응용통계연구 (2011)
24(5), 951–961
DOI: http://dx.doi.org/10.5351/KJAS.2011.24.5.951

제로팽창 음이항 회귀모형에 대한 베이지안 추론

심정숙[1] · 이동희[2] · 정병철[3]

[1]서울시립대학교 통계학과, [2]경기대학교 경영학과, [3]서울시립대학교 통계학과

(2011년 7월 접수, 2011년 8월 채택)

요 약

본 논문에서는 제로팽창 음이항(ZINB) 회귀모형에서 회귀계수에 대한 추론방법으로 마코프체인몬테카를로(MCMC) 기법을 이용한 베이지안 추론방법을 제안하였다. 본 연구에서 고려한 ZINB 회귀모형은 반응변수의 평균뿐만 아니라 제로팽창확률에 대한 회귀모형을 고려한 것으로서 Jang, et al.(2010)의 연구를 확장한 것이다. 아울러 실제사례에 본 연구에서 제안한 베이지안 추론방법을 적용하고 과대산포를 허용하지 않는 제로팽창 포아송(ZIP) 회귀모형과 적합결과를 DIC를 이용하여 비교하였다. 실제 사례분석 결과 ZINB 회귀모형의 DIC가 ZIP모형보다 작게 나타나 ZINB 회귀모형이 ZIP 회귀모형보다 잘 적합되었음을 알 수 있었다.

주요용어: 베이지안 모형선택, 잠재변수, 제로팽창 음이항 회귀모형, 마코프체인 몬테카를로(MCMC).

1. 서론

제로팽창 계수형 자료는 계수형 반응변수에서 0의 값이 가정된 분포에 비하여 많이 발생하는 형태의 자료를 의미하며, 실제 경제학 및 의학 등에서 흔히 발생하는 형태의 자료이다. 예를 들어 설문조사를 통하여 최근 1달 동안 병원에 방문한 횟수를 Y라 정의한다면 Y는 0의 값이 가정된 분포에 비하여 많이 발생하게 될 것이다. 이러한 제로팽창 계수형 자료에 대하여 Lambert (1992)는 제로팽창 포아송(Zero-inflated Poisson; ZIP) 회귀모형의 사용을 제안하였다. ZIP 모형은 포아송 분포의 성격상 평균과 분산이 동일한 자료에 대해서만 사용할 수 있다는 문제가 존재한다. 하지만 실제 얻어지는 계수형 자료는 평균에 비하여 분산이 커지는 과대산포가 발생하는 경우가 흔히 발생한다. 이와 같이 과대산포가 존재하는 계수형 자료에 대하여 Ridout 등 (1998, 2001) 및 Yau 등 (2003)은 제로팽창 음이항(Zero-inflated Negative Binomial; ZINB) 회귀모형을 제안하였다.

하지만 이러한 다양한 제로팽창 계수자료에 대한 연구는 주로 빈도주의의 관점에서 수행되어 왔다. 그러나 빈도주의적 관점 또는 고전적 추론은 대표본 근사를 사용하므로 제로팽창된 자료와 같이 매우 치우친 분포를 갖는 경우 표본크기가 크지 않으면 편의가 크고 작은 포함확률을 가지는 등 추정치의 신뢰도가 떨어진다는 단점이 있다 (임아경과 오만숙, 2008). 베이지안 분석방법은 기존의 고전적인 분석방법에 비하여 유용한 사전정보의 사용을 가능하게 한다. 모수를 추정할 때 점근적 근사를 사용하지 않고 모수의 실제 표본을 이용할 뿐만 아니라 (Gelman 등, 2004; Congdon, 2005; Rodrigues, 2003; Ghosh 등, 2006), 표본의 크기가 작을 때 편의와 정확성에 있어 상대적으로 최대우도방법보다 더 신뢰성 있는 분석

이 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임(No.2010-0024403).

[3]교신저자: (130-743) 서울시 동대문구 전농동 서울시립대로 163, 서울시립대학교 통계학과, 교수.
E-mail: bcjung@uos.ac.kr

을 할 수 있는 장점을 가지고 있다 (장학진 등, 2008; Ghosh 등, 2006). 이러한 베이지안 접근방법들은 모수의 사후분포의 기댓값을 수치적으로 구할 수 있는 마코프체인 몬테카를로(Markov Chain Monte Carlo; MCMC) 방법을 이용하고 있으며, 통계학의 여러 분야에서 베이지안 추론방법들이 활발히 사용되고 있다.

베이지안 관점에서 제로팽창 계수자료에 대한 모형을 연구한 연구로는 Ghosh 등 (2006), 임아경과 오만숙 (2008), 장학진 등 (2008) 및 Jang 등 (2010) 등을 들 수 있다. Ghosh 등 (2006) 및 임아경과 오만숙 (2008)은 ZIP모형에 대한 베이지안 추론방법을 제안하였고, 장학진 등 (2008) 및 Jang 등 (2010)은 제로팽창된 교통사고 자료에 대하여 ZINB 모형을 가정하고 베이지안 추론방법을 제안하였다. 장학진 등 (2008) 및 Jang 등 (2010)이 고려한 ZINB 회귀모형은 제로팽창확률을 하나의 고정된 모수로 가정하고 이 모형에 대한 베이지안 추론방법을 제안하였다. 그러나 제로팽창 자료에 대한 모형을 고려할 경우 어떤 요인들에 의하여 제로팽창확률들이 영향을 받는가를 살펴보는 것은 관련 분야의 연구자에게는 매우 중요한 문제일 수 있다. 하지만 그들의 연구는 이러한 부분을 모형에 고려하지 않았다는 관점에서 연구의 제한점이 존재한다.

