The Korean Journal of Applied Statistics (2019) DOI: https://doi.org/10.5351/KJAS.2019.32.1.029
32(1), 29–40

Sensitivity analysis of missing mechanisms for the 19th Korean presidential election poll survey

Seongyong Kim[a,1] · Dongho Kwak[a]

[a]Division of Big Data and Management Engineering, Hoseo University

(Received September 14, 2018; Revised December 3, 2018; Accepted December 3, 2018)

Abstract

Categorical data with non-responses are frequently observed in election poll surveys, and can be represented by incomplete contingency tables. To estimate supporting rates of candidates, the identification of the missing mechanism should be pre-determined because the estimates of non-responses can be changed depending on the assumed missing mechanism. However, it has been shown that it is not possible to identify the missing mechanism when using observed data. To overcome this problem, sensitivity analysis has been suggested. The previously proposed sensitivity analysis can be applicable only to two-way incomplete contingency tables with binary variables. The previous sensitivity analysis is inappropriate to use since more than two of the factors such as region, gender, and age are usually considered in election poll surveys. In this paper, sensitivity analysis suitable to an multi-dimensional incomplete contingency table is devised, and also applied to the 19th Korean presidential election poll survey data. As a result, the intervals of estimates from the sensitivity analysis include actual results as well as estimates from various missing mechanisms. In addition, the properties of the missing mechanism that produce estimates nearest to actual election results are investigated.

Keywords: imputation, sensitivity analysis, incomplete contingency table, missing mechanism

1. 서론

불완전 분할표는 무응답이 존재하는 범주형 자료를 분할표로 표현한 것으로, 완전히 관측된 분할표와 일부만 관측된 분할표로 이루어져 있다. 이와 같이 무응답이 포함된 불완전 분할표의 대표적인 예로는 선거여론조사를 들 수 있다. 선거여론조사에서는 표본 설계를 통해 지역, 성, 연령대등에 따라 표본을 할당하고, 전화면접 등을 통해 지지후보자에 대한 질문 등에 대한 조사를 진행한다. 그러나 많은 경우 무응답이 발생하며, 무응답으로 인해 조사결과 및 개표 결과 간에 차이가 발생할 수 있다. 따라서 선거여론조사에서는 무응답 추정의 중요성이 점점 커지고 있는데, 무응답의 추정은 무응답 메카니즘(missing mechanism)에 따라 달라지게 된다.

무응답 메카니즘은 크게 missing completely at random (MCAR), missing at random (MAR) 및 missing not at random (MNAR)이 있는데, MCAR은 무응답의 발생이 반응변수 및 설명변수에 의존하지

This research was supported by Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Education (NRF-2016R1D1A3B03930392).
[1]Corresponding author: Division of Big Data and Management Engineering, Hoseo University, 20, Hoseo-ro 79, Asan 31499, Korea. E-mail: yaba96@hoseo.edu

않는 것을 의미하며, MAR은 무응답의 발생이 관측된 설명변수에 의존하는 경우를 말한다. 마지막으로 MNAR은 무응답의 발생이 관측되지 않은 반응변수에 의존하는 경우이다 (Rittle과 Rubin, 2002). 이와 같은 무응답 메카니즘을 반영하여 무응답을 추정하기 위해 selection 모형, pattern mixture 모형, 일반화선형모형(generalized linear model; GLM) 등의 방법이 널리 이용되고 있다 (Fay, 1986; Baker와 Laird, 1988; Baker 등, 1992; Little, 1993; Little과 Rubin, 2002; Foster와 Smith, 1998). 그러나 무응답 메카니즘이 잘못 선택되어 무응답을 추정하는 경우 추정 결과가 달라질 뿐만 아니라 편향(bias)이 발생하며, 분산 역시 크게 추정되는 것으로 알려져 있다 (Clarke, 2002; Choi 등, 2009; Poleto 등, 2011). 따라서 관측된 자료의 무응답이 어떠한 무응답 메카니즘을 따르는지를 파악하는 것이 불완전 분할표 분석의 선결 과제이다.

기존에는 무응답 메카니즘의 판별이 모형의 적합도에 기반하여 이루어졌으나, Molenbeghs 등 (2008)은 관측된 자료만을 가지고는 무응답 메카니즘의 판별이 불가능함을 이론적, 실증적으로 보였다. 이에 따라 특정 무응답 메카니즘에 기반한 무응답 추정 결과를 제시하기보다, 다양한 모형의 추정 결과를 하나의 구간으로 제시하는 민감도 분석이 제안되었다 (Molenberghs 등, 2001; Baker 등, 2003; Vansteelandt 등, 2006; Kim과 Kim, 2018; Kim, 2016).

Molenberghs 등 (2001)은 범주의 수가 2개인 이원 불완전 분할표(two-way incomplete contingency table)에서의 민감도 분석을 제안하였으며, 이를 슬로베니아 독립에 대한 여론조사 자료에 적용하였다. 이들은 MAR 및 MNAR을 포괄하는 초과모수 모형(overparameterized model)을 설정하고, 이 가운데 MNAR을 나타내는 모수를 민감도 모수로 지정하여, 해당 모수의 변화에 따른 관심모수(예를 들어 독립 지지율 또는 후보별 지지율)의 변화를 무지의 구간(region of ignorance)과 확실성의 구간(region of uncertainty)이라는 두 개의 구간으로 표현할 것을 제안하였다. 무지의 구간은 민감도 모수의 변화에 따른 관심 모수의 예측치 변화를 구간으로 표현한 것이며, 불확실성의 구간은 예측치 뿐만 아니라 신뢰구간의 변화까지 반영한 구간이다. Kim (2016)은 Molenberghs 등 (2001)의 방법을 우리나라 18대 대선 자료에 적용하였다.

그러나 Molenberghs 등 (2001) 및 Kim (2016)은 범주의 수가 2개인 이원 분할표만을 대상으로 민감도 분석을 수행하였다. 우리나라의 선거여론조사의 경우, 지역 뿐만 아니라, 성, 연령 역시 지지후보자에 영향을 주는 것으로 알려져 있으며, 표본 설계 역시 이러한 성향을 반영하여 이루어진다. 따라서 본 논문에서는 기존 Molenberghs 등 (2001)의 방법론을 확장하여, 삼원 또는 사원 불완전 분할표에 적용할 수 있는 민감도 분석을 제안하고자 한다. 또한 해당 방법론을 우리나라의 19대 대선 여론조사 자료에 적용하여 민감도 분석을 실시하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 우리나라 19대 대선 여론조사 자료를 소개하고, 3장에서는 민감도 분석에 이용될 무응답 모형을 소개하도록 한다. 4장에서는 Molenberghs 등 (2001)의 민감도 분석을 확장하고, 이를 실제 자료에 적용해 보도록 한다. 5장에서는 결론을 제시한다.

