Steven Taylor

ADEC7460.02 Predictive Analytics & Forecasting Midterm

December 11, 2021

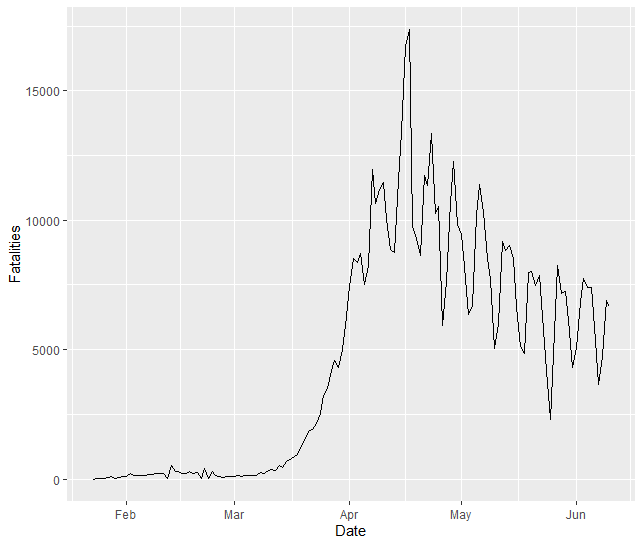
**Introduction**

COVID-19 has spread across the globe for nearly two years. As of December 1, 2021 COVID-19 is responsible for 780,131 deaths in the United States alone. Globally, the death count stands at 5,215,556. My goal is to use the available dataset to forecast COVID-19 deaths moving forward.

Not only has this virus proven deadly, but the response to this virus has dampened the health and well-being of those not infected with COVID-19. To date, the COVID-19 pandemic has ignited the deepest [global recession] since the end of World War II” (Brookings, 2021). This forecasting effort should be used as a resource to decision makers as they grapple with the difficult decisions that coincide with leading through a pandemic.

**Analysis and Results**

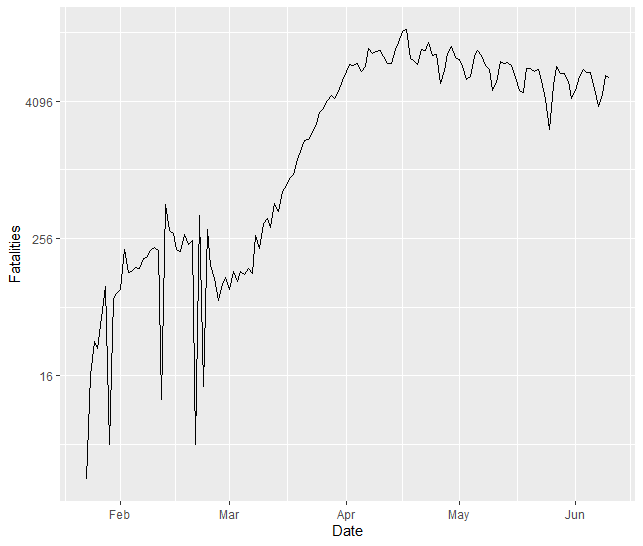
Our data consists of nearly one million rows of case and fatality counts, by country. The earliest data begins January 23, 2020 and runs through June 10, 2020. Upon receipt of the data I unpivoted the case and fatality variables to create unique columns for case and fatality counts. I then performed exploratory analysis to understand the variables and data completeness. The plot below indicates our fatality counts, by date, from January through June of 2020.



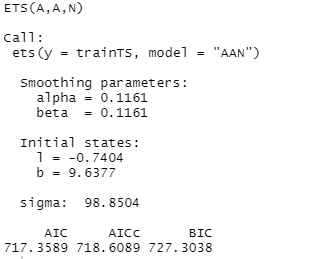
Early in 2020 the global COVID-19 fatalities were consistently low, hovering near zero. Half way through March, 2020 we see a steep increase in COVID-19 fatalities. This trend continues upward until our peak fatalities reached ~17,500 in the middle of April. From this date, the fatality counts have been on a steady downward trend.

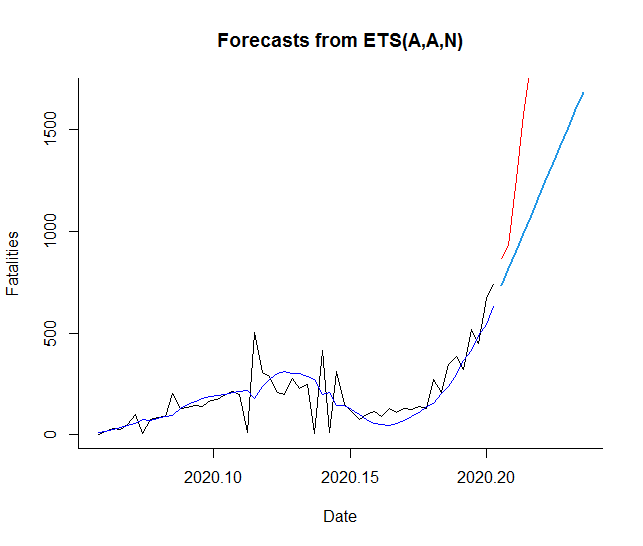
I applied several models and/or transformations to the data set in an attempt to create a model that seems reasonable and has a low RMSE.

