Nonlinear Time Series Models to Analysis and Predicting COVID-19 Cases in Holy Kerbala- Iraq

JassimNasirHussain Mohammed Saleh Hashem

jasim.nasir@uokerbala.edu.iqmohammed.saleh@s.uokerbala.edu.iq

Abstract: The emerging coronavirus (Covid-19) represents the last detected strain of the coronavirus. This virus appeared for the first time in the Chinese city of Wuhan in December / 2019. While in Iraq, it appeared for the first time on February 24, 2020. The Covid-19 virus is a global health problem that includes all the parts of the world without any exception, especially in Iraq and the governorate of holy Kerbala. The number of infected people from the beginning of the pandemic until the date of (5-8-2020) reached about (1108558) injuries and about (15741) deaths. This represents a very high rate, and it certainly represents a major economic, social, and health problem.

This instigated our interest in researching and studying this phenomenon, as well as predicting the number of infected people in the future. So, this study aims to predict the numbers of people infected with the Covid-19 virus in the holy city of Kerbala via utilizing the nonlinear time series models. In addition to choosing the best prediction model by utilizing some statistical criteria such as (AIC, BIC, H-Q).

Data for the numbers of people infected by the Covid-19 virus in the governorate of holy Kerbala were obtained from the official website of the Iraqi Ministry of Health for the period from (1-6-2020) to (30-9-2020).

Three nonlinear models have been utilized in order to predict this series (Exponential model, Logistic model, Gompertz model). Furthermore, the statistical criteria were utilized in order to compare these models and choose the best model that represents these data.

The results showed that the logistic model is the best model representing COVID-19 data, which gives the lowest values for all three criteria. Then the Gompertz model and the exponential model are coming after it.

1. Introduction

Several acute pneumonia cases were first discovered in China in December 2019, which were confirmed as acute and infectious acute respiratory diseases that were caused via the new Covid-19 virus. The infection of this virus spread widely and rapidly in China, especially in the popular markets for seafood through the early infected persons. Also, the infection spread to other regions in China and many other countries until it reached Iraq for the first time on (24/2/2020) and spread to many Iraqi governorates. The governorate of holy Kerbala recorded the first injuries on (3/3/2020). The number of injured people in the governorate of holy Kerbala, from the beginning of the pandemic until the date of (8/5/2020), reached about (1108558) injuries. This pandemic has caused great damage to human society. Furthermore, it is one of the most important factors that threaten human life and health greatly. In addition, this pandemic limits social and economic development and also imperils national security and stability. The improvement of the service and the health reality that provides a better life for humans was the main motive for the beginning of research related to the prediction of many dangerous infectious diseases and viruses. Our study aimed toanalyze and predict the Covid-19 virus cases at one of the Iraqi governorates. On the other hand, the topic of time series analysis represents as one of the important statistical topics which deal with the behavior of the phenomenon and its interpretation during a specific time period. The prediction has received great interest in many scientific fields. Because of the critical importance of forecasting, there are many methods have emerged in order to predict the future behavior of phenomena.

Many researchers have been tried to model and furcate the COVID-19 pandemic from the beginning of it in 2020. One of them Jia et al. (2020)[5], utilized three nonlinear models (Gompertz model, Logistic model, and Von Bertalanffy model). The epidemiological trend of the Covid-19 virus was first analyzed, as well as predictions utilizing these models. The results showed that the prediction is distinct according to the parameters and the different regions. Also, the results showed that the logistic model is the best model among the models that have been studied. In the same year, (Medina et al.)[7] presented their research which aims to test the validity of the Gompertz model and the Logistic model for predicting confirmed cases and deaths from the Covid-19 virus in Cuba. The results showed that both models have a good fit.

