응용통계연구 제 19권 3호, 2006년, pp. 505–519

영과잉 포아송 회귀모형에 대한 베이지안 추론: 구강위생 자료에의 적용*

임아경[1] 오만숙[2]

요 약

셀 수 있는 이산 자료(discrete count data)에 대한 분석은 여러 분야에서 활용되고 있지만 영(zero)을 과도하게 포함하고 있는 영과잉 자료는 자료의 성격상 포아송 분포를 따르지 못할 때가 있어 분석에 어려움이 따른다. Zero-Inflated Poisson(ZIP)모형은 이런 어려움을 극복하기 위하여 영에 대한 점확률을 가지는 분포와 포아송 분포를 합성하여 과도한 영과 영이 아닌 자료를 설명하는 모형이다. 설명변수가 존재할 때는 포아송 분포 부분에서 반응변수의 평균과 공변량사이에 로그선형 연결함수를 사용한 Zero-Inflated Poisson Regression(ZIPR) 모형이 사용될 수 있다. 본 논문에서는 Markov Chain Monte Carlo 기법을 이용한 ZIPR 모형의 베이지안 추론방법을 제안하고, 이를 실제 구강위생 자료에 적용하며 다른 모형들과 비교한다. 그 결과 베이지안 추론 방법을 적용한 영과잉 모형의 추정오차가 다른 모형들의 추정오차보다 작았고, 예측치가 더 정확했다는 점에서 우수함을 알 수 있었다.

주요용어: 포아송 회귀모형, 영과잉 자료, 합성분포, 몬테칼로

1. 서론

셀 수 있는 이산 자료(discrete count data)에 대한 분석은 여러 응용분야에서 활용되고 있다. 셀 수 있는 이산 자료에 대하여 가장 널리 적용되는 모형은 포아송 모형(poisson model) 또는 변형된 포아송 모형이다. 본 논문에서 우리의 관심은 이산자료 중 영이 과도하게 많은 경우이다. 예를 들면 제조 과정에서 얻어지는 제품의 불량품수의 경우 공정이 완벽에 가까우면 전체 자료의 80-90%가 영으로 관측된다(Lambert, 1992). 이렇듯 영의 비율이 본래의 포아송 모델에 의해 관측되는 영의 비율보다 아주 높은 비율을 차지하는 자료를 영과잉 자료(zero-inflated data)라고 한다. 이런 영과잉 자료의 예는 실제 생활에서 흔하게 접할 수 있다. 예를 들면, 한 마을에서 콜레라가 발생할 빈도(Dahiya and Gross, 1973)와 고속

* 본 연구는 한국과학재단 목적기초연구(R06-2002-012-01002-0)지원으로 수행되었음.
1) (120-750) 서울시 서대문구 대현동 21, 이화여자대학교 통계학과, 대학원생
E-mail: lak612@hanmail.net
2) (교신저자)(120-750) 서울시 서대문구 대현동 21, 이화여자대학교 통계학과, 교수
E-mail: msoh@mm.ewha.ac.kr

도로에서 자동차의 사고빈도(Lord, 2005) 등이다. 이 밖에 Gupta 외(1996), Umbach(1981), Yip(1988), Li 외(1999), Ghosh 외(2005) 등에서도 영과잉 자료를 찾아볼 수 있다.

변수들 간의 상관관계를 살펴보는데 있어 가장 많이 쓰이고 있는 모형으로 다중선형 회귀모형과 포아송 회귀모형이 있다. 그 중 다중선형 회귀모형은 수리적으로 간단해서 추정이 쉽고 변수들간의 관계 설명이 쉽다. 그러나 이 모형이 타당하기 위해서는 많은 가정들이 필요하다. 그 중에서도 오차가 정규분포를 따른다고 가정해야 하는데, 그 가정이 만족되기 위해서는 변수들이 연속형이어야 하고 대칭성을 가져야 한다. 하지만 영과잉 자료는 영을 과도하게 가지고 있는 치우친 자료이고 범주형에 가까운 카운트 자료이므로 영과잉 자료는 선형 회귀모형에 적합되기 어렵다. 다음으로 카운트 자료를 분석하는 도구로 잘 알려진 포아송 회귀모형 역시 많이 이용하고 있으나, 이 역시 자료 내에서의 큰 변동으로 인하여 과대산포(overdispersion) 현상이 일어나기 쉬워서 영과잉 자료에는 적합하기 힘들다.

기존에 이런 전통적인 추론방법을 이용한 모형들은 영을 과도하게 포함하고 있다는 자료의 성격을 고려하지 않고 일괄적으로 분석하여 자료의 정보 중에서 많은 부분을 잃게 된다. 이러한 단점을 보완하기 위하여 분포의 혼합 형태를 가지는 Zero-Inflated Poisson(ZIP) 모형이 제안되었고 설명변수가 존재할 때는 Zero-Inflated Poisson Regression(ZIPR)이 제안되었다(Lambert, 1992). ZIP 모형의 기본 아이디어는 영에 대한 점확률(point mass)을 가지는 분포와 포아송 분포를 합성하여 과도한 영과 영이 아닌 자료를 설명한다는 것이다.

영과잉 포아송 모형과 영과잉 포아송 회귀모형에 대한 분석은 주로 빈도론자들의 관점에서 수행되어 왔다. 그러나 빈도론자 또는 고전적 추론은 대표본 근사를 사용하므로 영과잉 자료와 같이 매우 치우친 분포를 갖는 경우 표본 크기가 매우 크지 않으면 추정치의 신뢰도가 떨어지는 단점이 있다. 이를 극복하기 위한 대안으로 ZIP 모형에 대한 베이지안 추론에 대한 연구가 수행되었으며(Angers and Biswas, 2003; Rodrigues, 2003; Ghosh, Mukhopadhyay and Lu, 2005) Markov Chain Monte Carlo 방법을 사용한 베이지안 추론은 표본크기가 작을 때에도 추정치의 신뢰구간을 잘 추정함을 보여주었다. 그러나 위의 베이지안 논문에서는 설명변수를 고려하지 않아 반응변수에 유의한 영향을 미치는 설명변수가 존재할 경우 이를 모형에 포함시키는 ZIPR 모형에 대한 베이지안 기법의 연구가 요구된다.

