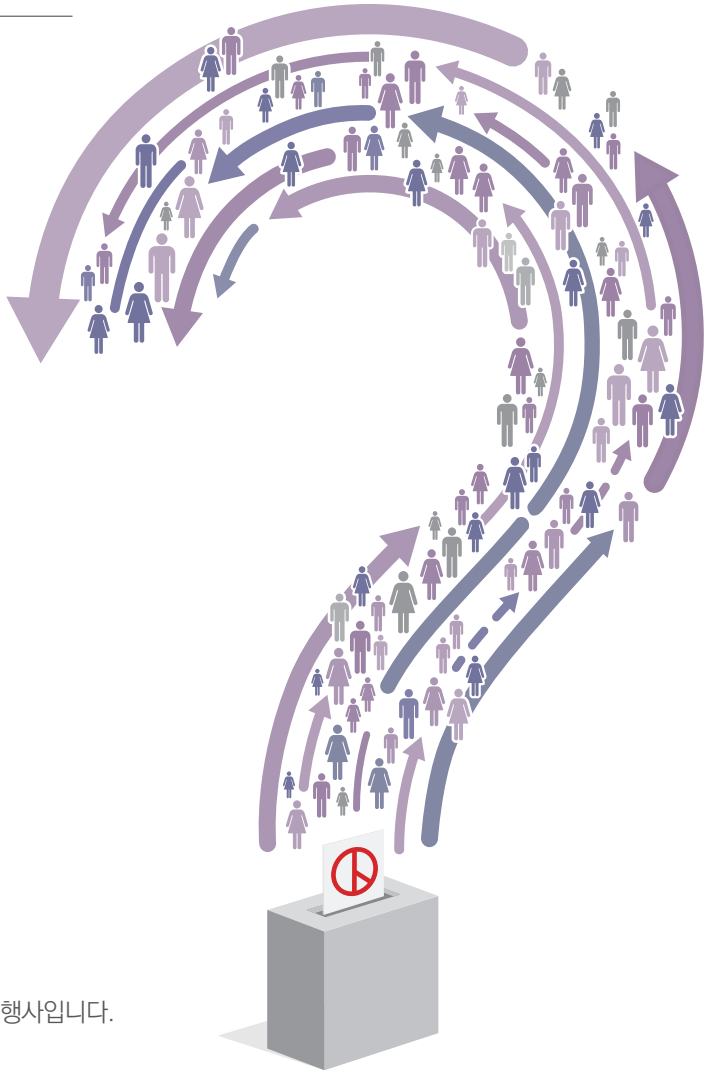
SNUAC
Seoul National University Asia Center
서울대학교아시아연구소

제8회

KOSSDA
데이터 페어

선거자료의 이해와 활용

선거여론조사 똑 소리 나게 활용하기



2020. 2. 5. (수)

14:00-17:00

서울대 아시아연구소 삼익홀

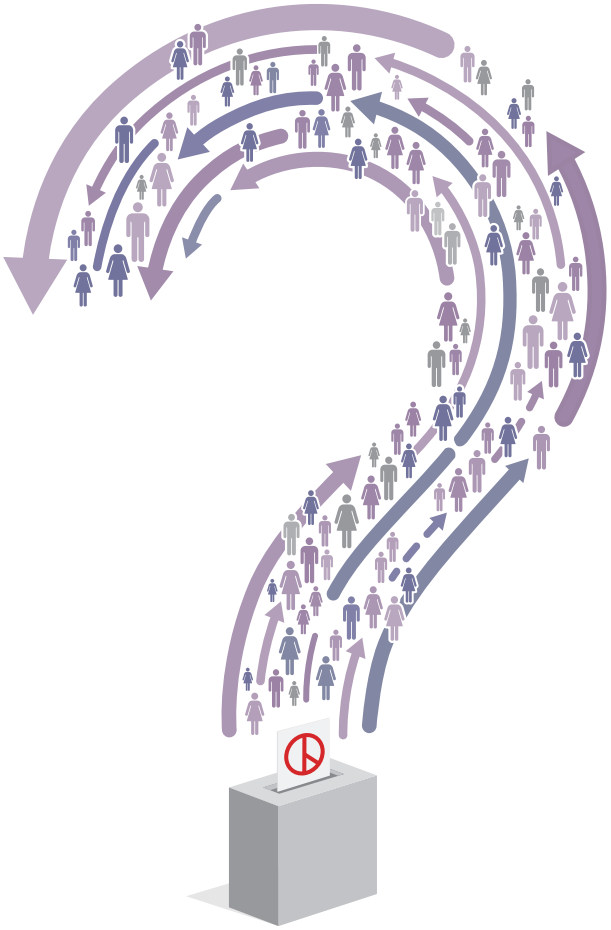
KOSSDA 데이터 페어는 서울대 사회과학대학
'미래 기초학문분야 기반조성사업'의 지원을 받아 개최되는 행사입니다.

제8회

KOSSDA데이터 페어

선거자료의 이해와 활용

선거여론조사 똑 소리 나게 활용하기



1부: 국내외 선거자료 소개

1. **국내 선거조사의 빛과 그림자**
   김석호 교수 (서울대 사회학과)

2. **선거자료 활용하기**
   하상응 교수 (서강대 정치외교학과)

2부: 선거자료 활용사례

1. **선거패널조사의 활용**
   길정아 박사 (고려대 정부학연구소)

2. **선거집계자료와 선거연구의 확장**
   박원호 교수 (서울대 정치외교학부)

3. **파이썬을 이용한 간단한 여론조사 분석기**
   황준식 연구원 (넥슨코리아)

SNUAC
Seoul National University Asia Center
서울대학교아시아연구소

SNUAC
Seoul National University Asia Center
서울대학교아시아연구소

1부 국내외 선거자료 소개

1. 국내 선거조사의 빛과 그림자

김석호 교수 (서울대 사회학과)

국내 선거자료의 빛과 그림자

김석호 교수 (서울대학교 사회학과)
한상효, 신규호 (서울대학교 사회발전연구소)

SNUAC
Seoul National University Asia Center
서울대학교 아시아연구소

목차

2

국내 선거자료

3

선거 전후 자료: 조사목적

- 선거 이전이나, 종료 후 해당 선거와 관련된 사람들의 태도, 행위 등을 파악
- 단순히 특정 후보의 승패보다 정치 행동과 태도에 관심을 두며 측정
  - 투표 참여에 대한 태도
  - 특정 정당을 지지하는 사람들이 공유하는 특징
  - 이슈가 되는 정책에 대한 태도
  - 특정 정치인에 대한 호감도 등
- 사회인구학적, 사회경제적, 사회 관계 등 다양한 변수를 함께 측정하여 학술적으로 활용 가능한 데이터를 구성

4

선거 전후 자료: 조사내용

예시) 2017 정치와 민주주의에 관한 의식 조사 (사회발전연구소)

- 응답자의 특성(성별, 연령, 거주지역, 소득, 주관적 계층의식, 종교, 혼인여부, 직업 등)
- 선거 관련 문항(투표 참여 여부 및 불참 이유, 투표 후보, 과거 투표 이력 등)
- 정치 태도(관심, 효능감, 투표에 대한 태도, 지지 정당, 이념 성향 등)
- 사회 참여(정치 대화 정도, 결사체 참여, SNS 활용 정도, 촛불·태극기 집회 참여 여부 등)
- 정치 현안에 대한 태도(국정운영평가, 정치온도계, 정책 별 찬반 정도 등)
- 기타 변수들(경제전망, 사회 신뢰, 감정 상태, 낙관주의적 태도 등)

5

선거 전후 자료: 조사설계

예시) 2017 정치와 민주주의에 관한 의식 조사 (사회발전연구소)

구분	주요 내용
모집단	• 전국 만 19세 이상 성인남녀(2017년 5월 9일 투표권이 있었던 사람에 한정)
표본크기	• 1,200명
표본오차	• ± 2.2% 포인트 (95% 신뢰수준)
표집방법	• 16개 광역시도별로 주민등록인구현황(2017년 4월 기준)에 따라 비례배분 후 층화 확률 비례계통추출
조사도구	• 구조화된 설문지
조사방법	• 대면면접조사(face to face interview)

6

선거 전후 자료: 주요 생산 기관 및 생산 자료(ㄱㄴㄷ순)

- EAI 동아시아연구원 (선거 패널조사)
- 서울대학교 사회발전연구소 (정치와 민주주의에 관한 의식 조사)
- 서울대학교 정치커뮤니케이션센터 (선거 온라인 패널조사)
- 서울대학교 한국정치연구소 (선거에 대한 국민의식 조사)
- 성균관대학교 서베이리서치센터 (한국종합사회조사, KGSS)
- 아산정책연구원 (총선대선패널조사, 월례 여론조사)
- 중앙선거관리위원회 (유권자 의식조사, 선거에 관한 여론조사)

7

선거 전후 자료: 주요 자료 목록

선거	조사기관	자료명
2012년 총선/대선	동아시아연구원	EAI 총선·대선 패널조사(KEPS)
	아산정책연구원	총선대선패널조사(1-7차)
	중앙선거관리위원회	제19대 국회의원선거 유권자 조사
	중앙선거관리위원회	제18대 대통령선거 관련 유권자 의식 조사
	한국정치연구소	2012년 정치와 민주주의에 관한 의식 조사
2014년 지선	한국정치연구소	2014년 지방선거에 대한 국민의식조사
	정치커뮤니케이션센터	제6회 전국동시지방선거 온라인 패널조사
	중앙선거관리위원회	제6회 전국동시지방선거 유권자 의식조사
2016년 총선	정치커뮤니케이션센터	제20대 국회의원 선거 온라인 패널조사
	중앙선거관리위원회	제20대 국회의원선거 유권자 조사
	한국정치학회	20대 국회의원 선거 유권자 조사
2017년 대선	중앙선거관리위원회	제19대 대통령선거 관련 유권자 의식조사
	사회발전연구소	2017년 정치와 민주주의에 대한 의식 조사
2018년 지선	한국정치연구소	2018년 지방선거에 대한 국민의식조사
	중앙선거관리위원회	제7회 전국동시지방선거 유권자 의식조사

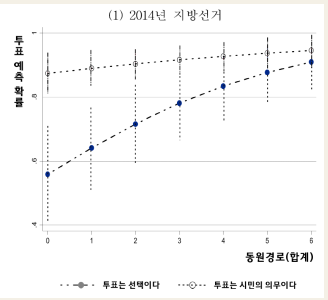
*선거 전후 자료는 한국사회과학자료원 (kossda.snu.ac.kr), 서베이리서치센터 (kgss.skku.edu), 한국사회과학데이터센터 (ksdcdb.kr) 등에서 다운로드 가능

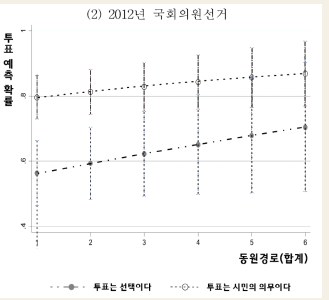
8

선거 전후 자료: 자료 활용 사례

사례) 김석호, 한수진. 2015. “지방선거와 국회의원선거에서 유권자들은 다른 이유로 투표하는가? 동원과 시민성의 선거 간 차별적 효과에 대한 연구”

- 동원(주변 사람들의 투표 권유)과 시민성(투표 의무 인식)은 모두 투표율을 높이는데 기여함
- 투표에 대한 책임감이 적은 지방선거에서, 시민성이 없는 사람들에게 동원의 효과가 강하게 나타남
- 즉, 동원이 없었다면 투표하지 않았을 사람들이 투표장에 가게 된 것임





9

여론조사 자료: 개요

- 특정 시기 유권자들의 의견을 파악하기 위한 도구
- 응답자 선정 및 문항 구성 과정에서 대표성과 객관성을 중시
- 선거 전후 사회조사에 비해 상대적으로 짧은 기간동안 이루어짐
- 여론조사 기관과 여론조사 결과는 여론조사심의위원회에서 관리
- 한국에서 수행되는 모든 선거여론조사는 여심위에 등록되어야 함

10

여론조사 자료: 조사설계의 특징

- 조사방식: 유선, 무선전화, 인터넷, 스마트폰 어플리케이션 조사 등 다양
- 표본추출방식: 임의전화걸기(RDD, Random Digit Dialing)를 주로 활용
  - 유·무선 전화 비율의 적절성, 응답률 문제, ARS의 신뢰성 등에 대해 논쟁이 있음
- 선거 및 지역구 크기에 따라 최소 표본 수가 정해져 있음
  - 대통령선거 1,000명; 광역단체장 800명, 시군구 500명, 지역구의회 300명 등
- 조사 편향과 응답률을 높이기 위해 다양한 질문 방법을 활용함
  - 지지하는 정당이 있는지? ↔ 호감을 가진 정당이 있는지?
  - 정당 제시 순서를 로테이션하여 묻는지, 지지정당이 없다고 할 때 다시 묻는지 등

11

여론조사 자료: 조사항목들의 특징

- 다음의 내용을 반드시 포함하여 자료를 공표해야 함

1. 조사의뢰자
2. 선거여론조사기관
3. 조사지역
4. 조사일시
5. 조사대상
6. 조사방법
7. 표본의 크기
8. 피조사자 선정방법
9. 응답률
10. 가중값 산출 및 적용 방법
11. 표본오차
12. 질문 내용

12

여론조사 자료: 2018년 제7회 지방선거 이전 1년 동안의 추세

- 총 조사기관: 26개
- 총 공표 조사 수: 1,648개 (지역 별 조사, 중복 공표 포함)
- 총 전국대상 조사 수: 207개 (중복 제외)

전체 응답자 별 조사 방법의 비율	* 총 응답자 305,155명
유선전화 – 면접원 조사	9.1%
유선전화 – 자동응답(ARS)	9.5%
무선전화 – 면접원 조사	36.8%
무선전화 – 자동응답(ARS)	42.2%
인터넷 조사	0.5%
스마트폰 어플리케이션 조사	1.9%

전체 응답자의 정치적 특성	* 총 응답자 305,155명
더불어민주당 지지자	49.4%
자유한국당 지지자	15.3%
정의당 지지자	5.3%

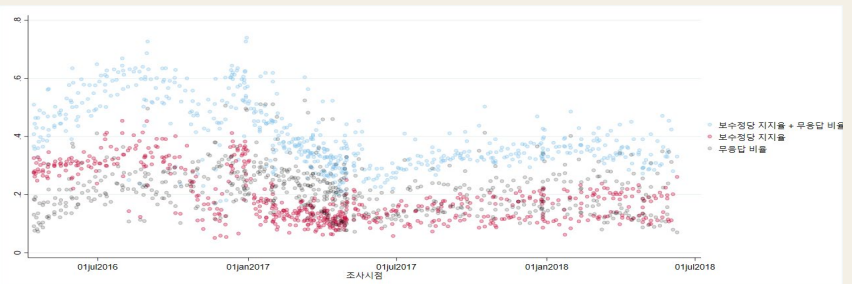
- 국민의당, 바른정당, 바른미래당, 민주평화당 등은 2018년 초 정당명이 바뀌어 제외함

13

여론조사 자료: 자료활용 사례

사례) 신규호. 2019. "'샤이 보수'는 실재하는가"

- 2016년 총선 ~ 2018년 지선까지의 전국 규모 여론조사 결과를 수집 및 분석한 결과
- 다양한 여론조사에서 보수정당 지지율에 편차가 나타나는데, 무응답 비율과 이를 합칠 경우 편차가 사라짐
- 이러한 편향은 ARS가 아니라, 조사원에 의해 수집된 응답일 때 발생, 이는 '사회적 바람직성 편향' 때문
- 즉, 면접원이 지지정당을 물을 때 응답하지 못하는 '샤이 보수'가 존재하며, 이것이 지지율 편차의 원인



14

출구조사 자료: 개요

- 1968년 미국 대선에서부터 CBS가 본격적으로 시작
- 한국에서는 1995년 제1회 지방선거에서 MBC가 도입
- 공직선거법으로 인해 투표소 50m 밖에서 시행되며 사전투표 대상자는 조사불가
- 20대 대선(2017)부터 후보 결정요인, 응답자 정치성향 등을 묻는 심층출구조사 도입

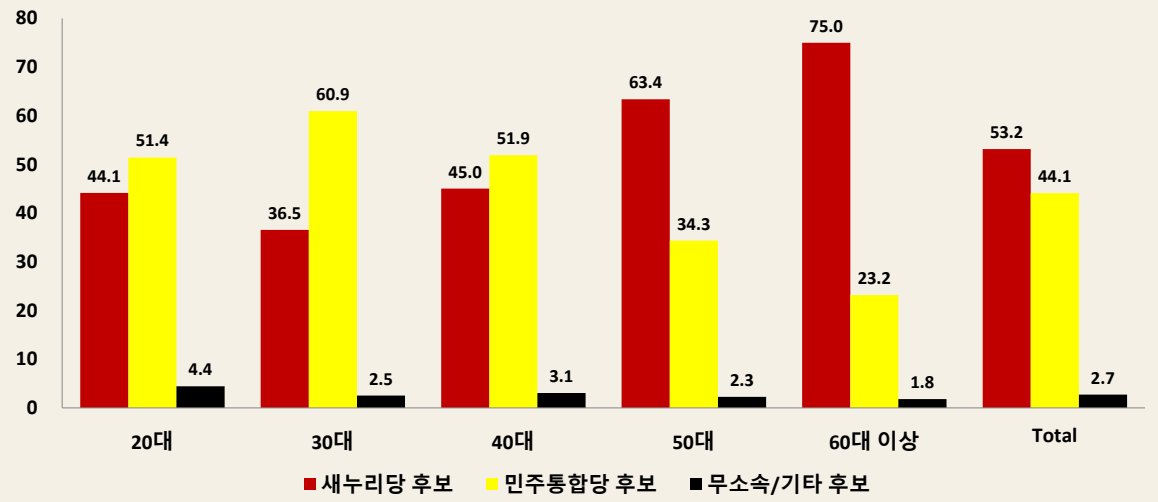
출구조사 자료: 자료활용 사례

사례) 박유성, 이선미. 2017. "'제19대 대선 출구조사 결과 분석과 의미"

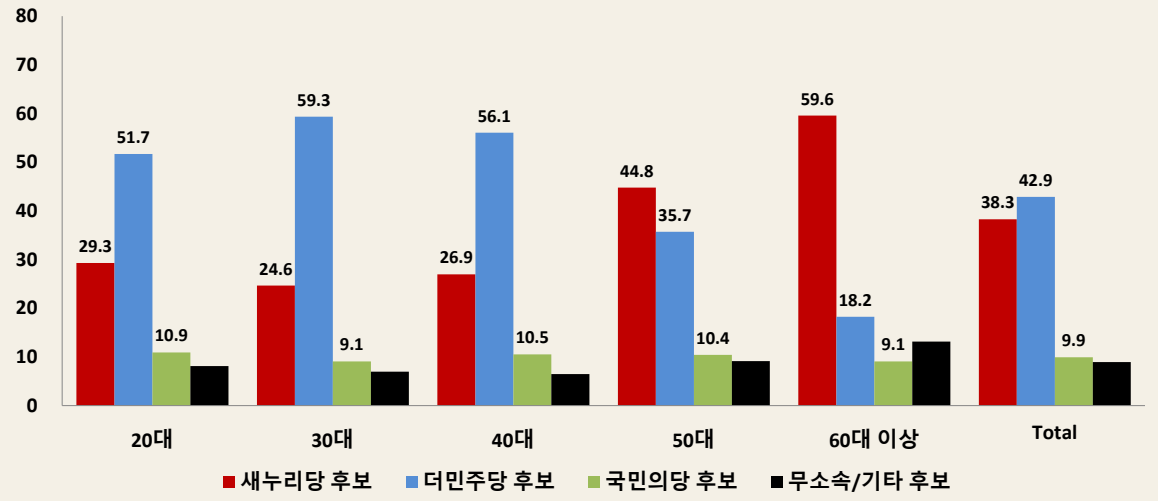
- 제19대 대선 출구조사는 후보자별 지지율 예측뿐만 아니라 지지율에 대한 인구사회학적 요인, 투표 결정요인, 차기 정부과제, 정치 성향 등을 함께 조사함
- 분석 결과, 투표자 연령과 지역 요인이 후보자 지지율에 결정적 영향을 미쳤으며, 투표자 정치 성향은 투표 결정요인과 차기 정부과제 우선순위에 가장 중요하게 영향을 준 것으로 나타남
- 문재인 후보 지지자의 정치 성향은 중도에서 약간 진보로 이동한 '중도적 진보'로 파악되었으며, 보수 성향 투표자의 분열과 중도성향 투표자의 지지가 문재인 후보의 당선 이유로 확인됨

