碩 士 學 位 論 文

응답유보층 보정을 통한 선거결과 예측 방법론 비교 연구

高 麗 大 學 校　大 學 院

統 計 學 科

金 廷 炅

2023年 2月

朴 敗 奎 敎 授 指 導

碩 士 學 位 論 文

응답유보층 보정을 통한 선거결과 예측 방법론 비교 연구

이 論文을 統計學 碩士學位 論文으로 提出함

2022年 10月

高 麗 大 學 校 大 學 院

統 計 學 科

金 廷 炅 (印)

金廷炅의 統計學 碩士學位論文
審査를 完了함

2022年　12月

委 員 長　박　민　규　(印)

委　員　신　승　준　(印)

委　員　김　경　희　(印)

응답유보층 보정을 통한
선거결과 예측 방법론 비교 연구

김 정 경

통 계 학 과

지도교수 : 박 민 규

초록

선거기간 중 지지율은 유권자와 각 정당에 영향을 미치는 가장 중요한 지표 중 하나로, 지지율을 파악하기 위해 선거여론조사를 실시하게 된다.
선거여론조사가 처음 실시, 공표된 이래로 정확성에 대한 논란은 매번 끊이지 않았고 무응답 및 조사거절로 인한 무응답오차의 발생은 여전히 해결할 수 없는 문제로 남아있다. 이를 해결하고자 여론조사 관련 기관과 학계에서는 여론조사의 결과 및 선거예측의 정확성 제고를 위해 지속적으로 연구해오고 있다.

본 연구에서는 선거여론조사 자료를 사용하여 '지지후보가 없다 & 모름/무응답'으로 응답한 응답유보층을 분류함으로써 득표율 예측의 정확도를 높이는 데 기여할 수 있는지 살펴보고, 분류 방법에 따라 결과가 어떻게 달라지는지 살펴보았다. 분류의 방법으로는 다항 로지스틱 회귀, 의사결정나무와 랜덤 포레스트 방법을 고려하였다.

상수항을 포함하지 않는 다항 로지스틱 회귀모형을 이용하여 응답유보층을 보정하는 것이 득표율 예측에 도움이 된다는 결과를 얻을 수 있었다.

주요 용어: 지지율, 선거예측, 응답유보층 보정, 분류모형

A comparative study of methodologies in prediction of election result by adjusting non-response

by Jungkyung Kim

Department of Statistics

under the supervision of Professor Mingue Park

Abstract

During the election, the approval rating is one of the most important indicators that affect voters and political parties. Election polls are conducted to provide an approval rating.

Since the election polls were first conducted and published, the controversy over accuracy has never ceased. The occurrence of non-response errors due to non-response and refusal to investigate remains an unsolvable problem. To solve this problem, polling agencies and academia have been continuously conducting research to improve the accuracy of poll results and election predictions.

In this study, i examined whether it can contribute to increase the accuracy of the prediction of the rate of votes by classifying the respondents who answered "There is no support candidate" & "I do not know / no response" using the election poll data and examined how the results vary depending on the classification method. Multinomial logistic regression, Decision tree and Random forest method were considered as methods of classification of non-response.

It was found that adjusting the non-response using a multinomial logistic regression model without intercept is helpful in predicting the rate of votes.

Keywords: Approval rating, Election prediction,
Adjusting non-response, Classification model

목 차

표 목차

그림 목차

제 1 장

서론

선거기간 중 지지율은 유권자와 각 정당에 영향을 미친다. 유권자는 자신이 지지하는 후보가 어느 정도의 지지를 받는지 알기 위해 지지율을 확인하고, 각 정당은 자신들이 내세운 후보자에 대한 지지율을 파악해 선거 전략을 세우거나 변경하기도 한다. (류제복, 2019)

이처럼 지지율은 가장 중요한 지표 중 하나로 쓰이게 되는데, 언론사와 여론조사 전문기관들은 지지율을 제공하기 위해 선거여론조사를 실시한다. (류제복, 2019)

선거여론조사의 정확성은 포함오차(coverage error : 표집틀과 모집단과의 차이로 발생하는 오차), 무응답오차(non-response error : 추출된 개체로부터 응답을 얻지 못해 발생하는 오차)와 측정오차(measurement error : 실제 관측해야 하는 참값을 측정하지 못해 발생하는 오차) 등 여러 요인에 영향을 받는다.

우리나라에서는 1987년 제 13대 대통령선거에서 처음으로 선거여론조사를 실시, 공표했는데 그 이후 진행되는 매 선거여론조사마다 정확성에 대한 논란은 끊이지 않았다. (류제복, 2019)

우리나라에서 수행되는 대부분의 여론조사는 일반적으로 전화조사를 통해 이루어지고 있으며, 가능한 한 전화조사 대상 유권자를 넓힘으로써 전화조사에서 발생하는 포함오차를 줄이는 방향으로 발전해 왔다. (곽은선·김영원, 2022)

초기 전화조사에서는 전화번호부를 표본추출틀로 이용한 조사가 이루어졌지만

미등재 가구의 누락에 따른 편향의 문제가 있었다. 이를 해결하기 위해 RDD(random digit dialing)방식이 도입되었다. 하지만 휴대전화 RDD의 경우 사용자의 거주지역을 파악할 수 없다는 단점이 있어 지방선거나 국회의원선거에서는 집전화에 의존하는 전화조사를 실시할 수밖에 없었고, 이로 인해 발생한 신뢰성 저하가 심각한 문제로 대두되었다. (곽은선·김영원, 2022)

이 문제점을 해결하기 위해 선거여론조사에서 휴대전화 가상번호를 사용할 수 있도록 2017년 2월 공직선거법이 개정되었고, 조사기관은 통신 3사로부터 휴대전화 가상번호를 제공받아 활용할 수 있게 되었다. 휴대전화 가상번호를 활용하면 거주 지역, 성별, 연령에 대한 정보를 함께 제공받을 수 있기 때문에 접촉률과 응답률을 높인다면 층화확률표본추출에 가까운 전화조사를 구현할 수 있는 조사환경을 마련하게 되었다. (곽은선·김영원, 2022)

그러나 휴대전화 가상번호를 활용하는 조사방식을 사용하더라도, 무응답 및 조사거절로 인한 무응답오차의 발생은 여전히 해결할 수 없는 문제로 남아있다.
이 문제는 선거여론조사 진행 시 응답유보층으로 인해 발생하게 된다. 응답유보층이란 '지지후보가 없다.'로 응답한 응답자와 '모름 또는 무응답'으로 응답한 응답자 층을 의미한다. 선거여론조사에는 상당 비율의 응답유보층이 존재하며, 응답유보층은 실제 투표에서 특정 후보에게 투표할 가능성이 있다. 따라서 응답유보층의 정보를 활용하여 실제로 어떤 후보에게 투표할 확률이 높은지를 파악한다면 득표율 예측의 정확도를 높일 수 있을 것이다.

본 연구에서는 응답유보층을 분류함으로써 득표율 예측의 정확도를 높이는 데 기여할 수 있는지를 살펴보고자 한다. 응답유보층 분류의 방법으로는 다항 로지스틱 회귀(multinomial logistic regression), 의사결정나무(decision tree)와 랜덤 포레스트(random forest) 방법을 고려하였고, 각 방법에 따라 결과가 어떻게 달라지는지를 살펴보고자 한다.

본 논문의 구성은 다음과 같다. 2장에서 선행 논문 2편을 요약하고, 3장에서 본 연구에서 사용할 NBS 전화조사 자료에 대해 소개하고 득표율 예측 및 방법론에 따른 결과를 비교해본다. 마지막으로 4장에서 결론을 끝으로 마무리한다.

제 2 장

논문 요약

다음 장에서 진행할 응답유보층을 보정한 득표율을 예측하기 전에 이에 대한 선행 논문 두 편을 제시한다. 두 논문은 김정훈·김지연·권은혜·김혁·강현철 (2017). 여론조사를 통한 선거예측에서 무응답층 보정의 효과와 곽은선·김영원 (2022). 현행 선거여론조사 방법의 정확성과 선거결과 예측 가능성이며, 응답유보층 보정 절차와 보정 결과에 중점을 두고 요약해보고자 한다.

2.1 여론조사를 통한 선거예측에서 무응답층 보정의 효과

위 논문에서는 여론조사 전문기관에서 2016년 4월 13일 제 20대 국회의원선거를 앞두고 선거일 7~10일 전에 실시된 여론조사 결과를 가지고 연구를 진행하였다. 여론조사 결과 중 20% 이상의 무응답층이 있고 실제 선거에서 1, 2위의 득표율 차이가 크지 않은 선거구를 선정해 10개에 대한 결과를 가지고 연구를 진행하였다. 무응답층을 예측하기 위하여 사용된 설명변수는 성별, 연령, 지역, 직업, 지지정당, 국정운영 평가, 투표의향 등이다.

(1) 무응답층 보정의 절차

1) 각 설명변수들에 대하여 빈도가 10% 미만인 범주는 지지후보의 패턴이 유사한 범주와 병합한다.

2) 데이터를 다음과 같이 3개의 집합으로 분리한다. D_1 : 여론조사에서 주요 후보(대략 10% 이상의 지지율을 가지는 후보)를 지지한 집합, D_2 : 여론조사에서 군소 후보(대략 10% 미만의 지지율을 가지는 후보)를 지지한 집합, D_3 : 지지후보를 밝히지 않은 집합 (무응답층)

3) 집합 D_1을 이용하여 무응답 대체에 대한 통계적 분류모형을 구축하고, 이를 집합 D_3에 적용하여 무응답층의 각 개체에 대하여 지지후보를 예측한다.

4) 집합 D_1과 집합 D_2에서 응답된 지지후보와 집합 D_3에 대하여 예측된 지지후보를 이용하여 득표율을 예측한다. 이때 성, 연령대, 지역에 의해 산출된 통상적인 가중치를 부여한다.

5) 단계 3)과 4)를 100번 반복하여 득표율의 평균을 계산한다.

(2) 결과

위 논문에서는 각 선거구별 1, 2위의 최종 득표율 격차와 무응답층의 보정을 통한 1, 2위의 예측 득표율 격차를 비교하여 무응답층의 보정이 득표율 예측의 정확도를 높이는 데 기여할 수 있는지 연구하였다.

