Global Scientific JOURNALS

GSJ: Volume 10, Issue 2, February 2022, Online: ISSN 2320-9186

www.globalscientificjournal.com

# MODELLING OF CONFIRMED CASES OF COVID-19 IN NIGERIA USING GENERALIZED ADDITIVE MODEL (GAM)

Saheed Bolaji SULEIMAN [1], Kamoru JIMOH[2] , Saheed Olalekan JABARU[3]

[1,2,3]Department of Physical Sciences (Statistics Unit), Al-Hikmah University, Ilorin, Nigeria.

E-mail: jimkaminsha@yahoo.com;

simplysaheed@gmail.com

saheedstat020677@yahoo.com

## *ABSTRACT*

*Coronavirus (Covid-19),the deadly pandemic continues to spread across the globe and there is need to study the nature of spread of the pandemic since researchers still do not know definitively the effectiveness of the said vaccine for the pandemic (Covid-19). Time series model was used to show the nature and behavior of confirmed cases of the pandemic in Nigeria. Firstly, a comparison of two models (GLM and GAM) was studied for the confirmed cases of covid-19 pandemic and GAM was suspected to be a better model for the data set with the minimum selection criterion. Secondly, for the numerical solution of the Generalized Additive Model (GAM), the* ***Cubic regression(cr)*** *spline was adopted and the knots were varied with seven days interval* ***(k= 7, 14 , 21 , 28). A GAM with k=28*** *was chosen among other GAM with different kth value since it has the best p-value, the k-index is greater than one(1), the* ***expected degree of freedom (edf)*** *departed widely from k-1and the diagnostic plot shows the normality assumption of the model. Finally, some graphical results were presented and the findings showed that, if all necessary preventives measures are abided by individuals in the country, there is tendency of the figures of Covid-19 pandemic to drop to insignificant figure in the future.*

**Key words:**Modeling, Additive,Splines, Knot, GAM.

## 1.0 Introduction

COVID-19 started in China in the City of Wuhan around 31st of December 2019. Due to numerous infected cases in the "Huanan Seafood Market" area, the situation became critical (Aljazeera, 2020). The government of China documented the incidents of the disease having symptoms like Pneumonia but the nature of the virus was still unknown. In a span of thirty days,the cases increased to about 40. History showed that, this Virus SARS (severe acute respiratory syndrome) started in China and took away the lives of around 770 people in the years 2002 and 2003.

The Virus is demonstrated as a respiratory disorder and ithas 2 to 14 days incubation period (Lauer et al., 2020). Its symptoms are: Fever, Runny nose, Diarrhea, Dry cough, Sore throat, Shortness of breath, Muscle joint Pains and Loss of senses like smell and taste in some cases.

The coronavirus is transmitted and spread in the following cases:

i.Person getting in contact with the droplets or secretions of someone sneezing or coughing and float down the respiratory tract.

ii. Physical contact with the virus through hands and taking it to the mouth.

Therefore, in order to decrease the risk of its spread, it is advised toavoid touching your face and keep washing your hands in order to eliminate the risk in case a person is exposed to it accidently. Using the alcoholic hand sanitizers can lead to increased chance of staying safe from the virus (Krisinformation, 2020).

People already experiencing respiratory disorders and elderly are mainly the target group of the virus.

## 1.1 CORONA VIRUS INFECTION IN NIGERIA

The methodology (Social distancing and masked up) adopted by Nigeria is based upon the voluntary cooperation and responsibility of the Nigeria individuals.

In the beginning, all the countries adopted the strict lockdown and had shut downed all the schools. But now, that situation has been relaxed and all the businesses now seemed to be opening again. All countries have a similar aim that is, to slow down the spread although the methodology might be different (The Local, 2020a).

## 1.2 AIM AND OBJECTIVES

The aim of this research is to establish a model for Coronavirus (COVID-19) data with respect to spread with time. However, the objectives are to:

- i. construct a generalize linear model (GLM);
- ii. construct a generalize additive model (GAM);
- iii. establish the tentative best fitted model for the data set;
- iv. diagnostic checking of the chosen model; and
- v. application of the best model to the prediction of Covid-19 Cases in Nigeria.

