응용통계연구 (2010)
23(3), 585–594

제로절단된 이변량 일반화 포아송 분포에서 산포모수의 효과 및 산포의 동일성에 대한 검정

이동희[1] · 정병철[2]

[1]경기대학교 경영학과, [2]서울시립대학교 통계학과

(2010년 2월 접수, 2010년 3월 채택)

요 약

본 연구에서는 제로절단된 이변량 일반화 포아송 분포에서 두 반응변수간 산포모수의 효과에 대하여 연구하였다. 모의실험 결과 두 반응변수가 서로 다른 산포를 갖는 경우 이를 무시하는 이변량 포아송 분포나 이변량 음이항 분포에 의한 모형적합은 효율성이 떨어지는 것으로 나타났다. 아울러 본 연구에서는 이와 같은 상이한 산포의 존재유무에 대한 가설검정에서 스코어 검정을 유도하고 우도비 검정과 효율성을 비교하였다.

주요용어: 이변량 일반화 포아송 분포, 제로절단, 이변량 음이항 분포, 스코어 검정, 우도비 검정.

1. 서론

사회현상이나 자연현상으로부터 2개의 변수를 관찰하거나 실험하는 경우가 흔히 발생한다. 관측된 2변량 자료가 이산형인 경우 그의 성격에 맞는 이변량 이산형분포를 사용해야 한다. 이변량 이산확률분포에 대해서는 Kocherlakota와 Kocherlakota (1992) 및 Johnson 등 (1997)에 자세히 나타나 있다. 이변량 계수형 자료는 연구의 특성상 가정된 분포에 비하여 제로의 값이 많이 발생하거나 제로의 값이 아예 발생할 수 없는 경우들이 존재한다. Hamdan (1972), Charambides (1984), Piperigou와 Papageorgiou (2003) 등은 이러한 자료에 대하여 제로절단된 이변량 포아송(Zero truncated bivariate Poisson; ZTBP)분포를 제안하고 이 분포의 성질 및 추정방법에 대하여 연구하였고, Jung 등 (2007)은 이 모형에서 독립성에 대한 검정통계량을 유도하였다. 사실 ZTBP 분포는 평균과 분산이 동일하게 나타나는 이변량 포아송분포를 이용하였으므로 반응변수에 대하여 동일한 산포를 가정한 모형이다. 하지만 실제 얻어지는 대부분의 계수형 자료에는 분산이 평균보다 크거나 작게 나타나는 과대 또는 과소산포의 문제가 존재한다. 만일 제로절단된 이변량 계수형 자료에 과대산포만 존재하는 경우라면 Marshall과 Olkin (1990)에 의하여 제안된 이변량 음이항 분포(bivariate negative binomial; BNB) 또는 Subramaniam과 Subramaniam (1973)에 의하여 제안된 BNB 분포 등을 이용한 제로절단된 BNB(Zero Truncated BNB; ZTBNB) 분포를 사용할 수 있다. 사실 정병철과 전희주 (2007)는 이 분포에서 과대산포에 대한 검정통계량을 유도하였다. 하지만 ZTBNB 모형에는 두 가지의 문제점이 존재한다. 첫 번째 문제점은 이들 모형은 두 반응변수의 과대산포만을 조절할 수 있고 과소산포는 조절하지 못한다는 문제점이다. 즉, 어느 한 반응변수의 분산이 평균보다 작아지는 과소산포의 문제가 발생하는 경우 이들 모형을

이 논문은 2006년도 정부재원(교육인적자원부 학술연구조성사업비)으로 한국학술진흥재단의 지원을 받아 연구되었음(KRF-2006-003-C00054).

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

이용한 적합은 효율적인 결과를 제공해주지 못하게 될 것이다. 사실 제로절단이 존재하는 경우 두 반응변수의 기댓값과 분산의 계산이 복잡해지므로 실제 자료에 과대산포가 존재하는지 또는 과소산포가 존재하는지를 자료에서 얻어진 표본통계량을 이용하여 알아내기는 쉽지 않은 문제이다. 이들 모형의 두 번째 문제점은 이들 모형은 반응변수간 산포의 동일성을 가정해야 한다는 점이다. 이들 모형에서 두 반응변수에 대하여 서로 다른 산포는 허용되지 않는다는 점이다.

본 연구에서는 과대 및 과소산포를 조절할 수 있고 각 반응변수에 서로 다른 산포를 허용할 수 있는 Famoye과 Consul (1995)에 의하여 제안된 이변량 일반화 포아송(Bivariate Generalized Poisson; BGP) 분포를 고려하고자 한다. 본 연구에서는 이 분포에서 제로절단된 BGP(Zero truncated BGP; ZTBGP) 분포를 제안하고, 반응변수에 존재하는 서로 다른 산포의 효과를 알아보고자 한다. 이를 위하여 서로 서로 다른 산포를 무시하는 ZTBP 모형과 ZTBNB 모형의 효율성이 실제 다른 산포를 갖는 자료에서 어떻게 나타나는가를 알아보고자 한다. 아울러 독립성 및 산포의 동일성을 검정하기 위한 스코어 검정통계량을 유도하고 모의실험을 통하여 얻어진 스코어 검정통계량의 소표본 성질을 LR 검정통계량과 비교하고자 한다.