이에 본 연구에서는 과대산포가 존재하는 제로팽창 계수형 자료에 대하여 제로팽창확률에 대한 모형과정을 포함하는 보다 일반적인 형태의 ZINB 회귀모형에 대한 베이지안 추론방법을 제안하고자 한다. 아울러 과대산포가 존재하는 제로팽창 계수형 형태의 실제 사례에 본 연구에서 제안한 추론방법을 적용하고 DIC(Deviance Information Criterion)를 이용하여 최적모형을 선택하는 방법론을 제안하고자 한다 (Kass와 Raftery, 1995; Spiegelhalter 등, 2002).

본 논문의 구성은 다음과 같다. 먼저 2장에서는 ZINB 회귀모형에 대하여 기술하고, 3장에서는 각 모수에 대한 사전분포를 제시하고 결합사후분포를 구한 후 이들의 베이지안 추론방법을 제안한다. 4장에서는 ZINB 회귀모형과 ZIP 회귀모형의 비교를 위한 DIC(Deviance Information Criterion)를 간략하게 설명하고, 5장에서는 실제 자료에 본 연구에서 제안된 베이지안 추론 방법을 적용하여 분석하고 DIC를 이용하여 모형선택을 실시하고자 한다. 마지막으로 6장에서는 본 연구의 결론에 대하여 기술한다.

2. ZINB 회귀모형

먼저 Y_i $(i = 1, 2, \ldots, N)$는 음이 아닌 정수 값을 갖는 반응변수로서 1의 확률로 0의 값이 발생하는 제로상태(Zero-state)와 음이항 분포를 따르는 음이항 상태(NB state)의 혼합으로 이루어지는 다음과 같은 확률분포를 갖는다고 가정해보자.

$$Y_i \sim \begin{cases} 0, & \text{with probability } \phi_i, \\ \text{NB}(\mu_i, \tau), & \text{with probability } 1-\phi_i, \end{cases} \tag{2.1}$$

여기서 ϕ_i는 제로상태에서 0이 관측될 팽창확률을 나타낸다. 그러므로 Y_i가 음이항 상태에서 관측될 확률은 $1-\phi_i$로 나타난다. 식 (2.1)에서 $\text{NB}(\mu_i, \tau)$는 전통적인 음이항 분포를 나타내며, 다음과 같은 형태를 갖는다.

$$f(y_i; \mu_i, \tau) = \frac{\Gamma(y_i+\tau)}{\Gamma(y_i+1)\Gamma(\tau)} \left(\frac{\tau}{\mu_i+\tau}\right)^{\tau} \left(\frac{\mu_i}{\mu_i+\tau}\right)^{y_i}, \tag{2.2}$$

여기서 $\tau > 0$는 0에 가까울수록 반응변수 Y의 평균에 비하여 분산이 커지게 되는 것을 나타내는 산포모수이다. 이와 같은 구조를 갖는 반응변수 Y_i 확률분포는 아래와 같이 정의할 수 있다.

$$P(Y_i = y_i) = \begin{cases} \phi_i + (1-\phi_i) f(y_i), & \text{if } \ y_i = 0, \\ (1-\phi_i) f(y_i), & \text{if } \ y_i = 0, 1, 2, \ldots, \end{cases} \tag{2.3}$$

여기서 $f(y_i)$는 식 (2.2)에 표현된 음이항 분포를 나타낸다.

이제 이와 같은 자료에 대한 회귀모형을 고려해보자. 먼저 식 (2.2)에 나타난 음이항 분포의 평균은 μ_i이며 항상 0보다 큰 값을 갖아야 하므로 μ_i에 대해서는 로그연결함수를 사용하고자 한다. 아울러 Y_i의 제로팽창확률 ϕ_i는 항상 0과 1사이의 값을 가짐으로 ϕ_i에 대해서는 로짓 연결함수(logit link function)를 이용하고자 한다. 그러므로 μ_i와 ϕ_i는 다음과 같이 설명변수들에 의하여 표현할 수 있다.

$$\log(\mu_i) = X_i^t\beta, \qquad \log\left(\frac{\phi_i}{1-\phi_i}\right) = Z_i^t\gamma, \tag{2.4}$$

여기서 X_i와 Z_i는 각각 $k_1 \times 1$과 $k_2 \times 1$차원의 설명변수 벡터를 나타내고, β와 γ는 각각 $k_1 \times 1$과 $k_2 \times 1$차원의 모수벡터를 나타낸다. 식 (2.4)와 같은 모형화에서 Jang 등 (2010)의 모형은 $\phi_i = \phi$, $(i = 1, \ldots, n)$로 정의하고 ϕ를 하나의 모수로 고려한(즉, 로짓연결함수를 고려하지 않는) 모형이다.

