Linear Predictions between COVID Cases and COVID Deaths

Kurt Stoneburner

7/4/2020

It all starts with correlation

We know that some portion of confirmed COVID cases will result in death. I spent my Saturday looking at the State of California COVID confirmed and Death numbers. The values are represent the number of people confirmed with COVID and those whose deaths are attributed to COVID. Any person dying on a given day, should be represented by a confirmed test on a previous day. This is an attempt to explore that relationship with the tools we have discovered so far.

Like any project the data must be collected and cleaned. The initial data values are date, county, total confirmed and total deaths. This paper focuses on statewide numbers, all values by county were summed to provide the statewide totals for the day.

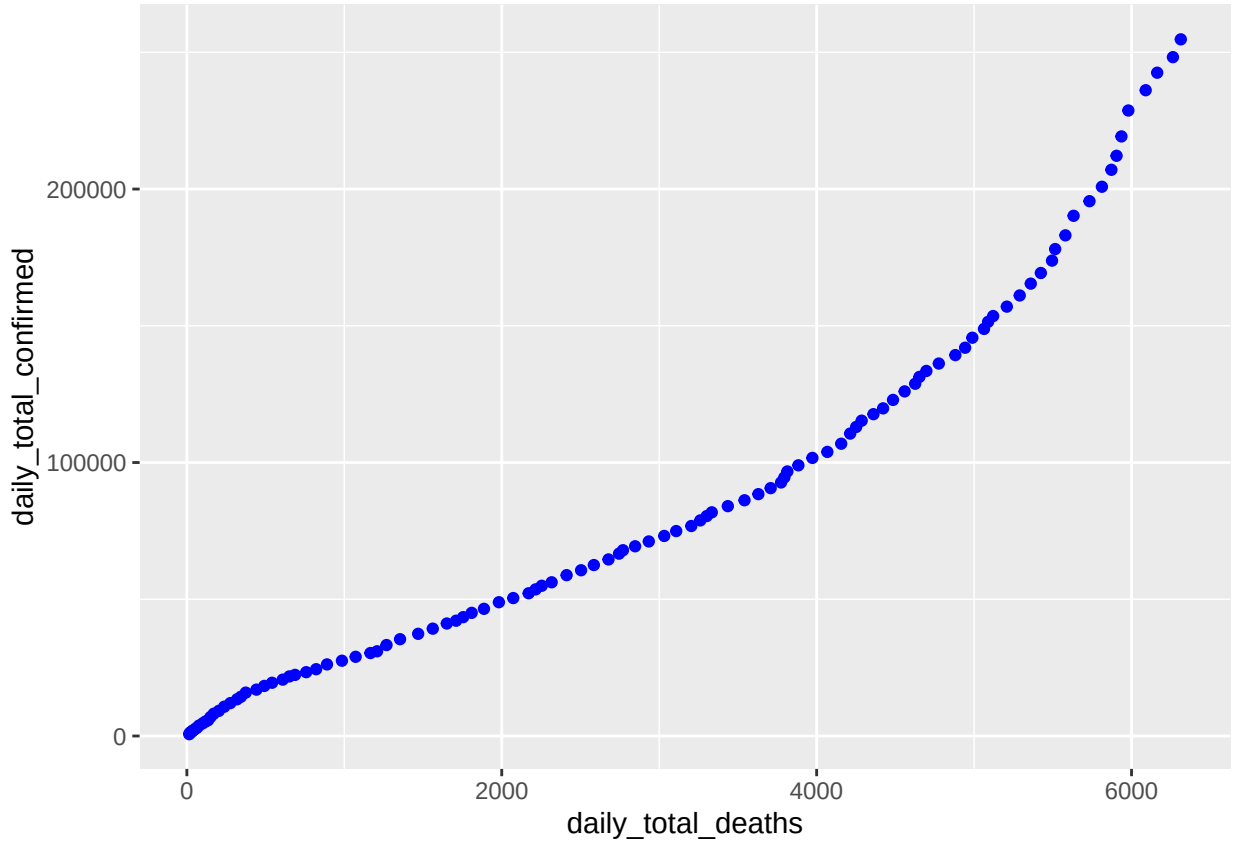
Looking at the numbers from the State of California [1]. There is a strong Pearson's correlation between confirmed cases and deaths.

```
cor(daily_covid_df$daily_total_confirmed,daily_covid_df$daily_total_deaths)
```

```
## [1] 0.969137
```

Plotting the total confirmed cases vs the total deaths shows a strong relationship between the values

[1] The official numbers from the State are available at:
https://data.ca.gov/group/covid-19



Baseline linear modeling

People don't die when they are confirmed as COVID positive [2]. Death generally occurs at some interval after testing. Can deaths be predicted from confirmed cases using a simple linear model?

