# Understanding the Impact of Socioeconomic and Regional Factors on Health and Educational Outcomes

Rishika Randev, Jenny Wu, Uzoma Uwazurike Jr., Shiyue Zhou

## Abstract

This study examines the relationships between socioeconomic, health, and education indicators in U.S. counties using data from the American Community Survey, the USDA Food Environment Atlas, small area income and poverty estimates, and other sources. We will explore regional influences and predictors of life expectancy, and also identify socioeconomic factors associated with school funding adequacy. By analyzing differences in life expectancy and education funding across socioeconomic and regional contexts, this study aims to gain insight into public well-being and economic inequality across the U.S. Through exploratory data analysis and regression modeling with interaction terms, we identify significant regional disparities in both life expectancy and school funding adequacy. The findings highlight that regions, particularly the West and Northeast, tend to have better health and education outcomes, emphasizing the critical role of location and economic conditions in shaping societal well-being.

## Introduction

### Background

Life expectancy and school funding are important indicators of public well-being because they capture fundamental aspects of health, education, and social equity that are essential for the development of individuals and communities. Life expectancy is an important health indicator that reflects not only the overall health of a society, but also its stability, economic resilience, and levels of inequality. For example, differences in life expectancy between regions and socioeconomic groups often reveal systemic inequalities, with marginalized communities experiencing shorter life spans due to limited access to health care, healthy living environments, and economic opportunities. Similarly, adequately funded schools mean that students

have equal access to learning opportunities and are an important investment in social mobility. Underfunded schools are often located in economically disadvantaged or segregated areas, exacerbating educational gaps and limiting students' potential to break the cycle of poverty. Because both of these outcomes are impacted by a complicated web of interactions between socioeconomic, political, and cultural factors, it is vital that we work towards better understanding the most influential out of these, so that policy and community intervention designed to improve life expectancy and school funding can take on a multifaceted, and thus a more effective approach.

### Research Questions

For this purpose, we focused our analysis on two key research questions. First, we investigated how socioeconomic factors—residential segregation, access to healthy foods, and health insurance coverage—impact average life expectancy. This analysis included a comparison across different regions to identify variations in the relationship between residential segregation and life expectancy. Second, we evaluated how residential segregation, property taxes, and median household income influence school funding adequacy. This question allowed us to assess educational investment disparities among regions with varying economic conditions. The observational unit for this analysis was the U.S. county.

### Data Overview

The dataset used in this study combined information from three primary sources.

The first source was the County Health Rankings & Roadmaps program, associated with the University of Wisconsin Population Health Institute, which provided us with 2024 datasets for all U.S. states. This program consolidates the latest county-level measurements of various population health, economic, demographic, and social factors from the American Community Survey, the USDA Food Environment Atlas, Small Area Income and Poverty Estimates, and other reliable government sources into publicly available datasets every year. Our specific variables of interest were originally collected as part of:

- National Center for Health Statistics - Natality & Mortality Files, 2019-2021 (average life expectancy)
- American Community Survey 5-year estimates, 2018-2022 (residential segregation index & median household income)
- USDA Food Environment Atlas, 2019 (percentage of population with limited access to healthy foods)
- Small Area Health Insurance Estimates, 2021 (percentage of adults uninsured)
- School Finance Indicators Database, 2021 (school funding adequacy)

The second source was IPUMS, which provided us with median poverty tax data for every county from 2018-2022.

The third source was the U.S. Census Bureau's Regions and Divisions of the United States, which maps every U.S. to one of four geographic regions, and also a division within the regions. We specifically included regions as a variable in this analysis. This data was loaded into R using a publicly available GitHub repository titled *Census Regions*.

