2011년도 국가통계 품질개선지원 연구용역

『국민건강영양조사』
품질개선지원 최종결과보고서

- 무응답 현황분석 및 대체 -

2011. 10.

통계청 STATISTICS KOREA

제 출 문

제 출 문

통계청장 귀하

본 보고서를 『국민건강영양조사』 품질개선지원 연구용역 과제의 최종 연구결과물로 제출합니다.

[개선지원 부문]: 무응답 현황분석 및 대체

2011년 10월 30일

청주대학교 바이오정보통계학과 교수 류제복 ㊞

연구진

책 임 연 구 원 류 제 복 (청주대학교 교수)

연 구 원 이 승 주 (청주대학교 교수)

요 약 문

최종결과보고서 요약문

연구과제명	「국민건강영양조사」 무응답 현황분석 및 대체
주 제 어	항목무응답, 연관성분석, 무응답대체군, 무응답대체
연 구 기 간	2011.06.27 ~ 2011.10.30 (4개월)
연 구 기 관	
연구진구성	류제복, 이승주

보건복지부에서 생산하고 있는 “국민건강영양조사”는 국민의 건강수준, 건강관련 의식 및 행태, 식품 및 영양섭취 실태를 파악하기 위한 조사이다. 이 통계는 2010년도 정기품질진단결과 품질이 우수한 것으로 평가되었고, 관련분야에서 널리 사용되고 있다. 그러나 항목무응답에 대한 조치가 미흡하여 이에 대한 개선지원을 하게 되었다. 본 연구에서 다룬 내용을 요약하면 다음과 같다.

○ 현행 국민건강영양조사의 무응답현황을 단위무응답과 항목무응답으로 분류하여 살펴보았다
○ 무응답 대체방법을 단위무응답과 항목무응답으로 나누어 소개하고, 이는 향후 무응답대체 시 활용할 수 있다.
○ 무응답을 대체할 8개의 목표변수와 관련된 보조변수들에 대한 특성을 분석하고 대체기준을 설정하였다.
○ 무응답대체군 설정을 위해서 CHAID 알고리즘을 사용하여 8개 목표변수와 관련 보조변수들 간의 연관성분석을 실시하였다.
○ 8개 변수 중 대체 가능한 2개 목표변수에 대한 대체를 실시하고 그 결과를 분석하였다.

본 연구에서 다룬 항목무응답에 대한 대체과정은 향후 증가할 것으로 예상되는 “국민건강영양조사”의 항목무응답 처리에 도움을 주어 전반적으로 통계품질향상에 기여할 것으로 기대된다.

최종결과보고서

목 차

- 표 목 차 -

- 그 림 목 차 -

1장. 서론

1절. 연구배경

보건복지부에서 생산하고 있는 “국민건강영양조사”는 2010년 통계청의 정기품질진단에서 우수한 통계로 평가받았다. 다만, 무응답현황에 대한 분석 및 대체방안 마련 등 몇 가지 사항에 대한 보완이 필요하다는 지적을 받은 바 있다. 한편 통계청 품질관리과는 통계작성기관을 대상으로 한 2011년도 Consulting 수요조사에서 본 통계를 지원 대상 통계로 선정하고 보완사항들에 대한 개선지원을 하게 되었다.

2절. 연구목적 및 필요성

“국민건강영양조사”의 무응답 특성분석 및 개선방안을 마련하기 위해서 살펴본 통계의 현황 및 연구목적과 필요성은 다음과 같다.

□ **현황**

- ○ “국민건강영양조사”의 단위무응답이 약 20%수준이고 항목무응답은 약 5% 수준임
- ○ 단위무응답은 일반적으로 널리 사용되는 가중치 조정방법을 사용하고 있으나 항목무응답의 경우 특별한 조치를 취하고 있지 않음
- ○ 항목무응답에 대한 특별한 조치를 취하지 않은 상태로 마이크로자료를 제공하고 있음
- ○ 전문가 자문회의 등에서 무응답에 대한 조치를 요구하고 있는 실정임

□ 연구목적 및 필요성

○ 무응답을 무시하고 응답자료 만을 사용할 경우 추정 결과가 편향되어 조사의 신뢰성이 떨어짐

○ 무응답에 대한 조치를 취하지 않은 불완전자료(incomplete data)는 자료 활용에 제한을 받아 통계분석이 불충분하게 됨

○ 무응답에 대한 적절한 대체를 하여 완전자료(complete data)를 만들면 자료의 손실을 줄여 보다 더 정확한 추정을 할 수 있을 뿐만 아니라 기존의 분석방법을 사용할 수 있음

○ 따라서 본 연구에서는 “국민건강영양조사”에서 발생하는 무응답의 특성을 파악하고 무응답대체 과정을 상세히 제시하며 대체 후 결과를 분석하여 향후 무응답대체 시 도움을 주고자 함

3절. 연구내용 및 방법

본 연구의 내용은 “국민건강영양조사”에서 발생하는 항목무응답의 현황을 분석하고 일반적으로 사용하고 있는 대체방법론을 제시하는 범위로 제한하였다. 2009년 “국민건강영양조사”는 건강설문조사(630개 문항), 영양조사(132개 문항), 검진조사(79개 문항) 등 3개의 조사로 구성되며, 2009년의 조사완료 표본수는 10,533명이다. 따라서 항목무응답 현황과 조사의 중요도 및 관심도 등을 종합적으로 고려하여 질병관리본부와 무응답을 대체할 변수를 선정한다. 한편 연구방법은 “국민건강영양조사”와 관련된 조사결과보고서, 품질진단결과보고서, 무응답대체와 관련 자료 등에 관한 문헌연구, 그리고 2009년도 조사 자료를 이용한 실증분석방법을 병행한다.

연구 내용을 요약하면 다음과 같다.

□ **2009년도 단위무응답과 항목무응답 수준파악**

○ 2009년도에 실시된 “국민건강영양조사” 자료로부터 무응답현황을 파악한다.

○ 무응답대체의 필요성 검토(대체변수의 선택).

□ **무응답 발생 변수들에 대한 연관성분석**

○ 무응답이 발생한 주요 변수들과 연관성이 있는 보조변수들을 찾는다.

○ 연관성분석을 통해 무응답을 대체할 적절한 대체군을 찾는다.

□ **무응답 대체 방안**

○ 무응답 대체방법들의 소개

○ 무응답 대체군 설정

○ 무응답 대체 실시

○ 대체 결과의 분석

□ **개선방안 및 기타**

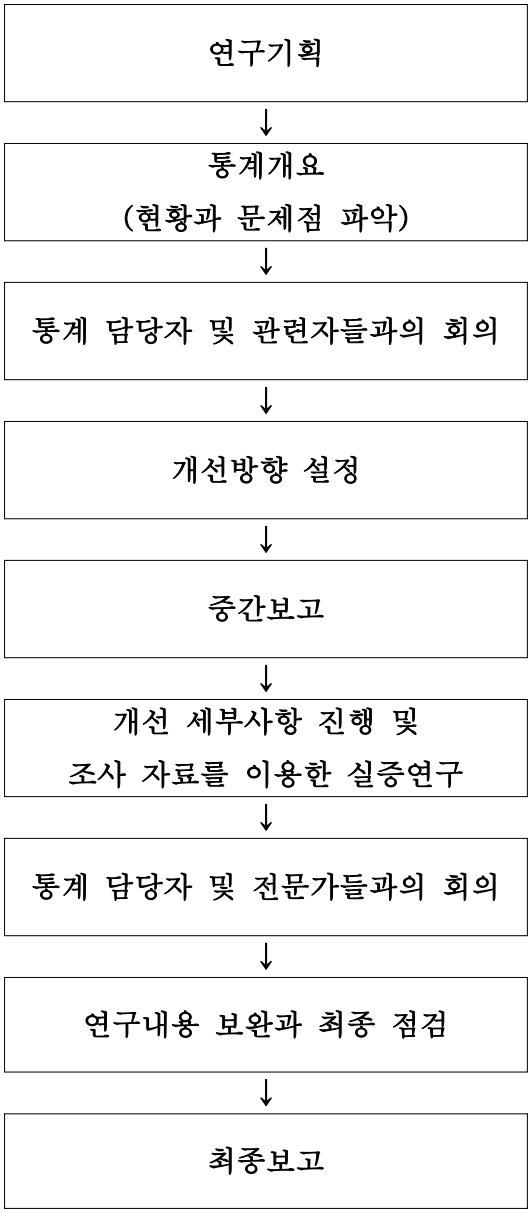
○ 본 조사에 적합한 항목무응답 대체과정 제시

○ 항목무응답 대체와 관련된 연구방향

○ 기타 제안사항

4절. 연구흐름도

본 연구용역의 수행을 위한 연구흐름도는 <그림 1>과 같다.



<그림 1> 연구흐름도

2장. 통계개요

"국민건강영양조사"는 지정통계(승인번호 11702)로 국민건강증진법 제 16조에 의해서 매년 조사결과가 공표된다. 1998년부터 3년 주기로 시행되던 본 조사는 제4기(2007-2009) 부터는 매년 조사하는 순환표본조사(rolling survey sampling)방법을 사용한다. 본 통계의 개요를 살펴보면 다음과 같다.

1절. 현황

□ 조사목적

○ 국민의 건강수준, 건강관련 의식 및 행태, 식품 및 영양섭취 실태에 대한 국가 및 시도 단위의 통계 산출

○ 만성질환 및 관련 위험요인의 시계열 추이 파악

○ 국민건강증진 종합계획의 정책목표 수립 및 평가, 건강증진 프로그램 개발 등 보건정책에 필요한 근거자료 제공

□ 조사대상

○ 제4기(2007-2009) 조사모집단은 2005년 인구주택총조사 결과의 모든 가구와 국민으로 정의

○ 제5기(2010-2012) 조사모집단은 2005년 인구주택총조사 결과의 노후화로 최신의 통합된 조사구모집단이 없어, 표본추출틀을 일반가구와 아파트 가구로 분리. 일반가구는 2009년 6월 주민등록조사 통/반/리별 목록을 활용하고 아파트가구는 KB 국민은행 아파트

시세 조사용 목록을 활용한 추출틀을 사용

□ **조사내용 및 방법**

○ “국민건강영양조사”는 건강설문조사, 영양조사, 검진조사로 구성

○ 건강설문조사와 검진조사는 이동검진센터에서 실시하고 영양조사는 대상 가구를 직접 방문조사

□ **조사 및 공표 주기 :** 1년

2절. 문제점

○ 2009년도 조사의 단위무응답률은 약 20%이고 항목무응답률은 약 5%정도이나 항목무응답에 대한 구체적인 현황 파악과 대체 방법이 미흡

○ 단위무응답은 가중치 조정방법을 사용하여 모수 추정에 활용하고 있으나 항목무응답에 대해서는 조치를 취하고 있지 않음

3절. 개선방안 개요

○ 2009년도 조사의 무응답현황 파악

○ 항목무응답 대체군 개발

○ 항목무응답 대체실시 및 대체결과 분석

3장. 세부 개선방안

1절. 연구협의 내용

본 연구 진행을 위해서 질병관리본부 연구진들과의 연구 협의 내용을 정리하면 다음과 같다.

