응용통계연구 (2008)
21(5), 845-854

선거 여론조사 자료의 표준적 요약과 시각화

허명회[1] · 이용구[2]

[1]고려대학교 통계학과; [2]중앙대학교 통계학과

(2008년 6월 접수, 2008년 7월 채택)

요 약

선거관련 여론조사에 대한 대부분의 요약 보고서는 지역, 성, 연령대 등 인구지리적 요인과 교육수준, 직업, 소득 등의 사회적 요인 별 후보지지율 교차표로 되어 있다. 그러나 인구지리적 요인과 사회적 요인들이 상호 연관되어 있어 각 요인별 지지율 분포에 그 외 요인들의 효과가 겹쳐 반영되므로 조사 자료의 해석에 있어 중복성의 문제가 야기된다. 본 사례 연구의 대강은 선거 여론조사의 결과보고에 있어 사회적 요인 수준별 지지율 분포를 인구지리적 요인으로부터 분리하여 추정할 것과 다수의 교차표를 동시에 시각화하는 배중 대응분석의 활용을 제안하는 내용이다.

주요용어: 선거여론조사, 직접 표준화, 대응분석, 배중기법(DOUBLING).

1. 연구 배경과 목적

2007년 12월의 제 17대 대통령 선거에 앞서 수많은 여론조사가 행하여졌다. 조사 하나 하나씩에 들어가는 비용이 만만치 않으므로 전체로는 수십억 원이 소요되었을 것이다. 그럼에도 불구하고 조사결과 보고서들은 인구지리·사회적 요인의 수준별 후보지지율을 보여주는 단순한 교차표로 가득 차 있을 뿐이다. 표 1.1, 1.2, 1.3을 보자. KBS · MBS 컨소시엄을 위해 한 조사기관이 2007년 12월 전화면접으로 시행한 조사에서 나온 것으로 권역별로 후보지지율을 보여주는 단순 교차표이다. 이 사례에서는, 권역에 따라 후보들의 지지율이 심한 편차를 보이고 있고 각 후보의 지지율이 성별로는 크게 차이가 나지 않지만 성별에 따라 무응답 비율에서 현격한 차이를 보인다. 연령대별로도 후보지지율이 상당히 달리 나타나고 있다. 예컨대 이명박은 나이가 많은 유권자에게 선호되는 경향이 있으나 이회창과 문국현은 그 반대이다. 이상과 같이 3개 교차표에서 인구지리적 요인과 지지후보 간 연관 패턴을 대략 파악해볼 수 있다.

우리나라에서 권역, 성, 연령대 간 상호 연관성은 약하기 때문에 후보지지율에 대한 3개 인구지리적 요인의 영향을 표 1.1, 1.2, 1.3과 같은 단순 교차표들에서 살펴본다고 해도 별 문제가 없다. 그러나 교육, 직업, 소득으로 나누어 후보지지율 분포를 각기 살펴볼 때는 중복성의 문제를 무시할 수 없다. 왜냐하면 교육과 나이, 직업과 성, 직업과 나이, 소득과 나이 등은 밀접한 연관 관계에 있기 때문이다 (부록 참조). 그러므로 교육수준, 직업, 소득에 따른 후보지지율을 각각의 단순 교차표로부터 구하는 현재 조사업계의 관행으로는 지지후보의 선택에 미치는 사회적 요인의 고유한 영향이 포착되지 않는다.

우리는 이 문제를 해결하는 방법으로 사회적 요인별로 인구지리적 요인의 효과를 보정한 직접표준화(direct standardization) 지지율을 제시하고자 한다. 아울러 배중(double) 대응분석(correspondence analysis)으로 다수의 교차표를 시각화해냄으로써 정치후보들의 포지션이 매핑된 지각도(perceptual map)의 활용을 제안하고자 한다. 이로써 조사결과 보고서가 보다 풍부한 정보를 담게 되길 기대한다.

