Archives of Current Research International

20(1): 6-12, 2020; Article no.ACRI.54471
ISSN: 2454-7077

A Robust Predictive Modelling of Nigeria's Population Growth Rate Using Partial Least Square Regression

Bright C. Offorha[1*], Chukwudike C. Nwokike[1], Okezie, Uche-Ikonne[2], Obubu Maxwell[3], Fidelia C. Onwunmere[1] and Chikezie Uche-Ikonne[4]

[1]Department of Statistics, Abia State University, P.M.B. 2000, Uturu, Nigeria.
[2]Department of Mathematics and Statistics, Lancaster University, Lancaster, UK.
[3]Department of Statistics, Nnamdi Azikiwe University, Awka, Nigeria.
[4]Department of Public Health, Abia State University, P.M.B. 2000, Uturu, Nigeria.

Authors' contributions

This work was carried out in collaboration among all authors. Author BCO conceived the presented idea, designed the study, performed the statistical analysis and wrote the first draft of the manuscript. Authors CCN and OUI managed the analyses of the study and co-wrote the first draft. Authors OM, FCO and CUI managed the literature searches and critically revised the first draft. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/ACRI/2020/v20i130167
Editor(s):
(1) Dr. Marco Muscettola, University of Bari, Italy.
Reviewers:
(1) Mubashir Mehmood, Shaheed Benazir Bhutto University, Pakistan.
(2) Irshad Ullah, Pakistan.
Complete Peer review History: http://www.sdiarticle4.com/review-history/54471

Original Research Article

Received 02 December 2019
Accepted 09 February 2020
Published 20 February 2020

ABSTRACT

Nigeria, a developing nation is experiencing the overwhelming effects of her exponentially ever-increasing population. The resultant effects are clearly evident for all stakeholders to see and feel. Researches have been carried out to study, explain and recommend solutions to this lurking epidemic. But unfortunately, numerous researchers have failed to address key issues in regression modelling as used in their studies, some of such issues are; using Wald's statistic as a variable selection tool rather than the much consensus purposeful variable selection techniques, ignoring the existence of multicollinearity and also missing data. These issues are enough to render the findings in most studies reviewed inadequate, invalid and misleading to be used as a policy-making tool. In this study, the aim is to build a robust predictive model of the Nigeria population growth rate taking into account the aforementioned issues in regression modelling hitherto ignored by some researchers who had used almost this same variables used in this current study. As it would have

**Corresponding author: Email: mrofforha@gmail.com;*

been expected, death rate, maternal deaths and infant deaths all had negative signs indicating an opposing relationship between these variables and Nigeria population growth rate. The assessment carried out showed that our model has high predictive power, hence, could be used to predict future Nigeria's population growth rate.

Keywords: PLSR; Nigeria; missing data; collinearity and multiple imputations.

1. INTRODUCTION

Nigeria, a developing nation is experiencing the overwhelming effects of its exponentially ever-increasing population. The resultant effects are evident in her teeming unemployed (and underemployed) populace, overstretched natural resources, increased crime rates, high cost of living, increased dependency ratio, low quality of life, environmental degradation and negative economic development [1-4].

In 1950 Nigeria was not among the 13 ranking highest populated countries in the world. But in 1996 Nigeria made the list and was ranked 10th, then only to overtake developed countries like Germany and Russia in 2016 to make the 7th position on the list [5] and according to United Nations projections, Nigeria is estimated to surpass the population of USA in 2050 to be the 3rd most populous country in the world with 733 million inhabitants [5]. This is bad for a developing nation, although with abundant natural resources but unable to harness them effectively to her own advantage.

Researches have been carried out to study, explain and recommend solutions to this lurking population growth epidemic by modelling the population growth of Nigeria, with the hope of arming policymakers adequately to make the right decision in population planning. But unfortunately, some researchers have failed to address key issues in regression modelling as used in their studies, such as; using Wald's test statistic as a variable selection tool [6-9], rather than the much consensus purposeful variable selection techniques [10,11] ignoring the existence of multicollinearity among the explanatory variables and also not addressing the issue of missing data. These issues are quite relevant enough to render the findings in most of the reviewed studies inadequate, invalid and misleading to be used as a policy-making tool [6,12].