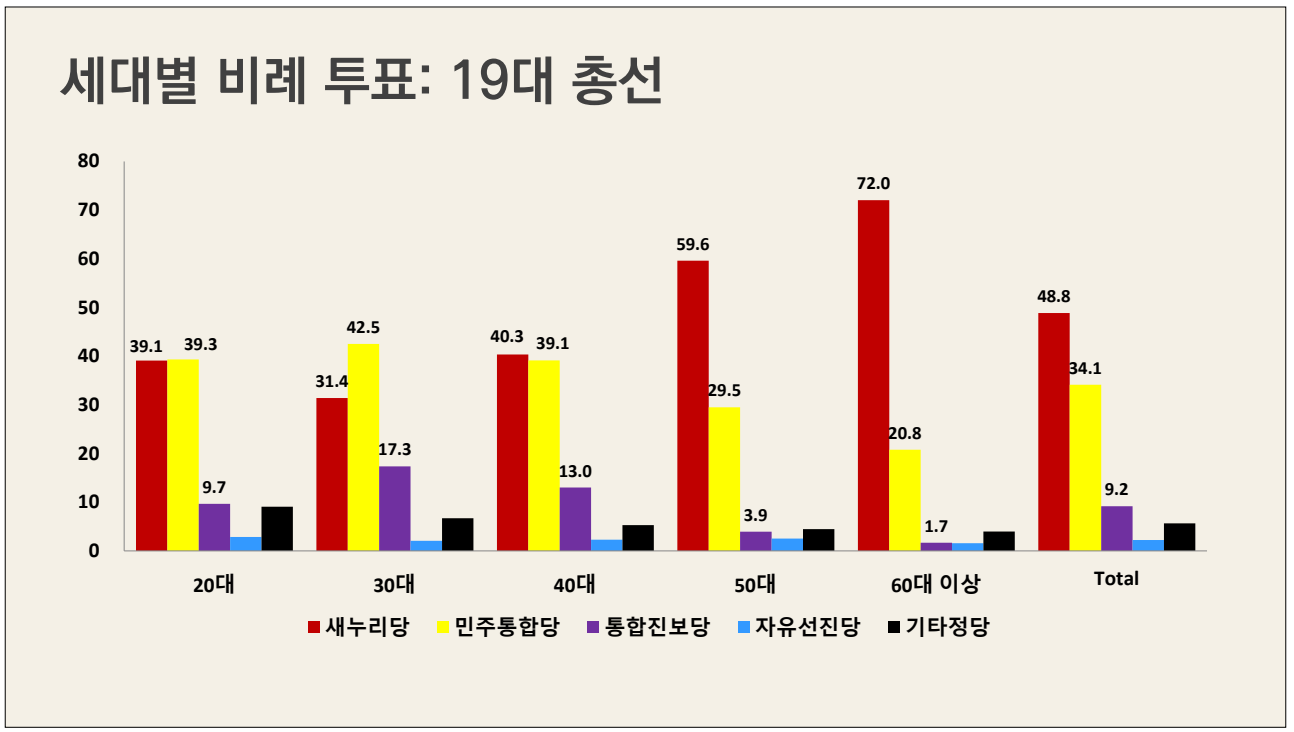
16

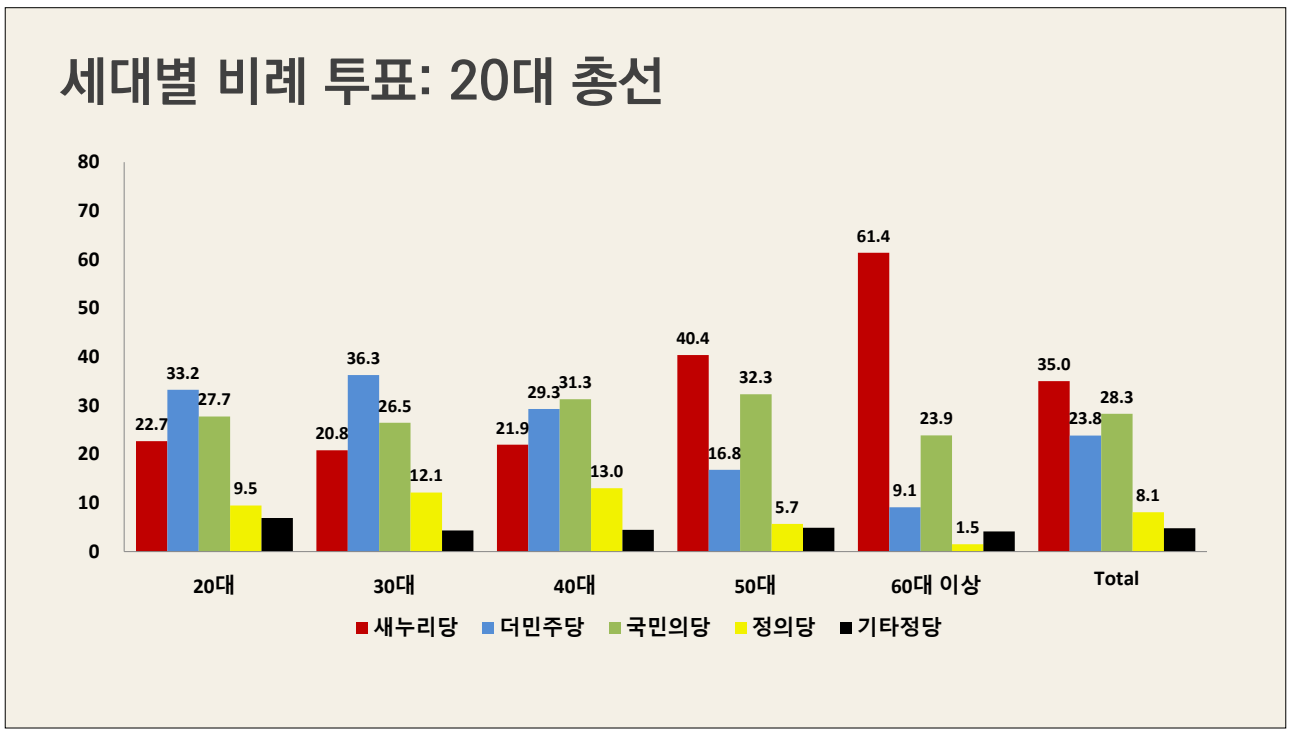
세대별 지역구 투표: 19대 총선



세대별 지역구 투표: 20대 총선




세대별 비례 투표: 19대 총선
80
70
60
50
40
30
20
10
0
39.1
39.3
9.7
31.4
42.5
17.3
40.3
39.1
13.0
59.6
29.5
3.9
72.0
20.8
1.7
48.8
34.1
9.2
20대
30대
40대
50대
60대 이상
Total
새누리당
민주통합당
통합진보당
자유선진당
기타정당


세대별 비례 투표: 20대 총선
80
70
60
50
40
30
20
10
0
22.7
33.2
27.7
9.5
20.8
36.3
26.5
12.1
21.9
29.3
31.3
13.0
40.4
16.8
32.3
5.7
61.4
9.1
23.9
1.5
35.0
23.8
28.3
8.1
20대
30대
40대
50대
60대 이상
Total
새누리당
더민주당
국민의당
정의당
기타정당

행정자료: 개요

- KOSSDA 시군구 통계, 2005-2013
- 지역 수준의 통계자료를 보다 편리하게 이용할 수 있도록 KOSSDA가 국가통계로 산출되는 지표 가운데 인구, 노동, 기반시설, 교통, 보건·복지, 정치·행정 분야의 35개 시군구 지표를 선정하여 시계열 자료로 재구성
- e-지방지표: KOSIS(국가통계포털)
- 부동산 실거래가 공개 시스템: http://rt.molit.go.kr/
- NHISS(국민건강보험공단 자료): https://nhiss.nhis.or.kr/bd/ab/bdaba015lv.do

21

행정자료

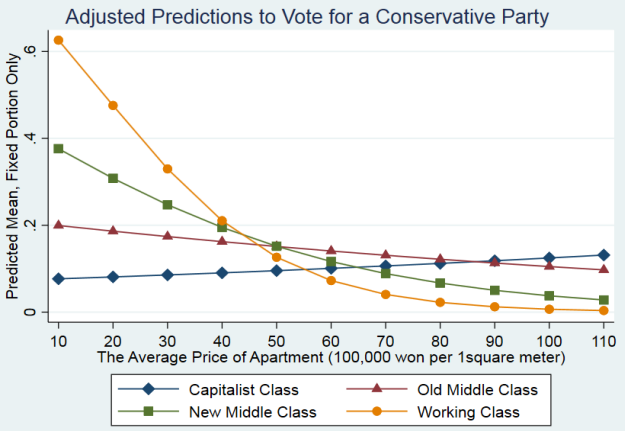
- 지역/선거구별 특성을 담은 자료로 선거자료와 결합하여 분석에 활용
- 대부분의 선거자료가 설문조사에 기반한 개인수준의 자료임에 비하여, 행정자료는 지역수준의 자료임으로 두 자료를 결합하면 다층모형(HLM)에 기반한 분석가능

22

행정자료: 자료활용 사례

사례) 한상효. 2017. “지역사회의 경제적 조건이 계급별 투표에 미치는 영향”

- 개인수준자료와 지역수준(시군구) 행정자료를 결합하여 계급(배반)투표의 원인에 대하여 탐색
- 개인수준자료인 KGSS 2010을 활용하여 계급 변수와 투표변수를 구성, 지역수준자료를 활용하여 시군구별 인구밀도, 시군구별 경제력 수준 및 불평등 수준 변수 구성
- 시군구별 인구밀도 변수는 KOSSDA자료 활용, 시군구별 경제력 수준과 불평등 수준 변수는 부동산 실거래가 공개 시스템을 활용하여 구성



해외 선거자료

ANES(American National Election Studies)

- 공식적으로 1977년 설립되었지만, 선거조사는 1948년부터 미시간대학교에서 실시되어 옴
- 전국단위 대표성 담보
- 주/시/선거구 단위 분석 불가능
- https://electionstudies.org/
- 홈페이지에서 회원 가입 후 자료 다운로드 가능

25

ANES(American National Election Studies)

- 조사주기: 전국단위 선거가 있는 해
- 조사방식: 1200-2500명 대상 면대면 방문 인터뷰

- 새로운 질문구성을 위한 전화 파일럿 인터뷰도 시행

- 흑인과 히스패닉의 경우, 통계적 정확성을 확보하기 위해 실제 인구비율보다 더 많이 포함되어 있음
- ANES 자료를 활용한 논문목록:

https://electionstudies.org/papers-documents/anes-bibliography/

26

ANES(American National Election Studies)

- 대표반복질문

Exhibit 1. Recurring question topic list for the ANES Time Series

I. Partisanship and attitudes towards parties
- Feelings about parties
- Party identification
- Closeness to parties
- Party performance
- Role of parties

II. Candidate and incumbent evaluations
- Feeling thermometers
- Candidate evaluations
- Traits of president, candidates
- Affect toward President, candidates
- Evaluation of Congressional candidates

IIA. Performance evaluations
- Performance of current president
- Retrospective evaluations
- Other government performance
- Candidate performance
- Congressional candidate performance

IIB. Contact with Congressional candidates/incumbents

III. Issues
- Social welfare issues
- Racial policy issues
- Economic issues
- Foreign relations
- Social issues
- Environment
- Federal spending

IV. Ideology and values
- Religious values
- Moral Traditionalism
- Equalitarianism and race
- Social trust and altruism
- Cognitive style
- Other values and predispositions

V. System support
- Trust in government
- Power of the federal government
- Efficacy and gov't responsiveness
- Patriotism
- Other system support

VI. Political participation and mobilization
- A. Civic participation
- B. Engagement and information
  - Interest
  - Discuss politics
  - Political knowledge
- C. Campaign activity
  - Respondent's campaign activity
  - Campaign contact/solicitation
- D. Registration, turnout, vote choice
  - Previous Presidential election
  - Vote intention
  - Turnout & registration
  - Vote choice
  - State primary/caucus

VII. Media
- Use of media

VIII. Social groups
- Group thermometers
- Group identification and closeness

IX. Personal and demographic data
- Education
- Employment status
- Religious identification
- Race/ethnicity
- DOB
- Mobility
- Income
- Social class
- Other personal and demographic data (labor union, home tenure, marital status, etc.)

27

ANES 자료: 자료활용 사례

- 사례: Barreto, Matt A., et al. "The racial implications of voter identification laws in America." American Politics Research 47.2 (2019): 238–249.

Article

The Racial Implications of Voter Identification Laws in America

American Politics Research
2019, Vol. 47(2) 238–249

DOI: 10.1177/1532673X18810012
journals.sagepub.com/home/apr
SAGE

Matt A. Barreto[1], Stephen Nuño[2], Gabriel R. Sanchez[3], and Hannah L. Walker[4]

Abstract
Over 40 states have considered voter identification laws in recent years, with several adopting laws requiring voters to show a valid ID before they cast a ballot. We argue that such laws have a disenfranchising affect on racial and ethnic minorities, who are less likely than Whites to possess a valid ID. Leveraging a unique national dataset, we offer a comprehensive portrait of who does and does not have access to a valid piece of voter identification. We find clear evidence that people of color are less likely to have an ID. Moreover, these disparities persist after controlling for a host of relevant covariates.

Keywords
voter ID laws, racial and ethnic politics

Introduction

Early challenges to voter identification laws equated them with poll taxes, given it costs money to obtain identification through the department of motor

[1]University of California, Los Angeles, USA
[2]Northern Arizona University, Flagstaff, USA
[3]University of New Mexico, Albuquerque, USA
[4]Rutgers University–New Brunswick, NJ, USA

Corresponding Author:
Hannah L. Walker, Assistant Professor of Political Science, Rutgers University, 89 George Street, New Brunswick, NJ 08901-8554, USA.
Email: hlwalker@polisci.rutgers.edu

28

CCES(Cooperative Congressional Election Study)

- YouGov에서 운영하는 50,000 명 이상이 참여하는 인터넷 선거 조사
- 선거가 있는 해에는 선거 전, 선거 후 2회 조사를 실시하며, 선거가 없는 해에는 1회만 실시
- 전국단위 대표성 담보
- 일부 주/시/선거구 단위 분석 가능
- https://cces.gov.harvard.edu/
- 홈페이지에서 자료 다운 가능

29

CCES(Cooperative Congressional Election Study)

- 주요 참여방식

-2018: 데스크탑(35%), 스마트폰(56%), 타블렛(9%)

-2012: 데스크탑(90%)

- 대표반복질문

- 미디어 및 소셜미디어 이용, 경제 평가

- 사회지도층(주지사, 대법원, 국회의원 등) 평가

- 사회 주요이슈(총기규제, 임신중단, 이주 등) 평가

- 자료 활용 사례: What Do Swing Voters Think? Meet @American__Voter

https://www.nytimes.com/2020/01/20/opinion/twitter-democratic-debate.html

30

CCES 자료: 자료활용 사례

- 사례: Hall, Andrew B., and Daniel M. Thompson. "Who punishes extremist nominees? Candidate ideology and turning out the base in US elections." *American Political Science Review* 112.3 (2018): 509–524.

Who Punishes Extremist Nominees? Candidate Ideology and Turning Out the Base in US Elections

ANDREW B. HALL *Stanford University*
DANIEL M. THOMPSON *Stanford University*

Political observers, campaign experts, and academics alike argue bitterly over whether it is more important for a party to capture ideologically moderate swing voters or to encourage turnout among hardcore partisans. The behavioral literature in American politics suggests that voters are not informed enough, and are too partisan, to be swing voters, while the institutional literature suggests that moderate candidates tend to perform better. We speak to this debate by examining the link between the ideology of congressional candidates and the turnout of their parties' bases in US House races, 2006–2014. Combining a regression discontinuity design in close primary races with survey and administrative data on individual voter turnout, we find that extremist nominees—as measured by the mix of campaign contributions they receive—suffer electorally, largely because they decrease *their party's share of turnout in the general election, skewing the electorate towards their opponent's party. The results help show how the behavioral and institutional literatures can be connected. For our sample of elections, turnout appears to be the dominant force in determining election outcomes, but it advantages ideologically moderate candidates because extremists appear to activate the opposing party's base more than their own.*

"The key data is this, and it's important to reemphasize if only to shut up the useless, overpaid political consultants who idiotically babble about 'moving to the center' or 'compromising with the other side'...What matters is turning out our voters. That's it. The Democrats win when we fire up and turn out our base."
—Blog post on Daily Kos[1]

"Democrats cannot win elections without capturing the votes of independent-minded swing voters."
—Commentary in the Wall Street Journal[2]

INTRODUCTION

The current state of American politics, characterized by high degrees of legislative polarization, brinksmanship, and gridlock (e.g., McCarty, Poole, and Rosenthal 2006), has raised new questions about the interplay of ideology and electoral success in US elections. The study of candidate ideology and electoral performance in US elections can be crudely divided into two literatures that seem fundamentally at odds with each other. On one side is what we might call the institutional literature, which uses election data to suggest that there is an electoral advantage for moderate candidates (e.g., Ansolabehere, Snyder, and Stewart 2001; Canes-Wrone, Brady, and Cogan 2002; Erikson 1971; Erikson and Wright 2000; Hall 2015). This literature is often associated with the idea that candidates must appeal to "swing voters" to win office. On the other side is what we might call the behavioral literature, which uses survey evidence to suggest that many voters are uninformed and ideologically inconsistent, casting doubt on whether swing voters are relevant or whether they even exist at all (e.g., Campbell et al. 1960; Converse 1964; Lenz 2012; Miller and Stokes 1963). This latter literature is often associated with the idea that turnout among the parties' bases determines election outcomes.[3] In the strongest version of this claim, voters are rigid partisans, "campaigns consist in large part of reminding voters of their partisan identities—'mobilizing' them to support their group at the polls," and more moderate candidates do no better than more extreme candidates because "election outcomes are essentially random choices among the available parties" (Achen and Bartels 2016, 311–312).

This disagreement rages on in the popular press, too, where pundits and campaign practitioners debate the relative merits of hypothetical moderate candidates who capture swing voters or ideologically committed

Andrew B. Hall is an Assistant Professor in the Department of Political Science at Stanford University, Stanford, CA 94305-6044 (andrewbenjaminhall@gmail.com), http://www.andrewbenjaminhall.com.
Daniel M. Thompson is a Ph.D. Student in the Department of Political Science at Stanford University, Stanford, CA 94305-6044 (danmckinleythompson@gmail.com), http://www.danmckinleythompson.com.
For helpful discussion, the authors thank Avi Acharya, Bob Erikson, Jim Fearon, Anthony Fowler, Stephen Pettigrew, Kevin Quinn, Ken Shotts, Brad Spahn, Danielle Thomsen, and participants of the MIT American Politics Conference and the Emory Institutions and Lawmaking Conference. For data, the authors thank Shigeo Hirano and Jim Snyder. For guidance using voter file data, the authors especially thank Brad Spahn. All remaining errors are the authors' sole responsibility. Replication files are available at the American Political Science Review Dataverse: https://doi.org/10.7910/DVN/9ZYFBX.
Received: May 18, 2017; revised: November 27, 2017; accepted: January 29, 2018. First published online: March 7 2018.

[1] http://www.dailykos.com/story/2014/11/5/1342347/-CRUSH-the-GOP-don-t-compromise-with-em-how-to-win-in-2016-and-what-not-to-do.

[3] Hill (2016) is one interesting article at the intersection of these literatures. The article combines administrative data on individual voter turnout with precinct-level vote returns to estimate swing voting. Under the assumptions of the article's model, partisan turnout is

CSES(Comparative Study of Electoral System)

- 1994년부터 진행 중인 전세계 국가 선거에 대한 공동 연구 프로젝트
- 미시건 대학교와 라이프니츠 사회과학연구의 공동 기여로 운영됨
- https://cses.org/

32

CSES(Comparative Study of Electoral System)

Macro Variables
- Political system characteristics
- Aggregate country-level data

District Variables
- Vote share and turnout
- Seats, candidates, and party lists

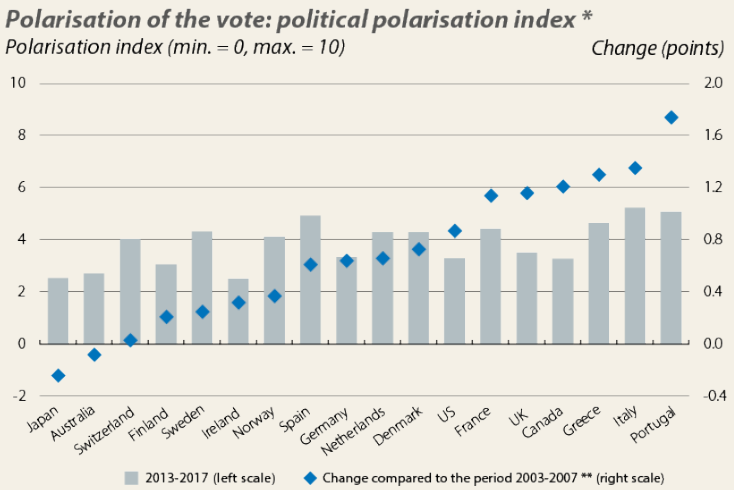
Micro Variables
- Voting items
- Module theme items

- 국가, 지역구, 유권자 등 세 다층 구조 데이터
- 매 5년 마다 새로운 모듈을 조사, 현재는 5차
  - 모듈1: 선거제도, 정치적 인식, 사회적 균열구조
  - 모듈2: 투표의 의미로서 책임성과 대표성
  - 모듈3: 후보 대립의 유의미성과 선거 경쟁의 통합성
  - 모듈4: 분배 정치와 사회보장에 대한 태도
  - 모듈5: 정치 엘리트와 '외집단'에 대한 태도

33

CSES(Comparative Study of Electoral System)

사례) Canals & Aldecoa, 2019. "The Causes of Political Polarization"



- 회색 지표가 높을수록 국회 내 정당 간 의견의 차이가 강한 것으로, 정당 간 양극화가 심한 것으로 이해
- 파란 점은 2003-07에 비해 정당 양극화가 얼마나 달라졌는지를 의미
- 그리스, 이탈리아, 포르투갈 등의 변화로 미루어 보아 금융위기와 정치의 양극화 간 일정한 관계가 유추될 수 있음

34

국내 선거자료의 문제점

35

국내 선거 전후 자료

1. 선거전후자료 구축 시급
   - 미국의 ANES, CCES, CSES처럼 동일한 목적과 내용, 그리고 조사 체계를 통해 지속적으로 축적되는 자료가 존재하지 않음
   - 문제는 예산, 누가 돈을 내지? 한국연구재단?
   - 국제 비교 자료 없음
   - 정치학 주도의 선거 연구 네트워크 필요

2. 자료 공유와 확산의 문제
   - 서울대학교 한국정치연구소와 사회발전연구소를 제외하고 외부 연구자들에게 공개하는 기관, 연구자 없음
   - 한국인의 투표 행태에 대한 확정적 지식이 존재하는가?

