Research article

On computational analysis of nonlinear regression models addressing heteroscedasticity and autocorrelation issues: An application to COVID-19 data

Mintodê Nicodême Atchadé a,b,*, Paul Tchanati P. a

a National Higher School of Mathematics Genius and Modelization, National University of Sciences, Technologies, Engineering and Mathematics, Abomey, Benin
b University of Abomey-Calavi/International Chair in Mathematical Physics and Applications (ICMPA: UNESCO-Chair), 072 BP 50 Cotonou, Benin

ARTICLE INFO

Keywords:
COVID-19
Prediction
Nonlinear least squares
Statistical modeling
Nonlinear regression

ABSTRACT

This paper develops a method for nonlinear regression models estimation that is robust to heteroscedasticity and autocorrelation of errors. Using nonlinear least squares estimation, four popular growth models (Exponential, Gompertz, Verhulst, and Weibull) were computed. Some assumptions on the errors of these models (independence, normality, and homoscedasticity) being violated, the estimates are improved by modeling the residuals using the ETS method. For an application purpose, this approach has been used to predict the daily cumulative number of novel coronavirus (COVID-19) cases in Africa for the study period, from March 13, 2020, to June 26, 2021. The comparison of the proposed model to the competitors was done using statistical metrics such as MAPE, MAE, RMSE, AIC, BIC, and AICC. The findings revealed that the modified Gompertz model is the most accurate in forecasting the total number of COVID-19 cases in Africa. Moreover, the developed approach will be useful for researchers and policymakers for predicting purpose and for better decision making in different fields of its applications.

1. Introduction

COVID-19 is a pandemic of infectious disease called Sars-CoV-2 coronavirus disease [1]. It appeared on November 16, 2019 in Wuhan in Hubei province (central China), before spreading around the world and declared as a pandemic on March 11, 2020. The spread of this disease is staggering. As of October 15, 2021, the world noticed 238,437,517 confirmed COVID-19 cases and 4,879,235 deaths [1]. This pandemic affects all socio-economic aspects of all countries of the world.

The socio-economic damage of this pandemic immediately motivated the use of mathematical models to understand the evolution of the epidemic and to plan effective response strategies accordingly [2, 3, 4, 5, 6, 7]. In this context, it is crucial to be able to predict how COVID-19 will develop to assist health structures in properly directing efforts. Several works have already been carried out with the aim of predicting the evolution of COVID-19 in the world. The initial research was conducted in Wuhan, China, where the epidemic originally emerged. The first modeling approaches were essentially built around compartmental models [8, 9, 10, 11, 12]. The fundamental problem when modeling an epidemic using compartmental models is the calculation of the RO (the average number of secondary infections caused by an infectious individual introduced into a fully susceptible population). Moreover, the hypothesis according to which the study population is homogeneous is not always verified in real life. To circumvent these difficulties, several models have been developed using artificial intelligence and linear regression [9, 13, 14, 15, 16]. In [17], the authors used deep learning methods to predict the number of reported novel coronavirus positive cases in 32 states and union territories of India. They noticed that their models were more accurate for short-duration forecasts. [18] used an exponential model and regression analysis to predict positive cases of COVID-19 in India. Besides, artificial intelligence is a powerful tool for forecasting. The difficulty of this approach is that there is no way to know which model is suitable to the data. Thus, several authors have worked on different methods to identify the one that best models the evolution of the COVID-19 epidemic. This study [19] conducted a comparative study of ARIMA models, exponential smoothing techniques, and Poisson counting models to find the best model for COVID-19 data from Chile. The results obtained suggest that the ARIMA model is the...

* Corresponding author at: National Higher School of Mathematics Genius and Modelization, National University of Sciences, Technologies, Engineering and Mathematics, Abomey, Benin.
E-mail address: nickpowerabc@gmail.com (M.N. Atchadé).

https://doi.org/10.1016/j.heliyon.2022.e11057
Received 5 April 2022; Received in revised form 29 July 2022; Accepted 7 October 2022

2405-8440/© 2022 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
most suitable for predicting the number of confirmed cases of COVID-19 while for the number of deaths the smoothing techniques seem more suitable. In [20], four models (Holt model, ARIMA, TBATS, and cubic spline smoothing) were investigated on COVID-19 data from USA and Italy. They concluded that all models provided satisfactory results, with ARIMA and cubic spline smoothing models giving better results.

