The Korean Journal of Applied Statistics (2018) DOI: https://doi.org/10.5351/KJAS.2018.31.2.287
31(2), 287–301

A Bayesian zero-inflated Poisson regression model with random effects with application to smoking behavior

Yeon Kyoung Kim[a] · Beom Seuk Hwang[a,1]

[a]Department of Applied Statistics, Chung-Ang University

(Received February 13, 2018; Revised February 16, 2018; Accepted February 17, 2018)

Abstract

It is common to encounter count data with excess zeros in various research fields such as the social sciences, natural sciences, medical science or engineering. Such count data have been explained mainly by zero-inflated Poisson model and extended models. Zero-inflated count data are also often correlated or clustered, in which random effects should be taken into account in the model. Frequentist approaches have been commonly used to fit such data. However, a Bayesian approach has advantages of prior information, avoidance of asymptotic approximations and practical estimation of the functions of parameters. We consider a Bayesian zero-inflated Poisson regression model with random effects for correlated zero-inflated count data. We conducted simulation studies to check the performance of the proposed model. We also applied the proposed model to smoking behavior data from the Regional Health Survey (2015) of the Korea Centers for disease control and prevention.

Keywords: Markov chain Monte Carlo, Metropolis algorithm, random effect, smoking behavior, zero-inflated count data

1. 서론

과도하게 많은 0의 값을 가지는 셀 수 있는 이산형 자료(zero-inflated count data)는 사회과학, 자연과학, 의학, 공학 등 여러 다양한 분야에서 흔히 사용되고 있다. 예를 들어, 어떤 제품의 불량품 개수, 자동차 보험금 청구 횟수, 1년 동안 병원 응급실 이용 횟수 등의 자료에서는 0의 값의 비율이 상당히 높게 나타난다. 일반적으로 셀 수 있는 이산형 자료에 대한 대표적인 모형인 포아송 모형에서는 평균과 분산이 동일하다는 가정이 성립해야 하지만, 0이 과도하게 많은 경우에는 평균에 비해 분산이 과도하게 커지는 현상이 나타나게 된다. 이와 같이 0의 값이 과도하게 많이 관측되는 자료를 분석하기 위해서 Cohen (1963)은 영과잉 포아송 모형(zero-inflated Poisson; ZIP)을 제안하였다. ZIP 모형은 0에 대한 점확률(point mass)과 기존의 포아송 분포를 혼합하여 영과잉 부분과 0이 아닌 부분을 설명하는 모형이다. Lambert (1992)는 공변량(covariates)을 모형에 도입하여 영과잉 포아송 회귀모형(ZIP regression; ZIPR)을 제안하였고, Greene (1994)은 음이항 분포(negative binomial)를 고려하여 영과잉 음이항 모형(zero-inflated negative binomial; ZINB)으로 확장하였다. 그 이후로 영과잉 포

This research was supported by the Chung-Ang University Research Scholarship Grants in 2016.
[1]Corresponding author: Department of Applied Statistics, Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul 06974, Korea. E-mail: bshwang@cau.ac.kr

아송 모형과 관련된 여러가지 방법론이 제안되고 다양한 분야에서 응용되어 왔다. Li 등 (1999)은 전자장치(electronic equipment)의 불량품을 찾아내는데 다변량 영과잉 포아송 모형을 제안하였고, Ridout 등 (2001)은 ZIPR 모형에 대한 score test 방법을 제안하였다.

이러한 모형의 분석방법으로 최대가능도 방법(maximum likelihood method)에 기반을 둔 빈도론자(frequentist)들의 접근 방법이 널리 사용되어 왔다. 하지만, 표본의 크기가 크지 않고, 영과잉 자료와 같이 매우 치우진 분포를 갖는 경우 최대가능도추정량(maximum likelihood estimator; MLE)은 점근적 정규성(asymptotic normality)이 성립하지 않는 단점을 가지고 있다. 이를 보안하기 위해 최근에는 베이지안 추론 방법이 많이 사용되어 왔다 (Angers와 Biswas, 2003; Rodrigues, 2003; Ghosh 등, 2006; Oh와 Lim, 2006; Jang 등, 2008; Shim 등, 2011; Lee 등, 2011a; Liu와 Powers, 2012). 이러한 베이지안 접근법들은 기본적으로 마코프체인 몬테카를로(Markov chain Monte Carlo; MCMC) 방법을 사용하여 모수의 사후분포(posterior distribution)를 찾아내려고 한다. MCMC 방법이 유용하게 널리 사용되고 있지만, 영과잉 이산형 자료같은 복잡한 형태의 혼합 자료의 경우에 잘 적합하지 않는 경우가 발생하기 때문에 이를 해결하기 위해서 다양하게 개선된 MCMC 방법이 개발되고 있다. Ghosh 등 (2006)은 데이터 확대(data augmentation) 방법을 MCMC에 도입했고, Lee 등 (2011a)는 역 베이즈 공식 표집기(inverse Bayes formula (IBF) sampler)를 ZIP 모형에 적용하였다.

영과잉 모형에 대한 또 다른 확장 모형으로서, 반복측정자료 또는 상관관계가 있는 자료에 대한 랜덤효과(random effect)를 도입한 모형들도 개발되어 왔다. 주로 빈도론자들의 접근방법이 제안되었는데, Hall (2000)은 랜덤절편(random intercept)을 포아송 분포 부분에 적용하였고, Yau와 Lee (2001)는 랜덤효과를 포아송과 베르누이 두 부분에 독립적으로 도입하였다. Min과 Agresti (2005)는 포아송과 베르누이 두 부분의 랜덤효과를 상관관계가 있다고 가정하였다. Lee 등 (2011b)은 주변 랜덤효과(marginalized random effects) 모형을 도입하여 영과잉 반복측정자료를 분석하였다. 또한, Neelon 등 (2010)은 경시적(longitudinal) 영과잉 자료에 대해 랜덤효과를 고려하여 베이지안 분석 방법을 사용하였다. 이처럼 다양한 분야에서 여러가지 방법론으로 영과잉 이산형 자료가 분석이 되고 있지만, 상관관계가 있는 자료, 특히 지역적 의존성(geographical dependency)을 가지는 자료에 대해 랜덤효과를 고려한 영과잉 모형은 그리 많이 발달하지 않아 왔다.

본 논문에서는 반응변수 사이에 상관관계가 존재하는 영과잉 이산형의 형태를 띠는 자료에 대해서 랜덤효과를 포함한 영과잉 포아송 모형을 고려하여 베이지안 접근방법을 사용하여 추론을 하려고 한다. 제시된 방법으로 전국의 20대 남자의 흡연 자료에 적용해본다. 구체적으로 2장에서는 영과잉 자료에 대한 모형을 차례로 설명하고, 3장에서는 베이지안 추론방법을 간략히 설명한다. 사전분포(prior distribution)와 사후분포(posterior distribution)를 계산하고, 이를 바탕으로 MCMC 방법의 단계를 설명한다. 모형 선택을 위해서 사용할 deviance information criterion (DIC)의 개념에 대해서 간략히 설명한다. 4장에서는 모의 실험을 통해 랜덤효과가 포함된 영과잉 모형의 적합성에 대해 살펴보고, 5장에서는 실제 흡연 데이터를 가지고 분석을 시도한다. 끝으로 5장에서는 본 논문을 요약 정리하고 향후 연구의 방향에 대해 논의한다.

2. 영과잉 자료에 대한 모형

2.1. 영과잉 포아송 모형

반응변수 Y_i $(i = 1, \ldots, n)$가 음이 아닌 정수값을 갖는 영과잉 이산형 자료의 형태를 띨 때, Y_i의 확률분포는 0의 값이 발생하는 점확률분포와 0보다 큰 정수값을 갖는 포아송 분포의 혼합분포 구조를 가지

고 있다. 이를 표현하기 위해 지시변수 Z를 다음과 같이 정의한다.

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

이때, Y는 다음과 같은 포아송-베르누이 혼합 구조를 갖는다.