□ 1차 협의 일시 및 장소

○ 일시 : 2011년 7월 14일(목), 오전 10시~12시

○ 장소 : 질병관리본부 건강영양조사과 회의실

○ 참석 :

- 질병관리본부 건강영양조사과 : 장명진(책임연구원), 조유미(선임연구원), 이나연(선임연구원), 양지은(기술연구원), 박선민(기술연구원)
- 청주대학교 : 류제복(연구책임자), 이승주(연구원)

□ 2차 협의 일시 및 장소

○ 일시 : 2011년 9월 1일(목), 오후 5시~6시 30분

○ 장소 : 청주대 류제복교수 연구실

○ 참석 :

- 질병관리본부 건강영양조사과 : 장명진(책임연구원), 조유미(선임연구원)
- 청주대학교 : 류제복(연구책임자), 이승주(연구원)

□ 협의내용

○ 조사현황

- 2009년도 표본수는 총 10,533명
- 3년 순환표본조사 실시

○ 조사방법

- 건강설문조사와 검진조사는 이동검진센터에서 실시
- 건강설문조사는 면접방식과 자기기입식(음주, 흡연 등) 사용하고 검진조사는 직접계측, 관찰, 검체분석 등을 사용
- 영양조사는 대상가구를 직접 방문조사

○ 무응답현황

- 단위무응답은 가중치조정 방법을 적용

※ 단위응답률은 보고서상에서 조사참여율로 간주. 이때 조사 대상 가구는 조사구내 가구목록에서 빈집을 제외하고 표본을 선정하며 응답을 거부한 경우는 예비표본가구로 대체한 후 조사

- 항목무응답은 특별한 조치를 취하고 있지 않음

※ 항목무응답의 주요 발생 원인은 응답 거부나 측정 자료가 자료처리의 기준을 벗어난 경우(예를 들어, 폐활량 측정에서 지나치게 작거나 높은 경우)

○ 무응답대체

- 항목무응답에 대한 대체를 검토

- 조사항목은 3개 조사에서 841개이나 변수 수는 3,000개 이상임
- 항목무응답률이 높거나 주요 변수에 대해서 무응답대체를 실시
- 통계청 등에서 무응답대체 방법으로 널리 사용되고 있는 핫덱(Hot-Deck) 대체를 중심으로 대체방법과 대체절차를 검토함
- 대체가 여의치 않으면 그 사유를 명시하고 대안을 제시
- 보조변수의 사용이 여의치 않으면 공통변수만 사용 검토

○ 자료

- 2009년도 Full data
- 항목무응답 대체를 위한 data(무응답 포함된 자료)
  (예; 기본변수 + 주요 무응답변수 + 무응답변수와 연관성이 있다고 판단되는 변수)

※ 질병관리본부의 관련분야 연구진들이 협의하여 주요 무응답항목(목표변수)과 해당 항목들과 연관성이 높다고 판단되는 변수(보조변수)들을 추가로 선정

2절. 무응답현황

"국민건강영양조사"는 국가 및 시·도 단위의 표본조사로 제1기(1998년)부터 제3기(2005년)까지 3년 주기로 실시되었다. 제4기(2007년-2009년)부터는 연중조사체계로 개편하여 매년 조사가 실시되고 있으며 현재 제5기(2010년-2012년)조사가 진행 중이다. 이 조사는 건강설문조사(630개 문항), 영양조사(132개 문항), 검진조사(79개 문항) 등 3개의 조사로 구성되며 총 설문 문항 수는 841개에 이르는 대규모 조사이다.

따라서 조사대상으로 선정된 표본들로부터 완전한 응답 자료를 얻는 것은 현실적으로 불가능하다. 즉, 표본으로 선정된 응답자들이 부재중이거나 응답 거부, 응답 불능, 또는 부적격 등의 요인으로 무응답이 발생하게 된다. 이러한 무응답은 조사결과에 상당한 영향을 미쳐 조사의 신뢰도를 떨어뜨린다.

무응답은 설문 문항 모두에 대해서 응답하지 않는 단위무응답(unit nonresponse)과 일부 항목에 대해서만 응답하지 않는 항목무응답(item nonresponse)으로 나눈다. 2009년 조사된 자료로부터 단위무응답과 항목무응답 현황을 살펴본다(참고: 복지부・질병관리본부, 2009).

(1) 단위무응답 현황

"국민건강영양조사"에서의 단위무응답 현황은 조사참여율로 정의되어 있다. 제4기(2007-2009)의 조사참여율은 <표 1>과 같다. 단위무응답(조사불참)에 대해서는 가중치조정을 통해서 추정에 반영해 준다.

<표 1> 제4기(2007-2009) 국민건강영양조사 참여율

	전체			검진 및 건강설문조사			영양조사		
	대상자	참여자	참여율	대상자	참여자	참여율	대상자	참여자	참여율
제4기 (2007-2009)	31,705	24,871	78.4	31,705	23,632	74.5	27,050	22,137	81.8
1차년도 (2007)	6,455	4,594	71.2	6,455	4,246	65.8	5,083	4,099	80.6
2차년도 (2008)	12,528	9,744	77.8	12,528	9,308	74.3	10,539	8,641	82.0
3차년도 (2009)	12,722	10,533	82.8	12,722	10,078	79.2	11,428	9,397	82.2

1) 전체조사 참여율 : 검진조사, 건강설문조사, 영양조사 중 1개 이상 참여한 비율
2) 영양조사 대상자 : 검진조사 또는 건강설문조사에 1인 이상 참여한 가구의 가구원 전체

(2) 항목무응답 현황

"국민건강영양조사"에서의 항목무응답은 <표 2>와 같다. 항목무응답은 무응답률이 낮아 대체를 고려하지 않았다. 그러나 향후 조사환경의 변화로 항목무응답의 발생이 증가할 것으로 예상된다. 이번 연구에서는 제4기 조사결과를 바탕으로 질병관리본부와의 협의를 통해 8개의 항목무응답 변수를 선택하였다. 이번에 선택한 변수의 항목무응답은 <표 2>에서와 같이 항목무응답률이 높은 것만을 기준으로 하지 않고 현재는 항목무응답률이 높지 않지만 조사결과에 영향이 크고 민감한 항목들을 선정하였다. 따라서 항목무응답에 대한 대체과정은 비록 선정변수가 변한다 하여도 큰 차이는 없을 것이다.

<표 2> 제4기(2007-2009) 국민건강영양조사 항목무응답 현황

조사부문		조사 항목	무응답률(%)		
			'07	'08	'09
건강설문조사	흡연	흡연시작연령	3.54	0.67	0.19
		직장실내 간접흡연노출여부	5.10	1.39	0.41
	음주	음주시작연령	5.13	0.90	0.03
		1년간 음주운전여부 - 오토바이	0.50	0.12	0.40
	신체활동	격렬한신체활동 시간	3.61	0.89	0.00
		걷기 시간	2.74	0.40	0.02
	정신건강	하루수면시간	3.79	0.10	0.41
		자살생각여부	0.34	0.12	0.44
	삶의 질	1달간 결근결석여부	0.80	0.25	0.53
	이환	1년간 숨쉴때 가슴에서 소리여부	1.34	0.47	0.57
	손상	지난1년간 손상발생여부	0.13	0.25	0.53
	의료이용	병의원미검진여부	0.13	0.22	0.52
	가구조사	가구총소득(개인단위)	4.66	3.11	1.12
	교육 및 경제활동	교육수준	0.17	0.26	0.04
		직업분류	1.35	0.12	0.07

<표 2> 제4기(2007-2009) 국민건강영양조사 항목무응답 현황 (계속)

조사부문		조사 항목	무응답률(%)		
			'07	'08	'09
검진조사	신체계측	신장, 체중	0.78	0.77	0.34
		허리둘레	0.80	1.02	0.33
	혈압	혈압	0.53	0.09	0.20
	구강검사	치아우식, 자연치아, 보철물	4.12	2.20	1.58
		치주질환	8.56	2.50	0.05
	임상검사	헤모글로빈, 헤마토크릿	3.54	5.05	6.41
		혈당	3.20	4.68	6.04
		지질, 크레아티닌, 인슐린	3.14	4.27	5.79
	안(눈)검사		-	0.74	0.41
	이비인후(귀, 코, 목)검사		-	1.73	0.13
	골밀도 및 체지방검사		-	3.14	3.36
영양조사	식품섭취 빈도조사	보리/잡곡	0.12	0.04	0.05
		토마토	0.12	0.07	0..04
	식생활조사	비타민 및 무기질제 섭취 경험	2.27	1.84	1.56
	영유아 식생활조사	출생시 신장	8.19	14.13	11.75
		일반우유 시작시기	1.64	0.80	0.50

1) 건강설문 무응답률 : 성인(만19세이상) 기준

3절. 무응답 대체방법

무응답이 발생한 경우 무응답을 대체하는 방법은 다양하다. 비록 같은 조사라 하더라도 대체할 변수의 특성과 보조변수에 따라 대체방법이 다를 수 있다. 본 절에서는 일반적으로 널리 사용되고 있는 대체방법을 단위무응답과 항목무응답으로 나누어서 살펴본다. 자세한 대체방법들은 김규성(2000), 김영원과 조선경(1996), 송주원과 안형진(2009), 조사통계연구회(2000), 그리고 최필근(2009a, 2009b, 2009c, 2010, 2011) 등을 참고하면 된다.

(1) 단위무응답 대체방법

□ 무응답자 교체(nonrespondent substitution)

초기 표본에 있는 무응답자(가구 또는 사업체)를 초기 표본에 포함되지 않은 모집단의 다른 구성원들로 교체하여 처음의 표본크기와 같게 해 주는 방법을 무응답자 교체라 한다. 일반적으로 교체응답자들은 확률적(random), 또는 비확률적(nonrandom)으로 얻는다.

확률적 교체는 무응답자와 같은 집단에 있는 구성원들로부터 선정되지만 비확률적 교체는 확률추출에 의하지 않고 사전에 정해진 기준을 적용해서 이루어진다. 예를 들어, 표본으로 선정된 단위가 응답을 하지 않으면 면접원은 즉시 그 지역에 거주하고 있는 무응답 가구와 유사한 특성을 갖고 있는 다른 단위로 교체한다. 교체 표본들이 교체해야 할 것들과 모든 면에서 동일하면 그 만큼 교체는 무응답편향을 줄여준다.

무응답자 교체의 장점은 목표로 하는 표본크기를 유지할 수 있어 표본오차로부터의 분산성분을 관리할 수 있다. 또한 무응답자들과 교체된 사람들이 주 연구변수에 관해 유사하다면, 무응답편향이 감소될 수 있다. 한편 단점으로는 교체가 가능하다는 점이 바로 응답을 얻으려는 면접원들의 노력을 줄여 기대응답률보다 초기표본에 대한 응답률이 낮게 된다. 그리고 조사의 응답률이 과대 추정될 가능성이 있다.

□ 가중치조정(weight adjustment)

가중치조정은 단위무응답을 다루는 가장 일반적인 방법으로 무응답에 대해 별도로 다른 값을 대체해주지 않는 대신에 응답된 자료들의 가중치를 조정해주어 무응답으로 인한 효과를 줄여주는 방법이다. 가

중치를 계산하는 방법으로는 이용 가능한 정보의 근원에 따라 “표본에 기초한 방법”과 “외부정보를 이용한 방법”이 있다.

○ 표본에 기초한 방법

이용 가능한 정보가 표본으로 한정되며, 전체 모집단에 대한 정보는 알 수 없다. 무응답 단위들의 기본가중치를 표본응답자들에게 배정하여 응답단위들에 대해 조정된 가중치의 합은 전체 표본단위들에 대한 기본가중치의 합이 된다.

조사 단위들은 응답그룹(그룹1), 무응답그룹(그룹2), 그리고 부적격 그룹(그룹3)중의 하나에 속하고 무응답자들은 모두 조사 적격자들이라고 가정한다. 표본자료에 의한 무응답 조정인자는 다음과 같다.

$$F_c = \frac{\sum_{i=1}^{n_1} w_i MOS_i + \sum_{i=1}^{n_2} w_i MOS_i}{\sum_{i=1}^{n_1} w_i MOS_i}$$

여기서 n_1은 응답자들의 수이고, n_2는 무응답자들의 수이다. w_i는 기본가중치이고 MOS_i는 동일하거나, 표본추출설계에 의해 결정되는 적절한 크기척도, 또는 특정한 형태의 추정치에 효율적이게 선정될 수 있다. 무응답 조정인자 F_c를 표본응답자들의 기본가중치에 곱하여 조정된 가중치를 얻는다. 즉, 표본에 기초한 조정된 가중치, $w_i^{(a)}$는 다음과 같이 계산된다.

$$w_i^{(a)} = \begin{cases} F_c w_i, & \text{그룹1} \\ 0, & \text{그룹2} \\ w_i, & \text{그룹3} \end{cases}$$

그리고 조정된 가중치들은 다음의 성질을 갖는다.

$$\sum_{i=1}^{n_1} w_i^{(a)} MOS_i + \sum_{i=1}^{n_3} w_i^{(a)} MOS_i = \sum_{i=1}^{n_1+n_2+n_3} w_i MOS_i$$

○ 외부정보를 이용한 방법

표본에 기초한 무응답 조정 가중치를 계산한 후 외부자료를 이용하여 사후층화, 갈퀴(raking)법, 또는 보정(calibration)등의 방법으로 이 가중치를 조정한다. 외부자료를 이용한 경우 표본자료에 의한 무응답 조정인자 F_c는 다음과 같이 수정된다.

$$F_c^* = \frac{MOS_c}{\sum_{i=1}^{n_1} w_i MOS_i}$$

여기서 MOS_c는 계급 c에 대해 외부데이터로부터 얻는 크기척도이며, 이 계급에서 응답자들에 대해 조정된 가중치를 합하면 MOS_c가 된다.

(2) 항목무응답 대체방법

항목무응답을 대체하는 방법은 무응답항목에 대해서 하나의 값을 대체하는 단일대체방법(single imputation)과 여러 개의 값을 대체하는 다중대체방법(multiple imputation)이 있다. 여기서는 단일대체방법에 대해서 다룬다. 한편 대체항목의 선정은 확률을 기반으로 한 확률적 대체(stochastic imputation)와 비확률적인 방법인 결정론적 대체(deterministic imputation)로 나룰 수 있다. 결정론적 대체방법으로 연역적대체, 평균대체, 비대체, 회귀대체 등이 주로 사용되고 확률적 대체방법으로는 랜덤핫덱대체, 응용핫덱대체, 가중핫덱대체, 랜덤회귀대

체 등이 널리 사용된다. 특히 범주형 자료의 경우 범주별로 응답자들의 분포에 따라 무응답자를 확률적으로 대체할 때 대체군을 사용하지 않는 경우와 대체군을 이용한 경우를 각각 단순임의 확률대체와 대체군을 이용한 확률대체로 나눈다.

