Journal of Big Data

RESEARCH Open Access

Bayesian negative binomial logit hurdle and zero-inflated model for characterizing smoking intensity

Mekuanint Simeneh Workie[1*] and Abebaw Gedef Azene[2]

*Correspondence: mekuanintsimeneh@gmail.com
[1] Department of Mathematical and Statistical Modeling (Statistics), Bahir Dar Institute of Technology-Bahir Dar University, Bahir Dar, Ethiopia
Full list of author information is available at the end of the article

Abstract

Smoking invariably has environmental, social, economic and health consequences in Ethiopia. Reducing and quitting cigarette smoking improves individual health and increases available household funds for education, food and better economic productivity. Therefore, this study aimed to apply the Bayesian negative binomial logit hurdle and zero-inflated model to determine associated factors of the number of cigarette smokers per day using the smoking intensity data of 2016 Ethiopia Demographic and Health Survey. The survey was a community-based cross-sectional study conducted from January 18 to June 27, 2016. The survey used two stage stratified sampling design. Bayesian analysis of Negative Binomial Logit Hurdle and Zero-inflated models which incorporate both overdispersion and excess zeros and carry out estimation using Markov Chain Monte Carlo techniques. About 94.2% of them never cigarettes smoked per day and the data were found to have excess zeros and overdispersion. Therefore, after considering both the zero counts and the enduring overdispersion, according to the AIC and Vuong tests, the Zero-inflated Negative Binomial and Negative Binomial Logit Hurdle model best fit to the data. The finding Bayesian estimation technique is more robust and precisely due to that it is more popular data analysis method. Furthermore; using Bayesian Zero-inflation and Zero hurdle model the variable: age, residence, education level, internet use, wealth index, marital status, chewed chat, occupation, the media were the most statistically significant determinate factors on the smoking intensity.

Keywords: Bayesian approach, Zero-inflated regression, Smoking, Markov chain Monte Carlo

Background

Smoking is currently considered one of the greatest problems in public health worldwide. It is one of the most preventable causes of death and a major known cause of noncommunicable diseases [1, 2]. Smoking invariably has social, economic, health, and environmental consequences in Ethiopia. Reducing cigarette smoking improves individual health and increases available household funds for food, education, and better economic productivity [3].

Globally, over 7 million people die each year due to smoking related to this, over 86% of deaths are from direct tobacco use, while around 13% are due to secondhand smoke. The low and middle economical income countries are especially facing an increasing burden of tobacco-related diseases. Africa, particularly sub-Saharan Africa, is experiencing increasing tobacco use [4]. Among Ethiopian men who smoke cigarettes daily, one-quarter (25%) smoke 5–9 cigarettes each day; 6% of daily cigarette smokers smoke 25 or more cigarettes each day. The percentage of men age 15–49 who don't smoke cigarettes has increased slightly since 2011, from 93 to 95% [5]. Despite the presence of studies on tobacco use and associated factors, no study assesses the smoking intensity among smokers. Evidence of smoking intensity is very important for planning and implementing behavior change interventions within the country.

Several studies have been conducted to determine related risk factors of smoking [6, 7]. However, few studies can be found on investigating the intensity of smoking in Ethiopia, which highlights the importance of considering the number of cigarettes smoked per day as a count response variable and investigating. There were no studies which conducted by considering the excess zeros and over dispersion on this outcome variable in Ethiopia. The analysis of count data with many zeros can be done in various fields, including medical and public health. Several models have been developed in recent decades to analyze this count data, especially smoking intensity data [8]. The count data are different from the categorization of count data to be used in crude rate and logistic regression will lead to loss of information.

Furthermore, treating count data is different from ordinal data, continuous variable in linear regression or dichotomous variable in logistic regression models is likely to bias the results [9]. The a Poisson regression or negative binomial regressions are commonly used to model count outcomes assuming Poisson distribution or negative binomial distribution. But the probability of zeroes based on Poisson regression or negative binomial regression cannot account for excess zero counts because excess zeroes will bias the estimation of parameters [10, 11]. Therefore, Hurdle and zero-inflated count models have been the best model so far to solve this issue concerning excess zeroes [12]. The choice of the model either of two is often depend on the nature of the problem and the data collection procedures [13]. The difference is hurdle model includes a mass at zero and a truncated distribution whereas the zero-inflated model is based on a mass at zero and a regular distribution, the inferential results is often very similar. Zero-inflated models and hurdle models provide a way of modeling the excessive proportion of zero values and allow for overdispersion [14, 15].

Consequently; the Negative Binomial–Logit Hurdle (NBLH) and Zero-inflated Negative Binomial (ZINB) Regression model are flexible models for dealing with zero-inflated and over dispersed count data [16]. Bayesian parameter estimation methods can be applied by the Markov chain Monte Carlo (MCMC) simulation that can generate random values with the Gibbs-sampling algorithm. Hence the Bayesian method is more flexible for parameters estimation [17]. Therefore, this study aimed to apply the Bayesian NBLH and Bayesian ZINB regression model to determine the associated factors of the number of cigarettes smoking per day and to identify the optimum model using the smoking intensity data of the 2016 Ethiopia Demographic and Health Survey (EDHS). The finding of this study intended to fill the gap of the model for excess zeros and over dispersed data. This study may use an alarm for policymakers by identifying the associated factors of smoking intensity per day and also it intended to bring to the improvement of the study design and participants of future studies.

Method

Source of data and study design

The dataset used for this study was obtained from 2016 EDHS which conducted from January 18 to June 27, 2016. The survey was a population-based cross-sectional study. The survey used two stage stratified sampling methods. In the first stage, a total of 645 clusters was randomly selected proportional to the household size from the sampling strata, and in the second stage, 28 households per cluster were selected using systematic random sampling. In this survey, a total of 12,645 smoking intensity people selected from 645 clusters.

Variables of the study

The dependent variable of this study was the number of cigarette smokers per day (the Smoking intensity is the number of cigarette sticks smoked in the last 24 h). Since a significant number of the daily smokers used manufactured cigarettes, this study adopted the number of manufactured cigarettes the participants used per day as an outcome variable to assess the smoking intensity in Ethiopia.

Independent variable

Age, marital status, residence, educational level, religion, occupational status, wealth index, frequency of media use, current chat chewing behavior and Internet use were the independent variables.

Statistical method

In this study, the variable of interest was count data. When the dependent variable is a count, it is appropriate to use non-linear models to describe the relationship between the dependent variable and a set of predictor variables. For count data, the standard framework for explaining the relationship between the dependent variable and a set of explanatory variables includes the Poisson, negative binomial regression, zero inflated and hurdle models. The advanced models for this study count data are the Bayesian zero-inflated regression model and Bayesian hurdle model [18]. In this study, excess zeros and overdispersion in the number of cigarettes smoked per day data exist. Therefore, the Negative binomial model cannot be used to handle the high number of zeros. To do this, four models are discussed that can deal with the excessive number of zeros, namely, the zero-inflated Poisson, Zero-inflated negative binomial, Poisson logit hurdle and negative binomial logit hurdle models can be alternatively used [19].

Zero inflated regression model

In Poisson data with too many zeros, the variance often exceeds the mean, causing overdispersion. Although the ZIP regression model can handle excess zeros for Poisson data, overdispersion may remain even after modeling excess zeros, and consequently ZIP parameter estimates can be severely biased. In such a case, the use of a zero-inflated negative binomial

(ZINB) distribution can be a good alternative regression model [13, 20]. Therefore, the ZIP regression model is more effective for excess zero outcomes than Poisson regression. While the ZINB regression model is more effective for excess zero outcomes in additional overdispersion than negative binomial regression model.

