STAT 153 Project

Yewen Zhou

12/12/2020

1 Executive Summary

The number of new cases of COVID is expected to follow a cyclic pattern in the next 10 days with a drop around day 64, a rise around day 66, and a drop around day 70. According to our SARIMA(q=1,d=1,D=1,Q=1,S=7) model, the lowest point is expected to occur on day 64 with around 565 new cases, and the highest point is expected to occur on day 66 with around 1130 new cases. The model suggests that the situation of COVID does not look promising as the trend will continue to increase and necessary measures should be taken in the Gotham city to slow down the spread of the virus.

2 Exploratory Data Analysis

In the COVID data, marked on the left in Figure 1, the number of new cases appears to have a quadratic relationship with the time. From the periodogram, it is shown that there are seasonal components at frequencies of 1/60, 8/60, and 9/60. Since the corresponding periods 60/8 and 60/9 are not integers, this might happen due to leakage. Additionally, there appears to be heteroskedasticity in the data. The standard deviation appears to be increasing linearly with the mean, which is implying that the variance appears to be increasing quadratically with the mean.

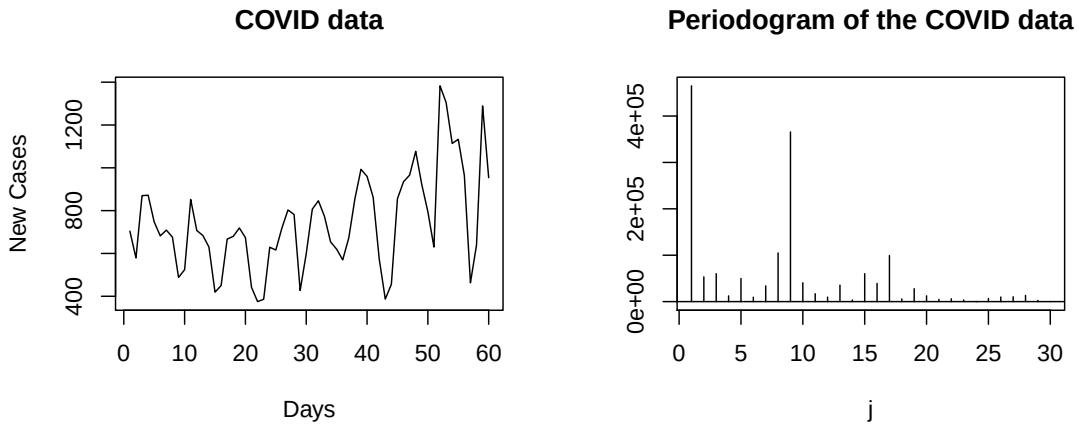

Figure 1: COVID New Cases and Periodogram.

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 are addressed with ARMA models in the following sections.

3.1 Parametric Signal Model

First, the heteroskedasticity 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. This deterministic model is detailed in the equation below, where $log($
$textCOVID_t)$ is the logarithemic transform of the COVID data and X_t is the additive noise term.

$$log(\text{COVID}_t) = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 cos(\frac{10\pi t}{60}) + \beta_4 sin(\frac{10\pi t}{60}) + \beta_5 cos(\frac{16\pi t}{60}) + \beta_6 sin(\frac{16\pi t}{60}) + \beta_7 cos(\frac{18\pi t}{60}) + \beta_8 sin(\frac{18\pi t}{60}) + \beta_9 cos(\frac{34\pi t}{60}) + \beta_{10} sin(\frac{34\pi t}{60}) + X_t$$

Figure 2 presents relevant plots including the spectral density of transformed data, the fitted model, residuals, and the spectral density of residuals.

