제1장

인구주택총조사 무응답 대체기법(II)
-2005년 인구주택총조사 자료를 대상으로-

이현정 · 최필근

제1절 서 론

1. 연구배경

개인정보 및 사생활 보호의식이 강화되는 등 조사환경이 악화됨에 따라 조사를 완벽하게 한다는 것이 점차 불가능하게 되었다. 그리고 여러 가지 예방책에도 불구하고 조사과정 중의 무응답은 거의 모든 통계조사에서 필연적으로 발생하게 된다(Lessler & Kalsbeek, 1992). 이러한 무응답 항목이 증가함에 따라 체계적인 무응답 처리의 필요성이 증대되었다.

무응답이란 조사표에 기입할 조사사항을 기입하지 않은 것을 말하며 응답자가 조사항목의 응답을 거부한 것과 조사자가 응답자를 조사기간에 전혀 만나지 못하여 조사표 전체를 공백으로 한 것을 포함한다. 인구주택총조사와 같은 면접조사, 인터넷 조사 및 응답자가 직접 기입하는 조사방법이 혼합형태를 이루고 있는 조사에서는 다양한 무응답 형태가 발생될 것으로 예상된다(송순관, 2005).

무응답은 조사 단위 자체가 응답을 하지 않는 단위무응답(unit non-response)과 일부 항목에 대해서만 응답을 하지 않거나 질문과 무관한 응답을 해서 불필요한 자료가 되는 항목무응답(item non-response)으로 나누어질 수 있다(Kalton & Kasprzyk, 1986). 단위무응답은 재조사나

가중값 조정방법을 통하여 처리하고 항목 무응답의 경우는 무응답 대체법(imputation)을 통하여 처리하는 것이 일반적이다.

대체법이란 조사 후에 무응답으로 인해서 발생한 결측값을 채워 넣는 방법을 의미한다. 결측값을 대체하는 이유는 결측값을 가능한 정확한 값으로 대체함으로써 무응답에 의한 문제를 줄이고, 조사결과의 신뢰성을 높이기 위해서라고 할 수 있다. 대체를 이용하면 무응답에 대한 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(김영원・이주원, 2003).

특히 인구주택총조사와 같은 대규모 조사의 경우, 전수 및 표본조사로 이루어지는 대규모 조사이기 때문에 조사불응 및 부재, 일부 항목의 누락 등 무응답이 발생할 가능성이 높다고 할 수 있다. 그래서 자료처리 시간을 단축하기 위해서는 전체 총조사 과정에 있어서 상당히 오랜 기간이 소요되었던 재조사 및 자료편집과정을 효율적으로 수행하는 것이 필요하다(이재원, 2000). 이에 따라 2000년 총조사에서는 1995년까지의 총조사에서는 적용하지 않았던 무응답 대체방법을 도입하는 것이 필요하게 되었다.

인구주택총조사의 자료처리는 통계조사 결과의 품질에 많은 영향을 미친다. 인구통계가 만들어질 때 사용하는 많은 기법이 이 과정에서 개발되고 있다. 경험과 기법이 동시에 중요시되므로 개인의 능력을 바탕으로 이루어지는 것보다 조직이 기능을 수행하는 방법으로 전환되어야 안정적인 기반이 구축된다 하겠다. 자료처리의 단계 중 결과집계 및 수준분석 단계에서 무응답 처리가 이루어진다. 오류내용을 제거하는 과정에서 무응답으로 인한 오류는 경험과 통계적 기법을 접목하여 추정하는 기법으로 해결해야 한다. 이것은 이론적 배경과 관련된 검토보고서가 필요하다. 항목간의 상관관계가 높을 것이라는 가정보다는 증거자료가 제시될 필요가 있다(2005 인구주택총조사 종합평가 보고서).

2. 연구목적 및 내용

본 연구는 지난 2008년에 진행되었던 『인구주택총조사 무응답 처리기법 연구(Ⅰ)』의 후속연구로서 가구원, 가구 및 주택에 관한 항목에 가장 적절한 대체군 및 대체방법을 제시하고자 함에 그 목적이 있다. 제시된 무응답 대체기법을 사용하여 여러 가지 원인으로 발생한 무응답들을 완전한 자료로 구성하고자 하는 것이다.

선행연구로는 이재원(2000)이 2000년 인구주택총조사를 위한 1999년 11월에 실시한 시험조사 자료를 이용하여 출생지, 5년 전 거주지, 혼인상태의 3개 항목에 무응답 대체기법을 적용하였고, 김영원과 조선경(2003)은 2000년 인구주택총조사를 위한 1999년 11월에 실시한 시험조사 자료를 이용하여 혼인상태 1개 항목에 적용하였다. 송순관(2005)은 2005년 인구주택총조사를 위한 5차 시험조사 자료를 이용하여 가구원, 가구, 주택과 관련한 주요한 항목에 적용하였으며, 이현정(2009)은 2010년 인구주택총조사를 위한 2차 시험조사 자료를 이용하여 가구원, 가구, 주택과 관련한 대부분의 항목에 적용하였다.

이번 연구에서는 2005년에 실시한 인구주택총조사 자료의 무응답을 처리할 수 있는 무응답 대체기법에 대해 살펴보고, 모의실험으로 무응답 처리를 실시한 후 그 효율성을 분석해 보고자 한다. 2절에서는 제2차 시험조사 자료를 이용하여 연구한 무응답 대체기법에 대한 문제점을 알아보고 본 연구의 기본방향을 검토해 보고자 한다. 3절에서는 카이제곱 검정의 연관성 분석 및 CHAID 알고리즘에 의한 대체군을 개발한다. 4절에서는 각각의 대체군을 이용하여 가구원, 가구 및 주택에 대한 항목 무응답 대체를 실시하고 그 효율성을 비교해 보고자 한다. 5절에서는 앞 절에서 모의실험한 무응답 처리결과에 대해 무응답 처리방법을 제안하고, 향후 연구방향을 제시하고자 한다.

제2절 선행연구 검토 및 연구 기본방향

1. 선행연구[1]에 대한 검토

가. 이용자료

이현정(2009)은 제2차 시험조사 자료를 이용하여 인구주택총조사 무응답 대체기법에 대하여 연구하였다. 조사대상은 2008년 4월 14일 0시 현재 조사지역내에 거주하는 모든 사람과 이들이 살고 있는 주택이었다. 조사규모는 505조사구에 30,025가구이며 조사대상 지역은 부산, 경기도, 강원도 6개 시군구의 6개 읍면동이었다. 조사항목은 총 44개 항목을 조사하였다.

〈표 1-1〉 제2차 시험조사 조사항목

구분	전수 개 항목	표본 개 항목
가구원 개 (24)	(18) ①성명 ②성별 ③나이 ④가구주와의 관계 ⑤교육정도 ⑥혼인상태 ⑦국적 및 입국 연도	①아동보육 ②출생년 (26 ③ 년)전 거주지 ④ 년 전 거주지 ⑤활동제약 ⑥통근·통학 여부 ⑦통근·통학 장소 ⑧이용교통수단 ⑨통근·통학 소요시간 ⑩경제활동상태 ⑪종사상 지위 ⑫산업 ⑬직업 ⑭근로장소 ⑮혼인년월 ⑯총 출생아수 ⑰고령자 생활비 원천
가구 개 (14)	①가구 구분 ②사용 방 수 ③주거시설 형태 ④건물 및 거주층 ⑤점유 형태 ⑥타지 주택 소유 및 주인가구 여부	①거주 기간 ②난방시설 ③식수 사용 형태 ④정보통신기기 보유현황 ⑤자동차 보유 대수 ⑥주차시설 ⑦임차료 ⑧가구소득
주택 개	①거처의 종류 ②주거용 연면적 ③총방수 ④건축 연도 ⑤편익시설수	①대지면적

자료(6제)차 시험조사 조사지침서

주 년 대비 신규항목은 밑줄로 표시된 개 항목이며 난방시설과 대지면적은 조사 표 종류를 변경 전수→표본

: 2005 6 ( )

1) 이현정(2009), 『인구주택총조사 무응답 처리기법 연구(Ⅰ)』에 대한 내용을 검토해 보고자 함

나. 무응답 대체를 위한 자료처리

선행연구는 제2차 시험조사를 실시한 후 1차 내검을 거친 자료를 이용하여 항목간 연관성 분석을 실시하고, 가장 적절한 대체군을 개발하여 그 결측치에 대한 가장 효율적인 무응답 대체방법을 찾아보고자 한 것이었다.

항목무응답 대체는 기존에 존재하는 자료 중 가장 유사한 자료로 그 항목을 대체하는 것이므로 전수조사 항목의 무응답은 전수조사 자료에서, 표본조사 항목의 무응답은 표본조사 자료에서 응답을 제공받기 위해서 전수조사 자료와 표본조사 자료를 분리하였다.

개별 항목의 무응답 대체를 하기 이전에 사전 자동내검 단계를 처리함으로써 실제적인 무응답 대체를 위한 원시입력자료를 작성하도록 했다. 조사를 실시하고 난 뒤 입력된 자료는 무응답과 '해당없음'에 해당하는 부분이 모두 공백으로 제공된다. 실제 무응답 된 자료를 대체하기 위해서는 '해당없음'으로 인한 공백과 무응답으로 인한 공백의 구분이 필요하다. '해당없음'에 해당하여 공백인 경우는 '999(일부는 99999)'로 입력하고 실제적으로 무응답인 것은 공백으로 둔다. 그리고 상식적인 값의 범위를 벗어나는 이상치의 경우는 공백으로 처리하여 대체하도록 한다.

다. 대체군(보조변수)의 개발

대체군 개발을 위한 알고리즘은 여러 가지가 있다. Breiman 등(1984)이 제안한 CART(Classification and Regression Tree)알고리즘, Kass(1980)가 제안한 CHAID(Chi-squared Automatic Interaction Detection) 등이 있고, 변수간의 연관성을 분석할 수 있는 카이제곱 독립성 검정(test of independence)도 있다.

선행연구에서는 대체군을 개발하는 데 있어서 범주형 자료분석의 한 방법인 카이제곱 독립성 검정을 실시하였다. 두 변수간의 독립성을 검정하고 연관성 측도를 이용하여 두 변수간의 연관성을 찾아내어 대체군

을 개발하는 것이다. 독립성 검정은 자료에 포함된 두 가지 특성 사이에 어떠한 연관관계가 있는지를 검정하는 방법으로 독립성 검정에서의 귀무가설은 "두 가지 특성들이 서로 독립(independent)이다"라는 것이다. 즉, 두 특성 사이에 연관관계가 없다는 것이다.

라. 항목무응답 대체방법의 효율성 비교

대체방법의 효율성 비교를 위해서는 2005년 인구주택총조사에서 사용한 기존의 항목대체방법과 주로 Hierarchical Hot-Deck방법을 이용하여 그 효율성을 비교해 보았다. 송순관(2005)은 Hierarchical Hot-Deck방법을 Hot-Deck방법과 Nearest Neighbor방법의 개념을 통합하여 응용된 방법으로 보았다. 한 카테고리에 존재하는 무응답자수에 비하여 응답자수가 적을 때 만약 응답자 선택을 비복원(without replacement)으로 수행한다면 모든 무응답자를 처리하기 위해서는 카테고리와 카테고리의 통합과정이 필요하게 된다. Hierarchical Hot-Deck방법은 이러한 카테고리를 적절히 통합하여 하나의 새로운 대체군(보조변수)으로 생성함으로써 자동적으로 무응답처리를 할 수 있도록 고안된 일종의 Hot-Deck방법이다.

마. 무응답 대체항목의 선정방법

인구주택총조사의 조사항목은 가구원에 관한 항목, 가구에 관한 항목, 주택에 관한 항목으로 구분할 수 있다. 무응답 처리대상 항목은 원칙적으로 모든 항목을 무응답 처리하는 것으로 간주한다. 성명, 산업 및 직업과 같이 주관식으로 응답하는 항목, 활동제약과 같이 왜곡될 가능성이 있는 항목, 국적 및 입국 연도와 같이 대체군 개발이 모호한 항목, 그리고 정확도 등을 검토하여 정확도가 낮은 항목 등은 무응답 대체항목에서 제외하도록 하였다.

바. 연구결과

선행연구에서는 2008년에 실시한 제2차 시험조사 자료에 대하여 무응답을 대체할 수 있는 세 가지 방법을 고려하고 있다. [방법1]은 2005년 인구주택총조사 자료 무응답 처리시 사용한 방법이고, [방법2]는 2008년 대체군과 2005년 항목대체방법을 적용한 것이며, [방법3]은 2008년 대체군과 2008년 항목대체방법을 적용한 것이다.

실제 결측치를 발생시킨 후 실시한 모의실험 결과 대체군은 카이제곱 독립성 검정을 통하여 얻어진 새로운 대체군을 사용하는 것이 효율적이라는 결론을 얻을 수 있었다. 또한 항목 대체방법을 Probability방법이나 Hot-deck방법을 변경하여 Hierarchical Hot-deck방법으로 적용하는 것이 꼭 필요하다는 것은 아니라는 사실을 알 수 있었다. 그러나 Hierarchical Hot-deck방법을 이용하면 도너(donor)수가 적어서 무응답 처리를 해도 결측값이 발생하는 일은 막을 수 있다는 장점이 있다. 대체의 정확도를 향상시키기 위해서는 항목 대체방법의 변경보다 가장 적절한 대체군을 발견하여 적용하는 것이 훨씬 더 큰 영향력이 있다는 결론을 얻을 수 있다.

선행연구 결과의 정확도는 주로 70% 이상으로 나타났다. 그러나 교육정도, 거주 기간, 가구소득 등은 기존의 조사문항으로는 대체군을 형성하는 데는 한계가 있다는 것을 보여준다. 이 항목들을 대체함에 있어서 정확도를 높이기 위해서는 연관성이 높은 새로운 문항의 조사가 필요할 것으로 판단된다. 무응답 대체가 이루어진 후에는 대체된 값이 다른 변수와의 연관관계에 있어서 논리적으로 문제가 없는지 반드시 확인하는 사후내검 단계를 수행해야 한다. 자세한 결과는 선행연구(이현정, 2009)를 참조하기 바란다.

