The Korean Journal of Applied Statistics (2014) DOI: http://dx.doi.org/10.5351/KJAS.2014.27.5.743
27(5), 743–753

Reanalysis of 2002 Donation Frequency Data: Corrections and Supplements

Byung Soo Kim[a] · Juhyung Lee[a] · Inyoung Kim[b] · Su-Bum Park[c] · Tae-Kyu Park[d,1]

[a]Department of Applied Statistics, Yonsei University; [b]Department of Statistics, Virginia Tech
[c]Gender Budget Research Center, Korea Women's Development Institute
[d]Department of Economics, Yonsei University

(Received July 29, 2014; Revised September 26, 2014; Accepted September 30, 2014)

Abstract

Kim *et al.* (2006) and Kim *et al.* (2009) reported a set of explanatory variables affecting donation frequency when they analyzed nationwide survey data on donations collected in 2002 by Volunteer 21, a nonprofit organization in Korea. The primary purpose of this paper is to correct computational errors found in Kim *et al.* (2006) and Kim *et al.* (2009), to rectify major results in the Tables and Figures and to supplement Kim *et al.* (2009) by providing new results. We add two logistic regressions to the ZIP and a mixture of two Poisson regressions of Kim *et al.* (2009). Through these two logistic regressions we could detect a set of explanatory variables affecting donation activity (0 or 1) and another set of explanatory variables, in which the volunteer (0, 1) variable is common, discriminating the infrequent donor group from the frequent donor group.

Keywords: Donation, EM algorithm, logistic regression, mixture of two Poisson, Poisson regression, ZIP.

1. 서론

2002년에 (사)볼런티어 21[현 (사)자원봉사문화]은 제주도를 제외한 전국에서 만 20세 이상 남녀 1456명을 개별 면접하여 지난 1년간의 기부횟수와 성별, 연령, 소득, 교육수준, 종교적 신념에 의한 기부 여부, 자원봉사(경험 여부)등과 같은 공변수를 조사하였다. 동 기부횟수 자료를 2002년 기부횟수 자료라고 부르기로 한다. Kim 등 (2006)은 2002년 기부횟수 자료의 히스토그램이 크게 2개 群, 즉 기부횟수가 작은 군(이하 작은 군)과 기부횟수가 큰 군(이하 큰 군)으로 대분되는 점에 착안하여 두개 포아송 분포의 혼합분포에 기초한 포아송 회귀모형을 사용하여 반응변수인 기부횟수에 유의적으로 영향을 미치는 설명변수를 보고 한 바 있다. 당시 익명의 심사위원은 0의 빈도가 특히 높은 점을 주목하여 0이 팽창된 포아송 분포(zero-inflated Poisson; ZIP)의 적용을 추후 연구로 제안한 바 있었다.

Kim 등 (2009)은 2002년 기부횟수 자료에 3개 포아송 분포의 혼합분포를 적합시키되, 첫번째 분포는 0에 집중된 분포를 고려함으로써 ZIP과 두개 포아송 분포의 혼합분포를 적용하였다. 동 적합의 결과로서 기부를 전혀 하지 않는 이른바 고정된 零群의 비율을 0.201로 추정하였고, 작은 군과 큰 군 각각에서 기부횟수에 영향을 미치는 설명변수는 공통적으로 소득과 자원봉사(0, 1)임을 보고하였다.

[1]Corresponding author: Department of Economics, Yonsei University, 50 Yonsei-Ro, Seodaemun-Gu, Seoul 120-749, Korea. E-mail: tkpark@yonsei.ac.kr

본 연구진은 Kim 등 (2006)과 Kim 등 (2009)의 계산 과정에서 오류를 발견하였다. 이에 오류를 수정하여 두 논문의 주요 계산 결과를 다시 보고하고자 한다. 아울러 Kim 등 (2009)에서는 다루지 않았던 내용중 기부행위(0, 1)에 영향을 미치는 설명변수의 검색과 작은 군과 큰 군을 구별하여 주는 설명변수를 파악하여 Kim 등 (2009)의 결과를 보완하고자 한다.

본고의 구성은 다음과 같다. 2장에서 ZIP과 두개 포아송 혼합모형을 소개하고, 동 모형 내에 일반화 선형모형의 정준 연결함수로서 포아송 회귀모형과 로지스틱 회귀모형을 구성한다. 3장에서는 Kim 등 (2006)과 Kim 등 (2009)의 주요 수정 내용을 표와 그림으로서 나타내고, 4장에서는 Kim 등 (2009)의 보완 결과를 보고한다. 5장에서 결론과 토의를 다루기로 한다.

2. ZIP과 두개 포아송 분포의 혼합모형

2002년 기부횟수 자료는 두가지 특징이 있다. 첫째 0의 관찰빈도가 포아송 분포의 기대빈도보다 훨씬 크다는 점이고, 둘째는 기부횟수가 작은 군과 큰 군으로 대분된다는 점이다. 이러한 두가지 특성을 포아송 모형에 반영하기 위하여 ZIP과 두개 포아송 분포의 혼합모형을 구성하기로 하겠으며 우선 몇가지 표기를 정의하기로 한다.

