International Journal of
Clinical Biostatistics and Biometrics

ORIGINAL RESEARCH

A Comparative Study of Probability Distribution Models for Number of Antenatal Visits

Ayodeji Isaac Akinjisola[1] and Emmanuel Teju Jolayemi[2]*

[1]*Department of Mathematical and Computing Science, Thomas Adewumi University, Kwara State, Nigeria*
[2]*Department of Statistics, University of Ilorin, Ilorin, Nigeria*

***Corresponding author:** Ayodeji Isaac Akinjisola, Department of Mathematical and Computing Science, Thomas Adewumi University, Oko, Kwara State, Nigeria, Tel: +2348167168312*

Abstract

There are several complications that arise as a result of lack or insufficient antenatal care received by pregnant women. Studies show that lack of education, poverty and several other factors give rise to low antenatal attendance in Nigeria and the consequences are grievous. This study was aimed at determining the factors that affect the frequency of antenatal visits in Nigeria. Data were obtained from the Demographic and Health Survey carried out in Nigeria. With these data, bar chart was used to describe the distribution of the data and the data showed evidence of zero inflation. Poisson, Negative Binomial, Zero-Inflated Poisson, Zero-Inflated Negative Binomial, Poisson Hurdle, Negative Binomial Hurdle distribution models were subjected to goodness of fit test to determine the distribution that fits the data most. The goodness of fit tests used were Vuong test and Akaike Information Criterion (AIC). The Zero-Inflated Negative Binomial model was selected out of the class of distribution models compared, and it was used to model the data. The results showed that level of education, wealth status, region and age affect the frequency of antenatal visits in Nigeria while the type of place of residence does not affect frequency of antenatal visits.

Keywords

Zero inflated, Antenatal care, Generalized linear model, Hurdle, Regression

Abbreviations

AIC: Akaike Information Criterion; ANC: Antenatal Care; ANOVA: Analysis of Variance; CI: Confidence Interval; GLM: Generalized Linear Model; NDHS: Nigeria Demographic and Health Survey; NARHS: National AIDS and Reproductive Household Survey; OR: Odds Ratio; ZIP: Zero-Inflated Poisson; ZINB: Zero-Inflated Negative Binomial

Introduction

Complications that happen during pregnancy and childbirth are the most leading causes of maternal mortality and morbidity among women whose age ranges from 15 to 49 in developing countries [1]. Annually around 287,000 women die secondary to pregnancy related causes in the globe. Among this figure, 99% of the maternal death is from underdeveloped countries. In developing countries, almost all pregnant women receive antenatal care at least once, but in sub-Saharan countries, the report is around 68%, where women take antenatal care (ANC) services at least once and majority of them visit the health institutions at third visit [2].

Antenatal care (ANC) is the care a pregnant woman receives during her pregnancy through a series of consultations with trained health care workers such as midwives, nurses, and sometimes a doctor who specializes in pregnancy and birth. ANC is a key strategy for reducing maternal morbidity and mortality directly by affording increased chances of the timely identification of high-risk pregnancies. It also represents an entry point for the integrated use of skilled health personnel. Empirical studies of preventive services have often found that regular monitoring of women during pregnancy is vital to reducing birth-related complications, providing supportive care, and promoting safer motherhood. In contrast, low health service utilization throughout the prenatal period breaks the critical link in the continuum of care and contributes to poor birth outcomes [3].

The World Health Organization's (WHO)

recommendation of at least four ANC visits spaced across regular intervals, and with a skilled attendant has been shown to improve health outcomes for both expectant mothers and infants. The WHO contends that the first antenatal care visit should occur with a skilled health attendant and as early as possible in the first trimester. For optimal health outcomes, comprehensive ANC should contain each of the following components: Identification of pre-existing conditions such as anemia, HIV or hypertension; Early detection of complications arising during pregnancy, such as gestational diabetes or preeclampsia; Health promotion and disease prevention, including vaccines, nutrition counseling and micronutrient supplements; Birth preparedness and complication planning, including breastfeeding counseling and antiretroviral therapy for HIV positive women [4].

Nearly half of all pregnancies in developing countries are not monitored by skilled healthcare professionals. Moreover, women who do receive some ANC generally have irregular visits, large spacing between visits and poor communication with healthcare providers throughout pregnancy. While ANC policies have improved globally, many low-income pregnant women do not receive the recommended number of ANC visits and often initiate and attend visits late in their pregnancies [5].

Two nationally representative surveys were conducted in Nigeria: Nigeria Demographic and Health Survey (NDHS) in 2013 and National AIDS and Reproductive household survey (NARHS) in 2012. The two surveys showed that the proportion of pregnant women who had not attended any ANC service in Nigeria was 33.9% and 34.9% respectively. According to the 2013 NDHS, only 60.9% among women of child bearing age (15-49 years) who had a live birth in the five years preceding the survey received ANC from a trained skilled ANC provider (i.e., a doctor, nurse or midwife, or auxiliary nurse or midwife). Only half (51.0%) reported making four or more ANC visits during the pregnancy. The questions are why are pregnant women not attending ANC in Nigeria? What are the limiting factors? What are the barriers?

