한국통계학회논문집 제3권 3호
1996년 12월 pp. 145-159

표본조사에서 항목 무응답 대체 방법[1)]

김 영 원[2)], 조 선 경[3)]

요 약

항목 무응답은 표본조사에서 비표본오차를 발생시키는 중요한 요인으로 지적되고 있다. 본 논문에서는 현재까지 통계조사의 분석과정에서 직관적으로 제시된 다양한 항목 무응답 대체방법들을 정리하고, 이런 방법들 간의 장·단점과 무응답의 발생 형태에 따른 대체 효과를 실제 사회조사 자료를 이용한 모의실험을 통하여 비교, 분석하였다.

1. 서론

각종 통계자료는 경제·사회현상을 파악하는데 중요한 근거로 이용되고 있다. 이러한 통계자료를 생산하기 위한 표본조사에는 표본오차와 비표본오차가 발생되는데, 비표본오차는 표본조사의 조사 기획 단계부터 실제 자료 분석 단계 등 전체 표본조사 과정에서 다양한 형태의 실수 또는 결함에 의하여 발생하는 오차이다. 실제 통계조사에서 비표본오차는 조사의 정확도를 결정하는 상당히 중요한 요인임에도 불구하고 이에 대한 통계적인 이론이 명확하게 정립되어 있지 않은 것이 현실이다. 특히 무응답은 비표본오차를 발생시키는 중요한 요인 중 하나이다.

무응답에 의한 오차는 두 가지로 나눌 수 있는데 하나는 단위 무응답(unit nonresponse)으로 조사단위로부터 얻어진 정보가 전혀 없는 경우를 의미하며 다른 하나는 항목 무응답(item nonresponse)으로 조사단위가 표본조사에 참가는 했지만, 질문 중 몇 가지 항목에 대한 대답을 얻는 것을 실패한 경우를 말한다.

이와 같은 무응답의 종류에 따라서 적절한 처리 방법을 고려해 볼 수 있는데, 일반적으로 단위 무응답에는 가중값 조정(weighting adjustment)을 항목 무응답에는 자료의 대체(imputation) 방법이 바람직한 것으로 알려져 있다. 물론 이러한 무응답 처리 방법보다 먼저 고려되어야 할 것은 무응답을 사전에 예방하여 조사 단계에서 완전한 자료를 얻는 것이다. 그러나 완전한 자료를 얻기 위한 여러 가지 예방책(Lessler와 Kalsbeek, 1992)에도 불구하고 거의 모든 통계조사에서는 항목 무응답이 발생한다.

특히 소규모 표본조사에서는 분석 단계에서 이런 무응답을 무시하는 경우가 대부분이며, 이런 경우 조사 결과에 있어서 많은 오류가 발생할 수 있다. 현재까지 항목 무응답을 대체하는 다양한 방법들이 제시되어 있지만 이들 방법들 간의 효율성은 극히 제한된 경우에 한해 이론적인 연구결

1) 본 연구는 숙명여자대학교 1996년도 교비연구비 지원에 의해 수행되었음.
2) (140-742) 서울 용산구 청파동 2가 숙명여자대학교 통계학과 부교수
3) (140-742) 서울 용산구 청파동 2가 숙명여자대학교 통계학과 강사

과들이 밝혀져 있고, 대규모 통계조사(예를 들어 미국의 CPS; Current Population Survey)에서 얻어진 자료를 이용한 실증분석을 통해 각 적용사례별로 그 효과가 비교되고 있다(Ernst, 1978; Herzog와 Lancaster, 1980; David 등, 1986). 하지만 흔히 무응답이 무시되는 소규모 사회조사나 시장조사에서도 이런 대체 방법들에 의한 효과를 기대할 수 있을 것이다. 효과적인 대체방법의 활용을 위해서는 관심변수와 높은 상관관계를 갖는 보조변수들을 확보하는 것이 핵심적인 사항인데, 일반적인 소규모 사회조사에서는 변수들 간의 상관관계가 별로 크지 않다는 사실에 유의해야 할 것이다. 따라서 본 연구에서는 현재 활용되고 있는 항목 무응답 대체방법들을 정리하고, 이런 방법들 간의 장·단점을 실제 소규모 사회조사 자료를 이용한 모의실험을 통하여 파악하고자 한다.

본 논문에서는 우선 무응답이 발생하는 경우 이에 대한 적절한 조치를 취하지 않을 때 발생하는 문제점에 대하여 2절에서 간단히 살펴보고, 3절에서는 본 논문에서 관심을 갖고 있는 항목 무응답의 경우에 결측값을 대체하는 기존의 다양한 방법들을 정리하였다. 4절에서는 다양한 무응답 형태를 갖는 불완전자료에 3절에서 소개한 방법들을 구체적으로 적용하여 대체 방법들 간의 장·단점과 무응답 형태에 따른 대체 효과를 실제 사회조사 자료를 이용한 모의실험을 통하여 비교 분석하였다.

2. 무응답의 영향과 대체

많은 통계조사에 있어서 실제적으로 결측값이 발생함에도 불구하고 결측값을 무시한 통계 분석이 흔히 시행되고 있다. 일반적인 표본조사에서 가장 관심이 되는 모평균, μ의 추정에 있어서 표본자료에서 발생한 결측값을 무시하고 분석을 하였을 때 발생하는 문제점을 Kalton(1983)은 다음과 같이 정리하였다.

간단하게 한 개의 질문만을 고려하면 전체 조사대상은 무응답자와 응답자로 분류된다. 첨자 r과 m을 사용하여 응답자와 무응답자를 구분하면, 모집단의 평균과 표본평균은 다음과 같이 나타낼 수 있다.

$$\mu = \overline{R}\,\mu_r + \overline{M}\,\mu_m, \quad \bar{x} = \bar{r}\,\bar{x}_r + \bar{m}\,\bar{x}_m$$

여기서 $\overline{R}$과 $\overline{M}$ ($\bar{r}$과 $\bar{m}$)은 각각 모집단(표본)에서의 응답률과 무응답률을 나타내고 μ_r과 μ_m ($\bar{x}_r$과 $\bar{x}_m$)은 각각 모집단(표본)에서 응답자층과 무응답자층의 평균을 나타낸다. 무응답에 대하여 어떠한 처리도 하지 않은 경우에는 응답자 평균, $\bar{x}_r$을 모평균을 추정하는데 사용하게 되므로 이때 발생하는 편의는 다음과 같다.

$$B(\bar{x}_r) = \overline{M}(\mu_r - \mu_m)$$

만약 무응답률이 매우 낮거나 $\mu_r \fallingdotseq \mu_m$이면 근사적으로 $\bar{x}_r$은 μ에 대하여 불편추정량이 되지만 일반적으로 이런 가정은 현실적이지 못하다.

