한국통계학회논문집
2010, 17권, 1호, 55-66

인터넷 선거여론조사 가중치보정을 위한 성향점수의 활용

김영원[1,a], 배예영[a]

[a]숙명여자대학교 수학통계학부

요 약

본 연구에서는 2007년에 실시한 17대 대통령 선거를 위한 NI Korea의 인터넷 패널조사와 KBS의 대선 패널 전화여론조사 결과를 토대로 인터넷조사와 전화조사의 차이를 비교하고, 인터넷조사의 활용 가능성을 검토해 보고자 한다. 인터넷조사는 조사대상자가 인터넷 사용자로 제한됨에 따라 발생하는 포함오차와 조사 참여 의사를 갖는 사람들만을 조사에 참여시킴으로써 발생하는 선택편향 등으로 인해 흔히 표본의 대표성이 문제점으로 지적되고 있다. 이런 문제점을 해결하기 위해 인터넷 사용자 표본이 전체 유권자 표본을 설명할 수 있도록 성향점수(propensity score)를 사용하여 가중치를 보정하는 방안을 제시한다. 17대 대선 자료를 기초로 한 사례분석을 통해, 적절한 성향점수보정 기법을 적용하는 경우 인터넷조사를 선거예측에 활용하는 것이 가능하다는 결론을 얻을 수 있었다.

주요용어: 가중치, 대통령선거조사, 성향점수모형, 인터넷조사.

1. 서론

최근 인터넷과 전자우편의 발전은 사람들의 의사소통 방법에 많은 변화를 주었고, 여론조사 방법에도 많은 변화를 가져왔다. 이런 사회 변화에 따라 최근에 많이 사용되는 인터넷조사는 확률추출법에 근거하지 않는 조사방법이기 때문에 많은 여론조사 전문가들은 인터넷조사를 신뢰할 수 있는 조사로 받아들이지 않고 있다. 그러나 유럽과 미국의 연구자들과 일부 조사기관에서는 선거결과 예측을 위한 몇몇 조사에서 할당추출법에 근거한 인터넷조사의 활용 가능성을 보여주고 있다 (Rosenbaum과 Rubin, 1984).

인터넷조사는 조사 대상자가 인터넷 사용자로 제한됨에 따라 발생하는 포함오차(coverage error)와 표본 선정과정에서 자발적인 참여 의사를 갖는 사람들만을 조사에 참여시킴으로써 발생하는 선택편향(selection bias) 등으로 인해 조사의 신뢰성에 한계가 있는 것으로 지적되고 있다. Lee (2006) 등은 이런 인터넷조사가 갖고 있는 문제점을 해결하기 위한 방안으로 성향점수보정(propensity score adjustment; PSA) 기법을 활용하는 방안을 제시하고 있다. Taylor (2000)와 Taylor 등 (2001)에 의하면 Harris Interactive에서는 미국 대통령 선거와 상원위원 및 주지사 선거예측을 위해 이런 기법을 적용한 것으로 알려져 있다. 국내에서는 김원용과 이홍철 (2003)이 선거예측조사를 목적으로 웹조사의 모집단 대표성을 확보하기 위해 성향가중모형을 적용하는 기초적인 연구를 수행하였다. 하지만 아직 우리나라에서 본격적으로 성향점수를 선거예측조사에 도입하기 위해 고려해야 할 가장 중요한 문제인 조사자의 어떤 특성이 성향점수 가중치 산출 모형에 반영되는 것이 요구되는지 등, 선거예측에 실제 이런 기법을 활용하기 위한 연구가 거의 이루어지지 못하고 있는 실정이다.

본 연구에서는 국내에서 아직 본격적인 연구가 미진한 선거예측조사에서 성향점수 가중치 적용 기법의 도입 가능성을 구체적으로 살펴보고자 한다. 이를 위해 우선 기존 연구에 제시되어 있는 인터넷

본 연구는 숙명여자대학교 2008학년도 교내연구비 지원에 의해 수행되었음.
[1] 교신저자: (140-742) 서울 용산구 효창원길 52, 숙명여자대학교 수학통계학부, 교수. E-mail: ywkim@sm.ac.kr

조사에서의 성향점수 보정방법을 정리한다. 우리나라 여론조사에서 이런 기법의 도입 가능성을 검토해 보기 위해 2007년 12월에 실시된 17대 대통령선거를 위한 여론조사 자료인 NI Korea의 인터넷 패널조사와 MBMR의 대선패널 전화조사를 비교하고, 현재 일반적으로 선거예측에 활용되는 전화조사를 참조조사(reference survey)로 활용하여, 인터넷조사의 편향을 제어하기 위해 우리나라 대통령선거예측조사에서 실제 활용할 수 있는 성향점수 가중치 보정방안을 제시한다.

2. 인터넷조사에서 성향점수보정

일반적인 인터넷(Internet) 또는 웹(Web) 조사의 자료수집 과정을 살펴보면, 먼저 사전에 자발적인 의사에 따라 대규모 패널을 구축하고 연구 목적에 따라 조사 대상자를 구축된 패널에서 추출해 표본을 구성한다. 추출된 표본을 대상으로 참여를 유도한 후, 최종 조사 자료는 이들 중 자발적으로 해당 조사에 응답한 사람으로부터 얻게 된다. 이와 같이 인터넷조사의 표본추출 및 조사과정은 비확률추출법(nonprobability sampling)을 기초로 하고 있으며 응답률이 높지 않기 때문에 표본 선택확률을 파악할 수 없을 뿐만 아니라 자발적 참여 및 무응답 등으로 인해 다양한 형태의 편향이 발생할 소지가 많다. 결과적으로 사전에 구축된 패널을 이용한 인터넷조사의 가장 큰 단점은 연구 대상 모집단을 대표할 수 있는 조사 자료를 확보할 수 없다는 것이다.

