응용통계연구 제 19권 1호, 2006년, pp. 1–12

포아송 분포의 혼합모형을 이용한 기부 횟수 자료 분석

김인영[1] 박수범[2] 김병수[3] 박태규[4]

요 약

본 논문에서는 2002년에 (사)볼런티어21에서 실시한 설문조사 자료를 이용하여 2001년에 우리나라 개인들이 기부한 횟수에 영향을 주는 유의한 변수들을 식별하였다. 기부횟수의 경험적 분포로 미루어 모집단은 기부를 적게 하는 집단과 많이 하는 집단으로 구성되며 따라서 모집단 분포를 두개 포아송 분포의 혼합분포로 모형화하였다. 이 모형에 기초하여 기부횟수에 영향을 미치는 변수들을 식별하였다. EM알고리즘을 이용하여 모수를 추정하고 2.5%와 97.5%에 기초한 백분위수 신뢰구간을 보완한 BCa(bias-corrected and accelerated) 신뢰구간을 계산하여 유의한 변수들을 찾았다. 연구결과 혼합 포아송 회귀모형에서는 기부횟수가 적은 집단("작은 군")과 기부횟수가 많은 집단("큰 군") 모두에서 소득과 자원봉사의 경험 유무(1:예, 0:아니오)가 기부횟수에 유의적으로 영향을 주는 변수로 밝혀졌다. 또한 두 변수 각각에서 회귀계수가 양수로 나타나 소득이 많을수록, 혹은 자원봉사의 경험이 있는 사람일수록 기부횟수가 증가하는 것을 알 수 있다. 그러나 소득과 자원봉사 변수의 회귀계수는 "작은 군"이 "큰 군"에 비해 더욱 크게 나타나고 있다. "작은 군"보다 "큰 군"의 사람들에게 기부가 생활화되어 있고, 따라서 소득과 자원봉사의 경험 유무가 기부횟수에 미치는 영향이 상대적으로 적은 것으로 파악된다.

주요용어: Akaike의 정보기준 (AIC), EM알고리듬, BC_a 붓스트랩 신뢰구간, 포아송 회귀, 포아송 분포의 혼합모형

1. 서론

개인의 기부참여의 여부에 영향을 주는 유의한 변수들을 식별하거나 기부금액에 영향을 주는 유의한 변수들을 식별하는 연구가 국내외에서 진행되어 왔다.(Smith *et al.*, 1995; Duncan, 1999; Park and Park, 2004). 그러나 외국의 경우와 달리 국내에서는 기부활동이 매우 간헐적으로 특별한 경우에 이뤄지고 있다. 이런 환경에서는 기부문화가 정착될 수 없기 때문에 기부의 참여여부와 더불어 기부참여의 빈도수에 영향을 미치는 변수의 식별이

1) Department of Epidemiology and Public Health, School of Medicine, Yale university, New Haven, CT 06520-8034, U.S.A. 박사후연구원
E-mail: kiy@yumc.yonsei.ac.kr
2) (156-800) 서울시 동작구 노량진동57-1 보건산업진흥원, 연구원
E-mail: avatar74@paran.com
3) (120-749) 서울시 서대문구 신촌동 134, 연세대학교 상경대학 응용통계학과, 교수
E-mail: bskim@yonsei.ac.kr
4) (교신저자) (120-749) 서울시 서대문구 신촌동 134, 연세대학교 상경대학 경제학과, 교수
E-mail: tkpark@yonsei.ac.kr

매우 중요한 의미를 가지게 된다. 그러나 기부횟수에 영향을 주는 변수를 식별하는 연구는 아직 보고된 바 없다. 따라서 본 논문에서는 2002년 (사)볼런티어21에서 실시한 설문조사 자료를 이용해 2001년에 우리나라 개인들이 기부한 횟수에 영향을 미치는 변수들을 식별하고자 한다.

자료는 2002년 7월 3일부터 7월 17일 까지 제주도를 제외한 전국에서 만 20세 이상 남녀 1456명을 대상으로 하였다. 면접원이 가구방문을 하여 개별 면접조사를 하고 기부에 대한 설문조사를 하였다. 이때 기부의 정의는 크게 넓은 의미의 기부와 좁은 의미의 기부로 구별할 수 있다. 좁은 의미의 기부는 교회, 성당, 절 등의 종교단체에 헌금을 제외한 자선적 목적의 물질적 헌납을 의미하고, 넓은 의미의 기부는 좁은 의미의 기부에 종교적 헌금을 추가하는 것으로 정의한다. 설문결과 설문지 응답자 중 50.6%인 737명이 2001년도에 한번 이상 좁은 의미의 기부를 했다고 응답하였다. 그림 1.1은 넓은 의미의 기부(W)와 좁은 의미의 기부(N)에 참여한 사람들의 비율을 나타낸다. 2001년에 넓은 의미의 기부 경험이 있는 사람들의 비율은 61.88%이고, 이것은 아직까지 우리나라에서는 종교적 헌금이 기부에 무시할 수 없는 비중을 차지하고 있음을 보여주는 것이라 하겠다.

