

연구논문

## 선거예측 개선을 위한 새로운 대체방법: 제19대 총선 출구조사를 중심으로

A New Imputation Method to Improve Election Prediction:  
From Exit Polls of the 19th General Election

전새봄<sup>a)</sup> · 박유성<sup>b)</sup>

Saebom Jeon · Yousung Park

출구조사 예측실패의 큰 원인 중 하나는 출구조사에서 발생하는 무응답이다. 이에 출구조사의 예측력을 개선하기 위하여, 무응답을 줄이기 위한 노력과 동시에 무응답을 사후층화(post-stratification)하거나 대체(imputation)하는 등 다양한 방법이 연구 및 적용되어 왔다. 그러나 무응답을 추정 및 대체하는 방법은 무응답 메커니즘에 따라 다름에도 불구하고, 추정에 앞서 그 메커니즘을 식별하는 방법은 제한적으로 연구되어 왔다. 본 연구는 메커니즘에 대한 식별이 불가하다고 알려진 우도기반 추정방법의 한계를 극복하고자, 새로운 추정법인 적률추정법을 제안하고 있다. 또한 각 응답 메커니즘 하에서의 변방값 발생 특성 및 무응답 메커니즘의 정의를 이용하여 메커니즘을 식별하는 방법을 제안하고 있다. 2012년 제19대 총선 출구조사 자료를 이용하여 실증 분석한 결과, 무응답 메커니즘을 식별 및 대체한 예측결과가 기존의 조사기관 예측결과 및 ML, 그리고 베이지안 방법보다 더 우수한 예측력을 보여, 본 연구 방법에서 제안한 방법이 출구조사 선거예측 정확도를 개선시키고 있음을 확인할 수 있었다.

**주제어:** 무응답 메커니즘, MAR, MNAR, 변방값 문제, 적률추정법

a) 고려대학교 세종캠퍼스 응용통계학과 초빙교수.

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

E-mail: yspark@korea.ac.kr

One of the biggest causes of failure in exit poll prediction is non-responses during exit polling. In this regard, there have been many efforts to reduce non-responses, and various researches and applications on imputation method including post-stratification. However, although these imputation methods depend on the missing mechanism, studies on the identification of mechanism prior to estimation are very limited. This research proposes a method of moment as a new estimation method to overcome the nonidentifiability of mechanism due to the same likelihood. Also, we suggest a way to identify the missing mechanism, using the definition of missing mechanism as well as the occurrence of boundary solution problem of moment estimates. From the empirical analysis of the exit polls of 19th general election in 2012, our method of moment imputation according to the identified missing mechanism shows better prediction than other methods including ML and Bayesian imputation. Therefore, our proposed imputation method can be a new alternative to improve the prediction accuracy of exit poll.

**Key words:** nonresponse mechanism, MAR, MNAR, boundary solution, method of moment

## I. 서론

출구조사(exit poll)란 선거 당일 투표소에서 투표를 막 마치고 나오는 유권자를 상대로 어느 후보를 선택하였는지를 조사하는 방법으로, 선거결과를 가장 빨리 예측할 수 있는 조사이기 때문에 여러 나라에서 시행하고 이를 법적으로 보장하고 있다. 출구조사가 투표 직후 유권자를 대상으로 즉시 시행되므로 선거 여론

조사보다 높은 정확도가 기대되는데 비해, 선거예측 오류에 대한 논란은 끊임없이 있어 왔다. 특히 총선 출구조사는 조사 표본이 상대적으로 적은 상태에서 지역구별 당선자를 예측해야 하는 한계 때문에 정확도가 대선 또는 광역 지자체 선거에 비해 상대적으로 취약한 면이 있다.

과거 15대 총선의 경우 여야 포함 총 39개 지역구에서 예측 오류가 발생하였고, 16대 총선의 경우 각 방송사마다 예상한 여당의 의석수는 다소 차이가 있으나 여야 포함 총 21~23개 지역구에서 예측을 실패하였다. 17대, 18대의 출구조사에서는 예상 여당 의석수 예측의 오차범위를 너무 넓게 설정하여 무의미한 예측결과를 발표하거나, 예측범위를 벗어나는 예상의석수 제시 등으로 인해 총선 출구조사에 대한 무용론이 제기되기도 하였다. 이후 19대 총선의 출구조사 예측력을 높이기 위해, 한국방송협회는 2010년 방송사공동예측조사위원회(KEP)를 설립하여 공동으로 전지역에 대해 출구조사를 시행하고, 출구조사의 정확도를 높이기 위하여 선거법 개정을 통해 출구조사를 투표소로부터 근거리에서 실시할 수 있도록 제한거리를 100m로 완화하는 등의 노력을 기울였다. 이에 2012년 19대 총선 출구조사 예측오류는 전체 246개 지역구 중 17개 지역으로 감소하였으며, 이 중 15개 지역은 오차범위 내였으므로 예측력 개선을 위한 여러 제도적 장치가 어느 정도 효과를 보였다고 생각할 수 있다.

출구조사 무용론을 주장하는 사람들은, 총선이 대선에 비해 지역구 간 특성이 다르고 지역구 내에서도 투표소별 모집단 분포가 동질하지 않을 수 있다는 점, 그리고 고령층 혹은 보수층이 정치적 의사표출 부담으로 출구조사에 덜 응하거나 솔직하게 응답하지 않았을 가능성을 지적한다. 이는 통계적 관점으로 보면, 표본에 대한 대표성 문제, 그리고 무응답에 대한 문제로 볼 수 있다. 즉 각 지역구에서 추출된 출구조사 투표소가 지역구의 의견을 대표하지 못함에서 발생하는 표본오차, 그리고 조사시 발생하는 무응답 등에 의한 실사오차로 인해 총선 출구조사 결과에 많은 오차가 발생할 수 있다는 점이다. 특히 무응답에 의한 오차는 선거예측 실패의 대표적 원인으로 꼽힌다.

그러나 Baustista et al.(2007)는 무응답률을 줄이더라도 예측 개선으로 이어지는 것은 아니라고 밝혔으며, 전세봄 외(2014) 역시 제19대 국회의원 선거 출구조사의 무응답률과 예측오차(Modified Within Prediction Error, 이하 MWPE)

와의 관계를 살핀 결과, 무응답이 약 17%를 차지하였지만 무응답률이 예측오차와 큰 관련이 없음을 파악하였다. 이러한 결과는 무응답을 줄이기 위한 여러 시간적, 비용적 노력보다는, 이미 발생한 무응답을 처리하는 사후 노력을 통해 예측력을 개선할 수 있음을 시사한다.

무응답 처리를 위한 사후 노력으로 여론조사 기관에서도 출구조사에서 조사되는 성별 및 연령 정보를 이용해 사후층화(poststratification)나 대체(imputation) 등을 통해 정확한 예측을 위해 노력하고 있다. 이때 무응답의 처리나 대체방법은 무응답의 발생체계가 완전임의결측(Missing Completely At Random, 이하 MCAR), 임의결측(Missing At Random, 이하 MAR), 그리고 비임의결측(Missing Not At Random, 이하 MNAR)인지에 따라 다르다(Little & Rubin 2002). 만약 지지후보에 대한 무응답이 *MCAR*을 따른다면, 무응답이 완전히 랜덤하게 발생하므로 지지후보자에 대한 응답이 없는 무응답 관측치는 삭제하고 응답자 자료만을 이용하여 지지율을 예측할 수 있으나, *MAR*이나 *MNAR*과 같이 무응답이 관측되거나 관측되지 않은 다른 변수에 의존한다면 이를 반영하여 무응답을 대체하는 것이 적절하다. *MAR*은 지지후보자에 대한 무응답이 관측된 변수에 의존하여 발생하는 경우이므로, 지지성향에 큰 영향을 미치는 성별 및 연령 정보를 이용하여 무응답을 대체할 수 있다. *MNAR*은 지지후보자가 누군지에 따라 무응답이 달라지는 침묵의 나선효과(spiral of silence effect)와 관련된 경우로, 지지하는 후보자가 열세를 보이거나 보수 성향인 경우 응답을 회피하는 경향이 있으므로 추격후보에 무응답을 더 부여하는 등의 대체방법을 사용한다.

<표 1> 19대 총선 출구조사 지지후보 응답자 수(서울 평창동 제2투표소) (단위: 명)

| Age   | 1위 후보 지지 | 2위 후보 지지 | 기타 후보 지지 | 무응답 |
|-------|----------|----------|----------|-----|
| 18-29 | 21       | 30       | 1        | 10  |
| 30-39 | 38       | 36       | 0        | 25  |
| 40-49 | 60       | 48       | 1        | 17  |
| 50-59 | 72       | 47       | 0        | 24  |
| 60+   | 88       | 10       | 1        | 43  |

이와 같이 무응답 메커니즘 가정에 따라 무응답 처리방법이 다른데, 이러한 메커니즘 가정이 선거예측 결과에 영향을 끼치는지 알아보기 위해 무응답 처리 예를 살펴보면 다음과 같다. <표 1>은 19대 총선 당시 출구조사 대상 투표소로 선정된 서울 평창동 제2투표소의 출구조사 응답결과이다. 19대 총선은 대체로 지역이나 이념 대결보다는 세대대결의 경향을 보인다고 알려졌으므로 지지후보에 대한 응답을 연령별로 파악하면 다음과 같다.

