International Journal of Statistical Distributions and Applications
2024; 10(4): 78-88
http://www.sciencepublishinggroup.com/j/ijsda
doi: 10.11648/j.ijsd.20241004.11
ISSN: 2472-3487 (Print); ISSN: 2472-3509 (Online)

Fitting COVID-19 Incidences in Kenya Using Fractional Polynomials and Linear Splines

Damaris Njoroge[1], Samuel Mwalili[2], Anthony Wanjoya[2]

[1]College of Pure and Applied Sciences, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya
[2]Department of Statistics and Acturial Sciences, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya

Email address:
damarisnjoroge56@gmail.com (Damaris Njoroge)

To cite this article:
Damaris Njoroge, Samuel Mwalili, Anthony Wanjoya. (2024). Fitting COVID-19 Incidences in Kenya Using Fractional Polynomials and Linear Splines. *International Journal of Statistical Distributions and Applications, 10*(4), 78-88.
https://doi.org/10.11648/j.ijsd.20241004.11

Received: 26 September 2024; **Accepted:** 28 October 2024; **Published:** 18 November 2024

Abstract: The study provides an in-depth analysis of COVID-19 infections in Kenya, aiming to model the non-linear trajectory of daily cases. The research explores two statistical techniques: fractional polynomials and linear splines, to fit the growth of infection rates over time. COVID-19, which first appeared in Kenya in March 2020, exhibited fluctuating trends in daily infections. The study utilizes infection data collected from March 13, 2020, to June 6, 2021. Descriptive statistics and exploratory data analysis revealed significant variability in daily cases, with the infection trajectory characterized by multiple waves. Fractional polynomial models, known for their flexibility in fitting non-linear relationships, were evaluated at varying degrees to identify the best model for COVID-19 incidence trends. The analysis showed that a second-degree fractional polynomial with powers (1, 2) provided the most accurate fit for the data. The closed test algorithm was applied to confirm the model's suitability. Additionally, linear spline models were employed, partitioning the data into segments and fitting linear splines at each knot point. The model with 19 knots demonstrated superior performance based on Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), outperforming the fractional polynomial model. The comparison of the two methods concluded that linear splines provided a more precise fit for the infection data, capturing the complex nature of COVID-19's spread in Kenya. The study's findings offer critical insights into the infection dynamics and can aid policymakers in resource allocation and mitigation planning during pandemics. The study recommends further analysis by incorporating more covariates and extending the models to other countries for a comparative understanding of pandemic management strategies.

Keywords: COVID-19, Fractional Polynomials, Linear Splines, Continous Data, Knots Placement, Best Fitting Model, Multivariate Regression Models

1. Introduction

1.1. Background of the Study

Coronavirus popularly abbreviated as COVID-19 continues to spread at an alarming rate and continues to penetrate globally across all continents. According to World Health Organization (WHO), Covid-19 is a respiratory illness which originated in Wuhan city, China on December 2019. The virus is caused by severe acute respiratory syndrome (SARS-CoV-2) whose symptoms include coughing, fever, headache and difficulty in breathing. On 13 March 2020, Kenya reported its first case of the novel coronavirus disease which saw the government move with speed to identify persons who came into contact with the first case. As of June 25, 2020, there were 9,519,482 confirmed cases, 483,959 fatalities and 5,169,767 recoveries globally. Out of these statistics, 5,206 cases were from Kenya with 132 fatalities and 1,823 recoveries [1]. These deaths greatly supersede the number of deaths associated with both severe acute respiratory syndrome coronavirus, SARS-CoV, and Middle East respiratory syndrome coronavirus (MERS-CoV). COVID 19 poses a huge threat to the global public health and the economy considering the outbreak continues to spread within the communities. The coincidence of the emergence of COVID-19 with the Spring Festival travel season was largely attributed for the rapid national and global spread of the virus especially across Europe and Asia [1]. The pandemic not only presents an imminent, but also a severe threat to the lives of the citizens globally, the global healthcare system as well as the global economy. If no serious measures are taken, it is expected that the escalating COVID-19 infection rate could result in significant morbidity and mortality for a large percentage of the global citizens in the future months [2]. The observed reproduction number and mortality rates of COVID-19 have been compared to those of the Spanish Flu of 1918. Recent reports indicate a high fatality rate of approximately 61.5% for critical cases which could be significantly high depending on the patient's age and any underlying comorbidities. Because of this severity, great pressure has been put on medical services which has further led to a shortage of intensive care resources both in developed economies like the United States and Italy as well as the developing economies like Kenya [3]. With the unavailability of prognostic biomarkers aimed at distinguishing patient's requiring immediate attention and consequently estimate their mortality rate, it is an urgent yet challenging necessity to easily identify and urgently accord necessary care to patients at imminent risk of death [4]. It is important to assess severity of COVID-19 in order to determine the appropriate mitigation strategies. This is especially important for a developing economy like Kenya as it enables planning for health-care needs with the increasing rate of infections. Considering the extent of the spread and the nature of its distribution, determining crude case fatalities only by dividing the number of deaths by the number of cases cannot offer a broader context and understanding the problem at hand [3]. This is further expounded by the failure to know the final clinical outcome of a majority of the reported cases during the growing pandemic hence making it impossible to accurately estimate the true case fatality ratio at the earlier stages of the pandemic. It is even difficult in the case of the country like Kenya where the proportion of the population tested is pretty low and cannot be considered to be a representative sample of the pandemic outbreak. This is not unique to Kenya only as data from the epicenter of the outbreak in Wuhan were mainly obtained through hospital surveillance which implies that the reported cases would mainly represent moderate and severe illnesses while the mild cases would be ignored [4]. This biased approach would ultimately lead to a relatively higher case fatality ratio. The spatio-temporal pattern exhibited by COVID-19 infections in Kenya shows scaling trends without illustrating a tendency of stabilization. This study, therefore, seeks to investigate a flexible approach to fitting non-linear trend of the COVID-19 incidence infections. This is to be achieved within the context of promising a success story in the fight against the diseases. This study uses the fractional polynomial (FP) method while maintaining the number of infections as a continuous variable. Fractional polynomial method was selected because of its ability to allow the data to determine the best fitting functional form for the infection rate without imposing a specific functional form [5]. This parametric method for modeling the evolution of the Kenyan Covid-19 infection-curve yields more accurate forecasting compared to the conventional polynomial methods. An alternative approach is the linear splines method which involves partitioning the data into different segments and fitting of linear splines at each partition. This method provides an easier way of interpreting the coefficients of the model which is quite difficult in the case of fractional polynomial method. The two methods are compared using AIC. These methods provide the ability to capture the growth trajectory of the infections in a compact, parsimonious model.

