**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 built a logistic regression model then….

**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 1 - Variable Summary by Obese Class

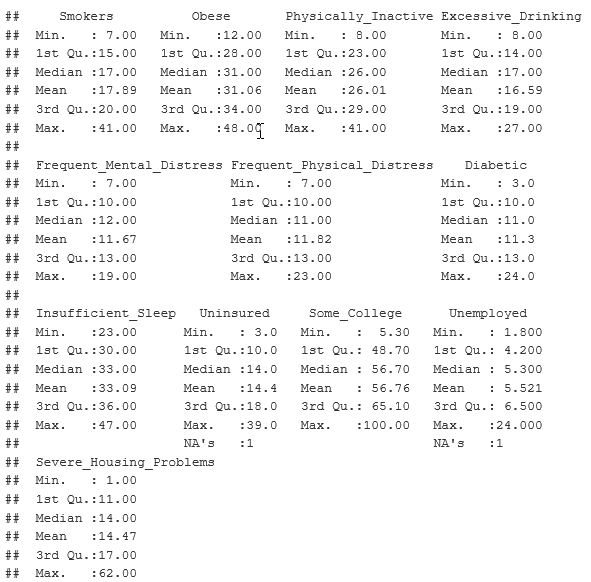


Figure 2 - Summary Statistics, Shows Missing Data

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.

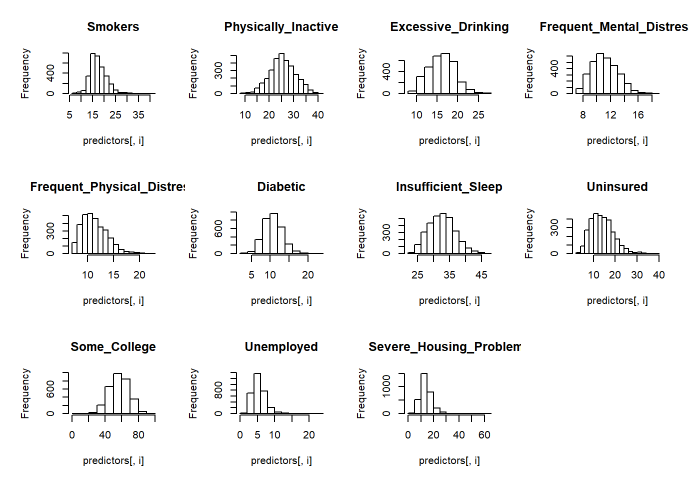


Figure 3 - Histogram of Variables

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 variable can be found in the Appendix.

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

Figure 4 - Fit Diagnostics and Box Plots for Unemployed

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

Figure 5 - 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 6 - Unemployed Outlier Before & After Plots

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

Figure 7 - Severe Housing Problems Outlier Before & After Plots

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

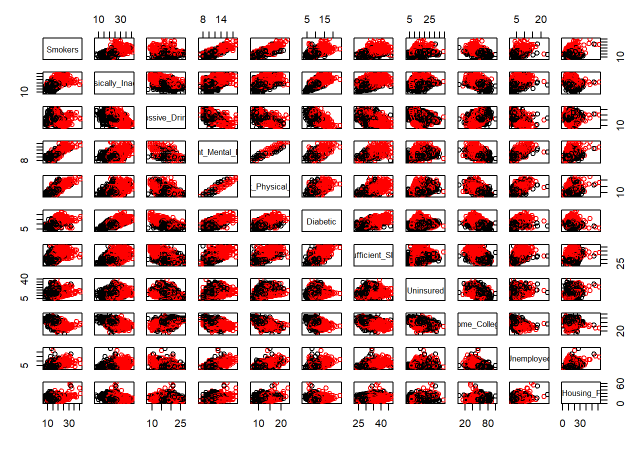


Figure 8 - Scatter Plot Matrix

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.

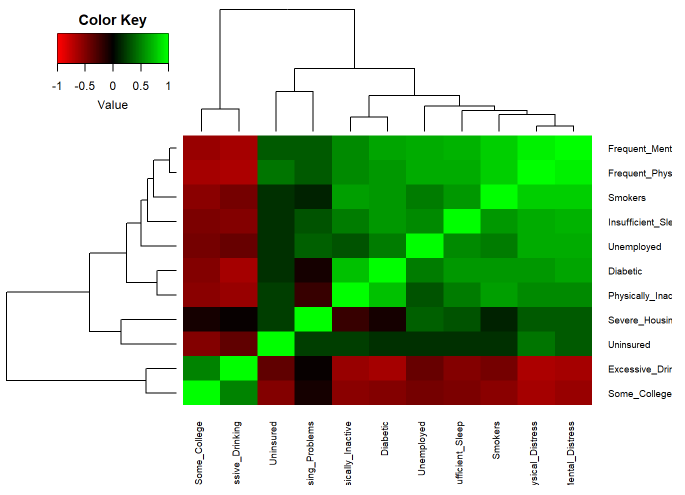


Figure 9 - Correlation Heatmap

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.

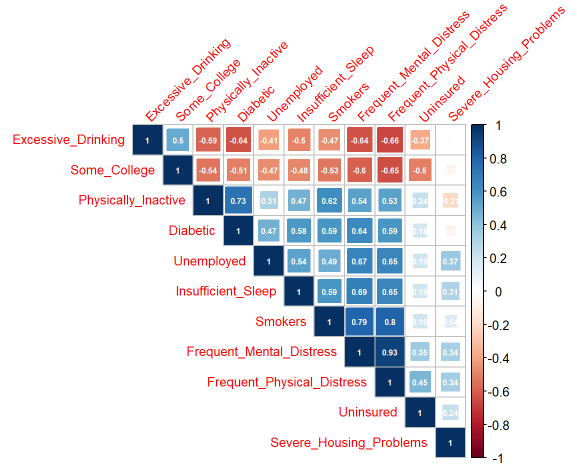


Figure 10 - Correlation Heatmap

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.

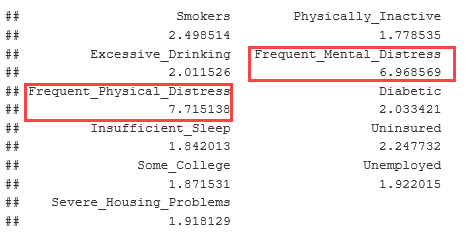


Figure 11 - Variable Inflation Factors

**PROBLEM STATEMENT**

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

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

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

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

**MODEL SELECTION**

We choose a

**Model Assumption Check**

**Interpretation**

**CONCLUSIONS**

**NOTES:**

PCA is just for EDA, basically just looking for separation (is my response separating out), are there outliers to investigate….don’t use for a model

**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**

SAS CODE

|  |  |
| --- | --- |
| 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.