□ 연역적 대체(deductive imputation)

연역적 대체방법은 비교적 간단하고 쉽게 항목무응답을 대체할 수 있는 방법이지만 논리적 제약조건이 명확하고 무응답과 관련된 보조자료가 충분해야 가능하다. 추론에 사용된 자료가 확실하지 않을 경우에는 무응답에 대한 대체값에 오류가 생기게 된다. 따라서 연역적 대체를 위해서는 결측값과 관련된 항목들에 응답한 자료들 간의 일치성을 점검해야 한다.

□ 평균대체(mean imputation)

평균대체는 무응답에 대한 대체값으로 대체군내의 응답값들의 평균을 사용하는 방법으로 간편하고 이용하기 쉽다. 특히 항목변수가 양적변수이고 구하고자 하는 통계량이 평균일 때 적절하다.

크기 N인 모집단에서 크기 n인 단순확률표본을 추출하였을 때 응답표본과 무응답표본을 각각 R과 R^c로 표기하자. 이때 응답단위는 r개이다. 그러면 평균대체는 $(n-r)$개의 무응답 단위의 대체값으로 r개의 응답단위들의 평균값을 사용한다. 즉, 대체후의 자료는 다음과 같다.

$$y_k^* = \begin{cases} y_k, & k \in R, \quad (\text{응답값}) \\ \bar{y}_r, & k \in R^c, \quad (\text{대체값}) \end{cases}$$

따라서 모집단 평균은 다음과 같이 관찰된 값들과 대체값들의 평균으

로 추정된다.

$$\bar{y}_I = \frac{1}{n}\left[\sum_{k\in R} y_k + \sum_{k\in R^c} \bar{y}_r\right] = \frac{1}{n}\left[r\bar{y}_r + (n-r)\bar{y}_r\right]$$
$$= \bar{y}_r$$

평균대체에 의한 모집단평균의 추정치인 $\bar{y}_I$의 분산은 표본평균보다 큰 분산을 갖는다. 그리고 무응답들에 모두 동일한 평균값을 대체하게 되므로 평균값의 빈도가 지나치게 높아서 관심변수의 경험적 분포가 왜곡되는 단점도 있다.

□ 비대체(ratio imputation)

무응답에 대한 대체값으로 관심변수와 관련이 있는 보조변수를 이용하여 다음과 같이 대체하는 방법을 비대체라 한다. 비대체 방법을 사용하기 위해서는 관심변수와 보조변수가 원점을 지나는 직선 관계를 갖고 분산이 보조변수에 비례하는 경우 효과적이다.

$$y_k^* = \begin{cases} y_k, & k\in R, \quad (\text{응답값}) \\ \left(\dfrac{\bar{y}_r}{\bar{x}_r}\right)x_k, & k\in R^c, \quad (\text{대체값}) \end{cases}$$

한편 랜덤비대체는 기존의 대체값에 확률오차인 e_k를 더해서 사용하는 방법이다. 즉,

$$y_k^* = \begin{cases} y_k, & k\in R, \quad (\text{응답값}) \\ \left(\dfrac{\bar{y}_r}{\bar{x}_r}\right)x_k + e_{k,} & k\in R^c, \quad (\text{대체값}) \end{cases}$$

이때 확률오차는 대체 전·후의 변동이 일정하도록 변동 폭을 계산해 주면 대체로 인한 변동의 감소를 조정해주는 효과를 얻을 수 있다.

□ **회귀대체(regression imputation)**

비대체방법과 유사하게 무응답에 대한 대체값으로 관심변수와 관련이 있는 보조변수를 이용한다. 이때 관심변수와 보조변수가 절편이 있는 직선 관계를 갖고 관심변수의 분산이 동일 할 때 유용하다.

$$y_k^* = \begin{cases} y_k, & k \in R, \quad (\text{응답값}) \\ \bar{y}_r + b(x_k - \bar{x}_r), & k \in R^c, \quad (\text{대체값}) \end{cases}$$

여기서 $b = \sum_{k \in R}(x_k - \bar{x}_r)(y_k - \bar{y}_r) / \sum_{k \in R}(x_k - \bar{x}_r)^2$이다. 한편 랜덤회귀대체는 랜덤비대체와 유사하게 기존의 대체값에 확률오차인 e_k를 더해준다. 즉,

$$y_k^* = \begin{cases} y_k, & k \in R, \quad (\text{응답값}) \\ \bar{y}_r + b(x_k - \bar{x}_r) + e_k, & k \in R^c, \quad (\text{대체값}) \end{cases}$$

회귀대체는 보조변수가 양적변수이든 질적변수이든 모두 사용가능하고, 질적변수인 경우에는 가변수(dummy variable)를 사용하면 된다. 한편 관심변수가 질적변수인 경우에는 대수선형모형이나 로지스틱모형을 적용할 수 있다.

□ **핫덱대체(hot-deck imputation)**

1962년 미국의 CPS(current population survey)에서 연간소득에 대한 결측값을 대체하는 데 처음으로 사용된 핫덱대체는 최근에 실제조사에서 가장 널리 사용되고 있다. 결측값에 대한 대체값으로 과거에 실시된 조사에서의 유사자료를 사용하는 콜덱(cold-deck)방법의 단점을 보완하기 위해서 같은 조사의 응답자들로부터 얻은 자료를 사용한다.

핫덱대체 시 별도의 대체군을 사용하지 않고 자료내의 무응답에 대해서 응답값을 임의로 선택하여 대체하는 단순임의 핫덱방법이 있다.

그러나 많은 경우 대체군을 이용한 핫덱대체 방법이 사용된다. 널리 사용되는 방법으로 랜덤핫덱(random hot-deck), 가중핫덱(weighted hot-deck), 응용핫덱(applied hot-deck), 그리고 계층적핫덱(hierarchical hot-deck) 등으로 구분할 수 있다.

랜덤핫덱은 무응답 대체를 위해서 대체군내의 응답값들 중에서 확률적으로 하나의 값을 선정해서 대체하는 방법이다. 즉,

$$y_k^* = \begin{cases} y_k, & k \in R, \quad (\text{응답값}) \\ y_i, & k \in R^c, \quad (\text{대체값}) \end{cases}$$

여기서 y_i는 응답값들 중에서 확률적으로 선정된 값이다. 랜덤핫덱 방법은 표본대체 후에도 표본의 분포가 그대로 유지되는 장점이 있고 통계량의 형태에 관계없이 사용될 수 있다. 랜덤핫덱에서 응답값들 중에 하나를 선정할 때 가중치를 달리해서 선정하면 가중핫덱이 된다. 응용핫덱은 랜덤핫덱의 응용으로 연속형 항목에서도 적용할 수 있도록 한 방법이다. 대체군 사용 시 범주형 항목은 항목값의 일치 여부를 판단해서 점수를 부여하고 연속형 범주는 신뢰구간을 이용해서 그 구간에 포함되면 비슷한 개체로 간주하여 점수를 부여한다. 모든 대체군의 항목들에 대해서 가장 점수가 높은 개체를 선택하고 그 중에서 하나를 확률적으로 선택해서 대체하는 방법이다. 한편 대체군 형성 시에 많은 보조변수를 사용하면 대체군 수가 증가해서 대체군내에서 기증자를 찾지 못하는 경우에 발생한다. 이런 경우 계층적 핫덱대체가 사용된다. 즉, 대체군 형성에 사용된 보조변수들 중에서 연관성이 적은 보조변수를 제거하고 새로이 대체군을 형성해서 앞서 기증자를 못 찾은 무응답에 대한 대체를 실시한다. 이러한 과정을 무응답대체가 이루어질 때까지 계속하는 방법을 계층적 핫덱대체라 한다.

4절. 무응답대체

□ 개요

제4기 3차년도(2009) 국민건강영양조사는 건강설문조사, 검진조사, 영양조사 등 세 가지 조사부문으로 구성되었다. 제4기 3차년도(2009) 자료의 참여자수는 조사별로 상이하고 가중치에 대한 평균, 표준편차, 최소값과 최대값은 <표 3>에 있다.

<표 3> 분야별 가중치 분포 및 참여자수

구분	모집단(명)	가중치변수	대상(명)	평균	표준편차	최소값	최대값
가구	16,916,966(가구)	wt_hs	3,975(가구)	4,271	1,483	446	9,188
건강설문조사	48,303,676	wt_itv	10,051	4,806	2,734	237	19,191
영양조사	48,303,676	wt_ntr	9,397	5,140	3,222	248	28,228
식품섭취조사 추가조사	48,303,676	wt_ntr_add	2,029	23,807	18,433	1,948	95,580
검진조사	48,303,676	wt_ex	10,078	4,793	2,725	241	19,191
검진조사 하반기신규도입	37,844,232	wt_ex1	2,866	13,205	9,173	1,578	43,693
체지방검사	43,838,503	wt_dw	7,739	5,665	3,719	247	28,787
폐기능검사	37,844,232	wt_pft	4,419	8,564	7,014	475	38,086
중금속검사	37,215,961	wt_hm	1,991	18,692	10,522	931	53,801

□ 목표변수와 보조변수

“국민건강영양조사”의 항목무응답을 대체할 목표변수와 대체에 사용할 보조변수들은 조사의 전문성을 고려해서 질병관리본부 건강영양조사과와 협의하여 선정하였다. 보조변수는 목표변수들에 공통적으로 사용될 공통보조변수와 각 목표변수에 적합한 보조변수들로 구성된다. 이들 보조변수들 중에서 무응답대체에 사용될 최종 보조변수들은 연관성분석을 통해서 결정한다. 목표변수들과 1차로 선정된 보조변수들의 특성은 <표 4>와 같다.

<표 4> 목표변수 및 보조변수들의 특성

목표변수			보조변수		
변수설명	해당연령	변수명	변수설명	해당연령	변수명
			·성별 ·연령 ·시도 ·동읍면 ·주택유형 ·소득수준 (공통보조변수)	1세이상	sex age region town_t apt_t incm
교육수준	1세이상	edu	·(성인)유년기환경 : 아버지 경제활동상태 ·(성인)유년기환경 : 아버지 직업분류코드 ·(성인)유년기환경 : 어머니 경제활동상태 ·(성인)유년기환경 : 어머니 직업분류코드 ·(성인)유년기환경 : 아버지 교육수준 ·(성인)유년기환경 : 어머니 교육수준	19세이상	EC_pjob_1 EC_pjob_3 EC_pjob_4 EC_pjob_6 EC_pedu_1 EC_pedu_2
월평균가구 총소득	1세이상	ainc	·결혼상태 ·교육수준 ·건강보험종류 ·민간의료보험가입여부	1세이상	marri_2 edu tins npins
			·(15세이상)경제활동 상태 ·(취업자)종사상지위 ·(취업자)종사상지위_임금근로자상세 ·(취업자)근로시간제 ·(취업자)사업체 규모 ·(취업자)표준산업분류 대분류 코드	15세이상	EC1_1 EC_stt_1 EC_stt_2 EC_wh EC_sz EC_ind
주관적 건강상태	1세이상	D_1_1	·최근2주간 몸이 불편했던 경험 유무 ·체질량지수 ·활동제한 여부	1세이상	D_2_1 HE_BMI LQ4_00
			·(성인)EuroQoL : 운동능력 ·(성인)EuroQoL : 자기관리 ·(성인)EuroQoL : 일상활동 ·(성인)EuroQoL : 통증/불편 ·(성인)EuroQoL : 불안/우울 ·현재흡연여부 ·월간음주여부 ·중등도신체활동 여부	19세이상	LQ_1EQL LQ_2EQL LQ_3EQL LQ_4EQL LQ_5EQL sm_presnt dr_month pa_mid
현재흡연 여부	19세이상	sm_presnt	·평생흡연여부 ·흡연시작연령 ·하루평균 흡연량 ·직장실내 간접흡연노출여부 ·가정실내 간접흡연노출여부	19세이상	BS1_1 BS2_1 BS3_2 BS8_2 BS9_2
체질량지수	1세이상	HE_BMI	·총(whole body total) 체지방률	10세이상	DW_WBT_pFT
			·주관적 체형인지	1세이상	BO1
			·식품섭취빈도 - 쌀 - 배추김치	12세이상	 F_RICE F_KCABB

<표 4> 목표변수 및 보조변수들의 특성 (계속)