한편 모평균에 대한 통계적인 추론을 위하여 필수적인 모분산 S^2의 추정에 있어서 무응답에 의하여 다음과 같은 문제가 발생한다. 무응답을 무시하는 경우 S^2에 대한 추정량으로 가장 손쉽게 생각할 수 있는 것은 전체 조사대상 중 응답된 r개의 자료만을 이용한 표본분산,

$s_r^2 = \sum_{i=1}^{r}(x_i - \bar{x}_r)^2/(r-1)$이다. 이 추정량을 사용할 때 발생하는 편의는 다음과 같다.

$$B(s_r^2) = \overline{M}(S_r^2 - S_m^2) - \overline{RM}(\mu_r - \mu_m)^2$$

위 식의 첫 번째 항은 응답자층의 모분산 S_r^2과 무응답자층의 모분산 S_m^2이 같다면 (이 가정은 크게 비현실적인 가정은 아니다.) 무시할 만하고 두 번째 항에서 $\mu_r \neq \mu_m$이면 s_r^2은 S^2을 과소 추정하게 된다.

이런 결과를 참고로 하면 표본조사에서 $\mu_r \neq \mu_m$인 경우와 같이 무시할 수 없는 무응답(nonignorable nonresponse)이 발생하면 의미 있는 통계분석을 위해서는 무응답에 대한 적절한 대처 방안을 강구하는 것이 필수적이다. 이런 관점에서 항목 무응답의 경우 결측값 대신 적절한 값을 삽입하는 대체 방법이 주로 사용된다.

대체 방법의 사용 목적은 결측값을 다른 값으로 대체한 완전한 자료(complete data set)를 구성하여 이에 따라 기존의 완전한 자료에 적용되는 통계분석기법을 그대로 적용할 수 있게 하는 것이다. 이를 위하여 바람직한 대체 방법은 첫째, 추정에서의 무응답 편의를 감소시켜 주며 둘째, 모집단의 분포로부터 표본자료의 분포가 크게 왜곡되는 것을 방지하여야 한다.

한편 대체된 자료를 사용하는 경우 다음과 같은 문제점들을 인지할 필요가 있다. 첫째, 대체한 후의 결과가 결측값이 있는 자료에 기초했을 때보다 편의가 항상 더 작아진다고 보장할 수 없다. 둘째, 분석자가 대체된 자료를 완전한 자료인 것처럼 간주하면 상관관계와 같은 이변량 또는 다변량 모수에 대한 추정에서 편의가 발생되기 쉽다(Santos, 1981). 셋째, 통계량의 실제 분산을 과소 추정하는 단점이 있다(Ford, 1976; Bailar와 Bailar, 1978). 또한 대체를 하는 경우에는 이로 인한 응답들 간의 불일치가 발생할 수도 있으므로 자료의 편집 과정이 반드시 필요하다. 그러므로 대체 방법을 통한 완전한 자료를 제공할 때는 잠재적인 위험이 있다는 것을 경고해야 하며 이를 위하여 최소한 대체된 값은 이를 표시하여 주는 것이 바람직하다.

효과적인 대체 방법을 수행하기 위해서는 일반적으로 대체층(imputation class)의 형태로 표본을 나누는 작업이 필요하다. 대체층이란 결측값과 통계적으로 밀접한 관계가 있는 변수인 보조 변수(control variables)들의 교차분류의 형태로 나타나는 층을 말한다. 따라서 많은 대체방법에 있어서 성공적인 대체층의 구성이 대체방법의 효율성을 좌우하게 된다,

각 대체층 내에서 $\mu_r \fallingdotseq \mu_m$이 되도록 하면 무응답 편의를 줄일 수 있기 때문에 대체층은 상호 배반적이며 각 층내에서는 관심 변수에 대하여 동질적인 소그룹으로 이루어지는 것이 바람직하다. 만약에 이에 사용된 보조 변수가 결측값과 밀접하게 연관이 되어 있고 대체층의 수가 작으며 층내에서 더 동질적일수록 무응답의 편의는 줄어들 것이다. 이 직관적인 내용을 Ford(1983)는 통계적으로 설명하였고, Welniak과 Coder(1980)는 미국의 CPS 자료에 대한 실증분석을 통하여 대체층 구성의 중요성을 확인하였다. 이런 관점에서 Sonquist 등(1971)은 바람직한 대체층을 결정하기 위한 탐색 알고리즘을 제안하였다.

바람직한 대체층을 형성하여 이에 따른 대체 효과를 기대하기 위해서는 대체층이 다음 두 가지의 조건을 어느 정도 만족시켜 주는 것이 필요하다. 하나는 각 층내에서 결측값이 랜덤하게 발생하여 대체를 위하여 값을 제공하는 조사 단위인 제공자(donor)가 결측값을 갖는 조사 단위인 수용자(recipient) 값을 대신할 수 있어야 한다는 것이며 또 하나는 층내에서의 제공자 값의 분산이 크지 않아야 한다는 것이다.

3. 대체 방법

현재 활용되고 있는 많은 대체 방법들은 실제적으로 통계 자료를 처리하는 과정 중에 직관적인 입장에서 제시된 방법이기 때문에 통계 이론적으로 그 효율성 규명이 어려운 경우가 많다. 아울러 이 방법들은 기본적인 방법 이외에 다양한 변형이 가능하며, 필요에 따라 몇 가지 방법이 단계적으로 혼합되어 사용되어진다. 일반적으로 이런 대체방법들에 대한 효율성에 대한 이론적인 연구는 극히 제한적인 가정 하에서 수행된 것(Bailar 등, 1978; Ford 1983; Platek과 Gray, 1983)이 있지만 이런 연구 결과를 활용하여 실제 분석에서 효율적인 대체 방법을 구현하는데는 많은 어려움이 있다. 따라서 현실적으로는 각 방법의 장·단점을 정확히 파악하여 실제 문제에 적합한 방법을 모색하는 것이 필요하다. 이런 관점에서 우선 현재 제안되어 있는 대체 방법들을 정리하고 다음 절에서 모의실험을 통하여 이런 방법들 간의 장·단점을 검토하고자 한다.

