응용통계연구 (2009)
22(4), 819–827

0이 팽창된 포아송 회귀모형을 이용한 기부회수 자료의 재분석

김인영[1] · 박태규[2] · 김병수[3]

[1]Department of Statistics, Virginia Tech, [2]연세대학교 경제학과, [3]연세대학교 응용통계학과

(2009년 4월 접수, 2009년 5월 채택)

요 약

김인영 등 (2006)은 두 개 포아송 분포의 혼합모형에 기초한 회귀모형으로써 2002년 (사)볼런티어 21에서 실시한 설문조사 자료를 분석하여 우리나라 개인들이 기부한 횟수에 영향을 미치는 유의적 변수들을 식별하였다. 본고에서는 김인영 등 (2006)에서도 언급하였듯이 기부횟수 0의 관찰 빈도와 예측 빈도간 차이가 유독 큰 점을 감안하여, 0이 팽창된 포아송(zero inflated Poisson: ZIP)을 기존의 두 개의 포아송 혼합분포에 추가하여 일종의 세 개 포아송 혼합분포 형태로 모집단 분포를 구성하며 동 모형의 회귀모형으로써 기부횟수 자료를 재분석하고자 한다. 회귀계수에 대한 추정은 두 단계 EM 알고리즘으로 이루어 졌고, 유의적 설명 변수의 검색은 김인영 등 (2006)과 같았으나 본 연구에서는 고정된 零군의 비율을 0.201로 추정할 수 있었으며, 두 가지 유의적 설명변수인 소득과 자원봉사 중에서 자원봉사가 기부 횟수를 늘리는 안정적 도구 변수로써 작용할 수 있음을 보고하고 있다.

주요용어: 0이 팽창된 포아송 분포, 포아송의 혼합분포, 기부횟수, EM알고리즘.

1. 서론

김인영 등 (2006)은 두 개 포아송 분포의 혼합모형에 기초한 회귀모형으로써 2002년 (사)볼런티어 21에서 실시한 설문조사 자료를 분석하여 우리나라 개인들이 기부한 횟수에 영향을 미치는 유의적 변수들을 식별하였다. 기부횟수의 경험적 분포로 미루어 모집단을 기부횟수가 적은 집단(작은 군)과 기부횟수가 많은 집단(큰 군)으로 구성하였고, 자연히 두 개 포아송 분포의 혼합분포로써 모집단 분포를 모형화 하였다. 두 개 포아송의 혼합분포에 기초한 회귀모형을 사용하여 반응변수인 기부횟수에 유의적으로 영향을 미치는 설명변수를 식별하였으며, EM알고리즘과 붓스트랩 방법으로 각 회귀계수와 회귀계수의 95% 신뢰구간을 계산하였다. 계산 결과 "작은 군"과 "큰 군" 모두에서 소득과 자원봉사의 경험 유무(이하 자원봉사로 칭함)가 유의적 변수로 확인되었으며, 두 변수 각각에서 회귀계수가 양수로 나타나 소득이 많을 수록, 자원봉사의 경험이 있는 사람 일수록 기부횟수가 증가하는 현상을 설명할 수 있었다. 한편 주목할 만한 점은 소득과 자원봉사의 회귀계수는 "작은 군"이 "큰 군"에 비하여 세배 정도 큰 값으로 나타나고 있는데, 이는 "작은 군"보다 "큰 군" 사람이 기부를 생활화 함으로써 소득과 자원봉사가 기부횟수에 미치는 영향이 상대적으로 적은 것을 뒷받침하여 준다. 본고에서는 김인영 등 (2006)에서도 언급하였듯이 기부횟수 0의 관찰 빈도와 예측 빈도간 차이가 유독 큰 점을 감안하여, 0이 팽창된 포아

이 논문은 2005학년도 연세대학교 학술연구비(2005-1-0179)의 지원에 의하여 이루어진 것임.
[3]교신저자: (120-749) 서울 서대문구 성산로 262 연세대학교 상경대학 응용통계학과, 교수.
E-mail: bskim@yonsei.ac.kr

