제3장

경제총조사 항목 무응답 대체방법 연구
-숙박 및 음식점업 조사항목 중심으로-

최 필 근

제1절 서 론

1. 연구배경 및 목적

통계는 각종 국가정책과 지역발전의 계획수립을 위한 중요한 자료이다. 특히 통계청에서 2011년 5월에 최초로 실시하는 경제총조사는 국가 전체 산업에 대한 구조와 분포, 경영실태 등에 관한 사항을 종합적으로 파악하는 대규모 통계(전수)조사이다. 조사결과는 경제 및 산업정책 수립의 기초자료, 각종 통계의 모집단자료, 연간 및 월간통계의 기준점(Bench-Mark)자료, 소지역단위(지역) 통계자료 등으로 활용될 예정이다. 이와 같이 경제총조사 자료는 매우 중요한 역할을 하므로 정확한 통계생산을 위하여 자료의 품질을 높이기 위한 노력이 계속적으로 이루어져야 할 것이다. 자료의 품질을 향상시키기 위해서는 표본의 설계에서부터 자료를 제공하기까지의 모든 단계에서 정확하게 이루어져야 한다. 하지만, 이러한 일련의 과정에서 의도하지 않은 오차들이 발생하기 마련이며 그 중 하나가 무응답으로 인해서 발생하게 된다.

최근의 조사환경을 고려하면 계속적으로 무응답 비율이 높아질 가능성이 있다. 1인·노인·맞벌이 가구의 증가와 같은 인구구조학적인 변화가 이루어지고 있으며, 민감한 항목에 대하여 노출하지 않으려는 의도가 강하여 항목의 특성에 따라 무응답 비율이 높아질 수 있음을 고려해야 할 것이다. 경제조사에서는 매출액과 같은 항목이 실제로 조사하기도 어려우며 무응답 비율도 높게 발생하고 있는 실정이다. 본 연구에서 사용할 2010년 기준 경제총조사 시범예행조사(숙박 및 음식점업)에서도 매출액 항목은 대략 20%의 무응답 비율을 보이고 있다. 조사 특성상 시범예행조사와 본조사와는 다소 차이가 있겠지만 민감한 항목에 대한 무응답 비율이 높아질 것이라는 점은 변함이 없을 것

으로 판단된다. 물론 인터넷조사를 비롯하여 응답률을 높이기 위한 조사방법론적인 연구를 하는 것도 중요하지만 모든 항목에의 응답에는 한계가 있으므로 근본적으로 해결할 수 있는 방안을 마련하기 위한 노력이 필요할 것이다. 이러한 방안 중의 하나가 무응답을 처리하기 위한 연구라고 할 수 있다.

한국보다 조사환경이 열악하다고 할 수 있는 미국, 캐나다, 호주 등에서는 무응답 처리를 위한 연구가 장기적인 계획 하에 진행 중에 있으며, 이미 체계적인 시스템이 구축된 나라도 적지 않다. 한국의 통계청에서도 2000년 이후 무응답 처리에 대한 관심이 지속적으로 증가하여 이와 관련된 연구가 계속 진행되고 있다. 특히, 가장 중요한 조사라고 할 수 있는 인구주택총조사와 농업총조사에 대한 무응답 대체기법이 연구되어 2010년 인구주택 및 농업총조사 자료의 무응답 처리를 체계적으로 할 수 있을 것으로 판단된다.

조사 자료에서 발생할 수 있는 무응답의 형태는 조사대상으로부터 얻은 정보가 전혀 없는 단위 무응답(unit nonresponse)과 특정 항목 값의 정보가 없는 항목 무응답(item nonresponse)으로 나누어진다. 무응답의 처리는 발생형태별로 적절하게 이루어져야 하는데 단위 무응답의 경우는 주로 가중치 조정 방법을 사용하고, 항목 무응답의 경우에는 적절한 값을 채워 넣기 위한 여러 가지 대체법을 이용하게 된다. 본 연구에서 진행될 경제총조사 무응답 처리 연구는 항목 무응답의 대체에 초점을 맞추고자 한다.

항목 무응답 연구를 위해서는 각 항목과 연관성이 높은 항목을 찾기 위한 과정이 선행되어야 한다. 이 항목들의 정보를 이용하여 대체하고자 하는 항목의 값을 찾게 되는데 이를 일반적으로 대체군이라 한다. 대체군은 모형을 추정할 때 보조변수와 같은 역할을 하며 대체의 정확도를 결정하는 중요한 요인이라 할 수 있다. 따라서 대체군을 찾기 위한 노력도 동시에 진행되어야 할 것이다. 주로 의사결정나무 분석의 CHAID 및 CART 알고리즘, 카이제곱 독립성 검정, 회귀분석의 변수선택 방법 등을 이용하여 분석할 수 있다. 그 다음 과정은 자료의 특성에 가장 적합한 대체기법을 선택하는 것이다. 경제총조사 항목의 경우 범주형과 연속형 항목이 혼합되어 있다. 일반적으로 대체하고자 하는 항목이 범주형인 경우에는 핫덱 대체 및 이와 유사한 최근방 기증자(donor) 대체방법을, 연속형인 경우에는 회귀 대체, 비 대체, 평균 대체 방법을 적용할 수 있다. 하지만, 이러한 방법들은 조사 자료의 특성, 무응답 비율, 대체군(보조변수) 사용여부 등 주어진 환경에 따라서 대체의 정확도가 다르게 나타날 수 있으므로 주어진 자료에 가장 적절하게 적용될 수 있는 방법이 선택되어져야 할 것이다. 이는 다양한 모의실험을 통하여 선택할 수 있을 것이다.

본 연구에서는 경제총조사 조사표 중에서 숙박 및 음식점업의 주요 항목에 대하여 무응답 대체를 위한 대체군 및 대체기법을 개발하고자 한다. 숙박 및 음식점업 이외의

조사표도 이와 거의 유사한 항목들로 구성되어 있으며 본 연구에서 제시된 방법을 사용할 수 있을 것으로 판단된다. 그리고 본 연구를 통해서 체계적이며 정확도 높은 무응답 대체기반을 마련하여 경제총조사 자료의 품질향상에 기여할 수 있기를 기대한다.

2. 연구내용 및 방법

본 연구의 목적은 경제총조사(숙박 및 음식점업)의 주요 항목(종사자수, 영업개월수, 휴무일수, 일일영업시간, 연면적, 매출액, 영업비용)의 무응답 대체를 위한 대체군 및 대체기법을 개발하는데 있다. 이러한 작업을 수행하기 위해서 2010년 기준 경제총조사 시범예행조사(숙박 및 음식점업) 자료를 이용할 것이다. 자료에 관한 내용은 제1절 후반부에 간략하게 언급하고자 한다.

첫 단계인 대체군(보조변수) 선택은 의사결정나무 방법인 CHAID 알고리즘을 이용해 연관성분석을 실시한 후 결정한다. 다른 방법들과의 비교는 이전의 연구에서 밝힌 바 있다. 『무응답 처리를 위한 방법론 연구(Ⅰ)(통계개발원, 2009)』의 『농업총조사 항목간 연관성 분석 및 대체군(보조변수) 개발(최필근)』과 『인구주택총조사 무응답 대체기법 연구(Ⅱ)(이현정, 최필근)』을 참조하면 본 연구에서 사용될 CHAID 알고리즘의 우수성을 알 수 있을 것이다. 다음 단계는 경제총조사 자료에 가장 적절한 대체기법을 선정하는 것이다. 대체하기 위한 항목들은 연속형 항목과 범주형 항목으로 구분된다. 따라서 각각의 형태에 적합한 대체방법들을 적용하고자 한다. 연속형 항목의 경우 평균 대체, 회귀 대체, 비 대체, 응용 핫덱 대체 방법들을 이용하며, 범주형 항목은 최빈값 대체, 응용 핫덱 대체방법들을 이용하여 대체 정확도를 비교·분석할 것이다. 이를 위하여 다양한 모의실험을 실시하고 각 방법들에 대하여 대체전후의 평균차이, 추정량의 표준편차를 평균값으로 나누어준 변동계수(CV), 구성비 변화 등을 고려하여 경제총조사 자료에 가장 적합한 대체기법을 제시할 것이다. 그리고 본 연구의 모든 결과를 이용하여 시범예행조사(숙박 및 음식점업) 자료에 적용해 볼 것이다.

본 연구를 위하여 제2절에서는 연관성 분석을 위한 CHAID 알고리즘을 설명하고 각 항목에 대한 분석결과 및 선정된 대체군을 제시한다. 제3절에서는 본 연구에 사용하고자 하는 무응답 대체방법을 자세하게 소개한다. 제4절에서는 각 항목에 대해 다양한 대체방법으로 모의실험을 실시하여 경제총조사에 적용할 대체방법을 제시한다. 제5절에서는 선정된 대체방법을 경제총조사 시범예행조사(숙박 및 음식점업) 자료에 적용하고 그 결과를 기술한다. 마지막으로 제6절에서는 연구의 최종적인 결론을 요약하고 본 보고서를 마무리 하고자 한다.

3. 자료 설명

본 연구에서 사용하게 될 2010년 기준 경제총조사 시범예행조사 자료는 2010년 6월 서울특별시(강서구), 부산광역시(부산진구), 제주도(제주시, 서귀포시) 3개 시·도(4개 시군구)에서 2,005개 조사구의 109천개 사업체를 대상으로 조사가 실시된 자료이다. 이 중 산업대분류 숙박 및 음식점업 자료를 이용하여 연관성분석 및 모의실험을 실시할 것이다. 숙박 및 음식점업 이외의 산업대분류에 대해서는 본 연구의 결과를 이용하여 적용할 수 있으므로 후에 논의하면 될 것으로 판단된다.

숙박 및 음식점업 경우의 총 자료(사업체)의 수는 22,866(개)이다. 이 중 대체하고자 하는 각 항목별 무응답 비율은 종사자수(1.2%), 영업개월수(58.7%), 휴무일수(58.6%), 일일영업시간(58.6%), 연면적(58.9%), 매출액(20.5%), 영업비용(59.6%) 정도이다. 시범예행조사의 목적 중의 하나가 조사 전 과정에서 발생하는 여러 가지 문제점을 파악하고, 이를 보완하여 본 조사를 성공적으로 수행하기 위함이므로 실제 본 조사의 무응답 비율은 높지 않을 것으로 생각된다. 통계청에서는 각 항목마다 차이는 있겠지만 10%는 넘지 않을 것으로 예상하고 있다. 따라서 현재의 자료 상태를 고려하여 본 연구를 위해서는 항목 무응답 부분을 제거하고 완전한 자료를 만든 후에 이를 바탕으로 연구를 진행할 것이다.

제2절 CHAID 알고리즘과 연관성분석 결과

경제총조사 자료의 항목 무응답 대체를 위한 대체군은 CHAID 알고리즘을 이용하여 연관성분석을 실시하여 결정한다. 이 알고리즘은 항목 간의 의미 있는 관계를 탐색하는데 효과적이라고 알려져 있다. 이 절에서는 CHAID 알고리즘을 소개하고 경제총조사 자료의 종사자수, 영업개월수, 휴무일수, 일일영업시간, 연면적, 매출액, 영업비용 항목에 대한 연관성분석 결과를 제시한다.

1. CHAID 알고리즘

의사결정나무의 분리 알고리즘 중의 하나인 CHAID는 목표변수가 범주형 자료인 경우에는 χ^2통계량에 의한 분할, 연속형 자료인 경우에는 F검정을 이용한 분할을 수행하는 분석방법이다. 구체적인 알고리즘을 살펴보면 다음과 같다.

step 1 : 각 설명변수에 대하여, 목표변수와 가장 유사성(p값으로 측정)이 큰 범주의 짝을 찾는다. p값을 계산하는 방법은 목표변수의 자료특성에 의해 결정된다.

이 때 목표변수가 범주형인 경우는 2×d 분할표를 통한 χ^2검정을 사용한다. 여기서 d는 목표변수의 범주 수이다.

(예시) 2×d 분할표에서의 p값 계산

	범주 1	범주 2	…	범주 d	합계
범주 1	f_{11}	f_{12}	…	f_{1d}	$f_{1.}$
범주 2	f_{21}	f_{22}	…	f_{2d}	$f_{2.}$
합계	$f_{.1}$	$f_{.2}$	…	$f_{.d}$	$f_{..}$

분할표에서 유사성 검정을 위한 카이제곱 통계량은

$$\chi^2 = \sum_{i,j} \frac{(f_{ij} - e_{ij})^2}{e_{ij}},$$

분포의 동일성 또는 독립성의 가설 하에서 계산된 기대도수는

$$e_{ij} = \frac{f_{i.} \times f_{.j}}{f_{..}}$$

와 같이 계산된다. 이 때 카이제곱의 값이 클수록 각 범주에 의하여 목표변수를 분리할 가능성이 커진다. 성별에 의한 선호도를 나타내고 있는 간단한 예를 살펴보면 다음과 같다.

case 1

	찬성	반대	계
남	40 (20)	10 (30)	50
여	0 (20)	50 (30)	50
계	40	60	100

case 2

	찬성	반대	계
남	30 (25)	20 (25)	50
여	20 (25)	30 (25)	50
계	50	50	100

( ) 안의 값은 각 셀에서의 기대도수를 나타냄

· case 1의 카이제곱 통계량 :

$$\chi^2 = \frac{(40-20)^2}{20} + \frac{(10-30)^2}{30} + \frac{(0-20)^2}{20} + \frac{(50-30)^2}{30}$$
$$= 66.67$$

· case 2의 카이제곱 통계량 :

$$\chi^2 = \frac{(30-25)^2}{25} + \frac{(20-25)^2}{25} + \frac{(20-25)^2}{25} + \frac{(30-25)^2}{25}$$
$$= 4$$

따라서 case 1의 경우의 카이제곱 통계량이 크기 때문에 case 2의 경우보다 성별이 분리될 가능성이 커짐을 알 수 있다. 목표변수가 연속형인 경우에는 2개 이상의 그룹의 평균차이를 검정하는 분산분석표의 F검정을 사용하여 분리한다.

step 2 : 가장 큰 p값을 가지는 설명변수 범주의 짝에 대하여 그 p값과 미리 정해놓은 α값을 비교한다.

· p값이 α값보다 클 경우에는 짝을 이루는 설명변수의 범주를 통합하고, 새로 생성된 범주에 대하여 step 1을 다시 실행한다.
· p값이 α값보다 작을 경우에는 step 3으로 간다.

step 3 : 조정된 각 설명변수의 범주에 대하여 새로운 p값을 계산하고, 가장 작은 p값을 가지는 설명변수를 선택해 그 p값과 미리 정해놓은 α값을 비교한다.

· p값이 α값보다 작거나 같을 경우에는 설명변수의 범주에 근거한 노드를 분리한다.
· p값이 α값보다 클 경우에는 노드를 분리하지 않으며, 이 노드는 최종노드가 된다.

