Kathleen Capella
Jared Diesslin
Dylan Greene

Relationship between Common Health Factors & Lifestyle Choices and Drug Overdose Mortality Rates on a County Level

1. Introduction

In 2017, the Centers for Disease Control and Prevention reported 70,237 deaths due to drug overdoses in the United States. This number is evidence of a growing trend of overdose deaths over the last two decades. Since 1999, the national age-adjusted rate of drug overdose deaths has been on an overall upwards trend in America, with increasing by an average of 10% per year from 1999 through 2006, 3% per year from 2006 through 2014, and 16% per year from 2014 through 2017 [1]. As these statistics indicate, drug overdose mortality has become a serious concern facing law enforcement, public health officials, and legislators today.

While this problem presents many issues on a national scale, the implementation of solutions often occurs on a more localized level. In a document attempting to present evidence-based strategies for combating the drug overdose crisis, the CDC cites examples of local efforts such as syringe services, target naloxone distributions, and medication-assisted treatment programs as models for other communities to follow [2]. Due to the complexity of the issue at hand, including many biological, psychological and social factors, it is increasingly important for research to be conducted with the goal of placing information in the hands of those who can make the most impact. For this reason, our project decided to examine the drug overdose mortality crisis on a county level. Our goal is to identify features on a county level that have a statistical significance in modeling drug overdose mortality rates. Specifically, we would like to determine whether population health factors such as household income, high school graduation rates, unemployment, HIV prevalence, poor mental health days, and excessive drinking, impact the model.

2. Related Work

One similar study observed drug overdose deaths in New Mexico from 1990-2005. In this study, researchers aimed “to determine the contribution of heroin, prescription opioids, cocaine and alcohol/drug combinations to the total overdose death rate and identify changes in drug overdose patterns among New Mexico subpopulations.” [3]
In addition to exploring demographic differences, this study also tried to see which type of drug contributed the most to the total death rate, whereas ours looks at the drug overdose mortality rate in as a general metric. While ours looked across the country, theirs adjusted for region. We specifically wanted to look at the differences at a granular geographic level between counties, so we looked at this as a targeted metric rather than adjusting for it. In terms of models, the report mentions the use of a log-linear regression model, while we decided to use separate logistic and linear models.

A separate study from Marion County in Indiana was similar to ours in that it looked at drug overdose mortality on a county level (albeit only one county as opposed to the national reach of our study). In this study, researchers also looked into certain demographic features and found, for the people for whom they had the appropriate information available, that those who had fatally overdosed were more likely to be unemployed and to have not completed a high school education [4]. In terms of methods, this study was more factual and did not involve any model building.

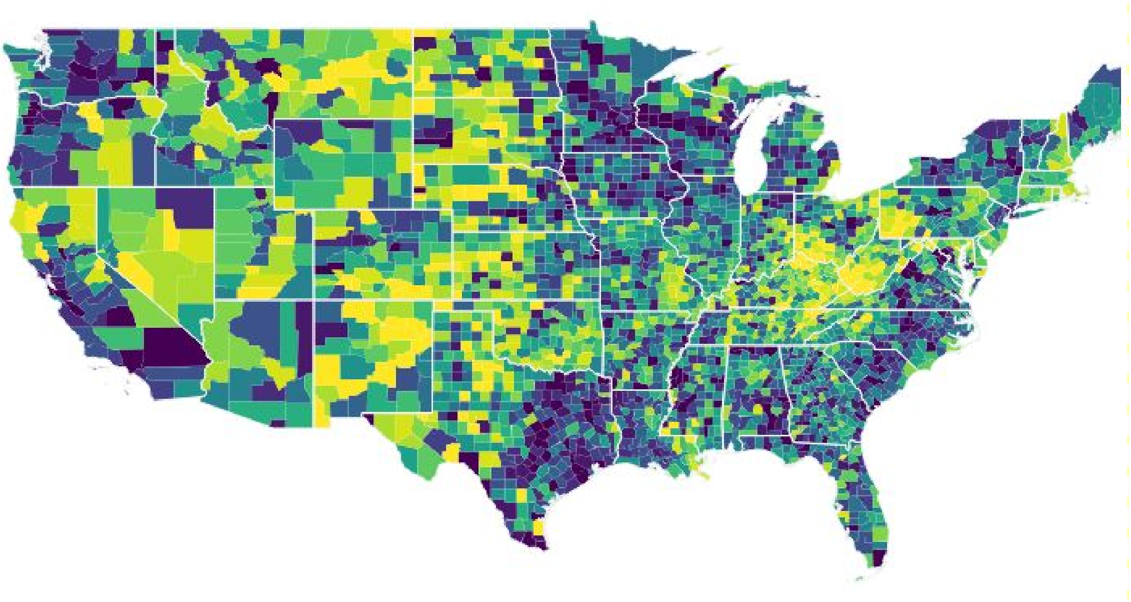
An article from the National Academy for State Health Policy demonstrates that there has been some interest expressed in looking at the drug overdose crisis within the lense of urban vs. rural, even making the claim that “death from opioid overdose is 45 percent higher in rural regions than urban areas.”[5]. This article points out that rural areas may need special attention in addressing the drug overdose crisis. This is part of the reason why we had decided to explore whether rural vs. urban made a difference to our model.

