응용통계연구 (2009)
22(2), 375–385

전화조사의 체계적 편향
- 2007년 대통령선거 여론조사들에 대한 메타분석 -

김세용[1] · 허명회[2]

[1]고려대학교 통계학과, [2]고려대학교 통계학과

(2008년 11월 접수, 2008년 12월 재택)

요 약

2007년 12월의 대통령선거 과정에서 행해진 수많은 전화 여론조사에서 이명박 후보가 일관되게 앞자리를 지켰다. 많은 수의 여론조사가 시행되는 경우 개별 여론조사의 표본추출오차는 상쇄되어 소멸한다. 그러나 일정한 편향은 많은 수의 조사가 실행된다 하더라도 잔존하므로 중요한 문제가 아닐 수 없다.

우리나라의 전화조사는 지역, 성과 연령대를 고려한 할당추출(quota sampling)을 사용하며 대부분 전화번호부를 표집 틀로 한다. 조사 거절률도 높은 편이다. 이에 따라 조사표본들이 할당변인 외의 인구사회적 속성에서 모집단을 잘 대표하지 못할 수 있다. 이 연구의 목적은 허명회 등 (2004)의 연구 방법론을 2007년에 수행된 대통령선거 여론조사 18개 사례에 적용하여 다음 물음에 답하는 데 목적이 있다.

- 물음 1. 각 후보에 대한 선호도 또는 지지율에 체계적 편향이 있지 않았는가?

- 물음 2. 편향이 있었다면, 그 원인이 어디에 있는가?

첫째 물음에 답하기 위하여 2007년 11월 이후 시행된 11개 사례 자료에 지역, 성과 연령대 외에 직업과 학력까지 고려한 반복비례가중법(rim weighting)을 적용해보았다. 그 결과, 이명박 후보의 지지율이 평균 1.4%P 과다 추정되었던 것으로 나타났다. 반면, 정동영 후보의 지지율은 평균 0.6%P 과소 추정되었고 이에 따라 두 후보간 지지율 차이가 2.0%P (= 1.4 + 0.6) 과다하게 추정되었던 것으로 보여진다. 둘째 물음에 답하기 위하여 위의 11개 사례 자료에서 이명박 후보 지지를 종속변수로 하는 로지스틱 회귀 분석을 하였다. 그 결과, 전화조사 표본에서의 저학력자 과소 및 가정주부의 과다가 이명박 편향의 원인이 되는 것으로 밝혀졌다.

주요용어: 전화조사, 대통령 선거, 메타분석, 체계적 편향, 반복비례가중법, 로지스틱 회귀.

1. 연구 배경 및 목적, 주요 결과

우리나라 사회조사의 대부분은 전화조사로 시행되고 있지만 크게 두 가지의 방법론적 문제점을 가지고 있다. 첫째는, 전화번호부를 표집 틀로 하는데 전화번호부에는 휴대폰만 가진 가구와 실제 유선전화를 가지고 있으나 전화번호부에 등재하지 않은 가구가 포함되지 않으므로 모집단 포함률(population coverage)이 높지 않다는 점이고 (강현철 등, 2008), 둘째는, 재통화(call-back) 없이 집중적으로 낮 시간대에 표본을 확보한다는 점이다. 지역·성별·연령대 할당을 적용하지만 이렇게 얻는 전화조사 표본에는 각 할당 내에서 낮 시간대 재택성향이 높은 응답자들이 과다할 수밖에 없다 (허명회와 황진모, 2006). 이에 따라, 재택 성향과 다수의 사회적 요인 간 연관성에 의하여 조사표본에 특정 사회적 계층이 과다하게 대표되고 일부 계층은 과소하게 대표됨으로써 체계적 편향이 발생할 위험성이 있다. 우리나라에서

[2]교신저자: (136-701) 서울시 성북구 안암동 5가, 고려대학교 통계학과, 교수. E-mail: stat420@korea.ac.kr

최근 여론조사에 대한 사회적 관심도가 높아 각종 선거과정에서 많은 수의 여론조사가 빈번하게 시행되지만 아무리 많은 수의 여론조사를 평균 낸다고 하더라도 편향으로 인한 오차는 없어지지 않는다. 이와 같은 정량적 정보의 왜곡이 선거과정에서 특정 후보를 부당하게 유리하게 하고 반면 다른 후보를 부당하게 불리하게 할 수 있으므로 선거여론조사에서 체계적 편향이 발생하고 있는가를 감시할 사회적 책임이 통계전문가들에 있다고 하겠다.

이 연구의 목적은 다음 물음에 답하는 데 목적이 있다.

- 물음 1. 각 후보에 대한 선호도 또는 지지율에 체계적 편향이 있지 않았는가?
- 물음 2. 편향이 있었다면, 그 원인이 어디에 있었는가?

