American Journal of Theoretical and Applied Statistics
2021; 10(1): 22-31
http://www.sciencepublishinggroup.com/j/ajtas
doi: 10.11648/j.ajtas.20211001.14
ISSN: 2326-8999 (Print); ISSN: 2326-9006 (Online)

Estimating Population Total Using Machine Learning Logistic Regression: COVID-19 Pandemic Challenges Perspective

Thomas Mageto

Department of Statistics and Actuarial Sciences, Jomo Kenyatta University of Agriculture and Technology, Juja, Kenya

Email address:
tttmageto@gmail.com

To cite this article:
Thomas Mageto. Estimating Population Total Using Machine Learning Logistic Regression: COVID-19 Pandemic Challenges Perspective. *American Journal of Theoretical and Applied Statistics*. Vol. 10, No. 1, 2021, pp. 22-31. doi: 10.11648/j.ajtas.20211001.14

Received: January 8, 2021; **Accepted**: January 15, 2021; **Published**: January 22, 2021

Abstract: The estimation of the population total in undeveloped and developing countries in the recent past has attracted a lot of interest to many researchers due to the sole purpose of planning resource allocation, personnel training and infrastructure in social, health, transport, communication and education. The comprehensive census survey in many countries are conducted every ten years but the government administration changes in many counties every four to five years due to the limit of government terms as per the constitution and therefore does not coincide with the time of census survey. Further, due to the emerging COVID-19 pandemic challenges that requires ministry of health protocols of social distance, the census survey in which the methods of questionnaire and personal interview are commonly used need to be avoided and therefore there is need to search for a better and reliable estimating models for estimating the population total which is the main focus of the study. The existing and developed methods of exponential and logistic class of population total estimating modes have been considered and compared. The main problem in the logistic models in estimating the population total is the estimation of the highest possible population that can be attained for each of the administrative units. In this study a machine learning logistic regression has been proposed and incorporated to search and estimate the constant using the supervised learning process. The performance of the methods have been compared using the Root Mean Square Error (RMSE) whose values were recorded as 1.062, 1.524, 0.477, 0.819 and 0.286 for the exponential, logistic I, Logistic II, logistic III and machine learning logistic (logistic IV) in which the proposed model performed better with the least square error value of 0.286. The proposed model was then used to project the population total and projected the population total for all regions as 51.00, 55.02, 62.50, 69.10, 74.65 and 79.14 in millions in the years 2024, 2029, 2039, 2049, 2059 and 2069 respectively.

Keywords: Population Total Estimates, Growth, COVID-19, Logistic Regression and Projection

1. Introduction

The determination and estimation of population total in undeveloped and developing countries is important for governments to be able to plan for development projects as promised and documented in the manifesto during national elections campaigns. The requirement for the population census for development as set out by the United Nations is to have a detailed population and housing census every ten years [1]. In Kenya for instance the elections are conducted every five years while the Kenya Population and Housing Census is conducted every ten years, this therefore means that estimating the statistics that includes the population total for this purpose is a priority in departments of Economic Planning and National Bureau of Statistics due to the key role that the statistics from the census survey plays in resource allocation for economic growth of countries [2, 3]. Since most of the data in census survey are collected using questionnaires and personal interviews, this pose danger of COVID-19 transmission as keeping the social distance and avoiding sharing of census survey materials will be hard to achieve. The development of better strategy and models for estimating the population total will go a long way in boosting the confidence of governments in obtaining better, reliable, cheaper and timely statistics especially during the time of COVID-19 pandemic as collecting such data is limited due to

the health protocols as advised by ministry of health officials and experts.

1.1. Statement of Problem

In planning for development, the census survey need to be carried to obtain the necessary statistics as a basis for resource allocation for the government projects. The collection of such statistics is collected after every ten years as it is expensive for most governments to conduct the census survey in shorter durations of time. In order to have reliable statistics especially when there is a change of government there is need to have timely and reliable data for the planning departments and this can only be achieved through estimates as conducting the census survey requires capital, equipment, well trained personnel and approvals which is only feasible and affordable after a longer duration of time in order to give amble time to source for the funds from partners, train personnel and procure equipment for the census survey. In the recent past an attempt has been made to achieve this requirement but the problem of estimating the optimum population that can be attained (constant k) in the logistic models has been a great challenge as the value is invalid in many populations. In this study therefore the focus will be to develop machine learning logistic regression technique that will search better strategy, estimate parameters and determine the projected population totals for the Kenya administrative regions.

1.2. Objectives

1.2.1. General Objective

The general objective of the research project is to develop machine learning logistic nonlinear regression model for estimating the interpolation and projected Kenya population total for the various administrative regions for development.

1.2.2. Specific Objectives

(i) Study the characteristics of the Kenya population using the population pyramid
(ii) Study the characteristics of the Kenya population using the population lifetables model.
(iii) Develop machine learning nonlinear logistic regression model for estimating the population total for various regions in Kenya.
(iv) Determine the parameters of the proposed logistic regression model.
(v) Compare the performance of the developed logistic regression model with other already existing logistic regression models.
(vi) Determine the interpolation and projected population total for existing administrative units using the proposed developed logistic regression model.

1.3. Hypothesis

The hypothesis in this research project is important as it will be used to guide in determining whether the estimating population total model is a better model compared to other models in the same class that is simply given and stated as machine learning logistic regression model is better nonlinear model in estimating population total. However the heuristic approach to the measure of performance of the model is to consider the Root Mean Square Error (RMSE) measure due to the large number of models being considered.

