Final R

Ananya Goyal, Karolina Hajkova, Francesca Pilia, Aurora Pulpito

2023-05-02

Loading Packages

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2
## ──
```

```
## ✔ tibble  3.1.8     ✔ purrr   1.0.1
## ✔ tidyr   1.3.0     ✔ stringr 1.5.0
## ✔ readr   2.1.3     ✔ forcats 0.5.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
library(magrittr)
```

```
##
## Attaching package: 'magrittr'
##
## The following object is masked from 'package:purrr':
##
##     set_names
##
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
library(readxl)
library(usmap)
library(plotly)
```

```
##
## Attaching package: 'plotly'
##
## The following object is masked from 'package:ggplot2':
##
##     last_plot
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following object is masked from 'package:graphics':
##
##     layout
```

```
library(infer)
library(broom)
library(boot)
library(stargazer)
```

```
##
## Please cite as:
##
##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Ta
bles.
##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

```
library(moderndive)
```

Introduction

Obesity is a major public health issue in the United States, and research has shown that poverty and education are two important factors that significantly influence its prevalence. Poverty can contribute to limited access to healthy food options, as well as a lack of accessible and secure places for physical activity. People living in poverty may consume more inexpensive, high-calorie, and nutritionally deficient foods, which can increase the risk of obesity. In addition, limited opportunities for physical activity can further exacerbate the issue. Lower levels of education can also contribute to the risk of obesity. People with inadequate knowledge about healthy eating and the harmful effects of obesity on their health may adopt unhealthy eating habits and sedentary lifestyles. Furthermore, people with lower levels of education may not fully understand the importance of exercise in maintaining a healthy weight.

Our project aims to investigate the relationship between poverty, education, and obesity in the United States using econometric methods applied to relevant datasets. Our main research question is to : *What extent are poverty and education associated with obesity risk*? By analyzing the available data, we aim to provide empirical evidence on the impact of poverty and education on obesity and to quantify the magnitude of this relationship. Through this analysis, we hope to contribute to a better understanding of the underlying drivers of obesity in the United States.

Literature Review

Levine (2011) found that while the rates of obesity in high-income countries are higher than those in middle- and low-income countries, in the US, where obesity is prevalent, people living in poverty-dense counties are most prone to obesity. Within social epidemiological research, income was found to be inversely associated with obesity, though this relationship can be interpreted in two directions: (1) the causation hypothesis that explains lower income as a cause for subsequent obesity and (2) the perspective of a reversed causality, in which obesity is not the result, but rather the cause for lower income, found Kim and von dem Knesebeck (2008). Although there is some suggestion that food deserts and hunger may be the cause, the sedentariness of people living in poverty-dense regions may also be a significant factor in their obesity.

Yoon et al. (2006) discovered that income had a greater impact on BMI and waist circumference than education. Nonetheless, according to research conducted by Cutler and Lleras-Muney (2006), individuals who have completed more years of education are less likely to engage in unhealthy behaviors such as smoking, excessive drinking, illegal drug use, and obesity. They are also more likely to engage in healthy behaviors like exercising and obtaining preventive care. The relationship between education and health is non-linear for obesity, with additional years of schooling having increasing effects. In a review conducted by Grossman and Kaestner (1997), years of formal education was identified as the most important factor associated with good health. Furthermore, Webbink et al. (2008) found a negative relationship between education and the probability of being overweight through cross-sectional estimates from a study of twins.

Exploring the Relationship Between Education and Obesity

Overweight and obesity rates have been increasing sharply over recent decades in all industrialized countries, as well as in many lower-income countries.The rise in obesity has reached epidemic proportions, with over 1 billion adults worldwide estimated to be overweight and at least 300 million of those considered to be clinically obese (WHO, 2003). The circumstances in which people have been leading their lives over the past 20-30 years, including physical, social and economic environments, have exerted powerful influences on their overall calorie intake, on the composition of their diets and on the frequency and intensity of physical activity at work, at home and during leisure time.

On the other hand, changing individual attitudes, reflecting the long-term influences of improved education and socio-economic status (SES) have countered to some extent environmental influences.(The Organization for Economic Co-operation and Development) Many OECD countries have been concerned not only about the pace of the increase in overweight and obesity, but also about inequalities in their distribution across social groups, particularly by level of education, socio-economic status and ethnic background. Inequalities across social groups appear to be particularly large in women (Wardle et al., 2002; Branca et al., 2007). Acting on the mechanisms that make individuals who are poorly educated and in disadvantaged socio-economic circumstances so vulnerable to obesity, and those at the other end of the socio-economic spectrum much more able to handle obesogenic environments, is of great importance not just as a way of redressing existing inequalities, but also because of its potential effect on overall social welfare. The current distribution of obesity appears particularly undesirable, as it is likely to perpetuate the vicious circle linking obesity and disadvantage by intergenerational transmission.

Economists have shown much interest in the estimation of the causal effect of education on wages and economic growth . Empirical studies, for example, suggest that education has a positive impact on health and well-being (Wolfe and Haveman 2002; Lleras- Muney 2005), particularly in poorer countries (Cutler and Lleras-Muney, 2006), reduces crime (Lochner and Moretti 2004) and water and air pollution (Appiah and McMahon 2002). The finding that education has positive externalities provides a rationale for government intervention. However, the causal nature of the link between education and health is still subject to a certain degree of scrutiny, and the precise mechanisms through which education may affect health are not yet fully understood. Lifestyles may be one of the keys to understanding such a relationship, as they are often significantly influenced by education and, at the same time, they contribute to health and longevity by affecting the probability of developing a wide range of diseases.

Obesity is a close marker of important aspects of individual lifestyles, such as diet and physical activity, and is also an important risk factor for major chronic diseases, such as diabetes, heart disease, stroke and certain cancers. Obesity is also associated with negative labour market outcomes, in term of both wages and employment, particularly for women.

Specific Objectives:

1. To explore the correlation between body mass index, and obesity, on one hand, and formal education, expressed in terms of years spent in full-time education, on the other, controlling for possible confounding factors. The main goal of this analysis is to determine whether the intensity of the relationship between education and obesity is constant, or whether it shows increasing or decreasing strength at either end of the education spectrum.
2. To assess the extent to which the correlations identified may reflect the influences of factors associated with individual education, such as socio-economic status and the level of education of household members.
3. To assess the extent to which the correlations identified may reflect causal links between education and obesity.
4. To explore what conceptual model of the role of education as a determinant of health is most consistently supported by the findings concerning the correlation between obesity and aspects of individual and group education.

Downloading Data

For this project, we are analyzing data from the United States on three variables - obesity, poverty, and education - from the years 2017 to 2021. The obesity variable is derived from the National Center for Chronic Disease Prevention and Health Promotion (NCCDPHP), which provides national and state-level prevalence estimates of obesity. Obesity is defined as having a body mass index (BMI) of 30 or higher. The poverty variable is obtained from Statista (using data from the US Census), which reports the percentage of the population living below the poverty line in each state. The poverty line is defined as the minimum level of income deemed necessary to achieve an adequate standard of living. Lastly, the education variable is sourced from the Federal Reserve Economic Data (FRED) website, which reports the percentage of the population holding a Bachelor's Degree or higher in each state.

We selected these variables, namely obesity, poverty, and education, because they are known to be related to each other and have significant impacts on public health. By analyzing the relationships between these variables, we can gain insights into how socioeconomic factors affect health outcomes in different regions of the United States.

```
obesity2021 = read.csv(file = "~/Desktop/obesity/2021data.csv")%>%
  select(ID, YearStart, Description, LocationDesc, Topic, Data_Value, Sample_Size)
obesity2020 = read.csv(file = "~/Desktop/obesity/2020data.csv")%>%
  select(ID, YearStart, Description, LocationDesc, Topic, Data_Value, Sample_Size)
obesity2019 = read.csv(file = "~/Desktop/obesity/2019 data.csv")%>%
  select(ID, YearStart, Description, LocationDesc, Topic, Data_Value, Sample_Size)
obesity2018 = read.csv(file = "~/Desktop/obesity/2018 data.csv")%>%
  select(ID, YearStart, Description, LocationDesc, Topic, Data_Value, Sample_Size)
obesity2017 = read.csv(file = "~/Desktop/obesity/2017 data.csv")%>%
  select(ID, YearStart, Description, LocationDesc, Topic, Data_Value, Sample_Size)

obesity = rbind(obesity2017, obesity2018, obesity2019, obesity2020, obesity2021)%>%
  filter(!is.na(ID) & (!(LocationDesc %in% c("Guam", "Puerto Rico", "Virgin Islands"))))
```

```
edu2021 = read_excel("~/Desktop/obesity/2021.xlsx")
edu2020 = read_excel("~/Desktop/obesity/2020.xlsx")
edu2019 = read_excel("~/Desktop/obesity/2019.xlsx")
edu2018 = read_excel("~/Desktop/obesity/2018.xlsx")
edu2017 = read_excel("~/Desktop/obesity/2017.xlsx")

education = rbind(edu2017, edu2018, edu2019, edu2020, edu2021)


california = read_excel("~/Desktop/states/california.xlsx")%>%
  mutate(Location_Desc = "California")
```

```
## New names:
## • `` -> `...3`
```

```
alabama = read_excel("~/Desktop/states/alabama.xlsx")%>%
  mutate(Location_Desc = "Alabama")
```

```
## New names:
## • `` -> `...3`
```

```
alaska = read_excel("~/Desktop/states/alaska.xlsx")%>%
  mutate(Location_Desc = "Alaska")
```

```
## New names:
## • `` -> `...3`
```

```
arizona = read_excel("~/Desktop/states/arizona.xlsx")%>%
  mutate(Location_Desc = "Arizona")
```

```
## New names:
## • `` -> `...3`
```

```
arkansas = read_excel("~/Desktop/states/arkansas.xlsx")%>%
  mutate(Location_Desc = "Arkansas")
```

```
## New names:
## • `` -> `...3`
```

```
colorado = read_excel("~/Desktop/states/colorado.xlsx")%>%
  mutate(Location_Desc = "Colorado")
```

```
## New names:
## • `` -> `...3`
```

```
connecticut = read_excel("~/Desktop/states/connecticut.xlsx")%>%
  mutate(Location_Desc = "Connecticut")
```

```
## New names:
## • `` -> `...3`
```

```
dc = read_excel("~/Desktop/states/DC.xlsx")%>%
  mutate(Location_Desc = "District of Columbia")
```

```
## New names:
## • `` -> `...3`
```

```
delaware = read_excel("~/Desktop/states/delaware.xlsx")%>%
  mutate(Location_Desc = "Delaware")
```

```
## New names:
## • `` -> `...3`
```

```
florida = read_excel("~/Desktop/states/florida.xlsx")%>%
  mutate(Location_Desc = "Florida")
```

```
## New names:
## • `` -> `...3`
```

```
georgia = read_excel("~/Desktop/states/georgia.xlsx")%>%
  mutate(Location_Desc = "Georgia")
```

```
## New names:
## • `` -> `...3`
```

```
hawaii = read_excel("~/Desktop/states/hawaii.xlsx")%>%
  mutate(Location_Desc = "Hawaii")
```

```
## New names:
## • `` -> `...3`
```

```
idaho = read_excel("~/Desktop/states/idaho.xlsx")%>%
  mutate(Location_Desc = "Idaho")
```

```
## New names:
## • `` -> `...3`
```

```
illinois = read_excel("~/Desktop/states/illinois.xlsx")%>%
  mutate(Location_Desc = "Illinois")
```

```
## New names:
## • `` -> `...3`
```

```
indiana = read_excel("~/Desktop/states/indiana.xlsx")%>%
  mutate(Location_Desc = "Indiana")
```

```
## New names:
## • `` -> `...3`
```

```
iowa = read_excel("~/Desktop/states/iowa.xlsx")%>%
  mutate(Location_Desc = "Iowa")
```

```
## New names:
## • `` -> `...3`
```

```
kansas = read_excel("~/Desktop/states/kansas.xlsx")%>%
  mutate(Location_Desc = "Kansas")
```

```
## New names:
## • `` -> `...3`
```

```
kentucky = read_excel("~/Desktop/states/kentucky.xlsx")%>%
  mutate(Location_Desc = "Kentucky")
```

```
## New names:
## • `` -> `...3`
```

```
louisiana = read_excel("~/Desktop/states/louisiana.xlsx")%>%
  mutate(Location_Desc = "Louisiana")
```

```
## New names:
## • `` -> `...3`
```

```
maine = read_excel("~/Desktop/states/maine.xlsx")%>%
  mutate(Location_Desc = "Maine")
```

```
## New names:
## • `` -> `...3`
```

```
maryland = read_excel("~/Desktop/states/maryland.xlsx")%>%
  mutate(Location_Desc = "Maryland")
```

```
## New names:
## • `` -> `...3`
```

```
massachusetts = read_excel("~/Desktop/states/massachusetts.xlsx")%>%
  mutate(Location_Desc = "Massachusetts")
```

```
## New names:
## • `` -> `...3`
```

```
michigan = read_excel("~/Desktop/states/michigan.xlsx")%>%
  mutate(Location_Desc = "Michigan")
```

```
## New names:
## • `` -> `...3`
```

```
minnesota = read_excel("~/Desktop/states/minnesota.xlsx")%>%
  mutate(Location_Desc = "Minnesota")
```

```
## New names:
## • `` -> `...3`
```

```
mississippi = read_excel("~/Desktop/states/mississippi.xlsx")%>%
  mutate(Location_Desc = "Mississippi")
```

```
## New names:
## • `` -> `...3`
```

```
missouri = read_excel("~/Desktop/states/missouri.xlsx")%>%
  mutate(Location_Desc = "Missouri")
```

```
## New names:
## • `` -> `...3`
```

```
montana = read_excel("~/Desktop/states/montana.xlsx")%>%
  mutate(Location_Desc = "Montana")
```

```
## New names:
## • `` -> `...3`
```

```
national = read_excel("~/Desktop/states/national.xlsx")%>%
  mutate(Location_Desc = "National")
```

```
## New names:
## • `` -> `...3`
```

```
nebraska = read_excel("~/Desktop/states/nebraska.xlsx")%>%
  mutate(Location_Desc = "Nebraska")
```

```
## New names:
## • `` -> `...3`
```

```
nevada = read_excel("~/Desktop/states/nevada.xlsx")%>%
  mutate(Location_Desc = "Nevada")
```

```
## New names:
## • `` -> `...3`
```

```
hampshire = read_excel("~/Desktop/states/new hampshire.xlsx")%>%
  mutate(Location_Desc = "New Hampshire")
```

```
## New names:
## • `` -> `...3`
```

```
jersey = read_excel("~/Desktop/states/new jersey.xlsx")%>%
  mutate(Location_Desc = "New Jersey")%>%
  filter(Year != 2019)
```

```
## New names:
## • `` -> `...3`
```

```
mexico = read_excel("~/Desktop/states/new mexico.xlsx")%>%
  mutate(Location_Desc = "New Mexico")
```

```
## New names:
## • `` -> `...3`
```

```
newyork = read_excel("~/Desktop/states/new york.xlsx")%>%
  mutate(Location_Desc = "New York")
```

```
## New names:
## • `` -> `...3`
```

```
carolina = read_excel("~/Desktop/states/north carolina.xlsx")%>%
  mutate(Location_Desc = "North Carolina")
```

```
## New names:
## • `` -> `...3`
```

```
dakota = read_excel("~/Desktop/states/north dakota.xlsx")%>%
  mutate(Location_Desc = "North Dakota")
```

```
## New names:
## • `` -> `...3`
```

```
ohio = read_excel("~/Desktop/states/ohio.xlsx")%>%
  mutate(Location_Desc = "Ohio")
```

```
## New names:
## • `` -> `...3`
```

```
oklahoma = read_excel("~/Desktop/states/oklahoma.xlsx")%>%
  mutate(Location_Desc = "Oklahoma")
```

```
## New names:
## • `` -> `...3`
```

```
oregon = read_excel("~/Desktop/states/oregon.xlsx")%>%
  mutate(Location_Desc = "Oregon")
```

```
## New names:
## • `` -> `...3`
```

```
pennsylvania = read_excel("~/Desktop/states/pennsylvania.xlsx")%>%
  mutate(Location_Desc = "Pennsylvania")
```

```
## New names:
## • `` -> `...3`
```

```
island = read_excel("~/Desktop/states/rhode island.xlsx")%>%
  mutate(Location_Desc = "Rhode Island")
```

```
## New names:
## • `` -> `...3`
```

```
scarolina = read_excel("~/Desktop/states/south carolina.xlsx")%>%
  mutate(Location_Desc = "South Carolina")
```

```
## New names:
## • `` -> `...3`
```

```
sdakota = read_excel("~/Desktop/states/south dakota.xlsx")%>%
  mutate(Location_Desc = "South Dakota")
```

```
## New names:
## • `` -> `...3`
```

```
tennessee = read_excel("~/Desktop/states/tennessee.xlsx")%>%
  mutate(Location_Desc = "Tennessee")
```

```
## New names:
## • `` -> `...3`
```

```
texas = read_excel("~/Desktop/states/texas.xlsx")%>%
  mutate(Location_Desc = "Texas")
```

```
## New names:
## • `` -> `...3`
```

```
utah = read_excel("~/Desktop/states/utah.xlsx")%>%
  mutate(Location_Desc = "Utah")
```

```
## New names:
## • `` -> `...3`
```

```
vermont = read_excel("~/Desktop/states/vermont.xlsx")%>%
  mutate(Location_Desc = "Vermont")
```

```
## New names:
## • `` -> `...3`
```

```
virginia = read_excel("~/Desktop/states/virginia.xlsx")%>%
  mutate(Location_Desc = "Virginia")
```

```
## New names:
## • `` -> `...3`
```

```
washington = read_excel("~/Desktop/states/washington.xlsx")%>%
  mutate(Location_Desc = "Washington")
```

```
## New names:
## • `` -> `...3`
```

```
wvirginia = read_excel("~/Desktop/states/west virginia.xlsx")%>%
  mutate(Location_Desc = "West Virginia")
```

```
## New names:
## • `` -> `...3`
```

```
wisconsin = read_excel("~/Desktop/states/wisconsin.xlsx")%>%
  mutate(Location_Desc = "Wisconsin")
```

```
## New names:
## • `` -> `...3`
```

```
wyoming = read_excel("~/Desktop/states/wyoming.xlsx")%>%
  mutate(Location_Desc = "Wyoming")
```

```
## New names:
## • `` -> `...3`
```

```
poverty = rbind(alabama, alaska, arizona, arkansas, california, colorado, connecticut, d
c, delaware, florida, georgia, hawaii, idaho, illinois, indiana, iowa, kansas, kentucky,
louisiana, maine, maryland, massachusetts, michigan, minnesota, mississippi, missouri, m
ontana, national, nebraska, nevada, hampshire, jersey, mexico, newyork, carolina, dakot
a, ohio, oklahoma, oregon, pennsylvania, island, scarolina, sdakota, tennessee, texas, u
tah, vermont, virginia, washington, wvirginia, wisconsin, wyoming)
```