## Methods

### Variable Selection

Predictor variables were selected *a priori* based on their perceived relevance to the two outcomes we wanted to better understand: average life expectancy and school funding adequacy. Nutrition and access to quality healthcare are inherently linked to overall well-being. To analyze the extent to which each contributes to improving life expectancy, we included two key predictors: the percentage of the population with limited access to healthy foods and the percentage of uninsured adults in a county. The former is defined by the US Department of Agriculture as "the percentage of the [county] population that is low-income and does not live close to a grocery store", where "close" is defined differently for rural and nonrural areas. The latter variable represents the percentage of adults in a county younger than 65 who lack health insurance. We also were interested in seeing how residential segregation would impact life expectancy among these predictors, given existing studies on the negative health effects of segregation, and whether this relationship would exhibit any variations across different regions. This is why we chose to include residential segregation index values. The residential segregation index, used by the County Health Rankings & Roadmaps program, measures how evenly two groups, such as Black and white residents, are distributed across geographic areas, with values ranging from 0 (complete integration) to 100 (complete segregation).

School funding adequacy is defined by the County Health Rankings program as "the average dollar gap between actual per-pupil spending and the amount needed for students to achieve national average test scores, considering districts' varying equity-based needs." Because a major source of public school funding is tax revenue, especially property tax, we chose to include this variable as a predictor, along with an indicator of the general economic conditions of a county—median household income. In addition, to explore how race and racial diversity interface with variations in school funding, we selected residential segregation to be a part of this regression as well.

### Exploratory Data Analysis & Data Preprocessing

2024 datasets for all U.S. states were combined and merged with median property tax and region mapping data, resulting in a final dataset with 3144 county observations and 9 differ-

ent variables (state, average life expectancy, residential segregation index, % limited access to healthy foods, % uninsured adults, school funding adequacy, median household income, median property tax, and region). Because school funding adequacy was given as a numerical value, and we wanted to focus on better understanding the dichotomy between school funding adequacy and inadequacy, this variable was converted into a binary categorical variable (where the value was set to 1 if the numerical value was positive, indicating the counties' schools were adequately funded, and set to 0 if the numerical value was negative, indicating underfunding).

Exploratory data analysis for the two outcome variables was conducted by 1) creating scatterplots of each individual predictor against the outcome (for life expectancy) and 2) creating pairwise plots between variables (for school funding). During the initial stages of analysis, we found that  of our observations were missing residential segregation index data, and in addition, certain states were missing nearly all of their segregation values. Because of this, we decided to drop all states missing over 50% of their values for any variable. This led to our analysis being limited to 33 states, and excluding the following: Alaska, Colorado, Idaho, Iowa, Kansas, Minnesota, Montana, Nebraska, North Dakota, South Dakota, Utah, Wyoming, and Vermont. We were left with 2379 total observations, and the missing percentage for residential segregation dropped to 22%.

After this, remaining missing values in all variables were imputed using the mice package and predictive mean matching to generate one complete dataset. The original school funding adequacy numerical variable was first imputed, and then converted into binary values for subsequent analysis.

### Model Fitting & Assessment

For our first research question, a multiple linear regression model was fit, regressing average life expectancy on residential segregation index, percentage with limited access to healthy foods, percentage uninsured adults, and region. Linear regression assumptions were assessed using diagnostic plots, especially residual vs. fitted and quantile-quantile plots. The adjusted R-squared value was used to evaluate the fit of the model, and the effect of region as an interaction with residential segregation was evaluated using nested F tests. We also calculated VIF to check for multicollinearity and Cook's distance to identify influential points.

For our second research question, a binary logistic regression model was fit with school funding adequacy (categorical) as the outcome and residential segregation index, median property tax, region, and median household income as predictors. The model was assessed using a confusion matrix, an ROC curve, and a comparison of deviance between the full model and a reduced model excluding region as a predictor.

## Results

### Data Overview

The below tables display summary statistics on average life expectancy (1) and school funding adequacy (2). Both the median and mean life expectancy were 75 years. School funding adequacy counts show that around 66% of the counties in our analysis had under funded schools.