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

그러므로, 반응변수 Y_i $(i = 1, \ldots, n)$가 영과잉 포아송 모형을 따를 때, $Y_i \sim \text{ZIP}(\lambda, p)$, 다음과 같은 확률질량함수(probability mass function)를 가진다.

$$P(Y_i = y_i|\lambda, p) = \begin{cases} p + (1-p)e^{-\lambda}, & \text{if } \ y_i = 0, \\ (1-p)\dfrac{e^{-\lambda}\lambda^{y_i}}{y_i!}, & \text{if } \ y_i > 0. \end{cases}$$

이 모형에서 0의 값은 두 개의 분포에서 각각 발생하고 있음을 알 수 있다. 즉, 영과잉 상태에서 발생하는 경우와 포아송 분포를 통해서 발생하는 경우로 나뉘어진다. 이와 같은 모형의 정의를 바탕으로 $\boldsymbol{Y} = (y_1, \ldots, y_n)$가 주어질 때, 가능도함수(likelihood function)는 다음과 같이 구해진다.

$$L(\lambda, p|\boldsymbol{y}, \boldsymbol{z}) = \prod_{i=1}^{n} \left[p + (1-p)e^{-\lambda}\right]^{z_i} \left[(1-p)\frac{e^{-\lambda}\lambda^{y_i}}{y_i!}\right]^{1-z_i}. \tag{2.1}$$

2.2. 영과잉 포아송 회귀모형

위에서 정의한 영과잉 포아송 모형에 공변량을 고려하게 되면 영과잉 포아송 회귀모형이 정의된다. 일반적으로 공변량은 베르누이 분포의 매개변수 p와 로짓 연결함수(logit link function)를 통해 연결이 되고, 포아송 분포의 매개변수 λ와 로그 연결함수(log link function)를 통해 연결이 된다. 구체적으로 다음과 같은 영과잉 포아송 회귀모형을 정의할 수 있다.

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha_0 + \alpha_1 X_{11i} + \alpha_2 X_{12i} + \cdots + \alpha_m X_{1mi} = \boldsymbol{X_{1i}\alpha}, \tag{2.2}$$

$$\log(\lambda_i) = \beta_0 + \beta_1 X_{21i} + \beta_2 X_{22i} + \cdots + \beta_l X_{2li} = \boldsymbol{X_{2i}\beta}, \tag{2.3}$$

여기서 $\boldsymbol{X_{1i}} = (1, X_{11i}, X_{12i}, \ldots, X_{1mi})$는 제로 단계에서의 m개의 공변량으로 이루어진 벡터이고, $\boldsymbol{X_{2i}} = (1, X_{21i}, X_{22i}, \ldots, X_{2li})$는 포아송 분포 단계에서의 l개의 공변량으로 이루어진 벡터이다. 또한, $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_m)^T$와 $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_l)^T$는 각각 거기에 상응하는 계수 벡터이다. 영과잉 포아송 회귀모형에서 가능도함수는 다음과 같이 구해진다.

$$\begin{aligned} L(\boldsymbol{\alpha}, \boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{z}) = \prod_{i=1}^{n} &\left[\frac{e^{\boldsymbol{X_{1i}\alpha}}}{1+e^{\boldsymbol{X_{1i}\alpha}}} + \left(1 - \frac{e^{\boldsymbol{X_{1i}\alpha}}}{1+e^{\boldsymbol{X_{1i}\alpha}}}\right) e^{-e^{\boldsymbol{X_{2i}\beta}}}\right]^{z_i} \\ &\times \left[\left(1 - \frac{e^{\boldsymbol{X_{1i}\alpha}}}{1+e^{\boldsymbol{X_{1i}\alpha}}}\right) e^{-e^{\boldsymbol{X_{2i}\beta}}} \frac{\left(e^{\boldsymbol{X_{2i}\beta}}\right)^{y_i}}{y_i!}\right]^{1-z_i}. \end{aligned} \tag{2.4}$$

2.3. 랜덤효과를 포함한 영과잉 포아송 회귀모형

제 2.2장에서 정의된 영과잉 포아송 회귀모형에서는 관찰치들이 서로 독립이라는 가정을 하고 있다. 하지만, 반응변수들이 서로 상관관계가 있거나 분산의 요인들이 다양하게 나타날 때, 랜덤효과를 고려한 회귀모형을 생각해볼 수 있다. 여기서는 Hall (2000)의 제안에 따라, 랜덤 절편이 포아송 모형 부분에만 포함된 모형을 살펴보려고 한다. 하지만, 향후에 포아송 모형과 베르누이 모형 두 부분에 랜덤효과를 도입한 보다 일반적인 모형도 생각해 볼 수 있을 것이다 (Yau와 Lee, 2001; Min과 Agresti, 2005; Lee 등, 2011b). i번째 그룹의 j번째 개체의 반응변수를 Y_{ij} $(i = 1, \ldots, I,\ j = 1, \ldots, J)$라고 정의하면 식 (2.2)와 (2.3)은 다음과 같이 변형된다.

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha_0 + \alpha_1 X_{11ij} + \alpha_2 X_{12ij} + \cdots + \alpha_m X_{1mij} = \boldsymbol{X_{1ij}\alpha}, \tag{2.5}$$

$$\log(\lambda_{ij}) = \beta_0 + \beta_1 X_{21ij} + \beta_2 X_{22ij} + \cdots + \beta_l X_{2lij} + \gamma b_i = \boldsymbol{X_{2ij}\beta} + \gamma b_i, \tag{2.6}$$

여기서 $\boldsymbol{X_{1ij}} = (1, X_{11ij}, X_{12ij}, \ldots, X_{1mij})$는 제로 단계에서의 m개의 공변량으로 이루어진 벡터이고, $\boldsymbol{X_{2ij}} = (1, X_{21ij}, X_{22ij}, \ldots, X_{2lij})$는 포아송 분포 단계에서의 l개의 공변량으로 이루어진 벡터이다. 또한, $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_m)^T$와 $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_l)^T$는 각각 거기에 상응하는 계수 벡터이다. $b_1, \ldots, b_I$는 독립적인 표준정규분포를 따르는 랜덤효과이고 그에 상응하는 계수는 γ라고 가정한다. 여기서 γ는 그룹 간의 변동을 의미한다. 이때, 랜덤효과를 포함한 영과잉 포아송 회귀모형의 완전데이터 가능도함수(complete data likelihood function)는 다음과 같다.

$$\begin{aligned} L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma | \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{b}) = \prod_{i=1}^{I}\prod_{j=1}^{J} & \left[\frac{e^{\boldsymbol{X_{1ij}\alpha}}}{1+e^{\boldsymbol{X_{1ij}\alpha}}} + \left(1 - \frac{e^{\boldsymbol{X_{1ij}\alpha}}}{1+e^{\boldsymbol{X_{1ij}\alpha}}}\right) e^{-e^{\boldsymbol{X_{2ij}\beta}+\gamma b_i}}\right]^{z_{ij}} \\ & \times \left[\left(1 - \frac{e^{\boldsymbol{X_{1ij}\alpha}}}{1+e^{\boldsymbol{X_{1ij}\alpha}}}\right) e^{-e^{\boldsymbol{X_{2ij}\beta}+\gamma b_i}} \frac{\left(e^{\boldsymbol{X_{2ij}\beta}+\gamma b_i}\right)^{y_{ij}}}{y_{ij}!}\right]^{1-z_{ij}} \\ & \times \prod_{i=1}^{I} f(b_i), \end{aligned} \tag{2.7}$$

여기서 $f(b_i)$는 표준정규분포의 확률밀도함수를 나타낸다.