10개 선거구의 여론조사 결과, 단순보정 결과(무응답층을 응답층의 비율대로 보정한 결과)와 세 가지 분류모형을 적용한 결과를 제시하였으며, 분류모형으로 로지스틱 회귀 모형, 선형판별함수, 성향점수를 이용하였다.

무응답층을 보정하여 득표율을 예측한 결과, 방법에 따라 다소의 차이는 있으나 어떤 방법을 사용하든 적어도 단순 보정의 결과보다는 더 좋은 예측 결과를 얻을 수 있었고, 세 가지 방법 중 선형판별함수 방법이 전체적으로 보다 우수한 결과를 나타내었다.

2.2 현행 선거여론조사 방법의 정확성과 선거결과 예측 가능성

위 논문에서는 방송협회와 지상파 3사(KBS·MBC·SBS)가 구성한 방송사공동예측조사위원회(KEP; Korea Election Pool)가 2021년 4월 7일 보궐선거를 앞두고 입소스, 코리아리서치인터내셔널, 한국리서치에 의뢰하여 서울과 부산에서 수행한 1~3차 여론조사 결과를 가지고 연구를 진행하였다.

KEP 전화조사는 통신 3사가 제공하는 휴대전화 가상번호를 100% 사용하는 전화면접 방식으로 수행되었으며, 응답률은 21.3% ~ 36.6%였다. 응답유보층을 예측하기 위하여 사용된 설명변수는 성별, 연령, 직업, 지지정당, 국정운영 평가, 투표의향 등이다.

(1) 응답유보층 보정의 절차

1) 데이터를 다음과 같이 3개의 집합으로 분리한다. D_1 : '누구에게 투표하시겠습니까?' 라는 문항에 기호 1번(더불어민주당) 또는 기호 2번(국민의힘) 후보자를 선택한 응답자 집합, D_2 : 위 문항에 기호 1번, 기호 2번을 제외한 후보자를 선택한 응답자 집합, D_3 : 위 문항에 없다 & 모름/무응답으로 응답한 응답자 집합 (응답유보층)

2) 집합 D_1을 이용해 응답성향 모형에 적합한다.

3) 성별, 연령과 함께 모든 조사에서 공통적으로 유의한 문항을 설명변수로 하는 응답성향 모형을 이용하여 집합 D_3의 각 개체에 대하여 지지후보를 예측한다.

4) 집합 D_1과 집합 D_2에서 응답된 지지후보와 집합 D_3에 대하여 예측된 지지후보를 이용하여 득표율을 예측한다.

(2) 결과

위 논문에서는 각 조사별 기호 1, 2번의 최종 득표율과 응답유보층의 보정을 통한 기호 1, 2번의 예측 득표율의 예측오차를 비교하여 응답유보층의 보정이 득표율 예측의 정확도를 높이는 데 기여할 수 있는지 연구하였다.

6개의 여론조사 결과, 단순보정 결과와 분류모형을 적용한 결과를 제시하였으며, 분류모형으로 로지스틱 회귀 모형을 이용하였다.

응답유보층을 보정하여 득표율을 예측한 결과, 단순 보정의 결과에 비해 예측오차가 상당 폭 줄어들어 상당히 신뢰할 수 있는 예측 득표율 산출이 가능하다는 결론을 얻을 수 있었다.

제 3 장

실증 데이터 분석

본 장에서는 전국지표조사(NBS; National Barometer Survey)에서 제 20대 대통령선거를 앞두고 실시된 53~67차 선거여론조사 자료를 사용하여 득표율을 예측해보고 방법론에 따른 결과를 비교해보고자 한다.

3.1 자료 소개

전국지표조사는 엠브레인퍼블릭, 케이스탯리서치, 코리아리서치, 한국리서치가 외부 기관의 의뢰를 받지 않고 자체적으로 시행·공표하는 정기 전화 여론조사이다. NBS 전화조사는 국내 전화조사에서 일반적으로 사용하는 할당추출(quota sampling) 방법이 아닌 모수추정이 가능한 확률적 표본추출방법인 층화확률추출(stratified random sampling) 방법을 사용한다. 각 시·도 구분과 성/연령을 이용한 세부 층화를 통해 전체 192개 층을 구성하고 유권자 수 기준 비례배분을 통해 세부 층별 목표 표본크기를 정한다. 이후, 국내 통신3사로부터 제공받은 휴대전화 가상번호를 세부 층별 목표 표본크기만큼 무작위 추출하여 컴퓨터를 이용한 전화면접조사(CATI; Computer Assisted Telephone Interview)를 실시한다. 이때, 전화를 받지 않거나 통화 중인 경우 접촉률을 높이기 위해 요일과 시간대를 달리하여 5차례 재통화를 실시한다. 재통화 실시에도 불구하고 조사가 불가능한 경우, 동일한 층 내에서 다른 휴대전화 가상번호 1개를 무작위추출하여 조사를 시도한다.

NBS 전화조사의 목표 표본크기는 <표 1>과 같고, 53~66차는 4개 조사기관이 로테이션 방식으로 매 회차당 2개 조사기관이 조사를 수행하였고 67차는 4개 조사기관이 모두 조사를 수행하였으며, 응답률은 20.3%~32.5%였다.

NBS 전화조사가 응답자에게 조사한 문항은 <표 2>와 같다. 이는 응답자의 기본정보에 해당하는 거주 지역, 성별, 연령, 직업, 학력을 묻는 문항과 응답자의 정치성향을 파악할 수 있는 대통령의 국정운영에 대한 의견, 지지하는 정당, 대통령선거에서 기대하는 결과 등의 문항을 포함하고 있다.

<표 1> NBS 53~67차 전화조사의 표본크기와 응답률

	조사기간	조사기관	표본크기	응답률
53차	11.08 ~ 11.10	코리아리서치, 케이스탯리서치	1,009명	32.5%
54차	11.15 ~ 11.17	케이스탯리서치, 한국리서치	1,004명	30.2%
55차	11.22 ~ 11.24	한국리서치, 엠브레인퍼블릭	1,004명	29.6%
56차	11.29 ~ 12.01	엠브레인퍼블릭, 코리아리서치	1,015명	29.0%
57차	12.06 ~ 12.08	케이스탯리서치, 한국리서치	1,004명	28.3%
58차	12.20 ~ 12.22	한국리서치, 엠브레인퍼블릭	1,000명	24.3%
59차	12.27 ~ 12.29	코리아리서치, 케이스탯리서치	1,000명	28.3%
60차	01.03 ~ 01.05	엠브레인퍼블릭, 코리아리서치	1,000명	27.9%
61차	01.10 ~ 01.12	코리아리서치, 케이스탯리서치	1,000명	29.3%
62차	01.17 ~ 01.19	케이스탯리서치, 한국리서치	1,000명	26.5%
63차	01.24 ~ 01.26	한국리서치, 엠브레인퍼블릭	1,000명	26.7%

계속..

	조사기간	조사기관	표본크기	응답률
64차	02.07 ~ 02.09	엠브레인퍼블릭, 코리아리서치	1,007명	29.7%
65차	02.14 ~ 02.16	코리아리서치, 케이스탯리서치	1,012명	20.3%
66차	02.21 ~ 02.23	케이스탯리서치, 한국리서치	1,004명	25.9%
67차	02.28 ~ 03.02	엠브레인퍼블릭, 케이스탯리서치, 코리아리서치, 한국리서치	2,013명	27.3%

<표 2> NBS 전화조사의 설문문항

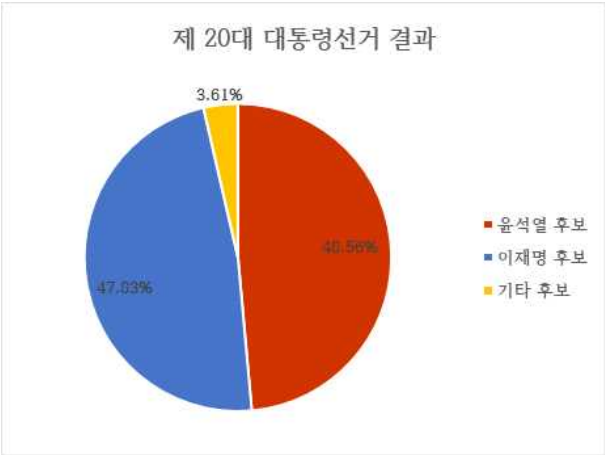
문항	비고
SQ1 : 지역	
SQ2 : 성별	
SQ3 : 연령	
Q1 : 선생님께서는 문재인 대통령이 대통령으로서 일을 잘하고 있다고 생각하십니까? 잘못하고 있다고 생각하십니까?	
Q2 : 선생님께서는 현재 어느 정당을 지지하십니까?	
Q3 : 선생님께서는 내년 3월 치러지는 제20대 대통령선거에서 투표하실 생각이십니까? 투표하지 않으실 생각이십니까?	

계속..