목표변수			보조변수		
변수설명	해당연령	변수명	변수설명	해당연령	변수명
			·성별 ·연령 ·시도 ·동읍면 ·주택유형 ·소득수준 (공통보조변수)	1세이상	sex age region town_t apt_t incm
총콜레스테롤	1세이상	He_Chol	·(성인)고지혈증 유병여부 ·(성인)고지혈증 의사진단여부 ·(성인)고지혈증 치료 ·(성인)고지혈증제 복용 ·비만유병여부 ·뇌졸증 유병여부 ·심근경색증 또는 협심증 유병여부 ·고혈압 유병여부	19세이상	DI2_lt DI2_dg DI2_pt DI2_2 HE_obe DI3_lt DI4_lt HE_HP
폐기능 판정결과	19세이상	HE_COPD	·현재흡연여부 ·(성인)천식 유병여부 ·(성인)만성폐쇄성 폐질환 유병여부 ·(성인)만성폐쇄성 폐질환 의사진단여부 ·(성인)폐결핵 유병여부 ·(성인)폐결핵 의사진단여부 ·(성인)기관지확장증 유병여부 ·(성인)기관지확장증 의사진단여부 ·(성인)만성폐쇄성폐질환 증상1_최근 1년간 3개월 이상 가래여부 ·(성인)만성폐쇄성폐질환 증상2_최근 1년간 3개월 이상 기침여부	19세이상	sm_presnt DJ4_lt DJ5_lt DJ5_dg DJ2_lt DJ2_dg DJ7_lt DJ7_dg DJ5_11 DJ5_21
			·천식 의사진단여부	1세이상	DJ4_dg
에너지	10세이상	N_EN	·체질량지수 ·조사1일전 아침 결식(아니오 응답) ·조사1일전 점심 결식(아니오 응답) ·조사1일전 저녁 결식(아니오 응답)	1세이상	HE_BMI L_BR1 L_LN1 L_DN1
			·식품섭취빈도 - 쌀 - 배추김치 - 라면	12세이상	F_RICE F_KCABB F_INSTND
			·주관적 체형인지	1세이상	BO1
			·중등도신체활동 여부	19세이상	pa_mid

□ 대체기준

"국민건강영양조사"의 마이크로자료에는 변수가 3천여 개에 이를 정도로 많고, 계층적구조로 이루어진 설문도 있다. 상위 질문에 해당사항

이 없는 경우 하위질문에서도 계속해서 비해당으로 처리되므로 무응답과는 다르다. 그리고 3개 조사의 대상수도 다르고 조사대상 연령도 상이하다. 항목무응답대체에 사용될 변수들의 특성도 상이하므로 무응답대체를 위한 몇 가지 기준을 정하였다.

○ 무응답 대체방법의 선정은 조사 자료에서 무응답을 제거한 응답자료를 대상으로 무응답 대체방법을 비교분석한 후 선정하는 것이 일반적이다. 그러나 본 조사에서 무응답을 제거한 응답 자료가 충분치 않아서 본 조사와 유사한 경우에 일반적으로 널리 사용되는 핫덱대체 방법을 사용하였다.

○ 무응답대체에 사용할 보조변수에 무응답이 있는 경우에는 우선적으로 보조변수의 무응답을 대체하고 목표변수의 대체를 실시해야 한다. 그러나 무응답이 발생한 보조변수의 대체를 위해 또 다른 보조변수를 선정하는 등 대체과정이 매우 복잡하고 번거로워 대체의 정확성이 떨어지게 되므로 여기서는 보조변수에 대한 무응답대체는 실시하지 않았다. 즉, 무응답이 있는 보조변수의 정보는 제외하였다.

○ 무응답을 대체할 목표변수의 조사대상 연령과 보조변수의 조사대상 연령이 다른 경우 이용 가능한 연령대로 하였다. 예를 들어 목표변수의 조사대상이 1세 이상이나 보조변수의 조사대상이 19세 이상인 경우 무응답대체 연령도 19세로 하였다.

상기 기준에 따라 목표변수와 보조변수들의 분석가능 연령은 <표 5>와 같다.

<표 5> 목표변수와 보조변수들의 분석 가능 연령

목표변수			보조변수					
변수설명	변수명	조사부문	변수설명	조사부문				분석가능 연령
				공통	건강	검진	영양	
			·성별	○				1세 이상
			·연령	○				
			·시도	○				
			·동읍면	○				
			·주택유형	○				
			·소득수준	○				
교육수준	edu	건강설문조사	·(성인)유년기환경 : 아버지 경제활동상태		○			19세 이상
			·(성인)유년기환경 : 아버지 직업분류코드		○			
			·(성인)유년기환경 : 어머니 경제활동상태		○			
			·(성인)유년기환경 : 어머니 직업분류코드		○			
			·(성인)유년기환경 : 아버지 교육수준		○			
			·(성인)유년기환경 : 어머니 교육수준		○			
월평균가구 총소득	ainc	건강설문조사	·결혼상태		○			19세 이상
			·교육수준	○				
			·건강보험종류		○			
			·민간의료보험가입여부		○			
			·(15세이상)경제활동 상태		○			
			·(취업자)종사상지위		○			
			·(취업자)종사상지위_임금근로자상세		○			
			·(취업자)근로시간제		○			
			·(취업자)사업체 규모		○			
			·(취업자)표준산업분류 대분류 코드		○			
주관적 건강상태	D_1_1	건강설문조사	·최근2주간 몸이 불편했던 경험 유무		○			19세 이상
			·체질량지수			○		
			·활동제한 여부		○			
			·(성인)EuroQoL : 운동능력		○			
			·(성인)EuroQoL : 자기관리		○			
			·(성인)EuroQoL : 일상활동		○			
			·(성인)EuroQoL : 통증/불편		○			
			·(성인)EuroQoL : 불안/우울		○			
			·현재흡연여부		○			
			·월간음주여부		○			
			·중등도신체활동 여부		○			
현재흡연 여부	sm_presnt	건강설문조사	·평생흡연여부		○			19세 이상
			·흡연시작연령		○			
			·하루평균 흡연량		○			
			·직장실내 간접흡연노출여부		○			
			·가정실내 간접흡연노출여부		○			
체질량지수	HE_BMI	검진조사	·총(whole body total) 체지방률			○		12세 이상
			·주관적 체형인지		○			
			·식품섭취빈도					
			- 쌀				○	
			- 배추김치				○	

<표 5> 목표변수와 보조변수들의 분석 가능 연령 (계속)

목표변수			보조변수					
변수설명	변수명	조사부문	변수설명	조사부문				분석가능 연령
				공통	건강	검진	영양	
총콜레스테롤	He_Chol	검진조사	·(성인)고지혈증 유병여부 ·(성인)고지혈증 의사진단여부 ·(성인)고지혈증 치료 ·(성인)고지혈증제 복용 ·비만유병여부 ·뇌졸증 유병여부 ·심근경색증 또는 협심증 유병여부 ·고혈압 유병여부		○ ○ ○ ○ ○ ○ 	 ○ ○	 	19세 이상
폐기능 판정결과	HE_COPD	검진조사	·현재흡연여부 ·(성인)천식 유병여부 ·(성인)만성폐쇄성 폐질환 유병여부 ·(성인)만성폐쇄성 폐질환 의사진단여부 ·(성인)폐결핵 유병여부 ·(성인)폐결핵 의사진단여부 ·(성인)기관지확장증 유병여부 ·(성인)기관지확장증 의사진단여부 ·(성인)만성폐쇄성폐질환 증상1_최근 1년간 3개월 이상 가래여부 ·(성인)만성폐쇄성폐질환 증상2_최근 1년간 3개월 이상 기침여부 ·천식 의사진단여부		○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○			19세 이상
에너지	N_EN	영양조사	·체질량지수 ·조사1일전 아침 결식(아니오 응답) ·조사1일전 점심 결식(아니오 응답) ·조사1일전 저녁 결식(아니오 응답) ·식품섭취빈도 - 쌀 - 배추김치 - 라면 ·주관적 체형인지 ·중등도신체활동 여부		 ○ ○	○ 	 ○ ○ ○ ○ ○ ○ 	12세 이상

(1) 무응답대체군 설정

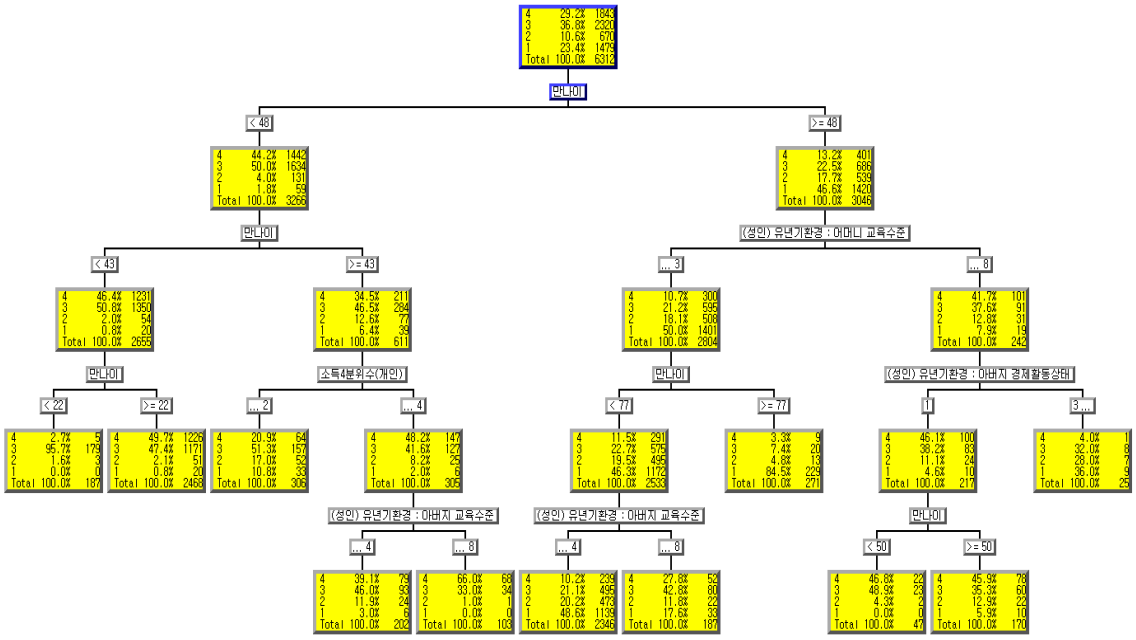
1) 연관성분석

제4기 3차년도(2009) 국민건강영양조사 자료에서 연관성분석을 실시할 항목은 교육수준, 월평균가구총소득, 주관적 건강상태, 현재흡연여

부, 체질량지수, 총콜레스테롤, 폐기능 판정결과, 에너지 등 8개 항목이다. 이 항목들을 이용하여 각 항목별로 대체에 사용할 보조변수(대체군)를 CHAID 알고리즘을 이용하여 선정한다.

□ 교육수준

교육수준 무응답은 건강설문조사에서 44명이나 19세 미만의 무응답자 2명(13세)은 연역적으로 대체가 가능하므로 19세 이상에서 무응답은 총 42명이다. 목표변수인 교육수준과 보조변수로 사용할 공통변수들은 해당연령이 1세 이상이나 6개의 보조변수들은 19세 이상 성인을 대상으로 조사한 변수들이므로 교육수준에 대한 연관성 분석의 기준 연령은 19세 이상 성인을 대상으로 하였다.



<그림 2> 교육수준에 대한 연관성모형(의사결정나무)

교육수준에 대한 연관성 분석결과는 <그림 2>와 <표 6>을 이용하여

자세히 설명하고 나머지 항목들은 간단히 설명한다. <그림 2>의 교육수준에 대한 연관성모형은 CHAID 알고리즘을 사용하여 형성된 의사결정나무이다. 교육수준에 대한 연관성분석 결과 연관성을 나타내는 깊이, 각 마디의 분리변수와 분리값, 그리고 중요도가 <표 6>에 있다. 분리변수의 괄호 안의 숫자는 각 가지에서의 마디 번호를 나타내며, 마지막 열의 중요도는 목표변수를 분리하는데 상대적으로 기여하는 정도를 나타낸다. 즉, 중요도가 큰 보조변수는 대체군을 형성하는데 기여하는 바가 크다.(참고; 최필근(2009a), Brieman et al.(1984))

<표 6> 교육수준에 대한 연관성분석

교육수준				
깊이 (연관성)	분리변수	분리값(좌)	분리값(우)	중요도
1	만나이	48미만	48이상	1.0000
2	만나이(1) 유년기 환경: 어머니 교육수준(2)	43미만 1, 2, 3	43이상 4, 5, 6, 7, 8	0.2982
3	만나이(1) 소득사분위수(개인)(2) 만나이(3) 유년기 환경: 아버지 경제활동상태(4)	22미만 1, 2, M 77미만 1	22이상 3, 4 77이상 2, 3	0.1413 0.0953
4	유년기 환경: 아버지 교육수준(4) 유년기 환경: 아버지 교육수준(5) 만나이(7)	1, 2, 3, 4, M 1, 2, 3, 4, M 50미만	5, 6, 7, 8 5, 6, 7, 8 50이상	0.2243

깊이 1

첫 번째 분리가 시작된다. 교육수준은 4개의 항목으로 구성되어 있으며 의사결정나무에서 첫 번째 마디인 부모마디에 나타난 성인 전체의 교육수준 분포(구성비)는 고등학교 졸업이 36.8%, 대졸이상이

29.2%, 초등학교 졸업이하가 23.4%, 중학교 졸업이 10.6%로 나타나 있다. 이를 분리기준에 따라 분리하여 자식마디를 형성하게 된다.