3. 표본의 대표성 문제
   - 자료수집 과정에 대한 정보 접근 불가능
   - 표본의 대표성 문제는 대부분 저비용과 관련되어 있음
   - 할당표집과 표본대체

4. 낮은 응답률

5. 학제간 연구의 부재

36

여심위 자료

1. 여심위는 과연 필요한가?
   - 민주주의의 기본 원칙에 관한 문제
   - 정치여론조사의 질은 여심위로 인해 개선되고 있는가?
   - 혁신적 조사 기법의 도입에 방해가 되는 것은 아닌가?
2. 표집틀과 표본추출의 문제
   - 안심번호 사용이 관건
3. 응답률 계산 방식
4. 전국 단위에서의 대표성이 모든 것을 해결하지는 않는다
   - 특정 선거구 단위
   - 집단별 비교 시 대표성
5. 선거법
   - 안심번호 사용 금지
   - 여론조사 공표 금지 기간
   - 전화조사가 기준?
6. 학술적 분석이 가능한가?
   - 20개 남짓의 변수
   - 조사기관마다 다른 목표모집단, 표집틀, 표본추출, 조사 원칙 등
7. 자료 공개

37

출구조사 자료

1. 연구 목적의 공개는 불가능한가? 공개한다면 다양한 방식으로 활용 가능
2. 개인 정보 보호와 문항 수의 딜레마(성, 연령대, 선거구, 지지후보 및 정당이 전부)
3. 높아지는 사전 투표율

38

행정자료

1. 법정동과 행정동이라는 두 가지 기준에 따른 자료들의 난립
2. 지역의 경제력을 보여줄 수 있는 주민 소득관련 자료의 부재
3. 지방자치단체별 주민의 중위소득을 공개하는 미국과 다르게 한국에는 주민소득관련자료 부재
4. GRDP는 광역자치단체 수준에서만 존재
5. NHISS 자료 활용 가능

39

MEMO

MEMO

SNUAC
Seoul National University Asia Center
서울대학교아시아연구소

1부 국내외 선거자료 소개

2. 선거자료 활용하기

하상응 교수 (서강대 정치외교학과)

선거자료 활용하기

제 8회 KOSSDA 데이터 페어
2020. 02. 05.
하상응 (서강대학교)

목차

선거자료의 종류

여론조사 자료
- 선거 전 여론조사
- 출구조사
- 선거 후 여론조사

투표행태/선거 자료
- 중앙선거관리위원회 (집합자료)

행정 자료 (보조 자료)
- 시군구/읍면동 단위

기타 비정형 자료
- 후보자 공약 목록
- 후보자 선거 운동 정보 (사진, 동영상, 음성)

여론조사 자료: 선거전 자료

선거 결과 예측 목적

과제: 대표성이 담보된 표본(representative sample)
- 표집 방식(sampling): 확률 표집(probability sampling), 편의 표집(convenience sampling)
- 조사 방식(modes of survey): 면접 조사, 전화 조사, 온라인 조사 등
- 조사 관행: 재접촉(recontacting), 표본 교체(sample replacement)
- 선거 종류

문제점: 표본의 대표성

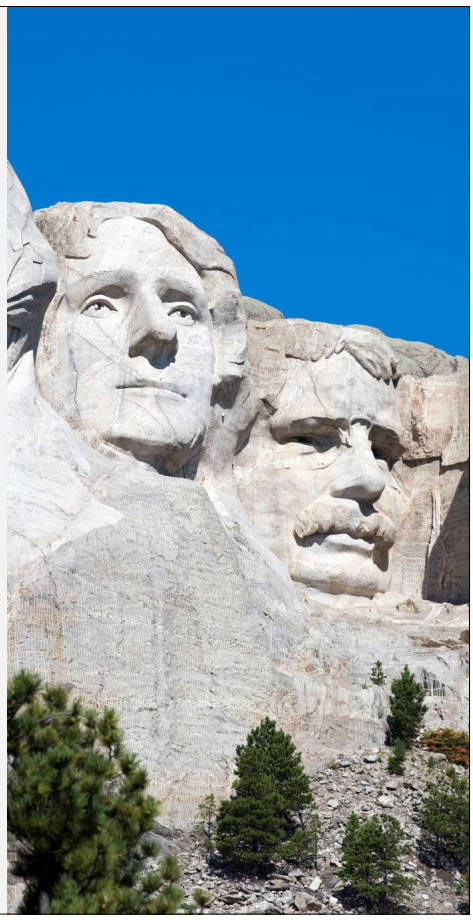
전국 단위에서 대표성이 있는 표본이 한 특정 선거구 단위에서의 대표성을 보장해 주지 않음

한 선거구 단위에서의 경쟁에 관심을 갖는다면, 그 선거구 소속 유권자를 모집단으로 한 표본을 별도로 구성해 설문 조사를 해야 함 (지역구 253개)

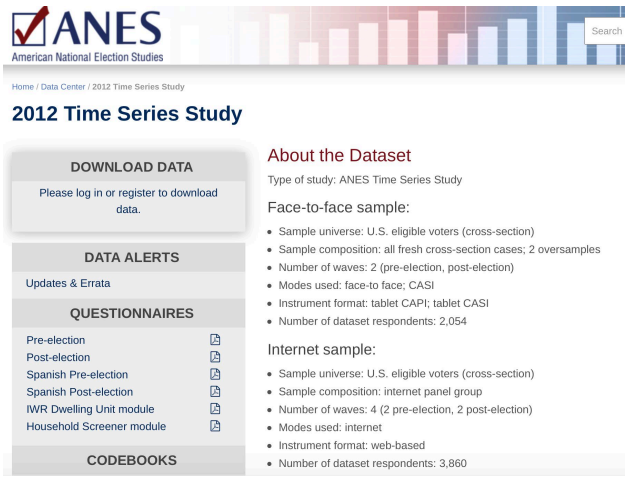
대통령 선거가 아닌, 국회의원 선거 및 지방자치단체장 선거가 갖는 큰 문제점

문제점: 표본의 대표성

- 미국의 경우
- American National Election Studies (ANES)
  - 전국 단위의 대표성 담보
  - 주/시/선거구 단위 분석 불가능
- Cooperative Congressional Election Study (CCES)
  - 전국 단위의 대표성 담보
  - 엄청나게 큰 표본을 만들어 일부 주/시/선거구 단위 분석 가능함



ANES



CCES

Cooperative Congressional Election Study

Contact

HOME Explore News FAQ People Publications

- CCES 2006-2018 Data/Guide
- CCES 2018 Data
- CCES 2017 Data/Guide
- CCES 2016 Data/Guide
- CCES 2010-14 Panel Study
- CCES 2015 Data/Guide
- CCES 2014 Data/Guide
- CCES 2013 Data/Guide
- CCES 2010-12 Panel Study
- CCES 2012 Data/Guide

WELCOME!

The CCES is a 50,000+ person national stratified sample survey administered by YouGov. Half of the questionnaire consists of Common Content asked of all 50,000+ people, and half of the questionnaire consists of Team Content designed by each individual participating team and asked of a subset of 1,000 people. In addition, several teams may pool their resources to create Group Content.

The survey consists of two waves in election years. In the pre-election wave, respondents answer two-thirds of the questionnaire. This segment of the survey asks about general political attitudes, various demographic factors, assessment of roll call voting choices, political information, and vote intentions. The pre-election wave is in the field from late September to late October. In the post-election wave, respondents answer the other third of the questionnaire, mostly consisting of items related to the election that just occurred. The post-election wave is administered in November.

In non-election years, the survey consists of a single wave conducted in the fall.

- Brace, P., Sims-Butler, K., Arceneaux, K., & Johnson, M. (2002). Public opinion in the American states: New perspectives using national survey data. *American Journal of Political Science*, 173-189.

Public Opinion in the American States: New Perspectives Using National Survey Data

Paul Brace Rice University
Kellie Sims-Butler Pennsylvania State University
Kevin Arceneaux Rice University
Martin Johnson Rice University

General measures of ideology and partisanship derived from national survey data concatenated to the state level have been extremely important in understanding policy and political processes in the states. However, due to the lack of uniform survey data covering a broad array of survey questions, we know little about how specific state-level opinion relates to specific policies or processes. Using the General Social Survey (GSS) disaggregated to the state level, we develop and rigorously test specific measures of state-level opinion on tolerance, racial integration, abortion, religiosity, homosexuality, feminism, capital punishment, welfare, and the environment. To illustrate the utility of these measures, we compare the explanatory power of each to that of a general ideology measure. We use a simulation to clarify conditions under which a national sample frame can produce representative state samples. We offer these measures to advance the study of the role public opinion plays in state politics and policy.

The public opinion-policy linkage is a crucial topic for democratic theorists and has preoccupied students of state government and politics for years. Without survey data at the state level, pioneering studies employed surrogates derived from demographic variables or simulations to judge the responsiveness of state policymaking to public preferences (Plotnick and Winters 1985; Weber and Shaffer 1972). Some ingenious studies also explore the causes and consequences of public opinion using national survey data disaggregated to subnational units (Gibson 1989, 1992, 1995; Miller and Stokes 1963; Norrander 2000).

Wright, Erikson, and McIver's research (1985) significantly advanced our understanding of the state public opinion and policy linkage by pooling 1976 through 1988 CBS/*New York Times* polls and disaggregating them to the state level to create reliable, stable, and valid measures of state ideology and partisanship (Erikson, Wright, and McIver 1993). A host of influential studies employ these measures (e.g., Hill and Hinton-Anderson 1995) to illustrate fundamental linkages between *general* mass political attitudes and the *general* choices of state policy makers. Yet, they represent only a first step in gauging the effects of opinion on state policy. The general nature of the ideology measure developed by Erikson, Wright, and McIver leaves open many remaining questions about how *specific* attitudes may influence *specific* political outcomes and processes in the states.

Paul Brace is Clarence L. Carter Professor of Political Science, Rice University, 6100 Main St., Houston, Texas 77005 (pbrace@rice.edu). Kellie Sims-Butler is Assistant Professor of Political Science, Pennsylvania State University, 107 Burrowes Building, University Park, PA 16802 (mycelle@rice.edu). Kevin Arceneaux is a Doctoral Candidate in Political Science, Rice University, 6100 Main St., Houston, Texas 77005 (kevina@rice.edu). Martin Johnson is a Doctoral Candidate in Political Science, Rice University, 6100 Main St., Houston, Texas 77005 (pressley@rice.edu).

문제점 (선거법 관련)

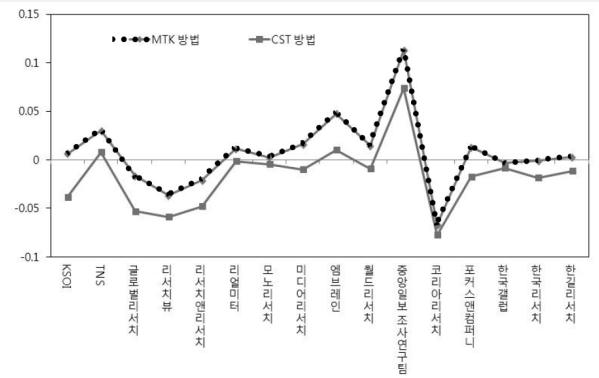
여론조사 공표 금지기간: 예측 정확성 낮춤

2020년 선거부터 만 18세 유권자(2002년 1월 1일부터 4월 16일 사이 출생자) 투표 가능: 모집단/표집틀 변화

준연동형 비례대표제: 여론과 선거 결과 간 불일치 가능성

여심위 자료 활용사례

최필선, 민인식. 2013. “18대 대통령 선거 여론조사의 기관별 정확성 측정 및 비교”



〈그림 2〉 MTK와 CST 방법에 의해 구한 조사오차 비교

- 두 가지 여론조사 예측 방법을 활용하여, 제18대 대통령선거에서 각 조사기관의 예측 결과가 실제 투표 결과에 비해 누구를 더 유리하도록 추정했는지 검증
- 조사오차가 클수록 박근혜 후보에게 편향된 조사결과를, 조사오차가 작을수록 문재인 후보에게 편향된 조사결과를 공표한 것임
- 즉, 중앙일보 조사연구팀은 박근혜의 지지율을 실제보다 높게, 코리아리서치는 문재인의 지지율을 실제보다 높게 발표한 것

12

여론 조사 자료: 출구조사

- 선거 결과 예측 목적 (단기)
- 대면 조사; 문항 수 제한됨
- 문제: 사전투표의 활성화

여론조사 자료: 선거후 자료

선거 결과 분석 목적

과제: 패널 자료 (panel; TSCS data)

- 선거 전 자료와 결합하여 사용하면 유권자의 선거 전후 태도 변화를 파악할 수 있음
- 투표 여부, 투표 선택을 제외한 현안에 대한 태도에 대한 답변은 선거 결과에 영향을 받을 수 있음

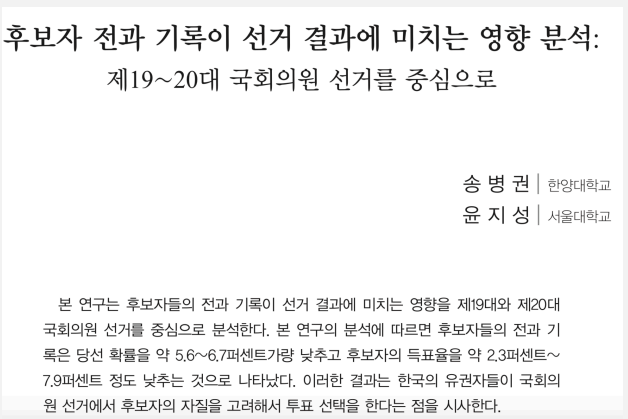
투표행태/선거 자료

중앙선거관리위원회

집합자료(aggregate data)

- 투표율, 득표율 (읍면동 및 투표구 별)
- 후보자 (성별, 연령, 학력, 소속 정당, 전과기록 등)

- 송병권, & 윤지성. (2016). 후보자 전과 기록이 선거 결과에 미치는 영향 분석: 제 19~ 20 대 국회의원 선거를 중심으로. *한국정치연구*, *25*(3). 85-107.



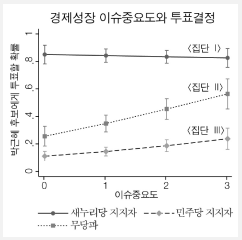
후보자 전과 기록이 선거 결과에 미치는 영향 분석:
제19~20대 국회의원 선거를 중심으로

송 병 권 | 한양대학교
윤 지 성 | 서울대학교

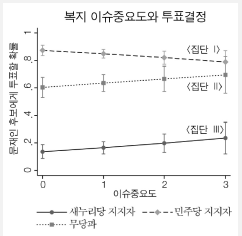
본 연구는 후보자들의 전과 기록이 선거 결과에 미치는 영향을 제19대와 제20대 국회의원 선거를 중심으로 분석한다. 본 연구의 분석에 따르면 후보자들의 전과 기록은 당선 확률을 약 5.6~6.7퍼센트가량 낮추고 후보자의 득표율을 약 2.3퍼센트~7.9퍼센트 정도 낮추는 것으로 나타났다. 이러한 결과는 한국의 유권자들이 국회의원 선거에서 후보자의 자질을 고려해서 투표 선택을 한다는 점을 시사한다.

선거자료 활용 사례

송진미, 박원호. 2014. “이슈선점과 정당일체감: 제18대 대통령선거를 중심으로”



- 경제성장이 중요한 이슈라고 응답한 무당파(집단II), 민주당 지지자(집단III)는 해당 이슈를 더 중요하다고 인식할수록 박근혜에게 투표했을 확률이 높아짐. 경제성장 이슈를 박근혜가 주도하여, 해당 이슈를 중요하게 여기는 경우 민주당 지지자도 박근혜에게 투표하게 됨



- 복지이슈가 중요하다고 응답한 사람 중 민주당 지지자(집단 I)는 복지가 중요한 이슈라고 생각할수록 문재인에게 투표할 확률이 줄어들었음.
- 그러나 복지가 중요한 이슈라고 응답한 무당파(집단II), 새누리당 지지자(집단III)는 해당 이슈를 더 중요하다고 인식할수록 문재인에게 투표했을 확률이 높아짐
- 집단 I과 집단 II, III 간의 모순되는 결과는 두 후보가 복지 이슈를 동등하게 선점했을 가능성을 시사

17

행정자료 (보조자료)

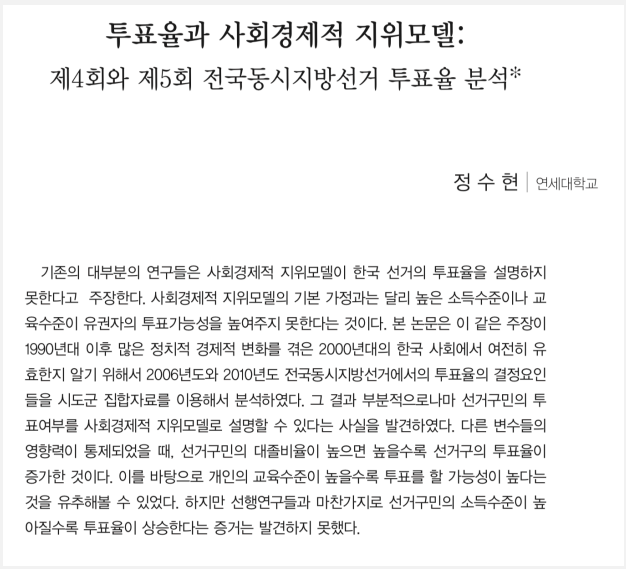
행정자료 + 선거 집합자료

- 분석단위: 시군구, 읍면동, 선거구 등
- 해석시 주의점: 생태학적 오류(ecological fallacy)

행정자료 + 선거 개인자료(설문조사)

- 분석단위: 시군구, 읍면동, 선거구 + 개인
- 분석방법: 다층모형(multi-level model), 위계선형모형(hierarchical linear model)

- 정수현. (2012). 투표율과 사회경제적 지위모델: 제 4 회와 제 5 회 전국동시지방선거 투표율 분석. *한국정치연구*. 21(1), 27-54.

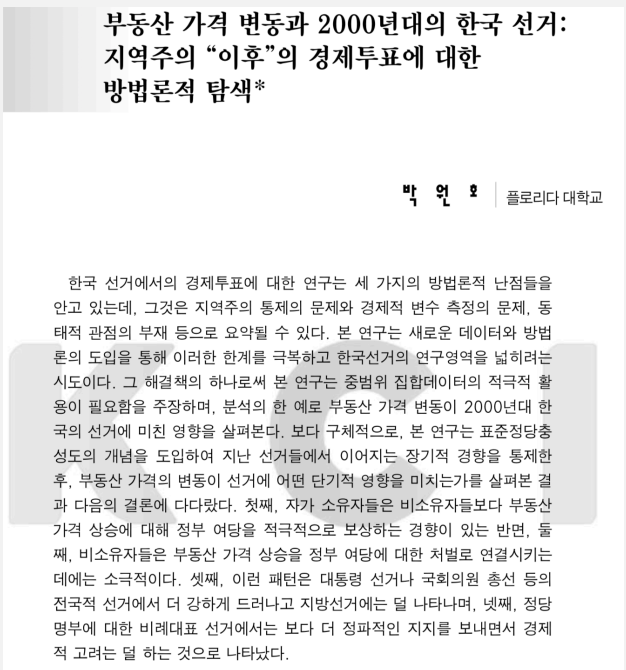


투표율과 사회경제적 지위모델:
제4회와 제5회 전국동시지방선거 투표율 분석*

정 수 현 | 연세대학교

기존의 대부분의 연구들은 사회경제적 지위모델이 한국 선거의 투표율을 설명하지 못한다고 주장한다. 사회경제적 지위모델의 기본 가정과는 달리 높은 소득수준이나 교육수준이 유권자의 투표가능성을 높여주지 못한다는 것이다. 본 논문은 이 같은 주장이 1990년대 이후 많은 정치적 경제적 변화를 겪은 2000년대의 한국 사회에서 여전히 유효한지 알기 위해서 2006년도와 2010년도 전국동시지방선거에서의 투표율의 결정요인들을 시도군 집합자료를 이용해서 분석하였다. 그 결과 부분적으로나마 선거구민의 투표여부를 사회경제적 지위모델로 설명할 수 있다는 사실을 발견하였다. 다른 변수들의 영향력이 통제되었을 때, 선거구민의 대졸비율이 높으면 높을수록 선거구의 투표율이 증가한 것이다. 이를 바탕으로 개인의 교육수준이 높을수록 투표를 할 가능성이 높다는 것을 유추해볼 수 있었다. 하지만 선행연구들과 마찬가지로 선거구민의 소득수준이 높아질수록 투표율이 상승한다는 증거는 발견하지 못했다.

- 박원호. (2009). 부동산 가격 변동과 2000 년대의 한국 선거: 지역주의" 이후" 의 경제투표에 대한 방법론적 탐색. *한국정치연구*. 18(3), 1-28.

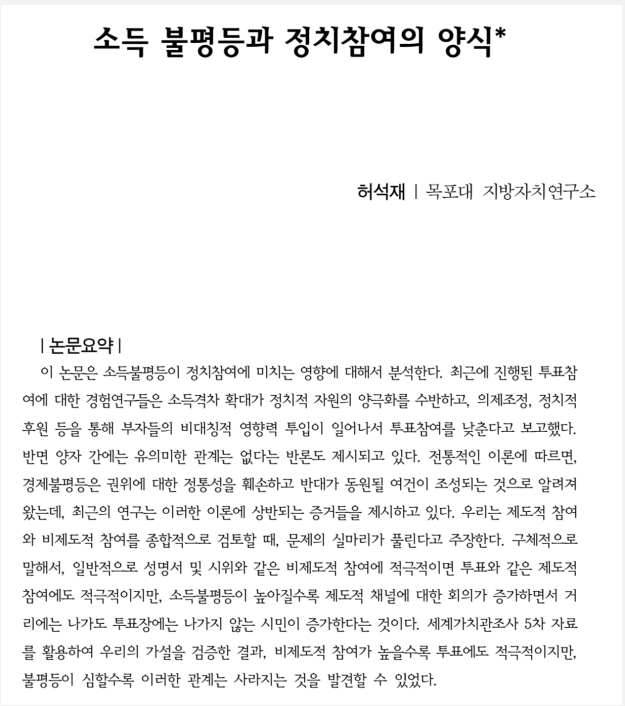


부동산 가격 변동과 2000년대의 한국 선거:
지역주의 "이후"의 경제투표에 대한
방법론적 탐색*

박 원 호 | 플로리다 대학교

한국 선거에서의 경제투표에 대한 연구는 세 가지의 방법론적 난점들을 안고 있는데, 그것은 지역주의 통제의 문제와 경제적 변수 측정의 문제, 동태적 관점의 부재 등으로 요약될 수 있다. 본 연구는 새로운 데이터와 방법론의 도입을 통해 이러한 한계를 극복하고 한국선거의 연구영역을 넓히려는 시도이다. 그 해결책의 하나로써 본 연구는 중범위 집합데이터의 적극적 활용이 필요함을 주장하며, 분석의 한 예로 부동산 가격 변동이 2000년대 한국의 선거에 미친 영향을 살펴본다. 보다 구체적으로, 본 연구는 표준정당충성도의 개념을 도입하여 지난 선거들에서 이어지는 장기적 경향을 통제한 후, 부동산 가격의 변동이 선거에 어떤 단기적 영향을 미치는가를 살펴본 결과 다음의 결론에 다다랐다. 첫째, 자가 소유자들은 비소유자들보다 부동산 가격 상승에 대해 정부 여당을 적극적으로 보상하는 경향이 있는 반면, 둘째, 비소유자들은 부동산 가격 상승을 정부 여당에 대한 처벌로 연결시키는 데에는 소극적이다. 셋째, 이런 패턴은 대통령 선거나 국회의원 총선 등의 전국적 선거에서 더 강하게 드러나고 지방선거에는 덜 나타나며, 넷째, 정당 명부에 대한 비례대표 선거에서는 보다 더 정파적인 지지를 보내면서 경제적 고려는 덜 하는 것으로 나타났다.

- 허석재. (2015). 소득 불평등과 정치참여의 양식. *한국정당학회보*, *14*(3), 41-67.



소득 불평등과 정치참여의 양식*

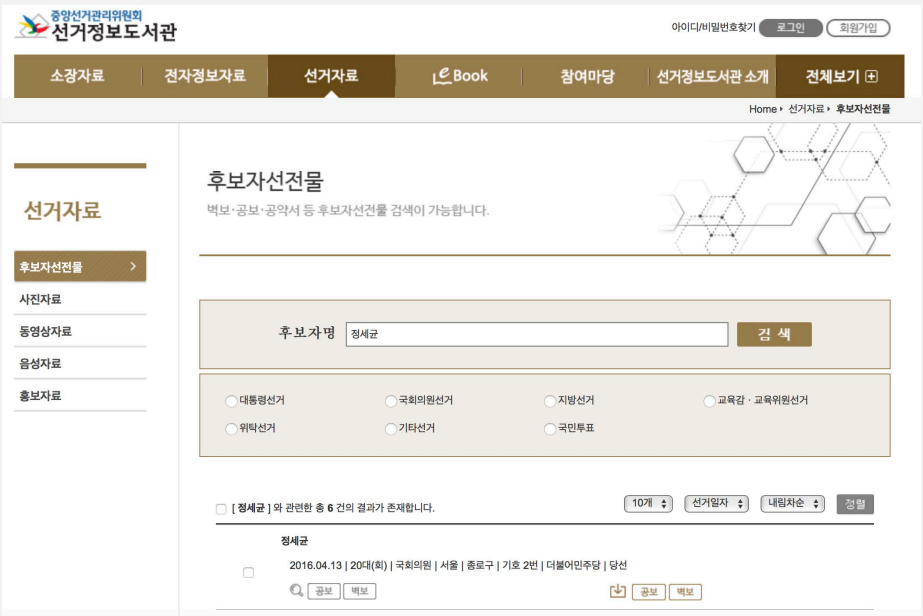
허석재 | 목포대 지방자치연구소

| 논문요약 |

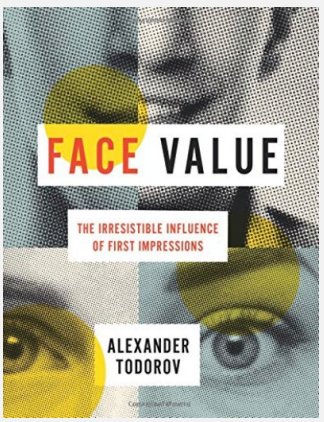
이 논문은 소득불평등이 정치참여에 미치는 영향에 대해서 분석한다. 최근에 진행된 투표참여에 대한 경험연구들은 소득격차 확대가 정치적 자원의 양극화를 수반하고, 의제조정, 정치적 후원 등을 통해 부자들의 비대칭적 영향력 투입이 일어나서 투표참여를 낮춘다고 보고했다. 반면 양자 간에는 유의미한 관계는 없다는 반론도 제시되고 있다. 전통적인 이론에 따르면, 경제불평등은 권위에 대한 정통성을 훼손하고 반대가 동원될 여건이 조성되는 것으로 알려져 왔는데, 최근의 연구는 이러한 이론에 상반되는 증거들을 제시하고 있다. 우리는 제도적 참여와 비제도적 참여를 종합적으로 검토할 때, 문제의 실마리가 풀린다고 주장한다. 구체적으로 말해서, 일반적으로 성명서 및 시위와 같은 비제도적 참여에 적극적이면 투표와 같은 제도적 참여에도 적극적이지만, 소득불평등이 높아질수록 제도적 채널에 대한 회의가 증가하면서 거리에는 나가도 투표장에는 나가지 않는 시민이 증가한다는 것이다. 세계가치관조사 5차 자료를 활용하여 우리의 가설을 검증한 결과, 비제도적 참여가 높을수록 투표에도 적극적이지만, 불평등이 심할수록 이러한 관계는 사라지는 것을 발견할 수 있었다.