이런 문제를 해결하기 위해 지역, 성별, 연령대 등의 인구통계학적 요소를 기준으로 한 기존의 사후층화 방법을 이용하는 방법을 고려해 볼 수 있지만, 이런 기존의 사후층화를 통해서는 인터넷 조사에서 발생하는 편향을 보정하는데 한계가 있다 (Vehovar와 Manfreda, 1999). 따라서 인터넷 조사가 갖고 있는 한계를 해결하기 위한 방편으로 최근 성향점수보정 기법이 활용되고 있다 (Lee, 2006). 이 방법은 성향점수를 이용한 가중치 조정과정을 통해 인터넷조사에서 흔히 발생하는 편향을 줄이는 방법이다. 원래 성향점수보정 기법은 관측연구(observation study)에서 비교집단 사이에 존재하는 개체 특성의 차이를 조정하여 각 집단별 개체의 선택절차(selection mechanism)에 따른 영향을 완화하는 방안으로 도입된 기법이다 (Rosenbaum와 Rubin, 1983, 1984; D'Agostino, 1998).

자발적 참여를 통해 구성된 패널을 기초로 한 인터넷조사를 위한 성향점수보정은 참조조사(reference survey)가 존재한다는 가정에서 시작한다. 참조조사는 인터넷조사와 비슷한 시점에 실행되어야 하고, 면접조사, 전화조사 등과 같이 신뢰성이 인정된 전통적인 조사방법에 의한 것으로 높은 응답률을 갖는 양질의 조사이여야 한다. 성향점수보정 기법은 인터넷조사가 갖고 있는 근본적인 한계를 극복하기 위해 참조조사를 벤치마킹 하는 것이다. 즉, 성향점수모형을 매개로 주요 공변량을 대상으로 인터넷조사의 표본분포가 신뢰할 수 있는 참조조사의 표본분포와 일치하도록 사후적으로 가중치를 조정하는 것이다. Taylor (2000)와 Taylor 등 (2001)에 의하면 관측연구에서 집단 비교에서 사용되고 있는 성향점수모형을 인터넷조사에 처음 활용한 것은 미국 대선 예측에서 Harris Interactive에 의해 것으로 볼 수 있다. Lee (2006)가 제시한 인터넷 패널 조사를 위한 성향점수보정 과정을 정리하면 다음과 같다.

우선 인터넷조사(Web조사라고도 함)와 참조조사에 대해 다음 기호를 사용하기로 한다.

s^W: 인터넷조사 표본, n^W: 인터넷조사 표본크기, d_j^W: 인터넷조사 기본가중치($j = 1, 2, \ldots, n^W$)

s^R: 참조조사 표본, n^R: 참조조사 표본크기, d_k^R: 참조조사 기본가중치($k = 1, 2, \ldots, n^R$).

성향점수보정을 위해 우선 두 표본을 결합하여 총괄표본(s)을 구성한다. 즉, 총괄표본 $s = (s^W \cup s^R)$이고, 표본크기는 $n = n^W + n^R$이다. 총괄표본(s)에서 성향점수를 계산하게 되는데, i번째 단위의 인터넷조사 참여 성향점수는 $e(x_i) = P(i \in s^W | x_i)$에 해당한다. 이는 주어진 조건(공변량 x_i라는 조건)에서 i번째 단위가 인터넷조사에 참여할 확률을 나타내며, 총괄표본에서 다음과 같은 로지스틱 모형을 이용해

성향점수를 추정할 수 있다.

$$\ln\left[\frac{e(x)}{1-e(x)}\right] = \alpha + \beta^T f(x),$$

여기서 x는 공변량 벡터, $f(x)$는 공변량 벡터의 함수를 나타낸다.

성향점수보정은 추정된 성향점수를 기초로 총괄표본의 조사단위들을 성향점수에 따라 몇 개의 계급(class)으로 구분하고, 각 계급에 대한 s^W와 s^R에서의 상대적인 비중이 같아지도록 가중치를 보정하여, s^W에서 보정된 가중치를 적용한 계급별 분포가 s^R에서 계급별 분포와 일치하도록 s^W의 가중치를 조정하는 과정을 말한다. 이런 과정은 단계별로 다음과 같이 정리될 수 있다.

우선, 총괄표본(s)에서 모든 단위를 추정된 성향점수 크기에 따라 정렬한 후 C개의 계급으로 분할한다. 계급을 구성함에 있어서 각 계급에 같은 수의 단위가 포함되도록 한다 (Cochran (1968)은 5분위수(quintile points)를 기초로 하여 다섯 개의 계급으로 분할하는 것을 권장하고 있음). 각 계급내의 모든 단위들이 동일한 성향점수를 갖는 것이 이상적이지만 실제 문제에서는 각 계급내의 단위들의 성향점수가 큰 차이가 없도록 한다. 구성된 C개의 계급을 $s_c = s_c^W \cup s_c^R$로 표기하면, c번째 계급은 $n_c = n_c^W + n_c^R$개의 단위로 구성된다. 여기서 $s_c^W(n_c^W)$와 $s_c^R(n_c^R)$은 c계급 내에서 s^W와 s^R와 연계된 표본(단위수)를 의미한다.

