Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Statistical Planning and
Inference 136 (2006) 1360–1375

journal of
statistical planning
and inference

www.elsevier.com/locate/jspi

Bayesian analysis of zero-inflated regression models

Sujit K. Ghosh*, Pabak Mukhopadhyay, Jye-Chyi(JC) Lu

Department of Statistics, North Carolina State University, Box 8203, Raleigh, NC 27695-8203, USA

Received 15 July 2003; accepted 13 October 2004
Available online 13 November 2004

Abstract

In modeling defect counts collected from an established manufacturing processes, there are usually a relatively large number of zeros (non-defects). The commonly used models such as Poisson or Geometric distributions can underestimate the zero-defect probability and hence make it difficult to identify significant covariate effects to improve production quality. This article introduces a flexible class of zero inflated models which includes other familiar models such as the Zero Inflated Poisson (ZIP) models, as special cases. A Bayesian estimation method is developed as an alternative to traditionally used maximum likelihood based methods to analyze such data. Simulation studies show that the proposed method has better finite sample performance than the classical method with tighter interval estimates and better coverage probabilities. A real-life data set is analyzed to illustrate the practicability of the proposed method easily implemented using `WinBUGS`.

MSC: 62F15; 62E10; 62P30

Keywords: Bayesian inference; Data augmentation; Gibbs sampling; Markov chain Monte Carlo; WinBUGS; Zero-inflated power series models

1. Introduction

Statistical methods for analyzing count data with numerous zeros are very important in various scientific fields including but not limited to industrial applications (e.g., Lambert,

* Corresponding author. Tel.: +1 919 515 1950; fax: +1 919 515 7591.
E-mail address: sghosh@stat.ncsu.edu (S.K. Ghosh).

0378-3758/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2004.10.008

Table 1
Summary of the NORTEL data

Defect counts				
Y:	0's $= 42$	1's $= 8$	2's $= 2$	3's $= 2$
Levels of the covariates				
X_1:	75	150	225	
X_2:	300	550	800	
X_3:	30	65	100	
X_4:	30	45	60	

1992) and biomedical applications (e.g., Heilbron and Gibson, 1990; Hall, 2000). As an illustration, consider the data set in Table 1, which presents the number of defects that resulted from an experiment for improving printed circuit board (PCB) manufacturing quality at Nortel, RTP, North Carolina. Out of 54 observations, 42 (78%) of them are zeros (no defects), while 8, 2 and 2 of them have one, two and three defects, respectively. In addition to such defect counts, we also obtained data on other controllable factors (covariates) that might explain the variation in the defect counts. Regression models with commonly used discrete distributions such as Poisson and Negative Binomial (see Miaou, 1994), may not fit these data well, and seriously underestimate the zero-defect probability, which is an important indicator of production quality. We use the Nortel data set as a motivating example to develop our models for count data with many zeros. However, by no means, the proposed methodology is limited to this specific data set.

The zero-inflated Poisson regression model proposed in Lambert (1992) is very useful to model discrete data with many zeros. We extend the models to include a broad class of distributions (e.g. power series distributions) and present an alternative approach to fit such models to "zero-inflated data" with a relatively small to moderate sample size. The zero-inflated model has the interpretation that when the production process is in near perfect state, zero defect can be observed with high probability. However, due to changes in manufacturing environment, the process moves to an imperfect state and defective outcomes are possible, but not inevitable. The environmental changes are usually unobservable and random. This causes the process to move randomly back and forth between the perfect and the imperfect states. If the production process is reasonably good, the data that counts the number of defects will consist of many zeros. Most data sets collected from Nortel manufacturing processes have this feature of many zeros. See Li et al. (1999) for an example. This poses a challenge in their quality improvement practice. For other interesting applications related to zero-inflated models see Dahiya and Gross (1973), Umbach (1981), Yip (1988), Gupta et al. (1996), Welsh et al. (1996), Gurmu (1997), Welsh et al. (1996), and Hinde and Demetrio (1998). An overview of zero inflated models can be found in Ridout et al. (1998).

Classical statistical methods based on the maximum likelihood estimate (MLE) and the likelihood ratio (LR) test for zero inflated Poisson regression can be found in Hall (2000). The approximation theory based on large samples usually serves as the basis for deriving classical inference for non-normal data and often requires the use of nonstandard asymptotic theory (see Self and Liang, 1987). It will be of interest to see how these procedures perform with finite samples, especially in estimating the zero-defect probability. In simulation studies based on a sample of size $n = 50$ (see Section 4.1.1), it was found that the classical procedure performs reasonably well in cases where the zero-defect probability ($\Pr(Y = 0)$) was not chosen close to unity. However, when $\Pr(Y = 0)$ was chosen closer to unity, the Bayesian estimates performed very well with respect to interval width and coverage probability. Motivated by such good finite sample performance of the Bayesian methods, this article develops Bayesian point and interval estimation methods for zero inflated regression models.

In a Bayesian approach, parameters are considered random and a joint probability model for both data and parameters is required. The joint posterior distribution of the parameters of the proposed models turns out to be analytically intractable, hence simulation-based methods (see Tierney, 1994) broadly known as Markov Chain Monte Carlo (MCMC) are required to obtain the point and interval estimates of the parameters. A simple code written in `WinBUGS` (Spiegelhalter et al., 1999), has been used to perform all the required computations (see Appendix). In Section 4.1, several models have been fitted to show that the computing time should not be a major obstacle to implementing the proposed Bayesian methods in real-life operations.

Section 2 presents parametric formulations of zero-inflated models with a particular emphasis on the zero inflated power series (ZIPS) model. Section 3 presents a Bayesian analysis for the ZIPS regression models. The process states (perfect and imperfect) are viewed as missing data and the data augmentation method (Tanner and Wong, 1987) is integrated into the MCMC procedure to generate samples from the posterior distribution of parameters of interest. Section 4 illustrates the procedure with real-life data. Results from some simulations are also presented to compare interval estimates from Bayesian and large sample chi-square approximation methods. Section 5 concludes this study and addresses a few areas of future work.

