Analysis of County-Level Obesity Rates in the United States

AUTHOR
Liz Rightmire

AFFILIATION
Middlebury College

Introduction

The WHO has identified obesity in America as an "escalating global epidemic." As of 2023, it affects one in three adults and one in six children in America. Obesity does not touch all Americans equally: non-Hispanic Black adults have the highest prevalence of obesity at 49.9%, and the rate is only slightly lower for Hispanic adults–45.6%. This is much higher than the rates for White (41.4%), and Asian adults (16.1%). Inequalities persist when looking at socioeconomic status: adults with college degrees have lower obesity prevalence compared to those with less education (Tiwari, 2024). Unfortunately, rates of obesity in the United States are only growing. In order to create successful mitigation strategies for this epidemic, it is important to understand the key factors at play.

Using publicly available county-level public health data from The CDC and the FDA, I aim to create a linear model to predict a county's age-adjusted obesity rate from other characteristics.

I will use data from two sources. First, the CDC PLACES dataset provides chronic disease and health-related data for 3111 U.S. counties. These data are derived from Behavioral Risk Factor Surveillance System data, Census population data, and American Community Survey data. From this dataset, I obtained the response variable for my model–obesity rate. This is an age-adjusted value, which is a statistical measurement that compares health outcomes in groups of people with different age structures. It's used to control for the effects of age differences on health event rates, and to remove confounding caused by age (CDC).

Most explanatory variables were drawn from the USDA's Food Atlas, which presents an overview of food access indicators for U.S. counties. These data were compiled from census population data and data from other government agencies. The variables I chose come from 4 different pages in the .xslx Food Atlas download: "Access," "Health," "Socioeconomic," and "Rural." These data can be joined together using a county's unique FIPS code. From the "Access" file came LowStoreAccess, which is the share of the population in the tract with limited access to a food store. This is defined as living more than 1 mile from a food store in urban areas or more than 10 miles in rural areas. From "Health" came the number of recreation facilities per 1,000 people, and from "Socioeconomic" came the county's median income. Finally, a Rural indicator variable was sourced from the CDC's Urban-Rural Classification Scheme. While the original data described rural-ness and urban-ness using a 1-9 scale, I made this variable binary by encoding counties with a score greater than 4 as urban, following the dataset's documentation.

Data

▼ Code

```
# Packages
library(tidyverse)
library(kableExtra)
library(broom)
library(GGally)
library(car)
```

▼ Code

```
# Read in the data
places <- read_csv("CDC_Places.csv")|>
  filter(MeasureId == "DIABETES", Data_Value_Type == "Age-adjusted prevalence")
|>
  select("LocationID", "Data_Value") |>
  rename("FIPS" = "LocationID",
         "Obesity_Rate" = "Data_Value")

access <- read_csv("access.csv") |>
  select("FIPS", "LACCESS_POP15") |>
  rename("LowStoreAccess" = "LACCESS_POP15")

health <- read_csv("health.csv") |>
  select("FIPS", "RECFACPTH16") |>
  rename("RecFacilities" = "RECFACPTH16")

socioeconomic <- read_csv("socioeconomic.csv") |>
  select("FIPS", "MEDHHINC15") |>
  rename("MedianIncome" = "MEDHHINC15")
```

▼ Code

```
# Create Rural Indicator Variable
RuralUrban <- read_csv("RuralUrban.csv") |>
  mutate("Rural" = case_when(RUCC_2023 <= 4 ~ 1,
                             RUCC_2023 > 4 ~ 0)) |>
  select("FIPS", "Rural")
```

▼ Code

```
# join the data
data <- places |>
  inner_join(access, by = "FIPS") |>
  inner_join(health, by = "FIPS") |>
  inner_join(socioeconomic, by = "FIPS") |>
  inner_join(RuralUrban, by = "FIPS")
```