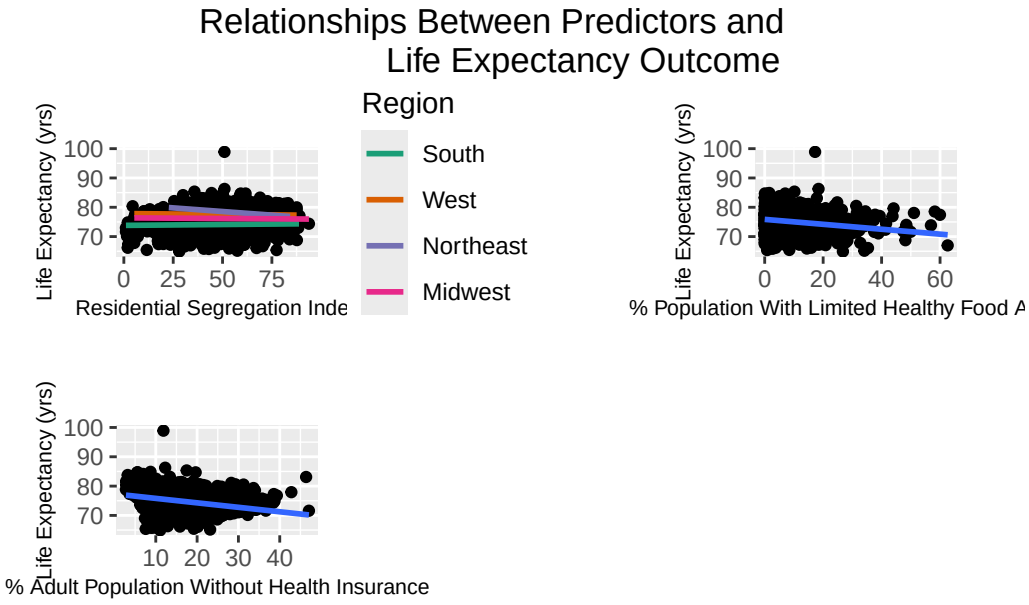
| | **Average Life Expectancy (years)** |
|---|---|
| Min | 65 |
| 25th percentile | 73 |
| Median | 75 |
| 75th percentile | 77 |
| Max | 99 |
| Mean | 75 |

| **0 (Under Funded)** | **1 (Adequately Funded)** |
|---|---|
| 1539 | 806 |

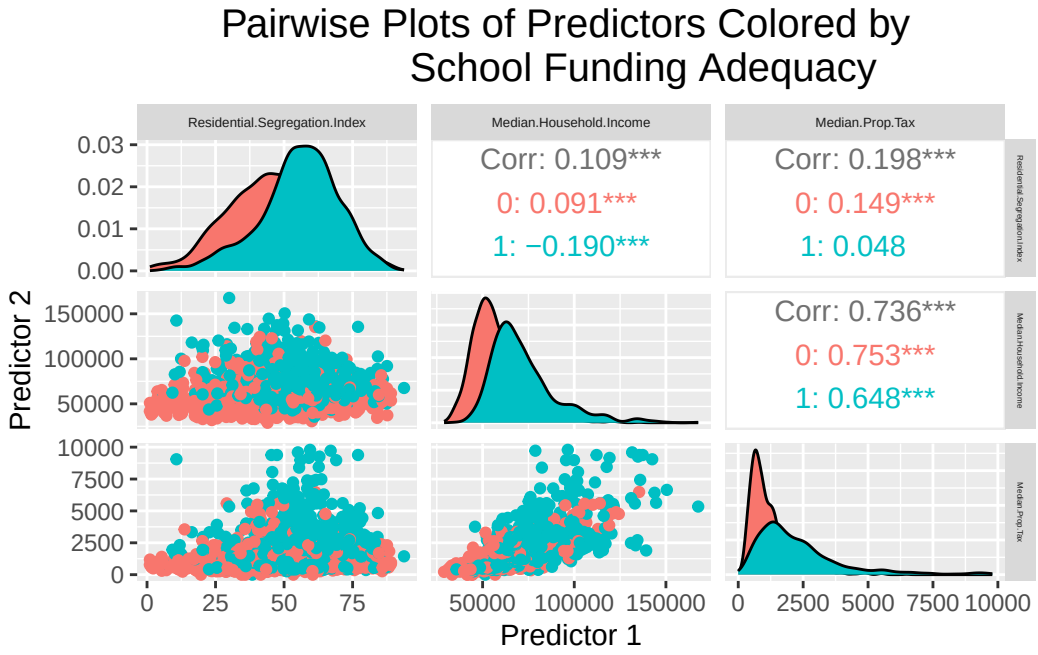
Additionally, the number of counties per region in this analysis indicate an imbalance between categories, with the South region being over represented (holding 60% of all counties), and the East and Northeast holding around 9% of the counties each.

