The Korean Journal of Applied Statistics (2017) DOI: https://doi.org/10.5351/KJAS.2017.30.6.943
30(6), 943–955

Causal study on the effect of survey methods in the 19th presidential election telephone survey

Ji-Hyun Kim[a,1] · Hyojae Jung[a]

[a]Department of Statistics and Actuarial Science, Soongsil University

(Received September 28, 2017; Revised November 16, 2017; Accepted December 7, 2017)

Abstract

We investigate and estimate the causal effect of the survey methods in telephone surveys for the 19th presidential election. For this causal study, we draw a causal graph that represents the causal relationship between variables. Then we decide which variables should be included in the model and which variables should not be. We explain why the research agency is a should-be variable and the response rate is a should-not-be variable. The effect of ARS can not be estimated due to data limitations. We have found that there is no significant difference in the effect of the proportion of cell phone survey if it is less than about 90 percent. But the support rate for Moon Jae-in gets higher if the survey is performed only by cell phones.

Keywords: causal graph, confounding, multivariate adaptive regression splines (MARS), telephone survey

1. 연구목적

2017년 4월 기준 이동전화 서비스 가입자 수가 6200만 명을 넘어 성인 대부분이 개인 휴대전화기를 보유하게 된 지금, 전화를 이용한 여론조사 방식도 집전화인 유선전화만 이용하는 방식에서 개인휴대전화인 무선전화와 병행하여 조사하는 방식으로 바뀌고 있다. 국내 전화여론조사방법에 대한 연구도 유선전화 임의번호걸기 방식에 대한 연구에서 (Kang 등, 2008; Huh와 Kim, 2008) 무선전화를 병행하는 방식에 대한 연구로 (Kim과 Woo, 2012; Lee 등, 2012; Jang과 Cho, 2013) 관심이 옮겨가고 있다. 하지만 아직 유선전화와 무선전화, 두 방식에 따른 조사 결과 차이와 두 방식의 적정 혼용비율에 대한 연구가 다양한 자료에서 충분히 이루어지지 않았다. 유선전화조사에서 조사된 표본의 특성이 목표하는 모집단의 특성과 다를 수 있는데, 성별과 나이, 지역으로 할당표집하고 가중치를 두어 보정하더라도 표본의 편향을 제거하기는 힘들다. 예를 들어 집전화로 조사된 특정 지역의 50대 남성 조사모집단이 그 지역의 50대 남성 목표모집단과 다른 특성을 갖는다면 조사결과를 모집단의 성별/나이/지역 비율에 따른 가중치를 이용하여 보정하더라도 편향이 제거되지 않는다. 무선전화를 이용하면 이런 조사모집단과 목표모집단의 괴리를 줄일 수 있으나 무선전화로 이루어진 조사는 또 다른 편향을 초래할 수 있다는 우려도 제기되었다 (Kim과 Woo, 2012). 또한, 유선전화와 무선전화에 따른 차이 말고 자동응답조사(ARS) 방식과 면접조사 방식에 따른 차이도 문제가 된다.

[1]Corresponding author: Department of Statistics and Actuarial Science, Soongsil University, Sangdo-Ro 369, Dongjak-Gu, Seoul 06978, Korea. E-mail: jxk61@ssu.ac.kr

이 연구의 목적은 크게 두 가지이다. 우선 유선/무선, ARS/면접과 같이 서로 다른 조사방법이 조사결과에 어떤 영향을 주는지를, 즉, 조사방법의 인과성(causation)을 19대 대선예측조사 자료에 근거해 살펴보고자 한다. 다음으로 인과관계를 밝히기 위해 회귀모형을 쓰는 경우 변수선택을 어떻게 해야 하는가에 대해서도 설명하고자 한다 (여기서 회귀모형은 반응변수를 설명변수의 함수로 나타내고자 하는 통계적 모형에 대한 일반적 용어로서 선형임의효과모형(linear random effect model)과 비모수회귀모형을 포함한다.). 흔히 회귀모형에 포함시킬 설명변수를 선택할 때 반응변수와 관련 있는 변수를 모두 포함시켜 모형의 설명력 또는 예측력을 높이고자 하는데, 특정 변수의 인과관계를 밝히고자 하는 것이 목적인 경우에는 변수 선택에 있어 조심해야 한다 (Pearl, 2009; Thepepomma와 Kim, 2016). 중요하지만 간과하기 쉬운 이 사실을 강조하여, 인과관계를 밝히고자 회귀모형을 쓰는 연구자들에게 참고가 되게 하였다.

논문 구성은 다음과 같다. 2절에서 분석에 쓰인 자료를 소개하고 탐색적 분석을 통해 파악한 자료의 특성을 설명한다. 3절에서 인과연구(causal study)에서 쓰이는 인과 그래프(causal graph)를 본 연구의 문제에 대해 그려보고 이를 이용하여 모형에 포함시켜야 하는 변수는 무엇이고 제외시켜야 하는 변수는 무엇인지에 대해 알아본다. 4절과 5절에서 조사방법이 조사결과에 미치는 영향을 다양한 회귀모형을 이용해 알아본다. 6절에서 결과를 요약하고 첨언하였다.

2. 분석자료

여론조사 결과는 공직선거법 및 선거여론조사기준에 따라 중앙선거여론조사심의위원회 홈페이지에 의무적으로 등록하게 되어 있다. 본 연구에 쓰인 자료는 여기에 등록된 19대 대선 관련 선거여론조사 자료이다. 5개 정당 중에서 가장 마지막으로 경선이 끝난 더불어민주당의 후보가 결정된 다음 날인 2017년 4월 4일부터 여론조사결과 공표 금지가 시작되기 전 날인 5월 2일까지 등록된 전화조사 자료를 분석에 썼으며, 조사 개수는 87개로서 5개 주요 정당 후보 지지율이 기록된 전화조사는 전부 포함시켰다.

5개 주요 정당 후보 지지율 중에서 분석에 쓴 반응변수는 당선자인 문재인 후보의 지지율이다. 다른 후보들은 선거 기간 중 지지율 등락이 심해 조사의 정확성을 판단하기에 어려움이 있어서 문재인 후보의 지지율을 이용하였다. 지지율과 같은 비율을 회귀모형의 반응변수로 쓸 때 역 사인 변환 $\sin^{-1}(\sqrt{p})$이 권장되기도 하지만, 비율이 0.2에서 0.8 사이이면 선형변환에 가까워 변환 효과가 없다. 문재인 후보 조사 지지율은 0.322에서 0.482 사이였다.

