Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v15n1p167

Application of Bayesian analysis on risk factors of coronary artery disease
By Ghosh, Samanta

Published: 20 May 2022

Electronic Journal of Applied Statistical Analysis
Vol. 15, Issue 01, May 2022, 167-186
DOI: 10.1285/i20705948v15n1p167

Application of Bayesian analysis on risk factors of coronary artery disease

Sarada Ghosh[a] and Guruprasad Samanta*[b]

[a]*Department of Statistics, Gurudas College, Phool Bagan, Kolkata-700054, India*
[b]*Department of Mathematics, Indian Institute of Engineering Science and Technology, Shibpur, Howrah-711103, India*

Published: 20 May 2022

Coronary Artery Disease is the leading cause of death globally. Coronary artery disease (or Ischemic heart disease) is caused by plaque buildup in the arteries that supply oxygen-rich blood to heart. Plaque causes a narrowing or blockage that could result in a heart attack. Some risk factors responsible for mortality and morbidity during IHD (Ischemic heart disease) are evaluated using suitable statistical models. In this work, due to count data, we propose Poisson, Negative Binomial and also utilize a flexible class of zero inflated models such as Zero-Inflated Poisson (ZIP) and Zero-Inflated Negative Binomial (ZINB) models estimated by the method of MLE and are compared to assess the most appropriate model for the underlying data. The forward and backward model selection procedures are also taken to permit the most significant factors associated with heart disease. The ZIP model is identified as the most appropriate one in this work. Moreover, a Bayesian estimation is chosen to account for prior on regression coefficients in a small sample size setting. This estimation also evolves as an alternative to traditionally used MLE based methods for such data. As per our simulation studies: the proposed method has better finite sample performance than the classical method with tighter interval estimates and better coverage probabilities. The simulation is based on *R*-software.

keywords: Zero inflated model; Bayesian inference; Gibbs sampling; Markov chain Monte Carlo; Log-likelihood.

*Corresponding author: gpsamanta@math.iiests.ac.in

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

1 Introduction

In the World, heart disease is the biggest killer of both men and women (WHO (2019)). One can die from heart disease about every minute in the United States alone and one in every four deaths is associated with heart disease. The risk of heart disease is increased by fat distribution across the body and other cardio metabolic conditions in women (presented at annual Radiological Society of North America meeting). It is presented in the 69^{th} Annual Conference of the Cardiological Society of India that the risk of heart disease increase by 500 percent with baldness and gray hair in men. In other news, a new study in the BMJ (British Medical Journal) is presented that smoking just one cigarette per day can increase the risk of heart disease. It is also suggested recently that restless leg syndrome (RLS) may also increase the risk of death from heart-related conditions, especially for older women. All of these mentioned problems gradually destroy the lining of the heart's blood vessels and make them become narrowed or blocked completely.

Heart disease is a general term that means that the heart is not working normally. A person can have heart disease but does not feel any illness. But some people with heart disease have symptoms (pain in chest, trouble for breathing, palpitations, swelling of feet or legs, cyanosis or feeling weak due to the body and brain are not getting enough blood for supplying them with oxygen). Nowadays, Coronary Artery Disease (CAD) is most common heart disease in the World. CAD is also known as ischemic heart disease (IHD). It normally happens for accumulating cholesterol on the artery walls and creating plaques. It is known as atherosclerosis. The arteries become narrow and reducing blood flow to the heart. Sometimes, a clot can obstruct for delivering blood to the heart muscle. If blood vessels connecting to the heart become very narrow, or somehow blood vessels are blocked partially or completely, then blood cannot flow through them normally. As a result, muscle cannot be able to work in normal capacity for supplying requisite amount of blood to the heart muscle. Heart muscle become sick and weak, in fact it can even die if blood flow stops.

There are four primary coronary arteries are detected on the heart's surface: (i) right main coronary artery, (ii) left main coronary artery, (iii) left circumflex artery, (iv) left anterior descending artery. These arteries are responsible for bringing oxygen and nutrient-rich blood to heart. A healthy heart can move daily approximately 3,000 gallons of blood throughout our body according to the Cleveland Clinic.

Several symptoms, in case of CAD, occur when heart does not get sufficient arterial blood. Angina (a type of chest pain due to insufficient blood flow to the heart) is the most common symptom of CAD. Some people describe such a discomfort in another way just like chest pain, tightness, burning, heaviness, squeezing etc. Apart from this, minor problems may also occur due to CAD: breathing problem, pain in shoulders or arm, dizziness etc. In spite of facing these types of problems, women also be disturbed with many symptoms (such as nausea, vomiting, back and jaw pain, shortness of breath without feeling chest pain) of CAD. If blood flow is decreasing than normal level of human body, heart may also become very weak and then abnormal heart rhythms (arrhythmia) or rates occur due to insufficiency of blood. Regional wall motion abnormality (RWMA) is a terminology used in echocardiography. This is commonly applicable for abnormalities

of motion of the left ventricular (lower muscular chamber of the heart) walls. If all segments of the left ventricle are contracting normally, then RWMA is absent in this case. It is noted that 30 patients out of 32 having RWMA can assess coronary artery disease. So, left ventricular regional wall motion abnormality (RWMA) predict the existence of significant coronary artery disease with 94% accuracy (Safford and Bove (1987)). Some risk factors are considered in this work which are most important predictors for CAD such as high blood pressure (HBP), pulse rate (PR), tobacco smoking (present or ex denoted by PTS and ETS respectively), diabetes mellitus (DM), obesity, body mass index (BMI), hyper tension (HT) etc. (Alizadehsani et al. (2013); Karaolis, Moutiris and Hadjipanayi (2010)). For diagnosing of CAD, a review of medical history, a physical examination, and other medical testing are required. So, some results during medical test are also included like chronic renal failure (CRF), dyslipidemia, weak peripheral pulse (WPP), lung rales (LR), typical chest pain (TCP), Dyspnea, Q-wave, left ventricular hypertrophy (LVH), fasting blood sugar (FBS), creatine, triglyceride, lipo-protein density (low and high denoted by LLPD and HLPD respectively). Generally, the risk for CAD also increases with age of people: men have a greater risk for the disease beginning at age 45 and women have a greater risk beginning at age 55 (Alizadehsani et al. (2013)). Among many tests, electrocardiogram is most important which can helps to determine whether human had a heart attack or not. So, it is necessary to reduce or control the risk factors and seek treatment to lower the chance of heart attack or vulnerable stroke, if diagnosed with CAD. Treatment also depends on patients current health condition, risk factors and overall well being. Our lifestyle should be changed in such a manner that decreases the risk of heart disease and stroke. As for examples, quit smoking tobacco, reduce or stop consumption of alcohol, exercise regularly, lose weight to a healthy level, eat a healthy diet (describe by doctor). Beside these, doctor may prescribe to the affected patients a suitable procedure for increasing blood flow to the heart.