3. 베이지안 추론

3.1. 사전분포와 사후분포

일반적인 베이지안 분석 방법을 따라서 미지의 모수들에 대해 사전분포를 선택한 후 결합사후분포(joint posterior distribution)를 유도하려고 한다. 랜덤효과를 포함한 ZIP 회귀모형의 세 모수 $\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma$에 대해 각각 다음과 같이 독립적인 사전분포를 고려할 수 있다.

$$\begin{aligned} \boldsymbol{\alpha} &\sim \text{MVN}\left(\boldsymbol{\mu_\alpha}, \boldsymbol{\Sigma_\alpha}\right), \\ \boldsymbol{\beta} &\sim \text{MVN}\left(\boldsymbol{\mu_\beta}, \boldsymbol{\Sigma_\beta}\right), \\ \gamma &\sim N(\mu_\gamma, \sigma_\gamma^2), \end{aligned} \tag{3.1}$$

여기서 $\text{MVN}(\boldsymbol{\mu}, \Sigma)$은 평균 $\boldsymbol{\mu}$와 공분산행렬 Σ를 가지는 다변량정규분포를 나타낸다. 사전분포의 평균인 $\boldsymbol{\mu_\alpha}$와 $\boldsymbol{\mu_\beta}$의 차원은 각각 $m \times 1$과 $l \times 1$이고, 공분산인 $\boldsymbol{\Sigma_\alpha}$와 $\boldsymbol{\Sigma_\beta}$의 차원은 각각 $m \times m$과 $l \times l$이다.

식 (2.7)에 정의된 완전데이터 가능도함수와 식 (3.1)의 사전분포를 이용하여 다음과 같이 결합사후분포를 구한다.

$$\begin{aligned}
p(\boldsymbol{\alpha},\boldsymbol{\beta},\gamma|\boldsymbol{y},\boldsymbol{z},\boldsymbol{b}) &\propto p(\boldsymbol{y},\boldsymbol{z},\boldsymbol{b}|\boldsymbol{\alpha},\boldsymbol{\beta},\gamma)p(\boldsymbol{\alpha})p(\boldsymbol{\beta})p(\gamma)\\
&\propto \prod_{i=1}^{I}\prod_{j=1}^{J}\left[\frac{e^{\boldsymbol{X_{1ij}\alpha}}}{1+e^{\boldsymbol{X_{1ij}\alpha}}}+\left(1-\frac{e^{\boldsymbol{X_{1ij}\alpha}}}{1+e^{\boldsymbol{X_{1ij}\alpha}}}\right)e^{-e^{\boldsymbol{X_{2ij}\beta}+\gamma b_i}}\right]^{z_{ij}}\\
&\quad\times\left[\left(1-\frac{e^{\boldsymbol{X_{1ij}\alpha}}}{1+e^{\boldsymbol{X_{1ij}\alpha}}}\right)e^{-e^{\boldsymbol{X_{2ij}\beta}+\gamma b_i}}\frac{\left(e^{\boldsymbol{X_{2ij}\beta}+\gamma b_i}\right)^{y_{ij}}}{y_{ij}!}\right]^{1-z_{ij}}\\
&\quad\times\exp\left[-\frac{1}{2}\sum_{i=1}^{I}b_i^2\right]\times|\Sigma_\alpha|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(\boldsymbol{\alpha}-\boldsymbol{\mu_\alpha})^T\Sigma_\alpha^{-1}(\boldsymbol{\alpha}-\boldsymbol{\mu_\alpha})\right]\\
&\quad\times|\Sigma_\beta|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu_\beta})^T\Sigma_\beta^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu_\beta})\right]\\
&\quad\times\left(\sigma_\gamma^2\right)^{-\frac{1}{2}}\exp\left[-\frac{1}{2\sigma_\gamma^2}(\gamma-\mu_\gamma)^2\right]. \qquad (3.2)
\end{aligned}$$

3.2. Markov chain Monte Carlo

MCMC 알고리즘을 사용하여 미지의 모수를 추론하기 위해 식 (3.2)에서 구한 결합사후분포로부터 각 모수와 랜덤효과에 대한 조건부 사후분포를 다음과 같이 계산한다.

$$\begin{aligned}
p(\boldsymbol{\alpha}|\boldsymbol{y},\boldsymbol{z},\boldsymbol{b},\boldsymbol{\beta},\gamma) &\propto \prod_{i=1}^{I}\prod_{j=1}^{J}\left[\frac{e^{\boldsymbol{X_{1ij}\alpha}}}{1+e^{\boldsymbol{X_{1ij}\alpha}}}+\left(1-\frac{e^{\boldsymbol{X_{1ij}\alpha}}}{1+e^{\boldsymbol{X_{1ij}\alpha}}}\right)e^{-e^{\boldsymbol{X_{2ij}\beta}+\gamma b_i}}\right]^{z_{ij}}\\
&\quad\times\left[\left(1-\frac{e^{\boldsymbol{X_{1ij}\alpha}}}{1+e^{\boldsymbol{X_{1ij}\alpha}}}\right)e^{-e^{\boldsymbol{X_{2ij}\beta}+\gamma b_i}}\frac{\left(e^{\boldsymbol{X_{2ij}\beta}+\gamma b_i}\right)^{y_{ij}}}{y_{ij}!}\right]^{1-z_{ij}}\\
&\quad\times\exp\left[-\frac{1}{2}(\boldsymbol{\alpha}-\boldsymbol{\mu_\alpha})^T\Sigma_\alpha^{-1}(\boldsymbol{\alpha}-\boldsymbol{\mu_\alpha})\right], \qquad (3.3)\\
p(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{z},\boldsymbol{b},\boldsymbol{\alpha},\gamma) &\propto \prod_{i=1}^{I}\prod_{j=1}^{J}\left[\frac{e^{\boldsymbol{X_{1ij}\alpha}}}{1+e^{\boldsymbol{X_{1ij}\alpha}}}+\left(1-\frac{e^{\boldsymbol{X_{1ij}\alpha}}}{1+e^{\boldsymbol{X_{1ij}\alpha}}}\right)e^{-e^{\boldsymbol{X_{2ij}\beta}+\gamma b_i}}\right]^{z_{ij}}\\
&\quad\times\left[\left(1-\frac{e^{\boldsymbol{X_{1ij}\alpha}}}{1+e^{\boldsymbol{X_{1ij}\alpha}}}\right)e^{-e^{\boldsymbol{X_{2ij}\beta}+\gamma b_i}}\frac{\left(e^{\boldsymbol{X_{2ij}\beta}+\gamma b_i}\right)^{y_{ij}}}{y_{ij}!}\right]^{1-z_{ij}}\\
&\quad\times\exp\left[-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu_\beta})^T\Sigma_\beta^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu_\beta})\right], \qquad (3.4)\\
p(\gamma|\boldsymbol{y},\boldsymbol{z},\boldsymbol{b},\boldsymbol{\alpha},\boldsymbol{\beta}) &\propto \prod_{i=1}^{I}\prod_{j=1}^{J}\left[\frac{e^{\boldsymbol{X_{1ij}\alpha}}}{1+e^{\boldsymbol{X_{1ij}\alpha}}}+\left(1-\frac{e^{\boldsymbol{X_{1ij}\alpha}}}{1+e^{\boldsymbol{X_{1ij}\alpha}}}\right)e^{-e^{\boldsymbol{X_{2ij}\beta}+\gamma b_i}}\right]^{z_{ij}}\\
&\quad\times\left[\left(1-\frac{e^{\boldsymbol{X_{1ij}\alpha}}}{1+e^{\boldsymbol{X_{1ij}\alpha}}}\right)e^{-e^{\boldsymbol{X_{2ij}\beta}+\gamma b_i}}\frac{\left(e^{\boldsymbol{X_{2ij}\beta}+\gamma b_i}\right)^{y_{ij}}}{y_{ij}!}\right]^{1-z_{ij}}\\
&\quad\times\exp\left[-\frac{1}{2\sigma_\gamma^2}(\gamma-\mu_\gamma)^2\right], \qquad (3.5)
\end{aligned}$$

$$p(b_i|b_{(-i)},\boldsymbol{y},\boldsymbol{z},\boldsymbol{\alpha},\boldsymbol{\beta},\gamma) \propto \prod_{j=1}^{J}\left[\frac{e^{\boldsymbol{X_{1ij}\alpha}}}{1+e^{\boldsymbol{X_{1ij}\alpha}}}+\left(1-\frac{e^{\boldsymbol{X_{1ij}\alpha}}}{1+e^{\boldsymbol{X_{1ij}\alpha}}}\right)e^{-e^{\boldsymbol{X_{2ij}\beta}+\gamma b_i}}\right]^{z_{ij}}$$
$$\times\left[\left(1-\frac{e^{\boldsymbol{X_{1ij}\alpha}}}{1+e^{\boldsymbol{X_{1ij}\alpha}}}\right)e^{-e^{\boldsymbol{X_{2ij}\beta}+\gamma b_i}}\frac{\left(e^{\boldsymbol{X_{2ij}\beta}+\gamma b_i}\right)^{y_{ij}}}{y_{ij}!}\right]^{1-z_{ij}}$$
$$\times\exp\left[-\frac{1}{2}b_i^2\right]. \tag{3.6}$$

식 (3.3)–(3.6)에 나타난 각 모수의 조건부 사후분포는 일반적으로 알려진 표준적인 분포의 형태를 띠고 있지 않다. 따라서 모수를 추정하기 위해 MCMC 알고리즘 중에서도 메트로폴리스-해스팅스(Metropolis-Hastings; MH) 알고리즘을 사용하고자 한다. 다음과 같은 단계를 거쳐 MCMC 알고리즘은 진행된다.