본 논문에서는 설명변수가 존재하는 경우 Markov Chain Monte Carlo 기법을 이용한 ZIPR 모형의 베이지안 추론을 제안하고 이를 실제 치과역학연구에서 얻어진 구강위생자료에 적용한다.

앞으로 2장에서는 혼합모형인 영과잉 포아송 회귀모형(ZIPR)을 기술하고, 3장에서는 영과잉 포아송 회귀모형(ZIPR)에 대한 베이지안 추론방법을 제안한다. 4장에서는 제안된 베이지안 기법을 사용하여 실제 구강위생 자료에 대한 분석을 수행한다. 또한 다중선형 회귀모형, 포아송 회귀모형, 영과잉 포아송 모형과 포아송 회귀모형의 모수추정 결과를 비교해 보겠다. 5장에서는 이 연구의 결론을 보일 것이다.

2. 영과잉 포아송 회귀모형

개체 i가 p개의 설명변수 $\mathbf{X}_i = (X_{i1}, ..., X_{ip})$와 0 이상의 정수값을 갖는 반응변수 Y_i를

가질 때 영과잉 포아송 회귀모형(ZIPR)은 Y_i의 분포를 다음과 같이 가정한다.

$$Y_i \sim \begin{cases} I_{\{0\}}(y_i) & \text{with probability} \quad \omega \\ p_i(y_i|\beta) & \text{with probability} \quad 1-\omega. \end{cases}$$

여기서 $I_{\{0\}}(y_i)$는 지시함수로 $y_i = 0$이면 1, 아니면 0값을 갖는다. $p_i(y_i|\beta)$는 평균 $\theta_i = e^{\mathbf{x}_i\beta}$를 갖는 포아송 분포의 확률밀도함수, 즉, 로그연결함수를 갖는 포아송 확률밀도함수로

$$p_i(y_i|\beta) = e^{-e^{\mathbf{x}_i\beta}} e^{y_i \cdot \mathbf{x}_i\beta}/y_i\,!$$

이며 β는 p차원 미지의 모수벡터이다.

위 모형에서 보면 $Y_i = 0$은 $I_{\{0\}}(y_i)$분포와 포아송 분포에서 모두 발생가능하므로 $Pr(Y_i = 0) = \omega + (1-\omega)p_i(0|\beta)$이다. 따라서 반응변수 Y_i의 확률 밀도 함수가 아래와 같이 주어진다.

$$Pr(Y_i = y_i) = \begin{cases} \omega + (1-\omega)p_i(0|\beta) & \text{if} \quad y_i = 0 \\ (1-\omega)p_i(y_i|\beta) & \text{if} \quad y_i = 1, 2, ... \end{cases} \tag{2.1}$$

따라서 β 와 ω의 우도함수는, $\mathbf{y} = (y_1, ..., y_n)$ 가 주어질 때,

$$l(\beta, \omega \mid \mathbf{y}) = \prod_{y_i=0, i=1}^{i=n} \{\omega + (1-\omega) \cdot p_i(0 \mid \beta)\} \times \prod_{y_i \neq 0, i=1}^{i=n} \{(1-\omega) \cdot p_i(y_i \mid \beta)\} \tag{2.2}$$

이다.

Markov Chain Monte Carlo 기법을 이용한 β 와 ω의 사후추론을 쉽게 하기 위하여 잠재변수(latent variable) J_i를 다음과 같이 정의한다.

$$J_i = \begin{cases} 1 & \text{if} \quad Y_i \sim I_{\{0\}}(y_i) \\ 0 & \text{if} \quad Y_i \sim p_i(y_i \mid \beta) \end{cases}$$

즉, $Y_i = 0$ 이 포아송분포로부터 나온 것인지 아니면 포아송 분포에 비하여 과도한 0을 나타내는 점확률분포 $I_{\{0\}}(y_i)$에서 나온 것인지 판별하는 지시변수이다. 0이 관측된 조건 하에서 $J_i = 1$일 확률, 즉, 관측된 0이 과도한 부분에서 나왔을 확률은

$$Pr(J_i = 1 \mid \beta, \omega) = \frac{\omega}{\omega + (1-\omega)p_i(0 \mid \beta)} \tag{2.3}$$

이다. 전체적으로 관측된 0의 갯수를 m, 그 중 점확률분포에서 나온 0의 갯수를 S라 정의하면, $S = \sum J_i$이며 $\mathbf{J} = \{J_i\}$가 포함된 우도함수는

$$\begin{aligned} & l(\beta, \omega, \mathbf{J} \mid \mathbf{y}) \\ & \quad = \prod_{y_i=0} \{\omega + (1-\omega) \cdot p_i(0 \mid \beta)\} \times \prod_{y_i \neq 0} \{(1-\omega) \cdot p_i(y_i \mid \beta)\} \\ & \quad = \omega^S \cdot (1-\omega)^{m-S} \prod_{y_i=0, J_i=0} p_i(0 \mid \beta) \times (1-\omega)^{n-m} \cdot \prod_{y_i \neq 0} p_i(y_i \mid \beta) \\ & \quad \propto \omega^S \cdot (1-\omega)^{n-S} \cdot \prod_{y_i=0, J_i=0} e^{-e^{\mathbf{x}_i\beta}} \cdot \prod_{y_i \neq 0} e^{y_i \cdot \mathbf{x}_i\beta} \cdot e^{-e^{\mathbf{x}_i\beta}} \end{aligned} \tag{2.4}$$

으로 $\mathbf{J}$가 주어진 상태에서 우도함수가 β와 ω의 함수로 분리됨을 알 수 있다. 다음 장에서는 ZIPR 모형의 베이지안 추론 과정에 대하여 알아본다.

3. 사전분포와 사후분포

미지의 모수 (β,ω)의 사전분포로 $\pi(\beta,\omega)$를 가정하면 $(\beta,\omega,\mathbf{J})$의 결합사후분포는

$$\pi(\beta,\omega,\mathbf{J}\mid\mathbf{y})\propto\pi(\beta,\omega)\cdot l(\beta,\omega,\mathbf{J}\mid\mathbf{y})$$

이다. 이 장에서는 사전분포를 가정하고 Markov Chain Monte Carlo 알고리즘을 적용하기 위한 각 모수의 조건부 사후분포를 유도해 본다.

