A Mathematical Model for COVID-19 to Predict Daily Cases using Time Series Auto Regressive Integrated Moving Average (ARIMA) Model in Delhi Region, India

Tarunima Agarwal, Modern School,
Barakhamba Road, New Delhi, India

Stavelin Abhinandithe K, Assistant Professor,
Division of Medical Statistics, Faculty of Life
Sciences, JSSAHER, Mysuru, Karnataka, India

Abstract:
INTRODUCTION: Coronavirus disease (COVID-19) is an infectious disease caused by a coronavirus that is circulating worldwide. Various countries have used various measures to combat the disease's spread. Many studies have adopted the mathematical modeling to predict the cases during the pandemic. In our study we have used Box- Jenkins's Auto Regressive Integrated Moving Average (ARIMA) time series mathematical model. MATERIALS AND METHODS: Publicly available data of daily COVID-19 confirmed cases along with Meteorological variables were considered using Expert Modeler in SPSS to predict and forecast COVID-19 cases in Delhi region, India. RESULTS: Spearman's correlation was used to find the relationship between COVID-19 cases with Meteorological variables. Humidity, rainy days and Average sunshine were found to be significant. ARIMA (0, 1, 14) model found to be best fitted model for the given data with R square value of fitted model is 0.920. Ljung-Box test value is 39.368 with p value showing significant, indicating that the fitted model is adequate to predict and forecast COVID-19 cases. CONCLUSION: ARIMA (0, 1, 14) mathematical model was selected as a best suited model to predict and forecast the incidence of COVID-19 cases in Delhi region, which would be useful for the policymakers for better preparedness.

Keywords:- *COVID-19, Mathematical Model, ARIMA Model, Meteorological Variables, Ljung-Box test.*

I. INTRODUCTION

The 2019 coronavirus pandemic (COVID-19) has spread and acquired worldwide significance after being initiated by a new coronavirus (SARS-CoV-2, previously known as 2019-nCoV). Burgeoning and resurging contaminants pose international public health challenges [1]. As of May 13, 2020 COVID-19 was verified in even more than 4.2 million people around the world having culminated in far more than 293,000 fatalities. And over 180 nations from all continents apart from Antarctica had also revealed laboratory-confirmed occurrences with COVID-19 [2]. As of May 21, 2020, the total numbers of COVID-19 cases in India were reported to be 1,12,442. India was one of 11 countries with more than a million cases but fewer deaths. Only 2.94 percent of Indians who tested positive were on oxygen, 3% were in intensive care units, and 0.45 percent was on ventilator support, according to reports. Throughout the present lockdowns, if the appropriate social distancing and preventive measures were not observed or lack implementation, the total number of cases could rise exponentially. The main emphasis should indeed be on increasing the COVID testing, obtaining the N95 masks as well as personal protective equipment for healthcare workers, and ventilators [3]. Countries with increase of COVID-19 cases including India have already implemented alternative strategies to slow the transmission of the infection. Many have thoroughly checked traced contacts, restricted transportation and public gatherings, facilitated social distancing and implemented lockdowns. Some have incorporated complete/partial shutdowns in cities with large rates of infection, with strict protocols in place [4-7].

Mathematical models have been developed during the pandemic to study the disease dynamics, forecast the number the cases, and strengthen the public health and social measures to limit the transmission of COVID-19 and reduce deaths. In our study we have used Box- Jenskin's Auto Regressive Integrated Moving Average (ARIMA) mathematical model.

II. MATERIALS AND METHODOLOGY

A. Study Area: The current research was carried out in Delhi, India's National Capital Territory (NCT), which is a city and union territory. On three sides, it is bordered by Haryana, and on the east, it is bounded by Uttar Pradesh. The NCT is 1,484 square kilometers in size (573 sq mi). According to the 2011 census, the city of Delhi had a population of over 11 million people, second to Mumbai in India, while the entire NCT area had a population of about 16.8 million. The National Capital Region (NCR), which includes the surrounding satellite cities of Ghaziabad, Faridabad, Gurugram, and Noida, has an estimated 2016 population of over 26 million people, making it the world's second-largest urban region, according to the United Nations.