분리변수

분리변수는 만나이로 목표변수인 교육수준의 분포를 가장 잘 구별해주는 보조변수이다. 즉, 목표변수와 가장 연관성이 높은 항목은 만나이로 중요도는 1이다.

분리값(좌)

만나이로 교육수준을 왼쪽으로 분리하는데, 만나이가 48세 미만인 성인들은 왼쪽으로 분리한다. 이때 성인들의 교육수준 분포는 고졸이 50.0%, 대졸이상이 44.2%, 중졸이 4.0%, 초졸이하가 1.8%로 부모마디에 비해 순수도(impurity, 목표변수의 특정 범주에 해당 마디의 개체들이 집중되어 있는 정도)가 크게 증가하였음을 알 수 있다. 이는 만나이가 교육수준을 잘 분리하고 있다는 의미다.

분리값(우)

만나이가 48세 이상인 성인들은 오른쪽으로 분리되며, 성인들의 교육수준 분포는 초졸이하가 46.6%, 고졸이 22.5%, 중졸이 17.7%, 대졸이상이 13.2%로 나타났다. 부모마디에 비해 순수도가 증가하여 만나이가 교육수준을 잘 분리하고 있음을 알 수 있다.

깊이 2

두 번째 분리가 시작된다. 두 번째 분리과정은 첫 번째 분리한 왼쪽과 오른쪽 마디를 분리기준에 따라 더 상세하게 분리한다.

2(1) 만나이 만나이로 교육수준을 첫 번째 왼쪽으로 분리한 3,266명을 다시 만나이로 분리한다는 것을 의미한다. 즉, 48세 미만인 성인 3,266명을 43세 미만은 왼쪽으로 분리하고 43세 이상은 오른쪽으로 분리하여 자식마디를 형성한다. 이 때 만나이가 43세 미만 성인의 교육수준 분포는 고졸이 50.8%, 대졸이상이 46.4%, 중졸이 2.0%, 초졸이하가 0.8%로 나타났다. 이는 첫 번째 분리와 큰 차이를 보이지 않으므로 추가 분리가 필요함을 알 수 있다. 만나이가 43세 이상 48세 미만 성인의 교육수준 분포는 고졸이 46.5%, 대졸이상이 34.5%, 중졸이 12.6%, 초졸이하가 6.4%이다.

2(2) 어머니 교육수준 첫 번째 마디에서 오른쪽으로 분리된 48세 이상 성인 3,046명을 두 번째로 연관성이 높은 유년기환경: 어머니 교육수준으로 다시 상세하게 분리한다. 어머니 교육수준(유년기환경)이 초등학교 졸업이하인 경우는 왼쪽으로 분리한다. 이때 성인의 교육수준 분포는 초졸이하가 50.0%, 고졸이 21.2%, 중졸이 18.1%, 대졸이상이 10.7%로 나타났다. 어머니 교육수준(유년기환경)이 중학교 이상은 오른쪽으로 분리되며 이때 성인의 교육수준 분포는 대졸이상이 41.7%, 고졸이 37.6%, 중졸이 12.8%, 초졸이하가 7.9%가 된다. 오른쪽 마디는 첫 번째 분리보다 순수도가 크게 증가하여 분리가 잘 되었음을 알 수 있다. 그러나 왼쪽 마디는 더 세부적으로 분리할 필요가 있다.

깊이 3

세 번째 분리과정은 두 번째로 분리된 왼쪽마디부터 차례로 분리기준에 따라 상세하게 자식마디를 형성한다.

3(1) 만나이 두 번째 분리된 첫 번째 마디의 43세 미만 성인들을 다시 만나이로 분류한다. 만나이가 22세 미만인 성인을 왼쪽으로 분리한다. 이때 교육수준이 고졸인 비율이 95.7%로 매우 높아지므로 분리가 아주 잘 되었음을 알 수 있다. 한편 만나이가 22세 이상 43세 미만은 오른쪽으로 분리되는데, 이때도 교육수준 분포는 대졸이상이 49.7%이고 초졸이 47.4%로 잘 분리되었음을 알 수 있다.

3(2) 소득사분위수 두 번째 분리가 완료된 두 번째 마디를 다시 세부적으로 분리한다. 즉, 만나이가 43세이상 48세미만인 성인을 소득사분위수(개인) 항목으로 다시 분리한다. 소득사분위수가 하, 중하, 결측인 성인들은 왼쪽 마디로 분리된다. 이때의 교육수준 분포는 고졸이 51.3%로 가장 높고 대졸이상은 20.9%, 중졸은 17.0%, 초졸이하는 10.8%로 나타났다. 소득사분위수가 중상, 상인 성인들은 오른쪽 마디로 분리되며 이때의 교육수준 분포는 대졸이상이 48.2%, 고졸이 41.6%로 구성되어 잘 분리되었음을 알 수 있다.

3(3) 만나이 두 번째 분리가 완료된 세 번째 마디를 다시 세부적으로 분리한다. 즉, 만나이가 48세 이상이고 어머니 교육수준이 초등학교 졸업 이하인 성인들의 교육수준을 만나이 항목으로 다시 분리한다. 만나이가 77세 미만인 성인들은 왼쪽 마디로 분리되며 이 경우의 학력수준의 분포는 초졸이하가 46.3%, 고졸이 22.7%, 중졸이 19.5%, 대졸이상이 11.5%의 비율로 나타났다. 그러나 왼쪽으로 분리된 것은 교육수준을 분리하는데 크게 도움이 되지 않으므로 이 마디에서 무응답을 대체할 때 오류의 가능성이 있다. 만나이가 77세 이상인 성인들은 오른쪽 마디로 분리된다. 이때의 학력수준은 초졸이하가 84.5%로 대부분을 차

지하므로 두 번째 분리 후의 구성비에 비해서 순수도가 크게 증가하여 분리가 잘 된 것으로 판단된다.

3(4) 아버지 경제활동 두 번째 분리가 완료된 네 번째 마디를 다시 세부적으로 분리한다. 즉, 만나이가 48세 이상이고 어머니 교육수준이 중학교 졸업 이상인 성인들의 교육수준을 유년기 아버지 경제활동상태 항목으로 다시 분리한다. 아버지 경제활동상태가 직업이 있는 경우는 왼쪽 마디로 분리되며 이 경우의 학력수준의 분포는 대졸이상이 46.1%, 고졸이 38.2%, 중졸이 11.1%, 초졸이하가 4.6%의 비율로 나타났다. 그러나 왼쪽으로 분리된 것은 교육수준을 분리하는데 크게 도움이 되지 않아 이 마디에서 무응답을 대체할 때 오류가 일어날 가능성이 높을 것으로 예상된다. 아버지 경제활동상태가 일자리가 없었거나 사망, 이혼/별거 등으로 아버지가 없는 경우는 오른쪽 마디로 분리된다. 이때의 학력수준 분포는 초졸이하가 36.0%, 고졸이 32.0%, 중졸이 28.0%로 구성된다.

깊이 4

세 번째 분리가 완료된 후 네 번째 분리가 시작된다. 네 번째 분리 과정은 세 번째로 분리된 왼쪽마디부터 차례로 분리기준에 따라 더 상세하게 자식마디를 형성해 나간다.

4(4) 아버지 교육수준 세 번째 분리가 완료된 네 번째 마디를 다시 세부적으로 분리한다. 즉, 만나이가 43세 이상 48세 미만이고 소득사분위수가 중상 또는 상인 성인들의 교육수준을 아버지 교육수준 항목으로 다시 분리한다. 아버지 교육수준이 고등학교 졸업이상이면 오른쪽으로 분리되며 이때의 교육수준 분포는 대졸이상이 66.0%, 고졸이

33.0%로 고졸이상의 구성비가 99.0%가 된다. 따라서 분리가 잘 되었음을 알 수 있다.

4(5) 아버지 교육수준 세 번째 분리가 완료된 다섯 번째 마디를 다시 아버지 교육수준으로 분리한다. 즉, 아버지 교육수준이 중학교 졸업이하 또는 결측이면 왼쪽으로 분리되며 고등학교 졸업이상이면 오른쪽으로 분리된다. 그러나 순수도가 크게 증가하지 않는다.

4(7) 만나이 세 번째 분리가 완료된 일곱 번째 마디를 다시 만나이 항목으로 분리한다. 만나이가 50세 미만은 왼쪽 마디로, 50세 이상은 오른쪽 마디로 분리하며, 교육수준 분포는 고졸이상이 각각 95.7%와 81.2%로 대부분을 차지한다. 이는 만나이가 교육수준을 잘 분리하고 있다는 것을 의미한다.

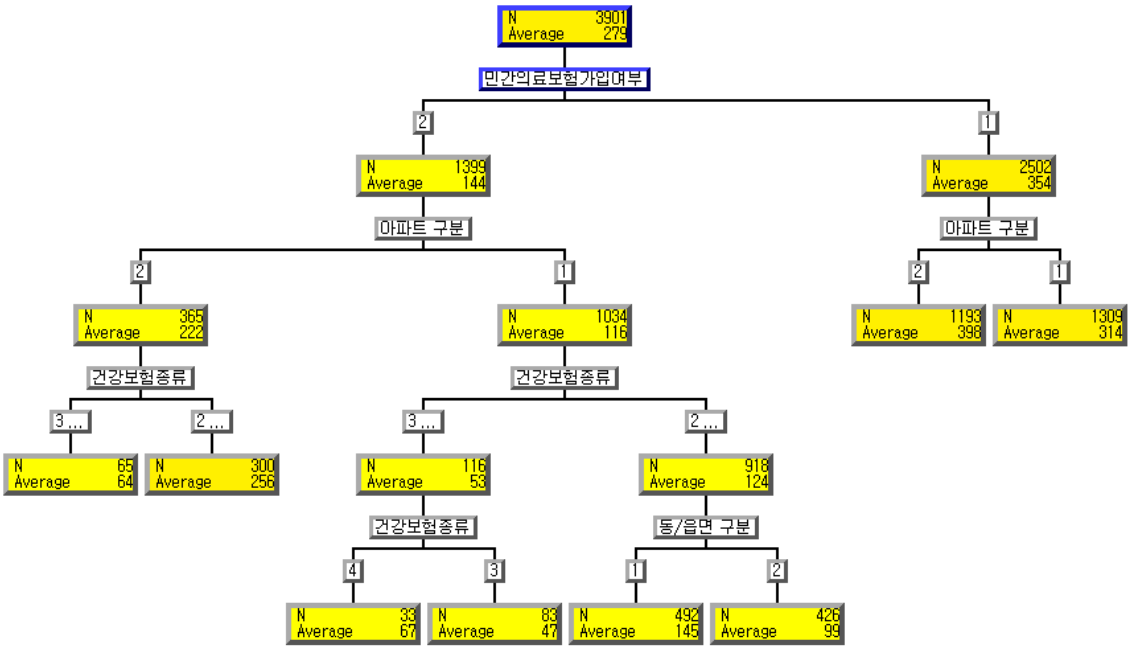
이와 같이 목표변수가 교육수준인 경우 대체에 사용할 대체군(보조변수)을 선택하는 과정을 의사결정나무를 통하여 자세히 설명하였다. 요약하면, 만나이, 어머니 교육수준, 아버지 교육수준, 소득사분위수, 아버지 경제활동 등은 교육수준을 분리하는데 연관성이 크다. 이 중에서 만나이는 제일 중요한 보조변수임을 알 수 있다. 분리과정에서 순수도가 증가하여 분리가 잘된 마디들은 대체 시 정확도가 높아지지만, 분리가 잘 안된 경우에는 대체의 정확도가 떨어질 것으로 예상된다.

월평균가구총소득, 주관적 건강상태, 현재흡연여부, 체질량지수, 총콜레스테롤, 폐기능 판정결과, 에너지 등 나머지 7개 목표변수들에 대한 연관성분석도 교육수준에 대한 연관성분석과 유사하므로 상세한 것은 생략한다.

□ 월평균가구총소득

월평균가구총소득은 건강설문조사에서 108명이 응답하지 않았다. 월평균가구총소득은 가구조사 자료로 가구내의 모든 가구원의 소득을 합산하여 조사하므로, 가구내의 모든 가구원들에게 동일한 가구소득금액들 부여한 자료구조를 가진다.