확률변수 Y가 다음 식 (2.1)의 확률질량함수(pmf)를 가질 때 Y는 평균 $\lambda > 0$를 갖는 포아송 분포를 이룬다고 하며 식 (2.1)의 pmf를 $\mathrm{Po}(y;\lambda)$로 나타내기로 한다.

$$\Pr(Y = y) = \frac{e^{-\lambda}\lambda^{y}}{y!} \equiv \mathrm{Po}(y;\lambda). \tag{2.1}$$

단, $y \in \{0\} \cup I^{+}$이고 I^{+}는 양정수의 집합을 나타낸다.

모집단 분포가 $\mathrm{Po}(y;\lambda_1)$과 $\mathrm{Po}(y;\lambda_2)$의 두개 성분으로 구성되고 각 성분의 혼합비율이 π_1, $1-\pi_1 (0 \le \pi_1 \le 1)$일때 모집단 분포는 두개 포아송 분포의 혼합분포를 구성하며 동 pmf는 다음 식 (2.2)와 같다.

$$\Pr(Y = y) = \pi_1 \mathrm{Po}(y;\lambda_1) + (1-\pi_1)\mathrm{Po}(y;\lambda_2). \tag{2.2}$$

단, $0 < \lambda_1 \le \lambda_2$. 이제 식 (2.2)의 모집단에다 세번째 성분을 추가하는데 그 분포가 0에 집중된 분포일 경우를 생각해 본다. 기부를 전혀 하지 않는 이른바 고정된 0군을 Z로 나타내고 고정된 0군의 비율을 $\pi_0 = \Pr(Y \in Z)$로 나타내기로 하자. 기부횟수가 0으로 관찰되는 개체는 Z에 속할 수도 있고 Z의 여집합인 Z^c에 속할 수도 있다. 0이 Z^c에 속할 경우 이를 확률적 0(random zero)이라고 부를 수 있다. 식 (2.2)처럼 작은 군과 큰 군으로 구성된 모집단에 세번째 성분으로 Z가 추가되면 모집단의 pmf는 식 (2.3)과 같고 이를 ZIP과 두개 포아송 분포의 혼합모형이라고 부르며 기호로는 ZIMP_2로 나타내기로 한다.

$$\Pr(Y = y) = \pi_0 I(Y \in Z) + (1-\pi_0)[\pi_1 \mathrm{Po}(y;\lambda_1) + (1-\pi_1)\mathrm{Po}(y;\lambda_2)]I(Y \in Z^c). \tag{2.3}$$

단, $I(\cdot)$는 지시함수를 나타낸다. 식 (2.3)을 다음 식 (2.4)와 같이 나타낼 수 있다.

$$\Pr(Y = y;\boldsymbol{\pi},\boldsymbol{\lambda}) = \begin{cases} \pi_0 + (1-\pi_0)\left[\pi_1 e^{-\lambda_1} + (1-\pi_1)e^{-\lambda_2}\right], & \text{if } y = 0, \\ (1-\pi_0)\left[\pi_1 \mathrm{Po}(y;\lambda_1) + (1-\pi_1)\mathrm{Po}(y;\lambda_2)\right], & \text{if } y > 0. \end{cases} \tag{2.4}$$

단, $\boldsymbol{\pi} = (\pi_0, \pi_1)^T$, $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)^T$이다. $\boldsymbol{x} = (x_1, \ldots, x_p)^T$를 공변수 벡터라고 할 때 혼합비율 π_0, π_1과 포아송 평균 λ_1, λ_2는 일반화 선형모형의 연결함수로서 다음 식 (2.5)–(2.8)과 같이 나타낼 수 있다.

$$\mathrm{logit}(\pi_0) = \boldsymbol{\gamma}_0^T \boldsymbol{x}, \tag{2.5}$$

$$\text{logit}(\pi_1) = \boldsymbol{\gamma}_1^T \boldsymbol{x}, \tag{2.6}$$

$$\log(\lambda_1) = \boldsymbol{\beta}_1^T \boldsymbol{x}, \tag{2.7}$$

$$\log(\lambda_2) = \boldsymbol{\beta}_2^T \boldsymbol{x}. \tag{2.8}$$

단, $\boldsymbol{\gamma}_0$, $\boldsymbol{\gamma}_1$, $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$는 회귀계수 벡터를 나타낸다.

식 (2.5)–(2.8)은 꼭 같은 공변수가 $\boldsymbol{\pi}_0$, $\boldsymbol{\pi}_1$, $\boldsymbol{\lambda}_1$, $\boldsymbol{\lambda}_2$에 영향을 미칠 수 있다는 것을 전제로 한다. 실제로 Lambert (1992)는 용접의 효율과 관련된 실험설계에서 포아송의 평균과 고정된 0군의 비율에 영향을 미치는 공변수가 서로 다른 경우를 고려하고 있다. 그러나, 본 기부자료의 경우 두 모수군에 영향을 미치는 요인이 다르다는 보고가 아직 없는 한, 이미 문헌을 통하여 보고된 사회경제적 요인들을 대상으로 유의적 변수를 검색하고자 한다.

Kim 등 (2006)은 2002년 기부횟수 자료에 식 (2.2)를 적합하였고, Kim 등 (2009)은 식 (2.4), 식(2.7)–(2.8)을 적합하였다. Kim 등 (2009)이 적합한 모형을 ZIP과 두개 포아송 회귀의 혼합모형으로 부르기로 하였다. 본고의 4절에서 보고할 내용은 식 (2.4)–(2.8)을 적합한 내용이다. 식 (2.4)–(2.8)의 모형을 ZIP과 두개 포아송-두개 로지스틱 회귀의 혼합모형으로 부르기로 하겠다.