| **Midwest** | **Northeast** | **South** | **East** |
|---|---|---|---|
| 552 | 203 | 1421 | 203 |

As mentioned earlier, EDA for the first research question involved generating scatterplots plotting each numerical predictor plotted against life expectancy, as shown below. The residential segregation index vs. life expectancy plot was also faceted by region to explore whether each region demonstrated a different relationship in these two variables. Because the slopes were largely similar, this suggested that the interaction effect might not be that strong. In general, all three scatterplots demonstrated considerable spread and did not show a very substantial linear relationship between the numerical predictors and average life expectancy.




For the second research question, pairwise plots were generated to show the relationship between predictors (as seen by the scatterplots below), and also to show the relationship between each numerical predictor and the school funding outcome variable (as seen by the density plots below). Points and plots colored in red represent counties where school funding is inadequate (=0), and points and plots colored in blue represent counties where funding is adequate (=1).




### Question 1

### Model Results

Based on the multiple linear regression results, percentage of uninsured adults, percentage of population with limited access to healthy foods, and region all emerged with statistically significant coefficient values. The interaction term between region and residential segregation, as well as the individual term of residential segregation, were not statistically significant. While percentage of uninsured adults and percentage with limited access to healthy foods were statistically significant, the coefficients themselves indicate very minor impacts on life expectancy (a 1 unit increase in either of these percentages leads to average life expectancy decreasing by around 0.05 years). The impact of region on life expectancy, however, was notable: life expectancy is on average 4.17 years higher in the West region compared to the South, 6.7 years higher in the Northeast compared to the South, and 2.41 years higher in the Midwest compared to the South.

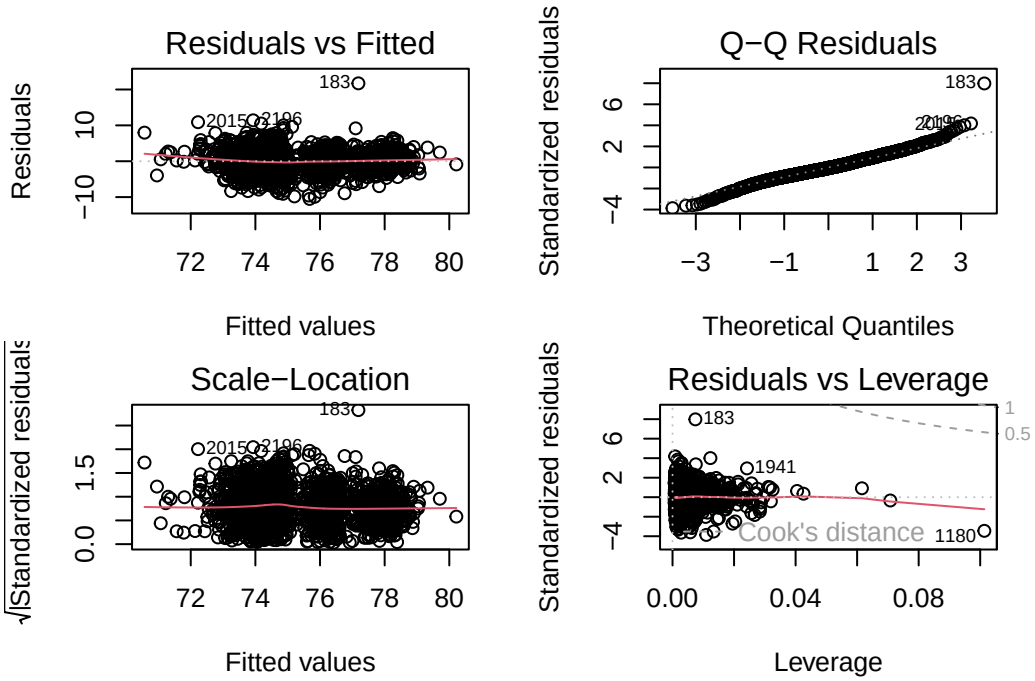
| | (1) | | | | |
|---|---|---|---|---|---|
| | Est. | S.E. | 2.5 % | 97.5 % | p |
| (Intercept) | 75.02 | 0.27 | 74.49 | 75.55 | <0.01 |
| Residential Segregation Index (RSI) | 0.0074 | 0.0046 | −0.0016 | 0.0165 | 0.11 |
| Region-West | 4.17 | 0.96 | 2.28 | 6.05 | <0.01 |
| Region-Northeast | 6.7 | 1.2 | 4.4 | 9.0 | <0.01 |
| Region-Midwest | 2.41 | 0.61 | 1.22 | 3.60 | <0.01 |
| Percentage With Access to Healthy Foods | −0.0552 | 0.0085 | −0.0718 | −0.0386 | <0.01 |
| Percentage of Uninsured Adults | −0.045 | 0.010 | −0.065 | −0.026 | <0.01 |
| RSI*Region-West | −0.018 | 0.016 | −0.049 | 0.014 | 0.28 |
| RSI*Region-Northeast | −0.059 | 0.019 | −0.097 | −0.021 | <0.01 |
| RSI*Region-Midwest | −0.0150 | 0.0104 | −0.0355 | 0.0054 | 0.15 |