1.4. Justification of the Study

The research project is a crucial and important study due to the contribution it makes in the development as the estimation of population total for planning in undeveloped and developing countries taking into consideration the COVID-19 pandemic challenges as the census surveys and other forms of surveys are not feasible and having reliable estimation models will fill the emerging gap as the government projects need to be budgeted and allocated resources and require the statistics for fair distribution of the available resources.

2. Literature Review

An attempt has been made by various researchers in modeling population total for instance investigation on population growth using exponential and hyperbolic modeling for the world population and estimated that the population is estimated to reach 100 billion in the year 2172 [4]. A study conducted by Gotelli has considered the geometric and exponential models and further considered incorporating the deaths and births [5]. Another study conducted by Kabareh et al considered the estimation of the population total using the birth and death process [6]. However, due to the unavailability of births and deaths data in most of the administrative units due to the beliefs, religion and nonresponse on the vital statistics, the estimation using such information is possible in developed countries where the records are available and complete.

In estimating the Kenya population total for the year 2019, the study by Kabareh et al estimated the population total using the Lagrange polynomial and estimated the value of forty eight thousand five hundred thirty three thousand five hundred eighty seven that is close but not good enough to the actual census survey value of forty seven thousand five hundred sixty for thousand two hundred ninety six [7]. The study on population total has also been investigated by Kabareh et al in which the piecewise polynomial approximation to the Newton backward difference polynomial approximation of the finite population total considered and results compared [8]. The results revealed that the piecewise polynomial is a better predicting model than the Newton backward polynomial. The logistic model has been considered in bounded population by Kabareh et al in which they considered the bounded population total using linear regression in the presence of auxiliary information [9, 10]. The results and conclusions from the past studies leaves the problem an open area for further investigation and in this study the investigation is intended to explore and develop a better model for estimating the population total.

3. Methodology

The methodology for the study problem considers the existing models for the estimation of population total that include the exponential and class of logistics models are considered, derived and discussed. The proposed machine learning logistic regression model will be developed and estimation of the parameters discussed.

3.1. Exponential Model

In the exponential model also referred to as logistic model by some authors makes the assumption that the population grows at a rate that is proportional to the population size, that is, in each unit of time, a certain percentage of the individuals produce new individuals. If the reproduction takes place more or less continuously, then this growth is represented by $\frac{dp}{dt} = rp$ where p is the population as a function of t, and r is the proportionality constant [11]. The differential equation for the logistic model can be solved by separate the variables and integrate both sides such that

$$\int_{-\infty}^{\infty} \frac{dp}{p} = \int_{-\infty}^{\infty} rdt \text{ then}$$

$$\ln p = rt + c \text{ such that } P_t = e^{rt + c} \text{ or}$$

$$P_t = c_1 e^{rt}$$

$$\text{Let } t = 0 \text{ then } p_0 = c_1 \text{ such that}$$

$$P_t = P_0 e^{rt} \tag{1}$$

3.2. Logistic I model

The logistic I model was originally developed by Verhulst and later studied by Pearl, R. and Read, L [12]. The curve in its simplest form takes the form given as

$$\hat{P}_t = \frac{k}{1+10^{a + bt}} \tag{2}$$

Where

$\hat{P}_t$ - Estimated population in some inter-censual year t.

$$\hat{k} = \frac{2y_0y_1y_2 - y_1^2(y_0 + y_2)}{y_0y_2 - y_1^2}, \hat{a} = \log\frac{k - y_0}{y_0} \text{ and } \hat{b} = \frac{1}{n}\log\frac{y_0(k - y_1)}{y_1(k - y_0)}$$

Such that t is the year for which the value has to be interpolated, y_0 is geometric mean of the first three years of the series, y_1 is geometric mean of the middle three years of the series, y_2 is geometric mean of the last three years of the series and n is the duration between successive censuses.

3.3. Logistic II Method

The logistic II method is a derivation from logistic I method in which we take into consideration the case of populations N_1, N_2, N_3 recorded at equidistance times of t_1, t_2 and $2t_2 - t_1$. Further let us now take $N_1 = y_o$, $N_2 = y_1$ and $N_3 = y_2$ such that

$$\hat{k} = \frac{2y_0y_1y_2 - y_1^2(y_0 + y_2)}{y_0y_2 - y_1^2} = \frac{2N_0N_1N_2 - N_1^2(N_0 + N_2)}{N_0N_2 - N_1^2}$$

$$= \frac{-N_0N_1^2N_2\left\{\frac{1}{N_0} + \frac{1}{N_2} - \frac{2}{N_1}\right\}}{-N_0N_2N_1^2\left\{\frac{1}{N_0N_2} - \frac{1}{N_1^2}\right\}} = \frac{\left\{\frac{1}{N_0} + \frac{1}{N_2} - \frac{2}{N_1}\right\}}{\left\{\frac{1}{N_0N_2} - \frac{1}{N_1^2}\right\}}$$

The constant a is the y-intercept in the plot of the graph of $\log(k/p - 1)$ against t, that is at $t = 0$

$$a = \log\left\{\frac{k - p_0}{p_0}\right\}$$

In determining the constant b, plot the graph of $\log(k/p - 1)$ against t and the value of b will be the gradient/slope

$$b = \frac{\log\left(\frac{k}{p_1} - 1\right) - \log\left(\frac{k}{p_0} - 1\right)}{t} = \frac{1}{t}\log\left\{\left(\frac{k}{p_1} - 1\right) / \left(\frac{k}{p_0} - 1\right)\right\} = \frac{1}{t}\log\left\{\frac{p_0(k - p_1)}{p_1(k - p_0)}\right\}$$