2. 19대 대선 선거여론조사 자료

본 연구에서는 전국 단위에서 지역, 성, 연령을 고려한 19대 대선 여론조사 자료의 지지후보에 대한 무응답을 추정하고, 이를 통해 후보자별 지지율을 예측하고자 한다. 지역의 경우 표본 설계와 동일하게 '서울', '인천/경기', '대전/충남/충북', '광주/전남/전북', '대구/경북', '부산/울산/경남' 및 '강원/제주'의 7개 권역으로 분류하였으며, 성별은 '남성' 및 '여성', 연령대 역시 표본 설계와 동일하게 '19세–29세', '30대', '40대', '50대' 및 '60대 이상'으로 분류하였다. 후보자의 경우 '문재인', '홍준표', '안철수', '유승민' 및 '기타'의 5 후보로 분류하였다.

Table 2.1. Survey result for the 19th presidential election in Seoul and Incheon/Gyeonggi-do

Region	Gender	Ages	$R=1$					$R=0$
			Moon	Hong	Ahn	Yu	Others	Missing
Seoul	Male	19–29	10.00	0.63	3.13	1.88	2.50	1.88
		30–39	11.74	2.47	1.24	1.24	1.24	3.09
		40–49	12.19	1.36	3.39	0.00	2.71	1.36
		50–59	8.28	4.83	3.45	0.00	0.69	2.76
		60+	0.73	9.53	3.67	0.00	0.00	8.07
	Female	19–29	1.33	0.00	0.00	3.33	6.67	8.67
		30–39	10.22	2.27	1.14	1.70	1.70	3.97
		40–49	12.29	0.00	4.53	1.29	0.00	3.88
		50–59	5.91	3.28	5.91	0.00	0.00	5.91
		60+	6.40	9.24	2.13	0.00	0.00	9.24
Incheon/ Gyeonggi	Male	19–29	12.60	1.32	5.04	2.63	3.29	6.13
		30–39	13.02	1.08	4.33	2.17	3.26	8.14
		40–49	20.53	4.83	4.77	1.36	1.38	4.13
		50–59	16.30	8.46	2.84	0.71	0.71	4.97
		60+	2.48	16.99	8.04	0.00	0.00	2.48
	Female	19–29	13.94	1.18	2.48	0.00	6.02	4.38
		30–39	20.29	2.14	1.43	0.00	1.43	5.71
		40–49	24.22	1.46	4.99	0.73	2.66	1.93
		50–59	13.99	3.30	6.55	1.63	1.63	4.92
		60+	6.68	12.62	8.14	0.78	2.93	5.85

Table 2.1은 '서울' 및 '인천/경기지역'의 성별 연령대별 지지후보 조사 결과로, 각 칸 빈도(cell count)는 지역, 성, 연령대별 표본 가중치를 반영한 값이다. 아래 테이블에서 알 수 있듯이, 표본 설계 시 고려된 지역, 성 및 연령대는 완전히 관측되는데 반해 지지후보의 경우 무응답이 존재한다. 전체 1100명의 조사 대상 가운데 무응답이 192명 가량으로, 무응답 비율이 17.50%를 차지하고 있다. 문재인 후보의 경우 38.5%, 홍준표 후보는 16.8%, 안철수 후보의 경우 15.7%, 유승민 후보의 경우 3.8%의 지지를 받는 것으로 나타났다.

무응답을 제외하고 전국 단위에서 지지후보에 대한 비율을 구해보면, 문재인 후보자의 경우 46.68%, 홍준표 후보자는 20.38%, 안철수 후보자는 19.08%, 유승민 후보자의 경우 4.64%, 기타 후보자는 9.22%의 지지를 받는 것으로 조사되었으나, 실제 개표 결과의 경우 문재인 후보자는 41.09%, 홍준표 후보자는 24.04%, 안철수 후보자는 21.42%, 유승민 후보자는 6.76%의 지지를 받은 것으로 나타난다. 이러한 차이는 무응답 층에 기인한 것으로 여겨지며, 선거여론조사에서 무응답 추정의 필요성을 알 수 있다.

3. 다차원 불완전 분할표에서의 무응답 모형

본 논문에서는 무응답 추정을 위한 모형으로 Molenberghs 등 (2001)이 제안하고 Kim (2016)이 이용한 selection 모형을 확장하고자 한다. Table 2.1에서와 같이 세 개의 설명변수인 X_1, X_2와 X_3를 고려하며, 반응변수는 Y라 하자. 본 논문의 선거여론조사 자료의 경우 지역을 X_1, 성별을 X_2, 연령을 X_3라 하고 지지후보자를 Y로 할 수 있다. X_1, X_2, X_3는 범주의 수가 각각 I, J, K인 범주형 변수로 완전히 관측되며, Y는 범주의 수가 L인 범주형 변수로 무응답이 존재한다. R은 Y의 관측여부를 나타내는

지시 변수로, Y가 완전히 관측되는 경우 $R = 1$이라 하고, 관측되지 않은 경우 $R = 0$라 하자. 그러면 $R = 0$일 때의 Y가 관측되었다는 가정하에서의 가상의 분할표는 $I \times J \times K \times L \times 2$ 차원의 분할표로 나타나지며, 이 때 $X_1 = i$, $X_2 = j$, $X_3 = k$, $Y = \ell$, $R = r$에서의 칸 확률을 $\pi_{ijk\ell r}$이라 하고, 빈도를 $y_{ijk\ell r}$이라 하자. 그러나 $R = 0$인 경우 지지후보자는 관측되지 않고 Y의 주변 합만 관측되므로, 실제로 관측되는 불완전 분할표는 Table 2.1과 같은 형태이다. Y가 무응답일 때 $X_1 = i$, $X_2 = j$, $X_3 = k$에서 관측된 주변합을 y_{ijk+0}이라 하고, 해당 칸 확률을 π_{ijk+0}라 하자. 여기서 아래 첨자 $+$는 해당 첨자의 모든 값을 합한 것을 나타낸다.