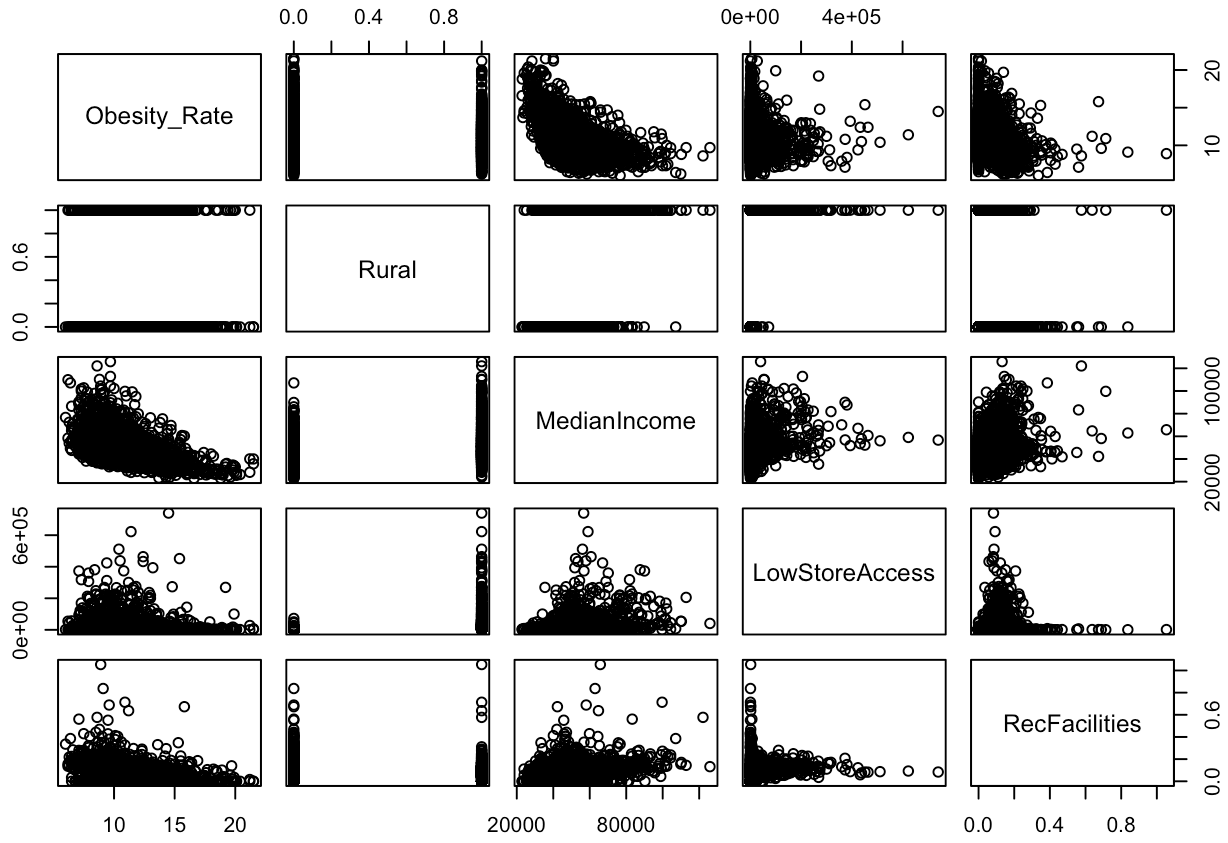
Methods

Feature Selection

To predict obesity rate from rural status, median income, store access, and recreation facility access, I will use a multiple linear regression model because the response variable, and all but one explanatory variable, are numerical. This model