Mosharraf Hossain
Time Series Analysis (STAT 460)

Forecasting Project

Abstract:

In this project, I wanted to analyze various forecasting methods. First, I performed Auto Regressive Integrated Moving Average (ARIMA), then Holt-Winters smoothing process to see what our forecasting model says. Furthermore, I performed Two-Sided Moving Average to see which one smooths better. I compared the model with that of CDC to visually analyze the accuracy of the model. Our forecasting model was quite accurate at predicting weekly deaths as well as daily deaths given by CDC. The only major difference was the confidence interval. While CDC had smaller 95% and 5% values, my forecasting model had very high confidence interval.

Background:

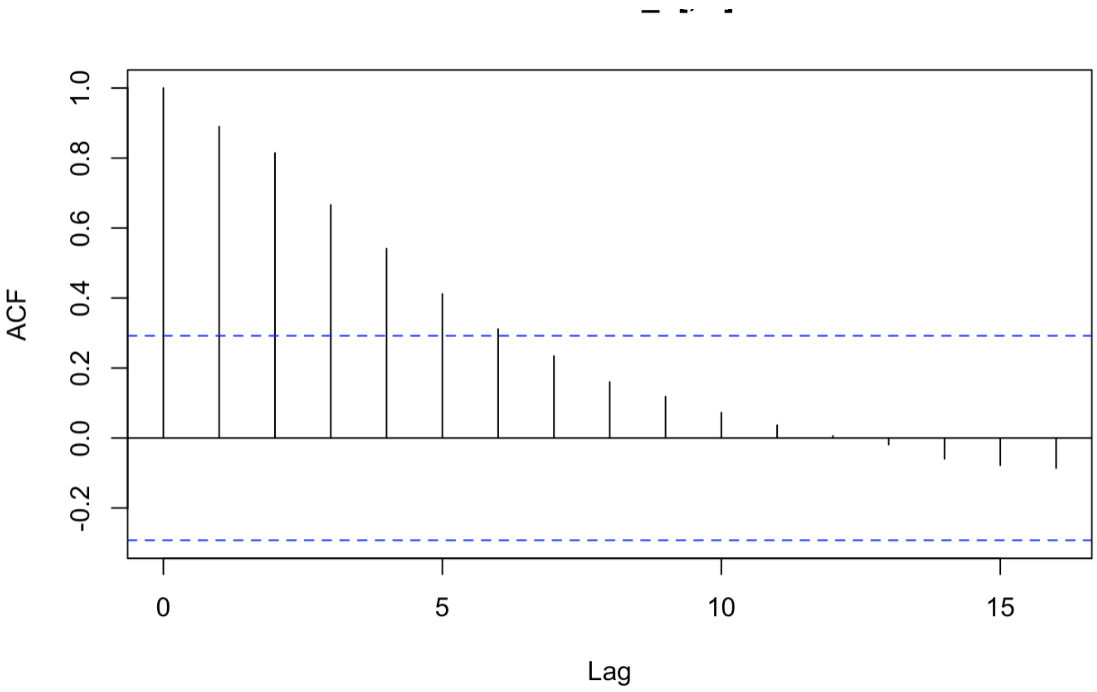
We are currently seeing the largest pandemic in our lifetime and during the beginning of 21st century. There has been SARS outbreak in 2002-2004, H1N1/09 pandemic in 2009 but none of the pandemic or outbreak was as severe as COVID-19.

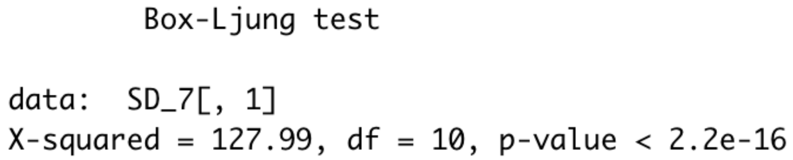
Objective:

My objective on this project is to compare various smoothing process and how the forecasting method works after we smooth. What ARIMA model to use and if we have seasonal pattern what to do.

Data and Methods:

I accumulated the data and converted into a time series object. The time series object helped me to find the Autocorrelation function which is crucial to determine whether the time series is stationary or non-stationary. Our ACF plot shows decreasing ACF as the lag increases which is typical of non-stationary time series and the Ljung-Box test shows that the p-value is less than 0.05, which cements our assumption that the time series is non-stationary.