Editing the data

We edited the data by first selecting only the relevant columns for poverty data (Year, percentage, and location description) and filtering the data to only include the years 2017-2021. We then merged this poverty data with the obesity data, using the state name and year as the matching criteria, and selected only the relevant columns (Year, state name, poverty percentage, ID, obesity percentage, and sample size). We also merged in the education data, using the year and state name as the matching criteria.

We removed any rows with a location description of “National”.Removing these row ensured that we were only analyzing data at the state level, which was consistent with the scope of our project. Similarly, we removed the row with ID “223574” to remove Florida 2021 from our dataset due to the unavailability of obesity data. Similarly we do not have the data for New Jersey 2019.

Furthermore, we added a new column called “obesity” because we needed a numeric variable that represents the obesity rate for each state and year combination. The original “Data_Value” column that contained the obesity rate values was in character format and needed to be converted to numeric format before we could use it in our analysis. By assigning the converted values to a new column called “obesity,” we created a standardized variable that is easier to work with and can be used in calculations and statistical analyses. Finally, we removed the original “Data_Value” column to avoid confusion and streamline the dataset by keeping only the necessary variables. We treated the column ‘edu’ for the same reason. These editing steps were performed to ensure that we had a clean and organized dataset with the relevant variables for our analysis.

After merging the datasets on poverty, obesity, and education by state and year, the final dataset consists of 253 observations and 7 variables. The variables include year, state, poverty rate, obesity rate, sample size, and the percentage of the population with a Bachelor’s degree or higher. The dataset provides information on the relationship between poverty, education, and obesity in the United States from 2017 to 2021. All missing values and invalid data have been removed from the dataset to ensure accuracy and consistency.

```
poverty = poverty%>%
  select(Year, per, Location_Desc)%>%
  filter (Year %in% c(2017, 2018, 2019, 2020, 2021))

merged <- merge(poverty, obesity, by.x = c("Location_Desc", "Year"), by.y = c("LocationD
esc", "YearStart"))%>%
  select(Year, Location_Desc, per, ID, Data_Value, Sample_Size)

merged = merge(merged, education, by.x = c("Year", "Location_Desc"), by.y = c("Year", "L
ocation_Name"))

merged = merged%>%
  filter(ID != "223574")%>%
  filter(Location_Desc != "National")
merged <- merged%>%
  mutate("obesity" = add_column(as.numeric(merged$Data_Value)))%>%
  select(!Data_Value)
```

```
## Warning: The `.data` argument of `add_column()` must be a data frame as of
## tibble 2.1.1.
```

```
## Warning: The `.data` argument of `add_column()` must have unique names as of tibble
## 3.0.0.
## ℹ Use `.name_repair = "minimal"`.
```

```
merged <- merged%>%
  mutate("education" = add_column(as.numeric(merged$Edu)))%>%
  select(!Edu)
```

Descriptive Data

Maps

Obesity Rate By States

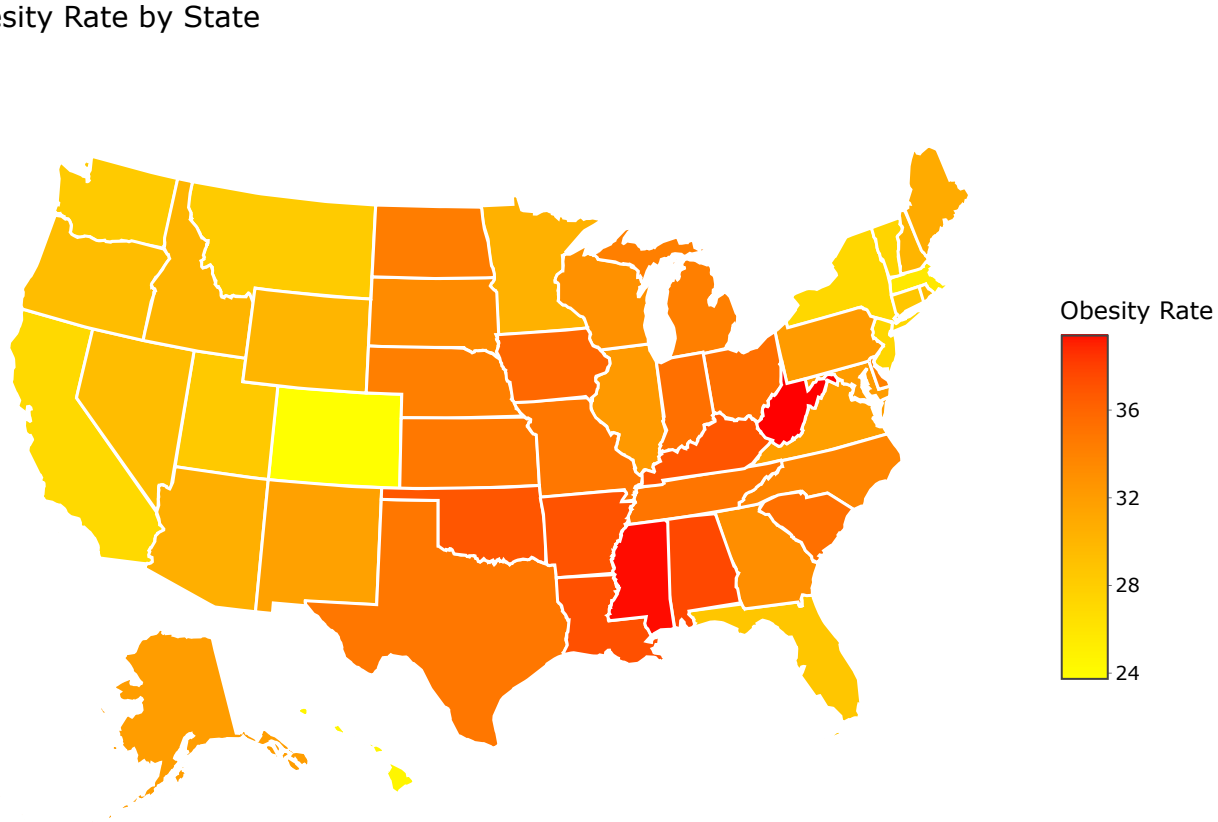
This map shows the obesity rate by state. The scale of the map is from 23.740 (Colorado) to 39.400 (West Virginia). The resulting map shows the obesity rate by state, with darker red shades indicating higher obesity rates. Hovering over each state displays its name and obesity rate.

```
obrate = merged%>%
  group_by(Location_Desc)%>%
  summarise(mean(obesity))%>%
  rename("obesity" = "mean(obesity)")%>%
  rename("state" = "Location_Desc")

obmap <- plot_usmap(data = obrate, values = "obesity", color = "white") +
  aes(text = state) + # Add state names as text aesthetic
  scale_fill_continuous(name = "Obesity Rate", label = scales::comma, , low = "yellow",
high = "red")+
  theme(legend.position = "right")+
  ggtitle("Obesity Rate by State")

obmap <- ggplotly(obmap)

obmap
```



Poverty Rate By States

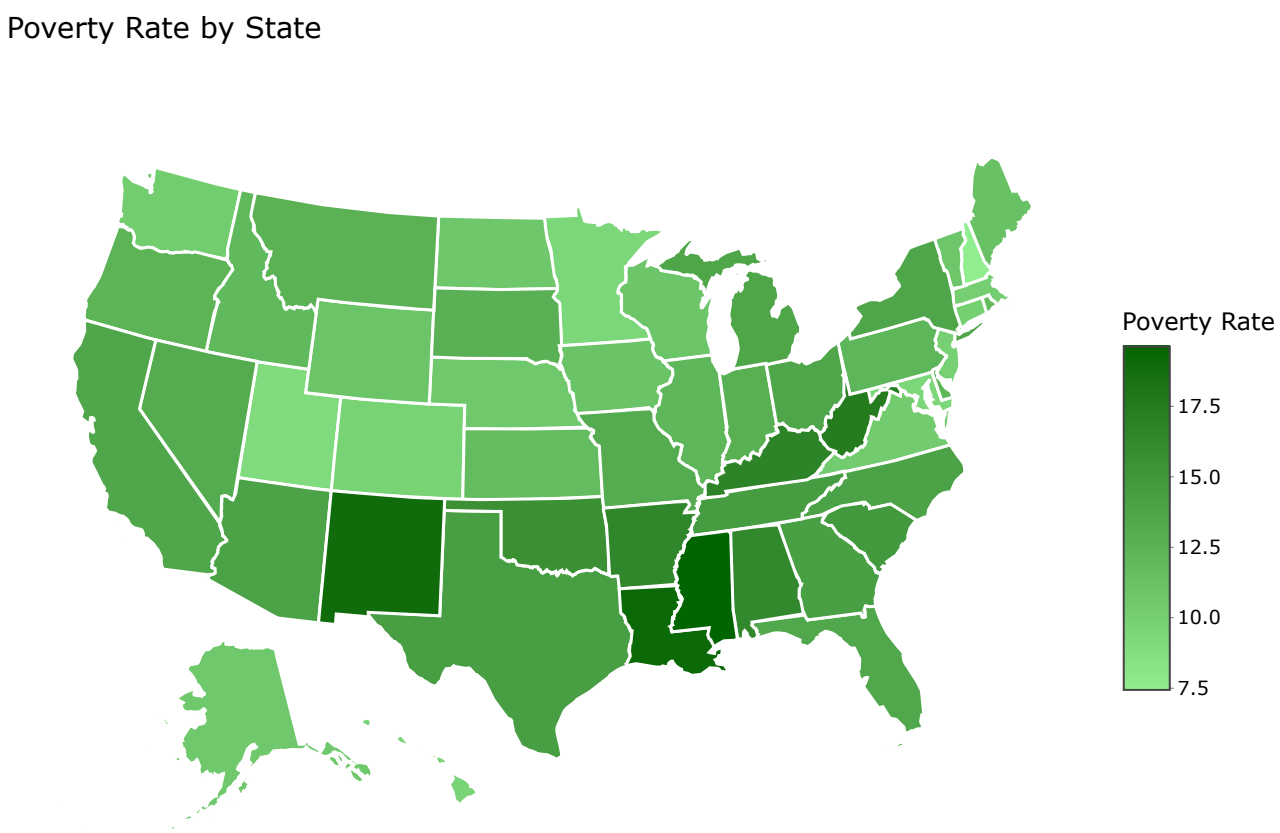
This map shows the poverty rate in each state of the US. The color scale ranges from light green (low poverty rate) to dark green (high poverty rate), with a legend to indicate the corresponding poverty rate values. The scale ranges from 7.44 (New Hampshire) to 19.62 (Mississippi).

```
povrate = merged%>%
  group_by(Location_Desc)%>%
  summarise(mean(per))%>%
  rename("state" = "Location_Desc")


povmap <- plot_usmap(data = povrate, values = "mean(per)", color = "white") +
  aes(text = state) +
  scale_fill_continuous(name = "Poverty Rate", label = scales::comma, , low = "lightgree
n", high = "darkgreen") +
  theme(legend.position = "right")+
  ggtitle("Poverty Rate by State")


povmap <- ggplotly(povmap)

povmap
```



Education Rate By States

This map shows the education rate in each state. The resulting map shows a gradient of blue colors, ranging from light blue to dark blue, with darker shades representing higher education rates. The scale ranges from 21.960 (West Virginia) to 60.800 (District of Columbia).