Zero-inflated Poisson regression model

The most commonly employed approach to dealing with heterogeneity associated with excess zeros, is to use a Poisson distribution that has been mixed with a point mass at zero to allow for the inclusion of additional structural zeros. In ZIP regression, the excess zero counts are assumed to occur with probability ω_i and follow a Poisson distribution with mean μ_i, with probability $1 - \omega_i$ where i = 0, 1, 2,..., n. ZIP model can thus be seen as a mixture of two-component distributions, a zero part, and no-zero components, given by [12]:

$$\mathrm{P}\left(\mathrm{Y}=\mathrm{y}\right) = \begin{cases} \omega_i + (1-\omega_i)e^{-\mu_i}, & y_{i=0} \\ (1-\omega_i)\frac{e^{-\mu_i}\mu_i^{y_i}}{y_{i!}}, & y_i = 1, 2, 3, \ldots.. \end{cases} \tag{1}$$

The first part of the Eq. (1) above is the zero part of the model and the second part is the non-zero count's part of the model. The mean and variance of the zero-inflated Poisson model are: $E(Yi) = \mu_i(1\text{-}\omega_i)$ and Var $(Yi) = \mu_i\,(1-\omega_i)(1+\mu_i\omega_i)$;

The most natural choice to model the probability of excess zeros is to use a logistic regression model with a logit link and count data excluding excess zeros can be modeled through Poisson regression:

$$\text{Logit }(\omega_i) = x_i^T\beta \text{ and Log }(\mu_i) = Z_i^T\gamma \tag{2}$$

where $x_i{}^T$ represents a vector of covariates and $\boldsymbol{\beta}$ a vector of parameters and γ are the vector coefficients Z_i^T.

Zero-inflated negative binomial regression model

The ZINB model is especially suited to dealing with both over dispersed and zero-inflated data. The ZINB distribution is a mixture distribution assigning a mass of ω_i to 'extra' zeros and a mass of $(1 - \omega_i)$ to a negative binomial distribution, where $0 \leq \omega_i \leq 1$. Based on the probability function of the zero-modified distribution, then the probability mass function for the ZINB regression model is:

$$\mathrm{P}(\mathrm{Y_i} = y_i) = \begin{cases} \omega_i + (1-\omega_i)\left(\dfrac{\phi}{\mu_i+\phi}\right)^{\phi}, & y_i = 0 \\ (1-\omega_i)\dfrac{\Gamma\left(y+\phi\right)}{\Gamma\left(\mathrm{y}+1\right)\Gamma(\phi)}\left(\dfrac{\phi}{\mu_i+\phi}\right)^{\phi}\left(\dfrac{\mu_i}{\mu_i+\phi}\right)^{y_i}, & y_i = 1, 2, 3\ldots \end{cases} \tag{3}$$

where ϕ^{-1}, μ_i and $\Gamma(.)$ representing dispersion, mean, and gamma function respectively. The mean and variance of the zero-inflated negative binomial regression model are:

$$\mathrm{E}\ (y_i) = (1-\omega_i)\mu_i \text{ and } \mathrm{Var}(y_i) = (1-\omega_i)\mu_i(1+\ \emptyset\mu_i + \omega_i\mu_i)$$

Assume that there are predictors for logistic regression function and negative binomial regression function. Hence, the ZINB regression model can be written as follow:

$$\text{Logit}\ (\omega_i) = x_i^T \beta \text{ and Log } (\mu_i) = Z_i^T \gamma \tag{4}$$

where β are the vector coefficients X_i^T and γ are the vector coefficients Z_i^T. The standard estimation technique for the ZIP and the ZINB is based on the maximum likelihood estimation. The Newton–Raphson algorithm can be used to maximize the log-likelihood functions [12].

Hurdle model

A hurdle model is a class of statistical models where a random variable is modelled using two parts, the first which is the probability of attaining value zero (a point mass at zero), and the second part models the probability of the non-zero values (a distribution that generates non-zero counts) from a zero-truncated distribution. The use of hurdle models is often motivated by an excess of zeros in the data, that is not sufficiently accounted for in more standard statistical models [4]. All zeros in the hurdle model are assumed to be "structural" zeros, i.e., they are generated from a single process, and are observed since the condition is absent. We explore two zero-truncated count distributions for the hurdle model specification [18]. The Hurdle Model of count data can be expressed as follows for the Poisson and Negative Binomial distribution. We consider a Poisson hurdle regression model in which the response variable y has the distribution:

$$\mathrm{P}\left(\mathrm{Y_i} = \mathrm{y_i}\right) = \begin{cases} \omega_i, & y_{i=0} \\ (1-\omega_i)\frac{e^{-\mu_i}\mu_i^{y_i}}{y_{i!}\left(1-e^{-\mu_i}\right)}, & y_i = 1, 2, 3, \ldots \end{cases} \tag{5}$$

A negative binomial hurdle distribution is given by:

$$\mathrm{P}\left(\mathrm{Y_i} = \mathrm{y_i}\right) = \begin{cases} \omega_i, & y_i = 0 \\ (1-\omega_i)\frac{\Gamma(y_i+\phi)}{\Gamma(y_i+1)\Gamma(\phi)}\frac{\mu^{y_i}\phi^{y_i}(1+\mu\phi)^{-(y_i+\phi)}}{1-(1+\emptyset\mu)^{\phi}}, & y_i = 1, 2, 3 \ldots \end{cases} \tag{6}$$

where $\phi \geq 0$ is a dispersion parameter that is assumed not to depend on covariates. Zero and truncated hurdle model:

$$\text{Logit}\ (\omega_i) = x_i^T \beta \text{ and Log } (\mu_i) = Z_i^T \gamma \tag{7}$$

where β are the vector coefficients X_i^T and γ are the vector coefficients Z_i^T. The parameter $\varnothing$ is a measure of dispersion.

Tests for ZIP and ZINB regression models

The test of overdispersion in ZIP regression model against ZINB alternatives, H_0: $a=0$ vs. H_1: $a>0$, can be performed using likelihood ratio test (LRT), $T=2(\ln L1-\ln L0)$, where $\ln L_1$ and $\ln L_0$ are the models 'log-likelihood under their respective hypotheses. Since the null hypothesis is on the boundary of parameter space, the LRT statistic is asymptotically distributed as a half of probability mass at zero and a half of chi-square with one degree of freedom. The Maximum Likelihood Estimation (MLE) method is used to estimate parameters in the count models. This study includes ZIP, ZINB, Hurdle Poisson, and NBLH to accommodate the excess zeros for the number of cigarettes smoked per day count data. In this paper, Akaike's information criteria (AIC) and log-likelihood values are used for model selection measures. It is also used dispersion parameters to test for overdispersion. AIC and log-likelihood are basic methods for assessing the performance of the models and model selection [12]. Feature Selection plays a very important role in machine learning algorithms. The feature selection of techniques tunes certain parameters to select only few features which are most essential and relevant and far away from the redundant information [21]. But in this article, one of important count regression variable selection by using forward variable selection algorithm.

Bayesian ZINB and negative binomial–logit hurdle model

The number of cigarette smokers daily is a count variable. For the modeling of count data, two-part models are applied in the presence of excessive zeros. Therefore, for a better fit an over-dispersed model that incorporates excessive zeros, i.e., the ZINB regression model is used. If the data is under dispersed, the ZINB model should not be used. The hurdle model is also flexible and can handle both under-dispersion and over-dispersion problem. The NBLH model is used on data with either excessive zero counts in the response or at times too few zero counts. In the case where there are too few zero counts, a zero-inflated model cannot be used. The hurdle model is a good way to deal with such data [18]. It uses two-part. The first part estimates zero elements the dependent variable are zero hurdle model and the second part estimates not zero elements (non-negative integer) from the dependent variable is called truncated negative binomial models [22]. Suppose that Y_1,..., Y_n is a random sample from either ZINB or NBLH distribution. Then the probability mass functions of ZINB or NBLH regression models are given respectively by:

$$\mathrm{P(Y_i =)} = y_i \begin{cases} \omega_i + (1-\omega_i)\left(\frac{\phi}{\mu_i+\phi}\right)^{\phi}, & y_i = 0 \\ (1-\omega_i)\frac{\Gamma(y+\phi)}{\Gamma(\mathrm{y}+1)\Gamma(\phi)}\left(\frac{\phi}{\mu_i+\phi}\right)^{\phi}\left(\frac{\mu_i}{\mu_i+\phi}\right)^{y_i}, & y_i = 1,2,3\ldots \end{cases} \tag{8}$$

$$\mathrm{P}\big(\mathrm{Yi} = \mathrm{yi}\big) = \begin{cases} \omega_i, & y_i = 0 \\ (1-\omega_i)\frac{\Gamma(y_i+\phi)}{\Gamma(y_i+1)\Gamma(\phi)}\frac{\mu^{y_i}\phi^{y_i}(1+\mu\phi)^{-(y_i+\phi)}}{1-(1+\emptyset\mu)^{\phi}}, & y_i = 1,2,3\ldots \end{cases} \tag{9}$$

where ϕ μ, and Γ(.) representing the dispersion parameter, mean, and gamma function, respectively. The most natural choice to model the probability of excess zeros is to use the logit link function and a negative binomial and truncated negative binomial model with log link function respectively.

