제3장

출생전후기 사망통계의 무응답 대체기법

최 필 근

제1절 서 론

1. 연구배경 및 목적

UN 등의 국제기구에서는 보건수준의 측정지표로 출생전후기 사망통계 자료를 이용하고 있으며, 한국의 통계청에도 이와 관련된 자료를 계속적으로 요청하고 있다. 하지만 기존의 영아사망 자료는 출생 및 사망 신고가 동시에 누락되는 경우가 많고, 사산 자료는 신고체계의 미비로 수집이 어려워 정확한 출생전후기 사망통계 자료를 작성·제공하기가 어려웠다. 이러한 이유로 통계청에서는 정확한 출생전후기 사망통계를 작성하기 위한 일환으로 1999년부터 화장장의 영아사망 및 사산 신고 자료와 모자보건법의 신생아 및 사산 신고 자료를 수집하였고, 2001년부터는 내부적으로 출생전후기 사망 및 사산 통계의 시산을 실시하고 있다. 또한 2008년부터는 국가중장기통계발전계획의 일환으로 출생전후기 사망통계의 개발을 추진하고 있으며 2010년에는 통계결과의 공표 및 국제기구에 자료를 제공할 예정이다.

현재 통계청에서는 출생전후기의 기간을 OECD 및 UN에서 정의하고 있는 임신 28주에서 생후 7일 미만으로 사용하고 있다. 그러므로 출생전후기 사망통계를 작성하기 위해서 영아사망 자료와 사산 자료를 수집해야 하는데 이러한 자료들은 조사가 아닌 사망신고, 화장신고, 모자보건신고와 같은 다양한 행정자료를 통하여 획득하고 있다. 그러나 신고서식이 상이한 세 가지 자료를 이용하여 필요한 자료를 수집하기 때문에 이 자료들을 하나로 통합할 때 관심의 대상이 되는 주요 항목들에 대한 누락이 높은 비율로 발생하게 된다. 2007년 영아사망 자료의 경우 모연령은 50.9%, 출생아체중은 31.9%, 재태기간의 경우에는 36.5%의 누락이 발생하였다. 이처럼 누락의 비율이 높은 경우 수집된 자

료만으로 분석을 하는 것은 분석결과에 상당한 오류를 초래할 수 있다. 따라서 누락된 항목들의 특성을 파악하여 가장 적절한 값을 대체함으로써 자료의 품질을 향상시키는 노력이 이루어져야 할 것이다.

출생전후기 사망통계 자료의 누락된 부분은 일반 조사통계의 항목 무응답의 개념으로 생각할 수 있다. 무응답의 형태는 조사대상으로부터 얻은 정보가 전혀 없는 단위 무응답(unit nonresponse)과 특정 항목값의 정보가 없는 항목 무응답(item nonresponse)으로 나누어진다. 무응답의 처리는 발생형태별로 적절하게 이루어져야 하는데 단위 무응답의 경우는 주로 가중치 조정 방법을 사용하고, 본 연구의 출생전후기 사망통계 자료처럼 항목 무응답의 경우에는 적절한 값을 채워 넣기 위한 여러 가지 대체법을 이용하게 된다.

항목 무응답 대체연구는 크게 두 가지 과정으로 나눌 수 있다. 하나의 과정은 대체하고자 하는 항목에 대한 대체군(보조변수)을 결정하는 것이다. 이는 대체하고자 하는 항목과 연관성이 높은 항목들을 사용함으로써 대체의 정확도를 향상시키기 위함이다. 실제로 가장 적합한 대체기법을 적용하더라도 대체군의 선택이 잘못된다면 대체결과는 좋지 않을 것이라고 판단된다. 따라서 대체항목과 관련이 높은 항목들을 통계적 기법을 이용하여 찾아야 할 것이다. 다른 하나의 과정은 자료(항목)의 특성에 가장 적합한 대체기법의 개발 또는 적용이다. 지금까지 주로 사용되고 있는 대체방법들은 일반적인 사회조사처럼 관심의 대상이 되는 항목이 범주형일 경우에는 핫덱대체와 이와 유사한 최근방 기증자(donor) 대체방법이, 경제관련조사와 같이 연속형 항목들이 주를 이루는 경우는 회귀대체, 최근방대체, 평균대체 방법이 적용되고 있다. 하지만, 이러한 방법들은 조사자료의 특성을 잘 파악하고 적용해야 하며 더 정확한 대체가 가능하다면 기존의 방법들을 응용하든지 아니면 새로운 방법을 개발할 수도 있을 것이다.

본 연구에서는 출생전후기 사망통계 자료의 주요 항목에 대하여 적절한 대체방법을 제시하고자 함에 있다. 이 자료는 통계청의 조사통계 자료에 비해서 무응답 비율이 상당히 높기 때문에 다양한 모의실험을 통해서 대체전후의 변화를 검토하고 이를 근거로 적절한 대체기법을 선택해야 할 것이다. 본 연구를 통해서 체계적이며 정확도가 높은 출생전후기 사망통계의 무응답 대체기반을 마련하여 통계 공표 결과의 정확성 향상과 높은 품질의 자료를 제공하는데 기여하고자 한다.

2. 연구내용 및 방법

본 연구에서는 출생전후기 사망통계의 주요 항목의 무응답 대체를 위한 대체군 및 대체기법을 개발하고자 한다. 연구 자료는 2007년 출산전후기 사망통계 자료이며 이를 구성하고 있는 두 부문 중에서 영아사망 자료에서는 모연령, 출생아체중 및 재태기간 항

목을, 사산 자료에서는 모연령 및 사산아체중 항목에 대한 연구가 진행될 것이다.

각 항목에 대한 대체군은 의사결정나무 방법인 CHAID 알고리즘을 이용하여 연관성 분석을 실시하여 이를 바탕으로 결정할 것이다. 이전의 연구인 『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. 본 알고리즘의 상세한 내용은 본문에서 자세하게 소개될 것이다.

무응답을 대체하기 위한 방법은 몇 가지 대체기법을 검토를 하고 난 후 결정을 하고자 한다. 본 연구에서 대체하고자 하는 항목들은 모두 연속형 자료이므로 일반적으로 사용되는 회귀 대체방법을 사용할 수도 있으나, 자료의 특성과 대체전후의 결과를 검토한 후에 적용해야 할 것이다. 왜냐하면 회귀 대체방법은 목표변수에 대한 적절한 모형을 만들 수 있는 보조변수가 부족하다면 왜곡된 대체결과를 초래할 수도 있다. 따라서 본 연구에서는 회귀대체, 평균대체와 2008년도에 개발한 응용 핫덱 대체방법을 함께 검토할 것이다. 이를 위하여 다양한 모의실험을 실시하고 각 방법들에 대하여 대체전후의 평균차이, 추정량의 표준편차를 평균값으로 나누어준 변동계수(CV), 구성비 변화 등을 고려하여 출생전후기 사망통계 자료에 가장 적합한 대체기법을 제시할 것이다. 그리고 본 연구의 모든 결과를 이용하여 2008년 출생전후기 사망통계 자료에 적용할 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 모의실험을 실시하여 회귀 대체방법, 평균 대체방법, 응용 핫덱 대체방법을 비교·분석하고, 출생전후기 사망통계에 적용할 대체방법을 제시한다. 제5절에서는 선정된 대체방법을 2008년 출생전후기 사망통계 자료에 적용하고 그 결과를 기술한다. 마지막으로 제6절에서는 연구의 최종적인 결론을 요약하고 본 보고서를 마무리 하고자 한다.

제2절 CHAID 알고리즘과 연관성분석 결과

출생전후기 사망통계 자료의 무응답 대체를 위한 대체군은 CHAID 알고리즘을 이용하여 연관성분석을 실시하여 결정한다. 이 알고리즘은 항목들 간의 의미 있는 관계를 탐색하는데 효과적이라고 알려져 있다. 이 절에서는 CHAID 알고리즘을 소개하고 영아사망 자료의 모연령, 출생아체중 및 재태기간 항목과 사산 자료의 모연령 및 사산아체중 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. 그리고 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.

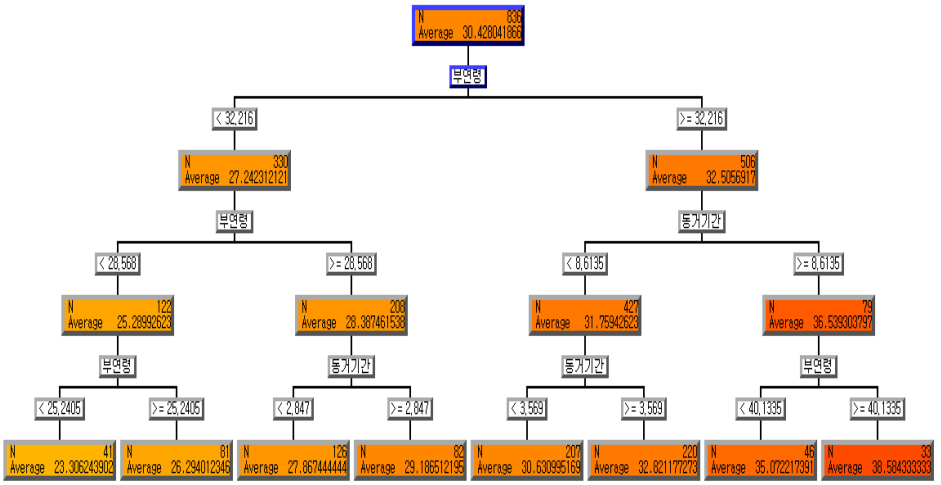
2. 각 항목에 대한 연관성분석 결과

출생전후기 사망통계 자료는 영아사망 자료와 사산 자료로 구성되어 있다. 본 연구에서는 구성하고 있는 총 항목들 중에서 사용 가능한 항목들을 선택하여 분석을 실시한다. 분석에 사용된 항목을 살펴보면 영아사망 자료의 경우에는 시도, 생존기간, 부학력, 부연령, 모학력, 모연령, 재태기간, 출생아체중, 총출산아수, 동거기간 10개 항목이며, 사산 자료의 경우에는 시도, 재태기간, 모연령, 사산아체중 4개 항목이다. 이를 바탕으로 각 항목에 대한 연관성분석의 결과를 제시하고 무응답 대체에 사용할 최종 대체군을 결정하고자 한다.

가. 영아사망 자료

1) 모(母)연령

모연령에 대한 연관성 분석의 결과를 [그림 3-1]과 <표 3-1>을 이용하여 설명하고자 한다. 나머지 항목들도 설명방식이 같기 때문에 하나의 경우에 대해서만 자세하게 설명하고, 나머지 항목들은 약식으로 설명을 할 것이다. 그리고 <표 3-1>에서 분리변수의 괄호 안의 숫자는 각 가지에서의 마디 번호를 의미한다.



