

한국 사회조사 단위무응답 처리를 위한 최적의 보조정보 구성에 관한 실증적 연구

정구현¹⁾ · 이승희²⁾

요약

모든 가구 표본조사는 어떠한 이유에서든 무응답을 필연적으로 수반할 수밖에 없다. 대부분 무응답은 비접촉과 응답거부 형태로 이루어진다. 사회조사는 이러한 무응답에 관한 문제를 제거하기 위해 조사현장에서 표본대체(substitution) 방법을 사용해 왔다. 하지만 표본대체 방법은 선택편향 및 특정 계층에서 체계적 오차 발생 등에 대한 오랜 논란이 있어, 이를 불식시키기 위해 2016년 조사부터는 무응답에 대한 탐색 및 처리를 위한 새로운 방법을 도입했다. 무응답 처리를 위해 필수적으로 필요한 보조정보(auxiliary information)를 파라데이터(paradata)와 행정자료로 부터 확보가 가능했기 때문이다. 이 연구에서는 사회조사에 적합한 무응답 처리 방법을 위해 모의실험을 전개한다. 2016년 사회조사 결과를 준모집단(pseudo-population)으로 간주하여 일부 표본자료를 300회 추출하고, 여기에서 6가지 무응답 메커니즘에 따라 무응답을 발생시키고 나서 가중값 조정법을 통해 무응답을 처리한다. 무응답 처리를 위해서는 무엇보다 응답성향과 관련 있고 동시에 조사변수와 연관성 높은 보조정보 구성이 중요하다. 보조정보 구성은 무응답 편향, 분산, 설계효과 등에 직접적 영향을 미치고 있음을 확인 할 수 있었다. 무응답 처리결과를 보면, 주요 조사변수와 관련성이 높은 파라데이터와 행정자료로부터 통합자료를 보조정보로 구성한 경우에 무응답 처리 성능이 더욱 우수하며, 특히 비임의결측(NMAR) 자료의 경우에도 유의미한 결과를 보여준다.

주요용어 : 보조정보, 무응답 메커니즘, 파라데이터, 행정자료, 가중값 조정법

1. 서론

가구 표본조사에서 무응답은 어떠한 이유에서든 필연적으로 발생하고 있다. 무응답은 주로 조사 참여를 거부하는 응답 거부(refusal), 조사기간에 접촉할 수 없는 비접촉(non-contact), 건강상의 이유로 한 조사불가(inability) 등으로 발생하며, 최근 다른 국가들과 마찬가지로 조사환경 악화에 따라 무응답 증가 추세가 더욱 빠르게 진행되고 있다(de Leeuw and de Heer, 2002; Stoop, 2005; Brick and Williams, 2013). 특히 조사기간이 짧은 표본조사의 경우, 1인가구나 맞벌이 가구 증가에 따른 주간부재는 응답률에 더욱 부정적인 영향으로 작용하고 있다(김서영과 박라나, 2013).

이에 따라 무응답을 감소시키기 위해 조사기관 및 조사설계자는 방문시간 연구, 인센티브 제공, 방문횟수 확대, 충분한 조사기간 확보 및 혼합조사(mixed-mode) 활용 등 다방면으로 노력하고 있으나(Weeks, 1988; 이승희, 2010), 무응답 감소는 한계에 직면한 듯하다. 반면, 최근 무응답에 대한 이론적 발전과 필수적 보조정보를 이용하여 효과적으

1) 교신저자. Institute for Social Research, University of Michigan, 객원연구원(통계청 과장). E-mail: goohyunj@umich.edu

2) Institute for Social Research, University of Michigan. E-mail: sungheel@umich.edu

로 무응답을 처리하기 위한 노력은 크게 증가하고 있다.

통계청 사회조사(social survey)는 2015년까지 표본가구에서 조사를 거부하거나 몇 차례 방문에도 접촉이 불가능한 경우, 가장 현실적인 방법으로 조사현장에서 대체(substitution) 조사를 진행하여 무응답 문제를 완화시키려 시도해왔다. 이는 조사가 성공할 것으로 기대되는 가구로만 조사가 집중되며, 거절 후 응답전환(refusal conversion)이 가능한 대상을 쉽게 불응으로 처리하는 경향을 보였다. 이러한 문제가 선택편향(selection bias)의 원인이 되고 있음을 학계와 전문가들 사이에서 지속적으로 논의되어 왔다(Kish, 1965; Little and Rubin, 1987; Sharon, 2010). 기존분석을 통해서도 주택유형과 점유형태 등이 표본가구와 대체가구에서 유의미한 차이를 발견할 수 있으며, 이는 전체 자료에도 크게 영향을 미치는 것으로 확인할 수 있다(통계청, 2016a).

무응답 처리를 위해서는 응답대상 뿐 아니라 무응답 대상에 대한 정보를 갖춘 보조정보의 확보가 매우 중요한 역할을 한다. 하지만 접촉 불가능한 경우나 응답을 거부한 가구에 대해 명확하고 다양한 정보를 얻는다는 것은 쉬운 일이 아니며, 무응답 가구에 대한 정보 부족이나 결핍은 무응답을 효과적으로 처리할 수 없게 한다. 무응답 정보는 표본추출틀, 보고자료 및 센서스 개별자료 등을 통해 확보가 가능하다(Kalton and Maligalig, 1991). 다만 확보된 자료가 기준시점과 활용시점간 사이가 커지면 커버리지 문제 등 활용에 어려움이 당연하게 따를 수밖에 없다. 다행히 사회조사는 2015년부터 조사무응답에 대한 관심 확대와 적절한 처리를 위한 한 방법으로 파라데이터(paradata) 수집을 진행하고 있으며, 또한 센서스 전수항목 작성을 위해 매년 지속적으로 확보되고 있는 행정자료를 주요 보조정보로 활용할 수 있게 됐다.

이와 함께 특정 그룹에서 무응답률에 대한 체계적인 차이는 무응답 편향과 관련된 다는 점에서 특성별 응답성향에 대해 세부적으로 분석할 필요가 있다(Brick 등, 2000). 무응답 발생이 특정 변수와 무관하게 무작위로 발생한다면, 어느 정도 무응답률 자체는 큰 문제가 되지 않을 수도 있다(Curtin 등, 2000). 하지만, 실제 조사에서는 무응답이 특정 변수와 밀접하게 관련하여 발생됨을 종종 확인할 수 있다. 무응답 발생여부가 보조변수와 관심변수의 의존여부에 따라 다양한 형태의 무응답 메커니즘을 가정할 수 있다(Rubin, 1976). 본 연구에서는 무응답에 영향을 미치는 유의미한 변수를 탐색하고, 무응답 메커니즘에 따라 구성된 보조정보의 유효성 검토를 위해 통계적 모의실험을 전개한다.

2. 기초자료 분석

2.1 사회조사 표본설계

2016년 사회조사는 2010년 인구주택총조사 10% 표본조사구와 최근까지 신축된 아파트 조사구를 더하여 4천9백 조사구를 표본추출틀(sampling frame)로 구축하여 표본설계를 진행했다. 지역별로 추정하기 위해 7대 특광역시, 세종과 9개 도지역에 대해서는 동부와 읍면부로 구분하여 27개 지역으로 층화(stratification) 작업을 실시했다. 27개 층별(stratum)로는 주택유형, 농가, 유배우, 1인가구, 60세인구, 자가 등의 비율을 분류지표로 구성하여 내재적 층화(implicit stratification)를 함으로써 층화효과를 높일 수 있었다(통계청, 2016b).

표본규모는 시도별로 공표수준의 목표오차를 감안하여 전국기준으로 변동계수(coefficient of variation, CV)가 2% 이내가 되도록 하지만, 역시 지역별로는 변동계수가 10% 이내가 되도록 표본을 배정했다. 표본추출은 지역별로 표본규모에 맞게 층화 2단계 집락추출 방법을 사용하고 있다. 먼저 1차 추출단위인 조사구 1,548개를 확률비례계통추출에 의해 선정한 후, 조사구내에서 2차 추출단위인 가구는 랜덤 추출을 했다.

2016년부터는 조사현장에서 표본 대체방법을 사용하지 않는 대신에, 전년도 관심변수 변동성 및 응답률 등 정도(precision)에 미치는 요인을 고려하여 표본크기는 전년에 비해 36% 증가한 25,233가구를 최종 방문가구로 채택하여 조사를 진행했다(Statistics Canada, 2003).

〈표 2.1〉 2015년 사회조사 표본규모 및 2016년 사회조사 표본규모 결정

표본 조사구수	2015년				2016년	
	표본 가구수	응답 가구수	대체율	변동계수	최종표본 규모	예상 변동계수
1,548	18,576	12,857	30.8	1.5	25,233	1.6

2.2 무응답률 및 분석

2016년 사회조사는 빈집 등 부적격을 제외한 적격 표본 22,642가구 중 전체 무응답률은 18.6%로 나타났다. 무응답을 분석할 수 있는 자료는 직접 조사방식을 통한 파라데이터(paradata) 또는 간접 방식인 행정자료 수집 등을 통해 얻을 수 있다. 파라데이터는 자료수집 과정, 공표, 유지, 보관 등 조사 전역에 걸친 광범위한 자료이다. 다만 사회조사의 파라데이터 수집은 아직 초기단계라 할 수 있으며, 조사목적에 맞는 핵심적인 보조정보를 얻기에는 다소 제약이 따른다. 조사원들이 파라데이터를 일종의 조사항목 일부분으로 간주하여 조사 부담을 갖고 있으며, 특히 응답거부 및 비접촉 가구에 무응답 정보를 얻기는 거의 불가능하며, 조사원을 전적으로 의존한 내용은 신뢰성이 낮다는 주장을 펴기도 한다(Kreuter 등, 2010). 이런 논쟁에도 파라데이터를 통해 조사관리 측면, 조사설계 측면, 총오차 측면 및 통계품질 측면 등 다양한 방면에서 활용성을 언급하면서 자료수집의 당위성을 강조하고 있다(김정섭과 임경은, 2010년).

2016년 수집된 파라미터는 연건평, 아파트 평형(소형 중형 대형), 무응답 사유, 주택유형, 방문일자, 방문시간, 지역(동부 읍면부) 등에 대해 응답가구 및 무응답가구에 대한 정보를 동시에 포함하고 있다. 이로부터 무응답률 18.6%는 다시 응답 거부 12.3%, 비접촉 5.7%, 기타 무응답 0.6%로 구분되는데, 최근 대부분 조사에서 나타나는 것처럼 응답 거부가 무응답의 높은 비중을 차지하는 현상을 볼 수 있다(Atrostic 등 2001, Brick and Williams 2013).

