Journal of the Korean Data & Information Science Society
2016, **27**(3), 587–598

http://dx.doi.org/10.7465/jkdi.2016.27.3.587
한국데이터정보과학회지

Monte-Carlo expectation-maximaization 방법을 이용한 무응답 모형 추정방법[†‡]

최보승[1] · 유현상[2] · 윤용화[3]

[1]고려대학교 응용통계학과 · [2]대구은행 마케팅부 · [3]대구대학교 전산통계학과

접수 2016년 3월 15일, 수정 2016년 4월 19일, 게재확정 2016년 4월 20일

요 약

각종 선거를 앞두고 여러 여론조사 기관들은 다양한 방법으로 선거 결과를 예측한다. 조사를 통한 선거 예측을 수행하는 데 있어서 발생할 수 있는 문제점 중 하나는 무응답이며 무응답 대체 방법에 따라 예측 결과는 완전히 다른 결과를 생산해 낼 수 있다. 본 연구에서는 무응답 대체의 방법으로 모형을 기반으로 한 대체 방법에 대하여 연구하였다. 특히, 최대 우도 추정 방법을 적용했을 때 무시할 수 없는 무응답 (non-ignorable non-response) 체계 하에서 발생할 수 있는 변방 값 문제를 해결하기 위해 Wei와 Tanner (1990)가 제안한 Monte Carlo EM 알고리즘을 적용하였다. 모의 실험을 통하여 MCEM 방법과 기존의 최대 우도 추정 방법, 베이지안 추정 방법 사이의 비교 연구를 진행하였고 그 결과 MCEM 방법이 기존 방법들에 대한 대안 방법으로 이용될 수 있음을 보였다. 또한 2012년에 시행된 제18대 대통령 선거 당일의 출구조사 자료를 적용하여 실증 분석을 수행하였다. 예측 결과를 비교하기 위해 Bautista 등 (2007)이 제안한 MWPE (modified within precinct error)를 이용하였다.

주요용어: 결측자료, 몬테카를로 EM, 무시할 수 없는 무응답, 무응답 대체, 일반화 선형 모형.

1. 서론

설문조사로 수집된 자료로 분석할 경우 가장 먼저 부딪치는 문제는 결측 값 또는 무응답 문제라고 할 수 있다. 각종 여론조사, 설문조사를 할 경우 무응답이 발생하지 않는 경우는 매우 드물기 때문에 본 논문은 예측의 정확도를 높이는 중요한 관건이라고 할 수 있는 무응답처리 문제를 다루고자 한다.

Hong과 Huh (2001)는 높은 무응답률은 무응답자의 지지후보를 무리하게 추측하고 조사회사마다 무응답자 판정방법이나 무응답자 포함 여부가 다르므로 예측결과가 제각각이게 되고, 최종 예상 득표율 산출에 상당한 영향을 미쳤을 것으로 보인다고 하였다. Kim과 Kwak (2010), Kwak 등 (2013)은 투표자 선정 및 응답 과정에서 오차를 일으키는 변수들 중에 무응답의 비율이 오차와의 상관관계가 높아 오차를 일으키는 중요한 원인이라고 하였다. 무응답 처리를 포함한 예측 문제에 대하여 국내외에 많은 연구가 진행되어 왔다. Crespi (1988)의 연구에서는 무응답 대체 방법으로 주요 2개 후보에게 비례 배분하는 방법, 주요 2개 후보에게 반으로 나누어 배분하는 방법, 현직 후보자가 있다면 그 외의 도전자 후보에

† 이 논문은 대구대학교 교내 연구비로 지원받아 수행된 연구임 (No.20130310).
‡ 이 논문은 제2저자 유현상의 석사학위 논문을 바탕으로 추가 연구하여 작성된 것임.
[1] (30019) 세종특별자치시 세종로 2511, 고려대학교 세종캠퍼스 응용통계학과, 조교수.
[2] (42123) 대구광역시 수성구 달구벌대로 2310, 대구은행 마케팅부, 계장.
[3] 교신저자: (38453) 경상북도 경산시 진량읍 대구대로 201, 대구대학교 전산통계학과, 교수.
E-mail: yhyoon@daegu.ac.kr

게 배분하는 방법, 무응답을 버리고 후보자들의 득표율을 재계산하는 방법 등 4가지 방법을 제시하였고, 그 가운데 비례배분이 가장 좋다는 의견을 제시하였다. Cho 등 (2008)과 Lee와 Kang (2012)은 설문조사에서 발생하는 무응답에 대한 종류와 대체 방법을 소개하였다. Lee 등 (2006)은 2002년 강원지역의 농가경제 자료를 예제로 하여 응답자들과 무응답자 사이에 지역적 상관관계를 이용한 대체방법을 보여주었다. Cho 등 (2008)은 농촌 생활지표조사에서의 무응답을 연속형과 범주형으로 나누어 무응답 대체를 비교하였다. 이러한 방법들은 표본조사의 정보를 이용하여 무응답 대체를 수행하는 방법이다. 이와 다른 대체방법으로 모형적 접근에 의한 무응답 대체방법이 연구되고 있다. Park과 Brown (1994), Park과 Lee (1998), Yoon과 Choi (2014)는 무응답모형을 추정하기 위한 방법으로 최대 우도 추정 방법을 사용할 때 나타날 수 있는 단점을 해결하기 위하여 Dirichlet 사전확률 분포를 할당하는 경험적 베이지안 방법을 제안하였다. Choi 등 (2007)은 사전분포를 할당하는 데 있어서 EM 알고리즘에서 발생하는 기대 빈도를 사전분포의 초모수로 할당하는 방법을 제안하고 1948년 미국에서 시행된 대통령 선거의 사전조사 결과를 이용하여 기존방법과 Dirichlet 사전확률 분포에 5가지 유형의 사전분포 초모수 할당을 통한 베이지안 방법을 제시하여 비교하였다. Park과 Choi (2010)는 사전분포의 초모수를 할당하는 데 있어서 관찰빈도와 칸 기대빈도에 대한 최대 우도 추정치를 동시에 이용하는 방법을 제시하였고 다양한 상황에서 모의실험을 통하여 제안한 방법에 대한 성능을 검증하였다. Choi와 Kim (2012)은 여러 무응답모형 가운데 적절한 무응답 모형을 선택하기 위한 방법을 제안하였는데 Ibrahim 등 (2008)이 제안한 EM 알고리즘을 이용한 모형 추정에서의 모형 선택 방법을 이용하였다. Yoon과 Choi (2012)은 2004년 국회의원 선거를 앞두고 실시한 여론조사 자료를 가지고 무응답을 가지고 있는 다차원 분할표의 범주형 자료에 대하여 무응답의 대체와 모수의 추정을 함께 수행하는 계층적 베이지안 방법을 제안하였고 조건부 사후분포로부터 모수를 추출하기 위한 MCMC 방법을 제시하였다.

