응용통계연구 제 **21**권 *3*호, *2008*년, *pp. 361-375*

무시할 수 없는 무응답을 가지고 있는 교체표본조사에서의 무응답 대체와 교체그룹 편향 추정*

최보승[1] 김대영[2] 김기환[3] 박유성[4]

요 약

본 논문에서는 패널의 일부를 규칙적으로 교체하는 4-8-4 교체표본설계에서 발생할 수 있는 항목 무응답을 대체하는 방법에 대하여 연구하였다. 특히 소득이나 취업과 같이 민감한 질문에 대하여 발생할 수 있는 무응답에 대하여 무시할 수 없는 무응답(nonignorable nonresponse) 체계하에서 발생하는 무응답을 가정하였다. 무응답들의 대체방법으로 모형에 기반한 대체방법을 고려하였으며 베이지안 방법을 이용하여 사후확률밀도함수를 최대화하는 최대사후우도추정량(maximum posterior likelihood estimator)을 구하였다. 그리고 대체된 자료를 이용하여 면접시점이 달라질 때 발생하는 편향을 추정하였으며 추정된 편향을 제거한 후 연속적인 두 조사기간에서의 각 칸의 확률과 고정된 시점에서의 주변확률을 계산하였다. 모의실험을 통해 최종적으로 도출된 결과를 평균제곱오차와 편향의 관점에서 비교하였다.

주요용어: 무응답 대체, 무시할 수 없는 무응답, 교체표본조사, EM 알고리즘.

1. 서론

사회를 구성하는 개인의 성향이 점차 다양해짐에 따라 사회구성원들의 특성이나 욕구를 파악하기 위한 다양한 조사가 시도되고 있다. 국가적인 차원에서도 각종 정책수립을 위한 기초자료를 확보하기 위해 가구소득분포 및 소비수준과 취업, 실업, 노동력 등과 같은 인구의 경제적 특성을 대규모의 비용을 투입하여 조사한다.

실제로 조사를 통해 얻는 자료에는 적지않은 무응답이 포함되어 있는 경우가 대부분이다. 무응답은 대표적인 비표본 오차 중 하나이다. 무응답을 제외하고 분석을 수행하게 되면 연구자의 의도와는 달리 단순히 응답한 집단 만을 대상으로 분석한 것이 되고 표본의 크기가 줄어 추정치의 효율이 떨어진다. 또, 무응답이 없는 자료를 바탕으로 하는 표준적인 통계적 방법을

* 본 연구는 고려대학교 교내연구교원지원사업에 의하여 수행되었음.

1) (136-701) 서울시 성북구 안암동 5가 1, 고려대학교 통계연구소, 연구교수.
E-mail: cbskust@gmail.com
2) Graduate student, Dept. of Statistics, Pennsylvania state university, PA 16802, U.S.A.
E-mail: sas2000@hanmail.net
3) (339-770) 충청남도 연기군 조지원읍 서창리 208 고려대학교 과학기술대학 정보통계학과, 부교수.
E-mail: korpen@korea.ac.kr
4) (136-701) 교신저자. 서울시 성북구 안암동 5가 1, 고려대학교 정경대학 통계학과, 교수.
E-mail: yspark@korea.ac.kr

적용하기가 어렵고, 무응답자들의 성향이 응답자들과 체계적인 차이를 보이게 되면 추정치에는 편향이 생기게 된다.

무응답을 형태에 따라 구분하면, 여러 개의 응답항목으로 이루어진 조사단위 전체에 답하지 않은 단위 무응답(unit nonresponse)과 조사단위에는 응답하였으나 그 조사단위를 구성하고 있는 몇 개의 중요한 항목에 대해 응답을 하지 않은 항목 무응답(item nonresponse)으로 나눈다. Little과 Rubin (2002)은 무응답이 발생하는 구조적인 체계를 응답확률의 행태에 따라 무시할 수 있는 무응답(ignorable nonresponse)과 무시할 수 없는 무응답(nonignorable nonresponse)으로 나누었다. 무시할 수 있는 무응답은 무응답이 무응답값과 관계가 없지만 무시할 수 없는 무응답일 때는 무응답이 무응답값과 상관이 있어서 무응답할 확률이 무응답값에 영향을 받는다. 가구소비실태조사나 경제활동인구조사의 경우 연령이나 성별에 대해 같은 정보를 가진 사람들 내에서 소득이나 취업에 대한 질문에서 발생하는 무응답은 소득금액이나 취업여부에 의존하는 경향이 있다. 이러한 사실은 Paul과 Lawes (1982), Fienberg와 Stasny (1983)에 의해 연구된 바 있다. 따라서, 이러한 질문들에서 발생하는 무응답은 무응답 발생 확률이 무응답 값에 영향을 받는 무시할 수 없는 무응답(nonignorable nonresponse)에 속하게 된다.

패널형태의 자료는 조사에서 다수의 질문을 연속적으로 반복하여 얻을 수 있는데 이 과정에서 무응답이 많이 발생할 수 있다. 따라서 패널평태의 자료에서 무응답 처리에 관한 연구는 다양하게 이루어져 왔다. David 등 (1986)은 Current Population Survey(CPS)의 소득자료를 대체하는데 있어 기존의 대체방법인 CPS 핫덱 방법을 회귀모형에 기초한 무응답 대체 방법과 비교하였다. Stasny (1986)는 캐나다 취업 조사자료에서 무응답을 비임의적으로 간주하였으며 무응답이 취업상태(취업,실업,비경제활동인구)나 시간과 관련이 있다는 것을 모형을 통해 고려하여 노동력의 흐름을 최대우도방법으로 추정하였다. 또, Stasny (1991)는 범죄조사자료를 통해 비임의적 무응답 가정 하에서 소지역내 의 조사대상들이 무응답을 할 확률을 계층적 모형과 경험적 베이즈 방법을 이용해서 추정하였다. Conaway (1992)는 반복 측정된 범주형 자료를 분석하기 위한 조건부 우도방법을 무시할 수 있는 또는 무시할 수 없는 무응답이 있는 경우에 적용하였다. Bonetti 등 (1999)은 임상시험를 통해 얻어지는 삶의 질자료에서 무시할 수 없는 무응답이 발생한다는 사실을 인지하고 이를 고려하기 위해 설계된 모형을 추정하는데 적률추정방법을 사용하였다.

본 논문에서는 CPS의 4-8-4 교체표본설계(rotation sampling design)에서 발생하는 두 가지 종류의 무응답을 대체하는 방법에 대해 연구하였고, 면접시점이 달라서 발생하는 편향(interview time bias)을 추정하였다. 또, 추정된 편향을 제거한 후에 연속적인 두 조사기간에 대한 각 칸의 확률과 고정된 시점에서의 주변확률을 추정하였다. 무응답들을 대체할 때 이용한 방법들은 최대우도추정량(maximum likelihood estimator: MLE)과 사후확률밀도함수를 최대화하는 최대사후우도추정량(maximum posterior estimator: MPE)이다. 특히 칸도수 추정치가 0이 되는 현상을 줄이기 위해서 사전분포의 모수에 대해 Clogg 등 (1991), Park과 Brown (1994)이 제안한 방법을 현실적인 상황에 부합되게 수정한 베이지안 방법 (최보승 등, 2007)을 이용하였다. 최종적으로 도출된 결과들을 평균제곱오차(mean squared error: MSE)와 편향(bias)의 관점에서 비교하였다.

