The Korean Journal of Applied Statistics (2014) DOI: http://dx.doi.org/10.5351/KJAS.2014.27.6.1003
27(6), 1003–1016

Analysis of Missing Data Using an Empirical Bayesian Method

Yong Hwa Yoon[a] · Boseung Choi[a,1]

[a]Department of Computer Science and Statistics, Daegu University

(Received August 14, 2014; Revised October 13, 2014; Accepted October 20, 2014)

Abstract

Proper missing data imputation is an important procedure to obtain superior results for data analysis based on survey data. This paper deals with both a model based imputation method and model estimation method. We utilized a Bayesian method to solve a boundary solution problem in which we applied a maximum likelihood estimation method. We also deal with a missing mechanism model selection problem using forecasting results and a comparison between model accuracies. We utilized MWPE(modified within precinct error) (Bautista *et al.*, 2007) to measure prediction correctness. We applied proposed ML and Bayesian methods to the Korean presidential election exit poll data of 2012. Based on the analysis, the results under the missing at random mechanism showed superior prediction results than under the missing not at random mechanism.

Keywords: Missing data, non-response, Empirical Bayesian, EM algorithm.

1. 서론

각종 여론조사를 시행하는 기본적인 목적은 조사의 대상이 되는 사람들이 가지고 있는 생각을 파악하고 예측하고자 하는데 있다. 특히 선거를 앞두고 시행되는 여러 사전조사에서는 예측이 더 중요한 목적이다. 공인된 대부분의 조사들은 그동안 구축된 과학적인 방법을 통하여 조사를 함으로써 여론조사를 통한 예측의 정확도를 높이고자 부단히 노력하고 있다. 하지만 많은 비용을 들이고 과학적으로 정교한 조사를 수행한 경우라 할지라도 실제값과 조사결과의 차이인 오차가 발생할 수 밖에 없다. 이러한 오차를 얼마나 잘 줄이는가 하는 문제는 조사의 정확도를 높이는 중요한 관건이라 할 수 있다. 조사 과정에서 발생할 수 있는 여러 오차 요인 가운데 가장 빈번하게 발생하는 문제는 무응답 또는 결측치의 문제라 할 수 있다. 따라서 무응답과 결측치에 대한 처리 문제는 조사의 결과의 정확도를 높이는데 있어서 매우 중요한 문제라 할 수 있다. 결측 처리를 위한 방법으로 그동안 다양한 방법이 제시되어 왔다. 주로 조사의 관점에서 결측 또는 무응답에 대한 대체 및 처리 연구가 진행되어 왔으며 다른 한편으로는 수리적인 모형에 기반하여 대체를 수행하고 추정하는 문제를 다루고자 하였다.

Kim과 Kwon (2009)는 통계청의 결제활동인구조사의 결과를 이용하여 조사현장에 발생하는 무응답 현황을 파악하고 무응답 유형별 오차를 평가하여 무응답 오차의 기여도가 큰 무응답 유형을 찾고자 하였

This research is supported by Daegu University Research Grant in 2012(No.21020499).
[1]Corresponding author: Department of Computer Science and Statistics, Daegu University, 201, Daegudae-ro, Gyeongsan-si, Gyeongsangbuk-do, 712-714, Republic of Korea. E-mail: bchoi@daegu.ac.kr

다. Kim과 Nam (2009)는 실제 조사자료를 이용하여 무응답 조정층의 구성에 따라 어떠한 구성이 무응답 대체에 효과적인지에 대한 실증 분석 및 모의실험을 진행하였다. Pak과 Shin (2010)은 패널자료에서 발생하는 무응답 처리를 위하여 시계열-횡단명 다중 대체법을 제안하였는데 이 방법은 시계열 대체법과 횡단명 무응답 대체법을 결합한 방법이라 할 수 있다. 조사에서 발생하는 결측자료를 대체하기 위한 방법 가운데 내제적인 방법에 의한 대체방법으로 널리 쓰이는 핫댁대체나 콜드덱대체 방법이 있다. Song (2011)은 핫덱 대체를 적용하는데 있어서 결측이 발생한 벼수 뿐만 아니라 결측 발생여부와 관련된 변수를 함께 고려하여 대체군을 생성하는 방법을 제안하였다. 또한 Song (2014)는 핫덱대체 방법을 다중응답에서 발생하는 무응답 문제에 적용하고자 하였다. Park 등 (2013)은 조사자료에서 발생하는 결측 가운데 특히 범부형 자료에 대한 결측 문제를 다루었다. 기존에 잘 알려져 있는 범주형 자료에 대한 결측방법에 대하여 실제 조사 자료를 이용한 모의실험을 통한 비교 분석을 수행하였다. 그 결과로는 상대적으로 단순한 방법을 적용한 대체 방법이 더 우수한 것으로 판명되었다. 본 연구에서는 조사나 자료수집 과정에서 발생할 수 있는 무응답 또는 결측치의 처리 문제를 다루고자 한다. 특히 범주형 자료에서 발생하는 무응답 혹은 결측에 대한 처리 문제를 다루고자 한다. 또한 본 연구에서는 모형에 기반한 결측 대체 및 모형의 추정 문제를 다루고자 한다.

결측 처리에 대한 문제를 들어가기 앞서 고려하여야 하는 사항은 결측 체계에 대한 가정이다. Little과 Rubin (2002)는 우도함수에 기반하여 결측 체계를 세 가지로 구분 하였다. 반응변수를 $Y = \{y_i\},\ i = 1, 2, \ldots, n$라 하자. Y를 다시 Y_{obs}와 Y_{mis}로 나눌 수 있는데 Y_{obs}는 완전히 관찰된 자료이고 Y_{mis}는 결측 자료를 나타낸다. 즉 $Y = (Y_{\text{obs}}, Y_{\text{mis}})$로 표현할 수 있다. 관찰자료 Y에 1대1로 대응하는 행렬 혹은 벡터 M을 함께 정의한다. M은 Y의 결측 여부를 나타내는 지시 변수로 이루어져 있다. 즉 i번째 관찰치에 대하여 y_i가 관찰된 자료인 경우 $m_i = 1$이 되고 결측인 경우 2이 된다. 결측 지시변수 M의 분포함수가 모수 벡터 θ에 대하여 다음의 식을 만족할 때

$$f(m|y, \theta) = f(m|\theta), \quad \text{for all } \ y_i,\ \theta,$$

결측 체계는 완전임의결측(MCAR; missing completely at random)이라 부른다. 반면에 다음의 식을 만족하면

$$f(m|y, \theta) = f(m|y_{\text{obs}}, \theta), \quad \text{for all } \ y_{\text{mis}},\ \theta, \tag{1.1}$$

결측 체계는 임의결측(MAR; missing at random)이 된다. 반면에 결결측 지시변수의 분포함수가 결측인 y_{mis}에 영향받는 경우 이를 비임의 결측(NMAR; not missing at random)이라 부른다.

