Thailand Statistician
July 2013; 11(2): 111-131
http://statassoc.or.th
Contributed paper

Bayesian Analysis of Zero–Altered Poisson Regression Models

Thipwan Kunthong* [a], Sujit K. Ghosh [b] and Saengla Chaimongkol [a]

[a] Department of Mathematics and Statistics, Thammasat University, Rangsit Center, Pathum Thani 12120, Thailand.

[b] Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203, USA.

* Corresponding author; e-mail: thipwan@grad.sci.tu.ac.th

Received: 15 October 2012
Accepted: 23 May 2013

Abstract

This study introduces the generalized zero-altered Poisson regression model with the suitable link functions of parameters. There are three different models based on the effects of parameters in the generalized zero-altered Poisson regression models. The Bayesian approach is studied which the prior distributions of regression parameters in both linear predictors are specified as independent normal distributions for the fixed effects and inverse-gamma distributions for the random effects. The Bayesian estimation method can be carried out using WinBUGS. Simulation study in the generalized zero-altered Poisson regression models illustrates that the Bayesian approach is satisfactory estimation method for fixed effects model (Model I) at large sample sizes and even larger sample sizes for the mixed models (Model II and Model III). For application, generalized zero-altered Poisson regression models are applied to the number of births of a reproductive woman in the south of Thailand. The results showed that number of births was significantly affected by age of women, number of household members, age at first birth, and number of additional children wanted in both linear predictors. In addition, the religion, place of residence, and education were significantly effects in the Poisson linear predictor. The age at first marriage was significantly effect in the dispersion linear predictor.

Keywords: Bayesian, GLMs, zero-altered Poisson distribution, dispersed probability models, WinBUGS.

1. Introduction

When we analyze count data, it is useful to investigate the patterns of dispersion using exploratory data analysis. In recent year there has been considerable interest in developing models for count data that allow for excess zeroes. Such high frequencies of zero counts often leads to over dispersion and over dispersed probability models, such as generalized or zero inflated Poisson distributions are generally used in practice [1-4]. Ridout et al. [5] presented an overview of zero inflated models. Winkelmann and Zimmermann [6] provided a thorough review of the statistical models for count data and also presented many potential applications with underdispersion.

Generalized or zero-inflated Poisson distributions are commonly used to model the count data. Such distributions can account for the overdispersion due to many zero counts. However, in many applications of count data show evidence of underdispersion such as airline failures, number of changes of employer, number of births by women [6]. As the case of underdispersion needs awkward parameter restrictions, the generalized or zero-inflated Poisson distributions are often inadequate. Recently, a new class of generalizations of the Poisson distribution that can account for both under and over dispersion has introduced [7]. However, Ghosh and Kim [8] pointed that such distributions are somewhat inflexible in practice. Therefore, they proposed the more flexible and advantage class of zero-altered distributions which can account for both types of dispersion and include other familiar models.

In modeling count data, other controllable factors (covariates) that might explain the variation in the counts. Regression models are very useful to model such data. Ghosh et al. [9] introduced Zero Inflated Poisson (ZIP) regression models which is a special case in the class of zero inflated models including other familiar models. A Bayesian estimation method is applied by using sampling-based methods. The proposed method has better finite sample performance than the classical method by simulation studies with tighter interval estimates and better coverage probabilities. They used WinBUGS to illustrate the performance of the proposed method by applying it to a real-life data set. Angers and Biswas [10] proposed a zero-inflated generalized Poisson model and a Bayesian analysis can be considered for some appropriate priors and the posteriors are obtained using Monte Carlo integration with importance sampling. The real-life data set is applied to these methods. This paper shows that the Poisson model is a misfit that badly underestimates the number of zero counts. The ZIP is a misfit that does not provide good estimates of the nonzero counts. Melkersson and Rooth [11] proposed a zero-and-two-inflated count data model illustrated a relative excess of zero and two children. The Poisson and gamma count distributions are used in the model using the fertility data of Swedish women. Wang and Famoye [12] studied the modeling for household fertility decisions by using a generalized Poisson regression model. The model is estimated by the maximum likelihood estimation method and discussed the suitable model by tests for dispersion and goodness-of-fit measures.

Bayesian approach is widely applied for fitting several models such as zero-inflated generalized Poisson model [10], zero-inflated regression model [9], and differential item functioning model [13, 14]. Gelman et al. [15] provided an excellent introduction to Bayesian data analysis. Usually the joint posterior distribution is complex and unavailable in closed form, thus simulation-based method broadly known as Markov Chain Monte Carlo (MCMC) [16] required to obtain the point and interval estimates of the parameters. MCMC algorithm can be used WinBUGS software to perform all the required computations. This research, the zero-altered Poisson regression models included the effect of covariates will be proposed and developed. Our proposed models turn out to be analytically intractability, hence a Bayesian approach is developed as an alternative to classical statistical methods based on the maximum likelihood estimate.

The article is organized as follows: In Section 2, we describe the generalized zero-altered Poisson regression models. In Section 3, we present the procedure of simulation study for the generalized zero-altered Poisson regression model analyzed by the Bayesian method using WinBUGS. In Section 4, the results of the simulation study for the generalized zero-altered Poisson regression model are demonstrated. In Section 5, the application of the generalized zero-altered Poisson regression model are presented. Conclusions are given in Section 6.

2. The Zero-Altered Poisson Regression Models

For a random sample of observations $y_1, y_2, \ldots, y_n$, where n is the sample size, the zero-altered Poisson regression model is applied with probability mass function [8].

$$f\left(y_i\,|\delta_i,\lambda_i\right)=\begin{cases}\delta_{i+}+\left(1-|\delta_i|\right)e^{-\lambda_i} & if\ y_i=0\\ \left(1-\delta_{i+}+\delta_{i-}\left\{\dfrac{e^{-\lambda_i}}{1-e^{-\lambda_i}}\right\}\right)\dfrac{\lambda_i^{y_i}e^{-\lambda_i}}{y_i\,!} & if\ y_i=1,2,\ldots\end{cases}$$

where $\delta_i\in\left(-1,\,1\right)$, $\lambda_i>0$, $\delta_{i+}=\max\left\{\delta_i,0\right\}$ and $\delta_{i-}=\max\left\{-\delta_i,0\right\}$. The mean and variance are

$$\mu_i = \mathrm{E}\left[Y_i\right] = \left(1-\delta_{i+}+\delta_{i-}\left\{\frac{e^{-\lambda_i}}{1-e^{-\lambda_i}}\right\}\right)\lambda_i = \omega(\delta_i,\lambda_i)\lambda_i$$

and

$$\sigma_i^2 = \mathrm{Var}\left[Y_i\right]= \omega(\delta_i,\lambda_i)\lambda_i+\frac{1-\omega(\delta_i,\lambda_i)}{\omega(\delta_i,\lambda_i)}\mu_i^2 \text{ , respectively.}$$

