한국통계학회논문집
2011, 18권, 5호, 571–579

DOI: http://dx.doi.org/10.5351/CKSS.2011.18.5.571

서로 다른 산포를 허용하는 이변량 영과잉 음이항 회귀모형

김동석[a], 정슬기[a], 이동희[1,b]

[a]경기대학교 수학과, [b]경기대학교 경영학과

요 약

본 연구에서는 두 반응변수에 서로 다른 산포를 허용하는 새로운 이변량 영과잉 음이항 회귀모형을 제안하고, Deb과 Trivedi (1997)에 나타난 헬스케어 자료를 이용하여 두 반응변수가 갖는 서로 다른 산포도를 무시한 Wang (2003)이 제안한 이변량 영과잉 음이항 회귀모형과의 효율성을 로그우도와 AIC의 관점에서 비교하였다. 모형적합결과, 본 연구에서 제안한 모형이 모형선택기준 관점에서 기존모형에 비하여 월등히 우수한 결과를 보여주었다.

주요용어: 과대산포, 서로 다른 산포, 영과잉, 이변량 음이항 모형.

1. 서론

Li 등 (1999)이 영과잉된 이변량 계수형 자료에 대한 최초의 연구를 실시한 이래로, Walhin (2001) 및 Wang 등 (2003)에 의해 영과잉 이변량 계수형 자료에 대한 여러 형태의 이변량 영과잉 포아송(bivariate zero-inflated Poisson; BZIP) 회귀모형이 제안되었다. 이들의 연구는 포아송 모형을 기본 모형으로 고려하였기 때문에 반응변수에서 발생할 수 있는 과대산포(overdispersion)의 존재를 고려하지 않은 연구라는 한계를 가지고 있는데, 실제로 관찰되는 이변량 영과잉 계수형 자료는 두 반응변수에서 평균에 비하여 분산이 커지는 과대산포의 문제가 흔히 발생한다. 이러한 과대산포를 반영하기 위하여 최근 Wang (2003)은 Marshal과 Olkin (1990)의 이변량 음이항(Bivariate negative binomial; BNB)분포를 이용한 이변량 영과잉 음이항(bivariate zero-inflated negative binomial based on Marshall and Olkin's; BZINBMO) 회귀모형을 제안하였다. BZINBMO모형은 함께 관찰된 이변량 계수형 반응변수에 대해 하나의 산포모수를 이용하기 때문에 이들 반응변수가 서로 다르게 가질 수 있는 산포(heterogeneous dispersion)를 허용하지 못한다는 단점을 가지고 있다. 그러나 실제로 이변량 계수형 자료에서는 서로 다른 산포와 영과잉을 동시에 갖고 있는 자료가 흔히 관찰된다. 예를 들어, Deb과 Trivedi (1997)가 연구한 미국의 1987–1988 National Medical Expenditure Survey(NMES) 자료를 살펴보자. 이 자료는 미국에서 4406명의 노인들을 대상으로 헬스케어 사용에 대하여 6가지 이용형태를 측정한 자료이다. 이러한 6개의 반응변수 중에서 외과의사 외래방문횟수(OPP, Y_1)와 병원에 입원한 일수(HOSP, Y_2)를 반응변수로 고려해보자. 이 경우 두 반응변수에서 모두 (0, 0)이 나타난 사람, 즉 두 서비스 모두 이용하지 않은 사람의 비율은 64.7%로 나타나 영과잉이 발생함을 알 수 있다. 이와 함께 Y_1의 평균과 분산은 각각 0.751과 13.343, Y_2의 평균과 분산은 0.296과 0.557로 나타났다. 이를 통해 살펴보면 Y_1의 평균대비 분산비가 Y_2의 평균대비 분산비에 비하여 훨씬 더 크게 나타나, Y_1의 과대산포의 정도가 Y_2에 비하여 매우 큰 것을 알 수 있다. 그러므로 두 반응변수 OPP와 HOSP는 이러한 기술통계량으로부터 영과잉은 물론 서로 다른 산포도를 갖고 있음을 예상할 수 있다.

[1] 교신저자: (443-760) 경기도 수원시 영통구 이의동 산 94-6, 경기대학교 경영학과, 조교수. E-mail: dhl@kgu.ac.kr

영과잉 이변량 계수형 자료에서 두 반응변수 각각에 존재하는 서로 다른 산포가 회귀계수의 추정에 어떻게 영향을 미치는지에 대해서는 아직 정확하게 알려지지 않고 있다. 다만 이동희와 정병철 (2010)의 연구가 보여주었듯이 이변량 계수형 자료에서 서로 다른 산포를 반영할 경우에는 그렇지 않은 경우보다 모형의 적합도가 우수할 것으로 예상된다.