1.2. Statement of the Problem

Since the COVID-19 pandemic started in Wuhan in December 2019, it has spread across the globe at a very alarming rate. Across both developed and developing countries, it has not only caused mortality, but also put considerable stress on health systems with a vast majority of the affected needing critical care including mechanical ventilation [3]. This is why there is urgency in determining an accurate estimate to effectively manage the rising case load while providing the highest quality of health care possible. The forecasts and scenarios for COVID-19 have largely been based on mathematical compartmental models. These models assume random mixing between all individuals in a given population. While results of these models are sensitive to stating assumptions and thus differ between models considerably, they generally suggest that given current estimates of the basic reproductive rate (the number of cases caused by each case in a susceptible population), 25% to 90% of the population could eventually become infected unless effective mitigation measures are put in place and adhered to the latter. A number of sub-Saharan African countries, including Kenya, are at moderate to high risk of novel coronavirus importation, measured by volume of air travel arriving from infected Chinese provinces [3]. And as such, the likelihood of a significant outbreak remains high, with potentially severe consequences for fragile regional health systems. The potentially high negative impact of a novel coronavirus outbreak in Kenya provides a strong motivation for forecasting studies of COVID-19 pandemic magnitude ahead of a serious outbreak under a number of plausible scenarios. This modeling study provides a baseline for continuous updating as improved data become available, e.g. time course of COVID-19 cases, updated mobility estimates, the proportion and infectivity of asymptomatic cases, and new intervention strategies proposed or implemented. It also provides the basis for studies of health service capacity.

1.3. Objectives of the Study

1.3.1. General Objective

To determine COVID-19 incidence infections in Kenya using fractional polynomials and linear splines.

1.3.2. Specific Objectives

1. To fit a fractional polynomial model for COVID-19 infections.

2. To fit a linear spline model for COVID-19 infections.
3. To evaluate the best fitting model for COVID-19 infections.

1.4. Justification of the Study

From the fatality and infections recorded and with the government testing more people, successful management of the pandemic calls for the need to understand the fatalities and infections to determine whether mitigation measures are effective. Considering the public health and economic repercussions brought about by the pandemic, there are high chances that mathematical modelling can help bring into perspective essential epidemiological parameters that determine the fate of the pandemic Other forecasting models focus on modelling the empirically observed COVID-19 population death rate curves which directly reflect both the transmission of the virus and the infection fatality rates in each specific community. In a situation for a country like Kenya, deaths offer more accurate estimates since there is limited testing capacity and the tests are prioritized for more severe ill patients. This is why the proposed methods are considered since they build on a recent conceptualization of detecting communities of connectivity in a time-series and develops a novel model based on the fractional polynomials and linear splines that guarantees more accurate and reliable forecasting [5]. Such an approach not only provides insight of good policy and decision-making practices, but also management that would greatly aid in decision-making and management of the available health resources in the fight against COVID-19 pandemic. Mathematical epidemiologists use models not only to support a broad range of policy questions, but also gain an understanding of the pandemic itself. In far reaching pandemics like COVID-19, mathematical models have been used widely for planning and identifying critical gaps and preparing plans aimed at detecting and responding to pandemic events. This study, therefore, plays a significant role in not only helping understand the transmission of the virus, but also forecast the likely mortality rate and by predicting where transmission is likely to happen and advice on where controls can be put in place.

1.5. Scope of the Study

The scope of this study encompasses the application of statistical modeling techniques to analyze the daily COVID-19 infection rates in Kenya. The research focuses on modeling the non-linear growth trajectory of the pandemic using two specific methods: fractional polynomials and linear splines. The data used spans from March 13, 2020, to June 6, 2021, capturing daily infection rates and enabling the exploration of trends across different phases of the pandemic. The study involves conducting exploratory data analysis to understand the underlying distribution of COVID-19 cases and to select the most appropriate modeling technique. The fractional polynomial model is applied to fit the non-linear trends in the infection data, with the best fitting model identified through a closed test algorithm. Additionally, linear spline models are employed, dividing the data into segments and estimating the most accurate model by determining optimal knot points. Model selection criteria, including AIC and BIC, are used to evaluate the performance of each model. This study aims to provide a comprehensive analysis of COVID-19 infection trends in Kenya and to identify the most suitable model for forecasting and policymaking in pandemic response efforts.

2. Literature Review

2.1. Introduction

This chapter on literature review assess previous literature on the subject and places each work in the context of its contribution to understanding the research problem under consideration. However, COVID-19 is a relatively new phenomenon and as such the literature review will seek to synthesize information from previous studies on other pandemics in a bid to identify new ways of interpreting prior research as well as locate this research within the context of existing literature. For this reason, this chapter reviews COVID-19 incidence infections in Kenya as well as a review of the fractional polynomials and linear splines modeling which is very popular for modeling and summarizing the relationship between any biological aspects in correlation to its causative factors.