월평균가구총소득의 보조변수는 15세 이상을 대상으로 하지만 질병관리본부와 협의하여 19세 이상의 자료를 사용하기로 하였다. 따라서 19세 이상 가구수는 3,945가구이지만 월평균가구총소득이 결측인 44가구를 제외한 3,901가구를 분석대상으로 하였다.



<그림 3> 월평균가구총소득에 대한 연관성모형(가구단위)

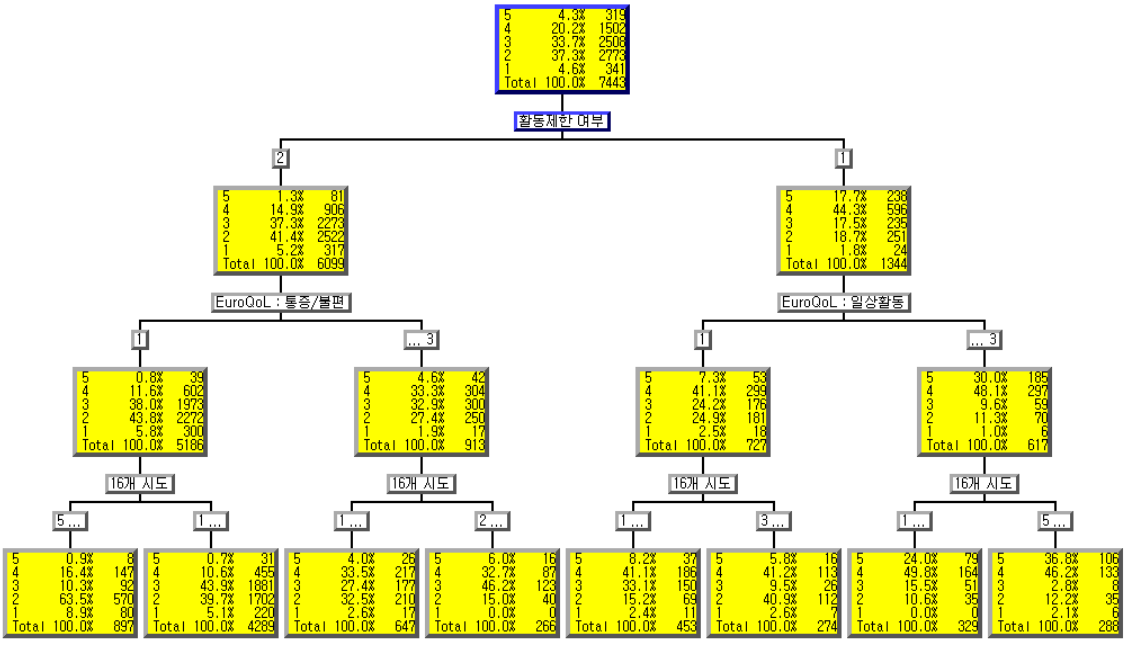
<표 7> 월평균가구총소득에 대한 연관성분석

월평균가구총소득				
깊이(연관성)	분리변수	분리값(좌)	분리값(우)	중요도
1	민간의료보험 가입여부	2, M	1	1.0000
2	아파트 구분(1) 아파트 구분(2)	2 2	1, M 1, M	0.4334
3	건강보험종류(1) 건강보험종류(2)	3, 4, 5, M 3, 4, M	1, 2 1, 2, 5	0.2510
4	건강보험종류(3) 동/읍면(4)	4, M 1, M	3 2	0.1099

월평균가구총소득 항목도 교육수준 항목에서 설명한 내용과 같은 방법으로 대체군을 설정한다. 가구단위 월평균가구총소득은 민간의료보험 가입여부와 가장 큰 연관성을 갖고 있다. 가구단위 월평균가구총소득의 전체평균은 279(만원)이며, 민간의료보험 가입여부, 아파트구분, 건강보험종류 등과 큰 연관성을 가진다. 민간의료보험에 가입(1)한 가구의 월평균가구총소득의 평균은 354(만원)이며, 민간의료보험에 가입하지 않은(2) 가구의 월평균가구총소득의 평균은 144(만원)이다. 민간의료보험 가입여부는 또한 아파트 구분에 따라 재분류된다. 민간의료보험에 가입하지 않고 아파트 구분이 아파트인 경우 월평균가구총소득은 222(만원), 일반인 경우는 116(만원)으로 가구단위 월평균가구총소득에서 상당한 차이가 있음을 알 수 있다. 민간의료보험에 가입하지 않은 경우, 아파트 구분과 건강보험종류에 의해서 월평균가구총소득이 상당한 차이를 나타내고 있다. 거주형태가 아파트인 경우 지역의료보험이나 사업장(직장)의료보험에 가입한 경우 256(만원), 의료급여 1종, 의료급여 2종, 미가입 또는 결측인 경우에는 64(만원)의 가구단위 월평균가구총소득이 있다. 아파트 구분에서 일반인 경우 지역의료보험, 사업

장(직장)의료보험에 가입하거나 또는 미가입한 경우 124(만원), 의료급여 1종, 의료급여 2종 또는 결측인 경우에는 월평균가구총소득이 53(만원)이다. 또한 지역의료보험, 사업장(직장)의료보험에 가입하거나 미가입하고 동/읍·면 구분이 동(1)인 가구가 읍·면(2)인 가구보다 가구단위 월평균가구총소득이 높은 것으로 나타났다. 민간의료보험에 가입한 경우에는 아파트 구분으로 다시 분리되며 가구단위 월평균가구총소득은 아파트가 일반인 경우보다 가구단위 월평균가구총소득이 높다. 자세한 결과는 <그림 3>과 <표 7>을 참조하면 된다.

□ 주관적 건강상태



<그림 4> 주관적 건강상태에 대한 연관성모형

주관적 건강상태는 건강설문조사에서 40명이 응답하지 않았다. 그러나 건강설문조사 응답자중 40명은 목표변수와 건강상태와 삶의 질을

나타내는 보조변수들에 모두 응답하지 않았으므로 항목무응답보다는 단위무응답으로 처리하는 것이 바람직하다. 그러나 향후 조사에서 주관적 건강상태 항목에 무응답이 발생할 경우 무응답대체를 위해서 19세 이상의 자료를 사용하여 보조변수를 선정한다.

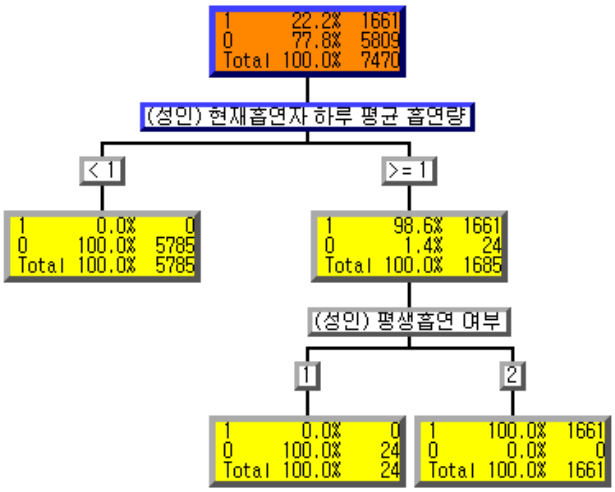
<표 8> 주관적 건강상태에 대한 연관성분석

주관적 건강상태				
깊이 (연관성)	분리변수	분리값(좌)	분리값(우)	중요도
1	활동제한 여부	2, M	1	0.9660
2	(성인) 통증/불편(1) (성인) 일상활동(2)	1, M 1, M	2, 3 2, 3	0.5052 0.3643
3	16개 시도(1) 16개 시도(2) 16개 시도(3) 16개 시도(4)	5, 6, 11~13 1, 3~6, 8~13, 16, M 1, 2, 4, 7~10, 14~16, M 1~4, 7~10, 14, 16, M	1~4, 7~10, 14~16, M 2, 7 14, 15 3, 5, 6, 11~13 5, 6, 11~13, 15	1.0000
4	16개 시도(2) 16개 시도(3) 만나이(4) 몸이 불편했던 경험유무(5) 체질량지수6) 만나이(8)	1, 3, 4, 8~10, 16, M 1, 3, 4, 8, 9, 16, M 60미만 2 19.2미만 79미만	2, 7, 14, 15 5, 6, 10~13 60이상 1, M 19.2이상, M 79이상	 0.1436 0.1751 0.0510

주관적 건강상태는 5개의 항목으로 구성되며 좋음이 37.3%, 보통이 33.7%, 나쁨이 20.2%로 세 항목이 91.2%를 차지하고 있다. 주관적 건강상태 항목은 16개 시도와 연관성이 가장 크며 활동제한 여부, 통증/불편, 일상활동 등과도 연관성이 높다. 주관적 건강상태 항목은 활동제한 여부에 따라 분리되며 아니오라고 응답한 경우에는 주관적 건강상태가 좋은 경우가 37.3%에서 41.4%로 증가하지만 활동제한 여부가 예라고 응답한 경우에는 주관적 건강상태가 나쁜 경우가 20.2%에서 44.3%로 증가한다. 또한 활동제한 여부는 통증/불편과 일상활동에 의

하여 세분된다. 통증/불편이 없음이라고 응답한 경우에는 주관적 건강상태가 좋은 경우가 43.8%로 증가하며, 통증/불편이 다소 있거나 매우 심한 경우에는 주관적 건강상태가 나쁜 경우가 33.3%로 증가한다. 또한 통증/불편과 일상활동은 16개 시도에 의하여 다시 분리되고 이는 또한 16개 시도, 최근 2주간 몸이 불편했던 경험 유무, 만나이 등에 의해 재분리 된다. 자세한 결과는 <그림 4>와 <표 8>을 참조하면 된다.

□ 현재흡연여부



<그림 5> 현재흡연여부에 대한 연관성모형

<표 9> 현재흡연여부에 대한 연관성분석

현재흡연여부				
깊이(연관성)	분리변수	분리값(좌)	분리값(우)	중요도
1	현재흡연자 하루 평균 흡연량	1미만	1이상	1.0000
2	평생흡연여부	1	2, M	0.1366

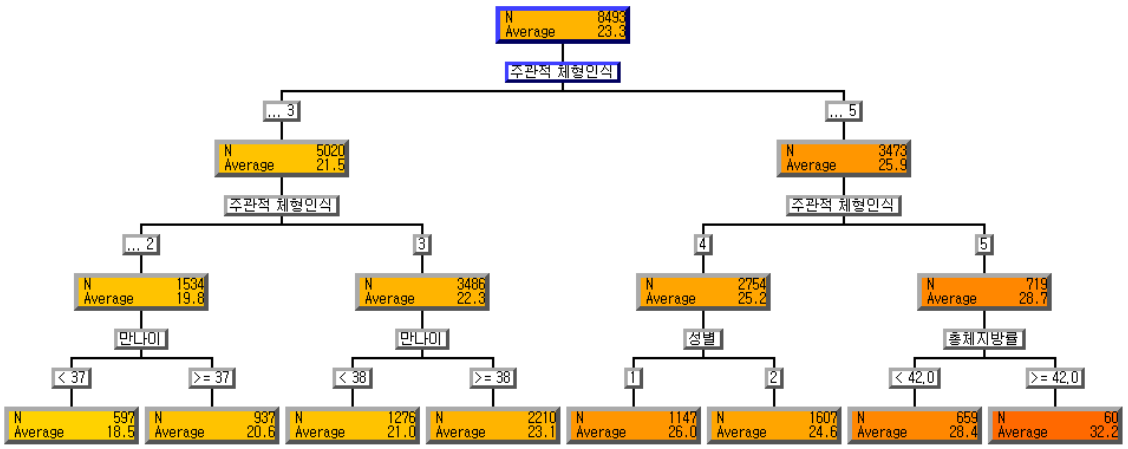
현재흡연여부는 5개의 항목으로 구성되며 현재흡연이 22.2%, 과거흡

연/비흡연이 77.8%의 비율을 차지하고 있다. 현재흡연여부 항목은 현재흡연자 하루 평균 흡연량과 가장 큰 연관성을 갖고 있으며 평생흡연여부와도 연관성이 높다. 현재흡연여부 항목은 현재흡연자 하루 평균 흡연량에 따라 먼저 분리되며 현재흡연자 하루 평균 흡연량이 1개비 이상과 1개비 미만으로 분리된다. 현재흡연자 하루 평균 흡연량이 1개비 이상인 경우에는 현재흡연인 경우가 22.2%에서 98.6%로 크게 증가하여 대다수를 차지하며 과거흡연/비흡연인 경우에는 77.8%에서 1.4%로 크게 감소하므로 분리가 아주 잘 되었다. 또한 현재흡연자 하루 평균 흡연량이 1개비 이상인 경우는 평생흡연여부에 의하여 잘 분리되었음을 알 수 있다. 자세한 결과는 <그림 5>와 <표 9>를 참조하면 된다.

