STAT 153 Project

Yewen Zhou

11/24/2020

1 Executive Summary

The Covid-19 dataset exhibits a pattern of quadratic trend and seasonality with a heteroskedastic behavior. We use log VST to minimize its heteroskedasicity. To pursue stationarity, we use two different approaches of parametric form and differencing. In both cases, we use periodograms to speculate possible frequencies and validate our model. With no clear "spikes" left in our residual (or filtered) plots, we have captured the trend and seasonality in the original data.

2 Exploratory Data Analysis

In the Covid dataset, marked by the plot on the left of Figure 1, the number of new cases appears to be quadratically related with days. By plotting the periodogram of the data, it is shown that there are conspicuous seasonal components in the data at Fourier frequencies of 1/60, and 9/60. Additionally, the data appears to be heteroskedastical - the variance increases linearly with the mean.

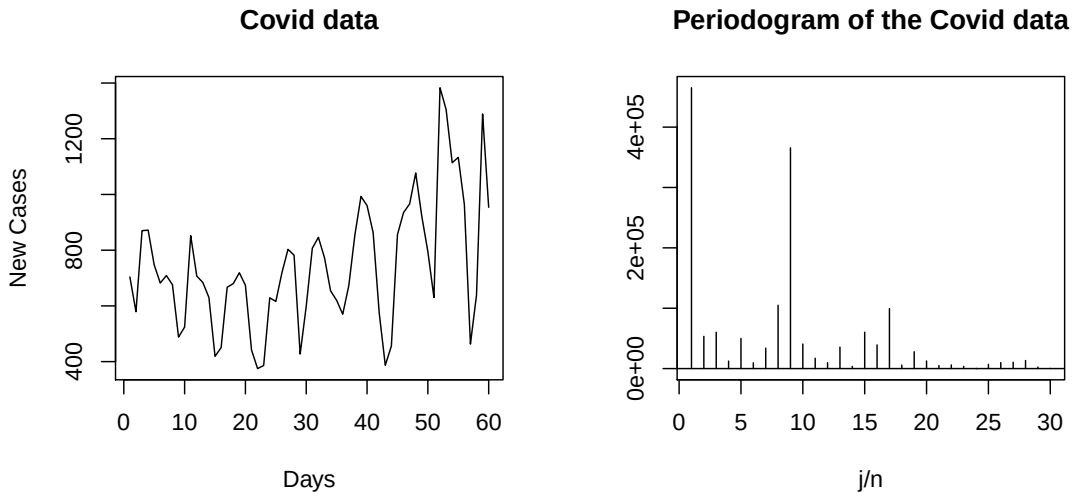

Figure 1: Covid-19 Cases and Periodogram

3 Models Considered

To model the natural signal in this data, both a parametric model and a differencing approach are used. The remaining stationary "noise" will be addressed in future iterations of this report using ARMA models.

3.1 Parametric Signal Model

First, the heteroskedacity is addressed by applying a log VST (Variance Stabalizing Tranform) on the Covid data. Then, the spectral density of the transformed data is plotted, which shows seasonal components at frequencies of 5/60, 8/60, 9/60, and 17/60. These information leads to a parametric signal model with quadratic trend and sinusoids at frequencies as above. The relevant plots are shown in Figure 2.

