Time Series Analysis of COVID-19

Jihee Yoon

December 14, 2020

1 Executive Summary

With the rapidly growing Covid cases around the world since early January 2020, the United States also experienced Covid cases rise between June and August 2020 as the pandemic reached large populations with a weekly seasonal pattern. August contributed to the especially large new cases of COVID-19 attributable with a stark incline due to the rapidly evolving spread of the virus. According to the parametric model with ARIMA(1,2,2)x(1,0,0)[7] noise, the new daily Covid rates seen between 8/19/20 to 8/28/20 will continue the same pattern of maintaining a slight increase and otherwise remaining relatively the same.

2 Exploratory Data Analysis

The new COVID-19 cases have been growing, as seen in Figure 1, and it displays seasonality and a trend. There is a strong seasonal pattern: new COVID cases spike during the weekdays and drop every weekend. The periodic cycles appeared to recur every 7 days, with a stark contrast between the weekends and weekdays. On a few Sundays and mostly Saturdays of each week, there was a drop in the number of new cases. The weekdays displayed a significant rise in the number of new cases. A hypothesis for this observation is that during the weekends, fewer labs may be open and working to process test results. In addition, more people may leave their homes to enjoy their weekend by traveling and meeting with other people, despite the social distance policy, and thus would require a few days before the first signs of the symptoms begin to show.

On the weekdays, there does not appear to be a clear correlation on the weekday and the number of new cases. The variance of the COVID cases gradually increase over time, and thus the Covid data appears to be slightly heteroskedastic.

The COVID data demonstrates an upward trend over the span of sixty days. From August 10th to 13th, there appears to have been the highest reports of new COVID-19 cases, with the peak on August 10th with 1383 new cases. This is because over time, the SARS-CoV-2 coronavirus is bound to grow in the number of cases and deaths as the pandemic continues over time.

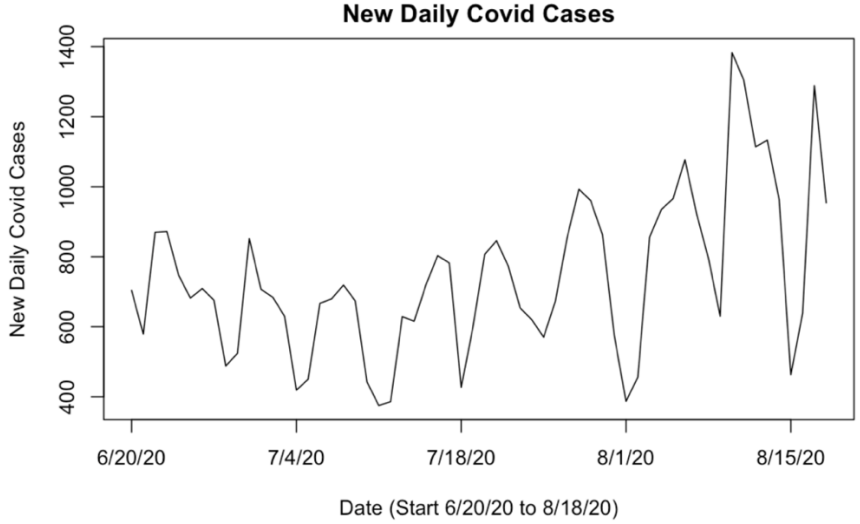

Figure 1: New COVID-19 cases each day from 6/20/20 to 8/18/20

3 Models Considered

Based on observations of the data, a variance stabilizing transform was conducted by taking a log of the original data. To model the natural signal in this data, both a parametric model and a differencing approach. The remaining stationary "noise" will be addressed using ARMA models in future iterations of this report. Going forward, all models are all taken with respect to the transformed data.

3.1 Parametric Model

A parametric model is considered to capture the trend. With the maximum frequency at 9 based on the periodogram (j=9) and the length of Covid new cases as n=60, the period was n/j = 60/9 = 6.6667. Thus, a sinusoid with a period of approximately 7 was used based on the weekly seasonality in the data observations, and this sinusoid interacted with time and the behavior of the days of the week. The parametric model captured the general seasonality and upward trend of the daily new Covid cases.