EM 알고리즘 문맥에서 완전 자료를 구성하기 위하여 j번째 개체 ($j = 1, \ldots, n$)에 대하여 다음 두개의 지시변수 z_{1j}와 z_{2j}를 식 (2.9)–(2.10)과 같이 정의하기로 한다 (Dempster 등, 1977; McLachlan과 Krishnan, 1997).

$$z_{1j} = \begin{cases} 1, & \text{if } Y_j \in Z, \\ 0, & \text{if } Y_j \sim \pi_{1j}\text{Po}(y_j;\lambda_{1j}) + (1-\pi_{1j})\text{Po}(y_j;\lambda_{2j}), \end{cases} \tag{2.9}$$

$$z_{2j} = \begin{cases} 1, & \text{if } Y_j \sim \text{Po}(y_j;\lambda_{1j}), \\ 0, & \text{if } Y_j \sim \text{Po}(y_j;\lambda_{2j}). \end{cases} \tag{2.10}$$

단, $\text{logit}(\pi_{0j}) = \boldsymbol{\gamma}_0^T \boldsymbol{x}_j$, $\text{logit}(\pi_{1j}) = \boldsymbol{\gamma}_1^T \boldsymbol{x}_j$, $\log(\lambda_{1j}) = \boldsymbol{\beta}_1^T \boldsymbol{x}_j$, $\log(\lambda_{2j}) = \boldsymbol{\beta}_2^T \boldsymbol{x}_j$이고, $\boldsymbol{x}_j^T$는 j번째 개체의 공변수 벡터이며 z_{1j}, z_{2j}는 각각 지시함수로서 j번째 개체가 고정된 0군에 속하는지, 작은 군에 속하는지를 나타내준다. n은 표본 크기로서 2002년 기부횟수 자료의 경우 1422였고, p는 절편항을 포함하여 7이다.

식 (2.4)–(2.8)에 기초한 EM 알고리즘을 개략적으로 기술하면 다음과 같다. 우선 $\boldsymbol{y} = (y_1, \ldots, y_n)^T$, $\boldsymbol{z_1} = (z_{11}, \ldots, z_{1n})^T$, $\boldsymbol{z_2} = (z_{21}, \ldots, z_{2n})^T$로 정의하면 완전자료 우도함수는 다음 식 (2.11)과 같이 얻어진다.

$$\begin{aligned} & L_c(\boldsymbol{\gamma}_0, \boldsymbol{\gamma}_1, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2 | \boldsymbol{y}, \boldsymbol{z_1}, \boldsymbol{z_2}) \\ &= \prod_{i=1}^{n} \pi_{0i}^{z_{1i}} \left[(1-\pi_{0i})\pi_{1i}\text{Po}(y_i;\lambda_{1i})\right]^{(1-z_{1i})z_{2i}} \left[(1-\pi_{0i})(1-\pi_{1i})\text{Po}(y_i;\lambda_{2i})\right]^{(1-z_{1i})(1-z_{2i})} \\ &= \prod_{i=1}^{n} \left(\frac{\pi_{0i}}{1-\pi_{0i}}\right)^{z_{1i}} (1-\pi_{0i}) \left[\frac{e^{-\lambda_{1i}}\lambda_{1i}^{y_i}}{y_i!}\right]^{(1-z_{1i})z_{2i}} \pi_{1i}^{(1-z_{1i})z_{2i}} (1-\pi_{1i})^{(1-z_{1i})(1-z_{2i})} \left[\frac{e^{-\lambda_{2i}}\lambda_{2i}^{y_i}}{y_i!}\right]^{(1-z_{1i})(1-z_{2i})}. \end{aligned} \tag{2.11}$$

단, $\pi_{0i} = e^{\gamma_0^T \boldsymbol{x_i}}/(1+e^{\gamma_0^T \boldsymbol{x_i}})$, $\pi_{1i} = e^{\gamma_1^T \boldsymbol{x_i}}/(1+e^{\gamma_1^T \boldsymbol{x_i}})$이다.

식 (2.11)에 로그를 취하면 다음 식 (2.12)와 같이 완전자료 로그우도 함수를 얻는다.

$$l_c(\boldsymbol{\gamma}_0, \boldsymbol{\gamma}_1, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2 | \boldsymbol{y}, \boldsymbol{z_1}, \boldsymbol{z_2})$$

Table 3.1. Corrected table of Table 4.2 of Kim *et al.* (2006).

group (proportion)	explanatory variable	estimated regression coefficient	standard error	p-value	significance
small group (0.6288)	intercept	−1.6443	0.1241	< 0.0001	
	income	0.3388	0.0423	< 0.0001	$*$
	volunteer	1.5591	0.1040	< 0.0001	$*$
	religious belief	−0.2584	0.1070	0.0158	$*$
	education	−0.0434	0.1111	0.6959	
	age	0.4113	0.1206	0.0006	$*$
	gender	0.4466	0.1114	0.0001	$*$
large group (0.3712)	intercept	1.6316	0.0383	< 0.0001	
	income	0.1036	0.0196	< 0.0001	$*$
	volunteer	0.4117	0.0442	< 0.0001	$*$
	religious belief	0.1420	0.0375	0.0001	$*$
	education	−0.0809	0.0410	0.0485	$*$
	age	0.0626	0.0484	0.1960	
	gender	0.0380	0.0365	0.2971	