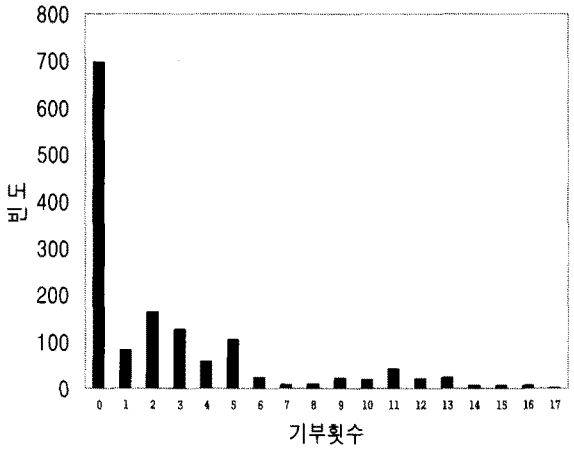

그림 **2.1.** 원자료에서 관찰된 기부 횟수의 분포

송(zero inflated Poisson: ZIP)을 기존의 두 개의 포아송 혼합분포에 추가하여 일종의 세 개 포아송 혼합분포 형태로 모집단 분포를 구성한다. 이러한 모집단 분포를 간략하게 "ZIP+ 두 개 포아송 혼합"으로 부르기로 한다. 본고에서는 ZIP+ 두 개 포아송 혼합분포의 회귀모형으로써 기부횟수 자료를 재분석하고자 한다. 0이 팽창된 포아송(ZIP) 모형은 Johnson 등 (1992)에서도 소개하고 있으며, Lambert (1992), Wang 등 (1996), Böhning 등 (1999), Jung 등 (2005), Lam 등 (2006) 등에서 다루고 있다. 특히 Böhning 등 (1999)은 0이 팽창된 포아송 회귀모형을 아동들의 충치 개수에 적합시켜 6가지 예방 조치에 대한 효과를 비교하였고, Wang 등 (1996)은 포아송 혼합 회귀모형의 추론 과정을 다루면서 포아송의 평균뿐 아니라 혼합비율도 공변수에 의존적인 모형을 구성하였다. 그러나, 필자가 알고 있는 한도에서 ZIP+ 두 개 포아송 회귀모형을 실제 자료에 적용한 보고는 아직 없었다. 본고의 구성은 다음과 같다. 2장에서는 김인영 등 (2006)에서 분석한 자료 및 결과를 간단히 소개하고, 3장에서 ZIP+ 두 개 포아송 혼합분포의 회귀모형을 소개하고, 회귀계수 추정을 위한 EM 알고리즘을 제시한다. 4장에서는 자료 분석 결과와 경제적 의미 그리고 토의사항을 논의한다.

2. 자료 및 김인영 등 (2006)의 분석결과

(사)볼런티어 21은 2002년에 제주도를 제외한 전국에서 만 20세 이상 남녀 1456명을 대상으로 개별 면접을 실시하였고, 지난 1년간 기부횟수와 연령, 소득, 교육정도, 종교, 지역, 직업, 자원봉사(경험 여부) 등을 조사하였다. 본 연구에서 사용하는 기부는 교회, 성당, 절 등의 종교 단체에 내는 헌금은 포함하지 않고, 기타 자선적 목적의 물질적 헌납만을 의미하고 있다. 총 1456명 중 소득에 대한 무응답자 34명을 제외한 1422명을 분석대상으로 하였다. 설문조사 시 얻어진 설명변수는 응답자의 월평균 가구 소득(income), 자원봉사 여부(vol; 1: 예, 0: 아니오), 기부에 대한 태도 변수인 종교적 신념에 의한 기부 여부(atti; 1: 예, 0: 아니오), 50세 이상과 미만 여부(age; 1: 예, 0: 아니오), 대졸 이상의 교육여부(edu; 1: 예, 0: 아니오), 남녀 성별(sex; 1: 남; 0: 여)이 있다. 월평균 가구 소득은 4개의 범주를 가지고 있고, 150만원 미만, 150만원 이상~250만원 미만, 250만원 이상~400만원 미만, 400만원 이상으로 구성되어 있다.