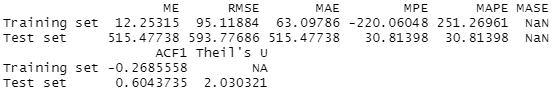
*Log transformation:* Taking the log of a variable will effectively change the fatality count from a unit change to a percent change. This is especially important when using medium to large datasets. Given the size of our dataset this transformation is worth exploring. When we apply a log transformation to the COVID-19 fatality data we notice high volatility in the early stages of 2020. This volatility is lost in the earlier plot due to the size of data were working with. Another divergence is the downtrend after our peak in April, 2020. This transformation indicates that the downtrend occurs at a much slower rate than the earlier chart indicates.



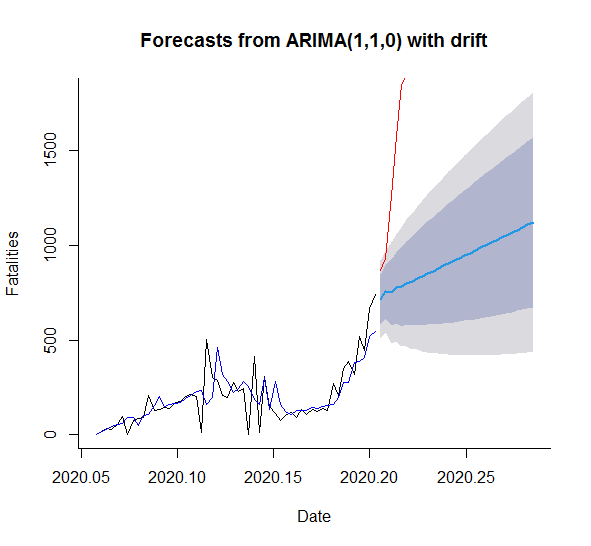
*ETS AAN:* ETS models work well with non-seasonal data, which is what we expect the COVID-19 data to represent. The ETS model produces a smoothing parameter for the slope which is estimated to be essentially zero, indicating that the trend is not changing over time. This model forecasts that COVID-19 fatalities will continue their upward trend. The RMSE on our training data is 95.1.

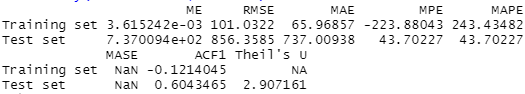




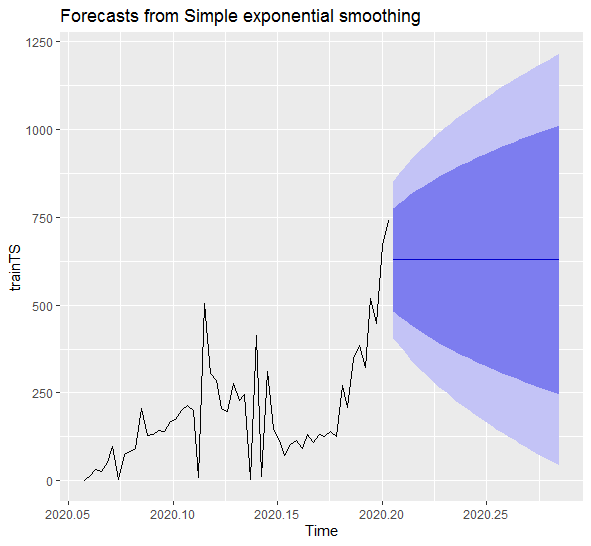


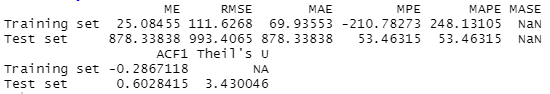
*Auto Arima:* ARIMA models aim to describe the autocorrelations in the data. The ARIMA forecast below indicates an increase in COVID-19 fatalities but the slope is more modest than our earlier ETS model. Unfortunately, the RMSE is 101 indicating this model is less precise than our earlier ETS model.





*Simple Exponential Smoothing (ses):* The simple exponential smoothing method is another that is designed for non-seasonal data. This method produces a flat forecast. From the chart below we see the expected fatalities to be near 675 fatalities. Our ses model has the largest RMSE at 111 indicating this is our least reliable model.





**Discussion**

There are several limitations occurring in this exercise. The first limitation is simply the duration of our data. The pandemic has been going on for nearly two years but we’re forecasting based on five months of data. We’re also missing any ancillary variables that may be impacting our COVID-19 volume. For example, average age of a population may be a large contributor to fatalities.

Another limitation is the completeness of our data. Although we have case and fatality counts down to the county level, many of these fields are not completely filled in (blanks). Further, there appears to be duplication or inconsistency in the geography. For example, US has data by state and county but there are also rows of US data where no state or county is indicated and the volume is unusual given the other rows of data.

Moving forward, we can continue to perform these forecasts as more data becomes available. The expectation is that the data will continue to grow in size and we’ll be able to have more confidence in the forecast. I would also like to find ancillary time series data sets to coincide with COVID-19 volume to see what other factors may be impacting cases and fatalities.

**Sources**

COVID Data Tracker. Centers for Disease Control and Prevention.

<https://covid.cdc.gov/covid-data-tracker/#datatracker-home>

Global COVID-19 Tracker. Kaiser Family Foundation.

<https://www.kff.org/coronavirus-covid-19/issue-brief/global-covid-19-tracker/>

Social and economic impact of COVID-19. Brookings.

<https://www.brookings.edu/research/social-and-economic-impact-of-covid-19/>