

연구논문

출구조사에서 무응답 메커니즘 식별 및 대체에 대한 연구:

제19대 국회의원 선거 출구조사에의 응용

A Study on the Nonresponse Mechanism and Imputation for the Exit Poll: The Application of Exit Polls of the 19th Assembly Election

전새봄^{a)} · 이수정^{b)} · 박유성^{c)}

Saebom Jeon · Soojung Lee · Yousung Park

무응답은 출구조사 예측실패의 가장 큰 원인 중 하나로, 이러한 무응답 문제를 해결하기 위한 많은 연구들이 진행되었다. 대표적인 무응답 대체(imputation)방법으로, 조사기관에서 이용하는 사후층화법, 무응답 메커니즘 가정에 따른 ML(maximum likelihood) 추정법, 그리고 베이지안(Bayesian) 추정법 등이 있다. 그러나 기존의 무응답 대체방법들은, 무응답 메커니즘을 판별할 수 있는 명확한 기준의 부재와 MNAR(missing not at random) 가정하에서 발생하는 변방값 문제(boundary solution)로 인해 한계가 있다. 본 논문에서는 무응답 메커니즘의 정의 및 무응답 메커니즘에 따른 변방값 문제 발생의 특성을 이용하여 MAR(missing at random)과 MNAR을 판별하는 새로운 방법을 제안하고 있다. 또한 MNAR 가정하에서 발생하는 변방값 문제를 해결하기 위한 새로운 ML 추정방법을 제안하고 있다. 이를 2012년 제19대 국회의원 출구조사 자료를 이용하여 실증분석하였다. 그 결과, 본 연구방법에서 제안한 메커니즘 판정 및 추정 결과가, 기존의 여론조사기관의 예측결과 및 ML, 베이지안 방법보다 더 우수한 예측 정확도를 보였다.

a) 고려대학교 세종캠퍼스 응용통계학과, 초빙교수

b) 고려대학교 통계학과 석사

c) 교신저자(corresponding author): 고려대학교 통계학과 교수 박유성.

E-mail: yspark@korea.ac.kr

주제어: 무응답 매커니즘, MAR, MNAR, 무응답 대체, 변방값 문제

Non-response is one of the biggest causes of failure in the exit poll prediction, and many studies have been in progress to solve this non-response problem. Typical non-response imputation methods include post-stratification used by research institutions, ML estimation and Bayesian method according to the non-response mechanism. However, these imputation methods have limits to determine the mechanisms of non-response in the absence of clear standards, and to solve the boundary solution occurred in the MNAR assumption. In this paper, we propose a way to distinguish MNAR from MAR using the definition of the non-response mechanism as well as the characteristics of the boundary solution problems under each mechanism. In addition, we also propose a new ML method to solve the boundary solution problem under MNAR assumption. We demonstrated the empirical analysis using the exit polls of 19th member of congress in 2012. As a result, the mechanism determination proposed in this research showed better prediction compared to the poll institution, ML and Bayesian imputation.

Key words: nonresponse mechanism, MAR, MNAR, imputation, boundary solution

I. 서론

출구조사란 선거 당일에 투표를 마치고 나온 유권자를 상대로 어느 후보를 선택하였는지를 묻는 선거 여론조사의 한 기법이다. 따라서 기본적으로 선거전 여론조사에 비해 정확도가 높을 것으로 기대되지만, 실제 개표결과와

비교해 보면 출구조사 예측이 빗나가는 경우가 종종 있다. 일반적으로 출구조사는 전체 투표소 중 표본 투표소를 선정한 다음 일정 간격의 투표자를 대상으로 성별, 연령, 지지후보자 등을 면접하는 등의 방식으로 조사한다. 한국에서는 1996년 제15대 국회의원 선거에서 전화를 통한 출구조사가 실시되었으며, 그 이후 출구조사의 정확도를 높이기 위해 선거법상 기존에는 500m이내에서 출구조사를 할 수 없었던 것을 선거법을 개정하여 그 제한거리를 300m(1999), 100m(2004)로 완화하였다. 특히 2000년 제16대 국회의원 선거에서 KBS·MBC·SBS 3사가 모두 참여한 사실상 첫 출구조사가 실시된 이래, 최근 2012년 제19대 선거에서는 방송3사가 공동으로 전국적인 출구조사를 실시하였다. 그러나 이러한 노력에도 불구하고 출구조사에서의 예측실패 논란은 끊임없이 제기되어 왔다. 이러한 예측실패는 표면적으로는 예측오차가 실제보다 과소추정되었기 때문이고, 보다 근본적인 원인으로는 출구조사를 시행할 때 발생하는 다양한 오차에 기인한다고 할 수 있다. 대선에 비해 지역구간 특성이 다르고 지역구 내에서도 투표소별 모집단 분포가 동질하지 않을 수 있다. 또한 고령층 혹은 보수층의 정치적 의사표출에 대한 부담으로 출구조사에 덜 응하거나 응답하지 않는 침묵의 나선효과도 예측실패의 원인으로 꼽힌다.

출구조사의 정확성을 평가하는 데에는 여러 척도가 존재하는데 보통 예측오차를 사용한다. 예측오차는 실제 득표율과 출구조사 자료로부터 추정한 득표율 사이에 얼마나 차이가 있는지를 측정하는 지표이다. 그런데, 출구조사 과정은 여러 단계에 걸쳐 진행되므로 각 과정에서 발생하는 오차를 단계별로 살펴, 전체 예측오차를 전체 투표소 중에서 출구조사를 시행할 표본 투표소를 추출하는 과정에서 생기는 투표소추출오차, 각 투표소에서 조사 대상자를 추출하는 과정에서의 표본추출오차, 무응답에 의한 오차 등으로 구분할 수 있다. 이에 김영원·곽은선(2010) 및 현경보(2005)는 출구조사에서 발생한 예측오차를 출구조사 절차에 따라 오차를 구분하여 출구조사의 정확성을 평가하였다. 즉, 선거구에서의 출구조사 예측치를 $\hat{p}$, 실제 득표율을 p 라 하고, 표본으로 추출된 투표소의 개표결과를 이용해 계산한 득표율을 $\tilde{p}$ 라 하여, 전체

오차($p - \hat{p}$)를 투표소추출오차($\tilde{p} - p$)와 실사오차($\hat{p} - \tilde{p}$)로 구분하였다(김영원 · 박은선 2010).