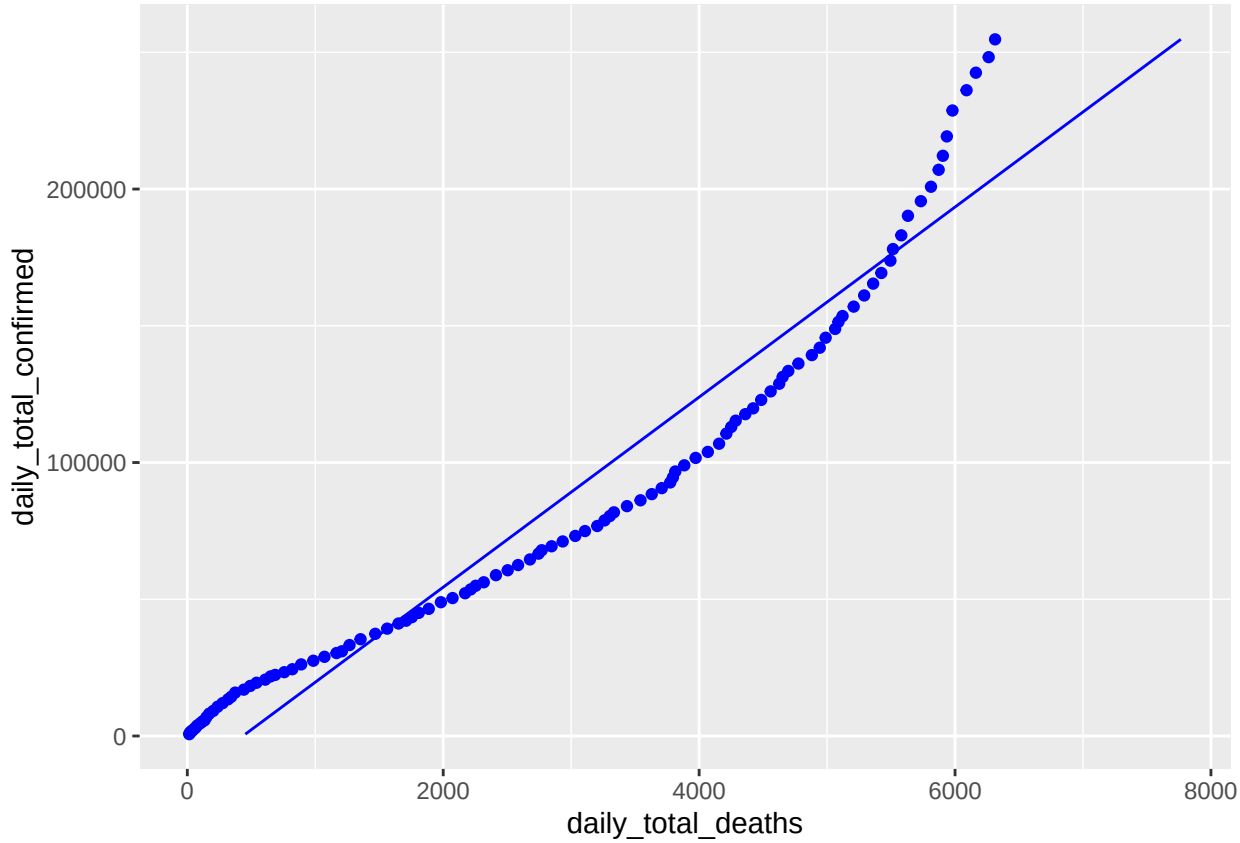
Using the unaltered data, the linear model has very significant R vale

```
### R value. R^2 taken from Summary
sqrt(.9467) ## R=.9729851
```

```
## [1] 0.9729851
```

This is a baseline plot. It's confirmed cases vs deaths (by date). The solid blue line is a base line prediction of deaths. Visually, we can tell the baseline model isn't very helpful in predicting deaths based on confirmed cases, Correlation is not causation.