[1]교신저자: (136-701) 서울시 성북구 안암동 5-1, 고려대학교 통계학과, 교수. E-mail: stat420@korea.ac.kr
[2](156-756) 서울시 동작구 흑석동 221, 중앙대학교 통계학과, 교수. E-mail: leeyg@cau.ac.kr

표 1.1. 제 17대 대통령 선거 여론조사 결과: 권역별 후보지지율

권역	표본크기	정동영	이명박	이회창	문국현	기타	무응답	합계
서울	328	12.8	51.8	10.7	6.7	2.4	15.5	100.0
부산경남	241	8.7	44.8	13.7	6.2	4.6	22.0	100.0
대구경북	167	7.8	56.3	18.0	3.6	1.8	12.6	100.0
인천경기	411	13.9	44.3	11.9	8.3	6.1	15.6	100.0
광주전라	161	44.7	11.2	3.7	11.8	7.5	21.1	100.0
대전충청	153	12.4	34.6	20.3	7.2	5.2	20.3	100.0
강원제주	61	11.5	27.9	13.1	4.9	6.6	36.1	100.0
전체	1522	15.2	42.2	12.6	7.2	4.7	18.1	100.0

*자료출처: KBS·MBS의 컨소시엄을 위하여 A 기관이 2007년 12월에 전화면접으로 시행하였다.
*기타후보: 권영길, 이인제, 정근모, 허경영, 전관, 금민 등.
*$\chi^2 = 228.4$, 자유도 $= 30$, p-값 < 0.001.

표 1.2. 제 17대 대통령 선거 여론조사 결과: 성별 후보지지율

성	표본크기	정동영	이명박	이회창	문국현	기타	무응답	합계
남성	746	15.8	44.1	13.5	7.6	5.2	13.7	100.0
여성	776	14.6	40.3	11.7	6.8	4.1	22.4	100.0
전체	1522	15.2	42.2	12.6	7.2	4.7	18.1	100.0

*$\chi^2 = 20.1$, 자유도 $= 5$, p-값 < 0.001.

표 1.3. 제 17대 대통령 선거 여론조사 결과: 연령대별 후보지지율

연령대	표본크기	정동영	이명박	이회창	문국현	기타	무응답	합계
20대−	310	12.9	35.2	17.7	15.2	3.9	15.2	100.0
30대	358	16.8	37.7	12.8	8.9	6.4	17.3	100.0
40대	347	19.6	42.7	12.1	5.8	4.9	15.0	100.0
50대	234	12.4	49.1	11.5	3.0	5.1	18.8	100.0
60대+	273	12.5	49.5	8.1	1.5	2.6	26.0	100.0
전체	1522	15.2	42.2	12.6	7.2	4.7	18.1	100.0

*$\chi^2 = 97.6$, 자유도 $= 20$, p-값 < 0.001.

2. 직접 표준화에 의한 나이 효과 보정

우리나라에서 연령대에 따라 교육수준이 크게 다르다. 이에 따라, 각 교육수준(중졸 이하, 고졸, 대학 이상) 집단에서 연령대 구성비가 다르게 된다. 표 2.1은 교육수준별로 연령대 구성비를 보여주는데 '중졸 이하' 학력자 집단에서는 60대 이상이 56.9%를 차지하지만 '대학 이상' 학력자 집단에서 60대 이상은 5.4%에 불과하다. 이에 따라 표 2.4의 학력수준별 후보지지율에는 학력수준에 의한 성향적 차이 뿐 아니라 나이 효과가 혼재되어 나타난다. 표 2.2와 2.3에서 볼 수 있듯이 직업과 소득도 나이와 연관이 있고 이에 따라 표 2.5와 2.6의 직업 및 소득 범주별 후보 지지율에 나이 효과가 개입될 수 밖에 없다. 직접 표준화(direct standardization)는 인구학 및 역학에서 오랫동안 활용되어 왔던 제 3 요인의 효과를 보정하는 방법이다. (이 사례에서 다음과 같이 무응답 대체를 하였다. 1) 학력수준 질문이 저학력 응답자에게 심리적 부담을 준다고 보고 학력수준에 대한 무응답은 '중졸 이하'로 대체하였다(총 31명, 2.0%). 2) 소득을 묻는 질문도 일부 계층 응답자에는 상당한 심리적 부담을 줄 것으로 생각된다. 무소득이면 100만원 이하에 해당되지만 전화조사 응답자들에게 그렇게 논리적인 응답을 기대하기 어렵다. 소득 무응답자 236명(전체의 15.5%)을 연령대로 구분해 보면 29세 이하가 89명(38%)으로 가장 많은데

