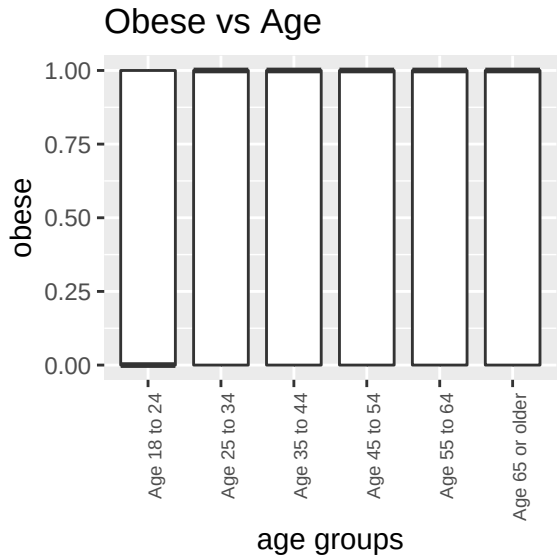
This section includes plots of EDA which helped analyze and understand the data. The plots are enlarged to show the points, because in smaller sizes, the points are clustered and it does not lead to an appropriate analysis.

Plot 1



In plot 1, the effect of exercise on whether a person has obesity is analyzed. Looking at the plot, it is clear that for no exercise, the proportion of people obese are higher. For the people that answered yes, it still seems that there is a split between obesity and non-obesity, further investigation is needed. It could be because of other lifestyle factors that cause the obesity in people who do exercise. The dataset has many observations, in a further enlarged plot, the results seem to look slightly different. It seems exercise is important to consider.

Plot 2

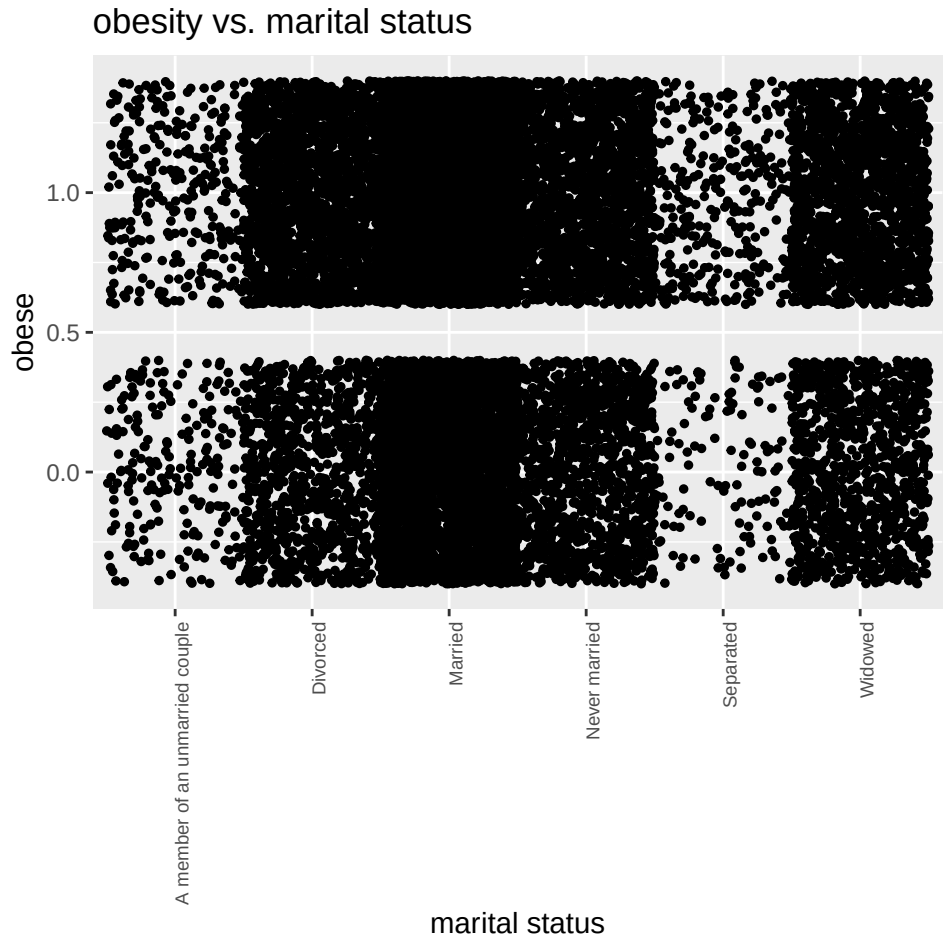


While analyzing plot 2, it seems that obesity is not impacted by age groups. For all age groups, there is a spread of people who are obese and not obese. This means that age groups may not be an important variable for the prediction model.

Plot 3

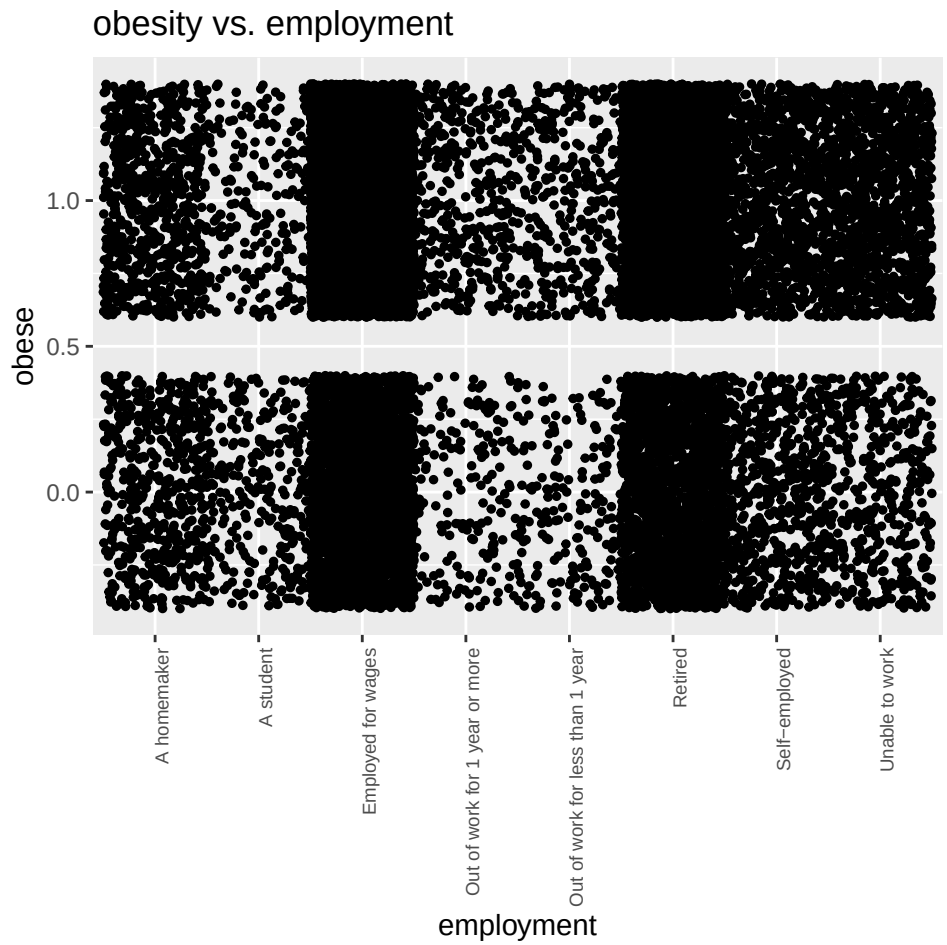
This plot is quite interesting. It can be noticed that for the "yes" level of having high blood pressure, there is a big split in the sample. It seems that most with high blood pressure are likely obese. The option of not being obese with high blood pressure is only shown with a small dot, thus, this an important risk factor to consider. For all other categorical levels, there is a spread of being obese and not being obese.

