Journal of the Korean Data & Information Science Society
2012, **23**(4), 627–641

http://dx.doi.org/10.7465/jkdi.2012.23.4.627
한국데이터정보과학회지

무응답을 가지고 있는 범주형 자료에 대한 모형 선택 방법[†]

윤용화[1] · 최보승[2]

[12]대구대학교 전산통계학과

접수 2012년 4월 23일, 수정 2012년 6월 5일, 게재확정 2012년 6월 15일

요약

본 연구는 다차원 분할표 형태로 정리된 범주형 자료가 결측치나 무응답을 가지고 있을 때 주어진 자료를 가장 잘 설명하고 예측의 정확도를 높일 수 있는 모형의 추정과 모형의 선택 문제를 다루었다. 무시할 수 없는 무응답 (non-ignorable non-response)체계하에서 최대우도 추정에서 발생할 수 있는 변방값 문제를 해결하기 위하여 계층적 베이지안 모형을 고려하였다. 또한 모형 적합도를 높이기 위한 변수 조합을 찾는 모형 선택의 문제를 함께 다루었다. 베이지안 접근하에서 모형 선택의 문제를 다루기 위하여 베이즈 인자 (Bayes factor)를 모형 선택의 기준으로 이용하였다. 제시된 방법은 2004년 실시된 우리나라 국회의원 선거를 앞두고 수행된 여론조사 데이터를 이용하여 실증분석을 수행하였다. 분석결과 무시할 수 없는 무응답 체계하에서 설명변수로 투표참여여부를 이용하는 것이 가장 적합한 모형으로 판명되었다.

주요용어: 모형선택, 무응답, 베이즈 인자, 선거, 마코프체인 몬테카를로.

1. 서론

2012년 대한민국은 국회의원 선거와 대통령 선거를 동시에 치루게 된다. 이들 선거는 과거 어느 선거보다도 치열하고 예측하기 어려운 상황으로 전개되고 있으며 2012년 대한민국에서는 선거를 앞두고 다양한 여론조사의 결과들이 끝임없이 생산되고 있다. 최근에는 기존의 집전화나 휴대폰 번호를 이용한 조사와 더불어 임의전화걸기방식 (RDD; Random Digit Dialing)방법이 도입되어 전화번호부에 등재되어 있지 않은 전화번호까지 포함하여 여론조사를 진행하고 있다. 이는 기존처럼 전화번호부에 등재된 번호만을 대상으로 여론조사를 시행하는 경우 편향이 발생할 수 있고 잘못된 예측을 하는 경우가 빈번하게 발생하기 때문이다. 새로운 전화조사 방법의 도입은 궁극적으로 보다 정확한 예측을 위하여 오류를 줄이고자 하는 노력의 일부라 할 수 있다.

일반적으로 여론조사를 수행하게 될 때는 단순히 지지하는 후보에 대한 질문 뿐 만 아니라 인구통계학적 변인을 포함한 여러 질문을 추가적으로 하게 된다. 추가적인 문항을 이용함으로써 세대간, 지역간 혹은 개개인의 정치 성향에 대한 후보의 지지도를 파악하고자 하는 의도 뿐 만 아니라 이들 정보를 이용하여 후보를 결정하지 않은 부동층의 향방을 예측하고자 하는데 이용하기 위해서이다. 추가적으로 수집되는 정보들은 인구통계적 측면에 따른 지지층의 분석 뿐만 아니라 여론조사에서 빈번하게 발생하는 무응답 또는 응답거부자 또는 미결정자의 지지후보 예측을 수행하고자 하는데

† 본 연구는 대구대학교 교내연구비로 수행된 연구임(과제번호 20110207).
[1] (712-714) 경상북도 경산시 진량읍 대구대로 201, 대구대학교 전산통계학과, 교수.
[2] 교신저자: (712-714) 경상북도 경산시 진량읍 대구대로 201, 대구대학교 전산통계학과, 조교수.
E-mail: bchoi@daegu.ac.kr

있어서도 중요한 요인으로 작용할 수 있다. 대부분의 여론조사 기관에서는 다양한 통계적 모형을 이용하여 무응답이나 응답거부자에 대한 예측을 수행하여 보다 정확한 선거예측을 하고자 노력하고 있다.

이렇듯 전화여론조사에서 수집되는 정보들은 단순히 지지후보에 대한 문제 뿐 만 아니라 추가적인 다양한 정보들이 취합된다. 그리고 그 정보들은 전화조사라는 특성상 범주형자료의 형태가 된다. 따라서 수집된 정보들은 대부분 분할표 형태로 정리된다. 분할표 형태로 정리된 자료를 분석하는데 있어서 두개의 변수만을 고려한 경우 2차원 분할표 형태로 자료를 정리하고 카이제곱 검정이나 우도비 검정등을 통하여 두 범주형 변수간의 연관성을 측정할 수 있다. 그러나 분할표를 구성하는 변수의 수가 증가하게 되면 카이제곱 검정이나 우도비 검정은 한계를 가지게 되고 일반적으로 로그선형모형을 이용하여 분석을 수행하는 경우가 대부분이다. 3차 이상의 고차원 분할표로 정리된 자료에 추가적으로 무응답에 대한 문제까지 고려하게 된다면 분할표의 차원은 더욱 증가하게 된다. 예를 들어 2차원 형태로 정리된 분할표에 자료에 대하여 분할표를 구성하는 두 변수에 모두 무응답이 발생하게 된다면 무응답에 의해 처리된 이후 확장된 분할표는 4차원 분할표의 형태가 된다 (Choi 등, 2009).

본 연구는 이와 같이 다차원 분할표의 형태로 정리될 수 있는 범주형 자료에 대하여 무응답을 가지고 있는 자료에 대한 처리 문제를 다루고자 한다. 특히 선거여론조사와 같이 분할표를 구성하는 변수 가운데 하나인 지지후보를 반응변수로 하고 함께 조사되어 분할표에 정리된 다른 변수들은 설명변수로 고려하고자 한다. 그리고 지지후보를 나타내는 반응변수에 대한 조사과정에서 발생할 수 있는 무응답이나 응답거부를 결측치로 고려하여 이에 대한 적절한 대체문제를 고려하고자 한다. 또한 궁극적으로 가장 적절한 예측을 수행하기 위한 예측모형을 구현하는 문제를 고려해 보기로 한다. 즉 본 연구의 목적은 다차원 분할표 형태로 정리된 범주형 자료가 결측치나 무응답을 가지고 있을 때 주어진 자료를 가장 잘 설명하고 예측의 정확도를 높일 수 있는 최적의 모형을 선택하는 모형 선택의 문제를 다루는 것이다.

일반적으로 조사 과정에서 발생하는 무응답은 그 발생체계에 따라 크게 3가지로 구분할 수 있다. Little과 Rubin (2002)의 구분에 따라 무응답은 완전임의결측 (MCAR; missing completely at random), 임의결측 (MAR; missing at random), 비임의결측 (MNAR; missing not at random)으로 구분할 수 있다. 먼저 완전임의결측은 무응답의 발생여부가 무응답을 발생한 변수나 혹은 자료수집과정에서 함께 수집된 그 어떤 변수에 의해서도 영향받지 않은 경우이다. 두 번째 임의결측은 무응답의 발생 여부가 무응답이 발생하지 않은 다른 변수에 의해서 영향받는 경우이다. 마지막 비임의결측은 무응답의 발생여부가 무응답이 발생한 변수 자신에 의하여 영향을 받는 경우로 무시할 수 없는 무응답 (non-ignorable non-response)라 부른다. 관찰된 자료에서 무응답이 발생하는 경우 무응답이 이 세가지의 무응답 체계 가운데 어떠한 것을 따르는가 하는 문제는 다양한 방법을 통하여 검증될 수 있다. 특히 선거와 같이 민감한 주제를 가지고 시행되는 여론조사의 경우 무응답이 무시할 수 없는 무응답을 따르는 경향이 강하다 (Chen과 Stasny, 2003; Rubin 등, 1995). 즉 지지하는 후보를 밝히지 않았을 때 그 이유가 바로 본인이 지지하는 후보가 열세에 놓여 있는 후보라면 이와 같은 상황이 무시할 수 없는 무응답이라 할 수 있다.

