Robert Cerda

Covid-19 Analysis

Applied Business Forecasting

Professor Cavagnaro

April 10, 2020

**Homework #4 Covid-19**

**Hypothesis/Introduction:**

I was assigned the task of trying to come up with the best method for predicting the Covid -19 outbreak. What I am going to try and do in this paper is have select screen shots of my graphs and data to show you the summarization of the work I completed on the excel workbook to add conciseness and clarity. Something I would like to note; the output of the data will be shortened so that I may fit the title and outcome on the same screen, the full set of data will be on the workbook.

What I plan to do is a lot of trial and error, what I propose will happen is I will find the number of cases and deaths per day and find an underlying trend and use that to create my best method. A few issues I think I will run into are unexplained data points that will be affected by outside factors, such as the social distancing coming into place, below you will see my findings.

1. Graph the cumulative number of cases over time, and the number of total deaths over time in the USA. Describe the time series patterns. Do the same each state and country in the data set.

**USA**

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

In each of these, we plotted just the given data in the data set up to April 9th. I created them separately so that we may be able to see them for what each of them are. The first point we can make is that the model is not linear. We can clearly see the slope continuously increasing as the time series does. The second point we can make is when added a trend line, the R-squared is extremely high. The reason for this is I applied a different trendline for both, the first is an Exponential trend line and the second is a Polynomial trend at level 3 line on the total deaths for comparison and to show how uncomplex it is. What this means is that the data is showing strong Exponential growth.

**Italy South Korea**

A screenshot of a cell phone

Description automatically generatedA close up of a map

Description automatically generated

**California New York**

A close up of a piece of paper

Description automatically generatedA picture containing screenshot, map

Description automatically generated

**Michigan**

A close up of a piece of paper

Description automatically generated

In order to complete the second part properly, I felt the best way was to complete the total of all countries using a data set I retrieved from Kaggle. I placed Confirmed on top and Deaths below. What I think this shows and confirms is again an Exponential Trend. A common relationship most countries have in common when comparing Cumulative Cases, with the exception of New York which enjoyed a Linear Trend. Most Deaths were particular to either Exponential or Linear Trend as well.

1. The first difference converts “levels” into “changes.” Compute the number of new cases over time in each region (USA, Italy, South Korea, California, New York, Michigan), and the number of new deaths in each region. Graph these time series and describe the patterns.

**USA Italy**

A close up of a map

Description automatically generatedA close up of a map

Description automatically generated

**South Korea California**

A close up of a map

Description automatically generatedA close up of a map

Description automatically generated

**New York Michigan**

A picture containing text, map

Description automatically generatedA close up of a map

Description automatically generated

Just creating the outcomes of this set of data, I realized how different It is from the previous set of data. The graphs were very clear cut and obvious. With these, you can notice a lot of variability in not only the direction of the data, but also in trends. With each graph, I tested out to see which had the higher R-squared, some would only work with a Polynomial trend, whereas others would be just fine with a Linear trend including New York, Italy, and the US; this however does not conclude that we have a significant linear trend, just that we are working towards linearity. I just had an epiphany and noticed that these states/countries are among the highest in total number of cases in this group, this could possibly lead to a correlation later in the research.

1. A log transformation can be used to linearize exponential growth. Create new columns that are the logarithm of the cumulative cases and total deaths in the USA. Graph these time series and note the time series patterns. Then, use an appropriate forecasting technique to forecast the cumulative cases each day through April 12. What are some assumptions that must hold for this forecast to be accurate so far into the future?

**USA**

A screenshot of a cell phone

Description automatically generated

Using the transformation of the log function in Cumulative Cases and Total Deaths, we can see that a significant linear trend is created, this can open the avenue for a proper forecast.

A close up of text on a white background

Description automatically generated A screenshot of a cell phone

Description automatically generated

The linearity that the log function in the previous step created an opportunity for us to be able to use Holts Linear Trend method to create a forecast. What I did to create the model above is I used the logarithmic data with the Holt’s Method within the real-stat function and optimized the Mean Squared Error to its lowest possible Error rate. Once this gave us an effective forecast in the logarithmic, I was able to copy the formula down x amount of times; I chose one week. Once I had the desired number of forecasts, I used the Exponential function in excel to transform the data back to its original Exponential form.

What the model is telling us is that the confidence it has in this forecast is roughly 38% (alpha of .62) but for the previous time frames, the model has been very accurate and should continue to be as such as well as continue its Exponential upward trend.

1. Repeat part 3 for the number of total deaths in the USA. The log of zero is undefined, so begin with the first non-zero period.

A close up of text on a white background

Description automatically generatedA screenshot of a cell phone

Description automatically generated

For this model, we did the same method as we did for the cumulative cases in the previous step. In order to try and separate this from the previous one, I wanted to add the 1st difference and note that it continuously increases in difference to the point that the model suggests 5423 people will pass on 4/16/20 in the U.S. alone.

