빅데이터를 활용한 조사모집단 구축 및 적응적 표본설계 방안 연구

(부제 : 행정자료 등을 이용한 한국 사회조사 단위무응답 처리를 위한 최적의 보조정보 구성에 관한 연구)

2018년 8월

통 계 청
정 구 현

순 서

1. 서 론 6

1.1 연구배경 6
1.2 연구방향 9
1.3 연구한계 및 기대효과 10

2. 기초자료 분석 12

2.1. 주요 조사 무응답 처리 현황 12
2.2 기존의 사회조사 분석 17
2.3 2016년 사회조사 표본설계 22
2.4 무응답률 및 파라데이터 분석 25
2.5 무응답 편향 탐색 34

3. 모의실험(simulation) 41

3.1 모집단 및 표본셋 구성 41
3.2. 무응답 발생 및 유의성 검토 43
3.3. 보조정보 구성 48
3.4 무응답 처리 52

순 서

4. 모의실험 결과 58

4.1 평가내용 58

4.2 가중값 적용 59

4.3 평가 척도 61

4.4. 처리결과 64

1) 편향(bias) 65

2) 상대편향(relative bias) 67

3) 평균제곱오차 제곱근(RMSE) 75

4) 설계효과(Design Effect) 77

5) 모형비교 78

6) 시도별 무응답 처리결과 79

5. 결 론 82

5.1 연구내용 요약 82

5.2 시사점 및 향후 과제 84

<부록>

1. 통계저널 논문게재(2018. 7. 2.) : 통계연구(2018), 제23권 제2호, 27-49
2. 미시간대학교 연구소 세미나 발표(2018. 2. 7.)
3. 북미 통계학회(Joint Statistical Meeting) 발표(2018. 7 .31.)

훈련기관 개요

Ⅰ. 기관 개요

○ 훈련국 : 미국

○ 훈련기관명 : ISR(Institute for Social Research) University of Michigan

○ 인터넷 웹주소 : http://home.isr.umich.edu/

○ 기타(주소 등) : P.O. Box 1248
426 Thompson St. Ann Arbor, MI 48106-1248
Voice: (734) 764-8354 Fax: (734) 647-4575 email: umisr-info@umich.edu

Ⅱ. 기관 소개

1. 설립목적(Mission)

ISR(Institute for Social Institute)은 미시간대학의 central campus에 위치해 있으며 세계의 사회과학 분야의 리드 연구기관으로 다음과 같은 미션을 수행

○ 미국의 인구, 사회, 경제 정책 등에 관한 연구의 구체적인 계획과 실행

○ 연구에서 발견한 것들의 전파

○ 차세대 사회과학자 훈련

5개의 부서인 정치연구센터(CPS: Center for Political Studies), 정치와 사회연구 지원부서(ICPSR: Inter-University Consorcium for Political and Social Research), 인구연구센터(PSC: Population Studies Center), 집단의 동태 연구센터(RCGD: Research Center for Group Dynamics), 조사연구센터(SRC: Survey Research Center)에서 미션 수행에 필요한 역할 담당

2. 연 혁(History)

2009년 60주년이었으며 그 동안의 발자취를 부서 생성과 중요한 조사 중심으로 살펴보았음

○ 1946년 : Survey Research Center 탄생

○ 1946년 Surveys of Consumers : 미국경제에 대한 소비자 태도와 기대를 조사하는 것으로 시작, 현재 이 조사는 매월조사

○ 1948년 The Summer Institute in Survey Research Techniques
- SRC는 Summer Institute 를 통해 대학원 수준의 조사설계, 실행, 분석 등의 수업제공, 60여년 동안 105개 이상 국가에서 1만명 이상의 학생이 수강

○ 1948년 Michigan Election Studies
- Harry Truman 과 Thomas Dewey 대통령 선거에서 예측이 거의 근사하게 일치함에 따라 선거태도에 대한 연구 시작

○ 1949년 ISR설립
- 초대 연구소장 : “Likert Scale"을 개발한 Rensis Likert

○ 1951년 The Detroit Area Study

○ 1955년 Growth of American Families Survey
- 현재 National Survey of Family Growth

○ 1961년 Sampling Program for Foreign Statisticians
- Leslie Kish : 대학원에 개도국 통계인을 위한 8주간의 summer program 제안

○ 1961년 Population Studies Center
- 인구연구에 대해 학생훈련 및 인구학자들의 공동연구의 장

○ 1962년 The Inter-university Consortium for Political and Social Research
- 세계에서 가장 큰 사회과학 데이터, 약 100만 연구 데이타 보유

○ 1965년 Survey Sampling
- Leslie Kish : 현장에서의 bible 인 샘플링 관련 지식 저술

○ 1968년 The Panel Study of Income Dynamics
- 세대 간 분석과 빈곤동태(dynamics of poverty)의 이해

○ 1970년 The Center for Political Studies
- 개인 선거 행동, 미디어와 정치, 정치와 개인행동에 관한 기관효과 등 연구

○ 1975년 Computer-Assisted Interviewing
- CATI(Computer-Assisted Telephone Interviewing)

○ 1975년 Monitoring the Future
- 매년 조사, 50,000명, 음주, 흡연, 마약 사용에 대한 트랜드, 청소년기의 가치 및 태도변화, 담배 판매 전략의 영향, 국가의 술 및 마약 보호 정책에 활용

○ 1976년 Program for Research on Black Americans
- 처음 조사이후 8, 9, 12년차에 동일 대상에 대해 재조사한 종단 데이터 제공

○ 1976년 Centralized Telephone Interviewing
- 조사방법에서 face-to-face를 대체하는 가장 경제적이고 효과적인 방법

○ 1980년 Research on the Census
- Reynolds Farly : 1980년 미국 센서스와 함께 30년의 센서스를 통해 미국인의 인구, 사회, 경제 특성들을 미국인들이 어떻게 변화하여 왔는지 비교

○ 1981년 Questions and Answers in Surveys
- Howard Schuman and Stanley Presser : 메릴랜드대학 소속, 저술

○ 1991년 University of Warsaw Collaboration
- ISR의 첫 자매결연, 폴란드 바르샤바 대학 연구소, 이후 Germany, South Africa, Japan, China, and Qatar 등으로 확대

○ 1992년 The Joint Program in Survey Methodology
- 메릴랜드, 미시간, Westat 합동의 훈련프로그램 구성

○ 1995년 Data Documentation Initiative
- 사회과학 문서들의 상호교환을 위해 국제표준 시도

○ 1998년 Population Studies Center Joins ISR

- Population Studies Center 를 ISR 에 통합, 건강, 장애, 사망, 인구동태, 고령화, 방법론 등 연구

○ 2001년 ICPSR Delivers Data over the World Wide Web

○ 2002년 Michigan Census Research Data Center

○ 2009: 60th anniversary

3. 훈련기관에 대한 설명

○ ISR(http://www.src.isr.umich.edu)

- 사회과학연구의 세계유수 기관으로 방법론, 데이터서비스, 차세대 인력 양성에서 리드역할
  - · CATI 처음 도입과 함께 집중화된 CA-기반의 조사방법 개발 및 효과 검증
  - · 2002년 데이터공유를 위한 센서스 데이터 센터 설치
  - · 통계작성기관의 재원으로 메릴랜드대학, Westat과 공동으로 통계종사자 양성 교육 제공 및 대학원 운영

- 교수(85명), 계약교수(36명), 직원(1,200명)

Regular Salaried Employees:		**Other Temporary Staff***	**130**
Primary Research Staff*	**150**	Off-Site Contingent Staff**	**546**
Research Fellows/GSRAs	**29**	On-Site Contingent Staff***	**109**
Regular Salaried Employees	**315**	**TOTAL TEMPORARY EMPLOYEES:**	**785**
TOTAL REGULAR SALARIED EMPLOYEES:	**494**	* *Appointments in SRC*	
		** *Active Field Interviewers*	
		*** *SSL Temporary Survey Technicians*	
TOTAL SRC PERSONNEL:	**1,279**		

1. 서론

1.1 연구배경

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)(<표 1.1>). 특히 조사기간이 짧은 표본조사의 경우, 1인가구나 맞벌이 가구 증가에 따른 주간부재는 응답률에 더욱 부정적인 영향으로 작용하고 있다(김서영과 박라나, 2013).

<표 1.1> 한국 통계청 표본조사 무응답률(%)

통계명	'10	'11	'12	'13	'14	'15	'16
경제활동인구조사	5.2	5.9	6.8	7.3	8.2	9.0	9.7
지역별고용조사	3.7	5.1	5.6	7.6	8.6	9.9	10.2
가계동향조사	18.6	19.6	20.2	20.5	22.5	23.2	24.1

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 효과적으로 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 집중

되며, 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 관심과 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[1](<표 1.2>).

이와 함께 특정 그룹에서 무응답률에 대한 체계적인 차이는 무응답 편향과 관련된 다는 점에서 특성별 응답성향에 대한 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 통

1) 1925년부터 실시해온 면접방법의 센서스를 2015년 등록센서스로 전환에 따라 매년 최신 행정자료를 수집하여 인구, 가구특성을 작성하여 공표(14개 기관 25종 행정자료 활용)

계적 모의실험을 전개한다.

<표 1.2> 2015년 인구주택조사 조사항목

구 분		등록센서스 항목(12)	표본조사 항목(52)	
UN 권고 항목 (38)	인구 (24)	①성명 ②성별 ③나이 ④가구주와의 관계 ⑤국적 ⑥입국연월	①성명 ②성별 ③나이 ④가구주와의 관계 ⑤국적 ⑥입국연월 ⑦종교 ⑧교육정도 ⑨전공계열 ⑩출생지 ⑪1년전 거주지 ⑫5년전 거주지	⑬활동제약 ⑭통근통학여부 ⑮통근통학장소 ⑯경제활동상태 ⑰종사상지위 ⑱산업 ⑲직업 ⑳근로장소 ㉑혼인상태 ㉒혼인연월 ㉓출산자녀수 ㉔자녀출산시기
	가구 (8)	①가구구분	①가구구분 ②사용방수 ③주거시설형태 ④난방시설	⑤주거전용 영업겸용 여부 ⑥점유형태 ⑦임차료 ⑧주인가구여부
	주택 (6)	①거처의 종류 ②주거용 연면적 ③건축연도 ④대지면적	①거처의 종류 ②총방수 ③주거시설수	④주거용 연면적 ⑤건축연도 ⑥대지면적
고유 항목 (15)	인구 (11)	①본관	①아동보육 ②이용교통수단 ③일상생활·사회생활 제약 ④통근통학소요시간 ⑤현직업 근무연수	⑥추가계획자녀수 ⑦결혼 전 취업 여부 ⑧경력단절 ⑨사회활동 ⑩고령자생활비원천
	가구 (4)		①거주기간 ②건물 및 거주층	③주차장소 ④타지주택소유여부

1.2 연구방향

무응답(nonresponse)은 표본가구에서 조사에 응답하지 않는 단위 무응답(unit nonresponse)과 조사에는 응답하나 일부 내용에 응답 하지 않는 항목 무응답(item nonresponse)으로 구분된다. 항목무응답은 대체로 무응답 대체 방법(imputation)을 통해 무응답 편향을 보정한다. 단위 무응답의 경우는 보통 무응답 가중값 조정을 통해 무응답을 보정하고 있다(Groves, 2006).

무응답 가중값 적용 방법은 응답과 무응답의 성향이 동일하다는 것을 가정 하에 시작한다. 이런 가정이 위배될 경우에 무응답 가중값을 사용한 조정은 오히려 더 큰 편향을 가져올 수 있다. 다만, 현실적으로 조사에서 응답과 무응답의 성향을 파악하는 것은 쉬운 일은 아니다. 단위 무응답 발생은 응답률을 낮추는 결과를 갖기 때문에 표본조사에서는 단위 무응답에 보다 큰 관심을 갖게 되고, 본 연구는 단위 무응답에 초점을 두어 진행한다.

통계청 사회조사는 무응답이 발생하면 이러한 문제를 완화시키기 위해 조사현장에서 표본대체 방법을 사용했으며, 이는 선택편향 등 비표본오차가 발생하는 원인이 되며, 조사에 응답한 결과만을 이용함으로써 무응답 편향을 가져와 통계 정확도에 영향을 주고 있는 것으로 확인됐다(Kish, 1965; Little and Rubin, 1987; Sharon, 2010). 따라서 표본대체율이 30.8%에 이르고 있는 2015년 사회조사 결과의 분석을 통해 대체가구 및 대체자료에서 유의미성을 판단할 필요가 있다.

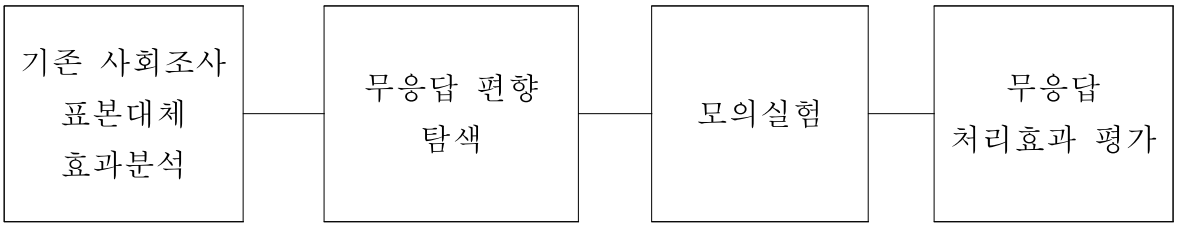
2016년부터는 응답 거부나 접촉할 수 없는 대상으로부터 발생되는 무응답에 대해서 대체하지 않고 무응답 처리방법을 통하여 처리할 예정으로 무응답에 영향을 미치는 주요 특성을 탐색하여 무응답 처리를 위한 보조변수(auxiliary variables)로 활용할 수 있다. 무응답을 탐색하기 위해 직접 조사를 통한 파라데이터와 간접 방식인 행정자료를 통해 얻는 방법이 있다.

조사변수의 무응답 발생여부(missingness)에 대한 구조를 이해하기 위해 Rubin(1976)이 제안한 3가지 메카니즘(mechanism)에 따라 모의실험(simulation)을 실시한다. 3가지 메카니즘은 완전임의결측(MCAR: Missing

Completely At Random), 임의결측(MAR: Missing At Random) 및 비임의결측(NMAR: Not Missing At Random)으로 구분되어 진행된다.

사회조사 무응답률은 2016년 18.6%이고 기존 표본대체율은 30%임을 감안하여 두 가지 무응답률에 따른 무응답 처리의 차이점을 비교하고자 한다.

가구조사 무응답 처리방법은 일반적으로 가중치 조정법(weighting adjustment)이 사용된다(Groves, 2006). 보조변수(auxiliary variables)의 주변분포에 맞게 조정하는 래이킹 비 조정법 및 캘리브레이션 보정방법, 보조변수의 주변분포 또는 결합분포에 맞게 조정이 가능하도록 발전한 일반화 회귀추정(GREG) 방법이 있다. 또한 보조변수의 교차표를 기초하여 무응답층을 결정하는 CHAID(Chi-squared automatic interaction detection) 알고리즘과 보조변수를 활용한 응답성향을 통해 가중값을 조정하는 로지스틱 분석이 있다.



<그림 1.1> 무응답 처리를 위한 연구진행 과정

1.3 연구한계 및 기대효과

본 연구는 몇 가지 한계점을 가지고 시작한다. 따라서 이러한 한계점 하에 연구가 진행되었고, 연구결과가 도출되었기 때문에 이를 이해하고 접근할 필요가 있다.

먼저, 파라데이터(paradata)에 관한 부분이다. 2015년부터 파라데이터를 수집해왔고, 2016년 조사결과에서 파라데이터를 이용한 보조정보 구성이 가능했다. 다만 파라데이터 내용은 지역(동부 읍면부), 연건평, 아파트 평형(소형 중형 대형), 주택유형, 무응답 사유, 방문일자, 방문시간 등으로 많은 부

분이 표본설계 표본추출틀(sampling frame)에서 직접 얻어지는 변수가 많아 내용면에서 충분하지 않다. 조사원이 직접 기록한 자료로 면접하기 어려울 것 같은 가구는 불능으로 처리할 수 있고, 방문횟수를 임의로 기록할 수 있으며 특히 불응과 불능의 구분이 명확하지 않다. 불능가구라고 판단되면 재방문이 필요치 않아 불응가구를 불능가구로 처리하는 등 조사원의 자의적인 판단에 의존할 수밖에 없다. 동시에 시스템이 아닌 엑셀 등을 이용함에 따라 수집된 자료가 완전하게 정돈되어 있지 못한 상태이다. 따라서 활용하는데 제한적일 수밖에 없다.

두 번째로, 행정자료(administrative data) 활용에 관련한 내용이다. 2015년 센서스 공표를 위해 13개 기관 24종의 자료를 수집(2015년 기준)하고 있고 이미 입수가 되어 있지만, 2016년 사회조사는 2010년 센서스 표본추출틀을 사용하고 있어 행정자료를 주요 정보로 사용할 수 없었다. 행정자료를 활용하기 위해서는 먼저 우리나라 모든 주소에 맞게 가구주 성별, 나이, 가구원수, 주택유형, 점유형태, 세대유형, 외국인 여부 등 미시적인 행정자료의 정보를 연결(matching)하는 작업이 선행돼야 한다. 기본적으로 2015년 센서스 표본추출틀은 개별 주소별로 상기 내용의 정보가 연결되어 구축되어 있다.

셋 번째로, 무응답 편향 및 처리를 위한 필요한 정보량의 부족에 따라 추가적인 추가 정보를 얻기 위해 응답가구만을 이용하여 응답성향(response propensity)를 구성하여 활용한다. 이러한 한계점에 있으나, 모의실험(simulation)은 원칙적으로 응답된 관측값에서 무응답을 발생하기 때문에 조사변수를 향후 수집 가능한 정보로 가정하여 진행할 수 있다.

본 내용을 통해 paradata의 체계적인 수집과 활용 가능한 형태의 중요성을 깨달을 수 있다. 이는 현장조사를 이해하는데 큰 도움을 주어 조사설계, 조사관리 개선, 조사원 평가 및 무응답 처리 등에 활용성이 높아 조사 정확성을 제고할 수 있는 가장 접근성 있는 효과적인 방법이라고 할 수 있다. 또한 통일된 기준으로 얻어진 파라데이터는 비표본오차(non-sampling error)를 축소하는데 유용하게 활용될 수 있다. 예로서, 적정 방문횟수 분석, 최적의 조사 기간 및 방문시간, 불응 최소화 전략 등 무응답을 최소화 할 수 있는 긴요한 자료로 활용할 수 있다.

행정자료는 응답자와 무응답자의 정보를 동시에 가지고 있으며, 무응답 처리를 위한 주요 보조정보로 활용할 수 있다. 무엇보다 응답성향이나 기억 등을 의존하지 않아 매우 정확한 자료라는 점에서 의미가 크다. 역시 표본설계에서도 층화변수로 활용이 가능하여 조사 전반에 활용성이 높다. 이를 통해 적응적 표본설계가 가능하다.

2. 기초자료 분석

2.1 주요 조사 무응답 처리 현황

1) 가계금융 · 복지조사

가계금융복지조사는 패널조사로 최초 무응답률은 17.2%로, 무응답 처리는 주로 단위무응답에서 이루어지고 있으며, 항목무응답의 경우는 재방문 및 전화질의 등에 의해 보완하고 있다.

무응답 처리방법은 SAS에서 제공하는 CHAID(chi-square automatic interaction detection) 모형에 따라 무응답 층을 구성하여, 층별로 응답 가구가 무응답 가구를 대표할 수 있도록 가중치를 조정(weighting adjustment)하는 방법을 사용한다.

보조변수(auxiliary variables)로 시도, 주택시도, 주택유형, 가구주 연령, 가구원수 및 교육정도를 이용하여 무응답 층을 구성하고 있으며, 주요 관심변수(target variables)로는 소득, 자산, 부채 등 금액변수 등을 추정한다.

2) 가계동향조사

가계동향조사는 주로 단위무응답을 처리하고 있으며, 항목 무응답은 가계금융 · 복지조사와 마찬가지로 전화를 통한 내용검토를 이용하여 처리하고 있다.

2016년 무응답율은 약 23%이며, 무응답처리방법은 CHAID(chi-square automatic interaction detection) 모형을 이용, 동질적인 50개 무응답층을 구성하여 셀조정 방법에 의해 처리한다.

무응답 처리를 위한 보조변수는 시도, 가구구분, 가구원수, 주택유형을 이용하고 있다.

3) 경제활동인구조사

경제활동인구조사는 2016년 무응답율은 약 10%이며, 가구원에 대한 단위무응답을 처리를 하며, 단위무응답 조정을 위해 로지스틱회귀분석의 성향점수를 이용하고 있다. 활용된 보조변수는 지역, 주택유형, 1인여부 및 농가 등을 사용한다.

4) 외국인 고용조사

외국인 고용조사 특성상 대상을 주소지 이동이 빈번하여 무응답이 발생시 대체허용를 허용하고 있으며, 2016년 기준으로 무응답율은 약 2.5%가 발생했다.

단위무응답 조정을 위해 로지스틱 회귀분석을 활용하여 무응답 조정을 하고 있으며, 활용되는 보조변수로는 지역, 국적 및 체류자격 등을 사용한다.

5) 영국 LFS(Labor Force Survey)

영국의 노동력조사의 무응답 처리방법은 로지스틱 회귀분석을 이용하여 처리를 하고 있으며, 활용 보조변수로 GOR(Government Office Regions), OAC group (Output Area Classification), household reference person age, sex, ethnicity, length of time at address, age of youngest dependent child 등이 이용된다.

사후보정을 위해서는 calibration 방법을 이용하여 지역, 성, 연령그룹에 맞게 벤치마킹하고 있다.

6) 미국 CPS(Current Population Survey)

미국의 노동력조사는 무응답 처리를 위해 127개 무응답 cluster를 구성하고 이를 세분화하여 214개 cell로 구성하여 가중치 조정을 수행한다.

214개 cell 별로 무응답 조정인자를 아래와 같이 산출하여 응답자가 무응답자를 대표할 수 있도록 가중값을 조정한다.

$F_{ij} = \frac{Z_{ij} + N_{ij}}{Z_{ij}}$, Z_{ij}는 i cluster, j cell 응답자, N_{ij}는 i cluster, j cell 무응답자

다만, 조정인자가 2를 초과하거나, 응답수가 50미만이면 인접 cell과 통합(collapsing)을 실시한다.

<표 2.1> 미국 CPS 무응답률

Nonresponse rate	1993	1994	1995	1996	2003
Overall Type A	4.69	6.19	6.86	6.63	7.25
Noncontact	1.77	2.30	2.41	2.28	2.58
Refusal	2.85	3.54	3.89	4.09	4.10
Other	.13	.32	.34	.25	.57

7) 미국 SIPP(Survey of Income Program Participation)

미국의 SIPP는 패널조사로 첫 회 조사에서는 무응답 층을 구성하기 위한 보조변수로 센서스 지역(MSA-central city, MSA-non-central city, other place), 가구주 인종(black, non-black), 점유형태(owner, renter), 가구원수(1, 2, 3, 4+ people) 및 빈곤층(high poverty, low poverty)으로 500개 이상 구성하여 층 내 가중치 조정했으며, 2008 Panel의 첫 조사 무응답율 19.4%로 나타났다.

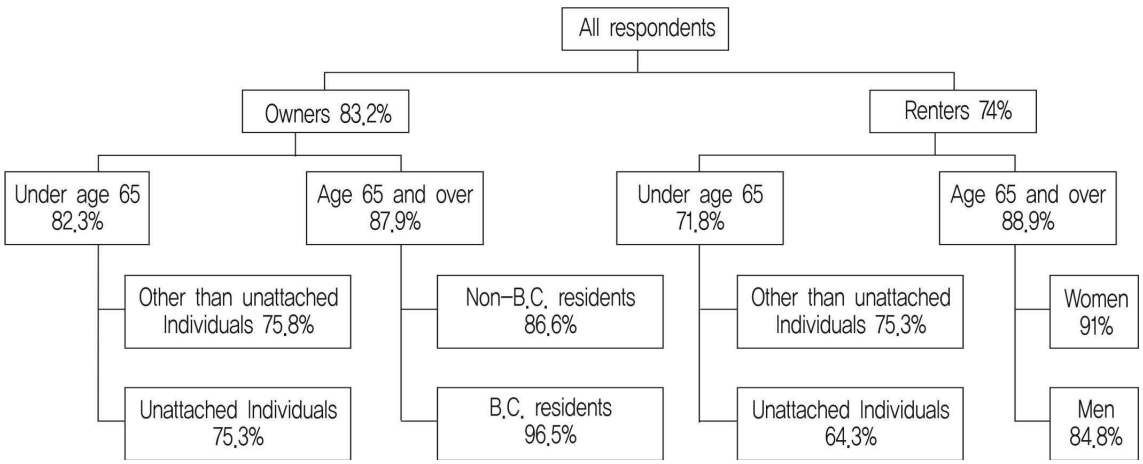
계속조사의 무응답 층을 구성하기 위한 점유형태(owner/renter occupied), 가구형태(female householder, no spouse present; 65+; other), 인종(race and Hispanic origin), 교육수준 및 가구원수(size), 빈곤상태(poverty status), 주요수입원(type of income), 재정상태(type of financial

assets) 등으로 층을 구성하여 진행한다.

이 또한 조정인자가 2를 초과하거나, 응답수가 30미만이면 인접 cell과 통합하여 층을 재조정하고 있다.

8) 캐나다 CIS(Canadian Income Survey)

캐나다 소득조사의 무응답 처리를 Chisquare automatic interacitve detection(CHID)를 이용하여 가중치 조정을 하며, 보조변수로 점유형태, 연령, 성별, 거주지역, 가족 동거여부 등을 보조변수로 사용한다.



<그림 2.1> 캐나다 소득조사 무응답 구성

9) 캐나다 LFS(Labour Force Survey)

캐나다 LFS의 경우 항목무응답에 대해서는 imputation를 사용하여 처리하고, 단위무응답에 대해서는 지역별로 가중치를 조정하여 보정한다.

이는 응답가구와 무응답가구의 특성이 유의미한 차이가 나지 않는다고 가정하에서 전개되며 무응답률은 약 10% 수준이다.

10) 호주 LFS(Labour Force Survey)

호주 LFS의 무응답율은 2008~2013년 5년 평균 96.5%으로 무응답률이 높지 않아 별도의 무응답 처리함 없이 지역, 성, 연령별 기준인구에 사후 보정만을 수행한다.

11) ESS(European Social Survey)

유럽국가 23개국이 참여하는 사회조사는 표본오차(sampling error) 및 무응답 편향(non-response bias) 제거를 위해 성별, 연령그룹, 교육정도, 지역에 맞게 사후보정 작업을 수행한다.

<표 2.2> ESS 사회조사 무응답 구성

Country	n	n_r	$\frac{n_r}{n}$	n_{int}	$\bar{attempts}$	$\bar{cases}$
AL	1602	1215	0.76	67	51.04	23.91
BE	3267	1869	0.57	155	69.29	21.08
BG	3200	2263	0.71	250	26.01	12.80
CH	2907	1529	0.53	106	114.36	27.42
CY	1588	1122	0.71	59	51.63	26.92
CZ	3010	2009	0.67	231	24.61	13.03
DE	8904	2964	0.33	200	151.99	44.52
DK	3372	1652	0.49	105	93.07	32.11
EE	3707	2382	0.64	144	63.76	25.74
ES	2868	1889	0.66	74	176.89	38.76
FI	3296	2197	0.67	119	152.97	27.70
FR	4200	1969	0.47	166	75.28	25.30
GB	4520	2287	0.51	151	142.39	29.93
HU	3194	2020	0.63	153	60.22	20.88
IE	4358	2633	0.60	121	87.31	36.02
IL	3230	2511	0.78	117	32.69	27.61
IT	2778	973	0.35	130	60.35	21.37
LT	4470	2109	0.47	136	26.82	32.87
NL	3537	1845	0.52	95	141.31	37.23
NO	3041	1628	0.54	96	70.92	31.68
PL	2706	1903	0.70	170	37.04	15.92
PT	3040	2152	0.71	77	86.51	39.48
RU	3772	2490	0.66	301	26.54	12.53
SE	3750	1863	0.50	91	465.12	41.21
SI	2250	1261	0.56	54	115.57	41.67
SK	2506	1853	0.74	138	41.41	18.16
UA	3692	2184	0.59	243	24.33	15.19
XK	2282	1295	0.57	72	55.40	31.69

n: Total sample size
n_r: Completed interviews
$\frac{n_r}{n}$: ~ response rate
n_{int}: Number of interviewers
$\bar{attempts}$: Attempts per interviewer
$\bar{cases}$: Cases per interviewers

2.2 기존 사회조사 분석

2015년까지 사회조사는 무응답이 발생된 경우, 조사현장에서 조사원이 직접 인접한 가구로 표본대체(substitution)를 실시하는 방식을 사용했다. 본 연구를 위해 원가구와 대체가구 간 가구 특성에 차이가 있는지 여부를 확인한 결과, 원가구의 자가 비율(62.8%)이 대체가구에 비해 높고, 전·월세 비율은 대체가구가 더 높은 것으로 확인되었다.