Plot 4



This plot shows the differences caused by marital status in the obesity status. It can be seen that for the married couple, there seems to be much more points in the obese range, compared to the non-obese range. The value of obese is only 0(not obese) or 1(for obese). The divorced or widowed categories also have more points that demonstrate increase in obesity for these categories.

Plot 5



While analyzing plot 5, it seems that for employed workers, retired individuals, self-employed and unable to work individuals, there is a higher proportion of people who are obese. This is interesting and important to consider for a prediction model. This seems to be an important predictor variable.

Methods

Cleaning

There were numerous steps taken to ensure a comprehensive analysis and to ensure that a very reliable and accurate model for prediction of whether an individual has obesity is found.

After analyzing the original dataset, it can be seen there are 491775 observations of 330 variables, for the year 2013. This was a very large, rich dataset. The first step was to analyze all variables included, and it was decided to select the variables that would be most important for my regression model to predict obesity. After reviewing, the dataset was modified to include 109 variables. A new "obese" variable was created, which was created from original variables in the dataset, indicating obesity in an individual based on the survey. This was created to aid in the regression analysis. A new modified dataset was created and saved, based on first procedure to clean the data.

The next step is to use this dataset, and further clean it. I realized that the dataset had many missing values in the responses, as individual chose not to respond to questions or did not have answers to certain questions. I decided to remove rows that had missing values for the "obese" outcome variable, since that was

an important variable in order to create an accurate model. Then, I decided to remove variables that had greater than 49 percent of missing values, because I did not want to impute those variables and create an inaccurate model. An option I considered was creating a regression model to estimate values for these missing variables, but there were too many missing values, so I did not choose to create a model. The variables I removed were college housing, blood pressure medicine, current as asthma related questions, some smoking variables, exercise related questions, prediabetes questions, blood sugar, pain, emotions, sugar drinks, sugar fruit drinks, sodium intake, aspirin intake, some reactions to race questions and self-identification, mental health and stigma, social context questions like rent or meals, and life satisfaction. These topics had numerous questions about them, however, due to many missing values and also considering its significance to the obesity model, I decided to remove the ones with large missing values.

For remaining variables with missing values less than 49%, I decided to use median or mode imputation. I felt many of the variables with missing values had a higher percentage of certain responses. For example, the responses could be 80% "yes", that is why I decided to use the median or mode imputation, because I thought that the numbers would still represent the entire population. I did feel that this would underestimate variance, and create some distortion, however, for many of the variables it seemed to fit in well. I found that the mean would still be the same, and was very easy to implement for the categorical variables.

Lastly, I observed that all the rows/observations in the data were independent, which is an important assumption of the logistic regression model. The response variable was "1/0" for the obese variable, which is also an assumption of the logistic regression model, since it followed the binomial distribution. These factors were the reason that a generalized linear model was used.

Modelling

After creating the final dataset for use in modelling, I decided to create train and test sets from the original data. The original data had 465,046 observations, however, I decided to choose a small train and test sample. In creating a regression model, it is very difficult to run tests and perform model selection in a large sample, which would not be very effective using the glm function in R. This will be discussed in the weaknesses section. I analyzed some of the variables, and found that some variables only have 1 factor. For example, for the variable "PVTRESDD1", 99.99% of respondents chose "Yes", for the question regarding if the individual was responding from a private residence. These variables had to removed, because they did not seem to have big impacts on the model, since most of the responses were the same. After removing variables, I created a sample set of the dataset which contained 20,000 observations. This was created by using simple random sampling without replacement. I created a training set of 16,000 observations, and test set of 4000 observations, both collected from the sample set using simple random sampling without replacement.

First, I created a logistic regression model using all the variables available, using the GLM function in Rstudio. I analyzed p-values to determine significant variables, and decided to remove variables that seemed insignificant to my model. There were also some variables causing perfect separation in the initial model, however, after reducing variables to the significant ones, that did not occur again. I performed AIC and BIC stepwise methods for model selection, and to try to obtain the most important predictors. I created a few more models, and continued using AIC and BIC model selection methods. The reason I continued performing model selection is due to the fact that the initial models had too many variables, and I did not want to overfit to the training dataset. In the initial models, factors such as vegetables or fruit intake did not seem to be significant. However, I felt there was correlation between many of my variables, and since the goal of my model was also to interpret important behavioral factors that predict obesity, I decided to include them in the later models. I performed AIC tests, and I compared 2 models using the ANOVA function. This helped compare the residuals and deviance in the models.

After numerous model selection procedures, the best model for prediction of whether a person was obese was chosen.The final model had a dependent variable which followed a binomial distribution, each observation was independent in the survey since individuals were interviewed only once. The model uses a GLM which uses the logit link.