3. Solution/Method

We used a few different tools and techniques to try and find significant outcomes for our study. One of these methods was Select K-best, which was able to take the dataframe from all of the counties and find the 20 most significant features in the dataframe. We then used these 20 features to try and find whether there was a difference in drug overdose mortality rate among the lower 50% and upper 50% of each feature. This was done by sorting each feature individually, finding the mean, splitting the feature data at that point, and taking the mean of the lower 50% and the mean of the upper 50% and comparing them.

Another method which we used in our experiment was logistic regression. This was accomplished by finding a binary outcome for both the top 25% and the bottom 25% of drug overdose mortality rate.

Along with these traditional methods of analysis, one innovative method that we used was trying to predict drug overdose mortality rates for the counties which did not have drug overdose mortality rates in their dataset. We were able to do this for the missing counties and show it on a heat graph, which shows which parts of the country have higher drug overdose mortality rates than others. Here is the resulting graph (yellow means higher drug overdose mortality rates):



4. Data and Experiments

In order to model and explore drug overdose mortality rates and significant factors, a large amount of data is needed. Specifically, modeling the effects of population health data required many data points in order to have a robust sample that could be used to infer generalizable information. While state data would have been easier to obtain and work with, more data points were desired to provide a better chance of identifying both geospatial and population health data trends. After some preliminary data discovery, county level data presented itself as the most fine-grained yet complete data available.

In particular, data from County Health Rankings[6] provided a robust feature set, including the desired target of Drug Overdose Mortality Rates, on a county level. The data from this site was collected using a shell script which requested the multi-sheet XLS file for each state. Once each file was retrieved, substantial pre-processing and data cleaning was required to convert the files into a workable format. These steps included, but were not limited to: conversion of each sheet to a csv and header formatting using a Python script, combining data from all states, checking for and removing features exhibiting multicollinearity, and dealing with missing data via imputation or removal. The cleaned data was then saved in pickled DataFrames for use in modelling and experimentation.

Following the completion of data acquisition and cleaning, experimentation was performed. Several experiments were used to evaluate feature significance in modeling drug overdose rates. The primary method used to perform this evaluation was to train linear regression models using subsets of the entire feature space. Once trained, model summary statistics were evaluated. Specifically, the feature coefficients for the linear model were viewed in conjunction with the confidence intervals for those coefficients. Using that information, it was easy to determine the significance (or lack of significance) of the feature by ensuring the confidence interval for the coefficient was either entirely positive or entirely negative. That is to say that at the given confidence, the model was able to confidently calculate of the direction of influence of the feature.

This method of examining feature significance should reliably determine if the feature provides statistically significant value. However, one area where reliability may have been compromised in this study was the impracticality of checking LINE assumptions for the linear model. Specifically, the large feature space made it prohibitive to check that the relationship is linear, that the errors have the same variance, that the errors are independent, and that the errors are normally distributed for each and feature that was being considered. Without explicitly checking each of these assumptions for every feature used in the linear models, it is likely that one or more of these assumptions was violated for at least one feature. The result of this may have been missing a pattern in the data that could have resulted in a model with a significantly better fit to the data. As such, given more time and resources, a follow up study in which these assumptions are confirmed would be valuable and an important next step.

Another form of experimentation was strictly exploratory. Specifically, several attempts to manually identify trends in the data were performed. These included generating visual representations of the data and summary statistics for certain features. The reliability of these exploratory experiments is purely subjective. However, they do provide valuable information and graphics which can be used to educate and provide a better intuitive understanding of the data for future modeling efforts.

5. Evaluation and Results

The primary goal of experimentation was to identify which population features are significant in modeling drug overdose mortality rates and to understand how those features vary with that target. Going into the experimentation, there were several features which were the focus. However, upon evaluating the statistical significance many of these features it was clear that many were not significant. Thus, the goal of understanding how those features vary with target was irrelevant. Perhaps the best example of this was the consideration of rural versus urban counties. Originally one of the main research interests, the rural versus urban distinction was observed to provide no statistically significant value. In totality, the features which were originally intended to be examined demonstrated little predictive power and significance; the linear model trained on those features had a terrible R-squared value of 0.169. This observation resulted in a pivot in the direction of experimentation.

Specifically, experimentation was pivoted to using and examining other features which had originally been intended to be left out. In the previous attempt, intuition was proven to be a poor selector of features to examine. As such, it was decided to attempt a computational approach to feature selection. After some initial research, it was decided to analyze the twenty best features based on the highest univariate correlations between each regressor and the target. This was accomplished using a feature selection procedure available in SciKit Learn and a scoring function which tested the individual effect of each of many regressors by calculating an F-score for each regressor. This method provided features which were then used for the remainder of experimentation.