기타 비정형 자료

- 후보자 포스터 및 공약집
- 후보자 토론회 영상


중앙선거관리위원회
선거정보도서관
아이디/비밀번호찾기
로그인
회원가입
소장자료
전자정보자료
선거자료
Book
참여마당
선거정보도서관 소개
전체보기
Home ▸ 선거자료 ▸ 후보자선전물
선거자료
후보자선전물
사진자료
동영상자료
음성자료
홍보자료
후보자선전물
벽보·공보·공약서 등 후보자선전물 검색이 가능합니다.
후보자명
정세균
검 색
대통령선거
국회의원선거
지방선거
교육감 · 교육위원선거
위탁선거
기타선거
국민투표
[정세균] 와 관련한 총 6 건의 결과가 존재합니다.
10개
선거일자
내림차순
정렬
정세균
2016.04.13 | 20대(회) | 국회의원 | 서울 | 종로구 | 기호 2번 | 더불어민주당 | 당선
공보
벽보
공보
벽보

후보의 인상 – 정치행태


FACE VALUE
THE IRRESISTIBLE INFLUENCE OF FIRST IMPRESSIONS
ALEXANDER TODOROV


페미니즘 서울시장
8
녹색당
신지예
DESIGN

Political Psychology

ISPP

Political Psychology, Vol. 38, No. 6, 2017
doi: 10.1111/pops.12403

Perceived Conflict and Leader Dominance: Individual and Contextual Factors Behind Preferences for Dominant Leaders

Lasse Laustsen
Aarhus University, Denmark

Michael Bang Petersen
Aarhus University, Denmark

Recent research finds that political candidates and leaders with dominant, masculine physical features are more preferred under conditions of conflict than of cooperation. Importantly, however, methodological limitations of past research have hindered the identification of whether this effect reflects that voters intuitively view (1) dominant leaders as more competent in solving problems of conflict, (2) nondominant leaders as more competent in solving problems of cooperation, or (3) both. In this article, we utilize recent advances in evolutionary psychology to form precise predictions on the nature of the underlying psychology and employ an unprecedented array of data types—including highly controlled experiments, natural experiments, and behavioral measures—to investigate the validity of these predictions. Using large approximately nationally representative surveys of 2,009 Poles and Ukrainians fielded during the Crimea crisis in 2014, we find that preferences for leader dominance are exclusively driven by the intuition that dominant leaders are better able to facilitate aggressive responses during social conflict and that these preferences are regulated by contextual conditions and individual predispositions related to such responses.

MEMO

SNUAC
Seoul National University Asia Center
서울대학교아시아연구소

2부 선거자료 활용사례

1. 선거패널조사의 활용

길정아 박사 (고려대 정부학연구소)

선거패널조사의 활용

2020년 2월 5일 14:00~17:00

제8회 KOSSDA 데이터 페어

길정아 (고려대학교 정부학연구소)

1

목차

2

KOSSDA
한 국 사 회 과 학 자 료 원

한국의 선거 패널 데이터

3

KOSSDA
한 국 사 회 과 학 자 료 원

KOSSDA 소장 선거 패널 데이터

1	2007 대선패널조사
2	2008 총선패널조사
3	EAI 총선대선패널조사, 2012
4	EAI·SBS·중앙일보·한국리서치 공동 2006 지방선거 패널조사 : 광주
5	EAI·SBS·중앙일보·한국리서치 공동 2006 지방선거 패널조사 : 부산
6	EAI·SBS·중앙일보·한국리서치 공동 2006 지방선거 패널조사 : 서울
7	EAI·SBS·중앙일보·한국리서치 공동 2006 지방선거 패널조사 : 전국
8	EAI·SBS·중앙일보·한국리서치 공동 2006 지방선거 패널조사 : 충남
9	아산정책연구원 총선대선패널조사, 2012 : 1-7차 [통합자료]
10	아산정책연구원 총선대선패널조사, 2012 : 1차
11	아산정책연구원 총선대선패널조사, 2012 : 2차
12	아산정책연구원 총선대선패널조사, 2012 : 3차
13	아산정책연구원 총선대선패널조사, 2012 : 4차
14	아산정책연구원 총선대선패널조사, 2012 : 5차
15	아산정책연구원 총선대선패널조사, 2012 : 6차
16	아산정책연구원 총선대선패널조사, 2012 : 7차
17	유권자의 대인커뮤니케이션과 태도변화에 관한 패널조사, 2007 : 1차
18	유권자의 대인커뮤니케이션과 태도변화에 관한 패널조사, 2007 : 2차
19	지방선거패널조사, 2010 : 경기
20	지방선거패널조사, 2010 : 경남
21	지방선거패널조사, 2010 : 서울
22	지방선거패널조사, 2010 : 전국
23	지방선거패널조사, 2010 : 전북
24	지방선거패널조사, 2010 : 충남
25	촛불집회 참여자 패널조사 : 3차, 2009

4

KOSSDA
한 국 사 회 과 학 자 료 원

패널 데이터의 구조

5

KOSSDA
한 국 사 회 과 학 자 료 원

데이터의 구조

- **횡단면 데이터(cross-sectional data)**

: 여러 개체에 대하여 동일한 시점에서 조사한 데이터. 개체(i)별로 변량이 발생한다.

- **시계열 데이터(time-series data)**

: 동일한 하나의 개체에 대하여 여러 시점에 걸쳐 조사한 데이터. 시간(t)별로 변량이 발생한다.

- **패널 데이터(panel data; time-series cross-sectional data)**

: 여러 개체에 대하여 여러 시점에 걸쳐 조사한 데이터. 개체(i)별로, 그리고 시간(t)별로 변량이 발생한다.

6

KOSSDA 한국사회과학자료원

Panel Data vs. Time-Series Cross-Sectional Data

- **Panel Data**
  - 개체(i)가 많고, 시간대(t)가 많지 않음
  - 대표적인 예: 동일한 응답자를 대상으로 한 선거 전, 선거 후 조사

- **Time-Series Cross-Sectional Data**
  - 시간대(t)가 많음 (보통 t>10)
  - 대표적인 예: OCDE 국가들을 대상으로 한 민주주의 지수와 GDP 데이터
  - 사회과학 영역에서 말하는 패널 데이터 분석 방법(FE, RE 등)은 실질적으로 이러한 TSCS 데이터를 대상으로 한 것

→ 편의상 두 가지를 구분하지 않고 패널 데이터로 통칭하는 것이 일반적

7

KOSSDA 한국사회과학자료원

Panel Data vs. Pooled Cross-Sectional Data

- **Panel Data (+ Time-Series Cross-Sectional Data)**
  - 동일한 내용을 동일한 여러 개체에 대하여 여러 시점에 걸쳐 조사한 데이터.
  - 각 관찰값들이 개체를 중심으로 서로 연관되어 있다. 서로 독립적이지 않다.
  - 데이터의 축적: merge (column으로 데이터를 추가. 나중에 long shape으로 변환)

- **Pooled Cross-Sectional Data**
  - 동일한 내용을 서로 다른 여러 개체에 대하여 여러 시점에 걸쳐 조사한 데이터.
  - 각 관찰값들이 개체를 중심으로 서로 연관되지 않는다. 서로 독립적이다.
  - 데이터의 축적: append (row로 데이터를 추가)

8

KOSSDA
한 국 사 회 과 학 자 료 원

패널 데이터 구조: wide shape

- 변수 설명
  - id: 개체별(응답자별)
  - inc: 2015년-2018년의 개인 소득 → 시변 변수(time-varying variable)
  - training: 2015년-2018년의 직업 훈련 연 수 → 시변 변수(time-varying variable)
  - male: 응답자 성별 (남성 or 여성) → 시불변 변수(time-invariant variable)
- 연도별 변수들을 merge 하여 wide shape의 형태를 가진 하나의 데이터셋으로 구축

id	inc2015	inc2016	inc2017	inc2018	training2015	training2016	training2017	training2018	male
1	200	200	207	210	10	11	12	13	1
2	300	320	320	330	10	10	11	11	1
3	300	305	330	320	7	8	9	10	0
4	200	220	230	280	12	13	13	13	1
5	400	430	430	450	12	12	12	13	0
6	150	170	200	210	5	6	7	8	0
7	500	510	520	530	12	13	14	15	0
8	300	310	330	360	10	11	11	12	1
9	200	250	260	260	8	9	10	11	1
10	200	200	200	220	8	8	9	10	0

9

KOSSDA
한 국 사 회 과 학 자 료 원

패널 데이터 구조: long shape

- **각 시점(t)별로 inc 변수와 training 변수가 하나의 분석 단위가 되도록 long shape으로 변형**
- **이 경우, n이 4배로 증가**
- **반드시 필요한 정보: 개체(i)와 시간(t)**
  - 각 분석 단위가 어떤 개체(i) 내에 포함(nested)되는지 표시: id 변수
  - 몇년도에 조사된 데이터인지 표시: year 변수
- **시변(time-varying) 변수 vs. 시불변(time-invariant) 변수**
  - inc 변수와 training 변수는 하나의 개체(i) 내에서 값이 변화 → time-varying
  - male 변수는 하나의 개체(i) 내에서 값이 불변 → time-invariant

id	year	inc	training	male
1	2015	200	10	1
1	2016	200	11	1
1	2017	207	12	1
1	2018	210	13	1
2	2015	300	10	1
2	2016	320	10	1
2	2017	320	11	1
2	2018	330	11	1
3	2015	300	7	0
3	2016	305	8	0
3	2017	330	9	0
3	2018	320	10	0
4	2015	200	12	1
4	2016	220	13	1
4	2017	230	13	1
4	2018	280	13	1
5	2015	400	12	0
5	2016	430	12	0
5	2017	430	12	0
5	2018	450	13	0

10

KOSSDA
한 국 사 회 과 학 자 료 원

패널 회귀분석의 기본 모형

11

KOSSDA
한 국 사 회 과 학 자 료 원

회귀분석 기본 모형

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

- y_i: 종속변수. 개체(i)에 따라 변량을 가진다.
- x_i: 독립변수. 개체(i)에 따라 변량을 가진다.
- α: 상수
- β: 회귀계수. x 한 단위의 변화가 가져오는 y의 변화량
- ε_i: 오차항. 개체(i)에 따라 변량을 가진다.

12

KOSSDA
한 국 사 회 과 학 자 료 원

"패널" 회귀분석 기본 모형

$$y_{it} = \alpha + \beta x_{it} + \varepsilon_{it}$$

- y_{it}: 종속변수. 개체(i)와 시간(t)에 따라 변량을 가진다.
- x_{it}: 독립변수. 개체(i)와 시간(t)에 따라 변량을 가진다.
- α: 상수
- β: 회귀계수. x 한 단위의 변화가 가져오는 y의 변화량
- ε_{it}: 오차항. 개체(i)와 시간(t)에 따라 변량을 가진다.

13

KOSSDA
한 국 사 회 과 학 자 료 원

"패널" 회귀분석 기본 모형

$$y_{it} = \alpha + \beta x_{it} + \varepsilon_{it}$$
$$\Rightarrow y_{it} = \alpha + \beta x_{it} + u_i + e_{it}$$

- y_{it}: 종속변수. 개체(i)와 시간(t)에 따라 변량을 가진다.
- x_{it}: 독립변수. 개체(i)와 시간(t)에 따라 변량을 가진다.
- α: 상수
- β: 회귀계수. X 한 단위의 변화가 가져오는 y의 변화량
- ε_{it}: 오차항. 개체(i)와 시간(t)에 따라 변량을 가진다.
- u_i: 개체(i)가 가지는 특수한 속성, 즉 개체마다의 관찰되지 않은 이질성(unit-specific unobserved heterogeneity)을 나타내는 오차항. 시간에 따라 변화하지 않는다.
- e_{it}: u_i를 제외한 오차항. ε_{it} 중에서 개체(i)와 시간(t)에 따라 변량을 가지는 부분만을 남겨둔 것.

14

KOSSDA
한 국 사 회 과 학 자 료 원

"패널" 회귀분석 기본 모형

$$y_{it} = \alpha + \beta x_{it} + u_i + e_{it}$$

- 총 오차(ε_{it}) = 개체 특성(u_i) + 고유오차(e_{it})
- $u_i \sim (0, \sigma_u^2)$, $e_{it} \sim (0, \sigma_e^2)$, $cov(u_i, e_{it}) = 0$, $cov(e_{it}, e_{is}) = 0$

- 추정 방법: Between Effect, Fixed Effect, Random Effect
- 특히, 고정효과 모형(FE)은 개체마다의 관찰되지 않는 이질성, 즉 u_i 을 통제(LSDV 모형으로 추정하여 절편으로 고정 혹은 within-group 추정으로 제거)함으로써, 불편추정량(unbiased estimator)이 됨
- 따라서 추정에는 t를 중심으로 발생하는 변량을 사용

15

KOSSDA
한 국 사 회 과 학 자 료 원

"선거" 패널 데이터의 특성

16

KOSSDA
한 국 사 회 과 학 자 료 원

“선거” 패널 데이터의 특성

- 많은 응답자를 대상으로 하여, 몇 차례의 반복적 조사. 따라서 i 〉 t
- 모든 wave에 모든 변수가 조사되지 않은 경우가 많다. (예: EAI 2012년 7차 조사에서 후보 호감도 변수가 2차, 5차, 7차에만 조사)
- 응답자 누락(attrition)의 문제가 국가 데이터보다 크다.
- 모든 조사에 응한 응답자들간에 공유되는 특성이 반영될 수 있다. (예: 교육수준 높은 이들이 보다 높은 응답률 → 교육수준 높은 이들이 평균적으로 보다 진보적 → 패널조사의 평균값이 보다 진보적)
- t가 작으므로, 일반적으로 패널 데이터 분석에 적용되는 FE 혹은 RE 추정(주로 t를 기준으로 발생하는 변량을 이용하여 추정)을 한 연구가 별로 없다.

17

KOSSDA
한 국 사 회 과 학 자 료 원

선거패널데이터 분석방법 소개

18

KOSSDA 한국사회과학자료원

선거패널데이터 분석방법

- 패널 데이터를 사용하지만 실제로는 횡단면 회귀분석을 수행
- Error Component Model: BE, FE, RE. 특히 FE
- Lagged Dependent Variable Model

→ 중요한 것은 연구의 질문과 가설의 독립변수가 무엇인지,

→ 그리고 i와 t중 무엇을 기준으로 발생하는 변량을 사용한 분석과 조응하는지를 명확히 하는 것

19

KOSSDA 한국사회과학자료원

선거패널 데이터 분석 논문 예시

20

KOSSDA 한국사회과학자료원

횡단면 회귀분석: 예시1

지병근. 2013. "호남에서 나타난 정당후보득표율의 지역편향: 제18대 대선 사례 분석." 『한국정당학회보』 제12권 제1호.

- 데이터: EAI 2012 총선/대선패널조사
- 분석 모형: 로지스틱 회귀분석, OLS 회귀분석
- 주요 내용: 2012년 대통령선거에서 후보에 대한 투표결정요인 검증

이 연구의 투표결정모델에서 종속변수는 투표한 후보이며, 이번 선거에서 가장 많이 득표한 박근혜 후보를 기본범주로 다항로지스틱회귀분석(multinomial logistic regression)이 활용되었다. 독립변수에는 지난 1년간 국가경제상황평가(1~5: 1 = 매우 나빠졌다, 5 = 매우 좋아졌다), 지지정당(새누리당 = 1, 기타 = 0; 민주통합당 = 1, 기타 = 0), 이념적 보수성(0~10), 연령(2~6: 2 = 20대 이하, 6 = 6 = 대 이상), 교육수준(1~4: 1 = 중졸 이하, 2 = 고졸, 3 = 대학재학, 4 = 대졸 이상), 지역(호남 = 1, 기타 = 0; 영남 = 1, 기타 = 0), 성별(남성 = 1, 여성 = 0) 등이 포함되었다. 국가경제상태에 대한 평가는 종속변수인 문재인 후보에 대한 투표가능성에 부정적인 영향을 미칠 것으로 예상된다. 새누리당 지지 또한 종속변수에 부정적인 영향을 미치는 반면, 민주통합당 지지는 긍정적인 영향을 미칠 것으로 예상된다. 이념적 보수성, 연령, 영남은 종속변수에 부정적인 영향을 미치는 반면, 호남, 학력, 남성은 긍정적인 영향을 미칠 것으로 예상된다. 정당선호결정모델에는 종속변수

KOSSDA 한국사회과학자료원

지병근. 2013. "호남에서 나타난 정당후보득표율의 지역편향: 제18대 대선 사례 분석." 『한국정당학회보』 제12권 제1호.

〈표 4〉 제18대 대선에서 유권자들의 정당선호 결정요인

변수	전국		호남	
	새누리당	민주통합당	새누리당	민주통합당
	모델 1	모델 2	모델 3	모델 4
국가경제	0.86***	-0.09	0.57	0.53
	(0.08)	(0.07)	(0.30)	(0.29)
남성	-0.22	0.05	-0.31	-0.06
	(0.13)	(0.12)	(0.43)	(0.41)
연령	0.46***	-0.10**	0.33	0.04
	(0.05)	(0.05)	(0.18)	(0.18)
교육수준	-0.26***	-0.06	-0.01	-0.30
	(0.07)	(0.06)	(0.22)	(0.21)
이념적 보수성	0.44***	-0.25***	0.11	-0.25**
	(0.033)	(0.03)	(0.12)	(0.11)
호남	-0.44	0.48**		
	(0.23)	(0.20)		
영남	0.52***	-0.45***		
	(0.15)	(0.13)		
상수	-0.70	6.81***	0.95	6.15***
	(0.41)	(0.36)	(1.29)	(1.24)
R2	0.36	0.10	0.11	0.08
관측치	1,301	1,296	120	120

괄호안의 수는 표준오차; *** p〈0.01, ** p〈0.05
자료출처: EAI 2012 총선/대선패널조사(7차)

〈표 5〉 문재인/안철수 후보단일화 추진 전후 지지후보결정요인

변수	전국(4차)		호남(4차)		전국(5차)	호남(5차)
	박근혜	안철수	박근혜	안철수	박근혜	박근혜
	모델 1	모델 2	모델 3	모델 4	모델 5	모델 6
국가경제	0.36**	-0.02	0.30	0.18	0.31**	0.40
	(0.15)	(0.13)	(0.57)	(0.40)	(0.14)	(0.45)
새누리당 지지	0.76***	0.06	0.60***	0.18	0.80***	0.63**
	(0.06)	(0.04)	(0.21)	(0.12)	(0.06)	(0.19)
민주통합당 지지	-0.55***	-0.15***	-0.47**	-0.24	-0.65***	-0.73***
	(0.06)	(0.04)	(0.21)	(0.13)	(0.06)	(0.21)
남성	-0.19	-0.27	-1.18	-1.35**	-0.20	-0.98
	(0.22)	(0.18)	(0.89)	(0.60)	(0.20)	(0.66)
연령	0.14	-0.14	-0.34	-0.26	0.10	-0.16
	(0.09)	(0.08)	(0.33)	(0.21)	(0.08)	(0.27)
교육수준	-0.15	0.02	0.92**	0.68**	-0.14	0.06
	(0.11)	(0.10)	(0.45)	(0.29)	(0.11)	(0.33)
이념적 보수성	0.18***	0.01	0.34	0.13	0.22***	0.28
	(0.06)	(0.05)	(0.22)	(0.13)	(0.06)	(0.19)
호남	-1.16***	0.30			-0.91***	
	(0.44)	(0.27)			(0.35)	
영남	0.04	-0.50**			0.22	
	(0.25)	(0.23)			(0.24)	
상수	-2.42***	1.39**	-3.80	0.28	-2.34***	-1.83
	(0.79)	(0.65)	(2.74)	(1.75)	(0.74)	(2.07)
Log likelihood	-1021.63		-88.29		-665.64	-68.99
LR chi2(18)	733.54		38.36		756.18	35.08
Prob 〉 chi2	0.00		0.01		0.00	0.03
Pseudo R2	0.26		0.18		0.36	0.20
관측치	1066		92		1028	98

괄호안의 수는 표준오차; *** p〈0.01, ** p〈0.05
자료출처: EAI 2012 총선/대선패널조사(4, 5차)

KOSSDA 한국사회과학자료원

횡단면 회귀분석: 예시2

이한수. 2013. "정책 선호가 후보자 선택에 미치는 영향력의 변화: 제19대 대통령 선거를 중심으로." 『사회과학연구』 제24권 제3호.

- 데이터: EAI 2017 대선 사전/사후조사
- 분석 모형: 다항로지스틱 회귀분석
- 주요 내용:
  - 사전조사, 사후조사 따로 2번의 다항로지스틱 회귀분석을 수행
  - 후보 선택에 있어, 어떤 정책에 대한 선호는 선거 전에 보다 가시적인 영향력을 미치나, 어떤 정책에 대한 선호는 선거 후에 보다 영향을 미친다는 것을 주장

이 연구의 투표결정모델에서 종속변수는 투표한 후보이며, 이번 선거에서 가장 많이 득표한 박근혜 후보를 기본범주로 다항로지스틱회귀분석(multinomial logistic regression)이 활용되었다. 독립변수에는 지난 1년간 국가경제상황평가(1~5: 1 = 매우 나빠졌다, 5 = 매우 좋아졌다), 지지정당(새누리당 = 1, 기타 = 0; 민주통합당 = 1, 기타 = 0), 이념적 보수성(0~10), 연령(2~6: 2 = 20대 이하, 6 = 6 = 대 이상), 교육수준(1~4: 1 = 중졸 이하, 2 = 고졸, 3 = 대학재학, 4 = 대졸 이상), 지역(호남 = 1, 기타 = 0; 영남 = 1, 기타 = 0), 성별(남성 = 1, 여성 = 0) 등이 포함되었다. 국가경제상태에 대한 평가는 종속변수인 문재인 후보에 대한 투표가능성에 부정적인 영향을 미칠 것으로 예상된다. 새누리당 지지 또한 종속변수에 부정적인 영향을 미치는 반면, 민주통합당 지지는 긍정적인 영향을 미칠 것으로 예상된다. 이념적 보수성, 연령, 영남은 종속변수에 부정적인 영향을 미치는 반면, 호남, 학력, 남성은 긍정적인 영향을 미칠 것으로 예상된다. 정당선호결정모델에는 종속변수

23

KOSSDA 한국사회과학자료원

이한수. 2013. "정책 선호가 후보자 선택에 미치는 영향력의 변화: 제19대 대통령 선거를 중심으로." 『사회과학연구』 제24권 제3호.