우선 각 계급에서 다음과 같이 조정인자(adjustment factor)를 계산한다.

$$f_c = \frac{\Sigma_{k\in s_c^R} d_k^R / \Sigma_{k\in s^R} d_k^R}{\Sigma_{j\in s_c^W} d_j^W / \Sigma_{j\in s^W} d_j^W}. \tag{2.1}$$

식 (2.1)에서 기본가중치가 선택확률의 역수라면 다음과 같이 표현될 수 있다.

$$f_c = \frac{\Sigma_{k\in s_c^R} d_k^R / \Sigma_{k\in s^R} d_k^R}{\Sigma_{j\in s_c^W} d_j^W / \Sigma_{j\in s^W} d_j^W} \equiv \frac{\hat{N}_c^R/\hat{N}^R}{\hat{N}_c^W/\hat{N}^W}.$$

c계급 내에서 인터넷 표본 단위 j에 대한 성향점수보정(PSA) 인자는 다음과 같다.

$$d_j^{W.PSA} = f_c d_j^W = \frac{\hat{N}_c^R/\hat{N}^R}{\hat{N}_c^W/\hat{N}^W} d_j^W. \tag{2.2}$$

만약, 단순확률추출과 같이 모든 단위들에 대한 기본가중치가 같거나 알 수 없는 경우에 다음의 조정인자를 사용할 수 있다.

$$f_c = \frac{n_c^R/n^R}{n_c^W/n^W}. \tag{2.3}$$

식 (2.2)에 의한 성향점수보정 가중치를 사용하면 성향점수로 구성한 계급에 대한 인터넷조사 표본에서 가중 분포와 참조조사 표본에서 가중 분포가 같아진다. 예를 들어 식 (2.2)의 성향점수보정 가중치를 사용하게 되면 인터넷조사에서 계급 c에 해당하는 모집단 크기에 대한 추정값은 다음과 같아진다.

$$\hat{N}_c^{W.PSA} = \sum_{j\in s_c^W} d_j^{W.PSA} = \hat{N}^W \frac{\hat{N}_c^R}{\hat{N}^R}.$$

한편, s^W로부터 얻어지는 특정 변수(y)에 대한 평균은 다음과 같이 추정되며, 여기서 추정값 $\hat{\bar{y}}^{W.PSA}$를 구할 때, 참조조사 표본은 사용되지 않는다는 점에 유의할 필요가 있다.

$$\hat{\bar{y}}^{W.PSA} = \frac{\Sigma_c \Sigma_{j\in s_c^W} d_j^{W.PSA} y_i}{\Sigma_c \Sigma_{j\in s_c^W} d_j^{W.PSA}}.$$

참고로 기존의 확률추출법에 의한 조사에서는 선택확률에 따른 가중치를 산출하여 비편향 추정량을 유도하는 것이 가능하기 때문에 선택편향을 조정하는 목적으로 성향점수보정 기법을 사용할 필요가 없었다. 하지만 포함오차 (Duncan과 Stasny, 2001), 무응답오차 (Smith 등, 2000; Vartivarian과 Little, 2003) 등에 따른 편향을 줄이기 위해 가중치를 조정하는 방법으로 사용되어 왔다.

3. 선거예측조사에서 성향점수보정 실증분석

3.1. 실증분석 대상 자료 개요

실증분석을 통해 자발적 패널을 이용한 인터넷조사에 PSA를 적용함으로써 인터넷조사의 선택편향을 어느 정도 보정할 수 있는지, 다시 말해 인터넷 조사에서 PSA의 활용을 통한 효과를 실제 대통령 선거 여론조사 자료를 통해 실증적으로 검토해 보고자 한다.

본 실증분석을 위한 인터넷 선거여론조사 자료는 2007년 8월 22일부터 23일까지 NI Korea에서 운영하는 패널사이트(PamiClub)의 20세 이상의 패널을 대상으로, 전체 패널 중 자발적으로 조사에 참여한 3,123명으로부터 얻은 응답결과이다. 이 자료에는 조사시점에 응답자가 지지하는 후보자와 함께 응답자의 성별, 연령, 교육수준, 거주지역 등의 정보가 포함되어 있다.

한편, 본 연구에서 참조조사(reference survey)로 사용된 전화여론조사는 KBS와 MBMR(이하 'KBS')가 공동으로 수행한 17대 대선패널 조사이다. KBS에서 실시한 대선패널 전화여론조사는 우리나라 전체 유권자들을 대상으로 패널을 구축했으며, 표본의 대표성 확보를 위해 현재 우리나라에서 선거조사에서 흔히 사용하는 성/연령 등에 따른 할당추출을 근간으로 하는 전화조사와는 달리 RDD(random digit dialing) 방법을 이용하여 표본을 추출했다. KBS 패널은 2007년 8월 10일부터 14일까지 5일 동안 총 7회에 걸친 콜백(call-back) 과정을 통해 구축했으며, 동일 표본 유권자를 대상으로 8월 15일부터 8월 17일까지 1차 조사가 이루어진 후, 대선 투표일까지 매달 조사가 수행되었다.

본 연구에서는 NI Korea의 인터넷 패널조사의 선택편향 보정을 위한 PSA 활용에 필요한 참조조사로 KBS의 대선패널 조사 자료를 활용한다. KBS 패널조사는 수차례 걸쳐 조사가 이루어졌기 때문에 인구사회통계 변수 이외에도 정치성향, 매체접촉현황, 전자우편을 포함하여 인터넷 사용 여부 등 다양한 정보를 얻을 수 있다는 특징을 갖고 있다. 따라서 인터넷조사의 PSA를 이용한 가중치 보정에 있어서 성별, 연령, 교육수준, 거주지역 등의 인구사회통계 관련 정보뿐만 아니라 선거예측에 있어서 많은 영향을 줄 수 있는 지지정당, 과거 투표성향 등의 정보를 활용할 수 있다는 장점을 갖고 있다.