### Model Assessment

The adjusted R-squared value for the model was around 0.2, indicating that only a small amount of the variability in life expectancy was explained by the predictors in the model. This points to the fact that are a plethora of other socioeconomic and environmental factors that have an effect on life expectancy, and expanding our model to include more of these would have likely yielded a better fit.

In conducting diagnostics on the multilinear regression model, we first checked to see if the model meets our assumptions of linearity, independence, normality, and homoscedasticity. The residuals vs. fitted plot showed relatively equal variance and no discernible pattern, meaning that the linearity and homoscedasticity assumptions were reasonably met. In addition, because each observation in the dataset represented a unique county in the U.S., independence can be reasonably assumed. Assessing the Q-Q Plot for normality, the residual distribution was approximately normal, despite a slight divergence on both tails of the plot. Finally, through observing the Residuals vs. Leverage plot (not pictured), we observed only a few points falling near or beyond the Cook's distance threshold, suggesting that there may be a few influential points that can be seen as significantly impacting the results of the regression. These points could affect the validity of the model.

With variance inflation factors of roughly less than 2 for all predictors, there was little to no multicollinearity amongst the predictors. Through a nested-f test comparing a full model with a reduced model that excludes the interaction term between region and residential segregation, we concluded that a full model fit the data substantially better. The reduction in RSS in the full model compared to the reduced model was highly significant, indicating that the interaction between region and residential segregation added meaningful explanatory power to the model.




## Question 2

Pivoting to the second research question, we sought to assess the impact residential segregation, median property taxes, the median household income, and the region of which we are assessing has on whether or not a school is adequately funded by using a logistic regression model.

### Model Results

In holding all other variables constant and using the South region as the baseline category for region, the table below displays the odds ratio results from the logistic regression model.

While all estimates are statistically significant at the 95% confidence level, exponentiated regression coefficients from the logistic model demonstrate that the level of residential segregation, the median household income, and median property tax of the county have very minimal impacts on the odds of a school being adequately funded.

Region, however, has a substantial impact on the odds of school funding adequacy. Being located in the Northeast or the Midwest has the greatest impact on how well the school is funded, with the odds of adequate school funding being over 200 times higher in a Northeast county compared to a Southern county, and the odds of adequate school funding being over 3 times higher in a Northeast county compared to a Southern county, holding all other variables constant. Such disparities across the country are caused by a number of factors, including: (1) capacity - how well off a state is based on its economy and resources, and (2) effort - the