## 2.0 GENERALIZED LINEAR MODEL

The generalized linear model (GLM), as an extension of linear regression model, is a framework for modelling the dependent variables that are bounded or discrete, formulated by Nelder and Wedderburn(2016). The differences between GLM and the general linear regression model are in two major aspects: first, unlike in linear regression model, the distribution of the dependent variable does not have to be continuous, and can be non-normal as long as it follows the exponential family of distributions such as the Poisson, binominal, gamma and normal distributions; second, the dependent variable values are predicted by a linear combination of independent variables, which are connected to the dependent variable via a link function. A GLM has its basic structure as follows:

$$\boldsymbol{g(\mu_i) = x_i\beta} \qquad \textbf{(2.1)}$$

where$\mu_i \equiv E(Y_i)$is the expectation of random dependent variable $Y_i$, $g$ is a smooth monotonic "link function", $x_i$ is the ith row of a model matrix, X and β is a vector of unknown parameters.

## 2.1 GENERALIZED ADDITIVE MODEL

The generalized additive model (GAM) was originally developed by Hastie et al, (1987), derived by the combination of additive model(AM) and generalized linear model(GLM). In general, GAM has a structure as follows:

$$\boldsymbol{g(\mu_i) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_k(x_k)} \qquad \textbf{(2.2)}$$

In other words, the purpose of GAM is to maximize the quality of prediction of a dependent variable from various exponential distributions, by estimating nonparametric functions of the independent variables whichare connected to the dependent variable via a link function.

GAM is introduced elaborately by Wood, S.N (2017), with a general structure as follows:

$$\boldsymbol{g(\mu_i) = A_i\theta + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + \cdots} \qquad \textbf{(2.3)}$$

where$\mu_i \equiv E(Y_i)$and $Y_i \sim EF(\mu_i, \varphi)$,$Y_i$ is a response (dependent) variable and $EF(\mu_i, \varphi)$denotes an exponential family of distribution with mean $\mu_i$ and scale parameter, $\varphi$, $A_i$ is a row of the Model matrix for any strictly parametric model components, θ is the corresponding parameter vector and the $f_j$ are smooth (non-parametric) functions of the covariates, $x_k$.

## 2.3.1 PRESENTATION OF FUNCTIONS WITH BASIC EXPANSIONS

The presentation and estimation of component functions of a GAM model is best introduced by considering a model containing one function and one covariate, just as follows:

$$\boldsymbol{y_i = f(x_i) + \varepsilon_i} \qquad \textbf{(2.4)}$$

where $\boldsymbol{y_i}$ is a response variable, $\boldsymbol{x_i}$ is a covariate, $f$ is a smooth function and the $\boldsymbol{\varepsilon_i}$**is**independent,$N(0, \sigma^2)$ random variables.

To estimate $f$ requires that f be represented in such a way that (2.4) becomes a linear model. This can be done by choosing some basic functions, defining the space offunctions of which $f$ is an element. If $b_j(x)$is the $j^{th}$of such basic function for somevalues of the unknown parameters, then,$f$ is assumed to have a representation as follows:

$$\boldsymbol{f(x_i) = \beta_0 + \sum_{j=1}^{k} f_j(x)\beta_j},\ \text{where} \qquad \textbf{(2.5)}$$

k is the basis dimension, which controls the degree of model smoothness.

### 2.3.2 SPLINES

According to wood et al (2017), representing the smooth model terms using spline basis is likely to obtain substantially reduced function approximation errors for a given dimension of smoothing basis. There are various types of splines in use, and among the most common onesis cubic regression splines.

### 2.3.2.1Cubic Regression Splines

Cubic spline is a piecewise polynomial with a set of extra constraints (continuity, continuity of the first derivative, and continuity of the second derivative). There are many equivalent bases that can be used to represent cubic splines. Cubic regression splines are the approach to parameterize the spline in terms of its values at the knots. In this case, apart from the continuity constraints, the spline should have zero second derivatives at the start and end knot.

Cubic regression spline takes both smoothness and local influence into consideration unlike other traditional methods such as polynomial regression or broken stick regression (Faraway 2016). This basis does not require any re-scaling of the independent variables before it is used to construct a GAM in order to have directly interpretable parameters, although we do have to choose the locations of the knots.