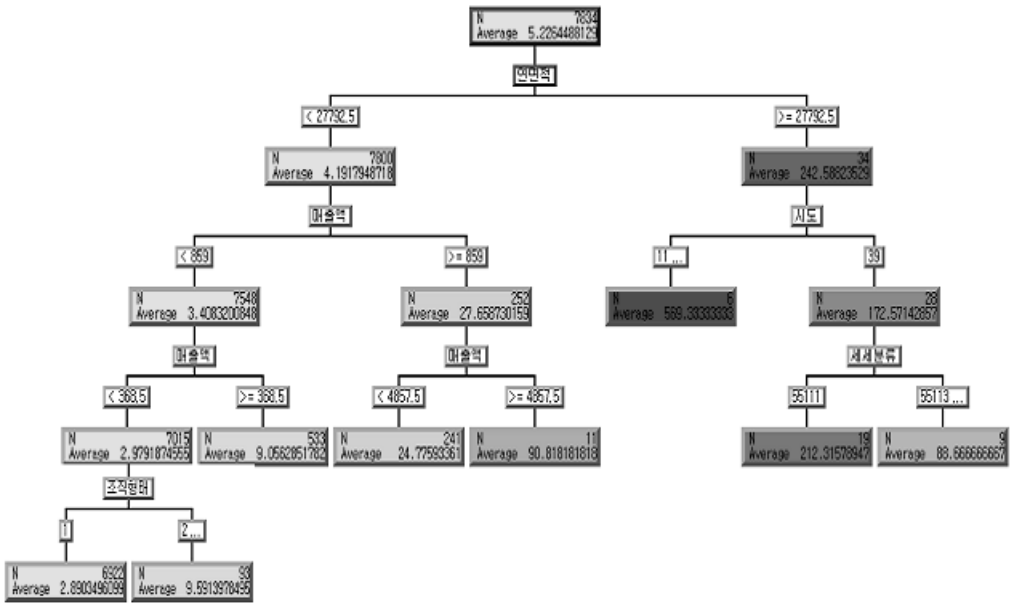
step 4 : 더 이상 분리할 노드가 없거나 정해진 정지규칙이 만족할 때까지, 위의 과정을 독립적으로 반복한다.

2. 각 항목에 대한 연관성분석 결과

경제총조사(숙박 및 음식점업) 자료에서 연관성분석을 실시할 항목은 종사자수, 영업개월수, 휴무일수, 일일영업시간, 연면적, 매출액, 영업비용 7개 항목이다. 이 항목을 이용하여 각 항목별 대체에 이용할 대체군(보조변수)을 결정하고자 한다.

가. 종사자수

종사자수에 대한 연관성 분석의 결과를 [그림 3-1]과 <표 3-1>을 이용하여 설명하고자 한다. 나머지 항목도 설명방식이 유사하기 때문에 종사자수 경우에 대해서만 자세하게 설명하고, 나머지 항목은 약식으로 설명할 것이다. 그리고 <표 3-1>에서 분리변수의 괄호안의 숫자는 각 가지에서의 마디 번호를 의미한다.



[그림 3-1] 종사자수에 관한 연관성 모형

〈표 3-1〉 종사자수에 관한 연관성 분석

종사자수			
깊이(연관성)	분리변수	분리지점(좌)	분리지점(우)
1	연면적	27792.5 미만	27792.5 이상
2	매출액(1) 시도(읍면동)(2)	859 미만 11, 21	859 이상 39
3	매출액(1) 매출액(2) 세세분류(4)	368.5 미만 4857.5 미만 55111	368.5 이상 4857.5 이상 그 외
4	조직형태(1)	1	2, 3, 4

1(연면적)

종사자수와 가장 연관성이 높은 항목을 의미한다. 전체 종사자수의 평균은 5.23(명)이며 연면적에 의하여 이를 분리하여 나간다. 연면적이 27792.5(m^2) 미만인 경우에는 좌로 분리되고, 이때의 종사자수의 평균은 4.19(명)로 전체 평균에 비하여 조금 낮아지는 것을 알 수 있다. 반면에 연면적이 27792.5(m^2) 이상인 경우에는 우로 분리되며 이때의 종사자수의 평균은 242.59(명)로 아주 큰 폭으로 증가하였음을 알 수 있다.

2_1(매출액)

첫 번째 분리가 끝난 후 두 번째 분리가 시작된다. 이 과정은 첫 번째 분리가 된 왼쪽 부분을 다시 세부적으로 분리한다. 매출액 항목으로 다시 분리가 된다. 연면적이 27792.5(m^2) 미만인 사업체 중에서 매출액이 859(백만원) 미만이면 좌로 분리되고 이때의 종사자수의 평균은 3.41(명)이 되며, 매출액이 859(백만원) 이상이면 우로 분리되고 이때의 종사자수의 평균은 27.66(명)이 된다.

2_2(시도)

이 과정은 첫 번째 분리가 된 오른쪽 부분을 다시 세부적으로 분리한다. 다음으로 연관성이 높은 시도 항목으로 다시 분리가 된다. 연면적이 27792.5(m^2) 이상인 사업체 중에서 시도가 11(서울), 21(부산)이면 좌로 분리되고 이때의 종사자수의 평균은 569.33(명)이 되며, 시도가 39(제주)이면 우로 분리되고 이때의 종사자수의 평균은 172.57(명)이 된다. 따라서 대도시에 종사자수가 많은 사업체가 있을 가능성이 높음을 분석결과 알 수 있다.

3_1(매출액)

두 번째 분리가 끝난 후 세 번째 분리가 시작된다. 이 과정은 두 번째 분리가 된 첫 번째 부분을 다시 세부적으로 분리한다. 매출액 항목에 의해 계속적으로 분리가 된다. 연면적이 27792.5(m^2) 미만인 사업체 중에서 매출액이 368.5(백만원) 미만이면 좌로 분리되고 이때의 종사자수의 평균은 2.98(명)이 되며, 매출액이 368.5(백만원) 이상 859(백만원) 미만이면 우로 분리되고 이때의 종사자수의 평균은 9.06(명)이 된다.

3_2(매출액)

이 과정은 두 번째 분리가 된 두 번째 부분을 다시 세부적으로 분리한다. 매출액 항목에 의해 계속적으로 분리가 된다. 연면적이 27792.5(m^2) 미만인 사업체 중에서 매출액이 859(백만원) 이상 4857.5(백만원) 미만이면 좌로 분리되고 이때의 종사자수의 평균은

24.78(명)이 되며, 매출액이 4857.5(백만원) 이상이면 우로 분리되고 이때의 종사자수의 평균은 90.82(명)가 된다. 이러한 연관성분석의 결과 종사자수는 연면적과 매출액의 크기에 상당히 비례하고 있음을 알 수 있다.

3_4(세세분류)

이 과정은 두 번째 분리가 된 네 번째 부분을 다시 세부적으로 분리한다. 세세분류에 의해 다시 분리가 된다. 연면적이 27792.5(m^2) 이상이고 제주 지역에 있는 사업체 중에서 55111(호텔업)이면 좌로 분리되고 이때의 종사자수의 평균은 212.32(명)가 된다.

4_1(조직형태)

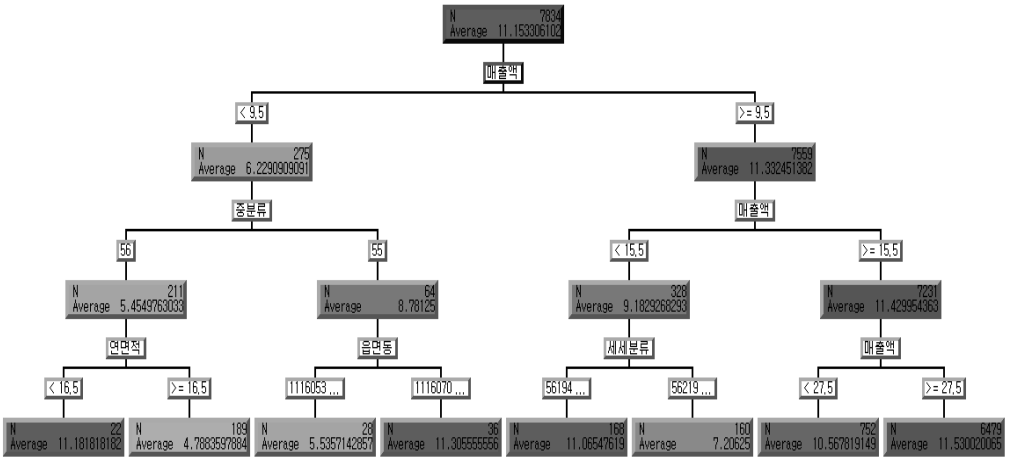
세 번째 분리가 끝난 후 네 번째 분리가 시작된다. 이 과정은 세 번째 분리가 된 첫 번째 부분을 다시 세부적으로 분리한다. 조직형태 항목으로 다시 분리가 된다. 연면적이 27792.5(m^2) 미만이고 매출액이 368.5(백만원) 미만인 사업체 중에서 조직형태가 1(개인사업체)이면 좌로 분리되고 이때의 종사자수의 평균은 2.89(명)가 되며, 조직형태가 1 이외의 사업체인 경우에는 우로 분리되고 이때의 종사자수의 평균은 9.59(명)가 된다.

이상으로 종사자수 항목을 분리하는 과정을 자세히 설명하였다. 요약하면, 연면적, 매출액, 시도(읍면동), 세세분류, 조직형태 항목이 종사자수와 높은 연관성이 있다는 것을 알 수 있다. 하지만, 실제 적용할 자료에 따라서 이 항목들을 대체군으로 사용할 수 있는지의 여부를 검토하여야 한다. 본 연구에서 사용될 시범예행조사의 경우 종사자수 항목이 무응답인 경우의 대부분이 연면적과 매출액 항목도 무응답으로 나타났다. 그러므로 종사자수를 대체할 경우 연면적과 매출액의 정보는 사용할 수가 없다는 것을 알 수 있다. 이런 경우 연면적과 매출액은 대체군으로부터 제외되어야 할 것이다. 물론 실제 경제총조사 자료를 가지고 무응답 관계를 고려했을 때 두 항목을 대체군으로 사용할 수도 있을 것이다. 따라서 연관성분석을 통해서 제시된 항목과 실제 대체군으로 사용될 항목은 다소 차이가 발생할 수 있음을 알리며, 최종 대체군의 결정은 경제총조사 자료를 검토한 후 결정하면 될 것으로 판단된다.

나. 영업개월수

다음으로 영업개월수 항목도 종사자수 항목과 같은 방법으로 대체군을 결정하고자 한다. 영업개월수 항목은 매출액과 가장 큰 연관성을 가진다. 영업개월수의 전체평균은 11.15(개월)이며, 매출액, 산업분류, 연면적과 큰 연관성을 가진다. 매출액이 9.5(백만원) 미만은 6.23(개월), 매출액이 9.5(백만원) 이상은 11.33(개월)의 영업개월수를 평균적으로

나타내고 있다. 그리고 매출액이 9.5(백만원) 미만인 경우에는 중분류에 의해서 나누어지는데 55(숙박업)는 8.78(개월), 56(음식점 및 주점업)은 5.45(개월)의 영업개월수를 가진다. 56(음식점 및 주점업)의 경우 연면적에 의해서 상당한 차이를 보이는데, 매출액이 9.5(백만원) 미만인 경우 연면적이 16.5(m^2)에 의해서 영업개월수가 4.79(개월)와 11.18(개월)로 나누어짐을 알 수 있다. 또한 읍면동 및 세세분류도 영업개월수와 연관성을 보인다. 자세한 결과는 [그림 3-2]와 <표 3-2>를 참조하기 바란다.



[그림 3-2] 영업개월수에 관한 연관성 모형

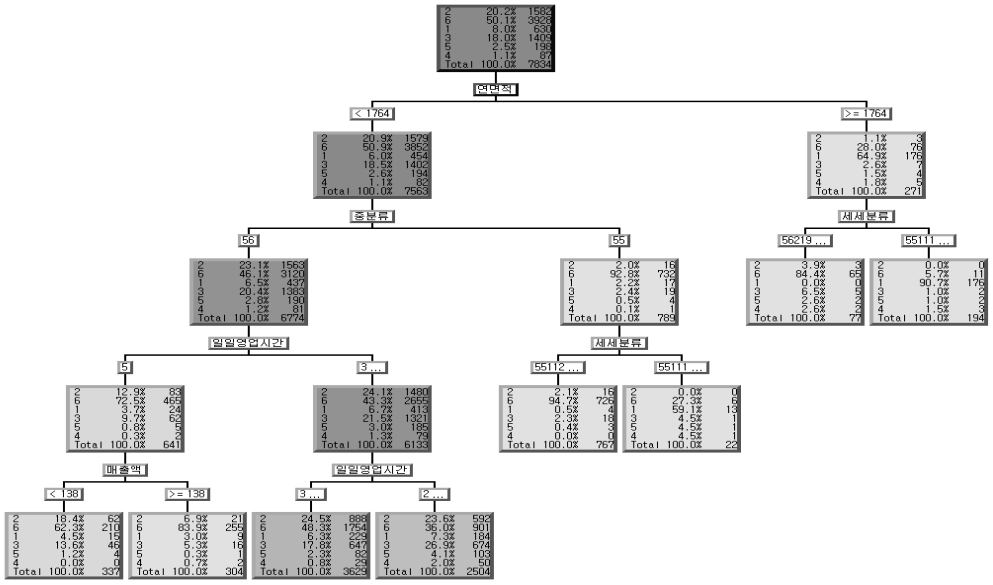
<표 3-2> 영업개월수에 관한 연관성 분석

영업개월수			
깊이(연관성)	분리변수	분리지점(좌)	분리지점(우)
1	매출액	9.5 미만	9.5 이상
2	중분류(1) 매출액(2)	56 15.5 미만	55 15.5 이상
3	연면적(1) 읍면동(2) 세세분류(3) 매출액(4)	16.5 미만 27.5 미만	16.5 이상 27.5 미만
4	종사자수(5)	2.5 미만	2.5 미만

다. 휴무일수

휴무일수는 6개의 항목으로 구성되는데 휴무없음이 50.1%, 월2-3일이 20.2%, 월4-5일

이 18.0%로 세 항목이 대략 90% 정도를 차지하고 있다. 휴무일수와 가장 연관성이 높은 항목은 연면적으로 연면적이 1764(m^2) 이상인 경우에는 휴무일수가 월1일인 경우가 64.9%로 상당히 높은 비율을 차지하게 된다. 이는 산업분류(중분류, 세세분류)에 의하여 더 세부적으로 나누어진다. 연면적이 1764(m^2) 미만이면서 중분류가 56(음식점 및 주점업)인 경우, 일일영업시간이 14시간 이상이면 휴무일수가 없는 경우가 전체의 72.5%를 차지하고 있다. 여기에서 매출액이 138(백만원) 이상이 되면 휴무없음이 83.9%로 더 높아짐을 알 수 있다. 자세한 결과는 [그림 3-3]과 <표 3-3>을 참조하기 바란다.