The Negative Binomial (NB) and Poisson models are two basic generalized linear model (GLM) which are widely applied to analyze count data (Agresti (2002); Ghosh and Samanta (2019)). The Poisson regression is identified by equal mean and variance and fits admirably for equidispersed data whereas the NB is utilized in the cases if over-dispersion is present in the response. Poisson regression model is compared with the NB model for identifying the best fitting model in this paper. However, these standard models fail when most of the observed counts are zeros. Then zero-inflated models have been utilized to address in such cases by modelling zero counts separately (Shankar, Milton and Mannering (1997)). This study seeks to investigate the most appropriate model to examine the uttermost important factors that significantly influence the region of regional wall motion abnormalities (RWMA) which is highly related to coronary artery disease. In section 2 we discuss about model derivation and preliminaries. Then the data description and simulation are discussed in the section 3. The concluding remarks and general discussion is held in the last section i.e., at section 4. This work also discusses the significance of modelling for excess-zero in count data structure in the context of Bayesian modelling. The simulation of this work is based on *R*-software (Zeileis, Kleiber and Jackman (2008)). Finally, the last section consists of the general discussions and conclusions of the paper.

2 Model derivation and preliminaries

2.1 Methodology for estimating model parameters

The conventional univariate Poisson model typically comes in mind as the outcomes are counts. The Negative Binomial and Zero-Inflated models are also considered because the data has an excess of zero counts and presence of over-dispersion. Since both can arise simultaneously, extensions such as Zero-inflated Poisson (ZIP) and Zero-inflated Negative binomial (ZINB) are obviously considerable models as well. Due to over-dispersion, we can proceed with the quasi-likelihood Poisson GLM and Negative Binomial (NB) models. The NB model is an extension of the Poisson regression and is generally used for addressing over-dispersion and is reviewed in this work.
Let $X = (X_1, X_2, ..., X_p)$ be the vector of p regressors and let $Y = (Y_1, Y_2, ..., Y_n)$ be the response vector with n as the number of observations. Let $Y_i \sim \text{Pois}(\theta_i)$, with random mean $\theta_i \sim \Gamma(\frac{\mu_i}{\lambda}, \lambda)$. Then the marginal distribution of $Y_i (i = 1, 2, ..., n)$ is the NB with probability function given by,

$$p(y, k, \mu) = \frac{\Gamma(y_i + k)}{\Gamma(k)\Gamma(y_i + 1)} \left(\frac{1}{\mu_i k^{-1} + 1}\right)^k \left(\frac{\mu_i k^{-1}}{\mu_i k^{-1} + 1}\right)^{y_i}, \text{ where } y_i = 0, 1, 2, ... \quad (1)$$

The simplest distribution for count data (i.e., data that take only non-negative integer value) is the Poisson distribution. Let Y denotes a count and let $\psi = E(Y)$. The Poisson probability mass function (pmf) for Y is

$$f(y; \theta) = \frac{e^{-\theta}\theta^y}{y!}, \text{ where } y = 0, 1, 2, ... \quad (2)$$

where $\theta(> 0)$ is both the mean and variance of the distribution, so it is described as equi-dispersed. The Poisson model is not sufficient in the case of excess of zeroes in the sample due to the violation of the equi-dispersion assumption. In contrast, sometimes many data are over-dispersed, whenever the variance exceeds their mean, so this reduces the usefulness of the Poisson distribution. In terms of the ensuing discussion, it is essential to recognize that the Poisson model and standard variants that permit for over-dispersion, cannot able for describing multi-modal data. The zero-inflated Poisson (ZIP) regression model is a modification of the familiar Poisson regression model which allows for an over-abundance of zero counts in the data (Mullahy (1986); Lambert (1992); Liu and Powers (2012); Lu et al. (2014)).

Firstly, we have defined the ZIGP (Zero inflated generalised poisson regression) such as follows (Czado et al. (2007); Wang, Shuangge and Wang (2015)):

$$P(Y_i = y_i | x_i, z_i) = \begin{cases} p_i(z_i) + (1 - p_i(z_i))\text{f}(\theta_i, x_i; 0), \text{ if } y_i = 0 \\ \\ (1 - p_i(z_i))\text{f}(\theta_i; y_i, x_i), \text{ if } y_i > 0 \end{cases} \quad (3)$$

where $f(\theta_i, x_i; y_i)$, y_i=0, 1, 2,... is GPR (generalised poisson regression) model and $0 < p_i < 1$.

In this work, we have considered ZIP model (Czado et al. (2007); Wang, Shuangge and Wang (2015)) whose objective is very straightforward, i.e., it assumes that outcomes emanate from two following processes (Lambert (1992)). It is mentioned that ZIP (i) models zero inflated by including a proportion $(1 - p_i)$ of extra zeroes and a proportion $p_i \exp(-\theta_i)$ of zeroes coming from the Poisson distribution and (ii) models the non-zero counts by using zero-truncated Poisson model. The ZIP model is as follows:

$$P(Y_i = y_i|x_i, z_i) = \begin{cases} p_i(z_i) + (1 - p_i(z_i))\text{Pois}(\theta_i; 0|x_i), \text{ if } y_i = 0 \\ \\ (1 - p_i(z_i))\text{Pois}(\theta_i; y_i|x_i), \text{ if } y_i > 0 \end{cases} \tag{4}$$

with z_i as a vector of covariates defining the probability θ_i, $\text{Pois}(\theta_i; 0|x_i) = \exp(-\theta_i)$, and $\text{Pois}(\theta_i; y_i|x_i) = \frac{e^{-\theta}\theta^{y_i}}{y_i!}$. The mean and the variance of ZIP are:

$$E(y_i|x_i, z_i) = (1 - p_i)\theta_i \tag{5}$$

and

$$\text{Var}(y_i|x_i, z_i) = (1 - p_i)(\theta_i + p_i\theta_i^2) \tag{6}$$

It is very clear that ZIP model changes into the classical Poisson model when $p_i = 0$. Otherwise, ZIP is over-dispersed since the variance exceeds the mean. This over-dispersion is not due to the heterogeneity of the data which can be handled using negative binomial model. Instead, it appears from the splitting of the dataset into the two statistical processes because of the excess of zeroes. For the independently distributed responses sampled from ZIP(p_i, θ_i), the commonly used link functions are given by,

$$\log(\theta) = z\beta \tag{7}$$

We can model $p_i(z_i)$ using a Logit model (Lambert (1992)) given by:

$$p_i(z_i) = \frac{\exp(z_i'\alpha)}{1 + \exp(z_i'\alpha)} \tag{8}$$

where z_i is a vector of covariates defining the probability p_i and α be a vector of its corresponding parameters. In this work the likelihood function of Y_i can be defined as:

$$L = \prod_{y_i=0} \left[p_i(z_i) + (1 - p_i(z_i)) \exp(-\theta_i)\right] \prod_{y_i \neq 0} \left[(1 - p_i(z_i))\frac{e^{-\theta}\theta^y}{y!}\right] \tag{9}$$

For statistically well-posed zero-inflated probability model, (Vuong (1989)) proposed a test which is a well suited method to compare ZIP regression to other non nested models for counts data. Suppose $P_N(y_i|x_i)$ be the predicted probability (from model N) of ith observed count and define m_i as:

$$m_i = \ln\left(\frac{P_1(Y_i|x_i)}{P_2(Y_i|x_i)}\right) \tag{10}$$

The test statistic for the Vuong's test (for the hypothesis $E(m_i) = 0$) is defined as follows:

$$V^* = \frac{\overline{m}\sqrt{n}}{S_m} \tag{11}$$

where $\overline{m}$ = mean of $m_i = \frac{1}{n}\sum_{i=1}^{n} m_i$ and S_m = standard deviation = $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(m_i - \overline{m})^2}$ and n is the sample size. It is mentioned that the Vuong's statistic is asymptotically normally distributed.