또한 이는 최종가구에 영향을 미쳐 자가의 비율을 낮추는 효과를 가져왔다. 여기서 원가구는 대체함 없이 당초 추출된 가구를, 대체가구는 무응답에 따라 조사하지 못하여 대체된 가구를 말하며, 최종가구는 조사가 완료된 가구로 원가구와 대체가구를 합한 것을 의미한다.

<표 2.3> 원가구와 대체가구의 가구 특성 비교(2015년)

단위 : 가구, %

		원가구		대체가구		최종가구	
			비율		비율		비율
주택유형	단독	5,420	42.16	2,425	42.40	7,845	42.23
	아파트	5,845	45.46	2,459	43.00	8,304	44.70
	연립/다세대	1,453	11.30	718	12.55	2,171	11.69
	기타주택	139	1.08	117	2.05	256	1.38
점유형태	자가	8,073	62.79	3,355	58.66	11,428	61.52
	전세	1,731	13.46	816	14.27	2,547	13.71
	월세	2,135	16.61	1,140	19.93	3,275	17.63
	사글세	333	2.59	158	2.76	491	2.64
	무상	585	4.55	250	4.37	835	4.50
가구소득	<100만	3,037	23.62	1,386	24.24	4,423	23.81
	<200만	2,851	22.17	1,289	22.54	4,140	22.29
	<300만	2,510	19.52	1,084	18.95	3,594	19.35
	<400만	1,716	13.35	772	13.50	2,488	13.39
	<500만	1,156	8.99	507	8.87	1,663	8.95
	<600만	653	5.08	261	4.56	914	4.92
	<700만	371	2.89	152	2.66	523	2.82
	<800만	563	4.38	268	4.69	831	4.47
가구원수	1인	3,418	26.58	1,592	27.84	5,010	26.97
	2인	4,012	31.20	1,772	30.98	5,784	31.14
	3인	2,450	19.06	1,053	18.41	3,503	18.86
	4인+	2,977	23.15	1,302	22.77	4,279	23.04
합계		12,857	100.00	5,719	100.00	18,576	100.00

주요 문항에 대해 원가구와 대체가구 간 응답 특성에 차이가 있는지 여부를 확인한 결과, 대부분 응답 결과에는 별다른 차이가 없으나, 주관적 만족감 및 고용의 안정성 등에서 유의미한 차이 발생함을 볼 수 있다.

<표 2.4> 원가구와 대체 가구의 주요 응답 특성 비교_ 개인사항 등(2015년)

단위 : 가구, %

조사부문	문항	항목	원가구		대체가구		최종가구	
개인사항 (가구주)	주관적 만족감	매우 만족	1,083	8.4	515	9.0	1,598	8.6
		약간 만족	2,809	21.8	1,185	20.7	3,994	21.5
		보통	6,319	49.1	2,824	49.4	9,143	49.2
		약간 불만족	2,045	15.9	949	16.6	2,994	16.1
		매우 불만족	601	4.7	246	4.3	847	4.6
복지	전반적 생활여건	많이 좋아짐	843	6.6	420	7.3	1,263	6.8
		약간 좋아짐	4,631	36.0	2,016	35.3	6,647	35.8
		변화 없음	5,630	43.8	2,509	43.9	8,139	43.8
		약간 나빠짐	1,340	10.4	599	10.5	1,939	10.4
		많이 나빠짐	390	3.0	164	2.9	554	3.0
	장애인 차별 정도	전혀 없다	420	3.3	153	2.7	573	3.1
		별로 없다	4,718	36.7	2,053	35.9	6,771	36.5
		조금 심하다	6,254	48.6	2,839	49.6	9,093	49.0
		매우 심하다	1,465	11.4	674	11.8	2,139	11.5
사회참여	기부여부	현금 기부	3,523	27.4	1,483	25.9	5,006	26.9
		물품 기부	215	1.7	105	1.8	320	1.7
		모두 기부	450	3.5	227	4.0	677	3.6
		기부 안함	8,669	67.4	3,904	68.3	12,573	67.7
	계층의식	상상	61	0.5	32	0.6	93	0.5
		상하	243	1.9	88	1.5	331	1.8
		중상	2,346	18.2	1,041	18.2	3,387	18.2
		중하	4,460	34.7	2,006	35.1	6,466	34.8
		하상	3,302	25.7	1,406	24.6	4,708	25.3
		하하	2,445	19.0	1,146	20.0	3,591	19.3
	계층이동	매우 높다	269	2.1	113	2.0	382	2.1
		비교적 높다	2,668	20.8	1,246	21.8	3,914	21.1
		비교적 낮다	5,252	40.8	2,307	40.3	7,559	40.7
		매우 낮다	2,565	20.0	1,096	19.2	3,661	19.7
		모르겠다	2,103	16.4	957	16.7	3,060	16.5
합계			12,857	100.0	5,719	100.0	18,576	100.0

<표 2.5> 원가구와 대체가구의 주요 응답 특성 비교_여가 부문 등(2015년)

단위 : 가구, %

조사부문	문항	항목	원가구		대체가구		최종가구	
여가	여가활용 만족도	매우 만족	640	5.0	294	5.1	934	5.0
		약간 만족	2,351	18.3	1,043	18.2	3,394	18.3
		보통	6,175	48.0	2,823	49.4	8,998	48.4
		약간 불만족	2,747	21.4	1,175	20.5	3,922	21.1
		매우 불만족	944	7.3	384	6.7	1,328	7.1
소득과 소비	소득 만족도	매우 만족	159	1.2	81	1.4	240	1.3
		약간 만족	1,176	9.1	517	9.0	1,693	9.1
		보통	5,097	39.6	2,265	39.6	7,362	39.6
		약간 불만족	4,497	35.0	1,957	34.2	6,454	34.7
		매우 불만족	1,601	12.5	748	13.1	2,349	12.6
	소비생활 만족도	매우 만족	183	1.4	113	2.0	296	1.6
		약간 만족	1,469	11.4	645	11.3	2,114	11.4
		보통	6,134	47.7	2,724	47.6	8,858	47.7
		약간 불만족	3,807	29.6	1,690	29.6	5,497	29.6
		매우 불만족	1,241	9.7	536	9.4	1,777	9.6
노동	고용의 안정성	매우 많이 느낀다	1,418	11.0	628	11.0	2,046	11.0
		느끼는 편이다	3,936	30.6	1,842	32.2	5,778	31.1
		느끼지 않는 편이다	3,073	23.9	1,280	22.4	4,353	23.4
		전혀 느끼지 않는다	823	6.4	339	5.9	1,162	6.3
	일자리 만족도	매우 만족	227	1.8	110	1.9	337	1.8
		약간 만족	1,397	10.9	583	10.2	1,980	10.7
		보통	3,085	24.0	1,435	25.1	4,520	24.3
		약간 불만족	1,103	8.6	495	8.7	1,598	8.6
		매우 불만족	236	1.8	82	1.4	318	1.7

표본가구를 현장에서 대체한 경우, 주택유형 및 점유형태 등 가구특성에서 보다 조사하기 쉬운 가구로 조사가 집중되는 선택편향(selection bias)이 발생함을 확인할 수 있다.

또한 특정 계층에 체계적 오차(systematic error)함에 따라 무응답 편향의 원인이 되며, 전체적인 응답결과에 유의미한 영향을 미치고 있음을 확인할 수 있었다. 이러한 이유로 학계 및 전문가들은 표본대체에 대한 문제점이 오랫동안 논란이 되어 왔다(Kish, 1965; Little and Rubin, 1987; Sharon, 2010). 미국 및 유럽 등 조사 선진 국가 대부분은 대체방법을 조사에서 사용하고 않는다.

여기서 현장 대체법(substitution)[2]에 대해 살펴 볼 필요가 있다. 현장 대체법은 전화 설문조사 경우에는 할당된 목표표본을 채우기 위해 어디에서나 일반적으로 사용하고 있는 방법이나, 가구 표본조사에서는 사용은 피하고 있다. 이는 표본단위에서 불응이나 비접촉이 발생되는 단위 무응답(unit nonresponse)을 대체하는 가장 편리한 방법으로, 일반적으로 무응답자와 유사한 조사 특성을 갖는 대상으로 교체하는 것이 원칙이다. 일례로 통계청 가계동향조사는 현장 표본대체는 매우 제한적이지만 대체를 위해서는 가구주 직업, 가구원수 등이 동일한 경우에 허용된다. 다만, 일반 가구표본 조사에서 무응답자 특성을 안다는 것은 현실적으로 어렵고, 만약 무응답자 특성을 여타 정보로부터 갖고 있다면 대체도 매우 효과적일 것이다.

현장 대체법(substitution)은 랜덤대체(random substitution)과 특정화 대체(specially designated substitution)로 구분할 수 있으며, 랜덤대체는 설계단계에서 층(strata)과 지역 내 무응답 대체를 위해 하나이상의 예비표본을 사전에 동일하게 설계한 후 확률적으로 선택하여 대체하는 방법이다. 특정화 대체는 기본적으로 랜덤대체와 유사하나 교체대상은 무응답자와 지리적으로 인접한 이웃, 표본추출틀에서 인접한 명부 혹인 특성을 갖는 단위로 교체하는 방법으로 조사기관에서 많이 사용하는 방법이다. 다만, 대체방법은 무응답 편향(nonresponse bias)을 감소하는데 크게 도움을 주지 못한다(변종석과 한혜은, 2011).

중요한 점은 다른 방법 보다 더욱 효과적인 대리값(proxy values)을 제공할 수 있는 지가 중요하며 특정 방법에 의한 현장 대채법은 평균대체(mean imputation) 보다 더 나은 대리값을 제공할 수 있으며, 무응답자에 대한 유용한 정보가 존재한다면 현장 대체법이 대체 보다 더 나은 값을 제공할 수 있다. 모집단에 비해 표본이 상대적으로 작은 단위로 층화 된 경우에 표본교체는 매우 적절하다(Chapman, 1979). 학교, 병원과 같은 시설에 대한 조사로 조사단위에 대한 층화정보가 정확하고, 표본당 조사비용이 비싸 표본크기를 작게 설계한 경우에 현장 대체법이 적절할 수 있다. 즉 병원, 학교는 무응답 시설과 유사한 특성의 교체가 가능하다. 다만, 대규모 조사나 무

2) 현장 대체법(substitution)은 앞장 표본대체로 사용했으나, 항목 무응답 처리방법에서 대체방법(impuatation)로 혼동될 수 있어 이 장에서는 현장 대체법으로 사용함

응답 단위에 대한 정보가 상대적으로 부족한 경우에는 현장 대체법은 가중층 대체방법 보다 편향을 감소하거나 정확한 분산추정 등 어떤 유용한 개선 결과를 제공하지 못하다.

현장 대체법(substitution)의 장단점을 다시 정리하면, 설계과정에 목표한 크기의 표본크기를 유지 할 수 있으며 특히 원표본의 추출확률과 거의 동일한 확률로 대체표본을 추출하는 경우는 원표본이 자체가중표본(self-weighting sample)이라면 최종표본도 자체가중표본일 가능성 높다. 또한 표본크기를 증가시켜 주므로 추정치의 분산에 대한 감소효과가 나타난다. 이러함에도 현장대체법은 원표본에 대한 조사 참여 및 자료수집에 대한 추가적인 노력을 하는 것에 비해 현장 표본대체가 매력적이지 못하다는 점이며, 연구자가 예비표본 단위(backup unit)를 원표본 처럼 인지하여 조사결과에 있어 편향을 크게 하는 원인이 된다.

또한 응답률 계산시 사용된 대체의 수준을 간과하거나 마치 원표본 처럼 응답률을 계산하는 경향이 있다. 응답률을 과대 추정하여 무응답 편향(bias)을 과소추정하고, 응답하기 쉬한 대상으로 조사가 집중되거나 응답 전환자를 무시하게 되어 구조적인 편향(systematic bias)을 가져올 수 있다. 또한 이는 분산의 과소추정(underestimation)의 원인이 될 수 있다(김규성, 2000).

2.3 2016년 사회조사 표본설계

2016년 사회조사는 사회적 관심사항, 삶의 질에 관한 사항 등 사회구성원의 주관적 관심사를 파악하여 사회개발정책의 기초 자료로 제공함을 목적으로 한다.

동 조사는 연간으로 주기적인 조사로 만 13세이상 모든 인구를 대상으로 하고 있다. 특히 매년 조사내용을 달리하여 조사하고 있으며 짝수 해에는 보건, 교육, 안전, 가족, 환경부문을 홀수 해에는 복지, 사회참여, 문화와 여가, 소득과 소비, 노동부문 등으로 구분하여 조사를 시행하고 있다.

조사방법은 일반적으로 조사원에 의한 가구방문 면접조사를 원칙으로 하고 있지만, 일부 원하는 가구에 대해서는 인터넷조사도 허용하고 있다. 다만 조사항목이 70문항 이상으로 방대하며 가구원 모두를 조사해야하는 특성상 인터넷 조사 보다는 현실적인 면접조사 방식을 취하고 있다.

조사결과에 대한 공표는 전국단위에서는 모든 항목에 대해서 다루고 있지만, 시도 단위 공표는 표본오차 등을 고려하여 제한적으로 발표하고 있다. 공표를 위한 주요 특성은 성별, 연령별, 교육정도별, 직업별, 산업별, 혼인상태별, 활동상태별, 종사상지위별, 가구소득별, 주관적 만족정도별 등을 살펴 볼 수 있다.

동조사의 모집단(population)은 조사시점 전국의 만 13세이상 모든 가구원이나, 조사를 위한 조사모집단(survey population)은 조사하기 어려운 기숙시설, 특수사회시설, 외국인거주지역에 거주하는 사람을 제외한 일반가구 내의 가구원을 대상으로 하고 있다.

표본추출틀(sampling frame)은 2010년 인구주택총조사 10% 표본조사구와 최근까지 신축된 아파트 조사구를 더하여 4천9백 조사구를 표본추출틀로 구축하여 표본설계를 진행했다. 앞서 살펴본 것처럼, 조사하기 어려운 섬지역이나 외국인 거주지역은 표본추출틀에서 실제 조사에서는 제외하였다.

<표 2.6> 표본추출틀의 포함률

(단위: 개, 가구, %)

조사모집단 일반조사구(1,2,A)		표본추출틀 일반조사구(1,A)			
조사구	가구	조사구	포함률	가구	포함률
49,637	2,782,095	49,331	99.4	2,765,763	99.4

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 분류지표[3](sub-stratification)로 구성하여 내재적 층화(implicit stratification)를 함으로써 층화효과를 높일 수 있었다(통계청, 2016b).

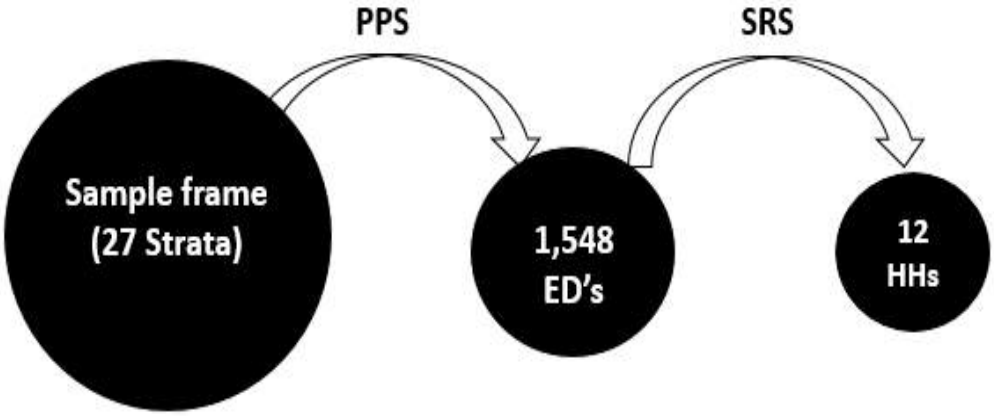
<표 2.7> 분류지표 적용 내역

구분	적용범위	사용지표	비 고
1차 분류	전국	주택유형	- 단독주택, 아파트, 기타주택 중 구성비가 높은 순
2차 분류	읍면부	농가 비율	- 동부에는 적용되지 않음
3차 분류 4차 분류	시도 (27개층)	유배우 비율 1인가구 비율 60세이상 인구비율 자가비율	- 응답성향에 최우선순위로 영향을 미치는 변수 - 차순위에 해당하는 분류지표를 각 25개 층별로 반영

※ 2012년 사회조사 표본설계 결과임

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.

3) 층화 후 조사구 특성을 반영하는 분류지표(보조층화지표)에 따라 조사구를 정렬한 후 계통 추출 방법으로 조사구를 추출 하면, 보조층화 지표도 사실상 층화 효과를 얻을 수 있음



<그림 2.2> 층화 2단계 집락추출

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). 다시 강조 하자면, Statistics Canada(2003)에서는 표본규모를 결정하기 위한 조건으로 우선 정도(precision)에 미치는 요소를 모집단 관심변수 변동성, 모집단 크기, 표본설계 디자인 효과, 응답률을 들고 있다. 우리나라에서 목표표본을 고려하여 표본을 추가하여 설계하는 것을 과대추출(over-sampling)로 사용하고 있으나, 일반적으로 과대추출은 특정 집단 즉, 고소득 층(high-income class) 또는 발생빈도가 희소한 집단에 대해 표본수를 추가 하여 추출한 경우로 한정하여 사용하고 있어 이에 대한 명확한 구별이 요구된다.

<표 2.8> 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

2.4 무응답률 및 파라데이터 분석

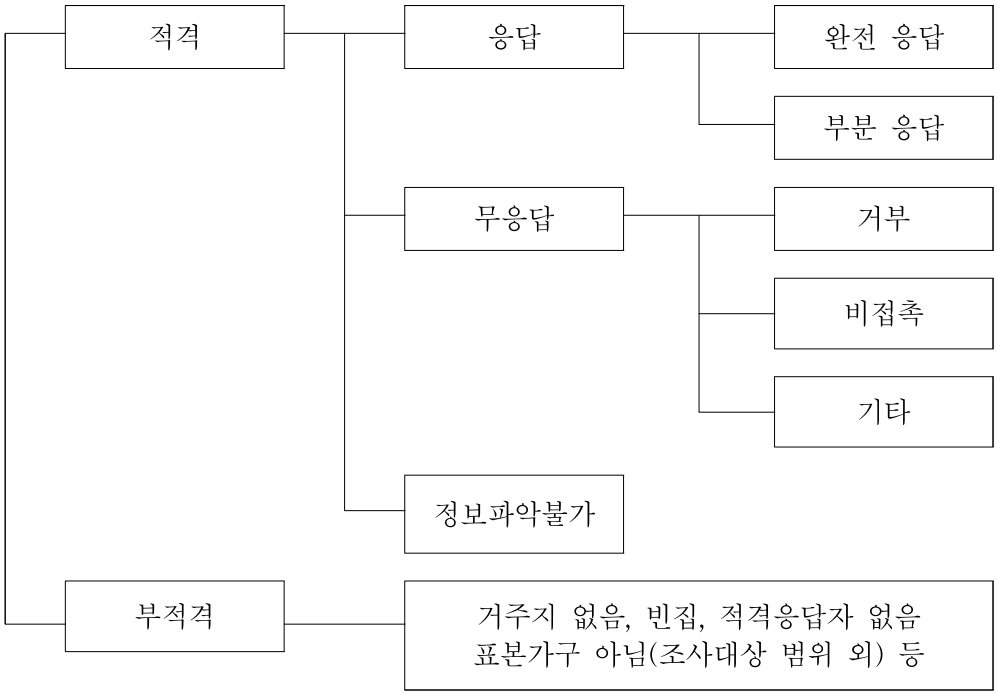
2016년 사회조사는 빈집 등 부적격을 제외한 적격 표본 22,642가구 중 전체 무응답률은 18.6%로 나타났다. 무응답은 응답 거부(refusal), 조사기간에 표본에 접촉할 수 없어서 발생되는 비접촉(non-contact) 및 건강상의 이유, 언어소통 문제, 사망 등의 사유로 응답을 얻지 못한 조사불가(inability)로 구분할 수 있다(AAPOR, 2009). 보다 세부적인 내용은 APPOR(2009)를 참조하면 좋겠다. 무응답 가구의 특성을 분석함으로 편향 탐색(bias detection) 및 무응답 조정 등에 활용이 가능하다.

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 파라데이터는 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 무응답은 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). 당초 무응답에 빈집 등 일부 부적격을 포함하여 분류했으나, APPOR(2009) 기준에 맞게 재분류 작업을 했음을 밝힌다.

<표 2.9> 유형별 응답형태(AAPOR, 2009)



2016년 사회조사 전체 응답률은 81.4%이고, 무응답률은 18.6%로 나타났다. 무응답은 응답 거부가 12.3%, 비접촉이 5.7%, 건강상의 이유 등 기타 무응답이 0.6%를 보였다. 무응답은 동부 지역, 아파트 거주 가구 및 방문횟수가 증가할수록 높게 나타남을 확인할 수 있다.

<표 2.10> 가구 특성별 응답 및 무응답 가구수 및 구성비

(단위 : 가구, %)

		전체	응답		무응답		응답 거부		비접촉		기타	
				%		%		%		%		%
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

지역별로는 동부는 78.8%, 읍면부는 88.9%를 보였다. 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 가구 및 이동이 잦은 1인 가구 등의 이유로 주간 부재율이 높아 조사기간 동안 접촉이 불가능한 비접촉률도 6.3%로 읍면부 지역 3.9%에 비해 상대적으로 높게 나타났다. 반면 방문횟수는 동부는 2.81회로 읍면부 2.14회에 비해 높았다.

<표 2.11> 지역별 응답구분

(단위 : 가구, %, 회)

	소계			응답			응답 거부			비접촉			기타		
	가구	비율	방문횟수	가구	비율	방문횟수	가구	비율	방문횟수	가구	비율	방문횟수	가구	비율	방문횟수
전국	22,642	100.0	2.64	18,422	81.4	2.43	2,788	12.3	3.58	1,287	5.7	3.57	145	0.6	2.59
동부	16,973	100.0	2.81	13,380	78.8	2.59	2,434	14.3	3.61	1,067	6.3	3.67	92	0.5	2.90
읍면부	5,669	100.0	2.14	5,042	88.9	2.00	354	6.2	3.40	220	3.9	3.12	53	0.9	2.04

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). 특히 대형아파트의 응답거부는 20.0%를 초과하여 단독주택 거주 가구에 비해 2.7배가 높게 나타났다. 한편 주택유형별로 조사기간 동안 비접촉한 비율은 큰 차이가 없음을 발견할 수 있다.

조사를 위해 방문횟수는 단독주택에 비해 아파트와 기타주택에서 상대적으로 높으며, 특히 응답거부와 비접촉이 발생한 경우는 응답한 가구에 비해 방문횟수가 상당히 높아짐을 알 수 있다.

<표 2.12> 주택유형별 응답구분

(단위 : 가구, %, 회)

	소계			응답			응답거부			비접촉			기타		
	가구	비율	방문횟수	가구	비율	방문횟수	가구,	비율	방문횟수	가구	비율	방문횟수	가구	비율	방문횟수
총계	22,642	100.0	2.64	18,422	81.4	2.43	2,788	12.3	3.58	1,287	5.7	3.57	145	0.6	2.59
단독주택	9,223	100.0	2.36	7,963	86.3	2.20	699	7.6	2.59	482	5.2	2.76	79	0.9	2.65
아파트	10,557	100.0	2.82	8,201	77.7	2.60	1,686	16.0	3.60	629	6.0	3.66	41	0.4	3.05
중소형	10,159	100.0	2.81	7,908	77.8	3.51	1,605	15.8	3.60	606	6.0	3.64	40	0.4	3.63
대형	398	100.0	2.98	293	73.6	3.41	81	20.4	3.67	23	5.8	3.43	1	0.3	3.71
기타주택*	2,862	100.0	2.86	2,258	78.9	2.23	403	14.1	3.05	176	6.1	3.00	25	0.9	2.96

* 연립/다세대주택 및 기타(비주거용 건물 내 주택, 오피스텔 등) 포함

앞서 동부거주 지역, 아파트 거주 가구의 응답거부와 비접촉이 높게 나타난 것을 살펴보았다. 지역과 주택유형을 2차원으로 살펴보면, 응답거부는 짐작할 수 있듯이 동부지역 아파트 거주 가구가 16.1%로 가장 높게 나타났고, 반면 읍면부 지역 단독주택 거주 가구는 3.1%로 가장 낮은 응답 거부율을 보였다(<표 2.13>).

또한 읍면부 지역 아파트 거부율도 14.8%로 나타나 아파트에 거주하는 가구의 협조도가 낮음을 볼 수 있다. 비접촉 비율은 동부지역이 읍면부 지역 보다 높게 나타났으나, 동부 내 및 읍면부 내 주택유형에서는 차이가 상대적으로 낮은 것을 볼 수 있다.

방문횟수는 응답한 가구에 비해 무응답 가구가 높게 나타났으며, 이를 세분화하면 응답거부는 동부 지역이나 읍면부 지역의 오피스텔 등에 거주하고 있는 기타가구의 방문횟수가 상대적으로 높게 나타났다. 만날 수 없는 가구는 동부지역의 기타주택에서 가장 높게 나타나 상대적으로 이동이 많은 가구에 방문횟수가 높아짐을 확인할 수 있다.

<표 2.13> 주택유형 및 동·읍면부별 응답구분

(단위 : 가구, %, 회)

구분		소계			응답			응답거부			비접촉			기타		
		가구	비율	방문횟수	가구	비율	방문횟수	가구	비율	방문횟수	가구	비율	방문횟수	가구	비율	방문횟수
전국		22,642	100.0	2.64	18,422	81.4	2.43	2,788	12.3	3.58	1,287	5.7	3.57	145	0.6	2.59
	단독주택	9,223	100.0	2.36	7,963	86.3	2.20	699	7.6	2.59	482	5.2	2.76	79	0.9	2.65
	아파트	10,557	100.0	2.82	8,201	77.7	2.60	1,686	16.0	3.60	629	6.0	3.66	41	0.4	3.05
	기타주택*	2,862	100.0	2.86	2,258	78.9	2.23	403	14.1	3.05	176	6.1	3.00	25	0.9	2.96
동부		16,973	100.0	2.81	13,380	78.8	2.59	2,434	14.3	3.61	1,067	6.3	3.67	92	0.5	2.90
	단독주택	5,282	100.0	2.68	4,331	82.0	2.49	578	10.9	3.59	338	6.4	3.61	35	0.7	2.71
	아파트	9,226	100.0	2.84	7,131	77.3	2.61	1,489	16.1	3.61	569	6.2	3.68	37	0.4	3.14
	기타주택*	2,465	100.0	2.94	1,918	77.8	2.75	367	14.9	3.63	160	6.5	3.76	20	0.8	2.80
읍면부		5,669	100.0	2.14	5,042	88.9	2.00	354	6.2	3.40	220	3.9	3.12	53	0.9	2.04
	단독주택	3,941	100.0	1.93	3,632	92.2	1.85	121	3.1	3.16	144	3.7	2.93	44	1.1	1.84
	아파트	1,331	100.0	2.69	1,070	80.4	2.50	197	14.8	3.51	60	4.5	3.55	4	0.3	2.25
	기타주택*	397	100.0	2.33	340	85.6	2.13	36	9.1	3.64	16	4.0	3.25	5	1.3	3.60

* 연립/다세대주택 및 기타(비주거용 건물 내 주택, 오피스텔 등) 포함

최종 방문횟수별로 무응답을 보면, 방문횟수가 높아질수록 예상하듯 응답을 얻기가 더욱 어려워짐도 확인할 수 있다.