표 **2.1.** 학력수준별 연령대 구성비율

학력수준	표본크기	20대−	30대	40대	50대	60대+	합계
중졸이하	313	1.6	2.6	12.8	26.2	56.9	100.0
고졸	487	11.9	28.1	28.3	20.1	11.5	100.0
대학이상	722	34.2	29.5	23.4	7.5	5.4	100.0
전체	1522	20.4	23.5	22.8	15.4	17.9	100.0

표 **2.2.** 직업별 연령대 구성비율

직업	표본크기	20대−	30대	40대	50대	60대+	합계
화이트	236	30.1	33.9	24.2	8.5	3.4	100.0
블루	115	16.5	32.2	27.0	17.4	7.0	100.0
자영업	273	6.6	24.5	38.1	23.8	7.0	100.0
농림어업	62	0.0	6.5	12.9	21.0	59.7	100.0
주부	566	8.8	28.1	24.2	16.8	22.1	100.0
학생	136	97.8	2.2	0.0	0.0	0.0	100.0
무직기타	134	14.2	6.0	7.5	15.7	56.7	100.0
전체	1522	20.4	23.5	22.8	15.4	17.9	100.0

표 **2.3.** 소득별 연령대 구성비율

소득	표본크기	20대−	30대	40대	50대	60대+	합계
100−	441	21.3	9.1	10.2	17.5	42.0	100.0
100+	264	16.7	22.3	21.2	19.3	20.5	100.0
200+	322	20.8	33.2	28.3	13.7	4.0	100.0
300+	206	19.4	33.0	35.0	9.2	3.4	100.0
400+	136	22.8	30.9	30.1	12.5	3.7	100.0
500+	153	22.2	27.5	27.5	17.0	5.9	100.0
전체	1522	20.4	23.5	22.8	15.4	17.9	100.0

이 층에서는 비경제 활동자가 많으므로 실제 소득이 100만원 이하일 것으로 볼 수 있다. 이에 따라 소득에 대한 무응답을 '100만원 이하'로 대체하였다.)

이해를 돕기 위하여 간단한 수치 예를 들어보도록 하겠다. 직접 표준화에 의한 학력수준별 나이 효과 보정 이명박 지지율은 다음과 같이 산출된다.

단계 1: A에는 전체집단에서 나이대별 구성비율이 입력된다. B1, B2, B3에는 각 학력수준과 연령대별 조합에서의 이명박 지지율이 입력된다.

	구성비율	중졸이하 지지율(%)		고졸 지지율(%)		대학이상 지지율(%)	
	A	B1	A*B1	B2	A*B2	B3	A*B3
20대−	0.204	0.0	0.00	27.6	5.63	37.7	7.69
30대	0.235	37.5	8.81	36.5	8.58	38.5	9.05
40대	0.228	27.5	6.27	45.7	10.42	43.8	9.99
50대	0.154	40.2	6.19	44.9	6.91	70.4	10.84
60대+	0.179	38.8	6.95	71.4	12.78	66.7	11.94
합계			28.22		44.32		49.51

표 **2.4.** 학력수준별 후보 지지율: 원자료

학력수준	표본크기	정동영	이명박	이회창	문국현	기타	무응답	합계
중졸−	313	15.7	37.1	7.0	2.6	5.1	32.6	100.0
고졸	487	16.8	43.7	13.6	4.9	5.3	15.6	100.0
대학+	722	13.9	43.4	14.4	10.8	4.0	13.6	100.0
전체	1522	15.2	42.2	12.6	7.2	4.7	18.1	100.0

$*\chi^2 = 87.2$, 자유도 $= 10$, p-값 < 0.001.

