Journal of the Korean Data & Information Science Society
2014, **25**(1), 53–64

http://dx.doi.org/10.7465/jkdi.2014.25.1.53
한국데이터정보과학회지

무응답모형에 기반한 출구조사의 예측 정확성 비교 연구[†]

곽정애[1] · 최보승[2]

[1]대구대학교 대학원 통계학과 · [2]대구대학교 전산통계학과

접수 2013년 11월 27일, 수정 2013년 12월 17일, 게재확정 2013년 12월 23일

요 약

조사를 통한 선거 예측을 수행하는 데 있어서 발생할 수 있는 문제점 가운데 하나는 무응답이라 할 수 있으며 무응답 대체에 대한 방법에 따라 예측 결과는 완전히 다른 결과를 생산해 낼 수 있다. 특히 대통령 선거와 같은 민감한 주제에 대한 선거에서는 무응답 대체가 더욱 더 중요하다. 본 연구에서는 무응답 대체의 방법으로 모형에 기반을 둔 대체 방법에 대하여 연구를 진행하였다. 모형에 기반을 둔 대체 방법에서는 무응답 체계의 가정에 따라 무응답 모형을 구축할 수 있으며 무응답 체계에 따라 각기 다른 대체 결과를 제공할 수 있다. 모형에 기반을 둔 무응답 대체 및 추정에서 적절한 무응답 체계의 가정은 정확한 모형 추정을 위한 매우 중요한 전제 조건이다. 그러나 무응답 체계의 가정에 대한 검증 절차는 아직 정확한 해법이 알려지지 않은 상황이다. 본 연구에서는 실제 자료를 이용한 모형 적합을 통하여 무응답 체계 가정에 대한 정확도를 비교하고자 하였다. 2012년에 시행된 18대 대통령 선거과정에서 수행된 출구조사 결과를 이용하여 무응답 체계의 가정에 대한 검증과 모형에 의한 예측 정확도를 비교하였다. 무응답 모형의 추정과 무응답 대체를 위하여 EM 알고리즘에 기반을 둔 최대우도 추정방법을 이용하였으며 예측 결과를 비교하기 위하여 Bautista 등 (2007)이 제안한 MWPE (modified within precinct error)를 이용하였다.

주요용어: 무응답 체계, 선거예측, 출구조사

1. 서론

대한민국에서는 지난 2012년에 18대 대통령 선거가 치러졌으며 투표 당일 공중파 방송 3사 (KBS, MBC, SBS)는 역대 최대규모의 출구조사를 수행하였다. 더욱 정확하고 빠른 선거예측 결과를 생산하기 위하여 대단위 출구조사가 막대한 비용과 인력을 투입하여 수행되었다. 그 어느때 보다 당선자 예측이 어려운 박빙의 선거이었기에 출구조사에 의한 예측은 더욱이 관심을 두게 되었다. 출구조사를 통한 예측 결과는 새누리당 박근혜 후보 50.1% 대 민주통합당 문재인 후보 48.9%의 지지율로 예측하였고 출구조사는 95%의 신뢰도의 표본오차는 ±0.8%이다. 전화 여론 조사를 시행한 YTN의 결과는 새누리당 46.1% 에서 49.9%, 민주통합당 49.7% 에서 53.5%로 발표되었다. 18대 대통령 선거의 최종 결과는 새누리당 51.6% 대 민주통합당 48% 로 새누리당 박근혜 후보가 18대 대통령에 당선되었다. 언론기관과 여론조사기관들은 비교적 정확한 예측 결과를 생성하였다고 할 수 있다.

그러나 막대한 비용과 인력이 투입된 출구조사에서도 일정 정도의 오차가 발생하였다. 오차가 발생하는 이유는 여러 가지가 있는데 그 이유로는 투표소 추출방법, 선거일에 투표소로부터 300미터 이내에서

[†] 본 연구는 대구대학교 교내 연구비를 지원 받아 수행된 연구임 (No.20120481).
[1] (712-714) 경상북도 경산시 진량읍 대구대로 201, 대구대학교 대학원 통계학과, 석사과정.
[2] 교신저자: (712-714) 경상북도 경산시 진량읍 대구대로 201, 대구대학교 전산통계학과, 조교수.
E-mail: bchoi@ daegu.ac.kr

는 조사가 금지인 법적 문제, 조사원 선정과 훈련, 응답자들의 선정, 조사 거절과 거짓응답, 출구 조사 방법, 무응답 등으로 여러 원인에 의해 예측 오차가 발생할 수 있다 (Ryu, 2000; Hong과 Huh, 2001; Hyun, 2005; Kim과 Kwak, 2010). 이러한 오차의 원인 가운데 출구 조사에서 조사 거절을 하는 무응답률의 경우 우리나라는 대략 15% 에서 20% 수준으로 무응답의 대체 결과에 따라 출구 조사 예측률이 크게 달라지기 때문에 오차가 발생할 수 있는 원인 중 큰 부분을 차지한다고 할 수 있다 (Rhee, 2004; Kim과 Kwak, 2010; Kim과 Choi, 2011). 무응답 처리를 포함한 예측문제에 대하여 그동안 국내외에서 많은 연구가 진행되어 왔다. Ryu (2000)와 Kim과 Kim (2007)은 무응답자의 경우 해당 성별·연령대별 지지율 그대로 해당 범주의 지지율로 무응답을 처리하였다. Ryu (2003)은 선거 예측 조사에서 얻어진 자료를 통해 가중치 조정에 대한 방법을 설명하였다. Crespi (1998)의 연구에서는 무응답 대체의 방법으로 주요 2개 후보에게 비례배분하는 방법, 주요 2개 후보에게 반으로 나누어 배분하는 방법, 현직 후보자가 있다면 그 외의 도전자 후보에게 배분하는 방법, 무응답을 버리고 후보자들의 득표율을 재계산하는 방법 등4가지 방법을 제시하였고, 그 가운데 비례배분이 가장 좋다는 의견을 제시하였다. Kim (2000)은 무응답 대체 방법과 대체 효과에 대하여 무응답 대체 방법을 결정적인 대체방법과 확률적인 대체 방법에 대한 여러 가지 방법을 설명하였다. Lee 등 (2006)에서는 2002년 강원지역의 농가경제 자료를 예제로 하여 공간상관을 이용한 무응답 대체 방법을 사용하였다. Cho 등 (2008)에서는 농촌 생활지표조사에서의 무응답을 연속형과 범주형으로 나누어 연속형 무응답일 경우 평균 대체, 회귀 대체등을 이용하였고 범주형 무응답 일 경우 최빈값 이용, 확률 대체 등을 이용하여 대체·비교 하였다. Lee와 Kang (2012)에서는 무응답을 단위 무응답과 항목 무응답으로 나누어 실제 사례를 이용하여 비교하였다. 이러한 방법들은 표본조사의 정보를 이용하여 무응답 대체를 수행 하는 것이다. 이와는 다른 접근 방법으로 모형적 접근에 의한 무응답 대체 연구가 진행되어 왔다. Baek 등 (2002)은 한국 노인 약물 역학 코호트 자료를 평가하기 위해 로그선형모형에서 모수의 최대우도 추정치를 구하고자 EM 알고리즘을 이용하여 추정하였다. Park과 Brown (1994), Choi 등 (2007), Choi 등 (2009), Park과 Choi (2010)등은 로그선형모형에서 다항 분포를 가정하고 각 칸의 기대확률에 대한 사전분포로 Direichlet 분포를 할당하는 경험적 베이지안 방법을 이용하였다. Yoon과 Choi (2012)에서는 무응답을 포함하는 다차원 분할표 형태에 대한 모형 추정 방법으로 계층적 베이지안 방법을 제안하였고 조건부 사후분포로부터 모수를 추출 하기 위한 MCMC 방법을 제시하였다.