[그림 3-1] 모연령에 관한 연관성 모형(1차)

〈표 3-1〉 모연령에 관한 연관성 분석의 세부내용(1차 모형)

모연령			
깊이(연관성)	분리변수	분리지점(좌)	분리지점(우)
1	부연령	32.22(세)미만	32.22(세)이상
2	부연령(1) 동거기간(2)	28.57(세)미만 8.61(년)미만	28.57(세)이상 8.61(년)이상
3	부연령(1) 동거기간(2) 동거기간(3) 부연령(4)	25.24(세)미만 2.85(년)미만 3.57(년)미만 40.13(세)미만	25.24(세)이상 2.85(년)이상 3.57(년)이상 40.13(세)이상

1

가장 연관성이 높은 항목을 의미한다. 모연령을 분리하는데 가장 연관성이 높은 항목은 부연령이다. 전체 모연령의 평균은 30.43세이며 이를 분리하여 나간다.

좌

부연령에 의하여 왼쪽으로 분리되는 것을 의미한다. 부연령이 32.22세 미만인 경우는 좌로 분리되고, 이때의 모연령의 평균은 27.24세로 전체 평균에 비하여 조금 낮아지는 것을 알 수 있다.

우

부연령에 의하여 오른쪽으로 분리되는 것을 의미한다. 부연령이 32.22세 이상인 경우는 우로 분리되고, 이때의 모연령의 평균은 32.51세로 전체 평균에 비하여 조금 높아지는 것을 알 수 있다.

2(1)

첫 번째 분리가 끝난 후 두 번째 분리가 시작된다. 이 과정은 첫 번째 분리가 된 왼쪽 부분을 다시 세부적으로 분리한다. 부연령 항목으로 다시 분리가 된다. 부연령이 28.57세 미만이면 왼쪽으로 분리되고 이때의 모연령의 평균은 25.29세가 되며 부연령이 28.57이상 32.22세 미만이면 오른쪽으로 분리되고 이때의 모연령의 평균은 28.39세가 된다.

2(2)

이 과정은 첫 번째 분리가 된 오른쪽 부분을 다시 세부적으로 분리한다. 두 번째로 연관성이 높은 동거기간 항목으로 다시 분리가 된다. 부연령이 32.22세 이상 중에서 동거기간이 8.61년 미만이면 왼쪽으로 분리되고 이때의 모연령의 평균은 31.76세가 되며 동거기간이 8.61년 이상이면 오른쪽으로 분리되고 이때의 모연령의 평균은 36.54세가 된다.

3(1)

두 번째 분리가 끝난 후 세 번째 분리가 시작된다. 이 과정은 두 번째 분리가 된 처음 부분을 다시 세부적으로 분리한다. 부연령 항목으로 다시 분리가 된다. 부연령이 25.24세 미만이면 왼쪽으로 분리되고 이때의 모연령의 평균은 23.31세가 되며 부연령이 25.24세 이상 28.57세 미만이면 오른쪽으로 분리되고 이때의 모연령의 평균은 26.29세가 된다.

3(2)

이 과정은 두 번째 분리가 된 두 번째 부분을 다시 세부적으로 분리한다. 동거기간 항목으로 다시 분리가 된다. 부연령이 28.57세 이상 32.22세 미만 중에서 동거기간이 2.85년 미만이면 왼쪽으로 분리되고 이때의 모연령의 평균은 27.87세가 되며 동거기간이 2.85년 이상이면 오른쪽으로 분리되고 이때의 모연령의 평균은 29.16세가 된다.

3(3)

이 과정은 두 번째 분리가 된 세 번째 부분을 다시 세부적으로 분리한다. 동거기간 항목으로 다시 분리가 된다. 부연령이 32.22세 이상 중에서 동거기간이 3.57년 미만이면 왼쪽으로 분리되고 이때의 모연령의 평균은 30.63세가 되며 동거기간이 3.57년 이상 8.61년 미만이면 오른쪽으로 분리되고 이때의 모연령의 평균은 32.82세가 된다.

3(4)

이 과정은 두 번째 분리가 된 네 번째 부분을 다시 세부적으로 분리한다. 부연령 항목으로 다시 분리가 된다. 동거기간이 8.61년 이상 중에서 부연령이 32.22세 이상 40.13세 미만이면 왼쪽으로 분리되고 이때의 모연령의 평균은 35.07세가 되며 부연령이 40.13세 이상이면 오른쪽으로 분리되고 이때의 모연령의 평균은 38.58세가 된다.

이상으로 모연령 항목을 분리하는 과정을 자세히 설명하였다. 요약하면, 부연령과 동

거기간 항목이 모연령과 상당히 높은 연관성이 있다는 것을 알 수 있다. 하지만 2007년 영아사망 자료에서는 모연령을 대체하기 위해서 부연령과 동거기간 항목을 사용할 수가 없다. 왜냐하면 모연령이 누락된 데이터 대부분은 부연령과 동거기간도 함께 누락이 되어 있기 때문이다. 그러므로 모연령의 대체군을 결정하기 위해서는 이 항목을 제외하고 추가 분석을 실시해야만 한다. 그리고 현재의 대체군은 자료의 수집과정에서 획득할 수 있을 경우에 사용한다면 더 좋은 대체가 될 수 있을 것이다. <표 3-2>에는 모연령의 대체시 사용할 수 없는 부연령과 동거기간 항목을 제외하고 분석한 결과가 주어져 있다.

〈표 3-2〉 모연령에 관한 연관성 분석의 세부내용(2차 모형)

모연령			
깊이(연관성)	분리변수	분리지점(좌)	분리지점(우)
1	총출산아수	1.5(명)미만	1.5(명)이상
2	모학력(1) 부학력(2)	2, 3, 4, 9 3, 4	7 2, 7, 9
3	생존기간(1) 부학력(2) 총출산아수(3) 총출산아수(4)	180.5(일)미만 4 2.5(명)미만 2.5(명)미만	180.5(일)이상 3, 7, 9 2.5(명)이상 2.5(명)이상

부연령과 동거기간 항목을 사용할 수 없는 경우에 모연령과 가장 연관성이 높은 항목은 총출산아수 항목이다. 총출산아수가 1명, 2명, 또는 3명 이상의 여부에 따라서 모연령의 차이가 존재한다는 것이다. 다음으로 모학력이나 부학력도 모연령과 연관성이 존재하고 있음을 알 수 있다. 모학력이 7(대졸이상)인 경우가 2(초등학교졸), 3(중학교졸), 4(고등학교졸)인 경우보다 모연령의 평균이 2.5세 정도 높으며, 부학력도 이와 비슷한 현상을 보인다. 마지막으로 영아의 생존기간도 모연령과 연관성이 있어 보이는데 생존기간이 180.5일보다 많은 경우와 적은 경우는 모연령의 평균이 3세 이상의 차이가 나타남을 알 수 있다. 따라서 부연령 및 동거기간을 사용할 수 없는 경우에 총출산아수, 모학력, 부학력 및 생존기간 항목을 이용하면 1차 모형보다는 다소 정확도가 떨어지나 좋은 대체가 가능할 것으로 판단된다.

그러나 2007년 영아사망 자료에서는 이전의 1차 모형과 유사하게 생존기간을 제외한 3개의 항목은 모연령과의 동시 누락으로 사용되지 못한다. 따라서 총출산아수, 모학력, 부학력을 대신할 정보가 있는지 이를 제외하고 다시 분석을 실시하였다.

〈표 3-3〉 모연령에 관한 연관성 분석의 세부내용(3차 모형)

모연령			
깊이(연관성)	분리변수	분리지점(좌)	분리지점(우)
1	생존기간	173.5(일)미만	173.5(일)이상
2	출생아체중(1)	2.83(kg)미만	2.83(kg)이상

<표 3-3>에는 1, 2차 모형에서 모연령의 대체군으로 사용할 수 없는 항목들을 제외하고 분석한 결과가 주어져 있다. 모연령과 연관성이 높은 대부분의 항목이 제외되었으므로 3차 모형은 매우 간단하게 나타난다. 2차 모형에서 선택되었던 생존기간에서 출생아체중의 항목이 추가되었다. 출생아체중 항목은 모연령과의 연관성이 매우 높은 것은 아니지만 대체를 위한 정보는 분명히 있다는 것을 알 수 있다. 출생아체중 항목도 누락이 많기 때문에 대체군으로 사용하기 위해서는 사전에 모든 누락부분이 대체가 되어야 할 것이다.

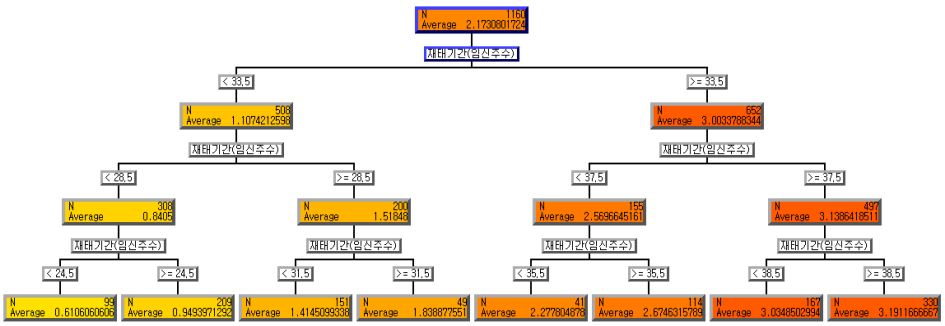
따라서 모연령의 대체군으로는 생존기간과 출생아체중 항목을 이용할 것이며, 출생아체중 항목을 대체하고 난후에 대체를 실시하여 보조정보의 활용을 최대한 이용하고자 한다. <표 3-4>에 모연령 항목에 대한 대체군의 구축과정을 정리하였다.

〈표 3-4〉 모연령 항목의 대체군

1차 대체군	부연령, 동거기간
2차 대체군	총출산아수, 모학력, 부학력, **생존기간**
3차 대체군	**생존기간, 출생아체중**

2) 출생아체중

출생시체중 항목도 모연령 항목과 같은 방법으로 대체군을 결정하고자 한다. 사용가능한 모든 항목들에 대한 연관성분석의 결과는 [그림 3-2]와 <표 3-5>에 주어져 있다.