If t_h is the time it takes to reach half-way the maximum population, then

$$\frac{1}{2}k = \frac{k}{1+10^{a + bt_h}} \text{ such that}$$

$$a + bt_h = 0 \text{ or } a = -bt_h \text{ then}$$

$$\hat{P}_t = \frac{k}{1+10^{b(t - t_0)}} \tag{3}$$

3.4. Logistic III Model

In logistic III model, consider the logistic form in logistic II. Now if we let $y = 10$ such that $\ln y = \ln 10$ then $y = e^{\ln 10}$ [13]. The logistic equation can now be written in the form

$$\hat{P}_t = \frac{k}{1+e^{b\ln 10\,(t - t_0)}} \text{ but } b = \frac{1}{t}\log\left\{\frac{p_0(k - p_1)}{p_1(k - p_0)}\right\} = -\ln 10\frac{1}{t}\log\left\{\frac{p_1(k - p_0)}{p_0(k - p_1)}\right\}$$

Since b is the rate of growth, then the logistic III method take the form

$$\hat{P}_t = \frac{k}{1+e^{-r(t - t_0)}} \tag{4}$$

where r is the growth rate.

3.5. Machine Learning Logistic Regression (Logistic IV) Model

In the presented and discussed models of logistic I, Logistic II and logistic III the problem has always been the determination of the highest population that can be attained. In quite a number of populations the value is invalid or null and therefore with such an invalid value in the logistic equation, the estimation of the population total is not possible. In this study we have proposed and developed the machine learning logistic regression method in which the search of the constant, estimation of parameters and determination of the population totals is carried out using an algorithm. The development of the algorithm is as illustrated and depicted in Figure 1.

The steps depicted in Figure 1 are translated and written into a complete step-by-step procedure commonly referred to as algorithm as written in Algorithm I below [14]. The steps developed for computing the highest population attained, population estimates and projections of populations that are

unambiguous in a finite number of steps represent the proposed machine learning logistic regression model that finally is written as a program code in R or MATLAB that is used to compute for the parameters and projection of the population total estimates for the administrative units.

