Summer 2020

Bayesian Zero-Inflated Model for Ordinal Data

Huizhong Yang

BAYESIAN ZERO-INFLATED MODEL FOR ORDINAL DATA

by

Huizhong Yang

Bachelor of Agriculture
Jiangxi Agricultural University, 2011

Master of Science
Jiangxi Agricultural University, 2014

Submitted in Partial Fulfillment of the Requirements

For the Degree of Master of Science in Public Health in

Biostatistics

The Norman J. Arnold School of Public Health

University of South Carolina

2020

Accepted by:

Bo Cai, Director of Thesis

Feifei Xiao, Reader

Robert Moran, Reader

Cheryl L. Addy, Vice Provost and Dean of the Graduate School

ACKNOWLEDGEMENTS

First, I would like to express my very great appreciation to my academic advisor, Dr. Feifei Xiao, and Director of Thesis, Dr. Bo Cai, for their direction, insight, and patience during my program. I wish to acknowledge the help provided by Dr. Robert Moran for his constructive suggestions and guidance. My grateful thanks are also extended to the faculty and staff of the Department of Epidemiology and Biostatistics and my fellow classmates for their support and encouragement throughout my studies.

I would like to offer my special thanks to Dr. Mary Kelley from Emory University for sharing her code which allowed me to do a model comparison. Her willingness to share her prior work so generously has been very much appreciated.

And lastly, I would like to acknowledge the emotional and academic support provided by my husband during my program and thesis.

ABSTRACT

Datasets with a relatively large number of zeros is commonly seen in medical applications. Although models like Zero-inflated Poisson (ZIP) model are proposed for counts data, there is still some issues with ordinal data which have excess zeros. In this paper, we developed a Bayesian approach to accommodate the excess zero in ordinal data. Intellectual disability (ID), also known as mental retardation (MR), is a disability characterized by below-average intelligence or mental ability and a lack of the learning necessary skills for daily life. A person with intellectual disability has intellectual functioning and adaptive behaviors limitations. Intellectual disability is a life-term disability and usually originates before birth. The ID data set contains numerus zeros since majority of children are normal, and the responses contain scaled levels. Motivated by a frequentist study using EM algorithm to maximize the log-likelihood iteratively for zero-inflated ordinal data, we apply a Bayesian method on the ID data set. The proposed method allows the unknown thresholds of latent variable be flexible and accommodate the excess zero at the same time. A simulation study is also conducted to evaluate the performance of the proposed method with comparison of the regular proportional odds model and frequentist zero-inflated proportional odds model.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ESE ..... Empirical Standard Error

ESMSE..... Empirical Squared Root of the Mean Squared Error

ID ..... Intellectual Disability

IQ ..... Intellectual Quotient

MCMC ..... Markov Chain Monte Carlo

MR ..... Mental Retardation

MSE ..... Mean Squared Error

PO ..... Proportional Odds

SSD ..... Sample Standard Deviation

ZINB ..... Zero-Inflated Negative Binomial

ZIP..... Zero-Inflated Poisson

ZIPO..... Zero-Inflated Proportional Odds

95CP..... 95% Coverage Probability

CHAPTER 1

INTRODUCTION

1.1 BACKGROUND

Intellectual disability (ID), also known as mental retardation (MR), is a disability characterized by below-average intelligence or mental ability and a lack of learning necessary skills for daily life. Person who with intellectual disability has intellectual functioning and adaptive behaviors limitations, which refers that he/she lack of practical skills, such as the ability to learn, make decisions, and solve problems, and also the lack of social skills for communicating with others effectively, interacting with others, and taking care of someone. Intellectual disability is a life-term disability and usually originates before birth. It can be caused by some factors interfering with normal brain development at any time during pregnancy and brain development period. The most common cause may include genetic conditions, malnutrition, infections, head injuries, and toxic substances exposures. Chemical exposures, sometimes emphasis on heavy metal exposures, such as arsenic (As), lead (Pb) and mercury (Hg) can cause developmental neurotoxicity which have been associated with causing neurobehavioral dysfunctions and sub-average intelligence. While relatively low levels of exposure seem to have a subtle effect, high doses these metals can cause ID (Goldman and Koduro 2000; Sullivan and Krieger 2001; Bellinger and Needleman 2003).

There are abundant studies showing that both prenatal and postnatal chemical metal exposure may affect neurodevelopment for children (Wasserman et al. 1994, 1997;

Counter et al. 2002; Liu et al. 2010). Plenty of evidence indicates that the chemical metals could cross the placenta and accumulate in fetal tissues (Gundacker and Hengstschläger 2012). Compared with adult toxicity, developmental neurotoxicity is potentially more severe and irreversible (Miodovnik 2011). The chemical metals could enter the human body by multiple methods including particulate inhalation, contact with contaminated soil or water, consumption of food sources grown or raised in contaminated fields, or even direct ingestion of contaminated soil or water. Oral and skin contact have been recognized as primary exposure pathway of chemical for pregnant women (Baghurst et al. 1992; Wassermann et al. 1994; Davidson et al. 2006). For children, hand to mouth contamination has been shown as a major route for chemicals such as Pb and As go through soil to blood (Goldman and Koduro 2000; Miodovnik 2011; Wang et al. 2007). Additionally, some prior studies have proven that urban areas exhibit increased metal concentrations in soil from industrial and transportation sources, while rural areas experience increased metal concentrations in soil of from natural geologic sources, pesticides, and industrial facilities (Li et al. 2004; Aelion et al. 2008; Davis et al. 2009).

This study data is from a retrospective cohort study of pregnant women who were insured by South Carolina Medicaid from 1996 through 2002 and resided in one of seven residential study areas during pregnancy. In this study, ID is categorized orderly by intellectual quotient (IQ), containing six levels: normal (0), very mild (1), mild (2), moderate (3), severe (4) and very severe (5). Additionally, the chemical concentration of the soil was measured in the residential study areas. Though only limited chemicals are discovered to have effect on ID generating, such as arsenic (As), lead (Pb) and mercury (Hg) (McDermott et al. 2011; Dufault et al. 2009), we would still like to assess whether

factors like demographics (mom's age or race, baby's birth weight and so on) and other chemicals (Chromium(Cr), Manganese(Mn), Barium(Ba), Copper(Cu), Nickel(Ni)) has association with ID. Figure 1.1 and Figure1.2 displayed the frequency of ID level for all seven sites and site 4 respectively, which showed how often each level occurred in the ID data. From Figure 1.1, we could tell that over 4000 responses in the aggregate ID data were normal (0) and the total subjects were 5016. While Figure 1.2 showed that the total subjects for site 4 were 1803 with over 1500 responses in ID data were normal (0). Since the data had so many zeros that we decided to use zero-inflated model.

1.2 MOTIVATION AND OUTLINE OF STUDY

Cumulative models are widely utilized for ordinal categorical data among regression models (LÄÄRÄ and MATTHEWS 1985; Albert and Chib 2001; Liu and Agresti 2005). In cumulative models, the ordinal responses are derived from categorizing a continuous latent variable by adjacent intervals on the continuous scale. The observed outcomes can be viewed as the result of a cumulative process in each category which can be reached in sequence. The most popular cumulative model for ordinal responses is proportional odds (PO) model (McCullagh 1980; Agresti 2013), which is a regression logistic model with each cumulative logit has its own intercept and slopes are equal.