기부 횟수의 분포는 그림 2.1에서 볼 수 있듯이 8회를 기준으로 두 개 군(작은 군, 큰 군)으로 나누어지는 것을 확인할 수 있었다. 김인영 등 (2006)은 이점에 착안하여 두 개 포아송 혼합분포에 기초한 회

표 **2.1.** 혼합 포아송 회귀모형을 사용하여 추정된 회귀계수, 붓스트랩 방법을 이용하여 얻은 회귀계수의 95% 신뢰구간

군(비율)	설명변수	추정된 회귀계수	표준편차	95% 신뢰 구간의 하계	95% 신뢰 구간의 상계	유의성 여부
작은 군 (0.698)	절편	−1.898	0.377	−2.761	−1.251	*
	소득	7.542	2.262	3.714	11.466	*
	자원봉사	0.973	0.209	−0.438	0.366	*
	종교적 신념에 의한기부	−0.038	0.216	−0.438	0.366	
	교육	0.020	0.202	−0.339	0.380	
	나이	0.143	0.183	−0.192	0.463	
	성별	0.123	0.213	−0.250	0.526	
큰 군 (0.302)	절편	1.487	0.124	1.248	1.740	*
	소득	2.337	0.893	0.591	4.386	*
	자원봉사	0.329	0.089	0.154	0.518	*
	종교적 신념에 의한기부	0.124	0.084	−0.056	0.254	
	교육	−0.017	0.095	−0.194	0.152	
	나이	0.040	0.091	−0.151	0.206	
	성별	−0.005	0.093	−0.182	0.183	

귀모형(이하 혼합 포아송 회귀모형이라 부름)을 사용하여 반응변수인 기부회수에 유의적으로 영향을 미치는 설명변수를 식별하였다. EM 알고리즘 (Dempster 등, 1977)을 사용하여 혼합 포아송 회귀모형의 모수를 추정하였으며 동 모수, 즉 회귀계수의 95% 신뢰구간은 붓스트랩 방법과 보완된 BCA(bias-corrected and accelerated; Efron과 Tibshirani, 1993)방법을 사용하였다. 이렇게 얻어진 회귀계수의 추정량과 동 회귀계수의 95% 신뢰구간은 다음 표 2.1과 같다. (단일) 포아송 회귀모형과 혼합 포아송 회귀모형중 어느 모형이 자료에 더 적합한 모형인지를 선택하는 기준으로 다음 식 (2.1)의 아카이케 정보 기준(Akaike Information Crieterion: AIC)를 사용할 수 있다.

$$\mathrm{AIC} = -2\log(L) + 2v \tag{2.1}$$

단, L은 우도 함수이고, ν는 모형에 있는 모수의 갯수를 나타낸다. 한편, AIC는 표본 크기가 작거나 모수의 개수가 많을 때는 과대 적합하는 경향이 있어서 이점을 수정한 다음 식 (2.2)의 AIC_c도 함께 계산을 하였다 (Hurvich와Tasi, 1989).

$$\mathrm{AIC}_c = \mathrm{AIC} + \frac{2v(v+1)}{n-v-1} \tag{2.2}$$

단일 포아송 회귀모형과 혼합 포아송 회귀모형의 AIC(AIC_c)의 값은 각각 8769.227(8769.306), 5955.486(5955.784)로 계산되었다.