[그림 3-2] 출생아체중에 관한 연관성 모형(1차)

〈표 3-5〉 출생아체중에 관한 연관성 분석의 세부내용(1차 모형)

출생아체중			
깊이(연관성)	분리변수	분리지점(좌)	분리지점(우)
1	재태기간	33.5(주)미만	33.5(주)이상
2	재태기간(1) 재태기간(2)	28.5(주)미만 37.5(주)미만	28.5(주)이상 37.5(주)이상
3	재태기간(1) 재태기간(2) 재태기간(3) 재태기간(4)	24.5(주)미만 31.5(주)미만 35.5(주)미만 38.5(주)미만	24.5(주)이상 31.5(주)이상 35.5(주)이상 38.5(주)이상

출생아체중 항목은 재태기간 항목과 매우 높은 연관성을 가진다. 즉, 재태기간을 알 수 있으면 출생아체중을 정확도 높게 추정을 할 수 있다는 것이다. 분석결과를 보면 재태기간이 24.5주 미만인 출생아의 체중은 평균 0.62kg이며, 24.5주 이상 28.5주 미만은 0.95kg, 28.5주 이상 31.5주 미만은 1.41kg, 31.5주 이상 33.5주 미만은 1.84kg, 33.5주 이상 35.5주 미만은 2.28kg, 35.5주 이상 37.5주 미만은 2.67kg, 37.5주 이상 38.5주 미만은 3.03kg, 38.5주 이상은 3.19kg의 평균 체중을 가짐을 알 수 있다. 하지만 출생아체중 항목도 모연령 항목처럼 연관성이 매우 높은 재태기간 항목을 사용할 수가 없기 때문에 2차 대체군을 다시 결정해야 한다. 이로써 대체의 정확도는 다소 떨어질 수 있다는 것을 알려둔다. <표 3-6>에는 출생아체중의 대체시 사용할 수 없는 재태기간 항목을 제외하고 분석한 결과가 주어져 있다.

〈표 3-6〉 출생아체중에 관한 연관성 분석의 세부내용(2차 모형)

출생아체중			
깊이(연관성)	분리변수	분리지점(좌)	분리지점(우)
1	생존기간	46.5(일)미만	46.5(일)이상
2	동거기간(1) 생존기간(2)	2.5(년)미만 90.5(일)미만	2.5(년)이상 90.5(일)이상
3	생존기간(2) 모연령(3)	0.5(일)미만 26.18(세)미만	0.5(일)이상 26.18(세)이상

재태기간 항목을 사용할 수 없는 경우에 출생아체중과 가장 연관성이 높은 항목은 생존기간이다. 생존기간이 긴 경우가 출생시의 체중이 클 가능성이 많다는 것과 동거기간이나 모연령도 출생아체중과 연관성이 존재하고 있음을 알 수 있다.

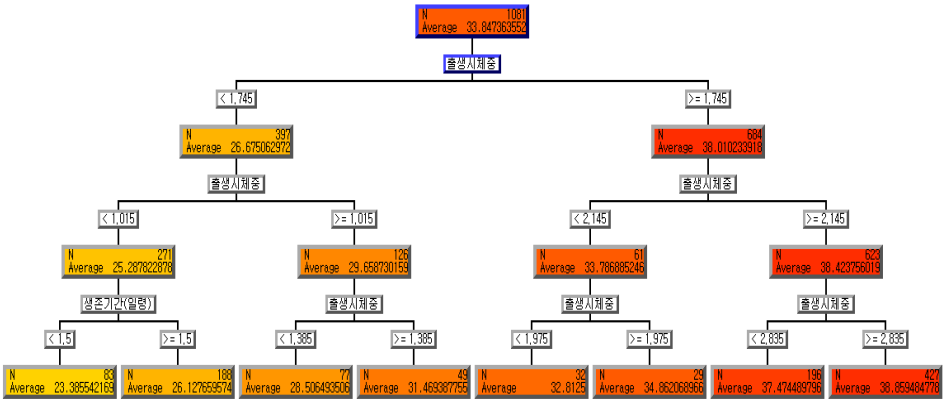
〈표 3-7〉 출생아체중 항목의 대체군

1차 대체군	재태기간
2차 대체군	**생존기간**, 동거기간, 모연령
3차 대체군	**생존기간**

그러나 동거기간과 모연령은 출생아체중과 동시에 누락이 되며, 더 이상 대체에 이용할 수 있는 항목이 없으므로 출생아체중의 대체군으로는 생존기간 하나의 항목만을 이용할 것이다. <표 3-7>을 참조하기 바란다.

3) 재태기간

재태기간 항목에 대한 연관성분석의 결과는 [그림 3-3]와 <표 3-8>에 주어져 있다.



[그림 3-3] 재태기간에 관한 연관성 모형

〈표 3-8〉 재태기간에 관한 연관성 분석의 세부내용

재태기간			
깊이(연관성)	분리변수	분리지점(좌)	분리지점(우)
1	출생아체중	1.75(kg)미만	1.75(kg)이상
2	출생아체중(1)	1.02(kg)미만	1.02(kg)이상
	출생아체중(2)	2.15(kg)미만	2.15(kg)이상
3	생존기간(1)	1.5(일)미만	1.5(일)이상
	출생아체중(2)	1.39(kg)미만	1.39(kg)이상
	출생아체중(3)	1.98(kg)미만	1.98(kg)이상
	출생아체중(4)	2.84(kg)미만	2.84(kg)이상

재태기간 항목은 출생아체중 항목과 매우 높은 연관성을 가진다. 즉, 출생아체중의 정보를 이용하면 재태기간을 정확도 높게 추정을 할 수 있다는 것이다. 다음으로 생존기간과도 연관성이 존재함을 알 수 있다. 분석결과를 보면 출생아체중이 1.02kg 미만인 경우 재태기간의 평균은 25.29주이며, 1.02kg 이상 1.39kg 미만은 28.51주, 1.39kg 이상 1.75kg 미만은 31.47주, 1.75kg 이상 1.98kg 미만은 32.81주, 1.98kg 이상 2.15kg 미만은 34.86주, 2.15kg 이상 2.84kg 미만은 37.47주, 2.84kg 이상은 38.86주의 평균 재태기간을 가짐을 알 수 있다. 하지만 재태기간 항목을 대체할 때 출생아체중의 정보를 사용할 수 있는 경우는 10분의 1 정도밖에 되지 않는다. 따라서 모연령의 경우와 마찬가지로 출생아체중의 누락된 부분이 대체가 된 후에 재태기간 항목의 대체를 실시할 것이다. 결론적으로 재태기간 항목의 대체군은 출생아체중과 생존기간으로 결정하고자 한다.

나. 사산 자료

1) 모(母)연령

사산자료의 모연령 항목에 대한 연관성분석의 결과는 <표 3-9>에 주어져 있다. 사산자료의 경우 보조정보가 될 수 있는 항목이 거의 없기 때문에 모형이 단순하며, 모연령의 경우 재태기간과의 연관성이 높지 않지만 부연령, 동거기간, 총출산아수 등 연관성이 높은 항목들을 사용할 수 없으므로 재태기간 항목만을 이용해야 할 것으로 판단된다.

〈표 3-9〉 모연령에 관한 연관성 분석의 세부내용(사산)

모연령(사산)			
깊이(연관성)	분리변수	분리지점(좌)	분리지점(우)
1	재태기간	31.5(주)미만	31.5(주)이상

2) 사산아체중

사산아체중 항목의 경우에는 누락된 비율이 99%에 이른다. 따라서 연관성분석 및 대체를 위한 과정에서 주어진 자료의 특성을 신중하게 검토하여야 할 것이다. <표 3-10>에 사산아체중 항목의 연관성분석 결과가 주어져있다. 분석결과를 살펴보면 재태기간이 19.5주 미만인 경우 사산아체중의 평균은 186.71g이며, 19.5주 이상 21.5주 미만은 371.71g, 21.5주 이상 31.5주 미만은 692.65g, 31.5주 이상은 1889.19g의 평균 체중을 가짐을 알 수 있다. 사산자료의 모연령과 같이 재태기간 하나의 항목이 선택되었지만 사산아

체중과 재태기간은 매우 높은 연관성을 가지므로 대체가능성을 고려할 때 긍정적인 결과라고 할 수 있다. 그러나 상당히 높은 누락으로 인한 손실은 충분히 감수해야 할 것이며 본 연구자는 1% 자료의 패턴이 전체자료와 유사하기를 바라며 이후의 연구를 진행할 것이다.

〈표 3-10〉 사산아체중에 관한 연관성 분석의 세부내용

사산아체중			
깊이(연관성)	분리변수	분리지점(좌)	분리지점(우)
1	재태기간	31.5(주)미만	31.5(주)이상
2	재태기간(1)	21.5(주)미만	21.5(주)이상
3	재태기간(1)	19.5(주)미만	19.5(주)이상

이상으로 연관성분석을 통하여 각 항목에 대한 대체군을 제시하였다. 2007년 출생전후기 사망통계 자료의 경우 주요 항목들이 동시에 누락이 되는 경우가 대부분이므로 좋은 보조정보의 사용이 제한되고 있다. 하지만 제시된 대체군은 현재 사용가능한 가장 적절한 보조정보라고 말할 수 있으며 향후 신고서식의 항목개정 등을 통하여 양질의 자료를 얻을 수 있게 되었으면 하는 바람이다.

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. 이 절에서는 일반적으로 쓰이고 있는 단일 대체방법에 대하여 간략하게 소개하고자 한다.

1. 결정적 대체방법

결정적 대체방법에서의 대체값은 실제 조사된 응답값 중의 하나를 선택할 수도 있고, 다른 보조변수를 이용하여 대체값을 만들어 적용할 수도 있다. 이때 대체되는 목표변수의 값이 연속형인 경우는 어느 방법을 이용해도 무방하지만 범주형인 경우에는 최종 대

체값이 이산형이 되도록 보정을 해주어야 하는 과정이 더 필요하다.

가. 연역적 대체방법(deductive imputation)

논리적인 제약조건이나 다른 기록에 의하여 확실하게 대체값을 지정하여 무응답을 대체하는 방법이다. 이 방법은 내검과정에서 응답값이 논리적으로 타당하지 않을 때에나 합계를 구성하고 있는 하나의 항목이 무응답일 경우 주로 이용되어진다.

나. 시기적 대체방법(historical imputation)

반복조사(repeated survey)에서 매우 유용하게 사용되는데, 만일 동일 항목의 응답값이 조사시점에 따라 안정된 값을 보이고 전회 조사값과 금회 조사값의 상관관계가 높으면 금회 조사의 무응답 항목에 전회 조사의 응답값을 대체하는 방법이다. 이 방법은 특히 응답패턴으로 인한 편향의 영향을 받지 않는 장점이 있다.

다. 평균 대체방법(mean imputation)

무응답 항목에 목표변수의 전체 평균을 대입하거나 또는 대체군 내의 평균을 대입하는 방법이다. 즉,

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

여기서 R은 응답 집합, R^c은 무응답 집합이며 $\bar{y}_r$은 응답값의 평균이다. 이 방법은 간단하여 이용되기 쉬운 장점이 있으며, 항목이 양적 변수이고 구하고자 하는 통계량이 평균일 때 유용하다. 그러나 대체 후의 값들은 평균값의 빈도수가 지나치게 많아져 응답값들의 분포가 왜곡되고, 중위수나 백분위수와 같은 평균이 아닌 통계량을 구할 때에는 효율이 저하되는 단점이 있다.

라. 축차 핫덱 대체방법(sequential hot-deck imputation)

데이터 파일을 구성할 때 직전에 응답한 단위의 항목값으로 대체하는 방법이다. 즉,

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

이 방법은 사회·인구통계조사에서 유용하다. 표본은 사회·인구통계적 지표에 의해서 자연스럽게 대체군으로 구분되며, 대체군 내의 항목값은 서로 유사할 가능성이 높다.