2.2. Review of COVID-19 Incidence Infections in Kenya

The government of Kenya started enforcing stringent measures on combating COVID-19 on March 13th immediately patient zero tested positive for the novel coronavirus. Since then, the number of cases to have tested positive and the number of deaths continue to increase gradually. As of May 22nd 2020, the Republic of Kenya had a total of 1,161 cases after testing a sample size of 55,074 according to [6]. Reference [6] also noted that by May 22nd 2020, they had discharged a total of 380 COVID-19 patients. Mapping the COVID-19 infections by counties, the Ministry of Health noticed that the most affected counties were Nairobi, Mombasa, Mandera, Kwale, and Kajiado. The country had also registered 50 fatalities at the aforementioned. The fatalities were from the confirmed cases of COVID-19. This means that from the tested and confirmed cases, the case fatality rate (CFR) stood at 4.3%. However, it is important to note that the country was not at its optimal position of testing for COVID-19 infections due to reduced testing capacity and human resources. This means that the infection fatality rate (IFR) could be higher than the CFR due to the possible discrepancies between the confirmed cases and potential infections that are yet to be confirmed due to some limitations. The COVID-19 tracking model used by the Kenyan Ministry of Health has been under review as it appears to have discrepancies in relation to models developed by other institutions. According to [7], as of May 5th 2020, the country had confirmed 490 COVID-19 cases after the PCR testing, which were only a small subset of infections. This is due to the fact that most of the cases were either asymptomatic or even mild. Reference [7] noted that the lack of testing kits and the technical personnel had inhibited Kenya's ability to test more individuals, which is largely to be blamed for the low numbers of confirmed cases posted against a relatively high number of projected and predicted infections in the country. By May 22nd 2020, the country had only tested 55,074 individuals for a period of more than two months which indicates a low testing rate considering that the country has a population of around 47 million persons. According to [7], the low testing levels and the high levels of asymptomatic cases, it becomes impossible to make inferences on the actual number of infections from the currently confirmed cases, which necessitates the need for the development of models that can paint the actual picture of the COVID-19 infections in Kenya. Reference [7] noted that a randomized test by a private organizer in Kibera, the largest slum in the country, confirmed 3 positive cases out of the 400 sample tested. The testing carried out at the end of April indicated that the actual current infections in the country should be around 79,000 devoid of the infection growth rate that exist in the country. Reference [7] also indicated that the IFR of COVID-19 stood at 0.66% of the infections, which is a projection of the country having at least 3700 COVID-19 infections by May 5th 2020. According to [8], the crude mortality ratio for COVID-19 is between 3-4% which means that in the worst case scenario the country will be recording 40 deaths for every 1,000 confirmed cases. Technically this appears to be true considering that the country has recorded 50 deaths out of the 1161 confirmed cases which translates to a crude mortality ratio of around 4.3%. However, [8] noted that the models being used by the government were unclear with the data sharing protocols being opaque. The government is yet to develop a way of modelling the pandemic. Reference [8] proposed the Susceptible, Exposed, Infectious, Recovered (SEIR) model in dealing with the pandemic. The model is essential in the development of interventions as well as enhancing the preparedness of the country in dealing with the pandemic.

2.3. Review of the Fractional Polynomials and Linear Splines Modeling

Either the fractional polynomials or linear splines can be used to model and summarize the relationship between the growth of any biological aspect (be it good or bad) in correlation with its causative factors. For instance, the two models can be used to summarize the correlation between age and height increase among children or the correlation between nutrition and death. According to [9], these models are "often used to describe trajectories, as they analyse repeated measures (level 1) clustered within individuals (level 2)"(pp.129). The two approaches are used to model a dependent factor against independent functions which are bound to be either linear or which are supposed to include polynomials. These models are used to develop the best fitting trajectory within the modeling framework with the fractional polynomials producing a smooth function and the linear splines generating a set of connected phases [9]. For the linear spline each phase will have a different growth rate although the connected phase will be within the best fitting trajectory. An example of how the two look like is shown in the figure below.

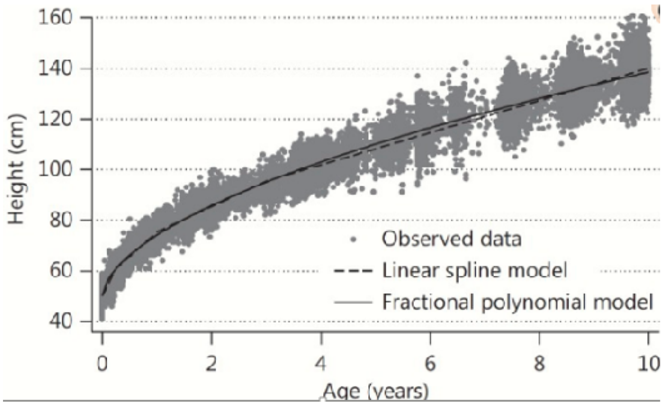

Figure 1. *Tilling, Macdonald-Wallis, Lawlor, Hughes & Howe, 2014.*

In their study on the relationship between obesity and mortality using BMI, [10] identified that the multivariate fractional polynomials are essential in addressing the asymmetric and nonlinear relationships between two biological aspects by "allowing the data itself to determine the functional form of the correlated factors and the other adjustment variables". Fractional polynomials have been identified to be a best fitting model due to its power transformation of the covariates. The fractional polynomial model is essential in creating symmetry in curves that are nonlinear and asymmetrical as a way of creating the best fit trajectory for two functional factors that are correlated [10]. Allowing the estimated trajectory curve to have a flexible shape is essential in ensuring that the fractional polynomial model is sensitive about the correlation between the dependent variable and the single continous covariate. Fractional polynomial being a flexible regression model provide a succint and accurate approximation of the relationship between the covariate and the response variable [11]. Whenever an analysis is being done using skewed data, there is a doubt on the linearity assumption, which necessitates the use of a logarithmic model that categorizes the continuous variables to enhance linearity [12]. In this case, the use of fractional polynomials comes in handy to enhance the model by the power transformation of continous variable. According to [13] there are two types of fractional polynomials FPI which has 8 models and FP2 which has 36 models. These two types of polynomial models have the ability to produce a range of curves which are essential in covering different continuous functions in health sciences.