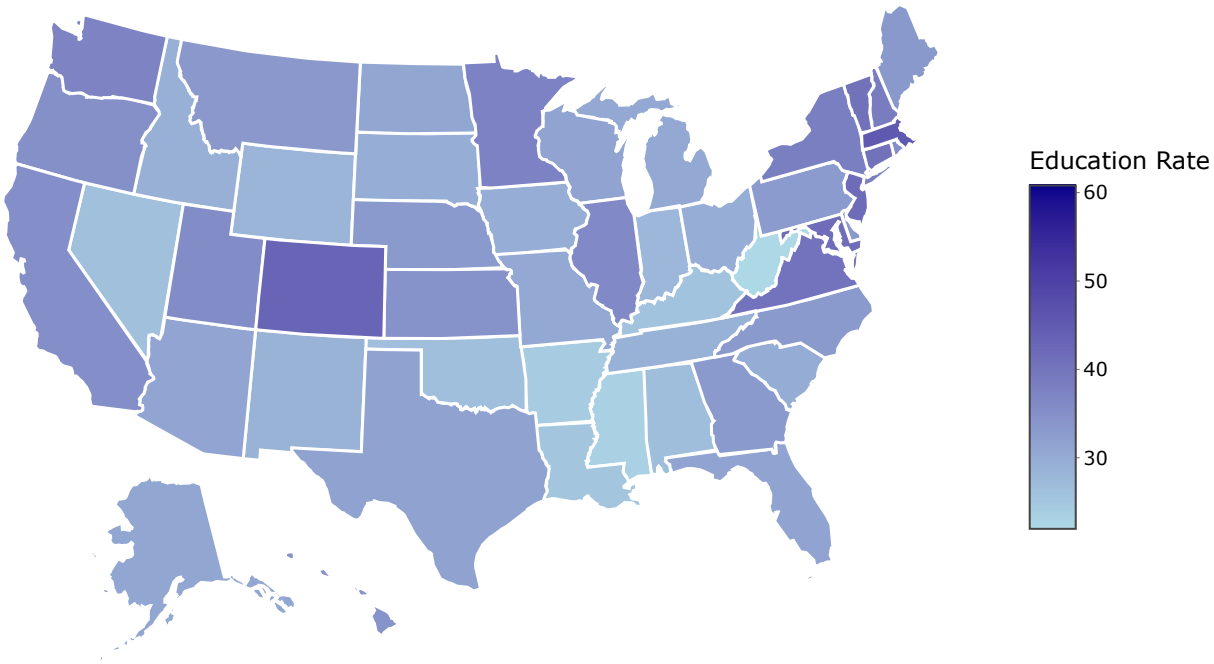
```
edurate = merged%>%
  group_by(Location_Desc)%>%
  summarise(mean(education))%>%
  rename("state" = "Location_Desc")

edumap <- plot_usmap(data = edurate, values = "mean(education)", color = "white") +
  aes(text = state) + # Add state names as text aesthetic
  scale_fill_continuous(name = "Education Rate", label = scales::comma, , low = "lightbl
ue", high = "darkblue") +
  theme(legend.position = "right") +
  ggtitle("Education Rate by State")

edumap <- ggplotly(edumap)

edumap
```

Education Rate by State



Statistics

Means and Medians

The results showed that the mean obesity rate was 31.88% for all years and states combined, the mean poverty rate was 12.72%, and the mean education rate was 33.09%. Additionally, the median obesity rate was 31.9%, the median poverty rate was 12.4%, and the median education rate was 31.9%. These results provide a general

understanding of the average education, obesity, and poverty rates for all the states and years in the dataset.

Minimums and Maximums (obesity)

Looking at the trend in obesity levels over time, the results from the second part of the code show that there is a clear variation in the minimum and maximum obesity levels across the five years in the dataset. The minimum obesity level increased from 22.6% in 2017 to 24.7% in 2021, while the maximum obesity level increased from 38.1% in 2017 to 40.8% in 2019, before decreasing slightly to 40.6% in 2021. These results suggest that the obesity rate in the US has been steadily increasing in recent years, with the peak being in 2019.

```
merged%>%
  summarise("Mean Education Rate" = mean(education),
            "Median Education Rate" = median(education),
            "Mean Obesity Rate" = mean(obesity),
            "Median Obesity Rate" = median(obesity),
            "Mean Poverty Rate" = mean(per),
            "Median Poverty Rate" = median(per))
```

```
##   Mean Education Rate Median Education Rate Mean Obesity Rate
## 1            33.09447                  31.9          31.88221
##   Median Obesity Rate Mean Poverty Rate Median Poverty Rate
## 1                31.9          12.72174                12.4
```

```
merged %>%
  group_by(Year)%>%
  summarise("Minimum Obesity Level" = min(obesity),
            "Maximum Obesity Level" = max(obesity))
```

```
## # A tibble: 5 × 3
##   Year  `Minimum Obesity Level` `Maximum Obesity Level`
##   <chr>                   <dbl>                   <dbl>
## 1 2017                     22.6                    38.1
## 2 2018                     23                      39.5
## 3 2019                     23.8                    40.8
## 4 2020                     24.2                    39.7
## 5 2021                     24.7                    40.6
```

Graphs

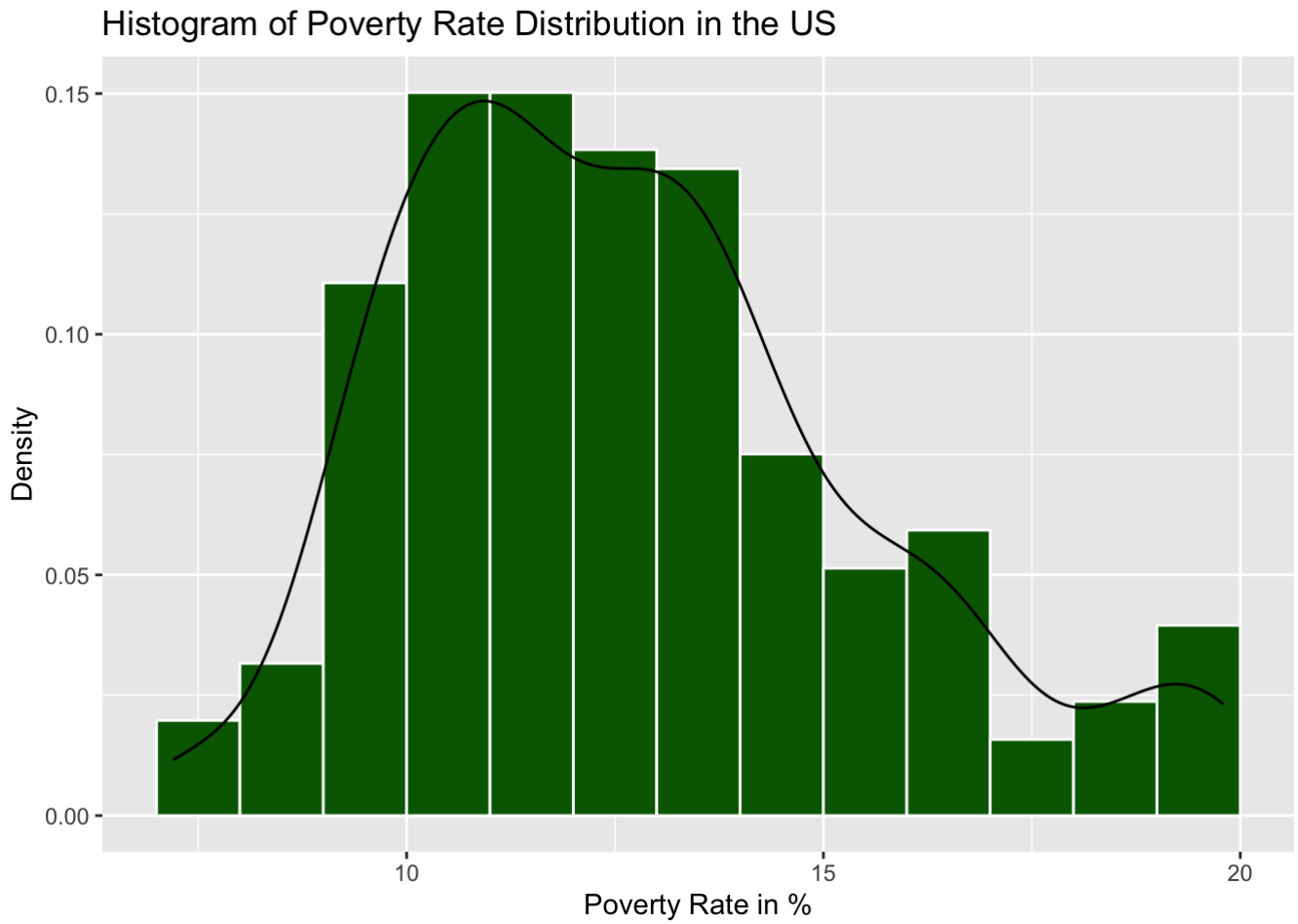
Histogram of Poverty Rate Distribution in the US

The histogram is constructed from a relationship composed of the percentage poverty rate on the x-axis, and the density by U.S. state on the y-axis. The first figure to be identified is the mean with a percentage rate of 12.72%, next at the rate of 12.4% falls the median which is the value xi below and above which 50% of the values in x lie.

Pr(X<=m)>=o.5 and Pr(X>=m)>=0.5

The trend in the graph shows that the percentage poverty rate at a level between 5 percent and 10 percent has an upward surge. This means that in U.S. states at a density level between 0.05 and 0.11, the poverty level varies greatly. While states with a percentage poverty level between 10% and 14% have a density ranging from 0.11 to 0.15. On the left side of the histogram there is a sudden drop in both variables with the poverty level ranging from 0.14% up to 20% , corresponds to an ever decreasing density per state. The histogram shows a unimodal distribution, as it has only a peak at a percentage poverty rate equal to 11. In the end this is a type of a normal distribution, in particular it's a normal curve slopes to the left, known as skewed distribution.

```
merged%>%
  ggplot(aes(x=per))+
  geom_histogram(boundary = 9, binwidth = 1, col = "white", fill = "darkgreen", aes(y=af
ter_stat(density)))+
  geom_density() +
  labs(title = "Histogram of Poverty Rate Distribution in the US", x = "Poverty Rate in
%", y = "Density")
```

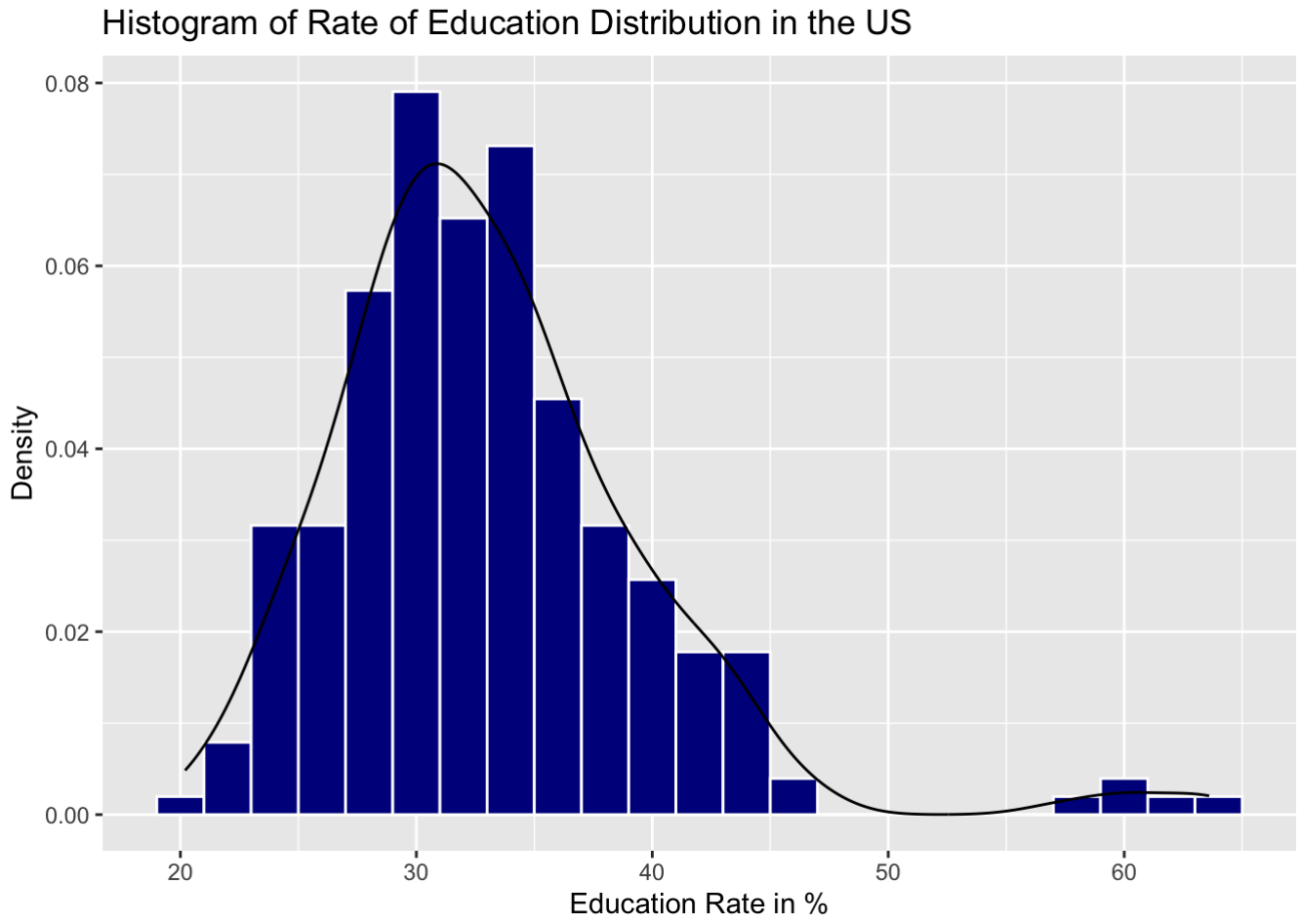


Histogram of Rate of Education Distribution in the US

The histogram shows a relationship composed with the education rate on the x-axis, and the density by U.S. state on the y-axis. The first figure to be identified is the mean with a percentage rate of 33.09%, next at the rate of 31.9% falls the median. The trend shows a positively skewed distribution with an unimodal distribution, as it has only a peak at an education rate equal to 30%. It is clear that as the density is higher the education rate

grows, even though , the highest rates of education falls in the interval between 27% and 35%, and this interval correspond to a range of density between 0.04 and 0.08. The lowest education rates are concentrated in the right part of the slope and they reach a education rate of 60%/75%

```
merged%>%
  ggplot(aes(x=education))+
  geom_histogram(boundary = 9, binwidth = 2, col = "white", fill = "darkblue", aes(y=aft
er_stat(density)))+
  geom_density() +
  labs(title = "Histogram of Rate of Education Distribution in the US", x = "Education R
ate in %", y = "Density")
```

