응용통계연구 제 21권 2호, 2008년, pp. 341-353

이변량 음이항 모형에서 붓스트랩 방법을 이용한 과대산포에 대한 검정*

전명식[1)] 정병철[2)]

요 약

본 연구에서는 이변량 음이항 분포에서 과대산포와 "내재적 상"의 존재유무에 대한 가설검정 문제를 다루었다. 과대산포에 대한 스코어 검정의 표준정규분포 근사는 명목유의수준을 과소추정한 반면 "내재적 상"에 대한 스코어 검정은 명목유의수준을 과대추정하고 있음을 보였다. 본 연구에서는 이와 같은 스코어 검정의 표준정규분포 근사의 문제점을 해결하기 위하여 붓스트랩 방법을 제안하였다. 스코어 검정에 대한 붓스트랩 방법은 두 검정에서 명목유의수준을 제대로 유지하고 검정력도 높게 나타나 스코어 검정의 표준정규분포 근사에 존재하는 문제를 해결하는 효율적인 대안으로 판단된다.

Keywords: 이변량 포아송 분포, 이변량 음이항 분포, 과대산포, 붓스트랩.

1. 서론

최근 경제학, 의학 등 다양한 학문분야에서 이변량 계수데이터(Bivariate Count Data)에 대한 활용성 및 중요성이 매우 높아지면서, 이 자료의 분석에 관한 여러 연구가 시도되고 있다. 계수자료를 모형화 함에 있어 계수가 어떤 속성으로 측정되었는지에 따라 다양한 분석모형의 적용 및 개발이 이루어질 수 있다. 그 중 이변량 포아송(Bivariate Poisson: BP) 분포는 이변량 계수데이터의 분석에 있어서 가장 광범위하게 사용되는 확률모형이다 (Holgate, 1964; Kocherlakota와 Kocherlakota, 2001). BP 모형은 두 확률변수 사이의 상관관계를 고려할 수 있는 반면 모형의 성격상 주변 확률변수의 평균과 분산이 동일하다는 제약조건이 있다. 그러나 실제 얻어지는 이변량 계수 데이터에서는 평균과 함께 분산도 증가하는 과대산포(overdispersion)의 형태를 갖는 경우가 흔히 발생한다. 이와 같이 과대산포가 존재하는 이변량 계수자료의 경우 BP 분포를 사용한 모형적합은 효율성이 떨어지므로 과대산포를 조절할 수 있는 이변량 음이항 (Bivariate Negative Binomial: BNB) 분포를 사용해야 한다 (Bates와 Neyman, 1952; Marshall과 Olkin, 1990; Subrahmaniam, 1966; Subrahmaniam과 Subrahmaniam, 1973). BNB 분포는 크게 2가지 형태가 제안되고 있다. 첫 번째는 Bates와

* 이 논문은 2007년도 정부재원 (교육인적자원부 학술연구조성사업비)으로 한국학술진흥재단의 지원을 받아 연구되었음 (KRF-2007-314-C00039).

1) (136-701) 서울시 성북구 안암동, 고려대학교 통계학과, 교수.
E-mail: jhun@korea.ac.kr

2) (130-743) 교신저자. 서울시 동대문구 전농동, 서울시립대학교 통계학과, 조교수.
E-mail: bcjung@uos.ac.kr

Neyman (1952)에 의해 제안된 이변량 음이항 분포로서 두 독립적인 포아송 분포를 따르는 확률변수와 감마분포를 따르는 확률변수의 혼합에 의해 만들어지는 분포이며, 두 번째 음이항 분포는 Subrahmaniam (1966)에 의하여 제안된 분포로서 두 변수간 상관이 존재하는 BP 분포와 감마분포의 혼합(compounding)을 이용하여 만들어지는 분포이다. 그러므로 Subrahmaniam (1966)의 BNB 분포가 Bates와 Neyman (1952)의 BNB 분포에 비하여 보다 넓은 의미를 갖는 분포라 할 수 있다. BP 분포와 BNB 분포에 대한 보다 자세한 설명을 위해서는 Kocherlakota와 Kocherlakota (1992) 및 Johnson 등 (1997)을 참고할 수 있다.

본 연구에서는 Subrahmaniam (1966) 형태의 BNB 분포에서 과대산포 및 "내재적 상관 (intrinsic correlation)"의 존재유무에 대한 가설검정을 다루고자 한다. 이때 "내재적 상관"은 BNB 분포를 만드는 혼합 이전에 BP 분포에서 발생하는 상관을 의미한다. 본 연구모형에서 과대산포는 존재하지 않고 내재적 상관만 존재하는 경우라면 BP 분포가 적합하며, 내재적 상관이 존재하지 않고 과대산포만 존재하는 경우라면 Bates와 Neyman (1952)이 제안한 BNB 분포가 적절할 것이다. 반면 과대산포와 내재적 상관이 모두 존재하는 경우라면 본 연구에서 고려하는 BNB 분포가 적합할 것이다. 그러므로 서로 상관된 두 계수형 확률변수가 주어지는 경우 어떤 모형이 타당할지에 대한 검정은 모형적합 이전에 필수적으로 실시해야만 한다. 사실 Subrahmaniam (1966)은 BNB 분포에서 내재적 상관의 존재유무에 대한 스코어 검정통계량을 유도하였고 Jung 등 (2007)은 동일한 모형에서 과대산포를 검정하기 위한 스코어 검정과 LR 검정통계량을 유도하였다. Jung 등 (2007)은 과대산포에 대한 스코어 검정은 명목유의수준을 과소추정하고 "내재적 상"에 대한 스코어 검정은 명목유의수준을 과대추정하고 있음을 보여주었다. 이에 본 연구에서는 이들 스코어 검정의 과대 및 과소추정의 문제를 해결하기 위하여 붓스트랩의 방법을 제안하고자 한다. 사실 과대산포에 대한 스코어 검정에서 붓스트랩 방법의 적용은 일변량 포아송 회귀모형에서 주로 사용되었다. Jung 등 (2005, 2006)은 제로팽창 및 중도절단된 일변량 포아송 회귀모형에서 과대산포의 검정에 붓스트랩 방법의 사용을 제안하였다. 이들에 의하여 제안된 붓스트랩 방법은 명목유의수준을 정확히 유지하고 검정력도 높게 나타나 과대산포의 검정에 효율적인 방법으로 나타났다. 본 연구에서는 일변량에서 Jung 등 (2005, 2006)의 연구를 이변량 계수자료 모형으로 확장 적용하고자 한다.