3. ZIP+ 두 개 포아송 혼합분포의 회귀모형

ZIP+ 두 개 포아송 혼합 회귀모형을 구성하기 위하여 우선 다음과 같은 표기를 정의하기로 한다. 확률변수 Y가 평균 λ의 포아송 분포를 따를 경우 Y의 확률 질량 함수는 다음 식 (3.1)과 같다.

$$\Pr(Y=y) = \frac{e^{-\lambda}\lambda^{y}}{y!} \equiv \mathrm{Po}(y;\lambda) \tag{3.1}$$

또한 모집단이 각각 평균이 λ_1, λ_2를 갖는 두 개의 포아송 분포의 혼합 분포로 구성되어 있는 경우 Y의 주변 분포는 다음 식 (3.2)와 같다.

$$\Pr(Y=y) = \pi_1 \mathrm{Po}(y;\lambda_1) + \pi_2 \mathrm{Po}(y;\lambda_2) \tag{3.2}$$

단, π_1은 첫 번째 부모집단의 비율이고 동 부모집단은 $\mathrm{Po}(y;\lambda_1)$을 따르고, $\pi_2 = 1-\pi_1$이며, 두 번째 부모집단은 $\mathrm{Po}(y;\lambda_2)$를 따른다. 이제 식 (3.2)에다 0에 질량이 집중된 세 번째 부모집단을 추가하여 다음 식 (3.3)과 같이 ZIP+ 두 개 포아송 혼합분포로써 모집단 분포를 구성할 수 있다.

$$\Pr(Y=y) = \pi_0 I(y \in Z) + (1-\pi_0)[\pi_1 \mathrm{Po}(y;\lambda_1) + \pi_2 \mathrm{Po}(y;\lambda_2)], \tag{3.3}$$

단, Z는 고정된 0(fixed zero)의 값을 갖는 개체들의 집합이며, λ_1, λ_2는 동 집합의 비율을 나타낸다. 식 (3.3)에서 ZIP+ 두 개 포아송 혼합분포의 확률 질량 함수를 다음 식 (3.4)와 같이 나타낼 수 있다.

$$\Pr(Y=y;\underline{\pi},\underline{\lambda}) = \begin{cases} \pi_0 + (1-\pi_0)[\pi_1 e^{-\lambda_1} + \pi_2 e^{-\lambda_2}], & y=0, \\ (1-\pi_0)[\pi_1 \mathrm{Po}(y;\lambda_1) + \pi_2 \mathrm{Po}(y;\lambda_2)], & y>0, \end{cases} \tag{3.4}$$

단, $\underline{\pi} = (\pi_0, \pi_1, \pi_2)$, $\underline{\lambda} = (\lambda_1, \lambda_2)$. 이제 식 (3.4)에 기초한 회귀모형을 구성하기 위하여 λ_1, λ_2를 각각 대수로 변환한 후 다음 식 (3.5), 식 (3.6)과 같이 공변수 벡타 $\underline{x} = (1, x_1, \ldots, x_6)$의 선형결합으로 나타낸다.

$$\log(\lambda_1) = \beta_1{}'\underline{x}, \tag{3.5}$$

$$\log(\lambda_2) = \beta_2{}'\underline{x}, \tag{3.6}$$

단, $\beta_1{}' = (\beta_{10}, \beta_{11}, \ldots, \beta_{16})$, $\beta_2{}' = (\beta_{20}, \beta_{21}, \ldots, \beta_{26})$는 각각 회귀계수 벡타를 나타낸다. 완전 자료를 구성하기 위하여 j번째 개체에 대하여 지시변수인 z_{1j}와 z_{2j}를 각각 다음 식 (3.7)–(3.8)과 같이 정의한다.

$$z_{1j} = \begin{cases} 1, & y_j \in Z, \\ 0, & y_j \sim \pi_1 \mathrm{Po}(\lambda_{1j}) + \pi_2 \mathrm{Po}(\lambda_{2j}), \end{cases} \quad j=1,\ldots,n. \tag{3.7}$$

$$z_{2j} = \begin{cases} 1, & y_j \sim \mathrm{Po}(\lambda_{1j}), \\ 0, & y_j \sim \mathrm{Po}(\lambda_{2j}), \end{cases} \quad j=1,\ldots,n. \tag{3.8}$$

