STAT 153 Final Project

Group: Tessa Williams, Rex Winn

May 26, 2020

Executive Summary:

This report aims to help visualize and model the coronavirus (COVID-19) outbreak within the country of a non-given country "Timeseria". We analyzed the total number of confirmed cases over a 66-day period, starting from the day of the first COVID-19 case to present. We used the following three models to analyze the data: 1) $ARIMA(1,2,0)$, 2) $ARIMA(0,2,2)$, and 3) $ARIMA(1,3,1)$. Based on cross validation, we selected the $ARIMA(1,3,1)$ model to forecast the total case count over the next 10 days. Using this models forecast, we anticipate the total number of COVID-19 cases will reach approximately 110,000 total cases 10 days from now.

1 Exploratory Data Analysis

Beginning from the first documented coronavirus (COVID-19) case in the country of Timeseria to present, we analyzed 66 days of COVID-19 case count data. The COVID-19 data is presented in the figure below (Figure 1).

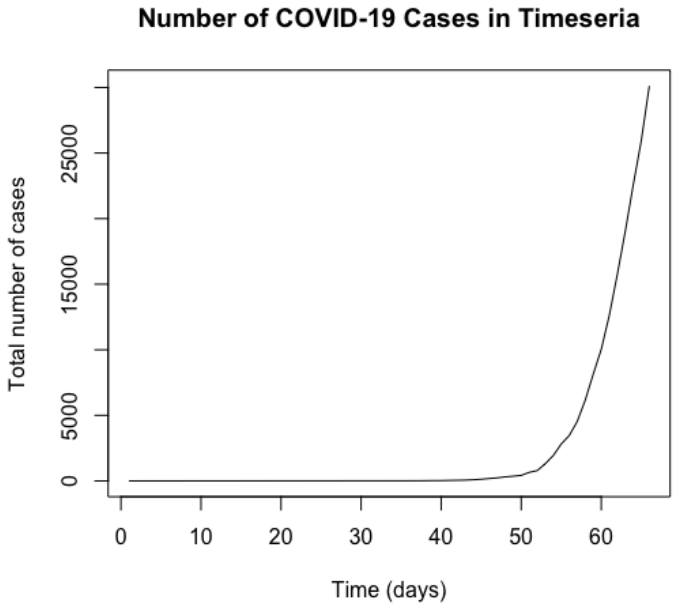

Figure 1: Total number of COVID-19 cases reported in Timeseria since beginning of outbreak.

While there is no noticeable noise or seasonality, the data is trending upward over time at what appears to be an exponential rate. The slope of the cumulative data appears to still be increasing, indicating the daily COVID-19 case count has not yet reached its peak. The total case count is currently at 30,116 cases and we anticipate this number will continue to increase over the next 10 days. Three models were developed to fit the data and pursue stationarity of the residuals. After detrending the data, the residuals were subsequently modeled using various ARMA models. Details of the three models are presented in the section below, followed by model comparison and a discussion of the results.

2 Models Considered

We developed three models to fit the COVID-19 data. All three models include differencing to remove the upward trend in the data. After detrending, the variance appeared heteroskedactic, so variance stabilizing transforms (VSTs) were used on the data. The VSTs used of model one consisted of taking the log of the data, and a Box-Cox transformation was used for model 2 and the model 3 with $\lambda = 0.05$. Various ARIMA components were then added to each of the models so that the residuals appeared to be consistent with a white noise process. An overview of the model development process is presented in the sections below. Further details, plots, and corresponding R code used to develop each model is presented in Appendix A at the end of this report.

2.1 Model 1

We decided to first detrend the data by taking a second order difference. The second order difference appeared to adequately detrend the data; however, the variance was increasing with time beginning at about 45 days. This lead us to apply a VST to the data prior to detrending. Because the variance appeared to be increasing exponentially with time, we took the log of the data as a VST.

After taking the second order difference of the transformed dataset, the residuals appeared to be stable. Upon further review of the autocorrelation function (ACF) and partial-autocorrelation (PACF) plots, we decided to add an AR(1) component to the model. This was selected based on the significant bar on the PACF plot at lag = 1. By adding an AR(1) component to the model, the residuals appeared to be stable. Based on the ACF plot of the model residuals and the Ljung-Box-Pierce test, we believe the residuals modeled as white noise is a valid assumption. Tessa's final model is an $ARIMA(1, 2, 0)$ model with a log VST. Plots of the model residuals, ACF of residuals, normal Q-Q standard residuals, and Ljung-Box statistic for model 1 are presented in the figure below.