이러한 모형적 접근에 의한 무응답 대체방법에는 무응답 모형을 적용하기 위해서 적절한 무응답 체계에 대한 가정이 필요하다. Little과 Rubin (2002)은 무응답을 발생 체계에 따라 크게 세 가지로 구분하였다. 첫 번째는 완전임의 결측 (missing completely at random; MCAR)으로 무응답의 발생 여부가 무응답을 가지고 있는 변수나 함께 조사된 다른 변수들에 아무 영향을 받지 않았을 경우이다. 두 번째는 임의 결측 (missing at random; MAR)으로 무응답의 발생 여부가 조사된 변수들 중에서 무응답을 가지고 있지 않은 관찰된 자료에 의해서만 영향을 받았을 경우이다. 이 두 가지의 가정은 무응답의 발생 여부가 무응답 자체에 영향을 받는 것이 아니므로 무시할 수 있는 무응답 (ignorable nonresponse)이라 한다. 세 번째는 비임의 결측 (not missing at random; NMAR)으로 무응답 발생 여부가 관찰된 자료 중에서 무응답을 가지고 있는 변수에서만 영향을 받았을 경우이다. 비임의 결측은 무응답의 발생 여부가 무응답 자체의 영향을 받으므로 무시할 수 없는 무응답 (non-ignorable non-response)이라 한다. 예를 들어 출구조사 시 자신의 지지하는 후보를 밝히지 않았을 때 특별한 이유가 없다면 무시할 수 있는 무응답이라고 할 수 있고 자신의 지지 후보가 그 지역의 열세 후보이기 때문에 밝히지 않았다면 무시할 수 없는 무응답이라고 할 수 있다. Shim과 Choi (1997)에서는 여론에 대한 무응답 현상을 사회문화적 특성과 연결시켜 인구 사회학적 특성에 의해 정보수준이나 인지 수준이 낮아 여론조사에서 “잘 모르겠다.” 고 무응답을 하는 경우는 미국과 같이 상대적으로 개방적이고 수평적인 사회에서 많이 나타나고, 정치상황적 요인에 의해 관여하고 싶지 않아 “잘 모르겠다.” 고 응답하는 경우는 한국과 같이 비교적 폐쇄적이고 수직적인 사회에서 많이 나타난다고 하였다. 무응답 체계에 대한 가정은 정확한 무응답 대체와 예측결과를 얻기 위해 매우 중요한 절차라 할 수 있다. 그러나 현재까지 무응답 모형의 대체방법에 대한 연구는 아직까지 확실한 답이 없으며 (Park과 Lee, 1998; Choi 등, 2007; Park과 Choi, 2010; Choi와 Kim, 2012; Yoon과 Choi, 2012; Yoon과 Choi, 2014), Ibrahim 등 (2008)에서는 모형적 선택의 방법을 이용하여 무응답 가정의 임의 결측과 비임의 결측을 직접 비교하는 것은 위험할 수 있다고 경고 하였다.

본 연구에서는 적절한 무응답 체계에 대한 가정에 따른 무응답 모형을 설정하고 이 무응답 모형을 이

용한 무응답 대체 방법에 대한 연구를 진행하였다. 무응답 모형을 통해 무응답을 대체 하는 방법으로 EM 알고리즘에 기반을 둔 최대 우도 추정방법을 이용하였다. 최대 우도 추정 방법에서 발생할 수 있는 문제를 Wei 와 Tanner (1990)가 제안한 방법을 적용하여 로그 선형모형의 모수에 직접 사전 분포를 할당하는 계층적 베이지안 방법으로 MCEM (Monte Carlo Expectation and Maximization)방법을 이용하였다. MCEM 알고리즘은 EM 알고리즘의 E-단계의 기댓값 계산과정을 Monte Carlo 방법으로 해결하여 무응답 자료를 구하고 이로부터 M-단계를 통하여 모수를 추정하는 방법이다. 본 연구에서는 모의실험을 통해 MCEM 알고리즘에 대해 알아본 후 실제 자료에 적용하였다.

본 연구의 구성은 다음과 같다. 2절에서는 무응답 모형의 추정방법을 소개한다. 결측 체계에 대한 가정에 따라 무응답 모형을 설정하고 EM 알고리즘과 MCEM 알고리즘을 이용한 무응답 추정 방법을 소개한다. 3절에서는 본 연구에서 소개하고 있는 MCEM 방법의 효용을 확인해 보기 위하여 진행한 모의실험 결과를 설명한다. 4절에서는 2012년에 실시된 대통령 선거의 출구 조사 결과를 이용한 실증 자료 분석을 설명하고 마지막 5절에서는 결론으로 본 연구 결과와 한계점을 정리하였다.

2. 무응답 모형 추정 방법

2차원 분할표 형태로 정리된 자료에 대하여 무응답을 가지고 있는 경우 주변 합을 가진 분할표 형태로 정의될 수 있다 (Kwak과 Choi, 2014). Park과 Choi (2010)이 사용한 로테이션을 이용하여 2차원 분할표에서 분할표를 구성하는 행, 열 변수 모두에 무응답이 발생하는 무응답 모형을 고려하여 보자. 행 변수를 X_1로, 열 변수를 X_2로 놓는다. 행과 열 변수는 각각 2개의 범주를 가지고 있다. 각 변수에 대한 무응답 지시변수를 R_1, R_2라 하고 무응답이 발생하지 않은 경우 1의 값을 가지며 무응답을 가지고 있는 경우 2의 값을 가진다. 분할표의 각 칸의 빈도는 y_{ijkl}로 나타낸다. 첨자 i, j는 행 변수와 열 변수의 수준을 나타내며 k, l은 무응답 지시변수의 수준을 나타낸다. 이제 무응답을 가지는 범주형 자료는 다음과 같이 주변합을 가지는 분할표 형태로 정리된다.

