Modeling the Effectiveness of California’s COVID Response

Kurt Stoneburner

8/4/2020

The unfolding COVID-19 pandemic is a generation defining event. How effective is California’s COVID-19 testing response to this unfolding pandemic? The answer lies in modeling the relationship between confirmed cases and deaths. By measuring the modeled error over time, the relative effectiveness of COVID-19 testing in California’s can be quantified.

### Unpacking the Data

California has excellent public health data reporting. Since March, the State publishes daily values for:

* Total Tests performed in a day
* New and total confirmed cases
* Hospitalizations
* ICU Utilization
* New and total deaths.

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.

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 contact tracing in California, it is 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. Current testing numbers indicate a selection bias towards people who have active 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 7,886,587 tests had been performed of which 500,130 tested positive for COVID-19. This gives a total test positivity rate of 6.34% or 1.3 out of 20 positive cases. This makes testing a poor indicator for actual total COVID-19 infections due to there being no reliable methods to capture asymptomatic cases short of testing entire populations.

In order to measure effectiveness, the outcomes of Hospitalization, ICU and Death must be modeled using confirmed cases. These outcomes should be reflected in the number of confirmed cases on some previous day. It is expected that some portion of the confirmed population be reflected in the outcomes for hospitalization, ICU and death (ie. some sick people get better, some die). The relationship between can be measured by generating a linear model where outcomes are predicted from the confirmed cases ranging 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 reflects confirmed totals for each prior day. This generates a row that has each confirmed count from the previous 30 days. Each date (with 29 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 to have difficulty where the data sets are less than 30 items, particularly when several of the values are zero. This typifies the data from many of the smaller counties early in the pandemic. 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 effect on the whole is low. This error is further mitigated by the algorithm selecting the best linear model for any given day.

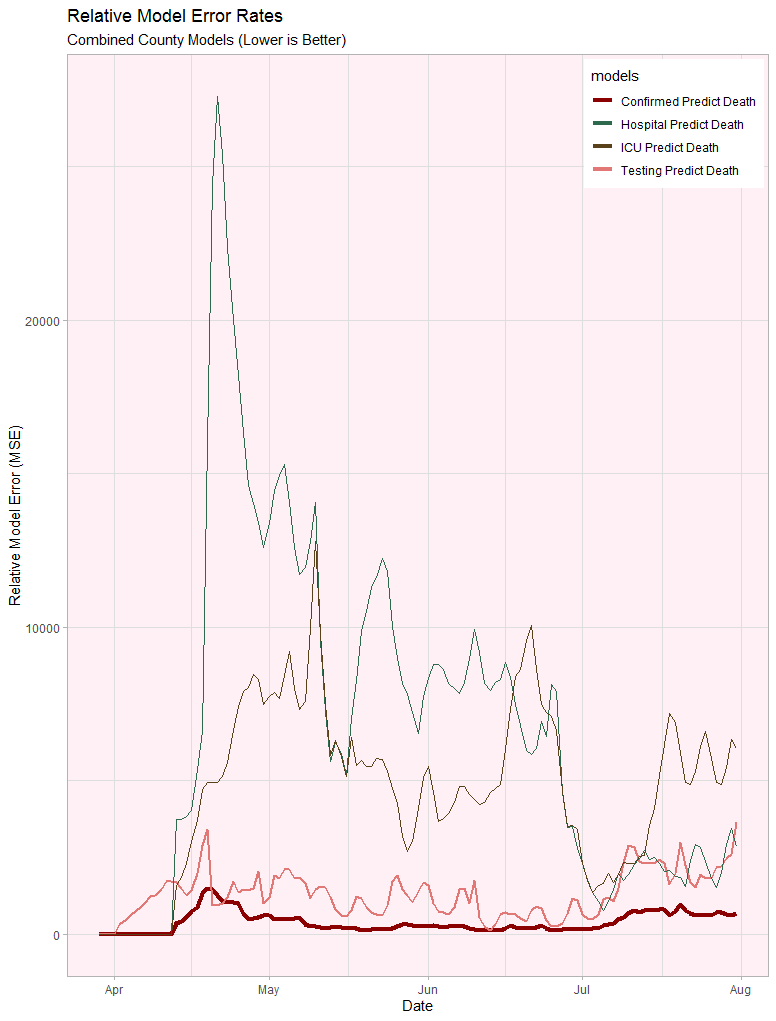
This methodology works very well at capturing daily variances since it generates a different model for each day based on the previous 15 days. However, there is significant variance in the data especially at the State level. California has a population of 39102839 spread across 58 counties very unevenly.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

