**Obesity Rate in the United States**

William Wu, Pavan Bapesh Vinnakota, Olivia Liang, Patrick Seng

**Introduction:**

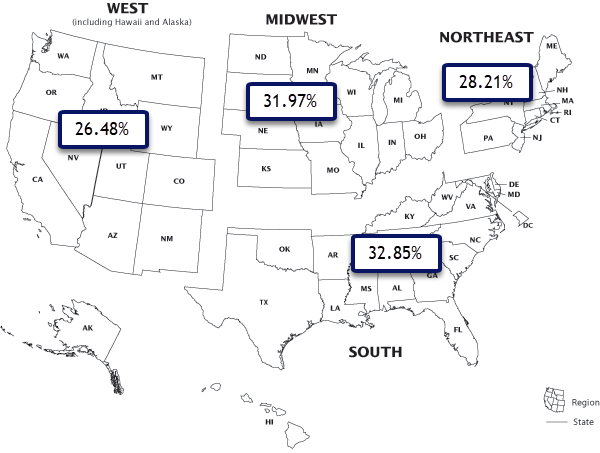
The obesity epidemic has become the emblematic problem of our time. With economic growth fueling easy access of processed food, better healthcare and sedentary lifestyles, obesity rates in the USA have skyrocketed from 30.5 percent in 1999-00 to 39.6 percent in 2015-16[[1]](#footnote-0). The estimated annual medical cost of obesity in 2015-16 in the United States was $147 billion[[2]](#footnote-1). Rising prevalence of obesity places a continued burden on the quality of life and the healthcare system across the nation.

The purpose of our project was to identify which factors are related to obesity rates across the United States to better understand what type of action might be taken to address this issue. We identified county level variables across various public datasets which could be related to obesity rates. We obtained the obesity rate data from cdc.gov, and the numbers of fast-food restaurants, full-service restaurants, and fitness centers from the Department of Agriculture’s Food Environment Atlas. Finally, population counts and other socio-economic factors which influence consumption were collected from US Census Data projections.

The next step was to map data from all of the datasets at a county level. We used county names and state names to match the datasets, while removing data unavailable in any particular dataset. This removal was minimal and primarily was related to US territories such as Puerto Rico, Guam etc. To account for the population differences across the various counties, we calculated all the interval variables per capita for 1000 residents. This ensured that the data was comparable between the various counties, while ensuring the numbers were not too low to interpret. We noticed that the response variable of obesity rate followed a normal distribution as expected with country level datasets. The range of data was between (12.7% - 48.0%) while the mean obesity rate at a county level was 31.33%.

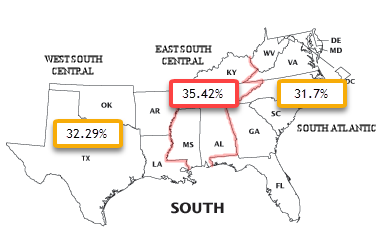
**Analysis:**

To begin our analysis, we explored how different regions of the country compared to one another. The purpose of this was to identify which areas showed unusually high rates of obesity, providing us with direction on where we might first look to implement changes to try and reduce the obesity rate in the country. We divided the United States into four regions defined by the US Census Bureau: West, South, Midwest, and Northeast, calculating the mean obesity rate for each region (Appendix 1 & Figure 1). From here, we were able to determine that there was a significant difference in regions due to the extremely low probability of seeing these results in the data if there was no relationship (Appendix 2). We then analyzed the specific differences between each region and found that and found that every region was in fact significantly different from one another (Appendix 3). Although their means appear relatively similar at a glance, the differences made sense, as there were a large number of counties per region.



**Figure 1 - Regional Obesity Rates**

Next, we used the census region separation again to segment the South region into the West South Central, East South Central, and South Atlantic sub regions. We found that there was a significant difference between these regions (Appendix 4). When comparing sub-regions to each other, we found that the West South Central and Atlantic sub regions both had significantly lower obesity rates than the East South Central region (Figure 2). This was interesting because we did not find overwhelming evidence that the two outer sub regions (West South Central and Atlantic) significantly differed from one another (Appendix 5). There was a 5.5% chance we would see the result we observed in our sample had there been no difference between the East and West most sub regions while there was a less than 1% we would see this result if they did not differ from the center most, East South Central sub region (Appendix 5).