$$= \sum_{i=1}^{n} \Big[z_{1i} \boldsymbol{\gamma}_0^T \boldsymbol{x_i} - \log\left(1 + e^{\boldsymbol{\gamma}_0^T \boldsymbol{x_i}}\right) + (1 - z_{1i}) z_{2i} \left(y_i \boldsymbol{\beta}_1^T \boldsymbol{x_i} - e^{\boldsymbol{\beta}_1^T \boldsymbol{x_i}} \right) + (1 - z_{1i}) z_{2i} \boldsymbol{\gamma}_1^T \boldsymbol{x_i}$$
$$- (1 - z_{1i}) \log\left(1 + e^{\boldsymbol{\gamma}_1^T \boldsymbol{x_i}}\right) + (1 - z_{1i})(1 - z_{2i}) \left(y_i \boldsymbol{\beta}_2^T \boldsymbol{x_i} - e^{\boldsymbol{\beta}_2^T \boldsymbol{x_i}} \right) - (1 - z_{1i}) log y_i! \Big]. \quad (2.12)$$

식 (2.12)의 완전자료 로그우도 함수는 식 (2.13)과 같이 네개의 성분으로 분리되므로 각 성분을 최대화함으로써 전체 로그우도 함수를 최대화시킬 수 있다.

$$l_c(\boldsymbol{\gamma}_0, \boldsymbol{\gamma}_1, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2 | \boldsymbol{y}, \boldsymbol{z_1}, \boldsymbol{z_2}) = l_c(\boldsymbol{\gamma}_0 | \boldsymbol{y}, \boldsymbol{z_1}, \boldsymbol{z_2}) + l_c(\boldsymbol{\gamma}_1 | \boldsymbol{y}, \boldsymbol{z_1}, \boldsymbol{z_2}) + l_c(\boldsymbol{\beta}_1 | \boldsymbol{y}, \boldsymbol{z_1}, \boldsymbol{z_2}) + l_c(\boldsymbol{\beta}_2 | \boldsymbol{y}, \boldsymbol{z_1}, \boldsymbol{z_2}) \quad (2.13)$$

식 (2.13)에서 EM 알고리즘을 통한 $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$의 추정절차는 Kim 등 (2009)의 식 (3.9)–(3.12)에서 자세히 기술하였고, $\boldsymbol{\gamma}_0$, $\boldsymbol{\gamma}_1$의 EM 알고리즘은 Lambert (1992)와 Morgan (1992, Eq. 2.15, p.71.)의 방법을 적용함으로써 얻어지므로 자세한 절차는 생략하기로 한다. EM 알고리즘의 초기값 설정은 다음과 같이 하였다. 원자료의 히스토그램 상에서 목측으로 고정된 0군, 작은 군, 큰 군을 구분하였고, 고정된 0군과 기부군(작은 군 ∪ 큰 군)을 대상으로 로지스틱 회귀모형을 구성하여 $\hat{\boldsymbol{\gamma}}_0$을 구하였다. 그리고 동 $\hat{\boldsymbol{\gamma}}_0$을 EM 알고리즘의 초기값으로 설정하였다. 작은 군을 대상으로 포아송 회귀모형을 구성하여 $\hat{\boldsymbol{\beta}}_1$을 구하였고, 동 $\hat{\boldsymbol{\beta}}_1$을 EM 알고리즘의 초기값으로 사용하였다. $\boldsymbol{\gamma}_1$과 $\boldsymbol{\beta}_2$의 초기값도 각각 유사한 방법으로 구하였다. EM 알고리즘의 최종 결과치는 작은 군과 큰 군을 구별하는 기준과는 무관하게 일치적인 결과를 보여주었고, 이는 바로 여러개 초기값에 대하여 일치적인 결과를 보여주고 있는 것이다. 모수 추정량의 표준오차는 관찰된 피셔의 정보행렬의 역행렬을 구함으로써 얻었다. R 프로그램의 수행시간은 HP Z220 CMT Workstation(Intel Xeon CPU, 3.20GHz)에서 4.4초 소요되었고, EM 알고리즘의 반복횟수는 허용 한계치 10^{-6}에서 38회였다.

3. 재분석으로 수정될 내용

Kim 등 (2006)의 Table 4.2는 다음 Table 3.1과 같이 바로 잡는다. Table 3.1은 Kim 등 (2006)에 비하여 작은 군의 비율이 0.698에서 0.629로 수정되었으며, 유의적 설명변수로 소득과 자원봉사 이외에