이런 물음들에 답하기 위하여 2007년 대통령선거 과정에서 시행된 여론조사 18개 사례의 조사자료를 수집하여 선행연구인 허명회 등 (2004)의 메타분석 방법론을 적용하기로 한다. 수집된 총 18개 조사자료 중 각 정당 후보자가 확정된 마지막 11개 조사자료에는 지역·성별·연령대 외에 학력과 직업까지 균형을 맞추기 위해 반복비례가중법(iterative proportional weighting, rim weighting)의 가중치를 적용하여 최종 추정치를 구하였다 (허명회 등, 2005). 이것을 지역·성별·연령대 칸 가중치를 적용해 얻은 조사기관들의 추정치와 비교해 보았다. 그 결과, 조사기관들의 이명박 후보의 지지율이 1.4%P 과다 추정되었고 반면 정동영 후보의 지지율이 0.6%P 과소 추정되었던 것으로 나타났다. 둘째 물음에 답하기 위하여 위의 11개 사례 자료에서 이명박 후보 지지를 종속변수로 하는 로지스틱 회귀 분석을 하였다. 그 결과, 조사표본에서의 저학력자 과소와 가정주부의 과다가 이명박 편향의 원인이 된 것으로 밝혀졌다.

2. 조사표본의 대표성

본 연구를 위해 우리나라의 조사업계를 주도하는 5개 조사전문기관(G, H, K, M, T)에서 2007년 3월부터 12월까지 실시한 18개의 대통령선거관련 전화조사 자료를 수집하였다. 18개 표본의 크기는 705에서 5,327 사이였고 그 중 각 정당후보가 확정된 11월과 12월에 조사된 사례가 11개였다.

조사기관별로 인구사회적 변수의 정의가 다소 다르기 때문에 통합된 메타 분석을 위해서는 일정한 기준으로 통일시킬 필요가 있다. 모두가 동의하기는 어렵겠지만 다음 기준으로 조사자료를 재정리하였다.

- 지역: 서울, 인천/경기, 강원/제주, 대전/충청, 광주/전라, 대구/경북, 부산/울산/경남
- 성별: 남, 여
- 연령대: 20대(19세 포함), 30대, 40대, 50대, 60대 이상
- 학력: 중졸이하, 고졸, 대재 이상
- 직업: 농/임/어업, 자영업/블루칼라, 화이트칼라, 가정주부, 학생, 기타/무직

50대와 60대 이상의 성향이 다를 수도 있을 것으로 판단되어 연령을 5개의 범주로 나눴으나 G사의 표본에 대하여는 50대와 60대 이상을 한 범주로 묶을 수밖에 없었다.

조사표본의 대표성을 평가하기 위하여는 모집단의 인구사회적 분포를 정확히 해둘 필요가 있다. 이 연구에서는 모집단 수치를 2005년 인구주택 총조사에서 구하였다. 이때 직업 분류가 가장 문제가 되는데, 현재 통계청에서 사용하고 있는 직업분류는 취업자를 (1) 의회의원, 고위임직원 및 관리자, (2) 전문가, (3) 기술공 및 준전문가, (4) 사무 종사자, (5) 판매 종사자, (6) 농업, 임업 및 어업 숙련 종사자, (7) 기능원 및 관련기능 종사자, (8) 장치, 기계조작 및 조립 종사자, (9) 단순 노무 종사자, (10) 기타 등으로 나눈다. 그러나 이러한 분류 체계는 여러 민간 조사기관들이 쓰고 있는 직업분류와 일치하지 않으므로 재정의를 통해 일치를 시도하였다. 선행연구와 이 연구에서 채택된 대응 관계는 표 2.1과 같다.

표 **2.1.** 조사전문기관과 통계청의 직업 분류 간 대응관계

조사전문기관 직업분류	통계청 표준 직업분류
농/임/어업농업	임업 및 어업 숙련종사자
자영업/블루칼라	서비스 종사자, 판매 종사자, 기능원, 장치/기계/조립/단순노무
화이트칼라	의회의원, 고위임직원 및 관리자, 전문가, 기술공 및 준전문가, 사무 종사자
가정주부	여자로서 일하지 않으며 일자리를 찾아보지 않은 사람(학생과 미혼 제외)
학생	학생
기타/무직	기타