For the best model results, each county is modeled individually then with the county results summed creating an aggregate statewide model. This process generates a significant number of models and becomes quite CPU intensive.

125 days to model \* 28 previous Confirmed days \* 58 counties = 203000 total models! That’s for each set of variables: Confirmed predict death, confirmed predict hospitalization, confirmed predict icu, plus a variety of other combinations. It turned out to be alot of models. Each one taking around 7 minutes to run. Typically, the models would built once daily and saved for later use.

### Model Error Rates

The total error in the models varies greatly. The model predicting deaths from prior confirmed cases was very accurate and generated a very low total error. Testing had more error but remains fairly accurate. Surprisingly, confirmed cases were not very good at predicting hospitalization and ICU.

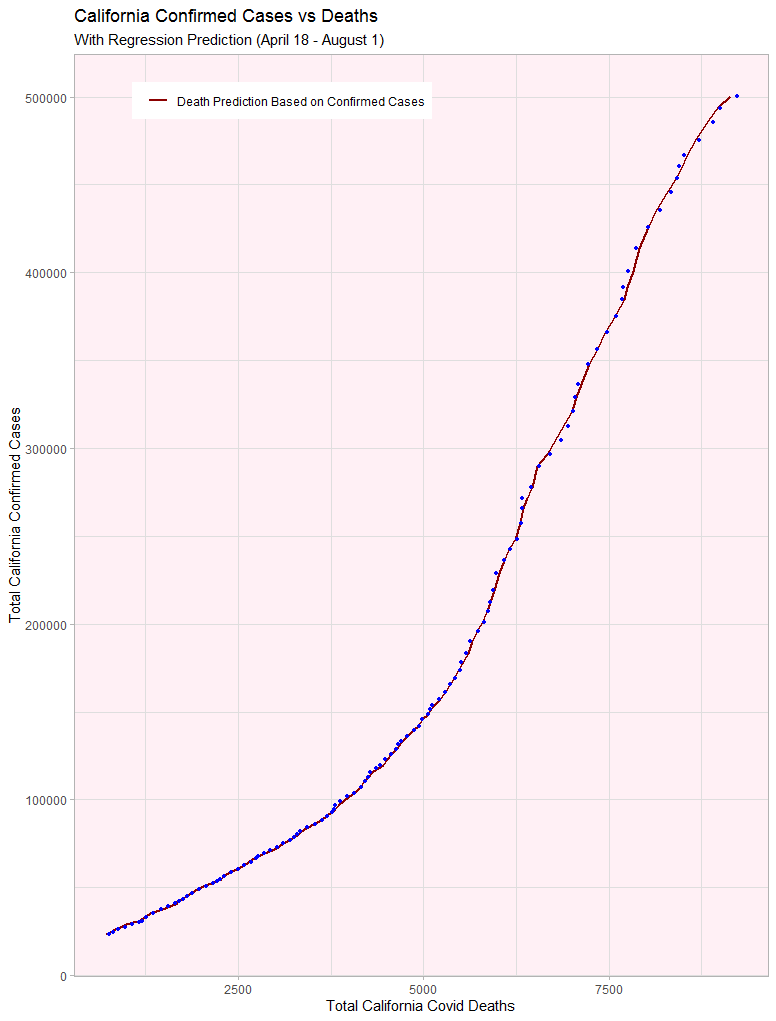


This is surprising since the assumption is those in hospital and ICU were more likely to die than the overall confirmed case population. The hospitalized and ICU population was also expected to die on a much shorter timeline than the confirmed cases. I expected much less variance with the hospitalization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

