DTSA Covid Report

Brent Lockee

6/6/2021

Covid-19 Report and Analysis

Data Import and initial cleaning

The data comes from Johns Hopkins github page and much of this initial loading and transformation is taken from the week three lecture series with permission from the project overview text. Full details on the origin and handling of the data set are available in the github repo.

In this report, I will focus on the data from the USA in an effort to provide a slightly more in-depth analysis.

```
cases <- read.csv("https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/c
deaths <- read.csv("https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/
```

```
us_cases <- cases %>%
  pivot_longer(cols = -c(UID, iso2, iso3, code3, FIPS, Admin2, Province_State, Country_Region, Lat, Long
  names_to = "date",
  values_to = "cases") %>%
  mutate(date = mdy(gsub('^.', '', date))) %>%
  select(-c(Lat, Long_, UID, iso2, iso3, code3, FIPS))

us_deaths <- deaths %>%
  pivot_longer(cols = -c(UID, iso2, iso3, code3, FIPS, Admin2, Province_State, Country_Region, Lat, Long
  names_to = "date",
  values_to = "deaths") %>%
  mutate(date = mdy(gsub('^.', '', date))) %>%
  select(-c(Lat, Long_, UID, iso2, iso3, code3, FIPS))
```

```
us <- us_cases %>%
  full_join(us_deaths)
```

```
## Joining, by = c("Admin2", "Province_State", "Country_Region", "Combined_Key", "date")
```

```
summary(us)
```

```
##     Admin2          Province_State     Country_Region     Combined_Key
##  Length:1674342     Length:1674342     Length:1674342     Length:1674342
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
```

```
## 
## 
##       date                cases           Population          deaths
##  Min.   :2020-01-22   Min.   :      0   Min.   :       0   Min.   :    0.00
##  1st Qu.:2020-05-26   1st Qu.:     13   1st Qu.:    9917   1st Qu.:    0.00
##  Median :2020-09-28   Median :    350   Median :   24892   Median :    6.00
##  Mean   :2020-09-28   Mean   :   3752   Mean   :   99604   Mean   :   76.33
##  3rd Qu.:2021-01-31   3rd Qu.:   1891   3rd Qu.:   64979   3rd Qu.:   37.00
##  Max.   :2021-06-05   Max.   :1244917   Max.   :10039107   Max.   :24400.00
```

Creating a Lagging New Case Variable

For my unique analysis and modeling, I'd like to see how predictive a lagging new case variable is on deaths. The code below creates new case and new death variables by subtracting each from the prior days respective value. From there, new case

According to the CDC, the median time from onset of illness to ICU admission was 9.5 - 12 days with the median hospital stay lasting 10 -13 days. This disease progression data informed my choice of using average daily case rates from 30 to 10 days prior.