### 2.4 RESULTS EVALUATION

There are various methods and tests for evaluating the regression results and generally applying only one of them is not enough to judge the model performance. Some of the methods are as follows:

### 2.4.1Adjusted R-square

The adjusted R-square, also known as the coefficient of determination, isthe proportion of the variance in the dependent variable that is predicted from the independent variable(s). $R^2$ normally rangesfrom 0 to 1, and the bigger, the better. The most general definition of $R^2$ is as follows:

$$\boldsymbol{R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \quad where}$$

$$\boldsymbol{SS_{tot} = \sum (y_i - \bar{y})^2 \ , SS_{reg} = \sum (f_i - \bar{y})^2 , SS_{res} = \sum (y_i - f_i)^2}$$

where $y_i$ are the observed values of the dependent variable, $f_i$ are the predicted values from the fitted regression line, $\bar{y}$ is the mean of the observed data, SStot is the total sum ofsquares (proportional to the variance of the data), $SS_{reg}$ is the regression sum of squares and $SS_{res}$ is the residual sum of squares.

### 2.4.2 Akaike Information Criterion

The Akaike information criterion (AIC) is an estimator of the relative quality of statistical models for a given dataset. Given a collection of models for the same data, AIC estimates the quality of each model, relative to each of the other models. Thus, AIC provides a direct means for model selection. The criterion is as follows:

$$\boldsymbol{AIC = -2l + 2p} \qquad \textbf{(2.6)}$$

where l is the maximized log likelihood for the model and p the number of modelparameters that have to be estimated. By using the R language, it is easy to calculate AIC and the model with the lowest AIC is the best.

## 3. ANALYSIS AND RESULTS

Covid-19 Cases data spanning the period from February to November, 2020 were compiled from the website of NCDC, which involves daily Covid-19 cases in Nigeria.

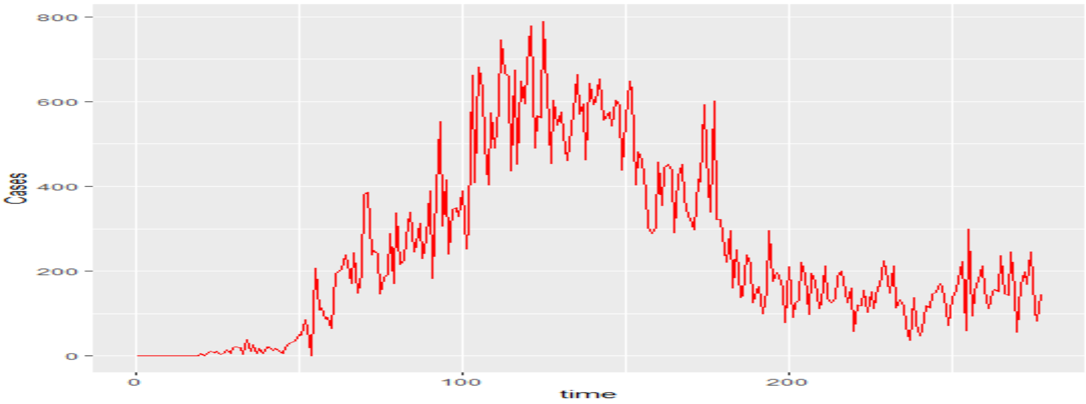

***Figure 3.1*** *above shows a non linear pattern of covid-19 cases in Nigeria from February to November, 2020.*

## RESULT

### Table 3.1 COMPARISON OF GLM AND GAM MODEL

| MODEL | AIC | BIC |
|---|---|---|
| **GLM** | 3730.1 | 3740.964 |
| **GAM** | 3177.319 | 3214.195 |

Based on the selection criteria (AIC and BIC), table 3.1 shows that GAM has the minimum value and it is the best model for the Covid-19 cases prediction in Nigeria.