There is a need to understand the reasons and in particular the limiting factors for the low attendance in ANC service in Nigeria. This research work used data from 3493 women. Analysis is designed to identify the factors that contribute significantly to the number of antenatal visits in Nigeria. In determining those factors, the statistical method to use is the regression model.

In biostatistics or health research, outcomes of interest often consist of count variables. For such count data, the standard framework for explaining the relationship between the outcome variable and a set of explanatory variables includes the Poisson and Negative Binomial regression models. However, the basic Poisson regression model forces the conditional variance of the outcome to equal the conditional mean, which is of limited use in real life. The Negative Binomial regression can be written as an extension of Poisson regression and it enables the model to have greater flexibility in modeling the relationship between the conditional variance and the conditional mean compared to the Poisson model. Also, an often encountered characteristic of count data, is that the number of zeros in the sample can exceed the number of zeros predicted by either Poisson or Negative Binomial model, and this is of interest because zero counts frequently have special status [6]. In general, Hurdle models and Zero-Inflated models are used for modeling count data with a preponderance of zeros. The Hurdle model is a two component model in which one component models the probability of zero counts and the other component uses a truncated Poisson/Negative Binomial distribution that modifies an ordinary distribution by conditioning on a positive outcome [7]. The Zero-Inflated model has a distribution that is a mixture of a binary distribution that is degenerate at zero and an ordinary count distribution such as Poisson or Negative Binomial. The Hurdle model considers the zeros to be completely separate from the non-zeros. The Zero-Inflated model is similar to the Hurdle model; however, it permits some of the zeros to be analyzed along with the non-zeros. The choice of the Zero-Inflated model in this research work is guided by the researcher's beliefs about the source of the zeros. There are two distinct processes driving zeros, one is sampling zeros which occur by chance and can be assumed a result of a dichotomous process, and the other one is structural zeros (true zeros) which are inevitable and are part of the counting process. Beyond this substantive concern, the choice should be based on the model providing the closet fit between the observed and predicted values [8].

The Zero-Inflated Poisson (ZIP) regression model was first introduced by Lambert [7] and she applied this model to the data collected from a quality control study, in which the response typically is the number of defective products in a sample unit. Further applications for the ZIP regression model can be found in dental epidemiology, occupational health, and children's growth and development. In practice, even after accounting for zero-inflation, the non-zero part of the count distribution is often over-dispersed. In this case, Greene (1994) described an extended version of the Negative Binomial model for excess zero count data, the Zero-Inflated Negative Binomial (ZINB), which may be more appropriate than the ZIP. It has been established that the ZIP parameter estimates can be severely biased if the non-zero counts are over-dispersed in relation to the Poisson distribution. The Zero-Inflated Negative Binomial (ZINB) has application in Biology.

Methods

Source of data collection

The source of data used for this research work is

secondary source of data. It was collected from the Demographic Health Survey archive based on survey conducted in Nigeria.

Generalized linear model

The Generalized linear model (GLM) is a flexible modeling framework which allows the response variables to have a distribution form other than normal. It also allows the linear model of several covariates to be related to a response variable via arbitrary choices of link functions.

Building a Generalized linear model consists of three steps:

a. Choosing a distribution for the response variable (Y)

b. Specifying covariates (X)

c. Choosing a link function between the mean of the response variable E(Y) and a linear combination of the covariates (βX)

Classical models such as analysis of variance (ANOVA) and ordinary least squares regression also belongs to the generalized linear model when Y is normally distributed. Y can also be specified as other distributional forms in exponential family such as a Binomial distribution, Poisson distribution, Negative Binomial distribution, and Gamma distribution. The link function brings together the response variable and the linear combination of the covariates.

Poisson regression model

Poisson regression is a generalized linear regression model used to model count data and contingency tables. Poisson regression assumes the response variable Y has a Poisson distribution, and assumes the logarithm of its expected value can be modelled by a linear combination of unknown parameters [9]. A Poisson regression model is sometimes known as Log-linear model, when used to model contingency tables. Therefore, the probability distribution of the Poisson random variable y can be written as

$$P(Y/Xi) = \frac{\mu^Y e^{-\mu}}{Y!}, \ y = 0, 1, 2, 3, \ldots$$

The mean of the above distribution is equal to the variance i.e. E(Y/Xi) = V(Y/Xi) = μ

Assumptions of Poisson regression

1. The response variable is a count data.

2. There are one or more explanatory variables which could be continuous, ordinal or nominal/dichotomous scale.

3. The observations must be independent of each other.

4. The count distribution must follow a Poisson distribution

5. The mean and variance are identical.

Negative Binomial regression model

The negative binomial regression is a generalization of poisson regression which loosens the restrictive assumption that the variance is equal to the mean made by Poisson model. It is used for over-dispersed count data. It has an extra parameter to model the over-dispersion. Therefore, the probability distribution of the Negative Binomial random variable y can be written as

$$P(Y/Xi) = \left(\frac{k}{k+\mu}\right)^k \frac{\Gamma(k+Y)}{\Gamma(Y+1)\Gamma k} \left(\frac{\mu}{k+\mu}\right)^Y, \ Y = 0, 1, 2, 3, \ldots; k, \mu > 0,$$

Where k is a shape parameter which quantifies the amount of overdispersion, and Y is the response variable. The Negative Binomial distribution approaches the Poisson distribution as k tends to infinity (no over-dispersion) [10].