In this study, the aim is to build a robust predictive regression model of the Nigeria population growth rate taking into account the aforementioned issues in regression modelling.

2. PRELIMINARY ANALYSIS

2.1 Missing Data

Missing data are common in most research, in the case of retrospective study it is either that the data was not available at the time of entry or it was omitted erroneously. Most studies do not handle missing data satisfactorily thereby leading to reduced statistical power and biased estimates, especially when the missing is not at random [13]. Table 1 shows the number (and percentage) of missing data in each of the variables considered in this study.

Maternal mortality and infant mortality appear to have a high percentage of missing values Table 1, but due to how relevant they are in this study they were however retained.

Under the assumption that the missing values are missing completely at random, we used the much adopted statistical technique, Multivariate Imputation by Chain Equation (MICE) to handle it. Furthermore, under MICE an approach, Multiple Classification and Regression Tree (CART) were employed [14].

After due procedures in imputing the missing data were followed, the data was observed to be complete and ready for further analysis.

2.2 Collinearity

Collinearity is a phenomenon used to describe a situation in multivariable regression where two variables are a linear combination of each other while multicollinearity occurs when more than two variables are involved. When this happens the design matrix is not of full rank, and the product of the design matrix and its transpose is singular and non-invertible hence the coefficients of the regressor variables cannot be computed. Different parametric ways of handling critical multicollinearity have been compared and principal component analysis was found to be a better method (Obubu, Nwokike, Virtus C., & Obite, 2019).

Table 1. Number (Percentage) of missing data in each variable

Variable	Pop. growth rate	Death rate	Fertility	GDP	Inflate rate	Life expectancy	Maternal mortality	Infant mortality
Total Complete (%)	59(100)	58(98.31)	58(98.31)	58(98.31)	59(100)	58(98.31)	26(44.07)	53(89.83)
Number missing (%)	0(0)	1(1.69)	1(1.69)	1(1.69)	0(0)	1(1.69)	33(55.93)	6(10.17)

Table 2. GVIF values from the multicollinearity test on each variable

Variable	Death rate	Fertility	GDP	Inflate rate	Life expectancy	Maternal mortality	Infant mortality
GVIF	422.72	2.60	1.49	1.24	397.05	1.56	1.48

To determine the presence of multicollinearity among the variables, the Generalized Variance Inflation Factor (GVIF) for each variable was obtained Table 2. This was done by carrying out a linear regression of that particular variable against all other variables and then obtaining the R^2 from the regression. The formula for *GVIF* is given as

$$GVIF = \frac{1}{(1-R^2)} \quad (1)$$

A GVIF value above 10 for any variable is treated as indicating multicollinearity [15,16]. The GVIF values of death rate and life expectancy are the highest Table 2, indicating a very high degree of correlation involving both variables. This is the basis of using partial least squares regression rather than the conventionally and popularly used o*rdinary* multiple least squares regression method in this current study. Although, another alternative would have been to drop these variables from further analysis, due to their relevance in predicting population growth they were however retained.

Research that used the same variables from the same source as this current study, did ignore these potential problems and still went ahead to model Nigeria's population growth rate using ordinary least squares regression models. This would have definitely led to reduced statistical power, misleading results, spurious findings, bias and invalid conclusions [6,12].

3. MATERIALS AND METHODS

3.1 Data Set

Data were sourced from the World Bank online database known as World Development Indicators, the data time frame spans from 1960 to 2018. It is common knowledge that the World Bank keeps high quality and credible data at both national and international levels.

3.2 Partial Least Square Regression

Partial least squares regression (PLSR) is commonly used in the situation where the variables are more than the observations, missing values are recorded and when multicollinearity exists in a dataset [17,18,19]. In this study, it was employed to handle the collinearity evident in the data set. Collinearity as explained in Section 2.2 violets the assumptions of ordinary least squares method that relies so much on the independence between the explanatory variables.

Principal Component Analysis (PCA) is similar to PLSR, but PLSR is a supervised technique as it explains the variations in both the response variable/$Y_{(n,1)}$ and the explanatory variables/$X_{(n,p)}$ (also called latent variables).