투표결과가 세대대결의 경향을 보인다는 것은 다시 말하면 지지후보에 대한 응답이 연령에 의존하는 성격을 지닌 *MAR*을 의미한다고 볼 수 있을 것이다. 그런데 박빙지역에 대한 사후분석 결과, 선거의 ‘숨은 표’가 여당 지지자들로 밝혀졌는데, 즉 이는 지지후보자에 대한 무응답이 *MNAR*이었음을 의미한다고 볼 수 있다. 이와 같이 출구조사에서의 무응답 메커니즘은 *MAR*일 수도 있고, 혹은 *MNAR*일 수도 있는데, 기존 연구에서는 *MAR*과 *MNAR*의 우도가 같아 이를 테스트할 수 없다고 밝혀 왔다(Molenberghs et al. 2008; Ibrahim et al. 2008). 그러나 득표율 예측결과는 <표 2>에서 볼 수 있듯이, 동일한 투표소에 대한 무응답 자료라 하더라도 무응답 발생이 *MCAR*인지, *MAR*인지, 혹은 *MNAR*인지의 가정에 따라 예측결과가 달라지는 것을 볼 수 있다.

<표 2> 무응답 메커니즘별 출구조사 예측치(서울 평창동 제2투표소) (단위: 명)

| Mechanism | <i>MCAR</i> |       |      | <i>MAR</i> |       |      | <i>MNAR</i> |       |      |
|-----------|-------------|-------|------|------------|-------|------|-------------|-------|------|
| Age       | 1위          | 2위    | 기타   | 1위         | 2위    | 기타   | 1위          | 2위    | 기타   |
| 18-29     | 26.52       | 37.88 | 1.26 | 23.41      | 33.44 | 1.11 | 31.00       | 30.00 | 1.00 |
| 30-39     | 47.98       | 45.46 | 0.00 | 44.21      | 41.88 | 0.00 | 63.00       | 36.00 | 0.00 |
| 40-49     | 75.76       | 60.61 | 1.26 | 74.44      | 59.55 | 1.24 | 77.00       | 48.00 | 1.00 |
| 50-59     | 90.91       | 59.35 | 0.00 | 90.91      | 59.35 | 0.00 | 96.00       | 47.00 | 0.00 |
| 60+       | 111.12      | 12.63 | 1.26 | 107.23     | 12.19 | 1.22 | 131.00      | 10.00 | 1.00 |

이때 <표 2>에 제시된 *MNAR*은 *ML(EM)* 추정결과이다. 이와 같이 무응답

메커니즘 가정에 따라 무응답 처리방법이 다르므로 득표율 예측결과는 달라질 수 있으며, 보다 정확한 선거예측을 위해서는 어떠한 무응답 메커니즘을 가정할지에 대한 적절한 식별이 필요하다. 이러한 관점에서 본 논문의 첫 번째 관심사는 출구조사에서 무응답 메커니즘을 식별할 수 있는가, 또한 식별된 메커니즘 가정이 적절한가이다. 이를 위해 본 논문은 Kim & Park(2015)과 Jeon et al.(2016)이 제안한 무응답 메커니즘별 변방값의 성격을 이용하여 무응답 메커니즘에 대한 식별 기준을 제시하고자 한다.

한편, 무응답 메커니즘을 적절히 식별한다 하더라도, 식별된 메커니즘 가정하에서 무응답을 대체하는 추정방법에 따라 예측결과는 달라질 수 있다. 다음의 <표 3>은 무응답 패턴을 *MNAR*로 가정하고 추정방법만 다를 때, 즉 *ML* 방법으로 추정한 결과(*MNAR\_EM*)와 베이지안 방법으로 추정한 결과(*MNAR\_BAYES*)를 나타낸다.

<표 3> MNAR 가정하에 추정방법별 출구조사 예측치(서울 평창동 제2투표소) (단위: 명)

| Estimation<br>Response | <i>MNAR_EM</i> |       |      |       |      |      | <i>MNAR_BAYES</i> |       |       |       |      |      |
|------------------------|----------------|-------|------|-------|------|------|-------------------|-------|-------|-------|------|------|
|                        | 응답             |       |      | 무응답   |      |      | 응답                |       |       | 무응답   |      |      |
| Age                    | 1위             | 2위    | 기타   | 1위    | 2위   | 기타   | 1위                | 2위    | 기타    | 1위    | 2위   | 기타   |
| 18-29                  | 21.73          | 30.00 | 1.00 | 9.27  | 0.00 | 0.00 | 20.83             | 29.17 | 2.81  | 4.38  | 4.56 | 0.24 |
| 30-39                  | 44.16          | 36.00 | 0.00 | 18.84 | 0.00 | 0.00 | 38.86             | 36.85 | 5.24  | 11.65 | 5.76 | 0.64 |
| 40-49                  | 53.98          | 48.00 | 1.00 | 23.02 | 0.00 | 0.00 | 58.49             | 46.68 | 7.88  | 5.35  | 7.30 | 0.29 |
| 50-59                  | 67.30          | 47.00 | 0.00 | 28.70 | 0.00 | 0.00 | 71.98             | 47.23 | 9.70  | 6.35  | 7.39 | 0.35 |
| 60+                    | 91.83          | 10.00 | 1.00 | 39.17 | 0.00 | 0.00 | 88.83             | 11.07 | 11.97 | 26.93 | 1.73 | 1.48 |

*MNAR* 가정하에 *ML* 즉 *EM*으로 추정한 결과, 총 119명의 무응답이 한 후보자로 쏠림 현상을 보이고 나머지 후보자에는 0의 확률이 부여되는 변방값 문제가 발생한 것을 볼 수 있다. 이처럼 *ML* 추정치는 변방값 문제에 빠지기 쉬우며, 변방값 문제가 발생하는 경우 추정결과가 불안정하다는 문제가 있다. 이를 해결

하기 위해 베이지안 추정치가 활발히 연구되었으며 대안으로 선호되었다(Foster & Smith 1998; Green & Park 2003; Choi et al. 2009; Park & Choi 2010; Clogg et al. 1991; Park & Brown 1994; Park 1998; Choi et al. 2009; Park & Choi 2010). <표 3>의 결과와 같이 베이즈 방법으로 변방값 문제를 어느 정도 해결할 수 있으나, 사전분포 선택에 따라 이를 완전히 해결하지 못하는 경우도 존재하며 사전분포에 대한 주관성 논쟁에서 자유롭지 못하다는 한계가 있다(Kim & Park 2015; 전세봄 외 2014; Jeon et al. 2016) 따라서 변방값 문제에서 자유로우면서 주관성이 배제된 새로운 추정방법이 요구된다. 이러한 관점에서 본 논문의 두 번째 관심사는 식별된 메커니즘 가정하에서 출구조사 예측력이 가장 우수한 추정법은 무엇인가, 그리고 그러한 추정법이 변방값 문제에서 자유로운가이다.

이 논문은 총 4개의 절로 구성되어 있다. 제2절에서는 무응답 메커니즘에 따른 변방값 성격을 이용하여 무응답 메커니즘을 식별하는 방안을 설명하고, 적률추정법을 이용한 무응답 대체방법을 제안한다. 제3절에서는 2012년 실시된 국회의원 출구조사 자료를 이용하여 각 지역구 및 투표소별 무응답 메커니즘을 식별하고 대체방법별 예측결과를 비교함으로써, 위의 두 관심사에 대한 실증분석 결과를 제시하고 논의할 것이다.

## II. 변방값 발생 특성을 이용한 무응답 메커니즘 식별

선거예측의 관심변수는 지지후보이므로 연령을 밝히지 않은 응답자는 제외하고, 연령이 파악된 응답자 중 지지후보를 밝히지 않은 응답자만 출구조사 예측에 활용한다고 하면, 출구조사 자료는 다음과 같은 삼차원 분할표로 요약가능하다.  $Y$ 를 응답자가 지지하는 후보에 대한 범주형 변수,  $X$ 를 응답자의 연령범주이며,  $R$ 을  $Y$ 에 대한 무응답 여부를 나타내는 변수라 할 때, 분할표는 연령범주의 수 ( $i = 1, \dots, I$ ), 지지후보 수 ( $j = 1, \dots, J$ ), 무응답 여부 ( $k = 1, 2$ )를 나타내는  $I \times J \times 2$  분할표로 나타낼 수 있다.

| $R$              |     | $R=1$     |           | $R=2$     |           |
|------------------|-----|-----------|-----------|-----------|-----------|
| $X \backslash Y$ | $Y$ | $Y=1$     | $Y=j$     | $Y=J$     |           |
| $X=1$            |     | $y_{111}$ | $\cdots$  | $y_{1J1}$ | $y_{1+2}$ |
| $X=i$            |     | $\cdots$  | $y_{ij1}$ | $\cdots$  | $y_{i+2}$ |
| $X=I$            |     | $y_{I11}$ | $\cdots$  | $y_{IJ1}$ | $y_{I+2}$ |

본 연구의 관심은 지지후보를 밝히지 않은 무응답 유권자의 표가 어느 후보를 지지하는지, 즉 숨은 표에 대한 지지후보를 적절히 추정하여 이를 대체함으로써 출구조사의 예측력을 보다 개선하고자 함이다. 선거에서 당선여부를 결정하는 것은 각 선거구 내의 득표율( $\pi_{ij+}$ )인데, 결합확률  $\pi_{ij+}$ 에 대한 추정은 무응답 유권자의 연령별 지지율  $\pi_{ij2}$ , 즉 지지후보를 밝히지 않은 숨은표에서 해당연령의 후보자별 득표 수  $y_{ij2}$ 에 대한 추정 및 대체 문제로 간주할 수 있다.