(1) 연역적 대체 방법

연역적 대체 방법(deductive method)이란 결측값에 대해서 현재 자료로부터 확실시되거나, 확신을 갖고 결측값을 유추할 수 있는 경우에 사용할 수 있는 방법이다. 예를 들어 가족들 중 가장의 인종에 대한 응답이 없는 경우에 다른 가족들이 모두 백인이라면 이 사람도 백인이라고 생각하여 백인으로 이 결측값을 대신하는 방법이다.

그러므로, 이 방법은 보조 변수가 결측값을 결정하는데 있어서 거의 오차를 가지고 있지 않다고 볼 수 있는 경우에 사용이 가능한 방법으로 만약 이 오차가 너무나 커서 무시할 수 없다면 이 방법 이외에 다른 방법이 적용되어야 한다.

(2) 정확한 대응 대체

정확한 대응 대체(exact match imputation)는 경우에 따라 결측된 정보를 다른 조사자료로부터 얻을 수 있는 경우 결측값과 동일한 조사단위에 해당하는 다른 외부 자료의 값으로 대체하는 방법을 의미한다. 이 방법은 대체값이 다른 외부 조사결과로부터 얻어지는 값이므로 그 외부 자료가 신뢰할 수 있어야 하고, 조사단위를 대응시키는데 발생하는 비용이 많이 들고 사용하기에 복잡하다는 제약이 있다. 아울러 일부 결측값에 대하여는 대응된 조사단위를 찾을 수 없는 경우가 빈번히 발생하는 문제점이 있다.

Schieber(1978)는 저소득자의 수입에 대한 조사에서 행정자료를 활용한 대응 대체 방법, 핫덱, 회귀 대체 방법을 비교한 경험적 연구 결과를 얻었다.

(3) 핫덱

핫덱(hot-deck)은 1962년 미국 CPS(Current Population Survey)에서 노동력 항목에 처음으로 사용되기 시작하였고 그 이후 다양한 형태로 변형, 발전되었다. 전통적인 핫덱의 수행 과정은 다음과 같다.

지역, 직업, 성별 등 적절한 기준에 따라 조사된 자료를 순서대로 입력한 자료화일을 작성하는데, 이 과정에서 만약 대체를 고려하는 변수가 무응답인 경우 자료화일상의 입력 순서에 입각하여 바로 앞에 입력되는 응답자에 대한 관측값으로 무응답을 축차적으로 대체한다. 자료화일의 순서에 따른 이런 전통적인 대체방법을 일반적으로 축차 핫덱(sequential hot-deck)이라 부른다.

이 방법은 다른 외부 자료를 사용하는 과거의 콜드덱의 치료책으로 등장한 것으로 특히 컴퓨터

기술이 낙후된 상황에서 자료 처리상의 편리성과 비용이 중요하게 고려된 방법이다. 만약 자료의 순서와 조사 항목의 값이 우연히 양의 상관관계이면 매우 효과적인 방법이나, 몇 가지의 심각한 단점을 가지고 있다. 첫째, 결측값을 할당하는데 있어서 확률 구조가 아닌 자료화일 상의 순서에 의존한다는 점과 둘째, 동일한 제공값을 여러 번 사용하게 될 수 있다는 것이다. 이런 단점을 극복하고자 Cox(1980)는 이 방법을 수정한 가중 축차 핫덱(weighted sequential hot-deck imputation procedure) 방법을 제안하였다.

Ernst(1980)는 제한된 조건하에서 핫덱에 의한 추정량의 분산과 관련된 연구결과를 정리하였고, Bailar와 Bailar(1983)는 평균 대체 방법과 이 방법의 편의를 특정한 세 가지 조건 하에서 이론적으로 비교하였는데 특히 표본 관찰값이 각각 독립적인 경우는 두 가지 방법 모두 불편추정량이 되며 자료가 연속 상관(serially correlated)인 경우에는 핫덱 방법이 더 좋다는 결과를 얻었다. 하지만 현실적인 조건하에서 이 방법의 적용상의 특성에 대해 단언할 수 있는 이론적인 규명은 아직 미진한 상황이다.

한편 컴퓨터 기술의 발전에 따라 자료 처리 비용과 관련된 제한이 약화됨에 따라 통계적인 기법에 의존하는 전통적인 핫덱을 보완한 다양한 방법들이 제시되었는데, 다음에 정리된 평균대체, 랜덤 대체, 거리함수 대응 방법 등이 이런 방법들로 현재의 자료를 활용한다는 측면에서는 광범위한 의미의 핫덱에 포함될 수도 있지만 본 논문에서는 이를 세분하여 다루었다.

(4) 평균 대체

평균 대체(mean-value imputation) 방법은 표본을 대체층으로 나눈 후에 각 층내에서 응답자들의 평균을 구하여 그 층의 모든 무응답에 이 평균을 삽입하는 것이다.

이 방법은 동일한 대체층내에서 결측값이 모두 한 개의 값 즉, 평균으로 대체됨으로 인해 관심변수의 경험적 분포가 상당히 왜곡된다는 큰 단점이 있지만 사용하기가 쉽고 평균이나 총합들과 같은 일변량 모수에 대한 추정에 있어 무응답 편의를 감소시키는 데는 상당히 효과적이어서 간단한 점추정이나 예산상의 제약이 있을 때 주로 사용되는 방법이다.

이 방법을 가장 간단하게 적용한 것으로는 대체층을 고려하지 않고 모든 결측값 대신에 전체 응답자의 평균을 대체하는 방법을 고려할 수 있다. 이 방법은 상당한 문제점을 내포하고 있지만 실제적으로 많은 사회과학조사에서 대체 방안으로 사용되고 있는 것이 현실이다.