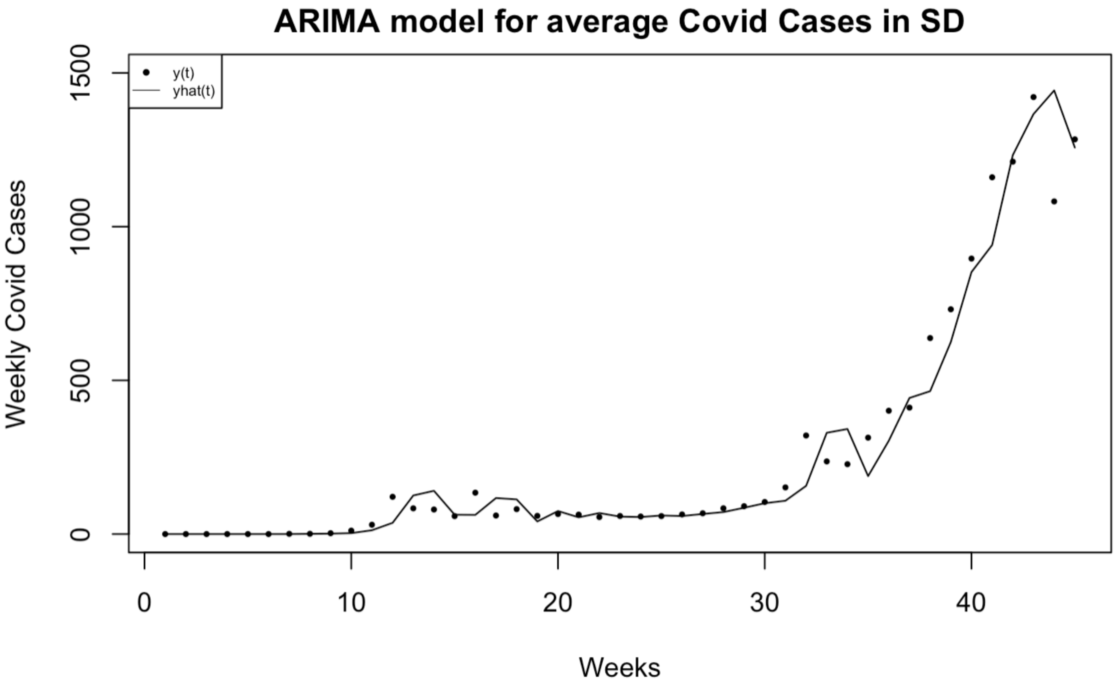


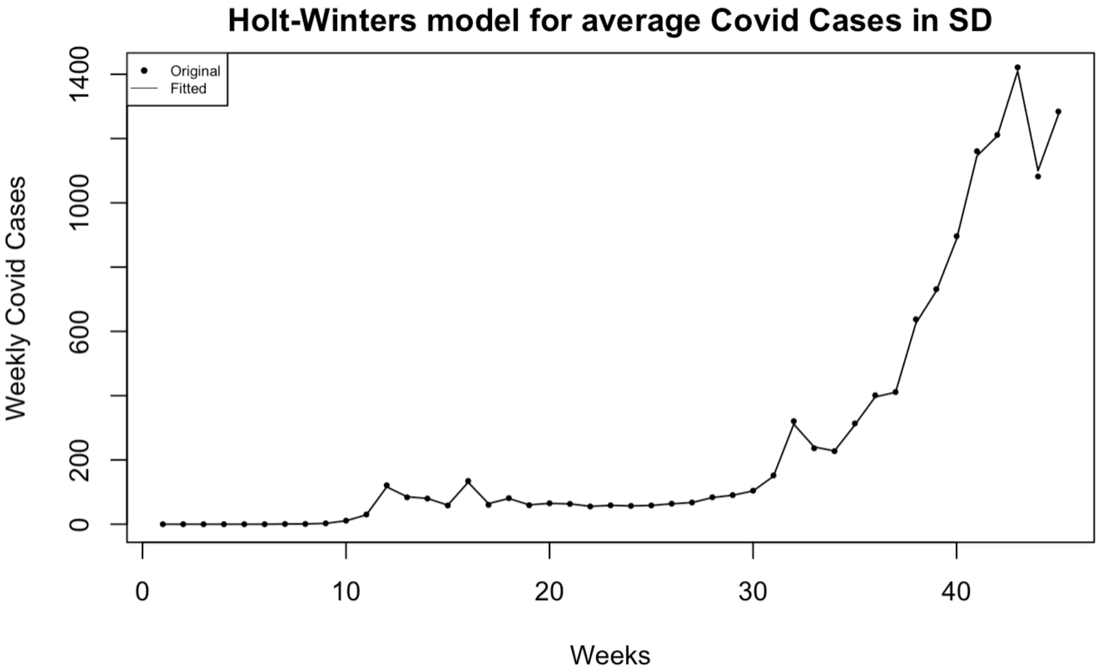

```
            Box-Ljung test

data:  SD_7[, 1]
X-squared = 127.99, df = 10, p-value < 2.2e-16
```

Upon figuring out the stationarity, I moved to fit ARIMA model. The best model (with the lowest Akaike Information Criterion) is ARIMA (2,2,1). Then I use Holt-Winters function which is another decomposition process. And I show both models on top of Weekly Covid case plot and it shows that the Holt-Winters have very high accuracy.

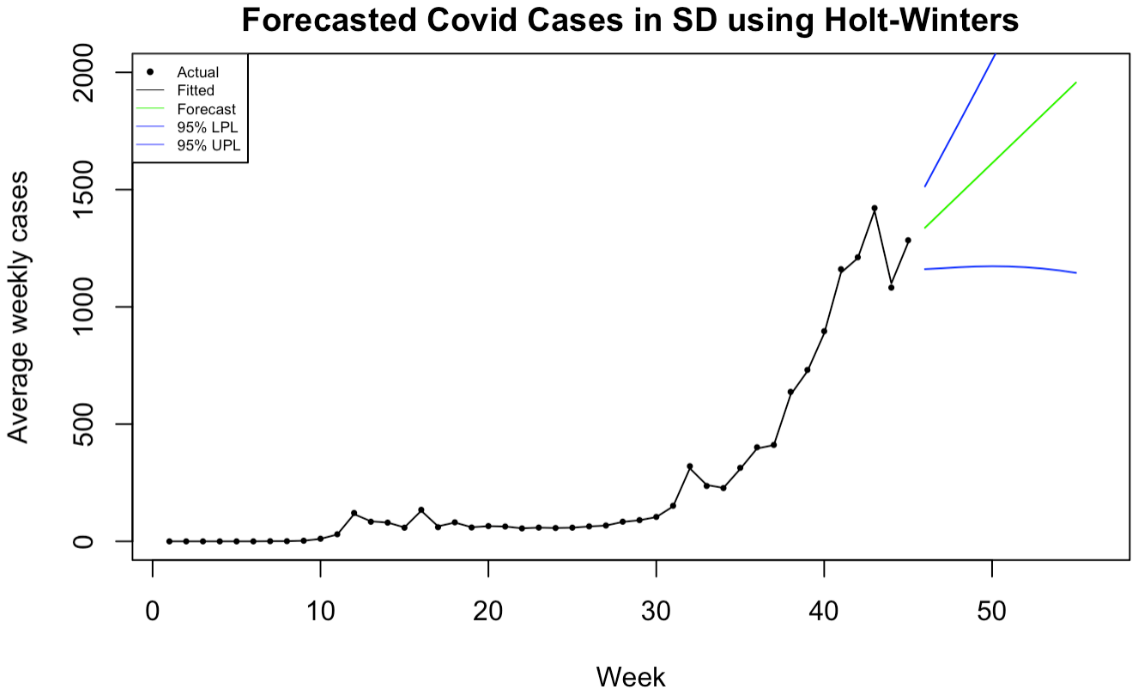
Mosharraf Hossain
Time Series Analysis (STAT 460)