표 **2.2.** 2000년과 2005년 인구주택 총조사에 의한 모집단의 인구사회적 속성 분포*

2000년			2005년		
지역	명	%	지역	명	%
-서울	7,050,649	22.3	-서울	7,665,000	21.5
-인천/경기	7,622,615	24.1	-인천/경기	9,446,113	26.6
-강원	1,028,287	3.3	-강원/제주	1,507,785	4.2
-대전/충청	3,168,052	10.0	-대전/충청	3,597,066	10.1
-광주/전라/제주	3,924,403	12.4	-광주/전라	3,758,740	10.6
-대구/경북	3,588,401	11.3	-대구/경북	3,850,447	10.8
-부산/울산/경남	5,249,006	16.6	-부산/울산/경남	5,767,255	16.2
성별	명	%	성별	명	%
-남	16,065,519	49.2	-남	17,465,323	49.1
-여	16,587,562	50.8	-여	18,127,083	50.9
연령대	명	%	연령대	명	%
-20대	7,945,164	24.3	-20대	7,971,593	22.4
-30대	8,280,181	25.4	-30대	8,209,067	23.1
-40대	6,948,359	21.3	-40대	8,023,940	22.5
-50대 이상	9,479,377	29.0	-50대	5,133,735	14.4
			-60대 이상	6,254,071	17.6
학력	명	%	학력	명	%
-중졸 이하	10,293,497	31.5	-중졸 이하	9,371,442	26.3
-고졸	12,094,814	37.1	-고졸	12,328,217	34.7
-대재 이상	10,254,855	31.4	-대재 이상	13,892,747	39.0
직업	명	%	직업	명	%
-농/임/어업	2,351,586	7.4	-농/임/어업	2,051,333	6.0
-블루칼라/자영업	9,253,164	29.0	-블루칼라/자영업	9,516,305	27.8
-화이트칼라	6,530,612	20.5	-화이트칼라	7,455,200	21.8
-가정주부	7,405,700	23.2	-가정주부	6,527,700	19.1
-학생	2,355,309	7.4	-학생	2,523,222	7.4
-기타/무직	3,974,947	12.5	-기타/무직	6,125,619	17.9

* 2000년 수치는 20세 이상. 2005년 수치는 19세 이상, 단 직업만 20세 이상.

표 2.2는 2000년과 2005년 인구주택 총조사에 의한 모집단의 인구사회적 속성을 나타내는 표이다. 2007년도 현재, 19세 이상에 선거권이 주어지므로 모집단을 19세 이상의 한국인 남녀로 정의할 수 있다. 단, 직업에 관해서는 인구주택 총조사 보고서에서 19세 이상에 대한 자료를 구할 수 없었기 때문에 20세 이상에 대한 통계로 대체하였다. 2000년 총 자료 수치는 20세 이상의 인구사회적 속성이다.

표 2.2에서 2000년과 2005년 모집단의 인구사회적 속성의 변화를 살펴보자. 지역별로는 2000년에 비해

표 2.3. 표본과 모집단의 구성비율 차이(%P)*

선행연구				본연구			
	평균	최소	최대		평균	최소	최대
지역				지역			
-서울	0.0	−0.9	1.2	-서울	−2.8	−15.4	0.1
-인천/경기	−1.2	−3.8	0.7	-인천/경기	−2.1	−10.6	0.7
-강원	0.1	−0.2	0.7	-강원/제주	1.1	−0.2	4.7
-대전/충청	0.3	−0.5	2.7	-대전/충청	1.9	−0.4	8.3
-광주/전라/제주	1.0	−0.5	3.2	-광주/전라	1.0	−0.9	6.5
-대구/경북	−0.1	−1.7	1.4	-대구/경북	−0.1	−1.5	3.4
-부산/울산/경남	0.0	−3.0	1.3	-부산/울산/경남	1.0	−0.4	11.3
성별				성별			
-남	0.4	−0.4	1.8	-남	−0.2	−4.0	1.2
-여	−0.4	−1.8	0.4	-여	0.2	−1.2	4.0
연령대				연령대			
-20대	−3.1	−10.3	3.6	-20대	−5.3	−12.3	0.6
-30대	1.3	−1.9	4.8	-30대	−1.9	−7.4	0.7
-40대	1.1	−3.0	4.2	-40대	1.1	−1.4	5.2
-50대 이상	0.7	−7.0	6.0	-50대	4.5	−0.2	20.1
				-60대 이상	4.5	−2.6	20.1
학력				학력			
-중졸 이하	−9.8	−16.7	−3.6	-중졸 이하	−5.7	−13.3	0.6
-고졸	1.5	−1.1	6.1	-고졸	−3.0	−8.3	3.6
-대재 이상	8.2	0.0	16.6	-대재 이상	8.7	−4.1	20.6
직업				직업			
-농/임/어업	−1.1	−4.9	2.2	-농/임/어업	−0.6	−2.3	3.3
-자영업/블루칼라	−2.6	−8.4	6.1	-자영업/블루칼라	−3.9	−6.8	2.1
-화이트칼라	−2.5	−7.3	7.9	-화이트칼라	−1.7	−10.3	4.7
-가정주부	8.9	2.0	13.6	-가정주부	11.3	4.0	18.1
-학생	−0.4	−4.3	3.2	-학생	0.9	−4.2	5.0
-기타/무직	−2.4	−6.6	1.1	-기타/무직	−6.0	−9.7	3.2

* 선행연구과 본 연구의 표본비율은 각각 2000년과 2005년 모집단 수치와 비교되었음.