$$\text{Logit } (\omega_i) = X_i^T\beta \text{ and Log } (\mu_i) = Z_i^T\gamma \tag{10}$$

where β are the vector coefficients X_i^T and γ are the vector coefficients Z_i^T. The parameter ∅ is a measure of dispersion. When ∅ = 0, the NBLH and ZINB model reduces to the Poisson regression model. For ∅ > 0, the NBLH and ZINB models can be used to fit over dispersed count data. When ∅ < 0, the NBLH model can be used to fit under dispersed count data. Suppose that Y_1,..., Y_n is a random sample from either a ZINB or NBLH distribution. Then the likelihood functions of the two models are given respectively by

$$L_1(\beta,\gamma,\emptyset) = \prod_{i=1}^{n}\left[I\left(y_i=0\right)\left\{\omega_i+(1-\omega_i)\left(\frac{\phi}{\mu_i+\phi}\right)^{\phi}\right\}\right. \\ \left. +I\left(y_i>0\right)\left\{(1-\omega_i)\frac{\Gamma\left(y+\phi\right)}{\Gamma\left(\mathrm{y}+1\right)\Gamma(\phi)}\left(\frac{\phi}{\mu_i+\phi}\right)^{\phi}\left(\frac{\mu_i}{\mu_i+\phi}\right)^{y_i}\right\}\right] \tag{11}$$

$$L_2(\beta,\gamma,\emptyset) = \prod_{i=1}^{n}\left[I\left(y_i=0\right)\{\omega_i\}+I\left(y_i>0\right)\left\{(1-\omega_i)\frac{\Gamma\left(y_i+\phi\right)}{\Gamma\left(y_i+1\right)\Gamma(\phi)}\frac{\mu^{y_i}\phi^{y_i}(1+\mu\phi)^{-(y_i+\phi)}}{1-(1+\emptyset\mu)^{\phi}}\right\}\right] \tag{12}$$

Let β and γ are the set of parameters for the above-mentioned model. We assume independent priors for these parameters**.** Since there is no prior information from historical data or previous experiments, then all parameters will use conjugate non-informative priors. The prior distribution for β and γ is assumed to be normal, while ϕ is assumed to be gamma-distributed. So, the joint prior distribution for either ZINB or NBLH regression parameters is:

$$f(\beta,\gamma,\emptyset) = f(\beta)\times f(\gamma)\times f(\emptyset)$$

$$f(\beta,\gamma,\emptyset) = \prod_{j=1}^{p}\left[\frac{1}{\sigma_{\beta_j}\sqrt{2\pi}}e^{-\frac{\left[\beta_j-\mu_{\beta_j}\right]^2}{2\sigma_{\beta_j}^2}}\right]\times\prod_{k=1}^{p}\left[\frac{1}{\sigma_{\gamma_k}\sqrt{2\pi}}e^{-\frac{\left[\gamma_k-\mu_{\gamma_k}\right]^2}{2\sigma_{\gamma_k}^2}}\right]\times\frac{1}{b^a\Gamma(\alpha)}\emptyset^{a-1}e^{-\frac{\emptyset}{b}} \tag{13}$$

where ∅~Gamma (a, b) with a = 0.001 and b = 0.001. but our a priori judgment was that knowledge of the slope parameter $\boldsymbol{\gamma}$ does not provide any information about β. Given that the prior distribution for parameters has been assessed, the next procedure is to combined the likelihood function prior to performing Bayesian inference. The posterior distribution can be written as

$$\mathrm{P}(\theta/y)\propto L\left(\theta;y\right)p(\theta)$$

where $\theta=(\beta,\gamma,\varnothing,)$, $L(\theta;y,X)$ are the likelihood function and $p(\theta)$ is the prior distribution. Then the fully conditional posterior distributions of parameters on the ZINB are given by:

$$P_1\left(\upbeta,\upgamma,\emptyset,\mathrm{y\ X}\right)\propto\prod_{i=1}^{n}\left[\begin{array}{l} I\left(y_i=0\right)\{\omega_i+(1-\omega_i)\left(\frac{\phi}{\mu_i+\phi}\right)^{\phi}\}+ \\ I\left(y_i>0\right)\left\{(1-\omega_i)\frac{\Gamma(y+\phi)}{\Gamma(\mathrm{y}+1)\Gamma(\phi)}\left(\frac{\phi}{\mu_i+\phi}\right)^{\phi}\left(\frac{\mu_i}{\mu_i+\phi}\right)^{y_i}\right\}\end{array}\right] \\ \times\prod_{j=1}^{p}\left[\frac{1}{\sigma_{\beta_j}\sqrt{2\pi}}e^{-\frac{\left[\beta_j-\mu_{\beta_j}\right]^2}{2\sigma_{\beta_j}^2}}\right]\times\prod_{k=1}^{p}\left[\frac{1}{\sigma_{\gamma_k}\sqrt{2\pi}}e^{-\frac{\left[\gamma_k-\mu_{\gamma_k}\right]^2}{2\sigma_{\gamma_k}^2}}\right] \\ \times\frac{1}{b^a\Gamma(\alpha)}\emptyset^{a-1}e^{-\frac{\emptyset}{b}} \tag{14}$$

The fully conditional posterior distributions of the parameter on the NBLH are given by $P_2\left(\beta, \gamma, \varnothing, \mathrm{y}\ \mathrm{X}\right)$

$$\begin{aligned}
&\propto \prod_{i=1}^{n}\left[I\left(y_i=0\right)\{\omega_i\}+I\left(y_i>0\right)\left\{(1-\omega_i)\frac{\Gamma\left(y_i+\phi\right)}{\Gamma\left(y_i+1\right)\Gamma(\phi)}\frac{\mu^{y_i}\phi^{y_i}(1+\mu\phi)^{-(y_i+\phi)}}{1-(1+\emptyset\mu)^{\phi}}\right\}\right]\\
&\times\prod_{j=1}^{p}\left[\frac{1}{\sigma_{\beta_j}\sqrt{2\pi}}e^{-\frac{\left[\beta_j-\mu_{\beta_j}\right]^2}{2\sigma^2_{\beta_j}}}\right]\times\prod_{k=1}^{p}\left[\frac{1}{\sigma_{\gamma_k}\sqrt{2\pi}}e^{-\frac{\left[\gamma_k-\mu_{\gamma_k}\right]^2}{2\sigma^2_{\gamma_k}}}\right]\\
&\frac{1}{b^a\Gamma(\alpha)}\emptyset^{a-1}e^{-\frac{\emptyset}{b}}
\end{aligned} \tag{15}$$

This procedure is implemented using the MCMC algorithm. The MCMC method is a general simulation method for sampling from posterior distributions and computing posterior quantities of interest. Each sample depends on the previous one, hence the notion of the Markov chain. The Markov chain method has been quite successful in modern Bayesian computing [23]. The two most common Markov Chain Monte Carlo algorithms are the Metropolis–Hastings and the Gibbs samplers. These two algorithms have many variants and extensions that have been developed. These forms and extensions are more advanced and sometimes more peculiar to some problems. In this section, we discuss in detail the Gibbs sampler together with its variants and extensions. The Gibbs sampling algorithm is a special case of the Metropolis–Hastings algorithm in which parameters are drawn from distributions with a 100% acceptance rate. It is an alternating conditional sampling. The joint posterior distribution is decomposed into a sequence of simpler conditional distributions, where the target is to generate a data point from the conditional distribution of each parameter, conditional on the current values of the other parameter. The Gibbs algorithm implements MCMC methods using the R software.