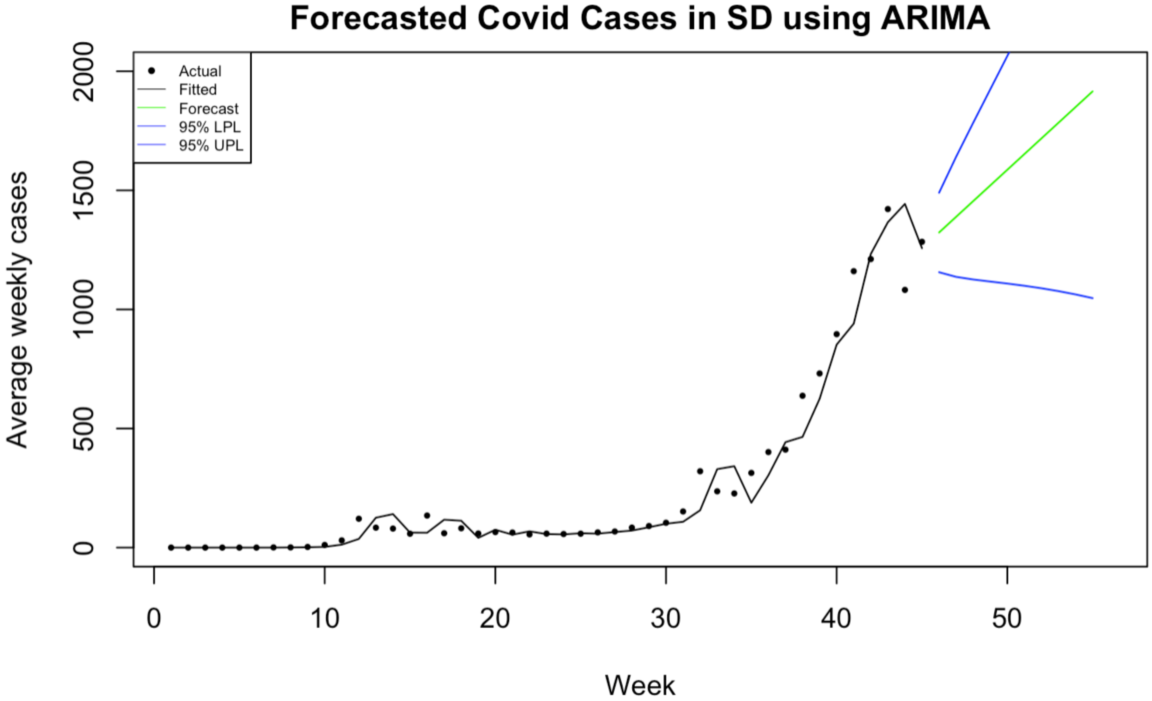




After that, I performed the forecast of 10 steps ahead, and ARIMA model shows that week 55 or mid-February, South Dakota will be reporting 1915 cases.

Mosharraf Hossain
Time Series Analysis (STAT 460)

