제6장

무응답 표본 가구의 표본교체 효과

김서영 · 안다영

제1절 서 론

1. 연구 배경

가. 무응답 가구 교체

사회 구조가 빠르게 변하면서 통계조사 환경은 과거에 비해 상당히 어려워졌다고 볼 수 있다. 특히 맞벌이가구, 노인가구, 1인가구 등의 증가는 표본가구의 조사 참여를 더욱 어렵게 만드는 요인이 되고 있다. 이러한 현상은 우리나라만의 문제는 아니고 유럽이나 북미 사회에서는 이미 1990년대 이전부터 심각한 문제로 대두되었다. 가구 대상 조사에서 무응답 가구의 증가는 통계의 정확성과 신뢰성을 위협하는 요인이 되고 있다. 이에 대해 국가통계 (official statistics)를 작성하는 기관들은 통계의 신뢰성 회복을 위해 무응답을 극복하려는 노력을 하고 있다. 통계조사에서 나타나는 무응답 현상은 조사에 협조하되 일부 조사항목에 대해서 응답을 거부하는 항목 무응답 (item nonresponse)과 조사단위 (개인 또는 가구)가 조사에 협조하지 않는 단위 무응답 (unit nonresponse)으로 구분된다. 본 연구는 단위 무응답에 대해서만 다루고자 한다.

단위 무응답은 적격 표본 (acceptable sample)으로 선택되었지만 현장조사 단계에서 응답을 하지 않아서 발생한 응답 단위의 결측 (missing)을 말한다. 이러한 무응답은 조사에서의 통계적 추론을 왜곡할 수 있다. 이 문제를 최소화하기 위한 방법들이 조사설계에서부터 현장조사 및 자료처리 등 매 단계에서 다양하게 쏟아져 나오고 있다. 대표적인 예로써, 현장에서의 표본교체 (field substitution), 더블 샘플링 (double sampling), 가중치 조정 (weighting adjustment), 사후층화 (post-stratification), 칼리브레이션 (calibration)과 같

은 방법들을 들 수 있다. 이들 각 방법은 무응답 발생 상황에 따라 다르게 적용될 수 있으며 방법에 따라서는 추가 비용 부담이 발생할 수 있다. 각 방법들의 특성을 다음과 같이 간략하게 살펴볼 수 있다.

현장에서의 표본교체 (이하 '표본교체')는 표본 크기에 영향을 주지 않기 위한 수단으로 널리 사용되었던 방법이다. 이 방법에서 표본교체는 응답대상가구가 목표 모집단에는 포함되어 있지만, 원래 표본으로 선택되지 않았던 단위로 무응답 단위를 교체하는 형태로 이루어진다. 표본교체의 경우 통계적 추론이 가능하려면 표본교체는 랜덤대체 (random substitution) 형태로 이루어져야 한다. 랜덤대체란 확률적으로 대체자를 선택하는 것을 의미한다. 대체자는 동일한 부그룹 (subgroup) 내에서 선택하거나 무응답 가구와 동일한 특성을 보장할 수 있는 그룹에서 선택하는 것이 일반적이다. 동일한 부그룹에서 선택된 대체자의 경우, 추정량의 편향은 핫덱 (hot-deck) 방법의 항목 무응답 대체(imputation)에 의해 얻어진 결과와 유사하다 (Lessler 와 Kalsbeek, 1992)고 알려져 있다.

더블 샘플링 (double sampling) 방법은 우편조사나 인터넷 조사에 널리 사용되고 있다 (Hansen 과 Hurwitz, 1946). 이 방법은 최초 추출한 표본의 일부가 응답하지 않을 경우, 응답하지 않은 표본들로부터 다시 표본을 뽑아서 처음의 방법과 다른 방법으로 재조사를 시도하는 것이다.

가중치 (weighting) 조정 방법은 응답자 그룹의 정보를 활용한 가중치에 의해 자료를 조정하는 방법이다. 이 방법은 표본을 몇 개의 그룹으로 나누고 각 그룹에 속한 모든 개체에게 응답률의 역수를 가중치로 사용한다. 이때 각 그룹에서의 가중치는 동일하게 부여되고, 이렇게 함으로써 그룹 내 무응답을 조정한다. 이 방법은 조정 그룹들 내의 속성이 내적으로 동일하면 편향이 0이 된다는 것이 증명되었다. 이것은 동일 그룹 내에서 응답자와 무응답자 간에 차이가 없다는 것을 의미한다. 가중치 조정을 하지 않은 상태에서 추정량의 편향 크기는 주어진 그룹 내에서 응답자와 무응답자로부터 계산된 추정량들 간의 차와 그 그룹 내의 응답률과의 함수 관계를 갖는다. 따라서 그룹은 응답률과 상관이 있는 요인들을 기초로 형성하는 것이 타당하고 응답자와 무응답자 간의 차이가 없도록 구성하는 것이 이상적이다.

사후층화에 의한 조정은 자료가 수집된 후에 사후적으로 조정을 위한 층을 구성하여 자료를 재조정하는 방법을 말한다. 조정에 필요한 변수는 조사 질문지에 포함된 변수로 정의되고 이 변수는 일반적으로 표본추출 시에는 이용할 수 없는 변수들이다. 이때, 각 층 내의 변수 값들의 변동은 표본 전체에서의 변수 값보다 그 변동성이 작다고 가정한다. 좋은 사후층화 변수는 관심 변수와 강한 상관관계에 있어야 하고, 사후층화 가중치를 계산하기 위해 반드시 보조정보를 알고 있어야 한다.

마지막으로 칼리브레이션 방법은 다른 벤치마킹 정보를 이용하여 조사 추정치를 조

정하는 방법을 말한다. 이 방법은 좋은 벤치마킹 변수가 있다면 편향을 줄이는데 매우 좋은 방법이지만, 사후층화와 마찬가지로 좋은 벤치마킹 변수를 찾는 것이 현실적으로 매우 어려운 작업이기도 하다.

지금까지 열거한 방법들 중에서 표본교체는 현장조사 단계에서 쉽게 적용할 수 있는 방법으로 현재 통계청의 일부 가구 단위 조사에서 사용하고 있는 방법이기도 하다. 따라서 본 연구는 다양한 무응답 조정 방법들 중 표본교체 방법에 대해서만 다룬다.

나. 표본교체

표본교체는 현장조사 단계에서 응답하지 않은 가구를 다른 가구로 교체하여 조사할 때 이루어진다. 따라서 표본교체는 무응답이 발생한 가구를 다른 가구로 복사하는 것과 유사하다. 즉, 무응답 가구를 새로운 가구로 갈아 끼우는 방법이라고 할 수 있다. 교체될 가구를 선택하는 방법은 여러 가지가 있다. 크게는 확률 추출과 비확률 추출에 의한 방법으로 나눈다. 그러나 표본교체를 통해 이루어진 조사의 통계적 추론을 위해서, 확률적 메커니즘 활용이 더 활발하게 논의되고 있는 것 같다. 조사방법론, 특히 표본추출에 관한 교과서들은 표본교체에 관해서는 아주 짧게 언급하거나 거의 언급하지 않고 있다 (Kish, 1965; Lessler 와 Kalsbek, 1991; Cochran, 1977; Groves, 1989). 일반적으로 이 문헌들은 비확률적 추출방법에 의한 표본교체자 선택을 선호하지 않는 경향이 있는 것 같다.

표본교체 방법에 관한 문헌들을 살펴보면, 1990년 이전까지는 표본교체 방법의 선호에 있어서 서로 다른 의견들이 있었다. Chapmam (1983)은 표본교체 방법에 관한 문헌고찰을 통해서 표본교체가 무응답 편향을 제거하지 못한다는 의견을 확신하기 어렵다고 하였다. 그는 4편의 경험적 연구를 통해 표본교체 방법의 장단점을 언급함으로써 대체로 표본교체 방법의 실용성을 증명할 어떠한 이론적 또는 경험적 근거는 없다고 주장하였다 (Chapman 과 Roman, 1985). 또한 이들은 집락추출에 의한 RDD (Random digital dialing)조사에서 표본교체 방법은 가중치 조정방법에 비해 분산측면에서는 효과가 있지만 편향이 더 관측되었다고 하였다. 미국 노동통계국에서도 유사한 연구 결과를 보고한 바 있다 (Biemer, Chapman 과 Alexader, 1990).

그러나 1990년을 기점으로 그 이후부터는 표본교체에 의한 무응답 조정 결과에 대한 우려의 목소리가 높아졌다고 볼 수 있다. Maliani 과 Pacei (1993)은 Italian Family Expenditure survey에서 표본교체가 갖는 심각한 문제점에 대해서 논의하였다. 그들은 실제로 표본교체 방법을 사용할 때 상당한 편향이 발생한다고 하였다. Slovenian Labour Force Survey (Vehovar, 1993)와 General Social Survey (GSS, Vehovar, 1995)에서도 표본교체 방법은 조사 추정치의 분산과 편향 면에서 정당화될 수 없다고 하였다. Verma 와 Gabilondo (1993)는 EU Labour Force Survey와 Family Budget Survey에 표본교체 방법을

적용했을 때 표본교체는 무응답률이 35%를 초과할 때 추천될 만하다고 하였다.

표본교체 방법이 국가통계에 활용된 사례는 주로, 남아프리카, 필리핀, 사우디아라비아 등의 통계청에서 살펴볼 수 있다. 그러나 이들 나라들에서 표본교체를 어떻게 사용하고 있는지에 대한 구체적 문서는 찾기 어려운 것이 사실이다 (Vehovar, 1999). 벨기에도 표본교체 방법을 사용하고 있고, 표본교체자 선택에 있어서 원 표본에서의 무응답가구와 그 속성이 유사한 가구를 선택하거나 랜덤교체 방법을 선택하고 있다 (Demarest 등, 2007). 일부 이런 국가들을 제외하면 대체로 학계를 포함하여 미국과 많은 유럽 국가들은 국가통계 조사에서 표본교체는 엄격하게 그 사용을 제한하고 있다고 볼 수 있다 (Vehovar, 1999).

종합적으로 표본교체는 학계에서는 이미 거의 추천되지 않은 방법이다. 매우 제한적이긴 하지만, 실질적 활용에 대해 표본교체 사용의 정당성에 대한 논의가 이루어졌고 지금은 대체로 자료품질에 부정적인 영향이 있다는 데 힘이 실리고 있다고 볼 수 있다. 표본교체에 대한 편향과 분산에 관한 이론적 논의는 많이 부족한 상태이긴 하나, 2000년 이후 최근 학술 저널들을 살펴보면 표본교체에 관한 논문은 찾아보기가 매우 드물다. 이처럼 이미 학계와 선진통계국가들의 국가통계 실무자들 사이에서 표본교체 방법은 연구분야로서 또는 자료품질 향상 방법으로서 추천받지 못하고 있는 것 같다 (Conference, 2010; e-mail 서신교환 등).

2. 연구 목적과 필요성

통계청의 일부 가구 단위 조사에서 무응답을 조정하기 위한 수단으로 표본교체 방법이 사용되고 있다. 이러한 표본교체는 표본의 축소 문제를 해결할 수 있다는 장점이 있다. 사회조사의 경우 표본으로 선택된 가구가 조사에 응답하지 않을 경우, 대체 표본을 통해 무응답 가구를 보완하고 있다. 지역별고용조사는 조사결과에 교체된 표본가구의 자료를 실제 결과 집계에 활용하지는 않지만, 그 활용성 연구 및 현장제어 수단으로 대체된 가구에 대해 정보를 수집하고 있다. 그렇지만 연구자들의 문헌연구에 따르면 최소한 대체결과가 조사 추정치에 어떤 영향을 미치는지에 대해서는 정확하게 밝혀진 바가 없는 것 같다. 실무자들 사이에서도 표본교체 활용에 관한 의견은 다양하지만, 구체적으로 문서화되었다기보다는 실무자들의 현장 경험에 의한 각자의 생각에 기인한 것이라 할 수 있다.