지역별로는 동부 무응답률은 21.2%를 기록한 반면 읍면부에서는 절반 수준인 11.1%으로 나타나, 동부의 거주 가구가 조사에 비협조적임을 알 수 있다. 또한 동부지역은 맞벌이 증가 등의 이유로 주간 부재율이 높아 조사기간 동안 접촉이 불가능한 비접촉률도 상대적으로 높았다. 주택유형별로는 아파트 가구의 무응답률은 22.3%로 단독주택 거주 가구 13.7%에 비해 무응답률이 집중되는 특징을 보인다. 특히 아파트 가구의 응답거부는 16.0%로 단독주택 가구의 7.6% 보다 2배 이상 높아 아파트 가구가 조사에 보다 비협조적임이 확인됐다. 아파트 가구의 소득이 단독주택 가구에 비해 1.7배 높게 나타나고 있어, 소득이 높은 가구의 무응답 성향이 높음을 알 수 있다(통계청, 2017년 가계금융복지조사). 최종 방문횟수별로 무응답을 보면, 방문횟수가 높아질수록 예상하듯 응답을 얻기가 더욱 어려워짐도 확인할 수 있다.

〈표 2.2〉 가구 특성별 응답 및 무응답 가구수 및 구성비(%)

	전체	응답		무응답		응답 거부		비접촉		기타		
		수	%	수	%	수	%	수	%	수	%	
전국	22,642	18,422	81.4	4,220	18.6	2,788	12.3	1,287	5.7	145	0.6	
지역	동부	16,973	13,380	78.8	3,593	21.2	2,434	14.3	1,067	6.3	92	0.5
	읍면부	5,669	5,042	88.9	627	11.1	354	6.2	220	3.9	53	0.9
주택 유형	단독주택	9,223	7,963	86.3	1,260	13.7	699	7.6	482	5.2	79	0.9
	아파트	10,557	8,201	77.7	2,356	22.3	1,686	16.0	629	6.0	41	0.4
	기타주택	2,862	2,258	78.9	604	21.1	403	14.1	176	6.1	25	0.9
방문 횟수	1회	4,784	4,627	96.7	157	3.3	71	1.5	43	0.9	43	0.9
	2회	5,436	5,264	96.8	172	3.2	109	2.0	48	0.9	15	0.3
	3회	5,602	4,500	80.3	1,102	19.7	734	13.1	322	5.7	46	0.8
	4회+	6,820	4,031	59.1	2,789	40.9	1,874	27.5	874	12.8	41	0.6

무응답 사유에 대한 세부적 구분은 무응답 처리를 위한 방법 및 효과와 직접 관련 되어 중요하다(Lin and Schaeffer, 1995; Steele and Durrant, 2011). 비접촉 비율은 30.5%로 가장 높고, 사생활 노출기피 23.4%, 바쁘고 귀찮아서 19.0% 순으로 나타났다. 비접촉률이 높은 이유는 조사기간이 짧게 진행되어 조사기간 내에 방문할 수 있는 방문횟수가 제한적인데 있다. 또한, 개인정보 노출을 꺼리고 사생활 보호의식이 강화되는 등의 최근 사회현상이 반영되어 사생활 노출기피가 높은 비율을 차지했다.

〈표 2.3〉 무응답 사유별 가구수 및 구성비(%)

		전체	동부	읍면부	전체	동부	읍면부
합계		4,220	3,593	627	100.0	100.0	100.0
응답 불응 (refuse)	사생활 노출기피	989	865	124	23.4	24.1	19.8
	가정 사정	74	62	12	1.8	1.7	1.9
	바쁘고 귀찮아서	801	689	112	19.0	19.2	17.9
	정부불신	289	250	39	6.8	7.0	6.2
	면접회피	635	568	67	15.0	15.8	10.7
비접촉(non-contact)		1,287	1,067	220	30.5	29.7	35.1
기타(incapacity)		145	92	53	3.4	2.6	8.5

2.3 무응답 편향 탐색

무응답 편향을 탐색하기 위해 무응답률과 부차 집단간 구성비의 차이를 이용하여 간접적으로 편향은 다음의 식처럼 표현할 수 있다(Groves, 1989; 김서영과 안다영, 2010).

$B(\bar{y}_r) = (\frac{m}{n})(\bar{y}_r - \bar{y}_m)$. 여기서 n 은 표본 가구, r 은 표본에서 응답 가구, m 은 무응답 가구를 의미한다. 파라미터 정보로부터 지역, 주택유형 및 방문횟수 등 부차 그룹에서 응답과 무응답 비율의 차이가 유의미했으며, 이들의 유의성 검정 결과 응답가구와 무응답 가구 특성이 명확히 달랐다. 동부 지역, 아파트, 방문횟수가 높은 경우에 무응답 가구의 비중이 응답가구보다 높게 나타나는데, 이 변수들은 무응답 조정에 활용한다.

〈표 2.4〉 무응답 유형별 차이 및 편향

		응답가구(A)		무응답가구(B)		차이(C)	편향(D)	χ^2 Value (Pr > Chisq)
		비중		비중		(= A - B)	(= $\frac{m}{n} * C$)	
지역	동부	13,380	0.73	3,593	0.85	-0.13	-2.33	286.40 (<0.001)
	읍면부	5,042	0.27	627	0.15	0.13	2.33	
주택유형	단독주택	7,963	0.43	1,260	0.30	0.13	2.48	256.13 (<0.001)
	아파트	8,201	0.45	2,356	0.56	-0.11	-2.11	
	기타주택	2,258	0.12	604	0.14	-0.02	-0.37	
방문횟수	1회	4,627	0.25	157	0.04	0.21	3.99	3,834.00 (<0.001)
	2회	5,264	0.29	172	0.04	0.25	4.57	
	3회	4,500	0.24	1,102	0.26	-0.02	-0.32	
	4회+	4,031	0.22	2,789	0.66	-0.44	-8.24	
합계		18,422	1.00	4,220	1.00	-	-	-

오즈비(odds ratio)는 파라미터를 보조정보로 로지스틱 회귀분석의 응답성향을 통해 얻어진다. 오즈비가 1보다 크면 기준이 되는 그룹보다 응답성향이 높고 반면 1보다 작으면 기준그룹 보다 응답성향이 낮음을 나타낸다(SIPP, 2017). 동부의 거주 가구는 읍면부 가구에 비해 응답성향이 8.9% 낮고, 아파트 가구는 기타주택에 비해 응답성향이 11.6% 낮게

나타났다. 반면, 최종 방문이 첫 회로 끝나는 가구는 최종 방문이 4회 이상인 가구에 비해 응답 가능성이 20배 이상 높게 나타나 대조를 보였다. 오즈비로부터 지역, 주택유형 및 방문횟수가 무응답에 영향력 있는 변수로 파악됐다.

〈표 2.5〉 로지스틱 회귀분석을 통한 오즈비(odds ratios)

	추정량	95% Wald 신뢰구간	
동부 vs 읍면부	0.911	0.822	1.011
단독주택 vs 기타주택	1.098	0.975	1.237
아파트 vs 기타주택	0.884	0.792	0.986
방문 1회 vs 방문 4회+	18.958	16.026	22.427
방문 2회 vs 방문 4회+	20.327	17.323	23.853
방문 3회 vs 방문 4회+	2.773	2.554	3.010

응답여부에 영향 미치는 추가 정보는 표본추출틀을 사용한 행정자료로부터 얻은 확보가 매우 중요하다. 최근 전 세계적으로 조사비용 절감 및 통계조사 응답부담 감소 등 통계생산 효율성이 강조됨에 따라 행정자료 활용이 점차 보편화되고 있다(UNECE, 2011). 통계청에서는 2015년부터 모든 가구 및 인구를 대상으로 행정자료를 정기적으로 수집하여 매년 센서스 전수 집계결과로 공표하고 있다. 2016년에는 성별, 연령, 성명, 국적, 입국시기, 가구구분, 거처종류, 연면적, 대지면적, 건축연도 등 13개 항목을 포함하고 있어, 이들을 주요 보조정보로 활용할 수 있다(통계청, 2017).

하지만 당시 표본추출틀 노후화로 표본 정보에 대한 내재된 불확실성과 행정자료 이용에는 시기적인 제약 등이 따랐다. 이를 극복하는 한 방법으로 파라미터 정보인 지역, 주택유형, 방문횟수를 반영하여 응답성향을 산정한 후, 응답가구만으로 가구특성별 평균 응답성향의 분포를 재구성하는 방법을 적용했다(〈표 2.6〉).

응답여부에 영향을 미치는 변수는 파라미터 정보를 포함하여 가구원수, 점유형태, 가구주 연령과 가구 소득 변수 등이 탐색됐다. 여기서 가구원수, 점유형태, 성별, 연령 등은 행정자료로 이미 수집되어 있어 근시일내 활용이 가능하다.

이처럼 응답여부에 영향을 미치는 변수들을 무응답 조정 변수로 구성하는 것이 바람직한데, 이는 무응답 편향을 효과적으로 제거할 수 있기 때문이다. 관련성이 약한 변수를 포함한 경우는 편향 축소에 그다지 도움이 되지 않고 분포의 안정성을 떨어뜨려 정확성을 저해할 수도 있다(Guillaume, 2013).

〈표 2.6〉 응답가구 특성별 평균 응답성향

가구특성		응답성향	F Value (Pr > F)	가구특성		응답성향	F Value (Pr > F)
지역	동부	0.82	1228.86 (<.0001)	가구원수	1인	0.86	123.00 (<.0001)
	읍면부	0.91			2인	0.86	
주택유형	단독주택	0.88	3인		0.82		
	아파트	0.81	4인+		0.82		
	기타주택	0.82	가구주	남자	0.84	11.73 (.0006)	
방문횟수	1회	0.98	성별	여자	0.85		
	2회	0.94		가구주 연령	20-29세	0.81	229.44 (<.0001)
	3회	0.82	30-39세		0.81		
	4회+	0.59	40-49세		0.82		
점유형태	자가	0.85	50-59세		0.83		
	비자가	0.83	60세이상	0.89			

3. 모의실험(simulation)

무응답에 대한 정확한 정보를 조사 현장으로부터 파악하는 데는 현실적으로 어려움이 있고 확보된 제한적인 정보만으로 특정변수에 대한 무응답 발생여부(missingness)를 결정 짓기는 쉽지 않다. 따라서 무응답 자료를 처리하기 위해서는 무응답 자료가 어떤 메커니즘을 따르는지를 살펴 볼 필요가 있다. 본 연구는 무응답 메커니즘 가정 하에 통계적 모의실험을 통해 실제 무응답 편향에 대한 처리효과를 확인할 것이다.

