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**.

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).

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).
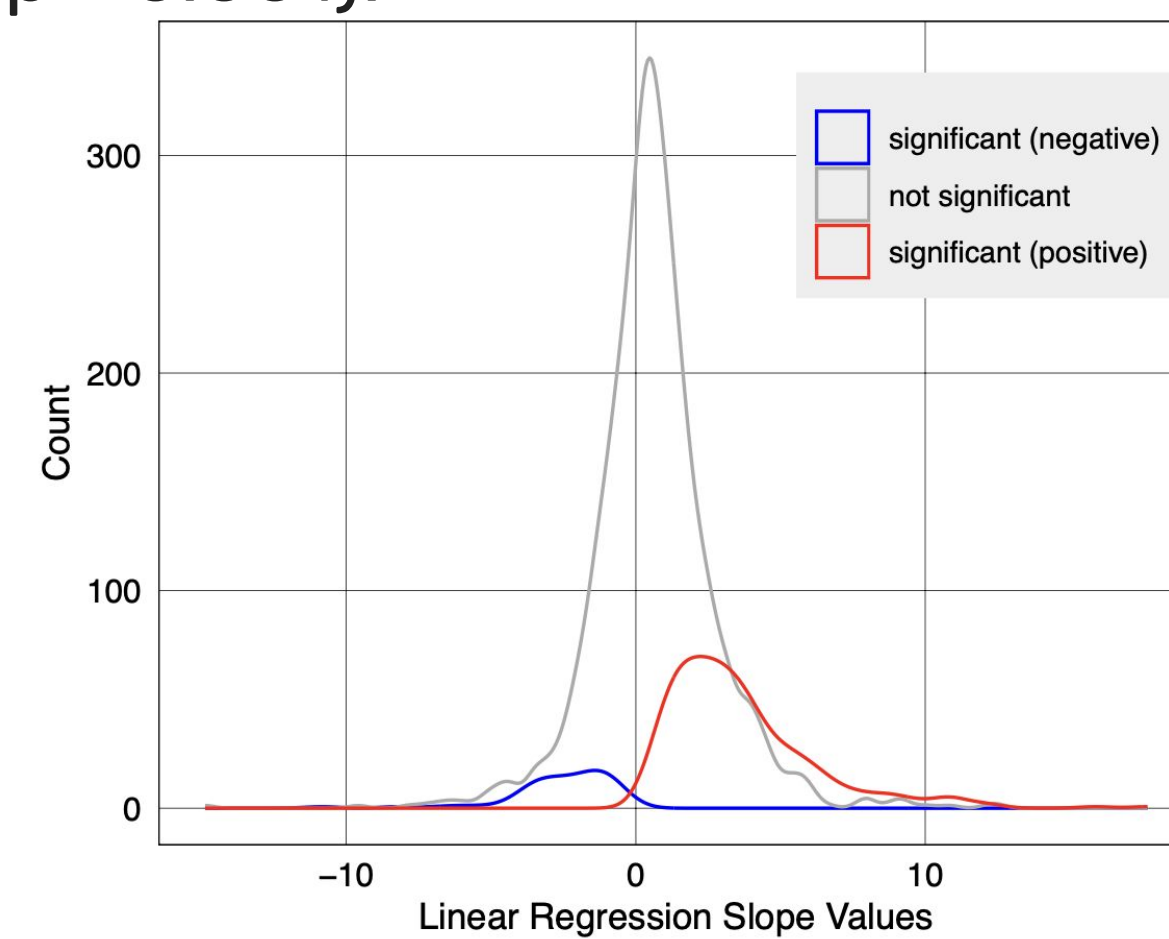
Figure 3: Distribution of county-level linear