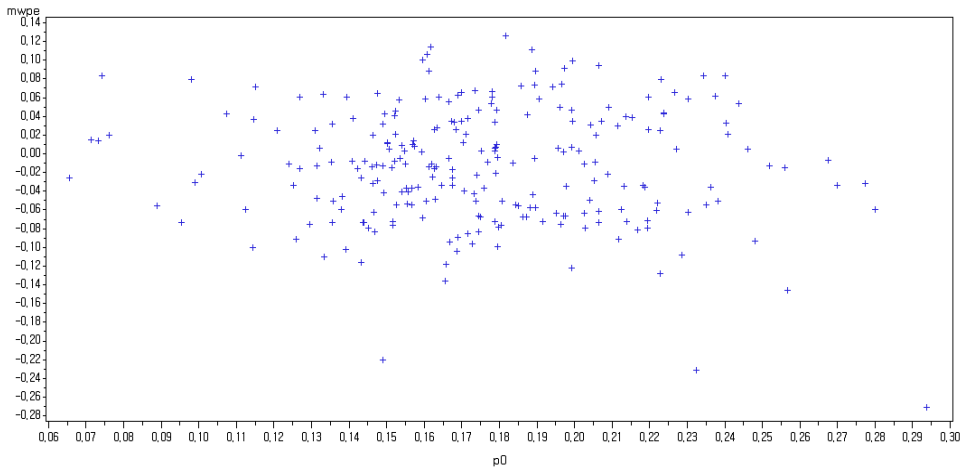
여기서 투표소추출오차가 작다는 것은 표본으로 추출된 투표소가 전체투표소를 설명하는 데에 있어 표본의 대표성이 우수하다는 것을 의미하며, 실사오차가 크다는 것은 표본 투표소에서의 출구조사 결과가 해당 투표소에서의 실제 득표율과 차이가 크다는 것을 의미한다.

그런데, 만약 이들 오차들 간에 서로 상쇄효과가 있다면, 전체 예측오차는 실제보다 작게 나타날 가능성이 있다. 그러므로 최종 예측결과의 차이만을 나타내는 예측오차로는 한계가 있어 이를 세분화하여 정확도를 평가하는 것이 바람직할 것이다.

실사오차는 표본오차(sampling error)와 무응답 등에 의한 비표본오차로 이루어져 있으며 여러 비표본오차 중에서 무응답은 예측실패의 가장 큰 원인 중의 하나로 알려져 있다. 지난 제19대 국회의원 선거 출구조사의 경우 총 817,158명의 출구조사 대상 중 142,432명이 무응답자로 나타나 무응답이 전체 응답의 약 17%를 차지하였다. 여기서 무응답은 우리의 주 관심사가 후보자 득표율임을 고려하여 성별과 연령과 같은 다른 변수는 제외하고 지지후보자에 대한 응답이 없는 경우만 무응답으로 간주하였다. 만약 무응답률을 줄인다면 예측의 정확도가 개선될까라는 물음에 답하기 위하여, Bautista et al. (2007)이 제안한 MWPE¹⁾ (Modified Within Prediction Error)를 사용하여 예측 지지율과 실제 지지율을 비교하고, 이러한 예측오차와 무응답률의 관계를 살펴보았다(Bautista et al. 2007).

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

이때 P_1 , P_2 는 선거구에서 1, 2등 후보자의 실제 공식 득표율이며 $\hat{P}_1$, $\hat{P}_2$ 는 출구조사를 이용한 추정치이다. α 가 1에 가까울수록 MWPE는 0에 가까워지고, 즉 예측지지율과 실제지지율의 차이가 없다고 할 수 있다.



〈그림 1〉 무응답률과 예측오차

〈그림 1〉은 2012년 국회의원 선거 출구조사의 246개 선거구별 무응답률과 MWPE(즉 예측오차)의 산점도이다. 산점도의 흩어진 정도로 보아 무응답률과 MWPE의 관계가 그리 크다고 볼 수 없다. 이는 낮은 무응답률이 더 정확한 예측결과로 이어지지 않는다는 선행 연구들의 결과와 일맥상통한 결과이다(Bautista & Callegaro 2007).

무응답을 줄이기 위해서는 체계화된 조사 시스템, 숙련된 조사원, 검증을 거친 설문지 등 여러 요소를 보완해야 하는데, 이런 시간적, 경제적 투자에 비해 예측오차를 크게 줄이지 못할 수 있다. 따라서 무응답률을 낮추려는 노력도 필요하지만, 불가피하게 발생하는 무응답에 대해 사후적으로 처리하는 방법이 더욱 중요하다. 무응답 처리방법에 따라 예측결과가 다르게 나타나게 되므로, 출구조사의 정확성을 높이기 위해서는 적절한 무응답 처리 및 대체가 보다 중요하기 때문이다.

일반적으로 출구조사를 시행할 때 지지후보자 이외에 성별 및 연령 등과 같은 정보도 함께 기입하는데, 여론조사 기관에서는 대부분 무응답을 처리하기 위해 이러한 정보를 이용하여 사후층화(poststratification)함으로써 보다

정확한 선거예측을 하고자 노력하고 있다. 이외에도 핫덱 대체(hot deck imputation)나 다중대체법(multiple imputation) 등 무응답을 처리하기 위한 다양한 방법들이 연구되고 있다. 보다 정교한 무응답 대체방법을 논의하기 위해, 무응답의 발생체계 및 그에 따른 대체방법을 살펴볼 필요가 있다.

조사 과정에서 발생하는 무응답은 발생체계에 따라 크게 완전임의결측(MCAR; Missing Completely At Random), 임의결측(MAR; Missing At Random), 그리고 비임의결측(MNAR; Missing Not At Random)의 세 가지 무응답 매커니즘으로 구분된다(Little and Rubin, 2002). 무응답 대체결과, 즉 출구조사 예측결과는 이러한 무응답 매커니즘 가정에 따라 달라지게 된다. MCAR을 가정하는 경우는 무응답을 완전히 무시하는 경우와 같으므로 지지 후보자에 대한 응답이 없는 관측치는 삭제하고 나머지를 이용하여 지지율을 계산한다. MAR은 무응답이 관측된 변수에 의존하여 발생하는 경우로 보통 연령, 성별, 인종 등을 무응답에 영향을 미치는 공변량(covariate)으로 두고 무응답을 대체한다. MNAR은 지지후보가 누구냐에 따라 무응답이 발생하는 경우로 일반적으로 추격후보(chasing candidate)에 무응답을 더 부여하는 등의 대체방법을 사용한다. 따라서 출구조사에서 발생하는 무응답을 처리하기 위해서는 먼저, 어떠한 무응답 매커니즘을 가정할 것인지가 우선적으로 고려되어야 할 것이다.

무응답 매커니즘을 적절히 식별하여 가정한 후에는, 해당 매커니즘 가정에 맞는 모형 및 추정방법을 선택해야 할 것이다. 일반적으로 무응답을 대체하기 위해, 무응답 건수를 공변량 및 무응답여부 등으로 설명하는 일반화선형모형(GLM; Generalized Linear Model)이 사용되며, 그에 대한 모수추정방법으로 최대우도 방법과 베이지안 방법이 고려될 수 있다. 그러나, 무응답 매커니즘 가정이 MNAR인 경우, 모수 추정 시 다음의 문제가 발생한다. 먼저, 최대우도추정방법의 경우 변방값(boundary solution) 문제가 발생할 수 있다. 변방값 문제란 무응답 대체 결과 특정 칸에 0의 확률이 부여되는 현상으로, 변방값 문제가 발생하면 Maximum Likelihood Estimator(MLE)는 유일한 해를 갖지 않고 그 결과가 불안정해질 수 있다(Park & Brown 1994). 이에 대

한 대안으로 MNAR 가정하에 변방값 문제를 피하기 위해 계층적 베이지안 및 EM을 이용한 베이지안 추정치가 선호되었다(Foster & Smith 1998; Green & Park 1998; Choi et al. 2009; Park & Choi 2010; Clogg et al. 1991; Park & Brown 1994; Park 1998; Choi et al. 2009; Park & Choi 2010). 그러나 베이지안 추정의 경우, 사전분포 선택이 주관적일 수 있다는 한계가 있다.