2. 모형

2.1. 이변량 일반화 포아송 분포

먼저 일변량에서 포아송 분포에 비하여 과대 및 과소산포를 조절할 수 있는 일반화 포아송(generalized Poisson; GP)분포는 Consul (1989)에 의하여 다음과 같이 제안되었다.

$$P(y;\theta,\lambda)=\frac{1}{y!}\left(\frac{\mu}{1+\alpha\mu}\right)^{y}(1+\alpha y)^{y-1}\exp\left[-\frac{\mu(1+\alpha y)}{1+\alpha\mu}\right],\quad y=0,1,\ldots. \tag{2.1}$$

확률변수 Y가 식 (2.1)의 확률분포를 갖는 경우 $Y\sim \mathrm{GP}(\mu,\alpha)$라 놓도록 한다. 식 (2.1)에서 α는 산포를 나타내는 모수로 $\alpha=0$인 경우 식 (2.1)은 평균이 μ인 포아송 분포가 되고, $\alpha>0$인 경우 $\mathrm{GP}(\mu,\alpha)$는 포아송 분포에 비하여 과대산포를 나타내며 $\alpha<0$인 경우 $\mathrm{GP}(\mu,\alpha)$는 포아송 분포에 비하여 과소산포를 나타낸다. 이때 $\alpha<0$인 경우 식 (2.1)의 확률분포가 음$(-)$이 아닌 값을 갖기 위하여 α값은 항상 $1+\mu\alpha>0$과 $1+\alpha y>0$을 만족해야 한다 (Consul, 1989). 이제 이와 같은 일변량에서 GP 분포를 이변량으로 확장해보자. Z_1, Z_2와 Z_3를 각각 $\mathrm{GP}(\mu_1,\alpha_1)$, $\mathrm{GP}(\mu_2,\alpha_2)$ 및 $\mathrm{GP}(\mu_3,\alpha_3)$를 따르는 독립적인 확률변수라 하고 Y_1과 Y_2를 다음과 같이 정의해보자.

$$Y_1=X_1+X_2,\qquad Y_2=X_1+X_3. \tag{2.2}$$

식 (2.2)와 같이 확률벡터 (Y_1,Y_2)를 정의하면 (Y_1,Y_2)는 이변량 일반화 포아송(BGP) 분포를 따르게 된다. $\lambda_j=\mu_j/(1+\alpha_i\mu_j)\,(j=1,2,3)$라 놓고 삼각소거법(trivariate reduction)을 적용하면 (Y_1,Y_2)의 결합확률분포는 다음과 같이 얻어진다.

$$P(Y_1=y_1,Y_2=y_2)=e^{-\lambda_1-\lambda_2-\lambda_3-\lambda_1\alpha_1y_1-\lambda_2\alpha_2y_2}\eta(y_1,y_2), \tag{2.3}$$

여기서 $\eta(y_1,y_2)$는 다음과 같다.

$$\eta(y_1,y_2)=\sum_{k=0}^{\min(y_1,y_2)}\frac{\lambda_1^{y_1-k}[1+\alpha_1(y_1-k)]^{y_1-k-1}}{(y_1-k)!}\frac{\lambda_2^{y_2-k}[1+\alpha_2(y_2-k)]^{y_2-k-1}}{(y_2-k)!}\frac{\lambda_3^{k}[1+\alpha_3k]^{k-1}}{k!}$$
$$\times\exp\Big(k(\lambda_1\alpha_1+\lambda_2\alpha_2-\lambda_3\alpha_3)\Big)$$

$$= \sum_{k=0}^{\min(y_1,y_2)} g_1(y_1-k)g_2(y_2-k)g_3(k)\exp\Big(k(\lambda_1\alpha_1+\lambda_2\alpha_2-\lambda_3\alpha_3)\Big). \tag{2.4}$$

식 (2.3)의 확률분포를 갖는 (Y_1, Y_2)에 대하여 $(Y_1, Y_2) \sim \text{BGP}(\mu_1, \mu_2, \mu_3, \alpha_1, \alpha_2, \alpha_3)$라 정의하자. 식 (2.3)에서 μ_3가 0인 경우라면 두 확률변수 Y_1과 Y_2는 서로 독립이 된다. 또한 $\alpha_1 = \alpha_2 = \alpha_3 = 0$인 경우 식 (2.3)에 나타난 Y_1과 Y_2의 확률분포는 이변량 포아송 분포로 축소됨을 쉽게 알 수 있다. 이 분포를 갖는 확률변수의 1차 및 2차 적률은 Famoye과 Consul (1995)에 나타나 있다.