The rest of this paper consists of five sections. Section 2 consists of the methodology, section 3 consists of the nonlinear time series models, section 4 consists of the Result and discussion,section 5 consists of theForecastingand section 6 consists of theConclusions

2. Methodology

Time series models are important in the field of analysis and forecasting. There are two types of these models: linear and nonlinear. The characteristics of the time series of the numbers of people who are infected with the Covid-19 virus in the Holy Kerbala Governorate from (1-6-2020) until (30-9-2020) will be studiedtodiscover the model which fit these data and estimating the parameters of the models which are utilized in this study. The first step in analyzing time series is to determine the characteristics of the time series through some of the following tests:

2.1 Augmented Dicky-Fuller test:

This test is used for the purpose of finding out whether that the time series is stationary or not. It can be expressed via the following equation:

$$\Delta Y_t = b_0 + b_1 T + \delta Y_{t-1} + a_i \sum_{i=1}^{n} \Delta Y_{t-i} + e_t$$

2.2 Mann-Kendall test:

This test is utilized in order to find out whether that the time series is linear or not. It can be expressed via the following equation:[2]

$$Z = \begin{cases} \dfrac{S-1}{\sqrt{\mathrm{var}\ (S)}} & \text{if } S > 0 \\ 0 & \text{if } S = 0 \\ \dfrac{S+1}{\sqrt{\mathrm{var}(S)}} & \text{if } S < 0 \end{cases}$$

2.3 Standard normal homogeneity test:

This test is utilized for the purpose of knowing whether that the data is homogeneous or not. It can be expressed via the following equation:[8]

$$T(u) = u\bar{z}_1 + (n-u)\bar{z}_2$$

2.4 The model parameters significance test:

The t-test is utilized to test the significance of the parameters of the estimated model via dividing the value of the estimated parameter by the standard deviation as follows:

$$tb = \frac{\hat{b}}{s\hat{b}} \quad . \quad tb_0 = \frac{\hat{b}_0}{s\hat{b}_0} \quad . \quad tb_1 = \frac{\hat{b}_1}{s\hat{b}_1} \quad \ldots\ldots\ldots\ldots. \frac{\hat{b}_j}{s\hat{b}_j}$$

2.5 The overall fitting test for the model:

The F test generally is utilized in order to test the morale of the model. The formula of this test is written as follows:

$$\mathrm{F} = \frac{\mathrm{RSS/k-1}}{\mathrm{SSE/n-k}} = \frac{\sum \mathrm{y_i^2/k-1}}{\sum \mathrm{e_i^2/n-1}}$$

2.6 Sequential residual test:

The Jarque-Bera test is utilized in order to determine whether that the residuals are of a normal distribution or not. This test is expressed via the following equation:

$$\mathrm{JB} = \frac{n}{6}B_1 + \frac{T}{24}(B_2 - 3)^2 \sim x^2_{\alpha(2)}$$

3. The nonlinear time series models

There are many nonlinear modelsto represent the nonlinear time series data. Some of them are chosen to model the time series of COVID-19 data, such as in the following subsections.

3.1 Exponential Model

This model is one of the most important nonlinear models in the time series. It is a mathematical model or expression which describes the process of increasing (such as the increase in the number of infections with the Covid-19 virus). Also, it is called the natural growth model. The exponential growth rate is an important measure of the severity of the epidemic. In general, this model contains only two parameters which namely (a and b). This model is expressed via the following equation:[6]

$$Y_t = \alpha e^{-bt} + e_t$$

$\mathrm{Y_t}$:is representing the number of infected individuals over the timet.

$(\alpha.\mathrm{b})$: represent the parameters of the model.

e_t: represents the random error.

3.2 Logistic Model

This model is one of the nonlinear models in the time series. It consists of three parameters (α, b, and k). The logistic model is mainly utilized in epidemiology. Moreover, this model was utilized first by the Belgian scientist Verhulst in 1845. The logistical equation is called carrying capacity. It is the maximum size to which the population can grow. At this stage, the population size stabilizes, and the growth rate increases. This model is expressed via the following equation:[1]

$$\mathrm{Y_t} = \alpha \mathrm{e}^{-\mathrm{be}^{-\mathrm{kt}}} + \mathrm{e_t}$$

$\mathrm{Y_t}$: represents the number of infected individuals over timet.