또한 이러한 조사는 지리적인 연속성을 가지므로 무응답이 발생하면 직전의 응답값으로 대체하는 것이 타당할 것이다. 하지만 동일한 값을 여러 번 사용하게 될 수 있다는 위험성이 있으며, 결측값을 할당하는데 있어서 확률구조가 아닌 자료파일 순서에 의존한다는 단점이 있다.

마. 비 대체방법(ratio imputation)

이용 가능한 보조변수가 있을 때 목표변수와 보조변수의 관계를 이용하여 무응답을 대체하는 방법이다. 만일 목표변수와 보조변수가 비례관계이면 다음과 같다.

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

이 방법은 목표변수와 보조변수가 원점을 지나는 직선관계이며, 분산이 보조변수에 비례하는 경우 효과적인 것으로 나타났다. 유한 모집단에서 예측이론에 의하면 비 대체방법이 매우 우수한 것으로 알려져 있으나 항목들이 양적 변수일 때에만 사용가능하다.

바. 회귀 대체방법(regression imputation)

비 대체방법과 유사하게 사용되며, 목표변수와 보조변수의 관계가 절편이 있는 직선관계이고 목표변수의 분산이 동일할 때 유용한 방법이다. 즉,

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

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. 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.

사. 최근방 대체방법(nearest-neighbor imputation)

보조변수를 이용하여 응답하지 않은 개체와 가장 유사한 응답개체를 찾아 대응되는 항목값을 대체하는 방법으로 현실적으로 이용 가능성이 높은 방법이다. 즉,

$$y_k^* = \begin{cases} y_k\,, & k \in R, \quad \text{응답값} \\ y_i\,,\ i : \min_{i \in R} \| x_k - x_i \|\,, & k \in R^c, \quad \text{대체값} \end{cases}$$

여기서 거리가 가장 가까운 응답값을 찾는 방법은 여러 가지가 있으며, 주로 절대거리를 많이 이용한다. 이 방법은 활용 가능성은 높지만 주어진 보조변수가 적절치 않거나 무응답률이 높은 경우 추정에 큰 편향을 가져올 수 있다.

2. 확률적 대체방법

결정적 대체방법은 대체값을 유일하게 결정하기 때문에 목표변수의 변동을 줄이는 경향이 있다. 이러한 단점을 보완하기 위하여 확률적 대체방법이 제안되었으며, 이 방법은 대체값에 확률적인 변동을 부여해 줌으로써 결정적 대체방법의 단점을 보완할 수 있다.

가. 랜덤 핫덱 대체방법(random hot-deck imputation)

대체군 내의 응답값 중에서 하나를 임의로 선택하여 대체하는 방법이다. 즉,

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

랜덤 핫덱 대체를 이용하면 무응답 대체 후에도 표본의 분포가 그대로 유지될 수 있다는 장점이 있다. 또한 평균 대체나 회귀 대체 등과는 달리 표본 분포가 유지되므로 통계량의 형태에 무관하게 이용될 수 있다. 그러나 응답 패턴이 목표변수와 무관한 경우에 더 적합하며 경제관련 조사보다는 사회관련 조사에 주로 사용되어진다.

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

랜덤 핫덱 방법을 응용한 것으로 연속형 항목에도 자유롭게 적용할 수 있게 한 방법이다. 대체군의 적용시 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. 이 방법은 농업총조사 무응답 대체에 적용하기 위해 개발한 것으로 다른 조사

의 무응답 대체에도 이용가능하다. 이 방법에 대한 자세한 설명은 2008년에 진행이 되었던 『농업총조사 무응답 대체기법 연구(Ⅰ)(최필근)』의 보고서에 있으니 참조하기 바란다.

다. 가중 핫덱 대체방법(weighted hot-deck imputation)

랜덤 핫덱 방법과 유사하나 응답값 중에서 하나를 선정할 때 가중값을 두어 선정하는 방법이다. 이 방법은 복합표본에서 층화나 집락화로 인하여 서로 다른 추출확률에 의하여 표본이 선정될 때 랜덤 핫덱 방법 대신에 주로 이용된다.

라. 랜덤 비 대체방법(random ratio imputation)

기존의 비 대체값에 확률오차를 포함시켜 대체값으로 이용하는 방법이다.

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

여기에 더해주는 확률오차 e_k는 대체로 인한 변동의 감소를 보정해 주는 효과가 있으며, 그 크기는 대체 전 변동과 대체 후 변동이 동일하게 되도록 변동의 폭을 계산하여 포함시킨다. 랜덤 회귀 대체방법 역시 같은 개념을 사용하여 대체를 실시한다.

본 절에서 설명한 대체방법들은 다양한 조사에서 사용되고 있는 방법들이다. 이러한 무응답 대체방법들은 자료의 특성(형태)에 따라서 적절하게 적용되어진다. 연속형 자료의 경우에는 회귀대체, 최근방대체, 평균대체 등의 방법을 사용하고, 범주형인 경우에는 확률적 대체방법이나 핫덱 대체방법 등을 주로 이용하게 된다. 출생전후기 사망통계 자료의 경우 대체할 항목과 대체군 항목 모두가 연속형 항목이므로 이 경우에 가장 많이 사용되는 회귀 대체방법과 사용하기 간편한 평균 대체방법 그리고 본 연구자가 개발한 응용 핫덱 대체방법 등으로 대체를 실시하면 될 것이다. 하지만 사용되는 방법들이 출생전후기 사망통계 자료에 얼마나 적합한지 검토가 필요하다. 따라서 4절에서는 이 방법들을 이용하여 대체를 실시하는 모의실험을 하고자 한다. 다양한 모의실험을 통하여 가장 적절한 대체방법을 선택할 것이며 향후 출생전후기 사망통계 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 2007년 출생전후기 사망통계자료이며 실험에 사용한 방법은 간단하게 사용할 수 있는 평균 대체방법, 연속형 항목에 가장 많이 이용되는 회귀 대체방법, 연속형과 범주형 항목 모두에 사용할 수 있게 개발된 응용 핫덱 대체방법이다. 데이터 파일은 각 항목별로 누락이 되지 않은 자료를 모아 새롭게 구성하였으며, 모의실험을 위해서 10%, 20%, 30%, 40%, 50%의 결측치를 임의로 발생을 시켰다. 단, 사산자료의 모연령은 누락된 부분이 작기 때문에 5%, 10%, 15%, 20%, 25%의 비율로 결측치를 발생시켰으며 각 실험마다 1,000을 반복하여 추정량을 계산하였다.

각 항목의 평균추정에 대한 편향정도를 알아보기 위하여 결측 전의 평균값과 3가지 방법에 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. 영아사망 자료

1) 출생아체중

연관성분석의 결과 출생아체중 항목은 생존기간 항목만을 대체군으로 사용해야 한다. 그리고 모연령과 재태기간 항목이 출생아체중 항목을 대체군으로 사용하기 때문에 출생아체중을 가장 먼저 대체하여 이후에 이 정보를 이용하도록 할 것이다. 출생아체중 항목은 총 영아사망 자료 1,703명 중에서 1,160명의 자료가 있어 대략 30%의 누락(무응

답)이 발생하고 있다. 따라서 총 1,160명의 자료를 모집단이라 가정하여 출생아체중에 대한 무응답 대체 실험을 실시하였다.

〈표 3-11〉 출생아체중 항목의 모의실험 결과(mean=2.1731 kg)

무응답 비율	통계량	대체방법		
		평균 대체	회귀 대체	응용 핫덱 대체
10%	MAE /오차비율	**0.0078**(0.35%)	**0.0079**(0.36%)	**0.0107**(0.49%)
	CV	0.0045	0.0045	0.0061
20%	MAE	**0.0120**(0.55%)	**0.0122**(0.56%)	**0.0167**(0.77%)
	CV	0.0068	0.0070	0.0097
30%	MAE	**0.0154**(0.71%)	**0.0156**(0.72%)	**0.0214**(0.98%)
	CV	0.0089	0.0090	0.0124
40%	MAE	**0.0190**(0.87%)	**0.0195**(0.90%)	**0.0267**(1.23%)
	CV	0.0108	0.0111	0.0153
50%	MAE	**0.0247**(1.14%)	**0.0255**(1.17%)	**0.0329**(1.51%)
	CV	0.0141	0.0145	0.0189

출생아체중에 대한 종합적인 대체결과가 <표 3-11>에 정리되어 있다. 세 가지 방법 모두가 평균을 추정하는데 있어 매우 좋은 대체가 되었다는 것을 알 수 있다. 오차의 비율은 무응답 비율이 높아질수록 같이 높아지는 것을 볼 수 있으나 상당히 작다는 것을 알 수 있다. 실제자료의 총 누락비율이 약 30%이므로 무응답 비율이 30%인 모의실험에 주목할 필요가 있다. 평균, 회귀, 응용 핫덱 대체방법의 오차비율은 0.71%, 0.72%, 0.98%로 1% 미만의 오차이므로 평균추정에 대해서는 세 방법 모두 좋은 대체결과를 보여주고 있다. 응용 핫덱 대체방법의 오차가 조금 더 크지만 큰 차이는 아닌 것으로 판단된다. 그리고 추정량의 안정성을 측정하기 위한 CV값 역시 매우 작은 값을 나타내므로 세 방법 모두 출생아체중의 평균값을 추정한 결과는 믿을 수 있을 것으로 생각된다.

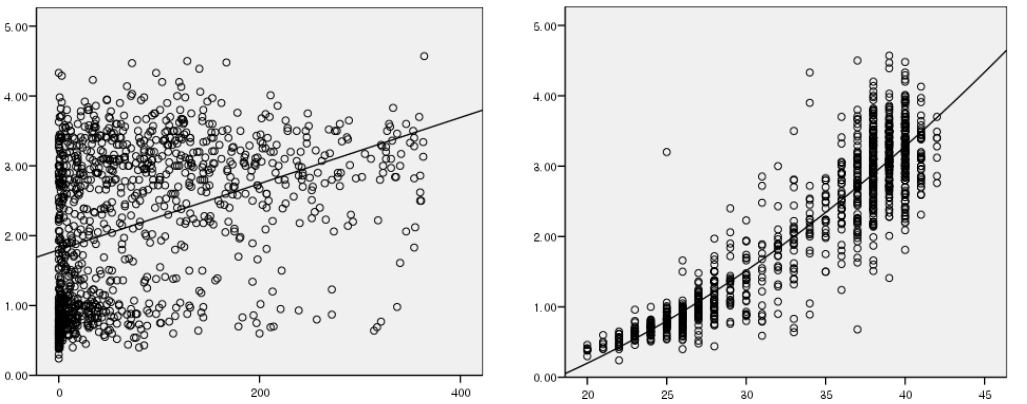
다음으로 대체를 하고 난 후의 구성비(분포)의 변화를 살펴보고자 한다. 대체를 한 후에 실제 구성비와 차이가 적어야 좋은 대체라고 말 할 수 있다. 평균 추정에서 정확한 대체가 된 경우에도 구성비 추정에서 상당히 왜곡되는 경우가 발생할 수 있다. 따라서 본 실험에서는 구성비 변화를 동시에 고려하였다. <표 3-12>에는 세 방법에 대해서 무응답 비율이 30%인 경우에 대체전후의 구성비 변화의 결과가 주어져 있다.

