**Obesity in the US: Model Comparisons Using Health and Demographics Predictors**

Authors: Evan Adams, Steve Bramhall, Andy Heroy, Adam Scheerer

**INTRODUCTION**

Obesity is a growing concern in the United States and is now considered a chronic disease by the American Medical Association, the American Association of Clinical Endocrinologists, the Obesity Society, the American Society of Bariatric Physicians, and the National Institute of Health [1]. Obesity is defined as having a body mass index (BMI) of 30kg/m2. There are many studies and statistics regarding various demographic and health contributors to BMI. This study uses the national dataset from the 2017 County Health Rankings (CHR) website [2]. CHR gathers various data related to health from different data sources/organizations.

Our approach to the project started with a review of the data and identifying variables of interest that include both health and demographic type variables. Then we performed some exploratory analysis to get a better understanding of the data and their relationships to each other. We ended problem 1 with a logistic regression using a LASSO technique, after reducing a number of different variables in the EDA to follow.

**DATA DESCRIPTION**

The CHR data includes various variables related to health by state and county. Since each state and county have varying populations, we chose to use data indicating the percentage of the population representing a certain health aspect. These values are not divided by 100 so the range is 0 to 100. All variables are continuous, so there is no concern for unbalanced categorical variables. The team decided to start with the following initial variables to model the probability of being obese (BMI > 30).

Health Variables

* Smokers
* Physically Inactive
* Excessive Drinking
* Frequent Mental Distress
* Frequent Physical Distress
* Diabetic
* Insufficient Sleep

Demographic Type Variables

* Uninsured
* Some College
* Unemployed
* Severe Housing Problems

**EXPLORATORY ANALYSIS**

We review some summary statistics first. The population has almost twice as many obese than non-obese.

|  |  |
| --- | --- |
|  |  |

Figure - Variable Summary by Obese Class

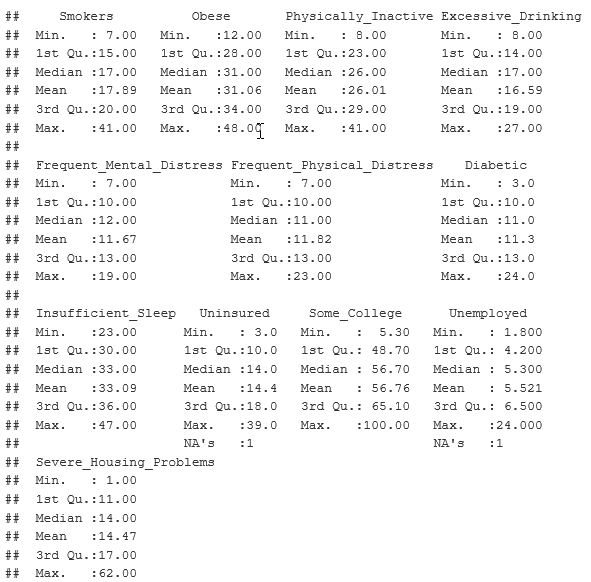


Figure - Summary Statistics, Shows Missing Data

The summary statistics show only one missing value for Unemployed and Uninsured. We impute the missing data with the median values since some distributions show a little skew when comparing the mean and medians. Figure 3 shows the histograms confirming some skew. This is not a concern since we have a relatively large data set and we are performing a logistic regression.

