STAT 153 Project Checkpoint 2

Danny Wu

3/10/2021

1. Executive Summary

COVID-19 has had a drastic impact to the daily lives of everyone around the globe. Having a prediction of the number of daily COVID-19 cases would be informative to both the health care sector as well as policy makers. Through our investigation, we fitted ...

2. Exploratory Data Analysis

Daily COVID-19 cases have dramatically increased since March of last year. As seen in the left panel of Figure 1, there is a very strong trend in the data. Cases grew from March to August then slightly declined until November, where it rapidly grew and exceeded the levels before the decline. There is also some seasonality based on the day of the week. We can see it in the fluctuations in the left panel as well as the right panel of Figure 1, where on average, cases are lower on Sunday and Monday. It is also clear that the data exhibits heretoscedasticity as the variance of daily COVID-19 cases has been increasing over time.

There are four dates in which there were zero cases counted which is likely erroneous data. Since the dates following the errors have strangely high counts, it can be assumed that the counts for the zero dates were accidentally moved to the next date. Thus we replace both with 1/2 of the count of the second day.

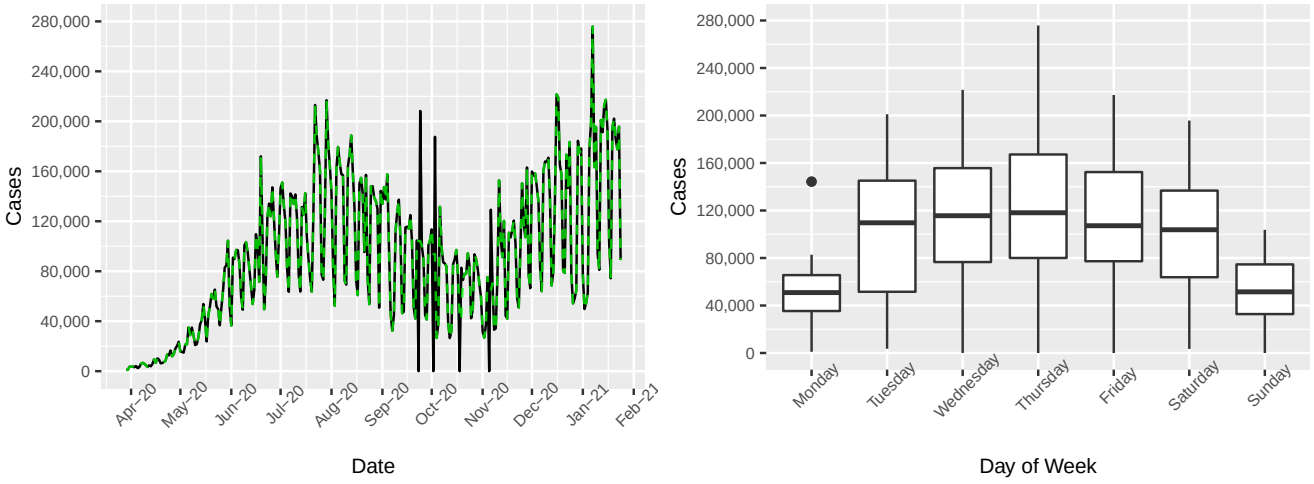

Figure 1: Left panel: Daily COVID-19 cases from March 29, 2020 to January 24, 2021. Dashed green line represents corrected dataset. Right panel: Box-plot of the distribution of daily cases grouped by the day of the week.

3. Models Considered

3.1 Parametric Signal Model

First we consider a parametric signal model. Using a periodogram we see that there are two dominant fourier frequencies at 2/302 and 43/302 corresponding to a period of 151 and 7.02 days. Thus we create a sinusoid with frequency 1/151 to model the larger trend, and then use indicators for the day of week to model the weekly seasonality. As it appears the amplitude of the weekly seasonality decreases in the troughs of the sinusoid, we include an interaction term between the sinusoid and the indicators. Lastly, we interact all the terms with time and time-squared in order to capture the changing amplitude of the sinusoid as time progresses. There are certain days where there are abnormally high counts as well as abnormally low counts too. Thus we add an indicator for these periods to help us better capture the underlying regularity of the data.

sorry I am working on the fancy equation that is neat and compact pls look at code for now :)

$$Y_t = \sum_{i=1}^{3} \beta_{2i-1} t^{i-1} cos\left(\frac{2\pi t}{151}\right) + \beta_{2i} t^{i-1} sin\left(\frac{2\pi t}{151}\right) + \sum_{j=7}^{9}\sum_{k=1}^{7} \beta_{7+(j-6)k} I_{weekday_k} t^{j-7}$$

$$+ \sum_{l=29} \text{interaction between sinusoid and weekdays AND the time function}$$

```
freq_ann = 2/302
model_para = lm(data = covid, cases_fixed ~
                  # big sinusoid
                  (cos(2*pi*t*freq_ann) + sin(2*pi*t*freq_ann)) * (1 + I(t) + I(t^2))

                # indicator for day of week
                  + dayofweek * (1 + I(t) + I(t^2))

                # day of week interaction with larger curve
                  + dayofweek:(sin(2*pi*t*freq_ann) + cos(2*pi*t*freq_ann)) * (1 + I(t) + I(t^2))

                # problem indicators
                  + index164_indicator
                  + index280_indicator
                  + index273_indicator
                  + index279_indicator
                  + index156_indicator
                  + index287_indicator
                )
```