For this model, the data used was a mix of numerical and categorical variables. For example, the "smoke100" variable, which asked whether one smoked in the last 100 days, was a variable representing yes or no. These variables were used as categorical, because the goal of the analysis was to interpret the model as well. It is important to see if certain behaviors can change the obesity outcome, and compare different behaviors. The age variable was split up into age groups in the dataset, and this was important, because certain age groups seem to have similar health. In categorical variables, the logistic regression uses one category as the baseline, and other categories are compared to the baseline. The GLM made sense according the data and objective, however, there will be discussion later regarding other models that may better handle data as big as this dataset. There are random forest, and complex machine-learning techniques with better prediction ability.

My model is based on the predictors that are self-perception of general health, whether a person was told by a medical professional that they have high blood cholesterol, whether a person was told if they have Chronic Obstructive Pulmonary Disease, emphysema or bronchitis, education, employment, sex, whether a person has difficulty walking, whether a person has smoked 100 cigarettes in their life, whether a person exercised in the last month, marital status, and whether a person ate vegetables during the past month. A note is that the last variable did not count "green", or "orange" colored vegetables as those were asked in other survey questions. The model converges and has no issues. The diagnostic checks performed will be discussed in the results section. The model passed diagnostic tests such as checking residual plots, or cross-calibration plots. The predictors were chosen using model selection methods, but also using characteristics of the survey. The predictors with responses that were representative of populations were used, and predictors which can be interpreted to make important decisions were used.

`The model I chose is shown below:`

$$logit(\pi_i) = \beta_0+\beta_1 x_{genhlth}+\beta_2 x_{toldhi2}+\beta_3 x_{chccopd1}+\beta_4 x_{educa}+\beta_5 x_{employ1}+\beta_6 x_{sex}+\beta_7 x_{diffwalk}+\beta_8 x_{smoke100}$$
$$+\beta_9 x_{exerany2} + \beta_{10} x_{marital} + \beta_{11} x_{vegetab1}$$

which is modeling

$$logit(\pi_i)$$

the log-odds of the probability of success for the response variable Y(either obese or not obese) dependent on the explanatory variables such as education, sex and the rest of the variables shown. The β_0 is the model intercept, and the $\beta_1, .., \beta_{11}$ are the model coefficients for each of the predictors.

Results

Table 1 shows the model results with the full summary table. This is also included in the appendix.

Table 1-Logistic Model for Obesity

```
##
## Call:
## glm(formula = as.factor(obese) ~ genhlth + toldhi2 + chccopd1 +
##     educa + employ1 + sex + diffwalk + smoke100 + exerany2 +
##     marital + vegetab1, family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2453  -1.2145   0.7062   0.9197   1.7476
##
```

```
## Coefficients:
##                                                   Estimate Std. Error z value
## (Intercept)                                     -0.4756471  0.1435367  -3.314
## genhlthFair                                      0.7737045  0.0701741  11.026
## genhlthGood                                      0.7711795  0.0525827  14.666
## genhlthPoor                                      0.3095397  0.0980463   3.157
## genhlthVery good                                 0.4598657  0.0489317   9.398
## toldhi2Yes                                       0.5217250  0.0383327  13.610
## chccopd1Yes                                     -0.2692489  0.0681670  -3.950
## educaCollege 4 years or more (College graduate) -0.3371930  0.0447054  -7.543
## educaGrade 12 or GED (High school graduate)     -0.0672944  0.0473452  -1.421
## educaGrades 1 through 8 (Elementary)             0.0021467  0.1195569   0.018
## educaGrades 9 though 11 (Some high school)      -0.0882976  0.0820449  -1.076
## educaNever attended school or only kindergarten  0.1457580  0.6753774   0.216
## employ1A student                                -0.3062977  0.1289604  -2.375
## employ1Employed for wages                        0.3563094  0.0747826   4.765
## employ1Out of work for 1 year or more            0.3567086  0.1275943   2.796
## employ1Out of work for less than 1 year          0.1493028  0.1274394   1.172
## employ1Retired                                   0.1209109  0.0764631   1.581
## employ1Self-employed                             0.1397631  0.0929176   1.504
## employ1Unable to work                            0.2914478  0.1024569   2.845
## sexMale                                          0.5686445  0.0376793  15.092
## diffwalkYes                                      0.4074043  0.0586794   6.943
## smoke100Yes                                     -0.1223825  0.0366214  -3.342
## exerany2Yes                                     -0.2039566  0.0434473  -4.694
## maritalDivorced                                  0.2257949  0.1158380   1.949
## maritalMarried                                   0.3253430  0.1089014   2.988
## maritalNever married                            -0.0557237  0.1144179  -0.487
## maritalSeparated                                 0.3310368  0.1619734   2.044
## maritalWidowed                                  -0.0034784  0.1185586  -0.029
## vegetab1                                         0.0004764  0.0001958   2.433
##                                                 Pr(>|z|)
## (Intercept)                                     0.000920 ***
## genhlthFair                                      < 2e-16 ***
## genhlthGood                                      < 2e-16 ***
## genhlthPoor                                     0.001594 **
## genhlthVery good                                 < 2e-16 ***
## toldhi2Yes                                       < 2e-16 ***
## chccopd1Yes                                     7.82e-05 ***
## educaCollege 4 years or more (College graduate) 4.61e-14 ***
## educaGrade 12 or GED (High school graduate)     0.155213
## educaGrades 1 through 8 (Elementary)            0.985674
## educaGrades 9 though 11 (Some high school)      0.281833
## educaNever attended school or only kindergarten 0.829130
## employ1A student                                0.017543 *
## employ1Employed for wages                       1.89e-06 ***
## employ1Out of work for 1 year or more           0.005180 **
## employ1Out of work for less than 1 year         0.241374
## employ1Retired                                  0.113810
## employ1Self-employed                            0.132540
## employ1Unable to work                           0.004447 **
## sexMale                                          < 2e-16 ***
## diffwalkYes                                     3.84e-12 ***
## smoke100Yes                                     0.000832 ***
```

```
## exerany2Yes                                    2.67e-06 ***
## maritalDivorced                                0.051268 .
## maritalMarried                                 0.002813 **
## maritalNever married                           0.626245
## maritalSeparated                               0.040976 *
## maritalWidowed                                 0.976594
## vegetab1                                       0.014967 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 20657  on 15999  degrees of freedom
## Residual deviance: 19413  on 15971  degrees of freedom
## AIC: 19471
##
## Number of Fisher Scoring iterations: 4
```

There are numerous results and interpretation obtained from table 1. This include the coefficients for the predictors, standard errors, and the z-values. The probabilities for the z-values are also shown, and the significant predictors can be seen. Lastly, there are some statistics included such as AIC, residual deviance, etc.The interpretation is further explained in the discussion section.