2장에서는 고려하고자 하는 모형 및 스코어 검정통계량을 다루고, 3장에서는 붓스트랩 방법을 제안하고자 한다. 4장에서는 모의실험을 통하여 붓스트랩 방법의 효율성을 체크하고, 5장에서는 실제 자료분석에 제안된 방법을 적용하고자 한다. 마지막으로 6장에서는 결론을 내리고자 한다.

2. BNB 모형 및 스코어 검정통계량

먼저 (Y_1, Y_2)에 대하여 다음과 같은 결합확률분포를 고려해보자 (Jung 등, 2007).

$$f(y_1, y_2) = \frac{\tau^{-\tau^{-1}}}{(\mu_1 + \mu_2 + \mu_3 + \tau^{-1})^{\tau^{-1}+y_1+y_2}} \frac{\Gamma(\tau^{-1} + y_1 + y_2)}{\Gamma(\tau^{-1})} S(y_1, y_2), \tag{2.1}$$

여기서

$$S(y_1, y_2) = \sum_{k=0}^{\min(y_1,y_2)} P(y_1, y_2, k)Q(k), \tag{2.2}$$

$$P(y_1, y_2, k) = \frac{\mu_1^{y_1-k}\mu_2^{y_2-k}\mu_3^k}{(y_1-k)!(y_2-k)!k!}, \qquad Q(k) = \frac{(1+\tau\mu_1+\tau\mu_2+\tau\mu_3)^k}{\Pi_{l=1}^k(1+\tau y_1+\tau y_2-l\tau)},$$

이며 μ_1, μ_2와 μ_3는 평균을 그리고 τ는 산포를 나타내는 모수이다. 그러므로 모든 모수들은 항상 0보다 크거나 같은 값을 가져야만 한다. 식 (2.1)의 확률분포는 BP 분포와 감마분포와의 혼합에 의하여 얻어지는 분포로 $(Y_1, Y_2) \sim \text{BNB}(\mu_1, \mu_2, \mu_3, \tau)$라 표기하자. 사실 Subrahmaniam (1966)은 BNB 분포를 만들기위한 혼합에 감마분포(α, β)를 사용하였다. 하지만 혼합이전과 같은 평균을 가지며 혼합이후 더 큰 분산을 갖는 모형을 만들기위하여 본 연구에서는 모수가 하나인 감마분포(τ^{-1}, τ^{-1})를 혼합에 사용하였다. 이와 같은 혼합방법은 일변량 계수형 자료에서 아주 흔하게 사용되는 방법이다 (Dean과 Lawless, 1989; Gurmu, 1991). Y_1과 Y_2에 대한 1, 2차 적률은 Jung 등 (2007)의 결과를 이용하면 다음과 같이 계산된다.

$$\begin{aligned}
E(Y_1) &= \mu_{10} = \mu_1 + \mu_3, \\
E(Y_2) &= \mu_{01} = \mu_2 + \mu_3, \\
\text{Cov}(Y_1, Y_2) &= \mu_{11} = \mu_3 + \tau(\mu_1+\mu_3)(\mu_2+\mu_3), \\
\text{Var}(Y_1) &= \mu_{20} = (\mu_1+\mu_3)[1+\tau(\mu_1+\mu_3)], \\
\text{Var}(Y_2) &= \mu_{02} = (\mu_2+\mu_3)[1+\tau(\mu_2+\mu_3)].
\end{aligned} \tag{2.3}$$

만일 $\tau \to 0$이면 식 (2.1)에 나타난 BNB 분포는 Holgate (1964)의 BP 분포로 수렴하고 식 (2.1)에 나타난 적률도 BP 분포의 적률로 표현된다. 그러므로 식 BP 모형 대 BNB 모형에 대한 가설검정은 다음과 같이 표현된다.

$$H_0^a : \tau = 0 \quad vs. \quad H_1^a : \tau > 0. \tag{2.4}$$

사실 식 (2.4)에 나타난 가설은 달리 표현하면 BP 모형에 과대산포가 존재하는지를 검정하는 가설이다. 가설 (2.4)에 대한 스코어 검정을 유도하는데 필요한 로그우도함수는 다음과 같이 구해진다.

$$\begin{aligned}
\log L &= \sum_{i=1}^n \phi_i \\
&= \sum_{i=1}^n \left[\sum_{l=1}^{y_{1i}+y_{2i}} \log(1+\tau y_{1i}+\tau y_{2i}-l\tau) - (y_{1i}+y_{2i})\log(\mu_1\tau+\mu_2\tau+\mu_3\tau+1) \right. \\
&\quad \left. + \log S(y_{1i}, y_{2i}) \right] - n\frac{\log(\mu_1\tau+\mu_2\tau+\mu_3\tau+1)}{\tau},
\end{aligned} \tag{2.5}$$

여기서 ϕ_i는 i번째 관측치의 확률함수에 대한 로그값을 나타낸다.

식 (2.5)의 로그우도함수를 이용하여 계산한 가설 (2.4)에 대한 스코어 검정통계량은 다음과 같다.

$$T_a = \widehat{S}(\tau)/\sqrt{\widehat{V}(\tau)}, \tag{2.6}$$

여기서 $\widehat{S}(\tau)$은 귀무가설하에서 계산된 τ에 대한 추정된 스코어 함수값이고 $\widehat{V}(\tau)$는 귀무가설하에서 계산된 $\widehat{S}(\tau)$의 점근적 분산을 나타낸다. 이때 $\widehat{S}(\tau)$은 다음과 같이 계산되며 $\widehat{V}(\tau)$는 정보행렬의 역행렬을 이용하여 계산할 수 있다. 사실 Jung 등 (2007)은 그레디언트의 외적(Outer-Product of Gradient)의 방법을 이용하여 정보행렬의 추정치를 계산하였다.