Table 2.1 Two-way contingency table with supplemental margins

		$R_2 = 1$		$R_2 = 2$
		$X_2 = 1$	$X_2 = 2$	
$R_1 = 1$	$X_1 = 1$	y_{1111}	y_{1211}	y_{1+12}
	$X_1 = 2$	y_{2111}	y_{2211}	y_{2+12}
$R_1 = 2$		y_{+121}	y_{+221}	y_{++22}

이제 관찰빈도 y_{ijkl}에 대하여 다항분포를 확률성분으로 하는 일반화 선형 모형으로부터 로그 우도 함수는 다음과 같이 정의된다.

$$\begin{aligned} \ell \quad \propto \quad & \sum_i \sum_j y_{ij11} \cdot \log(\pi_{ij11}) + \sum_i y_{i+12} \cdot \log(\pi_{i+12}) \\ & + \sum_j y_{+j21} \cdot \log(\pi_{+j21}) + y_{++22} \cdot \log(\pi_{++22}), \end{aligned} \tag{2.1}$$

여기서 $\pi_{ijkl} = Pr[X_1 = i, X_2 = j, R_1 = k, R_2 = l]$는 각 칸의 기대 확률이고 $N = \sum_{i,j,k,l} y_{ijkl}$으로 고정되어 있다고 가정한다. 행과 열의 범주가 각 두 개만 있음을 가정하였기 때문에 i, j 또한 1과 2의 값만을 가진다. 이제 일반화 선형 모형 가운데 하나인 로그 선형모형의 체계적 성분을 이용하여 무응답모형을 구축할 수 있다. Little과 Rubin (2002)의 무응답 체계의 가정 가운데 무응답 여부가 무응답이 발생하는 변수에 의존하는 무시할 수 없는 무응답 모형을 고려하여 보자. 이 모형은 또한 비 임의 결측 모

형이 된다.

$$\log(m_{ijkl}) = \beta_0 + \beta^i_{X_1} + \beta^j_{X_2} + \beta^k_{R_1} + \beta^l_{R_2} + \beta^{ik}_{X_1R_1} + \beta^{jl}_{X_2R_2} + \beta^{ij}_{X_1X_2} + \beta^{kl}_{R_1R_2} \tag{2.2}$$

이 무응답 모형 (2.2)를 모든 자료에 적용하여 행렬 모형으로 다음과 같이 표현 할 수 있다.

$$\log \mathbf{m} = \mathbf{Z}\boldsymbol{\beta}, \tag{2.3}$$

여기서 $\mathbf{m}$은 칸 기대빈도 벡터이고, $\boldsymbol{\beta}$는 모수벡터이다. 그리고 $\mathbf{Z}$는 계획행렬을 나타낸다. 무응답 혹은 결측치를 가지고 있는 범주형 자료에 대하여 로그 선형모형을 이용한 무응답 모형은 EM 알고리즘을 이용하여 무응답 대체를 포함한 모수의 추정을 수행할 수 있다 (Baker와 Laird, 1988). 그러나 무시할 수 없는 무응답 가정하에서 EM 알고리즘을 이용하여 최대 우도 추정을 수행하는 경우 변방값 문제가 발생할 수 있다. 변방값 문제란 다차원 분할표에서 무응답 빈도에 대한 추정치가 0 값을 가지게 되는 현상을 말한다. 이와 같은 변방값 문제가 발생하게 되면 모수의 추정 과정에서 유일하지 않은 해를 가질 수 있다 (Park과 Brown, 1994). 이를 해결하기 위한 방법으로 기대빈도에 사전분포를 할당하거나 (Choi 등, 2009; Park과 Choi, 2010) 일반화 선형 모형에서 체계적 성분의 모수에 사전분포를 할당하는 (Park 등, 2013) 베이지안 방법들이 제안되어 왔다. Park과 Choi (2010)에서 제안된 바와 같이 칸 기대빈도 또는 칸 기대확률에 사전분포를 할당하여 EM 알고리즘을 이용하는 절차는 다음과 같다.

관찰된 자료에 대하여 다항분포를 가정하였기 때문에 칸 기대확률 ($\pi_{ij11},\ \pi_{ij12},\ \pi_{ij21},\ \pi_{ij22}$)에 대하여 다음과 같이 Dirichelet 분포를 사전분포로 할당한다.

$$\prod_i \prod_j \pi_{ij11}^{\delta_{ij11}} \cdot \pi_{ij12}^{\delta_{ij12}} \cdot \pi_{ij21}^{\delta_{ij21}} \cdot \pi_{ij22}^{\delta_{ij22}}, . \tag{2.4}$$

할당된 사전분포 (2.4)과 로그 우도 함수 (2.1)을 가지고 다음과 같은 로그 사후 분포함수를 구할 수 있다.

$$\begin{aligned}
\ell_{pos} &= \sum_i \sum_j y_{ij11} \cdot (\mathbf{z}_{ij11} \cdot \boldsymbol{\beta}) - \sum_i \sum_j y_{ij11} \cdot \log\Big(\sum_{i,j,k,l} \exp(\mathbf{z}_{ijkl} \cdot \boldsymbol{\beta}) \Big) \\
&+ \sum_i y_{i+12} \cdot \log\Big(\sum_j \exp(\mathbf{z}_{ij12} \cdot \boldsymbol{\beta}) \Big) - \sum_i y_{i+12} \cdot \log\Big(\sum_{i,j,k,l} \exp(\mathbf{z}_{ijkl} \cdot \boldsymbol{\beta}) \Big) \\
&+ \sum_j y_{+j21} \cdot \log\Big(\sum_i \exp(\mathbf{z}_{ij21} \cdot \boldsymbol{\beta}) \Big) - \sum_j y_{+j21} \cdot \log\Big(\sum_{i,j,k,l} \exp(\mathbf{z}_{ijkl} \cdot \boldsymbol{\beta}) \Big) \\
&+ y_{++22} \cdot \log\Big(\sum_i \sum_j \exp(\mathbf{z}_{ij22} \cdot \boldsymbol{\beta}) \Big) - y_{++22} \cdot \log\Big(\sum_{i,j,k,l} \exp(\mathbf{z}_{ijkl} \cdot \boldsymbol{\beta}) \Big) \\
&+ \sum_{i,j,k,l} \delta_{ijkl} \cdot (\mathbf{z}_{ijkl} \cdot \boldsymbol{\beta}) - \sum_{i,j,k,l} \delta_{ijkl} \cdot \log\Big(\sum_{i,j,k,l} \exp(\mathbf{z}_{ijkl} \cdot \boldsymbol{\beta}) \Big).
\end{aligned} \tag{2.5}$$

