Lab 2: Great American Paradox

BMI relationship to household income in the U.S. using the National Health and Nutrition Examination Survey (NHANES), collected between August 2021 and August 2023

Datasci 203 Team 2 - G. Frimpong, K. Coppa, R. Schaefer, C. Schrupp

2024-12-13

This study investigates the relationship between income and Body Mass Index (BMI) to explore the "Great American Paradox," where economic prosperity coexists with high rates of obesity and health disparities. Using data from the National Health and Nutrition Examination Survey (NHANES) collected between 2021 and 2023, we analyze a nationally representative sample of 5,194 individuals. The dataset includes directly measured BMI and socioeconomic variables, ensuring robust and reliable analysis. Regression models evaluate how income-to-poverty ratios influence BMI while accounting for demographic factors such as age, gender, and race/ethnicity. Results highlight disparities in health outcomes across income levels, providing insights into the socioeconomic determinants of obesity. This research informs public health strategies aimed at reducing obesity-related disparities in the United States.

INTRODUCTION

Obesity is often viewed as a consequence of individual lifestyle choices, yet in the United States, it starkly reveals systemic inequalities—a phenomenon known as the "Great American Paradox." This paradox, where economic prosperity coexists with high rates of obesity and related health disparities, challenges assumptions about wealth as a protective factor for health. Surprisingly, obesity disproportionately affects lower-income groups, raising critical questions about how socioeconomic conditions interact with health behaviors and outcomes.

This study examines the relationship between income and Body Mass Index (BMI) to uncover the socioeconomic mechanisms underpinning this paradox. By leveraging robust, nationally representative data, we aim to identify patterns and disparities that contribute to obesity's persistence despite economic growth. The results will illuminate the broader social determinants of health, offering evidence to shape targeted public health interventions. In doing so, this research seeks to provide actionable insights into addressing one of America's most pressing public health challenges.

DESCRIPTION OF THE DATA SOURCE

To analyze the relationship between income and BMI, we utilize data from the National Health and Nutrition Examination Survey (NHANES) Disease Control and (CDC) (2021) , collected between August 2021 and August 2023. NHANES is a cornerstone of public health research in the United States, known for its rigorous methodology combining interviews and standardized physical examinations conducted by trained professionals. This ensures reliable, directly measured health data, particularly critical for BMI analysis.

DATA WRANGLING

Before we modified the data, the demographics file included 11,933 observations and 27 variables and the body measurements files included 8,860 observations and 22 variables. Both of these datasets include a unique ID that we used to join them. The joined dataset had 8,860 observations and 48 variables. We then made two modifications to the joined dataset to prepare it for analysis. The first was renaming the relevant columns to be more human readable. The second modification was updating the values in the categorical features. NHANES encodes all categorical variables with numeric indicators, so we replaced them with descriptive labels to make the data more human readable and to ensure the regression model treats these features as categorical variables. For the ethnicity variable, the reference was set to "White" to ensure stability in the regression model as it is the largest group.

OPERATIONALIZATION

Out of the 48 variables in the joined dataset, we kept 8 of them for our analysis. Three of them are used to ensure our data matches the complex stratification used by NHANES to be representative of the American population. Our dependent variable is BMI and our main independent variable is income, represented as a ratio of a person's income over the poverty line. We also kept the age, gender, and ethnicity variables as we hypothesized these demographic factors may also be significant predictors of BMI. Out of the 8,860 observations

in the joined dataset, 1,480 are missing data in at least one of the kept variables. We dropped these observations to ensure we could conduct our full analysis with all of the remaining data. We also dropped 2,186 observations where the subject's age is less than 20 years old. It is recommended to avoid using BMI for people under 20, so we removed these observations to get more reliable results on the remaining data Johnson et al. (2013). An income categorical variable is created to facilitate the examination of the nonlinear relationship between income and BMI, with a cutoff of 1.3 chosen to distinguish low-income groups at or near the poverty threshold, as defined by the income-to-poverty ratio Drewnowski (2009). We created a binary obesity indicator variable (BMI >= 30) to aid with visualizations (WHO) (2000). After selecting variables and dropping observations, our final dataset for analysis contained 5,194 observations and 10 variables.