Step 1: 모수의 초기값 $(\boldsymbol{\alpha}^{(0)},\boldsymbol{\beta}^{(0)},\gamma^{(0)})$을 설정하고, 랜덤효과 $b_i\,(i=1,\ldots,I)$를 표준정규분포에서 생성한다.

Step 2: t 시점의 값이$(\boldsymbol{\alpha}^{(t)},\boldsymbol{\beta}^{(t)},\gamma^{(t)})$로 주어졌을 때, $t+1$ 시점의 값을 식 (3.3)–(3.5)를 이용하여 다음과 같이 순차적으로 업데이트한다.

- 식 (3.3)을 바탕으로 MH 방법을 사용하여 $\boldsymbol{\alpha}$를 샘플링한다: $p(\boldsymbol{\alpha}^{(t+1)}|\boldsymbol{y},\boldsymbol{z},\boldsymbol{b},\boldsymbol{\beta}^{(t)},\gamma^{(t)})$
- 식 (3.4)를 바탕으로 MH 방법을 사용하여 $\boldsymbol{\beta}$를 샘플링한다: $p(\boldsymbol{\beta}^{(t+1)}|\boldsymbol{y},\boldsymbol{z},\boldsymbol{b},\boldsymbol{\alpha}^{(t)},\gamma^{(t)})$
- 식 (3.5)를 바탕으로 MH 방법을 사용하여 γ를 샘플링한다: $p(\gamma^{(t+1)}|\boldsymbol{y},\boldsymbol{z},\boldsymbol{b},\boldsymbol{\alpha}^{(t)},\boldsymbol{\beta}^{(t)})$

Step 3: 랜덤효과 $b_i\,(i=1,\ldots,I)$를 식 (3.6)을 바탕으로 MH 방법을 사용하여 업데이트한다.

Step 4: Step 2로 돌아가서 수렴할 때까지 반복한다.

여기서는 추정값의 수렴을 개선하기 위하여 구체적으로 분산조정 메트로폴리스(adaptive Metropolis) 알고리즘 (Haario 등, 2005)을 사용하였다. 즉, 각 단계에서 제안분포(proposal distribution)로서 정규분포를 사용할 때, 현재 모수의 값을 평균으로 사용하고 분산에 대해서는 경험적(empirical) 분산과 조정계수를 사용하여 매 단계마다 제안분포를 조정하였다. 예를 들어, $t+1$ 시점의 모수 $\theta^{(t+1)}$을 업데이트하기 위해서 다음과 같은 후보값(candidate value) θ^*를 사용한다.

$$\theta^* \sim N\left(\theta^{(t)},V^{(t)}\right)$$
$$V^{(t)}=\begin{cases}V^{(0)}, & \text{if } t\le t_0,\\ s\mathrm{Var}\left(\theta^{(0)},\ldots,\theta^{(t-1)}\right)+s\epsilon, & \text{if } t>t_0,\end{cases}$$

여기서 $V^{(0)}$는 모수 θ에 대한 제안분포의 초기 분산값이고, s는 후보값의 채택 비율(acceptance rate)을 최적값인 0.44로 유지하기 위해 곱해주는 조정 계수(d-차원에 대해 $2.4/\sqrt{d}$)이다 (Gelman 등, 2014). 또한, ϵ은 분산이 0이 되는 것을 막기 위한 아주 작은 상수값을 나타낸다. 알고리즘의 수렴 여부를 확인하기 위해서 골고루 퍼져있는 다양한 값들을 초기값으로 설정해서 MCMC를 실행한 후에 일반적인 베이지안 진단법(Bayesian diagnostics)을 사용한다. 여기서는 trace plot과 더불어 Gelman과 Rubin의 potential scale reduction factor, $\hat{R}$을 사용한다 (Gelman 등, 2014). $\hat{R}$은 마코프 체인 내 변동(within-chain variation)과 마코프 체인 간 변동(between-chain variation)을 비교한 측정치로서 그 값이 1.2보다 작을 때, 수렴이 잘 되고 있음을 나타내준다.

Table 4.1. Scenarios of simulation studies

Scenario	Number of groups	Number of subjects	Total N	True Values					Mean of zero proportion
				α_0	α_1	β_0	β_1	γ	
I	5	40	200	−2.0	0.5	0.3	0.7	0.6	0.401
II	5	40	200	−1.0	0.5	0.3	0.5	0.6	0.496
III	5	40	200	0.7	1.2	0.3	0.7	0.6	0.782
IV	10	100	1000	−2.0	0.5	0.3	0.7	0.6	0.405
V	10	100	1000	−1.0	0.5	0.3	0.5	0.6	0.503
VI	10	100	1000	0.7	1.2	0.3	0.7	0.6	0.782

3.3. Deviance information criterion

베이지안 분석에서 모형을 비교할 때 흔히 쓰이는 통계량으로 DIC가 있다 (Spiegelhalter 등, 2002). DIC는 다음과 같이 정의된다.

$$\mathrm{DIC} = \overline{D(\theta)} + p_D,$$

여기서 편차(deviance)는 $D(\theta) = -2\log f(y|\theta) + 2\log h(y)$로 정의되고, $\overline{D(\theta)} = E_\theta[-2\log f(y|\theta)|y] + 2\log h(y)$는 편차의 사후평균을 나타낸다. 그리고 $p_D = \overline{D(\theta)} - D(\hat{\theta}) = E[D(\theta)|y] - D(E[\theta|y])$는 모형에서 사용된 모수의 수(effective dimension)를 나타낸다. DIC는 모형의 상대적인 적합성에 대한 측정치인 $\overline{D(\theta)}$와 모형의 복잡한 정도에 대한 페널티를 나타내는 p_D로 구성된다. 따라서 더 작은 DIC를 가지는 모형이 상대적으로 데이터를 더 잘 적합하고 있다고 결론내릴 수 있다.

하지만, 랜덤효과가 포함된 모형에서는 모수 θ가 항상 식별 가능하지 않을 수가 있고, 그때의 사후평균 $\hat{\theta} = E_\theta[\theta|y]$은 좋지 않은 추정치가 되어 버린다. 따라서 Celeux 등 (2006)은 랜덤효과나 잠재변수가 포함된 모형에 대해서 사용할 수 있는 수정된 DIC를 다음과 같이 제안하였다.

$$\mathrm{DIC}_4 = -4E_{\theta,Z}\left[\log f(y,Z|\theta)|y\right] + 2E_Z\left[\log f(y,Z|E_\theta[\theta|y,Z])|y\right], \tag{3.7}$$

여기서 Z는 직접 관찰되지 않는 랜덤효과 또는 잠재변수를 나타내고, 수정된 DIC(DIC_4)는 Z를 포함한 완전데이터 가능도함수를 이용해서 계산한다. 본 논문에서는 모형의 비교를 위해 식 (3.7)에서 정의된 DIC(DIC_4)를 사용한다.