<표 2.14> 최종 방문회차별 응답구분

(단위 : 가구, %)

	소계		응답		응답거부		비접촉		기타	
	가구	비율	가구	비율	가구	비율	가구	비율	가구	비율
	22,642	100.0	18,422	81.4	2,788	12.3	1,287	5.7	145	0.6
1회	4,784	100.0	4,627	96.7	71	1.5	43	0.9	43	0.9
2회	5,436	100.0	5,264	96.8	109	2.0	48	0.9	15	0.3
3회	5,602	100.0	4,500	80.3	734	13.1	322	5.7	46	0.8
4회+	6,820	100.0	4,031	59.1	1,874	27.5	874	12.8	41	0.6

최종 방문횟수가 4회 이상인 경우, 동부의 응답 거부율이 27.9%로 읍면부 응답 거부율 24.2%에 비해 높게 나타나, 동부 가구의 조사 비협조율이

높음을 알 수 있다. 4회 이상 방문에도 만날 수 없는 가구는 동부 지역이 12.9%로 읍면부 12.1% 보다 다소 높게 나타났다.

<표 2.15> 방문회차 및 동·읍면부별 응답구분

(단위 : 가구, %)

		소계		응답		응답거부		비접촉		기타	
		가구	비율	가구	비율	가구	비율	가구	비율	가구	비율
동부		16,973	100.0	13,380	78.8	2,434	14.3	1,067	6.3	92	0.5
	1회	2,770	100.0	2,680	96.8	55	2.0	17	0.6	18	0.6
	2회	3,733	100.0	3,619	96.9	83	2.2	26	0.7	5	0.1
	3회	4,482	100.0	3,571	79.7	623	13.9	251	5.6	37	0.8
	4회+	5,988	100.0	3,510	58.6	1,673	27.9	773	12.9	32	0.5
읍면부		5,669	100.0	5,042	88.9	354	6.2	220	3.9	53	0.9
	1회	2,014	100.0	1,947	96.7	16	0.8	26	1.3	25	1.2
	2회	1,703	100.0	1,645	96.6	26	1.5	22	1.3	10	0.6
	3회	1,120	100.0	929	82.9	111	9.9	71	6.3	9	0.8
	4회+	832	100.0	521	62.6	201	24.2	101	12.1	9	1.1

응답가구의 조사방문 시간은 회차별 분포가 다소 상이한 것으로 확인할 수 있으며, 1회차에서 응답한 가구는 16~19시 이전(29.3%)에 응답한 비율이 가장 높았으며, 전체 응답가구의 응답 시점을 기준으로 볼 때도 16~19시 이전(28.4%)에 응답한 비율이 가장 높은 것으로 확인된다. 다만, 읍면부에서는 방문시간이 동부지역에 비해 빠르게 진행되고 있다.

<표 2.16> 응답가구의 조사방문 시간

(단위 : 가구, %)

	1회차 응답가구						최종 회차 응답가구					
	전국	비율	동부	비율	읍면부	비율	전국	비율	동부	비율	읍면부	비율
전체	4,627	100.0	2,680	100.0	1,947	100.0	18,422	100.0	13,380	100.0	5,042	100.0
10시 이전	122	2.6	52	1.9	70	3.6	512	2.8	350	2.6	162	3.2
10시~13시	1,047	22.6	511	19.1	536	27.5	4,603	25.0	3,158	23.6	1,445	28.7
13시~16시	1,118	24.2	564	21.0	554	28.5	4,686	25.4	3,245	24.3	1,441	28.6
16시~19시	1,355	29.3	828	30.9	527	27.1	5,230	28.4	3,847	28.8	1,383	27.4
19시~22시	960	20.7	707	26.4	253	13.0	3,315	18.0	2,716	20.3	599	11.9
22시 이후	25	0.5	18	0.7	7	0.4	76	0.4	64	0.5	12	0.2

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, 최근 개인정보 노출 꺼려 및 사생활 보호의식 강화 등에 따른 최근 사회현상이 반영되어 사생활 노출기피가 높은 비율을 차지했다.

무응답 사유를 지역별로 보면, 읍면부의 비접촉율이 상대적으로 높게 나타나는데 이는 연령층이 높은 노인들이 가족 등의 일시적인 방문으로 부재가 많은 일반적인 거주 현상을 살펴 볼 수 있다.

<표 2.17> 무응답 사유별 가구수 및 구성비

(단위 : 가구, %)

		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

방문회차에 따른 주택유형별 불응가구를 살펴보면, 대형아파트 거주 가구는 1회차 불응이 4회차 이상까지 지속적으로 높게 이어지는 것을 볼 수 있다.

이는 조사에 불응 전환(refusal conversion) 낮아 최종까지도 불응이 이어진 것을 보면 대형아파트 조사가 기타 주택에 비해 어렵다는 것을 짐작 할 수 있다.

<표 2.18> 응답 거부 가구 최종 방문횟수

(단위: 가구, %)

		응답 거부									
		전체		1회차		2회차		3회차		4회차 이상	
		가구	비율	가구	비율	가구	비율	가구	비율	가구	비율
전국		2,788	100.0	71	2.5	109	3.9	734	26.3	1,874	67.2
단독주택		699	100.0	21	3.0	26	3.7	225	32.2	427	61.1
아파트	중소형	1,605	100.0	42	2.6	70	4.4	383	23.9	1,110	69.2
	대형	81	100.0	4	4.9	0	0.0	17	21.0	60	74.1
기타주택*		403	100.0	4	1.0	13	3.2	109	27.0	277	68.7

* 연립/다세대주택 및 기타(비주거용 건물 내 주택, 오피스텔 등) 포함

불응사유를 세부적으로 보면, 사생활노출기피 35.5%, 바쁘거나 귀찮아서 28.7%, 면접회피 22.8% 순으로 나타나고 있다. 개인정보 노출을 꺼리고 사생활 보호의식이 강화되는 등의 최근 사회현상이 반영되어 사생활 노출기피가 높은 비율을 차지했다. 다만, 단독주택 거주 가구는 면접회피가 바쁘거나 귀찮아서 보다 높게 나타났다.

<표 2.19> 불응사유 현황

(단위 : 가구,%)

		전체		사생활 노출기피		가정사정		바쁘거나 귀찮아서		정부불신		면접회피	
		가구	비율	가구	비율	가구	비율	가구	비율	가구	비율	가구	비율
합계		2,788	100.0	989	35.5	74	2.7	801	28.7	289	10.4	635	22.8
동부		2,434	100.0	865	35.5	62	2.5	689	28.3	250	10.3	568	23.3
읍면부		354	100.0	124	35.0	12	3.4	112	31.6	39	11.0	67	18.9
단독주택		699	100.0	248	35.5	16	2.3	179	25.6	70	10.0	186	26.6
아파트	중소형	1,605	100.0	570	35.5	42	2.6	487	30.3	172	10.7	334	20.8
	대형	81	100.0	29	35.8	4	4.9	22	27.2	9	11.1	17	21.0
기타주택*		403	100.0	142	35.2	12	3.0	113	28.0	38	9.4	98	24.3

* 연립/다세대주택 및 기타(비주거용 건물 내 주택, 오피스텔 등) 포함

불능을 세분하면, 장기출타 등에 의한 비접촉 비율이 89.9%로 가장 높게 차지하고 다음으로 건강상 문제 4.6%, 언어소통문제 4.3% 순으로 나타났다. 이는 조사기간이 짧아 이동이 많은 1인가구나 맞벌이 가구를 만나지 못한 데 기인한 것으로 해석된다(<표 2.20>).

<표 2.20> 불능사유 현황

(단위: 가구, %)

	전체		비접촉		건강상 문제		언어소통 문제		기타	
	가구	비율	가구	비율	가구	비율	가구	비율	가구	비율
합계	1,432	100.0	1,287	89.9	66	4.6	62	4.3	17	1.2
단독주택	561	100.0	482	85.9	38	6.8	38	6.8	3	0.5
아파트	670	100.0	629	93.9	19	2.8	13	1.9	9	1.3
기타주택*	201	100.0	176	87.6	9	4.5	11	5.5	5	2.5

* 연립/다세대주택 및 기타(비주거용 건물 내 주택, 오피스텔 등) 포함

2.5 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; 이승희, 2010).

$B(\bar{y}_r) = (\frac{m}{n})(\bar{y}_r - \bar{y}_m)$. 여기서 n은 표본 가구, r은 표본에서 응답 가구, m은 무응답 가구를 의미한다. 파라데이터 정보로부터 지역, 주택유형 및 방문횟수 등 부차 그룹에서 응답과 무응답 비율의 차이가 유의미했으며, 이들의 유의성 검정 결과 응답가구와 무응답 가구 특성이 명확히 달랐다(<표 2.21>).

동부 지역, 아파트, 방문횟수가 높은 경우에 무응답 가구의 비중이 응답가구 보다 높게 나타나는데, 이 변수들은 무응답 조정에 활용한다.

<표 2.21> 무응답 유형별 차이 및 편향

		응답가구(A)		무응답가구(B)		차이(C) ($=A-B$)	편향(D) ($=\frac{m}{n}*C$)	χ^2 $Value$ ($Pr > Chisq$)
			비중		비중			
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 파라데이터를 보조정보로 로지스틱 회귀분석의 응답성향을 통해 얻어진다. 오즈비가 1보다 크면 기준이 되는 그룹보다 응답성향이 높고 반면 1보다 작으면 기준그룹 보다 응답성향이 낮음을 나타낸

다(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 방문횟수이 무응답에 영향력 있는 변수로 파악됐다.

<표 2.22> 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

사후층화(post-stratification)는 표본의 결합분포를 모집단의 결합분포에 맞게 벤치마킹하는 것으로 사후층화 보정은 표본의 포함률(coverage) 보정 및 조사 추정값의 정도를 보완하는 역할을 한다. 본 내용에서 무응답 편향에 대한 탐색은 조정전 설계가중값에 의한 외부 모집단 정보에 사후보정을 수행한 사후층화 가중값과 비교하는 방식을 취한다(정구현, 2014).

무응답 편향 탐색을 위한 방법으로 가중값의 효과를 평가하기 위해 설계가중값(w_1)로 사후층화 가중값(w_2)를 나눈 값을 조정계수(adjustment factor)라 이름하며 식(2.1)과 같이 계산한다.

$$\textit{Adjustment Factor} = \frac{w_{2h}}{w_{1h}} \tag{2.1}$$

여기서 h는 분석을 위한 층을 의미하며 평가기준(criteria)은 1.33보다 크면 표본이 크게 과소표현된 것으로, 0.75보다 작으면 표본이 크게 과대표현

된 것으로 평가한다(Ineke 등, 2010).

평가기준에 크게 유의한 정도는 아니지만, 방문횟수 4회 이상은 표본이 과소표현된 반면, 방문회수 1~2회는 표본이 과대표현된 것으로 나타났으며, 이러한 차이는 무응답 편향 뿐 아니라 표본설계, 표본프레임 등 다양한 원인이 있을 수 있다.

<표 2.23> 사후층화 조정계수

가구특성		조정계수	가구특성		조정계수
지역	동부	0.98	가구원수	1인	1.07
	읍면부	1.13		2인	0.93
주택유형	단독주택	1.01		3인	1.04
	아파트	0.99		4인+	0.98
	기타주택	1.00	가구주 성별	남자	0.99
방문횟수	1회	0.86		여자	1.02
	2회	0.86	가구주 연령	20-29세	1.07
	3회	0.96		30-39세	1.01
	4회+	1.31		40-49세	0.96
점유형태	자가	0.99		50-59세	1.01
	비자가	1.01		60세이상	0.98

모형에 의한 방법은 설계가중값과 사후층화 가중값을 적용하여 주요한 설명변수의 변화 여부를 확인하는 방법이다. 이를 위해 사회조사의 주요 관심변수인 소득수준을 종속변수로 하고 지역, 주택유형, 방문횟수를 독립변수로 다중회귀분석을 실시하였다. 지역은 동부와 읍면부로 구분하여 사용한다.

<표 2.24>에 따르면 설계가중값 및 사후가중값을 적용할 때 모든 독립변수가 유의미함을 알 수 있으며, 모형에는 큰 유의미한 편향을 발견할 수 없음 동부에 거주하거나, 기타주택 이외의 주택에 거주한 경우 소득수준이 높은 것을 보여준다.

모형의 변화는 표본추출과 무응답이 영향을 미치고 있음을 의미한다(Ineke et al., 2010). 모형의 결정계수(R^2) 증가는 관심변수와 사후층화 적용 후 독립변수 간에 관계가 증가됨을 의미한다. 모형의 결정계수가 증가함을 볼 수 있고, 이는 관심변수와 독립변수 간에 상관관계가 증가함을 볼 수 있

다. 다만, 편향은 총계 및 평균에 대한 추정값에 종종 심각한 영향을 미치지고 있지만, 변수간 관계에 관한 추정값은 크게 영향을 미치지 않는다(Bethlehem, 2002).

<표 2.24> 회귀모형에 의한 편향 검토

	설계가중치(w_1)				사후층화 가중치(w_2)			
	parameter estimate	SE	t-value	Pr> |t|	parameter estimate	SE	t-value	Pr> |t|
Intercept	2.47214	0.1233	20.04	<.0001	2.5183	0.1200	20.98	<.0001
동부	0.3108	0.0381	8.16	<.0001	0.2947	0.0366	8.04	<.0001
읍면부	0.0000	.	.	.	0.0000	.	.	.
단독주택	-0.68470	0.1173	-5.84	<.0001	-0.7132	0.1139	-6.26	<.0001
아파트	0.65860	0.1164	5.66	<.0001	0.6532	0.1130	5.78	<.0001
연립	-0.37600	0.120	-3.13	0.0017	-0.3678	0.1167	-3.15	0.0016
기타주택	0.0000	.	.	.	0.0000	.	.	.
방문횟수	0.2152	0.0126	17.07	<.0001	0.2018	0.0125	16.12	<.0001
R^2	0.138264				0.140828			

응답여부에 영향 미치는 추가 정보는 표본추출틀을 사용하든 행정자료로부터 얻든 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 전수집계결과로 공표하고 있다. 2016년에는 성별, 연령, 성명, 국적, 입국시기, 가구구분, 거처종류, 연면적, 대지면적, 건축연도 등 13개 항목을 포함하고 있어, 이들은 주요 보조정보로 활용할 수 있다(<표 2.26>; 통계청, 2017).

하지만 당시 표본추출틀 노후화로 표본 정보에 대한 내재된 불확실성과 행정자료 이용에는 시기적인 제약 등이 따랐다. 이를 극복하는 한 방법으로 파라데이터 정보인 지역, 주택유형, 방문횟수를 반영하여 응답성향을 산정한 후, 응답가구만으로 가구특성별 평균 응답성향의 분포를 재구성하는 방

법을 적용했다(<표 2.25>).

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

이와 같이 응답여부에 영향을 주는 변수들을 무응답 조정을 위한 변수로 구성하는 것이 바람직한데, 이는 무응답 편향을 효과적으로 제거할 수 있기 때문이다. 관련성이 약한 변수를 포함한 경우는 편향 축소가 그다지 도움이 되지 않고 분포의 안정성을 떨어뜨려 정확성을 저해할 수도 있다 (Guillaume, 2013).

<표 2.25> 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	541.801 (<.0001)		3인	0.82	
	아파트	0.81			4인+	0.82	
	기타주택	0.82		가구주 성별	남자	0.84	11.73 (.0006)
방문횟수	1회	0.98	620476 (<.0001)		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	110.34 (<.0001)		50-59세	0.83	
	비자가	0.83			60세이상	0.89	

<그림 2.3> 응답가구 특성별 평균 응답성향

<표 2.26> 행정자료를 활용한 인구총조사 공표내역

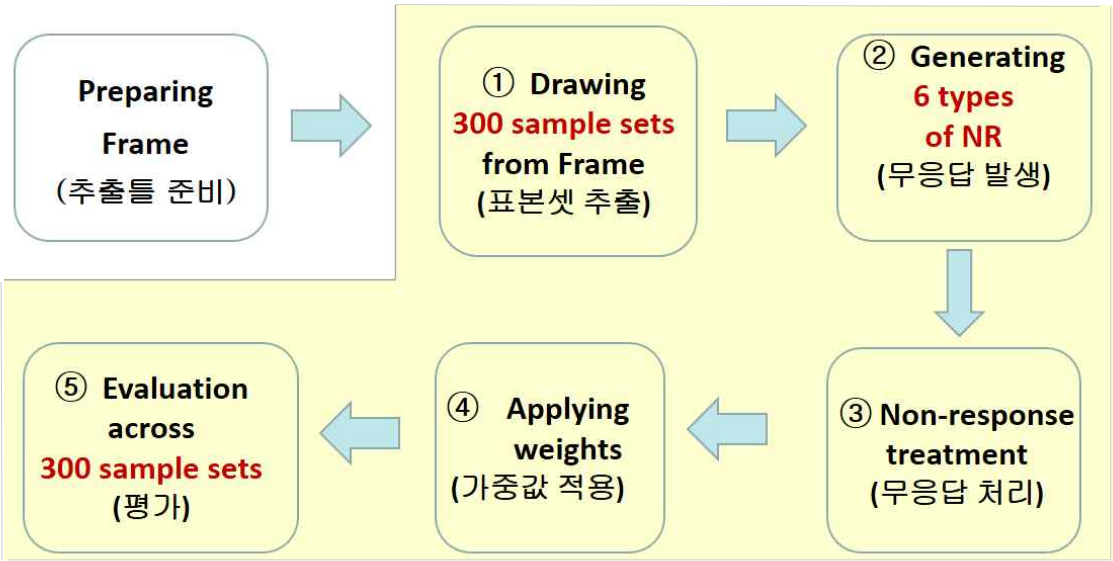
구 분	2015년 전수 항목(12)	2016년 전수항목(13)
인구	① 성명 ② 성별 ③ 나이 ④ 가구주와의 관계 ⑤ 국적 ⑥ 입국 연월 ⑦ 본관*	① 성명 ② 성별 ③ 나이 ④ 가구주와의 관계 ⑤ 국적 ⑥ 입국 연월 ⑦ 1년전 거주지 ⑧ 국적취득연도
가구	① 가구 구분	① 가구 구분
주택	① 거처의 종류 ② 주거용 연면적 ③ 건축연도 ④ 대지면적	① 거처의 종류 ② 주거용 연면적 ③ 건축연도 ④ 대지면적

* ‘본관’은 5년 주기의 항목

** 전수 신규통계항목

3. 모의실험(simulation)

무응답에 대한 정확한 정보를 조사 현장으로부터 파악하는 데는 현실적으로 어려움이 있고 확보된 제한적인 정보만으로 특정변수에 대한 무응답 발생여부(missingness)를 결정짓기는 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.



<그림 3.1> 모의실험 작업과정

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 결정

했다. 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. 원표본(full sample)은 300개 표본셋을 합하여 가중값을 부여하지 않고 계산된 특성별 비중에 해당된다.

<표 3.1> 원표본 구성

(단위 : %)

		모집단(가중전)				모집단(가중후)				원표본(가중전)			
		아파트	유배우	1인 가구	60세 이상	아파트	유배우	1인 가구	60세 이상	아파트	유배우	1인 가구	60세 이상
전국		44.5	65.6	27.8	38.1	48.2	65.3	27.5	32.2	49.5	65.6	27.5	32.2
동부	서울	38.3	64.6	25.0	31.1	37.5	62.4	27.5	29.8	37.8	62.4	27.1	27.8
	부산	46.4	59.1	29.1	43.1	49.7	60.7	27.7	39.7	49.2	56.6	28.3	37.0
	대구	50.3	67.6	23.8	35.5	51.5	66.1	25.9	33.3	55.2	68.7	23.8	29.8
	인천	50.2	64.2	22.7	28.3	53.7	64.3	24.3	25.6	53.1	66.1	22.2	25.5
	광주	66.2	64.5	26.9	27.5	70.2	64.7	28.0	24.2	72.0	63.6	27.2	23.0
	대전	52.4	66.2	27.6	29.0	52.4	65.4	29.6	26.8	52.5	66.3	29.5	23.8
	울산	52.5	69.0	23.4	31.8	55.6	69.5	23.5	28.6	55.7	68.9	23.1	25.4
	세종	100.0	84.2	12.1	12.9	100.0	78.4	20.1	15.5	100.0	78.1	19.5	14.1
	경기	56.4	71.2	21.1	24.3	56.7	69.2	23.6	23.2	55.9	70.8	21.9	22.7
	강원	59.0	67.7	28.3	34.5	60.9	64.9	33.0	31.0	63.4	70.9	27.6	29.9
	충북	56.4	66.9	26.9	33.6	57.5	63.7	31.2	32.2	60.7	64.8	27.9	32.8
	충남	58.9	66.2	25.6	23.9	57.3	61.9	30.8	21.9	57.7	63.1	28.9	19.5
	전북	62.8	70.2	23.8	36.1	62.8	66.7	28.7	34.2	64.9	72.3	24.3	31.1
	전남	63.4	64.5	29.1	37.5	71.7	65.1	28.2	31.8	72.6	64.5	28.2	29.8
	경북	45.8	66.1	30.3	30.9	51.4	65.2	31.7	28.1	51.9	68.2	29.9	26.6
	경남	52.7	67.2	25.0	32.5	55.6	66.8	26.8	30.3	54.9	68.6	25.1	26.9
	제주	27.3	61.0	30.5	35.0	32.5	64.4	26.5	31.0	34.0	58.3	27.2	28.2
읍면부	세종	49.6	66.5	28.5	38.8	48.4	54.9	43.5	39.9	47.4	64.9	33.0	37.1
	경기	41.2	68.4	25.7	43.9	53.0	70.5	23.3	35.7	56.5	67.2	23.1	32.8
	강원	16.3	61.1	39.5	56.3	27.8	64.0	31.8	50.0	27.6	63.8	36.2	50.0
	충북	23.6	64.0	34.0	57.2	38.4	65.1	31.8	46.5	40.3	62.7	31.3	47.0
	충남	24.2	61.3	35.9	53.3	35.6	65.5	31.2	43.3	38.3	61.0	33.1	43.5
	전북	8.2	55.2	43.2	71.0	17.4	61.1	35.6	60.8	16.2	58.5	40.1	61.3
	전남	0.0	60.0	39.2	74.6	0.0	62.3	36.6	68.9	0.0	63.0	35.6	66.7
	경북	15.9	64.0	35.4	62.9	28.6	63.5	34.8	52.3	30.7	65.7	32.9	50.0
	경남	26.3	62.2	37.1	55.3	40.5	66.8	31.5	45.2	44.5	71.2	31.5	42.5
	제주	0.0	63.2	30.9	52.6	0.0	64.0	30.0	51.2	0.0	57.7	30.8	47.4

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 무응답률을 처리했다. 무응답률은 가중값을 적용하지 않는 표본에 대한 비가중(unweighted) 무응답률을 사용하고 있다. 당초 무응답률에 따른 무응답 처리의 차이점을 비교하고자 했으나, 무응답률이 18.6% 경우에도 무응답률이 30%인 경우와 메카니즘에 따른 무응답 처리결과가 유사하게 나타나 본 연구에서는 무응답률 30.0%만을 최종적으로 다룬다(Sukasih 등, 2009).

무응답은 주요 관심 변수의 무응답 발생여부(missingness)가 다른 변수와 어떻게 관련되느냐에 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. 무응답이 관심변수나 응답자에 의존하지 않고 발생되므로 R_i을 i가구의 응답여부라고 할 때, 응답인 경우는 $R_i = 1$이 되고, 무응답을 한 경우는 $R_i = 0$이 된다. 이 때 $R_i \sim Bernoulli(P)$를 따르며, 무응답 확률은 30%를 따른다.

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

임의결측(MAR: Missing At Random) 자료는 무응답 발생여부가 보조변수에 의존함에 따라, 응답확률은 보조변수 값에 따라 달라진다. MAR1 무응답 발생은 파라데이터로부터 응답여부에 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 발생여부

가 보조변수(covariates)와 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. 즉, $P_i = \frac{1}{1+\exp(\alpha+\beta^* \, log_{10} x_i)}$ 소득 수준 등에 따라 상이한 무응답률을 갖지만 전체적으로는 30%가 되도록 제한하였다(Paul, 2012).

④ NMAR1 자료는 소득에 따라 무응답이 발생한다.

⑤ NMAR2 자료는 소득 및 방문횟수에 따라 무응답이 발생한다.

⑥ NMAR3 자료는 주관적 만족감에 따라 무응답이 발생한다.

한편, 앞서 살펴본 무응답 구조(missing mechanism)를 보다 더 이해하기 위해 무응답 편향에 의해 다르게 살펴 볼 필요가 있는데 무응답 편향은 식(3.1)와 같이 표현할 수 있다.

$$nonresponse \;\; bias(y_r) = \frac{\sigma_{y,p}}{\bar{p}} \tag{3.1}$$

여기서 $\sigma_{y,p}$는 응답과 무응답을 모두 포함하는 조사변수 y와 응답성향(response propensity) p의 공분산(covariance)이다(이승희, 2010). 응답성향 p과 x라는 보조변수(covariates) 세 가지로 무응답 구조로 구분할 수 있다.

첫째로 $\sigma_{y,p} \approx 0$으로 조사변수(y)와 응답성향(p)와 관계가 전혀 없는 완전랜덤결측(MCAR), 둘째는 $\sigma_{y,p}|x \approx 0$으로 x라는 다른 변수를 조건으로 할 때, y와 p라는 두 변수의 상관이 없어지는 임의결측(MAR)인데, 이들을 묶어서 무시할 수 있는 무응답(ignorable missingness)이라고 구분한다. 마지

막으로 $\sigma_{y,p}|x \neq 0$로 y와 p라는 두 변수가 다른 특정 변수를 조건으로 함에도 상관관계가 존재하는 무시할 수 없는 무응답(nonignorable missingness)으로 분류된다. 보통 무시할 수 없는 무응답은 복잡한 통계 모델을 이용하여 조정이 가능하나 현실적으로는 무응답 구조에 대해 정확히 알 수 없어 무응답 조정은 간단하지 않다. 무응답에 대한 깊이 있는 접근은 Little과 Rubin(1987)을 참조할 수 있다(이승희, 2010).

<표 3.2> 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	14	26.4	300	28.3	80	27.4	175	28.1	89	29.6	12
	아파트	30.0		32.2		31.2		31.8		31.0		30.1	
	기타주택	30.0		31.2		30.6		29.5		30.4		29.8	
방문횟수	1회	30.1	18	18.1	300	28.7	41	28.1	42	23.1	300	29.9	19
	2회	29.9		21.9		29.7		29.6		29.0		29.6	
	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	10	29.3	38	29.4	32	30.3	28	29.8	13	29.9	12
	비자가	30.1		30.9		31.0		29.1		29.9		29.8	
가구원수	1인	30.0	21	28.5	89	28.5	300	26.0	273	28.5	72	29.5	21
	2인	30.0		28.5		27.1		28.8		28.7		29.7	
	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	9	32.4	204	32.9	300	27.6	244	29.9	81	30.5	16
	30-39세	30.2		32.4		34.3		32.2		31.0		30.2	
	40-49세	29.9		31.7		33.6		32.4		31.0		29.9	
	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	12	25.9	156	26.1	150	21.5	300	27.0	300	29.2	22
	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		31.5		31.9		33.3		29.3		30.0	
	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	10	25.2	209	25.1	247	24.4	275	27.3	98	28.9	25
	중 졸	29.8		28.6		27.3		27.8		28.4		29.4	
	고 졸	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	12	31.4	13	31.7	25	32.3	83	31.8	30	34.6	300
	높음	30.0		30.1		30.6		31.2		30.3		31.4	
	보통	30.0		29.5		29.6		29.3		29.4		28.9	
	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 3.2>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 음영(shadow)으로 표시하여 통계적으로 응답여부의 유의미한 차이를 나타냈다. 또한 메커니즘별로 무응답 발생을 위해 활용된 변수는 별도 테두리(box)를 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 보이고 있다. 이 경우에는 무응답률 자체가 추정값에 위협적이지 않을 것으로 판단할 수 있다(Curtin 등, 2000).

임의결측(MAR)에서 NMAR1 자료는 지역, 주택유형 및 방문횟수 등 파라데이터(paradata) 정보로부터 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 행정자료(administrative)로부터 가구원수와 가구주 연령에 따라 무응답이 발생되어, 이들 변수에서 유의미한 차이를 보이며 가구주 학력에도 유의미한 결과가 도출됐다.