regression slopes. Linear regression coefficients with $p < 0.05$ are classified as significant.
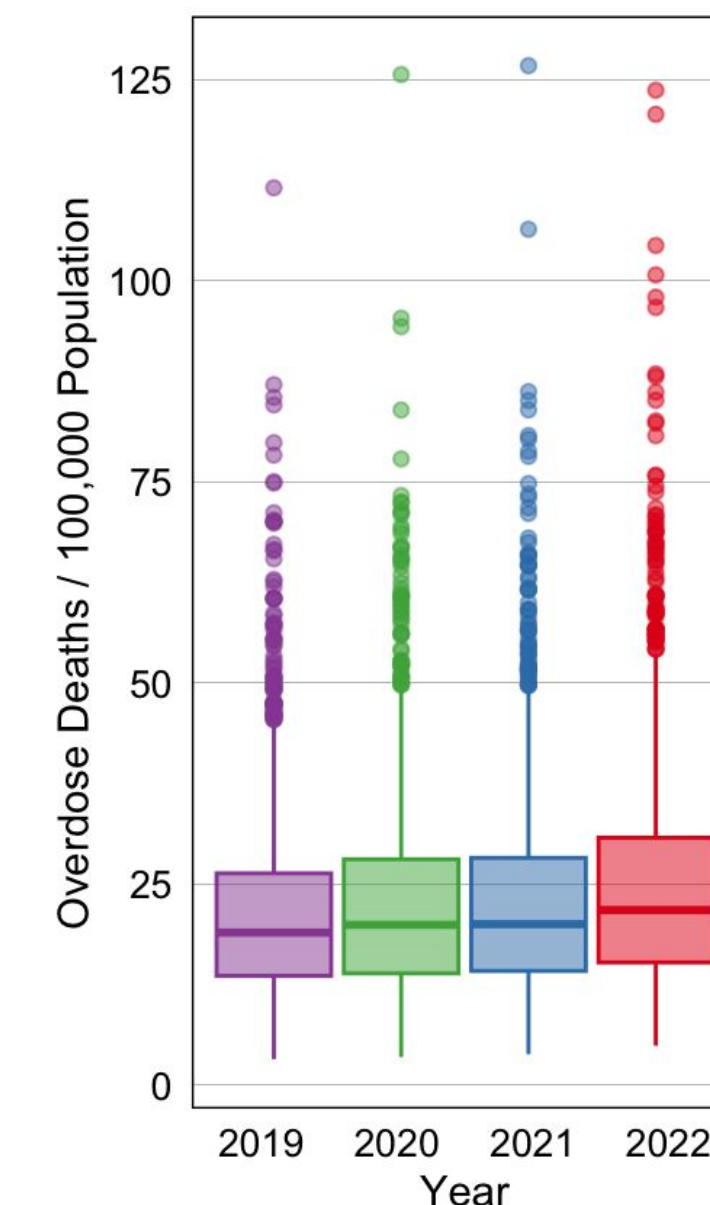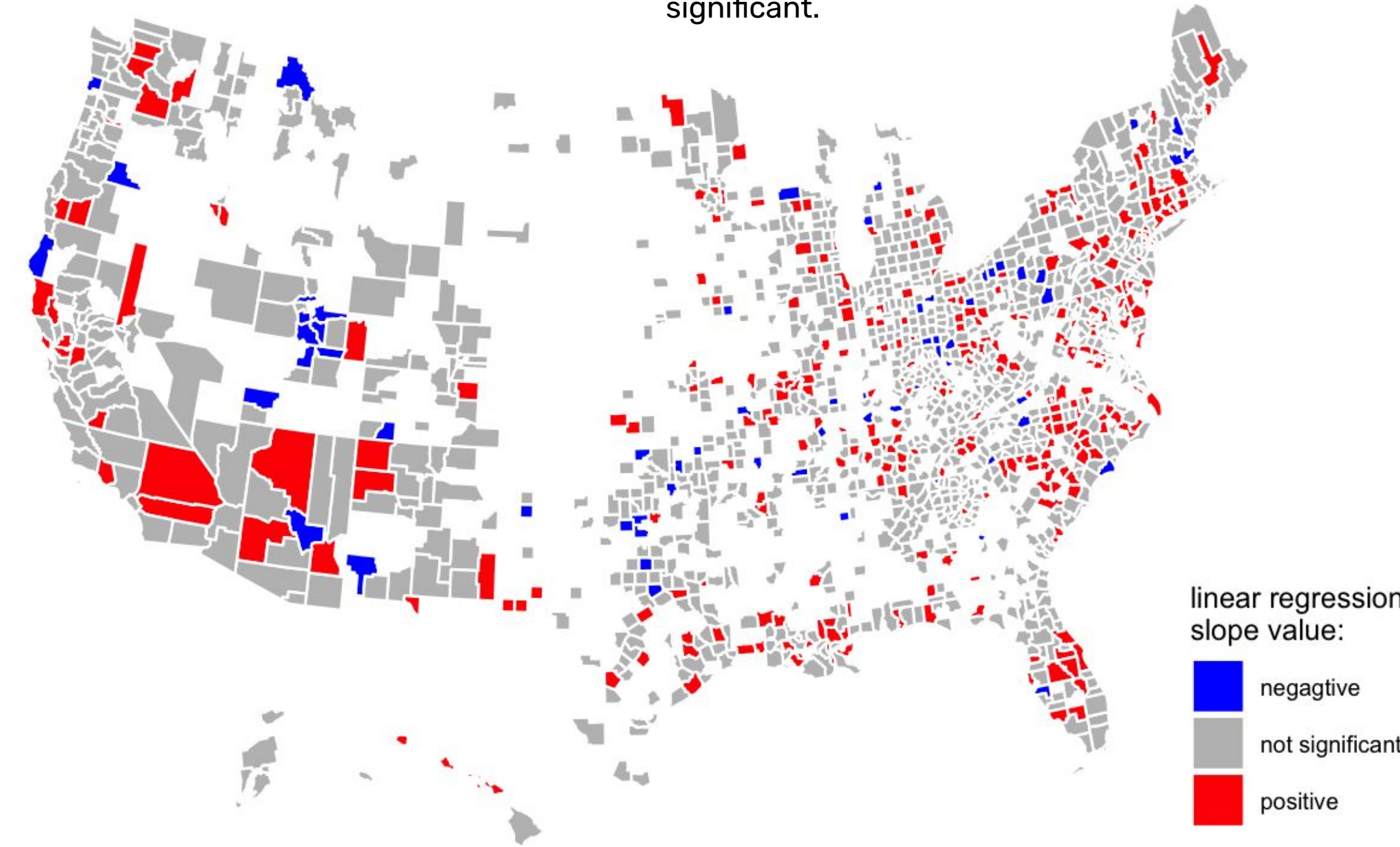
Figure 2: Yearly trend of overdose death rate through the US from 2019 to 2022