α		1								2								3								4	
g		1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8	1	2
t-1		u_8	u_7	u_6	u_5								u_4	u_3	u_2	u_1											
t			u_8	u_7	u_6	u_5								u_4	u_3	u_2	u_1										
	t+1			u_8	u_7	u_6	u_5								u_4	u_3	u_2	u_1									
M	t+2				u_8	u_7	u_6	u_5								u_4	u_3	u_2	u_1									
o	t+3					u_8	u_7	u_6	u_5								u_4	u_3	u_2	u_1								
n	t+4						u_8	u_7	u_6	u_5								u_4	u_3	u_2	u_1							
t	t+5							u_8	u_7	u_6	u_5							u_4	u_3	u_2	u_1							
h	t+6								u_8	u_7	u_6	u_5						u_4	u_3	u_2	u_1							
	t+7									u_8	u_7	u_6	u_5					u_4	u_3	u_2	u_1							
	t+8										u_8	u_7	u_6	u_5				u_4	u_3	u_2	u_1							
	t+9											u_8	u_7	u_6	u_5			u_4	u_3	u_2	u_1							
	t+10												u_8	u_7	u_6	u_5		u_4	u_3	u_2								
	t+11													u_8	u_7	u_6	u_5	u_4	u_3									
	t+12														u_8	u_7	u_6	u_5								u_4	

그림 2.1: 4-8-4 교체표본설계구조

2. 교체표본조사와 무응답

2.1. 교체표본조사

시간이 지나감에 따라 모집단의 특성이 변화하는데, 변화하는 모집단의 특성을 조사하는 방법으로 반복조사(repeated survey)가 사용된다. 그러나 조사를 계속 실시하게 되면 조사대상이 응답에 대한 부담을 느끼게 되며 이로 인해 수집된 자료의 신뢰성에 부정적인 영향을 미치게 된다. 그러나 조사결과의 안정성 때문에 조사대상을 매 조사시마다 바꿀 수도 없다. 이러한 문제점을 해결하기 위해 제안된 조사방법이 교체표본조사(rotation sample survey)이다. 교체표본조사는 조사대상을 계속적으로 패널에 포함시키지 않고 일정한 규칙에 따라 교체하는 방법으로, 조사대상이 일정한 기간동안 조사에 참여한 후 패널에서 제외되기 때문에 응답자의 부담을 줄일 수 있게 된다.

그림 2.1을 통해 본 연구에서 다루는 미국의 CPS 4-8-4 교체표본설계를 구체적으로 살펴보자. 그림 2.1에서 (α, g)는 g번째 교체그룹에 α번째 표본단위를 나타내는 지표이고, u_i는 특정한 달에 i번째로 면접하는 표본단위를 의미한다. 실제조사에서 교체표본조사의 가장 마지막 조사단위는 가구 및 가구 구성원인데, 표본교체 디자인을 구성하는 큰 틀은 동질적인 다수의 가구로 구성된 교체그룹들이다. 즉, 4-8-4 교체표본설계의 경우는 모두 8개의 교체그룹으로 구성되어 있고 g번째 교체그룹에서 추출된 표본단위는 t월에서 처음으로 조사되어 4개월 동안 조사에 참여하고 8개월 동안 조사에서 제외된 후 다시 4개월 동안 조사에 참여하는 과정을 1번 반복한다. 그리고 각 표본단위의 정보는 면접시점에만 보고 되며 모두 8번 조사된다. 4-8-4 교체표본설계의 특징은 8개의 교체그룹이 면접하는 시점에 모두 포함되어 있고, 면접시점이 수평적으로나 수직적으로 균형되어 있다는 점이다. 즉, 면접시점에 대한 수평적 균형은 조사되는 달에 1부터 8까지의 모든 교체그룹이 포함되는 것을 의미하며 수직적 균형은 8개월의 조사기간을 고려했을 때 각 교체그룹이 8번의 조사시점을 포함하고 있음을 나타낸다.

2.2. 교체표본조사에서 발생하는 무응답

두 조사기간 $(t, t+1)$에서 보면 실제로 조사대상의 응답거부나 정보부족 등의 이유로 여타 다른 조사처럼 무응답이 발생하는데, 교체표본설계의 특성으로 인해 발생하는 무응답도 있다. 즉, 항상 8개의 교체그룹들 가운데 6개의 교체그룹에서는 동일 표본단위의 동일 가구들이 연속적으로 조사되고 2개의 교체그룹에서는 동일한 그룹내의 서로 다른 표본단위에 속한 가구들이 조사된다(본 연구에서는 편의상 그룹 2와 6이 설계무응답이 발생하는 그룹이라고 하고 나머지 그룹 1, 3, 4, 5, 7, 8은 조사무응답이 발생하는 그룹이라고 하며, 두 기간 모두 무응답인 자료는 없다고 가정한다.).

현실적으로 관심있는 것은 연속적인 두 조사기간 $(t, t+1)$에서 노동력(취업 또는 미취업)의 흐름인데, 8개 그룹 중 연속적으로 관찰되지 않는 2개 그룹을 제외하는 것을 고려할 수 도 있다. 하지만 만약 설계로 인해 무응답이 발생하는 2개 그룹을 제외한다면 우선 경제적으로 비효율적일 뿐만 아니라 연구자나 조사자가 알고자 하는 목적을 위해 가용할 수 있는 정보를 잃게 된다. 또, 4-8-4 교체표본설계에서는 교체그룹마다 면접시점이 다르기 때문에 편향이 발생한다고 보는 것이 일반적이며 8개 그룹의 개별적인 편향은 다르지만 8개 그룹의 편향이 항상 0이라는 가정을 하고 있다. 즉, 6개 그룹만을 고려하면 6개 그룹이 가지는 편향의 합에 대한 정보가 없으므로 상이한 면접시점으로 인한 편향을 추정 할 수 없다. 따라서, 편향을 정확하게 추정하기 위해 8개 그룹을 사용해야 하며, 무응답이 발생 할 경우 조사시 발생하는 무응답 뿐 만 아니라 설계로 인해 발생하는 무응답도 고려해야 한다. 본 연구에서는 조사를 실시하면서 발생하는 무응답을 조사 무응답(survey nonresponse)으로 정의하고, 표본교체조사의 설계로 인해 발생하는 무응답을 설계 무응답(design nonresponse)으로 정의한다.

이러한 설계무응답이 발생하는 2개 그룹(그룹 2와 6)에서 $(t, t+1)$에 관찰되는 대상들은 면접시점도 8과 1, 4와 5로서 다르다. 따라서, 설계무응답이 발생하는 그룹들에 대해서는 표본단위별로 무응답을 대체시키고 그 결과로 만들어진 완전한 자료를 가중치에 의해 조절해서 해당 그룹을 대표하는 정보로 이용하도록 한다. 또, $(t, t+1)$에서 각 그룹의 면접시점이 모두 상이해서 각 그룹마다 편향을 가지고 있으므로, 올바른 칸의 확률을 추정할 수 없다(단, 8개 그룹이 가지고 있는 편향의 합은 0이다). 따라서 무응답을 대체한 후에 각 그룹의 편향을 추정하여 추정된 편향을 제거하고 칸의 확률을 추정하는 것이 바람직하다. 본 연구에서는 모형에 기반한 방법에 의한 무응답 대체 방법을 이용하였다. 이 방법은 부분적으로 발생한 무응답자료를 생성한다고 가정되는 하나의 초 모집단 모형을 정의하고, 정의된 모형하에서 우도를 기초로 추론한다. 이 때, 모수들은 최대우도추정법이나 베이지안 방법 등에 의해 추정된다.