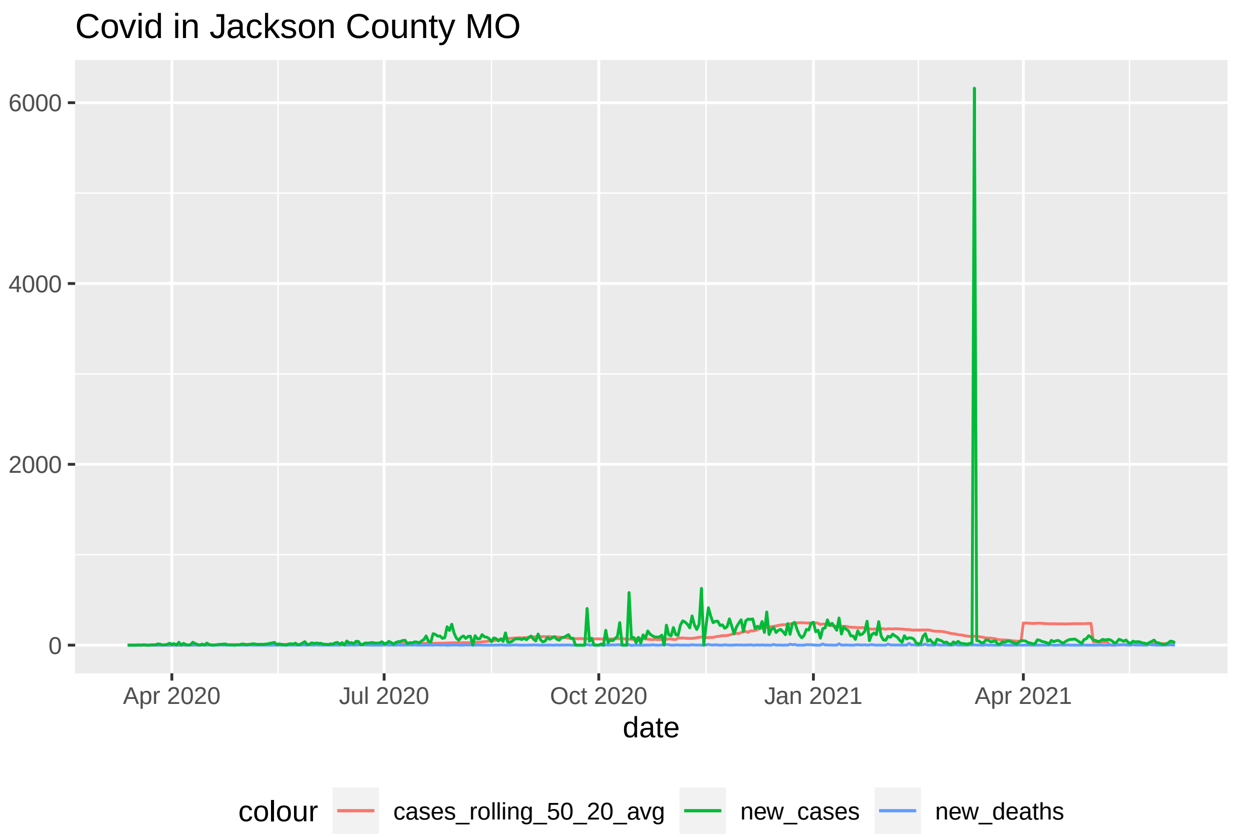
More information availble here: https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-guidance-management-patients.html

```
setDT(us)
setkeyv(us, c("Combined_Key", "date"))

us[, new_cases:= cases - shift(roll_sumr(cases, n=1)), by=Combined_Key]
us[, new_deaths:= deaths - shift(roll_sumr(deaths, n=1)), by=Combined_Key]

us[, cases_10:=shift(roll_sumr(new_cases, n=10)), by=Combined_Key]
us[, cases_30:=shift(roll_sumr(new_cases, n=30)), by=Combined_Key]
us[, cases_20:=shift(roll_sumr(new_cases, n=20)), by=Combined_Key]
us[, cases_50:=shift(roll_sumr(new_cases, n=50)), by=Combined_Key]

us <- us %>%
  mutate(cases_rolling_30_10_avg = (cases_30 - cases_10)/ 20) %>%
  mutate(cases_rolling_50_20_avg = (cases_50 - cases_20)/ 30) %>%
  filter(new_cases > -1)

summary(us)
```

