

STAT 153 Project

Yewen Zhou

11/24/2020

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.

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. By plotting the periodogram of the data, it is shown that there are seasonal components at frequencies of $1/60$, $8/60$, and $9/60$. Since the 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, so the variance appears to be increasing quadratically with the mean.

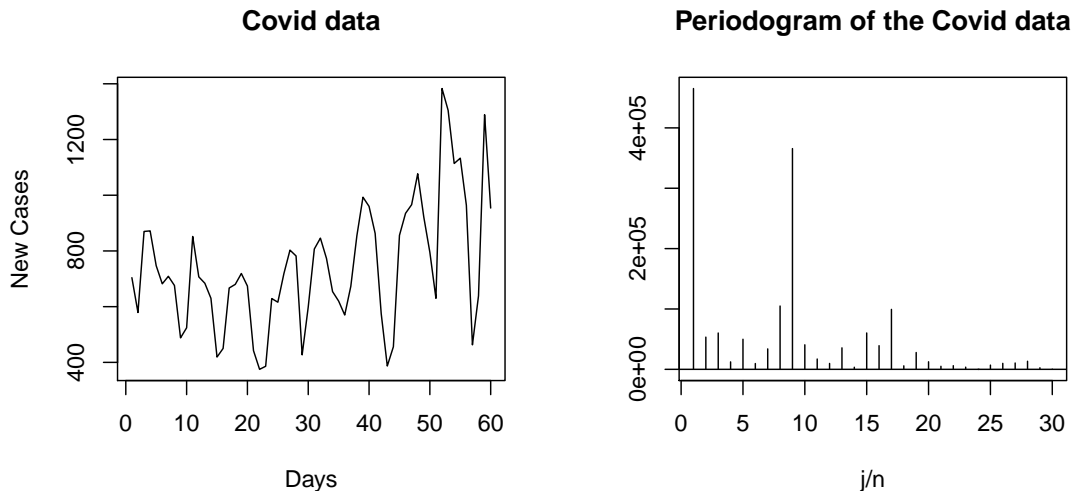


Figure 1: Covid-19 Cases and Periodogram

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 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 X_t is the additive noise term.

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

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

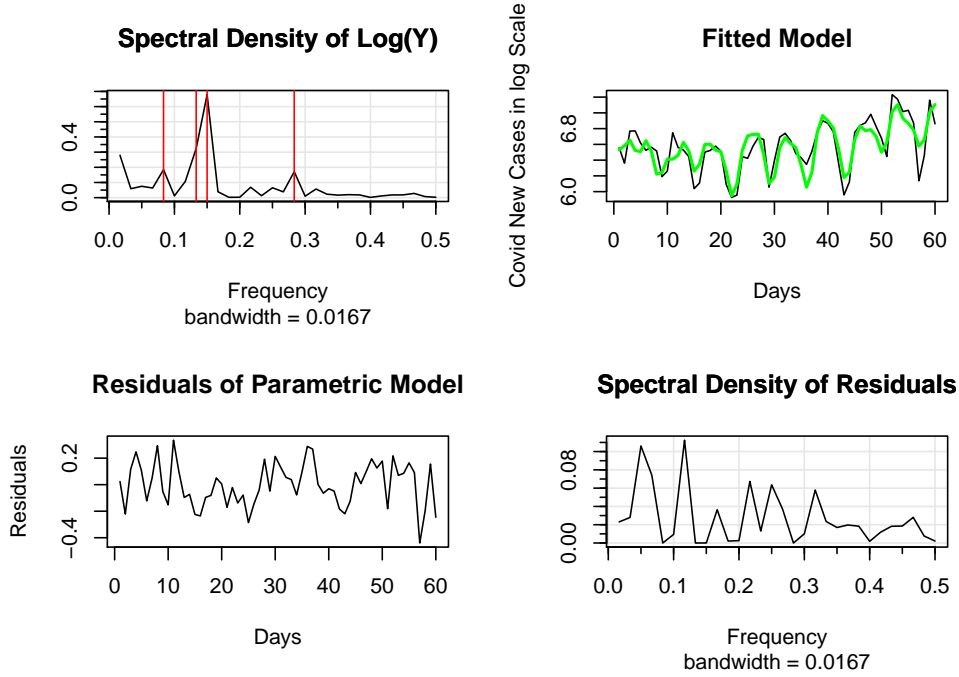


Figure 2: Spectral Density log(Covid), Fitted Model, Residuals, and Spectral Density of Residuals

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, $\text{ARMA}(0,0)$ 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.