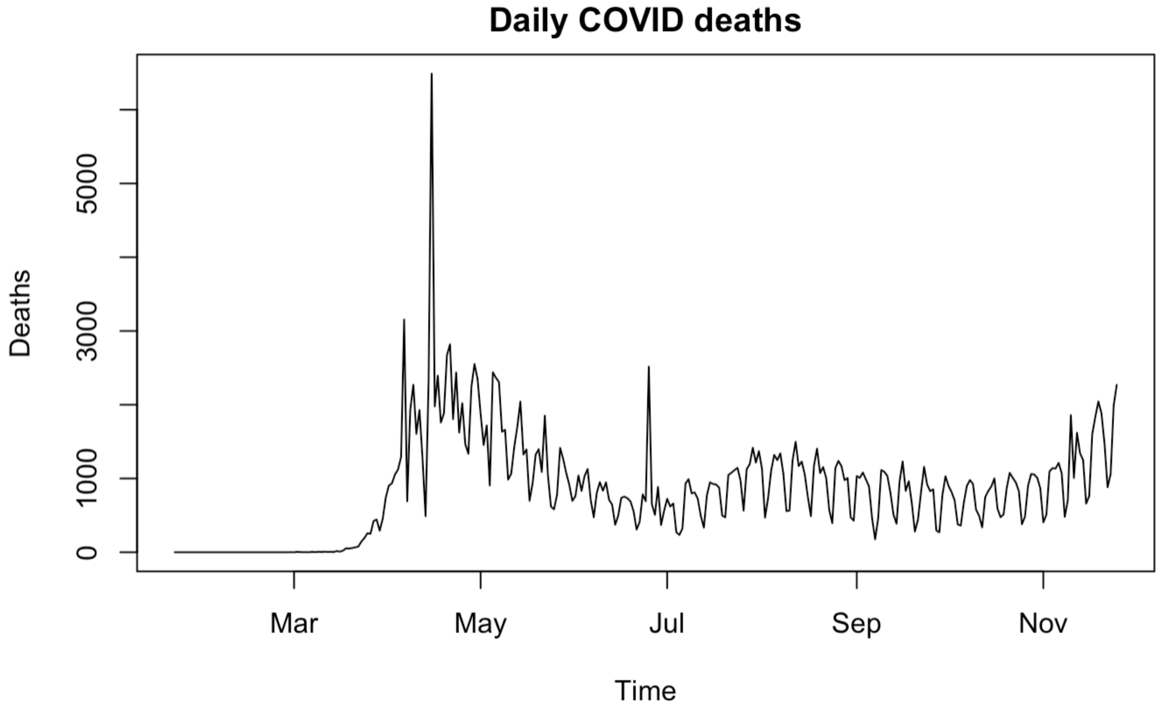




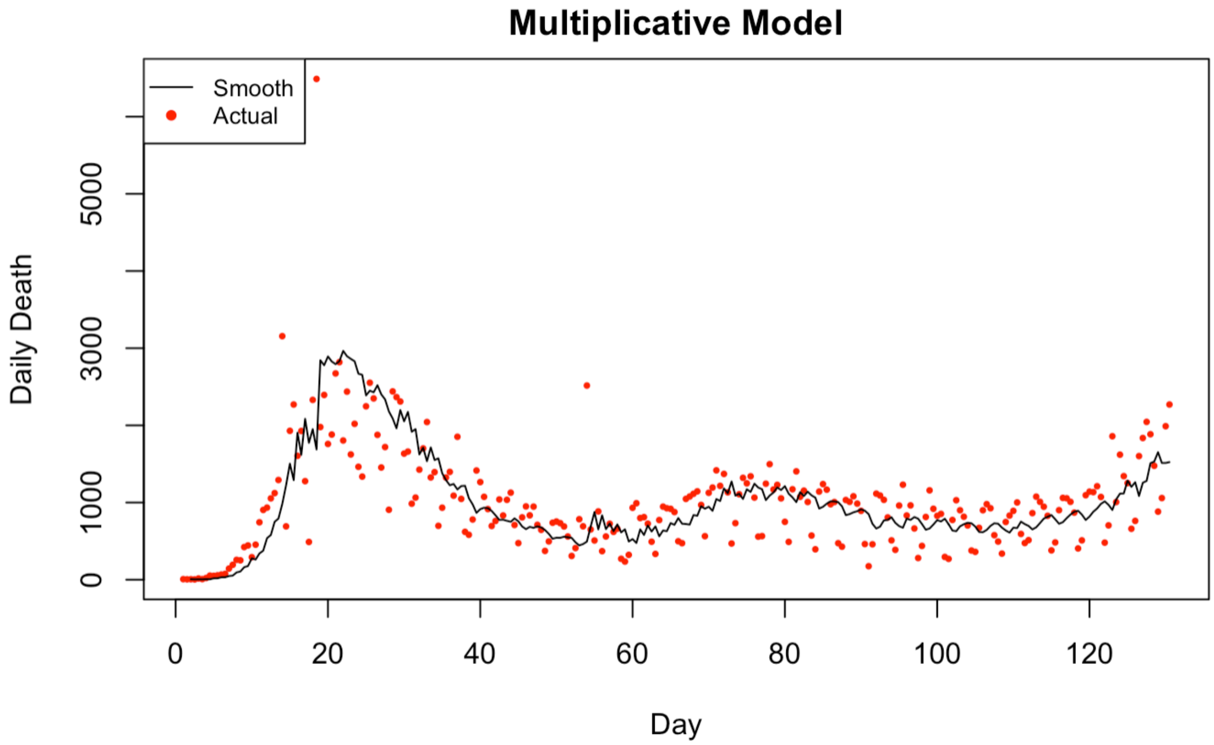
Both forecasting models showed similar forecasting. Holt Winters have a little bit steeper slop, which means Holt-Winters predicts there will be a little bit more cases than 1915 cases.

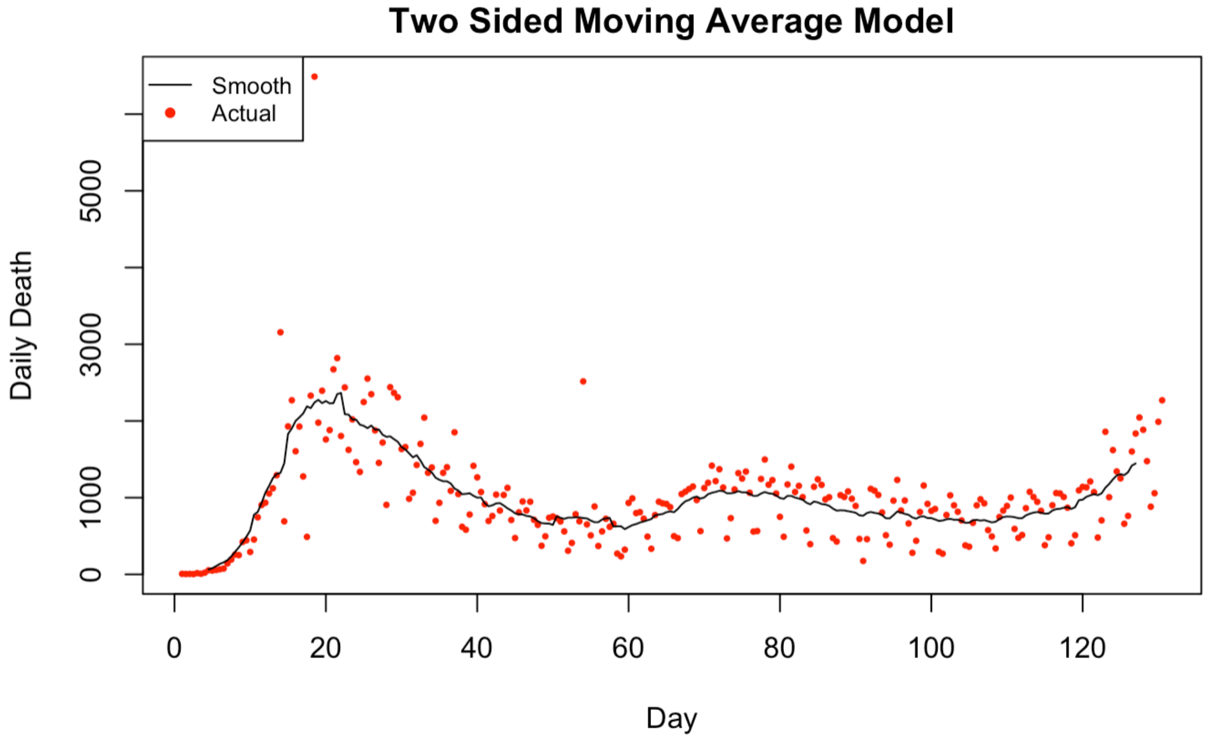
Mosharraf Hossain
Time Series Analysis (STAT 460)

For our next forecast, I used the CDC's daily death data. The plot looks like it has seasonal pattern with increasing trend and since it is actually a daily difference of total death data, and because of that, it is stationary time series.