무응답을 포함하는 범주형 자료의 분석은 일반적으로 EM 알고리즘 (Dempster 등, 1977)을 이용한 추정 방법이 주로 이용된다. Baker와 Laird (1988)과 Baker 등 (1992)는 로그선형모형에서 모수의 최대우도 추정치를 구하고자 EM 알고리즘을 이용한 추정 방법을 제안하였다. 그러나 무시할 수 없는 무응답 체계 가정하에서 최대우도추정치를 구하게 되면 변방값 (boundary solution) 문제가 발생하게 된다. 변방값 문제란 분할표상에서 추정된 무응답 빈도에 대한 확률이 특정 칸에서 0의 값을 가지는 현상을 말한다 (Choi 등, 2007). 변방값 문제가 발생하게 되면 최대우도 추정치가 유일한 해를 가지지

않게 되고 그 결과가 불안정해 질 수 있다 (Park과 Brown, 1994). 그리고 자유도가 0인 포화모형이라 하더라도 우도비 통계량 값이나 카이제곱 통계량 값이 0보다 큰 문제가 발생할 수 있다.

이와 같은 무응답 추정 과정에서 발생하는 변방값 문제를 해결 하기 위하여 여러가지 방법들이 제안되어 왔다. Park과 Brown (1994), Park (1998), Choi 등 (2009), Park과 Choi (2010) 등은 로그선형모형에서 랜덤성분에 다항분포를 가정하고 각 칸의 기대확률에 사전분포를 할당하는 경험적 베이지안 방법을 제안하였다. 랜덤성분으로 다항분포를 가정하였기 때문에 사전분포로는 기대확률에 켤레분포 관계에 있는 디리클레 (Dirichlet) 분포를 이용하였다. Choi (2007)과 Choi 등 (2008)은 유사한 무응답 추정문제를 다른 측면의 자료에 적용하였다. Choi (2007)은 무응답 추정 문제를 금융권 데이터에 적용하여 은행고객의 세분화 문제에 적용하였으며 Choi 등 (2008)은 교체표본조사에서 교체그룹의 편향을 추정하는데 있어서 조사과정에서 발생하는 무응답을 대체하는데 적용하였다. 이들에 의해서 제시된 방법들은 사전분포의 초모를 할당하는데 있어서 관찰된 자료에 의존하게하는 경험적인 베이지안 방법이라 할 수 있다.

이들과는 다른 접근방법으로 범주형 자료의 모형 추정에서 계층적 베이지안 방법을 이용하는 방법들도 제안되어 왔다. Cargnoni 등 (1997)은 범주형 자료를 반응변수로 하는 포아송 회귀모형에서 회귀계수에 사전분포를 할당하는 방법을 제안하였고 또한 Green과 Park (2003)도 분할표자료에 대한 로그선형모형에서 체계적 성분의 모수에 사전분포를 할당하는 계층적 베이지안 방법을 제안하였다. 이 때 두 방법간의 차이는 반응변수인 관찰빈도의 로그 기대값과 체계적 성분을 연결시키는데 있어서 전자는 연결함수를 이용하여 직접 연결하였고 후자는 잠재변수를 이용하여 잠재변수에 사전분포를 할당하는 방법으로 연결하였다.

무응답 혹은 결측을 가지는 범주형 자료에 대한 또 다른 연구로 Hong과 Jung (2011a)과 Hong과 Jung (2011b)는 이항반응 자료에 대하여 이 때 발생하는 결측의 문제를 해결하기 위하여 로지스틱 회귀모형 (Hong과 Jung, 2011a)와 이변량 프로빗 모형 (Hong과 Jung, 2011b)를 이용하였다. 또한 Chun 등 (2007)은 설문지자료에서 발생하는 결측의 문제를 해결하기 위하여 전체 자료를 크기순으로 재배열한 후 평균 대체와 중위수 대체등의 방법을 통하여 무응답을 처리하는 연구를 진행하였으며 Chung과 Han (2009)는 판별분석 모형에서 결측치가 블록의 형태로 발생할 때 붓스트랩 (bootstrap)방법을 이용하여 오차율에 대한 추정문제를 다루었다.

본 연구에서는 Green과 Park (2003)이 제안한 방법과 유사한 방법을 이용하여 무응답을 포함하는 범주형 자료에 대하여 무응답의 대체와 모수의 추정을 함께 수행하는 계층적 베이지안 모형을 이용한 모형의 추정문제를 다루고자 하였다. 조건부 사후분포로부터 모수를 반복적으로 추출하는 Markov Chain Monte Carlo (MCMC) 표본추출방법을 이용하여 모수에 대한 추정치를 구하고자 하였다. 여기서 우리는 두 가지 추가적인 사항을 고려하였다. 첫 번째는 무시할 수 없는 무응답 가정하에서 발생할 수 있는 변방값 문제의 발생이다. 그러나 이 문제는 사전분포를 할당하는 베이지안 방법에 의하여 해결될 수 있다. 두 번째로 고려하고자 하는 문제는 적절한 모형의 선택이다. 본 연구에서 이용하고자 하는 자료는 우리나라 국회의원 선거를 앞두고 실시된 사전 여론조사의 결과로서 반응변수인 후보지지도 뿐 만 아니라 추가적인 변수들이 함께 조사되었다. 고려된 모든 변수를 이용하여 분할표를 작성하게 되면 그 차원이 기하급수적으로 증가하게 되고 빈도가 0인 범주가 함께 증가하게 된다. 따라서 카이제곱 분포를 이용한 근사적 검정에 문제가 발생하게 된다. 따라서 모형 단순화의 측면에서 적절한 변수의 선택이 이루어 져야 한다. 그러나 결측치를 가지고 있는 범주형 자료에 대한 분석에서 우도함수에 근거한 고전적 선형모형하에서 행해지는 모형 선택의 방법에는 몇가지 문제가 따르게 된다. 전술한 바와 같이 변수의 수가 증가함에 따라 고차원의 분할표가 구성되며 이에 따라 빈도가 0인 범주가 급격하게 증가한다. 대표본에 근거한 추론에 문제가 발생하게 된다. 두 번째로는 무응답 대체에 따른 변방값 문제의 발생이다. 변방값 문제가 발생하게 되면 모형의 추정이 불안정하게 되며

포화모형임에도 우도비 통계량이 0보다 큰 문제가 발생하게 된다.

본 연구에서는 이와 같은 문제를 해결하기 위하여 계층적 베이지안 모형하에서 모형 선택을 위한 방법으로 베이즈 인자 (Bayes factor)를 이용한 모형 선택방법을 이용하였다. 베이즈 인자를 이용하여 모형 선택을 수행하게 되면 기본적으로 베이지안 접근방법으로 모수에 대한 추정을 수행하기 때문에 변방값 문제를 해결 할 수 있으며 계층적 구조에 놓여 있지 않은 모형들 간에도 모형의 비교를 수행할 수 있는 장점이 있다 (Kass와 Raftery, 1995). Kass와 Raftery (1995)는 베이즈 인자에 대한 자세한 정의 뿐만 아니라 관찰 데이터에 대한 주변 분포를 추정하는 다양한 방법을 제시하였다. 그러나 계층적 베이지안 모형에서 데이터에 대한 주변 분포를 추정하는 문제는 간단히 해결되는 문제가 아니다. 본 연구에서는 주변 분포의 계산을 위하여 Chib (1995)와 Chib과 Jeliazkov (2001)이 제안한 방법을 이용하였다.

본 연구의 이후 진행 과정은 다음과 같다. 다음 2절에서는 본 연구에서 제안하고 있는 계층적 베이지안 모형에 대한 모형 추정방법과 모형 선택을 위한 베이즈 인자의 계산방법을 소개한다. 다음 3절에서는 우리나라 선거 여론조사를 이용한 실증 분석의 결과를 제시한다. 마지막 4절에서는 결론으로 본 연구에서 제시하는 방법의 장단점과 한계에 대하여 논하고자 한다.