마지막으로 식 (2.1)과 같은 모형과정(modeling process)에서 실제로 Y_i의 값이 0으로 나타난 경우, 이 값이 제로상태에서 얻어진 값인지 음이상 상태에서 얻어진 값인지를 결정해주는 잠재변수(latent variable) J_i를 다음과 같이 정의하고자 한다.

$$J_i = \begin{cases} 1, & \text{if } \ y_i \text{ come from zero-state,} \\ 0, & \text{if } \ y_i \sim \mathrm{NB}(\mu_i, \tau). \end{cases} \tag{2.5}$$

이때 Y_i의 값이 제로 값으로 관측된 조건하에서 J_i는 다음과 같은 성공확률을 갖는 이항분포를 따르게 된다.

$$P(J_i = 1 | Y_i = 0, \beta, \gamma, \tau) = \frac{\phi_i}{\phi_i + (1-\phi_i)f(0)}, \tag{2.6}$$

여기서 $f(0)$는 음이항 분포에서 0이 나타날 확률, 즉 $f(0) = (\tau/(\mu_i + \tau))^\tau$를 나타낸다.

이상과 같은 모형 정의하여 반응변수 $Y_1, Y_2, \ldots, Y_n$이 주어졌을 때, 베이지안 추론에 필요한 모수에 대한 우도함수는 다음과 같이 구해진다.

$$\begin{aligned} L(\beta, \gamma, \tau, J|Y) &= \prod_{i=1}^{N} \left[I_{(y_i=0)} \left\{ \phi_i^{J_i} + \{(1-\phi_i)f(y_i)\}^{1-J_i} \right\} + I_{(y_i>0)} \left\{ (1-\phi_i)f(y_i) \right\} \right] \\ &= \prod_{y_i=0} \phi_i^{J_i} \prod_{y_i=0} \{(1-\phi_i)f(y_i)\}^{1-J_i} \prod_{y_i \neq 0} \{(1-\phi_i)f(y_i)\} \\ &= \prod_{y_i=0} \left(\frac{\phi_i}{(1-\phi_i)f(y_i)} \right)^{J_i} \prod_{y_i=0} \{(1-\phi_i)f(y_i)\} \prod_{y_i \neq 0} \{(1-\phi_i)f(y_i)\} \\ &= \prod_{y_i=0} \left(\frac{1}{f(y_i)} \right)^{J_i} \prod_{i=1}^{N} f(y_i) \prod_{y_i=0} \left(\frac{\phi_i}{1-\phi_i} \right)^{J_i} \prod_{i=1}^{N} (1-\phi_i), \end{aligned} \tag{2.7}$$

여기서 $I(\cdot)$은 Y_i가 제로일 때는 1의 값을 갖고 그렇지 않은 경우 0의 값을 갖는 지시함수를 나타낸다. 식 (2.7)에 나타난 우도함수는 J_i가 주어진 상태에서 β와 γ의 함수로 분리됨을 알 수 있다.

3. 사전분포와 사후분포

이 장에서는 추정하고자 하는 모수 β, γ 및 τ에 대한 사전분포를 정의하고 마코프체인 몬테카를로 알고리즘을 통해 관심모수를 추정하고자 각 모수의 조건부 사후분포를 유도하고자 한다. 먼저 차원이 $k_1 \times$

1인 모수벡터 β와 $k_2 \times 1$인 모수벡터 γ에 각각 독립적으로 다변량 정규분포를 사전분포로 가정하고자 한다.

$$\beta \sim \text{MVN}(\beta_0, \Sigma_\beta), \qquad \gamma \sim \text{MVN}(\gamma_0, \Sigma_\gamma), \tag{3.1}$$

여기서 사전분포의 평균인 β_0와 γ_0의 차원은 $k_1 \times 1$과 $k_2 \times 1$, 공분산인 Σ_β와 Σ_γ의 차원은 $k_1 \times k_1$과 $k_2 \times k_2$이다. 아울러 산포모수 τ에 대해서는 균일분포를 사전분포로 가정한다. 따라서 미지의 모수 (β, γ, τ)의 결합사전분포 $\pi(\beta, \gamma, \tau)$는 다음과 같이 구해진다.

$$\begin{aligned}\pi(\beta,\gamma,\tau) &\propto \pi(\beta) \times \pi(\gamma) \times \pi(\tau) \\ &\propto |\Sigma_\beta|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\beta-\beta_0)^t\Sigma_\beta^{-1}(\beta-\beta_0)\right) \times |\Sigma_\gamma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\gamma-\gamma_0)^t\Sigma_\gamma^{-1}(\gamma-\gamma_0)\right). \end{aligned} \tag{3.2}$$

식 (2.7)에 정의한 우도함수와 식 (3.2)의 사전분포를 이용한 결합사후확률분포는 베이즈 정리(Bayes Theorem)에 의하여 이는 다음과 같이 구해진다.

$$\begin{aligned}\pi(\beta,\gamma,\tau,J|y,x) &\propto \pi(\beta,\gamma) \times L(\beta,\gamma,\tau,J|y,x) \\ &\propto \prod_{y_i=0}\left(e^{Z_i^t\gamma}\right)^{J_i} \prod_{i=1}^{N}\left(\frac{1}{1+e^{Z_i^t\gamma}}\right) \prod_{y_i=0}\left(\left(\frac{\tau}{e^{X_i^t\beta}+\tau}\right)^{\tau}\right)^{-J_i} \\ &\quad \times \prod_{i=1}^{N}\frac{\Gamma(y_i+\tau)}{\Gamma(y_i)\Gamma(\tau)}\left(\frac{\tau}{e^{X_i^t\beta}+\tau}\right)^{\tau}\left(\frac{e^{X_i^t\beta}}{e^{X_i^t\beta}+\tau}\right)^{y_i} \\ &\quad \times |\Sigma_\beta|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\beta-\beta_0)^t\Sigma_\beta^{-1}(\beta-\beta_0)\right) \\ &\quad \times |\Sigma_\gamma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\gamma-\gamma_0)^t\Sigma_\gamma^{-1}(\gamma-\gamma_0)\right). \end{aligned} \tag{3.3}$$