비임의결측(NMAR)에서는 NMAR1 자료는 조사결과(survey outcomes) 소득에 따라 무응답이 발생하여, 소득수준별 응답성향은 유의한 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 조사변수인 주관적 만족감을 이용하여 무응답이 발생됐고, 이 변수에서만 유의미한 결과를 보였다.

<표 3.3> 시도별 메커니즘별 무응답 규모 및 무응답율

	원표본	무응답 규모					
		MCAR	MAR1	MAR2	NMAR1	NMAR2	NMAR3
전국	3,000	901	896	903	896	894	896
서울	280	84	94	86	87	88	85
부산	179	54	55	53	53	54	53
대구	151	45	46	45	45	46	44
인천	143	43	53	45	44	45	42
광주	144	43	47	45	43	44	43
대전	157	47	47	48	47	46	48
울산	159	48	55	49	50	51	48
세종	133	39	38	41	43	40	41
경기	362	109	112	111	111	110	108
강원	150	45	44	45	43	43	44
충북	153	45	46	45	42	45	45
충남	182	55	54	55	55	53	55
전북	174	53	46	50	50	50	52
전남	155	47	38	44	43	43	45
경북	178	54	42	52	51	49	53
경남	192	58	52	57	57	56	58
제주	108	32	27	32	32	31	32
	원표본	무응답률					
		MCAR	MAR1	MAR2	NMAR1	NMAR2	NMAR3
전국	3,000	30.0	29.9	30.1	29.9	29.8	29.9
서울	280	30.1	33.7	30.7	31.0	31.6	30.3
부산	179	30.1	30.9	29.5	29.6	30.2	29.6
대구	151	30.0	30.7	30.0	30.1	30.5	29.1
인천	143	29.8	36.8	31.1	30.6	31.5	29.4
광주	144	29.9	32.5	31.2	30.0	30.6	29.9
대전	157	29.9	30.1	30.6	30.2	29.5	30.4
울산	159	30.0	34.9	30.7	31.3	31.8	29.9
세종	133	29.6	28.4	30.8	32.0	30.3	30.6
경기	362	30.0	30.8	30.6	30.7	30.3	29.7
강원	150	29.9	29.5	29.8	28.7	29.0	29.5
충북	153	29.7	30.0	29.4	27.7	29.5	29.2
충남	182	30.0	29.6	30.1	30.0	29.4	30.1
전북	174	30.3	26.4	28.7	28.6	28.8	30.1
전남	155	30.0	24.4	28.1	27.8	27.7	29.2
경북	178	30.3	23.7	29.0	28.5	27.5	29.9
경남	192	30.0	27.0	29.9	29.5	29.0	30.3
제주	108	30.0	24.6	29.5	29.3	28.5	29.4

3.3 보조정보 구성

단위 무응답을 처리하기 위해서는 보조정보 구성이 중요하다. 보조정보는 응답 정보와 무응답 정보를 동시에 갖추어야 한다. 보조정보는 다양한 형태로 확보가 가능하다. 현장에서 얻은 파라데이터, 표본추출틀 또는 센서스 개별자료로부터 등 다양한 정보원으로부터 얻을 수 있다(Kalton and Maligalig, 1991). 다만 확보된 자료가 무응답 처리를 위해 충분한 보조정보를 갖추었는지 시의성 있는 최신 자료인지는 매우 중요하다. 또한 보조정보는 응답성향과 관련성이 있으면서 동시에 조사변수와 관련성도 높아야 한다(Rizzo, 1996; Matei and Ranalle, 2015). 보조변수를 구성하기 전에 변수간 상관관계를 살펴볼 필요가 있다.

변수 간의 상관관계는 상호 유의미한 관계를 보이나, 주관적 만족감, 건강평가, 스트레스 정도 등은 다른 만족도와 관련성이 높은 것을 볼 수 있다.

<표 3.4> 조사변수간 상관관계

	주관적 만족감	건강 평가	흡연 여부	음주 여부	의료서비스 만족도	스트레스 정도	학교교육 효과	사회전반 안전의식	결혼에 대한 견해	이혼에 대한 견해	가족 관계 만족도
주관적 만족감	1.000	0.445 <.0001	0.057 <.0001	-0.075 <.0001	0.125 <.0001	-0.283 <.0001	0.102 <.0001	0.133 <.0001	0.105 <.0001	0.068 <.0001	0.332 <.0001
건강평가	0.445 <.0001	1.000	-0.049 <.0001	-0.210 <.0001	0.034 <.0001	-0.201 <.0001	0.027 0.000	0.111 <.0001	-0.015 0.046	-0.044 <.0001	0.227 <.0001
흡연여부	0.057 <.0001	-0.049 <.0001	1.000	0.231 <.0001	0.082 <.0001	-0.100 <.0001	0.081 <.0001	0.004 0.547	0.066 <.0001	0.036 <.0001	0.045 <.0001
음주여부	-0.075 <.0001	-0.210 <.0001	0.231 <.0001	1.000	0.116 <.0001	-0.064 <.0001	0.085 <.0001	0.008 0.268	0.080 <.0001	0.103 <.0001	-0.014 0.055
의료서비스 만족도	0.125 <.0001	0.034 <.0001	0.082 <.0001	0.116 <.0001	1.000	-0.154 <.0001	0.148 <.0001	0.114 <.0001	0.106 <.0001	0.102 <.0001	0.148 <.0001
스트레스 정도	-0.283 <.0001	-0.201 <.0001	-0.100 <.0001	-0.064 <.0001	-0.154 <.0001	1.000	-0.110 <.0001	-0.142 <.0001	-0.100 <.0001	-0.072 <.0001	-0.188 <.0001
학교교육 효과	0.102 <.0001	0.027 0.000	0.081 <.0001	0.085 <.0001	0.148 <.0001	-0.110 <.0001	1.000	0.154 <.0001	0.117 <.0001	0.086 <.0001	0.077 <.0001
사회전반 안전의식	0.133 <.0001	0.111 <.0001	0.004 0.547	0.008 0.268	0.114 <.0001	-0.142 <.0001	0.154 <.0001	1.000	0.081 <.0001	0.054 <.0001	0.070 <.0001
결혼에 대한 견해	0.105 <.0001	-0.015 0.046	0.066 <.0001	0.080 <.0001	0.106 <.0001	-0.100 <.0001	0.117 <.0001	0.081 <.0001	1.000	0.384 <.0001	0.132 <.0001
이혼에 대한 견해	0.068 <.0001	-0.044 <.0001	0.036 <.0001	0.103 <.0001	0.102 <.0001	-0.072 <.0001	0.086 <.0001	0.054 <.0001	0.384 <.0001	1.000	0.096 <.0001
가족관계 만족도	0.332 <.0001	0.227 <.0001	0.045 <.0001	-0.014 0.055	0.148 <.0001	-0.188 <.0001	0.077 <.0001	0.070 <.0001	0.132 <.0001	0.096 <.0001	1.000

가구특성과 주요 관심변수와 상관관계는 모두 통계적으로 유의미한 관련성을 보이고 있다. 가구특성은 주요 보조정보(auxiliary information)로 활용될 수 있어, 이들의 자료를 파라데이터 또는 행정자료로부터 확보는 매우 중요하다. 또한 가구특성은 응답성향과도 관련성이 높아야 한다.

지역, 주택형태, 방문횟수는 직접 조사된 파라데이터로부터 확보가 가능하나, 기타자료는 행정자료를 통해서 수집이 가능하다. 다만, 소득, 가구주 학력 및 직업은 현재 행정자료로 수집될 수 없는 정보이다.

<표 3.5> 가구 특성과 관심변수와 상관관계(Spearman rank-order correlation)

	주관적 만족감	건강 평가	흡연 여부	음주 여부	의료서비스 만족도	스트레스 정도	학교교육 효과	사회전반 안전의식	결혼에 대한 견해	이혼에 대한 견해	가족 관계 만족도
지역	0.025	0.102	-0.042	-0.133	-0.136	0.068	-0.097	-0.001	-0.110	-0.124	-0.030
	0.001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.850	<.0001	<.0001	<.0001
주택형태	-0.063	-0.108	0.038	0.129	0.134	-0.051	0.099	0.038	0.094	0.087	-0.016
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.027
방문횟수	-0.023	-0.088	0.047	0.113	0.116	-0.056	0.096	0.029	0.096	0.087	0.016
	0.002	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.000	<.0001	<.0001	0.029
점유형태	0.100	-0.025	0.089	0.036	0.044	-0.084	0.043	0.005	0.151	0.147	0.048
	<.0001	0.001	<.0001	<.0001	<.0001	<.0001	<.0001	0.513	<.0001	<.0001	<.0001
가구원수	-0.151	-0.162	0.143	0.247	0.162	-0.115	0.118	0.040	0.024	0.001	-0.053
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.001	0.922	<.0001
가구주 성별	0.149	0.196	-0.319	-0.300	-0.099	-0.004	-0.079	0.058	0.063	0.049	0.040
	<.0001	<.0001	<.0001	<.0001	<.0001	0.6139	<.0001	<.0001	<.0001	<.0001	<.0001
가구주 연령	0.149	0.333	-0.217	-0.352	-0.189	0.124	-0.183	-0.037	-0.286	-0.249	0.113
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
소득	-0.278	-0.340	0.140	0.322	0.159	-0.052	0.120	-0.001	0.093	0.102	-0.091
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.871	<.0001	<.0001	<.0001
가구주 학력	-0.249	-0.365	0.109	0.299	0.151	-0.048	0.142	-0.004	0.178	0.178	-0.122
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.567	<.0001	<.0001	<.0001
가구주 직업	-0.024	-0.180	0.180	0.198	0.089	-0.066	0.094	0.019	0.051	0.046	0.002
	0.001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.010	<.0001	<.0001	0.790

다음은 가구특성별 상관관계를 살펴본다. 표두에 있는 가구특성은 무응답 발생을 위해 사용된 변수이고, 좌측에 있는 가구특성은 추후 무응답 처리를 위한 보조정보(auxiliary information)로 사용된 변수이다.

임의결측(MAR)에 활용되는 MAR1(지역, 주택유형, 방문횟수)와 MAR2(가구원수, 가구주 연령그룹)은 무응답 발생 변수와 보조정보 간에 상호 관련성이 높은 것을 알 수 있다.

비임의결측(NMAR)인 NMAR1의 무응답 발생으로 활용된 소득은 여러 보조정보와 관련성이 높아 소득을 직접 보조정보로 활용할 수 없더라도 다른 보조정보에 의해 무응답 처리가 높을 것으로 기대할 수 있다. 다만, NMAR3에 사용된 주관적 만족감은 타 변수와 관련성이 낮아 무응답 처리 효과가 낮을 것으로 판단된다.

<표 3.6> 가구특성별 상관관계
(Spearman rank-order correlation)

		무응답 발생 변수						
		지역	주택유형	방문횟수	가구원수	가구주 연령그룹	소득	주관적 만족감
보조정보	지역	1	-0.33889 <.0001	-0.2336 <.0001	-0.15457 <.0001	0.24347 <.0001	-0.23308 <.0001	0.025 0.001
	주택유형	-0.33889 <.0001	1	0.18427 <.0001	0.24236 <.0001	-0.23563 <.0001	0.28913 <.0001	-0.063 <.0001
	방문횟수	-0.2336 <.0001	0.18427 <.0001	1	0.12822 <.0001	-0.20712 <.0001	0.19609 <.0001	-0.023 0.002
	점유형태	-0.13416 <.0001	-0.05926 <.0001	0.07457 <.0001	-0.18093 <.0001	-0.28688 <.0001	-0.09906 <.0001	0.100 <.0001
	가구주 성별	0.04083 <.0001	-0.07774 <.0001	-0.03315 <.0001	-0.42932 <.0001	0.13599 <.0001	-0.34361 <.0001	0.149 <.0001
	가구원수	-0.15457 <.0001	0.24236 <.0001	0.12822 <.0001	1	-0.2815 <.0001	0.54589 <.0001	-0.151 <.0001
	가구주 연령그룹	0.24347 <.0001	-0.23563 <.0001	-0.20712 <.0001	-0.2815 <.0001	1	-0.40553 <.0001	0.149 <.0001

보조정보를 구성하기 위해 직접 파라데이터(paradata)를 활용하거나 수집된 행정자료(administrative data)를 활용할 수 있음을 앞서 살펴보았다. 먼저, 파라데이터로부터 연건평, 아파트 평형(소형 중형 대형), 무응답 사유, 주택유형, 방문일자, 방문시간, 지역(동부 읍면부) 등의 정보로부터 응답성향에 유의미한 정보인 지역, 주택유형, 방문횟수로 이루어진 3개 변수로 보조정보를 구성한다.

다음으로, 행정자료로부터 성별, 연령, 성명, 국적, 입국시기, 가구구분, 거처종류, 연면적, 대지면적, 건축연도 등 13개 항목을 포함하고 있으나 주요 보조정보로 활용하기 위해 응답성향에 영향을 미치는 주요 점유형태, 가구원수, 가구주 성별 및 연령 등 4개 변수를 선정하였다. 이렇게 파라데이터 3개변수를 더해 7개 변수를 다른 보조정보로 구성하여 통합자료라고 명명하고 이를 통해 무응답 조정을 전개한다(<표 3.7>).

<표 3.7> 무응답 메커니즘 및 보조정보에 활용된 변수

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

3.4 무응답 처리

표본조사에서 단위 무응답 처리를 위해 보편적으로 가중값 조정법이 사용된다(Groves, 2006). 보조변수(auxiliary variables)의 주변분포에 맞게 조정하는 래이킹 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 것 보다는 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 회귀방법 사용을 한다(통계청, 2016c).

본 연구에서는 가장 널리 사용되고 쉽게 적용할 수 있는 응답성향을 이용한 가중값 조정법인 로지스틱 회귀방법으로 접근하여 진행한다.

본격적인 무응답 처리를 진행하기에 앞서 무응답은 우선 조사에 응답을 거부하는 응답거부, 조사기간에 접촉할 수 없는 비접촉에 의한 단위 무응답(unit nonresponse)가 있고, 조사에는 응하나 민감한 질문인 약물(drug

abuse), 정치(politics), 소득(income), 사회적 바램(social desire) 등의 문항에 대해서는 답을 하지 않는 항목 무응답(item nonresponse)가 있다

먼저 단위 무응답을 살펴보면, 단위무응답은 주로 가중값 조정방법이 활용된다. 무응답 가중값 조정법의 기본원리는 표본추출에서 설계가중값을 부여한 것과 동일한 개념을 갖는다. 먼저 추출된 각 표본단위(i)는 추출확률(π_i)을 갖고 있으며, 추출확률의 역수를 고려해서 설계가중값을 작성한다. 다시 말해, 추출율의 역수로 가중 평균한 값은 HT(Horvitz-Thomson) 추정량 $\widehat{Y_{HT}} = 1/N \sum \pi_i^{-1} y_i$이다. 다만, 무응답이 있는 경우에는 표본으로 추출된 단위의 응답할 확률을 ϕ_i라고 하면 무응답이 있는 추정량은 식 (3.2)으로 표시된다(정구현, 2014).

$$\bar{y}_w = \frac{1}{N}\sum_{i=1}^{r}\frac{1}{\pi_i \phi_i} y_i \qquad (3.2)$$

응답률의 역수(ϕ^{-1})는 무응답 보정계수로 무응답을 조정하는 역할을 한다. 앞서 논의한 파라데이터 정보는 모두 응답여부와 관련이 높은 변수이며 조사변수와도 직접 연관됨을 논의했다. 보조변수는 무응답 조정을 위한 주변분포 또는 결합분포로 활용된다.

먼저 래이킹 비 조정법(raking ratio adjustment)은 미국 센서스 인구의 전수자료 및 표본자료의 불일치를 해결하기 위해 제안했다(Deming and Stephan, 1940). 미국 센서스국은 전수조사에서 얻은 주변분포(marginal distribution)를 이용하여 표본조사에서 나온 결합분포(joint distribution)를 추정한데 이를 활용하였다. 먼저 설계가중값을 한 쪽의 주변분포를 이용하여 조정한 후, 다음 다른 쪽의 주변분포를 이용하여 가중값을 조정한다. 일련의 작업을 지속적으로 반복하여 특정한 조건에 수렴할 때까지 진행된다. 보통 수렴조건은 모집단 총계와 조정가중값에 의한 총계의 차이를 이용하거나 최대 반복수를 기준으로 사용한다[4].

캘리브레이션 보정은 Deville과 Sarndal(1992)이 보조정보가 주어졌을 때, 모집단의 총합 추정량을 구하는 방법으로 정의한다. 이는 캘리브레이션

4) 통계청에서 사용하는 래이킹 비조정법은 차이는 0.0001, 최대 반복수는 100회를 사용한다.

가중값 w_k이 $\sum_s w_k x_k = \boldsymbol{t_x}$를 만족하면서 HT 추정량의 설계가중값 d_k와 거리함수를 최소화해야 해야 한다. 거리함수 조건에 따라 선형 캘리브레이션, 로짓 캘리브레이션 등 여러 가지 형태로 구분된다. 선형 캘리브레이션은 회귀추정량(Generalized regression; GREG)와 동일한 개념이다.

Bethlehem과 Keller(1987)은 일반화 회귀추정(generalized regression; GREG) 방법을 가중 추정량의 형태로 정리했다(식 (3.3)). 범주형 변수의 경우에 가중값은 비관측값(unobserved values) 사용하지 않고, 보조변수의 평균으로 적용이 가능하다.

$$\overline{\boldsymbol{y_w}} = n_R^{-1} \sum_{i=1}^{n_R} \boldsymbol{w_i y_i} \tag{3.3}$$

즉, 조정된 가중값은 $\boldsymbol{w_i} = \nu' X_i$가 되며 ν는 가중값 벡터이며, 이는 식 (3.4)으로 표현이 가능하다.

$$\boldsymbol{\nu} = n_R \left(\sum_{i=1}^{n_R} x_i x_i{}' \right)^{-1} \overline{X} \tag{3.4}$$

이는 또 다른 식 (3.5)와 같이 표현이 가능하다. 보다 자세한 내용은 Bethlehem과 Keller(1987), Bethlehem(2012)를 참조할 수 있다.

$$\boldsymbol{\nu} = \frac{n_R}{N} \left(\frac{N_1}{n_{r1}}, \frac{N_2}{n_{r2}}, \ldots, \frac{N_c}{n_{rc}} \right)' \tag{3.5}$$

따라서 회귀추정 방법을 활용함에도 종속변수 y를 의존하지 않고 추정이 가능하다. 장점은 주변분포만 있거나 결합분포가 많은 경우에 유용하게 활용 가능하다. 하나의 예로 지역(2)×주택유형(3)×점유형태(4)×가구주연령(4)은 96개 층이 요구되나, 지역+주택유형+점유형태+가구주연령으로 하면 13개 층으로 축소하여 처리가 가능하다.

무응답층을 결정하는데 있어 일반적으로 많이 사용되는 (CHAID: Chi-squared automatic interaction detection)은 가능한 모든 층의 응답률의

카이제곱 통계량을 최대로 하는 층의 분류를 찾아서 순차적으로 분류해 나가는 나무 모형(tree model)이다(Oh and Scheuren, 1983; Little, 1986).

로지스틱은 응답성향을 이용한 가중값 적용 방법으로, 응답과 불응으로 구성된 2진(binary) 지시변수를 종속변수로, 응답 및 무응답에 대한 공변수(covariates)를 독립변수로 하여 응답성향을 예측한다(Little, 1992). 즉, 응답성향을 예측하는 방법으로 응답자 가중값에 응답성향의 역수를 조정 팩터로 사용하여 곱하여 가중값을 작성한다(Iannacchione 등 1991; Lepkowski 등 1989; Czajka 등, 1992). 보조정보가 범주형일 경우는 래이킹 조정법과 유사한 방법이 된다.

지금까지는 단위 무응답에 대해 살펴보았는데, 조사에 응답을 하나 특정 문항에 대해서는 답을 하지 않는 항목 무응답(item nonresponse)에는 주로 대체방법(imputation methods)이 활용된다. 무응답이 있는 변수와 상관관계가 높은 있는 경우에 값을 대체하거나, 이를 통해 무응답을 예측하고 이를 대체하는 방법을 대체방법이라고 한다. 즉, 결측된 값을 그럴듯한 값으로 대체하는 통계적 기법을 대체(imputation)이라 할 수 있다.(통계교육원, 2010).

무응답 대체 방법은 크게 단일대체(single imputation)와 다중대체(multiple imputation) 방법으로 구분하고, 단일대체는 무응답값을 단일 값으로 대체된 값으로 작성하는 경우를 말하며, 이는 마치 실제 조사된 관측값처럼 분석하게 되면 추정값의 분산이 과소 추정되고 p-value가 작게 나타나는 문제가 발생된다. 이러한 무응답에 따라 발생되는 불확실성을 고려하기 위해 다중대체 방법이 사용된다.

단일대체(single imputation)는 결정적 대체(deterministic imputation)와 확률적 대체(stochastic imputation)으로 구분된다(김영원과 조선경, 1996; Kalton과 Kasprzyk, 1986). 결정적 대체는 무응답 항목에 대해 대체값이 단일 값으로 결정되는 대체방법으로, 연역적 대체, 시기적 대체, 평균대체, 축차 대체, 비대체, 최근방대체 등이 있다. 확률적인 대체 방법은 대체값이 확률적으로 정해지는 방법으로 핫덱 대체, 가중 핫덱 대체, 랜덤 비 대체, 랜

덤 회귀 대체 등이 있다(김규성, 2000).

연역적 대체(deductive imputation)는 다른 기록이나 논리적인 제약 조건에 의해 대체하는 방법이고, 시기적 대체(historical imputation)는 계속조사에서 동일 항목의 응답이 조사 시점에 따라 안정되고 이전 조사와 금번 조사 간에 관련성이 높은 경우에 이전 값으로 대체하는 방법이다.

평균대체(mean imputation)는 무응답 항목에 대체그룹 내에 있는 관측값 평균을 구하여, 무응답 값을 대체하는 방법이다. 이 방법은 간단하여 이용이 쉬운 장점이 있으며, 특히 항목이 연속형 변수일 때 유용한 방법이다. 다만, 대체 후 동일한 값이 많아져 응답자 분포가 왜곡되고 분산이 과소 추정되는 단점이 있다.

축차대체(sequential imputation)는 대체 그룹 내에서, 바로 이전에 응답값으로 대체하며, 인구 및 사회적 특성에 따라 대체 그룹이 잘 구성될 경우 효과적으로 사용할 수 있다.

비대체(ratio imputation)와 회귀 대체(regression imputation)는 보조변수를 이용하여 예측값을 구한 후 무응답을 대체하는 방법으로, 조사변수와 상관관계가 높고 연속형 변수에 효과적이다.

최근방 대체(nearest neighbour imputation)는 기타 변수를 이용하여 결측된 단위와 가장 유사한, 즉 절대 거리(absolute distance)가 가장 가까운 한 응답 조사 단위를 찾아 대체하는 방법으로 거리함수를 이용한다.

결정적 대체 방법은 대체 하나의 값으로 지정하여 대체를 함으로 무응답이 지닌 불확실성을 무시하는 경향이 있다. 이런 단점을 보완하기 위해 확률적 대체를 검토할 수 있다.

핫덱 대체(hot-deck imputation)는 대체 그룹내의 응답값 하나를 랜덤하게 결측값에 대체하는 방법으로, 무응답 대체 후에 표본의 분포가 유지되는 장점을 지닌다.

가중 핫덱 대체(weighted hot-deck imputation)는 응답값를 선정할 때 가중값을 두어 선정하는 방법으로 층화 또는 집락으로 이루어진 대체그룹의 추출확률이 다를 때 유용하게 활용할 수 있다.

랜덤 비대체 방법(random ratio imputation)이나 랜덤 회귀대체 방법(random regression imputation)은 추정된 예측값에 잔차를 포함하여 결측값을 대체하는 방법이다. 확률적 대체는 무응답의 변동성을 고려할 수 있다는 장점을 지닌다.

단일대체(single imputation)는 초기 응답값과 대체값에 대한 구별이 필요하다. 응답값은 모수에 대해 정확하게 측정한 값이지만, 대체값은 원래의 응답값과 다를 가능성이 높고, 정확한 대체모형이라고 하더라도 모두가 응답값과 동일할 가능성은 거의 없기 때문이다. 특히 모수에 대해 추론(inference)할 때 관측값의 정보만을 사용해야 한다. 이는 추정량의 분산을 과소 추정하고 추론에서 기각하지 않아야 하는 모수를 기각할 가능성도 있다.

다중대체(multiple imputation) 방법은 결측값을 2번 이상 대체하면 m개의 대체된 자료가 생성되고, 이들 대체값들은 무응답의 불확실성이 반영되어 이들의 평균 등을 대체값으로 사용하는 방법이다. 다중대체 방법은 Rubin(1978)에 의해 소개되었고, 자세한 내용은 Rubin(1987)을 참조할 수 있다. 대체값들에서 발생되는 차이가 반영되어 분산의 과소추정을 방지할 수 있고, 불확실한 결측값의 참값을 모형에 반영해 추정이 가능하다. 단점으로 단일대체에 비해 보다 복잡하고 각 자료들을 통합한 추론이 필요하다.

무응답 자료에 대해 다변량 정규분포를 가정하여 대체하는 프로그램은 SAS의 MI procedure, S-Plus의 missing data library, SPSS의 Missing Value Analysis(MVA) 모듈을 사용할 수 있다. 무응답 자료가 다른 형태의 변수들을 포함할 경우 대체를 위해 SAS 내에서 Macro를 사용하거나 독립적으로 실행할 수 있는 IVEware[5] 프로그램을 사용 가능하다(Raghunathan 등, 2002; 통계교육원, 2010).

5) https://www.src.isr.umich.edu/software/

4. 모의실험 결과

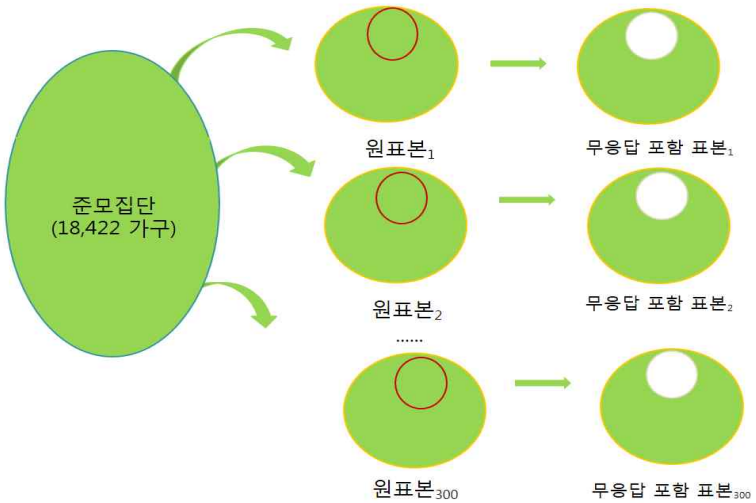
4.1 평가 내용

모의실험 결과를 측정하기 위해 2016년 사회조사의 보건, 교육, 안전, 가족, 환경 등 조사부문별로 모든 항목을 분석하였으나, 본 연구는 주관적 만족감, 건강평가, 흡연여부, 음주여부, 학교교육 효과, 결혼에 대한 견해, 이혼에 대한 견해 및 가족관계 만족도 등 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.

<표 4.1> 2016년 사회조사 조사부문 분석항목

개인	주관적 만족감
보건	건강평가, 흡연, 음주, 의료서비스 만족도, 스트레스 정도
교육	학교교육의 효과
안전	전반적인 사회 안전 인식, 교통안전시설 만족도
가족	결혼에 대한 견해, 이혼에 대한 견해, 전반적인 가족관계 만족도
환경	5년전 환경상황 변화

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.



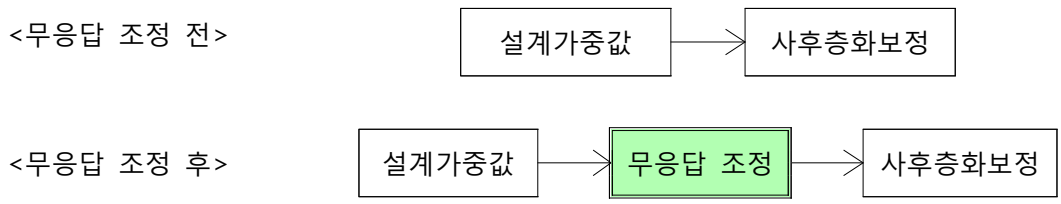
<그림 4.1> 원표본 및 무응답 표본셋

4.2 가중값 적용

원표본의 가중값은 설계가중값 및 사후층화 가중값 등 2단계로 가중값을 구성하여 산출된 추정량을 참값(true value) $\hat{\theta}_0$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.