The dataset with excess zeros is common in both manufacturing applications and medical applications. Appropriate statistical methods are very essential to investigate the inner logical connection for excess zeros data in most of the scientific fields. Some models were developed to accommodate the extra zeros issue like using Zero-inflated Poisson Regression (ZIP) model (Lambert 1992; Hall 1994) or using Zero-Inflated

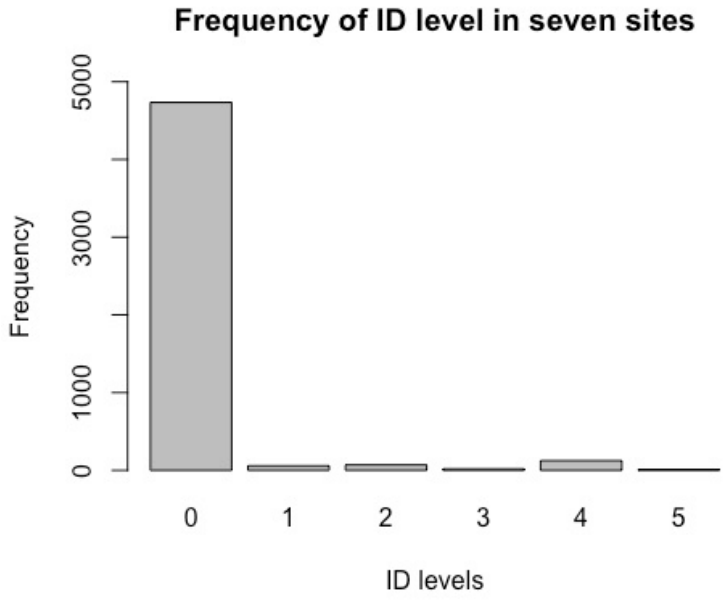

Figure 1.1 Frequency of ID levels in seven sites.

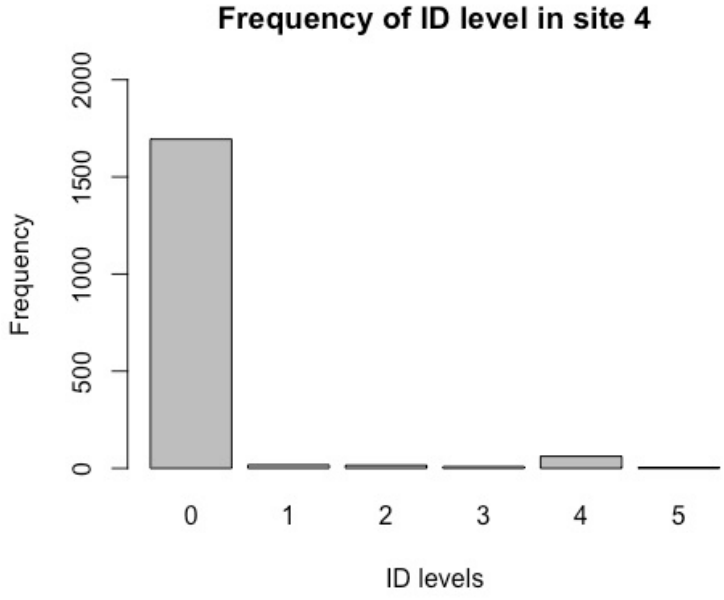

Figure 1.2 Frequency of ID levels in sites 4.

Negative Binomial (ZINB) model to avoid over-dispersion (Yau et al. 2003; Ridout et al. 2001). In contrast to the count data with excess zeros, the methods for ordinal data with excess zeros have been less developed, Kelley and Anderson (2008) proposed the zero-inflated proportional odds (ZIPO) model by using EM algorism to maximize the log-likelihood iteratively.

To our knowledge, the Bayesian methods for ordinal response data with excess zeros was yet developed, In this thesis we proposed a Bayesian methods for analysis of the ordinal data with excess zeros.

The contents of this study are as follows.

In Chapter 2, the proposed model (zero-inflated proportional odds model) was introduced along with the prior distributions being specified. Posterior distribution derivation and posterior computation by R and WinBUGS (R Development Core Team 2007; Lunn 2000) were done. Also, the "zeros trick" in WinBUGS was illustrated which can help create a new sampling distribution for the proposed mixture model.

In Chapter 3, we evaluated the proposed model by simulation study. At first, we explained how to generate the unusual excess zero data. Then the results from the proposed model were shown and compared to the results of the proportional odds model. Additionally, the model sensitivity analysis was performed.

In Chapter 4, ID data was used to do the real data analysis. The coefficients and 95% credible intervals are calculated for each covariate. The results from the proposed model were compared to the results of using the frequentist method (Kelley and Anderson 2008).

In Chapter 5, the work in this study was summarized and possible future research was discussed.

CHAPTER 2

THE ZERO-INFLATED PROPORTIONAL ODDS MODEL

In this chapter, we introduced structure of zero-inflated proportional odds model and how it fit the excess zero data. More specifically, in section 2.1, we explained the two distribution parts that constituted the zero-inflated proportional odds model. In section 2.2, we specified the prior of all the unknown parameters and prior reparameterization for threshold parameters. In section 2.3, we showed the work of getting the likelihood and illustrated the processing of posterior distribution computation. Also, we explained a "zeros trick" which helped us design a new distribution in WinBUGS that was not in the distribution selections of WinBUGS system.

2.1 MODEL DESIGN

Our goal was to produce a model that could be more accurate when fit to ordinal scale data in which not all the observations have the symptom or phenomenon being accessed/evaluated. The model was structured as follows. We created the model by regarding the response as two aspects: one was modelling the incidence, which means the symptom occurred or not; the other one was modelling the severity, i.e. the symptom levels or scale. In order to make it clearer in the subsequent explanation of the distributions, we called the process of modelling whether the symptom exist as "presence state" and call the process of modelling the level of the symptom as "scale state". By which, we could model the presence of the symptom and the scale of the symptom at the same time.

2.1.1 PRESENCE STATE

We used Bernoulli distribution to model whether the symptom occurred. We let y_i be either 0 or 1 on the ith subject (i = 0,1,…,n). We assumed that $y_i = 0$ (which means no symptom) is a success, then

$$y_i \sim \text{Bernoulli}(1 - p_i).$$

From which, we could get $y_i = 0$ with probability p_i.

Let $\boldsymbol{X}_i$ denote all of the covariates on the ith subject, we chose $\boldsymbol{x}_i$ to represent the subset of $\boldsymbol{X}_i$, which could include itself also. We used $\boldsymbol{x}_i$ to be the linear predictors for the probability of the presence state and $\boldsymbol{\alpha}$ as coefficients. Then the link function was

$$logit(p_i) = \boldsymbol{x}_i\boldsymbol{\alpha}.$$

2.1.2 SCALE STATE

For ordinal scales component, we used proportional odds model settings (McCullagh 1980; Agresti 2013; Congdon 2014). We let y_i be an ordinal scale response on the ith subject (i=0,1,…n) with level 0, 1,…,J. The scale response y often reflected the outcome of a latent continuous variable ω, and

$$\omega_i = \boldsymbol{x}_i\boldsymbol{\beta} + \varepsilon_i,$$

where, $\boldsymbol{x}_i$ is used as the subset of $\boldsymbol{X}_i$ as we mentioned before, and ε_i has a logistic distribution function. Thus y_i was obtained according to the location of ω_i on the scale, which means $y_i = j$ when $\theta_{j-1} \leq \omega_i < \theta_j$. Next, we imported an indicator M, which

$$M_i \sim Mult(1, p_{i0}, \ldots, p_{iJ}; \textstyle\sum_{j=0}^{J} p_{ij} = 1).$$

Multinomial distribution is a generalization of the Binomial distribution. Rather than only have two responses, “success” or “failure”, Multinomial distribution has $J + 1$ responses (0,….J level). If $y_i = j$, then $M_{ij} = 1$, otherwise, $M_{ij} = 0$. So that

$$
\begin{aligned}
p_{ij} &= \Pr(y_i = j), \\
&= \Pr\left(\theta_{j-1} \le \omega_i < \theta_j\right), \\
&= \Pr\left(\theta_{j-1} \le \boldsymbol{x}_i\boldsymbol{\beta} + \ \varepsilon_i < \theta_j\right), \\
&= \Pr\left(\theta_{j-1} - \boldsymbol{x}_i\boldsymbol{\beta} \le \ \varepsilon_i < \theta_j - \boldsymbol{x}_i\boldsymbol{\beta}\right), \\
&= \mathrm{P}\left(\theta_j - \boldsymbol{x}_i\boldsymbol{\beta}\right) - P\left(\theta_{j-1} - \boldsymbol{x}_i\boldsymbol{\beta}\right), \\
&= \gamma_{i,j} - \gamma_{i,j-1}\,,
\end{aligned}
$$

where

$$\gamma_{i,j} = \Pr(y_i \le \mathrm{j}) = \mathrm{P}\left(\theta_j - \boldsymbol{x}_i\boldsymbol{\beta}\right), \quad j = 0,\ldots, J-1.$$

Also, $\gamma_{i,j}$ was cumulative probability, therefore, $\gamma_{i,j} = p_{i1} + \cdots + p_{ij}$. So that

$$
\begin{aligned}
p_{i0} &= \gamma_{i,0}, \\
p_{ij} &= \gamma_{i,j} - \gamma_{i,j-1}, \quad j = 1, \ldots, J\text{-}1 \\
p_{iJ} &= 1 - \gamma_{i,J-1}.
\end{aligned}
$$

Since $P(\varepsilon)$ was logistic,

$$logit\left(\gamma_{i,j}\right) = \ \theta_j - \boldsymbol{x}_i\boldsymbol{\beta}.$$

$\boldsymbol{\theta} = (\theta_0, \ldots, \theta_{J-1})$ was the threshold for each categories 0,…,J-1, which made the difference of adjacent cumulative logits being independent for each category. $\boldsymbol{\beta}$ were identical coefficients for all categories.