4. 모의 실험

4.1. 모의 실험의 구성

반응변수들이 서로 상관관계가 존재하는 영과잉 이산형 데이터에 대해서 본 논문에서 제안된 랜덤효과를 포함한 ZIP 회귀모형의 적합성을 검증하기 위해 각기 다른 6개의 시나리오로 구성된 모의 실험을 시행하였다. 각 시나리오들은 분석의 간편성을 위해 하나의 공변량을 베르누이 분포 부분과 포아송 분포 부분에 공통적으로 사용한다고 가정한다. 즉, 식 (2.5)와 (2.6)에서 $m = l = 1$이고, $X_{11} = X_{21}$이다. 이때, 공변량 X는 표준정규분포에서 임의로 생성하였다. 영과잉 자료의 비율을 조정하기 위해 $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma)$의 참값을 다양하게 변화시켜서 베르누이 분포의 모수 p와 포아송 분포의 모수 λ의 값을 결정하였다. 또한, 표본크기가 미치는 영향을 보기 위해 전체 개체수를 200과 1,000으로 나누어서 시나리오를 구성하였다. 총 6개의 모의 실험 시나리오는 Table 4.1과 같고, 이를 바탕으로 각 시나리오별로 데이터셋을 100개씩 생성하였다. 총 6개의 모의 실험 시나리오에 랜덤효과를 포함한 ZIP 회귀모형과 랜

Table 4.2. Results of simulation studies

Scenario	Para meter	Truth	Model with random effects				Model without random effects			
			Estimate(CI)	CP	RMSE	DIC	Estimate(CI)	CP	RMSE	DIC
I	α_0	−2.0	−1.71(−2.53, −1.05)	0.90	0.437		−1.20(−1.85, −0.68)	0.43	0.903	
	α_1	0.5	0.26(−0.36, 0.93)	0.92	0.378		0.07(−0.43, 0.63)	0.65	0.522	
	β_0	0.3	0.30(−0.38, 0.86)	0.95	0.259	546.4	0.61(0.45, 0.76)	0.30	0.418	679.7
	β_1	0.7	0.68(0.54, 0.81)	0.95	0.076		0.65(0.52, 0.77)	0.73	0.118	
	γ	0.6	0.61(0.28, 1.15)	0.83	0.317		-	-	-	
II	α_0	−1.0	−0.99(−1.66, −0.45)	0.98	0.261		−0.63(−1.17, −0.19)	0.62	0.480	
	α_1	0.5	0.45(−0.07, 1.03)	0.98	0.231		0.30(−0.15, 0.79)	0.81	0.307	
	β_0	0.3	0.27(−0.41, 0.83)	0.96	0.273	510.8	0.57(0.39, 0.74)	0.33	0.425	598.5
	β_1	0.5	0.48(0.32, 0.65)	0.94	0.086		0.45(0.29, 0.61)	0.80	0.118	
	γ	0.6	0.59(0.19, 1.26)	0.82	0.363		-	-	-	
III	α_0	0.7	0.63(0.06, 1.15)	0.93	0.290		0.82(0.30, 1.29)	0.87	0.315	
	α_1	1.2	1.21(0.58, 1.96)	0.97	0.328		1.10(0.51, 1.79)	0.92	0.357	
	β_0	0.3	0.23(−0.56, 0.88)	0.95	0.306	489.4	0.51(0.22, 0.78)	0.49	0.423	522.6
	β_1	0.7	0.68(0.35, 1.00)	0.96	0.166		0.65(0.34, 0.95)	0.85	0.195	
	γ	0.6	0.48(−0.18, 1.27)	0.84	0.511		-	-	-	
IV	α_0	−2.0	−1.92(−2.40, −1.52)	0.94	0.222		−1.15(−1.44, −0.90)	0.06	0.891	
	α_1	0.5	0.45(0.12, 0.81)	0.95	0.176		0.09(−0.15, 0.34)	0.27	0.452	
	β_0	0.3	0.26(−0.10, 0.61)	0.92	0.207	2806.5	0.59(0.52, 0.66)	0.14	0.346	3356.9
	β_1	0.7	0.70(0.64, 0.76)	0.95	0.029		0.64(0.59, 0.70)	0.52	0.077	
	γ	0.6	0.61(0.44, 0.86)	0.80	0.171		-	-	-	
V	α_0	−1	−1.00(−1.31, −0.74)	0.92	0.154		−0.52(−0.73, −0.33)	0.16	0.515	
	α_1	0.5	0.49(0.25, 0.76)	0.94	0.134		0.27(0.08, 0.48)	0.37	0.269	
	β_0	0.3	0.27(−0.11, 0.62)	0.92	0.200	2626.5	0.60(0.52, 0.68)	0.13	0.370	3011.4
	β_1	0.5	0.49(0.42, 0.56)	0.99	0.033		0.45(0.38, 0.52)	0.66	0.071	
	γ	0.6	0.62(0.42, 0.90)	0.83	0.172		-	-	-	
VI	α_0	0.7	0.69(0.45, 0.92)	0.96	0.122		0.93(0.73, 1.13)	0.42	0.275	
	α_1	1.2	1.20(0.90, 1.52)	0.96	0.148		1.01(0.75, 1.30)	0.72	0.247	
	β_0	0.3	0.28(−0.16, 0.68)	0.97	0.189	2492.5	0.61(0.49, 0.72)	0.20	0.373	2632.0
	β_1	0.7	0.70(0.57, 0.82)	0.98	0.058		0.64(0.52, 0.76)	0.79	0.102	
	γ	0.6	0.61(0.37, 0.96)	0.84	0.232		-	-	-	

CI = credible interval; CP = coverage probability; RMSE = root mean squared error; DIC = deviance information criterion.

덤효과를 고려하지 않은 ZIP 회귀모형을 각각 적합시켜 그 결과를 비교하였다. 각 모수의 사후추정을 위해 사전분포의 영향을 최소로 한 무정보적인(noninformative) 사전분포를 다음과 같이 고려하였다.

$$\boldsymbol{\alpha} \sim \mathrm{MVN}(\mathbf{0}, 100I_2), \qquad \boldsymbol{\beta} \sim \mathrm{MVN}(\mathbf{0}, 100I_2), \quad \gamma \sim N(0, 100)$$

모수 추정방법으로는 3.2장에서 소개한 MCMC 알고리즘을 사용하였는데, 50,000번의 반복시행과 25,000번의 제거(burn-in)를 통해 얻은 표본을 바탕으로 추정값을 계산하였다.

4.2. 모의 실험의 결과

모의 실험을 통해서 두 모형을 평가하고 비교하기 위해 먼저 DIC를 계산하였다. 여기서는 100개의 데이터셋에 대해 DIC의 평균값을 제시하였다. 또한, 보다 구체적인 모수들의 비교를 위해 각 모수들

의 사후 평균을 100개의 데이터셋에 대해 평균값을 구한 후 모수의 참값과 비교하였다. 마찬가지로 95% 신용구간(credible interval; CI)의 100개에 대한 평균값을 제시하였다. 여기서 95% 신용구간은 MCMC를 통해 얻은 표본을 분위수(quantile-based) 방법을 사용하여 계산하였다. 또한, 95% 신용구간을 기준으로 한 포함확률(coverage probability; CP)을 계산하고, 제곱근평균제곱오차(root mean squared error; RMSE) RMSE $= \sqrt{E(\hat{\theta}-\theta)^2}$를 통하여 추정치의 편향(bias)과 분산을 측정하였다.