2. 계층적 베이지안 방법을 이용한 모형 추정 방법

2.1. 계층적 베이지안 모형

관찰된 모든 자료가 범주형인 자료를 고려하여 보자. 다차원 분할표를 구성하는 변수가운데 하나를 Y라 하고 Y는 J개의 범주로 구성되어 있다. 즉 $Y = j$, $j = 1, \ldots, J$가 된다. 그리고 이 변수만이 무응답 혹은 결측치가 발생한다고 가정한다. 여기서 이 변수의 결측여부를 나타내는 지시변수를 R이라 한다. 즉 $R = l$, $l = 1, 2$가 되며 결측이 발생하지 않으면 $l = 1$이 되고 결측이 발생한 경우 $l = 2$가 된다. 이제 다차원 분할표를 구성하는 나머지 변수들은 총 p개의 변수가 존재하고 이를 X=$\{X_1, X_2, \ldots, X_p\}$ 로 나타내며 각 X들도 모두 범주형 변수이고 각 변수들의 범주들은 첨자 $\{i_1, i_2, \ldots, i_p\}$로 구분되며 $\{I_1, I_2, \ldots, I_p\}$의 범주로 구성되어 있다. 변수 X들에 대해서는 결측이나 무응답이 발생하지 않는다고 가정한다. 변수 X들은 비록 p개의 범주로 구성되어 있지만 변수들의 모든 조합을 고려하여 총 $I = I_1 \times I_2 \times \cdots I_p$ 개의 범주를 고려하면 하나의 차원의 변수로 변형할 수 있으며 결국 Y와 X간의 2차원 분할표로 정리할 수 있다. 예를 들어 $I = 2$이고 $J = 2$인 분할표로 정리하면 다음 Table 2.1과 같다.

Table 2.1 Two-way contingency table with marginal sum

	$R = 1$		$R = 2$
	$Y = 1$	$Y = 2$	
$X = 1$	y_{111}	y_{121}	y_{1+2}
$X = 2$	y_{211}	y_{221}	y_{2+2}

여기서 y_{ijl}, $i = 1, 2, \ldots, I$, $j = 1, 2, \ldots, J$, $l = 1, 2$는 관찰된 칸 기대빈도를 나타내고 이 관찰된 빈도에 대하여 다항분포를 가정하면 우도함수는 다음의 식에 비례한다.

$$L \propto \prod_i \prod_j \pi_{ij|1}^{y_{ij1}} \prod_i \pi_{i+|2}^{y_{i+2}}. \tag{2.1}$$

여기서 $\pi_{ij|1} = Pr(X = i, Y = j|R = 1)$, $\pi_{i+|2} = Pr(X = i|R = 2)$이고 $N_1 = \sum_i \sum_j y_{ij1}$, $N_2 = \sum_i y_{i+2}$, $N = N_1 + N_2$으로 고정된 값으로 고려한다. 이제 각 다차원 분할표의 각 칸의 기

대빈도를 $\mu_{ij1} = N_1 \times \pi_{ij|1}$, $\mu_{ij2} = N_2 \times \pi_{ij|2}$라 하고 이 기대빈도에 대한 로그선형모형은 다음과 같다.

$$\log \boldsymbol{\mu} = \boldsymbol{Z}\boldsymbol{\beta}. \tag{2.2}$$

여기서 $\boldsymbol{Z}$는 계획행렬로 일반적인 실험계획법에서 계획행렬를 정의 하는 방법에 따라 설정할 수 있다. 본 연구에서는 각 변수에서 그 합이 0이 되도록 지정하였다. $\boldsymbol{\beta}$는 로그선형모형의 체계적 성분에 대한 모수 벡터가 된다. 만약 체계적 성분에 반응변수 Y와 결측지시변수 R 과의 상호작용 효과가 포함되면 무응답 체계는 비임의결측 (NMAR), 즉 무시할 수 없는 무응답이 되고 설명변수 X들과 결측지시변수 R간의 상호작용 효과만이 포함되면 무응답 체계는 임의결측 (MAR)이 된다. 만약 X들이나 Y와 R간의 어떠한 상호작용 효과도 모형이 포함되어 있지 않으면 이 때 무응답 체계는 완전임의결측 (MCAR)이 된다. 따라서 다양한 체계적 성분의 정의에 따라서 무응답 체계에 대한 지정을 할 수 있으며 모형에 포함하고자 하는 변수들의 형태도 지정할 수 있다. 이제 식 (2.1)과 (2.2)에 의하여 우도함수는 다음과 같이 표시될 수 있다.

$$L \propto \prod_i \prod_j \left(\frac{\exp(\boldsymbol{z}'_{ij1}\boldsymbol{\beta})}{\sum_i \sum_j \exp(\boldsymbol{z}'_{ij1}\boldsymbol{\beta})} \right)^{y_{ij1}} \prod_i \left(\frac{\sum_j \exp(\boldsymbol{z}'_{ij2}\boldsymbol{\beta})}{\sum_i \sum_j \exp(\boldsymbol{z}'_{ij2}\boldsymbol{\beta})} \right)^{y_{i+2}}. \tag{2.3}$$

본 연구에서는 계층적 베이지안 모형을 위하여 모수인 $\boldsymbol{\beta}$의 사전분포를 고려하기 이전에 보다 안정적으로 모수를 추출하기 위하여 Green과 Park (2003) 그리고 Chib (1995)가 제안한 방법에 따라 잠재변수를 추가적으로 고려하여 모수를 추출하는 방법을 이용하였다. Gelfand와 Smith (1990)은 이와 같은 기법을 Rao-Blackwellization이라 칭하였다. 분할표의 각 칸 기대빈도를 μ_{ijl}이라 할 때 기대빈도에 로그를 취한 값을 $\eta_{ijl} = \log \mu_{ijl}$이라 하고 이 잠재변수 η_{ijl}에 대한 사전분포는 Green과 Park (2003) 그리고 Chib (1995)의 제안에 따라 다음과 같은 정규분포로 고려한다.

$$\eta_{ijl}|\boldsymbol{\beta}, \sigma_l^2 \sim \ N\big(\boldsymbol{z}'_{ijl}\boldsymbol{\beta}, \sigma_l^2\big) \tag{2.4}$$

계층적 베이지안 모형을 위하여 모수인 $\boldsymbol{\beta}$에 다음과 같은 다변량 정규분포를 사전분포로 고려한다.

$$\boldsymbol{\beta} \propto N_q(\boldsymbol{B}_0, \Psi_{\boldsymbol{\beta}}). \tag{2.5}$$

여기서 q는 모수의 수를 나타낸다. η_{ijl}에 대한 분포를 정규분포로 할당함으로써 추가적인 모수 σ_1^2, σ_2^2가 발생한다. 이에 이 두 모수에 대한 사전분포로서 공액관계에 있는 역감마 (inverse gamma)분포를 사전분포로 할당한다.

$$\begin{aligned} \sigma_1^2 &\sim InverseGamma(a_1/2, a_1 b_1/2), \\ \sigma_2^2 &\sim InverseGamma(a_2/2, a_2 b_2/2). \end{aligned} \tag{2.6}$$

이제 y_{ij2}를 관찰되지 않은 무응답 빈도를 나타낸다고 하자. 따라서 로그선형모형 (2.5)로 부터 모수를 추정하기 위해서는 관찰되지 않은 빈도 y_{ij2}에 대한 적절한 대체가 수행되어야 한다. 전체적인 베이지안 체계하에서는 이 관찰되지 않은 빈도 y_{ij2}도 모수라 할 수 있다. 따라서 적절한 사전분포를 고려한 후 사후분포를 지정하여야 한다. 사전분포로는 상수에 비례하는 무정보적 (noninformative) 사전분포를 할당하고 관찰빈도에 대하여 다항분포를 가정하였기 때문에 사후분포로 다음과 같은 다항분포를 고려할 수 있다 (Green과 Park, 2003).

$$y_{ij2}|\boldsymbol{\beta}, y_{i+2} \sim \ Multi\big(y_{i+2}, \pi_{ij|2}\big). \tag{2.7}$$

여기서 $\pi_{ij|2} = \dfrac{\pi_{ij2}}{\pi_{i+2}} = \dfrac{\exp(\boldsymbol{z}'_{ij2}\boldsymbol{\beta})}{\sum_j \exp(\boldsymbol{z}'_{ij2}\boldsymbol{\beta})}$ 이다.