2.2 PRIOR SPECIFICATION AND REPARAMETERIZATION

We still needed to decide the prior distributions for all the unknown parameters so that we could complete the Bayesian specification of the proposed model. Assuming the different parameter vectors were independent, we chose Normal distribution as the prior

distribution for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, which was also the conventional choice pf priors for the regression coefficients. So that

$$\boldsymbol{\alpha} \sim Normal(0, \boldsymbol{\sigma}_{\boldsymbol{\alpha}}^2),$$

$$\boldsymbol{\beta} \sim Normal\left(0, \boldsymbol{\delta}_{\boldsymbol{\beta}}^2\right).$$

From the previous model setting, we got threshold $\boldsymbol{\theta} = (\theta_0, \dots, \theta_{J-1})$ which was unknown, and $\theta_0, \dots, \theta_{J-1}$ was monotone increasing ($\theta_1 \leq \theta_2 \leq \cdots \leq \theta_{J-1}$). We did not want to restrict the difference of each adjacent scale to be identical, so the reparameterization approach was used here. We assumed that

$$\theta_0 \sim Normal\left(0, \sigma_{\theta_0}^2\right),$$

$$\theta_1 = \theta_0 + \Delta_1,$$

$$\dots$$

$$\theta_{J-1} = \theta_{J-2} + \Delta_{J-1}.$$

For $\Delta_1, \dots, \Delta_{J-1}$, we adopted Bayesian Hierarchical data structure, which assumed that

$$\Delta_j \sim \exp\left(\lambda_j\right), \quad j = 1, \dots, J-1$$

$$\lambda_j \sim \exp(\tau). \quad j = 1, \dots, J-1$$

$\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_{J-1})$ was hyperprior for Δ.

2.3 POSTERIOR COMPUTATION

The full conditional posterior distributions for the unknown parameters could be derived through likelihood and prior distribution. Then we could do posterior computation based on the conditional posterior distribution and the prior distributions we chose previously. The model we proposed was a mixture distribution regression model, which cannot be found in the list of standard distribution. In the posterior computation process, we used a method called the “zeros trick” to restructure our model in WinBUGS.

2.3.1 FULL CONDITIONAL POSTERIOR DISTRIBUTIONS DERIVATION

From the model design section, we could easily get that

$$y_i = \begin{cases} 0 & \text{with probability } p_i + (1-p_i)\gamma_{i,0} \\ j & \text{with probability } (1-p_i)\left(\gamma_{i,i} - \gamma_{i,j-1}\right), \quad j = 1, \dots, J-1 \end{cases}$$. And from these, we could get likelihood

$$l(\alpha, \beta; y_i) = \prod_{i=1}^{n} \{[p_i + (1-p_i)\gamma_{i,0}][I(y_i = 0)] + (1-p_i)(\gamma_{i,j} - \gamma_{i,j-1})[I(y_i = j)]\}.$$

$$j = 1, \dots, J$$

And

$$l(\alpha, \beta; y_i) = \prod_{i=1}^{n} \{\frac{1}{1+\exp(\boldsymbol{x}_i\boldsymbol{\alpha})} \{[\exp(\boldsymbol{x}_i\boldsymbol{\alpha}) + \frac{1}{1+\exp(-\theta_0 + \boldsymbol{x}_i\boldsymbol{\beta})}][I(y_i = 0)]$$

$$+ (\frac{1}{1+exp(-\theta_j + \boldsymbol{x}_i\boldsymbol{\beta})} - \frac{1}{1+exp(-\theta_{j-1} + \boldsymbol{x}_i\boldsymbol{\beta})})[I(y_i = j)]\}\}.$$

$$j = 1, \dots, J$$

With the prior distributions, we could get conditional posterior distributions for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$:

$$P(\boldsymbol{\alpha}|\boldsymbol{\beta}, \boldsymbol{\theta}, y_i) \propto \frac{\exp(-\boldsymbol{\alpha}^2/\sigma_\alpha^2)}{1+\exp(\boldsymbol{x}_i\boldsymbol{\alpha})} \{[\exp(\boldsymbol{x}_i\boldsymbol{\alpha}) + \frac{1}{1+exp(-\theta_j + \boldsymbol{x}_i\boldsymbol{\beta})}][I(y_i = 0)] + [I(y_i = j)]\}$$

$$P(\boldsymbol{\beta}|\boldsymbol{\alpha}, \boldsymbol{\theta}, y_i) \propto \exp\left(-\boldsymbol{\beta}^2/\sigma_\beta^2\right) \{[\exp(\boldsymbol{x}_i\boldsymbol{\alpha}) + \frac{1}{1+exp(-\theta_j + \boldsymbol{x}_i\boldsymbol{\beta})}][I(y_i = 0)]$$

$$+ (\frac{1}{1+\exp\left(-\theta_j + \boldsymbol{x}_i\boldsymbol{\beta}\right)} - \frac{1}{1+\exp\left(-\theta_{j-1} + \boldsymbol{x}_i\boldsymbol{\beta}\right)})[I(y_i = j)]\}$$

2.3.2 POSTERIOR COMPUTATION

By multiplying the prior by the likelihood and taking samples from the posterior distributions through the iterative algorithm, the posterior computation used the Gibbs sampler to run iterations and update unknown parameters. After specifying initial values

for the parameters, the proposed MCMC algorithm proceeded by updating the unknown parameters consecutively. The implementation was processed by using R and WinBUGS. The proposed MCMC algorithm proceeds by updating the unknown parameters sequentially as shown below.

Step 1: Update $\boldsymbol{\theta}$.

Step 2: Update $\boldsymbol{\alpha}$ from its full conditional distribution.

Step 3: Update $\boldsymbol{\beta}$ from its full conditional distribution.

To specify a new sampling distribution in WinBUGS, we used the "zeros trick". Suppose we had observations y_i and its likelihood L_i, and

$$y_i \sim Poisson(\lambda_i),$$

then the observation of zeros had likelihood

$$L_i = \exp(-\lambda_i),$$

from which we could get

$$\lambda_i = -\log(L_i).$$

If responses of our data were a set of zeros, then λ_i should be set to $-\log(L_i)$. λ_i should always be larger than 0 as the property of Poisson distribution. As such we needed to add a large enough constant to make sure it was larger than 0. This trick can be demonstrated by the example code as shown below.

```
C <- 10000                          # set a large enough constant
for (i in 1:n){zeros[i] <-0         # create a set of zeros
phi[i] <- -log(L[i]) + C
zeros[i] ~ dpois(phi[i])}
```

By this approach, we could set likelihood L_i as whatever we needed to, especially when dealing with the truncated distributions and mixture distributions.

CHAPTER 3

SIMULATION STUDIES

To evaluate the performance of the proposed model, we conducted a simulation study with 100 replications. In section 3.1, we introduced how to generate data sets for the simulation since the data is quite exceptional rather than the common data sets we would use. Also, we introduced the parameters choosing and prior distributions in the simulation study. In section 3.2, we evaluated the performance of the proposed model through the estimation results and did model sensitivity analysis. In section 3.3, we compared the two different results from the proposed model and the proportional odds model to show the difference.

3.1 SIMULATION DATA GENERATION

3.1.1 DATA GENERATING PROCESS

In order to structure data to have the same properties as the excess zero data, we started from choosing the regression coefficients ($\boldsymbol{\alpha}$, $\boldsymbol{\beta}$) for both presence state and scale state. With the fixed covariates $\boldsymbol{x}_i$, we could get the Bernoulli probability p_i for presence state and Multinomial cumulative probability $\gamma_{i,0}, \dots, \gamma_{i,J-1}$ (levels were from 0 to J) for scale state.