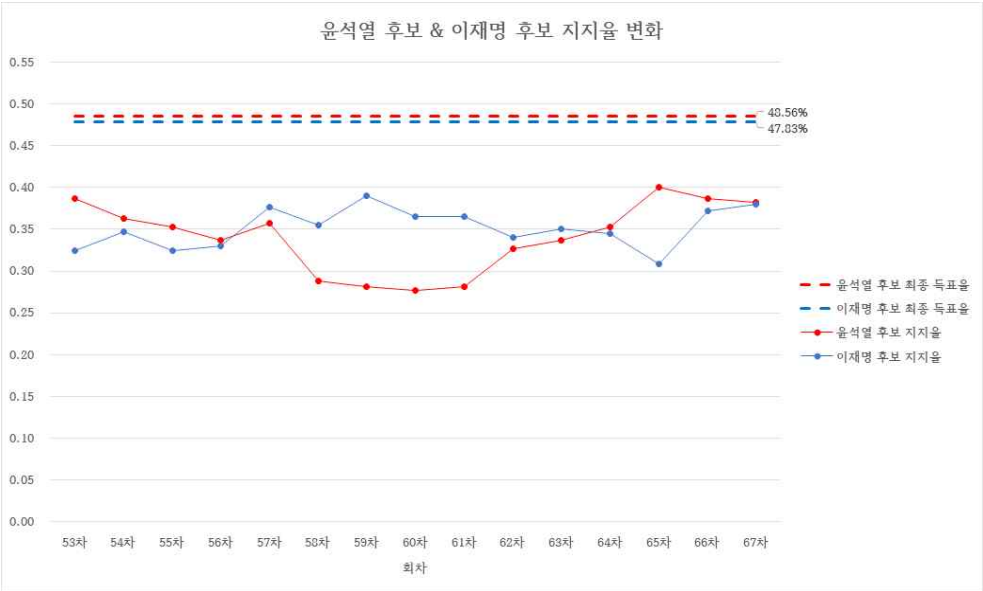
문항	비고
Q4 : 선생님께서는 이번 대선에 출마한 후보 중 누구를 지지하십니까?	
Q6 : 선생님 지지 여부와 상관없이, 선거 분위기나 주변 사람들의 반응을 볼 때 내년 대선에서 어느 후보가 당선될 것으로 보십니까?	
Q7 : 선생님께서는 내년 3월에 치러지는 제20대 대통령선거에 대한 다음 의견 중 어디에 더 공감하십니까?	① 안정적인 국정운영을 위해 여당후보가 당선되어야 한다. ② 국정운영에 대한 심판을 위해 여당후보가 당선되어야한다. ※ 60차 조사에만 포함되지 않은 문항
DQ1 : 직업	
DQ2 : 학력	
DQ3 : 개인의 소득뿐만 아니라 부동산, 예금 등 자산까지 종합적으로 고려한다면, 선생님 본인의 경제적 계층은 다음 중 어디에 속한다고 생각하십니까?	
DQ4 : 선생님의 이념성향은 다음 중 어디에 가깝다고 생각하십니까?	

제 20대 대통령선거는 유권자의 77.08%가 투표에 참여하였고, 최종 득표율은 <그림 1>과 같다. 윤석열 후보가 48.56%로 당선되었고, 이재명 후보는 47.83%를 득표했다.

<그림 1> 제 20대 대통령선거 결과



<그림 2> 윤석열 후보 & 이재명 후보 지지율 변화



<그림 2>는 NBS 전화조사 결과와 최종 득표율을 그래프로 표현한 그림이다. 각 후보의 NBS 전화조사 결과와 최종 득표율이 큰 차이를 보이고 있음을 볼 수 있다.

<표 3>은 NBS 전화조사 결과와 제 20대 대통령선거 결과의 오차이다. NBS 전화조사는 11.03%~24.22%의 응답유보층을 포함하고 있다.

NBS 전화조사 결과와 최종 득표율의 차이가 큰 것은 최종 득표율은 응답유보층을 포함하지 않은 상태에서 후보자별 득표율을 계산하기 때문이다. 따라서 응답유보층을 포함한 상태로 NBS 전화조사 결과를 최종 득표율과 비교하여 정확성을 평가하는 것은 올바른 평가방법이 아니라고 할 수 있다.

하지만 일반 유권자나 언론기관에서는 이런 잘못된 평가방식을 통해 여론조사의 정확성을 평가하고 비판하는 일은 흔히 일어난다. 따라서 NBS 전화조사 결과와 함께 응답유보층을 보정한 후보자별 득표율을 제공하는 것이 올바른 공표 방식이라고 생각한다.

<표 3> NBS 전화조사 결과와 선거결과 오차

회차	윤석열		이재명		기타		응답유보
	여론조사	오차	여론조사	오차	여론조사	오차	
53차	38.67	-9.89	32.47	-15.36	12.25	8.64	16.62
54차	36.27	-12.29	34.67	-13.16	11.48	7.87	17.59
55차	35.26	-13.30	32.47	-15.36	9.54	5.93	22.73
56차	33.64	-14.92	33.07	-14.76	10.40	6.79	22.89
57차	35.69	-12.87	37.68	-10.15	9.25	5.64	17.38

계속..

회차	윤석열		이재명		기타		응답유보
	여론조사	오차	여론조사	오차	여론조사	오차	
58차	28.80	-19.76	35.50	-12.33	11.49	7.88	24.22
59차	28.17	-20.39	39.07	-8.76	13.25	9.64	19.50
60차	27.69	-20.87	36.48	-11.35	15.92	12.31	19.90
61차	28.13	-20.43	36.57	-11.26	17.58	13.97	17.73
62차	32.72	-15.84	34.08	-13.75	16.74	13.13	16.48
63차	33.75	-14.81	35.05	-12.78	13.40	9.79	17.81
64차	35.29	-13.27	34.51	-13.32	14.17	10.56	16.02
65차	40.05	-8.51	30.90	-16.93	11.40	7.79	17.65
66차	38.65	-9.91	37.18	-10.65	13.18	9.57	10.99
67차	38.22	-10.34	37.97	-9.86	10.72	7.11	13.09

3.2 응답유보층 보정 절차

본 절에서는 선거여론조사의 응답유보층 보정을 위해 다음과 같은 절차를 제안한다.

1) 데이터를 다음과 같이 2개의 집합으로 분리한다. D_1 : Q4(선생님께서는 이번

대선에 출마한 후보 중 누구를 지지하십니까?) 문항에 지지후보를 밝힌 응답자 집합, D_2 : Q4 문항에 없다 & 모름/무응답으로 응답한 응답자 집합 (응답유보층)

2) 집합 D_1을 이용해 분류모형에 적합한다.

3) 2)에서 적합한 모형에 집합 D_2를 대입하여 득표율을 예측한다.

3.3 응답유보층 보정을 위한 분류모형

본 절에서는 응답유보층 보정에 이용할 분류모형에 대해 설명하고자 한다.

(1) 다항 로지스틱 회귀

반응변수가 3개 이상의 범주를 가질 때, 다항 로지스틱 회귀분석(multinomial logistic regression)을 사용한다. 반응변수 Y가 K개의 범주를 가진다고 가정한다. 반응변수 Y의 마지막 범주(K)를 기준 범주라고 할 때, 기준 범주 K에 대해 각 범주 k에 속할 확률의 로그를 기준범주 로짓이라고 하는데 이는 식 (3.1)과 같다.

$$\log\frac{P(Y=k)}{P(Y=K)} = \boldsymbol{\beta_k'}\boldsymbol{x}, \quad k=1, \cdots, K-1 \tag{3.1}$$

여기서 $\boldsymbol{x} = (1, x_1, \cdots, x_p)'$는 설명변수의 벡터이며, $\boldsymbol{\beta_k'} = (\alpha_k, \beta_{k1}, \beta_{k2}, \cdots, \beta_{kp})$는 그에 해당하는 회귀 계수 벡터를 의미한다. 위의 식 (3.1)로부터 범주 k에 속할 확률인 식 (3.2)를 구할 수 있다.

$$P(Y=k)=\frac{\exp(\boldsymbol{\beta}_{\boldsymbol{k}}^{'}\boldsymbol{x})}{1+\sum_{i=1}^{K-1}\exp(\boldsymbol{\beta}_{\boldsymbol{i}}^{'}x)},\quad k=1,\cdots,K-1 \tag{3.2}$$

(2) 의사결정나무

의사결정나무는 의사결정 규칙을 나무 구조로 나타내어 전체 자료를 몇 개의 소집단으로 분류(classification)하거나 예측(prediction)을 수행하는 분석방법이다. 나무 구조에 의해서 모형이 표현되기 때문에 직관적으로 이해하기 쉽다는 장점이 있다. 의사결정나무는 반응변수의 종류와 분류 방법에 따라 다양한 알고리즘이 존재하는데, 그 중에서 본 연구에 사용할 CART 알고리즘에 대해 살펴보고자 한다.

Algorithm 1 : CART의 분류 알고리즘

---

1) 각 설명변수에 대해 모든 가능한 분할 지점을 조사한다.
2) 1)에서 구한 분할 지점들 중 불순도의 향상도가 가장 큰 지점을 분류 집합으로 선택한다.
3) 각 관측치에 대해 2)에서 선택한 분류 집합에 속하면 왼쪽 자식노드로, 아니면 오른쪽 자식노드로 보낸다.
4) 위 과정을 모든 하위 노드에서 반복 시행하여 나무 구조를 만든다.

불순도의 향상도(Goodness of split)는 부모노드에 비해 자식노드들의 불순도(impurity)가 얼마나 감소하였는지를 나타내는 척도이다. 노드 내에 관측치들이 동일한 반응변수 값을 가질수록 불순도는 낮아지며, 불순도가 많이 개선되는 방

향으로 분할 규칙을 선택한다. 불순도의 향상도 공식은 식 (3.3)과 같다.

$$\triangle i(s,t) = i(t) - p_L \times i(t_L) - p_R \times i(t_R) \tag{3.3}$$

$i(t)$는 노드t의 불순도 함수이며, s는 분할 변수의 분할 지점을 의미한다. p_L과 p_R은 각각 부모노드 대비 왼쪽 자식노드와 오른쪽 자식노드의 비율을 의미한다. 불순도 함수는 지니 지수(Gini index)를 사용하며 식 (3.4)와 같다.

$$Gini(t) = 1 - \sum_{i=1}^{c} P_i^2 \tag{3.4}$$

P_i는 한 집단에 속한 데이터 중 범주 i에 속하는 관측치의 비율을 의미하고, c는 반응변수 범주의 수이다. 지니 지수는 관측치들이 동일한 반응변수 값을 가질수록 값이 작아진다.

(3) 랜덤 포레스트

랜덤 포레스트는 Leo Breiman(2001)에 의해 제안된 분석방법으로, 학습 과정에서 구성한 여러 개의 의사결정나무로부터 분류나 예측을 수행하는 앙상블(Ensemble) 기법이다. 앙상블 기법이란 주어진 데이터로부터 여러 개의 예측 모형을 생성한 후 이 모형들을 조합하여 최적의 예측 모형을 생성하는 기법을 말한다.