단, Z는 식 (3.3)에서 정의되었듯이 고정된 영의 값을 갖는 개체들의 집합이고, $x_j' = (1, x_{1j}, \ldots, x_{6j})$를 j번째 개체의 공변수 벡타라고 할 때 $\lambda_{1j} = \beta_1' x_j$, $\lambda_{2j} = \beta_2' x_j$이며 n은 표본크기로서 1422이다. 식(3.4)–(3.6)의 모형에서 회귀계수 β_1, β_2와 혼합비율 벡타 $\underline{\pi}$에 대한 추정은 두 단계 EM 알고리즘을 적용하여 계산할 수 있으며 평균 절차(E)와 최대화 절차(M) 각각을 두 단계로 구성하여 $E1$, $E2$, $M1$, $M2$로 표기하기로 한다. k번째 EM 반복후 얻어진 모수 추정량을 $(\pi_0^k, \pi_1^k, \beta_{10}^k, \ldots, \beta_{16}^k, \beta_{20}^k, \ldots, \beta_{26}^k)$, $\beta_1^{k\prime} = (\beta_{10}^k, \ldots, \beta_{16}^k)$, $\beta_2^{k\prime} = (\beta_{20}^k, \ldots, \beta_{26}^k)$로 나타내기로 한다. 이때 $(k+1)$번째 $E1$, $E2$ 단계는 각각 다음 식 (3.9)–(3.11)과 같이 유도될 수 있다.

$E1$:

$$E(z_{1j}=1|Y_j=y_j) = \begin{cases} \dfrac{\pi_0^k}{\pi_0^k + (1-\pi_0^k)\left[\pi_1^k \exp(-\lambda_{1j}^k) + \pi_2^k \exp(-\lambda_{2j}^k)\right]}, & y_j = 0, \\ 0, & y_j > 0 \end{cases} \tag{3.9}$$

$$\equiv \tau_{1j}$$

단, $\log(\lambda_{1j}^k) = \beta_1^{k'} x_1$, $\log(\lambda_{2j}^k) = \beta_2^{k'} x_2$이다.

$E2$:

$$E(z_{2j} = 1|Y_j = y_j) = \frac{\pi_1^k \mathrm{Po}(\lambda_{1j}^k)}{\pi_1^k \mathrm{Po}(\lambda_{1j}^k) + \pi_2^k \mathrm{Po}(\lambda_{2j}^k)} \equiv \tau_{2j.1}^k, \tag{3.10}$$

$$E(z_{2j} = 0|Y_j = y_j) = \frac{\pi_2^k \mathrm{Po}(\lambda_{2j}^k)}{\pi_1^k \mathrm{Po}(\lambda_{1j}^k) + \pi_2^k \mathrm{Po}(\lambda_{2j}^k)} \equiv \tau_{2j.2}^k \tag{3.11}$$

또한 $(k+1)$번째 반복에서 $M1$, $M2$ 단계는 각각 다음과 같다.

$M1$:

$$\pi_1^{k+1} = \frac{\sum_{j=1}^n \tau_{2j.1}^k}{n},$$

$$\pi_2^{k+1} = \frac{\sum_{j=1}^n \tau_{2j.2}^k}{n}$$

이며 $\theta \equiv (\beta_1', \beta_2')'$의 최대우도 추정량은 다음 식 (3.12)는 해로써 얻어진다.

$$\sum_{j=1}^n (1 - \tau_{1j}^k) \frac{\partial \left[\tau_{2j.1}^k \log \mathrm{Po}(y_j; \lambda_{1j}^k) + (1 - \tau_{2j.1}^k) \log \mathrm{Po}(y_j; \lambda_{2j}^k)\right]}{\partial \theta_i} = 0, \quad j = 1, \ldots, 14 \tag{3.12}$$

식 (3.12)는 비선형 방정식이므로 뉴튼-랩슨 방법을 사용하여 최대우도추정량을 계산할 수 있다.

$M2$:

$$\pi_0^{k+1} = \frac{\sum_{j=1}^n \tau_{1j}^{k+1}}{n}.$$

두 단계 EM 알고리즘에서 $|\theta^{k+1} - \theta^k| < 0.0001, |\pi_0^{k+1} - \pi_0^k| < 0.0001, |\pi_1^{k+1} - \pi_1^k| < 0.0001$을 만족할 때까지 반복을 계속한다.