조사방법을 나타내는 변수가 네 개 있는데 유선ARS, 무선ARS, 유선면접, 무선면접의 조사비율이다. 이 네 변수의 값을 합하면 항상 1이 된다. 무선ARS 조사비율이 1인, 즉 ARS 방식으로 무선전화에 대해서만 조사가 이루어진 것이 전체 87개 중에서 12개였다. 나머지 75개 조사는 두 가지 이상의 방식을 혼용하였다. 세 가지 방식을 혼용한 조사가 15개였는데 모두 유선ARS, 무선ARS, 무선면접 방식을 혼용하였다. 나머지 60개 조사는 두 가지 방식을 혼용하였다. 그리고 87개 모든 조사에서 무선조사가 반드시 이루어졌는데 무선ARS와 무선면접 조사비율을 합한 값의 최솟값이 0.46이었다. 유선전화로만 조사하지 않고 무선전화 조사를 병행했다는 것이 이번 19대 대선 전화여론조사 특징 중 하나이다.

조사기관도 조사결과에 영향을 미치는 변수인데 조사기관의 공신력과 축적된 조사기술 등이 종합된 변수이다. 조사기관 수는 총 19개이고, 단 한 번만 조사를 한 조사기관이 6개였다. 제일 많이 조사를 한 조사기관은 조사를 16번 실시하였다. 조사기관에 따라 조사방법이 달라지는 것을 알 수 있다. 전화면접 방식만 쓰는 조사기관이 있는가 하면 ARS 방식만 쓰는 조사기관도 있다. 하지만 그런 기관의 조사에서도 유무선 혼용비율에는 변동이 있었다.

Table 2.1. Variables and their distributional characteristics

변수명	변수 설명	분포특성
y	문재인후보 지지율	범위: 0.322–0.482, 평균: 0.410, 표준편차: 0.031
a_1	유선ARS 조사비율	범위: 0–0.497, 6개만 0.4보다 크고 81개는 0.2보다 작거나 같음, 64개가 0
a_2	무선ARS 조사비율	범위: 0–1, 평균: 0.295, 표준편차: 0.389, 52개가 0
a_3	유선면접 조사비율	범위: 0–0.54, 평균: 0.157, 표준편차: 0.156, 35개가 0
a_4	무선면접 조사비율	범위: 0–0.88, 평균: 0.489, 표준편차: 0.334, 20개가 0
org	조사기관	범주형 변수, 총 19개 기관, '리얼미터'에서 조사를 16번으로 최다 실시, 1번만 실시한 기관 수는 6개
time	조사시점	단위: 주, 범위: 1–5, 평균: 2.706, 표준편차: 1.296
size	표본크기	단위: 1,000명, 범위: 1.000–3.077, 평균: 1.293, 표준편차: 0.453
name	호명순서	범주형 변수, '기호순' 46개, '랜덤' 41개

조사가 실시된 시점도 조사결과에 영향을 미칠 수 있다. 조사시점을 나타내는 변수로 조사가 시작된 날짜와 선거가 실시된 5월 9일과의 차이를 7로 나눈 값을 썼다. 다른 변수가 갖는 값의 크기와 너무 다르지 않도록 주 단위로 나타내었다. 조사시점과 반응변수의 산점도와 평활곡선을 참고하여 조사시점을 나타내는 이차항을 고려하기로 하였다. 이차항은 일차항을 단순히 제곱하지 않고 R의 poly 함수를 이용해 일차항과 직교하도록 만들었다. 이렇게 만들어진 두 개의 직교하는 항을 $t1$과 $t2$로 부르기로 한다.

표본크기도 조사결과에 영향을 미칠 수 있는 변수이다. 표본크기는 최소 1,000명부터 최대 3,077명이었으며, 2,000명 이상 조사가 이루어진 경우가 12건 있었다. 다른 변수가 갖는 값 크기와 너무 다르지 않도록 천 명 단위로 나타내었다.

조사를 실시할 때 후보들 이름을 투표지 기호 순서대로 불러주는 경우도 있고 응답자 기억력에 의한 편향을 없애기 위해 매번 랜덤한 순서로 불러주는 경우도 있다. 전체 87개 조사마다 호명순서 방법이 정해져 있는데 이 변수를 회귀모형에 설명변수로 포함시켜 유의성을 갖는지를 보았다. 자료를 살펴본 결과 조사기관별로 호명순서가 고정된 경우가 많았으며, 전체 19곳 조사기관 중에서 두 호명 방법을 같이 쓴 곳은 8곳이었다.

지금까지 설명한 변수들의 분포를 탐색적으로 살펴보았으며 그 특징을 Table 2.1에 요약하였다. 그리고 조사결과에 영향을 미칠 수 있는 변수들로서 조사를 의뢰한 '의뢰기관'과 '응답률'이 있는데, 이 변수들은 회귀모형에 포함시키지 않았다. 그 이유를 다음 절에서 설명하고자 한다.

3. 인과 그래프와 변수선택

본 연구의 목적은 ARS와 면접, 무선과 유선 전화 조사비율을 조합해서 결정되는 조사방법이 조사결과에 미치는 영향을 알아보는 것이다. 즉, 관측 자료를 이용해 조사방법의 효과에 대한 인과연구를 하고자 한다. '조사방법'이라는 처리(treatment) 또는 개입(intervention) 변수의 효과를 추정하고자 할 때 조건화해야 하는 공변량이 있고 조건화하면 안 되는 공변량이 있다. 단순히 반응변수에 대한 예측력 또는 설명력을 높이고자 하는 것이 회귀모형의 목적인 경우 일반적인 수정결정계수나 Akaike information criterion (AIC)와 같은 변수선택기준으로 공변량을 선택하면 되지만, 처리변수의 인과효과(causal effect)를 추정하는 것이 목적인 경우 공변량을 선택할 때 조심해야 한다. 반드시 포함시켜야 하는 것이 있는가 하면 포함시키면 안 되는 것도 있다.