식 (3.3)에 나타난 결합사후확률분포로 부터 얻은 각 모수에 대한 조건부 사후분포는 다음과 같이 계산된다.

$$\begin{aligned}\pi(\beta|\tau,J,y) &\propto \prod_{y_i=0}\left(\left(\frac{\tau}{e^{X_i^t\beta}+\tau}\right)^{\tau}\right)^{-J_i} \prod_{i=1}^{N}\frac{\Gamma(y_i+\tau)}{\Gamma(y_i)\Gamma(\tau)}\left(\frac{\tau}{e^{X_i^t\beta+\tau}}\right)^{\tau}\left(\frac{e^{X_i^t\beta}}{e^{X_i^t\beta+\tau}}\right)^{y_i} \\ &\quad \times |\Sigma_\beta|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\beta-\beta_0)^t\Sigma_\beta^{-1}(\beta-\beta_0)\right), \\ \pi(\gamma|\tau,J,y) &\propto \prod_{y_i=0}\left(e^{Z_i^t\gamma}\right)^{J_i} \prod_{i=1}^{N}\left(\frac{1}{1+e^{Z_i^t\gamma}}\right) \times |\Sigma_\gamma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\gamma-\gamma_0)^t\Sigma_\gamma^{-1}(\gamma-\gamma_0)\right), \\ \pi(\tau|\beta,\gamma,J,y) &\propto \prod_{y_i=0}\left(e^{Z_i^t\gamma}\right)^{J_i} \prod_{y_i=0}\left(\left(\frac{\tau}{e^{X_i^t\beta}+\tau}\right)^{\tau}\right)^{-J_i} \times \prod_{i=1}^{N}\frac{\Gamma(y_i+\tau)}{\Gamma(y_i)\Gamma(\tau)}\left(\frac{\tau}{e^{X_i^t\beta}+\tau}\right)^{\tau}\left(\frac{e^{X_i^t\beta}}{e^{X_i^t\beta}+\tau}\right)^{y_i}. \end{aligned}$$

$$J_i|(y_i=0,\beta,\gamma,\tau) \sim \text{Bernoulli}\left(\frac{\phi_i}{\phi_i+(1-\phi_i)f(0)}\right). \tag{3.4}$$

식 (3.4)에 나타난 조건부 사후분포를 이용하여 각 모수들에 대한 사후추정을 할 수 있다. 그러나 식 (3.4)에 나타난 각 모수의 조건부 사후분포는 우리가 알고 있는 표준적인 분포의 형태로 간단하게 정리되지 않으므로 MCMC 알고리즘을 사용하여 모수를 추론하고자 한다. 본 연구에서는 MCMC 알고

리즘 중에서도 메트로폴리스 알고리즘(Metropolis Algorithm)과 깁스 알고리즘(Gibbs Algorithm)을 결합한 방법을 이용하고자 한다 (Peter, 2009, Chapter 10.5). 이 방법은 메트로폴리스 알고리즘의 규칙을 기반으로 각 단계마다 서로 다른 제안분포를 사용하고 각 단계에서 추정하고자 하는 모수를 분리하여 반복적으로 업데이트 한다. 이 때 메트로폴리스 알고리즘 기법 중에서도 마코프 체인으로 확률보행(random walk) 체인을 선택하고 현재의 값을 평균으로 하고 적절한 상수를 분산으로 갖는 정규분포로부터 후보 난수를 추출하는 방법을 사용하고자 한다. 각 모수의 새로운 값(candidate value)을 생성하기 위한 알고리즘은 다음과 같다.

Step 1: Update β_l, $l = 1, 2, \ldots, k_1$

1) β_l^*를 $N(\beta_l^{(t+1)}, \sigma_l)$로부터 생성.

2) 수락확률을 계산

$$\alpha_l = \min\left\{1,\ \frac{\pi\left(\beta_1^{(t+1)}, \ldots, \beta_{l-1}^{(t+1)}, \beta_{l+1}^{*}, \beta_{l+1}^{(t)}, \ldots, \beta_{k_1}^{(t)} \mid y\right)}{\pi\left(\beta_1^{(t+1)}, \ldots, \beta_{l-1}^{(t+1)}, \beta_{l+1}^{(t)}, \beta_{l+1}^{(t)}, \ldots, \beta_{k_1}^{(t)} \mid y\right)}\right\}.$$

3) 난수 $u_l \sim$ Uinform$(0,1)$에서 생성.

4) 만일 $\begin{cases} u_l \le \alpha_l, & \beta_l^{(t+1)} = \beta_l^*, \\ u_l > \alpha_l, & \beta_l^{(t+1)} = \beta_l^{(t)}. \end{cases}$

Step 2: Update γ_m, $m = 1, 2, \ldots, k_2$

1) γ_m^*를 $N(\gamma_m^{(t+1)}, \sigma_m)$로부터 생성.

2) 수락확률을 계산

$$\alpha_m = \min\left\{1,\ \frac{\pi\left(\gamma_1^{(t+1)}, \ldots, \gamma_{m-1}^{(t+1)}, \gamma_{m+1}^{*}, \gamma_{m+1}^{(t)}, \ldots, \gamma_{k_2}^{(t)} \mid y\right)}{\pi\left(\gamma_1^{(t+1)}, \ldots, \gamma_{m-1}^{(t+1)}, \gamma_{m+1}^{(t)}, \gamma_{m+1}^{(t)}, \ldots, \gamma_{k_2}^{(t)} \mid y\right)}\right\}.$$

3) 난수 $u_m \sim$ Uinform$(0,1)$에서 생성.