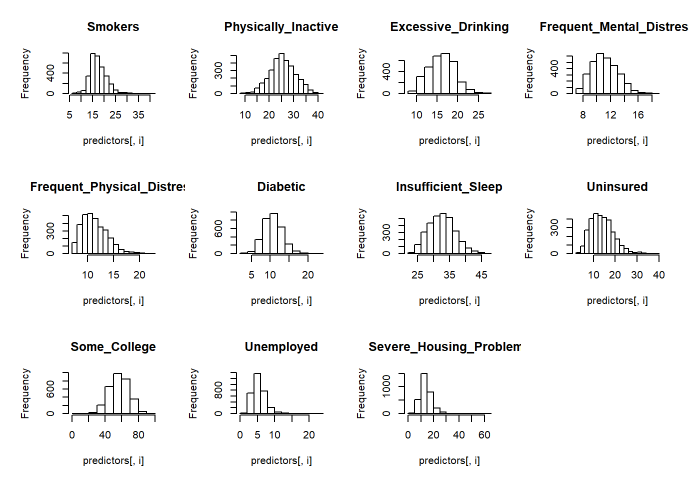


Figure - Histogram of Variables

We use SAS to help gain information about possible influential data points. The Cook’s D values were relatively low for all variables. However, there were two variables that had a few points with higher Cook’s D values compared to their respective observations. Those two variables were Unemployed and Severe\_Housing\_Problems. Figure 4 shows the fit diagnostics, box plots, and R boxplot information for Unemployed. Figure 5 shows the same for Severe\_Housing\_Problems (Renamed Housing\_Prob for short). The fit diagnostics for the other variables can be found in the Appendix 1

|  |  |
| --- | --- |
|  |  |

Figure - Fit Diagnostics and Box Plots for Unemployed

|  |  |
| --- | --- |
|  |  |

Figure - Fit Diagnostics and Box Plots for Severe\_Housing\_Problems

After further investigations, we removed 2 outliers for the unemployed variable. One of the outliers belonged to Yuma county in Arizona. The county is along the Mexico border and is predominately a farming community with migrant (seasonal) workers. This situation is uncommon and not typical of U.S. counties. We also remove the data for Imperial county in California for the same reasons. It is adjacent to Yuma county.

We removed 3 outliers for severe housing problems. We removed the data for Bethel, Northwest Arctic and Yukon-Koyukuk counties in Alaska. There are four factors that contribute to this category. They are housing units that lack complete kitchens, lack complete plumbing facilities, overcrowded, or severely cost burdened. These counties reside in Alaska where the cost to build is beyond what the residents can afford and therefore overcrowding is above normal compared to the rest of the United States. [Nathan Wiltse, Dustin Madden, 2018 Alaska Housing Assessment, Jan 17, 2018 [3]. The evidence for outlier removals are shown in Figure 6 and 7.

|  |  |
| --- | --- |
| Before Outlier Removal | After Outlier Removal |
|  |  |

Figure - Unemployed Outlier Before & After Plots

|  |  |
| --- | --- |
| Before Outlier Removal | After Outlier Removal |
|  |  |

Figure - Severe Housing Problems Outlier Before & After Plots

Figure 8 shows a scatter matrix of the variables colored by obese class.

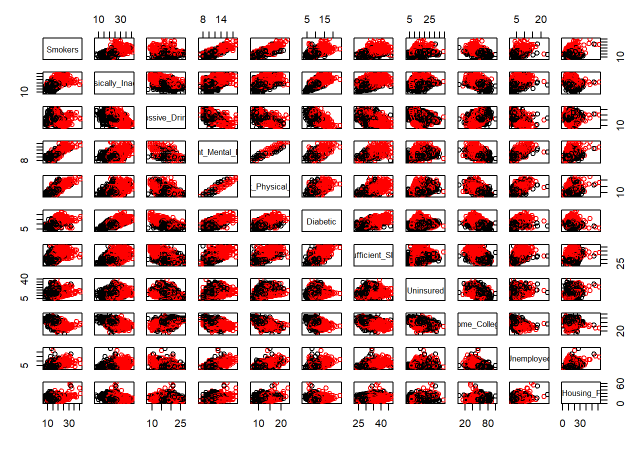


Figure - Scatter Plot Matrix

The scatter plot matrix shows strong correlation with the following:

* Frequenet\_Mental\_Distress, Frequent\_Physical\_Distress

There’s some visual correlation seen between the following:

* Smokers, Frequent\_Mental\_Distress, Frequent\_Physical\_Distress
* Diabetic, Physically\_Inactive, Insufficient Sleep, Frequent\_Mental\_Distress, Frequent\_Physical\_Distress.

The correlation heatmap in Figure 9 provides some addition insights.