2005년에 인천/경기 지역의 비율이 다소 증가한 것이 특징이고 성별로는 거의 차이가 없다. 그러나 연령대별로는 20대가 1.9%P 감소하고 30대가 2.3%P 감소한 반면 40대는 1.2%P 증가하고 50대 이상은 3.0%P 증가하였다. 학력별로는 중졸 이하와 고졸의 비율이 줄어든 반면 대재 이상은 크게 증가하였다. 직업별로 보면 가정주부가 감소하고 기타/무직이 증가하였다.

표 2.3은 선행연구와 본 연구(2절 참조)의 조사자료에서 표본과 모집단(2000년, 2005년)의 차이를 보여준다. 2005년도에 지역, 성별, 연령대, 학력 그리고 직업에서 조사표본과 모집단의 구성이 어떻게 다른지 살펴보면 다음과 같다.

지역별로는 서울(−2.8%P), 인천/경기(−2.1%P) 지역이 과소하게 표집되었고 가장 심하게는 서울에서는 15.4%P, 인천/경기에서는 10.6%P 과소 표집되었지만 이는 수도권에 크기에 비례하는 표본을 배정하게 되면 일부 지역의 표본이 지나치게 작아지게 되는 것을 막으려는 표본추출 설계상의 조치였던 것으로 보인다. 연령대로는 20대가 평균 5.3%P 과소하게 조사되었고 최대 12.3%P까지 과소하게 조사된 사례도 있었다. 학력별로는 대재 이상이 평균 8.7%P 과다 표집되었고 중졸 이하와 고졸은 각각

표 **2.4.** 지역, 성, 연령대 보정 후 학력과 직업에서의 조사표본과 모집단 차이(%P)*

선행연구				본 연구			
	평균	최소	최대		평균	최소	최대
학력				학력			
-중졸 이하	−10.4	−14.3	−6.4	-중졸 이하	−8.3	−12.1	−5.3
-고졸	1.1	−1.6	6.0	-고졸	−5.2	−8.8	−3.0
-대재 이상	9.3	3.4	14.9	-대재 이상	13.5	9.3	19.3
직업				직업			
-농/임/어업	−1.7	−6.1	1.7	-농/임/어업	−1.4	−2.8	0.7
-자영업/블루칼라	−2.7	−8.9	4.5	-자영업/블루칼라	−3.8	−6.9	0.3
-화이트칼라	−2.3	−7.1	5.8	-화이트칼라	−0.4	−6.4	4.6
-가정주부	8.7	1.8	13.0	-가정주부	9.4	4.2	17.5
-학생	0.6	−2.8	4.5	-학생	3.8	−2.4	6.0
-기타/무직	−2.6	−8.4	0.2	-기타/무직	−7.7	−9.9	−4.9

* 선행연구과 본 연구의 표본비율은 각각 2000년과 2005년 모집단 수치와 비교되었음.

5.7%P, 3.0%P 과소 표집되었다. 직업별로는 가정주부가 평균 11.3%P 과다 표집되었고, 화이트칼라, 자영업/블루칼라, 기타/무직이 과소 표집되었음을 볼 수 있다.

본 연구의 조사표본과 2005년의 모집단 간 5개 인구사회적 속성 각각의 구성비율의 차이를 선행연구의 조사표본과 2000년의 모집단 간 구성비율의 차이와 비교해 보면, 선행연구에 비해 본 연구에서 연령별 할당이 잘 되지 않았음을 알 수 있다. 그 사이에 젊은 층의 재택시간이 줄어든 데 원인이 있는 것으로 보인다. 학력별로 보면, 중졸 이하가 9.8%P 과소에서 5.7%P 과소로 바뀌어 다소 좋아졌지만 대재 이상은 8.2%P 과다가 8.7%P 과다로 바뀐 정도여서 별 차이가 없었다. 직업에서 주부는 선행연구에서 8.9%P 과다가 본 연구에서 11.3%P 과다로 바뀌고 기타/무직이 2.4%P 과소가 6.0%P 과소가 되는 등 오히려 조사표본의 대표성이 전반적으로 저하되었다.