[2]People are generally tested when displaying symptoms. Although some tests are performed post mortem.



The Hunt for a better correlation

Predicting deaths by the number of confirmed on the same day is clearly not a good way to go about it. The next step is to correlate deaths on a given day with confirmed on a previous. But over what range? I have been working under the assumption that death occurs around 21 days from infection.[3] Assuming this was true, symptoms don't typically present for at least 5 days. If someone was tested immediately, results may be reported at plus 7 days from infection. An offset of 14 or 15 is likely at the outer end of offsets. But for completeness I calculated from 21.

I wrote a function that addded offset columns from 1 to 21. Example: An offset of 5 would align the COVID Death value with a COVID confirmed value from 5 days prior. By plotting all the plausible offsets the resulting correlation values can be analyzed.

The resulting correlation table was quite interesting.

```
### Correlate Daily Total row
### Looking for column with highest correlation value
### Looks like offset_9
cor(offset_daily_df[2:length(colnames(offset_daily_df))])[2,]
```

```
## daily_total_confirmed    daily_total_deaths              offset_1
##             0.9691370             1.0000000             0.9705391
##              offset_2              offset_3              offset_4
##             0.9719240             0.9733846             0.9748377
```

[3] I could not find the original article where I sourced that information. Right or wrong it's a place to start

```
##              offset_5              offset_6              offset_7
##             0.9761435             0.9770342             0.9776848
##              offset_8              offset_9             offset_10
##             0.9782511             0.9786178             0.9788798
##             offset_11             offset_12             offset_13
##             0.9790107             0.9788002             0.9784087
##             offset_14             offset_15             offset_16
##             0.9778820             0.9771510             0.9762894
##             offset_17             offset_18             offset_19
##             0.9752687             0.9740991             0.9728062
##             offset_20             offset_21
##             0.9715116             0.9701439
```

It looks like offset_9 has the highest correlation between confirmed cases and deaths. This is worth examining further.

Let's look at the Linear Model and Predicted Data Frame.

```
### Build Model off of offset_9. This offset had the highest correlation value between confirmed and de
state_CD_offset_9_lm <-  lm(daily_total_deaths ~ offset_9, data=offset_daily_df)

### Created Prediction data frame using offset_15
### Build prediction based on Model
state_CD_offset_9_predict_df <- data.frame(
  daily_total_deaths = predict(state_CD_offset_9_lm,
                               newdata=offset_daily_df),
  daily_total_confirmed=offset_daily_df$daily_total_confirmed)
```

The R value has higher than the baseline R value

```
sqrt(.9619)
```

```
## [1] 0.980765
```

