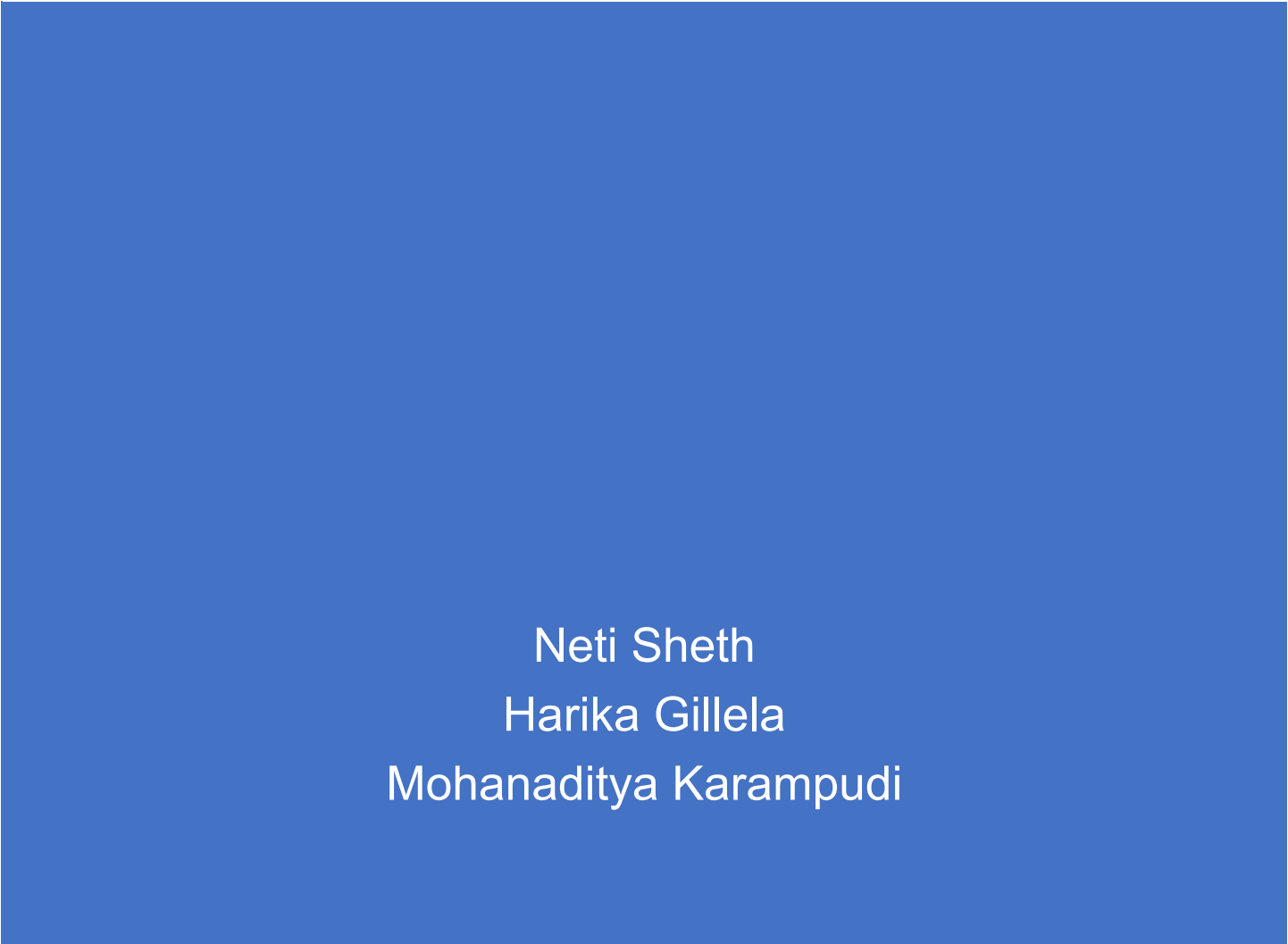




# EFFECT OF FOOD ENVIRONMENT ON OBESITY RATE



Neti Sheth  
Harika Gillela  
Mohanaditya Karampudi

# Table of Contents

		Page No
<b>1</b>	Executive Summary	2
<b>2</b>	Problem Definition and Significance	2
<b>3</b>	Prior Literature	2
<b>4</b>	Data Source and Preparation	3
<b>5</b>	Variable Choice	5
<b>6</b>	Exploratory Data Analysis and Visualization	7
<b>7</b>	Models	9
<b>8</b>	Quality Checks	12
<b>9</b>	Insights and Recommendation	13
<b>10</b>	References	13
<b>11</b>	Appendix	14

## **1. Executive Summary**

The obesity epidemic is affecting wealthy and poor countries alike. It's a complex and expensive health problem that needs to be tackled with collective effort. Everybody knows about human risk factors such as diet, physical activity, inactivity, drug use, and genetics. Many key factors in our society include education, skills, food marketing, affordability of healthy foods, physical activity, and food environment.

The food environment plays a major and frequently dominant role in nutrition choice, eating habits, and eventually energy intake. Through this project, we analyze the impact of food environments on adult obesity rates to give the county officials some actionable insights.

We merged sociodemographic, restaurants, food stores, and obesity data from the Food Environment Atlas - United States Department of Agriculture (USDA) and the US Census American Community Survey for all the US counties in the year 2010. We used Ordinary Least Square (OLS) regression to understand the cause and effect.

