Modeling the Effectiveness of California's COVID Response

Kurt Stoneburner

8/4/2020

The unfolding COVID-19 pandemic a generation defining event. How effective is California's COVID-19 testing initiative in response to the unfolding pandemic? The answer is to model deaths based on confirmed cases. By measuring the model error over time, California's relative response effectiveness can be meaured.

Sub-Heading

California has excellent public health data reporting. Every day since March, the State publishes Total Tests performed in a day, Confirmed cases, Hospitalizations, ICU Utilization and death. COVID-19 is a complicated disease, there is a wide spectrum of infection outcomes, ranging from from asymptomic cases (no visible signs of illness), to mild/moderate illness, hospitalization, ICU Utilization and death. There is also the possibility of long term and permanent health damage. Based on the data provided by the State of California, it is important to measure the effectiveness of COVID testing in California. This project looks at the relationship between confirmed cases and deaths in order to build a low error rate model where deaths are predicted based on the confirmed cases from some prior date.

Model Methodology

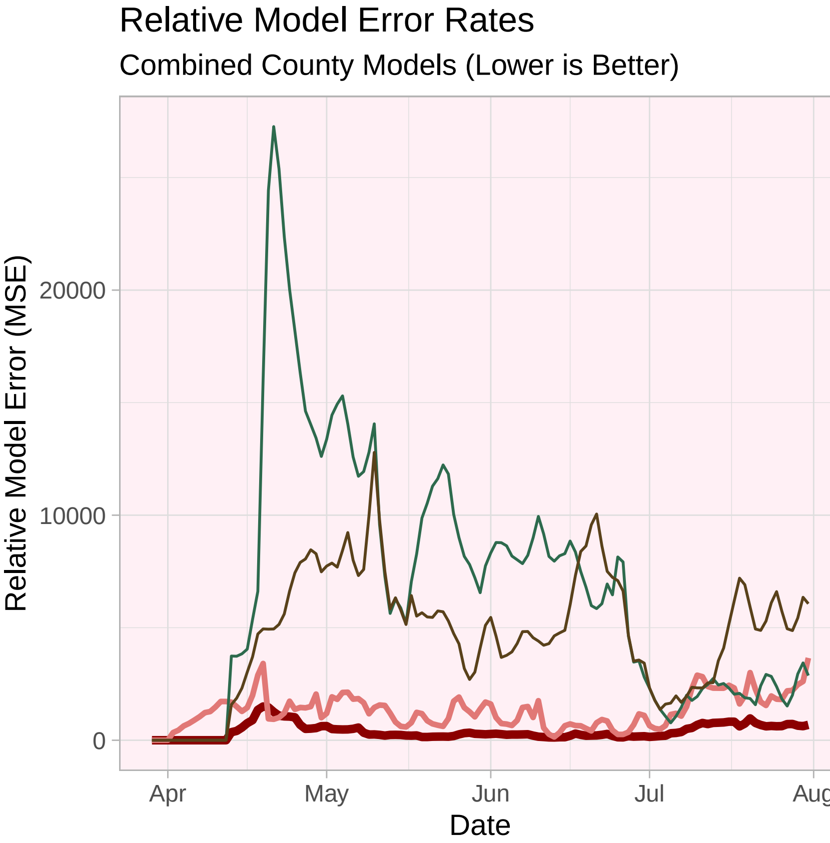
There is a general lack of information relating to COVID infection rates. With limited testing and tracing in California, it's difficult to measure the actual infection rate of COVID-19. Evaluating testing rates and confirmed cases generates a representation of some portion of the population that is infected. Testing numbers are indicating a selection bias toward testing people who have COVID-19 symptoms or have been likely exposed to someone with symptoms. This is reflected in the latest test data. On 2020-07-31(the last date of the data set). A total of 7886587 tests had been performed of which 5.0013×10^5 tested positive for COVID-19. This gives a total test positivity rate of 6.34% or 1.3out of 20 positive cases. This makes testing a poor indicator for actual total COVID-19 infections because there is no reliable data set to capture who haven't been infected.