Pearl (2009)은 인과연구에서 처리효과를 편향되지 않게 추정하려면 변수들 사이의 인과관계를 나타내는 인과 그래프를 먼저 그려야 한다고 주장하였다. 인과 그래프에 관한 용어와 주요 정리들을 Green-

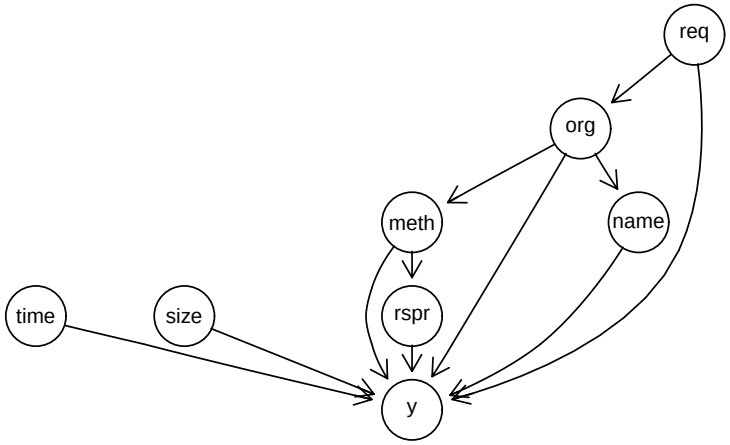

Figure 3.1. A causal graph for the effect of survey method. 'rspr' and 'meth' denote the response rate and the survey method, respectively. See Table 3.1 for the meaning of other abbreviations.

land 등 (1999)이 쉽게 요약하여 설명하였다. Thepepomma와 Kim (2016)은 인과효과 추정을 위한 변수선택 문제가 예측을 위한 변수선택 문제와 어떻게 다른지를 인위적인 자료 예를 들어 설명하였다.

전화조사 자료에 있는 변수들 사이의 인과관계를 Figure 3.1과 같은 인과 그래프로 그려 보았다. 응답을 얻은 회수를 전화가 연결된 총 회수로 나누어 구한 '응답률(변수명 rspr)'은 조사방법에 영향을 받으므로, 조사방법으로부터 조사결과에 이르는 경로는 '조사방법'→'응답률'→'조사결과'라는 경로와 '조사방법'→'조사결과'라는 두 경로가 있다. 만약 응답률을 회귀모형에 포함시켜 조건화하면 조사방법이 조사결과에 미치는 직접효과(direct effect)만 추정할 수 있고, 응답률을 거쳐 조사결과에 미치게 되는 영향인 간접효과(indirect effect)는 '조사방법'→'응답률'→'조사결과' 경로가 차단되어 추정할 수 없게 된다. 직접효과와 간접효과를 합한 것을 전체효과(total effect)라고 하는데, 우리가 추정하고자 하는 것은 조사방법이 조사결과에 미치는 전체효과이므로 응답률은 회귀모형에 포함시키면 안 되는 변수이다.

'조사기관(변수명 org)'은 반응변수인 '조사결과(y)'의 원인이 되는 변수인 동시에 처리변수인 '조사방법(변수명 meth)'의 원인 변수이다. 즉, 조사기관은 중첩변수(confounder)이다. 조사방법으로부터 조사결과에 이르는 경로는 '조사방법'→'조사결과', '조사방법'→'응답률'→'조사결과' 두 경로 외에 '조사방법'←'조사기관'→'조사결과', '조사방법'←'조사기관'←'의뢰기관'→'조사결과', '조사방법'←'조사기관'→'호명순서'→'조사결과'라는 세 개의 뒷문경로(backdoor path)가 있다. 이 세 개의 뒷문경로는 중첩변수인 조사방법을 조건화하면 모두 차단되며, 따라서 조사방법 효과에 대한 비편향추정량을 얻을 수 있다. 조사기관은 조사방법 효과를 제대로 추정하기 위해 반드시 포함시켜야 하는 변수이다.

조사를 의뢰하는 기관인 '의뢰기관'은 조사를 실시하는 조사기관과 다르지만 회귀모형에 포함시키지 않았다. 그 첫 번째 이유는 의뢰기관과 조사기관을 구별할 수 없는 경우가 상당히 있어서이다. 예를 들어 조선일보가 의뢰한 4건의 조사가 모두 하나의 조사기관에서 이루어졌으며, JTBC에서 의뢰한 6건의 조사도 하나의 조사기관에서 전부 이루어졌다. 또 다른 이유는 조사방법으로부터 조사결과에 이르는 '뒷문경로(back-door path)'에서 조사기관을 '차단(block)'하면 의뢰기관 효과도 차단되므로, 처리효과를 추정하고자 하는 것이 목적인 경우 조사기관을 회귀모형에 포함시킨다면 의뢰기관을 굳이 고려하지 않아도 된다. 한편, 지지하는 후보 이름을 불러주는 방법인 '호명순서(변수명 name)'는 조사기관에 따라 달라지는 경향을 확인하였는데 조사기관의 영향을 받는 변수로 보인다.

나머지 변수들인 조사시점(변수명 time)과 표본크기(변수명 size)는 반응변수에 영향을 미치는 변수로 볼 수 있다. 이 인과 그래프를 받아들인다면 조사기관은 반드시 조건화시켜야 하는 (또는 회귀모형에 포함시켜야 하는) 변수이며, 응답률은 포함시키면 안 되는 변수이다. 조사시점과 표본크기, 호명순서는 굳이 포함시키지 않더라도 조사방법의 효과를 편향되지 않게 추정하는 데에는 문제가 없으나 추정 오차를 줄여줄 수 있기 때문에 회귀모형에서 고려하기로 한다.

Figure 3.1로 나타낸 변수들 사이의 인과관계에 대해 동의하지 않을 수 있다. 어떤 경우에는 추가적인 가정이 문제가 되지 않을 수도 있다. 예를 들어 org→size나 org→time이 인과 그래프에 추가되더라도 org를 조건화하면 meth에서 y에 이르는 뒷문경로가 모두 차단되므로 meth의 효과추정에는 문제가 없다. 하지만 또 다른 경우에는 가정이 변경됨에 따라 조건화 또는 모형에 포함시켜야 할 변수가 달라져야 할 수도 있다. 하지만 이런 한계에도 불구하고 인과 그래프를 근거가 부족한 가정을 나타낸 그림이라고 무시하기보다, 통계학자가 주어진 인과 문제를 이해해서 표현하는 도구로, 또한 다른 의견을 가진 사람과 소통하기 위한 도구로 써야 한다고 생각한다.