In the zero-altered Poisson (ZAP) regression models, parameter vectors are denoted as $\boldsymbol{\delta}$ = $\left(\delta_1,\delta_2,\,\ldots,\delta_n\right)'$ and $\boldsymbol{\lambda}=\left(\lambda_1,\lambda_2,\,\ldots,\lambda_n\right)'$. From the random vector $\mathbf{Y}$ = $\left(Y_1,Y_2,\,\ldots,Y_n\right)'$ of size $n\times1$, be a discrete random vector with the zero-altered Poisson distribution: ZAP(δ_i, λ_i) for each random variable Y_i; i = 1, 2, …, n. For the independently distributed responses Y_i's from ZAP(δ_i, λ_i), the appropriately used link functions are considered as follows:

$$\log\left(\lambda_i\right)=\mathbf{x}_i'\boldsymbol{\beta} \quad \text{or} \quad \log\left(\lambda_i\right)=\mathbf{x}_i'\boldsymbol{\beta}+\mathbf{u}_i'\boldsymbol{\alpha}_i$$

where $\mathbf{x}_i'$ is the *i*th row of the $\mathbf{X}_{(n\times p)}$ covariate matrix with $p-1$ covariates when the intercept term included in the model. The dispersion parameter $\delta_i\in\left(-1,\,1\right)$ so we can specify Fisher's *z*' transformation function to be the linear predictor for δ_i, that is,

$$\frac{1}{2}\log\left(\frac{1+\delta_i}{1-\delta_i}\right)=\mathbf{z}_i'\boldsymbol{\gamma} \quad \text{or} \quad \frac{1}{2}\log\left(\frac{1+\delta_i}{1-\delta_i}\right)=\mathbf{z}_i'\boldsymbol{\gamma}+\mathbf{v}_i'\boldsymbol{\tau}_i$$

where $\mathbf{z}_i'$ is the *i*th row of the $\mathbf{Z}_{(n\times q)}$ covariate matrix with $q-1$ covariates when the intercept term included in the model. The vectors $\boldsymbol{\beta}$ = $\left(\beta_0,\beta_1,\,\ldots,\beta_{p-1}\right)'$ and $\boldsymbol{\gamma}$ =

$\left(\gamma_0, \gamma_1, \ldots, \gamma_{q-1}\right)'$ are the corresponding $p \times 1$ and $q \times 1$ vectors of unknown parameters (regression coefficients) associated with $\mathbf{X}_{(n\times p)}$ and $\mathbf{Z}_{(n\times q)}$, respectively. In general, the covariates $\mathbf{X}_{(n\times p)}$ and $\mathbf{Z}_{(n\times q)}$ may or may not be the same matrices. If the covariate matrices are the same, we need $2p$ regression parameters in the zero-altered regression model.

Then the parameters are modeled by the link functions

Model I: $\log\left(\lambda_i\right) = \mathbf{x}_i'\boldsymbol{\beta}$ and $\frac{1}{2}\log\left(\frac{1+\delta_i}{1-\delta_i}\right) = \mathbf{z}_i'\boldsymbol{\gamma}$,

Model II: $\log\left(\lambda_i\right) = \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{u}_i'\boldsymbol{\alpha}_i$ and $\frac{1}{2}\log\left(\frac{1+\delta_i}{1-\delta_i}\right) = \mathbf{z}_i'\boldsymbol{\gamma}$,

Model III: $\log\left(\lambda_i\right) = \mathbf{x}_i'\boldsymbol{\beta}$ and $\frac{1}{2}\log\left(\frac{1+\delta_i}{1-\delta_i}\right) = \mathbf{z}_i'\boldsymbol{\gamma} + \mathbf{v}_i'\boldsymbol{\tau}_i$

for covariate vectors $\mathbf{x}_i'$, $\mathbf{z}_i'$, $\mathbf{u}_i'$, and $\mathbf{v}_i'$ where i = 1, 2, …, n. $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ are respectively $p \times 1$ and $q \times 1$ vectors of unknown parameters (the fixed effects). $\boldsymbol{\alpha}_i$, $\boldsymbol{\tau}_i$ are respectively $r \times 1$ and $s \times 1$ vectors of unknown parameters (the random effects). $\alpha_{i1}, \ldots, \alpha_{ir}$ are independent and distributed as α_{ij} ~ Normal(α_j, $\sigma^2_{\alpha_j}$) for j = 1, 2, …, r, and $\tau_{i1}, \ldots, \tau_{is}$ are independent and distributed as τ_{ik} ~ Normal(τ_k, $\sigma^2_{\tau_k}$) for k = 1, 2, …, s. The parameters are estimated by Bayesian method using WinBUGS. So we get the estimates of δ_i and λ_i separated to 3 models as following:

Model I: $\hat{\lambda}_i = \exp\left(\mathbf{x}_i'\hat{\boldsymbol{\beta}}\right)$ and $\hat{\delta}_i = \frac{\exp\left\{2\left(\mathbf{z}_i'\hat{\boldsymbol{\gamma}}\right)\right\}-1}{\exp\left\{2\left(\mathbf{z}_i'\hat{\boldsymbol{\gamma}}\right)\right\}+1}$,

Model II: $\hat{\lambda}_i = \exp\left(\mathbf{x}_i'\hat{\boldsymbol{\beta}} + \mathbf{u}_i'\hat{\boldsymbol{\alpha}}_i\right)$ and $\hat{\delta}_i = \frac{\exp\left\{2\left(\mathbf{z}_i'\hat{\boldsymbol{\gamma}}\right)\right\}-1}{\exp\left\{2\left(\mathbf{z}_i'\hat{\boldsymbol{\gamma}}\right)\right\}+1}$,

Model III: $\hat{\lambda}_i = \exp\left(\mathbf{x}_i'\hat{\boldsymbol{\beta}}\right)$ and $\hat{\delta}_i = \frac{\exp\left\{2\left(\mathbf{z}_i'\hat{\boldsymbol{\gamma}} + \mathbf{v}_i'\hat{\boldsymbol{\tau}}_i\right)\right\}-1}{\exp\left\{2\left(\mathbf{z}_i'\hat{\boldsymbol{\gamma}} + \mathbf{v}_i'\hat{\boldsymbol{\tau}}_i\right)\right\}+1}$.

2.2 Bayesian Analysis

The likelihood function of $y_1, y_2, \ldots, y_n$ which are observations from the ZAP distribution in terms of the inverted link functions of dispersion and Poisson parameters is given by

$$
\begin{aligned}
L(\delta_i, \lambda_i) &= L\left(\frac{\exp\{2(\mathbf{z}_i'\boldsymbol{\gamma})\}-1}{\exp\{2(\mathbf{z}_i'\boldsymbol{\gamma})\}+1}, \exp(\mathbf{x}_i'\boldsymbol{\beta}) \right) \\
&= \prod_{i=1}^{n} f\left(y_i \left| \frac{\exp\{2(\mathbf{z}_i'\boldsymbol{\gamma})\}-1}{\exp\{2(\mathbf{z}_i'\boldsymbol{\gamma})\}+1}, \exp(\mathbf{x}_i'\boldsymbol{\beta}) \right. \right)
\end{aligned}
$$

$$
= \left(1 - \left(\frac{\exp\{2(\mathbf{z}_i'\boldsymbol{\gamma})\}-1}{\exp\{2(\mathbf{z}_i'\boldsymbol{\gamma})\}+1} \right)_{+} + \left(\frac{\exp\{2(\mathbf{z}_i'\boldsymbol{\gamma})\}-1}{\exp\{2(\mathbf{z}_i'\boldsymbol{\gamma})\}+1} \right)_{-} \left\{ \frac{e^{-\exp(\mathbf{x}_i'\boldsymbol{\beta})}}{1-e^{-\exp(\mathbf{x}_i'\boldsymbol{\beta})}} \right\} \right)^{n-n_0}
$$

$$
\times \frac{\lambda^{\sum_{y_i>0} y_i} e^{-(n-n_0)\exp(\mathbf{x}_i'\boldsymbol{\beta})}}{\prod_{y_i} y_i!}
$$

$$
\times \left(\left(\frac{\exp\{2(\mathbf{z}_i'\boldsymbol{\gamma})\}-1}{\exp\{2(\mathbf{z}_i'\boldsymbol{\gamma})\}+1} \right)_{+} + \left(1 - \left| \frac{\exp\{2(\mathbf{z}_i'\boldsymbol{\gamma})\}-1}{\exp\{2(\mathbf{z}_i'\boldsymbol{\gamma})\}+1} \right| \right) e^{-\exp(\mathbf{x}_i'\boldsymbol{\beta})} \right)^{n_0},
$$

where n_0 be the frequency of the zero count.

