Autoregressive (AR) model for Singapore COVID-19 2020 data

I used the COVID-19 Singapore time series data collected by the Ministry of Health (MOH) Singapore <https://data.world/hxchua/covid-19-singapore/workspace/file?filename=Covid-19+SG.xlsx>.

First, by making a regular time series data plot as seen in Figure 1, we can see that the COVID cases shot up in April and gradually decreased over the summer months, until another spike in cases in mid-August, and levelling back to close to 0 afterwards. There was a concentrated increase of COVID cases from the end of March till mid-September. There is no clear pattern that can be detected from this plot.

Chart, histogram

Description automatically generated

Figure 1: Time Series Plot of Original Data

I decided to log the data in case there might be any skewness in the data. Because there were a couple data points that were equal to 0, I added a constant, 1, to calculate the log function to keep the data points proportional since log(0) cannot be computed. Thus, by logging the data with log(x+1), I got the time series plot as shown in Figure 2.

Chart, line chart

Description automatically generated

Figure 2: Time Series Plot of logCases2

To choose the right ARIMA model parameters, I first used the AICc to see if the data needed any differencing:

If p+q = 1, d =0 with constant:

The AICc = 290log(59.5702/290)+2(1+2)(290/290-1-3)

=290log(0.20541448)+2(3)(290/286)

=-452.906334

If p+q = 1, d =1 with constant:

The AICc = 289log(53.9753/289)+2(1+2)(289/289-1-3)

= 289log(0.186766)+2(3)(289/285)

=-478.8286

If p+q = 1, d =2 with constant:

The AICc = 288log(103.572/288)+2(1+2)(288/288-1-3)

= 288log(0.359625)+2(3)(288/284)

=-288.451077

If p+q = 1, d =1 with no constant:

The AICc = 289log(53.9753/289)+2(1+1)(289/289-1-2)

= 289log(0.186766)+2(2)(289/286)

=-480.870853

If p+q = 1, d =2 with no constant:

The AICc = 288log(103.572/288)+2(1+1)(288/288-1-2)

= 288log(0.359625)+2(2)(288/285)

=-290.493479

From the calculations computed above, d = 1 had the most negative AICc. Thus, the AICc criterion assumes d = 1 with no constant is the best choice to go with compared to all other options for differencing in the ARIMA model.

In addition to AICc, I compared the autocorrelations and partial-autocorrelations of d = 0,1,2 to select the most attractive *d* in the ARIMA model. The ACF and PACF for the logCases2 show that the data needs to be differenced because the lags in the ACF are hanging around 1.0 in Figure 3; and the first lag of the PACF being close to 1.0, while the other lags are tiny only ranging ±0.2 at most from 0 in comparison as shown in Figure 4.

Chart, histogram

Description automatically generated

Figure 3: ACF logCases2

Chart

Description automatically generated

Figure 4: PACF logCases2

Thus, I tried differencing the logCases2 next. The ACF and PACF are much more attractive with the first lag of the ACF being ~-0.3 with no significant lags after; as well as the first lag of the PACF being around -0.3 with a cut-off at lag 3 (Figure 5, Figure 6 respectively). The slight negative lags show that the data does not need to be further differenced. Just to confirm that the data does not need to be differenced again, I included the ACF and PACF with d = 2 in Figure 7 and 8. You can see that the lags are less than -0.5 for both the ACF and PACF, indicating that the data was over differenced.

Chart

Description automatically generated

Figure 5: ACF of logCases2 with d =1

Chart

Description automatically generated

Figure 6: PACF of logCases2 with d =1

Chart

Description automatically generated

Figure 7: ACF of logCases2 with d =2

Chart

Description automatically generated

Figure 8: PACF of logCases2 with d =2

The parameter estimates for the model are:

A picture containing graphical user interface

Description automatically generated

Figure 9

Thus, the optimal model from my analysis would be ARIMA(1,1,0). The SE coefficient is 0.0558 for AR(1), with a T-value of -5.88, meaning there is a significant difference. The complete form of the fitted model would be yt=-0.3277t-1+et

From the Ljung-Box statistics, lags 12 and 24 are deemed to be inadequate because its p-value is less than 0.05. The p-values for the rest, 36 and 48 lags, are all greater than 0.05.

The ACF and PACF of the residuals (Figure 10, Figure 11) look like white noise, thus, the model is adequate.

A picture containing chart

Description automatically generated

Figure 10: ACF of Residuals

Chart

Description automatically generated

Figure 11: PACF of Residuals

The 95% forecast intervals (Figure 12) and forecast (Figure 13) are in the tables below. The forecast intervals seems a bit too wide. When I tried fitting the forecast model at origin time 150 or June 20,2 020, the forecast interval range was quite wide with the real values collected just hovering close to the mean. The data points are not too volatile; thus, the forecast interval seems excessively wide.

A picture containing text

Description automatically generated

Figure 12: 95% Forecast Intervals

Chart, scatter chart

Description automatically generated

Figure 13: Forecast with lead time 50 at time 290 (or November 7, 2020)

Chart, scatter chart

Description automatically generated

Figure 14: Forecast with lead time 50 at time 150 (or June 20, 2020)