Table 4.2는 모의 실험의 결과를 나타내고 있다. 먼저 두 모형의 적합성을 비교하기 위해 시나리오별로 DIC를 비교해보면, 모든 시나리오에서 랜덤효과를 포함한 ZIP 회귀모형이 랜덤효과를 고려하지 않은 모형보다 훨씬 작은 값들을 가지고 있음을 알 수 있다. 이를 통해 전자의 모형이 데이터를 더 잘 적합하고 있다고 결론내릴 수 있다. 랜덤효과를 포함한 ZIP 회귀모형에서 모수들의 추정치는 대체적으로 참값에 가까운 값을 얻을 수 있었다. 모수들에 대한 포함확률을 살펴보면, 모든 시나리오의 경우에 대해서 γ를 제외한 $\alpha_0, \alpha_1, \beta_0, \beta_1$이 95% 전후의 포함확률을 띠고 있음을 알 수 있다. 이는 표본크기와 영과잉 비율에 상관없이 성립한다. 반면에 랜덤효과를 고려하지 않은 ZIP 회귀모형에서는 모든 모수들에 대해 현저히 낮은 포함확률을 가지고 있음을 알 수 있다. 예를 들어, 표본크기 1,000을 가지는 시나리오 IV의 경우에 α_0의 추정치에 대한 포함확률은 랜덤효과 모형이 94%인데 반해, 랜덤효과 없는 모형은 단지 6%이다. RMSE 값에 대해서도 상당한 차이를 보이고 있다. $\alpha_0, \alpha_1, \beta_0, \beta_1$에 대한 RMSE가 랜덤효과 모형이 랜덤효과 없는 모형보다 더 작은 값을 가지고 있음을 알 수 있다. 또한, 표본크기가 커지면 같은 영과잉 비율 시나리오에 대해 RMSE가 작아지는 것으로 나타났다. 예를 들어, 영과잉 비율이 80% 정도되는 시나리오 III과 VI을 비교해보면, 모형에 상관없이 모든 모수들에 대해 RMSE가 현저히 줄어들고 있음을 확인할 수 있다. 전체적으로 반응변수들 간에 상관관계가 존재하는 영과잉 이산형 데이터에 대해서 랜덤효과를 고려한 모형이 그렇지 않은 모형에 비해서 더 잘 설명해주고 있음을 알 수 있었다.

5. 흡연 자료의 분석

5.1. 자료의 탐색

이 장에서는 2.3장에서 제안된 랜덤효과가 포함된 ZIP 회귀모형에 대한 베이지안 추론방법을 실제 자료에 적용하여 그 결과를 살펴보고자 한다. 본 연구에 사용된 자료는 한국 질병관리본부(Korea Centers for Disease Control and Prevention)에서 실시한 2015년 지역사회 건강조사 자료로서, 본 연구에서는 전국 17개 시·도 20대(만 20–29세) 남자의 흡연량에 대한 데이터를 사용하였다. 조사 응답 중 무응답/모름에 대한 데이터는 분석의 편의상 결측치로 간주하여 제외하였다. 반응변수 Y_{ij}는 i 지역의 j번째 사람의 평생흡연여부에 대한 응답으로 담배 개비로 표현하였다($i = 1, \ldots, 17$, $j = 1, \ldots, n_i$, n_i는 i 지역의 총 응답수). Table 5.1에 나와 있는 것처럼 20대 남자의 흡연량(Y)은 63% 이상이 0의 값을 가지고 있고, 그 0의 비율은 지역에 따라 많게는 0.689에서 적게는 0.548의 값을 가짐으로써 시도별로 상당한 차이가 나고 있음을 알 수 있다. 흡연량을 설명해주는 공변량으로써 신체질량지수(BMI), 음주여부(drink), 신체활동일수(phy.act), 가구소득(fm.income)이 분석에 사용되었다. 신체질량지수는 연속형 변수로서 25 이상일 경우 비만이라고 알려져 있다. 음주여부는 1(예)/2(아니오)를 나타내는 이진수 자료(binary data)이고, 신체활동일수는 지난 1주 동안의 활동일수를 나타낸다. 또한 가구소득은 8개의 카테고리로 나누어진 최근 1년 동안의 가구의 월평균 소득을 나타낸다.

5.2. 베이지안 모형

베이지안 추론을 위해 모의 실험에서와 마찬가지로 다음과 같은 무정보적인 사전분포를 사용하였다.

$$\boldsymbol{\alpha} \sim \mathrm{MVN}(\mathbf{0}, 100 I_5), \qquad \boldsymbol{\beta} \sim \mathrm{MVN}(\mathbf{0}, 100 I_5), \quad \gamma \sim N(0, 100).$$

Table 5.1. Description of smoking behavior data

지역	Total N	Number of $y = 0$	Number of $y > 0$	Proportion of $y = 0$
서울특별시	1359	900	459	0.662
부산광역시	750	483	267	0.644
대구광역시	392	263	129	0.671
인천광역시	423	256	167	0.605
광주광역시	262	173	89	0.660
대전광역시	270	186	84	0.689
울산광역시	265	154	111	0.581
경기도	2154	1331	823	0.618
강원도	429	235	194	0.548
충청북도	368	248	120	0.674
충청남도	422	254	168	0.602
전라북도	342	225	117	0.658
전라남도	490	309	181	0.631
경상북도	597	377	220	0.631
경상남도	572	366	206	0.640
제주특별자치도	130	85	45	0.654
세종특별자치시	45	29	16	0.644
전지역	9270	5874	3396	0.634

모수 추정방법으로는 3.2장에서 소개한 MCMC 방법 중 하나인 분산조정 메트로폴리스 알고리즘을 사용하였는데, 50,000번의 반복시행과 25,000번의 제거(burn-in)를 통해 얻은 표본을 바탕으로 추정치를 계산하였다. 각 모수 추정치의 수렴여부를 판단하기 위해 trace plot과 Gelman과 Rubin의 $\hat{R}$을 사용하였다. 모형의 비교를 위해 랜덤효과를 고려하지 않은 모형 역시 같은 방법으로 적합시켜 그 결과를 비교해 보았다.

5.3. 분석결과

랜덤효과를 포함한 ZIP 회귀모형에서 MCMC 알고리즘을 통해 추출된 각 모수 추정치들의 trace plot은 Figures 5.1과 5.2에서 볼 수 있다. 대체적으로 수렴이 잘 이루어지고 있음을 알 수 있고, 또한 각 모수들에 대한 Gelman과 Rubin의 $\hat{R}$값이 1에 가깝게 나오는 것을 확인할 수 있었다.

Table 5.2는 두 모형에 대해 각 모수의 추정치와 95% 신용구간 및 DIC를 나타낸다. 모수들의 추정치들이 두 모형간 크게 차이가 나지 않음을 알 수 있다. 하지만 모형의 적합도를 비교하기 위해 DIC를 살펴보면, 랜덤효과를 포함한 ZIP 회귀모형이 랜덤효과를 고려하지 않은 모형보다 더 작은 값을 가지고 있고 따라서 적합을 더 잘 시키는 모형이라고 판단할 수 있다. 이는 지역 간의 변동을 설명해주고 있는 γ의 추정치의 95% 신용구간이 0을 포함하고 있지 않기 때문에, 즉 γ가 유의적으로 0보다 커서 발생한 결과라고 해석할 수 있다.

Table 5.2에 나온 모수의 추정치에 대한 해석은 다음과 같다. 비흡연에 대한 오즈(odds)와 공변량들의 관계를 살펴보면, BMI가 증가함에 따라 흡연의 오즈는 유의적으로 증가하고, 가구소득의 증가는 흡연의 오즈를 유의적으로 감소시키고 있음을 알 수 있다. 또한, 음주를 하게 되면 흡연에 대한 오즈가 3.37($= e^{1.214}$)배 가량 증가하게 된다. 0보다 큰 담배량에 대한 포아송 모형에서의 추정치에 따르면, BMI가 증가하면서 흡연량은 늘어나고, 신체활동일수와 가구소득이 증가함에 따라 흡연량은 감소하는 것으로 나타났다. 증가와 감소에 대한 크기 자체는 작지만, 통계적으로 유의한 결과를 보여주고 있다.