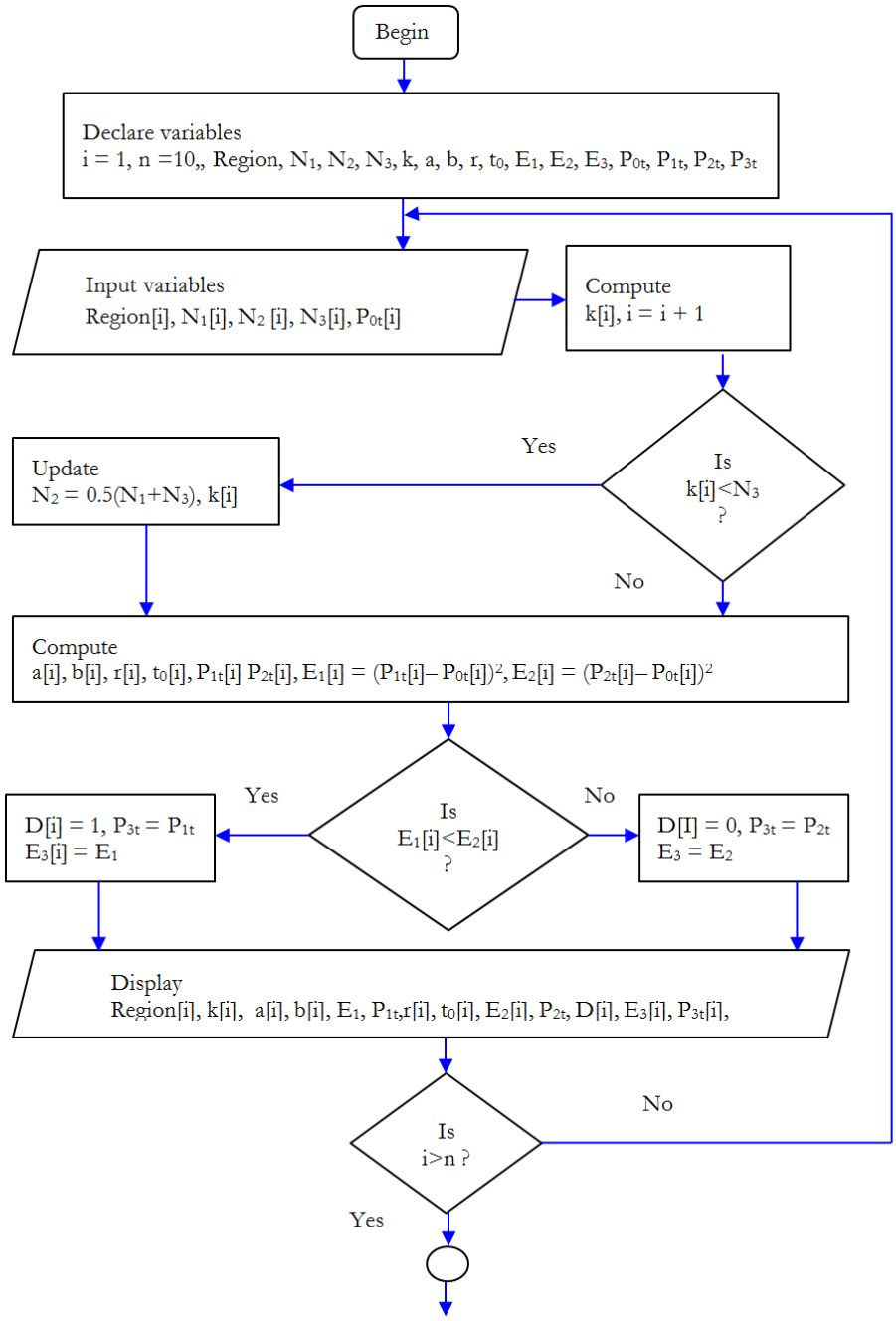

Figure 1. *Machine learning logistic regression estimation of parameters.*

Algorithm 1: Machine learning logistic regression estimation of parameters

1. Begin
2. Declare variables i = 1, n, Region, N_1, N_2, N_3, k, a, b, r, t_h, E_1, E_2, E_3, P_{0t}, P_{1t}, P_{2t}, P_{3t}
3. Input Region, N_1, N_2, N_3, P_{0t}
4. Compute maximum population that can be reached, k, i = i + 1
5. If k < N_3 update $N_2 = 0.5\ (N_1 + N_3)$ and k
6. Compute a, b, r, t_0, P_{1t}, P_{2t}, E_1, E_2,
7. If $E_1 < E_2$ D = 1 P_{3t}, = P_{1t}, $E_3 = E_1$ else D = 0, $P_{3t} = P_{2t}$, $E_3 = E_2$
8. Display Region, k, a, b, E_1, P_{1t}, r, t_0, E_2, P_{2t}, D, E_3, P_{3t}
9. If I <= N Repeat from step 3

The projection of the population would now be estimated using the new developed machine learning logistic regression model. The steps that have been agreed on after a series of simulation and testing are for the projection estimates are presented in a pictorial representation in Figure 2. The steps are then translated and written as in Algorithm II that could

be used to write R or MATLAB code for determining the projected population total 2024 - 2069.

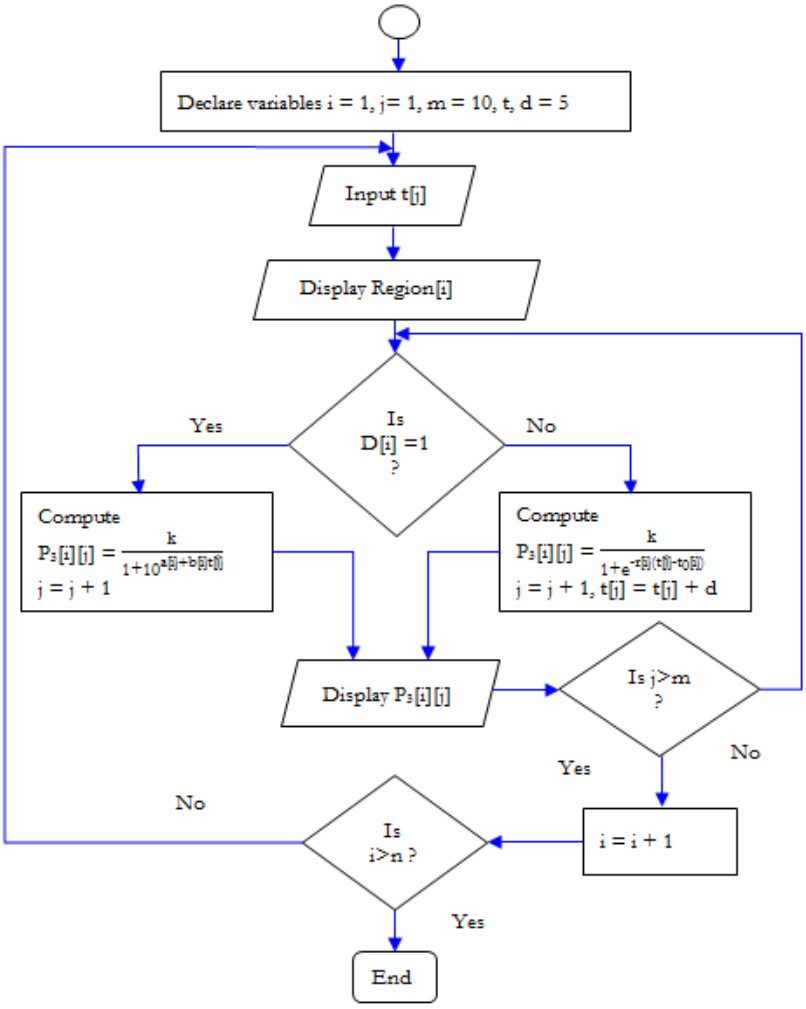

Figure 2. *Machine learning logistic regression population total projections.*

Algorithm II: Machine learning logistic regression population total projections

1. Declare variables i = 1, j = 1, m = 10, t, d = 5
2. Input t
3. Display Region
4. If D = 1 then Compute $P_3 = \frac{k}{1+10^{a+bt}}$ else Compute $P_3 = \frac{k}{1+e^{-r(t-t_0)}}$
5. Display P_3
6. If j <= m repeat from step 4 else execute step 7
7. i = i + 1
8. If i <=n repeat from step 2 else execute step 9
9. End

4. Analysis and Discussion of Results

In analysis and discussion of the results the logistic models are considered in estimating the population total of the smaller units of the Kenya population. In understanding the characteristics of the population that will help in explaining population dynamics the population pyramid and abridged life table are constructed for the population under consideration.

4.1. Population Pyramid

The age-sex pyramid of the population is the classification of the age structure by sex in the form of a histogram such that the base of the pyramid shows the lower age starting from zero while the top represents higher ages to a maximum of 89 years with age intervals of five years [15]. The Kenya National Population Census conducted in 2019 recorded the distribution of population by age and sex as recorded in Table 1 and age-sex pyramid was constructed as in Figure 2 [16].