B. Data: Publicly available data of daily COVID-19 confirmed cases along with Meteorological variables were considered using Expert Modeler in SPSS to predict and forecast COVID-19 cases in Delhi region, India. The day wise data of confirmed COVID-19 cases was collected from 12th March 2020 to 31st December 2020.

C. Data Analysis:

Correlation Analysis: Correlation quantifies the "strength" or the "degree" of a relationship between the different factors. We have adopted Spearman's correlation to know the relationship between COVID-19 with meteorological variables.

Time Series Analysis: A time series is a sequence of observations assumed over time. In regression analysis, time series models are most commonly used to explain the response variable's dependence on predictor variables such as covariates and maybe prior values in the series at each time.

During the last two decades, various time series models are used to predict and forecast infectious diseases. One such model is Auto Regressive Integrated Moving Average (ARIMA) models. This model was first introduced by Box-Jenkins. The Box-Jenkins ARMA (Autoregressive Moving Average) model is a mix of the AR (Auto Regressive) and MA (Moving Average) models. The Box-Jenkins model assumes that the temporal arrangement is fixed; hence it requires that non-fixed arrangements be differed at least several times to achieve stationary. As a result, an ARIMA model is created, with the "I" standing for "Integrated."

Autoregressive Integrated Moving Average (ARIMA)

In measurements, econometrics and building autoregressive coordinated moving normal (ARIMA) models are one of the most significant and generally utilized direct time arrangement models. The ARIMA model is main stream because of its attractive measurable properties just as the notable Box–Jenkins strategy (Box and Jenkins, 1976) in the model structure process. Furthermore, extraordinary exponential smoothing models can likewise be executed in ARIMA models. Despite the fact that ARIMA models are very adaptable as in they can speak to a few unique kinds of time arrangement and furthermore have the benefits of making precise forecast over a brief timeframe and gives simplicity of execution, their significant impediment is that they expect a direct structure for the model with no legitimization. ARIMA models accept that future estimations of a period arrangement have a characteristically straight relationship with present and past qualities just as with repetitive sound, subsequently approximations got by ARIMA models may not be seen as satisfactory on account of complex nonlinear genuine issues. They are regularly applied in situations where information show proof of non-stationary, where an underlying differencing step, which compares to the "coordinated" some portion of the model, can be applied to diminish the non stationarity. Nonetheless, numerous specialists have contended that numerous genuine frameworks are regularly nonlinear. These confirmations have empowered scholarly analysts and business experts to endeavor to grow more unsurprising forecast models than basic straight models. The ARIMA model is numerically or emblematically spoke to as an ARIMA (p,d,q) model, where the parameters p, d, and q are nonnegative numbers and allude to the sets of the autoregression, combination, and moving normal pieces of the model individually. ARIMA models structure a significant piece of the Box-Jenkins way to deal with time arrangement displaying. For instance, ARIMA(0,1,0) is same as I(1), and ARIMA(0,0,1) is only MA(1). The general ARIMA model combines three processes:

Autoregressive (AR) process

An autoregressive process of order p, denoted by the notation AR (p), is a stochastic process (Xt), where t is a white noise process. This condition officially speaks to a multiple regression, despite the fact that, in this occurrence the deciding variable isn't a free factor, however the memorable estimation of Xt itself. Adroitly, an autoregressive procedure is unified with a "memory", as in each value is associated with every single going before esteem. Following this translation, each incentive in an AR(p)- process is controlled by p going before values, where more established qualities will have a blurring impact. Low order forms, consequently, just have a "short memory". In an AR(1) process, additionally composed as ARIMA (1,0,0), the present worth is an element of the previous worth, which is a component of the one going before it, etc.