3. 모형 추정 방법

3.1. 무시할 수 없는 무응답모형

연속적인 두 기간동안 4-8-4 교체표본조사에서 얻을 수 있는 정보는 조사대상으로 선택된 단위들이 속한 교체그룹, 조사기간 $(t, t+1)$에 조사대상의 그룹화된 반응값 그리고 조사대상의 응답여부이다. 면접시점을 나타내는 교체그룹 X는 항상 관찰되며, 범주수는 $G(=8)$개이다. t월과 $t+1$월에서의 그룹화된 반응값을 나타내는 Y_t와 Y_{t+1}은 각각 I개, J개의 범주로 구

성되어 있다. 또, Y_t와 Y_{t+1}에서 응답여부를 나타내는 지시변수 R은 응답이면 $R=1$이고 무응답이면 $R=2$이다. 이 때, 항상 관찰되는 교체그룹 X는 조사대상자들의 면접시점에 대한 정보를 제공해 주고 있다. 즉, 한 조사대상자가 조사기간동안 모두 8번 응답할 기회가 생기는데 면접시점을 통해 조사대상자의 정보가 편향되어 있다는 사실을 알 수 있다.

동일한 교체그룹에 속한 조사대상자들에 대해 조사를 실시 할 때, 무응답여부는 질문자체에 의존하는 것이 일반적이다. 가령, 소득이나 취업여부에 대한 질문을 했을 때, 응답여부는 소득금액이나 취업여부에 대한 질문자체에 의존해서 발생한다. 이러한 경우 조사기간 $(t, t+1)$에서 조사대상자들의 반응변수의 값$(Y_t,\ Y_{t+1})$과 응답여부(R)의 상호작용 효과를 포함하는 무시할 수 없는 무응답모형을 고려해야 한다.

m_{gijr}를 g번째 교체그룹에서 t월에 반응변수 Y_t의 수준이 i, $t+1$에 반응변수 Y_{t+1}의 수준이 j, 응답여부가 r인 경우의 기대 칸 도수라 할 때, $\mathbf{m}$을 m_{gijr}로 이루어진 $c\times 1$의 열벡터라 하고, Z를 $c\times p$ 크기의 계획행렬 그리고 $\boldsymbol{\beta}$를 $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_p)$인 모수벡터라고 하면 X, Y_t, Y_{t+1} 그리고 R로 이루어지는 무응답모형은

$$\log \mathbf{m} = \mathbf{Z}\beta$$

로 정의된다.이 때, c는 전체 카의 수이고, p는 모수의 개수이다.

Baker와 Laird (1988)에 의해 칸의 확률 $\pi_{gij|r}$은 $\pi_{gij|r}=m_{gijr}/m_{+++r}$이므로 응답여부가 알려져 있을 때, 교체그룹(또는 면접시점) X, t달의 취업상태 Y_t 그리고 $t+1$달의 취업상태 Y_{t+1}의 결합확률밀도함수에 대한 로그우도함수는

$$\ell = \log L = \log\left(\prod_g\prod_i\prod_j \pi_{gij|1}{}^{n_{gij1}}\cdot\prod_g\prod_i \pi_{gi+|2}{}^{n_{gi+2}}\cdot\prod_g\prod_j \pi_{g+j|2}{}^{n_{g+j2}}\right)$$
$$= \sum_g^6\sum_i^I\sum_j^J n_{gij1}\ \log(\pi_{gij|1}) + \sum_g^8\sum_i^I n_{gi+2}\ \log(\pi_{gi+|2}) + \sum_g^8\sum_j^J n_{g+j2}\ \log(\pi_{g+j|2})$$

에 비례한다 (단, $\pi_{gij|1}=\Pr(X=g,\ Y_t=i,\ Y_{t+1}=j|R=1)$, $\pi_{gi+|2}=\sum_j \Pr(X=g,\ Y_t=i,\ Y_{t+1}=j|R=2)$, $\pi_{g+j|2}=\sum_i \Pr(X=g,\ Y_t=i,\ Y_{t+1}=j|R=2)$이고 n_{gij1}은 응답자들 중 $X=g,\ Y_t=i,\ Y_{t+1}=j$인 우의 도수이고, n_{gi+2}은 응답자들 중 $t+1$달에 무응답하고 $X=g,\ Y_t=i,\ Y_{t+1}=j$일 때 도수이며, n_{g+j2}은 응답자들 중 t달에 무응답하고 $X=g,\ Y_{t+1}=j$인 경우의 도수이다).

그러나 무시할 수 없는 무응답 가정하에서 최대우도추정량을 계산하면 반응변수의 범주들 가운데 적어도 한 범주에서 무응답에 대한 추정값이 0이 되는 변방값 문제(boundary solution problem)가 발생한다 (Baker와 Laird, 1988; Park과 Brown, 1994; Smith 등, 1999). 그 결과 추정치가 수렴하더라도 모수공간의 변방에서 좋지 못한 최대우도추정치가 나오거나 추정치가 유일하게 결정되지 못하는 문제가 발생하게 된다. 본 연구에서는 변방값이 발생하는 상황에 적절히 대처하기 위해 $\pi_{gij|2}$에 대해 사전분포를 고려하는 베이지안 방법을 사용하였다. 즉, 무응답 칸 $\pi_{gij|2}$에 다항분포와 켤레분포(conjugate distribution)인 Dirichlet분포를 사전분포로 고려하고 사후분포함수를 구해서 이것을 최대화시키는 최대사후추정량을 구하고자 하

였다. 무응답 칸 $\pi_{gij|2}$에 Dirichlet 사전분포를

$$\prod_g \prod_i \prod_j \pi_{gij|2}{}^{\delta_{ij}}$$

라 했을 때, 로그사후우도함수는

$$lp = \sum_g^6 \sum_i^I \sum_j^J n_{gij1} \log(\pi_{gij|1}) + \sum_g^8 \sum_i^I n_{gi+2} \log(\pi_{gi+|2})$$
$$+ \sum_g^8 \sum_j^J n_{g+j2} \log(\pi_{g+j|2}) + \sum_g^8 \sum_i^I \sum_j^J \delta_{ij} \log(\pi_{gij|2})$$

이 된다. 이 때, $\pi_{gij|r} = \pi_{gijr}/\pi_{+++r} = m_{gijr}/m_{+++r}$이다.

이제 Dempster 등 (1977)이 제안한 GEM 알고리즘을 이용하여 MPE를 계산하는 구체적인 방법을 알아보자.