Some authors have turned to nonlinear regression, in particular growth models, to investigate the evolution of COVID-19. [21] conducted a comparative study of three growth models (Gompertz, Richard, and Weibull) for the evolution of COVID-19 in Iraq. Weibull distribution was found to be the best for modeling the cumulative number of COVID-19 cases. In this paper [22], authors used the nonlinear Gompertz, Verhulst, and SIR models and proposed a combination of Gompertz and Verhulst models to predict active COVID-19 cases. The proposed model gives better results compared to other models. [23] studied the epidemic trends of COVID-19 with a logistic model and machine learning techniques. According to their mathematical estimates, the global epidemic will peak at the end of October 2020, with approximately 14.12 million people infected cumulatively. Growth models are used as modeling tools in many fields, including forestry, zoology, finance, epidemiology [24, 25, 26, 27]. The problem with using growth models is that we often encounter violations of nonlinear regression assumptions. Several approaches have already been proposed to solve the problem of autocorrelation and heteroscedasticity of errors in nonlinear regression. A method has been proposed for correcting the problem of autocorrelation of errors in nonlinear regression by assuming that these errors can be modeled by an Auto Regressive Moving Average (ARMA) model in [28]. To deal with heteroscedastic errors, researchers have proposed a variance function that specifies the conditional variance of the response given the covariates [29]. Moreover, [30] combined transformation and weighting into a general form to overcome heteroscedasticity issue in nonlinear regression.

This paper aims to use the ETS modeling approach in combination with the standard formulation of nonlinear regression to propose a method that solves the problem of autocorrelation and heteroscedasticity of errors. The rest of the paper is structured as follows: In Section 2, the source of the data and the used growth models as well as the ETS modeling approach are presented. In addition, in this section we provide the formulation of the proposed model. Section 3 presents the main results followed by the discussion and limitations of this work in Section 4. Section 5 summarizes the main contributions of this paper.

2. Methodology

2.1. Data

The data used is the total cumulative number of confirmed COVID-19 cases in Africa. We collected them from https://ourworldindata.org/explorers/coronavirus-data-explorer [31]. We have downloaded the full COVID-19 data available in csv version. We then filtered the column “location” using the value “Africa” to obtain the data for Africa. Thus, we considered the available daily total cases for the analysis. The data covers the period from March 13, 2020 to June 26, 2021 (95%) to train the model, and from June 27 to July 25, 2021 (5%) to validate the model.

2.2. Growth modeling

The mathematical formulation of growth models is mainly based on differential equations. Thus, we are interested in the variation of the physical variable that we are trying to model with respect to time. Depending on the nature of the differential equation used, there are several growth models. Suppose \( f \) is a function that traduces the evolution of the study’s physical variable. In the sequence, \( \frac{df}{dt} \) denotes the first derivative of \( f \) with respect to the time variable \( t \). In this work, we investigated COVID-19 spread with four of the most widely used growth models.

2.2.1. Exponential model

This model usually characterizes a quantity which increases more rapidly than at the beginning of its distribution and refers to the fact that the involved function is proportional to its growth rate. The model bears the name Thomas Malthus, who in 1798 published one of the oldest and most important works on population theory An Essay on the Principle of Population [32], one of the earliest and most influential books on population. The exponential model is characterized by the following differential equation (eq. (1)):

\[
\frac{df}{dt} = Kf, \quad K \in \mathbb{R}
\]  
(1)

and the solution is given by eq. (2)

\[
f(t) = \theta_1 e^{\theta_2 t}, \quad (\theta_1, \theta_2) \in \mathbb{R}
\]  
(2)

2.2.2. Gompertz model

Proposed by Benjamin Gompertz (1779–1865), it is a sigmoid type of mathematical model characterizing growth as being slowest at the start and end of a given period [33]. The differential equation of Gompertz model is given by eq. (3):

\[
\frac{df}{dt} = K(\ln(a) - \ln(f))f, \quad (a, K) \in \mathbb{R}
\]  
(3)

whose solution is given by eq. (4)

\[
f(t) = \theta_1 e^{-\theta_2 e^{-\theta_3 t}}, \quad (\theta_1, \theta_2, \theta_3) \in \mathbb{R}
\]  
(4)

2.2.3. Verhulst model

Firstly, published by Pierre Verhulst [34] as a logistic curve of population growth, it is a time-continuous model. Besides, based on the continuous equation, a discrete quadratic recurrence equation has been developed and widely used as the logistic map. Thus, the logistic model is described by the differential equation (eq. (5)):

\[
\frac{df}{dt} = Af(1 - \frac{f}{B}), \quad (A, B) \in \mathbb{R}
\]  
(5)

which is known as the logistic equation and has solution given in eq. (6)

\[
f(t) = \frac{\theta_1}{1 + \theta_2 e^{-\theta_3 t}}, \quad (\theta_1, \theta_2, \theta_3) \in \mathbb{R}
\]  
(6)

2.2.4. Weibull model

Initially presented as a statistical distribution model, the Weibull model, named after Waloddi Weibull (1951), is being used today. It is used to explain survival in circumstances of damage, disease, or in population dynamics studies. It also has various uses in population expansion and agricultural growth [35]. The function of the model is given by eq. (7):

\[
f(t) = \theta_1 - \theta_2 e^{-\theta_3 t}, \quad (\theta_1, \theta_2, \theta_3) \in \mathbb{R}
\]  
(7)

2.3. Standard formulation of nonlinear regression

The general form of the nonlinear regression model is:

\[
y_i = f(t; \theta) + \epsilon_i, \quad t = 1, ..., n
\]  
(8)

where:

. \( y_i \) is the dependent or response variable;
. \( t \) is the time variable;
. \( \theta \) is the vector of unknown \( p \)-parameters such that \( \theta = (\theta_1, \theta_2, ..., \theta_p) \);
. \( \epsilon_i \) is a random error term.