The mean of the above distribution E(Y/Xi) = μ and the variance, V(Y/Xi) = $\mu + \frac{\mu^2}{k}$.

Zero inflated regression models

Statisticians have developed new approaches to model zero-inflation in count data. The new approaches give rise to the Zero-Inflated models. In this model, two kinds of zeros are thought to exist in the data: "structural zeros" and "sampling zeros" [11]. The sampling zeros are due to the usual Poisson (or Negative Binomial) distribution,

which assumes that those zero observations happened by chance. Zero-Inflated models assume that the structural zeros are observed due to some specific structure in the data. For example, if a count of high-risk sexual behaviors is the outcome, some participants may score zero because they do not have a sexual partner; these are the structural zeros since they cannot exhibit unprotected sexual behavior. Other participants have sexual partners but score zero because they have eliminated their high-risk behavior. In Zero-Inflated models, the excess zeros (structural zeros) are generated through a separate process. This regression method consists of two models, the generalized count model and the Zero-Inflated models for predicting excess zeros. The excess zeros (structural zeros) are always assigned a probability π and the sampling zeros are assigned a probability 1-π. Two major Zero-Inflated models exist; The Zero-Inflated Poisson model and Zero-Inflated Negative Binomial model.

Zero-Inflated Poisson model: The Zero-Inflated Poisson (ZIP) regression is used for count data that exhibit overdispersion and excess zeros. The data distribution combines the Poisson distribution and the Logit distribution. The possible values of Y are the nonnegative integers: 0,1,2,3, and so on [12].

Suppose that for each observation, there are two possible cases. Suppose that if case 1 occurs, the count is zero. However if case 2 occurs, counts (including zeros) are generated according to a poisson model. Suppose that case 1 occurs with probability π and case 2 occurs with probability 1-π. Therefore, the probability distribution of the Zero-Inflated Poisson random variable y can be written as

$$P\left(Y/Xi\right) = \begin{cases} \pi + \left(1-\pi\right)e^{-\mu} & y = 0 \\ \left(1-\pi\right)\left(\dfrac{\mu^{Y}e^{-\mu}}{Y!}\right) & y > 0 \end{cases}$$

Where π is the logistic link function

The Poisson component can include an exposure time t and a set of k regressor variables (the x's). The expression relating these quantities is

$$\mu = \exp\left(\ln\left(t\right) + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \ldots + \beta_k x_{ki}\right)$$

Often, x_1 = 1, in which case β_1 is called the intercept. The regression coefficients β_1, β_2, β_3,..., β_k are unknown parameters that are estimated from a set of data.

The logistic link function π is given as

$$\pi = \frac{\lambda}{1+\lambda}$$

The mean of the above distribution, $E\left(Y/Xi\right) = \left(1-\pi_i\right)\mu_i$ and the variance $V\left(Y/Xi\right) = \left(1-\pi_i\right)\left(\mu_i + \pi_i\mu_i^2\right)$

Zero-Inflated Negative Binomial model: The Zero-Inflated Negative Binomial (ZINB) regression is used for count data that exhibit overdispersion and excess zeros. The data distribution combines the Negative Binomial distribution and the Logit distribution [13]. The possible values of Y are the nonnegative integers: 0,1,2,3, and so on.

Suppose that for each observation, there are two possible cases. Suppose that if case 1 occurs, the count is zero. However, if case two occurs, counts (including zeros) are generated according to a Negative Binomial model. Suppose that case 1 occurs with probability π and case 2 occurs with probability 1-π. Therefore, the probability distribution of the Zero-Inflated Negative Binomial random variable y can be written as

$$P\left(Y/Xi\right) = \begin{cases} \pi + \left(1-\pi\right)\left(\dfrac{k}{k+\mu}\right)^{k} & y = 0 \\ \left(1-\pi\right)\left(\dfrac{k}{k+\mu}\right)^{k}\dfrac{\Gamma\left(k+Y\right)}{\Gamma\left(Y+1\right)\Gamma k}\left(\dfrac{\mu}{k+\mu}\right)^{Y} & y > 0 \end{cases}$$

The Negative Binomial component can include an exposure time t and a set of k regressor variables (the x's). The expression relating these quantities is

$$\mu = \exp\left(\ln\left(t\right) + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \ldots + \beta_k x_{ki}\right)$$

Often, x_1 = 1, in which case β_1 is called the intercept. The regression coefficients β_1, β_2, β_3,..., β_k are unknown parameters that are estimated from a set of data.

The logistic link function π is given as

$$\pi = \frac{\lambda}{1+\lambda}$$

The mean of the above distribution, $E(Y/Xi) = (1-\pi_i)\mu_i$ and the variance

$$V(Y/X_i) = (1-\pi)\left(\mu_i + \mu_i^2\pi + \frac{\mu_i^2}{K}\right)$$

Where k is a shape parameter which quantifies the amount of overdispersion. This distribution approaches the Zero-Inflated Poisson and the Negative Binomial as $k \to \infty$ and $\pi \to \infty$ respectively.