To keep the model simpler and focus more on the study itself, we assumed the linear predictor $\boldsymbol{x}_i$ were the same for both states, which was not obligatory for the proposed model. Then we could gain the generated the ordinal responses y_i by the following steps:

Step1: Choose a sample size n, parameters ($\boldsymbol{\alpha}$, $\boldsymbol{\beta}$), and fixed covariates $\boldsymbol{x}_i$ (should be continuous or binary covariates).

Step2: For each observation i, generate the probabilities from the chosen linear predictors, which is through

$$p_i = \frac{exp(\boldsymbol{x}_i\boldsymbol{\alpha})}{1 + exp(\boldsymbol{x}_i\boldsymbol{\alpha})}$$

$$\gamma_{j,i} = \frac{exp(\theta_j - \boldsymbol{x}_i\boldsymbol{\beta})}{1+exp(\theta_j - \boldsymbol{x}_i\boldsymbol{\beta})} \qquad j = 0, \dots J-1.$$

Step 3: Generate variable B_i from the Bernoulli distribution with probability $1 - p_i$.

Step 4: Generate variable u_i from Uniform (0,1) distribution, and get categorical variable t_i through

$$t_i = j \text{ if } \gamma_{i,j-1} \leq u_i \leq \gamma_{i,j} \quad j = 0, \dots, J$$

$$\text{with } \gamma_{-1,i} = 0 \; and \; \gamma_{J,i} = 1.$$

Step 5: Generate categorical responses y_i through B_i and t_i:

$$y_i = B_i * t_i.$$

Step 6: Repeat the steps for n times.

Step 7: Generate the data sets with 100 replicates.

3.1.2 PARAMETERS CHOICES

We began with choosing sample size n=100, 500, 1000. Then we assumed there are five levels (0,1,2,3,4) for ordinal responses y_i. For fixed covariates $\boldsymbol{x}_i$, we chose to use x_1, x_2 where x_1 was a continuous variable randomly drawn from the Normal (0, 1) and x_2 was a binary variable randomly drawn from the Binomial (0.5). For parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, we used $\alpha_0, \alpha_1, \alpha_2$ and β_1, β_2, in which $\alpha_1, \alpha_2, \beta_1, \beta_2$ were all set to equal to 2. We

supposed the Bernoulli probability p = 0.001, 0.033, 0.27, 0.62 for presence state at the baseline (which means both x_1 and x_2 were equal to 0) where we could get α_0= -4.5951, -3.3777, -0.9946, 0.4895 for each p from

$$p = \frac{exp(\alpha_0)}{1+exp(\alpha_0)}.$$

Then we could get Bernoulli probability p_i for presence state of each individual by

$$p_i = \frac{exp(\boldsymbol{x_i\alpha})}{1+exp(\boldsymbol{x_i\alpha})},$$

and got an average of p_i that was equal to 0.1, 0.2, 0.5, 0.7. We assumed that the Multinomial cumulative probability $\gamma_{i,0}, \gamma_{i,1}, \gamma_{i,2}, \gamma_{i,3}$ for scale state were 0.1, 0.3, 0.6, 0.9 at the baseline. Then we could get $\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2, \theta_3)$ from

$$\gamma_{j,i} = \frac{exp(\theta_j)}{1+exp(\theta_j)} \qquad j = 0, \ldots J-1,$$

which was $\boldsymbol{\theta}$ = (-2.1972, -0.8473, 0.4054, 2.1972).

3.1.3 PRIORS CHOICES

Assuming the coefficients for two mixture parts were independent, we chose to set both $\boldsymbol{\sigma_\alpha^2}$ and $\boldsymbol{\delta_\beta^2}$ equal to 1000, which allowed $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ to be flexible. We chose $\sigma_{\theta_0}^2$=1000 also. For the hyper parameters $\boldsymbol{\lambda}$, which

$$\lambda_j \sim \exp(\tau),$$

we choose τ = 0.1, 0.5, 5.

3.1.4 MCMC SAMPLING SETTINGS

For each MCMC chain, we did 5,000 iterations with a setting thinner of 10 to reduce the autocorrelation and discarded the first 1000 as burn in.

3.2 SUMMARY OF SIMULATION RESULTS

We selected four estimations for each parameter to assess the model performance. The point estimate was the average of the 100 posterior means. The empirical standard error (ESE) was the average of the 100 estimated standard errors. The sample standard deviation (SSD) was the sample standard deviation of the 100 posterior means. The 95% coverage probability(95CP) was the percent of the 100 credible intervals for each parameter that contained the true value.

The estimation results of data with sample size 100, 500, and 1000 were showed in Table 3.1. We could see the proposed model fit reasonably well, such that where the larger the sample size, the better the fit. When the sample size was smaller than 100, the Bias, SSD, and ESE became larger and showed convergence issues. In order to check whether the model would be affected by the percentage of excess zeros in data, we chose several different averages of p_i for generating data. Table 3.2 showed the summary of estimation results when average of the Bernoulli probability p_i= 0.1, 0.2, 0.5, 0.7 with n =500 and τ = 0.5. Table 3.2 demonstrated that when the percentage of excess zeros in data became too large, the Bias and SSD increased and the estimates of scale state would be not as precise as the estimates when the probability was smaller. Although the situation happened as mentioned before, the estimates were still close to the true value, which means the proposed model was not be affected too much by the percentage of excess zeros in data as long as the percentage was not incredibly large. Also, with varying the τ value, we can do model sensitivity analysis to see if the model will be affected by hyper parameter changing. From Table 3.3, no big changes were revealed in

Table 3.1 Estimation results for simulation with 100, 500, and 1000 sample size. (τ = 0.5 and average of p_i= 0.2)

Parameters		Sample size n=100				
	True	Est.	Bias	SSD	ESE	95CP
α_0	-3.38	-6.24	-2.86	4.38	0.0513	85
α_1	2	2.07	0.07	2.05	0.0211	85
α_2	2	3.20	1.20	3.32	0.0404	84
β_1	2	4.01	2.01	0.40	0.0062	92
β_2	2	2.03	0.03	0.54	0.0085	96
Parameters		Sample size n=500				
	True	Est.	Bias	SSD	ESE	95CP
α_0	-3.38	-3.51	-0.13	0.48	0.0075	96
α_1	2	2.07	0.07	0.31	0.0047	96
α_2	2	2.09	0.09	0.42	0.0064	95
β_1	2	2.02	0.02	0.17	0.0026	92
β_2	2	1.97	-0.03	0.27	0.0036	89
Parameters		Sample size n=1000				
	True	Est.	Bias	SSD	ESE	95CP
α_0	-3.38	-3.44	0.06	0.36	0.0052	94
α_1	2	2.04	0.04	0.21	0.0033	97
α_2	2	2.05	0.05	0.28	0.0045	97
β_1	2	2.00	0.00	0.11	0.0018	95
β_2	2	2.00	0.00	0.18	0.0025	90

Table 3.2 Estimation results for simulation with average of the Bernoulli probability p_i= 0.1, 0.2, 0.5, 0.7 for presence state (n =500 and τ = 0.5)

Parameters		Average of probability p_i = 0.1				
	True	Est.	Bias	SSD	ESE	95CP
α_0	-4.60	-5.08	-0.48	0.86	0.0122	92
α_1	2	2.19	0.19	0.41	0.0060	94
α_2	2	2.29	0.29	0.62	0.0092	96
β_1	2	2.01	0.01	0.14	0.0023	97
β_2	2	1.99	-0.01	0.21	0.0034	96
Parameters		Average of probability p_i = 0.2				
	True	Est.	Bias	SSD	ESE	95CP
α_0	-3.38	-3.51	-0.13	0.48	0.0075	96
α_1	2	2.07	0.07	0.31	0.0047	96
α_2	2	2.09	0.09	0.42	0.0064	95
β_1	2	2.02	0.02	0.17	0.0026	92
β_2	2	1.97	-0.03	0.27	0.0036	89
Parameters		Average of probability p_i = 0.5				
	True	Est.	Bias	SSD	ESE	95CP
α_0	-0.99	-1.10	-0.11	0.29	0.0046	96
α_1	2	2.14	0.14	0.31	0.0046	92
α_2	2	2.12	0.12	0.33	0.0057	97
β_1	2	2.00	0.00	0.22	0.0036	97
β_2	2	1.99	0.01	0.30	0.0050	96
Parameters		Average of probability p_i = 0.7				
	True	Est.	Bias	SSD	ESE	95CP
α_0	0.49	0.27	-0.22	0.94	0.0052	91
α_1	2	2.11	0.11	1.01	0.0063	96
α_2	2	2.05	0.05	1.88	0.0094	91
β_1	2	1.89	-0.11	0.47	0.0051	93
β_2	2	1.83	-0.17	0.55	0.0074	93