2. 연구의 기본방향

가. 이용자료

본 연구에서는 2005년에 실시한 인구주택총조사 자료를 이용하고자 한다. 인구주택총조사는 우리나라의 인구・가구・주택에 대해 규모・구조・분포 및 특성을 파악하기 위한 조사이다. 2005년에 실시한 인구주택총조사는 조사원면접, 응답자 직접기입, 인터넷조사방법을 병행하여 실시하였고 조사항목은 전수 21개 항목과 표본 20개 항목으로 총 41개의 항목이었다. 그리고 시도별 특성항목 3개를 추가함으로써 총 44개의 항목을 조사하였다. 상세한 조사항목은 <표 1-2>와 같다. 이 조사의 결과로 47,278,951명의 인구수와 15,988,274 가구수, 그리고 13,222,641 주택수를 얻었다. 이 결과에서도 알 수 있듯이 이 인구주택총조사는 전수 및 표본조사로 이루어지는 대규모 조사이기 때문에 조사불응 및 부재, 일부 항목의 누락 등 무응답이 발생할 가능성도 높다고 할 수 있다.

〈표 1-2〉 2005 인구주택총조사 조사항목

구분	전수 개 항목	표본 개 항목
	성명 (21)	아동보육 (20)
가구원 개 (24)	(1)성별 (2)나이 (3)가구주와의 관계 (4)교육정도 (5)종교 (6)남북이산가족 (7)혼인상태 (8)	(1)년 전 거주지 (2)활동제약 (3)통근 통학 여부 (4)통근 통.학 장소 (5)이용교통수단 (6)통근 통학 소요시간 (7)경제활동상태 (8)종사상의 지위 (9)산업 (10)직업 (11)근로장소 (12)혼인년월 (13)

		총 출생아수 (14) 추가 계획 자녀 수 (15) 고령자 생활비 원천 (16)
가구 개 (11)	가구 구분 (1) 사용 방 수 (2) 주거 시설 형태 (3) 거주층 (4) 난방 시설 (5) 점유 형태 (6) 주인가구 및 주택 소유 여부 (7)	거주 기간 (1) 자동차 보유 대수 (2) 주차시설 (3) 임차료 (4)
주택 개 (6)	거처의 종류 및 건물층수 (1) 연건평 (2) 대지 면적 (3) 총 방 수 (4) 건축년도 (5) 편익 시설 수 (6)	
시도별 특성항목 시도별 (개 3)	자원봉사활동 (1) 자녀 출산시기 (2) 노후 준비방법 (3) 간호 수발자 (4) 치매 중풍시설 입소여부 (5) 현 시 도 거주사유 (6) 다른 시 도 이동사유 (7) 지역생활여건 만족도 (8) 식수사용형태 (9) 컴퓨터 보유대수 및 인터넷 사용여부 (10) 가구생활비 원천 (11) 최초 주택마련 시기 및 방법 (12)	

자료 인구주택총조사 조사지침서

: 2005

나. 연구의 기본방향

무응답 처리를 하기 위한 자료처리 단계에서는 개별 항목의 무응답 대체를 하기 이전에 사전 자동내검 단계를 처리함으로써 실제적인 무응답 대체를 위한 원시입력자료를 작성하도록 한다. 실제 무응답 된 자료를 대체하기 위한 '해당없음'으로 인한 공백과 무응답으로 인한 공백의 구분이 필요하기 때문이다. 또한 내검단계에서 처리할 수 있는 무응답은 미리 처리함으로써 imputation 프로그램을 이용하여 처리해야 할 무응답을 최소로 한다. 무응답 처리 순서는 가구원, 가구, 주택 순으로 처리한다. 전수자료와 표본자료는 분리하여 수행한다. 이번 연구에서는 외국인 조사표의 항목도 같이 대체하도록 한다.

대체군 개발방법은 여러 가지가 있지만, 본 연구에서는 카이제곱 검정의 연관성 분석에 의한 방법과 CHAID(Chi-squared Automatic Interaction Detection) 알고리즘에 의해 대체군을 개발하고자 한다. 선행연구에서는 전수조사 항목과 표본조사 항목이 동일한 경우 동일한 대체군을 이용하였으나, 이번 연구에서는 각각의 대체군을 별도로 개발하고자 한다.

항목무응답 대체방법의 효율성을 비교하기 위한 대체방법으로는 Hierarchical Hot-Deck방법을 이용하도록 한다. 선행연구 결과에서 알 수 있듯이 무응답 대체의 효율성은 대체군을 변경하는 것이 대체방법을 변경하는 것보다 더 높기 때문에 대체방법에 있어서는 단순화하고자 한다.

인구주택총조사의 조사항목은 가구원에 관한 항목, 가구에 관한 항목, 주택에 관한 항목으로 구분할 수 있다. 무응답 처리대상 항목은 성명, 산업 및 직업과 같이 주관식으로 응답하는 항목을 제외한 모든 항목을 무응답 처리하는 것으로 간주하여 무응답 대체를 실시하고자 한다.

제3절 대체군 개발

이 절에서는 인구주택총조사의 각 항목에 대하여 대체군을 개발하는 방법을 설명하고자 한다. 항목에 대한 연관성 분석방법으로 카이제곱 검정방법과 의사결정나무의 분리 알고리즘 중의 하나인 CHAID(Chi-squared Automatic Interaction Detection) 알고리즘을 이용한다. 두 방법에 의하여 개발된 대체군을 적용하여 적절한 모의실험을 실시한 후 가장 적합한 최종 대체군을 결정할 것이다.

1. 카이제곱 검정의 연관성 분석에 의한 대체군 개발

범주형 자료분석(categorical data analysis)의 대상이 되는 자료는 대부분 설문조사 결과 자료로서 도수분포표(frequency table) 또는 교차표(contingency table)의 형태로 먼저 정리되고, 교차표에서 분류 기준이 되는 변수에 따라 관측값들이 어떻게 분포하고 있는지를 판단하기 위한 분석방법으로 적합도 검정이나, 독립성 검정, 동질성 검정 방법이 사용된다. 이러한 분석방법들은 모두 카이제곱 분포를 따르는 검정통계량을 이용한다.

독립성 검정은 자료에 포함된 두 가지 특성 사이에 어떠한 연관관계가 있는지를 검정하는 방법으로 독립성 검정에서의 귀무가설은 “두 가지 특성들이 서로 독립(independent)이다”라는 것이다. 즉, 두 특성 사이에 연관관계가 없다는 것이다. 독립성 검정을 위해서 카이제곱 검정이 사용된다. 또한 분석결과 두 특성이 독립이 아닌 경우 어느 정도의 연관성을 가지고 있는지를 분석하기 위해서는 연관성 측도가 필요하다.

연관성 측도는 독립성 검정 또는 동질성 검정에서 설정된 귀무가설이 기각되는 경우, 즉, 두 변수들이 서로 독립이 아니고 연관성을 가지고 있다면, 어느 정도의 연관성을 가지고 있는지를 판단하기 위한 측도이다. 이 값이 클수록 두 범주형 변수 사이의 연관성이 크다는 것을 의미하지만, 실제로 연관성의 정도를 판단하기 위해서는 여러 가지 연관성 측도들이 이용될 수 있다. 이러한 연관성 측도는 변수가 순서형

(ordinal)인 경우와 명목형(nominal)인 경우에 따라 사용하는 측도가 다르다. 각 연관성 측도들은 자료의 형태에 따라 사용되는 경우가 다르므로 자료의 특성에 맞는 연관성 측도를 이용하여 두 변수 사이의 관계를 분석해야 한다.

가. 성별 항목에 대한 연관성 분석

카이제곱 검정을 실시한 1개의 항목에 대해 항목간 연관성 분석한 결과를 이용하여 'P값이 <0.0001'이고, '크래머 V값'이 큰 변수 중 7 ~ 8개 정도의 항목을 선택하여 대체군으로 선정하도록 한다.

성별 항목 이외의 모든 항목들도 같은 방식으로 설명되므로 성별 항목에 대해서만 자세하게 설명하고 나머지 항목들은 분석에 의한 결과만을 제시할 것이다.

카이제곱 검정 결과 성별과 연관성이 높은 항목은 가구주와의 관계, 가구원번호, 혼인상태, 교육정도이며 가장 연관성이 높은 항목은 가구주와의 관계로 나타났다.

〈표 1-3〉 성별에 대한 연관성 분석 결과

보조변수	카이제곱 값	유의확률	크래머의 V
가구주와의 관계			
가구원번호	1360557	<.0001	0.52981
혼인상태	1062455	<.0001	0.46819
교육정도	159999	<.0001	0.18169
	64053.5082	<.0001	0.115

그러나 인구주택총조사 자료는 경험적으로 시도, 시군구 등과 같은 행정구역에 밀접한 관련이 있는 것을 알기 때문에 시도, 시군구, 읍면동 코드로 결합된 행정구역코드를 모든 항목의 대체군으로 포함하도록 한다.

나. 각 항목별 연관성 분석결과 요약

1) 가구원에 관한 사항

〈표 1-4〉 가구원 항목들에 대한 대체군 요약(전수)

조사항목		대체군 카이제곱 ( )
성명		
성별		행정구역코드 시도 시군구 읍면동 가구원번호 가구주와의관계 (혼인상태 교육정도), , 행정구역코드 시도 시군구 읍면동 가구원번호
나이	만나이	( ) 혼인상태 교육정도 가구주와의관계 가구원수 , 종교여부 , , , , 행정구역코드 시도 시군구 읍면동 가구원번호
가구주와의 관계		( ) 성별 혼인상태 (, 만나이 교육상태 , 행정구역코드 시도 age2_group 읍면동 가구원번호
교육정도	교육정도	교육상태 혼인상태 , , 만나이) , 가구주와의관계 성별 age2_group, , 행정구역코드 시도, 시군구 읍면동 가구원번호
	교육상태	교육정도 (, 만나이, 혼인상태) , 가구주와의관계 age2_group, 가구원수 , , 행정구역코드 시도, 시군구 읍면동 가구번호
종교	종교여부	종교종류 (, 만나이, 혼인상태) , 가구주와의 관계 age2_group, , , 행정구역코드 시도 시군구 읍면동 종교여부 성별
	종교종류	혼인상태 (, ,), , , 행정구역코드 시도 시군구 읍면동
남북이산가족	남북이산가족 여부	남북이산가족출생지 (, , 만나이), 종교여부 , age2_group, ,
	남북이산가족 출생지	행정구역코드 시도 시군구 읍면동 남북이산가족여부 (, , 만나이), 혼인상태 행정구역코드 시도 시군구 age2_group 읍면동 가구원번호
	혼인상태	(, 만나이 가구주와의관계), , 교육상태 교육정도 age2_group, ,

주 : age2_group 은 만나이를 5 세 연령별로 그룹화한 변수임

〈표 1-4〉 가구원 항목들에 대한 대체군 요약(전수)_계속

조사항목	대체군 가이제곱 ( )
외국인여부	행정구역코드 시도 시군구 읍면동 외국인국적 외국인교육정도(외국인주된체류목적), 외국인직업 , ,
외국인 교육정도	행정구역코드 시도 시군구 읍면동 외국인직업 외국인주된체류목적 , 외국인거주기간 ( ), 혼인상태 , ,
외국인 국적	행정구역코드 시도 시군구 읍면동 외국인교육정도(외국인주된체류목적), 외국인직업 종교종류 교육정도 성별 조사구특성 , , , , ,
외국인 주된 체류목적	행정구역코드 시도 시군구 읍면동 외국인교육정도(외국인거주기간 외국인직업 혼인상태 , , ,
외국인 직업	행정구역코드 시도 시군구 읍면동 외국인교육정도(외국인주된체류목적), 외국인거주기간 혼인상태 ,
외국인 거주기간	행정구역코드 시도 시군구 읍면동 외국인주된체류목적 , 외국인교육정도), 외국인직업 혼인상태 , , ,
외국인 거주기간 년 월	외국인거주기간과 함께 대체함

주 : 대체군의 나열순서는 중요한 항목부터 나열된 것임

〈표 1-5〉 가구원 항목들에 대한 대체군 요약(표본)

조사항목		대체군 카이제곱 ( )
성명		-
성별		행정구역코드 시도 시군구 읍면동 가구원번호 ( ) 추가계획자녀여부 남자출생아수 여자출생아수 가구주와의관계 , , ,
나이	만나이	행정구역코드 시도 시군구 읍면동 가구원번호 ( ) 통근통학여부 혼인상태 어머니동거여부 , 혼인년도 , , maryy_group,
가구주와의관계		행정구역코드 시도 시군구 읍면동 가구원번호 ( ) 추가계획자녀여부 성별 혼안상태 통근통학여부 ,
교육정도	교육정도	행정구역코드 시도 시군구 읍면동 가구원번호 활동제약여부 (활동제약종류, 교육상태) , 어머니동거여부6,통근통학여부 5, ,
	교육상태	행정구역코드 시도, 시군구 읍면동 가구원번호 통근통학여부 활동제약여부 , 교육정도) , 활동제약종류 , 만나이 0, 행정구역코드 시도, 시군구 읍면동 종교종류
종교	종교여부	1, age2_group, 만나이 , , 혼인년도 , 통근통학여부, 행정구역코드 시도, maryy_group, 추가계획자녀여부, 남자출생아수 여자출생아수 , ,
	종교종류	행정구역코드 시도 시군구 읍면동 종교여부 ; 가구원수 추가계획자녀여부 ,), ,
남북이산가족	남북이산가족 여부	행정구역코드 시도 시군구 읍면동 , 남북이산가족출생지 , , 만나이), 고령자생활비원천 , age2_group, ,
	남북이산가족 출생지	행정구역코드 시도 시군구 읍면동 1 남북이산가족여부 (, , 만나이), 년전거주지행정구역코드 , age2_group, ,
아동보육	아동보육	행정구역코드 시도 시군구 읍면동 5 어머니동거여부 통근통학여부 (,), 취업여부 활동제약여부
	어머니동거여부	1 행정구역코드 시도 시군구 읍면동 아동보육 통근통학여부 취업여부 혼인상태 5 (,), 1,