3.1 모집단 및 표본셋 구성

모의실험을 위해 2016년 사회조사에 최종적으로 응답을 완료한 가구를 준모집단(pseudo-population)으로 정의했다. 즉, 2016년 조사를 마친 18,422가구를 표본추출틀(sampling frame)로 간주하여 27개 지역으로 층화한 후 지역별로 배분된 표본규모에 따라 총 3,000가구를 추출했다. 모의실험에서 표본규모가 큰 경우에는 무응답에 대한 조정 없이 사후층화 보정만으로도 그 효과를 충족할 수 있기 때문에, 무응답 처리효과를 명확히 분리하기 위해 전국 단위 공표수준인 3,000가구를 실험대상으로 결정했다. 이는 오차의 한계, 응답률 및 설계효과를 고려한 결과이다(Statistics Canada, 2003). 또한 2016년 조사에서 발생된 표본의 무응답(nonresponse) 및 비포함률(noncoverage)을 보완하여 최대한 모집단과 유사한 구조를 가지도록 표본 추출을 시도했다. 이를 위해 27개 지역별로 주택유형, 혼인상태, 1인가구, 60세이상 비중 등이 모집단과 유사하도록 계통추출 방법을 통해 300개 표본셋을 구성하여 모의실험을 전개한다.

3.2 무응답 발생 및 유의성 검토

2016년 사회조사 무응답에 관한 모의실험을 위해 먼저 빈집과 비거주지 건물 등 부적격을 제외한 무응답률 18.6%인 경우와 과거 표본 대체율인 30.0%인 경우를 고려하여 두 가지 조건에 따라 무응답을 처리했다. 무응답률은 가중값을 적용하지 않는 표본에 대한 비가중(unweighted) 무응답률을 사용하고 있다. 당초 무응답률에 따른 무응답 처리의 차이점 비교를 시도했으나, 무응답률이 18.6%인 경우에도 무응답률이 30%인 경우와 메카니즘에 따른 무응답 처리결과가 유사하게 나타나 본 연구에서는 무응답률이 30.0%인 경우만을 최종적으로 다룬다.

무응답은 주요 관심 변수의 무응답 발생여부(missingness)가 다른 변수와 어떻게 관련되는냐에 따라 여러 가지 무응답 메커니즘(mechanism)을 가정할 수 있다(Rubin, 1976). 앞서 추출된 300개 표본셋은 각각 아래의 6가지 메커니즘 가정을 기초로 진행된다.

완전임의결측(MCAR: Missing Completely At Random) 자료는 무응답 발생여부가 보조변수들을 의지하지 않고 모든 가구의 응답확률은 동일하다고 가정했다.

① MCAR 자료는 특정 변수와 무관하게 무응답 30%가 랜덤하게 발생한다.

임의결측(MAR: Missing At Random) 자료는 무응답 발생여부가 보조변수에 의존함에 따라, 응답확률은 보조변수의 값에 따라 달라진다. MAR1 무응답 발생은 파라미터로부터 유의미한 변수에 의해 발생하는 것으로 설정했으며, MAR2 무응답 발생은 행정자료로부터 응답여부에 직접 영향을 미치는 변수에 의해 발생하는 것으로 설정했다.

② MAR1 자료는 지역, 주택유형 및 방문횟수에 의존하여 무응답이 발생한다.

③ MAR2 자료는 가구원수 및 가구주 연령에 의존하여 무응답이 발생한다.

비임의결측(NMAR: Not Missing At Random) 자료는 무응답 발생여부가 보조변수와 조사변수의 비관측된 값(unobserved values)에 동시에 의존한다. 이런 이유로 무시할 수 없는(non-ignorable) 비임의결측(NMAR) 모델은 정확한 무응답 구조를 파악하는데 근본적인 어려움이 따르며(Rubin, 1976; Little and Rubin, 1987; Groves, 2006), 관측된 자료를 기반으로 한 임의결측(MAR)과 구별이 쉽지 않다(David 등, 1986; Molenberghs 등, 2008). 다만, 무응답 발생여부는 조사의 주제 및 측정항목 내용과 밀접한 관련이 있다고 알려져 있다(Bover, 2011). 무응답 발생여부가 조사주제에 종속된다는 가정 하에, 비임의결측(NMAR) 자료는 조사변수(survey variables)를 활용하여, 무응답률이 전체적으로 30%가 되도록 설정했다.

④ NMAR1 자료는 소득에 따라 무응답이 발생한다.

⑤ NMAR2 자료는 소득 및 방문횟수에 따라 무응답이 발생한다.

⑥ NMAR3 자료는 주관적 만족감에 따라 무응답이 발생한다.

〈표 3.1〉 무응답 메커니즘별 무응답률 및 유의성 검정(카이제곱 검정)

(단위: %, p -값 > 0.05(건수/300회))

		MCAR		MAR1		MAR2		NMAR1		NMAR2		NMAR3	
지역	동부	30.0	19	31.7	300	30.6	61	30.5	82	30.5	82	29.9	18
	읍면부	30.0		24.3		28.4		27.9		27.9		29.7	
주택유형	단독주택	30.0		26.4	300	28.3		27.4		28.1		29.6	
	아파트	30.0	14	32.2	300	31.2	80	31.8	175	31.0	89	30.1	12
	기타주택	30.0		31.2		30.6		29.5		30.4		29.8	
방문횟수	1회	30.1		18.1	300	28.7		28.1		23.1	300	29.9	
	2회	29.9	18	21.9		29.7	41	29.6	42	29.0		29.6	19
	3회	30.0		31.1		30.7		30.5		32.5		29.8	
	4회+	30.1		51.0		31.3		31.2		35.2		30.0	
점유형태	자가	29.9		29.3	38	29.4		30.3	28	29.8		29.9	12
	비자가	30.1	10	30.9		31.0	32	29.1		29.9	13	29.8	
가구원수	1인	30.0		28.5		28.5	300	26.0		28.5		29.5	
	2인	30.0	21	28.5	89	27.1		28.8	273	28.7	72	29.7	21
	3인	29.9		31.5		32.8		32.1		30.9		30.1	
	4인+	30.1		32.0		33.4		33.8		32.0		30.1	
가구주 연령	20-29세	29.8		32.4		32.9		27.6		29.9		30.5	
	30-39세	30.2		32.4		34.3		32.2		31.0		30.2	
	40-49세	29.9	9	31.7	204	33.6	300	32.4	244	31.0	81	29.9	16
	50-59세	30.0		30.8		31.1		31.3		31.2		29.9	
	60세이상	30.0		26.2		24.2		26.1		27.5		29.5	
가구 소득수준	<100만원	30.1		25.9		26.1		21.5		27.0		29.2	
	100-200	29.9		29.4		29.0		27.0		28.5		29.4	
	200-300	30.0		30.4		31.0		30.6		28.9		29.7	
	300-400	30.0	12	31.5	156	31.9	150	33.3	300	29.3	300	30.0	22
	400-500	29.9		31.8		32.5		35.5		29.0		30.6	
	500-600	30.1		31.1		32.4		37.2		38.4		30.5	
	600-700	30.2		34.1		33.2		38.4		39.0		31.1	
≥700만원	30.2		34.5		32.6		40.1		39.0		31.4		
가구주 학력	초졸이하	30.1		25.2		25.1		24.4		27.3		28.9	
	중졸	29.8	10	28.6	209	27.3	247	27.8	275	28.4	98	29.4	25
	고졸	30.0		30.3		31.0		30.1		29.8		29.8	
	대졸이상	30.0		32.2		32.4		33.0		31.6		30.5	
주관적 만족감	매우높음	30.2		31.4		31.7		32.3		31.8		34.6	
	높음	30.0		30.1		30.6		31.2		30.3		31.4	
	보통	30.0	12	29.5	13	29.6	25	29.3	83	29.4	30	28.9	300
	낮음	29.9		29.3		29.3		27.3		28.9		25.0	
	매우낮음	30.1		30.0		29.1		27.3		28.8		32.4	

〈표 3.1〉은 무응답 메커니즘별로 가구특성별 응답여부에 관한 정보를 얻고자 지역, 주택유형 등 300개 표본셋에 대해 각각 카이제곱 검정을 실시한 후, 누적빈도로 통계적 유의성을 측정했다. 다시 말해, 가구특성별로 유의한 빈도수가 200회 이상인 경우에 음영으로 표시하여 통계적으로 응답여부의 유의미한 차이를 나타냈다. 또한 메커니즘별로 무응답 발생을 위해 활용된 변수는 별도 테두리를 설정하여 구분했다.

완전임의결측(MCAR) 자료는 변수들과 관련 없이 랜덤한 조건으로 무응답이 발생했으며, 그 결과 9가지 모든 가구특성 변수에서 응답성향이 유의미한 차이를 발견할 수 없었다. 즉, 지역, 주택유형 등 모든 가구특성 변수에서 30% 수준의 동일한 무응답률을 보이

고 있다. 이 경우에는 무응답률 자체가 추정값에 위협적이지 않을 것으로 판단할 수 있다 (Curtin 등, 2000).

임의결측(MAR)에서 MAR1 자료는 지역, 주택유형 및 방문횟수 등 3개 변수에 따라 무응답이 발생했으며, 이들이 반영된 특성들의 부차그룹 응답성향은 통계적으로 유의미한 차이를 보인다. 이와 함께 가구주 연령과 학력에서도 응답성향이 유의미한 차이를 보여, 앞의 변수들과 가구주 연령 및 학력 간에 높은 상관관계가 존재함을 알 수 있다. MAR2 자료는 가구원수와 가구주 연령에 따라 무응답이 발생되어, 이들 변수에서 유의미한 차이를 보이며 가구주 학력에도 유의미한 결과가 도출됐다.

비임의결측(NMAR)에서는 NMAR1 자료는 소득에 따라 무응답이 발생하여, 소득수준별 응답성향은 유의미한 차이를 보이며, 동시에 가구원수, 가구주 연령 및 학력 등도 응답성향에서 유의성을 보였다. 즉, 소득은 가구원수, 가구주 연령 및 학력 등과 관련성이 높음을 엿볼 수 있다. NMAR2 자료는 소득과 방문횟수를 이용하여 무응답을 발생했으며, 이들 변수에서 통계적으로 유의미한 차이가 나타났다. NMAR3 자료는 주관적 만족감을 이용하여 무응답이 발생됐고, 이 변수에서만 유의미한 결과를 보였다.