□ 체질량지수

체질량지수의 전체평균은 23.3(kg/m^2)이며 체질량지수를 분리하는데 주관적 체형인식이 가장 중요한 역할을 한다. 주관적 체형인식이 매우 마른 편, 약간 마른 편, 보통이라고 응답한 응답자들의 체질량지수 평균은 21.5(kg/m^2), 약간 비만, 매우 비만이라고 응답한 응답자들의 체질량지수 평균은 25.9(kg/m^2)이다. 또한 주관적 체형인식이 매우 마른 편이거나 약간 마른 경우는 19.8(kg/m^2), 보통인 경우는 22.3(kg/m^2), 약간 비만인 경우는 25.2(kg/m^2)의 체질량지수 평균값을 가지며, 매우 비만인 경우의 체질량지수 평균은 28.7(kg/m^2)로 다른 체형인식에 비해 상당히 크다. 그리고 주관적 체형인식이 보통이하인 경우에는 만나이로 다시 세분류되며 약간 비만과 매우 비만인 경우에는 성별과 총체지방률로 다시 분리된다. 즉, 주관적 체형인식이 매우 비만이고 총체지방률이 42.0(%)이상인 경우의 체질량지수 평균은 32.2(kg/m^2)로 가장

높다. 자세한 결과는 <그림 6>과 <표 10>을 참조하면 된다.

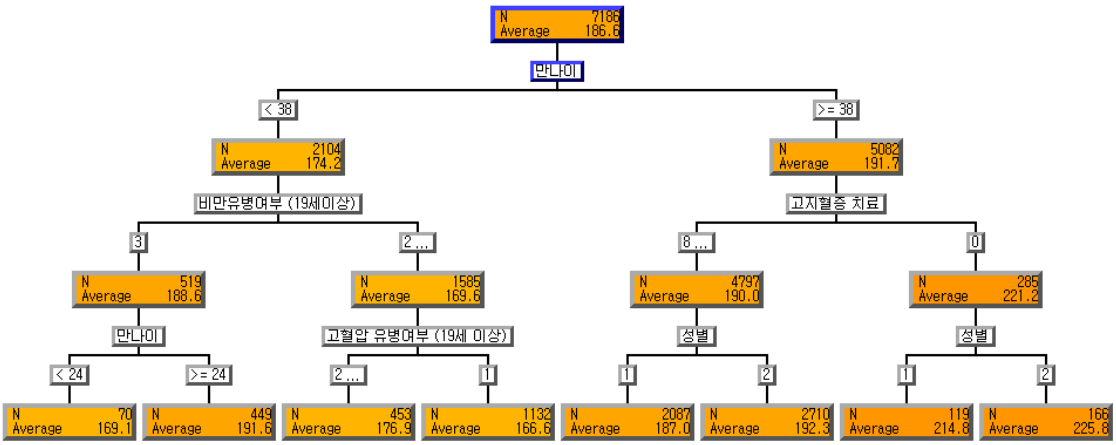


<그림 6> 체질량지수에 대한 연관성모형

<표 10> 체질량지수에 대한 연관성분석

체질량지수				
깊이(연관성)	분리변수	분리값(좌)	분리값(우)	중요도
1	주관적 체형인식	1, 2, 3, M	4, 5	1.0000
2	주관적 체형인식(1) 주관적 체형인식(2)	1, 2 4, M	3, M 5	
3	만나이(1) 만나이(2) 성별(3) 총체지방률(4)	37미만 38미만 1 42미만, M	37이상, M 38이상, M 2, M 42미만	0.3309 0.2352 0.2847
4	만나이(1) 총체지방률(2) 성별(3) 총체지방률(4) 만나이(5) 총체지방률(6) 성별(7)	17미만 29.4미만, M 1 35.3미만, M 19미만 36미만, M 1	17이상, M 29.4이상 2, M 35.3이상 19이상, M 36이상 2, M	

□ 총콜레스테롤



<그림 7> 총콜레스테롤에 대한 연관성모형

<표 11> 총콜레스테롤에 대한 연관성분석

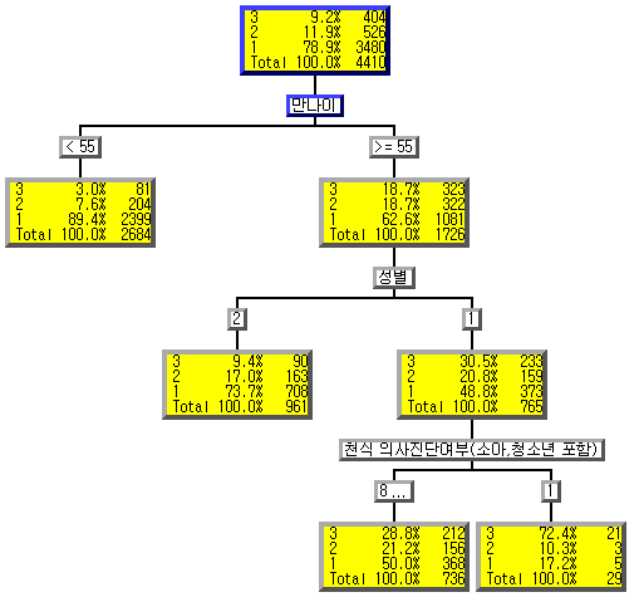
총콜레스테롤				
깊이(연관성)	분리변수	분리값(좌)	분리값(우)	중요도
1	만나이	38미만	38이상, M	1.0000
2	비만유병여부(1) 고지혈증 치료(2)	3,M 1, 8, M	1, 2 0	0.4384 0.5970
3	만나이(1) 고혈압 유병여부(2) 성별(3) 성별(4)	24미만 2, 3, M 1 1	24이상, M 1 2, M 2, M	 0.2152 0.2382
4	만나이(2) 만나이(3) 만나이(4) 만나이(5) 만나이(6)	27미만 29미만 59미만, M 49미만 52미만	27이상, M 29이상, M 59이상 49이상, M 52이상, M	

총콜레스테롤 수치의 전체평균은 186.6(mg/dL)이며 총콜레스테롤 항목은 만나이와 가장 연관성이 높다. 다음으로 고지혈증 치료, 비만유병여부, 성별, 고혈압 유병여부 순이다. 따라서 총콜레스테롤 항목은 가

장 중요한 역할을 하는 만나이에 따라 먼저 분리된다. 만나이가 38세 미만은 174.2(mg/dL), 만나이가 38세 이상은 191.7(mg/dL)의 총콜레스테롤 평균값을 갖는다. 또한 만나이가 38세 미만은 비만유병여부, 38세 이상은 고지혈증 치료로 다시 세분류된다. 만나이가 38세 이상이고 고지혈증 치료가 없으며 성별이 여자인 경우 평균 총콜레스테롤은 225.8(mg/dL)로 가장 높다. 자세한 내용은 <그림 7>과 <표 11>을 참조하면 된다.

□ 폐기능 판정결과

폐기능 판정결과는 3개의 항목으로 구성되며 폐기능 정상이 78.9%, 제한성 환기장애가 11.9%이고 폐쇄성 환기장애가 9.2%이다. 폐기능 판정결과와 가장 연관성이 높은 항목은 만나이이며, 다음으로 성별, 천식 의사진단여부 순이다. 따라서 폐기능 판정결과 항목은 연관성이 가장 높은 만나이로 먼저 분리된다. 만나이가 55세 미만인 경우에는 폐기능 정상 비율이 78.9%에서 89.4%로 증가하며 만나이가 55세 이상인 경우에는 폐기능 정상 비율이 78.9%에서 62.6%로 감소하고 제한성 환기장애와 폐쇄성 환기장애 비율이 각각 18.7%로 증가한다. 또한 만나이가 55세 이상인 경우 성별로 더 세분화되며 성별이 남자인 경우에는 폐기능 정상 비율은 48.8%로 크게 감소하지만 제한성 환기장애와 폐쇄성 환기장애 비율이 각각 20.8%와 30.5%로 크게 증가한다. 만나이가 55세 이상이고 성별이 남자이며 천식 의사진단여부가 있는 응답자들은 폐쇄성 환기장애 비율이 9.2%에서 72.4%로 크게 증가함을 알 수 있다. 자세한 결과는 <그림 8>과 <표 12>를 참조하면 된다.



<그림 8> 폐기능 판정결과에 대한 연관성모형

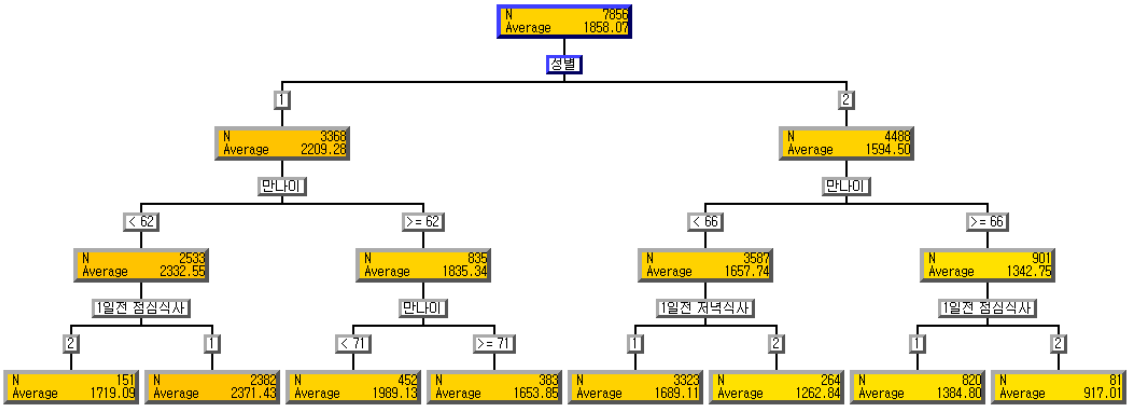
<표 12> 폐기능 판정결과에 대한 연관성분석

폐기능 판정결과				
깊이(연관성)	분리변수	분리값(좌)	분리값(우)	중요도
1	만나이	55미만, M	55이상	1.0000
2	성별(2)	2, M	1	0.6356
3	천식 의사진단여부(4)	0, 8, M	1	0.2752

□ 에너지

에너지의 전체평균은 1,858.07(Kcal)이고 에너지 항목은 성별과 가장 연관성이 높다. 다음으로 만나이, 조사 1일전 점심식사, 조사 1일전 아침식사, 조사 1일전 저녁식사 순이다. 에너지 항목은 가장 중요한 성별에 따라 먼저 분리된다. 에너지 평균은 남자 2,209.28(Kcal), 여자 1,594.50(Kcal)로 남자와 여자의 에너지 평균 간에 차이가 크다. 또한 성별은 만나이로 더 세분되며 남자이고 만나이가 62세 미만은 2,332.55(Kcal), 남자이고 만나이가 62세 이상은 1,835.34(Kcal), 여자이

고 만나이가 66세 미만은 1,657.74(Kcal), 그리고 여자이고 만나이가 66세 이상은 1,342.75 (Kcal)의 에너지 평균값을 갖는다. 또한 에너지는 조사 1일전 점심식사, 조사 1일전 아침식사, 조사 1일전 저녁식사, 그리고 배추와 연관성이 있다. 자세한 결과는 <그림 9>와 <표 13>을 참조하면 된다.



<그림 9> 에너지에 대한 연관성모형

<표 13> 에너지에 대한 연관성분석

에너지				
깊이(연관성)	분리변수	분리값(좌)	분리값(우)	중요도
1	성별	1	2, M	1.0000
2	만나이(1) 만나이(2)	62미만, M 66미만, M	62이상 66이상	0.5863
3	조사 1일전 점심식사(1) 만나이(2) 조사 1일전 저녁식사(3) 조사 1일전 점심식사(4)	2, M 71미만, M 1, M 1, M	1 71이상 2 2	0.4077 0.2472
4	조사 1일전 아침식사(1) 조사 1일전 아침식사(2) 조사 1일전 점심식사(3) 배추(4) 조사 1일전 점심식사(5) 조사 1일전 아침식사(6) 조사 1일전 아침식사(7)	2, M 1, M 1, M 0~6, M 1, M 1, M 1, M	1 2 2 7~9 2 2 2	0.2897 0.0961

2) 보조변수 선정결과

제4기 3차년도(2009) 국민건강영양조사 자료에서 연관성분석 실시 결과 무응답 대체변수와 선정된 보조변수는 <표 14>와 같다. 보조변수는 대체변수와 연관성이 높은 중요도 순으로 정렬하였다.