$(\alpha.\mathrm{b.k})$:represent the parameters of the model.

$\mathrm{e_t}$: represents the random error.

3.3 Gompertz Model

This model is another model of the nonlinear growth models in the time series. This model was formulated via BenjaminGompertz in 1825 to fit the death-rate tables. It has been utilized as the growth model, especially in epidemiological, medical, and biological studies. In the economic aspect, it is utilized to predict the development of the market and securities. It is also widely used in biology. In addition, it is utilized to describe the growth of animals and plants as well as the numbers and sizes of bacterial and cancer cells.

Moreover, this model is used to describe the law of the spread of infectious diseases and study the factors which control the spread of the Covid-19 virus.In general, the Gempertz model contains three parameters (α, b, and k). The model can be expressed via the following equation:[3]

$$Y_t = \alpha e^{-be^{-kt}} + e_t$$

Y_t: represents the number of the infected individuals over timet.

$(\alpha. b. k)$:represent the parameters of a model.

e_t: represents the random error.

4. Result and discussion

After studying the characteristics of the time series of numbers of people infected with the Covid-19 virus, the results showed that the series is unstable, nonlinear, and heterogeneous. Thus, the models that fit these data are the nonlinear models. In this research, three nonlinear models will be used, which are (the exponential model, the logistic model, and Gompertz model). The maximum likelihood method was used to estimate the parameters of these models. Since that these models are nonlinear, and it is difficult to estimate their parameters in the usual way. Therefore, one of the iterative methods, which is the Newton-Raphson iterative method, will be used to estimate the parameters of these models.

The results of estimating the exponential model show that the estimated model is as follows:

$$\widehat{Y}_t = 109.39e^{0.0058t}$$

Also, these results show that the P-value of the two estimated parameters (α.b) is (1.30E-14, 4.23E-5). Therefore, it is less than the significance level (0.05). Then, the two estimated parameters (α.b) were statistically significant.

Furthermore, these results show that the calculated value of the F-test of the exponential model was (235.6653), which is greater than the tabular value of the F-test table (3.69),which,is meaning that the model is statistically significant.

The results also showed that the P-value of the Jarque-Bera statistic of the exponential model is (0.1333) is greater than the 0.05 level of the significance, which is meaning that the residuals are subject to the normal distribution. The following figure (1) shows the expected values of the exponential model.

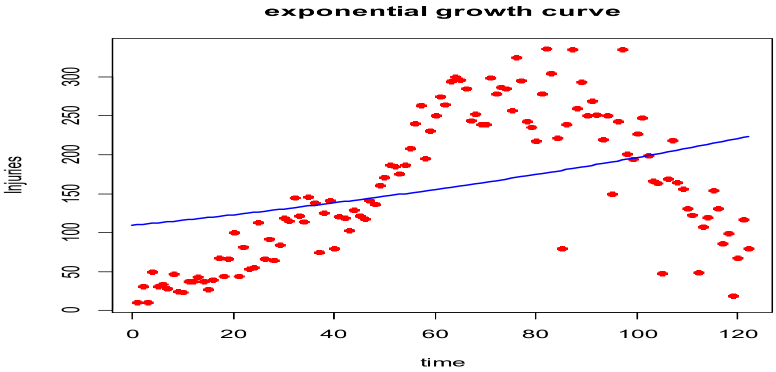

Figure (1): Represents the Expected Values of

Figure (1) shows the distance of the estimated values from the real values, which is meaning that the exponential model does not represent the data very well.

After estimating the parameters of the Gompertz model, then the model becomes as follows:

$$\widehat{Y}_t = 210.49e^{-6.11e^{-0.075t}}$$

Also, these results showed that the P-value of the estimated parameters (α.b.k) is (2.E-16, 4.78E-02, 9.1E-05). Therefore, it is less than the level of significance (0.05). So, the estimated parameters (α.b.k) are statistically significant.