Cox와 Folsom(1978)은 핫덱과 평균 대체 방법을 특정한 실제 자료에 적용해서 두 방법 간의 편의, 분산, 평균제곱오차를 비교 연구하였다.

(5) 랜덤 대체

랜덤 대체(random imputation) 방법은 대체층 내에서 제공값을 확률 추출(probability sampling)에 의해서 선택하여 그 값으로 결측값을 대체하는 것이다. 이 방법의 최대 장점은 평균 대체 방법을 사용하는 경우 분포를 왜곡시킨다는 문제점을 어느 정도 해결할 수 있다는 것이다.

Ford(1983)는 하나의 대체층만을 고려할 때 표본 자료가 서로 독립적인 경우에 평균 대체와 랜덤 대체, 핫덱의 분산을 비교하였는데 랜덤 대체의 분산이 평균 대체 분산보다 크고, 핫덱의 분산이 랜덤 대체의 분산보다 크다는 것을 보였다. 또한 Kalton과 Kish(1984)는 제공값을 선정하는 확률추출방법으로 복원 랜덤, 비복원 랜덤, 층화추출 등을 사용할 때 발생하는 대체에 위한 분산의 증가량에 대한 연구결과를 제시하였다. 한편 Cox와 Folsom(1978)는 그들의 경험적 연구에서 응답률이 매우 낮고 무응답 편의가 크다면 랜덤 대체에 의한 편의 감소 효과가 통계량의 분산을 증가

시키는 단점을 상쇄할 수 있다는 점을 강조하였다.

(6) 거리 함수 대응

거리 함수 대응(distance function matching)은 수용자와 제공자에 대하여 정의된 보조 변수의 거리가 가장 가까운 값으로 결측값을 대체하는 방법이다. 이 방법은 보조변수가 질적인 변수인 경우에는 핫덱에서 대체층을 구성하는 작업과 유사한 결과를 얻게 되므로 보조변수가 양적 변수인 경우 주로 활용된다. 간단한 예로 한 개의 양적 보조 변수가 있는 경우에 거리함수는 보조 변수의 절대 차이 또는 보조 변수를 변환한 값의 절대 차이로 정의될 수도 있다. 이 방법은 보조변수들의 특성에 따라 여러 가지 방법으로 변형되어 사용될 수 있다(Sande, 1979).

(7) 회귀대체

회귀대체(regression imputation)는 무응답이 있는 항목을 종속변수로 응답된 보조변수들을 독립변수로 하는 회귀모형을 적용하는 방법으로 독립 변수들은 질적 변수이거나 양적 변수이어도 상관이 없다. 관심 변수가 질적 변수인 경우에는 대수 선형 또는 로지스틱 모형을 적용할 수 있다. 이 방법은 전통적인 핫덱, 랜덤 대체, 거리함수대응 방법들과 달리 현 자료의 값을 그대로 대체값으로 사용하는 것이 아니라 회귀모형을 통해 얻게 되는 예측값을 사용한다는 측면에서 큰 차이가 있다.

Kalton(1983)은 이 방법에서 분포의 왜곡을 방지하기 위해 예측값에 오차항을 반영하는 방법들을 제시하였고, 사례연구를 통하여 랜덤 대체 방법과 비교해서 회귀대체방법으로 대체된 값과 실제값의 평균절대거리(mean absolute distance)를 비교해 보면 회귀대체가 약간 좋음을 보여주고 있다. 또한 질적인 독립 변수만을 사용하는 경우에 회귀대체의 결과는 오차의 가정에 따라서 평균대체나 랜덤 대체와 같아짐을 보였다.

이 방법은 대체적으로 바람직한 방법이지만 똑같은 보조변수 값을 갖는 결측값이 다수인 경우 같은 값이 여러 번 사용될 수 있으며 적합된 모형이나 가정이 맞지 않는 경우에는 분포의 왜곡을 초래하는 문제점이 있다.

4. 모의실험을 통한 대체 방법 비교

소규모 사회과학조사는 강제성을 띤다거나 응답의 중요성에 대한 인식의 부족으로 무응답이 발생하는 경우가 다른 통계조사에 비해 많으며 대체 방법을 수행하는데 꼭 필요한 보조변수를 찾기가 용이하지 않다는 특성 때문에 효과적인 대체 방법이 필요함에도 불구하고 무응답이 무시되기 쉽다. 그래서 본 절에서는 앞에서 소개한 대체 방법들을 실제 사회과학 자료를 이용한 모의실험을 통해 그 효과를 분석하고자 한다. 이를 위하여 모의실험에 사용된 표본자료가 모집단의 특성을 잘 대표한다고 가정하고 표본 자료 중 랜덤하게 발생시킨 항목 무응답값을 얼마나 효율적으로 복원하는가를 비교하여 각 대체 방법들의 장·단점을 비교하고자 한다.

4.1. 모의실험 자료

이 자료는 1991년도에 실시된 D연구소에서 실시한 주부의 가사노동 실태에 관한 조사로서 이

조사의 목적은 주부의 가사노동 가치평가에 관한 연구이다. 조사대상지역은 제주도를 제외한 전국이며 주부를 조사대상으로 하였다. 조사결과로 얻은 응답 자료의 수는 1460개이며, 그 중에 이상값이라고 생각되는 값은 제거한 1440개를 모의실험을 위한 기초 자료로 하였다.

이 자료를 이용한 모의실험을 통해 대체 효과를 분석하기 위하여 완전한 자료에서 일부 관측값을 랜덤하게 선택하여 무응답 값으로 간주하여 실제값과 대체값의 차이를 비교하고자 한다. 랜덤하게 결측값을 만든 변수는 응답자가 손쉽게 응답하기 곤란하다고 판단되는 주부 본인이 가사 노동의 임금으로 적당하다고 생각하는 금액에 관한 문항이며, 완전한 자료에서 이 변수를 4개의 수준으로 나누었을 때 분포 현황은 <표 1>과 같다.