두 모수 ω,β 에 대하여 각각 독립적으로 베타분포 $Beta(c,d)$ 와 다변량 정규분포 $N(\beta_0,\Sigma)$를 가정한다. 따라서

$$\pi(\beta,\omega)\propto e^{-\frac{1}{2}(\beta-\beta_0)'\Sigma^{-1}(\beta-\beta_0)}\omega^{c-1}(1-\omega)^{d-1}$$

이다. 사전분포와 우도함수를 결합하면 사후밀도함수는 다음과 같다.

$$\begin{aligned}
\pi(\beta,\omega,\mathbf{J}\mid\mathbf{y}) &\propto \pi(\beta,\omega)\cdot l(\beta,\omega,\mathbf{J}\mid\mathbf{y})\\
&\propto \omega^{c-1}(1-\omega)^{d-1}e^{-\frac{1}{2}(\beta-\beta_0)'\Sigma^{-1}(\beta-\beta_0)}\\
&\quad\cdot\omega^{S}(1-\omega)^{n-S}\cdot\prod_{y_i=0,J_i=0}\{e^{-e^{\mathbf{X}_i\beta}}\}\cdot\prod_{y_i\neq0}\{e^{y_i\cdot\mathbf{X}_i\beta}\cdot e^{-e^{\mathbf{X}_i\beta}}\}\\
&\propto \omega^{S+c-1}(1-\omega)^{n-S+d-1}\cdot e^{-\frac{1}{2}(\beta-\beta_0)'\Sigma^{-1}(\beta-\beta_0)}\\
&\quad\cdot\prod_{y_i=0,J_i=0}\{e^{-e^{\mathbf{X}_i\beta}}\}\cdot\prod_{y_i\neq0}\{e^{(y_i\cdot\mathbf{X}_i\beta-e^{\mathbf{X}_i\beta})}\} \qquad (3.1)
\end{aligned}$$

주어진 모수들에 대한 사후추정이 필요하나 식이 간단치 않아 수리적으로 가능치 않으므로 Markov Chain Monte Carlo기법을 사용하고자 한다. 구체적으로 각 모수들에 대한 조건부 사후분포(conditional posterior distribution)를 구해서 깁스 표본 기법을 이용한다. 그러나 모수 β는 조건부 사후분포를 수리적으로 구할 수 없으므로 메트로폴리스 해스팅 알고리즘(Metropolis-Hasting Algorithm)을 이용한다. 모수에 대한 조건부 사후분포를 정리해 보기로 한다.

식 (3.1)의 결합 사후분포로부터 ω에 대한 조건부 사후분포는 다음과 같이 구할 수 있다.

$$\begin{aligned}
\pi(\omega\mid\beta,\mathbf{y},\mathbf{J}) &\propto \omega^{S+c-1}(1-\omega)^{n-S+d-1}\\
&\sim Beta(S+c,n-S+d) \qquad (3.2)
\end{aligned}$$

식 (2.3)에서 $y_i=0$일때 $J_i=1$일 조건부 확률을 보면

$$P(J_i=1\mid\beta,\omega,y_i=0)=\frac{\omega}{\omega+(1-\omega)e^{-e^{\mathbf{X}_i\beta}}} \qquad (3.3)$$

임을 알 수 있다. 따라서

$$J_i \mid \beta, \omega, y_i = 0 \quad \sim \quad Bernoulli(\frac{\omega}{\omega + (1-\omega)e^{-e^{\mathbf{X}_i\beta}}})$$

이다.

β에 대한 식 ((3.1)은 매우 복잡하여 조건부 사후분포를 수리적으로 구할 수 없으므로 β는 메트로폴리스 해스팅(Metropolis-Hasting, MH) 알고리즘을 이용하여 추론하도록 한다. β 의 사전분포는 $\pi(\beta) \sim N(\beta_0, \Sigma)$이므로, β 의 조건부 사후분포를 구해보면

$$\pi(\beta|\omega, \mathbf{J}, \mathbf{y}) \propto e^{-\frac{1}{2}(\beta-\beta_0)'\Sigma^{-1}(\beta-\beta_0)} \cdot \prod_{y_i=0, J_i=0} \{e^{-e^{\mathbf{X}_i\beta}}\} \cdot \prod_{y_i \neq 0} \{e^{(y_i \cdot \mathbf{X}_i\beta - e^{\mathbf{X}_i\beta})}\} \qquad (3.4)$$

가 된다.

메트로폴리스 해스팅 기법에서 마코브 연쇄로 확률보행(random walk) 연쇄를 선택하고 현재의 값을 평균으로, 적절한 상수 δ를 원소로 갖는 대각행렬 δI를 분산으로 갖는 정규분포로부터 후보 난수를 추출하는 과정은 다음과 같다.

- Step1 : $\beta^{(0)} = \beta_0, S^{(0)} = \frac{1}{2}m.\ k = 1.$
- Step2 : $\omega^{(k)} \sim Beta(S^{(k-1)} + c, n - S^{(k-1)} + d)$
- Step3 : $J_i^{(k)}|y_i = 0 \sim Bernoulli(\frac{\omega^{(k)}}{\omega^{(k)} + (1-\omega^{(k)})e^{-e^{\mathbf{X}_i\beta^{(k-1)}}}}),\ i = 1, ..., m$
- Step4 : β^*를 $N(\beta^{(k-1)}, \delta I)$로부터 생성하고,

$$\pi(\beta|\mathbf{y}) = e^{-\frac{1}{2}(\beta-\beta_0)'\Sigma^{-1}(\beta-\beta_0)} \cdot \prod_{y_i=0, J_i=0} \{e^{-e^{\mathbf{X}_i\beta}}\} \cdot \prod_{y_i \neq 0} \{e^{(y_i \cdot \mathbf{X}_i\beta - e^{\mathbf{X}_i\beta})}\}$$

$$\alpha = min\{1, \frac{\pi(\beta^*|\mathbf{y})}{\pi(\beta^{(k-1)}|\mathbf{y})}\}$$

으로 정의한다. 난수 $u^{(k)}$ 를 $Uniform(0,1)$에서 생성하여

$$\beta^{(k)} = \begin{cases} \beta^* & \text{if} \quad u^{(k)} \leq \alpha \\ \beta^{(k)} = \beta^{(k-1)} & \text{if} \quad u^{(k)} > \alpha \end{cases}$$

로 놓는다.