### Table 3.2a: Parametric coefficients estimates

| MODEL | Parametric coefficients | | | |
|---|---|---|---|---|
| | | Estimate | Std. Error | t value |
| **1 (k=7)** | **Int.** | 244.271 | 4.459 | 54.78 |
| **2 (k=14)** | **Int.** | 244.271 | 4.301 | 56.8 |
| **3 (k=21)** | **Int.** | 244.271 | 4.224 | 57.83 |
| **4 (k=28)** | **Int.** | 244.271 | 4.062 | 60.14 |

### Table 3.2b: Smooth terms estimates

| Approximate significance of smooth terms | | | | | | |
|---|---|---|---|---|---|---|
| | Edf | Ref. df | F | p-value | Dev. Expld. % | R-sq.(adj) |
| **s(time)** | 5.905 | 5.996 | 297.3 | 2.00E-16 | 86.9 | 0.866 |
| **s(time)** | 12.07 | 12.83 | 151.6 | 2.00E-16 | 88.1 | 0.875 |
| **s(time)** | 16.06 | 18.3 | 110.7 | 2.00E-16 | 88.7 | 0.88 |
| **s(time)** | 21.63 | 24.65 | 89.93 | 2.00E-16 | 89.8 | 0.889 |

## DIAGNOSTIC CHECKING/MODEL ADEQUACY

### Hypothesis

Null hypothesis: Series are random.

Alternative hypothesis: series are not random.

Critical region: reject the null hypothesis at 0.05 level of significance if p-value less than 0.05.

**Table 3.3 GAM MODELS WITH VARIOUSE Kth VALUES**

| Kth VALUE | K'=k-1 | edf | K-index | p-value |
|---|---|---|---|---|
| 7 | 6.0 | 5.9 | 0.8 | 2e-16 *** |
| 14 | 13.0 | 12.1 | 0.88 | 0.03 * |
| 21 | 20.0 | 16.1 | 0.92 | 0.12 |
| 28 | 27.0 | 21.6 | 1.02 | 0.58 |

From table 3.3 it can be deduced that for k= 7, 14,and 21, the value of k-index is less than one and the expected degree of freedom (edf) is very close to the value of k-1 except for k= 28 where the k-index is greater than one and its expected degree of freedom is not close to k-1 value. Hence the GAM with k(knot)=28 is the best among others. Also, since P-values are greater than 0.05 for k=21 and K=28 then we conclude that the series are random but GAM with k=28 would be chosen.

### MODEL

$$y_i = 244.271 + f(time) + \varepsilon_i$$

**Test for Normalty ;**K=28

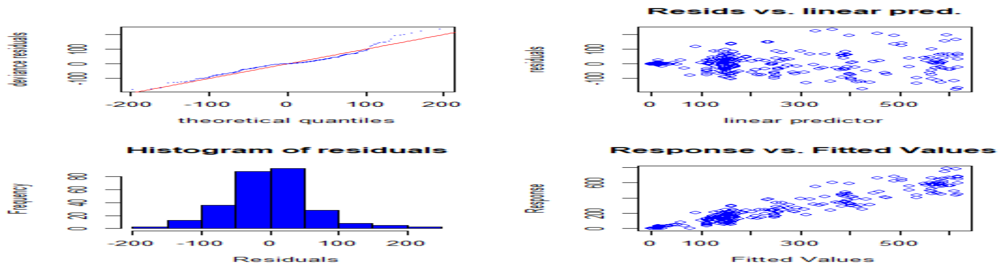

***Figure 3.2*** *above shows diagnostic plot for the model residual for k (Knot) =28.*The quantile-quantile plot(Q-Q Plot) are random which shows the Normality assumption.