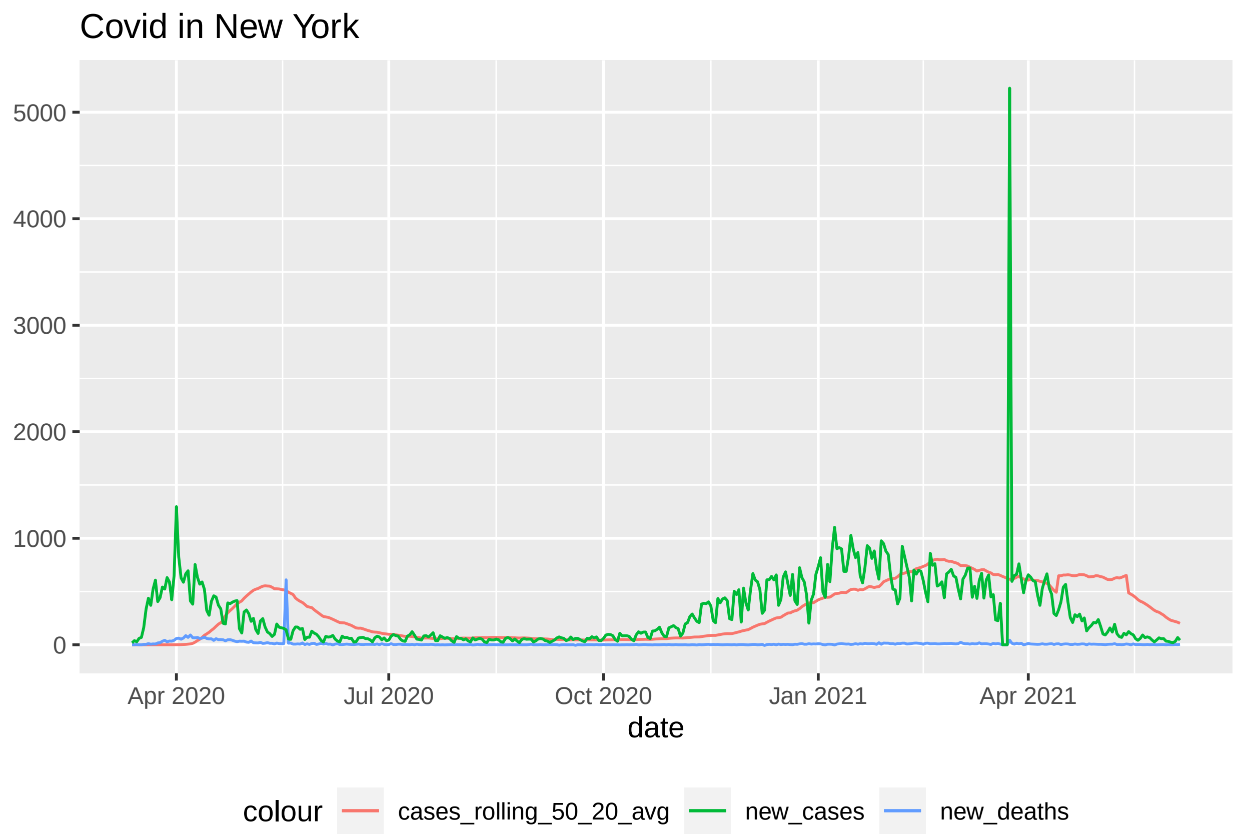
```
##     Admin2          Province_State     Country_Region     Combined_Key
##  Length:1644003     Length:1644003     Length:1644003     Length:1644003
##  Class  :character  Class  :character  Class  :character  Class  :character
##  Mode   :character  Mode   :character  Mode   :character  Mode   :character
## 
## 
## 
## 
##       date                cases           Population          deaths
##  Min.   :2020-01-23   Min.   :      0   Min.   :       0   Min.   :    0.00
##  1st Qu.:2020-05-25   1st Qu.:     12   1st Qu.:   10012   1st Qu.:    0.00
##  Median :2020-09-28   Median :    350   Median :   25069   Median :    6.00
##  Mean   :2020-09-27   Mean   :   3790   Mean   :  100633   Mean   :   76.97
##  3rd Qu.:2021-01-30   3rd Qu.:   1904   3rd Qu.:   65435   3rd Qu.:   37.00
##  Max.   :2021-06-05   Max.   :1244917   Max.   :10039107   Max.   :24400.00
##
##    new_cases          new_deaths           cases_10           cases_30
##  Min.   :    0.00   Min.   :-3962.000   Min.   :-69520.0   Min.   :-69520
##  1st Qu.:    0.00   1st Qu.:    0.000   1st Qu.:     2.0   1st Qu.:     9
##  Median :    1.00   Median :    0.000   Median :    21.0   Median :    76
##  Mean   :   20.59   Mean   :    0.362   Mean   :   205.7   Mean   :   639
##  3rd Qu.:    9.00   3rd Qu.:    0.000   3rd Qu.:   104.0   3rd Qu.:   337
##  Max.   :29423.00   Max.   : 1553.000   Max.   :151509.0   Max.   :431295
##                                         NA’s   :33420      NA’s   :100260
##     cases_20           cases_50      cases_rolling_30_10_avg
##  Min.   :-69520.0   Min.   :-69520   Min.   :-3476.00
##  1st Qu.:     5.0   1st Qu.:    22   1st Qu.:    0.20
##  Median :    47.0   Median :   147   Median :    2.35
##  Mean   :   418.7   Mean   :  1100   Mean   :   21.22
##  3rd Qu.:   217.0   3rd Qu.:   602   3rd Qu.:   11.00
##  Max.   :298035.0   Max.   :634681   Max.   :14901.75
##  NA’s   :66840      NA’s   :167089   NA’s   :100260
##  cases_rolling_50_20_avg
##  Min.   :-2317.33
##  1st Qu.:    0.27
##  Median :    2.57
##  Mean   :   21.75
##  3rd Qu.:   11.47
##  Max.   :14376.50
##  NA’s   :167089
```

