Drug and Alcohol Related Deaths: 2018-2021 CDC and Census Data Cleaning, Analysis, and RShiny App Development

Caitlin Seibel, Emily Geis, Casey Schmidt

December 15, 2023

Introduction

Illicit substance use is known to have several health related consequences, including cardiovascular disease, cancer, and death [1]. Opioid misuse has risen in the United States from the first wave of prescription opioid overdoses in the 1990s to the rise of synthetic opioid use (i.e. fentanyl) in 2013 [2], calling for concern on (potentially) increasing deaths from drug or alcohol related causes. We analyzed data obtained from the Centers for Disease Control and Prevention (CDC) [3] and the U.S. Census Bureau [4] to make inferences on drug and alcohol related death counts, adjusting for age, gender, racial strata, and an indicator of the COVID-19 pandemic.

In this project, we are primarily determining whether the rates of drug and alcohol related deaths are significantly associated with age, gender, race, and/or the COVID-19 pandemic. To do this, we created a merged dataset that contains data from multiple sources, as we did not originally receive population counts from the CDC. We ran a generalized linear model using the negative binomial distribution to determine which of our available variables were significantly associated with specific causes of death. Finally, we created an RShiny app that provides users with an interactive way to investigate the relationships between different variables in our dataset. In the app, customizable negative binomial GLM's can be ran and data visualizations can be created.

Table 1 shows counts of demographic variables we have available in this data set, as well as the counts for the cause of death. Note that 65% of deaths are from non-drug and non-alcohol induced causes since these include all deaths unrelated to drugs or alcohol. Also note that the standard deviation of the deaths is about 3800 with an approximate mean of 957 deaths, indicating high variability in the deaths. Plotting two visualizations of this shows that there is high variability in the death counts.

Table 1: Descriptive statistics

	Overall
n	12064
Deaths (mean (SD))	956.53 (3799.72)
Cause_of_Death (%)	
All other non-drug and non-alcohol causes	7879 (65.3)
Alcohol-induced causes	1821 (15.1)
Drug-induced causes	2364 (19.6)
Age (%)	
Under 5 years	436 (3.6)
5-9 years	194 (1.6)
10-14 years	193 (1.6)
15-19 years	439 (3.6)
20-24 years	606 (5.0)
25-29 years	763 (6.3)
30-34 years	854 (7.1)
35-39 years	896 (7.4)
40-44 years	889 (7.4)
45-49 years	944 (7.8)
50-54 years	979 (8.1)
55-59 years	992 (8.2)
65-69 years	927 (7.7)
70-74 years	862 (7.1)
75-79 years	769 (6.4)
80-84 years	717 (5.9)
85 years and over	604 (5.0)
Race (%)	
White	3960 (32.8)
American Indian or Alaska Native	1577 (13.1)
Asian	1437 (11.9)
Black or African American	3025 (25.1)
More than one race	1397 (11.6)
Native Hawaiian or Other Pacific Islander	668 (5.5)
Gender = Female (%)	5580 (46.3)
COVID = 1 (%)	5864 (48.6)

Methods

Data Preprocessing

Deaths from drug or alcohol use is defined as a death where drugs or alcohol was the main catalyst in a series of events that led to death [2]. 2018-2021 death count data obtained from the CDC was requested based on five-year age categories, gender (male/female), six racial groups (American Indian or Alaskan Native, Asian, Black or African American, Native Hawaiian or Other Pacific Islander, White, and More Than One Race), cause of death (Drug, Alcohol, and Non-drug/non-alcohol), and month of death. Population data for these strata groupings were unavailable.