- Step5 : $k = k + 1$ 로 하고 Step2 -Step4 를 반복한다.

위에 주어진 조건부 사후분포를 이용하여 Markov Chain Monte Carlo기법을 수행하면 모수들의 사후표본을 얻을 수 있고 이를 기반으로 원하는 사후 추정이 가능하게 된다.

4. 구강위생 자료의 분석

4.1. 자료탐색

3장에서 제안된 ZIPR에 대한 베이지안 추정법을 사용하여 2000년 6월 30일부터 2001년 2월 28일까지 실시된 '국민구강건강 실태조사'의 자료를 분석하였다. 표본선정 방법은 1995년 인구주택 총 조사의 조사구 중에서 시설단위 조사구를 제외한 보통조사구에서 총 200개의 표본 조사구를 추출하였다. 실제 조사는 사전에 조사자 훈련(calibration training)을 받은 전국 치과대학 예방치과 교수 및 전공치과의사로 이루어진 15개 구강 검사팀이 시행하였다. 18세 이상 성인 8,927명 중 본 분석에서 이용될 변수의 실제 응답자의 표본 8,716개를 사용하였다.

변수의 구성을 살펴보면 반응변수로는 간이 구강 위생지수(또는 치주질환 치료필요 지수)인 CPITN을 변환시켜 Y로 두었다. 반응변수에 이용된 간이 구강 위생 지표인 CPITN은 총 6분악 중 치주 조직 상태에 따라 점수를 표시한다. 이 지표에 따른 치주조직 검사 기준을 바탕으로 크기 3.5mm를 넘는 치주낭(CPITN에서는 3점 이상)을 가진 분악의 빈도($0 \le y \le 6$, 1인당)를 반응변수 Y로 보았다. 여기서 크기 3.5mm를 넘는 치주낭이 형성되면 치주염으로 판단하는데 구강위생 관리를 잘하면 3.5mm 이상의 치주낭은 쉽게 생기지 않는다. 따라서 반응변수는 영을 많이 포함하고 있다. 설명변수로는 치아우식경험 영구치 지수인 우식치(decay teeth, DT), 충전치(filled teeth, FT), 상실치(missing teeth, MT)를 X_1, X_2, X_3로 두었다. 설명변수 X_1은 우식치(decay teeth)의 빈도($0 \le DT \le 28$, 1인당)를 구한 값이고, X_2와 X_3는 충전치(filled teeth)와 상실치(missing teeth)의 빈도($0 \le FT, MT \le 28$, 1인당)를 구한 값이다. 그런데 X_1, X_2, X_3의 원래 관측단위를 사용하면 모수값이 지나치게 작게 나오므로 X_i의 단위를 1/100 로 축소하여 사용하였다.

그림 4.1은 전체 반응변수 Y (왼쪽)와 반응변수가 양의 값을 가질 때(오른 쪽)의 히스토그램을 그려보았다. 반응변수 Y를 살펴보면 영 값이 다른 값에 비해 약 74%로 월등히 높은 빈도를 차지하고 있음을 알 수 있고, 영을 제외한 Y의 히스토그램을 살펴보면 희박하게 발생하는 빈도의 형태라고 할 수 있다. 따라서 영을 제외한 Y가 앞에서 보여줬던 혼합 분포 중 하나인 포아송 분포를 따름을 알 수 있다.

표 4.1은 전체 Y와 $Y > 0$일 때 각각의 기초 통계량인데, Y의 중위수가 영이므로 절반 이상을 영 값이 차지하고 있음을 알 수 있고, Y와 $Y > 0$의 평균을 비교해 보면 각각 0.56,

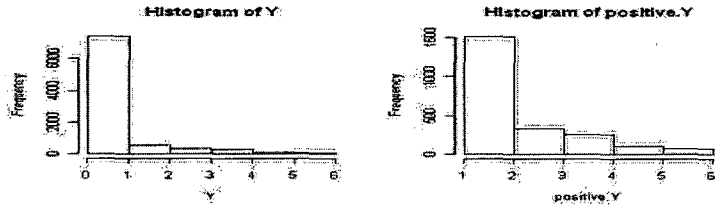

그림 4.1: 전체 반응변수와 반응변수가 양의 값을 가질 때의 히스토그램

표 4.1: 전체 반응변수와 반응변수가 양의 값을 가질 때의 기초통계량

	Y	$Y>0$
Min	0	1
1st Qu.	0	1
Median	0	2
3rd Qu.	1	3
Max	6	6
Total N	8716	2253
Mean	0.57	2.20
LCL Mean	0.55	2.15
UCL Mean	0.60	2.26
Variance	1.41	1.87
Std Dev.	1.19	1.37

표 4.2: 전체 반응변수와 반응변수가 양의 값을 가질 때의 정규성 검정

	Y		$Y>0$	
	통계량	유의확률	통계량	유의확률
Kolmogorov-smirnov	0.425	0.01	0.233	0.01
Cramer-von Mises	340.036	0.005	24.911	0.005
Anderson-Darling	1686.794	0.005	150.426	0.005

2.20로 전체 평균과 0을 제외한 평균이 큰 차이가 남을 알 수 있다.

표 4.2는 두 경우의 정규성 검정 결과이다. 모든 통계량의 p-value가 0.05보다 작으므로 정규성을 갖지 못하고, 표 4.1의 결과와 같이 비교해 보면 자료의 분포가 매우 치우쳐 있음을 알 수 있다.

4.2. 사후 추론

모수 ω와 β에 대한 사전분포는 다음과 같이 주도록 한다. Rodrigues(2003)가 설명변수가 없는 ZIP 모형에서 사용한 바와 같이 ω 에 대한 Jeffrey 무정보 사전분포(noninformative prior)를 가정하기 위해서 $c = d = 1/2$의 값을 주어 분석하였고, β의 사전분포인 정규분포의 평균(β_0)과 분산(Σ)으로는 과도한 영을 고려하지 않은 일반적인 포아송 회귀모형을 적용하여 최소제곱 추정법으로 구한 추정치를 이용하여 분석하였다.