$$\begin{aligned}
\frac{\partial \log L}{\partial \tau}\Big|_{H_0^a} &= \sum_{i=1}^{n} \frac{\partial \phi_i}{\partial \tau}\Big|_{H_0^a} = \widehat{S}(\tau) \\
&= \frac{1}{2}\sum_{i=1}^{n}\left[(y_{1i}+y_{2i}-\widehat{\mu}_1-\widehat{\mu}_2-\widehat{\mu}_3)^2-(y_{1i}+y_{2i})\right] \\
&\quad +\sum_{i=1}^{n}\left[\frac{\displaystyle\sum_{k=0}^{\min(y_{1i},y_{2i})} \widehat{P}(y_{1i},y_{2i},k)k\left((\widehat{\mu}_1+\widehat{\mu}_2+\widehat{\mu}_3)-(y_{1i}+y_{2i})+\frac{k+1}{2}\right)}{\displaystyle\sum_{k=0}^{\min(y_{1i},y_{2i})} \widehat{P}(y_{1i},y_{2i},k)}\right],
\end{aligned} \tag{2.7}$$

여기서

$$\widehat{P}(y_{1i},y_{2i},k) = \frac{\widehat{\mu}_1^{y_{1i}-k}\widehat{\mu}_2^{y_{2i}-k}\widehat{\mu}_3^{k}}{(y_{1i}-k)!(y_{2i}-k)!k!},$$

이며 $\widehat{\mu}_1$, $\widehat{\mu}_2$ 및 $\widehat{\mu}_3$는 귀무가설 (즉 BP 분포) 하에서 μ_1, μ_2 및 μ_3에 대한 ML 추정량을 각각 나타낸다. Holgate (1964)에 의하면 $\widehat{\mu}_1$과 $\widehat{\mu}_2$는 각각 $\bar{y}_{1.}-\widehat{\mu}_3$과 $\bar{y}_{2.}-\widehat{\mu}_3$로 얻어지며 $\widehat{\mu}_3$는 반복적인 방법으로 구할 수 있다. 여기서 $\bar{y}_{1.}=\sum_{i=1}^{n} y_{1i}/n$이며 $\bar{y}_{2.}=\sum_{i=1}^{n} y_{2i}/n$이다.

귀무가설 H_0^a가 맞다는 가정하에서 식 (2.6)에서 얻어진 스코어 검정통계량 T_a는 표본수가 커질때 점근적으로 표준정규분포를 따르게 된다. 그러므로 식 (2.6)에서 얻어진 T_a의 값이 표준정규분포의 $100(1-\alpha)$분위수의 값보다 크게되면 BP 분포를 기각하게 된다.

사실 양측(two-sided) 대립가설에 대한 전통적인 스코어 검정통계량은 $T_a^* = T_a^2$으로 얻어지며, 이의 점근적 분포는 $\chi^2(1)$분포를 따르게 된다. 하지만 본 연구에서 고려하는 바와 같이 대립가설의 형태가 단측인 경우 T_a^*를 이용한 추론은 효율성이 떨어지게 될 것이다. 그러므로 단측검정에서 스코어 검정통계량 T_a^*의 비효율성 문제를 해결하기 위하여 Verbeke와 Molenberghs (2003)는 일방향 스코어 검정통계량의 점근분포로 혼합 카이제곱 분포를 사용하는 검정방법을 제안하고 이에 대한 직관적인 해석방법을 제안하였다. 그들의 연구를 본 연구에 적용하면 본 연구에 대한 일방향 스코어 검정통계량은 다음과 같이 구해진다.

$$T_a^{**} = \begin{cases} T_a^2, & \text{if } \widehat{S}(\tau) \geq 0, \\ 0, & \text{if } \widehat{S}(\tau) < 0. \end{cases} \tag{2.8}$$

식 (2.8)에서 얻어진 T_a^{**}는 귀무가설이 맞다는 가정하에서 혼합 카이제곱 분포 $0.5\chi^2(0)+0.5\chi^2(1)$를 따르게 된다. 여기서 $\chi^2(0)$는 1의 확률로 0으로 수렴하는 값을 의미한다. 하지만

식 (2.8)에서 얻어진 T_a^{**}를 이용하는 것과 식 (2.6)에서 얻어진 T_a를 이용하는 것은 통계적으로 동일한 방법이라는 사실은 Verbeke와 Molenberghs (2003)를 참조하면 쉽게 알 수 있다. 즉, 이 분포의 각 유의수준하에서의 임계값은 표준정규분포에서 얻어진 임계값의 제곱으로 나타나므로 T_a 또는 T_a^{**}중 아무거나 사용하여도 동일한 검정결과를 얻을 수 있다.

본 연구모형에서 또 다른 검정은 공분산 모수인 μ_3의 존재유무이다. 사실 Subrahmaniam (1966)은 μ_3를 "내재적 상관(intrinsic correlation)"이라 칭하고 이의 존재유무에 대한 스코어 검정통계량을 유도하였다. 이 경우 다음과 같은 가설을 세울 수 있다.

$$H_0^b : \mu_3 = 0 \quad vs \quad H_1^b : \mu_3 > 0, \tag{2.9}$$

여기서 μ_3의 특성상 대립가설의 형태는 H_0^a와 마찬가지로 단측검정이 된다. 만일 귀무가설 H_0^b가 맞다면 (Y_{1i}, Y_{2i})의 결합확률분포는 다음과 같이 축소될 것이다.

$$f(y_{1i}, y_{2i}) = \frac{\Gamma(\tau^{-1} + y_{1i} + y_{2i})}{\Gamma(\tau^{-1}) y_{1i}! y_{2i}!} \mu_1^{y_{1i}} \mu_2^{y_{2i}} \tau^{-\tau^{-1}} (\tau^{-1} + \mu_1 + \mu_2)^{-(\tau^{-1} + y_{1i} + y_{2i})}. \tag{2.10}$$

식 (2.10)에 나타난 결합확률분포는 또다른 형태의 BNB 분포로서 두 개의 서로 독립인 포아송 분포를 따르는 확률변수와 감마분포의 혼합에 의하여 얻어지는 분포이다 (Bates와 Neyman, 1952; Marshall과 Olkin, 1990).