will allow me to determine the influence of each explanatory variable on obesity rate.

▼ Code

```
        pairs(Obesity_Rate ~ Rural * MedianIncome + LowStoreAccess + RecFacilities,
data = data)
```



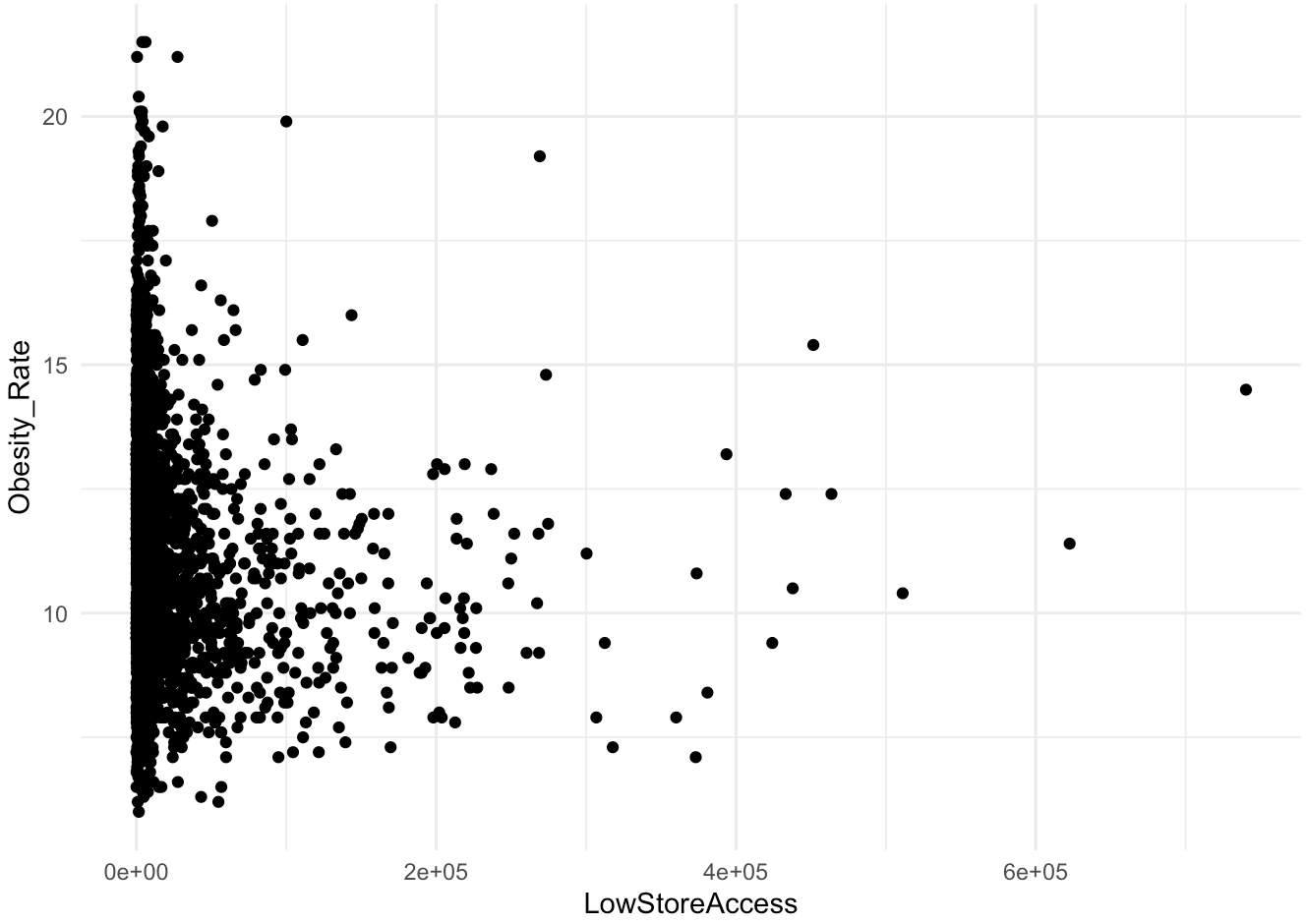
Graph 1: Checking Linearity between Response Variable and Each Predictor