The convergence of the algorithm

Flexible software for Bayesian analysis of complex statistical models by using MCMC methods. We use these tools to estimate the ZINB and NBLH regression models. MCMC is based on a combination of Markov chain and Monte Carlo estimation which eventually converges to the target distribution (the posterior distribution). If a chain becomes convergent means the produced sample from the target distribution has been obtained correctly. The Markov chain Monte Carlo (MCMC) method is a general simulation method for sampling from posterior distributions and computing posterior quantities of interest. MCMC methods sample successively from a target distribution. Several other aspects of the Markov chain method also contributed to its success. Most importantly, if the simulation algorithm is implemented correctly, the Markov chain is guaranteed to converge to the target distribution [24–27]. MCMC technique depends on the approximate distribution which is improved by a simulation of each step until a convergence of the posterior distribution is achieved.

Appropriate diagnostics such as; the Heidelberger Welch (stationarity test) and Heidelberger-Welch (halfwidth test).

Results

Information on the number of cigarettes smoked per day obtained from a total of 12,645 respondents in Ethiopia was studied. Figure 1 showed the percentage distribution of the number of cigarettes smoked in Ethiopia based on information from 12,645 respondents. In this study, 94.2% of them never were cigarettes smoked per day, whereas 0.6% of cigarettes smoking only once per day, 1.0% of cigarette smoking twice per day, 1.2% of cigarette smoking three times per day and 3.3% of cigarettes smoking at least four per day.

This indicates excess zero outcomes were large in numbers. However large numbers of cigarettes smoked per day were observed less frequently. This leads to a positively skewed distribution. It indicates that the data could be fitted better by a zero-inflated and hurdle model which takes into account excess zeroes. We visualized that an over-dispersion of the response variable. Since the histogram was highly peaked at zero, from this we can state that the overdispersion comes due to an excess zero. Due to a large number of excess zero outcomes, the histogram was highly picked at the very beginning (about the zero values).

Test of overdispersion

The Vuong test was used to test between the pairs of non-nested models. In this study, we compared ZIP versus PLH and ZINB versus NBLH to test if the over-dispersion in the smoke occurrence data was attributable to high frequencies of zero counts (excessive zeros). This to investigate if the excessive zeros were due to two sources (structure and sampling) or only one source (structure). Table 1 summary of the model comparisons

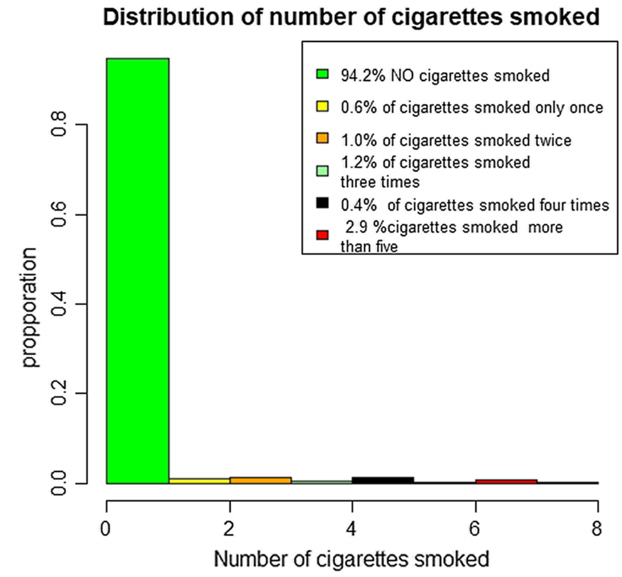

Fig. 1 Histogram of the number of cigarettes smoked per day in Ethiopia

Table 1 The likelihood ratio test (LRT) and Vuong test among the ZIP, ZINB, PLH, and NBLH models

Likelihood ratio Test for nested models				
Model	Hypothesis		p-value	Preferable model
ZIP = PLH vs ZINB = NBLH	H_0:∅ = 0 VS H_1:∅ > 0		< 2.2e−16	ZINB = NBLH
Vuong test for non-nested models				
Model	Hypothesis	Vuong statistics	p-value	Comment
ZIP vs PLH	Model1 = model2 vs Model1 > model2	1.216179	0.11196	ZIP = PLH
ZINB vs NBLH	Model1 = model2 vs Model1 > model2	1.234438	0.10852	ZINB = NBLH

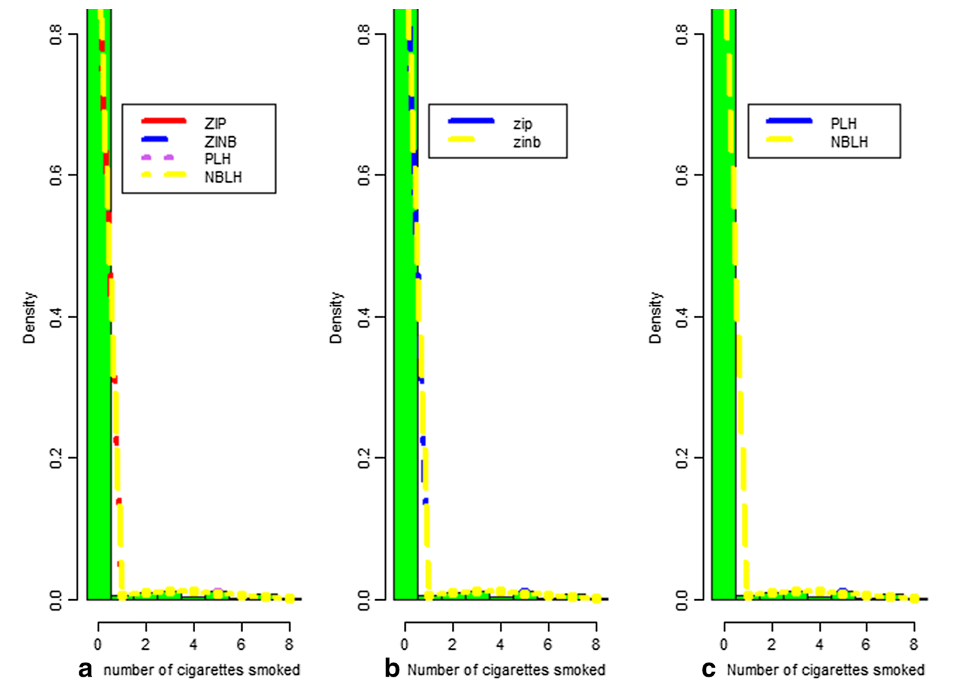

Fig. 2 Comparison of the densities of each model fits **a**–**c** number of cigarette smoking

based on Vuong's statistics for the four count regression models explored. States that if the corresponding p-value is bigger than a pre-specified critical value such as 0.05, then one can conclude that the four models fit the data equally well with no preference given to either model. But, if |V| yields a p-value smaller than the thresholds 0.05, then one of the models is better. Therefore, the ZINB, ZIP, NBLH, PLH was chosen as the best model. There was no difference between ZINB and NBLH models, indicating all models the excessive zeros equally well. The ZIP and ZINB models are nested so they can be compared by using the likelihood test for overdispersion to test H_0:∅ = 0, which provide evidence for preferring the ZINB over the ZIP (p-value < 2.2e−16).

In Fig. 2 zero-inflated and hurdle models captured almost all zero values. Based on predicted probabilities, the differences in model fit between the four models were remarkable. Still, the ZIP model and the PLH model did not fit the data reasonably well.