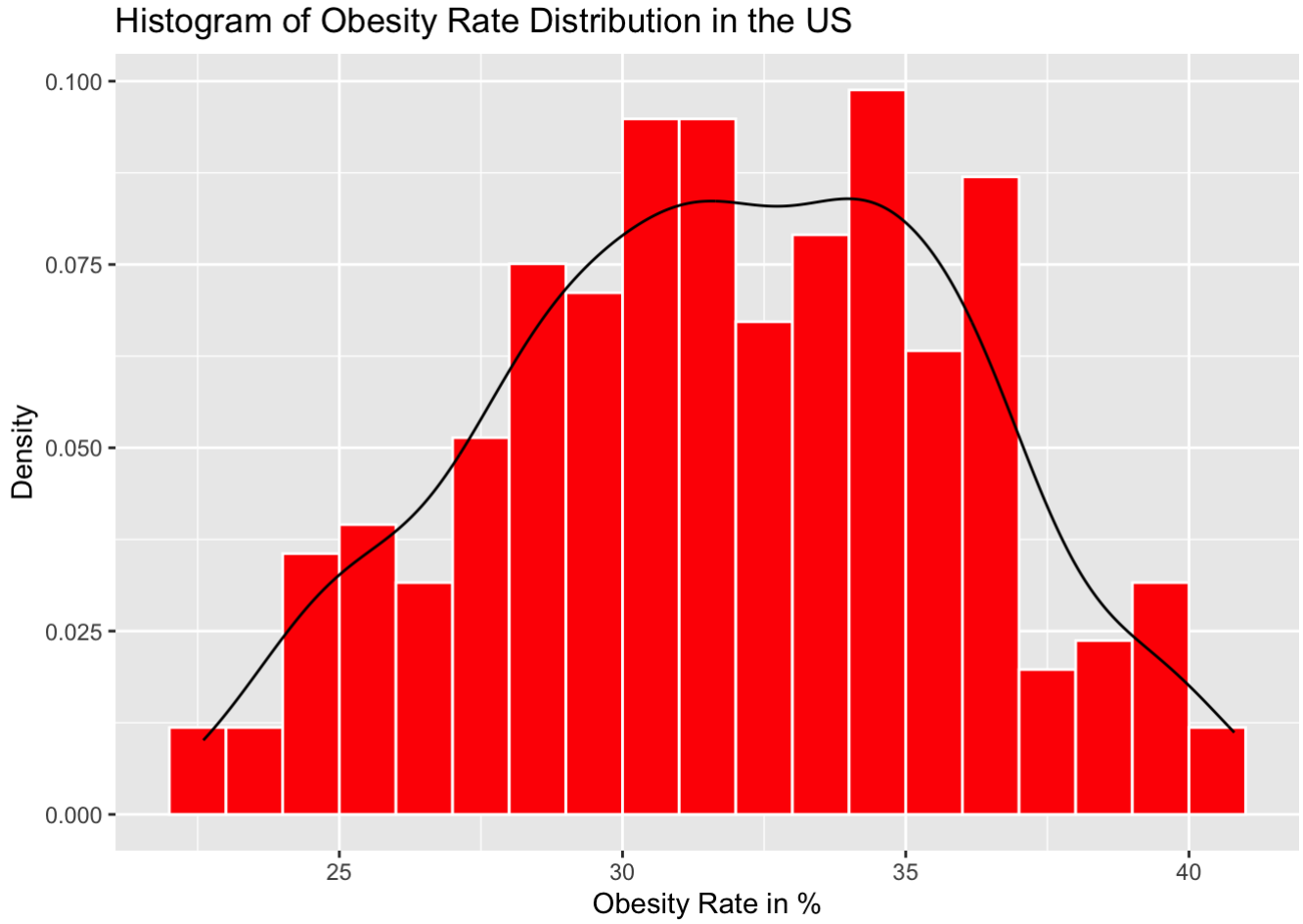


Histogram of Obesity Rate Distribution in the US

This histogram shows a flat normal distribution (or flattened Gaussian distribution), with on the x-axis the obesity rate in percentage, and the density by U.S. state on the y-axis. The first figure to be identified is the mean with a percentage rate of 31.88%, next at the rate of 31.9% falls the median. In general, we can infer that high levels of obesity correspond to high density states. Although the general trend in the graph communicates that at a level greater than or equal to 40 percent obesity-that is, the highest observed in our data-corresponds to a very low density around 0.020. The histogram shows a unimodal distribution, as it has only a peak at a percentage obesity rate that corresponds to the interval of 34%/35%, But it has a peak that is very close to the latter, which falls in the 30%/32% obesity range and reaches a density per state of 0.90.Histogram of obesity rate distribution in the

us This histogram shows a density plot of obesity rates in the US with boundary = 9 specifying that the first bin should start at 9 (assuming obesity rates are measured in percentages), binwidth = 1 specifying that each bin should have a width of 1%.

```
merged%>%
  ggplot(aes(x= obesity))+
  geom_histogram(boundary = 9, binwidth = 1, col = "white", fill = "red", aes(y=after_st
at(density)))+
  geom_density() +
  labs(title = "Histogram of Obesity Rate Distribution in the US", x = "Obesity Rate in
%", y = "Density")
```

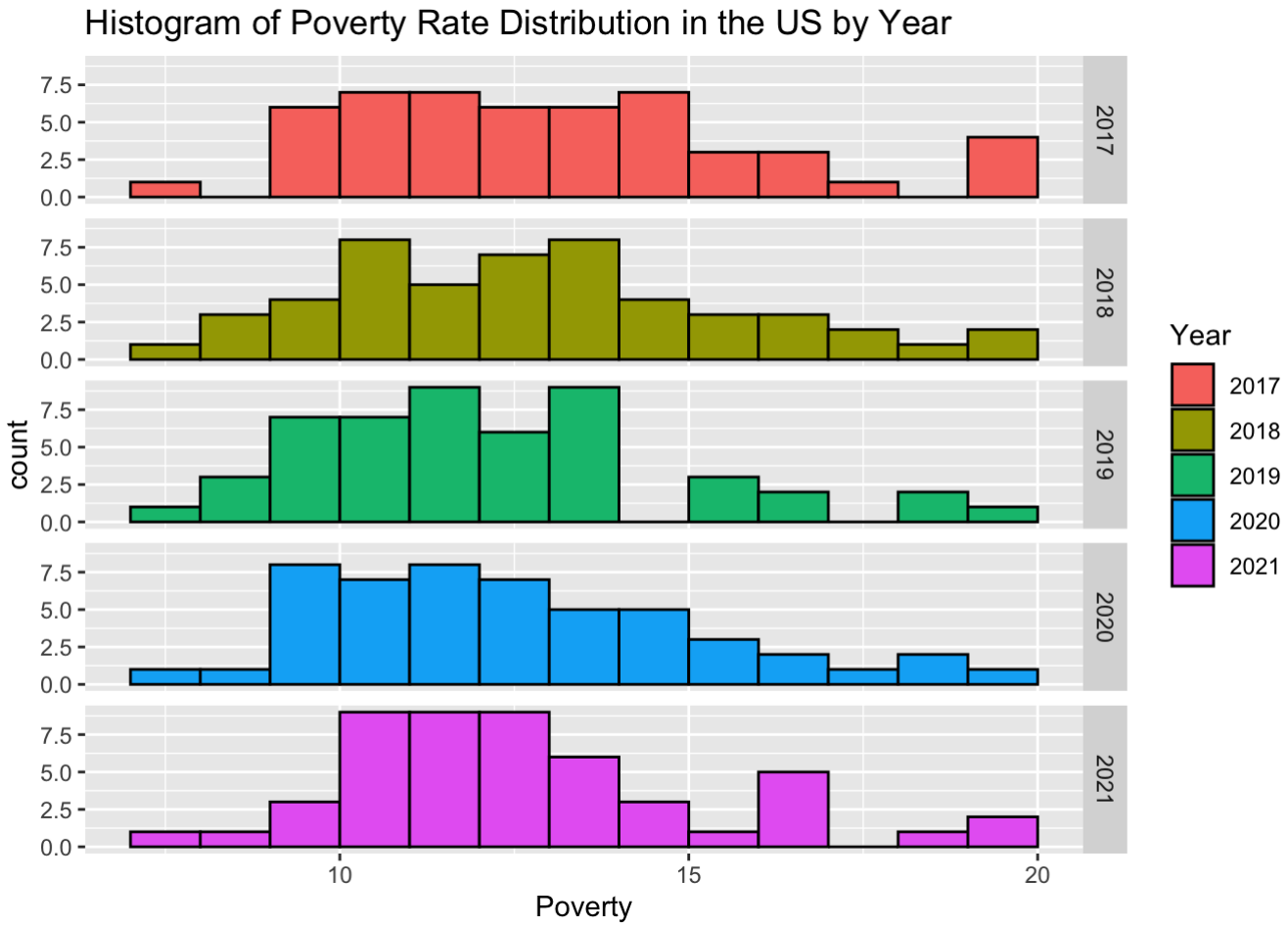


Histogram of Poverty Rate Distribution in the US by Year

The histogram trend on the distribution of education rate by year in the United States, over 5 years we can say that it varies.The poverty rate is a value that can be increased and decreased over time due to multiple factors. From 2017 to 2018 we see a change in the distribution, from a more even one we begin to see peaks at a l height of greater than 7.5. Indeed, in 2019 we see a more noticeable change by comparing it with the 2018 graph. The data are concentrated in the first half of the graph on the left, and there are even higher peaks of poverty.Over the past two years, no abysmal difference is found in the percentage poverty rate. Certainly we can see shifts in the rate and a growth in the right side of the graph from a minimum rate we reach 2.5. In 2021 we have a concentration of the rate evident in the right and middle part of the histogram where three bars at a poverty rate ranging from 10 to 13 reach over 7.5. Over the past two years, no abysmal difference is found in the percentage

poverty rate. Certainly we can see shifts in the rate and a growth in the right side of the graph from a minimum rate we reach 2.5. In 2021 we have a concentration of the rate evident in the right and middle part of the histogram where three bars at a poverty rate ranging from 10 to 13 reach over 7.5.

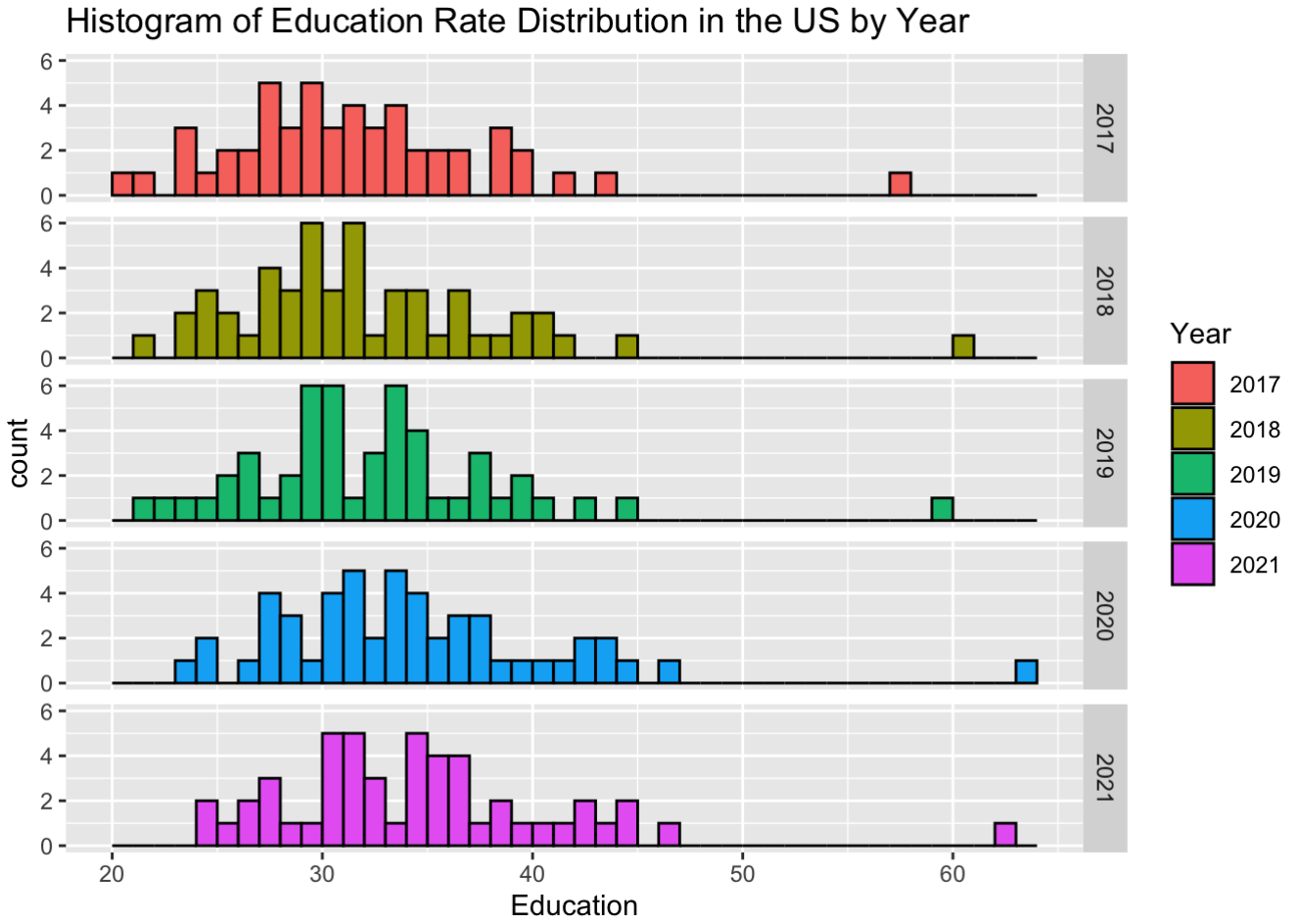
```
merged%>%
  ggplot(aes(x = per, color=, fill=Year))+
  geom_histogram(boundary = 9, binwidth = 1, col = "black")+
  facet_grid(rows = vars(Year))+
  labs(title = "Histogram of Poverty Rate Distribution in the US by Year", x = "Povert
y")
```



Histogram of Education Rate Distribution in the US by Year(density)

The rate of education over the years has not changed much, although in 2018 we have an increase at a rate of 27 and 30.One thing evident from the graphs is that we do not have an even distribution for all levels of education but they are concentrated over the years at a low average level ranging from 15 to 45 maximum and then we have over the years outliers more or less always located at an education level of 60 or more.It is much more common to find an average education rate between 30 and 40 than very low then before 20 or very high, even missing data ranging from 40 to 60.