4) 만일 $\begin{cases} u_m \le \alpha_m, & \gamma_m^{(t+1)} = \gamma_m^*, \\ u_m > \alpha_m, & \gamma_m^{(t+1)} = \gamma_m^{(t)}. \end{cases}$

Step 3: Update τ

1) τ를 Uinform$(0,1)$으로부터 생성.

2) 수락확률을 계산

$$\alpha = \min\left\{1,\ \frac{\pi\left(\tau^* | \beta_1^{(t+1)}, \ldots, \beta_{k_1}^{(t+1)}, \gamma_1^{(t+1)}, \ldots, \beta_{k_2}^{(t+1)}, y\right)}{\pi\left(\tau^{(t)} | \beta_1^{(t+1)}, \ldots, \beta_{k_1}^{(t+1)}, \gamma_1^{(t+1)}, \ldots, \gamma_{k_2}^{(t+1)}, y\right)}\right\}.$$

3) 난수 $u \sim$ Uinform$(0,1)$에서 생성.

4) 만일 $\begin{cases} u \le \alpha, & \tau^{(t+1)} = \tau^*, \\ u > \alpha, & \tau^{(t+1)} = \tau^{(t)}. \end{cases}$

Step 4: Update J_i

$$J_i|(Y_i = 0, \beta, \gamma, \tau) \sim \text{Bernoulli}\left(\frac{\phi_i}{\phi_i + (1-\phi_i)f(0)}\right),$$

여기서 $\phi_i = \exp(Z_i^t)/(1+\exp(Z_i^t))$이고, $P(y_i = 0) = (\tau/(\mu_i + \tau))^\tau$이다.

4. DIC(Deviance Information Criterion)를 이용한 베이지안 모형비교

DIC는 베이지안 분석에서 모형 비교 시 가능도비에 기반한 기존의 AIC(Akaike Infomation Criterion) 및 BIC(Bayesian Information Criterion)와 유사한 통계량으로 Spiegelhalter 등 (2002)에 의하여 제안되었다. AIC와 BIC의 경우 MCMC 시뮬레이션으로부터 손쉽게 얻어지지 않아 최대우도함수를 계산하여야 한다. 반면에 DIC는 베이지안 모형 선택의 경우 다른 기준들에 비하여 마코프체인 몬테카를로 시뮬레이션에 의해 생성된 표본들로부터 쉽게 계산되어진다. 따라서 DIC는 마코프체인 몬테카를로 시뮬레이션에 의해 얻어진 모형의 사후분포인 베이지안 모형을 선택하는 문제에서 다른 기준들에 비하여 유용하다. DIC는 시뮬레이션으로 얻어진 추정 값(θ)을 이용하여 계산된 $D(\theta)$의 평균인 $\bar{D}$와 추정 값의 평균에서 계산된 $D(\bar{\theta})$로 다음과 같이 쉽게 계산된다.

$$\text{DIC} = D(\bar{\theta}) + 2p_D = \bar{D} + p_D, \tag{4.1}$$

여기서 편차(Deviance)는 $D(\theta) = -2\log(f(y|\theta)) + C$와 같이 정의된다. 이때 C는 다른 모형과 비교하는 모든 계산상에서 소거되는 상수이다. 위의 식 (4.1)에서 알 수 있듯이 DIC는 두 부분으로 이루어져 있다. 자료가 모형에 얼마나 잘 적합(goodness of fit) 하는지를 측정하는 기댓값 $\bar{D} = E_\theta[D(\theta)]$와 모형의 복잡성(complexity)에 대한 패널티(penalty) 즉, 모형에서 사용된 모수의 수를 나타내는 $p_D = \bar{D} - D(\bar{\theta})$로 구분된다. 여기서 $\bar{\theta}$는 θ의 기댓값을 나타낸다. 따라서 DIC가 작게 나타날 수록 모형은 자료를 잘 적합시키게 된다.

5. 자료의 분석

5.1. 자료탐색

이 장에서는 본 연구에서 제안된 ZINB 회귀모형에 대한 베이지안 추정방법을 실제자료에 적용하여 그 결과를 살펴보고자 한다. 본 연구에 사용된 실제자료는 Gurmu와 Trivedi (1996)에 의하여 분석된 휴양항해(Recreational trips)자료이다. 이는 1980년 East Texas의 4개 호수(Conroe, Livingston, somervillr, and Houston) 주변 지역 23개 도시에 등록된 2,000명의 레저용 보트소유자 관리 자료이다. 이 중에서도 본 연구에서 분석된 자료는 Seller 등 (1985)에 의해 Somerville호수에서 휴양항해를 즐기는 659명에 대하여 부분표집된 자료이다. 자료의 조사항목을 살펴보면 1980년에 East Texas의 Somerville 호수에서의 휴양항해 횟수(TRIPS)를 나타내는 반응변수 Y와 설비의 주관적인 품질 순위(SO), 수상스키에 대한 응답자 기호(SKI), 연간가구소득(I), Somerville 호수에 연간사용료 지불하는 지에 대한 여부(FC3), Conroe 호수에 연간사용료 지불료(C3), Somerville 호수에 연간사용료 지불료(C3), Houston 호수에 연간사용료 지불료(C4)를 나타내는 7개의 설명변수로 이루어졌다. 본 연구에서는 반응변수 Y로 1980년에 East Texas의 Somerville 호수에서의 휴양항해 횟수(TRIPS)를 고려하였으며, 반응변수 Y의 평균 μ_i에 대한 설명변수 X로는 설비의 주관적인 품질 순위(SO), 수상스키에 대한 응답자 기호(SKI), 연간가구소득(I), Somerville 호수에 연간사용료를 지불하는 지에 대한 여부(FC3)를 X_1, X_2, X_3, X_4로 두었다. 또한 제로팽창확률 ϕ_i에 대한 설명변수 Z로는 설비의 주관적인