지역·성별·연령대 등 3개 요인의 균형을 위한 반복비례가중법 보정을 하더라도 학력과 직업에서 조사표본이 왜곡되는가? 표 2.4에서 지역·성별·연령대를 보정한 후 남는 표본과 모집단 간 학력과 직업의 구성비율 차이를 살펴보면, 조사표본이 모집단에 비해 중졸 이하의 저학력에서 평균 8.3%P 과소하였고 대재 이상에서 평균 13.5%P 과다하였으며 직업에서는 가정주부가 평균 9.4%P 과다하였고 자영업/블루칼라가 평균 3.8%P 과소하였다. 따라서 지역·성별·연령대를 가중치로 보정한다고 하더라도 학력과 직업에서의 표본 왜곡은 여전하다고 할 수 있다.

결론적으로, 할당추출에 의한 전화조사 표본들이 다음 두 가지 측면에서 대표성에 문제가 있다고 말할 수 있다.

- 학력별로는 저학력(중졸 이하)이 과소하고 고학력(대재 이상)이 과다하다.
- 직업별로는 가정주부가 과다하고 자영업/블루칼라, 기타/무직이 과소하다.

3. 반복비례가중치에 의한 후보 지지율 보정

조사표본의 불균형을 바로 잡기 위해 가중치 보정을 하는 경우 각 후보의 지지율이 얼마나 어떻게 달라지는가를 보자. 표 3.1과 3.2는 18개 사례 중, 주요 정당후보가 확정된 이후 시행된 11개 사례(2007년 11월과 12월)의 조사표본에 여러 종류의 가중법을 적용하여 얻은 이명박 후보와 정동영 후보의 지지율 추정치이다. 적용된 가중법은

표 3.1. 이명박 후보의 지지율 추정치와 추정치 간 차이(%, %P)

	None	cell[3]	rim[3]	rim[5]	차이1	차이2	차이3	차이4	차이5
K1	35.4	39.1	38.9	37.6	3.7	−0.2	−1.3	−1.5	2.2
K2	30.8	33.0	33.1	32.7	2.2	0.1	−0.4	−0.3	1.9
K3	35.8	35.6	35.4	33.8	−0.2	−0.2	−0.6	−1.8	−2.0
M1	37.7	32.4	31.9	32.0	−5.3	−0.5	0.1	−0.4	−5.7
M2	32.3	33.2	31.9	30.2	0.9	−1.3	−1.7	−3.0	−2.1
M3	33.7	33.6	33.5	33.1	−0.1	−0.1	−0.4	−0.5	−0.6
T1	38.7	37.9	37.2	35.8	−0.8	−0.7	−1.4	−2.1	−2.9
G1	49.9	45.4	45.3	44.2	−4.5	−0.1	−1.1	−1.2	−5.7
G2	48.8	44.7	44.6	42.6	−4.1	−0.1	−2.0	−2.1	−6.2
G3	39.7	38.3	37.7	36.7	−1.4	−0.6	−1.0	−1.6	−3.0
G4	40.3	38.7	39.3	37.2	−1.6	0.6	−2.1	−1.5	−3.1
평균	38.5	37.4	37.2	36.0	−1.1	−0.2	−1.2	−1.4	−2.5
최소	30.8	32.4	31.9	30.2	−5.3	−1.3	−2.1	−3.0	−6.2
최대	49.9	45.4	45.3	44.2	3.7	0.6	0.1	−0.3	2.2

- None: 원자료, 즉 가중치 적용하지 않음(none)
- cell[3]: 지역·성별·연령대 셀 가중법(cell weighting)
- rim[3]: 지역·성별·연령대 반복비례가중법(rim weighting)
- rim[5]: 지역·성별·연령대·학력·직업 반복비례가중법(rim weighting)

이다. 여기에 적용된 두 가중법은 간단히 설명하면 (허명회 등, 2004), 1) 셀 가중법은 각 다차원 셀에 대하여 모집단 빈도 대 표본 빈도의 비를 가중치로 적용하여 표본 분포를 모집단 분포에 일치시키는 방법이다. 따라서 셀의 모집단 빈도를 알고 있어야 쓸 수 있다. 이에 따라 셀을 구성하는 변수가 많아지면 이 방법은 현실적으로 불가능해진다. 2) 반복비례 가중법은 1회 1변수의 표본 분포를 단순 가중법으로 모집단 분포에 맞추되 이런 가중치 과정을 여러 변수에 대하여 순환적으로 반복하는 방법이다. 셀을 구성하는 변수가 다소 많아지더라도 이 방법은 적용이 가능하다.

각종 가중법의 적용에 의한 추정치 사이의 차이를 다음과 같이 표기하기로 한다.