Table 1. *Distribution of population by age and sex, 2019.*

Age	Male	Female	Age	Male	Female
0 - 4	3,006,344	2,986,769	45 - 49	916,334	869,871
5 - 9	3,116,951	3,085,516	50 - 54	662,981	645,585
10 - 14	3,209,760	3,136,142	55 - 59	546,963	571,105
15 - 19	2,686,264	2,599,442	60 - 64	419,441	450,553
20 - 24	2,112,690	2,334,778	65 - 69	311,345	346,818
25 - 29	1,839,543	2,014,859	70 - 74	235,974	278,548
30 - 34	1,698,678	1,871,887	75 - 79	119,295	163,820
35 - 39	1,348,195	1,301,828	80 - 84	82,933	120,961
40 - 44	1,157,154	1,102,014	85+	76,826	133,923

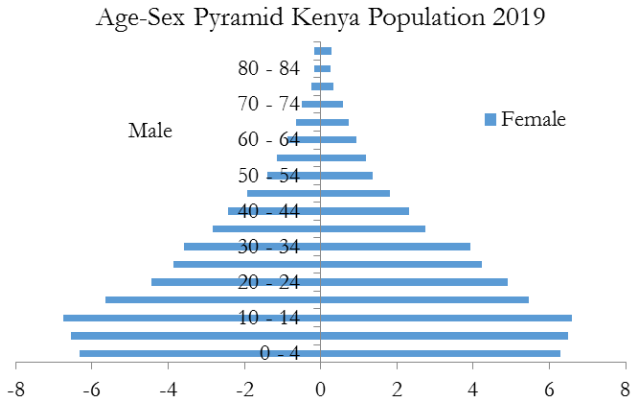

Figure 3. *Kenya population census 2019 pyramid.*

The pyramid in Figure 3 reflects the conditions of age-sex pyramid of developing countries whose main feature is a wide base. This is interpreted by demographers, economists and statisticians as marked decline in the death rate because the population is becoming younger and this is true description of Kenya as it is classified as a developing country as per the united Nations [17].

4.2. Life Table

This is an important technique in the field of health that characterizes the well-being of a population and has applications in insurance companies to predict how long an individual will live when determining premiums for persons who take life insurance policies and further used to predict the compensation payments to nuclear family members in case of disability or loss bread winner's life and therefore not economically active [18]. The life tables identify death rates experienced by a population over a given duration of time and in the recent past have been used make comparisons and estimate the biological limit to life.

Table 2. *Abridged life table for Kenya population 2019.*

x to x+n	p_x	l_x	L_x	T_x	e_x	$1/e_x$
0 - 1	0.962	100,000	97,360	5,773,323	57.73	0.0170
1 - 9	0.994	96,229	959,449	5,675,963	58.98	
10 - 19	0.926	95,661	921,054	4,716,514	49.30	
20 - 29	0.879	88,550	831,864	3,795,460	42.86	
30 - 39	0.885	77,823	733,615	2,963,596	38.08	
40 - 49	0.883	68,900	648,609	2,229,981	32.37	
50 - 59	0.000	60,822	1,581,372	1,581,372	26.00	

An initial population of 100, 000 new born for the Kenya population had the mortality experience monitored and the abridged life table has been prepared and summarized in Table 2 [19]. The life expectancy of the Kenya population has been determined and approximately given as 57.7 years with birth rate or death rate in the stationary population of 0.017, that is, 1700 per 100 000.

4.3. Estimation of the Logistic Models Parameters

In the estimation of the population totals the exponential, logistic I, logistic II, logistic III and logistic IV models will be considered and the performance of the models compared to determine a better performing model for estimating population total. The Kenya population was considered for the study that comprise of ten administrative units such that each region includes more than one county except Nairobi. The Kenya population and housing census survey recorded the population totals in the ten regions as recorded in Table 3 in the duration 1969 – 2019 [16].

Table 3. *Kenya population and housing census 1969 – 2019 in million.*

Region	1969	1979	1989	1999	2009	2019
Coast	0.944	1.343	1.829	2.487	3.325	4.329
North Eastern	0.246	0.374	0.371	0.962	2.311	2.490
Eastern	1.907	2.720	3.769	4.632	5.668	6.821
Central	1.676	2.346	3.117	3.724	4.384	5.482
Rift Valley L	0.317	0.378	0.518	0.902	1.592	1.859
Rift Valley C	0.911	1.347	2.056	2.787	3.791	4.679
Rift Valley U	0.981	1.515	2.407	3.298	4.624	6.215
Western	1.328	1.833	2.544	3.359	4.334	5.022
Nyanza	2.122	2.644	3.507	4.392	5.443	6.270
Nairobi	0.509	0.828	1.325	2.143	3.138	4.397

Table 4. *Estimated parameters for logistic I.*

	Logistic I					
Region	y_0	y_1	y_2	k	a	b
Coast	1.324	1.828	2.473	10.544	0.843	-0.016
North Eastern	0.324	0.511	0.938	0.438	-0.457	-0.039
Eastern	2.694	3.621	4.625	8.793	0.355	-0.020
Central	2.305	3.008	3.706	5.910	0.194	-0.021
Rift Valley L	0.396	0.561	0.906	0.168	-0.241	0.009
Rift Valley C	1.361	1.976	2.790	11.223	0.860	-0.019
Rift Valley U	1.530	2.291	3.323	13.101	0.879	-0.021
Western	1.836	2.502	3.333	12.532	0.765	-0.016
Nyanza	2.700	3.441	4.377	114.541	1.617	-0.011
Nairobi	0.823	1.329	2.073	9.608	1.028	-0.023