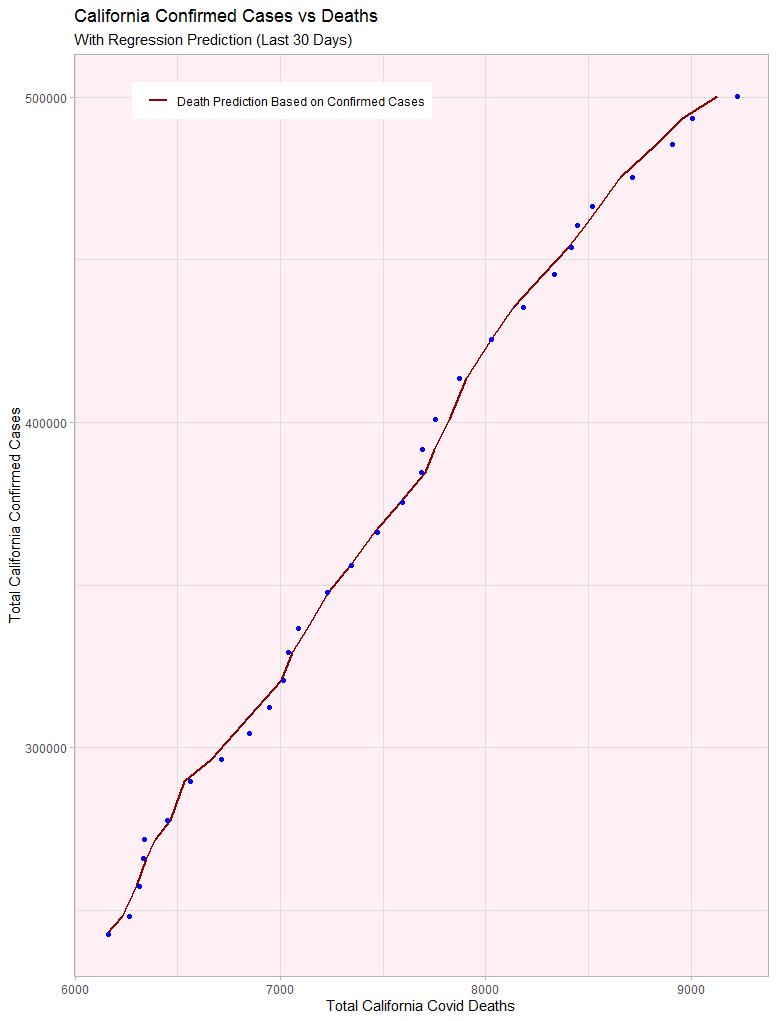
I suspect the issue lies in differing scales. A confirmed case and a death are only counted once. Hospitalization and ICU care involve multi-day stays before a patient recovers or dies. Being counted once, for confirmed or death vs multiple days in the hospital is a problem of scale and makes the values 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