앞에서 살펴본 바와 같이 표본교체가 현실적으로 합리적인 무응답 조정 방법인지에 대해서 논리적인 증거는 없는 것 같다. 그러나 대체로 많은 연구결과들에서 표본교체 방법에 대한 부정적인 견해가 더 많은 것은 사실이다. 대부분의 통계선진국가들은 공식적

으로 표본교체를 사용하지 않는 것으로 알려져 있다. 이러한 상황에서 표본교체를 공식적으로 사용하고 있는 우리나라의 현황을 살펴볼 필요가 있다. 이는 조사 추정치의 신뢰성 향상을 위해 반드시 이루어져야 할 검증 절차에 해당된다고 본다. 표본교체 효과가 명확하게 밝혀진다면, 우리 통계청의 무응답 대응 및 자료품질 향상을 위한 전략을 구축하는데 기여가 있을 것이라고 판단된다.

3. 연구 내용과 방법

본 연구는 가구단위 무응답에 대한 표본교체 효과를 검증하는데 그 목적이 있다. 연구 자료로는 '2010년 사회조사'와 '2009년 지역별 고용조사' 결과를 이용하였다. 본 연구에서는 각각의 자료를 분석하고, 그 결과로부터 표본교체가 조사 추정치에 어떤 영향을 미치는지를 살펴보고자 한다. 연구의 목적을 달성하기 위해 자료의 이용범위에 따라 <표 6-1>과 같이 4개의 모형을 설정하고, 각 모형에 의한 추정치들을 비교한다. 이때 관심변수는 사회조사에서는 분야별 의식과 관련된 몇 개 변수, 지역별 고용조사에서는 실업률 또는 고용률과 관련된 변수들과 인구학적 특성 변수를 중심으로 고려하고자 한다.

〈표 6-1〉 4가지 형태의 연구모형

연구모형	자료이용 내용 및 특성	표본 상황
모형 A	원표본 모두 이용한 추정 (무응답 가구에 대해 무응답 가중치 적용)	원표본 중 응답 + 무응답
모형 B	원표본 중 응답가구만 이용한 추정	원표본 중 응답
모형 C	표본교체 자료를 이용한 추정	원표본 중 응답 + 교체표본 중 응답
모형 D	교체된 자료 중 응답가구에 대해서만 추정	교체된 표본 중 응답

추가적으로 표본교체를 적용한 후의 최종 응답 가구만을 이용하여 2005 인구주택총조사 (센서스) 자료와 연결을 시도한다. 이를 통해 현재 시점에서 표본 추출틀과의 일치성 정도를 확인할 수 있을 것으로 기대한다. 이때 연계을 위한 key변수는 조사구 내 가구번호와 가구주 생년월일을 사용한다. 이렇게 할 경우 가구의 전출입에 따른 변동 없이 단지 가구주만의 변동에 따른 가구 변동사항은 확인하기 어렵다는 단점은 있다. 실제로 무응답 연구를 할 때 표본추출 프레임으로 사용되는 센서스 자료와의 연계는 조사 시점에서 무응답 가구의 특성을 분석하는데 중요한 의미가 있다. 그러나 우리나라와 같이 인

구이동이 빈번할 경우 센서스 자료와의 연계를 통해 무응답가구의 특성을 얼마나 파악할 수 있을지는 의문이다. 더구나 현재 표본조사의 경우 무응답 가구에 대한 어떠한 정보도 파악이 되지 않기 때문에 무응답 가구에 대한 연계는 불가능한 상태이다. 대안으로 최종 응답한 가구의 가구주 이름과 센서스의 가구주 이름과 연계함으로써 연계 비율이 어느 정도인지를 파악할 수 있다.

이러한 연계 결과는 향후 확률적 메커니즘에 의해 표본 대체자가 선택된다 하더라도 그 의미가 얼마나 희석될 수 있는지를 보여줄 수 있을 것이다. 또한 이 결과를 통해 향후 무응답 연구에 필요한 자료 수집의 중요성을 보일 수 있는 근거가 될 수 있을 것이다.

4. 기대효과 및 제약

두 가구 단위 조사 결과를 분석함으로써 가구 단위 무응답에 대한 표본대체 효과를 검증할 수 있을 것으로 기대한다. 또한 비확률적 표본대체가 통계 추정치에 어떤 영향을 미치는지 파악함으로써 통계청에서 향후 표본대체의 활용성 여부를 판단하는 중요한 근거가 될 수 있을 것으로 본다.

그러나 본 연구는 크게 두 가지 관점에서 제약이 있고, 이 점을 고려하여 연구의 결과를 이해할 필요가 있다. 하나는 이미 언급한 바와 같이 표본교체에 관련한 연구 문헌이 거의 없고, 있다고 하더라도 표본교체 사용여부에 대한 정확한 결론은 없다는 점이다. 다만 국제사회에서는 표본교체의 사용을 공식으로 허용하지 않는 분위기라는 것이다. 다른 하나는 무응답 가구의 특성을 파악할 수 있는 자료가 미약하다는 점을 들 수 있다. 이는 다른 나라의 상황도 비슷하긴 하지만, 우리나라의 경우, 지금 현재로써 무응답 가구에 대한 추적조사나 파라데이터 (paradata)에 관한 수집이 이루어지고 있지 않기 때문에 더 어려운 상황이라 할 수 있다. 이런 점에서 볼 때, 본 연구에 의한 표본교체 자료 분석결과도 직접적이기보다는 간접적인 결론을 유도한다고 볼 수 있을 것이다.

그렇지만, 본 연구는 통계조사에서 적절한 무응답 조정 방법을 찾기 위한 연구의 시작이라는 점에서 의미가 있다. 그리고 표본교체 방법이 하나의 무응답 방법으로 사용될 수 있는지의 여부도 판단해 볼 수 있을 것이다. 나라마다 통계작성 환경이 다르기 때문에 표본교체가 반드시 좋은 방법이 아니라고 보기는 어렵다. 표본교체 방법이 지닌 장단점을 파악함으로써 그 사용 가능성을 예측해 볼 수 있는 기회가 될 것이다. 뿐만 아니라 객관적이고 과학적인 방법으로 자료 품질을 향상시킬 수 있는 방법을 찾는데 기여할 수 있다는 점에서 본 연구의 의미가 있다고 볼 수 있다.

제2절 표본교체 현황 분석

1. 자료

본 절에서는 사회조사와 지역별고용조사에 대해 설명하고, 표본교체 효과를 측정하기 위해 자료에 대한 기초분석 결과를 설명하고자 한다.

가. 사회조사

사회조사는 국민 삶의 질에 관한 사항과 사회적 관심 사항 등 사회 구성원의 주관적 의식과 사회적 관심사를 조사하는 연간조사로, 삶의 수준과 국민생활의 모습 및 의식구조의 변화를 파악하여 사회개발정책의 기초 자료로 제공하는데 목적이 있다. 사회조사는 1979년 '한국의 사회지표' 체계구성을 목적으로 시행되었고, 조사부문은 총 10개 부문으로 가족, 노동, 보건, 환경, 교육, 소득과 소비, 복지, 문화와 여가, 안전, 사회참여로 구성되었다. 조사대상은 전국 만 15세 이상의 모든 가구원을 모집단으로 한다. 표본추출틀은 2005 인구주택총조사 표본조사구 중 아파트/보통 조사구만을 대상으로 하나, 해당 조사구 중 기숙사나 양로원 등 집단가구 및 현재 경상표본, 외부 승인통계, 지역통계 표본조사구는 제외한다. 이 표본추출은 전국 지역의 25개 (16개 시도, 동읍면) 지역별 층화 및 표본추출 변수에 따라 정렬한 후 층별로 가구수를 기준으로 확률비례계통추출방법을 이용하여 표본조사구를 추출한다. 이 표본조사구의 가구에 일련번호를 부여한 후 무작위로 최초 가구를 설정, 그 가구를 포함하여 연속 12가구를 조사하는 방식으로 약 17,000가구를 대상으로 한다. 그러나 무응답률을 낮추기 위하여 예비표본을 선정하게 되는데 이는 원래 표본조사구 (12가구)에서 무응답 가구 발생 시 가구를 대체하기 위한 표본으로 10가구의 예비표본을 선정한다. 실제 조사환경에서는 10개의 예비가구도 부족한 경우가 많은 실정이다.

<표 6-2>는 2010 사회조사에서 예비표본 가구의 사용실태를 나타낸 것이다. 표에서 가구번호가 1부터 12번까지는 표본추출 당시에 선택된 가구들이다. 원래 표본 12개 가구에서 무응답 가구 또는 부적격 가구가 발생할 경우 표본 대체가 이루어지고, 대체가구는 13번 이후 가구를 의미한다. 대체로 대체가구의 시작인 13번부터 15번 가구들은 원래 표본 가구들과 거의 유사한 규모로 방문되고 있음을 알 수 있다. 전체적으로 가구번호가 13번 이상인 대체 가구들은 전체 방문가구의 약 32%정도이며, 가구순번이 20번대 이상인 경우 전체 방문가구의 약 4%정도 된다.

〈표 6-2〉 사회조사의 예비표본 가구의 교체 실태

가구번호	가구수	백분율	가구번호	가구수	백분율
1	1,426	5.66	24	99	0.39
2	1,425	5.65	25	75	0.3
3	1,426	5.66	26	56	0.22
4	1,425	5.65	27	48	0.19
5	1,427	5.66	28	30	0.12
6	1,424	5.65	29	22	0.09
7	1,424	5.65	30	18	0.07
8	1,428	5.66	31	13	0.05
9	1,425	5.65	32	10	0.04
10	1,426	5.66	33	6	0.02
11	1,423	5.64	34	6	0.02
12	1,424	5.65	35	6	0.02
13	1,365	5.41	36	6	0.02
14	1,256	4.98	37	4	0.02
15	1,096	4.35	38	2	0.01
16	947	3.76	39	2	0.01
17	782	3.1	40	2	0.01
18	641	2.54	41	2	0.01
19	513	2.03	42	1	0
20	419	1.66	43	1	0
21	308	1.22	44	1	0
22	226	0.9	99	2	0.01
23	147	0.58	합계	25,217	100

나. 지역별고용조사

지역별고용조사는 지역 고용정책 수립에 필요한 시도, 시군의 고용구조 분석자료 및 산업·직업에 대한 세분화된 자료를 제공하기 위한 목적으로 작성되는 통계이다. 지역별고용조사 (당시 명칭은 시군구고용통계조사)는 2008년 10월 대면조사로 시작되었다. 지역별고용조사는 우리나라에 상주하는 만 15세 이상 인구를 대상으로 한다. 2005 인구주택총조사 조사구 중 조사구 내 집단가구 (기숙사, 보육원 및 양로원 등 시설가구)는 제외하고, 전국 8,800조사구의 약 175,000가구를 확률 추출하여 조사를 실시한다. 표본조사구는 여러 표본 군 중 모집단의 특성지표[1]와 가장 유사한 표본을 최종 표본 조사구로 선정한다. 그리고 표본 조사구의 가구에 일련번호를 부여한 후 무작위로 최초 가구를 설정한다. 그 초기 가구를 포함하여 연속 20가구를 표본 가구로 확정하고 조사원이 이 조사 대상가구를 직접 방문한다. 지역별고용조사의 예비표본은 5가구로 원표본의 20가구

1) 모집단 구조와 가장 흡사한 표본을 추출하기 위한 분석

에서 무응답이 발생하면 사회조사와 마찬가지로 연속된 다음가구부터 5가구를 지정하여 사용한다. 대부분 가구 단위 조사는 조사원이 직접 가구관리종합표를 작성하여 조사 진행 상황을 확인한다. 가구관리종합표의 일련번호 순으로 20가구가 표본 추출 당시의 조사대상가구이며, 21번부터 5가구는 예비가구이다. 현재 3회 이상 방문하여 면접했으나 도저히 조사가 불가한 경우 (부재, 거부, 기타 등에 의한 무응답, 불능) 대체할 수 있도록 하고 있다.