Moving Average (MA) Process

A stochastic process (Xt) is a moving average process of the order q, indicated as MA(q) or ARIMA (0,0,q), where εt is a white noise process. The distinction between an autoregressive procedure and a moving average procedure is unobtrusive however significant. Each incentive in a moving-normal arrangement is a weighted normal of the latest arbitrary unsettling influences, when each incentive in an autoregression is a weighted average of the ongoing estimations of the arrangement. Since these qualities thus are weighted midpoints of the past ones, the impact of a given aggravation in an autoregressive procedure wanes over the long haul. In a moving average procedure, an unsettling influence influences the framework for a limited number of periods (the request for the moving normal) and afterward unexpectedly stops to influence it.

Differencing-Integration

A time series that mirrors the total impact of some procedure is called coordinated. Such a period arrangement has a pattern and is in fixed. The stationary of an arrangement is fundamental for the estimation of AR and MA forms. Accordingly, time series that show a pattern ought to be differenced, until stationarity is practiced. When all is said in done first or second request differencing will be adequate for arrangement with a pattern to guarantee stationarity.

Stages of building model of the Time Series

In order to construct a Box-Jenkins time series model, there are four steps. These are model selection, model parameter estimation, residuals demonstrative check, model sufficiency, and determination.

D. Diagnosis

Identifying whether the time series is stationary or non-stationary is the first stage in building a Box–Jenkins model. The diagnostic phase includes the following steps:

- **Graphical Analysis**

This is the phase to introduce the information for example to draw the arrangement with the time, subsequent to drawing information that is the initial phase in investigation of whenever arrangement and through the drawing we would have a smart thought about the fixed, for example does the arrangement contain the occasional segment or the pattern or the strange qualities… and so on. We can come to know the non-fixed of arrangement just by the drawing information. Accordingly the sketch arrangement shows the need of right exchange with the goal that this arrangement can stable in its Mean and fluctuation before beginning any investigation.

- **Autocorrelation and Partial Autocorrelation Function Plots**

The autocorrelation function (ACF) and partial correlation function (PACF) are used to detect whether a time series is stationary or non-stationary, and the Ljung-Box Chi-Square test (Q – test) is used to test for significant autocorrelation coefficients. To determine whether the selected model is a statistically sufficient description of the time series, a variety of approaches can be used.

- **Identification**

Following the achievement of time-series stationary, the identification model process begins, which entails the use of data regarding how time-series are formed. The aim of getting an idea of the value p, d and q of the general linear model ARIMA and then get on preliminary estimates of the model parameters. This is done using ACF and PACF, and the following table shows the comparison between ACF properties and PACF properties. The order of the ARIMA model's separate processes is determined during the model's identification.

	ACF	**PACF**
AR(p)	Close to Zero gradually	Reaches to zero after the period of time (p)
MA(q)	Reaches to zero after the period of time(q)	Close to zero gradually
ARMA(p,q)	Close to zero gradually	Close to zero gradually

- **Estimation**

The estimation of the identified model's individual parameters is the next stage. The number of coefficients used to describe the model is proportional to the model's order. After then, appropriate algorithms are used to determine the coefficients. The final coefficients are used to calculate the values for the modeled series, the residual variance, and the related confidence intervals.

E. Model Structure Selection Criteria

Box and Jenkins (1976) proposed a non-seasonal ARIMA model that will be used to illustrate the various model structures with the obvious notations (p, d, q). For model selection, Schwarz (1978) established the Bayesian Information Criterion (BIC). The best fitted ARIMA model is chosen using selection criteria such as R-Square, Stationary R-Square, Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and Bayesian Information Criterion (BIC).