주 : age2_group 은 만나이를 5세 연령별로 그룹화한 변수임
maryy_group 은 혼인년도를 20년도별로 그룹화한 변수임

〈표 1-5〉 가구원 항목들에 대한 대체군 요약(표본)_계속

조사항목		대체군 카이제곱 ( )
년전 거주지	년전거주지 5	행정구역코드 시도 시군구 읍면동 ( ) 년전거주지 년전거주지행정구역코드 활동제약여부 5 5 6, 활동제약종류 어머니동거여부
	년전거주지 행정구역코드 5	행정구역코드 시도 시군구 읍면동 조사구특성 5 ( ) 년전거주지 혼인상태 가구주와의관계 교육정도 , 5 추가계획자녀여부 , 만나이 , , age2_group,
활동 제약	활동제약여부	대체 필요성 없음 단 모든 항목에 답을 하지 않은 경우에는 없음 에 답을 해야 함 ‘ ’
	활동제약종류	
통근통학여부		행정구역코드 시도 시군구 읍면동 취업여부 „ 통근통학장소 어머니동거여부 ( ), , 통근통학장소행정구역코드 , 행정구역코드 시도 시군구 읍면동)
통근통 학장소	통근통학장소	통근통학장소행정구역코드 이용교통수단 통근통학소요시간 분환산 1, tongtime_group, ( ),
	통근통학장소 행정구역코드	행정구역코드 시도 시군구 읍면동 통근통학장소 통근통학여부 구직여부 취업여부), ,
이용교 통수단	이용교통수단	행정구역코드 시도 시군구 읍면동 통근통학장소 통근통학여부 (, ,), , 통근통학소요시간 분환산 tongtime_group, 취업여부
통근통학소요시간 분환산 ( )		행정구역코드 시도 (시군구), 읍면동 통근통학장소 통근통학여부 이용교통수단 , 취업여부 1, , 행정구역코드 시도 시군구 읍면동 통근통학여부
경제활 동상태	취업여부	구직여부 어머니동거여부 혼인상태 ( ), , 행정구역코드 시도 시군구 읍면동 취업여부
	구직여부	취업가능성 통근통학여부 통근통학장소행정구역코드 ( ), , 행정구역코드 시도 시군구 읍면동 구직여부
	취업가능성	통근통학장소행정구역코드 통근통학여부 ( ), , 취업여부 가구원수 이용교통수단 , 통근통학소요시간 분환산 1,

주 은 통근통학소요시간을 분별로 그룹화한 변수임 ( )
: tongtime_group tongtime_group,

〈표 1-5〉 가구원 항목들에 대한 대체군 요약(표본)_계속

조사항목		대체군 카이제곱 ( )
종사상지위		행정구역코드 시도 시군구 읍면동 취업여부 ( ) 근로장소 통근통학장소 통근통학여부 , , ,
산업	직장사업체이름	-
	주된사업내용	-
직업	근무부서	-
	직책 직업 ( )	-
	하고있는일의종류	-
근로장소		행정구역코드 시도 시군구 읍면동 취업여부 ( ) 종사상지위 통근통학장소 통근통학여부 ,
혼인상태		행정구역코드' 시도 시군구 읍면동 가구원번호 어머니동거여부(, 만나이 취업여부 , 통근통학여부 , age2_group, , ,
혼인년월	혼인년도	행정구역코드 시도 시군구 읍면동 만나이 혼인상태 추가계획자녀여부 통근통학여부 , age2_group, , , ,
	혼인월	혼인년도와 함께 대체함
	양음력구분	혼인년도와 함께 대체함
총출생아수	남자출생아수	행정구역코드 시도 시군구 읍면동 추가계획자녀여부 (성별 혼인상태 여자출생아수),
	여자출생아수	행정구역코드 시도 시군구 읍면동 , 추가계획자녀여부 (성별 남자출생아수 혼인상태), , , ,
	남자자녀수	
	동거 비동거 (사망 , , 여자자녀수)	남자출생아수와 함께 대체함
	동거 비동거 (사망 , ,)	여자출생아수와 함께 대체함

〈표 1-5〉 가구원 항목들에 대한 대체군 요약(표본)_계속

조사항목		대체군 카이제곱 ( )
추가계획자녀수	추가계획자녀여부	행정구역코드 시도 시군구 읍면동 남자출생아수 ( ), 여자출생아수 추가계획자녀수 성별 , , ,
	추가계획자녀수	행정구역코드 시도 시군구 읍면동 (, ,) 추가계획자녀여부 , 혼인년도 남자출생아수 여자출생아수 , 성별 가구주와의 관계 , , ,
고령자생활비원천		행정구역코드 시도 시군구 읍면동 ( ), age2_group, 만나이 혼인상태 활동제약여부 6, marvy_group, 혼인년도, 활동제약종류 통근통학장소 주택여부 , 5, ,
외국인여부		행정구역코드 시도 시군구 읍면동 외국인국적 외국인교육정도(외국인주된체류목적), 외국인직업 ,
외국인 교육정도		행정구역코드 시도 시군구 읍면동 외국인직업 외국인주된체류목적 , 외국인거주기간), 혼인상태 ,
외국인 국적		행정구역코드 시도 시군구 읍면동 외국인교육정도 외국인주된체류목적 , 외국인직업 종교), 종류 교육정도 , 성별 조사구특성 , , , , ,
외국인 주된 체류목적		행정구역코드 시도 시군구 읍면동 외국인교육정도 외국인거주기간(외국인직업 , 혼인상태), , , ,
외국인 직업		행정구역코드 시도 시군구 읍면동 외국인교육정도 외국인주된체류목적 , 외국인거주기간), 혼인상태 ,
외국인 거주기간		행정구역코드 시도 시군구 읍면동 , 외국인주된체류목적 (, 외국인교육정도), 외국인직업 혼인상태 , , ,
외국인 거주기간 년 월		외국인거주기간과 함께 대체함

주 대체군의 나열(,)순서는 중요한 항목부터 나열된 것임

:

2) 가구에 관한 사항

〈표 1-6〉 가구 항목들에 대한 대체군 요약(전수)

조사항목		대체군 가이제곱
가구구분		행정구역코드 시도 시군구 읍면동 가구원수 (침실수 주인가구여부 거실수 점유형태 가구번호 거처의종류 , , , ,)
사용방수	침실수	행정구역코드 시도 시군구 읍면동 가구구분 (주인가구여부 가구원수 점유형태 거실수 조사구특성 거처의종류 건축년도 , , ,)
	침실이외의방수	행정구역코드 시도 시군구 읍면동 주인가구여부 (점유형태 단독주택의종류 거처의종류 건축년도 조사구특성 , , , ,)
	거실수	행정구역코드 시도 시군구 읍면동 단독주택의종류 (가구구분 침실수 점유형태 가구원수 화장실사용여부 조사구특성 , ,)
	식당수	행정구역코드 시도 시군구 읍면동 부엌사용여부 (수도형태 화장실형태 거처의종류 주인가구여부 난방시설 , , , ,)
주거시설형태	부엌형태	행정구역코드 시도 시군구 읍면동 부엌사용여부 (목욕시설사용여부 목욕시설형태 화장실형태 ,)
	부엌사용여부	행정구역코드 시도 시군구 읍면동 부엌형태 (목욕시설사용여부 화장실사용여부 점유형태 ,)
	수도형태	행정구역코드 시도 시군구 읍면동 단독주택의종류 (거처의종류 난방시설 화장실형태 건축년도 식당수 , ,)
	화장실형태	행정구역코드 시도 시군구 읍면동 화장실사용여부 (목욕시설형태 난방시설 목욕시설사용여부 , ,)
	화장실사용여부	행정구역코드 시도 시군구 읍면동 화장실형태 (목욕시설사용여부 부엌사용여부 목욕시설형태 , ,)
	목욕시설형태	행정구역코드 시도 시군구 읍면동 화장실사용여부 (목욕시설사용여부 난방시설 화장실형태 부엌형태 , , ,)
	목욕시설사용여부	행정구역코드 시도 시군구 읍면동 목욕시설형태 화장실사용여부 (부엌사용여부 화장실사용여부 부엌형태 , , , ,)

〈표 1-6〉 가구 항목들에 대한 대체군 요약(전수)_계속

조사항목		대체군 카이제곱( )
거주층	거주층구분	행정구역코드 시도 시군구 읍면동 가구번호 ( ) 거주층수 거처의 종류 주인가구여부 floor2_group, 건축년도 난방시설 점유형태 , , , ,
	거주층수	행정구역코드 시도 시군구 읍면동 조사구특성 ( ) 거주층구분 주인가구여부 거처의종류 건축년도 , , , ,
난방시설		행정구역코드 시도 시군구 읍면동 조사구특성 ( ) 목욕시설형태 화장설형태 단독주택의종류 , 주인가구여부 , , ,
점유형태	주거영업구분	행정구역코드 시도 시군구 읍면동 조사구특성 ( ) 거처의종류 단독주택의종류 부업사용여부 , ,
	점유형태	행정구역코드 시도 시군구 읍면동 가구번호 주인가구여부 거처의종류 건축년도), , ,
주인가구 및 주택소유여부	주인가구여부	행정구역코드 시도 시군구 읍면동 가구번호 거처의종류 점유형태 건축년도), ,
	타지주택소유	행정구역코드 시도 시군구 읍면동 조사구특성 가구번호 거처의종류 건축년도 주인가구여부 , 침실수 점유형태 , , ,

주 대체군의 나열순서는 중요한 항목부터 나열된 것임

: 은 거주층수를 층별로 그룹화한 변수임

floor2_group 5

〈표 1-7〉 가구 항목들에 대한 대체군 요약(표본)

조사항목		대체군 카이제곱 ( )
가구구분		행정구역코드 시도 시군구 읍면동 가구원수 ( ), 자동차보유여부 침실수 승용차수 주인가구여부 ' , , , , 주차시설
거주기간		행정구역코드 시도 시군구 읍면동 단독주택의 종류 ( ) 주인가구여부 수도형태 건축년도 화장실형태 ' , , , , 거처의종류
사용방수	침실수	행정구역코드 시도 시군구 읍면동 승용차수 ( ) 자동차보유여부 가구원수 가구구분 주인가구여부 침실이외의방수 , , , ,
	침실이외의 방수	행정구역코드 시도 시군구 읍면동 침실수 ( ) 주인가구여부 식당수 단독주택의종류 , 행정구역코드 시도 시군구 읍면동 주인가구여부
	거실수	침실수 자동차보유여부 목욕시설사용여부 ( ) , 목욕시설형태 부엌형태 거처의종류 , 행정구역코드 시도 시군구 읍면동 단독주택의종류
	식당수	거처의종류 난방시설 수도형태 조사구특성 거주기간 ( ) , 화장실형태 , , , , , ,
	부엌형태	행정구역코드 시도 시군구 읍면동 부엌사용여부 난방시설 목욕시설형태 목욕시설사용여부 ( ) , 행정구역코드 시도 시군구 읍면동 가구번호
주거시설형태	부엌사용여부	부엌형태 목욕시설사용여부 화장실사용여부 ( ) , 거처의종류 , , 행정구역코드 시도 시군구 읍면동 단독주택의종류
	수도형태	거처의종류 난방시설 거주기간 화장실형태 ( ) , 행정구역코드 시도 시군구 읍면동 화장실사용여부
	화장실형태	목욕시설형태 난방시설 목욕시설사용여부 ( ) , 행정구역코드 시도 시군구 읍면동 가구번호
	화장실사용여부	화장실형태 목욕시설사용여부 부엌사용여부 ( ) , 거처의종류 부엌형태 , , 행정구역코드 시도 시군구 읍면동
	목욕시설형태	목욕시설사용여부 난방시설 화장실형태 부엌형태 ( ) , 행정구역코드 시도 시군구 읍면동 목욕시설형태
	목욕시설사용여부	부엌사용여부 화장실사용여부 화장실형태 ( ) , , , ,

〈표 1-7〉 가구 항목들에 대한 대체군 요약(표본)_계속

조사항목		대체군 카이제곱 ( )
거주층	거주층구분	행정구역코드 시도 시군구 읍면동 ( ), floor2_group, 거주층수 가구번호 '거처의 종류 주인가구여부 건축년도 ' ' 전세보증금 ' ', , rent1_group,
	거주층수	행정구역코드 시도 시군구 읍면동 거주층구분 ( ) 조사구특성 주인가구여부 자동차보유여부 , 단독주택의종류 , ,
자동차 보유 대수	승용차보유대수	행정구역코드 시도 시군구 읍면동 자동차보유여부 ( ), 가구원수 침실수 주차시설 ,), , , , 1
	승합차보유대수	행정구역코드 시도 시군구 읍면동 자동차보유여부 주차시설 주거영업구분 ,), , 1,
	화물 및 기타 자동차보유대수	행정구역코드 시도 시군구 읍면동 자동차보유여부 주차시설 가구(원수, 단독주택의종류), , 1
	자동차보유여부	대체군 없음 승용 승합 트럭수 항목에 의해 결정 ( )
주차 시설	주차시설	행정구역코드 시도 '시군구 읍면동 조사구특성 자동차보유여부(승용차수 주인가구여부 , 1
	난방시설	행정구역코드 시도 시군구 읍면동 조사구특성 단독주택의종류(목욕시설형태), ,
점유 형태	주거영업구분	행정구역코드 시모 시군구 읍면동 조사구특성 거처의 종류 단독주택의 종류 주차시설 난방시설 ( ), 월세 사글세 , 거주층수 ,
	점유형태	행정구역코드 시도 (시군구), 읍면동 주인가구여부 rent2_group, floor2_group, 전세보증금 (,), 월세 사글세 , 사글세개월수 rent1_group, , rent2_group, ( ),
	사글세개월수	행정구역코드 시도 시군구 읍면동 점유형태 주인가구여부 전세보증금 (, ,), ,

