응용통계연구 (2008)
21(4), 603–613

영과잉 회귀모형에 대한 베이지안 분석

장학진[1] · 강윤희[2] · 이수범[3] · 김성욱[4]

[1]한양대학교 응용수학과; [2]한양대학교 응용수학과;
[3]서울시립대학교 교통공학과; [4]한양대학교 응용수학과

(2008년 2월 접수, 2008년 5월 채택)

요 약

셀 수 있는 이산 자료 중에서 일반적인 모형에 비하여 영의 빈도가 과도하게 많이 관측되는 자료가 있다. 이러한 경우에 포아송 또는 음이항회귀모형과 같은 일반적인 회귀모형에 의한 분석은 적절하지 못하다. 본 논문에서는 영과잉 포아송회귀모형과 영과잉 음이항회귀모형에 대하여 베이지안 분석을 하였다. 또한, 마코브 연쇄 몬테카를로 방법으로 계산한 베이즈 요인을 이용하여 모형선택을 하였다. 실제 교통사고 자료를 분석하여 이론적인 결과들을 뒷받침하였다.

주요용어: 영과잉 모형, 베이지안 모형선택, 베이즈 요인, 마코브 연쇄 몬테카를로.

1. 서론

셀 수 있는 이산형 자료에 대한 분석모형으로서 포아송회귀모형이 가장 널리 적용된다. 그러나 실제로 시행횟수 혹은 발생건수들의 분포는 흔히 분산이 평균보다 큰 경우가 많다. 포아송 분포의 경우 평균과 분산이 동일하므로 이와 같이 과분산된 자료에 대한 회귀모형으로서는 적절하지 않다. 이러한 경우에는 종종 포아송회귀모형 대신 음이항회귀모형을 사용한다.

교통사고 자료를 이용한 사고 예측 모형에 있어 일반 선형회귀모형이 음의 사고수를 예측한다는 문제점을 해결하기 위하여 Jovanis와 Chang (1986), Joshua와 Garber (1990) 등은 이산적 확률변수로 해석하는 포아송회귀모형을 도입하였다. 또한 분석 모형에 있어 실제 사고건수의 경우 분산이 평균보다 큰 과분산 문제가 발생하므로 분산과 평균이 같다는 기본 전제조건을 가지는 포아송회귀모형의 문제점을 해결하기 위하여 Miaou와 Lum (1993)은 분산이 평균보다 크다는 가정에서 출발하는 음이항회귀모형을 주장하였다. 이 두 모형은 사고빈도를 예측하는데 많이 사용되어져 왔었다. Shankar 등 (1995)은 포아송회귀모형과 음이항회귀모형으로 사고예측모형을 구축하고 반응변수는 사고건수 전체, 사고유형별 사고건수로 분리하여 모형을 구축하였다. Poch와 Mannering (1996)은 교차로 접근사고예측모형에서 음이항회귀모형의 적절성을 보였고, Milton과 Mannering (1998)도 도로기하구조, 교통관련요인과 사고건수 추정모형개발에서 음이항회귀모형이 포아송회귀모형보다 잘 적합함을 보였다.

이 논문은 한국과학재단 특정기초연구사업(R01-2005-000-10141-0) 지원으로 연구되었음.
[1](426-791) 경기도 안산시 상록구 사3동 1271, 한양대학교 응용수학과, E-mail: hjjang@hanyang.ac.kr
[2](426-791) 경기도 안산시 상록구 사3동 1271, 한양대학교 응용수학과, E-mail: skylark84@naver.com
[3](130-743) 서울특별시 동대문구 전농동 90, 서울시립대학교 교통공학과, 부교수. E-mail: mendota@uos.ac.kr
[4]교신저자: (426-791) 경기도 안산시 상록구 사3동 1271, 한양대학교 응용수학과, 부교수.
E-mail: seong@hanyang.ac.kr

그러나 기존의 포아송회귀모형이나 음이항회귀모형과 같은 전통적인 추론방법을 이용한 모형들은 어떤 특정한 값에서 기대도수보다 관측도수가 많이 나타나는 자료에 대하여 잘 설명할 수 없다. 교통사고 자료의 경우 영이 과도하게 많은 자료이다. 즉 영의 비율이 기존의 포아송회귀모형 또는 음이항회귀모형에 의해 기대되는 영의 비율보다 높게 관측되며, 이러한 자료를 영과잉 자료라고 한다.

Shankar 등 (1997)은 전통적인 포아송회귀모형이나 음이항회귀모형이 사고가 "0"인 지점을 적절히 설명하지 못하는 한계가 있다고 주장하였다. 이러한 단점을 보완하기 위하여 분포의 혼합 형태를 가지는 영과잉 포아송(Zero-Inflated Poisson: ZIP)회귀모형과 영과잉 음이항(Zero-Inflated Negative Binomial: ZINB)회귀모형이 제안되었다. 이러한 영과잉 모형은 자료에서 과도하게 나타나는 영을 설명하기 위하여 영에 대한 점확률을 가지는 분포를 기존의 분포와 합성한 것이다.

