Statistics and Applications {ISSN 2454-7395 (online)}
Volume 21, No. 2, 2023 (New Series), pp 99-106
http://www.ssca.org.in/journal

COVID-19 Cumulative Death Prediction in Two Most Populated Countries by Fitting ARIMA Model and Linear Regression

Shagun Sachdeva and Ravinder Singh
Department of Statistics, Central University of Haryana, Haryana, India

Received: 26 August 2022; Revised: 01 October 2022; Accepted: 08 November 2022

Abstract

COVID-19 (Coronavirus) has caused widespread disruption and hindered economic growth worldwide. Since the entire world has been hit by two dangerous waves of this epidemic, it has become increasingly vital for us to analyze COVID-19 casualties in order to forecast our future days. As a result, in this work, an attempt has been made to do a time series analysis and fit linear regression on the cumulative death of the two most populated countries, China and India. The research utilizes a simple yet powerful and objective approach, called autoregressive integrated moving average (ARIMA) to forecast the number of cumulative deaths. We have also fitted linear regression on the data to predict future values. The forecasted values have also been compared with the original cumulative death values. In conclusion, ARIMA model forecasted better results in comparison with regression model. As a result, ARIMA(0,2,1) and ARIMA(1,2,0) turns out to be the best model for China and India respectively. So, in the future, the government and health personnel can use these models to take desirable action to control the death count.

Key words: ARIMA model; Linear regression; Epidemic forecast; Cumulative deaths.

AMS Subject Classifications: 62J05, 62M10

1. Introduction

Coronavirus disease 2019 (COVID-19) is a serious, long-lasting contagious disease caused by the severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) that first arose in December 2019 in China. The pandemic has infected 511,749,262 people in 195 nations around the World, with 6,228,622 cases of death as of 28th April 2020. COVID-19's first case was found on 12th December 2019 which is now spreading exponentially and causing a large number of casualties. COVID-19 has been identified as the third largest cause of death in the World. More than half of the world's population lives in one of the 10 most populous countries in the world. China is the most populous country on earth with a population of around 1,397,715,000. The United States has the highest number of COVID-19

Corresponding Author: Ravinder Singh
Email: ravinderstats@cuh.ac.in

deaths (1,019,774), followed by India (523,693). Even though China is among the world's most populous countries, the number of fatalities by COVID-19 in China is lower. How a country faces such tough situations as COVID-19 gives an idea about the development of the country in terms of the medical sector, so it is important to study different countries and compare them.

To compare the most populated country in terms of COVID-19 deaths, we have used ARIMA model and simple linear regression. The regression and SEIR model were fitted on COVID-19 cases previously in (Panday *et al.*, 2020) and the SEIR model predicted closer results. Katoch and Sindhu (2021) also proposed a time series model based on genetic programming for the analysis of confirmed death and cases across the three most pretentious states of India - Maharashtra, Andhra Pradesh, Tamil Nadu, and Karnataka as well as for the whole India. Ding *et al.* (2020) studied the epidemic data from February 24 to March 30, 2020, and concluded that an inflection point was expected in early April in Italy. Bayyurt and Bayyurt (2020) compared the lag between COVID-19 cases and deaths with the help of the ARIMA model. Gambhir *et al.*(2020) applied regression on COVID-19 to study future patterns. An attempt was made by Hengjian and Tao (2020) to fit Non-linear regression on COVID-19 data. Batista (2020) studied the second phase of the coronavirus COVID-19 epidemic by the logistic model. It is vital to model and predict the deaths to deal with their consequences. Forecasting future COVID-19 deaths using statistical models is critical for breaking the transmission cycle in highly populated nations like China and India.

2. Mathematical background

We have used secondary data from World Health Organization (WHO) COVID-19 dashboard from 1st April 2021 to 30th June 2021 considering the number of cumulative deaths per day. The data includes confirmed cases and deaths along with their cumulative counts of all the countries. Excel was used in building the database of time series.