주 은 거주층수를 층별로 그룹화한 변수임 rent1_group,
은 전세보증금을 만원별로' 그룹화한 변수임
은 월세 사글세 를5 만원별로 그룹화한 변수임

: floor2_group
rent1_group 1,000
rent2_group ( ) 20

〈표 1-7〉 가구 항목들에 대한 대체군 요약(표본)_계속

조사항목		대체군 카이제곱 ( )
임차료	전세보증금	행정구역코드 시도 시군구 읍면동 주인가구여부 ( ), 점유형태 자동차보유여부 , 월세 사글세 , 단독주택의종류 , rent2_group, ( ),
	월세 사글세 ( )	행정구역코드 시도 시군구 읍면동 가구번호 ( ), 점유형태 주인가구여부 , 전세보증금 , 자동차보유여부 거처의종류 건축년도 , rent1_group, , ,
주인가구 및 주택소유여부	주인가구 여부	행정구역코드 시도 시군구 읍면동 가구번호 ( ), 거처의종류 점유형태 건축년도 , , ,
	타지주택 소유여부	행정구역코드 시도 시군구 읍면동 조사구특성 ( ), 승용차수 주차시설 , , 전세보증금 , 자동차보유여부 거처의종류 건축년도 , rent1_group, ,

주 : 대체군의 나열순서는 중요한 항목부터 나열된 것임

3) 주택에 관한 사항

〈표 1-8〉 주택 항목들에 대한 대체군 요약(전수)

조사항목		대체군 카이제곱 ( )
거처의 종류		행정구역코드 시도 시군구 읍면동 조사구특성 ( ) 단독주택의종류(, 건물층수, 방수 건축년도 , floor_group, , ,
거처의 종류	단독주택의 종류	행정구역코드 시도 시군구 읍면동 ( ), dae1_group, 대지면적 거처의종류 부엌수 독립된 출입구수
	건물층수	행정구역코드 시도 시군구 읍면동 조사구특성 ( ) 단독주택의종류(거처의종류 ,건축년도, ,
연건평 주거용연면적 ㎡ (,)		행정구역코드 시도 시군구 읍면동 조사구특성 ( ) 단독주택의종류(방수 거처의종류 화장실수 , 건물층수 , , , floor_group
대지면적		행정구역코드 시도 시군구 읍면동 단독주택의종류 거처의종류 (, 건물층수 건축년도 , 식당수 , floor_group, 연건평 화장실수 , size1_group,
총방 수	방수	가구부문 결과의 합으로 대체
	거실수	가구부문 결과의 합으로 대체
	식당수	가구부문 결과의 합으로 대체
건축년도		행정구역코드 시도 시군구 읍면동 조사구특성 단독주택의종류(거처종류 ,), 건물층수 , 화장실수 , , floor_group, ,
편익 시설 수	부엌수	행정구역코드 시도 시군구 읍면동 독립된 출입구수 화장실수 단독주택의종류 (, 거실수), ,
	화장실수	행정구역코드 시도 시군구 읍면동 부엌수 독립된 출입구수 (, 단독주택의종류 거실수), ,
	독립된	행정구역코드 시도, 시군구 읍면동 ,부엌수 화장실수
	출입구수	단독주택의종류 (, 거실수), , ,

주 대체군의 나열순서는 중요한 항목부터 나열된 것임

: 은 건물층수를 층별로 그룹화한 변수임
은 연건평 평 단위를 ㎡로 환산하여 평별을 그룹화한 변수임
은 대지면적 평 단위를 ㎡로 환산하여 평별을 그룹화한 변수임

floor_group	5
size1_group	5
dae1_group	5

〈표 1-9〉 주택 항목들에 대한 대체군 요약(표본)

조사항목		대체군 카이제곱 ( )
거처의 종류	거처의 종류	행정구역코드 시도 시군구 읍면동 조사구특성 ( ), 단독주택의종류 식당수 , 건물층수 , , floor_group,
거처의 종류	단독주택의 종류	행정구역코드 시도 시군구 읍면동 ( ), 대지면적 ㎡ ( ), 거처의종류 부엌수 독립된출입구수 , , dae1_group
거처의 종류	건물층수	행정구역코드 시도 시군구 읍면동 조사구특성 ( ), 식당수 단독주택의종류 거처의종류 , , ,
연건평 주거용연면적 ㎡ (,)		행정구역코드 시도 시군구 읍면동 식당수 거실수 ( ), 단독주택의종류 방수 거처의 종류 , , ,
대지면적 ㎡ ( )		행정구역코드 시도 시군구 읍면동 조사구특성 ( ), 단독주택의종류 거처의종류 식당수 , , 건물층수 건축년도 , , floor_group, ,
총방수	방수	가구부문 결과의 합으로 대체
총방수	거실수	가구부문 결과의 합으로 대체
총방수	식당수	가구부문 결과의 합으로 대체
건축년도		행정구역코드 시도 시군구 읍면동 식당수 거실수 거처의종류 (, 건물층수), , , floor_group,
편익시설수	부엌수	행정구역코드 시도 시군구 읍면동 독립된출입구수 ( ), 화장실수 단독주택의종류 식당수 , ,
편익시설수	화장실수	행정구역코드 시도 시군구 읍면동 독립된출입구수 ( ), 부엌수 단독주택의종류 방수 , ,
편익시설수	독립된 출입구수	행정구역코드 시도 시군구 읍면동 부엌수 화장실수 ( ), 단독주택의종류 방수 , , ,

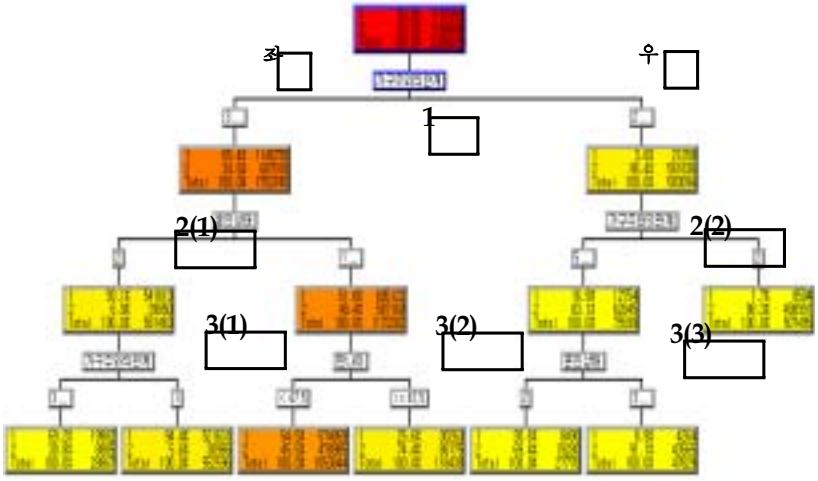
주 : 대체군의 나열순서는 중요한 항목부터 나열된 것임
floor_group 은 건물층수를 5층별로 그룹화한 변수임
size1_group 은 연건평 평 단위를 ㎡로 환산하여 5평별을 그룹화한 변수임
dae1_group 은 대지면적 평 단위를 ㎡로 환산하여 5평별을 그룹화한 변수임

2. CHAID 알고리즘에 의한 대체군 개발

의사결정나무의 분리 알고리즘 중의 하나인 CHAID는 목표변수가 범주형 자료인 경우에는 χ^2통계량에 의한 분할, 연속형 자료인 경우에는 F검정을 이용한 분할을 수행하는 분석방법이다. 카이제곱 검정과 가장 큰 차이점은 목표변수와 설명변수들간의 관계를 독립적으로 하나씩 고려하지 않고 모든 독립변수들을 동시에 고려하여 연관성 규칙을 찾아내는 것이다. CHAID 알고리즘은 2개 이상의 하위 나무구조를 반복적으로 분할하는데 이 때 설명변수의 범주쌍에 대하여 목표변수의 유의한 차이가 없으면 설명변수의 범주들을 병합하고, 유의적이지 않은 쌍이 나타날 때까지 분할을 계속한다. 그러나 일반적으로는 정지규칙을 정하고 그 규칙에 해당될 때까지만 분할을 계속한다. 이후에 각 목표변수에 대하여 가장 유의적으로 분할하는 설명변수를 선택하여 그 설명변수의 범주에 의하여 자료를 분할하는 방법이다. 구체적인 알고리즘은 『무응답 처리를 위한 방법론 연구(Ⅰ)(통계개발원, 2009)』의 『농업총조사 항목간 연관성 분석 및 대체군(보조변수) 개발(2008, 최필근)』을 참조하기 바란다.

가. 성별 항목에 대한 연관성 분석

CHAID 알고리즘에 의한 성별 항목의 연관성 분석의 결과가 <표 1-10>과 [그림 1-1]에 제시되어 있다. 성별 항목 이외의 모든 항목들도 같은 방식으로 설명되므로 성별 항목에 대해서만 자세하게 설명하고, 나머지 항목들은 분석에 의한 결과만을 제시할 것이다. <표 1-10>에서 깊이는 부여되는 순서를 의미하고 분리변수의 괄호안의 숫자는 각 가지에서의 마디 번호를 나타낸다.



[그림 1-1] 성별에 관한 연관성 모형

〈표 1-10〉 성별에 관한 연관성 분석의 세부내용

성별

(CHAID)

깊이 연관성	분리변수	분리지점 좌	분리지점 우
( )	가구주와의관계	( )	( )
1	혼인상태	1, 3, 7, 8, 10~14	2, 4~6, 9
2	가구주와의관계 (1)	2	1, 3, 4, M
	가구주와의관계 (2)	4~6, 9	2
3	만나이 (1)	3, 7, 8, 10~14	세 이상
	혼인상태 (2)	47.5	47.5
주 은 결측값을 의미함	(3)	2	3, 4, M

: M

1

가장 연관성이 높은 항목을 의미한다 성별을 분리하는데 가장 연관성이 높은 항목은 가구주와의 관계이다. 전체 남자와 여자의 구성비는 약 와 이며 이를 분리하여 나간다

50% 50% .

가구주와의 관계에 의하여 왼쪽으로 분리되는 것을 의미한다. 가구주와의 관계가 가구주(1), 자녀(3), 손자녀(7), 증손자녀(8), 형제자매(10), 형제자매의 자녀(11), 부모의 형제자매(12), 기타친인척(13), 기타동거인(14)인 경우는 좌로 분리되고, 이때의 구성비는 남자가 65.4%, 여자가 34.6%로 분리되고 있음을 알 수 있다.

가구주와의 관계에 의하여 오른쪽으로 분리되는 것을 의미한다. 가구주와의 관계가 배우자(2), 자녀의 배우자(4), 부모(5), 배우자의 부모(6), 조부모(9)인 경우는 우로 분리되고, 이때의 구성비는 남자가 3.6%, 여자가 96.4%로 분리가 되고 있다. 처음 구성비에서 확실한 차이를 나타내므로 매우 잘 분리된 것을 알 수 있다.

2(1)

첫 번째 분리가 끝난 후 두 번째 분리가 시작된다. 이 과정은 첫 번째 분리가 된 왼쪽부분을 다시 세부적으로 분리한다. 두 번째로 연관성이 높은 혼인상태 항목으로 다시 분리가 된다. 혼인상태가 배우자있음(2)인 경우는 왼쪽으로 분리되며 이때의 남녀의 구성비는 93.1%와 6.9%가 된다. 따라서 첫 번째 분리후의 구성비 65.4%와 34.6%에 비하여 상당히 많은 격차가 벌어진 것을 볼 수 있다. 그리고 미혼(1), 사별(3), 이혼(4)인 경우는 오른쪽으로 분리되며 이때의 남녀의 구성비는 51.6%와 48.4%가 된다.

2(2)

이 과정은 첫 번째 분리가 된 오른쪽부분을 다시 세부적으로 분리한다. 다시 가구주와의 관계 항목으로 분리가 되는데 가구주와의 관계가 자녀의 배우자(4), 부모(5), 배우자의 부모(6)인 경우는 오른쪽으로 분리되며 이때의 남녀의 구성비는 16.9%와 83.1%가 된다. 그리고 배우자(2)인 경우는 오른쪽으로 분리되며 이때의 남녀의 구성비는 1.7%와 98.3%로 매우 좋은 분리가 되었음을 알 수 있다.

3(1)

두 번째 분리가 끝난 후 세 번째 분리가 시작된다. 이 과정은 두 번째 분리가 된 처음부분을 다시 세부적으로 분리한다. 가구주와의 관계 항목으로 다시 분리가 된다. 가구주와의 관계가 자녀(3), 손자녀(7), 증손자녀(8), 형제자매(10), 형제자매의 자녀(11), 부모의 형제자매(12), 기타친인척(13), 기타동거인(14)인 경우는 왼쪽으로 분리되며 이때의 남녀의 구성비는 69.2%와 30.8%가 된다. 그리고 가구주(1)인 경우는 오른쪽으로 분리되며 이때의 남녀의 구성비는 94.4%와 5.6%로 이전에 비해서 조금 더 격차가 벌어졌다.

3(2)

이 과정은 두 번째 분리가 된 두 번째 부분을 다시 세부적으로 분리한다. 다음으로 연관성이 높은 만나이 항목으로 다시 분리가 된다. 만나이가 47.5세 미만인 경우는 왼쪽으로, 47.5세 이상이면 오른쪽으로 분리되며 이때의 남녀의 구성비는 25.6%와 74.4%가 된다. 두 번째 분리후의 구성비 51.6%와 48.4%에서 매우 좋은 분리가 되었음을 알 수 있다.

3(3)

이 과정은 두 번째 분리가 된 세 번째 부분을 다시 세부적으로 분리한다. 혼인상태 항목으로 다시 분리가 된다. 혼인상태가 배우자있음(2)인 경우는 왼쪽으로 분리되며 이때의 남녀의 구성비는 30.6%와 69.4%가 된다. 그리고 사별(3), 이혼(4)인 경우는 오른쪽으로 분리되며 이때의 남녀의 구성비는 8.9%와 91.1%로 좋은 분리가 되고 있다.