To make inferences about death-rate differences between substrata populations, we collected 2020-2022 data from the Census that used similar strata as the CDC. Two differences between the data were the characterizations of the five-year age groups and monthly/yearly population counts. The CDC characterized individuals who died before a year old as their own group; they also count all five year age groups up to 100 as unique groups. The Census collapses these into < 5 and > 85, with unique categories for five-year groups between these ages. CDC death count data was aggregated to match the Census strata. Further, Census data was only available on yearly 2020-2022 estimates, whereas we have monthly CDC data from 2018. We assumed the population is stable throughout a given year and that the rate of change within substrata is constant between years. By these assumptions, we estimated the substrata population counts for 2018 and 2019.

Figure 1: Distribution of Death Counts

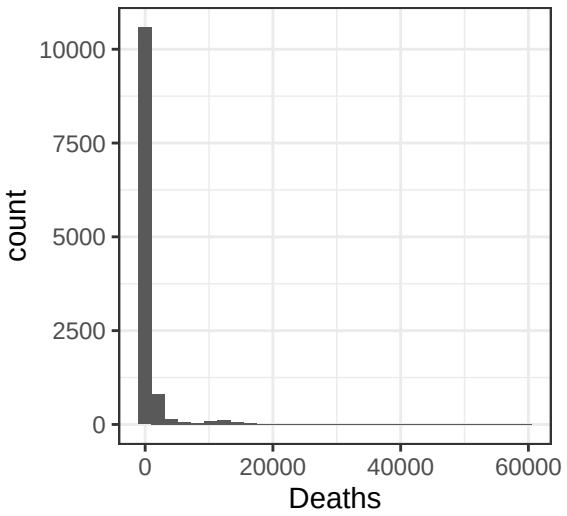
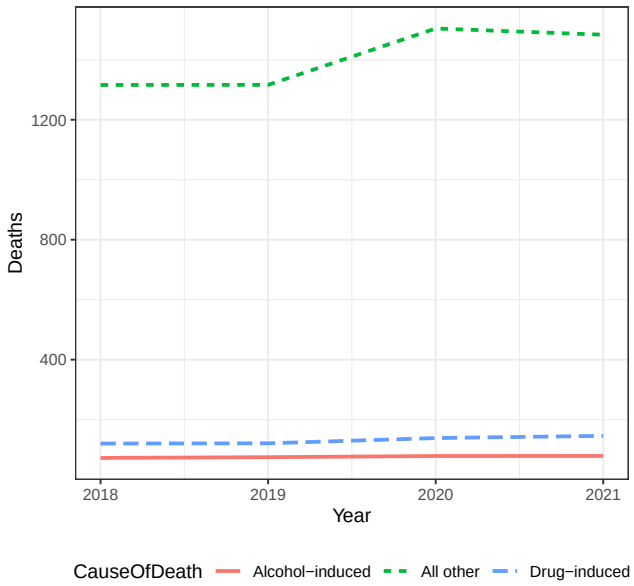


Figure 2: Average Death Counts over Time by Cause



Looking at the dates included in the model, we assumed that the COVID-19 pandemic will skew the death counts upwards. Plotting death counts over time, we see this assumption may hold, so we created an indicator to distinguish months before and after March 2020, which we are taking as the beginning of the COVID-19 pandemic in the U.S.

Model Building

To analyze our data, we choose to use a Generalized Linear Model (GLM). Our outcome variable of interest is number of deaths per month, referred to in our model simply as "Deaths". Our main predictors of interests are cause of death: drug related, alcohol related, or non-drug/alcohol related, and indicator of COVID-19 pandemic: 0 if before the start of the COVID-19 pandemic (defined here as March 2020), or 1 if after. Our other demographic covariates include age group, gender, and race. Our outcome is a count

variable so we initially assume a Poisson distribution. We use log(Population) as an offset for the model, since we want to model count rather than rate. Our initial GLM model with no interactions is as follows.

$$\hat{log(Deaths)} = \hat{\beta}_0 + log(Population) + \hat{\beta}_1 Cause + \hat{\beta}_2 Gender + \hat{\beta}_3 Age + \hat{\beta}_4 Race + \hat{\beta}_5 COVID \quad (1)$$