E – STEP: 무응답자들에 대한 정보인 주변합을 무응답 칸에 할당시켜서 의사-관찰빈도수를 정의한다. $n^{(1)}_{2ij2}, n^{(2)}_{2ij2}, n^{(1)}_{6ij2}$와 $n^{(2)}_{6ij2}$를 각각

$$n^{(1)}_{2ij2} = \frac{n_{2i+2}}{n_{2i+2}+\delta_{i+}}\left(n_{2i+2}\frac{m_{2ij2}}{m_{2i+2}}+\delta_{ij}\right), \quad n^{(2)}_{2ij2} = \frac{n_{2+j2}}{n_{2+j2}+\delta_{+j}}\left(n_{2+j2}\frac{m_{2ij2}}{m_{2+j2}}+\delta_{ij}\right),$$

$$n^{(1)}_{6ij2} = \frac{n_{6i+2}}{n_{6i+2}+\delta_{i+}}\left(n_{6i+2}\frac{m_{6ij2}}{m_{6i+2}}+\delta_{ij}\right), \quad n^{(2)}_{6ij2} = \frac{n_{6+j2}}{n_{6+j2}+\delta_{+j}}\left(n_{6+j2}\frac{m_{6ij2}}{m_{6+j2}}+\delta_{ij}\right)$$

로 표현하면

$$n^*{}_{gijk} = \begin{cases} n_{gij1}, & \text{if 응답 } (g=1,3,4,5,7,8), \\ n_{gi+2}\cdot\dfrac{\left(n_{gi+2}\cdot\dfrac{m_{gij2}}{m_{gi+2}}+n_{g+j2}\dfrac{m_{gij2}}{m_{g+j2}}+\delta_{ij}\right)}{n_{gi+2}+\sum_j n_{g+j2}\dfrac{m_{gij2}}{m_{g+j2}}+\delta_{i+}}, & \text{if } T+1\text{기 무응답} \\ & (g=1,3,4,5,7,8), \\ n_{g+j2}\cdot\dfrac{\left(n_{gi+2}\cdot\dfrac{m_{gij2}}{m_{gi+2}}+n_{g+j2}\dfrac{m_{gij2}}{m_{g+j2}}+\delta_{ij}\right)}{\sum_i n_{gi+2}\cdot\dfrac{m_{gij2}}{m_{gi+2}}+n_{g+j2}+\delta_{+j}}, & \text{if T기 무응답} \\ & (g=1,3,4,5,7,8), \\ w\times n^{(1)}_{2ij2}+(1-w)\times n^{(2)}_{2ij2}, & \text{if 무응답 } (g=2), \\ w\times n^{(1)}_{6ij2}+(1-w)\times n^{(2)}_{6ij2}, & \text{if 무응답 } (g=6) \end{cases} \tag{3.1}$$

이다. 식 (3.1)에서 정의한 E-STEP은 Park과 Brown (1994)의 연구처럼 $(t, t+1)$ 가운데 한 조사기간에 대해서만 무응답이 발생했을 때도 적용이 가능하다. 식 (3.1)의 E-STEP에서 정의

한 의사-관찰 빈도수는 앞서 제시한 $m_{+++2} = n_{+++2} + \triangle$보다 더 현실적인 가정인 $m_{+++2} = n_{+++2}$를 만족시키며 조사무응답과 설계무응답이 발생하는 그룹에서의 제약조건을 만족한다. 즉, 조사무응답이 발생할 때 $(t, t+1)$에 무응답인 경우는 각각 $m_{g+j2} = n_{g+j2}, m_{gi+2} = n_{gi+2}$ $(g = 1, 3, 4, 5, 7, 8)$를 만족하며 설계무응답이 발생하는 그룹 2와 6의 경우에는 각각의 제약조건 $m_{gi+2} = n_{gi+2}$, $m_{g+j2} = n_{g+j2}$ (g=2,6)을 만족시키면서 대체된다. 그리고, 표본단위(sample unit)별로 무응답을 대체시킨 2개 그룹(그룹 2와 6)은 대체된 결과로 만들어진 완전한 자료를 가중치에 의해 조절해서 해당 그룹을 대표하는 정보로 이용하도록 한다.

이제 무응답 칸 $\pi_{gij|2}$에 가정한 Dirichlet 사전분포의 모수 δ_{ij}를 구체화시키는 방법에 대하여 알아보자. 먼저 $\Delta = \sum_g \sum_i \sum_j \delta_{ij}$의 관계에서 Δ를 지정하여 δ_{ij}에 할당해줄 수 있다. Clogg 등 (1991)은 Δ를 고려하는 모형에서의 모수의 수 p로 부여하는 방법을 제안하였는데 이를 본 연구의 무응답 모형에 적용하면 $\delta_{ij} = (\Delta/G) \cdot (n_{+ij1}/n_{+++1})$이 된다. 본 연구에서는 이 방법을 관찰된 자료에 기초한 방법이라 하고 방법1(Method I)로 표기 하였다. 이 방법에 의하면 δ_{ij}는 고정되어 있으며 만약 최대우도추정량에 의해 변방값이 발생하면 관찰된 자료만으로 무응답칸의 기대돗수가 추정된다. Park과 Brown (1994), 박태성과 이승연 (1998)은 δ_{ij}를 할당하는데 있어서 해당 무응답 칸에 대응하는 관찰된 칸의 빈도에 비례하도록 할당하는 방법을 제안하였다. 그러나 현실적으로 4-8-4 교체표본조사에서 발생하는 무응답의 형태에 비추어 볼 때 이 방법은 항상 타당한 것은 아니다. 가령, 취업여부나 소득에 대한 정보를 얻는 조사가 진행 될 때 응답자들의 성향과 무응답자들의 성향은 다를 것이다. 응답자들 중에는 미취업된 사람들보다 취업된 사람들이 더 많겠지만, 무응답자들은 취업된 사람들보다 미취업된 사람들이 더 많을 것이다. 이처럼 응답자들과 무응답자들의 행태가 다른 상황에서 무응답칸에 주변합을 응답자들만의 정보를 이용해서 할당 할 경우에 현실적으로 부합되지 않는 결과를 초래할 수도 있다. 또, 무응답의 무시할 수 없는 정도가 커질수록, 무응답률이 커질수록 응답칸의 정보만을 이용하는 것보다 대체된 무응답칸의 정보까지 활용하는 것이 정확할 것으로 판단된다. 따라서 본 논문에서는 응답한 자료만을 이용하는 것이 아니라 반복을 통해 추정된 응답과 무응답 칸의 정보를 이용하는 방법을 추가적으로 고려하였다. 이는 Choi (2005)와 최보승 등 (2007)이 제안한 방법을 4-8-4 교체표본조사의 방법에 응용한 것으로써 δ_{ij}의 값을 할당하는데 있어서 δ_{ij}가 반복수행의 응답과 무응답의 대체값에 따라 변화되는

$$\delta_{ij} = \frac{K}{G} \frac{m_{+ij+}}{m_{++++}}$$

을 사용하였다. 이를 본 연구에서는 대체값에 기초한 방법이라 하고 방법2(Method II)로 표기하였다.