이제 이 로그 사후 분포함수에 EM 알고리즘을 적용하여 무응답 혹은 결측에 대한 대체와 모수에 대한 추정을 수행할 수 있다. 먼저 E-step에서는 무응답 값에 대한 기대값을 계산한다. E-step의 수행을 통하여 무응답 빈도에 대한 의사 관찰 빈도 y_{ij12}, y_{ij21}, y_{ij22}를 구하면 확장된 로그 사후 분포함수 (augmented log posterior function)는 다음과 같은 식을 가지게 된다. 이 식은 온전한 형태 (무응답이 발생하지 않은) 4차원 분할표에 대한 로그 선형모형에 비례하게 된다.

$$\begin{aligned}
\ell_{a.pos} = \quad & \textstyle\sum_i \sum_j (y_{ij11} + \delta_{ij11}) \log(\pi_{ij11}) + \sum_i \sum_j (y_{ij12} + \delta_{ij12}) \log(\pi_{ij12}) \\
& + \textstyle\sum_i \sum_j (y_{ij21} + \delta_{ij21}) \log(\pi_{ij21}) + \sum_i \sum_j (y_{ij22} + \delta_{ij22}) \log(\pi_{ij22}).
\end{aligned} \tag{2.6}$$

E-step에서는 식 (2.6)에서 의사 관찰 빈도에 대한 기대값 $E(y_{ij12}|\pi_{ijkl}, y_{i+12})$, $E(y_{ij21}|\pi_{ijkl}, y_{+j21})$, $E(y_{ij22}|\pi_{ijkl}, y_{++22})$을 계산하고 M-step에서는 관찰빈도와 E-step에서 계산된 의사 관찰빈도를 가지고 모수의 대한 최대 사후 추정치를 구하게 된다. 이와 같은 E-step과 M-step은 모수에 대한 사후 추정치가 수렴할 때까지 반복하면 된다. 만약 베이지안 방법을 이용하지 않고 EM 알고리즘을 이용하여 최대 우도 추정치를 구하고자 한다면 식 (2.6)에서 모든 δ_{ijkl}의 값에 0을 넣으면 된다.

완전한 베이지안 방법을 구축하기 위해서는 이와 같은 절차에 사전분포의 초모수인 δ_{ijkl}을 할당 방법을 추가적으로 고려하여야 한다. Park과 Brown (1994), Park (1998)은 사전분포의 초모수가 관찰된 빈도에 의존하는 방법을 제안하였고 Choi 등 (2009)와 Park과 Choi (2010)은 관찰빈도과 무응답 빈도를 모두 이용하여 초모수를 할당하는 방법을 제안하였다.

이와 같이 EM 알고리즘을 이용하는 방법은 로그 사후 분포의 기댓값이 쉽게 계산된다는 가정에서 유도되었지만, 실제로 기댓값의 계산이 어려운 경우, 즉 적분하기 어려운 경우가 발생할 수 있다. 또한 전술한 바와 같이 무시할 수 없는 무응답 가정하에서는 변방값 문제가 발생할 수 있다. 본 연구에서는 이와 같은 문제를 해결하기 위하여 Wei와 Tanner (1990)에 의해서 제안된 MCEM 알고리즘 (Monte Carlo EM algorithm)을 이용하고자 하였다. 이 알고리즘은 EM 알고리즘의 E-Step에서 Monte Calro 방법으로 기댓값을 계산하여 불완전한 자료로부터 모수를 추정할 수 있도록 하는 방법이다.

Wei와 Tanner (1990)에 의해 제안된 MCEM 알고리즘은 다음과 같은 세 단계의 절차로 이루어진다. 먼저 서술의 편의를 위하여 관찰빈도 $Y_{obs} = (\{y_{ij11}\})$와 관찰되지 않은 빈도 $Y_{mis} = (\{y_{ij12}\}, \{y_{ij21}\}, \{y_{ij22}\})$로 구분한다.

단계1: t번째 반복에서 최근 계산된 모수의 최대 추정치 $\boldsymbol{\beta}^t$가 주어졌을 때, 결측자료에 대한 조건부 확률 분포 $p(Y_{mis}|\boldsymbol{\beta}^t, Y_{obs})$로부터 표본 $Y_{mis}^{(1)}, \cdots, Y_{mis}^{(m)}$를 추출한다.

단계2: EM 알고리즘의 E step에서 수행하는 조건부 우도 함수의 기대값을 계산하는 대신에 수치 적분을 수행하기 위한 Monte Carlo 방법을 사용하여 다음과 같이 정의한다.

$$\hat{Q}(\boldsymbol{\beta}|\boldsymbol{\beta}^t) = \frac{1}{m}\sum_{t=1}^{m} \log(p(\boldsymbol{\beta}|Y_{mis}^t, Y_{obs}))$$

참고로 m이 1인 경우에 얻어지는 알고리즘을 "EM-형태 (EM type)"의 알고리즘이라 할 수 있다.

단계3: 마지막 M-Step으로 위의 Q-function을 최대로 만드는 $\boldsymbol{\beta}$에 의해 $\boldsymbol{\beta}^t$는 $\boldsymbol{\beta}^{t+1}$로 갱신된다. 갱신된 모수의 추정치로 결측치에 대한 조건부 확률 분포를 개선한다. 이 후의 단계는 EM 알고리즘과 같은 방법으로 Q-function이 수렴할 때까지 Monte Carlo 방법을 적용한 E-Step과 M-Step을 반복하여 Q-function을 최대화 시키는 최대 우도 추정치 $\hat{\boldsymbol{\beta}}$을 추정한다.

3. 모의실험

본 모의실험은 무시할 수 없는 무응답 모형에 대해 무응답 비율, 응답패턴, 초모수의 할당방법 등 총 32가지의 시나리오를 가지고 모의실험을 하였던 Park과 Choi (2010)의 모의실험 결과를 기초 자료로 하여 본 연구에서 적용한 EM 알고리즘 기법인 Monte Carlo EM 알고리즘과 Yoo (2015)의 결과를 함께 비교 하였다. 가능한 모든 무응답 모형을 고려하지 않고 본 모의 실험에서는 비임의결측 체계를 가정한 무시할 수 없는 무응답 모형에 대해서만 모의실험을 진행하였다. 따라서 모의실험에서 고려한 무응답 모형은 2절에서 제시된 식 (2.2)을 그대로 이용하였다.