이상으로 성별을 분리하는 과정을 자세히 설명하였다. 요약하면, 가구주와의 관계, 혼인상태, 만나이가 성별을 분리하는데 많은 연관성이 있다는 것을 알 수 있다. 그리고 세부적으로 분리가 잘 되는 부분에서는 대체의 정확도가 상당히 높을 것으로 판단되며, 반대의 경우에서 대체의 정확도는 떨어질 것이다. 따라서 현재 사용할 수 있는 대체군으로 어느 정도의 대체의 정확성이 있는지는 모의실험을 통해서 볼 수 있을 것이다.

나. 각 항목별 연관성 분석결과 요약

1) 가구원에 관한 사항

〈표 1-11〉 가구원 항목들에 대한 대체군 요약(전수)

조사항목		대체군 (CHAID)
성명		-
성별		행정구역코드 시도 시군구 읍면동 가구원번호 가구주와의관계 (혼인상태 ,), 만나이 , , , age2_group,
나이	만나이	행정구역코드 시도 시군구 읍면동 가구원번호 (,) 혼인상태 교육정도 교육상태 가구주와의관계 ,
가구주와의 관계		행정구역코드 시도 시군구 읍면동 가구번호 가구원번호 조사구특성 혼인상태 성별 , 만나이 , , , , age2_group,
교육정도	교육정도	행정구역코드 시도 시군구 읍면동 조사구특성 가구원번호 교육상태, 혼인상태), 만나이, 성별 , , , age2_group, ,
	교육상태	행정구역코드 시도 시군구 읍면동 가구원번호 (,) 만나이 교육정도, 가구주와의관계 , age2_group,
	종교여부	행정구역코드 시도 시군구 읍면동 종교종류
종교	종교종류	행정구역코드 시도 시군구 읍면동), 만나이 교육정도 (, ,), age2_group,
남북이산가족	남북이산가족 여부	행정구역코드 시도 시군구 읍면동 , 남북이산가족출생지 , 가구주와의관계), 만나이 , , age2_group,
	출생지	행정구역코드 시도 시군구 읍면동 만나이 남북이산가족여부 (, ,), age2_group, ,
혼인상태		행정구역코드 시도 시군구 읍면동 가구원번호 가구주와의관계 (, 만나이 성별), ,

주 은 만나이를 세 연령별로 그룹화한 변수임 age2_group, ,

: age2_group 5

〈표 1-11〉 가구원 항목들에 대한 대체군 요약(전수)_계속

조사항목	대체군 (CHAID)
외국인여부	행정구역코드 시도 시군구 읍면동 외국인국적 (, ,),
외국인 교육정도	행정구역코드 시도 시군구 읍면동 외국인직업 (,), 외국인국적 , 만나이 , age2_group,
외국인 국적	행정구역코드 시도 시군구 읍면동 가구번호 (,), 외국인체류목적 , 만나이 , age2_group, 외국인교육정도 외국인직업 가구주와의관계 , ,
외국인 주된 체류목적	행정구역코드 시도 시군구 읍면동 외국인직업 (, ,),
외국인 직업	행정구역코드 시도 시군구 읍면동 외국인주된체류목적 ,외국인교육정도), ,
외국인 거주기간	행정구역코드 시도 시군구 읍면동 외국인주된체류목적 , ,),
외국인 거주기간 년 월	외국인거주기간과 함께 대체함

주 : 대체군의 (,) 순서는 중요한 항목부터 나열된 것임

〈표 1-12〉 가구원 항목들에 대한 대체군 요약(표본)

조사항목		대체군 (CHAID)
성별		행정구역코드 시도 시군구 읍면동 가구원번호 (,) 여자출생아수 혼인상태 , 만나이 , 이용교통수단 , age2_group, ,
나이	만나이	행정구역코드 시도 시군구 읍면동 가구원번호 (,) 혼인상태 고령자생활비원천 , 교육정도 교육상태 , 1 년전거주지 , , 5
가구주와의관계		행정구역코드 시도 시군구 읍면동 가구원번호 (,) 혼인년도 조사구특성 성별 혼인상태 , maryy_group 만나이 고령자생활비원천 , , age2_group 행정구역코드 시도 시군구 읍면동 교육상태
교육정도	교육정도	혼인상태 (, 만나이,), , 어머니동거여부 고령자생활비원천 age2_group 행정구역코드 시도 시군구 읍면동 조사구특성
	교육상태	가구원번호 (, 만나이 취업여부 , 통근통학여부, age2_group, , ,
	종교여부	행정구역코드 시도 시군구 읍면동 종교종류
종교		행정구역코드 시도 시군구 읍면동 조사구특성 ,
	종교종류	만나이 가구주와의관계), 교육정도 , 교육상태 age2_group, , , ,
남북이산가족	남북이산가족 여부	행정구역코드 시도 시군구 읍면동 남북이산가족 출생지, , 만나이),
	출생지	행정구역코드 시도 시군구 읍면동, age2_group, 만나이 남북이산가족여부 가구주와의관계), age2_group,
아동보육	아동보육 어머니동거1 여부	행정구역코드 시도 시군구 읍면동 교육정도 어머니동거여부 (조사구특성 , 가구주와의관계), , 행정구역코드 시도 시군구 읍면동 , 가구주와의관계 (아동보육 ,), 행정구역코드 시도 시군구 읍면동 조사구특성
년전거주지	년전거주지 5	1 , 가구번호 년전거주지 , 만나이 , 행정구역코드 시도 시군구 읍면동 조사구특성 , 5 년전거주지 , age2_group, 년전거주지 혼인상태 가구주와의관계 교육정도 ,
	행정구역코드 5	(, ,), 추가계획자녀여부 , 만나이 , , 5

주 은 만나이를 세 연령별로 그룹화한 변수임
은 혼인년도를 년도별로 그룹화한 변수임 5 age2_group,

: age2_group 5
maryy_group 20

〈표 1-12〉 가구원 항목들에 대한 대체군 요약(표본)_계속

항목		대체군 (CHAID)
활동제약	활동제약여부	대체 필요성 없음 단 모든 항목에 답을 허지 않은 경우에는 없음 에 답을 해야 함
	활동제약종류	
통근통학여부		행정구역코드 시도 시군구 읍면동 통근통학장소 (, ,), , 근로장소
통근통학장소	통근통학장소	행정구역코드 시도 시군구 읍면동 ( ), 통근통학장소행정구역코드 이용교통수단 통근통학소요시간 분환산 1, tongtime_group)
	통근통학장소 행정구역코드	행정구역코드 시도 시군구 읍면동 ( ), 통근통학소요시간 분환산 1, tongtime_group)
이용교통수단	이용교통수단 1	행정구역코드 시도 시군구 읍면동 (통근통학장소 만나이 통근통학장소행정구역코드), , , 통근통학소요시간 분환산 tongtime_group)
통근통학소요시간 분환산 ( )		성별 종사상지위 ( ), 행정구역코드 시도 시군구 읍면동 통근통학장소 이용교통수단 (통근통학장소행정구역코드 , 행정구역코드 시도 시군구 읍면동 근로장소
경제활동상태	취업여부	구직여부 종사상지위 ,), ,
	구직여부	행정구역코드 시도 시군구 읍면동 취업가능성 ,
		행정구역코드 시도 시군구 읍면동),
	취업가능성	여자출생아수 교육정도 교육상태 활동제약여부 만나이 , , 6, 행정구역코드 시도 시군구 읍면동 근로장소 age2_group,
종사상지위		가구주와의 관계 (통근통학장소), , 만나이 , 통근통학소요시간 분환산 age2_group, 이용교통수단 tongtime_group, ( ),
근로장소		행정구역코드 시도 시군구 읍면동 종사상지위 1 교육정도 통근통학장소 성별 이용교통수단 (, ,), , 행정구역코드 시도 시군구 읍면동 가구원번호 , 1
혼인상태		혼인년도 추가계획자녀여부 (, ,), , 고령자생활비원천 가구주와의관계 maryy_group, , 만나이 1, , age2_group,
혼인년월	혼인년도	행정구역코드 시도 시군구 읍면동 만나이 추가계획자녀여부 (,),
	혼인월	혼인년도와 함께 대체함 age2_group,

주 은 통근통학소요시간을 분별로 그룹화한 변수임

: tongtime-group

〈표 1-12〉 가구원 항목들에 대한 대체군 요약(표본)_계속

항목		대체군 (CHAID)
총 출생아수	남자출생아수	행정구역코드 시도 시군구 읍면동 ( ) 추가계획자녀여부 ' ' 혼인년도 여자출생아수 고령자생활비원천 maryy_group, , 1
	여자출생아수	행정구역코드 시도 시군구 읍면동 ( ) maryy_group, 혼인년도 추가계획자녀여부 남자출생아수 , ,
	남자자녀수	남자출생아수와 함께 대체함
	동거 비동거 사망 ( ) 여자자녀수	여자출생아수와 함께 대체함
	동거 비동거 사망 (, ,)	행정구역코드 시도 시군구 읍면동 ( )
추가 계획 자녀수	추가계획자녀여부	혼인년도 남자출생아수 여자출생아수), maryy_group, , 만나이 , ,
	추가계획자녀수	행정구역코드 시도 시군구 읍면동 추가계획자녀 age2_group, 혼인년도 남자출생아수) 여자출생아수 , maryy_group,
고령자생활비원천 1		행정구역코드 시도 시군구 읍면동 조사구특성 취업여부 가구주와의관계 혼인상태) 종교여부 , 교육정도 , , , ,
외국인여부		행정구역코드 시도 시군구 읍면동 외국인국적 ( )
외국인 교육정도		행정구역코드 시도 시군구 읍면동 외국인직업 외국인국적 (, 만나이), , 행정구역코드, 시군구, age2_group, 읍면동 가구번호
외국인 국적		외국인주된체류목적 (, , 만나이), , 외국인교육정도 외국인직업 가구주와의관계 , age2_group,
외국인 주된 체류목적		행정구역코드 시도 시군구 읍면동 외국인직업
외국인 직업		행정구역코드 시도 시군구 읍면동), 외국인 주된 체류목적 외국인교육정도 (,),
외국인 거주기간		행정구역코드 시도 시군구 읍면동 외국인 주된 체류목적 (, ,),
외국인 거주기간 년 월		외국인거주기간과 함께 대체함

주 대체군의 나열순서는 중요한 항목부터 나열된 것임 (,)

:

2) 가구에 관한 사항

〈표 1-13〉 가구 항목들에 대한 대체군 요약(전수)

항목		대체군 (CHAID)
가구구분		행정구역코드 시도 시군구 읍면동 가구원수 ( ), 주거영업구분 점유형태 침실수 , 행정구역코드 시도 시군구 읍면동 가구원수 ( ), 침실이외방수 주인가구여부 거실수 난방시설 , 수도형태 가구구분 , , , 행정구역코드 시도 시군구 읍면동 침실수 ( ), 주인가구여부 거실수 거처의 종류 점유형태 , 건축년도 가구구분 , , , 행정구역코드 시도 시군구 읍면동 침실수 ( ), 침실이외방수 점유형태 식당수 건축년도 , 주인가구여부 , , , , 행정구역코드 시도 시군구 읍면동 난방시설
사용방수	침실수	
	침실이외방수	
	거실수	
	식당수	주인가구여부 거처의 종류 침실수), , 거주층수 , , , floor2_group, 행정구역코드 시도 시군구 읍면동 부엌사용여부 ( ), 목욕시설형태 난방시설 화장실형태), , 행정구역코드 시도 시군구 읍면동 ,
주거시설형태	부엌형태	
	부엌사용여부	목욕시설사용여부 거처의 종류 화장실사용여부 ( ), 화장실형태 , , , , 행정구역코드 시도 시군구 읍면동 거처의종류 ( ), 행정구역코드 시도 시군구 읍면동 목욕시설형태 , 건축년도 화장실사용여부 수도형태), 난방시설 , 부엌형태 , , , , 행정구역코드 시도 시군구 읍면동 , 목욕시설사용여부 주인가구여부 화장실형태 ( ), 부엌사용여부 거처의종류 , , 행정구역코드 시도 시군구 읍면동 , 목욕시설사용여부 난방시설 ( ), 행정구역코드 시도 시군구 읍면동 , 부엌사용여부 화장실사용여부 ( ), 은 거주층수를 층별로 그룹화한 변수임 ,
	수도형태	
	화장실형태	
	화장실사용여부	
	목욕시설형태	
	목욕시설사용여부	
주		

: floor2_group 5

〈표 1-13〉 가구 항목들에 대한 대체군 요약(전수)_계속

항목		대체군 (CHAID)
거주층	거주층구분	행정구역코드 시도 시군구 읍면동 (거주층수 주인가구여부 건축년도), floor2_group, , ,
	거주층수	행정구역코드 시도 시군구 읍면동 거처번호 (거처의종류 건축년도 난방시설), , , ,
난방시설		행정구역코드 시도 시군구 읍면동 조사구특성 (거처의종류 수도형태 건축년도 점유형태), , , ,
주거영업구분	주거영업구분	행정구역코드 시도 시군구 읍면동 거처의 종류 (거주층수 ,), , floor2_group,
	점유형태	행정구역코드 시도 시군구 읍면동 주인가구여부 (거실수 난방시설 , ,), ,
주인가구 및 주택 소유 여부	주인가구여부	행정구역코드 시도 시군구 읍면동 점유형태 (거처의 종류 가구번호 ,), 거주층수 , , , floor2_group,
	타지주택 소유여부	행정구역코드 시도 시군구 읍면동 침실수 (점유형태 거처종류 ,난방시설 건축년도), , ,

주 대체군의 나열순서는 중요한 항목부터 나열된 것임

:

〈표 1-14〉 가구 항목들에 대한 대체군 요약(표본)