- 차이1 = cell[3] − None
- 차이2 = rim[3] − cell[3]
- 차이3 = rim[5] − rim[3]
- 차이4 = rim[5] − cell[3] (= 차이2 + 차이3)
- 차이5 = rim[5] − None (= 차이1 + 차이4)

표 3.1에서 이명박 후보의 지지율 추정치 간 차이를 보자. 차이1(= cell[3] − None)은 평균 −1.1%P로서 cell[3]을 적용하면 대체로 지지율 추정치가 작아지는 경향이 있다. 차이2(= rim[3] − cell[3])는 대체로 작게 나타나 기준변수가 같은 경우 칸 가중법과 림 가중법의 차이는 작은 것으로 보인다. 한편, 차이3(= rim[5] − rim[3])과 차이4(= rim[5] − cell[3])는 거의 음의 값을 취하는 것을 볼 수 있다. 이것은 통상적으로 조사기관들이 지역·성별·연령대 균형을 위해 가중법을 적용하지만 추가적 보정을 통해 학력과 직업에 추가적인 균형을 취하는 경우 이명박 지지율 추정치가 대체로 하락하게 됨을 말한다. 구체적으로, rim[5] 추정치가 rim[3] 추정치와 cell[3] 추정치에 비해 각각 1.2%P와 1.4%P 작게 나타났는데,

표 **3.2.** 정동영 후보의 지지율 추정치와 추정치 간 차이(%, %P)

	None	cell[3]	rim[3]	rim[5]	차이1	차이2	차이3	차이4	차이5
K1	23.3	20.0	20.2	20.1	−3.3	0.2	−0.1	0.1	−3.2
K2	17.1	16.3	15.1	15.5	−0.8	−1.2	0.4	−0.8	−1.6
K3	17.1	16.9	16.9	17.2	−0.2	0.0	0.6	0.6	0.4
M1	23.7	20.6	17.6	21.2	−3.1	−3.0	3.6	0.6	−2.5
M2	15.1	14.4	15.6	15.9	−0.7	1.2	0.3	1.5	0.8
M3	14.7	14.9	14.7	16.7	0.2	−0.2	2.0	1.8	2.0
T1	13.3	13.9	14.3	13.7	0.6	0.4	−0.6	−0.2	0.4
G1	16.2	17.5	17.9	18.2	1.3	0.4	0.3	0.7	2.0
G2	15.9	17.3	16.6	16.8	1.4	−0.7	0.2	−0.5	0.9
G3	14.1	14.4	14.7	14.9	0.3	0.3	0.2	0.5	0.8
G4	13.6	13.1	13.5	14.9	−0.5	0.4	1.4	1.8	1.3
평균	16.7	16.3	16.1	16.9	−0.4	−0.2	0.8	0.6	0.2
최소	13.3	13.1	13.5	13.7	−3.3	−3.0	−0.6	−0.8	−3.2
최대	23.7	20.6	20.2	21.2	1.4	1.2	3.6	1.8	2.0

이것은 가중법으로 지역·성별·연령대 균형을 취하더라도 학력과 직업 분포의 왜곡이 여전한 현재의 전화조사 표본들로부터 구한 이명박 지지율 추정치에는 그 정도 크기의 체계적 편향이 있음을 보여주는 것이다. 차이5(= rim[5] − None)는 평균이 −2.5%P나 되었다.

표 3.2는 각종 가중법 적용에 따른 정동영 후보의 지지율 추정치 간 차이를 보여준다. 이명박 지지율 추정치 차이에 비해 대체적으로 작지만, 차이3과 차이4의 값들이 양인 경우가 많고, 차이3(= rim[5] − rim[3])이 평균 0.8%P, 차이4(= rim[5] − cell[3])가 평균 0.6%P로 나타났다. 따라서 2007년의 우리나라 대통령선거관련 여론조사에서 이명박 후보와 정동영 후보 간 차이 추정치에 체계적 편향이

$$2.0\%\text{P} \quad (= 1.2\%\text{P} + 0.8\%\text{P}, \quad \text{or} \quad 1.4\%\text{P} + 0.6\%\text{P})$$

정도 크기로 존재하였다고 볼 수 있다.

4. 체계적 편향의 원인

앞 절에서 나타난 이명박 지지율 추정치가 갖는 체계적 편향의 원인을 알아보기 위해 이명박 지지를 종속 변수로 하고 지역, 성별, 연령대, 학력, 직업을 설명 변수로 하는 로지스틱 회귀모형을 적합해보기로 한다.

표 4.1은 11개 사례 중 첫째 것인 K1 표본자료에서의 로지스틱 회귀모형 추정 결과를, 표 4.2는 11개 사례 각각에 대한 로지스틱 모형 결과들을 종합해 보여준다.