<표 14> 무응답 대체변수와 보조변수

항목	무응답률(%)	연관성분석 결과 보조변수
교육수준	0.438	· 만나이, 어머니 교육수준, 아버지 교육수준, 소득사분위수, 아버지 경제활동 상태
월평균가구총소득[1)]	1.182	· 민간의료보험 가입여부, 아파트 구분, 건강보험종류, 동/읍면
주관적 건강상태	0.398	· 16개 시도, 활동제한 여부, 통증/불편, 일상활동, 몸이 불편했던 경험 유무, 만나이, 체질량지수
현재 흡연여부[2)]	0.546	· 현재흡연자 하루 평균 흡연량, 평생흡연여부
체질량지수	0.467	· 주관적 체형인식, 만나이, 총체지방률, 성별
총콜레스테롤[3)]	5.786[3)] 4.670[2)]	· 만나이, 고지혈증 치료, 비만유병여부, 성별, 고혈압 유병여부
폐기능 판정결과[2)]	41.377[4)]	· 만나이, 성별, 만성폐쇄성 폐질환 증상, 기관지 확장증 유병여부
에너지	0.064	· 성별, 만나이, 조사 1일전 점식식사, 조사 1일전 아침식사, 조사 1일전 저녁식사, 배추

1) 가구단위 월평균가구총소득, 2) 19세 이상 성인, 3) 10세 이상
4) 폐기능 무응답 및 판정불가 비율

(2) 무응답대체 및 결과분석

제4기 3차년도(2009) 국민건강영양조사 자료에 대한 연관성분석을 실시하여 8개 목표변수에 사용할 보조변수들을 선정하고 각 항목별로 선정된 보조변수를 이용하여 대체군을 형성하였다. 이 절에서는 구성된 무응답 대체군을 이용하여 항목별 무응답을 대체한다. 그러나 대체할 변수의 항목 무응답률이 매우 낮거나 목표변수와 보조변수가 모두 무응답인 경우, 또는 실제로는 항목 무응답이 발생하지 않은 경우는 분석에서 제외하였다. 따라서 8개의 무응답을 대체할 목표변수 중에서 무응답률이 높은 폐기능 판정결과와 총콜레스테롤 2개 항목에 대해서만 대체를 실시하였다.

1) 폐기능 판정결과

폐기능검사는 만 19세이상 성인 7,538명을 대상으로 실시하였다. 그러나 검사과정이 매우 힘들어서 검사 자체를 거부하거나 제외(통증이 있는 귀감염, 동맥류/망막박리/뇌졸중/심장마비/기흉 진단경험, 3개월내 안과수술, 3개월내 개심술/개복술, 3개월내 뇌졸중/심장마비 경험, 결핵 진단 등)되는 비율이 높아 1차 무응답이 발생한다. 또한 검사에 참여했어도 검사의 정도관리 요건(적합성 만족 그래프 수 2개 이상, 재현성)을 충족하지 못하면 판정불능으로 분류되어 유병률 산출시 제외되어 2차 무응답이 발생한다. 1차 무응답과 2차 무응답을 합한 총무응답은 3,119명(약 41.4%)으로 매우 많아 폐쇄성폐질환 유병률 추정시 편향을 일으키고, 무응답층과 응답층은 대체변수별로 통계적으로 유의한 차이가 있게 된다. 그러므로 무응답률이 높은 변수의 경우 대체의 효

율이 감소할 수 있다. 하지만 향후 대체전후 변화에 대한 연구검토가 필요하여 포함하였다.

폐기능 판정결과는 3개의 항목으로 구성되며 무응답을 제외한 이용가능한 자료를 사용한 경우 폐기능 정상이 78.91%, 제한성 환기장애가 11.93%이고 폐쇄성 환기장애가 9.16%가 되었다. 연관성분석결과 폐기능 판정결과 항목은 만나이, 성별, 천식 의사진단 여부를 보조변수로 사용하여 대체군을 형성하였다. 이 대체군을 이용하여 폐기능 판정결과 항목무응답을 대체하였다. 무응답 대체는 만 19세 이상 성인 7,538명을 대상으로 실시하였다.

폐기능 판정결과에 대한 대체결과는 <표 15>에 정리되어 있다. 대체방법으로는 확률대체, 핫덱대체와 계층적 핫덱대체 방법을 사용하였다. 대체군을 이용한 확률대체와 핫덱대체를 실시할 때에는 대체군 내에서 응답자 수가 무응답자 수보다 커야 한다. 대체군 내에서 무응답자 수가 응답자 수보다 많은 경우에는 무응답을 대체할 기증자(donor)를 발견할 수 없는 경우가 생기므로 이때에는 계층적 핫덱대체 방법을 사용하는 것이 바람직하다.

<표 15> 무응답대체 후 폐기능 판정결과의 분포

응답항목	이용 가능한 자료	단순임의 확률대체	대체군을 이용한 확률대체	단순임의 핫덱대체	대체군을 이용한 핫덱대체	계층적 핫덱대체
1. 폐기능 정상	3,487 (78.91%)	5,950 (78.93%)	5,973 (79.24%)	5,950 (78.93%)	5,997 (79.56%)	5,975 (79.27%)
2. 제한성 환기장애	527 (11.93%)	904 (11.99%)	912 (12.10%)	903 (11.98%)	892 (11.83%)	882 (11.70%)
3. 폐쇄성 환기장애	405 (9.16%)	684 (9.07%)	653 (8.66%)	685 (9.09%)	649 (8.61%)	681 (9.03%)

2) 총콜레스테롤

연관성분석결과 총콜레스테롤은 만나이, 고지혈증 치료, 비만유병여부, 성별, 고혈압 유병여부 등 5개의 보조변수로 대체군을 형성할 수 있다. 이 대체군을 이용하여 총콜레스테롤 항목무응답을 대체하였다. 대체방법은 평균대체, 핫덱대체, 계층적 핫덱대체 방법을 사용하였다. 무응답대체는 만 19세 이상 성인 7,538명을 대상으로 실시하였고 대체결과는 <표 16>에 정리되어 있다.

총콜레스테롤은 3개의 항목으로 구성되며 무응답을 제외한 이용 가능한 자료의 표본평균은 186.5977(mg/dL), 표준오차는 0.4209(mg/dL)이다. 이용 가능한 자료를 사용한 경우와 대체방법들의 평균과 표준오차는 큰 차이를 보이지 않는다.

<표 16> 무응답대체 후 총콜레스테롤의 평균과 표준오차

분석(대체)방법	총콜레스테롤(HE_chol)	
	평균	표준오차
이용 가능한 자료 분석	186.5977	0.4209
평균대체	186.5977	0.4012
대체군을 이용한 평균대체	186.7054	0.4021
단순임의 핫덱대체	186.6016	0.4100
대체군을 이용한 핫덱대체	186.6940	0.4117
계층적 핫덱대체	186.4912	0.4100

3장 4절에서 언급한 바와 같은 대체조건 하에서 <표 15>와 <표 16>에 의하면 응답자료 만을 사용한 경우와 무응답을 대체한 경우의 차이는 미미하다. 따라서 본 자료에서의 대체방법은 연구자가 대체의 편리성을 감안해서 선정하는 것이 바람직하다.

4장. 결론 및 제안

보건복지부 질병관리본부에서 생산하고 있는 “국민건강영양조사”는 국민의 건강수준, 건강관련 의식 및 행태, 식품 및 영양섭취 실태를 파악하기 위한 조사이다. 본 통계는 2010년도 정기품질진단결과 품질이 우수한 것으로 평가되었고 관련 분야에서 매우 널리 사용되고 있다. 그러나 항목무응답 대체에 대한 조치가 미흡하므로 이에 대한 개선이 이루어지면 현재보다 나은 품질의 통계를 생산할 수 있다.

이를 위해서 제4기 3차년도(2009) 국민건강영양조사 자료를 이용해서 무응답 현황을 파악하고 대체방법을 검토하였다. 항목무응답 대체변수와 대체에 사용될 보조변수들을 선정하고 이들을 이용해서 대체군을 설정하였다. 2가지 목표변수에 대해서 무응답대체를 실시하고 그 결과를 검토하였으며, 향후 연구를 위한 몇 가지 사항들을 제안한다.

□ 결론

○ 현행 “국민건강영양조사”의 무응답현황을 단위무응답과 항목무응답으로 분류하여 살펴보았다

○ 무응답대체 방법을 단위무응답과 항목무응답으로 나누어 소개하고, 이는 향후 무응답대체 시 활용할 수 있다.

○ 무응답을 대체할 8개의 목표변수와 관련된 보조변수들에 대한 특성을 분석하고 대체기준을 설정하였다.

○ 무응답대체군 설정을 위해서 CHAID 알고리즘을 사용하여 8개 목표변수와 관련 보조변수들 간의 연관성분석을 실시하였다.

○ 8개 변수 중 대체 가능한 2개 목표변수에 대한 대체를 실시하고

그 결과를 분석하였다.

□ 제안

본 연구에서 다룬 내용과 더불어 다음 사항들에 대한 추가적인 연구는 “국민건강영양조사”의 항목무응답 대체에 도움을 주어 전반적으로 통계품질을 향상시킬 것으로 기대된다.

○ “국민건강영양조사”의 3개 조사에 대해서 무응답이 없는 완전자료(또는 모집단의 특성과 유사한 가상모집단)를 이용한 모의실험을 통해 대체방법들을 비교연구
○ 대체 후 추정치에 대한 분산추정의 문제를 검토하여 대체방법의 선택에 활용하고 가중치문제도 고려
○ 응답층과 무응답층의 특성별 차이를 나타내는 변수를 이용한 무응답대체 방법에 대한 검토
○ 기타 사항
  – 자료에서 결측과 모름의 의미를 혼동할 우려가 있으므로 결측으로 통일하여 원시자료와 코드북을 작성
  – 원시자료 이용지침서의 원시자료 구성에서 각 조사항목별로 실제 조사대상 연령을 명시(일부 조사항목은 설문지의 조사대상 연령과 실제 자료의 대상 연령이 상이함)
  – 조사 부문별로 거의 모든 항목이 무응답인 경우는 단위무응답으로 처리하고 원시자료에서 제외(단, 응답층과 무응답층의 특성 파악에 활용 가능)

참 고 문 헌

김규성(2000), “무응답 대체 방법과 대체 효과”, *조사연구*, 제1권 2호, 1-14.

김규성・이기재・김진(2005), “농어가경제조사에서 가중핫덱 무응답 대체방법의 활용”, *응용통계연구*, 제18권 2호, 311-328.

김영원・이주원(2003), “CART를 활용한 결측값 대체방법: 인구주택총조사 혼인상태 항목을 중심으로”, *조사연구*, 제4권 2호, 1-21.

김영원・조선경(1996), “표본조사에서 항목 무응답 대체 방법”, *한국통계학회논문집*, 제3권 3호, 145-159.

복지부・질병관리본부(2010), *2009 국민건강통계 - 국민건강영양조사 제4기 3차년도(2009)*

송주원・안형진(2009), *무응답 자료처리 및 분석*, 통계교육원

이현정(2009), “인구주택총조사 무응답 처리기법 연구(Ⅰ)”, *연구보고서*, 통계개발원.

이현정・최필근(2009), “인구주택총조사 무응답 처리기법 연구(Ⅱ)”, *연구보고서*, 통계개발원.

조사통계연구회(2000), *무응답오차*, 자유아카데미.

최필근(2009a), “농업총조사 항목간 연관성 분석 및 대체군(보조변수) 개발”, *연구보고서*, 통계개발원.

최필근(2009b), “농업총조사 무응답 대체기법 연구(Ⅰ)”, *연구보고서*, 통계개발원.

최필근(2009c), “농업총조사 무응답 대체기법 연구(Ⅱ)”, *연구보고서*, 통계개발원.

최필근(2010), “출생전후기 사망통계의 무응답 대체기법”, *연구보고서*, 통계개

발원.

최필근(2011), “경제총조사 항목 무응답 대체방법 연구”, *최종보고서*, 통계개발원.

통계청(2010), “국민건강영양조사 2010년 정기통계품질진단”, *연구용역최종보고서*.

Andridge, R. R. and Little, R. J. A.(2010), "A Review of Hot Deck Imputation for Survey Non-response", *International Statistical Review*, Vol. 78, No. 1, 40-64.

Brieman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J.(1984), *Classification and Regression Tress*, Chapman & Hall.

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

Norholt, E. S.(1998), "Imputation: method, simulation experiment and practical examples", *International Statistical Review,* Vol. 66, No. 2, 157-180.

Rao, J. N. K. and Shao, J.(1992), "Jackknife Variance Estimation with Survey Data under Hot Deck Imputation", *Biometrika,* Vol. 79, No. 4, 811-822.

Rubin, D. B. and Little, J. A.(1986), *Statistical Analysis with Missing Data*, John Wiley & Sons, New York.

Sande, I. G.(1979), “A Personal View of Hot Deck Imputation Procedures”, *Survey Methodology*, Vol. 5, 238-258.

SAS Institute Inc.(2004), *Getting Started with SAS Enterprise Miner 4.3*, Cary, NC, USA.