4. 결과, 해석 및 토의

ZIP+ 두 개 포아송 혼합 회귀모형을 이용하여 김인영 등 (2006)의 기부횟수 자료를 재 적합시킨 결과는 다음 표 4.1과 같다. 표 4.1에서 보듯이 고정된 0군의 비율을 0.201로 추정한 것은 본 모형의 특징적 성격에 기인한다. 본 모형에서 밝혀진 것처럼 기부참가에 대한 설문조사가 이루어진 당해연도 기부활동에 참여하지 않은 전체 응답자의 비율이 49%(697) 중 고정적으로 기부에 참여하지 않는 것으로 밝혀진 20.1%의 응답자를 제외한 나머지는 간헐적으로 기부에 참여하고 있으나 당해연도에는 기부에 참여하지 않은 경우이다. 따라서 이들은 잠재적으로는 기부활동에 참가할 수 있는 군으로 해석할 수 있으며 "작은 군"의 기부횟수를 증가시킬 수 있다면 기부활동의 참여자로 전환될 수 있는 군에 포함될 수 있다. 또한 김인영 등 (2006)에서도 지적하였지만 고정된 0군을 모형에 포함함으로써 "작은 군"의 소득의 회귀계수(8.349)가 표 2.1의 회귀계수(7.542)보다 더욱 큰 값을 가지게 되었고, "작은 군"이 소득에 대하여 민감하게 반응을 보인다는 김인영 등 (2006)의 결과를 뒷받침하여 주고 있다. 이는 0군을 추가로 도입한 본 모형에서 기부활동이 일상화된 "큰 군"에 속한 기부자들에 비해 간헐적으로 기부를 수행하는 "작은 군"에 속한 기부자들의 경우 기부활동의 횟수가 소득에 의해 더 크게 영향을 받는다는 결과로서 기존

표 4.1. ZIP+ 두 개 포아송 혼합 회귀모형을 기부횟수 자료에 적합한 결과

군(비율)	설명변수	추정된 회귀계수	표준편차	95% 신뢰 구간의 하계	95% 신뢰 구간의 상계	유의성 여부
固定 零 군 (0.201)						
작은 군 (0.698)	절편	−2.135	0.4331	−0.3077	−1.3573	*
	소득	8.349	2.5277	3.6278	13.6116	*
	자원봉사	1.045	0.2440	0.5837	1.5796	*
	종교적 신념에 의한 기부	−0.059	0.2572	−0.5722	0.4698	
	교육	0.030	0.2589	−0.4956	0.5254	
	나이	0.173	0.2405	−0.2951	0.6410	
	성별	0.155	0.2511	−0.3200	0.6710	
큰 군 (0.302)	절편	1.463	0.1443	1.1584	1.7524	*
	소득	2.287	0.9874	0.4098	4.1683	*
	자원봉사	0.331	0.0996	0.1353	0.5242	*
	종교적 신념에 의한 기부	0.128	0.0960	−0.0487	0.3293	
	교육	−0.017	0.1110	−0.2347	0.2070	
	나이	0.046	0.9987	−0.1573	0.2323	
	성별	−0.014	0.0966	−0.2158	0.1732	