At 5% level of significance:

(i) The first model is accepted if $V^* > 1.96$.

(ii) If $V^* < -1.96$, then the second one is preferred.

(iii) The two models are equivalent when $-1.96 < V^* < 1.96$.

The Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are used for choosing the best model from several fitted models. When a model is fitted to a set of observed data generated from the unknown true model, the AIC provides a clear idea of the information lost by considering the complexity of the model and its goodness of fit test. The AIC for a model with k parameters and likelihood L is defined as (at convergence):

$$\text{AIC} = 2\ln(L) + 2k \tag{12}$$

As with AIC, the BIC model choice criteria penalizes model with additional parameters:

$$\text{BIC} = 2\ln(L) + k\ln(n) \tag{13}$$

where n represents the sample size and k is the number of parameters. Out of various models fitted for a given dataset, the preferred model has smallest AIC and BIC values.

2.2 The Bayesian Inference

In Bayesian approach, prior information about the distribution of parameters is considered along with the likelihood of the observed data to establish a posterior distribution of relevant quantities for inference about unknown parameters as well as other predictors of interest including data with combinations of parameters. Bayesian analysis involves the specification of prior distribution for the parameters of interest (Ghosh et al. (2006); Zeng et al. (2014)). In Bayesian analysis, conjugate priors are often chosen for convenience to obtain the resulting posterior distribution in a closed form belonging to the same distributional family (Neelon (2018)). However Bayesian analysis is performed basically in the situations where multiple parameters are included (such as the ZIP model). In this context it is mentioned that ZIP (p, θ) consists of two stages: (i) a Bernoulli zero-inflated stage with parameter p and (ii) a Poisson count stage with parameter θ. The

pmf (probability mass function) of ZIP can also be written in the following form:

$$P(Y=0) = p + (1-p)\frac{b(0)}{c(\theta)} \tag{14}$$

If $Y = \kappa$, where $\kappa = 1, 2, ...$, then pmf is as follows:

$$P(Y=\kappa) = (1-p)\frac{b(\kappa)\theta^{\kappa}}{c(\theta)} \tag{15}$$

where $c(\theta) = \sum_{\kappa=0}^{\infty} b(\kappa)\theta^{\kappa}$, $0 \leq p < 1$ and $\theta > 0$. For a random sample $Y = (Y_1, Y_2, ..., Y_n)$ from the ZIP model, the form of likelihood function is as follows:

$$L(p, \theta|Y) \propto [pc(\theta) + (1-p)b(0)]^{G_0}(1-p_0)^{n-G_0}\frac{\theta^G}{c^n(\theta)} \tag{16}$$

where $G_0 = G_0(Y)$ is the number of $\{i : Y_i = 0\}$ & $G = G(Y) = \sum_{i=1}^{n} Y_i$. The assumption is that the parameters θ and p in the case of prior distributions are independent. We have to use the conjugate priors such as $p \sim \text{Beta}(b_1, b_2)$ & $\theta \sim \pi(\theta)$, where $\pi(\theta) \sim \theta^{a_1}/c[(\theta)]^{a_2}$. It is not necessary to imply the prior and posterior independence. It is also assume that the hyper parameters a_1, a_2, b_1, and b_2 are known. If both of the value of hyper parameters b_1 and b_2 equal 1, then it provides a uniform prior over (0,1) for the p parameter. Sometimes, the computation of a joint posterior is very difficult using a standard density. To overcome such a difficulty, simulations help create a skillful strategy. The Monte Carlo method is used to sample from the posterior distribution to get rid of such limitations. Generally, it is assumed that the prior distributions for p and θ are independent and we have to use the following conditional conjugate priors as $p \sim \text{Beta}(b_1, b_2)$ and $\theta \sim \pi(\theta)$, where $\pi(\theta) \sim \theta^{a_1}/c[(\theta)]^{a_2}$. It is noted that prior independence does not necessarily imply posterior independence. The hyper parameters a_1,a_2,b_1,b_2 are assumed to be known. In particular, if $b_1 = b_2 = 1$ then it gives the uniform prior over (0,1) for the parameter p. Small values of a_2 result in a non-informative (high variance) prior for θ. Sometimes, it is very tough to compute the joint posterior by using a standard density. Simulation methods offer a skilful strategy in this case. Monte Carlo simulation based techniques are used to sample from the posterior distribution to overcome such analytical limitations. Apart from this, in particular case, the Gibbs sampling method has been utilized to prevail a large number of random variates from the posterior distribution. Any distributional summary such as mean, median or quantiles etc of the posterior distribution can be approximated using their corresponding sample analogue.

3 Data description and simulation

The data of heart disease is provided by UCI Machine Learning Repository. A sample of 303 patients are considered in this work. Summary statistics on the number of people

affected RWMA mean (0.62) and variance (1.28) are different and degree of abnormalities of regional wall motion (i.e., normal, mild, moderate, severe and extreme) are also considered in this work. This gives an indication of possible over-dispersion in the counts.

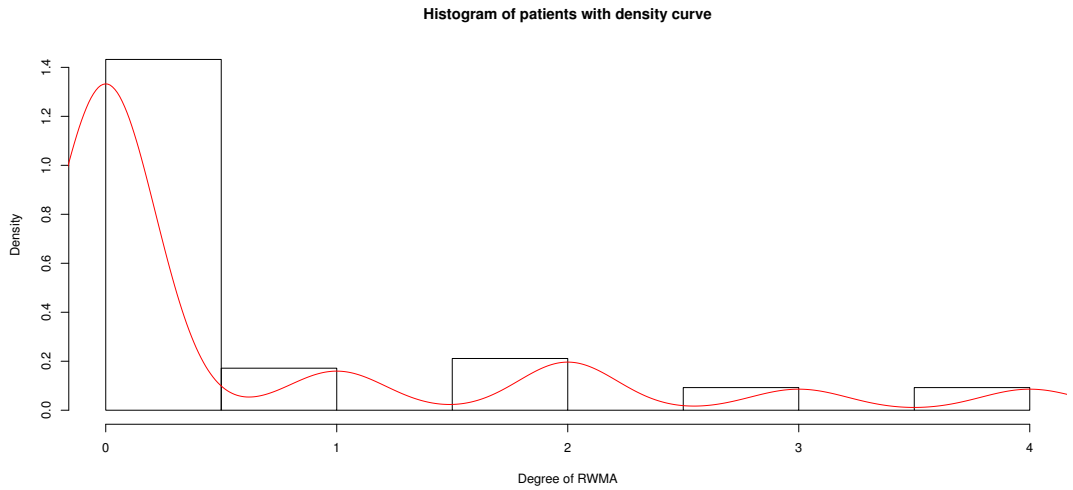

Figure 1: The frequencies of the degree of RWMA (from none to extreme)

Also, the histogram in Figure.1 shows a skew distribution with over 72% of zero observation. The two phenomena, over-dispersion and zero-inflated, exhibited by the data need to be accounted for by our model. Furthermore, a suitable model building process is required for selecting the most significant predictors of the 24 observations. Since the outcomes are counts, the conventional univariate Poisson model may be suitable. However, due to the presence of possible over-dispersion and has an excess of zero counts, the Negative Binomial and Zero-inflated models are also considered. Since both can occur simultaneously, extensions such as (i) Zero-inflated Poisson (ZIP) and (ii) Zero-inflated Negative binomial (ZINB) are obvious candidate statistical models as well.