<표 1> 관심 변수의 구간별 도수 단위(만원)

구간	도수	퍼센트
30이하	321	22.3
31-50	664	46.0
51-70	252	17.5
71이상	203	14.1
합 계	1,440	100.0

먼저 무시할 수 없는 무응답을 고려하기 위해 <표 1>의 구간을 기준으로 구간별 무응답률을 <표2>와 같이 설정하였다. 무응답이 구간에 따라 급격하게 증가하는 경우와 완만하게 증가하는 2가지로 구분하기 위하여 경우A와 B를 고려하였다. (이와 반대 방향으로 무응답률이 증가하는 경우에도 그 결과는 큰 차이가 없었음.) 또한 전체 무응답률에 따른 대체효과를 파악하기 위하여 경우A와 B의 수준별 무응답률을 유지하면서 전체 무응답률의 변화를 고려한 <표 3>에 제시된 6가지 경우를 최종적인 모의실험 대상으로 하였다.

6가지 경우에 대하여 랜덤하게 무응답 자료를 발생시켜 각 경우마다 결측값을 무시한 경우 얻게 되는 평균과 표준편차와 완전한 자료에서 얻게 되는 평균과 표준편차의 차이를 비교한 결과는 <표 3>과 같다. 참고로 완전한 자료의 평균은 51.57이고 표본표준편차는 20.52이었다.

이 결과는 예상되는 것과 같이 무응답률이 증가할 때, 또한 그 무응답률이 구간에 따라 급격하게 증가하는 경우에 평균의 차이가 커지게 되는 것을 확인할 수 있다. 이제 <표 3>의 각 경우별로 다양한 대체 방법들을 적용시켜 그 방법들 간의 대체 효과를 비교해 보겠다.

대체 방법을 수행하기 이전에 대체층을 나누기 위하여 보조 변수를 선택하여야 하는데 그 방법으로는 보조 변수가 질적 변수인 경우에는 관심 변수가 그 변수의 수준에 따라 차이가 있는지를 검정하는 분산분석을 통해서 통계적으로 유의적인 변수를 선택하였고, 양적 변수인 경우에는 변수간의 피어슨의 상관계수가 유의적인 변수를 선택하였다. 보조 변수로 선정된 변수들은 첫째, 가정관리인을 만약에 둔다면 적당하다고 생각하는 임금(PAY)에 관한 응답이다. 이 변수는 양적 변수이지만 대체층 구성을 위하여 사용되는 경우 2개의 범주를 갖는 분류 변수로 이용하였다. 둘째는 4개의 범주로 구성된 주부의 학력(SCH)이며 셋째는 4개의 범주로 구성된 거주하는 지역의 규모(SCALE)가 선정되었다.

<표 2> 구간별 무응답률 (%)

	구간 1	구간 2	구간 3	구간 4
경우 A	10	20	30	40
경우 B	15	20	25	30

<표 3> 각 경우별 평균, 표본표준편차의 편차

경우	전체 무응답률(%)	평균의 차이	표본표준편차의 차이
A-1	10	-1.03	-0.48
A-2	22	-2.67	-1.26
A-3	30	-3.93	-2.11
B-1	10	-0.58	0
B-2	21	-1.41	-0.63
B-3	30	-2.21	-0.99

대체층을 구성하기 위하여 앞에 제시된 3개의 보조 변수들의 범주를 이용한 일반적인 교차 분류에 의해서 대체층을 구성할 수 있다. 아울러 본 논문에서는 Sonquist 등(1971)이 제시한 탐색 알고리즘(searching algorithm)을 통해서 대체층을 만들었는데 그 결과는 <표 4>과 같다. 탐색 알고리즘은 전체 응답자를 우선 두 개의 그룹으로 나누는데, 이 두 그룹은 그룹내의 변동은 최소화시키며 그룹간의 변동은 최대화시킨다는 관점에서 나온 것이다. 그리고 그 나누어진 각각의 그룹을 독립적으로 다시 두 개의 그룹으로 나누는데 위의 원칙에 따라 통계적으로 유의적이지 않을 때까지 이 과정을 반복한다. 이 자료에서는 이 과정을 통해 5개의 대체층을 얻었다.

<표 4> 탐색 알고리즘을 통해서 얻은 5개의 대체층

대체층의 기준			층구분	도수	평균	표본표준편차
PAY<50	중학교 이하	중소도시 이하	1	534	42.93	17.19
	고등학교 이상	특별시	2	169	52.45	18.33
		광역시 이하	3	352	48.29	16.81
PAY≥50	중소도시 이상		4	308	67.64	21.02
	읍,면지역		5	77	60.29	19.26

4.2. 대체 방법의 적용

실증분석을 통한 모의실험을 위하여 적용된 대체 방법은 다음과 같으며 각 대체 방법을 쉽게 구분하기 위해 괄호 안의 기호를 사용하였다.

[1] 전체 응답자의 평균을 각 결측값에 할당한다.(GM)

[2] 대체층을 PAY와 SCH를 이용해서 8개의 층으로 나눈 후 각 층별로 응답자들의 평균을 각 층의 결측값에 할당하는 평균 대체 방법을 적용한다.(M8)

[3] 대체층을 [2]와 같이 나눈 후 각 층내에서 랜덤 대체 방법을 적용한다.(R8)

[4] [2]와 같이 나뉜 8개의 대체층을 SCALE의 구분에 따라 32개의 층으로 나눈 후에 각각의 층마다의 그 빈도가 25 이하이면 가장 가까운 층과 묶어서 총 20개의 층으로 구분한 다음 평균 대체 방법을 적용한다.(M20)

[5] [4]에서와 같이 대체층을 나눈 후 랜덤 대체 방법을 적용한다.(R20)

[6] 대체층을 탐색 알고리즘을 통해 <표 4>과 같이 5개의 층으로 나눈 후 각 층별로 응답자들의 평균을 각 층의 결측값에 대체한다.(MS5)

[7] 대체층을 탐색 알고리즘을 통해 5개의 층으로 나눈 후 각 층내에서 랜덤 대체 방법을 적용한다.(RS5)

[8] 응답한 자료만을 가지고 PAY, SCH, SCALE을 독립 변수로 하는 회귀 대체 방법을 적용한다.(REG)

[9] [8]과 같이 회귀 대체 방법을 적용하는데, 랜덤하게 오차를 만들어서 추정값에 오차를 더한 값을 사용하였다. 오차항은 랜덤하게 선택된 제공자의 실제값과 그 때의 예측값과의 차이를 이용하였다.(REGE)

