Indian Institute of Technology (BHU) Varanasi

Aryan Patel

Integrated Dual Degree (B.Tech + M.Tech) in Mathematics and Computing

Application of the ARIMA model on the COVID-19 epidemic dataset

June 2020

Department of Mathematical Sciences

Abstract

Coronavirus disease 2019 (COVID-19) is an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). It was first identified in December 2019 in Wuhan, China, and has resulted in an ongoing pandemic and a major worldwide threat. Several studies are being conducted using various mathematical models to predict the probable evolution of this epidemic. These mathematical models based on various factors and analyses are subject to potential bias. Here, a simple econometric model that could be useful to predict the spread of COVID-2019 has been proposed. Auto Regressive Integrated Moving Average (ARIMA) model has been performed on the Johns Hopkins epidemiological data to predict the epidemiological trend of the virus.

Contents

Chapter 1

Introduction

The COVID-19 pandemic commenced from December 2019 in Wuhan, China and has caused extreme damage in almost the entire world [6]. 2019-nCoV or COVID-19, commonly known as Coronavirus, is a novel highly contagious virus belonging to Coronaviridae family that has been suspected to be transmitted to humans from animals. This virus causes mild to severe respiratory illness and death.

This pandemic has engulfed 187 countries/regions in merely four months infecting 9,609,088 people and taking the death toll to 489,296 as of 25 June 2020 [3]. However, the premature cases show the infection is less severe as compared to other coronaviruses such as SARS-CoV (Severe Acute Respiratory Syndrome Corona Virus) and MERS-CoV (Middle East Respiratory Syndrome Corona Virus), the cases of rapid human-to-human transmission signify that 2019-nCoV is highly infectious than others. Although a local seafood market in Wuhan is believed to be the source of exposure, the scope of occurrence of this disease is not clear since its occurrence at present is very dynamic [6]. An apparent variation is present in epidemiological examinations and detection abilities performed by different countries for detecting infected cases. Presently, the highest cases of COVID-19 infections have been reported in the US, however, the cases are abruptly rising in Brazil, Russia, India and UK daily [3]. China, the place of origin of the disease, is now receiving a very few cases.

At present, there is neither a treatment nor a vaccination for the COVID-19 infection. In this circumstance, the only option is preventing the occurrence of infection and preparing our healthcare system for the probable future. With that in mind, it is vital to construct models that

are computationally competent as well as realistic so that they can help policy makers, medical personnel and also general public in decision making. Modeling the disease and providing future forecast of possible number of daily cases can assist the medical system in getting prepared for new patients.

The statistical prediction models are useful in forecasting as well as controlling the global epidemic threat. In this project, Auto-Regressive Integrated Moving Average (ARIMA) model has been employed for predicting the incidence of COVID-19 cases. The best ARIMA model has been identified and then the number the cases for the next 21 days are predicted. The main objective of the study is to find the best predictive model and apply it to forecast future incidence of COVID-19 cases.

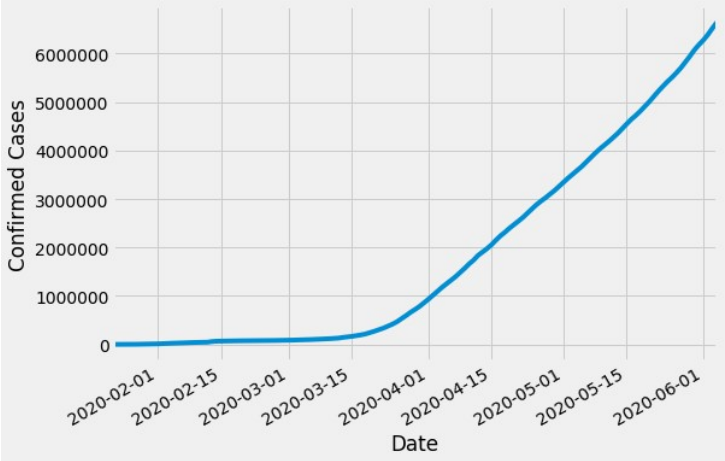

Figure 1.1: Plot of Confirmed COVID-19 cases vs Date (during 22 Jan 2020 - 4 June 2020)

Chapter 2

Methodology

2.1 Dataset

Data comprising of confirmed, recovered and death cases of worldwide COVID-19 infections during 22 January 2020 - 4 June 2020 is collected, as per World Health Organization classification, from the official website of Johns Hopkins University [4]. This data is used to build predictive models.

