STAT 153 Project Checkpoint 3

Danny Wu

3/31/2021

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.

The are a few anomalies in the data where a day has zero counts and the following day has a spike in cases (marked by the red and orange points respectively in the left panel of Figure 1). These anomalies are likely due to cases being miscounted and accidentally moved to the next day. There are four of these instances. To correct for them, we divide the number of counts on the second date in half and set that as the number of counts for both dates.

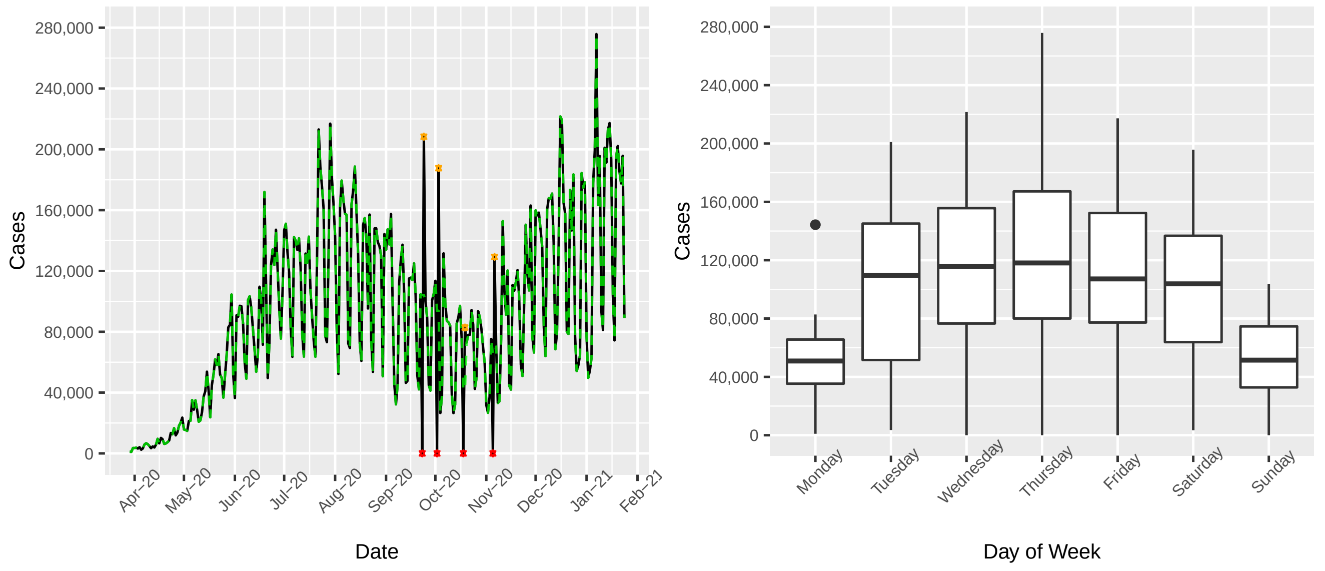

Figure 1: Left panel: Daily COVID-19 cases from March 29, 2020 to January 24, 2021. Dashed green line represents corrected dataset with red and orange points indicating erroneous data points. Right panel: Box-plot of the distribution of daily cases grouped by the day of the week.

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 also interact time, day of week indicator, and the larger sinusoid to capture any effects that might be from all three at the same time.

$$\text{Cases}_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \sum_{i=0}^{5} \left[\beta_{3+6i} I_{\text{weekday}_{it}} + \beta_{4+6i} t I_{\text{weekday}_{it}} + \beta_{5+6i} I_{\text{weekday}_{it}} \cos\left(\frac{2\pi t}{151}\right) \right. \tag{1}$$

$$\left. + \beta_{6+6i} I_{\text{weekday}_{it}} \sin\left(\frac{2\pi t}{151}\right) + \beta_{7+6i} t I_{\text{weekday}_{it}} \cos\left(\frac{2\pi t}{151}\right) + \beta_{8+6i} t I_{\text{weekday}_{it}} \sin\left(\frac{2\pi t}{151}\right) \right] \tag{2}$$

$$+ \sum_{j=0}^{2} \left[\beta_{39+2j} \cos\left(\frac{2\pi t}{151}\right) + \beta_{40+2j} \sin\left(\frac{2\pi t}{151}\right) \right] \tag{3}$$