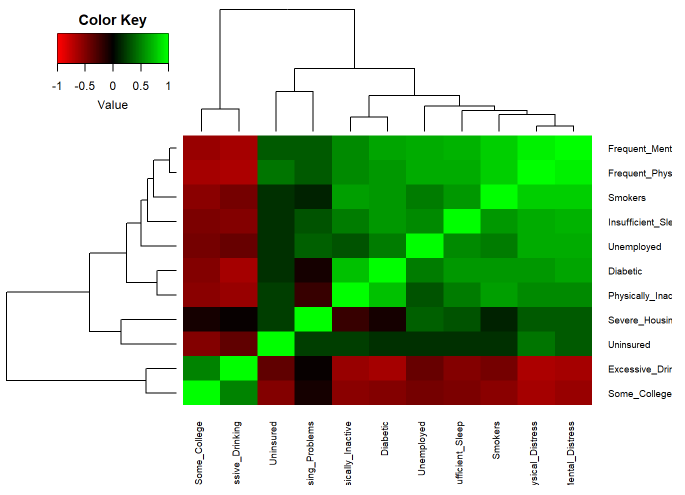


Figure - Correlation Heatmap

The dendogramed heatmap confirms the strong correlation previously seen with Frequenet\_Mental\_Distress and Physical\_Mental\_Distress.

Additional correlation is seen between the following:

* Unemployed, Insufficient Sleep
* Some\_College, Excessive\_Drinking
* Diabetic, Physically\_Inactive
* Smokers, Frequenet\_Mental\_Distress, Frequent\_Physical\_Distress
* Uninured, Severe\_Housing\_Problems

The correlations identified by the dendogram surprisingly all make practical sense. One would expect to lose sleep if they were unemployed. Drinking being correlated to college makes sense. Diabetic is not uncommon amongst physically inactive people. If someone is living in an area with severe housing problems, we might expect they would not be able to afford insurance.

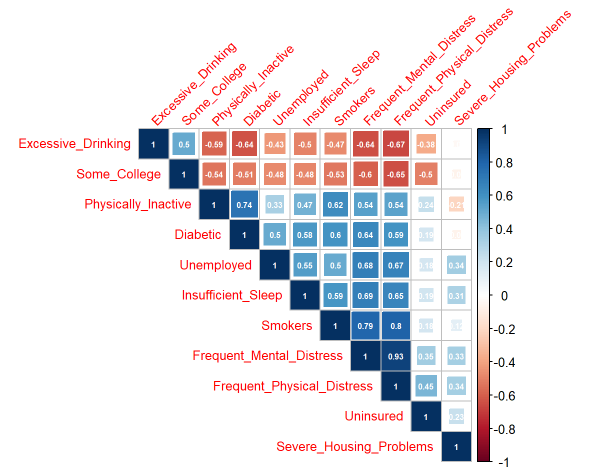


Figure - Correlation Heatmap

Based on the variable correlation heatmap in Figure 10, the order of correlated variables are:

1. Frequenet\_Mental\_Distress, Frequent\_Physical\_Distress
2. Smokers, Frequenet\_Physical\_Distress
3. Smokers, Frequenet\_Mental\_Distress
4. Diabetic, Physically\_Inactive
5. Frequent\_Mental\_Distress, Insufficient Sleep
6. Unemployed, Frequent\_Mental\_Distress
7. Unemployed, Frequent\_Physical\_Distress
8. Excessive\_Drinking, Frequent\_Physical\_Distress
9. Excessive\_Drinking, Frequent\_Mental\_Distress
10. Diabetic, Frequent\_Mental\_Distress

The variable inflation factors (VIFs) shown in Figure 11 and previous visual tools agree there is a strong relationship between Frequent\_Mental\_Distress and Frequent\_Physical\_Distress. We choose to remove Frequent\_Physical\_Distress.

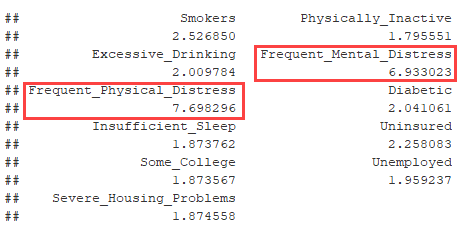


Figure - Variable Inflation Factors

We use PCA to visualize any other insights. It is fortunate to already have our data somewhat normalized on a percentage scale. It reduces the scale sensitivity seen with PCA. Figure 12 shows the PCA values.

