Team MB, MF, (PM)^-1
Final Write Up

The COVID-19 pandemic has left no stone unturned in dramatically changing the way in which we conduct our daily lives. Although leaving behind a wreckage of human casualties and exposing weaknesses in healthcare systems, this pandemic has shown that there are far reaching differences in our society. This team's analysis aimed to address the question of how the COVID-19 pandemic disproportionately affects poor communities and communities of color in the United States.

For this analysis, data was gathered in 3 different areas: demographics, economics, and pandemic health outcomes. The COVID-19 data used in this analysis was collected from the NYTimes' Github on April 15, so our analysis focuses on cases up to and on that date. Demographic data was gathered from the Census.gov website and is thus data reflecting demographic trends from the 2010 Census. Poverty data came from the Small Area Income and Poverty Estimates provided by the Census.gov website; this data is from the 2018 estimates. Further, population data from the R 'usmap' package was used for plotting purposes.

We used a multiple linear regression to model the deaths per capita in each county dependent on different characteristics of the county, mostly pertaining to the demographic or economic make-up of the county. We set the log deaths per capita in a county as our response variable and chose the initial explanatory variables so as to reduce multicollinearity in our model but still retain the majority of the data that we had gathered. We then performed a stepwise backwards selection in order to choose the most parsimonious and efficient features to predict the log number of deaths per county. We then added interaction effects for our continuous predictor variables and repeated this process. Afterwards, we checked for the normality, constant variance, independence, and linearity of our model.

Most interestingly, and in accordance with our research question, we found that the percentage of black individuals in a county is quite highly correlated with deaths per capita. Another interesting finding that we discovered was that there was a high correlation coefficient between the number of log cases in a county and the percent of county residents who are in poverty. This indicates that as the number of cases multiply in a county, the deaths per capita in impoverished counties increase at a rate far larger than the deaths per capita of wealthier counties. This again indicates that the coronavirus has more significant and substantial effects on communities of color and impoverished communities in America.

There are a few limitations in our modelling that we would like to address. First, as mentioned earlier, we do not believe that the number of deaths per capita in counties are independent data points. We discuss this further in our discussion of independence, but we believe that the number of deaths per capita are positively correlated with geographically neighboring counties which may have an effect on our modelling.

Next, our sample sizes are still small. Despite the fact that this virus has become a global pandemic, many states are at different stages of the pandemic right now and the pandemic could affect communities differently at different stages in its life cycle. Additionally, reporting seems to be a large issue. Many states are reporting deaths at different rates and have different practices for reporting deaths which can add variance to the data.