〈표 6-3〉 지역별고용조사의 예비표본 가구 교체 실태

가구번호	가구수	백분율	가구번호	가구수	백분율
1	5,990	4.17	32	478	0.33
2	5,644	3.93	33	342	0.24
3	5,545	3.86	34	283	0.20
4	5,562	3.88	35	209	0.15
5	5,592	3.90	36	177	0.12
6	5,588	3.89	37	132	0.09
7	5,574	3.88	38	114	0.08
8	5,543	3.86	39	99	0.07
9	5,527	3.85	40	63	0.04
10	5,519	3.85	41	62	0.04
11	5,498	3.83	42	50	0.03
12	5,551	3.87	43	51	0.04
13	5,556	3.87	44	37	0.03
14	5,486	3.82	45	35	0.02
15	5,492	3.83	46	27	0.02
16	5,511	3.84	47	28	0.02
17	5,532	3.86	48	29	0.02
18	5,552	3.87	49	25	0.02
19	5,453	3.80	50	20	0.01
20	5,464	3.81	51	22	0.02
21	5,030	3.51	52	14	0.01
22	4,737	3.30	53	16	0.01
23	4,269	2.98	54	13	0.01
24	3,796	2.65	55	11	0.01
25	3,292	2.29	56	13	0.01
26	2,521	1.76	57	6	0.00
27	1,978	1.38	58	11	0.01
28	1,539	1.07	59	9	0.01
29	1,169	0.81	60	8	0.01
30	892	0.62	61	4	0.02
31	624	0.43	합계	143,476	100

<표 6-3>은 2009년 지역별고용조사에서 예비표본 가구의 사용실태를 나타낸 표이다. 사회조사와 마찬가지로 최초 표본추출 당시 선택된 표본들 중 응답을 얻지 못할 경우 순번대로 표본교체가 이루어진다. 지역별고용조사는 가구번호 1번부터 20번까지가 최초 표본이다. 즉 21번 이후는 표본교체 가구라고 보면 된다. <표 6-3>을 보면 최초 표본의 경우 전체 방문가구의 약 77%이고 가구번호가 21번 이상인 교체가구들은 전체 방문가구의 약 22%를 차지한다. 특이한 점은 사회조사와 마찬가지로 최초표본가구 이후의 약 3~5가구는 최초표본 가구들과 거의 유사한 규모로 방문되고 있다는 점이다.

2. 표본교체 현황

본 절에서는 두 조사에서 실질적으로 이루어지고 있는 표본교체 현황을 분석한다. 무응답 가구에 대해 이를 대신할 새로운 가구로 교체하기 전과 교체한 후에 가구의 기본적 현황들의 변화를 살피고자 한다.

가. 사회조사의 경우

<표 6-4>는 사회조사에서의 표본교체에 따른 무응답 가구의 교체 현황을 나타낸 것이다. 2010 사회조사의 대상 표본가구는 17,103가구로 이 대상가구를 분석 자료로 사용한다. 이 17,103가구 중 12,091가구가 응답하였고 나머지 5,012가구는 응답하지 않았다(무응답과 불능 포함). 이 무응답 가구에 대해서는 가구 교체가 이루어졌고, 5,012가구의 교체를 목적으로 8,114가구가 현장에서 추가로 방문되었다. 이 추가 방문 가구 중 약 5,021가구가 응답하였고, 나머지 3,093가구는 여전히 응답하지 않았다. 추가로 조사된 교체대상 가구의 응답률은 약 61.2%로 높지 않은 것을 알 수 있다. 총체적으로 보면 원표본 가구의 무응답에 따라 새롭게 교체된 가구를 포함하여 조사기간 동안 방문한 전체 가구수는 25,217가구였고, 이 중에서 응답한 가구수는 17,112가구인 것으로 나타났다(<표 6-4> 참고).

〈표 6-4〉 사회조사의 표본 현황

	원표본	교체표본	합계
방문한 전체 가구수	17,103	8,114	25,217
· 응답한 가구수 (응답 가구원 수)	12,091 (26,200명)	5,021 (10,646명)	17,112 (36,846명)
· 무응답 가구수	5,012	3,093	8,105

무응답률 계산에 있어서 응답을 얻을 수 없는(빈집, 장애, 언어문제 등의 이유로 조사에 응할 수 없는) 가구, 즉 불능 가구는 일반적으로 응답률 계산에서 제외한다(APPOR, 2010). 본 연구의 경우, 사회조사 자료에서 무응답과 불능 가구를 구분하기 어렵기 때문에 응답률을 계산할 때 부적격 가구도 포함하였다. <표 6-5>에 따르면 사회조사의 전체 응답률은 67.9%, 가구 교체율은 29.3%인 것으로 나타났다. 즉, 사회조사는 가구를 대상의 가구원들의 의식을 조사하는 것으로써 응답해야 할 항목이 많고 내용이 광범위하기 때문에 응답을 얻기가 그만큼 어렵다고 이해할 수 있는 부분이다.

〈표 6-5〉 2010 사회조사의 응답률

총방문가구수 ①	부재불응불능 ②	응답가구수 ③	교체율[2] (②/③)*100	응답률 ③/①*100
25,217	5,012	17,112	29.3	67.9

<표 6-6>은 16개 시도별 표본교체 현황을 나타낸다. 실제 사회조사 결과가 지역별로 공표되기 때문에 지역별 현황을 살펴볼 필요가 있다. <표 6-6>에서 왼쪽 열의 응답 가구원수는 원표본의 응답가구원수와 교체표본의 응답가구원수를 나타내고, 합계는 전체 응답자수를 나타낸다. 표의 오른쪽 열의 응답가구는 마찬가지로 원표본가구의 응답가구수와 교체된 표본가구의 응답가구수를 나타내며 합계는 전체 응답가구수를 나타낸 것이다. 표에서 원표본(율)에 율은 '원표본 응답 수/합계' 에 의해 계산된 것이다.

<표 6-6>과 [그림 6-1]에 따르면 지역별로 볼 때 서울이 원표본 응답률이 낮고 교체비율이 가장 높다. 그 다음으로 인천, 광주, 경기, 울산 순으로 대도시권에서 거부, 비접촉으로 인한 불응이 많다고 볼 수 있다. 또한 재미있는 사실은 전체적으로 가구 교체 비율에 비해 가구원 교체 비율이 약간 낮은 것을 알 수 있다 (가구 원표본(율)이 낮음). 게다가 충북, 전북, 경남, 제주는 다른 지역들에 비해 가구원 교체 비율이 더 낮은 것으로 나타났다. 이러한 사실은 이들 지역들에서 교체된 가구의 가구원 수가 다른 지역들에 비해 적은 가구들이 추출되었다는 것을 의미할 수 있다. 이는 사회조사와 같이 가구 내 가구원을 대상으로 하는 조사에서 표본교체는 표본의 기본 구성분포가 흐트러질 수 있음을 암시할 수 있을 것이다.

2) 교체율(교체된 가구비율) = 교체표본 중 응답가구수 / {(원표본 응답가구수+교체표본 응답가구수)}
단, 원표본 무응답수≥교체표본 응답수

〈표 6-6〉 사회조사의 지역별 표본교체 현황

	응답 가구원수				응답가구수		
	원표본(율)	교체	합계		원표본(율)	교체	합계
서울	2,883(61.8)	1,784	4,667	서울	1,259(61.0)	805	2,064
부산	1,839(74.8)	619	2,458	부산	814(73.7)	290	1,104
대구	1,681(74.9)	563	2,244	대구	770(75.5)	250	1,020
인천	1,545(65.7)	808	2,353	인천	657(64.4)	363	1,020
광주	1,364(68.3)	632	1,996	광주	616(67.5)	296	912
대전	1,455(73.3)	531	1986	대전	662(72.6)	250	912
울산	1,109(69.2)	494	1,603	울산	498(69.2)	222	720
경기	3,067(69.0)	1,380	4,447	경기	1,342(68.2)	626	1,968
강원	1,457(76.8)	440	1,897	강원	731(76.2)	229	960
충북	**1,404(71.8)**	551	1,955	충북	**679(70.7)**	281	960
충남	1,527(77.0)	455	1,982	충남	731(76.2)	229	960
전북	**1,502(75.6)**	485	1,987	전북	**742(77.3)**	218	960
전남	1,490(75.2)	492	1,982	전남	715(74.5)	245	960
경북	1,442(72.0)	560	2,002	경북	724(71.8)	284	1,008
경남	**1,558(74.3)**	539	2,097	경남	**735(72.9)**	273	1,008
제주	**877(73.7)**	313	1,190	제주	**416(72.2)**	160	576
총합	26,200	10,646	36,846	총합	12,091	5,021	17,112

[그림 6-1] 16개 시도별 원표본과 교체표본의 응답률

다음으로 원표본과 교체표본 각각의 응답가구만을 대상으로 기본적인 가구 특성을 살

펴보자. 이를 통해 표본교체로 인한 교체가구의 기본적 특성과 원표본 가구의 특성 간의 차이를 파악할 수 있다. <표 6-7>에서 보면, 교체가구 중 응답가구와 원표본 가구 중 응답가구 간에 있어서 거처종류, 주택점유형태 및 학력은 그 구성비에 있어서 차이는 있는 것으로 나타났다. 특히, 점유형태가 '자기집'이라 응답한 경우는 원표본 가구 중 64.3%, 교체표본 가구 중 59.6%로 두 응답 결과에는 약 4%p 차이가 있는 것으로 나타났다. 즉, 교체표본의 응답가구는 원표본의 응답가구에 비해 자기집 비율이 낮고 전세 비율이 높다고 볼 수 있다. 학력 부분에 있어서는 교체표본의 응답가구는 원표본 응답가구에 비해 가구주의 학력이 약간 높게 나타났다. 전체적으로 원표본과 교체표본의 응답가구 특성에 있어서 교체표본은 아파트 비율이 낮고 전세비율이 높게 나타났다. 학력에 있어서는 큰 차이는 없지만 교체표본인 경우가 전문대졸 이상의 고학력자들이 약간 많은 것으로 나타났다. 그 밖에 가구주와의 관계와 혼인상태에 대해서는 큰 차이를 보이지 않았다.

〈표 6-7〉 사회조사 대상가구의 인구학적 특성에 따른 원표본과 교체표본별 응답 분포(%)

		모형 B (원표본 중 응답가구)	모형 D (교체표본 중 응답가구)
거처종류	단독주택	38.7	39.1
	아파트	46.7	44.4
	연립주택/기타	14.6	16.5
점유형태	자기집	64.3	59.6
	전세	17.6	20.1
	월세 및 기타	18.1	20.3
성별	남자	47.73	47.49
	여자	52.27	52.51
가구주와 관계	가구주	46.2	47.2
	배우자	29.8	29.5
	자녀	18.7	18.6
	부모(배우자쪽포함)	3.5	2.8
	기타	1.8	1.9
학력	무학	5.3	5.3
	초등학교	12.1	11.2
	중학교	10.7	9.9
	고등학교	35.7	35.6
	전문대졸	13.9	14.7
	대학 이상	22.3	23.3
혼인상태	미혼	23.6	23.9
	배우자있음	64.2	64.1
	기타(사별+이혼 등)	12.2	12.0

마지막으로, 2005년 인구주택총조사와 2010년 사회조사의 응답 대상자들에 대해서 표본의 일치율을 살펴보았다. 두 자료의 연계 키 (key)변수는 가구고유번호 (17자리)와 생년월일 (4자리)을 사용하여 가구 연계와 가구원 연계를 시도하였다.

우선 가구고유번호를 연계한 결과, 조사에 응답한 17,112가구 중 16,624가구 (약 97.1%)가 완전히 연계되었다. 즉, 2005년 인구주택총조사에 대해서 2010년 현재 조사에 응답한 가구들의 가구 일치율은 약 97.1%로 매우 높은 것으로 나타났다. 연계되진 않은 488가구는 재건축 또는 택지 변경으로 인한 변동에 기인한 것으로 짐작된다. 특히 연계가 되지 않은 가구들의 지역별 분포는 서울 (9.4%)과 경기 (12.9%)가 16개 시도 지역 중 높은 비중을 차지하였다. 가구고유번호에 의한 연계로 가구의 이동 및 변경 현황은 알 수 있지만, 2005년 인구주택총조사 당시와 현재의 가구의 속성이 동일한지의 여부는 판단하기 어렵다.