2.2. 제로절단된 BGP(ZTBGP) 분포

보통 이변량에서 제로절단은 세 가지 형태를 갖는다. 이는 두 변수 중 한 변수는 제로절단되고 나머지 한 변수는 제로절단되지 않은 경우, 두 변수 모두 제로절된된 경우 및 $(0,0)$칸에서만 절단된 경우 등이 있다. 논의를 단순화하기 위하여 세 가지 제로절단 중에서 한 가지만 고려해보자. 편의상 Y_1은 제로절단 되었고, Y_2는 그렇지 않다고 가정하자. 이 경우 Y_1과 Y_2의 결합확률분포는 다음과 같이 된다.

$$f_{T2}(y_1,y_2) = \frac{f(y_1,y_2)}{V_{T2}(y_1,y_2)}, \quad y_1 = 1,2,\ldots,\ y_2 = 0,1,2,\ldots, \tag{2.5}$$

여기서 $V_{T2}(y_1, y_2)$는 절단된 제로에 대한 수정항으로 다음과 같이 얻어진다.

$$V_{T2}(y_1,y_2) = \sum_{y_1=1}^{\infty}\sum_{y_2=0}^{\infty} f(y_1,y_2) = 1-\exp(-\lambda_1-\lambda_3). \tag{2.6}$$

이제 n개의 독립적인 관측치 (Y_{1i}, Y_{2i}) $(i = 1,\ldots,n)$를 이용하면 로그우도함수는 다음과 같이 구할 수 있다.

$$\log L = -n(\lambda_1+\lambda_2+\lambda_3) + \sum_{i=1}^{n}\left[-\lambda_1\alpha_1 y_{1i} - \lambda_2\alpha_2 y_{2i} + \log(\eta(y_{1i},y_{2i})) - \log(V_{T2}(y_{1i},y_{2i}))\right]. \tag{2.7}$$

각 모수 μ_1, μ_2, μ_3, α_1, α_2 및 α_3에 대한 최대우도 추정량(ML)은 식 (2.7)에 나타난 로그우도함수를 이용하여 반복적인 방법에 의하여 유도할 수 있다

2.3. 제로절단된 BP(ZTBP) 및 제로절단된 BNB(ZTBNB) 모형

반응변수에 존재하는 서로 다른 산포가 모형적합에 어떻게 영향을 주는지를 파악하기 위하여 산포모수를 고려하지 않는 모형과 두 반응변수에서 산포의 동일성을 가정하는 모형을 고려하고자 한다. 먼저 두 반응변수에서 산포를 고려하지 않는 모형으로는 ZTBP 모형을 들 수 있다 (Hamdan, 1972; Jung 등, 2007). 두 번째 모형으로는 Marshall과 Okin (1990) 형태의 ZTBNB 분포를 들 수 있다.

이 경우 ZTBP 모형과 ZTBNB 모형의 확률분포는 다음과 같다.

$$f^{*}_{BP}(y_1,y_2) = \frac{f_{BP}(y_1,y_2)}{V^{*}_{BP}(y_1,y_2)}, \quad y_1 = 1,2,\ldots,\ y_2 = 0,1,\ldots,$$
$$f^{**}_{BNB}(y_1,y_2) = \frac{f_{BNB}(y_1,y_2)}{V^{**}_{BNB}(y_1,y_2)}, \quad y_1 = 1,2,\ldots,\ y_2 = 0,1,\ldots, \tag{2.8}$$

여기서 $f_{BP}(y_1, y_2)$와 $f_{BNB}(y_1, y_2)$는 각각 제로절단되지 않은 BP와 BNB 분포의 결합확률분포를 나타내고, $V^{*}_{BP}(y_1, y_2)$와 $V^{**}_{BNB}(y_1, y_2)$는 제로절단에 대한 수정항을 나타낸 것으로 다음과 같이 얻어진다

(정병철과 전희주, 2007).

$$
\begin{aligned}
f_{BP}(y_1, y_2) &= \exp(-\mu_1 - \mu_2 - \mu_3) \sum_{k=0}^{\min(y_1, y_2)} \frac{\mu_1^{y_1-k}}{(y_1-k)!} \frac{\mu_2^{y_2-k}}{(y_2-k)!} \frac{\mu_3^{k}}{k!}, \\
V_{BP}^{*}(y_1, y_2) &= \exp(-\mu_1 - \mu_3), \\
f_{BNB}^{**}(y_1, y_2) &= \frac{\Gamma(\tau^{-1} + y_1 + y_2)}{\Gamma(\tau^{-1})\Gamma(y_1)\Gamma(y_2)} \mu_1^{y_1} \mu_2^{y_2} \tau^{-\tau^{-1}} (\tau^{-1} + y_1 + y_2)^{-(\tau^{-1}+y_1+y_2)}, \\
V_{BNB}^{**}(y_1, y_2) &= \sum_{y_1=1}^{\infty} \sum_{y_2=0}^{\infty} f_{BNB}(y_1, y_2) = 1 - (1 + \tau\mu_1)^{-\tau^{-1}}.
\end{aligned} \tag{2.9}
$$

식 (2.9)에 나타난 BNB 분포에서 τ는 과대산포를 나타내는 모수이다. BNB 분포에서 Y_1과 Y_2는 동일한 산포모수에 의존하므로 분포의 특성이 두 변수간 산포의 동일성을 가정하는 모형이다. 반면 BP 분포에는 산포모수가 존재하지 않으므로 두 반응변수에서 평균과 분산이 동일한 특성을 갖는 모형이다.