Cardiovascular disease (CVD) risk factor control is difficult to reduce the CVD risk of those who have already diabetes. Recently some clinical trials have demonstrated that the use of more aggressive targets for blood pressure and cholesterol control among individuals with diabetes results in reduced incidence of CVD events (Shepherd et al. (2006); Hansson et al. (1998); Collins (2003)). The individuals with diabetes have a larger decline in both (total and LDL cholesterol) than those without diabetes and similar declines in blood pressure levels. The improvement in risk factor levels is paralleled by an increase in different drug treatments, which likely contributes to the observed trends. Diabetes mellitus is still associated with approximately an overall two to three-fold increased risk of CVD mortality (Fox et al. (2004); Gregg et al (2007); Gu et al. (1999)). The aim of this work is to model the RWMA in presence of some covariates using the zero-inflated model specifications. Generally, maximum likelihood estimation (MLE) technique is used to obtain the estimates of parameters in these models. The basic

principle behind this method is to estimate parameters that maximize the likelihood of the observed data.

3.1 Estimation of model parameters

The Poisson, negative binomial, zero-inflated Poisson and zero-inflated negative binomial models are fitted to the data set under the present study. First we have to select which of these regressors are significant in explaining the number of people having RWMA and then apply both backward and forward model selection. In forward selection, the model-fitting process begins with only the intercept and then sequentially adds the effect for improving the fit. This process is terminated when adding an effect produces no significant improvement. At each step, the effect that is most significant is added. The process is terminated when the significance level for adding any effect is greater than some specified entry significance level whereas the backward elimination starts with the full model where all independent effects are included. Then effects are deleted one by one until a stopping condition is satisfied. At each step, the effect that makes 'the smallest contribution to the model' is deleted. The significance level of an effect determines whether to drop that effect. At any step, the least significant predictor is dropped, and the process continues until all effects that remain in the model are significant at a specified stay significance level. The forward and the backward model selection criteria are used to find the best subset of effects for the conditional mean of the underlying models. Here the selection criteria is chosen based on AIC, BIC for selecting the final model. There are more than 50 variables in the data set. In the case of backward elimination, first consider the full model and all the AIC which are more than 700.19 for each model and similarly in case of BIC we get larger BIC for each model than mentioned in Table 1. Therefore, we eliminate the covariates (variables which are least significant, i.e., one with large p-value) one by one and terminate the process until a stopping criterion is reached (Zhang (2016)). Apart from this, the log-likelihood, AIC and BIC of the final models are displayed in Table 1. Finally, the models are considered with 24 important predictors, i.e., high blood pressure, smoking, diabetes, obesity, hyper tension, chest pain, left ventricular hypertrophy, creatine etc. are chosen as covariates emerge as the best after forward and backward model selection and then compare among them as follows: The Poisson regression model has performed badly

Table 1: A comparison of log-likelihood, AIC and BIC values

Criterion	Poisson	NB	ZIP	ZINB
Log-likelihood	-315.41	-291.19	-279.6	-280.3
AIC	686.82	640.38	617.24	619.23
BIC	790.81	748.07	743.22	746.57

among all the models. The Pearson statistic suggests how well the model performs in predicting the observed count response when covariates are considered. It also helps to evaluate whether there is evidence of over-dispersion or not. Indeed, the large ratio (Pearson/DF) gives an indication of the presence of over-dispersion in the response that needs to be accounted for. Next, a NB model is fitted to the data. Some consistency is observed in the variable selection because the variables are significant for NB model. It is shown in our work, NB is preferable than Poisson model. But we are still facing with the problem of excess zeros in our data. So, ZIP model is fitted to handle the over-dispersion. The AIC for the ZIP is smaller than the case of the conventional Poisson. In addition, we performed the Vuong test($V^* = 2.223$),which is significant and emphasized the superiority of the ZINB model over the conventional NB model. Next, compared with the Poisson model, we observed that the Poisson had a larger AIC and BIC value with the ZINB performing better. Thus, it appears the ZINB model corrected for the excess zeroes and did adequately capture the dispersion in the data as compared to the Poisson and NB case. Finally, the ZIP model was considered. The AIC for the ZIP and that of the conventional Poisson model were significantly different. The Vuong test ($V^* = 3.454$) was significant at 0.05 significance level. This is an indication that the ZIP model performs quite better than the Poisson. Therefore, amongst all the models considered, we selected the ZIP model as the best fitting model for the given data set.

3.2 Model fitting with Bayesian approach (for no covariates)

First, we utilize the data set in absence of explanatory variables. Regular Negative Binomial and Poisson distributions are also fitted to the data set for comparing its performance with the zero-inflated models. Then generate samples from the posterior distribution of the parameters. The parameters p, θ, zero defect probability $\Leftrightarrow$ P(Y =0) and lastly the deviance are observed for evaluating the convergence of Markov Chain Monte Carlo (MCMC) method.

Deviance is a measurement of goodness-of-fit for a statistical model, basically used in hypothesis testing. e.g., for the no covariate case, deviance $= -2\log[L\{(p,\theta)|Y\}]$. Since in this situation deviance is only function of the parameters, its posterior distribution is simply derived from the MCMC iterations using WinBUGS software. In this work, Bayesian zero inflated regression models without and with covariates are also been involved. For each Bayesian estimation run, we have applied Gibbs sampling with 10,000 MCMC iterations and 3 chains to the zero inflated models by using the WinBUGS. For no covariate case, we first initialise with the assumption that a uniform $(0, 1)$ prior for parameter p and a gamma $(0.5, 0.5)$ prior for θ.

For negative binomial models, the posterior means of the deviance are about 241.87 and for Poisson models it is near about 244.92. Therefore, it indicates that negative binomial models have roughly better performance than Poisson models. But the posterior mean and median estimates of P(Y =0) from the zero inflated Poisson regression model are very close to empirical estimates of zero counts (0.72 in our data) which suggests better fitting of ZIP model compared to the regular Poisson or NB and Bayesian

Table 2: Posterior summary (without covariates): zero-inflated models

Model	**Parameter /quantity**	**Mean**	**Sd**	**Percentile**			**Average length of CI**
				25%	**50%**	**97.5%**	
ZIP(p,θ)	θ	**0.432**	**0.064**	**0.231**	**0.433**	**0.675**	**0.44**
	p	**0.786**	**0.354**	**0.322**	**0.792**	**0.893**	**0.57**
	μ = **(1-p)**θ	**0.558**	**0.267**	**0.145**	**0.556**	**0.764**	**0.62**
	P(Y=0)	**0.722**	**0.087**	**0.083**	**0.721**	**0.873**	**0.79**
	Deviance	**244.97**	**15.7**	**216.38**	**216.87**	**220.91**	**4.51**
ZINB(p,θ,r)	**1-**θ	**0.862**	**0.08**	**0.674**	**0.884**	**0.937**	**0.26**
	r	**4.12**	**2.43**	**3.389**	**3.87**	**8.989**	**5.6**
	p	**0.782**	**0.42**	**0.252**	**0.84**	**0.764**	**0.51**
	μ = **r(1-p)**θ**/ (1-**θ**)**	**0.562**	**0.344**	**0.342**	**0.561**	**0.734**	**0.39**
	P(Y=0)	**0.739**	**0.77**	**0.643**	**0.731**	**0.872**	**0.23**
	Deviance	**241.87**	**19.7**	**213.33**	**215.78**	**217.7**	**4.40**