Since we are using a Poisson distribution, we need to check for overdispersion, or a violation of the assumption of equal mean and variance. To check for this violation, we will run our model and calculate the deviance divided by the residual degrees of freedom. If the equal mean and variance assumption is met, we expect this value to be approximately 1. From our model, we calculate this metric as 111.4576, which is much larger than 1, so we have evidence of overdispersion. To account for this overdispersion, we will use a negative binomial outcome for our GLM. After running the same model from above with a negative binomial outcome, the deviance divided by the residual degrees of freedom is now 1.083095, which is approximately equal to 1, so we have successfully corrected the overdispersion.

We also want to investigate the possible interaction between cause of death and the indicator of COVID-19 pandemic. Thus, we add an interaction term to our model, seen below.

$$\hat{log(Deaths)} = \hat{\beta}_0 + log(Population) + \hat{\beta}_1 Cause + \hat{\beta}_2 Gender + \hat{\beta}_3 Age + \hat{\beta}_4 Race + \hat{\beta}_5 COVID + \hat{\beta}_6 Cause \times COVID \quad (2)$$

We use a likelihood ratio test to compare this more complex model to our previous model. This LRT tests the hypothesis: $H_0 : \beta_6 = 0, H_A : \beta_6 \neq 0$. We are able to reject the null hypothesis (p-value $<$ 2.2e-16), and conclude that there is a significant interaction between cause of death and before vs after COVID-19 pandemic, and thus we prefer the more complex model.

Model Diagnostics

To check for multicollinearity, we used the variance inflation factor (VIF). If there is no multicollinearity, we expect the VIF to be between 1 and 5. We calculate the following VIF values for our model with interaction term. We have two values that are slightly greater than 5, but not enough to cause concern. We can conclude that there are no issues with multicollinearity in our final model.

Table 2: VIF for final model

Covariate	VIF
Cause of Death	5.330553
Gender	1.012963
Age	1.257762
Race	1.527936
COVID	1.525917
Cause of Death × COVID	5.302398

Results

Model Results

From Table 3, we can conclude that all of our predictors are significantly associated with deaths per month. For causes of death related to drugs or alcohol, the log expected counts of deaths per month are lower than for non-drugs/alcohol related death. This is likely due to the fact that non-drugs/alcohol related deaths include all other causes of death, so intuitively this result makes sense. There is a significant

Table 3: ANOVA results for final model

Covariate	p-value
Cause of Death	< 2e-16 *
Gender	< 2e-16 *
Age	< 2e-16 *
Race	< 2e-16 *
COVID	< 2e-16 *
Cause of Death × COVID	< 2e-16 *

interaction between cause of death and COVID-19 pandemic when predicting log expected counts of death per month. For all causes of death, the log expected counts of deaths per month increased after the start of the COVID-19 pandemic. We would expect this from non-drug/alcohol related deaths, since this category would include deaths due to COVID-19, but it is interesting to see that drug and alcohol related deaths also increase significantly after the start of the COVID-19 pandemic. Multiple demographic variables are also significantly associated with counts of deaths per month, which could lead to further research into possible interactions between demographic variables.

RShiny App

To create an interactive platform through which users can visualize different trends in the data, we created an RShiny app that offers a couple different features and methods that can be used to analyze our data. There are two overarching options in our app; regression model building and trend visualization. The main page, titled "Analysis", is where a regression model can be built and the second page, titled "Exploratory Plots", is where data trend visualization can be performed.

The main page of the app features a couple of drop down lists to select different aspects of a regression model. You can first select the desired outcome variable from a list of all available variables in the dataset, and then you can select all of the covariates you're interested in including in the model. There is also an option to include interaction terms in the model, but this is not necessary to run the regression. Once these variables are selected, the app runs a GLM in the background with a negative binomial distribution assumption and an offset of the log of the total population (see above for further model description). The first tab on this page shows the summary of the regression model, with the β estimates, p-values, and other relevant model estimates. The second tab on this page shows the residual plot resulting from the model, with the fitted values on the x-axis and the residual values on the y-axis. The final tab on this page shows the ANOVA output of the model, giving relevant statistics for all selected variables and interaction terms.