In randomized controlled trials the adjustment of the prognostic covariates is essential in ensuring that the imbalances between the functional factors is eliminated, which leads to the development and specification of the nature of association between causative factor and the outcome [14]. The adjustment of the prognostic is essential in identifying whether the nature of association between the functional factors is logarithmic, quadratic, or linear. The adjustment of the continuous covariate becomes essential when the association between the outcome and the covariate is unknown whereby the fractional polynomials and the linear splines becomes very imperative [14]). The benefits of using the fractional polynomials is the ability to allow non-linear associations as well as keeping the data as continuous as possible. Fractional polynomials are crucial for modeling considering that it is available in a number of statistical packages. Using the linear splines is also equally important as it also keep data continuous and allow for non-linear associations. In randomized control trials, the linear splines are used by ensuring that knots are chosen and placed at specified percentiles of the data to ensure that a smoothened trajectory of the covariates and the outcome is developed according to [15]. According to [15] the application of the smoothing splines involves knots being placed at different data points though a penalized likelihood should be maximized to ensure that smoothed estimates and trajectory are derived. This makes linear splines an equally important model in epidemiological research. However, linear splines appear to be more detailed than the fractional splines considering that knots are inserted at every major data point in the estimated trajectory. According to [16], while both fractional polynomial and linear splines are essential in identifying the functional forms for continuous covariates and are similar in predicting performance, they differ to some extent. However, the fractional polynomial is better than the splines in terms of performance when the amount of information is medium, which means that the former is better in recovering simpler functions [17]. On the other hand, the linear splines are better in recovering the complex functions that contain large amounts of data with no localized structures and forming into explanatory models that be easily inference. Reference [18] noted " that spline modeling, while extremely flexible, can generate fitted curves with uninterpretable 'wiggles', particularly when automatic methods for choosing the smoothness are employed."

In their study, [19] applied multivariate linear regression model to provide an estimation of the H1N1 pandemic in different countries as a way of coming up with the mortality burden of the pandemic. By using these models, the study found out that the pandemic's mortality was tenfold higher than the WHO's laboratory-confirmed mortality. This means that the multivariate linear regression models of polynomial and splines are more accurate in determining the actual mortality of pandemics than any other model. Refence [19] used a model that included virology surveillance time series data and the linear secular trends so that the pandemic's mortality could be estimated. The use of the multivariate regression model that included the polynomials and the splines was essential in obtaining the missing data points, which explains why the studied countries were posting lower mortalities than the ones identified by this particular study. What this means is that the multivariate regression models are the best in modelling pandemics or other epidemiological diseases with high infections and mortality rates. Moreover, the modeling paints the correct picture of severity and mortality especially when a pandemic's data is continuous over a period of time.

3. Methodology

3.1. Introduction

This chapter introduces the proposed models for purposes of modeling and analyzing the data. The section introduces and justifies the suitability of fractional polynomials and linear splines. The chapter also outlines the model selection criteria where the best fitting model for the data is selected.

3.2. Fractional Polynomials

Proposed by Royston and Altman, fractional polynomials (FP) is a generalization of the polynomial function which provides a flexible parameterization of a continuous variable. The general fractional polynomial function of degree m is defined as;

$$\phi_m(X;\boldsymbol{\beta},\mathbf{p}) = \beta_0 + \sum_{i=1}^{m} \beta_i X^{(p_i)} \tag{1}$$

where m is a positive integer, $\mathbf{p}$ = $p_1 \leq p_2 \leq$..........$\leq p_m$ represents the real-valued vector of powers, $\beta_1, \beta_2, ..., \beta_m$ represents the real-valued coefficients. The round brackets notations denotes the Box-Tidwell transformation [12];

$$X^{(p_i)} = \begin{cases} X^{(p_i)}, & p_i \neq 0, \\ lnX, & p_i = 0 \end{cases}$$

Equation (1) can be extended to the case when $m > 1$. For m = 2 and $\mathbf{p}$ = (p_1, p_1);

$$\phi_m(X;\boldsymbol{\beta},\mathbf{p}) = \beta_0 + (\beta_1 + \beta_2)X^{(p_1)} \tag{2}$$

which is a fractional polynomial of degree 1. However, if $p_1 = p_2$, the second-degree fractional polynomial transformation is defined as;

$$X^{\mathbf{p}} = \begin{cases} X^{p_1}, X^{p_2} & p_1 \neq p_2 \\ X^{p_1}, X^{p_1} lnX & p_1 = p_2 \end{cases}$$

The functional form of the second power, for the repeated powers is as a result of the limit as p_2 tends to p_1 for

$$X^{p_1}(X^{p_2 - p_1})(p_2 - p_1)^{-1} \tag{3}$$

Such that FP2 model with repeated powers p_1 = p_2 can be defined as;

$$\phi_m(X;\boldsymbol{\beta},\mathbf{p}) = \beta_0 + \beta_1 X^{(p_1)} + \beta_2 X^{(p_1)} lnX \tag{4}$$

for $m > 2$ and $p_1 \leq p_2 \leq$..........$\leq p_m$, (4) can be generalised as follows;

$$\phi_m(X;\boldsymbol{\beta},\mathbf{p}) = \beta_0 + \beta_1 X^{(p_1)} + \sum_{i=2}^{m} \beta_i X_i^{(p_i)} (lnX_i)^{-1} \quad (5)$$

Given arbitrary powers $p_1 \leq p_2 \leq$**.........**$\leq p_m$, and setting $H_0(X)$ = 1 and p_0 = 0, we obtain the extended definition of the fractional polynomial function;

$$\phi_m(X;\boldsymbol{\beta},\mathbf{p}) = \sum_{i=0}^{m} \beta_i H_i(X) \quad (6)$$

where i = 1, 2,........,*m* and the recurrence relation $H_i(X)$ is defined;

$$H_i(X) = \begin{cases} X_i^{(p)}, & p_i \neq p_{i-1}, \\ H_i - (X)lnX, & p_i = p_{i-1} \end{cases}$$

The first degree fractional polynomial, FP1 (*m* = 1) performs eight tests with the power transformations given in the predefined set S = {-2, -1, -0.5, 0, 0.5, 1, 2, 3} where 0 represents the log transformation [11]. In FPI, *p* is a single power. As for a second-degree fractional polynomial, the model is fitted to each possible pair of powers from the set of powers, S. FP2 model fits 36 models.