The residual plot of the parametric model was centered at zero, without obvious seasonality or trend, so the parametric model is relatively stationary. The ACF plot of the residuals of the parametric model had one autocorrelation exceeding the blue bands so this parametric model did not appear to be entirely white noise. Thus, though parametric model roughly captured the trend, the model is nearly but not completely white noise.

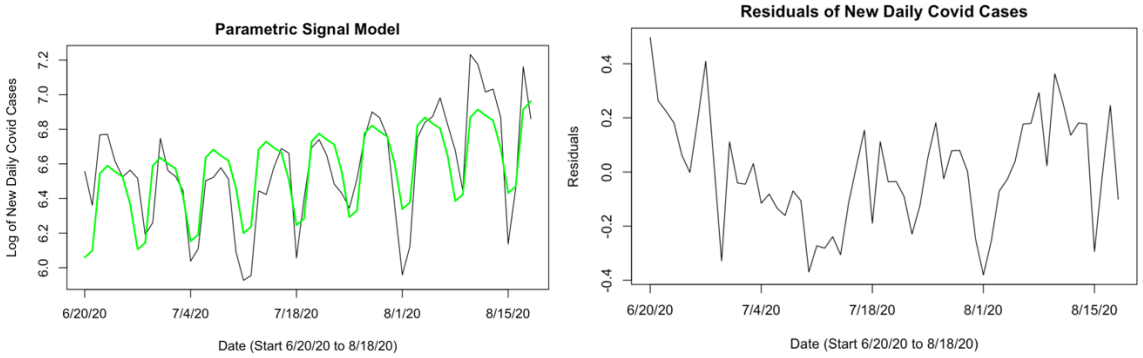

Figure 2: Parametric model fit (top left), Residuals of parametric model (top right)

3.1.1 Parametric signal with ARMA(p=1, q=0)[7]

The ACF and PACF plots for the parametric model residuals are shown in Figure 3. The lags with the largest magnitude ACF values of the original data occur at lags 1 and 7. This observation led to proposing the ARMA(p=1,q=0)[7] as a potential fit, with d=P=D=Q=0. This ARMA model implies an ACF indicated by the red dots. This ARMA model only had a ϕ_1 parameter, and the ACF and PACF of the residuals were within the blue bands and thus appeared to be white noise.

3.1.2 Parametric signal with ARMA(p=2, q=0)[7]

The ACF and PACF plots for the parametric model residuals are shown in Figure 4. The lags with the largest magnitude ACF values occur at lag 0. The ARMA(p=2,q=0)[7] was proposed as a potential fit, with d=P=D=Q=0. This ARMA model had ϕ_1, ϕ_2, and Φ_1 parameters, and the ACF and PACF values were within the blue bands and thus appeared to be white noise.

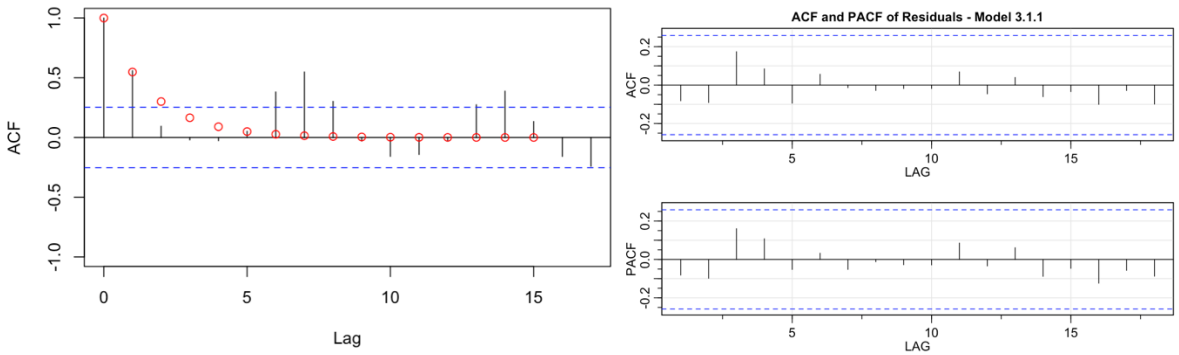

Figure 3: ACF of ARMA parameters (left), ACF & PACF of Residuals of Parametric Model 3.1.1 (right)