항목		대체군 (CHAID)
가구구분		행정구역코드 시도 시군구 읍면동 가구원수 ( ) 주거영업구분 점유형태 주차시설 침실수 , 1,
거주기간		행정구역코드 시도 시군구 읍면동 건축년도 ( ) 주인가구여부 거처종류 난방시설 조사구특성 , 전세보증금 , 월세 사글세 , rent1_group rent2_group ( )
사용방수	침실수	행정구역코드 시도 시군구 읍면동 가구원수 ( ), , 침실이외방수 주인가구여부 거실수 전세보증금 가구구분 , , rent1_group,
	침실이외방수	행정구역코드 시도 시군구 읍면동 침실수 ( ), , 주인가구여부 거실수 식당수 전세보증금 거처의종류 가구구분 rent1_group,
	거실수	행정구역코드 시도 시군구 읍면동 침실수 ( ) 침실이외방수 식당수 주인가구여부 거처의종류
	식당수	행정구역코드 시도 시군구 읍면동 난방시설 ( ), , 침실이외방수 침실수 거처의종류
주거시설형태	부엌형태	행정구역코드 시도 시군구 읍면동 목욕시설형태 ( ) 부엌사용여부 난방시설 화장실형태 건축년도 ,
	부엌사용여부	행정구역코드 시도 시군구 읍면동 , 목욕시설사용여부 거처의종류 화장실사용여부 화장실형태 , , ,
	수도형태	행정구역코드 시도 시군구 읍면동 거처의종류
	화장실형태	행정구역코드 시도 시군구 읍면동 목욕시설형태 ( ) 거주기간 건축년도 화장실사용여부 부엌형태 ,
	화장실사용여부	행정구역코드 시도 시군구 읍면동 , 주인가구여부 목욕시설사용여부 화장실형태 ( ) 부엌사용여부 , , ,
	목욕시설형태	행정구역코드 시도 시군구 읍면동 목욕시설사용여부 난방시설 부엌형태 ( )
	목욕시설사용여부	행정구역코드 시도 시군구 읍면동 부엌사용여부 화장실사용여부 ( ) ,

주 은 거주층수를 층별로 그룹화한 변수임
은 전세보증금을 만원별로 그룹화한 변수임
분 월세 사글세 를 5 만원별로 그룹화한 변수임
: floor2_group
rent1_group 1,000
rent2_group ( ) 20

〈표 1-14〉 가구 항목들에 대한 대체군 요약(표본)_계속

항목		대체군 (CHAID)
거주층	거주층구분	행정구역코드 시도 시군구 읍면동 ( ), floor2_group, 거주층수 주인가구여부 건축년도
	거주층수	행정구역코드 시도 시군구 읍면동 거처번호 ( ), , 거처의 종류 건축년도 난방시설
자동차 보유 대수	승용차수	행정구역코드 시도 시군구 읍면동 ( ) 자동차보유여부 침실수 주차시설 침실이외방수 1 , 가구구분 타지주택소유여부 거처의 종류
	승합차수	행정구역코드 시도 시군구 읍면동 ( ) 자동차보유여부 승용차수 트럭기타수 식당수
	기타트럭수	행정구역코드 시도 시군구 읍면동 , ( ) 자동차보유여부 승용차수 식당수 주차시설
주차 시설	자동차보유여부	대체군 없음 승용 , 승합 트럭수 항목에 의해 결정 1
	주차시설	행정구역코드 시도 시군구 읍면동 거처의종류 ( ) 수도형태 건축년도 , ,), , 1 행정구역코드 시도 시군구 읍면동 거처의종류
난방시설		수도형태 조사구특성 건축년도), , 거주층수 목욕시설형태 , , floor2_group, 행정구역코드 시도 시군구 읍면동 거처의종류
점유 형태	주거영업구분	거주층수 난방시설), , 월세 floor2_group, , 전세보증금rent2_group, 행정구역코드 시도 시군구 읍면동 주인가구여부 ( ), rent1_group,
	점유형태	월세 사글세 ,), 전세보증금 , 사글세개월수 ( ), rent1_group, rent2_group, 사글세개월수 행정구역코드 시도 시군구 읍면동 점유형태
		행정구역코드 시도 시군구 읍면동 점유형태 ( ),
임차료	전세보증금	난방시설 침실수 거처의종류 , ( ), 침실이외방수 , 건축년도 , , 행정구역코드 시도 시군구 읍면동 점유형태 ,
	월세	전세보증금 거주기간 주거영업구분 ( ), 난방시설 건축년도 침실수 , , rent1_group, , 행정구역코드 시도 시군구 읍면동 점유형태 ,
주인 가구 및 주택 소유 여부	주인가구여부	가구번호 거처의종류 (,), ,
	타지주택	행정구역코드 시도 시군구 읍면동 승용차수
	소유여부	전세보증금 건축년도 (,), , rent1_group,

주 대체군의 나열순서는 중요한 항목부터 나열된 것임 ,

:

3) 주택에 관한 사항

〈표 1-15〉 주택 항목들에 대한 대체군 요약(전수)

항목		대체군 (CHAID)
거처의 종류	거처의 종류	행정구역코드 시도 시군구 읍면동 조사구특성 (, ' ' 건물층수 , 독립된출입구수 , floor_group, ,
	단독주택의 종류	행정구역코드 시도 시군구 읍면동 부엌수 (건물층수 방수 독립출입구수 , floor_group, , ,
	건물층수	행정구역코드 시도 시군구 읍면동 거처의종류 건축년도 (, ,), ,
연건평 ㎡ ( )		행정구역코드 시도 시군구 읍면동 방수 대지면적 화장실수 , , 건물층수 , floor_group,
대지면적 ㎡ ( )		행정구역코드 시도 시군구 읍면동 연건평 (건물층수,), size1_group, floor_group,
총방수	방수	가구부문 결과의 합으로 대체
	거실수	가구부문 결과의 합으로 대체
	식당수	가구부문 결과의 합으로 대체
건축년도		행정구역코드 시도 시군구 읍면동 건물층수 화장실수 (, 연건평), floor_group, 거처의 종류, 거실수 , size1_group, ,
편익 시설수	부엌수	행정구역코드 시도 시군구 읍면동 방수 화장실수 독립된 출입구수 (,), 연건평 ,
	화장실수	행정구역코드 시도 시군구 읍면동 방수 독립된 출입구수 부엌수 (, ,), 연건평 , size1_group,
	독립된 출입구수	행정구역코드 시도, 시군구, 읍면동, 방수 화장실수 부엌수 (, ,), ,

주 : 대체군의 나열순서는 중요한 항목부터 나열됨

floor_group은 건물층수를 5개 층별로 그룹화한 변수임

size1_group은 연건평 평 단위를 ㎡로 환산하여 5개 평별을 그룹화한 변수임

dae1_group은 대지면적 평 단위를 ㎡로 환산하여 5개 평별을 그룹화한 변수임

〈표 1-16〉 주택 항목들에 대한 대체군 요약(표본)

항목		대체군 (CHAID)
거처의 종류	거처의 종류	행정구역코드 시도 시군구 읍면동 ( ), 단독주택의종류 조사구특성 ', 건물층수 독립된출입구수 ', , floor_group,
	단독주택의 종류	행정구역코드 시도 시군구 읍면동 부엌수 ( ) 건물층수' 방수 독립된 출입구수 ', floor_group, , ,
	건물층수	행정구역코드 시도 시군구 읍면동 거처의종류 (, ,), , 건축년도
연건평 ㎡ ( )		행정구역코드 시도 시군구 읍면동 방수 ( ), , 대지면적, 화장실수 dae1_group, , , floor_group, 건물층수
대지면적 ㎡ ( )		행정구역코드 시도 시군구 읍면동 연건평 (건물층수,), size1_group, , floor_group,
총방수	방수	가구부문 결과의 합으로 대체
	거실수	가구부문 결과의 합으로 대체
	식당수	가구부문 결과의 합으로 대체
건축년도		행정구역코드 시도 시군구 읍면동 건물층수 화장실수 (, , 연건평), floor_group, 거처의종류, 거실수 , size1_group,
편익 시설수	부엌수	행정구역코드' 시도 시군구 읍면동 방수 화장실수 독립된출입구수 (,), 연건평 , size1_group,
	화장실수	행정구역코드 시도 시군구 읍면동 방수 독립된출입구수 (부엌수 ,), 연건평 , size1_group,
	독립된 출입구수	행정구역코드 시도 시군구 읍면동 방수 화장실수 부엌수 (, ,), ,

주 대체군의 나열순서는 중요한 항목부터 나열됨

: 은 건물층수를 층별로 그룹화한 변수임
floor_group 5
은 연건평 평 단위를 ㎡로 환산하여 평별을 그룹화한 변수임
size1_group 5
은 대지면적 평 단위를 ㎡로 환산하여 평별을 그룹화한 변수임
dae1_group 5

제4절 모의실험을 통한 대체군의 효율성 비교

앞서 개발한 대체군들을 이용하여 가구원, 가구 및 주택에 관한 각 항목들에 대한 모의실험을 실시한다. 모의실험에서 사용할 2005년 인구주택총조사 자료는 전수조사 가구 15,988,274가구 중 10%인 1,756,609가구, 표본조사 가구 1,591,631가구 중 159,954가구이다. 실험을 위해 각 항목마다 임의로 5%의 결측치를 발생시켰고, 상세한 결측치 추출 현황은 <표 1-17>과 같다. 모의실험에서 사용한 항목 대체방법은 Hierarchical Hot-Deck방법이다.

카이제곱 검정의 연관성 분석 및 CHAID 알고리즘에 의해 대체군을 결정하고 모의실험을 실시하여 각 대체군들의 효율성을 검토한다. 효율성 측정에 있어서 범주형 자료는 원자료를 얼마나 정확하게 대체시키는지, 대체로 인한 분포의 왜곡문제는 없는지 살펴보며, 연속형 자료는 평균의 차이가 없는지를 측정하고자 한다.

〈표 1-17〉 결측치 추출현황

항목	전수자료			표본자료		
	원자료	모의실험 (10%)	결측치 (5%)	원자료	모의실험 (10%)	결측치 (5%)
가구원						
가구	46,062,984	4,926,365	246,400	5,042,490	504,341	25,300
주택	15,988,274	1,756,609	87,900	1,591,631	159,954	8,000
	15,988,274	1,756,609	87,900	1,591,631	159,954	8,000

1. 가구원에 관한 사항

가. 성명, 산업, 직업

성명, 산업, 직업은 직접 기입하는 주관식 항목이므로 무응답 대체가 필요 없다.

나. 성별

〈표 1-18〉 성별에 대한 대체결과

정확도 (%) \ 대체군	카이제곱 검정 (A)	알고리즘 CHAID (B)	차이 (B-A)
전수	69.8	70.2	0.4
표본	77.4	78.3	0.9

대체 전후의 분포변화 표본 ( )

실제 인원수	남	여
카이제곱	12,542	12,758
	12,663	12,637
CHAID	12,632	12,668

성별에 대한 무응답 처리를 할 때 [방법A]의 정확도는 전수자료 69.8%, 표본자료 77.4%이고, [방법B]처럼 대체군을 변경할 경우 전수 0.4%, 표본 0.9% 증가를 보이고 있지만 증가 효과는 미미하다. [방법A]와 [방법B]의 대체군 변경으로 인한 정확도의 차이는 거의 없다고 볼 수 있다. 대체 전후 항목분포의 변화를 살펴보면 [방법B]가 실제 인원수인 12,542명, 여자 12,758명에서 남자 12,632명, 여자 12,668명으로 대체하여 분포의 왜곡이 더 적은 편이라 할 수 있다.

다. 나이

〈표 1-19〉 만나이에 대한 대체결과

대체전평균		카이제곱검정 (A)	알고리즘 CHAID (B)	차이 (B-A)
전수	평균	34.55	34.55	-
	절대차이 오차비율	34.38	34.29	-0.09
	대체전평균 (평균)	0.17 (0.50%)	0.26 (0.76%)	0.09 (0.26%)
표본	평균	35.61	35.61	-
	절대차이 오차비율	35.60	35.49	-0.11
	( )	0.01 (0.02%)	0.12 (0.33%)	0.11 (0.31%)

대체 전후의 분포변화 표본 ( )

구성비 (%)	0-4	5-9	10-14	15-19	20-24	25-29	30-34	35-39	40-44
실제 인원수 분포변화비율 (%)	1230 (4.86)	1575 (6.23)	1703 (6.73)	2160 (8.54)	1981 (7.83)	1996 (7.89)	2083 (8.23)	2039 (8.06)	2030 (8.03)
카이제곱 검정									
대체후 인원수 분포변화비율 (%)	1230 (4.86)	1575 (6.23)	1704 (6.74)	2166 (8.56)	1972 (7.80)	1987 (7.86)	2067 (8.17)	2043 (8.08)	2056 (8.13)
절대차이	0	0	1	6	9	9	16	4	26
알고리즘 CHAID									
대체후 인원수 분포변화비율 (%)	1230 (4.86)	1575 (6.23)	1704 (6.74)	2163 (8.55)	1991 (7.87)	1994 (7.88)	2088 (8.25)	2054 (8.12)	2091 (8.27)
절대차이	0	0	1	3	10	2	5	15	61

대체 전후의 분포변화 표본 계속 ( )

구성비 (%)	45-49	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85-
실제 인원수 분포변화비율 (%)	2018 (7.98)	1518 (6.00)	1232 (4.87)	1067 (4.22)	1039 (4.11)	686 (2.71)	463 (1.83)	292 (1.15)	183 (0.72)
카이제곱 검정									
대체후 인원수 분포변화비율 (%)	2035 (8.05)	1515 (5.99)	1252 (4.95)	1053 (4.16)	1029 (4.07)	690 (2.73)	456 (1.80)	274 (1.08)	191 (0.76)
절대차이	17	3	20	14	10	4	7	18	8
알고리즘 CHAID									
대체후 인원수 분포변화비율 (%)	2037 (8.05)	1505 (5.95)	1239 (4.90)	1027 (4.06)	1002 (3.96)	679 (2.68)	452 (1.79)	272 (1.08)	192 (0.76)
절대차이	19	13	7	40	37	7	11	20	9

만나이에 대한 무응답 처리를 할 때 [방법A]의 평균의 오차비율은 전수 0.50%, 표본 0.02%이고, [방법B]처럼 대체군을 변경할 경우 전수 0.26%, 표본 0.31% 증가를 보이고 있지만 증가 효과는 미미하다. [방법A]와 [방법B]의 대체군 변경으로 인한 평균의 차이는 거의 없다고 볼 수 있다. 여기서 오차비율은 (|실제 평균 - 대체후 평균|/실제 평균)의 백분율을 한 것이다. 만나이는 연속형 값을 가지므로 범주화를 하여 대체 전후의 분포의 변화를 보고자 한다. 대체 전후의 분포의 변화를 살펴보면 [방법A]와 [방법B]가 비슷한 분포를 가지며 실제 분포 비율과도 큰 차이가 없음을 알 수 있다.