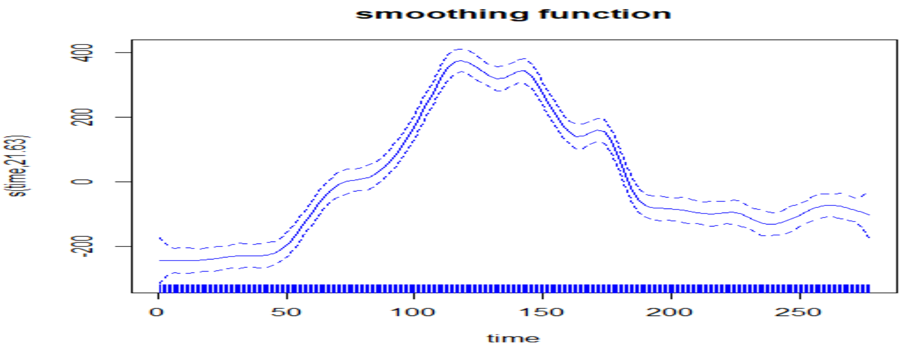

***Figure 3.3:*** Model of the Cases of Covid-19 pandemic as a smooth function of time.

The default plot shows several attributes. The solid line is the predicted value of the dependent variable as a function of time. The small lines along the x axis are the ”rug”, showing the location of the sample plots. The dashed lines are two times the standard errors of the estimates. The y axisis in the linear units, in which the values are centered on 0, and extend to both positive and negative values of the smooth function. This default plot also signifies the posterior means of the nonlinear effect of time of the confirmed cases of Covid-19 and its 95% C.I.

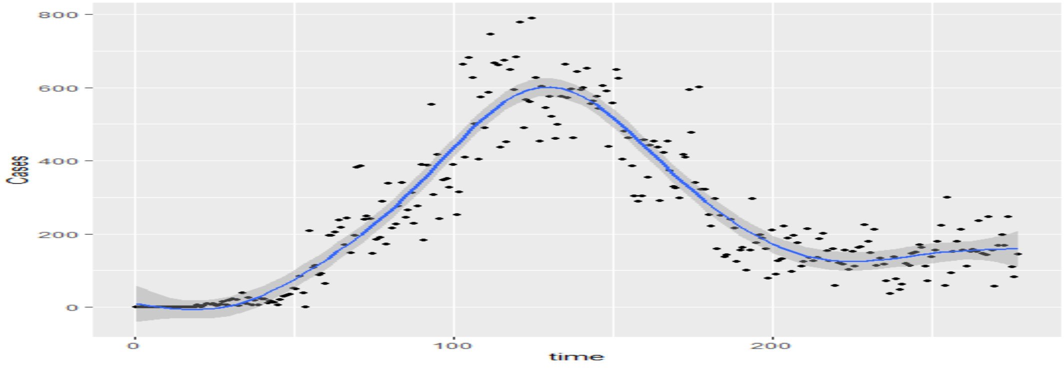

***Figure 3.4***Model of the Cases of Covid-19 pandemic as a smooth function of time.

The blue line is the predicted valueof the dependent variable as a function of the time. The small lines along the x axis are the “rug”, showing the location of the sample plots. The shaded portion shows the confidence interval (95%) lines which are two times the standard errors of the estimates.

## 4. FINDINGS

Based on the findings obtained, GAM is considered as the suitable model for the prediction of the spread of covid-19 cases in Nigeria. From the comparison made, GAM appears to be suitable for studying spread of epidemic. From table 4.3 GAM with **K=28** appears the best for the data set since it has the most satisfied selective condition, that is , it has the highest value of **deviance explained, R-square, P-value, while the k-index values is greater than 1 and also satisfied normality assumption.**

## 5. CONCLUSION

The pattern of the confirmed cases of covid-19 in Nigeria will either remains low or drop drastically to insignificant figure if all corrective and preventive measures are put in place by the agencies concern and every individual.

## REFERENCES

**Aljazeera. (2020, April 29).** Timeline: How the new coronavirus spread | Coronavirus pandemic News | Al Jazeera.

**Faraway, J.J. (2016),**Linear models with R: Chapman and Hall/CRC.

**Krisinformation. (2020).** Slow down the spread of Covid-19 - Krisinformation.se.

**Lauer, S. A., Grantz, K. H., Bi, Q., Jones, F. K., Zheng, Q., Meredith, H. R., Azman, A. S.,Reich, N. G., &Lessler, J. (2020).** The Incubation Period of Coronavirus Disease

**Nelder, J.A. and R.W. Wedderburn(2016)** Generalized linear models. Journal of theRoyal Statistical Society.

**2019 (COVID-19) From Publicly Reported Confirmed Cases:** Estimation and Application. Annals of Internal Medicine.

**The Local (2020).** The Nordic divide on coronavirus: Which country has the right strategy?

**Wood, S.N.( 2017)** Generalized additive models: an introduction with R: Chapman and Hall/CRC.