The model, as written in eq. (8), is valid under the following assumptions [36]: correct mean of function \( f \), variance homogeneity, independence of errors \( \epsilon_i \), normality of errors \( \epsilon_i \). In nonlinear regression, when
The assumptions are violated, this can lead to significant bias in the estimates [36]. Estimation methods that consider the autocorrelation of errors provide much better results than the ordinary least squares method [28]. In the absence of homoscedasticity of errors, the estimate of the parameters is not influenced but unreliable confidence intervals can be obtained [29, 36].

If \( f(t; \hat{\theta}) \) is the function which better fits the relation between \( y \) and \( t \), then we must have the minimum of the residuals. The objective is therefore to find the values \( (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_p) \) that minimize the residual sum of squares, namely:

\[
RSS = \sum_{t=1}^{n} \left[ y_t - f(t; \hat{\theta}) \right]^2
\]

(9)

The RSS becomes the minimum when the estimations of \( \theta \) are entered into eq. (9); the \( \hat{\theta} \) may then be calculated by setting the first partial derivative of RSS with respect to \( \theta \) to zero (\( \frac{\partial RSS}{\partial \theta} = 0 \)), which produces the \( p \) normal equations that must be solved for. Nonlinear ordinary least squares (OLS) and the maximum likelihood method (ML) are the two frequently used techniques for estimating the parameters of nonlinear models.

Among the recurrent issues in nonlinear modeling, one can cite the estimation of initial parameters’ values. However, by correctly comprehending the parameters’ definitions in the context of the phenomenon being modeled, the issue of specifying the beginning values of parameters can be resolved. The parameter \( \theta_1 \), which is straightforward to calculate, is defined in the growth models and provided as the dependent variable’s highest potential value. As a result, \( \theta_1 \) was set equal to the maximum number of COVID-19 cumulative cases in our simulations. The other parameters are determined by linearizing the nonlinear model for the study period.

2.4. ETS modeling

The main idea of ETS models is to decompose any time series into three components (error, trend and seasonal) that will be used to make short forecasts, which are suitable for highly dynamic phenomena [37]. There are 15 possible combinations, given in Table 1, of trend and seasonal components. Considering the fact that the error can be additive or multiplicative, 30 combinations in total are obtained. Each ETS model is linked to time series of its component that are described by the equations of state. For instance, ETS (A, AD, N) is defined by eq. (10):

\[
\begin{align*}
    y_{t} &= l_{t-1} + \phi b_{t-1} + \epsilon_t \\
    l_{t} &= l_{t-1} + \phi b_{t-1} + \alpha \epsilon_t \\
    b_{t} &= \phi b_{t-1} + \beta \epsilon_t
\end{align*}
\]

(10)

where \( l_t \) is the level component, \( b_t \) the trend component and \( \epsilon_t \) the error component of ETS model. \( 0 < \alpha, \beta, \phi < 1 \).

2.5. Proposed ETS-growth model formulation

In nonlinear regression, the choice of the function \( f \) is as important as the vector of the residuals. Indeed, the vector of residuals must follow a well-defined distribution to avoid modeling errors. In the case of nonlinear regression, several assumptions must be checked to ensure that the constructed model is robust and that it can be used for making prediction. In nonlinear regression, the absence of homoscedasticity and autocorrelation are necessary. In the implementation of growth models, one is often confronted with the violation of these assumptions and therefore with the problem of quality of the forecasts. We propose the following model (eq. (11)) to overcome the problem of heteroscedasticity and autocorrelation.

\[
\begin{align*}
    y_{t} &= f(t; \theta) + z_t + \mu_t \\
    z_t &= ETS
\end{align*}
\]

(11)

where \( \mu_t \) is a Gaussian white noise.

Assuming, for example, that the ETS (A, AD, N) process is the one that better fits the error of eq. (8) model, eq. (11) can be rewritten as follows (eq. (12)):

\[
\begin{align*}
    y_{t} &= f(t; \theta) + z_t + \mu_t \\
    z_t &= l_{t-1} + \phi b_{t-1} + \epsilon_t \\
    l_{t} &= l_{t-1} + \phi b_{t-1} + \alpha \epsilon_t \\
    b_{t} &= \phi b_{t-1} + \beta \epsilon_t
\end{align*}
\]

(12)

where \( z_t \) is an ETS process, \( l_t \) the level component, \( b_t \) the trend component and \( \epsilon_t \) the error component of ETS model. \( 0 < a, \beta, \phi < 1 \).