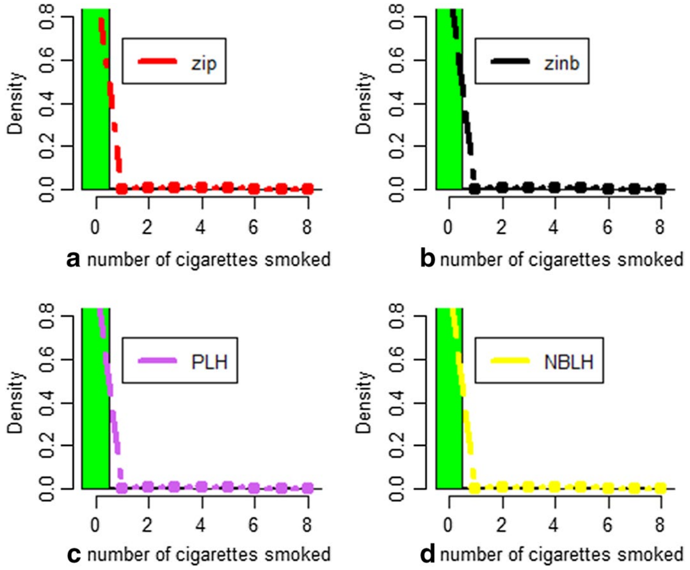

Fig. 3 Comparison of the densities of each model fits **a**–**d** observed and predicted frequencies of smoke occurrence for ZIP, ZINB, PLH, NBLH models

Table 2 Model comparison using AIC and log-likelihood

	ZIP	PLH	ZINB	NBLH
AIC	7591.194	7591.697	7590.496	7590.006
Log-Likelihood(df)	−3764.597 (df = 30)	−3764.849 (df = 30)	−3765.248 (df = 31)	−3765.003 (df = 31)

The NBLH and ZINB models gave a good prediction of zeros. Ninety four percent of the observed zeros were predicted by both models.

Figure 3 illustrates the observed frequency of smoking occurrence in the 12,645-validation data points (bar chart) and predicted frequencies of smoking occurrence from each of the four models. It was clear that the ZINB, PLH, and NBLH model yielded better predictions for both zero counts and positive counts than did the other models.

In Table 2 to compare the model improved convergence of model parameters, decreases the deviance and increases the predictive power. we can state that the zero-inflated Poisson model is better than the standard Poisson model, and also the zero-inflated negative binomial model is better then the negative binomial model. Since the **ZINB** and the **NBLH** model have the closest AIC values, we can conclude that these models perform best for our data set.

Results of Bayesian ZINB and NBLH Model

The main goal of this paper was to present a diagnosis of MCMC convergence and to investigate the significant predictors as well as characterizing smoking intensity in Ethiopia. Bayesian approach results, it is needed to check the convergence assessment, which involves checking that the sequence or chain has converged to and provides a

Table 3 Heidelberger and Welch Stationarity and half-width tests for the Bayesian chains used in the diagnosis of MCMC

Count model coefficients (NBLH)				Count model coefficients (ZINB)			
Parameters	Stationarity Test	P-Value	Half-width test	Parameters	Stationarity Test	P-Value	Half-width test
		C				C	
$\varnothing$	Passed	0.517	Passed	$\varnothing$	Passed	0.24	Passed
β_0	Passed	0.456	Passed	β_0	Passed	0.574	Passed
β_1	Passed	0.322	Passed	β_1	Passed	0.38	Passed
β_2	Passed	0.263	Passed	β_2	Passed	0.425	Passed
β_3	Passed	0.188	Passed	β_3	Passed	0.354	Passed
β_4	Passed	0.289	Passed	β_4	Passed	0.371	Passed
β_5	Passed	0.515	Passed	β_5	Passed	0.226	Passed
β_6	Passed	0.677	Passed	β_6	Passed	0.353	Passed
β_7	Passed	0.462	Passed	β_7	Passed	0.645	Passed
β_8	Passed	0.51	Passed	β_8	Passed	0.71	Passed
β_9	Passed	0.39	Passed	β_9	Passed	0.586	Passed
β_{10}	Passed	0.397	Passed	β_{10}	Passed	0.286	Passed
β_{11}	Passed	0.635	Passed	β_{11}	Passed	0.386	Passed
β_{12}	Passed	0.516	Passed	β_{12}	Passed	0.897	Passed
β_{13}	Passed	0.371	Passed	β_{13}	Passed	0.582	Passed
β_{14}	Passed	0.172	Passed	β_{14}	Passed	0.375	Passed
Zero hurdle model coefficients (binomial with logit link)					Zero-inflation model coefficients		
γ_0	Passed	0.398	Passed	γ_0	Passed	0.447	Passed
γ_1	Passed	0.272	Passed	γ_1	Passed	0.595	Passed
γ_2	Passed	0.281	Passed	γ_2	Passed	0.692	Passed
γ_3	Passed	0.094	Passed	γ_3	Passed	0.258	Passed
γ_4	Passed	0.151	Passed	γ_4	Passed	0.292	Passed
γ_5	Passed	0.294	Passed	γ_5	Passed	0.524	Passed
γ_6	Passed	0.3	Passed	γ_6	Passed	0.494	Passed
γ_7	Passed	0.489	Passed	γ_7	Passed	0.464	Passed
γ_8	Passed	0.365	Passed	γ_8	Passed	0.427	Passed
γ_9	Passed	0.218	Passed	γ_9	Passed	0.299	Passed
γ_{10}	Passed	0.377	Passed	γ_{10}	Passed	0.329	Passed
γ_{11}	Passed	0.378	Passed	γ_{11}	Passed	0.446	Passed
γ_{12}	Passed	0.546	Passed	γ_{12}	Passed	0.797	Passed
γ_{13}	Passed	0.658	Passed	γ_{13}	Passed	0.387	Passed
γ_{14}	Passed	0.659	Passed	γ_{14}	Passed	0.398	Passed

$C = \frac{C1+C2+C3}{3}$, C1= chain one, C2 = chain two, C3 = chain three

representative sample from the posterior distribution. The Heidelberg and Welch diagnostic calculates a test statistic (based on the Cramer-von Mises test statistic) to accept or reject the null hypothesis that the Markov chain Monte Carlo is from a stationary distribution. The diagnostic consists of two parts. In part one; if the null hypothesis is rejected, discard the first 10% of the chain. Repeat until the null hypothesis is accepted or 50% of the chain is discarded. If the test still rejects the null hypothesis, then the chain fails the test and needs to be run longer. If the chain passes the first part of the diagnostic, then it takes the part of the chain not discarded from the first part to test the second part. The halfwidth test calculates half the width of the 95% credible interval around the