기존의 연구에서는 2차원의 불완전 분할표만 고려한 반면, 본 논문에서는 4차원의 불완전 분할표를 고려하기 때문에 무응답 추정 모형의 모수가 많아지는 한계점이 존재한다. 이를 간략화하기 위해 본 논문에서는 각 지역에서의 성별, 연령별, 연령별 후보자 지지율이 서로 독립임을 가정한다. 즉,

$$\pi_{ijk\ell r} = \Pr[X_1 = i] \times \Pr[X_2 = j, X_3 = k, Y = \ell, R = r \mid X_1 = i] = \pi_i \times \pi_{jk\ell r|i} \tag{3.1}$$

이며, 서로 다른 지역에서의 지지확률인 $\pi_{ijk\ell r}$와 $\pi_{ijk\ell r'}$은 서로 독립을 가정한다. 각 지역별 투표 성향이 현저히 다른 우리나라의 상황을 고려할 때 이는 타당한 가정이라 판단되며, 모형을 간략화할 수 있는 장점이 있다. 각 지역의 비중을 나타내는 π_i는 여론조사 설계에서 이용되는 각 지역별 배분 비중으로 추정하도록 하며, 각 지역 내에서의 성별, 연령별 후보자 지지율은 Molenberghs 등 (2001)과 동일하게 아래와 같은 selection 모형으로 표현하도록 한다.

$$\begin{aligned}\pi_{jk\ell r|i} &= \Pr[X_2 = j, X_3 = k, Y = \ell \mid X_1 = i] \times \Pr[R = r \mid X_1 = i, X_2 = j, X_3 = k, Y = \ell] \\ &= p_{jk\ell|i}\phi_{r|ijk\ell}.\end{aligned} \tag{3.2}$$

식 (3.2)에서 $p_{jk\ell|i}$는 지역 i에서의 성별 j, 연령대 k 및 지지후보가 ℓ인 경우의 주변확률(marginal probability)이며, $\phi_{r|ijk\ell}$은 지역 i에서 성별 j, 연령대 k일 때 지지후보 ℓ에 대한 무응답 확률(non-response probability)이다. 따라서 지역 i에서 후보자 ℓ에 대한 지지율은 $p_{jk\ell|i}$를 성 및 연령에 대해 전부 합함으로써 구할 수 있으며, 전국 지지율은 각 지역별 비중을 지역 i에서 후보자 ℓ에 대한 지지율에 곱한 후 모든 지역에 대해 합함으로써 구할 수 있다. 이에 대해서는 추후 상세히 설명하기로 한다.

$p_{jk\ell|i}$는 $X_1 = i$에서 모든 X_2, X_3 및 Y에 대해 합이 1이 되어야 하므로, 로짓 변환을 이용하여 다음과 같이 모형화하며,

$$p_{jk\ell|i} = \begin{cases} \dfrac{1}{1 + \sum_{j,k,\ell} \exp\left(\gamma_{jk\ell|i}\right)}, & \text{for } j = J,\ k = K, \text{ and } \ell = L, \\ \dfrac{\exp\left(\gamma_{jk\ell|i}\right)}{1 + \sum_{j,k,\ell} \exp\left(\gamma_{jk\ell|i}\right)}, & \text{otherwise.} \end{cases} \tag{3.3}$$

무응답 확률인 $\phi_{r|ijk\ell}$ 역시 로짓 변환을 이용하여 아래와 같이 모형화한다 (Molenberghs 등, 2001).

$$\phi_{r|ijk\ell} = \frac{\exp\left[\beta_{jk\ell|i}(1 - r)\right]}{1 + \exp\left[\beta_{jk\ell|i}\right]}. \tag{3.4}$$

식 (3.4)에서 $\beta_{jk\ell|i}$를 어떻게 모형화하였는가에 따라 반영되는 무응답 메카니즘이 달라지게 되며, 추정결과 역시 달라진다. Table 3.1은 지역 i에서 $\beta_{jk\ell|i}$에 대한 각각의 모형 소개 및 모수의 수, 그리고 동일한 의미를 가지는 로그선형모형을 보이고 있다.

Table 3.1에서 M1의 경우 X_2, X_3 및 Y가 Y의 무응답 발생에 영향을 미치므로 MAR 메카니즘과 MNAR 메카니즘이 혼용된 형태이며, M2의 경우 X_2와 X_3 및 이들의 교호작용이 무응답에 영향을 미

Table 3.1. Nonresponse models

Symbol	Model	Number of parameters	Matched log-linear model
M1	$\beta_{jk\ell\|i} = \beta_{0\|i} + \beta_{j\|i} + \beta_{k\|i} + \beta_{\ell\|i}$	$J+K+L-2$	$< X_2R, X_3R, YR >$
M2	$\beta_{jk\ell\|i} = \beta_{0\|i} + \beta_{j\|i} + \beta_{k\|i} + \beta_{jk\|i}$	JK	$< X_2R, X_3R, X_2X_3R >$
M3	$\beta_{jk\ell\|i} = \beta_{0\|i} + \beta_{j\|i} + \beta_{\ell\|i} + \beta_{j\ell\|i}$	JL	$< X_2R, YR, X_2YR >$
M4	$\beta_{jk\ell\|i} = \beta_{0\|i} + \beta_{j\|i} + \beta_{k\|i} + \beta_{\ell\|i} + \beta_{jk\|i} + \beta_{j\ell\|i}$	$JK+JL-J$	$< X_2R, X_3R, YR, X_2X_3R, X_2YR >$

치므로 MAR 메카니즘을 따른다. M3의 경우 X_2와 Y 및 이들의 교호작용이 무응답에 영향을 미치는 모형으로 MNAR 메카니즘을 따른다. M4의 경우 M2 및 M3가 혼합된 모형이다. 이를 선거 자료에 대입하여 보면, M1 모형은 성, 연령 및 지지후보자가 무응답에 영향을 주며, M2 모형은 성, 연령 및 각 성에 따른 연령 효과가 무응답에 영향을 줌을 의미한다. M3 모형은 성, 지지후보자 및 성에 따른 지지후보자 효과가 무응답에 영향을 줌을 나타낸다.

Table 3.1의 각 모형의 모수 수와 식 (3.3)의 모수($\gamma_{jk\ell|i}$)의 수인 $JKL-1$을 함께 고려하면, M1은 $L=K$이고 $J=2$인 경우 포화모형(saturated model)이 되며, M2는 항상 포화모형이다. M3의 경우 $L=K$인 경우 포화모형이 되며, M4는 초과모수모형(overparameterized model)이다. 따라서 본 논문에서 고려하는 성, 연령 및 후보자의 수를 고려하면, 각 지역에서 M1, M2, M3 모두 포화모형이 됨을 알 수 있다.

