COVID-19 Analysis

Christopher Washington

2024-06-14

Introduction

The COVID-19 pandemic is something that has affected nearly every person on earth, and in such a way that requires statistical analysis. We will see in this report how the total number of cases and deaths in the state of California have progressed over time, and provide a method for predicting future totals based on the number of days since the first occurrence of COVID-19 in California.

Required R Libraries

The following R libraries are required to run the R code contained in this document:

```r
library(tidyverse)
library(knitr)
library(lubridate)

set.seed(10)
```

Dataset

The dataset that we will use for the following analysis is a dataset of time series data about the number of cases and deaths in each county of the United States of America. ## Data Retrieval We will begin by loading the dataset:

```r
base_url = "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_co

cases = read_csv(paste(base_url, "time_series_covid19_confirmed_US.csv", sep=""))

deaths = read_csv(paste(base_url, "time_series_covid19_deaths_US.csv", sep=""))
```

Before conducting our analysis, there are a number of preprocessing steps that we must do on this dataset. The following R code is responsible for pivoting the dataset so that each row corresponds to a specific date, as well as renaming columns for clarity and ease of use.

```r
cases = cases %>%
  pivot_longer(
    cols = -(UID:Combined_Key),
    names_to = "date",
    values_to = 'cases'
  ) %>% select(Admin2:cases) %>%
  mutate(date = mdy(date)) %>%
  select(-c(Lat, Long_)) %>%
  rename(county = Admin2) %>%
  rename(state = Province_State)

deaths = deaths %>%
  pivot_longer(
    cols = -(UID:Population),
    names_to = "date",
    values_to = 'deaths'
  ) %>% select(Admin2:deaths) %>%
  mutate(date = mdy(date)) %>%
  select(-c(Lat, Long_)) %>%
  rename(county = Admin2) %>%
  rename(state = Province_State) %>%
  rename(population = Population)
```

We then join the cases and deaths data frames.

```r
totals = cases %>%
  full_join(deaths) %>%
  select(county, state, date, cases, deaths, population)
```

The analysis that we will conduct is specifically related to the cases and deaths in the state of California, so we will retrieve the state totals for all states with the following R code

```r
totals = totals %>%
  group_by(state, date) %>%
  summarize(
    cases = sum(cases),
    deaths = sum(deaths))
```

We will then filter out all states other that California.

```r
totals = totals %>%
  filter(state == 'California') %>%
  filter(cases > 0)
```

Finally, we will add a column that represents the number of days since the first case occurred in California.

```r
totals = totals %>%
  group_by(state) %>%
  mutate(days_since_first_case = as.integer(date - min(date)))
```

After our pre-processing, the final date looks like the following.

```r
kable(head(totals))
```

state	date	cases	deaths	days_since_first_case
California	2020-01-26	2	0	0
California	2020-01-27	2	0	1
California	2020-01-28	2	0	2
California	2020-01-29	2	0	3
California	2020-01-30	2	0	4
California	2020-01-31	3	0	5

Model 1 Days Since First Case vs Cases in California

The first model that we will consider is a linear model of the number of cases in California as a function of the number of days since the first case in California. We will use the following linear model.