Table 2 is below, which displays the odds ratios for the variables.

Table 2-Summary Statistics-Odds Ratio

Characteristic	**OR**	**95% CI**	**p-value**
genhlth			
Excellent			
Fair	2.17	1.89, 2.49	<0.001
Good	2.16	1.95, 2.40	<0.001
Poor	1.36	1.13, 1.65	0.002
Very good	1.58	1.44, 1.74	<0.001
toldhi2			
No			
Yes	1.68	1.56, 1.82	<0.001
chccopd1			
No			
Yes	0.76	0.67, 0.87	<0.001
educa			
College 1 year to 3 years (Some college or technical school)			
College 4 years or more (College graduate)	0.71	0.65, 0.78	<0.001
Grade 12 or GED (High school graduate)	0.93	0.85, 1.03	0.2
Grades 1 through 8 (Elementary)	1.00	0.79, 1.27	>0.9
Grades 9 though 11 (Some high school)	0.92	0.78, 1.08	0.3
Never attended school or only kindergarten	1.16	0.34, 5.29	0.8
employ1			
A homemaker			
A student	0.74	0.57, 0.95	0.018
Employed for wages	1.43	1.23, 1.65	<0.001
Out of work for 1 year or more	1.43	1.11, 1.84	0.005
Out of work for less than 1 year	1.16	0.91, 1.49	0.2
Retired	1.13	0.97, 1.31	0.11
Self-employed	1.15	0.96, 1.38	0.13
Unable to work	1.34	1.10, 1.64	0.004
sex			
Female			
Male	1.77	1.64, 1.90	<0.001
diffwalk			
No			
Yes	1.50	1.34, 1.69	<0.001
smoke100			
No			
Yes	0.88	0.82, 0.95	<0.001
exerany2			
No			
Yes	0.82	0.75, 0.89	<0.001
marital			
A member of an unmarried couple			
Divorced	1.25	1.00, 1.57	0.051
Married	1.38	1.12, 1.71	0.003
Never married	0.95	0.75, 1.18	0.6
Separated	1.39	1.01, 1.92	0.041
Widowed	1.00	0.79, 1.26	>0.9
vegetab1	1.00	1.00, 1.00	0.015

This table is very important for interpretation of the predictors. It includes the odds ratio for each predictor, the 95% confidence interval, and p-value ranges. This will be explained in detail in the discussion section.

Prediction Error

There were numerous diagnosis tests that were performed. The test set was used to determine the prediction error of the GLM. The model was used on the test set, and the predict function was used to predict the response variable. If the probability of obesity was greater than 0.5, it was assumed that the individual was obese. If the probability of obesity was less than 0.5, it was assumed the individual was not obese. After prediction, the results were compared to the actual "obese" variable in the test set, to determine if the predictions were accurate. In the test set, the model was accurate about 67.2% of the cases.

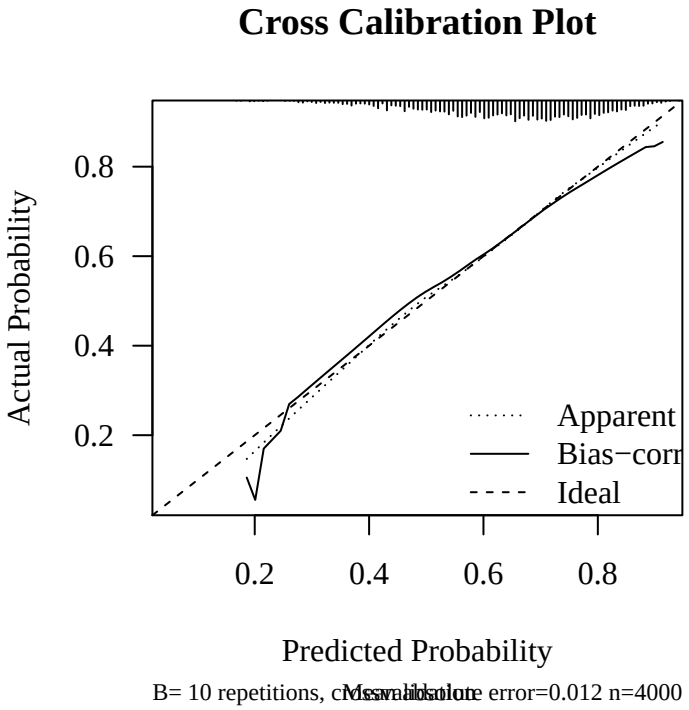
Diagnostics

These diagnostic methods will be further discussed in the discussion section.

The model was also used on the test set to check for the cross calibration plot. This is used to compared the predicted probabilities and see if they are seem to be accurate.

Figure 1-Calibration Plot on Test Set

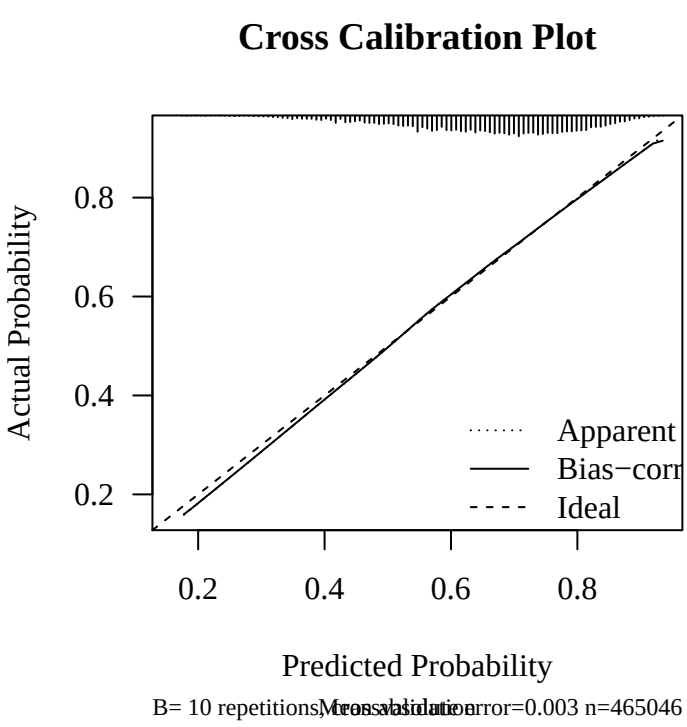
```
##
## n=4000   Mean absolute error=0.012   Mean squared error=0.00027
## 0.9 Quantile of absolute error=0.025
```



Looking at the plot, it seems that the model is very accurate and very close to the line, and the predicted and actual probabilities are very close to each other. The mean squared error is 0.00027.