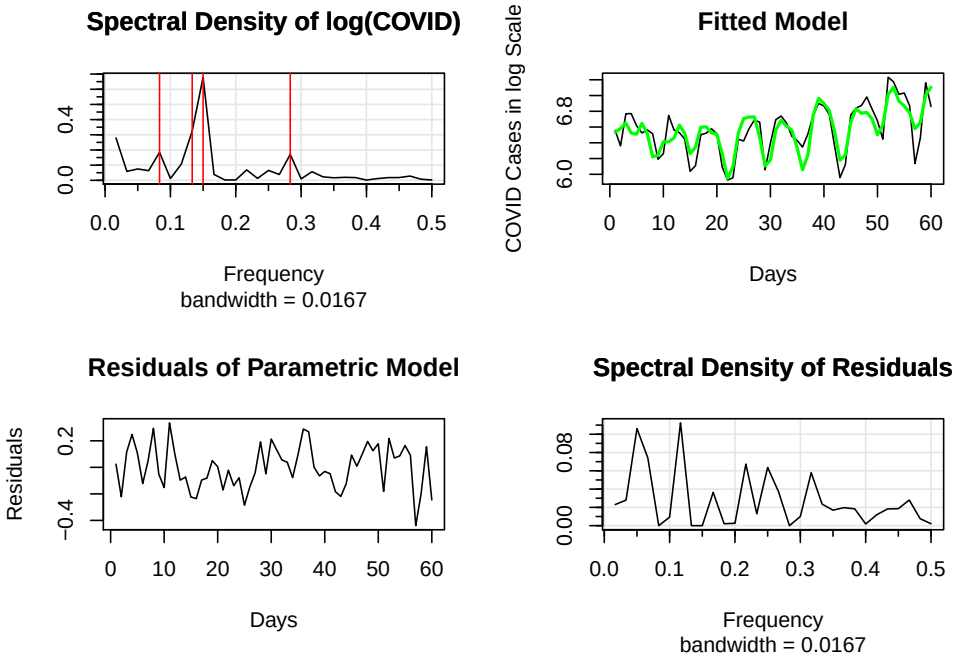

Figure 2: Spectral Density of $log(\text{COVID}_t)$, Fitted Model, Residuals, and Spectral Density of Residuals. $log(\text{COVID}_t)$ is the logarithemic tranform of the COVID data.

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 looks to be distributed evenly across the frequency domain, which is a good sign that major frequencies have been captured and implies that the residuals might be white noise.

3.1.1 Parametric signal + ARMA(0,0)

The ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function) plots for the parametric model's 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. The Ljung-Box Test of this model, marked on the bottom left in Figure 3,

shows that all p-values are above the rejection threshold, which implies that residuals is white noise and the model is a good fit.

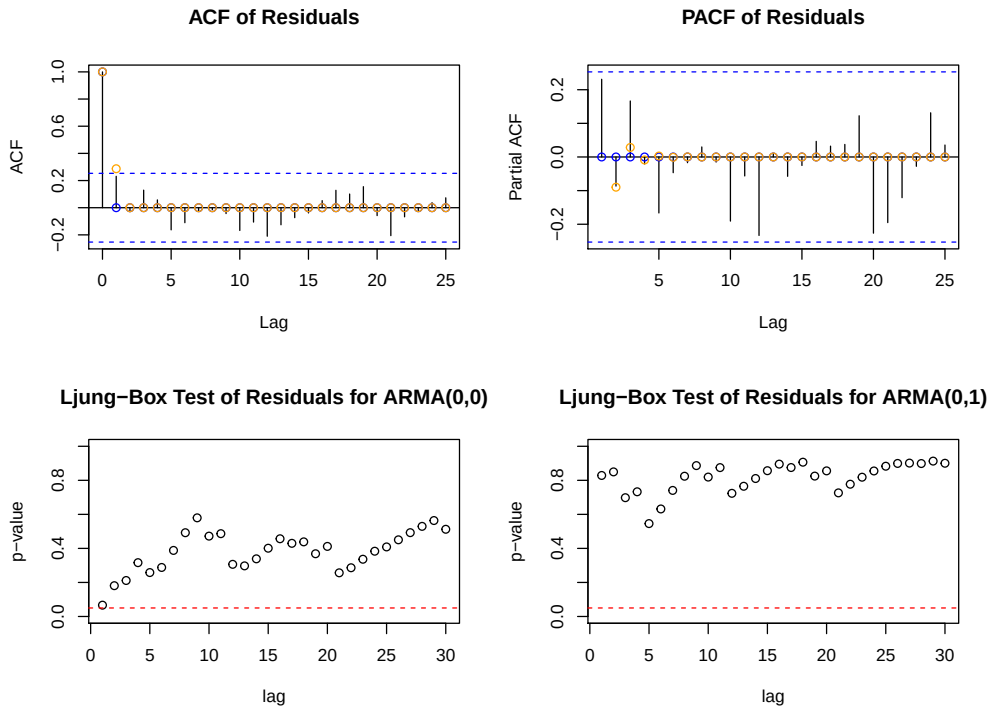

Figure 3: Autocorrelation function (ACF), partial autocorrelation function (PACF) values for the parametric signal model's residuals, and Ljung-Box Test of residuals for ARMA(0,0) and ARMA(0,1). Blue circles reflect the ARMA(0,0) model, while the orange circles reflect the ARMA(0,1) model.