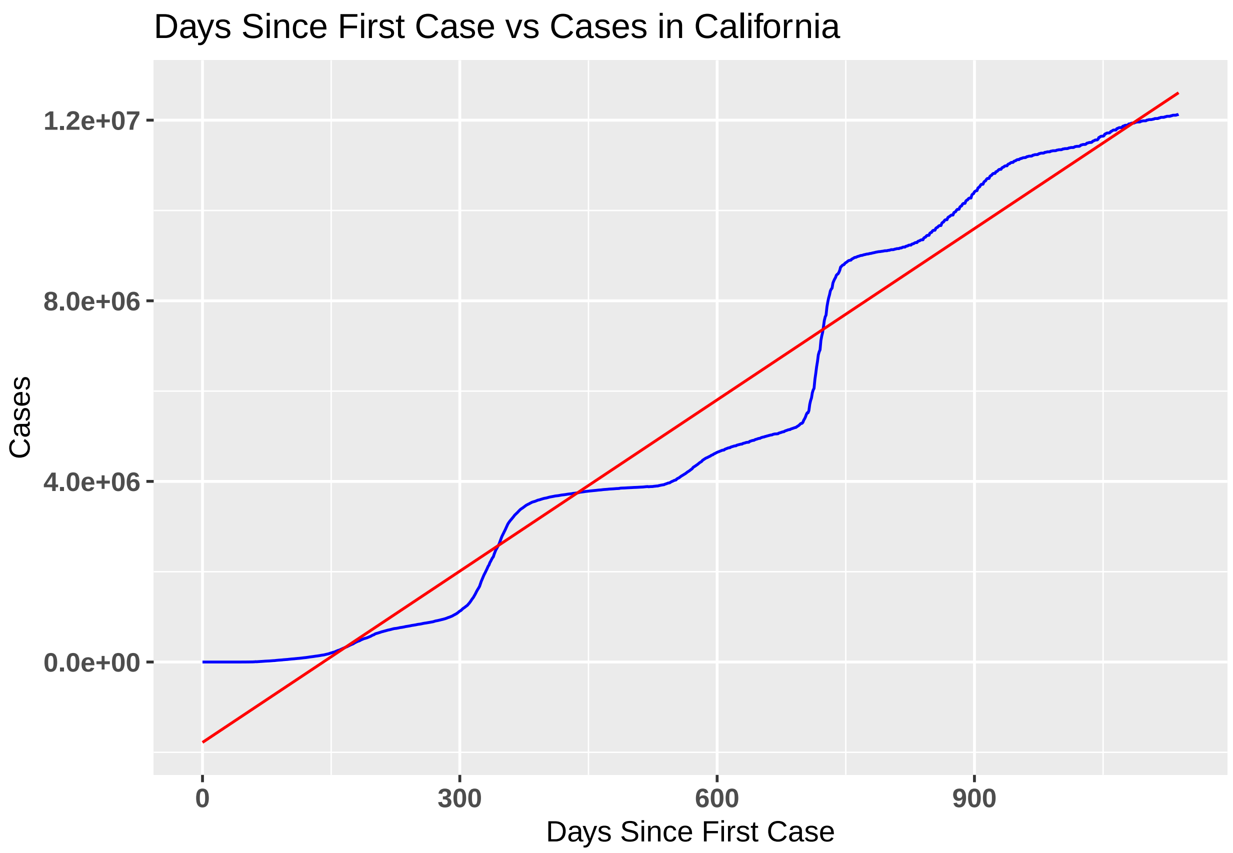
```r
model_cases = lm(
  cases ~ days_since_first_case,
  data = totals
)

summary(model_cases)
```

```
##
## Call:
## lm(formula = cases ~ days_since_first_case, data = totals)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1772399  -630661   103671   648586  1780709
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -1.781e+06  4.921e+04  -36.19   <2e-16 ***
## days_since_first_case  1.264e+04  7.488e+01  168.85   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 831000 on 1137 degrees of freedom
## Multiple R-squared:  0.9616, Adjusted R-squared:  0.9616
## F-statistic: 2.851e+04 on 1 and 1137 DF,  p-value: < 2.2e-16
```

We then append a column of predicted values to the dataset.

```r
totals = totals %>%
  mutate(pred_cases = predict(model_cases))
```

Finally, we have the following visualization of the actual data (in blue) and the predicted data (in red).



Model 2 Days Since First Case vs Deaths in California

The second model that we will consider is a similar model, but regarding the deaths in California instead of the cases. The following R code prepares the linear model.