본 논문에서는 지역간 성, 연령별 지지후보가 서로 독립임을 가정하였기 때문에, 각 지역별로 식 (3.3)의 모수인 $\gamma_{jk\ell|i}$ 및 Table 3.1의 $\beta_{jk\ell|i}$에 대한 모수들을 추정하도록 하며, 최우추정법을 이용하여 아래의 지역별 로그우도함수($\log L_i$)가 최대화되는 값을 추정치로 한다.

$$\log L_i = \sum_j \sum_k \sum_\ell y_{ijk\ell 1} \log \pi_{jk\ell 1|i} + \sum_j \sum_k y_{ijk+0} \log \pi_{jk+0|i}. \tag{3.5}$$

지역 i에서 $\gamma_{jk\ell|i}$를 집적한 모수 벡터를 $\boldsymbol{\gamma_i}$, $\beta_{jk\ell|i}$에 대한 모형의 모수를 집적한 모수벡터를 $\boldsymbol{\beta_i}$라 하고, $\boldsymbol{\theta_i} = (\boldsymbol{\gamma_i}'\ \boldsymbol{\beta_i}')'$라고 했을 때, $\boldsymbol{\theta_i}$의 분산공분산 행렬은

$$\mathrm{Cov}(\boldsymbol{\theta_i}) = \begin{pmatrix} \mathrm{Cov}(\boldsymbol{\gamma_i}) & \mathrm{Cov}(\boldsymbol{\gamma_i}, \boldsymbol{\beta_i}) \\ \mathrm{Cov}(\boldsymbol{\beta_i}, \boldsymbol{\gamma_i}) & \mathrm{Cov}(\boldsymbol{\beta_i}) \end{pmatrix} \tag{3.6}$$

로 표현될 수 있다. $\hat{\boldsymbol{\theta}}_i$이 식 (3.5)의 최우추정치라고 할 때, $\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\theta}}_i)$는 $\hat{\boldsymbol{\theta}}_i$에 대해 식 (3.5)의 음의 헤이지안 행렬(Hessian matrix)의 역행렬을 구함으로써 구할 수 있으며, 본 논문에서는 이를 수치해석적으로 구하였다.

$\hat{p}_{jk\ell|i}$는 추정된 $\hat{\boldsymbol{\gamma}}_i$를 식 (3.3)에 대입하여 구하며, $\hat{p}_{jk\ell|i}$의 분산공분산 행렬은 다음과 같이 추정한다 (Agresti, 2002). $\hat{\boldsymbol{p}}_i$를 지역 i에서의 $\hat{p}_{jk\ell|i}$를 집적한 벡터하고 하면,

$$\widehat{\mathrm{Cov}}(\hat{\boldsymbol{p}}_i) = \left[\mathrm{diag}(\hat{\boldsymbol{p}}_i) - \hat{\boldsymbol{p}}_i\hat{\boldsymbol{p}}_i'\right] \widehat{\mathrm{Cov}}(\hat{\boldsymbol{\gamma}}_i) \left[\mathrm{diag}(\hat{\boldsymbol{p}}_i) - \hat{\boldsymbol{p}}_i\hat{\boldsymbol{p}}_i'\right]' \tag{3.7}$$

이다. 지역 i에서 후보자 ℓ에 대한 추정 지지율은

$$\hat{p}_{++\ell|i} = \sum_j \sum_k \hat{p}_{jk\ell|i} \tag{3.8}$$

이며, 분산은 식 (3.7)의 결과를 이용하여 다음과 같이 추정할 수 있다.

$$\widehat{\mathrm{Var}}(\hat{p}_{++\ell|i}) = \sum_{j,j'} \sum_{k,k'} \widehat{\mathrm{Cov}}(\hat{p}_{jk\ell|i}, \hat{p}_{j'k'\ell|i}) \tag{3.9}$$

Table 3.2. Estimated cell counts in Seoul

Gender	Ages	$R=1$					$R=0$					
		Moon	Hong	Ahn	Yu	Others	Moon	Hong	Ahn	Yu	Others	Marginal
Male	19–29	10.00	0.62	3.12	1.88	2.50	1.03	0.06	0.32	0.19	0.26	1.88
	30–39	11.74	2.47	1.24	1.24	1.24	2.02	0.43	0.21	0.21	0.21	3.09
	40–49	12.19	1.35	3.39	0.00	2.71	0.84	0.09	0.23	0.00	0.19	1.36
	50–59	8.28	4.83	3.45	0.00	0.69	1.32	0.77	0.55	0.00	0.11	2.76
	60+	0.73	9.53	3.67	0.00	0.00	0.42	5.52	2.12	0.00	0.00	8.07
Female	19–29	1.33	0.00	0.00	3.33	6.67	1.02	0.00	0.00	2.55	5.10	8.67
	30–39	10.22	2.27	1.13	1.70	1.70	2.38	0.53	0.26	0.40	0.40	3.97
	40–49	12.29	0.00	4.53	1.29	0.00	2.63	0.00	0.97	0.28	0.00	3.88
	50–59	5.91	3.28	5.91	0.00	0.00	2.31	1.28	2.31	0.00	0.00	5.91
	60+	6.39	9.24	2.13	0.00	0.00	3.33	4.80	1.11	0.00	0.00	9.24

각 지역에서 추정된 후보자별 지지율 및 이의 분산을 이용하여, 전국 단위에서의 후보자별 지지율 및 분산을 구하게 되는데, 전국 단위에서의 후보자별 지지율은 식 (3.1) 및 (3.2)에 의해

$$\hat{\pi}_{++\ell+} = \sum_i \sum_j \sum_k \sum_r \hat{\pi}_{ijk\ell r} = \sum_i \hat{\pi}_i \times \hat{p}_{++\ell|i} \tag{3.10}$$

이며, 여기서 $\hat{\pi}_i$로는 여론조사 설계에서 이용되는 각 지역별 배분 비중을 이용하도록 한다. $\hat{\pi}_{++\ell+}$의 분산은 식 (3.10) 및 앞서 가정한 각 지역에서의 후보자 지지율의 독립성에 의해

$$\widehat{\mathrm{Var}}(\hat{\pi}_{++\ell+}) = \sum_i \hat{\pi}_i^2 \widehat{\mathrm{Var}}(\hat{p}_{++\ell|i}) \tag{3.11}$$

이다. Table 3.2는 서울 지역에 M2 모형을 적합한 결과이다. M2 모형 적합 결과, 서울 지역의 경우 문재인 후보의 지지율은 44.5%로 나타났으며, 95% 신뢰구간은 (37.2%, 51.9%)로 추정되었다. 전국 지지율의 경우 M2 모형 적합 결과 문재인 후보는 45.6%였으며, 홍준표 후보의 경우 20.6%, 안철수 후보의 경우 18.9%, 유승민 후보는 5.2%의 지지를 받는 것으로 추정되었다.