로지스틱 회귀계수에 대한 해석을 쉽게 하기 위하여 일본의 수량화 방법 1처럼 원 계수를 요인별로 중심화하여 제시하였다. 즉 k개 범주를 갖는 한 요인에 대한 범주별 원 계수를 $b_1, \ldots, b_k\,(=0)$라고 하고 각 범주의 주변 빈도를 $f_1, \ldots, f_k$라고 할 때 중심은

$$\bar{b} = \sum_{j=1}^{k} b_j f_j \Big/ \sum_{j=1}^{k} f_j$$

가 되고 중심화 계수는 $b_1 - \bar{b}, \ldots, b_k - \bar{b}$로 산출된다. 계수 값들의 범위(range), 즉

$$\text{Imp} = \max_{j=1,\ldots,k} b_j - \min_{j=1,\ldots,k} b_j$$

표 **4.1.** K1 (2007년 12월 17일) 사례에 대한 로지스틱 모형 결과 *

변수	값	로지스틱 모형 중심화값	범위	표본** 비율(%)	모집단 비율(%)	차이(%P)
성별	남	0.06				
	여	−0.05	0.11			
나이	20대	−0.46				
	30대	−0.50				
	40대	−0.02				
	50대	0.31				
	60대+	0.58	1.08			
지역	서울	0.60				
	인천/경기	0.44				
	강원/제주	0.15				
	대전/충청	−0.21				
	광주/전라	−1.72				
	대구/경북	0.98				
	부산/울산/경남	0.44	2.70			
학력	중졸 이하	−0.38		19.0	26.3	−7.3
	고졸	0.05		31.7	34.7	−3.0
	대재 이상	0.19	0.57	49.3	39.0	10.3
직업	농/임/어업	−0.27		4.7	6.0	−1.3
	자영업/블루칼라	−0.03		25.2	27.8	−2.6
	화이트칼라	0.04		15.4	21.8	−6.4
	가정주부	0.07		32.8	19.1	13.7
	학생	0.01		11.6	7.4	4.2
	기타/무직	−0.05	0.34	10.3	17.9	−7.6

* 여기서 모집단 수치는 2005년도 것임.

** rim[3] 가중치로 지역·성별·연령대 보정한 이후의 표본 응답자의 학력 및 직업의 비율임.

를 각 요인의 중요도로 보고 이들을 살펴보면 2007년 대통령 선거에서 가장 중요도가 큰 변수는 지역이었고(Imp = 2.70) 다음이 연령이었다(Imp = 1.08). 학력과 직업의 중요도는 나이의 절반 수준이었으나(Imp = 0.57, 0.34) 그렇다고 무시할 수는 없다. 왜냐하면 표 4.1의 마지막 3개 열에 제시된 표본 구성비율과 모집단 구성비율 간 부호화 차이와 회귀계수 중심화 값의 패턴을 보면, 전자가 음이면 후자도 음이고 전자가 양이면 후자도 양인 경향이 있어 그 곱이 항상 양이기 때문이다. 이런 메커니즘은 K1 조사표본에서 이명박 지지율을 실제보다 높게 추정하는 원인이 되었다. 즉, 이명박 지지율이 상대적으로 낮은 중졸 이하 계층이 조사표본에 과소하게 대표됨으로써 추정치 하락을 막고 이명박 지지율이 상대적으로 높은 대재 이상 계층은 조사표본에 과다하게 대표됨으로써 추정치 상승을 유발한 것이다. 또한 직업에서 이명박 지지율이 높은 가정주부는 조사표본에 과다하게 대표되어 있고 이명박 지지율이 낮는 기타/무직 계층은 조사표본에 과소하게 대표되어 있어 이명박 지지율 추정치가 커지는 결과를 초래하였다. 이 사례에서 차이3(= rim[5] − rim[3])가 −1.3%P, 차이4(= rim[5] − cell[3])가 −1.5%P로 나왔는데, 그 원인은 조사표본에서 학력과 직업의 구성비율이 제대로 잡히지 않은 데 있었다고 하겠다.

K1 사례에 작동한 이명박 편향 메커니즘은 다른 사례에서도 거의 유사하게 반복적으로 나타나지만 여기서는 개별 사례에 대한 로지스틱 모형 결과 대신 11개 사례에 대한 결과의 평균을 간략히 제시하기로 한다. 표 4.2를 보면, 이명박 지지율이 상대적으로 낮은 중졸 이하 계층이 조사표본에 과소하게 대표됨으

표 **4.2.** 11개 사례 (2007년 11월–12월)에 대한 로지스틱 모형 결과 종합 (평균)*

변수	값	로지스틱 모형 중심화값	범위	표본** 비율(%)	모집단 비율(%)	차이(%P)
성별	남	0.06				
	여	−0.06	0.12			
나이	20대	−0.53				
	30대	−0.47				
	40대	−0.05				
	50대	0.43				
	60대+	0.48	1.01			
지역	서울	0.33				
	인천/경기	0.23				
	강원/제주	−0.02				
	대전/충청	−0.27				
	광주/전라	−1.59				
	대구/경북	0.63				
	부산/울산/경남	0.27	2.22			
학력	중졸 이하	−0.40		18.2	26.3	−8.1
	고졸	0.05		31.8	34.7	−2.9
	대재 이상	0.17	0.57	50.0	39.0	11.0
직업	농/임/어업	−0.21		4.2	6.0	−1.8
	자영업/블루칼라	−0.04		23.0	27.8	−4.8
	화이트칼라	−0.11		20.0	21.8	−1.8
	가정주부	0.12		31.0	19.1	11.9
	학생	0.02		10.5	7.4	3.1
	기타/무직	0.04	0.33	11.2	17.9	−6.7

* 여기서 모집단 수치는 2005년도 것임.