식 (2.5)의 로그우도함수를 이용하여 가설 (2.9)에 대하여 유도한 스코어 검정통계량은 다음과 같다.

$$T_b = \widetilde{S}(\mu_3) / \sqrt{\widetilde{V}(\mu_3)}, \tag{2.11}$$

여기서 $\widetilde{S}(\mu_3)$는 귀무가설 H_0^b (즉, 식 (2.10)에 나타난 BNB 분포) 하에서 얻어진 μ_3에 대한 추정된 스코어 함수로서 다음과 같이 구해진다.

$$\begin{aligned} \frac{\partial \log L}{\partial \mu_3}\Big|_{H_0^b} &= \sum_{i=1}^{n} \frac{\partial \phi_i}{\partial \mu_3}\Big|_{H_0^b} = \widetilde{S}(\mu_3) \\ &= -\frac{\sum_{i=1}^{n} (\widetilde{\tau} y_{1i} + \widetilde{\tau} y_{2i} + 1)}{1 + \widetilde{\tau}\widetilde{\mu}_1 + \widetilde{\tau}\widetilde{\mu}_2} + \frac{1 + \widetilde{\tau}\widetilde{\mu}_1 + \widetilde{\tau}\widetilde{\mu}_2}{\widetilde{\mu}_1 \widetilde{\mu}_2} \sum_{i=1}^{n} \frac{y_{1i} y_{2i}}{1 + \widetilde{\tau} y_{1i} + \widetilde{\tau} y_{2i} - \widetilde{\tau}}, \end{aligned} \tag{2.12}$$

여기서 $\widetilde{\mu}_1$, $\widetilde{\mu}_2$ 및 $\widetilde{\tau}$는 귀무가설 H_0^b하에서 얻어진 μ_1, μ_2 및 τ에 대한 ML 추정량을 각각 나타낸다. Jung 등 (2007)에 의하면 $\widetilde{\mu}_1 = \bar{y}_{1.}$이고 $\widetilde{\mu}_2 = \bar{y}_{2.}$로 나타나며 $\widetilde{\tau}$는 반복적인 방법에 의하여 구할 수 있다. 더불어 식 (2.11)에 나타난 $\widetilde{V}(\mu_3)$는 H_0^b하에서 얻어진 $\widetilde{S}(\mu_3)$의 점근적 분산으로 정보행렬의 추정치의 역행렬을 통하여 얻을 수 있다. 귀무가설 H_0^b가 맞다는 가정하에서 식 (2.11)에 나타난 스코어 검정통계량 T_b는 점근적으로 표준정규분포를 따르게 된다.

2.1. 스코어 검정통계량 T_a와 T_b의 표준정규분포 근사의 적합성 파악

앞 장에서 설명한 각 가설에 대한 스코어 검정통계량 T_a와 T_b의 점근적 분포인 표준정규분포 근사의 적합성을 파악하고자 귀무가설 H_0^a가 참인 경우 및 H_0^b가 참인 경우를 고려하여 모의실험을 실시하였다. 모의실험 방법에 대한 자세한 설명은 4장에 나타나있다.

표 2.1: 여러 명목유의수준 하에서 계산된 T_a의 추정된 유의수준

				$N(0,1)$의 각 명목유의수준 하에서 계산된 추정된 유의수준							
				$n=100$				$n=300$			
μ_1	μ_2	μ_3	τ	0.20	0.10	0.05	0.01	0.20	0.10	0.05	0.01
0.5	1.0	0.00	0.0	0.1106	0.0439	0.0155	0.0017	0.1211	0.0548	0.0219	0.0027
		0.10	0.0	0.1567	0.0651	0.0262	0.0032	0.1756	0.0770	0.0332	0.0035
		0.22	0.0	0.1591	0.0701	0.0281	0.0025	0.1734	0.0780	0.0341	0.0049
		0.33	0.0	0.1608	0.0701	0.0263	0.0020	0.1773	0.0762	0.0321	0.0039
		0.40	0.0	0.1602	0.0707	0.0261	0.0034	0.1773	0.0795	0.0344	0.0052
1.5	1.0	0.00	0.0	0.1090	0.0455	0.0170	0.0016	0.1127	0.0494	0.0207	0.0027
		0.14	0.0	0.1518	0.0641	0.0245	0.0027	0.1709	0.0755	0.0327	0.0049
		0.27	0.0	0.1663	0.0717	0.0303	0.0036	0.1746	0.0758	0.0313	0.0044
		0.55	0.0	0.1559	0.0673	0.0265	0.0030	0.1710	0.0777	0.0332	0.0043
		0.82	0.0	0.1654	0.0696	0.0293	0.0034	0.1721	0.0763	0.0331	0.0041

표 2.2: 여러 명목유의수준 하에서 계산된 T_b의 추정된 유의수준

				$N(0,1)$의 각 명목유의수준 하에서 계산된 추정된 유의수준							
				$n=100$				$n=300$			
μ_1	μ_2	τ	μ_3	0.20	0.10	0.05	0.01	0.20	0.10	0.05	0.01
0.5	1.0	0.00	0.0	0.1821	0.1098	0.0698	0.0256	0.1632	0.0886	0.0522	0.0150
		0.05	0.0	0.2392	0.1418	0.0862	0.0302	0.2294	0.1296	0.0661	0.0211
		0.10	0.0	0.2446	0.1492	0.0969	0.0376	0.2307	0.1313	0.0758	0.0215
		0.15	0.0	0.2497	0.1514	0.0963	0.0371	0.2308	0.1277	0.0746	0.0243
		0.20	0.0	0.2598	0.1562	0.1011	0.0386	0.2400	0.1348	0.0795	0.0241
1.5	1.0	0.00	0.0	0.1766	0.1014	0.0659	0.0242	0.1659	0.0919	0.0531	0.0151
		0.05	0.0	0.2362	0.1462	0.0914	0.0337	0.2329	0.1325	0.0743	0.0246
		0.10	0.0	0.2467	0.1473	0.0913	0.0363	0.2368	0.1412	0.0780	0.0243
		0.15	0.0	0.2569	0.1618	0.1029	0.0414	0.2359	0.1350	0.0776	0.0223
		0.20	0.0	0.2619	0.1578	0.0982	0.0364	0.2357	0.1349	0.0751	0.0241

이를 간단히 요약해보면 T_a의 점근적 성질파악을 위해서 (Y_{1i}, Y_{2i})는 $H_0^a : \tau = 0$가 참인 경우, 즉 $\tau = 0$가 성립하는 $\mathrm{BP}(\mu_1, \mu_2, \mu_3)$에서 발생시켰다. 또한 T_b의 점근적 성실을 위해서 (Y_{1i}, Y_{2i})는 $H_0^b : \mu_3 = 0$가 참일 때, 즉 식 (2.10)에 나타난 $\mu_3 = 0$인 BNB 분포에서 생성하였다.

