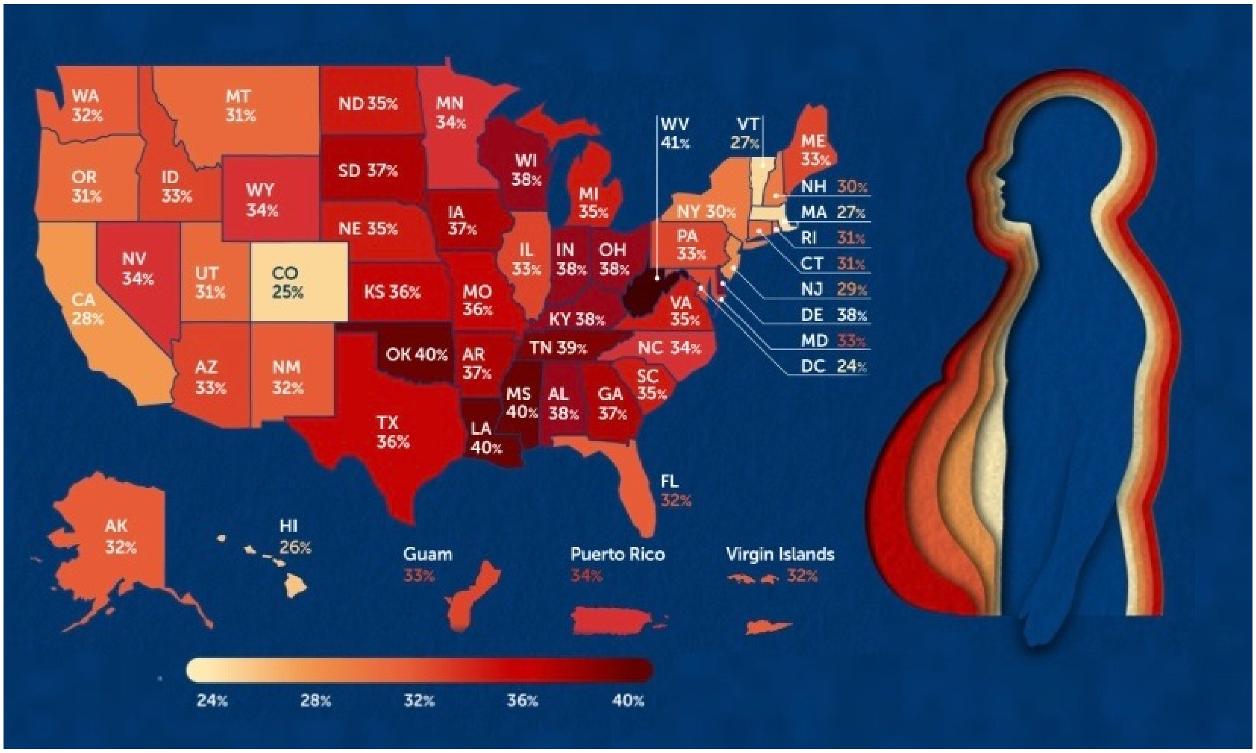
Identifying Predictors of Obesity in the US

coldspurs



Introduction

Over 40% of adults and 19% of children in the U.S. suffer from obesity, an epidemic with significant physiological, psychological, and economic impacts (NIH, 2024; CDC, 2022). This analysis explores key institutional factors influencing obesity in the U.S., examining obesity rates and related risk factors. The dataset includes data on the 50 states compiled from credible sources, including government databases and academic research. The 3 research questions are: 1) Do states with higher median incomes have lower obesity rates? A study conducted by the Harvard School of Public Health found that healthier food options were significantly more expensive than alternative food options. This question examines whether economic barriers contribute to obesity. 2) Do states with more fast food restaurants have higher obesity rates? The U.S. dominates the global fast food market, and this analysis investigates whether high densities of fast food restaurants in certain states are linked to higher obesity rates. 3) Do certain U.S. regions (Northeast, Midwest, West, South) have higher obesity rates? Factors such as food access, climate, and lifestyle may contribute to regional disparities in obesity, and this analysis aims to identify whether these differences are significant.