3장에서 주어진 깁스 표본 기법과 메트로폴리스 알고리즘을 이용하여 총 510,000번 난수생성을 한 후 그 중 안정된 표본만을 사용하기 위해서 초기의 표본 10,000개를 제거(burn-in period)하고, 자기상관(auto correlation)의 발생을 막기 위해서, 추출된 표본 중 50번째 표본만을 선택(thinnig)하였다. 따라서 최종적으로 수집된 표본의 수는 10,000개이다. 그림 4.2 - 그림 4.7은 각 모수마다 Time Sequence plot(TS plot)을 그려서 얻어진 표본들이 잘 수렴

되는지를 확인하고, 각 모수의 표본의 주변 사후분포를 그려본 것이다. 그림을 보면 알 수 있듯이 대체적으로 수렴이 잘 되었음을 알 수 있다.

이제 각 모수들의 값을 추정해 보도록 한다. 각 모수마다 10,000개씩의 표본을 얻었으므로, 그의 평균을 모수의 추정값으로 사용하였다. S, ω의 주변 사후분포의 그림을 살펴보면 한쪽으로 심하게 치우쳐져 있다. 이에 대한 결과는 다음 표 4.3과 같다. 전체 조사자 8,716명 중에서 구강 위생관리를 잘한다고 판단되는 집단이 약 $\omega = 67\%$를 차지하고, 그렇지 못한 집단이 $1-\omega = 33\%$를 차지한다고 나타났다. 총 8,716명의 67%에 해당하는 사람은 5,839명인데 이는 S의 추정치 5,830명과 근접하여 잠재변수에 대한 모수 S가 실제 자료에 맞게 추정된 것을 알 수 있다. 또한 3.5mm이상의 치주낭 발생분악이 발견되지 않았던 사람은 $m = 6,466$인데 이중 $S = 5,830$이 차지하는 비율이 약 90%로서, 3.5mm이상의 치주낭 발생분악이 발견되지 않았던 사람 중 90%는 실제로 위생관리를 잘 한 집단이나 나머지 10%는 위생관리를 잘하지 않았으나 치주낭 발생분악이 발견되지 않았음을 알 수 있다.

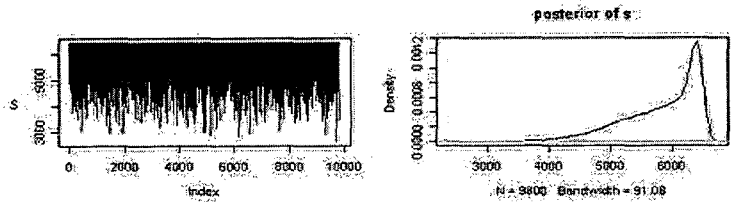

그림 4.2: S에 대한 time sequence plot과 주변 사후분포

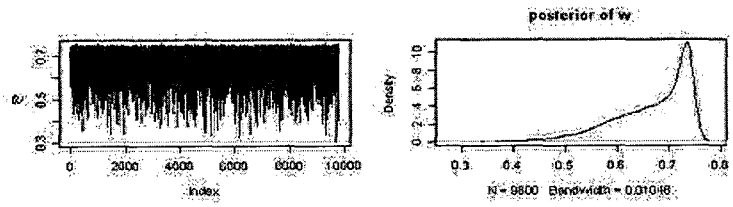

그림 4.3: ω에 대한 time sequence plot과 주변 사후분포

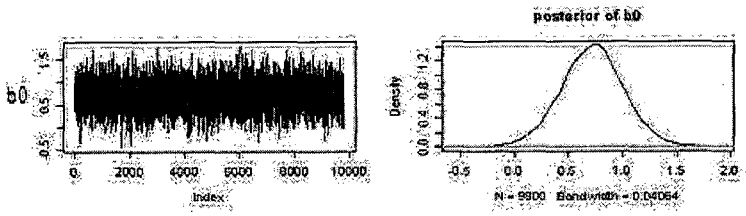

그림 4.4: β_0에 대한 time sequence plot과 주변 사후분포

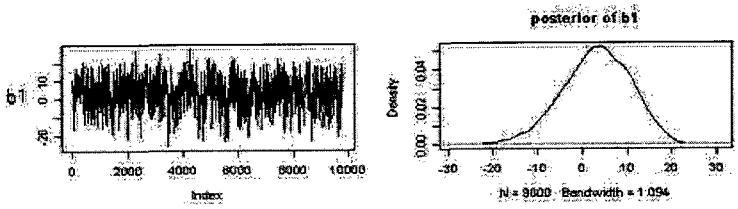

그림 4.5: β_1에 대한 time sequence plot과 주변 사후분포

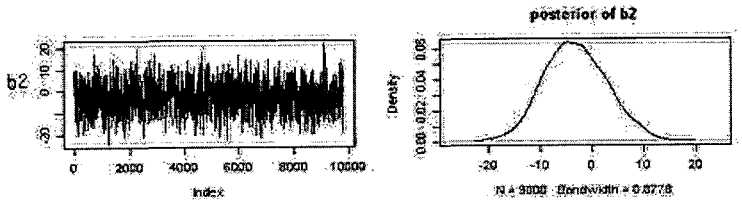

그림 4.6: β_2에 대한 time sequence plot과 주변 사후분포

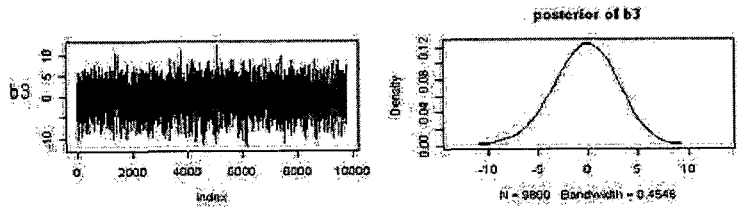

그림 4.7: β_3에 대한 time sequence plot과 주변 사후분포

표 4.3: 베이지안 추론법에 의한 모수 $S, \omega, \beta_0, \beta_1, \beta_2, \beta_3$의 추정 결과

	Mean	Std Err	95% C.I.	
			2.5%	97.5%
S	5830	6.5034	5817.25	5842.47
ω	0.6690	0.0023	0.6645	0.6734
β_0	0.7234	0.0029	0.7038	0.7434
β_1	2.9148	0.0793	2.3316	3.4979
β_2	-3.4214	0.0628	-3.5446	-3.2982
β_3	-0.3073	0.0321	-0.3702	-0.2445