영과잉 모형에 대한 분석은 주로 빈도론자(frequentist)들의 관심에서 수행되어 왔다. 그러나 빈도론자 또는 고전적 추론은 대표본 근사를 사용하므로 영과잉 자료와 같이 매우 치우친 분포를 갖는 경우 표본크기가 매우 크지 않으면 추정치의 신뢰도가 떨어지는 단점이 있다 (*cf.* 임아경과 오만숙, 2006).

베이지안 분석 방법은 기존의 고전적인 분석 방법에 비해 유용한 사전 정보의 사용을 가능하게 할 뿐 아니라 표본의 크기가 작을 때에 상대적으로 더 신뢰성 있는 분석을 할 수 있는 장점을 가지고 있다. 이 추론 과정은 대부분의 경우에 복잡한 적분 계산을 요구하고 있기 때문에 우도함수와 사전분포함수가 짝관계가 성립하는 몇몇 경우에 대해서만 국한적으로 사용되어 왔으며 실제 자료의 분석에는 많은 어려움이 있었다. 그러나 최근 들어 사후분포의 기대값을 수치적으로 구할 수 있는 마코브 체인 몬테카를로(Markov Chain Monte Carlo: MCMC) 방법이 널리 사용됨에 따라 베이지안 분석 방법은 여러 분야에서 사용되게 되었다. MCMC 방법은 모수들의 완전 조건부 사후분포들로부터 랜덤변량 생성을 통하여 모수들의 추정값을 구할 수 있다.

본 연구에서는 ZIP 회귀모형과 ZINB 회귀모형에 대하여 베이지안 추론을 이용한 사고예측 모형을 제안하며 이 모형들을 이용하여 베이지안 모형 선택을 하고자 한다.

본 논문의 구성은 다음과 같다. 먼저 2장에서는 ZIP 회귀모형과 ZINB 회귀모형을 기술하고, 각 모형에 대한 사전분포를 제시한 후 사후분포를 살펴본다. 3장에서는 베이즈 요인을 소개하고, 사전분포의 가정 하에 ZIP 회귀모형과 ZINB 회귀모형의 모형 선택을 위한 베이즈 요인의 계산법을 설명한다. 4장에서는 영과잉 모형에 대한 베이지안 추론 방법을 제안한다. 5장에서는 제안된 베이지안 기법을 사용하여 실제 교통 사고자료에 대하여 분석을 수행하고, 베이지안 모형 선택의 과정을 수행해 본다. 6장에서는 이 연구의 결론을 보일 것이다.

2. 영과잉 사고예측모형

2.1. 영과잉 모형의 우도함수

개체 i가 p개의 설명변수 $\mathbf{X}_i' = (X_{i1}, X_{i2}, \ldots, X_{ip})$와 음이 아닌 정수값을 갖는 반응변수 Y_i를 가질 때 Y_i는 ω의 확률로 0에 대한 점확률을 가지는 분포를 따르며, $1-\omega$의 확률로 포아송분포 또는 음이항 분포를 따른다고 하자.

$\mathbf{y} = (y_1, \ldots, y_n)$가 주어질 때, ZIP 회귀모형과 ZINB 회귀모형의 우도함수는 각각,

$$L(\beta, \omega) = \prod_{i=1}^{n} \Big[I_{(y_i=0)} \left\{ \omega + (1-\omega) \exp\left\{ -\exp(\mathbf{X}_i'\beta) \right\} \right\} + I_{(y_i>0)} (1-\omega) \frac{\exp\left\{ -\exp(\mathbf{X}_i'\beta) \right\} \exp\left\{ y_i (\mathbf{X}_i'\beta) \right\}}{y_i!} \Big],$$

$$L(\beta,\omega,\theta) = \prod_{i=1}^{n}\left[I_{(y_i=0)}\left\{\omega+(1-\omega)\left(\frac{\theta}{\theta+\exp(\mathbf{X}_i'\beta)}\right)^{\theta}\right\}\right.$$
$$\left.+I_{(y_i>0)}(1-\omega)\frac{\Gamma(\theta+y_i)}{\Gamma(\theta)\Gamma(y_i+1)}\left(\frac{\theta}{\theta+\exp(\mathbf{X}_i'\beta)}\right)^{\theta}\left(\frac{\exp(\mathbf{X}_i'\beta)}{\theta+\exp(\mathbf{X}_i'\beta)}\right)^{y_i}\right]$$

이다. 여기서 $I_{(\cdot)}$는 지시함수로서 사건이 사실일 때는 1이고, 다른 경우는 0의 값을 갖는다. 또한, $0 \le \omega \le 1$, $\lambda_i = \exp(\mathbf{X}_i'\beta)$이며, β는 $p\times 1$ 미지의 모수벡터, $\theta\,(\theta>0)$는 과분산계수의 역수이다.

2.2. 사전분포와 사후분포

ZIP 회귀모형의 두 모수 ω, β에 각각 독립적으로 ω에 대하여 초모수 δ_0와 η_0를 갖는 베타분포를 가정하고, β에 대해서는 평균이 μ이고 분산이 $\mathbf{\Sigma}$인 다변량 정규분포를 가정한다. 따라서 사전분포는

$$\pi(\omega) = \frac{\Gamma(\delta_0+\eta_0)}{\Gamma(\delta_0)\Gamma(\eta_0)}\omega^{(\delta_0-1)}(1-\omega)^{(\eta_0-1)},$$
$$\pi(\beta) = \frac{1}{(2\pi\mathbf{\Sigma})^{\frac{1}{p}}}\exp\left\{-\frac{1}{2}(\beta-\mu)'\mathbf{\Sigma}^{-1}(\beta-\mu)\right\}, \quad \beta\in\mathbf{R}^p$$

이다.