2.4. 모의실험

본 절에서는 두 반응변수에 서로 다른 산포가 존재할 때 이를 무시하는 모형의 모형적합 결과가 어떻게 나타나는지 알아보기 위하여 모의실험을 실시하였다. 모의실험은 Kocherlakota와 Kocherlakota (1985), Paul 등 (1989) 및 Jung 등 (2007)에서 BP 및 ZTBP 분포에서 독립성을 검정하기 위하여 사용한 모의실험 디자인과 유사한 방법을 적용하였다. 모의실험에서 반응변수 Y_1과 Y_2는 다음과 같은 2가지 분포에서 발생시켰다.

1) 첫 번째 모의실험에서 반응변수 Y_1과 Y_2는 Y_1이 제로절단되고 Y_2는 제로절단되지 않은 제로절단된 ZTBGP$(\mu_1, \mu_2, \mu_3, \alpha_1, \alpha_2, \alpha_3)$에서 발생시켰다.
2) 두 번째 방법에서 반응변수 Y_1과 Y_2는 산포모수를 고려하지 않는 ZTBP(μ_1, μ_2, μ_3)에서 발생시켰다.

두 모의실험 모두 표본수는 $n = 100$과 200을 사용하였다. 위와 같이 2개의 모의실험 디자인을 사용한 이유는 첫 번째 모형은 산포모수의 존재가 모형적합에 어떠한 영향을 미치는가를 알아보기 위함이었으며, 두 번째 모형을 사용한 이유는 산포모수가 존재하지 않는 제로절단된 이변량 계수형 자료에 산포모수를 고려한 모형의 효율성이 어떻게 달라지는가를 알아보기 위함이다. 첫 번째 방법에서는 $\mu_1 = \mu_2 = 1$로 고정하고 $\alpha_1 = \alpha_3 = 0.2$로 고정시킨 상태에서 상관모수 μ_3는 0.0, 0.22, 0.45 및 0.67로 변화시켰으며, Y_2의 산포모수 α_2는 0.2에서 2.0까지 0.6단위로 변화시켜가며 실험하였다. 두 번째 모의실험은 $\mu_1 = 1$로 고정시킨 상태에서 μ_2는 0.5에서 2.0까지 0.5단위로 변화시켰으며 μ_3의 값은 0.0, 0.22, 0.45, 0.67로 변화시켜가며 실험하였다. 각 모수의 조합에서 1,000번의 반복을 실시하였다. 각 반복에서는 각각 ZTBGP 모형, ZTBP 모형 및 ZTBNB 모형에 대한 ML 추정을 이용하여 모수를 추정하고, 모수 추정 후 각 모형의 모형적합기준인 AIC 값을 계산하였다.

다음 표 3.1은 $n = 200$인 경우 ZTBGP 분포에서 $\mu_1 = \mu_3 = 1.0$과 $\alpha_1 = \alpha_3 = 0.2$인 경우 μ_3와 α_2의 값에 따라 고려된 3가지 모형에서 1,000번의 반복에 의하여 얻어진 AIC 값들의 최소값, 최대값 및 평균값을 나타낸다. 본 연구에서는 과소산포($\alpha_1 < 0$ 또는 $\alpha_2 < 0$)가 존재하는 경우에 대한 모의실험도 실시하였다. 하지만 이의 결과는 과대산포($\alpha_1 > 0$ 또는 $\alpha_2 > 0$)가 존재하는 경우와 유사하여 이의 결과는 지면관계상 제시하지 않았다. 아울러 $n = 100$에서의 결과도 $n = 100$의 결과와 유사하여 지면관계상 제시하지 않았다.

표 **2.1.** $\mu_1 = \mu_3 = 1.0$과 $\alpha_1 = \alpha_3 = 0.2$인 경우 각 모형에서 얻어진 AIC의 평균, 최소값 및 최대값($n = 200$)