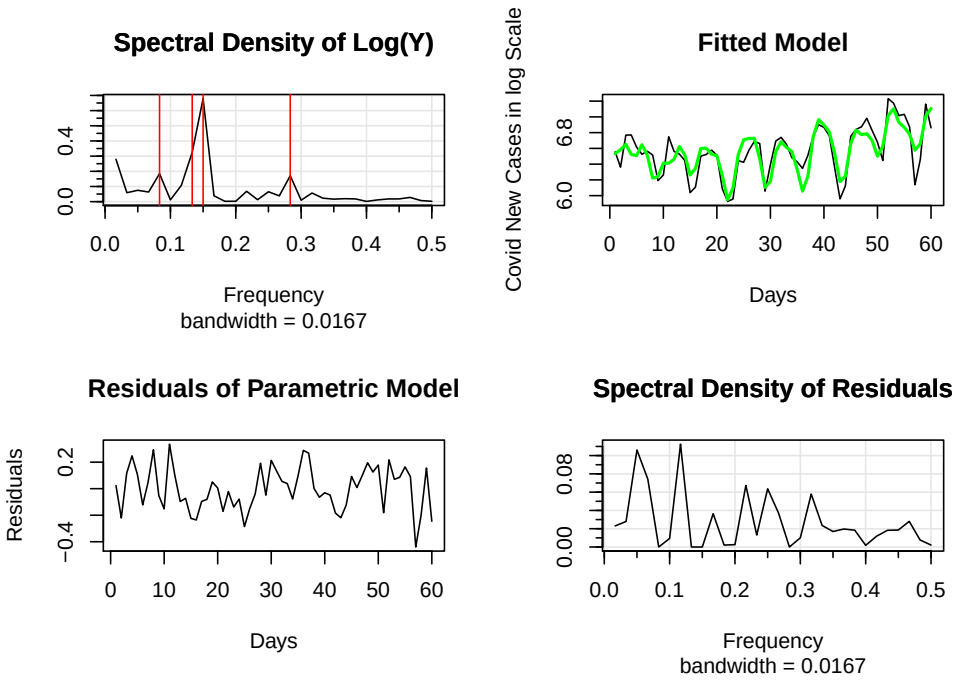

Figure 2: Parametric signal model fit, residuals, and periodograms

It can be seen from the plots that the model fits the transformed data well. Furthermore, the residuals appears to be stationary with a constant mean around 0 and a constant variance. The spectral density plot of residuals confirms that major seasonal components have been captured.

3.1.1 Parametric signal + ARMA(0,0)

The ACF and PACF plots for the parametric model residuals are shown in Figure 3. Because there are no lags with ACF/PACF magnitudes beyond the 95% confidence bands, the residuals is likely to be white noise. Thus, ARMA(0,0) is proposed, and this model implies the ACF and PACF indicated by the blue circles in Figure 3, which fit the general pattern of the sample autocorrelations.

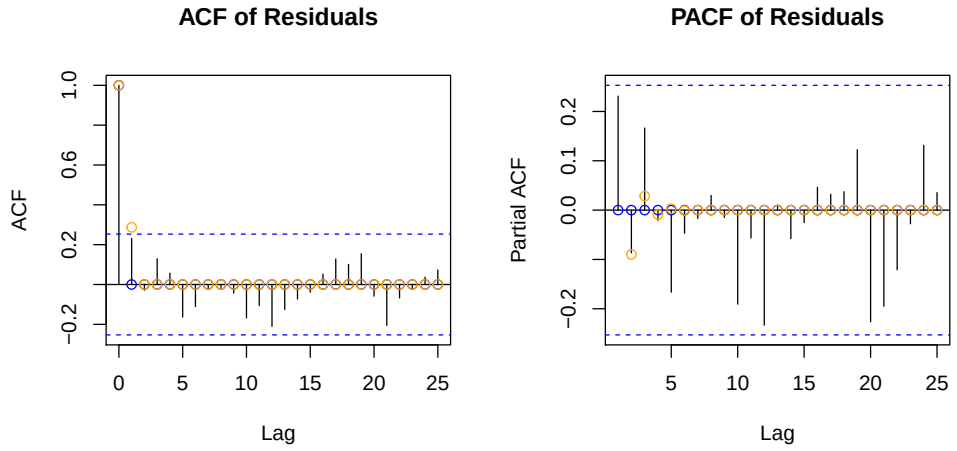

Figure 3: Autocorrelation function (ACF) and partial autocorrelation function (PACF) values for the parametric signal model's residuals. Blue circles reflect the ARMA(0,0) model, while the orange circles reflect the ARMA(0,1) model.