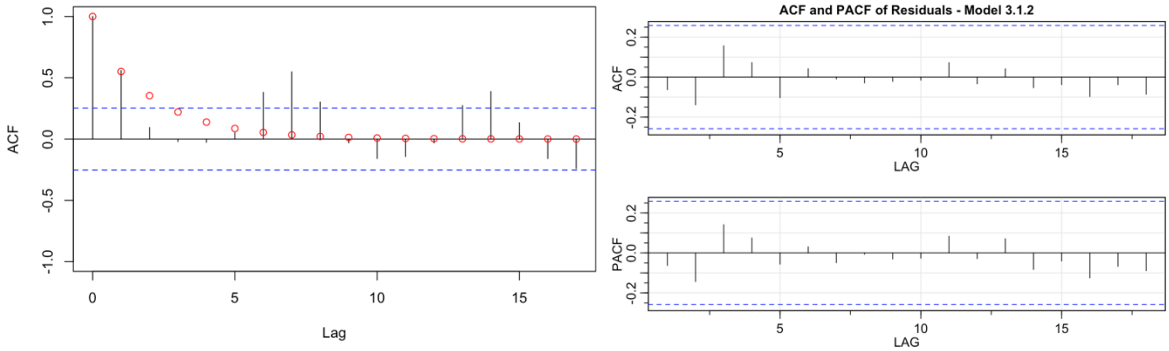

Figure 4: ACF of ARMA parameters (left), ACF & PACF of Residuals of Parametric Model 3.1.2 (right)

3.2 Differencing

To pursue stationarity, differencing is conducted to mathematically manipulate the raw time series so what remains can be assumed to be stationary. For the first order difference, lag-7 was used to eliminate the seasonality observed per week in the new daily Covid cases. In addition, lag-7 was the most effective compared to other alternatives because it yielded the lowest variance among the first order differences relative to lag-1, lag-14, lag-21, and lag-30. Afterwards, a second order difference with lag-1 was taken on the first order difference with lag-7, and this further removed the upward trend and seasonality without estimating it by differencing. Figure 5 below gives the time series of the differences which appears relatively stationary with no trend or seasonality, and the mean appears to be zero with constant variance.

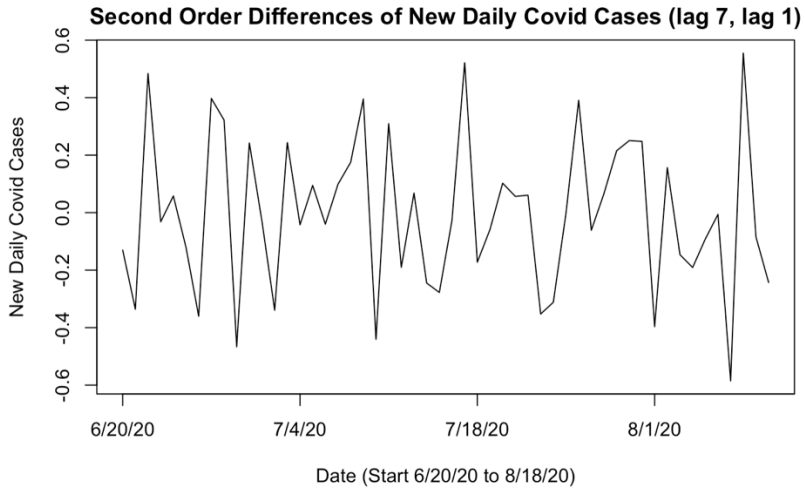

Figure 5: Second Order Differences of New Daily Covid Cases by Days

3.2.1 Differencing with ARIMA(p=1, d=1, q=1)x(P=0, D=1, Q=0)[7]

The ACF and PACF plots for the differencing models are shown in Figure 6. The lags with the largest magnitude ACF values occur at lags 1 and 7. This observation led to proposing the ARIMA(p=1, d=1, q=1)x(P=0, D=1, Q=0)[7] as a potential fit. Parameter d=1 due to the first order lag-7 and D=1 due to the second order lag-1. The SARIMA model with parameters p=q=d=D=1 and P=Q=0 yielded the ACF indicated by the red dots. This ARMA model had a ϕ_1 and $\theta_1$parameter, and lag-7 was the only lag that exceeded the blue bands on the ACF and PACF, which otherwise would appear to be white noise.

