Prediction of COVID-19 death rate and effect of policy responses in California

Lin Wang

Introduction

COVID-19 is a highly contagious disease which affected most states and counties in the United States. At the early stage of this pandemic, New York was worst hit. The residents living in areas with high population density, such as New York – which is a metropolitan city with large population and high population density – may have more chance to get into close contact with others through public transportation, gatherings and result in a surge of the infection.

Some countries have been very successful in preventing the spread of the disease, some might have been worse hit than others. Differences in government responses could explain part of this. In United States, six counties in Bay Area became the first in the nation to announce shelter-in-place orders on March 16, and on March19, California became the first to mandate a state-wide order. Since then, additional states across the country began implementing state-wide restrictions. To understand whether the government policy is effective against the outbreak, it is essential that we look at the timing and strictness of responses.

This project is trying to evaluate 1) Association between COVID death rate and population density in United States, 2) Effect of policy responses to COVID pandemic in California.

Methods

1. Data collection

The data was obtained from multiple source:

   a. National history of COVID-19 data – COVID19Tracking (https://covidtracking.com/data/download)
   b. State-level COVID-19 history data & 2018 state population data – New York Times(https://github.com/nytimes/covid-19-data)
   c. Lockdown data – Aura Vision's Lockdown Tracker (https://auravision.ai/covid19-lockdown-tracker)

2. Data cleaning and wrangling

`data.table` was used to download and read in large data sets. EDA checklist was done to check dimensions, headers, footers, variable names, and types. New variables and some outliers were created and corrected using `data.table`, `tidyverse` and `dplyr`, date variable was re-formatted into correct format by `lubridate`, `NA` in death and cases columns were kept because we would have missing values at the start of the outbreak. After merging from raw data, it was ready to do further analysis and visualization.

3. Data visualization

The tools used to visualize data are mainly from `ggplot`, `plotly`.

   a. Death rate prediction

   During this ongoing epidemic, some of the active cases already detected may subsequently die, leading to underestimation of CFR estimated before their death. To

mitigate the bias due to delays to case resolution during an ongoing outbreak, a naïve CFR and IFR was calculated by the following formula using national history data:

$$CFR, in\ \% = \frac{Number\ of\ deaths\ from\ disease}{Number\ of\ death\ from\ disease + Number\ of\ recovered\ from\ disease} \times 100,$$

However, in this case, CFR may be overestimated if people die quicker than they recover.

$$IFR, in\ \% = \frac{Number\ of\ deaths\ from\ disease}{Number\ of\ infected\ individuals} \times 100.$$

Then a hypothesis test of COVID death rate of most recent day was generated to see death rate across the nation.

A linear regression model was built to predict the association between death rate and population density, the model includes population density, population, cases, deaths, new cases, and new deaths.

b. Effect of policy responses

Policy responses were categorized into four types based on date variable: loose, first order, statewide order, and limited lockdown.