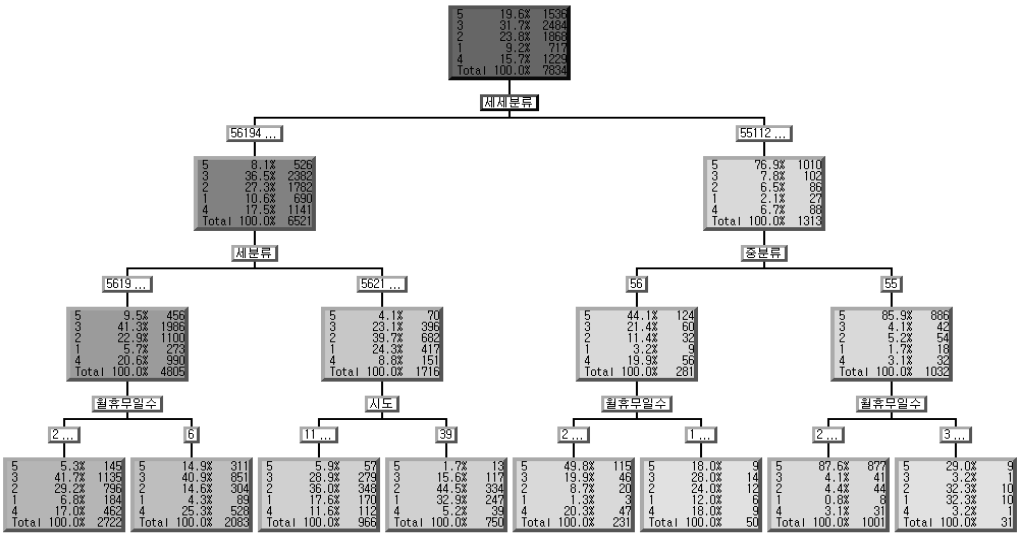
[그림 3-3] 휴무일수에 관한 연관성 모형

〈표 3-3〉 휴무일수에 관한 연관성 분석

휴무일수			
깊이(연관성)	분리변수	분리지점(좌)	분리지점(우)
1	연면적	1764 미만	1764 이상
2	중분류(1) 세세분류(2)	56	55
3	일일영업시간(1) 세세분류(2)	5	1, 2, 3, 4
4	매출액(1) 일일영업시간(2)	138 미만 3, 4	138 미만 1, 2

라. 일일영업시간

일일영업시간 항목을 살펴보면 8시간미만 9.2%, 8-10시간 23.8%, 10-12시간 31.7%, 12-14시간 15.7%, 14시간 이상 19.6%의 비율로 구성되어져 있다. 가장 연관성이 높은 항목은 산업분류(중분류, 세분류, 세세분류)인데 중분류가 55(숙박업)인 경우에는 14시간이상 영업하는 사업체가 85.9%로 상당히 높아짐을 알 수 있다. 그 외의 분류코드별 특성과도 연관성이 높다는 것을 분석결과를 통해 알 수 있다. 다음으로 연관성이 높은 항목은 휴무일수이다. 휴무일수가 없음 또는 적은 경우 일일영업시간은 더 많아지는 경향을 보인다. 자세한 결과는 [그림 3-4]와 <표 3-4>를 참조하기 바란다.



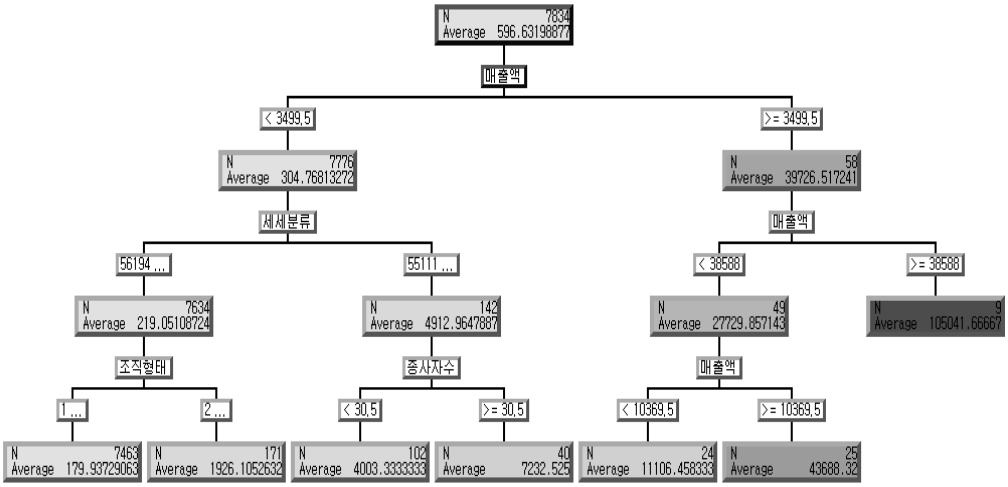
[그림 3-4] 일일영업시간에 관한 연관성 모형

〈표 3-4〉 일일영업시간에 관한 연관성 분석

일일영업시간			
깊이(연관성)	분리변수	분리지점(좌)	분리지점(우)
1	세세분류		
2	세분류(1) 중분류(2)	56	55
3	휴무일수(1) 시도(읍면동)(2) 휴무일수(3) 휴무일수(4)	1, 2, 3, 4, 5 2, 6 1, 2, 6	6 1, 3, 4, 5 3, 4, 5

마. 연면적

연면적의 전체평균은 596.63(m^2)이며, 연면적을 분리하는데 매출액이 가장 큰 역할을 한다. 매출액이 3499.5(백만원) 미만은 304.77(m^2), 매출액이 3499.5(백만원) 이상은 39726(m^2)의 연면적을 평균적으로 나타내고 있다. 또한 매출액이 38588(백만원) 이상인 경우의 연면적은 105041.67(m^2)로 상당히 큰 것을 알 수 있다. 그리고 매출액이 3499.5(백만원) 미만인 경우에는 세세분류에 의해서 나누어지는데 56194(분식 및 김밥 전문점)는 219.05(m^2)로 작고, 55111(호텔업)은 4912.96(m^2)로 큰 연면적을 가짐을 볼 수 있다. 또한 조직형태가 개인사업체인 경우는 연면적이 작은 것으로 나타났고, 종사자수도 연면적과 연관성을 보인다. 자세한 결과는 [그림 3-5]와 <표 3-5>를 참조하기 바란다.



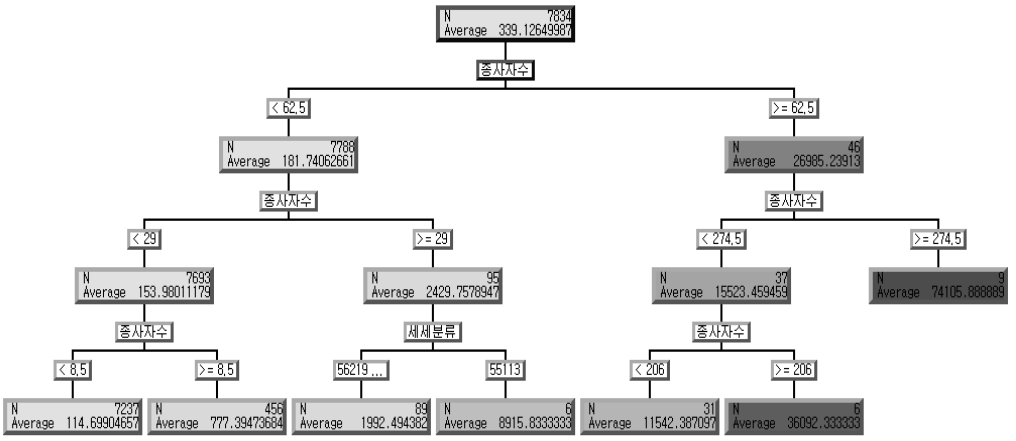
[그림 3-5] 연면적에 관한 연관성 모형

〈표 3-5〉 연면적에 관한 연관성 분석

연면적			
깊이(연관성)	분리변수	분리지점(좌)	분리지점(우)
1	매출액	3499.5 미만	3499.5 이상
2	세세분류(1) 매출액(2)	38588 미만	38588 이상
3	조직형태(1) 종사자수(2) 매출액(3)\	1, 3 30.5 미만 10369.5 미만	2, 4, 5 30.5 이상 10369.5 이상
4	읍면동(6)		

바. 매출액

매출액은 종사자수와 가장 큰 연관성을 가진다. 매출액의 전체평균은 339.13(백만원)이며, 종사자수가 62.5(명) 미만은 181.74(백만원), 종사자수가 62.5(명) 이상은 26985.24(백만원)의 매출액을 평균적으로 나타내고 있다. 또한 종사자수가 8.5(명) 이하인 경우에는 매출액이 114.7(백만원)로 줄어들고, 종사자수가 274.5(명) 이상인 경우에는 74105.89(백만원)로 상당히 늘어나는 것을 볼 수 있다. 그리고 세세분류에 의해서도 매출액의 특성이 잘 나타나고 있다. 시도(읍면동) 및 연면적도 매출액과 연관성이 있는 것으로 나타났다. 자세한 결과는 [그림 3-6]과 <표 3-6>을 참조하기 바란다.



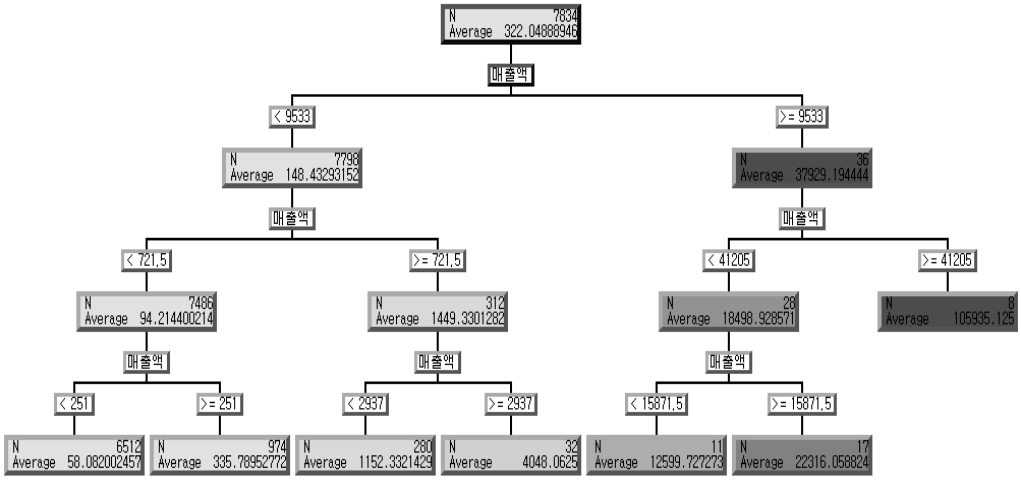
[그림 3-6] 매출액에 관한 연관성 모형

〈표 3-6〉 매출액에 관한 연관성 분석

매출액			
깊이(연관성)	분리변수	분리지점(좌)	분리지점(우)
1	종사자수	62.5 미만	62.5 이상
2	종사자수(1) 종사자수(2)	29 미만 274.5 미만	29 이상 274.5 이상
3	종사자수(1) 세세분류(2) 종사자수(3)	8.5 미만 206 미만	8.5 이상 206 이상
4	시도(읍면동)(1) 연면적(3) 연면적(5)	 788 미만 15856 미만	 788 이상 15856 이상

사. 영업비용

영업비용 항목은 매출액과 매우 높은 연관성을 가진다. 즉, 매출액을 알 수 있으면 영업이익을 정확도 높게 추정할 수 있다는 것이다. 분석결과를 살펴보면 매출액이 251(백만원) 미만인 사업체의 영업이익 평균은 58.08(백만원)이며, 251(백만원) 이상 721.5(백만원) 미만은 335.79(백만원), 721.5(백만원) 이상 2937(백만원) 미만은 1152.33(백만원), 2937(백만원) 이상 9533(백만원) 미만은 4048.06(백만원), 9533(백만원) 이상 15871.5(백만원) 미만은 12599.73(백만원), 15871.5(백만원) 이상 41205(백만원) 미만은 22316.07(백만원)의 평균 영업비용을 가짐을 볼 수 있다. 자세한 결과는 [그림 3-7]과 <표 3-7>을 참조하기 바란다.



[그림 3-7] 영업비용에 관한 연관성 모형

〈표 3-7〉 영업비용에 관한 연관성 분석

영업비용			
깊이(연관성)	분리변수	분리지점(좌)	분리지점(우)
1	매출액	9533 미만	9533 이상
2	매출액(1) 매출액(2)	721.5 미만 41205 미만	721.5 이상 41205 이상
3	매출액(1) 매출액(2) 매출액(3)	251 미만 2937 미만 15871.5 미만	251 이상 2937 이상 15871.5 이상

3. 대체군 선정 결과

연관성분석을 통하여 각 항목에 사용될 대체군(보조변수)이 선정되었다. 이 절에서는 완전자료에 의해 선정된 대체군과 시범예행조사 자료에 적용할 대체군을 각기 제시하고자 한다. 이러한 이유는 시범예행조사 자료의 무응답 비율이 매우 높아 각 항목마다 동시에 누락이 된 자료가 대부분이며, 자료의 총수도 실제 조사의 2~3% 정도밖에 되지 않기 때문이다. 따라서 적용을 위해서는 완전자료에 의해 제시된 대체군을 그대로 사용할 수 없다.

하지만 실제 경제총조사 자료는 시범예행조사보다 무응답 비율이 매우 낮을 것으로 보이며, 제시된 대체군의 사용여부는 조사 자료를 검토한 후에 결정해야 할 것으로 판단된다. <표 3-8>은 연관성분석 결과의 대체군과 시범예행조사 자료의 적용을 위한 대체군이 주어져 있다.

〈표 3-8〉 경제총조사 항목의 대체군

항목(무응답률)	연관성분석 결과 대체군	시범예행조사 자료에 적용할 대체군
종사자수(1.2%)	연면적, 매출액, 읍면동, 세세분류, 조직형태	시군구, 소분류, 매출액
매출액(20.5%)	종사자수, 세세분류, 읍면동, 연면적	시군구, 소분류, 종사자수
영업비용(59.6%)	매출액, (읍면동, 세세분류)	매출액, 시군구, 소분류
연면적(58.9%)	매출액, 세세분류, 조직형태, 종사자수, 읍면동	시군구, 소분류, 매출액, 종사자수
영업개월수(58.7%)	매출액, 세세분류, 연면적, 읍면동, 종사자수	시군구, 소분류, 매출액, 연면적, 종사자수
휴무일수(58.6%)	연면적, 세세분류, 일일영업시간, 매출액	연면적, 소분류, 매출액
일일영업시간(58.6%)	세세분류, 휴무일수, 읍면동	소분류, 휴무일수, 시군구

연관성분석 결과와 시범예행조사 자료에 적용할 대체군의 주된 차이는 자료의 수가 적어 읍면동이나 세세분류까지 사용하지 못한다는 것이다. 하지만, 이러한 문제는 실제 경제총조사 자료에서는 대부분 사용가능할 것으로 보인다. 또한 종사자수와 매출액은 대부분 항목의 대체군으로 사용될 수 있어 두 항목의 대체가 실행되면 다른 항목의 대체군으로 사용할 수 있을 것이다. 그리고 시범예행조사의 무응답 비율을 고려할 때 대체순서는 <표 3-8>의 항목 순서대로 하면 될 것이다.

제3절 무응답 대체방법

무응답 대체방법은 무응답 항목의 대체값으로 한 개의 값을 부여하는 단일 대체방법(single imputation)과 여러 개의 값을 대체하는 다중 대체방법(multiple imputation, Rubin(1987))으로 구분되며, 단일 대체방법은 무응답 항목에 유일하게 결정된 대체값을 대입하는 결정적 대체방법(deterministic imputation)과 대체값을 확률적으로 결정하여 대입하는 확률적 대체방법(stochastic imputation)으로 구분된다. 이러한 방법들에 대한 자세한 설명은 『무응답 처리를 위한 방법론 연구(Ⅰ)(통계개발원, 2009)』를 참조하기 바라며, 이 절에서는 본 연구에서 사용될 방법에 대해서 예를 들어 간략하게 소개하고자 한다. <표 3-9>에는 설명을 위한 자료를 임의로 구성하였다.

〈표 3-9〉 예제 자료

사업체	대체항목(연속형)	대체군1(범주형)	대체군2(연속형)
1	25	A	94
2	29	B	206
3	24	B	66
4	39	A	295
5	37	B	232
6	28	A	172
7	40	B	370
8	27	A	81
9	40	B	292
10	34	B	206
11	30	A	93
12	39	B	310
13	38	B	236
14	28	A	111
15	**missing**	B	301
16	38	B	303
17	40	B	315
18	26	A	166
19	24	A	77
20	34	B	222

본 예제는 대체하고자 하는 연속형 항목에 대하여 대체군은 범주형과 연속형 항목을 가지고 있는 경우이다. 예제 자료에서는 15번째의 사업체에서 무응답이 발생하여 이를 각각의 방법으로 대체하는 과정을 보여주고자 한다.