따라서 무응답 매커니즘과 변방값 문제의 관점에서 본 연구의 관심사를 정리하면 다음과 같다. 첫째, 무응답 매커니즘 가정이 맞는가? 무응답이 발생했을 때 그 무응답의 매커니즘이 MAR인지 혹은 MNAR인지를 검정할 수 있는 방법이 없다고 밝혀져 왔다(Molenbberghs et al, 2008; Ibrahim et al. 2008). 그러나 출구조사 예측이 가정하는 무응답 매커니즘에 따라 결과가 다르므로, 적절한 무응답 매커니즘을 정의하기 위한 방법이 요구된다. 본 논문에서는 무응답 매커니즘에 따른 변방값의 성격을 이용하여 MAR 모델과 MNAR 모델을 구분하였다(Kim & Park 2014).

둘째, MNAR 가정하에서 발생하는 변방값 문제는 ML이 아닌 베이지안 방법을 이용하면 정말 해결되나? ML 추정치는 변방값 문제에 빠지기 쉬우며, 이를 해결하기 위한 대안으로 EM 및 MCMC를 이용한 베이지안 추정치가 제안되어 왔다. 그러나 Kim & Park (2014)의 시뮬레이션 연구 결과, 모든 베이지안 접근법이 변방값 문제를 해결하는 것은 아니며, MNAR 가정하의 베이지안 추정치가 MAR 가정하의 MLE보다 좋지 않을 수 있음을 보였다. 이에 본 연구에서는 변방값 문제가 발생하는 경우 베이지안 추정치를 이용하는 대신 연령 카테고리 통합을 통해 변방값 문제에서 벗어나는 방법을 제안하고 이를 전통적인 베이지안 방법과 효율을 비교하고자 한다.

이 논문은 총 4개의 절로 구성되어 있으며 제2절에서는 변방값 성격 및 무응답 매커니즘의 정의를 이용하여 무응답 매커니즘 식별 방안을 제안한다. 또한 무응답 대체시 발생하는 연령구간의 선택문제, 모수추정 등을 위한 방법을 제안한다. 제3절에서는 2012년 실시된 국회의원 출구조사 자료를 이용하여 적절한 무응답 매커니즘을 식별하고 무응답 대체방법별 분석결과를 비교함으로써, 위의 두 가지 물음에 대한 결과를 제시하고 논의할 것이다.

II. 무응답 매커니즘과 변방값 문제

1. 무응답 매커니즘

무응답의 처리방법으로 가장 먼저 생각할 수 있는 방법은 무응답을 제거하고 응답한 사람의 자료만을 이용하는 것이다. 과거에 주로 이용했던 이 방법은 변수들 간의 관계를 고려하지 않고 단순히 제거함으로써 자료를 효과적으로 이용하지 못 한다는 단점이 있다. 또한 점점 무응답률이 증가하는 상황에서 단순 제거는 정보의 왜곡을 불러올 수 있다. 이에 최근에는 무응답 대체(imputation)를 통해 완전한 자료(complete data)로 만든 후 이를 분석에 이용하는 방법이 널리 이용되고 있다. 그런데, 어떤 무응답 매커니즘을 가정하느냐에 따라 대체결과가 달라지므로 무응답 매커니즘 결정이 우선되어야 한다.

Little & Rubin(2002)의 구분에 따라, 무응답 발생체계는 완전히 랜덤하게 무응답이 발생하는 완전임의결측(이하 MCAR), 무응답이 성별과 나이와 같은 관측된 변수에 의존하여 발생하는 임의결측(이하 MAR), 그리고 무응답이 관측되지 않은 변수에 의존하는 비임의결측(이하 MNAR)의 매커니즘으로 구분된다. MCAR의 경우 무응답이 완전히 랜덤하게 발생하므로 지지후보자에 대한 응답이 없는 관측치는 삭제하고 응답자 자료만을 이용하여 예측치를 계산한다. MAR은 무응답이 관측된 변수에 의존하여 발생하는 경우로 본 연구에서는 지지성향에 큰 영향을 미치는 연령과 성별을 공변량으로 고려하였다. MNAR은 무응답이 관측되지 않은 변수에 의존하는 경우로 침묵의 나선이론으로 설명될 수 있다. 침묵의 나선이론이란 사람들은 자신의 의견이 사회적으로 우세하거나 지배적이 여론과 일치되면 그것을 적극적으로 표현하지만 그렇지 않으면 침묵을 지키는 성향이 있다는 이론으로서, 출구조사의 경우 자신이 지지하는 후보자가 열세를 보이거나, 보수적 성향인 경우 응답을 회피하는 경향이 있다는 것으로 해석할 수 있다.

이들 무응답 매커니즘을 수식으로 표현하면 다음과 같다. I 개의 범주를 갖는 X 변수, J 개의 범주를 갖는 Y 변수, 그리고 Y 의 무응답 여부를 의미하는

지시변수 R 을 이용하면 X, Y, R 에 대해 셀 빈도수가 y_{ij2} 로 나타나는 $I \times J \times K$ 분할표가 만들어진다. 여기에서 X 는 항상 관찰되고 Y 가 관찰되면 $R=1$, 그리고 관찰되지 않으면 $R=2$ 이다. 하나의 변수에 대해 합한 주변합은 ‘+’ 첨자를 이용하여 나타낼 수 있으며, 따라서 조사단계에서 관측되는 변수는 y_{ij1} 과 y_{i+2} 이다. 분할표에서 관측된 각 셀의 빈도 y_{ijk} 는 모수 π_{ijk} , $N = \sum_{ijk} y_{ijk}$ 를 갖는 다항분포를 따르며 다항분포의 로그우도함수는 다음과 같다.

$$\log L = \sum_i \sum_j y_{ij1} \log \pi_{ij1} + \sum_j y_{+j2} \log \pi_{+j2}$$

일반적으로 MAR, MNAR과 같은 무응답 매커니즘을 설명하기 위해, 로그/로짓 링크함수를 이용한 일반화선형모형(GLM; General Linear Model)이 사용되며, π_{ijk} 의 연결함수(link function)로 로그함수를 이용하면 각 무응답 매커니즘을 다음과 같은 로그선형모형(Log Linear Model)으로 나타낼 수 있다. $m_{ijk} = N \times \pi_{ijk}$ 일 때,

$$\log m_{ijk} = \lambda_X^i + \lambda_Y^j + \lambda_R^k \quad (\text{MCAR}) \quad (2.1)$$

$$\log m_{ijk} = \lambda_X^i + \lambda_Y^j + \lambda_R^k + \lambda_{XY}^{ij} + \lambda_{XR}^{ik} \quad (\text{MAR}) \quad (2.2)$$

$$\log m_{ijk} = \lambda_X^i + \lambda_Y^j + \lambda_R^k + \lambda_{XY}^{ij} + \lambda_{YR}^{jk} \quad (\text{MNAR}) \quad (2.3)$$

여기서 m_{ijk} 는 관찰값 y_{ijk} 에 대해 대응되는 각 셀의 기대빈도이다. 예를 들어 MAR을 가정하는 경우는, 식 (2.2)의 로그선형모형을 이용하여 다항분포의 셀 빈도를 추정하고, 이를 통해 후보자의 득표율을 예측할 수 있다.