2.2 Model Development

For forecasting a time series, ARIMA modeling is one of the best modeling techniques. ARIMA models are always represented with the help of some parameters and the model is expressed as ARIMA (p, d, q). Here, p stands for the order of auto-regression, d signifies the degree of trend difference while q is the order of moving average. So, for parameters let's say = (2, 1, 1), the model would be:

$$ARIMA(2,1,1) : \triangle y_t = \alpha_1 \triangle y_{t-1} + \alpha_2 \triangle y_{t-2} + \beta_1 \epsilon_{t-1}, where \triangle y_t = y_t - y_{t-1} \tag{2.1}$$

Here, y_t is the predicted number of confirmed COVID-19 cases at tth day, α_1, α_2, β_1 are parameters, and ϵ_t is the residual term for tth day.

In this case, ARIMA model has been applied to the time series data of confirmed COVID-19 cases worldwide. A grid search is performed to obtain the best values of the parameters p, d, q. The grid search approach involves evaluating the model for different parameters and comparing

the error score in each case [2]. The parameters thus obtained are $(7, 2, 1)$. Plot diagnostics is performed to quickly generate model diagnostics and investigate for any unusual behavior [1]. This is to ensure that the residuals of our model are uncorrelated and normally distributed with zero-mean. If the seasonal ARIMA model does not satisfy these properties, it is a good indication that it can be further improved.

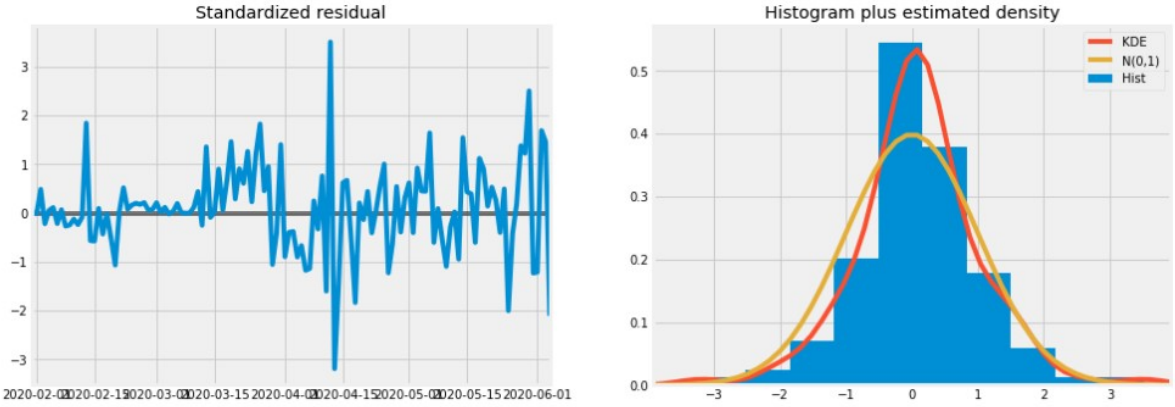

Figure 2.1: Plot Diagnostics (a)