4.3. 모의실험 결과

대체 효과를 비교하기 위한 척도로 대체값과 실제값의 차이의 평균(DM)과 평균절대편차(mean absolute deviation; MAD)를 사용하였다. DM, MAD를 수식으로 표현하면 아래 식과 같은데 $\hat{y_i}$은 대체값, y_i은 실제값, m은 무응답 자료 수이다.

$$DM = \frac{1}{m}\sum_{i=1}^{m}(\hat{y_i} - y_i)$$

$$MAD = \frac{1}{m}\sum_{i=1}^{m}|\hat{y_i} - y_i|$$

여기서 DM은 모평균 추정에 있어서 대체된 자료에서의 추정값과 완전한 자료에서 얻은 추정값 간의 차이를 설명하고, MAD는 개별적인 결측자료에 대한 대체값과 실제값간의 평균절대편차(Mean Absolute Deviation)를 나타낸다. 개별적인 대체값에서는 실제값과 차이가 발생하지만 크고 작은 대체값이 서로 상쇄되어 평균추정값에는 큰 차이가 발생하지 않는 경우 적은 DM을 갖으면서 MAD는 크게 나타날 수가 있다. 평균 대체와 랜덤 대체를 비교하면 대체로 이런 현상을 보여 주고 있다.

무응답을 랜덤하게 생성시킴으로 인해서 특이한 무응답 자료가 발생하여 생기는 문제점을 피하고 안정적인 모의실험 결과를 얻기 위하여 <표3>에 제시된 6가지 경우의 무응답 형태에 따라 각각 다섯 개의 무응답 자료들을 발생시켜 이 자료에 대해 제시된 대체방법들을 적용하여 DM, MAD를 구한 후 이들 다섯 개의 결과를 각각 단순 평균한 결과를 <표 5>에 수록하였다. 여기서 랜덤 대체의 경우에는 랜덤 대체에 따른 특이한 결과를 배제하기 위하여 각 경우에 5회의 다중대체를 수행하여 그 평균값을 사용하였다(실제적으로 각 반복적인 랜덤 대체간에 큰 차이는 없었다). <표 5>에서 GM에서 얻은 DM은 무응답을 무시한 분석에서 추정값과 완전한 자료에서의 추정값간의 차이와 같고 따라서 이 값은 다른 대체방법들에 의하여 얻어지는 무응답 편의에 대한 대체 효과를 확인하는 기준값에 해당한다.

<표 5> 각 대체 방법에서의 DM, MAD

대체 방법	경우 A-1		경우 A-2		경우 A-3		경우 B-1		경우 B-2		경우 B-3	
	DM	MAD	DM	MAD	DM	MAD	DM	MAD	DM	MAD	DM	MAD
GM	-10.12	17.90	-11.65	18.34	-12.91	19.04	-5.75	16.46	-6.55	16.38	-7.24	16.58
M8	-8.25	16.63	-9.65	16.59	-10.49	15.72	-4.93	15.30	-5.58	15.34	-5.52	15.18
R8	-8.26	22.29	-9.59	21.46	-10.53	21.60	-5.00	20.38	-4.97	20.42	-5.36	20.04
M20	-8.14	16.39	-9.57	16.64	-10.42	17.18	-5.10	15.02	-5.51	15.25	-5.39	15.06
R20	-8.61	21.61	-9.36	21.57	-10.67	21.34	-5.19	20.18	-5.45	20.43	-5.54	19.87
MS5	-8.46	16.38	-9.49	16.91	-10.50	17.01	-4.51	15.35	-4.98	15.31	-5.59	15.39
RS5	-8.14	21.56	-9.13	21.89	-10.15	21.61	-4.66	20.87	-4.97	20.47	-5.46	20.43
REG	-7.45	15.55	-8.62	15.54	-9.26	15.72	-4.51	13.97	-4.70	13.94	-4.91	14.08
REGE	-7.63	20.83	-8.88	20.44	-9.19	20.22	-5.05	20.05	-4.45	18.95	-4.99	19.12

<표 5>의 모의실험 결과를 통하여 다음과 같은 사실을 확인할 수 있다.

첫째, 무응답을 무시한 방법(GM)에서 발생하는 편의와 대체방법을 사용하는 경우의 편의를 비교하여 보면 예상되는 바와 같이 관심변수의 수준에 따라 무응답이 급격하게 증가하는 경우(경우 A)에 완만하게 증가하는 경우(경우 B)에 비하여 대체방법을 강구하는 것이 무응답에 의한 편의를 줄이는데 도움을 준다는 사실을 확인할 수 있다. 특히 전체 무응답률이 낮고 무응답률이 수준별로 큰 차이가 없는 B-1의 경우(무시할 수 있는 무응답(ignorable non-response)에 가까움) 실제적으로 대체효과가 거의 나타나고 있지 않다. 아울러 전체 무응답률이 증가함에 따라 대체 효과도 증대된다는 것을 알 수 있다.

둘째, 전반적으로 REG방법이 다른 방법에 비하여 무응답 편의를 줄이는 데 있어서 가장 효과적인 것으로 나타났다. 이는 본 모의실험에 사용된 자료에 일반적인 사회과학조사에서는 흔치 않게 관심변수와 상당히 높은 상관관계(피어슨의 상관계수: 0.532)를 갖는 양적 보조변수 PAY를 활용하기 때문에 발생한 결과라고 판단된다. 만약 이런 보조변수가 확보되지 않은 상황이라면 평균

대체인 M8과 MS5가 효과적인 대체방법이라고 판단된다. 한편 이와 동일한 대체층을 사용한 랜덤 대체에 해당하는 R8과 RS5 그리고 REG에 오차항을 고려한 REGE의 경우 무응답 편의를 보정한다는 측면에서 각각 앞의 방법들과 큰 차이를 보이고 있지 않지만 MAD가 상당히 증가한다는 사실에 유의할 필요가 있다. 뒤에 고려할 분산추정과 분포의 왜곡이라는 측면에서 오히려 후자의 방법들이 설득력을 갖는 방법이 될 수 있다.

