

BAYESIAN INFERENCE FOR ZERO-INFLATED POISSON REGRESSION MODELS

HUI LIU and DANIEL A. POWERS

Department of Sociology
Michigan State University
316 Berkey Hall
East Lansing
MI 48824
USA
e-mail: liuhu@msu.edu

Population Research Center
Department of Sociology
University of Texas at Austin
Austin, TX 78712
USA
e-mail: dpowers@austin.utexas.edu

Abstract

Count data with excess zeros are common in social science research and can be considered as a special case of mixture structured data. We exploit the flexibility of the Bayesian analytic approach to model the mixture data structure inherent in zero-inflated count data by using the zero-inflated Poisson (ZIP) model. We discuss the importance of modelling excess-zero count data in social sciences and review the distributional properties of zero-inflated count data, with special attention given to its mixture data structure in the context of Bayesian modelling. We illustrate the methodology using data from the Americans' changing lives (ACL) survey on cigarette smoking. Results from predictive checks suggest that the proposed Bayesian ZIP model provides a good fit to the

2010 Mathematics Subject Classification: 62F15, 62J12.

Keywords and phrases: Bayes, zero-inflated Poisson, regression analysis, count data.

Received March 27, 2012

specific case of zero-inflated count data. Simulation studies suggest that the proposed Bayesian method performs better than the maximum likelihood approach to estimate the ZIP model, with larger coverage probabilities and smaller bias, as measured by the root mean squared error (RMSE). This is especially true for small samples and in cases of very high or very low incidence of excess zero outcomes.

1. Introduction

Count data with excess zeros are common place in social science research. The Poisson regression model is a standard methodological tool for count data, and has been extended in numerous ways to accommodate data that depart from the usual assumptions of Poisson sampling with respect to the well-known mean variance equivalence. Zero-inflated count data, i.e., count data containing an overabundance of zeros, are often a source of this violation in assumption (Cameron and Trivedi [2, 3]; Dean and Lawless [7]; Greene [10]; Hausman et al. [12]; Heilbron [13]; King [15]; Lambert [16]; Mullahy [23]; and Long [18]). The zero-inflated Poisson (ZIP) model is an oftenutilized strategy to account for inflated zeros in count data, and has traditionally been estimated by using maximum likelihood estimator (MLE) (Long [18]).

The maximum likelihood estimator (MLE) is attractive in many respects as it has several desirable asymptotic properties. For example, as the sample size increases to infinity, the MLE tends to be an unbiased and efficient estimator (if it exists), with a normal distribution (Casella and Berger [4]). However, these asymptotic assumptions require large samples and do not necessarily hold in small samples. In addition, maximum likelihood methods typically use confidence intervals to quantify uncertainty in the estimator by subtracting and adding a certain amount of standard error around the point estimate. Such confidence intervals in classical statistics do not always have a straightforward interpretation and are often misinterpreted as the probability that the unknown parameter value falls within the interval, which is indeed consistent with the viewpoint of Bayesian intervals (Gelman et al. [8]).

More importantly, the classical approach of using confidence intervals to quantify uncertainty is often cast into doubt in the case of severely zero-inflated data and/or small sample sizes, when the asymptotic normality of the MLE is not satisfied (Ghosh et al. [9]). Bayesian analysis offers an alternative approach to address these deficiencies by providing the full joint distribution of the parameters of interest to fully account for various sources of uncertainty in the parameters, in addition to providing a commonsense interpretation of Bayesian intervals (Gelman et al. [8]).

This paper introduces a Bayesian alternative to account for the uncertainty in modelling zero-inflated count data. Section 2 explains the rationale for modelling this type of data in social science research. Section 3 reviews the mixture distribution of the zero-inflated count data. Section 4 formalizes the Bayesian analytic approach for the ZIP model. Section 5 illustrates the Bayesian ZIP model with an empirical example on smoking behaviour. The flexibility of evaluating ZIP models in a Bayesian framework is illustrated by using posterior predictive checking techniques. Section 6 conducts a simulation and demonstrates that the proposed Bayesian approach for the ZIP model performs better—in the sense of yielding larger coverage probabilities and smaller root mean squared errors—than the classical maximum likelihood approach, especially in the case of small sample size coupled with very high or very low incidence of inflated zeros in the data.

2. Modelling Count Data with Excess Zeros in Social Sciences

Zero -inflation is a commonly occurring characteristic in much of the count data used in social science research. This kind of data occurs frequently in *demographic* research in modelling the total number of children born to women in a developed country; in *medical sociology* when modelling the number of chronic health conditions; in *adolescent studies* of the number of times ever having sex among younger adolescents; in *health behaviour studies* modelling the frequency of

substance use; and in *criminology* when modelling the number of deviant behaviours. Response variables are measured as frequencies or counts of the occurrence of the specific behaviour over time or space, and for social behaviours such as these it is not uncommon to find many non-occurrences. In many cases, the zeros may actually dominate the distribution leading to considerable truncation.

Social researchers often opt to recode excess-zero count data to nominal or ordinal measures as a way to circumvent problems associated with truncated count data. For example, Umberson et al. [28] recoded drinking into three categories including non-drinkers, moderate drinkers, and heavy drinkers. Although in some cases more detailed information is available, it is commonplace to recode excess-zero count data into nominal or ordinal variables, possibly because modelling approaches for discrete or nominal data are more familiar than those associated with zero-inflated count data, or possibly because a comparison of particular response categories is substantively interesting. However, the potential loss of information from adopting a nominal or ordinal coding from a count measurement is a major concern when adopting this strategy, and may result in a model that is less efficient in terms of using available information in the data.

Techniques for modelling over-dispersed count data have been fully developed over the past several decades. About two decades ago the ZIP model was introduced by statistical engineers (Lambert [16]). Economists and sociologists began to utilize the ZIP model extensively beginning in the mid-1990s (Greene [10]; Long [18]; and Zorn [29]). Maximum likelihood based approaches are the most widely used estimation methods for the ZIP model in social sciences. One major criticism of such classical approaches from a Bayesian perspective is that uncertainty in parameters is not fully taken into account (Gelman et al. [8]). Moreover, in the case of data from small samples, the asymptotic normality properties of MLE may not be realized and inference based on the estimator's variance is often not tenable (Gelman et al. [8]).

Therefore, we believe it is important to offer an alternative Bayesian modelling approach to zero-inflated count data that may correct potential weaknesses associated with the maximum likelihood estimator in certain contexts. In particular, we think that Bayesian analysis allows social scientists to gain greater familiarity with, and insight into, specific aspects of their data. Additionally, the Bayesian approach takes account of various sources of uncertainty in the parameters and the full joint distribution of the parameters of interest can be generated by using this approach (Gelman et al. [8]). Furthermore, the mean values and interval estimates, or credible intervals, obtained from the estimated posterior distributions of parameters of interest are reliable regardless of the sample size in Bayesian paradigm (Gelman et al. [8]).

3. Mixture Distributions for Zero-Inflated Count Data

Classical statistics offers two strategies to model zero-inflated count data. One modelling strategy is the zero-inflated Poisson (ZIP) model (Lambert [16]; Greene [10]; Long [18]; Liu and Powers [17]). The other strategy is the Poisson Hurdle model (Heilbron [13]; King [15]; Mullahy [22]). These specifications are described in detail by Zorn [29] as specific cases of dual-regime models, in which a sequence of processes—a transition stage followed by an event count stage—determine the observed event counts. In the case of the Hurdle model, after making a transition from a zero state, the event counts are truncated at zero, implying that a nonzero count is inevitable once the Hurdle is crossed. In the case of the zero-inflated Poisson (ZIP) model, the second stage event counts are not truncated at zero, which implies that a subset of the population making the transition to the event count stage may have counts of zero. The ZIP model then assumes two different sources for zero values in the data: Some zeros are *structural zeros* in the sense that certain individuals may be “immune” to certain behaviours, and other zeros are *sampling zeros* that would occur by chance as determined by parameters underlying the distribution.