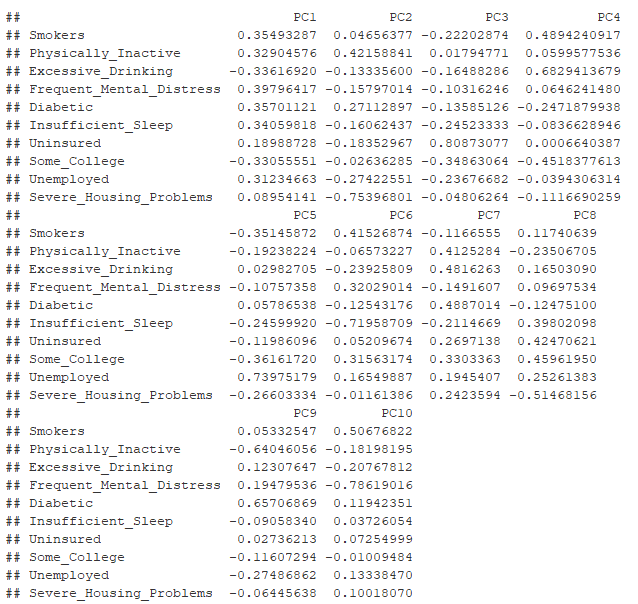


Figure - PCA Values

The Scree and Cumulative Proportion plots are shown in Figure 13. The plots show 6 PCs are needed to retain 90% of the total variation in the data.

|  |  |
| --- | --- |
|  |  |

Figure - Scree and Cumulative Proportion Plots

Next, we review the first few PCA plots shown in Figure 14.

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |

Figure - PCA Plots to View Separation

The PCA plots for PC1 and PC2 vs all PCs up to PC6 show fairly good separation. This tells us we should expect decent results if we used these for our model. However, we will use the original variables for our model. The PCA plots may show some possible outliers. However, our previous analysis for outliers using the Cook’s D information do not indicate any addition concerns driving further investigations. In other words, the possible outliers seen are not influential enough to drastically affect our modeling.

**PROBLEM STATEMENT**

We want to model the probability of being obese (BMI > 30) for the following variables.

1. Smokers
2. Physically Inactive
3. Excessive Drinking
4. Frequent Mental Distress
5. Diabetic
6. Insufficient Sleep
7. Uninsured
8. Some College
9. Unemployed
10. Severe Housing Problems

Our hypothesis statements are shown below.

Ho: There is no relationship between the predictor variables and whether someone is obese or not.

Ha: There is a relationship between the predictor variables and whether someone is obese or not.

**MODEL SELECTION**

We choose a logistic regression using LASSO for feature reduction.

Our first pass for a logistic regression using LASSO yields the results shown in Figure 15.

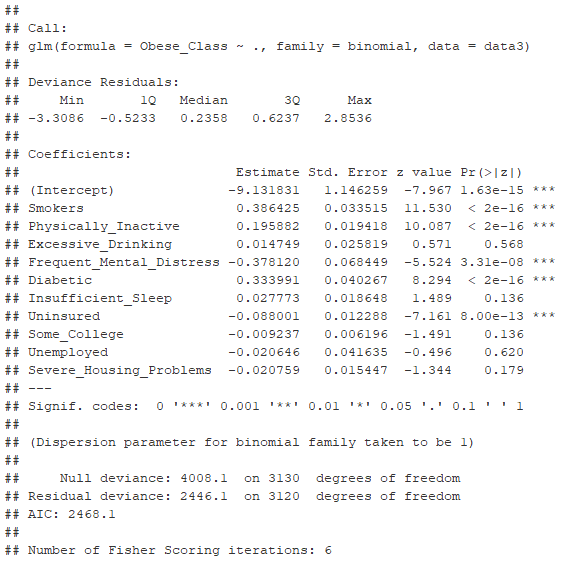


Figure - First Pass Results for Logistic Regression Using LASSO Feature Reduction

With p-values greater than 0.05, the model suggests removing the following insignificant predictors:

* Excessive Drinking
* Insufficient Sleep
* Some College
* Unemployed
* Severe Housing Problems

We tested other logistic regression libraries and data splits and agreement was seen with the above recommendations so we chose to remove the above predictors and reran the model.