Hurdle regression models

This is another type of model that is used to model over-dispersed data with excess zeros. This model is like the Zero-Inflated model, only that it assumes that all zero observations are structural zeros. The positive (non-zero) observations have sampling origin, following either Poisson or Negative Binomial distribution. The zero observations can come from only structural source. The two major Hurdle models are the Poisson Hurdle model and the Negative Binomial Hurdle model [14].

Poisson Hurdle regression model: The Poisson Hurdle regression is used for count data that exhibit overdispersion and excess zeros. The data distribution combines the Poisson distribution and the Logit distribution. The possible values of Y are the nonnegative integers: 0,1,2,3, and so on.

Suppose that for each observation, there are two possible cases. Suppose that if case 1 occurs, the count is zero. However if case 2 occurs, counts (excluding zeros) are generated according to a Poisson model. Suppose that case 1 occurs with probability π and case 2 occurs with probability 1-π. Therefore, the probability distribution of the Poisson Hurdle random variable y can be written as

$$P(Y/Xi) = \begin{cases} \pi & Y = 0 \\ (1-\pi)\left(\dfrac{\mu^Y e^{-\mu}}{(1-e^{-\mu})Y!}\right) & Y > 0 \end{cases}$$

Negative Binomial Hurdle regression model: The Negative Binomial Hurdle regression is used for count data that exhibit overdispersion and excess zeros [15]. The data distribution combines the Negative Binomial distribution and the Logit distribution. The possible values of Y are the nonnegative integers: 0,1,2,3, and so on.

Suppose that for each observation, there are two possible cases. Suppose that if case 1 occurs, the count is zero. However if case 2 occurs, counts (excluding zeros) are generated according to a Negative Binomial model. Suppose that case 1 occurs with probability π and case 2 occurs with probability 1-π. Therefore, the probability distribution of the Negative Binomial Hurdle random variable y can be written as

$$P(Y/Xi) = \begin{cases} \pi & Y = 0 \\ (1-\pi)\left(\dfrac{k}{k+\mu}\right)^k \dfrac{\Gamma(k+Y)}{\Gamma(Y+1)\Gamma k}\left(\dfrac{\mu}{k+\mu}\right)^Y \dfrac{1}{1-\left(\dfrac{k}{k+\mu}\right)^k} & Y > 0 \end{cases}$$

Test of goodness of fit

Six different regression models have been discussed above. Therefore, we need to test the goodness of fit for each distribution in order to determine the best model that fits the distribution. The goodness of fit tests used in this research work are the Vuong test and Akaike Information Criterion (AIC) value.

Vuong test: The Vuong test is used to compare a pair of non-nested model to determine which one is better and fit most [16]. It is a likelihood-ratio-based test for model selection. The Vuong test statistic is normally distributed. The test statistic is given as

$$Z = \frac{LR}{\sqrt{N_W}}$$ where $LR = L_1 - L_2 - \dfrac{K_1 - K_2}{2}\log N$, L_1 and L_2 are the corrected maximum likelihood of models 1

and 2 respectively, K_1 and K_2 are the numbers of parameters in models 1 and 2 respectively, w is defined by setting w^2 equal to the mean of the squares of the point-wise log-likelihood ratios.

H_0: The two models are equally close to the true data generating process

Vs

H_1: One model is closer than the other

The criterion for detecting the model that is closer is based on the value of the test statistic. If the test statistic is positive, model 1 is closer, otherwise, model 2 is preferred. The magnitude of the test statistic also is important as it can be used to determine how closer the preferred model is to the true data generating process than the other.

Akaike Information Criterion (AIC): The Akaike Information Criterion was used to determine the best model among the various models that we have, that fits the data. The model with minimum AIC value is considered as the best model for the data [17].

AIC = -2log L ($\hat{\theta}$) + 2k,

Where L ($\hat{\theta}$) is the maximum likelihood function for the estimated model and it offers summary information on how much discrepancy exists between the model and the data, where K is the number of free parameters in the model. AIC accesses both the goodness of fit of the model and the complexity of the model. It rewards the model fit by the maximized log-likelihood term, i.e., -2log L ($\hat{\theta}$), and also prefers a relatively parsimonious model by having K as a measure of complexity.

Data processing and analysis

The data collected were entered into SPSS version 20. The bar chart of the response variable was plotted to determine the shape of the distribution. Base on the bar chart, Vuong test, the goodness of fit test was conducted to determine the best model that fits the data. The AIC value for each model was calculated to further ascertain the best model. The overall best model was the Zero-Inflated Negative Binomial model. The response variable is the number of antenatal visits while the explanatory variables are wealth index, highest educational level, current age, type of place of residence and region. Wealth index, highest educational level, type of place of residence and region are all categorical variables while current age is a count variable. The categorical variables were coded as follow: wealth index (poorest, poorer, middle, richer, richest), highest educational level (no education, primary, secondary, tertiary), type of place of residence (urban, rural) and region (north central, north east, north west, south east, south south, south west). The reference category for the categorical variables are poorest, no education, urban, north central respectively.

The Zero-Inflated Negative Binomial regression model was used to fit the data to see the relationship between the influential factors and the number of antenatal visits. Level of significance and association of variables were tested at 95% Confidence Interval (CI).The independent variables which were not significant were removed from the model and the analysis was run again. The significant models were all included in the final model.