Methods and Analysis

The initial EDA indicated that the probability distribution of the response variable, obesity, is unimodal, approximately normal (with a slight skew to the left), and symmetric. Therefore, the response variable was deemed suitable for regression analysis. Next, 7 quantitative variables were tested for multicollinearity. A VIF value was calculated for each predictor, the largest of which was 2.796. We concluded that multicollinearity was not a concern for our data. All quantitative explanatory variables were then screened via stepwise screening, using a significance level of 0.1 for entry and removal into the model. The stepwise screening resulted in a removal of 3 insignificant quantitative variables (cell phone usage, fast food restaurant density, hospital density), leaving 3 quantitative variables (median age, median income, and smoking rate) remaining. We then entered the first stage of model building: testing the significance of the quantitative variables. From EDA, we determined there were no higher-order terms. Each remaining quantitative predictor was tested for significance through an individual t-test along with a global F-test for adequacy, and they were all determined to be significant. The second stage of model building involved the exploration of our 3 qualitative variables: insufficient sleep rate, region, and alcohol consumption levels. After testing for significance using nested F-tests, "Region" was the only qualitative variable remaining due to its significance. (For reference, the Midwest region is used as the base level for the Region variable). As two interaction plots indicated potential qualitative-qualitative interactions, they were eventually deemed insignificant upon testing. Moving into stage 3, we did not identify any potential quantitative-qualitative interactions from the EDA. However, based on intuition and our own background knowledge, we tested for interactions between region and median income as well as smoking rate and median age. Both were identified as insignificant. As a result, the final model contains the following four predictors: smoking rate, median income, median age, and region.

Upon establishing the final model, the three regression assumptions—lack of fit, homoscedasticity, and normality—were assessed using the Residual Plot, Residual vs Predicted Plot, and QQPlot, respectively. For independence, we assumed states were uncorrelated and observations independent. Each of the plots illustrated that all three assumptions were not violated, indicating that a transformation of the response variable was not necessary. Next, a residual analysis to identify influential observations and outliers was conducted. Based on Cook's Distance and hat values, observations 2, 6, 9, and 44 were identified as influential points. In terms of potential outliers, observations 6 and 8 had studentized residuals above the threshold of ±2, and observations 6, 8, and 45 had deleted studentized residuals above ±2. Following a holistic examination of all potential influential observations and outliers, observations 8 (Florida) and 44 (Utah) were removed from the dataset. Lastly, we implemented weighted least squares regression (WLSR) into our analysis. In the ordinary least squares regression, each state was treated in the same manner. However, in reality, there are considerable differences between the economies, environments, and demographics of states (e.g., New York and Alaska); therefore, they should not be treated with equal importance in our analysis. WSLR places greater influence on observations with lower variance and reduces the impact of outliers (that were not already removed in the previous step). The model was then refitted with the new data and method, resulting in our final prediction equation.

Results

Our final model is as follows:

$$\begin{gathered} E(obesity) = \beta_0 + \beta_1 SmokingRate + \beta_2 MedianAge + \beta_3 MedianIncome \\ + \beta_4 DumRegionNortheast + \beta_5 DumRegionSouth + \beta_6 DumRegionWest \end{gathered}$$

A global F-test yields an F-statistic of 44.077 (5, 44 DF) with a p-value of 5.752e-15, confirming model adequacy in predicting obesity rates. Additionally, the adjusted R-squared value of 0.82 indicates that about 82% of the variation in obesity rates is explained by this model. Lastly, we calculated a RSE of 1.295 percentage points. Taking all of these metrics into consideration, we conclude that the final prediction model is fairly accurate in predicting obesity rates across the 50 states.