Results

Prediction of COVID-19 death rate

Figure 1 Overview of COVID 19 CFR and IFR in US

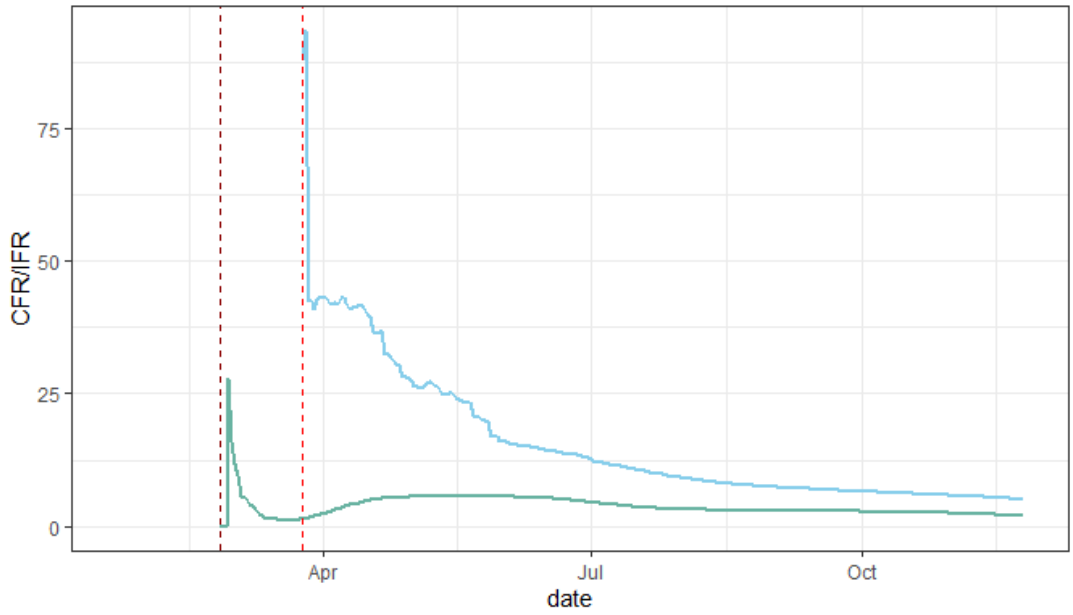


Figure 1 tells the developing trends of COVID 19 CFR and IFR in US. The green line is infection fatality ratio (IFR), which estimates the proportion of deaths among all infected individuals. The blue line is case fatality ratio (CFR), which estimates the proportion