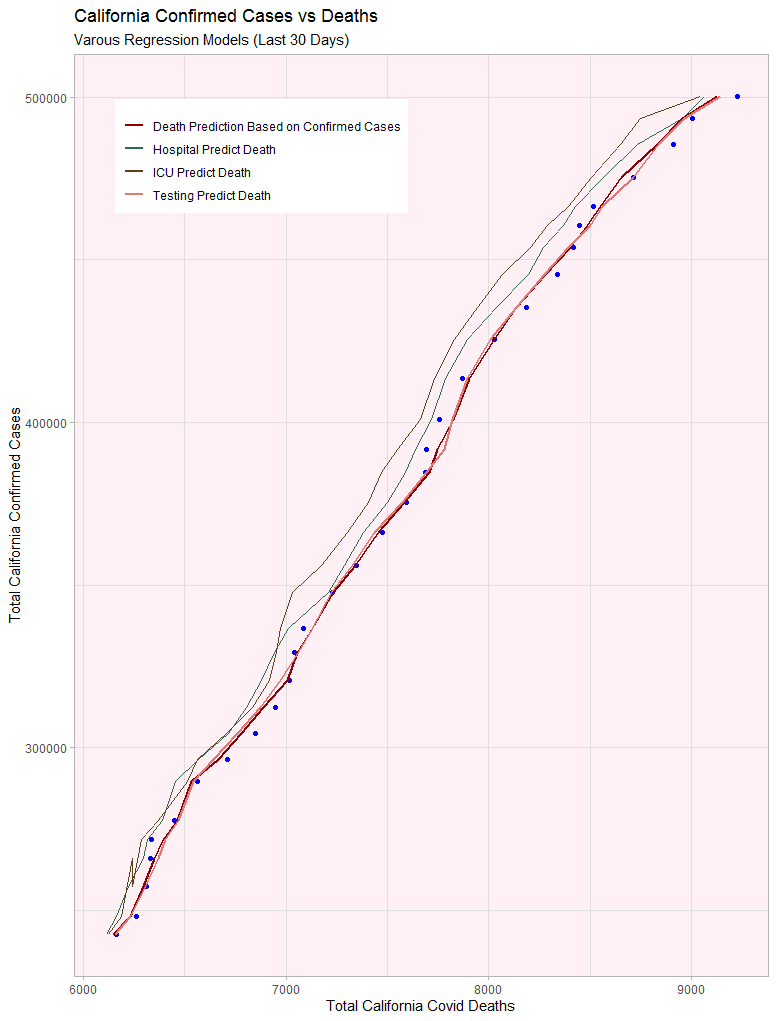
Plotting death predictions with a low error model is a bit of challenge. The regression predictions follow the actual data so closely that it is difficult to distinguish between the regression line and the actual deaths when plotting all deaths in California.



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



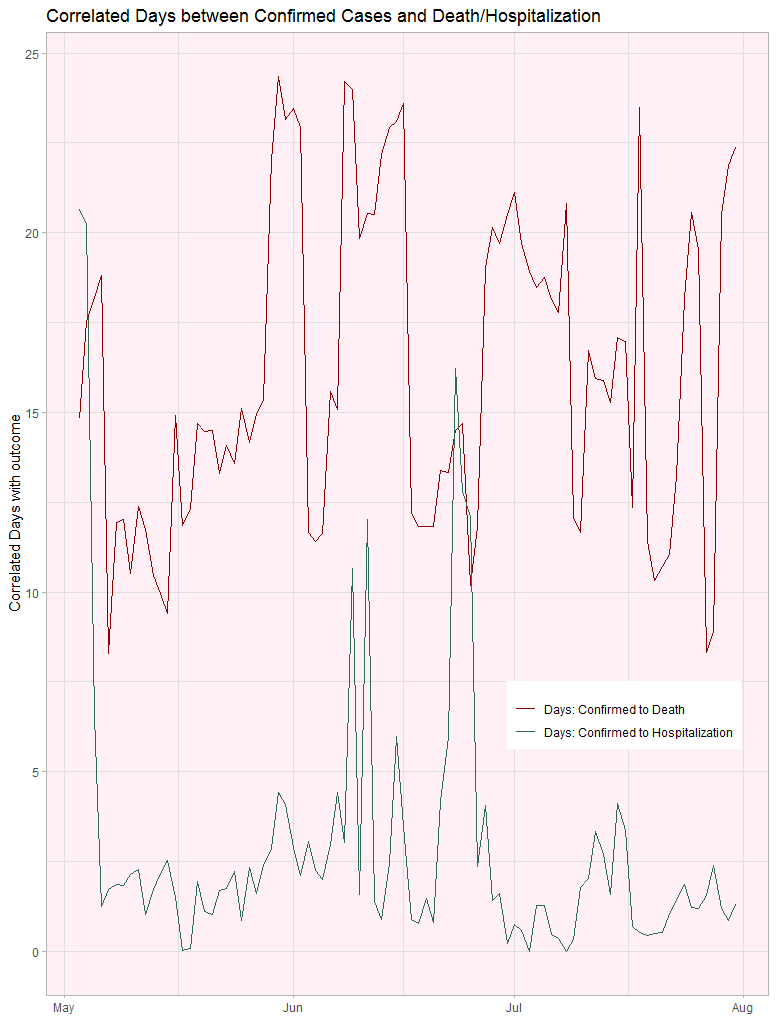
If machine learning were employed, there would be concerns about over-fitting. However, this algorithm was only effective for Testing and confirmed cases predicting death. The higher error models did get in the ballpark and could be be used for generalized predictions.