이러한 모형에 기반을 둔 무응답 대체를 수행하기 위해서는 적절한 무응답 체계에 대한 가정이 필요하다. Little과 Rubin (2002)은 무응답을 발생 체계에 따라 크게 세 가지로 구분하였다. 첫 번째는 완전임의결측 (missing completely at random; MCAR)으로 무응답의 발생 여부가 무응답을 가지는 변수나 함께 조사된 다른 그 어떤 변수에도 영향을 받지 않는 경우이다. 두 번째는 임의결측 (missing at random; MAR)은 무응답의 발생 여부가 무응답을 가지고 있지 않은 관찰된 자료에 의해서만 영향을 받는 경우이다. 이 두 가지 무응답들은 무응답의 발생 여부가 무응답 자체로부터 아무런 영향을 받지 않은 경우로 무시할 수 있는 무응답 (ignorable nonresponse)이라 한다. 세 번째는 비임의결측 (missing not at random; NMAR)으로 무응답 발생 여부가 무응답 자체에 영향을 받는 것으로 무시할 수 없는 무응답 (nonignorable nonresponse)이라 한다. 예를 들어 자신이 지지하는 후보를 밝히지 않았을 경우 특별한 이유가 없다면 무시할 수 있는 무응답이라 할 수 있고, 자신의 지지 후보가 그 지역의 열세 후보여서 밝히지 않았다면 무시할 수 없는 무응답이라고 할 수 있다.

적절한 무응답 체계에 대한 가정은 정확한 무응답 대체와 예측결과를 나타내기 위한 매우 중요한 절차라 할 수 있다. 그러나 현재까지 무응답 체계의 선택을 평가하는 방법에 대한 연구는 그리 많지 않으며 그 어떠한 연구에서도 확실한 해답을 제시하고 있지 않은 형편이다 (Choi 등, 2007; Choi 등, 2008; Choi와 Kim, 2012; Yoon과 Choi, 2012). 일반적인 모형 선택의 방법에 기반을 두어 무응답 체계에 대한 결과를 평가하는 연구가 진행된 경우가 있다. Choi와 Kim (2012)은 경험적 베이지안 모형을 이용

한 무응답 모형 추정에서 EM 알고리즘에 기반을 둔 모형 선택의 연구를 진행하였으며 Yoon과 Choi (2012)는 계층적 베이지안 모형을 이용하여 무응답 모형을 구축하였고 MCMC 방법에 따른 모형 추정과 베이즈 인자 (Bayes factor)의 계산을 통한 모형 선택의 방법을 제안하였다. 그러나 Ibrahim 등 (2008)의 연구에 의하면 모형 선택적 방법을 통하여 임의결측과 비임의결측의 비교를 직접 수행하는 것은 위험할 수 있다고 경고하고 있다.

본 연구에서는 무응답 체계에 대한 평가를 실제 자료를 이용한 모형적합을 통하여 예측 정확도에 기반을 둔 평가를 이용하여 진행하고자 한다. 전통적으로 표본을 기반으로 하여 자료 대체나 가중치를 두는 방법 (Ryu, 2000; Kim과 Kim, 2007)이 아니라 실제 관찰된 자료를 이용하여 무응답 모형을 기반으로 하여 연구해 보고자 한다. 모형을 기반으로 한 무응답은 무시할 수 있는 무응답과 무시할 수 없는 무응답으로 나눌 수 있으며 실제로 어떤 모형을 더 따르는지 판단하게 된다. 무응답 체계가 정확히 가정되었다면 그 가정에 기반을 둔 예측모형의 결과가 잘못된 가정에 의한 예측결과에 비하여 좋을 것으로 기대할 수 있다. 이를 검증하기 위하여 지난 2012년 대통령 선거 당일 수행된 출구조사 결과를 이용하여 예측 정확도를 평가하고 무응답 체계에 대한 가정 선택을 평가한다. 본 연구의 진행은 다음과 같다. 먼저 2절에서는 무응답 모형의 정의와 함께 무응답 대체 및 모형 추정 그리고 예측 결과의 평가 방법에 대하여 설명한다. 3절에서는 실제 자료를 이용한 분석 결과를 제시하고 마지막 4절에서는 결론으로 본 연구 방법의 한계점과 추후 진행방향에 대하여 논하고자 한다.

2. 무응답 모형의 추정방법및 평가방법

2.1. 무응답 모형

일반적으로 수집된 자료의 반응변수와 설명변수가 모두 범주형 자료라면 다차원 분할표로 표현할 수 있다. 이렇게 만들어진 다차원 분할표에 무응답이 포함되어 있다면 로그선형 모형을 이용하고 모수 추정하기 위해 EM 알고리즘을 사용한다. (Fay, 1986; Baker와 Laird, 1988; Chambers와 Welsh, 1993; Park과 Brown, 1994; Park과 Lee, 1998)

먼저 본 연구에서 사용한 모형에 대한 정의는 다음과 같다. 선거예측을 위한 수집된 관찰변수 중 지지후보는 반응변수 Y이며 J개의 범주를 가지고 있고 여기에 영향을 주는 요인으로 연령을 고려 할 때 이는 설명변수이며 X로 표시한다. 이때 X는 I개의 범주를 가지고 있다. 설명변수 X에 대해서는 결측치를 가지지 않고 반응변수 Y에서만 결측치를 가지고 있다고 가정한다. 그리고 지지후보의 무응답 여부는 R로 $R=1$이면 응답, $R=2$이면 무응답으로 나타낸다. X, Y, R로 구성되고 무응답이 있는 3차원 분할표가 된다. 변수 X가 i번째 범주를 가지고 있고 Y가 j번째 범주를 가질 때 이는 응답 자료를 나타내고 칸 빈도수를 y_{ij1}로 표시할 수 있다. 변수 X에서만 i번째 범주를 가지고 Y가 무응답을 가질 때 칸 빈도수를 y_{i+2}로 표시한다. 이때, X의 범주가 2개이고 Y의 범주가 2개일 때의 무응답이 있는 분할표는 Table 2.1 같이 나타낼수 있다.