실제 이들 두 개 조사는 별도의 목적으로 수행된 선거여론조사로 본 연구에서 다루는 PSA의 적용을 염두에 두고 기획된 것이 아니다. 따라서 본 실증분석에서는 NI Korea의 인터넷 조사가 수행된 시점에 해당하는 한나라당 경선(8월 19일)과 대통합민주신당 경선(10월 14일) 사이의 시점에서 대선 후보로 거론되고 있는 주요 후보들의 지지율 예측을 분석대상으로 하고, 이 시점과 비슷한 시기에 실시된 KBS 대선패널 2차 조사(응답자 수 2,162명)를 활용해 NI Korea 인터넷 조사 결과를 보정하는 경우 인터넷 조사와 참조조사의 후보자별 지지율에 있어서 어떤 차이가 발생하게 되는지 분석해 보고자 한다.

아울러 NI Korea의 인터넷조사 자료에는 무응답이 없지만, KBS 패널 조사(2차 조사)에서는 일부 항목 무응답이 있다. 주어진 변수 중 일부 변수에 결측값(무응답)이 존재하는 경우, PSA를 생성하기 위한 로지스틱 모형을 적합시키거나, 주요 변수를 선택하는 과정을 효과적으로 수행할 수 없게 된다. 따라서 KBS 패널 자료의 경우 일반적인 무응답 대체방법을 통해 항목 무응답을 먼저 처리하여 완벽한 자료를 만든 후 실증분석을 실시했다.

여기서는 실제 대선 선거여론조사를 대상으로 한 실증분석을 통해 우리나라 인터넷 선거여론조사의 편향 보정을 위해 PSA를 적용하는 경우 어떤 변수들을 포함하는 것이 효과적이며 또한 우리나라 선거여론조사에서 PSA 기법을 적용한 인터넷 조사의 활용 가능성을 검토해 본다.

표 1: KBS와 NI Korea 표본 지역/성별/연령 분포 비교 (단위: %)

구분		KBS 전체 표본	KBS 인터넷사용자	NI Korea 표본
지역	서울	21.42	21.75	31.73
	부산	7.45	8.11	7.78
	대구	4.76	5.10	5.19
	광주	2.78	2.59	3.43
	인천	5.23	5.10	5.54
	대전	3.01	3.43	3.65
	울산	1.90	2.03	1.34
	경기	22.25	22.03	20.94
	강원	3.33	2.94	2.24
	충북	3.61	3.36	2.53
	충남	3.61	3.57	2.40
	전북	4.16	3.71	2.72
	전남	3.70	3.36	2.15
	경북	4.76	4.62	3.59
	경남	6.57	6.64	4.26
	제주도	1.48	1.68	0.51
성별	남자	49.77	51.12	59.05
	여자	50.23	48.88	40.95
연령	20대	19.61	24.76	26.42
	30대	24.70	31.47	36.57
	40대	22.90	25.24	19.50
	50대이상	32.79	18.53	17.52
합계		100	100	100

3.2. KBS 전화조사와 NI Korea 인터넷조사 표본 구성

인터넷조사에서 발생하는 선택편향을 파악하기 위해서는 우선 자발적 참여에 의한 인터넷조사 표본과 전화조사 표본에서 발생하는 다양한 인구사회통계학적 특성에서 어떤 차이가 있는지 살펴보는 것이 중요하다. 아울러 KBS 전화조사 표본의 경우, 조사 문항 중 인터넷 사용여부가 포함되어 있기 때문에 전체 응답자 중 인터넷을 사용하는 유권자들을 구분하는 것이 가능하다. KBS 전화조사 전체 응답자 2,162명 중 실제 인터넷 조사가 가능하다고 볼 수 있는 인터넷 사용자는 응답자 중 66%에 해당하는 1,430명인 것으로 나타났다. 이를 통해 NI Korea의 인터넷 조사 표본과 같이 자발적인 의사에 의해 이루어지는 인터넷 조사에서의 표본과 RDD 방식을 통해 인터넷 사용자를 표본으로 추출하는 경우 얻을 수 표본의 특성을 비교하는 것도 의미 있는 결과가 될 수 있다.

KBS 전체 표본과 인터넷사용자 표본 그리고 NI Korea 인터넷조사 표본을 일반적인 여론조사에서 흔히 표본 할당(quota) 변수로 사용하는 지역, 성별, 연령대로 구분하여 분포를 비교해 보면 표 1과 같다. 지역 분포에 있어서는 NI Korea 인터넷조사 표본의 경우 서울이 차지하는 비중이 높은 것을 볼 수 있지만 다른 지역에 있어서는 구성 비율에 있어서 큰 차이를 보이지 않는다. 한편 NI Korea 인터넷조사 표본에서 '남자' 구성 비율이 높고, 예상대로 특히 '50대'의 구성 비율은 매우 낮은 것을 볼 수 있다. 한편 KBS 표본에서 인터넷 사용자 표본의 분포를 보면 전반적으로 KBS 전체 표본과 NI Korea 표본의 중간수준의 분포를 보이는 것으로 나타났다.