랜덤 포레스트는 과대적합 문제를 최소화하여 모델의 정확도를 향상한다는 장점을 가지고 있으며, 알고리즘은 다음과 같다.

Algorithm 2 : 랜덤 포레스트 알고리즘

1) $b = 1, \cdots, B$ 에 대하여
   가) 학습 데이터에서 크기가 N 인 부트스트랩 표본Z^*을 추출한다.
   나) 부트스트랩 표본Z^*에서 의사결정나무의 각 종단 노드(terminal node)가 최소 노드 크기인 n_{min}에 이를 때까지 다음 과정을 반복하며 의사결정나무 C_b 를 생성한다.
      a) p개의 설명변수에서 m개를 무작위로 선택한다.($m < p$)
      b) 선택한 m개의 설명변수 중 최적의 설명변수와 분기점(split point)을 찾는다.
      c) 노드를 두 개의 자식 노드(child node)로 분할한다.
2) B 개의 C_b 를 앙상블한 결과인 $\{C_b\}_1^B$ 를 도출한다.

분류에 있어 $\widehat{C}_b(x)$ 를 b번째 의사결정나무의 분류 예측이라고 한다면, 새로운 관측치 x에 대한 예측은 다음과 같이 결정된다.

$$\widehat{C}_{rf}^B(x) = majority\ vote\ \left\{\widehat{C}_b(x)\right\}_1^B$$

랜덤 포레스트는 무작위성(randomness)을 최대로 주기 위해서 각 부트스트랩 표본에서 설명변수를 무작위로 추출한다. 따라서 관측치 뿐만 아니라 변수에도 임의성을 적용함으로써 의사결정나무를 구축함에 있어 다양성을 확보하게 된다. 따라서 의사결정나무의 특징인 분산이 크다는 단점을 보완한다.

위 알고리즘에서 랜덤 포레스트의 의사결정나무의 수 B, 각 의사결정나무의 최소 노드 크기 n_{min}과 각 노드에서 임의로 선택할 설명변수의 수 m은 적절한 값

을 선택해야하는 초모수이다. 의사결정나무의 수는 랜덤 포레스트의 성능을 결정하는 주요 모수로 클수록 모형의 성능이 좋아진다. 최소 노드 크기는 나무의 깊이를 설정하는 모수로 클수록 얕은 나무를 생성한다. 각 노드에서 임의로 선택할 설명변수의 수는 랜덤 포레스트의 트리들 간의 연관성을 설정하는 모수로 클수록 비슷한 나무들을 생성한다.

분류의 경우, 일반적으로 n_{min}은 1을 사용하고 m은 $\sqrt{p}$를 사용한다.

3.4 분석 결과

본 절에서는 득표율을 예측해보고 방법론에 따른 결과를 비교해보고자 한다.

본 연구에서는 다항 로지스틱 회귀모형과 랜덤 포레스트에 상수항을 포함시키지 않고 연구를 진행하였다.

(1) 응답유보층 보정

3.2절에서 제안한 응답유보층 보정 절차를 따라 응답유보층을 보정하고 득표율을 예측해보았다. 비교를 위해서 단순보정 결과를 함께 제시했다. 단순보정이란 NBS 전화조사 결과에서 응답유보층을 제외하고 나머지 후보자들의 지지율 합이 100%가 되도록 조정한 것이다. 또한, 최종 득표율과의 오차를 산출해 함께 제시했다.

• 다항 로지스틱 회귀

<표 4>는 다항 로지스틱 회귀모형을 이용한 응답유보층 보정을 통해 득표율을 예측한 것이다.

분류모형을 이용한 응답유보층 보정의 예측오차가 단순보정의 예측오차보다 크게 나타났다. 따라서 분류모형을 이용한 응답유보층 보정이 단순보정에 비해 효과가 없는 것으로 보인다.

<표 4> 다항 로지스틱 회귀모형을 이용한 응답유보층 보정

회차	후보	선거 결과	예측		예측오차	
			단순 보정	분류분석 보정	단순 보정	분류분석 보정
53차	윤석열	48.56	46.37	44.39	-2.19	-4.17
	이재명	47.83	38.94	38.20	-8.89	-9.63
	기타	3.61	14.69	17.41	11.08	13.80
54차	윤석열	48.56	44.01	40.84	-4.55	-7.72
	이재명	47.83	42.07	40.54	-5.76	-7.29
	기타	3.61	13.93	18.62	10.32	15.01
55차	윤석열	48.56	45.63	43.24	-2.93	-5.32
	이재명	47.83	42.02	39.90	-5.81	-7.93
	기타	3.61	12.34	16.86	8.73	13.25

계속..

회차	후보	선거 결과	예측		예측오차	
			단순 보정	분류분석 보정	단순 보정	분류분석 보정
56차	윤석열	48.56	43.63	39.87	-4.93	-8.69
	이재명	47.83	42.89	43.42	-4.94	-4.41
	기타	3.61	13.48	16.71	9.87	13.10
57차	윤석열	48.56	43.20	41.79	-5.36	-6.77
	이재명	47.83	45.60	43.57	-2.23	-4.26
	기타	3.61	11.20	14.64	7.59	11.03
58차	윤석열	48.56	38.00	36.13	-10.56	-12.43
	이재명	47.83	46.84	43.39	-0.99	-4.44
	기타	3.61	15.16	20.47	11.55	16.86
59차	윤석열	48.56	35.00	32.39	-13.56	-16.17
	이재명	47.83	48.54	46.62	0.71	-1.21
	기타	3.61	16.46	20.99	12.85	17.38
60차	윤석열	48.56	34.58	32.34	-13.98	-16.22
	이재명	47.83	45.54	43.14	-2.29	-4.69
	기타	3.61	19.88	24.52	16.27	20.91

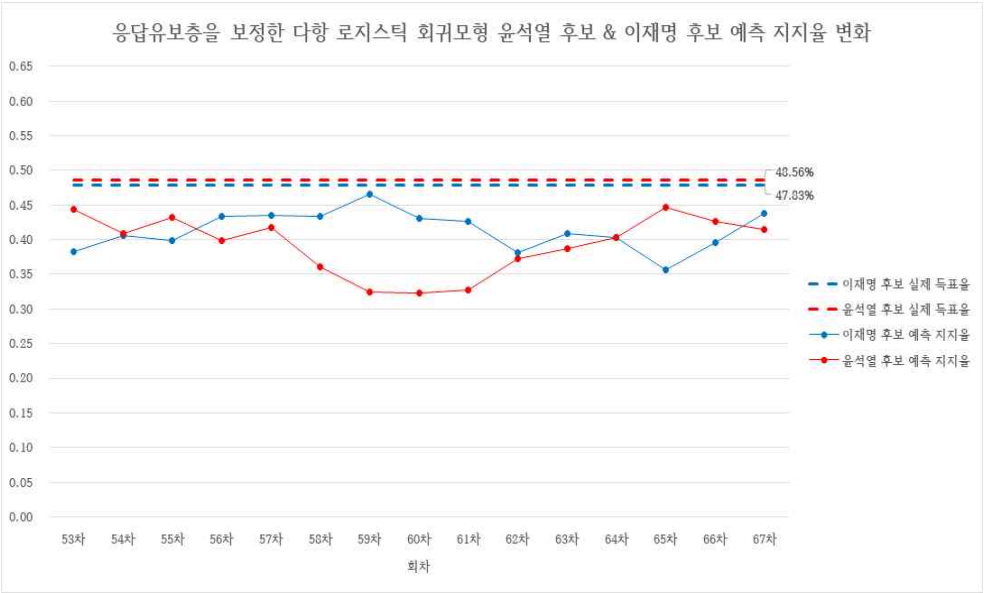
계속..

회차	후보	선거 결과	예측		예측오차	
			단순 보정	분류분석 보정	단순 보정	분류분석 보정
61차	윤석열	48.56	34.19	32.80	-14.37	-15.76
	이재명	47.83	44.45	42.69	-3.38	-5.14
	기타	3.61	21.36	24.50	17.75	20.89
62차	윤석열	48.56	39.17	37.24	-9.39	-11.33
	이재명	47.83	40.80	38.17	-7.03	-9.66
	기타	3.61	20.03	24.60	16.42	20.99
63차	윤석열	48.56	41.06	38.70	-7.50	-9.86
	이재명	47.83	42.64	40.85	-5.19	-6.98
	기타	3.61	16.30	20.45	12.69	16.84
64차	윤석열	48.56	42.03	40.30	-6.53	-8.26
	이재명	47.83	41.10	40.25	-6.73	-7.58
	기타	3.61	16.88	19.46	13.27	15.85
65차	윤석열	48.56	48.64	44.71	0.08	-3.85
	이재명	47.83	37.52	35.61	-10.31	-12.22
	기타	3.61	13.85	19.68	10.24	16.07

계속..

회차	후보	선거 결과	예측		예측오차	
			단순 보정	분류분석 보정	단순 보정	분류분석 보정
66차	윤석열	48.56	43.42	42.66	-5.14	-5.90
	이재명	47.83	41.77	39.61	-6.06	-8.22
	기타	3.61	14.81	17.72	11.20	14.11
67차	윤석열	48.56	43.97	41.47	-4.59	-7.09
	이재명	47.83	43.69	43.74	-4.14	-4.09
	기타	3.61	12.34	14.79	8.73	11.18

<그림 3> 응답유보층을 보정한 다항 로지스틱 회귀모형
: 윤석열 후보 & 이재명 후보 예측 지지율 변화



<그림 3>은 다항 로지스틱 회귀모형을 이용한 응답유보층 보정을 통해 득표율을 예측한 것을 그래프로 표현한 그림이다.

<그림 2>와 비교했을 때, 최종 득표율과의 격차는 줄어들었지만 마지막 조사의 예측 득표율이 윤석열 후보자는 41.47%, 이재명 후보자는 43.74%로 예측순위가 실제와 뒤바뀐 결과를 보여주었다.

• 의사결정나무

<표 5>는 의사결정나무를 이용한 응답유보층 보정을 통해 최종 득표율을 예측한 것이다.

분류모형을 이용한 응답유보층 보정의 예측오차가 단순보정의 예측오차보다 크게 나타났다. 따라서 분류모형을 이용한 응답유보층 보정이 단순보정에 비해 효과가 없는 것으로 보인다.