Table 5. *Estimated parameters for exponential, logistic II – IV.*

		Logistic IV				
	Exponential	Logistic II			Logistic III	
Region	r	k	a	b	r	t_h
Coast	0.0290	15.694	1.190	-0.016	0.036	2045.9
North Eastern	0.0876	12.314	1.791	-0.025	0.100	2023.6
Eastern	0.0202	49.642	1.373	-0.013	0.023	2100.0
Central	0.0163	10.949	0.717	-0.014	0.026	2024.6
Rift Valley L	0.0568	21.216	1.890	-0.018	0.060	2050.7
Rift Valley C	0.0308	71.272	1.874	-0.016	0.032	2098.3
Rift Valley U	0.0338	11.932	1.043	-0.021	0.050	2018.1
Western	0.0255	13.884	0.974	-0.016	0.035	2031.4
Nyanza	0.0214	24.517	1.028	-0.012	0.027	2055.8
Nairobi	0.0381	6.143	1.065	-0.027	0.067	2008.3

The estimated parameters of the models have been determined and recorded in Table 4 for logistic I and Table 5 for exponential, logistic II, logistic III and logistic IV models. The logistic I model recorded lower estimated population that would be reached in all the regions except Rift Valley U, Nyanza and Nairobi. In Nyanza the recorded value of 114.54 million is the highest recorded value in all the regions and the value recorded is extremely high, which is an indication of weakness in the model. The logistic I model further recorded a very low value of the constant k in North Eastern region which is also is also an indication of flaws in the model for estimating population total. However, further performance

test is required to support this claim.

The estimated population totals for the ten regions in 2019 were determined and recorded as in Table 6. The exponential model recorded the highest estimated population total followed by logistic I, logistic III and logistic II model. The performance of the models is determined using the root mean square (RMSE) that is given as

$$RMSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{T}_i - T_i)^2$$

The Root Mean Square Error (RMSE) recorded for exponential, logistic I, logistic II and logistic III were recorded as in Table 7. The logistic I model recorded the highest RMSE followed by exponential, logistic III and logistic II with values of 1.524, 1.062, 0.819 and 0.477 respectively.

Table 6. *Estimated Kenya Population 2019 in Million.*

Region	Exponential	Logistic I	Logistic II	Logistic III	Logistic IV
Coast	4.446	5.153	4.363	4.353	4.353
North Eastern	5.550	0.436	2.858	4.757	2.858
Eastern	6.936	7.170	7.648	6.901	6.901
Central	5.160	5.186	5.408	5.078	5.408
Rift Valley L	2.808	0.066	2.065	2.737	2.065
Rift Valley C	5.156	6.185	5.539	5.129	5.129
Rift Valley U	6.484	7.646	6.071	6.107	6.107
Western	5.593	6.596	5.522	5.446	5.446
Nyanza	6.744	8.865	6.747	6.661	6.661
Nairobi	4.596	5.577	3.983	4.120	4.120
Total	53.474	52.880	50.204	51.289	49.047

In the search for the parameters of logistic IV using the machine learning, first, take into consideration the two models that have better estimates using the RMSE measure that are selected as logistic II and logistic III models. Secondly, create a dummy variable that will be used to select the parameters using the classification or discriminant criterion that forms the whole process of developing the logistic IV model [20]. Consider a random variable d_i that takes the value 1 if logistic II model has lower square error and therefore its parameters are used for estimating the population total otherwise takes the value 0 if the square error is larger such that the logistic III model is used for estimating the population total [21]. It is observed that the logistic II model will be used in estimating the population totals for the North Eastern, Central and Rift Valley L while the logistic III method will be used in estimating the population total for the remaining regions.

Table 7. *Root Mean Square Error (RMSE) population 2019.*

Region	Exponential	Logistic I	Logistic II	Logistic III	Logistic IV	Dummy
Coast	0.014	0.678	0.001	0.001	0.001	0
North Eastern	9.361	4.220	0.136	5.141	0.136	1
Eastern	0.013	0.122	0.685	0.006	0.006	0
Central	0.104	0.088	0.006	0.163	0.006	1
Rift Valley L	0.902	3.212	0.042	0.772	0.042	1
Rift Valley C	0.227	2.269	0.739	0.202	0.202	0
Rift Valley U	0.072	2.045	0.021	0.012	0.012	0
Western	0.326	2.477	0.250	0.180	0.180	0
Nyanza	0.226	6.736	0.228	0.153	0.153	0
Nairobi	0.039	1.392	0.171	0.077	0.077	0
RMSE	1.062	1.524	0.477	0.819	0.286	

The estimated population totals for logistic IV model have been recorded in Table 6 while the Square Error (SE) and Root Mean Square Errors have been recorded in Table 7. It is observed that logistic IV model records the lowest population total estimate of 49.047 and lowest RMSE of 0.286 which is an indication that the developed logistic model performance is a better estimating model. The Square Errors of all regions for the models are depicted in Figure 4 in which exponential and logistic I records high Square Error values compared to logistic II, logistic III and logistic IV models.

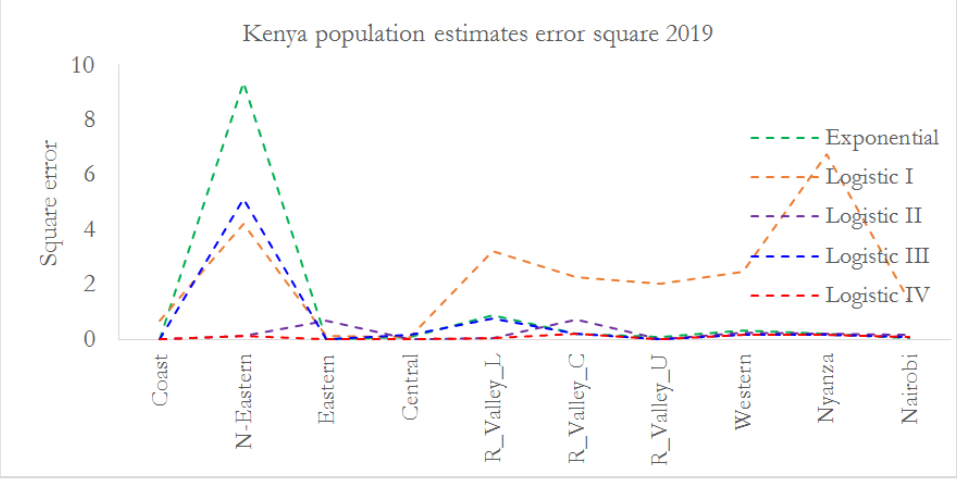

Figure 4. *Square error for ten Kenya regions population 2019.*

Since the logistic IV model has better performance compared to all the other models, it will be given higher priority in estimating the population total for the regions. The first step in defining the model is to estimate the parameters that have been estimated and recorded as in Table 8. The choice of the model and parameters are believed to bridge the gap of obtaining better population total estimates for development.