모의실험에 앞서 모형 식 (2.2)의 모든 X_1, X_2, R_1, R_2의 범주의 수는 2이다. 따라서 8개의 모수를 본 모형에서 정의하여야 한다. 무응답 모형에서의 모수 가운데 비임의결측의 정도를 나타내는 모수는

$\beta^{ik}_{X_1R_1}$, $\beta^{jl}_{X_2R_2}$이다. 이 모수들은 로그 선형 모형의 확률성분의 기대빈도와 다음과 같은 관계를 가진다.

$$\begin{aligned} 4\beta^{11}_{X_1R_1} &= \log(m_{1111}/m_{2111})(m_{1121}/m_{2121}), \\ 4\beta^{11}_{X_2R_2} &= \log(m_{1111}/m_{1211})(m_{1112}/m_{1212}). \end{aligned}$$

만약 $\beta^{11}_{X_1R_1} = \beta^{11}_{X_2R_2} = 0$이 되면 응답자와 무응답자간의 응답 패턴의 차이가 없음을 의미한다. 반면에 $\beta^{11}_{X_1R_1}$, $\beta^{11}_{X_2R_2}$의 값이 크면 클수록 응답자와 무응답자간의 응답 패턴의 차이가 크다는 것을 의미한다. 본 모의실험에서는 이 두 모수의 값을 0.2에서 0.8까지 0.2 간격으로 변화를 주었다. 나머지 모수들에 대해서는 $(m_{1111}/m_{1211})/(m_{2111}/m_{2211}) = 5$, $(m_{1111}/m_{1112})/(m_{1112}/m_{1122}) = 2$로 고정하고 자료 가운데 무응답 비율이 20%가 되도록 지정하였으며 총 빈도의 수는 $N = \sum_{ijkl} m_{ijkl} = 1,000$으로 고정하였다. 이와 같이 모수를 할당한 후 다항분포를 따르는 표본을 추출하였는데 표본 추출의 결과로 변방값이 발생하는 경우와 발생하지 않은 경우로 나누어 각각 1,000 셋트의 표본을 추출하였다. 추출된 표본에 대한 모형 적합으로 본 연구에서 소개한 MCEM 방법으로 모형 추정을 시도 하였고 그 결과를 Park과 Choi (2010)이 수행한 최대추도추정 (ML)과 베이지안 방법을 이용한 추정 (Type III)과 비교하였다.

Table 3.1은 변방값이 발생한 경우 MCEM 방법, 최대 우도 추정 방법, 베이지안 추정방법에 따른 모의실험 결과를 정리한 것이다. 모의실험 결과 가운데 칸 기대빈도 m_{11++}와 m_{1112}에 대한 평균제곱오차 (MSE) 값과 편향 (bias) 값을 정리한 것이다. 괄호안의 값이 편향을 나타낸다. 모의실험 결과를 살펴보면 응답과 무응답값의 응답 패턴 차이의 크기에 상관없이 최대 우도 추정 결과보다 MCEM이나 베이지안 결과가 좋음을 확인할 수 있다. MCEM 결과와 베이지안 결과간의 차이를 비교하기 위하여 추정된 통계량들을 가지고 짝진표본 t-검정을 수행하였다. Table 3.2는 편향에 대한 검정 결과이고 Table 3.3은 평균제곱오차에 대한 검정 결과이다. 두 결과 모두 매우 작은 t-통계량을 보이고 있으며 통계적으로 유의한 차이가 없다. 이 결과를 바탕으로 최대 우도 추정을 적용한 결과 변방값 문제가 발생한다면 MCEM 방법을 이용하는 것이 또 다른 대안이 될 수가 있을 것이다. 이 결과는 Yoo (2015)의 결과를 일부 참조한 것이다.

Table 3.1 MSE and biases for $m_{11++} = \sum_{kl} m_{11kl}$ and m_{1112} when the boundary solution occurs (the numbers in parentheses are biases).

	$\beta^{11}_{X_1R_1}$	$\beta^{11}_{X_2R_2}$	MCEM	ML	Bayesian
M_{11++}	0.2	0.2	335.8 (-11.1)	429.3 (-14.8)	327.9 (-11.2)
	0.2	0.4	335.6 (-7.9)	384.0 (-11.8)	296.2 (-8.0)
	0.2	0.6	382.2 (-6.1)	444.9 (-11.1)	341.3 (-6.3)
	0.2	0.8	361.0 (-3.1)	388.1 (-7.7)	320.4 (-3.3)
	0.4	0.4	383.2 (-7.3)	440.8 (-12.0)	334.9 (-7.5)
	0.4	0.6	374.3 (-6.8)	448.0 (-12.6)	316.4 (-7.1)
	0.4	0.8	386.1 (-2.7)	411.6 (-8.6)	321.8 (-3.1)
	0.6	0.6	391.4 (-4.1)	446.4 (-10.5)	325.6 (-4.5)
	0.6	0.8	401.4 (-0.8)	395.5 (-7.8)	301.1 (1.3)
	0.8	0.8	412.6 (2.1)	380.7 (-6.0)	314.9 (1.5)
M_{1112}	0.2	0.2	84.5 (-5.2)	122.3 (-7.0)	85.5 (-5.2)
	0.2	0.4	58.8 (-4.9)	95.9 (-7.3)	60.1 (-5.0)
	0.2	0.6	35.5 (-3.8)	74.9 (-6.9)	36.6 (-3.9)
	0.2	0.8	17.5 (-1.8)	40.7 (-4.9)	18.0 (2.0)
	0.4	0.4	75.8 (-3.7)	115.0 (6.0)	77.0 (-3.8)
	0.4	0.6	56.9 (-3.4)	105.1 (-6.8)	58.9 (-3.5)
	0.4	0.8	30.0 (-1.3)	58.1 (-4.7)	30.9 (-1.5)
	0.6	0.6	79.7 (-1.6)	128.8 (5.2)	82.1 (-1.9)
	0.6	0.8	49.8 (-0.2)	79.1 (3.9)	50.7 (0.5)
	0.8	0.8	76.1 (1.3)	104.5 (-2.8)	76.5 (1.0)

Table 3.2 Paired sample t-test of biases between MCEM and Bayesin estimation when the boundary solution occurs.

Method	N	Mean	Std	$\bar{d}$	S_d	t
MCEM	20	-3.626	3.234	-0.033	-0.002	-0.033
Bayesian	20	-3.592	3.236			

Table 3.3 Paired sample t-test of biases between MCEM and Bayesin estimation when the boundary solution occurs.