2.2. MCMC 과정을 통한 모수 추출

이제 조건부 사후분포로부터 모수를 추출하는 과정을 알아보자. 먼저 추출하고자 하는 모수에 대한 초기치를 설정하여야 한다. 첫 번째로 무응답 빈도에 대한 초기치는 식 (2.7)으로부터 무응답 빈도에 대한 기대확률을 관찰된 자료만을 이용하여 계산하였다. 예를 들면 y_{112}의 경우 $y^*_{112} = y_{1+2} \times (y_{111}/y_{1+1})$을 계산하여 초기치로 할당하였다. 다음으로 $\boldsymbol{\beta}$에 대한 초기치를 지정한다. 먼저 실제 관찰치 $\{y_{ij1}\}$과 결측에 대한 대체값 초기치 $\{y^*_{ij2}\}$를 모두 이용하여 일반적인 로그선형모형에서 모수를 추정하기 위한 방법 (Agresti, 2002, p.146)인 반복 재가중 최소제곱 (iterative reweighted least square) 방법을 이용하여 모수에 최대우도 추정치를 구하여 이를 초기치로 이용하였다. 마지막으로 σ^2_l. $l = 1, 2$의 초기치는 $\{y_{ij1}\}$, $\{y^*_{ij2}\}$, 그리고 $\boldsymbol{\beta}$의 초기치를 이용하여 계산하였다. 다음으로 각 모수에 대한 조건부 사후분포를 알아보자.

먼저 y_{ij2}의 경우 다항분포 (2.7)으로부터 추출한다. 다음으로 $\boldsymbol{\beta}$의 조건부 사후분포는 사전분포 (2.5)와 (2.4)를 이용하여 다음의 다변량 정규본포로부터 표본을 추출한다.

$$\boldsymbol{\beta} \propto MVN_q\Big(\boldsymbol{V}(Z^T\Sigma^{-1}\boldsymbol{\eta} + \Psi_{\boldsymbol{\beta}}^{-1}\boldsymbol{B}), \boldsymbol{V}\Big). \tag{2.8}$$

여기서 $\boldsymbol{V} = (Z^T\Sigma^{-1}Z + \Psi_{\boldsymbol{\beta}}^{-1})^{-1}$로 주어지며 $\Sigma = diag(\sigma_1^2 \times I_{I\times J}, \sigma_2^2 \times I_{I\times J})$이다. 이 때 $diag(*)$는 $*$를 대각원소로 하는 대각 행렬을 나타내며 $I_{I\times J}$는 크기가 $I \times J$인 단위행렬을 나타낸다.

다음으로 σ_1^2, σ_2^2에 대한 조건부 사후분포는 다음과 같은 역감마 분포를 가진다.

$$\begin{aligned}
\sigma_1^2 &\sim InverseGamma\Big(\tfrac{I\times J+a_1}{2}, \tfrac{a_1b_1+(\boldsymbol{\eta}_1-Z_1\boldsymbol{\beta})^T(\boldsymbol{\eta}_1-Z_1\boldsymbol{\beta})}{2}\Big),\\
\sigma_2^2 &\sim InverseGamma\Big(\tfrac{I\times J+a_2}{2}, \tfrac{a_2b_2+(\boldsymbol{\eta}_2-Z_2\boldsymbol{\beta})^T(\boldsymbol{\eta}_2-Z_2\boldsymbol{\beta})}{2}\Big).
\end{aligned} \tag{2.9}$$

여기서 $\boldsymbol{\eta}_1$과 Z_1는 관찰된 빈도에 대한 잠재변수와 계획행렬이고 $\boldsymbol{\eta}_2$과 Z_2는 무응답 대체에 의해 생성된 의사관찰빈도에 대한 잠재변수와 계획행렬을 나타낸다.

다음으로 η_{ijl}에 대한 조건부 사후분포를 알아보자. $\{\eta_{ijl}\}$에 대한 분포만을 고려한다면 다항분포로부터의 표본추출은 포아송분포로부터의 표본추출과 동일하므로 (Agresti, 2002, p.340; Green과 Park, 2003), 조건부 사후분포는 다음의 식에 비례한다.

$$\eta_{ijl} \propto \exp\left(y_{ijl}\eta_{ijl} - \exp(\eta_{ijl}) - \frac{(\eta_{ijl} - \boldsymbol{z}'_{ijl}\boldsymbol{\beta})^2}{2\sigma_l^2}\right) \tag{2.10}$$

그러나 η_{ijl}에 대한 조건부 사후분포 (2.10)은 일반적으로 알려진 형태의 분포가 아니므로 메트로폴리스-해스팅스 (Metropolis-Hastings) 알고리즘을 이용하여 표본을 추출할 수 있다. η_{ijl}를 추출하기 위한 알고리즘은 다음과 같다.

알고리즘 2.1 η_{ijl} 추출을 위한 알고리즘

(a) 다음의 제안분포 (proposal distribution)로부터 η^m_{ijl}를 추출한다.

$$J(\eta^m_{ijl}|\eta^{m-1}_{ijl}) = f(\eta^m_{ijl}|\hat{\eta}_{ijl}, c^2V_\eta, \nu_\eta). \tag{2.11}$$

여기서 f 는 자유도가 ν_η, 위치모수가 $\hat{\eta}_{ijl}$, 그리고 척도모수가 V_η인 비중심 t 분포를 나타낸다. 비중심 t분포를 제안분포로 고려함으로써 메트로폴리스-해스팅스 알고리즘의 효율성을 고려하는데 c^2 뿐만아니라 자유도인 ν_η를 함께 이용할 수 있는 장점이 있다.

(b) 다음식으로부터 비 $\alpha(\eta_{ijl}^{m-1}, \eta_{ijl}^{m})$를 계산한다.

$$\alpha(\eta_{ijl}^{m-1}, \eta_{ijl}^{m}) = \frac{p[\eta_{ijl}^{m}|\{y_{ijl}\}, \boldsymbol{\beta}, \sigma_l^2]/f(\eta_{ijl}^{m}|\hat{\eta}_{ijl}, c^2V_\eta, \nu_\eta)}{p[\eta_{ijl}^{m-1}|\{y_{ijl}\}, \boldsymbol{\beta}, \sigma_l^2]/f(\eta_{ijl}^{m-1}|\hat{\eta}_{ijl}, c^2V_\eta, \nu_\eta)}$$

여기서 $p[\eta_{ijl}^{m}|\{y_{ijl}\}, \boldsymbol{\beta}, \sigma_l^2]$ 은 식 (2.10)이 된다.

(c) 다음과 같은 확률을 가지고 새로운 η_{ijl}^{m}를 결정한다.

$$\eta_{ijl}^{m} = \begin{cases} \eta_{ijl}^{m} & \min(\alpha,1) \geq \text{uniform(0,1)} \text{ 일 때} \\ \eta_{ijl}^{m-1} & \min(\alpha,1) < \text{uniform(0,1)} \text{ 일 때} \end{cases}$$

2.3. 베이즈 인자의 계산과정

2.1절과 2.2절에서 제시된 조건부 사후분포로부터 각 모수들을 반복적으로 추출한 후 사후평균을 계산하여 모수들에 대한 최종 추정치를 계산할 수 있다. 이 과정에서 로그선형모형 (2.2)의 지정에 따라 다양한 후보모형을 지정할 수 있다. 본 절에서는 여러 후보모형 가운데 베이지인자를 이용하여 최적의 모형을 선택하는 과정을 설명한다.

서로 다른 두 후보 모형 M_1과 M_2에 대한 베이스 인자는 다음과 같이 정의된다.

$$\frac{p(\{y_{ijl}\}|M_2)}{p(\{y_{ijl}\}|M_1)} = \frac{\int p(\boldsymbol{\theta}_2|M_2)p(\{y_{ijl}\}|\boldsymbol{\theta}_2, M_2)d\boldsymbol{\theta}_2}{\int p(\boldsymbol{\theta}_1|M_1)p(\{y_{ijl}\}|\boldsymbol{\theta}_1, M_1)d\boldsymbol{\theta}_1}.$$

여기서 $\boldsymbol{\theta}_1$과 $\boldsymbol{\theta}_2$는 각 모형하에서의 전체 모수벡터를 나타낸다. 결과적으로 베이즈인자를 구하기 위해서는 각 모형하에서 주변우도 함수인 $p(\{y_{ijl}\}|M_2)$ 와 $p(\{y_{ijl}\}|M_1)$을 계산하여야 한다. 본 연구에서는 Chib (1995)와 Chib과 Jeliazkov (2001)에 의해 제안된 MCMC 표본추출 결과를 이용한 주변 우도함수 계산 방법을 응용하여 이용하였다. 먼저 적절하게 선택된 모수의 추정치 $\boldsymbol{\theta}^*$에 대하여 관찰된 전체 데이터 $\boldsymbol{y} = \{\{y_{ij1}\}, \{y_{i+2}\}\}$에 대한 주변 우도함수는 다음과 같이 정의될 수 있다.