Visualizing and Validating the Lagging Variable

```
us %>%
  filter(Combined_Key == "Jackson, Missouri, US" & !is.na(cases_rolling_50_20_avg)) %>%
  ggplot(aes(x = date, y = new_cases)) +
  geom_line(aes(y = cases_rolling_50_20_avg, color = "cases_rolling_50_20_avg")) +
  geom_line(aes(y = new_deaths, color = "new_deaths")) +
  geom_line(aes(color = "new_cases")) +
  theme(legend.position = "bottom") +
  labs(title = "Covid in Jackson County MO", y = NULL)
```



```
us %>%
  filter(Combined_Key == "New York, New York, US" & !is.na(cases_rolling_50_20_avg)) %>%
  ggplot(aes(x = date, y = new_cases)) +
  geom_line(aes(y = cases_rolling_50_20_avg, color = "cases_rolling_50_20_avg")) +
  geom_line(aes(y = new_deaths, color = "new_deaths")) +
  geom_line(aes(color = "new_cases")) +
  theme(legend.position = "bottom") +
  labs(title = "Covid in New York", y = NULL)
```



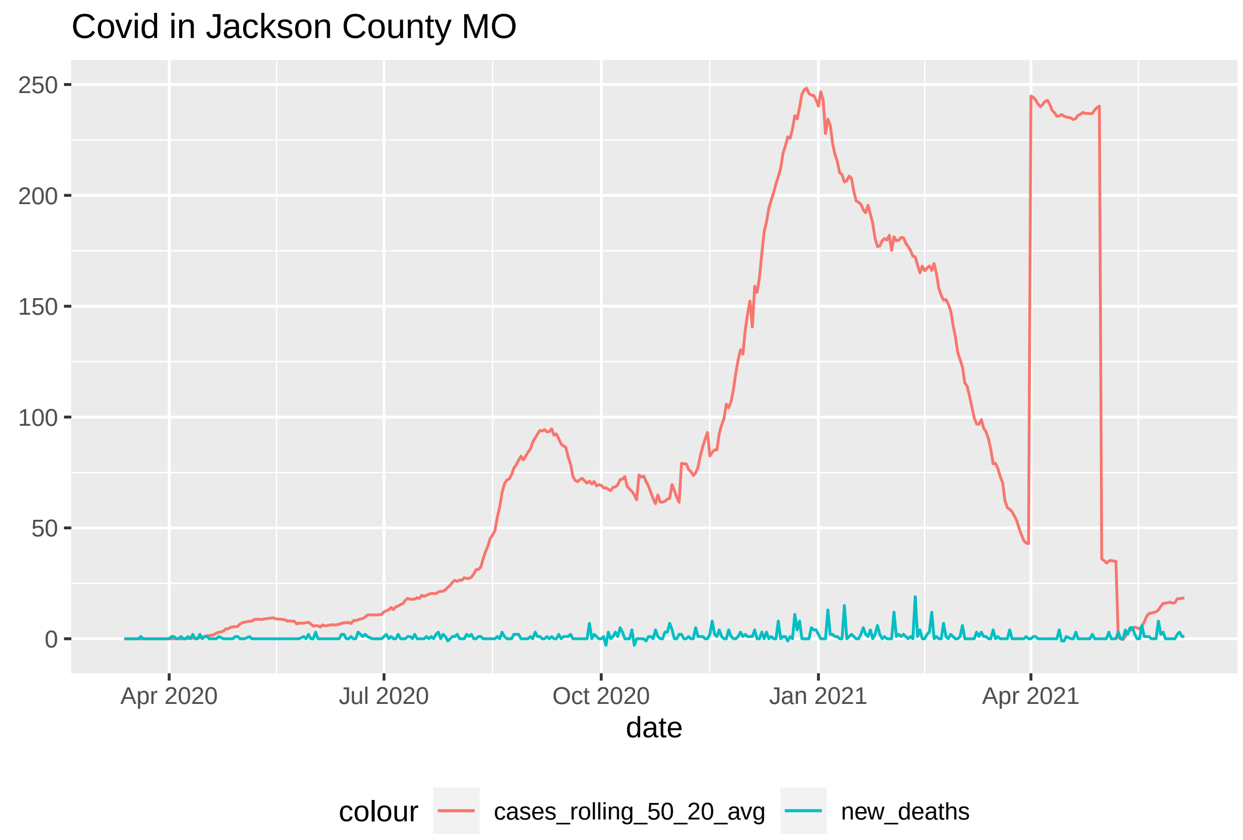
The lagging variable graphs as expected - it tracks the new_cases variable but shifted forward by 30 days. However, each location has one or two days with a much higher new case number than its neighbors. I suspect those are data dump days when then case count catches up over a holiday or some other reporting delay. You can see the problematic lines for each location below.

```
filter(us, Combined_Key == "New York, New York, US" & !is.na(cases_rolling_50_20_avg) & new_cases > 300
```

