Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Modelling children ever born using performance evaluation metrics: A dataset

Jecinta U. Ibeji*, Temesgen Zewotir, Delia North, Lateef Amusa

School of Mathematics, Statistics and Computer Science, University of KwaZulu Natal, Durban, South Africa

ARTICLE INFO

Article history:
Received 13 November 2020
Revised 14 April 2021
Accepted 15 April 2021
Available online 22 April 2021

Keywords:
Poisson regression
Negative binomial regression
Generalized poisson regression
Performance evaluation metrics

ABSTRACT

Predicting the number of total children ever born in a country is a key component for proper implementation of economic growth policy. Here, performance metrics were used to predict models that appropriately describe the factors that affect children ever born. A comparison of 60% training and 40% validation, 70% training and 30% validation, 80% training and 20% validation also 90% training and 10% validation was performed respectively to examine the three models' behaviours (Poisson regression, Negative Binomial regression and Generalized Poisson regression) with RMSE, R^2, MAE and MSE as performance metrics. Although all the three models had almost identical performance evaluation metrics, the Poisson regression was chosen as the most appropriate model because it is the simplest model.

DOI of original article: 10.1016/j.sciaf.2020.e00494

* Corresponding author.
E-mail address: jecintaanyaogu@gmail.com (J.U. Ibeji).

https://doi.org/10.1016/j.dib.2021.107077

Specifications Table

Subject	Statistics, Demography
Specific subject area	Statistics
Type of data	The raw data is available in SPSS format (sav). The analyzed data in this article are provided in tables and figures
How data was acquired	Secondary data was obtained from Nigeria Demographic and Health Survey (NDHS), covering all the regions
Data format	NDHS is a secondary data consisting of a refined primary data collected and collated
Parameters for data collection	The data were secondary data covering all regions of Nigeria Demographic Health Survey
Data source/ Location	Primary data source: http://www.dhsprogram.com/data/dataset_admin/login_main.cfm Abuja, Nigeria
Data accessibility	Data can be downloaded as excel file in supplementary (.xlsx)
Related research article	Jecinta U Ibeji, Delia North, Temesgen Zewotir, Lateef Amusa Modelling Fertility levels in Nigeria using Generalized Poisson regression-based Approach, Scientific Africa. https://doi.org/10.1016/j.sciaf.2020.e00494

Value of the Data

- The dataset gives information about the number of total children ever born in Nigeria, which is a key component for proper implementation of economic growth policy.
- Analysis of this dataset provides insight into the appropriate model describing the factors that affect children ever born in Nigeria.
- The dataset could be used to create integrated support tools for the government, health policymakers and international agencies concerned with fertility-associated problems.
- The information in this dataset will be valuable in planning and evaluation of fertility policies in Nigeria.

1. Data Description

Nigeria Demographic and Health survey (NDHS) 2013 was implemented by the national population commission, an agency saddled with the responsibility of collecting and collating demographic data. In 2013, data on fertility levels, marriage and fertility preference were collected. The target groups were women within the age of 15 and 49 years in randomly selected households across Nigeria. 30878 women who were within childbearing age were interviewed out of 30977 households selected. Children ever born are children born alive by married women from age 15 years and above. The data contains information on key indicators for urban and rural areas in Nigeria, the six geo-political zones, the 36 states and the federal capital territory. The data on childbearing patterns were collected in different forms. First, each woman was asked the number of daughters and sons living with her, the number born alive and later died and those living elsewhere. A complete history of all the women's children including the name, sex, month and year of birth, age, and survival of each of the children. Data was also collected for women ever been pregnant.

The secondary data containing total children ever born with the independent variables was partitioned into training and validation of different percentages to study the performance of the three models using the parameter estimates as seen in Tables 1–4.

Tables 1–4 show the predictive statistics of the dataset, while the inferential statistics of this dataset was discussed in our previous publication [1]. Table 1 contains a summary comparison of Poisson regression, Negative Binomial regression and Generalized Poisson regression using 60%:40% partitioning, while Tables 2–4 contain 70%:30%, 80%:20% and 90%:10%, respectively. All the variables used here can be seen in Table S1 in the supplementary information.

Table 1
Summary of Poisson, Negative Binomial and Generalized Poisson regression data analysis for 60%:40% training:validation dataset splitting.

	MAE	MSE	RMSE	R^2
Poisson				
Training	1.613814	4.313774	2.076963	0.3624504
Validation	1.600686	4.262919	2.064684	0.3604352
Negative Binomial				
Training	1.613813	4.313784	2.076965	0.3624491
Validation	1.600686	4.26293	2.064686	0.3604339
Generalized Poisson				
Training	1.613700	4.315765	2.077442	0.3622371
Validation	1.600644	4.264933	2.065171	0.3602402

Table 2
Summary of Poisson, Negative Binomial and Generalized Poisson regression data analysis for 70%:30% training:validation dataset splitting.

	MAE	MSE	RMSE	R^2
Poisson				
Training	1.605426	4.273806	2.067319	0.3588404
Validation	1.616214	4.344664	2.084386	0.366524
Negative Binomial				
Training	1.605426	4.273815	2.067321	0.3588393
Validation	1.616214	4.344675	2.084388	0.3665226
Generalized Poisson				
Training	1.605347	4.276152	2.067886	0.3585859
Validation	1.616177	4.347424	2.085048	0.3661882

Table 3
Summary of Poisson, Negative Binomial and Generalized Poisson regression data analysis for 80%:20% training:validation dataset splitting.