Table 4 Summary statistic of the posterior distribution of the model parameters

Para	Bayesian negative binomial-logit hurdle (BNBLH)								
	Count model coefficients				Para	Zero hurdle model coefficients			
	Mean	SD	HPD of 95% CrI	OR		Mean	SD	HPD of 95% CrI	OR
∅	0.029	0.016	(0.003, 0.065)	1.029					
β_0	0.869	0.163	(0.558, 1.189)	2.385	γ_0	−3.299	0.298	(−3.902, −2.753)	0.037
β_1	0.026	0.085	(−0.139, 0.193)	1.026	γ_1	1.022	0.156	(0.721, 1.334)	2.779
β_2	0.090	0.087	(−0.078, 0.257)	1.094	γ_2	1.434	0.164	(1.123, 1.766)	4.195
β_3	−0.161	0.073	(−0.304, −0.021)	0.851	γ_3	−0.709	0.141	(−0.992, −0.433)	0.492
β_4	−0.035	0.050	(−0.133, 0.064)	0.966	γ_4	0.369	0.102	(0.174, 0.573)	1.446
β_5	0.133	0.089	(−0.041, 0.305)	1.142	γ_5	−0.518	0.149	(−0.812, −0.223)	0.596
β_6	0.223	0.078	(0.075, 0.376)	1.25	γ_6	−0.099	0.138	(−0.366, 0.179)	0.906
β_7	0.056	0.067	(−0.077, 0.188)	1.058	γ_7	−0.250	0.134	(−0.515, 0.007)	0.779
β_8	−0.068	0.070	(−0.206, 0.064)	0.934	γ_8	−0.313	0.131	(−0.569, −0.057)	0.731
β_9	−0.080	0.069	(−0.217, 0.054)	0.923	γ_9	−0.887	0.137	(−1.162, −0.624)	0.412
β_{10}	0.027	0.094	(−0.156, 0.209)	1.027	γ_{10}	−0.371	0.181	(−0.731, −0.018)	0.69
β_{11}	0.105	0.082	(−0.050, 0.271)	1.111	γ_{11}	−0.249	0.151	(−0.547,0.045)	0.780
β_{12}	0.221	0.065	(0.096, 0.349)	1.247	γ_{12}	2.121	0.112	(1.908,2.345)	8.339
β_{13}	0.149	0.065	(0.025, 0.279)	1.161	γ_{13}	−0.383	0.128	(−0.629, −0.120)	0.682
β_{14}	0.005	0.047	(−0.088, 0.098)	1.005	γ_{14}	−0.227	0.095	(−0.410, −0.041)	0.797
Para	Bayesian zero-inflated negative binomial (BZINB)								
	Posterior count model coefficients				Para	Posterior zero-inflation model coefficients			
	Mean	SD	HPD of 95% CrI	OR		Mean	SD	HPD of 95% CrI	OR
∅	0.029	0.016	(0.003, 0.065)	1.029	γ_0	3.210	0.302	(2.570, 3.799)	24.78
β_0	0.901	0.163	(0.588, 1.215)	2.462	γ_1	−1.021	0.156	(−1.339, −0.725)	0.360
β_1	0.019	0.087	(−0.154, 0.185)	1.019	γ_2	−1.426	0.165	(−1.767, −1.110)	0.240
β_2	0.083	0.088	(−0.092, 0.251)	1.087	γ_3	0.697	0.136	(0.431, 0.965)	2.008
β_3	−0.170	0.075	(−0.321, −0.029)	0.844	γ_4	−0.378	0.103	(−0.577, −0.173)	0.685
β_4	−0.036	0.051	(−0.135, 0.066)	0.965	γ_5	0.529	0.152	(0.238, 0.828)	1.697
β_5	0.122	0.092	(−0.053, 0.306)	1.13	γ_6	0.121	0.143	(−0.148, 0.408)	1.129
β_6	0.213	0.082	(0.059, 0.376)	1.237	γ_7	0.254	0.132	(−0.007, 0.514)	1.289
β_7	0.053	0.068	(−0.080, 0.186)	1.054	γ_8	0.306	0.130	(0.053, 0.563)	1.358
β_8	−0.075	0.069	(−0.212, 0.059)	0.928	γ_9	0.889	0.133	(0.625, 1.152)	2.433
β_9	−0.088	0.072	(−0.233, 0.052)	0.916	γ_{10}	0.360	0.191	(−0.003, 0.748)	1.433
β_{10}	0.011	0.096	(−0.175, 0.200)	1.011	γ_{11}	0.251	0.156	(−0.050, 0.560)	1.285
β_{11}	0.098	0.082	(−0.063, 0.261)	1.103	γ_{12}	−2.096	0.113	(−2.315, −1.875)	0.123
β_{12}	0.221	0.065	(0.098, 0.353)	1.247	γ_{13}	0.406	0.127	(0.151, 0.651)	1.501
β_{13}	0.152	0.071	(0.016, 0.293)	1.164	γ_{14}	0.231	0.094	(0.050, 0.413)	1.26
β_{14}	0.004	0.047	(-0.088, 0.098)	1.004					

Intercept(β_0); age: 25–34 (β_1&γ_1), = > *35* (β_2&γ_2); Rural (β_3&γ_3); education level(β_4&γ_4); Religion*: Orthodox*(β_5&γ_5), Muslim (β_6&γ_6); Internet (β_7&γ_7); wealth index: Middle(β_8&γ_8), Rich (β_9&γ_9); marital status: single(β_{10}&γ_{10}), married(β_{11}&γ_{11}); chewed chat(β_{12}&γ_{12});Occupation(β_{13}&γ_{13}) and media (β_{14}&γ_{14}),∅ is dispersion parameter

mean. If the ratio of the half-width and the mean is significant, the chain passes the test. Table 3 shows the Heidelberger and Welch stationarity tests for the Bayesian Markov chain Monte Carlo. It shows the stationarity and convergence during the burn-in period.

Application of Bayesian negative binomial-logit hurdle model

Bayesian count model coefficients

In Table 4 the estimated Bayesian count model coefficient odds ratio comparing the urban, considering other variables are held constant in the model. The expected number of the smoker who takes manufactured cigarettes was 0.851 times less than for rural to urban, remaining all other variables constant in the model. A smoker affiliated with the Muslim religion smoked on average was 1.237 smoke cigarettes per day times higher than other religions, remaining all other variables constant in the model. The result also revealed that the expected number of smokers who were used chewed chat was 1.247 times higher than people who have not used chewed chat. The expected counts of smokers would be expected to increase by 1.161 times for those who have an occupation to the group who have not occupation.

Bayesian-zero hurdle model coefficients

This finding is the estimated Bayesian-zero hurdle regression coefficient's odds ratio comparing the age group below 24, considering other variables are held constant in the model. Estimated the odds of the non-zero smokers for age 25–34 years and $\geq$ 35 years were 2.779 and 4.195 times higher than as compared to age $\leq$ 24 years respectively. The Zero hurdle model indicated that the estimated odds of the number of non-zero smokers of people who lived in rural was 0.492 times less than those who lived in urban. Additionally, the effects of education level on the smoke of people, we found that education level people were (OR = 1.446; HPD CrI 0.174, 0.573) times more likely to the number of non-zero smokers compare to no educated people. Furthermore, as the level of education level increases, the odds of the number of non-zero smoke also increased by 1.446.

Though Orthodox are found to be highly likely to non-zero smoke than their counterparts. Compared to men who are affiliated with other religious groups, those affiliated with Muslims and other religions are associated with having a lower non-zero smoking probability. The results indicate that people in lower socioeconomic have a higher likelihood of non-zero smoke. Additionally, the people in the higher and middle wealth category are reported to be less likely to non-zero smoke than their counterparts in the lower wealth category.

Regarding the determinants of cigarette smoking intensity, the study finds that single in marital status who non-zero smoke cigarette have a 0.69 times lower likelihood of non-zero smoke quantities of a cigarette than their counterparts. The estimated odds that the number of non-zero smoke with people chewed chat was 8.339 times more as compared people non chewed chat. Media use is found one of the important significant predictors of smoke cigarettes. The estimated number of non-zero smoke cigarettes for people who were used media is about 0.797 times lower than people who were not used. Similarly, the estimated odds of the number of non-zero smoke those have occupation people had decreased by 31.8% as compared to the on the occupation (see Table 4).

Application of Bayesian ZINB model

Bayesian count model coefficients

The finding of this study also revealed that residence had a significant factor in the number of cigarette smokers. The expected number of daily cigarette smokers with rural was 0.844 times less than that of the expected number of cigarette smokers daily for urban smokers while holding all other variables in the model. This means urban smokers and rural smokers are not certain zero. But urban smokers are higher than rural smokers. Besides, the expected number of daily cigarette smokers with have occupation was 1.164 times higher than that of the expected number of cigarette smokers daily for not occupation smokers. Thus, the occupation, the higher the predicted daily. Revealed that religion has a significant factor in the number of cigarette smokers. The expected number of cigarette smokers for Muslims was 1.237 times higher than the expected number of cigarette smokers for other religions controlling other variables in the model. And also, the expected number of cigarette smokers for chewed chat increased by 24.7% as compared to the expected number of cigarette smokers for not chewed chat controlling other variables in the model (see Table 4).