In order to measure effectiveness, we have to measure testing results to the outcomes of Hospitalization, ICU and Death. These outcomes should be reflected in the number of confirmed cases on some earlier day. We expect to see some portion of the population reflected in the outcomes. The outcome percentage varies greatly, just like the disease. It can be measured by generating a linear model where deaths are predicted from the confirmed cases from 2 - 30 days prior. This is accomplished by creating a data frame that contains the actual data on a date (deaths, hospitalization, ICU, and testing), then adding 29 columns which points has the confirmed totals for each prior day. This generates a row that has each confirmed count from the previous 30 days. Each date (with 28 dependent variables) is modeled against the previous 15 days. This period was chosen after extensive trial-and-error and displayed the least variance in error. The linear models have tend to have some difficulty with data sets where the variables are less than 30 particularly as many of the values approach zero. This typifies the data from many of the smaller counties. Models from these low population counties tend to be more error prone, but they have low overall impact on the total model since their percentage of the whole is low, and the algorithm is selecting the best linear model for any given day.

This methodology works very well at capturing the daily variances since it generates a different model for each day based on the previous 15 days. However, there is significant variance in the data especially at the State level. California has a population of 39102839 spread across 58 counties very unevenly. For example, 24% of California's population resides in Los Angeles County. Most counties have less than 1% of the total population. There is significant social, economic, and political diversity throughout the state. The California county COVID-19 response generally reflects this diversity. Some county health departments aggressive pursued quarantine measures, others did not.

Model Error Rates

The total error in the models varied greatly. The model predicting deaths from prior confirmed cases was very accurate and generated a very low total error. Testing was slightly worse. Surprisingly, confirmed cases were not very good at predicting hospitalization and ICU.

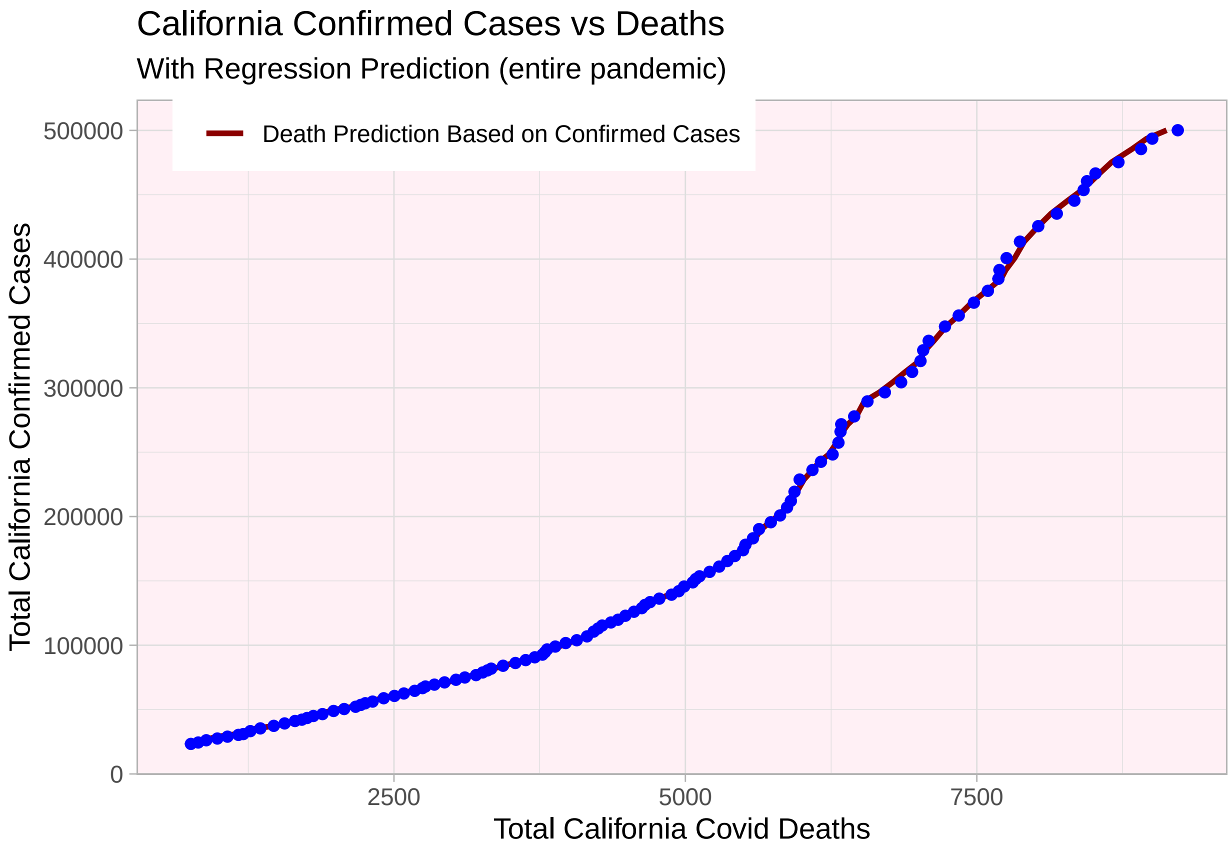


This is surprising since my assumption was that those in the hospital and ICU were more likely to die than the overall confirmed case population. The hospitalized and ICU population was also expected to die on a much shorter timeline than the confirmed cases. I expected much less variance with hospitalization and ICU.