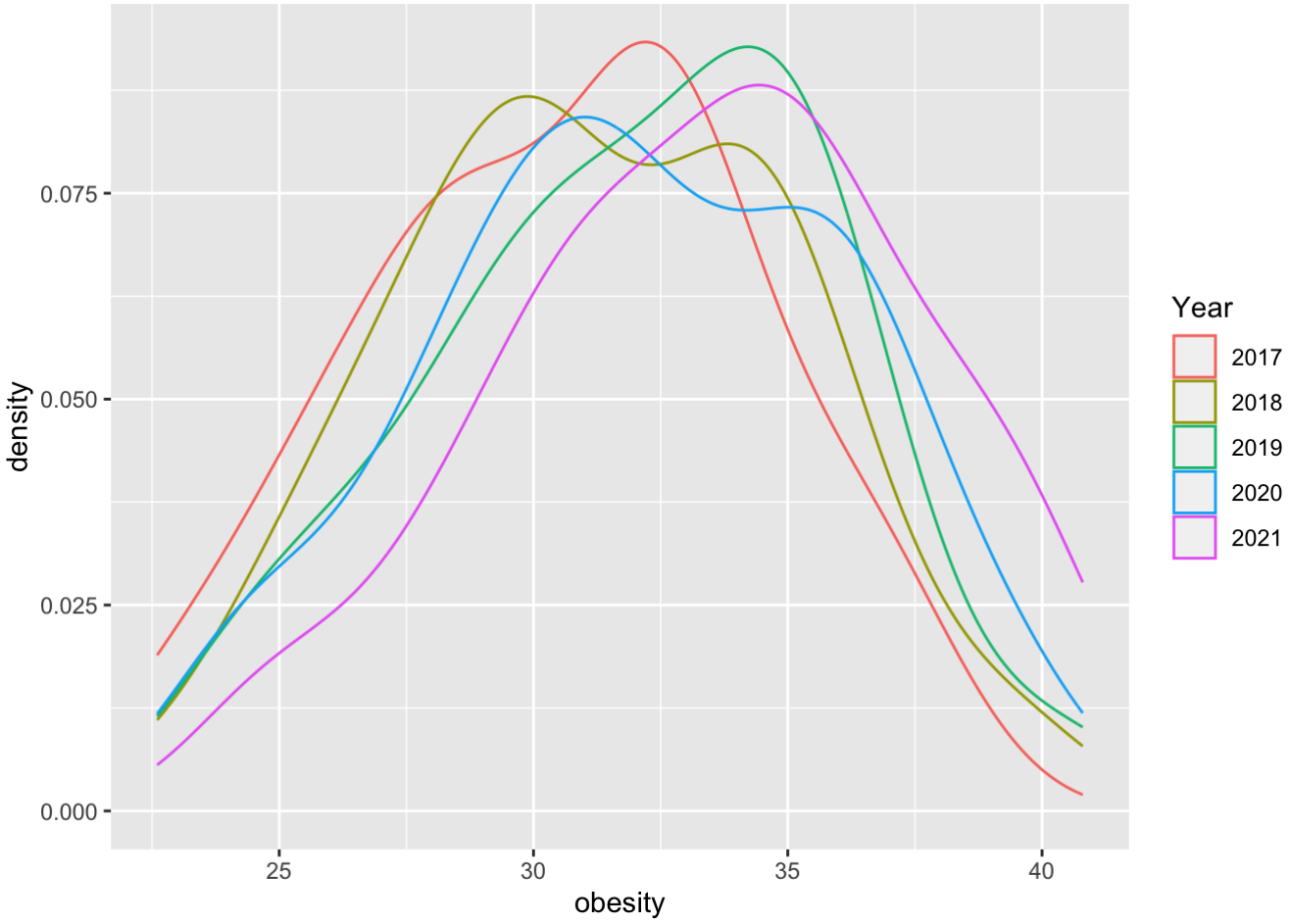
```
merged%>%
  ggplot(aes(x = education, color=, fill=Year))+
  geom_histogram(boundary = 9, binwidth = 1, col = "black")+
  facet_grid(rows = vars(Year))+
  labs(title = "Histogram of Education Rate Distribution in the US by Year", x = "Educat
ion")
```



Density Plot of Obesity Rate Distribution in the US by year

A density plot can be seen as an extension of the histogram. As opposed to the histogram, the density plot can smooth out the distribution of values and reduce the noise.It visualizes the distribution of data over a given period, and the peaks show where values are concentrated.As such density plots work better at determining the distribution shape because they're not affected by the number of bins. The peaks of the obesity distribution over the time period of five years display where values are concentrated over the interval. In 2017, the obesity rate peaks at 32.5 with a density reaching nearly 0.100. In 2018 the obesity rate peaks at 30 with a density reaching nearly 0.080. In 2019 the obesity rate peaks at 34 with a density that is very close to that of the previous year but does not quite reach it. In 2020, the obesity rate peaks at 31 with a density approaching 0.078. In 2021, the obesity rate peaks at 34.5 with a density that fully reaches 0.080. It can be said that the year in which the obesity rate in accordance with density reaches the absolute highest levels is 2017, next in order we find 2017,2021,2018,2020.This report shows the fluctuating trend in obesity rates over the years where the peaks all remain within a range of 30 to 35.

```
merged%>%
  ggplot(aes(x = obesity, color=Year))+
  geom_density()
```

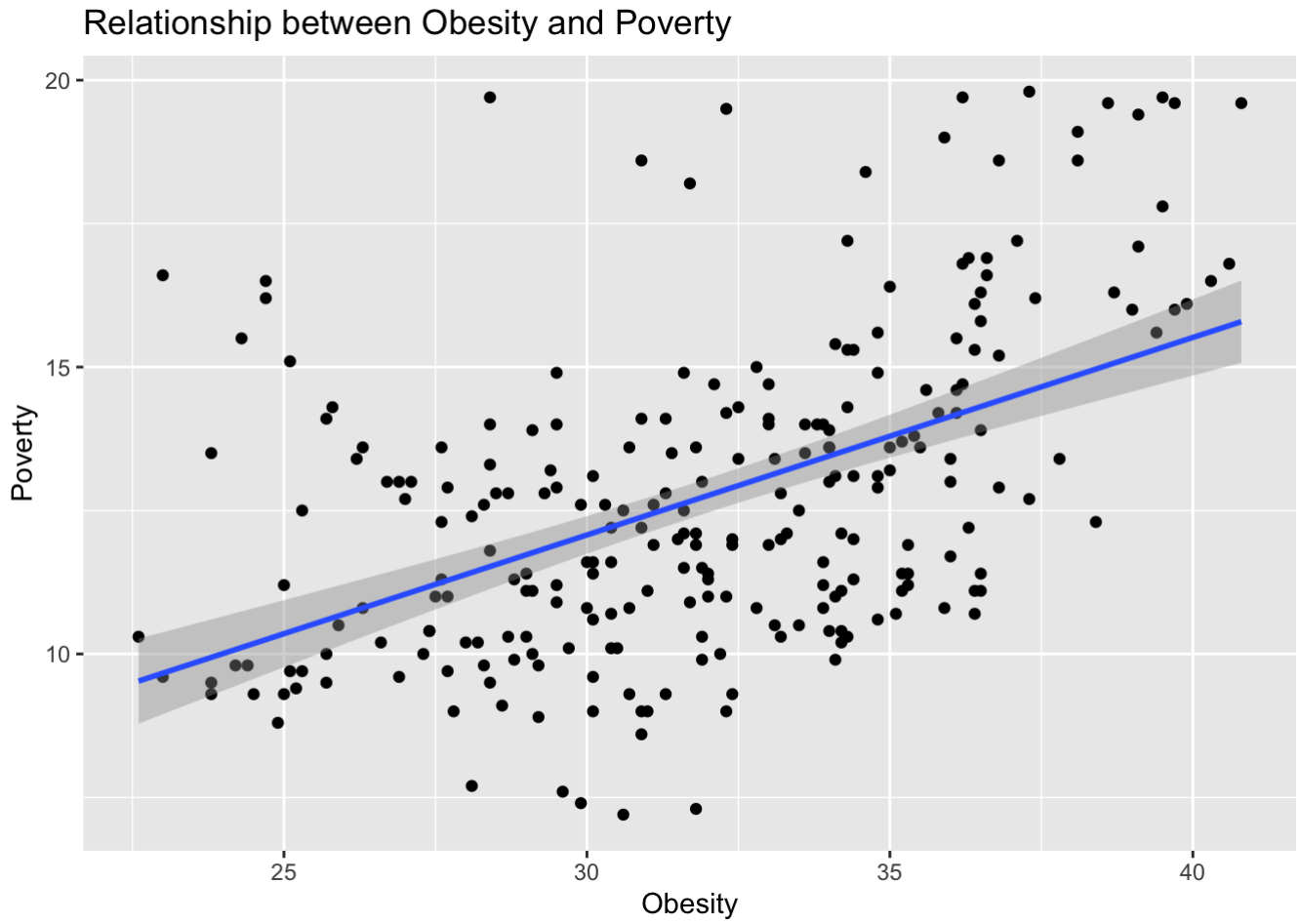


Relationship between Obesity and Poverty (scatter plot)

This scatter plot describe the correlation between Obesity(x) and poverty (y). The x-axis represents the independent variable and the y-axis represents the dependent variable. One important component to a scatter plot is the direction of the relationship between the two variables.Obesity and poverty have a positive association because the above-average values of obesity tend to accompany above-average values of poverty,and the below-average values also tend to occur tougher.The graph shows a positive relationship, as the value of obesity increase, the values of education also increase. States with the highest level of obesity (on the right of the graph) have a higher poverty rate, while those with a lower level of obesity have a lower poverty rate. The graph shows that those who suffer the most from obesity are in a condition of poverty, on the contrary on the right part of the graph, people who suffer less from obesity are in a better economical condition. The form of the relationship between the two variables is crucial to understand the graph,this is a linear relationship, this means that the points on the scatterplot closely resemble a straight line, and usually a relation is linear because obesity rate increases when poverty rate does the same. In addition, the strength of the relationship between the two variables says a lot. The poverty rate is positively associated with obesity rate, with a coefficient of 0.742 and a significance level of ($p < 0.01$). The relationship between two variables is generally considered strong when their r value is larger than 0.7 in this case there is a moderate ,positive, linear association between obesity and poverty.

```
merged%>%
  ggplot() +
  aes(x = obesity, y = per) +
  geom_point() +
  geom_smooth(method = lm) +
  labs (title = "Relationship between Obesity and Poverty", x = "Obesity", y = "Povert
y")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

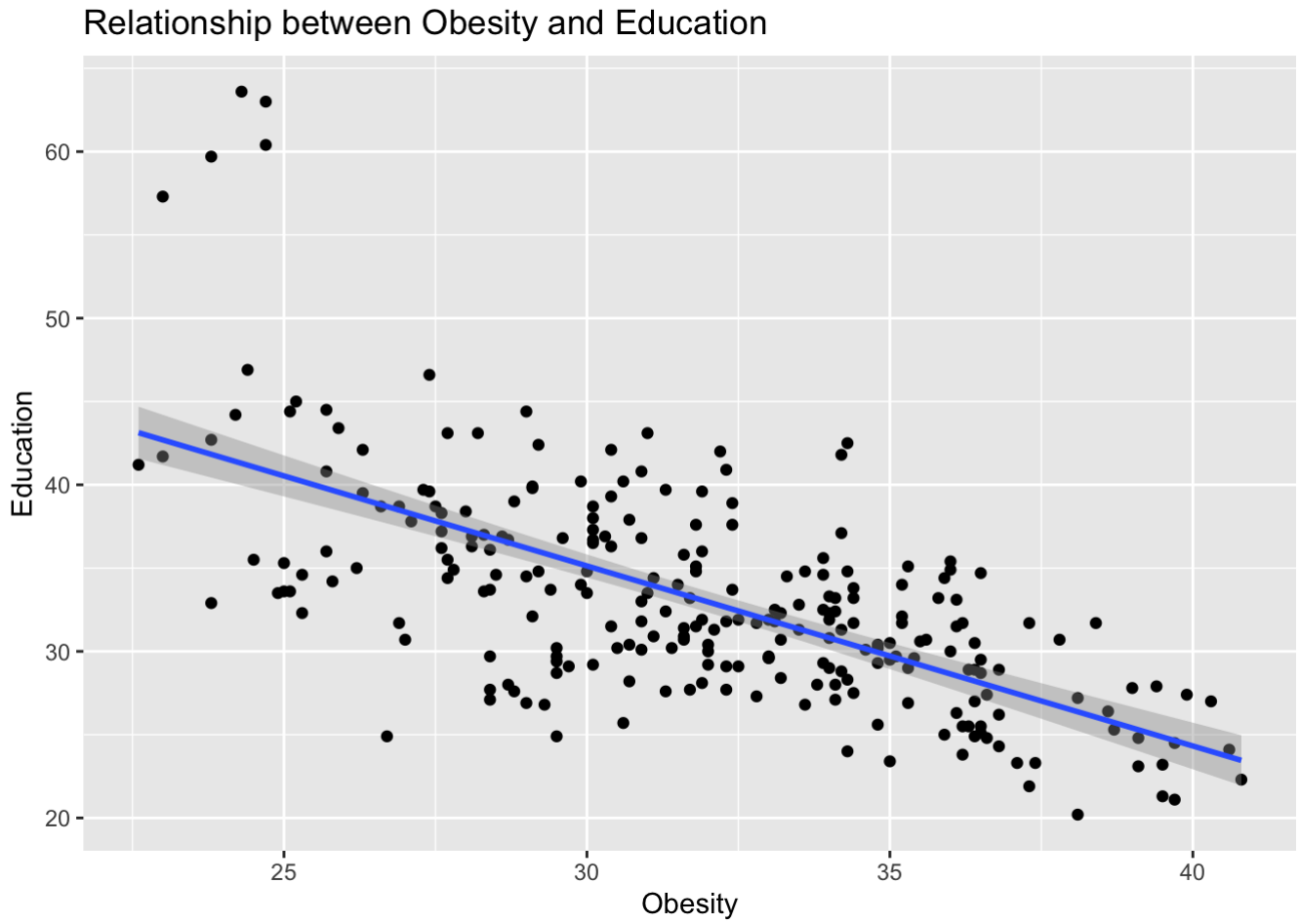


Relationship between Obesity and Education(scatter plot)

This scatterplot describe the correlation between Obesity(x) and Education (y). The x-axis represents the independent variable and the y-axis represents the dependent variable. The regression line shows us a negative association w because in general, as the obesity rate decrease their education level increases. This can be seen in the regression part as the coefficient of education is -0.400 , this is statistically significant with a p-value of less than 0.001, which means that states with the highest level of obesity (on the left of the graph) have a lower education rate, while those with a lower level of obesity have a higher education rate. The result makes perfect sense, when you consider that generally those who have an education are usually able to understand the risks associated with a lifestyle harmful to health, and are oriented towards a healthier diet;This is confirmed by the r-squared value. this is a linear relationship, this means that the points on the scatterplot closely resemble a straight line.