모든 모수조합에서 표본크기 $n = 100$ 및 300이 사용되었으며 총 10,000번의 독립적인 반복을 통하여 명목유의수준 0.01, 0.05, 0.10 및 0.20 하에서 두 검정 T_a와 T_b가 귀무가설을 채택 또는 기각하는지를 살펴보았다. 다음 표 2.1는 $H_0^a : \tau = 0$가 참인 경우 $\mu_1 = 0.5, 1.5$와 $\mu_2 = 1$인 경우 μ_3의 여러 조합에서 계산된 스코어 검정통계량 T_a의 $N(0,1)$ 근사에서 다양한 명목유의수준 하에서 얻어지는 추정된 유의수준을 나타낸다. 아울러 표 2.2는 $H_0^a : \mu_3 = 0$가 참인 경우 스코어 검정통계량 T_b의 추정된 유의수준을 나타낸다. 본 모의실험에서 μ_1과 μ_2의 다른 조합에 대해서도 비슷한 결과를 얻었지만 지면관계상 생략하였다.

표 2.1에 나타난 $H_0^a : \tau = 0$에 대한 단측 스코어 검정통계량 T_a의 표준정규분포 근사

는 소표본에서 명목유의수준을 과소추정하고 있음을 알 수 있다. 일례로 명목유의수준이 0.05인 경우를 고려해보자. 이항분포에서 정규분포로의 근사를 이용하면 10,000번의 반복을 통해 추정된 유의수준이 0.0457보다 작거나 0.0543보다 크게 나타날 가능성은 5% 미만이다. 하지만 표 2.1에서 $n = 300$인 경우라 할지라도 추정된 유의수준이 0.0207에서 0.0344사이에 나타나 명목유의수준을 과소추정하고 있다. 반면 표 2.2의 결과 $H_0^b : \mu_3 = 0$에 대한 스코어 검정통계량 T_b의 $N(0,1)$ 근사는 명목유의수준을 과대추정하고 있음을 알 수 있다. 그러므로 좀 더 정확한 가설검정을 위해서는 스코어 검정통계량 T_a와 T_b에 대한 적절한 보정이 필요하게 된다.

보통 일변량 포아송 회귀모형에서 과대산포에 대한 스코어 검정통계량은 H_0^a의 검정결과와 유사하게 명목유의수준을 과소추정하는 경향을 보인다 (Dean과 Lawless, 1989; Gurmu, 1991). 이와 같은 과소추정의 문제를 해결하기 위하여 Dean과 Lawless (1989)와 Gurmu (1991)는 스코어 검정통계량을 보정한 보정된 스코어 검정(adjusted score test) 통계량을 제안하였고 Jung 등 (2005, 2006)는 붓스트랩 방법의 사용을 제안하였다. 하지만 본 연구 모형은 설명변수가 존재하지 않고 더불어 이변량의 특성상 두 반응변수간 복잡한 상관구조가 존재하므로 Dean과 Lawless (1989)와 Gurmu (1991)의 연구처럼 보정된 스코어 검정통계량을 유도할 수 없다. 그러므로 본 연구에서는 Jung 등 (2005, 2006)의 아이디어를 H_0^a와 H_0^b의 검정에 적용하여 각 가설에서 얻어지는 스코어 검정의 명목유의수준에 대한 과대 및 과소추정의 문제를 해결하고자 한다.

이들이 제안한 일변량 포아송 회귀모형에서 모수적 붓스트랩 방법은 다음과 같이 요약될 수 있다. 사실 모수적 붓스트랩 방법은 귀무가설이 맞는 경우 반응변수는 제로팽창 및 중도절단된 포아송 회귀모형을 따른다는 사실을 이용하는 것이다. 즉, 주어진 표본을 이용하여 해당 포아송 회귀모형의 회귀계수에 대한 ML 추정량을 유도한 후, 이를 이용하여 반응변수에 대한 붓스트랩 표본을 생성하고 그 표본에 대한 스코어 검정통계량 값을 계산한다. 충분히 많은 B개의 붓스트랩 표본에서 구해진 스코어 검정통계량 값을 이용하여 붓스트랩 분포를 만들고 이를 이용하여 과대산포에 대한 가설검정에 이용하는 방법이다.

3. 붓스트랩 방법론

본 장에서는 스코어 검정통계량 T_a와 T_b의 표준정규분포 근사시 발생하는 명목유의수준의 과소추정 및 과대추정의 문제를 해결하기 위하여 모수적 붓스트랩(parametric bootstrap) 방법론의 사용을 제안하고자 한다. 본 연구모형에서 귀무가설 H_0^a가 맞는 경우라면 (Y_{1i}, Y_{2i})는 BP(μ_1, μ_2, μ_3)분포를 따르고 H_0^b가 맞는 경우라면 식 (2.10)에 나타난 축소된 BNB(μ_1, μ_2, τ)분포를 따른다는 사실을 알 수 있으므로 모수적 붓스트랩(parametric bootstrap) 방법의 사용이 적절할 것으로 생각된다. 먼저 가설 $H_0^a : \tau = 0$에 대한 가설검정에 사용되는 스코어 검정통계량 T_a에 대한 다음과 같은 붓스트랩 절차를 고려해보자.

Step 1. 주어진 표본자료 (Y_{1i}, Y_{2i}), $i = 1, \ldots, n$를 이용하여 H_0^a하 (BP 모형)에서 모수 μ_1, μ_2 및 μ_3에 대한 ML 추정량 $\widehat{\mu}_1$, $\widehat{\mu}_2$ 및 $\widehat{\mu}_3$를 구한 후 식 (2.6)에 나타난 스코어 검정통계량 $T_a = \widehat{S}(\tau)/\sqrt{\widehat{V}(\tau)}$를 계산한다.