of deaths among identified confirmed cases. The first vertical line indicates the date of first death, and the second line is the date when recovered cases were observed, so that is why the CFR is approaching very large at the beginning and became stable gradually.

Figure 2 Hypothesis test of COVID death rate

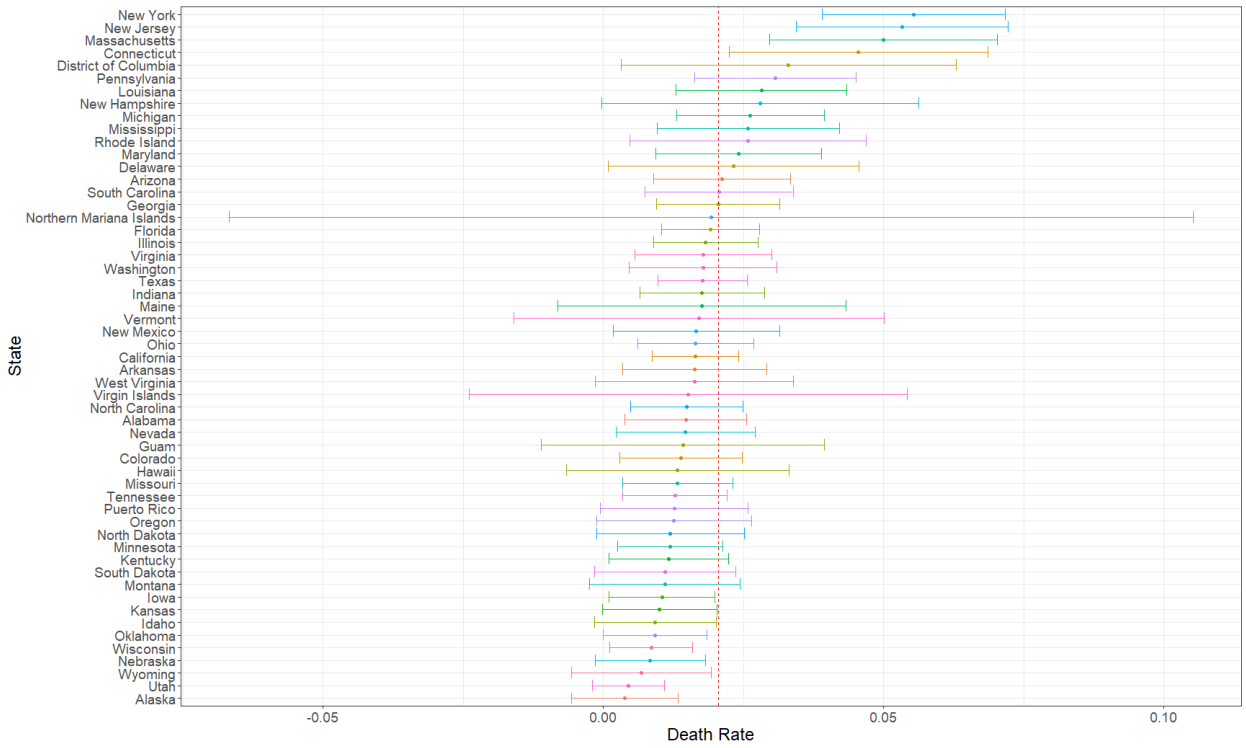
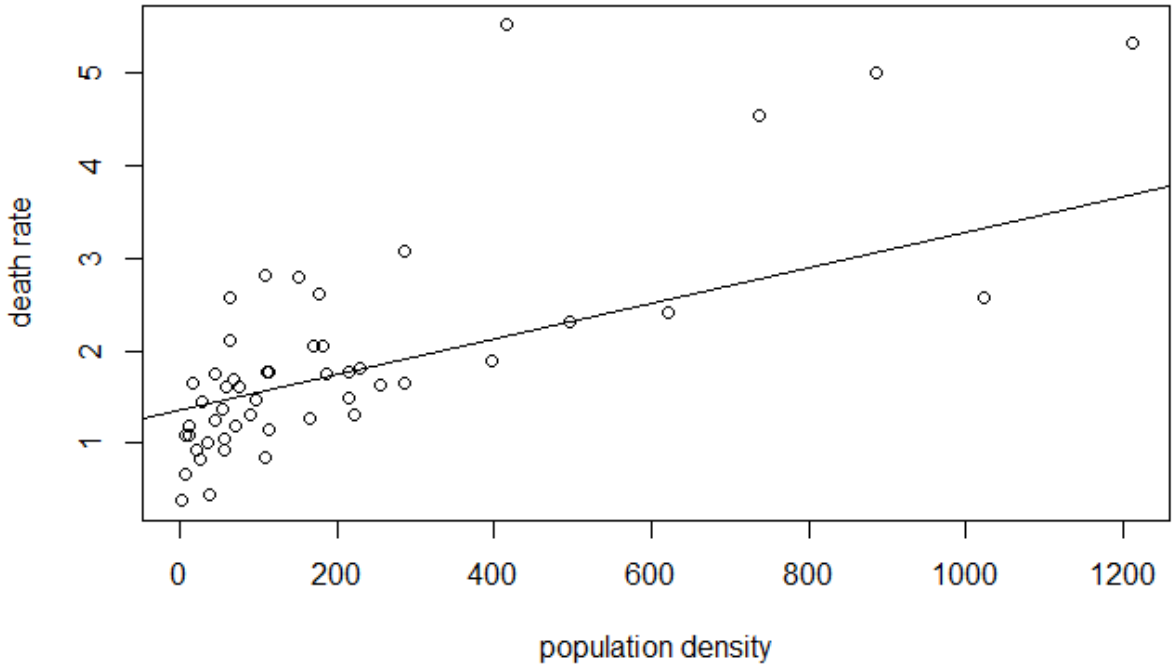