이와 같이 무응답 매커니즘 가정에 따라 모형, 모수추정 및 그로 인한 무응답 대체결과는 다르게 나타나는데, 정작 무응답이 발생했을 때 무응답의 매커니즘이 MAR인지 혹은 MNAR인지를 검정할 수 있는 방법이 없다고 알려져 왔다(Molenbberghs et al, 2008; Ibrahim et al. 2008). 이에 본 논문에서는 Kim & Park(2014)이 보여준 무응답 매커니즘에 따른 변방값의 성격 및

MAR과 MNAR의 정의를 이용하여 각 투표소의 무응답 매커니즘이 MAR인지 MNAR인지 구분하였다.

2. 무응답 매커니즘 식별

Kim & Park (2014)이 증명한 이론은 다음과 같다. 우선 X 의 임의의 두 범주 $i, i' (\neq i)$ 에 대해 다음과 같이 정의할 수 있다.

$$\hat{w}(i, i') = y_{i+2} / y_{i'+2}, \quad \hat{w}_j(i, i') = \hat{\pi}_{ij1} / \hat{\pi}_{i'j1}$$

이때, $\hat{\pi}_{ij1}$ 는 ML 추정치이다. 그리고 모든 j 에 대해 $\hat{w}_m(i, i') = \max \hat{w}_j(i, i')$, $\hat{w}_n(i, i') = \min \hat{w}_j(i, i')$ 라고 정의하면,

보조정리1.

MNAR 가정하에 MLE가 모수공간의 변방에 있지 않으면 모든 i, i' 에 대해 $\hat{w}_n(i, i') < \hat{w}(i, i') < \hat{w}_m(i, i')$ 를 만족한다.

MNAR 가정하에서 $I \geq J$ 인 경우, π_{ij1} 의 MLE는 y_{ij1}/N 이므로 관측된 값 y_{ij1} 만으로도 보조정리1에 따라 필요조건을 만족하는지 여부를 확인할 수 있다.

보조정리2.

(i) 실제로 MNAR이 맞는데, MNAR를 가정했다면, 임의의 j 에 대해 $\lambda_Y^j + \lambda_{YR}^{j2}$ 이 충분히 클 때, 모든 (i, i') 에서 $w_n(i, i') < w(i, i') < w_m(i, i')$ 와 $w_n(i, i') = w(i, i') = w_m(i, i')$ 를 만족한다.

(ii) 실제로는 MAR이 맞는데, MNAR를 가정했다면, $|\lambda_{XR}^{i2} + \lambda_{XR}^{i'2}|$ 을 만족하는 경우 몇몇 (i, i') 에서 $w(i, i') < w_n(i, i')$ 또는 $w_m(i, i') < w(i, i')$ 를 만족한다.

즉 실제 MNAR 무응답 매커니즘일 때, MNAR 모형을 올바르게 적용한다면, 보조정리1에서 주어진 필요조건은 항상 만족한다. 그리고 보조정리2의 (i)

에 의해 MNAR 무응답 매커니즘인 경우 MLE는 변방값 문제가 거의 발생하지 않는다. 반면에 실제로는 MAR 매커니즘인데도 불구하고 MNAR 모형을 잘못 적용하면 무응답이 공변량에 강하게 의존한다는 가정하에 MLE는 변방값 문제에 빠지게 된다. 따라서 MNAR 모형을 적용하였을 때 변방값 문제가 발생한다면 MNAR이 아니라 MAR이라 볼 수 있으며, 반대로 변방값 문제가 발생하지 않으면 MNAR이라 볼 수 있다. 이와 같이 Kim & Park(2014)은 MAR 가정하에서는 빈번하게 변방값 문제가 발생하지만 MNAR 가정하에서는 거의 변방값 문제가 발생하지 않음을 증명하였다. 따라서 이러한 무응답 매커니즘에 따른 변방값 발생 성격을 이용하여, 우선적으로 변방값 문제가 발생하는 경우는 MAR, 변방값 문제가 발생하지 않는 경우는 MNAR로 무응답 매커니즘을 잠정적으로 구분하고자 하였다.

그 다음, MAR과 MNAR의 정의를 이용하여 이를 한 번 더 재분류할 수 있다. MAR은 무응답이 연령 및 성별과 같은 설명변수에 의존하는 매커니즘을 의미한다. 따라서 무응답률이 연령에 의존한다는 뚜렷한 증거가 없는 경우는 MNAR이라 볼 수 있을 것이다. 설령 해당 투표소가 앞서 MNAR 가정하에서 변방값 문제가 발생하여 잠정적으로 MAR로 분류되었더라도 보다 근본적인 MAR 정의에 반하는 성격을 보인다면 해당 투표소는 MNAR로 보는 것이 적절할 것이다. 이때, 무응답이 연령에 의존하지 않는다는 판단을 내리는 기준에는 여러 가지가 있을 수 있다. 본 논문에서는 각 투표소에서의 연령별 무응답률 차이를 기준으로 삼았다. 먼저 각 투표소의 연령별 무응답률을 구하여, 5개 연령범주 간 무응답률 중 최대값과 최소값의 차이를 구한다. 그리고 이러한 투표소별 최대 무응답률과 최소 무응답률의 차이를 16개의 지역으로 구분하여 이 차이의 지역별 평균과 표준편차(σ)를 구한다. 지역별 평균에서 일정 크기 이상 벗어나게 작은 경우 지지율이 연령에 의존하지 않는다고 판단하고 MNAR 모델을 적용하기로 한다. 마지막으로 위의 두 기준에 의해 MNAR로 판정된 투표소들 중 무응답 개수가 지나치게 작아 변방값 문제 발생 여부 혹은 무응답률의 차이 비교로 MNAR이라 판정하기 힘든 투표소의 경우는, 이들을 재분류해 MAR이라 판정하였다.