**Figure 2 - South Sub Region Obesity Rates**

Before conducting a linear regression analysis, we explored collinearity between our predictor variables to ensure the integrity of our results. We found that the correlation between the median income and the poverty rate is below -70%, which indicated we must drop one of them so that their collinearity did not affect the accuracy of the model. The complete correlation matrix and plots are in Appendix 6. Although there was additional correlation between predictor variables present, we decided to include them in our initial model and proceed with caution.

We utilized multiple regression to explore which factors affect the obesity rate in the United States. Our null hypothesis was that there is no relationship between our predictor variables and the obesity rate. The first multi-regression model included all of our gathered predictor variables excluding poverty rate with obesity rate as the response variable, referenced in Appendix 7 (Figure 3). After reviewing the results of our initial model, with all predictors included, we found that the Grocery Stores per 1000 Residents variable was insignificant in explaining variability in obesity rates across the country and subsequently removed this predictor variable as we moved to our next model. The p-value of our next model, Model A, provided evidence of a relationship between our predictors and the obesity rate and explained 63% of the variation of the obesity rate across the country (Figure 3 & Appendix 8). Still, we were concerned that some of the previously mentioned pairs of predictor variables with moderate correlations were causing the model to be overfitted. To explore if removing these variables decreased error, we dropped Full Service Restaurants and Fitness Centers per 1000 Residents due to their high correlation with Fast Food per 1000 Residents to create our next model, Model B. This model explained slightly less of the variation in the obesity rate and introduced more error into the model (Appendix 7). The p-values of the predictor variables removed also indicated that they were significant. We decided to move forward with Model A, while acknowledging the sacrifice of simplicity and risk of collinearity in favor of keeping the significant factors that explain more of the variation of obesity rates (Figure 4).

Obesity Rate = 𝛃0 + 𝛃1(Fast Food Restaurants) + 𝛃2(Full Service Restaurants)

+ 𝛃3(fitness center) + 𝛃4(Grocery Stores) + 𝛃5(Percent Unemployed) + 𝛃6(Poverty Rate) + 𝝴

**Figure 3 - Initial Model**

Considering that the obesity rates might vary a lot among different states, we included states as factors into the model. For the purpose of showing the largest differences between the states, we assigned the state displaying the highest rate of obesity, West Virginia, as our base category (Appendix 8). In comparison with models including Region or Sub Region as factors, the model including States explained more variation in the obesity rate by county. The complete regression result of the final model is referenced in Appendix 8.

Obesity Rate = 𝛃0 + 𝛃1(Fast Food Restaurants) + 𝛃2(Full Service Restaurants)

+ 𝛃3(fitness center) + 𝛃4(Percent Unemployed) + 𝛃5(Median Income) + 𝛃6(State\*) + 𝝴

\*State is categorical factor representation of 50 states

**Figure 4 - Model A**

The p-value of our final model indicates that there is a very low chance that we would get our sample result if there was no relationship present between our predictor variables mentioned above and the obesity rate (Appendix 8). This probability was so small that we concluded that there is enough evidence to reject our null hypothesis of no relationship. In terms of the coefficients, we can conclude the following: On average, we expect 1.2% increase in the obesity rate for every additional fast food restaurant per 1000 residents of a county holding other factors constant. On the contrary, we see that on average, for each additional fitness center per 1000 residents in a county the obesity rate decreases by 2.73% holding other variables in the model constant. Similarly, for each increase of median income by $10,000 in a county we see on average, a decrease of approximately 1% in the obesity rate. The complete summary of the model and its coefficients is attached as Appendix 8 for reference.

Our analysis has also shown that states are significantly different from one another. In our earlier model building, we identified West Virginia as the state with the highest obesity rate. In our final model, we see how other states compare to WV. California is the most populous state in the nation and is in one of the least obese. It had an obesity rate 11.42% lower than West Virginia. Alabama is one of the states in our previously identified East South Central sub region, and only reflects a 0.39% lower obesity rate than West Virginia. A list of states and their obesity compared to West Virginia can be found in Appendix 8. This information may prove useful in targeting locations where strategies for reducing obesity rates can be first addressed.