본 연구의 설문 자료는 종교 단체의 헌금 액수만 보고하고, 헌금 횟수는 문의하지 않았다. 따라서 본 연구에서는 종교적 헌금을 제외한 좁은 의미의 기부 횟수를 분석대상으로 하며, 앞으로 언급되는 기부는 좁은 의미의 기부임을 밝힌다. 총 설문지 응답자의 성별, 연령별, 사회경제적 특징에 대한 분포와 성별, 연령별, 소득별 범주내에서 기부자의 비율은 표1.1에서 찾아볼 수 있다. 그리고 기부에 참여한 사람들이 기부한 단체의 종류와 비율은 표1.2에서 확인할 수 있는데, 한 사람이 여러 단체에 기부할 수 있으므로 비율의 합은 100%를 넘을 수 있다.

기부횟수가 반응변수이므로 기부횟수의 분포을 포아송 분포로 생각해 볼 수 있다. 그리고 일반화 선형모형에서 연결함수를 로그함수로 이용하여 기부횟수와 관련된 설명변수를 찾을 수 있다. 이 모형을 포아송 회귀모형이라고 하고 자세한 내용은 2절에 요약되어 있다.

총 1456명 중 34명이 소득에 대해서 무응답을 하였고 이들을 제거한 1422명만을 이용하여 자료 분석을 하였다. 이 자료로 부터 얻은 기부횟수의 분포는 그림 1.2과 같이 얻어졌다. 이 그림으로부터 기부횟수의 분포는 단봉 분포가 아님을 알 수 있었다. 즉, 기부횟수 8회를 기준으로 두개의 분포를 그리는 것을 확인할 수 있다. 따라서 기부횟수는 기부를 적게 하는 집단과 많이 하는 집단으로 구성됨을 알 수 있다. 우리는 기부를 적게 하는 집단을 "작은 군"으로 많이 하는 집단을 "큰 군"으로 간단히 부르기로 한다. 그리고 두개 포아송 분포의 혼합분포로 기부횟수를 모형화하고 이 모형에 기초하여 기부횟수에 영향을 주는 유의한 변수들을 찾았다. 우리는 유한개의 포아송 분포의 혼합분포로 반응변수를 모형화한 회귀모형을 간단히 혼합 포아송 회귀모형이라고 부르기로 한다. 이 혼합 포아송 회귀모형을 2절에서 개관하였다. 3절에서는 혼합 포아송 회귀 모형의 모수를 추정하는 방법을 설명하였다. 4절에서는 2002년에 (사)볼런티어 21에서 실시한 기부횟수에 대한 설문조사 자료를 이용하여 기부횟수에 영향을 주는 유의한 요인을 식별하였다. 5절에서는 결론과 토의 및 추후 연구 과제를 논의 하고 있다.

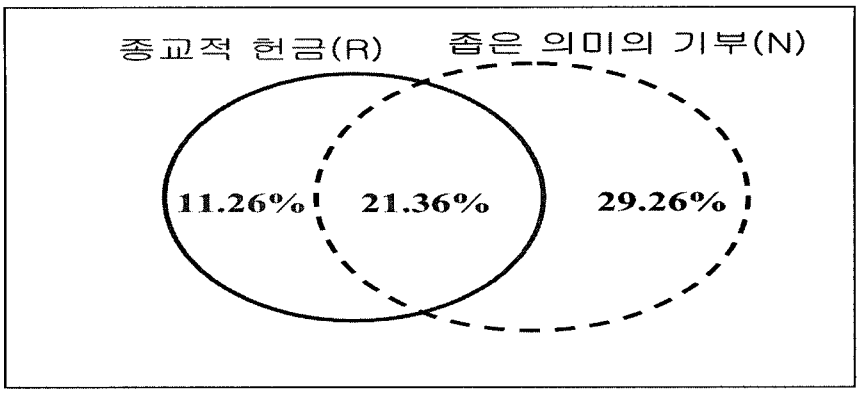

그림 1.1: 넓은 의미의 기부(W)와 좁은 의미의 기부(N)에 참여한 사람들의 비율

Note: 좁은 의미의 기부는 종교적 헌금(R)을 제외한 자선적 목적의 물질적 헌납으로 정의되며, 넓은 의미의 기부는 좁은 의미의 기부에 종교적 헌금을 추가한 것으로 정의 됨. 즉 W=R∪N. R-(R∩N)의 비율 11.26%는 종교적 헌금만 한 사람들의 비율을 나타내고 21.36%는 (R∩N)의 비율을, 29.26%는 N-(R∩N)의 비율을 나타냄.