3.2.2 Differencing with ARIMA(p=1, d=2, q=2)x(P=1, D=0, Q=0)[7]

The ACF and PACF plots for the differencing models are shown in Figure 7. The lags with the largest magnitude ACF values also occur at lags 1 and 7. This observation led to proposing the ARIMA(p=1,d=2,q=2)x(P=1,D=0,Q=0)[7] as a potential fit. This SARIMA model implies an ACF indicated by the red dots. This ARMA model had ϕ_1, θ_1, and θ_2 parameters, and lag-7 was also the only lag that exceeded the blue bands on the ACF and PACF, which otherwise appeared to be white noise.

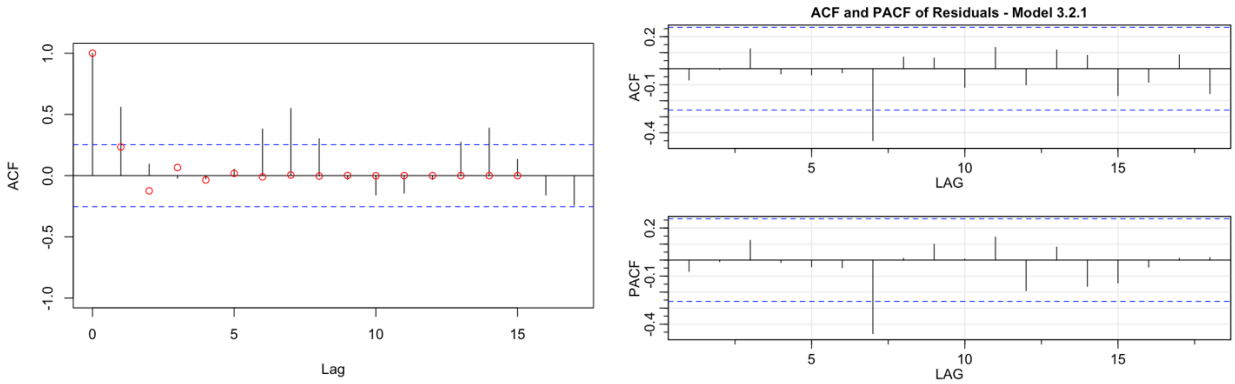

Figure 6: ACF plot of ARMA parameters fit (left),
ACF and PACF of Residuals of Differencing Model 3.2.1

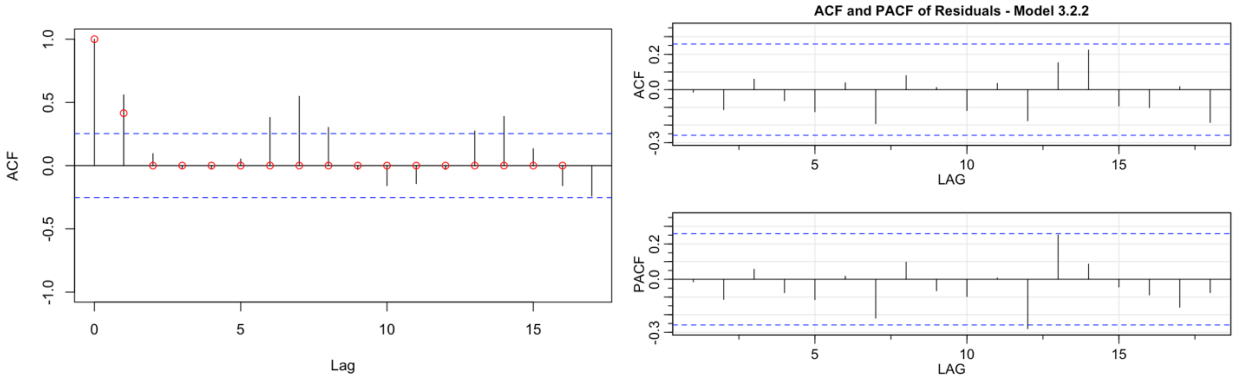

Figure 7: ACF plot of ARMA parameters fit (left),
ACF and PACF of Residuals of Differencing Model 3.2.2

4 Model Comparison and Selection

The cross-validation rolls through time starting from 40 days and it grows in increments in each for-loop of the cross validation. Cross validation was conducted to compute the sum squared errors (SSE) of the four models as a comparison metric of the best overall model.