3.3 무응답 처리

표본조사에서 단위 무응답 처리를 위해 보편적으로 가중값 조정법이 사용된다(Groves, 2006). 보조변수(auxiliary variables)의 주변분포에 맞게 조정하는 레이킹 비 조정법(raking ratio adjustment) 및 캘리브레이션 보정방법(calibration), 보조변수의 주변분포 또는 결합분포에 맞게 조정이 가능하도록 발전한 일반화 회귀추정(GREG) 방법이 있다(Deming and Stephan, 1940; Deville and Sarndal, 1992; Bethlehem and Keller, 1987; Bethlehem 등, 2011). 또한 보조변수의 교차표를 기초하여 무응답층을 결정하는 CHAID(Chi-squared automatic interaction detection) 알고리즘과 보조변수를 활용한 응답성향을 통해 가중값을 조정하는 로지스틱 분석이 있다(Oh and Scheuren, 1983; Little, 1992). 다만 많은 방법들이 상당히 비슷하고 방법 간에 상당한 관련이 있어, 특별한 방법을 선정하는 것보다는 관련된 보조변수 선정이 더욱 유의미하다 할 수 있다(Kalton and Flores-Cervantes, 2003).

통계청 조사별 선행연구로는, 가계동향조사(2016년)는 시도, 가구구분, 가구원수 및 주택유형 등을 보조변수로 활용하고 있으며, CHAID 방법을 통해 50개 무응답층을 구성한 후 무응답 조정을 하고 있다. 가계금융·복지조사(2011년)는 가계동향조사와 비슷하게 지역, 혼인상태, 입주형태, 가구주 산업 및 직업 등을 보조변수로 CHAID 모형을 통해 8개 무응답층을 구성하여 무응답을 처리하고 있다. 외국인고용조사(2016년)는 지역, 국적 및 체류자격 등을 보조변수로 하여 로지스틱 회귀방법을 사용하고 있으며, 경제활동인구조사는 주택유형, 소유형태, 지역, 1인가구 여부, 농가여부 등을 보조변수로 한 로지스틱 회귀방법을 사용한다(통계청, 2016c). 본 연구에서는 가장 널리 사용되는 응답성향을 이용한 가중값 조정법인 로지스틱 회귀방법으로 접근하여 진행한다.

무응답 조정을 위한 보조정보는 직접 파라데이터를 활용하거나 수집된 행정자료를 활용할 수 있음을 앞서 살펴보았다. 먼저, 파라데이터로부터 응답성향에 유의미한 정보인

지역, 주택유형, 방문횟수로 이루어진 3개 변수로 보조정보를 구성한다. 다음으로, 행정자료로부터 응답성향에 영향을 미치는 점유형태, 가구원수, 가구주 성별 및 연령 등 4개 변수와 파라미터 3개변수를 더해 7개 변수를 다른 보조정보로 구성하여 통합자료라고 명명한다(〈표 3.2〉).

〈표 3.2〉 무응답 메커니즘 및 보조정보에 활용된 변수

무응답 메커니즘		보조정보
① MCAR	-	① (파라미터) 지역, 주택유형, 방문횟수
② MAR1	지역, 주택유형, 방문횟수	
③ MAR2	가구원수, 가구주 연령	
④ NMAR1	소득	② (통합자료) 지역, 주택유형, 방문횟수, 점유형태, 가구원수, 가구주 성별, 가구주 연령
⑤ NMAR2	소득, 방문횟수	
⑥ NMAR3	주관적 만족감	

4. 모의실험 결과

4.1 평가 내용

모의실험 결과를 측정하기 위해 2016년 사회조사의 보건, 교육, 안전, 가족, 환경 등 조사부문 별로 모든 항목을 분석하였으나, 본 연구는 주관적 만족감, 건강평가, 흡연여부, 음주여부, 학교교육 효과, 결혼에 대한 견해, 이혼에 대한 견해 및 가족관계 만족도 등 9개 주요항목에 초점을 맞추어 전개한다. 또한 사회조사의 만족도와 견해 등은 주로 5점 척도로 구성되나, 지면을 고려하여 ‘매우만족’과 ‘만족’을 ‘만족’으로 묶는 등 3점 척도로 변환했다.

무응답 편향에 대한 평가 및 비교를 위해, 무응답이 없는 원표본(full-sample 또는 no unit missing)과 무응답이 포함된 표본셋을 구분하여 모수 추정 절차를 달리 진행한다.

원표본의 가중값은 설계가중값 및 사후층화 가중값 등 2단계로 가중값을 구성하여 산출된 추정량을 참값(true value) θ_0 로 정의했다. 원표본은 무응답이 없는 각각 표본셋을 총합하여 90,000가구로 구성하여 한 번에 가중값을 적용하는 방법과 각 표본셋 3,000가구에 가중값을 적용하여 평균을 사용하는 방법이 있다. 어느 방법을 사용하던 동일한 결과를 갖지만 총합을 이용한 경우, 계산을 단순화 한다는 장점이 있다. 설계가중값은 27개 각 지역층에 맞게 가구 추출률의 역수를 사용했다. 사후층화 가중값으로 통계청 가구추계에서 제공된 27개 지역에 4가지 가구원수 유형을 곱한 2차원 분포를 사용하여 전체적으로는 108개 사후층(post stratification)에 벤치마킹을 적용했다.

무응답을 가지고 있는 표본셋에 대해서는 설계가중값, 무응답 조정 및 사후층화 조정 등 3단계로 가중값 작업을 적용하여 300개 셋의 추정량을 산출했다. 설계가중값은 27개 지역층으로부터 직접 산출한 다음, 무응답 단계에서는 성향점수를 이용한 조정계수를 적용하여 무응답 조정 가중값이 작성되며, 무응답 조정가중값은 다시 108개 외부 벤치마크(external benchmarks)인 가구추계에 맞게 사후보정을 했다.

4.2 평가 척도

무응답 처리결과에 대한 평가를 위해 무응답 편향(bias), 상대편향(relative bias), 평균오차 제곱근(RMSE), 설계효과(design effect due to weighting) 등을 척도로 사용하고, 마지막으로 보조정보 간 모형비교(model comparisons)로 유의성을 점검한다. 편향은 원표본에 가중값을 적용한 추정량(참값)과 각 표본셋에서 산출된 추정량과의 차이로 식(4.1)과 같다. $i(=1,2,3,4,5,6)$ 는 메커니즘 방법별 자료를 의미하며, $r(=1,\dots,300)$ 은 300개 표본셋에 해당한다. 상대편향은 편향을 참값으로 나누어 백분위(percentage)로 표준화한 척도로서 식 (4.2)처럼 계산된다. 평균제곱오차 제곱근은 무응답 조정된 추정량의 편향과 변동성을 동시에 측정하는 척도로 식 (4.3)처럼 정의한다. 다만 처리효과를 측정하는데 편향에 더욱 높은 비중을 주고 있다(Groves, 2006; 이승희, 2010).

설계효과는 무응답 조정계수를 적용한 후, 추정값의 분산이 커지는 효과로 정의할 수 있다. 식 (4.4). 무응답 조정 후 편향은 감소하나, 변동성이 증가되어 정도(precision)는 낮아질 수밖에 없다. 설계효과는 정도의 손실을 측정하는 유용한 척도로 사용된다(Kish, 1992). 모형간 비교는 구성된 두 가지 보조정보에 대한 우도비 검정(likelihood-ratio test)을 통해 모형의 유의성을 확인한다.

$$\text{편향(Bias)} : \frac{1}{300} \sum_{r=1}^{300} (\hat{\theta}_{ir} - \hat{\theta}_0), \quad \hat{\theta}_0 = \text{참값} \quad (4.1)$$

$$\text{상대편향(Relative Bias)} : \frac{(\bar{\theta}_i - \hat{\theta}_0)}{\hat{\theta}_0} \times 100, \quad \bar{\theta}_i = \frac{1}{300} \sum_{r=1}^{300} \hat{\theta}_{ir} \quad (4.2)$$

$$\text{평균제곱오차 제곱근(RMSE)} : \sqrt{\frac{1}{299} \sum_{r=1}^{300} (\hat{\theta}_{ir} - \hat{\theta}_0)^2} \quad (4.3)$$

$$\text{설계효과(Design Effect due to weighting)} : \frac{1}{300} \sum_{r=1}^{300} (1 + CV(w_r)^2) \quad (4.4)$$

모형 간 비교(Model comparisons)

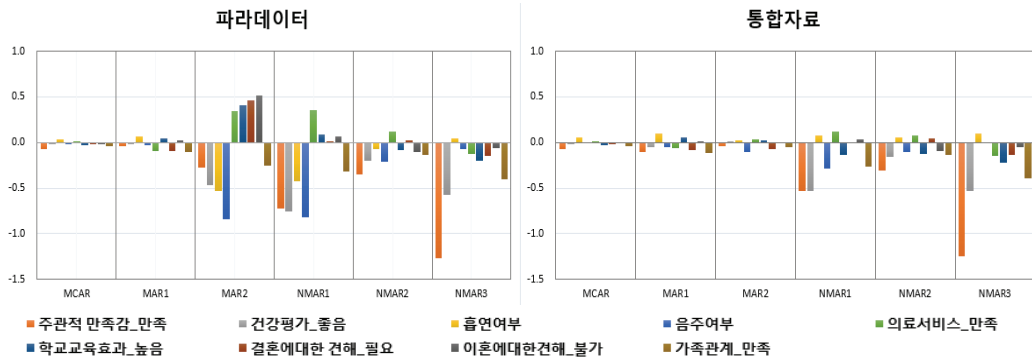
4.3 실험 결과

편향(bias)에 대해 처리결과는 대각선과 음영으로 설명된다(〈표 4.1〉). 대각선은 무응답 처리 후 편향이 감소되어 처리효과가 발생한 반면, 음영은 처리 후에도 편향이 일정 부분 남아 있어 처리효과가 낮은 것으로 평가할 수 있다. 편향의 감소효과 측정기준으로 주요변수의 평균 표준오차(standard error, SE)인 0.6%를 임계점(threshold)으로 사용한다. 완전임의결측(MCAR)의 경우는 당초 편향이 크지 않아 무응답 처리 후에도 큰 변화를 보이지 않지만 처리효과의 성능이 가장 우수하다. 파라미터를 보조정보를 사용하여 무응답을 조정한 경우, 무응답 발생여부와 직접 관련된 변수를 포함하고 있는 MAR1 자료와 NMAR2 자료는 상대적으로 처리효과가 높게 나타났다. 흥미로운 사실은 통합자료의 경우, 무응답 발생여부와 관련되거나 상관관계가 높은 대리 보조변수를 갖춤으로써, 무응답을 처리한 결과는 매우 성공적이다(Paul, 2012). 이로써 무응답 발생여부가 보조변수를 의존하거나 소득에 대한 상관관계가 높은 보조변수가 있는 경우에 비임의결측(NMAR) 자료도 성공적인 처리결과를 보였다. 다만, NMAR3 자료는 보조변수와 연관성이 없다는 점에서 처리효과가 낮게 나타났다.