M – STEP: E-STEP에서 구한 의사-관찰빈도수를 관찰된 빈도수로 간주하고 무시할 수 없는 무응답모형에 대한 사후확률밀도함수를 최대화시키는 단계이다.

$$\begin{aligned} &\sum_{g}^{6}\sum_{i}^{I}\sum_{j}^{J} z_{(gij1)v} \cdot {n^*}_{gij1} + \sum_{g}^{8}\sum_{i}^{I}\sum_{j}^{J} z_{(gij2)v} \cdot {n^*}_{gij2} \\ &= \sum_{g}^{6}\sum_{i}^{I}\sum_{j}^{J} z_{(gij1)v} \cdot m_{gij1} + \sum_{g}^{8}\sum_{i}^{I}\sum_{j}^{J} z_{(gij2)v} \cdot m_{gij2}. \end{aligned} \tag{3.2}$$

M-STEP에서 사용하는 모형은 $(XY_t, XY_{t+1}, Y_tR, Y_{t+1}R, Y_tY_{t+1})$이고 적합방법은 Bishop 등 (1975)이 제안한 반복비율적합(iterative proportional fitting: IPF) 알고리즘을 이용하였고 반복수행 과정에서 연속적으로 계산된 로그사후우도값의 차이가 10^{-6}일 때까지 반복 계산을 수행하였다.

Y_t와 Y_{t+1}이 오직 2개의 값만을 가지고 $Y_t = Y_{t+1} = 1$은 취업을 의미하고 $Y_t = Y_{t+1} = 2$는 미취업을 의미할 때, 취업상태와 무응답 지시 변수간의 상호작용 효과 β^{Y_tR}, $\beta^{Y_{t+1}R}$를 가지고 취업상태와 무응답 여부간의 관계를 살펴볼 수 있으며 무응답 체계가 무시할 수 없는 무응답을 따르는가에 대한 가정을 검증할 수 있다. 각 조사 시점 $(t, t+1)$에서 취업여부와 무응답여부의 오즈비를 θ라 할 때 $\beta_{11}^{Y_tR} = \log(\theta_{(g)1(j)1})/4$, $\beta_{11}^{Y_{t+1}R} = \log(\theta_{(gi)11})/4$가 된다. 이 때 $\beta_{11}^{Y_{t+1}R}$의 경우 오즈비는 $\theta_{(gi)11} = (m_{gi11}/m_{gi12})/(m_{gi21}/m_{gi22})$이 되는데 현실적으로 $i = 14$일 때 $(t, t+1)$에 (취업, 취업)이 될 확률은 무응답자보다 응답자가 더 높을 것이고, (취업, 미취업)이 될 확률은 무응답자가 응답자보다 더 높을 것이다. 그리고 $i = 2$이면 $(t, t+1)$에 (미취업, 취업)이 될 확률은 무응답자보다 응답자가 더 높을 것이고, (미취업, 미취업)이 될 확률은 무응답자가 응답자보다 더 높을 것이다. 따라서, 오즈비가 크면 클수록 무응답할 확률이 무응답값에 영향을 미칠 것이라는 무시할 수 없는 상황이 현실적으로 더 타당하다 할 수 있다.

3.2. 교체그룹 편향 추정

무응답에 대한 대체를 수행한 후에 교체그룹편향을 추정하고 $(t, t+1)$의 칸 확률과 고정된 시점에서의 주변확률을 추정하는 방법에 대하여 알아보자. $(t, t+1)$의 칸 확률 및 주변확률의 추정은 8개 교체그룹의 면접시점이 상이해서 발생하는 편향을 정확히 추정하고 편향이 제거된 효과를 얼마나 제대로 추정하는지를 의미한다. 따라서, 정확하게 교체그룹편향을 추정하고 그 편향을 제거한 효과를 제대로 추정하는 것은 매우 중요한 문제이다. 우선, 고려하는 모형 $(XY_t, XY_{t+1}, Y_tR, Y_{t+1}R, Y_tY_{t+1})$에서 면접시점과 관련 있는 편향과 편향이 제거된 효과를 구분해 보자. 편향이 제거된 효과는 주어진 (i, j)칸에 대해서

$$\begin{aligned} m_{gij+} &= m_{gij1} + m_{gij2} \\ &= \exp\left(\beta_g{}^X + \beta_{gi}{}^{XY_T} + \beta_{gj}{}^{XY_{T+1}}\right) \cdot (\text{Other Effect}_{ij}) \\ &= \exp(\text{Group Effect}_g) \cdot (\text{Other Effect}_{ij}) \end{aligned} \tag{3.3}$$

이다. 다음으로 $\log(m_{gij+})$를 모든 그룹에 대해서 합하면 $\sum_{g=1}^{G} \beta_g{}^X = \sum_{g=1}^{G} \beta_{gi}{}^{XY_T} = \sum_{g=1}^{G} \beta_{gj}{}^{XY_{T+1}} = 0$이므로

$$\begin{aligned} \sum_{g=1}^{G} \log(m_{gij+}) &= \sum_{g=1}^{G} \left(\beta_g{}^X + \beta_{gi}{}^{XY_T} + \beta_{gj}{}^{XY_{T+1}}\right) + G \cdot \log(\text{Other Effect}_{ij}) \\ &= G \cdot \log(\text{Other Effect}_{ij}) \end{aligned}$$

이다. 즉, 편향이 제거된 효과는 다음과 같이 추정 할 수 있다.

$$\log(\text{Other Effect}_{ij}) = \sum_{g=1}^{G} \frac{\log(m_{gij+})}{G}. \tag{3.4}$$

다음으로 면접시점과 관련있는 편향에 대해 살펴보자. 식 (3.3)와 (3.4)에서 그룹편향과 관련된 효과를

$$\beta_g{}^X + \beta_{gi}{}^{XY_T} + \beta_{gj}{}^{XY_{T+1}} = \log(m_{gij+}) - \log(\text{Other Effect}_{ij})$$

로 추정가능하다. 그리고, 추정된 $\log(m_{gij+}) - \log(\text{Other Effect}_{ij})$이 $\text{Group Effect}_{gij}$이므로 순수한 편향$(X)$은

$$\begin{aligned}
&\beta_g^X + \beta_{gi}{}^{XY_t} + \beta_{gj}{}^{XY_{t+1}} = \text{Group Effect}_{gij} \\
\Leftrightarrow\ &\sum_{i=1}^{I}\sum_{j=1}^{J} \beta_g{}^X + \beta_{gi}{}^{XY_t} + \beta_{gj}{}^{XY_{t+1}} = \sum_{i=1}^{I}\sum_{j=1}^{J} \text{Group Effect}_{gij} = \text{Group Effect}_{g++} \\
\Leftrightarrow\ &IJ\beta_g^X = \text{Group Effect}_{g++} \quad \left(\because \sum_{i=1}^{I} \beta_{gi}{}^{XY_t} = \sum_{j=1}^{J} \beta_{gj}{}^{XY_{t+1}} = 0\right) \\
\Leftrightarrow\ &\beta_g^X = \frac{\text{Group Effect}_{g++}}{IJ}
\end{aligned}$$

로 추정가능하다. 지금까지 편향과 편향을 제거한 효과추정에 대해 살펴보았다. 이제 식 (3.3)와 (3.4)에 의해 연속적인 두 조사기간에 대한 정확한 칸의 확률을 구할 수 있으며 그것을 바탕으로 고정된 시점에서 Y_t와 Y_{t+1}에 대한 주변확률을 구할 수 있다.