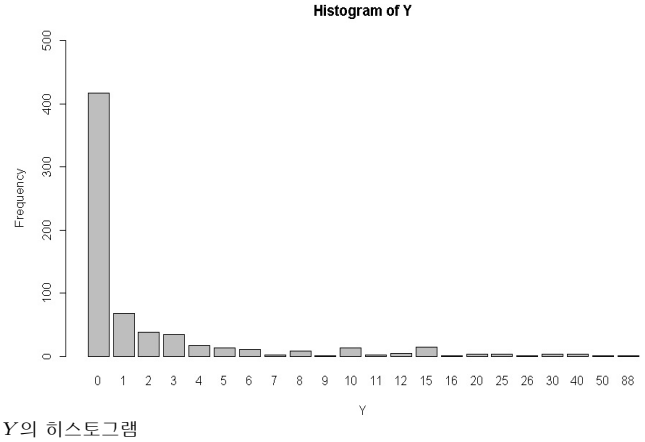


그림 5.1. 반응변수 Y의 히스토그램

품질 순위(SO)와 수입(I)만을 고려하여 Z_1, Z_2로 두었다. 설비의 주관적인 품질순위(SO)는 1점에서 5점 척도를 사용하여 조사하였으며, 수상스키에 대한 응답자 기호(SKI)는 호수에서 주로 수상스키를 즐기는 지에 대한 여부(예: 1, 아니오: 0)를 사용 하였다. 연간가구소득(I)는 9개의 범주로 범주화하였으며, 연간사용료 지불여부(FC3)는 더미변수(예: 1, 아니오: 0) 처리하였다. 그림 5.1은 반응변수 Y의 히스토그램을 나타낸 것이다. 그림 5.1를 살펴보면 반응변수 Y중에서 제로의 값이 높은 비중(63.3%)을 보이고, 이들의 중위수가 0으로 나타나 자료의 절반 이상을 제로 값이 차지하고 있음을 알 수 있다. 아울러 이 자료에 대한 기초통계량은 다음과 같이 구해진다.

$$\begin{aligned} \text{제로를 포함하는 경우:} \quad & \bar{Y} = 2.244, \quad S_y^2 = 39.595, \\ \text{제로를 포함하지 않는 경우:} \quad & \bar{Y} = 6.112, \quad S_y^2 = 84.373. \end{aligned} \tag{5.1}$$

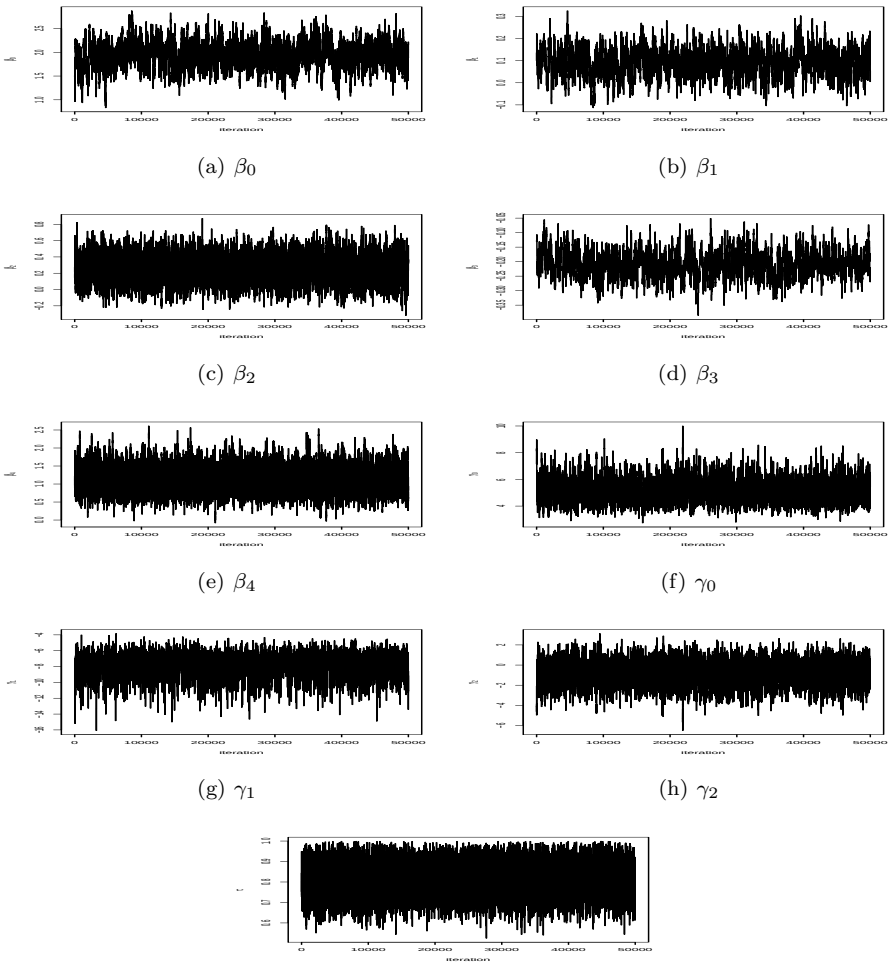
식 (5.1)에서 얻어진 제로를 포함하는 반응변수 Y와 제로를 포함하지 않는 반응변수 Y의 기초통계량을 살펴보면, 두 경우의 분산 값들이 각각의 평균값들에 비해 매우 크게 나타나 자료에 과대산포가 존재함을 확인 할 수 있었다.