## 1. 무응답 메커니즘에 따른 무응답 발생확률

결합확률 추정을 보다 간단히 하기 위해, 확률을 요소별로 분해(factorization)하는 선택모형(Selection Model: SM)과 패턴혼합모형(pattern Mixture Model: PMM)이 널리 사용된다(Little & Rubin 2002; Little 1994). SM을 고려하는 경우,  $X, Y, R$ 에 대한 결합확률  $P(X=i, Y=j, R=k)$ 는 다음과 같이 분해될 수 있다(Glynn et al. 1986; Little 1994).

$$P(X=i, Y=j, R=k) = P(X=i, Y=j) \times P(R=k|X=i, Y=j) \quad (1)$$

이때,  $P(X=i, Y=j, R=k)$ 를  $\pi_{ijk}$ 로,  $P(X=i, Y=j)$ 를  $\pi_{ij}$ 로,  $P(R=k|X=i, Y=j)$ 를  $\pi_{k|ij}$ 로 각각 정의하면, 식 (1)은  $\pi_{ijk} = \pi_{ij} \times \pi_{k|ij}$ 로 간단히 표현할 수 있다. Baker & Laird(1988)은 이  $\pi_{ij}$ 와  $\pi_{k|ij}$ 를 각각 ‘margin’ 모형과 ‘non-response’ 모형 두 개의 부분모형으로 정의하고, ‘margin’ 모형이 포화(saturated)모형이고 무응답(non-response) 확률이  $X$ , 혹은  $Y$ 에만 의존하는 경우  $\sum_j \lambda_{jk} = \sum_k \lambda_{jk} = 0$

하에서  $MAR$  및  $MNAR$  log-linear 모델을 다음과 같이 정의하였다.

$$MAR \text{ model: } \log(\pi_{ijk}) = \lambda_i + \lambda_j + \lambda_k + \lambda_{ij} + \lambda_{ik}$$

$$MNAR \text{ model: } \log(\pi_{ijk}) = \lambda_i + \lambda_j + \lambda_k + \lambda_{ij} + \lambda_{jk}$$

즉  $\pi_{ij}$ 이 포화모형이며  $Y$ 의 무응답  $R$ 이  $X$ , 혹은  $Y$ 에만 의존한다는 가정에 따라, 결합확률  $\pi_{ijk}$ 에 대한 추정문제는 조건부 확률  $\pi_{k|ij}$ 에 대한 추정, 즉 무응답 메커니즘에 따라  $\pi_{k|i}$  혹은  $\pi_{k|j}$ 에 대한 추정으로 축약될 수 있다(Little and Rubin 2002; Baker & Laird 1988; Clarke & Smith 2005). 만약  $Y$ 의 무응답 메커니즘이  $MCAR$ 을 따른다면  $Y$ 의 무응답 즉  $R$ 은  $X$ ,  $Y$ 에 무관하게 완전 임의적으로 발생할 것이므로  $\pi_{k|ij}$ 는  $X$ 와  $Y$ 에 의존하지 않는  $\pi_k$ 로 축약가능하다. 따라서 동일한 가정하에 식(1)은 무응답 메커니즘에 따라 다음과 같이 정리될 수 있다.

$$\pi_{ijk} = \begin{cases} \pi_{ij} \times \pi_{k|i}, & \text{for } MAR \quad Y, \\ \pi_{ij} \times \pi_{k|j}, & \text{for } MNAR \quad Y. \end{cases}$$

한편,  $PMM$ 의 경우,  $X$ ,  $Y$ ,  $R$ 에 대한 결합확률  $P(X=i, Y=j, R=k)$ 는 다음과 같이 분해될 수 있다(Smith et al. 1999).

$$P(X=i, Y=j, R=k) = P(R=k) \times P(Y=j|R=k) \times P(X=i|Y=j, R=k) \quad (2)$$

이때 결합확률  $P(X=i, Y=j, R=k)$ 를  $\pi_{ijk}$ 로,  $P(R=k)$ 는  $\pi_k$ 로,  $P(Y=j|R=k)$ 는  $\pi_{j|k}$ 로,  $P(X=i|Y=j, R=k)$ 는  $\pi_{i|jk}$ 로 각각 정의하면, 식(2)는  $\pi_{ijk} = \pi_k \times \pi_{j|k} \times \pi_{i|jk}$ 로 간단하게 표현할 수 있다. 이때  $\pi_k$ ,  $\pi_{j|k}$ ,  $\pi_{i|jk}$ 는  $SM$ 하에서의 조건부 확률  $\pi_{k|ij}$ 를 이용하여 추정이 가능하다. 즉,  $\pi_k = \sum_{i,j} \pi_{ijk} = \sum_{i,j} \pi_{ij} \times \pi_{k|ij}$ ,  $\pi_{j|k} = \frac{\sum_i \pi_{ijk}}{\sum_{i,j} \pi_{ijk}} = \frac{\sum_i \pi_{ij} \times \pi_{k|ij}}{\sum_{i,j} \pi_{ij} \times \pi_{k|ij}}$ ,  $\pi_{i|jk} = \frac{\pi_{ijk}}{\sum_i \pi_{ijk}} = \frac{\pi_{ij} \times \pi_{k|ij}}{\sum_i \pi_{ij} \times \pi_{k|ij}}$ 와 같이 표현되므로,  $SM$ 하에서  $\pi_{k|ij}$ 가 무응답 메커니즘에 따라 각각  $\pi_{k|i}$ ,  $\pi_{k|j}$ ,  $\pi_k$ 로 축약가능하였

던 것과 마찬가지로, *PM*하에서도 역시 무응답 발생확률은 무응답 메커니즘에 따라 달라진다. 이와 같이 *SM*과 *PMM*의 확률분해 과정을 통해, *Y*의 무응답이 *MAR*인 경우는 *X*범주에만 의존하고, *MNAR*인 경우에는 *Y*에만 의존하며, *MCAR*인 경우는 의존하지 않음을 확인할 수 있었다. 그런데 이렇듯, 무응답 메커니즘에 따라 무응답 발생확률이 다르게 정의됨에도 불구하고, 관측우도 관점에서 *MAR*과 *MNAR*의 무응답 메커니즘은 식별불가능하다고 알려져 있다 (Molenberghs et al. 2008). 이는 무응답 메커니즘을 식별하기 위해서는 우도기반 추정이 아닌 다른 추정법이 요구됨을 의미한다.

## 2. 적률추정법을 이용한 무응답 식별

일반적으로 분할표 내의 무응답에 대한 대표적인 모형은 무응답 건수를 공변량 및 무응답여부 등으로 설명하는 일반화선형모형(Generalized Linear Model: GLM)이나, 이러한 모형을 이용한 추정은 우도에 기반하고 있다. 따라서 본 논문은 특정 모형을 가정하지 않고 보다 일반적인 확률함수로 무응답을 대체하는 출구조사 예측방법을 제안하고자 한다.

Jeon et al.(2016)은 특정모형에 의존하지 않으며 관측정보를 이용하는 새로운 추정법인 적률추정법을 제안하였다. 즉 완전히 관측된  $\pi_{ij1}$ 와 부분적으로 관측된  $\pi_{i+2}$ 에 대한 적률추정치로 각각  $\hat{\pi}_{ij1} = y_{ij1} / \sum_{i,j,k} y_{ijk}$ ,  $\hat{\pi}_{i+2} = y_{i+2} / \sum_{i,j,k} y_{ijk}$ 을 제안하였으며, 또한  $I \times J \times 2$  분할표에서  $i$ 번째  $X$ ,  $j$ 번째  $Y$ 의 무응답 발생 비율에 대한 함수  $h(i,j) = \pi_{ij2} / \pi_{ij1}$ 를 정의하였다. 이때  $h(i,j)$  함수는 결합확률의 비(ratio)이므로 *SM* 및 *PMM*하에서 무응답 메커니즘에 따라 추약가능하다. 즉 Jeon et al.(2016)이 제안한 변수의존 무응답 메커니즘(variable-based missing mechanism) 정의에 따라, *Y*의 무응답 발생이 *MAR*를 따른다면,  $h(i,j)$ 는  $i$ 에만 의존하는  $h(i, \cdot) (= \pi_{i+2} / \pi_{i+1})$ 로 추약가능하고, 반대로 *Y*의 무응답 발생이 *MNAR*를 따른다면,  $h(i,j)$ 는  $j$ 에만 의존하는  $h(\cdot, j)$ 로 추약가능하다. 만약 *Y*가 *MCAR*을 따른다면  $h(i,j)$ 는  $i,j$ 에 의존하지 않는  $h(\cdot, \cdot) (= \pi_{++2} / \pi_{++1})$ 으로 추약가능하다. 이러한  $h(i,j)$ 를 통해  $\hat{\pi}_{ij2}$ 는 무응답 메커니즘에 따라 다음과