DATA VISUALIZATION

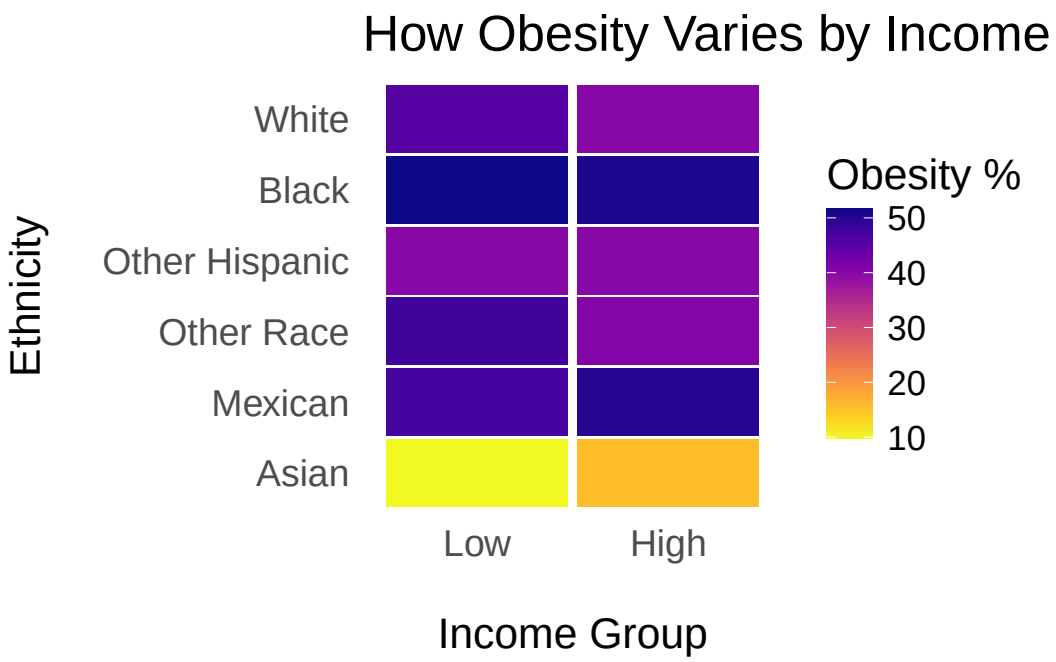

Figure 1: This heatmap shows how obesity prevalence varies by income and ethnicity. Asians are outliers with an extremely low prevalence of obesity (low-6.7%, high-11.7%) compared to the other ethnic groups. Black, Mexican, and Other Hispanic ethnicities have a higher prevalence of obesity in the high-income group.

MODEL SPECIFICATION

The regression models in this study explore the relationship between income and BMI, progressively adding demographic and interaction terms. **Model 1** serves as the baseline, using income as a continuous predictor. **Model 1a** replaces income with income categories to examine potential nonlinear effects. **Model 2** adds gender and its interaction with income to assess

gender-specific patterns. **Models 3** and **4** extend this by incorporating age and ethnicity, respectively, to explore additional moderating effects. Finally, **Model 5** includes a three-way interaction (income, gender, and ethnicity) but was excluded from analysis due to instability and lack of sufficient data in certain subgroups.

MODEL ASSUMPTIONS

The model was applied with careful attention to key assumptions and the complex survey design. Linearity was assessed through residual diagnostics, revealing no major deviations from expected patterns. The complex survey design, including weights, strata, and primary sampling units (PSUs), was properly specified using `svyglm()`, which also provides robust standard errors by default, helping to address potential heteroscedasticity and ensure valid inference and independence of observations (IID). Collinearity was evaluated and identified as a potential concern for multi-variable models but is not anticipated to pose significant issues in the current specifications. Two primary limitations were considered: the censoring of income data at 5.0, which restricts differentiation at higher income levels, and the exclusion of BMI data for participants under 20 years old, which ensures adult-specific BMI interpretations but reduces generalizability. These measures ensure the model's robustness while acknowledging areas for improvement.