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. The Ljung-Box Test of this model, marked on the bottom right in Figure 3 shows that all p-values are above the rejection threshold, which suggests that the residuals after fitting ARMA(0,1) is likely to be white noise. Moreover, the p-values are higher than those in the first suggested model. These observations imples that ARMA(0,1) might 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, which is essentially weekly effect and makes more sense. 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. Thus, the differencing procedure is shown in the equation below, where COVID_t represents the COVID time series data, X_t represents the log transformed data, and Y_t represents the filtered data after differencing.

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

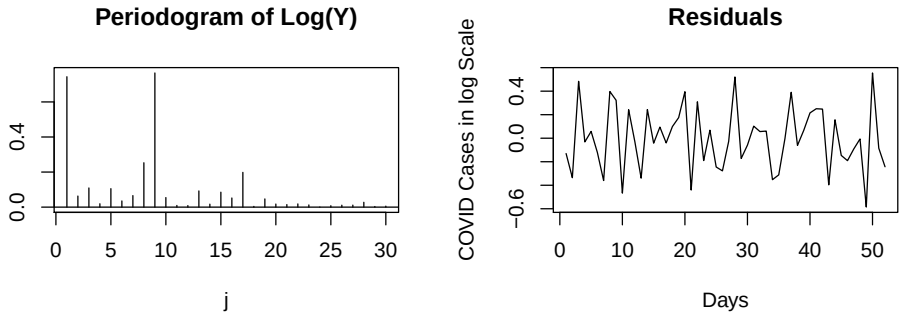

Figure 4: Periodogram of $log(\text{COVID}_t)$ and Residuals after differnecing. $log(\text{COVID}_t)$ is the logarithemic transform of the COVID data.

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. In the following sections, two MSARMA models are proposed to address the stationary residuals.

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

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

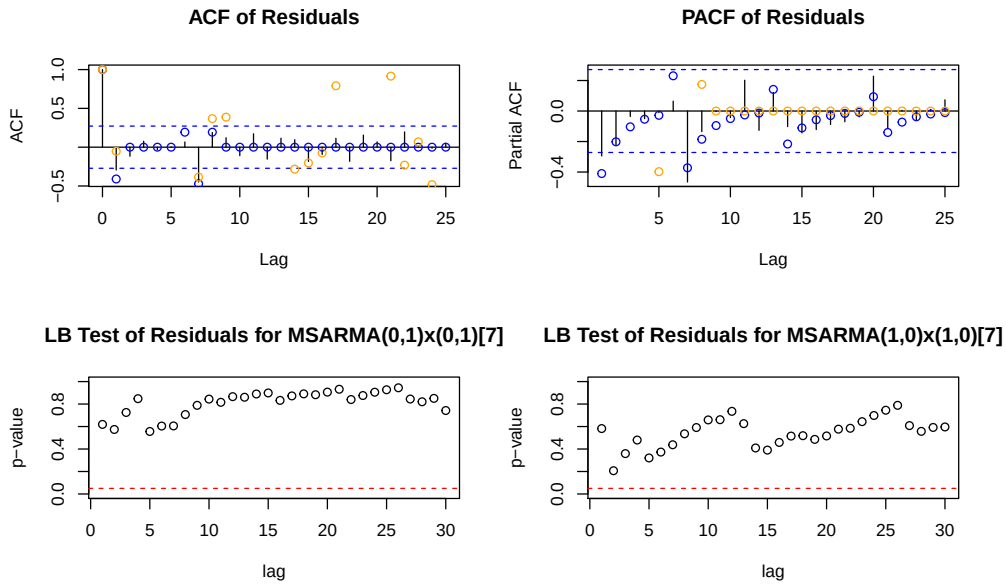

Figure 5: ACF, PACF of residuals, Ljung-Box Test of residuals for MSARMA(0,1)x(0,1)[7] and MSARMA(1,0)x(1,0)[7].

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, a multiplicative seasonal ARMA model, MSARMA(0,1)×(0,1)[7] is proposed. The theoretical ACF and PACF values are 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. The Ljung-Box Test of the residuals is shown on the bottom left in Figure 5. It shows all p-values are not significant, which suggests that the residuals is white noise and the model 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, the alternative model, MSARMA(1,0)×(1,0)[7] 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. The Ljung-Box Test for this model is plotted on the bottom right in Figure 5. Although all p-values are not significant, they are lower than the values in the first suggested model. Based on these observations, MSARMA(1,0)x(1,0)[7] is not as a good fit as MSARMA(0,1)x(0,1)[7].

4 Model Comparison and Selection

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 RMSPE of the SARIMA(q=1,d=1,D=1,Q=1,S=7) model, which is the same model proposed in section 3.2.1, is the best overall according to this cross-validation exercise, and therefore this model will be used for forecasting.