In fractional polynomials, duplication of powers does not reduce the degree of the model. The process of fitting a polynomial of degree one involves taking each power from the set to determine whether the fit of the model is improved by the transformations. From (4) the second term is multiplied the log since we will have two separate terms with the same powers. In fractional polynomials, the covariate X must be greater than zero to ensure that the interpretation of the intercept is meaningful.

Model selection criteria

Combination of different pairs of the polynomials will yield different model deviances which is examined to determine the best-fitting polynomial. The best fitting fractional polynomial is the model with the smallest deviance. However, in order to increase parsimony and stability of the selected models, [11] proposed the use of a closed test algorithm for model selection of fixed *m* and a level of significance, α. The procedure follows the steps:

1. Inclusion - For a covariate X, the best fitting second-degree fractional polynomial is compared with the null model using χ^2 with 4 d.f. at the α level. If the result is statistically significant then continue, otherwise drop the covariate X from the model.
2. Non-linearity - Test the best fitting second-degree fractional polynomial in comparison with the linear model using a χ^2 with 3 d.f. at the α level. If the test is significant then continue, otherwise we choose the linear model.
3. Simplification - Test the best fitting second-degree FP against the best fitting first-degree FP with 2 degrees of freedom. If the test is significant, second-degree FP is the best fitting model, otherwise the best fitting model is first-degree fractional polynomial.

3.3. Linear Splines

Linear splines also known as piecewise model or broken stick is an alternative approach in modelling COVID-19 infections by using a series of joined knot points. The method involves partitioning the data into different segments and fitting a linear spline at each segment where the corresponding coefficients of each spline describes the average slope between each knot point [9]. Linear splines models yields interpretable coefficients compared to the fractional polynomials model. The linear spline model with knots at t_k, k = 1,2,.....,K continous at each knot point will be of the form;

$$\phi(x) = \beta_0 + \beta_1 x + \sum_{k=1}^{K} \mu_k (x - t_k)_+ \quad (7)$$

where β_k denotes the weight of each linear segment and $(x - t_k)_+$ refers to the kth linear function with a knot at t_k [20]. The positive part of the function is defined as;

$$(x - t_k)_+ = \begin{cases} x - t_k, & x > t_k, \\ 0, & otherwise \end{cases}$$

The linear spline model can be extended to be a piecewise polynomial of *p*;

$$\phi(x) = \beta_0 + \beta_1 x + ... + \beta_p x^p + \sum_{k=1}^{K} \mu_k (x - t_k)_+^p \quad (8)$$

The basis function for the linear spine model is defined as;

$$[1, x, (x - \mathbf{t}_1)_+ + \ldots + (x - \mathbf{t}_K)_+]$$

Combining equation (3.8) with the basis function, the linear spline model can be written as follows;

$$\phi(x) = \beta_0 + \beta_1 x + \beta_2 (x - t_1)_+ + \beta_3 (x - t_2)_+ + ... \quad (9)$$

Determination of the number of knot points The number and position of the knot points is done by either placing the knots at the center (average) of the distribution or starting with a large number of knots then gradually reduce the number until a smooth curve is achieved [21]. However, increasing the number of knots points beyond the optimal point will lead to over fitting. In our study, the knot points were determined using AIC and the BIC to determine the optimal number of knot points [22]. A low value of AIC or BIC corresponds to the optimal number of knot points.

$$AIC = -2ln(L) + 2K \quad (10)$$

Where;
AIC - Akaike Information Criterion
L - the maximum likelihood
K - the number of parameters in the model

$$BIC = -2ln(L) + kln(n) \tag{11}$$

Where;
BIC - Bayesian Information Criterion
L - the maximum likelihood
n - number of observations
k - the number of parameters in the model
The best fitting fractional polynomial will be compared with the best fitting linear spline using the AIC.

4. Data Analysis and Results

4.1. Data

In this study, we used daily COVID-19 infections data from the period March 13, 2020 to June 6, 2021 (N = 450). The data was obtained from the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. As of June 13, 2021, there were 175,337 total cases with 3,410 total deaths and 120,208 total recoveries. R software was used for data analysis by loading several packages such as mfp, ggplot2, gridExtra, tidyverse, gam among others to help in statistical model building and plotting.

4.2. Exploratory Data Analysis

4.2.1. Description of COVID-19 Infections Data

Table 1. *Descriptive Statistics.*

	Daily infections
Length (N)	450
Minimum	0
Maximum	2,008
Mean	382.9
1st Quartile	105.0
Median	246.0
3rd Quartile	560.5

Table 1 presents the summary statistics for the daily COVID-19 infections, the average daily infections in Kenya for the past 15 months (450 days) is 382.9 with a minimum value of 0 cases and a maximum value of 2,008 cases. The median COVID-19 infections is 246.0 cases. The lower (1st) and upper (3rd) quartiles are 105.0 and 560.5 respectively.

4.2.2. Scatter Plot for the Number of Infections

Figure 2 shows the distribution of the daily infections in Kenya. It is observed that there is an increasing trend in number of new infections from the day the first case was reported with a peak experienced around mid August, 2020. There was a significant drop in the number of daily cases reported in September. In November, 2020, there was a significant upward trend in the number of cases with a decline reported in January 2021. The third wave is observed in March, 2021 characterized by a sharp rise in the number of daily infections.