<그림 4.2> 무응답 조정 전 및 조정후 가중값 적용

이를 보다 구체적으로 살펴보면, 표본추출을 위해 모집단에서 정의한 27개 지역층에서 추출확률의 역수를 적용하여 설계 가중값(base weight or design weight)을 식(4.1)처럼 계산한다.

$$\text{설계가중값(base weight)}: \ w_b = \frac{M_h}{m_h} \qquad (4.1)$$

· h : 지역층, $h = 1,2,\cdots,27$(세종~제주는 동부와 읍면부로 구분)
· M : 전체 가구수, · m : 표본 가구수

무응답 가중값을 작성하기 위해 앞서 살핀 두 가지 보조정보, 즉 파라데이터 지역, 주택유형, 방문횟수를 이용한 보조정보와 파라데이터에 행정자료 점유형태, 가구원수, 가구주 성별 및 연령을 더하여 구성된 보조정보 두 가지 셋을 이용하여 로지스틱 회귀모형에 의해 응답확률을 산출하여 응답확률의 역수로 무응답 조정계수를 식 (4.2)처럼 산출하였다. 산출된 응답확률은 응답성향(response probability) 또는 성향점수(propensity score)라고도 불린다. 무응답 가중값은 설계가중값에 무응답 조정계수를 곱하여 계산된다(식 4.2).

$$\text{무응답 조정계수}: \ w_{r,\,hi} = \frac{1}{\hat{p}_{hi}} \qquad (4.2)$$

· $w_{r,\,hi}$: h층 i가구의 무응답조정계수
· $\hat{p}_{hi}$: h층 i가구의 응답확률

$$\text{무응답 가중값}: \ w_{br,\,hi} = w_b \times w_{r,\,hi} \qquad (4.3)$$

· $w_{b,\,hi}$: h층 i가구의 설계가중값
· $h = 1,2,\cdots,27$: 설계층(지역), $i = 1,2,\cdots,n_i$: 가구

무응답 조정가중값은 다시 27개 지역과 지역내 가구원수를 곱한 108개 외부 벤치마크(external benchmarks)인 가구추계에 맞게 사후보정을 했다.

$$\text{가구 사후층화 조정계수}: \ w_{c,\,hi} = \frac{N_{c,\,gagu}}{\hat{N}_{c,\,gagu}} \qquad (4.4)$$

· $w_{c,hi}$: c 사후층 내 h층 i가구의 사후층화 조정계수
· $N_{c,\,gagu}$: 사후층 내 모집단 크기

· $\hat{N}_{c,\,gagu}$: 사후층 내 무응답 가중값의 합
· $c=1,2,\cdots,108$: 사후층

4.3 평가 척도

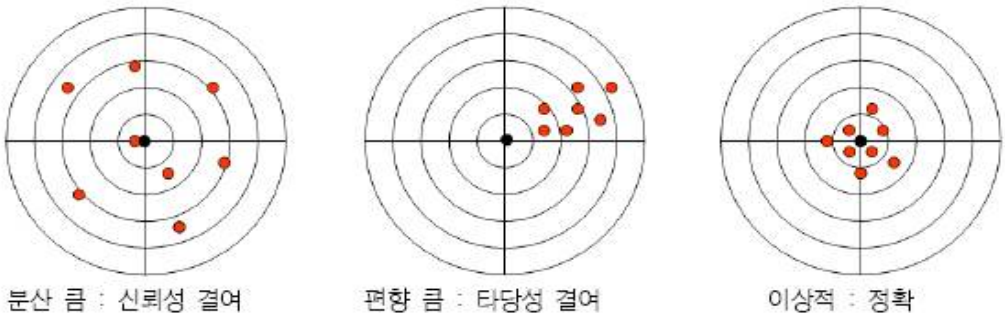
무응답 처리결과에 대한 평가를 위해 무응답 편향(bias), 상대편향(relative bias), 평균오차 제곱근(RMSE), 설계효과(design effect due to weighting) 등을 척도로 사용하고, 마지막으로 보조정보 간 모형비교(model comparisons)로 유의성을 점검한다.

편향(bias)은 원표본에 가중값을 적용한 추정량(참값)과 각 표본셋에서 산출된 추정량과의 차이로 식(4.5)에 의해 계산된다. $i(=1,2,3,4,5,6)$는 메커니즘 방법별 자료를 의미하며, $r(=1,\cdots,300)$은 300개 표본셋에 해당한다. 양(positive)의 값은 추정값이 모수 보다 작게 추정된 경우이며, 음(negative)의 경우는 반대에 해당 된다. 편향의 절대값이 작은 경우 정확도(precision)가 보다 높다고 판단할 수 있고 동시에 무응답 처리 성능이 우수하다 할 수 있다.

상대편향(relative bias)은 편향을 참값으로 나누어 백분위(percentage)로 표준화한 척도로서 식(4.6)처럼 계산된다. 이를 통하여 비중이 작은 경우에 낮게 나오는 편향에 대해 올바른 해석이 가능하다. 이 또한 편향과 마찬가지로 적은 값을 가질수록 처리효과가 높다고 평가할 수 있다.

평균제곱오차 제곱근(RMSE; root mean square error)은 무응답 조정된 추정량의 편향과 변동성을 동시에 측정하는 척도로 식(4.7)처럼 정의한다. RMSE는 분산(variance)과 편향(bias)라는 두 요소로 구성되어 있다(Kish, 1965; Andersen 등, 1979; Groves, 1989). 표본조사에 대한 품질을 평가하기 위해서는 흔히 생각되는 편향뿐 아니라 변동성도 같이 고려해야 한다. 표본조사에서 분산은 조사를 반복하거나 새로운 표본을 추출하여 조사한다면, 추정값의 변동성을 측정하는 것이고, 편향은 반복과는 무관하게 일정하게 조사에서 발생되는 체계적인(systematic) 문제로 야기된다. 이는 사회과학이

나 계량심리학(psychometrics)에서는 분산을 측정의 신뢰성(reliability) 또는 정도(precision)과 관련짓고, 편향은 타당성(validity)으로 설명한다. 오차(error)는 정확성(accuracy)은 편향과 분산이 모두 작을 때 우수하다 할 수 있다. 다만 처리효과를 측정하는데 편향에 더욱 높은 비중을 주고 있다(Groves, 2006; 이승희, 2010).



<그림 4.3> 편향 및 분산

설계효과(design effect due to weighting)는 무응답 조정계수를 적용한 후, 추정값의 분산이 커지는 효과로 정의할 수 있다(식 (4.8)). 여기서 가중값의 합이 1이 되도록 표준화 되어 있으며, $CV(w_i)$는 가중값 w_i의 변동계수이다. 가중값의 변동성이 커짐에 따라 추정량의 분산이 커지게 됨에 따라 이를 분산확대 계수(VIF; variance inflation factor)라고도 명명한다(Kalton and Flores-Cervantes, 2003).

다시 말해, 무응답 조정 후 편향은 감소하나, 변동성이 증가되어 정도(precision)는 낮아질 수밖에 없다. 설계효과는 정도의 손실을 측정하는 유용한 척도로 사용된다(Kish, 1992). 따라서 무응답과 관련이 없는 보조변수를 사용한 경우에 무응답 편향에는 큰 도움이 되지 않고 분산만 증가시키는 결과를 야기한다.

분산확대계수는 가중값에 의한 가중치 변동요인을 의미하며 예로서 VIF(Variance Inflation Factor)가 1.14면 모수 추정 시 가중값으로 인해 분산이 14%가 증가하고, $\sqrt{VIF}$는 표본오차가 7% 증가되었음을 의미한다. 또한 VIF가 2.5이상인 경우는 극단가중값(extreme weights)으로 판단되어 가중값 조정(weighting trimming)에 대한 추가 작업을 수행해야한다.

마지막으로 모형간 비교는 구성된 두 가지 보조정보에 대한 우도비 검정(likelihood-ratio test)을 통해 모형의 유의성을 확인한다.

편향(Bias) : $\frac{1}{300}\sum_{r=1}^{300}(\hat{\theta}_{ir}-\hat{\theta}_0)$, $\hat{\theta}_0$= 참값 (4.5)

상대편향(Relative Bias) : $\frac{(\overline{\theta}_i-\hat{\theta}_0)}{\hat{\theta}_0}\times 100$, $\overline{\theta}_i = \frac{1}{300}\sum_{r=1}^{300}\hat{\theta}_{ir}$ (4.6)

평균제곱오차 제곱근(RMSE) : $\sqrt{\frac{1}{299}\sum_{r=1}^{300}(\hat{\theta}_{ir}-\hat{\theta}_0)^2}$ (4.7)

설계효과(Design Effect) : $\frac{1}{300}\sum_{r=1}^{300}(1+CV(w_r)^2)$ (4.8)

모형 간 비교(Model comparisons)

.

4.4 처리 결과

무응답 처리는 2가지 보조정보(파라데이터, 통합자료)를 이용하여 6가지 무응답 메카니즘에 따라 진행되었다. 무응답 처리 결과를 살펴보기 전에 무응답 처리이전에 발생된 편향(bias)과 상대편향(relative bias)에 먼저 검토할 필요가 있다. 이는 무응답 처리 전과 처리 후를 비교 분석함으로써 처리효과에 대한 평가를 명확하고자 함이다. 메카니즘별로 완전임의결측(MCAR)은 편향과 상대편향이 모두 기준 이하로 나타났다. 임의결측(MAR)과 비임의결측(NMAR)은 여러 항목에서 편향과 상대편향이 기준점 이상으로 크게 나타난 것을 볼 수 있다.

본 내용은 3점 척도로 구성되며, 추정 비율과의 차이를 보여주고 있지만, 5점 척도로 이루어진 내용과 세부적인 추정 비중을 보려면 <표 4.5>를 참조할 수 있다.

<표 4.2> 무응답 처리 이전 추정값의 편향 및 상대적 편향 비교

		편향(Bias)						상대 편향(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

1) 편향(bias)

편향(bias)에 대해 처리결과는 대각선과 음영으로 설명된다(<표 4.3>). 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(<표 4.6>).

완전임의결측(MCAR)의 경우는 당초 편향이 기준점 이내에 있어 무응답 처리 후에도 큰 변화를 보이지 않지만 처리효과의 성능이 가장 우수하다. 두 가지 보조정보에서 큰 차이를 보이지 않는 것으로 확인된다.

파라데이터를 보조정보를 사용하여 무응답을 조정한 경우, 무응답 발생여부와 직접 관련된 변수를 포함하고 있는 MAR1 자료와 NMAR2 자료는 상대적으로 처리효과가 높게 나타났다. 반면 MAR2와 NMAR1은 무응답 발생이 파라데이터 정보와 관련이 없어 처리효과가 낮게 나타났다. NMAR3 주관적만족감의 불응여부는 무시할 수 없는(non-ignorable) 무응답으로 처리효과 낮다(Guillaume, 2013; Matei and Ranalli, 2015).

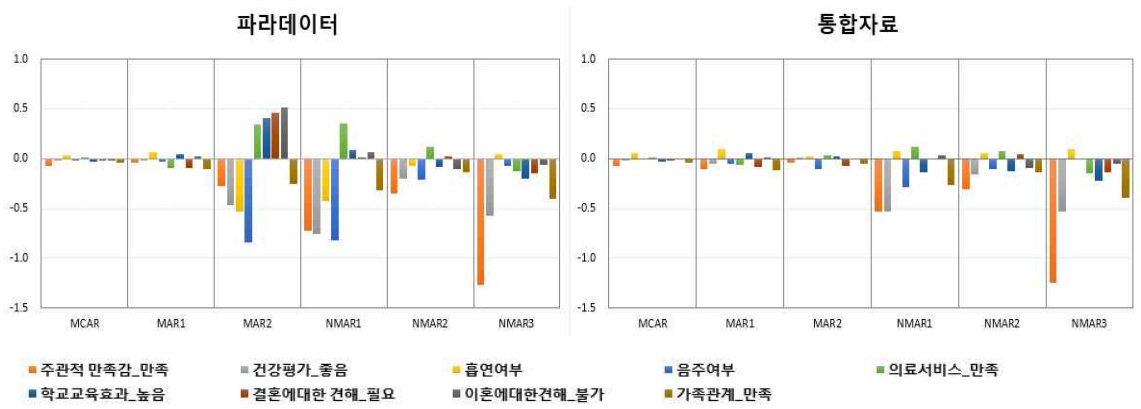
흥미로운 사실은 통합자료의 경우, 무응답 발생여부와 관련되거나 상관관계가 높은 대리 보조변수를 갖춤으로써, 무응답을 처리한 결과는 매우 성공적이다(Paul, 2012). 즉, 통합정보는 파라데이터를 포함하고 있어, 사실은 무응답 발생과 직접 관련된 MAR1, MAR2, NMAR1, NMAR2에서 처리효과가 높게 나타났다. 다만 소득의 경우는 보조정보로서 직접 사용되지는 않았지만 상관관계가 높은 변수를 보조정보로 활용함으로써 편향을 축소시키는데 기여하였다. 이로써 무응답 발생여부가 보조변수를 의존하거나 소득에 대한 상관관계가 높은 보조변수가 있는 경우에 비임의결측(NMAR) 자료도 성공적인 처리결과를 보였다.

다만, NMAR3 자료는 보조변수와 연관성이 없다는 점에서 처리효과가 낮게 나타났다.

<표 4.3> 추정값의 편향 비교

		파라데이터						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 4.4> 추정값의 편향 비교

2) 상대 편향(relative bias)

상대편향(relative bias)은 낮은 추정 비율을 갖는 경우에 과소 되는 편향을 상대적 백분위로 표준화하여 편향에 대한 해석을 명확하게 할 수 있다(<표 4.4>). 편향과 마찬가지로 완전임의결측(MCAR)의 경우는 상대편향 자체가 크지 않아 처리 후에도 크게 변화가 없지만 성능은 우수하게 나타났다. 따라서 MCAR은 별도의 무응답 처리가 큰 의미가 없음을 알 수 있다.

임의결측(MAR)인 MAR1은 무응답 발생이 파라데이터 변수에 의해 발생되어, 파라데이터 변수가 포함된 양쪽 보조정보에서 매우 우수한 처리효과를 보인다. MAR2는 무응답 발생이 행정자료 변수에 의해 발생되어 파라데이터로 구성된 보조정보에서는 처리 후 기준점 이상 편향이 발견되었지만, 통합자료를 이용한 보조정보에서는 모두 기준점 이하로 처리됨을 확인되었다.

비임의결측(NMAR)인 MAR1은 무응답 발생이 소득에 의해 발생되었고, 보조정보가 파라데이터인 경우는 소득과 보조정보 간에 상관관계가 있지만 소득을 직접 보조정보로 활용하지 않아 편향이 남아 있는 것을 볼 수 있다. 다만 통합자료를 활용한 경우에는 소득과 보다 관련된 변수가 존재하여 처리효과가 상대적으로 좋게 나타났다. MAR2는 무응답 발생과 관련된 방문횟수가 포함되고, 보조정보에도 포함됨에 따라 무응답 처리의 효과가 우수하게 나타났다. NMAR3는 보조정보와 관련이 없어 처리효과가 낮은 것으로 해석된다.

다시 정리하면, 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 상대 편향이 지속됐다. 상대편향 처리효과는 주요변수 평균 변동계수(CV) 1.5%를 임계점으로 평가했다.

<표 4.4> 추정값의 상대편향 비교

		파라데이터						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 4.5> 추정값의 상대편향 비교

<표 4.5> 무응답 자료별 편향 평가(파라데이터)

부문	문항	항목	원표본	MCAR		MAR1		MAR2	
				조정전	조정후	조정전	조정후	조정전	조정후
개인	주관적 만족감	매우 만족	10.86	10.87	10.84	10.68	10.83	10.72	10.74
		약간 만족	26.99	26.89	26.95	26.90	26.99	26.73	26.84
		보통	45.85	45.91	45.92	46.13	45.91	46.14	46.09
		약간 불만족	12.07	12.08	12.08	12.04	12.05	12.16	12.11
		매우 불만족	4.22	4.24	4.22	4.25	4.22	4.24	4.23
보건	건강평가	매우 좋다	6.54	6.57	6.54	6.45	6.52	6.42	6.41
		좋은 편	34.13	34.13	34.12	34.18	34.13	33.78	33.80
		보통	39.22	39.18	39.27	38.61	39.23	38.85	39.04
		나쁜 편	17.45	17.47	17.42	17.91	17.42	18.14	17.97
		매우 나쁨	2.66	2.66	2.66	2.86	2.69	2.81	2.78
	흡연율		31.16	31.31	31.20	31.16	31.23	30.72	30.63
	음주율		72.30	72.19	72.28	71.43	72.27	71.20	71.45
	의료서비스 만족도	매우 만족	13.98	14.02	13.99	14.54	14.02	14.45	14.33
		약간 만족	37.02	37.14	37.02	37.40	36.89	37.20	37.02
		보통	40.55	40.41	40.55	39.79	40.63	39.98	40.25
		약간 불만족	7.72	7.69	7.70	7.51	7.73	7.64	7.69
		매우 불만족	0.73	0.74	0.73	0.77	0.73	0.72	0.72
	스트레스 정도	매우 느낌	5.59	5.60	5.59	5.65	5.57	5.58	5.57
		느낀 편	48.26	48.27	48.27	47.72	48.14	48.00	48.07
		느끼지 않는 편	39.33	39.27	39.29	39.50	39.43	39.42	39.42
		전혀 느끼지 않음	6.82	6.86	6.86	7.13	6.86	7.00	6.95
교육	학교교육 효과	매우 효과적	7.94	7.86	7.93	7.99	7.90	8.02	8.06
		효과적	30.06	30.08	30.05	30.92	30.15	30.50	30.36
		보통	39.92	39.99	39.95	39.30	39.84	39.68	39.72
		별로 비효과적	17.49	17.44	17.47	17.23	17.50	17.27	17.35
		매우 비효과적	4.58	4.64	4.61	4.56	4.61	4.53	4.51
안전	사회전반 안전의식	매우 안전	1.27	1.26	1.27	1.22	1.27	1.29	1.31
		비교적 안전	12.93	12.97	12.92	12.98	12.88	13.02	12.94
		보통	43.52	43.55	43.57	43.49	43.57	43.39	43.42
		비교적 불안	34.67	34.63	34.67	34.99	34.74	34.84	34.86
		매우 불안	7.61	7.59	7.58	7.33	7.54	7.46	7.48
	교통안전 시설 만족도	매우 만족	9.06	9.00	9.04	9.25	9.04	9.23	9.26
		약간 만족	37.69	37.66	37.67	37.53	37.67	37.84	37.88
		보통	34.96	35.10	35.01	35.30	35.08	34.93	34.84
		약간 불만족	14.59	14.59	14.59	14.30	14.51	14.42	14.41
		매우 불만족	3.69	3.65	3.69	3.62	3.70	3.59	3.61
가족	결혼에 대한 견해	반드시 해야 함	16.05	16.03	16.08	16.57	16.08	16.61	16.56
		하는 것이 좋음	41.61	41.57	41.56	41.30	41.50	41.56	41.56
		해도 안 해도 좋음	38.71	38.78	38.72	38.51	38.83	38.26	38.30
		안하는 것이 좋음	2.07	2.09	2.10	2.03	2.04	2.03	2.04
		하지 말하야 함	0.31	0.31	0.31	0.31	0.31	0.30	0.31
		잘모름	1.24	1.22	1.23	1.27	1.24	1.24	1.23
	이혼에 대한 견해	절대 이혼은 안 됨	12.23	12.27	12.22	12.64	12.21	12.64	12.52
		가급적 안해야 됨	34.36	34.33	34.35	34.37	34.40	34.57	34.58
		해도 안해도 됨	38.26	38.19	38.26	37.74	38.22	37.67	37.83
		이유가 있으면 함	12.56	12.65	12.57	12.74	12.58	12.57	12.49
		잘모름	2.59	2.57	2.59	2.50	2.59	2.55	2.57
	가족관계 만족도	매우 만족	19.21	19.06	19.16	18.96	19.13	18.91	19.02
		약간 만족	34.87	34.93	34.87	35.21	34.84	34.88	34.81
		보통	41.04	41.12	41.10	40.85	41.13	41.27	41.27
		약간 불만족	3.97	3.99	3.96	4.11	4.00	4.01	3.97
		매우 불만족	0.91	0.91	0.91	0.87	0.90	0.93	0.93
환경	5년전 환경변화	매우 좋아짐	2.69	2.73	2.70	2.83	2.70	2.72	2.69
		약간 좋아짐	24.26	24.09	24.21	24.07	24.22	24.32	24.48
		보통	43.71	43.85	43.75	43.99	43.81	43.87	43.73
		약간 나빠짐	24.08	24.12	24.09	23.99	24.00	23.98	23.94
		매우 나빠짐	5.26	5.22	5.25	5.11	5.27	5.10	5.16

* 음영은 편향이 0.5% 초과, 대각선은 무응답 처리후 편향이 0.5% 이내로 축소

<표 4.5> 무응답 자료별 편향 평가(파라데이터) -계속-

부문	문항	항목	원표본	NMAR1		NMAR2		NMAR3	
				조정전	조정후	조정전	조정후	조정전	조정후
개인	주관적 만족감	매우 만족	10.86	10.52	10.49	10.59	10.63	10.20	10.17
		약간 만족	26.99	26.47	26.65	26.79	26.88	26.35	26.42
		보통	45.85	46.07	46.18	46.11	46.03	46.46	46.46
		약간 불만족	12.07	12.59	12.40	12.19	12.17	12.90	12.89
		매우 불만족	4.22	4.35	4.29	4.32	4.29	4.09	4.06
보건	건강평가	매우 좋다	6.54	6.50	6.45	6.50	6.51	6.37	6.36
		좋은 편	34.13	33.45	33.46	33.98	33.96	33.76	33.75
		보통	39.22	38.79	39.16	38.97	39.27	39.39	39.48
		나쁜 편	17.45	18.38	18.08	17.79	17.56	17.77	17.71
		매우 나쁨	2.66	2.87	2.84	2.76	2.70	2.70	2.70
	흡연율		31.16	30.67	30.73	31.12	31.09	31.31	31.20
	음주율		72.30	71.29	71.47	71.68	72.09	72.12	72.22
	의료서비스 만족도	매우 만족	13.98	14.53	14.32	14.35	14.12	14.00	13.95
		약간 만족	37.02	37.03	37.04	37.28	37.00	37.05	36.93
		보통	40.55	39.95	40.23	40.05	40.49	40.50	40.67
		약간 불만족	7.72	7.71	7.69	7.57	7.66	7.71	7.71
		매우 불만족	0.73	0.79	0.73	0.74	0.72	0.74	0.74
	스트레스 정도	매우 느낌	5.59	5.76	5.72	5.67	5.64	5.67	5.65
		느낀 편	48.26	47.99	48.18	48.06	48.24	48.54	48.56
		느끼지 않는 편	39.33	39.38	39.17	39.27	39.25	39.06	39.08
		전혀 느끼지 않음	6.82	6.87	6.93	7.00	6.88	6.74	6.72
교육	학교교육 효과	매우 효과적	7.94	8.01	8.07	7.97	7.96	7.83	7.89
		효과적	30.06	30.14	30.02	30.29	29.96	29.96	29.91
		보통	39.92	39.88	39.82	39.73	39.93	40.09	40.06
		별로 비효과적	17.49	17.46	17.56	17.51	17.63	17.53	17.56
		매우 비효과적	4.58	4.51	4.52	4.51	4.51	4.60	4.57
안전	사회전반 안전의식	매우 안전	1.27	1.31	1.30	1.25	1.27	1.24	1.25
		비교적 안전	12.93	12.86	12.95	13.06	12.99	12.87	12.83
		보통	43.52	43.58	43.38	43.49	43.52	43.52	43.53
		비교적 불안	34.67	34.69	34.81	34.65	34.58	34.71	34.75
		매우 불안	7.61	7.56	7.56	7.55	7.63	7.65	7.65
	교통안전 시설 만족도	매우 만족	9.06	9.15	9.19	9.15	9.10	9.03	9.08
		약간 만족	37.69	37.45	37.79	37.58	37.64	37.63	37.64
		보통	34.96	35.28	35.06	35.13	34.99	35.10	35.01
		약간 불만족	14.59	14.49	14.34	14.52	14.60	14.60	14.60
		매우 불만족	3.69	3.63	3.63	3.62	3.67	3.65	3.68
가족	결혼에 대한 견해	반드시 해야 함	16.05	16.30	16.36	16.32	16.15	15.97	16.02
		하는 것이 좋음	41.61	41.22	41.31	41.46	41.55	41.51	41.50
		해도 안 해도 좋음	38.71	38.74	38.62	38.60	38.68	38.86	38.80
		안하는 것이 좋음	2.07	2.15	2.12	2.05	2.07	2.11	2.12
		하지 말하야 함	0.31	0.27	0.31	0.31	0.31	0.31	0.31
		잘모름	1.24	1.32	1.28	1.25	1.25	1.24	1.25
	이혼에 대한 견해	절대 이혼은 안 됨	12.23	12.60	12.46	12.52	12.32	12.26	12.21
		가급적 안해야 됨	34.36	34.27	34.20	34.12	34.17	34.28	34.31
		해도 안해도 됨	38.26	37.84	38.14	38.08	38.30	38.23	38.30
		이유가 있으면 함	12.56	12.70	12.56	12.69	12.57	12.63	12.55
		잘모름	2.59	2.59	2.64	2.59	2.64	2.60	2.62
	가족관계 만족도	매우 만족	19.21	18.97	19.09	19.02	19.13	18.81	18.91
		약간 만족	34.87	34.70	34.67	34.99	34.81	34.81	34.76
		보통	41.04	41.18	41.26	41.03	41.13	41.41	41.38
		약간 불만족	3.97	4.18	4.05	4.06	4.01	4.06	4.03
		매우 불만족	0.91	0.96	0.93	0.90	0.91	0.91	0.92
환경	5년전 환경변화	매우 좋아짐	2.69	2.63	2.67	2.69	2.62	2.71	2.69
		약간 좋아짐	24.26	24.03	24.21	24.07	24.20	24.05	24.18
		보통	43.71	43.78	43.72	44.01	43.88	43.82	43.71
		약간 나빠짐	24.08	24.33	24.14	24.08	24.06	24.13	24.10
		매우 나빠짐	5.26	5.23	5.26	5.15	5.24	5.28	5.32