1. 평균 대체방법(mean imputation)

무응답 항목에 목표변수의 전체 평균을 대입하거나 또는 대체군 내의 평균을 대입하는 방법이다. 이 방법은 간단하여 이용하기 쉬운 장점이 있으며, 항목이 양적 변수이고 구하고자 하는 통계량이 평균일 때 유용하다. 그러나 대체후의 값들은 평균값의 빈도수가 지나치게 많아져 응답값들의 분포가 왜곡되고, 중위수나 백분위수와 같은 평균이 아닌 통계량을 구할 때에는 효율이 저하되는 단점이 있다.

예제의 경우 15번째 사업체는 대체군1은 B, 대체군2는 301의 값을 가진다. 따라서 이 두 값을 가지고 있는 사업체를 찾아 대체하고자 하는 값의 평균을 구하면 될 것이다. 여기서 한 가지 고려할 사항은 대체군2의 경우 연속형 항목이므로 301이 여러 개 존재하지 않는다면 301과 유사(±5%)한 값들을 하나의 집단으로 묶어 생각하면 될 것이다. 이것은 이후에 설명할 응용 핫덱 방법에서 사용한 것으로 효율성은 입증되었다.

이러한 절차를 통하여 최종적으로 선택되는 도너는 9, 12, 16, 17번째 사업체이며, 이들의 평균값 (40+39+38+40)/4=39.25(39)가 최종 대체값이 됨을 알 수 있다.

2. 회귀 대체방법(regression imputation)

회귀 대체는 대체하고자 하는 항목에 대하여 가능한 보조변수를 이용하여 회귀모형에 적합시키는 방법으로, 목표변수와 보조변수의 관계가 절편이 있는 직선 관계이고 목표변수의 분산이 동일할 때 유용한 방법이다. Kalton(1982)은 i번째 결측값의 대체값을 다음과 같이 구하였다.

$$\hat{y}= \widehat{\beta_0}+ \sum_{k=1}^{K} \widehat{\beta_k} x_{ki} + \epsilon_i$$

여기서 $\hat{\beta}$는 일반적 최소제곱법에 의해 추정된 계수이다. 이 방법은 미국의 인구조사(CPS: current population survey)에서 발생하는 결측값을 대체하기 위해서 이용하였으며, 그 결과 대체된 값과 실제값의 평균절대편차를 비교할 때 다른 대체 방법에 비해 매우 적절함을 보였다.

예제의 경우 대체군1이 B인 11개의 자료를 이용하여 회귀모형에 적합시킨다. 실제 자료에서는 11개보다 훨씬 많은 자료를 이용하므로 자료의 수에 대한 문제는 고려하지 않는다. 이러한 과정에 의해 추정된 모형은 $\hat{y}= 21.214 + 0.058x$가 된다. 따라서 x의 값에 301을 넣어 얻은 최종 대체값은 38.672(39)가 됨을 알 수 있다.

3. 비 대체방법(ratio imputation)

비 대체방법은 이용 가능한 보조변수가 있을 때 목표변수와 보조변수의 관계를 이용하여 무응답을 대체하는 방법으로 목표변수와 보조변수가 원점을 지나는 직선관계이며, 분산이 보조변수에 비례하는 경우 효과적인 것으로 나타났다. 유한 모집단에서 예측이론에 의하면 비 대체방법이 매우 우수한 것으로 알려져 있으나 항목들이 양적변수일 때에만 사용가능하다. 만일 목표변수와 보조변수가 비례관계이면 다음과 같다.

$$\hat{y} = \left(\frac{\bar{y}_R}{\bar{x}_R} \right) x_k$$

여기서 R은 응답이 있는 자료이다. 회귀 대체방법과 마찬가지로 대체군1이 B인 11개의 자료를 이용하여 대체값을 추정한다. 이 때 $\bar{y}_R$은 35.73이며, $\bar{x}_R$은 250.73이다. 따라서 15번째 사업체의 x값이 301이므로 추정값은 $y = (35.73/250.73)*301 = 42.89(43)$이 됨을 알 수 있다.

4. 응용 핫덱 대체방법(applied hot-deck imputation)

랜덤 핫덱 방법을 응용한 것으로 연속형 항목에도 자유롭게 적용할 수 있게 한 방법이다. 대체군의 적용 시 범주형 항목은 항목값의 일치여부를 판단하여 점수를 부여하고, 연속형 항목은 범주화하지 않고 신뢰구간의 개념을 이용하여 그 구간 안에 들어가면 비슷한 개체로 판단하여 점수를 부여한다. 모든 대체군의 항목들에 대하여 비교한 후 가장 점수가 높은 개체들을 선택하여 그 중에서 하나의 개체를 임의로 대체하는 방법이다.

예제의 경우 대체군1이 B인 사업체에 1점을 부여하고, 대체군2의 값이 286~316사이에 있는 사업체에 1점을 부여하게 된다. 이는 평균 대체방법에서 설명한 내용으로 대체군2의 경우 연속형 항목이므로 301이 여러 개 존재하지 않는다면 301과 유사(±5%)한 값들을 하나의 집단으로 묶어 같은 대체군으로 간주하였다. 마지막으로 점수가 가장 높은 사업체를 도너로 선택하게 되는데 9, 12, 16, 17번째 사업체가 2점으로 가장 높은 점수를 획득하였다. 이 때 최고 점수가 2개 이상이면 이들 중 랜덤하게 하나의 사업체를 선택하여 대체하게 된다. 따라서 15번째 사업체는 40, 39, 38, 40 중 하나의 값을 가지게 되며 40은 2개이므로 선택될 가능성이 더 높다는 것을 알 수 있다.

5. 최빈수 대체방법(mode imputation)

최빈수 대체방법은 평균 및 응용 핫덱 대체방법과 유사하게 도너를 선택한다. 하지

만 최종 도너를 선택하는 과정에서 최빈수를 찾아 무응답 항목에 대체하는 것이다. 그러나 이 방법도 평균 대체방법처럼 무응답 비율이 높아질수록 응답값들의 분포가 왜곡될 가능성이 커지는 단점을 가지고 있다.

예제의 경우 9, 12, 16, 17번째 사업체가 도너로 선택이 되며, 이 때 대체 가능한 값은 40, 39, 38, 40이므로 최빈수 40이 실제 대체가 됨을 알 수 있다.

본 절에서 설명한 대체방법들은 다양한 조사에서 사용되고 있는 방법들이다. 이러한 무응답 대체방법들은 자료의 특성(형태)에 따라서 적절하게 적용되어진다. 경제총조사의 경우 연속형 항목은 평균, 회귀, 비, 응용 핫덱 대체방법, 범주형 항목은 최빈수와 응용 핫덱 대체방법으로 모의실험을 실시할 것이다. 그리고 모의실험을 통하여 가장 적절한 대체방법을 선택할 것이며 향후 경제총조사 자료의 무응답 대체방법으로 사용할 것이다.

제4절 대체 항목에 대한 모의실험

1. 모의실험 개요

이 절에서는 개발된 대체군을 이용하여 항목별로 모의실험을 실시하여 대체의 정확도를 검토하고자 한다. 모의실험에 사용할 자료는 2010년 기준 경제총조사 시범예행조사 자료이다. 데이터 파일은 각 항목별로 누락이 되지 않은 자료를 모아 새롭게 구성하였으며, 모의실험을 위해서 2%, 5%, 10%의 결측치를 임의로 발생시켰다. 그리고 모든 실험은 1,000번을 반복하여 추정량을 계산하였다.

각 항목의 평균추정에 대한 편향정도를 알아보기 위하여 결측 전의 평균값과 대체방법에 의해서 대체된 후의 추정 평균값과의 평균절대오차(mean absolute error)를 계산하고, 추정량의 안정성을 측정하기 위해 추정량의 표준편차를 평균값으로 나누어준 변동계수(CV)를 계산하여 대체방법들의 효율성을 비교하였다. 또한 대체전후의 분포(구성비)변화를 살펴봄으로써 얼마나 실제 분포와 유사하게 대체가 되는지를 판단하고자 한다. 다음의 식은 평균절대오차를 구하는 식으로 참조하기 바란다.

$$MAE_{\text{mean}} = \frac{1}{R}\sum_{i=1}^{R} | M - \widehat{M}_i |$$

여기서 R은 반복횟수를 나타내며 M은 목표변수의 실제 평균이고 $\widehat{M}_i$는 i번째 반복에서 추정방법에 의해 대체된 후의 목표변수 추정 평균이다.

2. 각 항목에 대한 모의실험

가. 종사자수

연관성분석의 결과 종사자수 항목은 연면적, 매출액, 읍면동, 세세분류, 조직형태 항목을 대체군으로 이용할 수 있으나, 시범예행조사 특성상 시군구, 소분류, 매출액 항목만을 대체군으로 사용해야 한다. 매출액은 종사자수와 동시에 누락된 것이 대부분이지만 모의실험을 위해서 사용할 것이다. 종사자수와 매출액의 값이 모두 포함된 18,179개의 자료를 모집단이라 가정하여 종사자수에 대한 무응답 대체 실험을 실시하였다.

〈표 3-10〉 종사자수의 모의실험 결과(mean=3.47, 단위(명))

무응답 비율	통계량	대체방법				
		평균 대체	회귀 대체 (직선, 절편o)	회귀 대체 (직선, 절편x)	비 대체	응용 핫덱 대체
2%	MAE	**0.0157(0.45%)**	**0.0071(0.20%)**	**0.0199(0.57%)**	**0.0098(0.28%)**	**0.0093(0.27%)**
	CV	0.0054	0.0041	0.0061	0.0046	0.0048
5%	MAE	**0.0371(1.07%)**	**0.0122(0.35%)**	**0.0437(1.26%)**	**0.0174(0.50%)**	**0.0171(0.49%)**
	CV	0.0081	0.0057	0.0092	0.0070	0.0057
10%	MAE	**0.0724(2.09%)**	**0.0234(0.67%)**	**0.0854(2.46%)**	**0.0279(0.80%)**	**0.0274(0.79%)**
	CV	0.0108	0.0086	0.0115	0.0098	0.0096

종사자수에 대한 종합적인 대체결과가 <표 3-10>에 정리되어 있다. 연속형 항목을 대체하기 위한 평균, 회귀, 비, 응용 핫덱 방법에 절편이 없는 회귀 대체방법을 추가하여 분석하였다. 이는 절편이 있는 회귀 대체방법으로 모형을 구축할 경우 추정식에 의하여 종사자수가 음의 값으로 대체될 가능성이 있기 때문에 절편이 없는 모형으로 다시 추정을 하였다.

무응답 비율이 2%인 경우에는 모의실험에 사용된 방법들 모두 평균을 추정하는데 있어 오차가 작다는 것을 알 수 있다. 특히, 회귀(절편 있음), 비, 응용 핫덱 대체방법은 조금 더 정확한 대체가 되고 있음을 볼 수 있다. 하지만, 회귀(절편 있음) 대체는 대체된 값이 음일 가능성이 있다는 것을 명심해야 할 것이다. 이를 보정한 회귀(절편 없음) 대체는 오차가 더 커짐을 알 수 있다. 이러한 현상은 무응답 비율이 5%, 10%로 증가할수록 더 뚜렷하게 나타난다. 무응답 비율이 10%가 되면 평균 및 회귀(절편 없음) 대체와는 달리 비 및 응용 핫덱 대체방법의 오차비율은 0.8%와 0.79%로써 1% 이하의 오차비율

을 유지해 평균 추정에 있어서는 좋은 결과를 보여준다. 그리고 추정량의 안정성을 측정하기 위한 CV값 역시 매우 작은 값을 나타내므로 종사자수의 평균값을 추정한 결과는 믿을 수 있을 것으로 생각된다.

다음으로 대체를 하고 난 후의 구성비(분포)의 변화를 살펴보고자 한다. 모의실험에서 대체를 한 후에 실제 구성비와 차이가 적어야 좋은 대체라고 말 할 수 있다. 평균 추정에서 정확한 대체가 된 경우에도 구성비 추정에서 상당히 왜곡되는 경우가 발생할 수 있다. 따라서 본 실험에서는 구성비 변화를 동시에 고려하였다. <표 3-11>에는 무응답 비율이 2%인 경우에 대체전후의 구성비 변화의 결과가 주어져있다.

〈표 3-11〉 종사자수의 대체전후 구성 변화(무응답비율 2%인 경우)

범주(명)	1	2	3	4	5 이상
실제 사업체수	4529	7109	2907	1564	2070
평균 대체방법					
대체후 사업체수	4489	7164	2913	1557	2056
절대차이(오차)	**40(0.9%)**	**55(0.8%)**	**6(0.2%)**	**7(0.4%)**	**14(0.7%)**
회귀(직선, 절편 o) 대체방법					
대체후 사업체수	4479	7136	2921	1578	2065
절대차이(오차)	**50(1.1%)**	**27(0.4%)**	**14(0.5%)**	**14(0.9%)**	**5(0.2%)**
회귀(직선, 절편 x) 대체방법					
대체후 사업체수	4621	7052	2891	1557	2058
절대차이(오차)	**92(2.0%)**	**57(0.8%)**	**16(0.6%)**	**7(0.4%)**	**12(0.6%)**
비 대체방법					
대체후 사업체수	4595	7060	2884	1555	2085
절대차이(오차)	**66(1.5%)**	**49(0.7%)**	**23(0.8%)**	**9(0.6%)**	**15(0.7%)**
응용 핫덱 대체방법					
대체후 사업체수	4522	7103	2918	1574	2062
절대차이(오차)	**7(0.2%)**	**6(0.1%)**	**11(0.4%)**	**10(0.6%)**	**8(0.4%)**

무응답 비율이 2%인 경우에는 무응답 비율이 낮기 때문에 각 방법들에서의 대체전후의 구성비 변화 차이가 매우 크지는 않으나, 응용 핫덱 대체방법에서 가장 낮은 오차를 보이고 있다. 특히 종사자수가 1인 사업체의 구성비에서 다소 오차의 비율이 차이가 남을 알 수 있다. 대부분의 방법에서 1% 이상의 오차를 보이지만, 응용 핫덱 대체방법은 0.2%로 변화가 매우 적어 대체후의 구성비 변화 문제도 거의 일어나지 않음을 알 수 있다. 평균 추정에 있어서는 비 대체도 매우 좋은 대체방법이나 대체전후의 구성비 변화는

다소 커지는 것을 볼 수 있다. 이러한 현상은 무응답 비율이 커짐에 따라 더 확실하게 나타난다. <표 3-12>와 <표 3-13>에는 무응답 비율이 5%, 10%인 경우에 대체전후의 구성비 변화의 결과가 주어져있다.