2. Zero inflated power series (ZIPS) models

The random variable, Y in a zero-inflated model can be represented as $Y = V(1-B)$, where B is a Bernoulli(p) random variable and V independently to B has a discrete distribution such as Poisson(θ), NegBin(θ, r) or more generally power series, PS(θ). Notice that under this representation, the mean ($E(Y)$) and variance ($\mathrm{Var}(Y)$) are given by,

$$E(Y) = (1-p)E(V),$$
$$\mathrm{Var}(Y) = \frac{p}{1-p}[E(Y)]^2 + \delta E(Y),$$

where $\delta = \mathrm{Var}(V)/E(V)$ denotes the coefficient of dispersion of the latent random variable V. Thus, it follows that if the latent variable, V does not have an underdispersed distribution (i.e., $\delta \geqslant 1$), then the distribution of Y is overdispersed. On the other hand, if V has a underdispersed distribution (i.e., $\delta < 1$), then Y has a underdispersed distribution if and only if $E(V) < \frac{1-\delta}{p}$. In general, any discrete distribution could be used for V, however for some theoretical insights and parsimony we restrict our attention to a simple but broad class of discrete distribution for V, namely, the power series (PS) distribution. The probability mass function of the ZIPS distribution; henceforth represented as ZIPS(p, θ), is given by,

$$P(Y=0)=p+(1-p)\frac{b(0)}{c(\theta)},$$
$$P(Y=k)=(1-p)\frac{b(k)\theta^k}{c(\theta)}, \quad k=1,2,\ldots, \tag{1}$$

where $c(\theta)=\sum_{k=0}^{\infty} b(k)\theta^k$, $0\leqslant p<1$ and $\theta>0$. When $p=0$, the ZIPS distribution reduces to a regular PS(θ) distribution. Note that $E(Y)=(1-p)\mu(\theta)$, where $\mu(\theta)=\theta\frac{d\ \log c(\theta)}{d\theta}$ denotes the mean of a regular PS(θ) distribution. It follows that, even if the mean of the PS distribution, $\mu(\theta)$, is large, zeros can occur frequently if p is close to 1. In a ZIPS regression model, the covariates are usually linked to model parameters p and θ (e.g., Lambert, 1992 considers ZIP). Denote $\boldsymbol{p}=(p_1,\ldots,p_n)$ and $\boldsymbol{\mu}=(\mu(\theta_1),\ldots,\mu(\theta_n))$. For the independently distributed responses Y_i's sampled from ZIPS(p_i, θ_i), the commonly used link functions are given by,

$$\log(\boldsymbol{\mu})=\boldsymbol{Z\beta} \quad \text{and} \quad \text{logit}(\boldsymbol{p})=\log[\boldsymbol{p}/(1-\boldsymbol{p})]=\boldsymbol{W\gamma}, \tag{2}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are vectors of regression parameters associated with predictors (covariates), $\boldsymbol{Z}=(\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n)^{\mathrm{T}}$ and $\boldsymbol{W}=(\boldsymbol{w}_1,\ldots,\boldsymbol{w}_n)^{\mathrm{T}}$, respectively. In some applications, the design matrices, $\boldsymbol{Z}$ and $\boldsymbol{W}$ are chosen to be the same.

Given a random sample $\boldsymbol{Y}=(Y_1,Y_2,\ldots,Y_n)$ from ZIPS(p, θ) the likelihood function of (p,θ) is given by

$$L(p,\theta|\boldsymbol{Y})\propto[pc(\theta)+(1-p)b(0)]^{S_0}(1-p)^{n-S_0}\frac{\theta^S}{c^n(\theta)}, \tag{3}$$

where $S_0=S_0(\boldsymbol{Y})=\#\{i\ :\ Y_i=0\}$ and $S=S(\boldsymbol{Y})=\sum_{i=1}^n Y_i$ are sufficient statistics for (p,θ). The likelihood equations can be solved using the Newton–Raphson algorithm or the EM algorithm (Dempster et al., 1977). However, simulation-based methods are required to obtain the joint posterior distribution of (p,θ). Using Binomial expansion, Eq. (3) can be written as

$$L(p,\theta|\boldsymbol{Y})\propto\left[\sum_{j=0}^{S_0} w_j p^j(1-p)^{n-j}\right]\theta^S, \tag{4}$$

where $w_j=w_j(\boldsymbol{Y},\theta)=\binom{S_0}{j}\left(\frac{b(0)}{c(\theta)}\right)^{n-j}$. For a fixed value of θ, $L(p,\theta|\boldsymbol{Y})$ (viewed as a function of p only), is a "mixture of Beta's," and similarly for a fixed value of p, $L(p,\theta|\boldsymbol{Y})$ (viewed as a function of θ only) is a mixture of a density from exponential family. Thus, in the above case, when no covariates are present Gibbs sampling-type algorithms can be easily implemented. However, when covariates are present and we use a model as in (2), a little more sophisticated algorithm such as data augmentation is required. The implementation of such sampling has now become relatively easier with the advent of software like `WinBUGS` (see Appendix for a sample code). In Section 3, we describe a data augmentation method based on the $Y=V(1-B)$ representation of the zero inflated distribution.

3. Bayesian analysis

Bayesian analysis requires the specification of prior distribution for the parameters. Assume that the prior distributions for p and θ are independent, and use the following conditional conjugate priors:

$$p \sim \text{Beta}(b_1, b_2) \quad \text{and} \quad \theta \sim \pi(\theta),$$

where $\pi(\theta) \propto \theta^{a_1}/[c(\theta)]^{a_2}$ is a conjugate prior for the PS family. Note that prior independence does not necessarily imply posterior independence. The hyperparameters a_1, a_2, b_1, b_2 are assumed known. In particular, $b_1 = b_2 = 1$ gives the uniform prior on (0, 1) for p. Small values of a_2 result in a noninformative (high variance) prior for θ. In Section 4, we use $a_1 = a_2 = 0.001$ and $b_1 = b_2 = 0.5$ for our model fitting. With the likelihood as given in Eq. (3), the joint posterior distribution of (p, θ) has a nonstandard density. To overcome such analytical limitations, Monte Carlo simulation-based techniques are used to sample from the posterior distribution. In particular, the Gibbs sampling method (Gelfand and Smith, 1990) has been used to obtain a large number of random variates from the posterior distribution. Any distributional summary (such as mean, median or quantiles) of the posterior distribution can then be approximated by their corresponding sample analogue.