ZINB 회귀모형에 대해서는 세 모수 ω, β와 θ에 각각 독립적으로 ω와 β에 대해선 ZIP 회귀모형과 같은 베타분포와 다변량 정규분포를 가정하고, θ에 대해선 균일분포를 가정한다.

각각의 우도함수에 대하여 사전분포를 결합하면 두 모형의 결합사후분포는 각각 다음과 같다.

$$\pi(\beta,\,\omega\,;\,\mathbf{y}) \propto \prod_{i=1}^{n}\left[I_{(y_i=0)}\left\{\omega+(1-\omega)\exp\left\{-\exp(\mathbf{X}_i'\beta)\right\}\right\}\right.$$
$$\left.+I_{(y_i>0)}(1-\omega)\frac{\exp\left\{-\exp(\mathbf{X}_i'\beta)\right\}\exp\left\{y_i(\mathbf{X}_i'\beta)\right\}}{y_i!}\right]$$
$$\times\omega^{(\delta_0-1)}(1-\omega)^{(\eta_0-1)}\cdot\exp\left\{-\frac{1}{2}(\beta-\mu)'\mathbf{\Sigma}^{-1}(\beta-\mu)\right\},$$
$$\pi(\beta,\,\omega,\,\theta\,;\,\mathbf{y}) \propto \prod_{i=1}^{n}\left[I_{(y_i=0)}\left\{\omega+(1-\omega)\left(\frac{\theta}{\theta+\exp(\mathbf{X}_i'\beta)}\right)^{\theta}\right\}\right.$$
$$\left.+I_{(y_i>0)}(1-\omega)\frac{\Gamma(\theta+y_i)}{\Gamma(\theta)\Gamma(y_i+1)}\left(\frac{\theta}{\theta+\exp(\mathbf{X}_i'\beta)}\right)^{\theta}\left(\frac{\exp(\mathbf{X}_i'\beta)}{\theta+\exp(\mathbf{X}_i'\beta)}\right)^{y_i}\right]$$
$$\times\omega^{(\delta_0-1)}(1-\omega)^{(\eta_0-1)}\cdot\exp\left\{-\frac{1}{2}(\beta-\mu)'\mathbf{\Sigma}^{-1}(\beta-\mu)\right\}.$$

3. 베이지안 모형선택

3.1. 베이즈 요인

자료 $\mathbf{y}$가 주어졌을 때, q개의 서로 다른 모형 $M_1, M_2, \ldots, M_q$ 중 가장 그럴듯한 모형을 선택하고자 한다. 모형 $M_i\,(i=1,2,\ldots,q)$하에서 자료의 모수적 밀도함수를 $f_i(\mathbf{y};\,\theta_i)$라 하고, Θ_i를 θ_i의 모수공간이

표 **3.1.** 베이즈 요인의 해석

B_{ji}	모형 M_i에 비해 모형 M_j가 좋은 증거
$1.0 \sim 3.2$	선호하지만, 언급할 가치는 없다
$3.2 \sim 10$	긍정적으로 볼 수 있다
$10 \sim 100$	강하다
> 100	결정적이다

라 하자. 그리고 $\pi_i(\theta_i)$를 θ_i의 사전분포, $p(M_i)$를 모형 M_i의 사전확률이라 하자. 그러면 M_i의 사후확률은

$$P(M_i; \mathbf{y}) = \left[\sum_{j=1}^{q} \frac{p(M_j)}{p(M_i)} B_{ji}\right]^{-1}$$

로 주어진다. 여기서 B_{ji}는 모형 M_i에 대한 M_j의 베이즈 요인이며, 아래와 같이 정의 된다.

$$B_{ji}(\mathbf{y}) = \frac{m_j(\mathbf{y})}{m_i(\mathbf{y})} = \frac{\int_{\Theta_j} f_j(\mathbf{y}; \theta_j)\pi_j(\theta_j)d\theta_j}{\int_{\Theta_i} f_i(\mathbf{y}; \theta_i)\pi_i(\theta_i)d\theta_i},$$

이 때 $m_i(\mathbf{y})$는 모형 M_i하에서 $\mathbf{y}$의 주변(marginal) 혹은 예측(predictive)함수를 뜻한다. 또한, $B_{ii} = 1$이며 $B_{ij} = B_{ji}^{-1}\ (i, j = 1, \ldots, q)$임을 알 수 있다.

자료가 M_1과 M_2모형 중 어떤 모형을 지지하는지는 M_1과 M_2의 사후비율로 측정할 수 있다. 즉 비율은 사후확률로부터 다음의 식이 성립한다.

$$\frac{P(M_2; \mathbf{y})}{P(M_1; \mathbf{y})} = \frac{p(M_2)m_2(\mathbf{y})}{p(M_1)m_1(\mathbf{y})} = \frac{p(M_2)}{p(M_1)} B_{21}(\mathbf{y}).$$

베이지안 입장에서 볼 때 모형 선택은 직관적이라 할 수 있다. 모형을 견줘볼 때 자료의 지지도를 알아보는 방법으로 통상 베이즈 요인을 이용한다. Jeffreys (1961)는 베이즈 요인을 계산하고 해석하는데 지침을 주었다. 그것은 표 3.1과 같다.