<표 2> 정책 선호와 후보 지지 (사전조사)

변수	홍준표	안철수	유승민	심상정	기타
사드	2.43* (0.79) [11.45]	0.77* (0.25) [2.16]	−0.31 (0.49)	−0.00 (0.42)	0.75* (0.34) [2.12]
적폐/통합	−0.36 (0.36)	−1.07* (0.22) [0.34]	−1.05* (0.42) [0.34]	−0.43 (0.36)	−0.61* (0.30) [0.53]
교류/강경	−1.13* (0.43) [0.32]	−0.89* (0.24) [0.41]	−2.12* (0.52) [0.11]	−0.24 (0.42)	−1.02* (0.32) [0.35]
복지/성장	0.02 (0.39)	−0.58* (0.23) [0.55]	0.64 (0.44)	0.38 (0.41)	−0.21 (0.31)
탄핵찬반	−2.24* (0.51) [0.10]	−1.38* (0.45) [0.25]	0.41 (0.74)	12.39 (383.34)	−1.33* (0.51) [0.26]
이념	0.28* (0.08) [1.32]	0.09 (0.05)	0.17 (0.11)	0.00 (0.09)	0.04 (0.07)
중간 생략					
남성	−0.11 (0.34)	−0.03 (0.22)	−0.54 (0.40)	−0.49 (0.35)	···· (0.29) [0.57]
나이	0.01 (0.01)	0.00 (0.00)	−0.00 (0.01)	−0.01 (0.01)	−0.00 (0.01)
영남	0.42 (0.36)	−0.18 (0.25)	0.86* (0.41) [2.37]	−0.12 (0.40)	−0.06 (0.32)
호남	−0.88 (0.76)	−0.24 (0.30)	−1.07 (1.08)	−0.07 (0.48)	−0.10 (0.43)
상수	−0.24 (1.82)	2.33* (1.09)	−0.23 (1.85)	−12.37 (383.34)	3.69* (1.38)
N			1126		
Pseudo R2			0.41		

Note: 종속변수는 후보자 지지 (사전 조사), 기준(reference)은 문재인 지지, 표 안의 숫자는 회귀계수, 괄호 안의 숫자는 표준 오차, 괄호 안의 숫자는 상대적 위험 비율(relative risk ratio). 통계적 유의성 * < 0.10. IIA (Independence of Irrelevant Alternatives) 테스트(Suest-based Hausman test) 결과: 후보 지지는 상호 독립적.

<표 3> 정책 선호와 후보 선택 (사후조사)

변수	홍준표	안철수	유승민	심상정	기타
사드	1.14* (0.48) [3.14]	0.31 (0.26)	0.20 (0.34)	0.04 (0.34)	0.03 (0.70)
적폐/통합	−0.26 (0.33)	−0.76* (0.24) [0.46]	−0.61* (0.30) [0.54]	−0.28 (0.31)	−0.06 (0.59)
교류/강경	−1.41* (0.35) [0.24]	−0.84* (0.24) [0.42]	−1.15* (0.32) [0.31]	−0.53 (0.34)	−1.58* (0.67) [0.20]
복지/성장	−0.88* (0.33) [0.41]	−0.49* (0.24) [0.60]	−0.11 (0.31)	−0.15 (0.33)	−0.16 (0.57)
탄핵찬반	−2.90* (0.48) [0.05]	−1.03* (0.49) [0.35]	−0.03 (0.69)	−0.74 (0.84)	−3.10* (0.74) [0.04]
중간 생략					
남성	0.03 (0.31)	−0.25 (0.23)	−0.07 (0.29)	−0.37 (0.30)	−0.12 (0.56)
나이	0.02* (0.01) [1.03]	0.00 (0.01)	−0.02* (0.01) [0.97]	−0.03* (0.01) [0.96]	0.00 (0.02)
영남	0.49 (0.32)	0.00 (0.27)	0.44 (0.30)	−0.02 (0.34)	−0.23 (0.63)
호남	−1.40* (0.59) [0.24]	0.05 (0.30)	−1.04 (0.64)	0.00 (0.43)	−0.27 (0.90)
상수	2.36 (1.44)	1.00 (1.14)	1.33 (1.46)	1.57 (1.61)	0.52 (2.54)
N			942		
Pseudo R2			0.34		

Note: 종속변수는 후보자 선택 (사후 조사), 기준(reference)은 문재인 지지, 표 안의 숫자는 회귀계수, 괄호 안의 숫자는 표준 오차, 괄호 안의 숫자는 상대적 위험 비율(relative risk ratio). 통계적 유의성 * < 0.10. IIA (Independence of Irrelevant Alternatives) 테스트(Suest-based Hausman test) 결과: 후보 지지는 상호 독립적.

24

KOSSDA 한국사회과학자료원

횡단면 회귀분석: 예시3

강신구. 2013. “제19대 대통령 선거와 TV 토론회: 지지후보 변경에 미친 효과 .” 『OUGHTOPIA 』 제32권 제2호.

- 데이터: EAI 2017 대선 사전/사후조사
- 분석 모형: 프로빗 회귀분석, OLS 회귀분석
- 주요 내용:
  - 선거 전에 지지하던 후보를 선거에 실제로 투표했는지의 여부에 영향을 미치는 요인이 무엇인지를 검증
  - 후보자가 TV토론을 잘 했다고 생각하면, 그 후보에 대해 기존의 지지 및 평가를 변화하지 않을 것임을 주장

표에 제시되어 있는 것처럼, 세 개의 다변인 분석은 세 개의 서로 다른 종속변수를 대상으로 하여 실행되었다. <모형 1>의 종속변수는 지지후보변경으로 선거 전에 투표의사를 표명했던 후보에게 실제로 투표했는지의 여부를 보여주는 더미변수이다(일치하지 않으면 1). <모형 1>에서 주된 독립변수는 지지후보 TV 토론승리이다. 이 변수 역시 더미변수로서 선거 전에 투표의사를 표명했던 후보자가 TV 토론을 가장 잘했다고 답한 응답자에 대해서 1의 값을 부여하였다. <모형 2>는 <모형 1>의 독립변수, 즉 지지후보 TV 토론승리를 종속변수로 하여 이러한 TV토론 성적에 대한 평가에 미치는 유권자의 후보에 대한 기존 선호의 영향을 살펴보고자 하였다. 이에 따라 <모형 2>에서 가장 주목하는 독립변수는 지지후보에 대한 호감도[13]이다. 이 변수와 관련하여 특히 유의해야 할 점은 이것이 단순히 후보에 대한 호감도를 표시하는 것이 아니라, ‘지지하는 후보’, 즉 1차 조사에서 다른 후보를 제치고 투표할 의사를 표명했던 후보에 대한 호감도를 표시하고 있다는 점이다. 마지막의 <모형 3>은 <표 9>에서 살펴본 바 있는 설문을 활용하여 구성된 지지후보평가변화[14]를 종속변수로 하여 역시 지지후보에 대해 기존에 갖고 있는 호감도가 지지후보에 대한 평가 변화의 양상에 미치는 영향을 조금 더 세분하여 살펴보고자 하였다. 종속변수의 차이를 감안하여 모형 1, 2는 프로빗(probit), 모형 3은 일반최소자승(OLS) 모형의 분석방법을 이용하여 계수의 크기와 표준오차를 추정하였다.[15]

KOSSDA 한국사회과학자료원

강신구. 2013. “제19대 대통령 선거와 TV 토론회: 지지후보 변경에 미친 효과 .” 『OUGHTOPIA 』 제32권 제2호.

- 모형1의 분석 결과:
  - 지지후보 변경 “확률”에 영향을 미치는 요인검증
  - 이 연구에서 초점을 맞추고 있는, 지지후보가 TV토론에서 잘했다고 평가하면 지지후보를 변경할 확률이 감소한다.
- 개인들의 속성에서 나타나는 변량을 이용한 연구질문
  - 종속변수: 응답자의 “지지후보를 변경” 여부
  - 주요 독립변수: 응답자의 “TV토론 평가”

〈표 10〉 지지후보변경, 지지후보 TV토론 승리, 지지후보 평가변화에 대한 다변인 분석

	모형 1		모형 2		모형 3	
종속변수	지지후보 변경		지지후보 TV토론 승리		지지후보 평가변화	
분석모형	probit		probit		OLS	
	계수	표준 오차	계수	표준 오차	계수	표준 오차
여성	−0.180	0.097	0.144	0.097	−0.110	0.060
연령: 19~29세 이하	0.240	0.133	0.072	0.136	0.148	0.085
연령: 60세 이상	0.400**	0.148	−0.045	0.154	0.209*	0.096
교육수준	−0.062	0.054	0.064	0.055	0.014	0.034
가계소득	0.001	0.047	−0.060	0.049	0.056	0.030
주관적 계층인식	−0.006	0.080	−0.057	0.079	0.083	0.050
지지후보와 지지정당일치	−0.530***	0.097	0.420***	0.107	−0.317***	0.063
지지후보 지역거주	−0.342	0.181	0.089	0.162	−0.200	0.103
주요 정당후보 지지	−0.985***	0.253				
지지후보 TV토론승리	−0.750***	0.206				
주요정당후보지지* 지지후보 TV토론 승리	−0.403	0.379				
지지후보에 대한 호감도	−0.107***	0.029	0.098***	0.030	−0.080***	0.018
지지후보와의 이념적 거리	0.024	0.028	0.061	0.038	0.006	0.018
상수	1.784***	0.479	−1.981***	0.415	2.835***	0.254
사례수	930		940		913	

주: * p<0.05, ** p<0.01, *** p<0.001
출처: 2017년 EAI 대선패널 조사.

KOSSDA 한국사회과학자료원

Error Component Model (특히 FE): 예시1

김성연. 2017. "제18대 대통령 선거에서 이념의 영향: 패널 데이터 분석 결과." □의정연구□ 제23권 제2호.

- 데이터: EAI 2012 총선대선패널조사
- 분석 모형: Pooled OLS, Fixed Effect, Random Effect
- 주요 내용:
  - 후보 평가 혹은 투표선택에 대해 미치는 이념의 영향력을 검증
  - 이념적으로 가깝거나, 동일한 방향으로 이념적으로 분명한 후보에 대해 보다 호의적일 것임을 주장

이러한 점을 고려하여 이 연구에서는 다음과 같이 데이터 분석을 진행한다. 기본적으로, 2012년 4월 중순(2차), 11월 말(5차), 그리고 12월 하순(7차)에 조사된 세 차례의 패널 데이터(3-wave panel data)를 이용하여 이념 투표의 영향을 분석한다. 이 세 차례 패널 데이터는 전체 대통령 선거 시기를 포함할 뿐만 아니라 거의 모든 핵심 변수들을 포함한다. 유일한 문제는 응답자들의 이념 성향이 1, 3, 5, 7차 패널에서 조사되었기 때문에 2차 패널 조사에 포함되지 않았다는 것인데, 이것은 1차 패널(3월 말) 조사 결과로 대체할 수 있을 것이다. 왜냐하면 1차(3월 말)와 2차 패널(4월 중순)은 그 조사 시기가 약 2주밖에 차이가 나지 않으며, 불과 2주의 기간 동안 응답자들의 이념 성향이 크게 변할 가능성은 별로 없기 때문이다.[5] 이와 더불어, 분석의 엄밀성을 위해 추가적으로 5차(11월 말)와 7차 패널(12월 하순)에서 조사된 두 차례 패널 데이터(2-wave panel data)를 이용하여 이념 투표의 영향을 분석한다. 이 두 차례 패널 데이터 분석은 대통령 선거 마지막 채 한 달이 안 되는 시기(11월 말 – 12월 하순)만을 대상으로 한다는 점에서 일정한 한계를 갖는다. 그러나 다른 한편으로 이 시기에 대한 분석은 안철수 후보가 후보 등록 직전(11월 말) 사퇴한 후 주요 후보가 박근혜와 문재인 후보 두 명으로 압축되었다는 점에서 앞의 세 차례 패널 데이터 분석과 구분된다. 따라서 이 시기에 대한 분석은 이념 투표에 대한 분석의 엄밀성을 높이는 차원에서 의미가 있다.

27

KOSSDA 한국사회과학자료원

김성연. 2017. "제18대 대통령 선거에서 이념의 영향: 패널 데이터 분석 결과." □의정연구□ 제23권 제2호.

- 특히 고정효과 모형에 초점
- t를 중심으로 한 변량을 사용하여 추정
- 개인 고유의 속성을 통제하고, 3개의 time points(2차, 5차, 7차 조사)에 따라 나타나는 독립변수(이념)와 종속변수(후보 호감도)의 관계를 모형화

이 연구에서는 표준적인 패널 데이터 분석, 즉 합동 회귀 분석(pooled regression), 고정 효과(fixed effects) 분석, 그리고 변동 효과(random effects) 분석을 통해 이념 투표의 영향을 분석한다(Wooldridge 2010: 김성연 2016). 이 중에서 가장 중요한 것은 고정 효과 분석이다. 고정 효과 분석은 앞에서 설명한 패널 데이터의 특성을 충분히 활용하여 보다 엄밀한 분석 결과를 제시한다. 즉, 개별 유권자들의 모든 고정된 특성들(fixed individual heterogeniety), 특히 관측된(observed) 특성들뿐만 아니라 관측되지 않은(unobserved) 특성들까지 통제한 후 종속 변수와 설명 변수의 상관관계를 분석함으로써 내생성 오류의 가능성을 크게 줄이고 그에 따라 보다 신뢰할 수 있는 결과를 제시하는 것이다.

변동 효과 분석과 합동 회귀 분석 결과는 고정 효과 분석 결과와 비교할 때 상대적으로 신뢰성이 떨어진다고 할 수 있다. 구체적으로, 변동 효과 분석과 합동 회귀 분석은 예측 모형의 모든 설명 변수가 모든 시간에 걸쳐 외생적(exogenous)이라는 가정에 기초하고 있으며, 이 가정이 성립하지 않을 경우 일반적으로 그 분석 결과를 신뢰하기 어렵다. 그러나 이것은 매우 강력한 가정으로, 사실상 예측 모형에 생략된 설명 변수가 없다고, 즉 앞의 식 (1)에서 c^*가 존재하지 않는다고 가정하는 것이다. 따라서 만일 이 가정이 성립한다면 굳이 패널 데이터의 분석을 통하지 않고서도, 즉 횡단면 데이터의 분석만으로도 타당한(valid) 분석 결과를 얻을 수 있다.[6]

28

KOSSDA 한국사회과학자료원

김성연. 2017. “제18대 대통령 선거에서 이념의 영향: 패널 데이터 분석 결과.” □의정연구□ 제23권 제2호.

- 고정효과 모형의 해석:
  - 근접성 모형: 한 개인에게서 박근혜 후보와 이념 차이가 한 단위 가까워질수록 박근혜 후보에 대한 호감도가 0.08씩 증가
  - 방향성 모형: 한 개인에게서 박근혜 후보와 같은 보수적 이념을 가지고 있고 그 정도가 한 단위 분명해질수록 박근혜 후보에 대한 호감도가 0.02씩 증가
- 고정효과 모형은 개인들의 속성을 통제하고, 시간에 따른 변량(within-group variation)만을 사용, 해석에 유의

| 표 4 | 박근혜 후보 호오도에 대한 합동 회귀 분석, 고정 효과, 변동 효과 분석 결과 (2차, 5차, 7차 조사 데이터)

	근접성 모형			방향성 모형		
	합동 회귀 분석 (pooled regression)	고정 효과 (fixed effect)	변동 효과 (random effect)	합동 회귀 분석 (pooled regression)	고정 효과 (fixed effect)	변동 효과 (random effect)
5차 조사	−0.95** (0.09)	−0.77** (0.06)	−0.85** (0.06)	−0.98** (0.09)	−0.76** (0.06)	−0.85** (0.06)
7차 조사	−0.51** (0.09)	−0.32** (0.06)	−0.41** (0.06)	−0.55** (0.09)	−0.31** (0.06)	−0.42** (0.06)
민주통합당	−3.00** (0.10)	−0.46** (0.20)	−2.14** (0.11)	−2.99** (0.10)	−0.46** (0.14)	−2.11** (0.11)
기타 정당	−3.55** (0.15)	−0.32 (0.17)	−2.24** (0.14)	−3.68** (0.15)	−0.31 (0.17)	−2.22** (0.14)
무당파	−2.58** (0.10)	−0.62** (0.13)	−1.91** (0.10)	−2.44** (0.10)	−0.61** (0.13)	−1.83** (0.10)
이념 차이	−0.28** (0.02)	−0.08** (0.02)	−0.18** (0.01)	−0.07** (0.005)	−0.02** (0.005)	−0.04** (0.004)
나이	0.02** (0.003)		0.03** (0.004)	0.02** (0.003)		0.03** (0.004)
교육	−0.55** (0.09)		−0.60** (0.12)	−0.57** (0.09)		−0.61** (0.12)
지역 등 기타 통제 변수	...		...	...		...
N	4,078	4,235	4,078	4,078	4,235	4,078

주) 표준오차는 괄호 안에 나타냈으며, 유의도는 * $p<0.05$, ** $p<0.01$.

29

KOSSDA 한국사회과학자료원

Error Component Model (특히 FE): 예시2

김성연. 2016. ″한국 선거에서 경제 투표의 영향: 제18대 대통령 선거 패널 데이터 분석 결과.” 『한국정치학회보』 제50집 제2호.

- 데이터: EAI 2012 총선대선패널조사
- 분석 모형: Pooled OLS, Fixed Effect, Random Effect
- 주요 내용:
  - 후보 평가 혹은 투표선택에 대해 미치는 경제투표 행태를 검증
  - 경제적으로 긍정적인 평가를 할수록 여당의 박근혜 후보를 선택할 확률 혹은 박근혜 후보에 대한 호감도가 증가함을 주장

지금까지 설명한 변수들 중 후보 선택, 후보 호오도, 네 가지 경제 인식, 그리고 정당 지지는 모두 4월 말(2차 조사)과 12월 말(7차 조사)에 조사된 결과를 사용했으나, 이념 성향은 이와 달리 3월 말(1차 조사)와 12월 말(7차 조사) 조사 결과를 사용하였다. 이것은 유권자들의 이념 성향이 4월 말(2차 조사) 조사에 포함되지 않았기 때문이다. 그러나 1차와 2차 조사는 약 한 달 정도의 차이를 두고 진행되었기 때문에 이 기간 동안 이념 성향이 크게 변했을 가능성은 별로 없을 것이다. 유권자들의 나이, 교육 수준, 출신 지역 등 나머지 변수들은 조사 기간 동안 변하지 않는 변수들이므로 모두 1차에서만 한차례 조사되었으며 그 결과를 사용하였다.

이 연구에서는 합동 회귀 분석(pooled regression), 고정 효과(fixed effects) 분석, 그리고 변동 효과(random effects) 분석 등 표준적인 패널 데이터 분석 모형들을 적용하여 경제 투표의 영향을 분석한다. 일반적으로 패널 데이터는 여러 가지 방법으로 분석될 수 있다. 예컨대, 종속 변수의 이전 시기 관측 값(lagged dependent variable)을 설명 변수로 활용할 수도 있고, 특정 설명 변수의 현 시기 와 이전 시기의 관측 값을 모두 활용할 수도 있으며, 도구 변수(instrumental variable)를 이용하여 분석의 엄밀성을 높일 수도 있다. 그러나 주지하듯이, 패널 데이터의 표준적인 분석 모형은 합동 회귀 분석(pooled regression), 고정 효과(fixed effects) 분석, 그리고 변동 효과 (random effects) 분석이다.[5] 이 중에서도 가장 중요한 것은 고정 효과 분석인데, 그것은 고정 효과 분석이 패널 데이터의 장점을 최대한 활용함으로써 보다 엄밀한 분석 결과를 제공하기 때문이다.

30

KOSSDA 한국사회과학자료원

김성연. 2016. ″한국 선거에서 경제 투표의 영향: 제18대 대통령 선거 패널 데이터 분석 결과.” 『한국정치학회보』 제50집 제2호.

- 고정효과 모형의 해석:
  - 한 개인에게서 경제에 대해 긍정적으로 평가하는 정도가 증가하는 것은 후보의 호감도에 통계적으로 유의미한 영향이 없다.
- 변동효과 모형의 해석:
  - 한 개인에게서 혹은 개인간에 경제에 대해 긍정적으로 평가하는 정도가 증가하면 박근혜 후보의 호감도에 정(+)의 영향력을 미친다.
- 고정효과 모형은 개인들의 속성을 통제하고, 시간에 따른 변량(within-group variation)만을 사용한 추정
- 변동효과 모형은 개인들 내부의 변량(within-group variation)과 개인간 변량(between-group variation)을 모두 사용한 추정
- 해석에 유의

| 표 7 | 박근혜 후보와 문재인 후보 호오도 차이의 통합회귀분석(pooled OLS), 고정 효과(fixed effects), 변동효과(random effect) 분석 결과

	종속 변수: 박근혜와 문재인 후보 호오도 차이		
	통합 회귀 분석 (Pooled OLS)	고정효과 분석 (Fixed Effect)	변동효과 분석 (Random Effect)
패널7차	-1.10*** (0.12)	-0.75*** (0.10)	-1.00*** (0.10)
민주통합당	-4.53*** (0.18)	-1.18*** (0.29)	-4.03*** (0.18)
다른 정당	-4.31*** (0.39)	-0.90* (0.47)	-3.74*** (0.36)
통합진보당	-5.61*** (0.27)	-0.85** (0.37)	-4.52*** (0.26)
지지 정당 없음	-2.76*** (0.17)	-0.57** (0.26)	-2.44*** (0.17)
이념	0.28*** (0.03)	0.03 (0.04)	0.23*** (0.03)
가정 경제 회고적 평가	0.14 (0.09)	-0.01 (0.12)	0.11 (0.09)
가정 경제 전망적 평가	0.25*** (0.09)	-0.12 (0.11)	0.20** (0.09)
국가 경제 회고적 평가	0.34*** (0.08)	0.11 (0.10)	0.35*** (0.08)
국가 경제 전망적 평가	0.76*** (0.08)	0.14 (0.10)	0.62*** (0.08)
나이	0.02*** (0.01)		0.03*** (0.01)
교육 (대재 이상)	-0.48*** (0.14)		-0.56*** (0.17)
지역 등 기타 통제 변수	.....		.....
N	2,713	2,814	2,713

주) 표준오차는 괄호 안에 보고했으며 유의도는 * p<0.1, ** p<0.05, *** p<0.01.

31

KOSSDA 한국사회과학자료원

Lagged Dependent Variable Model: 예시1

김성연. 2015. “정치적 태도와 인식의 양극화, 당파적 편향, 그리고 민주주의: 2012년 대통령 선거 패널 데이터 분석.” □민주주의와 인권□ 제15권 3호.

