0.1 Executive Summary

The Opioid market size in the United States is valued at over 11 billion dollars with more than 150 million prescriptions written every year since 2015—a rate that would be over 50 prescriptions per 100 people. In 2018, more than 40,000 Americans died from opioid overdose; the year before, the United States Department of Health and Human Services (HHS) declared a public health emergency for the US Opioid Epidemic. Which demographics were hit the hardest? Could a model be made to accurately predict future death rates? Was the declaration of the opioid epidemic due to an unforecastable increase in opioid related overdose deaths? In this analysis, the data was collected from CDC's Wonder tool where it contains mortality and population count data from 1999 to 2018. Our analysis found that middle-aged African American and White males were hit the hardest and that a Bayesian multiple linear regression can be fit on the data with various degrees of prediction accuracy based on training data time span.

0.2 Motivation

We want to explore CDC's Wonder data to see which demographics have been impacted the most by the opioid epidemic. We also propose the overall crude rate (mortality rate) can be modeled using a Bayesian multiple linear regression, while comparing its predictive qualities to a regular frequentist multiple linear regression.

0.3 Data and EDA

The dataset used—which was downloaded from CDC's Wonder tool—contains 1382 observations (each observation is a specific cohort eg. African American, male, 20-25 years old) for a given year, along with its respective data like number of deaths, year, population and crude rate—the dependent variable in our analysis. Because deaths and population are directly calculated into the crude rate, they will not be included in the analysis, leaving only the variables Year, Gender, Race, Age, and Crude Rate; for this analysis, we also focused on data from ages 20 to 64.

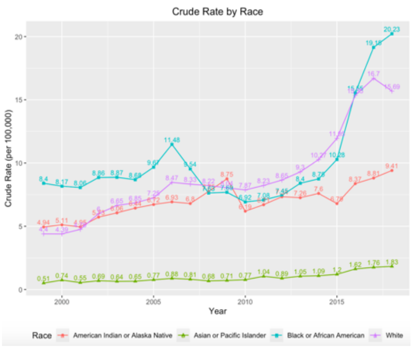


Figure 1: Crude rate of race by years

White Americans, both male and female had the highest number of deaths peaking in 2017 with 24,901 deaths, but adjusted for population, African Americans had the highest crude rate of 20.23 in 2018, compared to White American's 15.69 (Fig. 1). Males were also more likely to die from opioids as their crude rates were almost three times higher than female's: 23.44 for males and 8.4 for females in 2017. An interesting finding (Fig. 2) was that crude rates decreased during the 2008 Great Recession, dipping 2010, before exploding in 2013. This period also marked the beginning of the age group change where the highest crude rates from 2000 to 2010 were the 40-44 and 45-49 age groups before shifting to a younger demographic of 25-39 year-olds.

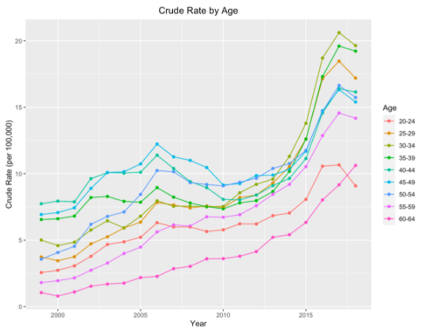


Figure 2: Crude rate of age group by years

0.4 Models

An initial frequentist multiple linear regression model was created to model CR (crude rate) using 10 years' data. The model with the best AIC was then used to predict the 2017 crude rate, and its respective confidence interval given 2017's data was (7.62, 9.86). Then JAGS was used to build the Bayesian linear regression model. Because we had not much prior domain knowledge in epidemiology and the opioid epidemic, weakly informative priors were chosen. The covariates coefficients' betas were given a prior of Normal(0, 1) and the variance was given a prior of Inverse-Gamma(1, 1). Model:

$$Y_i^* \mid \beta 0, \beta 1 \ldots \beta 15, x_i, \sigma \sim N(\beta 0 + \beta 1 x_{i,}\ \mathit{female} + \beta 2 x_{i,}\ \mathit{male} + \beta 3 x_{i,}\ \mathit{Native\ American} + \ldots. + \beta 7 x_{i,}\ \mathit{Age\ 20\text{-}24} + \ldots. + \beta 15 x_{i,}\ \mathit{Age\ 60\text{-}64}, \sigma)$$

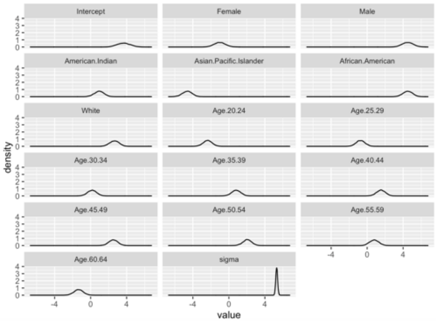


Figure 3: Distribution plot of posteriors

For sampling dependent samples from the posterior distribution of the parameters (Gender, Race, Age group), Markov Chain Monte Carlo Gibbs Sampling (MCMC) was used. Each parameter was sampled 5000 times along with a burn-in of 5000 samples; MCMC convergence was achieved for each parameter (Appx. 1). This produced sample distributions for each parameter (Fig. 3). The coefficients aligned with the EDA with males and African

Americans having a higher coefficient. Age group 25-29 had a surprisingly high coefficient though, and variance was quite high.