The Bayesian generalized linear models could be fitted by incorporating the prior information directly on the regression parameters through multivariate normal, i.e., $\pi(\boldsymbol{\beta})$~ Normal($\boldsymbol{\beta}_\mathbf{0}$, $\boldsymbol{\sigma}_\mathbf{0}^\mathbf{2}$). Choice of vague prior would be with $\boldsymbol{\beta}_\mathbf{0} = \mathbf{0}$ and $\boldsymbol{\sigma}_\mathbf{0}^\mathbf{2} = c\mathbf{I}$, where c is a very large number. From the structure of link functions and the previous study, the prior distributions for regression coefficients (fixed effects) are the normal distributions with mean 0 and variance 100 for regression coefficients ($\boldsymbol{\beta}$) in the linear predictor of Poisson parameter. For regression coefficients ($\boldsymbol{\gamma}$) in the linear predictor of dispersion parameter, the prior distributions are the normal distributions with mean 0 and variance 10. For the random effects, the prior distributions for the variance of the random terms in Model II and Model III are the inverse-gamma distributions.

3. Simulation Study

Let a model with intercept and covariates $x_{i1} = z_{i1}$, $x_{i2} = z_{i2}$, u_{i1} and v_{i1} were considered for linear predictors. The true values of parameters β_0, β_1, β_2, γ_0, γ_1, and γ_2 will be set for the expected values of λ_i^* = 1.0, 3.0, 5.0 and δ_i^* = –0.5, 0.0, 0.5 (represent for the under- equi- and over-dispersion). The steps of simulation study for the generalized zero-altered Poisson regression model would be:

Step 1: Specify the true values of regression parameters β_0, β_1, β_2, γ_0, γ_1, and γ_2 for setting the Poisson parameter values and the dispersion parameter values, respectively.

The values of these parameters are as following:

β_0 = 0.0, log 3.0, log 5.0, $\beta_1 = \beta_2$ = 1.0,

$\gamma_0 = -\frac{1}{2}$ log 3.0, 0.0, $\frac{1}{2}$ log 3.0, $\gamma_1 = \gamma_2$ = 1.0,

α_{i1} ~ Normal($\alpha_1, \sigma_{\alpha_1}^2$), τ_{i1} ~ Normal($\tau_1, \sigma_{\tau_1}^2$),

for $\alpha_1 = \tau_1 = 0$ and $\sigma_{\alpha_1}^2 = \sigma_{\tau_1}^2 = 1$. Assume the sample sizes n = 50, 100 and 300.

Step 2: Generate the values of the covariates x_{i1} ~ Bernoulli(0.5), x_{i2} ~ Normal(0, 1), u_{i1} ~ Normal(0, 1), and v_{i1} ~ Normal(0, 1) for i = 1, 2, …, n using R [17].

Step 3: Calculate the values of the Poisson and dispersion parameters from each model

Model I:
$$\lambda_i^* = \exp\left(\beta_0 + \beta_1\left(x_{i1} - 0.5\right) + \beta_2 x_{i2}\right),$$

and
$$\delta_i^* = \frac{\exp\left\{2\left(\gamma_0 + \gamma_1\left(x_{i1} - 0.5\right) + \gamma_2 x_{i2}\right)\right\} - 1}{\exp\left\{2\left(\gamma_0 + \gamma_1\left(x_{i1} - 0.5\right) + \gamma_2 x_{i2}\right)\right\} + 1},$$

Model II:
$$\lambda_i^* = \exp\left(\beta_0 + \beta_1\left(x_{i1} - 0.5\right) + \beta_2 x_{i2} + \alpha_{i1} u_{i1}\right),$$

and
$$\delta_i^* = \frac{\exp\left\{2\left(\gamma_0 + \gamma_1\left(x_{i1} - 0.5\right) + \gamma_2 x_{i2}\right)\right\} - 1}{\exp\left\{2\left(\gamma_0 + \gamma_1\left(x_{i1} - 0.5\right) + \gamma_2 x_{i2}\right)\right\} + 1},$$

Model III:
$$\lambda_i^* = \exp\left(\beta_0 + \beta_1\left(x_{i1} - 0.5\right) + \beta_2 x_{i2}\right),$$

and $$\delta_i^* = \frac{\exp\left\{2\left(\gamma_0+\gamma_1\left(x_{i1}-0.5\right)+\gamma_2 x_{i2}+\tau_{i1}v_{i1}\right)\right\}-1}{\exp\left\{2\left(\gamma_0+\gamma_1\left(x_{i1}-0.5\right)+\gamma_2 x_{i2}+\tau_{i1}v_{i1}\right)\right\}+1}.$$

by using the generated values of covariates and specified coefficients in Step 1 and Step 2.

Step 4: Generate y_i from the zero-altered Poisson distribution with the dispersion parameter values δ_i^* and the Poisson parameter values λ_i^* or ZAP(δ_i^* , λ_i^*) in Step 3 for i = 1, 2, …, n .

Step 5: Assume the prior independence distributions are the normal distribution for all regression parameters and the inverse-gamma distribution for the variance of random effects such that

β_0 ~ Normal(0, 100), β_1 ~ Normal(0, 100), β_2 ~ Normal(0, 100),

γ_0 ~ Normal(0, 10), γ_1 ~ Normal(0, 10), γ_2 ~ Normal(0, 10),

$\frac{1}{\sigma^2_{\alpha_1}}$ ~ Gamma(0.1, 0.1), $\frac{1}{\sigma^2_{\tau_1}}$ ~ Gamma(0.1, 0.1).

Step 6: Estimate the parameters by programming WinBUGS and setting a burn-in period of 10,000 iterations of 20,000 samples and 3 chains. The WinBUGS code can be requested from the corresponding author.

Step 7: Replicate the procedure from Step 2 to Step 6 that is the Monte Carlo (MC) simulated data 1,000 sets.

Step 8: Compute the performance of estimates considering the bias, standard error (S.E.), 2.5 percentile, the posterior median, 97.5 percentile. Illustrate the performance of estimates by the boxplots.

Posterior inferences can be evaluated using the concept of calibration of the posterior mean (or median), the Bayesian analogue to the classical notion of “bias.” For parameter θ, we label the posterior median as $\hat{\theta}$ and define the miscalibration of the posterior median as E($\hat{\theta}$) − θ, for any value of θ. If the prior distribution is true—that is, if the data are constructed by first drawing θ from p(θ), then drawing y from p(y|θ)—then the posterior median is automatically calibrated; that is its miscalibration is 0 for all values of θ.