To estimate the different parameters of eq. (11), we start by fitting the standard model to obtain the vector \( \theta \), then we test the different assumptions. If an assumption (homoscedasticity of errors, no autocorrelation of errors) is violated, the error is modeled by an ETS process to obtain the best ETS model. The ETS process used in the model is then incorporated to estimate the vector \( \theta \) again. The Fig. 1 summarizes the different steps of modeling.

2.6. Models comparison metrics

To assess the relative quality of statistical models, we computed the Akaike information criteria (AIC) and the Bayesian information criteria (BIC) [38]. In addition, for diagnostic purpose, we used mean absolute percentage error (MAPE), mean absolute error (MAE), and root mean squared error (RMSE) to evaluate the variation in the errors related to our forecast. These metrics are all increasing functions of the error. Consequently, a well-fitted model will have lower MAE, MAE, RMSE, AIC, and BIC. The formulas of those metrics are provided by the following equations (eqs. (13) to (17)):

\[
\text{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100
\]

(13)

\[
\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} e_i
\]

(14)

\[
\text{RMSE} = \sqrt{\frac{\text{RSS}}{n - p - 1}}
\]

(15)

\[
\text{AIC} = n \ln \left( \frac{\text{RSS}}{n} \right) + 2p
\]

(16)

\[
\text{BIC} = n \ln \left( \frac{\text{RSS}}{n} \right) + n \ln(n)
\]

(17)

As in linear regression, it is important to ensure the significance of the coefficients of the model, although the principle differs. In the case of nonlinear regression, we defined the null hypothesis as \( H_0 : \theta_j = 0, j = 1, \ldots, p \) with \( p \) the number of parameters and the alternative as \( H_1 : \exists j \theta_j \neq 0, j = 1, \ldots, p \). Besides, we computed the confidence interval (CI) at a% risk level. The null hypothesis is rejected if the confidence interval does not contain zero. In this case, the coefficient is significant at the threshold a [39].
3. Results

3.1. Initial values

For the exponential model we obtained, after linearization, $\hat{\theta}_1 = 41693.48$, and for the three other models, as we explained above, we take $\hat{\theta}_1$ as the maximum of daily cumulative COVID-19 cases, so $\hat{\theta}_1 = 5384872$. The parameter $\hat{\theta}_1$ is taken equal to $\hat{\theta}_1$ for Weibull model. The other values, shown in Table 3, are obtained by linearization.

3.2. Estimation of parameters

Using the Minpack.lm package in R, we estimated the parameters of the standard model using the Levenberg-Marquardt algorithm [40] because this method avoids some limitations of Newton-Gauss methods and gradient descent. Starting from the initial values provided in Table 3, the estimates are grouped in Table 4. The graphs in Fig. 2 give the appearance obtained by the different models with the real curve. We can see that the fit is of poor quality. We will test the hypotheses of homoscedasticity and no autocorrelation of errors with the standard formulation. We will first use a graphical and then an analytical approach. In Figs. 3 to 6, we suspect that the residuals are autocorrelated and do not appear to follow a normal distribution. In addition, the error variance seems to be not constant.

For the analytical approach, we used the Shapiro-Wilk test [41] to check the normality of errors, the Ljung-Box [42] for autocorrelation and Levene’s test [43] for homoscedasticity. The associated $p$-values provided in Table 2 are less than 5% threshold. Consequently, we reject the null hypotheses of normality, no autocorrelation and homoscedasticity of errors.

The assumptions of homoscedasticity and no autocorrelation of errors are violated in the standard formulation, we will then use the ETS method to see if our approach corrects this problem. Table 5 gives the estimated values of the smoothing parameters of the best ETS model for each growth model and the Table 6 gives the estimates of the coefficients from the proposed model. We noticed that for the study period, the most appropriate is the ETS (A, AD, N). Figs. 7 to 10, present the graphs of the residuals obtained with the proposed model. The graphs seem to indicate that hypotheses violation from the standard formulation has been corrected by the proposed model.

We notice that in Table 7 all the $p$-values associated with the Shapiro-Wilk test are less than the 5% threshold. So, we reject the null hypothesis of normality of errors. This result is not surprising because our variable of interest is a cumulative one [7]. With the proposed model, we can see that the homoscedasticity and no autocorrelation assumptions are satisfied. Consequently, we can use these models to make short-term forecasts, but first we need to test the goodness-of-fit to choose the best model. We note from Table 8 that the Gompertz and

---

**Table 2.** $p$-values associated with the Shapiro-Wilk, Ljung-Box and Levene tests (standard formulation).

| Model       | Shapiro-Wilk | Ljung-Box | Levene |
|-------------|--------------|-----------|--------|
| Exponential | 8.142e-14    | <2.2e-16  | <2.2e-16|
| Gompertz    | 4.667e-12    | <2.2e-16  | <2.2e-16|
| Verhulst    | 2.889e-13    | <2.2e-16  | <2.2e-16|
| Weibull     | 1.506e-09    | <2.2e-16  | 1.706e-09|