5.2. 분석결과

ZINB 모형에서 사후 추정하고자 하는 관심모수 β, γ, τ에 대한 사전분포는 다음과 같이 정의하였다. 우선 모수벡터 β와 γ의 경우에는 무정보적(noninformative)이지만 적절한(proper) 사전분포인 다변량 정규분포 MVN$(0, 10I)$로 정의하였고, 산포모수 τ의 사전분포로는 균일분포를 가정하였다. ZIP회귀모형에도 β와 γ에 대한 사전분포로 다변량 정규분포 MVN$(0, 10I)$를 정의하여 분석하였다. 앞에서 기술한 메트로폴리스와 깁스 알고리즘을 결합한 기법을 이용하여 총 51,000번의 랜덤변량을 생성한 후 그 중 수렴된 표본만을 사용하기 위해 초기의 표본 1,000개를 제거(burn-in)하였다. 따라서 사후추정에 사용된 표본의 개수는 51,000개이다.

그림 5.2는 ZINB 회귀모형에서 MCMC 알고리즘을 통해 추출된 각 모추정치들의 시퀀스 플롯(Sequence plot)을 나타낸 것이다. 이를 통해 각 추정치들이 대체적으로 잘 수렴되었음을 살펴볼 수 있다.

또한 각 모수마다 50,000개의 표본을 얻었으므로 이 표본을 이용하여 각 모수들을 추정하고자 한다. 표 5.1은 50,000개의 표본을 이용하여 추정한 각 모수의 추정치, 95%신용구간 및 DIC 값을 나타낸다. 먼



(a) β_0 (b) β_1

(c) β_2 (d) β_3

(e) β_4 (f) γ_0

(g) γ_1 (h) γ_2

(i) τ

그림 5.2. ZINB 회귀모형에서의 각각의 모수에 대한 추정 값들의 Sequence plot

저 DIC 값을 살펴보면 ZIP모형의 DIC는 7181.892로 나타나고 ZINB모형은 3927.776으로 나타났다. ZINB모형이 ZIP모형보다 적합결과가 훨씬 더 좋게 나타났다. 이제 회귀모형에 대한 설명을 고려해보자. ZINB모형의 적합이 ZIP모형보다 좋게 나타났으므로 ZINB 회귀식만 설명하고자 한다. ZINB 회귀모형에 대한 추정된 회귀식은 다음과 같다.

$$\log(\mu_i) = 1.9277 + 0.0894 * (\text{SO}) + 0.2556 * (\text{SKI}) + (-0.2084) * (I) + 1.1186 * (\text{FC3}),$$

$$\log\left(\frac{\phi_i}{1-\phi_i}\right) = 5.1034 + (-7.5266) * (\text{SO}) + (-0.9911) * (I). \tag{5.2}$$

추정된 회귀식 (5.2)에 따르면 휴양 항해 수의 평균은 설비의 주관적인 품질 순위가 한 단계 증가하면 약 1.09배 증가하고, 호수에서 주로 수상스키를 즐기는 경우에는 약 1.29배, 연간사용료를 지불하는 경우에

표 **5.1.** 베이지안 추정방법에 의한 ZIP과 ZINB 회귀모형의 추정 결과와 모형비교를 위한 DIC

모형	회귀식	회귀계수	추정 값	표준오차	95% C.I	DIC
ZIP	μ	Intercept(β_0)	2.1708	0.0005	(2.1699, 2.1717)	7181.892
		SO(β_1)	0.0133	0.0001	(0.0131, 0.0135)	
		SKI(β_2)	0.2856	0.0002	(0.2851, 0.2861)	
		I(β_3)	−0.1720	0.0001	(−0.1722, −0.1719)	
		FC3(β_4)	0.9236	0.0003	(0.9229, 0.9243)	
	ϕ	Intercept(γ_0)	3.0346	0.0011	(3.0323, 3.0368)	
		SO(γ_1)	−1.6785	0.0006	(−1.6798, −1.6773)	
		I(γ_2)	0.1131	0.0015	(0.1101, 0.1162)	
ZINB	μ	Intercept(β_0)	1.9277	0.0012	(1.9253, 1.9300)	3927.776
		SO(β_1)	0.0897	0.0003	(0.0888, 0.0899)	
		SKI(β_2)	0.2565	0.0007	(0.2552, 0.2578)	
		I(β_3)	−0.2084	0.0002	(−0.2088, −0.2080)	
		FC3(β_4)	1.1186	0.0015	(1.1157, 1.1215)	
	ϕ	Intercept(γ_0)	5.1034	0.0036	(5.0962, 5.1105)	
		SO(γ_1)	−7.5266	0.0061	(−7.5385, −7.5146)	
		I(γ_2)	−0.9911	0.0050	(−1.0008, −0.9814)	
	disp. para.	τ	0.8040	0.0003	(0.8033, 0.8047)	

는 약 3.06배 증가하는 것으로 나타났다. 또한 연간가구소득의 추정계수는 음의 값을 보이며 한 단계 증가하면 휴항 항해 수의 평균은 0.81배로 감소하는 것으로 나타났다. 또한 제로팽창확률의 로짓모형에서는 설비의 주관적인 품질 순위(SO)와 연간가구소득(I)의 추정된 회귀계수가 음의 값을 보여 설비의 주관적인 품질 순위(SO)와 연간가구소득(I)가 한 단계 커질수록 제로가 나타날 확률이 낮아지는 것으로 나타났다.