μ_3	α_2	ZTBP			ZTBNB			ZTBGP		
		평균	최소값	최대값	평균	최소값	최대값	평균	최소값	최대값
0.00	0.2	1052.36	928.11	1218.78	1041.16	921.08	1191.15	1036.26	918.66	1162.69
	0.8	1167.88	942.76	1436.04	1096.06	917.96	1287.92	1035.03	894.93	1167.17
	1.4	1260.04	968.30	1669.55	1113.78	928.16	1343.76	995.08	857.82	1145.31
	2.0	1337.59	962.95	1946.99	1120.81	913.73	1357.64	951.80	815.15	1089.89
0.22	0.2	1146.15	1003.93	1278.70	1134.80	1004.96	1230.44	1135.99	1009.59	1226.05
	0.8	1247.17	1073.56	1462.15	1186.13	1051.58	1332.20	1154.69	1033.24	1260.92
	1.4	1341.17	1096.28	1828.02	1216.26	1048.31	1445.91	1143.54	1019.09	1264.70
	2.0	1414.96	1051.46	2026.40	1228.40	991.78	1472.34	1121.55	947.83	1262.49
0.45	0.2	1225.14	1093.02	1391.54	1213.42	1084.75	1358.49	1215.32	1085.45	1360.44
	0.8	1324.99	1124.71	1516.02	1264.53	1113.53	1431.23	1245.36	1107.79	1365.82
	1.4	1407.94	1172.88	1763.57	1290.87	1134.92	1514.46	1235.83	1107.81	1388.78
	2.0	1476.16	1155.30	1959.62	1306.61	1108.77	1515.64	1221.79	1076.21	1371.90
0.67	0.2	1288.74	1156.34	1427.46	1278.52	1159.86	1391.97	1277.73	1160.59	1394.94
	0.8	1384.29	1203.45	1574.97	1324.94	1181.52	1456.31	1309.75	1176.43	1415.65
	1.4	1464.69	1229.31	1830.39	1352.36	1204.41	1511.47	1306.17	1179.94	1426.18
	2.0	1541.41	1239.19	2017.73	1371.58	1200.42	1591.49	1293.62	1157.54	1431.91

표 **2.2.** $\mu_1 = \mu_3 = 1.0$과 $\alpha_1 = \alpha_3 = 0.2$인 경우 각 모형에서 얻어진 AIC의 평균, 최소값 및 최대값($n = 200$)

μ_3	ZTBP			ZTBNB			ZTBGP		
	평균	최소값	최대값	평균	최소값	최대값	평균	최소값	최대값
0.00	934.23	831.96	1025.06	934.29	831.96	1021.60	938.90	835.54	1027.08
0.22	1040.63	952.56	1132.36	1043.81	964.57	1133.26	1045.27	955.33	1137.42
0.45	1120.39	1018.83	1223.26	1129.94	1030.18	1225.03	1124.93	1020.47	1222.03
0.67	1178.95	1090.09	1283.39	1197.20	1119.18	1293.04	1183.43	1093.33	1288.18

표 2.1의 결과를 살펴보면 고려한 모든 모수조합에서 두 반응변수의 산포를 무시한 ZTBP 모형의 적합이 평균 AIC의 기준을 적용할 때 가장 안 좋은 것으로 나타났다. 이와 같은 ZTBP 모형의 비효율성은 공분산 모수 μ_3의 값에는 크게 영향 받지 않은 반면 산포모수인 α_2의 값이 더 커질수록 좀 더 심해지는 것으로 나타났다. 두 반응변수에 동일한 산포를 가정하는 ZTBNB 모형의 적합은 ZTBP 모형보다는 좋게 나타난 반면 α_2가 0.2보다 큰 값을 갖는 경우(두 반응변수에 산포의 차이가 존재하는 경우) ZTBGP 모형에 비해서는 안 좋은 것으로 나타났다. 이와 같은 ZTBNB 모형의 비효율성은 α_2가 커질수록(두 반응변수의 산포차이가 커질수록) 더 심해지는 것으로 나타났다. 하지만 α_2가 0.2인 경우는 두 반응변수가 같은 산포를 갖는 경우이므로 ZTBNB 모형과 ZTBGP 모형의 적합은 서로 비슷한 것으로 나타났다.

다음 표 2.2는 ZTBP 분포에서 $\mu_1 = \mu_2 = 1$일 때 μ_3의 값에 따라 고려된 3가지 모형에서 1,000번의 반복에 의하여 얻어진 AIC 값들의 최소값, 최대값 및 평균값을 나타낸다.

표 2.2의 결과를 살펴보면 값의 변화에 따라 ZTBP 모형의 적합이 다른 모형을 사용하였을 때보다 약간 나은 것으로 나타났다. 하지만 그 차이는 아주 미미하여 3가지 중 어느 모형을 사용한다 할지라도 모형 적합 결과는 큰 차이가 없을 것이라 생각된다.

두 가지 모의실험을 통하여 다음과 같은 결론을 얻을 수 있다. 제로절단된 이변량 계수형 자료에서 만일 두 반응변수의 산포가 서로 다를 경우 이를 무시하는 ZTBP 모형이나 ZTBNB 모형을 이용한 모형

적합은 그 효율성이 무척 떨어지는 것으로 나타났다. 더불어 두 반응변수가 평균과 분산이 같은 동일산포 자료에서도 ZTBGP 모형의 효율성을 그다지 떨어지지 않는 것으로 나타났다. 그러므로 두 반응변수간 평균-분산 관계를 제대로 파악하지 못한 제로절단된 이변량 계수형 자료에 대한 모형화에 있어서 ZTBGP 모형을 이용한 모형적합은 어느 경우든 만족할만한 결과를 제공해 줄 것이라 생각된다.