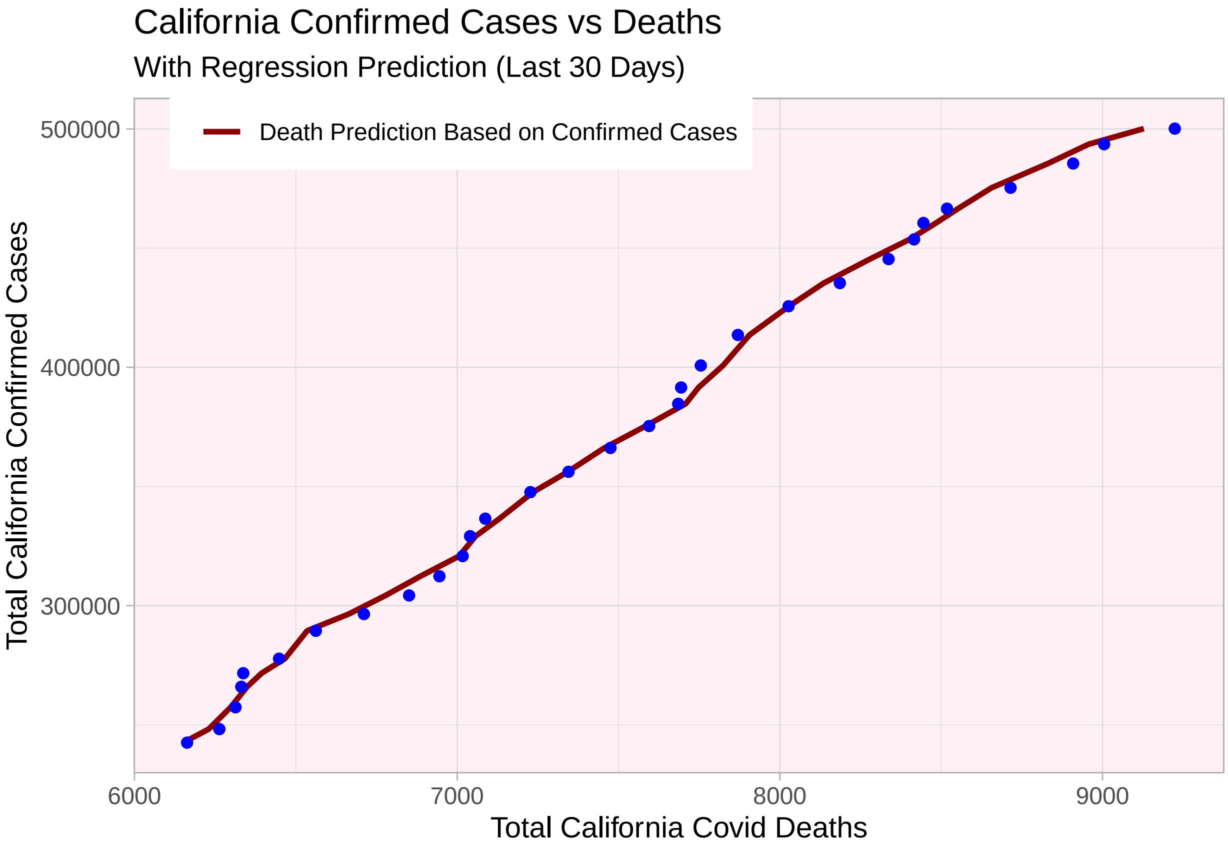
I suspect the issue lies in differing scales. A confirmed case and a death are only counted once each. Hospitalization and ICU care involve multi-day stays before patients recover or die. Being counted once, for confirmed or death vs multiple days in the hospital is a problem of scale and makes the values difficult to compare. It would be interesting to find data on active case (or model active cases) to compare with hospitalization and ICU.