3.1. The data augmentation method

The Gibbs sampling method based on full conditional distributions obtained from (4) is very efficient and attractive. However, it is not straightforward to extend it to the regression problem. To overcome this difficulty, a general procedure is proposed with the idea of augmenting to the data Y, the latent variables (V, B), using the representation $Y = V(1 - B)$ defined in Section 2. Instead of obtaining samples directly from the posterior of (p, θ) given the data, samples from the posterior of (p, θ, V, B), are obtained given the data Y. To implement this procedure within a Gibbs sampling algorithm, the conditional distribution of (V, B) given (Y, p, θ) is required. Sampling from this conditional distribution is sometimes referred to as the "data augmentation" step (Tanner and Wong, 1987) or the "imputation step." It is easy to see that,

$$P(V = v, B = 0|Y = y) = \begin{cases} \dfrac{(1-p)P(V=0)}{p + (1-p)P(V=0)} & \text{if } v = y = 0; \\ 1 & \text{if } v = y > 0; \\ 0 & \text{otherwise} \end{cases}$$

and

$$P(V = v, B = 1|Y = y) = \begin{cases} \dfrac{pP(V=v)}{p + (1-p)P(V=0)} & \text{if } y = 0; \\ 0 & \text{otherwise.} \end{cases}$$

That is, for $i = 1, \ldots, n$, if $Y_i = 0$ observed, a coin with probability of HEAD being p is flipped. If a HEAD lands, set $B_i = 1$ and draw V_i from the distribution of V with parameter θ. If a TAIL lands, set $B_i = 0$ and $V_i = 0$. If $Y_i > 0$ is observed, set $B_i = 0$ and $V_i = Y_i$. Then, sample

p from Beta$(\sum_{i=1}^{n} B_i + b_1, \sum_{i=1}^{n}(1 - B_i) + b_2)$ using the completed data consisting of V_i and B_i, obtained in the data augmentation step. Similarly, sample θ from a distribution from the exponential family. For instance, when $V \sim$ Poisson(θ), the conditional distribution of θ given the completed data is given by, $\theta \sim$ Gamma$(a_1 + \sum_{i=1}^{n} V_i, a_2 + n)$. With these sampled values of p and θ, return to the "data augmentation step" and iterate the whole process until the convergence of the sampling distribution to the posterior distribution of (p, θ, V, B). For most inferential procedures, only the sequence, $\{(p^{(l)}, \theta^{(l)}),\ l = 1, \ldots, N\}$, for some large N is saved. Note that N is not the sample size and can be chosen as large as possible (based on the convergence criteria and computing facility). This sample can be used to make inference about (p, θ) and any function of (p, θ), such as $\Pr(Y=0) = p + (1-p)\frac{b(0)}{c(\theta)}$, for the ZIPS model. The above sampling procedure can be easily implemented using `WinBUGS`.

3.2. Fitting ZIPS regression

In this article, the covariates that affect $\boldsymbol{p}$ and $\boldsymbol{\mu}$ are fixed. It is further assumed that the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ in the regression model (2) are a priori independent, and normal distributions have been used as priors for these parameters. That is,

$$\boldsymbol{\beta} \sim N_q(\boldsymbol{\beta}_0, \sigma_\beta^2 I_q) \quad \text{and} \quad \boldsymbol{\gamma} \sim N_r(\boldsymbol{\gamma}_0, \sigma_\gamma^2 I_r) \quad \text{are independent,}$$

where the constants $\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, \sigma_\beta^2$ and σ_γ^2 are assumed known. In particular, $\boldsymbol{\beta}_0 = 0, \boldsymbol{\gamma}_0 = 0$ and $\sigma_\beta^2 = \sigma_\gamma^2 = 10^3$ have been used to express prior ignorance. It may be noted that, the components of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are not independent a posteriori. Moreover, when informative prior distribution becomes available, one may replace the prior variance–covariance matrix of $\boldsymbol{\beta}$ and/or $\boldsymbol{\gamma}$ by some suitably structured (e.g., block diagonal) matrix instead of identity matrix.

A reasonable choice for the starting values of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ for the Monte Carlo simulation chain can be obtained by fitting standard logistic and Poisson regression models using any modern statistical software package such as `SAS` or `R`. In summary, the following algorithm can be used to generate samples from the conditional distributions.

- Start with a set of dispersed initial values of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$.
- *Data augmentation step*: Sample (V_i, B_i) using the data augmentation step based on the current value of p_i, μ_i(or equivalently θ_i) and the data $\boldsymbol{Y}$.
- Sample $\boldsymbol{\beta}$ using ARS (Adaptive Rejection Sampling, Gilks and Wild (1992)), given the sampled values of (V_i, B_i) and $\boldsymbol{\gamma}$.
- Sample $\boldsymbol{\gamma}$ using ARS, given the sampled values of (V_i, B_i) and $\boldsymbol{\beta}$.
- Return to the data augmentation step with the current sampled values of parameters $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ to update (V_i, B_i) until convergence.

Although these type of algorithms (involving data augmentation) are fairly standard in todays computing, we provide some more details of our implementation for the regression model when the latent variable has a Poisson distribution (see Appendix B). The modifications required for other distributions (in the PS family) is fairly straightforward and similar to the Poisson case. Alternatively, the code given in Appendix A can be modified to fit other distributions in the PS family.