Table 2.1 Two-way coitngency table with supplemental margin

	Response $R=1$		Non-response $R=2$
	Candidate A ($Y=1$)	Candidate B ($Y=2$)	
$X=1$	y_{111}	y_{121}	y_{1+2}
$X=2$	y_{211}	y_{221}	y_{2+2}

관찰된 빈도에 대하여 다항분포를 가정하고 로그우도함수는 다음의 식에 비례한다.

$$l \propto \sum_i \sum_j y_{ij1} \log(\pi_{ij1}) + \sum_i y_{i+2} \log(\pi_{i+2}). \tag{2.1}$$

여기서 $\pi_{ij1} = Pr(X = i, Y = j, R = 1)$, $\pi_{i+2} = Pr(X = i, R = 2)$이고 $N_1 = \sum_i \sum_j y_{ij1}$, $N_2 = \sum_i y_{i+2}$, $N = N_1 + N_2$이고 각 칸의 기대빈도를 $\mu_{ijk} = N \times \pi_{ijk}$ 이라 하고 이 기대빈도에 대한 로그선형모형은 다음과 같이 나타낼 수 있다.

$$\log(\mu_{ijk}) = \beta_0 + \beta_X^i + \beta_Y^j + \beta_R^k + \beta_{XY}^{ij} + \beta_{XR}^{ik}. \tag{2.2}$$

여기서 무응답 여부에 대한 지시변수 R과의 어떠한 상호작용이 포함되지 않은 모형은 완전임의결측 (MCAR)이 된다. 완전임의결측 모형에서 설명변수 X와 무응답 여부 R간의 상호작용 효과 (XR)가 포함이 된다면 임의결측 (MAR)모형이 된다. 이 두 모형은 모두 무응답 여부가 무응답과 관계가 없으므로 무시할 수 있는 무응답 모형 (ignorable nonresponse model)이라 할 수 있다. 여기서 만약 설명변수와 무응답 지시변수 간의 상호작용에 대한 모수 대신에 반응변수 Y와 무응답 지시변수 R의 상호작용 효과 YR이 모형에 포함되면 이는 비임의결측 (NMAR)이 되며, 이 때는 무응답 여부가 무응답과 관계가 있으므로 무시할 수 없는 무응답 모형 (nonignorable nonresponse model)이 된다 (Little과 Rubin, 2002). Table 2.1과 같은 자료에 대한 무응답 모형을 고려한 경우에는 모형 식별의 문제에 의하여 무응답 지시변수와 관련된 상호작용 효과 XR과 YR을 동시에 포함하는 모형을 고려할 수 없다.

이 로그선형모형은 다음과 같이 행렬모형으로 표현할 수 있다.

$$\log(\mu) = Z\beta$$

여기서 Z는 계획행렬 (design matrix)이고 β는 로그선형모형의 체계적 성분에 대한 모수 벡터가 된다. 무응답을 포함하는 자료의 분석은 일반적으로 EM 알고리즘을 이용한 추정 방법이 주로 이용된다. 무응답 추정을 위한 EM 알고리즘은 다음과 같다.

2.2. EM 알고리즘

다항분포 가정하에서의 랜덤성분 (2.1)과 체계적 성분의 선형모형 (2.2)을 결합하여 다음과 같은 우도함수를 정의할 수 있다.

$$\begin{aligned} l = &\sum_i \sum_j y_{ij1}(z_{ij1} \cdot \beta) - \sum_i \sum_j y_{ij1} \log(\sum_{ijk} \exp(z_{ijk} \cdot \beta)) \\ &+ \sum_i y_{i+2} \log(\sum_j \exp(z_{ij2} \cdot \beta)) - \sum_i y_{i+2} \log(\sum_{ijk} \exp(z_{ijk} \cdot \beta)). \end{aligned}$$

이제 로그우도함수를 최대화 시키는 β를 계산하기 위해 Dempster 등 (1977)이 제안한 GEM (generalized expectation and maximization) 알고리즘을 이용하였다. GEM 알고리즘의 E-step과 M-step은 다음과 같다.

E-step (expectation step): 먼저 모수벡터 β의 초기치가 주어지고 관찰된 주변합 y_{i+2}가 주어졌을 때 계산될 수 있는 무응답의 대체값 y_{ij2}, $i = 1, \cdots, I$, $j = 1, \cdots, J$와 무응답이 발생하지 않은 관찰치 y_{ij1}을 이용하여 확장된 (augmented) 사후분포함수는 다음과 같이 주어진다.

$$l_{augmented} = \sum_i \sum_j (y_{ij1}) \log(\pi_{ij1}) + \sum_i \sum_j (y_{ij2}) \log(\pi_{ij2}).$$

E-step에서는 확장된 사후분포함수식으로부터 무응답 빈도 y_{ij2}에 대한 기댓값을 계산한다. 반복수행 과정에서 이전 시점에서 계산된 칸 기대확률 π_{ijk}^{old}라 하고 이 확률과 관찰된 주변합 y_{i+2}이 주어졌을 때 $l_{augmented}$에 대한 조건부 기대값은 다음과 같다.

$$E[l_{augmented}] = \sum_i \sum_j (y_{ij1}) \log(\pi_{ij1}) + \sum_i \sum_j (E[y_{ij2}|\pi_{ijk}^{old}]) \log(\pi_{ij2}).$$

무응답 빈도 y_{ij2}는 주변합 y_{i+2}가 주어졌을 때 다항분포를 따름으로 무응답 빈도에 대한 기대값은 다음과 같이 계산되어 E-step을 완성한다.

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

여기서 $m_{ijk}^{old} = N \cdot \pi_{ijk}^{old}$로 계산되고 π_{i+2}와 m_{i+2}는 행 범주에 대한 주변합을 나타낸다.

