IDENTIFYING PREDICTORS OF OBESITY IN THE US

[redacted]

Introduction

- Over 40% of adults and 19% of children in the United States are obese. (NIH, 2024)
- Impacts range from physiological (diabetes, sleep apnea) to psychological (anxiety, depression) to economic (medical care costs). (CDC, 2022)
- This analysis examines the relationship between obesity rates (as a percentage of total state population) and potential risk factors in each of the 50 states.
  - Goal: identify the **most** significant predictors of obesity rates

Research questions

- 1. Do states with higher median incomes have lower obesity rates?
- 2. Do states with more fast food restaurants have higher obesity rates?
- 3. Do certain regions (as defined by the U.S. Census Bureau) have greater obesity rates?

Multicollinearity Check

- Potential issues with multicollinearity assessed using individual and average VIF values.
  - **Individual VIF values** for each quantitative variable is **below 10**
  - **Average VIF is below 3**

Multicollinearity is not a concern for this data.

Variable	VIF
income	2.669000
age	1.136000
smoking	2.796000
hospitals	1.285000
cellphoneuse	1.691000
fastfood	1.360000
Mean VIF	1.822833

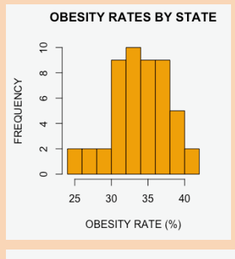
Variable Screening

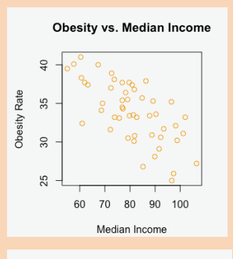
- **Stepwise regression** performed on seven quantitative variables resulting in the 3 quantitative variables remaining: **smoking rate, median age, and median income**.
  - p-ent = 0.1
  - p-rem = 0.1

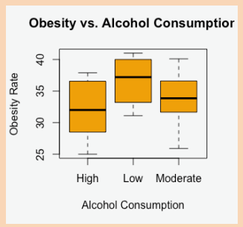
Stepwise Summary

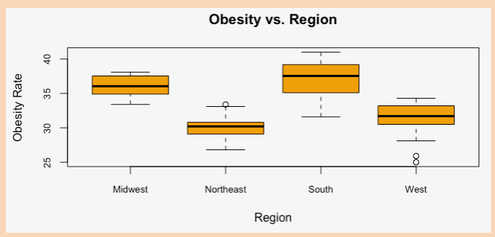
Step	Variable	AIC	SBC	SBIC	R2	Adj. R2
0	Base Model	274.746	278.529	NA	0.00000	0.00000
1	smoking (+)	235.093	240.769	NA	0.57260	0.56351
2	age (+)	232.271	239.838	NA	0.61267	0.59583
3	income (+)	229.912	239.371	NA	0.64563	0.62201

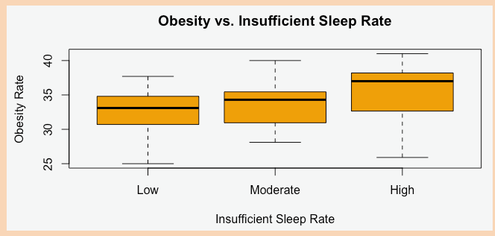
EDA











Data summary

Population of interest:
- American population
- Stratified by 50 states

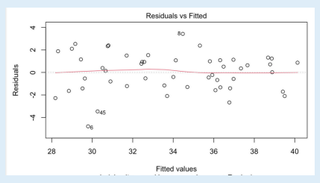
Data Sources:
- Data for 7 variables derived from government sources
- The other three variables contain data from credible organizations, databases, and academic research

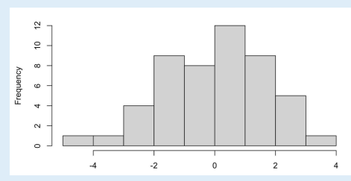
Data manipulations:
- Only one missing value in entire dataset -> dropped
- Rescaled income variable from dollars to thousands of dollars
- Grouped insufficient sleep rates into low, moderate, and high categories
- Grouped alcohol consumption data into low, moderate, and high consumption levels