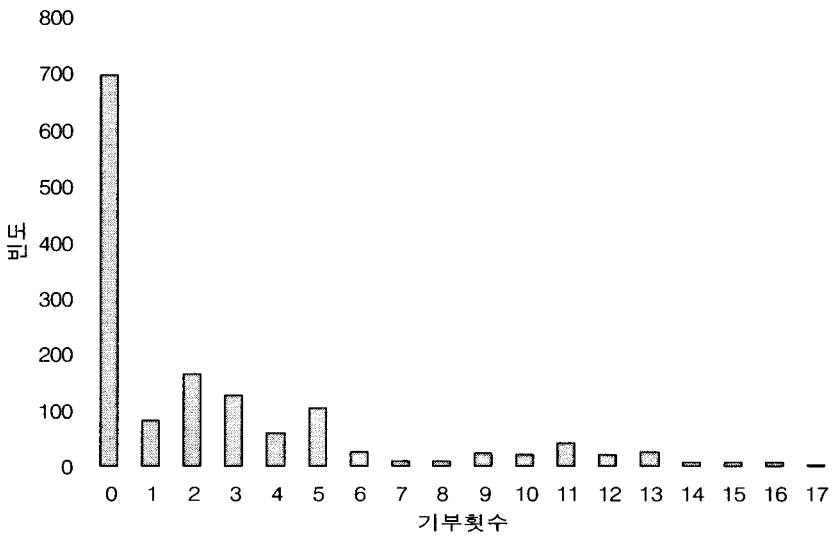

그림 1.2: 원자료에서 관찰된 기부횟수의 막대그래프

표 1.1: 설문지 응답자의 성별, 연령별, 그리고 사회 경제적 특징에 대한 분포와 성별, 연령별 소득별 범주내에서 기부자의 비율

변수	범주	응답자의 비율	기부자의 비율
성별	남	50.4%	52.7%
	여	49.6%	49.7%
연령	20대	22.6%	44.7%
	30대	31.9%	49.8%
	40대	24.4%	57.5%
	50대 이상	21.2%	50.3%
소득	150미만	27.4%	44.4%
(만원)	150이상-250미만	40.1%	49.5%
	250이상-400미만	24.8%	60.1%
	400이상	5.4%	53.9%
	미응답	2.3%	
교육	중학교 이하	18.3%	
	고등학교 이하	45.9%	
	대학교 이상	35.9%	
종교	불교	26.0%	
	기독교	20.0%	
	천주교	8.2%	
	무종교	45.8%	
지역	대도시	49.0%	
	중소도시	37.4%	
	읍/면	13.6%	
직업	농/임/어업	6.6%	
	자영업	18.9%	
	노동자	17.5%	
	영업	17.0%	
	가정주부	25.8%	
	학생	8.9%	
	기타	5.3%	

2. 혼합 포아송 회귀 모형

유한개의 소집단들이 있고 각 집단은 서로 다른 평균을 가지는 포아송 분포를 따른다고 하자. 즉, i번째 집단은 식 (2.1)과 같은 확률밀도함수를 따른다고 하자 $(i = 1, 2, ..., g)$.

$$f_i(y; \theta_i) = \frac{e^{-\mu_i}\mu_i^y}{y!} I_A(y). \tag{2.1}$$

단, A는 음수가 아닌 정수의 집합이고 $I_A(\cdot)$은 지시함수이다. 여기에서 θ_i를 다음으로 표기한다.

$$\theta_i = \log \mu_i = \beta_i^T \mathbf{x}.$$

단, β_i는 i번째 집단의 회귀계수 벡터이고, $\mathbf{x}$는 공변수 벡터이다. 그리고 i번째 집단의 비율을 π_i이라고 하면 g개의 포아송 분포의 혼합분포를 따르는 Y의 확률 밀도 함수는 식 (2.2)와

표 1.2: 2001년에 기부에 참여한 사람이 기부한 단체의 종류와 그 비율

단체	비율
종교단체를 통한 이웃돕기 기부 (수제민돕기, 불우이웃돕기 등)	43.0%
사회복지	23.2%
교육기관	8.4%
환경	4.5%
공익민간	4.1%
국제기구	3.7%
보건의료기관	2.7%
기업 및 민간 재단	2.8%
예술, 문화, 스포츠단체	2.2%
청소년단체	1.9%
오락	1.1%
기타	58.8%

같고 이를 혼합 포아송 회귀모형이라고 부르기로 하자.

$$f(y;\Theta) = \sum_{i=1}^{g} \pi_i f_i(y;\theta_i), \tag{2.2}$$

단, $\mathbf{\Theta} = (\theta_1, ..., \theta_g)$, $\pi_i \geq 0$, $i = 1, ..., g$, $\sum_{i=1}^{g} \pi_i = 1$.