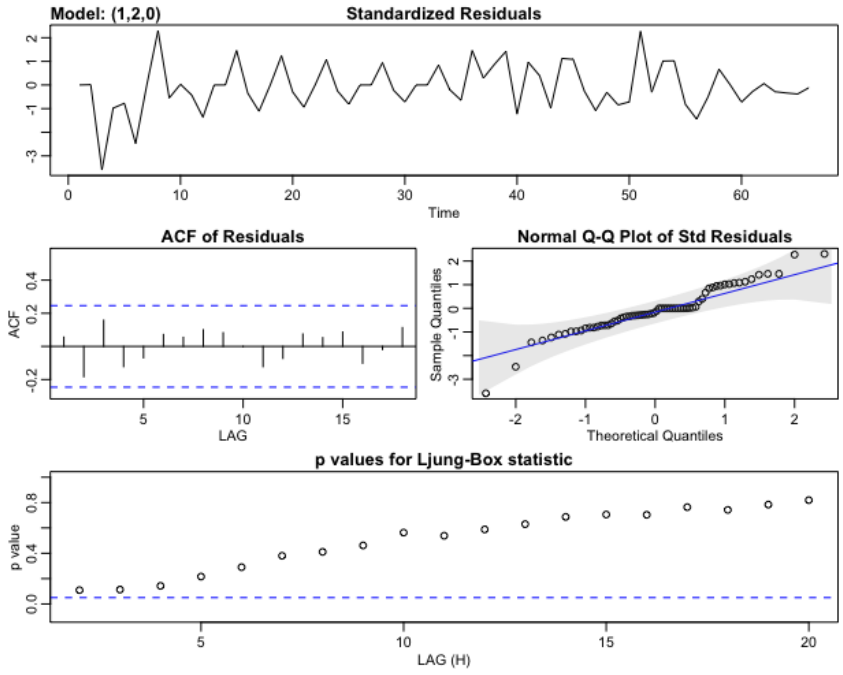

Figure 2: Model 1 ARIMA(1,2,0) diagnostic plots.

2.2 Model 2

Again we decided to take a second difference of the data to obtain a stationary process. To address the heteroskedasticity we decided to take a Box-Cox transformation of the data with a $\lambda = 0.05$. We decided upon this λ through iterating through λ = (-1,1) by increments of 0.01 and then selecting the λ that maximized variance stability.

After examining the ACF and PACF, we found large spikes at lag = 1 and 2 of the ACF. We decided on adding an MA(2) process to the model as the PACF looked as though it was decreasing exponentially.

After adding the MA(2) process, we looked at the periodogram of the residuals and found no obvious seasonal trends. Plots of the model residuals, ACF of residuals, normal Q-Q standard residuals, and Ljung-Box test are presented in the figure below.

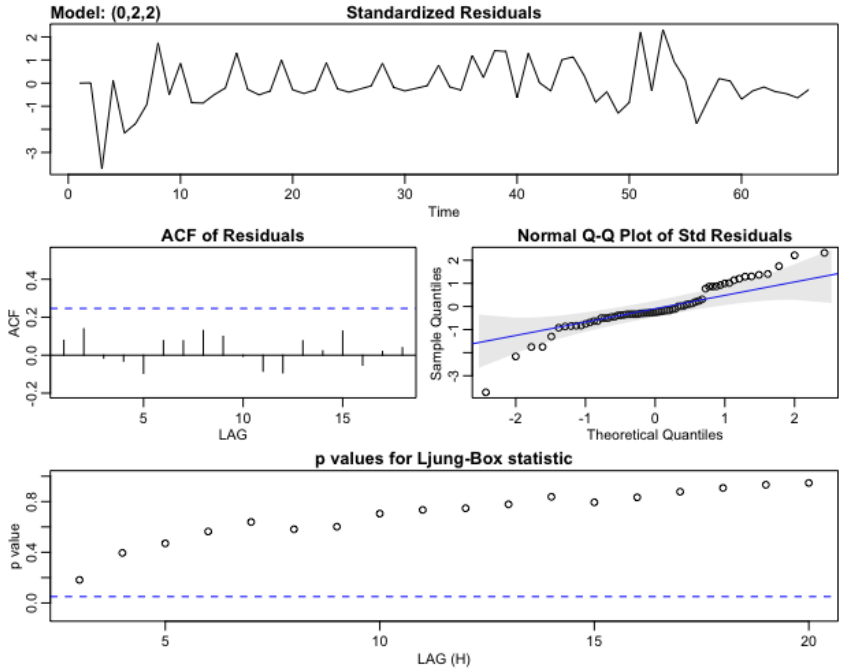

Figure 3: Model 2 ARIMA(0,2,2) diagnostic plots.