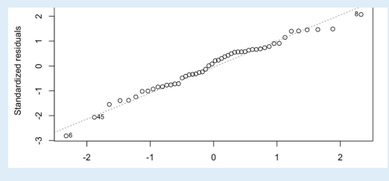
Variable Name	Abbreviated Name	Description
Obesity Percentage	obesity	Response Variable; The percent of the state population that is considered obese from the 2022 CDC BRFSS
Hospital Density	hospitals	Quantitative Variable; Number of hospitals per state
US Region	Region	Qualitative Variable; Region as defined by US Census Bureau
Cell Phone Usage	cellphone_use	Quantitative Variable; Average monthly search volume (containing phone-related search terms) per 100,000 residents.
Median Age	age	Quantitative Variable; Median age in the United States in 2022, by state. Data provided by the US Census Bureau.
Smoking Rate	Smoking	Quantitative Variable; Percentage of smokers , by state
Median Income	income	Quantitative Variable; Median income of state
Alcohol Consumption	Alcohol	Qualitative Variable; Alcohol Consumption, categorized into three levels
Insufficient Sleep Rate	Sleeping	Qualitative Variable; Percentage of adults who reported sleeping, on average, fewer than seven hours in a 24-hour period, categorized into three levels
Fast Food Restaurant Density	fast_food	Quantitative Variable; Fast food restaurants per 100,000 people

Residual Assumptions

- Lack of Fit:
  - No clear pattern showed from the residual plots
  - sum of the residuals are zero
  - **Mean zero assumption holds**
- Constant Variance:
  - In this residual vs. fitted plot, there are no "fanning" nor "funnel" shape.
  - The spread of the residuals seems consistent across fitted values
  - **Constant variance assumption appears to hold.**
- Normality:
  - Histogram of the residuals shows a largely normal distribution, with a symmetric bell-shaped curve.
  - Q-Q plot shows that most of the points lie on or near the diagonal line, indicating that the residuals are approximately normally distributed.
  - **Normality has not been violated.**

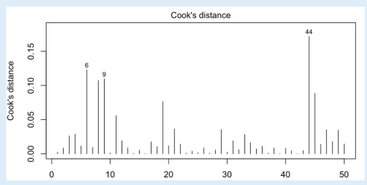


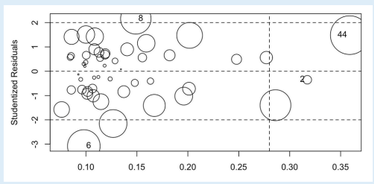


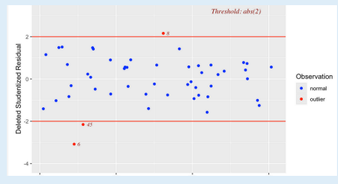


Residual Analysis

Removing Influential observations/outliers:
- Cook's distance (Influential points): 6, 9, and 44
- High-hat value(Influential points): 2,44
- High studentized residual(outliers): 6,8
- Deleted studentized residual (outliers): 4,8,45
- **Removed: 8 (Florida),44 (Utah)**







Analysis

Stage 1: Quantitative variables
- Initial model: E(Obesity) = β0 + β1(Smoking Rate) + β2(Median Age) + β3(Median Income)
- **Final model: E(Obesity) = β0 + β1(Smoking Rate) + β2(Median Age) + β3(Median Income)**

Stage 2: Qualitative variables

Initial model: E(Obesity) = β0 + β1(Smoking Rate) + β2(Median Age) + β3(Median Income)

Final model: E(Obesity) = β0 + β1(Smoking Rate) + β2(Median Age) + β3(Median Income) + β4(Dum_RegionNortheast) + β5(Dum_RegionSouth)+ β6(Dum_RegionWest)

**Dum_RegionNortheast = {0 if no, 1 if yes}. Dum_RegionSouth = {0 if no, 1 if yes}. Dum_RegionWest = {0 if no, 1 if yes}.*
***For this analysis, the Midwest region was used as the base level.*

Stage 3: Interactions
- Based on researchers' knowledge of topic, tests on the following interactions were performed: Region x Median Income and Smoking Rate x Median Age
- Both tests identified these interactions as insignificant. Therefore, the final model stayed the same:
  - **Final Model: E(Obesity) = β0 + β1(Smoking Rate) + β2(Median Age) + β3(Median Income) + β4(Dum_RegionNortheast) + β5(Dum_RegionSouth)+ β6(Dum_RegionWest)**