본 연구에서는 함께 관찰된 이변량 계수형 반응변수에 대해 이질적 산포를 허용하는 새로운 형태의 이변량 영과잉 음이항(bivariate zero-inflated negative binomial; BZINB) 회귀모형을 제안하고자 한다. 이 모형은 (0, 0)셀과 So 등 (2011)이 제안한 BNB모형의 혼합에 의하여 쉽게 얻어지는 모형이다. 2장에서는 본 연구에서 제안하는 모형과 추정방법을 설명하고, 3장에서는 Deb과 Trivedi (1997)에서 사용된 헬스케어 자료에 적합하고, 기존 모형과 결과를 비교하고자 한다. 4장에서는 이 논문의 결론을 하였다.

2. 모형과 추정

먼저 (Y_1, Y_2)를 계수값을 갖는 서로 상관된 이변량 확률변수라 하고, (y_{1i}, y_{2i}), $i = 1, \ldots, n$은 (Y_1, Y_2)의 관측치라 하자. 이때 두 확률변수 (Y_{1i}, Y_{2i})에 대하여 다음과 같은 결합확률분포를 고려해 보자.

$$\begin{aligned} P(Y_{1i} = 0, Y_{2i} = 0) &= \phi_i + (1 - \phi_i) f(0, 0) \\ P(Y_{1i} = y_{1i}, Y_{2i} = y_{2i}) &= (1 - \phi_i) f(y_{1i}, y_{2i}), \quad \text{if } (Y_{1i}, Y_{2i}) \neq (0, 0), \end{aligned} \tag{2.1}$$

이때 ϕ_i는 i번째 관측치의 (0, 0)셀에서의 팽창확률을 나타내고, $f(y_{1i}, y_{2i})$는 세 개의 음이항 확률변수들의 "삼각소거법"에 의하여 만들어지는 BNB 확률분포를 나타낸다. So 등 (2011)에 의하면 $f(y_{1i}, y_{2i})$는 다음과 같이 얻어진다.

$$f(y_{1i}, y_{2i}) = \prod_{j=1}^{3} (1 + \tau_j \mu_{ji})^{-\tau_j^{-1}} \sum_{k=0}^{\min(y_{1i}, y_{2i})} Q(i, k), \tag{2.2}$$

여기서

$$Q(i,k) = \prod_{l_1=1}^{y_{1i}-k} \frac{1 + \tau_1(y_{1i} - k + l_1)}{1 + \tau_1 \mu_{1i}} \prod_{l_2=1}^{y_{2i}-k} \frac{1 + \tau_2(y_{2i} - k + l_2)}{1 + \tau_2 \mu_{2i}} \prod_{l_3=1}^{k} \frac{1 + \tau_3(k - l_3)}{1 + \tau_3 \mu_3} \times \frac{\mu_{1i}^{y_{1i}-k} \mu_{2i}^{y_{2i}-k} \mu_3^k}{(y_{1i} - k)!(y_{2i} - k)! k!}$$

이며, τ_1, τ_2 및 τ_3는 산포모수를 나타내고, μ_3는 두 반응변수의 상관모수를 나타낸다.

식 (2.1)에 나타난 BZINB모형은 두 반응변수에 서로 다른 산포모수 τ_1과 τ_2를 사용하여 모형화하므로 이질적 산포를 허용한다는 점에서 Wang (2003)의 모형과 차별된다. 식 (2.1)의 모형을 Wang (2003)이 제안한 BZINBMO모형과 구별하기 위하여 BZINBDD모형이라 지칭하도록 하자.

이제 식 (2.1)의 분포를 갖는 BZINBDD 회귀모형을 고려해보자. 먼저 영과잉이 존재하지 않는 경우 두 반응변수 Y_{1i}과 Y_{2i}의 주변부 평균이 각각 $\mu_{1i} + \mu_3 = \mu_{1i}^*$와 $\mu_{2i} + \mu_3 = \mu_{2i}^*$로 얻어진다. 그러므로 이들에 대하여 로그 연결함수를 가정하고, 영과잉확률 ϕ_i는 항상 0에서 1사이의 값을 가지므로 ϕ_i에 대해서는 로짓 연결함수를 가정하면, 다음과 같다.

$$\mu_{1i}^* = \exp\left(\mathbf{x}_{1i}'\vec{\beta_1}\right), \quad \mu_{2i}^* = \exp\left(\mathbf{x}_{2i}'\vec{\beta_1}\right), \quad \text{and} \quad \log\left(\frac{\phi_i}{1 - \phi_i}\right) = \mathbf{z}_i'\vec{\gamma},$$

여기서 $\mathbf{x}_{1i}$, $\mathbf{x}_{2i}$와 $\mathbf{z}_i$는 각각 크기가 $k_1 \times 1$, $k_2 \times 1$와 $k_3 \times 1$인 설명변수 벡터를 나타내고, $\vec{\beta_1}$, $\vec{\beta_2}$ 및 $\vec{\gamma}$는 각각 $k_1 \times 1$, $k_2 \times 1$ 및 $k_3 \times 1$인 모수벡터를 나타낸다.