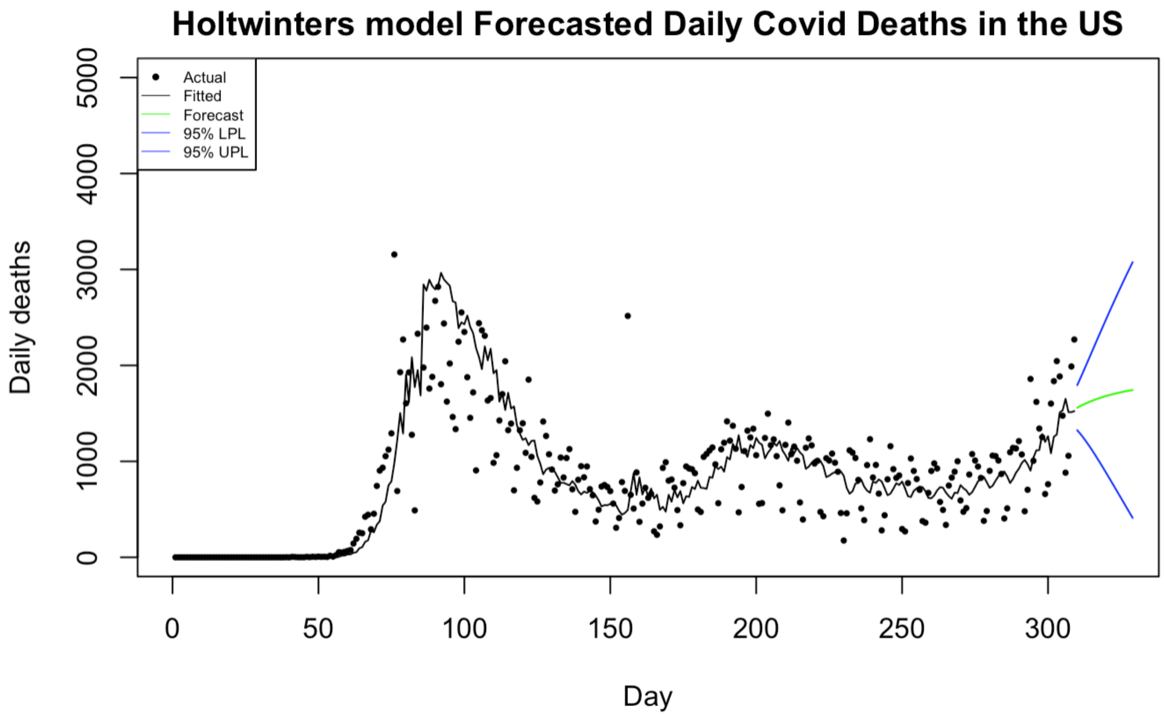
Which prompted me to use Holt-Winters' Multiplicative smoothing process [1]. This process smoothed the seasonality but two-sided weighted moving average (TS-MA) was able to smooth the seasonality much better.

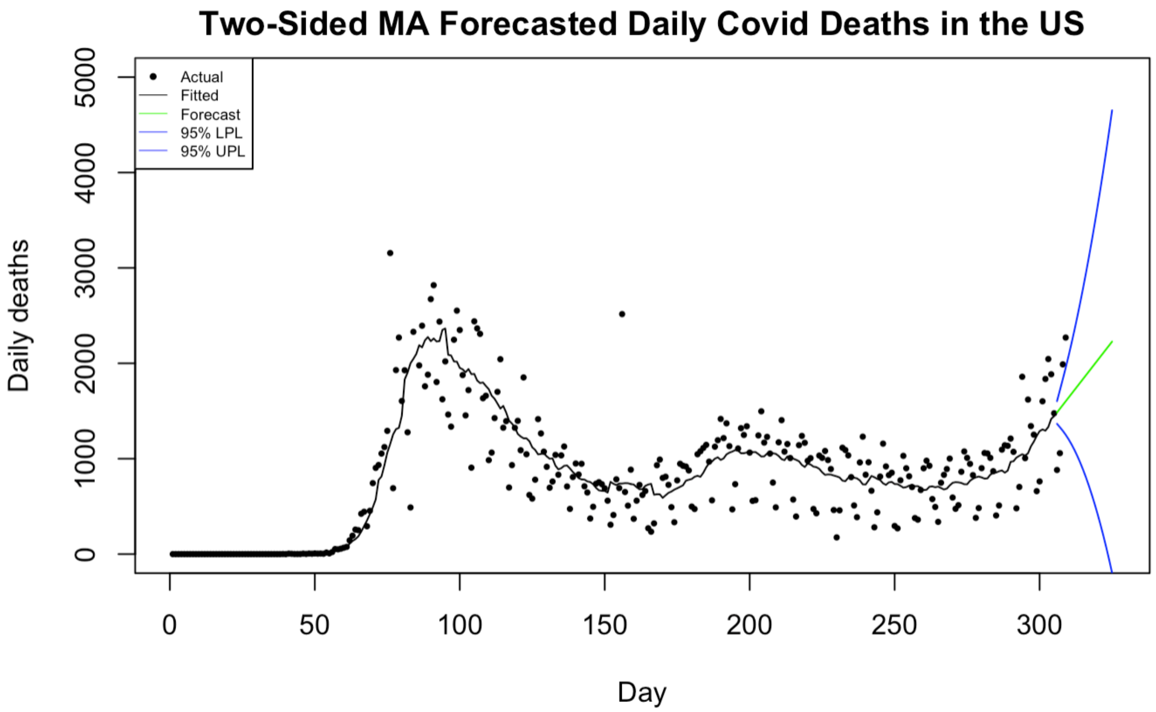




Therefore, I used the (TS-MA) process for forecasting accurately and the ARIMA model this time was ARIMA (1,1,1) [2].