〈표 3-12〉 종사자수의 대체전후 구성 변화(무응답비율 5%인 경우)

범주(명)	1	2	3	4	5 이상
실제 사업체수	4529	7109	2907	1564	2070
평균 대체방법					
대체후 사업체수	4439	7208	2938	1566	2038
절대차이(오차)	**90(2.0%)**	**99(1.4%)**	**31(1.1%)**	**8(0.5%)**	**32(1.5%)**
회귀(직선, 절편 o) 대체방법					
대체후 사업체수	4423	7175	2946	1589	2046
절대차이(오차)	**106(2.3%)**	**66(0.9%)**	**39(1.3%)**	**25(1.6%)**	**24(1.2%)**
회귀(직선, 절편 x) 대체방법					
대체후 사업체수	4752	6992	2876	1528	2031
절대차이(오차)	**223(4.9%)**	**117(1.6%)**	**31(1.1%)**	**36(2.3%)**	**39(1.9%)**
비 대체방법					
대체후 사업체수	4661	7003	2866	1548	2101
절대차이(오차)	**132(2.9%)**	**106(1.5%)**	**41(1.4%)**	**16(1.0%)**	**31(1.5%)**
응용 핫덱 대체방법					
대체후 사업체수	4551	7096	2890	1576	2066
절대차이(오차)	**22(0.5%)**	**13(0.2%)**	**17(0.6%)**	**12(0.8%)**	**4(0.2%)**

〈표 3-13〉 종사자수의 대체전후 구성 변화(무응답비율 10%인 경우)

범주(명)	1	2	3	4	5 이상
실제 사업체수	4529	7109	2907	1564	2070
평균 대체방법					
대체후 사업체수	4365	7304	2997	1515	1998
절대차이(오차)	**164(3.6%)**	**195(2.7%)**	**90(3.1%)**	**49(3.1%)**	**72(3.5%)**
회귀(직선, 절편 o) 대체방법					
대체후 사업체수	4318	7288	3011	1555	2007
절대차이(오차)	**211(4.7%)**	**179(2.5%)**	**104(3.6%)**	**9(0.6%)**	**63(3.0%)**
회귀(직선, 절편 x) 대체방법					
대체후 사업체수	4930	6954	2807	1492	1996
절대차이(오차)	**401(8.9%)**	**155(2.2%)**	**100(3.4%)**	**72(4.6%)**	**74(3.6%)**
비 대체방법					
대체후 사업체수	4774	6938	2782	1594	2091
절대차이(오차)	**245(5.4%)**	**171(2.4%)**	**125(4.3%)**	**30(1.9%)**	**21(1.0%)**
응용 핫덱 대체방법					
대체후 사업체수	4534	7135	2928	1544	2038
절대차이(오차)	**5(0.1%)**	**26(0.4%)**	**21(0.7%)**	**20(1.3%)**	**32(1.5%)**

무응답 비율이 증가할수록 대체전후의 구성비 변화가 크게 나타난다. 특히, 종사자수가 1인인 경우의 오차가 가장 큰 것으로 보인다. 평균 추정에서 정확한 대체를 보였던 비 대체방법의 경우 무응답 비율이 10%인 경우에는 5.4%까지 구성비 변화가 일어나는 것을 볼 수 있다. 그 외의 대체방법들도 2~6%에 해당되는 구성비 변화가 나타난다. 반면에 응용 핫덱 대체방법은 다른 대체방법에 비해서 상당히 안정적인 대체결과를 보여준다. 무응답 비율이 증가함에 따라 오차가 늘어나지만 그 폭은 작으며 1%정도의 구성비 변화가 유지되고 있음을 볼 수 있다. 따라서 종사자수 대체를 위해서는 응용 핫덱 대체방법이 매우 적절할 것으로 판단된다.

나. 매출액

연관성분석의 결과 매출액 항목은 종사자수, 세세분류, 읍면동, 연면적 항목을 대체군으로 이용할 수 있으나, 시범예행조사 특성상 시군구, 소분류, 종사자수 항목만을 대체군으로 사용해야 한다. 실제 매출액을 대체할 때에는 매출액을 대체하기 이전에 종사자수가 먼저 대체되므로 이 자료를 이용할 수 있을 것이다. 본 모의실험에서는 종사자수와 매출액의 값이 모두 포함된 18,179개의 자료를 모집단이라 가정하여 실험을 실시하였다.

<표 3-14> 매출액의 모의실험 결과(mean=166.92, 단위(백만원))

무응답 비율	통계량	대체방법				
		평균 대체	회귀 대체 (직선, 절편o)	회귀 대체 (직선, 절편x)	비 대체	응용 핫덱 대체
2%	MAE	**1.29(0.77%)**	**0.82(0.49%)**	**0.86(0.52%)**	**0.78(0.47%)**	**0.79(0.47%)**
	CV	0.0104	0.0083	0.0085	0.0076	0.0079
5%	MAE	**2.96(1.77%)**	**1.67(1.00%)**	**1.83(1.10%)**	**1.58(0.95%)**	**1.65(0.99%)**
	CV	0.0237	0.0145	0.0155	0.0119	0.0122
10%	MAE	**4.61(2.76%)**	**2.56(1.53%)**	**2.99(1.79%)**	**2.50(1.50%)**	**2.66(1.59%)**
	CV	0.0353	0.0204	0.0235	0.0201	0.0211

매출액에 대한 종합적인 대체결과가 <표 3-14>에 정리되어 있다. 평균 추정에 있어 평균 대체방법을 제외한 나머지 방법들은 큰 차이를 보이지 않는다. 회귀 대체방법도 적절한 추정을 하고 있지만, 종사자수 항목처럼 비 대체 및 응용 핫덱 대체가 조금 더 정확도가 높은 것으로 보인다. 무응답 비율이 10%로 증가한 경우에는 오차 비율이 1.5% 정도로 높아짐을 볼 수 있으나, 무응답 비율을 고려할 때 오차 비율이 높지는 않은 것으

로 판단된다. 또한 CV값 역시 매우 작은 값을 나타내므로 매출액의 평균값을 추정한 결과는 믿을 수 있을 것으로 생각된다. 다음으로 대체를 하고 난 후의 구성비(분포)의 변화를 살펴보고자 한다. <표 3-15>에는 무응답 비율이 2%인 경우에 대체전후의 구성비 변화의 결과가 주어져있다.

〈표 3-15〉 매출액의 대체전후 구성 변화(무응답비율 2%인 경우)

범주(백만원)	20 미만	20 - 50	50 - 100	100 - 200	200 이상
실제 사업체수	3298	5690	4378	2708	2105
평균 대체방법					
대체후 사업체수	3241	5659	4456	2723	2100
절대차이(오차)	**57(1.7%)**	**31(0.5%)**	**78(1.8%)**	**15(0.6%)**	**5(0.2%)**
회귀(직선, 절편 o) 대체방법					
대체후 사업체수	3264	5630	4450	2723	2112
절대차이(오차)	**34(1.0)**	**60(1.1%)**	**72(1.6%)**	**15(0.6%)**	**7(0.3%)**
회귀(직선, 절편 x) 대체방법					
대체후 사업체수	3239	5659	4442	2726	2113
절대차이(오차)	**59(1.8%)**	**31(0.5%)**	**64(1.5%)**	**18(0.7%)**	**8(0.4%)**
비 대체방법					
대체후 사업체수	3239	5655	4440	2731	2114
절대차이(오차)	**59(1.8%)**	**35(0.6%)**	**62(1.4%)**	**23(0.8%)**	**9(0.4%)**
응용 핫덱 대체방법					
대체후 사업체수	3310	5693	4379	2693	2104
절대차이(오차)	**12(0.4%)**	**3(0.1%)**	**1(0.02%)**	**15(0.6%)**	**1(0.05%)**

무응답 비율이 2%인 경우 응용 핫덱 대체방법을 제외한 나머지 방법들은 매출액이 20 미만인 범주에서 대체전후의 구성비 변화가 2% 가까이 나타나고 있다. 물론 다른 범주에서도 다소 큰 변화가 보이기도 한다. 하지만, 응용 핫덱 대체방법은 모든 범주에서 1%이하의 변화만을 보이며 안정적인 대체가 되고 있음을 알 수 있다. 응용 핫덱 이외의 대체방법에서는 무응답 비율이 높은 경우에 더 심각한 분포의 왜곡 현상이 나타나고 있다. <표 3-16>과 <표 3-17>에는 무응답 비율이 5%, 10%인 경우에 대체전후의 구성비 변화의 결과가 주어져있다.

무응답 비율이 5%가 되면 응용 핫덱 이외의 대체방법들은 1~3%정도의 구성비 변화가 일어나며, 무응답 비율이 10%가 되면 구성비 변화 역시 대략적으로 2배 정도로 늘어나고 있음을 알 수 있다. 각 범주에 대해 최대 6% 정도까지 변화가 일어나고 있다. 하지만, 응용 핫덱 대체방법의 경우에는 무응답 비율이 높아져도 대체전후의 구성비 변화가

1% 미만에서 유지되고 있는 것을 볼 수 있다. 이러한 결과는 응용 핫덱 대체방법의 큰 장점이라 할 수 있을 것이다.

〈표 3-16〉 매출액의 대체전후 구성 변화(무응답비율 5%인 경우)

범주(백만원)	20 미만	20 - 50	50 - 100	100 - 200	200 이상
실제 사업체수	3298	5690	4378	2708	2105
평균 대체방법					
대체후 사업체수	3226	5603	4493	2775	2082
절대차이(오차)	**72(2.2%)**	**87(1.5%)**	**115(2.6%)**	**67(2.5%)**	**23(1.1%)**
회귀(직선, 절편 o) 대체방법					
대체후 사업체수	3247	5584	4490	2739	2119
절대차이(오차)	**51(1.5)**	**106(1.9%)**	**112(2.6%)**	**31(1.1%)**	**14(0.7%)**
회귀(직선, 절편 x) 대체방법					
대체후 사업체수	3195	5617	4486	2770	2111
절대차이(오차)	**103(3.1%)**	**73(1.3%)**	**108(2.5%)**	**62(2.3%)**	**6(0.3%)**
비 대체방법					
대체후 사업체수	3210	5604	4467	2791	2107
절대차이(오차)	**88(2.7%)**	**86(1.5%)**	**89(2.0%)**	**83(3.1%)**	**2(0.1%)**
응용 핫덱 대체방법					
대체후 사업체수	3308	5697	4356	2723	2095
절대차이(오차)	**10(0.3%)**	**7(0.1%)**	**22(0.5%)**	**15(0.6%)**	**10(0.5%)**

〈표 3-17〉 매출액의 대체전후 구성 변화(무응답비율 10%인 경우)

범주(백만원)	20 미만	20 - 50	50 - 100	100 - 200	200 이상
실제 사업체수	3298	5690	4378	2708	2105
평균 대체방법					
대체후 사업체수	3132	5555	4637	2791	2064
절대차이(오차)	**166(5.0%)**	**135(2.4%)**	**259(5.9%)**	**83(3.1%)**	**41(1.9%)**
회귀(직선, 절편 o) 대체방법					
대체후 사업체수	3161	5543	4564	2799	2112
절대차이(오차)	**137(4.2)**	**147(2.6%)**	**186(4.2%)**	**91(3.4%)**	**7(0.3%)**
회귀(직선, 절편 x) 대체방법					
대체후 사업체수	3102	5563	4612	2822	2080
절대차이(오차)	**196(5.9%)**	**127(2.2%)**	**234(5.3%)**	**114(4.2%)**	**25(1.2%)**
비 대체방법					
대체후 사업체수	3112	5533	4606	2834	2094
절대차이(오차)	**186(5.6%)**	**157(2.8%)**	**228(5.2%)**	**126(4.7%)**	**11(0.5%)**
응용 핫덱 대체방법					
대체후 사업체수	3275	5720	4359	2726	2099
절대차이(오차)	**23(0.7%)**	**30(0.5%)**	**19(0.4%)**	**18(0.7%)**	**6(0.3%)**

다. 영업비용

연관성분석의 결과 영업비용 항목은 매출액의 정보만으로도 충분히 추정할 수 있을 것으로 판단된다. 따라서 매출액만을 대체군으로 사용한다. 실제 영업비용을 대체할 때에는 매출액이 먼저 대체되므로 이 자료를 이용할 수 있을 것이다. 본 모의실험에서는 영업비용과 매출액의 값이 모두 포함된 7,849개의 자료를 모집단이라 가정하여 실험을 실시하였다.

〈표 3-18〉 영업비용의 모의실험 결과(mean=249.93, 단위(백만원))

무응답 비율	통계량	대체방법				
		평균 대체	회귀 대체 (직선, 절편o)	회귀 대체 (직선, 절편x)	비 대체	응용 핫덱 대체
2%	MAE	**2.70(1.08%)**	**0.85(0.34%)**	**0.97(0.39%)**	**0.74(0.30%)**	**0.74(0.30%)**
	CV	0.0135	0.0055	0.0059	0.0052	0.0051
5%	MAE	**3.47(1.39%)**	**1.15(0.46%)**	**1.28(0.51%)**	**1.02(0.41%)**	**1.19(0.48%)**
	CV	0.0187	0.0065	0.0071	0.0060	0.0065
10%	MAE	**5.44(2.18%)**	**1.73(0.69%)**	**1.79(0.72%)**	**1.62(0.65%)**	**1.76(0.70%)**
	CV	0.0267	0.0092	0.0094	0.0088	0.0094

〈표 3-19〉 영업비용의 대체전후 구성 변화(무응답비율 2%인 경우)

범주(백만원)	20 미만	20 - 50	50 - 100	100 - 200	200 이상
실제 사업체수	1751	1942	1580	1231	1345
평균 대체방법					
대체후 사업체수	1745	1946	1583	1235	1340
절대차이(오차)	**6(0.3%)**	**4(0.2%)**	**3(0.2%)**	**4(0.3%)**	**5(0.2%)**
회귀(직선, 절편 o) 대체방법					
대체후 사업체수	1721	1954	1585	1241	1348
절대차이(오차)	**30(1.7)**	**12(0.6%)**	**5(0.3%)**	**10(0.8%)**	**3(0.2%)**
회귀(직선, 절편 x) 대체방법					
대체후 사업체수	1740	1947	1585	1237	1340
절대차이(오차)	**11(0.6%)**	**5(0.3%)**	**5(0.3%)**	**6(0.5%)**	**5(0.4%)**
비 대체방법					
대체후 사업체수	1739	1949	1585	1235	1341
절대차이(오차)	**12(0.7%)**	**7(0.4%)**	**5(0.3%)**	**4(0.3%)**	**4(0.3%)**
응용 핫덱 대체방법					
대체후 사업체수	1746	1945	1579	1235	1344
절대차이(오차)	**5(0.3%)**	**3(0.2%)**	**1(0.1%)**	**4(0.3%)**	**1(0.1%)**

영업비용에 대한 종합적인 대체결과가 <표 3-18>에 정리되어 있다. 종사자수와 매출액에 대한 결과와 유사함을 알 수 있다. 평균 대체방법은 다른 방법보다 다소 정확성이 떨어지는 것을 볼 수 있으며, 나머지 방법들의 차이는 크지 않음을 알 수 있다. 영업비용은 무응답 비율이 10%가 되어도 평균 추정에 대한 오차 비율이 0.7% 정도로 매우 정확한 추정이 되고 있다. 이는 영업비용과 매출액의 연관성이 매우 커서 한쪽 값을 알면 나머지 값을 매우 정확하게 추정할 수 있기 때문이다. 그리고 추정의 안정성을 알 수 있는 CV값 역시 매우 작은 값을 나타내므로 영업비용의 평균값을 추정한 결과는 믿을 수 있을 것으로 생각된다. 다음으로 대체를 하고 난 후의 구성비(분포)의 변화를 살펴보고자 한다. <표 3-19>에는 무응답 비율이 2%인 경우에 대체전후의 구성비 변화의 결과가 주어져있다.

무응답 비율이 2%인 경우 회귀(절편 있음) 대체방법을 제외하고는 대체전후의 구성비 변화는 크지 않은 것으로 판단된다. 대부분의 범주에서 1% 미만의 변화를 보이고 있는 것을 알 수 있다. 하지만 무응답 비율이 증가함으로써 응용 핫덱 대체방법을 제외한 방법들은 구성비 변화 비율이 증가하고 있음을 볼 수 있다. <표 3-20>과 <표 3-21>을 참조하기 바란다.