라. 가구원 항목에 대한 대체결과

<표 1-20>은 전수 및 표본조사의 가구원 항목에 대한 대체결과이다. 교육정도와 5년 전 거주지 행정구역코드 항목이 타 가구원 항목의 정확도에 비해 다소 낮다고 볼 수 있다.

〈표 1-20〉 가구원 항목에 대한 대체결과

대체군		전수			표본		
정확도 (%)	생월	카이제곱 검정 (A)	알고리즘 CHAID (B)	차이 (B-A)	카이제곱 검정 (A)	알고리즘 CHAID (B)	차이 (B-A)
만 나이	평균	69.8	70.2	0.4	77.4	78.3	0.9
	절대차이	34.38	34.29	-0.09	35.60	35.49	-0.11
	오차비율	0.17	0.26	0.09	0.01	0.12	0.11
가구주와의 관계		(0.50)	(0.76)	(0.26)	(0.02)	(0.33)	(0.31)
교육정도		89.5	89.7	0.2	86.6	84.5	-2.1
교육상태		47.8	52.2	4.4	46.9	54.1	7.2
종교여부		88.7	89.1	0.4	89.9	89.3	-0.6
종교종류		97.6	97.6	-	97.4	97.4	-
남북이산가족		66.5	66.9	0.4	70.7	70.3	-0.4
남북이산가족 출생지		99.5	99.5	-	99.3	99.3	-
아동보육		99.5	98.9	-0.6	99.4	98.6	-0.8
어머니 동거여부		x	x	x	93.2	93.7	0.5
1년 전 거주지		x	x	x	98.2	98.5	0.3
5년 전 거주지 행정구역코드		x	x	x	69.5	59.9	-9.6
시각·청각		x	x	x	21.0	24.0	3.0
활동 제약 여부	치매	x	x	x	95.7	95.7	-
	중풍				96.5	96.5	-
	육체적제약				96.4	96.4	-
	정신적제약				92.2	92.2	-
	없음				94.7	94.7	-
활동 제약 종류	배우기	x	x	x	99.1	99.1	-
	옷입기				94.1	94.1	-
	쇼핑				95.5	95.5	-
	취업활동				93.3	93.3	-
	없음				94.8	94.8	-
					99.2	99.2	-

〈표 1-20〉 가구원 항목에 대한 대체결과_계속

대체군 / 정확도 (%)	전수			표본		
	카이제곱 검정 (A)	CHAID 알고리즘 (B)	차이 (B-A)	카이제곱 검정 (A)	CHAID 알고리즘 (B)	차이 (B-A)
통근·통학 여부	x	x	x	96.2	94.2	-2.0
통근·통학 장소	x	x	x	93.5	93.4	-0.1
통근·통학 장소 행정구역코드	x	x	x	91.6	91.4	-0.2
이용교통수단	x	x	x	76.5	75.5	-1.0
통근·통학 소요시간	평균	x	x	x	25.56	24.18	-1.38
	절대차이 오차비율 (%)	x	x	x	0.08 (0.30)	1.46 (5.70)	1.38 (5.40)
취업여부	x	x	x	97.6	97.9	0.3
구직여부	x	x	x	98.8	80.8	-18.0
취업가능성	x	x	x	97.8	95.5	-2.3
종사상 지위	x	x	x	83.3	85.4	2.1
근로장소	x	x	x	89.2	89.1	-0.1
혼인상태	80.3	83.6	3.3	76.9	86.5	9.6
혼인년도	평균	x	x	x	1979.09	1978.84	-0.25
	절대차이 오차비율 (%)	x	x	x	0.94 (0.05)	1.19 (0.06)	0.25 (0.01)
남자 출생아수	x	x	x	73.5	72.5	-1.0
여자 출생아수	x	x	x	73.3	71.2	-2.1
추가계획 자녀여부	x	x	x	90.9	94.7	3.8
추가계획자녀수	x	x	x	98.8	98.8	-
고령자 생활비 원천	x	x	x	90.8	92.3	1.5
외국인여부	99.9	99.8	-0.1	99.9	99.9	-
외국인 교육정도	99.6	99.6	-	100.0	100.0	-
외국인 국적	99.5	99.4	-0.1	99.9	99.9	-
외국인주된 체류목적	99.8	99.7	-0.1	100.0	100.0	-
외국인 직업	99.6	99.6	-	100.0	100.0	-
외국인 거주기간	99.7	99.6	-0.1	100.0	100.0	-

주 혼인월과 양·음력 구분은 혼인년도와 함께 대체한다

: .

2. 가구에 관한 사항

가. 가구 구분

〈표 1-21〉 가구 구분에 대한 대체결과

정확도 (%) \ 대체군	카이제곱 검정 (A)	알고리즘 CHAID (B)	차이 (B-A)
전수	95.1	95.1	-
표본	95.9	96.1	0.2

대체 전후의 분포변화 표본 ( )

실제 가구수	가족으로 이루어진 가구	가족과 가족이외의 사람이함께 사는가구	인가구 1	가족이아닌 남남끼리함께사는 인 이하의가구5	가족이아닌 남남끼리함께사는 인 이상의가구6
카이제곱	6,206	27	1,628	100	6
	6,216	30	1,674	76	1
CHAID	6,221	22	1,668	84	2

대체 전후의 분포변화 표본 계속 ( )

실제 가구수	기숙사에 살고있는 집단가구	보육원등 아동복지 시설에 살고있는 집단가구	양로원등 노인복지 시설에 살고있는 집단가구	장애인 복지시설에 살고있는 집단가구	기타 복지시설에 살고있는 집단가구
카이제곱	21	4	3	2	3
	3	0	0	0	0
CHAID	3	0	0	0	0

가구 구분에 대한 무응답 처리를 할 때 [방법A]는 전수 95.1%, 표본 95.9%의 정확도를 보이고, [방법B]처럼 대체군을 변경할 경우는 전수 0%, 표본 0.2% 증가를 보이고 있지만 증가 효과는 미미하다. [방법A]와 [방법B]의 대체군 변경으로 인한 정확도의 차이는 거의 없다고 볼 수 있다.

나. 거주 층수

〈표 1-22〉 거주 층수에 대한 대체결과

		카이제곱 검정 (A)	알고리즘 CHAID (B)	차이 (B-A)
전수	대체전평균	3.49	3.49	-
	평균	3.89	3.85	-0.04
	절대차이	0.40	0.36	-0.04
	오차비율	(11.6%)	(10.4%)	(-1.2%)
	(대체전평균)			
표본	평균	4.17	4.17	-
	절대차이	4.42	4.41	-0.01
	오차비율	0.25	0.24	-0.01
	( )	(6.08%)	(5.76%)	(-0.32%)

대체 전후의 분포변화 표본 ( )

구성비 (%)	0-4	5-9	10-14	15-19	20-24	25-29	30-34	35-39	해당사항없음
실제 가구수									
분포변화비율	5506	1058	703	301	72	10	3	0	347
(%)	(68.83)	(13.23)	(8.79)	(3.76)	(0.9)	(0.13)	(0.04)	(-)	(4.34)
대체후 가구수									
분포변화비율	5419	1403	741	313	69	14	2	1	38
(%)	(67.74)	(17.54)	(9.26)	(3.91)	(0.86)	(0.18)	(0.03)	(0.01)	(0.48)
절대차이	87	345	38	12	3	4	1	1	309
대체후 가구수									
분포변화비율	5428	1398	738	314	69	14	1	0	38
(%)	(67.85)	(17.48)	(9.23)	(3.93)	(0.86)	(0.18)	(0.01)	(-)	(0.48)
절대차이	78	340	35	13	3	4	2	0	309

거주 층수에 대한 무응답 처리를 할 때 [방법A]의 평균의 오차비율은 전수 11.6%, 표본 6.08%이고, [방법B]처럼 대체군을 변경할 경우 전수 1.2%, 표본 0.32% 감소를 보이고 있지만 감소 효과는 미미하다. [방법A]와 [방법B]의 대체군 변경으로 인한 평균의 차이는 거의 없다고 볼 수 있다. 여기서 오차비율은 (|실제 평균 - 대체후 평균|/실제 평균)의 백분율을 한 것이다.

다. 가구 항목에 대한 대체결과

<표 1-23>은 전수 및 표본조사의 가구 항목에 대한 대체결과이다. 거주 기간 항목이 타 가구 항목의 정확도에 비해 다소 낮다고 볼 수 있다.

〈표 1-23〉 가구 항목에 대한 대체결과

대체군 / 정확도 (%) / 가구 구분		전수			표본		
		카이제곱 검정 (A)	CHAID 알고리즘 (B)	차이 (B-A)	카이제곱 검정 (A)	CHAID 알고리즘 (B)	차이 (B-A)
거주 기간		95.1	95.1	-	95.9	96.1	0.2
침실수		x	x	x	26.8	29.3	2.5
침실이외의 방수		65.2	68.4	3.2	55.5	60.7	5.2
거실수		79.1	83.7	4.6	70.2	71.2	1.0
식당수		83.3	84.6	1.3	90.4	91.3	0.9
부엌형태		84.5	85.5	1.0	82.5	83.7	1.2
부엌 사용여부		97.7	97.7	-	96.7	96.8	0.1
수도형태		99.5	99.2	-0.3	99.9	99.7	-0.2
화장 형태		95.1	94.5	-0.6	91.5	90.6	-0.9
화장실 사용여부		95.6	95.8	0.2	91.3	91.6	0.3
목욕시설형태		97.3	97.3	-	99.5	99.5	-
목욕시설 사용여부		99.3	99.2	-0.1	98.6	98.6	-
거주층 구분		99.2	97.0	-2.2	99.5	96.6	-2.9
거주층수	평균	98.2	98.1	-0.1	99.2	99.2	-
	절대차이	3.89	3.85	-0.04	4.42	4.41	-0.01
	오차비율 (%)	0.40 (11.59)	0.36 (10.37)	-0.04 (-1.22)	0.25 (6.08)	0.24 (5.76)	-0.01 (-0.32)

〈표 1-23〉 가구 항목에 대한 대체결과_계속

대체군 / 정확도 (%)	전수			표본		
	카이제곱 검정 (A)	CHAID 알고리즘 (B)	차이 (B-A)	카이제곱 검정 (A)	CHAID 알고리즘 (B)	차이 (B-A)
승용차 보유대수	x	x	x	77.8	78.5	0.7
승합차 보유대수	x	x	x	95.8	95.6	-0.2
화물 및 기타자동차 보유대수	x	x	x	88.1	89.2	1.1
자동차 보유여부	x	x	x	97.9	97.0	-0.9
주차 시설	x	x	x	86.2	87.1	-0.9
난방 시설	77.5	79.2	1.7	77.2	80.4	3.2
주거영업구분	94.2	94.5	0.3	95.4	95.6	0.2
점유형태	61.3	62.2	0.9	84.3	84.1	-0.2
사글세 개월수	x	x	x	99.4	99.5	0.1
전세·보증금	평균 절대차이	x	x	x	3368.99	3214.97	-154.02
	오차비율 (평균 %)				34.19 (1.03)	119.83 (3.59)	85.64 (2.56)
월세 사글세 ( ) 주인가구 여부	평균 절대차이	x	x	x	21.97	21.24	-0.73
	오차비율 (%)				0.05 (0.23)	0.67 (3.07)	0.62 (2.84)
타지 주택 소유	94.1	94.1	-	91.2	91.4	0.2
	86.0	86.0	-	81.7	81.3	-0.4

3. 주택에 관한 사항

가. 거처의 종류

〈표 1-24〉 거처의 종류에 대한 대체결과

대체군

정확도 (%)	카이제곱 검정 (A)	알고리즘 CHAID (B)	차이 (B-A)
전수	94.4	93.3	-1.1
표본	94.9	94.2	-0.7

대체 전후의 분포변화 표본 ( )

실제 주택수	단독주택	아파트	연립주택	다세대주택	비주거용 건물내주택
카이제곱	2,343	3,153	226	542	123
	2,305	3,163	225	552	86
CHAID	2,289	3,199	212	543	69

대체 전후의 분포변화 표본 계속 ( )

실제 주택수	오피스텔	호텔 여관등숙박업소의객실	기숙사및특수사회시설	판잣집 비닐하우스 움막	기타	해당 없음
카이제곱	63	5	45	15	9	1,476
	54	0	12	12	5	1,586
CHAID	52	0	16	12	4	1,604

거처의 종류에 대한 무응답 처리를 할 때 [방법A]는 전수 94.4%, 표본 94.9%의 정확도를 보이고, [방법B]처럼 대체군을 변경할 경우는 전수 1.1%, 표본 0.7% 감소를 보이고 있지만 감소 효과는 미미하다. [방법A]와 [방법B]의 대체군 변경으로 인한 정확도의 차이는 거의 없다고 볼 수 있다.

〈표 1-25〉 단독주택의 종류에 대한 대체결과

434

정확도 (%)	카이제곱 검정 (A)	알고리즘 CHAID (B)	차이 (B-A)
전수	94.4	96.1	1.7
표본	94.7	95.9	1.2

대체 전후의 분포변화 표본 ( )

실제 주택수	일반단독주택	다가구 단독주택	영업겸용 단독주택	해당없음
카이제곱	1,903	335	97	5,665
	1,905	329	90	5,676
CHAID	1,940	305	76	5,679

단독주택의 종류에 대한 무응답 처리를 할 때 [방법A]는 전수 94.4%, 표본 94.7%의 정확도를 보이고, [방법B]처럼 대체군을 변경할 경우는 전수 1.7%, 표본 1.2% 증가를 보이고 있지만 증가 효과는 미미하다. [방법A]와 [방법B]의 대체군 변경으로 인한 정확도의 차이는 거의 없다고 볼 수 있다.