Mosharraf Hossain
Time Series Analysis (STAT 460)





Mosharraf Hossain
Time Series Analysis (STAT 460)

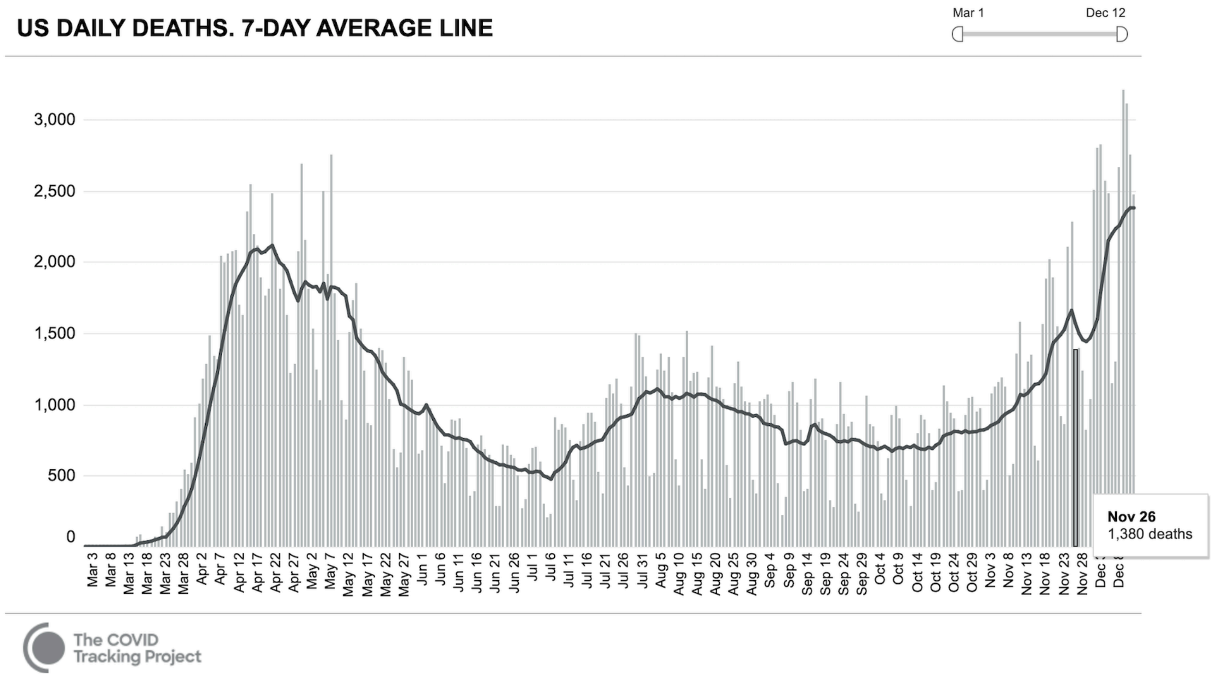
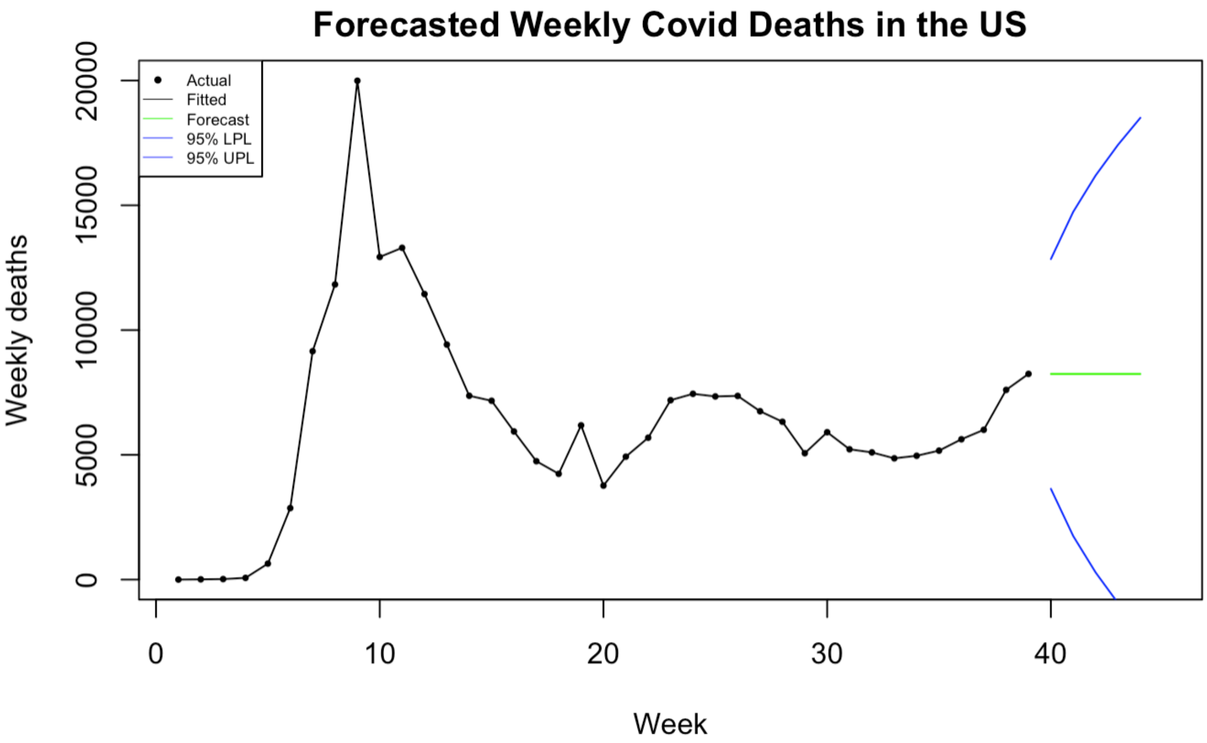


Figure: Actual Data for Daily Deaths till Dec 13th

When we look at the actual data, it shows that the slope is steeper after Nov. 28. We used data till November 28th for our 20 step ahead forecasting. Therefore, Two-Sided Moving Average did a good job at forecasting the COVID deaths and the slope seems to be pretty accurate too.