〈표 3-12〉 출생아체중의 대체전후의 구성비 변화(무응답비율 30%인 경우)

범주(kg)	1미만	1 - 2	2 - 3	3 - 3.5	3.5이상
실제 인원수	296	196	293	246	129
구성비	**25.5%**	**16.9%**	**25.3%**	**21.2%**	**11.1%**
평균 대체방법					
대체후 인원수	211	279	414	173	83
구성비	**18.2%**	**24.0%**	**35.7%**	**14.9%**	**7.2%**
회귀 대체방법					
대체후 인원수	211	312	355	196	86
구성비	**18.2%**	**26.9%**	**30.6%**	**16.9%**	**7.4%**
응용 핫덱 대체방법					
대체후 인원수	305	199	291	240	125
구성비	**26.3%**	**17.1%**	**25.1%**	**20.7%**	**10.8%**

총 1,160명의 출생아체중 구성비를 살펴보면 1kg미만은 25.5%, 1-2kg은 16.9%, 2-3kg은 25.3%, 3-3.5kg은 21.2%, 3.5kg이상은 11.1%이다. 30%의 임의 결측값을 대체하고 난 후의 구성비는 세 방법의 차이가 상당히 큰 것을 볼 수 있다. 평균 및 회귀 대체방법은 대체 후의 구성비가 상당히 왜곡이 된 반면에 응용 핫덱 대체방법은 구성비의 변화가 거의 일어나지 않는 것으로 보아 두 방법에 비해서 출생아체중 항목의 대체에 매우 적합한 방법으로 판단된다. 실제 평균 대체의 경우에는 대체군의 각 범주에 해당하는 평균값을 사용하기 때문에 평균의 추정에는 좋으나 구성비를 왜곡시킬 가능성이 많은 방법으로 알려져 있다. 따라서 이 방법은 평균의 추정값만을 사용할 때는 편리하게 이용할 수 있지만 그 외에는 사용에 앞서 구성비 문제를 확인해야 할 것이다.

회귀 대체방법은 모형이 잘 적합된 경우에는 평균 및 구성비 추정에 있어서 좋은 대체결과를 보여준다. 본 실험에서 큰 구성비의 변화를 가져온 이유는 모형구축에 사용된 보조변수인 생존기간 항목이 출생아체중 항목과의 상관관계가 높지 못해 발생한 것으로 판단된다. 이는 연관성분석의 결과 재태기간 항목이 출생아체중과 더 높은 연관성을 보이나, 이 정보를 사용할 수 없어 생존기간 항목을 대신 이용했기 때문에 회귀모형을 구축하는데 어려움이 따른다고 할 수 있다. [그림 3-4]를 참조하면 자세히 알 수 있을 것이다.



[그림 3-4] (출생아체중 vs. 생존기간) (출생아체중 vs. 재태기간) 산점도

첫 번째 그림은 출생아체중과 현재 사용되고 있는 보조변수 생존기간과의 산점도를 나타낸다. 두 항목 간의 선형관계가 크지 않으므로 회귀모형에 의한 추정결과가 평균 대체방법과 유사하다는 것을 알 수 있다. 그러므로 평균 추정 및 구성비 추정에서 평균 대체방법과 큰 차이가 나지 않으며, 구성비가 평균 근처로 집중되어 실제 구성비를 왜곡하는 현상이 발생됨을 보여준다. 반면에 두 번째 그림은 연관성분석에서 출생아체중과 연관도가 높은 재태기간 항목과의 산점도이다. 앞의 그림과 비교할 때 훨씬 더 좋은 모형을 구축할 수 있을 것으로 판단된다. 재태기간이 증가함에 따라서 출생아체중도 증가하는 추세가 있어 구성비 측면에서의 변화도 줄어들 것으로 생각되지만, 산점도의 폭이 다소 넓어 구성비의 변화는 있을 것으로 판단된다. 그리고 이전에 설명하였듯이 현재의 출생전후기 사망통계 자료는 출생아체중 항목의 대체에 재태기간 항목을 사용할 수 없으므로 회귀 대체방법으로는 본 실험의 결과 이상을 기대할 수는 없을 것이다. 따라서 회귀 대체값에 확률오차를 포함시켜 대체로 인한 변동의 크기를 보정해주는 랜덤 회귀 대체방법과의 비교실험을 추가로 실시하고자 한다.

<표 3-13>과 <표 3-14>에는 랜덤 회귀 대체방법을 사용했을 경우의 대체전후의 평균 및 구성비 변화의 결과가 주어져있다. 회귀 대체의 결과와는 다소 변화가 있음을 알 수 있다. 회귀 대체값에 확률오차가 포함됨으로써 평균의 추정에 있어서는 오차비율이 높아져 응용 핫덱 대체방법의 결과와 거의 유사하게 되었음을 알 수 있다. 이러한 결과는 적합된 회귀선에서 오차의 범위에서 임의로 값을 선택하기 때문에 평균 추정에서는 다소 오차비율이 커질 수 있음을 보여준다.

〈표 3-13〉 랜덤 회귀 대체방법과의 비교(평균)

무응답 비율	통계량	대체방법		
		평균 대체	랜덤 회귀 대체	응용 핫덱 대체
30%	MAE(오차비율)	**0.0152(0.70%)**	**0.0208(0.96%)**	**0.0210(0.97%)**
	CV	0.0088	0.0120	0.0121

〈표 3-14〉 랜덤 회귀 대체방법과의 비교(구성비)

범주(kg)	1미만	1 - 2	2 - 3	3 - 3.5	3.5이상
실제 인원수	296	196	293	246	129
구성비	**25.5%**	**16.9%**	**25.3%**	**21.2%**	**11.1%**
평균 대체방법					
대체후 인원수	211	277	411	173	88
구성비	**18.2%**	**23.9%**	**35.4%**	**14.9%**	**7.6%**
랜덤 회귀 대체방법					
대체후 인원수	264	232	299	219	146
구성비	**22.7%**	**20.0%**	**25.8%**	**18.9%**	**12.6%**
응용 핫덱 대체방법					
대체후 인원수	300	192	288	248	132
구성비	**25.9%**	**16.5%**	**24.8%**	**21.4%**	**11.4%**

이와는 반대로 구성비 변화 측면에서는 상당히 개선되었음을 알 수 있다. 회귀 대체의 경우 대체전후의 구성비 변화가 최대 10%까지 발생하던 것이 3%로 현저하게 줄어든 것을 볼 수 있다. 하지만 1% 미만의 변화를 보이는 응용 핫덱 대체방법과는 여전히 차이가 존재함을 알 수 있다. 따라서 본 실험의 결과를 정리하면 랜덤 회귀 대체방법을 사용하면 구성비 추정에는 오차가 줄어들 수 있으나, 평균 추정에는 오히려 오차가 늘어날 수도 있으며 또한 적절한 보조변수의 부재로 회귀모형을 적합시키기 어려운 경우에는 랜덤 회귀 대체방법을 사용한다 하더라도 한계가 있음을 알아야 할 것이다. 그러므로 본 자료의 경우에는 응용 핫덱 대체방법으로 대체를 실시하여 출생아체중의 평균 및 구성비를 추정하는 것이 실험한 다른 방법들에 비해서 보다 정확할 것으로 판단된다.

다음으로 다양한 범주에 대하여 구성비의 변화 정도가 일정한지 살펴보고자 한다.

〈표 3-15〉 출생아체중의 대체전후의 구성비 변화(2)

범주(kg)	0.7미만	0.7 - 1.4	1.4 - 2.1	2.1 - 2.8	2.8이상
실제 인원수	114	292	104	180	470
구성비	**9.8%**	**25.2%**	**9.0%**	**15.5%**	**40.5%**
평균 대체방법					
대체후 인원수	81	206	211	339	323
구성비	**7.0%**	**17.8%**	**18.2%**	**29.2%**	**27.8%**
랜덤 회귀 대체방법					
대체후 인원수	112	272	148	213	415
구성비	**9.6%**	**23.4%**	**12.8%**	**18.4%**	**35.8%**
응용 핫덱 대체방법					
대체후 인원수	108	283	110	182	477
구성비	**9.3%**	**24.4%**	**9.5%**	**15.7%**	**41.1%**

앞의 실험에서의 범주도 연구자가 임의로 정한 것이지만 범주가 변함에 따라서 그 결과가 일정한지도 확인할 필요성이 있을 것이다. 따라서 다른 두 가지의 범주에 대하여 기존의 실험을 반복하여 실시하였다. <표 3-15>에는 새로운 범주에 대하여 무응답 비율이 30%인 경우에 대체전후의 구성비 변화의 결과가 제시되어 있다. 총 1,160명의 출생아체중 구성비를 살펴보면 0.7kg미만은 9.8%, 0.7-1.4kg은 25.2%, 1.4-2.1kg은 9.0%, 2.1-2.8kg은 15.5%, 2.8kg이상은 40.5%이다. 30%의 임의 결측값을 대체하고 난 후의 구성비는 이전의 실험과 거의 유사함을 알 수 있다. 평균 대체방법은 대체후의 구성비가 상당히 왜곡되었고, 랜덤 회귀 대체방법은 회귀 대체방법에 비해서는 변화가 줄었지만 최대 5%정도의 오차가 있는 반면에 응용 핫덱 대체방법은 1%의 이하의 변화를 보이고 있어 대체로 인한 구성비의 변화문제는 크게 고려하지 않아도 될 것으로 판단된다.

<표 3-16>에는 또 하나의 새로운 범주에 대한 대체전후의 구성비 변화의 결과가 제시되어 있다. 이 경우도 이전의 결과와 거의 유사한 패턴을 보이고 있음을 알 수 있다.

따라서 다양한 실험의 결과를 고려할 때 랜덤 회귀 대체방법은 회귀 대체방법에 비해서 대체전후의 구성비 변화의 문제는 줄어들지만 평균 추정에 있어서 오차비율이 커진다는 것을 알 수 있으며, 각 범주의 변화에 관계없이 대체전후의 구성비 변화정도는 모든 방법이 일정하다고 판단된다. 그러므로 이후의 실험은 이 결과를 바탕으로 진행을 하도록 할 것이다.