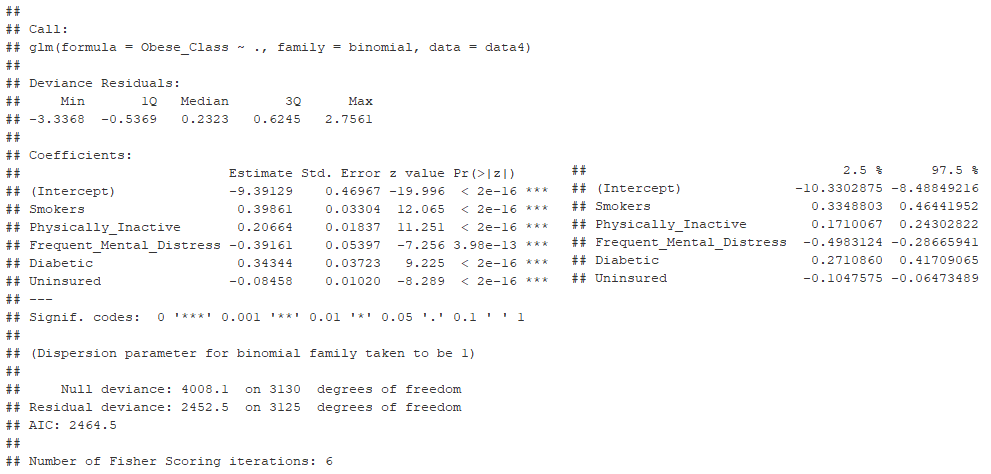


Figure - Final Results for Logistic Regression Using LASSO Feature Reduction

**Model Assumption Check**

During our EDA, removed a collinear variable give us independence with our predictor variables. We also identified some influential points and justification for removing those points. There is some skew with our data but since we have a fairly large dataset and we performed a logistic regression, there are no concerns with the skew. We did not have any categorical predictors so there is no need to perform a Chi square goodness of fit test.

Here, we choose the receiving operating characteristic (ROC) as our first measure of classier performance. Figure 17 shows the ROC curve for our base logistic regression model.

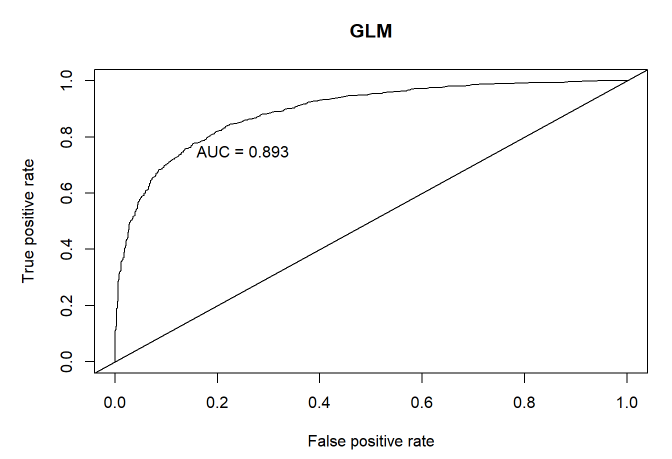


Figure - ROC Curve

Our area under the curve (AUC) = 0.885. Since this is about 0.8 our model does a good job discriminating between obese and not obese.

A logistic regression model provides a better fit to the data if it demonstrates an improvement over a model with fewer predictors. We used the likelihood ratio test. We create a model with two key predictors and compare against our base logistic regression model (“full model”). Figure 18 shows the results of our likelihood ratio test.

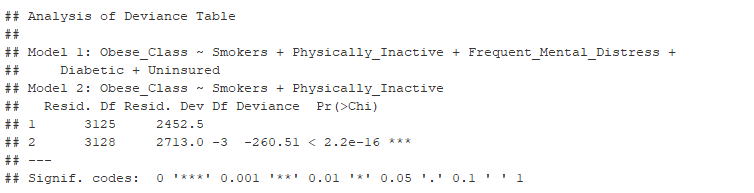


Figure - Likelihood Ratio Test Results

Ho: The reduced model glm.fitless is favored over a more full model (glmfit).   
Ha: The reduced model glm.fitless is not favored over a more full model (glm.fit).

We reject Ho. With an alpha of 0.05, the results show the observed difference in model fit is statistically significant with a p-value < 2.2e-16. The evidence suggests the glm.fit “full model” is favored.