표본을 직업, 교육수준, 가구소득에 따라 구분해 분포 현황을 비교해 보면 표 2와 같다. 표 2를 보면 자발적인 참여에 의해 구성된 NI Korea의 인터넷조사 표본에서는 전화조사 표본과 비교해 '사무/기술직'과 '경영/관리/전문직'의 비율이 매우 높고, 반면에 '일반노무직'과 '전업주부'는 비율이 매우 낮은 것을 알 수 있다. 교육수준에서는 인터넷조사의 경우 '대학교 재학' 이상의 고학력자의 구성 비율이

표 2: KBS와 NI Korea 표본의 직업/교육수준/가구소득 분포 비교 (단위: %)

구분		KBS 전체 표본	KBS 인터넷사용자	NI Korea 표본
직업	농/임/어/축산업	4.35	2.80	0.58
	자영업	15.36	16.22	10.34
	일반노무직	10.41	11.33	1.63
	사무/기술직	17.99	19.86	34.81
	경영/관리/전문직	2.59	2.87	19.53
	학생	10.18	10.77	10.41
	전업주부	30.11	29.58	11.40
	무직	8.33	5.73	3.87
	기타	0.69	0.84	7.43
교육수준	중학교졸업 이하	16.56	9.09	0.99
	고등학교 졸업	32.70	34.83	20.75
	대학교 재학/졸업	45.42	49.51	66.83
	대학원재학 이상	5.32	6.57	11.43
가구소득	100만원 이하	14.48	4.83	4.87
	101만원~150만원	6.29	3.99	8.29
	151만원~200만원	10.45	10.77	10.21
	201만원~250만원	10.18	11.61	11.14
	251만원~300만원	15.26	18.46	13.13
	301만원~400만원	17.48	22.10	21.81
	401만원~500만원	12.67	12.87	16.43
	501만원~600만원	6.52	7.48	6.50
	601만원~700만원	2.13	2.52	2.50
	700만원 이상	4.53	5.38	5.12
합계		100	100	100

높은 반면 저학력자의 구성 비율은 낮은 것으로 나타났다. 가구소득에서는 특히 월 소득 '100만원 이하'인 가구 구성 비율에서 큰 차이를 보이고 있다. 또 다른 특이한 점은 KBS 표본 중 인터넷 사용자들의 분포는 NI Korea 인터넷조사 표본보다는 KBS 전체 표본의 분포에 더 가깝다는 것이다. 이는 인터넷 조사를 하는 경우에도 RDD 전화조사 등을 통해 좀 더 대표성 있는 인터넷조사 표본을 추출하게 되면 자발적인 참여를 통해 구성되는 인터넷조사 표본과 상당히 다른 양상을 보일 수 있다는 것을 보여준다. 이런 사실은 향후 보다 대표성 있는 인터넷조사를 원하는 경우 어떤 방식의 표본추출방법이 도움이 될 수 있는지를 알려주는 의미 있는 결과로 보인다.

선거예측을 위한 성향점수보정 기법을 개발하기 위해서는 인구사회통계학적 변수만을 이용해서는 한계가 있다. 따라서 응답자의 투표 및 정치 성향을 파악할 수 있는 변수들을 PSA 모형에 반영하는 경우 보다 정교한 가중치 보정이 가능해질 수 있다. KBS 전화조사와 NI Korea 인터넷조사에서 공통적으로 조사된 정치성향과 관련된 항목들에 대한 분포 현황을 정리하면 표 3과 같다.

표 3에서 2002년 대선 투표성향을 보면 자발적인 참여를 통해 구성된 인터넷조사 표본인 NI Korea 조사에서 이회창후보에 대한 지지율이 낮아지는 현상을 볼 수 있다. 2004년 총선에 대한 경우 NI Korea 표본에서 한나라당 지지율이 낮고, 열린우리당에 대한 지지율이 높게 나타나고 있다. 아울러 응답자들이 지지하는 정당에 있어서도 NI Korea 표본에서 한나라당에 대한 지지율이 낮고, 민주노동당에 대한 지지율이 높은 것으로 나타났다. 이런 현상은 자발적인 인터넷조사 참여자 집단이 보다 진보적인 정치성향을 갖고 있다는 일반적인 예상과 일치한다.

3.3. 기본가중치 적용에 따른 지지율 차이

우리나라 대부분의 선거여론조사에서는 시도, 성별, 연령대에 따른 유권자 분포를 고려한 할당추

표 3: KBS와 NI Korea 표본의 정치성향에 따른 분포 비교 (단위: %)

구분		KBS 전체 표본	KBS 인터넷사용자	NI Korea 표본
2002년 대선 투표결과	이회창	34.78	31.33	28.59
	노무현	45.24	46.92	42.36
	권영길	2.17	2.73	3.23
	기타	0.23	0.21	2.18
	투표하지 않았다.	13.55	16.43	14.67
	말 할 수 없다.	4.02	2.38	8.97
2004년 총선 투표결과	열린우리당	21.05	23.01	27.79
	한나라당	43.39	41.40	32.69
	민주당	6.98	5.80	4.13
	민주노동당	4.35	5.17	8.87
	기타	2.45	2.10	1.99
	투표하지 않았다.	15.36	17.34	16.30
	말 할 수 없다.	6.43	5.17	8.23
지지정당	한나라당	55.04	54.13	44.89
	대통합민주신당	15.12	16.85	9.86
	중도통합민주신당	4.39	3.85	3.43
	민주노동당	5.41	6.43	14.06
	국민중심당	0.19	0.14	2.15
	기타	19.84	18.60	25.62
17대 대선 투표의향	반드시 투표할 것이다.	75.30	72.17	49.76
	웬만하면 할 것이다.	22.11	25.17	40.99
	별로 투표하고 싶지 않다.	2.08	2.24	8.20
	전혀 투표할 생각이 없다.	0.51	0.42	1.06
	합계	100	100	100