Method	N	Mean	Std	$\bar{d}$	S_d	t
MCEM	20	216.409	165.808	0.283	0.226	0.005
Bayesian	20	216.125	165.582			

Table 3.4는 변방값 문제가 발생하지 않은 경우에 대하여 MCEM, 최대 우도 추정, 베이지안 추정 방법을 이용하여 모형을 적합한 후 평균제곱오차와 편향을 비교한 결과를 정리한 표이다. Park과 Choi (2010)에서 언급된 바와 마찬가지로 비임의 결측 체계 (무시할 수 없는 무응답)를 가정하였다 하더라도 변방값 문제가 발생하지 않으면 베이지안 방법이 가지는 장점 없다. 오히려 편향에서는 베이지안 방법이 더 나쁜 결과를 주고 있다. 이와 유사한 결과를 MCEM에서도 찾아볼 수 있다. 그러나 일부 경우에 있어서는 MCEM의 결과가 평균제곱오차에서 더 작은 결과를 보여주고 있다. 세 결과 간의 통계적 유의성 여부를 검증하기 위하여 반복측정 분산분석을 실시하였다. Table 3.5는 편향에 대한 반복측정 분산분석결과이다. 세 방법 간의 차이를 나타내는 개체간 효과 (between)에 대한 검정 결과를 보면 통계적으로 유의한 차이를 보이고 있다. 편향의 평균과 표준편차를 정리한 Table 3.6을 보면 최대 우도 추정 방법의 편향이 작은 것을 볼 수 있다. 변방값이 발생하지 않은 경우이기 때문에 이는 타당한 결과 하겠다. Table 3.7과 Table 3.8은 평균제곱오차에 대한 반복측정 분산분석 결과와 평균과 표준편차를 정리한 표이다. 편향 만큼 크지는 않으나 이 또한 통계적으로 유의한 차이를 보이고 있는 것을 확인 할 수 있으며 특히 MCEM의 결과가 가장 작은 값을 보이고 있다.

Table 3.4 MSE and biases for $m_{11++} = \sum_{kl} m_{11kl}$ and m_{1112} when the boundary solution does not occur (the numbers in parentheses are biases).

	$\beta^{11}_{X_1R_1}$	$\beta^{11}_{X_2R_2}$	MCEM	ML	Bayesian
M_{11++}	0.2	0.2	270.2 (2.9)	270.3 (1.1)	254.9 (2.6)
	0.2	0.4	301.0 (5.0)	305.4 (2.2)	294.9 (4.2)
	0.2	0.6	382.2 (7.7)	389.4 (5.2)	371.4 (7.9)
	0.2	0.8	394.8 (8.6)	433.8 (6.1)	435.4 (9.1)
	0.4	0.4	324.2 (6.9)	364.4 (3.6)	354.2 (6.1)
	0.4	0.6	374.3 (8.8)	423.0 (6.4)	422.8 (9.7)
	0.4	0.8	425.2 (11.2)	474.0 (7.8)	495.7 (11.3)
	0.6	0.6	391.4 (11.5)	469.2 (7.7)	450.1 (11.4)
	0.6	0.8	467.6 (14.7)	536.3 (11.1)	596.2 (15.4)
	0.8	0.8	581.2 (18.9)	654.3 (14.0)	747.6 (19.0)
M_{1112}	0.2	0.2	50.0 (1.8)	50.5 (0.6)	44.0 (1.4)
	0.2	0.4	45.5 (2.7)	49.7 (1.6)	45.2 (2.7)
	0.2	0.6	35.5 (4.7)	59.6 (3.1)	60.2 (4.8)
	0.2	0.8	66.1 (6.0)	58.0 (4.1)	64.8 (5.9)
	0.4	0.4	62.9 (3.1)	75.4 (2.1)	68.8 (3.4)
	0.4	0.6	56.9 (5.0)	76.7 (3.0)	75.4 (4.9)
	0.4	0.8	80.4 (6.5)	74.6 (4.2)	80.9 (6.3)
	0.6	0.6	79.7 (6.0)	104.3 (4.0)	104.0 (6.0)
	0.6	0.8	112.7 (8.0)	108.2 (5.6)	118.1 (7.8)
	0.8	0.8	161.9 (9.5)	136.6 (5.4)	149.4 (8.9)

Table 3.5 Repeated measurement ANOVA of biases among MCEM, ML, and Bayesin estimation when the boundary solution does not occur

Source	DF	SS	MS	F	p
Between	19	904.4	47.6	113.2	<.0001
Within	2	84.1	42.0	100.1	<.0001
Error	38	15.9	0.4		
Corrected Total	59	1,004.5			

Table 3.6 Means and standard deviations for biases

Within	N	Mean	Std
MCEM	20	7.475	4.236
ML	20	4.945	3.336
Bayesian	20	7.440	4.401

Table 3.7 Repeated measurement ANOVA of MSE among MCEM, ML, and Bayesin estimation when the boundary solution does not occur

Source	DF	SS	MS	F	p
Between	19	2,151,965	113,261	175	<.0001
Within	2	9,036	4,518	7	0.0026
Error	38	24,524	645		
Corrected Total	59	2,185,527			

Table 3.8 Means and standard deviations for MSEs

Within	N	Mean	Std
MCEM	20	233.185	175.140
ML	20	255.685	197.515
Bayesian	20	261.700	211.815

4. 실증 자료 분석

자료 분석으로 이용된 자료는 2012년 12월 치러진 18대 대통령 선거 당일 수행된 출구조사의 자료로써 지역 (광역자치단체), 시·군·구 (기초 자치단체), 성별 (남, 여), 연령 (20대, 30대, 40대, 50대, 60대 이상), 지지후보 (새누리당 후보, 민주통합당 후보, 기타, 무응답) 등이 집계되었다. 출구 조사를 시행할 때에 성별, 연령, 지지후보를 자기기입식으로 유권자가 조사를 거부할 경우에는 나이와 성별만 기재하여 지지후보만 무응답을 가지게 된다. 본 연구는 Kwak과 Choi (2014)에서 이용된 자료를 가지고 분석을 진행하였다. 자료 분석으로 사용하기 위해 19대 국회의원 선거구에 맞춰 204개의 선거구별로 재조정하여 분석에 이용하였다. 출구 조사 때 같이 조사된 성별과 연령 중 성별은 고려하지 않고 연령대 변수만 이용하였다. 20대, 30대, 40대, 50대, 60대 이상으로 5개의 범주로 구성되어 있으며 무응답은 없다. 지지후보의 경우 기타후보에 대한 지지율이 매우 미약하여 기타후보에 대한 빈도는 분석 과정에서 제외 하였다. 따라서 지지후보에 대한 변수에 대해서만 무응답을 가지게 되므로 무응답 지시변수는 한개만 존재하게 되며 무응답 지시변수와 관찰된 변수의 상호작용 효과의 설정에 따라 완전임의 결측 (MCAR), 임의 결측 (MAR) 혹은 비임의 결측 (NMAR) 체계에 대한 무응답 모형을 모두 고려할 수 있다. 모의실험의 결과 MCEM 방법이 최대 우도 추정 방법이나 베이지안 추정 방법에 비하여 나쁘지 않은 결과를 보였기 때문에 세가지 무응답 모형에 대하여 MCEM 방법을 적용하여 모형 추정을 진행하였다.