F. Diagnostic check of the Residuals and the Model Adequacy

Following the evaluation of model parameters, we conducted a thorough quality trial at this point. This estimate and trial of the studied model was conducted to determine the accessibility of its own conditions through the use of random mistakes (residuals). The difference between the value that occurs and the value that is estimated for a certain timeframe is defined by residuals. Various procedures were used to check the evaluated model and parameters. The following are the diagnostic tools that were used:

- Over fitting
- Residuals of Autocorrelation Function (ACF)
- Residuals of Partial Autocorrelation Function (PACF)
- Forecasting

Forecasting is the final phase of study and examination of time-series models, and is the principal target of the investigation, and to estimating the future qualities to watch the time series and is unimaginable to expect to duplicate up to this level unless the underlying model run over the analytic tests which were introduced before, in the event that the model can't finish these assessments effectively, at that point the arrival to the main stage would be fundamental and that is (Diagnostic), and read autocorrelation function and partial autocorrelation function precisely and pick a subsequent introductory model, consequently the procedure is rehashed until we get the model meets the states of the tests productively. If the model is correct the other forecast will not be found that gives smaller mean square errors, then: The forecasting is the condition expected in the period (T+L) at the time (t). The best forecasting is when the estimate from the resulting error is modest and the variation is low.

Expert modeler of IBM SPSS 22 Software was used to determine the best fit time series model for the present study.

III. RESULTS

Model Statistics													
Model	**Number of Predictors**	**Model Fit Statistics**								**Ljung-Box Q(18)**			**Number of Outliers**
		Stationary R-Squared	R-Squared	RMSE	MAPE	MAE	MaxPE	MaxAE	Normalized BIC	Statistics	DF	Sig.	
Confirmed Cases-Model_1	0	0.331	0.92	552.267	29.321	326.365	766.186	4459.001	12.686	39.368	15	0.001	0

TABLE 1: ARIMA MODEL DESCRIPTION

Spearman's correlation was used to study the relationship between COVID-19 cases and meteorological Variables. Humidity, rainy days and Average sunshine were found to be significant.

In order to select the model, ACF and PACF plots are plotted using IBM SPSS software. Figure 1 shows the ACF and PACF plot in order to select AR and MA order.

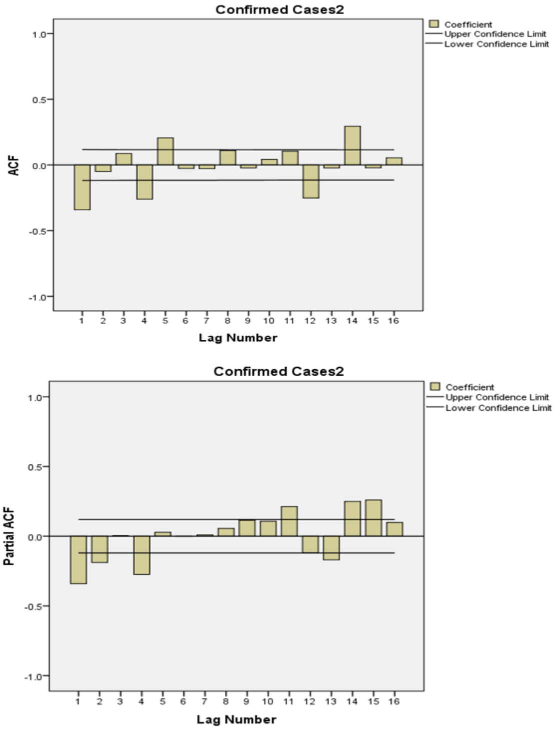


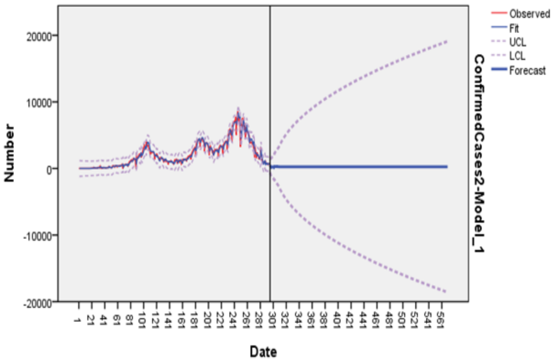
Fig 1: ACF and PACF Plots of Confirmed COVID-19 Cases of Delhi