Regression Model

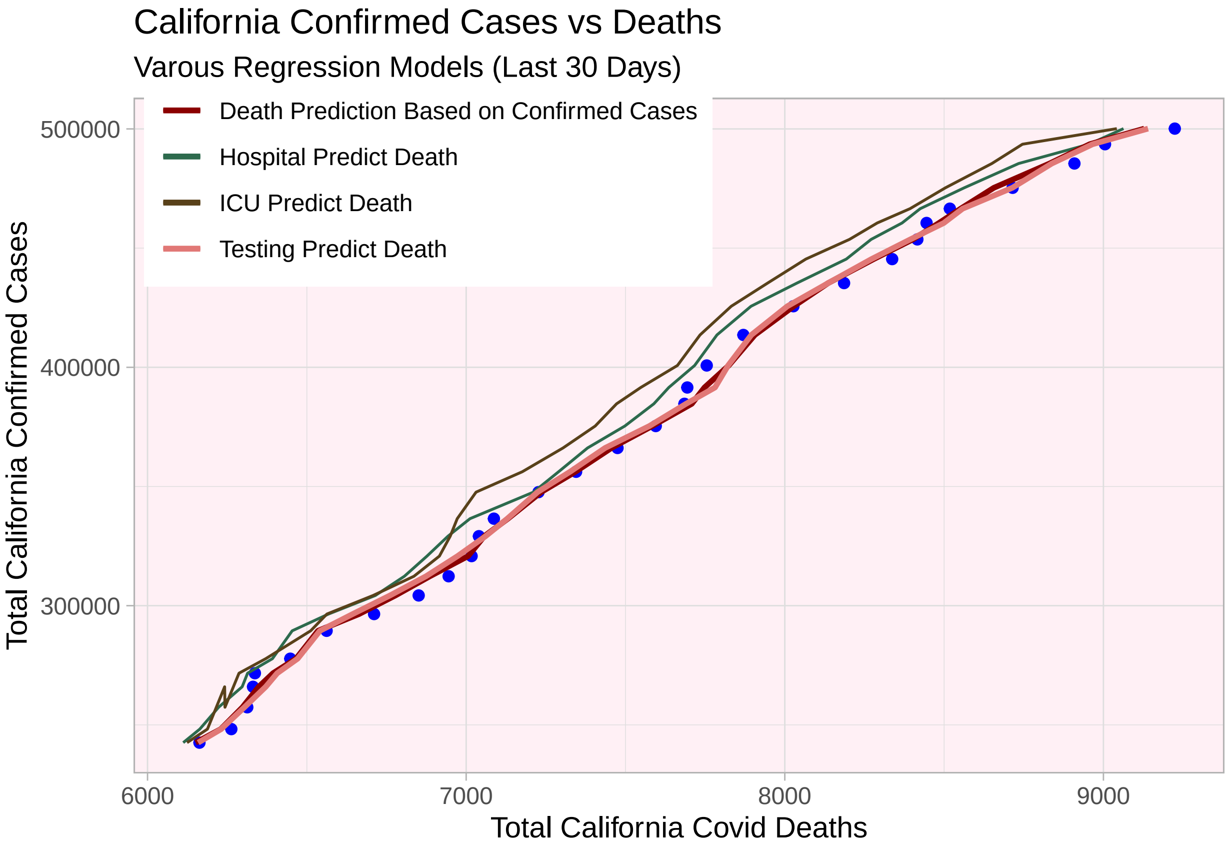
Plotting death predictions with a low error model is a bit of challenge. The regression predictions follow the actual data so closely that it is difficult to distinguish between the regression line and the actual deaths.



The differences between the regression and data points are distinguishable by plotting the last 30 days of data.