<표 5> 의사결정나무를 이용한 응답유보층 보정

회차	후보	선거 결과	예측		예측오차	
			단순 보정	분류분석 보정	단순 보정	분류분석 보정
53차	윤석열	48.56	46.37	45.80	-2.19	-2.76
	이재명	47.83	38.94	38.41	-8.89	-9.42
	기타	3.61	14.69	15.79	11.08	12.18
54차	윤석열	48.56	44.01	42.76	-4.55	-5.80
	이재명	47.83	42.07	41.46	-5.76	-6.37
	기타	3.61	13.93	15.78	10.32	12.17
55차	윤석열	48.56	45.63	43.76	-2.93	-4.80
	이재명	47.83	42.02	41.05	-5.81	-6.78
	기타	3.61	12.34	15.19	8.73	11.58

계속..

회차	후보	선거 결과	예측		예측오차	
			단순 보정	분류분석 보정	단순 보정	분류분석 보정
56차	윤석열	48.56	43.63	41.78	-4.93	-6.78
	이재명	47.83	42.89	43.90	-4.94	-3.93
	기타	3.61	13.48	14.31	9.87	10.70
57차	윤석열	48.56	43.20	43.32	-5.36	-5.24
	이재명	47.83	45.60	45.97	-2.23	-1.86
	기타	3.61	11.20	10.71	7.59	7.10
58차	윤석열	48.56	38.00	34.35	-10.56	-14.21
	이재명	47.83	46.84	49.07	-0.99	1.24
	기타	3.61	15.16	16.58	11.55	12.97
59차	윤석열	48.56	35.00	33.08	-13.56	-15.48
	이재명	47.83	48.54	47.15	0.71	-0.68
	기타	3.61	16.46	19.77	12.85	16.16
60차	윤석열	48.56	34.58	31.85	-13.98	-16.71
	이재명	47.83	45.54	43.90	-2.29	-3.93
	기타	3.61	19.88	24.25	16.27	20.64

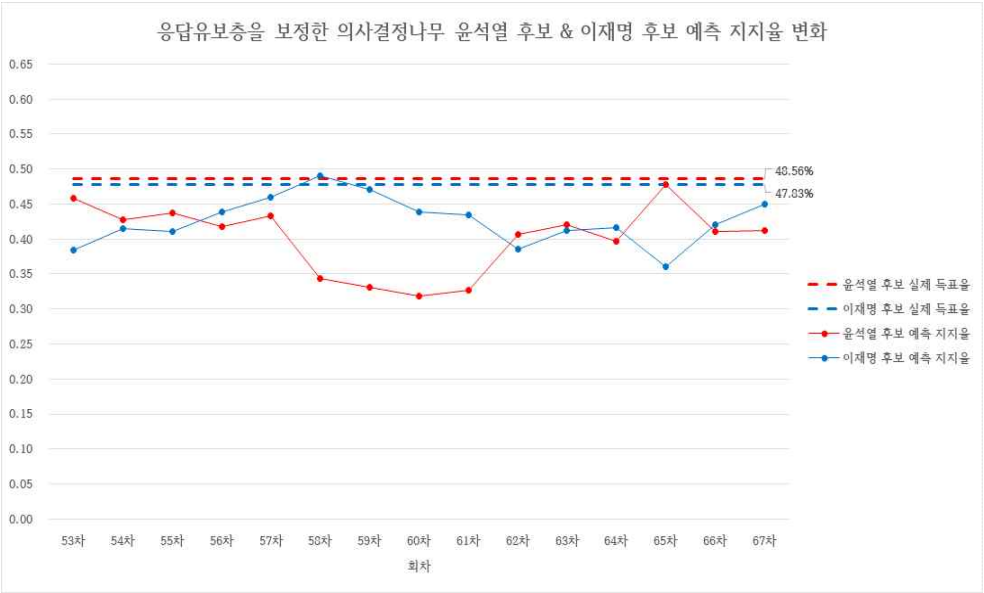
계속..

회차	후보	선거 결과	예측		예측오차	
			단순 보정	분류분석 보정	단순 보정	분류분석 보정
61차	윤석열	48.56	34.19	32.75	-14.37	-15.81
	이재명	47.83	44.45	43.44	-3.38	-4.39
	기타	3.61	21.36	23.81	17.75	20.20
62차	윤석열	48.56	39.17	40.64	-9.39	-7.93
	이재명	47.83	40.80	38.56	-7.03	-9.27
	기타	3.61	20.03	20.80	16.42	17.19
63차	윤석열	48.56	41.06	42.10	-7.50	-6.46
	이재명	47.83	42.64	41.27	-5.19	-6.56
	기타	3.61	16.30	16.63	12.69	13.02
64차	윤석열	48.56	42.03	39.65	-6.53	-8.91
	이재명	47.83	41.10	41.59	-6.73	-6.24
	기타	3.61	16.88	18.76	13.27	15.15
65차	윤석열	48.56	48.64	47.85	0.08	-0.71
	이재명	47.83	37.52	36.05	-10.31	-11.78
	기타	3.61	13.85	16.10	10.24	12.49

계속..

회차	후보	선거 결과	예측		예측오차	
			단순 보정	분류분석 보정	단순 보정	분류분석 보정
66차	윤석열	48.56	43.42	41.04	-5.14	-7.52
	이재명	47.83	41.77	42.12	-6.06	-5.71
	기타	3.61	14.81	16.84	11.20	13.23
67차	윤석열	48.56	43.97	41.19	-4.59	-7.37
	이재명	47.83	43.69	44.98	-4.14	-2.85
	기타	3.61	12.34	13.83	8.73	10.22

<그림 4> 응답유보층을 보정한 의사결정나무
: 윤석열 후보 & 이재명 후보 예측 지지율 변화



<그림 4>는 의사결정나무를 이용한 응답유보층 보정을 통해 득표율을 예측한 것을 그래프로 표현한 그림이다.

<그림 2>와 비교했을 때, 최종 득표율과의 격차는 줄어들었지만 마지막 조사의 예측 득표율이 윤석열 후보자는 41.19%, 이재명 후보자는 44.98%로 예측순위가 실제와 뒤바뀐 결과를 보여주었다.

• 랜덤 포레스트

<표 6>은 랜덤 포레스트를 이용한 응답유보층 보정을 통해 득표율을 예측한 것이다.

분류모형을 이용한 응답유보층 보정의 예측오차가 단순보정의 예측오차보다 크게 나타났다. 따라서 분류모형을 이용한 응답유보층 보정이 단순보정에 비해 효과가 없는 것으로 보인다.

<표 6> 랜덤 포레스트를 이용한 응답유보층 보정

회차	후보	선거 결과	예측		예측오차	
			단순 보정	분류분석 보정	단순 보정	분류분석 보정
53차	윤석열	48.56	46.37	44.98	-2.19	-3.58
	이재명	47.83	38.94	39.57	-8.89	-8.26
	기타	3.61	14.69	15.45	11.08	11.84
54차	윤석열	48.56	44.01	42.41	-4.55	-6.15
	이재명	47.83	42.07	41.98	-5.76	-5.85
	기타	3.61	13.93	15.61	10.32	12.00
55차	윤석열	48.56	45.63	43.18	-2.93	-5.38
	이재명	47.83	42.02	42.87	-5.81	-4.96
	기타	3.61	12.34	13.94	8.73	10.33

계속..

회차	후보	선거 결과	예측		예측오차	
			단순 보정	분류분석 보정	단순 보정	분류분석 보정
56차	윤석열	48.56	43.63	40.98	-4.93	-7.58
	이재명	47.83	42.89	44.83	-4.94	-3.00
	기타	3.61	13.48	14.19	9.87	10.58
57차	윤석열	48.56	43.20	41.62	-5.36	-6.94
	이재명	47.83	45.60	47.48	-2.23	-0.35
	기타	3.61	11.20	10.90	7.59	7.29
58차	윤석열	48.56	38.00	37.24	-10.56	-11.32
	이재명	47.83	46.84	45.32	-0.99	-2.51
	기타	3.61	15.16	17.44	11.55	13.83
59차	윤석열	48.56	35.00	32.20	-13.56	-16.36
	이재명	47.83	48.54	49.22	0.71	1.39
	기타	3.61	16.46	18.58	12.85	14.97
60차	윤석열	48.56	34.58	32.94	-13.98	-15.62
	이재명	47.83	45.54	44.37	-2.29	-3.46
	기타	3.61	19.88	22.69	16.27	19.08

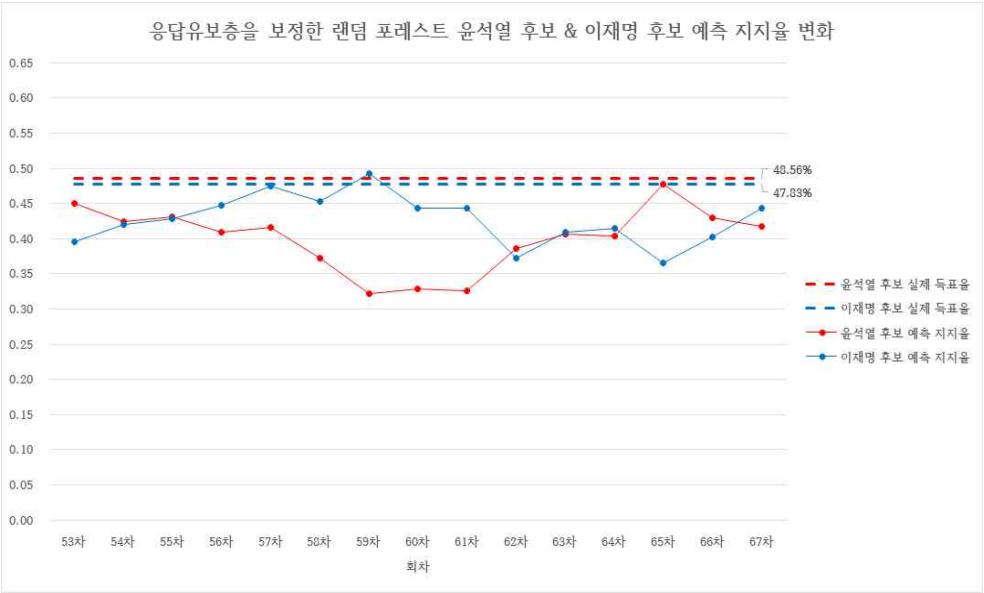
계속..