Zero-inflated count data may be viewed as a special case of a two-stage mixture distribution. It is convenient to think of this model in terms of a Poisson-Bernoulli mixture structure. As a very simple example to illustrate this mixture data structure, let us assume that Y is a zero-inflated count variable. Further, suppose that there are 10 observed cases for this variable denoted by \mathbf{y} . Each element in \mathbf{y} represents an individual observation equal to the specific value. For example, assume that \mathbf{y} represents the daily number of cigarettes smoked reported by the respondents in a health survey. Among the 10 respondents, 7 reported smoking no cigarettes

$$\mathbf{y} = (0 \ 5 \ 0 \ 0 \ 10 \ 0 \ 0 \ 0 \ 3 \ 0).$$

The mean of \mathbf{y} is 1.8. Therefore, if the data truly follow a Poisson distribution with a mean of 1.8, only about $e^{-1.8} \times 100\% = 16.5$ percent of counts would be expected to be zero. In other words, less than 2 of these 10 observations are expected to be zero under a Poisson distribution, which is clearly inconsistent with the data.

We can map the observed counts into two outcomes: (1) 1 for the unexpected/inflated zeros; and (2) 0 for all other observations consistent with an underlying Poisson distribution. We use Z to represent this inflated-zero index variable

$$Z = \begin{cases} 0, & \text{not inflated values,} \\ 1, & \text{inflated zeros.} \end{cases}$$

Then, the mixture structure of the zero-inflated count data Y is reflected in the following distributional assumptions:

$$\begin{aligned} Y|(Z = 0) &\sim \text{Poisson}(\lambda), \\ Y|(Z = 1) &\equiv 0, \\ Z &\sim \text{Bernoulli}(p). \end{aligned} \tag{1}$$

Therefore, the ZIP distribution consists of two stages: a Bernoulli zero-inflation stage with parameter p and a Poisson count stage with parameter λ . We denote the ZIP distribution as $Y \sim \text{ZIP}(\lambda, p)$, where λ is the Poisson mean parameter conditional on the observed value not being an inflated zero (i.e., conditional on $Z = 0$) and p is the distribution parameter of Z indicating the probability of being an inflated zero. The zero values in the ZIP distribution can be viewed as comprising two parts. One portion of the zeros arises from the Bernoulli distribution with parameter p indicating the probability of inflated zeros, whereas the other portion comes from what would be expected given a Poisson distribution with parameter λ . The probability mass function of $Y \sim \text{ZIP}(\lambda, p)$ is

$$\Pr(Y = y|\lambda, p) = \begin{cases} p + (1 - p)e^{-\lambda}, & \text{if } y = 0, \\ (1 - p)\frac{e^{-\lambda} \cdot \lambda^y}{y!}, & \text{if } y > 0. \end{cases} \quad (2)$$

It is easy to show that

$$\begin{aligned} E(Y) &= (1 - p)\lambda, \\ \text{Var}(Y) &= \lambda(1 - p)(1 + p\lambda). \end{aligned} \quad (3)$$

4. Bayesian Inference for Zero-Inflated Poisson Regression Models

In this section, we discuss how the Bayesian approach may be applied to a zero-inflated Poisson regression model taking account of the uncertainty in the parameters. We start with an unconditional ZIP model without covariates followed by the ZIP regression model including covariates. The Bayesian approach considers prior information about the distribution of parameters along with the likelihood of the observed data to construct a posterior distribution of relevant quantities for inference about unknown parameters as well as other features of interest including data, combinations of parameters, and combinations of parameters and data.

4.1. Unconditional ZIP model without covariates

Given an independent and identically distributed random sample $Y = (y_1, y_2, \dots, y_n) \sim \text{ZIP}(\lambda, p)$, the likelihood function is

$$f(Y|p, \lambda) = [p + (1-p)e^{-\lambda}]^k [(1-p)e^{-\lambda}]^{n-k} \prod_{j=k+1}^n \frac{\lambda^{y_j}}{y_j!}, \quad (4)$$

where n represents the total number of observations for Y among which a subset of the first k respondents have zero values. The y_{k+1}, \dots, y_n indicate the observations with nonzero values.

A major challenge of Bayesian statistics is the choice of an appropriate prior distribution for a parameter. Conjugate priors are often chosen for convenience to ensure that the resulting posterior distribution is a closed form in the same distributional family (Casella and Berger [4]). However, for many problems in Bayesian analysis, especially those involving multiple parameters such as the ZIP model, it is often very difficult to derive the joint posterior by using a standard density. Simulation methods offer a feasible strategy in this case. In general, which particular prior is chosen is based on past experience, influenced by personal preference, or chosen by convenience. For example, an often-used choice for the prior distribution of a parameter, which lies between 0 and 1 such as p in the ZIP model is a beta distribution with parameters a and b .

$$f(p|a, b) = \frac{1}{B(a, b)} p^{(a-1)} (1-p)^{(b-1)}, \quad a > 0, \quad b > 0. \quad (5)$$

In particular, $a = b = 1$ gives the uniform prior on $(0, 1)$.

The Poisson mean parameter, λ is often assumed to be drawn from a gamma distribution with shape and scale parameters c and d .

$$f(\lambda|c, d) = \frac{1}{\Gamma(c)d^c} \lambda^{(c-1)} e^{(-\lambda/d)}, \quad c > 0, \quad d > 0. \quad (6)$$

The parameters p and λ can be further assumed to be prior independent, yielding the joint prior distribution

$$f(p, \lambda|a, b, c, d) = \frac{1}{B(a, b)\Gamma(c)d^c} p^{(a-1)}(1-p)^{(b-1)}\lambda^{(c-1)}e^{(-\lambda/d)},$$

$$a, b, c, d > 0. \tag{7}$$

Applying Bayes rule to the unconditional ZIP model based on the likelihood function in Equation (4) and the prior distribution function in Equation (7) yields the following joint posterior distribution, which has a nonstandard density:

$$f(p, \lambda|Y) \propto f(Y|p, \lambda)f(p, \lambda)$$

$$\propto [p + (1-p)e^{-\lambda}]^k [(1-p)e^{-\lambda}]^{n-k} p^{(a-1)}$$

$$\times (1-p)^{(b-1)}\lambda^{(c-1)}e^{(-\lambda/d)} \prod_{j=k+1}^n \frac{\lambda^{y_j}}{y_j!}. \tag{8}$$