With the newly obtained feature vector, three different models were trained and evaluated. First, an Ordinary Least Squares model was trained. This new linear model vastly outperformed the previous; it had a modest, but greatly improved, R-square value of 0.574. This showed there was indeed some valuable information in the features. This was confirmed when the confidence intervals of the coefficients for each feature were examined as they overwhelmingly provided statistically significant effects to the model.

In an effort to reduce the noise and improve modeling, a Logistic Regression model was also created which only considered the lower and upper quartiles of the target. This method actually performed quite well (pseudo R-squared = 0.461). The complete list of significant variables included the following attributes:

Age_Adjusted_Mortality
Child_Mortality_Rate
Child_Mortality_Rate_Black
Child_Mortality_Rate_White
Infant_Mortality_Rate
Percent_Frequent_Physical_Distress
Percent_Uninsured_1
Segregation_index
Years_of_Potential_Life_Lost_Rate
Years_of_Potential_Life_Lost_Rate_Black
Mentally_Unhealthy_Days
Teen_Birth_Rate_White
Percent_Some_College
Household_Income
Percent_Rural
HIV_Prevalence_Rate
Percent_Frequent_Mental_Distress

Of the insignificant features, we were especially surprised to see that Graduation Rate was not significant, as this was something that other studies mentioned [7]. It is possible that since we did not have a specific way for controlling for age that this may have impacted Graduation Rate’s significance. However, we were interested to find that Percent Rural and Household Income were significant, although the odds ratio for each of these were close to 1, not indicating a very large difference from the base.

Another important area of exploration was analyzing the actual values of the significant features in order to understand the relationships. Most of the results of the lower 50% vs. upper 50% were not surprising for most of the features, but the results for Percent Uninsured were more surprising. The values for the lower 50% vs. upper 50% were almost identical, which we would have expected to be more different similar to the results of the rest of the features.

In addition to looking at the values split by percentile, it was decided that visualization would be a great way to convey the relationships. In order to have the greatest degree of visual separation, the data was again grouped by the lowest and highest drug overdose rate quartiles. Then, histograms were created which binned each feature into deciles. For the most statistically significant features the differences between lower and upper quartile histograms were striking. It was incredibly easy to observe the shift in the distributions of deciles in features such as Percent Unemployed. This was interesting to see in light of other studies that point to the link between unemployment and the drug overdose crisis[8].

Finally, it was decided to provide an interactive map of both the known overdose mortality rates and another of predicted rates. These maps can be used to get a visual understanding of clusters of high overdose rates. Ideally, with more time and resources these would also be used to further investigate the nationwide drug overdose epidemic. However, as it stands, they simply provide an interesting visual tool that demonstrates just how serious the drug abuse problem is in many counties nationwide and reaffirms the importance of further studies on the topic.

6. Works Cited

[1] Hedegaard, H., M.D., Minino, A. M., M.P.H, & Warner, M., Ph.D. (2018, November). Drug Overdose Deaths in the United States, 1999–2017. Retrieved December 11, 2018, from https://www.cdc.gov/nchs/data/databriefs/db329-h.pdf

[2] Carroll, J. J., PhD, MPH, Green, T. C., PhD, MSc, & Noonan, R. K., PhD. (2018). Evidence-Based Strategies for Preventing Opioid Overdose: What's Working in the United States. Retrieved from https://www.cdc.gov/drugoverdose/pdf/pubs/2018-evidence-based-strategies.pdf

[3] Shah, N. G., Lathrop, S. L., Reichard, R. R., & Landen, M. G. (2007, November 20). Unintentional drug overdose death trends in New Mexico, USA, 1990–2005: Combinations of heroin, cocaine, prescription opioids and alcohol. Retrieved from https://onlinelibrary.wiley.com/doi/full/10.1111/j.1360-0443.2007.02054.x

[4] Watson, D., PhD, Duwve, J., MD, MPH, Greene, M., PhD, Weathers, T., MPH, Huynh, P., MPH, & Nannery, R., MA. (2018, October). The changing landscape of the opioid epidemic in marion county and evidence for action.
Retrieved from https://www.rmff.org/wp-content/uploads/2018/06/Richard-M.-Fairbanks-Opioid-Report-October-2018.pdf

[5] Corso, C., & Townley, C. (2017, February 14). Intervention, Treatment, and Prevention Strategies to Address Opioid Use Disorders in Rural Areas. Retrieved from https://nashp.org/intervention-treatment-and-prevention-strategies-to-address-opioid-use-disorders-in-rural-areas/

[6] County Health Rankings. (n.d.). Retrieved from http://www.countyhealthrankings.org/

[7] Scommegna, P. (n.d.). Opioid Overdose Epidemic Hits Hardest for The Least Educated. Retrieved from https://www.prb.org/people-and-places-hardest-hit-by-the-drug-overdose-epidemic/

[8] Hollingsworth, Alex, Ruhm, J, C., Simon, & Kosali. (2017, February 23). Macroeconomic Conditions and Opioid Abuse. Retrieved from https://www.nber.org/papers/w23192