3. 혼합 포아송 회귀모형의 모수 추정 방법

혼합모형의 모수 추정은 EM 알고리즘을 이용하여 얻어진다 (Dempster *et al.*, 1977). Z_{ij}는 j번째 관찰된 Y_j가 i번째 집단에 있는지의 여부에 따라서 1 혹은 0 을 가지고 $\mathbf{Z}_j$는 Z_{ij}을 원소로 가지는 g차원 벡터를 나타낸다고 하자. Z_{ij}과 Y_j의 관찰값을 각각 z_{ij}와 y_j로 표기하기로 한다. 따라서 $\mathbf{Z}_j$를 원소로 갖는 $\mathbf{Z}$는 다항분포를 따른다. EM 알고리즘에서는 $\mathbf{Z}$을 결측치로 간주하여 관찰된 n차원 자료벡터 $\mathbf{Y}$를 미완성자료로 생각한다. 그리고 완성된 자료벡터 $\mathbf{Y}_c$를 $(\mathbf{Y}^T, \mathbf{Z}^T)^T$로 표기하기로 하면 완성자료의 로그 우도함수(L_c)는 다음 식 (3.1)과 같이 구할 수 있다.

$$\log L_c(\Psi) = \sum_{i=1}^{g} \sum_{j=1}^{n} z_{ij} \{\log \pi_i + \log f_i(y_j;\theta_i)\}. \tag{3.1}$$

여기서 $\mathbf{\Psi}$는$(\pi_1, ..., \pi_{g-1}, \pi_g, \Theta^T)^T$이라고 표기하기로 한다. EM 알고리즘은 $\mathbf{Y}$가 주어졌을 때 로그 우도 함수의 조건 기대값을 구하는 E(expectation)단계와 이 단계를 이용하여 최대 우도 추정량을 구하여 모수를 추정하는 M(maximization)단계로 이루어진다.

$\Psi^{(k)}$를 k번째 EM수행 후에 얻은 Ψ의 값이라고 할 때 $k+1$번째 E단계에서 Y_j가 y_j로 주어졌을 때 로그 우도 함수의 Z_{ij}에 대한 조건 기대값을 구하면 다음 식 (3.2)와 같다.

$$\begin{aligned} E_{\Psi^{(k)}}(Z_{ij}|Y_j = y_j) &= \Pr_{\Psi^{(k)}}\{Z_{ij} = 1|Y_j = y_j\} \\ &= \frac{\pi_i^{(k)} f_i(y_j;\theta_i^{(k)})}{\sum_{h=1}^{g} \pi_h^{(k)} f_h(y_j;\theta_h^{(k)})} \\ &= \tau_i(y_j;\Psi^{(k)}) \end{aligned} \tag{3.2}$$

따라서 π_i의 $k+1$번째 예측치인 $\pi_i^{(k+1)}$의 최대 우도 측정치는 식 (3.3) 과 같고

$$\pi_i^{(k+1)} = \frac{\sum_{j=1}^{n} \tau_i(y_j;\Psi^{(k)})}{n} \tag{3.3}$$

$\Psi^{(k+1)}$로 사용될 모수의 최대 우도 추정량은 식 (3.4)을 만족하는 근사해이다.

$$\sum_{i=1}^{g}\sum_{j=1}^{n} \tau_i(y_j;\Psi^{(k)}) \frac{\partial \log f_i(y_j;\theta_i)}{\partial \theta_i} = 0. \tag{3.4}$$

EM 알고리즘에 대한 자세한 내용은 Dempster *et al.* (1977)에 수록되어 있어서 생략한다.

4. 혼합 포아송 회귀모형을 이용한 자료분석

2002년에 (사)볼런티어21에서 기부에 대한 설문조사를 했다. 기부횟수에 영향을 주는 설명변수를 식별하는 것이 이 연구의 목적이다. 우리가 관심 있는 설명변수는 응답자의 월평균 가구소득 (*income*; 4개의 소득구간), 자원봉사의 여부(*vol*; 1:예, 0:아니오), 기부에 대한 태도변수인 종교적 신념에 의한 기부여부(*atti*; 1:예, 0:아니오), 50세 이상과 미만의 여부(*age*; 1:예, 0:아니오), 대졸이상의 교육여부(*edu*; 1:예, 0:아니요), 남녀성별(*sex*; 1:남, 0:여)들이 있다. 월평균 가구소득은 4개의 범주를 가지고 있고 나머지 변수들은 모두 2개의 범주를 가지고 있다.

4.1. 설명 변수의 선택 및 생성

설명변수의 선택과 생성에 대한 자세한 설명은 다음과 같다.