3. 산포의 동일성에 검정

앞 절에서 두 반응변수의 서로 다른 산포를 무시하는 경우, 적합 된 모형은 비효율적이라는 결과를 제시하였다. 그러므로 두 반응변수의 산포가 서로 동일한지 아닌지를 검정해보는 것은 모형적합 이전에 필수적으로 실시해야만 한다. 이때 사용되는 가설은 다음과 같다.

$$H_0 : \alpha_1 = \alpha_2 \quad vs. \quad H_1 : \alpha_1 \neq \alpha_2. \tag{3.1}$$

본 연구에서는 식 (3.1)을 검정하기 위한 스코어 검정통계량과 LR검정통계량을 유도하고자 한다. 먼저 $\alpha_1 = \alpha$로 놓고 $\alpha_2 = \alpha + \delta$로 놓는다면 식 (3.1)의 가설은 $H_0 : \delta = 0$과 같이 고쳐 쓸 수 있다.

3.1. 스코어 검정

식 (3.1)에 대한 스코어 검정을 유도하기 위해서는 귀무가설 하에서 계산된 각 모수에 대한 1차 편미분이 필요하다. 먼저 $\log L = \sum_{i=1}^{n} l_i$라 놓는다면, 귀무가설 하에서 l_i에 대한 δ의 1차 미분 값은 다음과 같이 구해진다.

$$\begin{aligned}\frac{\partial l_i}{\partial \delta}\Big|_{H_0} &= \frac{\hat{\mu_2}(1+\hat{\alpha}y_{2i})}{(1+\hat{\alpha}\hat{\mu_2})^2} - \hat{\lambda_2}y_{2i} \\ &\quad + \frac{\sum_{k=0}^{\min(y_{1i},y_{2i})} \hat{a_i}\left(-\hat{\lambda}_2(y_{2i}-k) + \frac{(y_{2i}-k-1)(y_{2i}-k)}{1+\hat{\alpha}(y_{2i}-k)} + \frac{k\hat{\mu}_2}{(1+\hat{\alpha}\hat{\mu}_2)^2}\right)}{\hat{\eta}(y_{1i},y_{2i})},\end{aligned} \tag{3.2}$$

여기서 $\hat{\lambda}_1 = \hat{\mu}_1/(1+\hat{\alpha}\hat{\mu}_1)$, $\hat{\lambda}_2 = \hat{\mu}_2/(1+\hat{\alpha}\hat{\mu}_2)$이고 $\hat{\mu}_1$, $\hat{\mu}_2$, $\hat{\mu}_3$, $\hat{\alpha}$ 및 $\hat{\alpha}_3$는 귀무가설하에서 얻어진 ML 추정량이다. 또한 $\hat{\eta}(y_{1i}, y_{2i})$는 식 (2.4)에 $\alpha_1 = \alpha_2(\delta = 0)$를 대입하면 얻을 수 있으며 $\hat{a}_i$는 다음과 같이 구해진다.

$$\hat{a}_i = \hat{g}_1(y_1 - k)\hat{g}_2(y_2 - k)\hat{g}_3(k)\exp\left(k\left(\hat{\lambda}_1\hat{\alpha} + \hat{\lambda}_2\hat{\alpha} - \hat{\lambda}_3\hat{\alpha}_3\right)\right).$$

다른 모수 μ_1, μ_2, μ_3, α 및 α_3에 대한 1차 편미분 값도 비슷하게 유도할 수 있다. 이제 $\theta = (\mu_1, \mu_2, \mu_3, \alpha, \alpha_3, \delta)'$이라 정의하자. 이 경우 귀무가설 하에서 로그우도함수에 대한 1차 편미분은 l_i에 대한 각 모수의 1차 편미분을 이용하면 $\partial \log L/\partial\theta_j = \sum_{i=1}^{n} \partial l_i/\partial\theta_j,\ j = 1, \ldots, 6$의 식을 이용하면 쉽게 얻을 수 있다. 이 경우 귀무가설 하에서 δ를 제외한 나머지 모수에 대한 1차 편미분 값은 0이 되므로 스코어 검정에는 다음과 같은 δ에 대한 1차 편미분 값만 이용하면 된다.

$$\frac{\partial \log L}{\partial \delta}\Big|_{H_0} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \delta}\Big|_{H_0} = \hat{S}(\delta). \tag{3.3}$$

스코어 검정통계량을 유도하기 위해서는 귀무가설 하에서 계산된 정보행렬(Information matrix)가 필요하다. 하지만 본 모형에서 정보행렬은 귀무가설이 맞는다 할지라도 정확하게 계산되지 않는다. 이 경우 정보행렬의 추정치로 사용할 수 있는 한 가지 방법은 스코어 벡터의 외적(Outer-Product of Gradient; OPG) 방법을 사용하는 것이다 (한상문과 정병철, 2009). 본 모형에서 OPG를 이용하면 다음과

같은 정보행렬의 추정치를 얻을 수 있다.

$$\hat{I}(\theta) = \sum_{i=1}^{n} \frac{\partial l_i^2}{\partial\theta\partial\theta'}. \tag{3.4}$$

식 (3.3)에 나타난 δ에 대한 1차 편미분과 식 (3.4)에 나타난 귀무가설 하에서 계산된 정보행렬을 이용하면 가설 (3.1)을 검정하기 위한 스코어 검정통계량은 다음과 같이 구해진다.

$$T = \hat{S}(\delta)^2 J^{\delta\delta}, \tag{3.5}$$

여기서 $J^{\delta\delta}$는 정보행렬 $\hat{I}(\theta)$의 역행렬에서 δ에 대응하는 대각원소를 나타낸다. 식 (3.5)와 같이 계산된 스코어 검정통계량은 귀무가설이 맞는다는 가정 하에서 근사적으로 자유도가 1인 카이제곱 분포를 따르게 된다.