Before fitting a model, I look to examine if the relationship between obesity rate and each predictor variable is linear. Two of the variables appear to have a linear relationship with obesity rate: rec facilities and median income. Unfortunately, there does not appear to be a linear relationship between store access and obesity rate. Let's look a closer look with a larger plotting window.

▼ Code

```
        #| label: graph-2
        #| fig-cap: "Figure 2: Non-Linear relationship between Store Access and Obesity
Rate"
```

```
data |>
  ggplot(aes(x = LowStoreAccess, y = Obesity_Rate)) +
  geom_point() +
  theme_minimal()
```



Even after applying a log or exponential transformation, the non-linearity is not solved. Therefore, the LowStoreAccess variable will be removed from the model.

With this adjustment, there does not appear to be collinearity between any of the predictor variables because the correlation values for each pair of predictor variables are very low in magnitude.

▼ Code

```
ggpairs(data = data %>% select(-FIPS, -LowStoreAccess))
```

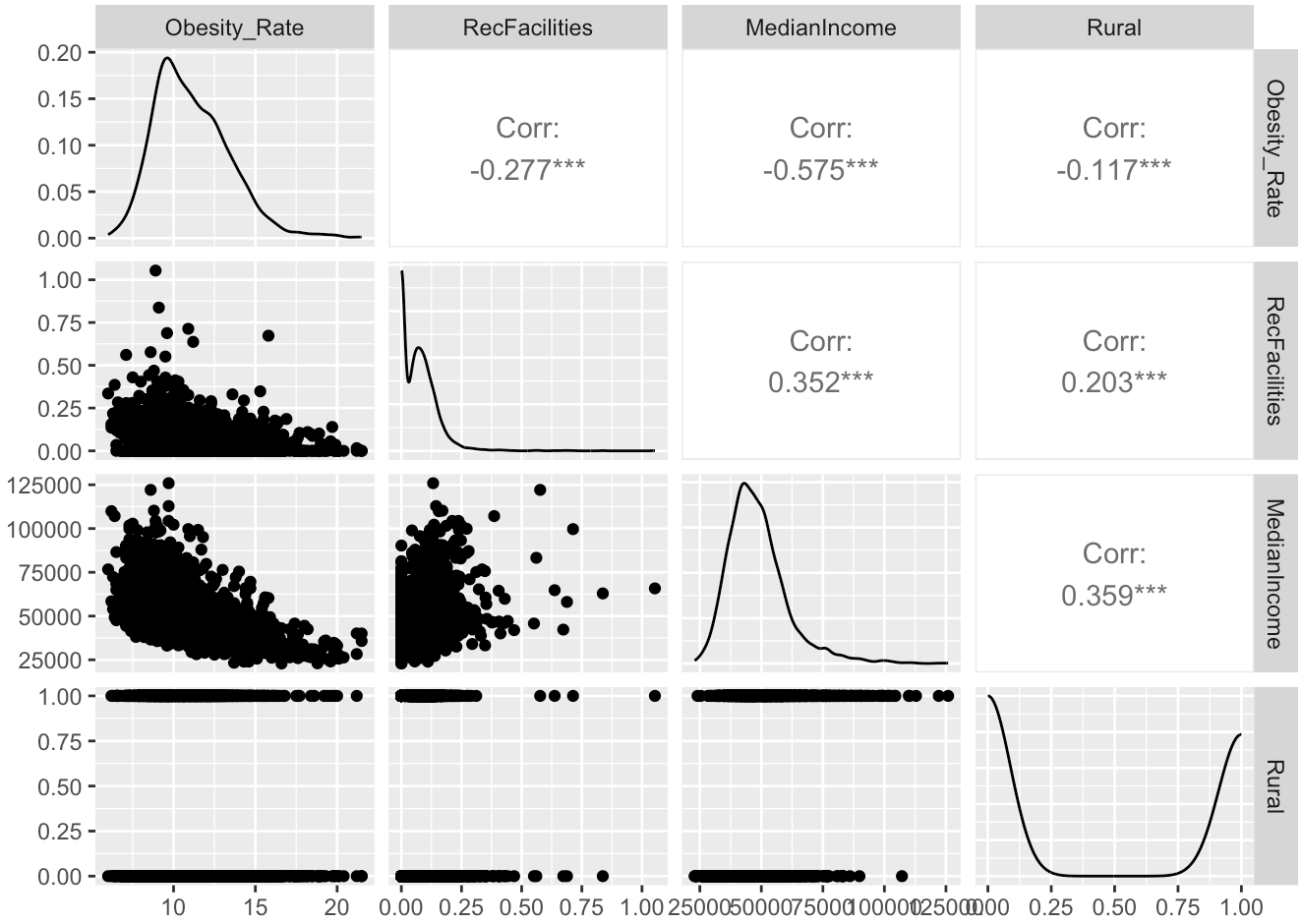


Figure 3: Checking for Collinearity between Predictor Variables

Regression Models

After assessing linearity, it is time to build some models. First, I will create a simple linear model with the 3 explanatory variables.

▼ Code

```
        simple_model <- lm(Obesity_Rate ~ Rural + MedianIncome +  RecFacilities, data =
data)
```

▼ Code

```
        summary(simple_model) %>%
           tidy() %>%
           mutate(p.value = ifelse(p.value < 0.0001, "<0.0001", round(p.value, 4))) %>%
           kbl(booktabs=TRUE, digits=2)%>%
           column_spec(1, monospace = TRUE)
```

term	estimate	std.error	statistic	p.value
(Intercept)	16.39	0.14	119.26	<0.0001

term	estimate	std.error	statistic	p.value
Rural	0.51	0.07	7.13	<0.0001
MedianIncome	0.00	0.00	-35.75	<0.0001
RecFacilities	-2.88	0.47	-6.14	<0.0001

Table 1: Simple Linear Model

▼ Code

```
        # AIC
        extractAIC(simple_model)
```

```
[1]    4.000 3883.391
```

▼ Code

```
        # Adjusted R^2
        round(summary(simple_model)$adj.r.squared,2)
```

```
[1] 0.35
```

The Adjusted-R^2 is 0.35, the AIC is 3883.39, and every term is significant with p values < 0.0001.

Next, I experiment with an interaction model because I believe there may be a different impact of income on the obesity rate depending on if a county is rural or urban. This is because more income is required in an urban population to maintain a similar quality of life to a rural environment.

▼ Code

```
        interaction_model <- lm(Obesity_Rate ~ Rural * MedianIncome + RecFacilities,
data = data |>na.omit())
```

▼ Code

```
        summary(interaction_model) %>%
           tidy() %>%
           mutate(p.value = ifelse(p.value < 0.0001, "<0.0001", round(p.value, 4))) %>%
           kbl(booktabs=TRUE, digits=2) %>%
           column_spec(1, monospace = TRUE)
```

term	estimate	std.error	statistic	p.value
(Intercept)	18.33	0.21	86.43	<0.0001
Rural	-2.71	0.29	-9.31	<0.0001
MedianIncome	0.00	0.00	-32.13	<0.0001
RecFacilities	-3.03	0.46	-6.54	<0.0001
Rural:MedianIncome	0.00	0.00	11.44	<0.0001

Table 2: Interaction Model

▼ Code

```
        # AIC
        extractAIC(interaction_model)
```

```
[1]    5.000 3720.887
```

▼ Code

```
        # Adjusted R^2
        round(summary(interaction_model)$adj.r.squared,2)
```

```
[1] 0.38
```

The Adjusted-R^2 for the interaction model is slightly larger at 0.38, and the the AIC is smaller at 3720.89. Once again, every term is significant with p values < 0.0001 (even the interaction term). Therefore, the interaction model is prefered and I will continue my analysis with this model.