* G사 사례들에 대하여는 50대와 60대에 동일 수치를 넣어 중심화 값을 산출하였음.

** rim[3] 가중치로 지역·성별·연령대 보정한 이후의 표본 응답자의 학력 및 직업의 비율임.

로써 추정치 하락을 막고 이명박 지지율이 상대적으로 높은 대재 이상 계층은 조사표본에 과다하게 대표됨으로써 추정치 상승을 유도한 것이다. 직업에서는 이명박 지지율이 높은 가정주부가 조사표본에 과다하게 대표됨으로써 이명박 지지율 추정치를 실제보다 커지게 한다. 기타/무직은 중심화 값의 부호와 표본구성비율과 모집단 구성비율 차이의 부호가 어긋나 추정 편향의 감소에 기여하는 것으로 나왔지만 나머지 5개 직업범주(농/임/어업, 자영영/블루칼라, 화이트칼라, 가정주부, 학생)에서는 모두 중심화 값의 부호와 표본 구성비율과 모집단 구성비율 차이의 부호가 일치하여 이명박 과다 추정을 유발하였다.

5. 맺음 말

2007년 대통령선거관련 여론조사에서 조사기관들이 발표한 이명박 후보과 정동영 후보의 지지율 차이에 2.0%P 크기의 체계적인 편향이 있었던 것으로 보인다. 이는 지역·성별·연령대를 보정 한 후에도 여전히 존재하는 학력과 직업의 왜곡이 이명박 후보의 지지율 추정에 일정한 방향의 영향을 주었기 때문이다. 허명회 등 (2004)의 연구에서 전화조사의 문제점으로 지적 되었던 표본 왜곡이 해결 되지 않고 여전히 문제점으로 남아 있다.

참고문헌

강현철, 한상태, 김지연, 정용찬, 허명회 (2008). RDD 전화조사와 주요 결과, <조사연구>, **9**, 1–22.

통계청 (2000). <인구주택총조사>, 통계청.

통계청 (2005). <인구주택총조사>, 통계청.

허명회, 윤영아, 김규성 (2005). 2차원 기준 반복비례법 연구, <통계연구>, **10**, 1–22.

허명회, 윤영아, 이용구 (2004). 사회조사에서 표본의 왜곡과 가중치 보정의 결과: 18개 사례연구, <조사연구>, **5**, 31–47.

허명회, 황진모 (2006). 전화조사를 위한 시간균형할당표본추출, <조사연구>, **7**, 39–52.

Systematic Bias of Telephone Surveys: Meta Analysis of 2007 Presidential Election Polls

SeYong Kim[1] · Myung-Hoe Huh[2]

[1]Department of Statistics, Korea University; [2]Department of Statistics, Korea University

(Received November 2008; accepted December 2008)

Abstract

For 2007 Korea presidential election, most polls by telephone surveys indicated Lee Myung-Bak led the second runner-up Jung Dong-Young by certain margin. The margin between two candidates can be estimated accurately by averaging individual poll results, provided there exists no systematic bias in telephone surveys.

Most Korean telephone surveys via telephone directory are based on quota samples, with the region, the gender and the age-band as quota variables. Thus the surveys may result in certain systematic bias due to unbalanced factors inherent in quota sampling. The aim of this study is to answer the following questions by the analytic methods adopted in Huh *et al.* (2004):

Question 1. Wasn't there systematic bias in estimates of support rates.

Question 2. If yes, what was the source of the bias?

To answer the questions, we collected eighteen surveys administered during the election campaign period and applied the iterated proportional weighting (the rim weighting) to the last eleven surveys to obtain the balance in five factors - region, gender, age, occupation and education level. We found that the support rate of Lee Myung-Bak was over-estimated consistently by 1.4%P and that of Jung Dong-Young was under-estimated by 0.6%P, resulting in the over-estimation of the margin by 2.0%P. By investigating the Lee Myung-Bak bias with logistic regression models, we conclude that it originated from the under-representation of less educated class and/or the over-representation of house wives in telephone samples.

Keywords: Telephone survey, presidential election, meta-analysis, systematic bias, rim weighting, logistic regression.

[2]Corresponding author: Professor, Department of Statistics, Korea University, Seoul 136-701, Korea. E-mail: stat420@korea.ac.kr