Our results suggest that active lifestyle, healthier food environment and choices are significant in controlling obesity.

## **2. Problem Definition and Significance**

Body Mass Index, or BMI, is used as a diagnostic tool for obesity. An ideal BMI for most adults is in the range 18.5 to 24.9. When it reaches 30, he or she is considered to be obese. The increase in obesity is linked to more than 60 chronic diseases including diabetes. About 35 percent of the US population is obese. It has a real and enduring impact on societies, countries, and most importantly, individuals, today, and through future generations.

Obesity is caused by an imbalance in energy consumption and spending at the most basic level but there are numerous, complex variables affecting this equation. Public health officials usually agree that the rise in obesity in the US occurred too fast to have genetic or biological significance as the root cause. This has prompted scientists to look at social developments as drivers of the obesity crisis, including improvements in the food environment, policies and the food production system.

The U.S. food system and food climate have developed over the past 40 years in a way that now provides a large supply of inexpensive, highly appetizing, energy-dense foods that are readily available, easy to eat, and heavily promoted. This form of environment encourages excess caloric intake and has led researchers to believe that the current food environment is the reason behind the increase in obesity levels.

Through our analysis, we want to help the county officials take preventive measures in order to encourage active living and provide a healthier food environment for all the residents.

## **3. Prior Literature**

To understand the various factors that affect obesity we have referred to many research papers. Based on our work we have found that obesity rates are higher in low-income households and minority groups. The link between where people live, and their risk of obesity led us to research the link between the food environment and health. Low-income and racial-ethnic groups are more likely to live close to unhealthy

food stores associated with poor diet than the Whites. Also, food outlets considered unhealthy (fast-food restaurants) are also more likely to be located in places with higher ethnic minority populations than whites. To capture the effect of healthy and unhealthy food outlets on obesity the food index is calculated as the ratio of unhealthy food access to healthy food access in a [paper](#) published by the US National Library of Medicine. In this study, the food environment was cited as a significant cause of the obesity epidemic. Further findings include increasing access to the stores that should be pursued along with other strategies such as improving diet quality, wellness activities.

From an [article](#) published in PMC we found that Energy balance was also another factor that affects obesity. Ecological models of obesity refer to energy balance as a function of factors “energy in” and “energy out”. The energy out is calculated as the calories burned during physical activities and the physical activities have a negative effect on obesity.

To have a better understanding of obesity, factors and consequences we referred to the data published on [CDC website](#), we found that genetic diseases and family history can also be factors responsible for obesity, but since the county-level data was not available we couldn’t consider these factors.

#### 4. Data Source and Preparation

We used 2010 data from USDA (United States Department of Agriculture, Economic Research Service) Food Atlas. It is a public dataset containing data for 3143 US counties. It has 275 indicators from different segments like Health, Insecurity, local, restaurants, stores, socio-economic data of each county. We also had the Supplemental data like population, meal programs in the USDA dataset. The data was in the form of a single excel file with data sorted in multiple sheets as per the category.

After carefully understanding each attribute, we shortlisted 40 attributes for our analysis. We merged all the different sheets data by looping through each sheet and selecting the needed attributes using state and county as the primary key. We dropped all the NULL values. We started with a correlation plot and found out that few of the attributes were highly correlated (e.g. grocery stores, supermarkets, recreation facilities). To handle the multicollinearity problem, we converted the relevant attributes to per 1000 population attributes. For example, instead of total grocery stores in a county, we took grocery stores per 1000 population as our attribute. In the end, we were left with 25 attributes.

From the US census bureau, we extracted the county-level data of education attainment and the commuting preferences for residents and merged it with our existing data.

Attribute	Definition	Data Source
<u>Obesity Rate</u> (Dependent Variable)	The estimate of the age-adjusted percentage of persons age 20 and older who are obese, where obesity is Body Mass Index (BMI) greater than or equal to 30 kilograms per meter squared.	Estimates are from the Centers for Disease Control and Prevention (CDC) using data from the Behavioral Risk Factor Surveillance System (BRFSS)
% Low access to store	Percentage of people in a county with low income and living more than 1 mile from a supermarket, supercenter, or large grocery store.	2012 report, Access to Affordable and Nutritious Food: Updated Estimates of Distances to