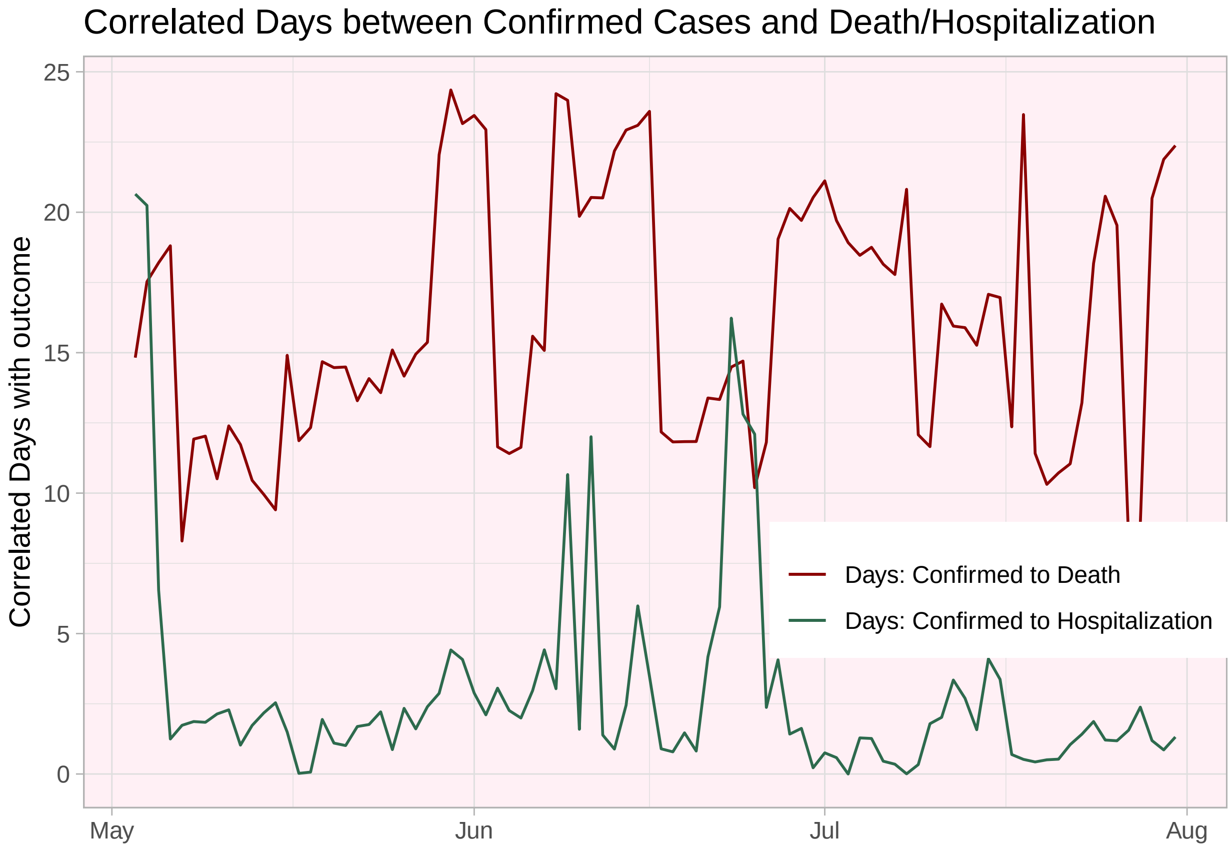
If machine learning were employed, there would be concerns about over-fitting. However, this algorithm was only effective for Testing and confirmed cases predicting death. The higher error models did get in the ballpark and could be be used for generalized predictions.



It is interesting that hospitalizations and ICU consistently predict higher than actual deaths. This supports the theory that the difference in variance is a difference in scale.

Relationship of outcomes

A big assumption about the progression of COVID-19 is the disease progression is linear and fairly consistent. If California's COVID-19 testing is meeting demand the expectation is to see a consistent correlation in the number of days between confirmed cases and deaths. There is significant variance in the number of days between confirmed cases being reported and death.