- 월평균 가구소득: 소득수준과 기부행위에 관련한 기존의 실증연구들은 대체적으로 소득수준이 기부행위에 유의미한 정의 효과를 갖는 것으로 보고되고 있다. Park and Park(2004)의 연구에 의하면 우리나라에서의 소득수준은 기부에의 참여여부와 참여수준 모두에 영향을 미치지 못하는 것으로 나타난 반면, Smith et al.(1995)의 연구에서는 소득수준이 기부의 결정에는 영향을 미치지 못하고 기부금액에만 영향을 미친다는 사실을 제시하였다. 본 연구에서는 우리나라의 경우 소득수준이 기부의 횟수에 영향을 미치는가를 확인해보기로 한다. 변수로 사용한 소득은 4개의 구간으로 나누어져 있으며 각 구간은 월 평균 가구소득 150만원 미만, 150만원 이상~250만원 미만, 250만원 이상~400만원 미만, 400만원 이상으로 구성되어 있다.

- 자원봉사의 여부: 민간의 자발적인 기부는 화폐적인 기부만 이루어지는 것이 아니라 시간의 기부(자원봉사)로도 이루어지기 때문에 자선이나 기부행위를 분석하기 위해서는 두가지 형태의 기부가 모두 고려되어야 한다. 즉 화폐적인 기부와 시간적인 기부가 서로 대체재의 관계인지 보완재의 관계인지를 확인할 필요가 있다. Park and Park(2004)의 연구에 의하면 우리나라에서 자원봉사는 기부여부 및 기부수준과 보완적인 관계에 있는 것으로 추정되었다. 따라서 본 연구에서는 자원봉사의 경험을 기부횟수에 영향을 미치는 변수로 선택하는 것은 타당하다고 판단된다. 변수는 자원봉사의 경험이 있으면 1, 없으면 0으로 처리되는 가변수(dummy variable)이다.

- 종교적 신념에 의한 기부여부: 우리나라에서 기부경험이 있는 사람들의 대다수가 종교단체에 기부한 것으로 나타나는 설문결과는 종교가 기부에 큰 영향을 미치고 있음을 반영하고 있다. 따라서 본 변수는 응답자들의 기부 동기가 종교적인 것에서 나왔는지 아닌지를 파악하기 위한 것으로, 종교적인 신념이 기부의 횟수에 영향을 미치는지를 파악하기 위해 선정되었다. 변수는 가변수로서 종교적인 신념에 의하여 기부를 하면 1, 그렇지 않으면 0으로 처리되었다. 단, 이 변수는 기부동기가 종교적인 신념에 의한 것인지 아닌지를 묻는 변수일 뿐 종교단체에 기부했는지의 여부를 나타내는 변수는 아니다.

- 연령: 기부에 관한 외국의 기존 연구에서는 노인일수록 관대하기 때문에 연령이 높을수록 기부에 유의한 영향을 미친다는 결과가 제시되었다. 그러나 Smith et al.(1995)의 연구에서는 연령이 기부의 결정에는 영향을 미치나 기부의 액수(수준) 에는 영향을 미치지 못한다고 하였으며, Park and Park(2004)에서는 우리나라의 경우 연령이 기부결정과 액수에 아무런 영향을 미치지 못하는 것으로 나타났다. 따라서 본 연구에서는 연령이 기부의 횟수에 영향을 주는가를 확인하기 위해 변수로 선정하였으며, 50세 이상이면 1, 미만이면 0인 가변수로 처리하였다.

- 교육수준: 설문에 응답한 사람들의 교육수준에 대한 변수이다. 대졸이상의 학력을 가진 경우 1, 그렇지 않는 경우 0의 값을 갖는 가변수이다. 이를 통해 교육수준이 높을수록 기부의 횟수가 높아질 것인가를 검정해보기로 한다. 일반적으로는 교육수준이 높을수록 사회적인 지위나 수준이 높으므로 자신의 편익을 위해 기부를 한다는 주장을 확인하기 위한 것이다.

- 성별: 남녀의 성별을 나타내는 변수로 남자인 경우 1, 여자인 경우 0으로 처리되는 가변수이다. 가정주부의 비중이 매우 높은 우리나라의 경우 성별이 기부의 횟수에 영향을 주는지를 검정하기 위해 포함하였다.