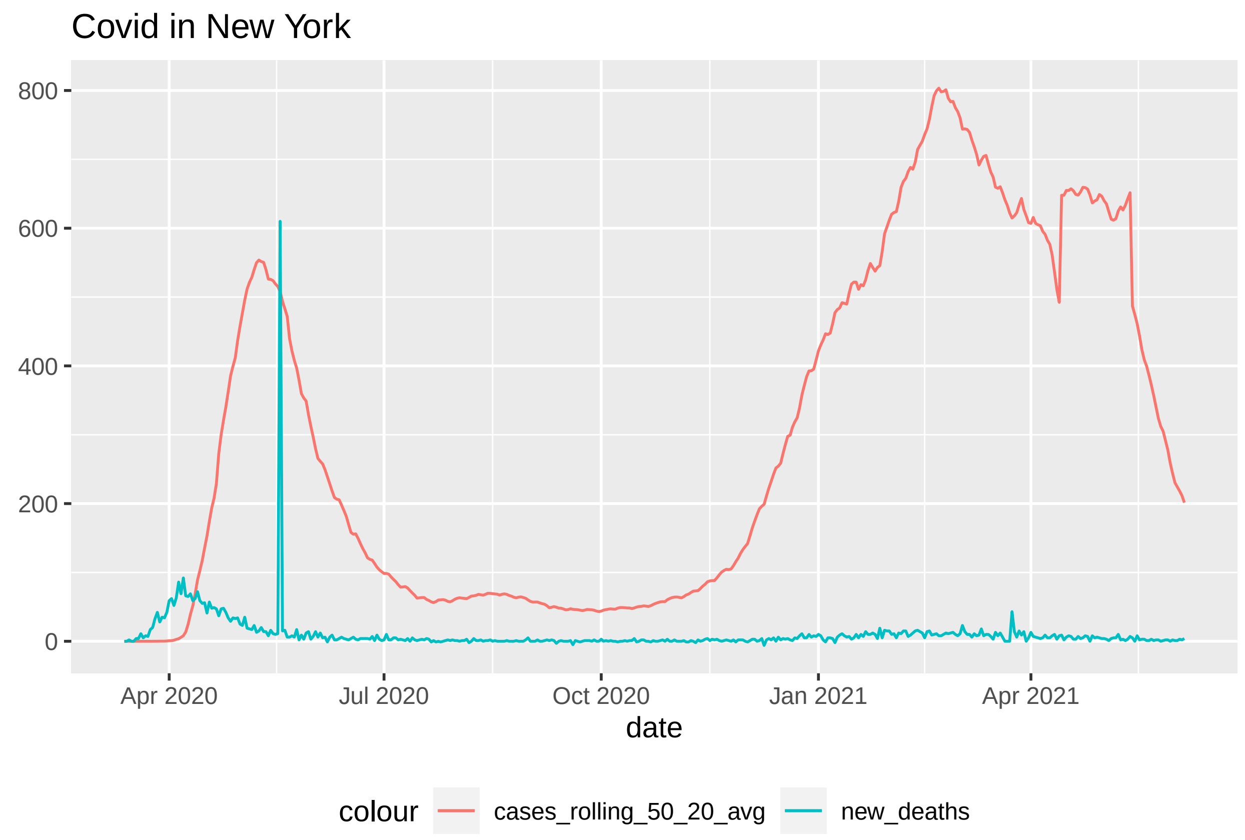
```
##      Admin2 Province_State Country_Region           Combined_Key       date
## 1: New York       New York             US New York, New York, US 2021-03-24
##     cases Population deaths new_cases new_deaths cases_10 cases_30 cases_20
## 1: 118986    1628706   4146      5226         43     3033    14771     8886
##    cases_50 cases_rolling_30_10_avg cases_rolling_50_20_avg
## 1:    27329                   586.9                614.7667
```

Repeating the above graphs without the new cases variable that makes the scale difficult to read.

```
us %>%
  filter(Combined_Key == "Jackson, Missouri, US" & !is.na(cases_rolling_50_20_avg)) %>%
  ggplot(aes(x = date, y = cases_rolling_50_20_avg)) +
  geom_line(aes(color = "cases_rolling_50_20_avg")) +
  geom_line(aes(y = new_deaths, color = "new_deaths")) +
  theme(legend.position = "bottom") +
  labs(title = "Covid in Jackson County MO", y = NULL)
```



```
us %>%
  filter(Combined_Key == "New York, New York, US" & !is.na(cases_rolling_50_20_avg)) %>%
  ggplot(aes(x = date, y = cases_rolling_50_20_avg)) +
  geom_line(aes(y = cases_rolling_50_20_avg, color = "cases_rolling_50_20_avg")) +
  geom_line(aes(y = new_deaths, color = "new_deaths")) +
  theme(legend.position = "bottom") +
  labs(title = "Covid in New York", y = NULL)
```