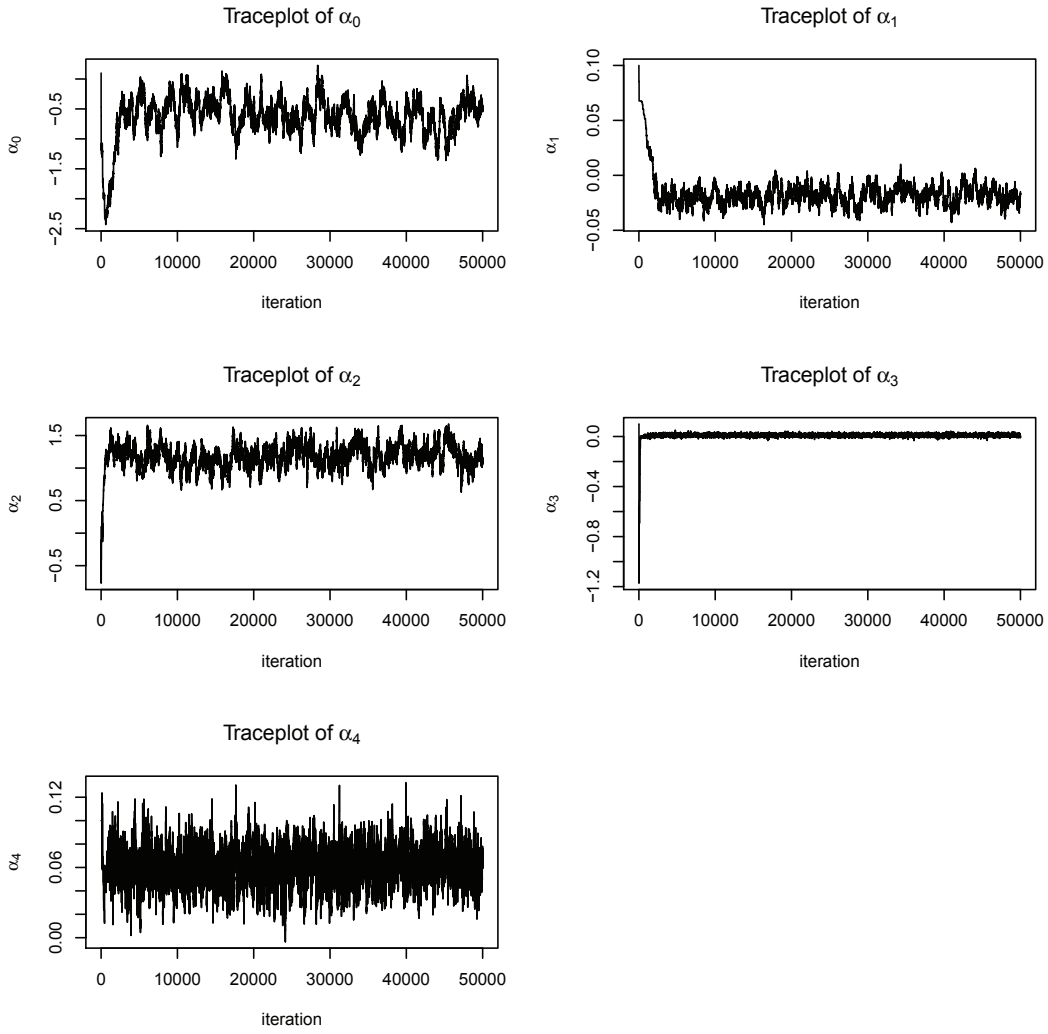

Figure 5.1. Traceplots of parameters, $\alpha_0, \alpha_1, \alpha_2, \alpha_3$, and α_4.

하지만, 음주를 하게 되면 담배량이 미미하게나마 증가하고 있음을 알 수 있다. 그 증가의 크기가 미미하지만, 이는 로짓 모형(logit model)에서의 α_2의 추정치에 대한 해석과 상반된 결과이고, 일반적으로 알려진 음주와 흡연과의 관계에도 역행하는 결과이다.

6. 결론

0이 과도하게 많이 나타나는 셀 수 있는 이산형 자료가 여러 다양한 분야에서 흔히 쓰이게 되면서 대표적인 분석방법인 ZIP 모형의 발달도 함께 이루어져 왔다. 그 중에서도 기존의 빈도론자들의 MLE이 가지고 있는 한계를 극복하고자 베이지안 추론 방법이 다양하게 발전되어 왔다. 본 논문에서는 반응변수들 사이에 상관관계가 존재하는 경우를 분석하기 위해 랜덤효과가 포함된 ZIP 회귀모형을 베이지안 추

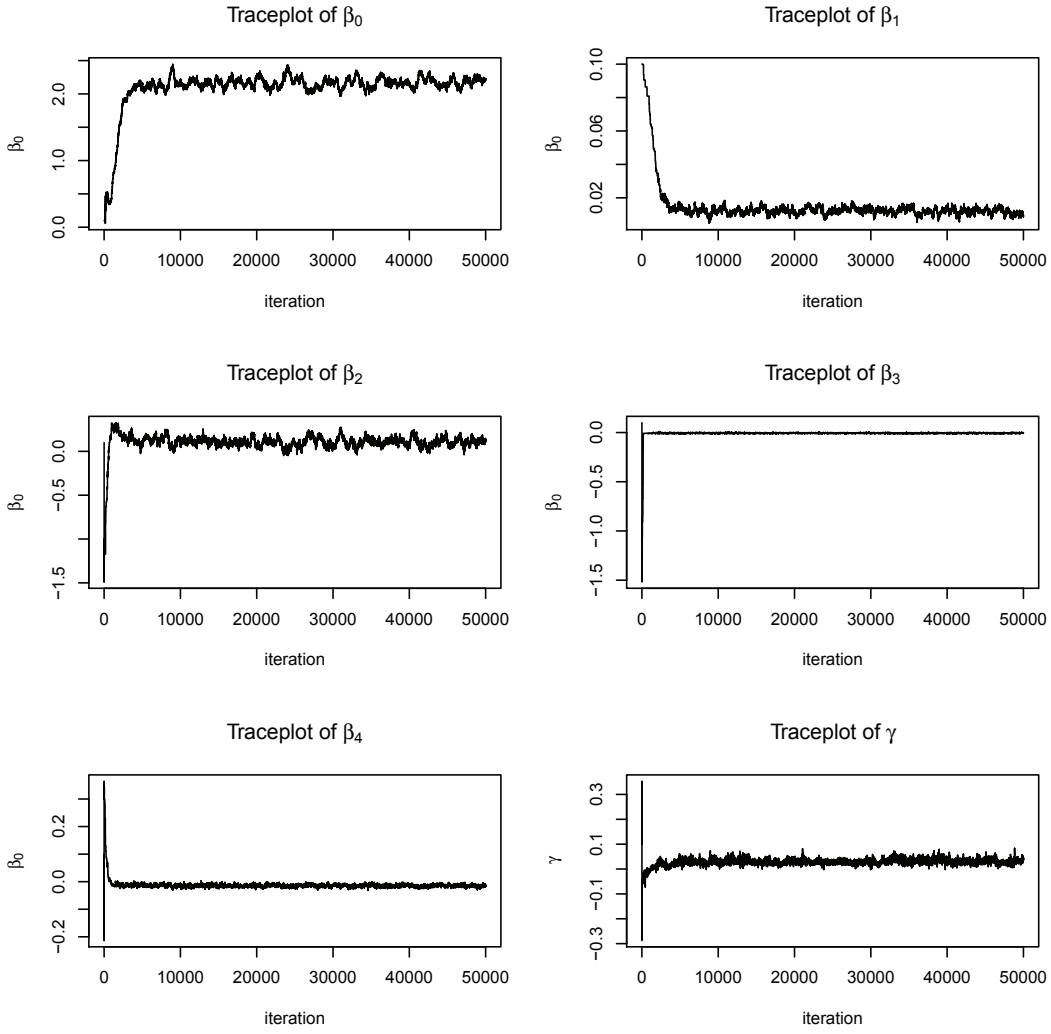

Figure 5.2. Traceplots of parameters, $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$, and γ.

론 방법을 토대로 제안하였다.

반응변수들 사이에 상관관계가 존재한다고 가정할 때, 랜덤효과를 고려한 모형이 얼마나 잘 적합하는지를 보여주기 위해 모의 실험에서 각기 다른 6가지의 시나리오를 바탕으로 분석하였다. 모의 실험 결과 모든 시나리오에서 랜덤효과를 포함한 ZIP 회귀모형이 랜덤효과를 고려하지 않은 모형보다 훨씬 작은 DIC를 가지고 있었다. 이를 통해 전자의 모형이 데이터를 더 잘 적합하고 있다고 결론내릴 수 있었다. 또한 랜덤효과를 포함한 ZIP 회귀모형을 사용하면 모수의 참값에 가까운 추정치들을 구할 수 있었고, 각각의 모수에 대해 95% 전후의 포함확률을 가지고 있음을 확인할 수 있었다. 반면 랜덤효과를 고려치 않은 ZIP 회귀모형의 경우 현저히 낮은 모수의 포함확률(6–92%)을 나타내었다. RMSE를 통한 모형 비교에서도 랜덤효과를 포함한 모형이 훨씬 작은 RMSE값을 가지고 있음을 알 수 있었다.