셋째, 대체층의 구성에 있어서 M8, R8, M20, R20와 같이 통계적으로 유의한 질적 변수를 사용한 일반적인 교차분류를 우선 고려할 수 있지만 대체 효과만을 고려한다면 Sonquist 등(1971)이 제시한 탐색 알고리즘에 의하여 구성된 MS5와 RS5가 대체층 구성에 있어서 효과적이라는 사실을 알 수 있다. 아울러 동질적인 대체층의 구성을 위하여 상당히 세분화된 대체층을 고려하기 쉬운데 M20과 R20을 참고하면 이런 방법이 대체 과정만 복잡하게 만들 뿐 큰 효과가 없다는 사실에 유의해야 한다.

한편 결측값을 대체할 때 또한 유의해야 하는 점은 대체값의 분포가 실제값의 분포를 크게 왜곡해서는 안된다는 것이다. 대체값들의 분포가 실제 분포와 얼마나 가까운가를 보기 위하여 경우 A-3에서 완전한 자료에서 무응답 값으로 제거되기 이전의 실제값과 각 대체 방법을 적용한 후 대체된 값들의 도수 분포를 정리한 결과는 <표6>과 같이 나타났다. 다른 경우에도 정도의 차이는 있지만 이와 거의 유사한 결과를 보였는데 이에 대한 모의실험 결과는 생략하였다.

<표 6> 각 방법의 대체값들의 도수 분포

구간	TV	GM	M8	R8	M20	R20	MS5	RS5	REG	REGE
1 - 10	2	.	.	10	.	1	.	3	.	1
11- 20	8	.	.	96	.	16	.	10	.	8
21- 30	33	.	.	70	.	83	.	89	6	34
31- 40	58	.	88	131	95	70	.	53	54	80
41- 50	119	438	97	58	90	171	278	144	160	107
51- 60	61	.	222	26	202	32	37	50	132	81
61- 70	40	.	31	24	51	19	123	29	61	58
71- 80	51	.	.	21	.	17	.	19	12	30
81- 90	6	.	.	1	.	.	.	2	7	21
91-100	59	.	.	27	.	27	.	38	6	6
101이상	1	.	.	1	.	2	.	1	.	12
합계	438	438	438	438	438	438	438	438	438	438

* TV는 제거되기 전의 실제값의 도수를 의미한다.

평균 대체와 오차를 고려하지 않은 회귀대체에 의한 대체값들의 분포는 원래 자료의 분포를 심하게 왜곡함을 알 수 있고, 이에 비하여 랜덤 대체 방법과 오차를 고려한 회귀모형 대체는 실제 분포와 유사한 분포를 유지하여 준다는 것을 알 수 있다.

아울러 이런 분포의 왜곡은 모평균에 대한 통계적인 추론에 있어서 필수적인 표준오차의 추정에 있어서도 문제점으로 지적되고 있다. 이에 대하여 Rubin(1987, 1996)은 다중대체에 의한 분산의 추정 방법을 제시하였는데 이에 대한 논의는 본 논문에서는 다루지 않는다. 이런 관점에서 완전한 자료에서의 표본표준편차와 대체된 자료에서의 표본표준편차의 차이를 비교해 본 결과는 <표 7>과 같다. <표 7>에서 NI는 무응답을 무시하고 응답자료만을 갖고 구한 표본표준오차를 의미한다.

<표 7> 각 대체 방법의 표본표준편차와 실제 표본표준편차의 차이

방법	경우 A-1	경우 A-2	경우 A-3	경우 B-1	경우 B-2	경우 B-3
NI	-0.28	-1.26	-2.11	-0.25	-0.63	-0.99
GM	-1.50	-3.62	-5.16	-1.32	-2.90	-4.23
M8	-1.13	-3.03	-4.38	-1.05	-2.36	-3.44
R8	-0.41	-1.20	-1.96	-0.25	-0.54	-0.90
M20	-1.25	-2.97	-4.00	-1.04	-2.22	-3.39
R20	-0.28	-1.07	-2.00	-0.30	-0.59	-0.95
MS5	-1.26	-2.94	-4.30	-1.07	-2.34	-3.47
RS5	-1.14	-1.01	-1.70	-0.18	-0.56	-0.78
REG	-1.13	-2.60	-3.70	-0.93	-2.07	-2.96
REGE	-0.42	-1.14	-1.72	-0.75	-0.50	-0.85

대체적으로 표본오차의 추정에 있어서 분포의 왜곡 문제와 마찬가지로 평균대체의 경우 표본표준오차를 상당히 과소추정 한다는 점에 유의하여야 하고, 이에 비하여 대체적으로 랜덤 대체 방법과 오차항을 고려한 회귀대체 방법이 큰 무리가 없다는 사실을 알 수 있다.

5. 결론

이상의 실증 분석을 통한 모의실험결과에서 볼 수 있듯이 무응답의 경향과 무응답률에 따라 대체 효과는 달라질 수 있음을 보았다.

무응답률이 그 수준에 따라 증가하는 비율이 완만한 경우와 무응답률이 작은 경우에는 대체의 효과가 크지 않음을 볼 수 있었고 수준에 따라 급격히 무응답률이 증가하는 경우와 전체 무응답

률이 큰 경우에 적절한 대체 방법의 사용이 보다 효과적임을 확인할 수 있다.

대체층을 사용하는 대체 방법들 중 대체층 안에서 평균 대체 방법보다는 랜덤 대체 방법을 사용하는 것이 분포 측면에서 실제값의 분포에 보다 가까운 대체값들을 얻을 수 있다는 장점이 있다. 앞 절의 결과에서 보면 평균 대체와 랜덤 대체를 통해서 얻은 각각의 대체값과 실제값의 차이의 평균(DM)을 비교해 보면 그 차이가 크게 나타나지 않지만 대체된 자료와 완전한 자료의 표본표준편차의 차이를 비교하거나 대체된 값들의 도수 분포를 실제 자료의 분포와 비교해보면 랜덤 대체가 평균 대체보다 상당히 만족할 만한 결과를 보여 주고 있다. 그러나 랜덤 대체는 평균 대체보다 MAD가 커지는 단점이 있다.