2.3 The Group Model

Finally for completeness, we opted for a third order differenced model of the data. We decided to include this model because we wanted to explore whether social distancing (that may have been implemented within the data's time frame) may have created a third order trend. We suspected that this model may not fit the early data as well, but may be more accurate in forecasting future COVID19 counts. After taking the third order difference, the variance appeared to still be increasing with time, though at a slower rate than the second differenced model; therefore, we decided to use a Box-Cox VST with $\lambda = 0.05$.

Looking at the ACF and PACF of the residuals of the transformed triple differenced data, we found large spikes at lag = 1 of the ACF and two significant spikes at lag = 1 and 2 of the PACF. We added an AR(2) process to the model but now saw significant spikes at lag = 1 and 2 of both the ACF and PACF. We now suspected that the model had a MA component and an AR component. Hence, we created an ARIMA(1,3,1) model of the transformed data and found no significant lags in the ACF or PACF plots, indicating that the residuals can be modeled as white noise. The Ljung-Box statistics confirms this notion with only lag one's statistic having at or below a 5 percent significance level. Plots of the model residuals, ACF of residuals, normal Q-Q standard residuals, and Ljung-Box test are presented in the figure below.

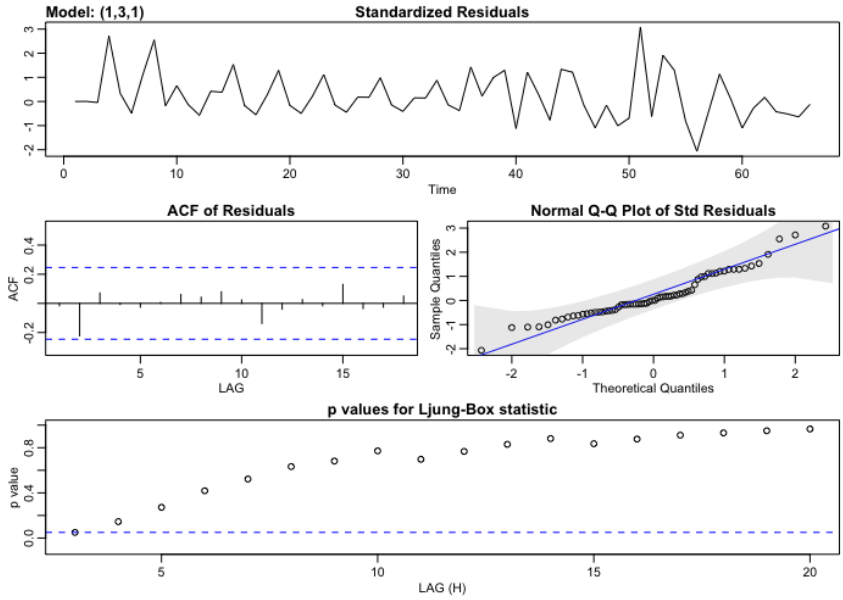

Figure 4: Model 3 ARIMA (1,3,1) diagnostic plots.

3 Model Comparison and Selection

To select a model for the 10-day prediction, we evaluated the out-of-sample fit of each of the three models. To determine which model performs best out of sample, we used cross validation by dividing the data into a 56-day training set and a 10-day test set. We then used the root mean square error (RMSE) as the metric of performance for each model. A summary of each model and the corresponding RMSE values are presented in the table below (Table 1).

Model Name	Description	RMSE
Model 1	ARIMA(1,2,0)	11,795
Model 2	ARIMA(0,2,2)	14,753
Model 3	ARIMA(1,3,1)	8,747

Table 1: Out-of-sample RMSE values for each model.