3.2. LR 검정

LR 검정은 제한모형(H_0)과 비제한 모형에서의 ML 추정량을 동시에 요구하므로 계산과정이 스코어 검정에 비하여 상대적으로 복잡하다. 본 모형에서 제한모형과 비제한 모형의 ML 추정량을 이용하면 가설 (3.1)에 대한 LR 검정통계량은 다음과 같이 구해진다.

$$\text{LRT} = -2[\log(\text{res}) - \log(\text{unres})], \tag{3.6}$$

여기서 $\log(\text{res}) - \log(\text{unres})$는 각각 제한모형과 비제한모형에서 얻어진 최대로그우도값을 나타낸다. 귀무가설이 맞다는 가정하에서 식 (3.6)에서 얻어진 LR 검정통계량은 근사적으로 자유도가 1인 카이제곱 분포를 따르게 된다.

3.3. 모의실험

본 연구에서 얻어진 동일산포에 대한 스코어 검정과 LR 검정의 효율성을 파악하기 위하여 모의실험을 실시하였다. 모의실험 디자인은 앞 절 모의실험에서 방법 1과 비슷한 방법을 적용하였다. 모의실험에서 반응변수 Y_1과 Y_2는 Y_1이 제로절단되고 Y_2는 제로절단되지 않은 ZTBGP($\mu_1, \mu_2, \mu_3, \alpha_1, \alpha_2, \alpha_3$)에서 발생시켰다. 각 모수에서 $\mu_1 = \mu_2 = 1$ 및 $\alpha_1 = \alpha_3 = 0.2$로 고정시킨 상태에서 μ_3의 값은 0.00, 0.22, 0.45, 0.67로 변화시켜가며 실험하였으며 α_2의 값은 0.2에서 1.0까지 0.2단위로 변화시켜가며 실험하였다. 이때 $\alpha_2 = 0.2$인 경우가 귀무가설을 만족하는 경우이다. 모든 실험조합에서 표본수 $n = 100$ 및 200을 사용하였으며, 1000번의 반복을 통하여 명목유의수준 0.05하에서 추정된 유의수준과 검정력을 계산하여 표 3.1에 나타내었다.

표 3.1의 결과에서 각 칸의 가장 윗 라인에 나타난 결과가 귀무가설($H_0 : \alpha_1 = \alpha_2$ 또는 $H_0 : \delta = 0$)이 참인 경우이다. 명목유의수준이 0.05인 경우 이항분포에서 정규분포로의 근사를 이용하면 1,000번의 반복을 통하여 추정된 유의수준이 0.036보다 작거나 0.064보다 크게 나타날 가능성은 5% 미만이다. 이와 같은 기준을 통해 살펴보면 본 연구에서 제안한 스코어 검정과 LR 검정은 모든 모수조합에서 명목유의수준을 제대로 유지하는 것으로 나타났다. 두 검정의 검정력은 LR 검정이 스코어검정에 비하여 약간 큰 것으로 나타났다. 두 검정모두 δ의 값이 커질수록(두 반응변수의 산포모수가 큰 차이를 보이는 경우) 검정력이 커지고 있음을 알 수 있다. 이상과 같은 모의실험 결과를 통하여 ZBBGP 분포에서 산포모수의 동일성에 대한 검정방법으로는 스코어 검정이나 LR 검정 모두 적절한 검정방법으로 나타났다. 하지만 LR 검정은 비제한 모형과 제한 모형에서의 MLE를 동시에 요구하기 때문에 스코어 검정보다 계산과정이 훨씬 복잡하다. 그러므로 본 연구에서는 산포모수의 동일성에 대한 검정시 계산이 간편하고 효율성도 떨어지지 않는 스코어 검정의 사용을 제안하고자 한다.

표 **3.1.** 산포의 동일성에 대한 각 검정의 추정된 유의수준과 검정력(명목유의수준 0.05, $\mu_1 = \mu_2 = 1.0$, $\alpha_1 = 0.2$, $\alpha_3 = 0.2$)