3.1.2 Parametric signal + ARMA(0,1)

Since there is no evidence for anything other than the white noise, the second model is selected as MA(1) with its theoretical ACF/PACF values shown as orange circles in Figure 3. This model captures the magnitude of ACF at lag=1 and the magnitude of PACF at lag=2 better than the first suggested model, thus appearing to be a better fit.

3.2 Differencing

In addition to fitting a parametric model on the signal, differencing is also considered. The periodogram of the log transformed data shows seasonal components at frequencies of 8/60 and 9/60, which correspond to periods at 60/8=7.5 and 60/9=6.67. This might occur due to leakage where the true period is 7. Furthermore, to remove the quadratic trend in the tranformed data, a second differencing with lag=1 is applied after the first differencing at lag=7.

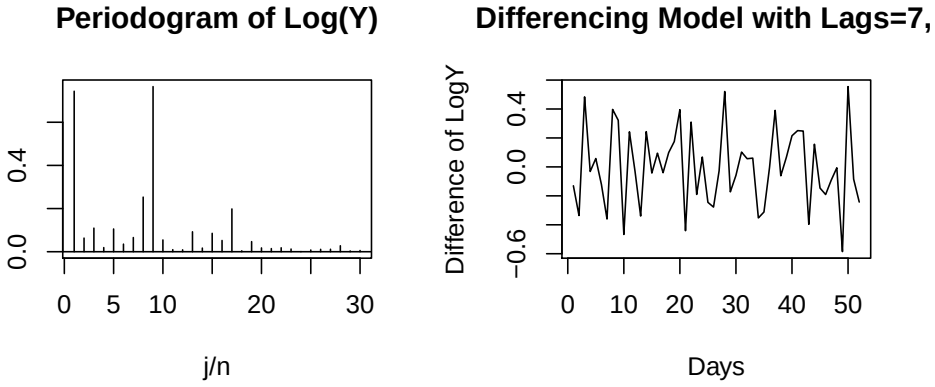

Figure 4: Differencing model and periodogram.

From Figure 4, we see that after differencing, the data appears to be stationary with a constant mean around 0 and a constant variance. Although there appears to be seasonal components in the filtered data, it is not deterministic.

3.2.1 Differencing + MSARMA(0,1)x(0,1)[7]

The ACF and PACF for differenced data are shown in Figure 5.

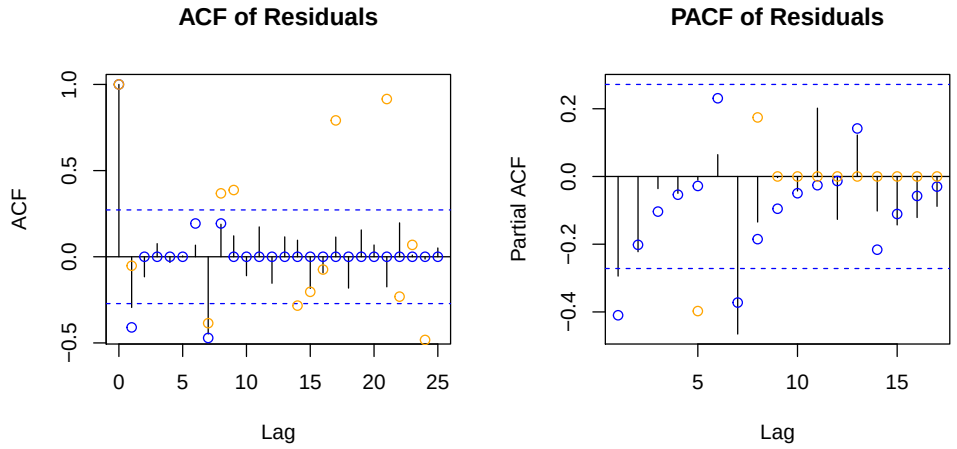

Figure 5: ACF and PACF of differencing model

The plots show that the magnitudes of ACF and PACF beyond the 95% confidence band occur at lag=1 and lag=7. The former magnitude at lag=1 suggests a possible ARMA model with q=1, the latter magnitude at lag=7 suggests a possible seasonal MA component with S=7 and Q=1. Based on the above observations, MSARMA(0,1)×(0,1)[7] is proposed. The theoretical ACF and PACF values is plotted as blue circles in Figure 5. These circles show that this model captures the general pattern of ACF and PACF well especially at lag=1 and lag=7, which indicates that it is a good fit.