Table 3.3 Estimation results for simulation with τ = 0.1, 0.5, 5. (n = 500 and average of p_i= 0.2)

Parameters		τ = 0.1				
	True	Est.	Bias	SSD	ESE	95CP
α_0	-3.38	-3.53	-0.15	0.48	0.0075	97
α_1	2	2.08	0.08	0.31	0.0047	96
α_2	2	2.09	0.09	0.42	0.0064	95
β_1	2	2.01	0.01	0.17	0.0026	92
β_2	2	1.96	-0.04	0.27	0.0036	88
Parameters		τ = 0.5				
	True	Est.	Bias	SSD	ESE	95CP
α_0	-3.38	-3.51	-0.13	0.48	0.0075	96
α_1	2	2.07	0.07	0.31	0.0047	96
α_2	2	2.09	0.09	0.42	0.0064	95
β_1	2	2.02	0.02	0.17	0.0026	92
β_2	2	1.97	-0.03	0.27	0.0036	89
Parameters		τ = 5				
	True	Est.	Bias	SSD	ESE	95CP
α_0	-3.38	-3.49	-0.11	0.48	0.0074	96
α_1	2	2.06	0.06	0.31	0.0046	96
α_2	2	2.07	0.07	0.41	0.0064	95
β_1	2	2.04	0.04	0.17	0.0026	93
β_2	2	1.99	-0.01	0.27	0.0036	89

the estimates for the different hyper parameters. The proposed model performance was stable.

3.3 COMPARISON WITH PROPORTIONAL ODDS MODEL

Except the simulation studies with the proposed model (Bayesian ZIPO model), we also ran the proportional odds model with the same 100 data sets as well. The proportional odds model was selected as the comparison have two main reasons. First, the proportional odds model has been the most commonly used model for ordinal responses data for decades. Second, the proportional odds model had a similar structure as the scale part in our proposed model. Through the comparison of the two models we could show more clearly whether it made a difference when the excess zeros part was not taken into consideration. The two models were compared based on estimate, bias, SSD, ESE, and 95CP. For each MCMC chain of the proportional odds model, we ran 5,000 iterations with a setting thinner of 10 to reduce the autocorrelation. Of these, the first 1000 were discarded as burn in, which was the same settings we used for proposed model.

Table 3.4 to Table 3.7 were the comparisons between proposed model (Bayesian ZIPO model) and proportional odds model with τ = 0.5, n = 500 and average of p_i= 0.1, 0.2, 0.5, 0.7. The proportional odds model had much greater Bias compared to the proposed model, which resulted in 0 of 100 95% credible intervals containing true value. When the percentage of excess zeros was small (average of p_i= 0.1, 0.2), the coefficients of parameters (β_1, β_2) from the proportional odds model were positive, which were the same direction as the proposed model, and contained true value. However, when the percentage of excess zeros increased (average of p_i= 0.5, 0.7), the coefficients of

Table 3.4 Comparison between proposed model and proportional odds model with τ = 0.5, n = 500 and average of p_i= 0.1.

Parameters		Bayesian ZIPO Model				
	True	Est.	Bias	SSD	ESE	95CP
α_0	-4.60	-5.08	-0.48	0.86	0.0122	92
α_1	2	2.19	0.19	0.41	0.0060	94
α_2	2	2.29	0.29	0.62	0.0092	96
β_1	2	2.01	0.01	0.14	0.0023	97
β_2	2	1.99	-0.01	0.21	0.0034	96
Parameters		Proportional Odds Model				
	True	Est.	Bias	SSD	ESE	95CP
α_0	-4.60					
α_1	2					
α_2	2					
β_1	2	0.75	-1.25	0.13	0.0015	0
β_2	2	0.78	-1.22	0.17	0.0026	0

Table 3.5 Comparison between proposed model and proportional odds model with τ = 0.5, n = 500 and average of p_i= 0.2.

Parameters		Bayesian ZIPO Model				
	True	Est.	Bias	SSD	ESE	95CP
α_0	-3.38	-3.51	-0.13	0.48	0.0075	96
α_1	2	2.07	0.07	0.31	0.0047	96
α_2	2	2.09	0.09	0.42	0.0064	95
β_1	2	2.02	0.02	0.17	0.0026	92
β_2	2	1.97	-0.03	0.27	0.0036	89
Parameters		Proportional Odds Model				
	True	Est.	Bias	SSD	ESE	95CP
α_0	-3.38					
α_1	2					
α_2	2					
β_1	2	0.20	-1.80	0.10	0.0014	0
β_2	2	0.19	-1.81	0.16	0.0026	0

Table 3.6 Comparison between proposed model and proportional odds model with τ = 0.5, n = 500 and average of p_i= 0.5.

Parameters		Bayesian ZIPO Model				
	True	Est.	Bias	SSD	ESE	95CP
α_0	-0.99	-1.10	-0.11	0.29	0.0046	96
α_1	2	2.14	0.14	0.31	0.0046	92
α_2	2	2.12	0.12	0.33	0.0057	97
β_1	2	2.00	0.00	0.22	0.0036	97
β_2	2	1.99	0.01	0.30	0.0050	96
Parameters		Proportional Odds Model				
	True	Est.	Bias	SSD	ESE	95CP
α_0	-0.99					
α_1	2					
α_2	2					
β_1	2	-0.48	-2.48	0.09	0.0015	0
β_2	2	-0.55	-2.55	0.18	0.0029	0

Table 3.7 Comparison between proposed model and proportional odds model with τ = 0.5, n = 500 and average of p_i= 0.7.

Parameters		Bayesian ZIPO Model				
	True	Est.	Bias	SSD	ESE	95CP
α_0	0.49	0.27	-0.22	0.94	0.0052	91
α_1	2	2.11	0.11	1.01	0.0063	96
α_2	2	2.05	0.05	1.88	0.0094	91
β_1	2	1.89	-0.11	0.47	0.0051	93
β_2	2	1.83	-0.17	0.55	0.0074	93
Parameters		Proportional Odds Model				
	True	Est.	Bias	SSD	ESE	95CP
α_0	-0.99					
α_1	2					
α_2	2					
β_1	2	-0.73	-2.73	0.11	0.0020	0
β_2	2	-0.88	-2.88	0.25	0.0038	0

parameters (β_1, β_2) from the proportional odds model were negative. It showed that if the percentage of excess zeros in data was small, though it had relatively large bias, the results from the proportional odds model could imply the correct effect direction of covariates towards the responses. However, when the percentage of excess zeros in data was large, the results from the proportional odds model may not have been reliable. This was not the case with the proposed model, which could indicate the precise results whether the percentage of excess zeros in the data was small or large. Table 3.8 showed the comparison between the proposed model and the proportional odds model with τ = 0.5 and average of p_i = 0.2 when the sample size was 100. The proportional odds model, unlike the proposed model, it did not have convergence issues when the sample size was small, but the coefficients of parameters (β_1, β_2) were negative, which were not consistent with the true value. These also implied that with more parameters than normal, the proposed model needed a larger sample size to reach the convergence. From Table 3.9, with the sample size equal to 1000, the Bias of estimate from the proposed model was slightly smaller than the Bias of estimate with the sample size equal to 500.