Model Description			
			Model Type
Model ID	Confirmed Cases2	Model_1	ARIMA(0,1,14)

TABLE 2: MODEL STATISTICS OF THE SELECTED ARIMA MODEL

From the Table 1 and 2, we can observe that ARIMA (0,1,14) was found to be the best model for the present study data with R-squared value 0.92, NBIC=12.686 and significant p value 0.001 ($p<0.05$). Despite the fact that time series modeler provides a variety of goodness of fit statistics, we choose stationary R-squared since it provides an estimate of the overall variation in the series that the model clarifies. When there is a pattern or seasonal pattern, such as in the current data, it is preferable to R-squared. Estimation implied that the model could clarify 92 % of the observed variations in the series. The forecasted model proposed ARIMA (0, 1, 14) which served useful in estimating COVID-19 occurrences for the future from January 2021 through August 2021. The forecast shows a minute decreasing trend in the forecasted values, this indicates that given the independent factors considered for the study remains constant as of December 2020, further COVID-19 cases in Delhi may show a decreasing trend. (Figure 2)

Fig 2: Observed and Forecasted Values of COVID-19 Cases



IV. CONCLUSIONS

According to the findings of time series analysis, the suggested ARIMA (0,1,14) model can forecast with sufficient goodness of fit and precision. Among all other models, the model was chosen for its low normalized BIC value, MAPE, and good R-Square. The model provides for a better understanding of COVID-19 dynamics, as well as forecasts that can be applied in public health planning and preparedness in COVID-19 pandemic response.

ACKNOWLEDGMENTS

I would like to thank my mentor Ms Stavelin Abhinandithe K, Assistant Professor, Division of Medical Statistics, Faculty of Life Sciences, JSS Academy of Higher Education & Research, Mysuru, Karnataka, India for giving me the invaluable opportunity to conduct this research, and for her constant guidance and motivation. I also express my gratitude to Dr. K. A. Raveesha, Professor and Head of Department, Faculty of Life Sciences, JSS Academy of Higher Education & Research, Mysuru, Karnataka, India for his constant support.

REFERENCES

[1]. Gao GF. From “A”IV to “Z”IKV: attacks from emerging and re-emerging pathogens. Cell 2018; 172: 1157-9.

[2]. David J Cennimo, Scott J Bergman, Keith M Olsen. What is the global and US prevalence of coronavirus disease 2019(COVID-19)? Medscape 2020.

[3]. Tabrez Ahmad. Scenario of the corona virus (COVID-19) in India. Research gate. Available on:http://www.researchgate.net/publication/340443256_scenario_of_the_Corona_Virus_Covid-19_in_India

[4]. Rajath Guptha, Anu Madgavkar. Coronavirus’ impact on India | McKinsey available on: http://www.mckinsey.com/featured-insights/india/getting-ahead-of-coronavirus-saving-lives-and-livelihood-in-india

[5]. Alzahrani, S.I.; Aljamaan, I.A.; Al-Fakih, E.A. Forecasting the spread of the COVID-19 pandemic in Saudi Arabia using ARIMA prediction model under current public health interventions. J. Infect. Public Health 2020, 13, 914–919. [CrossRef]

[6]. Kufel, T. ARIMA-based forecasting of the dynamics of confirmed Covid-19 cases for selected European countries. Equilibrium. Q.J. Econ. Econ. Policy 2020, 15, 181–204. [CrossRef]

[7]. Liu, Z.; Magal, P.; Webb, G. Predicting the number of reported and unreported cases for the COVID-19 epidemics in China, South Korea, Italy, France, Germany and United Kingdom. J. Theor. Biol. 2020. [CrossRef] [PubMed].