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 COVID-19 is the disease progresses in a linear and fairly consistent manner. 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. The data portrays a significant variance in the number of days between confirmed cases being reported and death.

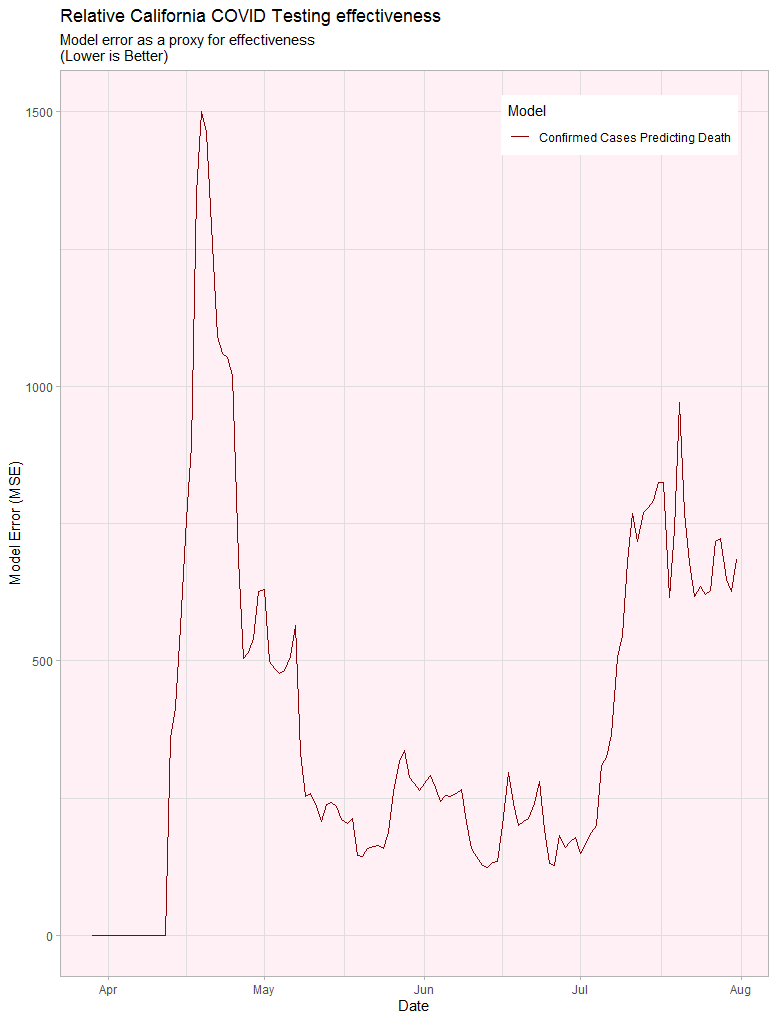


The confirmed cases from prior days ranges from 3 - 28 days. This variance is likely explained by testing not keeping up with demand. On days where the offset is high then low the following day, indicates a testing backlog: People who have been tested and but the results are delayed. Deaths captured in testing delays reflect the disease progressing at a normal date, then die shortly after results are released. This chart is a fairly good indicator of testing delays 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.

### California’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 in high demand and limited supply.
* *May & June:* California ramps up testing capacity and keeps up with demand.
* *July*: California starts the month well then demand quickly outstrips testing capacity. 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 with testing demand 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.