3.2.2 Differencing with MSARMA(1,0)x(1,0)[7]

To address the magnitude in PACF at lag=7, a possible seasonal AR component is also considered with P=1 and S=7. To address the magnitude in PACF at lag=1, p is chosen to be 1. Thus, MSARMA(1,0)×(1,0) is proposed. The fit of this model is shown as orange circles in Figure 5. Clearly, this model failed to capture the general pattern and didn't fit sample autocorrelations as well as the first suggested model.

4 Model Comparison and Selection

Table 1: Cross-validated out-of-sample root mean squared prediction error for the four models under consideration.

	RMSPE
Parametric Model + ARMA(0,0)	2.052124
Parametric Model + ARMA(0,1)	2.048344
SARIMA(q=1,d=1,D=1,Q=1,S=7)	1.118891
SARIMA(p=1,d=1,D=1,P=1,S=7)	1.267871

Using cross-validation that rolls through data from time stamp=40 to 50, each time forecasting the data in the next 10 days with root-mean-square prediction error, RMSPE, we have the following table 1. It shows that the SARIMA(q=1,d=1,D=1,Q=1,S=7) model noise is the best overall according to this cross-validation exercise, and therefore this model will be used for forecasting.

5 Results

Model:

$$\begin{aligned} X_t &= log(Z_t) \\ Y_t &= \nabla_7^1 \nabla^1 X_t \\ &= \nabla_7^1 (X_t - X_{t-1}) \\ &= X_t - X_{t-7} - X_{t-1} + X_{t-8} \end{aligned}$$

Where Z_t is the raw Covid data, X_t is the log transformed data, Y_t is the filtered data after applying differencing.

MSARMA(0,1)×(0,1)[7] Model:

$$\begin{aligned} X_t &= \Theta(B^7)\theta(B)W_t \\ &= (1 + \Theta B^7)(1 + \theta B)W_t \\ &= (1 + \Theta B^7)(W_t + \theta W_{t-1}) \\ &= W_t + \theta W_{t-1} + \Theta W_{t-7} + \Theta\theta W_{t-8} \end{aligned}$$

Where W_t is white noise process, B is the backshift operator, Θ is SMA parameter, and θ is the MA parameter.

5.1 Estimation of model parameters

Estimates for model SARIMA(0,1,1,0,1,1)[7] parameters:

	Estimate	SE
ma1	-0.5212	0.1536
sma1	-0.7064	0.1401
xmean	0.0044	0.0061

5.2 Prediction

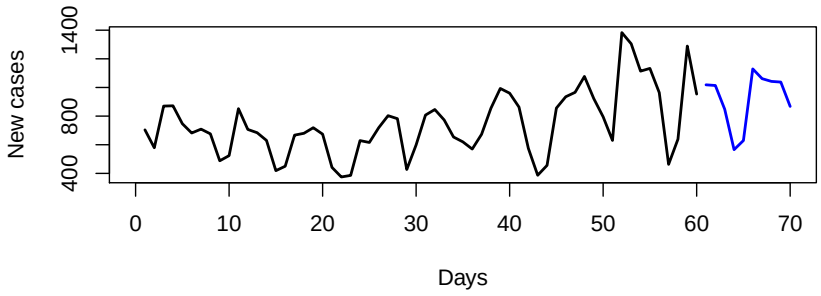

Figure 6: Covid with Prediction

The Covid raw data along with the prediction is plotted as Figure 6.