Figure 2-Calibration Plot on Dataset

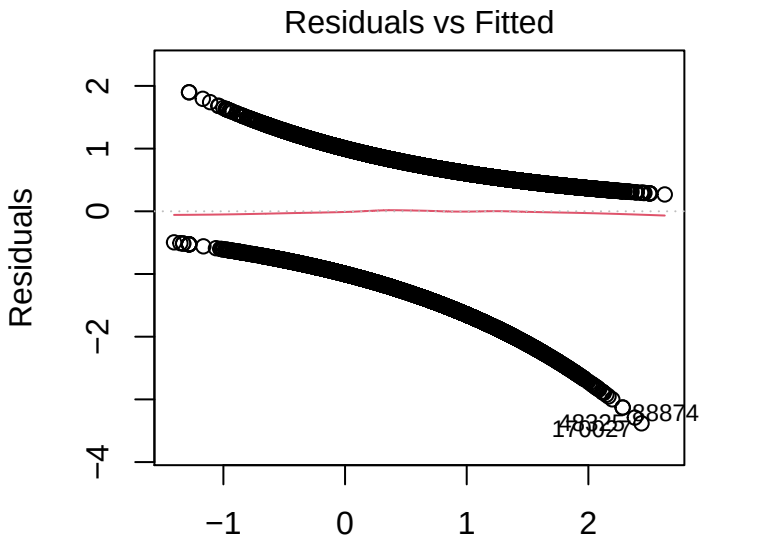
```
##
## n=465046   Mean absolute error=0.003   Mean squared error=2e-05
## 0.9 Quantile of absolute error=0.007
```

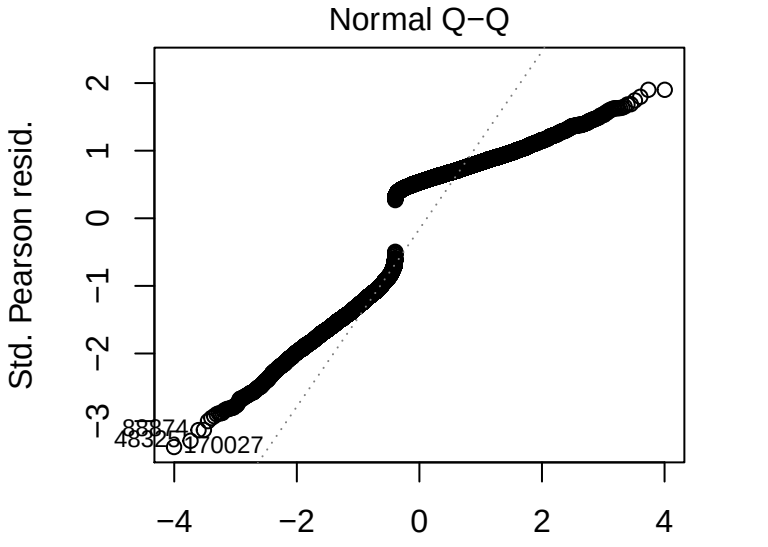


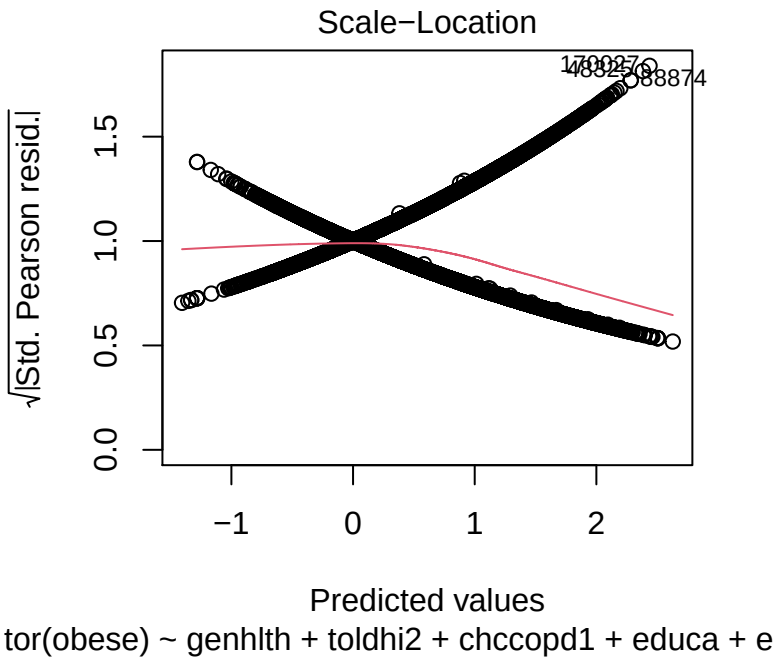
The calibration plot is created again, but this time using the entire dataset. The training set had 16,000 cases, and the test set had 4000 cases. It was very important to test the model on the entire dataset, because the entire set had 465,046 observations. It can help to further determine the accuracy of the regression model.

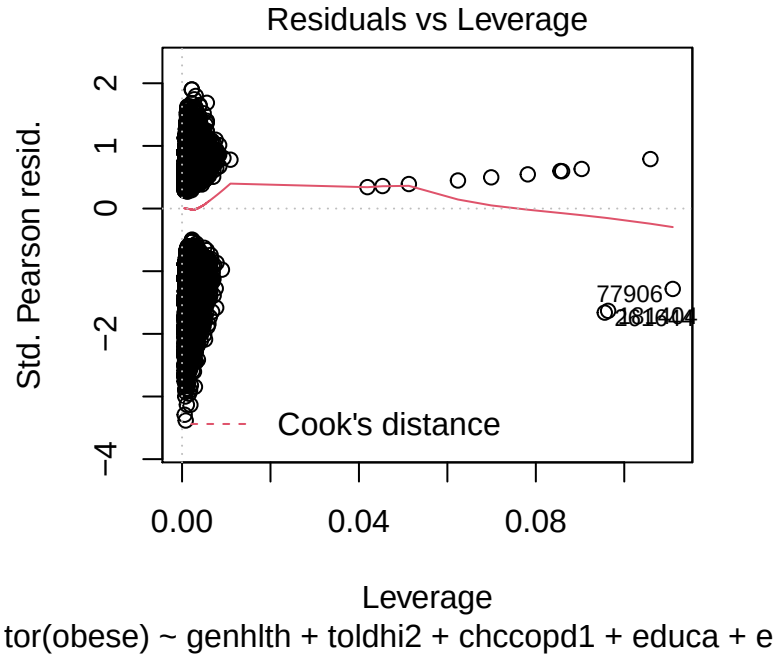
There were numerous other diagnostic performed as well. The figures below are other diagnostic performed to check the model.