연구에서의 결과를 다시 한 번 확인해주고 있다. 그림 4.1은 원자료의 분포와, 두 개 포아송 혼합 회귀모형에 의한 예측치, ZIP+ 두 개 포아송 혼합 회귀모형에 의한 예측치를 보여주고 있다. 두 개 포아송 혼합 회귀모형에 의한 0빈도의 예측치는 584인 반면 ZIP+ 두 개 포아송 혼합 회귀모형에 의한 0빈도의 예측치는 689로서 실제 관찰치인 697에 매우 근접함을 알 수 있다. 표 4.1의 결과는 표 2.1의 결과와 유의적 설명변수를 꼭 같이 검색하고 있으나 고정된 0군의 추가 도입으로 인하여 소득 변수와 자원봉사 변수의 회귀계수가 변한 것을 주목할 수 있다. 가령 소득에 대한 기부횟수의 탄력도는 동 회귀모형에서 회귀계수에 해당된다. "작은 군"에서 소득에 대한 기부횟수의 탄력도를 e_s^I, 자원봉사에 대한 기부횟수의 탄력도를 e_s^V라고 표기하고, e_b^I, e_b^V는 각각 "큰 군"에서 소득에 대한 기부횟수의 탄력도, 자원 봉사에 대한 기부횟수의 탄력도를 나타낸다고 하자. 이때 각 변수마다 두 군의 상대 탄력도를 e_s/e_b로 나타낼 수가 있는데 소득의 상대 탄력도는 3.64이고 자원봉사의 상대 탄력도는 3.18이다. 이 결과에 따르면 자원봉사변수가 기부횟수를 증가시키는데 있어, 즉, "작은 군"의 기부자들을 "큰 군"의 기부자로 전환시키는데 있어 소득보다 더욱 안정적인 변수로 작용할 수 있다는 것을 나타내 주고 있다. 따라서 기부문화를 정착하는데 필요한 기부횟수의 증가를 위해서는 자원봉사의 경험을 높이는 것이 중요하다는 것을 보여주고 있다. 즉, 자원봉사의 경험이 기부횟수를 증가시키기 위한 소위 "전략" 변수로서 사용될 수 있음을 말해주고 있다. 표 4.2는 (단일)포아송 회귀모형, 두 개 포아송 혼합 회귀모형, 그리고 ZIP+ 두 개 포아송 혼합 회귀모형, 세 개 포아송 혼합 회귀모형 그리고 ZIP+ 세 개 포아송 혼합 회귀모형 각각에 기초한 AIC, AICc를 나타내 주고 있으며, 다섯 개 모형 중 ZIP+ 두 개 포아송 혼합 회귀모형이 가장 적합도가 높은 것을 알 수 있다.

본고에서는 포아송 자료의 과산포 모형으로서 포아송의 혼합분포에 기초한 회귀모형을 고려하였고, 구체적으로는 ZIP+ 두 개 포아송 혼합 회귀모형으로 기부자료를 적합하였다. 그러나 경우에 따라서는 負의 이항분포에 기초한 회귀모형을 적용하여 볼 수도 있으리라 생각하며, 두 모형간 비교는 추후 연구 과제로 남겨 놓기로 한다.