**Table 3.** Initial values of parameters.

| Parameters   | $\theta_1$ | $\theta_2$ | $\theta_3$ | $\theta_4$ |
|--------------|------------|------------|------------|------------|
| Exponential  | 41693.48   | 0.013      | --         | --         |
| Gompertz     | 5384872    | 8.925      | 0.099      | --         |
| Verhulst     | 5384872    | 235.120    | 0.019      | --         |
| Weibull      | 5384872    | 5384872    | 1.178e-06  | 2.361      |
Verhulst models have relative low MAPEs (2.77% and 2.75%, respectively). The Gompertz model, on the other hand, has an AIC (8678.1) and a MAE (163325) lower than the corresponding Verhulst metrics (AIC: 8678.9, MAE: 168641). Let’s recall that these two models have substantially equal statistics so they are all adequate for modeling the dynamic behavior of COVID-19. However, the Gompertz model is the one that overperforms its competitors. We also computed the analysis of variance between the Gompertz model (the one with lower metrics) and Weibull model (the one with a higher number of parameters) and the outputs provided in Table 9 confirm the results obtained with metrics analysis. As we can see from Table 9, the $p$-value associated with the Fisher statistics is greater than 5% threshold, indicating that the modified Gompertz model outperforms its considered competitors. We will therefore use this model to make one step ahead predictions.

We have previously separated the data into two groups: the training data and test data to assess the quality of the adjustment. Now that the assessment is made and we know the best model, we will estimate the coefficients again on all observed data before making forecasts. By taking the estimate of the Gompertz model, we obtain the following system:

$$\begin{align*}
\dot{y}_t &= 8.635 \times 10^6 e^{-4.955 \times 5.092 \times 10^{-7} t} + \epsilon_t \\
\dot{e}_t &= l_{t-1} + 0.9557 h_{t-1} + \epsilon_t \\
\dot{h}_t &= l_{t-1} + 0.9557 h_{t-1} + 0.9999 \epsilon_t \\
\dot{\epsilon}_t &= 0.9557 h_{t-1} + 0.9999 \epsilon_t
\end{align*}$$

Using eq. (18), we can predict the cumulative cases of COVID-19 as the proposed approach is robust to heteroscedasticity and autocorrelation of errors.

4. Discussion

Although the coefficients from the standard models are significantly different from zero, the Fig. 2 shows some important differences between the estimated and the real values. With the proposed approach,
the gaps seem to be more reduced (Fig. 11). The metrics confirm this reduction of the gaps. The Ljung-Box test applied to the residuals of the four models has shown $p$-values greater than of 5% (exponential: 0.8003; Gompertz: 0.9421; Verhulst: 0.8764; Weibull: 0.8942). Thus, we cannot reject the null hypothesis of absence of autocorrelation. Levine’s test also output $p$-values greater than 5% (exponential: 0.08316; Gompertz: 0.06361; Verhulst: 0.08319; Weibull: 0.1067). These values confirm the assumption of homoscedasticity of errors based on Figs. 7 to 10. The confidence intervals obtained with the proposed model can be used for a more reliable interpretation as they do not contain zero and the corresponding $p$-values are less than 5% (Coefficients are significant). Let’s recall, that there is a need to test for homoscedasticity and absence of autocorrelation of errors before embarking on the prediction regardless significance of the coefficients estimated with the standard formulation. Using the proposed model, we obtained estimates that are not very far from the real values in view of the MAPE (0.027). These results are closer to the ones obtained by [6] considering the period over 100 days on the training set. We found that the Gompertz model is the one which best models COVID-19 growth in Africa. Working with nonlinear growth models, [21] found that Weibull model was the most...
suitable to COVID-19 growth in Iraq. The difference between the two studies shows that there is no single way to select immediately one model of COVID-19 cumulative cases. Weibull model has more parameters than the Gompertz one and when we have computed the analysis of variance between the two models, we have resulted to the choice of the Gompertz model as the best. We can therefore see that complex models are not always the most suited in COVID-19 data. We also found that ETS (A, AD, N) fits well the residuals from the standard model. Working on COVID-19 data, [44] used an Autoregressive (AR) process to correct the autocorrelation issue. They also found that the prediction is better with the model that considers the issue of autocorrelation. In this work, we focused on the four most widely used nonlinear models and concluded that the Gompertz model is the most adapted. A broader consideration of growth models may lead to a different conclusion. The limitations of this study are in line with the fact that the forecasts are done for a short period of time. This can be explained using the ETS model that is appropriate for short term forecasts. In sum, we have mainly worked on the computational aspects of the model. In our next work, we will expand the model’s theoretical aspects and apply it to a larger set of data (engineering, medicine, finance).
Table 6. Estimated parameters from the proposed model.