Next, we sought to create the best predictive Bayesian model with Bayesian Model Averaging using the BAS package. Using 10-year training data, whether we used highest probability model, median probability model, or best predictive model, the chosen parameters were the same: Year, Gender, Race, and age groups 30-34, 35-39, 40-44, 45-49, 50-54, and 55-59 (Fig. 4).

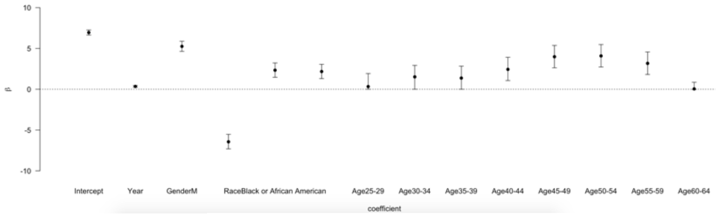


Figure 4: Coefficients for betas using BMA

The true crude rate CI for 2017 was (9.10, 14.83). The respective frequentist CI was (7.62, 9.86) while the BMA Bayesian regression's CI for 2017 was (7.64, 9.85)—not very different. This could be due to a relatively weak informative prior on top of larger sample size so priors have less weight. Using a 5-year training dataset (2012-2016) and 3-year training dataset (2014-2016), the respective predicted crude rate credible intervals are (9.35, 11.80) and (10.41, 13.19). Even though 3-year training dataset's model gave the best credible interval, this does not mean it is the best at forecasting future data; it only means that it captured the latest trend of the increasing crude rate compared to lower crude rates of a decade ago.

0.5 Conclusion

Given the results, we can conclude that White and African American middle-aged males are being impacted the most by the opioid crisis and that Bayesian linear regression is an option to model crude rate. While the Bayesian linear regression did not produce a closer predicted 2017 crude rate credible interval compared to frequentist linear regression, it does allow for setting a prior which is useful for smaller datasets and if there's prior domain knowledge, and Bayesian models give a distribution for each parameter while frequentist linear regression produces a point estimate for betas; future analysis could explore dynamic linear regressions to better fit the trend.

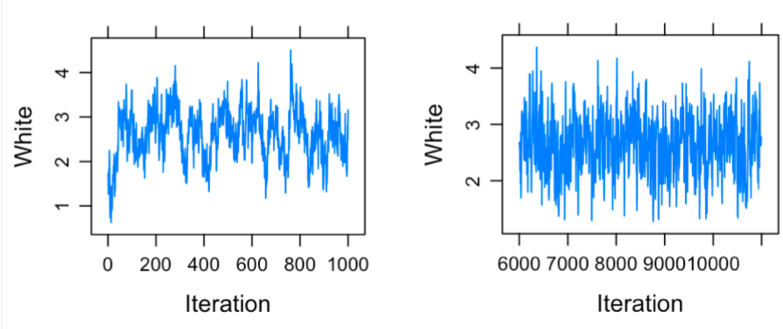
Because CDC's Wonder tool limits each data pull to 5 parameters, we were limited on what independent variables we could use. Future analysis could find a way to combine multiple datasets and data sources including stock prices of pharmaceutical companies that produce opioids along with the number of opioid prescriptions of each year. There is also a demographic change of older people having the highest crude rate to a much younger demographic after the 2008 Great Recession. From the crude rate trend along the years, we could see that there was a sharp increase in 2015 before decreasing in 2018. It would be interesting to see why that decrease happened along with what crude rates are in 2019 and 2020, especially with COVID-19 and the overburdening of the US medical system. Social quarantine could limit exposure to opioid sources which may lead to lower crude rates, but overcrowding of many hospitals could prevent people from seeking proper medical attention for opioid abuse.

0.6 References

Data: https://wonder.cdc.gov/ucd-icd10.html
https://cran.r-project.org/web/packages/BAS/vignettes/BAS-vignette.html
https://www.hhs.gov/opioids/about-the-epidemic/index.html
https://www.cdc.gov/injury/features/prescription-drug-overdose/index.html
https://www.grandviewresearch.com/industry-analysis/opioids-market

0.7 Appendix

MCMC convergence



Appendix 1: Traceplot of MCMC sampler for White covariate. Left is without burn-in and right is burn-in of 5000 samples