결측을 포함한 자료에 대하여 우도함수에 기반한 모수의 추정 문제를 고려하여 보자. 이 때 결측 체계는 다시 무시할 수 있는 결측 체계(ignorable missing mechanism)와 무시할 수 없는 결측 체계(non-ignorable missing mechanism)로 구분할 수 있다. 모수벡터 θ가 서로 구분되어 지는 두 부분의 모수벡터로 구분되어 질 때, 즉 $\theta = (\phi, \psi)$로 구분되어 질 때 모수에 대한 우도함수가 각각의 모수 벡터 ϕ와 ψ로 구분되어 지고 결측체계가 MAR이면 이를 무시할 수 있는 결측체계 혹은 무시할 수 있는 무응답이라 부른다. 즉 우도함수가 다음의 식을 만족할 때 이다.

$$L(\phi, \psi|y_{\text{obs}}m) \propto f(y_{\text{obs}}, m|\phi, \psi) = f(m|y_{\text{obs}}, \phi) f(y_{\text{obs}}|\psi). \tag{1.2}$$

식 (1.2)의 우변 첫 번째 식은 식 (1.1)과 같음을 볼 수 있다. 이와 달리 식 (1.2)에서 우변의 첫째식이 $f(m|y_{\text{mis}}, \phi)$이 되면 무시할 수 없는 결측체계를 따르게 된다.

모형에 기반하여 결측을 포함한 자료에 대한 모형 추정 문제는 많은 연구가 진행 되었다. Baker와 Laird (1988)은 다차원 분할표 형태로 정리된 범주형 자료에 대하여 결측을 포함하고 있을 때 로그선

형 모형을 이용한 최대우도추정 방법을 제안하였다. 이 때 결측치 추정을 위하여 EM 알고리즘 (Dempster 등, 1977)을 이용하였다.

최대우도추정 방법을 이용하여 결측 자료에 대한 모수 추정을 수행할 때 무시할 수 없는 결측 체계를 가정하여 모수를 추정하는 경우 변방값(boundary solution) 문제가 발생할 수 있다. 이는 모수의 추정치가 모수 영역의 변방에 걸리게 되어 국소 최대값을 가지는 문제를 말한다 (Baker와 Laird, 1988). 변방값 문제가 발생하게 되면 추정치의 결과가 불안정 해지고 분산이 발산하는 문제가 발생한다 (Park과 Brown, 1994; Choi 등, 2009). 또한 계수자료에 대한 결측치의 추정에 있어서 0의 값을 가지게 되거나 일방적으로 큰 값을 가지게 되는 문제가 발생하게 된다. Baker와 Laird (1988)은 2×2형태의 간단한 분할표 자료에 대하여 변방값 문제가 발생하는 조건을 제시하였다. Baker 등 (1992)는 변방값 문제가 발생하는 다양한 조건을 제시하고 또한 대안적인 모수 추정치를 구하는 방법을 제시하였다. Clarke (2002)는 변방값 문제가 발생하는 원인을 도표적 방법을 이용하여 설명하고자 하였다.

이러한 변방값 문제를 해결하여 모수를 추정하기 위한 방법으로 다양한 베이지안 방법이 제안되어 왔다. Park과 Brown (1994)는 결측을 가지는 분할표 자료에 대하여 각 칸의 기대확률에 사전분포로 공액(conjugate)관계에 있는 Dirichlet 분포를 할당하는 베이지안 방법을 이용하여 변방값 문제를 해결하고자 하였다. 이 때 사전분포의 할당은 결측이 발생한 칸의 기대 확률에만 부여하였으며 사전분포의 모수가 결측이 발생하지 않은 관찰된 자료에만 의존하도록 하였다. 이들은 제시한 방법은 일종의 경험적 베이지안 방법이라 할 수 있다. Park (1998)과 Park과 Lee (1998)은 유사한 자료에 대하여 결측이 발생한 칸 뿐만 아니라 결측이 발생하지 않은 칸의 기대확률에도 사전분포를 할당하는 경험적 베이지안 방법을 제시하였다. Choi 등 (2007)은 사전분포의 칸 기대확률에 사전분포를 할당하는데 있어서 EM 알고리즘에서 발생하는 기대 빈도를 사전분포의 초모수로 할당하는 방법을 제안하였다. Choi 등 (2009)와 Park과 Choi (2010)은 사전분포의 초모수를 할당하는데 있어서 관찰빈도와 칸 기대빈도에 대한 최대우도 추정치를 동시에 이용하는 방법을 제시하였고 다양한 상황하에서 모의실험을 통하여 제안한 방법에 대한 성능을 검증하였다. Choi 등 (2008)은 교체표본조사에서 발생하는 결측 문제를 해결하기 위하여 EM 알고리즘에 기반한 최대우도 추정 방법을 제안하였다.

베이지안 방법에 기반하여 문제를 해결하고자 하는 또 다른 흐름은 계층적 베이지안 모형을 이용하는 방법이 제안되었다. Forster와 Smith (1998)는 분할표 자료의 각 칸의 기대확률에 Dirichlet 분포를 사전분포로 할당한 후 계층적 베이지안 방법을 적용하고 Markov Chain Monte Carlo(MCMC) 기법을 사용하여 모수를 추정하고자 하였다. 이 때 결측체계의 가정에 따라 무시할 수 있는 결측체계의 경우 잘 알려진 형태의 조건부 사후분포부터 모수를 추출하는 Gibbs sampler 방법을 이용하였고 무시할 수 없는 결측체계에서는 Metropolis-Hatings 방법을 사용하여 조건부 사후분포로부터 모수를 추출하는 방법을 이용하였다.

이와는 조금 다른 접근 방법으로 Green과 Park (2003)는 분할표 형태의 관찰된 자료에 대하여 다항분포를 가정하고 로그 선형모형을 결합하는 일반화 선형모형을 이용하였다. 여기서 일반화 선형모형의 체계적 성분의 모수에 사전분포를 할당하는 계층적 베이지안 방법을 제안하였다. 체계적 성분의 모수에는 다변량 정규분포를 사전분포로 할당하였고 체계적 성분의 오차항을 도입하고 그 오차의 분산에 역감마분포를 사전분포로 할당하였다. 조건부 사후분포로부터 모수를 주출하기 위하여 기본적으로 Gibbs sampler 방법을 이용하였으며 잘 알려지지 않은 형태의 조건부 사후분포로부터 모수를 추출하기 위하여 Metropolis-Hastings 방법을 이용하였다. Park과 Choi (2010)은 Green과 Park (2003)의 방법을 확장하여 분할표 자료가 시계열 자료의 형태를 가지는 경우에 있어서 동적 베이지안 모형(Dynamic Bayesian)을 적용하여 조건부 사후분포로부터 모수를 추출하는 계층적 베이지안 방법을 제안하였다.

