연구논문

CART를 활용한 결측값 대체방법 : 인구주택총조사 혼인상태 항목을 중심으로*

Missing Value Imputation Method Using CART : For Marital Status in the Population and Housing Census

김 영 원** · 이 주 원***
Young−Won Kim · Ju−Won Lee

본 연구에서는 일반적인 사회조사에서 사용될 수 있는 효과적인 결측값 대체방법을 검토하기 위해 인구주택총조사 조사항목 중 혼인상태의 결측값을 대체할 수 있는 두 가지 방법을 제안하고 있다. 첫 번째 방법은 CART(Classification and Regression Tree)모형에서 얻어진 최대 예측확률을 기준으로 결측값을 대체하는 일종의 모형기반 접근법이고, 두 번째 방법은 CART 모형에서 얻어진 결과를 근거로 대체층을 구성하여 핫덱(hot−deck) 방법을 적용하는 대체방법이다. 효율성 비교를 위해 2000년 인구주택총조사를 위한 시험조사에서 얻어진 재조사 결과를 이용하여 오분류율을 검토해 본 결과 두 방법 중 CART 모형을 기반으로 핫덱 방법을 적용하는 것이 효율적이라는 결론을 얻을 수 있었다. 아울러 전국에 대해 동일한 모형을 설정한 경우와 거주지 특성에 따라 광역시·도의 동지역, 도의 읍·면지역으로 구분하여 대체방법을 적용하는 경우를 비교해 본 결과 지역 구분을 통한 효율성 향상 효과는 미흡한 것으로 파악되었다.

주제어: 인구주택총조사, 결측값 대체, 핫덱, CART 모형

We proposed imputation strategies for marital status in the Population and Housing Census 2000 in Korea to illustrate the effective missing value imputation methods for social survey. The marital status which have relatively high non−response rates in the Census are considered to develope the effective missing value imputation procedures. The Classification and Regression Tree(CART) is employed to construct the imputation cells for hot−deck imputation, as well as to predict the missing value by model−based approach. We compare two imputation methods which include the CART model−based imputation and the

* 본 연구는 숙명여자대학교 2001년도 교내연구비 지원에 의해 수행되었음.
** 교신저자(corresponding author) : 숙명여자대학교 수학통계학부 통계학전공 교수 김영원. E−mail : ywkim@sookmyung.ac.kr
*** 숙명여자대학교 통계학과 대학원

sequential hot−deck imputation based on CART. Also we check whether different modeling for each region provides the more improved results. The results suggest that the proposed hot−deck imputation based on CART is very efficient and strongly recommendable. And the results show that different modeling for each region is not necessary.

key words: Classification and Regression Tree, Hot−deck, Missing Value Imputation, Census

Ⅰ. 서론

통계조사에서 모든 조사대상으로부터 필요한 정보를 완벽하게 얻는다는 것은 쉬운 일이 아니다. 여러 가지 예방책에도 불구하고 조사과정 중의 무응답(non−response)은 거의 모든 통계조사에서 필연적으로 발생하게 된다(Lessler & Kalsbeek 1992). 통계조사에서 무응답이 생길 때 이를 무시하고 분석을 하면 추정결과에 편향(bias)이 발생하게 되고 따라서 조사결과에 대해 신뢰성이 떨어지게 된다.

무응답은 조사의 내용, 자료수집 방법, 조사단위들의 형태, 응답자의 태도 등에 많이 의존한다. 무응답을 분류하면 첫 번째는 단위 무응답(unit non−response)으로 조사단위로부터 얻은 정보가 하나도 없는 것을 의미하고, 두 번째로는 항목 무응답(item non−response)으로 응답을 해야 할 항목에 대해 응답을 하지 않거나 질문과는 무관한 응답을 함으로써 불필요한 자료가 되는 것을 의미한다(Kalton & Kasprzyk 1986).

무응답이 발생하였을 경우 효과적인 통계분석을 위해서는 적절한 방법을 통해 무응답을 처리하는 과정이 필요한데 단위 무응답의 경우는 가중값 조정방법을, 항목 무응답이 발생한 경우는 결측값을 채워 넣기 위해서 여러 가지 대체법(imputation)을 이용하는 것이 일반적이다. 여기서 대체법이란 조사 후에 무응답으로 인해서 발생한 결측값을 채워 넣는(fill−in, impute, replace) 방법을 의미한다. 결측값을 대체하는 이유는 결측값를 가능한 정확한 값으로 대체함으로써 무응답에 의한 문제를 줄이기 위해서

즉, 조사결과의 신뢰성을 높이기 위해서라고 할 수 있다.

대체를 이용하면 무응답에 의한 편향을 줄일 수 있고, 일반적인 통계분석기법을 그대로 적용할 수 있기 때문에 분석을 쉽고 간단하게 할 수 있으며, 또한 재조사 과정을 생략하여 조사시간을 단축시켜 주는 이점이 있다.

특히 이재원(2000)이 지적한 것과 같이 인구주택총조사와 같은 대규모 조사의 경우, 자료처리 시간을 단축하기 위해서는 전체 총조사 과정에 있어 상당히 오랜 기간이 소요되었던 재조사 및 자료편집 과정을 효율적으로 수행하는 것이 필요하다. 이에 따라 2000년 총조사에서는 1995년까지의 총조사에서는 적용하지 않았던 무응답 대체방법을 도입하는 것이 필요하게 되었다. 실제 총조사와 같은 대규모 통계조사에 있어서는 막대한 비용과 시간을 들여 무응답을 감소시키기 위한 노력을 하고 있지만, 필연적으로 많은 무응답이 발생하게 된다. 따라서 총조사 비용을 절감하는 동시에 신뢰할 수 있는 조사결과를 신속하게 도출하기 위해서는 각 항목에 따른 무응답을 대체할 수 있는 효율적인 기법을 개발하는 것이 필요하지만 아직 우리나라에서는 이에 대한 체계적인 연구가 매우 미흡한 실정이다.