4.2. ZIP regression models with covariates

Regression-type models are widely used in applied research to adjust for covariate effects and assess relationships between key predictors and the response. 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 λ . Thus, this allows simultaneous inferences to be made about the zero-inflation and count process. In the usual specification, covariates are related to λ through a loglinear model, and to p through a logit model. Specifically,

$$\log\left(\frac{p}{1-p}\right) = \mathbf{X}_1\boldsymbol{\alpha} = \alpha_0 + \alpha_1X_{11} + \alpha_2X_{12} + \dots + \alpha_mX_{1m},$$

$$\log(\lambda) = \mathbf{X}_2\boldsymbol{\beta} = \beta_0 + \beta_1X_{21} + \beta_2X_{22} + \dots + \beta_lX_{2l}, \tag{9}$$

where $\mathbf{X}_1 = (1, X_{11}, X_{12}, \dots, X_{1m})$ is the covariate vector included in the zero stage and $\mathbf{X}_2 = (1, X_{21}, X_{22}, \dots, X_{2l})$ is the covariate vector included in the Poisson stage. $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \dots, \alpha_m)^T$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_l)^T$ are the corresponding coefficient vectors. The observed data are assumed to be independent. For other situations, where the independence assumption is relaxed, a correlation structure can be modelled by introducing random terms (see Dagne [6]). With covariates included in the ZIP regression model, the likelihood function for a random sample

$$Y = (y_1, y_2, \dots, y_n) \sim \text{ZIP}(\lambda_j = e^{\mathbf{X}_{2j}\boldsymbol{\beta}}, p_j = \frac{e^{\mathbf{X}_{1j}\boldsymbol{\alpha}}}{1 + e^{\mathbf{X}_{1j}\boldsymbol{\alpha}}}), (j = 1, 2, \dots, n),$$

is reparameterized as

$$\begin{aligned} f(Y|\boldsymbol{\alpha}, \boldsymbol{\beta}) &= \prod_{j=1}^k \left[\frac{e^{\mathbf{X}_{1j}\boldsymbol{\alpha}}}{1 + e^{\mathbf{X}_{1j}\boldsymbol{\alpha}}} + \left(1 - \frac{e^{\mathbf{X}_{1j}\boldsymbol{\alpha}}}{1 + e^{\mathbf{X}_{1j}\boldsymbol{\alpha}}} \right) e^{-e^{\mathbf{X}_{2j}\boldsymbol{\beta}}} \right] \\ &\quad \times \prod_{j=k+1}^n \left[\left(1 - \frac{e^{\mathbf{X}_{1j}\boldsymbol{\alpha}}}{1 + e^{\mathbf{X}_{1j}\boldsymbol{\alpha}}} \right) e^{-e^{\mathbf{X}_{2j}\boldsymbol{\beta}}} \frac{(e^{\mathbf{X}_{2j}\boldsymbol{\beta}})^{y_j}}{y_j!} \right], \end{aligned} \quad (10)$$

where n represents the total number of observations for Y among which a subset of the first k respondents have zero values. The y_{k+1}, \dots, y_n indicate observations with nonzero values.

With covariates added, the prior is assigned on the regression coefficients for the log-linear model (i.e., the Poisson count stage) and the logistic model (i.e., the Bernoulli zero-inflation stage) as outlined earlier. Without prior knowledge about the distribution of the parameter, Bayesians may choose non-informative priors. In doing so, the role of the prior distribution is minimized and more weight is given to the data in determining the posterior distribution. For example, with an independent normal (μ, θ^2) prior for each element in $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ with a large value of θ to reflect the lack (or ignorance) of prior information, we have the following joint prior distribution:

$$f(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{j=0}^m \left[\frac{1}{\sqrt{2\pi\sigma_{\alpha j}}} e^{\left\{ \frac{-(\alpha_j - \mu_{\alpha j})^2}{2\sigma_{\alpha j}^2} \right\}} \right] \prod_{j=0}^l \left[\frac{1}{\sqrt{2\pi\sigma_{\beta j}}} e^{\left\{ \frac{-(\beta_j - \mu_{\beta j})^2}{2\sigma_{\beta j}^2} \right\}} \right]. \quad (11)$$

Applying Bayes rule to the ZIP regression model with covariates based on the likelihood function in Equation (10) and the prior distribution function in Equation (11) yields the following joint posterior distribution with a nonstandard density:

$$f(\boldsymbol{\alpha}, \boldsymbol{\beta} | Y) \propto f(Y | \boldsymbol{\alpha}, \boldsymbol{\beta}) f(\boldsymbol{\alpha}, \boldsymbol{\beta})$$

$$\begin{aligned} & \propto \prod_{j=1}^k \left[\frac{e^{\mathbf{X}_{1j}\boldsymbol{\alpha}}}{1 + e^{\mathbf{X}_{1j}\boldsymbol{\alpha}}} + \left(1 - \frac{e^{\mathbf{X}_{1j}\boldsymbol{\alpha}}}{1 + e^{\mathbf{X}_{1j}\boldsymbol{\alpha}}} \right) e^{-e^{\mathbf{X}_{2j}\boldsymbol{\beta}}} \right] \\ & \times \prod_{j=k+1}^n \left[\left(1 - \frac{e^{\mathbf{X}_{1j}\boldsymbol{\alpha}}}{1 + e^{\mathbf{X}_{1j}\boldsymbol{\alpha}}} \right) e^{-e^{\mathbf{X}_{2j}\boldsymbol{\beta}}} \frac{(e^{\mathbf{X}_{2j}\boldsymbol{\beta}})^{y_j}}{y_j!} \right] \\ & \times \prod_{j=0}^m e^{\left\{ \frac{-(\alpha_j - \mu_{\alpha j})^2}{2\sigma_{\alpha j}^2} \right\}} \prod_{j=0}^l e^{\left\{ \frac{-(\beta_j - \mu_{\beta j})^2}{2\sigma_{\beta j}^2} \right\}}. \end{aligned} \quad (12)$$

4.3. Gibbs sampling

Because of the difficulty in analytically determining the posterior distribution, simulation techniques are generally used to obtain posterior information. Markov Chain Monte Carlo (MCMC) estimation is the most commonly used simulation method for Bayesian models. Gibbs sampling is a special case of MCMC, and is based on sampling of the full conditional distribution. Gibbs sampling is a powerful, and in many cases, an attractive approach to Bayesian inference (see Gelman et al. [8]; Ghosh et al. [9]; and Lynch [21] for more detailed discussion about MCMC and Gibbs sampling). The Gibbs sampler follows the following iterative steps for the unconditional ZIP model:

- (1) Begin with a set of initial values of p and λ .
- (2) Sample from the full conditional distribution according to $(p | \lambda, \mathbf{Y})$.

- (3) Sample from the full conditional distribution according to $(\lambda|p, \mathbf{Y})$.
- (4) Return to step 2 until convergence.

In the case of the ZIP regression model with covariates, the Gibbs sampler follows the following iterative steps:

- (1) Begin with a set of initial values of $\alpha_0, \alpha_1, \dots, \alpha_m$ and $\beta_0, \beta_1, \dots, \beta_l$.
- (2) Sample from the full conditional distribution according to

$$(\alpha_0|\alpha_1, \dots, \alpha_m, \beta_0, \beta_1, \dots, \beta_l, \mathbf{Y}, \mathbf{X}).$$

- (3) Sample from the full conditional distribution according to

$$(\alpha_1|\alpha_0, \dots, \alpha_m, \beta_0, \beta_1, \dots, \beta_l, \mathbf{Y}, \mathbf{X}).$$

⋮

- $(m + l + 3)$ Sample from the full conditional distribution according to

$$(\beta_l|\alpha_0, \dots, \alpha_m, \beta_0, \beta_1, \dots, \beta_{l-1}, \mathbf{Y}, \mathbf{X}).$$

- $(m + l + 4)$ Return to step 2 until convergence.

Dramatic increases in computing power over the past two decades coupled with the development of special-purpose programs make it relatively straightforward to carry out Bayesian analysis by using MCMC. The *Bayesian inference using Gibbs sampling* (BUGS) project has developed flexible software for Bayesian analysis of complex statistical models by using MCMC methods (Spiegelhalter et al. [26]). We use these tools to estimate the ZIP regression models in this paper.