총 204개의 국회의원 선거구 자료에 대하여 모형 추정을 진행하여 추정된 결과를 살펴보았다. 지면 절약을 위하여 본 논문에서는 제시되지 않았으나 각 개별 선거구별 예측결과는 Kwak과 Choi (2014)의

결과와 큰 차이를 보이지 않았다. 추정된 결과를 이용하여 예측 결과의 정확도를 비교하기 위하여 Bautista 등 (2007)이 제안한 MWPE (modified within precinct error)를 이용하였다. 선거 예측결과이기 때문에 최종 선거가 치뤄진 이후에 실제 후보간의 지지율을 구할 수 있다. 선거 결과 당선된 후보와 차점을 차지한 후보의 지지율을 각각 P_1, P_2로 표시하고 모형으로부터 예측된 지지율을 $\hat{P}_1$, $\hat{P}_2$라 할 때 MWPE는 다음과 같이 주어진다.

$$MWPE = \frac{2P_1(1-\alpha)(P_1-1)}{P_1(1-\alpha)-1}, \quad \alpha = \frac{\hat{P}_1/\hat{P}_2}{P_1/P_2}$$

실제 지지율과 예측된 지지율간 차이가 없다면 $\alpha = 0$이 될 것이고 $MWPE$ 또한 0으로 접근할 것이다.

전체 204개 선거구 가운데 비임의 결측 체계 즉 무시할 수 없는 무응답 모형을 적용하여 먼저 최대 우도 추정을 진행하였다. 그 결과 변방값 문제가 발생한 94개 자료를 이용하여 다시 임의 결측과 비임의 결측 체계하에서 MCEM 방법을 적용하여 모형을 재 적합 하였다. 각각의 결과에 대하여 MWPE 값을 계산한 후에 짝진표본 t-검정을 수행하였다. 수행한 결과는 Table 4.1과 같다. MCEM 방법에 의하여 변방값 문제가 해결 되었음에도 불구하고 임의결측 (MAR)의 예측 정확도가 비임의결측 (NMAR)의 예측 정확도 보다 좋음을 알 수 있다. 우리나라 선거 예측에 있어서는 임의결측의 가정에 의한 예측이 보다 적절할 수 있다. 이는 Kwak과 Choi (2014), Yoon과 Choi (2014)에서와 유사한 결론을 내릴수 있다.

Table 4.1 Paired sample t-test of MWPE between MAR and NMAR

	N	Mean	Std	$\bar{d}$	S_d	t	p
MAR	94	-0.0073	0.1470	0.0933	0.0564	16.02	<.001
NMAR	94	-0.1006	0.1454				

5. 결론

본 연구의 첫 번째 목적은 무응답이나 결측이 발생하였을 때 무응답 대체를 수행하기 위한 것이다. 특히 무응답 체계에 대한 여러 가정 가운데 비임의 결측 (무시할 수 없는 무응답) 체계를 가정하였을 때 직면할 수 있는 변방값 문제를 해결하기 위하여 MCEM 방법을 제안하였다. 제안된 방법의 성능을 확인하기 위하여 모의실험을 통하여 무시할 수 없는 무응답 체계 가정 하에서 최대 우도 추정 방법, 베이지안 추정 방법간의 비교 실험을 진행하였다. 모의 실험 결과 변방값 문제가 발생하는 경우 베이지안 추정 방법이나 MCEM의 방법이 최대 우도 추정 방법에 비하여 평균제곱오차와 편향에서 모두 좋은 결과를 보였으며 베이지안 추정 방법과 MCEM 방법 간에는 유의한 차이를 보이지 않았다. 이에 반하여 변방값 문제가 발생하지 않은 경우 편향에서는 최대 우도 추정 방법이 가장 좋은 결과를 보였으며 평균제곱오차에서는 MCEM이 가장 좋은 결과를 보였다. 결론적으로 MCEM 방법이 변방값 문제를 해결하기 위한 새로운 대안 방법으로 이용될 수 있을 것이다. 또한, 실제 자료를 이용한 결과를 바탕으로 무응답 모형의 가정에 따른 예측력을 비교함으로써 무응답 체계에 대한 가정을 점검해 보고자 하였다. 이를 위하여 2012년 대선 출구조사의 자료를 이용하여 전국 204개 중 최대 우도 추정에서 변방값 문제를 발생시키는 94개 선거구에 대해서 모형 적합을 시도한 후 방법별 예측결과를 평가하기 위해서 MWPE 통계량을 이용하여 비교 분석을 진행하였다. 실제 데이터 분석에서는 무응답 체계에 대한 가정은 임의 결측과 비임의 결측으로 나누어 모형 적합을 수행하였고 모두 MCEM 방법을 적용하여 무응답 모형 추정을 수행하였다. 그 결과 임의 결측 가정에 의한 무시할 수 있는 무응답 모형에서 더 좋은 예측 결과를 보였다. 결론적으로 보았을 때 우리나라 선거 예측에 있어서는 무시할 수 있는 무응답 모형을 이용하여 예측을 수행하는 것이 더 적절하다 할 수 있겠다.

본 연구는 기본적으로 다차원 분할표 형태로 정리된 자료의 분석 기법을 다루고 있다. 다차원 분할표의 차원이 증가하거나 범주의 수가 증가하는 경우 추정하여야 하는 모수의 수가 급격하게 증가하게 되고 추정에 어려움이 따르게 된다. Dahinden 등 (2010)과 Nardi와 Rinaldo (2012)는 고차원 분할표에 대한 분해기법과 LASSO 방법을 적용하여 로그 선형 모형에 적용하여 고차원 분할표의 추정 문제를 다루었다. 추후 연구로 이와 같은 기법을 적용하여 고차원 분할표에서의 무응답 모형 추정에 대한 연구를 진행하고자 한다.

References

Baker, S. G. and Laird, N. M. (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of American Statistical Association*, **83**, 62-69.