Furthermore, these results show that the calculated value of the F-test for the Gompertz model is (295.456), which is greater than the tabular value of the F-test table (3.69), which is meaning that the model is statistically significant.

The results showed that the P-value of the Jarque-Bera statistic for the Gompertz model (0.0224) which is less than the 0.05 level of significance, which is meaning that the residuals are not subject to the normal distribution. The following figure represents the expected values of the Gompertz model.

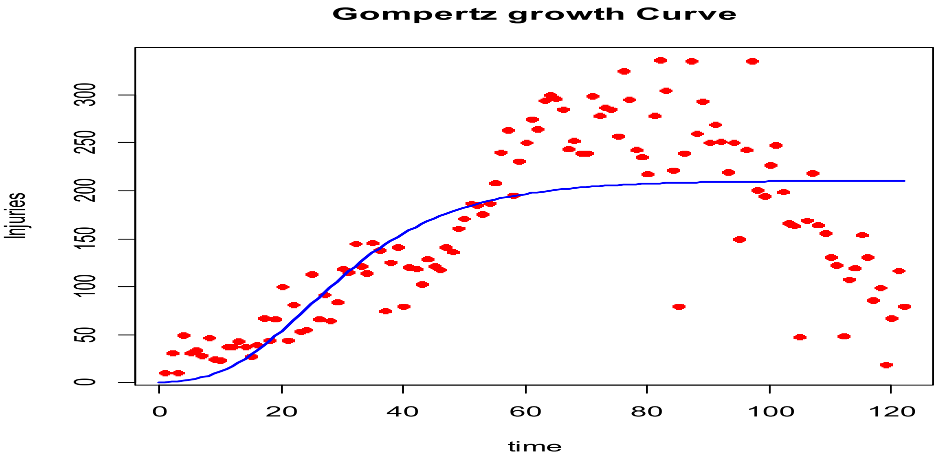

Figure (2): Represents theExpected Values of the Gompertz Model

Figure (2) shows that the model gives estimated values that are closer to the real data than the exponential model, but this model is still far from the good representation of the data.

After estimating the parameters of the logistic model, the model becomes as follows:

$$\hat{Y}_t = \frac{211.32}{1 + 22.07e^{-0.992t}}$$

It has been noticed that the P-Value of the estimated parameters (α.b.k) is (2.E-16, 1.62E-01, 2.89E-05). Therefore, it is less than the level of significance (0.05). So, the null hypothesis H_0 is rejected, and the estimated parameters (α.b.k) are statistically significant.

Through the results obtained, the arithmetic value of the F-test of the logistic model is (308,009), which is greater than the tabular value of the F-test table, which is (3.96). Therefore, the null hypothesis is rejected, which meaning that the model is statistically significant.

Also, the results that have been obtained, it was shown that the P-value of the Jarque-Bera statistic of the logistic model is (0.007), which is less than the level of significance of 0.05.

So, the null hypothesis H_0 is rejected, which meaning that the residuals are not subject to the normal distribution. The following figure (3) represents the estimated values of the logistic model.

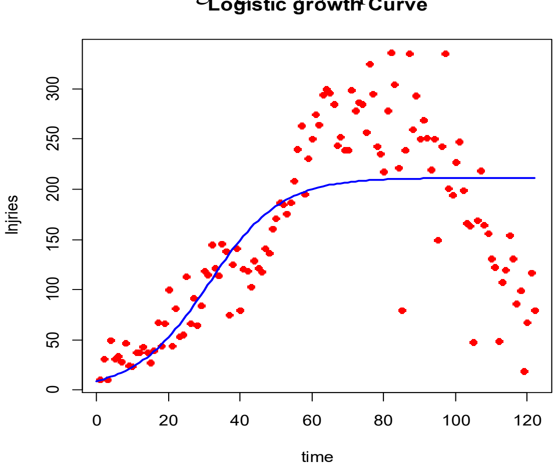

Figure (3): Represents the Expected Values of the Logistic Model

Figure (3) shows that the model gives estimated values that are closer to the real data more thanthe exponential model and the Gompertz model. Thus, it is the best model that represents the data well.