〈표 3-16〉 출생아체중의 대체전후의 구성비 변화(3)

범주(kg)	0.5미만	0.5 - 1.5	1.5 - 2.5	2.5 - 3.5	3.5이상
실제 인원수	25	391	181	434	129
구성비	**2.2%**	**33.7%**	**15.6%**	**37.4%**	**11.1%**
평균 대체방법					
대체후 인원수	18	282	246	524	90
구성비	**1.5%**	**24.3%**	**21.2%**	**45.2%**	**7.8%**
랜덤 회귀 대체방법					
대체후 인원수	30	358	235	407	130
구성비	**2.6%**	**30.9%**	**20.2%**	**35.1%**	**11.2%**
응용 핫덱 대체방법					
대체후 인원수	27	399	179	425	130
구성비	**2.3%**	**34.4%**	**15.4%**	**36.7%**	**11.2%**

2) 재태기간

연관성분석의 결과 재태기간 항목은 출생아체중과 생존기간 항목을 대체군으로 사용한다. 재태기간 항목은 총 영아사망 자료 1,703명 중에서 1,080명의 자료가 있어 대략 40%의 누락(무응답)이 발생하고 있다. 따라서 총 1,080명의 자료를 모집단이라 가정하여 재태기간에 대한 무응답 대체 실험을 실시하였다.

〈표 3-17〉 재태기간 항목의 모의실험 결과(mean=33.85주)

무응답 비율	통계량	대체방법		
		평균 대체	회귀 대체	응용 핫덱 대체
10%	MAE /오차비율	**0.0192**(0.06%)	**0.0205**(0.06%)	**0.0274**(0.08%)
	CV	0.0007	0.0008	0.0011
20%	MAE	**0.0292**(0.09%)	**0.0305**(0.09%)	**0.0436**(0.13%)
	CV	0.0011	0.0011	0.0016
30%	MAE	**0.0365**(0.11%)	**0.0403**(0.12%)	**0.0576**(0.17%)
	CV	0.0014	0.0015	0.0022
40%	MAE	**0.0481**(0.14%)	**0.0503**(0.15%)	**0.0738**(0.22%)
	CV	0.0018	0.0019	0.0029
50%	MAE	**0.0562**(0.17%)	**0.0607**(0.18%)	**0.0970**(0.29%)
	CV	0.0021	0.0023	0.0036

재태기간에 대한 종합적인 대체결과가 <표 3-17>에 정리되어 있다. 출생아체중과 마찬가지로 세 가지 방법 모두가 평균을 추정하는데 있어 매우 좋은 대체가 되었다는 것을 알 수 있다. 오차의 비율은 출생아체중의 경우보다 훨씬 더 작게 나타났다. 이는 연관성분석의 결과 재태기간과 연관성이 높은 출생아체중을 이용했기 때문이라고 볼 수 있다. 하지만 실제 대체에서는 출생아체중을 대체하고 난 후에 적용해야 하므로 실제 오차는 조금 더 커질 수도 있을 것으로 판단된다. 재태기간의 경우 실제자료의 총 누락비율이 약 40%이므로 무응답 비율이 40%인 모의실험에 주목할 필요가 있다. 평균, 회귀, 응용 핫덱 대체방법의 오차비율은 0.14%, 0.15%, 0.22%로 평균추정에 대해서 세 방법 모두가 좋은 대체결과를 보여주고 있다. 출생아체중의 경우처럼 응용 핫덱 대체방법의 오차가 조금 더 크지만 큰 차이는 아닌 것으로 판단된다. 그리고 CV값 역시 매우 작은 값을 나타내므로 세 방법 모두 출생아체중의 평균값을 추정한 결과는 믿을 수 있을 것으로 생각된다.

다음으로 대체를 하고 난 후의 구성비(분포)의 변화를 살펴보고자 한다. <표 3-18>에는 세 방법에 대해서 무응답 비율이 40%인 경우에 대체전후의 구성비 변화의 결과가 제시되어 있다.

〈표 3-18〉 재태기간의 대체전후의 구성비 변화(무응답비율 40%인 경우)

범주(주)	26미만	26 - 30	30 - 35	35 - 40	40이상
실제 인원수	154	178	116	439	193
구성비	**14.3%**	**16.5%**	**10.7%**	**40.6%**	**17.9%**
평균 대체방법					
대체후 인원수	133	212	100	522	113
구성비	**12.3%**	**19.6%**	**9.3%**	**48.3%**	**10.5%**
회귀 대체방법					
대체후 인원수	135	210	128	437	170
구성비	**12.5%**	**19.5%**	**11.8%**	**40.5%**	**15.7%**
응용 핫덱 대체방법					
대체후 인원수	150	180	106	445	199
구성비	**13.9%**	**16.7%**	**9.8%**	**41.2%**	**18.4%**

총 1,080명의 재태기간 구성비를 살펴보면 26주 미만은 14.3%, 26-30주는 16.5%, 30-35주는 10.7%, 35-40주는 40.6%, 40주 이상은 17.9%이다. 40%의 임의 결측값을 대체

하고 난 후의 구성비는 출생아체중의 경우보다는 변화가 작지만 평균대체는 약 2~8%, 회귀대체는 약 0~3%의 구성비 변화를 가져오므로 사용하기에는 다소 큰 것으로 판단된다. 이 경우에도 랜덤 회귀 대체방법을 이용할 시 구성비의 변화 정도는 줄어들지만 평균 추정에서의 오차비율이 늘어남을 보여 출생아체중의 경우와 같은 결론을 내릴 수 있을 것이다. 따라서 구성비의 변화가 1% 미만으로 유지되고 있는 응용 핫덱 대체방법이 두 방법에 비해서 재태기간 항목의 대체에 가장 적합한 것으로 판단된다. 평균 추정에 있어서의 오차비율 차이보다 구성비 추정에서의 오차의 차이가 현저하게 크므로 연구자는 응용 핫덱 대체방법을 추천하고자 한다.

다음으로 대체군 사용에 대한 내용을 검토하고자 한다. 재태기간의 경우 출생아체중과 생존기간을 대체군으로 사용을 한다. 하지만 재태기간과 출생아체중 항목은 자료의 대부분이 동시에 누락이 되기 때문에 대체군으로 사용을 하기 위해서는 출생아체중이 먼저 대체가 되어야 한다. 그러므로 대체가 된 출생아체중의 자료를 다시 재태기간 항목의 대체를 위해 사용함으로써 발생하는 오차도 고려하여야 할 것이다. 또한 출생아체중 항목을 대체군에서 제외를 하고 대체를 실시할 수도 있을 것이다. 그러나 재태기간과 출생아체중의 연관성이 매우 크므로 이로 인한 정보의 손실도 상당히 클 것으로 판단된다. 따라서 출생아체중 항목의 대체군 사용여부를 판단하기 위해서 모의실험을 실시하였다.

〈표 3-19〉 출생아체중의 대체군 포함여부에 따른 재태기간의 대체결과

무응답 비율	통계량	응용 핫덱 대체방법	
		대체군 포함	대체군 미포함
40%	MAE(오차비율)	**0.0750**(0.22%)	**0.1016**(0.30%)
	CV	0.0030	0.0042

<표 3-19>에는 출생아체중 항목의 대체군 포함여부에 따라서 재태기간 항목의 대체결과가 제시되어 있다. 대체군에 포함을 시킨 경우에는 평균추정에 대한 오차비율이 0.22%이며, 대체군에서 제외된 경우에는 0.30%로 나타났다. 비록 대체가 된 출생아체중의 정보를 이용하기 때문에 실제 정보와의 차이는 존재하지만 두 항목간의 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 3-20>에는 무응답 비율이 40%인 경우에 대체전후의 구성비 변화의 결과가 주어져 있다.

〈표 3-20〉 출생아체중의 대체군 포함여부에 따른 구성비 변화정도

범주(주)	26미만	26 - 30	30 - 35	35 - 40	40이상
실제 인원수	154	178	116	439	193
구성비	**14.3%**	**16.5%**	**10.7%**	**40.6%**	**17.9%**
대체군에 포함된 경우					
대체후 인원수	151	168	111	446	204
구성비	**14.0%**	**15.5%**	**10.3%**	**41.3%**	**18.9%**
대체군에 미포함된 경우					
대체후 인원수	143	164	132	429	212
구성비	**13.3%**	**15.2%**	**12.2%**	**39.7%**	**19.6%**

두 경우 모두 대체전후의 구성비 변화정도는 크지 않은 것으로 보인다. 출생아체중 항목을 대체군에 포함을 시킨 경우에는 최대 1% 정도의 변화가 있으며, 대체군에서 제외된 경우에는 최대 1.7% 정도로 나타났다. 이는 응용 핫덱 대체방법의 장점이 그대로 반영된 것으로 볼 수 있을 것이다. 하지만 대체군에 포함을 시킨 경우에 더 좋은 추정결과를 보이는 것은 평균 추정의 경우와 같으며, 본 자료에서는 재태기간 항목을 대체할 때에 출생아체중 항목을 대체군에 포함을 시키는 것이 더 적절할 것으로 판단된다.

3) 모(母)연령

연관성분석의 결과 모연령 항목은 생존기간과 출생아체중 항목을 대체군으로 사용한다. 모연령 항목은 총 영아사망 자료 1,703명 중에서 835명의 자료가 있어 대략 50%의 누락(무응답)이 발생하고 있는데 무응답 비율이 상당히 높은 것으로 나타났다. 따라서 총 835명의 자료를 모집단이라 가정하여 모연령에 대한 무응답 대체 실험을 실시하였다.

모연령에 대한 종합적인 대체결과가 <표 3-21>에 정리되어 있다. 앞의 두 항목처럼 세 가지 방법 모두가 평균을 추정하는데 있어 매우 좋은 대체가 되었다는 것을 알 수 있다. 모연령의 경우 실제자료의 총 누락비율이 약 50%이므로 무응답 비율이 50%인 모의실험에 주목할 필요가 있다. 평균, 회귀, 응용 핫덱 대체방법의 오차비율은 0.45%, 0.44%, 0.62%로 평균추정에 대해서 세 방법 모두가 좋은 대체결과를 보여주고 있다. 단지 응용 핫덱 대체방법의 오차가 조금 더 크지만 큰 차이는 아닌 것으로 판단된다. 그리고 CV값 역시 매우 작은 값을 나타내므로 세 방법 모두 출생아체중의 평균값을 추정한 결과는 믿을 수 있을 것으로 생각된다.

〈표 3-21〉 모연령 항목의 모의실험 결과(mean=30.43세)

무응답 비율	통계량	대체방법		
		평균 대체	회귀 대체	응용 핫덱 대체
10%	MAE /오차비율	**0.0451**(0.15%)	**0.0449**(0.15%)	**0.0616**(0.20%)
	CV	0.0019	0.0019	0.0025
20%	MAE	**0.0642**(0.21%)	**0.0641**(0.21%)	**0.0953**(0.31%)
	CV	0.0027	0.0027	0.0038
30%	MAE	**0.0891**(0.29%)	**0.0887**(0.29%)	**0.1213**(0.39%)
	CV	0.0037	0.0036	0.0049
40%	MAE	**0.1084**(0.36%)	**0.1066**(0.35%)	**0.1479**(0.49%)
	CV	0.0044	0.0044	0.0060
50%	MAE	**0.1372**(0.45%)	**0.1339**(0.44%)	**0.1892**(0.62%)
	CV	0.0057	0.0056	0.0077

다음으로 대체를 하고 난 후의 구성비(분포)의 변화를 살펴보고자 한다. <표 3-22>에는 세 방법에 대해서 무응답 비율이 50%인 경우에 대체전후의 구성비 변화의 결과가 제시되어 있다.