〈표 4.1〉 추정값의 편향 비교

		파라미터						통합자료					
		MCAR	MAR1	MAR2	NMAR1	NMAR2	NMAR3	MCAR	MAR1	MAR2	NMAR1	NMAR2	NMAR3
주관적 만족감	만족	-0.07	-0.04	-0.28	-0.73	-0.35	-1.27	-0.07	-0.10	-0.04	-0.53	-0.31	-1.25
	보통	0.06	0.06	0.24	0.33	0.18	0.60	0.06	0.09	0.02	0.18	0.15	0.60
	불만족	-0.01	-0.02	0.04	0.40	0.17	0.67	0.01	0.01	0.03	0.36	0.17	0.66
건강평가	좋음	-0.02	-0.02	-0.46	-0.76	-0.20	-0.57	-0.02	-0.06	0.00	-0.53	-0.16	-0.53
	보통	0.05	0.02	-0.18	-0.05	0.06	0.27	0.05	0.00	-0.09	0.11	0.09	0.28
	나쁨	-0.03	0.00	0.64	0.81	0.15	0.30	-0.03	0.05	0.09	0.42	0.06	0.26
흡연여부		0.04	0.07	-0.53	-0.43	-0.07	0.04	0.06	0.10	0.02	0.08	0.05	0.10
음주여부		-0.02	-0.03	-0.84	-0.82	-0.21	-0.07	-0.01	-0.05	-0.10	-0.29	-0.11	-0.01
의료서비스 만족도	만족	0.01	-0.09	0.34	0.35	0.12	-0.12	0.01	-0.06	-0.03	0.12	0.08	-0.14
	보통	0.01	0.08	-0.30	-0.32	-0.06	0.12	0.00	0.06	-0.05	-0.11	-0.02	0.14
	불만족	-0.01	0.01	-0.05	-0.03	-0.06	0.01	-0.01	0.00	0.02	-0.01	-0.06	0.01
학교교육 효과	높음	-0.03	0.05	0.41	0.09	-0.09	-0.20	-0.03	0.06	0.03	-0.13	-0.13	-0.22
	보통	0.02	-0.09	-0.20	-0.10	0.01	0.14	0.02	-0.10	-0.02	0.02	0.03	0.15
	낮음	-0.01	0.04	-0.21	0.01	0.08	0.07	0.01	0.04	-0.01	0.11	0.10	0.07
결혼에 대한 견해	필요	-0.02	-0.09	0.46	0.01	0.03	-0.14	-0.01	-0.09	-0.07	-0.01	0.04	-0.13
	중립	0.00	0.12	-0.41	-0.09	-0.03	0.09	0.00	0.10	0.07	-0.04	-0.04	0.08
	불필요	0.03	-0.03	-0.04	0.04	-0.01	0.04	0.03	-0.03	-0.02	0.02	-0.01	0.04
이혼에 대한 견해	불가	-0.02	0.02	0.51	0.07	-0.10	-0.06	-0.01	0.01	-0.01	0.04	-0.09	-0.05
	중립	0.00	-0.04	-0.43	-0.12	0.04	0.04	-0.01	-0.07	-0.03	-0.02	0.05	0.04
	가능	0.02	0.02	-0.08	0.05	0.06	0.03	0.01	0.06	0.04	-0.01	0.04	0.02
가족관계 만족도	만족	-0.04	-0.10	-0.25	-0.32	-0.13	-0.41	-0.04	-0.12	-0.05	-0.26	-0.13	-0.39
	보통	0.05	0.08	0.23	0.21	0.09	0.34	0.05	0.10	0.04	0.21	0.10	0.33
	불만족	-0.01	0.02	0.03	0.10	0.04	0.07	-0.01	0.02	0.01	0.05	0.03	0.06

* 무응답 처리 후 대각선은 편향이 0.6% 이내로 축소, 음영은 편향이 0.6% 이상



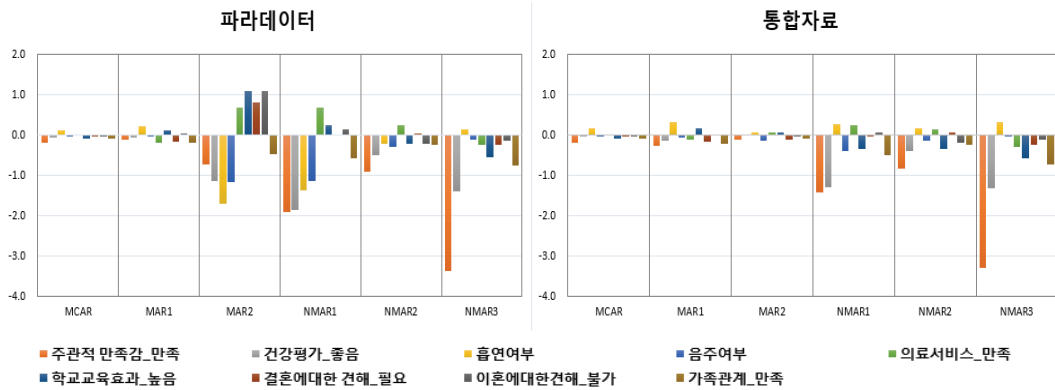
〈그림 4.1〉 추정값의 편향 비교

상대편향(relative bias)은 낮은 비중을 갖는 경우에 과소 되는 편향을 상대적 백분위로 표준화하여 편향에 대한 해석을 명확하게 할 수 있다. 무응답 처리 후 보조정보로 통합자료를 사용한 경우에 처리효과가 보다 우수하며, 무응답 발생과 직접 관련된 MAR1, MAR2 및 NMAR2 자료의 무응답 처리 효과가 높았다. 다만, NMAR1의 무응답 발생에 활용된 소득은 직접 보조정보로 사용되지 않았지만, 상관관계가 높은 가구원수 및 가구주 연령을 대리 보조변수로 사용하여 상대적으로 높은 처리효과를 보였다. NMAR3 자료는 무응답 발생여부가 보조변수와 관련성이 없어 상대적으로 높은 상대편향이 지속됐다. 상대편향 처리효과는 주요변수 평균 변동계수(CV) 1.5%를 임계점으로 평가했다.

〈표 4.2〉 추정값의 상대편향 비교

		파라미터						통합자료					
		MCAR	MAR1	MAR2	NMAR1	NMAR2	NMAR3	MCAR	MAR1	MAR2	NMAR1	NMAR2	NMAR3
주관적 만족감	만족	-0.19	-0.10	-0.74	-1.92	-0.92	-3.36	-0.18	-0.26	-0.11	-1.41	-0.82	-3.29
	보통	0.14	0.14	0.52	0.72	0.39	1.32	0.14	0.19	0.04	0.38	0.32	1.29
	불만족	0.05	-0.15	0.26	2.44	1.05	4.09	0.03	0.08	0.17	2.19	1.02	4.03
건강평가	좋음	-0.05	-0.05	-1.14	-1.86	-0.50	-1.40	-0.04	-0.14	0.00	-1.30	-0.38	-1.31
	보통	0.13	0.05	-0.46	-0.14	0.15	0.68	0.12	0.00	-0.24	0.28	0.24	0.70
	나쁨	-0.14	0.02	3.20	4.04	0.72	1.50	-0.15	0.27	0.46	2.08	0.32	1.27
흡연여부		0.13	0.22	-1.71	-1.38	-0.22	0.13	0.18	0.32	0.06	0.26	0.17	0.33
음주여부		-0.02	-0.04	-1.17	-1.14	-0.29	-0.10	-0.02	-0.07	-0.14	-0.40	-0.15	-0.01
의료서비스 만족도	만족	0.02	-0.18	0.68	0.69	0.24	-0.24	0.03	-0.12	0.07	0.24	0.15	-0.28
	보통	0.01	0.21	-0.73	-0.79	-0.14	0.29	0.00	0.15	-0.13	-0.28	-0.04	0.34
	불만족	-0.16	0.07	-0.55	-0.41	-0.75	0.06	-0.17	0.03	0.24	-0.08	-0.71	0.08
학교교육 효과	높음	-0.09	0.12	1.08	0.24	-0.23	-0.54	-0.08	0.16	0.08	-0.35	-0.33	-0.58
	보통	0.06	0.22	-0.50	-0.26	0.02	0.35	0.06	0.26	-0.04	0.06	0.07	0.37
	낮음	0.04	0.18	-0.96	0.07	0.35	0.30	0.04	0.19	-0.06	0.50	0.46	0.32
결혼에 대한 견해	필요	-0.04	-0.16	0.80	0.02	0.05	-0.25	-0.02	-0.15	-0.12	-0.01	0.07	-0.23
	중립	0.01	0.30	-1.07	-0.23	-0.08	0.23	0.00	0.26	0.19	-0.11	-0.09	0.22
	불필요	1.17	-1.32	-1.72	1.55	-0.36	1.86	1.21	-1.07	-0.63	0.79	-0.58	1.80
이혼에 대한 견해	불가	-0.04	0.04	1.10	0.15	-0.21	-0.14	-0.02	0.03	-0.02	0.08	-0.20	-0.12
	중립	0.00	-0.11	-1.12	-0.31	0.10	0.10	-0.02	-0.18	-0.09	-0.07	0.13	0.10
	가능	0.12	0.16	-0.56	0.31	0.40	0.17	0.09	0.37	0.28	-0.08	0.29	0.10
가족관계 만족도	만족	-0.08	-0.19	-0.47	-0.58	-0.24	-0.75	-0.08	-0.22	-0.09	-0.49	-0.25	-0.73
	보통	0.13	0.20	0.56	0.52	0.22	0.82	0.12	0.23	0.09	0.51	0.25	0.80
	불만족	-0.19	0.41	0.52	2.08	0.86	1.45	-0.18	0.49	0.18	1.08	0.67	1.30