본 연구에서 언급하고 있는 결측자료의 처리 방법들은 모두 모형에 기반한 방법들이며 또한 결측 체계

를 미리 가정하고 모형의 추정을 수행하는 방법들이다. 결측자료 발생시 주의깊게 관찰하여야 하는 상황 가운데 하나는 각종 조사에서 발생하는 무응답문제라 할 수 있다. 서론에서 언급하고 있는 대부분의 연구들은 조사 자료에 대한 결측 혹은 무응답 문제를 다루고 있다. 특히 선거를 앞두고 실시되는 여론조사와 같이 민감한 주제를 다루는 연구에서는 결측 체계에 대한 가정이 매우 중요하다. Rubin 등 (1995)은 이러한 상황하에서는 결측 체계 혹은 무응답 체계의 가정으로 무시할 수 없는 결측(무응답) 체계가 보다 적절하다고 언급하였다. 그러나 실제 상황에서 어떠한 결측체계가 더 좋은지를 판단하는 문제는 간단하지 않다. 결측 체계의 선택문제는 일종의 모형 선택의 문제로 확인해 볼 수 있다. Yoon과 Choi (2012)는 계층적 베이지안 방법에 의하여 추정된 모형에 대하여 Chib (1995)와 Chib와 Jeliazkov (2001)에 의하여 소개된 베이즈 인자(Bayes factor)를 계산하여 모형 선택에 대한 연구를 진행하였다. Choi와 Kim (2012)는 우리나라 선거 자료에 이용하여 Ibrahim 등 (2008)이 제안한 EM 알고리즘의 기반하에 수행된 모형 추정에서 모형 선택의 문제를 다루고자 하였다. 하지만 이들의 연구 또한 명확한 해답을 주지는 못하고 있다. Kwak과 Choi (2014)는 실제자료를 이용하여 각 결측 체계의 가정에 따른 모형 추정결과와 실제 투표 결과를 비교함으써 어떠한 결측 체계의 가정이 적합한지에 대한 연구를 진행하였다. 이 연구는 실증 자료에 기반한 경험적 연구라 할 수 있다. Kwak과 Choi (2014)의 연구는 EM 알고리즘의 이용한 최대우도 추정 결과에 기반하여 연구를 진행하였다. 본 연구는 Kwak과 Choi (2014)의 연구를 확장하여 최대우도 추정 방법이 가지는 문제를 다양한 베이지안 모형을 적용하여 해결한 후 이를 다시 실제자료에 적용하여 결측 체계 또는 무응답 체계의 선택 문제를 다루고자 한다.

본 연구의 진행 순서는 다음과 같다. 2장에서는 본 연구에서 적용한 여러 베이지안 방법을 소개 한다. 3장에서는 지난 2012년 말에 시행된 우리나라의 17대 대통령 선거에서 진행된 출구조사의 결과를 이용한 실증 분석 결과를 설명한다. 마지막 4장에서는 결론으로 본 연구 결과에 대한 정리와 한계점에 대하여 논하고자 한다.

2. 결측 체계에 따른 모형 추정 방법

본 장에서는 결측(무응답) 체계에 따른 모형을 정의하고 모형에 정의에 따라 EM 알고리즘을 이용한 모형 추정방법 그리고 베이지안 방법을 이용한 모형 추정 방법을 설명하고자 한다.

2.1. 결측 모형

결측 혹은 무응답을 가지고 있는 범주형 자료를 고려하여 보자. 2차원 분할표를 구성하는 두 변수를 각각 X_1, X_2라 하자. X_1은 R개의 범주로 구성되어 있고 X_2는 C개의 범주로 구성되어 있다. 그리고 두 변수 X_1, X_2 가운데 X_2에만 결결측이 발생한다고 가정하고 이에 대한 결측 지시변수를 M이라 한다. 변수 X_2가 관찰된 경우이면 $M = 1$이 되고 결측 이면 $M = 2$가 된다. 관찰된 자료만 고려한다면 $R \times C$ 분할표로 표현할 수 있으며 결측 자료까지 고려한 경우 부분적인 삼차원 분할표로 표시할 수 있다. 분할표의 각 칸의 빈도를 y라고 하자. 변수 X_1을 행 변수로 하고 X_2를 열 변수라 할 때 각 변수의 수준 i와 j에서 관찰된 빈도는 y_{ij1}으로 표시한다. 변수 X_2에서만 결측이 발생한다고 가정하였기 때문에 X_1만 관찰된 경우의 칸 빈도는 y_{i+1}로 표시할 수 있다. 여기서 열 변수의 수준을 $+$로 표시함으로 열 변수에 대한 주변합으로 표시하였다. 이와 같은 변수 정의에 대하여 $R = C = 2$인 경우 주어진 다음과 같이 주변합이 주어진 2×2 형태로 표현할 수 있다 (Table 2.1).

이제 주변합으로만 주어진 빈도에 적절한 결측 대체 방법을 적용하여 결측치에 대한 의사관찰빈도 y^*_{ij2}를 구하게 되면 이전의 Table 2.1은 다음과 같은 삼차원 분할표로 확장될 수 있다 (Table 2.2).

주어진 데이터가 Table 2.1가 같이 주어졌을 때 로그선형모형을 이용한 결측자료 추정을 위한 모형설정

Table 2.1. 2×2 contingency table with supplemental margin

	Observed ($M = 1$)		Missing ($M = 2$)
	$X_2 = 1$	$X_2 = 2$	
$X_1 = 1$	y_{111}	y_{121}	y_{1+2}
$X_1 = 2$	y_{211}	y_{221}	y_{2+2}

Table 2.2. $2 \times 2 \times 2$ contingency table after missing imputation

	Observed ($M = 1$)		Missing ($M = 2$)	
	$X_2 = 1$	$X_2 = 2$	$X_2 = 1$	$X_2 = 2$
$X_1 = 1$	y_{111}	y_{121}	y^*_{112}	y^*_{122}
$X_2 = 2$	y_{211}	y_{221}	y^*_{212}	y^*_{222}

은 다음과 같다. 분할표자료의 각 칸의 기대확률을 π_{ijl}이라 할때 관찰된 자료에 대하여 다항분포를 가정하면 우도함수는 다음의 식에 비례한다.

$$L \propto \prod_i \prod_j \pi_{ij1}^{y_{ij1}} \prod_i \pi_{i+2}^{y_{i+2}}. \tag{2.1}$$

여기서 $\pi_{ij1} = \Pr(X_1 = i, X_2 = j, M = 1)$, $\pi_{i+2} = \Pr(X_1 = i, M = 2)$이고, $N_1 = \sum_{i=1}^{R} \sum_{j=1}^{C} y_{ij1}$, $N_2 = \sum_{i=1}^{R} y_{i+2}$, $N_1 + N_2 = N$이다. 다항분포의 정의에 의하여 N_1, N_2, N는 고정된 값으로 가정한다. 확장된 3차원 분할표에 대하여 각 칸의 기대빈도를 $\mu_{ijl} = N \times \pi_{ijl}$라 할 때 이 기대빈도에 대한 로그선형모형에서의 선형 예측식은 다음과 같다.