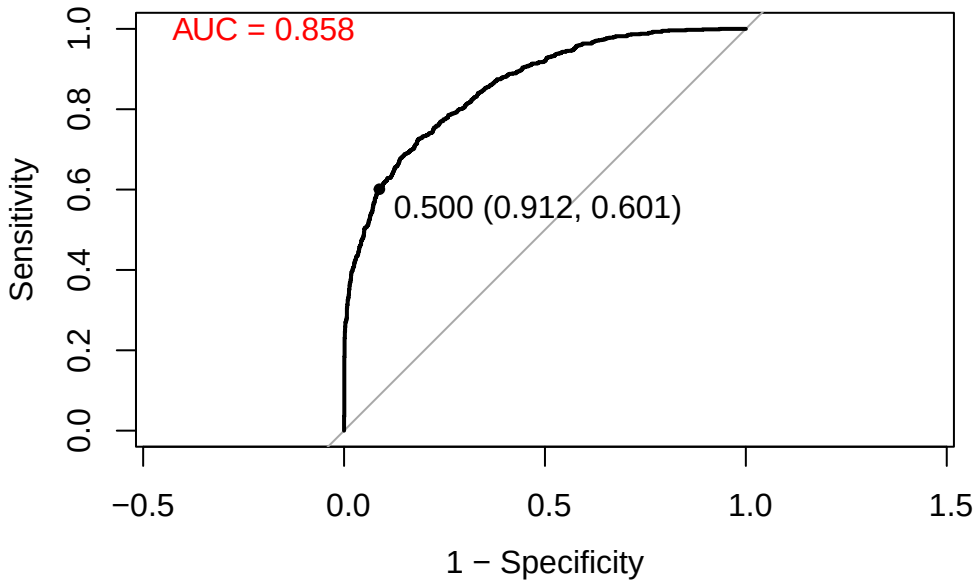
| | (1) | | | | |
|---|---|---|---|---|---|
| | Est. | S.E. | 2.5 % | 97.5 % | p |
| (Intercept) | 0.001 35 | 0.000 50 | 0.000 64 | 0.002 73 | <0.01 |
| Residential Segregation Index (RSI) | 1.0196 | 0.0041 | 1.0117 | 1.0277 | <0.01 |
| Median Household Income, USD | 1.0 | $5.5 \times 10^{-6}$ | 1.0 | 1.0 | <0.01 |
| Median Property Tax, USD | 1.0 | $8.5 \times 10^{-5}$ | 1.0 | 1.0 | 0.01 |
| Region-West | 1.83 | 0.35 | 1.26 | 2.66 | <0.01 |
| Region-Northeast | 232 | 141 | 82 | 974 | <0.01 |
| Region-Midwest | 3.84 | 0.52 | 2.95 | 5.02 | <0.01 |

states willingness to provide funding for education. Wealthier states with a high fiscal capacity, (typically those in the Northeast), have more funding available to spend on education than states with more limited resources (typically those in the South and the West).

### Model Assessment

Evaluating our model, we found that the logistic regression model demonstrated a strong overall performance with an accuracy rate of 80.33%. The model had a high sensitivity of 91%, indicating that it accurately predicted 91% of actual positive cases of inadequate school funding when evaluated on our dataset. However, it had a 60% specificity rate, highlighting it correctly identified cases of adequate school funding only 60% of the time. Moreover, looking at the ROC curve below, the AUC is 0.858, indicating a moderately strong discriminate ability.

In comparing our full model (with predictor variables for residential segregation, median household income, median property tax, and region) to a model with including all of these variables except for region, we found that there is statistical significance in the deviance between the residuals of the full model and reduced. As such, we can conclude that including region as a predictor refines and creates a better fit for our model.




### Conclusion

By analyzing these two research questions, we observed the significant impact of region on both life expectancy (a health indicator) and school funding (an education indicator), both of which are key components of societal well-being. These indices tend to be higher in the West and Northeast, which are more developed regions. The findings provide empirical evidence supporting the significant disparities in life expectancy across regions with differing economic conditions. They also provide insights into social policy investment decisions, highlighting the importance of fair resource allocation to people's well-being.

Our analysis faced limitations due to missing data, causing our dataset to be restricted to 33 states. Additionally, because The County Health Rankings program consolidates the latest from a variety of governmental sources, most of our variables were pulled from different years between 2018 and 2022; we assumed that trends remained largely consistent between these years, treating time as insignificant, but this may not actually be the case.

Future studies could incorporate more social variables to enhance the analysis and investigate the causal relationships between region and outcomes. Expanding the dataset to include more states and addressing missing data would also provide a more comprehensive understanding.