4. Results

This research presents the results of simulation study for three different ZAP regression models. Model I named “the fixed effects model” treats both linear predictors of Poisson and dispersion parameters as fixed effects. Model II and Model III named “the mixed effects model” comprises of either fixed or random linear predictor of Poisson and dispersion parts. In Model II, the linear predictor of Poisson parameters was randomized while the linear predictor of dispersion parameter was fixed, and vice versa for the Model III. For the fixed effects model (Model I), the performance of parameter estimates in the Poisson linear predictor at n = 300 is shown in Figure 1. The performance of parameter estimates in the dispersion linear predictor at n = 300 is shown in Figure 2. At n = 300, Figures 3-4 present the performance of variance component estimates for the random term in the Poisson linear predictor (Model II) and the dispersion linear predictor (Model III), respectively. In addition, the performance of variance component estimates for the random term in the dispersion linear predictor (Model III) at n = 500 is presented in Figure 4.

The results of simulation study can be summarized that the generalized zero-altered Poisson regression model in which both linear predictors were fixed (Model I), the Bayesian approach is an efficient estimation method. However, estimation of regression parameters in the Poisson linear predictor requires large sample sizes ($n \geq$ 100) at the small Poisson parameter value (λ = 1.0) and even larger sample sizes for estimation of regression parameters in the dispersion linear predictor at the underdispersion data set for all Poisson parameter values.

Similarly, for Model II (where the Poisson linear predictor was randomized while the dispersion linear predictor was fixed in the generalized zero-altered Poisson regression model), the Bayesian estimation approach is satisfactory for regression parameters and variance component in the Poisson linear predictor at very large sample sizes ($n \geq$ 300) for all combinations of parameters (λ, δ) in the ZAP model. The regression parameter estimation in the dispersion linear predictor requires large sample sizes for the underdispersion data set.

Table 1. Performances of regression coefficient estimates in the Poisson linear predictor for Model I.

n	Perfor-mances	underdispersion						equi-dispersion					
		λ = 1.0			λ = 5.0			λ = 1.0			λ = 5.0		
		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
50	Bias	-0.1637	0.0696	0.0332	-0.0155	0.0072	0.0048	-0.2480	0.0641	0.0886	-0.0283	0.0018	0.0120
	S.E.	0.2589	0.4702	0.2867	0.0826	0.1562	0.1039	0.3398	0.6480	0.3908	0.1033	0.2023	0.1402
	2.50%	-0.7350	0.2086	0.5057	1.4257	0.7024	0.8100	-1.0155	-0.1571	0.4098	1.3686	0.6144	0.7478
	Median	-0.1637	1.0696	1.0332	1.5939	1.0072	1.0048	-0.2480	1.0641	1.0886	1.5811	1.0018	1.0120
	97.50%	0.2814	2.0574	1.6312	1.7494	1.3162	1.2171	0.3155	2.3907	1.9423	1.7736	1.4092	1.2965
100	bias	-0.0683	0.0414	0.0192	-0.0133	0.0097	0.0050	-0.1063	0.0789	0.0277	-0.0157	0.0046	0.0026
	S.E.	0.1645	0.2834	0.1805	0.0590	0.1070	0.0663	0.2020	0.3952	0.2741	0.0675	0.1437	0.0854
	2.50%	-0.4170	0.5036	0.6789	1.4769	0.8009	0.8789	-0.5398	0.3466	0.5122	1.4567	0.7244	0.8387
	Median	-0.0683	1.0414	1.0192	1.5961	1.0097	1.0050	-0.1063	1.0789	1.0277	1.5937	1.0046	1.0026
	97.50%	0.2298	1.6171	1.3877	1.7084	1.2214	1.1388	0.2530	1.8998	1.5873	1.7215	1.2888	1.1725
300	bias	-0.0212	0.0125	0.0091	-0.0030	-0.0004	0.0015	-0.0344	0.0072	0.0114	-0.0060	0.0006	0.0031
	S.E.	0.0908	0.1563	0.0966	0.0319	0.0600	0.0401	0.1080	0.2116	0.1257	0.0355	0.0738	0.0513
	2.50%	-0.2082	0.7116	0.8246	1.5429	0.8819	0.9246	-0.2560	0.5967	0.7718	1.5324	0.8551	0.9050
	Median	-0.0212	1.0125	1.0091	1.6064	0.9996	1.0015	-0.0344	1.0072	1.0114	1.6034	1.0006	1.0031
	97.50%	0.1489	1.3254	1.2039	1.6680	1.1178	1.0816	0.1672	1.4275	1.2643	1.6714	1.1452	1.1055

Table 1. (continued) Performances of regression coefficient estimates in the Poisson linear predictor for Model I.

n	Perfor-mances	overdispersion					
		λ = 1.0			λ = 5.0		
		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
50	Bias	-0.4534	-0.0863	0.0153	-0.0405	0.0099	0.0263
	S.E.	0.6272	1.1959	0.6373	0.1413	0.2725	0.2045
	2.50%	-1.9821	-1.4640	-0.0929	1.2744	0.4766	0.6414
	Median	-0.4534	0.9137	1.0153	1.5689	1.0099	1.0263
	97.50%	0.4555	3.2299	2.4089	1.8282	1.5480	1.4405
100	bias	-0.2254	0.0038	0.0627	-0.0226	0.0022	0.0038
	S.E.	0.3338	0.6540	0.3842	0.0930	0.1877	0.1295
	2.50%	-0.9846	-0.2931	0.3600	1.3954	0.6301	0.7578
	Median	-0.2254	1.0038	1.0627	1.5868	1.0022	1.0038
	97.50%	0.3218	2.2803	1.8680	1.7604	1.3674	1.2640
300	bias	-0.0716	-0.0180	0.0052	-0.0057	0.0006	0.0009
	S.E.	0.1498	0.3061	0.1983	0.0484	0.0997	0.0662
	2.50%	-0.3848	0.3806	0.6212	1.5057	0.8056	0.8733
	Median	-0.0716	0.9820	1.0052	1.6037	1.0006	1.0009
	97.50%	0.2030	1.5828	1.3989	1.6958	1.1970	1.1320

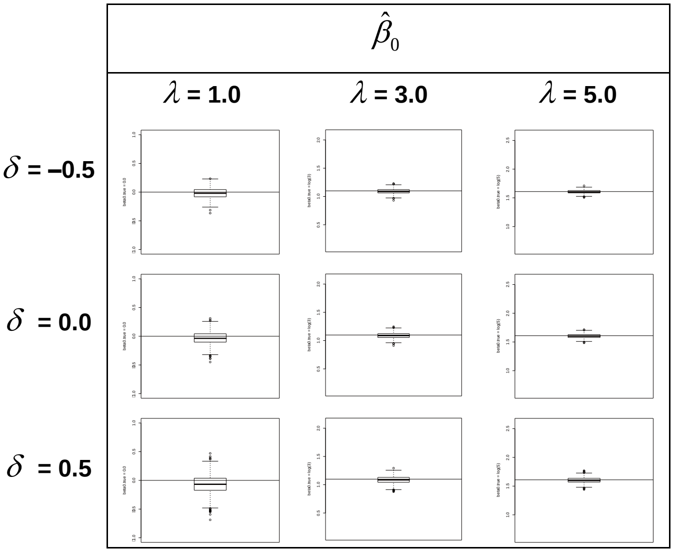

Figure 1. Performance of regression coefficient estimates ($\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$) in the Poisson linear predictor for Model I at n = 300.