4. 조사방법이 문재인후보지지율에 미치는 효과

조사방법 효과를 추정하기 위해 조사방법이라는 속성을 나타내는 변수를 정의해야 한다. 주어진 자료에서 조사방법은 a_1, a_2, a_3, a_4의 조합으로 정해지는데, 다음에 설명할 이유로 이 네 개의 변수 대신에 다른 변수를 쓰고자 한다. $a_1 + a_2 + a_3 + a_4 = 1$이기 때문에 선형회귀모형을 쓴다면 이 중 한 변수를 제외시켜야 하는데, 한 변수를 제외시키더라도 나머지 변수의 해석이 명확하지 않다. 예를 들어 a_1은 유선ARS 조사비율인데, 이 변수의 효과가 유의하다고 할 때 그것이 유선전화 효과인지 ARS 효과인지가 구분되지 않기 때문이다. 조사방법을 나타내는 새로운 변수로 무선전화 조사비율을 나타내는

$$\text{mob} = a_2 + a_4$$

와 ARS 조사비율을 나타내는

$$\text{ars} = a_1 + a_2$$

를 쓰기로 한다. 각 변수가 무선 대 유선, ARS 대 면접의 효과를 각각 나타내고, 무선전화 효과 크기가 ARS 조사비율 값이 얼마인가에 따라 달라지는지를 보려면 이 두 변수의 상호작용 항이 유의한가를 확인하면 되므로, a_1, a_2, a_3, a_4 대신 mob과 ars를 조사방법을 나타내는 변수로 쓰기로 한다.

하지만 자료의 한계 때문에 조사방법을 결정하는 변수 중에 ars는 분석에서 제외하기로 하였다. ARS 조사비율을 나타내는 ars는 조사기관을 나타내는 org와 구별이 잘 되지 않는데, ars의 분포를 살펴보면 19곳의 조사기관 중에서 2곳을 제외하고는 모두 ars 값 개수가 하나이다. 예를 들어 일곱 번 조사를 실시한 기관인 '칸타코리아'는 일곱 번 모두 면접 방식으로만 조사해 ars 값이 모두 0이었다. 또한 기관 '리서치뷰'는 5번 모두 ARS 방식으로만 조사해 ars 값이 모두 1이었다. 복수 개의 ars 값을 갖는 두 개의 기관 중에서 6번 조사를 실시한 '조원씨엔아이'는 1개만 0, 나머지 5개는 모두 1이었으며, 16번 실시한 '리얼미터'만 0.45에서 1 사이의 다양한 ars 값을 가진다. 따라서 인과효과를 추정하기 위해 반드시 포함시켜야 하는 중첩변수 org를 모형에 포함시킨다면 ars는 org와 구별이 거의 되지 않으므로 모형에 포함시키는 것이 바람직하지 않다. 만약 ars를 모형에 포함시킨다면 19개 조사기관 중에서 2개 조사기관의 ars 효과만 반영하게 되는 셈이다. ars와 달리 mob은 각 조사기관별로 다양한 값을 갖는다. 자료의 한계 때문에 무선전화조사 비율을 나타내는 mob이 조사방법을 나타내는 유일한 변수가 되었다.

조사방법을 나타내는 연속형 변수 mob의 효과를 추정하기 위해 회귀모형을 쓰기로 한다. 세 종류의 회귀모형을 적용하였는데, 최소제곱법을 쓰는 선형회귀모형(ordinary least squares linear regression

Table 4.1. Least square regression models in the order of AIC

Ordinary least squares regression model	AIC	BIC	Adj. R^2
(1) $y \sim$ mob + $t2$ + size + name + org + mob:name + mob:$t2$	−461.7	−397.6	0.771
(2) $y \sim$ mob + org + name + $t2$ + size + mob:$t2$	−460.5	−398.8	0.766
(3) $y \sim$ mob+$t1$+$t2$+size+name+org+mob:name+mob:$t1$+mob:$t2$	−459.9	−390.8	0.769
(4) $y \sim$ mob + $t1$ + $t2$ + size + name + org + mob:$t1$ + mob:$t2$	−459.5	−392.9	0.767
(5) $y \sim$ mob + org + name + $t2$ + size	−453.3	−394.2	0.744
(6) $y \sim$ mob + org	−426.9	−375.2	0.646

AIC = Akaike information criterion; BIC = Bayesian information criterion.

model)과 임의효과모형 중 하나인 선형임의절편모형(linear random intercept models), 그리고 multivariate adaptive regression splines (MARS) (Friedman, 1991)라고 부르는 회귀모형을 모두 적용해 보았다. MARS 모형은 비선형 항뿐만 아니라 다른 변수와의 상호작용 항까지 중요한 항을 자동적으로 찾아주는 유연한 회귀모형이다. 각 모형은 그에 따른 가정들이 있는데 여러 모형을 동시에 적용해봄으로써 모형이 달라지더라도 효과가 공통적으로 나타나는지 또는 모형에 따라 결과가 달라지는지를 보고자 하였다.

4.1. 선형회귀모형

먼저 최소제곱 선형회귀모형을 적용해 보았다. 단계적 선택(stepwise selection) 방법을 적용해서 잠정적으로 모형을 선택하고 대안적으로 생각해볼 수 있는 모형들과 비교해 보았다. 우리의 주된 관심인 처리변수 mob과 꼭 포함되어야 하는 변수인 org만 있는 제일 단순한 모형

$$y \sim \text{mob} + \text{org}$$

을 출발 모형으로 지정하고, mob의 일차항뿐만 아니라 이차항, 그리고 다른 변수와의 상호작용 항, time의 일차항과 이차항, size, name을 모두 포함시킨 모형

$$y \sim \text{mob} + I(\text{mob}^\wedge 2) + t1 + t2 + \text{size} + \text{name} + \text{org} + \text{mob:}t1 + \text{mob:}t2 + \text{mob:size} + \text{mob:name} + \text{mob:org}$$

을 최대 허용 모형으로 지정해서 단계적 선택 방법으로 모형을 선택했다. 그 결과 얻어지는 모형이 Table 4.1의 첫 번째 모형이다 (통계적 모형을 R (R Core Team, 2016)의 모형식(model formula) 형식으로 표현하였다.).