Table 8. *Parameters for projection of Kenya population 2024 – 2069.*

	Logistic IV				
Region	**k**	**a**	**b**	**r**	**t_h**
Coast	13.004	1.089	-0.015	0.037	2037.6
North Eastern	2.501	1.231	-0.045	0.297	2000.6
Eastern	18.350	0.903	-0.014	0.028	2037.7
Central	9.206	0.616	-0.014	0.048	2011.0
Rift Valley L	1.933	0.910	-0.035	0.167	1999.8
Rift Valley C	6.334	0.802	-0.024	0.064	2002.8
Rift Valley U	15.637	1.143	-0.018	0.045	2028.2
Western	5.849	0.583	-0.026	0.075	1995.0
Nyanza	7.709	0.462	-0.020	0.060	1994.3
Nairobi	12.381	1.338	-0.020	0.048	2031.3

4.4. Population Projection for Regions in Kenya 2024 - 2069

The estimation of population total in the future is important for economic and infrastructure planning for governments to keep the phase of population growth and economic growth. The projection of the population in all regions have been determined and recorded as in Table 9 using the developed logistic IV model. It is projected that the population total for the nation will be 55.02, 62.50, 69.10, 74.65 and 79.14 million in 2029, 2039, 2049, 2059 and 2069 respectively.

Table 9. *Projected population 2024 – 2069 in Million.*

Region	**2024**	**2029**	**2034**	**2039**	**2044**	**2049**	**2054**	**2059**	**2069**
Coast	4.884	5.466	6.064	6.671	7.274	7.864	8.432	8.969	9.929
North Eastern	2.369	2.421	2.452	2.472	2.483	2.490	2.494	2.497	2.499
Eastern	7.432	8.059	8.698	9.341	9.982	10.615	11.235	11.835	12.961
Central	5.494	5.852	6.194	6.518	6.820	7.099	7.354	7.586	7.980
Rift Valley L	1.754	1.809	1.848	1.875	1.894	1.907	1.915	1.921	1.928
Rift Valley C	5.040	5.339	5.579	5.767	5.912	6.022	6.104	6.165	6.244
Rift Valley U	7.078	7.959	8.836	9.688	10.495	11.243	11.920	12.522	13.500
Western	5.255	5.428	5.553	5.643	5.706	5.750	5.781	5.802	5.827
Nyanza	6.586	6.843	7.047	7.206	7.329	7.423	7.495	7.549	7.620
Nairobi	5.104	5.842	6.589	7.325	8.029	8.685	9.280	9.808	10.657
Total	51.00	55.02	58.86	62.50	65.92	69.10	72.01	74.65	79.14