Conclusions

The final prediction equation is as follows:

$$\begin{aligned}\widehat{obesityrate} = 44.077 + 0.624 SmokingRate - 0.287 MedianAge - 0.065 MedianIncome \\ -2.578 DumRegionNortheast + 0.043 DumRegionSouth - 3.715 DumRegionWest\end{aligned}$$

The model indicates a positive association between obesity rate and smoking rate, along with negative associations for the median age and median income predictors. As for the region variable, our interpretation of each dummy variable is in relation to the base level, the Midwest. Therefore, the model shows that the Northeast and West are associated with lower obesity rates, on average, than the Midwest. The South is associated with a slightly higher obesity rate than the Midwest on average. The global F-test and R-squared values indicate that this model is proficient in predicting obesity rates across the US.

To see this model being used in practice, we calculated a prediction interval for Wisconsin. Using Wisconsin's data, the model predicts an interval of (32.563, 38.184) with a point estimate of 35.374. The actual obesity rate for Wisconsin in 2022 was 37.7, which is included in the interval and results in a residual of 2.326 percentage points.

This model. however, has its limitations. First, our initial transformation of two quantitative variables into qualitative variables with three levels/categories may have impacted our findings. Secondly, the findings are not necessarily novel or groundbreaking. For example, the positive association between median income and obesity rate was already likely to be significant considering the Harvard study findings. Also, the positive association between smoking rate and obesity rate could be interpreted as rather intuitive. Future studies should explore the significant predictors further. Possible research questions include: 1) What factors drive regional differences in obesity rates? 2) Which predictors remain significant when analyzing data at the state or county level? 3) Are the same predictors significant on a continental or global scale?

Appendix A: Data Dictionary

Variable Name	Abbreviated Name	Description
Obesity	obesity	Response Variable. Indicates the percentage of obese individuals (Obesity defined ashaving a BMI of at least 30.0) per state in 2024. Write the description of the variable
Alcohol consumption	Alcohol	Qualitative variable. Refers to the alcohol consumption per capita from all beverages in the U.S. in 2022, by state (in gallons of ethanol). Thresholds are defined as follows: Low: Alcohol consumption below 2 liters per capita. Moderate: Alcohol consumption between 2 and 3 liters per capita. High: Alcohol consumption above 3 liters per capita.
Region in America	Region	Qualitative variable. Indicates region (as defined by US Census British), including Northeast, Midwest, South, and West.

Variable Name	Abbreviated Name	Description
Insufficient sleep reported by citizen	Sleep	Qualitative variable. Represents the percentage of adults reporting insufficient sleep per state in 2022. With a short sleep duration being defined as an average of getting less than 7 hours of sleep every 24 hours. It is indicative of sleep-related health behaviors. In the study, the variable was defined as qualitative with the threshold of: Low: Insufficient sleep rate below 33%. Moderate: Insufficient sleep rate between 33% and 37%. High: Insufficient sleep rate above 37%. Referred throughout the study as “insufficient sleep”
Median age	age	Quantitative variable. Represents the median age of residents per state in 2022, reflecting the age demographics. Referred throughout the study as “age”