		Supermarkets Using 2010 Data.
Poverty Rate	Percentage of county residents with household income below the poverty threshold.	U.S. Census Bureau, 2010 Census.
% Bachelors or higher	Percentage of county residents who have graduated with at least a bachelor's degree.	U.S. Census Bureau, 2010 Census.
% Public Transportation	Percentage of county residents commuting to work using public transport	U.S. Census Bureau, 2010 Census.
% Walk	Percentage of county residents commuting to work by walking	U.S. Census Bureau, 2010 Census.
Metro/ non-metro counties	Classification of counties by metro or nonmetro definition, where 1=metro county; 0=nonmetro county. Nonmetro counties have no cities with 50,000 residents or more.	USDA's Economic Research Service—Rural Classifications.
% white	Percentage of county resident population that is non-Hispanic White.	U.S. Census Bureau, 2010 Census.
Recreation & fitness facilities/1,000 pop	The number of “fitness and recreation centers” in a county divided by the number of county residents. It includes recreational sports facilities featuring exercise and other active physical fitness conditioning or recreational sports activities, such as swimming, skating, or racquet sports.	U.S. Census Bureau, 2010 Census.
Natural Amenity Index	Index of natural amenities constructed by USDA’s Economic Research Service, ranging from 1 to 6, where 1=lowest and 6=highest. It is based on the premise that people are drawn to areas with varied topography; lakes, ponds, or oceanfront; warm, sunny winters; and temperate, low-humidity summers. The index measures a county’s natural amenities score as a standard deviation from the all-county mean value.	Natural Amenities Drive Rural Population Change, AER-781, USDA, ERS, and the ERS Natural Amenities Scale.
Low fat milk : soda	The ratio of the regional average price of low-fat milk to the regional average price of sodas relative to the national average price ratio.	ERS estimates using the Quarterly Food-at-Home Price Database, QFAHPD-2
Fast-food restaurants/ 1,000 pop	The number of limited-service restaurants in the county per 1,000 county residents.	U.S. Census Bureau, 2010 Census.
Full-service restaurants/ 1,000 pop	The number of full-service restaurants in the county per 1,000 residents. It includes establishments primarily engaged in providing food services to patrons	U.S. Census Bureau, 2010 Census.

	who order and are served while seated and pay after eating.	
Grocery stores/1,000 pop	The number of supermarkets and grocery stores in the county per 1,000 county residents.	U.S. Census Bureau, 2010 Census.
Supercenters & club stores/1,000 pop	The number of supercenters and warehouse club stores in the county per 1,000 county residents.	U.S. Census Bureau, 2010 Census.
Specialized food stores/1,000 pop	The number of specialized food stores in the county per 1,000 county residents. It includes establishments primarily engaged in retailing specialized lines of food, such as retail bakeries, meat and seafood markets, dairy stores, and produce markets.	U.S. Census Bureau, 2010 Census.
Farmers' markets/1,000 pop	The number of farmers' markets in the county per 1,000 county residents.	U.S. Census Bureau, 2010 Census.

## 5. Variable Choice

We started with initial assumptions and used data to validate these assumptions. The variables we have decided are the economic factors, food availability, stores, education, public services and recreation facilities:

Attribute	Relation to obesity rate.
% Low access to store	Low access to stores can make it hard for people to have healthy food. People with low income cannot afford healthy food which will affect their health and can be a positive indicator of obesity rate
Poverty Rate	People with income less than the poverty threshold cannot afford healthy food and will find alternative unhealthy foods that can affect their health.
% Bachelors or higher	Education will increase awareness among the people and help them lead a healthier lifestyle.
% Public Transportation	Commuting by public transportation is a less convenient option than traveling by car. People might have to walk to their stations, and it increases physical activity.
% walk	Commuting to work by walk indicates a healthier lifestyle and can help reduce obesity.

Metro/ non-metro counties	People living in metro counties can have a different lifestyle compared to people living in non-metro counties. We have considered this variable to understand how the county status affects the obesity rate.
% white	Disparities can exist in the obesity rate based on ethnicity and race.
Recreation & fitness facilities/1,000 pop	Fitness activities will help people in the county stay active, maintain energy balance, and have a healthy lifestyle.
Natural Amenity Index	Higher amenity index promotes an active lifestyle in the county so county administrators can take measures to preserve nature and maintain the same index.
Low fat milk: soda (milk - healthy product soda - unhealthy product)	It is the ratio of prices of healthy products to unhealthy products. It can be useful to check how the affordability of healthy food to people affects obesity.
Fast-food restaurants/ 1,000 pop	The food from fast-food restaurants is referred to as unhealthy food which can have a positive effect on the obesity rate.
Full-service restaurants/ 1,000 pop	Full-service restaurants are a healthier alternative to fast-food restaurants and offer nutritious food. They can help control obesity.
Grocery stores/1,000 pop	The number of supermarkets and grocery stores in the county indicates the accessibility of healthy food to the people.
Supercenters & club stores/1,000 pop	Supercenters have both healthy and unhealthy food. Often, unhealthy food is cheaper and people can be more exposed to it and can buy it. Accessibility to unhealthy food increases and it can have a positive impact on obesity.
Specialized food stores/1,000 pop	Accessibility to retail bakeries, meat and seafood markets, dairy stores, and produce markets can help control obesity. These food items are rich in protein content and are considered healthy foods.
Farmers' markets/1,000 pop	The farmers market will have fresh vegetables and fruits and can be considered as healthy food and this will have a positive effect on the health of people.