2.1. ARIMA model

ARIMA model in equations:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_n Y_{t-n} \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + \cdots + \phi_q \epsilon_{t-q} \tag{1}$$

ARIMA, short for 'Auto-Regressive Integrated Moving Average' explains the time series based on its previous values (Box *et al.*, 2015). In ARIMA *(p,d,q)* the *'p'* denotes the order of the 'Auto Regressive' (AR) term which refers to the number of lags of Y, which is used as predictors. The 'Moving Average' (MA) term's order is *'q'* showing the number of lagged forecast errors that can be included in the ARIMA model. The value of *'d'* is the smallest number of differencing required to get the series stationary. There are different ARIMA models that we can fit, to select the best one we can choose the criteria like AIC (Akaike Information Criteria), and log-likelihood. To check the accuracy of the model, we have calculated the change percentage of both ARIMA and linear regression and compared them. We have also plotted the Auto Correlation Function (ACF) and Partial Auto Correlation Function (PACF) of residuals in order to check the normality.

2.1.1. Analysis of data - I

An attempt has been made in this study to analyze the data collected with the COVID-19 cumulative deaths in China and India. In this case, the methodology described below was applied, and conclusions were drawn from the study. The study was of about 91 days covering second wave data starting from 1st April 2021 to 30th June 2021. During the second wave, India faced more death losses when compared to China. The data is plotted in time series shown in Figure 1 and 2, demonstrating stochastic trends.

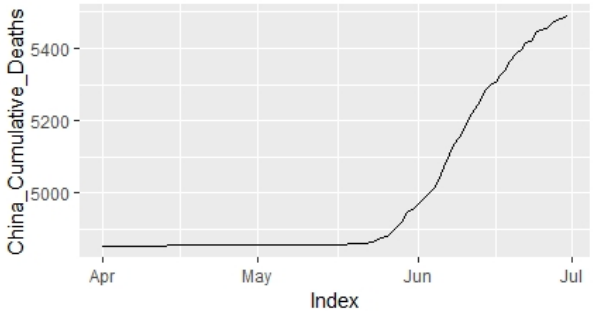

Figure 1: China cumulative death plot

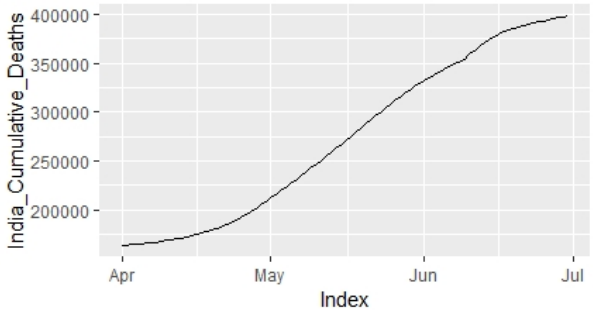

Figure 2: India cumulative death plot

Figure 1 and 2 depict that the COVID-19 peak started to reach India in April whereas for China it was after two months in June. With the use of software R, we plotted the Auto Correlation Function (ACF) and Partial Auto Correlation Function (PACF) to test the stationarity of China and India's time series graphically, as shown in Figure 3 to 6.