**Conclusion:**

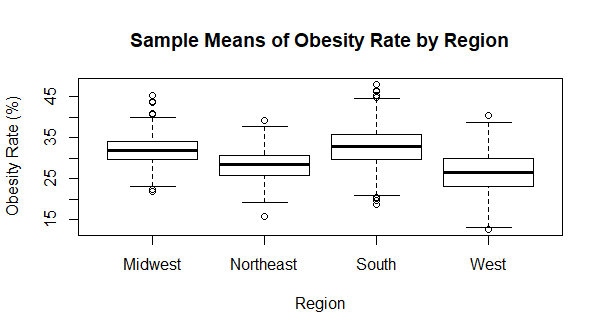
In our analysis, we wanted to find out what factors could predict obesity rate in order to begin thinking of ways to address the issue. Our model found that fast food restaurants, full service restaurants, and fitness centers per 1000 residents, as well as unemployment, median income, and the state of residence were all significant predictors for obesity. While it may be difficult to do anything to directly address median income and unemployment, we may be able to recommend actions related to the other predictors.

With fast food restaurants, we may be able to begin introducing quotas on the amount of fast food restaurants per capita that are allowed. Alternatively, we may be able to incentivize full service restaurants with subsidies or tax breaks. Fitness centers can similarly be incentivized as well to increase opportunities for exercise and subsequently reduce obesity rate. Since, the obesity level varies across states, we may be able to adjust these methods on a more granular level. We might also be able to compare practices of states that reflect higher and lower obesity rates and can learn from these differences. There are many actions that lawmakers could take with this information, and there are many more factors that they could look investigate in future research.

Our analysis found six factors that could be used to predict the obesity rate across counties in the United States. However, the model we generated only accounted for 63.22% of the variation, as demonstrated by the R-squared value in Appendix 8. Perhaps information on metabolism rates, physical activity levels, genetics, more in depth dietary information, or medical conditions could be used to further explain the obesity rate. There are many other factors we could have considered if we had access to the data, but the six that we found are a good starting block towards reducing America’s obesity problem.

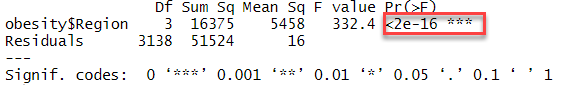
**Appendix:**

**Appendix 1 - Boxplot of Sample Means**

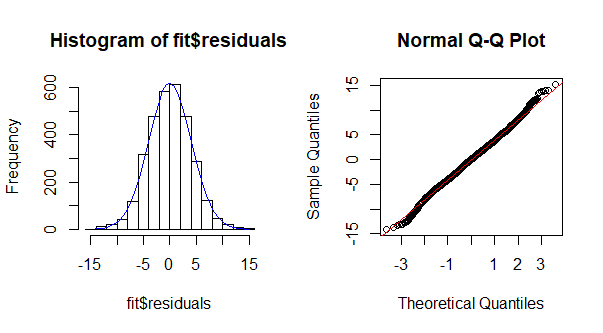


**Appendix 2 - Regional ANOVA Results**

P-value significant at significance level a = .01

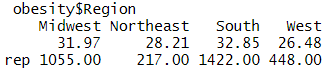


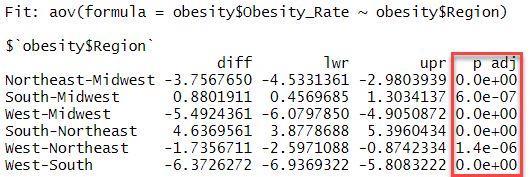
Regional Residual Checks (Shapiro-Wilks test excluded due to large n). Residuals show no evidence of rejecting our normality assumption.