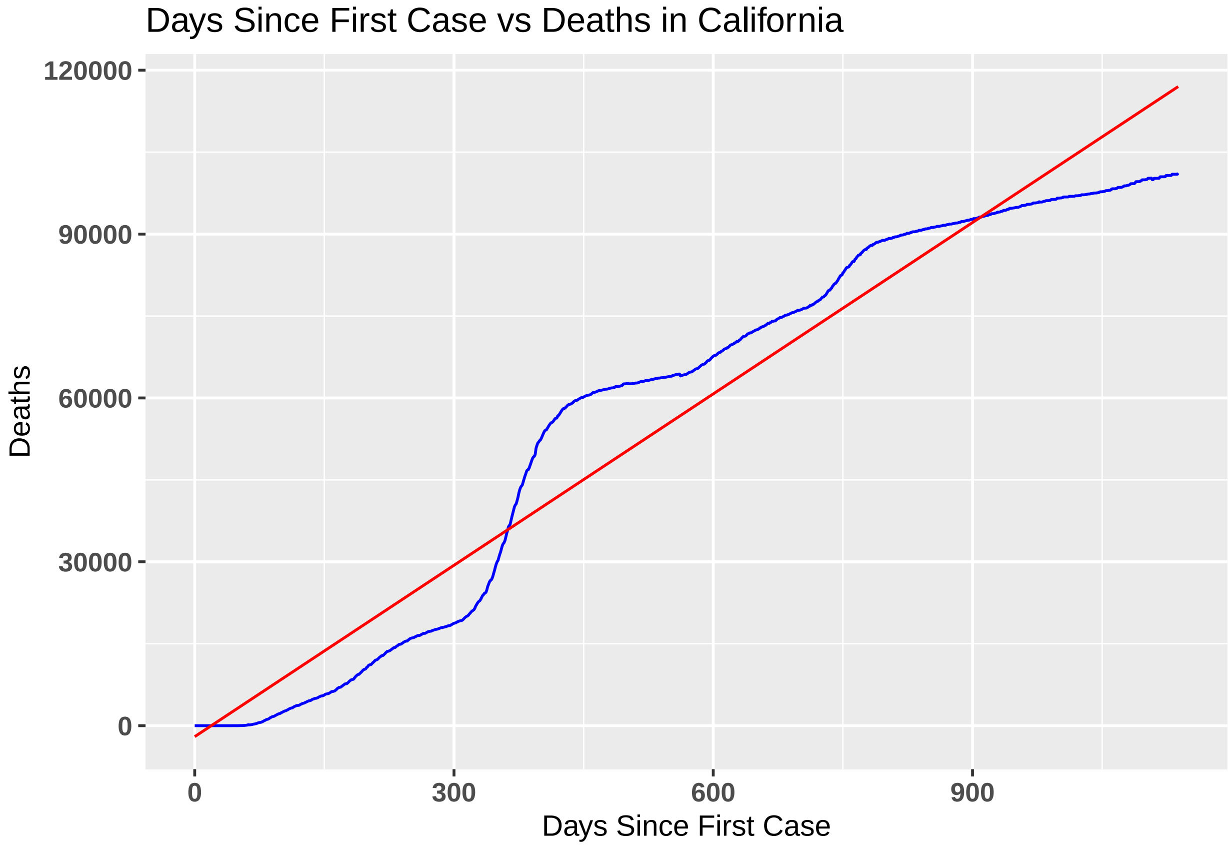
```
model_deaths = lm(
  deaths ~ days_since_first_case,
  data = totals
)

summary(model_deaths)
```

```
##
## Call:
## lm(formula = deaths ~ days_since_first_case, data = totals)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15945.9  -7909.7   -104.6   6731.9  15542.5
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -2008.3075   496.4637  -4.045 5.58e-05 ***
## days_since_first_case   104.5806     0.7555 138.433  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 8383 on 1137 degrees of freedom
## Multiple R-squared:  0.944,  Adjusted R-squared:  0.9439
## F-statistic: 1.916e+04 on 1 and 1137 DF,  p-value: < 2.2e-16
```

```
totals = totals %>%
  mutate(pred_deaths = predict(model_deaths))
```

We then have the following visualization displaying the actual data (in blue) and the predicted data (in red).



Bias Identification

There are a few immediate sources of bias in the dataset that need to be addressed. The first is the reporting of cases in each county. Since the dataset only contains cases of COVID-19 that have been confirmed by professional medical personel, the data almost certainly undercounts the real number of cases in each county. It is not farfetched to believe that there are numerous individuals in each county who contracted the illness, but never sought medical attention for it, thereby underrepresenting them in the dataset. Another source of bias in the dataset is the way that the number of deaths due to COVID-19 in each county has been computed. Throughout the duration of the pandemic there seemed to be many metrics for determining that the cause of an individual's death was COVID-19, and this can also vary from county to county. At this point, long after the fact, there is no solution to remedy these biases, but we should certainly be aware of them in our analysis.

Conclusion

In this report we analyzed a dataset of time series data of COVID-19 cases and deaths in each county of the United States of America. In particular, we considered the total cases and deaths in the state of California. We built two linear models to predict the number of cases and deaths given the number of days since the first case appeared, and both models obtained a high R^2 value indicating the performance of the linear models is good. Finally, we have identified the biases present in the underlying dataset.