Ryu 등(2001)은 우리나라 인구주택총조사에서 항목 무응답 대체방법을 개발하기 위한 기초적인 연구결과를 제시한 바 있다. 이들의 연구에서는 총조사를 위한 시험조사 중 일부 표본조사 자료를 이용하여 개략적인 결측값 대체방안을 제시하고 있으나 제한적인 표본자료만을 사용함에 따라 오분류 판정에 반영된 결측건수 등에 있어서 한계를 갖고 있다. 참고로 통계청에서 실시한 2000년 인구주택총조사 본 조사에서는 이들의 연구결과를 확대 적용하여 다양한 전수 및 표본조사 항목에 대한 결측값 대체가 이루어졌다.

한편 일반적으로 손쉽게 사용될 수 있는 항목 무응답 대체방법으로는 핫덱(hot－deck), 평균대체, 회귀대체 방법 등이 있으며, 김영원과 조선경(1996)은 실제 자료분석을 통해 이들 방법의 특성 및 효율성에 대한 비교·분석결과를 제시하고 있다. 하지만 이들 방법 중 평균대체와 회귀대

체 방법은 그 특성 상 관심변수가 연속형인 경우에 적용할 수 있는 방법이기 때문에 일반적인 사회조사에서 주로 관심대상이 되는 범주형 조사항목의 경우 핫덱이 가정 적합한 대체방법으로 알려져 있다.

핫덱을 사용하여 효율적인 결측값 대체를 하기 위해서는 각 대체층(imputation cell)내에서 가능하면 무응답이 MAR(missing at random)이 되도록 대체층을 구성하는 과정이 매우 중요하다. 이런 과정을 수행하기 위해 Sonquist 등(1971)은 탐색 알고리즘(searching algorithm)을 제시한 바 있으며, 미국, 캐나다 등의 통계청에서는 심층적인 연구를 통해 독자적인 대체층 구성방법을 갖고 있다.

하지만 일반적인 사회조사에 있어서 이런 심층적인 연구과정을 통해 대체층을 구성한다는 것은 현실적으로 실현되기 어렵기 때문에 현재까지 우리나라에서 수행된 각종 사회조사의 경우 결측값을 해결하기 위한 적절한 대처방안을 강구하고 있지 못하고 있는 것이 현실이다.

제시된 대체방법에서는 대용량의 자료로부터 변수들간의 의미 있는 관계를 탐색하는 데 효과적이라고 알려져 있는 데이터 마이닝 기법 중 Breiman 등(1984)이 제안한 CART(Classification and Regression Tree)를 도입하여 활용한다. 특히 이 방법은 흔히 접할 수 있는 기존의 SAS 또는 SPSS 등의 데이터 마이닝 모듈을 이용하기 때문에 다양한 분야의 사회조사에 있어서도 필요에 따라 얼마든지 손쉽게 적용 가능하다는 장점을 갖고 있다. 실제 CART는 Breiman 등(1984)이 지적한 것과 같이 무응답 대체층 구성을 목적으로 Sonquist 등(1971)이 제시한 탐색 알고리즘을 개량하여 체계화한 것으로 볼 수 있다.

일반적인 사회조사의 경우 대부분의 조사항목이 범주형으로 조사되고 있다는 측면을 반영하기 위해 총조사 조사항목 중 무응답률이 높은 대표적인 범주형 조사항목인 혼인상태를 중심으로 하여 본 연구를 통하여 적절한 결측값 대체방법을 제안하고 그 효율성을 검증하고자 한다.

본 연구에서는 혼인상태에 대한 적합한 대체방법을 구현하기 위해 범주형 항목의 경우 CART에서 얻어지는 최대 예측확률(maximum prediction

probability)을 이용하는 통계적 예측방법과 CART를 대체층 구성에 활용한 순차적 핫덱(sequential hot-deck)을 이용한 결측값 대체방법의 효율성을 비교해 본다. 아울러 이런 CART 모형을 활용하는 데 있어서 지역특성별로 CART 모형을 차별화하는 방안에 따른 효율성 향상 효과도 동시에 검토하고자 한다.

Ⅱ. 총조사에서 혼인상태에 대한 대체방법의 적용

1. 인구주택총조사를 위한 시험조사 개요 및 대체 적용 필요성

통계청에서는 1999년 11월에 2000년 인구주택총조사를 위한 시험조사를 실시하였다. 시험조사의 주요 목적은 사전검사 설문지와 조사절차의 문제점을 사전에 확인하기 위한 것이었다. 시험조사 설문지는 전수조사의 경우 22개 항목, 표본조사의 경우 34개 항목으로 구성되어 있었다. 전수조사는 16개의 시·도에서 추출된 1,058개의 조사구를 대상으로 실시되었다. 조사항목은 각 가구의 구성원과 관련된 문항으로 구성된 인구부문 문항과 주택과 관련된 문항으로 구성된 주택부문 문항으로 구분되어 있었다. 그리고 가구 또는 인구부문에 하나 또는 그 이상의 결측값이 있는 경우 일관된 값을 얻기 위해서 사전에 동의를 얻어 조사원들이 가구를 직접 방문하여 재조사하였다. 시험조사 결과 8개 항목은 전체 56개의 항목 가운데 상대적으로 높은 무응답률을 보여 주었다. 특히 인구부문에서는 남녀의 혼인상태를 묻는 항목의 무응답률이 매우 높게 나타났다.