Bayesian zero-inflation model coefficient

The zero-inflated negative binomial part also indicated that the estimated odds that the number of cigarette smokers becomes zero with chewed chat decreased by 87.7% (OR = 0.123; HPD of 95% CrI -2.315, -1.875) as compared to not used chewed chat. In other words, the lower the used chewed chat, the less likely the cigarette smokers are a certain zero. As shown place of residence has a significant effect on the probability of being the always not cigarette smokers. The odds of being in the always not cigarette smokers for rural was 2.008 times higher than compared to those in urban centers controlling other variables in the model. The analysis further indicated that the estimated odds the number of cigarette smokers becomes zero with age 25–34 and age ≥ 35 were 0.36 and 0.24 times less than that of age 15–24 respectively. Revealed that education has a significant factor in the probability of being an excess zero cigarette smokers. The odds of being in the always zero cigarette smokers decreased by a factor of 0.685 for educated as compared to not educated holding all other variables constant. According to the findings of this study, the wealth index of the household has a significant influence on the probability of being an excess zero cigarette smokers. The odds of being in the always zero cigarette smokers for medium and rich households were increased by a factor of 1.358 and 2.433 times higher than that of poor households, respectively while holding all other variables in the model constant. As a show that currently working had statistically significant on the number of excess zero cigarette smokers. The estimated odds that the number of cigarette smokers becomes zero with currently working increased by factor 1.501 as compared to not currently working. Similarly, that the estimated odds that the number of cigarette smokers becomes zero in media was 1.26 times that of the not use media (see Table 4).

Discussion

Smoking intensity is defined usually as the number of cigarettes smoked per day. It can be considered as an important factor in establishing many serious smoking-related diseases. Count regression models are the first-line models that may be wanting to determine factors related to smoking intensity as a counter response. As a result, we want to construct a model that may handle the existence of excess zero counts and also the over-dispersed phenomenon. Motivated by these facts, with excess zeros and high variability of non-zero outcomes, the NBLH and ZINB model was found to be better fitted. Therefore, there is no or little difference in AIC between the ZINB model and the NBLH model [28, 29]. However, it should be noted that NBLH which allows for over-dispersion and also accommodates the presence of excess zeros, is more appropriate among all zero-adjusted models [30].

In this article considered several count data models to examine the factors associated with the number of cigarette smokers using a dataset with over-dispersion and inflated with zeros from 2016 EDHS. The NB model has more flexibility to capture additional variability and it fits the highly variable cigarette data better than the Poisson. However, the model underestimated zeros in the data, suggesting a model appropriate for zero-inflated data and hurdle model may be indicated. The dispersion parameters from the ZINB and NBLH were significant, suggesting the equidispersion assumption in the count portions of ZIP and PLH was violated. Due to this, after considering both the zero counts and the enduring overdispersion, the ZINB and NBLH fit the data best according to the AIC and Vuong tests [31].

We used two models to fit the data and computed Bayesian estimates for existing parameters with the non-informative prior. The competing models were the NBLH and ZINB models. We used the Gibbs sampler to obtain samples from the full conditional distributions. The algorithm was run for 10,000 iterations after the initial 5000 iterates were discarded as a burn-in. We computed the posterior means and 95% highest posterior density (HPD) intervals of the regression coefficients and other relevant parameters [8].

The ZINB and NBLH models deal with zero-inflation and over-dispersion at the same time. These models have become a bit popular lately and to analyze the number of cigarettes smoked per day. The best thing about using these models more specific when handling zero-inflation is that they are doing decrease biases that result from the acute non-normality [32]. Interpretation between the Bayesian ZINB and NBLH are similar but with an important distinction, particularly concerning the logit component. Estimates from the NBLH logit suggest increased age was significantly associated with non-zero cigarette smoking, whereas the ZINB logit model suggests age was positively associated with being a structural zero that is, belonging to the non-smokers latent class. Therefore, in this study older age daily smokers had a higher likelihood to smoke cigarettes intensively, with the number of cigarette smokers becomes zero with older age were less than that of age 15–24 respectively. The age increases with an increased number of cigarette smokers in Ethiopia [33]. So, policymakers and programmers need to select and implement interventions reaching elderly smokers.

According to the findings of this study, the wealth index of the household had a significant influence on the probability of being an excess zero cigarette smokers. Zero cigarette smokers for medium and rich households were higher than that of poor households. This finding is consistent with those reported in Ethiopia, Ghana, and Kenya [34–36]. Generally, improving the economic status of the community is one of a positive policy direction.

As shown place of residence has a significant effect on the probability of being the always not cigarette smokers. The odds of being always not cigarette smokers for rural was higher than compared to those in urban. This finding is in line with a study done in Ethiopia [34]. This could be explained by the fact the differences in the availability and accessibility of manufactured cigarettes between the urban and rural areas.

The zero-inflated negative binomial part also indicated that the number of cigarette smokers becomes zero with chewed chat was less likely than not used chewed chat. In other words, the lower the used chewed chat, the less likely the cigarette smokers are a certain zero. This finding is similar with the study done in Ethiopia [37]. This study result revealed that tobacco control interventions need to be tailored to address other substance use behaviors.

The association of religion with smoking. Patterns of smoking are known to vary significantly by religion but less is known about how this association is affected by other factors. The link between religion and smoking can vary significantly across different religious communities and must be deployed with careful attention to community norms if it is to be effective. Revealed that religion has a significant factor in the number of cigarette smokers. The number of cigarette smokers for Muslims was higher than the number of cigarette smokers for other [38].

On the other hand, Mass media exposure is found one of the important significant predictors of smoke cigarettes. The estimated number of non-zero smoke cigarettes for people who were used media is lower than people who were not used. The estimated number of non-zero smoke cigarettes for people who were used media is lower than people who were not used. This finding is similar with the study done Nigeran [40]. Entertainment media outlets, for example television, movies, and radio are influential in initiating or sustaining the smoking habits of consumers.

Besides, smokers who had education smoked a higher number of cigarettes than those who were no education level. This is similar with a study done in Ghana and different from studies done in Nepal [36, 41]. In the same way number of cigarette smokers becomes zero with currently working was higher than not currently working.

Strengths and limitations

The main strength of the Bayesian negative binomial logit Hurdle and zero-inflated model used in this case accounts for survey design features such as weighting, clustering and stratification since a failure to account for design features leads to invalid statistical inference such as standard errors and over or under estimation. In general, hurdle and zero-inflated models have a wide range of application, and a Bayesian approach should appeal to investigators seeking more flexible alternatives to classical model-fitting procedures. The data on the number of cigarettes smoked per day was self-reported which may have recall bias and social desirability bias, leading to an underreporting of the rates. Finally, the study used the number of manufactured cigarettes used per day to measure the smoking intensity.

Conclusion

In this article the count regression models are the first-line models that can be used to determine factors associated with the number of cigarette smoking using a set with overdispersion and inflated with zeros from 2016 EDHS. Four count regression models were compared in terms of AIC and voung test. The model comparison identified that ZINB and NBLH models were better fitted for modeling the observed data with excess zeros and overdispersion. In this study, we proved that that the Bayesian negative binomial logit hurdle (NBLH) and zero inflated negative binomial (ZINB) models are more appropriate methods for the analysis of data with an excess of zeros and overdispersion. In estimating the parameters, Bayesian methods were used. Complex calculations using the Bayesian method in the estimation parameters can be solved by the Markov Chain Monte Carlo simulation that can generate random values with the Gibbs-sampling algorithm.

Furthermore, using Bayesian ZINB and NBLH helps to select the most significant factor. The variable: age, residence, education level, internet use, wealth index, marital status, chewed chat, occupation, media were the most determinate factors on the smoking intensity in Ethiopia. It is suggested that future research considers other proportions of zeros and event stage distributions, under dispersion adjustments, different optimization procedures. This research should also be extended to an approach for direct Bayesian marginal inference. The aim is to develop a Bayesian marginalized model for zero-inflated univariate count outcome in the presence of overdispersion.