만일 $\phi_i = 0$라면, 식 (2.1)에 나타난 BZINBDD모형은 So 등 (2011)에서 나타난 BNB모형으로 축소될 것이고, 만일 $\tau_1 \to 0, \tau_2 \to 0$ 및 $\tau_3 \to 0$이라면, 이 모형은 Wang 등 (2003)이 제안한 BZIP모형이 될 것이다. 식 (2.1)의 BZINBDD모형을 따르는 두 반응변수 Y_{1i}과 Y_{2i}의 상관계수는 다음과 같이 구해진다.

$$\mathrm{Corr}(Y_{1i}, Y_{2i}) = \frac{\mu_3(1+\tau_3\mu_3) + \phi_i\mu_{1i}^*\mu_{2i}^*}{\sqrt{\left[\mu_{1i}^*\left(1+\phi_i\mu_{1i}^*\right) + \tau_1\mu_{1i}^2 + \tau_3\mu_3^2\right]\left[\mu_{2i}^*\left(1+\phi_i\mu_{2i}^*\right) + \tau_2\mu_{2i}^2 + \tau_3\mu_3^2\right]}}. \tag{2.3}$$

이와 같은 식 (2.3)을 보면 본 연구에서 제안한 BZINBDD모형은 항상 양의 상관만을 허용하고 있다는 사실을 알 수 있다. 그러나 대부분의 이변량 계수형 자료는 음의 상관이 발생하는 경우는 극히 드물게 나타난다는 점에서 본 연구에서 제시한 모형의 유용성은 충분히 존재한다는 판단이다 (Marshall과 Olkin, 1990; Wang, 2003).

각 모수에 대한 최대우도(Maximum Likelihood; ML) 추정량을 얻기 위하여 n개의 독립적인 표본을 이용하여 얻은 로그우도함수는 다음과 같다.

$$\begin{aligned}\log L = &\sum_{i=1}^{n} I_{(y_{1i},y_{2i})=(0,0)} \log\left(\exp\left(\mathbf{z}_i'\vec{\gamma}\right) + (1+\tau_1\mu_{1i})^{-\tau_1^{-1}}(1+\tau_2\mu_{2i})^{-\tau_2^{-1}}(1+\tau_3\mu_3)^{-\tau_3^{-1}}\right)\\ &+ \sum_{i=1}^{n} I_{(y_{1i},y_{2i})\neq(0,0)} \log\left(\sum_{k=0}^{\min(y_{1i},y_{2i})} Q(i,k)\right) - \sum_{i=1}^{n}\log\left(1+\exp\left(\mathbf{z}_i'\vec{\gamma}\right)\right),\end{aligned} \tag{2.4}$$

여기서 $I_{(\cdot)}$은 조건이 사실일 경우 1의 값을 갖고, 그렇지 않은 경우 0의 값을 갖는 지시함수이다. 식 (2.4)에 나타난 로그우도함수를 이용한 최대우도추정량은 일반적인 뉴튼-랩슨(Newton-Rapshon)방법과 같은 수치적인 방법을 통하여 구할 수 있으며, EM 알고리즘 (Dempster 등, 1977)과 같은 다른 추정방법도 적용가능하다.

본 연구에서 제안한 BZINBDD모형과 동일한 산포모수를 사용하는 BZINBMO모형의 적합결과를 비교해보는 것은 제안하는 모형의 효율성 측면에서 매우 유용할 것이라 생각한다. 사실 Wang (2003)은 Marshall과 Olkin (1990)의 BNB 분포에 근거한 이변량 영과잉 음이항(BZINBMO) 회귀모형을 제안하였는데, 이때 사용된 Marshall과 Olkin (1990)의 BNB 분포는 두 독립적인 포아송 분포를 갖는 확률변수의 감마혼합에 의하여 쉽게 얻어지는 분포로서 통계학 및 계량경제학에서 제안하는 모형에 대한 벤치마크로서 널리 사용되는 확률모형이다 (Munkin과 Trivedi, 1999; Gurmu와 Elder, 2000; Cameron 등, 2004).

이와 같은 Wang (2003)이 제안한 BZINBMO모형의 확률분포는 다음과 같다.

$$\begin{aligned}P(Y_{1i}=0, Y_{2i}=0) &= \phi_i + (1-\phi_i)g(0,0)\\ P(Y_{1i}=y_{1i}, Y_{2i}=y_{2i}) &= (1-\phi_i)g(y_{1i},y_{2i}), \quad \text{if } (Y_{1i},Y_{2i}) \neq (0,0),\end{aligned} \tag{2.5}$$

여기서

$$g(y_{1i},y_{2i}) = \frac{\Gamma(\tau^{-1}+y_{1i}+y_{2i})}{\Gamma(\tau^{-1})\Gamma(y_{1i}+1)\Gamma(y_{2i}+1)}\mu_{1i}^{y_{1i}}\mu_{2i}^{y_{2i}}\tau^{-\tau^{-1}}(\tau^{-1}+\mu_{1i}+\mu_{2i})^{-(\tau^{-1}+y_{1i}+y_{2i})}, \tag{2.6}$$