```
merged%>%
  ggplot() +
  aes(x = obesity, y = education) +
  geom_point() +
  geom_smooth(method = lm) +
  labs (title = "Relationship between Obesity and Education", x = "Obesity", y = "Educat
ion")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Regression

```
Regression1 = lm(obesity ~ education, merged)
Regression2 = lm(obesity ~ per, merged)
Regression3 = lm(obesity ~ per*education, merged)
```

Regression 1

The first regression analysis was conducted to investigate the relationship between obesity rates and education levels across all states and years from the merged dataset. The results showed that the coefficient of education was -0.400, indicating that there is a negative relationship between education and obesity rates. This coefficient is statistically significant with a p-value of less than 0.001(***).Therefore, as the education level increases, the

obesity rate tends to decrease. The constant term in the regression equation was 45.111, which represents the predicted obesity rate when education level is equal to zero. This number is statistically significant with a p-value of less than 0.001 (***). The R-squared value was 0.432, indicating that the model explains 43.2% of the variation in obesity rates. The adjusted R-squared value was 0.430, which is a slightly lower value than the R-squared value. This suggests that the additional variable (education) did not improve the model significantly. The residual standard error was 3.073, which represents the standard deviation of the residuals or the amount of variation in the data that is not explained by the model. The F-statistic was 190.882 with a p-value of less than 0.001(***), indicating that the model is statistically significant overall.

Overall, the results of the regression model suggest that education is a significant predictor of obesity rates across all states and years of the dataset. The negative coefficient of education indicates that as education levels increase, obesity rates tend to decrease. However, the relatively low R-squared value suggests that there may be other important factors that contribute to obesity rates that are not accounted for in the model.

```
stargazer(Regression1, type = "text", title = "Regression Results")
```

```
##
## Regression Results
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                obesity
## ------------------------------------------------
## education                     -0.400***
##                                (0.029)
##
## Constant                      45.111***
##                                (0.977)
##
## ------------------------------------------------
## Observations                     253
## R2                              0.432
## Adjusted R2                     0.430
## Residual Std. Error       3.073 (df = 251)
## F Statistic           190.882*** (df = 1; 251)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

Regression 2

In this regression, we used the poverty rate (per) as an independent variable to predict the obesity rate (obesity) in the merged dataset. The resulting regression table, displayed using the stargazer function, shows that the poverty rate is positively associated with obesity rate, with a coefficient of 0.742 and a significance level of *** ($p < 0.01$). This suggests that as the poverty rate increases, so does the obesity rate. The standard error for the poverty rate coefficient is 0.080, indicating the precision of the estimated coefficient. This indicates that the estimated coefficient for per may be off by around 0.080 units from the true population coefficient. The relatively low standard error of 0.080 suggests that the coefficient estimate for per is likely to be relatively close to the true population coefficient, and we can have a reasonable degree of confidence in the estimate. Additionally, the constant term was 22.445, which represents the expected obesity rate when the poverty rate is zero. The regression model explained 25.5% of the variance in obesity rates, as indicated by the R-squared value. The

adjusted R-squared value, which adjusts for the number of independent variables in the model, was 0.252. The residual standard error was 3.518, indicating the average distance between the predicted and actual obesity rates. The F statistic for the model is 86.050 with a significance level of *** ($p < 0.01$), indicating that the model is significant overall. This model suggests that poverty rate is a significant predictor of obesity rate, with an increase in poverty rate associated with an increase in obesity rate.

These results suggest that there is a positive relationship between poverty rate and obesity rate, and that poverty rate is a statistically significant predictor of obesity rate in the merged dataset. However, the model's low R-squared value indicates that other factors not included in the model, such as genetics or lifestyle choices, also play a role in obesity rates. Additionally, the coefficient of determination suggests that poverty rate alone may not be enough to accurately predict obesity rates, and further research is needed to identify other factors that may impact obesity rates.

```
stargazer(Regression2, type = "text", title = "Regression Results")
```

```
##
## Regression Results
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                obesity
## ------------------------------------------------
## per                            0.742***
##                                (0.080)
##
## Constant                       22.445***
##                                (1.041)
##
## ------------------------------------------------
## Observations                     253
## R2                              0.255
## Adjusted R2                     0.252
## Residual Std. Error       3.518 (df = 251)
## F Statistic            86.050*** (df = 1; 251)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

Regression 3

Regression 3 is a multiple linear regression model that includes both poverty rate and education rate as independent variables, as well as an interaction term between the two. The results show that the coefficient for poverty rate is 0.754, which is statistically significant at the 0.05 level, indicating a positive relationship between poverty rate and obesity rates. The standard error for this coefficient is 0.318, which suggests that there is some degree of uncertainty around the true value of the coefficient. The coefficient for education rate is -0.157, which is not statistically significant at the 0.05 level, indicating that there is no strong relationship between education rate and obesity rates when poverty rate is taken into account. The standard error for this coefficient is 0.136, which is relatively small compared to the standard error for the poverty rate coefficient, indicating that the estimate for the education coefficient is more precise. The coefficient for the interaction term between poverty rate and education rate is -0.012, which is also not statistically significant at the 0.05 level. This suggests that there is no significant interaction effect between poverty rate and education rate on obesity rates. The standard error for this coefficient

is 0.009, which is relatively small, indicating a more precise estimate. The constant term in the model is 32.502, which represents the predicted obesity rate when poverty rate and education rate are both zero. This constant is statistically significant at the 0.01 level, indicating that it is a meaningful component of the model. The R-squared value for the model is 0.48, which means that the model explains 48% of the variance in obesity rates. The adjusted R-squared value, which takes into account the number of independent variables in the model, is 0.474. The residual standard error for the model is 2.952, which represents the average difference between the predicted and actual values of obesity rate.

In terms of the change in significance level of education rate in the third regression, one possible explanation is the presence of collinearity between education rate and poverty rate. Collinearity occurs when two or more independent variables in a regression model are highly correlated with each other. In this case, education rate and poverty rate may be negatively correlated, as higher education rates could lead to lower poverty rates. This can make it difficult to determine the individual effect of each variable on the dependent variable (obesity rate in this case). Additionally, the inclusion of the interaction term between poverty rate and education rate may have also affected the significance level of the education rate coefficient. Overall, the results suggest that when considering both education rate and poverty rate, it is poverty rate that has a stronger relationship with obesity rates.

```
stargazer(Regression3, type = "text", title = "Regression Results")
```

```
##
## Regression Results
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                obesity
## ------------------------------------------------
## per                             0.754**
##                                 (0.318)
##
## education                       -0.157
##                                 (0.136)
##
## per:education                   -0.012
##                                 (0.009)
##
## Constant                       32.502***
##                                 (4.766)
##
## ------------------------------------------------
## Observations                      253
## R2                               0.480
## Adjusted R2                      0.474
## Residual Std. Error       2.952 (df = 249)
## F Statistic            76.600*** (df = 3; 249)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

Dummy variable

In order to create a binary indicator variable based on poverty rate, we first calculated the 25th and 75th percentiles using the quantile function. We then used the mutate function from the “dplyr” package to create a new variable called “indicator” that takes on a value of 1 if the poverty rate is less than or equal to the 25th percentile, and a value of 0 if the poverty rate is greater than or equal to the 75th percentile. Any observations with poverty rates falling between the 25th and 75th percentiles were excluded from the analysis. This binary indicator variable can be used to investigate differences in obesity rates between areas with high and low poverty rates. This was done in order to control for any strong correlation between the two independent variables (poverty rate and obesity rate) that could affect the regression results. This approach was taken because it was observed that there was a significant difference in the regression results when using only one independent variable versus two.

```
q25 <- quantile(merged$per, 0.25)
q75 <- quantile(merged$per, 0.75)

mergeddummy <- merged %>%
  mutate(indicator = case_when(
    per <= q25 ~ 1,
    per >= q75 ~ 0,
    TRUE ~ NA_real_
  )) %>%
  filter(!is.na(indicator))

Regressiondummy1 = lm(obesity ~ education, mergeddummy)
Regressiondummy2 = lm(obesity ~ indicator, mergeddummy)
Regressiondummy3 = lm(obesity ~ indicator*education, mergeddummy)
```

Dummy regression 1

The regression table shows that there is a negative relationship between education and obesity rates. The model explains 58.2% of the variance in obesity rates, and the constant term is statistically significant at the 0.01 level. The model is statistically significant overall, with a p-value of less than 0.01. The coefficient estimate for education is statistically significant at the 0.01 level, indicating that it is a meaningful predictor of obesity rates in the model. The standard error for the coefficient estimate suggests some uncertainty around the true value of the coefficient.

The estimated coefficient for education in Regression 1 (-0.438) is slightly lower than in the second table (-0.400), which suggests that controlling for poverty through the dummy variable may have a small impact on the relationship between education and obesity. The adjusted R-squared values are also slightly different, with the first table showing a slightly higher value (0.579) compared to the second table (0.430). Overall, adding the dummy variable appears to have a modest effect on the regression results.

```
stargazer(Regressiondummy1, type = "text", title = "Regression Results")
```

```
##
## Regression Results
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                 obesity
## ------------------------------------------------
## education                      -0.438***
##                                 (0.032)
##
## Constant                       46.713***
##                                 (1.124)
##
## ------------------------------------------------
## Observations                      133
## R2                               0.582
## Adjusted R2                      0.579
## Residual Std. Error       3.085 (df = 131)
## F Statistic           182.683*** (df = 1; 131)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

Dummy regression 2

This regression table summarizes the model where obesity is the dependent variable and indicator is a binary variable representing the top and bottom 25% of the poverty rate. The coefficient estimate for the indicator variable is -5.411, showing a significant negative relationship between poverty rates and obesity rates. The constant term in the model is 34.645 and is statistically significant at the 0.01 level. The R-squared value for the model is 0.326, indicating that the model explains 32.6% of the variance in obesity rates. The F-statistic value for the model is 63.388, with a p-value of less than 0.01, indicating that the overall model is statistically significant. The coefficient estimate for the indicator variable is statistically significant at the 0.01 level, suggesting that it is a meaningful predictor of obesity rates in the model.

In Regression 2, the independent variable is the poverty rate itself, in this model, the independent variable is the binary variable indicating whether the poverty rate is in the top or bottom 25%. While both models are statistically significant, the R-squared value is higher in this model (0.326) than in Regression 2, indicating that this model explains more of the variance in obesity rates.

```
stargazer(Regressiondummy2, type = "text", title = "Regression Results")
```

```
##
## Regression Results
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                obesity
## ------------------------------------------------
## indicator                     -5.411***
##                                (0.680)
##
## Constant                      34.645***
##                                (0.479)
##
## ------------------------------------------------
## Observations                    133
## R2                             0.326
## Adjusted R2                    0.321
## Residual Std. Error       3.919 (df = 131)
## F Statistic           63.388*** (df = 1; 131)
## ================================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

Dummy regression 3

This model shows the results of a linear regression where the dependent variable is obesity and the independent variables are an indicator variable representing the top and bottom 25% of the poverty rate, education level, and the interaction between the indicator variable and education level. The coefficient estimate for the indicator variable is -4.359, which suggests that there is a negative relationship between poverty rates and obesity rates. The coefficient estimate for education is -0.380, indicating that there is a significant negative relationship between education level and obesity rates. The interaction term between the indicator variable and education level is positive but not statistically significant, with a coefficient estimate of 0.060. The constant term in the model is 45.800, which represents the predicted obesity rate when the poverty rate is not in the top or bottom 25% and education level is at the reference value. The R-squared value for the model is 0.621, which means that the model explains 62.1% of the variance in obesity rates. The F-statistic value for the model is 70.581, with a p-value of less than 0.01, indicating that the overall model is statistically significant.

```
stargazer(Regressiondummy3, type = "text", title = "Regression Results")
```

```
##
## Regression Results
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                 obesity
## ------------------------------------------------
## indicator                       -4.359
##                                 (3.324)
##
## education                      -0.380***
##                                 (0.041)
##
## indicator:education             0.060
##                                 (0.090)
##
## Constant                       45.800***
##                                 (1.265)
##
## ------------------------------------------------
## Observations                      133
## R2                               0.621
## Adjusted R2                      0.613
## Residual Std. Error       2.960 (df = 129)
## F Statistic            70.581*** (df = 3; 129)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

Bootstrapping Poverty

Through bootstrapping we quantify the uncertainty associated with a given estimator. Having a set of observations from which we want to derive estimates for a parameter T, whose estimate we denote at t.

The steps we took to apply the bootstrap are: 1. We have generated 1000 bootstrap samples of the data using the specified response and explanatory variables. The random samples have a size of 253. Next we calculated the slope coefficient of each bootstrap sample. In order to do so, we distributed their slope into a graph and then added 95% confidence intervals. The resulting slope coefficient estimates are stored in “bootstrap_poverty” 2. We created a histogram of the bootstrap sample slope with a specified binwidth of 0.05 and a boundary of 9. The aes()function specifies the x-axis variable as stat, which corresponds to the slope coefficient estimates. 3. The resulting plot shows the distribution of the bootstrap sample slope estimates, which can be used to estimate the uncertainty in the slope coefficient of the linear regression model.