$$\log \mu_{ijl} = z'_{ijl}\boldsymbol{\beta}. \tag{2.2}$$

여기서 $\boldsymbol{\beta}$는 선형 예측식의 모수 벡터이고 z_{ijl}은 계획행렬 Z에서 μ_{ijl}에 대응하는 행을 나타낸다. 이제 식 (2.2)의 모수 벡터 $\boldsymbol{\beta}$의 정의에 따라 결측모형을 가정할 수 있다. Little과 Rubin (2002)의 정의에 따른 완전임의결측(MCAR), 임의결측(MAR), 비임의결측(NMAR)은 다음과 같이 정의할 수 있다.

$$\begin{aligned}
\text{MCAR Model}: \quad & \log \mu_{ijl} = \beta_0 + \beta^i_{X_1} + \beta^j_{X_2} + \beta^l_M + \beta^{ij}_{X_1X_2}, \\
\text{MAR Model}: \quad & \log \mu_{ijl} = \beta_0 + \beta^i_{X_1} + \beta^j_{X_2} + \beta^l_M + \beta^{ij}_{X_1X_2} + \beta^{il}_{X_1M}, \\
\text{NMAR Model}: \quad & \log \mu_{ijl} = \beta_0 + \beta^i_{X_1} + \beta^j_{X_2} + \beta^l_M + \beta^{ij}_{X_1X_2} + \beta^{jl}_{X_2M}.
\end{aligned} \tag{2.3}$$

식 (2.3)의 두 번째 식은 결측이 발생하지 않은 변수 X_1과 결측지시변수 M과의 상호작용 효과 $\beta^{il}_{X_1M}$를 모형에 포함하고 있으므로 임의결측 모형이 되고 세 번째 식은 결측이 발생한 변수 X_2와 결측지시변수 M과의 상호작용 효과 $\beta^{jl}_{X_2M}$를 모형에 포함하고 있으므로 비임의결측 모형이 된다.

2.2. EM 알고리즘을 이용한 결측모형 추정

관찰된 자료에 대하여 다항분포 가정에 의한 로그선형모형의 로그우도함수는 랜덤 성분 식 (2.1)과 체계적 성분 식 (2.2)를 결합하여 다음의 식에 비례한다.

$$\begin{aligned}
\log L = l \propto & \sum_i \sum_j y_{ij1} \left(z'_{ij1}\boldsymbol{\beta} \right) - \sum_i \sum_j y_{ij1} \log \left(\sum_{ijl} \exp \left(z'_{ijl}\boldsymbol{\beta} \right) \right) \\
& + \sum_i y_{i+2} \log \left(\sum_j \exp \left(z'_{ij2}\boldsymbol{\beta} \right) \right) - \sum_i y_{i+2} \log \left(\sum_{ijl} \exp \left(z'_{ijl}\boldsymbol{\beta} \right) \right).
\end{aligned}$$

로그우도함수를 최대화 하는 모수의 추정치를 구하기 위하여 Dempster 등 (1977)이 제안한 GEM (genralized expectation maximization) 알고리즘을 이용하였다. GEM 알고리즘의 절차는 다음과 같다. 먼저 적절하게 모수 벡터 $\boldsymbol{\beta}$의 초기치가 주어졌을 때 E-step에서는 결측이 발생한 빈도에 대하여 의사 관찰빈도의 기대값을 계산하게 된다. 열변수 X_2의 결측에 대하여 주변합 y_{i+2}가 주어졌을 때 이전 단계에서 계산된 $\boldsymbol{\beta}^{\text{old}}$를 가지고 재 계산된 칸 기대확률 π_{ij2}/π_{i+2}를 이용하여 의사관찰빈도를 계산하게 된다.

$$y_{ij2}^{*} = E\left(y_{ij2}|\pi_{ijk}^{\text{old}}, y_{i+2}\right) = y_{i+2}\frac{\pi_{ij2}^{\text{old}}}{\pi_{i+2}^{\text{old}}} = y_{i+2}\frac{m_{ij2}^{\text{old}}}{m_{i+2}^{\text{old}}}.$$

여기서 m_{ij2}^{old}는 식 (2.2)로 부터 계산된 추정치 이다.

$$m_{ijl}^{\text{old}} = \hat{\mu} = \exp\left(z_{ijl}'\boldsymbol{\beta}^{\text{old}}\right).$$

M-step에서는 E-step에서 계산된 결측 빈도에 대한 의사관찰빈도를 마치 관찰된 빈도로 고려하여 데이터가 Table 2.2와 같은 3차원 분할표 자료와 같이 주어졌다는 가정하에 선형 예측식의 모수에 대한 최대우도 추정치를 구한다. 이 때 모수의 최대우도추정치를 구하는 방법은 일반적인 로그선형모형에서의 모수 추정 방법을 이용할 수 있으며 본 연구에서는 반복적 재가중 최소제곱법(iterative re-weighted least method) (Agresti, 2002)을 이용하여 모수에 대한 최대우도추정치를 계산하였다.

이와 같은 E-step과 M-step을 번갈아가며 반복 수행한다. 반복 수행 과정에서 이전단계에서 계산된 모수의 추정치와 현 단계에서 계산된 모수의 추정치 가운데 가장 작은 값의 차이가 10^{-10}보다 작아질 때까지 반복수행을 진행하였다.

2.3. 베이지안 방법

2.2절에서 소개한 방법은 결측치를 포함하고 있는 범주형 자료에 대하여 모수를 추정하고 결측치를 대체하기 위한 매우 일반적인 방법이라 할 수 있다. 일반화선형모형의 선형 예측식으로 식 (2.3)의 적용에 따라 결측체계의 가정에 따른 결측 모형을 선택하여 모수를 추정할 수 있다. 그러나 결측모형의 추정 가운데 비임의결측에 대한 가정하에 즉 식 (2.3)의 세 번째 식을 선택하여 EM 알고리즘을 이용하여 최대우도추정을 수행하는 경우 변방값 문제가 발생할 수 있다. 변방값 문제를 해결하기 위하여 여러 연구가 진행되어 왔는데 많은 연구들이 베이지안 방법을 적용하여 변방값문제를 해결하고자 하였다. 모수에 대하여 적절한 사전분포를 할당한 후 사후분포로부터 모수에 대한 추정을 수행하고자 하였는데 모수에 대한 사전분포를 할당하는 방법으로는 각 칸의 기대확률에 직접 사전분포를 할당하는 방법과 로그선형모형의 선형 예측식의 모수에 사전분포를 할당하는 방법이 있다. 본 연구에서는 칸 기대확률에 사전분포를 할당하는 방법을 이용하였다.