같이 추정가능하다.

$$\hat{\pi}_{ij2} = \begin{cases} \hat{h}(i, \cdot) \times \hat{\pi}_{ij1}, & \text{for } MAR \text{ } Y, \\ \hat{h}(\cdot, j) \times \hat{\pi}_{ij1}, & \text{for } MNAR \text{ } Y. \end{cases}$$

본 논문에서 고려하는 출구조사 자료의 경우  $X$ 는 항상 관측되고 무응답은  $Y$ 에서만 발생하므로,  $Y$ 의 무응답 발생이  $MCAR$  혹은  $MAR$ 인 경우  $\hat{\pi}_{ij1}$ ,  $\hat{\pi}_{++1}$ ,  $\hat{\pi}_{++2}$ ,  $\hat{\pi}_{i+1}$ ,  $\hat{\pi}_{i+2}$  모두 적률추정법을 이용하여 관측치로부터 바로 추정가능하다. 따라서  $Y$ 의 무응답 발생이  $MCAR$  혹은  $MAR$ 인 경우  $\hat{\pi}_{ij2}$ 는 항상 추정가능하다. 그러나  $Y$ 의 무응답 발생이  $MNAR$ 인 경우,  $\hat{\pi}_{+j2}$ 는 관측되지 않으므로  $\sum_j \hat{h}(\cdot, j) \times \hat{\pi}_{ij1}$  방정식 관계를 통해 풀어야 한다. 이때 만약  $X$ 범주의 수가  $Y$ 범주의 수보다 크다면, 즉  $I > J$ 라면, 추정해야 하는 모수의 수보다 방정식의 수가 많아 유일(unique) 해를 보장 못하므로, 이런 경우  $X$ 의 범주들을 합치거나  $X$ 의  $I$ 개 범주 중  $J$ 개 범주를 선택하여 여러 셋의  $J \times J$  형태로 바꾸는 등의 방법을 통해 해를 구할 수 있다.

이러한 무응답 비  $h(i, j)$ 는 무응답 메커니즘을 식별하는 데에도 사용될 수 있다. 분할표에서 각 셀빈도  $y_{ij1}$ ,  $y_{i+2}$ 가 0이 아니라면  $\hat{\pi}_{i+2}$ 은 0이 아니어야 하지만,  $MNAR$ 하에서는 위와 같은 방정식의 해를 구하는 과정에서  $\hat{\pi}_{i+2}$ , 즉  $\hat{h}(\cdot, j)$ 가 0 혹은 음(-)수로 추정될 수 있는 변방값 문제(boundary solution problem)가 발생할 수 있다(Park et al. 2014). Park et al.(2014)와 Kim & Park(2015)는  $\pi_{ij1}$ 과  $\pi_{+j2}$ 의 비율인 응답 오즈(response odds) 및 무응답 오즈(nonresponse odds)를 이용하여 변방값 발생 특성을 정의하여, 실제  $MNAR$ 을 따르는  $Y$ 에  $MNAR$ 을 가정하여 추정시, 셀빈도가 0이 아닌 한 변방값 문제는 발생하지 않음을 증명하였다. 그러나 실제  $MAR$ 을 따르는  $Y$ 에 대해  $MNAR$ 을 가정하여 추정한다면, 무응답 오즈가 응답오즈의 최소값과 최대값을 벗어나는 변방값 문제가 발생함을 보였다. 즉 무응답 메커니즘에 따라 추정치의 변방값 발생 특성은 다르게 나타나므로, 이를 역으로 생각하여 변방값 발생특성 및 추정치의 변방값 발생여부를 이용하면 무응답 메커니즘은 어느 정도 식별가능하다.

정리 1. 무응답 메커니즘이 알려지지 않은  $Y$ 에  $MNAR$ 을 가정한  $\hat{h}(\cdot, j)$ 를 추정할 때,

- (a) 변방값 문제가 발생한다면,  $Y$ 의 무응답 메커니즘은  $MNAR$ 보다는  $MAR$ 일 가능성이 높다. 즉  $Y$ 의 무응답 메커니즘을  $MNAR$ 로 가정한  $\hat{h}(\cdot, j)$  추정결과 변방값 문제가 발생한다면,  $Y$ 의 무응답 메커니즘이 실제로는  $MAR$ 을 따르는데  $MNAR$ 로 잘못 적용하였을 가능성이 높다는 것을 의미하므로,  $\hat{h}(\cdot, j)$  추정시 변방값 문제가 발생하는 경우  $Y$ 의 무응답 메커니즘은  $MAR$ 로 식별할 수 있다.
- (b) 변방값 문제가 발생하지 않는다면,  $Y$ 의 무응답 메커니즘은  $MAR$ 보다는 다른 메커니즘 즉  $MNAR$ 이나  $MCAR$ 일 가능성이 높다. 그런데 본 논문에서  $MNAR$ 은  $Y$ 의 무응답  $R$ 이 연령이나 성별과 같은  $X$ 에는 의존하지 않고  $Y$ 에만 의존한다고 가정하였으므로,  $X, Y$ 와  $R$ 의 관계여부에 따라  $MNAR$ 과  $MCAR$ 을 식별할 수 있을 것이다.
  - (b-1) 만약  $X$ 와  $Y$ 가 통계적으로 유의한 연관성이 없다면,  $Y$ 의 무응답  $R$ 과  $Y$ 와의 관계는  $X$ 와는 무관한 순수한 관계라 할 수 있으므로  $R$ 이  $Y$ 에만 의존하는 이러한 경우를  $MNAR$ 로 식별할 수 있다.
  - (b-2) 만약  $X$ 와  $Y$ 가 통계적으로 유의한 연관성이 있다면,  $Y$ 의 무응답  $R$ 과  $Y$ 의 연관성은  $X$ 의 연관성에 기인하였을 수 있으므로  $MNAR$ 로 식별할 수 없으며, 이 경우  $MCAR$ 로 식별할 수 있다.

이에 대한 자세한 증명은 Jeon et al. (2016)에 제시되어 있으며 관심있는 독자는 저자에게 요청 및 참고하길 바란다. 한편,  $\hat{h}(\cdot, j)$  추정시,  $I > J$ 의 경우, 추정모수의 유일(unique) 해를 보장하기 위해  $X$ 의 범주들을 합쳐서  $I = J$  조건이 맞도록 바꾸거나,  $I$ 개 범주 중  $J$ 개 범주를 선택하여 여러 셋의  $J \times J$  형태로 바꾸는 등의 방법을 통해 해를 구할 수 있는데, 본 논문에서는 후자의 방법을 이용하여 변방값 발생을 정의하였다.  $X$ 와  $Y$ 에 대한 통계적으로 유의한 연관성에 대한 기준 역시 연구자에 따라 차이가 있을 수 있으나 본 논문에서는  $X, Y$  분할표에서의 카이제곱값을 기준으로 하였다. 이처럼 무응답 메커니즘에 따른 변방값 발

생 특성 및  $X, Y, R$  변수 간의 연관성 혹은 독립성 검정을 이용함으로써  $Y$ 의 무응답 메커니즘은 어느 정도 식별이 가능하다.

### 3. 무응답 대체 가중치

무응답의 편향을 줄이기 위한 대체방법의 하나로 무응답은 삭제하고 응답자에 가중치를 적용하는 방법이 제안되었으며, 이때 가중치는 출구조사에서 조사되는 표본이 모집단을 충분히 반영하는 동질적 분포라는 가정하에 등록된 선거명부를 이용하거나 실제 투표자 수를 이용하여 계산하고, 이것을 이용하여서 대체결과를 비교 및 보정하는 방법이다(Pavia 2010). 대표적인 가중치로는 대표표본(Representative Sample: RS) 가중치와 표본설계(Sampling Design: SD) 가중치가 있으며, 본 논문에서는 표본투표소의 투표율을 반영할 수 있는 SD 가중치를 이용하여 각 지역구별 후보자의 득표율은 다음의 식(4)와 같이 조정한다.

$$\hat{\pi}_j^{SD} = \sum_l \frac{N_l}{N\phi_l} \hat{\pi}_{jl} / \sum_l \frac{N_l}{N\phi_l}. \quad (4)$$

단,  $\hat{\pi}_{jl}$ 은  $j$ 번째 후보자에 대한  $l$ 번째 투표소의 추정치이며,  $N_l$ 은  $l$ 번째 투표소 선거명부에 등록된 유권자의 수이고,  $\phi_l$ 는  $l$ 번째 투표소에 등록된 유권자  $N_l$  중 실제 투표한 사람  $n_l$ 의 비율( $= n_l / N_l$ )이다. 즉, SD 가중치는 선거구 내 등록된 선거인 중 실제 투표한 사람들의 투표율을 반영하여 각 후보자에 대한 투표자 수를 합산하는 방법이다.