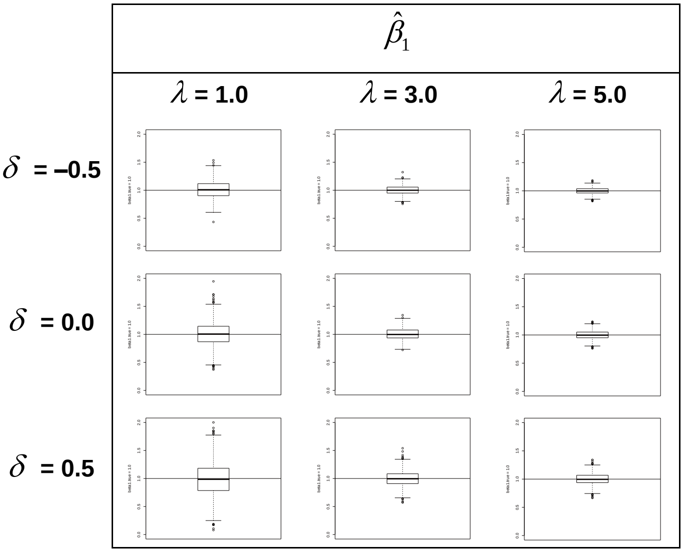

Figure 1. (continued) Performance of regression coefficient estimates ($\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$) in the Poisson linear predictor for Model I at n = 300.

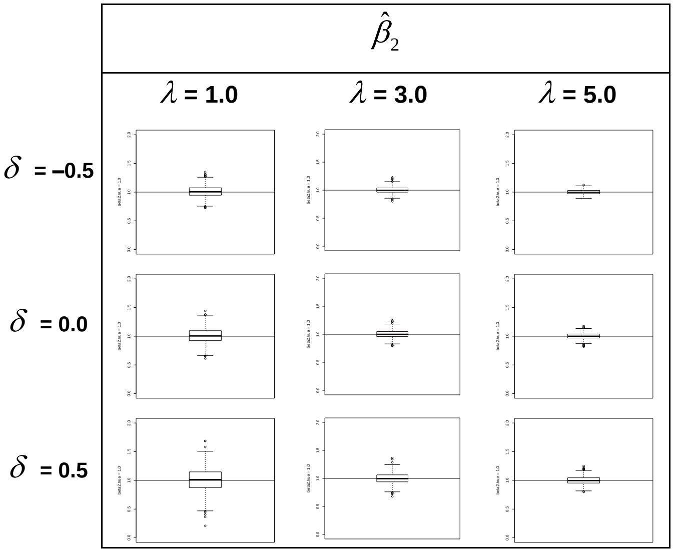

Figure 1. (continued) Performance of regression coefficient estimates ($\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$) in the Poisson linear predictor for Model I at n = 300.

Table 2. Performances of regression coefficient estimates in the dispersion linear predictor for Model I.

n	**Perfor-mances**	**λ = 1.0**								
		δ = -0.5			**δ = 0.0**			**δ = 0.5**		
		$\hat{\gamma}_0$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_0$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_0$	$\hat{\gamma}_1$	$\hat{\gamma}_2$
50	**bias**	-0.2229	0.2130	0.2089	-0.1115	0.0958	0.1505	-0.0846	-0.0072	0.0545
	S.E.	0.3001	0.4940	0.3207	0.2087	0.4163	0.2775	0.2491	0.4693	0.3185
	2.50%	-1.4480	0.3533	0.6752	-0.5380	0.3188	0.6857	0.0051	0.1038	0.4946
	Median	-0.7722	1.2130	1.2089	-0.1115	1.0958	1.1505	0.4647	0.9928	1.0545
	97.50%	-0.2729	2.2906	1.9283	0.2819	1.9521	1.7698	0.9810	1.9412	1.7415
100	**bias**	-0.1105	0.1123	0.1023	-0.0465	0.0657	0.0574	-0.0424	0.0282	0.0344
	S.E.	0.1847	0.2898	0.1965	0.1363	0.2735	0.2004	0.1696	0.3232	0.1936
	2.50%	-1.0579	0.5839	0.7656	-0.3221	0.5482	0.6962	0.1799	0.4094	0.6769
	Median	-0.6598	1.1123	1.1023	-0.0465	1.0657	1.0574	0.5069	1.0282	1.0344
	97.50%	-0.3343	1.7206	1.5330	0.2130	1.6181	1.4807	0.8456	1.6741	1.4374
300	**bias**	-0.0341	0.0300	0.0333	-0.0132	0.0210	0.0252	-0.0126	0.0096	0.0067
	S.E.	0.0968	0.1575	0.1044	0.0757	0.1558	0.0955	0.0915	0.1723	0.1097
	2.50%	-0.7815	0.7278	0.8445	-0.1636	0.7216	0.8493	0.3598	0.6741	0.7967
	Median	-0.5834	1.0300	1.0333	-0.0132	1.0210	1.0252	0.5367	1.0096	1.0067
	97.50%	-0.4020	1.3457	1.2523	0.1333	1.3315	1.2232	0.7189	1.3474	1.2276

Table 2. (continued) Performances of regression coefficient estimates in the dispersion linear predictor for Model I.

n	Performances	λ = 5.0								
		δ = -0.5			δ = 0.0			δ = 0.5		
		$\hat{\gamma}_0$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_0$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_0$	$\hat{\gamma}_1$	$\hat{\gamma}_2$
50	**bias**	-0.8694	0.8905	0.7559	-0.1251	0.3659	0.3118	-0.0041	0.4058	0.3986
	S.E.	0.6769	0.8205	0.6371	0.2443	0.5288	0.3938	0.1869	0.4746	0.4231
	2.50%	-3.0381	0.6442	0.8279	-0.6726	0.5244	0.7263	0.1973	0.6521	0.7576
	Median	-1.4187	1.8905	1.7559	-0.1251	1.3659	1.3118	0.5452	1.4058	1.3986
	97.50%	-0.4204	3.8316	3.2969	0.2851	2.5952	2.2574	0.9343	2.5062	2.4015
100	**bias**	-0.3076	0.2965	0.3138	-0.1205	0.2367	0.2431	0.0075	0.1282	0.1492
	S.E.	0.3274	0.3791	0.3448	0.1657	0.3559	0.2763	0.1136	0.2812	0.2319
	2.50%	-1.6245	0.6719	0.7789	-0.4804	0.6355	0.8060	0.3513	0.6425	0.7650
	Median	-0.8569	1.2965	1.3138	-0.1205	1.2367	1.2431	0.5568	1.1282	1.1492
	97.50%	-0.3432	2.1610	2.1284	0.1684	2.0301	1.8813	0.7964	1.7412	1.6701
300	**bias**	-0.1005	0.1259	0.1104	-0.0222	0.0506	0.0582	0.0092	0.0277	0.0321
	S.E.	0.1462	0.2114	0.1701	0.0632	0.1484	0.1304	0.0582	0.1398	0.1161
	2.50%	-0.9679	0.7517	0.8186	-0.1479	0.7772	0.8243	0.4501	0.7693	0.8223
	Median	-0.6498	1.1259	1.1104	-0.0222	1.0506	1.0582	0.5585	1.0277	1.0321
	97.50%	-0.3945	1.5821	1.4853	0.0992	1.3589	1.3332	0.6783	1.3165	1.2759