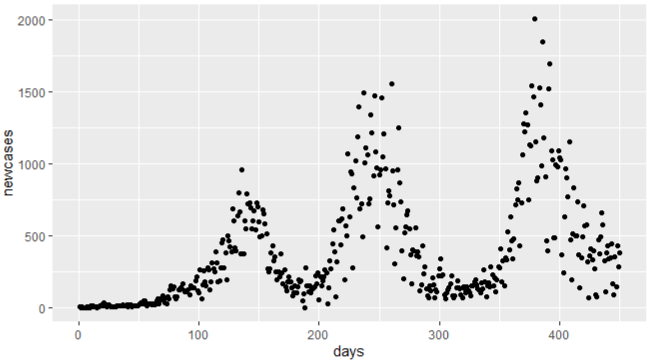

Figure 2. *Scatter plot for the raw data of COVID-19 Infections.*

4.3. Fractional Polynomial Model

The first method to modelling COVID-19 daily infections is the fractional polynomial model which involves power transformation taking powers from the predefined subset of powers S = {-2, -1, -0.5, 0, 0.5, 1, 2, 3}. Table 2 shows the summary of the first degree fractional polynomials (FP1);

Table 2. *Summary for FP1 model.*

Coefficient	**Estimate**	**Std. Error**	**t value**	**Pr($>\mid t\mid$)**
Intercept	-128.07	48.07	-2.664	0.008
I((days/100)$^{0.5}$)	360.75	32.01	11.269	$< 2\times10^{-16}$

From Table 2 above, the power of FP1 is 0.5. The scaled continuous predictor variable (days) has one significant term $I(days/100)^1)$ with the scaling parameter taken as 100. The FP1 model is of the form;

$$FP_1 = -128.07 + 36.075 \times days^{0.5} \tag{12}$$

Next, the second degree fractional polynomial (FP2) is of power; *p* = (1, 2).

The summary for FP2 is shown in Table 3;

Table 3. *Summary for FP2 model.*

Coefficient	**Estimate**	**Std. Error**	**t value**	**Pr($>\mid t\mid$)**
Intercept	-81.02	47.68	-1.699	0.0899
I(days/100)^1)	355.20	48.82	7.275	$< 1.56\times10^{-}12$
I((days/100)^2)	-49.76	10.48	-4.747	$< 2.79\times10^{-6}$

The summary for the second degree fractional polynomial shows that the scaled continuous predictor variable (days) have two significant terms $I(days/100)^1)$ and $I(days/100)^2)$. The FP2 model is of the form;

$$FP_2 = -81.02 + 3.552 \times days^1 - 0.004976 \times days^2 \tag{13}$$

Figure 3 shows different polynomial models.

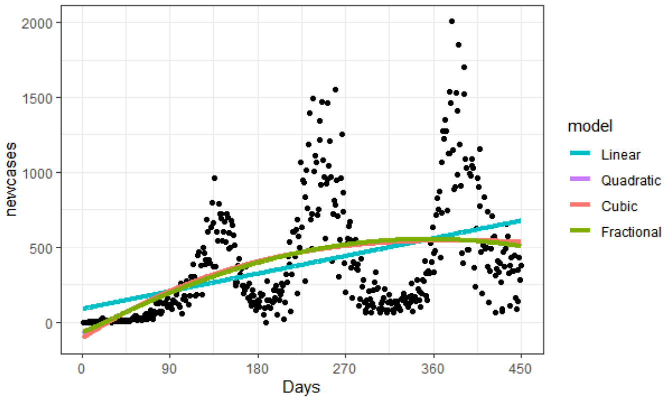

Figure 3. *FP1 and FP2.*

Figure 4 shows that FP2 provides a better fit for the daily COVID-19 infections since it yields a smoother curve compared to the FP1 curve.

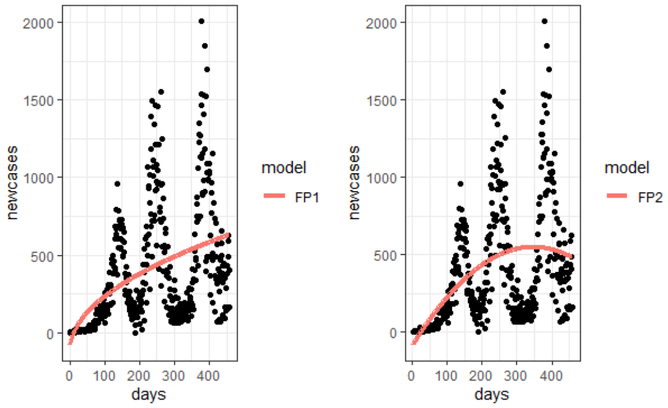

Figure 4. *FP1 and FP2.*

Estimating the best fitting fractional polynomial

The best fitting fractional polynomial is obtained using the closed test algorithm which combines variable transformation and model selection. Table 4 provides a step-by-step illustration of the model selection process with a significance level; α = 0.05. Since the three steps are statistically significant with p-values less than 0.05, the best fitting model for predicting the number of daily COVID-19 infections is the second-degree (FP2) with powers (1, 2). The coefficients of the final model are; β_0 = -81.02, β_1 = 3.552 and β_2 = -0.004976.

Table 4. *Application of the closed test algorithm.*

Model	Deviance	Powers	Step	Comparison	*p*-value
FP2	50355608	1, 2	1	FP2 vs Null	1.86×10^{-27} (df = 4)
FP1	51331225	0.5	2	FP2 vs Linear	8.29×10^{-5} (df = 3)
linear	52893900	1	3	FP2 vs FP1	0.0161 (df = 2)
Null	65881646				

4.4. Linear Spline Model

After examining the best fitting fractional polynomial, we fit an alternative model to fit the daily COVID-19 infections. Linear spline model fits a curve at different segments also known as knots, polynomial regression coefficients are then obtained at each knot. The first step is to determine the optimal knot points using AIC and BIC.

Determination of the number of knot points

The optimal number of knots is 19 because it has a lower AIC and BIC as shown in table 5. Different linear spline models with different number of knot points are shown in Figure 5.