The results from mean(bootstrap_poverty$stat) indicate that the mean bootstrap sample slope estimate for the relationship between obesity and poverty is 0.739. This means that, on average, for every unit increase in poverty, we expect a 0.739 unit increase in obesity. The results from sd(bootstrap_poverty$stat) indicate that the standard deviation of the bootstrap sample slope estimates is 0.086. This means that the bootstrap sample slope estimates varied by an average of 0.086 around their mean of 0.739. In comparison, the results from mean(bootstrap_education/$stat) indicate that the mean bootstrap sample slope estimate for the relationship between obesity and education is -0.403. This means that, on average, for every unit increase in education, we expect a -0.403 unit decrease in obesity. The results from sd(bootstrap_education$stat) indicate that the standard

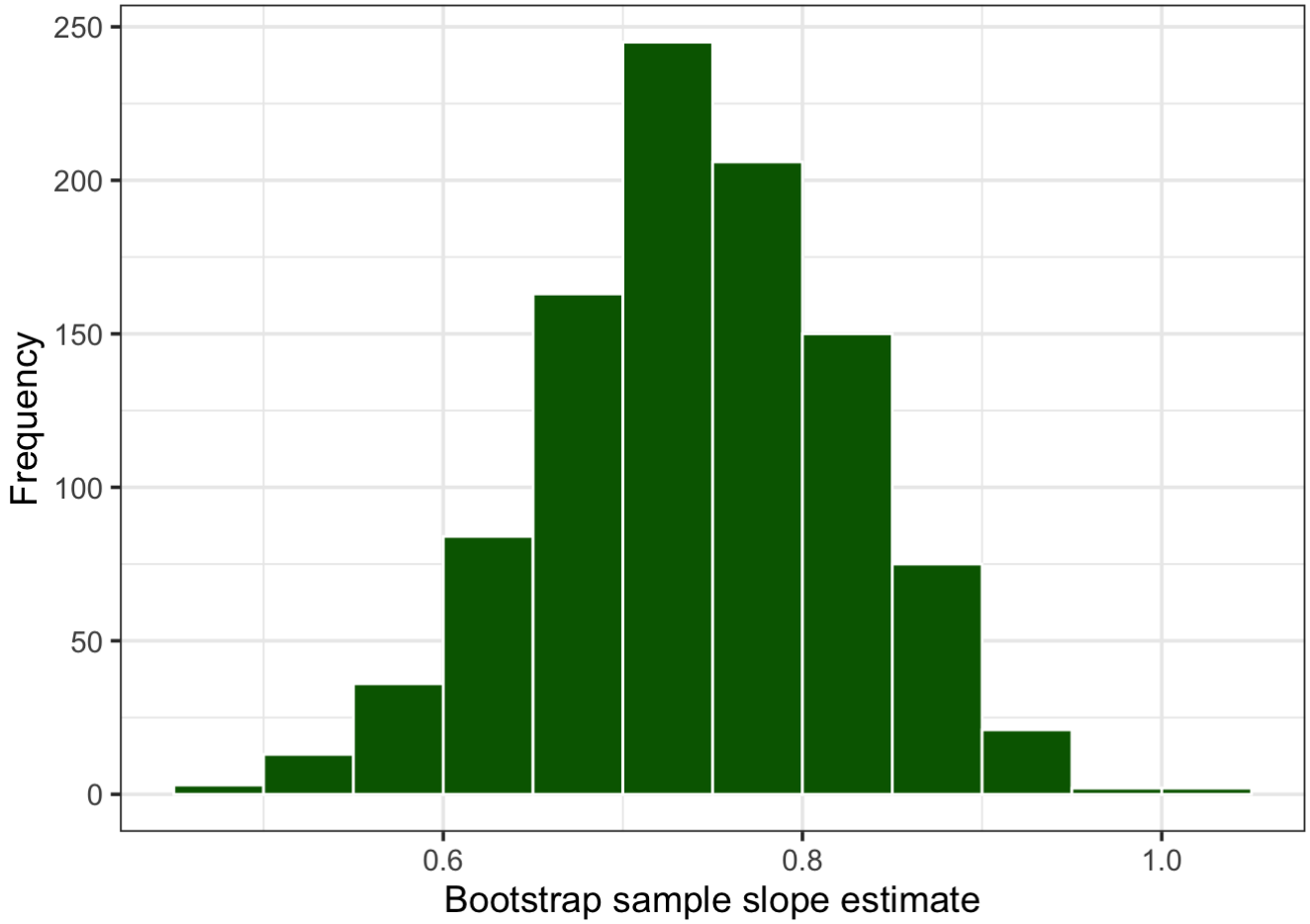
deviation of the bootstrap sample slope estimates is 0.033. This means that the bootstrap sample slope estimates varied by an average of 0.033 around their mean of -0.403. These two results suggest that the relationship between obesity and poverty is positive, while the relationship between obesity and education is negative. However, it's important to note that these results only provide evidence of an association and not necessarily a causal relationship.

Ci-pctile-poverty: we calculated the lower and upper bounds of a 95% confidence interval for the slope coefficient estimate by using the quantile function which calculates the 2.5th and 97.5th percentiles of the bootstrap sample slope estimates.
ci_theory_poverty: by calculating the 95% confidence interval for the slope coefficient estimate using theoretical methods. This is done using the tidy() function applied to a linear regression model (lm()) with obesity as the response variable and per as the explanatory variable. The conf.int = TRUE and conf.level = 0.95 arguments are used to request the confidence interval, with a confidence level of 95%. The resulting object ci_theory_poverty is then filtered to select only the row with the term equal to "per", which corresponds to the slope coefficient estimate. By comparing the confidence intervals obtained from the bootstrap analysis and theoretical methods, we can assess the agreement between these two methods of estimating the uncertainty in the slope coefficient. In this case, the 95% confidence interval obtained from the bootstrap analysis is 0.589 and 0.886, while the 95% confidence interval obtained from theoretical methods is 0.584 and 0.899. These two intervals are fairly similar, which suggests that the bootstrap analysis provides a reasonable estimate of the sampling distribution of the slope coefficient.

This R code generates a histogram to compare the bootstrap and theoretical methods of estimating the 95% confidence interval for the slope coefficient of a linear regression model.The aes() function specifies that the histogram should be based on the stat variable from the bootstrap_poverty object, which contains the bootstrap sample slope estimates.The geom_vline() function adds vertical lines to the plot to indicate the lower and upper bounds of the 95% confidence interval obtained from each method. The x intercept argument specifies the location of the vertical lines, which is obtained from the lower and upper columns of the ci_pctile_poverty object for the percentile method, and the conf.low and conf.high columns of the ci_theory_poverty object for the theoretical method. Overall, this plot allows us to visually compare the sampling distribution of the slope coefficient estimated by the bootstrap method with the 95% confidence intervals obtained from the percentile and theoretical methods. It also allows us to compare the agreement between the two methods in terms of the width and location of the confidence intervals.

```
bootstrap_poverty <- merged %>%
  specify(response = obesity, explanatory = per) %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "slope")

bootstrap_poverty %>%
  ggplot(aes(x = stat)) +
  geom_histogram(boundary = 9, binwidth = 0.05, col = "white", fill = "darkgreen") +
  labs(
    x = "Bootstrap sample slope estimate",
    y = "Frequency"
  ) +
  theme_bw(base_size = 14)
```




```
mean(bootstrap_poverty$stat)
```

```
## [1] 0.7419884
```

```
sd(bootstrap_poverty$stat)
```

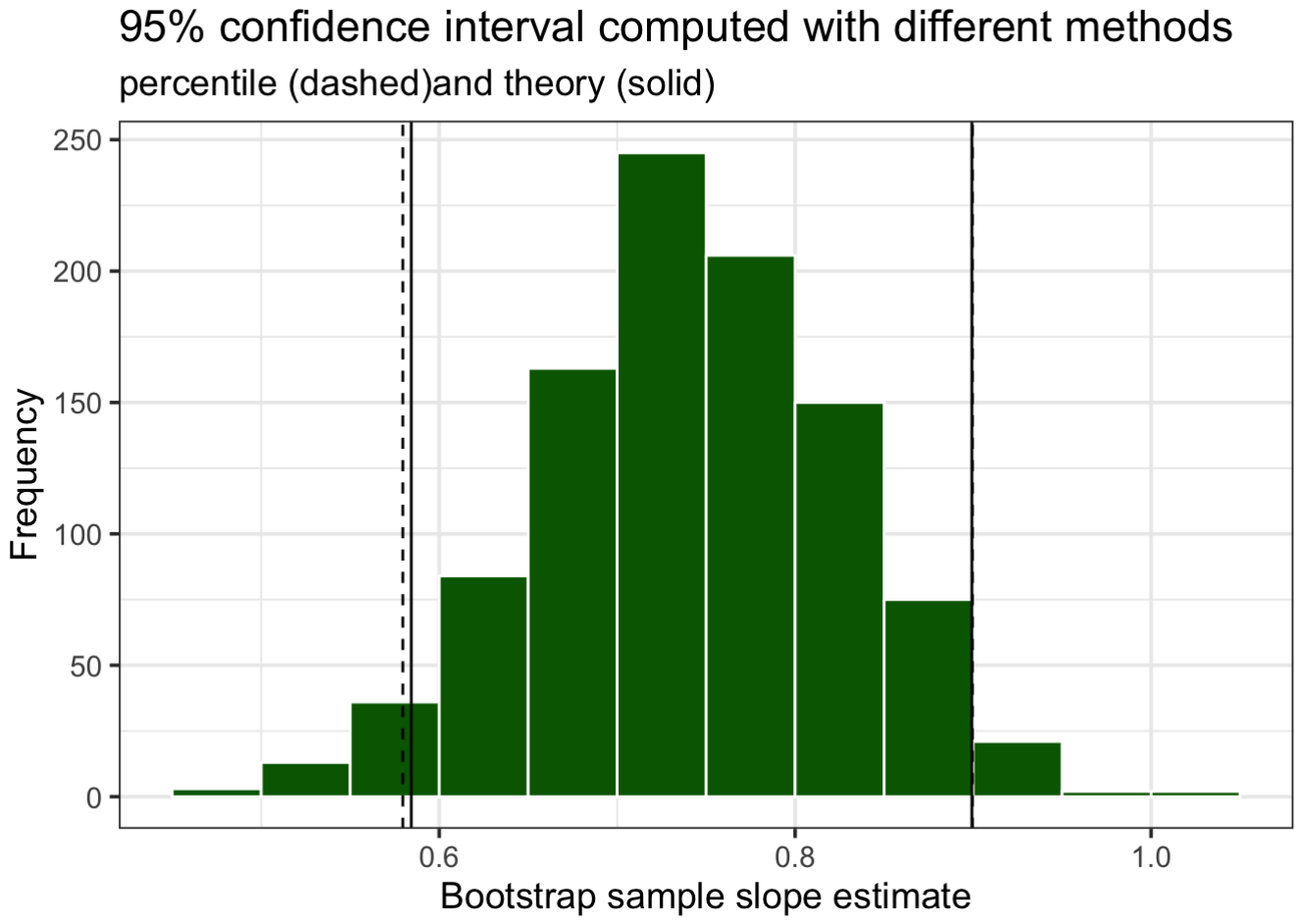
```
## [1] 0.08486685
```

```
ci_pctile_poverty = bootstrap_poverty %>%
  summarise(
    lower = quantile(stat, 0.025),
    upper = quantile(stat, 0.975)
  )

ci_theory_poverty <- tidy(lm(obesity ~ per, merged),
                  conf.int = TRUE, conf.level = 0.95) %>%
  filter(term == "per") %>%
  select(term, conf.low, conf.high)
ci_theory_poverty
```

```
## # A tibble: 1 × 3
##   term  conf.low conf.high
##   <chr>    <dbl>     <dbl>
## 1 per      0.584     0.899
```

```
bootstrap_poverty %>%
  ggplot(aes(x = stat)) +
  geom_histogram(boundary = 1, binwidth = 0.05, col = "white", fill = "darkgreen") +
  labs(
    x = "Bootstrap sample slope estimate",
    y = "Frequency",
    title = "95% confidence interval computed with different methods",
    subtitle = "percentile (dashed)and theory (solid)"
  ) +
  geom_vline(xintercept = c(ci_pctile_poverty$lower, ci_pctile_poverty$upper), linetype
= "dashed", show.legend = TRUE) +
  geom_vline(xintercept = c(ci_theory_poverty$conf.low, ci_theory_poverty$conf.high)) +
  theme_bw(base_size = 14)
```



Bootstrapping education

With this bootstrap analysis we estimated the slope of the linear regression model that describes the relationship between obesity rate and the education level in the dataset. We have generated 1000 bootstrap samples of the data using the specified response and explanatory variables. The random samples have a size of 253. The resulting bootstrap estimates of the slope are stored in the object bootstrap_education.

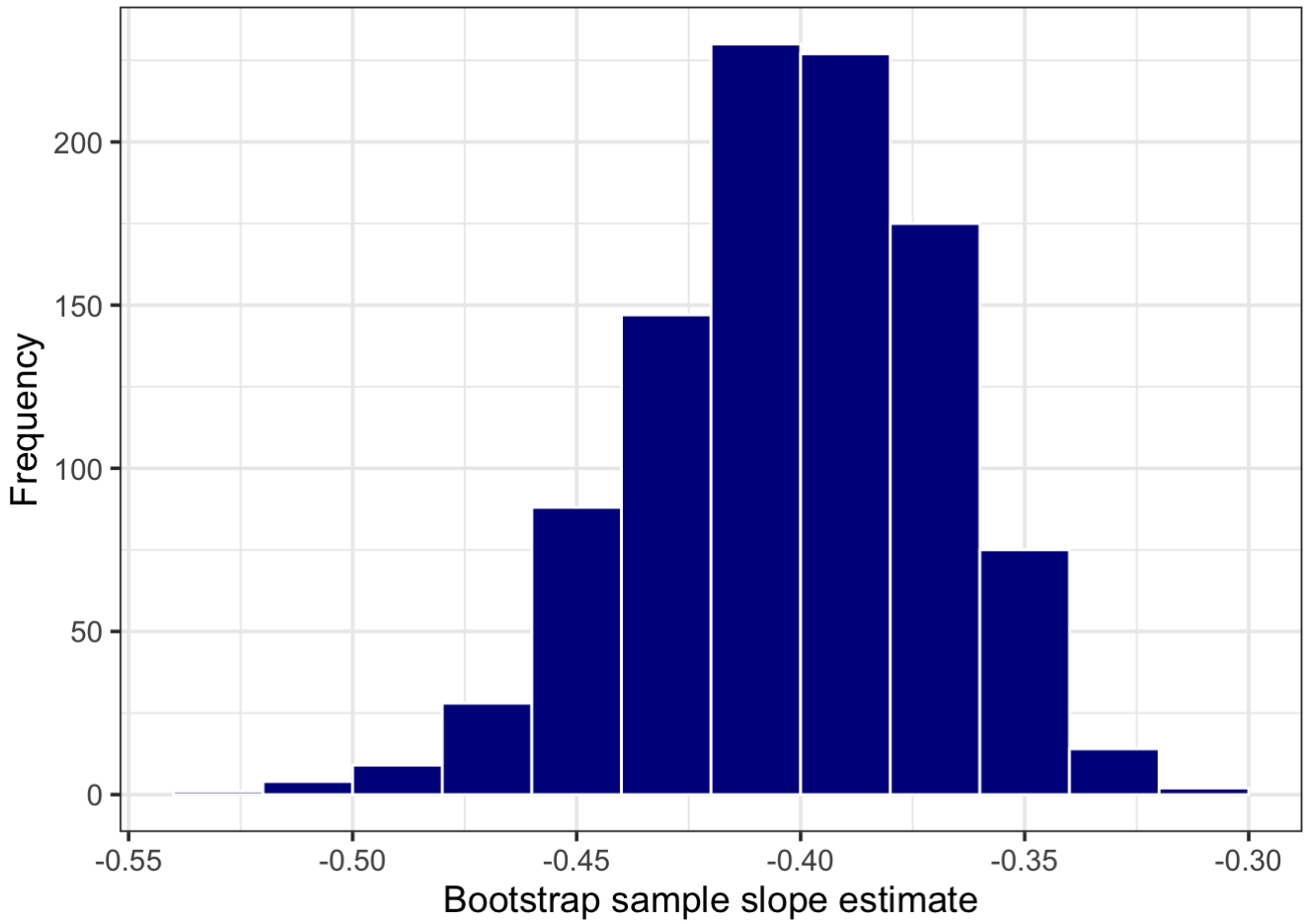
The code mean(bootstrap_education$stat) calculates the mean of the bootstrap sample slope estimates for the relationship between the obesity variable and the education variable in the merged dataset. The result of this calculation is -0.403. The code sd(bootstrap_education$stat) calculates the standard deviation of the bootstrap sample slope estimates for the same relationship. The result of this calculation is 0.033. These results suggest that there is a negative relationship between education and obesity in the merged dataset, as the mean slope estimate is negative. Additionally, the standard deviation of the slope estimates is relatively small, suggesting that the estimates are relatively consistent across the bootstrap samples.The theoretical confidence interval for the slope estimate using the education variable is [-0.474, -0.338]. The mean of the bootstrap sample slope estimates (-0.4030399) is within this confidence interval, suggesting that there is strong evidence for a negative relationship between education and obesity. Additionally, the relatively small standard deviation of the slope estimates supports the notion that the estimates are consistent and stable.