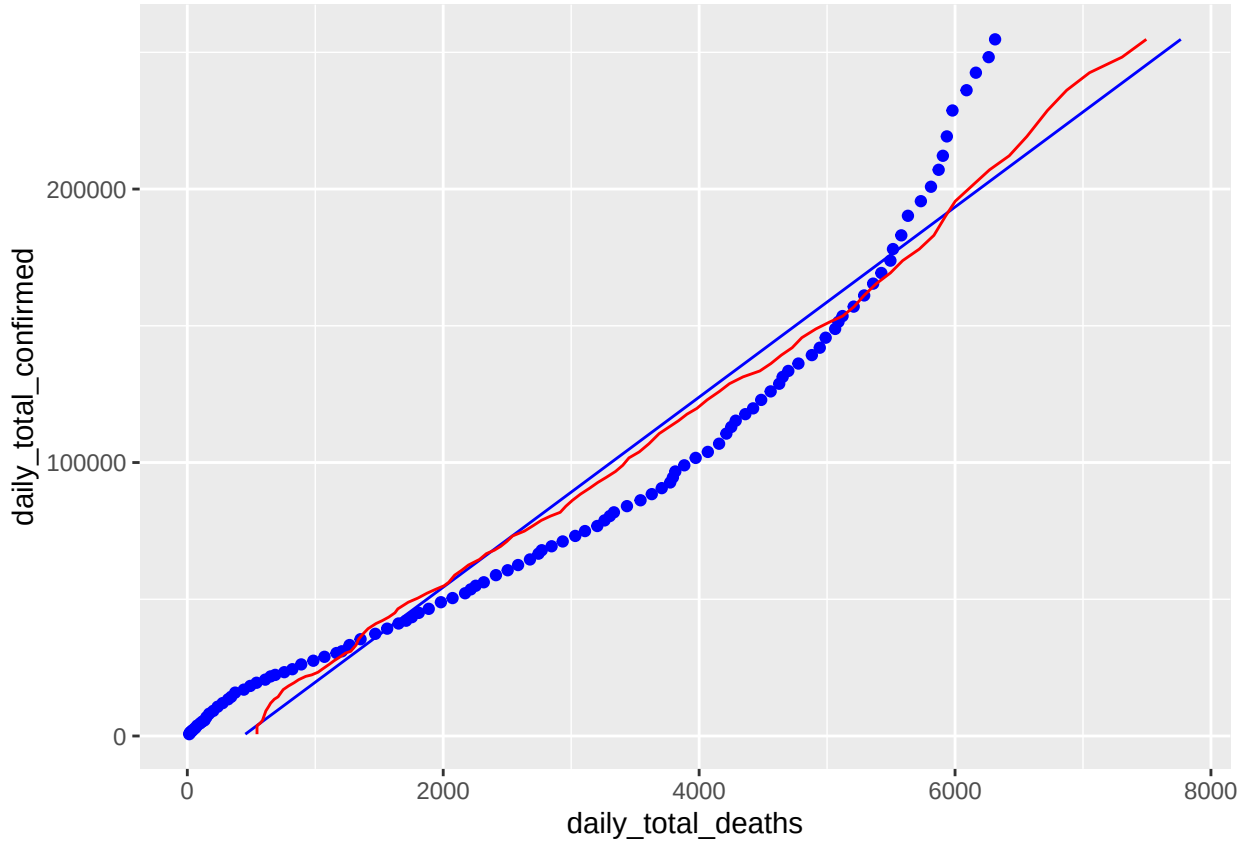
Looking at the residuals (which is a total error value). The Offset 9 residuals have a smaller error value than the baseline prediction.

```
## [1] "Base Residual: 26944499.31"
```

```
## [1] "Offset 9 Residual: 18757274.47"
```

```
## [1] "offffset 9 < Base: TRUE"
```

Plotting the results. Blue is the baseline prediction, red is our offset 9 linear model prediction.



The resulting predictions are better. Not great and certainly not definitive. One major assumption is for this model is that correlated offset value remains constant. The testing situation has improved significantly over time. Let's explore a subset of the data that is more indicative of the present situation.

"I like California in June"

The Offset 9 prediction intercepts the actual data around the 5000 death mark, which is in the middle of June. Subsequently the data deviates significantly from the model. This requires building a data frame with just the June data and running a correlation test to validate the offset.

```
cor(june_offset_df[2:length(colnames(june_offset_df))])[2,]
```

```
## daily_total_confirmed    daily_total_deaths              offset_1
##             0.9814765             1.0000000             0.9815363
##              offset_2              offset_3              offset_4
##             0.9817678             0.9830717             0.9851963
##              offset_5              offset_6              offset_7
##             0.9875528             0.9893667             0.9907568
##              offset_8              offset_9             offset_10
##             0.9921750             0.9933800             0.9948098
##             offset_11             offset_12             offset_13
##             0.9963907             0.9972745             0.9976784
##             offset_14             offset_15             offset_16
##             0.9978578             0.9975115             0.9972116
##             offset_17             offset_18             offset_19
```

```
##             0.9971950             0.9972744             0.9971628
##             offset_20             offset_21
##             0.9970676             0.9967216
```

The correlation table reveals the offset has shifted from 9 (for the whole data set) to 13 or 14. Since 13 and 14 are so close, we'll calculate the residuals for each and go with the smallest value.

```
## [1] "June Offset 13 residuals: 53640.3928940913"
```

```
## [1] "June Offset 14 residuals: 49500.5512911547"
```