MODEL RESULTS AND INTERPRETATION

Model 1 and 1a:

Income alone shows a significant (negative) relationship with BMI (p=0.002). Using income categories, the "High" category is associated with a modest but statistically significant lower BMI than the baseline category (p=0.04). Results from these two models suggest that once focusing on adults, there is a subtle socioeconomic gradient.

Models 3-5:

Income is negatively associated with BMI among females but less so (or even opposite in direction) for males. This indicates gendered patterns in how socioeconomic status translates into BMI outcomes. Including age did not significantly alter the relationship between income and BMI or reveal interaction effects. Age might already be indirectly accounted for via the adult-only sample, and no strong differential effects were detected in this linear form (p=0.5). Ethnicity affects baseline BMI levels, with Asian participants exhibiting notably lower BMIs at the reference income level (as shown in Figure 1). Yet, the income-BMI slope does not significantly differ across most ethnic groups, suggesting that while baseline risk may differ by ethnicity, socioeconomic gradients are less distinct. Attempting to model three-way interactions leads to a lack of degrees of freedom, rendering p-values undefined and estimates unreliable. This complexity suggests the need for a more parsimonious approach or a larger sample.

Conclusion

Model 1 reveals a statistically significant if modest socioeconomic gradient. Gender differences are evident, while age and ethnicity primarily influence baseline BMI rather than modifying the income relationship. Overly complex models with multiple interactions (gender, ethnicity,

income) exceed the data's capacity, leading to unstable results. A few limitations exist for this study. BMI alone does not perfectly capture body composition differences. These findings emphasize the importance of judicious model selection, and acknowledging the limitations of BMI as a measure and top-coded income, especially for subpopulations. Further research may refine these methods, incorporate additional variables, or explore non-linear models to better understand the nuanced relationship between income and obesity risk.

APPENDIX A - DATA SOURCE

1. Demographics File (DEMO_L): Provides our key independent variable: ratio of family income to poverty (INDFMPIR), Includes essential demographic controls: age, gender, race/ethnicity, Contains necessary survey design variables (weights, strata, PSUs) for proper statistical inference, Offers contextual socioeconomic information crucial for understanding income patterns.

   Link: https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2021/DataFiles/DEMO_L.xpt

2. Body Measurements File (BMX_L): Contains our dependent variable: Body Mass Index (BMI), Includes height and weight measurements taken by trained health technicians, Ensures standardized measurement protocols across all participants, Provides the physical data necessary to explore our central research question.

   Link: https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2021/DataFiles/BMX_L.xpt

APPENDIX B - STATISTICAL MODEL COMPARISON

Comparison of Models

	Dependent variable:				
	BMI				
	Model 1	Model 1a	Model 2	Model 3	Model 4
	(1)	(2)	(3)	(4)	(5)
Income	-0.33***		-0.64***	-0.47	-0.30
	(0.09)		(0.15)	(0.27)	(0.16)
Income Category: High		-0.81**			
		(0.36)			
Gender: Male			-2.60***		
			(0.80)		
Age			0.67***		
			(0.20)		
Ethnicity: Asian				0.01	

				(0.01)	
Income x Gender				0.003	
				(0.005)	
Income x Age					-4.63***
					(0.73)
Income x Ethnicity					1.56
					(1.14)
ethnicityMexican					-0.75
					(0.76)
ethnicityOther Hispanic					-0.74
					(0.93)
ethnicityOther Race					0.04
					(1.17)
income:ethnicityAsian					0.02
					(0.21)
income:ethnicityBlack					-0.02
					(0.29)
income:ethnicityMexican					0.57
					(0.40)
income:ethnicityOther Hispanic					0.29
					(0.31)
income:ethnicityOther Race					0.21
					(0.43)
Constant	30.75***	30.39***	31.92***	30.29***	30.71***
	(0.32)	(0.41)	(0.57)	(0.88)	(0.58)
Note:					*p<0.1; **p<0.05; ***p<0.01

Table 1: The stargazer table compares five models exploring the relationship between income and BMI, progressively incorporating additional variables and interactions. Model 5 is not included.