표 **2.5.** 직업별 후보 지지율: 원자료

직업	표본크기	정동영	이명박	이회창	문국현	기타	무응답	합계
화이트	236	15.3	44.5	13.6	8.5	5.1	13.1	100.0
블루	115	17.4	36.5	18.3	5.2	7.0	15.7	100.0
자영업	273	18.3	44.7	12.5	7.7	3.7	13.2	100.0
농림어업	62	29.0	32.3	9.7	1.6	6.5	21.0	100.0
주부	566	14.5	40.8	11.0	5.7	3.7	24.4	100.0
학생	136	11.0	39.7	16.9	16.9	5.1	10.3	100.0
무직기타	134	7.5	50.7	10.4	5.2	6.7	19.4	100.0
전체	1522	15.2	42.2	12.6	7.2	4.7	18.1	100.0

$*\chi^2 = 83.6$, 자유도 $= 30$, p-값 < 0.001.

표 **2.6.** 소득별 후보 지지율: 원자료

소득	표본크기	정동영	이명박	이회창	문국현	기타	무응답	합계
100−	441	14.5	35.8	10.4	2.9	4.3	32.0	100.0
100+	264	17.8	41.7	11.7	6.1	6.1	16.7	100.0
200+	322	15.2	44.4	15.5	9.0	5.3	10.6	100.0
300+	206	17.0	43.7	15.5	7.8	3.4	12.6	100.0
400+	136	14.0	48.5	8.8	14.7	5.9	8.1	100.0
500+	153	11.1	49.0	13.7	10.5	2.6	13.1	100.0
전체	1522	15.2	42.2	12.6	7.2	4.7	18.1	100.0

$*\chi^2 = 118.5$, 자유도 $= 25$, p-값 < 0.001.

단계 2: A와 B1을 곱하고 최종 합을 구한다. 이것이 이명박에 대한 중졸 이하 집단의 직접 표준화 지지율 28.2%이다. 마찬가지 방식으로 고졸 집단과 대학 이상 집단의 표준화 지지율로 44.3%와 49.5%가 얻어진다. 결과적으로 표준화 비율이 단순 비율에 비하여 학력수준 집단별로 −8.9%P, 0.6%P, 6.1%P 차이가 난다.

표 2.7은 직접 표준화에 의한 학력수준별 나이효과 보정 후보지지율을 보여준다. 표 2.4의 단순 지지율과 비교하여 볼 때 나이에 의한 표준화 보정을 함으로써 중졸 이하 집단에서는 이명박에 대한 지지율이 내려가고 대학 이상 집단에서는 지지율이 높아짐으로써 최대값과 최소값 차이가 커짐을 볼 수 있다. 표 2.8과 2.9는 이와 같은 직접 표준화로 나이 효과를 보정한 직업 및 소득 범주별 후보지지율을 보여준다.

3. 직접 표준화에 의한 인구지리적 요인 효과 보정

앞 절에서는 직접 표준화 기법으로 나이 효과를 보정한 바 있다. 더 나아가, 지지후보 선택에 있어 나이뿐만 아니라 권역과 성의 효과까지 보정함으로써 교육수준·직업·소득 등 사회적 요인의 고유한 효과를 살펴보기로 하자. 이를 위하여 우리는 다음과 같이 직접 표준화 기법을 확장하여 적용하고자 한다.

표 **2.7.** 학력수준별 직접표준화 지지율: 나이 효과 보정

학력수준	표본크기	정동영	이명박	이회창	문국현	기타	무응답	합계
중졸−	313	18.0	28.2	8.3	2.5	6.4	36.6	100.0
고졸	487	16.2	44.3	13.2	5.6	4.7	16.0	100.0
대학+	722	13.3	49.5	13.5	7.8	3.1	12.8	100.0
전체	1522	15.2	43.5	12.4	6.0	4.3	18.7	100.0