Table 3: Simulation for c.p. based on 95% C.I. and RMSE

Study	Parameter/quantity	Classical method					Bayesian method					%avarage reduction
		Std.dev	CI	CP	ALCI	RMSE	Std.dev	CI	CP	ALCI	RMSE	on ALCI
Study I	p = 0.1	0.023	(0.233,0.789)	0.928	0.56	0.803	0.0733	(0.302,0.874)	0.944	0.57	0.801	86%
	θ = 1	0.0302	(0.010,0.887)	0.942	0.88	0.052	0.201	(0.019,0.879)	0.958	0.86	0.051	
	P(Y=0)=0.43	0.0643	(0.019,0.984)	0.902	0.98	0.006	0.001	(0.201,0.741)	0.52	0.55	0.005	
Study II	p = 0.5	0.2827	(0.301,0.792)	0.962	0.49	0.403	0.0535	(0.286,0.839)	0.952	0.55	0.406	24%
	θ = 1	0.107	(0.112,0.805)	0.931	0.69	0.071	0.023	(0.080,0.901)	0.943	0.82	0.072	
	P(Y=0)=0.68	0.033	(0.011,0.787)	0.97	0.78	0.019	0.0042	(0.049,0.682)	0.91	0.63	0.016	
Study III	p = 0.9	0.2062	(0.333,0.822)	0.954	0.49	0.015	0.0553	(0.341,0.789)	0.938	0.45	0.016	35%
	θ = 1	0.643	(0.182,0.769)	0.924	0.59	0.154	0.204	(0.201,0.755)	0.946	0.55	0.155	
	P(Y=0)=0.94	0.032	(0.008,0.984)	0.982	0.98	0.043	0.035	(0.101,0.837)	0.986	0.74	0.031	

estimation helps for getting such information.

3.3 Simulation studies for comparison

Now, for showing validation of Bayesian method, we have to evaluate ZIP model (with appropriate statistics) without covariates for better understanding of this model. Here three simulation studies based on ZIP model are presented.

In simulation for study-1, $p = 0.1$, $\theta = 1$ and P(Y=0)=0.43, in simulation for Study-2, $p = 0.5$, $\theta = 1$ and $P(Y = 0) = 0.68$ and similarly in simulation for Study-3, $p = 0.9$, $\theta = 1$ and $P(Y = 0) = 0.94$ are used. For each case sample size is considered as 100 and is repeated 10,000 times. The results are based on c.p.(coverage probabilities with 95% interval estimates) in Table 3. The classical 95% confidence intervals are derived by inverting the LRT (likelihood ratio tests) based on the large sample chi-square distribution. The frequency and Bayesian estimates of each of the parameters

θ and p are very close (but in Study-3 the estimation of θ are different) and these suggests the Bayesian method performed better for estimating $P(Y = 0)$ (see Table 3). As for example, the reductions of the average length of the intervals are 86% and 35% for Study-1 and Study-3 respectively. It is concluded (from Table 3) that the Bayesian intervals are very competitive for coverage probabilities and that the average lengths of the Bayesian intervals can be significantly shorter than that obtained from the classical methods, when $P(Y = 0)$ is very near to one. For each case (Study-1, 2 and 3) sample size is considered as 50 and proceeding in the similar manner, we get the results (based on c.p.) for Study-1: 0.724, 0.853 and 0.822 respectively (for classical) and 0.969, 0.943 and 0.901 respectively (for Bayesian). For Study-2: 0.892, 0,919, 0.890 respectively (for classical) and 0.923, 0.911, 0.853 respectively (for Bayesian). Similarly, for Study-3: 0.874, 0.853, 0.822 respectively (for classical) and also get 0.917, 0.889, 0.867 (for Bayesian) respectively. So, as per the simulations: Bayesian method (specially for Study-1 and Study-3) gets high probability of convergence than classical method. Next, for fulfilling of comparison: root mean square error (RMSE) have also been evaluated in terms of classical and Bayesian estimations.

In Study-1, the RMSE and standard deviation (std.dev.) of the estimates based on ML are 0.803 (std.dev.= 0.023), 0.052 (std.dev.= 0.0302) and 0.006 (std.dev.= 0.0643) for the parameters p, θ and P (Y = 0) respectively whereas the corresponding RMSE and standard deviation of the Bayesian estimates are 0.801 (std.dev.= 0.0733), 0.051 (std.dev.= 0.201) and 0.005 (std.dev.= 0.001) for the parameters p, θ and P(Y =0) respectively. In Study-2, the RMSE and standard deviation of the MLE estimates are 0.403 (std.dev.= 0.2827), 0.071 (std.dev.= 0.107) and 0.019 (std.dev.= 0.033) and the RMSE and standard deviation of the Bayesian estimates were 0.406 (std.dev.= 0.0535), 0.072 (std.dev.= 0.023) and 0.016 (std.dev.= 0.0042) for the parameters p, θ and P(Y =0) respectively. Lastly in Study-3, the RMSE and standard deviation of the MLE estimates are 0.015 (std.dev.= 0.2062), 0.154 (std.dev.= 0.643) and 0.043 (std.dev.= 0.032) and the corresponding Bayesian estimates are 0.016 (std.dev.= 0.0553), 0.155 (std.dev.= 0.204) and 0.031 (std.dev.= 0.035) for the parameters p, θ and P(Y =0) respectively.

By the similar way, we get the results for Study-1: 0.801, 0.168 and 0.159 respectively (for classical) and 0.772, 0.141 and 0.143 respectively (for Bayesian). For Study-2: 0.356, 0,321, 0.280 respectively (for classical) and 0.369, 0.301, 0.299 respectively (for Bayesian). Similarly, for Study-3: 0.240, 0.501, 0.272 respectively (for classical) and also get 0.147, 0.442, 0.258 (for Bayesian) respectively for the small sample (i.e., n=50). The values of RMSE from the two approaches are also quite similar when the sample size is large. But it is shown by the simulations that the Bayes estimator also has smaller values of RMSE in the case of small sample, i.e., $n = 50$ and $p = 0.1$ or $p = 0.9$ than the MLE. The simulation results suggest that the Bayesian method performs better in terms of larger coverage probabilities and smaller RMSE than maximum likelihood, especially in the case of small samples along with either very high or very low incidence of zero inflated outcomes.

Our simulation studies indicate that the Bayesian approach performs better because it yields larger coverage probabilities and smaller bias than the classical maximum likeli-

hood method, particularly in the case of small samples with either very high or very low incidence of excess zeros outcomes. When n is sufficiently large: MLE and Bayesian both perform well and the difference between the two approaches are almost identical (for this reason we get almost same RMSE in both cases). Therefore, MLEs and the Bayesian estimates behave very similar and they have the same asymptotic normal distribution (when n is very large).