3. 무응답 대체 가중치

Pavia(2010)는 무응답에서 발생하는 편향을 줄이기 위한 방법으로, 무응답을 삭제하고 응답자에 대한 가중치를 적용하는 방법을 제안하였다. 본 연구에서는, 출구조사 데이터가 인구의 형태를 충분히 반영한다는 가정하에서, 선거명부에 등록된 선거인 수를 이용한 가중치 적용 방법(RS; Representative Sample)과 실제 투표한 투표자 수를 이용한 가중치 적용 방법(SD; sampling Design)을 이용하여 대체 결과를 비교 및 보정하였다. l 투표소에서 후보자 i 의 RS와 SD 가중치는 각각 다음과 같다.

$$\text{RS 가중치: } \hat{\pi}_i^{RS} = \frac{\sum_l N_l \hat{\pi}_{il}}{\sum_l N_l}, \quad \text{SD 가중치: } \hat{\pi}_j^{SD} = \frac{\sum_l \frac{N_l}{N\phi_l} \hat{\pi}_{jl}}{\sum_l \frac{N_l}{N\phi_l}}$$

이때 $\hat{\pi}_{jl}$ 은 ML 또는 베이지안 추정치이며 N_l 은 l 투표소에 등록되어 있는 유권자의 수이다. 또한 SD 가중치의 $\phi_l(=n_l/N_l)$ 은 l 투표소에 등록되어 있는 유권자(N_l) 중 투표한 사람의 수(n_l)이다. 즉, RS 가중치는 선거구에 등록된 선거인 수를 고려하고, SD 가중치는 실제 투표소에 투표한 사람 즉 표본에 포함될 확률이 서로 다를 수 있음을 고려한다.

III. 자료 분석

1. 2012 제 19대 국회의원 선거 출구조사 현황

본 논문에서 무응답 대체방법의 비교를 위해 사용한 데이터는 2012년 제 19대 국회의원 선거 출구조사 자료이다. 제19대 국회의원 선거 출구조사는 지상파 방송 3사가 공동조사에 합의하여 구성한 한국방송협회 산하의 방송사 공동예측조사위원회(KEP; Korea Election Poll)에서 미디어리서치, 코리아리서치센터, 그리고 TNS-RI 세 개의 조사기관에 의뢰하여 실시되었다. 2012년 4월 11일 오전 6시부터 오후 5시까지 246개 선거구 내 전체 13,507개 투표소

중 2,496개 투표소에서 이루어졌으며, 응답자 추출방법은 표본투표소로 추출된 투표소에서 출구를 나오는 투표자 기준으로 매 5번째 투표자를 등간격으로 조사한 체계적 추출방법이다.

출구조사 총 응답자 수는 858,324명이며 그 중 성별, 연령 결측치를 제외한 817,158개의 자료를 분석에 이용하였다. 이때 연령은 20대, 30대, 40대, 50대, 60대 이상의 총 5개의 범주로 구분되어 있으며 각각 18~29, 30~39, 40~49, 50~59, 60세 이상을 의미한다. 또한 지지후보자의 경우, 최종 당선 여부가 결국 1, 2위를 제외한 나머지 후보들은 관심사항이 아니므로, 분석에 앞서 이들 1, 2위 후보자를 제외한 나머지 후보자들은 모두 하나의 후보자로 통합하여 총 세 개의 범주로 나누었다. 지지후보자를 밝히지 않은 무응답자는 총 817,158명의 응답자 중 142,432명으로, 전체 응답의 약 17%가 무응답이었다.

서론의 <그림 1>에서 제시하였던 무응답률과 MWPE의 관계에서 보듯, 무응답률이 예측오차에 직접적인 관련이 없으므로, 무응답률을 낮추기보다는 이를 대체하는 것이 중요할 것이다. 한편 예측오차($\hat{p}-p$)를 구성하는 투표소 추출오차($\tilde{p}-p$)와 실사오차($\hat{p}-\tilde{p}$)의 부호가 서로 달라 상쇄된다면 실제 예측오차에 비해 작게 나타날 수 있으므로, 각 오차에 절댓값을 취해 총조사오차(total survey error)를 정의하는 것이 바람직할 것이다. 이때 정확성을 평가하기 위해서 주요 관심대상인 1, 2위 후보자에 대한 예측오차를 함께 고려해야 하므로(Martin et al. 2005), 1, 2위 후보자의 투표소추출오차와 실사오차를 종합한 선거구별 총조사오차

$$(|\hat{p}_1 - \tilde{p}_1| + |\hat{p}_2 - \tilde{p}_2| + |\tilde{p}_1 - p_1| + |\tilde{p}_2 - p_2|)$$

를 정의하였다. 그 결과, 총 246개 선거구의 총조사오차 평균은 6.41%였으며, 각 선거구의 총조사오차 중 실사오차($|\hat{p}_1 - \tilde{p}_1| + |\hat{p}_2 - \tilde{p}_2|$)가 차지하는 평균비율은 67.24%로 실사오차가 총조사오차의 많은 부분을 차지함을 확인할 수 있었다. 이러한 결과는 무응답과 같은 실사오차가 총조사오차를 결정하는 중요한 요인이라는 김영원·곽은선(2010)의 결과와도 일치한다. 따라서 출구조사

자료의 경우, 무응답으로 인한 오차가 상당한 비중을 차지하므로 무응답 대체를 통해 예측결과를 개선할 필요가 있음을 확인할 수 있다.

2. 무응답 매커니즘 판정

2012년 국회의원 선거 출구조사의 무응답률의 평균은 약 17%이며 선거구별로 6.6%에서 29.4%까지 분포되어 있다. 연령과 성별을 기준으로 세분화하여 살펴보면 <표 1>과 같다. 여자일수록, 그리고 연령이 높을수록 무응답률이 높은 것으로 보아 무응답률이 연령과 성별에 의존할 것으로 여겨진다. 따라서, 출구조사시 관측된 연령과 성별을 공변량으로 고려하여 MAR 모형을 적용하였다. 그러나 MNAR 모형 내에 연령과 성별을 모두 고려하게 되면 분할표에서 각 셀의 크기가 지나치게 작아져 대체가 어려워지고 변방값 문제가 발생하여 그에 따라 부정확한 결과를 도출할 수 있어 연령을 공변량으로 하는 모형을 적용하였다.

MNAR 가정하에서 MLE는 유일한 해를 갖지 않으므로 보통 MNAR 모델 하에서 변방값 문제를 해결하는 전통적인 방법은 베이지안 기법을 이용하는 것이다. 베이지안 접근법을 이용하는 방법으로는 계층적 베이지안 모델(Foster & Smith 1998; Green & Park 1998; Choi et al. 2009; Park & Choi 2010)과 EM 알고리즘을 이용하여 사후분포의 최빈값을 추정하는데 베이지안 모델을 이용하는 방법(Clogg et al. 1991; Park & Brown 1994; Park 1998; Choi et al. 2009; Park & Choi 2010)이 있다. 하지만 Kim and Park(2014)은 시뮬레이션을 통해 모든 베이지안 접근법이 변방값 문제를 해결하는 것은 아니며, 베이지안 추정치는 MNAR 모델이라 하더라도 MAR 가정하의 MLE보다 좋지 않을 수 있음을 보였다.