What this model shows is a one week forecast for total deaths in the United States. I optimized the MSE again to get the lowest error rate possible and the output was a very high alpha of .785 and a somewhat high beta of .434. The Confidence level is low, but I feel with the error rate in the previous dates, that we have a fairly strong model.

1. Repeat parts 3 and 4 for each of California, New York, Massachusetts, Michigan, Italy, South Korea, and India. Is a log transformation appropriate for each of these regions?

**Italy**

**A close up of a map

Description automatically generatedA screenshot of a cell phone

Description automatically generated**

**South Korea**

**A close up of a map

Description automatically generatedA screenshot of a computer

Description automatically generated**

**California**

**A close up of a map

Description automatically generatedA close up of text on a white background

Description automatically generated**

**New York**

**A close up of a map

Description automatically generatedA close up of text on a white background

Description automatically generated**

**Michigan**

**A close up of a map

Description automatically generatedA screenshot of a cell phone

Description automatically generated**

**Massachusetts**

**A close up of a map

Description automatically generatedA close up of a device

Description automatically generated**

**India**

**A close up of a map

Description automatically generated**A close up of a device

Description automatically generated

When observing each of the outputs and graphs above, there are a variety of points to note. First, I would like to address the question, “Is a log transformation appropriate for each of these regions?” Well, if you notice the appropriate models before this step, you would acknowledge that the majority of them have been following an exponential trend, in each one of these graphs, you can see that they (although some slightly less than others) quite obviously follow a linear trend, which tells us that the log transformation is the most appropriate to be able to use Holt’s method.

Now, is the model we created using Holt’s Linear Trend Method? Well, we wanted to optimize the Mean Squared Error and the vast majority of them are well below .1 if not in the general vicinity, which gives me a strong feeling that the model is very close with each prediction. What turns me away from the model slightly is the alpha being very high, some reaching as high as 1 showing 0 confidence in the model. That is combatted with a low probability of making a type 2 error (beta) for each high remark. Although I would prefer a model to reduce the level of alpha, I would say that we have a very strong method that we can continue to use.

1. The first difference of log-transformed data is the percent change. Compute the first difference of the log-transformed cumulative cases in the USA. Graph the time series and choose an appropriate method to forecast the value for each day through April 12. Undo the log transformation to convert this forecast back to a percent change. Interpret this forecast.

**USA**

A close up of text on a white background

Description automatically generatedA screenshot of a cell phone

Description automatically generated

Looking at this time series, it seems that we now no longer have a linear trend or any trend for that matter and cannot continue to use Holt’s Linear Trend Model to create a forecast. In order to determine which method I should use, I will place the data in a Correlogram to see what I should do with it.

A screenshot of a cell phone

Description automatically generated

Above is the ACF output that I have created with 40 lags. The first thing we can notice is that the data is significant as the first 4 lags breach the upper threshold. The second thing we can see is that the Correlogram seems to slowly decreases to zero, this indicates that my prior observation was incorrect and that we do have a trend. The last thing we notice is that there seems to be seasonality with significant peaks and troughs, so we just need to find the number of seasons. This I find that I have the hardest trouble with because the time series isn’t as clear, but I believe I found the spikes and dips at lags 0,10, and 21 which would give me the indication that this has an odd 11 seasons.

Because we have determined above that we have both trend and seasonality, I believe the best method to use in order to incorporate both of those would be Winter’s Multiplicative Method.

A screenshot of a cell phone

Description automatically generatedA close up of text on a white background

Description automatically generated

This almost broke my spirit when I saw the outcome. I really believed what I had done was correct, but I might haveto start back at square one. Before I did, I decided to look at the Correlogram again and see if I was able to determine a different cyclical pattern for seasonality, I tried a myriad of choices before I noted that there are subtle changes after each bunch of 3. So, I tried 3 periods for seasonality and lowered the MSE and sure enough I came up with a lot better model. But still was not satisfied, after some digging and some more research, I found that the MAPE is actually preferred to optimize when creating this model, therefore I chose to optimize the MAPE and created a stronger model that you can see below.

A close up of a map

Description automatically generated A close up of text on a white background

Description automatically generated

I am not very satisfied with this model. The forecast doesn’t follow the actual very well, it hits a minimal amount of points but does continue to get closer to the actual towards the end. Even after to optimization change from MSE to MAPE, it is still not a close to the actual. So, I reanalyzed the Correlogram again and came to the conclusion that aside from the first 4, the bars are fairly consistent in length, never really trying to approach zero, this has to be the problem with the previous charts, I have been trying to account for a trend that isn’t there.