3.4 Simulation studies in presence of covariates (using Bayesian approach)

For regression cases, normal distribution is assumed as prior for the regression parameters α and β. In particular, for the assumption of normal distribution we have considered mean as 0 with a very large variance 1000. A reasonable choice for the starting values of α and β for the Monte Carlo simulation chain can be prevailed by fitting the underlying models using statistical software (Ghosh et al. (2006)). In Table 4, posterior summary for zero-inflated model is presented. The 2.5% and 97.5% contribute an equal tail 95% posterior interval estimate for the parameters. In addition to such defect counts, we have also obtained data on other covariates that might explain the variation in the defect counts. Regression models with commonly used discrete distributions such as Poisson and Negative Binomial (Miaou (1994)), may not fit these data well, and seriously underestimate the zero-defect probability, which is an important indicator of heart disease. In a ZIP regression model, the covariates are usually linked to model parameters p and θ (Lambert (1992)). However, when covariates are present, a model having little more sophisticated algorithm such as data augmentation is required. Regression-type models are widely used in applied research to adjust for covariate effects and assess relationships between key predictors and the responses. While conventional regression models contain only one set of predictors for inference about a single response, covariates typically enter a ZIP regression model at both the Bernoulli zero-inflation and Poisson count stages, yielding two sets of parameters corresponding to p and θ. Thus, this allows simultaneous inferences to be made about the zero-inflated and count process. In the usual specification, covariates are related to θ through a log-linear model, and to p through a logit model. In previous section 3.2, we have considered no covariate cases for comparing the effects and assess relationships between key predictors and the response in this section 3.3. Since important predictors have already been chosen in Section 3.1, so we are considering only those covariates which are uttermost important for our current work.

The zero inflated models are fitted to link count data to mentioned covariates. In previous section, WinBUGS is used for fitting both regression models. Next, a five number summary (mean, std.dev., 2.5%, median and 97.5%) of parameters from zero inflated models have been presented in Table 4. A positive intercept in Table 4 of 11.456 with 95% posterior interval [0.386,40.643] indicates that the chance of being in the zero state is higher, moreover the sample mean of zero defect probability is 0.728 (which is

Table 4: Posterior summary of parameters (with covariates): ZIP model

Model	Parameter	Predictors	Mean	Sd	Percentile			ALCI
					25%	50%	97.5%	
ZIP (p, θ)	p	Intercept	11.456	2.34	3.053	11.289	19.496	
		Age	0.192	0.008	0.017	0.184	12.194	12.2
		Obesity	0.234	0.184	-1.002	0.198	22.279	23.3
		Weight	0.112	0.029	-0.142	0.117	15.122	15.2
		BMI	0.027	0.083	-2.132	0.017	19.024	21.5
		Sex	-0.7	0.532	-3.456	-0.865	10.701	14.1
		HBP	-0.005	0.01	-5.019	-0.098	11.13	16.1
		PR	-0.024	0.017	-1.724	-0.127	8.924	10.6
		PTS	1.22	1.07	0.001	0.034	21.28	21.2
		ETS	-0.407	0.45	-4.432	-0.977	17.412	21.8
		DM	-0.768	0.521	-6.768	-1.597	9.768	15.5
		HT	-0.119	0.428	-3.113	-0.154	8.169	11.2
		CRF	1.209	1.32	0.034	1.201	18.209	18.2
		Dylipidemia	0.392	0.351	-0.092	0.192	21.392	22.4
		WPP	0.707	1.399	-0.049	0.678	17.997	18.0
		LR	-0.855	1.031	-3.899	-0.987	13.855	17.7
		TCP	-0.761	0.355	-3.989	-0.861	21.761	24.6
		Dyspnea	0.392	0.351	-2.398	0.378	22.392	24.6
		Q-wave	-1.27	0.826	-4.297	-1.23	9.279	13.5
		LVH	1.212	0.75	0.002	1.103	19.212	19.2
		FBS	0.004	0.004	-2.004	0.001	8.654	10.6
		Creatine	0.224	0.664	-3.334	0.225	18.984	21.4
		Trygyceride	-0.005	0.003	-4.875	-0.012	8.005	12.9
		LLPD	0.007	0.005	-3.543	0.005	12.345	15.8
	θ	HLPD	0.005	0.017	-2.329	0.005	11.987	13.3
		Intercept	-2.393	0.098	-4.987	-2.389	-0.879	
		Age	2.57	0.002	0.985	2.55	1.939	0.95
		Obesity	0.556	0.104	0.234	0.508	2.279	1.9
		Weight	1.179	0.012	0.166	1.173	5.767	5.6
		BMI	1.132	0.033	0.182	1.132	3.398	3.2
		Sex	0.076	0.421	-0.019	0.065	2.991	0.07
		HBP	-0.002	0.009	-0.419	-0.002	3.751	4.1
		PR	1.693	0.012	0.677	-0.112	6.924	6.2
		PTS	0.192	0.938	-0.042	0.032	7.28	7.3
		ETS	-0.449	0.333	-2.121	-0.897	5.401	7.5
		DM	0.188	0.334	-1.178	-1.566	3.768	4.9
		HT	0.076	0.428	-2.006	-0.154	8.169	6.1
		CRF	0.062	1.292	-3.061	1.234	7.75	10.8
		Dylipidemia	0.607	0.281	-2.637	0.159	5.598	8.2
		WPP	-1.355	1.01	-3.336	0.787	4.998	7.3
		LR	-2.461	1.017	-5.437	-0.987	4.574	1.0
		TCP	0.892	0.223	-2.972	-0.909	4.912	6.8
		Dyspnea	-1.266	0.351	-3.284	0.895	8.354	11.6
		Q-wave	2.931	0.546	0.274	-1.242	8.75	8.5
		LVH	0.104	0.497	-1.892	0.103	8.643	9.5
		FBS	0.233	0.004	2.277	0.23	7.25	4.9
		Creatine	-1.005	0.562	-4.937	0.195	5.565	9.5
		Trygyceride	0.005	0.003	-2.744	0.005	7.586	10.1
		LLPD	0.031	0.002	-1.771	0.029	3.999	5.6
		HLPD	0.762	0.009	-2.712	0.654	3.909	6.7
	Sample Mean P(Y=0)		0.728	0.078	0.093	0.730	0.899	0.8
	Deviance		216.93	15.701	201.54	216.87	220.91	18.7

very close to the empirical percentage of the zero counts) with 95% posterior interval [0.683,0.752] but in the case of ZINB model it does not happen (sample mean of zero defect probability is 0.746). Apart from this, for ZINB model, our simulation studies indicate that the deviance is changed slightly to 234.87 from the deviance mentioned in Table 2 but it is shown in Table 4 that the deviance is dropped much to 216.93 from the deviance of Table 2 for the case of ZIP model.