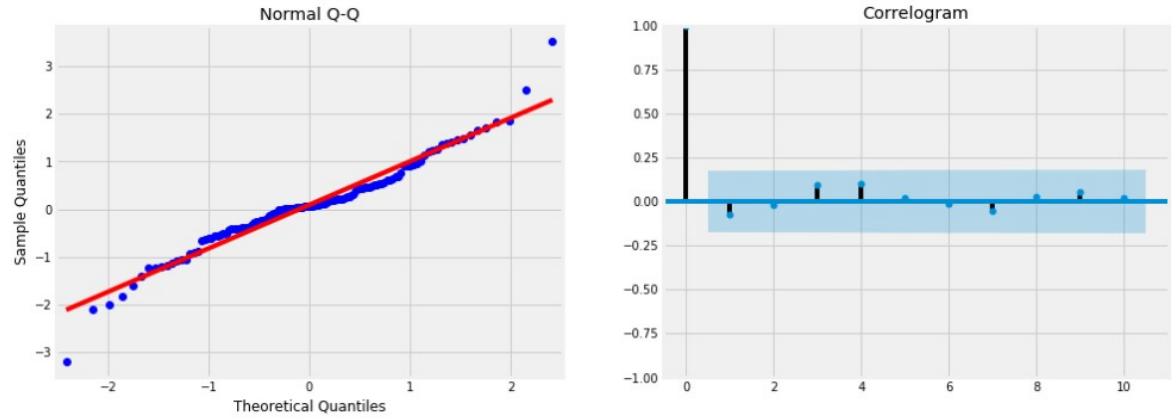

Figure 2.2: Plot Diagnostics (b)

In this case, the model diagnostics suggests that the model residuals are normally distributed based on the following:

- In the top right plot, we see that the red KDE line follows closely with the N(0,1) line (where N(0,1)) is the standard notation for a normal distribution with mean 0 and standard deviation of 1). This is a good indication that the residuals are normally distributed.
- The qq-plot on the bottom left shows that the ordered distribution of residuals (blue dots) follows the linear trend of the samples taken from a standard normal distribution with

N(0, 1). Again, this is a strong indication that the residuals are normally distributed.

- The residuals over time (top left plot) don't display any obvious seasonality and appear to be white noise. This is confirmed by the autocorrelation (i.e. correlogram) plot on the bottom right, which shows that the time series residuals have low correlation with lagged versions of itself.

Those observations lead us to conclude that our model produces a satisfactory fit that could help us understand our time series data and forecast future values. The built model is employed to forecast confirmed COVID-19 cases for the next 21 days, i.e. 5 June 2020 to 25 June 2020.

The trend of forthcoming incidences can be estimated from the previous cases and a time series analysis is performed for this purpose. Time series forecasting refers to the employment of a model to forecast future data based on previously observed data. In the present study, time series analysis is used to recognize the trends in confirmed COVID-19 cases over the period of 22 January 2020 to 4 June 2020 and to predict future cases from 5 June 2020 till 25 June 2020. A graph is plotted for actual confirmed cases and predicted confirmed cases with respect to time to verify the efficiency of the model (Figure 3.1). To get an idea of the recovery and death trends, a graph is plotted with respect to time (Figure 3.2). All the model developments, computations and comparisons have been performed in Python 3.7 using Jupyter notebook [5].

Chapter 3

Result and Discussion

Parameters are estimated for the ARIMA $(7, 2, 1)$ model which are displayed in Figure 3.1. Hence, the workable model obtained after the substitution of estimated parameters is used to forecast confirmed COVID-19 cases for the next 21 days, i.e. 5 June 2020 to 25 June 2020. The forecast for cases is presented in Table 3.1.