〈표 6-8〉 2005 인구주택총조사와 2010 사회조사의 표본일치 결과

key변수	연계 결과	
	가구수	%(전체 17,112가구 대비)
가구고유번호	16,624	97.1
가구고유번호+생년월일	4,613	27.0

따라서 가구고유번호와 생년월일을 연계 키 변수로 사용하여 두 자료를 연계하였다. 그 결과, 전체 17,112 응답가구 중 4,613가구만이 완전히 연계되었다. 이는 2005 인구주택총조사 시점에서부터 2010년 현재까지 5년 동안 가구 속성이 변하지 않고 남아 있는 가구가 전체 응답가구의 약 27% (4,613/17,112)에 해당됨을 의미한다. 이러한 연계 비율은 상당히 낮은 수치라 할 수 있으며 인구주택총조사 기준의 표본구조가 현재 시점에서는 표본추출 당시에 비해 다소 흐트러질 수 있다고 볼 수 있다. 이러한 현상은 5년 주기로 인구주택총조사를 실시하고 그것을 기반으로 표본을 추출하는 경우에도 어쩔 수 없는 현상이라 할 수 있다.

또한 가구의 속성이 동일하더라도 경우에 따라서는 가구주의 변동도 상당히 발생하고 있었다. 실제로 5년간 동일한 가구에 거주하고 있는 가구의 경우, 2005년에 가구원이었으나 2010년 현재는 가구주로 변동된 가구는 약 12% (556/4,613)나 되었다. 이는 가구주의 사망 또는 이혼・별거 등의 이유로 가구주의 변동이 있었을 것으로 이해할 수 있다. 이처럼 시간이 흐름에 따라 자연스럽게 가구주의 변동이 발생할 수 있지만, 가구주의 직업・산업・교육 등의 정보를 표본추출 변수로 사용하는 경우에는 이러한 변동성이 표본 구조에 영향을 줄 수 있을 것으로 생각된다.

지금까지 사회조사의 표본교체 현황을 살펴보았다. 종합하면 2010년도 사회조사 대상가구는 2005년 인구주택총조사 시점에 비추어 인구학적 특성 분포가 약간 달라져 있음을 알 수 있다. 또한 표본교체 비율은 약 29.3%로 높은 편이며, 특히 서울과 경기와 같은 대도시에서 교체비율이 더 높았다. 한편 충북, 전북, 경남과 같은 소도시 권에서는 가구 교체에 대해 가구원 교체율이 다른 지역들에 비해 낮아서, 이들 지역들에서 원표본의 가구의 가구원수보다 적은 가구원수로 구성된 가구들로 교체되고 있다고 볼 수 있다.

나. 지역별고용조사

<표 6-9>는 2009년 지역별고용조사의 표본가구 교체 현황을 나타낸다. <표 6-9>에서 보면 알 수 있듯이, 전체 약 175,000가구 중 경제활동인구조사의 표본을 제외한 143,352가구를 분석대상으로 사용하였다. 이 143,352가구를 원표본의 가구수라고 하자. 지역별고용조사 분석을 위해 사용된 자료는 원자료와 가구관리종합표 자료를 혼합하여 사용하였다. 이렇게 함으로써 가구관리종합표에 기록된 정보를 이용하여 지역별고용조사 조사표 결과를 다양한 측면에서 분석할 수 있다는 이점이 있다. 예를 들면 조사표에는 주택유형 항목은 없지만 지역별고용조사의 가구관리종합표에는 그 항목이 있다.

<표 6-9>와 <표 6-10>에서 보면, 교체표본의 응답률은 59.9% (32,013/53,404)로 원표본의 응답률 76.8% (110,137/143,352)보다 훨씬 낮게 나타났다. 놀라운 사실은 지역별고용조사의 경우, 표본교체를 추정에 반영하더라도 전체 응답률은 72.2%로, 응답률이 그다지 높지 않았다. 한편 원표본 가구에서 발생한 무응답 33,215가구를 새로운 가구로 교체하여 조사하기 위해 53,404가구를 추가로 조사하였다. 그 중 32,013가구가 조사에 응답한 것으로 나타났다. 교체에 의한 응답 가구를 모두 조사추정에 활용할 경우 교체된 가구의 비율은 22.5%이다. 또한 교체 성공률은 96.4% (32,013/33,215)로 매우 높다.

〈표 6-9〉 2009 지역별고용조사의 표본가구 교체 현황

	원표본(%)	교체표본(%)	합계(%)
방문한 전체 가구수	143,352(100)	53,404(100)	196,756(100)
응답한 가구수	110,137(76.8)	32,013(59.9)	142,150(72.2)
무응답 가구수	33,215(23.2)	21,391(40.1)	54,606(27.8)

〈표 6-10〉 지역별고용조사의 교체율과 응답률

방문가구수 A	부재불응불능가구수 B	조사가구수 C	교체율[3]	응답률 C/A
196,756	54,606	142,150	22.5	72.2

다음으로 지역별고용조사의 지역별 가구 교체 현황을 살펴보았다 (<표 6-11>). 사회조사와 같이 서울의 교체 비율이 가장 높고, 인천, 제주, 경기 순으로 나타났다. 반대로 대전과 울산, 경남은 그 비율이 다른 지역에 비해 상대적으로 낮은 것으로 나타났다.

〈표 6-11〉 지역별고용조사의 16개 시도별 표본교체 현황

	표본전체			응답가구만		
	원표본	교체	합계	원표본	교체	합계
서울	8,899	4,671	13,570	6,432	2,350	8,782
	65.58	34.42		73.24	26.76	
부산	5,719	1,688	7,407	4,556	1,118	5,674
	77.21	22.79		80.3	19.7	
대구	2,459	,856	3,315	1,882	531	2,413
	74.18	25.82		77.99	22.01	
인천	3,100	1,333	4,433	2,390	689	3,079
	69.93	30.07		77.62	22.38	
광주	1,118	462	1,580	837	267	1,104
	70.76	29.24		75.82	24.18	
대전	1,056	343	1,399	837	223	1,060
	75.48	24.52		78.96	21.04	
울산	1,200	309	1,509	1,037	151	1,188
	79.52	20.48		87.29	12.71	
경기	26,045	9,693	35,738	19,975	5,787	25,762
	72.88	27.12		77.54	22.46	
강원	12,715	5,520	18,235	9,588	3,095	12,683
	69.73	30.27		75.6	24.4	
충북	8,349	2,933	11,282	6,302	1,855	8,157
	74	26		77.26	22.74	
충남	12,254	4,449	16,703	9,369	2,783	12,152
	73.36	26.64		77.1	22.9	
전북	10,259	3,801	14,060	7,785	2,380	10,165
	72.97	27.03		76.59	23.41	
전남	15,833	5,680	21,513	12,072	3,586	15,658
	73.6	26.4		77.1	22.9	
경북	17,291	6,213	23,504	13,421	3,836	17,257
	73.57	26.43		77.77	22.23	
경남	15,855	4,988	20,843	12,735	3,081	15,816
	76.07	23.93		80.52	19.48	
제주	1,200	465	1,665	919	281	1,200
	72.07	27.93		76.58	23.42	
총합	143,352	53,404	196,756	110,137	32,013	142,150

3) 교체율 = 교체표본 중 응답가구수 / {(원표본 가구수 - 원표본 무응답가구수 +교체표본 응답가구수)}
단, 원표본 무응답수≥교체표본 응답수

<표 6-12>는 원표본과 교체표본의 응답가구만을 대상으로 기본적인 특성을 정리한 것이다. <표 6-12>를 보면 거처종류와 학력 부문에서 두 모형 간 차이가 있는 것으로 나타났다. 우선 원표본 응답가구의 단독주택비율이 약 53.3%인데 반해 모형 D의 주택비율은 약 59.2%로 6%p 차이를 보인다. 아파트비율도 약 6%p 차이를 보인다 (모형 B가 큼). 즉, 교체표본은 원표본에 비해 주택구성비가 높고 아파트구성비가 낮다. 학력을 보면 중학교 이하의 구성비는 모형 D가 높고, 고등학교 이상의 구성비는 원표본이 높게 나타나 전체적으로 고학력 비중은 원표본이 약간 높은 것으로 확인할 수 있다. 그 외 가구주와의 관계나 혼인상태는 거의 차이가 없는 것으로 나타났다.

〈표 6-12〉 지역별고용조사 대상가구의 인구학적 특성에 따른 원표본과 교체표본별 응답 분포(%)

		모형 B (원표본 중 응답가구)	모형 D (교체표본 중 응답가구)
거처종류	단독주택	53.29	59.24
	아파트	37.79	31.39
	연립주택/기타	8.91	9.38
가구주와 관계	가구주	47.58	48.28
	배우자	30.34	29.97
	자녀	16.07	15.47
	부모(배우자쪽포함)	4.7	4.9
	기타	1.31	1.39
학력	무학	9.86	11.2
	초등학교	17.48	18.84
	중학교	12.35	12.87
	고등학교	33.3	31.59
	전문대졸	8.17	7.4
	대학이상	18.85	18.11
혼인상태	미혼	19.89	19.57
	배우자있음	65.36	64.81
	기타(사별+이혼 등)	14.75	15.62

제3절 실제 자료를 이용한 표본교체 효과 분석

본 절에서는 사회조사와 지역별고용조사의 추정치를 통해 표본교체 효과를 살펴보았

다. 표본교체 효과는 앞에서 설명한 교체표본 포함 여부를 고려하여 4가지 모형을 가정하고, 각각으로부터 추정치를 비교하였다. 또한 4가지 각 모형에 대한 추정치의 분산을 통해 각 추정치의 정도를 비교하였다.

1. 사회조사의 경우

우선 사회조사에서 4가지 모형별로 분석에 사용된 가구수와 가구원수 현황을 살펴보면 <표 6-13>과 같다. 원표본 가구 중 5,012가구가 조사에 응답하지 않았고, 이 무응답 가구는 새로운 5,021가구로 교체되었다. 5,021가구로부터 교체된 가구원 수는 10,646명으로 원표본에서 무응답한 가구원 수와 비슷하거나 약간 적은 규모로 교체되었을 것으로 판단된다 (<표 6-4>의 사회조사의 표본 현황 참고).

〈표 6-13〉 4가지 모형별 가구수와 가구원수

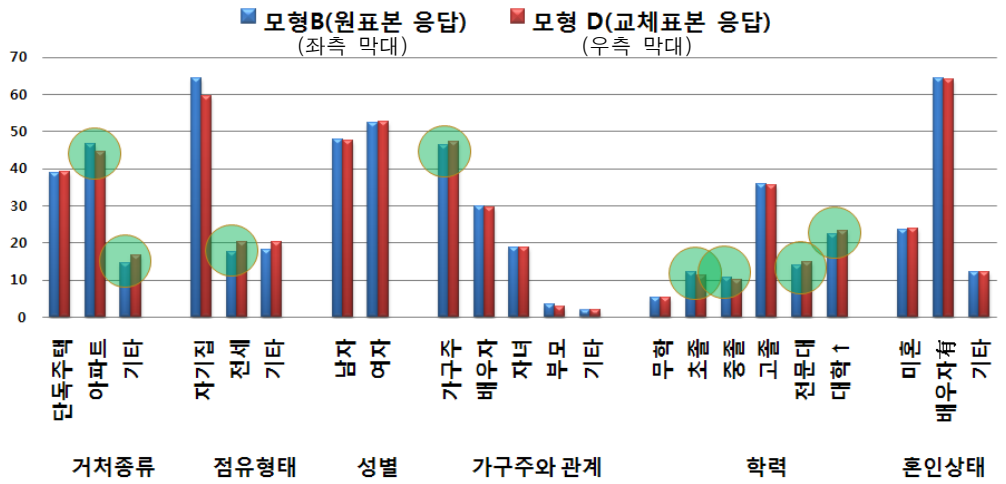
모형	구성	가구수	가구원수
A	원표본 (응답+무응답)	17,103	-
B	원표본 중 응답	12,091	26,200
C	교체 후 전체	17,112	36,846
D	교체표본 중 응답	5,021	10,646

위의 4개 모형을 자료에 적용한 후, 각각에 대해 조사 추정치를 구해 보았다. 모형 A는 원표본 내에 무응답 가구가 포함된 상태로, 16개 시도를 무응답층으로 사용하여 무응답 가중치를 적용하였다. 실제로 사회조사의 경우 무응답 가구의 특성을 파악할 수 있는 자료가 거의 없기 때문에 무응답 층을 형성하기가 쉽지 않다. 우선 전국 단위에서 위의 4가지 모형에 대해 추정치를 구하고 서로를 비교하였다. 여기서 추정치를 계산할 때, 자료에 대한 기본 가중치는 적용하지 않았다. 왜냐하면, 사회조사는 사후가중치가 존재하고, 확률비례계통추출방법에 의해 표본을 추출함으로써 자체가중 (self-selection)이 만족된다고 가정하였다.