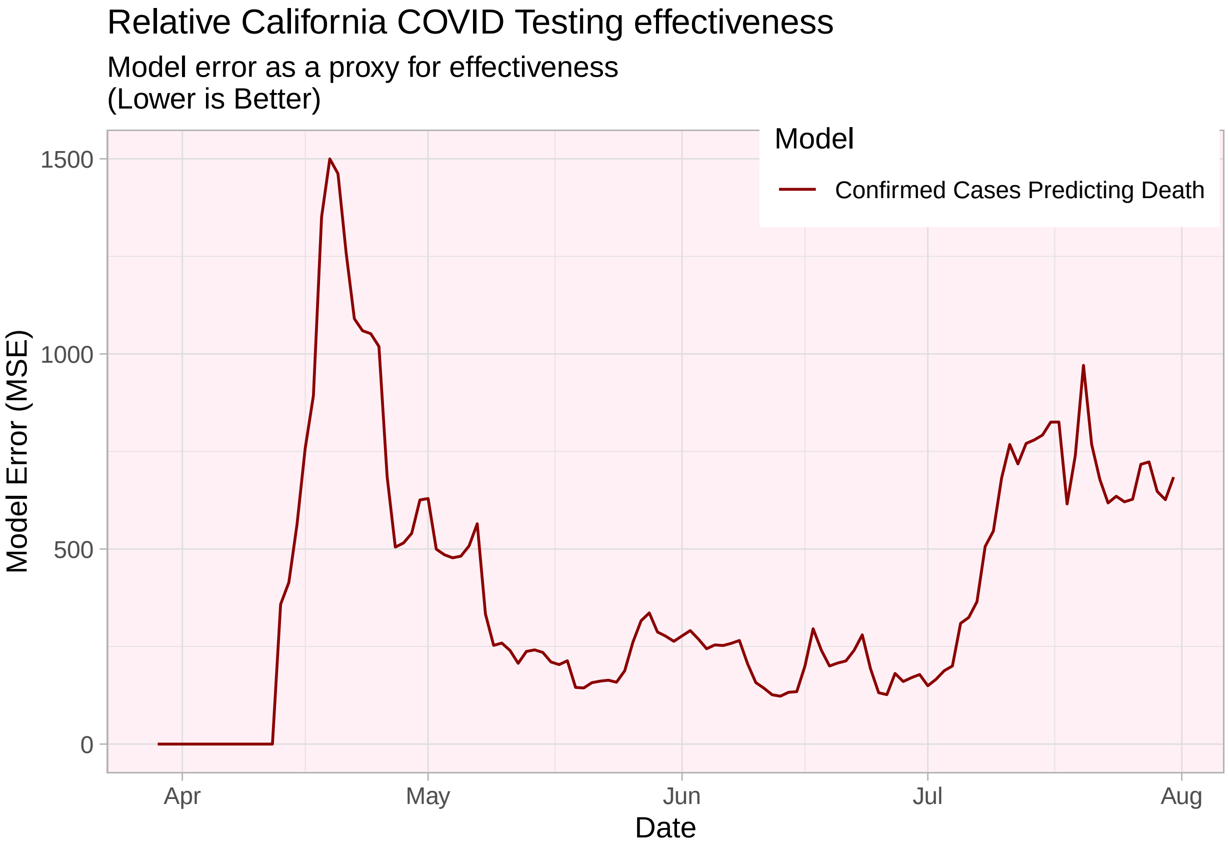


The confirmed cases from some days before ranges from 2 - 28 days. This variance is likely explained by testing not keeping up with demand. On days where the offset is high then low the following day, indicates a testing backlog. People who have been tested and not processed. Deaths from those cases are expected to continue along the regular course of the disease, but cannot be accurately captured in the data. This chart is a fairly good indicator of how inconsistent the testing response has been throughout the pandemic.

There are a few notable points about the days between hospitalization and death. Although the regression model was not very accurate at predicting deaths, the correlated days between hospitalization and death are pretty consistent (with the exception of one big spike likely due to model error). This indicates that hospitalization has been meeting demand in California. If the model was significantly off, it could indicate significant deaths were not occurring in hospitals or that people were at capacity and only received the sickest patients. Fortunately, the data does not reflect this.

Califnornia's relative effectiveness

The low error rate of the confirmed cases predicting deaths model can be used as a relative measure of California's COVID-19 testing initiative. The model captures the variance between confirmed cases and deaths. If every person who dies from COVID-19 was tested in a timely manner then the model will predict deaths with very high accuracy. Therefore, the variance not captured by the model reflects COVID-19 deaths that were not tested in a timely fashion.



The error tells the story of California's COVID-19 response.

- *April:* California is unprepared the pandemic. Testing is limited supply.
- *May & June:* California ramps up testing capacity is keeping up with demand.
- *July*: California starts the month well. Testing capacity quickly outstrips demand. This is likely due to 'quarantine fatigue' (people are tired of being locked down), combined with the start of summer, and the July 4th holiday weekend.

This analysis can be viewed as the State of California not responding to needs of the people. This is a very common narrative. As a California resident, I disagree with this view. The State was unprepared, like the rest of the country. California responded to the crisis and increased its testing response along with locking down the State. The biggest problem is that demand is exceeding capacity. Testing capacity is still limited. Therefore it is up to individual citizens to control the spread of the pandemic. The State can only do so much.

Looking Forward

There is still so much to do. The hardest part of this project was stopping to report my results. - Can the models be improved to lower the offset (days back) correlation days? - Improve county model ensembling to get better derived statewide numbers. Specifically, to generate daily linear model coefficients and intercepts. This would provide daily accurate prediction information. - Apply models to each (or at least the top 10) county(ies) in the State. What can be learned by examining the differences in county responses. - Accurately model 15 and 30 deaths counts. Use these models to estimate deaths in the next 15 and 30 days. - Improve the models for confirmed cases predicting Hospitalization and ICU by Finding or modeling active cases. - Examine the relationship of the model error to model outcome. Is there a correlation?

I fully intend to continue working with this evolving data set throughout my academic track.