4. 다차원 불완전 분할표에서의 민감도 분석

4절에서는 3절에서 소개된 모형을 기초로, 다차원 불완전 분할표에서의 민감도 분석을 소개하고자 한다. 선거 여론 조사에서는 관심의 대상이 되는 것은 각 후보에 대한 지지율이므로, 본 논문에서 소개되는 민감도 분석의 관심 모수 역시 각 지역에서의 후보별 지지율을 나타내는 $\hat{p}_{++\ell|i}$ 또는 전국 단위의 후보별 지지율인 $\hat{\pi}_{++\ell+}$이다.

서론에서 밝혔듯이, 관측된 자료만을 이용해서는 무응답 메카니즘, 혹은 모형의 선택이 불가능하다. 이러한 문제를 해결하기 위해 Molenberghs 등 (2001)은 다양한 무응답 모형을 이용하여 관심모수가 가질 수 있는 값을 구간으로 표현하는 민감도 분석을 제안하였다. 이들은 다양한 무응답 모형을 포괄하는 초과모형모수를 기반으로, 하나 혹은 두 개의 민감도 모수가 변화함에 따른 관심 모수의 변화를 파악하고자 하였다.

Molenberghs 등 (2001)은 관심모수가 가질 수 있는 값을 무지의 구간(region of ignorance)과 불확실성의 구간(region of uncertainty)으로 나타낼 것을 제안하였는데, 무지의 구간은 민감도 모수가 변함에 따른 관심모수의 변화를 구간으로 나타낸 것이며, 불확실성의 구간은 관심모수의 변화 뿐 아니라 신뢰구간의 변화 역시 구간으로 나타낸 것이다. Molenberghs 등 (2001)에서 제안한 민감도 분석은 2차원 불

완전 분할표를 대상으로 했으며, 각 변수의 범주 수 역시 2개에 그친다. 본 논문에서는 이를 우리나라의 선거여론조사 자료에 적합하도록 확장하도록 한다. 본 논문에서 제시하는 민감도 분석은 지역간 독립 가정으로 인해 각 지역별 후보지지율에 대한 민감도 분석이 가능할 뿐만 아니라, 이를 통합하여 전국 단위에서의 후보지지율에 대한 민감도 분석 역시 가능하다는 장점이 있다.

먼저 각 지역별 민감도 분석을 고려하도록 한다. 민감도 분석의 대상이 되는 초과모수모형으로는 Table 3.1의 M4 모형을 선택하고, MNAR 메카니즘에 해당하는 모수들을 민감도 모수로 설정한다. M4 모형의 경우, 모형의 성격을 규정하는 $\beta_{jk\ell|i}$는 다음과 같다.

$$\beta_{jk\ell|i} = \beta_{0|i} + \beta_{j|i} + \beta_{k|i} + \beta_{\ell|i} + \beta_{jk|i} + \beta_{j\ell|i}.$$

위 식은 MAR 메카니즘을 나타내는 M2 모형에 MNAR 메카니즘을 나타내는 $\beta_{\ell|i}$ 및 $\beta_{j\ell|i}$ 모수가 추가된 것으로, 본 논문에서는 $\beta_{\ell|i}$ 및 $\beta_{j\ell|i}$를 민감도 모수로 설정하도록 한다. 이는 M2 모형의 경우 포화모형으로, Table 3.2에서 보듯이 관측치와 추정치가 완전히 일치(perfect fit)이 될 뿐 아니라, MNAR 메카니즘 하에서 발생하는 경계점 문제(boundary solution) 역시 존재하지 않기 때문이다. 민감도 분석은 지역 단위의 민감도 분석과, 이를 이용한 전국단위의 민감도 분석으로 진행되는데, 지역 i에서 후보자 ℓ에 대한 민감도 분석은 다음과 같은 절차를 통해 시행된다.

Step 1: 민감도 모수 $\beta_{\ell|i}$ 및 $\beta_{j\ell|i}$의 범위를 정한다.

Step 2: $\beta_{\ell|i}$ 및 $\beta_{j\ell|i}$에 설정한 범위 내의 특정 값을 부여하고, 식 (3.5)를 최대화하는 $\boldsymbol{\gamma}_i$ 및 민감도 모수를 제외한 $\boldsymbol{\beta}_i$를 추정한다.

Step 3: 추정된 모수 및 지정된 민감도 모수를 이용하여 식 (3.8)의 $\hat{p}_{++\ell|i}$ 및 식 (3.9)의 $\widehat{\mathrm{Var}}(\hat{p}_{++\ell|i})$를 추정한다.

Step 4: Step 3의 결과를 이용하여 $\hat{p}_{++\ell|i}$의 신뢰구간을 추정한다.

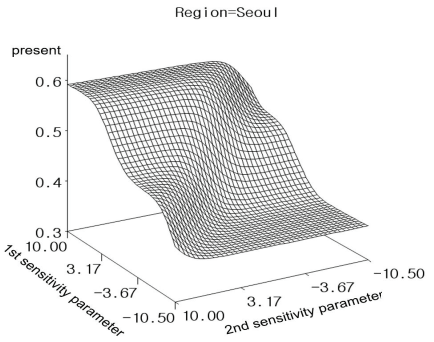
Step 5: 민감도 모수 $\beta_{\ell|i}$ 및 $\beta_{j\ell|i}$의 범위 내의 각 값에 대해 Step 2부터 Step 4를 반복한다.

Step 6: 반복 결과 도출된 $\hat{p}_{++\ell|i}$의 최대값 및 최소값으로 무지의 구간을 산출하며, $\hat{p}_{++\ell|i}$의 신뢰하한들의 최소값, 신뢰상한들의 최대값으로 불활실성의 구간을 구한다.