* 음영은 | 원 표본 추정값 - 사후보정 후 추정값 | >0.5%, 대각선은 무응답 처리후 편향이 0.5% 이내로 축소

<표 4.5> 무응답 자료별 편향 평가(통합자료) -계속-

부문	문항	항목	원표본	MCAR		MAR1		MAR2	
				조정전	조정후	조정전	조정후	조정전	조정후
개인	주관적 만족감	매우 만족	10.86	10.87	10.84	10.68	10.82	10.72	10.86
		약간 만족	26.99	26.89	26.95	26.90	26.94	26.73	26.95
		보통	45.85	45.91	45.91	46.13	45.94	46.14	45.87
		약간 불만족	12.07	12.08	12.08	12.04	12.08	12.16	12.08
		매우 불만족	4.22	4.24	4.22	4.25	4.22	4.24	4.24
보건	건강평가	매우 좋다	6.54	6.57	6.54	6.45	6.52	6.42	6.55
		좋은 편	34.13	34.13	34.12	34.18	34.10	33.78	34.12
		보통	39.22	39.18	39.26	38.61	~~39.22~~	38.85	39.12
		나쁜 편	17.45	17.47	17.42	17.91	17.47	18.14	~~17.52~~
		매우 나쁨	2.66	2.66	2.66	2.86	2.70	2.81	2.68
	흡연율		31.16	31.31	31.22	31.16	31.26	30.72	31.18
	음주율		72.30	72.19	72.29	71.43	~~72.25~~	71.20	~~72.19~~
	의료서비스 만족도	매우 만족	13.98	14.02	13.99	14.54	~~14.04~~	14.45	14.04
		약간 만족	37.02	37.14	37.03	37.40	36.90	37.20	36.99
		보통	40.55	40.41	40.55	39.79	~~40.61~~	39.98	~~40.49~~
		약간 불만족	7.72	7.69	7.70	7.51	7.73	7.64	7.72
		매우 불만족	0.73	0.74	0.73	0.77	0.73	0.72	0.75
	스트레스 정도	매우 느낌	5.59	5.60	5.59	5.65	5.58	5.58	5.59
		느낀 편	48.26	48.27	48.26	47.72	~~48.10~~	48.00	48.35
		느끼지 않는 편	39.33	39.27	39.29	39.50	39.45	39.42	39.22
		전혀 느끼지 않음	6.82	6.86	6.86	7.13	6.87	7.00	6.83
교육	학교교육 효과	매우 효과적	7.94	7.86	7.93	7.99	7.90	8.02	7.93
		효과적	30.06	30.08	30.05	30.92	~~30.17~~	30.50	30.11
		보통	39.92	39.99	39.95	39.30	~~39.82~~	39.68	39.91
		별로 비효과적	17.49	17.44	17.47	17.23	17.49	17.27	17.47
		매우 비효과적	4.58	4.64	4.61	4.56	4.62	4.53	4.59
안전	사회전반 안전의식	매우 안전	1.27	1.26	1.27	1.22	1.26	1.29	1.28
		비교적 안전	12.93	12.97	12.92	12.98	12.89	13.02	12.94
		보통	43.52	43.55	43.57	43.49	43.57	43.39	43.42
		비교적 불안	34.67	34.63	34.67	34.99	34.75	34.84	34.78
		매우 불안	7.61	7.59	7.58	7.33	7.53	7.46	7.58
	교통안전 시설 만족도	매우 만족	9.06	9.00	9.04	9.25	9.03	9.23	9.09
		약간 만족	37.69	37.66	37.67	37.53	37.63	37.84	37.70
		보통	34.96	35.10	35.01	35.30	35.10	34.93	34.94
		약간 불만족	14.59	14.59	14.59	14.30	14.52	14.42	14.60
		매우 불만족	3.69	3.65	3.69	3.62	3.71	3.59	3.67
가족	결혼에 대한 견해	반드시 해야 함	16.05	16.03	16.09	16.57	~~16.09~~	16.61	~~16.05~~
		하는 것이 좋음	41.61	41.57	41.56	41.30	41.49	41.56	41.55
		해도 안 해도 좋음	38.71	38.78	38.71	38.51	38.81	38.26	38.79
		안하는 것이 좋음	2.07	2.09	2.10	2.03	2.05	2.03	2.05
		하지 말하야 함	0.31	0.31	0.31	0.31	0.31	0.30	0.32
		잘모름	1.24	1.22	1.22	1.27	1.25	1.24	1.25
	이혼에 대한 견해	절대 이혼은 안 됨	12.23	12.27	12.22	12.64	12.23	12.64	12.22
		가급적 안해야 됨	34.36	34.33	34.36	34.37	34.37	34.57	34.36
		해도 안해도 됨	38.26	38.19	38.25	37.74	38.19	37.67	~~38.23~~
		이유가 있으면 함	12.56	12.65	12.57	12.74	12.60	12.57	12.58
		잘모름	2.59	2.57	2.59	2.50	2.61	2.55	2.61
	가족관계 만족도	매우 만족	19.21	19.06	19.16	18.96	19.12	18.91	19.17
		약간 만족	34.87	34.93	34.87	35.21	34.83	34.88	34.86
		보통	41.04	41.12	41.09	40.85	41.14	41.27	41.08
		약간 불만족	3.97	3.99	3.96	4.11	4.01	4.01	3.97
		매우 불만족	0.91	0.91	0.91	0.87	0.90	0.93	0.92
환경	5년전 환경변화	매우 좋아짐	2.69	2.73	2.70	2.83	2.70	2.72	2.68
		약간 좋아짐	24.26	24.09	24.21	24.07	24.19	24.32	24.30
		보통	43.71	43.85	43.75	43.99	43.84	43.87	43.69
		약간 나빠짐	24.08	24.12	24.09	23.99	24.00	23.98	24.06
		매우 나빠짐	5.26	5.22	5.25	5.11	5.26	5.10	5.27

* 음영은 | 원 표본 추정값 - 사후보정 후 추정값 | >0.5%, 대각선은 무응답 처리후 편향이 0.5% 이내로 축소

<표 4.5> 무응답 자료별 편향 평가(통합자료) -계속-

부문	문항	항목	원표본	NMAR1		NMAR2		NMAR3	
				조정전	조정후	조정전	조정후	조정전	조정후
개인	주관적 만족감	매우 만족	10.86	10.52	10.60	10.59	10.66	10.20	10.18
		약간 만족	26.99	26.47	26.72	26.79	26.89	26.35	26.43
		보통	45.85	46.07	46.03	46.11	46.00	46.46	46.44
		약간 불만족	12.07	12.59	12.36	12.19	12.17	12.90	12.88
		매우 불만족	4.22	4.35	4.29	4.32	4.29	4.09	4.06
보건	건강평가	매우 좋다	6.54	6.50	6.49	6.50	6.52	6.37	6.37
		좋은 편	34.13	33.45	33.65	33.98	34.00	33.76	33.77
		보통	39.22	38.79	39.33	38.97	39.31	39.39	39.49
		나쁜 편	17.45	18.38	17.77	17.79	17.49	17.77	17.68
		매우 나쁨	2.66	2.87	2.76	2.76	2.68	2.70	2.69
	흡연율		31.16	30.67	31.24	31.12	31.22	31.31	31.27
	음주율		72.30	71.29	72.01	71.68	72.19	72.12	72.29
	의료서비스 만족도	매우 만족	13.98	14.53	14.12	14.35	14.09	14.00	13.93
		약간 만족	37.02	37.03	37.00	37.28	36.99	37.05	36.93
		보통	40.55	39.95	40.43	40.05	40.53	40.50	40.69
		약간 불만족	7.72	7.71	7.70	7.57	7.66	7.71	7.71
		매우 불만족	0.73	0.79	0.74	0.74	0.73	0.74	0.74
	스트레스 정도	매우 느낌	5.59	5.76	5.72	5.67	5.64	5.67	5.65
		느낀 편	48.26	47.99	48.37	48.06	48.27	48.54	48.57
		느끼지 않는 편	39.33	39.38	39.06	39.27	39.23	39.06	39.07
		전혀 느끼지 않음	6.82	6.87	6.85	7.00	6.87	6.74	6.71
교육	학교교육 효과	매우 효과적	7.94	8.01	7.97	7.97	7.94	7.83	7.88
		효과적	30.06	30.14	29.90	30.29	29.93	29.96	29.90
		보통	39.92	39.88	39.95	39.73	39.95	40.09	40.07
		별로 비효과적	17.49	17.46	17.62	17.51	17.65	17.53	17.57
		매우 비효과적	4.58	4.51	4.56	4.51	4.52	4.60	4.57
안전	사회전반 안전의식	매우 안전	1.27	1.31	1.28	1.25	1.27	1.24	1.25
		비교적 안전	12.93	12.86	12.92	13.06	12.98	12.87	12.83
		보통	43.52	43.58	43.47	43.49	43.55	43.52	43.53
		비교적 불안	34.67	34.69	34.73	34.65	34.56	34.71	34.73
		매우 불안	7.61	7.56	7.60	7.55	7.64	7.65	7.65
	교통안전 시설 만족도	매우 만족	9.06	9.15	9.10	9.15	9.08	9.03	9.07
		약간 만족	37.69	37.45	37.68	37.58	37.62	37.63	37.62
		보통	34.96	35.28	35.08	35.13	35.00	35.10	35.01
		약간 불만족	14.59	14.49	14.47	14.52	14.63	14.60	14.61
		매우 불만족	3.69	3.63	3.68	3.62	3.68	3.65	3.68
가족	결혼에 대한 견해	반드시 해야 함	16.05	16.30	16.17	16.32	16.12	15.97	16.01
		하는 것이 좋음	41.61	41.22	41.49	41.46	41.59	41.51	41.52
		해도 안 해도 좋음	38.71	38.74	38.67	38.60	38.68	38.86	38.80
		안하는 것이 좋음	2.07	2.15	2.10	2.05	2.06	2.11	2.11
		하지 말하야 함	0.31	0.27	0.31	0.31	0.31	0.31	0.31
		잘모름	1.24	1.32	1.27	1.25	1.25	1.24	1.25
	이혼에 대한 견해	절대 이혼은 안 됨	12.23	12.60	12.38	12.52	12.31	12.26	12.21
		가급적 안해야 됨	34.36	34.27	34.25	34.12	34.18	34.28	34.33
		해도 안해도 됨	38.26	37.84	38.24	38.08	38.31	38.23	38.30
		이유가 있으면 함	12.56	12.70	12.49	12.69	12.55	12.63	12.54
		잘모름	2.59	2.59	2.65	2.59	2.64	2.60	2.62
	가족관계 만족도	매우 만족	19.21	18.97	19.11	19.02	19.13	18.81	18.92
		약간 만족	34.87	34.70	34.70	34.99	34.82	34.81	34.77
		보통	41.04	41.18	41.25	41.03	41.14	41.41	41.37
		약간 불만족	3.97	4.18	4.02	4.06	4.00	4.06	4.02
		매우 불만족	0.91	0.96	0.91	0.90	0.91	0.91	0.92
환경	5년전 환경변화	매우 좋아짐	2.69	2.63	2.66	2.69	2.62	2.71	2.69
		약간 좋아짐	24.26	24.03	24.11	24.07	24.17	24.05	24.17
		보통	43.71	43.78	43.76	44.01	43.89	43.82	43.71
		약간 나빠짐	24.08	24.33	24.20	24.08	24.08	24.13	24.10
		매우 나빠짐	5.26	5.23	5.27	5.15	5.24	5.28	5.32

* 음영은 | 원 표본 추정값 - 사후보정 후 추정값 | >0.5%, 대각선은 무응답 처리후 편향이 0.5% 이내로 축소

<표 4.6> 무응답 자료별 표준오차(파라데이터)

부문	문항	항목	MCAR	MAR1	MAR2	NMAR1	NMAR2	NMAR3
개인	주관적 만족감	매우 만족	0.81	0.85	0.83	0.82	0.82	0.80
		약간 만족	1.14	1.18	1.15	1.14	1.15	1.14
		보통	1.28	1.33	1.30	1.29	1.30	1.29
		약간 불만족	0.90	0.85	0.83	0.85	0.84	0.85
		매우 불만족	0.53	0.58	0.57	0.58	0.58	0.56
보건	건강평가	매우 좋다	0.63	0.66	0.64	0.63	0.64	0.63
		좋은 편	1.23	1.26	1.23	1.22	1.23	1.22
		보통	1.26	1.31	1.28	1.28	1.28	1.27
		나쁜 편	0.96	0.97	0.97	0.98	0.97	0.97
		매우 나쁨	0.41	0.39	0.39	0.40	0.39	0.40
	흡연율		1.19	1.23	1.20	1.19	1.21	1.20
	음주율		1.14	1.15	1.15	1.14	1.14	1.13
	의료서비스 만족도	매우 만족	0.98	0.96	0.96	0.97	0.97	0.96
		약간 만족	1.39	1.41	1.38	1.38	1.38	1.38
		보통	1.42	1.45	1.41	1.41	1.42	1.41
		약간 불만족	0.73	0.81	0.79	0.78	0.77	0.77
		매우 불만족	0.29	0.23	0.23	0.24	0.23	0.23
	스트레스 정도	매우 느낌	0.58	0.60	0.59	0.59	0.59	0.59
		느낀 편	1.29	1.33	1.30	1.30	1.30	1.30
		느끼지 않는 편	1.26	1.30	1.27	1.27	1.27	1.26
		전혀 느끼지 않음	0.64	0.66	0.66	0.65	0.65	0.64
교육	학교교육 효과	매우 효과적	0.71	0.70	0.70	0.69	0.69	0.69
		효과적	1.20	1.23	1.21	1.20	1.20	1.20
		보통	1.29	1.35	1.31	1.31	1.31	1.31
		별로 비효과적	1.05	1.06	1.03	1.03	1.04	1.03
		매우 비효과적	0.56	0.60	0.59	0.58	0.59	0.58
안전	사회전반 안전의식	매우 안전	0.30	0.29	0.30	0.30	0.30	0.29
		비교적 안전	0.88	0.90	0.88	0.87	0.87	0.87
		보통	1.27	1.32	1.29	1.29	1.29	1.28
		비교적 불안	1.23	1.26	1.24	1.23	1.24	1.23
		매우 불안	0.71	0.73	0.70	0.70	0.72	0.71
	교통안전 시설 만족도	매우 만족	0.76	0.75	0.75	0.75	0.74	0.75
		약간 만족	1.25	1.30	1.27	1.26	1.27	1.26
		보통	1.24	1.27	1.24	1.24	1.24	1.23
		약간 불만족	0.92	0.92	0.91	0.91	0.91	0.90
		매우 불만족	0.46	0.49	0.47	0.47	0.48	0.48
가족	결혼에 대한 견해	반드시 해야 함	0.92	0.95	0.94	0.94	0.93	0.93
		하는 것이 좋음	1.27	1.31	1.29	1.28	1.29	1.28
		해도 안 해도 좋음	1.26	1.30	1.27	1.27	1.27	1.26
		안하는 것이 좋음	0.38	0.41	0.40	0.42	0.41	0.41
		하지 말하야 함	0.16	0.13	0.13	0.13	0.13	0.13
		잘모름	0.31	0.25	0.24	0.26	0.25	0.25
	이혼에 대한 견해	절대 이혼은 안 됨	0.81	0.83	0.84	0.84	0.84	0.83
		가급적 안해야 됨	1.22	1.26	1.24	1.23	1.24	1.23
		해도 안해도 됨	1.26	1.30	1.26	1.26	1.27	1.26
		이유가 있으면 함	0.87	0.90	0.88	0.89	0.88	0.88
		잘모름	0.43	0.42	0.40	0.40	0.40	0.39
	가족관계 만족도	매우 만족	1.01	1.05	1.02	1.02	1.02	1.02
		약간 만족	1.23	1.25	1.23	1.22	1.23	1.22
		보통	1.28	1.32	1.29	1.28	1.29	1.28
		약간 불만족	0.49	0.53	0.53	0.54	0.53	0.53
		매우 불만족	0.28	0.25	0.25	0.26	0.25	0.24
환경	5년전 환경변화	매우 좋아짐	0.43	0.43	0.43	0.42	0.42	0.43
		약간 좋아짐	1.11	1.15	1.12	1.12	1.12	1.12
		보통	1.28	1.32	1.29	1.29	1.29	1.28
		약간 나빠짐	1.09	1.12	1.10	1.10	1.10	1.09
		매우 나빠짐	0.59	0.65	0.63	0.63	0.63	0.64

<표 4.6> 무응답 자료별 표준오차(통합정보) -계속-

부문	문항	항목	MCAR	MAR1	MAR2	NMAR1	NMAR2	NMAR3
개인	주관적 만족감	매우 만족	0.83	0.85	0.84	0.83	0.83	0.80
		약간 만족	1.15	1.18	1.15	1.15	1.15	1.14
		보통	1.29	1.33	1.30	1.29	1.30	1.29
		약간 불만족	0.83	0.85	0.83	0.85	0.84	0.86
		매우 불만족	0.58	0.58	0.57	0.58	0.58	0.55
보건	건강평가	매우 좋다	0.64	0.67	0.65	0.64	0.64	0.62
		좋은 편	1.23	1.26	1.24	1.23	1.23	1.23
		보통	1.27	1.31	1.28	1.28	1.28	1.28
		나쁜 편	0.96	0.97	0.95	0.97	0.96	0.96
		매우 나쁨	0.39	0.39	0.38	0.39	0.39	0.39
	흡연율		1.20	1.24	1.21	1.21	1.21	1.20
	음주율		1.13	1.15	1.13	1.13	1.13	1.13
	의료서비스 만족도	매우 만족	0.96	0.97	0.95	0.96	0.97	0.95
		약간 만족	1.38	1.41	1.38	1.38	1.39	1.38
		보통	1.41	1.45	1.42	1.42	1.42	1.41
		약간 불만족	0.78	0.81	0.79	0.78	0.77	0.78
		매우 불만족	0.23	0.23	0.24	0.25	0.23	0.25
	스트레스 정도	매우 느낌	0.59	0.60	0.59	0.59	0.59	0.59
		느낀 편	1.30	1.33	1.31	1.30	1.30	1.30
		느끼지 않는 편	1.27	1.30	1.27	1.27	1.27	1.26
		전혀 느끼지 않음	0.65	0.66	0.65	0.64	0.65	0.65
교육	학교교육 효과	매우 효과적	0.70	0.70	0.69	0.69	0.69	0.69
		효과적	1.20	1.23	1.21	1.20	1.20	1.20
		보통	1.31	1.35	1.32	1.31	1.31	1.31
		별로 비효과적	1.02	1.06	1.04	1.04	1.04	1.02
		매우 비효과적	0.59	0.61	0.60	0.59	0.59	0.59
안전	사회전반 안전의식	매우 안전	0.30	0.29	0.29	0.30	0.30	0.29
		비교적 안전	0.87	0.90	0.89	0.87	0.87	0.87
		보통	1.28	1.32	1.29	1.29	1.29	1.28
		비교적 불안	1.23	1.26	1.24	1.24	1.24	1.23
		매우 불안	0.70	0.73	0.71	0.71	0.72	0.71
	교통안전 시설 만족도	매우 만족	0.75	0.75	0.74	0.75	0.74	0.74
		약간 만족	1.26	1.30	1.27	1.26	1.27	1.26
		보통	1.24	1.27	1.24	1.24	1.24	1.24
		약간 불만족	0.91	0.93	0.92	0.92	0.92	0.91
		매우 불만족	0.48	0.49	0.48	0.48	0.48	0.48
가족	결혼에 대한 견해	반드시 해야 함	0.93	0.95	0.92	0.93	0.93	0.92
		하는 것이 좋음	1.28	1.32	1.29	1.29	1.29	1.28
		해도 안 해도 좋음	1.26	1.30	1.28	1.27	1.27	1.27
		안하는 것이 좋음	0.41	0.41	0.40	0.41	0.40	0.41
		하지 말하야 함	0.14	0.13	0.13	0.13	0.13	0.14
		잘모름	0.25	0.25	0.25	0.25	0.25	0.25
	이혼에 대한 견해	절대 이혼은 안 됨	0.83	0.83	0.83	0.84	0.84	0.83
		가급적 안해야 됨	1.23	1.26	1.24	1.24	1.24	1.23
		해도 안해도 됨	1.26	1.30	1.27	1.26	1.27	1.26
		이유가 있으면 함	0.89	0.91	0.89	0.88	0.88	0.89
		잘모름	0.39	0.42	0.41	0.40	0.40	0.39
	가족관계 만족도	매우 만족	1.02	1.05	1.03	1.02	1.02	1.01
		약간 만족	1.23	1.26	1.23	1.23	1.23	1.22
		보통	1.28	1.32	1.29	1.29	1.29	1.29
		약간 불만족	0.53	0.54	0.53	0.54	0.53	0.53
		매우 불만족	0.24	0.25	0.24	0.25	0.25	0.23
환경	5년전 환경변화	매우 좋아짐	0.43	0.43	0.43	0.42	0.42	0.43
		약간 좋아짐	1.12	1.15	1.12	1.12	1.12	1.12
		보통	1.28	1.32	1.29	1.29	1.29	1.28
		약간 나빠짐	1.10	1.13	1.10	1.11	1.10	1.10
		매우 나빠짐	0.63	0.65	0.64	0.63	0.63	0.64

3) 평균오차 제곱근(RMSE)

평균오차 제곱근(RMSE)은 무응답 처리 후 추정값의 편향(bias)과 변동성(variability)을 동시에 측정 가능한 척도이다. 평균오차 제곱근은 정확성(accuracy)과 관련 되며, 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).

무응답 메커니즘별로 완전임의결측(MCAR)에 가장 적은 값을 가져 무응답 처리효과가 우수함을 알 수 있다. 특히 완전임의결측은 보조정보와 무관하게 적은 값을 가져 별도의 무응답 처리가 필요하지 않는 것으로 판단된다(<표 4.7>).

임의결측(MAR)도 완전임의결측(MCAR) 보다 높은 값을 갖지만 처리효과가 우수하며, 통합자료를 보조정보로 사용할 경우에 더욱 무응답 처리 성능이 높은 것을 알 수 있다. 특히, MAR2의 경우는 통합자료를 보조정보로 사용한 경우에 MAR1 보다 더욱 낮은 수치를 갖는다.

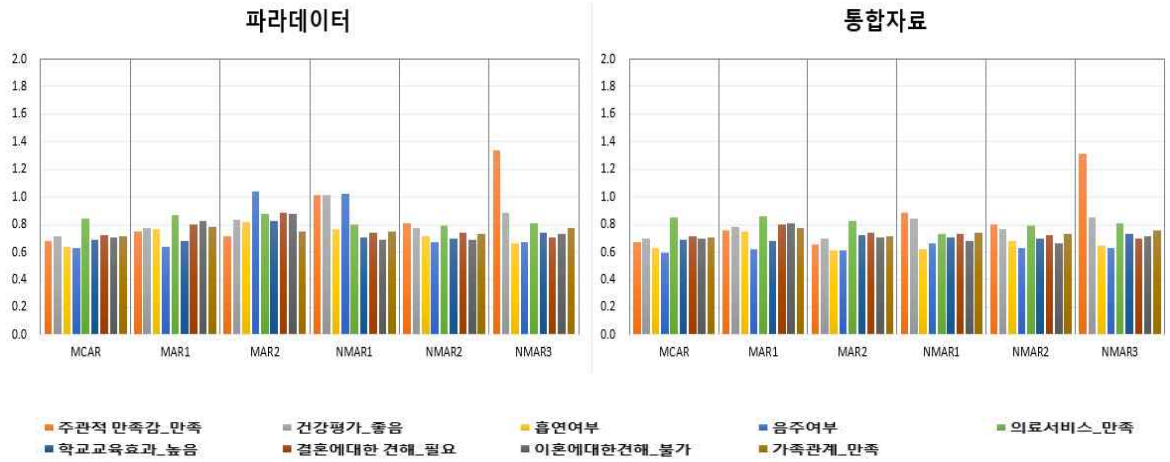
비임의결측(NMAR)은 상대적으로 높은 수치를 보이나, 통합자료를 보조정보로 사용한 경우에는 처리효과가 좋게 나타나는 것을 볼수 있다. 다만 NMAR3 경우는 편향이 상대적으로 큰 이유로 RMSE도 높게 나타났다.

상대편향의 결과와 마찬가지로 통합자료를 보조정보로 활용한 경우에 상대적으로 적은 값을 가져 적은 편향과 추정값의 분포가 안정된 점을 알 수 있다.

<표 4.7> 추정값의 평균오차 제곱근 비교

		파라데이터						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



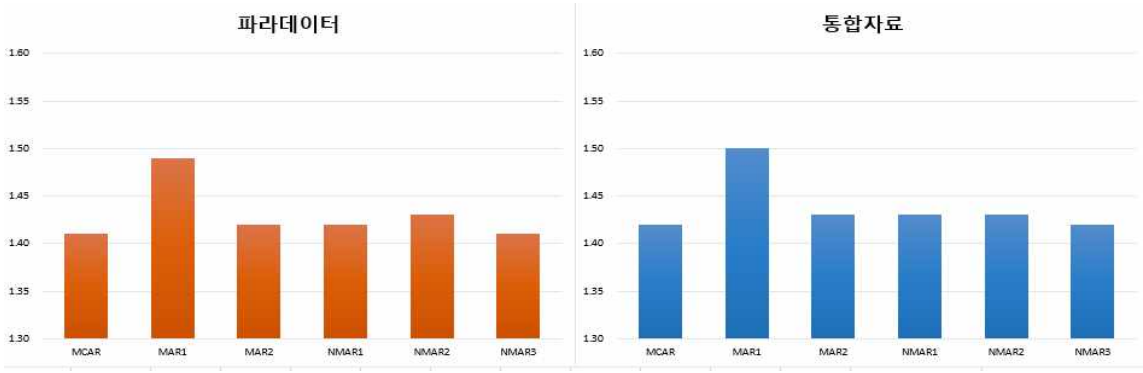
<그림 4.6> 추정값의 평균오차 제곱근 비교

4) 설계효과(design effect)

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 메커니즘별로 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 세분화된데 기인한다. 즉, 세분화된 층에 비해 상대적으로 표본규모가 작게 배정되어 분산의 확대가 커진 것으로 평가할 수 있다.

<표 4.8> 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 4.7> 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 확인했다.

5) 모형 비교

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.

가설1) $H_0 : log(\frac{\pi_k}{\pi_1}) = \beta_0\ (intercept-only)$, $H_1 : log(\frac{\pi_k}{\pi_1}) = \beta_0 + \beta_i X_i$, i는 파라데이터

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

가설3) $H_0 : log(\frac{\pi_k}{\pi_1}) = \beta_0 + \beta_i X_i$, i는 파라데이터, $H_1 : log(\frac{\pi_j}{\pi_1}) = \beta_0 + \beta_j X_j$, j는 통합자료

보조변수의 유의성 여부는 귀무가설 모델에 대한 대립가설 모델의 적합성 검정을 통해 확인이 가능하다. 가설1은 귀무가설이 기각됨에 따라 파라데이터를 이용한 모형이 통계적으로 유의한 것으로 나타났으며, 가설2는 통합자료를 이용한 모형이 크게 유의미한 결과를 보였다(<표 4.5>). 가설3은 파라데이터를 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 4.9> 우도비 검정(likelihood ratio test)을 통한 모형 비교

(단위: 건수/ 300회)

	$P-value < 0.05$					
	가설 1		가설 2		가설 3	
주관적 만족감	$Reject\ H_0$	300	$Reject\ H_0$	300	$Reject\ H_0$	300
건강평가	$Reject\ H_0$	300	$Reject\ H_0$	300	$Reject\ H_0$	300
흡연여부	$Reject\ H_0$	161	$Reject\ H_0$	300	$Reject\ H_0$	300
음주여부	$Reject\ H_0$	300	$Reject\ H_0$	300	$Reject\ H_0$	300
의료서비스 만족도	$Reject\ H_0$	300	$Reject\ H_0$	300	$Reject\ H_0$	300
학교교육 효과	$Reject\ H_0$	300	$Reject\ H_0$	300	$Reject\ H_0$	300
결혼에 대한 견해	$Reject\ H_0$	300	$Reject\ H_0$	300	$Reject\ H_0$	300
이혼에 대한 견해	$Reject\ H_0$	300	$Reject\ H_0$	300	$Reject\ H_0$	300
가족관계 만족도	$Reject\ H_0$	263	$Reject\ H_0$	300	$Reject\ H_0$	300

6) 시도별 무응답 처리결과

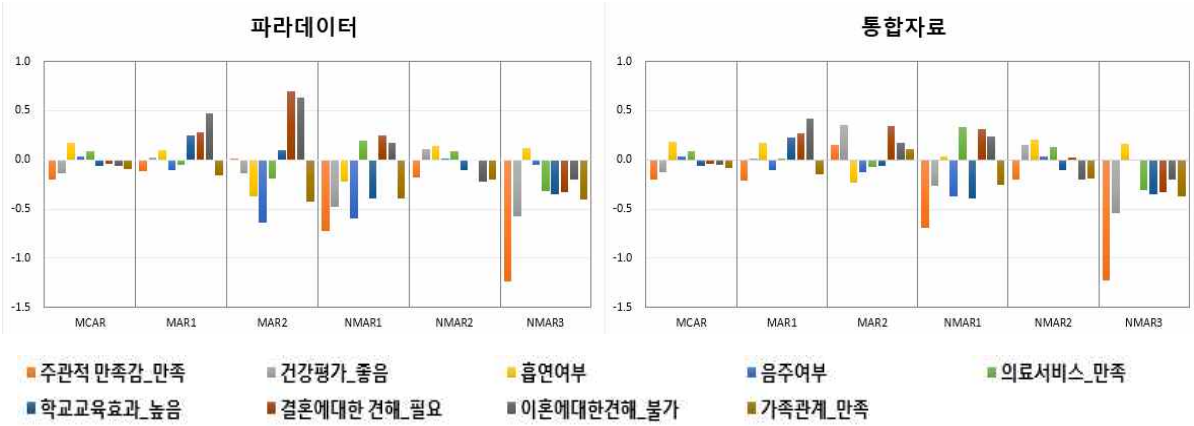
시도별 자료는 일정규모 표본규모를 갖춘 서울과 경기지역을 중심으로 살펴본다(<표 3.3>).