통계청에서는 이런 결측값을 채우기 위해 3번의 재조사를 계획하였다. 하지만 적절한 대체방법을 강구하여 3번의 재조사 과정 중 2차 및 3차 재조사 과정을 생략할 수 있다면 비용뿐만 아니라 시간이란 측면에서 커다란 이득을 얻을 수 있다. 1차 재조사 과정은 조사된 자료의 오류 수정, 항목간 연관성 검토, 논리적인 방법에 의한 에디팅(editing) 등이 포함된

필수적인 재조사 과정으로 현실적으로 생략할 수 없다. 따라서 본 연구에서는 1차 재조사 후 적절한 대체방법을 적용하는 경우 원래 계획했던 3차에 걸친 재조사 과정 중 2차 및 3차 재조사 과정을 축소할 수 있는지에 대한 가능성을 검토하고, 아울러 주어진 여건 하에서 가장 효율적인 대체방법을 구현하는 것을 목적으로 한다.

2. 혼인상태에 대한 CART 모형의 적용

2000년 인구주택총조사 시험조사에서 인구부문에서 높은 무응답률을 보이고 있는 혼인상태에 대한 결측값을 효과적으로 대체할 수 있는 방법을 구현하기 위해 Breiman 등(1984)에 의해 제안된 CART 방법을 활용한다.

무응답 대체를 위해 CART 모형을 활용하는 방법으로 우선 CART 모형에서 얻어진 최대 예측확률에 의거해 결측값을 대체하는 모형기반 대체법, 그리고 CART에 의해 얻어진 대체층을 활용한 핫덱 대체법을 비교해 보기로 한다. 아울러 혼인상태에 대한 결측값 대체에 있어서 응답자의 거주지에 따라 광역시·도지역의 동지역과 읍·면지역으로 각각 구분하여 거주지 특성에 따라 별개의 CART 모형을 적용하는 경우와 지역별 특성을 고려하지 않고 전국 응답자에 대해 동일한 CART 모형을 적용하는 경우 효율성에 있어 어떤 차이가 있는지 검토해 본다. 한가지 유의할 점은 총조사의 전수조사 항목에는 표본조사에 포함되어 있는 총 자녀수에 대한 조사항목이 없기 때문에 Ryu 등(2001)에서 사용된 대체층과는 다른 형태의 대체층을 사용해야 한다는 점에 유의할 필요가 있다.

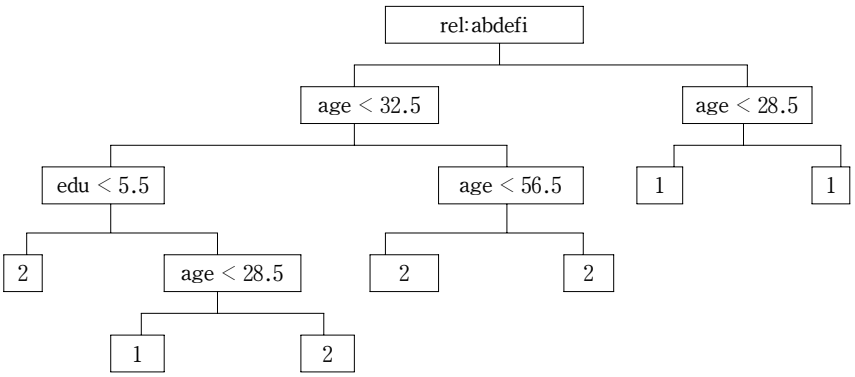
총조사를 위한 시험조사에서 인구부문에서 무응답률이 가장 높았던 15세 이상의 남녀별 혼인상태의 1차 재조사 후 응답률은 <표 1>과 같다. <표 1>을 보면 응답률의 경우 성별에서는 남성보다는 여성의 응답률이, 지역구분에 따라서는 남녀 모두 도의 동지역 응답률이 높다는 것을 알 수 있다.

〈표 1〉 혼인상태에 대한 성별과 지역에 따른 응답률 (단위: 명)

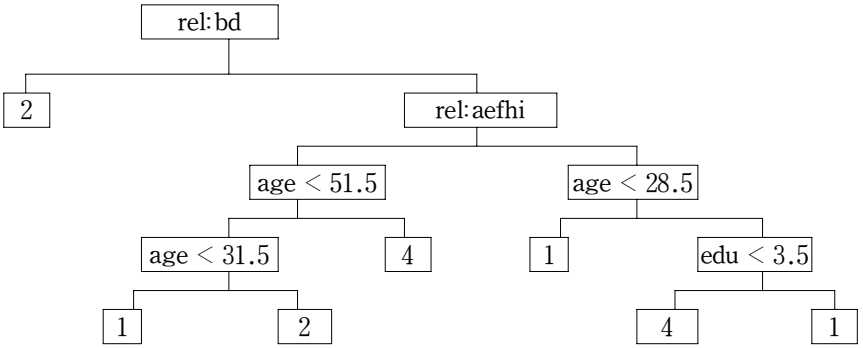
	남 성	여 성
전국	88.1%(58,075/65,891)	88.6%(60,904/68,755)
광역시	87.9%(29,031/33,018)	88.7%(31,069/35,011)
도(동지역)	97.2%(19,680/20,238)	97.6%(20,593/21,092)
도(읍 · 면지역)	74.1%(9,364/12,635)	73.1%(9,242/12,652)

본 연구에서는 우선 1차 재조사 결과로 얻어진 자료 중 혼인상태 항목에서 발생한 결측값에 대해 성별로 별도의 대체방법을 적용하여 완전한 데이터를 얻는다. 대체방법의 정확성을 검토하기 위해 대체과정을 통해 얻은 완전한 데이터와 통계청에서 계획했던 3차에 걸친 재조사를 실제로 완료한 후 얻어진 데이터를 비교하여 대체방법의 효율성을 검토한다. 여기서 혼인상태를 나타내는 값은 범주형(categorical)변수이다. 따라서 CART 모형에서 얻어지는 결과는 분류(classification)형태로 나타나게 된다.