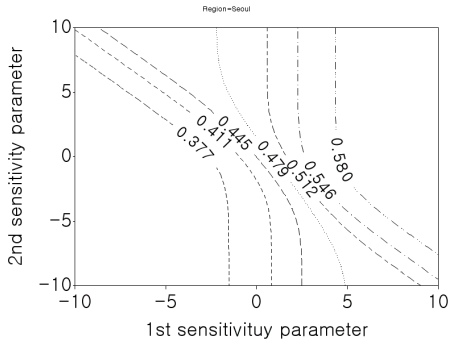
다음으로 전국 단위에서 후보자 ℓ의 지지율에 대한 민감도 분석을 해보도록 하자. 우선 각 지역의 민감도 분석을 전부 수행한 후, 설정된 민감도 모수의 각 값에서 추정된 $\hat{p}_{++\ell|i}$ 및 $\widehat{\mathrm{Var}}(\hat{p}_{++\ell|i})$를 식 (3.10) 및 (3.11)에 대입하여 $\hat{\pi}_{++\ell+}$ 및 $\widehat{\mathrm{Var}}(\hat{\pi}_{++\ell+})$를 구한다. 이 때 무지의 구간은 각 민감도 모수에서 추정된 $\hat{\pi}_{++\ell+}$의 최대값 및 최소값이며, 불확실성의 구간은 $\hat{\pi}_{++\ell+}$의 신뢰하한들의 최소값, 신뢰상한들의 최대값이다.

위에서 설명한 민감도 분석을 우리나라의 19대 대선 선거여론조사 자료에 적용하여 보도록 한다. 먼저 문재인 후보 지지율에 대한 지역별 민감도 분석을 실시하고, 이를 통해 전국 단위의 민감도 분석을 수행하였다. 민감도 모수인 $\beta_{\ell|i}$ 및 $\beta_{j\ell|i}$는 -10부터 10까지의 값을 설정하였으며, 하한값부터 0.5씩 값을 증가시켰다. Figure 4.1은 민감도 모수의 변화에 따른 서울 지역의 문재인 후보자에 대한 지지율인 $\hat{p}_{++\ell|i}$를 3차원 그림 및 등고선으로 나타낸 것이다.

Figure 4.1에서 볼 수 있듯이 $\beta_{\ell|i}$(Figure 4.1에서 첫 번째 민감도 모수)이 커질수록, 즉 문재인 후보자를 지지하는 경우의 무응답 확률이 커질수록, 지지율인 $\hat{p}_{++\ell|i}$가 상승하는 것을 볼 수 있다. 또한 여성 유권자 효과를 나타내는 $\beta_{j\ell|i}$(Figure 4.1에서 두 번째 민감도 모수)가 커질수록 $\hat{p}_{++\ell|i}$가 상승하는 것을 알 수 있다. Figure 4.2는 Figure 4.1의 그림들을 2차원으로 표현한 것이다.

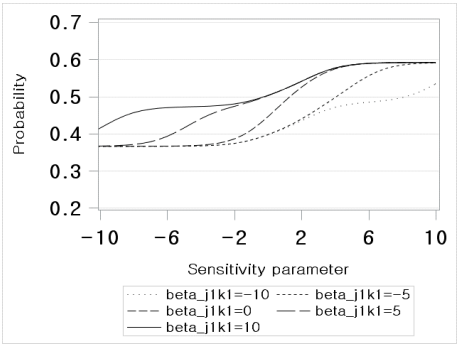


(a) 3-dimensional plot

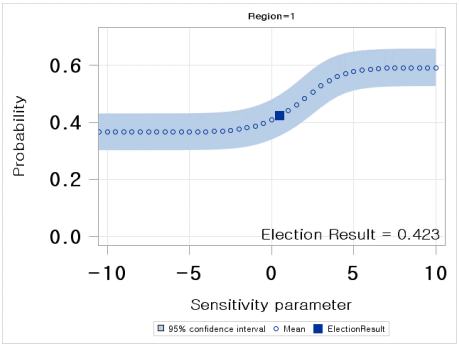


(b) contoure plot

Figure 4.1. Sensitivity plots.



(a) intervals of ignorance by $\beta_{j\ell|i}$'s



(b) interval of uncertainty at $\beta_{j\ell|i} = -2$

Figure 4.2. Sensitivity plots sliced by $\beta_{j\ell|i}$.

Figure 4.2(a)는 Figure 4.1의 3차원 그림을 2차원으로 투영한 것으로 무지의 구간을 나타내고 있다. 여기서 X축은 $\beta_{\ell|i}$이며 Y축은 지지율인 $\hat{p}_{++\ell|i}$을 나타내고, $\beta_{j\ell|i}$의 각 값별로 $\beta_{\ell|i}$의 변화에 따른 $\hat{p}_{++\ell|i}$의 변화를 실선 및 점선 등으로 나타내고 있다. $\beta_{j\ell|i}$의 값이 커짐에 따라 $\hat{p}_{++\ell|i}$를 나타내는 선들이 빠르게 상승하는 양상을 보이고 있다. 또한 두 민감도 모수들의 값이 커지거나 작아질 때, 지지율의 최대값 및 최소값은 일정한 값으로 수렴함을 알 수 있다. 이와 같이 수렴된 최대값 및 최소값으로 무지의 구간을 생성하게 되는데, 서울 지역 문재인 후보 지지율의 무지의 구간은 $(0.366, 0.592)$이다.

Figure 4.2(b)는 $\beta_{j\ell|i}$가 -2의 값을 가질 때 $\beta_{\ell|i}$의 각 값에서 추정된 $\hat{p}_{++\ell|i}$의 95% 신뢰구간을 나타내고 있다. 여기서 동그란 점은 $\hat{p}_{++\ell|i}$의 값, 네모로 표현된 점은 실제 서울 지역의 개표 결과이다. $\beta_{\ell|i}$의 값이 0.5일 때 M4 모형의 예측치는 개표 결과인 0.423과 일치하였다.

Table 4.1은 각 지역 및 전국에서의 문재인 후보자에 대한 개표 결과와 M2 모형 예측치 및 95% 신뢰구간, 민감도 분석 결과인 무지의 구간 및 불확실성의 구간을 나타내고 있다. 인천/경기 지역, 광주/전라 지역의 경우 개표결과가 M2 모형으로부터 구한 신뢰구간 밖에 존재하는 것을 볼 수 있다. 이는 해당 지역의 경우 무응답의 발생이 MNAR 메카니즘을 따름을 나타낸다. 해당 지역의 경우 무지의 구간도 개표결과를 포함하고 있지 않지만, 불확실성의 구간은 해당 지역의 개표결과를 포함하고 있다.