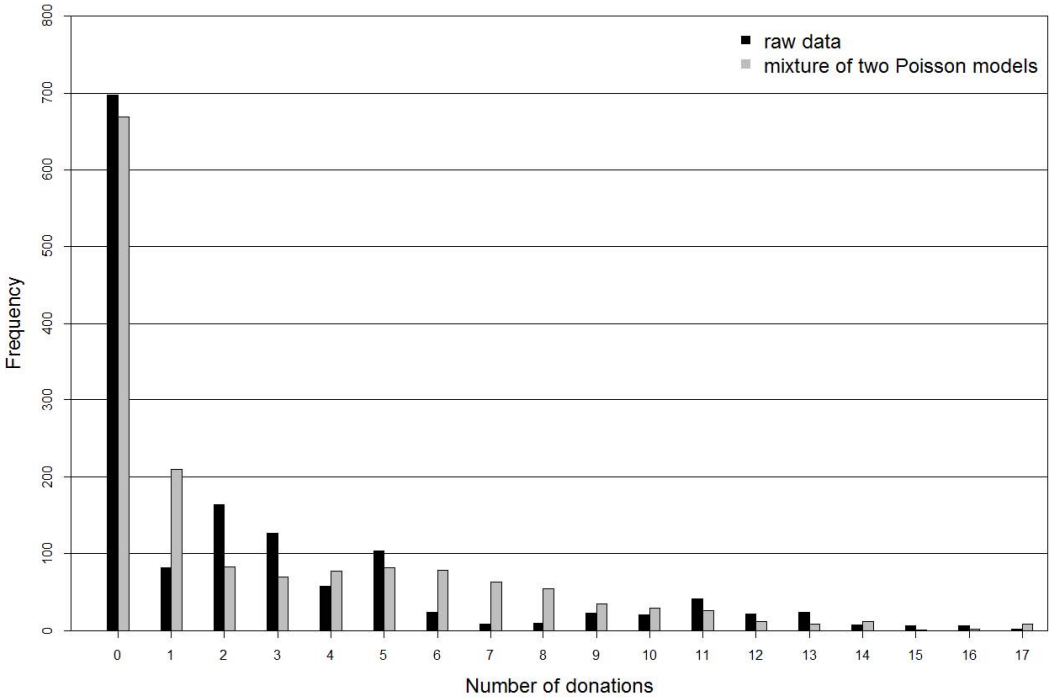

Figure 3.1. Corrected figure of Figure 4.1 of Kim *et al.* (2006).

작은 군에서 종교적 신념에 의한 기부(religious belief), 나이, 성별이 추가되었고, 큰 군에서는 종교적 신념에 의한 기부와 교육이 추가되었다. Kim 등 (2006)의 4.3의 서술적 내용은 그대로 적용이 된다. 그러나 Table 3.1의 작은 군과 큰 군에서 종교적 신념에 의한 기부의 회귀계수 부호가 서로 반대이고, 큰 군에서 교육의 회귀계수 부호가 음인 것은 조금 더 관찰이 필요한 부분일 수 있으며, 동시에 Figure 3.2와 Table 3.3에서 볼 수 있듯이 ZIMP_2 모형에 비하여 적합도가 떨어지는 모형이므로 적합 결여에 의한 현상으로 볼 수도 있다.

Kim 등 (2006)의 Figure 4.1은 다음 Figure 3.1로 바로 잡는다. Kim 등 (2006)의 Figure 4.1과 Figure 3.1은 기부횟수 0과 1에 해당되는 예측치가 상당한 차이를 보이고 있다. 단일 포아송 분포와 두개 포아송 분포의 혼합분포의 AIC와 AICc 값은 Table 4.3에서 찾아 볼 수 있다. AIC와 보정된 AIC인 AICc (Hurvich and Tsai, 1989)는 각각 다음과 같이 정의된다.

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

단, L은 우도함수이고, ν는 모수의 개수를 나타낸다.

Kim 등 (2009)의 Table 4.1과 Figure 4.1은 각각 Table 3.2와 Figure 3.2로 바로 잡는다. Table 3.2는 Kim 등 (2009)의 Table 4.1과 상당한 차이를 나타낸다. 우선 고정 0군과 기부 군 중 작은 군의 비율이 각각 0.201, 0.698에서 0.464와 0.763으로 수정되었다. Kim 등 (2009)에서는 유의적 설명변수로 작은

Table 3.2. Corrected table of Table 4.1 of Kim *et al.* (2009). Estimated regression coefficients of ZIP and a mixture of two Poisson regression models.

group (proportion)	explanatory variable	estimated regression coefficient	standard error	p-value	significance
fixed zero group (0.4636)					
small group (0.7633)	intercept	1.0760	0.0513	< 0.0001	
	income	0.0112	0.0261	0.6684	
	volunteer	0.0839	0.0622	0.1773	
	religious belief	0.0986	0.0504	0.0502	
	education	−0.0721	0.0548	0.1884	
	age	−0.2371	0.0682	0.0005	∗
	gender	−0.0811	0.0506	0.1086	
large group (0.2367)	intercept	2.3062	0.0515	< 0.0001	
	income	0.0371	0.0236	0.1154	
	volunteer	0.0553	0.0524	0.2915	
	religious belief	0.0767	0.0495	0.1215	
	education	0.0163	0.0549	0.7664	
	age	−0.0793	0.0593	0.1810	
	gender	0.0000	0.0477	0.9997	

군, 큰 군 모두에서 소득, 자원봉사가 검색되었으나 Table 3.2에서는 작은 군에서 나이가 유일하게 유의적 설명변수로 검색되었다. 4절에서 보고되는 ZIP과 두개 포아송 회귀-두개 로지스틱 회귀의 혼합모형을 적용하면, 로지스틱 회귀 모형을 통하여 소득, 자원봉사, 종교적 신념이 유의적 설명변수로 검색이 되어 기부행위를 설명할 수 있는 의미있는 결과를 확인할 수 있다. 이 부분을 포함한 모형의 전반적인 해석은 4절에서 다루기로 하겠다.