**Appendix 3 - Regional Tukey Results**

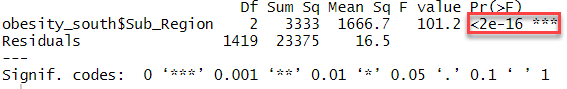
Table of means followed by Tukey test results. All p-values significant at significance level alpha = .01



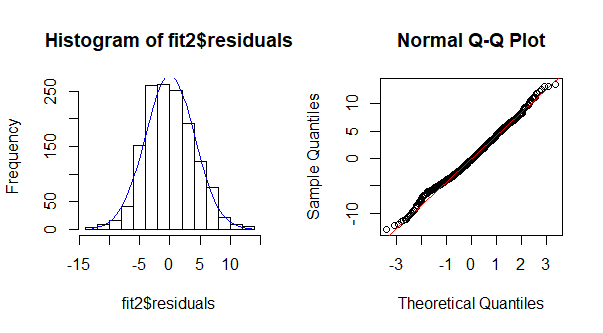


**Appendix 4 - Sub Regional ANOVA Results**

P-value significant at significance level a = .01

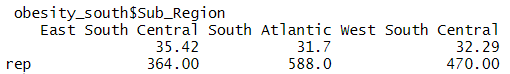
****

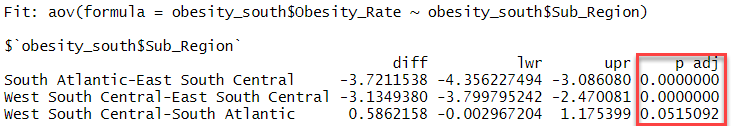
Regional Residual Checks (Shapiro-Wilks test excluded due to large n). Residuals show slight irregularity left of center, however not enough to outright reject our normality assumption.



**Appendix 5 - Sub Regional Tukey Results**

Table of means followed by Tukey test results. West South Central - South Atlantic insignificant at a = .05

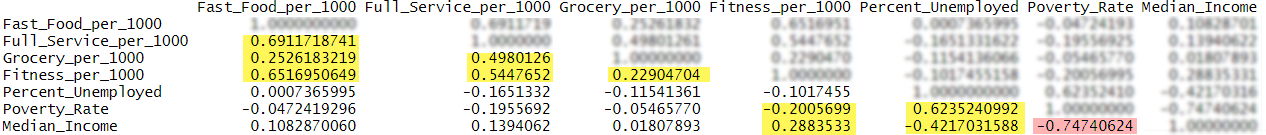
****

****

**Appendix 6 - Correlation Matrix**

We see that Poverty\_Rate and Median\_Income (Red) enough correlation to remove one from the model immediately.

Other pairs raise caution (Yellow), however we decided to keep them in our initial model as we believed they might be valuable forces in increasing explanatory power and reducing MSE.



**Appendix 7 - Model Selection Table**

Initial Model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Predictors** | **k** | **R2** | **s** | **Highest p-value** |
| Fast\_Food\_per\_1000 Full\_Service\_per\_1000  Grocery\_per\_1000 Fitness\_per\_1000 Percent\_Unemployed Median\_Income  state | 7 | 0.6322 | 2.844 | 0.670482  Grocery\_per\_1000 |

Model A - Removed Grocery

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Predictors** | **k** | **R2** | **s** | **Highest p-value** |
| Fast\_Food\_per\_1000 Full\_Service\_per\_1000  Fitness\_per\_1000 Percent\_Unemployed Median\_Income  state | 6 | 0.6322 | 2.843 | 0.000276  Fitness\_per\_1000 |

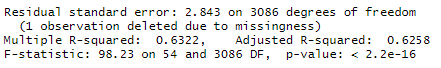
Model B - Removed Fitness and Full Service

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Predictors** | **k** | **R2** | **s** | **Highest p-value** |
| Fast\_Food\_per\_1000 Percent\_Unemployed Median\_Income  state | 4 | .0.6148 | 2.909 | 0.010935  Fast\_Food\_per\_1000 |