3.2. 모형선택을 위한 베이즈요인의 계산

주변함수는 몇몇 간단한 적분의 경우 해석적(analytic)으로 계산된다. 그러나 종종 해석적인 방법으로 계산이 어려운 경우에 있어서는 수치적인 방법을 이용하여야 한다. 수치적으로 이용 가능한 소프트웨어가 개발되어 있지만 일반적으로 이러한 적분에는 비효율적이기 때문에 사용을 많이 하지 않는다. 비효율적인 이유로서 첫 번째로 표본수가 적당히 클 때 적분값은 최대값 근처에서 치솟지만, 수치적인 방법으로는 최대값에 대한 어떠한 사전 지식 없이 피적분함수의 질량이 축적되는 곳을 찾기 어렵다. 두 번째로 차원이 높기 때문이다. 따라서 이러한 경우에 적분을 필요로 하지 않는 베이지안 방법인 몬테카를로 적분 방법을 통계적 방법에 맞게 조정하여 사용한다. 주변함수에 대한 간단한 몬테카를로 적분 추정 값은 다음과 같다.

$$\hat{m}(\mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n} f\left(\mathbf{y}; \theta^{(i)}\right),$$

여기서 $\theta^{(i)}(i = 1, \ldots, n)$는 사전분포의 표본이며, 이 식은 모수 값에 의해 추출된 우도함수의 평균이다. 이는 Raftery와 Banfield (1991)에 의해 처음으로 언급되었으며, McCulloch와 Rossi (1991)에 의해 몇몇 특별한 경우에 대한 자세한 연구가 이루어 졌었다.

주변함수의 산정에 있어 우도함수 값이 작아서 생기는 비효율성을 줄이기 위하여 우도함수의 역함수를 이용한 방법을 사용한다 (*cf.* Newton과 Raftery, 1994). 이때, 주변함수의 추정값은 다음과 같다.

$$\hat{m}(\mathbf{y}) = \left\{ \frac{1}{n} \sum_{i=1}^{n} f^{-1} \left(\mathbf{y}; \theta^{(i)} \right) \right\}^{-1},$$

여기서, n의 값이 무한대에 가까워질수록 $\hat{m}(\mathbf{y})$의 값은 실제 값인 $m(\mathbf{y})$에 수렴한다. 따라서 이러한 주변함수의 추정값을 이용하여 베이즈 요인을 계산한다.

4. 추정 절차

4.1. 깁스 샘플러

베이지안 추론에서 널리 사용되는 MCMC 기법 중 하나인 깁스 샘플러(Gibbs Sampler)는 Geman과 Geman (1984)에 의해 제안되었다. 그들은 이미지 분석과 관련하여 격자의 깁스 분포를 분석하기 위해 표집을 사용하였다. Gelfand와 Smith (1990)에서는 통계적 모형에 잘 맞는지에 대한 접근법과 이전에 추정이 가능하지 않았던 사후분포의 추정의 도구로 어떻게 MCMC 방법을 베이지안 통계학에서 사용할 것인지 설명했다.

모수 θ를 $1 \times p$ 미지의 모수벡터라고 가정하자. 결합 사후확률분포(joint posterior distribution)로부터 추출하고자 하는 하나의 모수 $\theta_i \, (i = 1, \ldots, p)$ 만의 분포를 고려하자. 단, 나머지 모수들은 고정된 값으로 간주한다. 그러면 θ_i에 대응하는 확률분포 $\pi_i \, (\theta_i; \, \theta_{-i}, \, \mathbf{y})$로부터 랜덤변량을 생성 할 수 있다. 이때 하나의 모수 θ_i만의 분포인 $\pi_i \, (\theta_i; \, \theta_{-i}, \, \mathbf{y})$는 완전 조건부 사후분포라 한다. 현재 θ의 상태를 $\theta^{(\mathbf{t})} = (\theta_1^{(t)}, \ldots, \theta_p^{(t)})$이라 한다면, 깁스 샘플러 알고리즘은 아래와 같다.

1단계 : $\pi_1(\theta_1; \, \theta_2^{(t)}, \ldots, \theta_p^{(t)}, \, \mathbf{y})$ 분포로부터 $\theta_1^{(t+1)}$를 생성한다.

2단계 : $\pi_2(\theta_2; \, \theta_1^{(t+1)}, \, \theta_3^{(t)}, \ldots, \theta_p^{(t)}, \, \mathbf{y})$ 분포로부터 $\theta_2^{(t+1)}$를 생성한다.

$$\vdots$$

p단계 : $\pi_p(\theta_p; \, \theta_1^{(t+1)}, \ldots, \theta_{p-1}^{(t+1)}, \, \mathbf{y})$ 분포로부터 $\theta_p^{(t+1)}$를 생성한다.