회차	후보	선거 결과	예측		예측오차	
			단순 보정	분류분석 보정	단순 보정	분류분석 보정
61차	윤석열	48.56	34.19	32.59	-14.37	-15.97
	이재명	47.83	44.45	44.34	-3.38	-3.49
	기타	3.61	21.36	23.08	17.75	19.47
62차	윤석열	48.56	39.17	38.66	-9.39	-9.91
	이재명	47.83	40.80	37.27	-7.03	-10.56
	기타	3.61	20.03	24.08	16.42	20.47
63차	윤석열	48.56	41.06	40.72	-7.50	-7.84
	이재명	47.83	42.64	40.87	-5.19	-6.96
	기타	3.61	16.30	18.40	12.69	14.79
64차	윤석열	48.56	42.03	40.33	-6.53	-8.23
	이재명	47.83	41.10	41.53	-6.73	-6.30
	기타	3.61	16.88	18.14	13.27	14.53
65차	윤석열	48.56	48.64	47.83	0.08	-0.73
	이재명	47.83	37.52	36.63	-10.31	-11.20
	기타	3.61	13.85	15.54	10.24	11.93

계속..

회차	후보	선거 결과	예측		예측오차	
			단순 보정	분류분석 보정	단순 보정	분류분석 보정
66차	윤석열	48.56	43.42	43.02	-5.14	-5.54
	이재명	47.83	41.77	40.22	-6.06	-7.61
	기타	3.61	14.81	16.77	11.20	13.16
67차	윤석열	48.56	43.97	41.80	-4.59	-6.76
	이재명	47.83	43.69	44.36	-4.14	-3.47
	기타	3.61	12.34	13.84	8.73	10.23

<그림 5> 응답유보층을 보정한 랜덤 포레스트
: 윤석열 후보 & 이재명 후보 예측 지지율 변화



<그림 5>는 랜덤 포레스트를 이용한 응답유보층 보정을 통해 득표율을 예측한 것을 그래프로 표현한 그림이다.

<그림 2>와 비교했을 때, 최종 득표율과의 격차는 줄어들었지만 마지막 조사의 예측 득표율이 윤석열 후보자는 41.59%, 이재명 후보자는 43.85%로 예측순위가 실제와 뒤바뀐 결과를 보여주었다.

응답유보층을 보정해 득표율을 예측한 결과, 단순보정의 결과가 분류모형을 이용한 응답유보층 보정의 결과보다 더 정확한 예측 결과를 보여주었다.

분류모형을 이용하여 응답유보층을 보정한 경우, 마지막 조사의 예측순위가 실제와 뒤바뀐 결과를 보여 분류모형을 이용한 응답유보층 보정이 올바른 예측을 하지 않은 것으로 보인다.

이는 여론조사 공표금지 기간에 일어난 윤석열 후보와 안철수 후보의 단일화로 인해 응답유보층만을 보정한 것이 원인이 된 것으로 추측된다.

(2) (응답유보층 + 안철수 후보 지지자) 보정

2022년 3월 9일 제 20대 대통령 선거가 시행되기 전, 2022년 3월 3일 윤석열 후보와 안철수 후보가 단일화를 선언했다. 공직선거법 제108조에 의하면 선거일 6일 전인 3월 3일부터 선거일의 투표마감시각까지 선거에 관하여 정당에 대한 지지도나 당선인을 예상하게 하는 여론조사의 경위와 그 결과를 공표하거나 인용하여 보도할 수 없다. 따라서 단일화로 인한 지지율 변화를 바로 확인할 수 없다는 점에서 단일화가 가져올 결과에 대해 예측하기 어려운 상황이 되었다.

본 연구에서 사용할 NBS 전화조사의 모든 회차도 단일화 선언 전에 조사가 완료되어 안철수 후보가 사퇴한 이후 안철수 후보를 지지했던 유권자들이 실제 투표에서 어떤 후보자를 지지하게 될지 파악할 수 없게 되었다.

따라서 본 논문의 3.2 응답유보층 보정 절차 1)의 집합 D_2를 Q4 문항에 '안철수 후보를 지지한다.'로 응답한 응답자, '지지후보가 없다.'로 응답한 응답자와 '모름 또는 무응답'으로 응답한 응답자 집합, 즉 (응답유보층 + 안철수 후보 지지자)로 변경하여 분석을 진행하고자 한다.

• 다항 로지스틱 회귀

<표 7>은 다항 로지스틱 회귀모형을 이용한 (응답유보층 + 안철수 후보 지지자) 보정을 통해 득표율을 예측한 것이다.

분류모형을 이용한 응답유보층 보정의 예측오차가 단순보정의 예측오차보다 작게 나타났다. 따라서 분류모형을 이용한 응답유보층 보정이 단순보정에 비해 효과가 있는 것으로 보인다.

<표 7> 다항 로지스틱 회귀모형을 이용한 (응답유보층+안철수 후보 지지자) 보정

회차	후보	선거 결과	예측		예측오차	
			단순 보정	분류분석 보정	단순 보정	분류분석 보정
53차	윤석열	48.56	46.37	48.10	-2.19	-0.46
	이재명	47.83	38.94	39.64	-8.89	-8.19
	기타	3.61	14.69	12.26	11.08	8.65
54차	윤석열	48.56	44.01	45.22	-4.55	-3.34
	이재명	47.83	42.07	42.53	-5.76	-5.30
	기타	3.61	13.93	12.25	10.32	8.64
55차	윤석열	48.56	45.63	45.30	-2.93	-3.26
	이재명	47.83	42.02	44.31	-5.81	-3.52
	기타	3.61	12.34	10.39	8.73	6.78

계속..

회차	후보	선거 결과	예측		예측오차	
			단순 보정	분류분석 보정	단순 보정	분류분석 보정
56차	윤석열	48.56	43.63	43.42	-4.93	-5.14
	이재명	47.83	42.89	45.74	-4.94	-2.09
	기타	3.61	13.48	10.84	9.87	7.23
57차	윤석열	48.56	43.20	44.95	-5.36	-3.61
	이재명	47.83	45.60	46.01	-2.23	-1.82
	기타	3.61	11.20	9.04	7.59	5.43
58차	윤석열	48.56	38.00	40.57	-10.56	-7.99
	이재명	47.83	46.84	47.11	-0.99	-0.72
	기타	3.61	15.16	12.32	11.55	8.71
59차	윤석열	48.56	35.00	37.92	-13.56	-10.64
	이재명	47.83	48.54	49.69	0.71	1.86
	기타	3.61	16.46	12.40	12.85	8.79
60차	윤석열	48.56	34.58	37.85	-13.98	-10.71
	이재명	47.83	45.54	49.71	-2.29	1.88
	기타	3.61	19.88	12.43	16.27	8.82

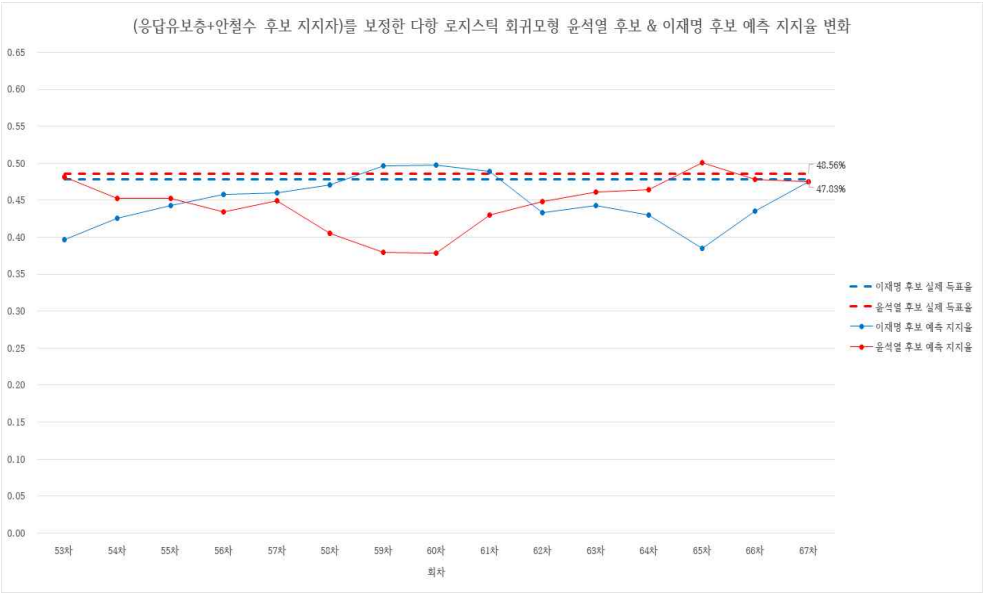
계속..

회차	후보	선거 결과	예측		예측오차	
			단순 보정	분류분석 보정	단순 보정	분류분석 보정
61차	윤석열	48.56	34.19	42.99	-14.37	-5.57
	이재명	47.83	44.45	48.90	-3.38	1.07
	기타	3.61	21.36	8.11	17.75	4.50
62차	윤석열	48.56	39.17	44.78	-9.39	-3.78
	이재명	47.83	40.80	43.37	-7.03	-4.46
	기타	3.61	20.03	11.85	16.42	8.24
63차	윤석열	48.56	41.06	46.06	-7.50	-2.50
	이재명	47.83	42.64	44.26	-5.19	-3.57
	기타	3.61	16.30	9.67	12.69	6.06
64차	윤석열	48.56	42.03	46.40	-6.53	-2.16
	이재명	47.83	41.10	43.03	-6.73	-4.80
	기타	3.61	16.88	10.57	13.27	6.96
65차	윤석열	48.56	48.64	50.11	0.08	1.55
	이재명	47.83	37.52	38.52	-10.31	-9.31
	기타	3.61	13.85	11.37	10.24	7.76

계속..