4. Data analysis and simulation studies

In this section, we analyze the data listed in Table 1 to illustrate the procedures mentioned in Section 3 and also present some results based on a simulation study to compare the performance of the Bayes procedures to that with its frequentist counter parts. We use Poisson and Negative Binomial (NB) distributions from the PS family for all our illustrations. Notice that a NB(r, θ) distribution belongs to the PS family (as presented in (1)) with $b(k) = \binom{r+k-1}{k}$ and $c(\theta) = (1-\theta)^{-r}$ with $\mu(\theta) = r\theta/(1-\theta)$. Although when r is treated as an unknown parameter, the NB distribution no longer belongs to the PS family, we decided to use a flat prior (discrete uniform on $\{1, 2, \ldots, 15\}$) for r, instead of fixing its value.

4.1. Data analysis: no covariate case

First, we analyze the Nortel data (see Table 1) without making use of the explanatory variables (X's). Regular Poisson and NB distributions are also fitted to the data to compare its performance with the ZIPS distributions.

The software `WinBUGS` was used to generate samples from the posterior distribution of the parameters. To check convergence of the MCMC method, the parameters p, θ, zero-defect probability, $\Pr(Y = 0)$, and the deviance were monitored. Note that a deviance is defined as the negative of twice the log-likelihood function, e.g., for the no covariate case, we write,

$$\text{deviance} = -2\,\log(L((p, \theta)|\mathbf{Y})).$$

Given the data, the deviance is a just function of the parameters, and hence its posterior distribution can be easily computed from the MCMC samples of the parameters.

The convergence of the Gibbs sampler was checked using a Splus code `CODA` (also available from the same website as `WinBUGS`). In particular, from the initial runs with three dispersed starting values, the dependence factor (I) of the Raftery–Lewis convergence diagnostic was found to be 1.05, 0.951, 1.05 and 1.04 for θ, p, $\Pr(Y = 0)$ and deviance, respectively. Also the 97.5% shrink factors of the Gelman–Rubin diagnostics (based on three parallel chains) were all unity for θ, p, $\Pr(Y = 0)$ and deviance. Thus, these summary diagnostics indicate no potential problem with the convergence of the sampler. For more details on these numerical diagnostics, the reader is referred to the `CODA` manual (see also Cowles and Carlin, 1996). These diagnostics should not be taken as a proof of convergence of the chains, however if there were any problems, usually the diagnostic factors point to some potential problems.

Based on the above diagnostic results, a burn-in of 5000 samples was used with three parallel chains and then every 5th sample was kept, until 1000 observations were obtained from each chain. Thus, a total of 3000 observations were generated from the joint posterior distribution of the parameters. The posterior summary of the parameters, based on the final 3000 observations, is presented in Table 2. `WinBUGS` took about 7–30 s on a Dell Pentium IV 2.0 GHz PC, to perform the job for the ZIPS models.

In Table 2, a five-number summary, consisting of the posterior mean, sd (standard deviation), 2.5 percentile, median and 97.5 percentile, has been presented. Note that the 2.5

Table 2
Posterior summary for ZIPS distribution

Poi(θ)	Mean	sd	2.5%	Median	97.5%
$\mu=\theta$	0.335	0.079	0.194	0.330	0.506
$\Pr(Y=0)$	0.717	0.056	0.602	0.719	0.823
Deviance	86.5	1.4	85.5	85.9	90.6
NB(θ, r)					
$1-\theta$	0.865	0.090	0.683	0.892	0.973
r	3.677	2.815	1.000	3.000	10.000
$\mu=r\theta/(1-\theta)$	0.337	0.084	0.191	0.331	0.524
$\Pr(Y=0)$	0.734	0.053	0.628	0.736	0.834
Deviance	83.6	2.0	80.9	83.7	88.5
ZIP(p, θ)					
θ	0.331	0.080	0.192	0.324	0.506
p	0.746	0.254	0.136	0.837	0.999
$\mu=(1-p)\theta$	0.251	0.104	0.046	0.254	0.459
$\Pr(Y=0)$	0.791	0.083	0.635	0.783	0.963
Deviance	86.5	1.5	85.5	85.9	90.9
ZINB(p, θ, r)					
$1-\theta$	0.858	.093	0.674	0.883	0.971
r	3.548	2.783	1.000	2.000	10.000
p	0.752	0.250	0.155	0.840	0.999
$\mu=r(1-p)\theta/(1-\theta)$	0.257	0.101	0.046	0.257	0.476
$\Pr(Y=0)$	0.799	0.079	0.655	0.794	0.963
Deviance	83.6	1.9	80.9	83.6	87.9

and 97.5 percentiles provide an equal-tail 95% posterior interval estimate for the parameters. Alternatively, the highest posterior density (HPD) interval can also be constructed (especially if the distribution is bimodal, which was not the case for our data).

Posterior means of the deviance for Poisson models are about 86.5, whereas those for NB models are about 83.6, indicating slight better performance of the NB models. However the posterior mean (and median) estimate of $\Pr(Y=0)$ from the ZIPS distribution is closer to the empirical percentage of zero-defects, which is 0.78 for our data. This indicates a better fit of the ZIP distribution compared to the regular Poisson or NB distribution. Thus, these Bayesian estimates indicate that the ZIPS distribution provides estimates closer to empirical estimate (of zero counts).

4.1.1. Simulation studies: compare Bayes and frequentist methods

To establish validity of the proposed Bayesian method, this section presents two simulation studies based on the ZIP model without regressors to keep the model simpler and the study more focused. In simulation study-I, true parameter values were fixed at $p=0.5$, $\theta=1$, which results into $\Pr(Y=0)=0.6839$. In simulation study-II, $p=0.9, \theta=1$ is used,

Table 3
Simulation studies: coverage probability based on 95% C.I.