**Interpretation**

Our final fitted model for Problem 1 is shown below.

log(p/(1-p)) = logit(p) = -10.18016 + 0.42933(Smokers) + 0.26783(Physically\_Inactive) - 0.37621(Frequent\_Mental\_Distress) + 0.03314(Insufficient\_Sleep)

Ho: There is no relationship between the predictor variables and whether someone is obese or not.   
Ha: There is a relationship between the predictor variables and whether someone is obese or not.

We reject the null hypothesis with multiple predictors having p-values < 0.05.

There is sufficient evidence at the alpha = 0.05 level of significance to suggest that there is a relationship between predictor variables and whether someone is more prone to being obese. Interpretation of the coefficients below describe each of the variables influence on our regression model.

Beta0 = -10.1812 with 95% CI (-11.1835,-9.2105) The odds of being obese with all variables =t 0 is exp(-10.1812)=0.000038.

Beta1 = 0.42933 with a 95% CI (0.37007,0.49044) This fitted model says, holding all other variables constant, the odds of being obese for smokers is exp(0.42933)=1.5362 over the odds of not being obese. In terms of percent change, the odds for being obese are 53.62% higher than the odds for not being obese.

Beta2 = 0.26783 with a 95% CI (0.23894, 0.29762) The coefficient for Physically\_Inactive, holding all other variables constant, says the odds of being obese is exp(0.26783)=1.3071 over the odds of not being obese. In terms of percent change, the odds for being obese are 30.71% higher than the odds for not being obese.

Beta3 = -0.37621 with a 95% CI (0.-45264, -0.302) The coefficient for Physically\_Inactive, holding all other variables constant, says the odds of being obese is exp(-0.37621)=0.6865 over the odds of not being obese. In terms of percent change, the odds for being obese are 31.35% lower than the odds for not being obese.

Beta4 = 0.03314 with a 95% CI (0.00246, 0.06386) The coefficient for Physically\_Inactive, holding all other variables constant, says the odds of being obese is exp(0.03314)=1.0337 over the odds of not being obese. In terms of percent change, the odds for being obese are 3.37% higher than the odds for not being obese.

**CONCLUSIONS**

The team identified seven health variables and four demographic variables to analyze and evaluate for modeling the probability of being obese. After EDA and using LASSO for feature reduction our model ended up with four health variables and one demographic variable. The variables are shown below.

Health Variables

* Smokers
* Physically Inactive
* Frequent Mental Distress
* Diabetic

Demographic Type Variables

* Uninsured

It is possible the model could be improved with additional predictors, but the data chosen did not have all possible variables currently collected by the CHR. Data could be brought in from other sources not currently targeted by the CHR. Our analysis targeted obesity, so our variable selection was centered around those that affected an average US. resident’s everyday life. Thus, through common sense and our LASSO regression, we arrived at the variables above. Our study would be classified as observational, as collection methods of each individual county are likely to be different. There is, however, significant evidence (p<<0.5) that obesity is related to the Health and Demographic variables to predict obesity in the United States.

**OBJECTIVE 2**

Now that we’ve gone through Problem 1 and fit our base model, its time to see if we can improve upon that accuracy with some different analytic approaches. We ran another logistic regression with an interaction term, a KNN, and random forest to compare to our baseline model. For simplicities sake, we’ll start with the most automated of process’, the random forest.   
  
**Model Selection**

We begin our second analysis with the simplest regression which is random forest. We will include all relevant variables to be included with the dataset.

**REFERENCES**

1. Obesity Facts

[<https://www.endocrineweb.com/conditions/obesity/obesity-america-growing-concern>](https://www.nimh.nih.gov/health/statistics/suicide.shtml)

1. County Health Rankings & Roadmaps Data Source

<http://www.countyhealthrankings.org/explore-health-rankings/rankings-data-documentation/national-data-documentation-2010-2017>

1. <https://www.ahfc.us/download_file/view/5124/853>

**APPENDIX**

1. SAS CODE above

|  |  |
| --- | --- |
| proc means data=data2;  var Smokers Phys\_Inactive Excess\_Drinking Mental\_Stress Phys\_Stress Diabetic Insufficient\_Sleep Uninsured College Unemployed Housing\_Prob;  by Obese;  run; | |
|  |  |