출을 하고 있으며, 각 그룹별로 할당된 표본을 채우지 못하는 경우 시도/성별/연령대별 유권자수를 기준으로 한 가중치를 적용하는 것이 일반적이다. 본 연구에서 분석대상으로 하는 KBS 전화조사의 경우 후보자별 지지율 예측을 위해 통계청의 주민등록인구통계를 기준으로 작성된 시도별, 성별, 연령별 구성비를 기준으로 산출한 가중치를 사용하고 있다. 한편 NI Korea 인터넷조사에서는 시도별, 성별, 연령별 및 교육수준을 고려하기 위해 2005년 인구주택총조사 통계를 기준으로 한 가중치를 사용하고 있다(지금부터 이들 가중치를 기본가중치라고 함). 두 조사에서 기본가중치를 적용하는 경우 KBS와 NI Korea 조사의 후보자별 지지율 추정결과에 있어서 표 1에 나타난 지역/성별/연령대 분포가 다르기 때문에 발생하는 차이는 상쇄된 것으로 볼 수 있다.

KBS 전화조사와 NI Korea 인터넷조사에서 사용하는 기본가중치를 적용한 추정결과와 가중치를 적용하지 않고 산출된 단순 평균에 의한 후보자별 지지율 추정결과를 비교해 보면 표 4와 같다. 기본가중치를 적용한 경우에도 이명박 후보의 지지율이 KBS의 경우 59.43%, NI Korea의 경우 46.79%를 나타내는 등 큰 차이를 보이고 있다. 특히 두 조사에서 기본가중치를 적용했을 때 단순평균의 경우보다 이명박후보에 대한 지지율 차이가 더 벌어지는 것을 볼 수 있다. 이런 차이는 다른 요인에 의한 영향도 있겠지만 NI Korea의 경우 자발적으로 참여를 원하는 인터넷 사용자만을 조사대상에 포함했기 때문에 발생하는 선택편향에서 그 원인을 찾을 수 있을 것이다.

이런 현상은 동일한 성별과 연령대의 유권자라고 해도 인터넷 사용 여부에 따라 투표 성향에 있어서 차이가 있다는 것을 시사한다. 따라서 인터넷조사의 경우 인터넷 사용자만을 표본에 포함함으로써 전체 유권자 모집단에 대한 추정값을 얻는데 한계가 있다는 것을 확인할 수 있다. 이런 문제점을 해결하기 위해서는 인터넷조사 결과에 대한 추가적인 보정이 필요하며, 이런 보정을 위해서는 단순히 성별과 연령대 같은 인구통계학적 변수만을 고려해서는 소기의 목적을 다를 수 없다. 결국 선거예측을 위

표 4: KBS 표본과 NI Korea 표본의 후보자별 지지율 (단위: %)

후보	단순 평균		기본가중치 적용	
	KBS	Ni Korea	KBS	Ni Korea
이명박	59.39	48.61	59.43	46.79
손학규	6.15	11.88	6.16	9.87
이인제	1.30	0.96	1.28	0.99
정동영	8.23	3.81	8.26	4.71
권영길	1.85	1.95	1.77	1.73
기타	8.93	12.97	8.96	13.78
모름	14.15	19.82	14.14	22.14
합계	100	100	100	100

표 5: 로지스틱 회귀모형을 이용한 변수 선택 결과

구분	변수		유형	p-value	선택 변수
인구사회통계학적 변수	area	지역	범주형	0.3779	
	gender	성별	범주형	< .0001	√
	age	연령	범주형	0.0341	
	job	직업	범주형	0.0813	
	edu	교육수준	범주형	0.7745	
	inc	가구소득	순서형	< .0001	√
정치성향 변수	p2002	2002년 대선 투표후보	범주형	< .0001	√
	p2004	2004년 총선 투표정당	범주형	0.6282	
	vote	17대 대선 투표의향	범주형	< .0001	√
	party	지지정당	범주형	< .0001	√

한 PSA 기법의 활용에 있어서는 표 3과 같은 유권자의 정치성향을 나타내는 변수를 모형에 반영하는 것이 필요하다는 것을 추론할 수 있다.

3.4. PSA 적용을 위한 변수 선택

KBS 전화조사와 NI Korea 인터넷조사에서 공통적으로 조사된 변수들 중에서 후보자 지지율에 영향을 줄 것으로 예상되는 변수를 정리하여 보면 인구사회통계학적 변수 6개와 정치성향과 관련된 변수 4개가 포함되어 있다. 따라서 본 연구에서는 이들 10개 변수를 활용하도록 한다. 이들 변수 중에서 후보자 지지율에 대한 설명력이 높은 주요변수를 파악하기 위해 두 자료를 합하여 다항 로지스틱 회귀모형을 이용하여 주요 변수를 선택했다. 변수선택은 단계별 선택방법을 적용하였으며, 변수 선택기준으로 유의수준 $\alpha = 0.05$를 사용했다.

변수선택 결과를 정리하면 표 5와 같으며, 인구사회통계학적 변수 중에서는 '성별'과 '가구소득'이 선택 되었고, 정치성향 변수 중에서는 '2002년 대선 때, 투표한 후보', '17대 대선 투표의향' 및 '지지정당'이 선택되었다.

실제 PSA 가중치 보정을 수행함에 있어서 정치성향 관련 변수를 모형에 포함하는 것이 효과적인지 또는 변수선택에 의한 주요 변수만을 모형에 반영하는 것이 효과적인지 등에 대해서는 실증분석 과정을 통해 확인해 볼 필요가 있다.

