Examining the Influence of Socioeconomic Factors on Premature Death Across Racial Groups

Tasnim Rida, Carnegie Mellon University
Lissandro Alvarado, Bucknell University
Oreoluwa Williams, Towson University
Naomi James, University of Buffalo

Carnegie Mellon University
Statistics & Data Science

Background

- A person's health is determined more by their zip code than by genetics or family history (Morgan, 2019).
- Low socioeconomic status is accompanied with limited access to healthcare resources.
- Premature deaths are often **preventable**; examining this measure alongside socioeconomic factors can reveal specific areas for improvement in the United States healthcare system.

Research Question

Do income inequality, unemployment, high school completion rates affect the number of premature deaths of certain racial groups at the county level?

We hypothesize that these socioeconomic factors will emerge as significant predictors of premature deaths.

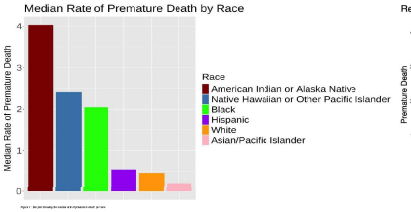
Data

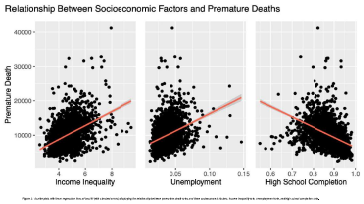
Dataset sourced from County Health Rankings Data by the University of Wisconsin Population Health Institute.

Variables of Interest

- Income Inequality: Ratio of household income at the 80th percentile to income at the 20th percentile
- Unemployment: Percentage of population ages 16 and older unemployed but seeking work
- High School Completion: Percentage of adults ages 25 and over with a high school diploma or equivalent
- Percent Population: Percentages of the population by county.
- Premature Death: Years of potential life lost before age 75 per 100,000 population (age-adjusted)

Exploratory Data Analysis



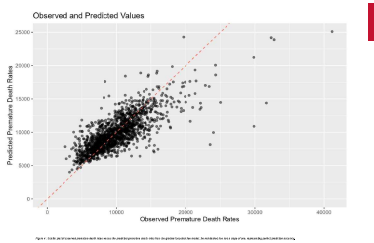


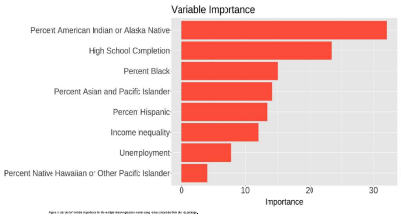
- We normalized the population size for each races to avoid any biases in the difference in population
- AIAN tend to experience more premature deaths with a median rate of 4 years lost for every 100,000 people in this ethnicity
- This bar chart reveals that there are some clear patterns of premature deaths within races

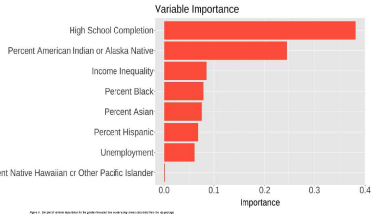
- Figure 2 shows us reasonably linear relationships between each of the predictor variables and premature death rate.
- As the income inequality in a given county increases, the premature deaths also increases.
- Similarly, as the unemployment in a given county increases, the premature death also increases.
- As the high school completion rate in a given county increases, the premature death rate decreases.

Results

- We fit a multiple linear regression and a gradient boosted tree; based on cross-validated RMSE values we selected the **gradient boosted tree as our final model**.
- The predicted values of premature death from the gradient boosted tree closely resemble the actual values of the response variable.
- High school completion emerged as the most important variable; NHOPI was the least important in both models.







Methods

We built a multiple linear regression model and a Gradient Boosted Tree to explore the relationship between premature death and socioeconomic factors.

Multiple Linear Regression

- Assumes linear relationships between predictors and response, the errors are normally distributed, homoscedasticity, and independence.
- Estimates coefficients using the least squares method.

Gradient Boosted Tree:

- Non-parametric model to capture complex relationships.
- Boosting improves prediction accuracy by focusing on largest residuals.

Discussions

Discussion

- The gradient boosted tree model has the greatest predictive accuracy.
- High school completion is the most significant predictor of premature deaths
- Insights from the gradient boosted tree model contrast with the results of the multiple linear regression model.

Limitations

- County-level analysis does not provide insights for individual data
- Incomplete racial breakdown data for socioeconomic factors

Future Work

- Obtain racial breakdowns for income inequality, unemployment, and high school completion
- Include age and gender as additional predictors