M-step (maximization step): E-step에서 무응답 빈도에 대한 대체 값이 계산되었기 때문에 확장된 로그우도함수 $l_{augmented}$는 3차원 분할표 자료에 대한 로그우도함수의 형태가 된다. 이에 로그 변환된 우도함수를 최대화시키는 모수에 대한 추정치를 일반적인 로그선형모형에서의 모수 추정방법을 이용하여 계산할 수 있다. 기대빈도를 $E[l_{augmented}]$에 대입한 후 반복적 가중 최소제곱법 (iterative re-weighted least method; Agresti, 2002)을 사용하여 모수에 대한 최대우도추정치를 계산한다.

모수 추정치가 수렴할 때까지 E-step과 M-step을 반복 수행한다. 반복수행단계에서는 계산된 모수의 추정치 가운데 가장 작은 차이가 10^{-10}보다 작아질 때까지 반복수행하였다.

2.3. MWPE (modified within precinct error)

EM알고리즘을 이용하여 계산된 모수 β의 추정치를 대입한 후 로그선형모형 $\log(\mu_{ijk})$를 각 칸의 기대도수 μ_{ijk}로 구할 수 있다.

$$\mu_{ijk} = \exp(\beta_0 + \beta_X^i + \beta_Y^j + \beta_R^k + \beta_{XY}^{ij} + \beta_{XR}^{ik} + \beta_{YR}^{jk}) \tag{2.3}$$

이렇게 구한 μ_{ijk}로 무응답이 있는 3차원 분할표는 무응답이 없는 일반적인 3차원 분할표 자료를 얻을 수 있다. 예로 든 무응답이 있는 3차원 분할표 Table 2.1을 Table 2.2와 같이 무응답이 없는 3차원 분할표로 표현할 수 있다.

Table 2.2 Three-way coningency table after missing imputation

	Response $R=1$		Non-Response $R=2$	
	Candidate A ($Y=1$)	Candidate B($Y=2$)	Candidate A ($Y=1$)	Candidate B ($Y=2$)
$X=1$	y_{111}	y_{121}	y_{112}^*	y_{122}^*
$X=2$	y_{211}	y_{221}	y_{212}^*	y_{222}^*

이처럼 구성된 3차원 분할표를 가지고 후보별 최종 예측치를 구할 수 있다. 각 칸의 추정된 빈도는 식 (2.3)로 계산될 수 있으며 이를 $\hat{y}_{ijk}$라 하고 후보자 A와 후보자 B의 추정된 지지율을 각각 $\hat{P}_1$과 $\hat{P}_2$라 하면 이들은 각각

$$\hat{P}_j = \frac{\sum_i \sum_k \hat{y}_{ijk}}{\sum_i \sum_j \sum_k \hat{y}_{ijk}}, \quad j = 1, 2. \tag{2.4}$$

로 계산될 수 있다.

추정된 모형 결과로부터 실제 결과와 모형간의 차이를 비교하는 방법은 여러 가지가 있다. 본 연구에서는 지지후보에 대한 무응답과 예측오류의 관계가 있다고 생각하여 Bautista 등 (2007)이 제안한 MWPE (modified within precinct error)를 사용하여 예측 지지율과 실제 지지율을 비교하였다.

$$MWPE = \frac{2P_1(1-\alpha)(P_1-1)}{P_1(1-\alpha)-1}, \quad \alpha = \frac{\hat{P}_1/\hat{P}_2}{P_1/P_2}. \tag{2.5}$$

여기서 P_1 과 P_2는 실제 지지율로써 당선자와 지지율이 두 번째로 높은 후보자이며 $\hat{P}_1$과 $\hat{P}_2$는 P_1과 P_2의 추정치로 식(2.4)으로부터 계산된다. α가 1로 가면 MWPE는 0에 가까워지고 MWPE이 0이라는 것은 출구 조사 예측률과 실제 지지율의 차이가 없다고 할 수 있다.

3. 자료분석

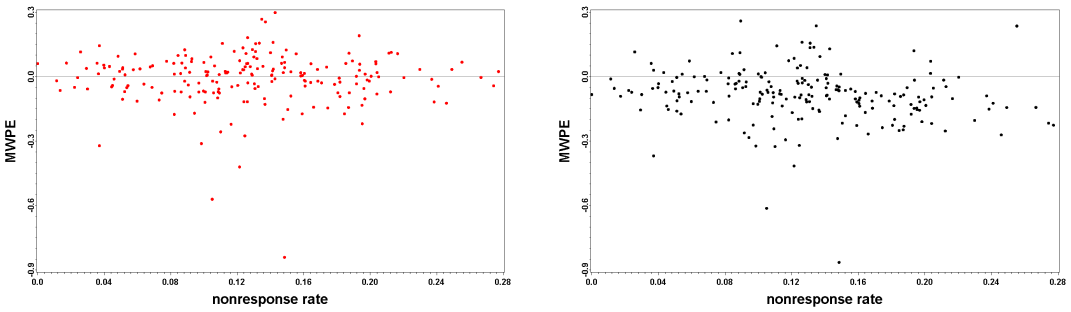
분석에 이용된 자료는 18대 대선 당시 전국 360개의 투표소에서 수행된 출구조사의 자료로써 지역 (광역자치단체), 구 (기초자치단체), 성별, 나이, 지지후보 (새누리당 후보, 민주통합당 후보, 기타, 무응답) 등이 집계되었다. 본 연구에서는 이 자료를 19대 국회의원 선거구에 맞추어 재조정하였다. 19대 국회의원 선거의 선거구는 총 242개이다. 자료의 조정에서 대치되지 않은 일부 선거구를 제외하고 총 204개의 선거구별로 자료를 재조정하여 분석에 이용하고자 하였다.