4.2. 혼합모형의 모수 추정

탐색적 자료 분석으로 전체 자료를 기부 빈도가 8이하인 군("작은 군")과 8을 초과 하는 군("큰 군")으로 나누고, 소득, 자원봉사, 종교적 신념에 의한 기부, 교육, 나이, 성별 각각을 두군간 비교를 하였고 분할표 분석의 카이제곱 분석을 하였다. 분석결과는 표 4.1에 주

어져 있다. 이 결과로 부터 소득, 자원봉사, 그리고 종교적 신념에 의한 기부가 유의한 변수들로 나타났다. 따라서 "작은 군"과"큰 군"이 있다고 판단하여 다음의 혼합 포아송 회귀모

표 4.1: 기부횟수가 8이하인 군("작은 군")과 8을 초과하는 군("큰 군")의 분할표 분석 결과

변수	카이제곱 통계량	P-값
소득	23.374	0
자원봉사	17.661	0
종교적 신념에 의한 기부	13.020	0.0003
교육	0.007	0.9347
나이	0.019	0.8913
성별	0.293	0.5886

형에서 g가 2인 경우에 EM알고리즘을 이용하여 모수추정을 하였다.

$$f(y;\Theta) = \sum_{i=1}^{g}\sum_{j=1}^{n}\pi_i f_i(y_j;\theta_i),$$

$$\theta_i = \beta_{i0} + \beta_{i1}(income) + \beta_{i2}(vol) + \beta_{i3}(atti) + \beta_{i4}(age) + \beta_{i5}(edu) + \beta_{i6}(sex).$$

그리고 1,000번의 붓스트랩표본을 추출하여 각 모수의 표준편차와 2.5%와 97.5%에 기초한 백분위수 신뢰구간을 보완한 BC_a (bias-corrected and accelerated) 신뢰구간을 구성하였다. BC_a에 대한 자세한 내용은 Efron and Tibshirani(1993)에서 찾을 수 있다. 추정된 값들은 표4.2에서 찾아 볼 수 있다.

4.3. 결과 및 집단에 대한 해석

두 개의 혼합 포아송 회귀분석 결과로 추정된 π_1과 π_2는 각각 0.698과 0.303이었다. 즉 기부횟수가 "작은 군"의 비율이 0.698이고 "큰 군"의 비율이 0.303으로 추정되었다. "작은 군"과 "큰 군" 각각에서 소득과 자원봉사의 경험 유무가 기부횟수에 유의적인 영향을 주는 것으로 나타났다. 또한 양쪽 군 모두에게서 두 변수의 회귀계수의 부호가 양수이므로 소득이 많은 사람들과 자원봉사의 경험이 있는 사람들이 더 자주 기부에 참여한다는 것을 확인할 수 있었다. 이중에서 자원봉사의 경험 유무를 나타내는 변수의 회귀계수가 양이라는 것은, 우리나라의 경우 화폐에 의한 기부와 시간에 의한 기부(자원봉사)가 서로 보완적인 관계에 있음을 보여준다고 하겠다. 즉, 자원봉사의 경험이 있는 사람들은 그렇지 못한 사람들에 비해 기부의 의미를 더욱 잘 파악하고 있으며, 이것이 기부횟수에 영향을 준 것으로 판단된다. 그러나 소득변수와 자원봉사 변수 각각의 회귀계수가 "큰 군"에 비해 "작은 군"이 훨씬 크게 나타나는 것을 알 수 있다. 이것은 다음과 같은 해석이 가능하다. 기부횟수가 "작은 군"에서는 기부에 대한 인식과 의미가 확고하지 않기 때문에 자원봉사의 경험을 가진 경우가 그렇지 않은 경우에 비해 기부횟수의 증가에 기여를 많이 하지만, "큰 군"의 경

표 4.2: 두 개 혼합 포아송 회귀분석을 통해서 추정된 모수, 붓스트랩 방법을 이용하여서 얻은 모수의 표준편차와 2.5%와 97.5%에 기초한 백분위수 신뢰구간을 보완한 BC_a (bias-corrected and accelerated) 신뢰구간

군 (비율)	변수	추정된 모수	표준편차	95% 붓스트랩 하위경계	신뢰구간 상위경계	유의 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	0.639	1.389	*
	종교적 신념에 의한 기부	-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	*
	종교적인 신념에 의한 기부	0.124	0.084	-0.056	0.257	
	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	

Note: 두 개($g = 2$)의 혼합 포아송 회귀모형에서 "작은 군"은 기부횟수를 적게 하는 집단, "큰 군"는 기부를 많이 하는 집단, 변수들은 월평균 가구소득(4개의 소득구간), 자원봉사의 여부(1:예, 0:이니요), 종교적 신념에 의한 기부여부(1:예, 0:아니요), 50세 이상의 여부(1:예, 0:아니요), 대졸이상의 교육여부(예:1, 0:아니요), 남녀성별(1:예, 0:아니요)들을 나타내고 *은 유의한 변수를 나타냄.

우는 이미 기부에 대한 의미를 깨닫고, 기부가 이미 생활화되어 있기 때문에 자원봉사가 기부횟수에 미치는 영향의 정도가 상대적으로 낮은 것으로 해석할 수 있다. 그리고 두 군간 소득의 회귀계수의 차이도 비슷한 맥락에서 해석할 수 있다.