### 4. 출구조사 예측 정확도

출구조사 예측의 정확도는 예측오차를 이용하여 평가할 수 있으며, 이때 예측오차는 당선 여부를 결정하는 주요 관심대상인 1, 2위 후보자에 대해 함께 고려하는 것이 바람직할 것이다(Martin et al., 2005; 전세봄 외 2014). 따라서 본 논문은  $i$ 번째 지역구의 1, 2위 득표율의 차이에 대한 예측오차를 다음과 같이 평균절대

오차(Man Absolute Error: MAE), 평균절대비율오차(Man Absolute Percentage Error, 이하 MAPE)로 평가하였으며 이를 수식으로 나타내면 각각 식(5)와 식(6)과 같다. 즉  $MAE$ 가 득표율 차이의 예측오차를 실제값과 예측값의 편의(bias) 측면에서 평가하는 데 비해,  $MAPE$ 는 이러한 편의를 실제 득표율 차이로 나누어 예측오차를 평가하므로 박빙인 경우에 더욱 효과적인 지표라 할 수 있다. 한편 1, 2위 득표율 차이에 대한 예측오차에 존재할 수 있는 상쇄효과를 피하기 위해 본 논문에서는 각 후보자에 대한 예측오차를 반영하는 수정-평균절대오차(Adjusted-MAE: Adj-MAE)와 수정-평균절대비율오차(Adjusted-MAPE: Adj-MAPE)를 정의하였으며, 각각 식(7)과 식(8)과 같다. 즉  $Adj-MAE$ 가 각 후보자의 예측오차를 실제값과 예측값의 편의(bias) 측면에서 평가하는 데 비해,  $Adj-MAPE$ 는 이러한 편의를 실제 득표율로 나누어 예측오차를 평가한다. 본 논문에서는 이러한 4가지 통계량을 이용하여 무응답 처리방법별 예측력을 비교하고자 한다.

$$MAE = \sum_{i=1}^I \left| (\hat{p}_{i1} - \hat{p}_{i2}) - (p_{i1} - p_{i2}) \right| / I, \quad (5)$$

$$MAPE = \sum_{i=1}^I \left| \frac{(\hat{p}_{i1} - \hat{p}_{i2}) - (p_{i1} - p_{i2})}{(p_{i1} - p_{i2})} \right| / I, \quad (6)$$

$$Adj-MAE = \sum_{i=1}^I (|\hat{p}_{i1} - p_{i1}| + |\hat{p}_{i2} - p_{i2}|) / I, \quad (7)$$

$$Adj-MAPE = \sum_{i=1}^I \left\{ \left| \frac{\hat{p}_{i1} - p_{i1}}{p_{i1}} \right| + \left| \frac{\hat{p}_{i2} - p_{i2}}{p_{i2}} \right| \right\} / I. \quad (8)$$

한편, 출구조사의 예측오차는 표본투표소 추출 과정이나 투표소 내 출구조사 대상자 추출 과정, 또는 무응답 혹은 자료처리 과정 등 어떠한 출구조사 단계에서도 발생할 수 있으며, 만약 이러한 오차 간에 상쇄효과가 있다면 예측오차는 작게 나타날 가능성도 있으므로 이를 발생 요소별로 세분화하여 예측력을 평가하는 것이 바람직하다고 지적되고 있다(김영원 & 곽은선 2010; 현정보 2005). 이들은 예측오차  $\hat{p} - p$ 는 출구조사 절차에 따라 발생하는 오차를 구분하여 실사오차

$\hat{p}-\tilde{p}$ 와 투표소 추출오차  $\tilde{p}-p$ 로 정의하였으며, 이때  $p$ 는 실제 개표결과 나타난 득표율을 의미하고,  $\hat{p}$ 은 출구조사 자료로부터 예측한 득표율을,  $\tilde{p}$ 은 표본으로 추출된 투표소들의 개표결과로 계산한 득표율을 의미한다. 투표소 추출오차는 각 지역구 내 출구조사 대상 투표소의 실제 득표율의 차이를 알 수 있으므로 표본으로 추출된 투표소의 대표성이 우수한지 여부를 판단할 수 있으며, 실사오차를 통해 표본 투표소로부터의 출구조사 예측결과가 실제 득표율과 얼마나 차이가 나는지 등을 판단할 수 있다. 따라서 본 논문에서는 선거예측에 실패한 지역구의 경우, 투표소 추출오차와 실사오차 측면으로 나누어 세부적으로 예측오차를 평가하였다.

### III. 실증 분석

#### 1. 제 19대 총선 출구조사

본 논문은 2012년 제19대 총선 출구조사 자료를 이용하여 제안된 무응답 메커니즘에 대한 식별방안 및 무응답 대체방법을 평가하였다. 제19대 총선에서는 지상파 방송 3사가 공동으로 한국방송협회 산하 방송사공동예측조사위원회(Korea Election Poll: KEP)를 구성하고 미디어리서치, 코리아리서치센터, TNS-RI에 의뢰하여 출구조사를 실시하였다. 출구조사는 투표당일인 2012년 4월 11일 오전 6시부터 오후 5시까지 246개 선거구 내 총 13507개 투표소 중 표본으로 추출된 2495개 투표소에서 조사되었으며, 실제 응답자는 표본투표소에서 나오는 매 5번째 투표자를 등간격으로 체계적으로 추출하여 조사하였다. 출구조사 총 응답자 수는 858,324명이었으며 성별, 연령 결측치를 제외한 817,158명의 자료가 분석에 사용되었다. 연령은 19~29, 30~39, 40~49, 50~59, 60세 이상으로 총 5개의 연령범주로 구분하였으며, 지지후보자의 경우, 최종 당선 여부가 1, 2위 중에 결정되므로 나머지 후보자들은 모두 기타로 통합하여 총 세 개의 범주로 나누었다. 총 817,158명의 응답자 중 지지후보자를 밝히지 않은 무응답자는 142,432명으로 전체 응답의 약 17%를 차지하며, 실사오차의 상당부분을 차지하고 있어 무응답

대체를 통해 예측결과를 개선할 필요가 있음을 확인할 수 있었다.

## 2. 무응답 메커니즘 식별 및 대체

앞장에서 우리는 선거예측시, 무응답 메커니즘 대체는 가정하는 메커니즘에 따라 결과가 달라지므로 무응답 메커니즘의 식별이 우선되어야 함을 논의하였다. 따라서 본 장에서는 정리 1의 무응답 메커니즘 식별기준에 따라 총 2,495개 투표소에 대한 무응답 메커니즘을 식별하였다. 그런데 출구조사에서 관측된 연령범주는 5개로 대체하고자 하는 지지후보의 범주 수인 3개 보다 많아  $I > J$ 이므로,  $\hat{h}(\cdot, j)$ 를 추정함에 있어 유일해(unique solution)를 보장하지 못한다. 따라서 본 논문에서는,  $\hat{h}(\cdot, j)$ 의 유일해를 구하기 위해  $I$  범주 중  $J$ 를 선택할 수 있는 가능한 모든 연령범주 조합  ${}_5C_3$ 에서, 모든 선택범주 조합에서  $\hat{h}(\cdot, j)$  추정치에 변방값이 발생한 경우를 *MAR*로 식별하였다. *MAR*이 아닌 나머지 투표소 역시 정리 1에서 제안한  $X, Y, R$ 에 대한 독립성검정 결과에 따라 *MNAR* 및 *MCAR*로 식별할 수 있었다. 그 결과 *MAR*로 식별된 투표소는 2,297개였으며, *MNAR*로 식별된 투표소는 52개, *MCAR*로 식별된 투표소는 147개였다.

일반적으로 출구조사의 예측결과는 정당별 예상의석 수 등, 즉 1위 당선자를 올바르게 가려내는지 여부가 가장 우선적으로 평가된다. 이때 1, 2위 순위 예측도 중요하지만 예측범위 등 통계적 유의성을 파악하기 위해서는 1, 2위 후보자의 득표율을 얼마나 정확하게 예측하는지가 중요하다. 이에 본 논문에서는 무응답에 대한 처리방법을 평가함에 있어, 순위에 대한 예측오류뿐 아니라 득표율 추정치에 대한 예측오류를 확인하였다.

<표 4>는 무응답 대체방법별 순위 및 득표율에 대한 예측오류를 나타낸다. (0,0)은 순위를 올바르게 맞추는 순위예측에도 성공하고 득표율에 대한 구간예측 역시 성공하였음을 나타낸다. (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 구간예측모두 실패한 선거구의 수를 나타낸다. 이때 구간예측은 1, 2위 후보자 모두 실제 득표율이

신뢰구간 내에 포함되는 경우만을 구간예측 성공으로 정의하였다.