〈표 3-20〉 영업비용의 대체전후 구성 변화(무응답비율 5%인 경우)

범주(백만원)	20 미만	20 - 50	50 - 100	100 - 200	200 이상
실제 사업체수	1751	1942	1580	1231	1345
평균 대체방법					
대체후 사업체수	1732	1954	1590	1223	1349
절대차이(오차)	**19(1.1%)**	**12(0.6%)**	**10(0.6%)**	**8(0.6%)**	**4(0.3%)**
회귀(직선, 절편 o) 대체방법					
대체후 사업체수	1683	1948	1624	1237	1357
절대차이(오차)	**68(3.9)**	**6(0.3%)**	**44(2.8%)**	**6(0.5%)**	**12(0.9%)**
회귀(직선, 절편 x) 대체방법					
대체후 사업체수	1721	1953	1596	1238	1341
절대차이(오차)	**30(1.7%)**	**11(0.6%)**	**16(1.0%)**	**7(0.6%)**	**4(0.3%)**
비 대체방법					
대체후 사업체수	1720	1948	1594	1239	1348
절대차이(오차)	**31(1.8%)**	**6(0.3%)**	**14(0.9%)**	**8(0.6%)**	**3(0.2%)**
응용 핫덱 대체방법					
대체후 사업체수	1754	1937	1580	1236	1342
절대차이(오차)	**3(0.2%)**	**5(0.3%)**	**0(0.0%)**	**5(0.4%)**	**3(0.2%)**

〈표 3-21〉 영업비용의 대체전후 구성 변화(무응답비율 10%인 경우)

범주(백만원)	20 미만	20 - 50	50 - 100	100 - 200	200 이상
실제 사업체수	1751	1942	1580	1231	1345
평균 대체방법					
대체후 사업체수	1721	1960	1592	1222	1354
절대차이(오차)	**30(1.7%)**	**18(0.9%)**	**12(0.8%)**	**9(0.7%)**	**9(0.7%)**
회귀(직선, 절편 o) 대체방법					
대체후 사업체수	1632	1961	1648	1251	1357
절대차이(오차)	**119(6.8)**	**19(1.0%)**	**68(4.3%)**	**20(1.6%)**	**12(0.9%)**
회귀(직선, 절편 x) 대체방법					
대체후 사업체수	1708	1951	1606	1246	1338
절대차이(오차)	**43(2.5%)**	**9(0.5%)**	**26(1.6%)**	**15(1.2%)**	**7(0.5%)**
비 대체방법					
대체후 사업체수	1697	1944	1604	1250	1354
절대차이(오차)	**54(3.1%)**	**2(0.1%)**	**24(1.5%)**	**19(1.5%)**	**9(0.7%)**
응용 핫덱 대체방법					
대체후 사업체수	1747	1938	1585	1242	1337
절대차이(오차)	**4(0.2%)**	**4(0.2%)**	**5(0.3%)**	**11(0.9%)**	**8(0.6%)**

각 방법별 및 범주별로 차이는 있지만 무응답 비율이 5%인 경우에는 대략 1~2%, 10%인 경우에는 1~3% 정도의 변화를 보이고 있다. 하지만, 응용 핫덱 대체방법은 1% 미만의 비율을 유지하고 있는 것을 볼 수 있다. 영업비용의 경우 대체의 정확도 측면에서 각 방법별로 큰 차이를 보이지는 않지만 응용 핫덱 대체방법을 통해서 가장 안정적인 대체를 할 수 있을 것으로 생각된다.

라. 연면적

연관성분석의 결과 연면적 항목은 매출액, 세세분류, 조직형태, 종사자수, 읍면동 항목을 대체군으로 이용할 수 있으나, 시범예행조사 특성상 시군구, 소분류, 매출액, 종사자수 항목만을 대체군으로 사용해야 한다. 실제 대체에서는 연면적을 대체하기에 앞서 종사자수와 매출액이 대체되므로 이 자료를 이용할 것이다. 본 모의실험에서는 연면적, 종사자수 및 매출액의 값이 모두 포함된 7,830개의 자료를 모집단이라 가정하여 실험을 실시하였다.

〈표 3-22〉 연면적의 모의실험 결과(mean=529.73, 단위(m^2))

무응답 비율	통계량	대체방법				
		평균 대체	회귀 대체 (직선, 절편o)	회귀 대체 (직선, 절편x)	비 대체	응용 핫덱 대체
2%	MAE	**4.64(0.88%)**	**2.31(0.44%)**	**2.51(0.47%)**	**3.02(0.57%)**	**2.49(0.47%)**
	CV	0.0110	0.0064	0.0075	0.0097	0.0070
5%	MAE	**10.13(1.91%)**	**4.98(0.94%)**	**5.59(1.06%)**	**6.10(1.15%)**	**5.46(1.03%)**
	CV	0.0152	0.0104	0.0115	0.0124	0.0112
10%	MAE	**20.19(3.81%)**	**8.95(1.69%)**	**10.33(1.95%)**	**11.02(2.08%)**	**9.62(1.82%)**
	CV	0.0309	0.0204	0.0243	0.0244	0.0217

연면적에 대한 종합적인 대체결과가 <표 3-22>에 정리되어 있다. 연면적 평균 추정은 회귀 및 응용 핫덱 대체방법이 조금 더 정확한 것으로 나타났다. 평균 대체방법은 다소 정확성이 떨어지는 것을 볼 수 있다. 무응답 비율이 2%, 5%, 10%로 높아질수록 오차 비율도 약 0.5%, 1%, 2%로 높아지는 것을 알 수 있다. 이전의 항목들에 비해서는 오차가 크다고 생각되나 무응답 비율을 고려하면 큰 문제는 되지 않을 것으로 판단된다. 그리고 CV값은 안정적으로 나타나고 있다.

〈표 3-23〉 연면적의 대체전후 구성 변화(무응답비율 2%인 경우)

범주(m^2)	50 미만	50 - 100	100 - 150	150 - 200	200 이상
실제 사업체수	1928	2413	1055	724	1710
평균 대체방법					
대체후 사업체수	1910	2411	1080	718	1711
절대차이(오차)	**18(0.9%)**	**2(0.1%)**	**25(2.4%)**	**6(0.8%)**	**1(0.1%)**
회귀(직선, 절편 o) 대체방법					
대체후 사업체수	1899	2425	1069	719	1718
절대차이(오차)	**29(1.5%)**	**12(0.5%)**	**11(1.0%)**	**5(0.7%)**	**8(0.5%)**
회귀(직선, 절편 x) 대체방법					
대체후 사업체수	1942	2403	1056	714	1715
절대차이(오차)	**14(0.7%)**	**10(0.4%)**	**1(0.1%)**	**10(1.4%)**	**5(0.3%)**
비 대체방법					
대체후 사업체수	1943	2390	1059	717	1721
절대차이(오차)	**15(0.8%)**	**23(1.0%)**	**4(0.4%)**	**7(1.0%)**	**11(0.6%)**
응용 핫덱 대체방법					
대체후 사업체수	1932	2407	1056	726	1709
절대차이(오차)	**4(0.2%)**	**6(0.2%)**	**1(0.1%)**	**2(0.3%)**	**1(0.1%)**

〈표 3-24〉 연면적의 대체전후 구성 변화(무응답비율 5%인 경우)

범주(m^2)	50 미만	50 - 100	100 - 150	150 - 200	200 이상
실제 사업체수	1928	2413	1055	724	1710
평균 대체방법					
대체후 사업체수	1891	2422	1105	715	1697
절대차이(오차)	**37(1.9%)**	**9(0.4%)**	**50(4.7%)**	**9(1.2%)**	**13(0.8%)**
회귀(직선, 절편 o) 대체방법					
대체후 사업체수	1874	2447	1076	713	1720
절대차이(오차)	**54(2.8%)**	**34(1.4%)**	**21(2.0%)**	**11(1.5%)**	**10(0.6%)**
회귀(직선, 절편 x) 대체방법					
대체후 사업체수	1966	2398	1042	707	1717
절대차이(오차)	**38(2.0%)**	**15(0.6%)**	**13(1.2%)**	**17(2.3%)**	**7(0.4%)**
비 대체방법					
대체후 사업체수	1973	2384	1042	712	1719
절대차이(오차)	**45(2.3%)**	**29(1.2%)**	**13(1.2%)**	**12(1.7%)**	**9(0.5%)**
응용 핫덱 대체방법					
대체후 사업체수	1924	2408	1065	718	1715
절대차이(오차)	**4(0.2%)**	**5(0.2%)**	**10(0.9%)**	**6(0.8%)**	**5(0.3%)**

〈표 3-25〉 연면적의 대체전후 구성 변화(무응답비율 10%인 경우)

범주(m^2)	50 미만	50 - 100	100 - 150	150 - 200	200 이상
실제 사업체수	1928	2413	1055	724	1710
평균 대체방법					
대체후 사업체수	1845	2435	1156	718	1676
절대차이(오차)	**83(4.3%)**	**22(0.9%)**	**101(9.6%)**	**6(0.8%)**	**34(2.0%)**
회귀(직선, 절편 o) 대체방법					
대체후 사업체수	1804	2484	1105	716	1721
절대차이(오차)	**124(6.4%)**	**71(2.9%)**	**50(4.7%)**	**8(1.1%)**	**11(0.6%)**
회귀(직선, 절편 x) 대체방법					
대체후 사업체수	2001	2394	1033	712	1690
절대차이(오차)	**73(3.8%)**	**19(0.8%)**	**22(2.1%)**	**12(1.7%)**	**20(1.2%)**
비 대체방법					
대체후 사업체수	2009	2372	1038	719	1692
절대차이(오차)	**81(4.2%)**	**41(1.7%)**	**17(1.6%)**	**5(0.7%)**	**18(1.1%)**
응용 핫덱 대체방법					
대체후 사업체수	1913	2431	1046	727	1713
절대차이(오차)	**15(0.8%)**	**18(0.7%)**	**10(0.9%)**	**3(0.4%)**	**3(0.2%)**

다음으로 대체를 하고 난 후의 구성비(분포)의 변화를 살펴보고자 한다. <표 3-23>에는 무응답 비율이 2%인 경우에 대체전후의 구성비 변화의 결과가 주어져있다. 대체전후의 구성비 변화 비율은 영업비용 항목과 유사한 패턴으로 나타났다. 응용 핫덱 대체방법을 제외한 방법들은 무응답 비율이 2%인 경우 각 범주별로 1% 전후의 변화를 보이고 있으며, 응용 핫덱 대체방법은 모든 범주에서 1% 미만의 변화를 보이고 있는 것을 알 수 있다. 무응답 비율이 5%, 10%로 증가할수록 무응답 대체전후의 구성비 변화 비율도 증가하는데 각 방법별 및 범주별로 차이는 있지만 무응답 비율이 5%인 경우에는 대략 1~3%, 10%인 경우에는 1~5% 정도의 변화를 보이고 있다. 하지만, 응용 핫덱 대체방법은 1% 미만의 비율을 유지하고 있는 것을 볼 수 있다. 자세한 내용은 <표 3-24>와 <표 3-25>를 참조하기 바란다.

마. 영업개월수

연관성분석의 결과 영업개월수 항목은 매출액, 세세분류, 연면적, 읍면동, 종사자수 항목을 대체군으로 이용할 수 있으나, 시범예행조사 특성상 시군구, 소분류, 매출액, 연면적, 종사자수 항목만을 대체군으로 사용해야 한다. 실제 대체에서는 영업개월수 대체전에 종사자수, 매출액 및 연면적이 대체되므로 이 자료를 이용할 것이다. 본 모의실험에서는 영업개월수, 연면적, 종사자수 및 매출액의 값이 모두 포함된 7,830개의 자료를 모집단이라 가정하여 실험을 실시하였다.

영업개월수 항목은 범주형으로 간주하고 분석하였다. 따라서 모의실험을 위한 대체방법으로 최빈수 대체방법과 응용 핫덱 대체방법 2가지 방법을 고려하였다. 확률에 의한 대체방법을 생각할 수 있으나, 이는 응용 핫덱 대체방법의 원리와 거의 유사하여 제외하도록 한다. 그리고 범주형 항목에서는 평균 추정은 의미가 없으므로 대체전후의 구성비 변화 측면에서만 정확도를 평가하고자 한다. <표 3-26>에는 무응답 비율이 2%인 경우에 대체전후의 구성비 변화의 결과가 주어져있다.

〈표 3-26〉 영업개월수의 대체전후 구성 변화(무응답비율 2%인 경우)

범주(개월)	1	2	3	4	5	6	7	8	9	10	11	12
실제 사업체수	93	126	100	101	82	98	88	122	97	101	59	6763
최빈수 대체방법												
대체후 사업체수	92	124	101	98	81	97	88	121	96	98	58	6776
절대차이	**1**	**2**	**1**	**3**	**1**	**1**	**0**	**1**	**1**	**3**	**1**	**13**
응용 핫덱 대체방법												
대체후 사업체수	93	124	102	100	81	97	88	122	97	102	58	6766
절대차이	**0**	**2**	**2**	**1**	**1**	**1**	**0**	**0**	**0**	**1**	**1**	**3**

무응답 비율이 2%인 경우 영업개월수가 1~11개월의 구성비는 두 방법 모두 거의 차이가 나지 않음을 볼 수 있다. 하지만, 영업개월수가 12개월의 경우에서 차이가 보인다. 이것은 최빈수 대체방법이 도너들 중에서 가장 빈도수가 높은 값으로 대체하기 때문에 총 빈도수에서 비율이 매우 큰 12개월의 값을 주로 대체하게 되는 특성 때문에 나타나게 된다. 또한 이러한 현상은 무응답 비율이 높아질수록 확연히 보인다. <표 3-27>과 <표 3-28>에는 무응답 비율이 5%, 10%인 경우의 대체결과가 주어져있다.

〈표 3-27〉 영업개월수의 대체전후 구성 변화(무응답비율 5%인 경우)

범주(개월)	1	2	3	4	5	6	7	8	9	10	11	12
실제 사업체수	93	126	100	101	82	98	88	122	97	101	59	6763
최빈수 대체방법												
대체후 사업체수	90	127	87	98	78	97	84	118	89	96	55	6811
절대차이	**3**	**1**	**13**	**3**	**4**	**1**	**4**	**4**	**8**	**5**	**4**	**48**
응용 핫덱 대체방법												
대체후 사업체수	91	130	96	102	81	99	87	121	94	99	58	6772
절대차이	**2**	**4**	**4**	**1**	**1**	**1**	**1**	**1**	**3**	**2**	**1**	**9**

〈표 3-28〉 영업개월수의 대체전후 구성 변화(무응답비율 10%인 경우)

범주(개월)	1	2	3	4	5	6	7	8	9	10	11	12
실제 사업체수	93	126	100	101	82	98	88	122	97	101	59	6763
최빈수 대체방법												
대체후 사업체수	82	114	91	93	73	91	86	111	92	93	55	6849
절대차이	**11**	**12**	**9**	**8**	**9**	**7**	**2**	**11**	**5**	**8**	**4**	**86**
응용 핫덱 대체방법												
대체후 사업체수	87	121	96	100	80	98	92	119	101	97	61	6778
절대차이	**6**	**5**	**4**	**1**	**2**	**0**	**4**	**3**	**4**	**4**	**2**	**15**

최빈수 대체방법은 무응답 비율이 높아질수록 12개월의 빈도수 변화가 더 커지고 있음을 볼 수 있다. 그 외의 범주에서도 응용 핫덱 대체방법보다 조금 더 높은 구성비 변화가 나타나고 있다. 반면에 응용 핫덱 대체방법은 무응답 비율이 증가하더라도 대체전후의 구성비 변화가 크지 않으며 안정적인 대체가 되고 있음을 알 수 있다. 따라서 영업개월수 항목의 대체에 응용 핫덱 대체방법이 적절할 것으로 판단된다. 그리고 최빈수 대체방법은 각 범주의 총 빈도수의 차이가 클수록 대체 오차가 커질 수 있음을 명심해야 할 것이다.