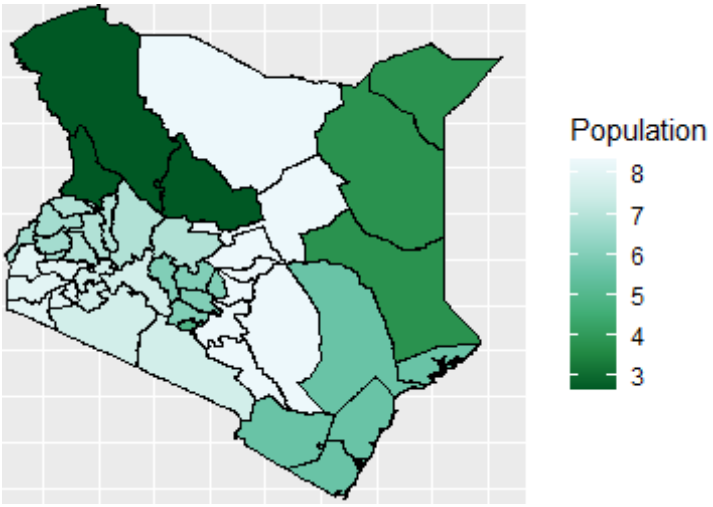

Figure 5. *Projection of Kenya population 2029 for regions in millions.*

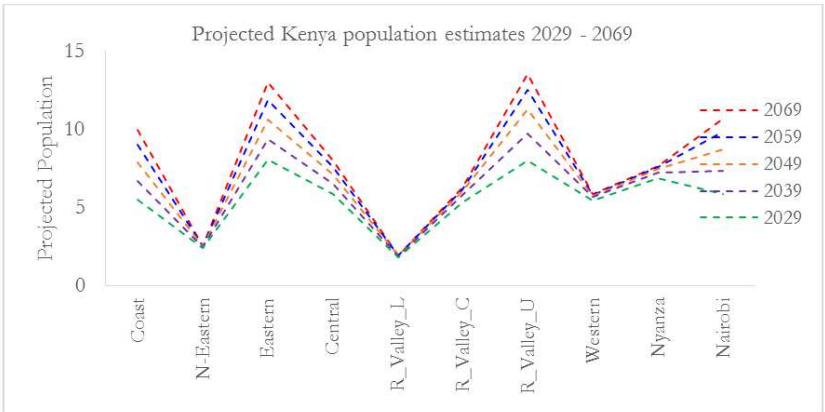

Figure 6. *Projection of Kenya population 2029 - 2069 for regions in millions.*

The projected population for 2029 has been represented on the Kenya map to show the aggregate population for the ten regions as shown in Figure 5. The Eastern region has highest projected population, Rift valley U is the second highest, Nyanza the third highest while Rift Valley L has the least projected population. The projected Kenya population in 2029 – 2069 for the ten regions have been computed and presented in Figure 6. The regions have recorded mixed projection where the regions North Eastern, Rift Valley lower and Western recorded lower increase of population total projections over the duration under consideration while the regions of Coast, Eastern, Rift Valley upper recoded higher increase of the population total projections in the same duration.

5. Conclusion, Recommendations and Further Research

5.1. Conclusion

In this research project, the machine learning logistic regression model has been developed that may be used for estimating population total. The machine learning regression model that is referred to as logistic IV model recorded a better performance when compared to the other models in the same class. When using the RMSE a value of 0.286 was recorded that is lower than the other models that recorded 1.062, 1.524, 0.477 and 0.819 for the exponential, logistic I, logistic II and logistic III models respectively. In estimating the population in the future, the machine learning logistic model projects that the current population total of all the ten regions of 47.564 million will nearly double in 40 years, that is, the projected population total will be 74.655 million in 2059.

5.2. Recommendation

The developed machine learning logistic regression model would be recommended to be used in estimation of the population total for smaller administrative units of countries for instance the forty seven counties in Kenya that have not fully developed the statistics database for planning.

5.3. Further Research

In improving the machine learning logistic regression model an estimation of the parameters for every projection, that is, before the next projection, the parameters (growth rate, time to reach half-way highest population possible and highest possible population size possible) are updated after every interpolation or projection of the population. Further, consideration of estimating the population totals for smaller administrative units for instance the counties, districts or divisions probably will lead to an overall improvement of the estimated population totals.

References

[1] Osaki-Tomita, K., Mrkic, S., Mbogoni, M., Tadesse, S., & Demirci, M. (2017). *Principles and Recomendations for Population and Housing Censuses.* New York: United Nations Publication.

[2] Wesley, E., & Peterson, F. (2017). The Role of Population in Economic Growth. *SAGE Open*, 01 15.

[3] Heady, H., & Hodge, A. (2009). The Effect of Population Growth on Economic Growth: A Meta-Regression Analysis of the Micro-Economic Literature. *Population and Development Review*, 35, 221-248.

[4] Hathout, D. (2013). Modeling Population Growth: Exponential and Hyperbolic Modeling. *Applied Mathematics*, 4, 299-304.

[5] Gotelli, N. J. (2001). *A Primer of Ecology.* Sunderland: Sinauer Associates.

[6] Kabareh, L., & Mageto, T. (2018). Estimation of Finite Population Total Using Birth and Death Process. *International Journal of Engineering, Science and Mathematics*, 7 (3), 33-48.

[7] Kabareh, L., Mageto, T., & Mwema, B. (2017). Approximation of Finite Population Totals Using Lagrange Polynomial. *Open Journal of Statistics*, 7, 689-701.

[8] Kabareh, L., & Mageto, T. (2017). Comparison of the Piecewise Polynomial Approximation to the Newton Backward Difference Polynomial Approximation of Finite Population Totals. *International Journal of Engineering, Science and Matheamtics,* 6 (7), 12-26.

[9] Kabareh, L., & Mageto, T. (2017). Estimation of Bounded Population and Carrying Capacity with the Logistic Model. *Open Journal of Statistics*, 7, 936-943.

[10] Kabareh, L., & Mageto, T. (2018). Estimating Bounded Population Total Using Linear Regression in the Presence of Supporting Information. *International Journal of Mathematics and Computational Science*, 4 (3), 112-117.

[11] Kulkami, S., Kulkami, S., & Patil, S. (2014). Analysis of Population Growth of India and Estimation for Future. *International Journal of Innovative Research in Science, Engineering and Technology*, 3 (9), 15843-15850.

[12] Agarwal, B. L. (1991). *Basic Statistics.* New Delhi: Wiley Eastern Limited.

[13] Keyfitz, N., & Caswell, H. (2005). *Applied Mathematical Demography.* New York: Springer Science Business Media, Inc.

[14] Berman, A. K., & Paul, L. J. (2008). *Algorithms.* New Delhi: Centage Learning India Private Limited.

[15] Jhingan, M., Bhatt, B., & Desai, J. (2007). *Demography.* Delhi: Vrinda Publications (P) LTD.

[16] Mwangi, Z. (2019). *2019 Kenya Population and Housing Census Volume III: Distribution of Population by Age, Sex and Administration Units.* Nairobi: Kenya National Bureau of Statistics.

[17] Secretariat, U. N. (2014). Country Classification. *Data Sources, Country Classification and Aggregation Methodology*, pp. 1-8.

[18] Pagano, M., & Gauvreau, K. (2008). *Principles of Biostatistics.* New Delhi: Cengage Learning India Private Limited.

[19] *Kenya Infant Mortality Rate 1960 - 2020*. (2020). Retrieved from Macrotrends: https://www.macrotrends.net/countries/KEN/kenya/infant-mortality-rate.

[20] Hair, J., Black, W., Babin, B., & Anderson, R. (2014). *Multivariate Data Analysis.* Harlow: Pearson Education Limited.

[21] Cochran, G. W. (1992). *Sampling Techniques.* New Delhi: Wiley Eastern Limited.