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.2978	0.123	2.420	0.016	0.057	0.539
ar.L2	-0.0453	0.076	-0.593	0.553	-0.195	0.104
ar.L3	-0.1611	0.101	-1.589	0.112	-0.360	0.038
ar.L4	0.1535	0.123	1.248	0.212	-0.088	0.395
ar.L5	-0.0835	0.120	-0.697	0.486	-0.318	0.151
ar.L6	0.1702	0.096	1.780	0.075	-0.017	0.358
ar.L7	0.5054	0.133	3.810	0.000	0.245	0.765
ma.L1	-0.6263	0.132	-4.761	0.000	-0.884	-0.369
sigma2	3.869e+07	1.24e-09	3.11e+16	0.000	3.87e+07	3.87e+07

Figure 3.1: Parameters of the ARIMA Model

Table 3.1: Figures for forecast confirmed COVID-19 cases and their lower and upper limits for 21 days (5 June 2020 to 25 June 2020)

Date	Forecast	Lower limit	Upper limit
2020-06-05	6.771272e+06	6.759081e+06	6.783464e+06
2020-06-06	6.902952e+06	6.879206e+06	6.926698e+06
2020-06-07	7.025310e+06	6.989489e+06	7.061130e+06
2020-06-08	7.138065e+06	7.090721e+06	7.185409e+06
2020-06-09	7.260835e+06	7.200744e+06	7.320926e+06
2020-06-10	7.396612e+06	7.323081e+06	7.470142e+06
2020-06-11	7.532790e+06	7.443887e+06	7.621693e+06
2020-06-12	7.669385e+06	7.560295e+06	7.778474e+06
2020-06-13	7.806015e+06	7.672873e+06	7.939157e+06
2020-06-14	7.937388e+06	7.778209e+06	8.096567e+06
2020-06-15	8.062954e+06	7.876985e+06	8.248922e+06
2020-06-16	8.194327e+06	7.980204e+06	8.408449e+06
2020-06-17	8.335151e+06	8.091432e+06	8.578871e+06
2020-06-18	8.478927e+06	8.203263e+06	8.754591e+06
2020-06-19	8.621981e+06	8.310802e+06	8.933161e+06
2020-06-20	8.763665e+06	8.413483e+06	9.113847e+06
2020-06-21	8.901818e+06	8.510311e+06	9.293326e+06
2020-06-22	9.036816e+06	8.602534e+06	9.471097e+06
2020-06-23	9.175440e+06	8.696814e+06	9.654065e+06
2020-06-24	9.320984e+06	8.796169e+06	9.845800e+06
2020-06-25	9.469875e+06	8.896414e+06	1.004334e+07

Time series analysis of the COVID-19 dataset provides meaningful statistical information. Figure 3.2 shows series graph of the active infected COVID-19 cases from 22 January 2020 to 25 June 2020. It is clear from this figure that the time series is not stationary. An increasing trend is displayed by the time series suggesting a high rise in COVID-19 cases.

For comparing the actual and forecast confirmed COVID-19 cases, a time series graph is plotted starting from 5 June 2020 till 25 June 2020. The plot is represented by Figure 3.3. The similarity of forecast data with actual data is clear from these plots. This comparison reveals the precision of the model in forecasting.