Table 1 below shows that the differencing method with ARIMA(1,2,2)x(1,0,0)[7] from 3.2.2 is the best overall according to the cross-validation, and therefore this model is used for forecasting.

	SSE
Parametric Model + ARMA(1,0)[7]	1.795060
Parametric Model + ARMA(2,0)[7]	1.634140
Weekly Differencing + ARIMA(1,1,1)x(0,1,0)[7]	1.561430
Weekly Differencing + ARIMA(1,2,2)x(1,0,0)[7]	1.247138

Table 1: Cross-validated Sum Squared Error of the 4 Models

Note that, as in Table 2 below, the BIC values of Model 1 and Model 2 are the BIC values of the residuals of parametric models, whereas the BIC of Model 3 and Model 4 are the BIC values of the wholistic model from the differencing method. Thus, the BIC values are incomparable since it measures the BIC of the residuals for the first two models and the BIC of the entire model for the last two models. Thus, the SSE was used to determine the best model as opposed to the lowest BIC value.

BIC of Model 1 Residuals	BIC of Model 2 Residuals	BIC of Model 3	BIC of Model 4
-0.5223632	-0.3904235	0.2705102	0.197329

Table 2: BIC Values of Models

Figure 8 below shows the diagnostic plots for this model. The differencing method with ARIMA(1,2,2)x(1,0,0)[7] from 3.2.2 has residuals which follow a roughly normal Q-Q plot with points at the end which deviate slightly from the blue line of normality. One of the lags on the ACF of the residuals exceeds the blue bands, meaning that this model is not white noise. Several of the p-values for the Ljung-Box statistic hover above the blue dashed lines so this indicates that the SARIMA model is a good fit since high p-values mean it is likely that we do not reject the null hypothesis.

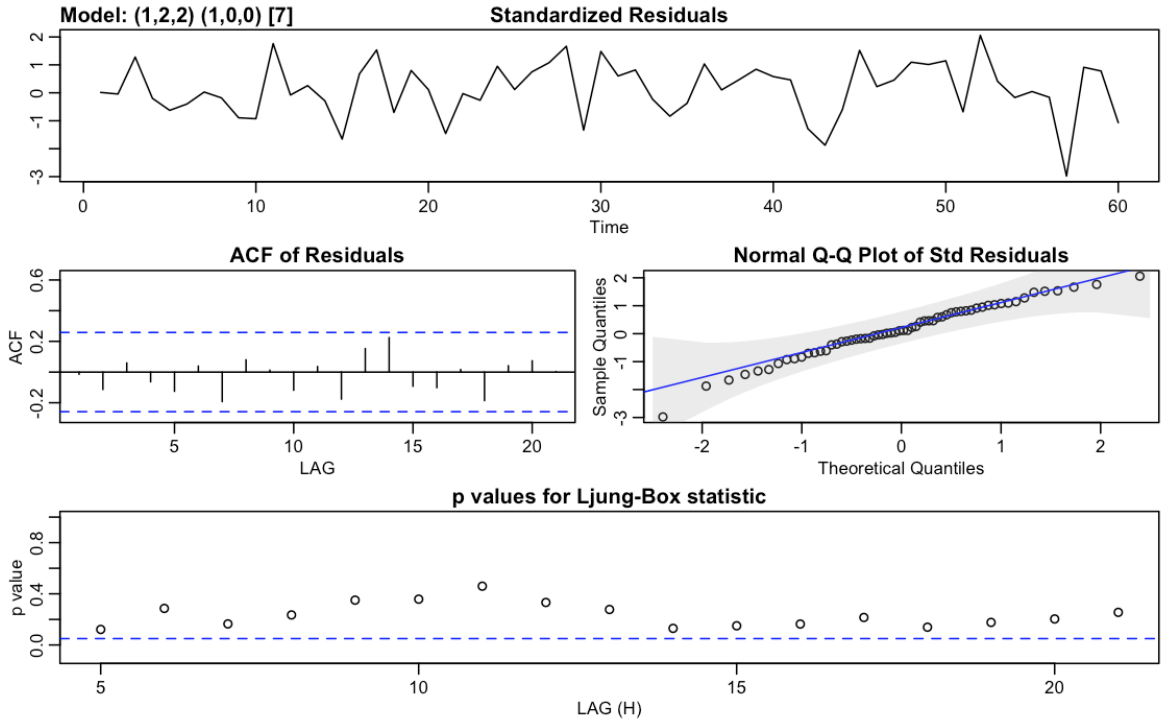

Figure 8: Diagnostic plots for ARIMA(1,2,2)x(1,0,0)[7]