이런 측면에서 본 연구에서는 모형에 포함된 설명변수를 달리 하는 6개의 로지스틱 회귀모형(M1~M6)을 이용한 PSA 보정을 수행해 보고, 어떤 모형이 인터넷조사가 구조적으로 갖고 있는 한계를 극복하는데 도움이 되는지 검토해 보기로 한다. 본 연구에서 검토한 6개의 모형에 포함된 변수들을 정리하면 표 6과 같다. 6개의 로지스틱 회귀모형 중, 주요변수만을 포함하는 모형(M1)을 예로 들면,

표 6: 6개 로지스틱 회귀모형에 포함된 변수

변수		[M1] 변수선택	[M2] 인구사회	[M3] 정치성향	[M4] 변수선택 & 인구사회	[M5] 변수선택 & 정치성향	[M6] 전체 변수
인구사회 통계학적 변수	area		√		√		√
	gender	√	√		√	√	√
	age		√		√		√
	job		√		√		√
	edu		√		√		√
	inc	√	√		√	√	√
정치성향 변수	p2002	√		√	√	√	√
	p2004			√		√	√
	vote	√		√	√	√	√
	party	√		√	√	√	√

표 7: 모형별 PSA 가중치 보정에 따른 인터넷조사 후보자별 지지율 (단위: %)

후보	KBS 기본가중치	[M1] 변수선택	[M2] 인구사회	[M3] 정치성향	[M4] 변수선택 & 인구사회	[M5] 변수선택 & 정치성향	[M6] 전체 변수
이명박	59.43	54.27	50.23	55.28	58.13	54.21	59.47
손학규	6.16	9.67	8.38	10.67	8.38	9.88	9.00
이인제	1.28	0.79	1.41	0.66	0.94	0.75	0.95
정동영	8.26	4.87	4.83	4.62	5.12	4.80	4.85
권영길	1.77	1.02	1.49	0.92	1.12	1.00	1.15
기타	8.96	12.65	13.02	12.76	10.81	12.71	10.88
모름	14.14	16.72	20.65	15.09	15.51	16.64	13.70
합계	100	100	100	100	100	100	100

PSA는 다음과 같은 로지스틱 회귀모형을 이용하여 개인별 인터넷 사용여부에 대한 성향 점수를 산출하여 활용한 것이다. 여기서 g는 인터넷 사용여부를 나타낸다.

$$\ln\left[\frac{\Pr(g=1)}{1-\Pr(g=1)}\right] = \alpha + \beta_1[\text{성별}] + \beta_2[\text{가구소득}] + \beta_3[\text{지지정당}] + \beta_4[\text{2002년 대선 투표후보}] + \beta_5[\text{17대 대선 투표의향}].$$

3.5. PSA 적용에 의한 가중치 보정 결과

각 모형에서 개인별 성향점수를 산출해 Cochran (1968)이 제안한 방법에 따라 5분위수(quintile points)를 기초로 다섯 개의 계급으로 분할하고 PSA 가중치 보정 후, 후보자별 지지율을 추정한 결과는 표 7과 같다. 기본가중치를 적용한 KBS 전화조사에서 후보자별 지지율과 각 모형을 이용해 PSA 가중치 보정을 통해 NI Korea 인터넷조사에서 산출한 후보자별 지지율의 차이는 표 4의 기본가중치를 적용한 결과에 비해 상당히 줄어든 것을 볼 수 있다.

전반적으로 정치성향 변수가 추가됨에 따라 KBS 전화조사와 지지율 차이가 작아지는 것을 알 수 있으며, 모든 변수를 포함한 경우(M6) KBS 전화여론조사와의 차이가 가장 작은 것을 알 수 있다. 전화조사와의 차이를 최대한 줄인다는 관점에서는 M6과 같이 모든 변수를 포함하는 것이 효과적일 수 있다. 비용 대비 효율성을 생각했을 때 변수선택 과정을 통해 선정된 변수만을 모형에 포함 하는 것(M1)이 하나의 방안이 될 수 있다. 한편, 인구사회통계학적 변수만을 포함한 모형(M2) 보다 정치성

향 관련 변수만을 포함한 모형(M3)이 KBS 전화여론조사와 차이가 작아지고 있다는 점에 유의할 필요가 있으며, 이는 선거예측을 목적으로 하는 경우 PSA의 적용에 있어서 과거 선거에서의 투표성향이나 지지정당 관련 변수가 모형에 포함되는 것이 필수적이라는 것을 보여주고 있다.

4. 결론 및 향후과제

흔히 자발적인 참여를 통해 수행되는 인터넷 선거여론조사의 경우 포함오차와 선택편향 등으로 인해 일반적인 유권자 모집단을 설명하는 데 한계가 있다. 본 연구결과를 통해 이런 인터넷 선거예측조사가 갖고 있는 한계는 지역, 성별, 연령대 등 일반적인 여론조사에서 사용되는 인구통계학적 변수를 벤치마킹하는 사후층화 과정을 통해 해결될 수 없다는 것을 볼 수 있었다. 또한 인터넷조사를 선거예측에 활용하기 위해서는 인터넷조사 응답자의 정치적인 성향을 파악할 수 있는 정보가 필요하며, 이런 정보가 확보되는 경우 사전에 적절한 PSA 모형을 설정하고, 이를 이용한 가중치 보정과정을 통해 최소한 참조조사(reference survey)와 유사한 수준의 정확성을 갖는 선거예측조사가 가능하다는 것을 본 연구결과를 통해 확인할 수 있었다. 이런 연구가 향후 좀 더 체계적으로 이루어진다면, 우리나라에서도 PSA의 활용을 통해 정확성을 담보할 수 있는 선거예측을 위한 인터넷조사의 도입이 가능할 것으로 보인다.