바. 휴무일수

연관성분석의 결과 휴무일수 항목은 연면적, 세세분류, 일일영업시간, 매출액 항목을 대체군으로 이용할 수 있으나, 시범예행조사 특성상 연면적, 소분류, 매출액 항목만을 대체군으로 사용해야 한다. 일일영업시간 항목은 휴무일수를 대체한 후 대체가 되므로 사용할 수 없다. 본 모의실험에서는 휴무일수, 연면적 및 매출액의 값이 모두 포함된 7,830개의 자료를 모집단이라 가정하여 실험을 실시하였다. 휴무일수도 범주형 항목으로 영업개월 수와 같은 방법으로 모의실험을 실시하였다. <표 3-29>에는 무응답 비율이 2%인 경우에 대체전후의 구성비 변화의 결과가 주어져있다.

〈표 3-29〉 휴무일수의 대체전후 구성 변화(무응답비율 2%인 경우)

범주(일)	1	2~3	4~5	6~7	8 이상	없음
실제 사업체수	626	1582	1409	87	198	3928
최빈수 대체방법						
대체후 사업체수	618	1565	1390	87	199	3971
절대차이	**8**	**17**	**19**	**0**	**1**	**43**
응용 핫덱 대체방법						
대체후 사업체수	625	1578	1406	91	197	3933
절대차이	**1**	**4**	**3**	**4**	**1**	**5**

영업개월수 항목과 유사하게 최빈수 대체방법은 범주별 총 빈도수가 큰 범주에서 구성비 차이가 나고 있음을 보여준다. 특히 없음의 범주에서 응용 핫덱 대체방법은 5개 차이가 나는데 반면 최빈수 대체방법은 43개로 많은 차이를 보여주고 있다. 무응답 비율이 높아질수록 더 많은 차이가 나는 것을 볼 수 있다. <표 3-30>과 <표 3-31>에는 무응답 비율이 5%, 10%인 경우의 대체결과가 주어져있다.

〈표 3-30〉 휴무일수의 대체전후 구성 변화(무응답비율 5%인 경우)

범주(일)	1	2~3	4~5	6~7	8 이상	없음
실제 사업체수	626	1582	1409	87	198	3928
최빈수 대체방법						
대체후 사업체수	601	1547	1366	89	191	4036
절대차이	**25**	**35**	**43**	**2**	**7**	**108**
응용 핫덱 대체방법						
대체후 사업체수	628	1574	1404	90	202	3932
절대차이	**2**	**8**	**5**	**3**	**4**	**4**

〈표 3-31〉 휴무일수의 대체전후 구성 변화(무응답비율 10%인 경우)

범주(일)	1	2~3	4~5	6~7	8 이상	없음
실제 사업체수	626	1582	1409	87	198	3928
최빈수 대체방법						
대체후 사업체수	589	1483	1337	75	185	4161
절대차이	**37**	**99**	**72**	**12**	**13**	**233**
응용 핫덱 대체방법						
대체후 사업체수	634	1569	1404	82	205	3936
절대차이	**8**	**13**	**5**	**5**	**7**	**8**

최빈수 대체방법은 무응답 비율이 높아질수록 2~3일, 4~5일, 없음의 범주에서 대체전후의 빈도수 변화가 응용 핫덱 대체방법에 비해 훨씬 더 커지고 있음을 볼 수 있다. 없음 항목은 무응답 비율이 10%가 되면 무려 233개의 빈도수의 차이가 발생하게 된다. 하지만 응용 핫덱 대체방법에서는 대체전후의 구성비 변화가 크지 않으며 안정적인 대체가 되고 있음을 알 수 있다.

사. 일일영업시간

연관성분석의 결과 일일영업시간 항목은 세세분류, 휴무일수, 읍면동 항목을 대체군으로 이용할 수 있으나, 시범예행조사 특성상 소분류, 휴무일수, 시군구 항목을 대체군으로 사용할 것이다. 휴무일수 항목은 일일영업시간을 대체하기 전에 대체를 하므로 사용 가능하다. 본 모의실험에서는 일일영업시간 및 휴무일수의 값이 모두 포함된 7,830개의 자료를 모집단이라 가정하여 실험을 실시하였다. <표 3-32>에는 무응답 비율이 2%인 경우에 대체전후의 구성비 변화의 결과가 주어져있다.

〈표 3-32〉 일일영업시간의 대체전후 구성 변화(무응답비율 2%인 경우)

범주(시간)	8 미만	8~10	10~12	12~14	14 이상
실제 사업체수	717	1868	2484	1229	1532
최빈수 대체방법					
대체후 사업체수	705	1855	2539	1210	1521
절대차이	**12**	**13**	**55**	**19**	**11**
응용 핫덱 대체방법					
대체후 사업체수	714	1863	2487	1234	1532
절대차이	**3**	**5**	**3**	**5**	**0**

무응답 비율이 2%인 경우에도 두 방법에서의 차이가 크다는 것을 알 수 있다. 총 빈도수가 가장 큰 10~12시간 범주에서 응용 핫덱 대체방법은 3개 차이가 나는데 반면 최빈수 대체방법은 55개로 많은 차이를 보이고 있다. 그리고 최빈수 대체방법은 모든 범주에서 대체전후의 구성비 변화가 큰 것으로 판단된다. 이러한 변화는 무응답 비율이 높아질수록 더 심하게 나타나며 <표 3-33>과 <표 3-34>에서 확인할 수 있다.

〈표 3-33〉 일일영업시간의 대체전후 구성 변화(무응답비율 5%인 경우)

범주(시간)	8 미만	8~10	10~12	12~14	14 이상
실제 사업체수	717	1868	2484	1229	1532
최빈수 대체방법					
대체후 사업체수	675	1872	2608	1174	1501
절대차이	**42**	**4**	**124**	**55**	**31**
응용 핫덱 대체방법					
대체후 사업체수	712	1866	2492	1233	1527
절대차이	**5**	**2**	**8**	**4**	**5**

〈표 3-34〉 일일영업시간의 대체전후 구성 변화(무응답비율 10%인 경우)

범주(시간)	8 미만	8~10	10~12	12~14	14 이상
실제 사업체수	717	1868	2484	1229	1532
최빈수 대체방법					
대체후 사업체수	654	1874	2722	1099	1481
절대차이	**63**	**6**	**238**	**130**	**51**
응용 핫덱 대체방법					
대체후 사업체수	710	1861	2496	1237	1526
절대차이	**7**	**7**	**12**	**8**	**6**

무응답 비율이 10%인 경우 10~12시간 항목의 대체전후의 변화가 매우 커짐을 볼 수 있다. 실제 사업체 수보다 238개가 증가하였으며 이러한 이유로 다른 항목들은 많은 사업체가 줄어든 것을 알 수 있다. 따라서 최빈수 대체방법은 대체후에 총 빈도수가 큰 범주의 구성비는 증가하고 반대의 경우에는 감소하는 현상을 볼 수 있다. 이와는 반대로 응용 핫덱 대체방법은 총 빈도수의 확률에 근거하여 대체가 이루어지므로 최빈수 대체방법의 단점을 보완하여 대체전후의 구성비가 유지되도록 한다.

이상의 모의실험을 통하여 응용 핫덱 대체방법의 정확성과 효율성을 제시하였다. 따라서 응용 핫덱 대체방법은 연속형 및 범주형 항목 모두에 적용 가능하고, 대체의 정확

성(특히, 대체전후의 구성비)도 다른 대체방법에 비해 우수하므로 경제총조사 항목 무응답 대체를 위해 사용하는데 매우 적절할 것으로 판단된다.

3. 추가검토 내용

가. 대체변수의 대체군 포함여부 검토

대체변수의 대체군 사용에 대한 내용을 검토하고자 한다. 시범예행조사 자료에서 연면적의 경우 시군구, 소분류, 매출액, 종사자수를 대체군으로 사용해야 한다. 하지만 연면적이 무응답일 경우 매출액과 종사자수가 무응답인 경우도 많이 발생하고 있다. 매출액은 20%, 종사자수는 1.2% 정도가 무응답으로 나타났다. 따라서 이 항목들을 대체군으로 사용을 하기 위해서는 매출액과 종사자수가 먼저 대체되어야 한다. 이 경우에는 대체된 자료를 다시 연면적 항목의 대체를 위해 사용하기 때문에 새로운 오차가 발생하게 될 것이다. 반면에 대체된 항목을 대체군에서 제외하고 시군구와 소분류 정보만으로 대체를 실시할 수도 있을 것이다. 그러나 매출액과 연면적의 연관성이 매우 크므로 이로 인한 정보의 손실도 상당히 클 것으로 판단된다. 따라서 매출액 및 종사자수 항목의 대체군 사용여부를 판단하기 위해서 모의실험을 실시하였다.

〈표 3-35〉 매출액 및 종사자수의 대체군 포함여부에 따른 연면적의 대체결과

무응답 비율	통계량	응용 핫덱 대체방법	
		대체군 포함	대체군 미포함
10%	MAE(오차비율)	**10.46(1.97%)**	**14.78(2.80%)**
	CV	0.0241	0.0358

<표 3-35>에는 매출액과 종사자수 항목의 대체군 포함여부에 따라서 연면적 항목의 대체결과가 제시되어 있다. 대체군에 포함을 시킨 경우에는 평균추정에 대한 오차비율이 1.97%이며, 대체군에서 제외된 경우에는 2.80%로 나타났다. 비록 대체(추정)가 된 매출액과 종사자수의 정보를 이용하기 때문에 실제 정보와의 차이는 존재하지만, 연면적과의 연관성이 상당히 높기 때문에 대체군에 포함을 시키는 것이 더 정확한 대체가 됨을 실험을 통해서 알 수 있다. 물론 연관성이 높지 않은 항목의 경우에는 본 실험의 결과와 반대로 나올 수도 있음을 알려둔다. 그리고 <표 3-36>에는 무응답 비율이 10%인 경우에 대체전후의 구성비 변화의 결과가 주어져 있다.

〈표 3-36〉 연면적의 대체군 포함여부에 따른 구성 변화

범주(m^2)	50 미만	50 - 100	100 - 150	150 - 200	200 이상
실제 사업체수	1928	2413	1055	724	1710
대체군에 포함된 경우					
대체후 사업체수	1909	2424	1068	728	1701
절대차이(오차)	**19(1.0%)**	**11(0.5%)**	**13(1.2%)**	**4(0.6%)**	**9(0.5%)**
대체군에 미포함된 경우					
대체후 사업체수	1907	2428	1070	721	1704
절대차이(오차)	**21(1.1%)**	**15(0.6%)**	**15(1.4%)**	**3(0.4%)**	**6(0.4%)**

대체전후의 구성비 변화정도는 크지 않은 것으로 보인다. 매출액과 종사자수 항목을 대체군에 포함을 시킨 경우에는 최대 1.2% 정도의 변화가 있으며, 대체군에서 제외된 경우에는 최대 1.4% 정도로 나타났다. 두 경우에서 구성비 변화의 차이는 거의 없으며 이는 응용 핫덱 대체방법의 장점이 그대로 반영된 것으로 볼 수 있을 것이다. 결론적으로 매출액과 종사자수를 대체군에 포함시킨 경우에 평균 추정에서 정확도가 더 높으므로 이 두 항목을 대체군에 포함시키는 것이 더 적절할 것으로 판단된다.

나. 무응답 비율이 30%인 경우의 모의실험

경제총조사의 무응답 비율은 항목마다 차이가 있겠지만 통계청에서는 10%가 넘지 않을 것으로 예상하고 있다. 하지만 민감한 항목의 경우 더 많은 무응답이 발생할 가능성도 배재할 수 없을 것이다. 따라서 영업비용과 매출액 항목에 대하여 무응답 비율이 30%인 경우에 대한 모의실험을 응용 핫덱 대체방법을 이용하여 추가로 실시하였다.

〈표 3-37〉 영업비용의 추가 모의실험(무응답비율 30%인 경우)

무응답 비율(10%)의 MAE 및 CV			무응답 비율(30%)의 MAE 및 CV		
1.76(0.70%), 0.0094			**6.44(2.58%), 0.0305**		
범주(백만원)	20 미만	20 - 50	50 - 100	100 - 200	200 이상
무응답 비율 10%인 경우 대체후 구성 변화					
절대차이(오차)	**4(0.2%)**	**4(0.2%)**	**5(0.3%)**	**11(0.9%)**	**8(0.6%)**
무응답 비율 30%인 경우 대체후 구성 변화					
절대차이(오차)	**11(0.6%)**	**19(0.9%)**	**30(1.9%)**	**16(1.3%)**	**6(0.4%)**

영업비용의 추가 모의실험 결과가 <표 3-37>에 정리되어 있다. 무응답 비율이 10%인 경우 평균추정에 대한 오차 비율이 0.7% 정도였으나, 30%로 높아지면 오차 비율 역시 2.58%로 커지고 있음을 볼 수 있다. 또한 대체전후의 구성비 변화도 대부분의 범주에서 1% 미만이었으나, 2% 정도까지 발생하고 있다. <표 3-38>의 매출액에서도 유사한 결과를 볼 수 있다. 오차 비율이 1.59%에서 4.18%로 상당히 높아지고 있으며, 구성비도 최대 0.7%에서 1.6%까지 변화됨을 알 수 있다.

〈표 3-38〉 매출액의 추가 모의실험(무응답 비율 30%)

무응답 비율(10%)의 MAE 및 CV			무응답 비율(30%)의 MAE 및 CV		
2.66(1.59%), 0.0211			**6.98(4.18%), 0.0437**		
범주(백만원)	20 미만	20 - 50	50 - 100	100 - 200	200 이상
무응답 비율 10%인 경우 대체후 구성 변화					
절대차이(오차)	**23(0.7%)**	**30(0.5%)**	**19(0.4%)**	**18(0.7%)**	**6(0.3%)**
무응답 비율 30%인 경우 대체후 구성 변화					
절대차이(오차)	**30(0.9%)**	**49(0.9%)**	**43(1.0%)**	**29(1.1%)**	**33(1.6%)**

본 모의실험에서 알 수 있듯이 무응답 비율이 높아질수록 대체로 인한 오차가 커지고 있음을 볼 수 있다. 따라서 경제총조사 항목 중에서 무응답 비율이 높은 항목(무응답 비율이 10%이상)에 대해서는 대체오차 정도를 감안하여 대체 여부를 결정하여야 할 것이다. 연구자의 판단으로는 항목에 따라 차이는 있겠지만 무응답 비율이 20% 정도까지는 대체를 고려할 수 있을 것으로 생각된다.

제5절 경제총조사 시범예행조사 자료에의 적용

앞에서 제시한 각 항목의 대체군 및 응용 핫덱 대체방법을 이용하여 2010년 기준 경제총조사 시범예행조사(숙박 및 음식점업) 자료의 무응답 부분을 대체하고자 한다. 이를 통하여 획득한 자료만 사용하여 분석한 결과와 대체한 부분을 추가하여 분석한 결과를 비교해보고자 한다. 현재 시범예행자료의 무응답 비율이 매우 높으므로 적용한 결과의 정확성은 낮을 수 있으나, 실제 경제총조사 자료의 무응답 비율은 10%를 넘지 않을 것으로 예상되어 실제 적용에서는 모의실험의 결과에서 제시된 오차정도를 감안하여 판단하면 될 것이다.

<표 3-39>에는 종사자수 항목에 적용한 결과가 제시되어 있다. 종사자수의 무응답 비율은 1.2%로 매우 낮아 무응답 부분을 제외한 결과와 대체한 후의 결과 차이는 거의 나

타나지 않음을 볼 수 있다. 무응답 부분이 제외된 경우의 종사자수의 평균은 3.48(명), 대체후의 평균은 3.47(명)로 변화가 없음을 알 수 있으며 이와 같은 결과는 구성비에서도 확인할 수가 있다.