관찰된 자료에 대하여 다항분포를 가정하였기 때문에 각 칸의 기대확률에 대한 사전분포로는 공액(conjugate) 관계에 있는 Dirichlet 분포를 사전분포를 할당하는 것이 타당하다. 사전분포는 다음과 같이 주어진다.

$$\prod_{i=1}^{2}\prod_{j=1}^{2}\prod_{l=1}^{2}\pi_{ijl}^{\delta_{ijl}}. \tag{2.4}$$

이 식에서 δ_{ijl}은 사전분포의 초모수를 나타낸다. 이제 우도함수 식 (2.1)과 사후분포함수 식 (2.4)를 결합한 후 사후분포함수를 구할 수 있다. 사후분포함수에 log를 취한 로그 사후분포함수는 다음 식에 비례

Table 2.3. Hyper parameters for Dirichlet prior

	δ_{ij1}	δ_{ij2}
BA1	$\nabla_1 \frac{y_{ij1}}{y_{++1}}$	$\nabla_2 \frac{y_{ij1}}{y_{++1}}$
BA2	$\nabla_1 \frac{m_{ij1}}{m_{++1}}$	$\frac{\nabla_2}{2}\left(\frac{m_{ij2}}{m_{++2}} + \frac{1}{R \times C}\right)$

한다.

$$
\begin{aligned}
l_{\text{pos}} \propto & \sum_i \sum_j y_{ij1} \left(z'_{ij1}\boldsymbol{\beta}\right) - \sum_i \sum_j y_{ij1} \log\left(\sum_{ijl} \exp\left(z'_{ijl}\boldsymbol{\beta}\right)\right) \\
& + \sum_i y_{i+2} \log\left(\sum_j \exp\left(z'_{ij2}\boldsymbol{\beta}\right)\right) - \sum_i y_{i+2} \log\left(\sum_{ijl} \exp\left(z'_{ijl}\boldsymbol{\beta}\right)\right) \\
& + \sum_i \sum_j \sum_l \delta_{ijl} \left(z'_{ijl}\boldsymbol{\beta}\right) - \sum_i \sum_j \sum_l \delta_{ijl} \log\left(\sum_{ijl} \exp\left(z'_{ijl}\boldsymbol{\beta}\right)\right). \qquad (2.5)
\end{aligned}
$$

이제 이 사후분포 식 (2.5)로부터 최대사후추정치를 구하기 위한 GEM 알고리즘에 대하여 알아보자. 2.2절에서와 마찬가지로 먼저 적절한 $\boldsymbol{\beta}$초기치가 주어지고 주변합 y_{i+2}가 주어졌을 때 E-step에서는 의사관찰빈도에 대한 기대값을 계산한다. E-step을 수행하기 위한 확장된(augmented) 로그 사후분포 함수는 다음과 같이 주어진다.

$$
\begin{aligned}
l_{\text{a.pos}} \propto & \sum_i \sum_j \left(y_{ij1} + \delta_{ij1}\right) \left(z'_{ij1}\boldsymbol{\beta}\right) - \sum_i \sum_j \left(y_{ij1} + \delta_{ij1}\right) \log\left(\sum_{ij2} \exp\left(z'_{ij2}\boldsymbol{\beta}\right)\right) \\
& + \sum_i \sum_j \left(y^*_{ij2} + \delta_{ij2}\right) \left(z'_{ij2}\boldsymbol{\beta}\right) - \sum_i \sum_j \left(y^*_{ij2} + \delta_{ij2}\right) \log\left(\sum_{ij2} \exp\left(z'_{ij2}\boldsymbol{\beta}\right)\right). \qquad (2.6)
\end{aligned}
$$

E-step에서는 이 확장된 로그사후분포함수 식 (2.6)으로부터 의사관찰빈도에 대한 기대값을 계산하고 M-step에서는 사후분포함수를 최대화하는 최대사후추정치(MPE; maximum posterior estimate)를 계산한다. E-step과 M-step의 계산은 2.2절에서 설명한 방법과 동일하다.

2.4. 사전분포의 초모수 할당

베이지안 방법의 완성을 위하여 사전분포인 Dirichlet 분포에 적절한 모수를 할당하여야 한다. 사전분포의 모수 할당과 관련된 연구로는 Park과 Brown (1994), Park (1998), Park과 Lee (1998), Forster와 Smith (1998), Choi 등 (2007, 2009), Park과 Choi (2010) 등의 연구가 있다. 이 가운데 본 연구에서는 선행연구의 모의실험에서 상대적으로 좋은 성능을 보여온 Park (1998)의 방법과 Choi 등 (2009)의 방법을 이용하고자 하였다. 다음 Table 2.3 에서 BA1으로 표시한 Park (1998)의 방법은 사전분포의 모수를 할당하는데 있어서 결측이 발생하지 않은 칸의 빈도를 계산하여 그 빈도에 비례하도록 하였다. 또한 전체 초모수의 합이 선형예측식에서의 모수의 갯수에 비례하도록 하였다. 또한 BA2로 표시한 Choi 등 (2009)의 방법은 사전분포의 모수를 할당하는데 있어서 칸 기대빈도의 최대우도추정치를 함께 이용하고자 하였다. 본 연구에서 이용한 사전분포의 모수 할당방법은 Table 2.3에 정리하였다. 여기서 $\nabla_i = q \times y_{++I}/y_{+++}$, $i = 1, 2$이고, 다시 q는 결측모형에 따른 선형예측식 (2.3)의 모수의 갯수이다. 그리고 m_{ijl}은 최대우도추정방법에서 계산된 각 칸 기대빈도의 추정치이다.

마지막으로 베이지안 방법의 적용에서는 E-step에서 결측빈도에 대한 기대값을 계산한 후에 행의 수준에 따라 관찰된 행 주변합 y_{i+2}와 기대빈도의 주변합이, 그리고 전체 빈도의 합과 기대빈도의 합이 일치하도록 다음과 같이 재 계산한 $\tilde{y}_{ijl}$를 최종 기대빈도로 하였다.

$$\tilde{y}_{ij1} = y^*_{ij1} \times \frac{y_{++1}}{y_{++1} + \delta_{++1}},$$
$$\tilde{y}_{ij2} = y^*_{ij2} \times \frac{y_{i+2}}{y_{i+2} + \delta_{i+2}}.$$