본 연구에서 제시된 PSA 기법이 실제 인터넷조사를 기반으로 한 선거예측에 있어서 얼마나 효과적인지를 제대로 평가하기 위해서는 실제 선거가 임박한 시점에 추가적으로 실시된 인터넷조사에 본 연구에서 제시된 PSA 기법을 적용해 선거예측 정확성을 평가해 보는 것이 필요하다. 하지만 여기서는 현실적인 여건상 추가적인 인터넷조사가 이루어지지 않아서 결국 최종 선거예측 과정에 제시된 방법을 적용해 사후적으로 본 연구에서 제안된 방법의 정확성을 평가하는 과정을 수행하지 못했다는 한계를 갖고 있다는 점을 밝혀둔다.

제시된 PSA 기법을 통해 산출된 가중치 보정 방법은 향후 유사한 인터넷조사에서도 활용될 수 있을 것이며, 이를 통해 인터넷조사가 갖고 있는 한계를 상당 폭 넘어설 수 있을 것으로 예상된다. 물론 사회적인 이슈를 다루는 일반적인 인터넷 여론조사에 대한 PSA 가중치 보정을 위한 모형은 본 연구에서 제시한 모형과 상당히 다른 양상을 보일 수 있다는 점에 유의할 필요가 있다.

한편, 표 1과 2를 보면 KBS 표본 중 인터넷 사용자들의 분포는 NI Korea 인터넷조사 표본보다는 KBS 전체 표본의 분포에 더 가깝다는 것을 볼 수 있다. 따라서 만약 인터넷 조사를 수행하는 경우에도 NI Korea 인터넷조사와 같이 사전에 구성된 대규모 패널에서 자발적으로 조사에 참여하는 응답자들로 인터넷조사 표본을 구성하는 대신 RDD 방법과 같은 표본추출 이론에 따른 과학적인 표본추출 과정을 통해 인터넷 사용자 표본을 구성한다면 보다 대표성 있는 표본을 확보하는 것이 가능할 것으로 판단된다. 따라서 보다 정확성 있는 인터넷조사를 원하는 경우 어떤 방식의 표본추출방법이 현실적으로 도입이 가능하고 효과적인지 보다 심층적인 연구가 수행될 필요가 있으며 이런 연구들을 통해 인터넷조사의 활용도를 높일 수 있을 것이다.

참고 문헌

김원용, 이흥철 (2003). 웹조사의 모집단대표성 확보를 위한 성향가중 모형의 적합성 검증, <방송연구>, 여름호, 143–166.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies, *Biometrics*, **24**, 295–313.

D'Agostino, R. B. Jr. (1998). Propensity score methods for bias reduction for the comparison of a treatment to a non-randomized control group, *Statistics in Medicine*, **17**, 2265–2281.

Duncan, K. B. and Stasny, E. A. (2001). Using propensity scores to control coverage bias in telephone surveys, *Survey Methodology*, **27**, 121–130.

Lee, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel web surveys, *Journal of Official Statistics*.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects, *Biometrika*, **70**, 41–55.

Rosenbaum, P. R. and Rubin, D. B. (1984). Recucing bias in observational studies. Using subclassification on the propensity score, *Journal of the American Statisical Association*, **79**, 516–524.

Smith, P. J., Rao, J. N. K., Battaglia, M. P., Daniels, D. and Ezzati-Rice, T. (2000). Compensating for nonresponse bias in the national immunization survey using response propensities, In *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 641–646.

Taylor, H. (2000). Dose internet research Work? Comparing online survey result with telephone survey, *International Journal of Market Research*, **42**, 58–63.

Taylor, H., Bremer, J., Overmeyer, C., Siegel, J. W. and Terhanian, G. (2001). The record of internet-based opinion polls in predicting the results of 72 races in the November 2000 US Elections, *International Journal of Market Research*, **43**, 127–135.

Vartivarian, S. and Little, R. (2003). On the formation of weighting adjustment cells for unit nonresponse, University of Michigan Department of Biostatistics Working Paper Series.

Vehovar, V. and Manfreda, K. L. (1999). Web surveys: Can the weighting solve the problem? *Proceedings of American Statistical Association, Section on Survey Research Methods*, 962–967.

2009년 11월 접수; 2009년 12월 채택

Propensity Score Weighting Adjustment for Internet Surveys for Korean Presidential Election

Young-Won Kim[1,a], Ye-Young Be[a]

[a]Department of Statistics, Sookmyung Women's University

Abstract

Propensity score adjustment(PSA) has been suggested as approach to adjustment for volunteer internet survey. PSA attempts to decrease the biases arising from noncoverage and nonprobability sampling in volunteer panel internet surveys. Although PSA is an appealing method, its application for internet survey regarding Korea presidential election and its effectiveness is not well investigated. In this study, we compare the Ni Korea internet survey with the telephone survey conducted by MBMR and KBS for 2007 Korean presidential election. The result of study show that the accuracy of internet survey can be improved by using PSA. And it is critical to include covariates that highly related to the voting tendency and the role of nondemographic variables seems important to improving PSA for Korea presidential election prediction.

Keywords: Internet survey, presidential election survey, propensity score adjustment(PSA), weight.

This Research was supported by the Sookmyung Women's University Research Grants 2008.

[1] Corresponding author: Professor, Department of Statistics, Sookmyung Women's University, Seoul 140-742, Korea. E-mail: ywkim@sookmyung.ac.kr