To ensure that there are not any excessive variables in the interaction model, I perform a backward selection procedure.

▼ Code

```
        interaction_model |>
          step(direction = "backward")
```

```
Start:  AIC=3720.89
Obesity_Rate ~ Rural * MedianIncome + RecFacilities

                     Df Sum of Sq   RSS    AIC
<none>                             10255 3720.9
- RecFacilities       1    141.33 10396 3761.5
- Rural:MedianIncome  1    432.03 10687 3847.3


Call:
lm(formula = Obesity_Rate ~ Rural * MedianIncome + RecFacilities,
    data = na.omit(data))

Coefficients:
       (Intercept)              Rural       MedianIncome       RecFacilities
         1.833e+01         -2.712e+00         -1.519e-04         -3.026e+00
Rural:MedianIncome
         6.739e-05
```

The smallest AIC value is achieved when all variables are included.

Verifying Assumptions

Linearity

I previously assessed the linearity between obesity rate and each explanatory variable. After removing the non-linear predictors, this model meets the linearity assumption.

Independence

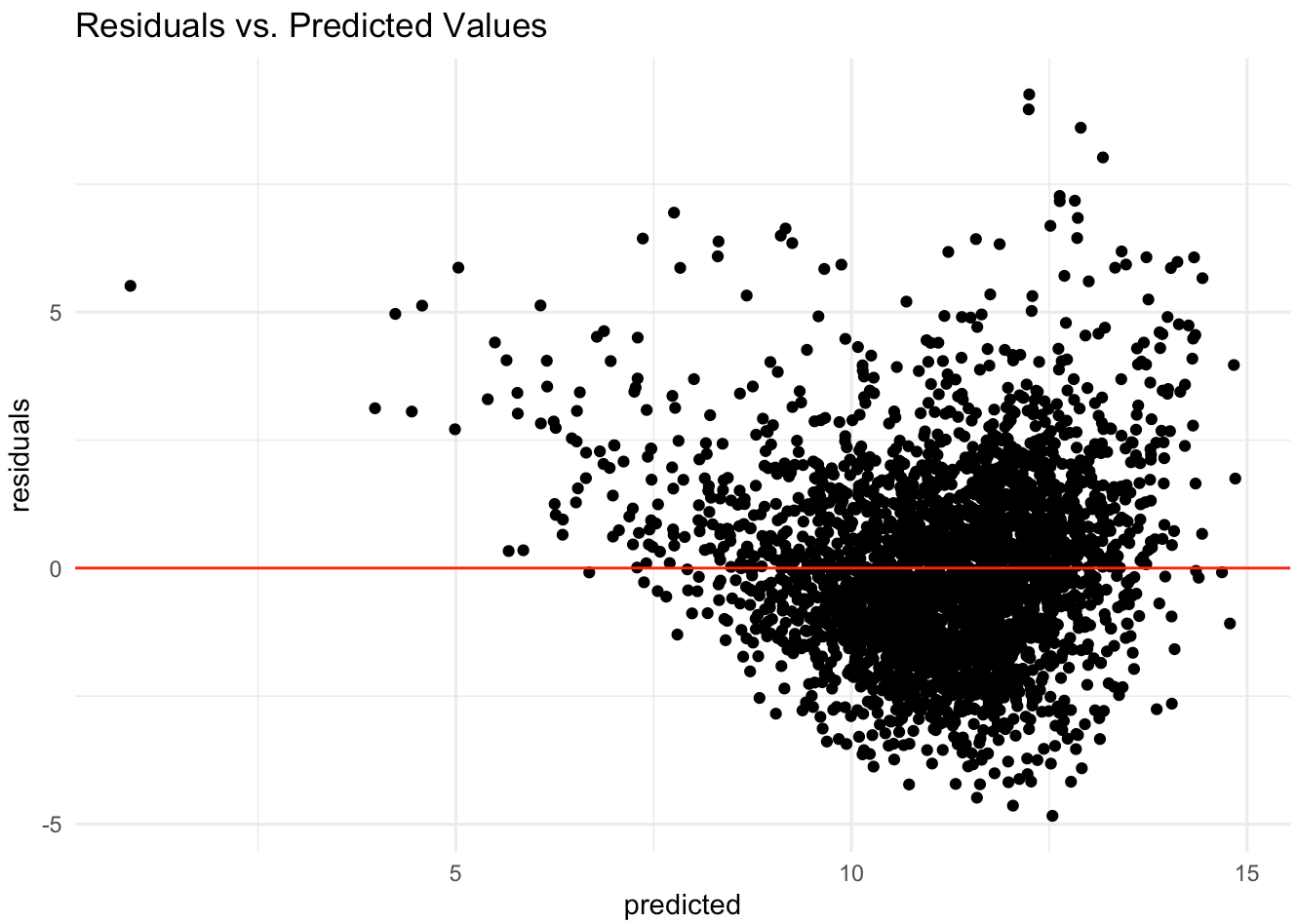
Each observation comes from a different US county. While geographically proximate counties will exhibit similar characteristics, county lines are strategically drawn to isolate communities. Therefore, the independence assumption is reasonably met