서울은 전국자료와 같이 파라데이터를 보조정보로 한 경우는 무응답 발생변수(e.g. 방문횟수)를 포함한 경우인 MAR1 자료와 NMAR2 자료는 상대적으로 처리효과가 높게 나타났다. 무응답 발생변수를 포함하지 않는 경우는 편향이 남아 있음을 알 수 있다. 통합자료를 보조정보로 한 경우는 무응답 발생변수를 포함하거나, 소득과 연관성 높은 보조정보를 갖춤으로써 무응답 처리효과가 높게 나타남을 볼 수 있다. NMAR3 주관적만족감의 불응여부는 보조정보와 연관성이 없어 처리효과 낮게 나타났다.

경기는 전국이나 서울자료와 같이 파라데이터를 보조정보로 한 경우는 무응답 발생변수를 포함한 경우에 처리가 높게 나왔으나, 포함하지 않는 경우는 편향이 남아있다. 반면 통합자료를 보조정보로 사용한 경우는 무응답 발생변수가 보다 많이 포함되거나 상관관계가 높은 변수가 포함되어 처리효과가 우수한 것으로 확인된다.

<표 4.10> 시도별 추정값의 편향 비교(서울)

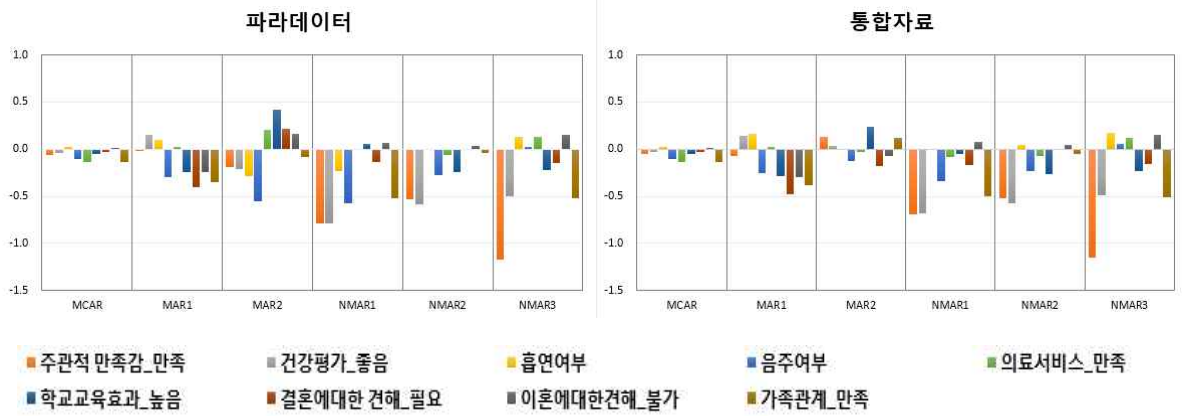
		파라데이터						통합자료					
		MCAR	MAR1	MAR2	NMAR1	NMAR2	NMAR3	MCAR	MAR1	MAR2	NMAR1	NMAR2	NMAR3
주관적 만족감	만족	-0.20	-0.12	0.00	-0.73	-0.18	-1.24	-0.20	-0.21	0.15	-0.69	-0.20	-1.23
	보통	0.13	-0.11	0.14	0.30	0.02	0.93	0.12	-0.04	-0.05	0.27	0.03	0.93
	불만족	0.07	0.23	-0.14	0.43	0.16	0.30	0.07	0.25	-0.10	0.42	0.17	0.30
건강평가	좋음	-0.14	0.02	-0.14	-0.48	0.11	-0.58	-0.13	0.01	0.35	-0.27	0.15	-0.54
	보통	0.17	0.10	-0.01	0.08	-0.08	0.38	0.16	0.05	-0.41	-0.10	-0.13	0.34
	나쁨	-0.03	-0.12	0.16	0.40	-0.03	0.21	-0.04	-0.06	0.06	0.37	-0.02	0.20
흡연여부		0.17	0.10	-0.37	-0.22	0.14	0.12	0.18	0.17	-0.23	0.03	0.20	0.16
음주여부		0.03	-0.10	-0.64	-0.59	0.01	-0.05	0.03	-0.10	-0.13	-0.37	0.03	0.00
의료서비스 만족도	만족	0.09	-0.05	-0.19	0.20	0.09	-0.32	0.09	0.01	-0.07	0.34	0.13	-0.30
	보통	-0.11	-0.06	0.10	-0.17	-0.08	0.46	-0.11	-0.13	-0.05	-0.33	-0.12	0.44
	불만족	0.03	0.13	0.11	-0.01	0.01	-0.13	0.03	0.13	0.13	0.00	0.00	-0.13
학교교육 효과	높음	-0.07	0.24	0.09	-0.40	-0.11	-0.35	-0.06	0.22	-0.06	-0.40	-0.11	-0.35
	보통	0.11	-0.33	0.08	0.12	-0.02	0.31	0.11	-0.30	0.25	0.12	-0.03	0.31
	낮음	-0.04	0.09	-0.18	0.28	0.14	0.04	-0.05	0.08	-0.19	0.27	0.14	0.04
결혼에 대한 견해	필요	-0.04	0.28	0.69	0.24	-0.01	-0.33	-0.04	0.27	0.34	0.31	0.02	-0.33
	중립	-0.03	-0.17	-0.70	-0.37	0.03	0.30	-0.04	-0.16	-0.33	-0.40	0.01	0.30
	불필요	0.08	-0.11	0.00	0.12	-0.01	0.03	0.08	-0.12	0.00	0.09	-0.03	0.03
이혼에 대한 견해	불가	-0.06	0.47	0.63	0.17	-0.22	-0.20	-0.05	0.42	0.17	0.24	-0.20	-0.20
	중립	-0.02	-0.38	-0.53	-0.18	0.18	0.03	-0.02	-0.38	-0.28	-0.26	0.16	0.03
	가능	0.07	-0.09	-0.09	0.01	0.04	0.17	0.07	-0.04	0.12	0.02	0.04	0.17
가족관계 만족도	만족	-0.10	-0.16	-0.43	-0.40	-0.20	-0.41	-0.09	-0.15	0.10	-0.26	-0.19	-0.38
	보통	0.03	0.03	0.44	0.41	0.12	0.39	0.03	0.03	-0.14	0.25	0.11	0.36
	불만족	0.06	0.13	-0.02	-0.02	0.07	0.01	0.06	0.12	0.03	0.00	0.07	0.01



<그림 4.8> 추정값의 편향 비교(서울)

<표 4.11> 시도별 추정값의 편향 비교(경기)

		파라데이터						통합자료					
		MCAR	MAR1	MAR2	NMAR1	NMAR2	NMAR3	MCAR	MAR1	MAR2	NMAR1	NMAR2	NMAR3
주관적 만족감	만족	-0.06	-0.02	-0.19	-0.79	-0.53	-1.17	-0.05	-0.07	0.13	-0.69	-0.52	-1.15
	보통	0.11	0.10	0.09	0.44	0.29	0.48	0.11	0.10	-0.09	0.38	0.29	0.47
	불만족	-0.05	-0.08	0.10	0.35	0.23	0.69	-0.06	-0.03	-0.04	0.32	0.24	0.68
건강평가	좋음	-0.04	0.16	-0.21	-0.78	-0.58	-0.50	-0.03	0.14	0.04	-0.68	-0.57	-0.48
	보통	0.10	-0.19	-0.20	0.17	0.45	0.34	0.10	-0.17	-0.10	0.22	0.46	0.36
	나쁨	-0.06	0.03	0.40	0.61	0.13	0.16	-0.07	0.03	0.06	0.46	0.11	0.12
흡연여부		0.02	0.10	-0.28	-0.23	-0.01	0.13	0.02	0.16	-0.01	0.00	0.05	0.17
음주여부		-0.10	-0.29	-0.55	-0.57	-0.27	0.02	-0.10	-0.25	-0.12	-0.34	-0.23	0.06
의료서비스 만족도	만족	-0.14	0.03	0.21	-0.01	-0.06	0.13	-0.14	0.02	-0.03	-0.08	-0.07	0.12
	보통	0.27	0.18	-0.17	-0.02	0.22	-0.08	0.27	0.18	0.01	0.04	0.23	-0.07
	불만족	-0.13	-0.21	-0.04	0.04	-0.16	-0.05	-0.13	-0.20	0.02	0.04	-0.16	-0.05
학교교육 효과	높음	-0.05	-0.25	0.41	0.05	-0.25	-0.22	-0.05	-0.29	0.23	-0.05	-0.27	-0.24
	보통	0.00	-0.08	-0.44	-0.06	0.18	0.14	0.00	-0.06	-0.25	0.01	0.18	0.16
	낮음	0.05	0.33	0.03	0.01	0.08	0.08	0.05	0.34	0.01	0.04	0.09	0.09
결혼에 대한 견해	필요	-0.03	-0.40	0.22	-0.13	0.00	-0.14	-0.03	-0.47	-0.18	-0.17	0.00	-0.15
	중립	0.00	0.34	-0.15	0.10	0.04	0.05	0.00	0.40	0.23	0.14	0.04	0.06
	불필요	0.04	0.07	-0.07	0.04	-0.03	0.10	0.04	0.07	-0.05	0.04	-0.03	0.10
이혼에 대한 견해	불가	0.00	-0.24	0.16	0.06	0.03	0.15	0.01	-0.30	-0.07	0.08	0.04	0.15
	중립	-0.03	0.13	-0.18	-0.02	-0.12	-0.09	-0.04	0.15	0.15	0.04	-0.12	-0.07
	가능	0.04	0.11	0.02	-0.04	0.10	-0.06	0.03	0.16	-0.07	-0.12	0.08	-0.08
가족관계 만족도	만족	-0.14	-0.35	-0.09	-0.52	-0.05	-0.52	-0.14	-0.39	0.12	-0.50	-0.06	-0.51
	보통	0.16	0.27	0.08	0.38	0.00	0.53	0.16	0.32	-0.09	0.39	0.01	0.54
	불만족	-0.02	0.06	0.01	0.13	0.04	-0.03	-0.03	0.06	-0.04	0.10	0.04	-0.04



<그림 4.9> 시도별 추정값의 편향 비교(경기)

5. 결론

본 연구는 2016년 사회조사 자료를 이용하여 무응답 처리를 위한 전체 과정을 단계적으로 살펴보고, 발생 가능한 모든 가정에 따라 모의실험 및 무응답 처리를 수행했다. 가용할 수 있는 다양한 자료원을 활용하여 실증적으로 무응답을 처리했으며, 작업과정에서 발견된 시사점 및 과제를 논의한다.

5.1 연구내용 요약

본 연구에서 무응답과 관련하여 논의된 내용을 요약하면, 무응답 편향에 영향을 미치고 있는 변수 탐색, 무응답 구조를 이해하기 위한 무응답 메커니즘별 시뮬레이션 및 무응답 처리결과 분석 등 다음과 같다.

첫째, 무응답 처리를 위해서는 일반적으로 무응답에 영향을 미치는 특성(변수)을 탐색해야 한다. 이러한 특성을 파악하기 위해서는 응답자와 무응답자의 정보를 동시에 가지고 있는 자료가 요구된다. 그러나 무응답에 대한 정확한 정보 수집은 현실적으로 매우 어렵다. 이러한 정보는 파라데이터, 표본추출틀 및 행정자료 등을 통해 수집이 가능하다(Kalton and Maligalig, 1991). 여러 가지 제약조건에도, 무응답률과 부차 그룹의 응답/무응답 구성비 차이를 이용한 편향 탐색(Groves, 1989), 오즈비(odd ratios)를 이용한 응답성향에 영향을 미치는 변수의 탐색(SIPP, 2017), 설계가중값과 사후 가중값의 비(ratio)인 사후층화 조정계수를 이용한 탐색(Ineke 등, 2010), 회귀모형에 의한 편향 탐색(Bethlehem, 2002), 응답가구 특성별 평균 응답성향 등을 이용하여 무응답 편향에 대한 탐색을 시도하였다.

이들 자료로부터 응답성향에 유의미한 차이가 있으면서, 무응답 편향의 갖는 변수로 지역(동부, 읍면부), 주택유형(단독주택, 아파트, 기타주택), 방문횟수(1회, 2회, 3회, 4회+), 점유형태(자가, 비자가), 가구원수(1인, 2인, 3인, 4인+), 가구주 성별(남자, 여자), 가구주 연령(20대, 30대, 40대, 50대, 60대+)로 찾을 수 있었다. 이렇게 탐색된 특성은 무응답 조정을 위한 보조정보로 활용이 가능하다.

둘째, 조사변수에 대한 무응답 구조를 이해하기 위해, 다양한 방법으로 무응답 메커니즘을 가정하여 접근했다. 무응답률의 변화가 무응답 편향에 유의한 영향을 주지 않을 수도 있고(Curtin 등, 2000; Keeteret 등 2000), 무응답률로 무응답 편향을 예측하지 못할 수도 있다(Groves, 2006). 즉, 무응답률 하락이 무응답 편향을 작게 하는 것이 아니라, 무응답 편향은 관심변수와 응답성향과의 관련성에 따라 달라진다 할 수 있다. 이러한 변동성에 의해 의존하는 무응답 메커니즘은 임의결측(MAR)과 비임의결측(NMAR)을 구분하는 데는 상당한 어려움이 따른다(David 등, 1986; Molenberghs 등, 2008). 다만, 무응답 메커니즘에 관련된 보조변수 확보 여부에 따라 비임의결측(NMAR)을 무시할 수 있는 임의결측(MAR)로도 바꿀 수 있다. 본 연구에서는 통계청 사회조사에서 발생되는 단위 무응답이 여러 형태의 메커니즘을 갖을 수 있다는 가정 하에 무응답을 발생하는 모의실험(simulation)을 실시했다.

무응답 발생은 6가지로 구분하여 진행했으며, 완전임의결측(MCAR)은 특정 변수와 관련 없이 무응답을 발생했다. 비임의결측(MAR)은 파라데이터를 이용한 MAR1과 행정자료를 이용한 MAR2로 구분하여 발생했다. 임의결측(NMAR)은 조사변수의 관측값(observed values) 뿐 아니라 비관측된 값(unobserved values)도 의지함에 따라 조사변수(survey variables)을 활용하여 NMAR1은 소득 수준에 따라 무응답 발생을 했으며, NMR2는 조사변수인 소득과 보조변수인 방문횟수를 결합하여 무응답을 발생했으며, NMAR3는 조사변수인 주관적 만족감에 따라 무응답을 발생시켰다. 여기에서 소득은 다른 변수(covariates)와 관련성이 높지만, 주관적 만족감은 보조정보와 관련성이 나타나지 않았다.

셋째, 두 가지 보조정보(auxiliary information)를 통해 각각 무응답 처리하고, 그 결과를 편향(bias), 상대편향(relative bias), 평균오차 제곱근(RMSE), 설계효과(design effect) 및 모형비교(model comparisons)를 통해 평가했다.

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. 두 가지 보조정보에 의한 무응답 처리에 대한 차이는 미미하다.

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 됨 없이 정확성을 낮추는 결과를 가져와 노력의 의미가 없어질 수 있다. 따라서 보조정보는 응답성향(응답/무응답)과 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).

5.2 시사점 및 향후 과제

표본 조사는 가구를 방문해서 개인을 통하여 응답을 얻어야 하므로 어느 시대에서도 어느 정도의 무응답이 포함될 수밖에 없다. 또한 조사의 고유한 특성 때문에 다양한 오차를 갖기 때문에 이들을 이해하고 총합적인 관리가 필요하다. 본 연구에서 발견된 시사점과 도출된 과제는 다음과 같다.

첫째, 최근 무응답 동향을 보면 조사환경이 지난 시점과 비교해 그다지 우호적이진 않다. 단위 무응답(unit non-response)이 발생하는 이유는 다양하다. 이를 위해 무응답을 유형화하여 구분할 필요가 있다(Groves and Couper, 1998; Lynn and Clarke, 2002). 응답을 시작했으나 도중에 그만 두

는 부분응답(partial response), 표본이 조사에 응답을 거부하는 응답 거부(refusal), 조사기간 동안 접촉이 불가능한 경우인 비접촉(noncontact), 언어나 건강상의 이유로 인하여 표본이 응답을 할 수 없는 기타 무응답(inability)으로 구분할 수 있다.

다만, 최근 경향은 사생활 노출기피, 바쁘다는 이유로 응답 거부가 무응답의 높은 비중을 차지하는 현상을 볼 수 있으며(Atrostic 등 2001, Brick and Williams 2013), 무응답은 여러 가지 이유로 지속적인 증가 추세를 보이고 있다.

다행히도 우리나라 전체 무응답은 미국 또는 유럽에 비해 낮아 조사환경은 아직은 좋은 편이라 할 수 있다. 무응답은 조사의 품질에 영향을 미치므로 가능한 최소화할 필요가 있다(Alreckand Settle, 1995). Babbie(2007)는 사회과학 연구를 위해 70% 이상의 응답률이 요구되며, Singleton과 Straits(2005)는 면접조사는 85% 이상의 응답률 유지를 제안하고 있다.

기본적으로 널리 알려진 무응답을 줄이는 방법으로는 조사를 시작하기 전에 보내는 사전 안내문(advance letter)을 보내는 것과 조사 협조를 위한 조사에 응답 여부와 관계없이 답례품(financial incentive)을 제공(Curtin 등, 2007; Cantor 등, 2008), 설문지 구성을 쉽게 하고 길지 않게 구성(Dillman 등, 1993), 방문(접촉) 횟수를 높이는 것(Dillman, 1978, 2000), 추가 조사가구 방문시점을 바꾸는 것(Politz and Simmons, 1949, 1950; Weeks 등, 1980), 자료 수집방법을 바꾸는 것(Peytchev 등, 2009), 조사 기간에 대해 충분한 제공(Weeks, 1988), 경험이 풍부한 조사원 활용 등이 있다(이승희, 2010).

부재 가구와 불응 가구에 대해 접촉횟수를 증가시켜 부재율을 감소할 필요가 있고, 불응가구에 대해서는 응답 전환(refusal conversion)을 높일 필요가 있다(김서영과 박라나, 2013)[6]. 특히 불응가구에 대해서는 경험이 많은 조사원이나 슈퍼바이저(supervisor)를 투입하거나 별도의 인센티브를 제공하는 방법도 필요하다. 또한 응답자가 원하는 조사방법을 선택할 수 있도록

6) 1인 가구, 미혼가구 등에서 부재성향이 높게 나타나고, 불응가구와 부재가구가 특성에 차이를 보이고 있어 무응답을 처리에서 이를 구분하여 처리하는 방법도 생각해 볼 수 있다.

인터넷 조사 방식 사용 또는 혼합조사(mixed-mode)를 적극 활용하는 것도 무응답을 줄일 수 있는 좋은 방법이 될 것이다. 정기적인 조사를 마치고 무응답자를 대상으로 추가조사 하는 방법도 있다(Biemer, 2001).

조사 설계자와 현장담당자의 무응답 축소를 위한 노력에도 조사환경은 더욱 악화되어 무응답 증가는 지속될 것으로 전망된다(Kalton, 2018). 현장에서 더 이상 처리되지 않는 무응답은 통계적인 방법을 이용해 처리하여 무응답 편향을 줄일 수 있다. 가중값을 이용한 여러 가지 통계적 보정방법(Montaquila 등, 2008; Lee and Valliant, 2007; Kalton and Flores-Cervantes, 2003; Bethlehem, 1998, 2002; Little, 1986)은 대부분의 표본조사에서 무응답으로 발생되는 편향을 줄이는 데 사용된다. 최근에는 반응적 표본설계(responsive survey design)를 이용하여, 표본추출틀이나 방문으로 수집할 수 있는 표본 단위의 특성을 사용하여, 조사와 동시에 무응답 오차를 수량화하여 이를 최소화하는 것으로 설계특성을 표본 가구에 맞추는 방법이다.(Groves and Heeringa, 2006).

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. 또한, 국제 무응답 학회(International Workshop on Household Survey Nonresponse)를 통해서도 무응답 연구에 대한 국제적 협력과 새로운 접근 방식을 같이 할 수 있을 것으로 판단된다.

둘째, 무응답에서 적정한 보조정보 확보는 조사를 통해 응답 정보를 얻는 것만큼이나 중요하다. 보조정보는 응답성향과 관련성이 높으면서 동시에 조사변수와 관련성도 높아야 한다(Rizzo, 1996; Matei and Ranalle, 2015). 이를 보다 통계학적으로 접근하여 설명하자면 다음과 같다.

표본조사에서 발생되는 대부분의 무응답 처리를 위해 무응답 메커니즘

을 임의결측(MAR)이라 가정한다. 이는 먼저 모든 특성에서 동일한 확률로 무응답이 발생한다는 완전임의결측(MCAR)은 현실적이지 못하기 때문이다. 또한, 무응답 메커니즘이 임의결측(MAR)으로 판단될 경우에 보조정보에 의해 관심 모수에 대한 추론이 가능하기 때문이다(통계교육원, 2010). 이런 이유로 무응답 처리 기법을 적용하기 위해서는 무응답 메커니즘이 임의결측(MAR)을 따르도록 무응답 발생과 연관성 높은 보조변수를 포함하여 한다. 관련 보조정보가 없거나 결핍된 경우에는 전통적인 무응답 처리방법인 가중값 조정이 유효하지 않을 수 있다(Guillaume, 2013).

다만, 무응답 메커니즘을 결정하기 어려운 이유는 메커니즘 자료가 어떤 변수를 포함하고 있는지에 따라 바뀔 수 있는 불확실성 때문이다(통계교육원, 2010). 따라서 무응답 자료에 대한 분석에서는 무응답의 발생과 관련된 변수를 포함하여 무응답 메커니즘이 비임의결측(NMAR) 보다는 임의결측(MAR)이 되도록 만들어야 한다.

예로서 소득이 무응답 발생과 관련이 된다면 소득과 밀접하게 관련된 세금, 자산, 지출 등의 정보를 포함하여 무응답 메커니즘을 비임의결측(NMAR)을 임의결측(MAR)으로 바꿀 필요가 있다. 한편, 무응답 발생여부는 소득뿐 아니라 지나치게 개인적이거나 민감한 내용을 포함한 조사내용을 포함한 경우에 발생되므로 무시할 수 없는 비임의결측(NMAR)에 대한 연구도 지속적으로 함께 해야 한다. 보다 상세한 내용은 Zhang(2002), Qin, Leung and Shao(2002), Beaumont(2000), Copas and Farewell(1998), Little and Rubin(1987), Greenlees, Reece and Zieschang(1982) 등을 참조할 수 있다.

셋째, 주요 보조정보 자료원 중 하나인 파라데이터 수집에 투자를 강화하고 활용성 제고에 주력해야 한다. 파라데이터(paradata)는 조사과정에서 일어나는 모든 과정을 포함하는 자료로 무응답 처리에 활용할 뿐 아니라, 조사설계 및 조사과정 개선, 조사원 평가 등 다양하게 활용할 수 있다. 특히 조사원과 가구간의 발생되는 현장 조사의 기록으로 무응답 사유, 방문시간, 응답시간, 방문횟수, 협조도, 응답전환(refusal conversion) 정보 등을 얻을 수 있어 조사전반을 이해하고 조사품질을 높이는데 활용 가능하다.

이런 이유로 모든 국가들이 파라데이터 수집과 활용을 위해 노력을 하며, 특히 컴퓨터를 이용한 조사현장의 추적파일(trace file) 또는 로그파일(log file)을 통한 면접별 진행시간, 조사 과정에서의 의미 있는 정보들을 다양한 형태로 탐색되고 있다(김정섭과 임경은, 2010). 다만, 우리의 경우는 파라데이터 수집을 아직 컴퓨터를 이용한 자동시스템이 아닌, 조사원의 최종 기록을 수집하고 있고, 입력된 자료의 내용에 대해 체계적인 내검과 정리가 다 하지 못한 상태이다.

다른 국가의 파라데이터 수집 현황을 보면, 미국 센서국은 PANDA (Performance and Data Analysis) 시스템을 통해 지역, 면접 대상, 하루 동안 진행된 면접 시간, 면접 결과, 이상치(outlier), 조사와 관련하여 조사원들에 의해 허위 기록이 가능한 내용, 즉 빈집 비율이 너무 높거나 단독가구가 많은 경우, 일반적 시간이 아닌 자정~오전 8시에 이루어진 면접, 면접 시간이 너무 짧은 경우 등을 확인 할 수 있다.

미국 건강센터는 Blaise 시스템을 통해 접촉횟수, 접촉시도(방문, 전화), 접촉날짜(시간), 성공여부, 조사거부 이유(바쁨, 무관심 등), 성공 전략, 실패 사유, 빈집 여부, 재방문 약속방법(약속 카드, 메모 등) 등을 수집하고 있다.

캐나다 통계청은 횡단면, 사회조사 및 농업조사, 장기조사, CATI에 의한 RDD 등 광범위한 조사에서의 파라데이터를 이용한 연구를 진행하며, 시스템 작업, 연락 시도 횟수, 접촉비율, 전화 거는 양상, 그리고 투입과 산출 간의 관계 등에 대한 분석을 하고 있다.

일반적인 표본조사는 무응답 오차 뿐 아니라 다양한 오차를 포함하고 있다(<표 5.2>). 표본오차(sampling error)를 제외하고는 나머지 비표본오차(non-sampling error)는 측정이 어렵고 이론적인 설명이나 추론 또한 어렵다. 비표본오차는 포함오차, 측정오차, 처리오차 등이 있다. 이러한 비표본오차를 축소할 수 있도록 조사과정 자료인 파라데이터는 매우 중요한 역할을 한다. 즉, 파라데이터는 조사관리 측면, 조사설계 측면, 총오차 측면 및 통계품질 측면 등 다양한 방면에서 활용할 수 있다(김정섭과 임경은, 2010).

조사관리 측면을 보면 조사원의 기록에 대한 활용 부문이다. 접촉하기 위

해 사용한 기록, 방문횟수, 실제조사 수행과정, 재방문을 위한 조치(협조 공문, 메모, 사전 전화 등)등을 통해 효과적인 사전조치, 방문시간, 적정 소요시간 및 적정 방문횟수(비용, 시간 등)등을 파악하여 조사 관리의 전략수립에 활용이 가능하다.

조사설계 측면에서는 조사표에서의 각 문항마다 응답하는 시간을 파라데이터를 통해 수집된다면, 응답자의 인지 수준과 문항 위치(순서), 워딩(wording), 조사표 구성 등에서 발생된 문제점에 대한 원인 파악 및 해결방안을 찾을 수 있다.

총오차 측면에서 활용은 발생되는 무응답을 세부적으로 살펴, 표본설계나 조사에 활용할 수 있다. 예로서 다문화 가족에 대한 조사표 준비, 과대표집, 적정 조사시간 등을 활용하여 무응답을 최소화하는 등 전체적인 비표본오차를 감소할 수 있다.

통계품질 측면은 파라데이터는 궁극적으로 통계품질을 향상을 위해 사용된다. 조사표 개선, 무응답 최소화 등의 작업은 정확성(accuracy), 접근가능성(accessibility), 관련성(relevance) 등을 제고할 수 있고, 적절한 측정도구 사용은 통계의 비교가능성(compatability)과 일관성(coherence)에도 영향을 줄 것이다.

파라데이터의 수집의 중요성과 다양한 활용성을 충분히 인지하고 있음에도 통계청 사회조사의 파라데이터 수집방식은 아직은 초기단계로 조사원이 모든 과정을 기록하기 보다는 종이로 기록해온 자료에 대해 최종 결과만을 입력(coding)되어 조사과정을 전반적인 과정을 이해하기 어렵다. 또한 분석자료로서 활용하기 보다는 참고자료로 활용하고 있어 체계적인 입력과 관리가 부족한 실정이다.

파라데이터의 체계적인 수집과 정리에 대한 재정립이 필요하다(West, 2011). 컴퓨터 등 정보화 기술을 기반으로 한 자료 입력과정을 자료로 만들어 평가함으로써 조사과정의 품질을 개선할 필요가 있다(Biemer and Caspar, 1994). 예를 들어, 시스템을 통한 자동생성(log), 각 문항 당 걸리는 시간, 전체 설문을 작성하는데 걸린 시간 등을 DB화하여, 조사의 품질 및

신뢰도 등을 측정하는 자료로 이용할 수 있어야 한다. 만약 특정 문항에 대해 지나치게 응답 시간이 길거나 짧다면, 관련 문항은 조사표 재설계하는 과정에서 수정이 필요할 것이다. 또한 특정 문항에 있어 응답자가 이해하는데 어려움이 있거나 오해할 여지가 있다면 해당 문항에 대해 수정을 하거나 없애는 등의 조치가 요구되며, 조사표 문항의 순서를 변경하는 것도 고려할 수 있다(김정섭과 임경은, 2010).