	Classical	Bayesian
Study-I		
$p = 0.5$	94.4	94.5
	(0.153, 0.698)	(0.378, 0.703)
$\theta = 1.0$	95.9	96.2
	(0.528, 1.689)	(0.687, 1.722)
$\Pr(Y = 0) = 0.684$	98.2	92.8
$p + (1 - p)e^{-\theta}$	(0.241, 0.802)	(0.602, 0.815)
Study-II		
$p = 0.9$	93.0	94.3
	(0.507, 0.972)	(0.634, 0.973)
$\theta = 1.0$	91.8	94.1
	(0.207, 2.610)	(0.867, 1.961)
$\Pr(Y = 0) = 0.937$	97.7	97.1
$p + (1 - p)e^{-\theta}$	(0.366, 0.984)	(0.655, 0.985)

which results into a higher $\Pr(Y = 0) = 0.9368$. For both simulation studies, data were generated from ZIP(p, θ), with a sample of size $n = 50$ and the procedure was repeated 600 times.

The results based on the average 95% interval estimates along with coverage probabilities are presented in Table 3. The classical 95% C.I.'s are computed by inverting the likelihood ratio tests based on the large sample chi-square distribution. The results given in Table 3 indicates that the Bayesian and Frequentist estimates of the parameters p and θ are very similar (except for the θ estimate in study-II), and that the proposed Bayesian method performed better in estimating the $\Pr(Y = 0)$. For example, the reductions of the average length of the intervals are 62% and 47% for study-I and study-II, respectively. Further, the study-II indicates that the average lower bound (which is 0.3659) for the classical estimate of $\Pr(Y = 0)$ may be too low for practical usages in applications with high $\Pr(Y = 0)$. In summary, we conclude from Table 3, that the Bayesian intervals are very competitive in terms of maintaining the nominal coverage probabilities and that the average lengths of the Bayesian intervals can be significantly shorter than that obtained from the classical methods, when $\Pr(Y = 0)$ is close to one.

Bayes estimates are not unbiased (under squared error loss), but for completeness a comparison of the classical and Bayesian methods were also done in terms of average bias and standard errors (s.e.). In study-I, the average biases (truevalue-mle) and s.e. of the estimates based on ML were 0.0337 (s.e. = 0.0155), 0.0040 (s.e. = 0.0302) and −0.0022 (s.e. = 0.0614) for the parameters p, θ and $\Pr(Y = 0)$, respectively. Whereas the corresponding average biases (truevalue − posterior mean) and s.e. of the Bayes estimates were −0.0666 (s.e. = 0.0723), 0.1478 (s.e. = 0.2231) and −0.0331 (s.e. = 0.0502) for the parameters p, θ and $\Pr(Y = 0)$, respectively. In study-II, the average biases and s.e. of the MLE estimates were 0.1571 (s.e. = 0.2928), 0.0926 (s.e. = 0.6773) and 0.0007 (s.e. = 0.0355) for the parameters p, θ and $\Pr(Y = 0)$, respectively. Whereas the average biases and s.e. of the Bayes estimates were 0.0676 (s.e. = 0.1476), −0.16 (s.e. = 0.4301) and −0.0067 (s.e. = 0.0338) for the parameters p, θ and $\Pr(Y = 0)$, respectively.

In addition to the above simulation studies, some preliminary small-scale studies were also performed with other parameter values and with varying sample sizes. In general, it was observed that Bayesian methods performed better with relatively small sample size (say $n=$ 30, 40, etc). Also, when the value of zero-defect probability was chosen close to one (e.g., as in study-II), it was observed that the classical estimates had large bias and poor coverage probabilities. However, when n was chosen relatively large (e.g., 200 or 250) both methods (MLE and Bayes) performed well and the difference between the two approaches were almost indistinguishable. These findings are not surprising, since when n is large, the Bayesian estimates and the MLEs behave very similar and they have the same asymptotic normal distribution.

From above studies, it follows that the strength of the Bayesian procedure is not in bias reduction, but rather in producing tighter interval estimates with good coverage probability, based on the finite sample distribution. The slight increase of computing time might be compensated by the use of freely available software packages.

4.2. Data analysis: regression case

The regular Poisson and ZIP regression models were fitted to link the count data to four controllable variables (covariates), each at three levels (see Table 1). For notational convenience, denote the level of each variable with the parenthesis such as $X_1(1)$ for level 1 $= 75$ of variable X_1 etc. From Fig. 1, note that the number of defects was zero more often when all the controllable variables were in their level 2. In the regression model, the level 2 was treated as 0 and the contrast of level 1 (and level 3) to level 2 were estimated. If more distinct levels are observed in the design, the covariates should be treated as continuous variables. For ZIP regression, the link functions in (2) were used with $\boldsymbol{z}=\boldsymbol{w}=\boldsymbol{X}=(X_1, X_2, X_3, X_4)^{\mathrm{T}}$.

As before, `WinBUGS` was used to fit both regression models and similar convergence diagnostics were used as described in Section 4.1. The posterior summary values, presented in Tables 4 and 5 are based on the final 3000 samples. `WinBUGS` took about 88 and 265 s to perform the job for Poisson and ZIP regression models, respectively. The code given in Appendix A is quite self-explanatory, but requires the knowledge of running the software. For more details the reader is advised to read the manual or to contact the first author.

Based on 95% posterior intervals given in Table 4, it is evident that for each of the four factors, the levels 1 and 3 do not have a statistically significant effect on the average number of defects (i.e. μ_i's). The intercept term was found to be significant with posterior mean −1.359 and 95% posterior interval [−3.274, −0.057]. The negative sign indicates that the average number of defects is less when the variables are set at their level 2. This is also evident from Fig. 1, where we see a large number of zero defects at level 2 for each of the factors. Note that the deviance for this regression model (with posterior mean 86.7) is close to the earlier models (in Section 4.1).