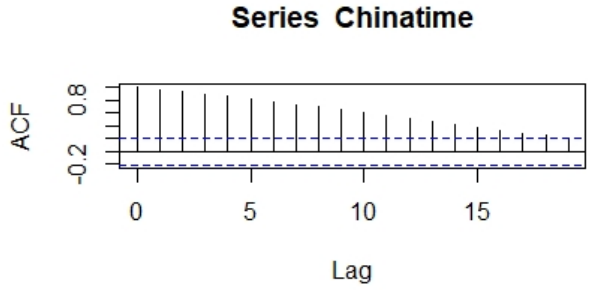

Figure 3: ACF China

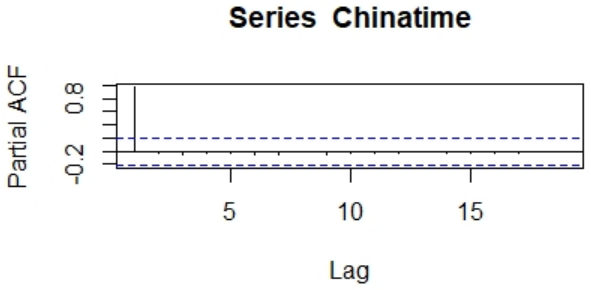

Figure 4: PACF China

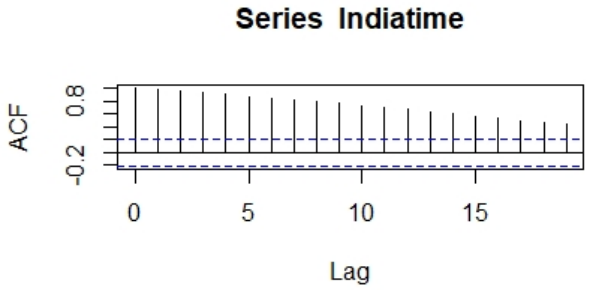

Figure 5: ACF India

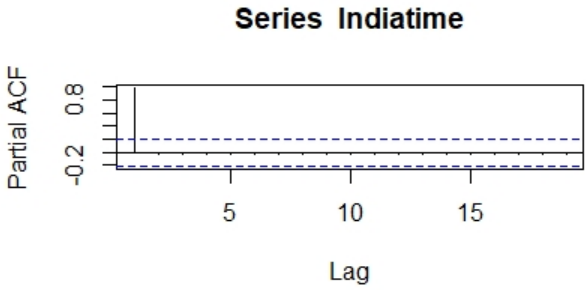

Figure 6: PACF India

Figure 3 and 5 depict the ACF of China and India respectively, since most of the bars in the ACF plot are above the upper limit so it has been concluded that data depends on the past value and thus it is not stationary. In the PACF plot, (Figure 4 and 6) one bar is not within the limit which also suggests the non-stationarity of data.

To check statistically whether data is already stationary or not, we have also applied Augmented Dickey-Fuller Test (ADF) and the results are demonstrated below in Table 1:

Table 1: Augmented Dickey- Fuller test results

Country	**Hypothesis**	**Test statistic**	***p*-value**
China	H_0 : The data is not stationary H_1 : The data is stationary	-1.9889	0.5808
India	H_0 : The data is not stationary H_1 : The data is stationary	-1.9825	0.5834

Table 2: ARIMA models

Country	**Model**	**ARIMA order**	**AIC**
China	1.1	ARIMA (2,2,2)	530.772
	1.2	ARIMA (0,2,0)	538.1234
	1.3	ARIMA (1,2,0)	528.0987
	1.4	**ARIMA (0,2,1)**	**525.1874**
	1.5	ARIMA (1,2,1)	526.9417
	1.6	ARIMA (0,2,2)	526.9914
	1.7	ARIMA (1,2,2)	528.8983
India	**2.1**	**ARIMA (1,2,0)**	**1374.284**
	2.2	ARIMA (2,2,1)	1377.411
	2.3	ARIMA (1,2,1)	1375.423
	2.4	ARIMA (0,2,1)	1375.489
	2.5	ARIMA (2,2,0)	1375.602
	2.6	ARIMA (0,2,0)	1394.637
	2.7	ARIMA (2,2,2)	1379.341

By checking the p-values we can conclude whether the data is stationary or not. The null hypothesis can be rejected if the p-value is less than 0.05; else, the null hypothesis will stand. In both China and India p-value is greater than 0.05 so we can't reject the null hypothesis and conclude that the time series of both countries is not stationary. Now we have to make them stationary for that we can take differences or log them. Here we are using R Studio which suggests various models and we can select the best ARIMA Model according to AIC (Akaike Information Criteria).

The model that has the minimum AIC (mentioned in bold) will be the best fit for our data. For India ARIMA model (1,2,0) and the China ARIMA model (0,2,1) satisfies this criterion as shown in Table 2. So we have selected these models and use them to forecast future cumulative death counts. In both China and India, the value of $d = 2$ means that the data is differenced two times to make it stationary. To ensure that the ARIMA model's residuals are normal, we have plotted the ACF and PACF of the Model's residuals, shown from Figure 7 to 10. The bars were coming within the limit which concludes that residuals follow Normal distribution.