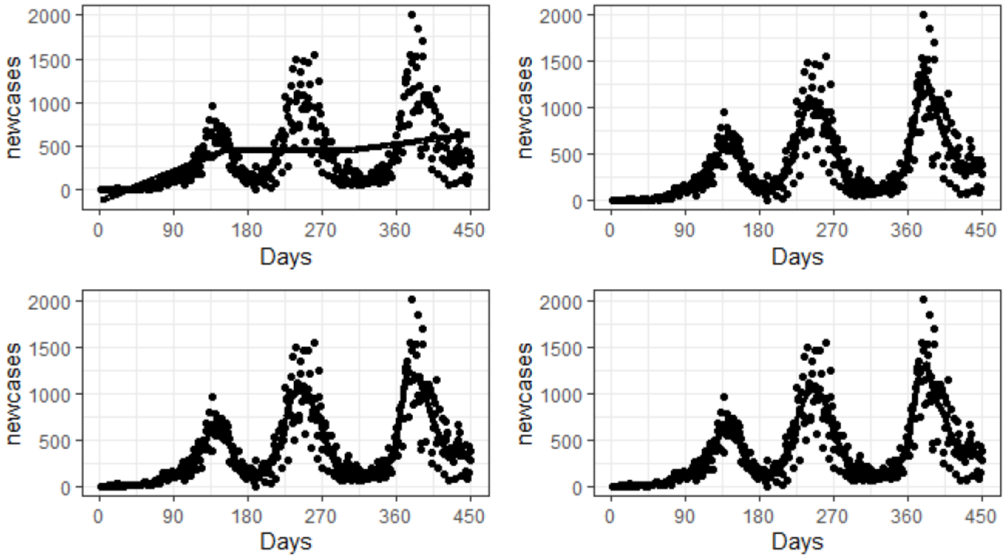

Figure 5. *Linear spline models with 3 segments (top left), 19 segments (top right), 29 segments (bottom left) and 40 segments (bottom right).*

Table 5. *AIC and BIC for linear spline models with different number of knots.*

Number of knots	AIC	BIC
40	5972	6145
35	5965	6116
29	5962	6089
22	5972	6071
20	5962	6052
19	5935	6022
15	6024	6094
11	6104	6158
9	6155	6201
3	6512	6533

4.5. Comparison Between Fractional Polynomials and Linear Splines

After determining the best fitting fractional polynomial and the linear spline with different knot point, it is important to evaluate the best fitting model for the daily COVID-19 infections. We used AIC to determine the best model. AIC values for the two models are given in table 6;

Table 6. *AIC for best fitting fractional polynomial and linear spline model.*

Model	AIC
Fractional polynomial (FP2)	6516
Linear spline	5935

Linear spline model with 19 knot points has a lower AIC = 5935 compared to the second degree fractional polynomial which indicates that the linear spline model provides a better fit for the daily COVID-19 infections.

4.6. Discussion

In this study, we modelled the trend of the daily COVID-19 cases in Kenya. The number of daily COVID-19 cases depicted a non-linear growth trajectory depicting the different stages of transmissions. The initial stages of the pandemic was as a result of imported cases which was followed by local transmission of the disease among people who came into contact with the first patients. The cases continued to increase rapidly as a result of community transmission with the government acting with speed to implement mitigation measures that were aimed at curbing further spread of the virus.The aim of this study to develop the most accurate model for fitting the number of cases in Kenya. Fractional polynomials and linear spline models have the ability to capture the non-linear trend. The two models were applied to the COVID-19 data in Kenya with the help of R software to determine the best fitting model for the data.

Further, fractional polynomials of different degrees were fitted to the data with the closed test algorithm being applied to determine the best fitting fractional polynomial. Many linear spline models with different number of knots were also applied to the COVID-19 data and AIC and BIC were used to determine the most accurate model. That is, the linear spline model with minimum value of AIC and BIC provides the most accurate fit for the data.

Among the fractional polynomials fitted, the second-degree fractional polynomial with powers (1, 2) provided the best fitting model. The finding is in agreement with [10] who identified that fractional polynomials are essential in addressing the asymmetric and non-linear relationships by allowing the data itself to determine the functional form of the covariates. On the other hand, a linear spline with an optimal number of knot points (knots = 19) provided the most accurate fit for the COVID-19 data. These findings concur with the results of [15] which noted that linear splines ensured a smoothened trajectory of the covariates by ensuring that the knots are selected and placed at specified data percentiles.

Moreover, comparing the two models for fitting the number of cases the best fitting models were compared using AIC. The best fitting model for the data has the lowest AIC; the linear spline has the lowest AIC. Thus, the model provides a more accurate fit for the daily COVID-19 infections in Kenya.

5. Conclusion and Recommendation

5.1. Conclusion

With the emergence of the COVID-19 pandemic globally, there is an urgent need for development of statistical models for accurate forecasting of the trend pattern of the virus. Despite the fact that the pandemic is still ongoing, models for accurate forecasting using the available data are pertinent in facilitating the medical stakeholders and governments to develop strategies for decision making and resource management aimed at curbing further spread among communities. This study employed fractional polynomial and linear spline models for modelling the number of daily COVID-19 cases. Based on the daily COVID-19 data in Kenya, the proposed methods captured the non-linearity trend exhibited by the number of daily cases. The fractional polynomial method fitted models of different degrees allowing the data to determine the power of the covariate. The linear spline approach partitioned the data into segments and fitted a linear function at each segment.

This analysis showed that second-degree fractional polynomial provided an accurate fit for the data compared to the first-degree fractional polynomial model. However, the linear spline model with 19 knots outperformed the fractional polynomial models which had the lowest value of AIC and BIC. Therefore, the linear spline model is proposed for fitting the number of daily COVID-19 cases.

5.2. Recommendation

This study relied solely on the COVID-19 data from Kenya, which has recorded relatively lower number of daily cases. Similarly, the study involved a single covariate whereas there are other factors that have a positive effect on the number of daily cases. The proposed model might need further examination to incorporate new covariates as well as comparison with other countries. This study recommends modelling for other countries especially those that were badly affected by the pandemic as well as countries within the East Africa region. This will provide an insight on the efficiency of different countries in managing the spread of the virus based on the different mitigation strategies adopted such as quarantine and social distance.