우선 거주지 지역특성을 반영하지 않고 전체 자료를 남성과 여성으로 구분하고, 무응답이 발생하지 않은 완전한 자료에서 적합한 대체모형을 찾기 위해 CART를 적용하여 변수들간의 연관관계를 분석한 결과는 〈그림 1〉 및 〈그림 2〉와 같다.



〈그림 1〉 전국 남성 혼인상태 CART



〈그림 2〉 전국 여성 혼인상태 CART

혼인상태를 위한 CART를 구성해 본 결과 가구주와의 관계(rel), 나이(agel), 교육받은 년 수(edu) 등이 분류변수로 사용되었다. 〈그림 1〉과 〈그림 2〉에서 혼인상태를 나타내는 “1”, “2”, “3”, “4” 는 각각 “미혼”, “배우자 있음”, “이혼”, “사별”을 의미한다. 그리고 마지막 노드(node)는 최대 예측확률(maximum prediction probability)을 갖는 혼인상태 범주를 나타내고 있다. 또한 그림에서 가구주와의 관계(rel)를 나타내기 위해 사용된 기호들은 다음과 같은 관계를 의미한다.

a : 가구주
b : 가구주의 배우자
c : 자녀
d : 자녀의 배우자
e : 가구주의 부모
f : 배우자의 부모
g : 손자녀・그 배우자
h : 증손자녀・그 배우자
i : 조부모
j : 형제자매・그 배우자
k : 형제자매의 자녀・그 배우자
l : 기타 친인척
m : 기타 동거인

참고로 CART 모형그림에서는 각 마디에 주어진 조건(예를 들어, agel < 32.5)이 맞는 경우 왼쪽 가지로, 그렇지 않은 경우 오른쪽 가지로 분류된다는 것을 나타낸다. 예를 들어, 여성의 경우 〈그림 2〉에서 가

구주와의 관계(rel)가 가구주의 배우자(b) 또는 자녀의 배우자(d)인 경우 혼인상태는 2(배우자 있음)로 분류된다는 것을 의미한다(이 경우는 논리적으로 오류가 없는 분류로 연역적 대체에 해당함).

한편 범주형 자료에 대한 CART에서는 마지막 노드(node)의 범주가 같더라도 실제 해당 범주에 할당될 확률을 나타내는 예측확률(prediction probability)은 다르다. 예를 들어 <그림 1>의 마지막 노드에서의 혼인상태 범주별 예측확률은 <표 2>과 같다.

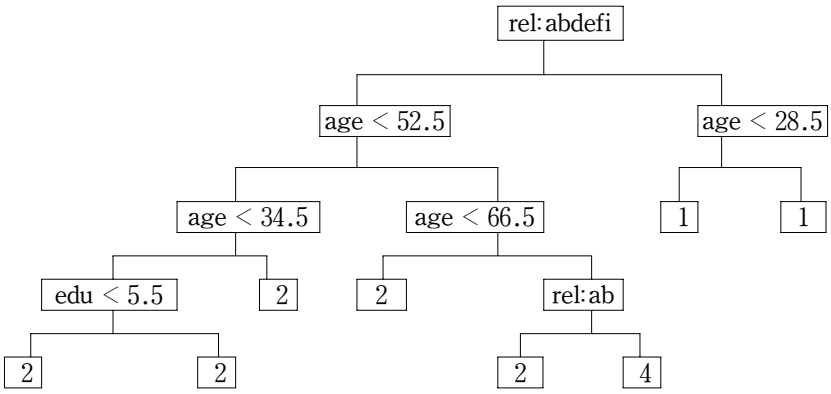
<표 2> 전국 남성 혼인상태의 각 최종 노드 범주별 예측확률

최종 노드 범주	미혼	배우자 있음	이혼	사별
배우자 있음(2)	0.20	0.79	0.01	0.00
미혼(1)	0.65	0.35	0.00	0.00
배우자 있음(2)	0.21	0.79	0.00	0.00
배우자 있음(2)	0.04	0.93	0.02	0.01
배우자 있음(2)	0.00	0.90	0.01	0.08
미혼(1)	0.98	0.02	0.00	0.00
미혼(1)	0.76	0.21	0.03	0.01

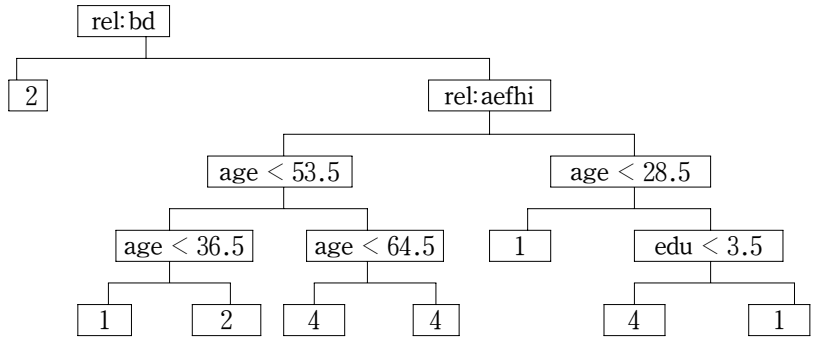
<표 2>를 보면 CART 모형의 마지막 노드의 범주는 최대 예측확률을 갖는 범주를 나타내며, 그림의 최종 노드에서 동일한 것으로 나타나는 경우라도 실제 이 범주로 분류될 예측확률은 노드마다 모두 다르다는 것을 알 수 있다. 여기서 예측확률은 자료에서 최종노드로 분류된 관측값들의 각 혼인상태 범주별 비율에 해당한다.