단계적 선택에서 $t1$은 선택되지 않았는데 $t1$의 포함여부에 따라 모형선택기준인 AIC 값이 어떻게 달라지는가를 보았다. 그리고 유의성이 높지 않은 상호작용 항인 mob:name 변수가 포함되어야 하는가도 살펴보았다. 비교해본 모형들과 AIC 값을 Table 4.1에 AIC 값 순서에 따라 정리하였으며 Bayesian information criterion (BIC) 값과 수정결정계수 값도 같이 보고하였다. 단계적 선택 방법으로 선택된 모형의 AIC 값이 그 중에서 제일 작았다. AIC 뿐만 아니라 수정결정계수 기준에서도 최적인 첫 번째 모형을 적합시킨 결과를 Table 4.2에 요약하였다.

mob의 효과는 mob의 상호작용 항들이 유의하기 때문에 일차항 계수만으로 해석하면 안 된다. 예를 들어 $t2$와 size 값이 각각 중앙값인 −0.009393, 1.023이고, name 값이 최빈값인 '기호순'일때 mob의 계수의 부호는

$$-0.10815 + (-0.30246) \times (-0.009393) \doteq -0.1054$$

Table 4.2. Estimated regression coefficients of model (1) in Table 4.1

Coefficients	Estimate	Standard error	t-value	p-value
(Intercept)	0.60943	0.06641	9.18	3.7e−13 ***
mob	−0.10815	0.06474	−1.67	0.09988 .
$t2$	0.17446	0.09377	1.86	0.06756 .
size	−0.00541	0.00442	−1.22	0.22554
name랜덤	−0.06186	0.02959	−2.09	0.04066 *
org리얼미터	−0.06301	0.02225	−2.83	0.00622 **
org마크로밀엠브레인	−0.13368	0.02716	−4.92	6.7e−06 ***
org여의도연구원	−0.15163	0.03961	−3.83	0.00030 ***
org조원씨엔아이	−0.12781	0.03129	−4.08	0.00013 ***
org㈜디오피니언	−0.05776	0.03080	−1.88	0.06548 .
org㈜리서치뷰	−0.05147	0.01683	−3.06	0.00328 **
org㈜리서치앤리서치	−0.12124	0.02105	−5.76	2.8e−07 ***
org㈜리서치플러스	−0.14670	0.03427	−4.28	6.6e−05 ***
org㈜메트릭스코퍼레이션	−0.10975	0.02553	−4.30	6.2e−05 ***
org㈜알앤써치	−0.05919	0.01770	−3.34	0.00141 **
org㈜에스티아이	−0.03139	0.02128	−1.48	0.14520
org㈜에이스리서치	−0.10401	0.02505	−4.15	0.00010 ***
org㈜한국리서치	−0.11693	0.02063	−5.67	4.0e−07 ***
org중앙일보조사연구팀	−0.12481	0.02693	−4.63	1.9e−05 ***
org칸타코리아(칸타퍼블릭)	−0.16596	0.03269	−5.08	3.8e−06 ***
org타임리서치	−0.06257	0.02176	−2.88	0.00552 **
org한국갤럽조사연구소	−0.11038	0.02302	−4.80	1.1e−05 ***
org한국사회여론연구소(KSOI)	−0.08878	0.02183	−4.07	0.00014 ***
mob:name랜덤	0.06380	0.04165	1.53	0.13061
mob:$t2$	−0.30246	0.11701	−2.58	0.01211 *

로서 음수이지만, $t2$ 값이 최솟값 −0.1354, name 값이 '랜덤'이면 mob 계수의 부호는

$$-0.10815 + 0.0638 + (-0.30246) \times (-0.1354) \doteq -0.0034$$

로서 0에 가까워진다. 따라서 mob 효과는 단순하지 않다. mob의 주효과 항만 있는 모형 (5)와 (6)에서 mob의 회귀계수에 대한 p-값이 각각 0.41, 0.74로서 mob 효과는 유의하지 않았다.

자료가 전부 87개이지만 한 조사기관에서 여러 조사를 실시했기 때문에 서로 다른 조사기관 수는 19개이다. 19개 조사기관에 18개 가변수를 할당해서 계수를 추정하는 대신에 조사기관을 나타내는 org을 임의절편(random intercept)으로 간주하는 선형혼합효과모형(linear mixed effect models)을 적용해 보았다. 이렇게 하면 추정해야 하는 계수 수를 줄일 수 있다. 모형에 포함시키는 설명변수를 달리 하면서 적용해본 결과를 Table 4.3에 정리하였다. AIC 기준에서 최적인 모형 (a)는 고정효과 mob만 있는 제일 단순한 모형으로서 mob의 계수는 0.0966이고 표준오차는 0.0281이었다. 모형 (a)에 따르면 mob 값이 커질 때, 즉 무선전화 조사비율이 높아질 때 문재인후보지지율이 높게 조사된다고 할 수 있다. 두 선형회귀모형인 최소제곱회귀모형과 선형임의절편모형에서 mob의 효과에 대한 해석은 같지 않았다. mob에 대한 해석은 MARS 모형에서 다시 다루기로 한다.

4.2. Multivariate adaptive regression splines 모형

마지막으로 자료에 MARS 모형을 적용해 보았다. 앞의 두 모형은 주어진 항들의 선형함수로 자료가 생

Table 4.3. Linear random intercept models in the order of AIC

선형임의절편모형(linear random intercept model)	AIC	BIC
(a) $y \sim$ mob	−392.5	−382.8
(b) $y \sim$ mob + $t2$ + size + name	−386.1	−369.2
(c) $y \sim$ mob + $t2$ + size + name + mob:$t2$ + mob:name	−379.8	−358.4
(d) $y \sim$ mob + $t1$ + $t2$ + size + name	−379.3	−360.1
(e) $y \sim$ mob + $t1$ + $t2$ + size + name + mob:$t1$ + mob:$t2$ + mob:name	−368.3	−342.4

AIC = Akaike information criterion; BIC = Bayesian information criterion.

Table 4.4. Multivariate adaptive regression spline models in the order of GCV

Multivariate adaptive regression splines model	GCV($\times 10^3$)
(i) $y \sim$ mob + $t1$ + $t2$ + size + name + org, degree = 1	0.3031
(ii) $y \sim$ mob + time + size + name + org, degree = 1	0.3164
(iii) $y \sim$ mob + $t1$ + $t2$ + size + name + org, degree = 2	0.3556
(iv) $y \sim$ mob + time + size+ $\sim$ name + org, degree = 2	0.3576
(v) $y \sim$ mob + org, degree = 1	0.3789
(vi) $y \sim$ mob + org, degree = 2	0.4116

GCV = generalized cross validation.