3.2.1 PLSR algorithm

Using orthogonal score approach [20,19], the design matrix $\underline{X}$ and the response matrix $\underline{Y}$ are assumed to be centred and possibly scaled (i.e. normalized) as an initial step to perform PLSR. If A represent the number of important components for prediction and $1 \leq A \leq p$, $a = 1,2,\ldots,A$, then the algorithm works iteratively as:

1. Compute the loadings weight as

   $w_a = X_{a-1}^T y_{a-1}$

   The weights define the direction in the space spanned by X_{a-1} of the maximum covariance with y_{a-1}. Normalize to loading weights to have a length equal to 1 by

$w_a \leftarrow \frac{w_a}{w_a}$

2. Obtain the scores t_a as

$t_a = X_{a-1} w_a$

3. Obtain the $\underline{X}$ and $\underline{Y}$ loadings p_a and q_a respectively, given as

$p_a = X_{a-1}^T \frac{t_a}{t_a^T t_a}$

$q_a = y_{a-1}^T \frac{t_a}{t_a^T t_a}$

4. Lastly, the data matrices are deflated by subtracting the contribution of t_a via:

$X_a = X_{a-1} - t_a p_a^T$

$y_a = y_{a-1} - t_a q_a$

- If $a < A$ return to 1.
- The vectors w_a, p_a , q_a, are stored in the matrices/vectors W, P and Q respectively.
- Then the PLSR estimates of the regression equation coefficients are obtained using these estimators: $\hat{\beta} = W(P^T W)^{-1} q$ and $\hat{\alpha} = \bar{y} - \bar{x}\hat{\beta}$

Hence, the regression model for PLSR can be mathematically expressed as

$$g(x) = \hat{\alpha} + \hat{\beta}_a (x_a - \bar{x}_a), \qquad a = 1, \dots, A \qquad 1$$

Where a is a particular component in the range $1 \leq A \leq p$.

4. RESULTS

For the purpose of model building and model evaluation, the data was divided into two sub-data. The sub-data used to develop the model spanned from 1960 to 2009 (50 observations) while that for testing spanned from 2010 to 2018 (9 observations). Hence, the original data spanned from when Nigeria gained her independence (1960) to when last the database was updated (2018), making it a total of 59 observations.

The RMSEP for the k-segments cross-validation and the variance explained were the two measures used to assess the best components (also called latent variables) for best predictions Table 3. In the RMSEP we look out for the components with the least CV or adj. CV, evidently the third (3) components have the least values (0.124 or 0.123 respectively) Table 3. In the variance explained we are to look out for that component that best explains the variation in both the response and explanatory variables respectively.

There is a trade-off at the 3 components in variance explained in both the response and explanatory variables, 78.04% and 63.23% respectively, beyond that it appears that there is a little improvement in the variance explained for the response variable (Nigeria population growth rate).

Table 4 shows the influence of each explanatory variable on the Nigeria population growth rate variable, the signs indicate which variable is positively or negatively related to the response variable.

The death rate, maternal deaths and infant deaths variables all have negative signs while the rest variables have positive signs Table 4. It is clear in Table A5 that some variables do have a negligible effect on the response variable, such as inflation, infant deaths and gross domestic product.

Table 3. RMSEP and variance explained (%) for each PLSR components

Components	Root mean squared error of prediction (RMSEP)		Variance explained (%) in the training section	
	Cross validation (CV)	Adjusted cross validation (adj. CV)	Explanatory	Response
1 comps	0.224	0.217	30.34	60.64
2 comps	0.128	0.127	45.98	76.27
3 comps	0.124	0.123	63.23	78.04
4 comps	0.133	0.131	76.43	78.13
5 comps	0.148	0.148	88.71	78.13
6 comps	0.149	0.146	100.00	78.13
7 comps	0.149	0.146	100.00	78.25

Table 4. Estimated PLSR coefficients for each variable

Variable	(Intercept)	Death	Fertility	GDP	Inflation	Life expectancy	Maternal deaths	Infants deaths
Coefficient	2.4981	-0.1009	0.1409	0.0192	0.0009	0.1145	-0.0188	-0.0025