한편 1차 재조사 자료를 응답자의 거주지 특성에 따라 광역(특별)시 및 도지역의 동지역과 읍·면지역으로 분류하여 각각 성별로 CART 모형을 적용한 결과는 <그림 3>부터 <그림 8>까지와 같다. 여기서 지역특성에 따

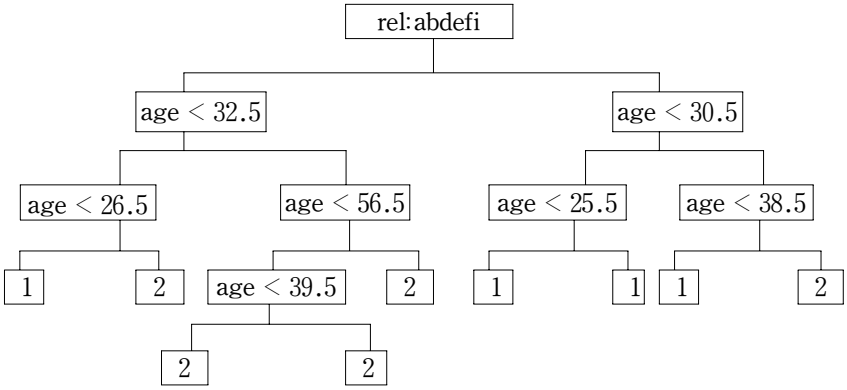
라 구현된 CART 모형들은 사용된 분류변수에 있어서는 차이가 없지만 분류기준에 있어서는 차이를 보이고 있다.



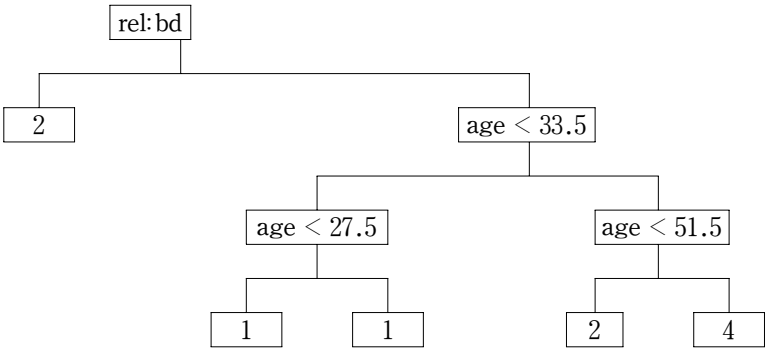
〈그림 3〉 광역시 남성 혼인상태 CART



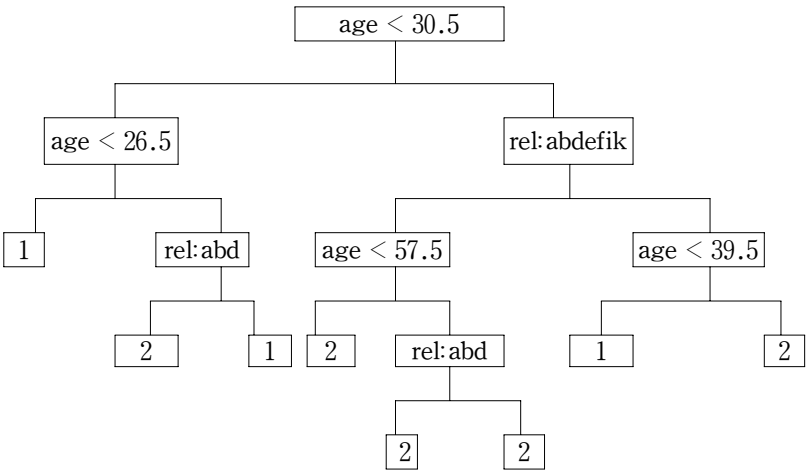
〈그림 4〉 광역시 여성 혼인상태 CART



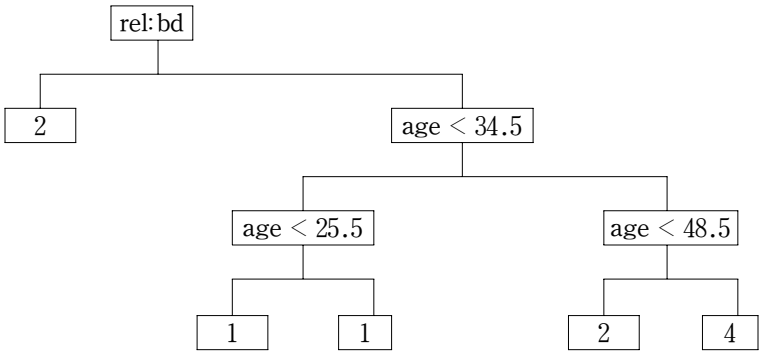
〈그림 5〉 도(동지역) 남성 혼인상태 CART



〈그림 6〉 도(동지역) 여성 혼인상태 CART



〈그림 7〉 도(읍·면지역) 남성 혼인상태 CART



〈그림 8〉 도(읍·면지역) 여성 혼인상태 CART

3. CART 모형을 이용한 결측값 대체방법

본 연구에서는 앞에서 적용된 CART 모형을 활용한 대체방법으로 다음 2가지 방법을 적용하여 그 효율성을 비교한다.