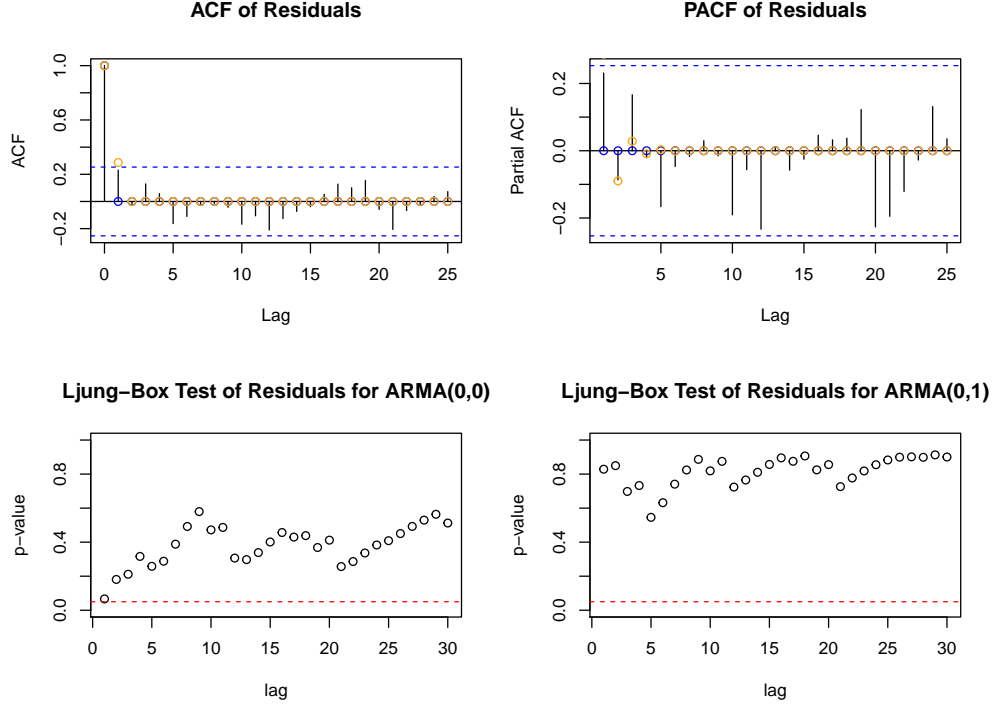


Figure 3: Autocorrelation function (ACF), partial autocorrelation function (PACF) values for the parametric signal model's residuals, and Ljung-Box Test of residuals for $\text{ARMA}(0,0)$ and $\text{ARMA}(0,1)$. Blue circles reflect the $\text{ARMA}(0,0)$ model, while the orange circles reflect the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

3.1.2 Parametric signal + $\text{ARMA}(0,1)$

Since there is no evidence for anything other than the white noise, the second model is selected as $\text{MA}(1)$ 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 $\text{ARMA}(0,1)$ is likely to be white noise. Moreover, the p-values are higher than those in the first suggested model. These observations implies that $\text{ARMA}(0,1)$ is 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.

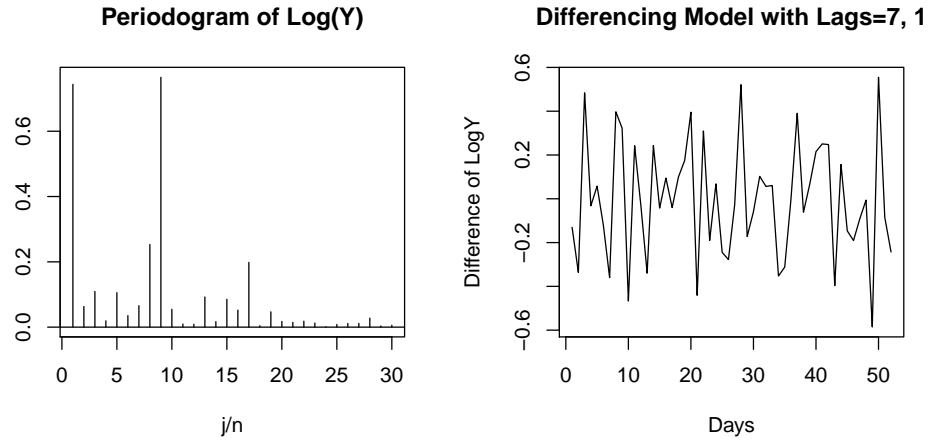


Figure 4: Differencing model and periodogram.

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.