이며, 이때 $\tau(\geq 0)$는 두 반응변수에 공통으로 사용되는 산포모수이다. 식 (2.6)의 확률분포를 따르는 두 확률변수 $E(Y_{1i})$과 $E(Y_{2i})$의 주변부 평균은 각각 μ_{1i}와 μ_{2i}로 얻어지므로, 이들에 대하여 다음과 같은 로그 연결함수를 가정하고, 영과잉확률 ϕ_i에 대해서는 로짓 연결함수를 사용한다.

$$\mu_{1i} = \exp\left(\mathbf{x}_{1i}'\vec{\beta_1}\right), \quad \mu_{2i} = \exp\left(\mathbf{x}_{2i}'\vec{\beta_1}\right), \quad \text{and} \quad \log\left(\frac{\phi_i}{1-\phi_i}\right) = \mathbf{z}_i'\vec{\gamma}.$$

이때 식 (2.6)에 나타난 BZINBMO모형을 따르는 두 확률변수 Y_{1i}와 Y_{2i}의 상관계수는 다음과 같다.

$$\mathrm{Corr}(Y_{1i}, Y_{2i}) = \frac{\mu_{1i}\mu_{2i}(\phi_i + \tau)}{\sqrt{\mu_{1i}\mu_{2i}\left[1 + \mu_{1i}(\tau + \phi_i)\right]\left[1 + \mu_{2i}(\tau + \phi_i)\right]}}. \tag{2.7}$$

식 (2.7)을 통해, BZINBMO모형의 상관계수 역시 본 연구에서 제안한 BZINBDD모형과 마찬가지로 항상 양의 상관만을 허용하는 모형이라는 사실을 알 수 있다.

3. 실제자료분석

이제 앞에서 제안한 BZINBDD모형을 Deb과 Trivedi (1997)에서 살펴본 미국의 1987–1988 National Medical Expenditure Survey(NMES) 자료에 적합시키고 그 결과를 살펴보고자 한다. 이 데이터는 Journal of Applied Econometrics 1997 Data Archive에서 얻은 자료로서 미국에서 4406명의 노인들에 대한 헬스케어 사용에 대하여 6가지 측정도구를 이용하여 관찰한 자료이다. Deb과 Trivedi (1997)는 이 데이터에 대하여 여러가지 일변량 계수자료에 대한 모형을 이용하여 헬스케어 사용에 대한 모형화를 처음으로 실시하였고, Munkin과 Trivedi (1999)는 이변량 포아송-로그정규혼합 모형에 대한 결합모형을 고려한 후, 이때 회귀계수에 대한 모의최대우도(simulated maximum likelihood; SML) 추정량을 사용하여 적합한 바 있다.

본 연구에서는 병원에 입원한 일 수(HOSP, Y_2)와 외과의사 외래방문 횟수(OPP, Y_1)를 반응변수로 고려하고자 한다. 앞에서 살펴봤듯이 NMES자료에서 두 서비스 모두 사용하지 않는 사람의 비율은 64.7 %로 나타나 두 반응변수에서 (0, 0)셀이 비교적 높게 나타나는 영과잉이 존재함을 알 수 있다. 또한 Y_1의 평균과 분산은 각각 0.751과 13.343으로 나타나고, Y_2의 평균과 분산은 0.296과 0.557로 나타났다. 이를 통해 살펴보면 Y_1의 평균과 분산비가 Y_2의 평균과 분산비에 비하여 훨씬 더 크게 나타나, Y_1의 과대산포의 정도가 Y_2에 비하여 매우 큰 것을 알 수 있다. 그러므로 두 반응변수 OPP와 HOSP는 이러한 기술통계량으로부터 영과잉응 물론 서로 다른 산포를 갖고 있는 자료로 생각할 수 있다. 아울러 두 반응변수의 표본상관계수는 0.111로 나타나 두 반응변수 사이에 약한 양의 상관이 존재함을 알 수 있다.

사회경제학적 변수, 보험 및 건강상태를 포함하는 16개의 변수를 Y_1과 Y_2의 주변부 평균 및 (0, 0)셀 팽창확률에 대한 독립변수들로 고려하였다. 다음의 표 1은 이들 16개 독립변수들에 대한 설명을 나타낸 것이다. 이 외에 반응변수 및 독립변수에 대한 보다 자세한 설명 및 기초통계량에 대해서는 Deb과 Trivedi (1997)를 참고하기 바란다.