6. 결론

본 연구에서는 과대산포와 제로팽창이 동시에 존재하는 계수형 자료에 대한 베이지안 추론방법에 대하여 연구하였다. 이러한 자료를 설명하는 모형으로 ZINB 회귀모형을 제시하였고, 회귀계수를 마코프체인 몬테카를로 기법을 이용하여 추론하는 베이지안 추론방법을 제안하였다. 또한 기존의 논문과 달리 팽창확률을 하나의 모수로 가정하지 않고 모형을 좀 더 일반화시키기 위해 제로팽창확률에도 로짓 연결함수를 사용하여 다양한 설명변수들에 의해 영향을 받는 것으로 가정하였다. 아울러 제로팽창과 과대산포가 존재하는 실제자료에 본 연구모형을 적용하고 과대산포를 고려하지 않는 ZIP 회귀모형과 DIC를 이용하여 모형적합 정도를 비교해보았다. 분석 결과 제로팽창 음이항 회귀모형의 DIC가 제로팽창 포아송 회귀모형의 DIC 비해 더 작은 값을 보이므로 제로팽창 음이항 회귀모형이 위의 분석자료를 잘 설명해주고 있음을 알 수 있었다.

참고문헌

임아경, 오만숙 (2006). 영과잉 포아송 회귀모형에 대한 베이지안 추론: 구강 위생 자료에의 적용, <응용통계연구>, **16**, 505–519.

장학진, 강윤희, 이수범, 김성욱 (2008). 영과잉 회귀모형에 대한 베이지안 분석, <응용통계연구>, **21**, 603–613.

Cameron, A. C. and Trivedi, P. K. (1998). *Regression Analysis of Count Data*, Cambridge University Press, Cambridge.

Congdon, P. (2005). *Bayesian Models for Categorical Data*, John Wiley & Sons, New York.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004). *Bayesian Data Analysis*, Chapman & Hall, BocaRaton.

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

Gurmu, S. and Trivedi, P. K. (1996). Excess zeros in count models for recreational trips, *Journal of Business and Economic Statistics*, **14**, 469–477.

Jang, H. J., Lee, S. B. and Kim, S. W. (2010). Bayesian analysis for zero-inflated regression models with the power prior: Aapplications to road safety countermeasures, *Accident Analysis and Prevention*, **42**, 540–547.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors, *Journal of the American Statistical Association*, **90**, 773–795.

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

Peter, D. H. (2009). *A First Course in Bayesian Statistical Methods*, Springer.

Ridout, M. S., Demetrio, C. G. B. and Hinde, J. P. (1998). Models for count data with many zeros, *Proceedings of the XIXth International Biometrics Conference, Cape Town, Invited Papers*, 179–192.

Ridout, M. S., 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, J. (2003). Bayesian analysis of zero-inflated distributions, *Communications in Statistics, Theory and Methods*, **32**, 281–289.

Seller, C., Stoll, J. R. and Chavas, J. P. (1985). Validation of empirical measures of welfare change: A comparison of nonmarket techniques, *Land Economics*, **61**, 156–175.

Spiegelhalter, D. J., Best, N. G., Carlin, 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.

Winkelmann, R. (2003). *Econometric Analysis of Count Data*, Springer, New York.

Yau, K. K. W., Wang, K. and Lee, A. H. (2003). Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros, *Biometrical*, **45**, 437–452.

Bayesian Inference for the Zero Inflated Negative Binomial Regression Model

JungSuk Shim[1] · Dong-Hee Lee[2] · Byoung Cheol Jung[3]

[1]Department of Statistics, University Of Seoul
[2]Department of Business Administration, Kyonggi University
[3]Department of Statistics, University Of Seoul

Abstract

In this paper, we propose a Bayesian inference using the Markov Chain Monte Carlo(MCMC) method for the zero inflated negative binomial(ZINB) regression model. The proposed model allows the regression model for zero inflation probability as well as the regression model for the mean of the dependent variable. This extends the work of Jang *et al.* (2010) to the fully defiend ZINB regression model. In addition, we apply the proposed method to a real data example, and compare the efficiency with the zero inflated Poisson model using the DIC. Since the DIC of the ZINB is smaller than that of the ZIP, the ZINB model shows superior performance over the ZIP model in zero inflated count data with overdispersion.

Keywords: Bayesian Model Selection, Latent variable, ZINB(Zero-Inflated Negative Binomial), MCMC (Markov Chain Monte Carlo).

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2010-0024403).
[3]Corresponding author: Professor, Department of Statistics, University Of Seoul, 163, Seoulsiripdaero, Dongdaemun-gu, Seoul 130-743, Korea. E-mail: bcjung@uos.ac.kr