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.

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).

Figure 5: Correlation plot for variables

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.

Demographics	Social & Economic	Health Behaviors	Health Outcomes
% 65 and older	Injury deaths	Adult smoking	HIV prevalence
% non-Hispanic White	Severe housing cost	Excessive drinking	COVID-19 death
% non-Hispanic Black	Long commute	Teen births	Premature death
% Asian raw value	Childcare centers	Motor vehicle crash deaths	Physical distress
% American Indian & Alaska Native	Firearm fatalities		
	Suicides		

Table 1: Variables that are commonly selected across models, with the direction of the coefficient estimates. (Red indicates positive association, blue indicates negative association, and gray indicates no significant association.)

Sensitivity Analysis

Variable	Coefficient	p-value
Physical inactivity	12.0451	< 0.01
Excessive drinking	12.0853	< 0.01
Some college	2.9830	0.01
Long commute driving alone	-3.6498	< 0.01
Diabetes prevalence	63.6848	< 0.01
Insufficient sleep	-22.4822	< 0.01
Traffic volume	-0.0024	0.03
Homeownership	4.6322	< 0.01
Population	-0.0002	< 0.01
65 and older	-6.3788	0.05
not proficient in English	33.1160	< 0.01

Table 2: Variables characterizing counties with NA death rate

- We study the effect of missing values by imputing the values 0, 5, and then 9 in three separate analyses (Figure 6).
- **As long as most counties has some non-zero value, excluding NA would still give a valid and working model.**