넷째, 조사에 있어서 행정자료의 역할 확대 및 새로운 가치 제공을 들 수 있다. 행정자료는 모든 단위를 파악할 수 있어 포함률(coverage) 높고, 가구의 접근성 및 조사원의 기록에 의존한 것에 비해 좋은 품질의 자료를 낮은 비용(less costly)에 이용할 수 있다는 장점이 있다. 아울러 또한 행정자료는 조시작 단계인 표본추출틀 구축에도 활용되며 다양한 정보를 사용하여 최적의 조사환경을 만드는 적응적 표본설계(adaptive sampling) 수행, 무응답 조정 및 사후층화 보정에도 활용이 가능하다.

특히 행정자료의 무응답에 대한 활용성은 앞서 살펴본 것처럼 소득 등 민감한 항목은 응답여부에 영향을 미치고 있어, 이에 대한 보조정보가 없거나 부족한 경우에는 전통적인 무응답 처리기법을 활용할 수 없다. 소득에 대한 직접적인 정보를 활용할 수 없다면 이와 밀접하게 연관성 있는 세금, 공시지가 및 실거래가 등은 무응답 편향 축소에 크게 기여할 것이다. 비록 무시할 수 없는 무응답 메커니즘인 비임의결측(NMAR)을 행정자료를 통한 보조변수 확보에 따라 무시할 수 있는 임의결측(MAR)으로도 바꿀 수 있어, 다양한 행정자료 확보를 위한 노력은 매우 중요한 과제이다.

행정자료가 표본조사에 미치는 가장 중요한 역할은 매년 최신의 표본추출틀(sampling frame)을 갱신(update)할 수 있다는 것이다. 표본추출틀을 최신의 정확한 개별단위(인구, 가구, 주택 등)로 구축함으로써 다양한 추출단위(리스트, 영역) 및 추출방법 등 표본설계, 무응답 조정 및 사후층화에 적용이 가능하다. 따라서, 기존에 설정된 조사구(enumeration district : ED)라는 공간 정보(공간 DB)에 도로명 주소로 구성된 거처와 주택(등록 DB) 등을 연결하여 매년 가구 및 인구에 관한 정보를 최신화(update)을 하고 있다. 이를 위해 통계청은 행정자치부, 국세청 정부기관들과 협력을 통해 다

양한 자료를 입수하고 있고, 특히 2015년부터는 14개 기관으로부터 25종 행정자료를 활용하여 매년 센서스 전수조사를 공표하고 있다(통계청, 2017).

<표 5.1> 공간 DB 및 등록 DB 보완

공간DB	등록DB
• 행정 및 조사구역 경계변경 • 신축 및 철거 건축물 반영 • 공동주택 전개도 생성 • 조사구요도 제작 등	• 등록기반 주택, 가구, 인구 자료 연계 • 표본조사자료를 이용한 속성정보 생성 • 전체 조사구역에 표본정보 공유 • 조사구역(공간DB)별 속성 분석 등

미국의 사례를 보면, 표본추출틀은 센서스 주소를 우편국의 행정자료 등 여러 경로를 통해 매년 보완 후 MAF(master address file) 구축하고, 지역, 센서스 트랙, 블록 그룹 및 블록 등을 매칭키로 이용하여 TIGER (Typologically Integrated Geographic Encoding and Referencing System) 시스템과 연계했다(Census Bureau, 2000, 2015). 이후 미 센서스국은 지속적으로 주택변수 및 가구변수를 추가하여 보강된 MAF(Augmented Master Address File)를 구축하여 미국 인구조사(ACS) 및 다른 조사에 활용하고 있다(Citro, 2014). 미국의 표본추출틀 구축 및 보완에 대해서는 <BOX 1>을 참조할 수 있다.

<BOX 1> 미국의 표본추출틀 구축 및 보완

① USPS Files : 우편국으로부터 DSF를 받아서 1년에 두 번 보정

- Delivery Sequence File(DSF) : 우편배달 master list(Address, ZIP code)
  ⇨ DSF update가 MAF에 없으면 MAF에 new record 생성
- Locatable Address Conversion System(LACS) 비교*
  ⇨ New record가 단순 주소 변경 또는 신축인지 확인 또한 동일 주소로 중복 레코드인지 확인
  * 도로명 street, road, terrace, lane, court, place 등 매칭

② Address Updates from ACS Personal Visit

⇨ 조사원(FRs)이 ACS를 위해 가구(HU)를 직접 방문시 수정 보완

③ **Demographic Area Address Listing Files(DAAL)**

⇨ CPS, SIPP, NCVS, AHS, ACS 등의 타조사에서 얻은 정보로 보완*

* 조사원의 랩톱 컴퓨터에 GPS 위도 경도 위치 수집 및 수정이 가능 ; MTdb 보완

④ **Community Address Updating System(CAUS)**

- 주소는 있으나 DSF에 나타나지 않는 경우
- MAF에는 없고 DSF에 있는 비도시 주소
- DSF와 MAF는 있으나, ACS 프레임에 없는 경우
- 이미 MAF에 사라진 주소이나, ACS 프레임에 있는 경우

⇨ ACS 프레임을 보정, 특히 비도시지역(non-city) 대상

⑤ **Geographic Support Systems Initiative(GSS-I)**

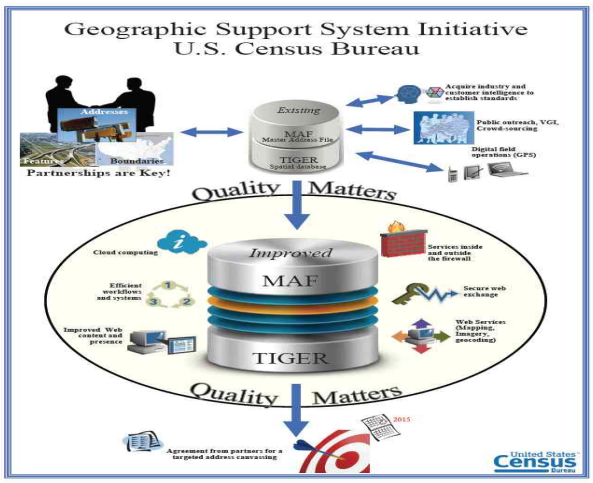
⇨ Address coverage, road update, MAF/TIGER 품질평가를 위해 연방, 주(State), 지방(Local), 부족(Tribal) 정부 등이 참여하여 MTdb 보완

⑥ **Commercial Data Sources(Third-party address data)**

⇨ Address간 매칭되지 않는 비율이 높으나, MAF에 추가되는 경우는 GSS-I 보다 낮음

⑦ **Administrative Records**

* Indian Health Services Registration, Housing and Urban Development public etc.



<그림12> GSS 운영과정

⑧ **2010 Census Field Operation(현장조사에서 보완)**

< Address List Building Operations >

- Address Canvassing
- Group Quarters Validation(GQV)

< Enumeration Operations >

- Group Quarters Enumeration
- Housing Unit Enumeration
- Enumeration of Transitory Locations
- Non-response Follow-up
- Filed verification

또한, 최근 행정자료를 이용하여 직접 통계를 생산하거나, 조사통계와 연계하여 통계를 작성하는 움직임이 활발하다(Citro, 2014). 행정자료는 저비용이며 높은 포함률을 지니고 응답자 입장에서는 응답부담(response burden)가 없고, 응답오차(response error)가 없다는 장점이 있다. 또한 표본조사를 행정자료로 대체할 수 있다는 주장도 있다(Hand, 2018). 다만, 이러한 행정자료의 매력에도 자료의 확인(cleaning) 작업이 필요하고, 행정을 위해 필요한 대상만이 수집되어 있을 수 있으며, 제한된 일부 항목만 존재하여 전체적으로 볼 수 없으며 행정자료 고유한 오차가 포함되어 있다. 또한 개인정보와 관련하여 비밀보호(confidentiality)가 요구되고, 연계를 위해 확률적(probabilistic) 연계 작업이 필요한 경우가 빈번하다. 이런 이유로 표본조사는 응답자로부터 많은 내용과 행정자료로 얻을 수 없는 주관적 내용 및 태도에 대해 파악이 가능하므로, 행정자료가 완전히 표본조사를 대체하는 것은 한계가 존재한다(Kalton, 2018).

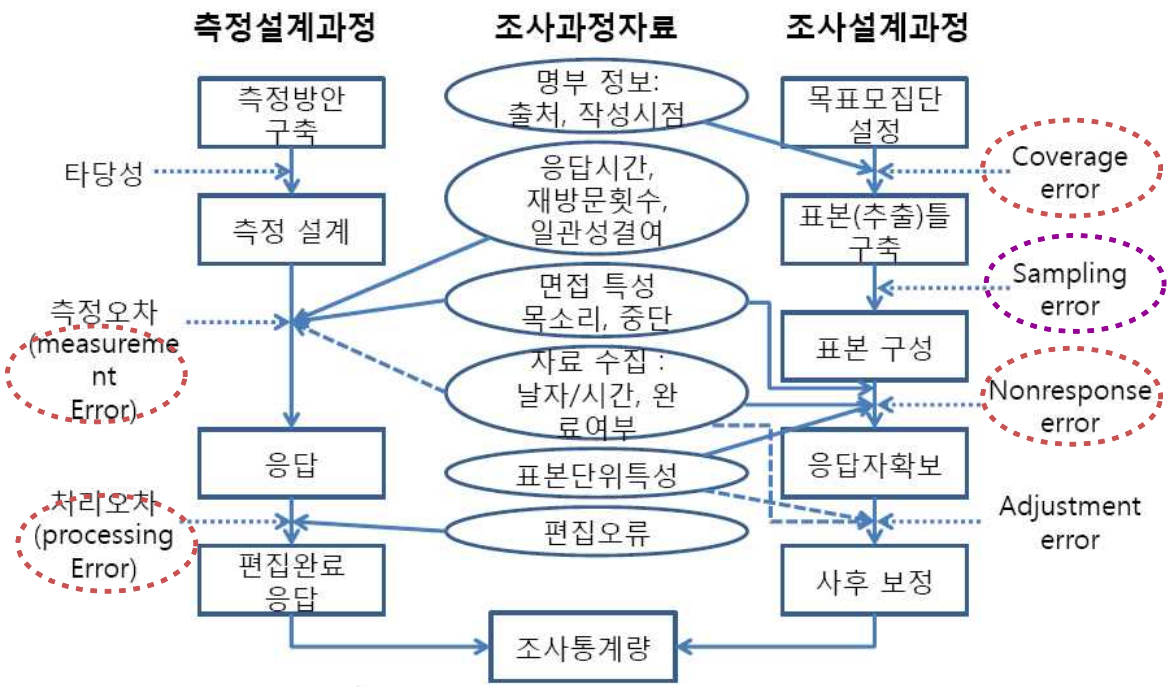
마지막으로, 본 연구는 무응답 오차(nonresponse error) 및 처리에 관한 측면만을 지나치게 강조함이 없지 않다. 하지만, 총조사 오차(total survey error)를 구성하는 표집오차(sampling error), 포함오차(coverage error), 측정오차(measurement error), 처리오차(processing error) 등도 함께 관심을 기울려 전체적으로 조사의 완전성을 높일 필요가 있다(<표 5.2>).

표본조사에서 총오차(또는 총조사 오차)는 크게 표집오차(sampling error)와 비표집오차(non-sampling error)로 구분할 수 있고, 비표집오차는 포함오차, 무응답 오차, 측정오차, 처리오차 등으로 나눌 수 있다.

표집오차(sampling error)로서 센서스와 같은 전수조사를 하지 않고 일부 표본을 선택해서 표본조사를 함으로 발생되는 오차로 표본크기에 따라 표집오차를 계산할 수 있어, 측정 가능한 오차이다. 다만, 표본조사라고 하더라도 비확률표본의 경우는 표집오차를 계산할 수 없다. 이를 위한 연구도 꽤 활발히 진행되고 있다(Abadie and Imbens, 2002; Schonlau 등, 2004, 2009; Terhanian 등, 2000; Lee and Valliant, 2009).

포함오차(coverage error)는 목표 모집단(target population)과 실제 조사에 활용되는 조사 모집단(survey population) 또는 표본추출틀(sampling frame)과의 포함률(coverage) 차이에 의해 발생되는 오차이다. 특히, 인터넷 조사의 포함률 문제는 오랫동안 지속돼 온 문제로, 사용자와 비사용자 간에 특성에서 차이가 확연하다(Couper, 2000; Couper and Miller, 2008; Norris, 2001 Rice, 2003). 또한 전화조사에서도 비등재되어 있는 경우가 많아 똑 같은 포함률 문제를 안고 있다.

<표 5.2> 표본조사에서 발생되는 오차



무응답오차(nonresponse error)는 앞서 내용과 같이, 응답 거부, 비접촉, 건강 및 언어 등의 이유에 의한 기타 무응답 등에 의해 발생되는 단위 무응답과 일부 항목에서 발생되는 항목 무응답에 의해 발생되는 오차이다. AAPOR(2009)[7]는 핵심질문은 전체를 응답한 경우와 그 외 질문은 80% 이상응답을 한 경우만 완전응답으로 여기고, 응답한 문항이 전체 문항의 50% 이상이 된 경우 부분응답이고 나머지는 응답 포기 또는 무응답으로 판다한다.

측정오차(measurement error)는 조사를 진행하는 과정에서 측정도구를 면접조사, 전화, 인터넷으로 또는 자기기입 방식이나 조사원 기입에 따라 인식정도 크게 달라진다(Bishop 등, 1986). 조사표 내용, 표현방법, 순서에 따라 이해 정도는 크게 달라져 이에 따른 영향을 받는다(Tourangeau 등, 2006; Belson, 1981; Mathiowetz and Wunderlich, 2000). 응답 척도를 5점 척도로 사용하는지 또는 다지 선다형을 사용하는지에 따라 측정 오차가 다르게 발생된다(Couper 등, 2004). 응답자가 쉽게 허위로 응답이 가능한 사회적 바람(social desirability)은 측정오차에 영향을 주고 있다(Phillips and Clancy, 1972).

처리오차(process error)는 수집된 자료를 처리하는 과정에서 일어나는 오차로 입력 오류와 에디팅 오류 등에 의해 발생한다.

조사과정에서 발생되는 모든 오차는 조사의 품질과 신뢰성에 영향을 미친다. 따라서 조사설계자와 조사현장에서 총오차를 줄이기 위해 총합적인 전략이 필요하고, 각 오차별로 최적의 체계적인 축소 방안을 모색해야 한다.

7) APPOR(2009)는 ① 응답해 야 할 문항 중 50% 미만(중도포기), 50~80% 미만(부분응답), 80% 이상(완전응답)으로 간주, ② 전제문항 중 50% 미만(중도포기), 50~80% 미만(부분응답), 80% 이상(완전응답)으로 간주, ③ 핵심 질문 중 응답 한 항목이 전체 핵심 질문의 50% 미만(중도포기), 50~99% 이하(부분응답), 100%(완전응답)으로 간주하고 있음

참고문헌

김규성 (2000), 무응답 대체 방법과 대체 효과, <조사연구>, 1(2), 1-14.

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.

변종석, 한혜은 (2011). 무응답편향: Callbacks and Substitutions, 국가통계방법론 심포지엄(통계청), 발표자료.

이승희 (2010). 총조사 오차의 패러다임으로 이해하는 표본조사, <통계연구> 15(1), 44-74.

정구현 (2014). 가계금융・복지조사에서 패널무응답 처리를 위한 가중 층 조정방법 연구, 박사학위 논문.

정구현, 이석훈, 변종석 (2013). 가계금융 ・ 복지조사의 무응답 처리를 위한 유용한 보조정보 선정, <조사연구>, 14(1), 69-91.

통계교육원 (2010). 무응답 자료 처리 및 분석.

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>

AAPOR(2009), Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys, The American Assocition for Public Opinion Research.

Abadie, A. and Imbens, G. (2002). Simple and Bias-Corrected Matching Estimators. Technical Report, Department of Economics, UC Berkeley.

Alreck, P.L. and R. B. Settle (1995). *The Survey Research Handbook: Guidelines and Strategies for Conducting a Survey*, second edition.

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.

Babbie, E.R. (2007), *The Practice of Social Research,* Wadsworth Publishing Company.

Belson, W. (1981). *The Design and Understanding of Survey Questions*. Aldershot: Gower Publishing.

Beaumont, J.-F. (2000). An estimation method for nonignorable nonresponse. *Survey Methodology*, 26(2), 131-136.

Bethlehem, J.G. (2012). Using response probabilities for assessing representativity, Discussion Paper, Statistics Netherlands.

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.

Biemer, P.P. (2001). Nonresponse Bias and Measurement Bias in a Comparison of Face to Face and Telephone Interviewing, *Journal of Official Statistics,* 17, 295-320.

Bishop, G., Hippler, H.J., Schwarz, N., and Strack, F. (1988). A Comparison of Response Effects in Self-Administered and Telephone Surveys, In: Groves, R.M., et al., eds. *Telephone Survey Methodology.* 321-40. New York: Wiley.

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.

Cantor, D., O'Hare, B.C., and O'Connor, K.S. (2008). The Use of Monetary Incentives to Reduce Nonresponse in Random Digit Dial Telephone Surveys, In: Lepkowski, J.M., et al., eds. *Advances in Telephone Survey Methodology*. 471-498. New York: John Wiley.

Citro C. (2014). From multiple modes for surveys to multiple data sources for estimates. *Survey Methodology*. 40(2), 137 - 161.

Copas, A.J., and Farewell, V.T. (1998). Dealing with non-ignorable non-response by using an 'enthusiasm-to-respond' variable. *Journal of the Royal Statistical Society*, Series A, 161, 385-396.

Couper, M.P. (2000). Web Surveys: A Review of Issues and Approaches, *Public Opinion Quarterly*, 64, 464-494.

Couper, M.P., and Miller, P.V. (2008). Web Survey Methods: Introduction, *Public Opinion Quarterly*, 72, 831-835.

Couper, M.P., Kenyon, K., and Tourangeau, R. (2004). Picture This! An Analysis of Visual Effects in Web Surveys, *Public Opinion Quarterly*, 68(2), 255-266.

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.

Czajka, J. L., Hirabayashi, S.M., Little, R.J.A., and Rubin, D.B. (1992). Projecting from Advance Data Using Propensity Modeling: An Application to Income and Tax Statistics. *Journal of Business & Economic Statistics, American Statistical Association*, 2, 117-31.

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.

Dillman, D.A. (1978). *Mail and Telephone Surveys: The Total Design Method.* New York: Wiley-Interscience.

Dillman, D.A., Sinclair, M.D., and Clark, J.R. (1993). Effects of QuestionnaireLength, Respondent-Friendly Design, and a Difficult Question on Response Rates for Occupant-Addressed Census Mail Surveys, *Public Opinion Quarterly*, 57, 289-304.

Dillman, D.A. (2000). *Mail and Internet Surveys: The Tailored Design Method. 2nd Edition.* New York: John Wiley.

Greenlees, J.S., Reece, W.S., and Zieschang, K.D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association*, 77, 251-261.

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.

Groves, R.M., and Heeringa, S.G. (2006). Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs, J*ournal of Royal Statistical Society* Series A., 169(3), 439-457.

Guillaume, O. (2013). Dealing with Non-ignorable Non-response using Generalised Calibration: A Simulation Study based on the Luxemburgish Household Budget Survey, Working Papers.

Hand, D.J. (2018). Statistical challenges of administrative and transaction data, *Journal of the Royal Statistical Society.*

Iannacchione, V.G., Milne, J.G., and Folsom, R.E. (1991). Response Probability Weight Adjustment Using Logistic Regression. *Proceedings of the American Statistical Association, Section on Survey Research Methods,* 637-642.

Ineke, S., Jaak, B., Achim, K., and Rory, F. (2010). *Improving Survey Response: Lessons learned from the European Social Survey*, Wiley.

Kalton, G. (2018). The Past, Present, and Future of Social Survey, 2018 JSPM Distinguished Lecture.

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.

Keeter, S., Miller, C., Kohut, A., Groves, R.M., and Presser, S. (2000). Consequences of Reducing Nonresponse in a National Telephone Survey, *Public Opinion Quarterly* 64,125 - 48.

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.

Peytchev, A., Baxter, R.K., and Carley-Baxter, L.R. (2009). Not All Survey Effort is Equal: Reduction of Nonresponse Bias and Nonresponse Error, *Public Opinion Quarterly*. 73(4), 785-806.

Phillips, D.L. and Clancy, K.J. (1972). Some Effects of 'Social

Desirability' in Survey Studies, *The American Journal of Sociology,* 77, 921-940.

Politz, A.N. and Simmons, W.R. (1950). An Attempt to Get the 'Not at Homes' into the Sample without Call-backs, *Journal of the American Statistical Association*, 45, 136-137.

Qin, J., Leung, D. and Shao, J. (2002). Estimation with survey data under nonignorable nonresponse or informative sampling. *Journal of the American Statistical Association*, 97, 193-200.

Lee, S. and Valliant, R. (2009). Estimation for Volunteer Panel Web Surveys Using Propensity Score Adjustment and Calibration Adjustment, *Sociological Methods and Research*, 37, 319-343.

Lepkowski, H., Kalton, G., and Kasprzyk, D. (1989). Weighting Adjustimets for Partial Nonresponse in the 1984 SIPP Panel. *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 296-301.

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.

Macarthur, C., Dougherty, G., and Pless, I.B. (1997). Reliability and Validity of Proxy Respondent Information about Childhood Injury: An Assessment of a Canadian Surveillance System, *American Journal of Epidemiology*, 145, 834-841.

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.

Norris, P. (2001). Digital Divide: Civic Engagement, Information Poverty, and Internet Worldwide. New York: Cambridge University Press

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.

Raghunathan, T. E., Solenberger, P., and Hoewyk, J. V., and (2002) IVEware: Imputation and Variance Estimation Software User Guide, Survey Research Center, Institute for Social Research, University of Michigan, available at http://www.isr.umich.edu/src/smp/ive/.

Rice, M.F. (2003). Information and Communication Technologies and the Global Digital Divide Technology Transfer, Development, and Least Developing Countries, *Comparative Technology Transfer and Society*, 1.1, 72-87.

Rizzo, L., Kalton, G., and Brick, M. (1996). A comparison of some weighting adjustment methods for panel nonrseponse, *Survey Methodology*, 22, 43-53.

Rubin, D.B. (1976). Inference and Missing Data. *Biometrika* 63, 581-592.

Rubin, D.B. (1978). Multiple imputation in sample survey. P*roceeding in Survey Research Methodology, American Statistical Association*, 20-34.

Rubin, D.B. (1987). *Multiple imputation in sample survey.* John Wiley: New York.

Schonlau, M., Van Soest, A., Kapteyn, A., and Couper, M. (2009).

Selection Bias in Web Surveys and the Use of Propensity Scores, *Sociological Methods & Research*, 37, 291-318.

Schonlau, M., Zapert, R., Simon, L.P., Sanstad, K., Marcus, S., Adams, J., Spranca, M., Kan, H., Turner, R., and Berry, S. (2004). A Comparison Between a Propensity Weighted Web Survey and an Identical RDD Survey, Social Science Computer Review, 22, 128-38.

Sharon L. Lohr. (2010). *Sampling: design and analysis.* 2nd Ed. Brooks/Cole, Cengage Learning.

Singleton Jr., R.A., and Straits, B.C. (2005). *Approaches to social research* (4th edition). New York, NY: Oxford University Press.

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. I*nternational 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

Terhanian, G., Bremer, J., Smith, R., and Thomas, R. (2000). *Correcting Data from Online Surveys for the Effects of Nonrandom Selection and Nonrandom Assignment*. White paper, Harris Interactive, Rochester, NY.

Tourangeau, R., Conrad, F., Arens, Z., Fricker, S., Lee, S., and Smith, E. (2006). Everyday Concepts and Classification Errors: Judgments of Disability and Residence, *Journal of Official Statistics*, 22(3), 385-418.

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).

Zhang, L.C. (2002). A method of weighting adjustment for survey data subject to nonignorable nonresponse. DACSEIS research paper no. 2, http://w210.ub.unituebingen.de/dbt/volltexte/2002/451.

http://kosis.kr/statisticsList/statisticsListIndex.do?menuId=M_01_01&vwcd=MT_ZTITLE&parmTabId=M_01_01#SelectStatsBoxDiv(통계청 2017년 가계금융·복지조사).

http://www.nonresponse.org/(국제 무응답 학회 홈페이지)

https://www2.census.gov/geo/pdfs/education/Uhl_CAS_2011.pdf(Building and Maintaining the Master Address File, US Census Bureau)

https://www.census.gov/pred/www/rpts/O.3.PDF(Census 2000 Housing Unit Coverage Study, US Census Bureau).

https://www.census.gov/geo/reference/mtfcc.html.

https://www2.census.gov/cac/nac/meetings/2015-10/2015-Blumerman.pdf(2020 Census Operational Plan Overview and Operational Areas, US Census Bureau).

https://www.census.gov/programs-surveys/decennial-census/2020-census/research-testing/maps-and-address-lists.html(Updating Existing Maps and Address Lists, US Census Bureau).

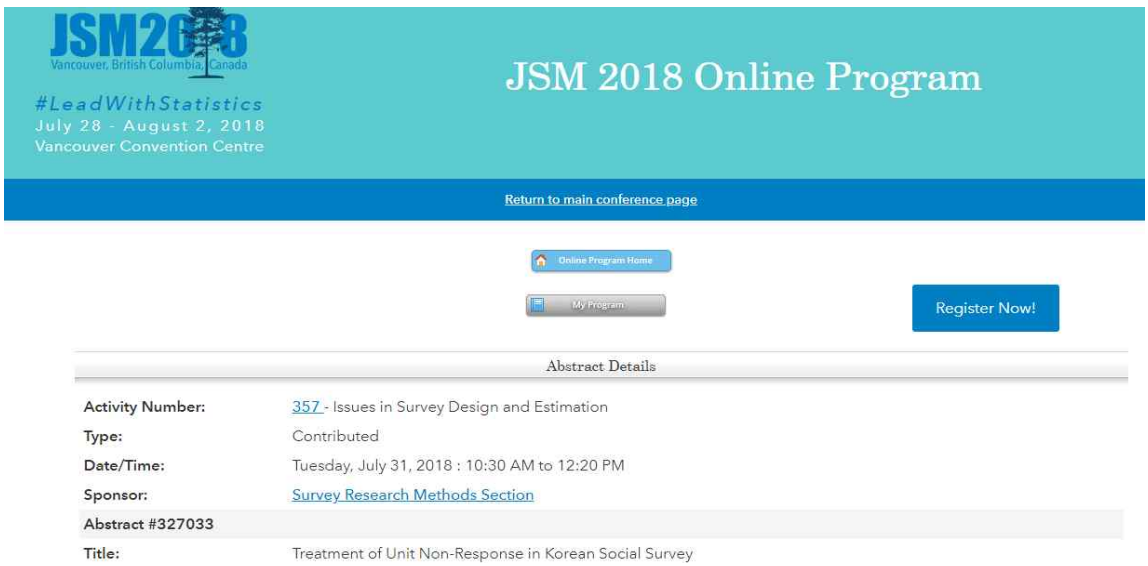
부록

1. 통계저널 논문게재(2018. 7. 2.) : 통계연구(2018), 제23권 제2호, 27-49

http://kostat.go.kr/edu/sri_kor_new/2/3/index.board

1. 생존모형을 활용한 한국의 출산력 변동 분석 : 우해봉, 장인수
2. 한국 사회조사 단위무응답 처리를 위한 최적의 보조정보구성에 관한 실증적 연구 : 정구현, 이승희
3. 이주상태(migration status)별 세분화된 데이터 생산을 위한 제언: 조사통계를 중심으로 : 박효민, 박영실, 김월화, 이영미
4. 소득집중도와 소득분포함수 : 윤종인
5. 상품공간 구조의 동태적 변화에 관한 연구 : 김대용

2. 미시간대학교 연구소 세미나 발표(2018. 2. 7.)

3. 북미 통계학회(Joint Statistical Meeting) 발표(2018. 7 .31.)

http://ww2.amstat.org/meetings/jsm/2018/onlineprogram/AbstractDetails.cfm?abstractid=327033