표 **2.8.** 직업별 직접표준화 지지율: 나이 효과 보정

직업	표본크기	정동영	이명박	이회창	문국현	기타	무응답	합계
화이트	236	14.0	52.5	12.1	5.9	4.3	11.2	100.0
블루	115	16.5	34.9	19.6	4.9	5.9	18.3	100.0
자영업	273	17.9	46.1	12.0	8.2	2.5	13.3	100.0
농림어업	62	28.7	31.4	6.5	5.9	11.2	16.2	100.0
주부	566	14.3	39.2	11.9	7.0	3.2	24.3	100.0
학생	136	18.6	42.1	9.5	5.6	3.4	20.8	100.0
무직기타	134	7.6	44.0	9.2	6.5	9.6	23.2	100.0
전체	1522	15.4	42.6	11.9	6.7	4.4	19.1	100.0

표 **2.9.** 소득별 직접표준화 지지율: 나이 효과 보정

소득	표본크기	정동영	이명박	이회창	문국현	기타	무응답	합계
100−	441	15.4	32.6	11.2	3.4	5.0	32.4	100.0
100+	264	18.5	40.0	12.0	6.5	5.7	17.3	100.0
200+	322	13.5	51.2	14.1	8.0	4.4	8.8	100.0
300+	206	18.8	47.8	12.4	6.8	2.7	11.6	100.0
400+	136	13.7	51.3	11.0	12.2	4.8	7.0	100.0
500+	153	11.0	51.4	13.4	10.6	2.2	11.4	100.0
전체	1522	15.4	43.5	12.3	6.9	4.4	17.6	100.0

부집단 l에서의 후보 C_g에 대한 표준화 지지율

$$= \sum_{i=1}^{7}\sum_{j=1}^{2}\sum_{k=1}^{5} m_{ijk}p_{ijkl}(C_g), \quad l = 1,\ldots,L;\ g = 1,\ldots,6, \tag{3.1}$$

여기서 m_{ijk}은 (권역: i, 성: j, 연령대: k) 칸의 구성비율이고 $p_{ijkl}(C_g)$는 (i, j, k, l) 칸에서 후보 C_g에 대한 지지율이다(후보 수 $= 6$). 그리고

$$\sum_{i=1}^{7}\sum_{j=1}^{2}\sum_{k=1}^{5} m_{ijk} = 1, \qquad \sum_{g=1}^{6} m_{ijk}p_{ijkl}(C_g) = 100(\%), \quad \text{각 } i, j, k, l\text{에 대하여.}$$

식 (3.1)은 2절에서 사용된바 있는 직접 표준화 산식의 확장이다. 그러나 문제의 소지가 있는데 조사 자료에 (i, j, k, l) 칸에 해당하는 케이스가 없는 경우이다. 예컨대 이 사례에서는 (권역, 성, 연령대, 교육수준)의 총 조합 칸의 수는 210개($= 7 * 2 * 5 * 3$)나 되므로 비는 칸이 있기 십상이다. 이런 경우에는 다음과 같은 결측 값 대체를 제안한다.

$$p_{ijkl}(C_g) \quad \leftarrow \quad \hat{p}_{ijkl}(C_g), \tag{3.2}$$

표 3.1. 학력수준별 직접표준화 지지율: 권역, 성, 나이 효과 보정

학력수준	표본크기	정동영	이명박	이회창	문국현	기타	무응답	합계
중졸−	313	18.4	31.5	8.1	4.5	6.2	31.3	100.0
고졸	487	16.7	42.4	14.3	5.9	5.0	15.8	100.0
대학+	722	13.1	47.6	14.5	7.9	3.2	13.7	100.0
전체	1522	14.9	42.9	13.2	6.6	4.4	17.9	100.0

*결측값 대체 빈도: 124례(8.1%).

표 3.2. 직업별 직접표준화 지지율: 권역, 성, 나이 효과 보정

직업	표본크기	정동영	이명박	이회창	문국현	기타	무응답	합계
화이트	236	15.7	44.3	13.5	5.8	5.6	15.1	100.0
블루	115	16.8	37.6	13.8	4.1	7.0	20.8	100.0
자영업	273	17.7	43.3	12.7	11.6	2.3	12.4	100.0
농림어업	62	27.6	30.4	9.3	3.9	10.1	18.7	100.0
주부	566	13.3	44.0	12.6	5.8	2.8	21.5	100.0
학생	136	10.9	49.1	11.5	8.2	8.0	12.4	100.0
무직기타	134	9.9	37.9	8.1	10.3	12.1	21.6	100.0
전체	1522	15.0	42.9	12.1	7.4	4.7	17.9	100.0

*결측값 대체 빈도: 594례(39.0%).