| Model    | Parameters | Estimates | Std. Error | 95% confidence interval |
|----------|------------|-----------|------------|-------------------------|
| Exponential | $\theta_1$ | 5.167e+05 | 9.425e+01 | 5.164e+05 - 5.168e+05 |
|           | $\theta_2$ | 5.251e+03 | 4.685e+07 | 5.249e+03 - 5.253e+03 |
| Gompertz  | $\theta_1$ | 7.081e+06 | 1.921e+03 | 7.077e+06 - 7.084e+06 |
|           | $\theta_2$ | 5.362e+00 | 1.486e+03 | 5.358e+00 - 5.366e+00 |
|           | $\theta_3$ | -6.067e-03 | 2.002e-06 | -6.071e-03 - -6.063e-03 |
| Verhulst  | $\theta_1$ | 5.523e+06 | 7.221e+02 | 5.521e+06 - 5.525e+06 |
|           | $\theta_2$ | 4.070e+01 | 2.208e+02 | 4.065e+01 - 4.074e+01 |
|           | $\theta_3$ | 1.288e-02 | 2.731e-06 | 1.287e-02 - 1.289e-02 |
| Weibull   | $\theta_1$ | 6.575e+06 | 3.087e+03 | 6.568e+06 - 6.583e+06 |
|           | $\theta_2$ | 6.582e+06 | 3.247e+03 | 6.576e+06 - 6.588e+06 |
|           | $\theta_3$ | 3.057e-06 | 1.155e-08 | 3.034e-06 - 3.079e-06 |
|           | $\theta_4$ | 2.137e+00 | 7.564e-04 | 2.135e+00 - 2.138e+00 |

Fig. 7. (a) residuals plot, (b) standardized residuals plot, (c) plot of autocorrelation, (d) Q-Q plot of the proposed exponential model.

Table 7. p-values associated with the Shapiro-Wilk, Ljung-Box and Levene tests (proposed model).

| Models   | Shapiro-Wilk | Ljung-Box | Levene |
|----------|--------------|-----------|--------|
| Exponential | 8.142e-14   | 0.8003    | 0.08316 |
| Gompertz  | 1.917e-13    | 0.9421    | 0.06361 |
| Verhulst  | 3.412e-13    | 0.8764    | 0.08319 |
| Weibull   | 1.506e-09    | 0.8942    | 0.10670 |

5. Conclusion

In this work, we studied the evolution of COVID-19 using four of the most widely used growth models, namely the exponential, the Gompertz, the Verhulst and the Weibull models. Using the Levenberg-Marquardt algorithm, the standard nonlinear regression approach converged for these four models but with violation of the assumptions. A new approach has been proposed to correct the issues of heteroscedasticity and autocorrelation by taking the errors of the standard approach and modeling them with the ETS method. The ETS (A, AD, N) is the one that best characterizes the error dynamics that we obtained with the standard nonlinear regression approach. The proposed method corrects the issues of heteroscedasticity and autocorrelation of errors in nonlinear regression. We predicted the cumulative number of COVID-19 cases over the period June 27, 2021, to July 25, 2021, and compared the considered four models using metrics such as MAPE, MAE, RMSE, AIC, and BIC. We lead to the conclusion that the Gompertz model was more suitable for modeling the evolution of COVID-19 in Africa. This study does not only offer a robust model for predicting the cumulative number of COVID-19 in Africa, but a tool for correcting heteroscedasticity and autocorrelation of errors in nonlinear regression. Researchers and policymakers can therefore use this new approach to predict the number of COVID-19 cases and for better decision making in public health.

Declarations

Author contribution statement

Mintodé Nicodème Atchadé: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Paul Tchanati: Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.
Fig. 8. (a) residuals plot, (b) standardized residuals plot, (c) plot of autocorrelation, (d) Q-Q plot of the proposed Gompertz model.

Fig. 9. (a) residuals plot, (b) standardized residuals plot, (c) plot of autocorrelation, (d) Q-Q plot of the proposed Verhulst model.

Table 8. Comparison metrics from the proposed model.

| Metrics | Exponential | Gompertz | Verhulst | Weibull |
|---------|-------------|----------|----------|---------|
| MAPE    | 1.799019e-01 | 2.774732e-02 | 2.759458e-02 | 2.812591e-02 |
| MAE     | 1.070642e+06 | 1.633259e+05 | 1.686418e+05 | 1.659755e+05 |
| RMSE    | 1.071117e+06 | 1.836820e+05 | 2.038066e+05 | 1.865757e+05 |
| AIC     | 8.679653e+03 | 8.678044e+03 | 8.678885e+03 | 8.678906e+03 |
| AICc    | 8.679705e+03 | 8.678130e+03 | 8.678971e+03 | 8.679035e+03 |
| BIC     | 8.692118e+03 | 8.694664e+03 | 8.695504e+03 | 8.699680e+03 |
**Table 9.** Analysis of variance output between Gompertz and Weibull models.

| Model   | Res.Df | Res.Sum Sq | Df | Sum Sq | F-Statistics | p-value |
|---------|--------|------------|----|--------|--------------|---------|
| Gompertz| 468    | 2722234093 | -  | -      | -            | -       |
| Weibull | 467    | 2715660014 | 1  | 6574079| 1.1305       | 0.2882  |

Res.Df: residuals degrees of freedom, Res.Sum Sq: residuals sum of squares, Df: degrees of freedom.