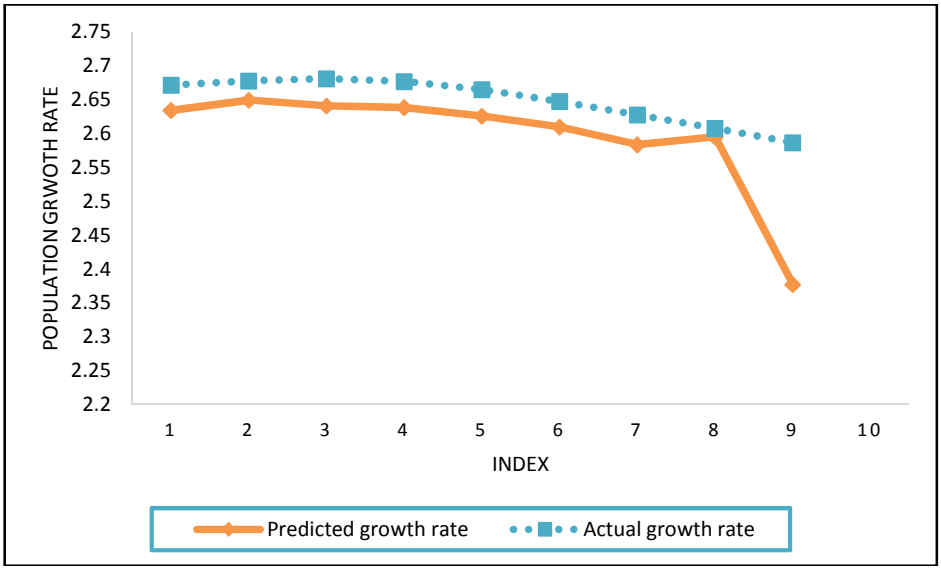

Fig. 1. Comparison between predicted growth rate and actual growth rate values

In assessing the predictive power of the model developed, line plots of predicted Nigeria population growth rate values using the PLSR model built against actual growth rate values (test sub-data) were compared. There is clearly a close agreement between the values of both plots except for the 9th values Fig. 1.

5. DISCUSSION

In this empirical study, the interest is to build a robust predictive model of Nigeria population growth rate by taking into account technical issues, which had hitherto been ignored by researchers. Due to the presence of collinearity the PLSR model was considered rather than the commonly used multivariate least squares regression model.

All variables were retained as prediction is the interest here, but the influence they assert on Nigeria population growth rate (response) variable are different in direction and size Table 4. Our model showed that of these variables; death rate, maternal deaths and infant deaths have negative signs which supports a natural phenomenon in which a decrease in these variables is expected to be accompanied with an increase in the population growth rate, and vice versa. This shows that our model is adequate and reasonably explains much of the variation in the Nigeria population growth rate variable.

Furthermore, to ascertain the robustness of our model we investigated its predictive ability using a test sub-data. Fig. 1, shows a close agreement between the values of the test data and the predicted values using the PLSR model built, hence we could say that all steps followed in Section 2 led to building a robust predictive regression model for Nigeria's population growth rate.

6. CONCLUSION

In this study, we set out to address issues usually ignored by most researchers while modelling Nigeria's population growth rate, making it impossible to build a robust model. To a large extent, we were successful.

Missing data were handled using MICE method to impute estimates for the missing values. Collinearity was detected which necessitated a switch from the popularly used ordinary multivariate least squares regression method to partial least squares regression method.