- 데이터: EAI 2012 총선대선패널조사
- 분석 모형: lagged DV model
- 주요 내용:
  - 2012년 대통령 선거를 기점으로, 유권자들은 자신의 당파성에 일치하는 방향으로 자신의 정치적 태도를 강화
  - 패널 데이터의 종단적 특성을 이용
  - 이전 시점(t-1)의 정치적 태도가 이후 시점(t)의 정치적 태도에 얼마나 영향을 미치는지를 검증

구체적으로, 이 연구는 지난 2012년 4월 말에서 12월 말 사이에 나타난 유권자들의 정치적 태도와 인식의 변화에 초점을 맞춘다. 수십 년에 걸쳐 나타나는 장기적인 정치적 태도와 인식의 변화도 중요하지만, 대통령 선거 시기와 같이 유권자들의 정치에 대한 관심이 높은 시기에 나타나는 단기적인 변화 또한 중요하다는 것은 명백하다. 또한 이 연구는 패널 데이터를 이용하여 이러한 정치적 태도와 인식의 변화를 추적하고 여기에 영향을 미치는 요인들을 분석할 것이다. 정치적 태도와 인식의 시간에 따른 변화 및 이에 영향을 미치는 요인들의 분석은 패널 데이터를 이용하지 않고서는 엄밀하게 이루어지기 어렵다. 이를 위해서는 동일한 유권자들의 정치적 태도와 인식 및 이에 영향을 미치는 요인들이 시간의 차이를 두고 여러 차례에 걸쳐 관측되어야 하기 때문이다.[1]

32

KOSSDA 한국사회과학자료원

김성연. 2015. "정치적 태도와 인식의 양극화, 당파적 편향, 그리고 민주주의: 2012년 대통령 선거 패널 데이터 분석." □민주주의와 인권□ 제15권 3호.

이 연구의 목적에 비추어 볼 때, 『EAI 총선대선패널조사, 2012』는 몇 가지 중요한 장점들을 갖고 있다. 가장 중요한 것은 이 데이터가 동일한 응답자들의 정치적 태도와 인식이 시간이 지남에 따라 어떻게 변하는지를 관측한 결과를 담고 있는 패널 조사 결과라는 점이다. 구체적으로, 응답자들의 이념 성향, 후보들에 대한 호오도, 후보의 이념 성향과 국정운영능력 등 자질에 대한 인식, 그리고 국가 경제와 가정 살림에 대한 평가와 전망 등이 반복적으로 조사되었다. 이것은 정치적 태도와 인식의 변화에 대한 보다 엄밀한 분석을 가능하게 한다. 즉, 여러 개의 횡단면(cross-sectional) 자료들을 분석하는 것과 달리, 동일한 응답자들의 정치적 태도와 인식이 시간이 지남에 따라 어떻게 변화했는지를 추적하고 이러한 변화에 이들의 정당 지지, 이념 성향, 출신 지역 등이 미친 영향을 분석할 수 있는 것이다.

$$정치적\ 태도_t = \beta_0 + \beta_1 정치적\ 태도_{t-1} + \beta_2 정당\ 지지_{t-1} + \beta_3 이념_{t-1} + \beta_4 SNS_{t-1} + 기타\ 변수 + \epsilon_t$$

이 식에서 종속 변수인 정치적 태도t는 시기 t에 측정된 유권자들의 정치적 태도와 인식으로 박근혜와 문재인 후보에 대한 호오도, 후보의 이념과 자질에 대한 인식, 그리고 국가 경제와 가정 살림에 대한 평가와 전망 등을 포함한다. 설명 변수들 중 정치적 태도t-1는 이전 시기(t-1)에 측정된 종속 변수이고, 정당 지지t-1, 이념t-1, 그리고 SNSt-1는 t-1에 측정된 정당지지, 이념 성향, 그리고 SNS 사용 정도이며, 기타 변수는 응답자들의 나이, 교육 수준, 소득, 그리고 출신 지역 등 표준적인 사회경제적 특징들을 포함한다.

위 모형은 일정 기간 동안 발생한 종속 변수의 변화를 일련의 설명 변수들로 설명하는 표준적인 통계분석 모형이다. 즉, 이 모형은 t-1과 t 시기 사이에 발생한 정치적 태도와 인식의 변화를 t-1에 관측된 정당 지지, 이념, 그리고 여타 설명 변인들로 설명한다. 이 식에서 β0는

KOSSDA 한국사회과학자료원

김성연. 2015. "정치적 태도와 인식의 양극화, 당파적 편향, 그리고 민주주의: 2012년 대통령 선거 패널 데이터 분석." □민주주의와 인권□ 제15권 3호.

- 분석 결과:
  - 4월 시점의 정치적 태도가 12월 시점의 정치적 태도에 영향을 미쳤다.
  - 본래 가지고 있던 정치적 태도가 12월에 더욱 강화되었다.

〈표 3〉 국가 경제와 가정 살림에 대한 평가와 전망 (2012년 4월 말 - 12월 말)

	국가 경제 (12월 말)		가정 살림 (12월 말)	
	전망	평가	전망	평가
상수	1.92 (0.24)**	1.31 (0.20)**	1.72 (0.21)**	1.47 (0.17)**
전망 (4월)	0.38 (0.03)**		0.35 (0.03)**	
평가 (4월)		0.38 (0.03)**		0.40 (0.02)**
새누리당 지지	0.26 (0.08)**	0.23 (0.07)**	0.20 (0.07)**	0.21 (0.06)**
민주통합당 지지	−0.06 (0.08)	0.01 (0.07)	−0.09 (0.07)	0.11 (0.06)†
기타 정당	−0.10 (0.09)	−0.04 (0.08)	−0.13 (0.08)†	−0.01 (0.07)
응답자 이념	0.01 (0.01)	0.00 (0.01)	0.01 (0.01)	−0.01 (0.01)
나이	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
교육 (대재 이상)	−0.16 (0.06)**	−0.04 (0.05)	−0.11 (0.05)*	0.01 (0.04)
소득	0.00 (0.01)	0.00 (0.01)	0.03 (0.01)**	0.03 (0.01)**
여성	0.22 (0.06)**	0.01 (0.06)	−0.02 (0.06)	0.09 (0.05)†
SNS 이용	−0.03 (0.02)	−0.01 (0.02)	0.00 (0.02)	0.00 (0.01)
인천/경기	0.11 (0.09)	0.03 (0.08)	0.03 (0.08)	−0.13 (0.07)†
대전/충청	0.06 (0.09)	0.10 (0.07)	0.03 (0.07)	0.10 (0.06)
광주/호남	−0.09 (0.09)	0.05 (0.07)	0.02 (0.08)	−0.03 (0.06)
대구/경북	−0.01 (0.09)	0.02 (0.07)	0.01 (0.07)	0.07 (0.06)
부산/경남	−0.03 (0.08)	0.13 (0.07)†	0.02 (0.07)	0.13 (0.06)*
기타 통제 변수				
Adj. R^2	0.26	0.23	0.22	0.25
N	1118	1142	1136	1151

주) P-value (two-tailed): † 〈 0.1, * 〈 0.05, ** 〈 0.01. 자료: 『EAI 총선대선패널조사, 2012』

KOSSDA 한국사회과학자료원

Lagged Dependent Variable Model: 예시2

길정아. 2019. "제4장: 이념거리 인식의 당파적 편향." *한국 유권자의 당파적 편향: 양극화의 미시적 토대*. 서울대학교 박사학위논문.

- 데이터: EAI 2017 대선 사전/사후조사
- 분석 모형: cross-lagged variable model
- 주요 내용:
  - 후보와의 이념 거리를 인식함에 있어 당파적 호감도의 영향을 받는지를 검증
  - 이념 거리 인식과 당파적 호감도는 상호간에 영향을 주고받는 관계
  - 이념적으로 가깝게 인식하는 후보에 대한 호감도가 높아지는지, 아니면 호감도가 높은 후보에 대해 이념적으로 더 가깝게 느끼는지를 비교 분석

본 연구의 종속변수는 유권자가 인식하는 자신과 정당들, 그리고 자신과 후보들의 상대적 이념 거리이다. 두 명의 주요 후보를 중심으로, 이념적 거리 변수는 다음과 같이 조작화 하여, 상대적인 거리로 측정한다.[9]

18대 대선: 정당자 상대적 이념거리$_i$ = (응답자$_i$-민주통합당$_i$)^2 - (응답자$_i$-새누리당$_i$)^2

후보간 상대적 이념거리$_i$ = (응답자$_i$-문재인$_i$)^2 - (응답자$_i$-박근혜$_i$)^2 [10]

19대 대선: 후보간 상대적 이념거리$_i$ = (응답자$_i$-홍준표$_i$)^2 - (응답자$_i$-문재인$_i$)^2 [11]

유권자가 인식하는 이념적 거리에 영향을 미치는 독립변수로서 두 정당, 그리고 두 후보에 대한 상대적 호감도를 다음과 같이 조작화 한다.

18대 대선: 정당간 상대적 호감도$_i$ = 새누리당 호감도$_i$-민주통합당 호감도$_i$

후보간 상대적 호감도$_i$ = 박근혜 호감도$_i$-문재인 호감도$_i$ [12]

19대 대선: 후보간 상대적 호감도$_i$ = 문재인 호감도$_i$-홍준표 호감도$_i$ [13]

KOSSDA 한국사회과학자료원

길정아. 2019. "제4장: 이념거리 인식의 당파적 편향." *한국 유권자의 당파적 편향: 양극화의 미시적 토대*. 서울대학교 박사학위논문.

분석 모형은 다음과 같다. 먼저, [가설1-1]과 [가설1-2]를 검증하기 위해, "교차-지연(cross-lagged)" 변수를 조작화 하여 다음과 같은 회귀분석 모형을 구성하였다.[15] 이는 다음의 [모형1] 그리고 [모형2]와 같다.

[모형1]

$$\text{상대적 이념거리}_{it} = \alpha_1 + \boldsymbol{\beta_1 \times \text{후보자 호감도}_{it-1}} + \gamma_1 \times \text{상대적 이념거리}_{it-1} + \Sigma\delta_j \times \text{Control}_{ji} + \varepsilon_{1it}$$

[모형2]

$$\text{후보자 호감도}_{it} = \alpha_2 + \boldsymbol{\beta_2 \times \text{상대적 이념거리}_{it-1}} + \gamma_2 \times \text{후보자 호감도}_{it-1} + \Sigma\delta_k \times \text{Control}_{ki} + \varepsilon_{2it}$$

15) 바텔스(Bartels 2002), 그리고 카시와 레이먼(Carsey and Layman 2006)이 정당일체감과 이슈에 대한 태도간의 동적(dynamic) 관계를 살펴보기 위해 패널 데이터를 활용하여 이러한 "교차-지연(cross-lagged)" 변수를 활용한 모형을 분석한 바 있다. 모형에 대한 자세한 설명은 핀켈(Finkel 1995)을 참고할 수 있다.

KOSSDA 한국사회과학자료원

길정아. 2019. “제4장: 이념거리 인식의 당파적 편향.” *한국 유권자의 당파적 편향: 양극화의 미시적 토대*. 서울대학교 박사학위논문.

- 이전 시기에 이념적으로 가깝게 여기는 후보를 이후 시기에도 가깝게 여기는 정도: 0.304
- 이전 시기에 선호하는 후보를 이후 시기에도 선호하는 정도: 0.594

- 이전 시기에 이념적으로 가깝게 여기는 후보를 이후 시기에 선호하는 정도: 0.070
- 이전 시기에 선호하는 후보를 이후 시기에 이념적으로 가깝게 여기는 정도: 0.216

- 후보와의 이념거리 인식이 유지되는 것보다, 후보에 대한 호감도가 더 안정적으로 유지된다.
- 당파적으로 선호하는 후보에 대해 이념적으로 보다 가깝게 여기는 정도가 더 크다.

[표4-4] 이념거리와 호감도의 교차 효과 검증: 19대 대통령선거

	종속변수: 이념거리$_t$			종속변수: 후보자호감도$_t$		
	Coef.	beta	Std. Err.	Coef.	beta	Std. Err.
후보자호감도$_{t-1}$	1.501***	0.216	0.277	0.563***	0.594	0.025
이념거리$_{t-1}$	0.302***	0.304	0.033	0.010**	0.070	0.003
더불어민주당	8.651**	0.129	2.985	1.731***	0.188	0.262
자유한국당	-11.573**	-0.114	3.737	-1.767***	-0.127	0.329
국민의당	6.344†	0.056	3.642	0.591†	0.038	0.319
바른정당	-5.255	-0.039	4.064	-0.068	-0.004	0.361
정의당	5.210	0.048	3.712	1.588***	0.105	0.326
성별(남성=1)	-3.574*	-0.053	1.681	-0.638***	-0.069	0.148
연령	-0.025	-0.010	0.073	-0.012†	-0.035	0.006
교육수준	0.273	0.005	1.733	-0.180	-0.022	0.151
가구소득	-0.026	-0.002	0.366	-0.026	-0.014	0.032
부산/울산/경남	-0.507	-0.005	2.531	-0.162	-0.012	0.222
대구/경북	1.754	0.015	3.015	-0.098	-0.006	0.263
광주/전라/전북	-0.897	-0.008	3.022	0.395	0.025	0.268
대전/세종/충청	0.993	0.009	2.819	0.134	0.009	0.248
강원/제주	4.623	0.026	4.416	0.504	0.021	0.384
상수	3.053	-	7.236	2.996***	-	0.635
N	998			1,016		
R-square	0.4083			0.7513		
adj. R-square	0.3986			0.7473		

*** p<0.001, ** p<0.01, * p<0.05, † p<0.1 (양측검정)

37

KOSSDA 한국사회과학자료원

Wrap-up

- 선거 패널 데이터의 구조를 이해하고,
- 연구의 주장과 가설, 이에 따른 독립변수와 종속변수가 어떤 내용을 담고 있는지를 명확히 하는 것,
- 즉 이 연구가 i와 t중 무엇을 기준으로 발생하는 변량을 사용해 추정하는 모형과 조응하는지를 명확히 하는 것이 가장 중요

38

KOSSDA
한 국 사 회 과 학 자 료 원

감사합니다

39

MEMO

MEMO

SNUAC
Seoul National University Asia Center
서울대학교아시아연구소

2부 선거자료 활용사례

2. 선거집계자료와 선거연구의 확장

박원호 교수 (서울대 정치외교학부)

선거집계자료와 선거연구의 확장

2020. 2. 5.
KOSSDA DATA FAIR
서울대학교 정치외교학부 박원호

Rethinking Election Data

선거여론조사 ≠ 선거데이터

- 선거여론조사가 선거데이터를 대체한 것은 비교적 새로운 현상
- 서베이의 근본적 한계
  - (사회학 버전?) 서베이만을 통해서 사회현상을 연구할 수 있는 것은 아니다
  - (인류학 버전?) 연구대상들이 자발적으로 직접 이야기하는 것을 다 믿을 수는 없다

3

정치학 버전(?)

"The public opinion polls give some indication of the composition of parties and the nature of political trends in the United States, but they are no substitute for the election results."

Harold F. Gosnell, *Grass Roots Politics* (1942)

4

선거여론조사가 던지는 난제들

- 비밀투표라는 존재론적인 문제
  - People forget and people lie: bandwagon/social dsblty
- 서베이가 닿지 않는 시공간적 영역
  - 1991년 이전 한국의 선거와 선거史 복원 프로젝트
  - 비서구국가들, 과거의 선거, 신생/비민주 국가의 선거
- 동학에 대한 접근: 패널자료, but 근본적 한계
- 객관적 상황을 '인식'이나 '평가'에 의존해야
- Yes, and boring, too.

5

집계자료의 희망

- 자료의 availability
  - 선관위, 통계청, UN, 지방정부, 연구단체 등 (KOSSDA 협약)
  - 개인정보에 대한 우려 없음
  - 강력한 중앙관료국가의 전통
  - 다양하고 새로운 집계단위 가능, '빅데이터'와의 친화성
- 잘 정의된 unit만 있다면
  - 시군구, 읍면동, 국가, Dyad; 그리고 Time horizon
  - Cross-unit merge를 통한 새로운 data set 정의 가능
  - GIS를 이용한 새로운 변수들과 insight
- 응답자의 인식이 아닌 "real thing" 사용 가능
  - 경제에 대한 인식이 아니라 경제지표 사용 가능

6

집계자료의 절망

- 무한한 가능성은 부담이기도 하다
  - 새로운 연구 의제를 발굴할 부담
  - Boring topic means at least you are not academically 'lonely.'
- Actor와 변수에 대한 고민이 필요
  - 생태학적 오류(ecological fallacy)
  - 투표는 유권자가, 그러나 자료 관측은 집계 unit에서
- 정교하고 세련된 방법론적 모델이 필요
  - 생태학적 추론(ecological inference)
  - 시계열적 방법(time-serial techniques)
  - 회귀불연속 디자인(regression discontinuity)

예컨대 다음과 같은 질문들

- 한국 민주화 과정 주도세력들은 누구이며 그들의 투표행태는 어떠했는가?
- 한국에서 부동산 가격 변동은 정당지지에 어떠한 영향을 미치는가?
- 독일 노동자 및 사회당원들은 히틀러를 얼마나 지지했는가?
- 미국 소수인종 유권자들의 투표참여율 및 후보자 지지
- 분할투표(split-ticket voting)는 누가, 얼마나, 왜 하는가?
- 한국에서 현직자 프리미움(incumbent advantage)은 얼마나 될까?

8

집계자료의 예시

- 중앙선거관리위원회 자료
  - 투개표결과: 투표소(약 18,000); 읍면동(3,500); 시군구(300)
  - 후보자 자료(대선, 총선, 지선 각 선거구)
- 통계청 센서스
  - 읍면동/시군구 별 인구학적 구성, 직업 및 경제상황
- 부동산 114
  - 읍면동 별 지가 변동 자료

9

Selected "Literature"

박원호. 2013. "생태학적 추론(Ecological Inference) 서설: 통계학적 연금술, 혹은 선거사 연구의 희망?" 『평화연구』 21(2): 395–426.

Park, Won-ho, Michael J. Hanmer, and Daniel R. Biggers. 2014. "Ecological Inference under Unfavorable Conditions: Straight and Split-Ticket Voting in Diverse Settings and Small Samples." *Electoral Studies* 36: 192–203.

Choi, Jae-in, and Won-ho Park. 2012. "Conditional Pocketbook Voting and Clarity of Responsibility in Korea." *Korean Political Science Review* 46(6): 85–107.

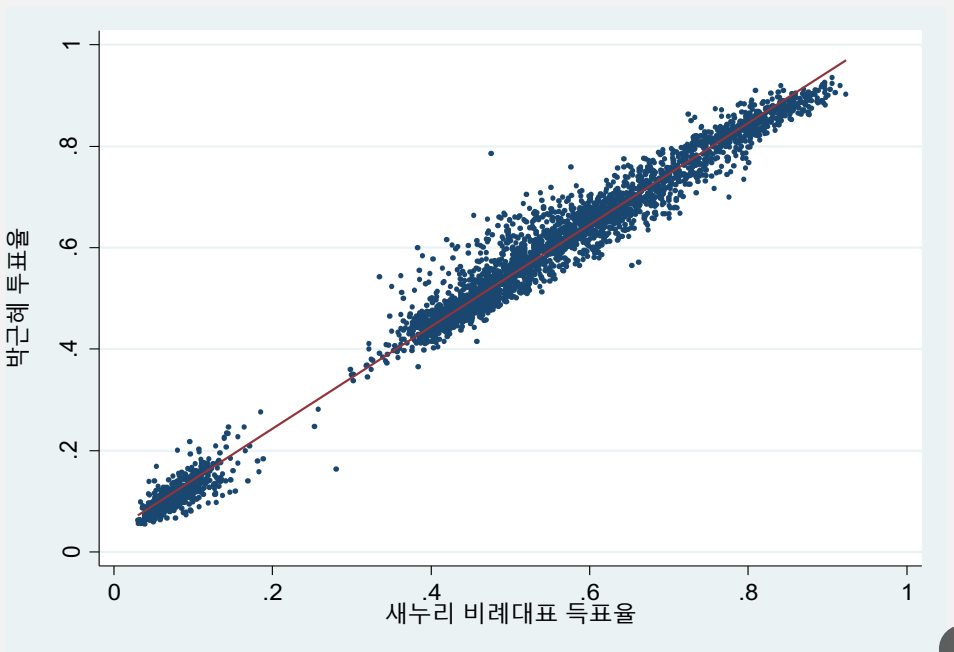
Kang, Woo Chang, Won-ho Park, and B. K. Song. 2018. "The Effect of Incumbency in National and Local Elections: Evidence from South Korea." *Electoral Studies* 56: 47–60.

박원호. 2009. "부동산 가격 변동과 2000년대의 한국 선거: 지역주의 '이후'의 경제투표에 대한 방법론적 탐색." 『한국정치연구』 18(3): 1–28.

10

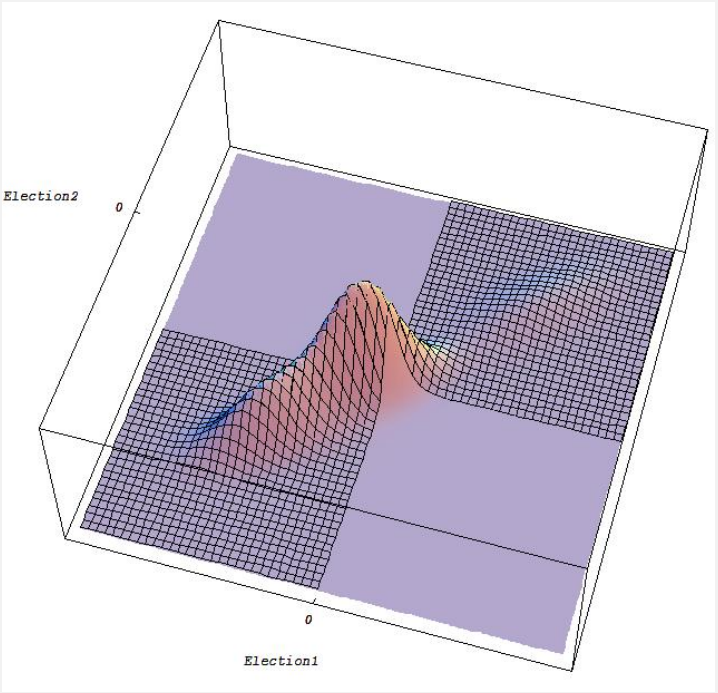
Selected Examples

2012년 총선과 대선: 읍면동 득표율



12

The Latent Structure and Estimation



13

투표자 전환율의 추정

Parameter	Panel	Ecol. Regression	King	Partisan Model
새누리→박근혜(*p*)	.95	1.05	.98	.96
(표준오차)	(.009)	(.004)	(.002)	(.002)
야권*→박근혜(*q*)	.15	.05	.09	.12
(표준오차)	(.016)	(.001)	(.002)	(.002)
N	992	3449	3449	3449

* 야권은 민주당과 통합진보당 투표자들을 합친 것임.