5. Example: Modelling Smoking using the Bayesian ZIP Regression Models

Health behaviours such as smoking are topics of interest for social scientists because they can reflect many socio-economic disparities. Smoking varies by gender, race, age, socio-economic status as well as

marital status groups (Pampel [25]). Previous studies on smoking behaviours are mostly about heavy smokers and less attention is given to light smokers. Recent research suggests that light smoking has a significant impact on health and morality (Bjartveit and Tverdal [1]). Smoking one to four cigarettes daily triples the risk of dying of heart disease or lung cancer (Bjartveit and Tverdal [1]). In this study, we focus on light smokers and examine how socio-demographic characteristics predict light smoking behaviours.

Previous studies have usually recoded smoking into binary or multinomial or ordinal category variables, leading to a loss of detailed smoking count measurement information. As discussed earlier, this treatment of zero-inflated count data is not fully-efficient in terms of using all information from the data. We apply the Bayesian ZIP model to study smoking as an illustrative example of this methodology. To illustrate the flexibility of model evaluation in Bayesian analysis, we conduct posterior checks to evaluate the adequacy of the Bayesian ZIP model in fitting particular data values.

5.1. Data and variables

Data for this study come from the Americans' changing lives (ACL) survey of individuals in the contiguous United States (House [14]). We use the first wave of this survey, which was conducted in 1986. We use number of cigarettes smoked per day as the measure of smoking. The observed smoking data exhibit considerable heaping at values of 10, 20, etc., most likely due to the tendency among heavier smokers to base their estimated daily consumption in units of "half-packs" and "packs". This results in considerable round-off of the actual number of individual cigarettes smoked. We focus on light smoking and restrict the analysis to individuals, who reported less than 10 cigarettes smoked per day. The number of smoking in the final analyzed sample is highly zero-inflated and about 93.5 percent of the respondents report no daily cigarette use. There are 2,736 individuals included in our analysis.

We use socio-demographic covariates to predict smoking behaviour. These covariates include: age (in years), gender (1 = male, 0 = female), years of schooling, family income (in \$10,000) and marital status (1 = unmarried, 0 = married). We include all these covariates in the prediction of both the odds of being a non-smoker (using a logistic model) and the amount smoked (using a loglinear model) as specified in Equation (9). Note, the ZIP model can be made more flexible by allowing different covariates to predict the odds of being a non-smoker and the amount smoked. Table 1 provides descriptive statistics of the variables analyzed in this study. The mean number of cigarettes smoked per day was 0.269 in this sample. The average age of respondents in this sample was 55.681, and 35.3 percent of them were male. Sample respondents averaged 11.448 years of schooling and the mean family income was \$23,521. The unmarried accounted for 44.5 percent of the sample.

Table 1. Descriptive statistics of variables analyzed, $N = 2,736$

Variable	Mean	Std. Dev.	Min.	Max.
Number of cigarettes smoked	0.269	1.156	0	9
Age (in 10 years)	5.568	1.783	24	96
Male	0.353	0.478	0	1
Black	0.319	0.466	0	1
Education (years of schooling)	11.448	3.625	0	17
Income (\$10,000)	2.352	2.261	0.250	11.000
Unmarried	0.445	0.497	0	1

5.2. Results from the Bayesian ZIP regression models

We apply Gibbs sampling (20,000 MCMC iterations, with 3 chains) to the ZIP regression models by using the BUGS software (Spiegelhalter et al. [26]).¹ We estimate two Bayesian ZIP models without and with covariates included, respectively. For the unconditional ZIP model without covariates, we assume a uniform $(0, 1)$ prior for p , and a

¹ OpenBUGS (O'Hara et al. [24]) is an alternative.

gamma(0.5, 0.5) prior for λ . For the full ZIP model with all basic socio-demographic covariates controlled, we use a less informative prior, i.e., a diffuse normal distribution centered at 0 with a very large variance of 1000 to reflect the lack (or ignorance) of prior information about the distribution of parameters in the regression model. Specifically, we assume that $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are *a-priori* independent and normally distributed as

$$\begin{aligned}\alpha_j &\sim \text{Normal}(0, 1000), \\ \beta_j &\sim \text{Normal}(0, 1000), \\ j &= 1, 2, \dots, 7.\end{aligned}\tag{13}$$

We conducted some sensitivity tests by applying alternative priors on $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, such as uniform and truncated normal distributions, which yielded no significant different results as shown here. The BUGS codes for fitting ZIP models are provided in the Appendix.

Table 2 shows the MCMC estimates for the ZIP regression models for smoking. The posterior mean and median are reported in Table 2 for each regression parameters together with the standard deviation and the 2.5 and 97.5 percentiles of the posterior distributions.² The *p*-values (reported in Table 2 as a grouped level) are one-tailed empirical probabilities and they are calculated as the proportion of the posterior parameter values that are as extreme as zero (Lynch and Western [20]). The upper panel of Table 2 shows the results from the unconditional ZIP model. These results suggest that the estimated mean probability of being a non-smoker is 0.933, which is very close to the observed sample distribution, where 93.5 percent of the sample are non-smokers. These results are expected due to the fact that the ZIP model is specifically parameterized to predict zeros. The estimated mean number of cigarettes smoked per day *for the smokers* from the unconditional ZIP model is 4.035, which is

² The posterior mean and median are the two Bayesian rules, which minimize the posterior squared error loss and absolute error loss, respectively (Casella and Berger [4]).

close to the observed sample mean of 4.117. The full posterior distributions of the estimated unconditional ZIP model parameters are displayed in Figure 1.

Results from the full Bayesian ZIP model with all basic socio-demographic covariates included, shown in the lower panel of Table 2, suggest that older individuals are more likely to be non-smokers; but older smokers tend to smoke more heavily than younger smokers. Blacks and the unmarried have lower odds of being non-smokers than whites and the married. Smokers with higher family income tend to smoke less than smokers with lower income. The full posterior distributions of the estimated regression coefficients from the full ZIP regression model with covariates included are displayed in Figure 2.

Table 2. Bayesian estimates for the ZIP model: ACL smoking, $N = 2,736$

	Mean	Std.Dev	Median	2.5 Percentile	97.5 Percentile
Unconditional ZIP model					
p	0.933	0.005	0.933	0.923	0.942
λ	4.035	0.155	4.032	3.738	4.344
Deviance	802.600	15.340	800.100	775.500	837.600
Full ZIP model with covariates					
<i>Log odds of non-smoker</i>					
Intercept	1.690*	0.505	1.688	0.706	2.679
Age	0.192*	0.049	0.192	0.096	0.289
Black	-0.449*	0.169	-0.448	-0.782	-0.122
Male	-0.188	0.171	-0.188	-0.521	0.148
Education	0.032	0.026	0.032	-0.020	0.084
Income	0.052	0.052	0.051	-0.047	0.159
Unmarried	-0.582*	0.178	-0.581	-0.933	-0.235
<i>Log of smoking count</i>					
Intercept	1.161*	0.218	1.165	0.733	1.595
Age	0.054*	0.024	0.054	0.008	0.100
Black	0.149	0.081	0.146	-0.008	0.314
Male	0.076	0.081	0.076	-0.081	0.234
Education	-0.001	0.012	-0.001	-0.024	0.023
Income	-0.062*	0.032	-0.061	-0.127	-0.001
Unmarried	-0.057	0.090	-0.057	-0.230	0.126
Deviance	798.500	16.640	797.100	770.200	835.000

* $p < 0.05$.