[방법1] CART 모형기반(model based) 대체

혼인상태가 무응답인 경우 무응답자가 CART 모형에 따라 어떤 최종 노드에 속하는지 분류하여 최대 예측확률을 갖는 범주에 속하는 것으로 무응답을 대체

[방법2] CART 대체층을 활용한 순차적 핫덱 대체

CART 모형에서 구성된 최종 노드를 대체층으로 사용하고, 전체 자료를 대체층별로 분류한 후 무응답자의 경우 동일 대체층에 속하는 응답자 중 가장 인접한 위치에 있는 응답자의 관측값으로 대체

여기서 [방법1]의 경우는 실제적으로 결측값을 대체층 내에서 최빈값(mode)으로 대체하는 것에 해당하고, 이는 연속형 변수의 경우 대체층의 평균을 사용하는 평균대체와 개념상 유사하다. 한편 [방법2]의 경우는 대체층 구성을 위해 CART 기법을 사용하고, 대체층 구성 후에는 결측값이 발생하면 대체층 내에서 자료입력 순서 상 가장 인접해 있는 관측값을 대체값으로 사용하는 가장 일반적인 무응답 대체방법인 순차적 핫덱(sequential hot−deck)을 적용한 것이다.

한편 범주형 자료에 대한 또 다른 모형기반 대체방법으로 제시된 [방법1] 대신에 최종 노드에서 각 범주에 대한 예측확률을 그대로 반영하여 대체범주를 확률적으로 정하여 대체하는 방법도 고려해 볼 수 있다. 이를 위해서는 대체과정에서 난수(random number)를 발생시켜 이를 근거로 예측확률에 따라 대체값을 결정하는 추가적인 과정을 수행해야 하기 때문에 그 과정이 상당히 복잡해지게 된다. [방법1]의 경우 일반적으로 평균대체를 적용하는 경우에도 문제가 되는 대체 후 자료에서 분포의 왜곡 현상이 발생하게 된다. 하지만 [방법1]을 변형한 이런 대체방법은 비록 오

분류율은 증가할 수 있으나 [방법1]의 적용에 따른 분포의 왜곡 문제를 해결할 수 있는 하나의 대안이 될 수 있다. 하지만 이런 방법은 적용과정이 매우 복잡하기 때문에 실제 사용 상에 많은 어려움이 있다고 판단되어 연구대상에 포함하지 않았다.

Ⅲ. 대체방법의 효율성 비교

2000년 인구주택총조사에서 실시하는 3번에 걸친 재조사는 조사 상의 오류를 수정하는 동시에 무응답을 줄이기 위해 계획된 것이다. 결과적으로 조사과정에 따라 4개 형태의 자료를 확보하게 되는데 첫 번째는 초기조사에서, 나머지 3번은 각각 순차적인 재조사 및 이에 따른 보완과정을 통해 얻어진 것이다. 이 연구에서는 마지막 2번의 재조사 과정을 적절한 결측값 대체과정을 통해 생략하는 경우 어떤 결과를 얻게 되는지 검토하고자 한다.

이에 따라 첫 번째 재조사 후 발생한 결측자료에 제시된 대체방법을 적용시켜 얻은 대체값과 원래 통계청에서 계획한 대로 3차 재조사 후에 완성된 데이터에서 1차 재조사시 무응답 자료 중 3차 재조사 후 실제 관측된 값을 서로 비교하여 대체방법의 효율성, 다시 말해 대체값의 정확성을 검토한다. 대부분의 대체방법의 효율성분석에 관한 연구에서는 관측된 자료 중 일부를 인위적으로 결측값으로 만든 후 대체방법의 효율성을 검토하고 있다는 것과 비교해 볼 때, 본 연구에서 사용한 시험조사 자료는 실제 재조사가 이루어진 후에 대체값의 정확도가 어느 정도 되는지 정확히 확인을 할 수 있다는 점에서 대체방법의 효율성 검증을 위한 자료로 매우 적합하다고 볼 수 있다.

우선, 지역과 성별에 따른 1차 재조사 자료에서 결측현황과 결측자료 중 3차 재조사를 통해 조사가 완료된 현황을 정리해 보면 〈표 3〉과 같다.

〈표 3〉에서 보는 바와 같이 1차 재조사 후 결측은 성별에 관계없이 도

의 동지역이 가장 낮다는 것을 알 수 있으며, 무응답 중 3차 재조사를 통해 조사가 완료된 경우는 광역시가 성별에 관계없이 가장 높은 것을 알 수 있다. 통계청 관계자의 의견에 따르면 이는 인구주택총조사를 담당하는 지방 조직이었던 읍·면·동에서 주관하던 통계기능을 시·군·구로 이관하는 작업이 추진됨에 따라 읍·면·동을 통한 무응답 오류 자료의 재조사가 제대로 이루어지지 않다고 설명될 수 있다.

〈표 3〉 1차 재조사 후 결측 및 3차 재조사 완료 현황 (단위: 명)

		1차 재조사 결측 현황	3차 재조사 완료 현황
남성	전국	11.9% (7,816/65,891)	37.2% (2,009/7,816)
	광역시	12.1% (3,987/33,018)	47.3% (1,887/3,987)
	도(동지역)	2.8% (558/20,238)	10.2% (57/558)
	도(읍·면지역)	25.9% (3,271/12,635)	2.0% (65/3,271)
여성	전국	11.4% (7,851/68,755)	29.1% (2,284/7,851)
	(광역)시	11.3% (3,942/35,011)	54.7% (2,155/3,942)
	도(동지역)	2.4% (499/21,092)	12.0% (60/499)
	도(읍·면지역)	27.0% (3,410/12,652)	2.0% (69/3,410)

1차 재조사 후 대체방법으로 대체된 값과 3차 조사 후 실제로 관측된 값이 일치하지 않는 경우 대체결과에 오류가 있다고 판단할 수 있다. 따라서 대체방법의 효율성을 비교하기 위해 성별, 지역별 구분에 따라 대체건수에 따른 대체방법별 오분류율을 계산한 결과는 〈표 4〉와 같다. 〈표 4〉에서 전국은 지역구분 없이 하나의 CART 모형을 사용한 경우를 나타내고, 광역시·도(동지역), 도(읍·면지역)은 각각 거주지역 특성에 따라 자료를 분류하여 별도의 CART 모형을 적용하고 각각에 대해 대체방법을 적용한 경우를 나타낸다.