The bar chart was plotted using SPSS. R program was used to carry out the Vuong test, compute the various AIC values and to fit the Zero-Inflated Negative Binomial regression model using the functions voung( ), AIC( ) and zeroinfl( ) respectively.

The fitted model

$$Log(Y) = \beta_0 + \beta_1 A_i + \beta_2 WP_i + \beta_3 WM_i + \beta_4 WRr_i + \beta_5 WRt_i + \beta_6 R_i + \beta_7 \Pr_i +$$
$$\beta_8 S_i + \beta_9 T_i + \beta_{10} NE_i + \beta_{11} NW_i + \beta_{12} SE_i + \beta_{13} SS_i + \beta_{14} SW_i$$

Taking the exponential of both sides gives

$$Y = \exp (\beta_0 + \beta_1 A_i + \beta_2 WP_i + \beta_3 WM_i + \beta_4 WRr_i + \beta_5 WRt_i + \beta_6 R_i + \beta_7 \Pr_i +$$
$$\beta_8 S_i + \beta_9 T_i + \beta_{10} NE_i + \beta_{11} NW_i + \beta_{12} SE_i + \beta_{13} SS_i + \beta_{14} SW_i)$$

Where Y is the response variable,

A_i = Age

WP_i = Poorer Class

WM_i = Middle Class

WRr_i = Richer Class

WRt_i = Richest Class

R_i = Rural

Pr_i = Primary

S_i = Secondary

T_i = Tertiary

NE_i = North East

NW_i = North West

SE_i = South East

SS_i = South South

SW_i = South West

β_0, β_1, β_2, β_3, ..., β_{14} are regression coefficients.

Results

From the bar chart above, it is obvious that the data is zero inflated. The bar for zero is taller than every other bar. We therefore run the appropriate goodness of fit test to determine the best model that fits the data (Figure 1).

The Zero-Inflated Negative Binomial model fits the data best among the class of regression models used in this research work. The Negative Binomial Hurdle is the second choice because it generated the lowest test statistic of 2.199229 and the p-value is 0.013931, which is closer to 0.05 than the p-value of the other models. It was almost competing with the Zero-Inflated Negative Binomial model (Table 1).

In the Table 2 of AIC above, Zero-Inflated Negative Binomial model has the least AIC value. This is an evidence that the Zero-Inflated Negative Binomial model is the best model, among the models used in this research work, that fits the data. The second choice is the Negative Binomial Hurdle model which has the second smallest AIC value.

Figure 1: Bar chart of the response variable (number of antenatal visits).
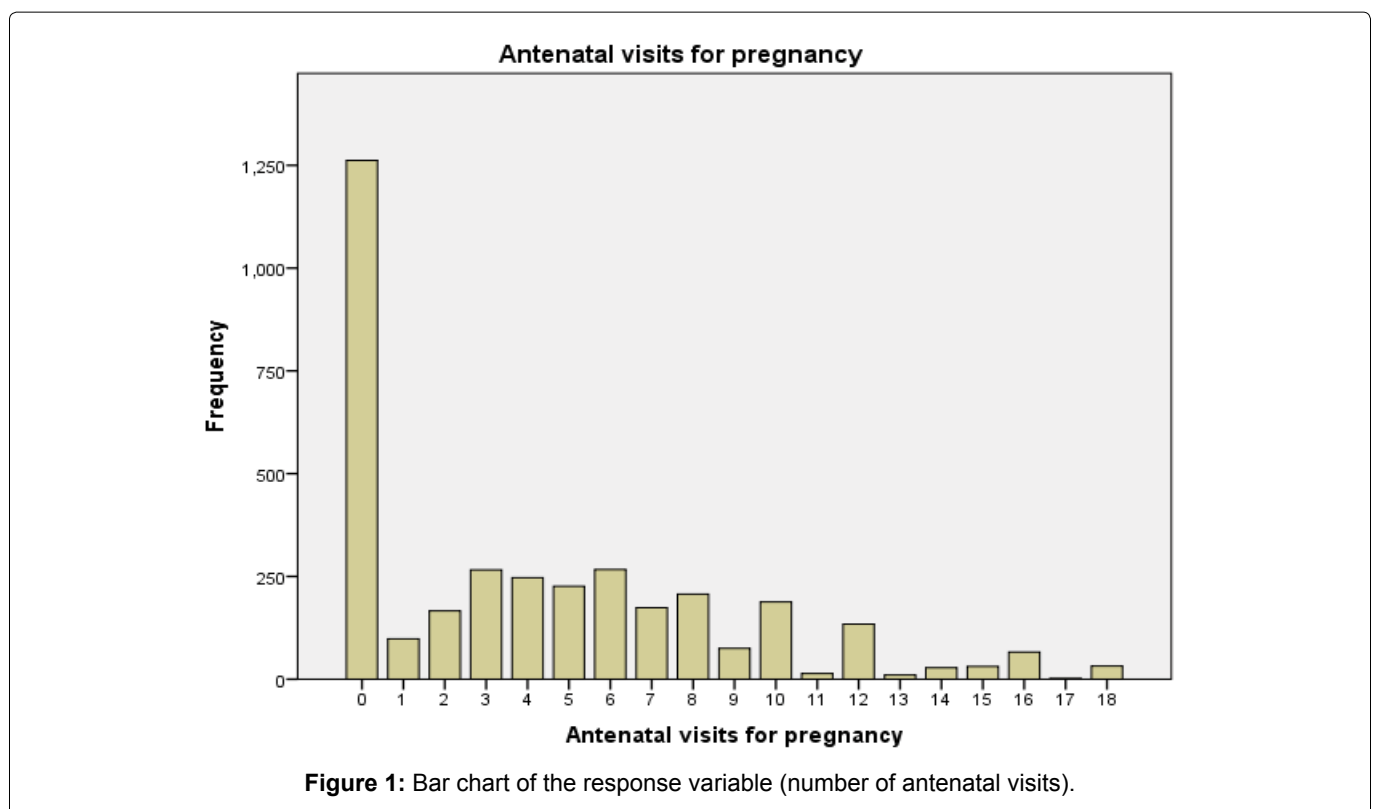
Table 1: Vuong non-nested test results.