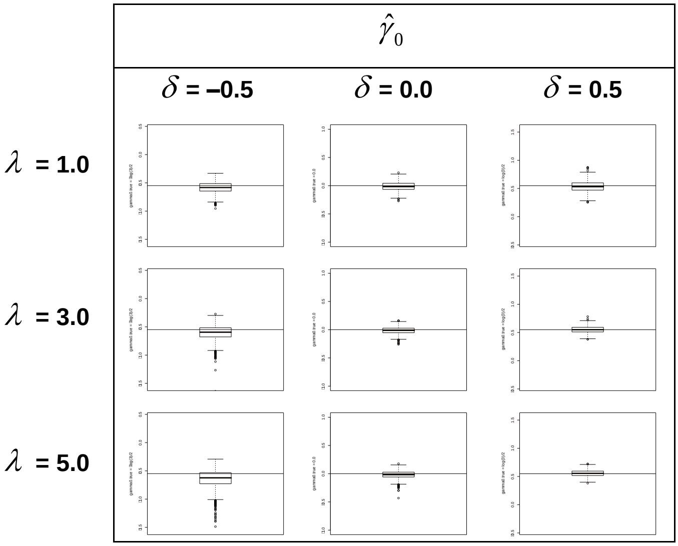

Figure 2. Performance of regression coefficient estimates ($\hat{\gamma}_0, \hat{\gamma}_1, \hat{\gamma}_2$) in the dispersion linear predictor for model I at n = 300.

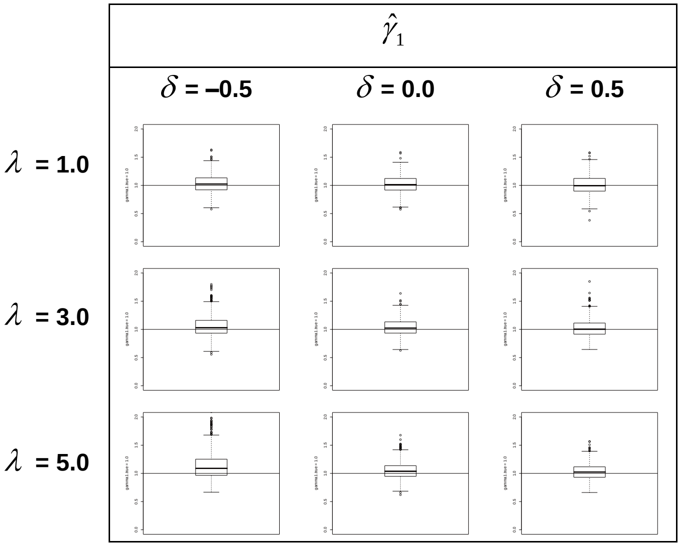

Figure 2. (continued) Performance of regression coefficient estimates ($\hat{\gamma}_0, \hat{\gamma}_1, \hat{\gamma}_2$) in the dispersion linear predictor for model I at n = 300.

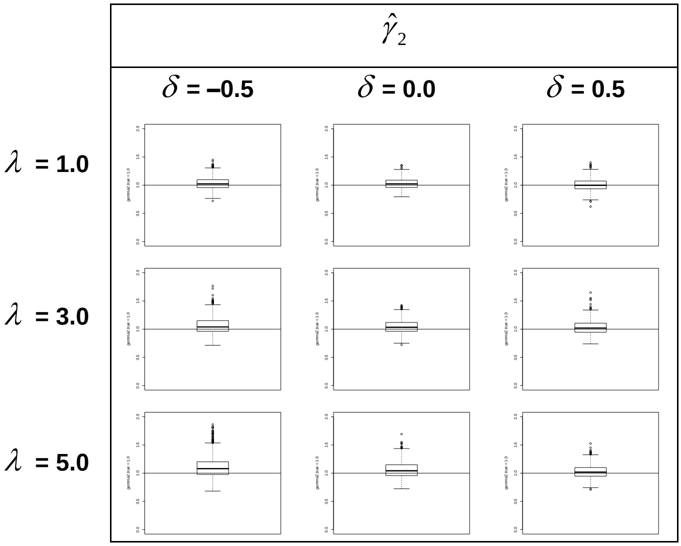

Figure 2. (continued) Performance of regression coefficient estimates ($\hat{\gamma}_0, \hat{\gamma}_1, \hat{\gamma}_2$) in the dispersion linear predictor for model I at n = 300.

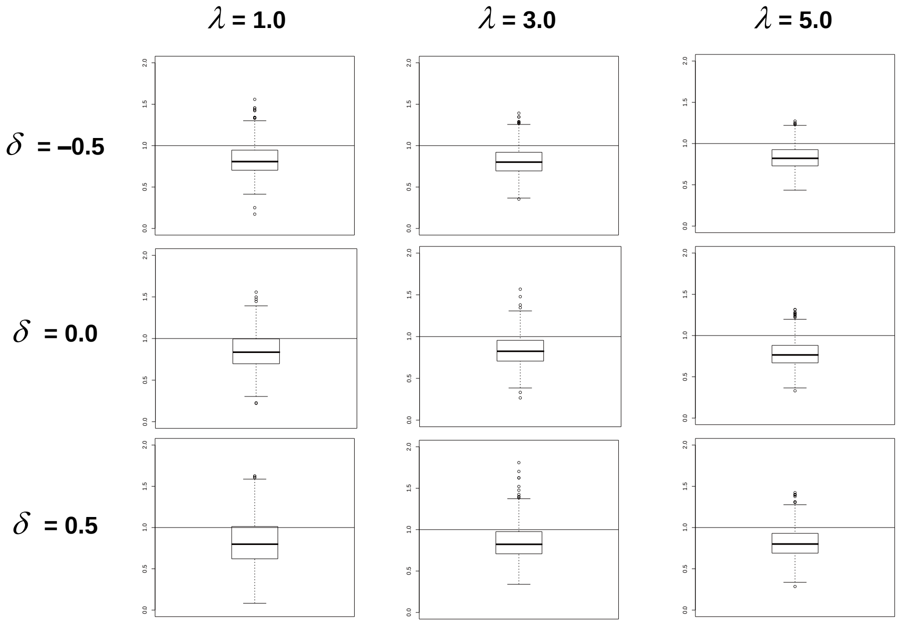

Figure 3. Performance of variance component estimates ($\hat{\sigma}^2_{\alpha_1}$) in the Poisson linear predictor for model II at n = 300.

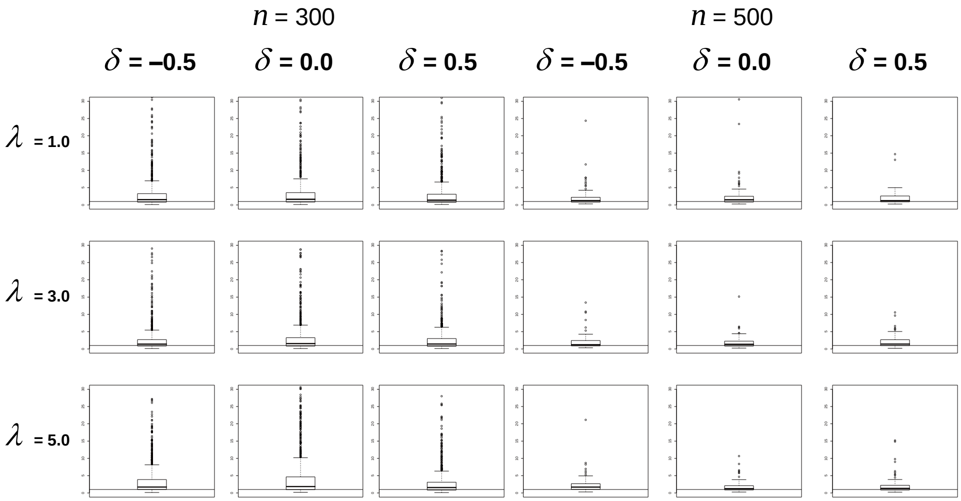

Figure 4. Performance of variance component estimates ($\hat{\sigma}^2_{\tau_1}$) in the dispersion linear predictor for model III at n = 300 and n = 500.