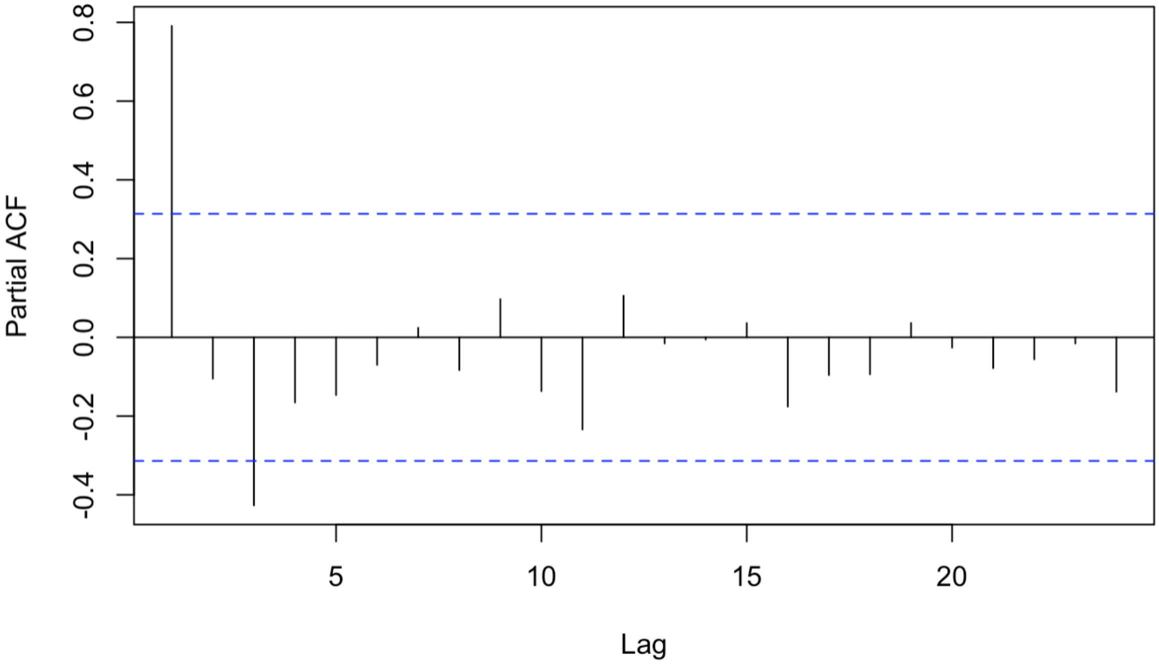
The next plots, I summarized weekly deaths and I wanted to see if the model can predict exactly like how CDC forecasts. When summarized and done a generalized forecasting model without any specification or model fit, the forecasting doesn't do a good job at it.

Mosharraf Hossain
Time Series Analysis (STAT 460)

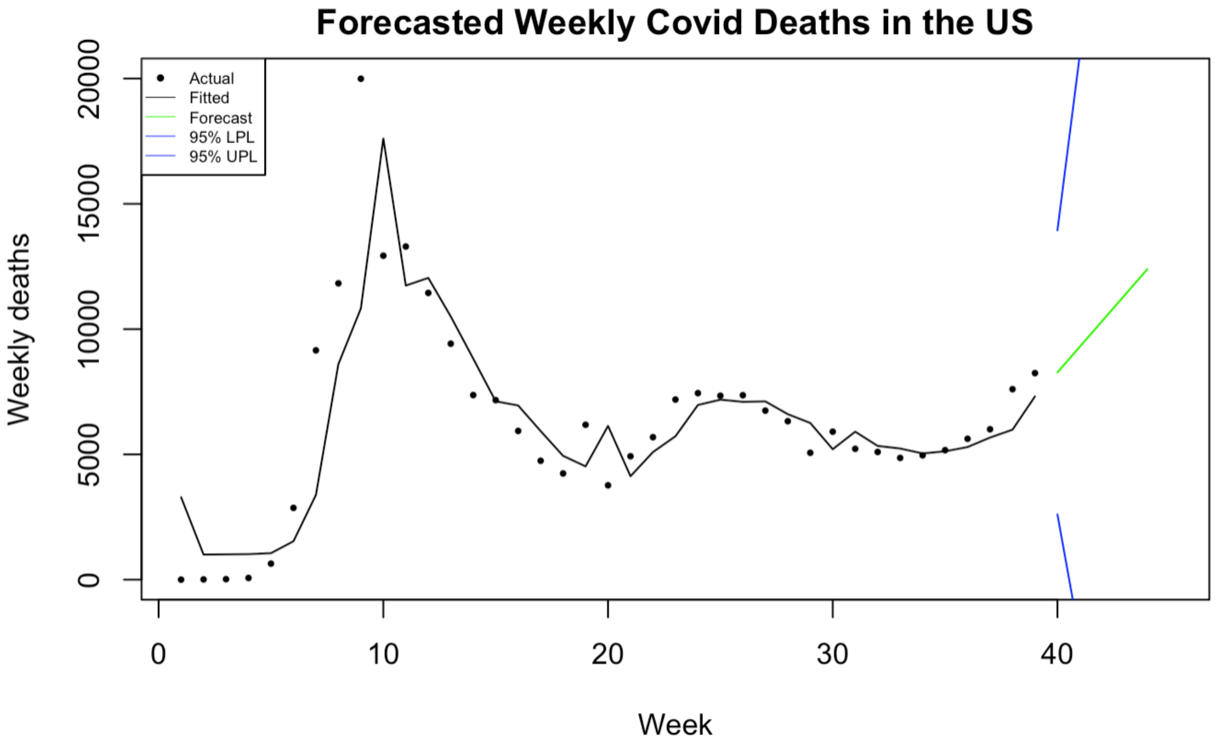


It only shows a straight line.

We look at the partial autocorrelation function of the weekly death data and it shows that one of the partial ACF lag spike is not significant.