Bautista, R., Callegaro, M., Vera, J. A. and Abundis, F. (2007). Studying nonresponse in mexican exit pollsm. *international Journal of Public Opinion Research*, **19**, 492-503.

Cho, Y. S., Chun, Y. M. and Hwang. D. Y. (2008). An imputation for nonresponses in the survey on the rural living indicators. *Korean Journal of Applied Statistics*, **21**, 95-107.

Choi, B., Choi, J. W. and Park, Y. S. (2009). Bayesian methods for an incomplete two-way contingency table with application to the Ohio (Buckeye state polls). *Survey Methodology*, **35**, 37-51.

Choi, B. and Kim, K. M. (2012). A model selection method using em algorithm for missing data. *Journal of the Korean Data Analysis Society*, **14**, 767-779.

Choi, B., Park, Y. S. and LEE, D. H. (2007). Election forecasting using pre-election survey data with nonignorable nonresponse. *Journal of the Korean Data Analysis Society*, **9**, 2321-2333.

Crespi, I. (1988). *Pre-election polling: Sources of accuracy and error*, Russel Sage, New York.

Dahinden, C., Kalisch, M. and Buhlmann, P. (2010).Decomposition and model selection for large contingency tables. *Biometrical Journal*, **52**, 233-252.

Hong, N. R. and Huh, M. H. (2001). A post-examination of forecasting surveys for the 16th general election. *Survey Research*, **2**, 1-35.

Ibrahim, J. G., Zhu, H. and Tang, N. (2008). Model selection criteria for missing-data problems using the EM algorithm. *Journal of American Statistical Association*, **103**, 1648-1658.

Kim, Y. W. and Kwak, E. S. (2010). A total survey error analysis of the exit polling for general election 2008 in Korea. *Survey Research*, **11**, 33-55.

Kwak, E. S., Kim, J. Y. and Kim, Y. W. (2013). Analysis of forecasting error of the exit poll for the general election of 2012 in Korea. *Survey Research*, **14**, 1-7.

Kwak, J. A. and Choi, B. (2014). A comparison study for accuracy of exit poll based on nonresponse model. *Journal of the Korean Data & Information Science Society*, **25**, 53-64.

Lee, H. J. and Kang, S. B. (2012). Handling the nonresponse in sample survey. *Journal of the Korean Data & Information Science Society*, **23**, 1183-1194.

Lee, J. H., Kim. J. and Lee, K. J. (2006). Missing imputation methods using the spatial variable in sample survey. *Korean Journal of Applied Statistics*, **19**, 57-67.

Little, J. A. and Rubin, D. B. (2002). *Statistical analysis with missing data*, second edition, Wiley, New York.

Nardi, Y. and Rinaldo, A. (2012). The log-linear group-lasso estimator and its aymptotic properties. *Bernoulli*, **13**, 945-974.

Park, T. (1998). An approach to categorical data with nonignorable nonresponse. *Biometrics*, **54**, 1579-1690.

Park, T. and Brown, M. B. (1994). Models for categorical data with nonignorable nonresponse. *Journal of American Statistical Association*, **89**, 44-52.

Park, T. S. and Lee, S. Y. (1998). General research papers : analysis of categorical data with nonresponses. *Korean Journal of Applied Statistics*, **11**, 83-95.

Park, Y. S. and Choi, B. (2010). Bayesian analysis for incomplete multi-way contingency tables with nonignorable nonresponse. *Journal of Applied Statistics*, **37**, 1439-1453.

Park, Y. S., Kim, K. W. and Choi, B. (2013). Dynamic Bayesian analysis for irregularly and incompletely observed contingency tables. *Journal of the Korean Statistical Society*, **42**, 277-289.

Shim, M. S. and Choi, H. C. (1997). Studies on non-response cases of election polls. *The Journal of Communication Science*, **14**, 137-162.

Wei, G. C. G. and Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of American Statistical Association*, **85**, 699-704.

Yoo, H. S. (2015). *A model selection method for non-response model based on empirical Bayesian method*, Master Thesis, Daegu University, Gyeongbuk.

Yoon, Y. H. and Choi, B. (2012). Model selection method for categorical data with non-response. *Journal of the Korean Data & Information Science Society*, **23**, 627-641.

Yoon, Y. H. and Choi, B. (2014). Analysis of missing data using an empirical Bayesian method. *Korean Journal of Applied Statistics*, **27**, 1003-1016.

Journal of the Korean Data & Information Science Society
2016, **27**(3), 587–598

http://dx.doi.org/10.7465/jkdi.2016.27.3.587
한국데이터정보과학회지

An estimation method for non-response model using Monte-Carlo expectation-maximization algorithm[†‡]

Boseung Choi[1] · Hyeon Sang You[2] · Yong Hwa Yoon[3]

[1]Department of Applied Statistics, Korea University
[2]Marketing Department, Daegu Bank
· [3]Department of Statistics and Computer Science, Daegu University

Received 15 March 2016, revised 19 April 2016, accepted 20 April 2016

Abstract

In predicting an outcome of election using a variety of methods ahead of the election, non-response is one of the major issues. Therefore, to address the non-response issue, a variety of methods of non-response imputation may be employed, but the result of forecasting tend to vary according to methods. In this study, in order to improve electoral forecasts, we studied a model based method of non-response imputation attempting to apply the Monte Carlo Expectation Maximization (MCEM) algorithm, introduced by Wei and Tanner (1990). The MCEM algorithm using maximum likelihood estimates (MLEs) is applied to solve the boundary solution problem under the non-ignorable non-response mechanism. We performed the simulation studies to compare estimation performance among MCEM, maximum likelihood estimation, and Bayesian estimation method. The results of simulation studies showed that MCEM method can be a reasonable candidate for non-response model estimation. We also applied MCEM method to the Korean presidential election exit poll data of 2012 and investigated prediction performance using modified within precinct error (MWPE) criterion (Bautista et al., 2007).

Keywords: Generalized linear model, imputation, MCEM, missing data, non-ignorable non-response.

[†] This research is supported by Daegu University research grant in 2013 (No.20130310).
[‡] This research is written by addtional research based on the master's dissertation of the second author, You.
[1] Assistant professor, Department of Applied Statistics, Korea University, Sejong 30019, Korea.
[2] Staff, Marketing Department, Daegu bank, Deagu 42123, Korea.
[3] Corresponding author: Professor, Department of Statistics and Computer Science, Daegu University, Gyeongbuk 38453, Korea. E-mail: yhyoon@daegu.ac.kr