The population has almost twice as many obese people than non-obese.

|  |  |
| --- | --- |
| proc glm data=data2 plot=diagnostics;  class Obese;  model Smokers Phys\_Inactive Excess\_Drinking Mental\_Stress Phys\_Stress Diabetic Insufficient\_Sleep Uninsured College Unemployed Housing\_Prob = Obese;  manova h = Obese / printe printh summary;  run; | |
| SMOKERS | |
|  |  |

The smokers distributions shows a little skew and with all observations having a Cook’s D under 0.0125, there are no influential points to be concerned about.

|  |  |
| --- | --- |
| PHYSICALLY INACTIVE | |
|  |  |

The physically inactive distribution is quite normal and all Cook’s D values are very low, under 0.008 so no influential points to worry about.

|  |  |
| --- | --- |
| EXCESS DRINKING | |
|  |  |

The excess drinking distribution is quite normal and all Cook’s D values are very low, under 0.005 so no influential points to worry about.

|  |  |
| --- | --- |
| MENTAL STRESS | |
|  |  |

The mental stress distribution is quite normal and all Cook’s D values are very low, under 0.004 so no influential points to worry about.

|  |  |
| --- | --- |
| PHYSICAL STRESS | |
|  |  |

The physical stress distribution exhibits a slight skew and all Cook’s D values are very low, under 0.01 so no influential points to worry about.

|  |  |
| --- | --- |
| DIABETIC | |
|  |  |

The diabetic distribution is fairly normal with a slight skew and all Cook’s D values are very low, under 0.008 so no influential points to worry about.

|  |  |
| --- | --- |
| INSUFFICIENT SLEEP | |
|  |  |

The insufficient sleep distribution is quite normal with a slight skew and all Cook’s D values are very low, under 0.007 so no influential points to worry about.

|  |  |
| --- | --- |
| UNINSURED | |
|  |  |

The uninsured distribution is fairly normal with a slight skew and all Cook’s D values are very low, under 0.008 so no influential points to worry about.

|  |  |
| --- | --- |
| COLLEGE | |
|  |  |

The college distribution is fairly normal with a slight skew and all Cook’s D values are very low, under 0.012 so no influential points to worry about.

|  |  |
| --- | --- |
| UNEMPLOYED | |
|  |  |

The unemployed distribution has some skew and possibly a couple influential points that will be reviewed.

|  |  |
| --- | --- |
| HOUSING PROBLEMS | |
|  |  |

The housing problems distribution has some skew and possibly a couple influential points that will be reviewed.

We remove 2 outliers for unemployed. We remove the data for Yuma county in Arizona due to the outlier it creates for unemployment. The county is along the Mexico border and is predominately a farming community with migrant (seasonal) workers. This situation is uncommon and not typical of U.S. counties. We also remove the data for Imperial county in California for the same reasons. It is adjacent to Yuma county.

We remove and 3 outliers for housing problems. We remove the data for Bethel, Northwest Arctic and Yukon-Koyukuk counties in Alaska for Severe Housing Problems. There are four factors that contribute to this category. They are housing units that lack complete kitchens, lack complete plumbing facilities, overcrowded, or severely cost burdened. These counties reside in Alaska where the cost to build is beyond what the residents can afford and therefore overcrowding is above normal compared to the rest of the United States. [Nathan Wiltse, Dustin Madden, 2018 Alaska Housing Assessment, Jan 17, 2018, https://www.ahfc.us/download\_file/view/5124/853]

|  |  |
| --- | --- |
|  |  |

After removing the outliers, there is no change in the fact that the population has almost twice as many obese people than non-obese.

|  |  |
| --- | --- |
| UNEMPLOYED | |
|  |  |

After removing the outliers, the unemployed distribution does not change. It still has some skew. All observations have a Cook’s D below 0.22 and this is deemed acceptable.

|  |  |
| --- | --- |
| HOUSING PROBLEMS | |
|  |  |

After removing the outliers, the housing problems distribution does not change. It still has a slight skew. All observations have a Cook’s D below 0.0125 and this is deemed acceptable.

We proceed with the dataset omitting the five data points.

1. Placeholder
2. Placeholder