<표 6-14>는 사회조사의 항목 중 응답자의 일반특성 항목 3개와 의식관련 항목 4개에 대해 위에서 정의한 4개 모형별로 추정치와 표준오차를 계산한 것이다.

〈표 6-14〉 사회조사에서 일반적 특성 변수에 대한 모형별 추정치와 표준오차

모형	추정치 (%)				표준오차			
	A	B	C	D	A	B	C	D
종사상지위								
1.임금근로	65.150	64.840	65.548	67.328	0.397	0.397	0.334	0.619
2.사업주	6.501	6.436	6.368	6.197	0.207	0.204	0.172	0.318
3.자영자	20.932	21.161	20.885	20.192	0.338	0.340	0.286	0.530
4.무급가족	7.417	7.564	7.199	6.284	0.216	0.220	0.182	0.320
점유형태								
1.자기집	63.976	64.321	62.932	59.590	0.439	0.436	0.369	0.693
2.전세	17.928	17.625	18.356	20.116	0.352	0.347	0.296	0.566
3.보증부월세	7.028	7.055	7.334	8.006	0.233	0.233	0.199	0.383
4.월세	7.792	7.659	7.971	8.723	0.246	0.242	0.207	0.398
5.무상	3.276	3.341	3.407	3.565	0.161	0.163	0.139	0.262
거처종류								
1.단독	38.162	38.698	38.809	39.076	0.442	0.443	0.373	0.689
2.아파트	46.729	46.671	45.997	44.374	0.455	0.454	0.381	0.701
3.연립	5.854	5.732	5.505	4.959	0.216	0.211	0.174	0.306
4.다세대	7.939	7.601	8.246	9.799	0.252	0.241	0.210	0.420
5.기타	1.316	1.299	1.443	1.793	0.105	0.103	0.091	0.187



[그림 6-2] 사회조사에서 원표본과 교체표본 응답자간 일반적 특성 분포

<표 6-14>에서 각 모형별로 짝지어서 추정치가 차이가 어떻게 발생하는지 설명해 보기로 한다. 우선 '종사상 지위' 변수에 대해 설명해 보자. 1) 표본교체를 하지 않고 원표본 중 응답자만 대상으로 하였을 때 임금근로자 비율은 64.84%이고, 표본교체 후에는 임금근로자 비율이 65.55%로, 두 추정치 간에 약간 차이가 있는 것으로 나타났다. 이는 무응답이 발생하였을 경우 무응답 조정이 없을 경우에 대한 편향이 0.71%포인트 발생한다고 볼 수 있다. 원표본의 응답자 (B)와 교체표본의 응답자 (D)의 임금근로자 비율은 64.84%와 67.33%로 교체된 표본에서 응답자의 임금근로자 비율이 약 2.5%포인트 더 높은 것을 알 수 있다. 이로부터 원표본 응답자와 교체표본 응답자 간에 응답 속성에 차이가 있음을 알 수 있다. 2) 원표본에 대해 무응답 가중치를 적용한 것 (A)과 표본교체 후 (C)의 임금근로자 비율은 각각 65.15%와 65.55%로 매우 근사하게 추정되었음을 알 수 있다. 이는 나머지 범주에서도 유사한 경향을 보인다. 결과적으로 종사상 지위에 대한 추정치의 경우 무응답 조정을 하기 전후에는 분명히 편향이 발생하고, 표본교체에 의한 응답자는 원표본 응답자 특성과 다를 수 있다. 또한 무응답 조정을 무응답 가중치를 적용한 경우와 표본교체를 한 경우 이 변수에 대한 추정치 차이는 거의 없다고 볼 수 있다.

각 모형별 분산비교는 각 범주별 표준오차 (standard error)를 사용하였다. 무응답가중치를 적용한 모형 A는 무응답가중치를 적용하지 않은 모형 B에 대해 표준오차가 같거나 비슷한 수준이었다. 한편, 표본교체 후 모형 C는 무응답가중치를 적용한 모형 A에 비해 표준오차가 작았다. 교체 표본 중 응답가구 모형인 D는 4가지 모형 중 표준오차가 가장 큰 것으로 나타났다. 이는 모형 D가 다른 모형들에 비해 표본크기가 작다는 것이 비율추정치의 분산을 높게 추정한다고 볼 수 있다. 본 연구는 주로 모형 A와 모형 C에 관심을 두고 있기 때문에 이 두 모형을 위주로 설명하였다. <표 6-14>의 '점유형태'와 '거쳐종류' 변수에 대해서도 동일한 방법으로 설명할 수 있다. 여기서는 구체적인 설명은 생략하기로 한다.

<표 6-15>는 사회조사 항목 중 임의로 선택한 의식관련 4개 항목에 대한 모형별 추정치와 표준오차를 나타낸 것이다. <표 6-14>의 응답자의 일반 특성 항목과 비슷하게 각 변수 내 범주에서 모형 D를 제외한 모형 A, B, C의 추정치 차이는 크지 않다. 재미있는 것은 무응답가중치를 적용한 모형 A는 원표본의 응답자만 사용한 모형 B와 교체표본을 포함한 모형 C 추정치 사이 값을 대체로 취한다. 이는 무응답 가중치를 적용함으로써 모형 B의 응답자만 이용한 추정치를 개선할 수 있다고 볼 수 있다. 게다가 모형 C는 모형 B와 모형 D 추정치 사이 값을 취한다. 이로부터 교체표본을 포함한 모형 C는 원표본의 응답 성향과 다소 차이가 있는 교체 표본 응답자들의 영향을 받는다고 볼 수 있다.

한편, 모형별 표준오차는 <표 6-14>의 결과와 매우 유사하다. 즉, 무응답 가중치 조정 전과 후인 모형 A와 B의 표준오차는 매우 유사하거나 동일한 값을 나타낸다. 무응답 가

중치 조정한 모형 A는 교체표본을 포함한 모형 C에 비해 표준오차가 약간 증가한 것을 알 수 있고, 변수 내 범주에 따라서는 상당히 크게 증가하는 것도 있음을 알 수 있다.

〈표 6-15〉 사회조사 의식관련 항목에 대한 모형별 추정치와 표준오차

	추정치 (%)				표준오차			
Model	A	B	C	D	A	B	C	D
결혼견해								
1.반드시	23.331	23.531	23.259	22.591	0.261	0.262	0.220	0.405
2.하는편이	42.344	42.313	42.618	43.368	0.306	0.305	0.258	0.480
3.그저그래	29.783	29.618	29.637	29.683	0.284	0.282	0.238	0.443
4.안해도	2.624	2.611	2.592	2.546	0.099	0.099	0.083	0.153
5.안해야	0.490	0.492	0.502	0.526	0.043	0.043	0.037	0.070
6.잘모름	1.428	1.435	1.392	1.287	0.073	0.074	0.061	0.109
주관적만족도								
1.매우만족	6.861	6.813	6.815	6.820	0.157	0.156	0.131	0.244
2.만족	22.522	22.492	22.895	23.887	0.259	0.258	0.219	0.413
3.보통	46.369	46.401	46.143	45.510	0.309	0.308	0.260	0.483
4.불만족	19.157	19.218	19.066	18.693	0.243	0.243	0.205	0.378
5.전혀불만족	5.092	5.076	5.081	5.091	0.136	0.136	0.114	0.213
스트레스정도								
1.매우	11.540	11.492	11.176	10.398	0.198	0.197	0.164	0.296
2.느낀편	57.641	57.592	57.990	58.971	0.306	0.305	0.257	0.477
3.별로	27.842	27.901	27.889	27.860	0.277	0.277	0.234	0.435
4.전혀안느낌	2.976	3.015	2.945	2.771	0.105	0.106	0.088	0.159
주관적건강평가								
1.매우좋은편	9.238	9.168	9.146	9.093	0.180	0.178	0.150	0.279
2.좋은편	34.463	34.324	34.986	36.615	0.295	0.293	0.249	0.467
3.보통	38.256	38.241	37.820	36.784	0.301	0.300	0.253	0.467
4.나쁜편	15.178	15.366	15.152	14.625	0.221	0.223	0.187	0.343
5.매우나쁜편	2.866	2.901	2.896	2.884	0.103	0.104	0.087	0.162

따라서 <표 6-14>와 <표 6-15>의 결과를 종합해 보면, 무응답 가중치 조정을 통해 추정치를 개선할 수 있으며, 이러한 무응답 가중치 조정을 통해 교체표본을 포함한 자료의 추정과 매우 근사해지는 것을 알 수 있다. 일반적으로 가중치 조정은 분산의 증가를 초래할 수 있고, 우리 분석결과에서도 무응답 가중치 조정으로 인해 약간의 분산 증가가 있음을 알 수 있다. 그러나 표준오차의 차이가 소수 둘째 자리에서 발생한다는 점을 고

려할 때, 추정치의 정도를 크게 훼손하는 정도는 아니라고 판단된다. 특히, 무응답으로 인한 원 표본크기를 유지하기 위해 추가로 방문하는 가구에 소요되는 경제적 측면과 표본크기 증가에 따른 비표본오차의 증가를 고려할 때, 분산의 증가가 있더라도 실용성 측면에서 크게 장점이 있을 것으로 보인다.

2. 지역별고용조사의 경우

지역별고용조사에 4가지 모형을 적용하여 추정치를 계산하였다. 적용된 4개 모형의 분석대상 가구 수는 다음 <표 6-16>과 같다. 관심추정치는 모형별 실업률, 고용률, 주택유형별 거주율이다 (참고로 모형 A에서 무응답 층은 16개 시도 변수를 사용).

〈표 6-16〉 지역별고용조사의 4개 모형

모형	구성	가구수	가구원수
A	원표본 (응답+무응답)	143,476	-
B	원표본 중 응답	111,179	233,061
C	교체 후 전체	143,476	299,823
D	교체표본 중 응답	32,297	66,762

* 응답자료와 가구종합관리표 자료를 key변수(가구고유번호)로 연계한 혼합자료이므로 가구관리종합표를 기준으로 정리한 <표 6-9>의 빈도와 다름, 차이는 1% 미만

위의 <표 6-16>을 보면 원표본가구 중 32,297가구가 무응답하여 이 무응답가구가 모두 교체되었고, 원표본 및 교체표본가구 중 응답가구로만 구성된 자료의 가구수는 143,476가구이다. 이를 정리하면 응답률은 약 77.5%라고 볼 수 있다.

관심변수에 대한 4개 모형별 추정치와 표준오차는 <표 6-17>과 같다. 우선 추정치 측면에서 모형결과를 해석하면 다음과 같다. 주택유형의 경우 원표본의 응답가구 (B)와 교체표본의 응답가구 (D)의 주택거주율은 각각 53.29%와 59.24%로 차이가 큰 것으로 나타났다. 또한 원표본의 응답가구 (B)와 표본교체 후 (C)의 주택거주율은 53.29%와 54.62%로 표본교체 후 (C) 응답가구의 주택거주 비율은 높게 나타났다. 또한 무응답 가중치에 의해 조정한 후 (A)의 주택거주 비율은 53.28%로 표본교체 후 (C)의 주택거주율 54.62%에 매우 근사해진다는 것을 알 수 있다.