$$p(\boldsymbol{y}) = \frac{L(\boldsymbol{y}|\boldsymbol{\theta}^*)\pi(\boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta}^*|\boldsymbol{y})}.$$

그리고 로그변환된 주변 우도함수의 추정치는 다음과 같다.

$$\log \hat{p}(\boldsymbol{y}) = \log L(\boldsymbol{y}|\boldsymbol{\theta}^*) + \log \pi(\boldsymbol{\theta}^*) - \log \hat{\pi}(\boldsymbol{\theta}^*|\boldsymbol{y}).$$

이 함수를 2.1절과 2.2절에 제시된 방법에 따라 계산하는 절차에 대하여 알아보자. 먼저 모수에 대한 사후추정치 $\boldsymbol{\theta}^*$는 MCMC 표본추출에 의하여 추출된 표본을 이용하여 사후평균을 계산하여 이용할 수 있다. 이로부터 함수 $\log L(\boldsymbol{y}|\boldsymbol{\theta}^*)$은 식 (2.3)을 이용하여 다음과 같이 계산한다.

$$\log L(\boldsymbol{y}|\boldsymbol{\theta}^*) = \log \sum_i \sum_j y_{ij1} \log \left(\frac{\exp(\boldsymbol{z}'_{ij1}\boldsymbol{\beta}^*)}{\sum_i \sum_j \exp(\boldsymbol{z}'_{ij1}\boldsymbol{\beta}^*)} \right) + \sum_i y_{i+2} \log \left(\frac{\sum_j \exp(\boldsymbol{z}'_{ij2}\boldsymbol{\beta}^*)}{\sum_i \sum_j \exp(\boldsymbol{z}'_{ij2}\boldsymbol{\beta}^*)} \right).$$

식 $\log \pi(\boldsymbol{\theta}^*)$에 대한 계산은 식(2.5)와 (2.6)의 분포함수에 $\boldsymbol{\beta}^*$와 σ_1^{2*}와 σ_2^{2*}에 대체하여 계산할 수 있다. 이제 마지막으로 $\log \hat{\pi}(\boldsymbol{\theta}^*|\boldsymbol{y})$의 계산 방법에 대하여 알아보자.

2.1절과 2.2절에서 제시된 방법에 따라 고려되는 모형하에서 모수 벡터는 $\boldsymbol{\theta}^*$=$\{\{y^*_{ij2}\}, \boldsymbol{\beta}^*, \sigma_1^{2*}, \sigma_2^{2*}\}$ 이고 $\hat{\pi}(\boldsymbol{\theta}^*|\boldsymbol{y})$는 다음과 같이 주어진다.

$$\begin{aligned}\log\hat{\pi}(\{y^*_{ij2}\}, \boldsymbol{\beta}^*, \sigma_1^{2*}, \sigma_2^{2*}|\boldsymbol{y}) &= \log\hat{\pi}(\{y^*_{ij2}\}|\boldsymbol{y}) + \log\hat{\pi}(\boldsymbol{\beta}^*|\{y^*_{ij2}\}, \boldsymbol{y}) \\ &+ \log\hat{\pi}(\sigma_1^{2*}|\{y^*_{ij2}\}, \boldsymbol{\beta}^*, \boldsymbol{y}) + \log\hat{\pi}(\sigma_2^{2*}|\{y^*_{ij2}\}, \boldsymbol{\beta}^*, \boldsymbol{y}) \quad (2.12)\end{aligned}$$

이제 식 (2.12)의 우변의 각 항들의 계산 절차에 대하여 알아보자. 첫 번째로 $\hat{\pi}(\{y^*_{ij2}\}|\boldsymbol{y})$의 계산방법을 알아보자. 3.1절과 3.2.절에서 설명한 MCMC 과정을 통하여 충분히 많은 수를 표본을 얻었을 때 이 가운데 마지막 G개를 추출하여 다음의 식을 계산한다.

$$\hat{\pi}(\{y^*_{ij2}\}|\boldsymbol{y}) = \frac{1}{G}\sum_{g=1}^{G}\prod_{i=1}^{I}\frac{y_{i+2}!}{\prod_j y^*_{ij2}!}\left(\frac{\exp(\boldsymbol{z}'_{i12}\boldsymbol{\beta}^{(g)})}{\sum_j \exp(\boldsymbol{z}'_{ij2}\boldsymbol{\beta}^{(g)})}\right)^{y^*_{i12}}\cdots\left(\frac{\exp(\boldsymbol{z}'_{iJ2}\boldsymbol{\beta}^{(g)})}{\sum_j \exp(\boldsymbol{z}'_{ij2}\boldsymbol{\beta}^{(g)})}\right)^{y^*_{iJ2}}$$

두 번째로 $\hat{\pi}(\boldsymbol{\beta}^*|\{y^*_{ij2}\}, \boldsymbol{y})$를 계산하기 위하여 $\{y^*_{ij2}\}$가 주어진 상태 하에서 식 (2.8), (2.9)과 알고리즘 2.1을 G번 반복수행 하여 $\sigma_1^{2(g)}$, $\sigma_2^{2(g)}$, $\{\eta^{(g)}_{ijl}\}$ 표본을 G개 추출한다. 추출된 표본을 이용하여 사후분포함수의 추정치를 다음과 같이 계산한다.

$$\hat{\pi}(\boldsymbol{\beta}^*|\{y^*_{ij2}\}, \boldsymbol{y}) = \frac{1}{G}\sum_{g=1}^{G} MVN_q\left(\boldsymbol{V}^{(g)}(Z^T\Sigma^{-1(g)}\boldsymbol{\eta}^{(g)} + \Psi_{\boldsymbol{\beta}}^{-1}\boldsymbol{B}), \boldsymbol{V}^{(g)}\right).$$

여기서 $\boldsymbol{V}^{(g)} = (Z^T\Sigma^{-1(g)}Z + \Psi_{\boldsymbol{\beta}}^{-1})^{-1}$로 주어지며 $\Sigma^{(g)} = diag(\sigma_1^{2(g)} \times I_{I\times J}, \sigma_2^{2(g)} \times I_{I\times J})$이다.

마지막으로 $\log\hat{\pi}(\sigma_1^{2*}|\{y^*_{ij2}\}, \boldsymbol{\beta}^*, \boldsymbol{y})$과 $\log\hat{\pi}(\sigma_2^{2*}|\{y^*_{ij2}\}, \boldsymbol{\beta}^*, \boldsymbol{y})$를 계산한다. $\{y^*_{ij2}\}$와 $\boldsymbol{\beta}^*$가 주어진 상태하에서 식 (2.9)과 알고리즘 2.1을 G번 반복수행 하여 $\{\eta_{ijl}\}$ 표본을 G개 추출한다. 추출된 표본을 이용하여 사후분포함수의 추정치를 다음과 같이 계산한다.

$$\hat{\pi}(\sigma_1^{2*}|\{y^*_{ij2}\}, \boldsymbol{\beta}^*, \boldsymbol{y}) = \tfrac{1}{G}\textstyle\sum_{g=1}^{G} InverseGamma\left(\frac{I\times J+a_1}{2}, \frac{a_1b_1+(\boldsymbol{\eta}_1^{(g)}-Z_1\boldsymbol{\beta}^*)^T(\boldsymbol{\eta}_1^{(g)}-Z_1\boldsymbol{\beta}^*)}{2}\right),$$

$$\hat{\pi}(\sigma_2^{2*}|\{y^*_{ij2}\}, \boldsymbol{\beta}^*, \boldsymbol{y}) = \tfrac{1}{G}\textstyle\sum_{g=1}^{G} InverseGamma\left(\frac{I\times J+a_2}{2}, \frac{a_2b_2+(\boldsymbol{\eta}_2^{(g)}-Z_2\boldsymbol{\beta}^*)^T(\boldsymbol{\eta}_2^{(g)}-Z_2\boldsymbol{\beta}^*)}{2}\right).$$

3. 자료분석

분석에 이용된 데이터는 지난 2004년 총선을 앞두고 실시된 여론조사 가운데 서울시 광진구갑 선거구에 대한 조사 결과를 이용하였다. 917명에 대하여 전화번호부를 이용한 전화여론조사를 실시하였다. 이 가운데 4개의 변수를 이용하여 자료를 정리하였다. 먼저 변수 Y는 반응변수로서 지지후보를 나타낸다. 1='한나라당 홍희곤 - Hong (Han)', 2='열린우리당 김영춘 - Kim (Yeol)', 3='기타 - Etc.'를 나타낸다. X_1은 연령을 나타내는 변수로 1='20 대', 2='30대', 3='40대', 4='50대 이상'을 나타낸다. X_2는 교육정도를 나타내는 변수로 1='중졸이하', 2='고졸', 3='대졸이상'을 나타낸다. 마지막 X_3은 투표참여 의사를 나타내는 변수로 1='참여하겠다', 2='참여하지 않겠다'를 나타낸다. 각 변수들 가운데 오직 반응변수 Y만이 무응답이 발생하였고 나머지 변수들에 대해서는 무응답이 발생하지 않았다. Table 3.1은 관찰된 자료를 정리한 표이다. 총 917명의 응답자 가운데 60명이 지지후보에 대하여 응답하지 않아 무응답률은 6.5%이다. 응답자 가운데 43.5%는 한나라당 홍희곤 후보를 지지한다 응답하였고, 48.7%는 열린우리당 김영춘 후보를 지지한다고 응답하였다. 나머지 7.8%는 기타 후보를 지지한다고 응답하였다.