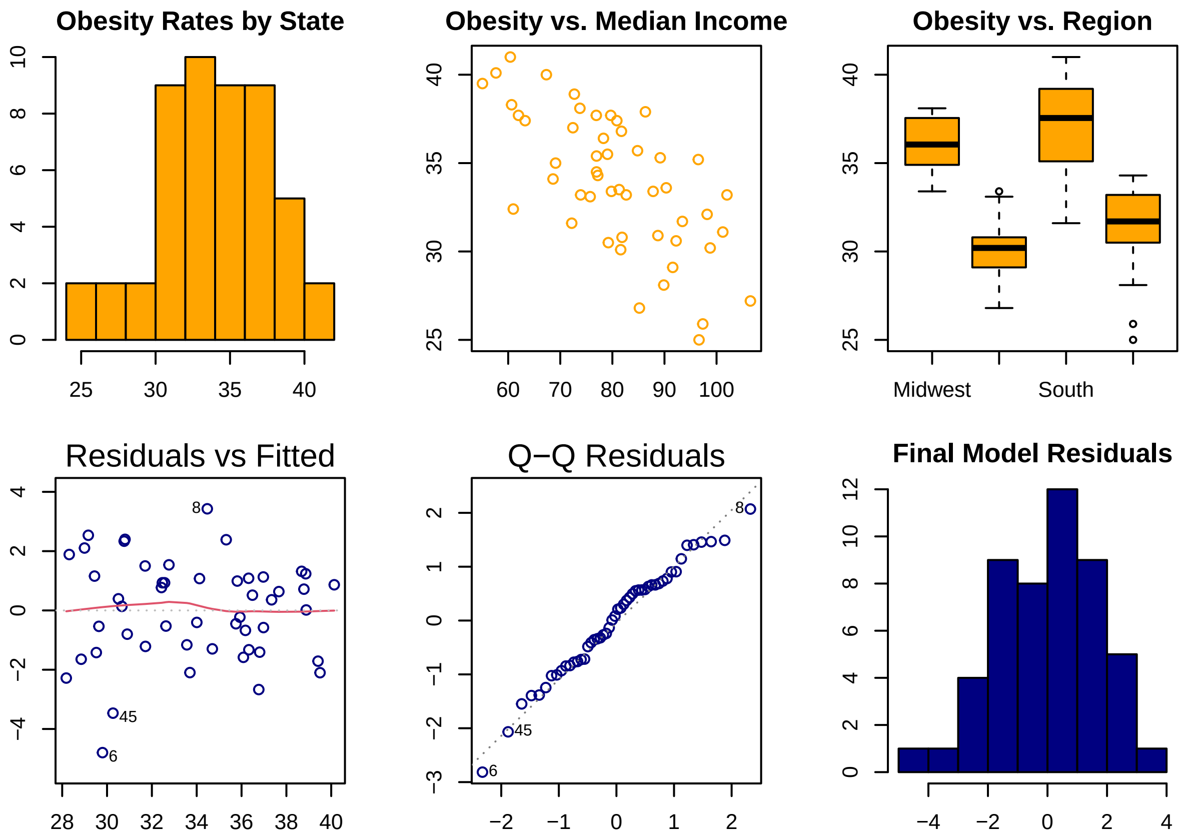
Variable Name	Abbreviated Name	Description
Percentage of adults who smoke	smoking	Quantitative variable; Indicates the percentage of adults who smoke in each state in 2022, providing insight into smoking prevalence. Referred throughout the study as “Smoking rate”
Median Household income	income	Quantitative variable; Indicates the average household income per state in 2023. Household income was defined to include all income received by each perso in the household who is aged 15 and older, excluding certain receipts such as capital gains. Money income is pretax and does not include stimulus payments and tax credits such as those from the American Rescue Plan Act (ARPA). Referred to throughout the study as “average income” or “average household income.”
Hospital number	hospital	Quantitative variable; Reflects the number of hospitals available per state. Referred throughout the study as “ Number of hospitals per state”

Variable Name	Abbreviated Name	Description
Average Monthly Search Volume Per 100k	cellphoneuse	Quantitative variable; Shows the average monthly search volume related to cellphone use per 100,000 people in 2024. Referred to throughout the study as “cellphone use” .
Fast food restaurant per 100,000 people	fastfood	Quantitative variable; Denotes the number of fast food restaurants per 100,000 people per state in 2024, serving as a proxy for fast food accessibility. Referred throughout the study as “fast food consumption”

Appendix B: Data Rows

```
           X obesity  Alcohol Region    Sleep  age smoking income hospitals
1    Alabama    38.3 Moderate  South     High 39.6      14  60.66        90
2     Alaska    32.1 Moderate   West     High 35.9      14  98.19        11
3    Arizona    33.2 Moderate   West Moderate 38.8      11  82.66        83
4   Arkansas    37.4      Low  South     High 38.9      17  63.25        52
5 California    28.1 Moderate   West Moderate 37.9       9  89.87       336
6   Colorado    25.0     High   West      Low 37.7      10  96.64        60
  cellphoneuse fastfood
1       343.89     81.7
2       242.08     61.9
3       352.50     67.9
4       318.54     69.9
5       380.75     82.3
6       369.63     75.7
```