Constant Variance

▼ Code

```
# Residuals vs. Predicted Values
data <- data |>
  na.omit() |>
  mutate(predicted = predict.lm(interaction_model),
         residuals = resid(interaction_model))

ggplot(data = data, aes(x = predicted, y = residuals)) +
  geom_point() +
  geom_hline(yintercept = 0, color = "red") +
  labs(title = "Residuals vs. Predicted Values") +
  theme_minimal()
```



Graph 4: Relationship between Residuals and Predicted Values for Chosen Model

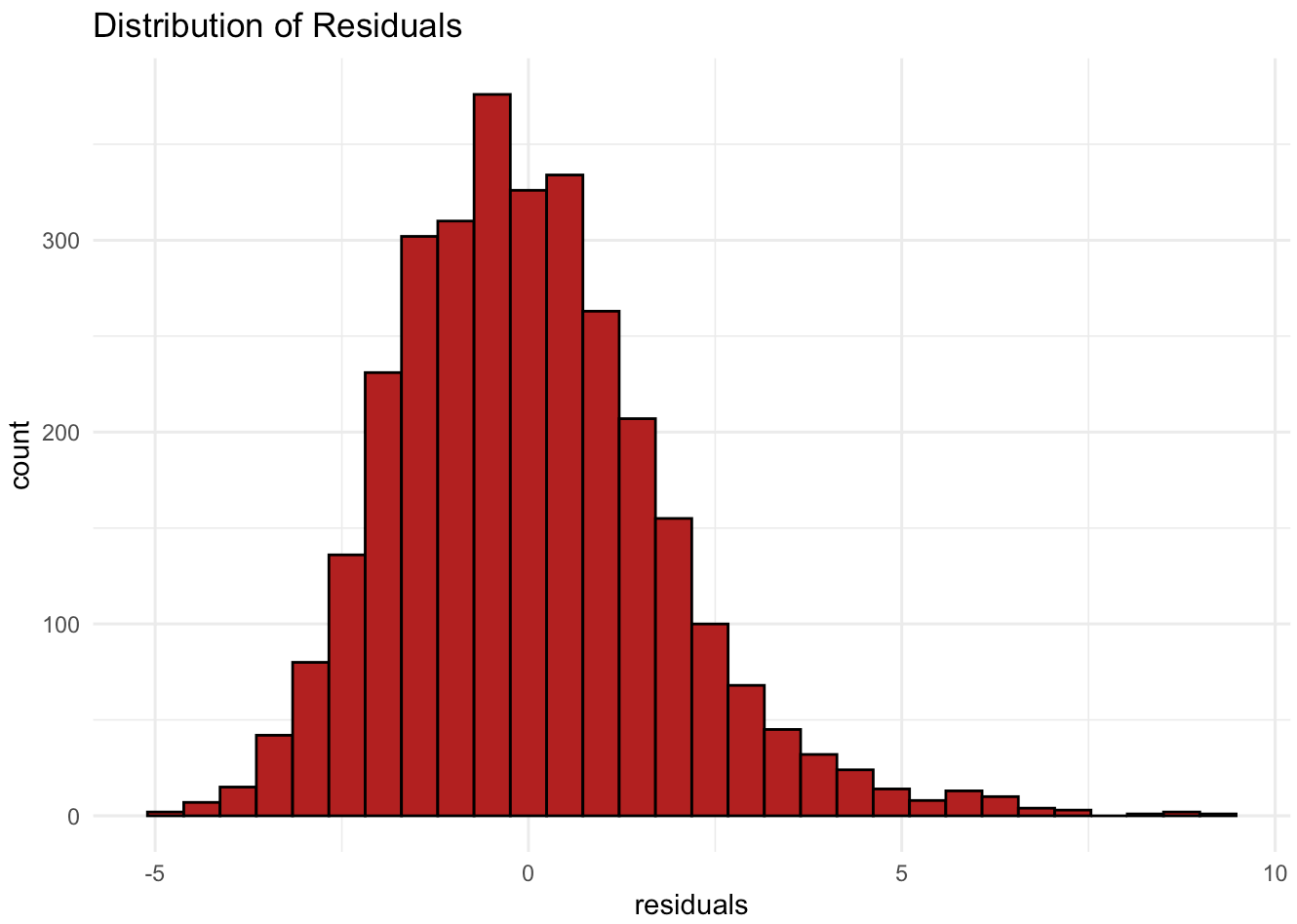
There are slightly more positive residuals than negative residuals. However, the residuals vs. predicted values plot generally appears to be a random cloud of points with very little discernible pattern, so I would argue that the constant variance assumption is reasonably met.