3. 자료분석

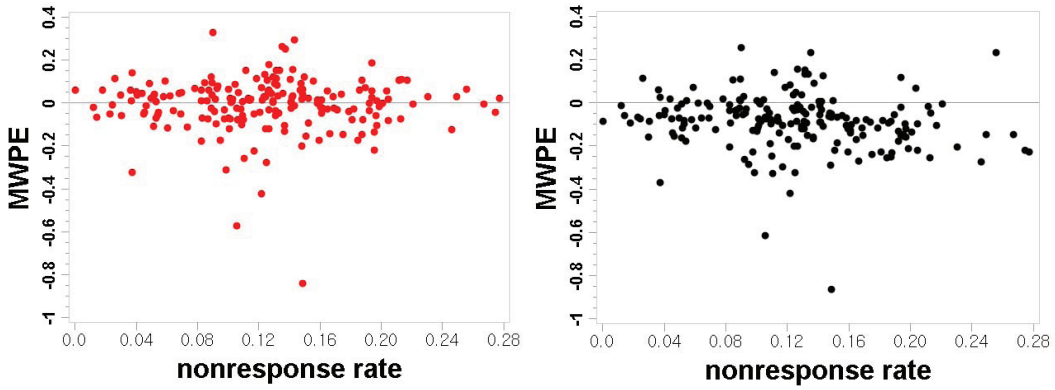
3장에서는 2장에서 소개한 방법을 실제 자료에 적용하여 변방값 문제가 발생한 경우 결측체계의 가정이 따른 모형에 의한 예측의 정확도를 비교해 보고자 한다. 분석에 사용된 자료는 2012년 대한민국의 18대 대통령 선거 당일날 진행된 방송 3가 동시 진행한 출구조사 자료이다. 전체 데이터는 전국에서 선별된 투표소에서 직접 조사된 자료로 성별, 나이, 지지후보(새누리당, 민주통합당, 기타, 무응답)가 조사되었다. Kwak과 Choi (2014)는 같은 자료를 이용하여 분석을 진행하였는데 투표소별로 집계된 자료를 대한민국의 국회의원 선거구별로 재집계하고 그 효과가 거의 없었던 기타 후보의 자료를 제거하고 분석에 이용하였다. 우리나라 전체 국회의원 선거구 가운데 204개의 선거구를 대상으로 하여 분석을 진행하였다. 본 연구에서는 그들의 분석을 확장하고자 하였다. 투표한 후보와 함께 조사된 항목은 성별과 연령이었다. 연령은 20대부터 10세 단위로 범주화 시켰으며 마지막 범주는 60대 이상으로 하였다. 본 분석에서는 성별은 고려하지 않고 연령대 만을 이용하였다. Table 2.1로 대체하여 본다면 X_1은 연령대가 되고 $r = 5$가 된다. X_2는 지지후보가 되고 $c = 2$가 된다. 관찰된 자료에 대하여 식 (2.3)의 3가지 결측모형을 고려할 수 있다. 그러나 최종적으로 본 분석을 통하여 예측하고자 하는 것은 지지후보에 대한 부분이다. 즉 변수 X_2의 주변합으로 예측결과를 정리하는 경우 MAR모형과 MCAR 모형은 같은 결과를 주기 때문에 최종적으로는 MAR모형과 NMAR모형만을 이용하였다. 두 개의 결측모형에 대하여 2장에서 소개한 EM알고리즘 방법을 이용하여 모형 추정을 수행하였으며 무시할 수 없는 결측모형, 식 (2.3)에서의 NMAR 모형의 경우 변방값 문제가 발생할 수 있다. 무시할 수 없는 무응답(혹은 결측) 체계 가정하에서 발생할 수 있는 변방값 문제를 보완하고자 2장에서 언급하였던 베이지안 방법을 적용하고자 하였다.

모형의 정확도를 비교하기 위하여 여러가지 통계량을 이용할 수 있다. 분석에 사용된 자료는 두 명의 후보에 대한 지지율만을 고려 하기로 하였기 때문에 이를 적절히 반영할 수 있는 방법을 이용하고자 하였다. 이를 위하여 Bautista 등 (2007)이 제안하였고 Kwak과 Choi (2014)가 이용하였던 MWPE(modified within precinct error)를 이용하였다. 두 명의 후보에 대한 실제 지지율을 각각 P_1, P_2라 하자. 모형으로부터 추정된 지지율을 각각 $\widehat{P}_1$, $\widehat{P}_2$라 할 때 $\widehat{P}_1$과 $\widehat{P}_2$ 각 칸 빈도에 대하여 모형으로부터 추정된 값을 이용하여 $\widehat{P}_j = \hat{y}_{+j+}/\hat{y}_{+++}$, $j = 1, 2$로 계산될 수 있다. 실제 지지율과 모형으로부터 계산된 예측 지지율을 이용하여 MWPE는 다음과 같이 정의된다.

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

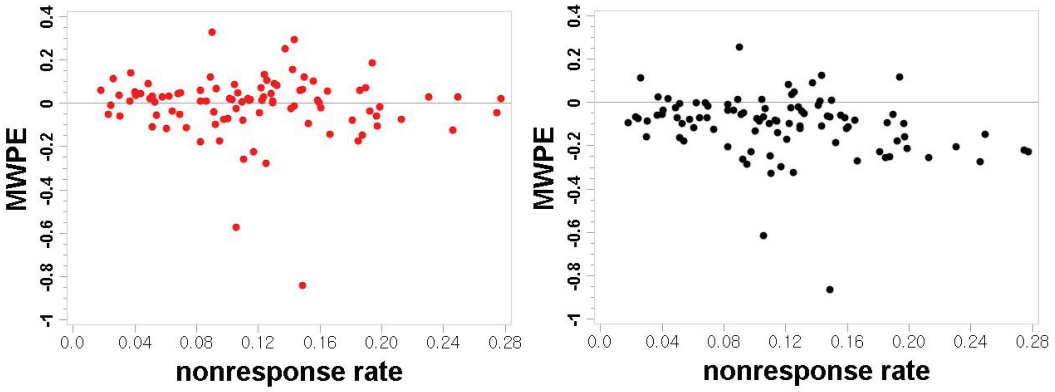
모형으로부터 예측된 지지율의 비율과 실제 지지율의 비율이 비슷해져 감에 따라 α는 1로 접근하게 되며 MWPE는 0으로 접근하게 된다. 즉 MWPE가 0에 가까울 수록 모형에 의한 예측력이 좋다고 할 수 있다.

먼저 무응답률과 MWPE 통계량과의 관계를 보자. Figure 3.1에서 가로축은 무응답률, 세로축은 MWPE 통계량을 나타낸다. 그리고 (a)는 MAR모형에 대한 결과이고 (b)는 NMAR 모형에 대한



(a) MAR (b) NMAR

Figure 3.1. MWPE and nonresponsee rates for MAR and NMAR: Whole cases



(a) MAR (b) NMAR

Figure 3.2. MWPE and nonresponsee rates for MAR and NMAR: Cases with boundary solutions