## 6. Exploratory Data Analysis and Visualization

Obesity Rates in United States

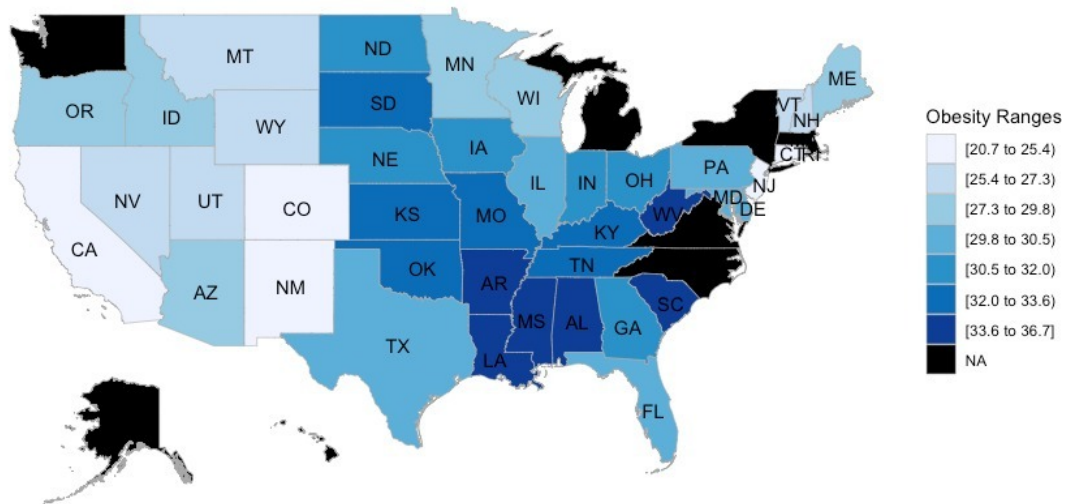


Fig: 1

Our dependent variable is obesity rate. We can observe from the map above (Fig 1) that all states have more than 20% of adults with obesity. The South - East part of the US has more prevalence than the other regions. In at least 12 states, prevalence of obesity was greater than 32%. Colorado has the least obesity rate and Alabama has the highest.

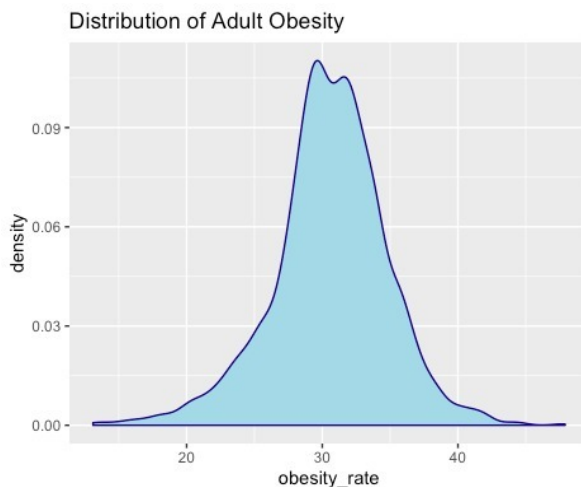


Fig : 2

The distribution of adult obesity rate (Fig 2) is normal and has less spread. The obesity rate ranges from 13.10% to 47.9% with a mean of 30.57%. This suggests that on average, the prevalence of obesity in the United States is 30.57%. 1 out of 3 adults, were found to be obese in 2010.



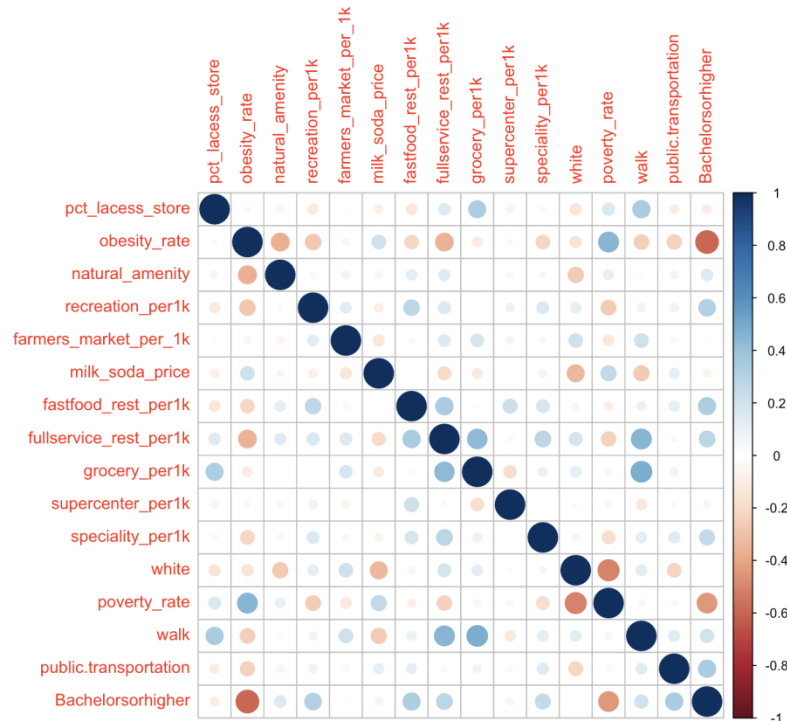


Fig : 3

From the correlation plot (Fig:3) of all the attributes, we can observe that all the attributes are not highly correlated with a few exceptions. The obesity rate is negatively correlated with % of bachelors, natural amenity, recreation facilities per 100, and full-service restaurants per 100 suggesting that these attributes can help lower obesity. The obesity rate is positively correlated with poverty rate which is expected.