- Logistic regression is used to see which variables are significantly associated with whether the death rate is NA for a specific county (Table 2).
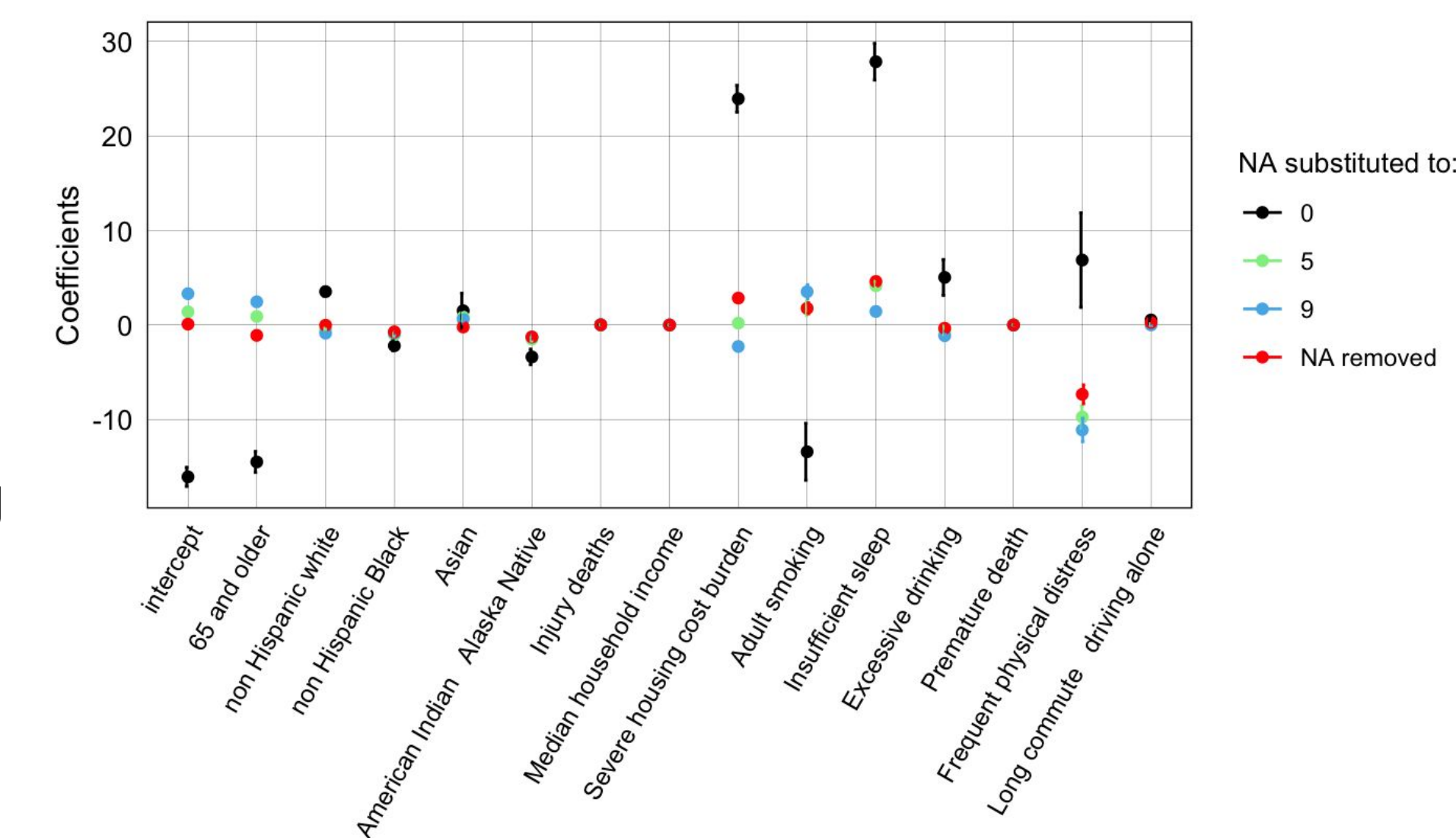
Figure 6: Comparison of linear regression