표 3.3. 소득별 직접표준화 지지율: 권역, 성, 나이 효과 보정

소득	표본크기	정동영	이명박	이회창	문국현	기타	무응답	합계
100−	441	15.3	32.4	11.9	3.2	4.5	32.7	100.0
100+	264	19.7	39.0	13.4	5.9	6.5	15.4	100.0
200+	322	14.1	50.0	15.0	7.4	4.2	9.3	100.0
300+	206	20.5	43.6	12.9	6.8	3.3	12.8	100.0
400+	136	12.5	51.0	9.3	13.3	4.3	9.6	100.0
500+	153	12.8	50.3	13.9	9.3	1.9	11.8	100.0
전체	1522	16.0	41.4	12.8	6.4	4.4	19.0	100.0

*결측값 대체 빈도: 102례(6.7%).

여기서 $\hat{p}_{ijkl}(C_g)$는 (i, j, k, l) 칸에서 후보 C_g에 대한 모형적합 지지율이다. 이 사례에서는 다항로짓 모형(multinomial logit model)을 적용할 것이다.

표 3.1, 3.2, 3.3은 이와 같은 직접 표준화로 권역, 성, 나이 효과를 보정한 학력수준, 직업, 소득 범주별 지지율을 보여준다. 이제 이명박 지지율이 1) 학력수준에 따라 높아지고, 2) 직업으로는 블루칼라 · 농림어업 · 무직기타에서 낮으며, 3) 소득 수준에 따라 높아지는 경향이 확인된다. 즉, 이명박은 사회적 상층이 지지하는 후보인 것이다. 그러나 원 자료에서 나온 단순 지지율로는 이런 점이 명확히 드러나지 않는다. 즉 표 2.4, 2.5, 2.6에서는 이명박 지지율이 1) 고졸 집단과 대학 이상 집단에서 비슷하였고, 2) 직업으로는 무직기타에서 가장 높게 나타났었다. 다만, 소득 수준에 따라 지지율이 높아지는 경향은 마찬가지였다.

4. 배중 대응분석에 의한 시각화

이제 표 1.1, 1.2, 1.3의 교차표와 표 3.1, 3.2, 3.3의 교차표에 정리된 지지율 프로파일을 시각화하여 보자. 길이가 q인 비율 프로파일들은 합이 1로 고정값을 가지므로 $q-1$ 차원 심플렉스 초평면에 위치하

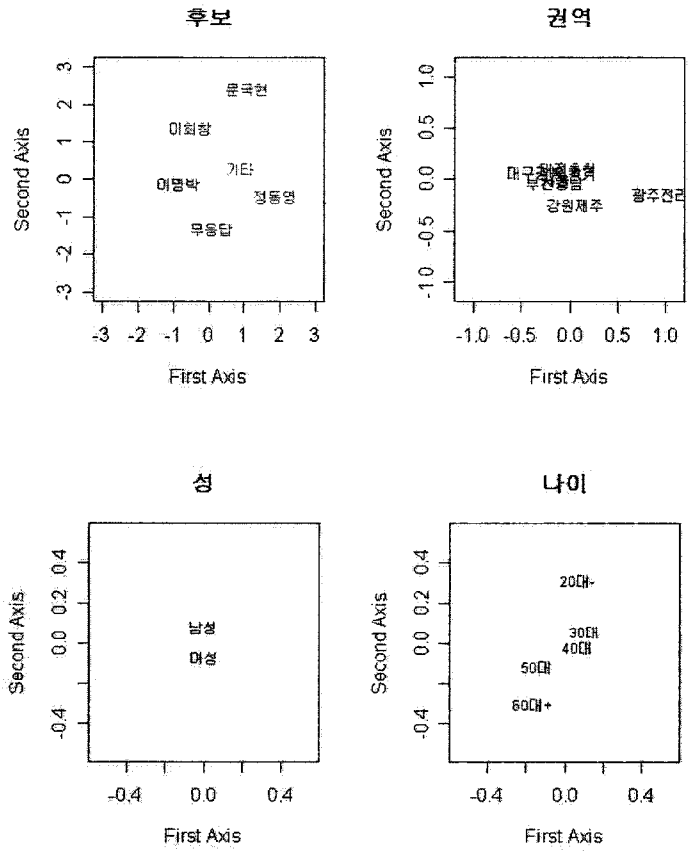

그림 **4.1.** 인구지리적 요인에 따른 지지후보 양상: 배중 대응분석(근사도 80.0%)