만일 반복수(replication) n이 적당히 크면, θ_i의 사후평균은

$$\hat{\theta}_i = \frac{1}{n} \sum_{j=1}^{n} \theta_i^{(j)}$$

을 이용하여 근사시킬 수 있다. n번의 랜덤변량을 생성 한 후 그 중 안정된 표본만을 사용하기 위해 초기의 표본 몇 개는 제거(burn-in)한다. 알고리즘에서 모수의 업데이트 순서는 고정한다. 깁스 샘플러는 많은 방법으로 생성될 가능성이 있기 때문에 순서를 고정하는 것이 모든 경우에 항상 성립하는 것은 아니며, 무작위로 업데이트 순서를 가정하는 것도 가능하다.

4.2. 완전 조건부 사후분포

주어진 모수들에 대한 사후추정이 필요하나 식이 간단하지 않아 수리적으로 가능하지 않으므로 MCMC 기법 중 하나인 깁스 샘플러를 사용하고자 한다. 깁스 샘플러 기법을 사용하기 위해서는 각각의 모수에 대해 완전 조건부 사후분포가 필요하다. 각각의 모수에 대한 완전 조건부 사후분포는 결합사후분포에서 각 모수에 의존하지 않는 요소는 무시하고 얻은 함수이다.

ZIP 회귀모형의 모수인 β와 ω에 대한 완전 조건부 사후분포는 다음과 같이 나타낼 수 있다. 즉,

$$\begin{aligned}\pi(\beta_l;\,\beta_{-l},\,\omega,\,\mathbf{y}) \propto \prod_{i=1}^{n}\Bigg[& I_{(y_i=0)}\left\{\omega+(1-\omega)\exp\left\{-\exp(\mathbf{X}_i'\beta)\right\}\right\} \\ & +I_{(y_i>0)}(1-\omega)\frac{\exp\left\{-\exp(\mathbf{X}_i'\beta)\right\}\exp\left\{y_i(\mathbf{X}_i'\beta)\right\}}{y_i!}\Bigg] \\ & \times\exp\left\{-\frac{1}{2}(\beta-\mu)'\boldsymbol{\Sigma}^{-1}(\beta-\mu)\right\}\end{aligned}$$

이고

$$\begin{aligned}\pi(\omega;\,\beta,\,\mathbf{y}) \propto \prod_{i=1}^{n}\Bigg[& I_{(y_i=0)}\left\{\omega+(1-\omega)\exp\left\{-\exp(\mathbf{X}_i'\beta)\right\}\right\} \\ & +I_{(y_i>0)}(1-\omega)\frac{\exp\left\{-\exp(\mathbf{X}_i'\beta)\right\}\exp\left\{y_i(\mathbf{X}_i'\beta)\right\}}{y_i!}\Bigg] \\ & \times\omega^{(\delta_0-1)}(1-\omega)^{(\eta_0-1)}\end{aligned}$$

이다. 여기서, β_l은 모수 β에서 l번째 요소를 나타낸다. 반면 β_{-l}은 β에서 l번째 요소만 제외한 $(p-1)\times 1$ 벡터로써, 즉, $\beta_{-l}=(\beta_1,\ldots,\beta_{l-1},\,\beta_{l+1},\ldots,\beta_p)'$이다.

또한, ZINB 회귀모형의 모수인 β, ω와 θ에 대한 완전 조건부 사후분포는 다음과 같이 나타낼 수 있다.

$$\begin{aligned}\pi(\beta_l;\,\beta_{-l},\,\omega,\,\theta,\mathbf{y}) \propto & \prod_{i=1}^{n}\left[I_{(y_i=0)}\left\{\omega+(1-\omega)\left(\frac{\theta}{\theta+\exp(\mathbf{X}_i'\beta)}\right)^{\theta}\right\}+I_{(y_i>0)}(1-\omega)A_i\right] \\ & \times\exp\left\{-\frac{1}{2}(\beta-\mu)'\boldsymbol{\Sigma}^{-1}(\beta-\mu)\right\}, \\ \pi(\omega;\,\beta,\,\theta,\,\mathbf{y}) \propto & \prod_{i=1}^{n}\left[I_{(y_i=0)}\left\{\omega+(1-\omega)\left(\frac{\theta}{\theta+\exp(\mathbf{X}_i'\beta)}\right)^{\theta}\right\}+I_{(y_i>0)}(1-\omega)A_i\right] \\ & \times\omega^{(\delta_0-1)}(1-\omega)^{(\eta_0-1)}, \\ \pi(\theta;\,\beta,\,\omega,\,\mathbf{y}) \propto & \prod_{i=1}^{n}\left[I_{(y_i=0)}\left\{\omega+(1-\omega)\left(\frac{\theta}{\theta+\exp(\mathbf{X}_i'\beta)}\right)^{\theta}\right\}+I_{(y_i>0)}(1-\omega)A_i\right],\end{aligned}$$

여기서

$$A_i=\frac{\Gamma(\theta+y_i)}{\Gamma(\theta)\Gamma(y_i+1)}\left(\frac{\theta}{\theta+\exp(\mathbf{X}_i'\beta)}\right)^{\theta}\left(\frac{\exp(\mathbf{X}_i'\beta)}{\theta+\exp(\mathbf{X}_i'\beta)}\right)^{y_i}$$

이다. 이러한 각각의 모수들에 대한 완전 조건부 사후분포를 구하여 깁스 샘플러 기법을 이용한다.