Table 5.2. Results of real data analysis

Parameter	Estimate(CI)	
	Model with random effects	Model without random effects
Log odds of non-smoker		
Intercept(α_0)	$-0.631(-1.161, -0.144)$	$-0.512(-1.207, -0.028)$
BMI(α_1)	$-0.018(-0.034, -0.002)$	$-0.021(-0.036, -0.004)$
Drink(α_2)	1.214(0.881, 1.532)	1.171(0.833, 1.528)
Phy.Act(α_3)	$0.010(-0.006, 0.027)$	$0.010(-0.006, 0.028)$
Income(α_4)	0.062(0.032, 0.093)	0.061(0.029, 0.092)
Log of smoking count		
Intercept(β_0)	2.166(2.024, 2.309)	2.143(2.050, 2.239)
BMI(β_1)	0.012(0.008, 0.016)	0.013(0.010, 0.016)
Drink(β_2)	0.108(0.014, 0.214)	0.118(0.037, 0.195)
Phy.Act(β_3)	$-0.005(-0.010, -0.001)$	$-0.006(-0.009, -0.002)$
Income(β_4)	$-0.015(-0.023, -0.006)$	$-0.015(-0.022, -0.008)$
Random Effect(γ)	0.032(0.016, 0.052)	-
DIC	36230.02	36259.05

CI = credible interval; DIC = deviance information criterion.

실제 흡연 자료를 대상으로 분석했을 때에도 랜덤효과를 포함한 ZIP 회귀모형이 그렇지 않은 모형보다 분석자료를 더 잘 설명하고 있음을 알 수 있었다. 모형의 비교를 위해 사용한 DIC값을 바탕으로 했을 때, 랜덤효과를 포함한 ZIP 회귀모형이 랜덤효과를 고려하지 않은 모형보다 더 좋은 모형이라는 결론을 내릴 수 있었다. 그리고, 모수들의 추정치를 해석할 때에도, 일반적으로 흡연에 미치는 영향과 대체적으로 비슷한 해석을 내릴 수 있었다.

본 논문에서는 공변량을 베르누이 분포 부분과 포아송 분포 부분에 공통적으로 사용한다고 가정하였고, 랜덤효과를 위해 랜덤 절편이 포아송 모형 부분에만 포함된다고 가정하였다. 서로 다른 공변량을 도입하고, 랜덤 기울기를 포함한 랜덤효과를 베르누이와 포아송 분포 두 부분에 모두 포함시킨다면 보다 일반적인 모형으로 발전시킬 수 있을 것이다. 또한 본 논문에서 사용한 실제 자료는 표본의 크기가 상당히 크기 때문에 베이지안 추론 방법의 장점을 쉽게 보여줄 수가 없었다. 보다 적합한 실제 자료를 찾아서 분석해 보는 것도 향후 과제 중 하나가 될 수 있을 것이다.

References

Angers, J. F. and Biswas, A. (2003). Bayesian analysis of zero-inflated generalized Poisson model, *Computational Statistics and Data Analysis*, **42**, 37–46.

Celeux, G., Forbes, F., Robert, C. P., and Titterington, D. M. (2006). Deviance information criterion for missing data models, *Bayesian Analysis*, **1**, 651–674.

Cohen, A. C. (1963). Estimation in mixtures of discrete distributions. In *Proceedings of the International Symposium on Discrete Distributions*, Montreal, 373–378.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian Data Analysis*, CRC Press, New York.

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

Greene, W. H. (1994). Accounting for excess zeros and sample selection in Poisson and negative binomial regression models, *NYU Working Paper*, No. EC-94-10.

Haario, H., Saksman, E., and Tamminen, J. (2005). Componentwise adaptation for high dimensional MCMC, *Computational Statistics*, **20**, 265–273.

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

Jang, H., Kang, Y., Lee, S., and Kim, S. W. (2008). Bayesian analysis for the zero-inflated regression models, *The Korean Journal of Applied Statistics*, **21**, 603–613.

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

Lee, J., Choi, T., and Woo, Y. (2011a). Bayesian approaches to zero inflated Poisson model, *The Korean Journal of Applied Statistics*, **24**, 677–693.

Lee, K., Joo, Y., Song, J. J., and Harper, D. W. (2011b). Analysis of zero-inflated clustered count data: a marginalized model approach, *Computational Statistics and Data Analysis*, **55**, 824–837.

Li, C. S., Lu, J. C., Park, J., Kim, K., Brinkley, P. A., and Peterson, J. P. (1999). Multivariate zero-inflated Poisson models and their applications, *Technometrics*, **41**, 29–38.

Liu, H. and Powers, D. A. (2012). Bayesian inference for zero-inflated Poisson regression models, *Journal of Statistics: Advances in Theory and Applications*, **7**, 155–188.

Min, Y. and Agresti, A. (2005). Random effect models for repeated measures of zero-inflated count data, *Statistical Modelling*, **5**, 1–19.

Neelon, B. H., O'Malley, A. J., and Normand, S. L. T. (2010). A Bayesian model for repeated measures zero-inflated count data with application to outpatient psychiatric service use, *Statistical Modelling*, **10**, 421–439.

Oh, M. S. and Lim, A. K. (2006). Bayesian analysis of a zero-inflated Poisson regression model: An application to Korean oral hygienic data, *The Korean Journal of Applied Statistics*, **19**, 505–519.

Ridout, M., Hinde, J., and Demetrio, C. G. B. (2001). A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives, *Biometrics*, **57**, 219–223.

Rodrigues, R. (2003). Bayesian analysis of zero-inflated distributions, *Communications in Statistics*, **32**, 281–289.

Shim, J., Lee, D. H., and Jung, B. C. (2011). Bayesian inference for the zero inflated negative binomial regression model, *The Korean Journal of Applied Statistics*, **24**, 951–961.

Spiegelhalter, D. J., Best, N. G., Carline, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit, *Journal of the Royal Statistical Society: Series B*, **64**, 583–639.

Yau, K. K. W. and Lee, A. H. (2001). Zero-inflated Poisson regression with random effects to evaluate an occupational injury prevention program, *Statistics in Medicine*, **20**, 2907–2920.

랜덤효과를 포함한 영과잉 포아송 회귀모형에 대한 베이지안 추론: 흡연 자료에의 적용

김연경[a] · 황범석[a,1]

[a]중앙대학교 응용통계학과

(2018년 2월 13일 접수, 2018년 2월 16일 수정, 2018년 2월 17일 채택)

요 약

0이 과도하게 많이 나타나는 자료는 여러 다양한 분야에서 흔히 볼 수 있다. 이러한 자료들을 분석할 때 대표적으로 영과잉 포아송 모형이 사용된다. 특히 반응변수들 사이에 상관관계가 존재할 때에는 랜덤효과를 영과잉 포아송 모형에 도입해서 분석해야 한다. 이러한 모형은 주로 빈도론자들의 접근방법으로 분석되어왔는데, 최근에는 베이지안 기법을 사용한 분석도 다양하게 발전되어 왔다. 본 논문에서는 반응변수들 사이에 상관관계가 존재하는 경우 랜덤효과가 포함된 영과잉 포아송 회귀모형을 베이지안 추론 방법을 토대로 제안하였다. 이 모형의 적합성을 판단하기 위해 모의 실험을 통해 랜덤효과를 고려하지 않은 모형과 비교 분석하였다. 또한, 실제 지역사회 건강조사 흡연 자료에 직접 응용하여 그 결과를 살펴보았다.

주요용어: 랜덤효과, 마코프체인 몬테카를로, 메트로폴리스 알고리즘, 영과잉 이산형 자료, 흡연자료

이 논문은 2016년도 중앙대학교 연구장학기금 지원에 의한 것임.

[1]교신저자: (06974) 서울특별시 동작구 흑석로 84, 중앙대학교 응용통계학과. E-mail: bshwang@cau.ac.kr