* 무응답 처리 후 대각선은 상대편향이 1.5% 이내로 축소, 음영은 상대편향이 1.5% 이상



〈그림 4.2〉 추정값의 상대편향 비교

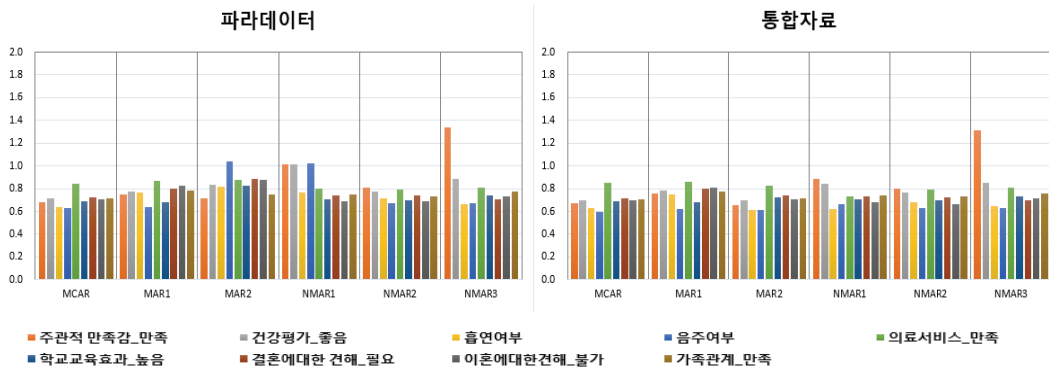
평균오차 제곱근(RMSE)은 무응답 처리 후 추정값의 편향과 변동성을 동시에 측정 가능한 척도이다. 평균오차 제곱근은 정확성(accuracy)과 관련되며, 이는 편향과 관련된 타당성(validity)과 분산과 관련된 정도(precision)로 나눌 수 있으며, 편향과 분산이 모두 작을 때 정확성이 우수하다 할 수 있다(Sukasih 등, 2009; 이승희, 2010).

상대편향의 결과와 마찬가지로 통합자료를 보조정보로 활용한 경우에 상대적으로 적은 값을 가져 적은 편향과 추정값의 분포가 안정된 점을 알 수 있다. 무응답 메커니즘별로 완전임의결측(MCAR)과 임의결측(MAR)에서 처리효과의 우수성을 보이며, 비임의결측(NMAR)은 상대적으로 높은 수치를 보이며, 특히 NMAR3 경우는 더욱 그렇다.

〈표 4.3〉 추정값의 평균오차 제곱근 비교

		파라미터						통합자료					
		MCAR	MAR1	MAR2	NMAR1	NMAR2	NMAR3	MCAR	MAR1	MAR2	NMAR1	NMAR2	NMAR3
주관적 만족감	만족	0.68	0.75	0.71	1.02	0.81	1.34	0.67	0.76	0.66	0.89	0.80	1.31
	보통	0.72	0.73	0.71	0.79	0.76	0.72	0.72	0.74	0.68	0.74	0.76	0.71
	불만족	0.51	0.56	0.51	0.67	0.54	0.73	0.51	0.56	0.52	0.65	0.55	0.72
건강평가	좋음	0.72	0.77	0.83	1.01	0.78	0.89	0.70	0.78	0.70	0.84	0.77	0.86
	보통	0.68	0.74	0.76	0.72	0.79	0.77	0.68	0.75	0.75	0.73	0.79	0.77
	나쁨	0.52	0.56	0.84	0.95	0.56	0.59	0.51	0.56	0.52	0.63	0.54	0.55
흡연여부		0.64	0.77	0.82	0.77	0.71	0.66	0.63	0.75	0.61	0.63	0.68	0.65
음주여부		0.63	0.64	1.04	1.02	0.67	0.67	0.60	0.62	0.61	0.66	0.63	0.63
의료서비스 만족도	만족	0.85	0.87	0.88	0.81	0.80	0.81	0.85	0.86	0.83	0.73	0.79	0.81
	보통	0.77	0.89	0.83	0.78	0.77	0.78	0.78	0.89	0.79	0.72	0.77	0.78
	불만족	0.46	0.54	0.45	0.47	0.42	0.43	0.46	0.54	0.45	0.47	0.42	0.43
학교교육 효과	높음	0.69	0.68	0.83	0.70	0.70	0.74	0.69	0.68	0.72	0.71	0.70	0.73
	보통	0.73	0.77	0.73	0.71	0.72	0.73	0.73	0.77	0.71	0.70	0.72	0.73
	낮음	0.59	0.62	0.65	0.64	0.67	0.60	0.59	0.62	0.62	0.65	0.67	0.60
결혼에 대한 견해	필요	0.72	0.80	0.88	0.74	0.74	0.70	0.72	0.80	0.74	0.74	0.73	0.70
	중립	0.70	0.77	0.86	0.71	0.74	0.68	0.69	0.77	0.74	0.71	0.73	0.68
	불필요	0.23	0.24	0.23	0.24	0.25	0.22	0.23	0.24	0.23	0.24	0.25	0.22
이혼에 대한 견해	불가	0.71	0.83	0.88	0.69	0.69	0.73	0.70	0.81	0.71	0.68	0.67	0.72
	중립	0.71	0.81	0.84	0.69	0.71	0.67	0.70	0.80	0.73	0.66	0.70	0.67
	가능	0.51	0.58	0.51	0.51	0.55	0.48	0.51	0.59	0.50	0.52	0.55	0.48
가족관계 만족도	만족	0.71	0.78	0.75	0.75	0.73	0.77	0.71	0.77	0.71	0.74	0.73	0.76
	보통	0.72	0.78	0.74	0.73	0.70	0.74	0.72	0.77	0.71	0.73	0.70	0.73
	불만족	0.30	0.35	0.33	0.30	0.34	0.32	0.30	0.35	0.33	0.29	0.34	0.31

* 음영은 평균제곱오차 제곱근이 1이상인 경우

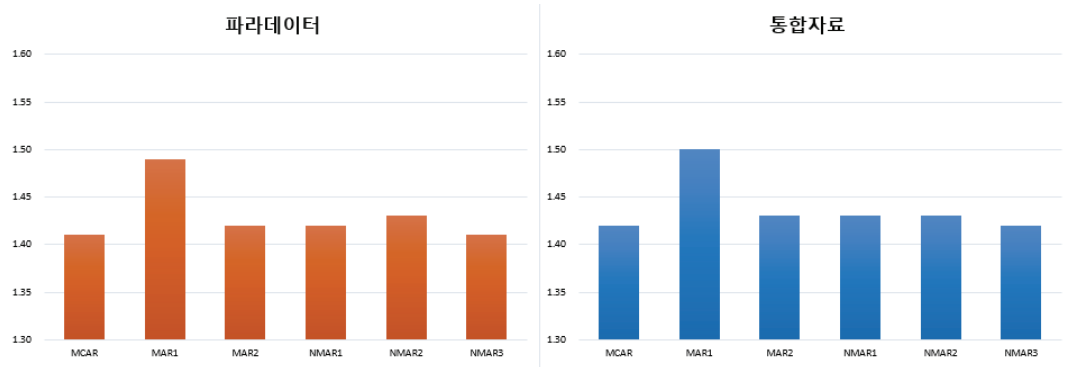


〈그림 4.3〉 추정값의 평균오차 제곱근 비교

설계효과(design effect due to weighting)는 무응답 조정이나 비포함(noncoverage) 보정 등 가중값 조정을 하고 나면 추정값의 분산이 커지는 효과를 살펴보는 측정도구로, 특히 무응답 조정은 응답 가구 가중값의 변동성이 추가적으로 더욱 커질 수밖에 없다(Kish, 1992; Rizzo 등, 1996). 활용된 보조정보 간에는 큰 차이를 보이지 않는 것으로 확인됐다. 무응답 메커니즘 별로 살펴보면, 최종적으로 가중값 분산은 40~50% 증가를 보이며, 표본오차로는 19~20% 증가했다. 특히 MAR1 자료의 분산 증가가 상대적으로 높은 이유는 상대적으로 층(strata)이 지나치게 세분화되는데 기인한다. 즉, 세분화된 층에 비해 상대적으로 표본규모가 작게 배정되어 분산의 확대가 커진 것으로 평가할 수 있다.

〈표 4.4〉 설계효과 비교

	파라데이터						통합자료					
	MCAR	MAR1	MAR2	NMAR1	NMAR2	NMAR3	MCAR	MAR1	MAR2	NMAR1	NMAR2	NMAR3
$1 + CV^2$	1.41	1.49	1.42	1.42	1.43	1.41	1.42	1.50	1.43	1.43	1.43	1.42
$\sqrt{1 + CV^2}$	1.19	1.22	1.19	1.19	1.20	1.19	1.19	1.22	1.20	1.20	1.20	1.19



〈그림 4.4〉 설계효과 비교

지금까지 무응답 처리를 위한 무응답 메커니즘 별로 보조변수의 관련여부와 처리효과를 살펴보았다. 그 결과 무응답 메커니즘이 보조변수와 직접 관련된 경우에 처리효과가 우수한 것으로 나타났으며, 또한 응답성향과 관련된 관심변수와 상관관계가 높은 보조정보를 가진 경우도 상대적으로 성능이 우수함을 확인할 수 있었다.

끝으로, 구성된 보조변수의 유용성을 보기 위해 다항 로짓 모형(baseline-category logit model)을 제안하고, 모형 적합성은 우도비 검정(likelihood ratio test)을 통해 유의미성 여부를 논의한다. 특정 관심변수를 종속변수(dependent variable)로, 보조정보를 독립변수(independent variables)로 사용하여 보조정보가 관심변수에 미치는 영향에 대한 유의미 여부를 검정하기 위해 3가지 가설을 설정했으며, 이는 다음과 같다.

$$\text{가설1) } H_0 : \log\left(\frac{\pi_k}{\pi_1}\right) = \beta_0 \text{ (intercept-only)}, \quad H_1 : \log\left(\frac{\pi_k}{\pi_1}\right) = \beta_0 + \beta_i X_i, \text{ } i \text{는 파라데이터}$$

가설2) $H_0 : \log\left(\frac{\pi_k}{\pi_1}\right) = \beta_0$ (*intercept-only*), $H_1 : \log\left(\frac{\pi_k}{\pi_1}\right) = \beta_0 + \beta_j X_j$, j 는 통합자료

가설3) $H_0 : \log\left(\frac{\pi_k}{\pi_1}\right) = \beta_0 + \beta_i X_i$, i 는 파라미터, $H_1 : \log\left(\frac{\pi_j}{\pi_1}\right) = \beta_0 + \beta_j X_j$, j 는 통합자료

보조변수의 유의성 여부는 귀무가설 모델에 대한 대립가설 모델의 적합성 검정을 통해 확인이 가능하다. 가설1은 귀무가설이 기각됨에 따라 파라미터를 이용한 모형이 통계적으로 유의한 것으로 나타났으며, 가설2는 통합자료를 이용한 모형이 크게 유의미한 결과를 보였다(〈표 4.5〉). 가설3은 파라미터를 이용한 모형에 비해 통합자료를 이용한 모형이 적합성이 우수함을 우도비 검정을 통해 확인할 수 있다. 즉, 행정자료로부터 활용된 보조정보 중 적어도 하나의 변수는 종속변수에 영향을 미치고 있어 행정자료를 사용한 통합자료 모델이 조사변수의 설명력을 높이는 것으로 확인됐다. 개별 변수가 종속변수의 변동성을 설명하는 세부적인 유의성 점검(Type III analysis of effects)도 살펴 볼 수 있으나, 사전에 상관관계가 높은 변수를 선정했으며 반복하는 모의실험인 점을 감안할 때, 개별변수를 찾는 것은 본 연구를 넘어선 것으로 판단된다.