5. 실제 교통사고자료의 분석

5.1. 자료탐색

2장에서 제안된 ZIP 회귀모형과 ZINB 회귀모형에 대한 베이지안 추정법을 사용하여 2002년부터 2005년까지 총 4년간 발생한 사고건수의 자료를 분석하였다. 또한 두 모형에 대한 베이즈 요인을 이용하여 베이지안 모형 선택의 과정을 수행하였다. 사고자료는 클로버 인터체인지의 루프연결로에서 얻은 27개의 자료이다.

표 **5.1.** 전체 종속변수 $\mathbf{Y}$와 $\mathbf{Y} > 0$일 때의 기초통계량

	$\mathbf{Y}$	$\mathbf{Y} > 0$
Total N	27	11
Mean	0.704	1.727
95% CL Mean	(0.296, 1.111)	(1.120, 2.335)
Variance	1.063	0.818
Std Dev.	1.031	0.905
Median	0	2

변수의 구성을 살펴보면 종속변수 $\mathbf{Y}$는 인터체인지의 루프연결로에서 2002년부터 2005년까지 발생한 사고건수의 총합이다. 각 인터체인지에서 발생하는 사고는 많지 않으므로, 종속변수는 영을 많이 포함하고 있다. 독립변수로는 곡률 차(curvature gap), 속도 차(speed difference), 연결로타입(ramp type), 교통량(traffic volume)을 각각 $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4$로 두었다. 곡률 차는 곡선이 2개로 구성된 루프연결로의 각각의 곡률에 대한 차를 사용하였다. 속도 차는 인터체인지의 연결로와 고속도로 본선의 제한속도의 차이를 사용하였다. 원래 관측치를 사용하면 모수 추정치가 작아지므로 1/10으로 축소하여 사용하였다. 연결로 타입은 유출입여부를 나타내는 가변수로 연결로의 유출입여부에 따라, 유입연결로의 경우 "1", 유출연결로의 경우 "0"으로 처리하여 사용하였다. 교통량은 인터체인지의 연결로 교통량에 log값을 취한 교통량을 사용하였다.

표 5.1은 전체 종속변수 $\mathbf{Y}$와 $\mathbf{Y} > 0$일 때 각각의 기초 통계량이다. 총 자료의 개수를 비교해 보면 16개의 지점에서 0이 나타나고 있음을 확인할 수 있으며 이는 전체자료의 절반이 넘는 개수이다. 평균을 비교해 보면 각각 0.704, 1.727로 0을 포함하고 있을 때와 포함하고 있지 않을 경우 큰 차이가 남을 알 수 있다.

5.2. 분석결과

영과잉 모형의 모수들에 대한 사전분포에 대하여 ω의 사전분포인 베타분포의 초모수 (δ_0, η_0)에 $(5, 5)$값을 주어 분석하고, β의 사전분포인 다변량 정규분포의 초모수 평균(μ)과 공분산($\mathbf{\Sigma}$)에는 일반적인 포아송 및 음이항회귀모형을 적용하여 추정한 추정치를 이용하여 분석하였다.

4장에서 주어진 깁스 샘플러 기법을 이용하여 총 11,000번 랜덤변량을 생성 한 후 그 중 안정된 표본만을 사용하기 위해서 초기의 표본 1,000개를 제거(burn-in)하였다. 따라서 최종적으로 수집된 표본의 수는 10,000개이다. 그림 5.1과 5.2는 ZIP 회귀모형과 ZINB 회귀모형의 각 모수들의 얻어진 표본을 그린 것이다. 그림을 보면 알 수 있듯이 대체적으로 수렴이 잘 되었음을 알 수 있다.

각 모수마다 10,000개씩의 표본을 얻었으므로, 그의 평균을 모수의 추정값으로 사용하였다. 모수의 추정값과 표준오차, 모수에 대해 95% HPD(highest posterior density)구간을 구한 값이 표 5.2에 나와 있다. 자료 27개중에 16개가 0의 값을 갖기 때문에, 실제자료에서의 0의 비율은 0.5926이다. 추정치를 살펴보면 ZIP 회귀모형의 경우 0.5374이며, 이는 실제 자료 중 53.7%가 0에 대한 점확률을 가지는 자료임을 의미한다. ZINB 회귀모형의 경우가 0.5239로 추정되어 실제 자료 중 52.4%가 0에 대한 점확률을 가지는 자료임을 알 수 있다.

추정된 모수값을 이용하여 영과잉 모형에서의 종속변수가 0인 자료에 대한 예측력을 일반적인 포아송회귀모형이나 음이항회귀모형과 비교하였다. 예측력을 평가하는데는 여러 가지 방법이 있다. 그 중 일반적인 절대오차평균(average absolute error: AAE)을 비교해보았다 (*cf.* Szabo와 Khoshgoftaar, 2000).