Figure 3-Additional Model Diagnostics


Residuals vs Fitted
Residuals
Predicted values
tor(obese) ~ genhlth + toldhi2 + chccopd1 + educa + e


Normal Q−Q
Std. Pearson resid.
Theoretical Quantiles
tor(obese) ~ genhlth + toldhi2 + chccopd1 + educa + e





The first plot in figure 3 is the Residuals vs. Fitted Plot. It is not extremely informative in the logistic regression, but residuals can still be analyzed. The second plot is the normal QQ-plot, and shows standardized pearson residuals vs the theoretical quantiles. It can be a diagnosis method to check for the normality

assumption. The third plot is the scale-location plot, and is used to analyze the residuals like plot 1. The fourth plot is the residuals vs leverage plot, which is used to determine outliers.

Discussion

Discussion of Results

While analyzing the table 1(summary of the logistic regression), there are many interesting results. For example, for a general health value of "fair", versus the general health value of "excellent"(baseline category), the change in the log odds of obesity is 0.789. For the told high cholesterol predictor(toldhi2), the value of "Yes", versus the value of "no"(baseline category), the change in the log odds of obesity is 0.534.There are similar interpretations for the rest of the predictors shown in the table 1. It can be noted that the variables have very significant p-values, and low standard error values. The AIC value is 19477. For certain categorical variables, it seems that some categories have much more significance than other categories. For example, in marital status, the married category seems to be more significant for the model.

The table of odds ratio(table 2) leads to many important results, and important interpretation as well. For the ratio of odds of general health being "fair", versus the ratio of odds of general health being "excellent", the odds of being obese increase by a factor of 2.20. In "poor" health, it increases by a factor of 1.32. Looking at education, for the education level less than college (additional 1-4 years), it seems that the odds of being obese increase by greater factors. Looking at employment, the odds of being obese as an employed worker increase by a factor of 1.47 compared to the odds of being obese in the baseline category of "homemaker". That is an interesting result, and further investigation is needed. A few questions that arise are that if working people do not have enough time to manage their health, or if those who are homemaker are more active and fit during the day?

There are many interesting observatios in table 2. The odds of being obese as a male increase by the factor of 1.78, versus of the odds of being obese as the female(baseline category). Thus, it seems more attention should be given to the male sex, since it seems that they are at a greater risk of obesity. If one is experiencing difficulty walking, the odds of being obese increase by a factor of 1.52, compared to odds of not experiencing difficulty walking. The p-value for this predictor is also extremely significant, which I noticed during the AIC/BIC model selection process as well. Thus, this variable should be an important question to consider as an insurer or doctor. It seems that exercising increasing the odds of being obese by a factor of 0.8, compared to the odds of being obese while not exercising. Thus, it seems to beneficial, since the factor is <1, and should decrease the odds of obesity. Looking at martial status, being divorced has an odds of being obese increase by a factor of 1.26, compared to the odds of being obese as an unmarried couple baseline category. Being married increases the odds of being obese by a factor of 1.37, versus the baseline category of being an unmarried couple. For the vegetable variable, eating other vegetables has an odds ratio of 1, and this means that intake of vegetables can be equally likely for both study groups; individual being obese or not. This is quite interesting, since it seems according to scientific research, eating healthy has numerous benefits for the body. There should be further investigation, and more questions regarding junk food vs healthy food, to further understand the differences. This survey did have some questions regarding sugar, salt, and sugar drinks; however, they were not extremely significant in my model for prediction.

The prediction error is 67.2%, and that means my model is quite good for prediction. This is a high number for prediction, and the model is working quite well. The cross calibration plots show that the estimates are very close to the true value, since they are near the 45 degree line. It shows the model works well.

Discussion of Diagnostics

For the residuals vs fitted plot, it seems that the points are quite close to 0 in the range of -4 to 2. There seems to be a non-linear pattern, but if the plot is analyzed while zoomed out, the pattern does not seem too substantial. This plot is not very informative. Next, while analyzing the normal QQ-plot, there does

seem to be a minor pattern, which requires further investigation. Additionally, the errors still seem to be near the 45 degree line, which shows the errors are normally distributed. For the scale-location plot, the residuals again seem to be very close to 0, but there does seem to be some sort of pattern while looking at their red trendline. This can be attributed to the fact that this is a logistic regression. While analyzing the residuals vs leverage plot, there seems to be a big cluster of points near 0 on the x-axis, with very small y-values. There can be some cook's distance points seen, and a few points that are away from the cluster as one moves to the right on the x-axis. I did not remove these points that can seem to be outliers, because that would risk overfitting the model. There are too many points in the dataset, and the training set may categorize some as outliers, but they may not be in the real dataset of 465,046 cases, thus, these points were not removed.

Weaknesses and Next Steps

There were some weaknesses encountered during this analysis. There was a problem of missing data in some of the predictors. The predictors that had greater than 49% missing values were removed, since using imputation on these predictors seemed like they would lead to a false analysis in some cases. It could have been possible that some of these predictors were significant to the prediction model. For the predictors with less than 49% missing data, median or mode imputation was used. This was especially important in categorical variables; however, this substantially can decease the variance. The mean remains the same using this imputation, and it is simple and efficient. However, there are issues with assuming the median and mode can be used for all missing values, since it does not effectively account for other responses from the survey. For the final predictor variables, it seemed like this did not cause a big problem, but there can be further investigation on this topic. A possible next step to improve this model is to create another prediction model for the missing variables, and try to use that to obtain a more complete dataset for use in modelling. Additionally, mainly household numbers were contacted in the survey process. This was in 2013, but for an analysis like this in 2020, it could be the case that more people have made the switch to using mobile phone numbers, thus, this should be considered in further survey's and more weight could be given to mobile phone numbers.