Creating a Model to Predict New Daily Deaths

```
mod <- lm(new_deaths ~ cases_rolling_50_20_avg, data=us)
summary(mod)
```

```
##
## Call:
## lm(formula = new_deaths ~ cases_rolling_50_20_avg, data = us)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3962.1    -0.2    -0.1    -0.1  1534.5
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             7.919e-02  4.039e-03   19.61   <2e-16 ***
## cases_rolling_50_20_avg 1.486e-02  3.259e-05  455.81   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.832 on 1476912 degrees of freedom
##   (167089 observations deleted due to missingness)
## Multiple R-squared:  0.1233, Adjusted R-squared:  0.1233
## F-statistic: 2.078e+05 on 1 and 1476912 DF,  p-value: < 2.2e-16
```

Comparing to a model that doesn’t look so far back - using the 30 to 10 day window instead.

```
mod <- lm(new_deaths ~ cases_rolling_30_10_avg, data=us)
summary(mod)
```

```
##
## Call:
## lm(formula = new_deaths ~ cases_rolling_30_10_avg, data = us)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3962.1    -0.2    -0.1    -0.1  1542.8
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)             6.261e-02  3.838e-03   16.31   <2e-16 ***
## cases_rolling_30_10_avg 1.519e-02  3.106e-05  489.18   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 4.698 on 1543741 degrees of freedom
##   (100260 observations deleted due to missingness)
## Multiple R-squared:  0.1342, Adjusted R-squared:  0.1342
## F-statistic: 2.393e+05 on 1 and 1543741 DF,  p-value: < 2.2e-16
```

The rolling case average from 50 to 20 days ago and 30 to 10 days ago are both highly significant when predicting daily new deaths for a location. Interestingly, each model only achieves an R squared of ~0.13. The model using the 30 to 10 days prior window performed slightly better as a whole. Even in our two locations, the new deaths data did seem highly variable from day to day, making it harder to predict.

Testing the addition of a county’s population to see if it has any impact on the model accuracy. It seems plausible that large and small counties may have different mortality rates.

```
mod <- lm(new_deaths ~ cases_rolling_30_10_avg + Population, data=us)
summary(mod)
```

```
##
## Call:
## lm(formula = new_deaths ~ cases_rolling_30_10_avg + Population,
##     data = us)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3962.0    -0.1     0.0     0.0  1541.0
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -2.324e-02  3.947e-03  -5.887 3.94e-09 ***
## cases_rolling_30_10_avg  1.305e-02  3.923e-05 332.527  < 2e-16 ***
## Population               1.305e-06  1.463e-08  89.179  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
```

```
## Residual standard error: 4.686 on 1543740 degrees of freedom
##   (100260 observations deleted due to missingness)
## Multiple R-squared:  0.1386, Adjusted R-squared:  0.1386
## F-statistic: 1.242e+05 on 2 and 1543740 DF,  p-value: < 2.2e-16
```

The population variable is also significant but does not much to the R-squared value of the model.

Next Steps, Bias Sources and Conclusion

The spike days of massive jumps in new cases (and deaths) may be causing issues with the analysis and may need to be removed. However, if they represent legitimate but delayed data, it may be proper to keep them in the dataset given that this analysis smooths the big jump in cases out over the rolling average windows.

These large jumps in new cases and deaths likely represent reporting biases, or at least delays. This data has been collected from a huge variety of institutions and there were many real world challenges in getting and aggregating this data. Even more fundamentally, cases during the beginning of the pandemic were likely under-reported due to lack of testing. Finally, the decision to use 30-10 and 50-20 lagging case windows could include some personal bias. I made what I felt was a logical, defensible decision based on the CDC website on Covid progression.

With rolling average daily new case counts and deaths lined up like this, a mortality analysis over time and by location would also be a logical next step.

This analysis has transformed the Covid data to show daily new case and death counts and then used those to create rolling, lagging new case averages. From there, models were created to assess the predictive power of those lagging case averages. The rolling new case average from 30 to 10 days prior turned out to be slightly more predictive than the 50 to 20 day window.