Appendix C: Tables and Figures



```
Call:
lm(formula = obesity ~ smoking + age + income + Region, data = obesity_data_v2,
    weights = wt)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-2.6831 -0.8746  0.1367  0.9267  1.9731

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     44.07700    6.61674   6.661 4.96e-08 ***
smoking          0.62363    0.13311   4.685 3.07e-05 ***
age             -0.28726    0.15049  -1.909   0.0633 .
income          -0.06532    0.03074  -2.125   0.0397 *
```

```
RegionNortheast -2.57823    1.02900  -2.506   0.0163 *
RegionSouth      0.04348    0.63412   0.069   0.9457
RegionWest      -3.71475    0.72642  -5.114 7.77e-06 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 1.295 on 41 degrees of freedom
Multiple R-squared:  0.843, Adjusted R-squared:   0.82
F-statistic: 36.69 on 6 and 41 DF,  p-value: 5.752e-15
```

Appendix D: References

Background

Centers for Disease Control and Prevention. (2022, July 15). Consequences of obesity. Centers for Disease Control and Prevention. https://www.cdc.gov/obesity/basics/consequences.html

Centers for Disease Control and Prevention. (2024, September 12). New CDC data show adult obesity prevalence remains high. Centers for Disease Control and Prevention. https://www.cdc.gov/media/releases/2024/p0912-adult-obesity.html

Eating healthy vs. unhealthy diet costs about $1.50 more per day. News. (2014, January 13). https://www.hsph.harvard.edu/news/press-releases/healthy-vs-unhealthy-diet-costs-1-50-more/

Fast Food Consumption by Country 2024. Fast food consumption by country 2024. (n.d.). https://worldpopulationreview.com/country-rankings/fast-food-consumption-by-country

U.S. Department of Health and Human Services. (2024, March 8). Research in context: Obesity and metabolic health. National Institutes of Health. https://www.nih.gov/news-events/nih-research-matters/research-context-obesity-metabolic-health

Why obesity is a disease: Unpacking the controversy and causes. Obesity Medicine Association. (2023, December 30). https://obesitymedicine.org/blog/why-is-obesity-a-disease/

World Health Organization. (2024, March 1). Obesity and overweight. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

Data

Amerisleep. (2023, December 13). Cellphone use statistics by state. https://amerisleep.com/blog/cellphone-use-statistics-by-state/

America's Health Rankings. (n.d.). Explore insufficient sleep in the United States. https://www.americashealthrankings.org/explore/measures/sleep

American Hospital Directory. (n.d.). Hospital statistics by state. https://www.ahd.com/state_statistics.html

KFF. (2024, October 18). Adults who report smoking by sex. https://www.kff.org/other/state-indicator/smoking-adults-by-sex/

Lake County Illinois GIS. (2024, September 20). Lake County, Illinois - National obesity by state. https://catalog.data.gov/dataset/national-obesity-by-state-d765a

National Institute on Alcohol Abuse and Alcoholism. (n.d.). Surveillance reports. https://www.niaaa.nih.gov/publications/surveillance-reports

NiceRx. (n.d.). The fast food capitals of America. https://www.nicerx.com/fast-food-capitals/

Trust for America's Health. (2023). Obesity report. https://www.tfah.org/wp-content/uploads/2023/09/TFAH-2023-ObesityReport-FINAL.pdf

U.S. Census Bureau. (2024, August 30). Historical income tables: Households. https://www.census.gov/data/tables/time-series/demo/income-poverty/historical-income-households.html

U.S. Census Bureau. (n.d.). Census regions and divisions of the United States. https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf

U.S. Census Bureau. (n.d.). Explore census data. https://data.census.gov/table/ACSDT5Y2020.B01002

Supplemental Code and Analysis Help

Zach. (2020, December 31). How to Perform Weighted Least Squares Regression in R. Statology. https://www.statology.org/weighted-least-squares-in-r/