게 된다. 대응분석(correspondence analysis)은 이러한 심플렉스 상에서의 주성분분석으로 볼 수 있는데 이 때 각 점에는 표본크기에 비례하는 가중치가 주어지고 카이제곱으로 두 프로파일 간 거리를 정의한다 (Greenacre와 Hastie, 1987).

통상적인 대응분석은 단일 이차원 교차표의 행 비율 프로파일을 시각화하지만 여기서는 표 1.1, 1.2, 1.3의 14개($= 7 + 2 + 5$) 인구지리적 범주별 프로파일을 동시에 고려한 시각화를 해보기로 하겠다. 즉 3개의 교차표를 위·아래로 붙이는 배증기법(doubling)을 확대하여 대응분석해보기로 한다 (Greenacre, 1993).

그림 4.1은 표 1.1, 1.2, 1.3에 대한 대응분석 결과이다. 1) 정동영 후보가 지리적으로 광주전라에 대응하고, 2) 여성이 지지후보 무응답과 대응하며, 3) 문국현-이회창-기타-이명박-정동영-무응답의 순서로 나이와 대응함을 볼 수 있다 (연소에서 연장의 방향으로).

그림 4.1이 인구지리적 요인에 따른 지지후보 양상을 보여준다면, 표 3.1, 3.2, 3.3의 표준화 지지율을 배증 대응분석으로 얻어낸 그림 4.2는 사회적 요인에 따른 지지후보 양상을 보여준다. 제 1축은 우에서 좌로 학력수준으로는 '중졸 이하'-'고졸'-'대학 이상'이 자리 잡고 있고 소득으로는 '100-'부터 '400+'와 '500+'까지 나타나므로 사회적 계층 순서를 나타낸다. 사회적 계층의 높은 곳에 이명박과 문국현·이회

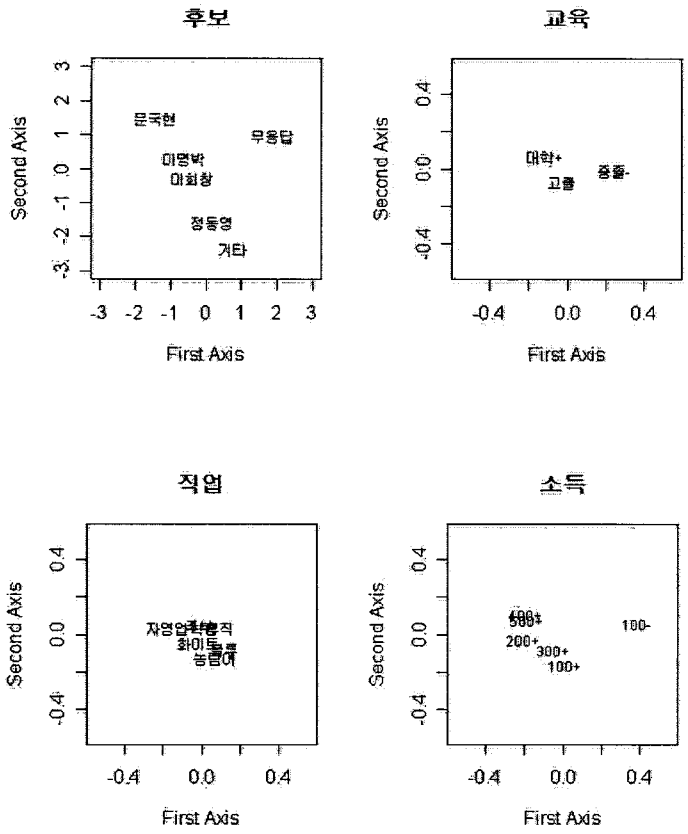

그림 **4.2.** 사회적 요인에 따른 지지후보 양상: 배중 대응분석(근사도 82.1%)