반응변수 Y는 지지후보로서 새누리당과 민주통합당 그리고 무응답으로 나타냈다. 지지후보의 경우 기타 후보에 대한 항목이 있었으나 모든 선거구에 대하여 그 크기가 매우 미비하였다. 이에 분석의 편의를 위하여 모든 기타후보의 빈도를 제외하고 분석을 수행하였다. 설명변수 X는 연령대를 나타내는 변수로 20대, 30대, 40대, 50대, 60대 이상으로 구분하였다. 지지후보에 대한 응답 여부 R은 $R = 1$일 때 응답이고 $R = 2$일 때는 무응답으로 나타낸다. 이 변수들을 이용하여 무응답 발생 체계에 따라 3가지 모형을 고려할 수 있다.

$$\begin{aligned}
&\text{Model 1} \ : \ \log(\mu_{ijk}) = \beta_0 + \beta_X^i + \beta_Y^j + \beta_R^k + \beta_{XY}^{ij} && (MACR)\\
&\text{Model 2} \ : \ \log(\mu_{ijk}) = \beta_0 + \beta_X^i + \beta_Y^j + \beta_R^k + \beta_{XY}^{ij} + \beta_{XR}^{ik} && (MAR)\\
&\text{Model 3} \ : \ \log(\mu_{ijk}) = \beta_0 + \beta_X^i + \beta_Y^j + \beta_R^k + \beta_{XY}^{ij} + \beta_{YR}^{jk} && (NMAR)
\end{aligned}$$

총 204개의 지역구 자료에 대하여 각각 이 세 가지의 무응답 모형의 모형 적합을 수행하였고 각 결과로부터 후보별 지지율 (2.4)와 적합도 비교 통계량 MWPE (2.5)를 계산하였다. Model 1은 완전임의결측, Model 2는 임의결측으로 무응답 발생 여부가 무응답 여부와 상관이 없으므로 무시할 수 있는 무응답 모형이라 할 수 있다. Model 3은 비임의결측으로 무응답을 가지고 있는 반응변수와 무응답 여부의 상호작용이 포함되었기 때문에 무시할 수 없는 무응답 모형이라 할 수 있다. EM 알고리즘을 이용한 최대우도추정을 수행하여 최종적으로 적합 된 자료를 2명의 지지후보에 따라 정리하여 최종 예측 결과를 구축하였다. Model 1, Model 2의 두 모형 간의 지지후보가 최종 예측 결과는 모두 같게 나왔다.

Model 1에서는 다른 어떤 변수에도 영향을 받지 않아서 무응답의 대체가 연령대에 상관없이 응답자의 후보 지지율에 비례하여 이루어진다. Model 2에서는 무응답을 가지고 있지 않은 자료로부터 영향을 받는데 이 연구에서는 설명변수가 나이 하나이기 때문에 Model 2 역시 연령에 따른 비율로 무응답을 채우게 된다. 결과적으로 지지후보만을 가지고 정리된 상황에서는 같은 지지율을 보이게 된다. 따라서 결과 정리에서는 Model 2에 의한 결과만을 제시하였다. Model 3은 비임의결측에 대한 모형으로 무응답 체계에 대해서는 무시할 수 없는 무응답 모형이라 할 수 있다.



(a) MAR (b) NMAR

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

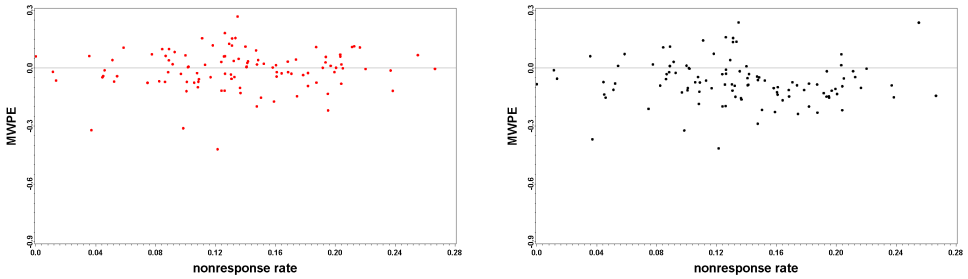
이제 분석 결과를 살펴보자. Figure 3.1은 임의결측과 비임의결측을 이용하여 수행한 MWPE값과 각 204개의 국회의원 선거구별로 무응답률을 계산하여 나타낸 그림이다. 왼편은 임의결측 (MAR)에 의한 분석 결과이고 오른편은 비임의결측 (NMAR)에 의한 분석 결과이다. 각 도표에서 X축은 무응답률을 나타내고 Y축은 MWPE값을 나타낸다. 전반적으로 보았을 때 임의결측 모형의 MWPE값이 비임의결측 모형의 MWPE값 보다 0에 더 많이 모여 있는 것을 볼 수 있다. 그러나 무응답률에 따라 임의결측과 비임의결측의 차이는 특별하게 보이지 않는다. 즉 무응답률이 높다고 하여 모형의 예측력이 떨어질 것이라고는 이야기할 수 없다. 예측력 비교를 위한 통계적 검정을 수행하기 위하여 짝진 표본 t-검정을 수행하였고 그 결과는 다음 Table 3.1에 정리되어 있다. Table 3.1의 Mean과 Std는 데이터의 MWPE에 대한 평균과 표준편차를 나타내고 d는 임의결측 MWPE 값과 비임의결측 MWPE값의 차로 $\bar{d}$와 S_d는 두 값의 차에 대한 평균과 표준 편차를 나타낸다.

Table 3.1 Paired t-test: Whole cases

	N	Mean	Std	$\bar{d}$	S_d	t
MAR	204	-0.0075	0.1243	0.0798	0.06	19.01
NMAR	204	-0.0873	0.1273			

Table 3.1로부터 204개의 국회의원 선거구별 MPWE값의 평균과 표준편차의 값을 보면 204개의 임의결측 평균값은 -0.0075, 표준편차는 0.1243이고 204개의 비임의결측 평균값은 -0.0873, 표준편차는 0.1273이다. MPWE값은 0에 가까울수록 높은 예측률을 가지고 있다고 할 수 있는데 임의결측의 평균이 비임의결측 평균보다 더 0에 가까우므로 임의결측이 더 좋은 결과를 가지고 있다고 할 수 있다. 이때 임의결측과 비임의결측의 짝진 표본 t-검정을 시행한 결과 검정통계량은 19.01이고 p-값은 $p < .001$로 두 평균의 차이는 통계적으로도 유의하다 .

Model 3의 무시할 수 없는 무응답 모형에서 EM 알고리즘 이용한 최대우도 추정에서 변방 값 문제가 발생할 수 있다. 변방 값 문제란 분할표에서 추정된 무응답 빈도에 대한 확률이 특정 칸에서 0의 값을 가지는 현상을 말한다. 변방 값 문제가 발생하게 되면 최대우도추정치가 유일한 해를 가지지 않게 되고 그 결과가 불안정해질 수 있다 (Baker와 Laird, 1988). 변방 값 문제를 해결하기 위해서 베이지안 접근법에 기반을 둔 방법들이 제안되어 왔다(Park과 Brown, 1994; Choi 등, 2009; Park과 Choi, 2010; Yoon과 Choi, 2012). 따라서 변방 값 문제가 발생한 경우 무응답 체계에 대한 비교를 수행하게 되면 무시할 수 없는 무응답 모형에 대한 결과가 왜곡될 가능성이 존재한다. 이러한 모형의 왜곡되는 결과를 배제하기 위하여 총 204개의 모형 결과 가운데 변방 값 문제가 발생하지 않은 경우만을 다시 정리하여 보았다. Figure 3.2는 변방 값 문제가 발생하지 않은 총 107개의 결과만을 가지고 Figure 3.1과 같은 도표를 작성한 것이다.