모수 $\beta_0, \beta_1, \beta_2, \beta_3$를 살펴보고자 한다. 먼저 이 추정된 모수들이 유의한지 알아보았다. Wald 통계량 $z^2 = (\frac{\hat{\beta}}{ASE})^2$ 이 자유도 1을 갖는 카이제곱분포를 따름을 이용하여 $H_0 : \beta = 0$을 검정한 결과, 각각의 검정통계량 값은 62258.85, 1351.049, 2968.168, 91.646이고 p-value가 모두 0.001보다 작아서 유의한 결과를 나타냈다. 따라서 구강 위생을 잘 관리하지 못한 집단($1-\omega$의 확률을 갖는 집단)에서 관측된 치주낭 발생 분악수는 모수 θ인 포아송 분포를 따른다고 가정하였으므로, 그 집단에 한하여 포아송 회귀모형을 나타내면,

$$\begin{aligned} \log\hat{\theta} &= 0.7236 + 2.9148x_1 - 3.4214x_2 - 0.3073x_3 \\ \hat{\theta} &= \exp(0.7236 + 2.9148x_1 - 3.4214x_2 - 0.3073x_3) \end{aligned}$$

이다. 설명변수의 단위를 1/100로 축소했던 것을 복원하여 원래의 단위를 사용하면, 우식치(decay teeth)가 한 개 증가하면 3.5mm 이상 치주낭이 생긴 평균 분악수가 1.03($= e^{0.0292}$)배 증가하고, 그에 따른 신뢰구간은 $(\exp(0.023), \exp(0.035)) = (1.024, 1.036)$이다. 즉 우식치가 한 개 증가함에 따라 추정된 치주낭 발생 분악수는 평균적으로 약 3%증가하고 적게는 약 2%에서 많게는 약 4%까지 증가한다. 충전치(filled teeth)가 한 개 많아지면 약 0.97배가 되며 그에 따른 신뢰구간은 $(0.965, 0.968)$이다. 이는 충전치가 한 개 많아지면 평균적으로 3.5mm 이상인 치주낭 발생 분악수가 평균적으로 3% 줄어드는 것을 의미한다. 그리고 상실된 치아(missing teeth)가 한 개 증가하면 약 0.99배의 승법효과가 있어 약 1% 줄어든다. 이에 따른 신뢰구간은 $(0.996, 0.998)$이다.

그림 4.8은 영과잉 포아송 회귀모형에 대한 잔차의 히스토그램과 정규 Q-Q 그림이다. 이 모형 역시 잔차가 정규성을 갖지는 못하지만 일반적인 포아송 회귀모형에 대한 잔차의 히스토그램과 정규 Q-Q그림과 비교하여 보았을 때 영과잉 포아송 회귀모형의 잔차가 표준정규분포에 더 가까웠고, 정규 Q-Q 그림에서도 거의 직선에 가까운 형태를 나타내고 있다.

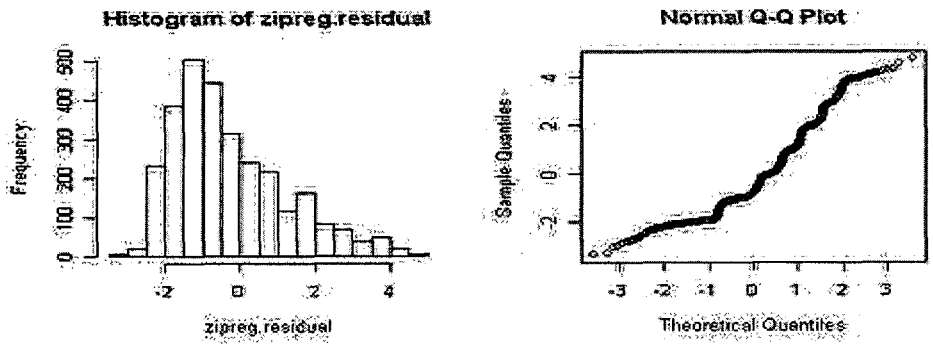

그림 4.8: ZIPR 모형에 대한 잔차의 히스토그램과 정규 Q-Q 그림

4.3. 다른 모형과의 비교

이 자료에 대하여 다중선형 회귀모형, 포아송 회귀모형, 베이지안 추론법을 적용한 영과잉 모형 (ZIP)과 영과잉 포아송 회귀모형(ZIPR)을 비교해 보도록 하자. 표 4.4는 각 회귀모형을 비교하기 위해 요약한 표이다. 포아송 회귀모형과 영과잉 포아송 회귀모형에서 $Y=0$일 때의 확률을 각각의 i에서 확률 Y_i의 평균으로, 즉 $\overline{Pr(Y=0)}=\frac{1}{n}\sum_{i=1}^{n}Pr(Y_i=0)$으로 구할 수 있다. 그 결과 고전적인 추론방법을 적용한 포아송 회귀모형의 $Pr(Y=0)$은 0.584로 추정되었고 베이지안 추론방법을 적용한 영과잉 포아송 회귀모형은 0.718로 추정되었다. 실제자료에서 영이 차지하는 비율이 0.74인 것과 비교해보면 베이지안 추론방법을 적용한 영과잉 포아송 회귀모형의 모수가 실제값에 매우 근접한 값으로 추정되었음을 알 수 있다. 추정치 옆에 괄호 안에 있는 구간은 95% 신뢰구간을 나타내는데 이 신뢰구간에서도 영과잉 포아송 회귀모형이 더 작은 신뢰구간의 범위를 가지고 있어 전통적인 추론법을 이용한 포아송 회귀모형의 모수보다 더 안정적이고 적절하다고 판단된다.

표준오차는 추정된 모수가 얼마나 정확한지 판단할 수 있는 도구인데, 포아송 회귀모형과 영과잉 포아송 회귀모형에서 $Pr(Y=0)$의 표준오차는 각각 0.00124와 0.000146으로 나와서 영과잉 포아송 회귀모형(ZIPR)이 월등히 작은 값을 갖는다. 그림 4.9는 포아송 회귀모형과 영과잉 포아송 회귀모형에서 $Pr(Y=0)$의 분포이다. 포아송 회귀모형보다 영과잉 포아송 회귀모형에서 큰 변동없이 평균값을 추정하고 있음을 보여준다.