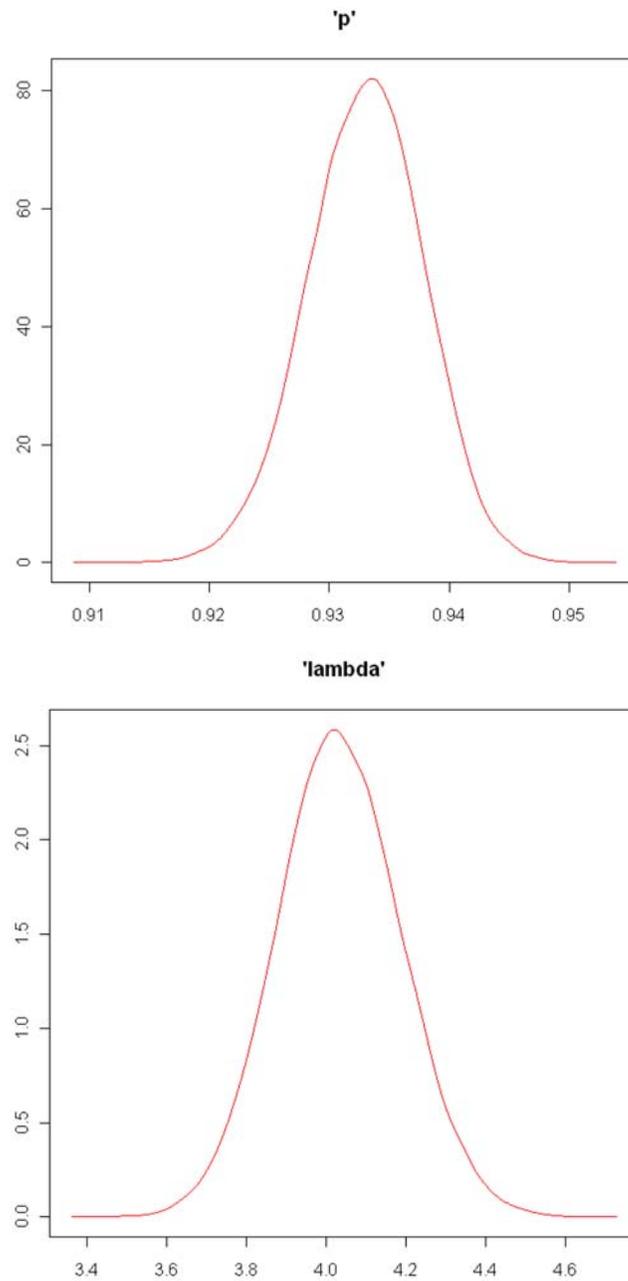


Figure 1. Posterior density plots of estimated parameters for the unconditional ZIP model.

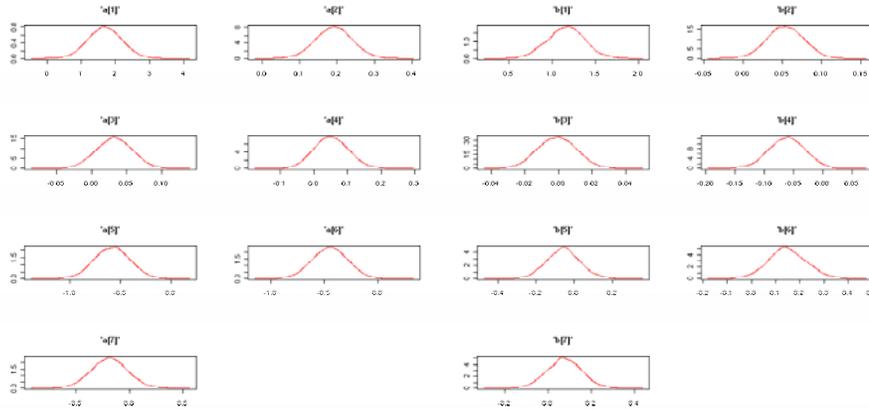


Figure 2. Posterior density plots of estimated regression coefficients from the full ZIP regression model with covariates. “a” represents coefficients in the Bernoulli zero-inflation stage and “b” represents coefficients in the Poisson count stage.

Bayesian analysis offers flexible tools for checking model fit that are not readily available in classical methods (Lynch and Western [20]). In particular, posterior checks can be informative about overall model fit on the entire data distribution as well as the extent to which the model fits specific aspects of the data. For example, we can evaluate how well the model fits the higher and/or lower end of the data distribution or a particular data value. This may be especially important in the case of zero-inflated count data because zero is a predominant feature of the distribution. Here, we utilize the relatively straightforward posterior predictive checking technique to evaluate the adequacy of the Bayesian ZIP models in fitting smoking data from the ACL survey.

The similarity of the observed and predicted data indicates a good model fit. We first compare the observed data distribution with the posterior predictive distribution of Y from both the unconditional and full ZIP models. We show the distributions for values greater than 0 in order to highlight the performance of the ZIP model in prediction of this range of data. Figure 3 shows the histograms of the posterior predictive smoking data (nonzero values) from the ZIP models and the observed

smoking data (nonzero values). Clearly, the histogram of posterior predictive smoking data from both the unconditional and full ZIP models are consistent with the distribution of observed data. The visually consistent shape of the posterior predictive distribution and observed distribution indicates that the ZIP models (either unconditional or full) are plausible for these data.

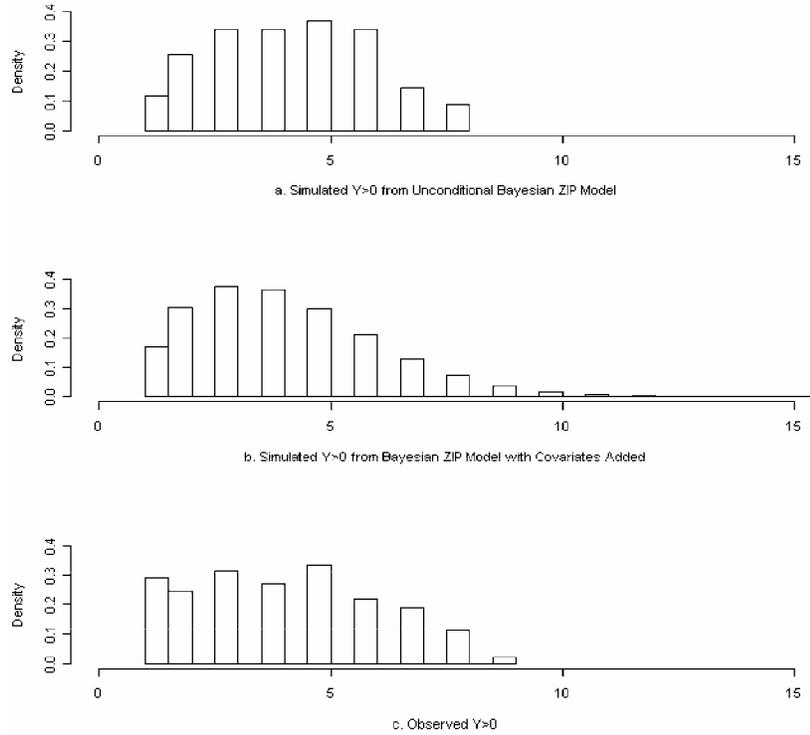
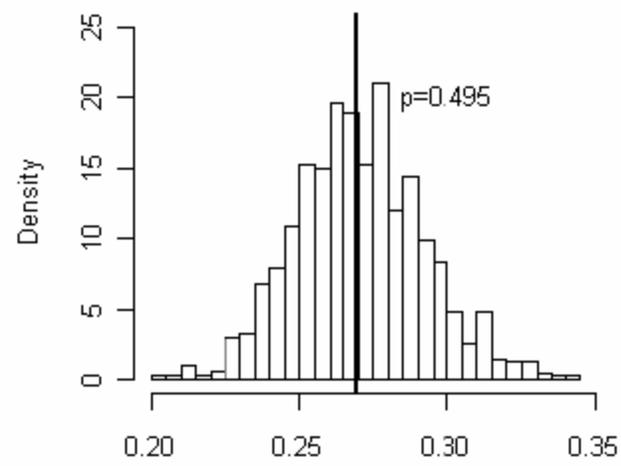
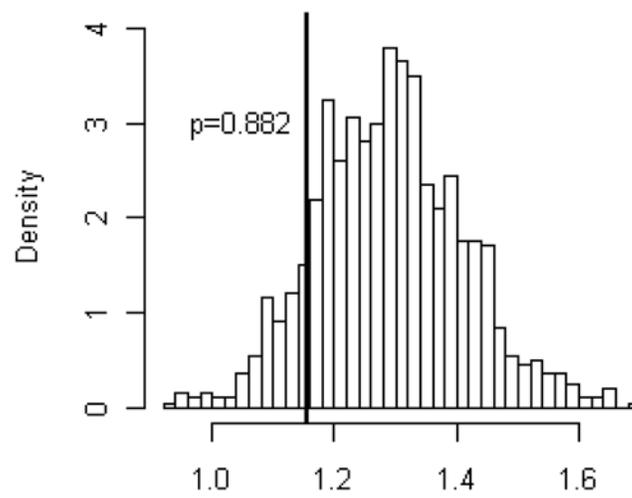