(a) MAR (b) NMAR

Figure 3.2 MWPE and nonresponsee rates for MAR and NMAR: Cases without boundary solution only

Table 3.2 Paired t-test: Cases without boundary solution only

	N	Mean	Std	$\bar{d}$	s_d	t
MAR	107	-0.0078	0.1028	0.0648	0.0608	11.02
NMAR	107	-0.0726	0.1116			

Table 3.2로부터 변방 값 문제가 발생하지 않은 107개 국회의원 선거구별 MPWE값의 평균과 표준편차의 값을 보면 107개의 임의결측 평균값은 -0.0078, 표준편차는 0.1028이고 107개의 비임의결측 평균값은 -0.0726, 표준편차는 0.1116이다. 임의결측이 비임의결측보다 MWPE 평균값이 0에 가까우므로 더 높은 예측률을 가지고 있다. 이때 임의결측과 비임의결측의 짝진 표본 t-검정을 실시한 결과 검정 통계량은 11.02고 p-값은 $p < .001$로 두 평균의 차이는 통계적으로도 유의하다. 전체 204개의 국회의원 선거부별 결과와 같은 결과이다. 하지만 전체 204개의 자료와 변방 값 문제가 발생하지 않은 107개 자료의 임의결측과 비임의결측의 평균차의 값을 비교하면 변방 값 문제가 발생하지 않은 107개 자료의 평균차가 더 작다. 이는 Table 3.1과 Table 3.2의 임의결측과 비임의결측의 MWPE평균을 봤을 때 변방 값 문제가 발생하지 않은 비임의결측의 MWPE평균이 0에 더 가까워진 것을 확인할 수 있다.

모형 적합에 대한 조금 더 상세한 비교를 위하여 MWPE를 기준으로 하여 그 결과를 임의결측의 예측력이 더 우세한 경우와 비임의결측이 더 우세한 경우로 나누어 보았다. 이와 동시에 이를 다시 변방 값 문제가 발생한 경우와 발생하지 않은 경우로 구분하여 정리하여 보았다. 다음 Table 3.3의 결과를 살펴보자. 두 번째 열인 boundary는 변방 값 문제의 발생 여부로 "1"인 경우 변방 값 문제가 발생한 경우이고 "2"인 경우는 변방 값 문제가 발생하지 않은 경우이다. 표의 가장 위쪽에 차지한 "all"은 지역 구분을 하지 않은 전국의 204개의 선거구별 정리한 결과이다. 97개의 선거구결과에서 변방 값 문제가 발생하였으며 107개의 선거구에서 변방 값 문제가 발생하지 않았다. 괄호 안의 숫자들은 행 백분율을 나타낸다. MWPE에 의한 예측 정확도를 살펴보면 전국적으로 보았을 때 변방 값 문제의 발생 여부에 상관없이 model 2를 이용한 무시할 수 있는 무응답 가정하에 모형이 무시할 수 없는 무응답 가정하에 모형보다 높은 예측 정확도를 보였다. 즉 실제 결과에 더욱 근접한 결과를 보여 주고 있다고 할 수 있다. 이 결과에 기반을 두어 보았을 때 우리나라 선거에서는 전반적으로 무시할 수 있는 무응답 가정하에서 무응답 모형을 추정하는 것이 더 적합하다 할 수 있다.

Table 3.3 Comparison result for MAR and NMAR

	boundary	MWPE		Total
		MAR (rate)	NMAR (rate)	
all	1	65 (67.01)	32 (32.99)	97
	2	71 (66.36)	36 (33.64)	107
Seoul Metrolitan Area	1	25 (65.79)	13 (34.21)	38
(seoul, Gyeonggi, Incheon)	2	35 (67.31)	17 (32.69)	52
Yeongnam area	1	20 (66.67)	10 (33.33)	30
(Daegu, Gyeongbuk, Gyeongnam, Busan, Ulsan)	2	13 (56.52)	10 (43.48)	23
Honam area	1	11 (68.75)	5 (31.25)	16
(Gwangju, Jeonbuk, Jeonnam, Jeju)	2	7 (58.85)	6 (46.15)	13
Chungcheong	1	6 (60)	4 (40)	10
(Daejeon, Chungbuk, Chungnam)	2	12 (85.71)	2 (14.29)	14
Gangwon	1	3 (100)	0 (00.00)	3
	2	4 (80.00)	1 (20.00)	5

다음으로 전체 결과를 지역으로 구분하여 살펴보았다. 전체 지역을 서울, 경기, 인천을 포함한 수도권 지역, 충청지역, 영남지역 (대구, 경북, 경남, 울산, 부산), 호남지역 (광주, 전북, 전남, 제주) 그리고 강원지역으로 구분하여 결과를 비교하였다.

전체지역의 결과와 지역을 나눈 결과에서는 변방 값 문제와 상관없이 임의결측이 비임의결측보다 더 좋은 예측 정확도를 가지고 있다. 그러나 영남과 호남의 경우 그 결과가 전체 결과의 흐름과 비교하였을 때 조금 다른 양상을 보이고 있다. 변방 값 문제가 발생한 경우 영남과 호남지역에서 무시할 수 있는 무응답 모형 가정에서 더 정확한 결과를 보인 비율은 66.67%와 68.75%로 전체 지역에 대한 결과인 67.01%와 큰 차이가 없다. 그러나 변방 값 문제가 발생하지 않은 경우 영남의 경우 무시할 수 없는 무응답 모형에서 예측 정확도가 높은 경우가 43.48%이고 호남의 경우 46.15%로 전체 지역에서의 비율인 33.64% 보다 10%이상 높은 것을 볼 수 있다.