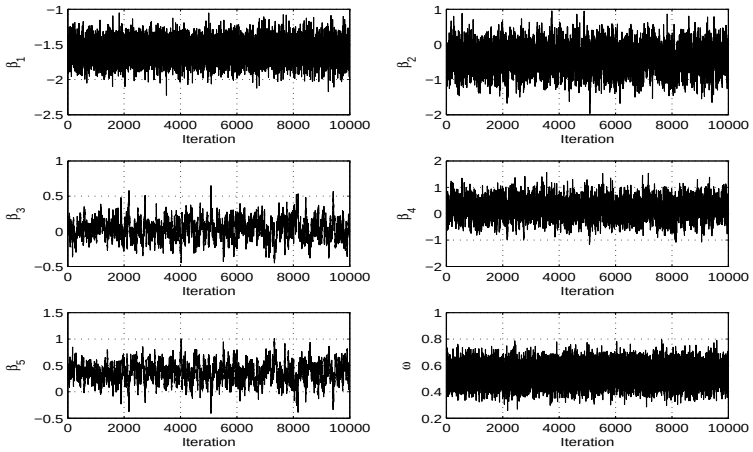

그림 **5.1.** ZIP 회귀모형에서의 각각의 모수에 대한 추정값

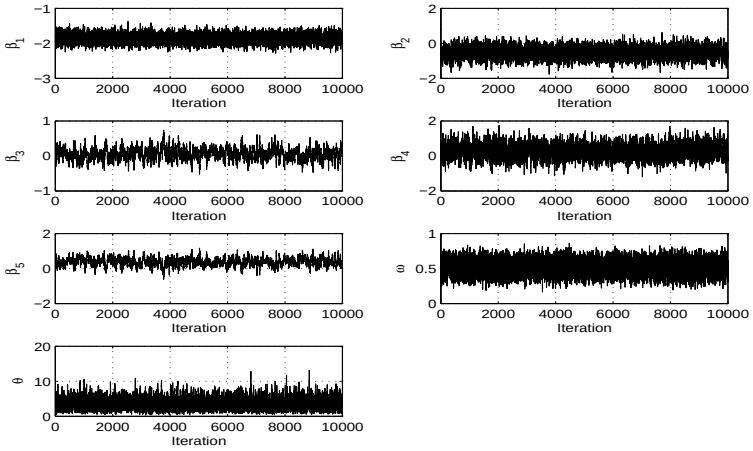

그림 **5.2.** ZINB 회귀모형에서의 각각의 모수에 대한 추정값

AAE는 다음과 같이 정의된다.

$$AAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|\,,$$

여기서 n은 실제 자료 중 종속변수가 0인 자료의 수이며, y_i와 $\hat{y}_i$는 종속변수의 실제값과 예측값을 나타낸다. AAE의 값이 더 작으면 예측력이 더 좋다는 것을 나타낸다. 각각의 사고예측모형에서의 0 자료에 대한 AAE는 표 5.3에 나와있다. 이 결과를 통하여 영과잉 모형이 일반적인 포아송회귀모형이나 음이항회귀모형보다 0에 대한 예측력이 좋다는 것을 알 수 있다.

영과잉 모형인 ZIP 회귀모형과 ZINB 회귀모형에 대하여 베이지안 모형선택을 해보자. ZINB 회귀모형을 M_1모형으로, ZIP 회귀모형을 M_2모형으로 정의하고 두 모형의 주변함수를 계산하면 ZIP 회귀모형의 경우 7.904×10^{-15}의 값이 나왔고, ZINB 회귀모형은 2.355×10^{-16}의 값이 나왔다. 베이즈 요인

표 **5.2.** 베이지안 추정법에 의한 각 모형에서의 모수추정 결과

	모수	평균	표준오차	95% HPD구간
ZIP 회귀모형	ω	0.5374	0.0772	(0.3862, 0.6841)
	β_1	−1.5877	0.1487	(−1.8807, −1.3058)
	β_2	−0.3901	0.3622	(−1.1043, 0.3217)
	β_3	0.0298	0.1432	(−0.2380, 0.3175)
	β_4	0.1873	0.3621	(−0.5213, 0.9023)
	β_5	0.3693	0.1850	(0.0018, 0.7151)
ZINB 회귀모형	ω	0.5239	0.1069	(0.3119, 0.7282)
	β_1	−1.8640	0.1218	(−2.1014, −1.6223)
	β_2	−0.5246	0.3117	(−1.1401, 0.0802)
	β_3	0.0559	0.1656	(−0.2748, 0.3802)
	β_4	0.2880	0.3919	(−0.4731, 1.0514)
	β_5	0.3820	0.2289	(−0.0699, 0.8341)

표 **5.3.** 사고예측모형에서의 0 자료에 대한 AAE

	AAE
포아송회귀모형	0.5591
음이항회귀모형	0.5530
ZIP 회귀모형	0.5077
ZINB 회귀모형	0.5072

B_{21}을 계산하면

$$B_{21} = \frac{\hat{m}_2(\mathbf{y})}{\hat{m}_1(\mathbf{y})} = \frac{7.904 \times 10^{-15}}{2.355 \times 10^{-16}} = 33.56$$

의 값이 나온다. 이 값은 표 3.1에 따르면 위의 교통자료가 M_1모형보다는 M_2모형을 선호하는 증거가 강력 하다는 것이다. 따라서 ZIP 회귀모형이 ZINB 회귀모형보다 위의 교통자료를 더 잘 설명해 주고 있음을 알 수 있다.