In addition, through the results that are shown in Table (1) Where three statistical criteria were utilized which namely (AIC, BIC, H-Q). As these criteria agreed that the logistic model is the best one that represents the data on the numbers of infections with the Covid-19 virus, which gives the least value for these criteria. Then, the Gompertz model and the exponential model come next.Table (1) represents the statistical criteria for comparing models

	Exponential model	Gompertz model	Logistic model
AIC	1081.152	1017.026	1012.542
BIC	1068.76	1025.438	1020.945
H-Q	1083.43	1020.443	1015.959

Table (1): The Statistical Criteria for Comparing Models

5. Forecasting:

The best model, which is the logistic model utilized in order to predict the numbers of people infected with the COVID-19 virus in the holy city of Karbala. The following Table (2) represents the predictive values of the logistic model.

21/9/2020	**211.3**	**26/9/2020**	**211.31**
22/9/2020	**211.3**	**27/9/2020**	**211.31**
23/9/2020	**211.3**	**28/9/2020**	**211.31**
24/9/2020	**211.3**	**29/9/2020**	**211.31**
25/9/2020	**211.31**	**30/9/2020**	**211.31**

Table (2): Represents the predictiveValues of the Logistic Model

From the results shown in Table (2), it has been noted that there is a convergence between the real and the predictive values of the logistic model. This indicates the efficiency of the proposed model for forecasting.

6. Conclusions:

- The statistical criteria (AIC, BIC, H-Q) were utilized to choose the best model among the nonlinear models, which are the exponential model, Gompertz model, an exponential model. The results showed that the logistic model is the best model that fits the data of COVID-19.
- The logistic model was utilized to predict the number of people infected with the COVID-19 virus. The results showed that there is a convergence between the real values and the predicted values. Then, this indicates the quality of the model utilized for the forecasting.

7.References:

1. A.J.Clark, L.W.Lake, T.W.Patzek. "Production Forecasting with Logistic Growth Models." *Society of Petroleum Engineers* 30 10 2011.

2. Ammar Salman Dawood, Ammar Ashour Akesh, Ahmed Sagban Khudier. "Study of Surface Water Quality and Trends Assessment at Shatt Al-Arab River in Basrah Province." *journal University of Kerbala* 16 11 2018.

3. Jaheen, Zeinhum F. "Prediction of Progressive Censored Data from the Gompertz Model ." *George Mason University* 23 2 2013.

4. Juan Felipe Medina-Mendieta MS, Manual Cortes-cortes PhD, Manual Cortes-lgiesias Ms. "Covid-19 Forecasts for Cuba Using Logistic Regression and Gompertz Curves." *MEDICC review* 29 4 2020.

Lin Jia, Kewen Li, Yu Jiang, Xin Guo, Ting Zhao. "Prediction and analysis of Coronavirus Disease 2019." *arXiv Preprint arXiv* 15 1 2020.

5. Ma, Junling. "Estimating epidemic exponential growth rate and basic reproduction number ." *Infectious Disease Modelling* 1 8 2020.
6. Medina. (2020, 4 29). Covid-19 Forecasting for Cuba Using Logistic Regression and Gompertz Curve. MEDICC review.
7. Omar M. A. Mahmood Agha, S Cagatay Bagcac, Nermin Sarlak. "Homogenneity Analysis of Precipitation Series in North Iraq." *Jornal of Applied Geology and Geophysics* 6 2017.
8. Yogesh Hole et al 2019 J. Phys.: Conf. Ser. 1362 012121