Socioeconomic Effects on Health and Well-being Using U.S. County Data

Lily Shaw | Steven Womack | Minh Le | Advisor: Erdogan Dogdu

College of Science and Engineering
Department of Computer Science

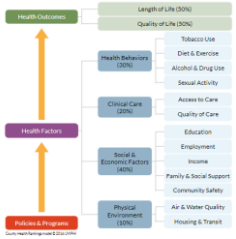
With rising disparities in socioeconomic factors, how are health outcomes being affected in the United States?

Using a publicly available, county-level data on social demographics, behaviors and health outcomes, we explore select questions and correlations between factors. The questions focus on the possible correlations between socioeconomic factors, race/ethnicity, and health outcomes for counties in which data was provided. Findings are presented with correlational matrices and heatmaps of the US counties for visualizations. We attempt to answer questions like:

Do areas with higher negative economic factors have worse health outcomes than areas with better economic factors? Is there a correlation between educational attainment and health outcomes? How does income inequality, childhood poverty, affect health behaviors such as smoking, alcohol intake, etc.? How do unemployment and lower insurance rates affect mental health versus physical health outcomes? Do areas with higher ratios of primary care physicians and mental health providers have larger amounts of teen births? As unemployment and educational attainment rise or fall, do health outcomes change as a result? How does the physical environment of an area (county) affect the health outcomes of the population?

Dataset

-Publicly collected and available for use
-Data on all +3000 US counties
-Social behaviors and health outcomes
-Breakdowns by race/ethnicity



Methodology

Most work was done using Python3 on Google Colab with standard libraries for visualizations and calculations.
Libraries Include:

- Pandas
- Altair
- Seaborne
- Numpy
- SKLearn
- XGBoost
- MatPlotLib

Correlations were calculated using standard Pearson correlations included with the Pandas library. Predictive models were developed using standard linear regression and data cleaning practices.

Data Cleaning Methods:

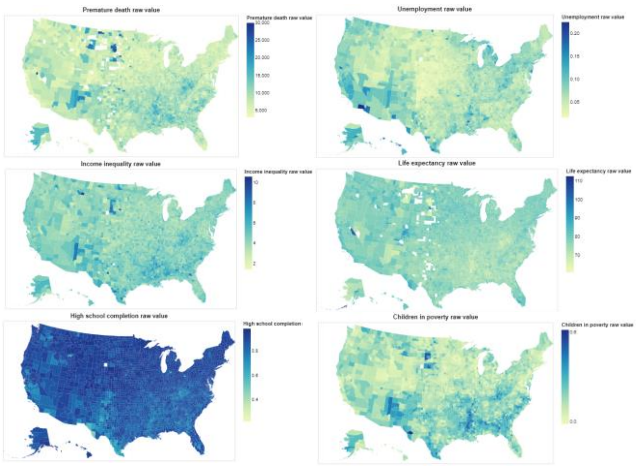
- Box plots to eliminate outliers
- Dropping incomplete data
- Removing overall totals and aggregates

Overall out of 3,143 counties, roughly 150-200 were dropped. With a drop rate of 5%-7% for incompleteness, we feel confident in our data being representative.
A test/train split of 50% ensured that the model was not over adjusted to the training data. Experimentation with lowering our MSE (Mean Square Error) using different models and different ratios for the train/test split confirmed our current ratio is ideal. Rescaling data was done for certain categories that were explored. This ensured that no MSE was over or under reported due to data scale.

Findings – Data Analysis

Findings are primarily in the form of heatmaps and correlational matrices to show geographical distribution of trends and possible correlations in specific regions.

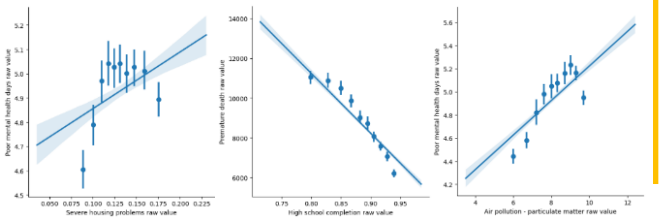
- Strong inverse correlation between educational level and health outcomes
- Emphasizes the criticality of education in well-being

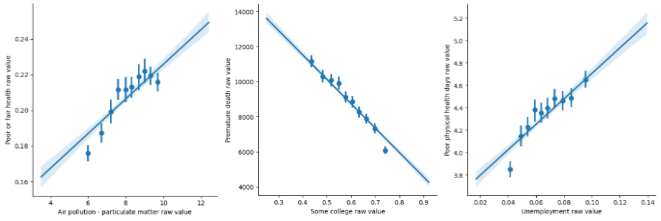


- Regionalized trends and correlations appear more than national correlations
- Regions with higher negative economic factors appear to have corresponding negative health outcomes
- The higher an area has in negative social factors, the higher the premature death rate and lower life expectancy
- National trends tend to dilute correlations and cloud findings

Findings – Predictive Modeling

Findings are primarily in the form of linear regressions and lines of best fit. Lines are of predicted values from the given dataset. Explored topics are unemployment, educational attainment, and health outcomes/effects.





Using standard linear regressions, we can accurately predict certain outcomes given certain socioeconomic factors. While we had success in our findings with certain factors giving strong models, many more factors gave results that were mixed and inconclusive.
Some predictive models found were:

- Higher educational attainment resulting in a lower premature death rate
- Higher air pollution rates result in more negative health outcomes and results

Some unreliable models were:

- Severe housing problems and mental health problems
- Educational attainment and birth weight

Conclusions

Overall findings showed correlations between socioeconomic factors and health outcomes. While certain trends were regionalized, such as unemployment and life expectancy, many factors showed strong correlations nationally. The difference in those types can be attributed to clustering of data near the mean.

Resources

- Vega Datasets – County FIPS Code https://github.com/vega/vega- datasets
- County Health Data and Rankings https://www.countyhealthrankings.org/explore-health-rankings/rankings-data-documentation