Abbreviations

CrI: Credible interval; CI: Confidence interval; EDHS: Ethiopia Demography and Health Survey; MLE: Maximum likelihood estimation; OR: Odds ratio; NBLH: Negative binomial-logit hurdle; ZIP: Zero inflated Poisson; ZINB: Zero-inflated negative binomial; PLH: Poisson logit hurdle; NB: Negative binomial; HPD: Highest posterior density; MCMC: Markov chain Monte Carlo.

Acknowledgements

The authors would like to acknowledge the Measure Demographic and Health Survey for providing online permission to use the Ethiopia Demographic and Health Survey 2016 data set.

Authors' contributions

MSW was responsible for the formulation of the methodology. MSW and AGA made substantial contributions to the conception and design of the study, analyzed and interpreted the data, and was a major contributor in writing the manuscript. MSW and AGA drafted the manuscript. Both authors read and approved the final manuscript.

Funding

No funding was received for this study.

Availability of data and materials

The data that support the findings of this study are available from the Measure DHS website (www.measuredhs.com but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of Measure DHS.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1] Department of Mathematical and Statistical Modeling (Statistics), Bahir Dar Institute of Technology-Bahir Dar University, Bahir Dar, Ethiopia. [2] Department of Epidemiology and Biostatistics, School of Public Health, College of Medicine and Health Science, Bahir Dar University, Bahir Dar, Ethiopia.

Received: 25 January 2021 Accepted: 13 April 2021
Published online: 21 April 2021

References

1. Reda AA, et al. Determinants of cigarette smoking among school adolescents in eastern Ethiopia: a cross-sectional study. Harm Reduct J. 2012;9(1):39.
2. Pešut D, Basara-Hadži Z. Cigarette smoking and lung cancer trends in Serbia-a ten-year analysis. Medicinski pregled. 2006;59(5–6):225–9.
3. Handebo S, et al. Smoking intensity and associated factors among male smokers in Ethiopia: further analysis of 2016 Ethiopian Demographic and Health Survey. BioMed Res Int. 2020;2020:4141370.
4. Bhore SJJC, Coletiva S. World no tobacco day: tobacco is a threat to the one health and sustainability. Ciência & Saúde Coletiva. 2020;25:4347–50.
5. EDHS, *ETHIOPIA Demographic and Health Survey 2016*. 2016.
6. Sharareh P, et al. Determining correlates of the average number of cigarette smoking among college students using count regression models. Sci Rep. 2020;10(1):1–10.
7. Karadoğan D, Önal Ö, Kanbay YJP. Prevalence and determinants of smoking status among university students: Artvin Çoruh University sample. Plos ONE. 2018;13(12):e0200671.
8. Jang H, et al. Bayesian analysis for zero-inflated regression models with the power prior: applications to road safety countermeasures. Accid Anal Prev. 2010;42(2):540–7.
9. Cheung YB. Zero-inflated models for regression analysis of count data: a study of growth and development. Stat Med. 2002;21(10):1461–9.
10. Prasetijo J, Musa WZ. Modeling Zero–Inflated Regression of Road Accidents at Johor Federal Road F001. in MATEC web of conferences. 2016. EDP Sciences.
11. Hall DBJB. Zero-inflated Poisson and binomial regression with random effects: a case study. Biometrics. 2000;56(4):1030–9.
12. Hilbe JM. Negative binomial regression. Cambridge: Cambridge University Press; 2011.
13. Hilbe JM. Modeling count data. Cambridge: Cambridge University Press; 2014.
14. Hofstetter H, et al. Modeling caries experience: advantages of the use of the hurdle model. Caries Res. 2016;50(6):517–26.
15. Sarul LS, Sahin S. An application of claim frequency data using zero inflated and hurdle models in general insurance. J Bus Econ Fin. 2015;4(4):732–43.
16. Bhaktha N. Properties of Hurdle Negative Binomial Models for Zero-Inflated and Overdispersed Count data. Ohio: The Ohio State University; 2018.
17. Shafira, S.A. and D. Lestari, *Bayesian Zero Inflated Negative Binomial Regression Model for The Parkinson Data*.
18. Hilbe JM, De Souza RS, Ishida EE. Bayesian models for astrophysical data: using R, JAGS, Python, and Stan. Cambridge: Cambridge University Press; 2017.
19. Hlongwane ZS. Zero-inflated regression models with application to water quality data from Umgeni Water. 2017.
20. Lim HK, et al. Score tests for zero-inflation and overdispersion in two-level count data. Comput Stat Data Anal. 2013;61:67–82.
21. Verma C, Stoffová V, Illés Z. Prediction of residence country of student towards information, communication and mobile technology for real-time: preliminary results. Proc Comput Sci. 2020;167:224–34.
22. Yuli Rusdiana R, Zain I, Wulan Purnami S. Censored Hurdle Negative Binomial Regression (Case Study: Neonatorum Tetanus Case in Indonesia). J Phys. 2017;855(1):012039.
23. Gelman A, et al. Bayesian data analysis. Boca Raton: CRC Press; 2013.
24. Chen M-H, Shao Q-M, Ibrahim JG. Monte Carlo methods in Bayesian computation. Berlin: Springer; 2012.
25. Congdon P. Bayesian Statistical Modelling. New York: John Wiley & Sons; 2001.
26. Congdon P. Bayesian statistical modelling, vol. 704. Hoboken: John Wiley & Sons; 2007.
27. Geman S, Geman DJ, Intelligence M. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans Pattern Anal Mach Intell. 1984;6:721–41.
28. Loeys T, et al. The analysis of zero-inflated count data: Beyond zero-inflated Poisson regression. Br J Math Stat Psychol. 2012;65(1):163–80.
29. Cameron AC, Trivedi PK. Regression analysis of count data, vol. 53. Cambridge: Cambridge University Press; 2013.
30. S Gurmu, PK Trivedi, E Statistics. Excess zeros in count models for recreational trips. J Bus Econ Stat. 1996;14(4):469–77.
31. Pittman B, et al. Models for analyzing zero-inflated and overdispersed count data: an application to cigarette and marijuana use. Nicotine Tob Res. 2020;22(8):1390–8.
32. Schunck R. and BG Rogge, No causal effect of unemployment on smoking? A German panel study. Int J Public Health. 2012. 57(6): p. 867–874.
33. Rachiotis G et al. Factors associated with adolescent cigarette smoking in Greece: results from a cross sectional study (GYTS Study). BMC Public Health. 2008. 8(1): 1–7.
34. Guliani H, Gamtessa S, Çule M. Factors affecting tobacco smoking in Ethiopia: evidence from the demographic and health surveys. BMC Public Health. 2019. 19(1): 1–17.
35. Tang S, et al. Prevalence of smoking among men in Ethiopia and Kenya: a cross-sectional study. Int J Environ Res Public Health. 2018;15(6):1232.
36. Nketiah-Amponsah E, Afful-Mensah G, Ampaw S. Determinants of cigarette smoking and smoking intensity among adult males in Ghana. BMC Public Health. 2018. 18(1): 1–10.
37. Kassa A, Deyno S. Prevalence and determinants of active and passive cigarette smoking among undergraduate students at Hawassa University, Hawassa, Ethiopia. J Trop Dis. 2014. 2(145): p. 2.
38. Hussain M et al. Smoking and religion: untangling associations using English survey data. J Relig Health. 2019;58(6):2263–76.
39. Khanal V, Adhikari M, Karki S. Social determinants of tobacco consumption among Nepalese men: findings from Nepal Demographic and Health Survey 2011. Harm Reduct J. 2013. 10(1): p. 1–10.
40. Tafawa AO et al., Mass media exposure, social stratification, and tobacco consumption among Nigerian adults. Cancer Causes Control. 2012. 23(1): p. 45–55.

41. Shrestha N et al. A nationally representative study on socio-demographic and geographic correlates, and trends in tobacco use in Nepal. Sci Rep. 2019. 9(1): p. 1–11.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.