이 자료에 대하여 BZINBDD모형과 BZINBMO모형을 적합하기 이전에 (0, 0)셀에서 영과잉이 존재하는 지를 먼저 살펴보아야 한다. 이를 위하여 식 (2.1)과 (2.5)에 나타난 모형을 다음과 같이 수정해 보자.

$$\begin{aligned} P(Y_{1i} = 0, Y_{2i} = 0) &= \phi + (1 - \phi) f^*(0, 0) \\ P(Y_{1i} = y_{1i}, Y_{2i} = y_{2i}) &= (1 - \phi) f^*(y_{1i}, y_{2i}), \quad \text{if } (Y_{1i}, Y_{2i}) \neq (0, 0), \end{aligned} \tag{3.1}$$

여기서 ϕ는 (0, 0)셀의 과잉확률을 나타내는 모수로 제반의 독립변수에 의존하지 않는다고 가정하자. 만일 식(3.1)에서 $f^*(y_{1i}, y_{2i})$가 식 (2.1)에 나타난 $f(y_{1i}, y_{2i})$와 같다면 식 (3.1)의 모형은 영과잉확률에 독립변수가 존재하지 않는 BZINBDD모형이 될 것이고, $f^*(y_{1i}, y_{2i})$가 식 (2.6)에 나타난 $g(y_{1i}, y_{2i})$이라면 식 (3.1)의 모형은 영과잉확률에 독립변수가 존재하지 않는 BZINBMO모형이 될 것이다.

이와 같은 정의 하에서 Van den Broek (1995) 및 Lee 등 (2007)의 제안에 따라 ϕ를 $\phi = \psi/(1 - \psi)$로 재모수화(reparameterization) 해보자. 그렇다면 영과잉확률의 존재유무에 대한 가설검정은 다음과 같

표 1: 분석에 사용된 독립변수에 대한 설명

변수명	설명
EXCLHLTH	자기진단 결과 건강상태가 좋은 것으로 판단된 경우 1의 값
POORHLTH	자기진단 결과 건강상태가 나쁜 것으로 판단된 경우 1의 값
NUMCHRON	만성질환의 수
ADLDIFF	일상생활에 문제를 가지고 있는 경우 1의 값
NOREAST	미국 북동부에 거주하는 경우 1의 값
MIDWEST	미국 중서부에 거주하는 경우 1의 값
WEST	미국 서부에 거주하는 경우 1의 값
AGE	연령을 10으로 나눈 값
BLACK	흑인인 경우 1의 값
MALE	남성인 경우 1의 값
MARRIED	기혼인 경우 1의 값
SCHOOL	교육년수
FAMINC	가족의 연수입(단위: $10,000)
EMPLOYED	현재 취업중인 경우 1의 값
PRIVINS	개인 의료보험에 가입된 경우 1의 값
MEDICAID	미국의 고령자 및 장애자 의료보험에 적용되는 경우 1의 값

은 가설로 나타낼 수 있다.

$$H_0 : \psi = 0 \quad vs. \quad H_1 : \psi \neq 0. \tag{3.2}$$

식 (3.2)에 대한 검정방법으로는 우도비 검정(likelihood ratio test)과 스코어 검정(score test)이 널리 사용된다. 이 가운데 스코어 검정은 귀무가설하에서 최대우도추정량만을 필요로 하기 때문에 우도비 검정에 비하여 계산이 쉽다는 장점이 있다는 점에서 영과잉에 대한 가설검정방법으로 널리 사용된다. Van den Broek (1995)과 Gupta 등 (2004)는 여러 가지 일변량 영과잉 계수자료에 대한 회귀모형에서 스코어 검정에 대하여 연구하였고, Lee 등 (2007)은 그들의 연구를 BZIP모형으로 확장한 바 있는데, 본 연구에서는 이들의 연구 결과에 따라 가설 (3.2)에 대한 스코어 검정통계량을 유도하고 예제자료에 적용하였다. NMES자료에 적용결과 스코어 검정통계량 각각 BZINBDD모형에서 3.89가 얻어지고 BZINBMO모형에서는 117.40의 값이 각각 얻어졌다. 이 결과는 NMES자료에 $(0, 0)$셀 과잉확률이 존재함을 의미하므로, 본 연구에서 고려하는 BZINBDD모형과 BZINBMO모형 설정이 타당하다는 것을 보여준다.

표 2는 본 논문에서 제안한 BZINBDD모형의 회귀계수 추정치, 최대로그우도값 및 AIC값을 나타낸다. 더불어 BZINBDD모형과 비교의 목적으로 Wang (2003)이 제안한 BZINBMO모형의 결과도 표 2에 함께 제시하였다.