회차	후보	선거 결과	예측		예측오차	
			단순 보정	분류분석 보정	단순 보정	분류분석 보정
66차	윤석열	48.56	43.42	47.79	-5.14	-0.77
	이재명	47.83	41.77	43.59	-6.06	-4.24
	기타	3.61	14.81	8.62	11.20	5.01
67차	윤석열	48.56	43.97	47.54	-4.59	-1.02
	이재명	47.83	43.69	47.50	-4.14	-0.33
	기타	3.61	12.34	4.96	8.73	1.35

<그림 6> (응답유보층 + 안철수 후보 지지자)를 보정한 다항 로지스틱 회귀모형 : 윤석열 후보 & 이재명 후보 예측 지지율 변화



<그림 6>은 다항 로지스틱 회귀모형을 이용한 (응답유보층 + 안철수 후보 지지자) 보정을 통해 득표율을 예측한 것을 그래프로 표현한 그림이다.

<그림 2>와 비교했을 때, 최종 득표율과의 격차가 줄어들었고 마지막 조사의 예측 득표율이 윤석열 후보자는 47.54%, 이재명 후보자는 47.50%로 예측순위가 실제와 같은 결과를 보여주었다.

• 의사결정나무

<표 8>은 의사결정나무를 이용한 (응답유보층 + 안철수 후보 지지자) 보정을 통해 득표율을 예측한 것이다.

분류모형을 이용한 응답유보층 보정의 예측오차가 단순보정의 예측오차보다 작게 나타났다. 따라서 분류모형을 이용한 응답유보층 보정이 단순보정에 비해 효과가 있는 것으로 보인다.

<표 8> 의사결정나무를 이용한 (응답유보층 + 안철수 후보 지지자) 보정

회차	후보	선거 결과	예측		예측오차	
			단순 보정	분류분석 보정	단순 보정	분류분석 보정
53차	윤석열	48.56	46.37	51.72	-2.19	3.16
	이재명	47.83	38.94	38.90	-8.89	-8.93
	기타	3.61	14.69	9.39	11.08	5.78
54차	윤석열	48.56	44.01	45.52	-4.55	-3.04
	이재명	47.83	42.07	47.91	-5.76	0.08
	기타	3.61	13.93	6.58	10.32	2.97
55차	윤석열	48.56	45.63	43.82	-2.93	-4.74
	이재명	47.83	42.02	50.02	-5.81	2.19
	기타	3.61	12.34	6.16	8.73	2.55

계속..

회차	후보	선거 결과	예측		예측오차	
			단순 보정	분류분석 보정	단순 보정	분류분석 보정
56차	윤석열	48.56	43.63	46.86	-4.93	-1.70
	이재명	47.83	42.89	44.80	-4.94	-3.03
	기타	3.61	13.48	8.34	9.87	4.73
57차	윤석열	48.56	43.20	44.78	-5.36	-3.78
	이재명	47.83	45.60	48.75	-2.23	0.92
	기타	3.61	11.20	6.48	7.59	2.87
58차	윤석열	48.56	38.00	41.41	-10.56	-7.15
	이재명	47.83	46.84	52.28	-0.99	4.45
	기타	3.61	15.16	6.31	11.55	2.70
59차	윤석열	48.56	35.00	39.47	-13.56	-9.09
	이재명	47.83	48.54	51.47	0.71	3.64
	기타	3.61	16.46	9.06	12.85	5.45
60차	윤석열	48.56	34.58	40.53	-13.98	-8.03
	이재명	47.83	45.54	54.28	-2.29	6.45
	기타	3.61	19.88	5.19	16.27	1.58

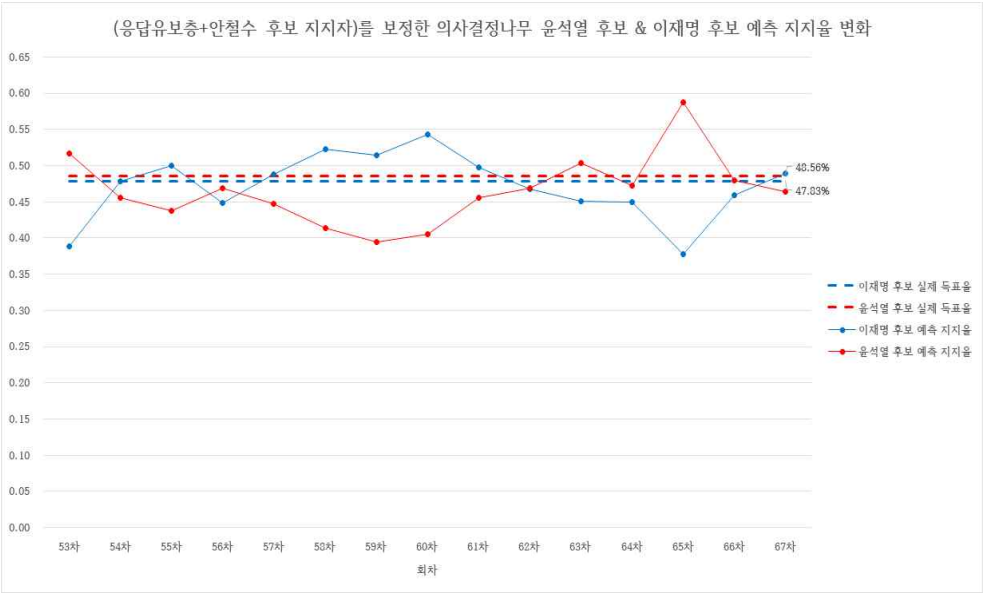
계속..

회차	후보	선거 결과	예측		예측오차	
			단순 보정	분류분석 보정	단순 보정	분류분석 보정
61차	윤석열	48.56	34.19	45.63	-14.37	-2.93
	이재명	47.83	44.45	49.78	-3.38	1.95
	기타	3.61	21.36	4.60	17.75	0.99
62차	윤석열	48.56	39.17	46.90	-9.39	-1.66
	이재명	47.83	40.80	46.82	-7.03	-1.01
	기타	3.61	20.03	6.28	16.42	2.67
63차	윤석열	48.56	41.06	50.35	-7.50	1.79
	이재명	47.83	42.64	45.13	-5.19	-2.70
	기타	3.61	16.30	4.52	12.69	0.91
64차	윤석열	48.56	42.03	47.20	-6.53	-1.36
	이재명	47.83	41.10	45.00	-6.73	-2.83
	기타	3.61	16.88	7.80	13.27	4.19
65차	윤석열	48.56	48.64	58.74	0.08	10.18
	이재명	47.83	37.52	37.80	-10.31	-10.03
	기타	3.61	13.85	3.47	10.24	-0.14

계속..

회차	후보	선거 결과	예측		예측오차	
			단순 보정	분류분석 보정	단순 보정	분류분석 보정
66차	윤석열	48.56	43.42	47.98	-5.14	-0.58
	이재명	47.83	41.77	45.89	-6.06	-1.94
	기타	3.61	14.81	6.13	11.20	2.52
67차	윤석열	48.56	43.97	46.41	-4.59	-2.15
	이재명	47.83	43.69	48.87	-4.14	1.04
	기타	3.61	12.34	4.72	8.73	1.11

<그림 7> (응답유보층 + 안철수 후보 지지자)를 보정한 의사결정나무
: 윤석열 후보 & 이재명 후보 예측 지지율 변화



<그림 7>은 의사결정나무를 이용한 (응답유보층 + 안철수 후보 지지자) 보정을 통해 득표율을 예측한 것을 그래프로 표현한 그림이다.

<그림 2>와 비교했을 때, 최종 득표율과의 격차는 줄어들었지만 마지막 조사의 예측 득표율이 윤석열 후보자는 46.41%, 이재명 후보자는 48.87%로 예측순위가 실제와 뒤바뀐 결과를 보여주었다.

• 랜덤 포레스트

<표 9>는 랜덤 포레스트를 이용한 (응답유보층 + 안철수 후보 지지자) 보정을 통해 득표율을 예측한 것이다.

분류모형을 이용한 응답유보층 보정의 예측오차가 단순보정의 예측오차보다 작게 나타났다. 따라서 분류모형을 이용한 응답유보층 보정이 단순보정에 비해 효과가 있는 것으로 보인다.

<표 9> 랜덤 포레스트를 이용한 (응답유보층 + 안철수 후보 지지자) 보정

회차	후보	선거 결과	예측		예측오차	
			단순 보정	분류분석 보정	단순 보정	분류분석 보정
53차	윤석열	48.56	46.37	48.26	-2.19	-0.30
	이재명	47.83	38.94	42.26	-8.89	-5.57
	기타	3.61	14.69	9.48	11.08	5.87
54차	윤석열	48.56	44.01	48.49	-4.55	-0.07
	이재명	47.83	42.07	44.24	-5.76	-3.59
	기타	3.61	13.93	7.28	10.32	3.67
55차	윤석열	48.56	45.63	47.34	-2.93	-1.22
	이재명	47.83	42.02	45.65	-5.81	-2.18
	기타	3.61	12.34	7.02	8.73	3.41

계속..