4. 모의실험

4-8-4 교체표본설계를 이용한 조사에서 발생하는 두 가지 형태의 무응답을 세 가지 방법으로 대체하고 대체 이후의 결과를 비교해 보고자 한다. 현실적으로 교체표본조사의 면접시점, 각 조사대상자들이 속한 그룹번호 그리고 응답여부에 대한 자료를 구하는 것은 매우 어려운 일이므로 취업여부를 조사하는 상황으로 가정하고 모의실험을 한다. 즉, t월과 $t+1$월에 조사대상의 그룹화된 반응값 Y_t, Y_{t+1}은 각각 취업이면 1, 미취업이면 2로 표기하도록 한다.

조사에서 얻을 수 있는 네 가지 변인(조사대상자들이 속한 그룹번호 또는 면접시점, t월의 취업상태, $t+1$월의 취업상태 그리고 응답여부)을 고려하기 위해 $8 \times 2 \times 2 \times 2$ 형태의 자료를 모형 $(XY_t, XY_{t+1}, Y_tR, Y_{t+1}R, Y_tY_{t+1})$에서 생성한다. 실제로 CPS에서는 인구통계적 항목, 노동력 항목, 산업 및 직업항목, 소득항목의 항목 무응답률이 각각 1.54%, 1.46%, 3.75%, 12.44%였고 이와 같은 상황을 모의실험에 적절히 반영하기 위하여 무응답률을 3%~4%, 5%~6%, 7%~10%으로 가정하였다. 취업과 관련된 무응답은 무시할 수 없는 무응답이므로 대수선형모형에서의 계수 β^{Y_tR}, $\beta^{Y_{t+1}R}$에 해당하는 오즈비가 10배, 5배, 2배가 되도록 하였다. 또, 특정 면접시점에서 8개 그룹은 각각 면접시점이 다르고 면접시점이 클수록 편향이 심할 것이다. 따라서 그룹 1에서 그룹 4까지, 그룹 5에서 그룹 8까지의 편향이 커졌다가 감소

표 4.1: 무응답율 7%~10%에서 교체그룹효과에 대한 평균제곱오차(MSE)와 편향(Bias)

가중치	$\theta_{(gi)11}$ $\theta_{(g)1(j)1}$	그룹	최우추정		방법1		방법2	
			MSE	Bias	MSE	Bias	MSE	Bias
0.5	10 배	1	1.8997	0.7012	0.0427	0.0613	0.0410	0.0574
		2	14.0806	−1.4317	0.0568	0.0593	0.0554	0.0367
		3	1.8349	0.5884	0.0602	0.0443	0.0583	0.0405
		4	2.0596	0.4450	0.1126	−0.0053	0.0773	0.0337
		5	1.9492	0.6960	0.0474	0.0551	0.0447	0.0512
		6	16.8518	−1.8708	0.1867	−0.2241	0.2010	−0.2542
		7	1.8936	0.6040	0.0655	0.0492	0.0639	0.0456
		8	1.5278	0.2678	0.1307	−0.0399	0.1102	−0.0111
	5 배	1	1.0822	0.5483	0.0286	0.0438	0.0289	0.0512
		2	7.8267	−1.2895	0.0400	0.0538	0.0423	0.0233
		3	1.0528	0.5249	0.0374	0.0354	0.0377	0.0436
		4	1.1478	0.4672	0.0569	0.0275	0.0547	0.0343
		5	1.0870	0.5467	0.0288	0.0451	0.0299	0.0527
		6	9.8345	−1.6654	0.1303	−0.2140	0.1452	−0.2380
		7	1.0912	0.5224	0.0390	0.0316	0.0386	0.0394
		8	0.7924	0.3453	0.0788	−0.0220	0.0742	−0.0060
	2 배	1	1.3485	0.6426	0.0189	0.0409	0.0201	0.0495
		2	11.0138	−1.6987	0.0281	0.0474	0.0299	0.0289
		3	1.3207	0.6268	0.0215	0.0251	0.0224	0.0336
		4	1.4068	0.6216	0.0314	0.0222	0.0325	0.0277
		5	1.3825	0.6482	0.0186	0.0470	0.0198	0.0555
		6	12.7273	−2.0442	0.0993	−0.2150	0.1165	−0.2410
		7	1.3744	0.6322	0.0243	0.0313	0.0254	0.0398
		8	1.2150	0.5712	0.0449	0.0012	0.0460	0.0065

하는 형태를 띄므로 이에 부합되게 그룹 주효과를 $\mathbf{X} = (0.3, 0.1, -0.1, -0.15, 0.3, 0.1, -0.1, -0.1, -0.35)$로 부여하였다.

이러한 상황에 맞게 해당되는 $\beta = (\beta^X, \beta^R, \beta^{Y_t}, \beta^{Y_{t+1}}, \beta^{Y_tY_{t+1}}, \beta^{Y_tR}, \beta^{Y_{t+1}R})$를 지정해 주고 각 칸의 확률 $\pi=(\pi_{1111}, \pi_{1121}, \pi_{1211}, \pi_{1221},\cdots,\pi_{8112}, \pi_{8122}, \pi_{8212}, \pi_{8222})$를 구하였다. 그리고, 그 확률을 바탕으로 표본수 N이 2400인 다항분포 Multinomial(N,π) 를 따르는 확률변량 $\mathbf{n} =(n_{1111}, n_{1121}, n_{1211}, n_{1221},\cdots n_{8112}, n_{8122}, n_{8212}, n_{8222})$을 생성하였다. 이렇게 자료를 생성한 후에는 실제로 관찰되는 빈도수 $\mathbf{n}_{obs} = (n_{1111}, n_{1121}, n_{1211}, n_{1221},\cdots, n_{8111}, n_{8121}, n_{8211}, n_{8221}; n_{11+2}, n_{12+2}, n_{1+12}, n_{1+22},\cdots,n_{81+2}, n_{82+2}, n_{8+12}, n_{8+22})$ 가운데 $(t, t+1)$의 각각 무응답을 $n_{g112}/2 + n_{g122}/2 = n_{g1+2}$, $n_{g212}/2 + n_{g222}/2 = n_{g2+2}$; $(n_{g112} - n_{g112}/2) + (n_{g212} - n_{g212}/2) = n_{g+12}$, $(n_{g122} - n_{g122}/2) + (n_{g222} - n_{g222}/2) = n_{g+22}$로 생성하였다. 반복수는 표본수 N에 따라 각각 500번으로 하였다.