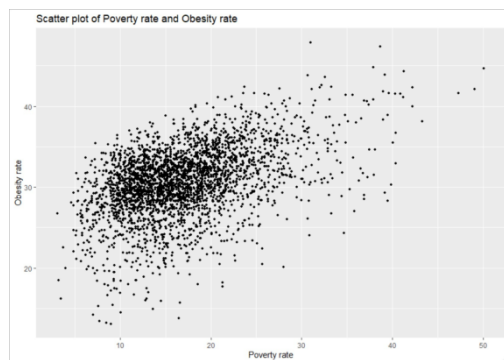


Fig: 4

A higher poverty rate signifies that more people will have lesser means to healthier food and lifestyle, in turn, increases the risk of obesity. From the scatterplot (Fig 4), we can observe the same relationship - obesity increases with poverty.

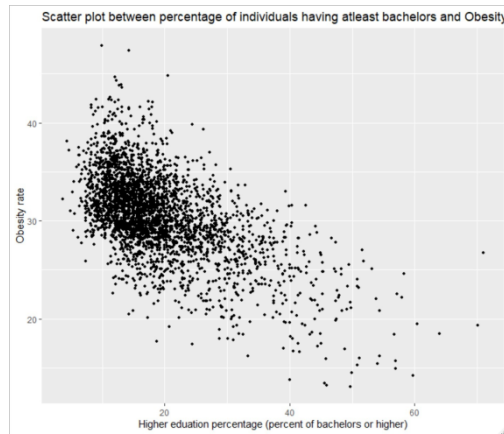


Fig 5

Education can increase awareness about the importance of high nutrition diets and the obesity epidemic and prevention. From the scatter plot (Fig 5), we can observe that an increase in the percentage of bachelors, reduces obesity.

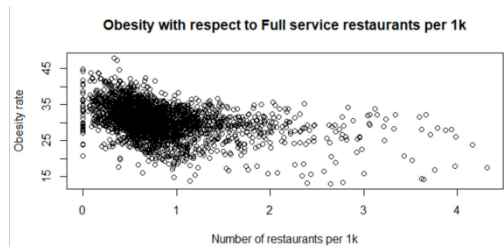


Fig 6

The scatter plot (Fig 6) suggests that having more full-service restaurants does not increase obesity. This can be because they have a more balanced and healthier menu. To control obesity, we don't have to stop eating out. We just have to find healthier alternatives.

## 7. Models

The dependent variable obesity rate has a normal distribution, so we implemented a multiple linear regression model to understand the effect of the various factors such as food environment, socio-economic, accessibility, health, and wellness activities on the obesity rate.

Model 1 is the “main effects” model and initial model that explains the effect of individual independent variables on obesity.

Model 2 considers independent variables that a county administrator can control or take necessary actions to reduce the obesity rate in the county.

Model 3 contains interactions of the individual variables and explains the collective effect of those interacted variables on the obesity rate.

Among the 3 models. Model 3 is our final model as the coefficients of the estimates are more closely aligned with the real world and the interaction terms provide more explanation on the obesity rate variation across counties.

Variable	Model 1	Model 2	Model 3
% Low access to store	-0.001	-0.003	0.003
Natural Amenity Index	-1.417		
% recreation/1000 pop	-2.655	-3.123	-3.394
Fast-food restaurants/ 1,000 pop: Specialized food stores/1,000 pop			1.928
Fast-food restaurants/ 1,000 pop: Recreation & fitness facilities/1,000 pop			0.257
Grocery stores/1,000 pop: % Low access to store			-0.016
Low-fat milk: soda	3.142	3.504	3.456
metro county	0.575		
Fast-food restaurants/ 1,000 pop	0.284	0.292	0.121
Full-service restaurants/ 1,000 pop	-0.559	-1.155	-1.189
Grocery stores/1,000 pop	0.094	0.163	0.491
Supercenters & club stores/1,000 pop	10.543	9.184	9.668
Specialized food stores/1,000 pop	-0.239	-0.008	-1.358
Farmers' markets/1,000 pop	1.913	2.077	2.010
%white	-0.019		
Poverty Rate	0.157	0.128	0.126
%walk	-0.069	-0.064	-0.062
% Bachelors or higher	-0.185	-0.198	-0.198
% public transportation	-0.146	-0.120	-0.123
constant	35.551	30.170	30.287

### Understanding Model 3

Low Accessibility:

- If the percentage of the population with low income and low access is increased by 1% then obesity will be increased by 0.003% which is almost 0% - no effect as the standard error is 0.008.

- With a 1% increase in population of people in the county with low income and low access to stores, the effect of the additional 1 unit of the grocery stores for 1000 population on obesity decreases by 0.016%.

#### Healthy food affordability:

- With every 1% increase in the price of healthy food products (low-fat milk) relative to the unhealthy food products (soda), obesity will increase by 3.45%. With a very low value of standard error (0.49) the attribute is statistically significant.

#### Recreation and Fitness Centers:

- 1 unit increase in the number of recreational centers per 1000 population, reduces obesity by 3.39%. Also, with just 0.6 of standard error, this attribute is statistically significant and has a large coefficient.

#### Restaurants:

- If the number of fast food restaurants per 1000 population increases by 1, obesity will increase by 0.12%.
- If the number of full-service restaurants per 1000 population increases by 1, obesity will reduce by 1.18% which specifies that full-service restaurants serve healthy food. With coefficient of -1.155 and St. Error (0.1), the attribute is statistically significant and highly relevant is vital in modeling
- As fast-food restaurants per 1000 population increases by 1,
  - the effect of the additional 1 unit of the specialized store for 1000 population on obesity increases by 1.9%
  - the effect of the additional 1 unit of the recreational facilities for 1000 population on obesity rate increases by 0.2%.