Model comparison	Vuong Test Statistic (p-value)	Preferable Model
Poisson Vs. Negative Binomial	-19.33141 (2,22e-16)	Negative binomial
Negative binomial Vs. ZIP	-16.55746 (2,22e-16)	ZIP
ZIP Vs. ZINB	-7.913903 (1.2472e-15)	ZINB
ZINB Vs. Poisson Hurdle	7.990348 (6.6613e-16)	ZINB
ZINB Vs. NB Hurdle	2.199229 (0.013931)	ZINB

Table 2: Model fit comparison using Akaike information criterion (AIC).

	Poisson	NB	ZIP	ZINB	P Hurdle	NB Hurdle
AIC	19848.84	16447.28	14685.94	14392.90	14763.04	14396.75
-2Log-likelihood	19818.84	16415.28	14613.94	14330.90	14703.04	14334.75
Df	15	16	36	31	30	31

Table 3: Descriptive Analysis for all the explanatory variables.

Explanatory variables	Frequency	Percentage
Age	3493	
Wealth Index		
Poorest	836	23.9%
Poorer	758	21.7%
Middle	702	20.1%
Richer	657	18.8%
Richest	540	15.5%
Type of place of residence		
Urban	1201	34.4%
Rural	2292	65.6%
Highest educational level		
No education	1785	51.1%
Primary	838	24.0%
Secondary	742	21.2%
Tertiary	128	3.7%
Region		
North central	597	17.1%
North east	859	24.6%
North west	1106	31.7%
South east	269	7.7%
South south	353	10.1%
South west	309	8.8%

Table 4: Estimated regression coefficients and their corresponding p-value (ZINB count model).

Explanatory variables	Parameter Estimate	P-value
Intercept	1.374462	2e-16
Age	0.006847	2.26e-05
Wealth Index		
Poorest (ref)		
Poorer	0.100488	0.017414
Middle	0.154440	0.000147
Richer	0.255305	5.20e-10
Richest	0.334807	7.10e-14
Type of place of residence		
Urban (ref)		
Rural	-0.019749	0.460765
Highest educational level		
No education (ref)		
Primary	0.017186	0.575848
Secondary	0.134973	4.25e-05
Tertiary	0.241642	3.34e-06
Region		

North central (ref)		
North east	-0.255579	2.38e-11
North west	-0.056276	0.123849
South east	0.056332	0.179101
South south	0.156398	9.27e-05
South west	0.465846	< 2e-16

Table 5: Estimated regression coefficients and their corresponding p-value (ZINB Zero-Inflation model).

Explanatory variables	Parameter Estimates	P-value
Intercept	0.706197	0.00797
Age	-0.013155	0.03290
Wealth Index		
Poorest (ref)		
Poorer	-0.332695	0.00756
Middle	-1.073954	5.34e-16
Richer	-1.691682	< 2e-16
Richest	-3.297337	7.25e-15
Type of place of residence		
Urban (ref)		
Rural	0.176345	0.17017
Highest educational level		
No education (ref)		
Primary	-1.266913	< 2e-16
Secondary	-1.758563	< 2e-16
Tertiary	-2.894437	0.00468
Region		
North central (ref)		
North east	0.328485	0.01975
North west	1.104337	1.08e-15
South east	-3.214429	9.31e-05
South south	0.340919	0.09149
South west	-1.865211	4.84e-08

AIC value for Negative Binomial Hurdle model is very close to that of Zero-Inflated Negative Binomial model. It is almost competing with the Zero-Inflated Negative Binomial model. The poorest model for this data is the Poisson model, which has the highest AIC value. The result gotten here is perfect with that of the Voung test.

The Table 3 above presents the descriptive analysis of respondents. Among the respondents of about 3493, 836 (23.9%) belong to the poorest household, 758 (21.7%) belong to the poorer household, 702 (20.1%) were in the middle class, 657 (18.8%) were in the richer class, and 540 (15.5%) were in the richest household. Respondents that resided in the urban areas are found to be 1201 (34.4%), while 2292 (65.6%) were in the rural areas. A larger number of the respondents resided in the rural areas. Findings show that greater number of the respondents had no education 1785 (51.1%), 838 (24.0%) had primary education, 742 (21.2%) had secondary education, and 128 (3.7%) had tertiary education. Respondents from north central were found

to be 597 (17.1%), 459 (24.6%) were from north east, 1106 (31.7%) were from north west, 269 (7.7%) were from south east, 353 (10.1%) were from south south, 309 (8.8%) were from south west.