Normality

I assess normality using two methods: a histogram of the residual values, and a QQ-plot.

▼ Code

```
        ggplot(data=data,aes(x=residuals)) +
geom_histogram(fill="firebrick",color="black") +
          labs(title="Distribution of Residuals") +
          theme(plot.title=element_text(hjust=0.5,size=16)) +
          theme_minimal()
```

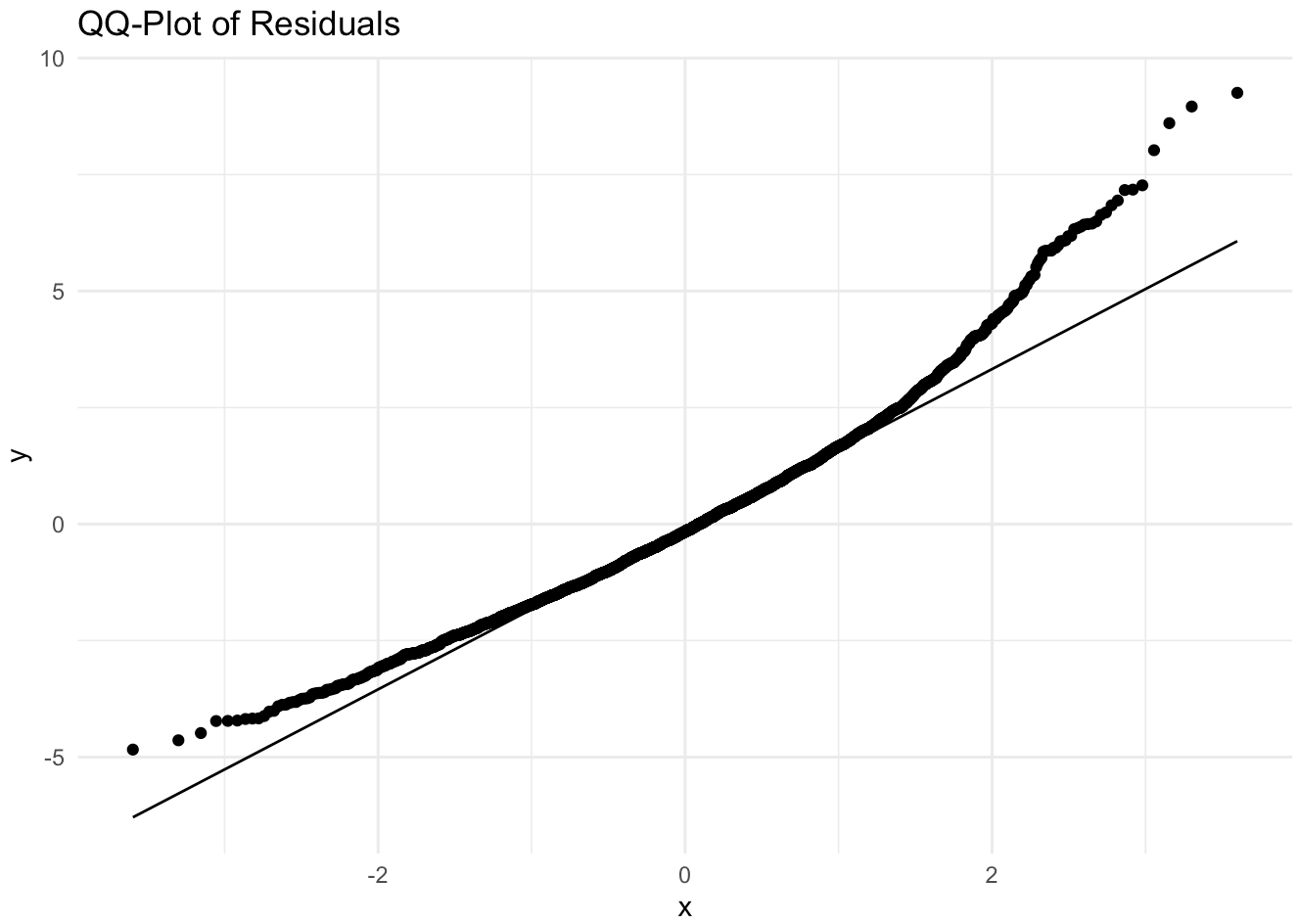


Graph 5: Histogram to Determine Normality of Model Residuals

The histogram shows a normal distribution centered at 0, with negligeable right skew.

▼ Code

```
ggplot(data=data,aes(sample=residuals)) + stat_qq() + stat_qq_line() +
  labs(title="QQ-Plot of Residuals") +
  theme(plot.title=element_text(hjust=0.5,size=16)) +
  theme_minimal()
```



Graph 6: QQ Plot to determine Normality of Model Residuals

The QQ plot shows that the residuals of the model mostly follow a normal distribution. Most of the points fit closely to the y=x line, but the values depart slightly at very small and large x values.