	MAE	MSE	RMSE	R^2
Poisson				
Training	1.611256	4.305445	2.074957	0.3598963
Validation	1.594595	4.218068	2.053794	0.3722824
Negative Binomial				
Training	1.611255	4.305455	2.074959	0.359895
Validation	1.594595	4.218074	2.053795	0.3722818
Generalized Poisson				
Training	1.611217	4.307683	2.075496	0.3596542
Validation	1.594352	4.218566	2.053915	0.3722467

Table 4
Summary of Poisson, Negative Binomial and Generalized Poisson regression data analysis for 90%:10% training:validation dataset splitting.

	MAE	MSE	RMSE	R^2
Poisson				
Training	1.609629	4.299476	2.975103	0.3615511
Validation	1.603307	4.213045	2.999453	0.3655643
Negative Binomial				
Training	1.609629	4.299486	2.975114	0.3615499
Validation	1.603307	4.213054	2.999464	0.3655631
Generalized Poisson				
Training	1.609559	4.301766	2.977169	0.3613071
Validation	1.603503	4.216727	3.001452	0.3650468

Based on the mean absolute error and root mean square error for Poisson, Negative Binomial and Generalized Poisson regression model, the performance evaluation for the training sample is higher than the validating sample, although with a slight difference [2,3]. Tables 1–4 identified Poisson as the most appropriate predictive model for validating samples.

In the predictive modeling, all the three models showed almost identical performance evaluation metrics while the Poisson regression was chosen as the most appropriate as it is the simplest model. This is because the root mean square error, mean squared error and the mean absolute error of the three models showed almost identical performance metrics.

Comparing the root mean square error, mean squared error, R-squared and mean absolute error for training and validating sample of each model, showed that all the three models had almost identical performance evaluation metrics. The Poisson regression was chosen as the most appropriate because it is the simplest model. This is important because it balances the goodness of fit with simplicity and predicts the probability of the outcome. Complex models adapt their shape to fit the data, but the additional parameter may not represent anything useful.

2. Experimental Design, Materials and Methods

In this work, Secondary data was obtained from Nigeria Demographic and Health Survey 2013, covering all the regions containing all analyzed primary data. The Secondary data was filtered, and the variables of interest was chosen. One major issue in fitting a model is how well it performs when applied to new data. To solve this problem, the data needs to be partitioned into a training set, which is used to create the model; a validation set, which is used to evaluate the model performance; and a test set, which is used to assess how well the algorithm was trained using the training dataset. Using SAS version 9.4, a comparison of 60% training and 40% validation, 70% training and 30% validation, 80% training and 20% validation, and 90% training and 10% validation was performed respectively to examine the three models behaviours (Poisson regression, Negative Binomial regression, and Generalized Poisson regression). Furthermore, the variations in the training performance evaluation metrics under each partition was examined as follows. First, the model is fit on the training dataset using a supervised learning method. The training dataset is then run with the current model, and this is used to compare the target for each input vector in the training dataset. Based on this and the specific learning algorithm being used, the models' parameters were adjusted, while variable selection and parameter estimation can be included in the model fitting [4]. Subsequently, in the validation dataset, the fitted model was used to predict the responses. While tuning the model's hyperparameters, the validation dataset provides an unbiased evaluation of a model fit on the training dataset [5].

The mean absolute error (MAE), Mean squared error (MSE), root mean square error (RMSE) and coefficient of determination (R^2) are the performance evaluation metrics used. The formulas are presented below,

Root Mean Square Error (RMSE) is given as:

$$RSME = \sqrt{\frac{\Sigma_{t=1}^{N}(Predicted_i - Actual_i)^2}{N}}$$

Mean Absolute Error (MAE) is given as:

$$MAE = \frac{\Sigma_{t=1}^{N}|predicted_i - actual_i|}{N} = \frac{\Sigma_{t=1}^{N}|e_i|}{N}$$

Mean squared error (MSE) is given as:

$$MSE = \frac{1}{N}\Sigma_{t=1}^{N}(predicted_i - actual_i)^2$$

Where N is the total number of observations.

Coefficient of determination (R^2):

$$R^2 = cor(actual_1, predicted_1)^2$$

CRediT Author Statement

Jecinta U. Ibeji: Conceptualization, methodology, Data curation, Writing - Original draft preparation; **Temesgen Zewotir:** Conceptualization and supervision; **Delia North:** Reviewing and Editing; **Lateef Amusa:** Visualization, Investigation Data search and Editing.

Declaration of Competing Interest

The authors do declare that there is no conflict of interest.

Acknowledgement

Authors are grateful to the University of KwaZulu-Natal, Durban for operational and infrastructural support and DHS for granting the author access to NDHS 2013 data.

Funding

This work was not funded.

Supplementary Materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.dib.2021.107083.

References

[1] J.U. Ibeji, T. Zewotir, D. North, L. Amusa, Modelling fertility levels in Nigeria using Generalized Poisson regression-based approach, Scientific African 9 (2020) e00494.
[2] W. Aertsen, V. Kint, J. Van Orshoven, K. Özkan, B. Muys, Comparison and ranking of different modelling techniques for prediction of site index in Mediterranean mountain forests, Ecol. Model. 221 (8) (2010) 1119–1130.
[3] D. Onoro-Rubio, R.J. López-Sastre, Towards perspective-free object counting with deep learning, in: European Conference on Computer Vision, Springer, 2016, pp. 615–629.
[4] D. Chicco, Ten quick tips for machine learning in computational biology, BioData Mining 10 (1) (2017) 35.
[5] G. James, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical Learning, Springer, 2013.