The Table 4 above gives the regression coefficients of the count model. Findings show that the respondents' types of place of residence do not contribute significantly to the model. Therefore, it is removed from the model.

The Table 5 above gives the regression coefficients of the Zero-Inflation model. Findings show that the respondents' types of place of residence do not contribute significantly to the model. Therefore, it is removed from the model.

The results above show that as age increases, the number of antenatal visits increases with 1.00687 (Table 6). Findings also show that women from poorer households are likely to have 11% (OR = 0.11) increase in number of antenatal visits compared to women from the poorest households, Middle class households were 17%

Table 6: The odds ratio and p-value of the reduced ZINB count model.

Explanatory variables	Odd Ratio(OR)	P-value
Intercept	3.95295	2e-16
Age	1.00687	2.26e-05
Wealth Index		
Poorest	1	
Poorer	1.10571	0.017414
Middle	1.16700	0.000147
Richer	1.29086	5.20e-10
Richest	1.39767	7.10e-14
Highest educational level		
No education	1	
Primary	1.01733	0.575848
Secondary	1.14451	4.25e-05
Tertiary	1.27334	3.34e-06
Region		
North central	1	
North east	0.77447	2.38e-11
North west	0.94528	0.123849
South east	1.05795	0.179101
South south	1.16929	9.27e-05
South west	1.59336	< 2e-16

(OR = 0.17) more likely to visit maternity for antenatal care. Women from richer households were 29% (OR = 0.29) more likely to go for antenatal care, while women from the richest households were 40% (OR = 0.4) more likely to have more antenatal visits. Thus, as wealth increases, the preference for more antenatal visits increases. The number of antenatal visits is 2% (OR = 0.02) higher among respondents with primary education compared to those with no education. Women with secondary education were 15% (OR = 0.15) likely to have more children, the number of antenatal visits is 27% (OR = 0.27) more for women with tertiary education. Therefore, the higher the educational level, the higher the desire for more antenatal visits. The number of antenatal visits of women in the north east is 23% (OR = 0.23) less compared with women in the north central. Women in the northwest were 6% (OR= 0.06) less likely to go for antenatal care. South eastern women were 6% (OR = 0.06) more likely to go for antenatal care. Women in the South South were 17% (OR = 0.17) more likely to have more antenatal visits. The South western women were 59% (OR = 0.59) more likely to visit maternity for antenatal care. Therefore, the women in the South were more likely to have more antenatal visits. If all the predictor variables in the model are evaluated at zero, the predicted number of antenatal visits would be 3.953 which is approximately 4.

As age increases, the odds of women having zero antenatal visits increase by 0.98693. Women from poorer households were 28% (OR = 0.28) less likely to have zero number of antenatal visits compared to women from the poorest households. Middle class households were 66% (OR = 0.66) less likely to have zero number of antenatal visits. Women from richer households were 82% (OR = 0.82) less likely to have zero number of antenatal visits while women from the richest households were 96% (OR = 0.96) less likely to have zero number of antenatal visits. Thus, as wealth increases, the likelihood of having zero number of antenatal visits decreases. Women with primary education were 72% (OR = 0.72) less likely to have zero number of antenatal visits compared to women with no education. Women with secondary education were 83% (OR = 0.83) less likely to have zero number of antenatal visits. Women with tertiary education were 95% (OR = 0.95) less likely to have zero number of antenatal visits. Thus, the more educated a woman is, the less likely she has zero number of antenatal visits. Women in the north east were 39% (OR = 0.39) more likely to have zero number of antenatal visits compared with women in the north central. Women in the North West were 202% (OR = 2.02) more likely to have zero number of antenatal visits. South eastern women were 96% (OR = 0.96) less likely to have zero number of antenatal visits. Women in the south South were 41% (OR = 0.41) more likely to have zero number of antenatal visits. The South western women were 85% (OR = 0.85) less likely to have zero number of antenatal visits. Thus, northern women were more likely to have zero number of antenatal visits. If all the predictor variables in the model are evaluated at zero, the odds of women having zero number of antenatal visits 2.0263 which is approximately 2 (Table 7).

Table 7: The odds ratio and p-value of the reduced ZINB Zero-Inflation model.

Explanatory variables	Odds Ratio(OR)	P-value
Intercept	2.02627	0.00797
Age	0.98693	0.03290
Wealth Index		
Poorest	1	
Poorer	0.71699	0.00756
Middle	0.34166	5.34e-16
Richer	0.18421	< 2e-16
Richest	0.03698	7.25e-15
Highest educational level		
No education	1	
Primary	0.28170	< 2e-16
Secondary	0.17229	< 2e-16
Tertiary	0.05533	0.00468
Region		
North central	1	
North east	1.38886	0.01975
North west	3.01722	1.08e-15
South east	0.04018	9.31e-05
South south	1.40624	0.09149
South west	0.15486	4.84e-08

Discussion

In this study, the aim was to determine the factors that affect the frequency of antenatal visits in Nigeria. This could be achieved by fitting a regression model. Due to over-dispersion and presence of excess zeros in the data, regular count models, Zero-Inflated models and Hurdle models were compared to determine the one that fit the data most. Two different goodness of fit test were used. The result of Akaike Information Criterion (AIC) shows that the Zero-Inflated Negative Binomial model fit the data most. The Vuong test further confirmed the choice of Zero-Inflated Negative Binomial model as the best distribution model that fits the data.