Finally, the Bayesian approach can be used to estimate parameters in the generalized zero-altered Poisson regression model in which the Poisson linear predictor was fixed while the dispersion linear predictor was randomized, known as “Model III”. The estimates of regression parameters in the Poisson linear predictor become more efficient as sample size increases (n ≥ 100) for λ = 1.0. The satisfactory estimates of regression parameters in the dispersion linear predictor and variance component require very large sample sizes (n > 300) for both accuracy and efficiency.

5. Modeling the 2009 Thai fertility data

In this section, the generalized zero-altered Poisson regression model was applies to study the correlation between the number of births (Y) and nine covariates from the 2009 fertility data set [18] which based on previous literature [19, 20]. The nine covariates, five continuous covariates (X_1, X_6, X_7, X_8, X_9) and four category covariates (X_2, X_3, X_4, X_5), were used in the full model. The descriptions of explanatory variables or covariates are presented in Table 3.

The data was divided into two parts: 80% (n = 4,118) was used for investigating the estimated model, the best estimated model was applied to the other 20% of data (n = 1,030). The values of bias are obtained from the difference between the predicted and observed numbers of births ($\hat{Y}_i - Y_i$). The practicability of model will be illustrated by the plot and histogram of biases.

The description of variables in the first part of data (80%) showed in Table 3. It can be observed that the slightly overdispersion is detected in this sample which the sample mean and variance are 2.2972 and 2.6263, respectively. The average women age was approximately 40 and approximately one-third (29.82%) of women was muslim. Approximately 50% of women were under compulsory education and lived in the municipal areas. The 76.52% of women are working women. The mean number of household members was 4.1297 while the minimum and maximum numbers of household members were 1 and 25, respectively. The average age at first marriage and first birth were 21.8817 and 24.2652 years, respectively. The mean number of additional children wanted was very small (0.3023).

Table 3. Description and statistics of variables (n = 4,118).

Variables	Mean	S.D.	Max.	Min.
number of births (Y)	2.2972	1.6206	0	13
age of women (X_1)	39.237	10.6661	15	59
religion (X_2) 1: Muslim, 0: others	0.2982	0.4575	0	1
place of residence (X_3) 1: municipal areas, 0:others	0.4949	0.5000	0	1
education (X_4) 1: under compulsory education, 0: others	0.4738	0.4994	0	1
occupation (X_5) 1: woman is working, 0: others	0.7652	0.4239	0	1
number of household members (X_6)	4.1297	1.9120	1	25
age at first marriage (X_7)	21.8817	4.7441	12	51
age at first birth (X_8)	24.2652	5.2331	13	44
number of additional children wanted (X_9)	0.3023	0.6665	0	6

In this study, the nine covariates were used in both linear predictors. All parameters in the generalized zero-altered Poisson regression model were estimated by the Bayesian approach using WinBUGS based on three parallel Markov chains with an initial burn-in of 10,000 iterations followed by 10,000 samples per chain, giving us a total 30,000 approximate samples from the posterior distributions of parameters. The prior distributions for regression coefficients were the normal distributions with mean 0 and variance 100 for regression coefficients (β_0 to β_9) in the linear predictor of Poisson parameter and the normal distributions with mean 0 and variance 10 for regression coefficients (γ_0 to γ_9) in the linear predictor of dispersion parameter. The posterior summary of the full and reduced models are presented in Table 4. The posterior credible intervals for β_5, β_7 in the Poisson linear predictor and γ_2, γ_3, γ_4, γ_5 in the dispersion linear predictor include zero. That is, occupation and age at first marriage are not significant in Poisson linear predictor. While religion place of residence, education, and occupation are also not significant in the dispersion linear predictor. Therefore, these covariates are excluded from the full model and the predictive models are represented as the reduced model.

Table 4. Posterior summary of parameters in the full and reduced models.

Para-meters	Full model					Reduced model				
	Mean	S.E.	2.50%	Median	97.50%	Mean	S.E.	2.50%	Median	97.50%
β_0	0.6073	0.0365	0.5360	0.6075	0.6785	0.6239	0.0252	0.5743	0.6242	0.6733
β_1	0.0340	0.0014	0.0312	0.034	0.0368	0.0340	0.0014	0.0312	0.034	0.0368
β_2	0.2957	0.027	0.2434	0.2955	0.3485	0.2941	0.0262	0.244	0.2943	0.3464
β_3	-0.0885	0.0257	-0.1395	-0.0881	-0.0385	-0.0929	0.0252	-0.142	-0.0927	-0.0437
β_4	-0.1296	0.0325	-0.1933	-0.1295	-0.0667	-0.1287	0.0311	-0.1896	-0.129	-0.0676
β_5	0.0184	0.0287	-0.0390	0.0189	0.0737	-	-	-	-	-
β_6	0.0886	0.0056	0.0775	0.0886	0.0994	0.0882	0.0055	0.0774	0.0882	0.0988
β_7	-0.0012	0.0066	-0.0137	-0.0013	0.012	-	-	-	-	-
β_8	-0.0471	0.0065	-0.0602	-0.0469	-0.0348	-0.0479	0.003	-0.0538	-0.0479	-0.042
β_9	-0.3182	0.0361	-0.3896	-0.3177	-0.2499	-0.3207	0.0364	-0.3914	-0.3209	-0.2497
γ_0	-2.5091	0.1968	-2.8880	-2.5090	-2.1220	-2.4245	0.1427	-2.7230	-2.42	-2.161
γ_1	-0.0369	0.0069	-0.0508	-0.0366	-0.0236	-0.0357	0.0062	-0.0485	-0.0355	-0.0238
γ_2	0.0555	0.1158	-0.1703	0.0553	0.2825	-	-	-	-	-
γ_3	0.0635	0.1035	-0.1404	0.0626	0.2643	-	-	-	-	-
γ_4	-0.0268	0.1162	-0.2489	-0.0255	0.2010	-	-	-	-	-
γ_5	0.0425	0.1149	-0.1888	0.0427	0.2597	-	-	-	-	-
γ_6	-0.223	0.0352	-0.2939	-0.2221	-0.1562	-0.2252	0.0358	-0.2957	-0.2247	-0.1557
γ_7	-0.0758	0.0106	-0.0967	-0.076	-0.0549	-0.0752	0.0092	-0.0940	-0.075	-0.0576
γ_8	0.2808	0.0163	0.2501	0.2805	0.3135	0.2786	0.0153	0.2490	0.2784	0.3093
γ_9	0.1510	0.0621	0.0313	0.1504	0.2718	0.1528	0.059	0.0413	0.1528	0.2706

Summary statistics for Bayesian estimation and predicted sample distributions are shown in Table 5. The reduced model estimates approximately 32% of the sample observations correctly (bias = 0), and about 70% correctly if we allow for an error of ± one child (bias = -1, 0, 1) which are larger than the full model. The DIC of the reduced model is smaller than the full model. Thus the reduced model yields the better fit than the full model for this data set. The predicted sample distributions that mostly close to resemble observed distribution is the reduced model according to the DIC. Figure 5 represents more clearly efficiency of the reduced model in the other 20% of the 2009 fertility data. The plot and histogram indicate that most biases of sample are approximately zero.