<표 4>에 제시된 무응답 대체방법의 결과는 10가지이다. 첫 번째는, 본 연구 결과의 비교를 위해 *KEP* 조사기관의 추정득표율을 기준으로 제시하였고, 그 이하는 무응답 메커니즘의 분류 및 추정방법에 따라 차례대로 제시한 결과이다. 우선 두 번째부터 여섯 번째까지는 투표소들의 무응답 메커니즘을 식별하지 않고 모든 투표소에 *MCAR*을 가정하거나, *MAR*을 가정하거나, *MNAR*을 가정한 결과들을 의미한다. 두 번째 *MCAR*는 무응답이 완전히 랜덤하게 발생한다고 가정하여 무응답을 제거한 *MCAR* 가정하의 대체 결과를 의미한다. 세 번째 *MAR*는 무응답이 성별과 연령에 의존하는 *MAR* 가정하의 결과이다. *MCAR* 및 *MAR*의 경우 추정방법에 따라 그 결과에 차이가 없는 반면, *MNAR*의 경우는 변방값 발생으로 인해 추정방법에 따라 결과가 다르게 나타나므로 이들 결과를 비교하기 위해 *EM* 추정결과 및 베イズ 추정결과를 제시하였다. 우선 네 번째 *MNAR<sub>EM</sub>*는 무응답이 지지후보에 의존하는 *MNAR* 가정하의 결과로, *MNAR* 모형을 *EM* 방법으로 추정한 *ML* 결과이다. 다섯 번째 *MNAR<sub>EM+Bayes1</sub>*와 여섯 번째 *MNAR<sub>EM+Bayes2</sub>* 결과는 *MNAR* 가정하에서 변방값 문제가 발생한 투표소는 베イズ 추정한 결과이다. 이때 Bayes1과 Bayes2는 각각 여러 베이지안 방법 중 Kim & Park(2015) 연구에서 가장 우수하였던 Clogg et al.(1991)의 사전분포와 Park & Brown(1994)의 사전분포를 의미한다.

일곱 번째 이하 *MECHS*는 앞서 제안한 메커니즘 식별결과에 따라 식별된 각 무응답 메커니즘별로 추정 및 대체하였음을 나타낸다. 우선 일곱 번째 *MECHS<sub>MM</sub>*는 각 투표소마다 식별된 무응답 메커니즘에 따라 *MM* 대체한 결과이고, 여덟 번째 *MECHS<sub>MM+sd</sub>*는 *MM* 대체결과에 sd 가중치를 적용한 경우이다. 마지막으로 *MECHS<sub>MM+Bayes1</sub>*, *MECHS<sub>MM+Bayes2</sub>* 결과는 각각 *MNAR*로 식별된 투표소에 한해 *MM* 대신 베イズ 추정 Bayes1과 Bayes2를 사용한 경우이다. 즉 마지막 두 방법은 다른 메커니즘 추정결과는 동일하고 *MNAR* 투표소에 대해서만 베イズ 추정법으로 바꾼 경우이므로, *MECHS<sub>MM</sub>* 결과와 비교하면 *MNAR*투표소에 대한 *MM* 추정과 베イズ 추정의 예측 성능을 비교할 수 있다.

〈표 4〉 무응답 대체방법별 예측에 실패한 선거구 수(0=예측성공, 1=예측실패)(단위: 개)

| 무응답 대체방법                             | (순위예측, 구간예측) |       |       |       | 순위실패<br>(1,+) | 구간실패<br>(+,1) |
|--------------------------------------|--------------|-------|-------|-------|---------------|---------------|
|                                      | (0,0)        | (0,1) | (1,0) | (1,1) |               |               |
| 1 <i>KEP</i>                         | 184          | 45    | 13    | 4     | 17            | 49            |
| 2 <i>MCAR</i>                        | 146          | 78    | 6     | 16    | 22            | 94            |
| 3 <i>MAR</i>                         | 184          | 44    | 13    | 5     | 18            | 49            |
| 4 <i>MNAR<sub>EM</sub></i>           | 95           | 105   | 4     | 42    | 46            | 147           |
| 5 <i>MNAR<sub>EM+ Bayes1</sub></i>   | 108          | 98    | 4     | 36    | 40            | 134           |
| 6 <i>MNAR<sub>EM+ Bayes2</sub></i>   | 101          | 103   | 3     | 39    | 42            | 142           |
| 7 <i>MECHS<sub>MM</sub></i>          | 201          | 30    | 13    | 2     | 15            | 32            |
| 8 <i>MECHS<sub>MM+ sd</sub></i>      | 197          | 33    | 14    | 2     | 16            | 35            |
| 9 <i>MECHS<sub>MM+ Bayes1</sub></i>  | 197          | 30    | 17    | 2     | 19            | 32            |
| 10 <i>MECHS<sub>MM+ Bayes2</sub></i> | 196          | 31    | 17    | 2     | 19            | 33            |

〈표 4〉에서 *MECHS<sub>MM+ sd</sub>*를 제외한 나머지 모든 대체방법의 결과는 투표율에 관계없이 지역구 내 각 후보자의 투표소별 추정득표 수를 모두 합해 투표자수로 나누어주는 방식으로 선거구별 후보자의 추정득표율을 산출한 후 실제득표율과 비교하여 순위 및 구간예측 성공 및 실패여부를 나타내었다. 예측에 실패한 선거구의 수 측면에서 볼 때, 각 투표소의 무응답 메커니즘을 식별하지 않고 모두 *MCAR*로 가정하여 대체하거나 모두 *MNAR*로 가정하여 대체하는 경우는, *KEP*가 예측한 결과에 비해 순위 및 구간 예측 정확도가 상당히 떨어지는 것을 볼 수 있다. 이는 특히 *MNAR* 가정하에 *EM*으로 예측 시 변방값 문제에 빠져 올바른 추정을 하지 못했기 때문으로 보인다. 이에 따라 변방값 문제에 빠진 경우 베이스 추정법을 고려한 결과, 정확도를 어느 정도 개선시키고 있으나 여전히 예측력은 떨어지는 것을 볼 수 있다. 모두 *MAR*로 가정하는 경우는 꽤 우수한

정확도를 보이고 있으나, 투표소별로 무응답 메커니즘을 식별하여 각각의 적률 추정법으로 대체한 *MECHS* 방법들의 예측 정확도가 더 개선된 것을 볼 수 있다.

순위예측에 실패한 선거구 (1,+) 및 구간예측에 실패한 선거구 (+,1) 측면에서 보면, *MECHS<sub>MM</sub>*이 모든 처리방법 중 가장 최소 수준의 예측실패를 보이고 있다. 순위예측과 구간예측을 모두 실패한 (1,1)의 경우, 7~10번의 메커니즘별 적률 추정결과들은 단 2개 투표소에 불과하므로 본 논문에서 제시하는 *MM* 추정법 및 무응답 메커니즘 식별기준이 선거예측 결과를 개선시키고 있음을 볼 수 있다. 특히 *MECHS<sub>MM</sub>*의 경우, 순위예측뿐 아니라 구간예측 측면 역시 예측실패 수가 최소수준이었다. 예측에 성공한 선거구의 수 측면에서 보면, 1,2위 후보자의 순위와 신뢰구간을 모두 정확히 예측한 (0,0)의 경우는, 전체 무응답 처리방법 중 *MECHS<sub>MM</sub>*이 201개로 가장 많았으므로 19대 총선 출구조사 예측에 가장 우수한 무응답 처리방법이라고 할 수 있다. 마지막으로, *MNAR*에서의 추정방법만 *MM* 대신 베이지 방법으로 바꾼 *MECHS<sub>MM+Bayes1</sub>*, *MECHS<sub>MM+Bayes2</sub>* 경우는 오히려 정확도가 떨어진 것으로 보아 *MNAR*에 대한 추정법으로도 본 연구에서 제안하는 적률추정법이 더 우수한 것을 확인할 수 있다.

이러한 연구결과는 기존의 전세봄 외(2014)에서 제안하였던 연령별 지지율 차이에 따라 무응답 메커니즘을 식별한 후 *ML* 추정을 통해 무응답을 대체하였던 방법과 비교해 볼 때, 순위예측 및 구간예측의 정확도 측면에서 더 우수한 결과를 보이고 있었다. 순위와 신뢰구간을 모두 정확히 예측한 (0,0)의 경우, 전세봄 외(2014)에서 제안하였던 식별 및 추정결과(*MAR+MNAR\_MLE+rs*의 경우, 195개 선거구)에 비해 본 연구방법의 *MECHS<sub>MM</sub>*는 201개 선거구로 정확도가 개선되었으며, 특히 구간예측을 실패한 (+,1) 선거구의 개수가 42개에서 32개로 상당히 줄어들어 구간예측 정확도가 개선되었음을 알 수 있었다. 일반적으로 선거예측에서의 어려움은 지지율의 차이가 오차범위 내인 경우 예측실패 가능성이 커진다는 점에서 볼 때, 본 논문의 무응답 메커니즘의 식별 및 추정방법으로 구간예측을 상당히 개선할 수 있으므로 선거예측의 정확도를 높이는 방법으로 적합하다고 판단된다.