행정자료를 포함한 통합자료를 보조정보로 구성한 경우, 무응답 처리에 높은 효과를 보인 것과 같이 모형에 대해서도 설명력이 높아짐을 확인할 수 있었다.

〈표 4.5〉 우도비 검정(likelihood ratio test)을 통한 모형 비교

(단위: 건수/ 300회)

	<i>P</i> -value < 0.05					
	가설 1		가설 2		가설 3	
주관적 만족감	<i>Reject H₀</i>	300	<i>Reject H₀</i>	300	<i>Reject H₀</i>	300
건강평가	<i>Reject H₀</i>	300	<i>Reject H₀</i>	300	<i>Reject H₀</i>	300
흡연여부	<i>Reject H₀</i>	161	<i>Reject H₀</i>	300	<i>Reject H₀</i>	300
음주여부	<i>Reject H₀</i>	300	<i>Reject H₀</i>	300	<i>Reject H₀</i>	300
의료서비스 만족도	<i>Reject H₀</i>	300	<i>Reject H₀</i>	300	<i>Reject H₀</i>	300
학교교육 효과	<i>Reject H₀</i>	300	<i>Reject H₀</i>	300	<i>Reject H₀</i>	300
결혼에 대한 견해	<i>Reject H₀</i>	300	<i>Reject H₀</i>	300	<i>Reject H₀</i>	300
이혼에 대한 견해	<i>Reject H₀</i>	300	<i>Reject H₀</i>	300	<i>Reject H₀</i>	300
가족관계 만족도	<i>Reject H₀</i>	263	<i>Reject H₀</i>	300	<i>Reject H₀</i>	300

5. 요약 및 결론

본 연구는 2016년 사회조사 결과를 기초로 응답성향에 영향을 미치는 유의미한 변수를 탐색하고 모의실험을 통해 무응답 처리에서 적정한 보조정보 구성의 중요성을 확인했다. 본 연구의 하나의 목적은 유용한 보조정보를 구성하여 무응답 편향을 효과적으로 줄이는데 있다. 응답성향에 영향을 미치는 변수로는 지역, 주택유형, 방문횟수, 점유형태, 가구원수, 가구주 성별 및 연령 등이 발견되었으며, 이러한 변수는 무응답 처리를 위한 주요 보조정보로 활용이 가능했다. 이렇게 얻어진 정보원으로부터, 파라미터 변수만으로 이루어진 파라미터 정보와 행정자료로부터 입수되는 변수까지 포함한 통합자료 정보로

구분하여 연구를 진행했다.

조사변수에 대한 무응답 구조를 이해하기 위해, 여러 가지 무응답 메커니즘을 가정하여 접근했다. 완전임의결측(MCAR) 자료는 무응답 발생이 특정 변수와 관련성이 없어, 별도 무응답 처리가 요구되지 않는 것으로 판단된다. 무응답 처리 후 추정값의 점추정 및 분산에서 모두 효과가 우수했다. 임의결측(MAR) 자료는 모든 무응답 처리연구에서 알려진 것처럼 매우 성공적이다. 무응답 발생여부가 보조정보를 의존함에 따라 당연하게도 편향 축소에 우수한 효과를 충족했다. 응답 여부와 관련하여 상관관계가 높으며, 보다 많은 정보량을 가진 통합자료의 효과가 더욱 우수함을 확인할 수 있다. 우리의 관심사인 비임의결측(NMAR) 자료의 무응답 발생여부는 근본적으로 무시할 수 있는 무응답인지를 결정하는 것이 쉽지 않아 보인다. 본 연구와 관련하여 소득과 상관관계가 높은 가구원수, 가구주 연령 등 보조정보를 활용함에 따라 무응답 조정이 성공적으로 작용함을 확인할 수 있었다. 보조정보의 양적 구성도 중요하지만, 무응답과 관련성이 낮은 보조변수 사용은 편향 감소에 큰 도움 없이 정확성을 낮추는 결과를 가져와 노력의 의미가 없어질 수 있다. 따라서 보조정보는 응답성향과 관련되며, 동시에 추정해야 하는 조사변수와도 관련이 있어야 한다(Rizzo, 1996).

소득과 관련된 조사변수의 무응답 발생여부는 무시할 수 없는(non-ignorable) 비임의결측(NMAR) 메커니즘을 따를 가능성이 높으며, 관련 보조정보가 없는 경우에는 전통적인 처리방법이 유효하지 않을 수 있다(정구현 등, 2013; Guillaume, 2013). 소득과 밀접한 세금, 공시지가 및 실거래가 등 관련 행정자료 수집은 무응답 편향 축소에 크게 기여할 것이다. 행정자료의 적극적 수집 확대와 함께 파라데이터의 체계적인 수집과 정리에 대한 재정립이 필요하다(West, 2011). 파라데이터는 조사대상에 대한 접근성 및 조사원 의지를 크게 의존하지만, 무응답 사유, 방문횟수, 면접시간, 협조도, 응답전환(refusal conversion) 정보 등 행정자료에서 확보할 수 없는 긴요한 정보 수집이 가능하며, 조사관리, 조사설계 개선, 조사원 평가 및 무응답 처리 등에 활용도가 높다는 점에서 정보화 기술을 이용한 자료 수집 및 적합한 형태로 정리는 매우 중요하다.

또한 조사에 따라 내용과 결과가 달라질 수 있어 무응답 처리를 위해 표본설계 및 무응답에 대한 처리 경험 등을 바탕으로 체계적인 연구가 요청된다. 일례로 미국의 SIPP(survey of income and program participation) 무응답 편향에 대한 기록을 통해 무응답과 관련된 잠재된 편향(potential bias)의 효과 분석 및 무응답에 대한 경험적 통찰력 등을 축적하여 조사환경에 따라 달라지는 무응답을 효과적으로 대처하고 있음을 볼 수 있다.

본 연구는 무응답 오차에 관한 측면만을 지나치게 강조함이 없지 않다. 하지만, 총조사 오차(total survey error)를 구성하는 표집오차(sampling error), 포함오차(coverage error), 측정오차(measurement error), 처리오차(processing error) 등도 함께 관심을 기울여 전체적으로 조사의 완전성을 높일 필요가 있다.

(2018년 3월 19일 접수, 2018년 4월 16일 수정, 2018년 4월 24일 채택)

감사의 글

본 연구가 진행될 수 있도록 자료제공 및 여러 가지 도움을 준 통계청 표본과 임경은 서기관님과 김연정 주무관님께 감사드립니다.

부록

〈부표 1〉 무응답 처리 이전 추정값의 편향 및 상대편향 비교

		편향(Bias)						상대편향(Relative Bias)					
		MCAR	MAR1	MAR2	NMAR1	NMAR2	NMAR3	MCAR	MAR1	MAR2	NMAR1	NMAR2	NMAR3
주관적 만족감	만족	-0.10	-0.28	-0.40	-0.90	-0.47	-1.31	-0.25	-0.73	-1.05	-2.37	-1.25	-3.46
	보통	0.06	0.28	0.29	0.40	0.26	0.61	0.13	0.61	0.63	0.87	0.57	1.33
	불만족	0.03	0.00	0.11	0.50	0.21	0.70	0.20	-0.03	0.68	3.07	1.32	4.29
건강평가	좋음	0.02	-0.04	-0.47	-0.79	-0.20	-0.54	0.05	-0.11	-1.17	-1.94	-0.48	-1.33
	보통	-0.04	-0.61	-0.37	-0.29	-0.25	0.18	-0.10	-1.55	-0.95	-0.73	-0.63	0.45
	나쁨	0.02	0.65	0.85	1.08	0.44	0.36	0.08	3.25	4.21	5.35	2.20	1.81
흡연여부		0.14	0.00	-0.45	-0.32	-0.04	0.14	0.46	-0.01	-1.73	-1.02	-0.13	0.46
음주여부		-0.10	-0.87	-1.10	-1.11	-0.61	-0.18	-0.14	-1.21	-1.52	-1.53	-0.85	-0.24
의료서비스 만족도	만족	0.16	0.94	0.65	0.69	0.63	0.04	0.32	1.84	1.27	1.36	1.24	0.08
	보통	-0.14	-0.76	-0.57	-0.64	-0.49	-0.04	-0.35	-1.88	-1.39	-1.57	-1.22	-0.11
	불만족	-0.02	-0.17	-0.08	-0.06	-0.14	0.00	-0.23	-2.07	-0.97	-0.67	-1.64	0.03
학교교육 효과	높음	-0.07	0.90	0.51	0.22	0.25	-0.22	-0.18	2.38	1.34	0.59	0.65	-0.57
	보통	0.06	-0.63	-0.24	-0.17	-0.19	0.16	0.16	-1.57	-0.61	-0.43	-0.49	0.40
	낮음	0.01	-0.28	-0.27	-0.05	-0.05	0.05	0.03	-1.26	-1.21	-0.22	-0.24	0.25
결혼에 대한 견해	필요	-0.07	0.21	0.50	0.04	0.12	-0.18	-0.12	0.36	0.87	0.08	0.20	-0.32
	중립	0.07	-0.20	-0.45	-0.12	-0.11	0.15	0.19	-0.51	-1.16	-0.31	-0.28	0.39
	불필요	0.01	-0.05	-0.05	0.02	-0.02	0.03	0.52	-1.95	-2.29	1.62	-1.03	1.82
이혼에 대한 견해	불가	0.00	0.42	0.62	0.16	0.06	-0.05	0.00	0.91	1.32	0.35	0.12	-0.10
	중립	-0.07	-0.52	-0.59	-0.31	-0.18	-0.03	-0.17	-1.36	-1.55	-0.80	-0.48	-0.08
	가능	0.07	0.10	-0.02	0.14	0.13	0.08	0.43	0.63	-0.16	0.95	0.84	0.52
가족관계 만족도	만족	-0.09	0.10	-0.28	-0.36	-0.07	-0.46	-0.17	0.18	-0.52	-0.67	-0.12	-0.85
	보통	0.07	-0.19	0.22	0.22	-0.02	0.37	0.18	-0.47	0.54	0.53	-0.04	0.89
	불만족	0.02	0.09	0.06	0.14	0.08	0.09	0.37	1.93	1.25	2.95	1.70	1.89