두개 혼합 포아송 모형을 이용하여 예측한 기부횟수의 막대그래프와 원자료에서 관찰된 기부횟수의 막대그래프가 그림 4.1에 주어져 있다. 이 그림을 보면 기부횟수 8회를 기준으로 두 개의 분포를 그리는 것을 확인할 수 있다. 기부횟수가 "작은 군"이 훨씬 높은 빈도수를 보이고 있는데, 이는 우리나라에서 아직 기부문화가 본격적으로 정착되지 못하여 비정기적으로 기부하는 사람들의 비중이 훨씬 높다는 사실과 일치한다 하겠다. 반면에 빈도수는 작지만 많은 기부횟수를 갖는 군은 정기적으로 기부를 하는 사람들로 분류할 수 있을 것이다.

포아송 모형과 두개 혼합 포아송 모형 중 어느 모형이 더 적합한지를 선택하기 위해서 다음의 Akaike의 정보기준(AIC)을 사용할 수 있고 이 값이 작은 모형을 선택한다(Akaike, 1974).

$$AIC = -2\log L + 2\nu.$$

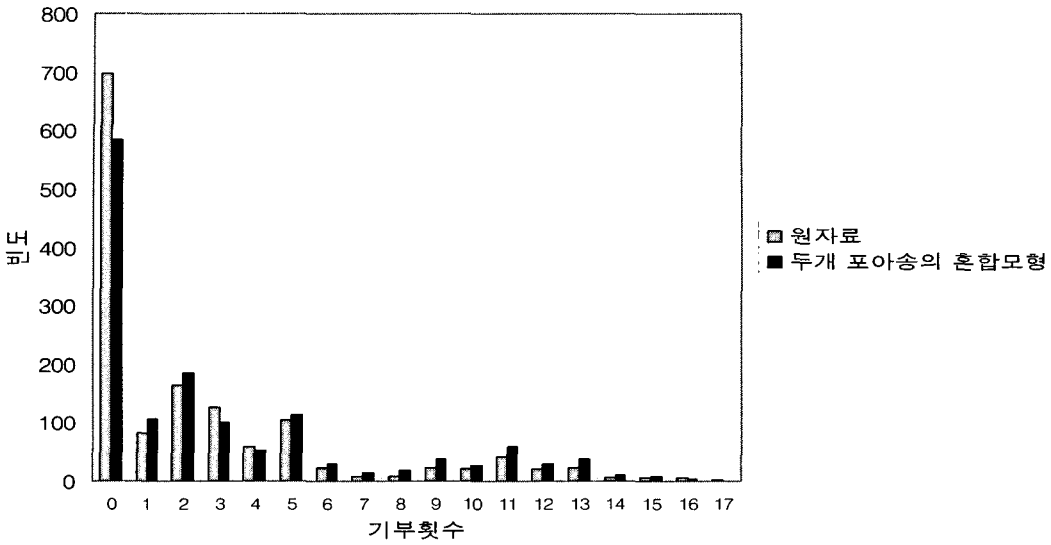

그림 4.1: 원자료에서 관찰된 기부횟수와 포아송 혼합 모형을 이용하여 예측한 기부횟수의 막대그래프

여기서 L는 우도함수이고 ν는 모형에 있는 모수의 수를 나타낸다. 그러나 Akaike의 정보기준(AIC)은 특히 표본의 수가 작거나 모수가 많을 때는 과대적합하는 경향이 있어서 이 점을 수정한 정보이론(AICc)을 계산한다(Hurvich and Tsai, 1989).

$$AICc = AIC + \frac{2\nu(\nu+1)}{n-\nu-1}.$$

단일 포아송 모형과 두개 혼합 포아송 모형의 $AIC(AICc)$ 값은 각각 8769.227(8769.306), 5955.486(5955.784)으로 계산되었다. 세개 혼합 포아송 모형은 이 값이 5949.281(5949.941)로 계산되었으나, 굳이 세개 혼합 포아송 모형을 적용할 이론적 근거가 희박하므로 두개 혼합 포아송 모형을 사용하기로 한다.

5. 요약 및 토의

본 연구에서는 혼합 포아송 회귀모형을 이용하여 기부횟수에 대한 분석을 모형화하였으며, 이 모형을 기초로 하여 기부횟수에 영향을 미치는 설명변수들을 식별하였다. 연구결과 소득과 자원봉사의 경험 유무가 모든 군에 유의적인 영향을 미치는 것으로 나타났다. 국내의 선행 연구에서 개인 기부 참여 여부 또는 기부금액(수준)에 유의적인 영향을 준 것으로 밝혀졌던 자원봉사의 경험 여부는 역시 이번 연구에서도 기부횟수가 "작은 군"과 "큰 군" 모두에 영향을 준 것으로 나타났다. 그러나 기존에 우리나라에서 기부에의 참여 여부

와 기부수준에 영향을 주지 못하는 것으로 밝혀졌던 소득은 기부횟수에는 영향을 주는 것으로 밝혀졌다 (Park and Park, 2004).