3.5 Interpretations of the results related to heart disease

In the underlying model (ZIP), there may be a number of groups with different number of parameters related to heart disease. Numerical results of parameter estimation of model clearly demonstrate the efficacy of the proposed approach. With the reference level as male, we observe a negative coefficient -0.679 and a p-value 0.02 for female. The decrement is represented by a percentage of 49.29%, so the results suggest that the probability of heart disease counts is reduced for female as compare with male. Here it is also mentioned that there is a myth that coronary artery disease (CAD) is less common and less severe in women. But in our work, although we get that men are more affected than women but the heart disease due to CAD is not negligible in case of women. Since, a 50-year-old woman's risk of dying from CAD is 10 times more than her mortality risk from hip fracture and breast cancer combined. Although mortality from ischaemic heart disease (IHD) has declined but as per observations it is of lesser magnitude in women as compared to men of a similar age (Heron et al. (2006)). It is obtained that the estimated coefficient of 0.5158 (for obesity) is highly significant ($p < 0.01$), we deduce that the increasing effect of obesity on the expected number of CAD affected persons is about 67.50%. Because there is a positive association between obesity and cholesterol level (Veghari et al. (2013)). Diabetes mellitus (DM) is one of the highly risk factor of heart disease. In our work, it is observed that persons having diabetes mellitus are affected more in heart disease (CAD). Among adults (with DM) there is a prevalence of $70\% - 80\%$ for elevated low density lipoprotein (LDL), $60\% - 70\%$ for obesity and $75\% - 85\%$ of hypertension (Preis et al. (2009)). Diabetes mellitus (DM) is associated with increased mortality risk of heart disease. More than 70% of people older than 65 years with diabetes mellitus die from heart disease or stroke (Berry, Tardif and Bourassa (2007)). Similarly, it is observed that current smokers are more than 20% more risk of CAD than ex-smoker. Typical chest pain is one of the uttermost important risk factor in our studies. The increment of typical chest pain of a person is affected more in regional wall motion abnormalities. Besides, some secondary factors (i.e., creatine, dyslipidemia, congestive heart failure, Q-wave etc.) have deep influences on heart disease.

4 Concluding Remarks

In this work, we have analyzed the effect of most significant predictors that could explain the number of affected people in serious heart conditions. Various suitable models are fitted and through a careful modeling selection process, the ZIP model is recognized as the best among them (from Table 1) based on the data that have been used. The most

appropriate model (ZIP) is taken to examine the important factors that significantly influences the region of regional wall motion abnormalities related to coronary artery disease. Zero inflated Poisson model is best among all, although Poisson is worst among underlying models. We have obtained the significance of modelling for excess-zero in count data structure in the context of Bayesian method. Bayesian method has better finite sample performance than the classical method with tighter interval estimates and better coverage probabilities. It can also be concluded that Bayesian approach performs better than the classical maximum likelihood estimation in the sense of yielding larger coverage probabilities and smaller root mean square error. Besides, in this work, we have analyzed the effect of risk factors that could explain the number of victims for heart disease and get fruitful conclusions from the interpretations of the results related to heart disease. This analysis evidently endorses the perception of people concerning heart disease. The major risk factors for coronary heart disease are obesity, diabetes mellitus, cholesterol, smoking etc. Apart from this some secondary risk factors are also influenced to heart failure. As well, female reduce the risk of fatality as compare with male but the rate of female due to CAD is not imperceptible. In previous work it was only demonstrated assessing the importance of cardiovascular risk factors with various approaches (Pencina et al. (2019)). But in this work, we not only analyze the effect of risk factors but also perform the Bayesian approach. Besides, it is shown that Bayesian method has better finite sample performance than the classical method which was not performed in earlier work (Ghosh et al. (2006)). It was shown in earlier that if negative binomial is better than Poisson (for excess zero count data set) then ZINB is better than ZIP (Wiafe et al. (2018)). But in our work, we have shown that ZIP model is best among all, although Poisson is worst among all.

The zero inflated models have been evolved and utilized to manage such count data and is estimated conventionally using maximum likelihood estimator. Apart from this, Bayesian methods have been utilized for estimating ZIP model because these methods provide various advantages in compare with maximum likelihood estimation for this model. In this context, it is mentioned that Bayesian intervals (also known as credible intervals) give a strong impetus to adopt a Bayesian perspective (Gelman et al. (2004)). It is very influential (in case of ZIP model) that Bayesian analysis can provide full joint distribution of the parameters (in which we are interested) and to account for various sources of uncertainty in modelling zero-inflated count data, which is not easy to achieve in traditional maximum likelihood methods (Gelman et al. (2004)). Zero-inflated model has the following advantages: (i) it is useful for modeling outcomes of manufacturing processes and different situations where count data has excess zeros, (ii) it is also very useful for process optimization in presence of covariates. In this work, Bayesian analysis has been used to model such type of count data (with excess zeros) using sampling-based methods. From simulation studies, it can be also concluded that the proposed method is very effective for inferences based on small samples. We have also performed simulations based on small sample ($n = 25, 50$ etc.) which assess that the proposed Bayesian approaches provide better results than the maximum likelihood method for estimating the ZIP model, with larger c.p. (coverage probabilities) and smaller bias measured from root mean squared error (RMSE). It is very special case

for small samples. Because such cases (very high or very low incidence) are the mirrors of the presence of excess zeros outcomes. In case of small samples with parameters close to the boundary of the Bayesian intervals can result: (i) extra uncertainty in parameters and (ii) failure of the asymptotic assumptions. Both are critical in maximum likelihood estimation. These situation may create uncertainty on inferences about model parameters when consider the excess zeros count data based on the variance of the estimator under MLE approaches (Gelman et al. (2004)). In our work, the Bayesian approach performs better than the classic maximum likelihood estimation in the sense of providing larger c.p.(coverage probabilities) and smaller bias (from Tables 3 and 4). Moreover, Bayesian methods provide the inference of combinations of parameters, data, or both which is a major advantage of the Bayesian approach. Finally, for ductility of Bayesian analysis in modelling, mixture data has specific relevance in ZIP model. Count data with excess zeros is a special case of a two-stage mixed structure which is very natural candidate for Bayesian analysis.

The Poisson, negative binomial, zero-inflated Poisson and zero-inflated negative binomial models are fitted to the count-data for forward and backward model selection criteria. If the data has lots of zero, then between Poisson and negative binomial models, sometimes Poisson gives better result than negative binomial and sometimes negative binomial gives better result, after fitted the data to the model. If negative binomial gives better result than Poisson, then generally it was shown zero-inflated negative binomial is the best model of all: Poisson, negative binomial, zero-inflated Poisson and zero-inflated negative binomial, for excess zeros data (Wiafe et al. (2018)). On the other hand if Poisson gives better than negative binomial, then zero-inflated Poisson model is the best model among all (i.e., poisson, negative binomial, zero-inflated Poisson and zero-inflated negative binomial) for excess zeros data (Ghosh et al. (2006); Neelon (2018)). But in this work, Poisson model is worst among all but in spite of that zero-inflated Poisson model (ZIP) is the best fitted among the models (considered here) which have been verified with various procedures in this work.

The ZIP model is specifically suitable for modelling zero-inflated count data. Zero-inflated is a potential mechanism that generates over-dispersion in count data. Although the ZIP model is used as an example of a Bayesian approach to model zero-inflated count data, the limitation of the model in the presence of over-dispersion can result in biased parameter estimates. Apart from this, the interactions between available covariates is not contemplated in work. Other models such as the zero-inflated negative binomial (ZINB) model provides additional corrections for over-dispersion. This paper illustrates the superiority of Bayesian analytic model of count data characterized by excess zeros.

Acknowledgements

The authors are grateful to the learned reviewers, Prof. Maurizio Carpita (Editor) and Dr. Enrico Ciavolino (Editor) for their careful reading, valuable comments and helpful suggestions, which have helped them to improve the presentation of this work

significantly.

Data Availability Statement

The data used to support the findings of the study are available within the article.

Conflict of Interest

The authors declare that they have no conflict of interest regarding this work.