〈표 4〉 지역과 성별에 따른 대체방법별 오분류율

대체방법		[방법 1]	[방법 2]
남성	전국	0.327 (657/2,009)	0.193 (387/2,009)
	광역시	0.325 (614/1,887)	0.218 (411/1,887)
	도(동지역)	0.263 (15/57)	0.263 (15/57)
	도(읍·면지역)	0.292 (19/65)	0.277 (18/65)
여성	전국	0.306 (699/2,284)	0.198 (452/2,284)
	광역시	0.307 (662/2,155)	0.200 (431/2,155)
	도(동지역)	0.267 (16/60)	0.267 (16/60)
	도(읍·면지역)	0.377 (26/69)	0.319 (22/69)

〈표 4〉에서 대체방법의 오분류율을 살펴보면 대체적으로 CART에서 얻어진 최대 예측확률을 기준으로 한 모형기반 대체방법인 [방법1]보다 CART 모형을 이용해 구성한 대체층을 이용하여 순차적 핫덱을 적용한 대체방법인 [방법2]가 효율성 측면에서 우수하다는 결론을 내릴 수 있다. 아울러 지역특성을 반영하는 경우와 지역특성을 고려하지 않는 경우 거의 비슷한 오류율을 보여 주고 있는 것으로 보아 혼인상태의 무응답 대체에 있어서는 굳이 광역시·도의 동 또는 읍·면지역 등과 같은 지역특성에 따라 별도의 CART 모형을 설정할 필요는 없어 보인다.

참고로 순차적 핫덱을 적용하는 [방법2]의 경우, 자료의 입력순서가 지역코드에 따라 수행되기 때문에 대체층 구성에 있어서 지역특성을 반영하지 않아도 대체층 내에서는 자연스럽게 지역적으로 가장 인접한 응답자의 값으로 결측값이 대체되는 특성이 있다는 점에 유의할 필요가 있다.

한편 전국을 대상으로 CART 모형을 설정하고 [방법2]를 사용하는 경우 발생하는 오분류 현황을 좀더 구체적으로 살펴보자. 남성(여성)의 경우 분석대상인 2,009건(2,284건)을 재조사 결과에서 나타난 혼인상태별로 구분하여 오분류 현황을 정리하면 각각 〈표 5〉와 〈표 6〉과 같다.

〈표 5〉 전국 CART 모형에서 [방법2] 남성 혼인상태별 오분류율

재조사 결과 \ 대체 결과	미혼	배우자 있음	이혼	사별	합계	오분류율
미혼	663	119	0	0	782	0.152
배우자 있음	35	880	44	0	959	0.082
이혼	26	59	50	5	140	0.643
사별	10	58	31	29	128	0.773
합계	734	1,116	125	34	2,009	

〈표 6〉 전국 CART 모형에서 [방법2] 여성 혼인상태별 오분류율

재조사 결과 \ 대체 결과	미혼	배우자 있음	이혼	사별	합계	오분류율
미혼	285	31	5	0	321	0.112
배우자 있음	11	905	12	4	932	0.029
이혼	21	66	68	3	158	0.569
사별	14	69	216	574	873	0.343
합계	331	1,071	301	581	2,284	

〈표 5〉와 〈표 6〉에서 비대각칸에 나타난 건수가 오분류 건수에 해당하며, 여기서 재조사 결과 혼인상태별 오분류율(오분류 건수/재조사 건수)은 전체 남성과 여성 모두 재조사 결과에서 혼인상태가 “3”(이혼) 또는 “4”(사별)인 경우에 “1”(미혼) 또는 “2”(배우자 있음)에 비해 상대적으로 오분류율이 매우 높게 나타나고 있다는 것을 볼 수 있다. 이런 결과는 비록 대체층을 구성하더라도 핫덱 과정에서 결측값에 대한 대체값으로 사용하게 되는 전체 응답결과 중 다른 경우에 비해 혼인상태가 “1” 또는 “2”인 경우가 월등히 많기 때문에 불가피하게 발생하는 현상으로 해석할 수 있다.

Ⅳ. 결론 및 제언

본 연구에서는 총조사에서 인구부문의 혼인상태의 결측값을 대체할 수 있는 두 가지 방법을 고려하고 있다. 첫 번째 방법은 CART 모형에서 얻어진 최대 예측확률을 기준으로 결측값을 대체하는 일종의 모형기반 접근법이고, 두 번째 방법은 CART 모형에서 얻어진 결과를 근거로 대체층을 구성하여 핫덱 방법을 적용하는 대체방법이다.

실제 재조사 결과를 이용하여 오분류율을 검토해 본 결과 두 방법 중 CART 모형을 기반으로 대체층을 구성하여 핫덱 방법을 적용하는 방법이 효율적이라는 결론을 얻을 수 있었다. 아울러 전국에 대해 동일한 모형을 설정한 경우와 거주지 특성에 따라 광역시·도의 동지역, 도의 읍·면지역으로 구분하여 대체방법을 적용하는 경우를 비교해 본 결과 남녀 모두 굳이 지역을 구분하여 대체방법을 사용하는 것이 필요하지 않다는 사실을 알 수 있다. 특히 지역별 특성을 반영하는 것이 효과가 없는 이유는 대체과정에서 자연스럽게 인접 가구의 관측값으로 결측값이 대체되는 순차적 핫덱의 특성으로 설명될 수 있다.