5 Results

The mathematical expression of the second order differences for model 4 from 3.2.2 is expressed as the second order differences with lag-7 and lag-1 of the log of the Covid data, as represented below in Figure 9. Its SARIMA model equation only has ϕ_1, θ_1, θ_2 , and $\Phi_1$parameters and is expressed in Figure 10.

$$Y_t' = \nabla_7 \nabla_1 log(Y_t)$$

Figure 9: Math Expression of Second Order Difference of Log Covid Data

$$X_t = \phi X_{t-1} + \Phi X_{t-1} + W_t + \theta_1 W_{t-1} + \theta_2 W_{t-2}$$

Figure 10: ARIMA Model Equation

5.1 Estimation of Model Parameters

This model only had one theta value for the MA and the estimation of this parameter is in Table 3.

Parameter	Estimate	Std. Error
ϕ_1	0.5324	0.1619
θ_1	-1.9533	0.1290
θ_2	0.9544	0.1255
Φ_1	0.6014	0.1131

Table 3: Parameter Estimate and Standard Error for ARIMA model for Model 4 from 3.2.2

5.2 Prediction

Model 4 from 3.2.2 was the best model because it earned the smallest sum of the squared errors relative to the other three models because its predictions were the closest to the actual values. Reversing the log transform using exponentiation to preserve the original form, the model predicts that the forecasted number of daily new cases for the next 10 days, 8/19/20 to 8/28/20, remains relatively the same. The number of cases appear to slightly decrease compared to the preceding weeks, but it still maintains a relatively high case of new daily Covid cases and the drops in Covid cases maintain the same pattern, every 7 days.

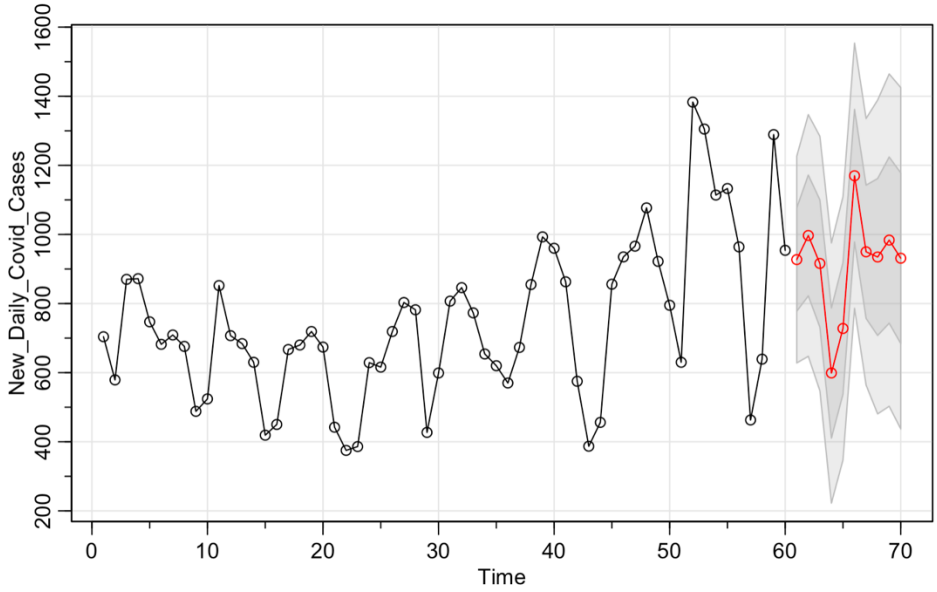

Figure 11: Forecasts of Daily New Covid Cases Including Next 10 Days