**Fig. 10.** (a) residuals plot, (b) standardized residuals plot, (c) plot of autocorrelation, (d) Q-Q plot of the proposed Weibull model.

**Fig. 11.** Real and estimated values from the proposed models: (a) exponential, (b) Gompertz, (c) Verhulst, (d) Weibull.
Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability statement

Data associated with this study has been deposited at https://ourworldindata.org/.

Declaration of interests statement

The authors declare no conflict of interest.

Additional information

Supplementary content related to this article has been published online at https://doi.org/10.1016/j.heliyon.2022.e11057.

Acknowledgements

We thank the editor and the anonymous reviewers for their careful reading of our manuscript and their many insightful comments and suggestions. The authors are also grateful to the lecturers at the International Chair in Mathematical Physics and Applications (ICMPE-UNESCO Chair), who were quite helpful in the improvement of this paper. We thank Mr. Yves Morel Sokadjio, who helped to proof-read this paper.

References

[1] Nouveau coronavirus (2019-nCoV), https://www.who.int/fr/emergencies/diseases/novel-coronavirus-2019, 2019.
[2] M.N. Atchadé, Y. Sokadjio, The influence of passenger air traffic on the spread of Covid-19 in the world, Transp. Res. Interdiscip. Perspect. 8 (2020).
[3] X. Jiang, B. Zhao, J. Cao, Statistical analysis on Covid-19, Biomed. J. Sci. Tech. Res. 26 (2020) 19716–19727.
[4] S. Zhao, H. Chen, Modeling the epidemic dynamics and control of Covid-19 outbreak in China, Quant. Biol. (Beijing, China) 8 (2020) 11–19.
[5] C.Y. Shen, Logistic growth modelling of Covid-19 proliferation in China and its international implications, Int. J. Infect. Dis. 96 (2020) 582–589.
[6] M.N. Atchadé, Y.M. Sokadjio, A.D. Mounsa, S.V. Kurisheva, M.V. Bochenova, Cross-validation of Covid-19 forecast models, SN Comput. Sci. 2 (2021) 296.
[7] M.N. Atchadé, Y.M. Sokadjio, Overview and cross-validation of Covid-19 forecasting univariate models, Eng. Appl. 61 (2022) 3021–3036.
[8] K.D.M. Mainuma Majumder, Early Transmissibility Assessment of a Novel Coronavirus in Wuhan, China, 2020, 3254675.
[9] B.M. Ndziaye, L. Tendeng, D. Seck, Analysis of the Covid-19 pandemic by sir model and machine learning techniques for forecasting, arXiv: Popul. Evol. (2020).
[10] J. Yang, Z. Zeng, K. Wang, S.S. Wong, W. Liang, M. Zanin, P. Liu, X. Cao, G. Zao, Z. Mai, J. Jiang, X. Liu, S. Li, Y. Li, F. Ye, W. Guan, Y. Yang, F. Li, S. Luo, Y. Xie, B. Liu, Z. Wang, S. Zhang, Y. Wang, N. Zhong, J. He, Modified sirf and ai prediction of the epidemics trend of Covid-19 in China under public health interventions, J. Thorac. Dis. 12 (2020) 165–174.
[11] R. Singh, R. Adhikari, Age-structured impact of social distancing on the Covid-19 epidemic in India, arXiv preprint, arXiv:2003.12055, 2020.
[12] G. Agossou, M.N. Atchadé, A. Djibril, Modeling the effects of preventive measures and vaccination on the Covid-19 spread in Benin republic with optimal control, Results Phys. 31 (2021) 104969.
[13] R. Sujath, J.M. Charterjee, A.E. Hassanien, A machine learning forecasting model for Covid-19 pandemic in India, Stoch. Environ. Res. Risk Assess. 34 (2020) 1–14.
[14] S.J. Fong, G. Li, N. Dey, R. Gonzalez-Crespo, E. Herrera-Viedma, Finding an accurate early forecasting model from small dataset: a case of 2019-ncov novel coronavirus outbreak, Int. J. Interact. Multimed. Artif. Intell. 6 (2020) 132–140.
[15] P. Wang, X. Zheng, G. Ai, D. Liu, B. Zhu, Time series prediction for the epidemic trends of Covid-19 using the improved LSTM deep learning method: case studies in Russia, Peru and Iran, Chaos Solitons Fractals 140 (2020) 110214.
[16] S. Lalamawanwa, J. Hussain, L. Chakhchhuak, Applications of machine learning and artificial intelligence for Covid-19 (Sars-CoV-2) pandemic: a review, Chaos Solitons Fractals 139 (2020) 110059.
[17] P. Arora, H. Kumar, B.K. Parigrahi, Prediction and analysis of Covid-19 positive cases using deep learning models: a descriptive case study of India, Chaos Solitons Fractals 139 (2020) 110017.