비록 타 지역의 경우 M2 모형으로부터 추정한 95% 신뢰구간이 개표 결과를 포함하고 있지만, 무응

Table 4.1. Sensitivity analysis of Moon

Region	Election result	Results from M2		Interval of ignorance	Interval of uncertainty
		Estimates	Confidence interval		
Seoul	0.423	0.445	(0.372, 0.519)	(0.366, 0.592)	(0.301, 0.657)
Incheon/Gyeonggi	0.419	0.516	(0.458, 0.574)	(0.439, 0.587)	(0.382, 0.647)
Daejeon/Chungcheong	0.404	0.464	(0.365, 0.563)	(0.384, 0.565)	(0.297, 0.655)
Gwangju/Jeolla	0.620	0.501	(0.400, 0.603)	(0.406, 0.589)	(0.308, 0.770)
Daegue/Gyeongbuk	0.217	0.311	(0.217, 0.406)	(0.244, 0.446)	(0.156, 0.538)
Busan/Ulsan/Gyeongnam	0.378	0.444	(0.365, 0.523)	(0.392, 0.537)	(0.318, 0.611)
Gangwon/Jeju	0.374	0.346	(0.202, 0.490)	(0.278, 0.405)	(0.464, 0.405)
Country	0.411	0.456	(0.376, 0.536)	(0.381, 0.556)	(0.309, 0.637)

Table 4.2. The estimates of MNAR model nearest to the election result

Region	Election result	M2	MNAR nearest to the election result			
			$\beta_{\ell\|i}$	$\beta_{j\ell\|i}$	Estimates	CI
Seoul	0.423	0.445	0.5	−2	0.423	(0.352, 0.495)
Incheon/Gyeonggi	0.419	0.516	−10	0	0.439	(0.385, 0.492)
Daejeon/Chungcheong	0.404	0.464	−5.5	3.5	0.404	(0.313, 0.495)
Gwangju/Jeolla	0.620	0.501	10	0	0.589	(0.499, 0.680)
Daegue/Gyeongbuk	0.217	0.311	−10	0	0.244	(0.165, 0.323)
Busan/Ulsan/Gyeongnam	0.378	0.444	−10	0	0.392	(0.319, 0.464)
Gangwon/Jeju	0.374	0.346	−3.5	5.5	0.374	(0.234, 0.513)
Country	0.411	0.456			0.417	(0.343, 0.491)

MNAR = missing not at random.

답 메카니즘이 MNAR 메카니즘을 따를 수 있다는 가능성을 여전히 배제할 수 없다 (Molenberghs 등, 2009). 이에 따라 본 논문에서는 민감도 모수가 어떠한 값을 가질 때 M4 모형의 예측 결과가 개표 결과와 가장 유사한지 파악함으로써 각 지역별로 나타나는 MNAR 메카니즘의 특징을 규명해 보고자 한다. 이를 향후 선거 여론조사의 무응답 추정에 반영하면, 예측의 정밀성을 높일 수 있을 것으로 여겨진다.

Table 4.2는 실제 개표결과와 가장 유사한 M4 모형의 예측치 및 이에 해당하는 민감도 모수의 값을 보이고 있다. 서울 지역의 경우 $\beta_{\ell|i}$의 값은 0.5, $\beta_{j\ell|i}$의 값은 -2일 때의 예측치가 0.423으로 실제 개표결과와 일치하였다. 이는 문재인 후보 지지자 가운데, 남성은 무응답할 확률이 MAR 메카니즘에 비해 더 높은데 반해, 여성의 경우 무응답할 확률이 더 낮음을 의미한다. 광주/전라 지역의 경우 $\beta_{\ell|i}$는 10, $\beta_{j\ell|i}$가 0의 값을 가질 때 개표결과와 가장 가까웠으며, 이는 문재인 후보 지지자의 경우 무응답할 확률이 MAR보다 더 높으며, 성별에 따른 차이는 없는 것을 의미한다. 대전/충청 지역과 강원/제주 지역의 경우 $\beta_{\ell|i}$는 음의 값을 가지며, $\beta_{j\ell|i}$는 양의 값을 가지는데, 이는 문재인 후보 지지자 가운데 남성은 무응답할 확률이 MAR 메카니즘보다 더 낮으나, 여성의 경우 더 높음을 의미한다. 그 외 지역의 경우 남녀 차이는 없었으며, 문재인 후보 지지자의 경우 무응답할 확률이 MAR 메카니즘보다 더 낮았다.

Table 4.2에서 제시된 실제 개표결과와 가장 유사한 지역별 M4 모형의 예측치를 이용하여 문재인 후보에 대한 전국 지지율을 예측한 결과, 예측치는 0.417이며, 95% 신뢰구간은 $(0.343, 0.491)$로, M2 모형의 예측치보다 개표 결과에 더 가까운 것을 알 수 있다.

본 논문에서는 지면 관계상 문재인 후보 지지율에 대한 민감도 분석 결과 및 개표결과와 가장 유사한 민감도 모수의 값만을 제시하였음을 밝힌다. 타 후보자에 대해서도 이와 동일하게 민감도 분석을 진행할 수 있다. 전국 단위에서의 민감도 분석 결과만을 제시하면, 홍준표 후보자의 경우 무지의 구

간은 $(0.168, 0.343)$, 불확실성의 구간은 $(0.113, 0.433)$이었으며, 안철수 후보자의 경우 무지의 구간은 $(0.157, 0.331)$, 불확실성의 구간은 $(0.097, 0.404)$이었다.

5. 결론

불완전 분할표 분석에서는 무응답 메카니즘에 따라 추정 결과가 달라지며, 무응답 메카니즘이 잘못 지정되는 경우 편의가 발생하거나 분산 추정치가 크게 추정되는 문제점이 존재한다. 그러나 관측된 자료를 이용하여 무응답 메카니즘을 판별하는 것은 불가능하며, 이러한 문제를 해결하기 위한 방안으로 민감도 분석이 제안되었다. Molenberghs 등 (2001)은 MAR 및 MNAR 메카니즘을 포괄하는 모형을 설정하고, 이 가운데 MNAR 메카니즘을 나타내는 모수를 민감도 모수로 하여 관심모수의 변화를 두 개의 구간으로 제시하는 민감도 분석을 제안하였다. 그러나 이들의 방법론은 2원 분할표에서 각 변수의 범주수가 2개인 경우로 제한되어 있다.