표 2의 결과 중에서 서로 다른 산포를 가정한 모형의 타당성을 살펴보자. BZINBDD모형에서 산포의 동일성에 대한 가설은 $H_0 : \tau_1 = \tau_2$가 될 것이며, 이 가설에 대한 검정은 우도비 검정을 통하여 수행할 수 있다. NMES 자료의 이 가설에 대한 우도비 검정통계량값은 $151.0 (p \leq 0.0001)$으로 얻어져 두 산포모수가 같지 않고 서로 다른 값을 보임을 알 수 있다. 사실 표 2의 결과를 살펴보면 τ_1과 τ_2는 각각 1.616과 5.646으로 추정되어 매우 큰 차이가 있음을 알 수 있다. 이제 BZINBDD모형과 BZINBMO모형의 적합결과를 비교해보자. 사실 두 모형은 서로 지분되어(nested) 있지 않으므로 직접적인 검정을 통하여 모형을 비교할 수는 없다. 하지만 로그우도값과 AIC값을 이용하여 간접적으로 비교해보면 BZINBDD모형의 적합이 BZINBMO모형의 적합결과에 비하여 월등히 우수하다는 사실을 알 수 있다.

BZINBDD모형과 BZINBMO모형의 추정결과를 살펴보면, 두 모형의 회귀계수 추정치는 서로 비

표 2: BZINBDD모형과 BZINBMO모형의 추정결과

변수명	BZINBDD			BZINBMO		
	HOSP	OPP	ϕ_i	HOSP	OPP	ϕ_i
Int.	−2.165(0.532)[a]	2.124(0.609)[a]	8.287(2.270)[a]	−2.707(0.680)[a]	2.198(0.473)[a]	9.429(2.541)[a]
EXCLHLTH	−0.455(0.212)[a]	−0.342(0.230)	0.174(0.527)	−0.602(0.283)[a]	−0.378(0.171)[a]	0.184(0.538)
POORHLTH	0.508(0.110)[a]	0.490(0.120)[a]	−0.308(0.605)	0.537(0.140)[a]	0.396(0.096)[a]	−0.732(0.845)
NUMCHRON	0.185(0.031)[a]	0.189(0.034)[a]	−1.117(0.217)[a]	0.218(0.041)[a]	0.197(0.029)[a]	−1.211(0.239)[a]
ADLDIFF	0.217(0.093)[a]	−0.173(0.103)	−1.675(0.725)[a]	0.182(0.118)	−0.119(0.081)	−2.461(1.361)
NOREAST	−0.015(0.113)	0.077(0.115)	−0.036(0.337)	−0.071(0.144)	0.165(0.094)	0.143(0.356)
MIDWEST	0.078(0.094)	0.040(0.100)	−0.383(0.327)	0.088(0.116)	0.117(0.075)	−0.314(0.347)
WEST	0.024(0.104)	0.106(0.130)	−0.989(0.461)[a]	0.017(0.136)	0.157(0.098)	−0.914(0.480)
AGE	0.070(0.064)	−0.400(0.074)[a]	−0.918(0.296)[a]	0.115(0.083)	−0.411(0.058)[a]	−1.070(0.337)[a]
BLACK	0.065(0.134)	0.880(0.124)[a]	−0.279(0.382)	0.157(0.167)	0.860(0.104)[a]	−0.168(0.396)
MALE	0.105(0.088)	0.048(0.098)	−0.887(0.314)[a]	0.171(0.116)	0.038(0.078)	−0.985(0.334)[a]
MARRIED	−0.099(0.089)	0.078(0.119)	−0.455(0.294)	−0.068(0.121)	0.161(0.087)	−0.276(0.307)
SCHOOL	−0.007(0.011)	0.034(0.012)[a]	−0.093(0.040)[a]	−0.003(0.014)	0.027(0.009)[a]	−0.099(0.043)[a]
FAMINC	0.020(0.018)	0.026(0.022)	0.117(0.035)[a]	0.030(0.023)	0.029(0.016)	0.133(0.038)[a]
EMPLOYED	−0.041(0.133)	−0.604(0.197)[a]	−0.464(0.495)	−0.018(0.177)	−0.645(0.139)[a]	−0.823(0.586)
PRIVINS	0.101(0.109)	−0.378(0.128)[a]	−0.345(0.343)	0.119(0.138)	−0.428(0.092)[a]	−0.568(0.357)
MEDICAID	0.209(0.157)	0.012(0.170)	0.764(0.445)	0.250(0.193)	−0.093(0.131)	0.586(0.479)
τ_1		1.616(0.179)[a]			-	
τ_2		5.646(0.374)[a]			-	
τ					2.576(0.153)[a]	
μ_3		0.031(0.006)[a]			-	
τ_3		4.166(4.062)			-	
로그우도		−6864.4			−7442.6	
AIC		13838.8			14989.2	

[a] 5 % 수준에서 유의한 변수. 괄호안은 추정된 표준오차를 나타냄.