그림 4.1에서 0에 대한 관찰 빈도는 697이고, 두개 포아송의 혼합 모형으로 적합된 예측빈도는 584이다. 관찰빈도가 예측빈도보다 큰 이러한 현상은 "0 이 추가된 포아송 모형" (zero inflated Poisson; ZIP)의 고려를 제시하고 있다. 최근 들어 ZIP모형이 포아송 모형의 과산포를 해결하는 방안으로 제시되기도 하였다 (van den Broek, 1995; Böhning *et al.*, 1999). 그러나, 본 분석 자료의 경우 0에 대한 관찰빈도와 예측빈도간 차이가 ZIP모형을 고려한 선행 분석 자료보다 훨씬 작고, 또한 ZIP모형에 대한 경제학적 해석이 용이치 않으므로 본고에서는 ZIP모형에 대한 구체적 논의는 하지 않기로 한다. 단일 포아송 모형에 ZIP 모형을 고려하는 것은 두개 포아송 혼합모형의 특수한 경우가 된다. 비슷한 맥락에서 두개 포아송 혼합모형에 ZIP을 추가하는 것은 결국 세개 포아송 혼합모형을 구성하는 것이 된다. 두개 포아송 혼합 모형으로 부터 ZIP을 고려한 세개 포아송 혼합모형으로의 이탈을 검색하는 문제는 중요한 통계적 문제를 구성하며, 이에 대한 논의는 추후 연구과제로 남겨 놓는다.

참고문헌

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716-723.

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 A*, 162, 194-209.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1997), Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B*, 39, 1-38.

Duncan, B. (1999), Modeling charitable contributions of time and money. *Journal of Public Economics*, 77, 213-242.

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

Hurvich, C. M. and Tsai, C. L. (1989), Regression and Time Series Model Selection in Small Samples. *Biometrika*, 76, 297-307.

McLachlan, G. and Peel, D. (2000), *Finite Mixture Models*, Wiley, New York.

Park, T. and Park, S. (2004), An economic study on charitable giving of individuals in Korea: some new findings from 2002 survey data, presented at the 6th ISTR conference, Toronto, Canada, 11-14 July 2004.

Smith, V. H., Kchoe, M. R. and Creamer, M. E. (1999), The private provision of public goods: Altruism and voluntary giving. *Journal of Public Economics*, 77, 213-242.

van den Broek, Jan. (1995). A score test for zero inflation in a Poisson regression. *Biometrics*, 51, 738-743.

[2005년 8월 접수, 2005년 8월 채택]

The Analysis of the Number of Donations Based on a Mixture of Poisson Regression Model

Inyoung Kim[1] Su-Bum Park[2] Byung Soo Kim[3] Tae-Kyu Park[4]

ABSTRACT

The aim of this study is to analyze a survey data on the number of charitable donations using a mixture of two Poisson regression models. The survey was conducted in 2002 by Volunteer 21, an nonprofit organization, based on Koreans, who were older than 20. The mixture of two Poisson distributions is used to model the number of donations based on the empirical distribution of the data. The mixture of two Poisson distributions implies the whole population is subdivided into two groups, one with lesser number of donations and the other with larger number of donations. We fit the mixture of Poisson regression models on the number of donations to identify significant covariates. The expectation-maximization algorithm is employed to estimate the parameters. We computed 95% bootstrap confidence interval based on bias-corrected and accelerated method and used then for selecting significant explanatory variables. As a result, the income variable with four categories and the volunteering variable (1: experience of volunteering, 0: otherwise) turned out to be significant with the positive regression coefficients both in the lesser and the larger donation groups. However, the regression coefficeitns in the lesser donation group were larger than those in larger donation group.

Keywords: Akaike's Information, Bootstrap Confidence Interval Based on Bias-Corrected and Accelerated Method, Expectation-Maximization Algorithm, Mixture of Poisson Regression models, Poisson Regression

1) Postdoctoral Research Associate, Department of Epidemiology and Public Health, School of Medicine, Yale university, New Haven, CT 06520-8034, USA.
E-mail: kiy@yumc.yonsei.ac.kr
2) Researcher, Korea Health Industry Development Institute 57-1 Noryangjindong, Donjakgu, Seoul, Korea
E-mail: avatar74@paran.com
3) Professor, Department of Applied Statistics, College of Business and Economics, Yonsei University, Korea
E-mail: bskim@yonsei.ac.kr
4) (Corresponding author) Professor, Department of Economics, College of Business and Economics, Yonsei University, Korea
E-mail: tkpark@yonsei.ac.kr