4. 재분석을 통한 보완 내용

Kim 등 (2009)은 식 (2.4), 식 (2.7)–(2.8)로 구성된 ZIP과 두개 포아송 회귀의 혼합모형을 적합하였으나, 본고에서는 동 모형을 확장하여 식 (2.4)–(2.8)로 구성된 ZIP과 두개 포아송-두개 로지스틱 회귀모형을 적합하였으며 적합 결과는 Table 4.1–4.2와 같다. Table 4.1에서 고정 0군의 추정비율이 $\bar{\hat{\pi}}_0$로 표시된 것은 식 (2.9)에서 보듯이 j번째 개인의 $\hat{\pi}_{0j}$를 산술 평균하였기 때문이다.

본고에서 고려한 3가지 모형, 즉 두개 포아송 혼합모형, ZIP과 두개 포아송 회귀의 혼합모형, 그리고 ZIP과 두개 포아송-두개 로지스틱 회귀의 혼합모형 각각의 예측치와 원자료를 그림으로 나타내면 Figure 4.1과 같다.

로지스틱을 회귀모형을 포함한 재분석의 결과에서 기부에 참여하지 않은 고정된 零群과 기부군 (작은 군, 큰 군 모두 포함)을 구분하는 설명변수로 소득, 자원봉사가 모두 유의적인 변수임을 알 수 있었다. 즉, 소득이 높을 수록, 그리고 자원봉사의 경험이 있는 경우가 그렇지 않은 경우에 비해 기부활동에 참여하는데 긍정적인 영향을 미치고 있는 것으로 나타났다. 또한, 기부군 내에서도 기부를 자주하는 큰 군과 기부를 간헐적으로 하는 작은 군을 구분해주는 설명변수로 자원봉사 경험과 종교적 신념이 유의적으로 나타났다. 이는 자원봉사의 경험이 있는 경우가 그렇지 않은 경우에 비해 큰 군에 속하는 경우가 높

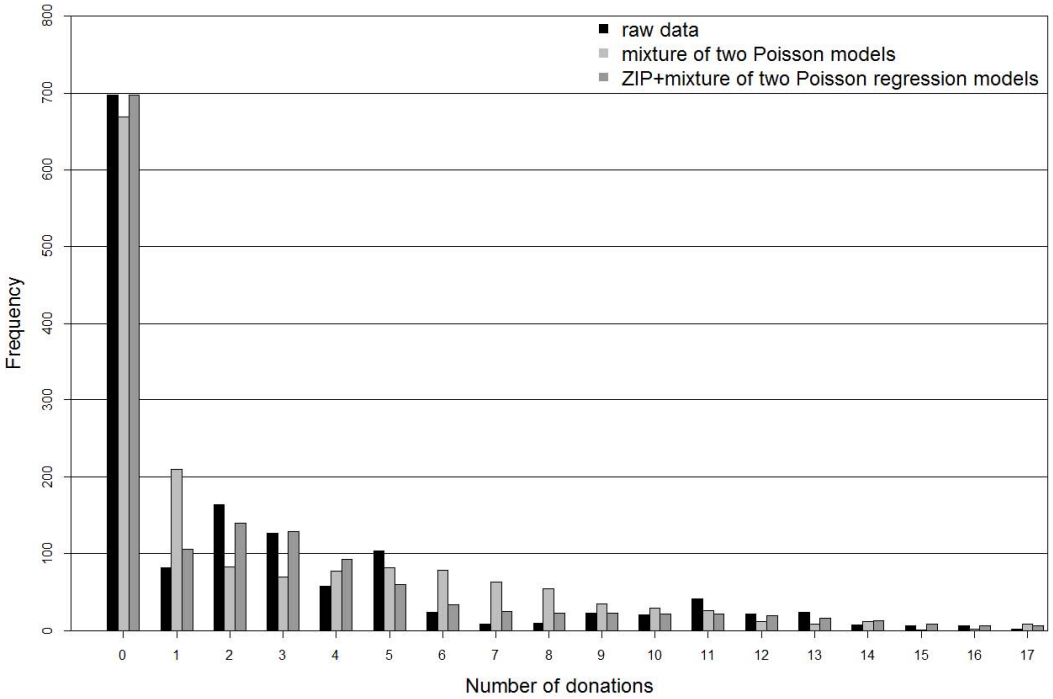

Figure 3.2. Corrected figure of Figure 4.1 of Kim *et al.* (2009).

다는 것을 의미한다. 즉, 자원봉사의 경험이 있는 경우 기부활동의 의미를 잘 인식함으로써 기부활동에 자주 참여하게 된다는 사실을 나타낸다. 또한, 종교적 신념을 가지고 있는 경우가 그렇지 않은 경우에 비해 기부활동에 자주 참여하는데 기여하고 있다는 것을 의미한다. 자원봉사의 경험과 종교가 기부군 내에서도 기부활동에 자주 참여하는지의 여부를 결정하는 설명변수로 확인되었다. 이는 기부행위에 대한 경제학적 분석의 의미로 볼 때 기부행위 여부와 기부횟수가 큰 군과 작은 군을 구분해주는 중요한 변수로서 자원봉사 경험과 종교적 신념의 의미를 재확인할 수 있는 결과이다 (Park과 Park, 2004; Son과 Park, 2008).