설계 무응답이 발생하는 그룹에 대해서는 표본단위별로 대체시키고 해당 그룹의 정보로 사용하기 위해 가중치로 조절하였는데, 모의실험에서 정한 가중치는 0.1, 0.5, 0.9이다. 그런데, 가중치에 따라 제시된 결과에는 큰 차이를 보이지 않았으므로 t달과 $t+1$달의 대체값

표 4.2: 가중치 0.5일 때 편향을 제거한 칸 확률에 대한 평균제곱오차(MSE)와 편향(Bias)

무응답률	$\theta_{(gi)11}, \theta_{(g)1(j)1}$	최우추정		방법1		방법2	
		MSE	Bias	MSE	Bias	MSE	Bias
3%~4%	10배	4.6052	0.6920	0.1946	0.1967	0.1730	0.1904
	5배	3.9463	0.6674	0.1315	0.1469	0.1381	0.1529
	2배	3.8127	0.6522	0.0610	0.0822	0.0729	0.0896
5%~6%	10배	7.2585	0.7717	0.1689	0.1931	0.1573	0.1889
	5배	5.2426	0.6972	0.1342	0.1506	0.1383	0.1557
	2배	1.6414	0.5312	0.0930	0.1025	0.1068	0.1090
7%~10%	10배	7.6707	0.7817	0.2156	0.2294	0.2003	0.2184
	5배	4.1765	0.5957	0.1094	0.1366	0.1121	0.1397
	2배	5.3910	0.6608	0.0657	0.0735	0.0789	0.0778

표 4.3: 무응답율 7%~10%에서 주변확률에 대한 균제곱오차(MSE)와 편향(Bias)

가중치	$\theta_{(gi)11}$ $\theta_{(g)1(j)1}$	(time,status)	최우추정		방법1		방법2	
			MSE	Bias	MSE	Bias	MSE	Bias
0.5	10 배	(T:E)	0.0004261	0.0178866	0.0001419	0.0088277	0.0000825	0.0063982
		(T:U)	0.0004261	−0.0178870	0.0001419	−0.0088280	0.0000825	−0.0063980
		(T+1,E)	0.0004669	0.0187617	0.0001761	0.0099869	0.0001052	0.0073463
		(T+1,U)	0.0004669	−0.0187620	0.0001761	−0.0099870	0.0001052	−0.0073460
	5 배	(T:E)	0.0003497	0.0149984	0.0001037	0.0068391	0.0000869	0.006138
		(T:U)	0.0003497	−0.0149980	0.0001037	−0.0068390	0.0000869	−0.0061380
		(T+1,E)	0.0003976	0.0164530	0.0001351	0.0085281	0.0001138	0.0077334
		(T+1,U)	0.0003976	−0.0164530	0.0001351	−0.0085280	0.0001138	−0.0077330
	2 배	(T:E)	0.0005394	0.0183962	0.0000932	0.0054664	0.0000939	0.0056085
		(T:U)	0.0005394	−0.0183960	0.0000932	−0.0054660	0.0000939	−0.0056080
		(T+1,E)	0.0006713	0.0218277	0.0001333	0.0083711	0.0001373	0.0086054
		(T+1,U)	0.0006713	−0.0218280	0.0001333	−0.0083710	0.0001373	−0.0086050

에 대한 정보를 동등하게 사용하고자 0.5로 선택하였다. 무시할 수 없는 무응답모형으로 무응답을 대체한 후 계산된 결과들이 표 4.1에서 4.4까지 제시되어있다. 표 4.1은 무응답률이 7%~10%일 때 교체그룹효과에 대한 평균제곱오차와 편향을 정리한 것이다. 무시할 수 없는 무응답 상황에서 무시할 수 없는 무응답모형을 이용한 경우 최대우도방법의 평균제곱오차와 편향은 베이지안 방법을 이용한 경우보다 크다는 것을 알 수 있다. 또, 무시할 수 없는 무응답 상황을 나타내는 오즈비가 커질수록 방법2는 방법1에 비해 평균제곱오차와 편향이 작아짐을 볼 수 있다. 이러한 경향들은 무응답률이 증가하는 경우에도 동일하게 나타난다.

표 4.2는 편향을 제거한 칸의 확률(또는 효과)에 대한 결과로서 네 칸에 대한 평균제곱오차와 편향의 평균값이다. 무시할 수 없는 무응답 상황에서 무시할 수 없는 무응답모형을 이용한 경우 최대우도방법의 평균제곱오차나 평향은 베이지안 방법을 이용한 경우보다 크다는 것을 알 수 있다. 또, 무시할 수 없는 정도가 커질수록 방법2는 방법1에 비해 평균제곱오차와 편향이 작아짐을 볼 수 있다.

표 4.4: 설계무응답이 발생하는 그룹의 (미취업,미취업)칸에서 변방값 발생회수

$\theta_{(gi)11}$ $\theta_{(g)1(j)1}$	가중치	설계무응답이 발생하는 그룹								
		3%~4%			5%~6%			7%~10%		
		ML	M-I	M-II	ML	M-I	M-II	ML	M-I	M-II
10배	0.1	158	52	16	135	52	10	124	39	1
	0.5	142	56	17	122	47	6	116	37	1
	0.9	146	58	15	139	45	4	123	32	1
5배	0.1	139	13	7	139	6	0	135	3	0
	0.5	127	13	9	118	6	1	119	1	0
	0.9	146	14	9	135	8	4	137	2	0
2배	0.1	229	0	0	153	1	3	206	0	0
	0.5	174	0	0	121	0	2	177	0	0
	0.9	201	0	0	136	0	0	199	0	0

$\theta_{(gi)11}$ $\theta_{(g)1(j)1}$	Weight	8개 교체 그룹								
		3%~4%			5%~6%			7%~10%		
		ML	M-I	M-II	ML	M-I	M-II	ML	M-I	M-II
10배	0.1	0	0	0	0	0	0	0	0	0
	0.5	0	0	0	1	0	0	1	0	0
	0.9	0	0	0	1	0	0	1	0	0
5배	0.1	5	0	0	2	0	0	0	0	0
	0.5	6	0	0	0	0	0	1	0	0
	0.9	4	0	0	2	0	0	0	0	0
2배	0.1	43	0	0	23	0	0	8	0	0
	0.5	29	0	0	15	0	0	5	0	0
	0.9	34	0	0	16	0	0	10	0	0

ML: 최우추정, M-I: 방법1, M-II: 방법2

표 4.3에 제시되어 있는 주변확률에 대한 추정결과를 살펴보면 무시할 수 없는 상황에서 무시할 수 없는 무응답모형을 적합시킨 결과 최대우도방법을 사용하였을 때의 평균제곱오차와 편향이 상대적으로 베이지안 방법을 이용하였을 때보다 크다는 것을 알 수 있다. 결론적으로 무시할 수 없는 상황에서 베이지안을 이용한 방법1과 방법2는 최대우도추정법보다 작은 평균제곱오차와 편향을 제시하고 있으며 무시할 수 없는 정도가 커질수록 방법2의 평균제곱오차와 편향이 방법1 보다 더 작다.

표 4.4는 무응답 대체이후에 $(t, t+1) =$ (미취업, 미취업)칸에서 발생하는 변방값 회수를 나타내고 있다. 이를 통해서 면접시점에 의해 발생한 편향과 그 편향을 제거한 후 칸 확률을 추정할 때 최대우도방법이 베이지안 방법에 비해 큰 평균제곱오차와 편향을 제공하는 근거를 알 수 있다. 표 4.4에서 위쪽의 표는 설계무응답이 발생하는 교체그룹에서 최대우도방법을 사용할 경우 변방값 발생회수를 정리한 것으로 (미취업, 미취업)칸의 추정값을 0으로 부여하는 회수가 베이지안에 비해 월등히 많음을 알 수 있다. 또, 무응답률이 높아질수록 베이지안 방법은 (미취업, 미취업)칸의 추정값을 0으로 부여하는 회수가 적어지고 특히 방법2는 방법1에 비해 매우 작음을 알 수 있다. 따라서, 무응답률이 높은 경우 설계 무응답의 관점에서는 방

법2가 더 적절하다고 판단된다. 표 4.4의 아래쪽 표는 8개 모든 교체 그룹에 대해 무응답을 대체한 후 변방값이 발생하는 회수가 제시되어 있는데, 두 가지 베이지안 방법은 최대우도추정법과는 달리 500번 반복 중에 한번도 (미취업, 미취업)칸의 추정값을 0으로 부여하지 않음을 볼 수 있다. 즉, 설계무응답이 생기는 교체그룹에 대해, 변방값을 자주 발생시키는 최대우도방법의 불안정성을 응답과 무응답의 기대값에 기초하는 방법2를 이용해서 해결 할 수 있고 방법1보다 방법2가 더 안정적임을 알 수 있었다.