The ci_pctile_education variable uses the bootstrap method to estimate the confidence interval, while the ci_theory_education variable uses the theoretical method. The results of ci_theory_education show that the theoretical confidence interval for the slope of the linear regression model between obesity and education is [-0.457, -0.343]. This means that if we were to estimate the slope using the entire population instead of a sample, we would expect it to be within this range with 95% confidence. The ci_pctile_education variable shows the 95% confidence interval for the slope of the linear regression model between obesity and education based on the bootstrap method. The confidence interval is estimated to be [-0.405, -0.387]. This means that if we repeatedly sampled from the population and estimated the slope using the bootstrap method, 95% of the resulting confidence intervals would contain the true population slope. Comparing the two confidence intervals, we can see that they overlap but the bootstrap interval is narrower than the theoretical interval. This suggests that the bootstrap method provides a more precise estimate of the confidence interval than the theoretical method. It is also worth noting that the estimated slope of the linear regression model for education and obesity based on the bootstrap method is -0.403, with a standard deviation of 0.033. This indicates that there is a strong negative relationship between education and obesity, and this relationship is statistically significant. In conclusion,Comparing the results of the bootstrap method for education and poverty, we can see that the mean slope estimate for poverty is 0.739, which is positive and indicates a positive relationship between poverty and obesity. This is in contrast to the negative relationship between education and obesity. The standard deviation of the slope estimate for poverty is also higher (0.086) than that of education (0.033), indicating more variability in the estimates.

This code generates a histogram of the distribution of bootstrap sample slope estimates for the relationship between the ‘education’ variable and the ‘obesity’ response variable. It is created using `ggplot`, with the x-axis representing the slope estimates and the y-axis representing the frequency of those estimates. The dashed line represents the 95% confidence interval computed using the percentile method based on the bootstrap samples, while the solid line represents the 95% confidence interval computed using the theoretical method based on the assumption of a linear regression model. The plot allows us to visually compare the two methods of computing the confidence interval and assess their agreement. We can see that the confidence intervals based on both methods overlap, suggesting that there is no significant difference between the two methods for this data set.

```
bootstrap_education <- merged %>%
  specify(response = obesity, explanatory = education) %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "slope")

bootstrap_education %>%
  ggplot(aes(x = stat)) +
  geom_histogram(boundary = 9, binwidth = 0.02, col = "white", fill = "darkblue") +
  labs(
    x = "Bootstrap sample slope estimate",
    y = "Frequency"
  ) +
  theme_bw(base_size = 14)
```




```
mean(bootstrap_education$stat)
```

```
## [1] -0.4018058
```

```
sd(bootstrap_education$stat)
```

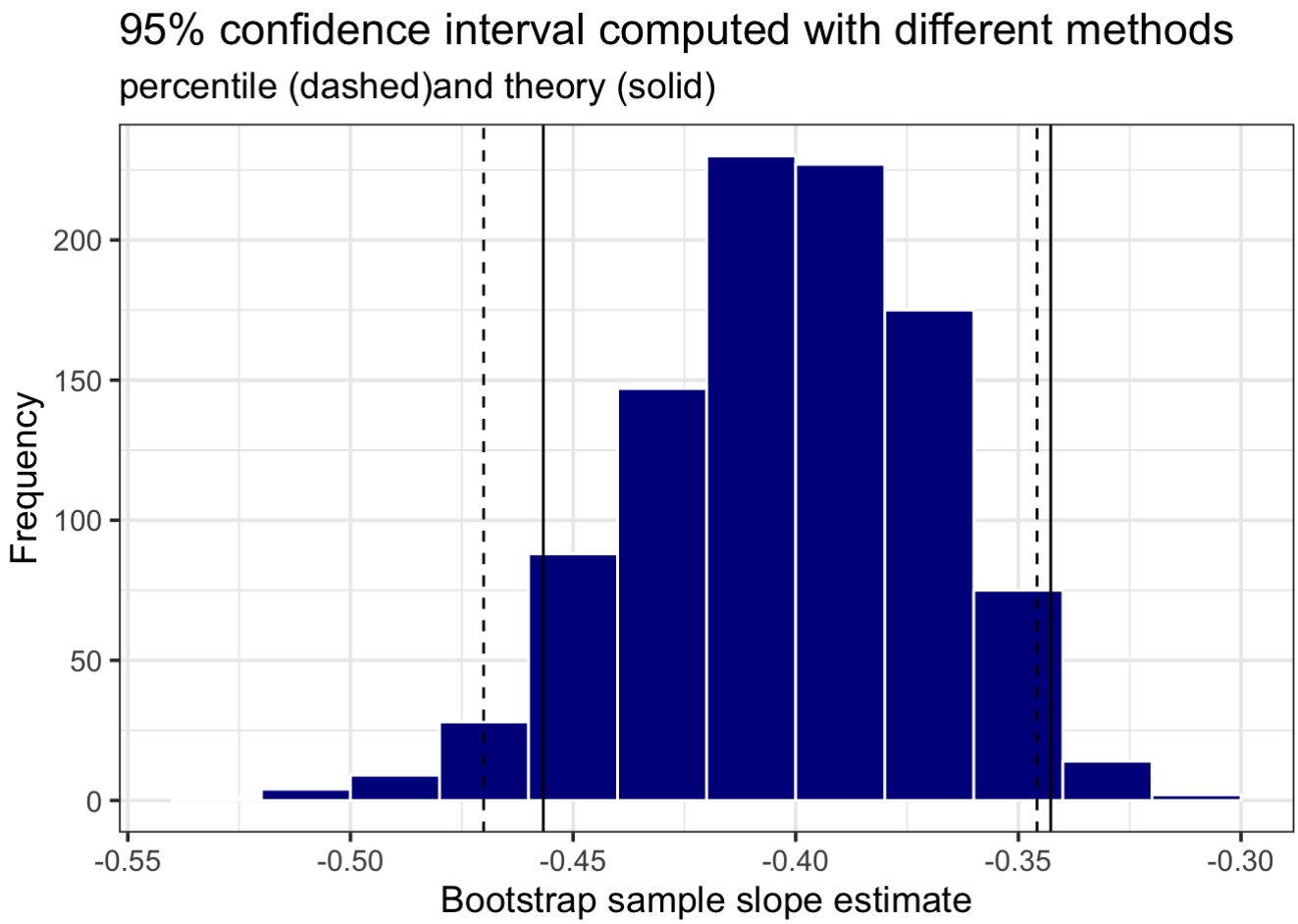
```
## [1] 0.03262841
```

```
ci_pctile_education = bootstrap_education %>%
  summarise(
    lower = quantile(stat, 0.025),
    upper = quantile(stat, 0.975)
  )

ci_theory_education <- tidy(lm(obesity ~ education, merged),
                            conf.int = TRUE, conf.level = 0.95) %>%
  filter(term == "education") %>%
  select(term, conf.low, conf.high)
ci_theory_education
```

```
## # A tibble: 1 × 3
##   term      conf.low conf.high
##   <chr>        <dbl>     <dbl>
## 1 education   -0.457    -0.343
```

```
bootstrap_education %>%
  ggplot(aes(x = stat)) +
  geom_histogram(boundary = 1, binwidth = 0.02, col = "white", fill = "darkblue") +
  labs(
    x = "Bootstrap sample slope estimate",
    y = "Frequency",
    title = "95% confidence interval computed with different methods",
    subtitle = "percentile (dashed)and theory (solid)"
  ) +
  geom_vline(xintercept = c(ci_pctile_education$lower, ci_pctile_education$upper), linet
ype = "dashed", show.legend = TRUE) +
  geom_vline(xintercept = c(ci_theory_education$conf.low, ci_theory_education$conf.hig
h)) +
  theme_bw(base_size = 14)
```



Causality

One of the factors that we are interested in exploring in our analysis is the minimum wage, as it is a potential policy lever that could affect the outcomes we are studying. Specifically, we want to investigate whether changes in minimum wage could be linked to changes in obesity rates. To gain some preliminary insights, we looked at changes in minimum wage and obesity rates in all states over the five-year period. We observed that, for example, in Arizona, the minimum wage increased by $1.5 from 2018 to 2019, which is a substantial change compared to other states, and during the same period, the obesity rate increased by 2%. On the other hand, in Connecticut, there was no change in minimum wage from 2017 to 2018, but the obesity rate still increased by 0.5%. A three-year insight into the relationship between minimum wage and obesity rate in West Virginia confirms this: The minimum wage was constant at $8.75 from 2017 to 2019, yet the obesity rate increased from 38.1% to 39.7%. These findings suggest that minimum wage increases may not be strongly linked to changes in obesity rates. However, we also found that our preliminary notion that a state's poverty rate influences the rate of obesity stands. For instance, West Virginia has one of the lowest minimum wages ($8.75 in 2017) in the United States and also one of the highest obesity rates (38.1% in 2017). In contrast, states with higher minimum wages, such as California ($10 in 2017), tend to have lower obesity rates (25.1% in 2017). While this correlation suggests a potential link between minimum wage and obesity rates, we must be cautious in drawing causal conclusions from this evidence alone. There may be other confounding factors at play, and further analysis is necessary to establish a causal relationship.

A possible reverse causality in the relationship between minimum wage and obesity rates could be that high obesity rates are driving down wages in certain states. Individuals who are already obese may have limited job opportunities or face discrimination in the job market, leading them to accept lower paying jobs. This could result in a higher concentration of low-wage workers in states with high obesity rates. Similarly, individuals with obesity

may be more likely to have health problems that limit their ability to work, making it more difficult for them to earn higher wages. For example, a study by Lee at al. (2019) found that obese individuals were more likely to have chronic health conditions that interfered with their ability to work, leading to lower wages and greater risk of unemployment. If a state has a large population of obese individuals, it could lead to increased healthcare costs and decreased productivity, which may in turn lead to lower wages. This could also lead to a higher prevalence of low-paying jobs, which could contribute to the high obesity rates observed in those states. These factors could contribute to a reverse causality scenario. In other words, rather than changes in minimum wage causing changes in obesity rates, it could be the other way around: high obesity rates could be leading to lower wages, which could then make it more difficult for individuals to afford healthier food options or engage in physical activity, ultimately perpetuating the cycle of obesity.

Limitations

There are several limitations to our analysis that should be considered. First, it is possible that other variables, such as cultural and environmental factors, may play a role in the relationship between poverty rates and obesity rates. Regarding the relationship between educational attainment and obesity, our analysis did not include other factors that may be associated with education, such as income or occupation. Additionally, the data used in this study was collected at the state level, which may not not accurately reflect the experiences of individuals within those states. Moreover, the models used in this study did not take into account all possible predictors of obesity, such as physical activity levels, access to healthy food options, and genetic factors. Furthermore, our models did not include other potential confounding variables that could impact the results such as age, race/ethnicity, and sex. Finally, the sample size of this study was relatively small, which may limit the applicability of the findings to the larger population. Further more the difference in sample size between the various states means of course that the reliability of the results varies from state to state.

Conclusion

The project analyzed the relationship between poverty rates and obesity rates in the United States. Based on the data analysis using multiple regression models, poverty rates were found to be a significant predictor of obesity rates in the United States. This suggests that poverty is a risk factor for obesity, with individuals living in poverty being more likely to develop obesity than those who are not. The relationship between poverty and obesity may be due to various factors such as limited access to healthy food options, lack of safe and convenient places for physical activity, and increased stress levels. While education levels were not found to be significant predictors in the models tested, it is important to note that education may still play a role in the development of obesity. This may be due to the fact that education is often associated with higher income levels and access to resources that promote healthy behaviors. However, the models tested in this analysis did not include other factors that may be associated with education. Therefore, more complex models that take into account a broader range of factors may be needed to fully understand the role of education in the development of obesity.

Overall, this project highlights the importance of addressing poverty as a risk factor for obesity and the need for interventions that promote healthy behaviors and access to resources in low-income communities.

Bibliography

1. “Changes in Basic Minimum Wages in Non-Farm Employment under State Law: Selected Years 1968 to 2019 | U.S. Department of Labor.” 2019. Dol.gov. 2019. https://www.dol.gov/agencies/whd/state/minimum-wage/history (https://www.dol.gov/agencies/whd/state/minimum-wage/history).

2. Devaux, Marion, Franco Sassi, Michele Cecchini, Francesca Borgonovi, and Jody Church.

2011. “Exploring the Relationship between Education and Obesity.” OECD Journal: Economic Studies 2011 (1): 1–40. https://doi.org/10.1787/eco_studies-2011-5kg5825v1k23 (https://doi.org/10.1787/eco_studies-2011-5kg5825v1k23).

3. Dinsa, G. D., Y. Goryakin, E. Fumagalli, and M. Suhrcke.

2012. “Obesity and Socioeconomic Status in Developing Countries: A Systematic Review.” Obesity Reviews 13 (11): 1067–79. https://doi.org/10.1111/j.1467-789x.2012.01017.x (https://doi.org/10.1111/j.1467-789x.2012.01017.x).

4. Lee, Hyeain, Rosemary Ahn, Tae Hyun Kim, and Euna Han. 2019. “Impact of Obesity on Employment and Wages among Young Adults: Observational Study with Panel Data.” International Journal of Environmental Research and Public Health 16 (1): 139. https://doi.org/10.3390/ijerph16010139 (https://doi.org/10.3390/ijerph16010139).

5. Levine, James A.

2011. “Poverty and Obesity in the U.S.” Diabetes 60 (11): 2667–68. https://doi.org/10.2337/db11-1118 (https://doi.org/10.2337/db11-1118).

6. National Center for Health Statistics.

2016. “Healthy People 2020 Mid-Course Review.” https://www.cdc.gov/nchs/data/hpdata2020/HP2020MCR-C29-NWS.pdf (https://www.cdc.gov/nchs/data/hpdata2020/HP2020MCR-C29-NWS.pdf).

7. Ogden, Cynthia L., Tala H. Fakhouri, Margaret D. Carroll, Craig M. Hales, Cheryl D. Fryar, Xianfen Li, and David S. Freedman. 2017. “Prevalence of Obesity among Adults, by Household Income and Education — United States, 2011–2014.” MMWR. Morbidity and Mortality Weekly Report 66 (50): 1369–73. https://doi.org/10.15585/mmwr.mm6650a1 (https://doi.org/10.15585/mmwr.mm6650a1).

8. OECD Journal: Economic Studies. https://www.oecd-ilibrary.org/economics/oecd-journal-economic-studies_19952856 (https://www.oecd-ilibrary.org/economics/oecd-journal-economic-studies_19952856). Consultato 1 maggio 2023.