〈표 3-39〉 종사자수 항목에의 적용결과

무응답 부분이 제외된 평균			**3.48(명)**		
무응답 부분이 대체된 후의 평균			**3.47(명)**		
범주(명)	1	2	3	4	5 이상
무응답 부분이 제외된 구성비					
구성비	**25.0%**	**38.5%**	**15.8%**	**8.6%**	**12.1%**
무응답 부분이 대체된 후의 구성비					
구성비	**25.1%**	**38.4%**	**15.8%**	**8.6%**	**12.1%**

매출액 항목은 전체 자료의 20.5%가 무응답을 이루고 있다. 다소 무응답 비율이 높지만 실제 조사에서는 무응답 비율이 많이 낮아질 것으로 기대하고 있다. 매출액에의 적용 결과는 무응답 비율이 10%인 경우의 모의실험 결과와 비교할 때 2배 정도로 예상할 수 있을 것이다. 무응답 부분이 제외된 경우의 매출액의 평균은 162.77(백만원)이며 대체후의 평균은 157.92(백만원)로 대략 3%정도가 감소된 것을 볼 수 있다. 이는 매출액이 162.77(백만원)보다 작은 사업체에서 높은 비율로 무응답이 발생한 것으로 판단할 수 있다. 구성비의 변화는 거의 나타나지 않았으나 매출액이 작은 구간에서 0.1% 정도의 구성비가 증가한 것을 볼 수 있다. <표 3-40>의 매출액 항목에 적용한 결과를 참조하기 바란다.

〈표 3-40〉 매출액 항목에의 적용결과

무응답 부분이 제외된 평균			**162.77(백만원)**		
무응답 부분이 대체된 후의 평균			**157.92(백만원)**		
범주(백만원)	20 미만	20 - 50	50 - 100	100 - 200	200 이상
무응답 부분이 제외된 구성비					
구성비	**18.1%**	**31.3%**	**24.1%**	**14.9%**	**11.6%**
무응답 부분이 대체된 후의 구성비					
구성비	**18.2%**	**31.4%**	**24.1%**	**14.8%**	**11.5%**

<표 3-41>에는 영업비용 항목에 적용한 결과가 제시되어 있다. 영업비용의 무응답 비율은 59.6%로 적용을 하는 것이 무의미 할 수도 있으나, 결과의 정확도보다는 적용을 해

서 어떠한 결과를 보이는지 살펴보고자 한다. 영업비용 역시 실제조사에서는 무응답 비율이 10%보다 낮을 것으로 기대한다. 무응답 부분이 제외된 경우의 영업비용의 평균은 248.06(백만원)이며 대체후의 평균은 140.70(백만원)로 대략 43%정도가 감소된 것을 볼 수 있다. 따라서 영업비용이 작은 사업체에서 많은 부분 무응답이 발생하였음을 알 수 있다. 이와 같은 결과는 구성비 변화에서도 확인할 수가 있다. 영업비용이 20(백만원)미만은 23.1%에서 31.8%로, 20-50(백만원)도 25.0%에서 29.8%로 큰 폭으로 늘어난 반면 100-200(백만원)은 14.9%에서 10.3%, 200(백만원) 이상은 17.5%에서 9.6%로 상당부분 줄어든 것을 볼 수 있다. 일반적으로 영업비용은 매출액보다 적은 경향이 있는데 적용 후의 자료는 적절한 것으로 판단된다.

〈표 3-41〉 영업비용 항목에의 적용결과

무응답 부분이 제외된 평균			**248.06(백만원)**		
무응답 부분이 대체된 후의 평균			**140.70(백만원)**		
범주(백만원)	20 미만	20 - 50	50 - 100	100 - 200	200 이상
무응답 부분이 제외된 구성비					
구성비	**23.1%**	**25.0%**	**19.5%**	**14.9%**	**17.5%**
무응답 부분이 대체된 후의 구성비					
구성비	**31.8%**	**29.8%**	**18.5%**	**10.3%**	**9.6%**

〈표 3-42〉 연면적 항목에의 적용결과

무응답 부분이 제외된 평균			**550.80(m^2)**		
무응답 부분이 대체된 후의 평균			**294.69(m^2)**		
범주(m^2)	50 미만	50 - 100	100 - 150	150 - 200	200 이상
무응답 부분이 제외된 구성비					
구성비	**24.3%**	**30.9%**	**13.1%**	**9.0%**	**22.7%**
무응답 부분이 대체된 후의 구성비					
구성비	**6.8%**	**26.0%**	**33.6%**	**18.5%**	**15.1%**

무응답 부분이 제외된 경우의 연면적의 평균은 550.80(m^2)이며 대체후의 평균은 294.69(m^2)로 대략 46%정도가 감소된 것을 볼 수 있다. 따라서 연면적이 550.80(m^2)보다 작은 사업체에서 응답을 많이 하지 않은 것으로 판단된다. 이와 같은 결과는 구성비 변화에서도 확인할 수가 있는데 연면적이 100-150(m^2)은 13.1%에서 33.6%로, 150-200(m^2)은 9.0%에서 18.5%로 큰 폭으로 늘어난 반면 200(m^2) 이상에서는 22.7%에서 15.1%로 많이 줄어든 것을 볼 수 있다. <표 3-42>의 연면적 항목에 적용한 결과를 참조하기 바란다.

〈표 3-43〉 영업개월수 항목에의 적용결과

범주(개월)	1	2	3	4	5	6	7	8	9	10	11	12
무응답 부분이 제외된 구성비												
구성비(%)	**1.32**	**1.54**	**1.28**	**1.24**	**1.02**	**1.32**	**1.01**	**1.51**	**1.15**	**1.15**	**0.69**	**86.77**
무응답 부분이 대체된 후의 구성비												
구성비(%)	**1.37**	**1.74**	**1.54**	**1.71**	**1.20**	**1.59**	**1.18**	**1.52**	**1.30**	**1.19**	**0.73**	**84.93**

<표 3-43>에는 영업개월수 항목에 적용한 결과가 제시되어 있다. 본 항목의 경우 무응답 부분이 제외된 구성비와 대체된 후의 구성비 차이는 12(개월)에서 2% 줄어든 것 이외의 차이는 거의 나타나지 않았다. 이러한 결과는 무응답 비율이 높지만 무응답의 특성과 영업개월수와의 관련성은 거의 존재하지 않기 때문인 것으로 판단된다.

휴무일수 항목의 경우 휴무없음의 구성비가 43.5%에서 40.2%로 3.3%, 1(일)은 15.9%에서 14.3%로 1.6% 줄어든 반면 4-5일의 경우가 17.5%에서 20.8%로 3.3% 늘어난 것을 볼 수 있다. 나머지 범주에서는 큰 변화는 보이지 않는다. 일일영업시간 항목의 경우 14(시간) 이상의 구성비가 20.3%에서 17.0%로 3.3% 줄어들었으며, 나머지 범주에서는 약간 늘어난 것으로 보인다. 하지만, 휴무일수와 일일영업시간도 영업개월수와 마찬가지로 무응답 패턴이 크게 나타나지 않는 것으로 판단된다. <표 3-44>와 <표 3-45>의 휴무일수와 일일영업시간 항목에 적용한 결과를 참조하기 바란다.

〈표 3-44〉 휴무일수 항목에의 적용결과

범주(일)	1	2~3	4~5	6~7	8 이상	없음
무응답 부분이 제외된 구성비						
구성비	**15.9%**	**17.8%**	**17.5%**	**2.9%**	**2.4%**	**43.5%**
무응답 부분이 대체된 후의 구성비						
구성비	**14.3%**	**18.3%**	**20.8%**	**3.4%**	**3.0%**	**40.2%**

〈표 3-45〉 일일영업시간 항목에의 적용결과

범주(시간)	8 미만	8~10	10~12	12~14	14 이상
무응답 부분이 제외된 구성비					
구성비	**9.5%**	**23.6%**	**30.9%**	**15.7%**	**20.3%**
무응답 부분이 대체된 후의 구성비					
구성비	**10.2%**	**24.9%**	**32.0%**	**15.9%**	**17.0%**

제6절 결 론

본 연구에서는 2010년 기준 경제총조사 시범예행조사 자료를 이용하여 2011년 5-6월에 실시되는 경제총조사(숙박 및 음식점업)의 주요 항목(종사자수, 매출액, 영업비용, 연면적, 영업개월수, 휴무일수, 일일영업시간)에 대하여 연관성 분석을 실시하여 대체군을 제시하였으며, 여러 대체방법(평균 회귀, 비, 응용 핫덱, 최빈수)을 이용한 모의실험을 실시하여 대체의 정확성을 검토하였다. 또한 연구 결과를 바탕으로 시범예행조사 자료의 실제 무응답 부분에 대하여 대체를 실시하여 대체전후의 평균 및 구성비 변화를 살펴보았다.

본 연구에서 사용된 시범예행조사 자료는 3개 시·도(4개 시군구)로 한정되어 있으며, 숙박 및 음식점업인 경우 총 자료의 수가 22,866(개)이나 이 중 많은 부분이 무응답을 포함하고 있어 연구에 많은 제약이 발생하였다. 특히, 무응답 부분을 제외하고 만든 완전자료의 수가 적어서 연관성분석의 결과로 제시된 대체군 모두를 모의실험에서 사용하기 어려워 대체로 인한 오차가 실제보다 더 커졌을 가능성도 있어 보인다. 또한 연관성분석의 결과가 경제총조사 자료를 이용했을 경우와 조금의 변화가 발생할 수도 있을 것이다. 하지만 항목 간의 연관성이 높고 낮은 정도의 문제이지 대체군 포함여부에는 큰 변화가 없을 것으로 판단된다. 그러므로 대체군을 설정할 때, 지역을 읍면동으로 산업분류를 세세분류로 하여도 경제총조사 자료가 방대하므로 충분히 대체가 가능할 것으로 생각된다. 그리고 본 연구에서 제시되지 않은 타 조사표 항목에 대해서는 이 결과를 근거로 대체군 설정 작업을 추가로 하여야 할 것이다. 그러나 대부분의 조사표에서 주요 항목은 유사하므로 어려운 작업은 되지 않을 것이다.

본 연구의 모의실험에서는 각 항목별로 응답된 자료를 이용하였고, 실험을 위해 각각의 목표변수에 대해서 다양한 비율로 무응답을 발생시켰다. 그리고 평균 대체, 회귀대체, 비 대체, 최빈수 대체, 응용 핫덱 대체방법을 이용하여 항목별로 무응답을 모두 대체하고난 후에 평균의 차이정도와 임의로 범주화된 분포변화 비율을 살펴보았다. 모의실험을 실시한 항목들은 유사한 결과를 보여주고 있다. 평균 추정에 대해서는 평균 대체방법을 제외한 회귀, 비, 응용 핫덱 대체방법들이 비슷한 수준의 오차비율을 보이고 있다. 또한 무응답 비율이 높아질수록 오차비율도 높아짐을 알 수 있다. 반면에 구성비 변화 측면에서는 상대적으로 응용 핫덱 대체방법이 훨씬 더 정확한 추정을 하고 있음을 알 수 있다. 무응답 비율이 10%인 경우 응용 핫덱 대체방법은 1%전후의 오차를 유지하고 있으나, 타 방법들은 3~5%정도까지 오차가 커지는 것을 볼 수 있다. 따라서 이러한 모의실험 결과를 고려할 때 경제총조사 무응답 대체방법으로 응용 핫덱 대체방법이 가장 적절할 것으로 판단된다.

연구결과를 바탕으로 2010년 기준 경제총조사 시범예행조사 자료의 무응답 부분을 대체한 결과, 영업비용이 적고 연면적이 작은 사업체에서 무응답이 많이 발생한 것으로 보인다. 그러므로 이러한 무응답 부분을 제외한 결과를 사용하는 것은 상당히 왜곡된 결과를 가져올 수도 있다는 것을 보여준다. 모든 항목들에 대한 자세한 적용결과는 제5절을 참조하기 바란다.

본 연구의 모의실험은 무응답 비율이 10%인 경우에 한해서 진행을 하였으며 매출액과 영업비용 항목에 대해서는 무응답 비율이 30%일 때의 실험을 추가로 실시하였다. 무응답 비율이 높아질수록 평균추정에 대한 오차비율과 대체전후의 구성비 변화 비율이 커지는 것을 볼 수 있으며, 경제총조사 자료에의 적용을 위해서는 실제 자료의 분석을 통하여 항목과 무응답 비율을 고려하여 대체여부를 결정하여야 할 것이다.

마지막으로 모든 연구는 경제총조사 자료가 아닌 시범예행조사 자료를 이용하였으므로 연구결과의 내용 및 정확성 측면에서 경제총조사 자료를 이용했을 경우와 차이를 보일 수도 있다. 하지만 본 연구의 진행 시점에서는 경제총조사 자료를 이용할 수 없었으므로 향후 보완 작업도 이루어져야 할 것이다. 최근 들어 통계조사의 환경이 계속적으로 악화되고 있어 각종 조사의 무응답 비율이 높아지고 있다. 이와 같은 현실에서 무응답 대체기법 연구는 매우 중요하며, 본 연구가 향후 경제총조사의 품질을 향상시키는데 많은 부분 도움이 되기를 기대한다.

참고문헌

김규성(2000), “무응답 대체 방법과 대체 효과”, 「조사연구」, 제1권 2호, pp.1-14.
김규성・이기재・김진(2005), “농어가경제조사에서 가중핫덱 무응답 대체방법의 활용”, 「응용통계연구」, 제18권 2호, pp.311-328.
김영원・이주원(2003), “CART를 활용한 결측값 대체방법: 인구주택총조사 혼인상태 항목을 중심으로”, 「조사연구」, 제4권 2호, pp.1-21.
김영원・조선경(1996), “표본조사에서 항목 무응답 대체 방법”, 「한국통계학회논문집」, 제3권 3호, pp.145-159.
김진(2004), “농가경제조사에 대한 대체법 비교”, 통계청,「통계연구」, 제9권 2호, pp.133-145.
송순관(2005), 「2005 인구주택총조사 무응답 처리방법 연구 및 읍면동 통계작성 가능성 검토」, 통계청 인구조사과.
이현정(2008), “인구주택총조사 무응답 처리기법 연구(Ⅰ)”, 연구보고서, 통계개발원.
이현정・최필근(2009), “인구주택총조사 무응답 처리기법 연구(Ⅱ)”, 연구보고서, 통계개발원.
최필근(2008), “농업총조사 항목간 연관성 분석 및 대체군(보조변수) 개발”, 연구보고서, 통계개발원.
최필근(2008), “농업총조사 무응답 대체기법 연구(Ⅰ)”, 연구보고서, 통계개발원.
최필근(2009), “농업총조사 무응답 대체기법 연구(Ⅱ)”, 연구보고서, 통계개발원.
최필근(2010), “출생전후기 사망통계의 무응답 대체기법”, 연구보고서, 통계개발원.
통계교육원(2005), 「무응답처리 실무론」.
_____(2009), 「무응답 자료처리 및 분석」.

Afifi, A. A. and R. M. Elashoff(1966), “Missing Observations in Multivariate Statistics Ⅰ: Review of the Literature", J. Am. Statist. Assoc., Vol. 61, pp.595-604.
Berry, M. J. A. and G. S. Linoff(1997), Data Mining Techniques, John Wiley & Sons, New York.
Kalton, G. and D. Kasprzyk(1986), “The Treatment of Missing Survey Data”, Survey Methodology, Vol. 12, pp.1-16.
Kass, G.(1980), “An Exploratory Technique for Investigating Large Quantities of Categorical Data”, Applied Statistics, Vol. 29, No. 2, pp.119-127.
Lessler and Kalsbeek(1992), Nonsampling Error in Surveys, John Wiley & Sons, New York.
Quinlan, J. R.(1986), “Induction of Decision Tree”, Machine Learning, 1, pp.81-106.
Rubin, D. B. and J. A. Little(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, pp.238-258.