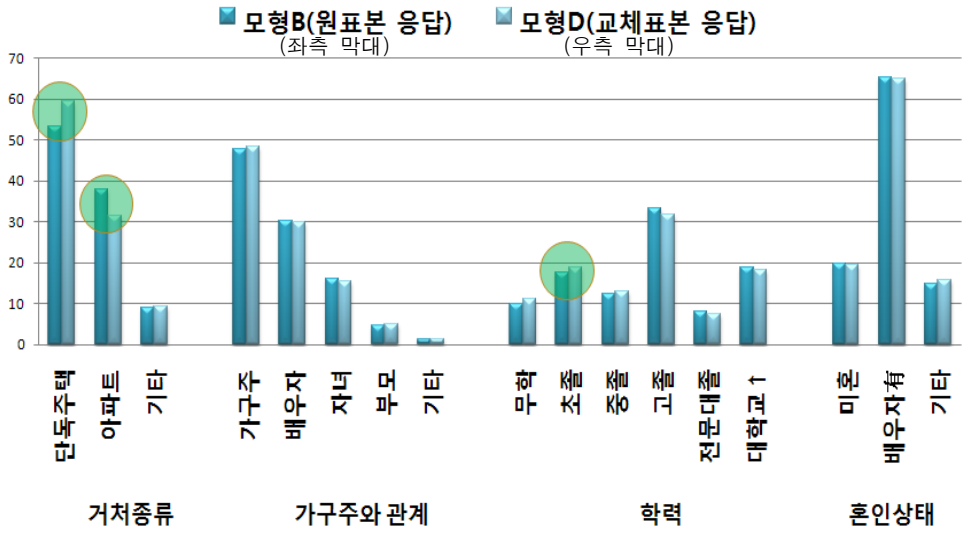
추정치의 정도 측면에서는 보면, 교체가구를 포함한 모형 C의 표준오차가 가장 작고, 교체가구의 응답가구만을 대상으로 한 모형 D의 표준오차가 가장 크다. 무응답 가중치를 적용한 모형 A와 원표본에서 무응답을 제외한 모형 B의 표준오차는 거의 유사한 결

과를 보인다. 이때 모형 A는 모형 C에 대해 표준오차가 거의 유사하거나 약간 크게 나타나는 경향이 있다. 따라서 사회조사의 결과와 마찬가지로 무응답 가중치 조정으로 인해 추정치의 분산이 약간 증가하는 것을 알 수 있다.

<표 6-17>의 나머지 변수들에 대해서도 동일한 방법으로 해석할 수 있고, 그 경향은 '주택유형' 과 같이 무응답 가중치 적용으로 추정치의 개선효과가 있으며, 대신에 분산의 증가를 초래하는 경향이 있음을 알 수 있다. 본 연구에서는 나머지 변수들에 대한 구체적인 해석은 생략하기로 한다.

〈표 6-17〉 지역별고용조사의 관심변수에 대한 모형별 추정치(%)와 표준오차

	추정치 (%)				표준오차			
변수	A	B	C	D	A	B	C	D
주택유형								
주택	53.278	53.294	54.616	59.235	0.104	0.104	0.091	0.191
아파트	37.755	37.791	36.366	31.387	0.101	0.101	0.088	0.180
기타	8.968	8.915	9.018	9.378	0.060	0.059	0.053	0.113
가구주관계								
가구주	47.565	47.576	47.732	48.276	0.104	0.103	0.091	0.193
배우자	30.338	30.344	30.261	29.971	0.095	0.095	0.084	0.177
미혼자녀	14.617	14.610	14.464	13.954	0.073	0.073	0.064	0.134
기혼자녀	1.133	1.134	1.145	1.183	0.022	0.022	0.019	0.042
손자녀	0.325	0.325	0.327	0.334	0.012	0.012	0.010	0.022
부모	4.573	4.567	4.608	4.751	0.043	0.043	0.038	0.082
조부모	0.131	0.130	0.133	0.145	0.008	0.008	0.007	0.015
미혼형제	0.670	0.666	0.684	0.746	0.017	0.017	0.015	0.033
기타	0.647	0.648	0.646	0.640	0.017	0.017	0.015	0.031
교육수준								
무학	9.841	9.860	10.159	11.203	0.062	0.062	0.055	0.122
초등학교	17.468	17.478	17.780	18.837	0.079	0.079	0.070	0.151
중학교	12.340	12.349	12.464	12.867	0.068	0.068	0.060	0.130
고등학교	33.285	33.302	32.920	31.587	0.098	0.098	0.086	0.180
전문대	8.154	8.166	7.995	7.399	0.057	0.057	0.050	0.101
4년대학	16.934	16.878	16.711	16.130	0.078	0.078	0.068	0.142
대학원이상	1.639	1.630	1.633	1.643	0.026	0.026	0.023	0.049
혼인상태	0.340	0.337	0.337	0.334	0.012	0.012	0.011	0.022
미혼	19.918	19.894	19.821	19.566	0.083	0.083	0.073	0.154
유배우	65.349	65.359	65.237	64.810	0.099	0.099	0.087	0.185
사별	11.927	11.943	12.130	12.781	0.067	0.067	0.060	0.129
이혼	2.806	2.804	2.813	2.843	0.034	0.034	0.030	0.064



[그림 6-3] 고용조사에서 원표본과 교체표본 응답자간 일반적 특성분포 비교

<표 6-18>은 지역별고용조사의 고용통계 작성과 관련한 항목을 임의로 선정하여 해당 변수에 대한 추정치와 표준오차를 나타낸 것이다. 실업률과 고용률은 가구 내 평균 실업률과 고용률을 나타낸다. '월평균 임금'은 조사표에서 '희망하는 월평균 소득수준'을 질문한 항목으로써 원표본 응답자 (모형 B)와 교체표본 응답자 (모형 C)의 추정치 간에는 약간 차이가 있는 것을 알 수 있다. 즉, 모형 B의 추정치가 모형 C 추정치에 비해 모든 범주에서 약간 낮다. 특히 200만 원 이상 높은 수준으로 갈수록 두 추정치 간 차이가 크게 나타나고 있다. 그리고 '10번 문항'은 지난 4주 내에 직업을 구해 본 비율을 구한 것으로 교체표본은 구직활동을 한 경험 있는 응답자비율이 원표본에 비해 더 낮은 것으로 나타났다. 이는 원표본 응답자가구의 실업률이 교체표본 응답자 가구의 실업률보다 더 크다는 것을 의미한다고 볼 수 있다. 한편, 응답가구만 이용한 모형 D는 고용률과 실업률이 다른 모형들에 비해 약간 낮게 추정되는 경향이 있다. 무응답 가중치 적용 전과 후의 모형 A와 B는 고용률과 실업률 추정치는 거의 동일한 값을 갖는다. 모형 A는 교체표본을 포함한 모형 C에 대해 아주 근소한 차이지만 약간 높게 추정되는 경향이 있다. 모형 C는 모형 D의 특성을 반영하고, 모형 A는 무응답 가중치 적용효과로 인해 모형 B의 추정치를 개선한다고 볼 수 있다.

이처럼 지역별고용조사의 교체표본 응답자는 원표본 응답자가구에 비해 취업가구가 더 많이 포함되어 있다는 것을 미루어 짐작할 수 있다. 따라서 이러한 교체된 표본이 집계자료로 사용될 경우, 지역별고용조사에서 실업률과 고용률이 모두 낮게 추정될 가능

성이 높다. 실제로 우리 분석결과에서도 이러한 교체표본이 포함된 모형 C의 실업률 추정치가 모형 A, B 추정치보다 더 낮게 추정된 것으로 나타났다. 물론, 교체표본이 확률적 메카니즘에 의해 추출된 것이라면 크게 문제될 것은 아니다. 그러나 실제로 교체표본 선정에 있어서 확률적 메카니즘을 사용하기는 그렇게 수월해 보이지는 않는다. 그렇다면 이렇게 교체표본의 응답 결과를 추정에 사용한다는 것은 그만큼 조사통계의 편향을 유도할 가능성이 커 보인다.

종합적으로 보면, 사회조사 경우와 마찬가지로, 무응답을 조정한 후와 전은 그 추정치들에 있어서 명백한 차이를 보였고, 교체된 가구는 이들을 포함한 전체자료의 추정치에 영향을 미치는 것을 알 수 있다. 또한 무응답 가중치 조정과 표본교체에 의한 추정치는 대체로 유사한 결과를 보였다. 무응답 가중치 적용은 분산의 증가를 초래하는 대신에, 추정치의 편향을 줄이는 효과가 있다고 볼 수 있다. 즉, 무응답 가중치 적용 후 추정치는 무응답이 없는 자료의 추정치 (모형 C)에 더 가까워짐을 알 수 있다.

〈표 6-18〉 고용통계 관련 변수에 대한 모형별 추정치(%)와 표준오차(230개 가중치)

	추정치 (%)				표준오차			
	A	B	C	D	A	B	C	D
고용률	58.439	58.415	58.316	57.973	0.001	0.001	0.001	0.001
실업률	2.046	2.048	2.032	1.972	0.000	0.000	0.000	0.000
월평균임금[4]	186.392	186.308	185.525	182.587	0.458	0.454	0.402	0.867
구직활동[5]								
했다	3.141	3.144	3.084	2.876	0.056	0.056	0.049	1.100
안했다	96.860	96.856	96.916	97.124	0.056	0.056	0.049	1.100
희망소득[6]								
~50만원	3.552	3.602	3.401	2.622	0.229	0.231	0.201	0.390
50~100만원	20.976	21.041	20.575	18.772	0.505	0.506	0.447	0.953
100~150만원	27.736	27.674	27.804	28.308	0.557	0.555	0.496	1.100
150~200만원	21.962	21.949	22.067	22.527	0.515	0.514	0.459	1.020
200~250만원	15.818	15.777	16.098	17.342	0.454	0.452	0.407	0.924
250~300만원	6.077	6.064	5.994	5.721	0.297	0.296	0.263	0.567
300~400만원	2.572	2.586	2.716	3.218	0.196	0.197	0.180	0.431
400만원~	1.307	1.308	1.346	1.490	0.141	0.141	0.127	0.296

4) 조사표 29번 문항 '최근 3개월간 직장에서 받은 월평균 임금 또는 보수는 얼마였습니까?'
5) 조사표 10번 문항 '지난 4주 내에 직장을 구해 보셨습니까?'
6) 조사표 19번 문항 ' 희망하는 월평균 소득 수준은 얼마입니까?'

이러한 현상은 소지역 단위로 갈수록 더 크게 나타날 것으로 예상된다. 특히 지역별 고용조사는 소지역통계 생산을 목적으로 한다는 점에서 최소한 시도 또는 시군별로 이러한 변화 양상을 살펴볼 필요가 있겠다. 물론 지역에서의 추정치와 표준오차 경향은 전국 수준과 거의 유사할 것으로 예상된다. 16개 시도에 대해, 고용률과 실업률에 대한 추정치와 표준오차를 모형별로 비교하였다. 16개 시도 추정치 계산에서 모형 A에 필요한 가중치는 230개 시군구를 무응답층으로 하여 무응답 가중치를 사용하였다.

<표 6-19>와 <표 6-20>은 지역별고용조사의 시도별 고용률과 실업률 그리고 각 추정치의 표준오차를 나타낸다. 이들 표로부터 알 수 있는 사실은 우선, 전국 추정치에 비해 16개 시도 추정치에서 모형별 추정치의 차이가 더 커진다는 것이다. 특히 모형 D와 모형 B의 추정치 간 차이는 전국 단위에 비해 훨씬 커졌다. 전국 고용률이 모형 D에서 57.973%인 것에 비해 서울은 52.635%로 약 5.3%p 낮게 추정되었다. 서울에서 모형 D와 C 추정치 간 차이는 1.3%p로 전국 단위 차이인 0.343%p보다 훨씬 높게 나타났다. 그러나 무응답 가중치 조정 후에는 서울의 추정치 간 차이는 (모형 A-모형 C=) 0.209%p로 전국 단위 차이인 0.123%에 크게 다르지 않다.

<표 6-20>은 <표 6-19>와 마찬가지로 16개 전체 시도에 걸쳐 볼 때, 교체표본 응답자(모형 D)들의 실업률이 대체로 낮고, 원표본 응답자에 교체표본 응답자 자료가 포함되면(모형 C) 실업률이 약간 낮아지는 경향이 있다. 하지만, 이러한 현상은 서울, 부산, 대구, 인천과 같은 대도시 지역에서는 오히려 반대의 현상이 나타나기도 하였다. 즉, 이들 지역에서는 교체표본의 실업이 높은 반면, 원표본에 교체표본이 포함되면 오히려 실업률이 낮아지는 경향이 있다. 이를 통해 이들 지역에서 실업가구가 전국에 비해 더 많을 것이라는 것을 짐작할 수 있다. 또한 무응답가중치 조정 후 (모형 A)는 추정치의 편향이 훨씬 줄어드는 것을 알 수 있다.