14

Voter Transition in Florida, 2000

Senate Race

	Rep	Dem	Others
Bush	0.367	0.052	0.009
Gore	0.042	0.497	0.017
Nader	0.005	0.008	0.004

Ballot Images, N = 2.8 Million

Senate Race

	Rep	Dem	Others
Bush	0.376	0.041	0.011
Gore	0.031	0.502	0.021
Nader	0.007	0.009	0.004

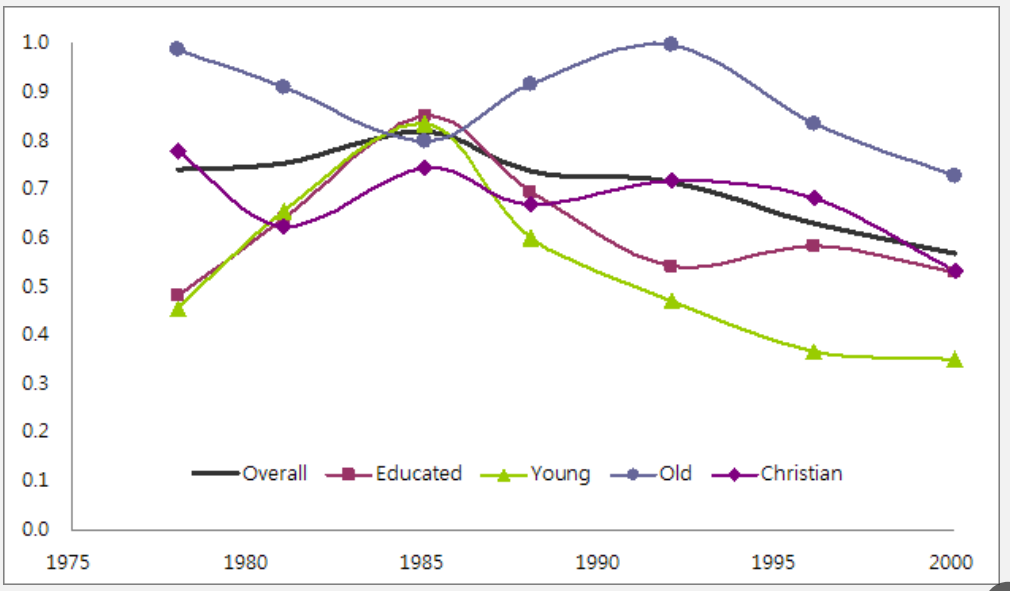
Precinct Data, N = 2,894

Voter Transition Estimates between Presidential and Senatorial Races, Ten Florida Counties

"Punch Card Counties": Broward, Highland, Hillsborough, Lee, Marion, Miami-Dade, Palmbeach, Pasco, Pinellas, Sarasota

15

한국 총선 집단별 투표 참여율: 1978~2000년, Ecological Estimates



16

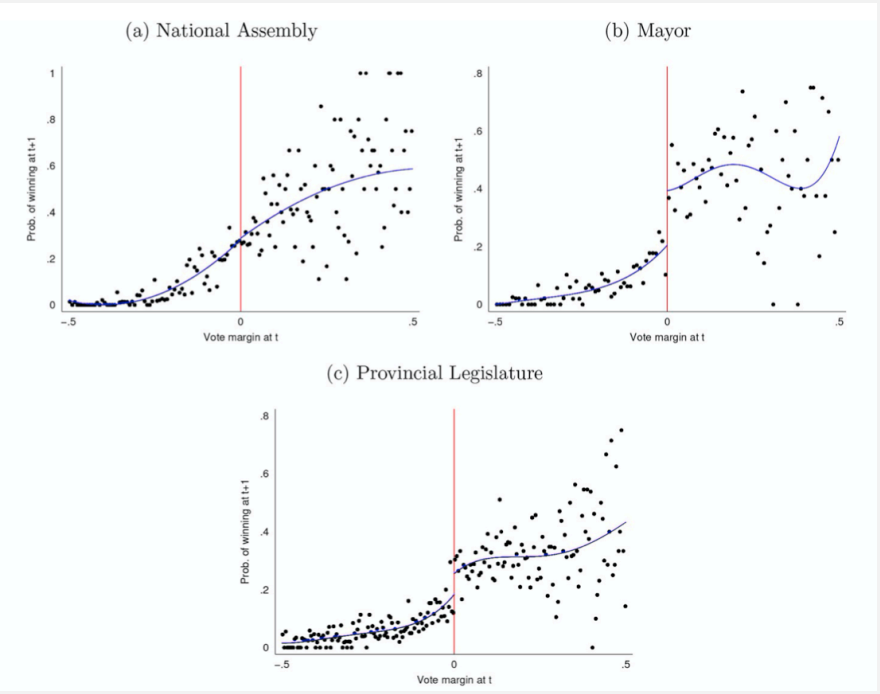
부동산 가격변동과 2000년대의 한국선거

대통령선거			지방선거		
	자가 소유	미소유		자가 소유	미소유
2002(제16대)(민)	.017**	.025*	2002(광역장)(민)	.112*	.059
2007(제17대)(민)	.121	-.032**	2002(광역의회)(민)	.066*	.036*
			2002(광역비례)(민)	.079	.052**
국회의원선거			2002(기초단체장)(민)	.139*	-.217
	자가 소유	미소유			
2004(지역구) (민)	.104*	-.026*	2006(광역장)(민)	-.000*	-.119*
2004(비례대표) (민)	.121	-.070	2006(광역의회)(민)	.003	-.104
2008(지역구) (한)	.082*	.006**	2006(광역비례)(민)	.021*	.020
2008(비례대표) (한)	.169**	.032**	2006(기초단체장)(민)	-.044	-.132

값들은 부동산 가격이 100% 상승했을 때 당시 여당 득표에 미칠 영향을 추정한 것임.

17

한국선거에서의 현직자 효과



MEMO

SNUAC
Seoul National University Asia Center
서울대학교아시아연구소

2부 선거자료 활용사례

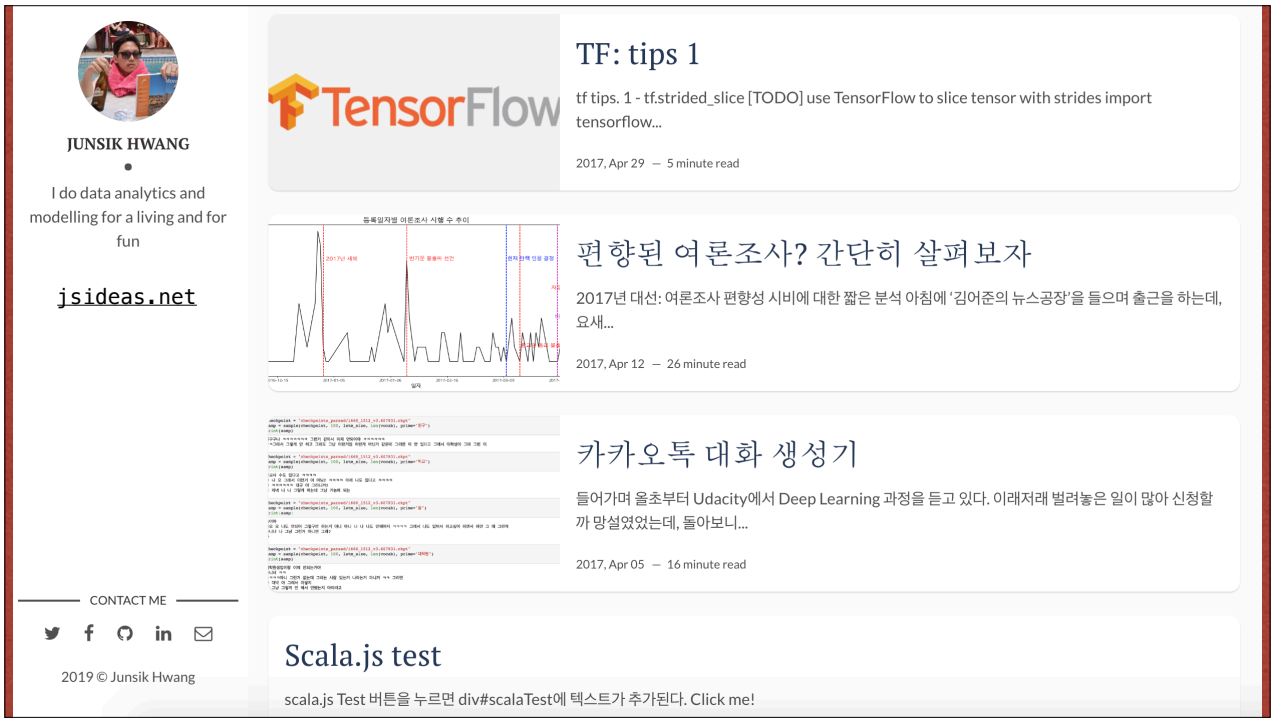
3. 파이썬을 이용한 간단한 여론조사 분석기

황준식 연구원 (넥슨코리아)

파이썬을 이용한 간단한 여론조사 분석기

황준식 / 넥슨코리아

2020.02.05



JUNSIK HWANG

I do data analytics and modelling for a living and for fun

jsideas.net

CONTACT ME

2019 © Junsik Hwang

TF: tips 1

tf tips. 1 - tf.strided_slice [TODO] use TensorFlow to slice tensor with strides import tensorflow...

2017, Apr 29 — 5 minute read

편향된 여론조사? 간단히 살펴보자

2017년 대선: 여론조사 편향성 시비에 대한 짧은 분석 아침에 '김어준의 뉴스공장'을 들으며 출근을 하는데, 요새...

2017, Apr 12 — 26 minute read

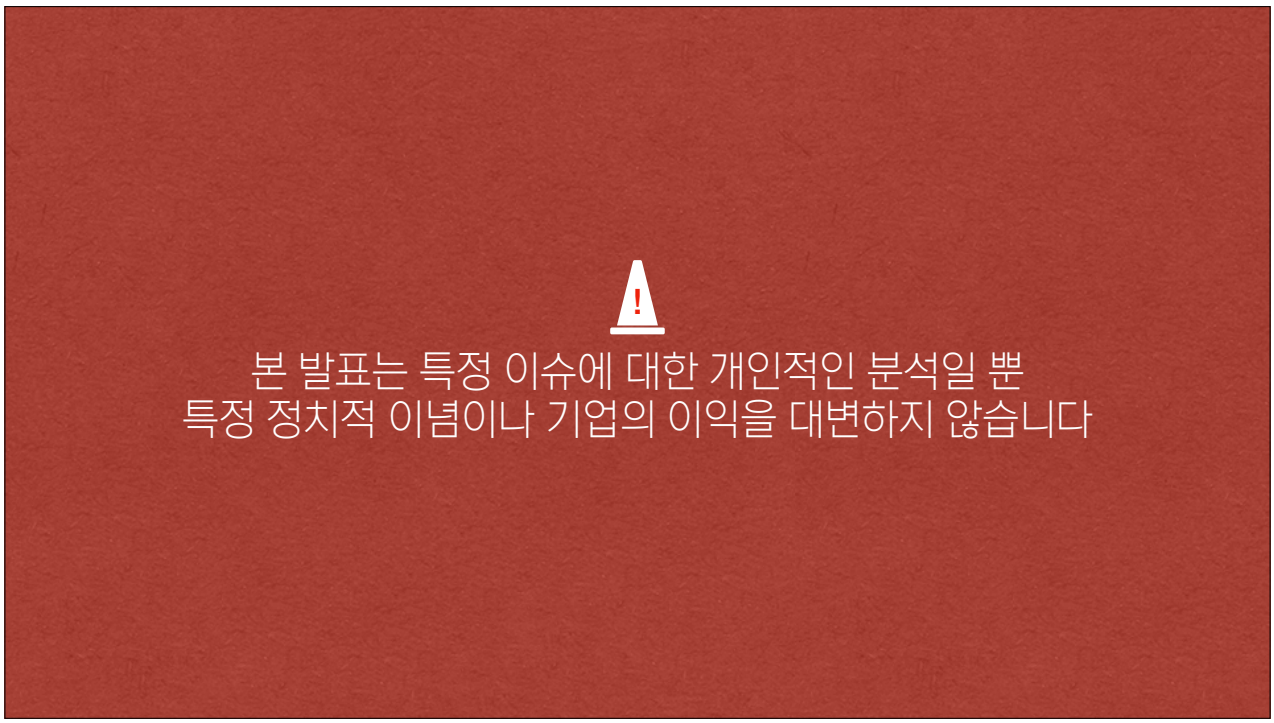
카카오톡 대화 생성기

들어가며 올초부터 Udacity에서 Deep Learning 과정을 듣고 있다. 이래저래 벌려놓은 일이 많아 신청할까 망설였었는데, 돌아보니...

2017, Apr 05 — 16 minute read

Scala.js test

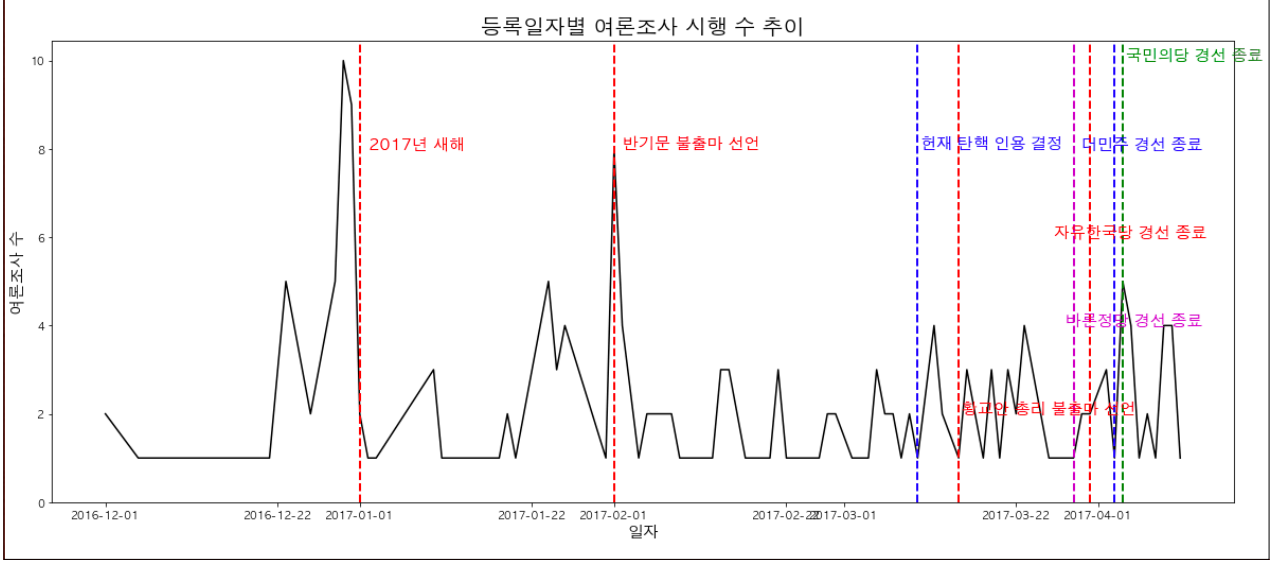
scala.js Test 버튼을 누르면 div#scalaTest에 텍스트가 추가된다. Click me!


본 발표는 특정 이슈에 대한 개인적인 분석일 뿐
특정 정치적 이념이나 기업의 이익을 대변하지 않습니다


분석 배경
여론조사가 수상하다?

탄핵 후 치러진 2017년 대선: 국민은 누구를 지지하는가?

국민은 누구를 지지하는가? 여론조사로 알아보자–



여론조사가 수상하다? 〈출근길 들었던 라디오 방송〉

김어준 :
그런데 이 조사에서 표본추출, 샘플링과정에 문제가 있다고 지적하셨는데
이 문제점은 어떻게 발견하셨는지, 그리고 어떤 지점이 이상한지를 좀 차근차근 설명해 주십시오.

김재광 :
사실 이번 대선은 굉장히 쉬운 것처럼 돼가지고 사람들이 별로관심이 없었는데요,
그런데 갑자기 최근에 4월 초쯤에 갑자기 여론조사결과가 반전하는 양상을 보이기 시작했는데요.
그런데 가만히 보니까 이게 하루 조사만으로 발표하는 것을 보고서 그건 좀 이상하다는 생각을 했었습니다.
보통 2, 3일은 걸려야 되는데 하루 만에 조살 하면 결과가 좀 더 보수적으로 나올 수 있거든요. … (중략)

출처: TBS [김어준의 뉴스공장]

여론조사가 수상하다? 〈출근길 들었던 라디오 방송〉

수상한 여론조사는 어떤 특징이 있는가?
- “하루 만에 조사했다 그러면 장히 이상하다고 느껴지거든요” -〉 여론조사 기간이 짧다
- “조사대상 규모가 확 줄어들었지만 2000명 응답은 그대로” -〉 응답율이 높다
- “무선전화조사에서 국번이 60개만 조사된다” -〉 국번당 조사된 번호 수가 많다
- “비적격이 상당히 걸러지고 나서 컨택이 됐다” -〉 낮은 비적격 비율

“선관위 산하에 여심위라고..” “홈페이지에 등록이 되게 돼있죠”

〉〉 정말 그런지 데이터로 살펴보면 어떨까?

출처: TBS [김어준의 뉴스공장]

데이터 수집

여론조사심의위원회 with BeautifulSoup

여론조사심의위원회 〉〉 깔끔하게 정리된 등록여론조사

중앙선거여론조사심의위원회 | 제도안내 | 알림마당 | 도움마당 | 등록마당 | 참여마당 | 위원회소개

Total 6077건

등록번호	조사기관명	조사의뢰자	여론조사명칭	등록일	지역	결정사항
6470	(주)한국리서치	KBS	전국 국회의원선거 비례대표국회의원선거 정당지지도 국정운영평가 및 현안	2020-01-23	전국	-
6469	(주)한국리서치	KBS	전국 국회의원선거 비례대표국회의원선거 정당지지도 국정운영평가 및 현안	2020-01-23	전국	-
6468	(주)코리아리서치인터내셔널	제주MBC, 제주CBS, 제주신보, 제주의소리	제주도 서귀포시 국회의원선거 정당지지도 기타 서귀포시선거구 제21대 총선, 사회 및 지역현안 등	2020-01-23	제주도	-
6467	(주)한국리서치	KBS	전국 국회의원선거 비례대표국회의원선거 정당지지도 국정운영평가 및 현안	2020-01-23	전국	-
6466	(주)코리아리서치인터내셔널	제주MBC, 제주CBS, 제주신보, 제주의소리	제주도 제주시 국회의원선거 정당지지도 기타 제주시을선거구 제21대 총선, 사회 및 지역현안 등	2020-01-23	제주도	-
6465	(주)코리아리서치인터내셔널	제주MBC, 제주CBS, 제주신보, 제주의소리	제주도 제주시 국회의원선거 정당지지도 기타 제주시갑선거구 제21대 총선, 사회 및 지역현안 등	2020-01-23	제주도	-
6463	(주)알앤써치	남도일보,뉴스1광주전남취재본부	전라남도 해남군 완도군 진도군 국회의원선거 정당지지도	2020-01-22	전라남도	-
6462	(주)알앤써치	남도일보,뉴스1광주전남취재본부	전라남도 여수시 국회의원선거 정당지지도	2020-01-22	전라남도	-
6461	(주)알앤써치	남도일보,뉴스1광주전남취재본부	전라남도 영암군 무안군 신안군 국회의원선거 정당지지도	2020-01-22	전라남도	-

여론조사 결과등록
가상번호 신청등록
불공정 여론조사 신고
TOP

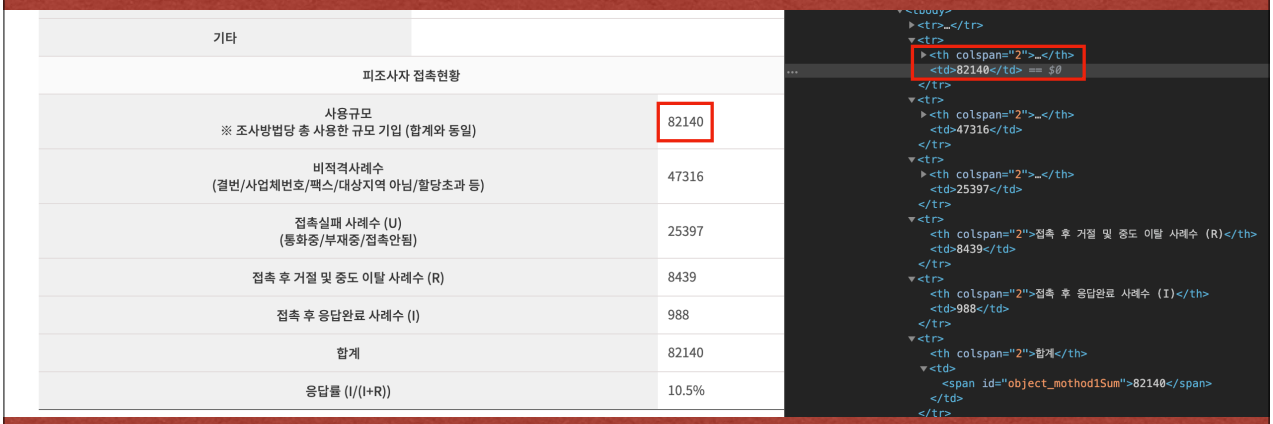
조사 방법까지도 조회 가능! -〉 데이터 분석 가능!

• 조사방법

조사방법 1			유선전화면접
			23.5 %
피조사자 선정방법			
조사대상			전국에 거주하는 만 19세 이상 남녀
표본 추출틀	전체	추출틀	유선전화번호 기타
		규모	48,046개
		구축방법	31,159개 KT DB 국번을 근거로 0000-9999까지 랜덤생성
표본추출방법			RDD 6,336개 번호 사용
기타			
피조사자 접촉현황			
사용규모 ※ 조사방법당 총 사용한 규모 기입 (합계와 동일)			6336
비적격사례수 (결번/사업체번호/팩스/대상지역 아님/할당초과 등)			2417

하나하나 클릭해서 엑셀에 적기 귀찮다.. 그럴땐 BeautifulSoup

하나하나 클릭해서 엑셀에 적기 귀찮다.. => BeautifulSoup



기타	
피조사자 접촉현황	
사용규모 ※ 조사방법당 총 사용한 규모 기입 (합계와 동일)	82140
비적격사례수 (결번/사업체번호/팩스/대상지역 아님/할당초과 등)	47316
접촉실패 사례수 (U) (통화중/부재중/접촉안됨)	25397
접촉 후 거절 및 중도 이탈 사례수 (R)	8439
접촉 후 응답완료 사례수 (I)	988
합계	82140
응답률 (I/(I+R))	10.5%

```
<tr>…</tr>
<tr>
  <th colspan="2">…</th>
  <td>82140</td> == $0
</tr>
<tr>
  <th colspan="2">…</th>
  <td>47316</td>
</tr>
<tr>
  <th colspan="2">…</th>
  <td>25397</td>
</tr>
<tr>
  <th colspan="2">접촉 후 거절 및 중도 이탈 사례수 (R)</th>
  <td>8439</td>
</tr>
<tr>
  <th colspan="2">접촉 후 응답완료 사례수 (I)</th>
  <td>988</td>
</tr>
<tr>
  <th colspan="2">합계</th>
  <td>
    <span id="object_mothod1Sum">82140</span>
  </td>
</tr>
```

정보가 구조적으로 잘 정돈되어 있는 경우
웹크롤링으로 작업시간 단축 가능!