참고로 본 연구결과에서 나타난 오분류율은 Ryu 등(2001)의 연구결과와 비교해 보면 특히 여성의 경우 약간 크게 나타나고 있다. 그 이유는 Ryu 등(2001)의 연구에서는 총조사의 시험조사 자료 중 표본조사에 해당하는 제한적인 일부 자료만을 활용하고 있기 때문에 여성의 경우 혼인상태를 잘 설명할 수 있는 “총 자녀의 수”라는 조사항목을 추가적으로 사용하고 있다는 것으로 설명할 수 있다. 이 조사항목은 본 연구의 대상인 총조사의 전수조사의 경우에는 조사항목에 포함되어 있지 않아 전수조사 자료에 대한 대체방법에서는 이 변수를 사용할 수 없다는 한계가 있다.

아울러 본 연구에서 적용한 모형기반 예측방법에 있어서는 각 범주에 할당된 확률자체는 무시하고 가장 예측확률이 큰 범주로 결측값을 대체하고 있다. 이에 따라 대체 후의 자료는 범주별 분포가 일부 왜곡되어 차후

의 자료분석에 영향을 줄 수 있다. 비록 오류율은 약간 증가할 수 있지만 이런 분포의 왜곡에 따른 문제점을 해결하기 위해서는 최종 노드에서 모형에서 얻어진 범주별 예측확률에 따라 확률적으로 결측값을 대체하는 모형기반 대체방법을 적용하는 것이 효과적이라고 판단된다.

본 연구에서 제시된 방법을 일반적인 사회조사에서 실제 적용하는 데 있어서 참고가 될 만한 몇 가지 사항을 정리하면 다음과 같다.

첫째, CART 모형에 나타난 분류변수에 결측이 발생하는 경우 대체방법 적용에 문제가 발생하게 된다. 이런 경우 일반적인 해결방안은 우선 대체층 구성에 필요한 변수를 적절한 방법으로 대체한 후 관심변수를 대체하는 과정을 거치게 된다. 특히 사회조사의 경우 많은 관심변수를 동시에 고려해야 하기 때문에 이런 경우 주요 변수 중 무응답이 적은 변수부터 순차적으로 대체해 나가는 것이 하나의 방안이 될 수 있다. 예를 들어 제시된 모형에서 교육년수와 혼인상태가 동시에 무응답인 경우 우선 나이를 기준으로 교육년수를 대체하고 순차적으로 혼인상태를 대체하는 방안을 고려할 수 있다.

둘째, 결측값 대체가 이루어진 후에는 항상 대체된 값이 다른 변수과 연관관계에 있어서 논리적으로 문제가 없는지 반드시 확인하는 에디팅 과정을 수행해야 한다.

셋째, 결측값을 대체하여 완전한 자료를 구성한 후에는 가중값을 적용한 추정값 산출 등에 있어 기존의 일반적인 통계분석 기법을 그대로 적용할 수 있다는 이점이 있다. 하지만 이 경우 계산되는 추정오차는 과소 추정되는 경향이 있다는 점에 유의할 필요가 있다. 이와 관련된 분산추정 방법에 대한 내용은 Rao와 Shao(1992) 등을 참고하기 바란다.

현재 우리나라의 경우 다양한 분야에서 많은 사회조사가 이루어지고 있지만 체계적인 결측값 대체방법을 적용하고 있는 사례는 많지 않다. 특히 각종 연구소 및 조사회사 등에서 수행되는 각종 사회조사에서는 주요 관심 항목에 무응답이 발생하는 경우 이를 자료에서 제외하는 것이 일반적인 경향이다. 그 이유는 우리나라 조사연구 관계자들이 무응답 대체는

손쉽게 수행할 수 없는 복잡한 과제라는 인식을 갖고 있는 것이 가장 큰 원인이라고 생각된다. 어떤 통계조사에 있어서나 무조건 결측값을 무시하고 분석하는 것보다는 약간의 신뢰성을 향상시키기 위해서라도 결측값을 줄이는 노력을 최대한 기울이는 것이 필요하다.

본 연구에 제시된 방법을 살펴보면, 어느 정도 조사자료 분석경험을 가진 연구자들은 큰 무리 없이 제시된 방법과 유사한 과정을 거쳐 얼마든지 효과적인 무응답 대체과정을 수행할 수 있다는 것을 알 수 있다. 따라서 본 연구결과를 참고로 하여 사회조사 분야에 종사하는 많은 조사연구자들이 결측값 대체방법의 적용이 큰 어려움 없이 수행될 수 있다는 점을 인식하기 바라고, 동시에 향후 다양한 분야의 사회조사에서 결측값을 적절하게 처리한 좀더 양질의 통계자료들이 생산될 수 있기를 기대해 본다.

참고문헌

김영원 · 조선경. 1996. “표본조사에서 항목 무응답 대체방법.” 《한국통계학회논문집》 3(3) : 145－159.

이재원. 2000. “무응답 및 오류자료의 Imputation 적용 결과.” 《무응답오차》 : 131－145. 조사통계연구회. 자유아카데미.

Breiman, L., J. H. Friedman, R. A. Olshen and C. J. Stone. 1984. *Classification and Regression Trees*. Chapman & Hall.

Kalton, G. and D. Kasprzyk. 1986. “The Treatment of Missing Survey Data.” *Survey Methodology* 12 : 1－16.

Lessler, J. T. and W. D. Kalsbeek. 1992. *Nonsampling Error in Surveys*. New York, Wiley.

Rao, J. N. K. and J. Shao. 1992. “Jackknife Variance Estimation with Survey Data under Hot Deck Imputation.” *Biometrika* 79 : 811－822.

Ryu, Jae－Bok, Young－Won Kim, Jin－Woo Park and Jae－Won

Lee. 2001. “Imputation Methods for the Population and Housing Census 2000 in Korea.” *Bulletin of the International Statistical Institute*, 53rd Session of ISI, August, 2001, Seoul, Korea : 421−422.

Sonquist, J. A., E. L. Baker and J. N. Morgan. 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.