일반적으로 회귀 대체를 하는 경우보다는 랜덤 대체의 사용이 더 안전하다. 왜냐하면 회귀 대체의 경우에 회귀모형에 필요한 가정이 맞지 않으면, 대체값들의 왜곡을 초래할 수 있다. 이런 경우 일반적으로 랜덤 대체를 사용하면 이러한 왜곡을 피할 수 있다. 본 연구의 실증분석에서는 회귀대체가 무응답 편의를 줄이는데 상당히 효과적이었고, 특히 대체값의 분포 측면을 고려하면 오차항을 고려한 회귀 대체를 통하여 좋은 결과를 얻을 수 있었다. 반면 랜덤 대체 방법은 단지 대체층 안에서 랜덤하게 결측값이 발생한다는 가정만이 필요하며, 적절한 대체층의 구성이 매우 중요하다. 실증 분석의 결과에서 대체층을 여러 가지로 만들어서 비교해 보았을 때 탐색 알고리즘을 통한 대체층이 보조 변수들의 단순한 교차분류 형태인 다른 대체층보다 만족할 만한 대체 효과를 가져다 준다는 사실을 알 수 있었다.

각 자료의 특성에 따라 대체 효과는 달라질 수 있기 때문에 어떤 방법이 항상 가장 효과적이라고 말할 수는 없지만 대체 방법들 간의 장·단점을 파악하여 이에 따라 각 사례별로 효과적인 대체 방법을 선정하는 작업이 무응답 대체에 있어서 매우 중요하다는 사실에 유의하여야 한다. 아울러 통계조사를 설계하는 단계에서 높은 무응답률이 예상되는 항목에 대해서는 무응답 대체를 위한 항목을 설문지에 사전에 포함하는 방안을 고려해 볼 수 있을 것이다.

참고문헌

[1] Bailar, J. C. III and Bailar, B. A. (1978). Comparison of Two Procedures for Imputing Missing Survey Values. *American Statistical Association 1982 Proceedings of the Section on Survey Research Methods*, 462-467.

[2] Bailar, B. A. and Bailar, J. C. III (1983). Comparison of the Biases of the Hot deck Imputation Procedure with an 'Equal-Weight' Imputation Procedure. In William G. Madow and Ingram Olkin, eds. *Incomplete Data in Sample Surveys, Vol. 3, Proceedings of the Symposium.* New York: Academic, 209-311.

[3] Bailar, B. A., Bailey, L., and Corby, C. (1978). A Comparison of Some Adjustment and Weighting Procedures for Survey Data. In N. Krishnan Namboodiri, ed. *Survey Sampling and Measurement.* New York: Academic, 175-198.

[4] Cox, B. G. (1980). The Weighted Sequential Hot deck Imputation Procedures. *American Statistical Association 1980 Proceedings of the Section on Survey Research Methods*, 721-726.

[5] Cox, B. G. and Folsom, R. R. Jr. (1978). An Empirical Investigation of Alternative

Non-response Adjustment. *American Statistical Association 1978 Proceedings of the Section on Survey Research Methods*, 444-449.

[6] David, M., Little, J. A., Samuhel, M. E. and Triest, R. K. (1986). Alternative Methods for CPS Income Imputation. Journal of *American Statistical Association, 81, 29-41.*

[7] Ernst, L. R.. (1978). Weighting to Adjust for Partial Nonresponse. *American Statistical Association 1978 Proceedings of the Section on Survey Research Methods*, 468-473.

[8] Ernst, L. R.. (1980). Variance of the Estimated Mean for Several Imputation Procedures. *American Statistical Association 1980 Proceedings of the Section on Survey Research Methods*, 716-720.

[9] Ford, B. L. (1976). Missing Data Procedures; A Comparative Study. *American Statistical Association 1976 Proceedings of the Social Statistics Section, Part 1,* 324-329.

[10] Ford, B. L. (1983). An Overview of Hot-deck Procedures. In William G. Madow and Ingram Olkin, eds. *Incomplete Data in Sample Surveys Vol 2. Theory and Bibliographics.* New York: Academic, 185-207.

[11] Herzog, T. N. and Lancaster, C. (1980). Multiple imputation of Individual Social Security Benefit Amounts-Part1. *American Statistical Association 1980 Proceedings of the Section on Survey Research Methods*, 398-403.

[12] Kalton, G. (1983). *Compensating for Missing Survey Data. Research Report Series.* Ann Arbor, MI: Institute for Social Research, University of Michigan.

[13] Kalton, G. and Kish, L. (1984). Some Efficient Random Imputation Methods. *Communications in Statistics, Theory and Methods, 13,* 1919-1939.

[14] Lessler and Kalsbeek (1992). *Nonsampling Error in Surveys.* New York: Wiley.

[15] Platek, R. and Gray, G. B. (1983). Imputation Methodology: Total Survey Error, Part V. In William G. Madow and Ingram Olkin, eds. *Incomplete Data in Sample Surveys Vol 2. Theory and Bibliographics.* New York: Academic, 249-333.

[16] Rubin, D. B. (1987). Multiple Imputations for Nonresponse in Surveys. *John Wileys & Sons,* New York.

[17] Rubin, D. B. (1996). Multiple Imputations after 18+ Years. *Journals of the American Statistical Association, 91,* 473-489

[18] Sande, I. G. (1979). A Personal View of Hot Deck Imputation Procedures. *Survey Methodology, 5,* 238-258.

[19] Santos, R. L. (1981). Effects of Imputation on Complex Statistics. *Income Survey Development Program, Survey Development Research Center in Nonresponse and Imputation Report on Additional Task 2,* Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan.

[20] Schieber, S. J. (1978). A Comparison of Three Alternative Techniques for Allocating Unreported Social Security Income on the Survey of the Low-Income Aged and Disabled. *American Statistical Association 1978 Proceedings of the Section on Survey Research Methods,* 212-218.

[21] Sonquist, J. A., Baker, E. L. and Morgan, J. N. (1971). *Searching for Structure : An Approach to Analysis of Substantial Bodies of Micro-Data Documentation for a Computer Program.* Ann Arbor, MI: Institute for Social Research, University of Michigan.

[22] Welniak, E. J. and Coder, J. F. (1980). A Measure of the Bias in the March CPS Earnings Imputation System. *American Statistical Association 1980 Proceedings of the Section on Survey Research Methods,* 421-425.