#### Stores:

- The number of grocery stores per 1000 population has a positive impact on the obesity rate. For an increase in each unit of grocery stores the obesity increases by 0.49% but has high 0.53 it is not a statistically significant attribute and can be ignored.
- For an increase in each supercenter & club store for 1000 population, the obesity will increase by 9.7%. This effect can be because they offer great discounts on bulk packs of unhealthy food items (like chips and soda) in addition to offering healthy food items like fruits and vegetables.
- For an increase in each specialized store for 1000 population the obesity will be reduced by 1.35%.
- Farmer's market which sells fresh fruits and vegetables makes healthier food available to people. We expect this field to reduce the obesity but the coefficient here represents that for an increase in every unit of farmer's market per 1000 population increases the obesity rate by 2%.

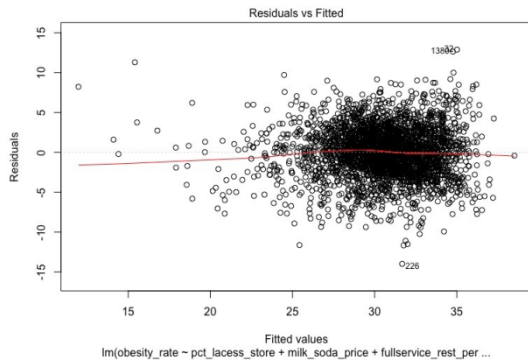
#### Socio-demographic:

- An increase of 1% of the poverty rate of the county, increases obesity by 0.12%.
- For every 1% increase in the percentage of the people who commute to work by walking will reduce the obesity of the county by 0.06%.
- For an increase in every unit percentage of the population with a bachelor's degree or higher in the county, the obesity reduces by 0.19%
- For a 1% increase in the county population that uses public transportation for commuting to work, obesity decreases by 0.12%.

## 8. Quality Checks

As we have implemented Linear regression, we have primarily checked for four basic assumptions:

1. The dependent and independent attributes have linear relation: In the residual vs Fitted, the passing of the red line along the horizon and no clear pattern of the points suggests that the model is linear.



2. Independence test: This assumption is not applicable to the data as each county's data is independent to other counties.

The Durbin-Watson test reveals that the model suffers from autocorrelation.

```
> dwtest(ols3)
```

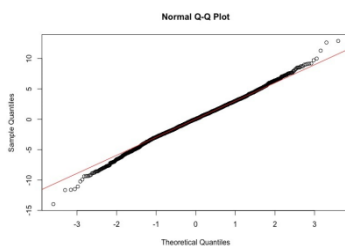
Durbin-Watson test

data: ols3

DW = 1.2403, p-value < 2.2e-16

alternative hypothesis: true autocorrelation is greater than 0

3. Normality - This assumption states that the residuals should confine to normal distribution. From the QQ plot it is relevant that most of the residuals are normally distributed with exception of few extreme points.



4. Residuals have constant variance - In an ideal case, the residuals should have equal variance across all the points (Homoscedasticity). But, the Breusch-Pagan test shows that there is Heteroscedasticity in the residuals.

```
> bptest(ols3)
```

```
studentized Breusch-Pagan test
```

```
data: ols3
```

```
BP = 249.69, df = 16, p-value < 2.2e-16
```

## 9. Insights and Recommendation

### Insights:

- Recreational facilities promote healthy and active living and can help to reduce obesity. 1% increase will reduce obesity by 3.39%.
- Full-service restaurants are a healthier alternative to fast food restaurants as they provide well-cooked nutritious, low-calorie healthy food.
- The affordability of healthy food items has a great impact on controlling obesity. If the prices are 1% less expensive compared to unhealthy products like soda, obesity will reduce by 3.45%.
- Specialized food stores like retail bakeries, meat and seafood markets, dairy stores, and produce markets are popular and often visited by people. Obesity can be reduced by 1.35%, if the number of specialized stores per 1000 population increase by 1.
- Even though grocery stores and farmers markets sell healthy food items, they don't help to reduce obesity.
- Though low accessibility to stores is a hurdle for having access to healthy food, it does not have any effect on obesity.
- If more people travel to work by public transportation, obesity can be controlled. 1% increase can reduce obesity by 0.12%.
- Education helps to create more awareness about the importance of healthy and active living and can help to reduce obesity. A 1% increase in bachelor graduates, can reduce obesity by 0.19%.

### Recommendation to County Officials to reduce obesity:

- Make healthy food products more affordable with the help of food assistance programs like SNAP(S) (Supplemental Nutrition Assistance Program). Increase of taxes on unhealthy products and subsidies for healthy ones can also help.
- Open more recreational facilities for every 1000 population to promote active living.
- Open more healthy food outlets like full-service restaurants and specialized stores (retail bakeries, meat and seafood markets, dairy stores, and produce markets) for every 1000 population.
- Control the number of supermarkets and club stores per 1000 population. They make food products like soda, instant and processed food, more convenient and readily available. It increases obesity risk.