Since there is no trend, I decided to try my hand at some moving averages. Below I created the first model that is a 3-Period Simple Moving Average.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

Below I created a 3 Period Weighted Moving Average Model to compare and contrast, as you can see the results are very similar.

A screenshot of a social media post

Description automatically generated A close up of text on a white background

Description automatically generated

I did the same thing for a 5-Period Moving Average to see how it compares.

A screenshot of a cell phone

Description automatically generated A screenshot of a cell phone

Description automatically generated

It seems as if the 3-Period Moving Average is the better model, I created a one-week (days) prediction using this method. Overall, I am very satisfied with the model, what I feel this method creates is more flexibility within the model which I feel makes the forecast stronger.

1. Repeat part 6 for each of California, New York, Michigan, Italy, and South Korea. In which state and in which country is the number of cases increasing the fastest? In which state and in which country is the number of cases increasing the slowest?

**Italy**

A screenshot of a cell phone

Description automatically generated A screenshot of a cell phone

Description automatically generated A close up of a map

Description automatically generated

Just to see how the previous findings would look on a different country, I tested the data to see how it would look graphed, in a Correlogram, and in Winter’s Method. The results were the same as the previous and I can determine that Simple Moving Average is the best model to use.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

This is the outcome of the Simple Moving Average with the Percent Change.

**South Korea**

A close up of a map

Description automatically generatedA screenshot of a cell phone

Description automatically generated

Here is the graph of the data and the Correlogram, these are very easy to distinguish that there is not a trend or seasonality.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

Here is the outcome of the Simple Moving Average with the Percent Change.

**California**

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

Here is the graph of the data and the Correlogram, these again are very easy to distinguish that there is not a trend or seasonality.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

At this point I think it is fairly obvious that I am going to need to do the same method for the remainder of the states.

**New York**

A close up of a map

Description automatically generatedA close up of a map

Description automatically generatedA screenshot of a cell phone

Description automatically generated

Here is the outputs of the Simple Moving Average Method and the percent change for New York.

**Michigan**

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

Here is the outputs of the Simple Moving Average Method and the percent change for New York.

  

  

Above are all of the averages of the Percent change for the Countries and States above, and from this data, we can see that the country with the most change is the USA. The state with the most change is Michigan which also has the most change out of all of the data sets.

1. Repeat parts 6 and 7 for using the number of total deaths instead of the number of cumulative cases. In which state and in which country is the number of total deaths increasing the fastest? In which state and in which country is the number of total deaths increasing the slowest?

**USA**

A screenshot of a cell phone

Description automatically generatedA screenshot of text

Description automatically generated

Here are the outputs of the Simple Moving Average with the percent change for the USA.

**Italy**

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

Here are the outputs of the Simple Moving Average with the percent change for Italy.

**South Korea**

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

Here are the outputs of the Simple Moving Average with the percent change for South Korea.

**California**

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

Here are the outputs of the Simple Moving Average with the percent change for California.

**New York**

A screenshot of a cell phone

Description automatically generatedA close up of text on a white background

Description automatically generated

Here are the outputs of the Simple Moving Average with the percent change for New York.

**Michigan**

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

Here are the outputs of the Simple Moving Average with the percent change for Michigan.

  

  

Above is the average percent change for each country/state in the data set. As we can see, the U.S. is the highest again for the Countries category and New York is the highest in the States Category as well as the highest in all categories.

1. Write a memo describing your analysis, justifying each step (including why you chose the method you did), and detailing your findings. How might we know when we are “flattening the curve,” based on this kind of forecast?

**Conclusion:**

Observing my original hypothesis, I can say that I did find a trend, a trend that I created using the logarithmic function. When calculating cumulative and total deaths, I really felt that I had found a very strong model using Holts method. Since this report took the course of several days to complete, I was able to compare it with real time data and found that I was a day or two off.

Upon realizing this, and you might have noticed in my graphs for the 1st difference Simple Moving Averages in step 8, I set the Forecasted time series back. When graphing them I found that if I started the forecast at the 1st data point and not at the time interval that it entered in at, I got a lot tighter fit to the model and a lot stronger model. Of course, this in turn would decrease m forecasts from 7 forecasted values to 4, but I believe a stronger model is worth that.

Overall, finding the methods were the toughest part about this project, but once found I can confidently say that I have the strongest model to predict the future outcome with the data set give. Now, to answer the question, “How might we know when we are “flattening the curve,” based on this kind of forecast?” The easy answer would be when the % change continues to decrease under 1.0. The hard to find answer would be when we find a cure. I stated in the beginning that a big problem in this project was going to be in the external factors effecting this data. Unfortunately, my forecasting knowledge will not be able to accommodate for this type of effect, but I am confident that this report gives a strong indication to where it’s heading.