5. 결론 및 토의

본 연구의 목적은 4-8-4 교체표본조사에서 발생하는 항목무응답을 대체하고, 면접시점이 다르기 때문에 발생하는 편향, $(t, t+1)$의 칸의 확률 및 고정된 시점에서의 주변확률을 추정하는 것이다.

두 가지 종류의 항목무 응답을 대체하기 위해 모형에 기반한 방법을 사용하였으며 무응답이 발생하는 항목의 현실성을 고려해서 무시할 수 없는 무응답으로 가정하였다. 그리고 GEM 알고리즘을 이용해서 무응답을 대체할 때, M-STEP에서는 최대우도추정방법, 베이지안 방법을 사용하였는데, 베이지안 방법으로 관찰값에 기반한 추정방법(방법1), 기대값에 기초한 추정방법(방법2)을 사용하였다. 무응답을 대체한 후 편향과 편향이 제거된 칸의 확률(효과)을 추정한 결과 무시할 수 없는 상황에서는 베이지안 방법의 평균제곱오차와 편향이 최대우도방법보다 더 작았으며, 무시할 수 없는 정도가 커질수록 방법1보다 방법2가 더 정확하게 추정하였다. 그리고 표본의 크기가 커질수록 베이지안 추정방법 방법1과 방법2는 균제곱오차와 편향이 감소하였지만 최대우도방법은 그러한 경향을 보이지 않았다. 또, 고정된 시점에서의 취업/비취업률에 대한 주변확률을 추정한 결과 최대우도추정법보다 베이지안 방법이 더 작은 균제곱오차와 편향을 나타냈고, 무시할 수 없는 정도가 클 경우에 방법2가 더 정확한 결과를 제공하였다.

설계무응답이 발생하는 그룹의 (미취업, 미취업)칸 추정치를 구할 때 최대우도추정법을 적용할 경우 변방값이 종종 발생함을 볼 수 있었다. 이러한 최대우도법의 불안정성을 본 연구에서 제안한 방법2를 이용해서 해결 할 수 있었다. 특히, 무응답률이 커질수록 방법2는 최대우도추정법이나 방법1에 비해 변방값이 발생하는 회수가 매우 적었다.

현실적으로 몇개의 중요한 항목에 대해 응답을 하지 않은 항목 무응답이 무시할 수 있는 것인지 아니면 무시할 수 없는지를 판단하는 절대적인 기준은 없으므로 무시할 수 없는 무응답의 정의에 따라 조사중에 발생하는 무응답의 종류를 판단해야 할 것이다. 그리고 무시 할 수 없는 무응답이 발생할 경우 조사를 통해 얻은 응답자료와 무응답 자료를 기초로 무시 할 수 없는 정도를 측정 할 수 있는 방법에 대한 연구가 진행되어야 할 것이다.

참고문헌

박태성, 이승연 (1998). 무응답을 포함하는 범주형 자료의 분석, <응용통계연구>, **11**, 83–95.

최보승, 박유성, 이동희 (2007). 무시할 수 없는 무응답을 갖는 예비조사자료를 이용한 선거 예측, *Journal of the Korean Data Analysis Society*, **6**, 785–794.

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.

Bishop, Y. M., Fienberg, S. E. and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*, MIT Press, Cambridge

Bonetti, M., Cole, B. F. and Gelber, R. D. (1999). A method-of-moments estimation procedure for categorical quality-of-life data with nonignorable missingness, *Journal of the American Statistical Association*, **94**, 1025–1034.

Choi, B. (2005). *Bayesian analysis for incomplete two-way categorical table with nonignorable nonresponse*, unpublished Ph.D dissertation, Korea University, Dept. of Statistics.

Clogg, C. C., Rubin, D. B., Schenker, N., Schultz, B. and Weidman, L. (1991). Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression, *Journal of the American Statistical Association*, **86**, 68–78.

Conaway, M. R. (1992). The Analysis of repeated categorical measurement subject to nonignorable nonresponse, *Journal of the American Statistical Association*, **87**, 817–824.

David, M., Little, R. J. A., Samuhel, M. E. and Triest, R. K. (1986). Alternative methods for CPS income imputation, *Journal of the American Statistical Association*, **81**, 29–41.

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, Series B*, **39**, 1–38.

Fienberg, S. E. and Stasny, E. A. (1983). Estimating monthly gross flows in labour force Participation, *Survey Methodology*, **9**, 77–102.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd edition, John Wiley & Sons, New York.

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

Paul, E. C. and Lawes, M. (1982). Characteristics of respondent and nonrespondent households in the Canadian labour force survey, *Survey Methodology*, **8**, 48–85.

Smith, P. W. F., Skinner, C. J. and Clarke, P. S. (1999). Allowing for non-ignorable nonresponse in the analysis of voting intension data, *Applied Statistics*, **48**, 563–577.

Stasny, E. A. (1986). Estimating gross flows using panel data with nonresponse: An example from the Canadian labour force survey, *Journal of the American Statistical Association*, **81**, 42–47.

Stasny, E. A. (1991). Hierarchical models for the probabilities of a survey classification and nonresponse: An example from the national crime survey, *Journal of the American Statistical Association*, **86**, 296–303.

[2008년 2월 접수, 2008년 3월 채택]

Nonignorable Nonresponse Imputation and Rotation Group Bias Estimation on the Rotation Sample Survey*

Bo-Seung Choi[1] Dae Young Kim[2] KeeWhan Kim[3] You Sung Park[4]

ABSTRACT

We propose proper methods to impute the item nonresponse in 4-8-4 rotation sample survey. We consider nonignorable nonresponse mechanism that can happen when survey deals with sensitive question (*e.g.* income, labor force). We utilize modeling imputation method based on Bayesian approach to avoid a boundary solution problem. We also estimate a interview time bias using imputed data and calculate cell expectation and marginal probability on fixed time after removing estimated bias. We compare the mean squared errors and bias between maximum likelihood method and Bayesian methods using simulation studies.

Keywords: Imputation, nonignorable nonresponse, rotation sampling survey, EM algorithm.

* This research was supported by a Korea University Grant.

1) Research professor, Institute of Statistics, Korea University, Seoul 136-701, Korea.
E-mail: cbskust@korea.ac.kr

2) Graduate student, Dept. of Statistics, Pennsylvania state university, PA 16802, U.S.A.
E-mail: sas2000@hanmail.net

3) Associate Professor, Dept. of Information and Statistics, Korea University, Chung-Nam 339-700, Korea.
E-mail: korpen@korea.ac.kr

4) Corresponding author. Professor, Dept. of Statistics, Korea University, 1 Anam-Dong, Sungbuk-Gu, Seoul 136-701, Korea.
E-mail: yspark@korea.ac.kr