Step 2. 크기가 n인 붓스트랩 표본 (Y_{1k}^*, Y_{2k}^*), $k = 1, \ldots, n$을 BP($\widehat{\mu}_1$, $\widehat{\mu}_2$, $\widehat{\mu}_3$)로부터 복원 추출 방법에 의하여 추출한다.

Step 3. 각 붓스트랩 표본 (Y_{1k}^*, Y_{2k}^*)로부터 BP 모형에서 모수 μ_1, μ_2 및 μ_3에 대한 ML 추정량을 유도한 후 다음과 같은 붓스트랩 스코어 검정통계량 T_a^*를 유도한다.

$$T_a^* = \widehat{S}_a^*(\tau)/\sqrt{\widehat{V}^*(\tau)}.$$

Step 4. Step 2와 3를 독립적으로 B번 반복시행 한다. 충분히 큰 B회의 붓스트랩 반복에 의하여 얻어진 T_a^*의 값을 이용하여 T_a^*의 $100(1-\alpha)$번째 분위수 $T_a^*(1-\alpha)$를 계산한다.

Step 5. 만일 표본에서 얻어진 스코어 검정통계량 T_a의 값이 $T_a^*(1-\alpha)$보다 크다면 유의수준 α에서 귀무가설 H_0^a를 기각한다.

$H_0^b : \mu_3 = 0$의 가설검정에 사용되는 스코어 검정통계량 T_b에 대한 붓스트랩 절차도 위와 비슷한 방법을 이용하여 실시할 수 있다.

4. 모의실험

앞 장에서 설명한 각 가설에 대한 스코어 검정통계량 T_a와 T_b 및 그에 대응되는 붓스트랩 스코어 검정통계량 T_a^*와 T_b^*의 효율성을 파악하기 위하여 모의실험을 실시하였다. 이때 모의실험은 Jung 등 (2007)에서 사용한 방법과 동일한 방법을 사용하였다. 즉, $\mu_2 = 1$로 고정한 상태에서 μ_1의 값은 0.5에서 2사이에 0.5단위씩 증가시키며 실험하였으며 공분산 모수인 μ_3의 값은 $\mu_3 \le \min(\mu_1, \mu_2)$의 조건을 만족하는 값으로 0과 1사이에서 선택하였다. μ_3 값의 선택방법은 Kocherlakota와 Kocherlakota (1985)가 BP 분포에서 독립성을 검정하기 위하여 사용한 값을 사용하였다. 더불어 산포모수인 τ의 값은 0과 0.2사이에서 0.05단위씩 증가시키며 실험하였다.

반응변수 (Y_{1i}, Y_{2i}), $(i = 1, \ldots, n)$는 주어진 모수조합을 이용하여 BNB(μ_1, μ_2, μ_3, τ) 분포에서 생성하였다. 각 반복에서 가설 H_0^a와 H_0^b에 대한 스코어 검정통계량 T_a와 T_b의 표준정규분포 근사에 의한 유의확률 및 그에 대응되는 붓스트랩 방법의 유의확률을 계산하였다. 이때 붓스트랩 반복은 $B = 1000$을 사용하였다. 모든 모수조합에서 표본크기 $n = 50$, $n = 100$과 200인 경우 총 1,000번의 반복을 통하여 유의수준 $\alpha = 0.01$, 0.05 및 0.10인 경우 각 검정의 추정된 유의수준과 검정력을 계산하였다.

다음 표 4.1은 명목유의수준 $\alpha = 0.05$일때 $\mu_1 = 1.5$이고 $\mu_2 = 1.0$인 경우 $H_0^a : \tau = 0$에 대한 스코어 검정의 $N(0,1)$ 근사와 붓스트랩 방법에 의하여 얻어진 추정된 유의수준과 검정력을 나타내며, 표 4.2는 $H_0^b : \mu_3 = 0$에 대한 스코어 검정의 $N(0,1)$ 근사와 붓스트랩 방법의 추정된 유의수준과 검정력을 나타낸다. 명목유의수준이 $\alpha = 0.01$ 또는 $\alpha = 0.10$인 경우에도 표 4.1과 4.2와 비슷한 결과를 얻었지만 지면관계상 이를 제시하지 않았다.

먼저 표 4.1에 나타난 $H_0^a : \tau = 0$에 대한 스코어 검정통계량 T_a의 효율성에 대하여 살펴보자. 표 4.1에서 각 블럭의 첫 번째 라인에 나타난 값이 귀무가설 H_0^a가 참인 경우이다.

표 4.1: 각 표본수와 모수조합에 따른 $H_0^a : \tau = 0$에 대한 T_a와 T_a^*의 추정된 유의수준 및 검정력 (명목유의수준 $\alpha = 0.05$)

				$n=50$		$n=100$		$n=200$	
				$N(0,1)$	Bootstrap	$N(0,1)$	Bootstrap	$N(0,1)$	Bootstrap
μ_1	μ_2	μ_3	τ	T_A	T_A^*	T_A	T_A^*	T_A	T_A^*
1.5	1.0	0.00	0.00	0.016	0.037	0.024	0.057	0.024	0.044
			0.05	0.040	0.091	0.083	0.138	0.129	0.211
			0.10	0.129	0.207	0.250	0.343	0.444	0.543
			0.15	0.223	0.339	0.395	0.518	0.686	0.758
			0.20	0.344	0.464	0.598	0.689	0.871	0.916
		0.14	0.00	0.018	0.048	0.023	0.050	0.040	0.062
			0.05	0.071	0.128	0.121	0.194	0.163	0.242
			0.10	0.153	0.257	0.312	0.386	0.471	0.556
			0.15	0.288	0.404	0.475	0.577	0.750	0.807
			0.20	0.386	0.507	0.643	0.733	0.915	0.932
		0.27	0.00	0.022	0.050	0.024	0.061	0.028	0.049
			0.05	0.066	0.134	0.119	0.179	0.207	0.275
			0.10	0.153	0.245	0.318	0.411	0.531	0.608
			0.15	0.290	0.396	0.503	0.597	0.791	0.836
			0.20	0.423	0.529	0.722	0.784	0.936	0.947
		0.55	0.00	0.028	0.051	0.036	0.061	0.021	0.042
			0.05	0.099	0.160	0.126	0.194	0.256	0.332
			0.10	0.188	0.277	0.340	0.433	0.577	0.655
			0.15	0.355	0.465	0.596	0.677	0.886	0.918
			0.20	0.485	0.615	0.798	0.865	0.975	0.985
		0.82	0.00	0.023	0.040	0.028	0.046	0.028	0.049
			0.05	0.101	0.168	0.161	0.237	0.277	0.361
			0.10	0.227	0.326	0.440	0.541	0.695	0.757
			0.15	0.401	0.530	0.685	0.755	0.917	0.945
			0.20	0.550	0.666	0.861	0.910	0.991	0.996