3.4 COMPARE BAYESIAN AND FREQUENTIST METHODS

We did a comparison with a frequentist method of zero-inflated proportional odds model (refer to Kelly and Anderson, 2008) which used EM algorithm to maximize the log-likelihood iteratively. In the comparison, we used the data with sample size n = 500 and average of p_i = 0.1, 0.2, 0.5, 0.7. In the Bayesian ZIPO model, we chose the hyper parameter τ = 0.5. For each MCMC chain, we performed 5,000 iterations where the first 1000 were discarded as burn in and used a setting thinner of 10 to reduce the autocorrelation. For the EM algorithm in the frequentist ZIPO model the iteration was

Table 3.8 Comparison between proposed model and proportional odds model with τ = 0.5, n = 100 and average of p_i= 0.2.

Parameters		Bayesian ZIPO Model				
	True	Est.	Bias	SSD	ESE	95CP
α_0	-3.38	-6.24	-2.86	4.38	0.0513	85
α_1	2	2.07	0.07	2.05	0.0211	85
α_2	2	3.20	1.20	3.32	0.0404	84
β_1	2	4.01	2.01	0.40	0.0062	92
β_2	2	2.03	0.03	0.54	0.0085	96
Parameters		Proportional Odds Model				
	True	Est.	Bias	SSD	ESE	95CP
α_0	-3.38					
α_1	2					
α_2	2					
β_1	2	-0.38	-2.38	0.23	0.0034	0
β_2	2	-0.46	-2.46	0.40	0.0064	0

Table 3.9 Comparison between proposed model and proportional odds model with τ = 0.5, n = 1000 and average of p_i= 0.2.

Parameters		Bayesian ZIPO Model				
	True	Est.	Bias	SSD	ESE	95CP
α_0	-3.38	-3.44	0.06	0.36	0.0052	94
α_1	2	2.04	0.04	0.21	0.0033	97
α_2	2	2.05	0.05	0.28	0.0045	97
β_1	2	2.00	0.00	0.11	0.0018	95
β_2	2	2.00	0.00	0.18	0.0025	90
Parameters		Proportional Odds Model				
	True	Est.	Bias	SSD	ESE	95CP
α_0	-3.38					
α_1	2					
α_2	2					
β_1	2	0.20	-1.80	0.07	0.0010	0
β_2	2	0.21	-1.79	0.12	0.0018	0

4000. Table 3.10 - Table 3.13 were the summary of the comparison. There was a little difference between results from Bayesian ZIPO and frequentist ZIPO models. We also calculated the ESMSE, which is empirical squared root of the mean squared error (MSE). The ESMSEs from the frequentist ZIPO model were smaller than those of the Bayesian ZIPO model. However, the ESEs of the Bayesian ZIPO model were much smaller than the ESEs of the frequentist ZIPO model.

Table 3.10 Summary of results from Bayesian and frequentist methods with average of p_i= 0.1.

Parameters		Bayesian ZIPO Model				
	True	Est.	Bias	SSD	ESE	ESMSE
α_0	-4.60	-5.08	-0.48	0.86	0.0122	0.98
α_1	2	2.19	0.19	0.41	0.0060	0.45
α_2	2	2.29	0.29	0.62	0.0092	0.68
β_1	2	2.01	0.01	0.14	0.0023	0.14
β_2	2	1.99	-0.01	0.21	0.0034	0.21
Parameters		Frequentist ZIPO Model				
	True	Est.	Bias	SSD	ESE	ESMSE
α_0	-4.60	-4.79	-0.19	0.78	0.0317	0.80
α_1	2	2.08	0.08	0.38	0.0160	0.39
α_2	2	2.13	0.13	0.57	0.0241	0.59
β_1	2	2.02	0.02	0.14	0.0066	0.14
β_2	2	2.00	0.00	0.21	0.0095	0.21

Table 3.11 Summary of results from Bayesian and frequentist methods with average of p_i= 0.2.

Parameters		Bayesian ZIPO Model				
	True	Est.	Bias	SSD	ESE	ESMSE
α_0	-3.38	-3.51	-0.13	0.48	0.0075	0.50
α_1	2	2.07	0.07	0.31	0.0047	0.32
α_2	2	2.09	0.09	0.42	0.0064	0.43
β_1	2	2.02	0.02	0.17	0.0026	0.17
β_2	2	1.97	-0.03	0.27	0.0036	0.27
Parameters		Frequentist ZIPO Model				
	True	Est.	Bias	SSD	ESE	ESMSE
α_0	-3.38	-3.38	0.00	0.46	0.0202	0.46
α_1	2	1.99	-0.01	0.30	0.0127	0.30
α_2	2	2.00	0.00	0.40	0.0176	0.40
β_1	2	2.03	0.03	0.14	0.0072	0.18
β_2	2	1.98	-0.02	0.27	0.0102	0.27

Table 3.12 Summary of results from Bayesian and frequentist methods with average of p_i= 0.5.

Parameters		Bayesian ZIPO Model				
	True	Est.	Bias	SSD	ESE	ESMSE
α_0	-0.99	-1.10	-0.11	0.29	0.0046	0.31
α_1	2	2.14	0.14	0.31	0.0046	0.34
α_2	2	2.12	0.12	0.33	0.0057	0.35
β_1	2	2.00	0.00	0.22	0.0036	0.22
β_2	2	1.99	0.01	0.30	0.0050	0.30
Parameters		Frequentist ZIPO Model				
	True	Est.	Bias	SSD	ESE	ESMSE
α_0	-0.99	-1.04	-0.05	0.27	0.0123	0.27
α_1	2	2.06	0.06	0.29	0.0125	0.29
α_2	2	2.05	0.05	0.32	0.0155	0.32
β_1	2	2.02	0.02	0.22	0.0101	0.22
β_2	2	2.02	0.02	0.30	0.0141	0.30

Table 3.13 Summary of results from Bayesian and frequentist methods with average of p_i= 0.7.

Parameters		Bayesian ZIPO Model				
	True	Est.	Bias	SSD	ESE	ESMSE
α_0	0.49	0.27	-0.22	0.94	0.0052	0.96
α_1	2	2.11	0.11	1.01	0.0063	1.02
α_2	2	2.05	0.05	1.88	0.0094	1.88
β_1	2	1.89	-0.11	0.47	0.0051	0.48
β_2	2	1.83	-0.17	0.55	0.0074	0.58
Parameters		Frequentist ZIPO Model				
	True	Est.	Bias	SSD	ESE	ESMSE
α_0	0.49	0.43	-0.05	0.27	0.0117	0.28
α_1	2	2.08	0.08	0.31	0.0154	0.32
α_2	2	2.12	0.12	0.46	0.0189	0.47
β_1	2	1.98	-0.02	0.33	0.0142	0.33
β_2	2	1.93	-0.07	0.46	0.0206	0.47

CHAPTER 4

REAL DATA ANALYSIS

A retrospective cohort study was conducted to identify associations of heavy metal soil concentrations and intellectual disability (ID) in children for urban and rural residential neighborhoods of South Carolina. This was a cohort study of pregnant women who lived in one of seven residential study areas during pregnancy from 1996 through 2002 and were insured by South Carolina Medicaid. Medicaid provided health insurance coverage for people living under the federal poverty level, including eligible low-income adults, children, pregnant women, elderly adults and people with disabilities. Medicaid is administered by each state while following federal requirements. In South Carolina, pregnant women living below an income level that was 185% of the Federal poverty level were eligible for Medicaid, which contributed to 50% of the births in South Carolina during this study period (McDermott 2011). In this study, the pregnant women were followed from pregnancy to delivery. Then, the newborns were continuously followed to identify if the child received a diagnosis of ID. After merging the Medicaid files for the mothers and records for children, we could get the data of ID responses with 8–12 years of follow-up time.

Because of the confidentiality agreement and uncontrollable landscape, the collecting and testing of soil samples conducted using grid intersection points samples throughout whole the residential study strips. Eight chemicals were measured in the soil samples, which included arsenic (As), chromium (Cr), mercury (Hg), lead (Pb),

manganese (Mn), barium (Ba), copper (Cu), and nickel (Ni) and. Other demographic covariates included mother's age (which ranged from age 12 to 42), mother's ethnicity (white and non-white), mother's alcohol consumption during pregnancy (yes/no), number of prior births (parity) (0, 1, 2 and 3 or more), child birth weight, child gender, and weeks of gestation (which ranged from 23 to 42 weeks). The response of this study, ID, was scaled orderly with six levels: normal (0), very mild (1), mild (2), moderate (3), severe (4) and very severe (5).