슷한 결과를 보인다. 반면 회귀계수 추정치의 표준오차는 반응변수에 따라 서로 다르게 나타남을 알 수 있다. 좀 더 자세히 살펴보면, BZINBMO모형에서 HOSP의 평균에 대한 회귀계수의 표준오차는 모든 설명변수에서 BZINBDD모형의 그것에 비하여 약간 크게 나타난 반면 OPP의 평균에 대한 회귀계수의 표준오차는 BZINBDD모형의 표준오차에 비하여 약간 작게 나타나고 있다. 이러한 표준오차의 차이는 BZINBMO모형이 서로 다른 산포를 허용하지 않기 때문에 발생한 것으로 판단된다. 그러므로 서로 다른 산포가 존재하는 NMES 자료에 대한 BZINBMO모형에서 회귀계수에 대한 추론은 표준오차의 편이(bias) 때문에 부정확할 가능성이 있다. 예를 들어 BZINBMO모형에서 OPP에 대한 EXCLHLTH변수는 5% 수준에서 유의했지만, BZINBDD모형에서는 유의하지 않게 나타나고 있다.

이제 BZINBDD모형에서 각 설명변수의 효과를 살펴보자. 회귀계수 추정치의 부호를 통하여 EXCLHLTH는 HOST의 평균을 감소시키는 반면, POORHLTH, NUNCHRON와 ADLDIFF가 커질수록 HOST의 평균이 커지는 것으로 나타났다. 또한 독립변수 POORHLTH, NUMCHRON, BLACK 및 SCHOOL이 커질수록 OPP의 평균이 커지는 반면, AGE, EMPLOYED 및 PRIVINS가 커질수록 이 값을 감소시키는 것으로 나타났다. ϕ_i의 경우, 독립변수 NUMCHRON, ADLDIFF, WEST, AGE, MALE 및 SCHOOL이 커질수록 두 서비스(OPP, HOSP) 모두 이용하지 않을 확률은 감소하는 것으로 나타났으며, 독립변수 FAMINC이 커질수록 이 확률은 증가하는 것으로 나타났다.

표 2에 나타난 추정결과를 이용하여 얻은 BZINBDD모형과 BZINBMO모형의 적합도수(fitted frequency)와 실제 얻어진 관측도수(observed frequency)는 표 3에 정리되어 있다. 이때 적합도수는 다음과 같은 과정에 의하여 계산하였다. 먼저 $\hat{p}(c_{1i}, c_{2i})$, $i = 1, \ldots, n$; $c_{1i}, c_{2i} = 0, 1, \ldots$를 (Y_{1i}, Y_{2i})가 (c_1, c_2)셀을 갖을 적합확률이라 정의하자. 그렇다면 (c_1, c_2) 셀의 적합도수는 $\sum_{i=1}^{n} \hat{p}(c_{1i}, c_{2i})$, $c_{1i}, c_{2i} = 0, 1, \ldots$ 식에 의하여 얻을 수 있다.

표 3에서 볼 수 있듯이, BZINBMO모형은 $Y_1 \leq 1$이거나 $Y_2 \leq 1$인 경우 관측도수를 과대예측 또는 과소예측하는 경향을 보여주고 있다. 반면 BZINBDD모형은 관측도수에 대한 적절한 적합결과를 제

표 3: 관측도수 및 BZINBDD모형과 BZINBMO모형의 적합도수

HOSP (Y_1)	모형	OPP (Y_2)								
		0	1	2	3	4	5	6	7	8+
0	관측도수	2851	369	149	51	35	25	14	4	43
	BZINBDD	2854.8	282.1	131.1	76.0	48.8	33.3	23.7	17.3	68.3
	BZINBMO	2919.9	421.4	153.1	66.8	32.3	16.7	9.2	5.3	9.0
1	관측도수	394	105	39	16	6	11	8	1	19
	BZINBDD	394.2	110.7	33.7	18.9	12.0	8.2	5.9	4.3	18.0
	BZINBMO	184.9	120.8	72.0	42.6	25.5	15.6	9.7	6.2	13.2
2	관측도수	103	34	8	6	5	0	3	4	13
	BZINBDD	109.6	26.3	12.9	5.7	3.6	2.5	1.8	1.3	6.0
	BZINBMO	31.4	33.4	26.9	19.7	13.9	9.7	6.7	4.7	12.3
3	관측도수	33	6	4	1	1	0	0	0	3
	BZINBDD	36.3	8.2	3.7	2.2	1.3	0.9	0.6	0.5	2.3
	BZINBMO	6.9	9.8	9.7	8.3	6.7	5.2	3.9	2.9	9.5
4	관측도수	7	7	2	1	2	0	0	0	1
	BZINBDD	13.8	3.0	1.4	0.8	0.5	0.3	0.3	0.2	1.0
	BZINBMO	1.8	3.1	3.6	3.5	3.1	2.6	2.1	1.7	6.7
5+	관측도수	9	5	2	1	2	0	0	1	2
	BZINBDD	11.5	2.4	1.1	0.6	0.4	0.3	0.2	0.2	1.0
	BZINBMO	0.8	1.7	2.4	2.7	2.8	2.7	2.5	2.2	14.2

공해주고 있다. 전반적으로 BZINBDD모형이 BZINBMO모형에 비하여 좀 더 나은 적합결과를 보여주고 있음을 확인할 수 있다.