μ_3	$\alpha_2(\delta)$	$n = 100$		$n = 200$	
		Score test	LR test	Score test	LR test
0.00	0.2(0.0)	0.045	0.048	0.057	0.056
	0.4(0.2)	0.162	0.166	0.260	0.266
	0.6(0.4)	0.311	0.327	0.498	0.502
	0.8(0.6)	0.444	0.451	0.644	0.654
	1.0(0.8)	0.540	0.542	0.800	0.810
0.22	0.2(0.0)	0.045	0.050	0.052	0.054
	0.4(0.2)	0.134	0.144	0.214	0.210
	0.6(0.4)	0.307	0.321	0.497	0.500
	0.8(0.6)	0.478	0.498	0.701	0.712
	1.0(0.8)	0.609	0.628	0.842	0.860
0.45	0.2(0.0)	0.047	0.045	0.044	0.050
	0.4(0.2)	0.142	0.150	0.239	0.244
	0.6(0.4)	0.323	0.334	0.598	0.602
	0.8(0.6)	0.522	0.538	0.821	0.832
	1.0(0.8)	0.699	0.708	0.936	0.940
0.67	0.2(0.0)	0.044	0.042	0.056	0.046
	0.4(0.2)	0.150	0.152	0.268	0.272
	0.6(0.4)	0.336	0.345	0.598	0.606
	0.8(0.6)	0.555	0.564	0.866	0.874
	1.0(0.8)	0.729	0.747	0.954	0.956

4. 결론

본 연구에서는 제로절단된 이변량 계수형 자료에서 두 반응변수간 산포모수의 효과에 대하여 연구하였다. 두 반응변수에 서로 상이한 산포가 존재하는 경우 이를 무시하는 제로절단된 이변량 포아송 분포 및 이변량 음이항 분포의 효율성을 알아보았다. 모의실험 결과 두 반응변수간 상이한 산포를 갖는 경우 제로절단된 이변량 포아송 분포나 이변량 음이항 분포에 의한 모형적합은 효율성이 떨어지는 것으로 나타났다. 아울러 본 연구에서는 두 반응변수에 존재하는 산포모수의 동일성에 대한 가설검정을 다루었다. 스코어 검정통계량과 LR 검정통계량을 유도하고 모의실험을 통하여 두 검정의 효율성을 비교하였다. 모의실험 결과 스코어 검정과 LR 검정은 명목유의수준을 제대로 유지하고 검정력도 높게 나타나 산포모수의 동일성에 검정에 적절한 검정방법으로 나타났다.

참고문헌

정병철, 전희주 (2007). 제로 절단된 이변량 음이항 분포에서 독립성에 대한 검정, <Journal of the Korean Data Analysis>, **9**, 2947–2957.

한상문, 정병철 (2009). 제로팽창된 이변량 음이항 분포에서 제로팽창에 대한 가설검정, <Journal of the Korean Data Analysis>, **11**, 1041–1050.

Charambides, C. A. (1984). Minimum variance unbiased estimation for zero class truncated bivariate poisson and logarithmic series distribution, *Metrika*, **31**, 115–123.

Consul, P. C. (1989). *Generalized Poisson Distributions: Properties and Applications*, Marcel Dekker, New York.

Famoye, F. and Consul, P. C. (1995). Bivariate generalized poisson distribution with some applications, *Metrika*, **42**, 127–138.

Hamdan, M. A. (1972). Estimation in the truncated bivariate poisson distribution, *Technometrics*, **14**, 37–45.

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1997). *Discrete Multivariate Distributions*, John Wiley & Sons, New York.

Jung, B. C., Han, S. M. and Lee, J. (2007). Score tests for testing independence in the zero-truncated bivariate poisson models, *Journal of Statistical Computation and Simulation*, **36**, 599–611.

Kocherlakota, K. and Kocherlakota, S. (1985). On some tests for independence in nonnormal situations: Neyman's test, *Communications in Statistics - Theory and Methods*, **14**, 1453–1470.

Kocherlakota, S. and Kocherlakota, K. (1992). *Bivariate Discrete Distributions*, Marcel Dekker, New York.

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**.

Paul, S. R., Liang, K. Y. and Self, S. G. (1989). On testing departure from the binomial and multinomial assumptions, *Biometrics*, **45**, 231–236.

Piperigou, V. E. and Papageorgiou, H. (2003). On truncated bivariate discrete distributions: A unified treatment, *Metrika*, **58**, 221–233.

Subrahmaniam, K. and Subrahmaniam, K. (1973). On the estimation of the parameters in the bivariate negative binomial distribution, *Journal of the Royal Statistical Society B*, **35**, 131–146.

The Effects of Dispersion Parameters and Tests for Equality of Dispersion Parameters in Zero-Truncated Bivariate Generalized Poisson Models

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

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

Abstract

This study, investigates the effects of dispersion parameters between two response variables in zero-truncated bivariate generalized Poisson distributions. A Monte Carlo study shows that the zero-truncated bivariate Poisson and negative binomial models fit poorly wherein the zero-truncated bivariate count data has heterogeneous dispersion parameters on dependent variables. In addition, we derive the score test for testing the equality of the dispersion parameters and compare its efficiency with the likelihood ratio test.

Keywords: Bivariate generalized Poisson distribution, bivariate negative binomial distribution, likelihood ratio test, score test, zero-truncation.

This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD, Basic Research Promotion Fund; KRF-2006-003-C00054).

[2]Corresponding author: Associate professor, Department of Statistics, University of Seoul, Jeonnong-Dong 90, Dongdaemun-Gu, Seoul 136-743, Korea. E-mail: bcjung@uos.ac.kr