Figure 2 presented a hypothesis test of COVID death rate across the United States, most states had a death rate below the mean. States had a large confidence interval probably have a few cases. And New York and New Jersey were worst hit from this plot.

Table 1 Summary of linear regression model results

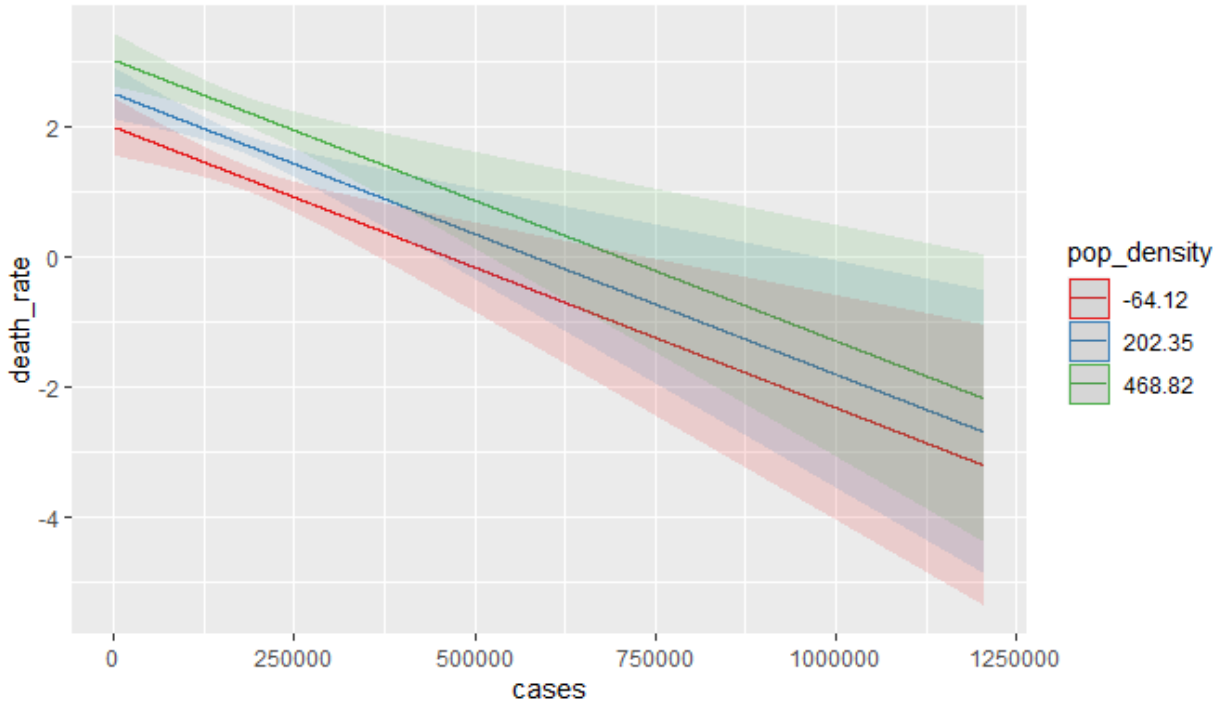
	death_rate		
Predictors	*Estimates*	*Statistic*	*p*
(Intercept)	1.35	12.53	**<0.001**
pop_density	0.00	6.61	**<0.001**
deaths	0.00	8.17	**<0.001**
population	0.00	0.84	0.405
cases	-0.00	-4.01	**<0.001**
newcases	0.00	0.17	0.866
newdeaths	0.00	0.94	0.351
Observations	50		
R^2 / R^2 adjusted	0.853 / 0.833		