&lt;표 5&gt; 무응답 대체방법별 예측력(246개 선거구, 1, 2위 후보만 고려)

|    | 무응답 대체방법                          | <i>MAE</i> | <i>MAPE</i> | <i>Adj - MAE</i> | <i>Adj - MAPE</i> |
|----|-----------------------------------|------------|-------------|------------------|-------------------|
| 1  | <i>KEP</i>                        | 0.0335     | 0.6898      | 0.0354           | 0.0864            |
| 2  | <i>MCAR</i>                       | 0.0431     | 0.6087      | 0.0451           | 0.1081            |
| 3  | <i>MAR</i>                        | 0.0348     | 0.4528      | 0.0368           | 0.0900            |
| 4  | <i>MNAR<sub>EM</sub></i>          | 0.0757     | 1.4829      | 0.0776           | 0.1824            |
| 5  | <i>MNAR<sub>EM+ Bayes1</sub></i>  | 0.0732     | 1.4491      | 0.0754           | 0.1775            |
| 6  | <i>MNAR<sub>EM+ Bayes2</sub></i>  | 0.0739     | 1.4648      | 0.0759           | 0.1788            |
| 7  | <i>MECHS<sub>MM</sub></i>         | 0.0345     | 0.4356      | 0.0365           | 0.0891            |
| 8  | <i>MECHS<sub>MM+ sd</sub></i>     | 0.0354     | 0.4155      | 0.0374           | 0.0911            |
| 9  | <i>MECHS<sub>MM+ Bayes1</sub></i> | 0.0353     | 0.4479      | 0.0371           | 0.0902            |
| 10 | <i>MECHS<sub>MM+ Bayes2</sub></i> | 0.0353     | 0.4479      | 0.0371           | 0.0902            |

<표 5>는 무응답 대체방법 간의 성능을 비교하기 위해 예측오류율을 평가한 결과이다. 우선 모든 오류율 측면에서 무응답 메커니즘을 모두 *MNAR*을 가정하여 *EM* 추정한 방법이 가장 나쁘고, 베イズ추정으로 변방값 문제를 해결한다 하더라도 전체적인 예측력은 크게 개선하지 못하고 있음을 볼 수 있다. 그 외의 대체방법들 중, *MAE*와 *Adj - MAE* 관점에서 가장 우수한 방법은 *KEP*라 할 수 있으며, *MECHS<sub>MM</sub>*도 이와 비슷한 낮은 수준의 예측 오류율을 보이고 있다. *MAPE* 관점에서 가장 최소의 오류율을 보이는 대체방법은 투표소별 투표율을 반영한 *MECHS<sub>MM+ sd</sub>*와 *MECHS<sub>MM</sub>*이었다. 그런데 *MAE*가 후보자 간 득표율 차이에 대한 예측오차 즉 편의(bias)를 나타내는 데 비해, *MAPE*는 이러한 예측 오차를 후보자 간 득표율 차이로 나누어 평가하므로 후보자들 간에 박빙인 경우 예측오류를 판단하기에 더 효과적이라 할 수 있다. 출구조사를 이용한 선거예측의 어려움은 두 후보자의 득표율 차이가 많이 나는 경우보다 박빙인 경우에 있으므로, 이러한 경우 표본을 더 확보하거나 대체하는 등의 노력을 기울인다는 점에서 볼 때, *MAPE* 측면에서 본 논문의 방법이 최소의 오류율을 보인다는 점은

의미가 있다고 할 수 있을 것이다. 한편  $MECHS_{MM+sd}$ 는  $MECHS_{MM}$  방법에  $sd$  가중치를 부여한 점만 다르므로 두 대체방법의 결과차이는 무응답 메커니즘 식별 및 추정방법에 기인한 것이 아니라 가중치 계산 즉 투표소별 투표율에 기인한 것이라 할 수 있다. 그밖의 대체방법별 결과는 투표율에 관계없이 지역구 내 각 후보자의 투표소별 추정득표 수를 모두 합해 투표자 수로 나누어주는 방식으로 선거구별 후보자의 추정득표율을 산출하였다.

<표 6> 대체방법별 예측력( $MNAR$  투표소를 갖는 42개 선거구, 1, 2위 후보만 고려)

| 무응답 대체방법               | $MAE$  | $MAPE$ | $Adj - MAE$ | $Adj - MAPE$ |
|------------------------|--------|--------|-------------|--------------|
| 1 $KEP$                | 0.0349 | 0.1798 | 0.0371      | 0.0920       |
| 4 $MNAR_{EM}$          | 0.0738 | 0.8279 | 0.0758      | 0.1819       |
| 5 $MNAR_{EM+Bayes1}$   | 0.0707 | 0.7854 | 0.0731      | 0.1757       |
| 6 $MNAR_{EM+Bayes2}$   | 0.0716 | 0.8052 | 0.0737      | 0.1773       |
| 7 $MECHS_{MM}$         | 0.0367 | 0.1046 | 0.0386      | 0.1000       |
| 8 $MECHS_{MM+sd}$      | 0.0350 | 0.1256 | 0.0380      | 0.1011       |
| 9 $MECHS_{MM+Bayes1}$  | 0.0376 | 0.1059 | 0.0402      | 0.1030       |
| 10 $MECHS_{MM+Bayes2}$ | 0.0374 | 0.1131 | 0.0407      | 0.1040       |

<표 6>의 결과는 전체 246개 선거구가 아닌, 선거구 내에  $MNAR$  투표소가 하나 이상 존재하는 42개 선거구를 대상으로 한 예측 정확도이다. 즉  $MNAR$  투표소를 포함하는 선거구에 한해 예측오류를 비교한 결과를 의미하므로,  $MNAR$ 을 가정한 추정결과들이 <표 5>에 비해 개선된 것을 볼 수 있다. 그러나 여전히 여러 지표들에서  $MNAR_{EM}$ ,  $MNAR_{EM+Bayes1}$ ,  $MNAR_{EM+Bayes2}$ 의 오류율이 높게 나타나고 있어, 무응답 메커니즘을 고려하지 않고 모든 투표소에  $MNAR$ 을 고려할 경우 예측력이 떨어질 것을 예상할 수 있다. 이러한 현상은 1, 2위만이 아니라 1, 2위 및 기타후보를 모두 예측하였을 때도 역시 마찬가지였으며, 따라서 본 논문에서는 기타후보까지 고려한 결과는 생략하였다.

&lt;표 7&gt; 순위 예측실패 선거구의 지역구 및 표본투표소 득표율 차이와 추정결과

| 선거구  | 선거구명         | $p_1 - p_2$ | $\tilde{p}_1 - \tilde{p}_2$ | $\hat{p}_1 - \hat{p}_2$ | 구간예측<br>실패여부 | KEP<br>당선자<br>예측오차 | MM추정<br>당선자<br>예측오차 |
|------|--------------|-------------|-----------------------------|-------------------------|--------------|--------------------|---------------------|
| 0122 | 서울시 은평구을     | 0.47        | 0.56                        | -1.11                   | 0            | -2.2               | -0.74               |
| 0124 | 서울시 서대문구을    | 0.41        | -0.50                       | -4.86                   | 0            | -3.0               | -2.64               |
| 0127 | 서울시 양천구갑     | 1.66        | 1.38                        | -2.94                   | 0            | -2.0               | -2.03               |
| 0130 | 서울시 강서구을     | 0.17        | 0.50                        | -1.18                   | 0            | -1.2               | -0.68               |
| 0145 | 서울시 송파구을     | 3.84        | 5.71                        | -2.45                   | 1            | -2.2               | -3.11               |
| 0147 | 서울시 강동구갑     | 3.72        | 3.35                        | -0.88                   | 0            | -2.1               | -2.14               |
| 0213 | 부산시 사하구갑     | 3.39        | 2.36                        | -0.40                   | 0            | -1.7               | -1.52               |
| 0501 | 광주광역시 동구     | -0.66       | 1.31                        | 3.19                    | 0            | -2.2               | 0.96                |
| 0601 | 대전광역시 동구     | 1.38        | 0.85                        | -2.65                   | 0            | -0.5               | -1.08               |
| 0810 | 경기도 의정부시을    | 3.31        | 4.11                        | -0.63                   | 0            | -2.2               | -1.74               |
| 0828 | 경기도 고양시 덕양구을 | 0.06        | -0.63                       | -4.10                   | 0            | -3.2               | -1.99               |
| 0850 | 경기도 광주시      | 1.41        | -0.59                       | -0.42                   | 0            | -1.2               | -0.56               |
| 0901 | 강원도 춘천시      | 4.62        | 4.45                        | -4.92                   | 1            | -2.4               | -4.58               |
| 1107 | 충청남도 논산계룡금산  | -2.17       | -1.02                       | 2.66                    | 0            | -3.1               | 1.48                |
| 1513 | 경상남도 거제시     | -2.35       | -2.84                       | 0.88                    | 0            | -2.1               | -0.30               |

\*  $p_1 - p_2$ : 실제 득표율 차이,  $\tilde{p}_1 - \tilde{p}_2$ : 표본투표소 득표율 차이,  $\hat{p}_1 - \hat{p}_2$ : MM추정 득표율 차이

마지막으로 순위를 가장 잘 맞춘  $MECHS_{MM}$  방법으로도 순위예측에 실패한 15개 선거구의 예측실패 원인을 분석하고 평가하고자 한다. <표 7>에서 15개 선거구는 모두 ‘오차 범위 내에서 경합’으로 판정된 지역구이다. 선거조사에서 경합이란 순위에 있어 우열을 가릴 수 없다는 의미이며, 오차범위 내 경합이란 우열을 가림에 있어 통계적으로 유의미한 차이를 판단하기 힘들다는 의미이다. 이들 선거구에 대한 예측실패가 무응답 처리 등 통계적 방법에 기인한 것인지, 혹은 표본투표소 추출 등에 기인하는 것인지 원인을 파악하기 위해, 예측오차를 실사오차와 표본오차로 나누어 세부적으로 파악하였다. 우선  $p_1 - p_2$ 는 1, 2위

후보의 선거구의 실제 득표율 차이,  $\tilde{p}_1 - \tilde{p}_2$ 는 1, 2위 후보의 표본투표소 득표율의 차이를 의미하며,  $\hat{p}_1 - \hat{p}_2$ 는 1,2위 후보에 대한  $MM$ 추정치에 rs 가중치로 보정한 추정 득표율 차이를 의미한다.