As it would have been expected, death rate, maternal deaths and infant deaths all had negative signs indicating an opposing relationship between these variables and Nigeria population growth rate. A plot comparing predicted population growth rate obtained using the model estimated and the actual growth rate values obtained from the test sub-data indicated a very close agreement between both values, this indicates that the model has a strong predictive accuracy.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Emmanuel CO. Another look at the impact of Nigeria's growing population on the country's development. African Population Studies. 2006;21(1):1-18.
2. Nwosu C, Dike AO, Okwara K. The effects of population growth on economic growth. The International Journal of Engineering and Science. 2014;3(11):7-18.
3. Oramah IT. The effects of population growth in Nigeria. Journal of Applied Sciences. 2006;6:1332-1337.
   DOI: 10.3923/jas.2006.1332.1337
4. John C. Nigeria faces a crippling population boom. Retrieved from Council on Foreign Relations; 2018.
   Available:https://www.cfr.org/blog/nigeria-faces-crippling-population-boom
5. United Nations. Department of Economic and Social Affairs, Population Division. World Population Prospects 2019: Highlights. ST/ESA/SER.A/423; 2019.
   Available:https://population.un.org/wpp/Publications/Files/WPP2019_Highlights.pdf
6. Eli HT, Mohammed ID, Amade P. Impact of population growth on economic growth in Nigeria (1980-2010). Journal of Humanities and Social Science. 2015; 20(4):115-123.
   DOI: 10.9790/0837-2045115123
7. Olatayo TO, Adeboye NO. Predicting population growth through births and deaths rate in Nigeria. Mathematical Theory and Modeling. 2013;3(1):96-101.
8. Azuh D, Matthew AO, Fasina FF. The determinants of population growth in Nigeria: A co-integration approach. The International Journal of Humanities & Social Studies. 2016;4(11):38-44.
9. Nyoni T. Determinants of population growth: Empirical evidence from Pakistan. Munich Personal RePEc Archive; 2018.
10. Hosmer D, Lemeshow S. Applied logistic regression. New York: Wiley; 2000.
11. Bewick V, Cheek L, Ball J. Statistics review 14: Logistic regression. Critical Care. 2005;9(1).
12. Ogunleye OO, Owolabi OA, Mubarak M. Population growth and economic growth in Nigeria: An appraisal. International Journal of Management, Accounting and Economics. 2018;5(5):282-299.
13. Kang H. The prevention and handling of the missing data. Korean Journal of Anesthesiology. 2013;64(5):402-406.
    DOI: 10.4097/kjae.2013.64.5.402
14. Geeta C, Vasudha V, Jayanthi R. A comparison of multiple imputation methods for data with missing values. Indian Journal of Science and Technology. 2017;10(19): 1-7.
    DOI: 10.17485/ijst/2017/v10i19/110646
15. Paul DA. Logistic regression using SAS system: Theory and applications. Cary, North Carolina: Cary, N.C.: SAS Institute; 2012.
16. Fox J, Monette G. Generalized collinearity diagnostics. Journal of American Statistical Association. 1992;87(417):178-183.
17. Philippe B, Vincenzo EV, Michel T. PLS generalised linear regression. Computational Statistics & Data Analysis. 2005;48:17-46.
18. Noraini I, Antoni W. Partial least squares regression-based variables selection for water level predictions. American Journal of Applied Sciences. 2013;10(4):322-330.
19. Tahir M, Kristian HL, Lars S, Solve S. A review of variable selection methods in partial least squares regression. Chemometrics and Intelligent Laboratory Systems. 2012;118:62-69.
20. Wold S, Martens H, Wold H. The multivariate calibration problem in chemistry solved. Conference Proceeding Matrix Pencils. 1983;286-293.

Appendix A

Table 5. Variable description

Variable	Description
Pop. Growth rate (annual %)	Is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship.
Death rate, crude (per 1,000 people)	Crude death rate indicates the number of deaths per 1,000 midyear population.
Fertility rate, total (births per woman)	Total fertility rate represents the number of children that would be born to a woman if she were to live to the end of her childbearing years and bear children in accordance with age-specific fertility rates of the specified year.
GDP growth (annual %)	The annual percentage growth rate of GDP at market prices based on constant local currency. Aggregates are based on constant 2010 U.S. dollars.
Inflation, GDP deflator (annual %)	Inflation, as measured by the annual growth rate of the GDP implicit deflator, shows the rate of price change in the economy as a whole.
Life expectancy at birth, total (years)	Life expectancy at birth indicates the number of years a new-born infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life.
Maternal deaths	A maternal death refers to the death of a woman while pregnant or within 42 days of termination of pregnancy, irrespective of the duration and site of the pregnancy, from any cause related to or aggravated by the pregnancy or its management but not from accidental or incidental causes.
Infant deaths	The number of infants dying before reaching one year of age.

Source: World Bank online database

Peer-review history:
The peer review history for this paper can be accessed here:
http://www.sdiarticle4.com/review-history/54471