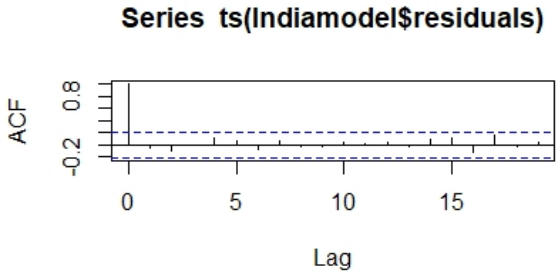

Figure 7: ACF of India residuals

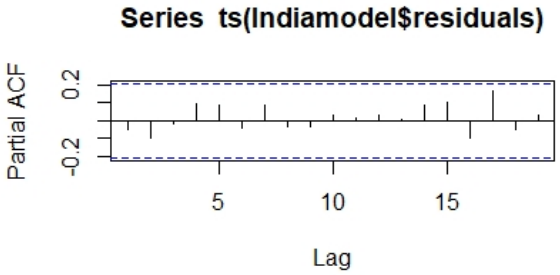

Figure 8: PACF of India residuals

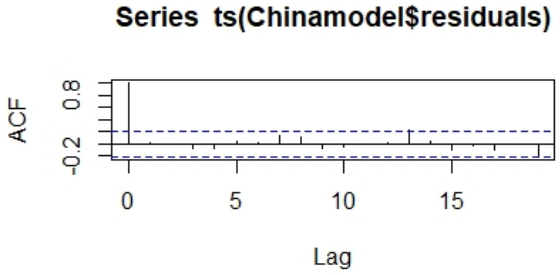

Figure 9: ACF of China residuals

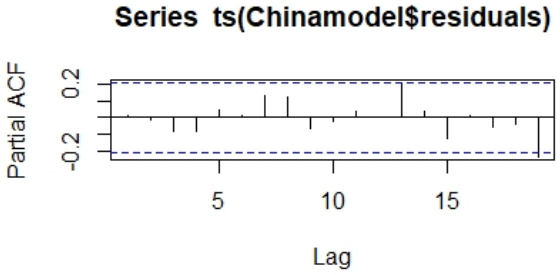

Figure 10: PACF of China residuals

2.2. Regression model

Simple linear regression is a method for predicting a response Y by using only one predictor variable X. The regression model estimates independent variables by using continuous dependent variables provided in the list. The model can be written as :

$$Y = \alpha + \beta\ X + \epsilon \tag{2}$$

where X is the independent variable that is used to estimate the value of Y by following the above relation. In our study cumulative deaths is the dependent variable whose values are being estimated by taking days as X variable or independent variable. In the above equation, α and β are the intercept term and slope parameter respectively, which are also known as regression coefficients. The unobservable error component ϵ indicates the gap between the true and observed values of Y and accounts for the failure of data to lie on a straight line. For statistical inferences we assume ϵ to be an independent and identically distributed random variable with mean zero and constant variance σ^2. We have used R Studio to fit the simple regression model for both China and India.

2.2.1. Analysis of data - II

COVID-19 is spreading exponentially in the world and many people are losing their close ones. The number of deaths due to COVID-19 was increasing in counts as the days were passing especially during the second wave of COVID-19. This study includes the COVID-19 second wave data set from WHO. The collected data was analyzed in R Studio. In India the number of deaths has been increasing day by day while in China the loss of deaths due to COVID-19 has been almost constant, this fact matches our calculated result.

3. Results and conclusion

In ARIMA time series analysis ARIMA (0,2,1) and ARIMA(1,2,0) turn out to be the best model for China and India respectively. The predicted values for the models are shown in Figure 11 and 12. Referring to Table 3 and 4, it has been concluded that ARIMA is the better model for forecasting future cumulative death as the change percentage for the ARIMA model is only 0.01 to 0.2% and 0.01 to 0.3 % for China and India respectively but on the other hand we can see a higher percentage difference for regression model of about 3.4 to 3.7 % for China and 4.0 to 9.9% for India. It is also observed that the second wave of COVID-19 has turned out to be less severe for China as compared to India. The government and associated departments can use ARIMA model to forecast COVID-19 deaths rather than regression model. The comprehensive study designed on the cumulative deaths of COVID-19 can help to see the severity of the situation by predicting the mortality rate. This will help the policymakers to take preventive measures and actions such as fulfilling the oxygen and vaccination demand, arranging beds and medical experts for controlling the COVID-19 situation. This research will also help policymakers to keep a track of how different decisions such as quarantine, lockdown, vaccination *etc.* are helpful in reducing the death count. This can be done by monitoring the difference between predicted deaths and actual deaths taking place after the implementation of policies.