Weighted Least Squares Regression (Added Technique)

- Weighted Least Squares Regression (WLS) was implemented into this analysis for several reasons:
  - In ordinary least squares regression, each observation (state) is treated equally. However, all states are not equal, meaning they are not all equally reliable in our analysis
  - WLS places greater importance (weight) on observations with lower variance
  - In other words, observations with greater reliability now have greater influence on the regression model.
    - With WLS, the analysis now accounts for the relative significance of each state.
  - This method also reduces the impact of the outliers (even after the previous residual analysis and removal of influential points)

- *WLS Impact:*
  - Adj. R-Sq increases: 0.8071 -> 0.82
  - RSE decreases: 1.704 -> 1.295
  - F-statistic increases: 33.78 -> 36.69 | p-value decreases: 2.309e-14 -> 5.752e-15
  - **Overall, model becomes a better fit for our data after implementing WLS**

Conclusion

Final Prediction Equation (after WLS and removal of outliers/influential points):

$$\widehat{Obesity} = 44.077 + 0.62363(\text{Smoking\_Rate}) - 0.28726(\text{Age}) - 0.06532(\text{Income}) - 2.578(\text{Dum\_RegionNortheast}) + 0.04348(\text{Dum\_RegionSouth}) - 3.715(\text{Dum\_RegionWest}))$$

Model Efficacy
- Global F-test for Adequacy:
  - F-statistic: 36.69
  - p-value: <0.0001
  - **Conclusion: This model is adequate in predicting obesity rates for American states**
- Adj. R2 value: 0.82
  - **Interpretation: 82% of the variation in obesity rates is explained by this model**
- RSE: 1.295

Beta Interpretations
- The true average obesity rates in the **Northeast** are **2.578 percentage points** less than that of the Midwest region, holding all other variables constant.
- The true average obesity rates in the **South** are **0.04348 percentage points greater** than that of the Midwest region, holding all other variables constant.
- The true average obesity rates in the **West** are **3.715 percentage points less** than that of the Midwest region, holding all other variables constant.
- For a 1-unit increase in **median age**, the expected obesity rate **decreases by 0.28726 percentage points**
- For a $1000 increase in **median income**, the expected obesity rate **decreases by 0.06532 percentage points**
- For a 1-percentage point increase in **smoking rate**, the expected obesity rate **increases by 0.62363 percentage points**

Usage (Prediction):
- To see how our model fares in practice, we can test it with an actual observation
- Predicted obesity rate for Wisconsin (using final model): 35.3735
- Actual obesity rate for Wisconsin: 37.7
- **Residual: 37.7-35.3735 = 2.3264 percentage points**
- The actual obesity rate for Wisconsin is included in the 95, 98, and 99% prediction intervals.

Future Improvements

Limitations
- Data manipulations of a quantitative predictor to qualitative (i.e. insufficient sleep rates, and alcohol consumption levels) may have impacted findings.
- Additionally, the final model is limited in its usefulness as this analysis is preliminary in nature.
  - The factors that have been identified as significant have already been discovered (median income — Harvard School of Public Health) or are rather intuitive (smoking rate).
  - Analysis should be complemented by further research that delves deeper into each of the identified predictors of obesity in the US (discussed below)

Future Research
1. What are some of the factors that impact into regional differences in obesity rates?
   - How do cultural, environmental, and demographic factors play a role in obesity and overall health across different regions?
2. Since we identified a positive relationship between obesity rates and smoking rates, what are the most prominent predictors of high smoking rates in the US?
3. Which predictors hold their significance when performing analysis within states? Regions? Counties?
4. Perform the same analysis with a larger population (i.e. North America, more developed/less developed countries, analysis of global trends)

Works Cited

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

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

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

Centers for Disease Control and Prevention. (2022, July 15). *Consequences of obesity.* Retrieved from 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.* Retrieved from https://www.cdc.gov/media/releases/2024/p0912-adult-obesity.html

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

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

KFF. (2024, October 18). *Adults who report smoking by sex.* Retrieved from 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.* Retrieved from https://catalog.data.gov/dataset/national-obesity-by-state-d765a

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

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

Trust for America's Health. (2023). *Obesity report.* Retrieved from 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.* Retrieved from 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.* Retrieved from https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf

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

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

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

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