Table 3.1 Pre-election survey results for 2004 Korean Assembly of KJK district

X_1	X_2	X_3	Y			
			Hong (Han)	Kim (Yeol)	Etc.	Non-response
1	1	1	0	1	0	0
1	1	2	0	0	0	0
1	2	1	3	6	1	1
1	2	2	1	3	1	0
1	3	1	20	47	4	5
1	3	2	7	5	2	2
2	1	1	1	1	1	2
2	1	2	2	4	0	0
2	2	1	17	47	4	9
2	2	2	3	10	0	4
2	3	1	20	84	8	8
2	3	2	2	8	0	4
3	1	1	8	15	2	1
3	1	2	1	2	1	1
3	2	1	40	61	12	4
3	2	2	8	1	2	4
3	3	1	44	37	6	2
3	3	2	5	3	0	4
4	1	1	78	28	12	2
4	1	2	5	2	1	1
4	2	1	65	33	8	3
4	2	2	3	0	0	0
4	3	1	38	17	2	2
4	3	2	2	2	0	1

이제 Table 3.1의 자료를 분석하기 위한 로그선형모형을 고려하여보자. 이 표의 자료를 그대로 이용하는 경우 총 정보의 수는 $4 \times 3 \times 2 \times 4 - 1 = 95$가 되고 무응답 지시변수 R과의 상호작용 효과를 전혀 고려하지 않는 모형 즉 완전임의결측모형을 고려한다 하더라도 추정하여야 할 모수는 모두 72개이다. $\boldsymbol{\beta}$에 대한 조건부 사후분포로부터 표본 추출을 수행하는 경우 72-변량 정규분포로부터 표본 추출을 수행하여야 한다. 이와같은 모형의 복잡성을 완화시키기 위하여 Table 3.1의 자료를 두 개의 독립변수만을 고려하는 표로 분할하였다. 즉 X_1X_2Y, X_1X_3Y, X_2X_3Y의 변수만을 고려하는 3개의 분할표로 자료를 재 정리한 후 로그선형모형을 이용한 모형 추정및 모형 선택을 수행 하였다. 본 연구에서 비교하고자 하는 모형은 총 31개로 다음과 같다. 여기서 $A|B$는 변수 A와 B간의 주효과와 가능한 모든 상호작용 효과를 포함함을 의미한다. 실제 분석에서는 $(X_1|X_3|Y,\ X_1|Y|R)$ 모형과 $(X_2|X_3|Y,\ X_2|Y|R)$모형도 이용하였으나 우도함수의 자유도가 음수인 관계로 비교에서는 제외 하였다.

(1) X_1X_2Y 분할표

M_{1-1}: $X_1|X_2|Y,\ X_1|Y|R$

M_{1-2}: $X_1|X_2|Y,\ X_2|Y|R$

M_{1-3}: $X_1|X_2|Y,\ X_1|X_2|R$

M_{1-4}: $X_1|X_2|Y,\ X_1|R,\ X_2|R,\ Y|R$

M_{1-5}: $X_1|X_2|Y,\ X_1|R,\ Y|R$

M_{1-6}: $X_1|X_2|Y,\ X_2|R,\ Y|R$

M_{1-7}: $X_1|X_2|Y,\ Y|R$

M_{1-8}: $X_1|X_2|Y,\ X_1|R,\ X_2|R$

M_{1-9}: $X_1|X_2|Y,\ X_1|R$

M_{1-10}: $X_1|X_2|Y,\ X_2|R$

M_{1-11}: $X_1|X_2|Y,\ R$

(2) X_1X_3Y 분할표

M_{2-2}: $X_1|X_3|Y,\ X_3|Y|R$
M_{2-3}: $X_1|X_3|Y,\ X_1|X_3|R$
M_{2-4}: $X_1|X_3|Y,\ X_1|R,\ X_3|R,\ Y|R$
M_{2-5}: $X_1|X_3|Y,\ X_1|R,\ Y|R$
M_{2-6}: $X_1|X_3|Y,\ X_3|R,\ Y|R$
M_{2-7}: $X_1|X_3|Y,\ Y|R$
M_{2-8}: $X_1|X_3|Y,\ X_1|R,\ X_3|R$
M_{2-9}: $X_1|X_3|Y,\ X_1|R$
M_{2-10}: $X_1|X_3|Y,\ X_3|R$
M_{2-11}: $X_1|X_3|Y,\ R$

(3) X_2X_3Y 분할표

M_{3-2}: $X_2|X_3|Y,\ X_3|Y|R$
M_{3-3}: $X_2|X_3|Y,\ X_2|X_3|R$
M_{3-4}: $X_2|X_3|Y,\ X_2|R,\ X_3|R,\ Y|R$
M_{3-5}: $X_2|X_3|YX,\ _2|R,\ Y|R$
M_{3-6}: $X_2|X_3|Y,\ X_3|R,\ Y|R$
M_{3-7}: $X_2|X_3|Y,\ Y|R$
M_{3-8}: $X_2|X_3|Y,\ X_2|R,\ X_3|R$
M_{3-9}: $X_2|X_3|Y,\ X_2|R$
M_{3-10}: $X_2|X_3|Y,\ X_3|R$
M_{3-11}: $X_2|X_3|Y,\ R$

고려대상인 전체 모형에서 대하여 2.1절과 2.2절에서 제안한 방법에 의하여 조건부 사후분포로부터 표본을 추출하고 다시 2.3절에서 제안한 방법에 의하여 베이즈 인자를 구하기 위하여 주변 우도함수를 계산하였다. 사전분포에 대한 초모수 (hyper parameter)를 할당하기 위하여 먼저 식 (2.5)의 $\boldsymbol{\beta}$의 사전분포에 대한 분산-공분산행렬에 비정보적 사전분포를 할당하기 위하여 $\Psi_{\boldsymbol{\beta}} = I \times 10^6$을 할당하였다. σ_1^2와 σ_2^2의 사전분포 (2.6)의 할당을 위하여 모형에 따라 다른 초모수를 할당하였다. 모형 구분에 상관없이 a_1과 a_2에는 10을 할당하였다. 모형 M_{1-1}부터 모형 M_{1-11}, 모형 M_{2-1}부터 모형 M_{2-11}, 모형 M_{3-1}부터 모형 M_{3-11}에 대하여 b_1은 각각 1.67, 1.70, 1.80을 할당하였으며 b_2에 대해서는 3.3, 3.4, 3.6을 할당하였다. 이는 Green과 Park (2003)이 제안한 방법에 따라 관찰된 자료의 산술평균과 표본분산을 계산하여 이에 비례하도록 하였다. 그리고 무응답 빈도에 할당된 사전분포의 척도모수가 응답 빈도에 할당된 척도모수보다 큰 값을 가지도록 할당하였다. 다음으로 잠재변수 $\{\eta_{ijl}\}$ 추출을 위한 알고리즘, 즉 알고리즘 2.1에 대한 설정은 다음과 같다. 식(2.11)의 척도모수와 위치모수는 Green과 Park (2003)이 제안한 방법에 따라 다음과 같이 할당하였다.

$$\hat{\eta_{ijl}} = V_n\left(\frac{z_{ijl}^T\boldsymbol{\beta}}{\sigma_l^2} + \frac{\log(y_{ijl})}{y_{ijl}^{-1}}\right),$$

$$V_n = \left(\frac{1}{\sigma_l^2} + \frac{1}{y_{ijl}^{-1}} \right)^{-1}.$$

계속해서 비중심 t분포의 자유도는 15로 할당하였고 상수 c는 경우에 따라 1.7에서 2.8을 할당하였다. 이는 Gelman 등 (2004)의 제안에 따라 알고리즘 2.1의 (c)에서의 채택확률이 0.44가 되도록 조절하였다. MCMC 표본추출은 먼저 40,000번을 초기치로 수행하여 버린 후 다시 40,000번을 추출하여 각 모수에 대한 사후분포의 표본을 구축하였다. 이 40,000번의 표본으로 부터 계산된 사후분포의 표본평균을 계산하여 이를 각 모수에 대한 베이지안 추정치로 고려하였다. 이 계산된 추정치를 이용하여 베이즈 인자 계산을 위한 표본추출을 위하여 2.3절에 설명된 각 단계별로 다시 총 $G = 40,000$의 표본을 추출을 추출하였다. 이와 같은 계산과정을 거쳐 계산된 주변 우도함수의 로그 값은 다음 Table 3.2와 같다.