결과이다. 두 그림 (a)와 (b)에서는 큰 차이점을 찾을 수 없다. 무응답률이 증가함에 따라 예측력이 떨어진다고 볼 수 없다. 다음 Figure 3.2는 최대우도추정에서 변방값 문제가 발생한 경우만을 대상으로 다시 똑같은 그림을 그린 것이다. 이 그림에서 무응답률이 큰 경우 MWPE 값의 절대값이 (a)의 MAR 모형에서 보다 (b)의 NMAR 모형에서 커지는 경향이 있음을 볼 수 있다. 예를 들어 살펴보자. 서울의 노원구을 선거구의 출구조사 결과를 보면 실제 두 후보에 대한 지지율은 박근혜 후보 46.7% 문재인 후보 53.3%로 나타났다. 그러나 출구조사 자료를 이용한 예측 결과를 보면 비임의결측 가정하에서의 예측결과는 박근혜 후보 57.1%, 문재인 후보 42.9%이다. 실제결과와 예측결과가 매우 큰 차이를 보이고 있다. 관찰된 자료 가운데 무응답률은 약 20%에 달하여 상대적으로 높은 무응답률을 보이고 있다. 특히 전체 무응답자 가운데 60대 이상의 비율이 44%로 매우 높은 편이다. 연령 문제를 고려한 다면 임의결측을 가정해 볼 수 있을 것이다. 20%에 달하는 무응답빈도에 대한 대체 결과를 보면 변방값 문제가 발생하여 모든 무응답 빈도를 박근혜 후보로 할당하였다. 이러한 변방값 문제에 의하여 실제결과와 예측결과가 완전히 뒤바뀌어 있음을 확인할 수 있다. 이와같이 변방값 문제가 발생한 경우 2.3절과 2.4절에 설명한 바와 같이 베이지안 방법을 적용하여 그 문제를 해결한 후에 다시 결측체계에 대한 비교 선택

Table 3.1. Comparison result for MAR and NMAR

	MWPE		Total
	MAR(rate: %)	NMAR(rate: %)	
Seoul Metropolitan area	25(65.79)	13(34.21)	38
Yeongnam area	20(66.67)	10(33.33)	30
Honam area	11(68.75)	5(31.25)	16
Chungchong	6(60.00)	4(40.00)	10
ALL	62(65.96)	32(34.04)	94

Table 3.2. Comparison result for MAR and BA1

	MWPE		Total
	MAR(rate)	BA1(rate)	
Seoul Metropolitan area	23(60.53)	15(39.47)	38
Yeongnam area	20(66.67)	10(33.33)	30
Honam area	9(56.25)	7(43.75)	16
Chungchong	6(60.00)	4(40.00)	10
ALL	58(61.70)	36(38.30)	94

문제를 수행할 수 있다. 본 연구에서는 총 204개의 모든 분석단위에 대하여 식 (2.3)의 결측모형에 대한 적합 뿐만 아니라 NMAR모형에 대해서도 두 가지의 베이지안 방법을 모두 적용하여 분석을 시도하였다.

다음 Table 3.1은 204개 전체 선거구별 분석 결과 가운데 변방값 문제가 발생한 선거구 94개에 대하여 최대우도추정 결과를 MWPE 통계량을 기준으로 정리한 것이다. 지역별 편차가 있는지를 확인하기 위하여 수도권(Seoul Metropolitan area), 영남(Yeongnam area), 호남(Honam area), 충청(Chungchong area)로 나누어 정리하였다. 실제 분석에서는 강원 지역과 제주지역또한 분석되었으나 이 두 지역의 선거구별 분석 결과에서는 변방값문제가 발생한 결과가 없었기 때문에 Table 3.1에는 생략 되었다. 총 94개의 선거구 가운데 MWPE를 기준으로 하였을 때 MAR의 적합결과가 더 우수한 경우가 62건으로 비율은 65.96%이다. 지역별로 나누어 보았을 때도 전체 결과와 크게 달라지지 않는다. 전반적으로 보았을 때 MAR모형의 적합이 예측 정확도에서 보다 높은 결과를 보여 주고 있음을 볼 수 있다. 베이지안 추정에 의하면 예측정확도에 있어서 어느 정도 개선이 이루어지는지 알아보자. 이 94개의 선거구에 대한 베이지안 추정 결과를 살펴보자. 먼저 MWPE 값의 변화를 살펴보면 먼저 MAR 모형의 경우 MWPE 값의 평균은 -0.0073이고 표준편차는 0.1470이다. NMAR 모형의 경우 평균이 -0.1030이고 표준편차는 0.1437이다. 표준편차는 거의 비슷하나 평균은 MAR 모형이 더 우수하다. 베이지안 모형의 적용 결과를 보면 BA1으로 표시한 Park (1998)의 방법을 적용한 경우 평균은 -0.0793이고 표준편차는 0.1426이다. 표준편차는 큰 변화가 없으나 최대우도 추정보다 평균이 개선된 것을 볼 수 있다. Choi 등 (2009)의 방법인 BA2의 결과를 보면 평균은 -0.0916이고 표준편차는 0.1409로 역시 최대우도 추정보다 평균이 조금 더 개선되었다.

다음 Table 3.2와 Table 3.3은 변방값 문제가 발생한 경우 베이지안 방법을 이용하여 모형을 재 추정한 후 결과를 정리한 표들이다. 각각 BA1 방법과 BA2 방법을 적용하고 예측된 결과를 바탕으로 MWPE 값을 계산하여 이를 MAR 방법의 결과와 비교한 것이다. Table 3.1 과 비교 하였을 때 비임의결측의 가정에 의한 모형의 추정 방법이 더 정확한 예측 결과를 보인 경우 34%에서 38%로 증가한 것을 볼 수 있다. 그러나 그 차이는 크지 않음을 볼 수 있다. 결론적으로 보았을 때 지난 대통령 선거의 출구조사를 기반으로 분석한 경우 무응답 혹은 결측 체계에 대한 가정에서 임의결측을 가정한 것이 비임의결측을 가

Table 3.3. Comparison result for MAR and BA2

	MWPE		Total
	MAR(rate)	BA2(rate)	
Seoul Metropolitan area	23(60.53)	15(39.47)	38
Yeongnam area	19(63.33)	11(36.67)	30
Honam area	10(62.50)	6(37.50)	16
Chungchong	6(60.00)	4(40.00)	10
ALL	58(61.70)	36(38.30)	94

정한 경우보다 그 정확도가 높다고 할 수 있다. 이는 향후 선거 예측을 수행하는 데 있어서 결측체계에 대한 가정으로 임의결측을 적용하는 것이 보다 적절하다 할 수 있을 것이다.