In Table 5, the five-number posterior summary of parameters from the ZIP regression model has been presented. It is interesting to note that the deviance dropped to 56.7 with

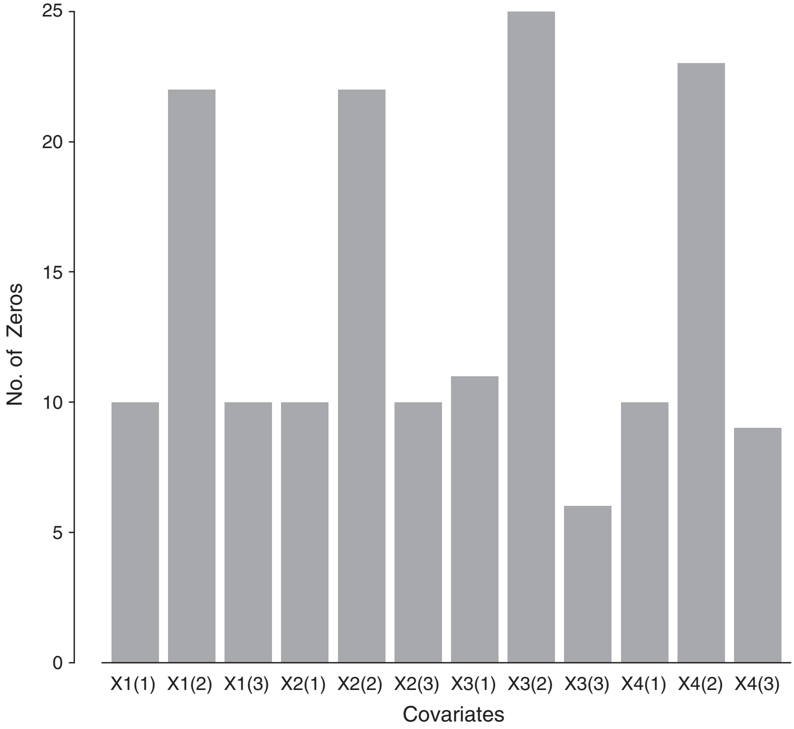

Fig. 1. Bar plot of the number of zeros at each level of the controllable factors X_1, X_2, X_3, X_4.

Table 4
Posterior summary of parameters for the Poisson regression model (see Section 4.2)

	Mean	sd	2.5%	Median	97.5%
Intercept	−1.36	0.82	−3.27	−1.25	−0.06
$X_1(1)$	0.24	0.65	−1.07	0.23	1.53
$X_1(3)$	−0.36	0.79	−2.02	−0.33	1.15
$X_2(1)$	0.18	0.65	−1.05	0.19	1.53
$X_2(3)$	−0.93	0.93	−3.05	−0.87	0.72
$X_3(1)$	−1.77	1.36	−4.85	−1.59	0.42
$X_3(3)$	0.82	0.58	−0.28	0.81	2.05
$X_4(1)$	−0.01	0.70	−1.41	0.00	1.35
$X_4(3)$	0.00	0.70	−1.37	0.00	1.36
$\overline{\Pr(Y=0)}$	0.75	0.04	0.66	0.76	0.84
Deviance	86.7	4.5	79.9	86.0	97.7

95% posterior interval [44.2, 72.2]. More interestingly, the interval does not contain the 2.5 percentile deviances from previous models (see Tables 2 and Table 5). This definitely indicates a better fit to the data. Similar to the case of regular Poisson regression model, it

Table 5
Posterior summary of parameters for the ZIP regression model (see Section 4.2)

		Mean	sd	2.5%	Median	97.5%
p	Intercept	18.39	13.57	0.11	15.92	49.97
	$X_1(1)$	−22.74	28.90	−74.41	−21.58	25.85
	$X_1(3)$	−6.60	21.35	−43.37	−8.54	41.30
	$X_2(1)$	18.70	33.42	−41.62	15.26	80.21
	$X_2(3)$	−29.29	21.55	−68.36	−30.18	17.40
	$X_3(1)$	−6.79	22.03	−46.10	−8.19	41.34
	$X_3(3)$	34.23	21.66	−3.75	32.69	78.74
	$X_4(1)$	−29.87	21.59	−70.66	−30.45	17.13
	$X_4(3)$	7.09	20.96	−34.68	6.96	50.99
λ	Intercept	−1.37	0.80	−3.21	−1.28	−0.04
	$X_1(1)$	6.17	7.41	−1.45	2.60	24.79
	$X_1(3)$	0.63	1.56	−2.13	0.49	3.93
	$X_2(1)$	−4.19	7.30	−22.670	−0.78	3.43
	$X_2(3)$	1.51	1.920	−2.01	1.58	5.13
	$X_3(1)$	−0.804	2.23	−5.31	−0.84	3.65
	$X_3(3)$	−0.72	1.56	−3.87	−0.73	2.03
	$X_4(1)$	2.56	1.73	−0.53	2.63	5.85
	$X_4(3)$	0.19	1.51	−2.95	0.23	3.06
	$\overline{\Pr(Y=0)}$	0.81	0.03	0.74	0.82	0.87
	Deviance	56.73	6.97	44.21	56.38	72.22

is again evident that for each of the four factors, the levels 1 and 3 do not have statistically significant effects on the average number of defects (μ) and the probability of being in the zero-state (p). However, the intercepts are significant. A positive intercept of 18.39 (with posterior 95% interval [0.1097, 49.97]) for regression via p, indicates that the chance of being in the zero-state is higher when the variables are set at their level 2, which is also evident from Fig. 1. Similarly, a negative intercept of −1.37 (with 95% posterior interval [−3.21, −0.04]) indicates that the average number of defects is less when the variables are set at their level 2, again a conclusion consistent with Fig. 1. Moreover, the sample mean zero-defect probability, $\overline{\Pr(Y=0)} = \frac{1}{n}\sum_{i=1}^{n} \Pr(Y_i = 0)$, is estimated to be 0.81 (with 95% posterior interval [0.74, 0.87]), which is close to the empirical percentage of the zero-counts of 0.78.

We have also fitted the NB regression models which resulted into higher deviance estimates as compared to the ZIP regression models and hence the results are not reported in this article. In addition, one of the reviewers remarked that the Nortel data appears "almost binary" as there are only four data points that have counts of two or three. We fitted a logistic regression model (treating the counts with two or three as one) which resulted into a good fit as well with deviance estimate of about 59.1 and the posterior mean of $\overline{\Pr(Y=0)}$ turned out to be 0.75 (with 95% posterior interval [0.66, 0.84]), reasonably close to the empirical estimate of 0.78.