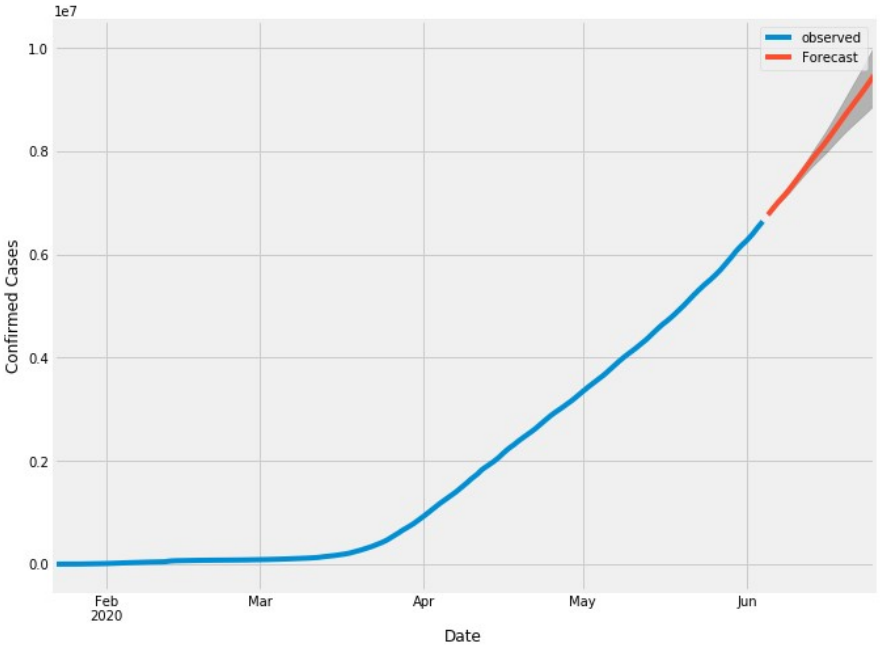

Figure 3.2: Times series plot for forecast confirmed COVID-19 infections from 22 January 2020 to 25 June 2020

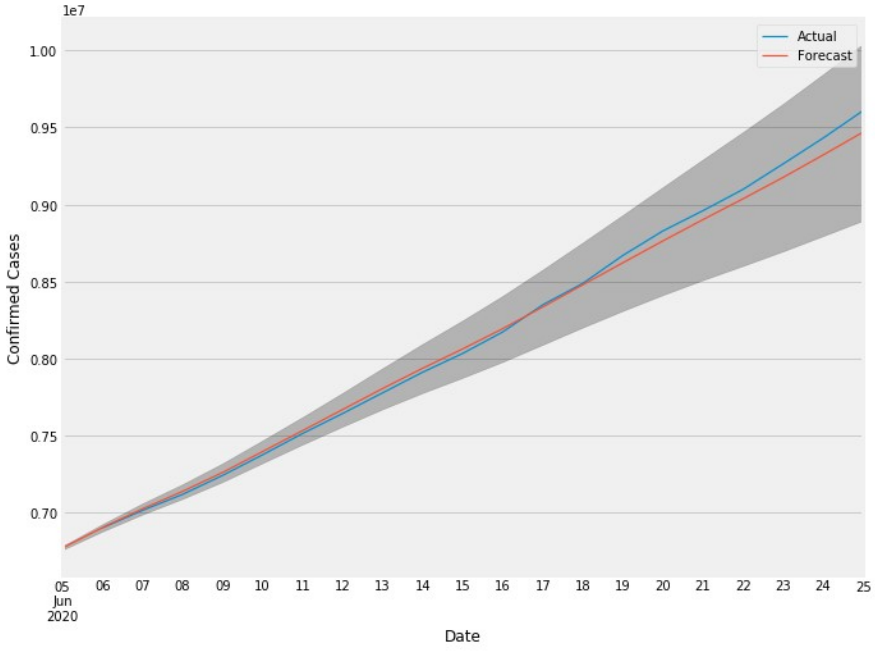

Figure 3.3: Comparison of Actual and Forecast Confirmed COVID-19 Cases from 5 June 2020 to 25 June 2020

Trend for the number of recovery and death cases with respect to time due to COVID-19 infections is depicted in Figure 3.4. It is observed that the number of recoveries as well as deaths increase with time, however the rate of recovery is higher than the death rate. Thus, a low mortality rate could be expected from the disease.