표 4.1을 살펴보면 T_a의 표준정규분포 근사는 고려된 모든 표본 수 n에서 명목유의수준을 과소 추정하고 있는 반면 붓스트랩 방법은 이를 제대로 유지하고 있음을 알 수 있다. 더불어 모든 실험에서 검정력도 스코어 검정에 대한 붓스트랩 방법이 표준정규분포 근사보다 높게 나타나고 있다. 표 4.2에 나타난 $H_0^b : \mu_3 = 0$에 대한 모의실험 결과 T_b의 표준정규분포 근사는 명목유의수준을 과대 추정하는 반면 붓스트랩 방법은 이를 제대로 유지하고 있음을 알 수 있다. 검정력에 있어서는 T_b에 대한 붓스트랩 방법이 $N(0,1)$ 근사에 비하여 약간 낮게 나타난다. 이는 T_b의 $N(0,1)$ 근사가 귀무가설 H_0^b를 과도하게 기각하는 경향이 존재하기 때문인 것으로 생각된다. 그러므로 붓스트랩 방법은 T_b의 $N(0,1)$ 근사가 범하는 과도한 기각을 조정하고 있음을 보여주는 것이다.

이상과 같은 모의실험 결과를 토대로 본 논문에서 제안한 붓스트랩 방법은 스코어 검정의 표준정규분포 근사에 존재하는 명목유의수준의 과대 또는 과소 추정 문제를 해결하는 효율적인 대안인 것으로 나타났다.

표 4.2: 각 표본수와 모수조합에 따른 $H_0^b : \mu_3 = 0$에 대한 T_b와 T_b^*의 추정된 유의수준 및 검정력 (명목유의수준 $\alpha = 0.05$)

				$n=50$		$n=100$		$n=200$	
				$N(0,1)$	Bootstrap	$N(0,1)$	Bootstrap	$N(0,1)$	Bootstrap
μ_1	μ_2	τ	μ_3	T_B	T_B^*	T_B	T_B^*	T_B	T_B^*
1.5	1.0	0.00	0.00	0.092	0.053	0.062	0.039	0.059	0.039
			0.14	0.232	0.134	0.235	0.167	0.329	0.273
			0.27	0.393	0.246	0.480	0.365	0.665	0.570
			0.55	0.702	0.538	0.880	0.770	0.990	0.979
			0.82	0.896	0.764	0.986	0.956	1.000	0.999
		0.05	0.00	0.117	0.064	0.099	0.061	0.083	0.049
			0.14	0.247	0.134	0.263	0.165	0.381	0.288
			0.27	0.418	0.263	0.511	0.360	0.706	0.601
			0.55	0.708	0.526	0.886	0.791	0.993	0.980
			0.82	0.906	0.762	0.984	0.970	1.000	0.999
		0.10	0.00	0.117	0.058	0.086	0.045	0.083	0.045
			0.14	0.277	0.137	0.281	0.181	0.364	0.273
			0.27	0.405	0.258	0.497	0.337	0.675	0.565
			0.55	0.699	0.535	0.849	0.739	0.993	0.983
			0.82	0.900	0.781	0.984	0.946	0.998	0.997
		0.15	0.00	0.126	0.050	0.111	0.062	0.079	0.048
			0.14	0.274	0.150	0.263	0.162	0.351	0.262
			0.27	0.414	0.230	0.505	0.346	0.684	0.573
			0.55	0.680	0.483	0.876	0.773	0.979	0.964
			0.82	0.867	0.734	0.982	0.951	1.000	1.000
		0.20	0.00	0.126	0.058	0.107	0.053	0.089	0.051
			0.14	0.259	0.144	0.288	0.165	0.341	0.238
			0.27	0.375	0.208	0.512	0.351	0.656	0.556
			0.55	0.692	0.497	0.882	0.769	0.981	0.961
			0.82	0.871	0.736	0.982	0.943	1.000	0.999

5. 실제 사례분석

여기서는 Kocherlakota와 Kocherlakota (1992, p.295)에 나타난 이변량 계수형 자료를 과대산포와 "내재적 상"에 대한 검정에 적용하고자 한다. 다음 표 5.1의 자료는 Cresswell과 Froggatt (1963)에 처음 나타난 것으로서 아일랜드에서 겹치지 않은 시간 범위동안 발생한 버스운전자들의 사고건수를 나타내는 자료이다.

본 연구에서는 표 5.1의 자료에 과대산포가 존재하는지 또는 "내재적 상관"이 존재하는 지를 앞 절에서 유도한 검정법들을 이용하여 검정하고자 한다. 먼저 표 5.1의 자료에 대한 기초통계량은 다음과 같이 얻어진다.

$$\bar{Y}_{1.} = 1.671, \quad \bar{Y}_{2.} = 1.911, \quad S_1^2 = 1.890, \quad S_2^2 = 2.261, \quad r = 0.420,$$

여기서 S_1^2과 S_2^2은 각각 Y_1과 Y_2의 표본분산을 나타내고 r은 두 변수간 표본상관계수를 나

표 5.1: 79명 버스운전자들의 사고건수

Y_1	Y_2								Total
	0	1	2	3	4	5	6	7	
0	5	6	4	1	1	0	0	0	17
1	4	9	3	4	3	0	0	0	23
2	2	5	5	4	2	0	0	0	18
3	1	6	4	1	1	2	1	0	16
4	0	0	0	2	0	0	0	0	2
5	0	0	1	0	0	0	0	0	1
6	0	0	0	0	1	0	0	1	2
Total	12	26	17	12	8	2	1	1	79

타낸다. 위의 기초통계량들을 살펴보면 두 변수 모두 표본분산이 표본평균보다 약간 크게 나타나고, 두 변수간 상관도 0보다 크게 나타나고 있다.