APPENDIX C - RESIDUALS VS FITTED VALUES PLOT

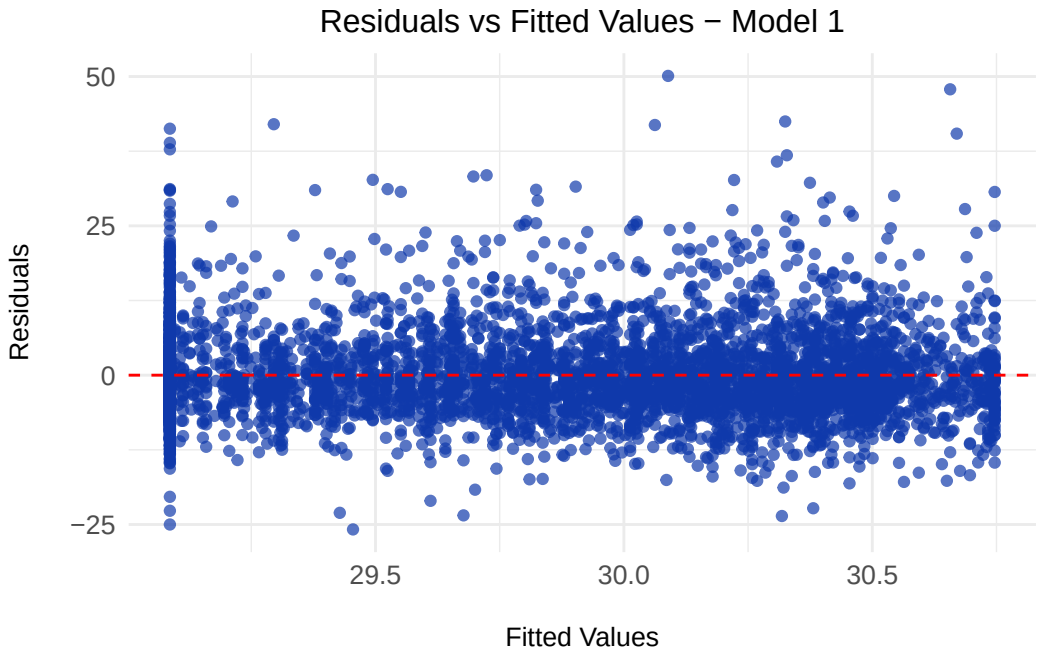

Figure 2: The Residuals vs. Fitted values plot for Model 1 shows residuals scattered around the zero line, with minor heteroscedasticity at higher fitted values. The censoring cap at 5 for the income variable likely contributes to this variability, suggesting room for model improvement.

REFERENCES

Disease Control, Centers for, and Prevention (CDC). 2021. *National Health and Nutrition Examination Survey Analytic Guidelines*. https://www.cdc.gov/nchs/nhanes/index.htm.

Drewnowski, A. 2009. "Obesity, Diets, and Social Inequalities." *Nutrition Reviews* 67 (suppl_1): S36–39.

Johnson, C. L., R. Paulose-Ram, C. L. Ogden, and et al. 2013. "National Health and Nutrition Examination Survey: Analytic Guidelines, 1999–2010." *Vital Health Stat* 2 (161).

(WHO), World Health Organization. 2000. *Obesity: Preventing and Managing the Global Epidemic*. WHO Technical Report Series 894.