〈표 3-22〉 모연령의 대체전후의 구성비 변화(무응답 비율 50%인 경우)

범주(세)	26미만	26 - 30	30 - 32	32 - 34	34이상
실제 인원수	140	251	127	150	167
구성비	**16.8%**	**30.0%**	**15.2%**	**18.0%**	**20.0%**
평균 대체방법					
대체후 인원수	62	143	443	94	93
구성비	**7.4%**	**17.1%**	**53.1%**	**11.3%**	**11.1%**
회귀 대체방법					
대체후 인원수	62	199	411	70	93
구성비	**7.4%**	**23.9%**	**49.2%**	**8.4%**	**11.1%**
응용 핫덱 대체방법					
대체후 인원수	130	265	133	135	172
구성비	**15.6%**	**31.7%**	**15.9%**	**16.2%**	**20.6%**

총 835명의 모연령의 구성비를 살펴보면 26세 미만은 16.8%, 26-30세는 30.0%, 30-32세는 15.2%, 32-34세는 18.0%, 34세 이상은 20.0%이다. 50%의 임의 결측값을 대체하고 난 후의 구성비는 세 방법의 차이가 상당히 큰 것을 볼 수 있다. 평균 및 회귀 대체방법은 대체후의 구성비가 상당히 왜곡이 된 반면에 응용 핫덱 대체방법은 구성비의 변화가 2% 미만으로 유지되고 있다. 이러한 결과는 출생아체중의 경우와 비슷한 원인으로 보인다. 모연령과 연관성이 높은 부연령, 동거기간, 총출산아수 등과 같은 항목들을 사용할 수 없으므로 회귀모형으로 대체를 실시하는 것은 어려움이 따른다고 생각된다. 비록 평균 추정에서는 좋은 값을 추정할 수는 있으나 자료의 퍼짐정도가 너무 넓어 구성비가 중심으로 몰리기 때문에 심각한 문제가 발생하는 것이다. 따라서 모연령 항목도 응용 핫덱 대체방법을 이용하여 대체를 실시하여야 할 것으로 판단된다.

나. 사산 자료

1) 모(母)연령

사산 자료의 모연령은 연관성분석의 결과 재태기간 항목만을 대체군으로 사용해야 한다. 사산 자료는 보조변수로 쓸 수 있는 것이 재태기간 항목밖에 없기 때문에 두 항목간의 연관성은 낮으나 대체에 이용하기로 한다. 하지만 모연령 항목은 총 사산 자료 8,572명 중에서 7,854명의 자료가 있어 대략 10%의 누락(무응답)이 발생하고 있는데 다른 항목들에 비해서는 무응답 비율이 낮다고 할 수 있다.

〈표 3-23〉 모연령 항목의 모의실험 결과(mean=29.01세)

무응답 비율	통계량	대체방법		
		평균 대체	회귀 대체	응용 핫덱 대체
5%	MAE /오차비율	**0.0120**(0.04%)	**0.0116**(0.04%)	**0.0179**(0.06%)
	CV	0.0005	0.0005	0.0008
10%	MAE	**0.0211**(0.07%)	**0.0192**(0.07%)	**0.0281**(0.10%)
	CV	0.0009	0.0008	0.0012
15%	MAE	**0.0266**(0.09%)	**0.0228**(0.08%)	**0.0364**(0.13%)
	CV	0.0011	0.0010	0.0015
20%	MAE	**0.0305**(0.11%)	**0.0255**(0.09%)	**0.0415**(0.14%)
	CV	0.0013	0.0012	0.0018
25%	MAE	**0.0356**(0.12%)	**0.0293**(0.10%)	**0.0463**(0.16%)
	CV	0.0015	0.0014	0.0020

총 7,854명의 자료를 모집단이라 가정하여 모연령에 대한 무응답 대체 실험을 실시하였다. 사산 자료의 모연령에 대한 종합적인 대체결과가 <표 3-23>에 정리되어 있다. 영아사망 자료의 항목들과 유사하게 세 가지 방법 모두가 평균을 추정하는데 있어서는 매우 좋은 대체가 되었다는 것을 알 수 있다. 모연령의 경우 실제자료의 총 누락비율이 약 10%이므로 무응답 비율이 10%인 모의실험의 결과를 보면 평균, 회귀, 응용 핫덱 대체방법의 오차비율은 0.07%, 0.07%, 0.10%로 상당히 낮은 오차비율을 가짐을 볼 수 있다. 또한 응용 핫덱 대체방법의 오차가 조금 더 크지만 큰 차이는 아닌 것으로 판단되며 CV값 역시 매우 작은 값임을 알 수 있다.

그리고 <표 3-24>에는 세 방법에 대해서 무응답 비율이 10%인 경우에 대체전후의 구성비 변화의 결과가 제시되어 있다. 총 7,854명의 모연령의 구성비를 살펴보면 26세 미만은 27.3%, 26-30세는 22.8%, 30-32세는 13.4%, 32-34세는 13.1%, 34세 이상은 23.4%이다. 10%의 임의 결측값을 대체하고 난 후의 구성비는 세 방법에서 많은 차이가 발생함을 알 수 있다. 평균 및 회귀 대체방법은 임의구간에서 대체후의 구성비가 대략 7%가 왜곡이 된 반면에 응용 핫덱 대체방법은 구성비의 변화가 0.2% 미만으로 상당히 좋은 대체가 되고 있음을 보여준다. 이러한 결과는 이전의 항목과 비슷한 원인으로 보인다. 따라서 모연령 항목도 응용 핫덱 대체방법을 이용하여 대체를 실시하여야 할 것으로 판단된다.

〈표 3-24〉 모연령의 대체전후의 구성비 변화(무응답 비율 10%인 경우)

범주(세)	26미만	26 - 30	30 - 32	32 - 34	34이상
실제 인원수	2142	1792	1055	1031	1834
구성비	**27.3%**	**22.8%**	**13.4%**	**13.1%**	**23.4%**
평균 대체방법					
대체후 인원수	1930	2353	1001	932	1638
구성비	**24.6%**	**29.9%**	**12.7%**	**11.9%**	**20.9%**
회귀 대체방법					
대체후 인원수	1930	2326	1008	942	1648
구성비	**24.6%**	**29.6%**	**12.8%**	**12.0%**	**21.0%**
응용 핫덱 대체방법					
대체후 인원수	2137	1802	1053	1042	1820
구성비	**27.2%**	**22.9%**	**13.4%**	**13.3%**	**23.2%**

2) 사산아체중

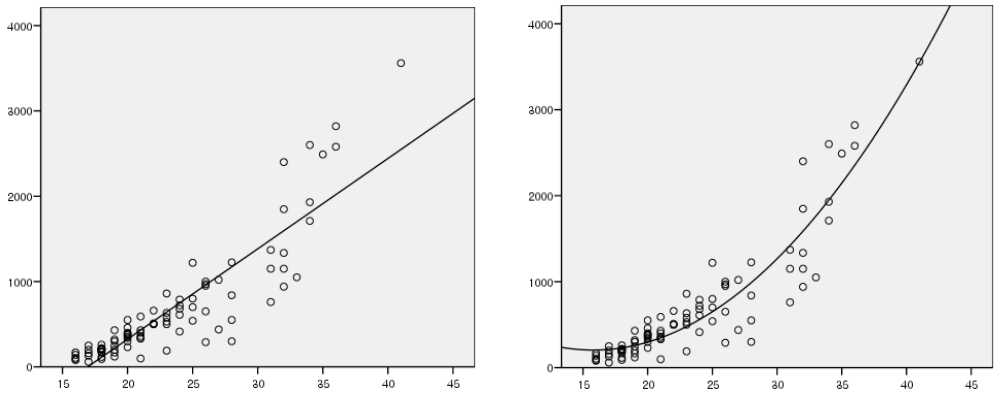
연관성분석의 결과 사산아체중 항목은 재태기간 항목만을 대체군으로 사용해야 한다. 사산아체중의 경우 총 사산 자료 8,572명 중에서 95명의 자료만 있어서 대략 99%의 누락(무응답)이 발생하고 있는데 이 항목의 무응답 대체 가능 여부를 잘 판단해야 할 것이다. 일반적으로 이런 경우에 무응답 대체를 한다는 것은 의미가 없을 것으로 생각되지만 본 연구에서는 모의실험의 결과를 통하여 가능성 여부를 제시하고자 한다.

한 가지 희망적인 것은 사산아체중은 보조변수로 사용할 재태기간과 매우 높은 연관성이 있기 때문에 좋은 모형을 구축할 수 있을 것으로 생각되며 이로 인해서 추정의 정확도가 매우 나쁘지는 않을 것으로 기대한다. 모의실험은 총 95명의 자료를 모집단이라 가정하여 사산아체중에 대한 무응답 대체를 실시하였다. 사산아체중에 대한 종합적인 대체결과가 <표 3-25>에 정리되어 있다.

〈표 3-25〉 사산아체중 항목의 모의실험 결과(mean=687.62 g)

무응답 비율	통계량	대체방법				
		평균 대체	회귀 대체 (직선, 절편o)	회귀 대체 (직선, 절편x)	회귀 대체 (2차 곡선)	응용 핫덱 대체
10%	MAE	**11.34**(1.65%)	**10.03**(1.46%)	**18.97**(2.76%)	**9.51**(1.38%)	**13.32**(1.94%)
	CV	0.0210	0.0185	0.0272	0.0176	0.0243
20%	MAE	**17.72**(2.58%)	**15.17**(2.21%)	**28.87**(4.20%)	**14.61**(2.12%)	**20.18**(2.93%)
	CV	0.0334	0.0284	0.0423	0.0271	0.0372
30%	MAE	**23.33**(3.39%)	**20.51**(2.98%)	**42.21**(6.14%)	**19.73**(2.87%)	**27.34**(3.98%)
	CV	0.0429	0.0375	0.0551	0.0359	0.0486
40%	MAE	**29.46**(4.28%)	**24.79**(3.61%)	**55.65**(8.09%)	**23.80**(3.46%)	**33.86**(4.92%)
	CV	0.0547	0.0457	0.0701	0.0443	0.0602
50%	MAE	**37.43**(5.44%)	**31.58**(4.59%)	**67.49**(9.82%)	**30.95**(4.50%)	**40.94**(5.95%)
	CV	0.0688	0.0576	0.0885	0.0564	0.0725

이전의 모의실험에 비해서 두 가지 방법이 추가가 되었다. 그 이유는 회귀 대체방법으로 모형을 구축한 경우 추정식에 의하여 사산아체중의 대체값이 음의 값일 가능성이 있기 때문에 절편이 없는 모형으로 다시 추정을 하였다. 또한 모형구축결과 1차 모형보다 2차 모형이 조금 더 적절한 것으로 판단되어 이를 모의실험에 추가하였다. 이에 관한 내용은 [그림 3-5]를 참조하면 자세히 알 수 있을 것이다.