본 논문에서는 이 방법론을 다차원 분할표에 적용되도록 확장하였으며, 개발된 방법론을 우리나라 19대 대선 선거여론 조사에 적용하였다. 본 논문에서는 각 지역간 성, 연령에 따른 후보자 지지가 서로 독립이라는 가정 하에서 각 지역별로 민감도 분석을 실시하였으며, 이를 통합하여 전국에 대한 민감도 분석 결과를 도출하였다. 분석 결과 인천/경기 지역, 광주/전라 지역의 경우 개표 결과가 MAR 가정으로부터 구한 신뢰구간 밖에 존재하였으나, 불확실성의 구간은 개표 결과를 포함하고 있었다. 이는 해당 지역의 경우 무응답의 발생이 MNAR 메카니즘을 따름을 의미한다. 타 지역의 경우 역시 Molenberghs 등 (2008)이 언급한 바와 같이 MNAR 메카니즘을 배제할 수 없으므로, 개표 결과와 가장 유사한 값을 갖는 NMAR 모형의 추정치 및 해당 값에서의 민감도 모수 값을 구하였다. 이를 통해 각 지역별 MNAR 메카니즘의 특징을 파악할 수 있었으며, 이를 향후 대선 여론조사에 적용하여 예측의 정밀성을 높일 수 있을 것으로 기대된다. 그러나 이를 위해서는 다양한 대선 여론조사 자료를 분석함으로써, 민감도 모수의 변화 추이를 파악하는 것이 선행되어야할 것으로 여겨진다.

References

Agresti, A. (2002). *Categorical Data Analysis* (2nd Ed), Wiley & Sons, New York.

Baker, S. G., Ko, C., and Graubard, B. I. (2003). A sensitivity analysis for nonrandomly missing categorical data arising from a national health disability survey, *Biostatistics*, **4**, 41–56.

Baker, S. G. and Laird, N. M. (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse, *Journal of American Statistical Association*, **83**, 62–69.

Baker, S. G., Rosenberger, W. F., and Dersimonian, R. (1992). Closed-form estimates for missing counts in two-way contingency tables, *Statistics in Medicine*, **11**, 643–657.

Choi, B. S., Choi, J. W., and Park, Y. (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 boundary solutions and identifiability in categorical regression with non-ignorable non-response, *Biometrical Journal*, **44**, 701–717.

Fay, R. E. (1986). Causal models for patterns of nonresponse, *Journal of American Statistical Association*, **81**, 354–365.

Forster, J. J. and Smith, P. W. F. (1998). Model-based inference for categorical survey data subject to nonignorable nonresponse, *Journal of the Royal Statistical Society: Series B*, **60**, 57–70.

Kim, S. (2016). Sensitivity analysis for uncertainty of missing mechanisms in an incomplete contingency table, *Journal of the Korean Data Analysis Society*, **18**, 1845–1855.

Kim, S. and Kim, D. (2018). Assessment of nonignorable log-linear models for an incomplete contingency table, *Statistica Sinica*, **28**, 1887–1905.

Kim, S., Park, Y., and Kim, D. (2015). On missing-at-random mechanism in two-way incomplete contingency tables, *Statistics and Probability Letters*, **96**, 196–203.

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

Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data, *Journal of American Statistical Association*, **88**, 125–134.

Molenberghs, G., Beunckens, C., Sotto, C., and Kenward, M. G. (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.

Molenberghs, G., Kenward, M. G., and Goetghebeur, E. (2001). Sensitivity analysis for incomplete contingency tables: the Slovenian plebiscite case, *Journal of the Royal Statistical Society: Series C*, **50**, 15–29.

Poleto, F. Z., Singer, J. M., and Paulino, C. D. (2011). Missing data mechanisms and their implications on the analysis of categorical data, *Statistics and Computing*, **21**, 31–43.

Vansteelandt, S., Goetghebeur, E., Kenward, M. G., and Molenberghs, G. (2006). Ignorance and uncertainty regions as inferential tools in a sensitivity analysis, *Statistica Sinica*, **16**, 953–979.

19대 대선 여론조사에서 무응답 메카니즘의 민감도 분석

김성용[a,1] · 곽동호[a]

[a]호서대학교 빅데이터경영공학부

(2018년 9월 14일 접수, 2018년 12월 3일 수정, 2018년 12월 3일 채택)

요 약

선거여론조사 자료의 경우 무응답이 흔히 관측되며, 이와 같이 무응답이 존재하는 범주형 자료는 불완전 분할표로 표현된다. 불완전 분할표로 표현된 선거여론조사 자료에서 후보자 지지율을 추정하는 경우, 지지율은 무응답이 어떤 메카니즘을 따르는가에 따라 다르게 추정되며, 따라서 자료가 어떠한 무응답 메카니즘을 따르는지에 대한 판별이 분석에 선행되어야 한다. 그러나 최근 연구에 따르면, 관측된 자료를 이용해서는 무응답 메카니즘을 판별할 수 없음이 밝혀졌다. 이러한 문제를 해결하기 위해 다양한 무응답 메카니즘을 반영할 수 있는 민감도 분석이 제안되었다. 그러나 기존에 제안된 민감도 분석의 경우, 이원 분할표에서 각 변수의 범주 수가 두 개인 경우만을 대상으로 한다. 우리나라 선거여론조사에서 고려되는 요인이 지역, 성, 연령 등임을 감안할 때, 기존 방법론으로 민감도 분석을 시행하기에는 한계점이 존재한다. 이에 따라 본 논문에서는 기존의 민감도 분석을 다차원 불완전 분할표에 적용할 수 있도록 확장하고, 이를 우리나라 19대 대선 여론조사 자료에 적용하였다. 분석 결과, 민감도 분석의 구간이 실제 지지율을 포함하고 있을 뿐 아니라, 다양한 무응답 메카니즘의 결과를 포괄하고 있으며, 실제 지지율과 가장 가까운 예측치의 경우 후보자에 대한 지지가 무응답의 발생에 영향을 미침을 알 수 있었다.

주요용어: 무응답 대체, 민감도 분석, 불완전 분할표, 무응답 메카니즘.

이 논문은 2016년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NRF-2016 R1D1A3B03930392).

[1]교신저자: (31499) 충청남도 아산시 배방읍 호서로 79번길 20, 호서대학교 과학기술융합대학 빅데이터경영공학부. E-mail: yaba96@hoseo.edu