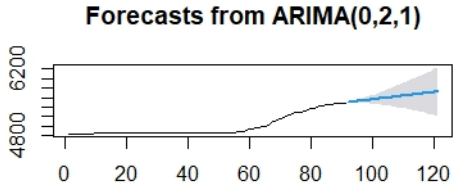

Figure 11: China forecast plot

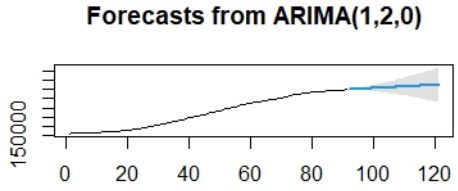

Figure 12: India forecast plot

Table 3: Forecasted and actual value comparison of China

Date	ARIMA actual data	ARIMA forecasted data	ARIMA percentage change	Regression forecasted data	Regression percentage change
01-07-2021	5495	5497.672	0.048626024	5305.5117	-3.448376706
02-07-2021	5508	5505.344	-0.04822077	5312.3607	-3.551911765
03-07-2021	5523	5513.016	-0.18077132	5319.2097	-3.689847909
04-07-2021	5533	5520.688	-0.222519429	5326.0587	-3.740128321
05-07-2021	5535	5528.36	-0.119963866	5332.9077	-3.651170732
06-07-2021	5537	5536.032	-0.017482391	5339.7567	-3.562277407
07-07-2021	5554	5543.704	-0.185379906	5346.6057	-3.73414296
08-07-2021	5563	5551.376	-0.208952004	5353.4547	-3.766767931
09-07-2021	5566	5559.048	-0.124901186	5360.3037	-3.695585699
10-07-2021	5578	5566.72	-0.202223019	5367.1527	-3.77998028
11-07-2021	5584	5574.392	-0.172063037	5374.0017	-3.760714542
12-07-2021	5588	5582.064	-0.106227631	5380.8507	-3.707038296
13-07-2021	5589	5589.736	0.013168724	5387.6997	-3.601723027
14-07-2021	5595	5597.407	0.043020554	5394.5487	-3.582686327
15-07-2021	5601	5605.079	0.072826281	5401.3977	-3.563690412

Table 4: Forecasted and actual value comparison of India

Date	ARIMA actual data	ARIMA forecasted data	ARIMA percentage change	Regression forecasted data	Regression percentage change
01-07-2021	399459	399313.2	-0.036499365	413929.61	4.01359566
03-07-2021	401050	401001.3	-0.012143124	420171.85	4.550959328
04-07-2021	402005	401845.6	-0.039651248	423292.97	5.029133841
05-07-2021	402728	402692	-0.008939036	426414.09	5.554715605
06-07-2021	403281	403537.4	0.063578497	429535.21	6.112236992
07-07-2021	404211	404383.3	0.042626252	432656.33	6.574578488
08-07-2021	405028	405229	0.049626199	435777.45	7.056227898
09-07-2021	405939	406074.7	0.033428668	438898.57	7.509609794
10-07-2021	407145	406920.4	-0.055164622	442019.69	7.889849884
11-07-2021	408040	407766.2	-0.067101265	445140.81	8.334623375
12-07-2021	408764	408611.9	-0.037209735	448261.93	8.811350542
13-07-2021	410784	409457.6	-0.322894757	451383.05	8.994367423
14-07-2021	411408	410303.4	-0.268492591	454504.17	9.482018614
15-07-2021	411989	411149.1	-0.203864666	457625.29	9.972414331

References

Batista, M. (2020). Estimation of the final size of the second phase of the coronavirus COVID-19 epidemic by the logistic model. *medrxiv.* DOI:10.1101/2020.02.16.20023606

Bayyurt, L. and Bayyurt, B. (2020). Forecasting of COVID-19 cases and deaths using ARIMA models. *medrxiv.* DOI:10.1101/2020.04.17.20069237

Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015). *Time Series Analysis: Forecasting and Control*. John Wiley Sons.

Ding, G., Li, X., Jiao, F., and Shen, Y. (2020). Brief analysis of the ARIMA model on the COVID-19 in Italy. *medRxiv.* DOI:10.1101/2020.04.08.20058636

Gambhir, E., Jain, R., Gupta, A., and Tomer, U. (2020, September). Regression analysis of COVID-19 using machine learning algorithms. *International Conference on Smart Electronics and Communication (ICOSEC),* 65-71. DOI:10.1109/ICOSEC49089.2020.9215356

Hengjian, C. and Tao, H. (2021). Nonlinear regression in COVID-19 forecasting. *Scientia Sinica Mathematica*, **51**. DOI:10.1360/SSM-2020-0055

https://covid19.who.int/data

Katoch, R. and Sidhu, A. (2021). An application of ARIMA model to forecast the dynamics of COVID-19 epidemic in India. *Global Business Review.* DOI:10.1177/0972150920988653

Pandey, G., Chaudhary, P., Gupta, R., and Pal, S. (2020). SEIR and regression model based COVID-19 outbreak predictions in India. *arXiv preprint arXiv:2004.00958.* DOI:10.48550/arXiv.2004.00958