* 음영은 편향이 0.6% 이상 또는 상대편향이 1.5% 이상

참고문헌

- 김서영, 박라나 (2013). 무응답가구의 특성 분석 사례연구, <조사연구>, 14(1), 31-67.
- 김서영, 안다영 (2010). 가구 면접조사에서 무응답률과 무응답 편향, <2010년 상반기 연구보고서 제Ⅱ권>, 111-164.
- 김정섭, 임경은 (2010). 총오차 축소를 위한 Paradata 수집방안, <2010년 상반기 연구보고서 제Ⅱ권>, 62-110.
- 이승희 (2010). 총조사 오차의 패러다임으로 이해하는 표본조사, <통계연구> 15(1), 44-74.
- 정구현, 이석훈, 변종석 (2013). 가계금융·복지조사의 무응답 처리를 위한 유용한 보조정보 선정, <조사연구>, 14(1), 69-91.
- 통계청 (2016a). 2015년 사회조사 조사과정 자료(Paradata) 분석, <통계청 내부자료>
- 통계청 (2016b). 2016년 사회조사 표본설계 보고서, <통계청 내부자료>
- 통계청 (2016c). 표본설계 기준 및 지침, <통계청 내부자료>
- 통계청 (2017). 2016 인구주택총조사 전수 집계결과 보도자료, <통계청 보도자료>
- Atrostic, B.K. Bates, N. Burt, G. and Silberstein, A. (2001). Nonresponse in U.S. Government Household Surveys: Consistent Measures, Recent Trends, and New Insights. *Journal of Official Statistics*, 17, 209-226.
- Bethlehem, J.G. and Keller, W.J. (1987). Linear Weighting of Sample Survey Data, *Journal of Official Statistics*. 3: 141-153.
- Bethlehem, J. Cobben, F. and Schouten, B. (2011). *Handbook of Nonresponse in Household Surveys*. John Wiley & Sons.
- Bover, O. (2011). The Spanish Survey of Household Finances (EFF): Description and Methods of the 2008 wave. Bank of Spain, Working Paper.
- Brick, J.M. J. Bruke, and T. Le. (2000). Analysis of Nonresponse Bias in the Base Year, Early Childhood Longitudinal Study: Kindergarten Class of 1998-99. Unpublished Report. Washington, DC: National Center For Education Statistics.
- Brick, J.M. Williams, D. (2013). Explaining Rising Nonresponse Rates in Cross-Sectional Surveys. *The ANNALS of the American Academy of Political and Social Science*, 645(1), 36-59.
- Curtin, R. Presser, S. and Singer, E.(2000). The Effects of Response Rate Changes on the Index of Consumer Sentiment, *Public Opinion Quarterly*, 64, 413-428.
- David, M. Little, R.J.A. Samuhel, M. and Triest, R. (1986). Alternative Methods for CPS Income Imputation. *Journal of the American Statistical Association*, 81, 29-41.
- de Leeuw, E. & de Heer, W. (2002). Trends in household nonresponse: A longitudinal and international comparison. In R. Groves, D. Dillman, J. Eltinge & R. Little (Eds.), *Survey nonresponse*, 41-54. New York: Wiley

- Deming, W.E. and Stephan, F.F. (1940). On a Least Square Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known, *Annals of Mathematical Statistics*, 11, 427-444.
- Deville, J.C. and Sarndal, C-E. (1992). Calibration Estimators in Survey Sampling, *American Statistical Association*, 87, 376-382.
- Dillman, D.A. (2000). *Mail and Internet Surveys: The Tailored Design Method. 2nd Edition*. New York: John Wiley.
- Groves, R.M. (1989). *Survey Errors and Survey Costs*. New York: John Wiley & Sons.
- Groves, R.M. (2006). Nonresponse Rates and Nonresponse Bias in Household Surveys. *The Public Opinion Quarterly*, 70(5), 646-675.
- Guillaume, O. (2013) Dealing with Non-ignorable Non-response using Generalised Calibration: A Simulation Study based on the Luxemburgish Household Budget Survey, Working Papers.
- Kalton, G. and Flores-Cervantes, I. (2003). Weighting methods, *Journal of Official Statistics*. 19(2): 81-97.
- Kalton, G. and Maligalig, D.S. (1991). A Comparison of Methods of Weighting Adjustments for Nonresponse, paper presented at the 1991 Census Research Conference.
- Kish, L. (1992). Weighting for unequal Pi. *Journal of Official Statistics*. Vol 8(2), 183-200.
- Kish, L. (1965). *Survey Sampling*. John Wiley & Sons. New York.
- Kreuter, F. Olson, K. Wagner, J. Yan, T. Ezzati-Rice, T.M. Casas-Cordero, C. Lemay, M. Peytchev, A. Groves, R.M. and Raghunathan, T.E. (2010). Using Proxy Measures and Other Correlates of Survey Outcomes to Adjust for Non-Response: Examples from Multiple Surveys. *Journal of the Royal Statistical Society, Series A*, 173, 389-407.
- Lin, I.-F. and Schaeffer, N.C. (1995). Using Survey Participants to Estimate the Impact of Nonparticipation. *Public Opinion Quarterly*, 59, 236-258.
- Little, R.J.A. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. John Wiley & Sons, New York.
- Little, R.J.A. (1992). Models for Nonresponse in Sample Surveys, *Journal of the American Statistical Association*, 77, 237-250.
- Little, R.J.A. and S. Vartivarian. (2003). On weighting the rates in nonresponse weights, *Statistics in Medicine*, 22, 1589-1599.
- Matei, A. and Ranalli, G. (2015). Dealing with non-ignorable nonresponse in survey sampling: A latent modeling approach, *Survey Methodology*, 41(1), 145-164.
- Molenberghs, G. Beunckens, C. Sotto, C. and Kenward, M.G. (2008). Every Missingness not at Random Model has a Missingness at Random

- Counterpart With Equal Fit. *Journal of the Royal Statistical Society: Series B*, 70, 371–388.
- Oh, H.L. and Scheuren, F.S.(1983). Weighting adjustment for unit nonresponse, In *Incomplete data in sample surveys*. Madow, W.G. Olkin, I. and Rubin, D.B.(eds), New York : Academic Press, 143–184.
- Olson, K. (2006). Survey Participation, Nonresponse Bias, Measurement Error Bias, and Total Bias. *Public Opinion Quarterly*, 70, 737–758.
- Paul, D.A. (2012). Handling Missing Data by Maximum Likelihood, SAS Global Forum, 312–2012.
- Rizzo, L. Kalton, G. and Brick, M. (1996). A comparison of some weighting adjustment methods for panel nonresponse, *Survey Methodology*, 22, 43–53.
- Rubin, D.B. (1976). Inference and Missing Data. *Biometrika* 63, 581–592.
- Sharon L. Lohr. (2010). *Sampling: design and analysis*. 2nd Ed. Brooks/Cole, Cengage Learning.
- Statistics Canada (2003). Survey Methods and Practices.
- Steele, F. and Durrant, G.B. (2011). Alternative Approaches to Multilevel Modelling of Survey Non-Contact and Refusal. *International Statistical Review*, 79, 70–91.
- Stoop, I.A.L. (2005). The Hunt for the Last Respondent: Nonresponse in Sample Surveys. The Hague: Social and Cultural Planning Office.
- Sukasih, A. Jang, D. Vartivarian, S. Cohen, S. and Zhang, F. (2009). A Simulation Study to Compare Weighting Method for Nonresponse in the National Survey of Recent College Graduates, JSM 2009.
- Survey of Income and Program Participation 2014 Nonresponse Bias Report (2017). https://www2.census.gov/programs-surveys/sipp/tech-documentation/complete-documents/2014/2014_SIPP_Wave_1_Nonresponse_Bias_Report_v1.0.2017.05.16.pdf
- UNECE (2011). Using Administrative and Secondary Sources for Official Statistics, A Handbook of Principles and Practices, UNECE.
- Weeks, M.F. (1988). Call Scheduling with CATI: Current Capabilities and Methods. In: Groves, R.M. et al. eds. *Telephone Survey Methodology*. 403–420. New York: Wiley.
- West, B.T. (2011). Paradata in Survey Research: Examples, Utility, Quality, and Future Directions. *Survey Practice*, 4(4).
- [http://kosis.kr/statisticsList/statisticsListIndex.do?menuId=M_01_01&vwcd=MT_ZTITLE&parmTabId=M_01_01#SelectStatsBoxDiv\(통계청 2017년 가계금융·복지조사\)](http://kosis.kr/statisticsList/statisticsListIndex.do?menuId=M_01_01&vwcd=MT_ZTITLE&parmTabId=M_01_01#SelectStatsBoxDiv(통계청 2017년 가계금융·복지조사))

Treatment of unit non-response in Korean Social Survey

Goohyun Jung¹⁾ · Sunghee Lee²⁾

Abstract

Unit non-response is inevitable in most surveys. Statistics Korea had used substitution traditionally as a way to mitigate non-response issues. Due to serious concerns associated with selection bias in substitution, since 2016, Statistics Korea no longer uses substitution and has made attempt to investigate the ways to detect and handle non-response bias. In particular, there is considerable amount of paradata and administrative data available at Statistics Korea that can be used for non-response adjustment.

This study compares different adjustment approaches for non-response in the Korean Social Survey. Using the 2016 Korean Social Survey as a pseudo-population, a simulation study was carried out where samples were generated under various non-response mechanisms and non-response was corrected for using paradata and administrative data through weighting. Results indicates that weighting adjustment is effective in reducing non-response bias in the Korean Social Survey. In particular, including administrative data in the adjustment improves the effectiveness.

Key words : Auxiliary Information, Non-response Mechanism, Paradata, Administrative data, Weighting Adjustment

1) (Corresponding author) Visiting Scholar, Institute for Social Research, University of Michigan.
E-mail: goohyunj@umich.edu

2) Institute for Social Research, University of Michigan. E-mail: goohyunj@umich.edu