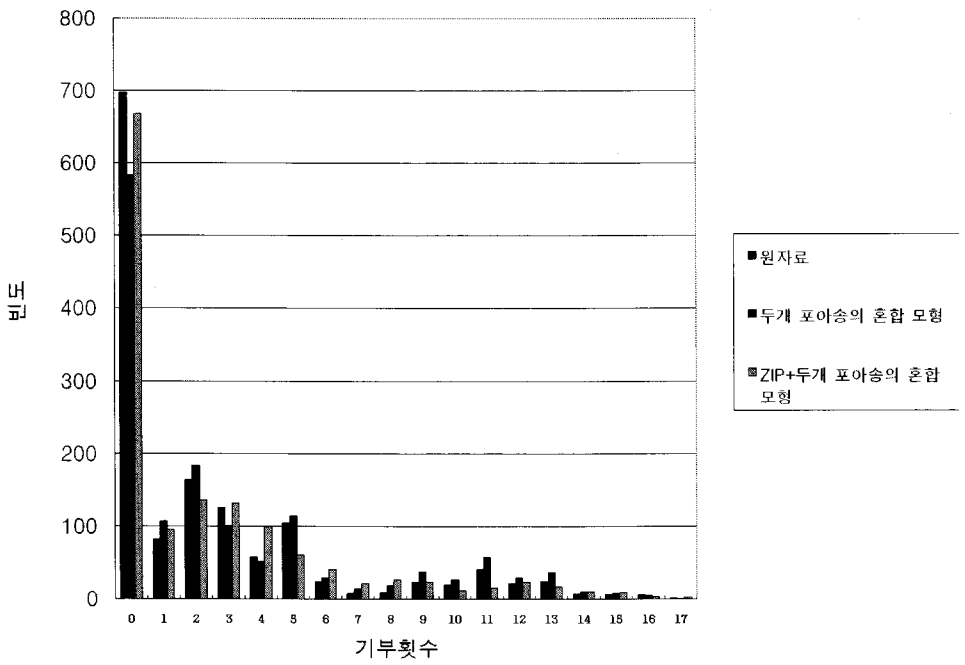

그림 **4.1.** 원자료의 분포, 두 개 포아송 혼합 회귀모형, ZIP+ 두 개 포아송 혼합 회귀모형 각각에 기초한 예측치

표 **4.2.** 네 가지 가능한 포아송 회귀모형에 기부횟수 자료를 적합 시켰을 때 얻어진 AIC, AIC_c.

	(단일)포아송 회귀모형	두 개 포아송 혼합 회귀모형	ZIP+ 두 개 포아송 혼합 회귀모형	세 개 포아송 혼합 회귀모형	ZIP+ 세 개 포아송 혼합 회귀모형
AIC	8769.2	5955.5	3138.3	5949.3	3518.3
AIC_c	8769.3	5955.8	3138.7	5949.9	3519.3

참고문헌

김인영, 박수범, 김병수, 박태규 (2006). 포아송 분포의 혼합모형을 이용한 기부 횟수 자료 분석, <응용통계연구>, **19**, 1–12.

Böhning, D., Dietz, E., Schlattmann, P., Mendonca, L. and Kirchner, U. (1999). The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **162**, 195–209.

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

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*, Chapman & Hall, New York.

Hurvich, C. M. and Tasi, C. L. (1989). Regression and time series model selection in small samples, *Biometrika*, **76**, 297–307.

Johnson, N. L., Kotz, S. and Kemp, A. W. (1992). *Univariate Discrete Distributions*, Second Ed. Wiley, New York.

Jung, B. C., Juhn, M. and Lee, J. W. (2005). Bootstrap tests for overdispersion in a zero-inflated Poisson regression model, *Biometrics*, **61**, 626–629.

Lam, K. F., Xue, H. and Cheung, Y. B. (2006). Semiparametric analysis of zero-inflated count data, *Biometrics*, **62**, 996–1003.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing, *Technometrics*, **34**, 1–14.

Wang, P., Puterman, M. L., Cockburn, I. and Le, N. (1996). Mixed Poisson regression models with covariate dependent rates, *Biometrics*, **52**, 381–400.

The Reanalysis of the Donation Data Using the Zero-Inflated Possion Regression

Inyoung Kim[1] · Tae-Kyu Park[2] · Byung Soo Kim[3]

[1]Department of Statistics, Virginia Tech; [2]Department of Economics, Yonsei University;
[3]Department of Applied Statistics, Yonsei University

(Received April 2009; accepted May 2009)

Abstract

Kim *et al.* (2006) analyzed the donation data surveyed by Voluneteer 21 in year 2002 at South Korea using a Poisson regression based on the mixture of two Poissons and detected significant variables for affecting the number of donations. However, noting the large deviation between the predicted and the actual frequencies of zero, we developed in this note a Poisson regression model based on a distribution in which zero inflated Poisson was added to the mixture of two Poissons. Thus the population distribution is now a mixture of three Poissons in which one component is concentrated on zero mass. We used the EM algorithm for estimating the regression parameters and detected the same variables with Kim et al's for significantly affecting the response. However, we could estimate the proportion of the fixed zero group to be 0.201, which was the characteristic of this model. We also noted that among two significant variables, the income and the volunteer experience(yes, no), the second variable could be utilized as a strategric variable for promoting the donation.

Keywords: EM algorithm, mixture of Poisson distributions, number of donations, zero inflated Poisson(ZIP).

This work was supported in part by Yonsei University Research Fund of 2005(2005-1-0719).
[3]Corresponding author: Professor, Department of Applied Statistics, Yonsei University, 262 Seongsanro, Seodaemun-gu, Seoul 120-749, Korea. E-mail: bskim@yonsei.ac.kr