Figure 3. Posterior predictive distributions of $Y > 0$ from the unconditional and full Bayesian ZIP models and observed $Y > 0$.

To further check the adequacy of the model, we use the mean and standard deviation as two discrepancy statistics to check the plausibility of the observed data under the ZIP models by comparing the posterior predictive distributions of those statistics with the observed values (Lynch and Western [20]). We compute the mean and standard deviation

for the last 1,000 draws of the posterior predictive distribution. Figure 4 shows the posterior predictive distributions of the two discrepancy statistics from the ZIP models. The bold lines in the plots represent the observed sample mean and standard deviation. The closer the posterior predictive distribution around the bold line (i.e., the observed data), the better the model fit. The p -values reported in the plots for posterior predictive checks on the mean and standard deviation are computed as the proportion of posterior predictive values as large as the observed data. Extreme values of p (i.e., close to 0 or 1 such as a p -value < 0.05 or a p -value > 0.95) indicate poor model fit.

Panels (a) and (b) of Figure 4 display the predictive distributions from the unconditional ZIP model and panels (c) and (d) display the predictive distributions from the full ZIP model. From panels (a) and (c) of Figure 4, we can see that the observed mean falls in the middle of the posterior predictive distributions from both the unconditional and full ZIP models. Specifically, 495 out of 1,000 (i.e., from the last 1,000 draws in MCMC) posterior predictive mean values of smoking from the unconditional ZIP model are no smaller than the observed mean of 0.269 (i.e., p -value = 0.495); 543 out of 1,000 posterior predictive mean values of smoking from the full ZIP model with covariates are no smaller than the observed mean (i.e., p -value = 0.543). Panels (b) and (d) show that 882 out of 1,000 posterior predictive values for standard deviation of smoking from the unconditional ZIP model are no smaller than the observed standard deviation of 1.156 (i.e., p -value = 0.882); 522 out of 1,000 posterior predictive values for standard deviation of smoking from the full ZIP model with covariates are no smaller than the observed standard deviation (i.e., p -value = 0.522). These results suggest that adding socio-demographic covariates into the ZIP model results in better prediction of the dispersion of the smoking data. Nevertheless, the posterior predictive data, conditional on either the unconditional or full ZIP model, are reasonably plausible.

a. Unconditional ZIP: $T(y)=\text{mean}(y)$ b. Unconditional ZIP: $T(y)=\text{sd}(y)$

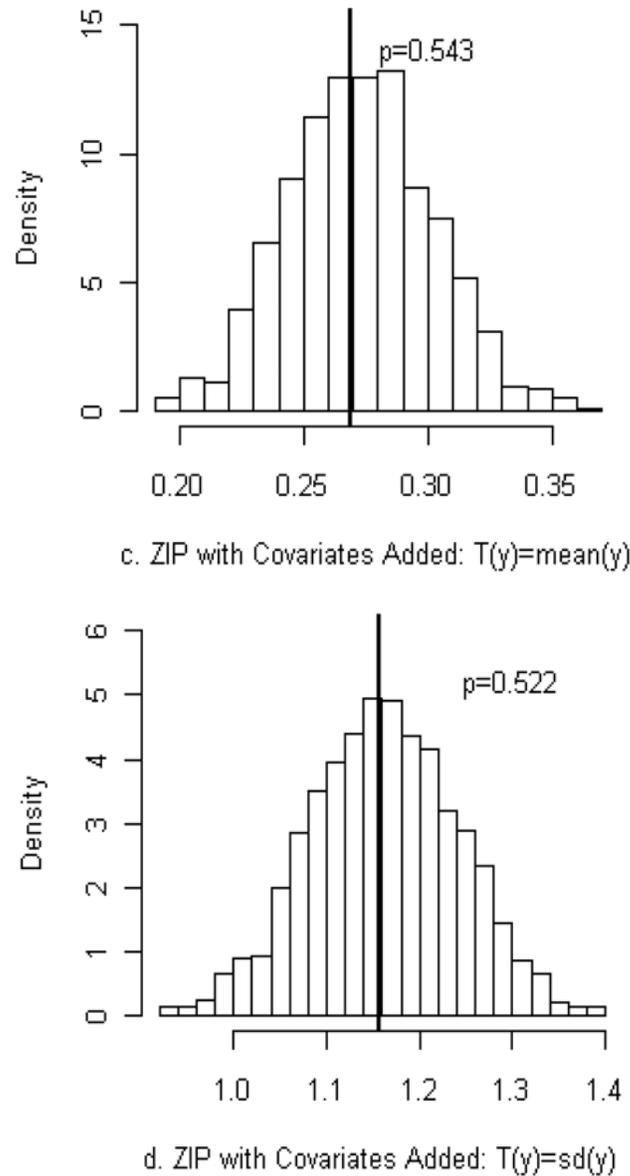


Figure 4. Posterior predictive checks on mean and standard deviation for the unconditional and full ZIP models. p -values are one-tailed empirical probabilities that the posterior predictive values are as large as the observed values from the original data. Bold vertical lines are the observed values from the original data.

For the sake of comparison, we also report the results from the classical ZIP model estimated based on maximum likelihood in Table 3. Notice that the classical and Bayesian ZIP estimations generally reveal similar results³ in this particular case with only one exception. That is, the estimated regression coefficient of family income on log of smoking account is significant in the Bayesian model in Table 2, while it is not significant in Table 3 based on MLE. This suggests that the Bayesian approach reveals more certainty in identifying another socio-demographic factor (i.e., family income) affecting smoking behaviour, which was not captured in the classical approach.

³ The ML estimate of p from the unconditional ZIP model is $\frac{e^{2.640}}{1 + e^{2.640}} = 0.933$. The ML estimate of λ from the unconditional ZIP model is $e^{1.398} = 4.047$. These ML estimates of p and λ are very similar to the Bayes estimates in Table 2.