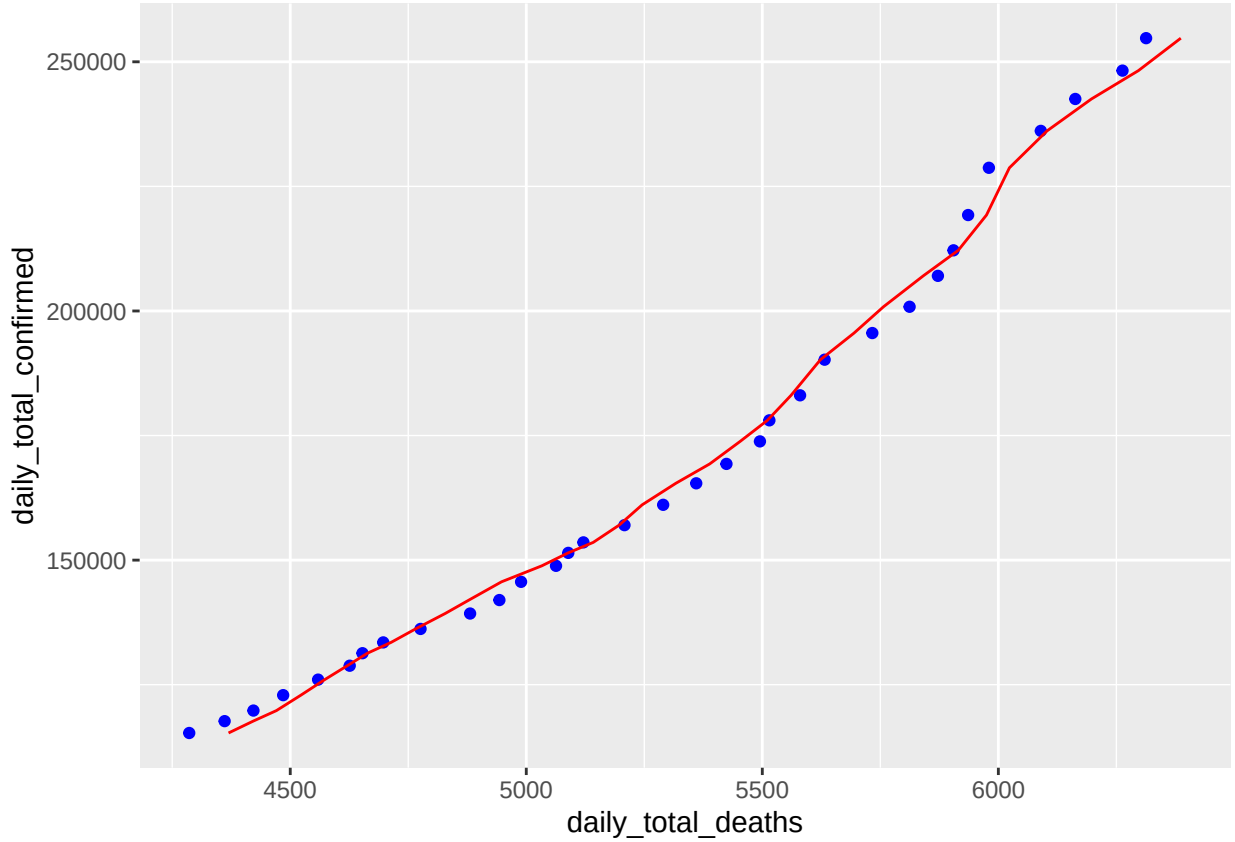
```
## [1] "June Offset 15 residuals: 57490.9794889714"
```

```
## [1] "June offset 14 is less than offset 13? TRUE"
```

```
## [1] "June offset 14 is less than offset 15? TRUE"
```

Offset 14 has the smallest residual which corresposnds with the highest correlation value in the offset table. This is further supported by calculating the residuals of offset 13 and 15 which were both higher. Since we are using a linear model offset 14 appears to be the best choice.

Here's the plot of June confirmed and deaths. The red line is the prediction.



That result stopped me in my tracks. There appears to be a very tight correlation between predicted deaths and the reported confirmed cases from 14 days prior. The data also demonstrates the offset days shifted during the pandemic. Likely due to an increase in testing and people not waiting as long to get tested. This is an interesting find and I intend to keep picking at it.