4. 결론

본 논문에서는 두 반응변수에 서로 다른 산포를 허용하는 새로운 형태의 이변량 영과잉 음이항(BZINBDD) 회귀모형을 제안하였다. 이 모형은 서로 독립적인 3개의 음이항 분포를 따르는 확률변수들의 "삼각소거법"에 의하여 만들어지는 이변량 음이항 확률분포와 (0, 0)셀에서의 확률과의 혼합에 의하여 쉽게 얻어지는 모형이다. 본 논문에서 제안한 BZINBDD모형을 Deb과 Trivedi (1997)에서 사용한 헬스케어 사용에 대한 실제자료에 적합한 결과, BZINBDD모형은 서로 다른 산포를 허용하지 않는 Wang (2003)이 제안한 BZINBMO모형의 결과에 비하여 최대로그우도값, AIC 및 적합빈도의 관점에서 월등히 나은 모형으로 나타났다. 본 연구에서 제안한 BZINBDD모형에 기반하여 영과잉확률에 대한 다양한 검정통계량의 효율성을 파악하고, 산포의 동일성에 대한 검정 및 두 반응변수의 독립성에 대한 검정 등은 추후 연구과제로 남겨둔다.

참고 문헌

이동희, 정병철 (2010). 코폴라를 활용한 이변량 제로팽창 일반화 포아송 회귀모형, *Journal of the Korean Data Analysis Society*, **12**, 1473–1484.

Cameron, A. C., Li, T., Trivedi, P. K. and Zimmer, D. M. (2004). Modelling the differences in counted outcomes using bivariate copula models with application to mismeasured counts, *Econometrics Journal*, **7**, 566–584.

Deb, P. and Trivedi, P. K. (1997). Demand for medical care by the elderly: A finite mixture approach, *Journal of Applied Econometrics*, **12**, 313–336.

Dempster, A. P., Laird, N. M., Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discusiion), *Journal of the Royal Statistical Society B*, **39**, 1–38.

Gupta, P. L., Gupta, R. C. and Tripathi, R. C. (2004). Score test for zero inflated generalized Poisson regression model, *Communications in Statistics - Theory and Methods*, **33**, 47–64.

Gurmu, S. and Elder, J. (2000). Generalized bivariate Count data regression models, *Economics Letters*, **68**, 31–36.

Lee, J., Jung, B. C. and Jin, S. H. (2007). Tests for zero inflation in a bivariate zero-inflated Poisson model, *Statistica Neerlandica*, **63**, 400–417.

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

Marshall, A. W. and Olkin, I. (1990). Multivariate distributions generated from mixtures of convolution and product families, In H.W. Block, A.R. Sampson and T.H. Savits(eds), *Topics in Statistical Dependence*, 372–393. IMS Lecture Notes - Monograph Series, **16**.

Munkin, M. K. and Trivedi, P. K. (1999). Simulated maximum likelihood estimation of multivariate mixed-Poisson regression models, with application, *Econometrics Journal*, **2**, 29–48.

So, S., Chun, H. and Jung, B. C. (2011). Bivariate negative binomial regression model with heterogeneous dispersions, *Communications in Statistics - Theory and Methods*, Submitted.

Van den Broek, J. (1995). A score test for zero inflation in a Poisson distribution, *Biometrics*, **51**, 738–743.

Walhin, J. F. (2001). Bivariate ZIP models, *Biometrical Journal*, **43**, 147–160.

Wang, K., Lee, A. H., Yau, K. K. W. and Carivick, P. J. W. (2003). A bivariate zero-inflated Poisson regression model to analyze occupational injuries, *Accident Analysis and Prevention*, **35**, 625–629.

Wang, P. (2003). A bivariate zero-inflated negative binomial regression model for count data with excess zeros, *Economics Letters*, **78**, 373–378.

2011년 7월 접수; 2011년 8월 채택

Bivariate Zero-Inflated Negative Binomial Regression Model with Heterogeneous Dispersions

Dongseok Kim[a], Seulgi Jeong[a], Dong-Hee Lee[1,b]

[a]Department of Mathematics, Kyonggi University
[b]Department of Business Administration, Kyonggi University

Abstract

We propose a new bivariate zero-inflated negative binomial regression model to allow heterogeneous dispersions. To show the performance of our proposed model, Health Care data in Deb and Trivedi (1997) are used to compare it with the other bivariate zero-inflated negative binomial model proposed by Wang (2003) that has a common dispersion between the two response variables. This empirical study shows better results from the views of log-likelihood and AIC.

Keywords: Bivariate negative binomial distribution, heterogeneous dispersion, overdispersion, zero-inflation.

[1] Corresponding author: Associate Professor, Department of Business Administration, Kyonggi University, San 94-6, Iui-Dong, Yeongtong-Gu, Suwon-Si, Kyonggi-Do 443-760, Korea. E-mail: dhl@kgu.ac.kr