Table 5. Observed and estimated sample distributions (n = 1,030).

Number of births	Observed proportion	Estimated values	
		Full model	Reduced model
0	0.0835	0.0699	0.0786
1	0.235	0.299	0.2845
2	0.3398	0.2515	0.2495
3	0.1874	0.166	0.1806
4	0.0689	0.1049	0.0961
5	0.0369	0.0417	0.0534
6	0.0194	0.0252	0.0252
7	0.0165	0.0204	0.0184
>7	0.0126	0.0214	0.0136
Correctly predicted proportion		**0.3146**	**0.3204**
Predicted ± 1 proportion		**0.6932**	**0.7019**
DIC		**11,230.1**	**11,219.2**

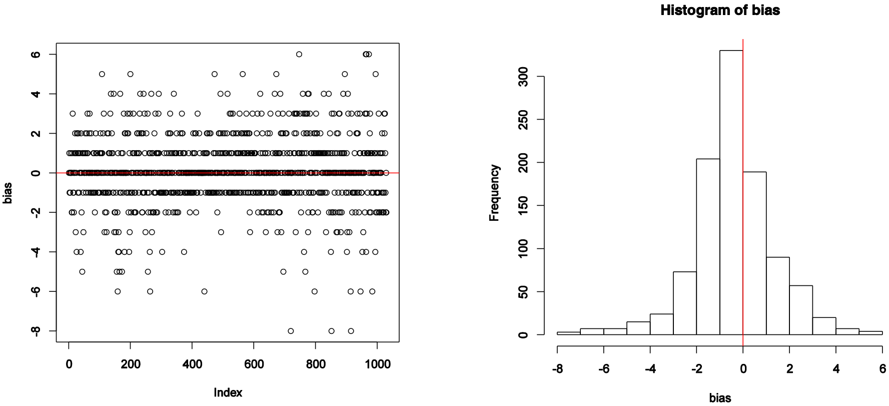

Figure 5. The performance of bias from the reduced model.

6. Conclusions

This study introduces the generalized zero-altered Poisson regression model with the suitable link functions of parameters and the Bayesian approach can be carried out for parameter estimation method using WinBUGS. Simulation studies in the zero-altered Poisson regression models illustrate that the Bayes estimator has a good performance with small bias and standard error. It can be concluded that the Bayesian approach is satisfactory estimation method for fixed effects model (Model I) at large sample sizes and even larger sample sizes for the mixed models (Model II and Model III). The results from Bayesian analysis of the generalized zero-altered Poisson regression model showed that number of births was significantly affected by age of women, number of household members, age at first birth, and number of additional children wanted in both linear predictors. In addition, the religion, place of residence, and education were significantly effects in the Poisson linear predictor. The age at first marriage was significantly effect in the dispersion linear predictor.

In the future research, other methods for parameter estimation could be used in the generalized zero-altered Poisson regression models and compared to the Bayesian estimation method. Moreover, random terms could be included in both linear predictors in the generalized zero-altered Poisson regression model and studied in the same simulation cases as Model I-III. The efficiency of the generalized zero-altered Poisson regression model should be compared to the other count models, i.e. the generalized Poisson regression model, the zero-inflated Poisson regression model for under-, equi- and over-dispersion data sets. Moreover, the generalized zero-altered Poisson regression model could be applied to real data sets in several fields, i.e. epidemiology, ecology, and economics and compared to the other count models in case of real data sets.

Acknowledgments

This research was supported by grants from Thammasat University, Thailand. We are grateful to the referees and the editor for valuable comments and constructive suggestions that improved this article.

References

[1] Hall, D.B., Zero-Inflated Poisson and binomial regression with random effects: a case study, *Biometrics*, 2000; 56: 1030-1039.

[2] Hinde, J., and Demetrio, C.G.B., Overdispersion: Models and estimation, *Computational Statistics and Data Analysis*, 1998; 27(2): 151-170.

[3] Gupta, P. L., Gupta, R.C., and Tripathi, R.C., Analysis of zero-adjusted count data, *Computational Statistics and Data Analysis*, 1996; 23(2): 207-218.

[4] Lambert, D., Zero-inflated Poisson regression with an application to defects in manufacturing, *Technometrics*, 1992; 34(1): 1-14.

[5] Ridout, M., Demetrio, C., and Hinde, J., Models for count data with many zeros. In: *Proceedings of the 19th International Biometrics Conference*, Cape Town, 1998; 179-190.

[6] Winkelmann, R., and Zimmermann, K.F., Recent developments in count data modeling: Theory and applications. *Journal of Economic Surveys*, 1995; 9: 1–24.

[7] Castillo, J., and Pérez-Casany, M., Overdispersed and underdispersed Poisson generalizations, *Journal of Statistical Planning and Inference*, 2005; 134: 486–500.

[8] Ghosh, S. K., and Kim, H., Semiparametric inference based on a class of zero-altered distributions, *Statistical Methodology*, 2007; 4(3): 371–383.

[9] Ghosh, S. K., Mukhopadhyay, P., and Lu, JC., Bayesian analysis of zero-inflated regression models. *Statistical Planning and Inference*, 2006; 136: 1360–1375.

[10] Angers, J.–F., and Biswas, A., A Bayesian analysis of zero–inflated generalized Poisson model, *Computational Statistics & Data Analysis*, 2003; 42: 37–46.

[11] Melkerson, M., and Rooth, O-D., Modeling female fertility using inflated count data models, *Population Economics*, 2000; 13(2): 189-203.

[12] Wang, W., and Famoye, F., Modeling household fertility decisions with generalized Poisson regression, *Population Economics*, 1997; 13: 273-283.

[13] Chaimongkol, S., Huffer, W.F., and Kamata A., A Bayesian approach for fitting a random effect differential item functioning across group units, *Thailand Statistician*, 2006; 4: 27–41.

[14] Chaimongkol, S., Huffer, W.F., and Kamata A., An explanatory differential item functioning model by the WinBUGS 1.4. *Songklanakarin Journal of Science and Technology*, 2007; 29(2): 449–458.

[15] Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B., *Bayesian Data Analysis*, London: Chapman and Hall, 1995.

[16] Spiegelhalter, D., Thomas, A., and Best, N., *WinBUGS Version 1.4 User Manual*, Medical Research Council Biostatistics Unit, Cambridge, 2003.

[17] R Development Core Team. R: Language and Environment for Statistical and Computing, “http://www.R-project.org”. Retrieved December 21, 2011.

[18] Thailand, Ministry of Information and Communication Technology, National Statistical Office. The 2009 Reproductive Health Survey. Retrieved September 15, 2010.

[19] Nipattra Wanotayaroj, Factors affecting fertility in the South of Thailand, Unpublished master’s thesis, Mahidol University, Faculty of Graduate Studies, Population, Reproductive Health and HIV/AIDS Research, 1997.

[20] Santos Silva, J. M. C., and Covas, F., A modified hurdle model for completed fertility. *Journal of Population Economics*, 2000; 13: 173–188.