The dataset after cleaning had about 465,065 cases. This was an extremely large dataset. Initially, a train set of 372,036 cases and test set of 93,009 cases had been created. However, these were too large to be time efficient. This process was stopped due to the fact it was too inefficient. At this point, there was research done on the glm function. There are numerous other methods to work with big data, that can perhaps be more time efficient, as it seems that the glm function cannot handle large amounts of data. It may take long hours to work on the full dataset. Thus, the train and test sets were made into smaller samples by using simple random sampling without replacement from the full dataset. The train set had 16,000 cases, and test set had 4,000 cases. This helped create an efficient process. There methods such as boosting, bagging, random forest, and advanced machine learning techniques that seem to work efficiently with big data, and can be able to be great at prediction. However, an important part of this analysis was to allow interpretation such that important decisions can be made after analyzing the results. That is why a generalized linear model was used, but comparing results with advanced methods can be very beneficial.

The 2013 survey had been made with very careful consideration and contribution by all the states in the USA. A few improvements could be made to these surveys. There could be more questions directed to youth, or questions that could ask adults about youth in the household. Obesity is becoming very common among youth, and solely focusing on adults may limit the results and may not prove effective. There could be more results obtained if youth situations were analyzed as well. The survey could have more questions asking about junk food intake, since there were quite a few questions on nutritional foods, but comparing to junk food intake and seeing the difference in health conditions would be very interesting. There were missing values as in any dataset, but it felt most responses were quite full. That was a very positive outcome of this survey, and it should be continued to keeping obtaining complete response. Additionally, there was a large dataset with rich predictors and numerous observations, which was very beneficial in the analysis.

In terms of the model, this analysis did not remove outliers. There were very few outliers available, and due to the small training set used compared to the big complete data set, it did not seem right to remove them,

due to the fact that the issue could be caused from the small sample of training data. Thus, a next step would be to compare a model with outliers removed, and compared the results, and determine if there are significant differences.

Appendix

The full logistic model is shown below.

```
##
## Call:
## glm(formula = as.factor(obese) ~ genhlth + toldhi2 + chccopd1 +
##     educa + employ1 + sex + diffwalk + smoke100 + exerany2 +
##     marital + vegetab1, family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2453  -1.2145   0.7062   0.9197   1.7476
##
## Coefficients:
##                                                  Estimate Std. Error z value
## (Intercept)                                     -0.4756471  0.1435367  -3.314
## genhlthFair                                      0.7737045  0.0701741  11.026
## genhlthGood                                      0.7711795  0.0525827  14.666
## genhlthPoor                                      0.3095397  0.0980463   3.157
## genhlthVery good                                 0.4598657  0.0489317   9.398
## toldhi2Yes                                       0.5217250  0.0383327  13.610
## chccopd1Yes                                     -0.2692489  0.0681670  -3.950
## educaCollege 4 years or more (College graduate) -0.3371930  0.0447054  -7.543
## educaGrade 12 or GED (High school graduate)     -0.0672944  0.0473452  -1.421
## educaGrades 1 through 8 (Elementary)             0.0021467  0.1195569   0.018
## educaGrades 9 though 11 (Some high school)      -0.0882976  0.0820449  -1.076
## educaNever attended school or only kindergarten  0.1457580  0.6753774   0.216
## employ1A student                                -0.3062977  0.1289604  -2.375
## employ1Employed for wages                        0.3563094  0.0747826   4.765
## employ1Out of work for 1 year or more            0.3567086  0.1275943   2.796
## employ1Out of work for less than 1 year          0.1493028  0.1274394   1.172
## employ1Retired                                   0.1209109  0.0764631   1.581
## employ1Self-employed                             0.1397631  0.0929176   1.504
## employ1Unable to work                            0.2914478  0.1024569   2.845
## sexMale                                          0.5686445  0.0376793  15.092
## diffwalkYes                                      0.4074043  0.0586794   6.943
## smoke100Yes                                     -0.1223825  0.0366214  -3.342
## exerany2Yes                                     -0.2039566  0.0434473  -4.694
## maritalDivorced                                  0.2257949  0.1158380   1.949
## maritalMarried                                   0.3253430  0.1089014   2.988
## maritalNever married                            -0.0557237  0.1144179  -0.487
## maritalSeparated                                 0.3310368  0.1619734   2.044
## maritalWidowed                                  -0.0034784  0.1185586  -0.029
## vegetab1                                         0.0004764  0.0001958   2.433
##                                                 Pr(>|z|)
## (Intercept)                                     0.000920 ***
## genhlthFair                                      < 2e-16 ***
```

```
## genhlthGood                                          < 2e-16 ***
## genhlthPoor                                         0.001594 **
## genhlthVery good                                     < 2e-16 ***
## toldhi2Yes                                           < 2e-16 ***
## chccopd1Yes                                         7.82e-05 ***
## educaCollege 4 years or more (College graduate) 4.61e-14 ***
## educaGrade 12 or GED (High school graduate)     0.155213
## educaGrades 1 through 8 (Elementary)            0.985674
## educaGrades 9 though 11 (Some high school)      0.281833
## educaNever attended school or only kindergarten 0.829130
## employ1A student                                0.017543 *
## employ1Employed for wages                       1.89e-06 ***
## employ1Out of work for 1 year or more           0.005180 **
## employ1Out of work for less than 1 year         0.241374
## employ1Retired                                  0.113810
## employ1Self-employed                            0.132540
## employ1Unable to work                           0.004447 **
## sexMale                                              < 2e-16 ***
## diffwalkYes                                         3.84e-12 ***
## smoke100Yes                                         0.000832 ***
## exerany2Yes                                         2.67e-06 ***
## maritalDivorced                                 0.051268 .
## maritalMarried                                  0.002813 **
## maritalNever married                            0.626245
## maritalSeparated                                0.040976 *
## maritalWidowed                                  0.976594
## vegetab1                                        0.014967 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 20657  on 15999  degrees of freedom
## Residual deviance: 19413  on 15971  degrees of freedom
## AIC: 19471
##
## Number of Fisher Scoring iterations: 4
```