The second page provides the option to create exploratory plots to visualize trends in the data. There are two drop-down lists from which the y variable and the variable on which to stratify on can be selected. There are four choices in the drop-down menu for the y-variable. The options are the total population, the total death count, only the alcohol-related deaths, and only the drug-related deaths. If either the alcohol-related deaths or the drug-related deaths are chosen, the app works in the background to create a subset of the full data set that consists only of the relevant cause of death. The time variable on the x-axis is automatically set as "Years". After creating the plot, the trends in the selected y variable are visible and separated for each group in the selected stratification variable. The output is color-coded and a key is provided.

Discussion

Conclusions

We found that there are many significant predictors for death counts, including the cause of death, gender, race, age, COVID-19, and a cause of death and COVID-19 interaction term. After running a GLM with the

negative binomial distribution, all of these variables were found to be significantly associated with the death counts, indicating that they are indeed predictors for death counts. We were also able to successfully create an RShiny app that allows users to create specialized GLM's and data visualizations.

Limitations

The CDC only allows for 5 grouping variables to be entered for data extraction, meaning we had to determine which four variables to stratify on to get drug/alcohol death counts. Because of this, we are not able to adjust for other potential confounding factors such as the state of residence. We were also limited to only choosing variables that were available through the CDC's website, so it is possible that there are other confounders that weren't even measured, like the estimated proportions of drug or alcohol users within each strata.

We also made assumptions about the populations in order to estimate the data due to the limitations of what data was available to us. This may bias our answers slightly, since the quantity used to estimate the proportion of deaths in a population is estimated. Since these populations are very large, we don't expect this to heavily impact our substantive conclusions.

Finally, our RShiny app is still limited in its functions, as there are still a few aspects that prevent full control over the models and visualizations. When creating a model, it is only possible to include one interaction term, as when more than two variables are selected from the drop down list, it creates a triple interaction term where all three variables are interacting. This limits the number of interaction terms that can be included in the model at the same time. Also when creating a model, it is not possible to select which levels of each categorical variable to treat as the reference group. The reference group is currently hard-coded into the app for each categorical variable, but it would provide even more flexibility and control if it were possible for the user to select which level to use as reference. Finally, the model-building portion of our app is hard-coded to be a negative binomial GLM with the log of the population as the offset, so it is not possible to run the regression with some of the variables listed as options for the outcome variable. It would provide more flexibility and control if the type of GLM were able to be selected, but this poses challenges as well. If we were to allow the distribution to be selected, we would be assuming that all of our app users have a strong understanding of GLM's which won't always be the case.

Author Contributions

Emily: created RShiny app in r, wrote rshiny app section of paper, formatted paper in rmarkdown, portion of limitations

Casey: Found and extracted data sets, introduction, data preprocessing, and limitations

Caitlin: R coding for model building and diagnostics, writing model building section of report

Github Link

https://github.com/caschmi/625-final-project.git

References

[1] National Institute on Drug Abuse. What are the other health consequences of drug addiction? 2020, URL: https://nida.nih.gov/publications/drugs-brains-behavior-science-addiction/addiction-health

[2] Centers for Disease Control and Prevention [CDC]. Understanding the Opioid Epidemic. 2023. URL: https://www.cdc.gov/opioids/basics/epidemic.html

[3] Centers for Disease Control and Prevention [CDC], Alcohol Use. 2023. URL: https://www.cdc.gov/nchs/fastats/alcohol.htm

[4] United States Census Bureau [USCB]. National Population by Characteristics 2020-2022. 2023. https://www.census.gov/data/tables/time-series/demo/popest/2020s-national-detail.html