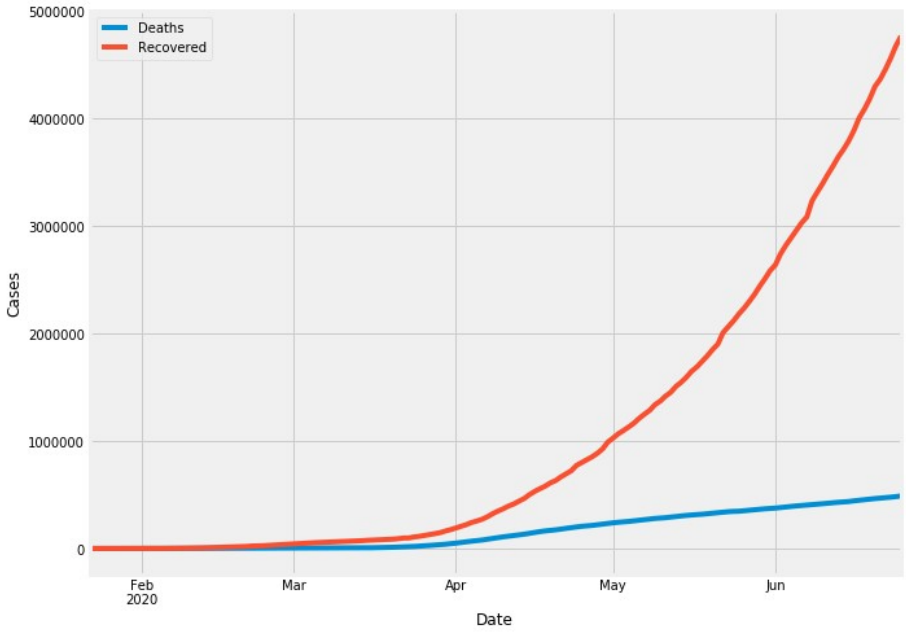

Figure 3.4: A comparative trend for the number of recoveries and deaths due to COVID-19 infections from 22 January 2020 to 4 June 2020

It has been established that measures like lockdown, quarantine, and sanitization, social distancing, and mask usage can decrease human exposure and control this pandemic [7]. Thus, these measures should be imposed firmly and strict actions must be taken against those people who violate the rules and fail to recognize the severity of the situation. Although a large amount of data helps provides a more exhaustive prediction and explanation, in the current scenario, these models could be valuable in anticipating near future cases of infection if the pattern of virus spread does not change in an unusual manner. Clearly, this virus is novel and has the capability to be transmitted and mutated acutely, which can affect predictions too far into the future.

Chapter 4

Conclusion

The novel coronavirus disease (COVID-19) has been declared as pandemic by WHO and is currently a major global threat [7]. This project aims to examine a decently accurate model for the prediction of observed COVID-19 infection cases and to employ said model for forecasting future COVID-19 cases in order to support the fight against this pandemic and assist in the preparation of healthcare services. As per the model's forecasts, the confirmed cases are expected to rise significantly in the coming days. The time series analysis shows an exponential enhancement in the infected cases. However, concerted government and civilian efforts such as lockdown, social distancing, mask use, or a cure, which seems far fetched as of now, can lead to a decline in the cases after at least couple of month. The prediction models will help the government and medical workforce to be prepared for the upcoming situation and have more readiness in healthcare systems. For further comparison or for future perspective, case definition and data collection must be maintained in real time.

References

[1] *A Guide to Time Series Forecasting with ARIMA in Python 3*. Digital Ocean. 2020. URL: https://www.digitalocean.com/community/tutorials/a-guide-to-time-series-forecasting-with-arima-in-python-3 (visited on 06/27/2020).

[2] Jason Brownlee. *Introduction to Time Series Forecasting With Python*. 2017.

[3] *Coronavirus Report on Kaggle by imdevskp*. 2020. URL: https://www.kaggle.com/imdevskp/corona-virus-report (visited on 06/27/2020).

[4] *COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University*. Johns Hopkins University. 2020. URL: https://github.com/CSSEGISandData/COVID-19 (visited on 06/27/2020).

[5] *Jupyter Notebook*. Project Jupyter. URL: https://jupyter.org/.

[6] *Wikipedia, 2019-20 coronavirus outbreak*. Wikipedia. 2020. URL: https://en.wikipedia.org/wiki/2019-20_coronavirus_outbreak (visited on 06/27/2020).

[7] *World Health Organization, Coronavirus disease (COVID-19) outbreak*. World Health Organization (WHO). 2020. URL: https://www.who.int/emergencies/diseases/novel-coronavirus-2019 (visited on 06/27/2020).