표 5.1의 자료에 대하여 앞 절에서 유도한 두 검정을 적용해보자. 먼저 과대산포에 대한 스코어 검정통계량 값은 $T_a = 1.032$로 나타나 $N(0,1)$ 근사를 이용한 유의확률은 $p = 0.1511$로 나타났으며 T_a에 대한 붓스트랩 근사를 이용한 유의확률은 $p = 0.1369$로 각각 나타났다. 이때 유의확률을 계산하기 위한 붓스트랩 반복은 $B = 10,000$을 사용하였다. 그러므로 T_a를 이용한 검정결과 표 5.1의 자료에서는 귀무가설 H_0^a를 기각할만한 충분한 증거가 발견되지 않았다. 두 번째로 "내재적 상관"의 존재유무에 대한 스코어 검정통계량 값은 $T_b = 3.041$로 나타났다. T_b에 대한 $N(0,1)$근사에서 얻어지는 유의확률은 $p = 0.0012$로 나타나고 붓스트랩 방법의 의한 유의확률은 $p = 0.0015$로 각각 나타났다. 그러므로 표 4.1의 자료에서는 귀무가설 H_0^b를 기각할만한 충분한 증거가 발견되었다.

이상과 같은 검정결과를 통하여 표 5.1의 자료에는 과대산포는 존재하지 않고 "내재적 상"만 존재하는 이변량 포아송분포를 이용한 적합이 가장 적절할 것으로 생각된다.

6. 결론

본 논문에서는 BNB 분포에서 과대산포와 "내재적 상"을 검정하는데 사용되는 스코어 검정통계량의 표준정규분포 근사가 명목유의수준을 과대 또는 과소추정하는 문제가 있음을 보이고, 이 문제를 해결하기 위한 방법으로 붓스트랩 방법의 사용을 제안하였다. 본 연구에서 제안된 붓스트랩 방법은 명목유의수준을 제대로 유지하여 스코어 검정의 표준정규분포 근사에 존재하는 명목유의수준의 과대 또는 과소 추정 문제를 해결하는 효율적인 대안인 것으로 나타났다.

참고문헌

Bates, G. E. and Neyman, J. (1952). Contributions to the theory of accident proneness, I: An optimistic model of correlation between light and severe accidents, *University of*

California Publications in Statistics, **1**, 215–254.
Cresswell, W. L. and Froggatt, P. (1963). *The Causation of Bus Driver Accidents*, Oxford University Press, London.
Dean, C. and Lawless, J. F. (1989). Tests for detecting overdispersion in poisson regression models, *Journal of the American Statistical Association*, **84**, 467–472.
Gurmu, S. (1991). Tests for detecting overdispersion in the positive poisson regression model, *Journal of Business & Economic Statistics*, **9**, 215–222.
Holgate, P. (1964). Estimation for the bivariate poisson distribution, *Biometrika*, **51**, 241–245.
Johnson, N. L., Kotz, S. and Balakrishnan, N. (1997). *Discrete Multivariate Distributions*, John Wiley & Sons, New York.
Jung, B. C., Jhun, M. and Han, S. M. (2007). Score tests for overdispersion in the bivariate negative binomial models, *Journal of Statistical Computation and Simulation*, To appear.
Jung, B. C., Jhun, M. and Lee, J. W. (2005). Bootstrap tests for overdispersion in a zero inflated poisson regression model, *Biometrics*, **61**, 626–628.
Jung, B. C., Jhun, M. and Song, S. H. (2006). Testing for overdispersion in a censored poisson regression model, *Statistics*, **40**, 533–543.
Kocherlakota, S. and Kocherlakota, K. (1992). *Bivariate Discrete Distributions*, Marcel Dekker, New York.
Kocherlakota, S. and Kocherlakota, K. (2001). Regression in the bivariate poisson distribution, *Communications in Statistics - Theory and Methods*, **30**, 815–825.
Kocherlakota, K. and Kocherlakota, S. (1985). On some tests for independence in non-normal situations: Neyman'S $C(\alpha)$ test, *Communications in Statistics - Theory and Methods*, **14**, 1453–1470.
Marshall, A. W. and Olkin, I. (1990). Multivariate distributions generated from mixtures of convolution and product families, by H.W. Block, A.R. Sampson and T.H. Savits, In *Topics in Statistical Dependence Institute of Mathematical Statics Lecture Notes - Monograph Ser.*, **16**, 372–393.
Subrahmaniam, K. (1966). A test for "intrinsic correlation" in the theory of accident proneness, *Journal of the Royal Statistical Society*, Ser. B, **28**, 180–189.
Subrahmaniam, K. and Subrahmaniam, K. (1973). On the estimation of the parameters in the bivariate negative binomial distribution, *Journal of the Royal Statistical Society*, Ser. B, **35**, 131–146.
Verbeke, G. and Molenberghs, G. (2003). The use of score tests for inference on variance components, *Biometrics*, **59**, 254–262.

[2007년 10월 접수, 2008년 1월 채택]

Testing for Overdispersion in a Bivariate Negative Binomial Distribution Using Bootstrap Method*

Myoungshic Jhun[1)] Byoung Cheol Jung[2)]

ABSTRACT

The bootstrap method for the score test statistic is proposed in a bivariate negative binomial distribution. The Monte Carlo study shows that the score test for testing overdispersion underestimates the nominal significance level, while the score test for "intrinsic correlation" overestimates the nominal one. To overcome this problem, we propose a bootstrap method for the score test. We find that bootstrap methods keep the significance level close to the nominal significance level for testing the hypothesis. An empirical example is provided to illustrate the results.

Keywords: Bivariate poisson, bivariate negative binomial, overdispersion, bootstrap.

---

* This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD, Basic Research Promotion Fund) (KRF-2007-314-C00039).
1) Professor, Dept. of Statistics, Korea University, Seoul 136-701, Korea.
E-mail: jhun@korea.ac.kr
2) Corresponding author. Assistant Professor, Dept. of Statistics, University of Seoul, Seoul 136-743, Korea.
E-mail: bcjung@uos.ac.kr