BeautifulSoup으로 쉽고 편하게 데이터를 긁어보세요

```
def meta_crawler(page_num):
  """
  여론조사심의위원회 – 여론조사결과 등록현황 페이지에서 여론조사 명칭과 링크를 수집한다
  """

  for page in range(page_num):
    page = page + 1
    url = "http://www.nesdc.go.kr/portal/bbs/…/…/&pageIndex={}".format(page)
    r = requests.get(url)
    data = r.text
    soup = BeautifulSoup(data, "html5lib")
    bd_list = soup.find("div", "bd_list") # 지정한 문서 영역을 파싱하는 부분
    …
```

BeautifulSoup 공식 문서

데이터 전처리

선별 & 정제 with Pandas

500개 여론조사 데이터 확보

	area	client	client_type	inst	method_ratio_sum	methods	participants	period	period_we_bool	regDate	regNo	total_response_ratio	vote	vote_title
495	전국	매일경제·MBN'레이더P'	전국단위	신문 리얼미터	100.0	[{'method': '유선', 'ratio': '15%', 'sample_size...	1010	2	False	2016-11-15	3187	11.2	기타	정례조사(전국 정례조사 2016년 11월 11일 일간집계)
496	전국	매일경제·MBN'레이더P'	전국단위	신문 리얼미터	100.0	[{'method': '유선', 'ratio': '15%', 'sample_size...	1011	2	False	2016-11-10	3186	11.1	기타	정례조사(전국 정례조사 2016년 11월 10일 일간집계)
497	전국	(주)데일리안	인터넷언론	(주)알앤써치	100.0	[{'method': '무선', 'ratio': '100%', 'sample_siz...	1510	2	True	2016-11-10	3185	8.3	제19대 대통령선거(정당지지율조사)	정기조사(전국 정기조사 정당지지율조사 알앤써치바로미터11월2주차)
498	전국	매일경제·MBN'레이더P'	전국단위	신문 리얼미터	100.0	[{'method': '유선', 'ratio': '15%', 'sample_size...	1521	3	False	2016-11-10	3184	13.1	기타	정례조사(전국 정례조사 2016년 11월 2주 주중집계)
499	전국	매일경제·MBN'레이더P'	전국단위	신문 리얼미터	100.0	[{'method': '유선', 'ratio': '15%', 'sample_size...	1012	2	False	2016-11-10	3183	13.4	기타	정례조사(전국 정례조사 2016년 11월 9일 일간집계)

대통령 선거 관련 여론조사 & 2016.12 이후만 필터링

pandas 공식 문서

```
import pandas as pd


## 2016년 12월 1일 이후의 데이터만 보기로 한다.
df.regDate = df.regDate.map(lambda x: pd.to_datetime(x))
df = df[df.regDate >= '2016-12-01']


## 대통령, 대선 관련 여론조사만 확보한다.
def isPresidential(title):
    if ("대통령" in title) | ("대선" in title):
        return True
    else:
        return False

df['isPresidential'] = df.vote_title.map(isPresidential)
df = df[df.isPresidential == True]


len(df) # 183
```

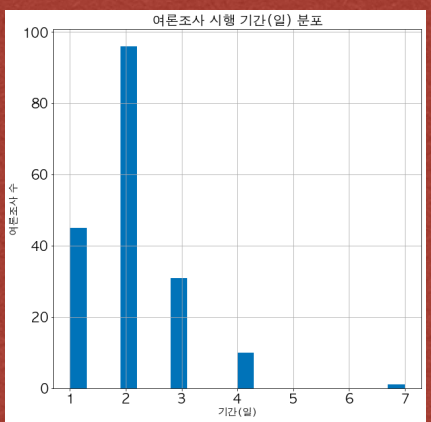
유선 비율과 국번 당 응답 횟수를 계산

```
## 여론조사의 유선 비율 측정
def wire_ratio(a_list):
    k = {m['method']: m['ratio'] for m in a_list}
    if '유선' in k.keys():
        res = float(k['유선'].replace("%", "")) / 100
    else:
        res = 0
    return res

df['wire_ratio'] = df.apply(lambda row: wire_ratio(row['methods']), axis=1)


## 국번 당 응답 성공 횟수
def success_to_gb_r_calc(a_list):
    gb_size = np.sum([m['gb_size'] for m in a_list])
    success_size = np.sum([m['success_size'] for m in a_list])
    return round(success_size / gb_size, 4)

df['gb_ratio'] = df.apply(lambda row: success_to_gb_r_calc(row['methods']), axis=1)
```

데이터 분석

여론조사는 정말 수상했나 with matplotlib

파이썬에서 차트를 그리는 가장 쉬운 방법 matplotlib

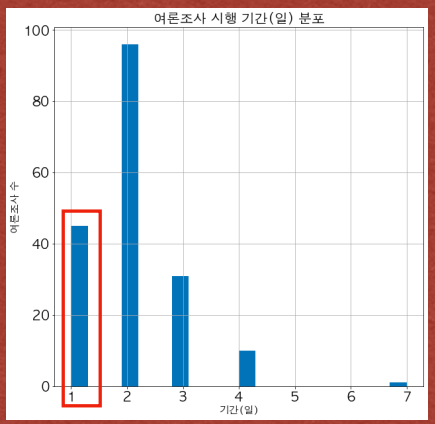


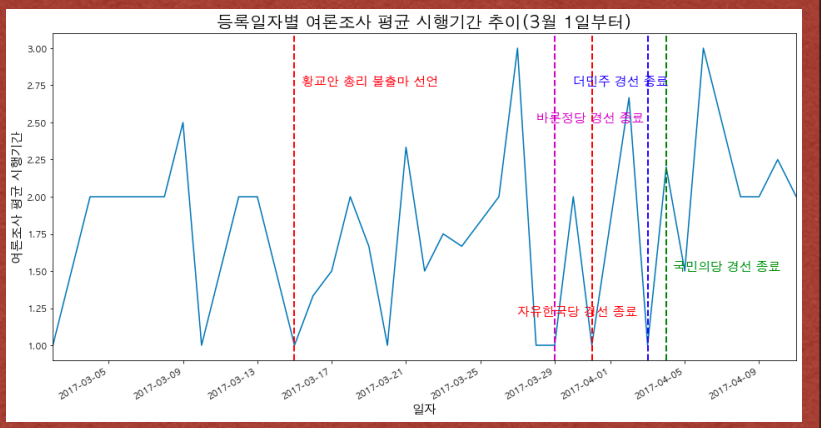

```
import matplotlib.pyplot as plt

plt.figure(figsize=(10, 10), facecolor='white')
df.period.hist(bins=20)

plt.tick_params(axis='both', which='major', labelsize=20)
plt.title("여론조사 시행 기간(일) 분포", fontsize=20)
plt.xlabel("기간(일)", fontsize=15)
plt.ylabel("여론조사 수", fontsize=15)

plt.show()
```

가설: 수상한 여론조사는..

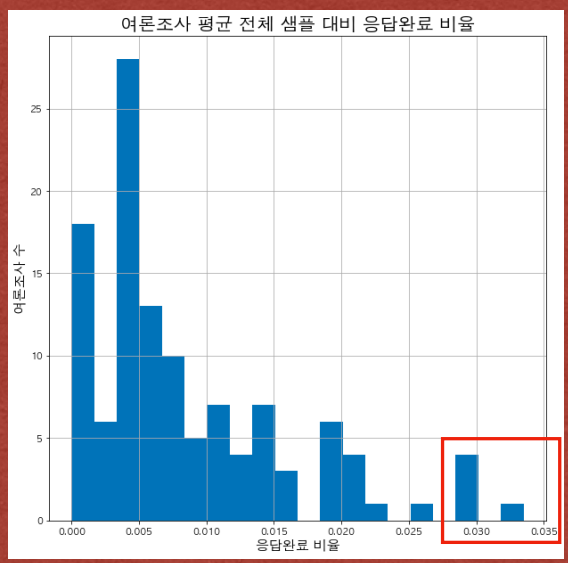
- 여론조사 기간이 짧다 (하루)
- 전체 대상 중 응답 완료 사례가 지나치게 높다
- 국번당 응답 성공 횟수가 지나치게 높다
- 전체 대상 중 비적격 사례의 비율이 지나치게 낮다

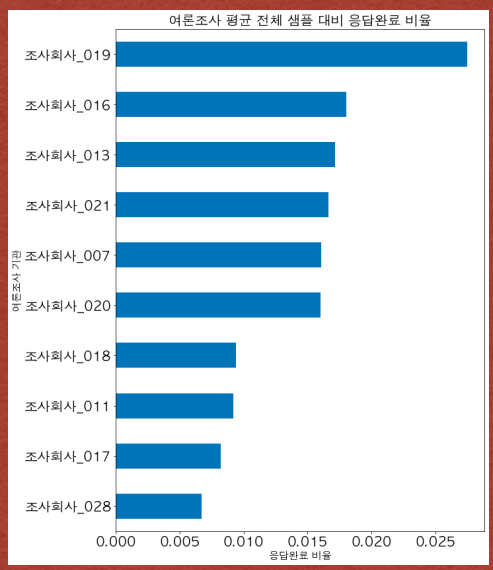
조사 기간: 하루짜리 조사도 많고 긴급 이벤트에 의해 시행되기도-



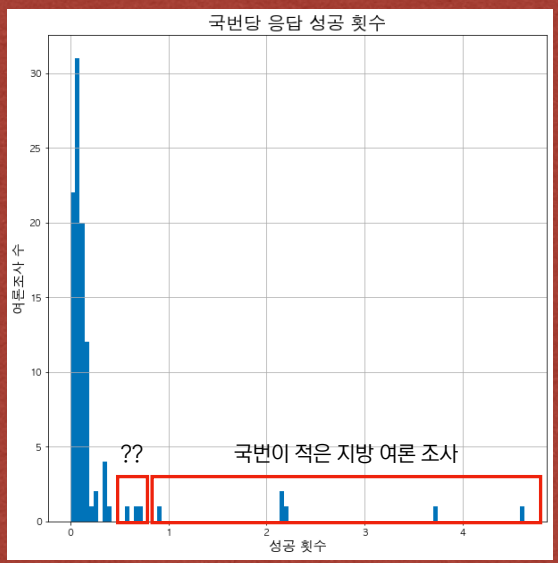


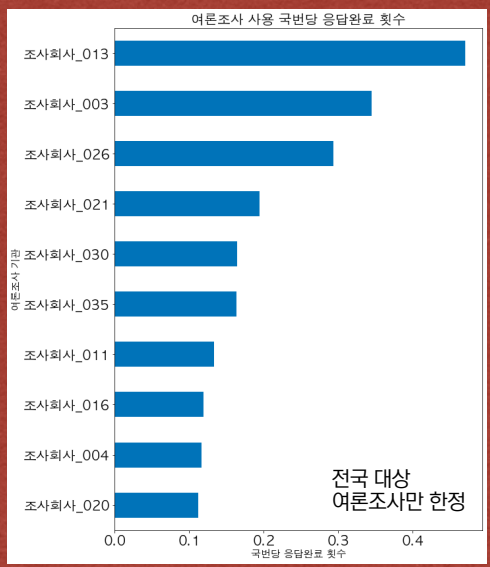
응답 완료 비율: 일부 높은 비율 발견됨. 효율성 vs. 부정의 소지



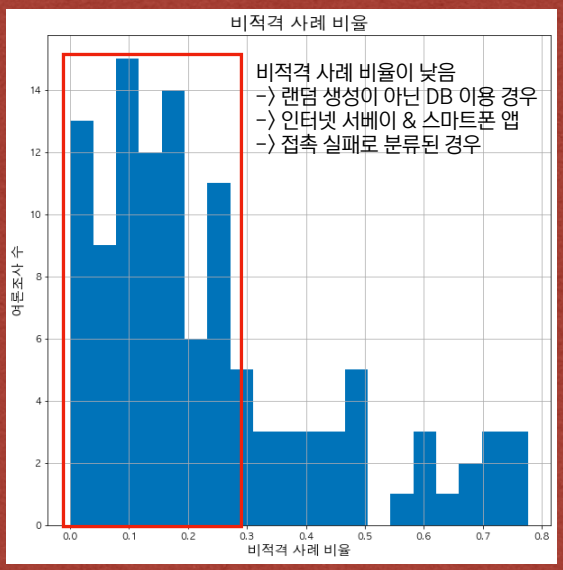


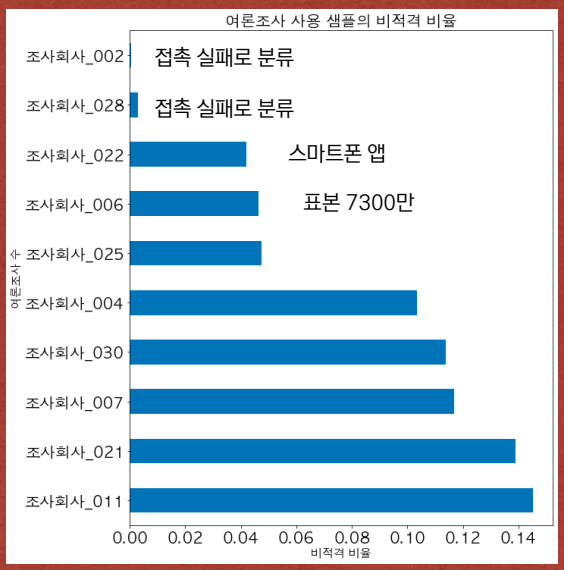
국번당 응답 성공 횟수: 전국 여론조사에서도 횟수가 높은 일부 조사들





비적격 비율: 0.3 미만이 많음. 비적격 수와 샘플 크기 기록 일관성 문제





가설: 수상한 여론조사는..

- ~~여론조사 기간이 짧다 (하루)~~
- 전체 대상 중 응답 완료 사례가 지나치게 높다: 일부 발견
- 국번당 응답 성공 횟수가 지나치게 높다: 일부 발견
- 전체 대상 중 비적격 사례의 비율이 지나치게 낮다: 일부 발견 & 데이터 일관성 부족

문제점: 기관별 평균을 내었기 때문에, 수상한 여론 *조사*가 묻힌 것이 아닐까?

김재광 : 사실 동일한 회사에서 이게 매월 정기조사거든요 동일한 회사에서 동일한 매달하는 거였습니다.
그러면 사실 보통 중간에 그 방법을 잘 바꾸지 않거든요,
왜냐하면 그게 추세를 봐야 하기 때문에 그런데 이 조사는 이상하게도 3월 조사하고 너무 방식이 바뀐 겁니다.

〉〉 기관별로 여론조사 간 similarity를 보자!

기관별 여론조사 간 similarity

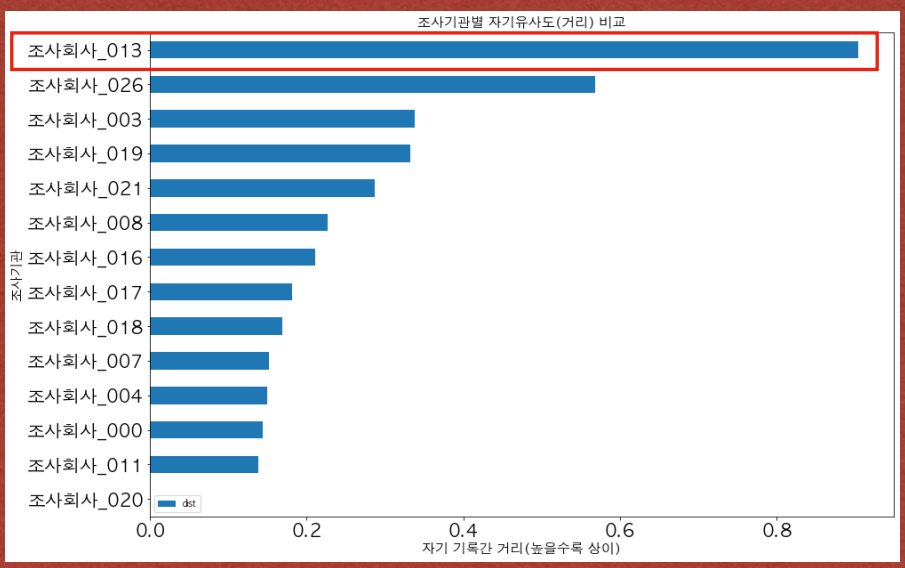
- 개별 여론조사를 {유선 비율, 응답 성공율, 국번당 성공 횟수, 부적격 비율} 벡터로 표현
  - ex) {0.235, 0.0114, 0.0256, 0.0809}
- 기관별로 여론조사 쌍 조합을 만든 후, 조합간 euclidean distance를 구함
  - ex) A 기관의 여론조사가 3개인 경우, AB, BC, CA로 3개 조합 생성
- 평균 거리가 높을수록 조사 방식이 달라졌음을 의미
- 요소가 NaN인 여론조사는 제외

추가된 조건

-〉 “전국” 여론조사로 한정

-〉 4개 요소간 스케일 차이를 보정하기 위해 0~1 사이로 MinMaxScaling을 수행

조사회사_013의 여론조사가 가장 변화가 컸던 것으로 드러나..



조사회사_013의 여론조사가 가장 변화가 컸던 것으로 드러나..

- 같은 클라이언트에 약 한달 간격으로 수행한 대선 여론조사
- 4월 9일자 여론조사: 이전에 비해 응답 성공율이 크게 오르고, 비적격 비율이 크게 낮아짐

	inst	client	regNo	regDate	wire_ratio	ssr	gb_ratio	inapp_ratio
376	조사회사_013	KBS·연합뉴스	3456	2017-02-06	0.45	0.0090	0.5672	0.4731
190	조사회사_013	KBS·연합뉴스	3555	2017-03-12	0.44	0.0090	0.1823	0.5924
105	조사회사_013	KBS·연합뉴스	3644	2017-04-09	0.40	0.0335	0.6604	0.0852

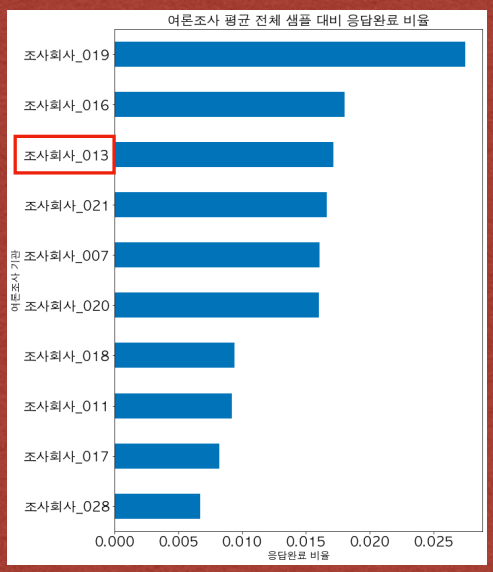
표본 추출 규모가 크게 달라짐

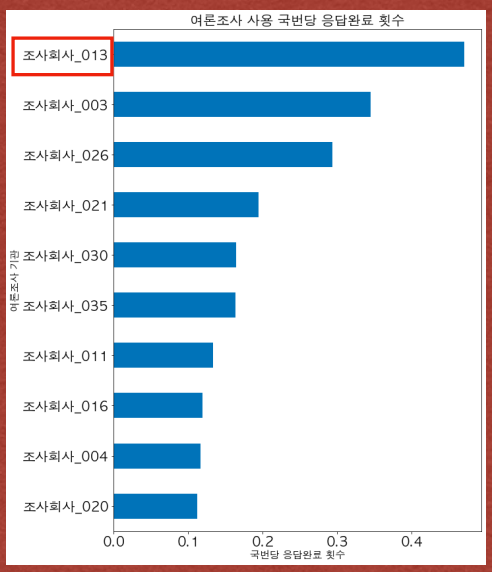
			2/6	3/12	4/9
조사방법 2			무선전화면접	무선전화면접	무선전화면접
			55 %	56 %	60 %
피조사자			선정방법	선정방법	선정방법
조사대상			전국에 거주하는 만 19세 이상 성인 남녀	전국에 거주하는 만 19세 이상 성인 남녀	전국에 거주하는 만 19세 이상 성인 남녀
표본 추출틀	전체	추출틀	무선전화번호 기타	무선전화번호 기타	무선전화번호 기타
		규모	100,000	120,001	30,000
		구축방법	82개 국번별 0001~9999까지 랜덤으로 생성	8,031개 국번별 0001~9999까지 랜덤으로 생성	60개 국번별 0001~9999까지 랜덤으로 생성
표본추출방법			RDD	RDD	RDD
기타			88,375개 사용	90,753개 사용	12,625개 사용

비적격 사례 수의 상대적 크기도 크게 달라짐

	2/6	3/12	4/9
사용규모 ※ 조사방법당 총 사용한 규모 기입 (합계와 동일)	88375	90753	12625
	8:5	9:6	12:2
비적격사례수 (결번/사업체번호/팩스/대상지역 아님/할당초과 등)	53287	62775	2650
접촉실패 사례수 (U) (통화중/부재중/접촉안됨)	28157	21555	2979
접촉 후 거절 및 중도 이탈 사례수 (R)	5816	5284	5787
접촉 후 응답완료 사례수 (I)	1115	1139	1209
합계	88375	90753	12625
응답률 (I/(I+R))	16.1%	17.7%	17.3%

응답완료 비율과 국번당 응답 완료 횟수에서 상위에 오르기도-





그 후

여론조사는 정말 수상했던 것인가

선관위, 여론조사업체 '조사회사_013'에 과태료 1500만원 부과

중앙일보 2017.04.19 이가영 기자

중앙선거관리위원회 여론조사심위원회(여심위)가 19일 여론조사업체 조사회사_013(익명)에 과태료 1500만원을 부과했다. 여론조사를 의뢰한 KBS는 해당 업체에 엄중히 경고하는 공문을 보내기로 했다.

(중략)

선관위에 따르면 조사회사_013은 KBS와 연합뉴스가 의뢰, 지난 8일부터 9일까지 실시한 여론조사의 표본추출틀 전체 규모가 유선전화 7만6500개, 무선전화 5만개임에도 유선과 무선 각각 3만개를 추출, 사용했다고 여심위 홈페이지에 사실과 다르게 등록했다.

(중략)

또 비적격 사례 수도 유선 2만5455개, 무선 1만4983개이고 접촉실패 사례 수는 유선 1만1863개, 무선 2만4122개지만 여심위 홈페이지에는 비적격 사례 수가 유선 2460개, 무선 2650개고 접촉 실패 사례 수도 유선 2766개, 무선 2979개라고 사실과 다른 내용을 등록했다.

(중략)

여심위는 다만 "해당 여론조사에서 당초 제기됐던 무선전화 국번 수와 비적격 사례 수 등의 과소함을 이유로 자체구축 DB를 사용했다는 주장의 경우 확인한 결과 특정 DB를 사용한 흔적은 없었다"고 밝혔다.

중앙일보 기사 원문

Takeaway

- 데이터는 그 자체로 완벽하지 않습니다. 수집 방법, 가공 방법에 따라 해석이 얼마든지 달라질 수 있어요.
  그렇기에 정치적/사회적 의미가 무거운 여론조사는 더 엄정한 관리와 감독이 필요하지 않을까요?
- 우리도 많은 관심을 갖고 여론조사심의위원회에 공개된 여론조사 자료를 살펴봅시다.
- 여론조사심의위원회에 기록된 여론조사의 정보는 일부 누락되거나 다르게 기재된 부분이 있어 주의하세요.
- 파이썬으로 하면 더 쉽고 빠르고 재현가능하답니다.
  - BeautifulSoup: 웹크롤링
  - pandas, numpy: 데이터 전처리 및 가공
  - matplotlib: 데이터 시각화

감사합니다.

MEMO

MEMO