We chose site 4 for the applications. The data was consisted of 1803 individual samples, included 7 demographic covariates of pregnancy and newborns' information, as well as chemical concentration information. In the data set of site 4, 1694 individuals of 1803 were identified as normal (0), 18 individuals of 1803 were identified as very mild (1), 16 individuals of 1803 were identified as mild (2), 9 individuals of 1803 were identified as moderate (3), 62 individuals of 1803 were identified as severe (4), 4 individuals of 1803 were identified as very severe (5). Figure 4.1 showed the locations of ID responses in Site 4. From the Figure 4.1, we can tell that the majority of the ID responses is normal (0). It also meant the data contained abundant excess zeros where we could use our proposed zero-inflated model to accommodate this problem.

We chose similar priors for the coefficients as we did in simulation studies with a hyper parameter $\tau = 0.5$. We ran the MCMC sampling through WinBUGS for 6000 iterations after discarding 1000 burn in with thinner of 10. For scale state, we chose partial covariates rather than full covariate vectors, which included birth weight, child gender, and the metals As, Hg, Pb, Mn, and Cu.

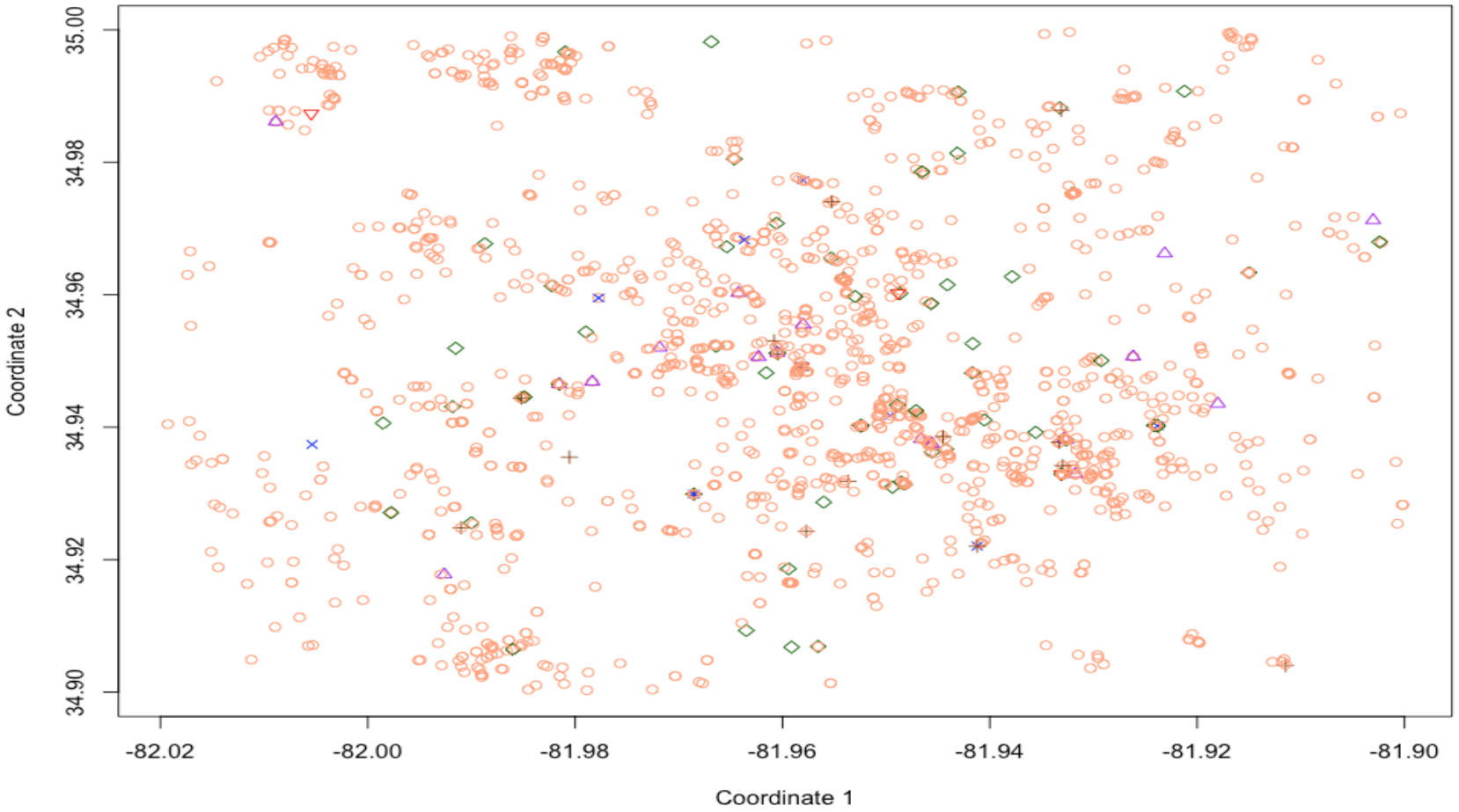

Figure 4.1 The locations for ID responses in Site 4. ‘O’ denotes ‘normal (0)’, ‘△’ denotes ‘ very mild (1)’, ‘＋’ denotes ‘mild (2)’, ‘×’ denotes ‘moderate (3)’, ‘◇’ denotes ‘severe(4)’ and ‘▽’denotes ‘very severe (5)’.

4.1 APPLICATION RESULTS

Table 4.1 indicated the summary of regression coefficients for the demographic and chemical concentration covariates of the presence state in the model. The coefficient for mothers' age showed the greater of the mothers' age, the less probability the children would be in normal level of ID. Though some studies implied that mothers whose age are over 35 would have higher chance to have a baby with ID, the result for the coefficient of mothers' age was not significant. This may be due to the percentage of mothers whose age was over 35 were only 1.55% in our data. The alcohol consumption of mothers was under the same condition. The coefficient of mothers' alcohol consumption showed that the child whose mother did not have alcohol during pregnancy had higher probability in the normal level of ID than the child whose mother did had alcohol. However, the coefficient was not significantly increasing, which again maybe due to the percentage of the mothers having alcohol during pregnancy being only 0.72% in our data. These were also consistent under the environment that doctors and media suggest women do not have alcohol during pregnancy.

Table 4.2 implied that lower birth weight, male gender and greater concentration of Pb exposure would significantly increase the risk of children having a higher ID level rather than normal. These are consistent with the conclusions of other previous studies (Cai et al. 2016).

4.2 COMPARISON WITH PROPORTIONAL ODDS MODEL

The proportional odds model also fitted with the same ID data. We ran the MCMC sampling through WinBUGS for 6000 iterations after discarding 1000 for burn-in with a thinner of 10. Because the proportional odds model did not have the presence

Table 4.1 The summary of regression coefficients for the demographic and chemical concentration covariates of the presence state in the model.

Covariates	Posterior Means	95% Credible Interval
Mother's age	-18.48	(-57.15, 16.09)
Mother's race (White)	2.18	(-57.38, 51.54)
Birth weight (kg)	4.63	(-13.32, 27.05)
Male child	-41.24	(-86.48, 2.66)
Alcohol consumption (no)	1.84	(-62.07,65.68)
Parity	-38.49	(-83.65, 9.69)
Gestational age (weeks)	26.36	(-21.44, 68.15)
As	-11.55	(-45.94, 16.15)
Cr	-7.61	(-51.4, 31.5)
Hg	-15.90	(-53.8, 17.45)
Pb	11.05	(-27.15, 41.86)
Mn	8.49	(-47.46, 58.0)
Ba	-11.11	(-51.05, 32.76)
Cu	-2.81	(-43.12, 32.14)
Ni	-13.7	(-62.9, 41.13)

Table 4.2 The summary of regression coefficients for the demographic and chemical concentration covariates of the scale state in the model.

Covariates	Posterior Means	95% Credible Interval
Birth weight (kg)	-0.28	(-0.46, -0.10)
Male child	0.57	(0.14, 0.99)
As	-0.04	(-0.28, 0.17)
Hg	-0.10	(-0.33, 0.12)
Pb	0.28	(0.03, 0.52)
Mn	-0.08	(-0.37, 0.19)
Cu	0.01	(-0.26, 0.33)

state, we only use the partial covariates in scale state. In Table 4.3, we saw that even though the results from two different models were greatly similar for most items, they were not exactly the same. The results from the proportional odds model implied that male gender and greater concentration of Pb exposure would significantly increase the risk of children having a higher ID level other than normal. However, the lower birth weight has no significant effect in the results of the proportional odds model. Also, in the proposed model, the odds for male gender having the risk of a higher ID level was smaller.