4. 결론

본 연구에서는 결측 혹은 무응답이 발생한 자료에 대하여 적절한 결측 체계에 대한 가정하에 결측모형을 추정하는 문제를 다루고자 하였다. 이는 일종의 모형 선택 문제로 볼 수가 있다. 그러나 어떠한 결측 체계를 가정하는가에 대한 문제는 지극히 연구자의 주관적인 견해에 의하여 결정될 수 있는 것이며 자료 분석의 측면 보다는 다양한 사회적인 문제들을 우선적으로 고려하여 결정할 수도 있을 것이다. 본 연구에서는 이러한 결측체계에 대한 선택의 문제를 주어진 자료에 기반한 실증적인 분석 결과에 기반하여 다루고자 하였다. 이를 위하여 대한민국의 전체 선거구를 대상으로하여 수집된 자료를 기반으로 하여 분석을 수행하였다. 또한 비임의결측 체계 또는 무시할 수 없는 결측 체계 가정 하에서 최대우도추정을 수행하는 경우 발생할 수 있는 변방값 문제를 해결하기 위하여 베이지안 방법을 적용한 후 실제적인 비교를 수행하고자 하였다. 본 연구에서는 제시되지 않았으나 Kwak과 Choi (2014)는 변방값 문제가 발생하지 않은 경우 임의결측과 비임의결측 간의 모형 비교의 문제를 다루었으며 본 연구에서는 변방값 문제가 발생한 경우 베이지안 방법 적용후에 비교 문제를 다루었다. Kwak과 Choi (2014)의 결과에서도 본 연구의 결과와 비슷하게 임의결측하에서의 예측결과가 비임의결측체계하에서의 결과보도 우수한 것을 볼 수 있었다. 두 결과를 종합하여 보았을 때 대한민국의 선거 예측을 위해서는 비임의결측 체계에 대한 가정 보다는 임의결측 체계의 가정하에서 결측 모형을 추정하는 것이 보다 높은 예측 정확도를 보이고 있다 할 수 있다.

본 연구에서 제안하고 있는 베이지안 방법은 다항분포의 가정에서 각 칸 확률에 사전분포를 할당하고 사전분포의 모수를 관찰된 자료로부터 정보를 얻는 일종의 경험적 베이지안 방법이라 할 수 있다. 이와는 다른 접근 방법으로 로그선형모형의 모수에 직접 사전분포를 할당하는 계층적 베이지안 방법을 적용할 수 있다. 이 방법은 사후분포가 일반적으로 알려진 형태의 분포가 되지 않으며 Markov Chain Monte Carlo 방법을 이용하여야 한다. 상대적으로 더 복잡한 방법이 될 수 있다. 이와 관련된 연구는 추후 연구로 진행될 수 있을 것이다.

References

Agresti, A. (2002). *Categorical Data Analysis*, Second edition, John Wiley & Sons Inc., New Jersey.

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

Baker, S. G., Rosenberger, W. F. and Dersimonian, R. (1992). Closed-form estimates for missing counts in two-way contingency tables, *Statistics in Medicine*, **11**, 643–657.

Bautista, R., Callegaro, M., Vera, J. A. and Abundis, F. (2007). Studying nonresponse in Mexican exit polls, *International Journal of Public Opinion Research*, **19**, 492–503.

Chib, S. (1995). Marginal likelihood from the Gibbs output, *Journal of the American Statistical Association*, **90**, 1313–1321.

Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the Metropolis-Hastings output, *Journal of the American Statistical Association*, **96**, 270–281.

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, G. M. (2012). A model selection method using EM algorithm for missing data, *Journal of the Korean Data Analysis Society*, **14**, 767–779.

Choi, B., Kim, D. Y., Kim, K. W. and Park, Y. S. (2008). Nonignorable nonresponse imputation and rotation group bias estimation on the rotation sample survey, *The Korean Journal of Applied Statistics*, **21**, 361–375.

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.

Clarke, P. S. (2002). On boundary solutions and identifiability in categorical regression with non-ignorable non-response, *Biometrical Journal*, **44**, 701–717.

Dempster, A. P., Laird, N. M. and Rubin, D. M. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society B*, **4**, 1–38.

Forster, J. J. and Smith, P. W. (1998). Model-based inference for categorical survey data subject to non-ignorable non-response, *Journal of the Royal Statistical society, Series B*, **60**, 57–70.

Green, P. E. and Park, T. (2003). A Bayesian hierarchical model for categorical data with nonignorable nonresponse, *Biometrics*, **59**, 886–896.

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

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

Kim, S. Y. and Kwon, S. P. (2009). The effect of survey refusal and noncontact on nonresponse error: For economically active population survey, *The Korean Journal of Applied Statistics*, **22**, 667–676.

Kim, Y. W. and Nam, S. J. (2009). Forming weighting adjustment cells for unit-nonresponse in sample surveys, *Communications for Statistical Applications and Methods*, **16**, 103–113.

Kwak, J. 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.

Pak, G. D. and Shin, K. I. (2010). Non-response imputation for panel data, *Communications for Statistical Applications and Methods*, **17**, 899–907.

Park, J. S., Kang, C., and Kim, K. K. (2013). A simulation study of imputation methods for transportation corporation's survey data, *Journal of the Korean Data Analysis Society*, **15**, 1903–1912.

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

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

Park, T. S. and Lee, S. Y. (1998). Analysis of categorical data with nonresponses, *The Korean Journal of Applied Statistics*, **11**, 83–95.

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

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.

Rubin, D. B., Stern, H. S. and Vehovar, V. (1995). Handling "Don't know" survey responses: The case of the Slovenian Plebiscite, *Journal of the American Statistical Association*, **90**, 822–828, nonresponse, *Journal of Applied Statistics*, **37**, 1439–1453.

Song, J. (2011). Selection of variables to form imputation classes in Hotdeck imputation, *Journal of the Korean Data Analysis Society*, **13**, 1321–1329.

Song, J. (2014). A comparison of imputation methods for multiple response questions, *Journal of the Korean Data Analysis Society*, **16**, 691–701.

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.

경험적 베이지안 방법을 이용한 결측자료 연구

윤용화[a] · 최보승[a,1]

[a]대구대학교 전산통계학과

(2014년 8월 14일 접수, 2014년 10월 13일 수정, 2014년 10월 20일 채택)

요 약

조사를 통하여 수집된 자료에 기반하여 분석을 수행하는데 있어서 결측값에 대한 적절한 대체 방법은 보다 정확한 결과를 얻기 위한 매우 중요한 절차이다. 본 연구에서는 모형에 기반하여 결측자료에 대한 대체방법과 모형 추정방법을 다루었다. 특히 최대우도추정 방법의 적용에서 발생할 수 있는 변방값 문제(bounday soluntion problem)를 해결하기 위하여 베이지안 방법을 적용하였다. 분석된 결과를 바탕으로 하여 예측을 수행한 후 결측체계에 따른 정확성 비교를 수행하여 결측체계에 따른 결측모형의 선택 문제를 다루었다. 예측의 정확도를 측정하기 위하여 Bautista 등 (2007)이 제안한 MWPE(modified within precinct error) 이용하여 비교를 수행 하였다. 본 연구에서 제시된 방법들은 2012년에 시행된 제 18대 대통령 선거 당일 시행된 출구조사의 자료를 적용하여 분석을 수행하였다. 분석 결과 임의결측체계의 가정에 따른 결과가 비임의체계 가정에 따른 결과보다 예측의 정확도가 더 높았다.

주요용어: 결측자료, 무응담, EM 알고리즘, 경험적 베이지안 방법.

본 연구는 대구대학교 교내연구비로 수행된 연구임(과제번호 21020499).

[1]교신저자: (712-714) 경북 경산시 진량읍 대구대로 201 대구대학교 전산통계학과. E-mail: bchoi@daegu.ac.kr