Figure 3 COVID-19 death rate vs. population density



Death rate has a significant association with population density, deaths, cases ($p < 0.001$) in the linear regression model and is not significantly associated with population ($p = 0.405$), new cases ($p = 0.866$) and new deaths ($p = 0.351$) (Table 1, Figure 3).

Figure 4 Predicted values of death rate vs. cases



Death rate is high in the region with high population density, but it has a trend of decreasing with more cases (Figure 4).

Effect of policy responses in California

On March 19, California mandate a state-wide safe-at-home order, and on May 25, the order ends and many communities are into reopening plans. New cases were into a surge under a loose policy. Therefore, on June 18, Governor Newsom ordered all Californians to wear face coverings in public places. On July 13, statewide closures were announced. All bars and indoor dining at restaurants must close completely. It seems this restriction achieved effective results at that stage with daily new cases decreasing. However, as winter coming and counties move towards easing restrictions, a new surge is coming, the latest daily new cases has reached at an unprecedented peak. California government reinforced the importance of staying at home and announced a limited lockdown from November 21 to December 21 with a probable extension or revision.

Figure 5 COVID new cases development in California under different policy

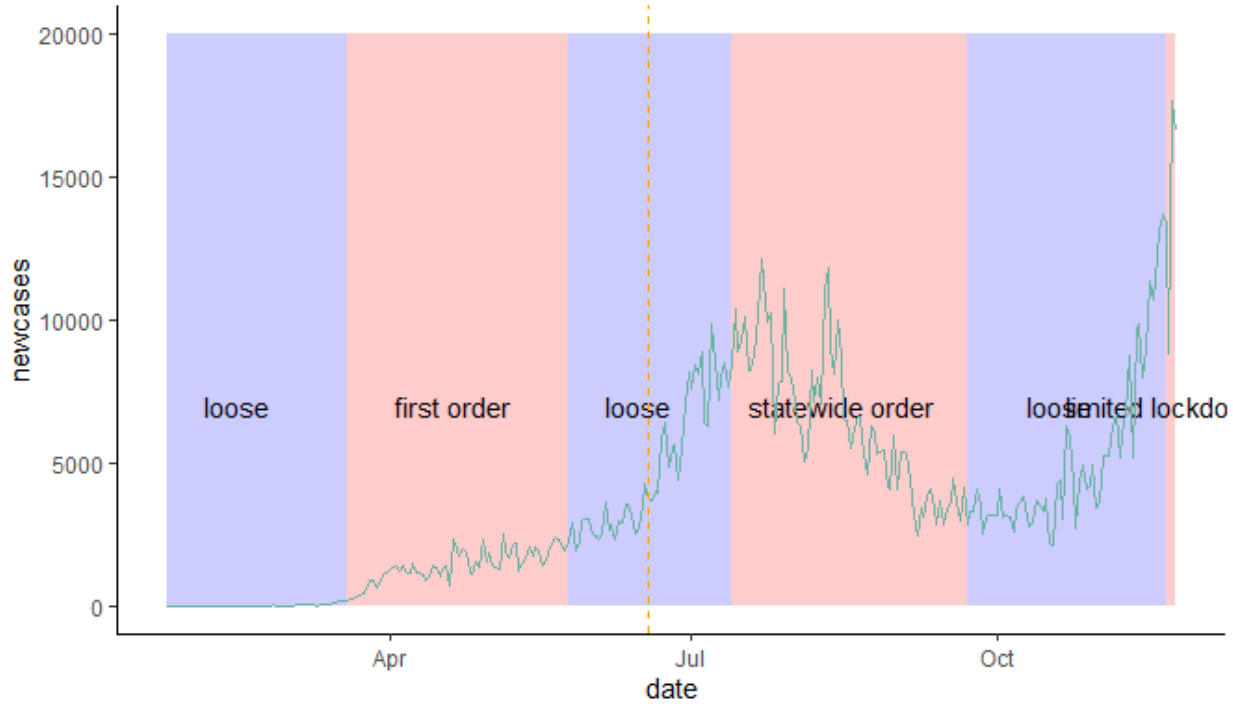
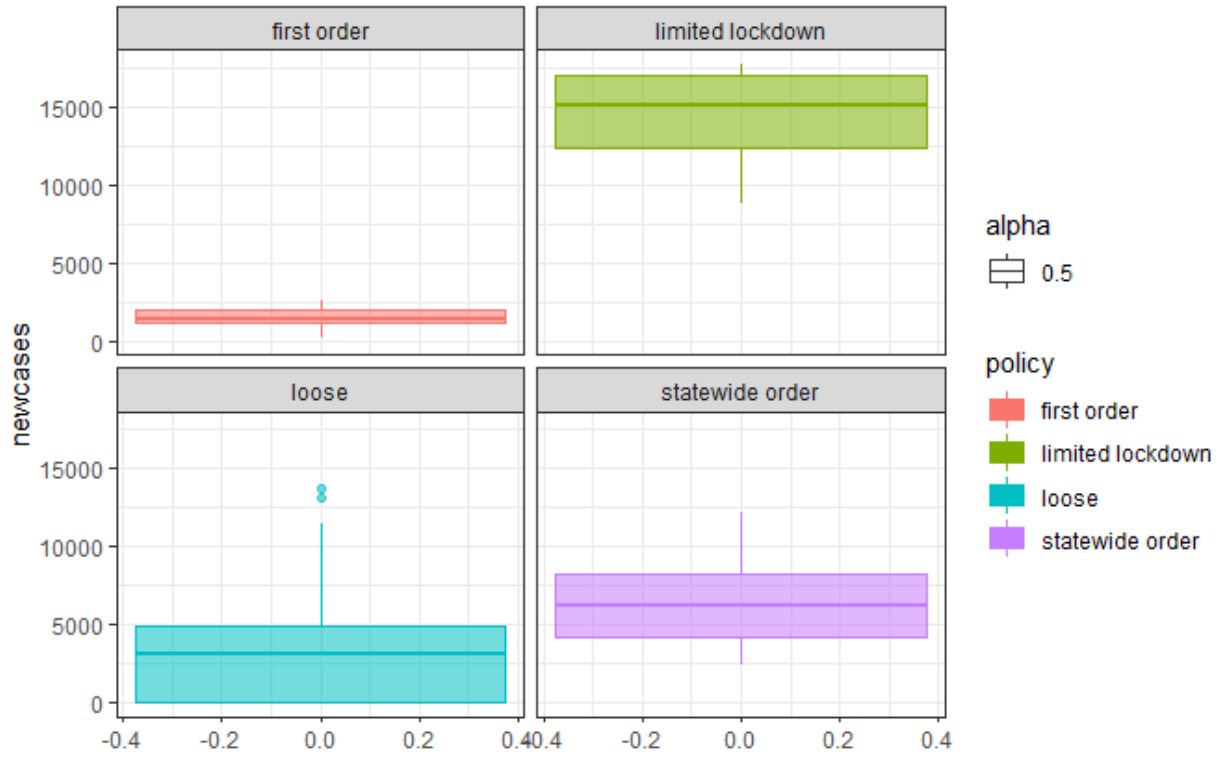


Figure 6 COVID new cases distribution in California under different policy



New cases fluctuate with a decreasing trend after restrictions effected but increasing rapidly once the policy become loose (Figure 5). During statewide order period, daily new cases has an average on nearly 6,000 but the latest data show at the beginning of limited lockdown time, it reached at 15,000 (Figure 6).

Conclusion and Summary

In conclusion, COVID death rate is significantly associated with population density. Deaths and cases were included in the model because they are the factors used in death rate calculation. The model had an adjusted R-squared of 0.833, and residuals were scattered around zero, which indicated the model has a good fit. However, the data set had limitations to build a more reliable model: CFR was sophisticated to estimate, the association need more detailed adjustment on factors such as location, humidity, people behavior, hospital capacity, etc.

And policy responses had a positive effect against COVID pandemic. The downside of this analysis is that lockdown data is not up to date because policy varies, resulting in lack of data to explicit the pattern across states.

In brief, practicing social distance and staying at home help people stay away from this contagious disease.