Table 4.3 Comparison with proportional odds model.

	Proposed Model		Proportional Odds Model	
Covariates	Posterior Means	95% Credible Interval	Posterior Means	95% Credible Interval
Birth weight (kg)	-0.28	(-0.46, -0.10)	-0.32	(-0.50,0.14)
Male child	0.57	(0.14, 0.99)	0.77	(0.37, 1.20)
As	-0.04	(-0.28, 0.17)	-0.04	(-0.27, 0.17)
Hg	-0.10	(-0.33, 0.12)	-0.10	(-0.31, 0.10)
Pb	0.28	(0.03, 0.52)	0.24	(0.03, 0.45)
Mn	-0.08	(-0.37, 0.19)	-0.09	(-0.34, 0.15)
Cu	0.01	(-0.26, 0.33)	0.03	(-0.21, 0.24)

CHAPTER 5

CONCULSION AND DISCUSSION

5.1 CONCLUSION

In this study, we have used the mixture model to accommodate zero inflation in ordinal data rather than use choose zero-inflated Poisson (ZIP) or zero-inflated negative Binomial (ZINB) which were more designed for counts data. The simulation study showed that the proposed model fit the data reasonably well. Due to having a greater number of the parameters compared to that of other simple models, the proposed model still had some limitations under some certain conditions. For example, the model showed convergence concern when the sample size was small or had too many predictors. However, this would not be a problem for large cohort studies and we also considered reducing some covariates for the applications in order to improve performance in these conditions.

5.2 DISCUSSION

For the application study, we used the proposed model to estimate the coefficients of parameters through estimating the unknown thresholds of the latent variable. By adopting the hyper parameter, we could keep the estimation of the thresholds flexible with the data which can help the model fit more effectively.

Applications of the proposed model should not be limited to just cohort studies. Data sets with many zeros are also common in clinical trials and manufacturing processes. Currently, there are few studies of zero-inflated models using Bayesian

method for categorical data, or more specifically ordinal data. This study is a good start in that direction. For the further research, we can also add spatial and temporal information to the data for prediction.

REFERENCES

Aelion CM, Davis HT, McDermott S, Lawson AB. 2008. Metal concentrations in rural topsoil in South Carolina: potential for human health impact. *Sci Total Environ.* 402: 149–156.

Agresti A. 2013. *Categorical Data analysis*. Third ed. New Jersey: John Wiley & Sons.

Albert J H, Chib S. 2001. Sequential Ordinal Modeling with Applications to Survival Data. *Biometrics.* 57: 829-836.

Baghurst PA, McMichael AJ, Wigg NR, Vimpani GV, Robertson EF, Roberts RJ, Tong SL. 1992. Environmental exposure to lead and children's intelligence at the age of seven years. The Port Pirie Cohort Study. *N Eng J Med*. 327: 1279–1284.

Bellinger DC, Needleman HL. 2003. Intellectual impairment and blood lead levels. *New Engl J Med*. 349: 500–502.

Cai B, Lawson AB, McDermott S, Aelion CM. 2016. A Bayesian semiparametric approach with change points for spatial ordinal data. *Stat Methods Med Res.* 25(2): 644-58.

Congdon P. 2014. *Applied Bayesian modelling.* Second ed. United Kingdom: John Wiley & Sons.

Counter SA, Buchanan LH, Ortega F, Laurell G. 2002. Elevated blood mercury and neuro-otological observations in children of the Ecuadorian gold mines. *J Toxicol Environ Health A.* 65: 149–163.

Davidson PW, Myers GW, Weiss B, Shamlaye CF, Cox C. 2006. Prenatal methyl mercury exposure from fish consumption and child development: a review of evidence and perspectives from the Seychelles Child Development Study. *Neurotoxicology.* 27: 1106–1109.

Davis HT, Aelion CM, McDermott S, Lawson AB. 2009. Identifying natural and anthropogenic sources of metals in urban and rural soils using GIS-based data, PCA, and spatial interpolation. *Environ Pollut.*157: 2378–2385.

Dufault R, Schnoll R, Lukiw WJ, LeBlanc B, Cornett C, Patrick L, Wallinga D, Gilbert SG & Crider R. 2009. Mercury exposure, nutritional deficiencies and metabolic disruptions may affect learning in children. *Behav Brain Funct*. 5: 44.

Goldman LR, Koduru S. 2000. Chemicals in the environment and developmental toxicity in children: a public health and policy perspective. *Environ Health Perspect.* 108 (Suppl. 3): 443–448.

Gundacke C, Hengstschläger M, 2012. The role of the placenta in fetal exposure to heavy metals. *Wien Med Wochenschr.* 162: 201–206.

Hall DB. 1994. Zero-Inflated Poisson and Binomial Regression with Random Effects: A Case Study. *Biometrics.* 56: 1030-1039.

Kelley ME, Anderson SJ. 2008. Zero inflation in ordinal data: Incorporating susceptibility to response through the use of a mixture model. *Stat Med.* 27(18): 3674–3688.

LÄÄRÄ E, MATTHEWS JNS, 1985. The equivalence of two models for ordinal data, *Biometrika*. 72: 1, 206–207.

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

Liu I, Agresti A. 2005. The analysis of ordered categorical data: An overview and a survey of recent developments. *Test*. 14: 1–73.

Liu Y, McDermott S, Lawson AB, Aelion CM. 2010. Analysis of soil concentrations of arsenic, mercury and lead and child outcomes of mental retardation and developmental delay. *Int J Hyg Environ Health.* 213: 116–123.

Li X, Lee S, Wong S, Shi W, Thornton I. 2004. The study of metal contamination in urban soils of Hong Kong using a GIS-based approach. *Environ Pollut.* 129: 113–124.

Lunn DJ, Thomas A, Best N, Spiegelhalter D. 2000. WinBUGS — a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*. 10:325–337.

McCullagh P. 1980. Regression Models for Ordinal Data. *J R Statist Soc.* B, 42(2): 109-142.

McDermott S, Wu J, Cai B, Lawson A, Aelion CM. 2011. Probability of intellectual disability is associated with soil concentrations of arsenic and lead. *Chemosphere*. 84, 31-38.

Miodovnik A. 2011. Environmental neurotoxicants and developing brain. *Mount Sinai J Med.* 78(1): 58–77.

R Development Core Team. 2007. *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. http:// www.R-project.org.

Ridout M, Hinde J, Demétrio CGB. 2001. A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. *Biometrics*. 57(1): 219–223.

Sullivan JB, Krieger GR. 2001. *Clinical Environmental Health and Toxic Exposures.* Second ed. Philadelphia: Lippincott Williams & Wilkins.

Wang SX, Wang ZH, Cheng XT, Li J, Sang ZP, Zhang XD, Han LL, Qiao XY, Wu ZM, Wang ZQ. 2007. Arsenic and fluoride exposure in drinking water: children's IQ and growth in Shanyin county, Shanxi province, China. Environ *Health Perspect,* 115(4): 643–647.

Wasserman GA, Liu X, Lolacon NJ, Factor-Litvak P, Kline JK, Popovac D, Morina N, Musabegovic A, Vrenezi N, Capuni-Paracka S, et al. 1997. Lead exposure and intelligence in 7-year-old children: the Yugoslavia Prospective Study. *Environ Health Perspect.* 105: 956–962.

Wasserman GA, Graziano JH, Factor-Litvak P, Popovac D, Morina N, Musabegovic A, Vrenezi N, Capuni-Paracka S, Lekvic V, Preteni-Redjepi E, et al. 1994. Consequences of lead exposure and iron supplementation on childhood development at age 4 years. *Neurotoxicol Teratol.,* 16: 233–240.

Yau KKW, Wang K, Lee AH. 2003. Zero-Inflated Negative Binomial Mixed Regression Modeling of Over-Dispersed Count Data with Extra Zeros. *Biometrical J.* 45(4): 437–452.