Table 3.2 $\log \hat{p}(\boldsymbol{y}|M)$ for model selection

Model	DF	$\log \hat{p}(\boldsymbol{y}\|M)$
M_{1-1}: $X_1\|X_2\|Y,\ X_1\|Y\|R$	0	-454.9
M_{1-2}: $X_1\|X_2\|Y,\ X_2\|Y\|R$	3	-449.1
M_{1-3}: $X_1\|X_2\|Y,\ X_1\|X_2\|R$	0	-466.6
M_{1-4}: $X_1\|X_2\|Y,\ X_1\|R,\ X_2\|R,\ Y\|R$	4	-429.0
M_{1-5}: $X_1\|X_2\|Y,\ X_1\|R,\ Y\|R$	6	-412.3
M_{1-6}: $X_1\|X_2\|Y,\ X_2\|R,\ Y\|R$	7	-416.1
M_{1-7}: $X_1\|X_2\|Y,\ Y\|R$	9	-398.4
M_{1-8}: $X_1\|X_2\|Y,\ X_1\|R,\ X_2\|R$	6	-425.3
M_{1-9}: $X_1\|X_2\|Y,\ X_1\|R$	8	-405.6
M_{1-10}: $X_1\|X_2\|Y,\ X_2\|R$	9	-410.5
M_{1-11}: $X_1\|X_2\|Y,\ R$	11	-402.5
M_{2-2}: $X_1\|X_3\|Y,\ X_3\|Y\|R$	2	-316.3
M_{2-3}: $X_1\|X_3\|Y,\ X_1\|X_3\|R$	0	-329.4
M_{2-4}: $X_1\|X_3\|Y,\ X_1\|R,\ X_3\|R,\ Y\|R$	1	-309.5
M_{2-5}: $X_1\|X_3\|Y,\ X_1\|R,\ Y\|R$	2	-325.7
M_{2-6}: $X_1\|X_3\|Y,\ X_3\|R,\ Y\|R$	4	-291.9
M_{2-7}: $X_1\|X_3\|Y,\ Y\|R$	5	-317.9
M_{2-8}: $X_1\|X_3\|Y,\ X_1\|R,\ X_3\|R$	3	-303.9
M_{2-9}: $X_1\|X_3\|Y,\ X_1\|R$	4	-318.6
M_{2-10}: $X_1\|X_3\|Y,\ X_3\|R$	6	-296.8
M_{2-11}: $X_1\|X_3\|Y,\ R$	7	-306.5
M_{3-2}: $X_2\|X_3\|Y,\ X_3\|Y\|R$	0	-229.4
M_{3-3}: $X_2\|X_3\|Y,\ X_2\|X_3\|R$	0	-240.9
M_{3-4}: $X_2\|X_3\|Y,\ X_2\|R,\ X_3\|R,\ Y\|R$	0	-223.8
M_{3-5}: $X_2\|X_3\|YX,\ X_2\|R,\ Y\|R$	1	-244.2
M_{3-6}: $X_2\|X_3\|Y,\ X_3\|R,\ Y\|R$	2	-216.5
M_{3-7}: $X_2\|X_3\|Y,\ Y\|R$	3	-228.0
M_{3-8}: $X_2\|X_3\|Y,\ X_2\|R,\ X_3\|R$	2	-225.3
M_{3-9}: $X_2\|X_3\|Y,\ X_2\|R$	3	-237.4
M_{3-10}: $X_2\|X_3\|Y,\ X_3\|R$	4	-222.9
M_{3-11}: $X_2\|X_3\|Y,\ R$	5	-219.4

Table 3.2의 결과를 살펴보자. 가장 큰 주변우도함수값을 가지는 모형은 $(X_2|X_3|Y,\ X_3|R,\ Y|R)$ 모형 (M_{3-6})이다. 유사하게 모형 M_{2-2}에서 모형 M_{2-11}사이에서도 가장 큰 주변우도함수값을 가지는 모형은 $(X_1|X_3|Y,\ X_3|R,\ Y|R)$모형 (M_{2-6})이다. 이를 종합하여 볼 때 변수 X_3인 투표 참여여부가 다른 두 변수 X_1, X_2, 즉 연령과 학력에 비하여 모형적합에 보다 적절한 요인이라 할 수 있다. 즉 투표참여여부가 당선자를 예측하는데 가장 중요한 변수라 할 수 있다. 그리고 이 두 모형은 모두 반응변수 Y와 무응답 지시변수 R간의 상호작용 효과를 포함하고 있는 모형이고 무응답체계는 무시할수 없는 무응답 모형이 된다. 이 상호작용 효과를 포함하지 않고 이 두 모형에

Table 3.3 Forecasting result of model $X_2|X_3|Y,\ X_3|R,\ Y|R$ (M_{3-6})

		Respondents			Non-response	Non-respondents		
X2	X3	Hong (Han)	Kim (Yeol)	Etc.		Hong (Han)	Kim (Yeol)	Etc.
1	1	87.0	45.0	14.9	5.0	1.9	1.7	1.9
1	2	8.0	8.0	1.9	2.0	0.9	0.9	0.9
2	1	125.1	147.0	25.1	17.0	5.0	6.0	6.3
2	2	15.0	14.0	3.0	8.0	2.7	2.7	2.9
3	1	122.1	185.1	20.1	17.0	5.5	5.8	5.9
3	2	16.0	18.1	2.1	11.0	3.6	4.0	3.7

대응되는 모형은 $(X_2|X_3|Y,\ X_3|R)$모형 (M_{3-10})과 $(X_1|X_3|Y,\ X_3|R)$모형 (M_{2-10})이다. Congdon (2002, p.471)이 제안한 기준에 따라 $2(\log \hat{p}(\boldsymbol{y}|M_{3-6}) - \log \hat{p}(\boldsymbol{y}|M_{3-10}))$=2(-216.5-(-222.9))=12.8이고 $2(\log \hat{p}(\boldsymbol{y}|M_{2-6}) - \log \hat{p}(\boldsymbol{y}|M_{2-10}))$=2(-291.9-(-296.8))=9.8로 무시할수 없는 무응답 모형이 더 지지된다 할 수 있다.