Kim 등 (2009)의 Table 4.2를 수정하고, ZIP과 두개 포아송-두개 로지스틱 회귀의 혼합모형을 추가하여 AIC, AICc를 계산하면 Table 4.3과 같다.

5. 결론과 토의

2002년 기부횟수 자료에서 고정된 0군의 비율은 0.465, 전체 모집단에서 작은 군[2]의 비율은 0.413, 그리고 큰 군의 비율은 0.122로 추정되었으며, 작은 군과 큰 군은 각각 연평균 2.8회, 10.6회를 기부하는 것으로 추정되었다.

Table 4.1은 두개 포아송 회귀 혼합모형의 회귀계수 추정 결과를 나타내는데 유의수준 0.05에서 나이가

[2] 이제까지 논의한 작은 군의 비율은 기부군(Z^c)에서 상대적 비율로 정의하였다. 전체 모집단에서 작은 군의 비율은 $\Pr(Y \in Z^c)\Pr(\text{작은 군}|Z^c) = (1-0.465)(0.773) = 0.413$이 된다.

Table 4.1. Estimates of regression coefficients of two Poisson regressions in the ZIP and a mixture of two Poisson-two logistic regression model.

group (proportion)	explanatory variable	estimated regression coefficient	standard error	p-value	significance
		$\hat{\beta}_1$			
fixed zero group ($\bar{\hat{\pi}}_0$=0.4647) ($\bar{\hat{\lambda}}_1$=2.8300)	intercept	1.1293	0.0505	< 0.0001	
	income	−0.0080	0.0261	0.7592	
	volunteer	0.0113	0.0634	0.8585	
	religious belief	0.0431	0.0500	0.3886	
	education	−0.0799	0.0546	0.1433	
	age	−0.2563	0.0686	0.0002	∗
	gender	−0.0737	0.0504	0.1434	
		$\hat{\beta}_2$			
small group ($\bar{\hat{\pi}}_1$=0.7725) ($\bar{\hat{\lambda}}_2$=10.630)	intercept	2.3531	0.0528	< 0.0001	
	income	0.0310	0.0237	0.1909	
	volunteer	0.0259	0.0519	0.6172	
	religious belief	0.0225	0.0503	0.6538	
	education	0.0201	0.0553	0.7160	
	age	−0.1030	0.0592	0.0819	
	gender	0.0143	0.0478	0.7656	

Table 4.2. Estimates of regression coefficients of two logistic regressions in the ZIP and a mixture of two Poisson-two logistic regression model.

group (proportion)	explanatory variable	estimated regression coefficient	standard error	p-value	significance
		$\hat{\gamma}_0$			
fixed zero group ($\bar{\hat{\pi}}_0$=0.4647)	intercept	0.1668	0.1099	0.1290	
	income	−0.1784	0.0585	0.0023	∗
	volunteer	−0.8246	0.1591	< 0.0001	∗
	religious belief	−0.1757	0.1096	0.1090	
	education	−0.0125	0.1224	0.9188	
	age	−0.1177	0.1399	0.4000	
	gender	−0.1295	0.1108	0.2425	
		$\hat{\gamma}_1$			
small group ($\bar{\hat{\pi}}_1$=0.7725)	intercept	1.5117	0.1921	< 0.0001	
	income	−0.0541	0.0902	0.5484	
	volunteer	−0.4696	0.2011	0.0195	∗
	religious belief	−0.5219	0.1825	0.0042	∗
	education	0.1477	0.1997	0.4595	
	age	−0.1960	0.2168	0.3659	
	gender	0.1759	0.1782	0.3233	

유일하게 작은 군에서 유의적인 설명변수로 검색되고 있다. 또한 유의수준 0.1에서 큰 군과 작은 군 모

Table 4.3. AIC and AICc of four possible Poisson based models. The first three columns correct the first three columns of Table 4.2 of Kim *et al.* (2009).

	Single Poisson Regression	Mixture of two Poisson regressions	ZIP+mixture of two Poisson regressions	ZIP+mixture of two Poisson-two logistic regressions
AIC	8777.0	5944.7	5443.5	5409.8
AICc	8777.1	5945.0	5443.8	5411.0