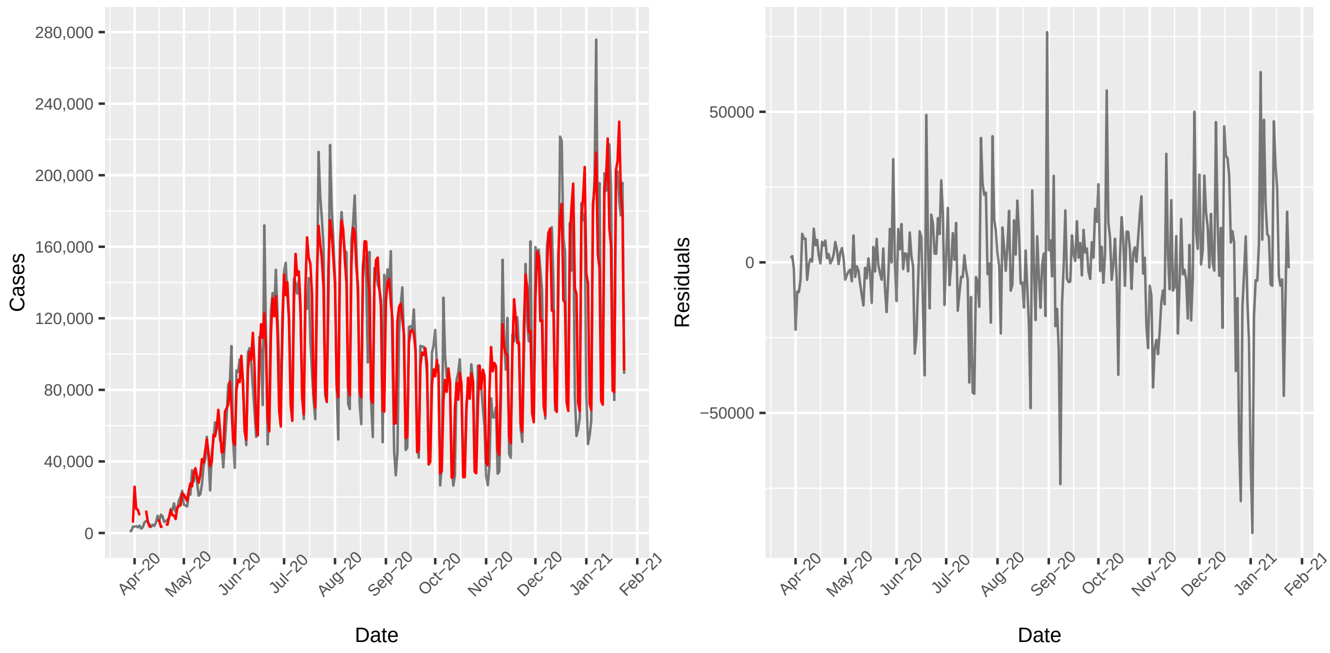

Figure 2: Left panel: Our parametric signal model with the fitted values plotted in red. Right panel: Residuals of the aforementioend parametric model

3.2 Differencing model

We now try a differencing approach. Because there exists heteroskedasticity, we apply a VST by taking the fourth root of all values. Since there is weekly seasonality, we applying differencing with a lag of 7. 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.

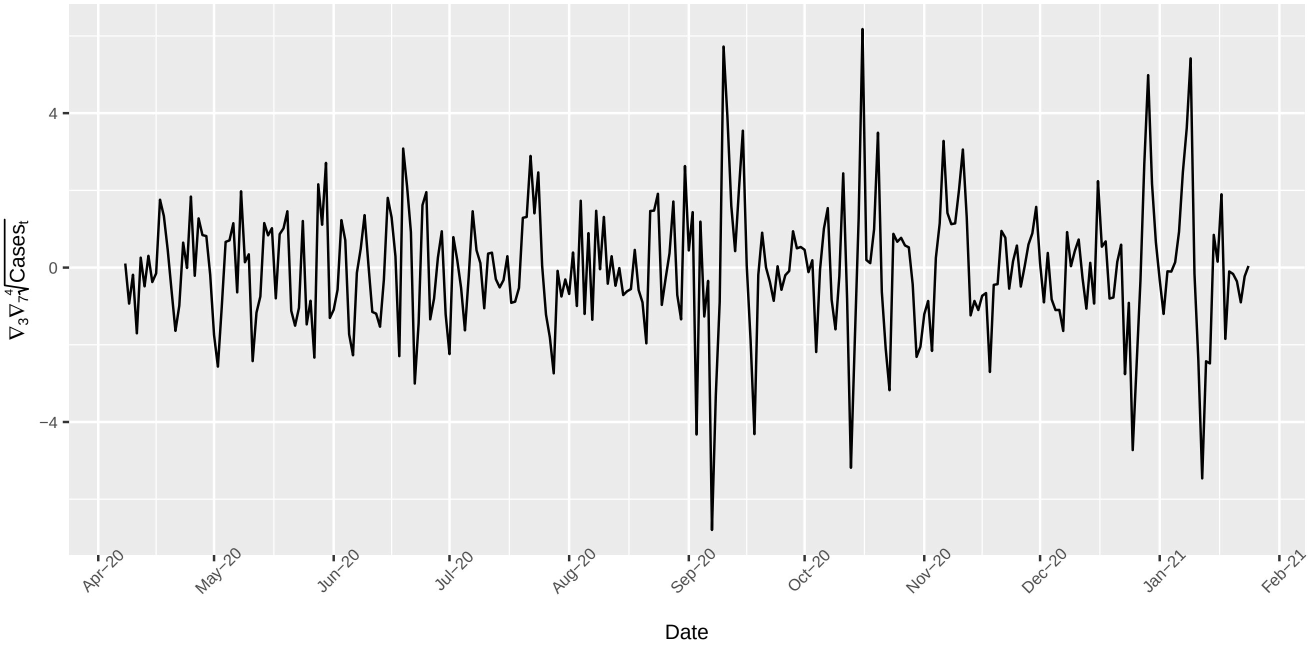

Figure 3: Plot of VST transformed data with lag 7 and lag 3 differencing applied. The result looks relatively stationary.