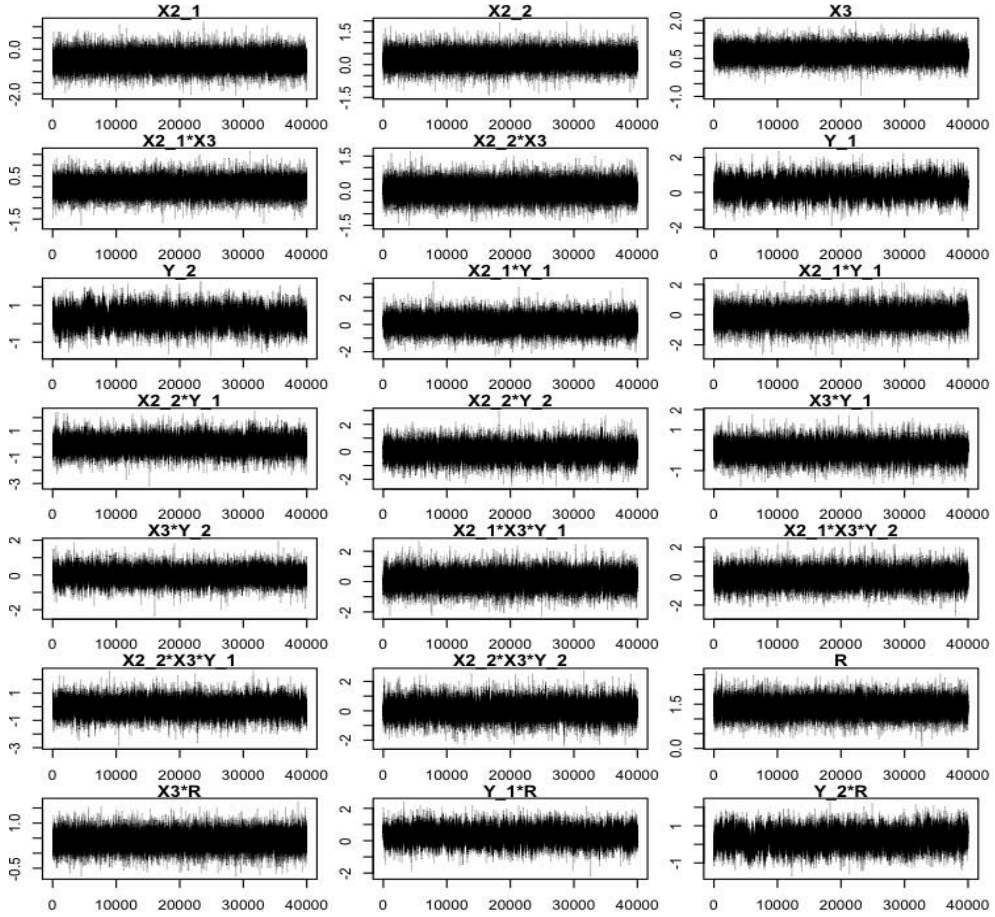

Figure 3.1 The trace plots for β of model $(X_2|X_3|Y,\ X_3|R,\ Y|R)$

최종적으로 선택된 $(X_2|X_3|Y,\ X_3|R,\ Y|R)$ 모형의 예측결과를 살펴보자. Table 3.3은 조건부 사후분포로 추출된 결과를 정리하여 모수의 사후분포에 대한 평균을 구하여 정리한 표이다. 표의 3∼5열은 응답자들의 추정결과이고 7∼9열은 열 6로 주어진 무응답값을 대체한 결과이다. 무시할 수 없는 무응답 모형이기 때문에 응답자와 무응답자들간의 응답패턴을 다른 것을 볼 수 있다. 또한 특정범주에서 0값이 대체되는 변방값 문제도 발생하고 있지 않음을 볼 수 있다. 최종적인 예측 결과를 보면 한나라당, 열린우리당, 기타 후보간 지지율은 42.7%, 47.7%, 9.6%로 예측되었다. 이에 반하여 실제 투표결과는 40.7%, 50.7%, 8.6%로 나타났다.

마지막으로 최종 선택된 모형 $(X_2|X_3|Y,\ X_3|R,\ Y|R)$에 대하여 MCMC 과정에서 모형의 수렴여부를 확인하기 위하여 선형예측식의 모수인 β들에 대한 시도표를 작성하였다. Figure 3.1은 $(X_2|X_3|Y,\ X_3|R,\ Y|R)$ 모형의 모수 β 가운데 절편항을 제외한 나머지 변수들에 대한 시도표이다. 모든 경우에서 안정적으로 수렴하고 있는 것을 볼 수 있다.

4. 결론

본 연구에서 우리는 무응답을 포함하고 있는 다차원 분할표 형태로 정리된 범주형 자료에 대한 모형 추정 방법으로 계층적 베이지안 방법을 제시하였다. 조건부 사후분포로부터 모수를 추출하기 위한 MCMC 방법을 제시하였으며 보다 효율적인 조건부 사후분포부터의 모수 추출을 위하여 잠재변수를 이용하는 방법을 이용하였다. 본 논문에서는 제시하지 않았으나 MCMC 표본추출에 의한 모형 수렴결과는 모두 비교적 안정적으로 수렴하는 것으로 판정 되었다. 베이지안 접근하에서의 모형 비교를 위하여 베이즈 인자를 통한 모형 비교를 수행하였으며 모형 수행과정에서 무응답 체계에 대한 비교 또한 함께 수행하였다. 제안된 방법은 우리나라 국회의원 선거를 앞두고 실시된 사전 여론조사 결과 자료를 이용하여 분석을 수행하였다. 분석결과 3개의 설명변수들 가운데 투표참여여부가 반응변수에 가장 밀접한 연관을 가지는 것으로 판명 되었으며 무응답 체계는 무시할 수 없는 무응답 모형이 보다 적합한 것으로 판명 되었다. 실제 예측결과와 비교해 보면 약간의 차이가 있음을 볼 수 있으나 전체적인 표본수가 917인 것을 감안하면 충분히 좋은 결과라 볼 수 있다. 본 연구에서 제시된 방법은 앞으로 시행되는 여론조사에서 보다 정확한 예측을 수행하는데 이용될 수 있을 것이로 기대된다.

참고문헌

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.

Cargnoni, C., Miller, P. and West, M. (1997). Bayesian forecasting of multinomial time series through conditionally Gaussian dynamic models. *Journal of the American Statistical Association*, **92**, 640-647.

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.

Chen, Q. L. and Stasny, E. A. (2003). *Handling undecided voters: Using missing data methods in election forecasting*, Technical Report, Department of Statistics, The Ohio State University.

Choi, B. (2007). A study of customer segmentation method using nonresponse model. *Journal of the Korean Data Analysis Society*, **9**, 1849-1860.

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.

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., 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.

Chun, Y. M., Son, H. K. and Chung, S. S. (2007). Treatment of missing data by decomposition and voting with ordinal data. *Journal of the Korean Data & Information Science Society*, **18**, 585-598.

Chung, H. C. and Han, C. P. (2009). Bootstrap confidence intervals for classification error rate in circular models when a block of observation is missing. *Journal of the Korean Data & Information Science Society*, **20**, 757-764.

Congdon, P. (2002). *Bayesian statistical modelling*, first edition, John Wiley & Sons Ltd., Chichester.

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*, **39**, 1-38.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling based to calculating marginal densities, *Journal of the American Statistical Association*, **85**, 398-409.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004). *Bayesian data analysis*, second edition, Chapman & Hall/CRC, Florida.

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

Hong, J. S. and Jung, M. H. (2011a). Undecided inference using logistic regression for credit evaluation. *Journal of the Korean Data & Information Science Society*, **22**, 149-157.

Hong, J. S. and Jung, M. S. (2011b). Undecided inference using bivariate probit models. *Journal of the Korean Data & Information Science Society*, **22**, 1017-1028.

Kass, R. E. and Raftery, E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773-795.

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

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 the American Statistical Association*, **89**, 44-52.

Park, Y. S. and Choi, B. (2010). Bayesian analysis for incomplete multi-way contingency tables with non-ignorable 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.

Journal of the Korean Data & Information Science Society
2012, **23**(4), 627–641

http://dx.doi.org/10.7465/jkdi.2012.23.4.627
한국데이터정보과학회지

Model selection method for categorical data with non-response[†]

Yong Hwa Yoon[1] · Boseung Choi[2]

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

Received 23 April 2012, revised 5 June 2012, accepted 15 June 2012

Abstract

We consider a model estimation and model selection methods for the multi-way contingency table data with non-response or missing values. We also consider hierarchical Bayesian model in order to handle a boundary solution problem that can happen in the maximum likelihood estimation under non-ignorable non-response model and we deal with a model selection method to find the best model for the data. We utilized Bayes factors to handle model selection problem under Bayesian approach. We applied proposed method to the pre-election survey for the 2004 Korean National Assembly race. As a result, we got the non-ignorable non-response model was favored and the variable of voting intention was most suitable.

Keywords: Bayes factor, election, MCMC, model selection, non-response.

[†] This research was supported by Daegu University Research Grant in 2011 (No. 20110207).
[1] Professor, Department of Statistics and Computer Science, Daegu University, Gyeongbuk 712-714, Korea.
[2] Corresponding author: Assistant professor, Department of Statistics and Computer Science, Daegu University, Gyeongbuk 712-714, Korea. E-mail: bchoi@daegu.ac.kr