<표 6-19>와 <표 6-20>의 표준오차는 전국단위에서와 유사한 경향을 보였다. 즉, 세부단위별로 보면 전국 단위에 비해 표준오차가 약간 커지는 경향이 있다. 이는 세부단위별로 내려갈수록 자료의 변동성이 커지는 데서 비롯된 당연한 현상이라 할 수 있다. 원표본에 교체표본이 포함된 경우 (모형 C)는 무응답이 없는 완벽한 자료로서 4개 모형 중 표준오차가 가장 작다. 무응답 가중치 조정 전 (모형 B)와 조정 후 (모형 A)에는 표준오차의 차이가 거의 동일한 것으로 나타났다. 이처럼 지역별고용조사와 같이 세부 단위 지역별로 통계를 작성하는 경우에도 무응답층을 잘 구성한다면 표본교체 방법에 의한 추정치에 근사한 값을 추정할 수 있을 것으로 판단된다. 시군구 단위에 분석도 향후 더 추가적으로 분석하고 비교하면 더 좋을 것이다.

〈표 6-19〉 지역별고용조사의 16개 시도 단위 4개 모형별 고용률 비교

	추정치(%)				표준오차			
	A	B	C	D	A	B	C	D
서울	53.525	53.567	53.316	52.635	0.003	0.003	0.002	0.004
부산	48.861	48.855	48.945	49.312	0.003	0.003	0.003	0.007
대구	52.098	52.076	52.365	53.344	0.005	0.005	0.004	0.010
인천	57.422	57.283	57.181	56.818	0.005	0.004	0.004	0.009
광주	52.172	52.246	52.303	52.482	0.008	0.008	0.007	0.015
대전	53.527	53.505	54.004	55.991	0.007	0.007	0.007	0.016
울산	57.474	57.409	57.542	58.430	0.006	0.006	0.006	0.016
경기	54.998	55.013	54.927	54.629	0.001	0.001	0.001	0.003
강원	58.889	58.949	58.555	57.306	0.003	0.003	0.002	0.005
충북	59.503	59.420	59.233	58.593	0.003	0.003	0.003	0.006
충남	60.285	60.265	60.106	59.561	0.003	0.003	0.002	0.005
전북	60.697	60.607	60.781	61.354	0.003	0.003	0.003	0.006
전남	65.904	65.935	65.640	64.630	0.002	0.002	0.002	0.004
경북	62.729	62.616	62.423	61.723	0.002	0.002	0.002	0.004
경남	57.804	57.717	57.907	58.723	0.002	0.002	0.002	0.005
제주	67.221	67.204	67.338	67.785	0.007	0.007	0.006	0.013

〈표 6-20〉 지역별고용조사의 시도 단위 4개 모형별 실업률 비교

	추정치(%)				표준오차			
	A	B	C	D	A	B	C	D
서울	3.226	3.220	3.261	3.374	0.001	0.001	0.001	0.002
부산	3.779	3.781	3.798	3.867	0.002	0.002	0.002	0.004
대구	3.115	3.133	3.288	3.819	0.002	0.002	0.002	0.005
인천	4.043	4.075	4.101	4.194	0.002	0.002	0.002	0.005
광주	3.815	3.802	3.743	3.548	0.004	0.004	0.004	0.008
대전	3.060	3.056	2.967	2.597	0.003	0.003	0.003	0.006
울산	2.911	2.909	2.713	1.380	0.003	0.003	0.003	0.006
경기	2.417	2.429	2.345	2.052	0.001	0.001	0.001	0.001
강원	1.022	1.007	1.032	1.113	0.001	0.001	0.001	0.001
충북	1.806	1.813	1.865	2.049	0.001	0.001	0.001	0.002
충남	1.963	1.958	1.939	1.874	0.001	0.001	0.001	0.002
전북	1.547	1.555	1.552	1.544	0.001	0.001	0.001	0.002
전남	1.151	1.142	1.110	0.998	0.001	0.001	0.001	0.001
경북	1.633	1.632	1.596	1.461	0.001	0.001	0.001	0.001
경남	1.632	1.653	1.626	1.504	0.001	0.001	0.001	0.001
제주	1.261	1.278	1.378	1.709	0.002	0.002	0.002	0.004

제4절 분석결과

1. 요약

지금까지 사회조사와 지역별고용조사를 대상으로 현장에서의 표본교체 효과를 살펴보았다. 전체적으로 교체된 표본은 원표본과 응답자 특성이 다른 것으로 나타났다. 변수별 구성분포를 보면, 인구학적 특성 변수에 대해서는 원표본과 교체표본의 응답자 간 응답 분포에 큰 차이가 없는 것으로 나타났다. 그렇지만 사회조사의 의식관련 변수 또는 지역별고용조사의 고용관련 변수의 경우는 대체로 차이가 있었고 경우에 따라서는 매우 큰 차이를 보이기도 하였다. 이와 같은 사실은 사회조사와 지역별고용조사 자료에서 모형별 추정치 간에 발생한 차이로부터 설명할 수 있다. 즉, 두 조사에서 원표본의 응답가구만 이용한 경우와 교체표본의 응답가구만 이용한 경우의 추정 간에 큰 차이가 있다. 또한 원표본의 응답가구만 사용했을 때와 원표본과 교체표본 응답자를 모두 사용했을 경우도 큰 차이를 보였다 (<표 6-14>, <표 6-15>, <표 6-17>, <표 6-18>).

앞에서 설명한 바와 같이, 학술계를 비롯한 국제사회의 실무통계 작성에 있어서, 현장에서의 표본교체는 대체로 권장되지 않는 분위기다. 표본교체에 대해서 가장 우려스러운 것은 원표본에서의 무응답자 가구와 대체된 가구의 속성이 다르다는 것이다. 이러한 교체자를 사용하게 되면 조사추정치의 편향을 유도할 수 있기 때문이다. 또한 표본교체자는 조사현장에서 표본교체가 가능하다는 사실을 알고 있다면 원표본에 대해 협조를 받으려는 노력을 게을리 할 수 있는 이유가 되기도 한다. 그렇지만, 표본교체가 무응답 가구와 유사한 속성을 갖는 가구로 대체될 수 있다면, 이는 표본의 크기 유지로 인한 자체가중 문제가 해결되면서 표본의 대표성을 유지할 수 있는 좋은 대안이 될 수 있다.

한편, 현장에서 100% 응답을 얻기가 어렵고, 무응답 가구와 거의 유사한 교체 가구를 선택하기가 어렵다면 이와 다른 대안이 필요하게 된다. 이런 측면을 고려할 때 실용성을 고려한 가장 현실적인 대안은 무응답 가중치 적용 방법을 고려할 수 있다. 본 연구에서는 무응답 가중치를 적용한 후 그 결과를 현장에서의 표본교체 결과와 비교하였다. 결과적으로 사회조사와 지역별고용조사 모두에서 무응답 가중치 조정은 조정 전에 비해 표본교체 결과에 더 근사해지는 것으로 나타났다. 즉, 표본교체 된 자료는 무응답이 없는 완벽한 자료라고 가정했을 때, 조사자료 내에 무응답 오차는 없다고 가정할 수 있다. 따라서 무응답 가중치 조정은 무응답 편향을 조정함으로써 기준 통계치 (교체표본을 포함한 자료: 모형 C)에 더 근사해지게 된다. 결국 무응답 가중치 조정은 무응답 편향을 줄임으로써 추정치의 정확성을 향상시키는 역할을 하게 된다. 이처럼 가구단위 무응답에 대해 이 무응답에 대한 고려가 없을 경우, 즉 무응답 조정을 하지 않고 응답가구만으

로 추정하게 되면, 이는 추정치 결과의 편향을 초래할 가능성이 크다고 볼 수 있다.

따라서 현장에서의 표본교체는 원표본의 크기를 유지하면서 무응답 편향이 없는 추정치를 얻을 수 있다. 그러나 사실, 교체된 표본이 원표본의 무응답한 가구의 특성을 그대로 대체하기 어렵다는 점을 감안한다면 어떤 오차가 추정치에 포함되었는지는 측정하기 어렵다. 본 연구에서 살펴본 바에 따르면 원표본의 응답가구와 교체표본의 응답가구의 특성이 대체로 유사하지만, 주요항목에 있어서는 약간 차이를 보인다는 점에서, 교체표본 활용은 매우 신중할 필요가 있을 것이다. 그렇지만 교체표본을 사용할 경우, 무응답 가중치 조정에 비해 적은 분산을 갖는 추정치를 얻을 수 있다는 장점도 있다.

현장에서의 표본교체는 본 연구에서의 자료분석 결과에 비추어 볼 때, 무응답 가중치 조정방법에 비해 무응답 편향과 분산측면에서 더 좋은 추정치를 유도할 수도 있다. 단, 교체표본이 원표본의 무응답 가구의 특성과 주요변수에 대해서 유사한 응답을 할 것이라는 대전제가 필요하다. 실제로 이러한 전제조건은 어떤 핵심변수에 대해서는 만족되기 어려울 것으로 보인다. 실제로 사회조사와 지역별고용조사의 표본교체는 조사구 안에서 이루어지는 것이 일반적이다. 그리고 표본추출단위인 조사는 아파트조사구와 주택조사구에 구분되어 있어서, 표본교체가 발생하더라도 주택유형이나 성과 같은 일반적인 인구학적 특성 분포는 그렇게 많이 바뀌지 않게 된다.

무응답이 없는 자료를 얻기 위해 즉, 원표본 크기를 유지하는 것이 교체표본을 사용하는 목적이라면 이것은 비용과 비표본오차 측면에서 그렇게 좋은 방법이라 할 수 없을 것이다. 왜냐하면, <표 6-2>와 <표 6-3>에 따르면 사회조사와 지역별고용조사는 원표본에서 무응답이 발생하면 조사구마다 평균적으로 약 5가구 이상은 추가 방문하는 것을 알 수 있다. 2010 사회조사의 경우, 무응답가구는 5,012가구였다. 이 가구를 교체하기 위해 8,114가구를 추가로 조사하였다. 8,114가구는 원표본 17,103가구의 약 47%에 해당되는 크기이다. 또한 지역별고용조사에서는 부재, 거부, 불능의 이유로 무응답한 가구의 수는 32,013가구였다. 이 가구를 응답가구로 교체하기 위해 53,404가구가 추가로 조사되었다. 원표본 크기 143,352가구의 약 37%에 해당하는 크기이다. 이 가구를 추가로 방문하기 위해 소요되는 경제적, 시간적 비용크기는 충분히 예상 가능할 것이다. 뿐만 아니라 짧은 시간에 추가 방문까지 완료해야 하는 조사원들의 입장에서 보면 현재 두 조사에 주어진 조사기간은 매우 짧다 (사회조사 13일간, 지역별고용조사 12일간). 그만큼 비표본오차가 발생할 가능성이 클 것이다. 이것은 조사원들의 불성실 내지는 능력의 문제라기보다는 물리적으로 조사를 정확하게 완성해내는 것이 쉽지 않다는 것이다.

우리는 본 연구로부터 세 가지 흥미로운 사실을 발견할 수 있었다. 1) 무응답을 대체하기 위해 선택된 교체가구는 원표본의 응답가구 특성과 사뭇 다르다는 것이다. 2) 무응답 조정은 반드시 필요하다는 것이다. 즉, 무응답 조정 전과 후의 관심변수의 추정치가

달라진다는 점에서 그 이유를 들 수 있다. 3) 표본교체에 의한 무응답 조정은 무응답 가중치 조정에 의한 추정치와 크게 다르지 않다는 것이다. 특히 세 번째 사실은 표본교체의 목적이 무응답률을 줄이는 데 있다면 오히려 무응답 가중치 조정이 좋은 대안이 될 수 있을 것이다. 사실 표본교체에 의한 추가 표본 조사는 비용과 조사현장에서의 부담이 매우 크게 작용하기 때문이다. 또한 통계적 무응답 가중치 조정은 좋은 무응답층을 잘 구성하거나 무응답 가중치 모형을 잘 설정할 수 있다면 표본교체에 비해 편향을 줄이는 효과가 훨씬 크다는 것은 이론적으로 알려진 사실이다. 통계적 방법에 의한 무응답 가중치 작업에 관한 연구는 향후 과제로 남긴다.

2. 표본교체의 장단점과 가이드라인

표본교체 방법이 갖는 장단점과 실제 적용에 따른 가이드라인을 Vehover (1999)와 본 연구 결과를 토대로 정리하면 다음과 같다. 특히 Vehover (1999)가 제시한 내용은 본 연구결과에서도 일부를 제외하고는 상당부분 검증되었다고 볼 수 있다.