Because the concerns with the QQ plot are minor and the histogram presents normal behavior, the normality assumption is met.

Results

▼ Code

```
        summary(interaction_model) %>%
           tidy() %>%
            mutate(p.value = ifelse(p.value < 0.0001, "<0.0001", round(p.value, 4)))
%>%
           kbl(booktabs=TRUE, digits=5) %>%
           column_spec(1, monospace = TRUE)
```

term	estimate	std.error	statistic	p.value
(Intercept)	18.32945	0.21207	86.43034	<0.0001
Rural	-2.71221	0.29142	-9.30677	<0.0001

term	estimate	std.error	statistic	p.value
MedianIncome	-0.00015	0.00000	-32.13447	<0.0001
RecFacilities	-3.02576	0.46247	-6.54256	<0.0001
Rural:MedianIncome	0.00007	0.00001	11.43905	<0.0001

The equation of my final model is:

$$\hat{\text{Obesity Rate}} = 18.33 - 2.71(\text{Rural}) - 0.00015(\text{Median Income})$$
$$-3.03(\text{Rec Facilities}) + 0.00007(\text{Rural * Median Income})$$

The baseline obesity rate is 18.33, and differing conditions of the predictor variables adjust this rate.

To begin, a β_3 of -3.03 indicates that each additional recreation facility in a county reduces the obesity rate by 3.03% when controlling for the median income and rural status.

If a county is rural, the model predicts its obesity rate to be 2.71% smaller than an urban county with similar conditions.

I would have expected higher median incomes to result in lower obesity rates, and this is the case: each additional dollar in median income reduces that county's obesity rate by 0.00015%. While this seems like an insignificant number, just a $1,000 increase in median income decreases the obesity rate by 0.15%, and a $10,000 increase in median income decreases obesity rate by 1.5%.

The impact of median income on obesity rate does change depending on if the county is rural or not. In rural counties, median income has less of an impact on obesity rate: each additional dollar reduces obesity rate by 0.00008% as opposed to a reduction of 0.00015%.

Each variable had a p-value < 0.0001. Therefore, even variables with very minute beta values have a statistically significant impact on obesity rate.

Obesity rate in a rural county can be modeled by:

$$\hat{\text{Obesity Rate}}_{rural} = 15.62 - 0.00008(\text{Median Income}) - 3.03(\text{Rec Facilities})$$

And obesity rate in an urban county can be modeled by:

$$\hat{\text{Obesity Rate}}_{urban} = 18.33 - 0.00015(\text{Median Income}) - 3.03(\text{Rec Facilities})$$

Written in this way, it is clear that the baseline obesity rate for a rural county is lower than that of an urban county, but median income has less of a decremental effect on obesity rates in a rural counties.

Discussion

The goal of this analysis was to predict a county's age-adjusted obesity rate using factors from the PLACES and Food Atlas datasets. In creating this final model, I gained insight into the complexity of this national public health crisis.

One key finding is: after controlling for the number of recreation facilities, income has a different impact on obesity rate depending on if it is a rural or urban county. Counties with lower income have lower obesity rates if they are classified as rural, but the lowest obesity rates exist in high-income urban environments. There is a point at which income has a stronger impact than population density.

The number of rec facilities in a county has the greatest impact on obesity rate. The final model predicts that each additional recreation facility decreases the obesity rate by about 3%! This finding supports the creation of more fitness centers, especially free or subsidized programs like YMCAs, to lessen the impact of income on obesity rate.

The Adjusted-R^2 value for thie model was 0.38. This means that only 38% of the variance in obesity rate can be explained by median income, population density, and the number of recreation facilities.

As a result, this model serves an investigative role but cannot be used in the broader context of mitigating obesity because of the complexity of the issue. The factors used in this model were selected purely out of curiosity for their impact. This model was not controlled for enough factors for it to have meaningful predictive power.

As far as future work, I would like to do more to address the constant variance assumption. Additionally, I would be curious to look into existing studies of obesity in America to better inform the factors chosen for this model. There are many county-level descriptors in the PLACES and Food Atlas datasets that could be used to garner interesting insights into this public health crisis.

References

CDC, 2015. "PLACES: Local Data for Better Health." *Centers for Disease Control*. https://www.cdc.gov/places/about/index.html

USDA, 2019. "Food Access Research Atlas." *USDA Economic Research Service*. https://www.ers.usda.gov/data-products/food-access-research-atlas/

NCHS, 2013. "Urban_Rural Classification Scheme for Counties." *National Center for Health Statistics* https://www.cdc.gov/nchs/data-analysis-tools/urban-rural.html.

Tiwari A, Balasundaram P. Public Health Considerations Regarding Obesity. Jun 5, 2023. StatPearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK572122/.