References

Agresti, A. (2002). An introduction to categorical data analysis. John Wiley and Sons, New York.

Alizadehsani, R., Habibi, J., Hosseini, M. J., Mashayekhi, H., Boghrati, R., Ghandeharioun, A., Bahadorian, B. and Sani, Z. A. (2013). A data mining approach for diagnosis of coronary artery disease. Computer Methods and Programs in Biomedicine, 111(1): 52-61.

Berry, C., Tardif, J. C. and Bourassa, M. G. (2007). Coronary heart disease in patients with diabetes: part II: recent advances in coronary revascularization. J Am CollCardiol, 49: 643-656.

Collins, R., Armitage, J., Parish, S., Sleigh, P. and Peto, R. (2003). MRC/BHF Heart Protection Study of cholesterol-lowering with simvastatin in 5963 people with diabetes: a randomised placebo-controlled trial. 361:2005-2016.

Czado, C., Erhardt, V., Min, A. and Wagner, S. (2007). Zero-inflated generalized Poisson models with regression effects on the mean, dispersion and zero-inflation level applied to patent outsourcing rates Statistical Modelling. journals.sagepub.com, 7(2): 125-153.

Fox, C. S., Coady, S., Sorlie, P. D., Levy, D., Meigs, J.B., D'Agostino, R.B., Wilson, P. W. and Savage, P. J. (2004). Trends in cardiovascular complications of diabetes. JAMA, 292: 2495-2499.

Gelman, A., Carlin, J. B., Stern, H. S., Rubin and D. B. (2004). Bayesian data analysis (2nd edition). Chapman and Hall/CRC, Boca Raton.

Gregg, E. W., Gu, Q., Cheng, Y. J., Narayan, K. M. and Cowie, C. C. (2007). Mortality trends in men and women with diabetes, 1971 to 2000. Ann Intern Med, 147: 149-155.

Ghosh, S. K., Mukhopadhyay, P. and Lu, J. C. (2006). Bayesian analysis of zero-inflated regression models. Journal of Statistical Planning and Inference, 136: 1360-1375.

Ghosh, S. and Samanta, G. P. (2019). Statistical modelling for cancer mortality. Letters in Biomathematics. 2019; Volume 6, No. 2. Available online: http://dx.doi.org/10.1080/23737867.2019.1581104

Gu, K., Cowie, C. C. and Harris, M. I. (1999). Diabetes and decline in heart disease mortality in US adults. JAMA, 281: 1291-1297.

Hansson, L., Zanchetti, A., Carruthers, S. G., Dahlof, B., Elmfeldt, D., Julius, S., Menard, J., Rahn, K. H., Wedel, H. and Westerling, S. (1998). Effects of intensive blood-pressure lowering and low-dose aspirin in patients with hypertension: principal results of the Hypertension Optimal Treatment (HOT) randomised trial. HOT Study Group. Lancet, 351: 1755-1762.

Heron, M., Hoyert, D. L., Murphy, S. L., Xu, J., Kochanek, K. D. and Tejada-Vera, B. (2006). Deaths: preliminary data for 2006. Natl Vital Stat Rep 2008, 56: 1-52.

Karaolis, M. A., Moutiris, J. A and Hadjipanayi, D. (2010). Assessment of the Risk Factors of Coronary Heart Events Based on Data Mining with Decision Trees. IEEE Transactions on Information Technology in Biomedicine, 14: 559-566.

Lambert, D. (1992). Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. Technometrics, 34(1): 1-14.

Liu, H. and Powers, D. A. (2012). Bayesian Inference for Zero-Inflated Poisson Regression Models. Journal of Statistics: Advances in Theory and Applications, 7(2): 155-188.

Lu, L., Fu, Y., Chu, P. and Zhang, X. (2014). A Bayesian Analysis of Zero-Inflated Count Data: An Application to Youth Fitness Survey. Tenth International Conference on Computational Intelligence and Security, Kunming, 699-703, doi:10.1109/CIS.2014.125.

Miaou, P. (1994). The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. Accident Analysis and Prevention, 26(4): 471-482.

Mullahy, J. (1986). Specification and testing of some modified count data models. Journal of Econometrics, 33(3): 341-365.

Neelon, B. (2018). Bayesian Zero-Inflated Negative Binomial Regression Based on Pólya-Gamma Mixtures. Bayesian Analysis, 14(3): 829-855.

Pencina, M. J., Navar, A. M., Wojdyla, D., Sanchez, R. J., Khan, I., Elassal, J., D'Agostino, R. B., Peterson, E. D. and Sniderman, A. D. (2019). Quantifying Importance of Major Risk Factors for Coronary Heart Disease. Circulation, 139: 1603-1611, DOI:10.1161/CIRCULATIONAHA.117.031855.

Preis, S. R., Pencina, M. J., Hwang, S. J., D'Agostino, R. B. Sr., Savage, P. J., Levy, D. and Fox, C. S. (2009). Trends in cardiovascular disease risk factors in individuals with and without diabetes mellitus in the Framingham Heart Study, Circulation, 212-220.

Safford, R. E. and Bove, A. A. (1987). Prediction of coronary artery disease by left ventricular regional wall motion abnormalities in patients with stenosis of the aortic valve. Br Heart J, 57, 237-41.

Shankar, V., Milton, J. and Mannering, F. (1997), Modelling Accident Frequencies as Zero-Altered Probability Processes: An Empirical Inquiry. Accid. Anal. and Prev., 29(6): 829-837.

Shepherd, J., Barter, P., Carmena, R., Deedwania, P., Fruchart, J. C., Haffner, S., Hsia,

J., Breazna, A., LaRosa, J., Grundy, S. and Waters, D. (2006). Effect of lowering LDL cholesterol substantially below currently recommended levels in patients with coronary heart disease and diabetes: the Treating to New Targets (TNT) study. Diabetes Care, 29: 1220-1226.

Veghari, G., Sedaghat, M., Joshghani, H., Banihashem, S., Moharloei, P., Angizeh, A., Tazik, E. and Moghaddami, A. (2013). Obesity and risk of hypercholesterolemia in Iranian northern adults. ARYA Atheroscler, 9(1): 2-6.

Vuong, Q. H. (1989). Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. Econometrica, 57(2): 307-333.

Wang, Z., Shuangge, M. and Wang, C. Y. (2015). Variable selection for zero-inflated and overdispersed data with application to health care demand in Germany. Biom J, 57(5): 867-884.

Wiafe, E. S., Kumi, A. A., Nortey, E. N. N. and Idd, S. (2018). Modelling vehicular crash mortalities in Ghana. Model Assisted Statistics and Applications, 13: 287-295, DOI10.3233/MAS-180433, IOS Press.

World Health Organization (2019). Available online: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death.

Zeileis, A., Kleiber, C. and Jackman, S. (2008). Regression Models for Count Data in R, Journal of Statistical Software, 27(8): 1-25.

Zeng, P., Wei, Y., Zhao, Y., Liu, J., Liu, L., Zhang, R., Gou, J., Huang, S. and Chen, F. (2014). Variable selection approach for zero-inflated count data via adaptive lasso. Journal of Applied Statistics, 41(4): 879-894.

Zhang, Z. (2016). Variable selection with stepwise and best subset approaches. Ann Transl Med, 136, doi:10.21037/atm.2016.03.35.