## 10. References

<https://www.cdc.gov/obesity/adult/causes.html>

<https://www.cdc.gov/obesity/strategies/index.html>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2708156/>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5708005/>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4283210/>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4201352/>  
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4977027/>  
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1447325/>  
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1449238/>  
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1470485/>

## 11. Appendix

### Variable Selection

```
rm(list=ls())
library(readxl)
library(dplyr)
library(tidyverse)
# POPULATION
data_population = read_excel("August2015.xls", sheet = "Supplemental
Data - County")
data_population$FIPS = as.numeric(data_population$FIPS)
data_population <- data_population[,c("FIPS", "State", "County", "2010
Census Population")]
# ACCESS
data_access = read_excel("August2015.xls", sheet = "ACCESS")
data_access$FIPS = as.numeric(data_access$FIPS)
data_access <- data_access[, c("FIPS", "State", "County",
"PCT_LACCESS_LOWI10")]
# HEALTH
data_health = read_excel("August2015.xls", sheet = "HEALTH")
data_health$FIPS = as.numeric(data_health$FIPS)
data_health <- data_health[,
c("FIPS", "State", "County", "PCT_OBESE_ADULTS10", "RECFAC07", "NATAMEN",
"RECFACPTH07")]
# LOCAL
data_local = read_excel("August2015.xls", sheet = "LOCAL")
data_local$FIPS = as.numeric(data_local$FIPS)
data_local <- data_local[, c("FIPS", "State", "County", "FMRKT09",
"FMRKTPTH09")]
# PRICES AND TAXES
data_prices = read_excel("August2015.xls", sheet = "PRICES_TAXES")
data_prices$FIPS = as.numeric(data_prices$FIPS)
data_prices <- data_prices[,
c("FIPS", "State", "County", "MILK_SODA_PRICE10")]
# RESTAURANTS
data_restaurants = read_excel("August2015.xls", sheet = "RESTAURANTS")
data_restaurants$FIPS = as.numeric(data_restaurants$FIPS)
data_restaurants <- data_restaurants[,
c("FIPS", "State", "County", "FFR07", "FFRPTH07", "FSRPTH07", "FSR07")]
# STORES
data_stores = read_excel("August2015.xls", sheet = "STORES")
data_stores$FIPS = as.numeric(data_stores$FIPS)
```

```

data_stores <-
data_stores[,c("FIPS","State","County","GROC07","SUPERC07","CONVS07","
SPECS07",
               "GROCPTH12", "SUPERCPTH07",
               "CONVSPTH07", "SPECSPTH07")]
# SOCIOECONOMIC
data_sc = read_excel("August2015.xls", sheet = "SOCIOECONOMIC")
data_sc$FIPS = as.numeric(data_sc$FIPS)
data_sc <-
data_sc[,c("FIPS","State","County","PCT_NHWHITE10","PCT_HISP10","PCT_N
HBLACK10","PCT_NHASIAN10","PCT_NHNA10",
           "PCT_NHPI10","MEDHHINC10","POVRATE10","METRO13")]
# Merging Categories
mergeCols <- c("FIPS","State","County")
df_final <- Reduce(function(x, y) merge(x, y, by = mergeCols),
list(data_population,data_access, data_health, data_prices,
data_local,

data_prices,data_restaurants, data_stores, data_sc ))
# Gini Index
gini_index = read.csv("GiniIndex.csv")
gini_index <- gini_index[,c("FIPS","Gini.Index")]
# County Commute
commute = read.csv("county_commute.csv")
commute <- commute[,c("FIPS","car.alone","walk",
"public.transportation")]
# Education in County
education = read.csv("County_Education.csv")
education <- education[,c("FIPS","Bachelorsorhigher")]

households = read.csv("households.csv")
households <- households[,c("FIPS", "occupied_housing_units")]

# Adding Gini Index and commute data to final dataset
df_final<-merge(df_final,commute, by = "FIPS", all.x = TRUE)
df_final<-merge(df_final,gini_index, by = "FIPS", all.x = TRUE)
df_final <- merge(df_final,education, by = "FIPS", all.x = TRUE)
df_final <- merge(df_final,households, by = "FIPS", all.x = TRUE)
dim(df_final)

rm(data_access,data_health,data_local,data_prices,data_stores,
   data_restaurants, data_sc,gini_index,commute, education)

df_final$MILK_SODA_PRICE10.x <- NULL

na_count <-sapply(df_final, function(y) sum(length(which(is.na(y)))))
na_count <- data.frame(na_count)
na_count
# Missing values of MILK_SODA_PRICE
df_final = within(df_final, MILK_SODA_PRICE10.y[State == 'AK'] <-
mean(df_final[df_final$State == "AL",]$MILK_SODA_PRICE10.y))

