Investigation of US Drug Overdose Death Rates

By: Jintong Chang, Wenxin Lan, Li Li, Henry Wu, Angela Zhu
Advisor: Peter Freeman

Introduction

- Drug overdose deaths have long been a major concern for the US healthcare system. Health service companies, like Optum (our client), are seeking to understand the socio-economic factors related to overdose fatalities.
- The goal of this research project is to detect trends in the US drug overdose death rate over the last four years while also identifying the potential predictors of the death rate.

Trend Analysis

- Across the US, the overdose death rate increases monotonically from 2019 to 2022, as shown in Figure 2. An ANOVA test shows that the death rate is not constant as a function of time (F=59.9 for 1 and 7180 degrees of freedom, p < 0.001).
- We apply linear regression to the death rate data county-by-county and find that for a majority of counties, the rate is increasing with time (Figure 3 and 4).

EDA

- We perform EDA on eight variables that are of interest to our client: Gender, English Non-Proficiency, Race, Rural Population, Mental Health Providers, Education, Unemployment, and Income inequality.
- Most variables show strong skewness, and appear highly concentrated.
- There are associations among several variables (Figure 5).

Modeling

- Best subset, Lasso, and stepwise regressions are performed on the most updated data of 2022 to find variables that are the most associated with drug overdose death rate. The commonly selected variables are listed in Table 1.
- All linear models perform well on the test dataset, MSE ranging from 37.60 to 37.97. The best subset model is the simplest, with only 23 predictors.
- Decision tree and random forest are outperformed by the linear models.

Sensitivity Analysis

- Logistic regression is used to see which variables are significantly associated with whether the death rate is NA for a specific county (Table 2).

Conclusion

- We discover a rising trend in the U.S. drug overdose death rate in the years 2019 to 2022, both overall and at a county level.
- We obtain useful features that correlate with high drug overdose on a county level and investigate the sign of their correlations.
- Our analysis indicates that our qualitative findings are not strongly affected by missing data values.

Data

- Our data is retrieved from the University of Wisconsin Population Health Institute website and the data are presented at the county level. Each year’s data consists of 86 raw values of socio-economic variables for each of the 3142 counties including the drug overdose death rate, which is expressed as deaths per 100,000 population.
- Drug overdose deaths fewer than 10 are suppressed and the suppressed data makes up approximately 40 percent of the data. Our EDA and modeling exclude those observations, and a sensitivity analysis is conducted to study its effect on our result.

Figure 1: The amount of suppressed data (red) versus present data (blue).
Figure 2: Yearly trend of overdose death rate through the US from 2019 to 2022
Figure 3: Distribution of county-level linear regression slopes. Linear regression coefficients with p < 0.05 are classified as significant.
Figure 4: Yearly trend of overdose death rate on US map at county level reflected in linear regression. The white region is where the data are suppressed.
Figure 5: Correlation plot for variables.
Figure 6: Comparison of linear regression coefficients (with error bars) taking in different substitutions of NA in the death count.