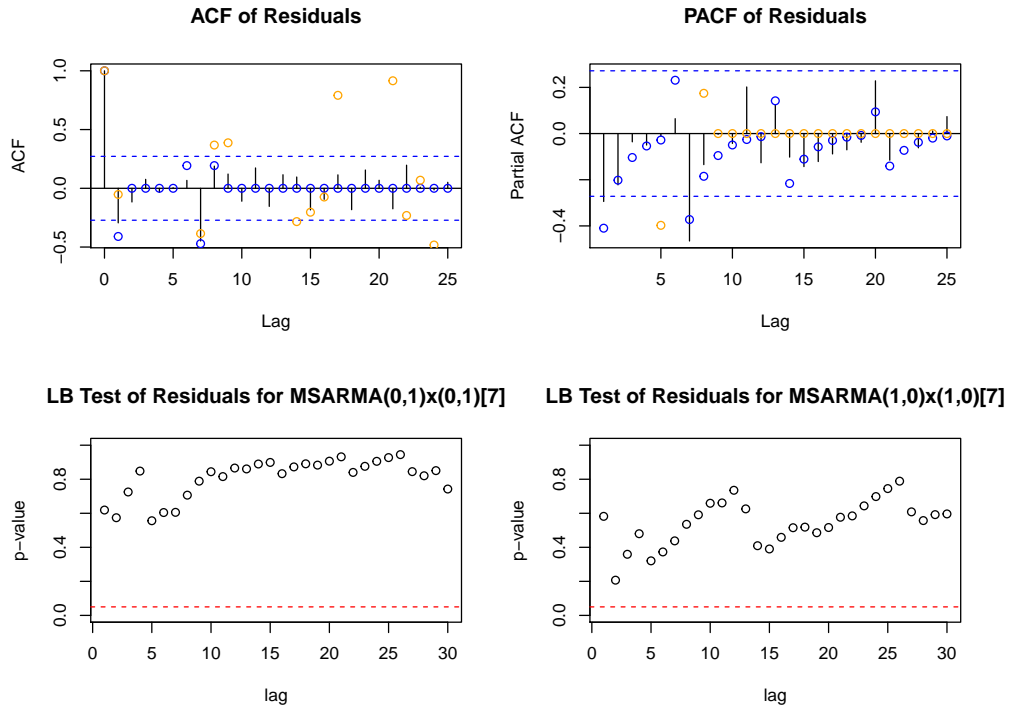


Figure 5: ACF, PACF of differencing model and Ljung-Box Test of residuals for MSARMA(0,1)x(0,1)[7] and MSARMA(1,0)x(1,0)[7]

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, $MSARMA(0,1) \times (0,1)[7]$ 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. 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 $MSARMA(1,0) \times (1,0)[7]$

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, $MSARMA(1,0) \times (1,0)$ 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. 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. These observations concludes that $MSARMA(1,0) \times (1,0)[7]$ is not as a good fit as $MSARMA(0,1) \times (0,1)[7]$.

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 $SARIMA(q=1, d=1, D=1, Q=1, S=7)$ model noise is the best overall according to this cross-validation exercise, and therefore this model will be used for forecasting.

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

To forecast Covid new cases in next 10 days, a differencing model will be used. Let Z_t be Covid new cases on day t , X_t be the log transformed data, Y_t be the filtered data after differencing. W_t is white noise process with variance σ_W^2 .

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

$$Y_t = W_t + \theta W_{t-1} + \Theta W_{t-7} + \Theta \theta W_{t-8}$$

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. 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.

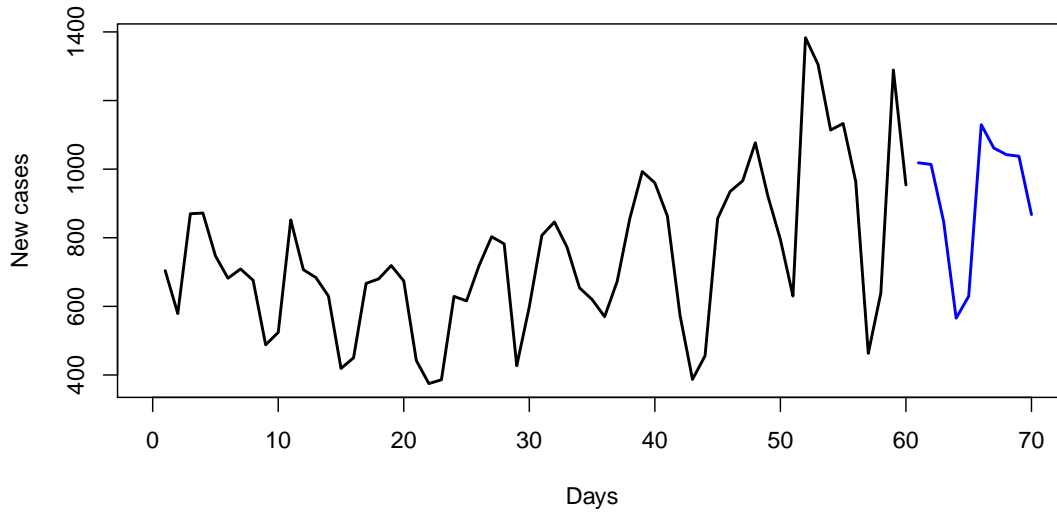


Figure 6: Forecasts of Covid new cases in next 10 days