<표 1> 제19대 총선 출구조사의 성별, 연령별 무응답률

(단위: %)

무응답률	20대	30대	40대	50대	60대 이상
남자	9.70	12.18	12.15	17.29	19.93
여자	11.59	14.39	15.75	20.11	30.72

이에 본 논문에서는 변방값 문제를 해결하여 MLE를 사용하는 방법을 제안하고자 한다. 연령을 통합함으로써 변방값 문제가 발생하지 않도록 만들어 준 후, MNAR 가정하의 MLE를 구해 추정치를 계산하는 방법이다. 연령을 통합할 때에는 투표소별로 변방값 문제가 발생하지 않는 조합을 찾아 그 조합대로 연령을 묶어 준다. 기존에 MAR 모델 적용 시 사용하던 각 투표소별 5개의 연령그룹과 3명의 지지후보자로 이루어진 3×5 분할표에서, 연령 카테고리를 통합하여 만들어진 3×3 분할표는 변방값 문제가 발생하지 않으므로 MNAR 모델하에서 MLE를 사용할 수 있다(Baker et al. 1992). 우리의 주 관심사는 연령이 아닌 지지율이므로 이 방법을 이용하면 지지율에 대한 정보손실 없이 변방값 문제를 해결할 수 있다. 게다가 성별이나 연령의 경우 조사원이 눈으로 짐작하여 적는 경우가 많아 정확하지 않다는 면에서 정보의 손실이 적다. 특히 인접한 연령인 경우 더욱 그러하다. 따라서 연령그룹은 측정오차를 가질 수밖에 없는 구조이기에 연령을 통합하여 변방값 문제를 해결하는 것은 정보의 큰 손실 없이 분석을 가능하게 한다.

우선, 무응답 매커니즘과 관련한 변방값 발생 측면에서 보면, 2012년 국회의원 선거 출구조사에서는 총 2,495개 투표소 중 2,389개의 투표소에서 변방값 문제가 발생하고 106개의 투표소에서 변방값 문제가 발생하지 않았으므로, 이들 106개의 투표소를 잠정적으로 MNAR으로 나머지 투표소를 MAR이라고 잠정적으로 판단할 수 있다. 그 다음 MAR과 MNAR의 정의를 이용하여 이들 매커니즘을 재식별하였다. 즉, 변방값 문제가 발생하더라도 무응답율이 연령에 의존한다는 뚜렷한 증거가 없는 경우 MNAR을 적용하였다. 무응답이 연령에 의존한다는 판단을 내리는 기준에는 여러 가지가 있을 수 있으나, 본 논문에서는 각 투표소에서의 연령별 무응답율 차이를 기준으로 삼았다. 즉, 각 투표소의 5개 연령별 무응답률을 구하고 각 투표소별 최대 무응답률과 최소무응답률의 차이가 16개 지역별 평균보다 유의하게 차이가 나게 적다면 지지율이 연령에 의존하지 않는다고 판단하고 MNAR로 재분류하였다. 유의하게 차이가 나는지는 지역 평균에 어느 정도의 불확실성을 감안하여 ‘지역평균-지역표준편차($\mu - \sigma$)’보다 연령별 무응답률 차이를 판단 기준으로 선택하였고, 제19대 국회의원 출구조사에서는 전체 2,495개의 투표소 중 374

개가 이 기준에 해당되었다. 마지막으로 위의 두 기준에 의해 MNAR로 판정된 투표소들 중 무응답 개수가 지나치게 작아 변방값 문제 발생 여부 혹은 무응답율의 차이 비교로는 MNAR이라 판단하기 힘든 투표소들은 MAR로 재분류하였다. 결과적으로, 위의 절차들을 통해 MNAR로 판정된 405개의 투표소 중 62개의 투표소는 MAR 모델로 재분류되었다.

3. 무응답 대체결과 비교

출구조사 자료를 이용한 예상득표율의 정확성을 비교할 때 가장 우선적으로 고려하는 것은 1위와 2위를 올바르게 가려내는지의 여부이다. 따라서 당선자 예측도 중요하지만 통계학적인 관점에서는 1위, 2위 후보자의 득표율을 얼마나 정확하게 예측하였는지가 중요하다. 이에 본 연구에서는 순위를 올바르게 맞췄는지 여부와 함께 추정치의 신뢰구간이 참값을 포함하는지의 여부도 확인하였다. 신뢰구간 내 참값의 포함 여부에 대해서는 1위 후보자와 2위 후보자 각각 추정치의 신뢰구간이 참값을 포함하는지 확인하고, 두 후보자 모두 참값이 신뢰구간 내에 포함되는 경우만을 신뢰구간 내에 정확하게 예측하였다고 판단했다. <표 2>는 무응답 대체방법별 순위 및 신뢰구간 내 예측에 성공 및 실패한 선거구의 수를 나타낸다. 즉, (0,0)은 순위를 올바르게 맞추는 순위예측에 성공하였으며, 신뢰구간 내에 참값을 포함하는 구간예측 역시 성공하였음을 나타낸다. 마찬가지로, (0,1)은 순위예측은 성공하였는데 반해 구간예측은 실패한 선거구의 수를 의미하고, (1,0)은 순위예측은 실패하고 구간예측은 성공한 선거구의 수를 의미한다. 마지막으로, (1,1)은 순위예측과 구간예측 모두 실패한 선거구의 수를 나타낸다.

<표 2>에 제시된 무응답 대체방법의 결과는 9가지이다. 우선, 본 연구결과와의 비교를 위해 KEP 조사기관의 추정득표율을 기준으로 제시하였고, 그다음은 무응답 매커니즘의 분류에 따라 차례대로 제시하였다. 두 번째는 무응답이 완전히 랜덤하게 발생한다고 가정하여 무응답을 제거한 MCAR 가정하의 대체결과를 의미한다. 세 번째와 네 번째 결과는 MCAR 대체결과에 Pavia(2010)가 제안한 가중치를 적용한 것으로 각각 MCAR + rs, MCAR + sd

〈표 2〉 무응답 대체방법별 선거구 수 비교

(0=예측성공, 1=예측실패, 단위: 개)

Imputation Method	(Rank, Confidence Interval)				#of Rank Error
	(0,0)	(0,1)	(1,0)	(1,1)	
KEP	184	45	13	4	17
MCAR	146	78	6	16	22
MCAR + rs	151	75	6	14	20
MCAR + sd	146	78	8	14	22
MAR	184	44	13	5	18
MAR + MNAR(MLE)	193	39	11	3	14
MAR + MNAR(MLE) + rs	195	37	10	4	14
MAR + MNAR(MLE) + sd	190	43	8	5	13
MAR + MNAR(Bayesian)	185	43	12	6	18

로 표기하였다. 다섯 번째는 무응답이 성별과 연령에 의존하는 MAR 가정하의 결과이다. 여섯 번째 이하 결과는 앞서 제안한 매커니즘 식별 결과에 따른 것으로, 2,495개 투표소 중 MAR로 분류된 투표소는 성별과 연령을 고려한 MAR 모형을 적용하고, MNAR로 판정된 투표소에는 추정방법에 따라 각각 달리 표기하였다. 우선 여섯 번째 결과는 연령구간을 선택하여 변방값 문제를 해결한 MLE 결과로 MAR + MNAR(MLE)로 표기하였다. 일곱 번째, 여덟 번째 결과는 MAR + MNAR(MLE) 대체결과에 rs, sd 가중치를 적용한 것이다. 마지막 아홉 번째 결과는 식별된 무응답 매커니즘 가정에 대해 베이지안 추정방법을 사용한 결과이다. 여러 베이지안 방법중 Kim & Park (2014)의 연구에서 가장 우수한 Clogg 방법을 이용하였다.