In summary, the ZIP regression model fits much better compared to the regular regression models (such as Poisson, NB and Binary logistic) for the type of data that we have analyzed here. Thus, as a part of practical guidelines it is recommended that a practitioner should fit both the regular and the ZIP regression model to data sets that contain large number of zeros. Then choose the model that has significantly lower estimates of deviance or other model choice criteria. For the data presented in Table 1, after comparing the posterior distributions of the deviances, it is quite clear that the ZIP regression model fits better. If such a distinct difference is not observed, it is possibly parsimonious to use the regular regression models (that do not account for zero inflation).

5. Conclusions

Zero-inflated models have been shown to be useful for modeling outcomes of manufacturing processes and other situations where count data with many zeros are encountered. In the presence of covariates, zero-inflated regression model has been found to be useful for process optimization. In this article Bayesian methodologies have been used to model such data, using sampling-based methods. From simulation studies, it is also evident that the proposed methods are quite effective in drawing inferences based on small samples. From the attached WinBUGS code it can be seen that the proposed method can be implemented easily and also it can be extended to other zero-inflated distributions involving other probability distributions (e.g., NB distribution).

Application of the proposed method is not limited to data sets from a manufacturing process, but there is a broad range of situations where such data sets with many zeros are encountered. For example, there are many zeros in Heilbron's (1994) drug abuse study. Saei and McGilchrist (1997) described a data set of the chemotherapy use during 1987 for each of the 39 counties of Washington state. There were numerous zeros in that data, and the non-zero component of the data was described by a random threshold model. This is similar to our motivation of the "perfect" and the "imperfect" states of manufacturing processes.

It is natural to extend these methods to bivariate (see Lu and Bhattacharyya, 1991) and multivariate cases where there are different types of defects and we want to assimilate information from these sources to optimize the process improvement. However, formulation of the regression problem for multivariate ZIP models can be tricky! A Bayesian approach to this problem is under study by the authors at present.

Acknowledgements

We thank Spiegelhalter et al. of MRC Biostatistics unit, Institute of Public Health, Cambridge, UK for providing the program BUGS (Bayesian inference Using Gibbs Sampling) and its window version WinBUGS, available free of cost from their web site http://www.mrc-bsu.cam.ac.uk/bugs. The WinBUGS code for the ZIP regression model is given in the Appendix. Finally we thank the editor, the associate editor, and the referees for their constructive comments and valuable suggestions that improved this article. S.K. Ghosh's and J.-C. Lu's research was supported by the NSF under award DMS-9803767.

Appendix A. `WinBUGS` code for ZIP regression

```
model{
 #Likelihood:
 for(i in 1:n){
 count[i] ~ dpois(mu[i])
 mu[i] <- u[i]*lambda[i]
 u[i] ~ dbern(p[i])
 logit(p[i]) <- alpha00+alpha[1,X[i,1]]+alpha[2,X[i,2]]
                +alpha[3,X[i,3]]+alpha[4,X[i,4]]
 log(lambda[i]) <- beta00+beta[1,X[i,1]]+beta[2,X[i,2]]
                   +beta[3,X[i,3]]+beta[4,X[i,4]]
 zdp[i] <- 1-p[i]+p[i]*exp(-lambda[i])}
 mzdp <- mean(zdp[])
#Priors:
 for(j in 1:4){
 alpha[j,1] ~ dnorm(0, 0.001) alpha[j,2] <- 0
 alpha[j,3] ~ dnorm(0, 0.001)
 beta[j,1] ~ dnorm(0, 0.001) beta[j,2] <- 0
 beta[j,3] ~ dnorm(0, 0.001)}
 alpha00 ~ dnorm(0, 0.001)
 beta00 ~ dnorm(0, 0.001)
 }
```

Appendix B. Data augmentation for ZIP regression

In case of zero inflated poisson (ZIP) regression models, we have $c(\theta)=e^{\theta}$, $b(k)=1/k!$ and hence $\mu=\mu(\theta)=\theta$. The regression model in (2) can be inverted to write as,

$$\theta_i = \exp(z_i^{\mathrm{T}}\boldsymbol{\beta}) \quad \text{and} \quad p_i = [1+\exp(-\boldsymbol{w}_i^{\mathrm{T}}\gamma)]^{-1}. \tag{5}$$

Similar to methods proposed in Section 3.1, a data augmentation technique can be used to generate samples from the joint posterior of $\boldsymbol{\beta}$ and γ. The data augmentation step is carried out in a similar manner as before but with the following modifications. The probability of obtaining a HEAD becomes $1/\{1+\exp[-\sum_{i=1}^{n}(\boldsymbol{w}_i^{\mathrm{T}}\gamma]\}$ and the mean parameter for the Poisson distribution becomes $\exp(\sum_{i=1}^{n} z_i^{\mathrm{T}}\boldsymbol{\beta})$. The posterior distribution of $(\boldsymbol{\beta},\gamma)$ can be obtained by multiplying the likelihood function (based on "complete data") and the prior,

$$\begin{aligned}&\exp\left\{\left(\sum_{i=1}^{n} V_i \boldsymbol{z}_i\right)^{\mathrm{T}} \boldsymbol{\beta}\right\} \exp\left\{\left(\sum_{i=1}^{n} B_i \boldsymbol{w}_i\right)^{\mathrm{T}} \boldsymbol{\gamma}\right\} \prod_{i=1}^{n} \frac{\exp(-\exp(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\beta}))}{(1+\exp(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{\gamma}))}\\ &\quad\times (\sigma_\beta)^{-q} \exp\left\{-\left(\frac{1}{2}\sigma_\beta^{-2}\right)(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right\}\\ &\quad\times (\sigma_\gamma)^{-r} \exp\left\{-\left(\frac{1}{2}\sigma_\gamma^{-2}\right)(\boldsymbol{\gamma}-\boldsymbol{\gamma}_0)^{\mathrm{T}}(\boldsymbol{\gamma}-\boldsymbol{\gamma}_0)\right\}. \end{aligned} \tag{6}$$