6. 결론

셀 수 있는 이산형 자료에 대하여 포아송 분포와 음이항회귀분포를 적용시키지만 영이 과도하게 높은 비율을 차지하고 있는 영과잉 자료의 경우 정규화된 분포를 따르지 못한다. 이런 자료들을 적합시키기 위하여 ZIP 회귀모형과 ZINB 회귀모형이 제시되었고, 이러한 영과잉 모형을 베이지안 관점에서 분석하여 각각의 모수를 추정하였다. 또한, 베이즈 요인을 적용하여 계산한 결과에 따라 모형 선택을 하였다.

분석 결과, 영에 대한 점 확률을 가지는 분포와 포아송 분포를 합성하여 자료를 설명해 줌으로써 과도하게 차지하고 있는 영의 부분을 잘 보완해 주었음을 알 수 있다. 또한 이러한 영과잉 모형 중 ZIP 회귀모형이 ZINB 회귀모형보다 자료에 잘 부합하는 모형이라고 결론을 내릴 수 있었다.

참고문헌

임아경, 오만숙 (2006). 영과잉 포아송 회귀모형에 대한 베이지안 추론: 구강위생 자료에의 적용, <응용통계연구>, **19**, 505–519.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities, *Journal of the America Statistical Association*, **85**, 389–409.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

Jeffreys, H. (1961). *Theory of Probability*, (Third edition), Oxford University Press, Oxford.

Joshua, S. C. and Garber, N. J. (1990). Estimating truck accident rate and involvements using linear and poisson regression models, *Transportation Planning and Technology*, **15**, 41–58.

Jovanis, P. P. and Chang, H. L. (1986). Modelling the relationship of accidents to miles traveled, *Transportation Research Record*, **1068**, 42–51.

McCulloch, R. and Rossi, P. E. (1991). A bayesian approach to testing the arbitrage pricing theory, *Journal of Econometrics*, **49**, 141–168.

Miaou, S. P. and Lum, H. (1993). Modeling vehicle accidents and highway geometric design relationships, *Accident Analysis and Prevention*, **25**, 689–709.

Milton, J. C. and Mannering, F. L. (1998). The relationship among highway geometrics, traffic-related elements and motor-vehicle accident frequencies, *Transportation*, **25**, 395–413.

Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap, *Journal of the Royal Statistical Society, Series B*, **56**, 3–48.

Poch, M. and Mannering, F. (1996). Negative binomial analysis of intersection-accident frequencies, *Journal of Transportation Engineering*, **122**, 105–113.

Raftery, A. E. and Banfield, J. D. (1991). Stopping the Gibbs Sampler, the use of morphology and other issues in spatial statistics, *Annals of the Institute of Statistical Mathematics*, **43**, 32–43.

Shankar, V., Mannering, F. L. and Barfield, W. (1995). Effect of roadway geometrics and environmental factors on rural freeway accident frequencies. *Accident Analysis and Prevention*, **27**, 371–389.

Shankar, V., Milton, J. C. and Mannering, F. L. (1997). Modeling accident frequencies as zero-altered probability process: An empirical inquiry, *Accident Analysis and Prevention*, **29**, 829–837.

Szabo, R. M. and Khoshgoftaar, T. M. (2000). Exploring a poisson regression fault model: A comparative study, Technical Report TR-CSE-00-56, Florida Atlantic University.

Bayesian Analysis for the Zero-inflated Regression Models

Hakjin Jang[1] · Yunhee Kang[2] · S. Lee[3] · Seong W. Kim[4]

[1]Division of Applied Mathematics, Hanyang University;
[2]Division of Applied Mathematics, Hanyang University;
[3]Division of Transportation Engineering, The University of SEOUL;
[4]Division of Applied Mathematics, Hanyang University

(Received February 2008; accepted May 2008)

Abstract

We often encounter the situation that discrete count data have a large portion of zeros. In this case, it is not appropriate to analyze the data based on standard regression models such as the poisson or negative binomial regression models. In this article, we consider Bayesian analysis for two commonly used models. They are zero-inflated poisson and negative binomial regression models. We use the Bayes factor as a model selection tool and computation is proceeded via Markov chain Monte Carlo methods. Crash count data are analyzed to support theoretical results.

Keywords: Zero-inflated model, Bayesian model selection, Bayes factor, Markov chain Monte Carlo.

This research was supported by grant No. R01-2005-000-10141-0 from the Basic Research program of the Korea Science and Engineering Foundation.

[1]Division of Applied Mathematics, Hanyang University, 1271 Sa 3-Dong, Sangnok-Gu, Ansan 426-791, Korea. E-mail: hjjang@hanyang.ac.kr

[2]Division of Applied Mathematics, Hanyang University, 1271 Sa 3-Dong, Sangnok-Gu, Ansan 426-791, Korea. E-mail: skylark84@naver.com

[3]Professor, Division of Transportation Engineering, The University of SEOUL, 90 Jeonnong-Dong, Dongdaemun-Gu, Seoul 130-743, Korea. E-mail: mendota@uos.ac.kr

[4]Corresponding author: Professor, Division of Applied Mathematics, Hanyang University, 1271 Sa 3-Dong, Sangnok-Gu, Ansan 426-791, Korea. E-mail: seong@hanyang.ac.kr