또한 세 모형의 진단을 위해서 Pearson 적합도 통계량 $X^2=\sum\frac{(y_i-\hat{y}_i)^2}{\hat{y}_i}$을 구해서 비교해 보았다. 다중선형 회귀모형과 포아송 회귀모형은 영과잉 포아송 회귀모형과 비교하여 매우 큰 값을 가지고 있다. 즉, 영과잉 포아송 회귀모형에서 추정된 Y의 예측치가 실제자료와 가장 근접한 것을 알 수 있다.

표 4.4: 각 회귀모형에 대한 비교 요약

	Real Data	Linear Reg	Poisson Reg	ZIP Reg
$Pr(Y=0)$	0.74	-	0.584(0.582, 0.587)	0.718(0.717, 0.718)
Minimum $\hat{Y}$	0	-0.282	0.097	0
Maximum $\hat{Y}$	6	1.898	2.951	3.722
Pearson Chi-square	-	21089.151	21281.642	14667.498

다중선형 회귀모형에서는 회귀 관계의 존재와 회귀모수가 모두 유의했지만 잔차 분석의 측정치로 쿡의 거리(Cook's D)를 살펴보면, Y의 관측치가 2 이상인 값들을 모두 이상치(outlier)로 판단(전체자료의 약 10%)한다. 그리고 예측치(predicted value)의 최소값과 최대값이 각각 -0.28에서 1.9로, Y의 최대값이 6인 것에 비하면 예측력도 크게 떨어진다. 또한 포아송 회귀모형에서는 모델에 대한 적합도 검정이 유의하지 않았고, 잔차 분석의 측정치로서 수정잔차를 기준으로 살펴본 결과, Y의 관측치가 2이상인 값들을 이상치로 판단(전체자료의 약 11%)한다. 그리고 예측치의 최소값과 최대값이 각각 0.08부터 2.95로 다중선형 회귀모형보다는 범위에 맞게 예측이 되었지만, 영 과잉 자료에 대한 예측치의 최소값으로 0.08의 값을 갖는 것은 적절치 못하다. 반면 영과잉 포아송 회귀모형은 전체 자료의 약

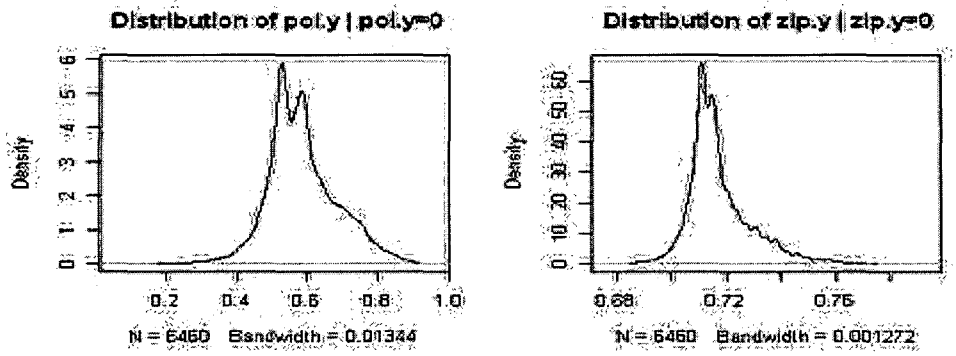

그림 4.9: 포아송 회귀모형과 ZIP 회귀모형에서 $Pr(Y_i = 0)$

2%만을 이상치로 판단하고 예측치의 최소값과 최대값으로 각각 0부터 3.722로 다른 모형들 보다 범위에 맞게 예측했다.

여기서 중요한 점은 다중선형 회귀모형과 포아송 회귀모형이 Y가 2 이상인 값들을 모두 이상치로 판단한다는 것이다. Y가 영을 과도하게 포함하고 있지만 다중선형 회귀모형과 포아송 회귀모형은 자료의 특성을 고려하지 않고 일괄적으로 분석한다. 이로 인하여 모수의 값을 더 작은 쪽으로 끌어당기려는 지렛대의 영향을 받게 되어 모수를 더욱 과장되게 추정하는 현상이 나타나는 것이다. 이에 반해 영과잉 포아송 회귀모형에 베이지안 추론 방법은 영과잉 자료의 특성을 고려하여 혼합모형으로 만들어 분석을 하기 때문에 이상치의 영향을 덜 받을 수 있다는 점에서 혼합모형의 베이지안 추론의 정확성을 알 수 있다.

표 4.5: 각 회귀모형에 대한 비교 요약

	ZIP		ZIP Reg	
	Mean	Std Err	Mean	Std Err
S	6043	0.27571	5830	6.5034
ω	0.6690	0.00006	0.6690	0.0023
θ	1.8602	0.00033	-	-
β_0	-	-	0.7236	0.0029
β_1	-	-	2.9148	0.0973
β_2	-	-	-3.4214	0.0628
β_3	-	-	-0.3073	0.0321

영과잉 포아송 모형(ZIP)과 영과잉 포아송 회귀모형(ZIPR)의 모수비교 결과인 표 4.5를 살펴보자. 참고로 ZIP 모형은 설명변수를 고려하지 않은 모형이고 ZIPR은 로그 연결함수를 이용하여 설명변수를 고려한 모형이다. 같은 모수에 대해서 추정된 표준오차가 설명변수가 추가되었을 때보다 설명변수가 없는 모형에서 표준오차 값이 더 작게 나타났다. 또한

두 모형에서 추정된 모수 ω는 설명변수가 있는 모형과 없는 모형, 모두 0.669로 같은 값으로 추정되었지만 S는 설명변수가 있는 모형에서 6043, 없는 모형에서 5830으로 약 200의 차이를 보인다. ω는 관측치가 두 집단에서 어느 한 집단에서 관측될 확률이었고 S는 잠재변수로서 영의 값을 갖는 모든 관측치 중에서 영만 있는 집단에서 관측된 빈도이므로 이 두 변수를 비교하여 모수의 정확성을 평가할 수 있다. 따라서 전체 관측치 8,716명 중에서 67%는 5,831명이므로, 설명변수가 있는 모형에서 추정된 S가 5830명으로서 실제 자료에 맞게 추정된 것을 알 수 있었다. 이는 설명변수를 추가하여 모형을 세우고 분석하여 자료의 결과를 더 정확하게 추정하였다고 할 수 있다.