가. 장점

① 단순성 : 표본교체는 자체가중 표본의 성격을 유지할 수 있다. 자체가중 표본이 갖는 장점을 충분히 살릴 수 있다.

② 표본크기 통제 : 다소 약한 장점이다. 왜냐하면 표본교체 시 현장에서 제어할 수 없는 부적격 표본의 발생에 따른 표본크기의 변동이 발생하기 때문이다. 더구나 조사원들에 따라서 업무가중과 무응답자체가 각 조사원들의 방문회수에서 어떤 차이를 유발할 수 있다. 이런 문제를 줄이기 위해서 Kish (1965)는 약간 복잡하긴 하지만 부가 표본을 사용하여 원표본과 동일하게 현장을 통제할 수 있다고 하였다. 즉, 현장에서 조사원이 교체 표본을 선택하는 과정에 참여하지 않아야 한다.

③ 무응답 편향 제거 : 무응답 조정을 하지 않은 경우에 비해 표본교체는 무응답 편향을 줄이는 데 기여한다. 즉, 무응답이 특정 지역, 도시-비도시 등과 같이 특정 그룹 수준에서 발생할 때 편향이 제거된다. 물론 이것은 무응답 조정으로도 충분히 얻을 수 있는 결과이다.

④ 최적 표본 구조 : 표본교체 방법은 표본의 구조 유지를 보장한다. 특히 복합표본 설계에서 작은 층이나 작은 군이 사용되었을 때 유용한 방법이다. 게다가 결측을 피하게 됨으로써 표본설계의 기본 특성을 유지할 수 있다는 점에서 추정치의 정도 (precision)를 보장할 수 있다. 위의 3가지 장점은 그다지 의미 있는 것은 아니다. 그러나 최적의 표본 구조를 유지할 있다는 것은 표본교체의 가장 큰 이점이라 할 수 있다.

나. 단점

① 현장조사 통제 : 대면조사인 경우, 표본교체는 현장조사 통제를 어렵게 한다. 그 밖의 전화조사나 웹 조사의 경우 이런 문제는 상당부분 없앨 수 있다.
② 무응답이 제거된다는 환각 : 표본교체로 무응답이 없어진다는 환상은 매우 강할 수 있다. 또한 이런 환상 때문에 단위 무응답 문제를 처리하고자 하는 노력이 줄어들 수 있다.
③ 높은 무응답률 : 접촉하기 어려운 가구가 조사 제외 대상 가구로 분류되고 대체 가구로 교체될 수 있다는 사실을 조사원이 알고 있다면 조사원의 노력은 감소될 수 있다. 이러한 사실은 Vehovar (1994)의 실험 연구를 통해 밝혀진 바 있다. 실제로 본 연구에서는 지역별고용조사의 경우 추가조사가구의 응답률은 약 60%, 사회조사의 경우 약 61.7%로 원표본 대상 조사의 응답률보다 훨씬 낮은 것을 확인할 수 있다.
④ 현장조사 기간의 연장 : 표본교체로 인해 현장조사 기간이 길어질 수 있다.
⑤ 조사비용 증가 : 추가표본의 응답률의 그렇게 높지 않다는 점을 감안하면, 원표본의 무응답 가구를 완벽하게 교체하기 위해서는 실제 무응답가구수보다 훨씬 더 많은 가구를 더 방문해야 한다.

다. 적용상의 가이드라인

1) 표본교체는 대규모의 확률표본에 대한 교체인 경우 적절하지 않다. 대규모 확률표본은 다음의 특성 중 최소한 한 가지를 만족하는 경우에 해당된다.
○ 현장조사에 투입될 수 있는 시간이 짧다.
○ 표본교체 편향이 크다는 증거가 있다.
○ 현장조사 과정에 대한 통제가 약하거나 이를 위한 비용이 적다.

2) 다음의 현실적 이유는 표본교체 사용을 정당화 할 수 있다.
○ 자체가중 표본에 대한 필요성이 매우 강하다. 그러나 이런 경우, 다음의 표본교체 절차에 대해서 유효해야 한다.
- 가중치 작업에 대한 다른 이론적 이유가 없다.
- 표본교체는 무응답 편향을 제거할 수 있다. 적어도 대안적 방법만큼은 효과가 있어야 한다.
- 많은 표본 층에서 무응답 때문에 관측치가 없으면 위험하다.
- 표본교체에 의해 추정의 정도가 향상된 것에 대한 이점이 있다.

현실적으로 표본교체에 의한 추정치의 정도 향상은 적다. 그러나 표본 층이 작거나 무응답률이 높거나 층간 상관성이 약하다면 표본교체로 인한 이점은 상당히 있을 것이다. 그렇다 하더라도 통계청에서 실시되고 있는 조사기간은 대략 경상조사의 경우 7일, 분기, 연간조사의 경우 10일~13일 정도이다. 통계청의 대부분 조사가 대규모의 어려운 조사라는 점을 고려하면 그렇게 충분한 조사기간이라 보기는 어렵다. 또한 통계청은 확률표본에 근거하고 있기 때문에 비확률 교체표본의 사용은 추정치의 신뢰성 측면에서 상당히 우려되는 부분이기도 하다.

제5절 제언 및 향후 연구

현장에서의 표본교체는 높은 무응답률과 표본층에서 매우 적은 표본수를 갖는 조사라면 조건부적으로 정당화될 수 있을 수 있겠다. 조사규모가 작고 현장제어가 적절하게 이루어질 수 있는 조사에서 표본교체 대한 편향과 분산 이슈는 그렇게 중요하지 않을 수 있다. 이 경우는 현장조사 제어가 어렵다는 것과 함께 자료수집기간의 연장이 표본교체의 주요 단점이 될 수 있다. 대규모 표본조사에서 표본교체는 추정치의 정도 면에서 약간의 이점이 있을지라도 교체표본 편향은 그만큼 개입될 수 있다는 우려가 있다. 즉, 대규모조사에서 표본교체로 인한 추정치의 정도 향상은 크지 않다는 것이 일반적이다.

현실성 측면에서 교체표본 활용은 큰 장점은 없어 보인다. 조사비용과 함께 비표본오차가 상대적으로 커질 수 있기 때문이다. 이에 대해 무응답 가중치 조정은 표본교체와 달리 조사현장 제어, 조사비용 부담 및 비표본 오차 증가에 대한 부담 없이 추정치의 편향을 줄일 수 있는 좋은 대안이 될 수 있다. 그러나 무응답 가중치 적용은 편향을 줄이는 대신 분산이 커질 수 있다는 단점이 있다. 본 연구에서 살펴보았듯이 매우 간단한 무응답 가중치 방법을 적용하였음에도 실제로 무응답 편향은 확실히 줄어든다는 것을 확인하였다. 분산의 증가가 있었지만, 교체표본의 분산에 비교했을 때 그렇게 큰 차이는 아니라는 점에 주목할 필요가 있다. 즉, 무응답 가중치 방법을 적용함으로써 교체표본을 활용한 추정치에 근사하게 추정할 수 있다.

따라서 본 연구에 따르면 사회조사와 지역별고용조사의 경우 표본교체보다는 무응답 가중치 조정방법을 사용하는 것도 추천할 만하다. 이는 추정치면에서 표본교체 결과와 거의 유사하고, 현장조사에 대한 추가 노력과 부담이 필요하지 않다는 점에서 표본교체보다는 장점이 많은 것 같다. 특히, 지역별고용조사와 사회조사는 대규모 표본조사이고 통계의 정확성을 요하는 조사, 무응답률이 그렇게 높지 않다는 점을 고려할 때 무응답 가중치 조정으로 충분히 정도 높은 추정치를 얻을 수 있을 것으로 기대한다.

그렇다고 해서 현재 통계청에서 일부 진행되고 있는 표본교체 방법을 당장 바꾸기에는 현실적으로 어려운 부분이 있을 것이다. 표본교체 방법이 교체된 표본에 대해 랜덤성만 유지할 수 있다 하더라도 무응답 보정방법으로서 큰 역할을 할 수 있다. 현재 사용되는 있는 표본교체 방법이 제대로 현장에서 진행될 수 있도록 그 절차를 보완해서 사용하는 것도 좋은 대안이 될 수 있다. 표본교체 방식을 당장 다른 방법으로 바꾸게 된다면 응답률 계산 및 현장조사 절차 등 다양한 측면에서 부차적인 문제점들이 발생하게 될 것이다. 따라서 현장의 표본교체 방식을 최대한 보완해서 사용하되 서서히 가중치 적용과 같은 다른 방식으로 바꿔가는 전략이 필요할 것으로 보인다.

예를들면 벨기에는 현장에서 표본교체를 하고 있는 몇 안 되는 국가 중 하나이다 (Demarest 등, 2002). 벨기에는 거부 가구에 대해 그 가구와 속성이 일치된 가구를 선택하거나 랜덤방식으로 교체 가구를 선택하고 있다. 교체된 가구는 가구주 연령, 가구원 수 그리 거주지역 등 기본적인 특성변수에 대해서 초기 선택된 가구와 연계가 되도록 선택하고 있다. 방법은 초기에 선택된 모든 가구에 대해 연계가 되는 가구를 3가구를 선택하고, 초기가구와 연계가구를 포함한 4가구는 동일한 특성을 갖는 그룹을 형성하게 된다. 만약 그룹 내에서 모든 가구가 무응답하게 되면 새로운 교체군 (substitution cluster)이 생성되고, 이때 교체군은 랜덤하게 선택된다. 이처럼 벨기에의 경우 최대한 무응답 가구와 유사한 가구로 교체함으로써 두 표본 간의 차이를 줄이고자 노력하고 있음을 알 수 있다.

무응답 가중치 방법의 경우, 현실적인 대응방법 외에도 방법론적으로 향후 이 분야에 대한 많은 연구가 필요하다. 현재 통계청 가구대상 조사의 경우 가계동향조사는 대표적으로 무응답 가중치 방법에 의해 단위 무응답을 조정하고 있다. 사회조사나 지역별고용조사의 경우도 각 조사의 특성을 고려하여 무응답 가중치 방법을 찾는다면 방법론적으로 활용하는데 매우 유용할 것이다. 본 연구에서 적용한 무응답 가중치 방법은 사회조사의 경우 16개 시도, 지역별고용조사의 경우 230개 시군구를 무응답층으로 하여 단순한 응답률의 역수를 사용하였다. 이론적으로 보면, 무응답층을 보다 세밀하게 구성하거나 모형을 이용한 가중치 조정방법을 사용한다면 훨씬 더 정도 높은 추정치를 얻을 수 있을 것이다. 따라서, 각 조사별로 무응답층으로 사용할 만한 정보를 추가로 탐색하거나, 각 자료에 적합한 모형을 이용하는 방법을 연구할 필요가 있을 것이다. 지역별고용조사의 경우는 가구관리종합표 자료를 통해 원표본 대상가구의 응답, 부재, 거부, 기타, 불능 등에 관한 정보를 파악할 수 있기 때문에 이러한 응답자들의 성향을 고려한 무응답 조정이 가능할 것이다.

이에 본 연구는 향후 응답자의 성향을 고려한 성향점수 모형이나 좀더 세분화된 무응답층을 고려하여 가중치 조정 방법을 찾을 계획이다. 가능하다면, 사회조사와 지역별고용조사를 시작으로 단위 무응답 가중치 적용 연구를 수행하고, 이를 가구단위 대상 조사에 확대 적용하는 연구가 진행될 수 있기를 기대한다.

참고문헌

Champman, D. and Roman, A. (1985). An Investigation of substitution for an RDD survey. Processings fo the Survey Research, 45-61.

Cochran, W. (1977). Sampling techniques. New York: Wiley.

Demarest, S. Gisle, L., and Van der Heyden, J. (2007). Playing hard to get: field substitutions in health surveys.

Giommi, A. and Rocco, E. (2003). Evaluation of the impact of substitutions for unit nonresponse in the labour force survey of the municipality of Florence.

Groves, R. (1989). Survey errors and survey costs. New York: Wiley.

Kish, L. (1965). Survey sampling. New York: Wiley.

Lessler, L. and Kalsbeek, W. (1991). Non-sampling errors in surveys: Wiley.

Vehovar, V. (1999). Field substitution and unit nonresponse. Journal of Official Statistics, 15(2), 335-350.