Abbreviations

COVID-19	Coronavirus Disease
WHO	World Health Organization
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
CFR	Case Fatality Rate
BMI	Body Mass Index
SARS-CoV-2	Severe Acute Respiratory Syndrome
MERS-CoV	Middle East Respiratory Syndrome Corona Virus
FP	Fractional Polynomial
IFR	Infection Fatality Rate
SEIR	Susceptible Exposed Infectious Recovered Model
H1N1	Swine Flu
CSSE	Center for Systems Science and Engineering
FP1	First Degree Fractional Polynomial
FP2	Second Degree Fractional Polynomial

Acknowledgments

I thank the Almighty God, without whose grace all these would not be possible. My sincere gratitude goes to all those who participated in one way or the other for the accomplishment of my research project. I would like to take this opportunity to express my special gratitude to my supervisors Prof. Mwalili and Dr. Wanjoya for their valuable knowledge and constructive advice that made this work a success. Finally, my appreciation extends to my fellow students, friends and relatives for their moral support and prayers that made the journey worthwhile.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Xu, B., Gutierrez, B., Mekaru, S. et al. (2020) "Epidemiological data from the COVID-19 outbreak, real-time case information", *Sci Data, 7(106)* https://doi.org/10.1038/s41597-020-0448-0

[2] Read, J. M. et al. (2020). Novel coronavirus 2019-nCoV: early estimation of epidemiological parameters and epidemic predictions. *MedRxiv*, pp. 1-11.

[3] Gilbert, M. et al. (2020). Preparedness and vulnerability of African countries against importations of COVID-19: a modelling study. *The Lancet, 395(10227)*, pp. 871-877.

[4] Anastassopoulou, C., Russo, L., Tsakris, A., Siettos, C. (2020). *Data-based analysis, modeling and forecasting of the COVID-19 outbreak*, PLoS ONE, 15(3), e0230405

[5] Keeling, M. J. & Rohani, P. (2008).*Modeling Infectious Diseases in Humans and Animals*, Princeton University Press.

[6] Ministry of Health. (2020). Kenya COVID-19 cases hits 1,161 Nairobi , Friday May 22, 2020 - MINISTRY OF HEALTH. Retrieved 24 May 2020, from *https://www.health.go.ke/kenya-covid-19-cases-hits-1161-nairobi-friday-may-22-2020/*

[7] Winn, H. (2020). How Kenya is tracking against our COVID-19 model. Retrieved 24 May 2020, from https://rescue.co/update-5th-may-how-kenya-is-tracking-against-our-covid-19-model/

[8] Nanyingi, M. (2020). Predicting COVID-19: what applying a model in Kenya would look like. Retrieved 25 May 2020, from https://theconversation.com/predicting-covid-19-what-applying-a-model-in-kenya-would-look-like-134675

[9] Tilling, K., Macdonald-Wallis, C., Lawlor, D., Hughes, R., & Howe, L. (2014). Modelling Childhood Growth Using Fractional Polynomials and Linear Splines. *Annals Of Nutrition And Metabolism, 65(2-3)*, 129-138. https://doi.org/10.1159/000362695

[10] Wong, E., Wang, B., Garrison, L., Alfonso-Cristancho, R., Flum, D., Arterburn, D., & Sullivan, S. (2011). Examining the BMI-mortality relationship using fractional polynomials. *BMC Medical Research Methodology, 11(1)*. https://doi.org/10.1186/1471-2288-11-175

[11] Royston, P. (2017). Model selection for univariable fractional polynomials. *The Stata Journal, 17(3)*, 619-629.

[12] Baneshi, M., Nakhaee, F., & Law, M. (2020). On the Use of Fractional Polynomial Models to Assess Preventive Aspect of Variables: An Example in Prevention of Mortality Following HIV Infection. *International Journal Of Preventive Medicine*, 4(4), 414-419.

[13] Duong, H & Volding, D. (2014). Modelling continuous risk variables: Introduction to fractional polynomial regression. *Vietnam Journal of Science, (1)*. 1-5.

[14] Kahan, B., Rushton, H., Morris, T., & Daniel, R. (2016). A comparison of methods to adjust for continuous covariates in the analysis of randomised trials. *BMC Medical Research Methodology, 16(1).* https://doi.org/10.1186/s12874-016-0141-3

[15] Harrell Jr FE. (2001). Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. New York: Springer.

[16] Binder, H., Sauerbrei, W., & Royston, P. (2012). Comparison between splines and fractional polynomials for multivariable model building with continuous covariates: a simulation study with continuous response. *Statistics In Medicine*, 32(13), 2262-2277. https://doi.org/10.1002/sim.5639

[17] Sauerbrei, W., Royston, P., & Binder, H. (2007). Selection of important variables and determination of functional form for continuous predictors in multivariable model building. *Statistics In Medicine, 26(30)*, 5512-5528. https://doi.org/10.1002/sim.3148

[18] Simonsen, L., Spreeuwenberg, P., Lustig, R., Taylor, R., Fleming, D., & Kroneman, M. et al. (2013). Global Mortality Estimates for the 2009 Influenza Pandemic from the GLaMOR Project: A Modeling Study. Plos *Medicine, 10(11)*, e1001558. https://doi.org/10.1371/journal.pmed.1001558

[19] Royston, P., Ambler, G., & Sauerbrei, W. (1999). The use of fractional polynomials to model continuous risk variables in epidemiology. *International journal of epidemiology, 28(5)*, 964-974.

[20] Hansen, M. H., Huang, J., Kooperberg, C., Stone, C. J., & Truong, Y. K. (2001). *Statistical Modeling with Spline Functions Methodology and Theory*.

[21] Wang, Y. (2011). *Smoothing splines: methods and applications.* CRC Press.

[22] Likhachev, D. V. (2017). *Selecting the right number of knots for B-spline parameterization of the dielectric functions in spectroscopic ellipsometry data analysis*. Thin Solid Films, 636, 519-526.