창이 자리 잡고 있다. 한편, 제 2축은 기타후보(권영길, 이인제, 정근모, 허경영, 전관, 금민) 지지와 관계있는데 교육수준과 소득과는 관계가 적으나 직업적으로 기타무직·농림어업과 연관이 있는 것으로 보여진다. 이명박과 이회창이 사회적 지각도에서 같은 포지션에 있으므로 경합 관계임을 확인할 수 있다.

5. 맺음말

표 2.7, 2.8, 2.9와 3.1, 3.2, 3.3의 직접 표준화 지지율이 어느 정도의 표준오차를 갖는지, 4절의 식 (3.2) 외에 결측 값 대체에 더 좋은 방법이 있는지에 대한 후속 연구가 필요하다. 향후 조사결과 보고서가 다양한 시각적 정보를 담게 되길 기대한다.

부록: 인구지리적 요인과 사회적 요인 간 연관성

범주형 변수 간 연관성 측도에는 여러 가지가 있으나 본 연구에서는 두 변수의 수량화 변환 간 상관계수를 구하여 활용하였다. 두 범주형 변수를 행과 열에 넣어 교차표 F를 구성하여 F의 행 주변과 열 주변을 각각 r과 c라고 할 때, 1이 아닌 $D_r^{-1/2}FD_c^{-1/2}$의 가장 큰 비정칙값(singular value)이 행 수량화 변

환과 열 수량화 변환 간 상관계수와 일치한다 (허명회, 1999). 즉, F에 대한 대응분석으로 행 요인과 열 요인 간 연관성을 구하였는데, 이렇게 얻은 권역, 성, 나이, 교육수준, 직업, 소득 간 연관성 행렬은 다음과 같다.

	권역	성	나이	교육	직업	소득
권역	1	0.014	0.082	0.191	0.221	0.202
성		1	0.059	0.146	0.759	0.080
나이			1	0.610	0.641	0.479
교육				1	0.479	0.503
직업					1	0.384
소득						1

이로부터 알 수 있는 사실은 1) 권역, 성, 나이 간 연관성은 뚜렷하지 않으나, 권역은 직업 및 소득과 약하게, 성은 직업과 강하게 결합되어 있다는 것, 2) 나이는 교육, 직업, 소득 등 사회적 요인들과 강하게 연관되어 있다는 것, 3) 교육과 직업, 교육과 소득 간 연관성도 강한 편이나 직업과 소득 간 연관성은 그보다 약한 편이라는 것이다.

참고문헌

허명회 (1999). <다변량 수량화>, 자유아카데미.

Greenacre, M. (1993). *Correspondence Analysis in Practice,* Academic Press, London.

Greenacre, M. and Hastie, T. (1987). The geometric interpretation of correspondence analysis, *Journal of the American Statistical Association,* **82**, 437-447.

Standardizing and Visualizing Descriptive Summaries of Election Survey Data

Myung-Hoe Huh[1] · Yonggoo Lee[2]

[1]Dept. of Statistics, Korea University; [2]Dept. of Statistics, Chung Ang University

(Received June 2008; accepted July 2008)

Abstract

Survey reports of election opinions consist of numerous cross-tabulations between socio-demographic variables and political opinions including preferred candidates. Since socio-demographic variables are related each other, duplicate interpretations arise. The aim of this study is twofold: The first is to separate the effects of socio variables such as education, occupation and income from the effects of demographic variables such as region, sex and age. The second is the visualization of multiple cross-tabulations in low-dimensional space by extended doubling technique of correspondence analysis. Survey researchers may get some help from this study to present their survey results more lucidly and visually.

Keywords: Opinion survey on election, direct standardization, correspondence analysis, doubling.

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

[2]Professor, Dept. of Statistics, Chung Ang University, Seoul 156-756, Korea. E-mail: leeyg@cau.ac.kr