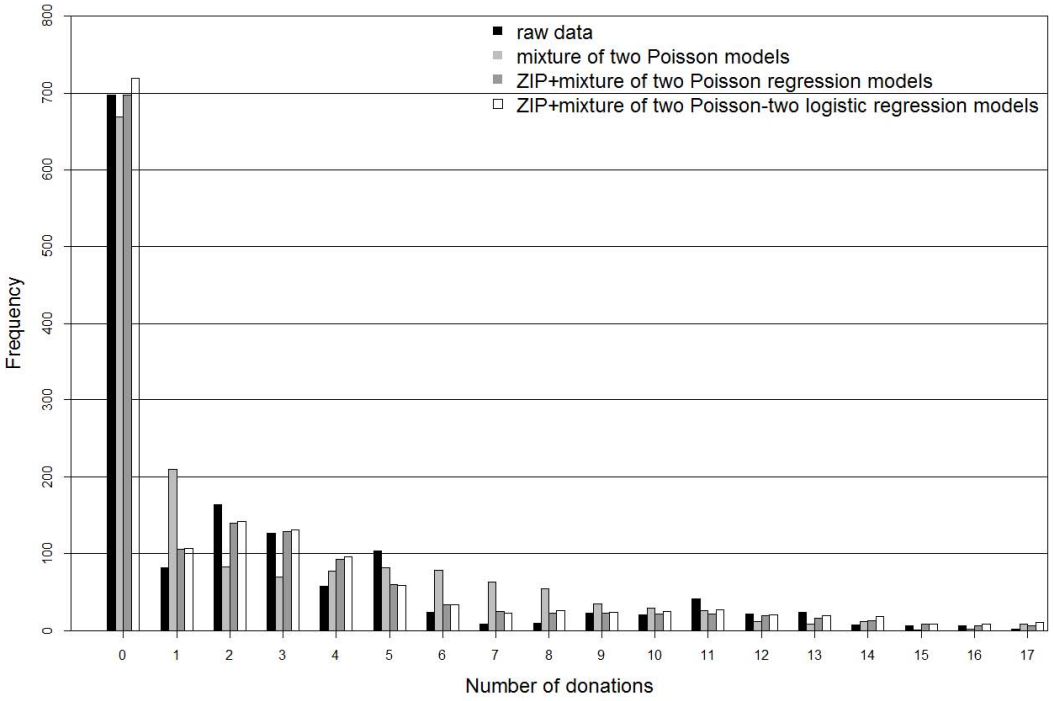

Figure 4.1. Raw data and predicted values under three different models.

두에서 나이가 비교적 어릴 수록(50세 미만) 기부횟수가 더 많은 것으로 나타나고 있다.[3] Table 4.2의 결과는 고정 0군과 기부군을 구별지어 주는 특성과 작은 군과 큰 군을 구별하는 특성을 잘 나타내고 있다. Table 4.2에서 고정 0군의 자원봉사 회귀계수의 절댓값이 작은 군의 자원봉사 회귀계수의 절댓값보다 1.8배 큰 것은 자원봉사가 기부행위에 결정적 역할을 하고 있음을 말해주고 있다. 고정 0군에서 종교적 신념에 의한 기부는 p값이 0.11로서 한계적 유의성을 띄고 있는데 작은 군의 동 변수의 p값과 회귀계수를 비교하여 볼때 종교적 신념에 의한 기부는 작은 군과 큰 군을 구별지어 주는 변수이지 기부행위 자체에 영향을 미치는 변수는 아닌 것으로 판명된다. 비슷한 논리로 소득은 기부행위에는 영향을 미치지만 작은 군과 큰 군을 구별짓는 유의적 설명변수는 되지 않음을 알 수 있다.

[3]최근 들어 소액기부를 정기적으로 실시하는 젊은 층이 늘어나는 추세 (한겨레신문, 2011.1.30, http://www.hani.co.kr/arti/society/society_general/461351.html)이고, 동 추세와 평행한 결과가 나타났으므로 p값을 다소 높히는 것이 정당화될 수 있다고 판단한다.

References

Dempster, A. P., Larid, 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.

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

Kim, I., Park, S. B., Kim, B. S. and Park, T. K. (2006). The analysis of the number of donations based on a mixture of Poisson regression model, *The Korean Journal of Applied Statistics*, **19**, 1–12.

Kim, I., Park, T. K. and Kim, B. S. (2009). The reanalysis of the donation data using the zero-inflated Poisson regression, *The Korean Journal of Applied Statistics*, **22**, 819–827.

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

McLachlan, G. J. and Krishnan, T. (1997). *The EM Algorithm and Extensions*, New York: Wiley.

Morgan, B. J. T. (1992). *Analysis of Quantal Response Data*, New York: Chapman & Hall.

Park, T. K. and Park, S. B. (2004). An economic study on charitable giving of individuals in Korea: Some new findings from 2002 survey data, presented at 6^{th} conference of ISTR, Toronto, Canada.

Son, W. and Park, T. K. (2008). *A Study on Private Donation in Korea*, Seoul: Korea Institute of Public Finance.

2002년 기부횟수 자료의 재분석: 수정 및 보완

김병수[a] · 이주형[a] · 김인영[b] · 박수범[c] · 박태규[d,1]

[a]연세대학교 응용통계학과, [b]Department of Statistics, Virginia Tech,
[c]한국여성정책연구원 성인지 예산센터, [d]연세대학교 경제학과

(2014년 7월 29일 접수, 2014년 9월 26일 수정, 2014년 9월 30일 채택)

요 약

Kim 등 (2006)과 Kim 등 (2009)은 2002년에 (사)볼런티어 21에서 조사한 설문자료에 기초하여 우리나라 개인의 기부횟수에 영향을 주는 유의적 설명변수를 보고한 바 있다. 본고에서는 Kim 등 (2006)과 Kim 등 (2009)의 계산 오류를 발견하여 이를 수정하고, 아울러 Kim 등 (2009)이 적용한 0이 팽창된 포아송 모형에 로지스틱 회귀모형을 추가하였다. 동 로지스틱 모형으로 기부행위(0, 1)에 영향을 주는 설명변수를 식별하고, 아울러 기부횟수가 작은 群과 큰 群을 판별하여 주는 설명변수를 식별하고자 한다.

주요용어: 0이 팽창된 포아송(ZIP), 기부, 두개 포아송의 혼합, 로지스틱 회귀, 포아송 회귀, EM 알고리즘.

[1]교신저자: (120-749) 서울 서대문구 연세로 50, 연세대학교 경제학과. E-mail: tkpark@yonsei.ac.kr