성되었다는 가정을 한다. 즉 각 항이 갖는 값의 전체 범위에서 일정한 기울기로 증가하거나 감소한다는 가정을 하고 있는데, 이런 강한 가정을 하지 않고 자료에 존재하는 관계를 좀 더 유연하게 찾을 수 있게 MARS 모형을 적용해 보았다. R의 earth package (Milborrow, 2017b)를 이용하였으며 Table 4.4에 있는 모형 식과 같이 선언해주면 모형에 선언된 각 변수들이 갖는 값 범위에서 부분적으로 다른 기울기를 갖는 것을 허용하며, 'degree = 2'라고 선언해주면 필요한 2차 상호작용 항을 자동적으로 찾아준다.

MARS 방법은 훈련자료에 적합한 모형을 찾은 다음 훈련자료가 아닌 독립적인 자료에 더 적합한 모형을 찾기 위해 가지치기(pruning)를 한다. 이 때 generalized cross validation (GCV)라고 부르는 기준을 이용하는데, GCV는 추정 모형에 불필요한 항이 추가되는 것에 대해 불이익을 줘서 훈련자료만 잘 설명하는 현상을 막고자 하는 장치로서 선형모형의 AIC나 수정결정계수와 같은 역할을 하는 모형선택 기준이다. 모형에 포함되는 설명변수와 모형의 차수(degree)를 변화시켜가면서 다양한 모형을 적합시켜 보았는데, 그 결과를 GCV 값 순서대로 Table 4.4에 정리하였다.

GCV 기준에서 최적인 모형 (i)을 추정된 항과 계수로 표현하면 다음 식과 같다.

$$\begin{aligned}\hat{y} = {} & 0.4058 - 0.8587(\text{mob} - 0.84)^+ + 1.3191(\text{mob} - 0.86)^+ - 0.7618(t1 - 0.1077)^+ \\ & - 0.0682(0.0601 - t1)^+ + 0.3013(t2 - 0.0914)^+ + 0.0447\, I(\text{org} = \text{리얼미터}) \\ & + 0.0796\, I(\text{org} = \text{디오피니언}) - 0.0282\, I(\text{org} = \text{칸타코리아}) \\ & + 0.0346\, I(\text{org} = \text{한국사회여론연구소}).\end{aligned}$$

위 식에서 $(x-a)^+$는 $x > a$일 때 $x - a$이고 아니면 0인 함수이며, $I(A)$는 A가 참일 때 1이고 거짓이면 0인 지시함수(indicator function)이다. 위 식에서 mob의 효과를 쉽게 알아보기 위해 Figure 4.1과 같은 그림을 그릴 수 있는데, plotmo package (Milborrow, 2017a)를 이용하면 된다. 이 그림을 그릴 때 mob을 제외한 나머지 변수들의 값을 자동적으로 고정시키는데 정량적 변수인 $t1, t2$의 값은 중앙값, 범주형 변수의 값은 첫 번째 범주로 고정시킨다. Figure 4.1과 식에 의하면 mob 값이 최솟값인 0.46에서 0.84까지는 효과가 없다가 0.84에서 0.86까지 $\hat{y}$를 작아지게 한다. 0.86보다 커지면 $\hat{y}$를 다시 커지게 하는 효과가 있다. 그런데 전체 87개 mob의 값 중에서 0.84보다 큰 것은 30개인데, 이 중에서 0.9 이하인

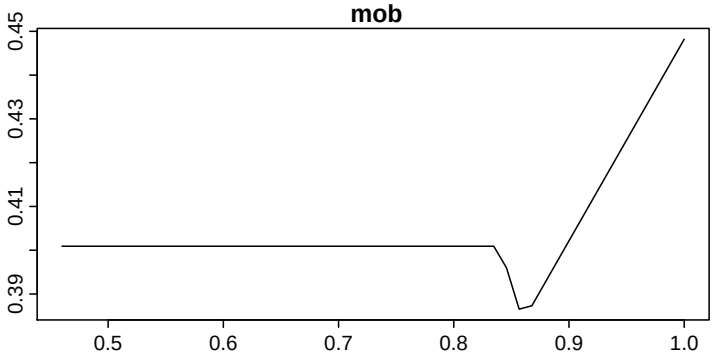

Figure 4.1. The effect of 'mob', the proportion of cell phone survey, estimated by the optimal multivariate adaptive regression splines model.

것이 18개, 0.9보다 크고 1.0보다 작은 것은 하나도 없고 1.0인 것이 12개이다. 따라서 mob이 0.84에서 0.9 사이일 때 보이는 크지 않은 변화를 무시한다면, 무선조사비율이 1.0일 때 그렇지 않은 경우에 비해 문재인후보지지율이 높아진다고 대략적으로 말할 수 있다. (유난히 낮은 문재인후보지지율을 보이는 자료 2개를 제외하고 적합시키면 0.84에서 0.9 사이에서 보이는 mob과 $\hat{y}$ 사이의 다소 복잡한 관계는 변화가 없는 수평 관계로 단순화된다. 그러므로 이 범위에서 보이는 다소 복잡한 관계에 큰 의미를 두지 않아도 될 것으로 보인다.) 모형 (i)은 mob의 효과가 특정한 범위에서만 나타난다는 점과 상호작용 항은 선택되지 않았다는 점에서 앞에서 찾은 두 선형회귀모형과 사뭇 달라 보인다.

선형회귀모형은 mob이 갖는 값의 전체 범위 내에서 일정한 기울기를 갖는다는 강한 가정을 하고 있는 반면, MARS 모형은 값의 부분적 범위에서 다른 기울기를 허용하므로 실제에 가까운 관계를 탐지해낼 가능성이 높다고 할 수 있다. MARS 모형으로 효과 크기를 계산해 보았다. 추정모형 (i)이 참이라고 할 때, 무선조사비율(mob)이 1이면 0.84 이하인 경우에 비해 (또는 0.9 이하인 경우에 비해)

$$-0.8587 \times (1 - 0.84) + 1.3191 \times (1 - 0.86) \doteq 0.0473.$$

즉, 약 4.7% 포인트만큼 문재인후보지지율이 높아진다고 할 수 있다.

조사시기와 조사기관도 중요한 변수이지만 해석을 하지 않았다. 왜냐하면 유의한 관련성을 인과성으로 결론내릴 수 없고 이들 변수의 효과 규명에 관심을 갖지도 않았기 때문이다. 예를 들어 조사기관의 경우, 모형에서 의뢰기관을 포함시키지 않았으므로 조사기관에서 의뢰기관을 경유하여 조사결과에 이르는 뒷문이 차단되지 않아 조사기관의 효과가 제대로 추정되지 않는다.