Table 3. Maximum likelihood estimates for the ZIP model: ACL smoking, $N = 2,736$

	Estimate	Std. Error	95% C. I.
Unconditional ZIP model			
$\log\left(\frac{p}{1-p}\right)$	2.640*	0.077	[2.488, 2.792]
$\log(\lambda)$	1.398*	0.038	[1.323, 1.472]
Deviance	2091		
Full ZIP model with covariates			
<i>Log odds of non-smoker</i>			
(Intercept) z	1.680*	0.499	[0.702, 2.658]
Age	0.191*	0.049	[0.095, 0.287]
Black	-0.447*	0.168	[-0.776, -0.118]
Male	-0.192	0.169	[-0.523, 0.140]
Education	0.033	0.026	[-0.018, 0.084]
Income	0.050	0.052	[-0.052, 0.151]
Unmarried	-0.582*	0.178	[-0.931, -0.234]
<i>Log of smoking count</i>			
Intercept	1.163*	0.227	[0.718, 1.609]
Age	0.054*	0.025	[0.006, 0.102]
Black	0.150	0.082	[-0.010, 0.310]
Male	0.080	0.082	[-0.081, 0.240]
Education	-0.001	0.012	[-0.025, 0.023]
Income	-0.060	0.032	[-0.122, 0.002]
Unmarried	-0.057	0.093	[-0.239, 0.125]
Deviance	2023		

* $p < 0.05$.

"C. I." indicates confidence interval.

6. Comparing Maximum Likelihood and Bayesian Estimation of the ZIP Model

In order to further compare the performance of the proposed Bayesian approach for the ZIP model to the maximum likelihood approach, we carry out six simulation studies. For simplicity, we use the unconditional ZIP model without covariates in the simulation studies. We define various sample sizes, n , and true parameter values for the Poisson and Bernoulli parameters, λ and p , in the six studies in order to compare the performance of Bayesian estimation and maximum likelihood in fitting ZIP models under various scenarios. The six studies are based on the following combinations of true parameter values and sample sizes: (1) $n = 100$, $p = 0.1$, and $\lambda = 1$; (2) $n = 100$, $p = 0.5$, and $\lambda = 1$; (3) $n = 100$, $p = 0.9$, and $\lambda = 1$; (4) $n = 1000$, $p = 0.1$, and $\lambda = 1$; (5) $n = 1000$, $p = 0.5$, and $\lambda = 1$; and (6s) $n = 1000$, $p = 0.9$, and $\lambda = 1$.

The simulation process is based on the following steps:

- (1) Set the true parameter values of p and λ .
- (2) Generate $\mathbf{Y} = (y_1, y_2, \dots, y_n) \sim \text{ZIP}(\lambda, p)$.
- (3) Fit the ZIP model with maximum likelihood.
- (4) Fit the ZIP model with Bayesian simulation.
- (5) Repeat steps 2-4 for 500 times.

For each Bayesian estimation run, we apply Gibbs sampling with 10,000 MCMC iterations and 3 chains to the ZIP model by using the BUGS software. We assume a uniform(0, 1) prior for p , and a gamma(0.5, 0.5)⁴ prior for λ . The simulation process generated six sets of results reported in Table 4. We compare the coverage probabilities

⁴ The mean of gamma(0.5, 0.5) distribution is 0.25, which is different from the true λ value of 1 set in the simulations. This discrepancy reflects our non-informative prior knowledge.

based on 95 percent interval estimates⁵ and root mean squared error (RMSE)⁶ from Bayesian and MLE approaches in these six cases. Results in Table 4 show that when the sample size is 1,000, the interval estimates of p and λ from both the Bayesian and maximum likelihood approaches cover the true parameter values about 95 percent of the time. The values of RMSE from the two approaches are also quite similar when the sample size is 1,000. However, for cases with a sample size of 100, the interval estimates of p from the maximum likelihood approach provide lower coverage of the true parameter values when compared to Bayesian intervals, especially for the cases with values of p close to the boundary of the parameter space (i.e., $p = 0.9$ and $p = 0.1$). The Bayes estimator also has smaller values of RMSE in the case of $n = 100$ 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.

⁵ Wald confidence intervals were used in the simulation study. The likelihood ratio confidence intervals were suggested to be more accurate for the ZIP model, although more difficult to compute (Lambert [16]).

⁶ Root mean squared error (RMSE) is one of the most often-used measures for bias of an estimator, which is estimated as the square root of the average squared difference between the estimator $\hat{\theta}$ and the parameter θ : $\text{RMSE} = \sqrt{E(\hat{\theta} - \theta)^2}$.

Table 4. Simulation results: MLE vs. Bayes to estimate ZIP models

	Study-I $n = 100$		Study-II $n = 100$		Study-III $n = 100$		Study-IV $n = 1000$		Study-V $n = 1000$		Study-VI $n = 1000$	
True values	$p = 0.1$ $\lambda = 1$		$p = 0.5$ $\lambda = 1$		$p = 0.9$ $\lambda = 1$		$p = 0.1$ $\lambda = 1$		$p = 0.5$ $\lambda = 1$		$p = 0.9$ $\lambda = 1$	
Bayes	C.P.	RMSE										
p	0.976	0.764	0.934	0.460	0.902	0.173	0.944	0.801	0.950	0.406	0.938	0.017
λ	0.966	0.138	0.938	0.209	0.852	0.495	0.958	0.051	0.940	0.071	0.946	0.154
MLE	C.P.	RMSE										
p	0.728	0.802	0.954	0.433	0.872	0.259	0.928	0.803	0.962	0.403	0.954	0.016
λ	0.924	0.152	0.936	0.216	0.860	0.537	0.942	0.054	0.940	0.071	0.944	0.154

“C.P.” indicates coverage probability based on 95 percent interval estimates.

“RMSE” indicates root mean squared error.

7. Conclusion

Count data with excess zeros are commonplace in social science research. The ZIP regression model has been developed and utilized to handle data of this type, and is traditionally estimated by using maximum likelihood. This paper introduces a Bayesian alternative to estimate the ZIP model, as we believe it provides several advantages when compared to maximum likelihood estimation for this model. For example, as interval estimates receive increasing emphasis in social science research, the common-sense interpretation of Bayesian intervals (i.e., credible intervals) provides a strong impetus to adopt a Bayesian perspective (Gelman et al. [8]).

Perhaps more importantly for the case of the ZIP model, Bayesian analysis can provide the full joint distribution of the parameters of interest and can account for various sources of uncertainty in modelling zero-inflated count data, which is not easily achieved in classical maximum likelihood approaches (Gelman et al. [8]). Small samples along with parameters close to the boundary may result in extra uncertainty in parameters and failure of the asymptotic assumptions, which are critical in maximum likelihood estimation. These conditions may cast doubt on statistical inferences about model parameters for count data with excess zero values based on the variance of the estimator under maximum likelihood approaches (Gelman et al. [8]). Our simulation studies demonstrate that the Bayesian method performs better in the sense of yielding larger coverage probabilities and smaller bias than the classic maximum likelihood approach, especially in the case of small samples with either very high or very low incidence of excess zero outcomes.

Moreover, Bayesian approaches provide additional opportunities for inference that may not be available in classical methods. Researchers can choose the quantities of interests to monitor in order to reflect any aspect of the model, such as the posterior predictive distribution of Y , as we illustrated in the smoking example. In particular, the focus of inference

could be combinations of parameters, data, or both. This is a major advantage of the Bayesian approach. Depending on the goals of inference, more complex quantities can be simulated and monitored in the process of modelling. These could include constructed parameters and/or data that may not be simultaneously identified in the standard parameterizations of classical models. Yet, once these quantities are sampled, inference can be carried out in a straightforward way. This is in sharp contrast to a classical model-based analysis, where output usually consists of a set of regression coefficients and ancillary parameters of interest (i.e., variance components, threshold parameters, etc.), and where quantities of interest—beyond those reported in the output—must be generated in a post-estimation phase. The kinds of post-estimation inferences that can be carried out in this context depend to a great extent on the complexity of the resulting quantities of interest and the assumptions about their distributions (i.e., the extent to which standard errors can be analytically derived for these quantities).