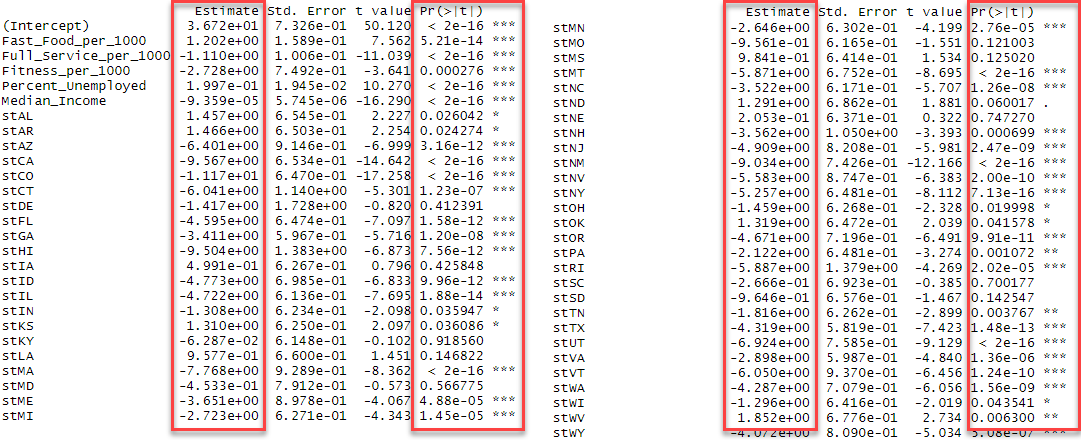
**Appendix 8 - Final Fitted Model (Model A)**

Obesity Rate = (36.72) + 1.20(Fast Food Restaurants) - 1.11(Full Service Restaurants) - 2.73(Fitness Centers) + .20(Percent Unemployed) - .000093(Median Income) + 𝛃5(State)

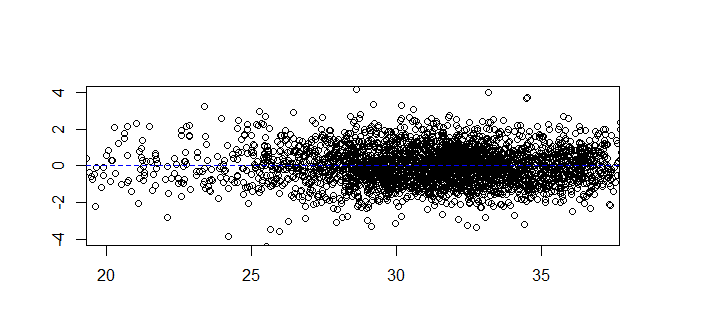
Model Summary - R-squared and MSE

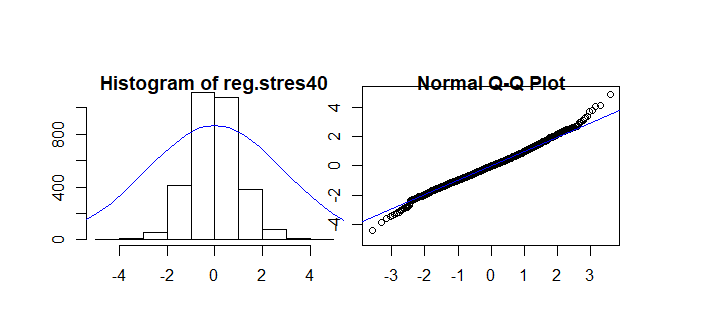


Model Summary - Coefficients:



Final Model (Model A) Residual Checks. Residuals appear to reflect normality as they are distributed evenly above and below 0 and show no strong signs of heteroscedasticity. Additionally, our histogram and qqplot reflect a normal distribution around 0.

****

****

1. "US Obesity Rates Have Hit An All-Time High [Infographic] - Forbes." 16 Oct. 2017, <https://www.forbes.com/sites/niallmccarthy/2017/10/16/u-s-obesity-rates-have-hit-an-all-time-high-infographic/>. Accessed 16 Aug. 2019. [↑](#footnote-ref-0)
2. "Adult Obesity Facts | Overweight & Obesity | CDC." <https://www.cdc.gov/obesity/data/adult.html>. Accessed 16 Aug. 2019. [↑](#footnote-ref-1)