5. 조사방법이 조사정확도에 미치는 효과

앞 절에서 조사방법에 따라 문재인후보지지율이라는 조사결과가 어떻게 달라지는가에 대해 살펴보았다. 하지만 보다 관심 있는 것은 조사방법에 따라 조사정확도가 어떻게 달라지는가 하는 것이다. 이를 살펴보기 위해서는 조사의 '정확도'에 대한 정량적 정의가 있어야 한다. 조사를 통해 얻은 조사지지율과 실제득표율(41.1%)의 차이를 예측오차로 정의하면 이 예측오차를 줄이는 조사방법이 조사의 정확도를 높이는 방법이라고 간주할 수 있다. 예측오차를

$$y = |\text{문재인후보지지율} - 0.411|$$

와 같이 정의한다. 앞 절에서는 문재인후보지지율을 반응변수 y로 두었으나, 이 절에서는 예측오차를 y로 두고, 앞 절에서와 같이 전체 87개 자료를 이용해 $E[y]$에 관한 모형을 구축해서 조사의 예측오차 또는 정확도에 미치는 조사방법의 효과를 추정해 보았다.

최소제곱 선형회귀모형 중에서 AIC 기준 최적모형은 R의 통계적 모형을 위한 식으로 표현했을 때

$$y \sim \text{mob} + \text{name} + t2 + \text{org} + \text{mob}{:}t2 + \text{mob:name}$$

으로서, mob의 효과는 4절에서와 같이 $t2$와 name의 값에 따라 달라진다. org를 임의절편으로 둔 선형 임의절편모형 중에서 AIC 기준 최적모형은 고정효과 mob만 있는 단순한 모형이었으며 mob의 계수의 부호는 양수이었지만 p-값이 0.43으로서 유의하지 않았다.

GCV를 최소로 하는 최적 MARS 모형은 3차 상호작용까지 허용하는 모형이었으며 모형의 추정 식은 다음과 같았다

$$\begin{aligned}\hat{y} = {} & 0.0175 + 0.0918(\text{mob} - 0.818)^{+} + 0.0455\, I(\text{org} = \text{디오피니언}) \\ & + 0.0205\, I(\text{org} = \text{칸타코리아}) + 0.1894\, t1\, I(\text{org} = \text{칸타코리아}) \\ & + 0.0682\, t2\, I(\text{name} = \text{random}) + 0.0315\, I(\text{name} = \text{random})\, I(\text{org} = \text{리서치앤리서치}) \\ & - 0.2268\, t2\, I(\text{name} = \text{random})\, I(\text{org} = \text{리얼미터}).\end{aligned}$$

위 모형에 3차 상호작용 항이 있지만 mob에 관한 항은 주효과 항만 있다. 이 모형을 이용하여 무선조사비율이 예측오차에 미치는 영향을 다음과 같이 해석할 수 있다. 무선조사비율이 약 81.8%보다 높아지면 예측오차가 증가하기 시작하며, 무선조사비율이 100%가 되면 81.8% 이하인 경우에 비해 예측오차가 $0.0918 \times (1 - 0.818) \doteq 0.0167$로서 1.7% 포인트 정도 높아진다고 할 수 있다. 이 때 관측된 mob 값 중에서 0.9보다 크고 1.0보다 작은 값이 없으므로 이 구간에서 예측오차의 변화에 대해 제대로 판단할 근거가 부족하다는 점에 주의해야겠다.

정확도 또는 예측오차에 관한 분석 결과를 간략하게만 보고한 이유가 있다. 먼저 예측오차에 관한 모형이 자료를 잘 설명하지 못한다는 점이다. 문재인후보지지율에 관한 최적 MARS 모형의 경우 $R^2 = \text{Corr}^2(y, \hat{y}) = 0.809$이었으나 예측오차에 관한 최적 MARS 모형의 경우 0.546이었다. 하지만 보다 중요한 이유는 예측오차의 정의에 대한 타당성 문제이다. 특정한 조사시점에서 얻어진 조사결과는 그 시점의 실제지지율에 대한 추정값이지 최종득표율에 대한 추정값이 아니다. 그런데 '조사시점의 문재인후보지지율과 최종득표율의 차이'라고 예측오차를 정의했으므로, 실제지지율이 시간에 따라 계속 변화하였다면 조사가 정확히 이루어졌다 하더라도 예측오차는 작아지지 않는다. 따라서 문재인 후보에 관한 실제지지율이 조사기간 동안 변하지 않고 일정했다는 가정이 없으면 예측오차의 정의에 대한 타당성도 확보되지 못한다.

예측오차를 나타내는 측도로 Martin 등 (2005)이 제안하고 Kim과 Hwang (2014)이 지방선거 전화조사 분석에 사용했던 로그 오즈비(odds ratio) 측도가 있다. 상위 두 후보 지지율을 이용하는 이 측도는 부동층의 크기 변화에 영향을 덜 받는다는 장점을 갖지만, 상위 두 후보 실제지지율 비가 조사기간 동안 일정하다는 가정이 필요하다. 조사기간 동안 2위 후보 지지율 변화가 심했던 이번 대선 자료의 특성 때문에 이 측도도 한계를 가지므로 따로 적용해보지 않았다.

6. 요약과 결론

중앙선거여론조사심의위원회 홈페이지에 등록된 87개의 19대 대선 전화조사 자료를 이용하여 유선/무선, ARS/면접이라는 '조사방법'이 후보자의 지지율이나 예측정확도라는 '조사결과'에 미치는 영향을

알아보고자 하였다. 자료의 한계로 ARS와 면접 방식의 효과에 대해서는 알아보지 못했고 무선조사비율의 효과에 대해서만 알아보았다.

문재인후보지지율과 예측오차라는 두 조사결과에 대해 각각 살펴보았다. 실용적인 관심은 조사의 정확도에 있지만 '정확도' 또는 예측오차에 대한 정의의 타당성 때문에 이에 대한 분석 결과에 큰 비중을 두지 않았다. 그러나 정확도가 아닌 특정후보 지지율이 조사방법에 따라 어떻게 달라지는가를 알아보는 것도 중요하다. 왜냐하면 조사방법에 따라 조사결과에 반영되는 정치성향의 차이를 알아볼 수 있기 때문이다. 한편, 특정후보 지지율은 조사 때마다 달라지는 부동층의 크기에 영향을 받는다는 문제점이 있다. 특정후보 지지율 대신에 상위 두 후보의 지지율 비를 고려하는 Martin 등 (2005)의 로그 오즈비 측도를 고려해 볼 수 있다.