In order to implement the modified version of the Gibbs sampling algorithm, the following full conditional densities are needed: $[\boldsymbol{\beta}|\text{rest}]$ and $[\boldsymbol{\gamma}|\text{rest}]$, where "rest" refers to the vector of all parameters and data excluding the former argument, e.g., $[\boldsymbol{\beta}|\text{rest}] = [\boldsymbol{\beta}|\text{V's, B's}, \boldsymbol{\gamma}]$. The full conditional distributions of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are not standard distributions. Their densities are as follows:

$$\begin{aligned}\pi[\boldsymbol{\beta}|\text{rest}] \propto (\sigma_\beta)^{-\mathrm{q}} &\exp\left\{-\left(\frac{1}{2}\,\sigma_\beta^{-2}\right)(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right\}\\ &\times \exp\left\{\left(\sum_{i=1}^{n} V_i \boldsymbol{z}_i\right)^{\mathrm{T}} \boldsymbol{\beta}\right\} \prod_{i=1}^{n} \exp(-\exp(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\beta}))\end{aligned} \tag{7}$$

and

$$\begin{aligned}\pi[\boldsymbol{\gamma}|\text{rest}] \propto (\sigma_\gamma)^{-r} &\exp\left\{-\left(\frac{1}{2}\,\sigma_\gamma^{-2}\right)(\boldsymbol{\gamma}-\boldsymbol{\gamma}_0)^{\mathrm{T}}(\boldsymbol{\gamma}-\boldsymbol{\gamma}_0)\right\}\\ &\times \exp\left\{\left(\sum_{i=1}^{n} B_i \boldsymbol{w}_i\right)^{\mathrm{T}} \boldsymbol{\gamma}\right\} \prod_{i=1}^{n} (1+\exp(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{\gamma}))^{-1}.\end{aligned} \tag{8}$$

The above densities are obtained by retaining only those terms in (6) that involve $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, respectively. It is easy to see that both of the densities in (7) and (8) are log-concave. So the adaptive rejection sampling (ARS) (see Gilks and Wild, 1992) can be used to sample $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ from their respective full conditional distributions given in (7) and (8).

References

Cowles, M.K., Carlin, B.P., 1996. Markov Chain Monte Carlo convergence diagnostics: a comparative review. J. Amer. Statist. Assoc. 91, 833–904.

Dahiya, R.C., Gross, A.J., 1973. Estimating the zero class from a truncated Poisson sample. J. Amer. Statist. Assoc. 68, 731–733.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data vis the EM algorithm (with discussion). J. Roy. Statist. Soc. B 39, 1–38.

Gilks, W.R., Wild, P., 1992. Adaptive rejection sampling for Gibbs sampling. Appl. Statist. 41, 337–348.

Gelfand, A.E., Smith, A.F.M., 1990. Sampling based approaches to calculating marginal densities. J. Amer. Statist. Assoc. 85, 398–409.

Gupta, P.L., Gupta, R.C., Tripathi, R.C., 1996. Analysis of zero-adjusted count data. Comput. Statist. Data Anal. 23, 207–218.

Gurmu, S., 1997. Semiparametric estimation of hurdle regression models with an application to Medicaid utilization. J. Appl. Econometrics 12, 225–242.

Hall, D.B., 2000. Zero-inflated Poisson and binomial regression with random effects: a case study. Biometrics 56, 1030–1039.

Heilbron, D.C., 1994. Zero-altered and other regression models for count data with added zeroes. Biometrical J. 36, 531–547.

Heilbron, D.C., Gibson, D.R., 1990. Shared needle use and health beliefs concerning AIDS: regression modeling of zero-heavy count data. Poster session. Sixth International Conference on AIDS, San Francisco, CA.

Hinde, J., Demetrio, C., 1998. Overdispersion: models and estimation. Comput. Statist. Data Anal. 27, 151–170.

Lambert, D., 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics 34, 1–14.

Li, C.S., Lu, J.C., Park, J., Kim, K.M., Brinkley, P.A., Peterson, J., 1999. A multivariate zero-inflated Poisson distribution and its inference. Technometrics 41 (1), 29–38.

Lu, J.C., Bhattacharyya, G.K., 1991. Inference procedures for a bivariate exponential model of Gumbel based on life test of system and components. J. Statist. Plann. Inference 27, 383–396.

Miaou, S.P., 1994. The relationship between truck accidents and geometric design of road sections. Poisson versus negative binomial regressions. Accident Anal. Prevention 26, 471–482.

Ridout, M., Demetrio, C.G.B., Hinde, J., 1998. Models for count data with many zeros. International Biometric Conference, Cape Town.

Saei, A., McGilchrist, C.A., 1997. Random threshold models applied to zero class data. Austral. J. Statist. 39, 5–16.

Self, S.G., Liang, K-Y., 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. J. Amer. Statist. Assoc. 82, 605–610.

Spiegelhalter, D.J., Thomas, A., Best, N.G., 1999. WinBUGS Version 1.2 User Manual, MRC Biostatistics Unit.

Tanner, M., Wong, W., 1987. The calculation of posterior distributions by data augmentation (with discussion). J. Amer. Statist. Assoc. 82, 528–550.

Tierney, L., 1994. Markov chains for exploring posterior distributions (with discussions). Ann. Statist. 22, 1701–1762.

Umbach, D., 1981. On inference for a mixture of Poisson and a degenerate distribution. Commun. Statist. Ser. A 10, 299–306.

Welsh, A., Cunningham, R., Donnelly, C., Lindenmayer, D., 1996. Modeling the abundance of rare species—statistical models for count with extra zeros. Ecol. Model. 88, 297–308.

Yip, P., 1988. Inference about the mean of a Poisson distribution in the presence of a nuisance parameter. Austral. J. Statist. 30, 299–306.