```



```

df_final = within(df_final, MILK_SODA_PRICE10.y[State == 'HI'] <-
mean(df_final[df_final$State == "AL",]$MILK_SODA_PRICE10.y) + 0.5)
# Farmer's Market
df_final$FMRKT09[is.na(df_final$FMRKT09)] <- 0

# removing missing values
df_final = df_final[complete.cases(df_final), ]

## Renaming columns
df_final <- df_final %>%
  rename(
    'population_2010' = '2010 Census Population',
    'fullservice_restaurants' = 'FSR07',
    'fastfood_restaurants' = 'FFR07',
    'farmers_mrkt' = 'FMRKT09',
    'conv_stores' = 'CONVS07',
    'super_center' = 'SUPERC07',
    'specialized_strs' = 'SPECS07',
    'grocery_stores' = 'GROC07',
    'pct_laccess_store' = 'PCT_LACCESS_LOWI10',
    'obesity_rate' = 'PCT_OBESE_ADULTS10',
    'natural_amenity' = 'NATAMEN',
    'milk_soda_price' = 'MILK_SODA_PRICE10.y',
    'recreation_fac' = 'RECFAC07',
    'poverty_rate' = 'POVRATE10',
    'metro_county' = 'METRO13',
    'median_income' = 'MEDHHINC10',
    'white' = 'PCT_NHWHITE10',
    'black' = 'PCT_NHBLACK10',
    'hisp' = 'PCT_HISP10',
    'asian' = 'PCT_NHASIAN10',
    'alaska' = 'PCT_NHNA10',
    'giniindex' = "Gini.Index",
    'onlycar' = "car.alone",
    'hawaiian' = 'PCT_NHPI10',
    'grocery_per1k' = 'GROCPTH12',
    'supercenter_per1k' = 'SUPERCPTH07',
    'convenience_per1k' = 'CONVSPTH07',
    'speciality_per1k' = 'SPECSPTH07',
    'fastfood_rest_per1k' = 'FFRPTH07',
    'fullservice_rest_per1k' = 'FSRPTH07',
    'farmers_market_per_1k' = 'FMRKTPTH09',
    'recreation_per1k' = 'RECFACPTH07'
  )

write.csv(df_final, "final_data.csv", row.names = FALSE)

```

## Data Modeling

```
library(DataExplorer)
```

```

library(car)
library(dplyr)
library(corrplot)
library(stargazer)
library(MASS)
rm(list=ls())
df = read.csv("final_data.csv")

# Correcting data type
df$metro_county = as.factor(df$metro_county)
df$metro_county = relevel(df$metro_county, 1)

# Removing unnecessary variables
df$FIPS <- NULL
df$State <- NULL
df$County <- NULL
df$black <- NULL
df$hispanic <- NULL
df$asian <- NULL
df$alaska <- NULL
df$hawaiian <- NULL
df$median_income <- NULL
# df$walk <- NULL
df$giniindex <- NULL
df$onlycar <- NULL
df$total_stores <- NULL
df$log_occupied_housing <- NULL
df$farmers_mrkt <- NULL
df$fastfood_restaurants <- NULL
df$fullservice_restaurants <- NULL
df$grocery_stores <- NULL
df$super_center <- NULL
df$conv_stores <- NULL
df$specialized_stores <- NULL
df$occupied_housing_units <- NULL

df$recreation_fac <- NULL

# May be needed later
df$population_2010 <- NULL
df$convenience_per1k <- NULL

attach(df)

# Plot correlation plot
plotCorrelation <- function(data ){
  cor_df = cor(data[sapply(data, is.numeric)])
  corrplot(cor_df, title="Correlation Plot")
}

plotHistogram <- function(df) {
  for (i in colnames(df)){

```

```

    print(i)
    hist(df[,i],main=paste("histogram curve of column:",i))
  }
}

library(ggplot2)
plotScatterPlot <- function(df) {
  for (i in colnames(df)){
    plot(df[,i], df$obesity_rate, main=paste("Scatter plot of:", i))
  }
}

plotCorrelation(df)
plotHistogram(df)
plotScatterPlot(df)

colnames(df)
attach(df)

# Basic OLS model
ols1 <- lm(obesity_rate ~ pct_laccess_store + natural_amenity +
recreation_perlk +
          milk_soda_price + metro_county + fastfood_rest_perlk +
          fullservice_rest_perlk + grocery_perlk + supercenter_perlk
+ speciality_perlk +
          farmers_market_per_1k + white + poverty_rate + walk +
          Bachelorsorhigher + public.transportation, data = df)
# vif(ols1)

# OLS model with only control variables
ols2 <- lm(obesity_rate ~ pct_laccess_store + recreation_perlk +
          milk_soda_price + fastfood_rest_perlk +
fullservice_rest_perlk +
          grocery_perlk + supercenter_perlk + speciality_perlk +
walk +
          farmers_market_per_1k + poverty_rate +
          Bachelorsorhigher + public.transportation, data = df)
# vif(ols2)

# OLS model with control variables and interaction terms
ols3 <- lm(obesity_rate ~ pct_laccess_store + milk_soda_price +
fullservice_rest_perlk +
          farmers_market_per_1k + walk + poverty_rate +
grocery_perlk +
          Bachelorsorhigher + public.transportation +
supercenter_perlk +
          fastfood_rest_perlk*speciality_perlk +
          recreation_perlk*fastfood_rest_perlk +
          grocery_perlk*pct_laccess_store, data = df)
plot(ols3)

```

```
stargazer(ols1, ols2, ols3, type="text", title = "Comparision Models",  
out="output.doc")
```

```
hist(ols3$res)  
qqnorm(ols3$res)  
qqline(ols3$res, col="red")  
shapiro.test(ols3$res)           # Residuals not MV normal
```

```
plot(ols3$res ~ ols3$fit)  
bartlett.test(list(ols3$res, ols3$fit))  
norm <- rnorm(630)  
bartlett.test(list(ols3$res, norm)) # Residuals heteroskedastic
```

```
library("car")  
vif(ols3)           # No multicollinearity in data
```