이들 두 지역은 지역색이 매우 강한 지역으로 선거 결과 특정 정당의 후보로 지지율이 밀집되는 현상을 지속적으로 보여왔다. 이런 지역적 특수성 하에서 상대적으로 열세인 후보를 지지하고 있을 때 본인이 지지하고 있는 후보를 선뜻 밝히는 것은 어려운 일이 될 수 있다. 이러한 상황에서는 무시할 수 없는 무응답의 가정이 더 적절하다 할 수 있으며 실제 결과에서도 무시할 수 없는 무응답 가정 하에서의 추정결과의 정확도가 상대적으로 높이 나왔음을 볼 수 있다. 강원 지역의 경우 그 빈도가 상대적으로 작으므로 논의에서 제외 하였다.

Table 3.4 Comparion between actual support rates and forecast support rates of MAR and NMAR model for Saenuri party candidate (unit: %)

Constituency	real	MAR	NMAR	Constituency	real	MAR	NMAR
Busan Yeongdogu	59.0	62.5	63.0	Gyeongnam Jinjugap	68.0	73.8	77.8
Busan Namgu	60.7	68.4	71.6	Gyeongnam Tongyeong	70.1	81.1	77.8
Busan Bukgueul	57.1	56.8	57.3	Gyeongnam Sacheon etc.	68.1	74.0	73.8
Busan Sahagu	58.5	59.4	59.0	Gyeongnam Gimheaeul	52.4	54.4	59.2
Busan Yeonjegu	60.4	59.3	63.6	Gyeongnam Uiryeong etc.	73.7	77.1	78.2
Busan Suyeonggu	62.0	61.8	67.1	Gwangju Seogu	8.1	5.1	12.3
Daegu Namgu	81.5	78.3	78.4	Gwangju Bukgueul	7.4	7.2	7.4
Daegu Donggu	80.1	76.8	68.3	Jeonbuk Wansangugap	12.3	8.1	13.5
Daegu Bukgu	79.7	83.3	82.5	Jeonbuk Deokjingu	11.8	8.6	13.4
Daegu Suseonggu	78.9	77.9	77.3	Jeonbuk Iksangap	13.6	10.0	12.7
Ulsan Namgu	61.7	63.1	65.3	Jeonbuk Jeogeup	11.7	12.3	19.3
Ulsan Bukgu	54.0	69.6	70.2	Jeonbuk Wanjugun	13.2	11.2	12.7
Gyeongbuk Pohangbukgu	79.7	83.5	83.4	Jeonbuk Jinan etc.	16.4	13.4	20.6
Gyeongbuk Gyeongju	79.4	81.5	78.8	Jeonnam Naju*Hwacheon	8.9	6.8	14.5
Gyeongbuk Gimcheon	83.9	87.1	86.6	Jeonnam Goheung	11.2	7.9	10.5
Gyeongbuk Gumieul	80.6	74.7	73.8	Jeju Jejugap	50.0	51.0	55.3
Gyeongbuk Uisunggun	86.9	88.2	89.2	Jeju Jejueul	50.0	50.6	54.4
Gyeongnam Jinhaegu	64.6	61.7	65.4	Jeju Seogwipo	52.8	56.4	58.9

마지막으로 선거구별 예측 정확도를 비교해 보고자 하였다. 분석에 이용된 전체 204개 선거구 가운데 무시할 수 없는 무응답 모형 가정하에서 변방값 문제가 발생하지 않은 107개 선거구 결과 가운데 다시 주요 관심 지역인 영남과 호남지역 36개 선거구의 실제 결과와 모형에 의한 예측결과를 직접 비교하여 보았다. Table 3.4는 그 결과를 정리한 것으로 선거 당시 새누리당 후보의 실제 실제 득표율 (real), 무시할 수 있는 무응답 가정 모형 (MAR), 무시할 수 없는 무응답 가정 모형 (NMAR)에 의한 예측결과를 각각 표시하여 비교해 보았다. 무시할 수 있는 무응답 가정하일 때 20개의 선거구 추정결과가 정확도가 높았고 무시할 수 없는 가정하일 때 16개의 선거구 추정결과가 정확도가 높았다. 무시할 수 있는 무응답 가정하일 때 20개의 선거구 중 부산 수영구의 추정결과와 실제 결과의 차이가 0.17%이고 무시할 수 없는 무응답 가정하일 때 16개의 선거구 중 광주 북구을의 추정결과와 실제 결과의 차이가 0.013%로 실제 결과와 거의 비슷하다고 할 수 있다.

4. 결론

본 연구에서는 무응답이 발생한 경우 무응답 대체를 포함한 추정 방법에서 모형에 기반을 둔 방법을 이용하고자 하였고 이때 필요한 무응답 모형의 체계에 대한 가정의 정확성 여부를 평가하고자 하였다. 즉 무응답 체계에 대한 두 가지 가정인 무시할 수 있는 무응답 가정과 무시할 수 없는 무응답 가정하에서 어떠한 가정이 보다 적절한가에 대한 평가를 수행하기 위하여 모형 선택 방법에 대한 기준을 이용하지 않고 실제 자료를 이용한 분석 결과를 바탕으로 하여 더 정확한 예측력을 비교함으로써 무응답 체계에 대한 가정을 점검해 보고자 하였다. 이를 위하여 전국 204개 지역의 출구조사 결과를 가지고 무응답 체계에 따른 모형 적합을 수행하였으며 모형 정확도의 평가를 위하여 MWPE 통계량을 이용하였다. 그 결과 무시할 수 있는 무응답 모형을 이용한 적합 결과가 무시할 수 없는 무응답 모형에 대한 적합 결과보다 더 정확한 예측력을 보였다. 선거와 같이 민감한 사항에 대한 사전 조사에서 발생한 무응답이 무시할 수 없는 무응답 일 것이라는 일반적인 가정과는 다른 결과를 보이고 있다. 그러나 선거 결과 지역색이 매우 강한 영남과 호남 지역의 경우 무시할 수 있는 무응답 모형 가정 하에서의 정확도가 상대적으로 높은 것을 보였다. 이는 지지하는 후보가 해당 지역에서 상대적으로 열세인 후보일 때 그 지지후보를 밝히지 않을 것이라는 가정에 보다 적합한 결과라 할 수 있다.

본 연구에서 이용된 자료는 기본적으로 국회의원 선거구에 따라 정리된 결과이다. 그러나 같은 지역 내에서도 특정 후보에 대한 지지 여부가 다르게 분포되는 경우가 빈번하다고 볼 수 있다. 이에 따라 보다 세분화된 지역구분에 따라 분석을 수행하여야 할 필요가 있으며 앞으로 연구에서는 보다 세분화된 지역의 자료를 이용하여 상대적으로 세밀한 지역 구분에 따라, 예를 들면 서울 강남지역, 강북지역 등과 같은 구분에 따라 분석을 수행할 필요가 있다.