Regression analysis was run on the data using Zero-Inflated Negative Binomial model. The result shows that as age increases, the number of antenatal visits also increases. As women grow older, their need and desire for more antenatal visits when pregnant increases. This could be as a result of complications due to old age. Also women from richer families go for more antenatal visits than women from poorer families. This probably is as a result of financial capability women from richer families have, to use antenatal care services. The result also shows that as level of education of women increases, the number of antenatal visits increases. So the level of education of women also affects their attitude to antenatal care services. More educated women tend to turn out for antenatal care services than their less educated counterpart. The result shows that women in the northern part of the country are less likely to visit maternity for antenatal care than women in the southern part of the country. This is so due to high number of uneducated women and high level of poverty in the northern part of the country. Types of place of residence do not significantly affect the number of antenatal visits.

Conclusions

Antenatal care is very important for pregnant women in order to avoid complications that come with pregnancy and child birth. Unfortunately, many pregnant women do not utilize the antenatal care services. There is need to determine the factors that affect the frequency of antenatal visits. In this study, data on antenatal visits were gotten from the demographic and health survey. Poisson distribution, Negative Binomial distribution, Zero-Inflated Poisson distribution, Zero-Inflated Negative Binomial distribution, Poisson Hurdle distribution, Negative Binomial Hurdle distribution were subjected to goodness of fit test, using Vuong test and Akaike Information Criterion(AIC), to determine the one that fit the data most.

The best model was Zero-Inflated Negative Binomial model and it was used to model the data. The result shows that age, level of education, wealth status, region of women affect the frequency of antenatal visits.

References

1. Tessema GA, Laurence CO, Melaku YA, Misganaw A, Woldie SA, et al. (2017) Trends and causes of maternal mortality in Ethiopia during 1990-2013: Findings from the

Akinjisola and Jolayemi. Int J Clin Biostat Biom 2024, 10:055

• Page 11 of 12 •

Global Burden of Diseases study 2013. BMC Public Health 17: 160.

2. Okedo-Alex IN, Akamike IC, Ezeanosike OB, Uneke CJ (2019) Determinants of antenatal care utilisation in sub-Saharan Africa: A systematic review. BMJ Open 9: e031890.

3. Hijazi HH, Alyahya MS, Sindiani AM, Saqan RS, Okour AM (2018) Determinants of antenatal care attendance among women residing in highly disadvantaged communities in northern Jordan: A cross-sectional study. Reprod Health 15: 106.

4. Respress ET, Jolly PE, Osia C, Williams ND, Sakhuja S, et al. (2017) A cross-sectional study of antenatal care attendance among pregnant women in Western Jamaica. J Pregnancy Child Health 4: 341.

5. Warri D, George A (2020) Perceptions of pregnant women of reasons for late initiation of antenatal care: A qualitative interview study. BMC Pregnancy Childbirth 20: 70.

6. Muoka AK, Ngesa OO, Waititu AG (2016) Statistical models for count data. Science Journal of Applied Mathematics and Statistics 4: 256-262.

7. Lambert D (1992) Zero-inflated Poisson regression with an application to defects in manufacturing. Technometrics 34: 1-14.

8. Mullahy J (1986) Specification and testing of some modified count data models. Journal of Econometrics 33: 341-365.

9. Imrey PB (2000) 14 - Poisson Regression, Logistic Regression, and Loglinear Models for Random Counts, Handbook of Applied Multivariate Statistics and Mathematical Modeling 2000: 391-437.

10. Joe H, Zhu R (2005) Generalized poisson distribution: The property of mixture of poisson and comparison with negative binomial distribution. Biom J 47: 219-229.

11. Tang W, He H, Wang WJ, Chen DG (2018) Untangle the structural and random zeros in statistical modelings. J Appl Stat 45: 1714-1733.

12. Monod A (2014) Random effects modeling and the zero-inflated poisson distribution. Communications in Statistics - Theory and Methods 43: 664-680.

13. Mwalili SM, Lesaffre E, Declerck D (2008) The zero-inflated negative binomial regression model with correction for misclassification: An example in caries research. Stat Methods Med Res 17: 123-139.

14. Feng CX (2021) A comparison of zero-inflated and hurdle models for modeling zero-inflated count data. J Stat Distrib Appl 8: 8.

15. Bhaktha N (2018) Properties of hurdle negative binomial models for zero-inflated and overdispersed count data [Doctoral dissertation, Ohio State University]. OhioLINK Electronic Theses and Dissertations Center.

16. Schneider L, Chalmers RP, Debelak R, Merkle EC (2020) Model selection of nested and non-nested item response models using Vuong tests. Multivariate Behav Res 55: 664-684.

17. Arnold TW (2010) Uninformative parameters and model selection using Akaike's information criterion. The Journal of Wildlife Management 74: 1175-1178.