Finally, the flexibility of Bayesian methods in modelling mixture data has particular relevance for the ZIP model. Count data with excess zeros are a special case of a two-level mixture structure that is a natural candidate for Bayesian analysis. It is often easier and more realistic to model complex data structures by thinking in terms of a mixture—specifically a Poisson-Bernoulli structure in the case of zero-inflated count data.

The ZIP model is specifically suited for modelling zero-inflated count data. Zero-inflation is one potential mechanism that generates overdispersion 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 overdispersion can result in biased parameter estimates. Other models such as the zero-inflated negative binomial (ZINB) model provide additional correction for overdispersion (see Greene [10]; Long [18]; Hall [11]; Long and Freese

[19]; see also Chin and Quddus [5], for a discussion comparing count models).⁷ Careful exploratory analysis to investigate additional sources of overdispersion using alternative models is recommended.

Bayesian analysis presents some challenges for researchers due to the computational intensiveness of this approach. Analysis using complex models and large data sets often take several hours or longer to converge to a stationary Markov chain. Further, Bayesian estimation problems usually must be programmed, thus requiring some sophistication on the part of the user. Aside from these complexities, recent developments in simulation methods and advances in computing power together provide researchers with powerful tools for estimation, inference, and prediction. Increasing availability of user-friendly software (e.g., WinBUGS, OpenBUGS, etc.) for applying Bayesian approach opens the door for social scientists to readily implement these methods. This paper illustrates the superiority of Bayesian analytic model of count data characterized by excess zeros, which is a data structure often observed but less proficiently explored in sociological research.

⁷ Specifically, the ZINB model adds a gamma-distributed multiplicative random effect to the Poisson model to account for an extra source of individual-level variation.

References

- [1] K. Bjartveit and A. Tverdal, Health consequences of smoking 14 cigarettes per day, *Tob Control* 14 (2005), 315-320.
- [2] A. Cameron and P. K. Trivedi, Econometric models based on count data: Comparisons and applications of some estimators and tests, *Journal of Applied Econometrics* 1 (1986), 29-53.
- [3] A. Cameron and P. K. Trivedi, Regression-based tests for overdispersion in the Poisson model, *Journal of Econometrics* 46 (1990), 347-364.
- [4] George Casella and Roger Berger, *Statistical Inference*, Second Edition, Duxbury/Thomson Learning, 2002.
- [5] H. C. Chin and M. A. Quddus, Modelling count data with excess zeroes: An empirical application to traffic accidents, *Sociological Methods and Research* 32 (2003), 90-111.
- [6] G. A. Dagne, Hierarchical Bayesian analysis of correlated zero-inflated count data, *Biometrical Journal* 46 (2004), 653-663.
- [7] C. Dean and J. F. Lawless, Tests for detecting overdispersion in Poisson regression models, *Journal of the American Statistical Association* 84 (1989), 467-472.
- [8] A. Gelman, J. B. Carlin, H. S. Stern and D. B. Rubin, *Bayesian Data Analysis*, Second Edition, Boca Raton, Fla, 2004.
- [9] S. K. Ghosh, P. Mukhopadhyay and J. C. Lu, Bayesian analysis of zero-inflated regression models, *Journal of Statistical Planning and Inference* 136 (2006), 1360-1375.
- [10] W. H. Greene, Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models, Working Paper EC-94-10, Department of Econometrics, New York University, 1994.
- [11] D. B. Hall, Zero-inflated Poisson and binomial regression with random effects: A case study, *Biometrics* 56 (2000), 1030-1039.
- [12] J. Hausman, D. B. Hall and Z. Griliches, Econometric models for count data with an application to the patents R & D relationship, *Econometrica* 52 (1984), 909-938.
- [13] D. Heilbron, Zero-altered and other regression models for count data with added zeros, *Biometrical Journal* 36 (1994), 531-547.
- [14] J. S. House, *Americans' Changing Lives: Waves I, II, and III, 1968, 1989, and 1994*, [Computer file], ICPSR version, Ann Arbor, MI: University of Michigan, Institute for Social Research, Survey Research Center [Producer], 2002; Ann Arbor, MI: Inter-university Consortium for Political and Social Research [Distributor], 2003.
- [15] G. King, Event count models for international relations: Generalizations and applications, *International Studies Quarterly* 33 (1989), 123-147.

- [16] D. Lambert, Zero-inflated Poisson regression with an application to defects in manufacturing, *Technometrics* 34 (1992), 1-14.
- [17] H. Liu and D. A. Powers, Growth curve models for zero-inflated count data: An application to smoking behaviour, *Structural Equation Modelling* 14(2) (2007), 247-279.
- [18] J. S. Long, *Regression Models for Categorical and Limited Dependent Variables*, *Advanced Quantitative Techniques in the Social Sciences*, Sage Publications, Inc., 1997.
- [19] J. S. Long and J. Freese, *Regression Models for Categorical Dependent Variables using Stata*, Second Edition, College Station, TX: Stata Press, 2006.
- [20] S. Lynch and B. Western, Bayesian posterior predictive checks for complex models, *Sociological Methods and Research* 32 (2004), 301-335.
- [21] S. Lynch, *An Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*, Springer Science and Business Media, LLC, 2007.
- [22] J. Mullahy, Specification and testing of some modified count data models, *Journal of Econometrics* 33 (1986), 341-365.
- [23] J. Mullahy, Heterogeneity, excess zeros, and the structure of count data models, *Journal of Applied Econometrics* 12 (1997), 337-350.
- [24] B. O'Hara, U. Liggs and S. Sturtz, Making BUGS Open, *R. News* 6 (2006), 12-17.
- [25] F. C. Pampel, Age and education patterns of smoking among women in high-income nations, *Social Science and Medicine* 57 (2003), 1505-1514.
- [26] D. J. Spiegelhalter, A. Thomas, N. G. Best and W. Gilks, *Bayesian Inference using Gibbs Sampling Manual Volume (Version II)*, Cambridge: MRC Biostatistics Unit, Institute of Public Health, 1996.
- [27] D. J. Spiegelhalter, N. G. Best, B. P. Carline and A. Van der Linder, Bayesian method of model complexity and fit (with discussion), *Journal of the Royal Statistical Society B* 64 (2002), 583-639.
- [28] D. Umberson, M. D. Chen, J. S. House, K. Hopkins and E. Slaten, Social relationships and their effects on psychological well-being: Are men and women really so different?, *American Sociological Review* 61 (1996), 836-856.
- [29] C. J. W. Zorn, An analytic and empirical examination of zero-inflated and Hurdle Poisson specifications, *Sociological Methods and Research* 26 (1998), 368-400.

**Appendix: BUGS Programming Codes for the
Zero-Inflated Poisson Regression Models**

The unconditional ZIP model:

```
model {for(j in 1 : J) {y[i] ~ dpois(mu[j]) mu[j] <- (1 - b[j])
  * lambda b[j] ~ dbern(p)}
```

```
# Priors: lambda ~ dgamma(0.5, 0.5) p ~ dunif(0, 1)}.
```

The ZIP model with covariates:

```
model {for(j in 1 : J) {smk2[j] ~ dpois(mu[j]) mu[j] <- (1 - u[j])
  * lambda[j] u[j] ~ dbern(p[j])
```

```
logit(p[j]) <- a[1] + a[2] * age[j] + a[3] * male[j] + a[4] * black[j] + a[5]
  * educ[j] + a[6] * income + a[7] * unmar[j],
```

```
log(lambda[j]) <- b[1] + b[2] * age[j] + b[3] * male[j] + b[4] * black[j] + b[5]
  * educ[j] + b[6] * income[j] + b[7] * unmar[j]}
```

```
for(k in 1 : 7){a[k] ~ dnorm(0, 0.001) b[k] ~ dnorm(0, 0.001)}.
```