본 연구는 19대 대선 선거여론조사의 특성에 관한 단순한 사례연구가 아니다. 실험이 아닌 관측 자료를 이용하는 인과연구(causal study)에서 '변수선택'에 주의해야 한다는 점을 선거여론조사 자료를 통해 강조하고자 하였다. 확률화실험(randomized experiment)이 아닌 관측연구에서 얻어진 자료로부터 원인 변수의 효과를 알아보고자 할 때, 모형에 원인 변수와 함께 반드시 포함시켜야 할 변수와 포함시키면 안 되는 변수가 있다는 점을 간과하기 쉽다. 본 연구에서 '조사기관'은 반드시 포함시켜야 하는 변수이며 '응답률'은 포함시키면 안 되는 변수임을 역설하였다.

효과의 정확한 추정을 위해서 변수선택도 중요하지만 적용하는 모형의 종류도 중요하다. 최소제곱선형모형과 선형임의절편모형과 함께 MARS 모형을 적용해 비교해 보았는데 다른 모형도 적용할 수 있을 것이다. 보다 유연하면서도 과대적합(overfitting)을 방지하는 장치를 갖고 있으며 추정된 모형을 식으로 표현해주는 MARS 방법 결과를 이용해 무선전화 조사비율의 효과에 대한 해석을 하였다. 조사기관, 조사시점, 표본크기, 호명순서 등의 조건이 같더라도 무선전화로만 100% 조사를 실시하게 되면 유선전화와 혼용해서 조사한 경우에 비해 약 4.7 퍼센트 포인트만큼 문재인 후보 지지율이 높아지는 것으로 추정되었다. 전체 조사를 무선전화로만 조사를 실시하면 '예측오차'도 커지게 된다. 예측오차의 정의에 대한 타당성 문제로 무선전화의 적정 혼용비율을 제시하기에 조심스럽지만, 약 80% 이하이기만 하면 예측오차에 별 차이가 없는 것으로 보이며, 전체 조사를 무선전화로만 실시하면 예측오차가 커지는 것으로 보인다. 무선전화 조사비율 효과에 대한 본 연구 결과는 19대 대선 전화조사 자료에서 추정된 것이며, ARS 방식 조사비율 효과를 포함한 전반적인 조사방법의 효과를 알아보기 위해 더 많은 자료들에 대한 연구가 필요하다.

끝으로 전화조사 현장 경험이 없는 저자에게 유용한 정보와 건설적 심사의견을 주신 심사위원께 감사를 드린다.

References

Friedman, J. H. (1991). Multivariate adaptive regression splines, *Annals of Statistics*, **19**, 1–141.

Greenland S., Pearl J., and Robins, J. M. (1999). Causal diagrams for epidemiologic research, *Epidemiology*, **10**, 37–48.

Huh, M. H. and Kim, Y. W. (2008). RDD sample versus directory-based sample for telephone surveys: the case of 2007 presidential election forecasting in Korea, *Survey Research*, **9**, 55–69.

Jang, D. H. and Cho, S. K. (2013). The validity of mobile RDD survey methodology: case study of poll data for 2012 presidential election, *Survey Research*, **14**, 19–47.

Kang, H. C., Han, S. T., Kim, J. Y., Jung, Y. C., and Huh, M. H. (2008). Random digit dialing telephone survey and major findings, *Survey Research*, **9**, 1–22.

Kim, J. Y. and Woo J. Y. (2012). Cell phones and political polls in Korea, *Journal of the Korean Association of Party Studies*, **11**, 225–246.

Kim, Y. and Hwang, D. (2014). Forecasting error and bias of telephone survey for 2014 local election, *Survey Research*, **15**, 1–32.

Lee, K., Lee, H., and Hyun, K. (2012). A study on mixed-mode survey which combine the landline and mobile telephone interviews: the case of special election for the mayor of Seoul, *Survey Research*, **13**, 135–158.

Martin, E. A., Traugott, M. W., and Kennedy, C. (2005). A review and proposal for a new measure of poll accuracy, *Public Opinion Quarterly*, **69**, 342–369.

Milborrow, S. (2017a). plotmo: Plot a Model's Response and Residuals. R package version 3.3.4. https://CR AN.R-project.org/package=plotmo

Milborrow, S. (2017b). Derived from mda:mars by Trevor Hastie and Rob Tibshirani. Uses Alan Miller's Fortran utilities with Thomas Lumley's leaps wrapper. earth: Multivariate Adaptive Regression Splines. R package version 4.5.1. https://CRAN.R-project.org/package=earth

Pearl, J. (1995). Causal diagrams for empirical research, *Biometrika*, **82**, 669–688.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*, Cambridge University Press, Cambridge.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Thepepomma, S. and Kim, J. (2016). Covariate selection criteria for controlling confounding bias in a causal study, *The Korean Journal of Applied Statistics*, **29**, 849–858.

19대 대선 전화조사에서 조사방법 효과에 대한 인과연구

김지현[a,1] · 정효재[a]

[a]숭실대학교 정보통계보험수리학과

(2017년 9월 28일 접수, 2017년 11월 16일 수정, 2017년 12월 7일 채택)

요 약

전화를 이용한 19대 대선 선거예측조사에서 ARS 조사비율과 무선전화 조사비율을 달리함에 따라 조사결과가 어떻게 달라지는가를 보았다. 조사방법이 조사결과에 미치는 효과를 추정하는 인과연구를 시도하였으며, 이를 위해 변수들 사이의 인과관계를 가정하는 인과 그래프를 그린 다음 모형에 포함시켜야 할 변수와 포함시키면 안 되는 변수를 판단하였다. 조사를 실시한 조사기관은 중첩변수로서 모형에 포함시켜야 하는 변수이며 응답률은 모형에 포함시키면 안 되는 변수임을 설명하였다. ARS 조사비율의 효과는 자료 한계 때문에 추정할 수 없었으며, 무선전화 조사비율이 약 90%를 넘지 않으면 효과에 별 차이가 없으나 전체 조사를 무선전화로만 실시하면 문재인후보지지율이 높아진다.

주요용어: 인과 그래프, 중첩, MARS, 전화조사

[1]교신저자: (06978) 서울시 동작구 상도로 369, 숭실대학교 정보통계보험수리학과. E-mail: jxk61@ssu.ac.kr