각 대체방법에 따라 전체 246개 선거구를 1, 2위 후보자의 순위와 신뢰구간을 정확히 예측한 (0,0)의 경우는, MAR + MNAR(MLE) + rs이 가장 많다. 즉, 본 연구에서 제안한 매커니즘 식별방법 및 MNAR하에서 ML추정치에 rs 가중치를 적용한 결과, 1, 2위 순위를 맞춘 동시에 1, 2위 후보자 지지율의 추정치가 신뢰구간 내에 참값을 포함한 선거구가 전체 246개 중 195개이다. 이 경우, 두 후보자의 추정치 모두 신뢰구간 내에 참값을 포함하지만 순위를 맞

추지 못한 선거구는 10개이며, 반대로 신뢰구간 내에 참값을 포함하지는 못하지만 순위는 맞춘 선거구는 37개이다. 그리고 신뢰구간도 벗어나고 순위도 맞추지 못한 선거구는 4개이다.

순위에 대한 예측에 실패한 선거구의 수 역시, 본 연구에서 제안한 결과가 가장 정확성이 높다. 조사기관의 경우 1, 2위 후보자를 맞추지 못한 선거구가 17개 선거구인데 비해, $MAR + MNAR(MLE) + rs$, $MAR + MNAR(MLE) + sd$ 의 경우 각각 14, 13개 선거구로 비교적 더 잘 맞추었다. 신뢰구간 내 참값의 포함 여부를 기준으로 비교해도, 본 연구에서 제안하는 $MAR + MNAR(MLE) + rs$, $MAR + MNAR(MLE) + sd$ 이 각각 205(195+10), 198(190+8)개로, 조사기관 197개(=184+13)에 비해 정확성이 높다. 또한, 본 연구에서 제안한 MNAR 모형하의 MLE와 베이지안 결과를 비교하더라도 본 연구에서 제안한 방법이 더 우수함을 알 수 있다. 순위를 정확히 맞추는지와 신뢰구간 내에 포함하는지를 기준으로 볼 때, $MAR + MNAR(Bayesian)$ 은 모든 투표소를 MAR로 적용한 경우와 비슷한 성능을 보여 MNAR의 효과를 얻을 수 없었다.

다음으로 무응답 대체방법 간의 성능을 비교하기 위하여 대체방법별 예측 오류율을 평가하여 그 결과를 <표 3>에 제시하였다.

<표 3>의 두 번째 열은 선거구별 예측오차의 평균이며 세 번째 열은 예측오차를 실제 득표율 차이로 나누어 단위의 효과를 없앴다. 네 번째 열은 1, 2위 후보자 각각의 예측오차를 합하여 평균을 구했고, 마지막 열에서는 각 후보자의 예측오차를 참값으로 나누 후 합한 값의 평균을 구했다. 단순한 예측오차 혹은 예측오차의 합보다는 참값으로 나누어 오차의 비율을 비교하는 것이 더 정확한 비교라고 볼 수 있다.

<표 2>와 <표 3>의 결과는 Kim & Park (2014)에서의 모의실험 결과인, MNAR을 가정으로 하는 베이지안 추정결과가 MAR 가정하의 MLE보다 좋지 않을 수도 있다는 결과와 일맥상통한다. 그 결과, 본 연구에서 제안하는 연령 통합을 통한 변방값 문제 해결방안이 베이지안 추정방법보다 더 오류건수 및 오류율이 작은 것을 볼 수 있었다. 또한 본 연구방법이 조사기관의 성, 연령, 지역별 자료를 이용한 사후층화 보정방법보다 정확성이 뛰어나함을 확인할 수 있었다.

<표 3> 무응답 대체방법별 오차의 평균 비교

Imputation Method	$ \hat{p}_1 - \hat{p}_2 - (p_1 - p_2) $	$\frac{ \hat{p}_1 - \hat{p}_2 - (p_1 - p_2) }{(p_1 - p_2)}$	$ \hat{p}_1 - p_1 + \hat{p}_2 - p_2 $	$\frac{ \hat{p}_1 - p_1 }{p_1} + \frac{ \hat{p}_2 - p_2 }{p_2}$
KEP	0.03346	0.68980	0.03541	0.08645
MCAR	0.04314	0.60866	0.04507	0.10810
MCAR+rs	0.04219	0.61743	0.04393	0.10550
MCAR+sd	0.04305	0.59965	0.04490	0.10752
MAR	0.03480	0.45278	0.03677	0.08999
MAR+MNAR(MLE)	0.03220	0.32539	0.03422	0.08413
MAR+MNAR(MLE)+rs	0.03122	0.27188	0.03318	0.08183
MAR+MNAR(MLE)+sd	0.03220	0.24983	0.03400	0.08364
MAR+MNAR(Bayesian)	0.03347	0.42027	0.03542	0.09250

다음으로는 순위를 가장 잘 맞춘 MAR + MNAR(MLE) + sd 방법으로도 맞추지 못한 13개 선거구의 예측실패 원인을 분석하고 해결방안을 제시하고자 한다. 이들 13개 선거구는 모두 ‘오차 범위 내에서 경합’으로 판정된 선거구이다. 선거조사에서 경합이란 우열을 가릴 수 없게 엇치락뒤치락하며 겨루어 나가는 상황으로 통계학적으로 설명하면 1등 후보자의 신뢰구간과 2등 후보자의 신뢰구간이 서로 겹쳐 확실하게 어느 쪽이 더 우열하다고 판단을 내리기 힘든 경우이다. 예측실패의 원인이 무응답이 아닌, 투표소추출오차 ($|\tilde{p}_1 - p_1| + |\tilde{p}_2 - p_2|$) 등에 기인한 것은 아닌지 확인하기 위하여 각 선거구의 실제 득표율과 표본투표소의 개표결과를 비교하였다. <표 4>에서 p_1, p_2 는 각각 1, 2위 후보의 실제득표율, $\tilde{p}_1, \tilde{p}_2$ 는 각각 표본으로 추출된 투표소에서 1, 2위 후보의 득표율을 의미한다.

p_1, p_2 의 대소관계, 즉 순위와 $\tilde{p}_1, \tilde{p}_2$ 의 순위를 비교하면 서울특별시 서대문구을, 광주광역시 동구, 경기도 고양시 덕양구갑, 경기도 고양시 덕양구을, 경기도 광주시 다섯 선거구의 경우는 실제값의 순위와 표본투표소로 뽑힌 투표소들로 구한 실제값의 순위가 반대임을 알 수 있다. 이는 해당 선거구 내에서 표본투표소를 추출하는 과정에서 생기는 오차이므로 이 과정에서 이미 1, 2