나. 연건평

〈표 1-26〉 연건평(주거용 면적)에 대한 대체결과

대체전평균		카이제곱 검정 (A)	알고리즘 CHAID (B)	차이 (B-A)
전수	평균	80.35	80.35	-
	절대차이	78.54	77.65	-0.89
	오차비율 대(체전평균)	1.80 (2.24%)	2.70 (3.36%)	0.9 (1.12%)
표본	평균	81.49	81.49	-
	절대차이	77.64	76.15	-1.49
	오차비율 ( )	3.85 (4.72%)	5.33 (6.55%)	1.48 (1.83%)

대체 전후의 분포변화 표본

면적 ㎡ ( )

실제 주택수 분포변화비율 (%)	-16	17-32	33-49	50-65	66-82	83-98	99-115
	0 (-)	215 (2.69)	735 (9.19)	1806 (22.58)	994 (12.43)	1424 (17.8)	452 (5.65)
대체후 주택수 분포변화비율 절대차이 (%)	2 (0.03)	218 (2.73)	748 (9.35)	1881 (23.51)	1006 (12.58)	1483 (18.54)	420 (5.25)
	2	3	13	75	12	59	32

카이제곱 검정

알고리즘 CHAID							
대체후 주택수	2	204	773	1804	1007	1523	404
분포변화비율 (%)	(0.03)	(2.55)	(9.66)	(22.55)	(12.59)	(19.04)	(5.05)
절대차이	2	11	38	2	13	99	48

대체 전후의 분포변화 표본 계속 ( )

면적 ㎡ ( )							
실제 주택수	116-131	132-148	149-164	165-181	182-197	198-214	215-320
분포변화비율 (%)	244 (3.05)	176 (2.20)	90 (1.13)	77 (0.96)	42 (0.53)	65 (0.81)	24 (0.30)
대체후 주택수							
분포변화비율 (%)	217 (2.71)	172 (2.15)	78 (0.98)	69 (0.86)	44 (0.55)	38 (0.48)	18 (0.23)
절대차이	27	4	12	8	2	27	6
대체후 주택수							
분포변화비율 (%)	233 (2.91)	162 (2.03)	79 (0.99)	73 (0.91)	29 (0.36)	34 (0.43)	14 (0.18)
절대차이	14	14	1	4	13	31	10

대체 전후의 분포변화 표본 계속 ( )

면적 ㎡ ( )								
실제 주택수	231-247	248-263	264-280	281-297	298-313	314-330	331-	해당없음
분포변화비율 (%)	32 (0.40)	12 (0.15)	12 (0.15)	12 (0.15)	10 (0.13)	15 (0.19)	44 (0.55)	1519 (18.99)
대체후 주택수								
분포변화비율 (%)	15 (0.19)	11 (0.14)	8 (0.105)	10 (0.13)	2 (0.03)	15 (0.19)	23 (0.29)	1522 (19.03)
절대차이	17	1	4	2	8	0	21	3
대체후 주택수								
분포변화비율 (%)	15 (0.19)	5 (0.06)	4 (0.05)	4 (0.05)	1 (0.01)	10 (0.13)	7 (0.09)	1613 (20.16)
절대차이	17	7	8	8	9	5	37	94

연건평(주거용 면적)에 대한 무응답 처리를 할 때 [방법A]의 평균의 오차비율은 전수자료 2.24%, 표본자료 4.72%이고, [방법B]처럼 대체군을 변경할 경우 전수 1.12%, 표본 1.83% 증가를 보이고 있지만 증가 효과는 미미하다. [방법A]와 [방법B]의 대체군 변경으로 인한 평균의 차이는 거의 없다고 볼 수 있다. 여기서 오차비율은 (|실제 평균 - 대체후 평균|/실제 평균)의 백분율을 한 것이다.

다. 주택 항목에 대한 대체결과

<표 1-27>은 전수 및 표본조사의 주택 항목에 대한 대체결과이다. 건축년도 항목이 타 주택 항목의 정확도에 비해 다소 낮다고 볼 수 있다.

〈표 1-27〉 주택 항목에 대한 대체결과

대체군 / 정확도 (%)		전수			표본		
		카이제곱 검정 (A)	CHAID 알고리즘 (B)	차이 (B-A)	카이제곱 검정 (A)	CHAID 알고리즘 (B)	차이 (B-A)
거처의 종류		94.4	93.3	-1.1	94.9	94.2	-0.7
단독주택의 종류		94.4	96.1	1.7	94.7	95.9	1.2
건물 층수	평균 절대차이	9.16	9.12	-0.04	8.52	8.36	-0.16
	오차비율 (평균 %)	0.05 (0.60)	0.01 (0.16)	-0.04 (-0.44)	0.21 (2.56)	0.05 (0.62)	-0.16 (-1.94)
연건평	절대차이	78.54	77.65	-0.89	77.64	76.15	-1.49
	오차비율 (평균 %)	1.80 (2.24)	2.70 (3.36)	0.9 (1.12)	3.85 (4.72)	5.33 (6.55)	1.48 (1.83)
대지 면적	절대차이	269.12	272.74	3.62	279.30	286.69	7.39
	오차비율 (평방수 %)	4.78 (1.75)	1.17 (0.43)	-3.61 (-1.32)	3.13 (1.11)	4.26 (1.51)	1.13 (0.40)
거실수 대청마루		89.1	89.9	0.8	78.4	78.4	-
식당수(부엌이)		97.1	97.0	-0.1	98.4	98.4	-
딸린 식당 포함 (건축년도)		96.7	96.7	-	99.1	99.1	-
부엌수		52.5	58.2	5.7	61.2	67.4	6.2
화장실수		95.8	94.1	-1.7	94.7	92.7	-2.0
독립된 출입구수		65.7	82.1	16.4	79.2	84.2	5.0
		94.5	92.5	-2.0	93.7	91.4	-2.3

4. 정확도 범위 현황

〈표 1-28〉 모의실험 실시 후 정확도 범위 현황(CHAID 알고리즘 기준)

표본조사 항목	가구원	가구	주택	합계
총 세부항목수	44	25	9	78
정확도 범위	24.0% ~ 100.0%	29.3% ~ 99.7%,	67.4% ~ 99.1%	24.0% ~ 100.0%
90% 이상 ~ 이상 미만	30	15	6	51
80 ~ 90% 이상 미만	6	6	1	13
70 ~ 80% 이상 미만	5	2	1	8
60 ~ 70% 이상 미만	0	1	1	2
50 ~ 60% 이상 미만	2	0	0	2
40 ~ 50% 이상 미만	0	0	0	0
30 ~ 40% 이상 미만	0	0	0	0
20 ~ 30%	1	1	0	2

주 : 연속형 자료의 항목은 제외

항목 중 가구원번호는 가구원항목의 무응답 대체에 있어서 중요한 부분으로 나타났다. 따라서 표지항목인 가구원 번호를 기입할 때는 가구주와의 관계의 순으로 기입순서를 지켜주는 것이 좋다고 볼 수 있다. 또한 50% 미만의 정확도를 가지는 항목이 5년 전 거주지 행정구역코드 24.0%, 거주 기간 29.3%이다. 이러한 항목들의 정확도를 향상시키기 위해서는 연관성 높은 항목의 추가 개발이 필요하다고 본다.

제5절 결론 및 제안

본 연구에서는 2005년 인구주택총조사 항목에 대하여 카이제곱 검정 및 CHAID 알고리즘을 통하여 대체군을 개발하였다. 그리고 범주형과 연속형 항목에 관계없이 Hierarchcal Hot-Deck방법으로 모의실험을 실시하여 대체 후의 정확성을 비교해 보았다. 만나이 및 거주층수 등과 같은 연속형 변수는 그룹화 하여 범주형 변수로 변경하여 모의실험을 하였다.

본 연구의 모의실험에서는 2005년 인구주택총조사 전수조사 자료는 1,756,609 가구를 추출하여 이용하였고 표본조사 자료는 159,954 가구를 추출하였다. 실험을 위해 임의로 5%의 결측치를 발생시켜 전수 87,900 가구, 표본 8,000가구의 무응답을 발생시켰다. 그리고 항목별로 무응답 대체를 모두 하고 난 뒤 대체의 정확도, 평균의 차이, 분포비율, 오차비율 등을 살펴보았다. 연속형 변수의 평균의 차이는 적은 것으로 나타났다. 대체 후의 정확도 측면에서는 항목마다 차이를 보이고 있다. 몇 항목들은 다른 항목에 비해서 정확도가 조금 낮으나 큰 문제는 되지 않는 것으로 판단되며, 대체후의 분포변화도 거의 나타나지 않음을 알 수 있다. 모든 항목들에 대한 자세한 모의실험 결과는 제4절을 참조하기 바란다.

본 연구의 결과를 고려하여 2010년 인구주택총조사를 위해서 다음과 같은 제안을 한다.

첫째, 대체군은 CHAID 알고리즘을 이용하여 개발한 대체군을 사용하도록 한다. 모의실험 결과 정확도 측면에서는 카이제곱검정과 CHAID 알고리즘을 이용한 대체군이 큰 차이가 없으므로 대체군의 수가 적은 CHAID 알고리즘을 이용하는 것이 바람직하다고 할 것이다. 그러나 구직여부와 같이 대체군 1개의 항목이 정확도에 큰 영향을 미치는 경우는 그 1개의 항목이 존재하지 않을 경우 정확도가 낮아지게 된다. 이러한 경우는 예외적으로 카이제곱 검정의 대체군을 이용하여 정확도를 향상시키는 것이 바람직하다. 2010년 인구주택총조사의 항목에서도 본 연구결과인 CHAID 알고리즘을 이용한 대체군을 사용하도록 한다.

각 항목에 대한 대체군은 제3절의 <표 1-11>에서 <표 1-16>까지 참조하기 바란다. 2010년 인구주택총조사와 2005년 인구주택총조사의 중복되는 조사 항목에 대해서는 본 연구결과인 대체군과 대체방법을 적용하고, 2010년에 추가되는 항목에 대해서는 최종 시험조사 자료를 이용하여 대체군을 개발하도록 한다.

둘째, 항목 대체방법에 있어서는 여러 가지 방법을 사용하기보다는 확률적 대체방법 중 Hierarchical Hot-Deck 방법을 사용하는 것이 타당하다고 본다. 비복원을 가정한다면 무응답이 응답보다 많을 경우는 대체하지 못하는 경우가 발생하므로 Hierarchical Hot-deck을 사용하는 것이 좋다고 본다. 정확성 측면에서 Probability, Hot-deck, Hierarchical Hot-deck은 큰 차이를 보이지 않았기 때문이다. 농업총조사에서 사용하는 응용Hot-deck 대체방법은 연속형 자료일 경우 유용하다고 할 수 있다. 인구주택총조사 자료는 대규모이면서 범주형자료로 이루어지기 때문에 Hierarchical Hot-deck과 응용Hot-deck 모두 타당하므로 어떤 방법을 사용하여도 무관하다고 본다.

마지막으로, 모의실험 결과 정확도의 범위를 고려할 때 무응답 대체를 하는 것이 가능하다고 본다. 전수자료의 정확도 범위는 가구원 항목이 47.8% ~ 99.9%이고, 표본자료의 정확도 범위는 21.0% ~ 100.0%이다. 본 연구에서 모의실험결과 정확도 50% 이상이 총 78개 세부항목 중 76개로 약 97.4%이다. 그러나 실제 무응답을 대체하였을 경우는 이 결과보다는 정확도가 떨어질 수 있다. 그럼에도 불구하고 본 연구에서 제시된 대체군과 대체방법은 인구주택총조사 무응답 대체를 위한 하나의 효율적인 방법임에 틀림없을 것으로 확신한다.

응답거부 등 조사환경이 갈수록 악화됨에 따라 그 변화에 능동적으로 대처하여 고품질의 통계작성을 위한 연구를 수행하는 것이 점차 필요하게 되었다. 이러한 측면에서 본 연구는 인구주택총조사의 품질을 향상시키는 데 많은 도움이 될 것이라고 생각하며, 또한 많이 기여할 수 있게 되기를 기대한다.

참고문헌

김영원・이주원 (2003), “CART를 활용한 결측값 대체방법 인구주택총조사 혼인상태 항목을 중심으로”, 「조사연구」 제 권 제 호, pp.1-21.

김영원・조선경 (1996), “표본조사에서 항목 무응답 대체 방법”, 「한국통계학회논문집」 제 권 제 호, pp.145-159.

김재광・한근식・윤연옥 (2004), “가계조사 무응답 대체기법 연구”, 통계청 「통계연구」 제 권 제 호, pp.79-102.

김진 (2004), “농가경제조사에 대한 대체법 비교”, 통계청 「통계연구」 제 권 제 호, pp.133-145.

송순관 (2005), “무응답 처리방법 연구 및 읍・면・동 통계작성 가능성 검토”,

이기성・한성실・정기문 (2007), “한글 통계자료분석 자유아카데미 이내성 회귀분석을 이용한 SPSS 기법활용 연구”, 통계청 (2008), “국가통계의 품질제고를 위한 Imputation 방법론 연구” 년, 제 호,

, 2008 2 , pp.44-105.

이재원 (2000), “무응답 및 오류자료의 적용 결과 「무응답오차 조사통계연구회 Imputation 자유아카데미”, pp.134-145.

이현정 (2009), “인구주택총조사 무응답 대체기법 연구Ⅰ. 통계청 「무응답 처리를 위한 방법론 연구」 년 제 호”,

조사통계연구회 (2007), 무응답 오차 자유아카데미 2 , pp.243-322.

최필근 (2008), “농가조사 항목 간 연관성 분석 및 대체군 보조변수 개발”, 통계청 「무응답 처리를 위한 방법론 연구」 (년 제 호)

”, , , 2009 2 , pp.4-141.

통계청 (2006), “2005 인구주택총조사 종합평가보고서 내부자료”.

통계청 (2006), “ 제 차 시험조사 조사지침서 내부자료”.

(2008), “ 2 ”.