5. 결론

셀 수 있는 이산 자료(discrete count data)에 대하여 주로 포아송 과정(poisson precess)을 적용시키지만 영이 과도하게 높은 비율을 차지하고 있는 영과잉 자료의 경우 포아송 분포를 따르지 못한다. 이런 자료들을 적합시키기 위해서 혼합모형(mixed model)이 제시되었고, 그 중에서 본 논문에서는 영과잉 포아송 회귀모형(ZIPR)을 베이지안 관점에서 분석하였다. 이 영과잉 자료에 대하여 네 모형을 비교하여 보았는데, 다중선형 회귀모형, 포아송 회귀모형과 혼합 모형인 영과잉 포아송 모형(ZIP)과 영과잉 포아송 회귀모형(ZIPR)으로 각 모수를 추정하였다. 다중선형 회귀모형과 포아송 회귀모형은 전통적인 추론 방법인 최소제곱법을 이용하였고 영과잉 포아송모형과 영과잉 포아송 회귀모형은 베이지안 관점에서 분석하여 네 모형에 의해 추론된 모수를 비교해 보았다.

그 결과 각 회귀모형의 모수들은 모두 유의하게 나타났다. 그러나 영이 아주 높은 비율을 차지하고 있어, 다중선형 회귀모형의 잔차에 대한 가정들을 만족하지 못했고 잔차의 산점도를 살펴보았을 때, 포아송 회귀모형에 더 적합한 듯 보였다. 그러나 포아송 회귀모형도 자료내의 큰 변동으로 인하여 과대산포 현상이 일어나서 영과잉 자료에 적합하지 않았다. 그리고 회귀모형들 중 가장 적합한 모형을 알아보기 위해 반응변수의 예측치를 구해서 Pearson 적합도 통계량으로 비교해본 결과, 베이지안 추론법을 적용한 영과잉 포아송 회귀모형에 의해 추정된 예측치와 실제 반응변수 값과 비교하였을 때 가장 근접한 것으로 나타났다. 또한 각 모형에서 추정된 예측치의 최소값과 최대값도 영과잉 포아송 회귀모형이 다른 모형에 비해서 범위에 가장 알맞게 추정되었다. 이는 다중선형 회귀모형과 포아송 회귀모형이 영을 과도하게 포함하고 있는 자료의 특성을 고려하지 않고, 이상치로 판단되는 값들에 의해 모수와 표준오차를 크게 추정하여, 그에 따른 예측치 또한 불안정하게 나타난 것이다. 따라서 구강 위생관리가 잘된 집단에 영에 대한 점확률을 가지는 분포와 잘 안된 집단에 포아송 분포를 합성하여 자료를 설명해줌으로써 과도하게 차지하고 있는 영의 부분을 보완해 주었고, 모형 안에서 각 모수에 대한 조건부 분포를 구하여 각각 모수를 추정함으로써 교호작용으로 인해 생길 수 있는 어려움을 베이지안 추론 방법을 적용하여 보다 안정적이고 정확한 추론을 하였다.

본 논문에서는 ZIPR 을 구성하는 점확률 분포와 포아송 분포 중 포아송 분포에만 설명변수를 도입하였다. 점확률분포의 확률 ω에도 로짓 연결함수를 사용하여 설명변수를 도입

하고 포아송 분포의 회귀계수 β 에 적용한 것 처럼 머트로폴리스 헤스팅스 기법을 이용하면 로짓모형의 회귀계수 추정이 가능할 것이다.

참고문헌

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

Dahiya, R. C. and Gross, A. J. (1973). Estimating the zero class from a truncated poisson sample, *Journal of the American Statistical Association* **68**, 731-733.

Ghosh S. K. and Mukhopadhyay P. (2005). Bayesian analysis of zero-inflated regression models, *Journal of statistical planning and inference* In press, 1-16.

Gupta, P. L. and Gupta, R. C. and Tripathi R. C. (1996). Analysis of zero-adjusted count data, *Computational Statistics and Data Analysis* **23**, 207-218.

Lambert, D. (1992). Zero-Inflated Poisson Regression With an application to Defects in Manufacturing, *Technometrics* **34**, 1-14.

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

Lord, D. and Washington, S. P. and Ivan, J. N.(2005). Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes:balancing statistical fit and theory, *Accident Analysis and Prevention* **37**, 35-46.

Rodrigues, R. (2003). Bayesian Analysis of Zero-Inflated Distributions, *Communications in Statistics* **32**, 281-289.

Umbach, D. (1981). On inference for a mixture of Poisson and a degenerate distribution, *Communications in Statistics: theory and methods* **10**, 299-306.

Yip, P. (1988). Inference about the mean of a Poisson distribution in the presence of nuisance parameter, *Australian Journal of Statistics* **30**, 299-306.

[2005년 12월 접수, 2006년 7월 채택]

Bayesian Analysis of a Zero-inflated Poisson Regression Model: An Application to Korean Oral Hygienic Data*

Ah-Kyoung Lim[1] Man-Suk Oh[2]

ABSTRACT

We consider zero-inflated count data, which is discrete count data but has too many zeroes compared to the Poisson distribution. Zero-inflated data can be found in various areas. Despite its increasing importance in practice, appropriate statistical inference on zero-inflated data is limited. Classical inference based on a large number theory does not fit unless the sample size is very large. And regular Poisson model shows lack of fit due to many zeroes. To handle the difficulties, a mixture of distributions are considered for the zero-inflated data. Specifically, a mixture of a point mass at zero and a Poisson distribution is employed for the data. In addition, when there exist meaningful covariates selected to the response variable, loglinear link is used between the mean of the response and the covariates in the Poisson distribution part. We propose a Bayesian inference for the zero-inflated Poisson regression model by using a Markov Chain Monte Carlo method. We applied the proposed method to a Korean oral hygienic data and compared the inference results with other models. We found that the proposed method is superior in that it gives small parameter estimation error and more accurate predictions.

Keywords: Poisson regression model; Zero inflated data; Mixture model; Monte Carlo.

* This work was supported by grant No.(R06-2002-012-01002-0) from the Basic Research Program of the Korea Science and Engineering Foundation

1) Graduate Student, Dept. of Statistics, Ewha Womans University, Seoul 120-750, Korea
E-mail: lak612@hanmail.net

2) (Corresponding author) Professor, Dept. of Statistics, Ewha Womans University, Seoul 120-750, Korea
E-mail: msoh@mm.ewha.ac.kr