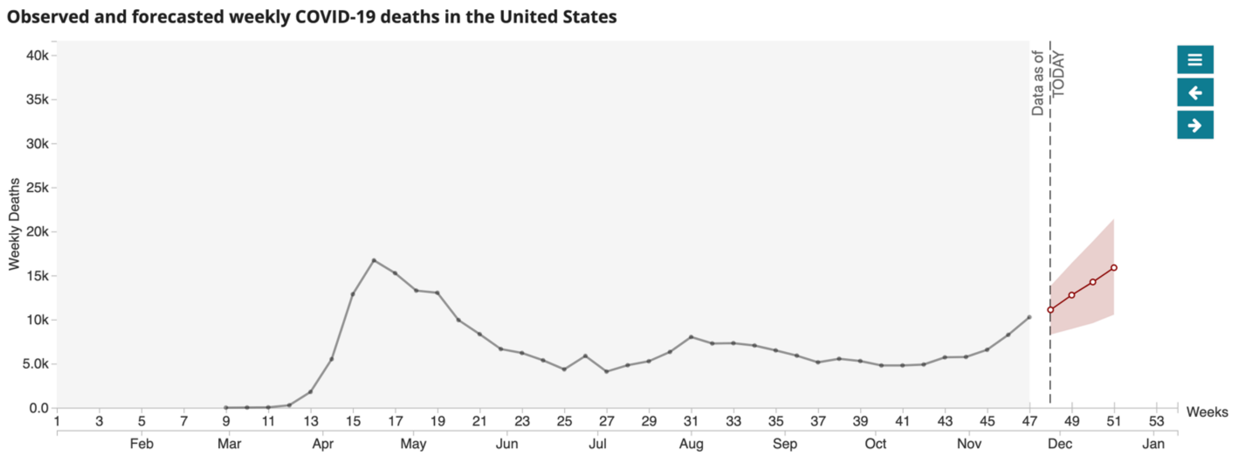
Which means, we will use AR(1) model and our auto.arima() function also shows that the best model is ARIMA (1,0,0) or AR(1). Therefore, we fit the AR(1) model which smooths the graph and we use the forecasting model. The forecasting model shows upward linear trend, and it is much more gullible than what we had before.



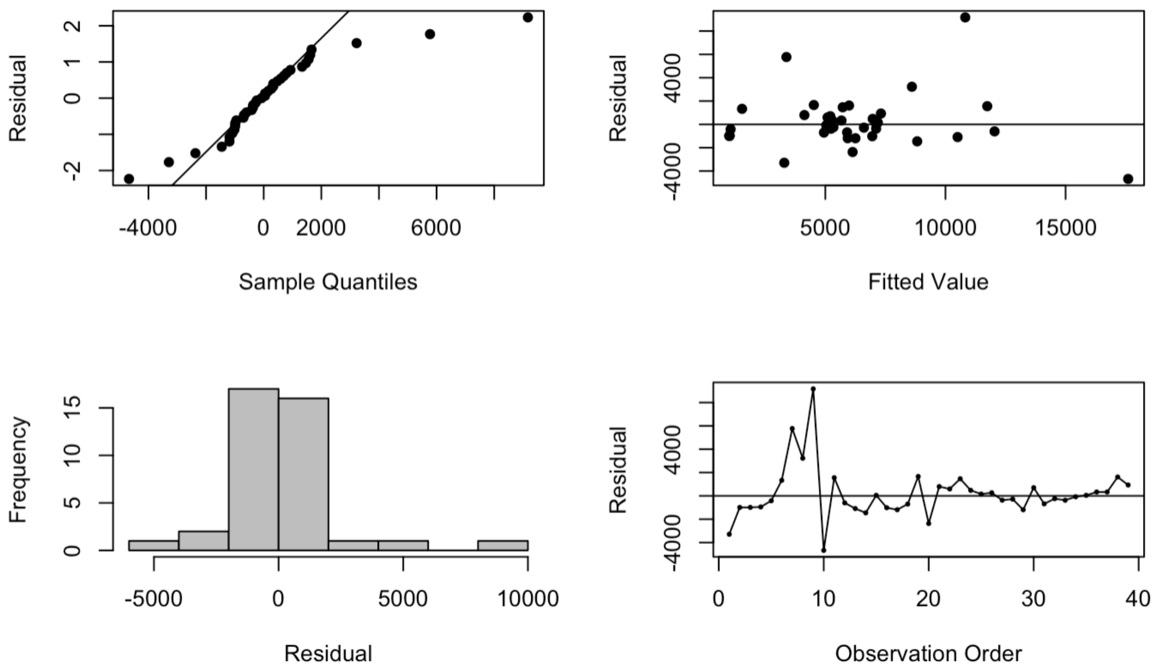
RESULT:

The forecasted model performed much better when we applied smoothing process. The trends are upward rather than having a straight line. The methodology for accurate forecasting is to fit a smoothed line based on the least AIC we get by applying arima() function in R. Then we use that fitted model to forecast the number of steps ahead. Generalized forecasting doesn't do a good job. It results in a straight line. The above plot is can be used to compare with the below plot given by CDC.

Mosharraf Hossain
Time Series Analysis (STAT 460)


Observed and forecasted weekly COVID-19 deaths in the United States

Each plot shows similar forecasting. CDC estimates 13000 deaths, and my model gave me around 12392 deaths and each model forecasts it will be increasing. The only difference is CDC plot has smaller confidence interval. Therefore, their model has lower bias estimation, and they were able to remove more noise in their data. Diagnostic plot of the forecast model shows that the there are some outliers present on our model.



The QQ plot shows that the forecast residuals are normally distributed, and histogram shows that the residuals don't have any non-normality. It is quite bell shaped. Residual vs Fitted Value plot shows that the residuals are scattered throughout the fitted values. It is a little bit congested between 5000-7000. Nonetheless, there is mostly equality of variance. We can use our forecast model for prediction since the diagnostic plots don't show anything out of ordinary. If we can remove the outliers from our main data set, it would forecast a little bit more accurately.

Conclusion:

The forecasts match with those of CDC's therefore it can be used to determine what would happen in the future and take appropriate measure to curb the deaths and cases. Right now we broke the record for daily deaths in the US which happened to be back in April.

Reference

1. Hyndman, R.J., & Athanasopoulos, G. (2018) *Forecasting: principles and practice*, 2nd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2. Accessed on 8 Dec. 2020.

2. *Halloween Time Series Workshop*, ramikrispin.github.io/halloween-time-series-workshop/. Accessed on 8 Dec. 2020