회차	후보	선거 결과	예측		예측오차	
			단순 보정	분류분석 보정	단순 보정	분류분석 보정
56차	윤석열	48.56	43.63	45.29	-4.93	-3.27
	이재명	47.83	42.89	46.41	-4.94	-1.42
	기타	3.61	13.48	8.30	9.87	4.69
57차	윤석열	48.56	43.20	44.43	-5.36	-4.13
	이재명	47.83	45.60	48.98	-2.23	1.15
	기타	3.61	11.20	6.59	7.59	2.98
58차	윤석열	48.56	38.00	41.98	-10.56	-6.58
	이재명	47.83	46.84	51.05	-0.99	3.22
	기타	3.61	15.16	6.96	11.55	3.35
59차	윤석열	48.56	35.00	38.68	-13.56	-9.88
	이재명	47.83	48.54	52.01	0.71	4.18
	기타	3.61	16.46	9.31	12.85	5.70
60차	윤석열	48.56	34.58	39.75	-13.98	-8.81
	이재명	47.83	45.54	54.18	-2.29	6.35
	기타	3.61	19.88	6.07	16.27	2.46

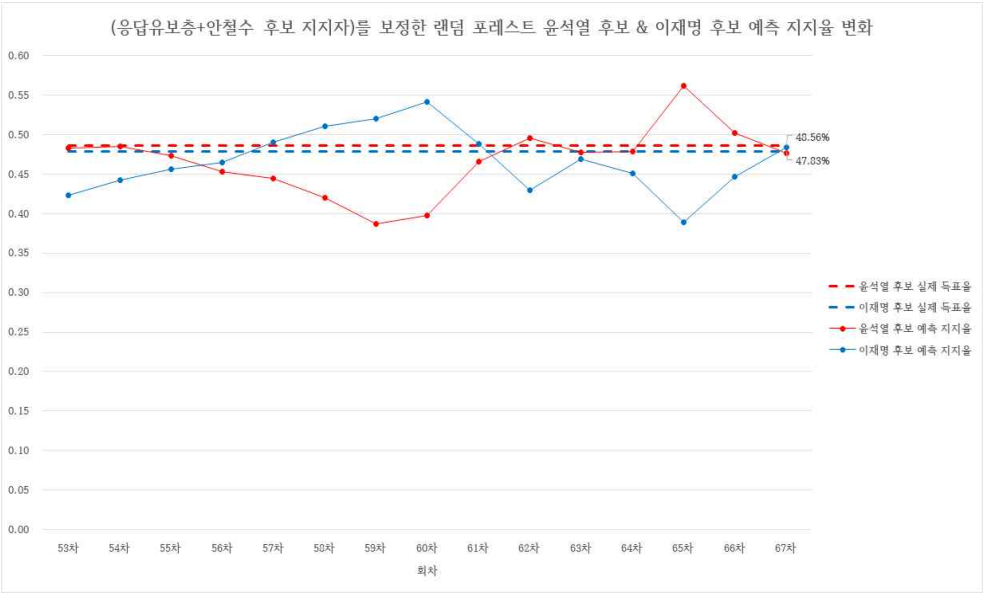
계속..

회차	후보	선거 결과	예측		예측오차	
			단순 보정	분류분석 보정	단순 보정	분류분석 보정
61차	윤석열	48.56	34.19	46.59	-14.37	-1.97
	이재명	47.83	44.45	48.81	-3.38	0.98
	기타	3.61	21.36	4.60	17.75	0.99
62차	윤석열	48.56	39.17	49.58	-9.39	1.02
	이재명	47.83	40.80	42.95	-7.03	-4.88
	기타	3.61	20.03	7.47	16.42	3.86
63차	윤석열	48.56	41.06	47.76	-7.50	-0.80
	이재명	47.83	42.64	46.88	-5.19	-0.95
	기타	3.61	16.30	5.36	12.69	1.75
64차	윤석열	48.56	42.03	47.89	-6.53	-0.67
	이재명	47.83	41.10	45.12	-6.73	-2.71
	기타	3.61	16.88	6.99	13.27	3.38
65차	윤석열	48.56	48.64	56.15	0.08	7.59
	이재명	47.83	37.52	38.89	-10.31	-8.94
	기타	3.61	13.85	4.96	10.24	1.35

계속..

회차	후보	선거 결과	예측		예측오차	
			단순 보정	분류분석 보정	단순 보정	분류분석 보정
66차	윤석열	48.56	43.42	50.18	-5.14	1.62
	이재명	47.83	41.77	44.68	-6.06	-3.15
	기타	3.61	14.81	5.14	11.20	1.53
67차	윤석열	48.56	43.97	47.65	-4.59	-0.91
	이재명	47.83	43.69	48.42	-4.14	0.59
	기타	3.61	12.34	3.93	8.73	0.32

<그림 8> (응답유보층 + 안철수 후보 지지자)를 보정한 랜덤 포레스트
: 윤석열 후보 & 이재명 후보 예측 지지율 변화



<그림 8>은 랜덤 포레스트를 이용한 (응답유보층 + 안철수 후보 지지자) 보정을 통해 득표율을 예측한 것을 그래프로 표현한 그림이다.

<그림 2>와 비교했을 때, 최종 득표율과의 격차가 줄어들었고 마지막 조사의 예측 득표율이 윤석열 후보자는 47.35%, 이재명 후보자는 47.52%로 예측순위가 실제와 뒤바뀐 결과를 보여주었다.

(응답유보층 + 안철수 후보 지지자)를 보정해 득표율을 예측한 결과, 분류모형을 이용한 보정의 결과가 단순보정의 결과보다 더 정확한 예측 결과를 보여주었다.
다항 로지스틱 회귀모형을 이용한 경우, 마지막 조사의 예측순위가 실제와 같은 결과를 보여 다항 로지스틱 회귀모형을 이용한 (응답유보층 + 안철수 후보 지지자) 보정이 효과가 있다는 결론을 내릴 수 있었다.
반면에 의사결정나무와 랜덤 포레스트를 이용한 경우, 마지막 조사의 예측순위가 실제와 뒤바뀐 결과를 보여 의사결정나무와 랜덤 포레스트를 이용한 (응답유

보층 + 안철수 후보 지지자) 보정이 올바른 예측을 하지 않은 것으로 보인다.

다항 로지스틱 회귀모형과 랜덤 포레스트의 경우 상수항을 분류모형에 포함하지 않았기 때문에 모든 독립변수가 NA일 때 종속변수를 예측할 수 없지만, 의사결정나무의 경우 상수항이 분류모형에 포함되어 모든 독립변수가 NA일 때 종속변수를 예측할 수 있게 된다. 예를 들면, 67차 조사에서 의사결정나무를 이용하여 (응답유보층 + 안철수 후보 지지자)를 예측하려는 경우 모든 독립변수가 NA일 때 이재명 후보를 예측하게 된다. 즉, 모든 질문에 응답하지 않은 응답자가 이재명 후보를 지지할 것이라고 예측하게 된다.

다항 로지스틱 회귀모형의 경우 교호작용을 분류모형에 포함하지 않았으나 의사결정나무와 랜덤 포레스트의 경우 교호작용이 분류모형에 포함되어 하나의 독립변수가 다른 독립변수의 종속변수에 대한 효과에 영향을 주게 된다.

따라서 세 방법의 결과가 다를 뿐만 아니라 예측의 정확도 역시 달라진 것으로 추측된다.

제 4 장

결론

본 연구에서는 NBS에서 실시된 53~67차 선거여론조사 자료를 사용하여 응답유보층을 분류함으로써 득표율 예측의 정확도를 높이는 데 기여할 수 있는지를 살펴보았다. 응답유보층 분류의 방법으로는 다항 로지스틱 회귀, 의사결정나무와 랜덤 포레스트 방법을 고려하였고, 각 방법에 따라 결과가 어떻게 달라지는지를 살펴보았다.

응답유보층을 보정해 득표율을 예측한 결과, 단순보정의 결과가 분류모형을 이용한 보정의 결과보다 더 정확한 예측 결과를 보여주었다. 여론조사 결과와 비교했을 때, 최종 득표율과의 격차는 줄어들었지만 마지막 조사의 예측순위가 실제와 뒤바뀐 결과를 보여주었다. 이는 여론조사 공표금지 기간에 일어난 윤석열 후보와 안철수 후보의 단일화로 인해 안철수 후보를 지지했던 유권자들이 실제 투표에서 어떤 후보를 지지하게 될지 파악할 수 없게 되었기 때문이다.

본 연구에서 사용할 NBS 전화조사 자료 역시 모든 회차가 단일화가 이뤄지기 전에 조사가 완료되어 응답유보층을 (응답유보층 + 안철수 후보 지지자)로 변경하여 분석을 진행하였다.

(응답유보층 + 안철수 후보 지지자)를 보정해 득표율을 예측한 결과, 분류모형을 이용한 보정의 결과가 단순보정의 결과보다 더 정확한 예측 결과를 보여주었다. 다항 로지스틱 회귀모형을 이용한 경우, 마지막 조사의 예측순위가 실제 순위와 같은 결과를 보여주었다. 반면에 의사결정나무와 랜덤 포레스트를 이용한 경우,

마지막 조사의 예측순위가 실제와 뒤바뀐 결과를 보여주었다. 이는 의사결정나무를 이용한 경우 분류모형에 상수항이 포함되게 되어 모든 질문에 응답하지 않아도 한 후보자를 지지한다는 예측 결과가 나오기 때문이다. 또한, 의사결정나무와 랜덤 포레스트를 이용한 경우 분류모형에 교호작용이 포함되게 되어 하나의 독립변수가 다른 독립변수의 종속변수에 대한 효과에 영향을 주게 되기 때문이다.

따라서 본 연구에서는 상수항을 포함하지 않는 다항 로지스틱 회귀모형을 이용하여 응답유보층을 보정하는 것이 득표율 예측에 도움이 된다는 결론을 내릴 수 있었다.

하지만 본 연구의 결과는 제 20대 대통령 선거에 대한 여론조사 결과 중 NBS 전화조사의 결과를 이용하여 도출한 것이기 때문에 일반적인 현상이라고 할 수 없다. 또한, 본 연구에서와 같이 단일화로 인한 후보자의 사퇴 등이 득표율 예측에 영향을 줄 수 있다.

따라서 향후 여론조사의 결과 및 선거예측의 정확성 제고를 위해 응답유보층 보정에 대한 다양한 연구가 필요하다고 판단된다.

참고 문헌

류제복. (2019). 선거예측의 대안적 방법 - 2017년 대통령선거를 중심으로, 청주대학교 산업과학연구 제 36권 2호 : 8.

곽은선. and 김영원. (2022). 현행 선거여론조사 방법의 정확성과 선거결과 예측 가능성, 조사연구 제 23권 1호 : 131-153.

김정훈., 김지연., 권은혜., 김혁. and 강현철. (2017). 여론조사를 통한 선거예측에서 무응답층 보정의 효과, 한국자료분석학회 제 19권 1호 : 107-117.

박민규. (2011). 여론조사방법론 개선에 대한 논의, Issue BRIEF.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). Classification and Regression Trees. Wadsworth.

Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32.