[18] G. R. P. Sk, P. G, A comprehensive analysis of Covid-19 outbreak situation in India, medRxiv (2020).
[19] C. Barzit-Sandoval, G. Ferreira, K. Benz-Parrza, P. López-Flores, Prediction of confirmed cases of and deaths caused by Covid-19 in Chile through time series techniques: a comparative study, PLoS ONE 16 (2021) 1–16.
[20] E. G deceit, A. Zidati, R.D. Szczenniak, Forecasting Covid-19 confirmed cases, deaths and recoveries: revisiting established time series modeling through novel applications for the USA and Italy, PLoS ONE 16 (2021) 1–11.
[21] B.G. Al-Ani, Statistical modeling of the novel Covid-19 epidemic in Iraq, Epidemiol. Methods 10 (2021) 20200025.
[22] C. Mahanty, R. Kumar, B.K. Mishra, D.J. Hemanth, D. Gupta, A. Khanna, Prediction of Covid-19 active cases using exponential and non-linear growth models, Expert Syst. (2020).
[23] P. Wang, X. Zheng, J. Li, B. Zhu, Prediction of epidemic trends in Covid-19 with logistic model and machine learning techniques, Chaos Solitons Fractals 139 (2020) 110059.
[24] Y. Wen, K. Liu, H. Liu, H. Cao, H. Mao, X. Dong, Z. Yin, Comparison of nine growth curve models to describe growth of partridges(alectores chukar), J. Appl. Anim. Res. 47 (2019) 195–200.
[25] C.F.R. Toviscood, B.E. Lokonon, R.G. Kakai, On the use of growth models to understand epidemic outbreaks with application to Covid-19 data, PLoS ONE 15 (2020) 1–14.
[26] A. Ahmedy, Y. Fadaei, M. Shirani, F. Rahmani, Modeling and forecasting trend of Covid-19 epidemic in Iran until may 13, 2020, Med. J. Islam. Repub. Iran 34 (2020) 27.
[27] H.M. Al-Saedli, H.H. Hameed, Mathematical modeling for Covid-19 pandemic in Iraq, J. Interdiscip. Math. 24 (2021) 1407–1427.
[28] A.R. Gallant, J.J. Goebel, Nonlinear regression with autocorrelated errors, J. Am. Stat. Assoc. 71 (356) (1976).
[29] Raymond J. Carroll, David Ruppert, Transformation and Weighting in Regression, Monographs on Statistics and Applied Probability, Springer US, 1988.
[30] D. Ruppert, Transformation and weighting, in: The Work of Raymond J. Carroll, Springer, 2014, pp. 155–194.
[31] Hannah Ritchie, Edouard Mathieu, Lucas Rodríguez-Gurino, Cameron Apple, Charlie Giattino, Esteban Ortiz-Orozco, Joe Hasell, Bobbie Macdonald, Saloni Dutani, Max Roser, Coronavirus pandemic (Covid-19), our world in data, https://ourworldindata.org/coronavirus, 2020.
[32] T. Malinva, An Essay on the Principle of Population, Printed for J. Johnson, in St. Paul's Church-Yar 1798.
[33] B. Gompertz, On one uniform law of mortality from birth to extreme old age, and on the law of sickness, J. Inst. Actuar. Assur. Mag. 16 (5) (1871).
[34] M. Schickelze, Pierre-Francois Verhulst (1804-1849). La première découverte de la fonction logistique, Population 36 (3) (1981).
[35] D.J. Mahanta, M. Borah, Parameter estimation of Weibull growth models in forestry, Int. J. Math. Trends Technol. 8 (2014) 157–163.
[36] C. Ritz, J. Streibig, Nonlinear regression with R (use R!), 2008.
[37] R.J. Hyndman, Forecasting with Exponential Smoothing: the State Space Approach, Springer Series in Statistics, Springer, Berlin Heidelberg, 2008.
[38] H. Akaie, A new look at the statistical model identification, IEEE Trans. Autom. Control 19 (1974) 716–723.
[39] D. Fekedulegn, J. Colbert, Parameter estimation of nonlinear growth models in forestry, Silva Fenn. 33 (1999).
[40] D.W. Marquardt, An algorithm for least-squares estimation of nonlinear parameters, J. Soc. Ind. Appl. Math. 11 (1963) 431–441.
[41] S.S. Shapiro, M.B. Wilk, An analysis of variance test for normality (complete samples), Biometrika 52 (1965) 591–611, http://www.jstor.org/stable/2333709.
[42] R.A. Brockwell, P.J. Davis, Introduction to time series and forecasting by P.J. Brockwell, R.A. Davis, Biometrics 54 (3) (1998).
[43] J.B. Sprague, Guideline Document on Statistical Methods for Environmental Toxicity Tests, Environment Canada, Ottawa, ON, 2005. OCLC: 769249127.
[44] R. Pérez Arceu C., S. Estrada, H. de-la Torre-Gutiérrez, A two-step polynomial and nonlinear growth approach for modeling Covid-19 cases in Mexico, Mathematics 9 (2021).