본 연구가 가지는 또 다른 한계점은 변방 값 문제에 대한 해결 여부이다. 본 연구에서 수행된 결과에서도 거의 50%에 해당하는 97개의 결과에서 변 방값 문제가 발생하였다. 그러나 본 연구의 주목적이 변방 값 문제에 대한 해결이 아니므로 이에 대한 논의를 수행하지 않았다. 앞으로 연구를 통하여 변방 값 문제에 대한 해결을 한 후에 무응답 체계에 대한 비교를 함께 수행해 볼 수 있을 것이다.

References

Agresti, A. (2002). *Categorical data analysis,* second edition, John Wiley & Sons Inc., New Jersey.

Baek, J. E., Kang, W. C., Lee, Y. J. and Park, B. J. (2002). An approach to survey data with nonresponse: evaluation of KEPEC data with BMI. *Journal of Preventive Medicine and Public Health*, **35**, 136-140.

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.

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

Chambers, R. L. and Welsh, A. H. (1993). Log-linear models for survey data with non-ignorable nonresponse. *Journal of Royal Statistical Society B*, **55**, 157-170.

Cho, Y. S., Chun, Y. M. and Hwang, D. Y. (2008). An imputation for nonresponses in the survey on the rural living indicators. *The Korean Journal of Applied Statistics*, **21**, 95-107.

Choi, B., Choi, J. W. and Park, Y. S. (2009). Bayesian methods for an incomplete two-way contingency table with application to the Ohio (Buckeye state polls). *Survey Methodology*, **35**, 37-51.

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

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

Choi, B., Park, Y. S. and Lee, D. H. (2007). Election forecasting using pre-election survey data with nonignorable nonresponse. *Journal of the Korean Data Analysis Society*, **9**, 2321-2333.

Crespi, I. (1988). *Pre-election polling: Sources of accuracy and error*, Russel Sage, New York.
Dempster, A. P., Laird, N. M. and Rubin, D. M. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, **4**, 1-38.
Fay, R. E. (1986). Causal models for patterns of nonresponse. *Journal of the American Statistical Association*, **81**, 354-365.
Hong, N. R. and Huh, M. H. (2001). A post-examination of forecasting survey for the 16th general election. *The Korean Association for Survey Research*, **2**, 1-35.
Hyun, K. B. (2005). A study on the election poll and its accuracy in th 17th general election. *Korea Regional Communication Research Association*, **5**, 301-336.
Ibrahim, J. G., Zhu, H. and Tang, N. (2008). Model selection criteria for missing-data problems using the EM algorithm. *Journal of the American Statistical Association*, **103**, 1648-1658.
Kim, K. S. (2000). Imputation methods for nonresponse and their effect. *The Korean Association for Survey Research*, **1**, 1-14.
Kim, Y. W. and Choi, Y. J. (2011). Systematic forecasting bias of exit poll: Analysis of exit poll for 2010 local elections. *The Korean Association for Survey Research*, **12**, 25-48.
Kim, Y. W. and Kim, J. H. (2007). An overview of exit polls for the 2006 local elections. *The Korean Association for Survey Research*, **8**, 55-79.
Kim, Y. W. and Kwak, E. S. (2010). A total survey error analysis of the exit polling for general election 2008 in korea. *The Korean Association for Survey Research*, **11**, 33-55.
Lee, H. J. and Kang, S. B. (2012). Handling the nonresponse in sample survey. *Journal of the Korean Data & Information Science Society*, **23**, 1183-1194.
Lee, J. H., Kim, j. and Lee, K. J. (2006). Missing imputation methods using the spatial variable in sample survey. *The Korean Journal of Applied Statistics*, **19**, 57-67.
Little, J. A. and Rubin, D. B. (2002). *Statistical analysis with missing data*, second edition, Wiley, 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.
Park, T. S. and Lee, S. Y. (1998). Analysis of categorical data with nonresponses. *The Korean Journal of Applied Statistics*, **11**, 83-95.
Park, Y. S. and Choi, B. (2010). Bayesian analysis for incomplete multi-way contingency tables with nonignorable nonresponse. *Journal of Applied Statistics*, **37**, 1439-1453.
Rhee, J. W. (2004). Problems of the election forecasting in the 2004 korean general election. *Journal of Communication Research*, **41**, 110-135.
Ryu, J. B. (2000). A plan of improving the reliability of the electon forecasting survey - A case of the 16th general election. *The Korean Association for Survey Research*, **1**, 15-34.
Ryu, J. B. (2003). A history and th improvable direction of exit poll. *The Korean Association for Survey Research*, **4**, 31-48.
Yoon, Y. H. and Choi, B. (2012). Model selection method for categorical data with non-response. *Journal of the Korean Data & Information Science Society*, **23**, 627-641.

Journal of the Korean Data & Information Science Society
2014, **25**(1), 53–64

http://dx.doi.org/10.7465/jkdi.2014.25.1.53
한국데이터정보과학회지

A comparison study for accuracy of exit poll based on nonresponse model[†]

Jeongae Kwak[1] · Boseung Choi[2]

[1]Department of Statistics, Daegu University
[2]Department of Statistics and Computer Science, Daegu University

Received 27 November 2013, revised 17 December 2013, accepted 23 December 2013

Abstract

One of the major problems to forecast election, especially based on survey, is nonresponse. We may have different forecasting results depend on method of imputation. Handling nonresponse is more important in a survey about sensitive subject, such as presidential election. In this research, we consider a model based method of nonresponse imputation. A model based imputation method should be constructed based on assumption of nonresponse mechanism and may produce different results according to the nonresponse mechanism. An assumption of the nonresponse mechanism is very important precondition to forecast the accurate results. However, there is no exact way to verify assumption of the nonresponse mechanism. In this paper, we compared the accuracy of prediction and assumption of nonresponse mechanism based on the result of presidential election exit poll. We consider maximum likelihood estimation method based on EM algorithm to handle assumption of the model of nonresponse. We also consider modified within precinct error which Bautista (2007) proposed to compare the predict result.

Keywords: Election prediction, exit poll, nonresponse mechanism.

[†] This research was supported by Daegu University Research Grant in 2012 (No.20120481).
[1] Graduate student, Department of Statistics, Daegu University, Gyeongbuk 712-714, Republic of Korea.
[2] Corresponding author: Assistant professor, Department of Statistics and Computer Science, Daegu University, Gyeongbuk 712-714, Republic of Korea. E-mail: bchoi@daegu.ac.kr