coefficients (with error bars) taking in different substitutions of NA in the death count.
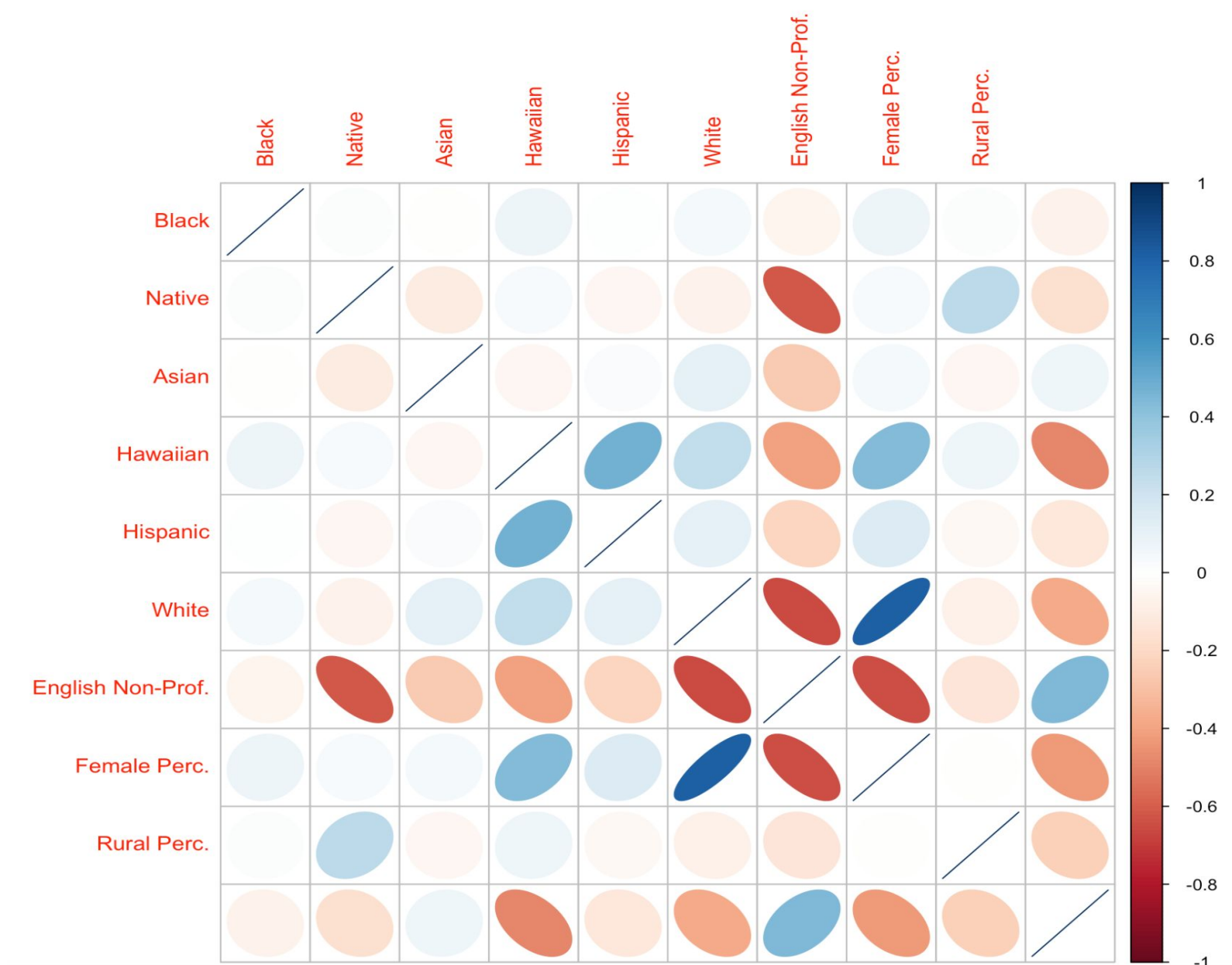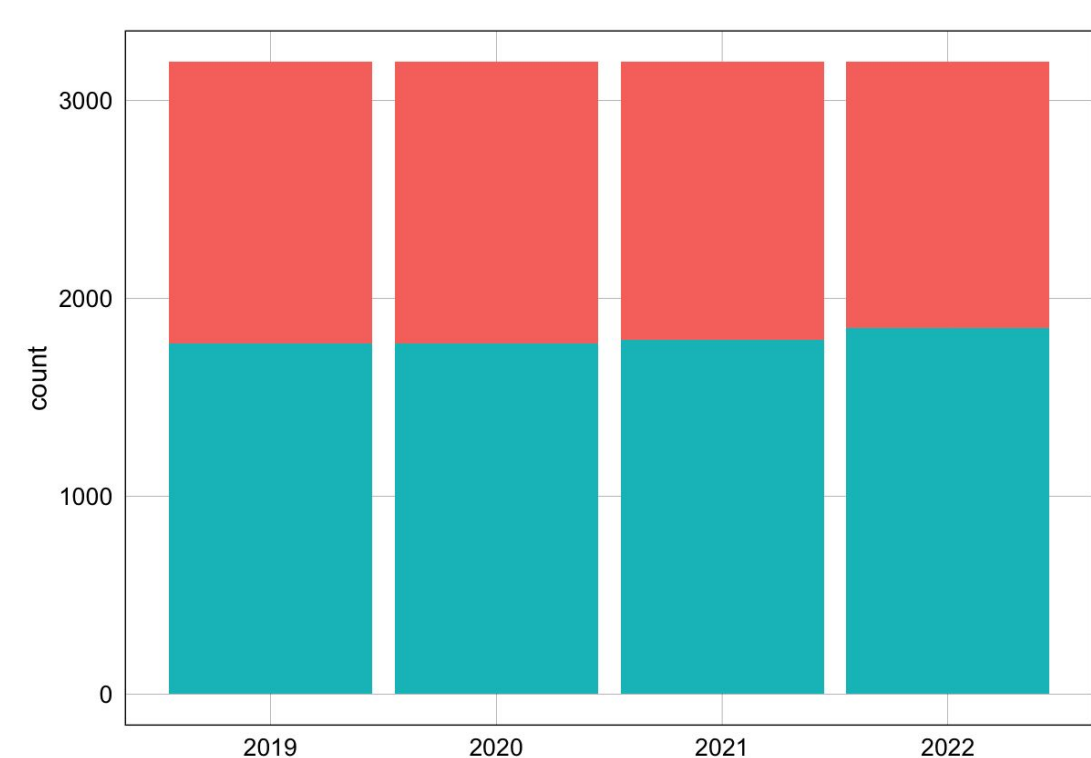
Conclusion

- 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.