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

5 Results

The SARIMA(q=1,d=1,D=1,Q=1,S=7) model will be used to forecast COVID new cases in the next 10 days, and the detailed equation is listed below where COVID_t represents the number of new cases of COVID on day t, X_t represents the log transformed data, Y_t represents the filtered data after differencing, and W_t represents the white noise process with variance σ_W^2.

$$X_t = log(\text{COVID}_t)$$
$$X_t = X_{t-1} + X_{t-7} - X_{t-8} + W_t + \theta W_{t-1} + \Theta W_{t-7} + \Theta\theta W_{t-8}$$

θ, Θ are coefficients which will be estimated in the next subsection.

5.1 Estimation of model parameters

Estimates of the model parameters are given in Table 2.

Table 2: Model Estimates

	Estimate	SE	Coefficient Description
θ	-0.5212	0.1536	MA coefficient
Θ	-0.7064	0.1401	Seasonal MA coefficient
σ_M^2	0.0386671599450724		Variance of White Noise

5.2 Prediction

Figure 6 shows the forecast values of COVID new cases in next 10 days in log scale and the original scale. The model predicts that the new cases will follow a cyclic pattern with a drop up to day 64, a rise up to day 66, then a drop up to day 70. The lowest point is expected to occur on day 64 with around 565 new cases, and the highest point is expected to occur on day 66 with around 1130 new cases. The trend will be slowly going upwards, which suggests that necessary measures should be taken to help reduce the spread of the virus.

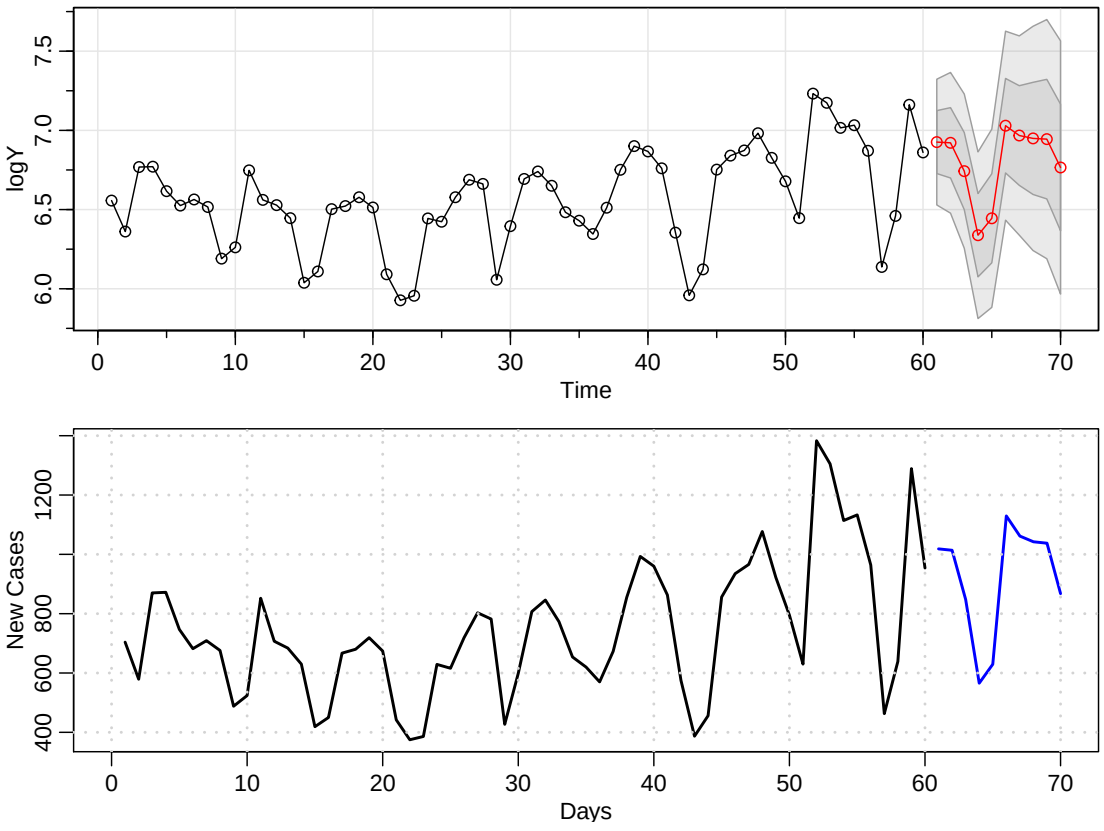

Figure 6: Forecasts of COVID new cases in next 10 days in log scale and original scale.