This is a list of missing values observed in the variables of the dataset.

```
##                  miss
## X_state   0.000000e+00
## pvtresd1  2.676753e+01
## colghous  1.000000e+02
## nummen    2.677037e+01
## numwomen  2.677037e+01
## genhlth   4.036399e-01
## physhlth  2.228051e+00
## menthlth  1.754258e+00
## poorhlth  4.944395e+01
## hlthpln1  3.871689e-01
## persdoc2  3.662244e-01
## medcost   2.482843e-01
## checkup1  1.282090e+00
```

```
## sleptim1  1.502110e+00
## bphigh4   2.887499e-01
## bpmeds    5.962096e+01
## bloodcho  1.832139e+00
## cholchk   1.488038e+01
## toldhi2   1.457211e+01
## cvdinfr4  5.260536e-01
## cvdcrhd4  8.993950e-01
## cvdstrk3  2.983072e-01
## asthma3   3.170149e-01
## asthnow   8.671344e+01
## chccopd1  5.539118e-01
## addepev2  4.654568e-01
## diabete3  1.691831e-01
## veteran3  1.516954e-01
## marital   6.954400e-01
## children  4.624066e-01
## educa     4.624066e-01
## employ1   6.885263e-01
## income2   1.452412e+01
## weight2   4.204362e+00
## height3   1.550099e+00
## renthom1  1.717859e+00
## sex       1.423415e-03
## qlactlm2  2.303696e+00
## decide    2.641452e+00
## diffwalk  2.595496e+00
## diffdres  2.464948e+00
## diffalon  2.714656e+00
## smoke100  3.033908e+00
## smokday2  5.632312e+01
## stopsmk2  8.447379e+01
## lastsmk2  7.208520e+01
## usenow3   2.850491e+00
## alcday5   3.994510e+00
## avedrnk2  5.298968e+01
## drnk3ge5  5.295999e+01
## maxdrnks  5.387789e+01
## fruitju1  7.362717e+00
## fruit1    6.872655e+00
## fvbeans   7.623812e+00
## fvgreen   7.149001e+00
## fvorang   7.434904e+00
## vegetab1  7.919679e+00
## exerany2  6.919628e+00
## exract11  3.272798e+01
## exeroft1  3.338214e+01
## exerhmm1  3.424229e+01
## exract21  3.376320e+01
## exeroft2  5.657730e+01
## exerhmm2  5.767902e+01
## strength  8.096996e+00
## pdiabtst  5.433318e+01
## prediab1  5.244533e+01
```

```
## diabage2  9.306492e+01
## insulin   9.259926e+01
## bldsugar  9.272676e+01
## painact2  9.990077e+01
## qlmentl2  9.990016e+01
## qlstres2  9.990097e+01
## qlhlth2   9.990544e+01
## hlthcvrg  3.576493e+01
## drvisits  3.134604e+01
## medscost  2.900208e+01
## medbills  2.934208e+01
## ssbsugar  7.901581e+01
## ssbfrut2  7.906116e+01
## wtchsalt  7.396635e+01
## longwtch  8.564364e+01
## dradvise  7.397448e+01
## cvdasprn  7.230075e+01
## rlivpain  9.050074e+01
## rrclass2  9.920594e+01
## rrcognt2  9.921692e+01
## rratwrk2  9.968746e+01
## rrhcare3  9.931127e+01
## rrphysm2  9.919821e+01
## rremtsm2  9.919841e+01
## misnervs  9.265009e+01
## mishopls  9.264704e+01
## misrstls  9.267206e+01
## misdeprd  9.265802e+01
## miseffrt  9.270805e+01
## miswtles  9.267002e+01
## misnowrk  9.269747e+01
## mistmnt   9.268080e+01
## mistrhlp  9.297870e+01
## misphlpf  9.299639e+01
## scntmony  8.736028e+01
## scntmeal  8.656296e+01
## scntwrk1  9.341935e+01
## lsatisfy  9.762676e+01
## mscode    2.778567e+01
## X_bmi5cat 5.434803e+00
## X_rfbmi5  5.435209e+00
## obese     5.435209e+00
```

References

- "CDC - BRFSS 2013 Survey Data and Documentation. 2013 BRFSS Overview." Centers for Disease Control and Prevention, Behavioral Risk Factor Surveillance System , 23 July 2013, www.cdc.gov/brfss/annual_data/annual_2013.html.
- "CDC - BRFSS 2013 Survey Data and Documentation. 2013 BRFSS Codebook Report." Centers for Disease Control and Prevention, Behavioral Risk Factor Surveillance System , 23 July 2013, www.cdc.gov/brfss/annual_data/annual_2013.html.

- Daniel D. Sjoberg, Michael Curry, Margie Hannum, Karissa Whiting and Emily C. Zabor (2020). gtsummary: Presentation-Ready Data Summary and Analytic Result Tables. R package version 1.3.5. https://CRAN.R-project.org/package=gtsummary
- Frank E Harrell Jr (2020). rms: Regression Modeling Strategies. R package version 6.0-1. https://CRAN.R-project.org/package=rms
- Hao Zhu (2020). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.2.1. https://CRAN.R-project.org/package=kableExtra
- H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
- JJ Allaire and Yihui Xie and Jonathan McPherson and Javier Luraschi and Kevin Ushey and Aron Atkins and Hadley Wickham and Joe Cheng and Winston Chang and Richard Iannone (2020). rmarkdown: Dynamic Documents for R. R package version 2.3. https://rmarkdown.rstudio.com.
- LPLenka. 2020, November. BRFSS Data. The Behavioral Risk Factor Surveillance System (BRFSS) 2013. Version 3.https://www.kaggle.com/lplenka/brfss-data.
- Matthijs Meire, Michel Ballings and Dirk Van den Poel (2016). imputeMissings: Impute Missing Values in a Predictive Context. R package version 0.0.3. https://CRAN.R-project.org/package=imputeMissings
- "Obesity and Overweight." World Health Organization, World Health Organization, 1 Apr. 2020, www.who.int/news-room/fact-sheets/detail/obesity-and-overweight.
- "Obesity in Canada." Public Health Agency of Canada(PHAC), and Canadian Institute for Health Information(CIHI). Canada.ca, Government of Canada, 23 June 2011, www.canada.ca/en/public-health/services/health-promotion/healthy-living/obesity-canada.html.
- R Core Team (2020). R: A language and environment for statistical computing. R, Foundation for Statistical Computing, Vienna, Austria. URL, https://www.R-project.org/.
- Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686