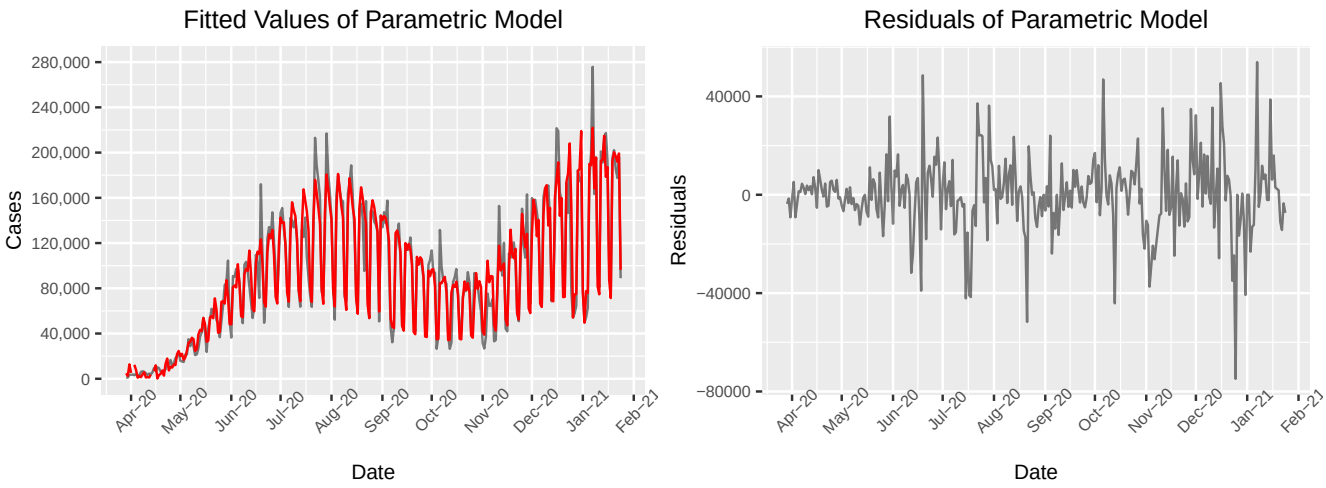

Figure 2: Left panel: Our parametric signal model with fitted values in red. Right panel: Residuals of the model

3.2 Differencing model

We now try a differencing approach. Because there exists heteroskedasticity, we apply a VST transformation to the data by raising all values to the 0.15 power. Since there is weekly seasonality, we difference it with a lag of 7 to take care of that. Looking at a periodogram, we see there is still a rather large frequency at period 3, so we take another difference of lag 3. The resulting plot looks relatively stationary with a few problem values that have big spikes and falls.

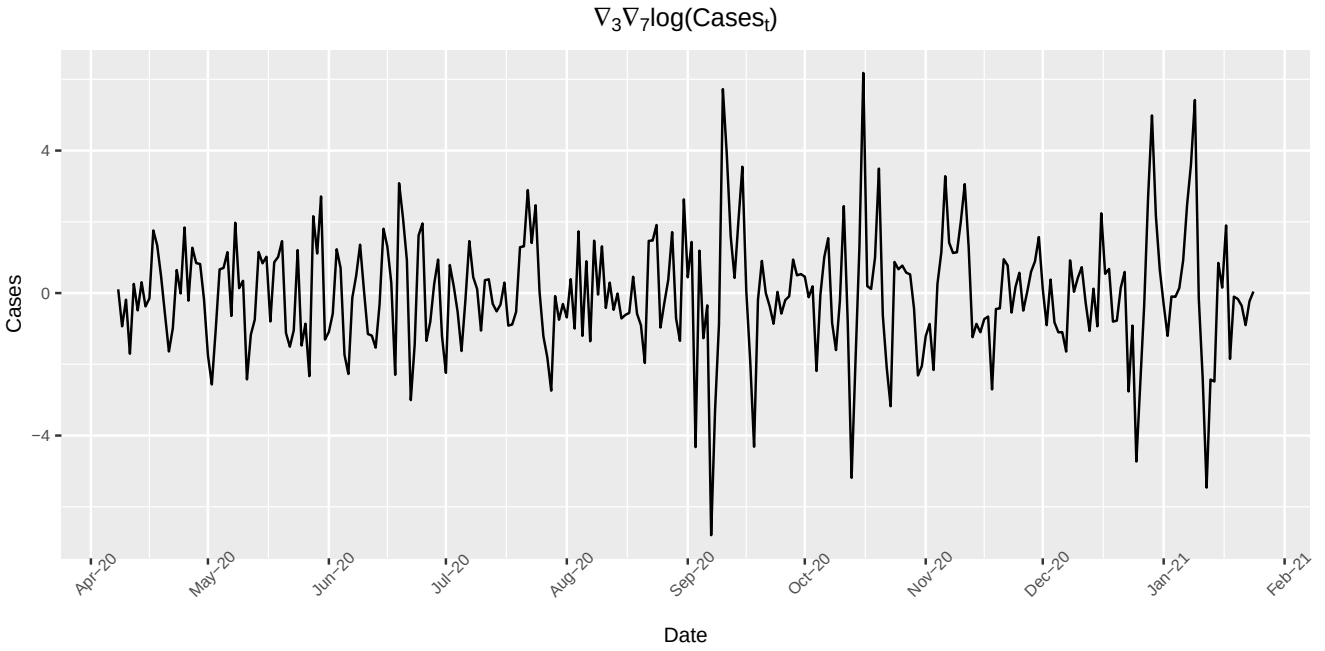

Figure 3: Differencing 'signal model' which looks relatively stationary