세 득표율 차이의 부호를 이용하여 순위예측의 실패를 확인 및 검토한 결과, 서울시 서대문구을, 경기도 고양시 덕양구을, 경기도 광주시, 그리고 광주광역시 동구 등 네 선거구의 경우, 표본으로 뽑힌 투표소의 추정치 차이가 실제 개표결과 득표율 차이와 부호가 서로 반대임을 알 수 있다. 이는 해당 선거구에서 추정할 득표율의 참값과 표본에 기인한 추정치의 순위가 서로 다르다는 의미로, 이렇듯 표본 자체가 잘못 추출되어 1, 2위 순위가 뒤바뀐 경우에는  $MAR$ ,  $MNAR$  등의 무응답 처리를 하더라도 방향을 뒤집기는 힘들다. 즉 이들 세 선거구의 예측실패는 무응답 처리방법에 기인하기보다는 표본투표소 추출에 기인한 것임을 알 수 있다. 이들을 제외한 나머지 순위예측에 실패한 선거구는 12개였으며, 이들 12개 선거구 중 10개 선거구는 모두  $MM$  추정에 기반한 구간예측은 성공하였으므로 통계적으로 올바른 예측을 수행하였음을 알 수 있다.

따라서 본 논문에서 제시한 무응답 메커니즘 식별 및 대체방법으로 순위예측 및 구간예측을 모두 실패한 선거구는 서울 송파구을과 강원도 춘천시 단 2개에 불과하므로 본 논문의 선거예측 개선안이 실제적으로 예측력 개선에 기여했음을 알 수 있다. 마지막으로 당선자 예측오차 측면에서 본 논문의  $MM$  추정결과와  $KEP$ 에서 예측한 당선자 득표율의 예측오차를 비교한 결과,  $MM$  추정결과가  $KEP$  예측결과보다 총 순위실패 수를 감소시켰을 뿐 아니라 순위에 실패한 개별 지역구의 당선자 예측오차 역시 대체로 줄어드는 것으로 보아 예측력을 개선시켰다고 할 수 있다.

#### IV. 결론 및 제언

본 연구는 출구조사의 예측력을 개선하고자 제19대 총선 자료를 이용하여 각 지역구 내 투표소들의 무응답 메커니즘을 식별하고 식별 결과에 따라 무응답을

대체(imputation)하는 새로운 방안을 제안하고 있다.

본 논문의 첫 번째 관심사였던 출구조사에서 무응답 메커니즘이 식별가능한가, 또한 식별된 메커니즘 가정이 적절한가에 대한 답을 위하여 본 연구는 무응답 메커니즘별 변방값 발생 특성 및 무응답 메커니즘의 정의를 이용하였다. 즉 무응답 메커니즘이 *MAR*일때 *MNAR* 모형을 적합하면 변방값 문제가 발생한다는 특성을 이용하여 변방값 문제의 발생여부에 따라 *MAR*여부를 판단하고, 무응답 메커니즘 정의에 따라 무응답이 순수하게 지지후보에만 의존하는 경우를 *MNAR*으로 식별, 그렇지 않은 경우를 *MCAR*로 식별하고 있다. 즉, 출구조사에서 연령별 투표 및 무응답 패턴을 이용하여 투표소의 무응답 메커니즘은 어느 정도 식별 가능하였으며, 이러한 무응답 메커니즘 식별에 따른 대체 결과 예측력이 개선됨을 확인할 수 있어, 식별된 메커니즘 가정이 적절하였음을 확인할 수 있었다.

두 번째 관심사였던, 식별된 메커니즘 가정하에서 출구조사 예측력이 가장 우수한 추정법이 무엇인가, 그러한 추정법이 변방값 문제에서 자유로운가에 답하기 위해 본 연구는 *MNAR* 메커니즘에서 흔히 발생하는 변방값 문제를 해결하기 위한 새로운 대안으로 적률추정치를 제안하였다. 그 결과 기존의 대체방법보다 출구조사 예측의 정확성이 향상된 것을 확인할 수 있었다.

한편, 본 논문에서 사용하는 출구조사 데이터는 투표 마감 전인 오후 5시까지의 데이터로, 5시 이후의 투표결과는 반영하고 있지 않아 실제 투표결과와는 다소 차이가 있을 수 있다. 비록 한시간 가량의 자료 부족으로 인한 결과 차이는 크지 않을 것으로 예상되나, 이러한 차이로 인해 사후분석 결과가 완벽하지 않을 수 있으므로 예측결과 자체에 대한 평가보다는 적절한 예측 개선을 위한 무응답 대체방법으로 활용될 수 있을 것으로 기대한다. 또한 변방값 발생을 정의함에 있어, 선택범주 중 몇 개 이상을 변방값으로 정의하는지에 따라 식별결과는 차이가 있을 수 있으므로 본 식별률에 대한 결과를 일반화하기에는 한계가 있을 수 있다. 그러나 본 논문은 무응답 메커니즘에 대한 고려없이 대체하거나, 혹은 특정 무응답 메커니즘 가정하에 추정하였던 기존의 대체방법에 비해, 무응답 메커니즘에 대한 식별 가능성을 제시하였다는 점에서 의의가 있다.

## 참고문헌

- 김영원·곽은선. 2010. “2008 총선 출구조사의 총조사오차 분석.” 《조사연구》 11(3): 33-55.
- 현경보. 2005. “선거예측조사의 정확도와 그 영향요인 연구: 17대 총선 예측조사 결과의 사후 분석.” 《언론과학연구》 5(1): 301-336.
- 전세봄·이수정·박유성. 2014. “출구조사에서 무응답 메카니즘 식별 및 대체에 대한 연구: 제19대 국회의원 선거 출구조사에서의 응용.” 《조사연구》 15(3): 61-83.
- Baker, S.G. and N.M. Laird. 1988. “Regression Analysis for Categorical Variables with Outcome Subject to Nonignorable Nonresponse.” *Journal of the American Statistical Association* 83: 62-69.
- Bautista, R., Callegaro M., Vera, J.A., and Abubdis, F. 2007. “Studying Non-responses 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.
- Clarke. P.S. 2002. “On the Boundary Solution and Identifiability in Categorical Regression with Non-ignorable Non-response.” *Biometrical Journal* 44(6): 701-717.
- Clarke, P.S. and W.F. Smith 2005. “On Maximum Likelihood Estimation for Loglinear Models with Non-ignorable Non-responses.” *Statistics & Probability Letters* 73: 441-448.
- Clogg, C.C., D.B. Rubin, N. Schenker, B. Schultz, and L. Weidman. 1991. “Multiple Imputation of Industry and Occupation Codes in Census Public-use Samples Using Bayesian Logistic Regression.” *Journal of the American Statistical Association* 86(413): 68-78.

- 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
- Green, P.E. and T. Park. 2003. "A Bayesian Hierarchical Model for Categorical Data with Nonignorable Nonresponse." *Biometrics* 59(4): 886-896.
- Glynn, R.J., N.M. Laird, and D.B. Rubin. 1986. *Selection Modeling Versus Mixture Modeling with Nonignorable Nonresponse*. Springer New York.
- 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.
- Jeon, S., T. Kwon, and Y. Park. 2016. "Identification of Missing Mechanism and Moment Estimators in Incomplete Two-way Contingency Tables with Unit Missing Cases." *Journal of the American Statistical Association*. (Submitted)
- Kenward, M.G. and G. Molenberghs. 1999. "Parametric Models for Incomplete Continuous and Categorical Longitudinal Data." *Statistical Methods in Medical Research* 8(1): 51-83.
- Kim, S. and Y. Park. 2015. "Power-linear Models for Incomplete Contingency Tables with Nonignorable Nonresponses." *Statistics* 49(6): 1189-1203.
- Little, R.J. 1994. "A Class of Pattern-mixture Models for Normal Incomplete Data." *Biometrika* 81(3): 471-483.
- Little, J.A. and D.B. Rubin. 2002. *Statistical Analysis with Missing Data*(2nd ed.), Wiley, New York.
- 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.
- Park, Y., D.Kim, and S. Kim. 2014. "Identification of the Occurrence of Boundary Solutions in a Contingency Table with Nonignorable Nonresponse." *Statistics & Probability Letters* 93: 34-40.
- Pavia, J.M. 2010. "Improving Predictive Accuracy of Exit Polls." *International Journal of forecasting* 26: 68-81.
- Smith, P.W.F., C.J. Skinner, and P.S. Clarke. 1999. "Allowing for Non-ignorable Non-response in the Analysis of Voting Intention Data." *Applied Statistics* 48: 563-577.

<접수 2016/04/05, 수정 2016/06/02, 게재확정 2016/06/24>