<표 4> 예측실패한 선거구의 참값과 표본 투표소에서의 추정치

Electoral Districts		p_1	p_2	$\tilde{p}_1$	$\tilde{p}_2$
0105	서울특별시 성동구을	48.61%	50.00%	49.12%	49.58%
0124	서울특별시 서대문구을	49.20%	48.79%	48.64%	49.15%
0127	서울특별시 양천구갑	50.83%	49.17%	50.69%	49.31%
0145	서울특별시 송파구을	49.96%	46.12%	50.95%	45.24%
0501	광주광역시 동구	31.12%	31.78%	31.15%	29.84%
0601	대전광역시 동구	34.93%	33.55%	35.08%	34.23%
0827	경기도 고양시 덕양구갑	48.93%	49.79%	49.97%	48.52%
0828	경기도 고양시 덕양구을	48.35%	48.29%	47.97%	48.60%
0850	경기도 광주시	47.37%	45.96%	46.61%	47.20%
0901	강원도 춘천시	49.61%	44.99%	49.61%	45.16%
0903	강원도 원주시을	49.24%	46.02%	47.99%	47.15%
1107	충청남도 논산계룡금산	40.11%	42.28%	39.27%	40.29%
1513	경상남도 거제시	33.24%	35.59%	31.29%	34.13%

위 후보자가 바뀐 경우에는 아무리 MAR, MNAR 기법을 이용해도 방향을 뒤집기 힘들다. 따라서 이 다섯 선거구는 무응답 대체방법의 문제점보다는 투표소 추출과정의 오류라고 할 수 있다.

13개의 선거구 중 표본투표소 추출과정에서의 오류로 인해 순위를 맞추지 못한 다섯 선거구를 제외하고 나머지 8군데 중 ‘성동구을’과 ‘송파구을’은 전체 투표소에 성별, 연령별 MAR 가정하에서 1, 2위를 정확히 맞춘 곳이다. 하지만 연령별 무응답률을 고려하는 과정에서 MNAR로 잘못 적용되어 들어간 몇몇 투표소로 인해 뒤집혀버린 것임을 확인했다. 이러한 문제는 연령별 무응답률을 고려한 MNAR 판정과정에서의 한계점으로 보인다. 본 연구에서는 최대, 최소 무응답률의 차이가 지역별 평균보다 1σ 작은 경우에 연령별 차이가 없다고 판단하고 MNAR로 적용하였다. 하지만 본 연구에서 사용한 기준 이외에 연령별로 무응답률에 차이가 없음을 판단해 주는 더 나은 기준을 찾는다면 이 문제를 보완할 수 있을 것이다. 또 다른 해결방법으로는 사전정보

의 이용에 있다. ‘송파구을’의 경우 방송 3사가 TNS 등 여론조사기관을 통해 총선 9일 전 지역별 유권자 500명씩 임의전화걸기 방식으로 여론조사를 한 결과 1위 후보자가 40.4%의 지지율로 2위 후보자(29.7%)를 크게 앞섰다. 만약 이러한 여론조사의 결과를 반영한다면 무응답 대체과정에서 순위가 바뀌었을 때 이를 올바르게 조정하여 정확하게 예측할 수 있을 것이다. 따라서 선거 전에 이루어지는 여론조사나 지역 전문가 등을 통한 사전정보를 충분히 이용한다면 MAR과 MNAR을 고려하는 과정에서 발생하는 위의 두 곳과 같은 선거구의 예측실패를 예방할 수 있을 것이다.

IV. 결론 및 제언

본 연구에서는 출구조사 무응답 처리의 정확성을 높이하고자 2012년 제19대 국회의원 선거 자료를 이용하여 무응답을 대체(imputation)하는 새로운 방안을 제안하였다. 먼저 변방값의 발생특성과 무응답 매커니즘의 정의를 이용하여 무응답 매커니즘을 식별 및 판정하였다. MNAR 가정 시 발생하는 변방값 문제를 해결하기 위한 대안으로 기존의 베이지안 방법이 아닌 연령범주를 활용한 새로운 MLE 방법을 제안하였다. 그 결과 조사기관의 사후증화법을 이용한 무응답 보정이나 베이지안 방법보다 예측의 정확성이 더욱 향상됨을 볼 수 있었다. 향후 출구조사 관련 연구에 본 연구결과가 하나의 대안 혹은 도움이 되기를 기대한다.

참고문헌

- 김영원 · 곽은선. 2010. “2008 총선 출구조사의 총조사오차 분석.” 《조사연구》 11(3): 33-55.
- 현경보. 2005. “선거 예측조사의 정확도와 그 영향요인 연구: 17대 총선 예측조사 결과의 사후 분석.” 《언론과학연구》 5(1): 301-336.

- Baker, S.G., W.F. Rosenberg, and R. Dersimonian. 1992. "Closed-Form Estimates for Missing Counts in Two-way Contingency Tables." *Statistics in Medicine* 11: 643-657.
- Bautista, R., M. Callegaro, J.A. Vera, and F. Abubdis. 2007. "Studing Nonresponses in Mexican Exit Polls." *International Journal of Opinion Research* 19: 492-503.
- Choi, B.S., J.W. Choi, and Y. Park. 2009. "Bayesian Methods for an Incomplete Two-way Contingency Table with Application to the Ohio (Buckeye State) Polls." *Survey Methodology* 35: 37-51.
- Foster, J.J. and P.W.F. Smith. 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.
- Ibrahim, J.G., H. Zhu., and N. Tang. 2008. "Model Selection Criteria for Missing-data Problem Using the EM Algorithm." *Journal of the American Statistical Association* 103: 1648-1658.
- Kim, S. and Y. Park. 2014. "Power-linear Models for Incomplete Contingency Tables with Nonignorable Non-responses." *Statistics*(ahead-of-print), 1-15.
- Little, J.A. and D.B. Rubin. 2002. *Statistical Analysis with Missing Data*. (2nd ed.), New York: Wiley.
- Martin, E.A., M.W. Traugott, and C. Kenedy. 2005. "A Review and Proposal for a New Measure of Poll Accuracy." *Public Opinion Quarterly* 69(3): 342-369.
- Molenberghs, G., C. Beunckens, C. Sotito, and M.G. Kenward. 2008. "Every Missingness not at Random Model Has a Missingness at Random Counterpart with Equal Fit." *Journal of the Royal Statistical Society: Series B* 70: 371-388.
- Park, T. and M.B. Brown. 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, Y. and B.S. Choi. 2010. “Bayesian Analysis for Incomplete Multi-way Contingency Tables with Nonignorable Nonresponse.” *Journal of Applied Statistics* 37: 1493-1453.
- Pavia, J.M. 2010. “Improving Predictive Accuracy of Exit Polls.” *International Journal of forecasting* 26: 68-81.

<접수 2014/6/18, 수정 2014/8/21, 게재확정 2014/8/25>