Based on the out-of-sample fit, the model 3 performed the best, so we decided to use this model for our predictions. This model will look best if the third order trend that becomes more apparent in the later part of the time series continues through the prediction interval.

4 Results

The generic time series model generally consists of three main components as defined in equation (1).

$$f(Y_t) = m_t + s_t + X_t \tag{1}$$

In this equation $f(Y_t)$ is the Box-Cox transformed COVID-19 case count, m_t is the trend of the data, s_t is the seasonality of the data, and X_t models the residuals. We accounted for the trend (m_t) of the transformed data was by taking a third order difference. After detrending the data, there was no apparent seasonality, so we set this term (s_t) to 0. We then modeled the residuals (X_t) as an $ARMA(1,1)$ process. The final resulting model is an $ARIMA(1,3,1)$ model defined by equation (2).

$$(1+\phi B)\nabla_3 X_t = (1+\theta B)W_t \tag{2}$$

In this equation, ϕ is the AR(1) coefficient, θ is the MA(1) coefficient, ∇_3 removes the trend with a third order difference of the transformed data $(f(Y_t))$, and W_t is a white noise process. The model parameter estimates, as well as a 10-day forecast are presented in the following sections.

4.1 Estimation of model parameters

We used maximum likelihood to estimate the parameters of the $ARIMA(1, 3, 1)$ model, the equation of which is presented in the previous section (equation (2)). The estimated model parameters and the corresponding standard errors are presented in the table below (Table 2).

Parameter	Estimate	(s.e)
ϕ	-0.7096	(0.1046)
θ	-0.9999	(0.0877)

Table 2: Parameter estimates and corresponding standard errors for the ARIMA model in equation 2.

4.2 Prediction

Based on the ARIMA model in equation (2), we anticipate the total COVID-19 case count will continue to increase reaching approximately 110,000 total cases 10 days from now. Our estimated 10-day forecast is shown on the figure below (Figure 5).

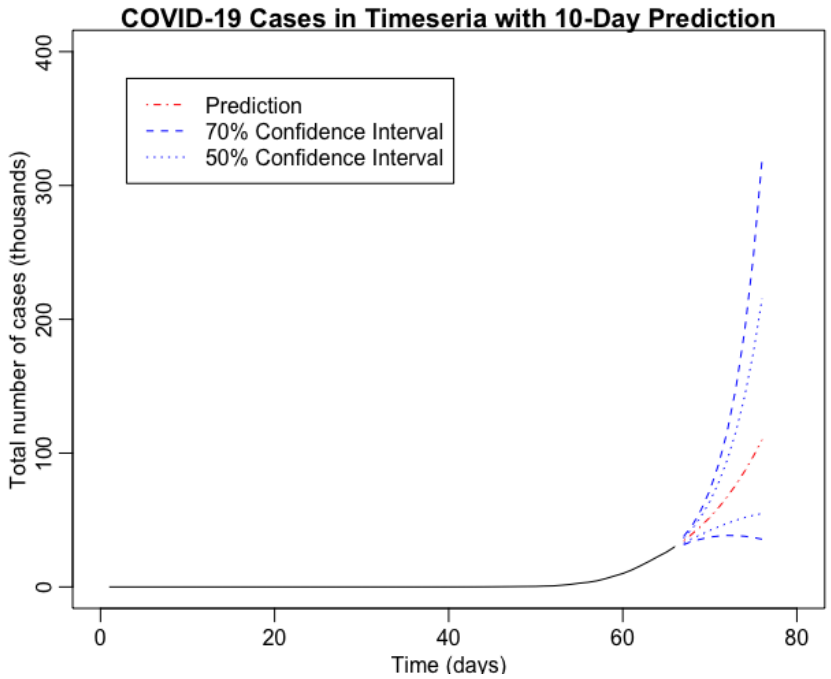

Figure 5: Estimated number of COVID-19 cases predicted to occur in the next 10 days.

The 70 percent and 50 percent confidence intervals are also presented on Figure 5. The 70 percent confidence intervals indicates the number of total COVID-19 cases 10 days from now could range from approximately 38,000 to 320,000. The upper level of our 95 percent confidence interval shows the total number of cases could read up to about 789,000 cases. Our best 10-day estimate shows more than a 200 percent increase in total COVID cases from now, demonstrating the importance of lowering the rate of COVID-19 exposure within the country of Timeseria.