[그림 3-5] 사산아체중과 재태기간의 산점도(1차 모형 vs. 2차 모형)

<표 3-25>의 결과를 살펴보면 이전의 경우에 비해서는 다소 오차비율이 높다는 것을 알 수 있다. 이것은 사용할 수 있는 자료의 수가 너무 작기 때문에 모형을 구축하든지 적절한 도너(donor)를 선택하든지 어려움이 따른다는 것을 나타낸다. 하지만 99%가 누락된 상황에서 이 정도로 추정을 할 수 있다면 긍정적으로 생각할 수도 있을 것이다. 사산아체중의 경우 실제자료의 총 누락비율이 너무 높아 모의실험의 여건상 무응답 비율이 50%인 경우의 결과를 고려하기로 한다. 평균, 회귀(직선 및 절편 있음), 회귀(직선 및 절편 없음), 회귀(2차 곡선), 응용 핫덱 대체방법의 오차비율은 5.44%, 4.59%, 9.82%, 4.50%, 5.95%로 2차 곡선의 형태로 추정한 경우가 실제 평균과 가장 근사하게 추정을 하고 있음을 보여준다. 하지만 절편 없는 회귀 대체를 제외한 나머지 방법들은 이전의 경우와 마찬가지로 큰 차이는 보이지 않음을 알 수 있다.

그리고 <표 3-26>에는 다섯 방법에 대해서 무응답 비율이 50%인 경우에 대체전후의 구성비 변화의 결과가 제시되어 있다. 총 95명의 사산아체중의 구성비를 살펴보면 250g 미만은 28.4%, 250-500g은 23.2%, 500-750g은 17.9%, 750-1000g은 9.5%, 1000g 이상은 21.0%이다. 자료의 수가 작기 때문에 실제 구성비와 추정 구성비의 차이가 크게 보일 수 있으나, 실제 자료의 수는 8,572명이므로 모의실험의 결과보다는 줄어들 것으로 판단된다. 50%의 임의 결측값을 대체하고 난 후의 구성비의 차이는 응용 핫덱 대체방법이 가장 작다는 것을 알 수 있으나, 회귀(2차 곡선) 대체방법도 응용 핫덱 대체방법과 많은 차이는 발생하지 않는다. 따라서 사산아체중 항목은 평균 추정은 회귀(2차 곡선) 방법이 우수하며 구성비 추정은 응용 핫덱 방법이 우수하므로 이 두 방법 중에서 어느 방법으로 대체를 실시하여도 큰 차이가 없을 것으로 보인다. 평균 추정과 구성비 추정 중에서 더 중점적으로 생각되는 방향으로 방법을 선택할 수도 있으며, 원 자료의 제공을 고려한다면 같은 값이 반복되는 응용 핫덱 대체방법보다 각각 다른 값이 추정되는 회귀(2차 곡선) 대체방법이 더 바람직할 수도 있을 것이다. 또한 모든 항목을 하나의 방법으로 통일

하여 응용 핫덱 대체방법을 사용하는 것도 의미가 있을 것이다.

〈표 3-26〉 사산아체중의 대체전후의 구성비 변화(무응답비율 50%인 경우)

범주(g)	250미만	250 - 500	500 - 750	750 - 1000	1000이상
실제 인원수	27	22	17	9	20
구성비	**28.4%**	**23.2%**	**17.9%**	**9.5%**	**21.0%**
평균 대체방법					
대체후 인원수	12	49	7	5	22
구성비	**12.5%**	**51.6%**	**7.4%**	**5.3%**	**23.2%**
회귀(직선, 절편 o) 대체방법					
대체후 인원수	22	24	13	10	26
구성비	**23.2%**	**25.3%**	**13.7%**	**10.5%**	**27.3%**
회귀(직선, 절편 x) 대체방법					
대체후 인원수	12	13	24	22	24
구성비	**12.5%**	**13.7%**	**25.3%**	**23.2%**	**25.3%**
회귀(2차 곡선) 대체방법					
대체후 인원수	22	25	17	9	22
구성비	**23.2%**	**26.3%**	**17.9%**	**9.4%**	**23.2%**
응용 핫덱 대체방법					
대체후 인원수	25	24	17	11	18
구성비	**26.3%**	**25.3%**	**17.9%**	**11.6%**	**18.9%**

본 연구자는 응용 핫덱 대체방법으로 모든 항목을 일괄적으로 처리하는 것에 더 무게를 두고 제안하고자 하나 담당자 의견을 충분히 반영하여 결정하고자 한다. <표 3-27>에는 전체 모의실험의 결과가 요약되어 있으니 참조하기 바란다. 그리고 다음 절에서는 선정된 대체방법을 2008년 출생전후기 사망통계 자료에 적용하고 그 결과를 기술할 것이다.

〈표 3-27〉 전체 모의실험의 결과요약

자료	항목	평균 오차비율	구성비 변화비율	선정된 대체방법
영아사망	출생아체중	**0.98%**	**0.8%**	**응용 핫덱**
	재태기간	**0.22%**	**0.9%**	**응용 핫덱**
	모연령	**0.62%**	**1.8%**	**응용 핫덱**
사산	모연령	**0.10%**	**0.2%**	**응용 핫덱**
	사산아체중	**5.95%**	**2.1%**	**응용 핫덱** / 회귀(2차)

제5절 출생전후기 사망통계 자료에의 적용

앞에서 제시한 각 항목의 대체군 및 응용 핫덱 대체방법을 이용하여 2007년, 2008년 출생전후기 사망통계 자료의 누락부분에 대하여 대체를 실시하고자 한다. 이를 통하여 획득한 자료만 사용하여 분석한 결과와 누락된 부분을 추가하여 분석한 결과를 비교해 보고자 한다. 실제 누락된 부분의 값은 알 수 없으므로 이전의 모의실험의 결과에서 제시된 오차정도를 감안하여 대체결과를 판단해야 할 것이다.

1. 2007년 출생전후기 사망통계 자료

2007년 출생전후기 사망통계 자료는 영아사망 1,703명, 사산 8,572명의 자료로 구성되어 있으며 각 항목에 따라서 10%~99%까지의 누락 부분을 포함하고 있다.

가. 출생아체중(영아사망 자료)

출생아체중 항목의 총 자료의 수는 1,160명으로 전체 자료의 30% 정도가 누락이 되어 있다. 이 항목의 모의실험 결과를 살펴보면 평균 추정에 대한 오차비율은 0.98%, 구성비 변화비율은 1% 미만으로 나타났다. <표 3-28>에는 누락된 부분을 제외한 자료로 작성한 결과와 누락된 부분을 대체한 후의 전체 자료로 작성한 결과가 제시되어 있다.

〈표 3-28〉 출생아체중 항목에의 적용결과(누락비율 30%)

누락부분이 제외된 평균			**2.17(kg)**		
누락부분이 대체된 후의 평균			**2.03(kg)**		
범주(kg)	1미만	1 - 2	2 - 3	3 - 3.5	3.5이상
누락부분이 제외된 구성비					
구성비	**25.5%**	**16.9%**	**25.3%**	**21.2%**	**11.1%**
누락부분이 대체된 후의 구성비					
구성비	**31.8%**	**19.1%**	**22.1%**	**18.1%**	**8.9%**

누락부분이 제외된 경우의 출생아체중의 평균은 2.17(kg)이었으나, 대체가 된 후의 평균은 2.03(kg)으로 약 6.5%정도가 감소된 것을 볼 수 있다. 따라서 체중이 적게 나가는 출생아 자료가 많은 부분 누락이 되었음을 알 수 있다. 이와 같은 결과는 누락된 부분이 대체된 후의 구성비 변화에서도 확인할 수가 있다. 1kg미만은 25.5%에서 31.8%로 증가

하였고, 1-2kg도 16.9%에서 19.1%로 늘어남을 볼 수 있다. 반면에 2-3kg은 25.3%에서 22.1%, 3-3.5kg은 21.2%에서 18.1%, 3.5kg 이상은 11.1%에서 8.9%로 조금씩 감소가 된 것을 알 수 있다. 이는 대체군으로 사용되었던 생존기간의 값이 작은 신생아의 체중이 많이 누락이 되었기 때문인 것으로 판단된다.

나. 재태기간(영아사망 자료)

재태기간 항목의 총 자료의 수는 1,080명으로 전체 자료의 40% 정도가 누락이 되어 있다. 이 항목의 모의실험 결과를 살펴보면 평균 추정에 대한 오차비율은 0.22%, 구성비 변화비율은 1% 미만으로 나타났다. <표 3-29>에는 누락된 부분을 제외한 자료로 작성한 결과와 누락된 부분을 대체한 후의 전체 자료로 작성한 결과가 제시되어 있다.

〈표 3-29〉 재태기간 항목에의 적용결과(누락비율 40%)

누락부분이 제외된 평균			**33.85(주)**		
누락부분이 대체된 후의 평균			**32.54(주)**		
범주(주)	26미만	26 - 30	30 - 35	35 - 40	40이상
누락부분이 제외된 구성비					
구성비	**14.3%**	**16.5%**	**10.7%**	**40.6%**	**17.9%**
누락부분이 대체된 후의 구성비					
구성비	**20.3%**	**19.4%**	**11.9%**	**34.4%**	**14.0%**

누락부분이 제외된 경우의 재태기간의 평균은 33.85(주)이었으나, 대체가 된 후의 평균은 32.54(주)로 약 3.9% 정도가 감소된 것을 볼 수 있다. 따라서 재태기간이 짧았던 출생아 자료가 많은 부분 누락이 되었음을 알 수 있으며 누락된 부분이 대체된 후의 구성비 변화에서도 확인할 수가 있다. 26주 미만은 14.3%에서 20.3%로, 26-30주는 16.5%에서 19.4%로 높은 비율로 늘어난 반면에, 35-40주는 40.6%에서 34.4%, 40주 이상은 17.9%에서 14.0%로 줄어든 것을 볼 수 있다. 이와 같은 결과도 출생아체중과 마찬가지로 대체군으로 사용되었던 출생아체중과 생존기간의 값이 작은 신생아의 재태기간이 많이 누락이 되었기 때문인 것으로 판단된다.

다. 모연령(영아사망 자료)

모연령 항목의 총 자료의 수는 835명으로 전체 자료의 50%정도가 누락이 되어 있다. 이 항목의 모의실험 결과를 살펴보면 평균 추정에 대한 오차비율은 0.62%, 구성비 변화

비율은 2%미만으로 나타났다. <표 3-30>에는 누락된 부분을 제외한 자료로 작성한 결과와 누락된 부분을 대체한 후의 전체 자료로 작성한 결과가 제시되어 있다.

〈표 3-30〉 모연령 항목에의 적용결과(누락비율 50%)

누락부분이 제외된 평균			**30.43(세)**		
누락부분이 대체된 후의 평균			**30.60(세)**		
범주(세)	26미만	26 - 30	30 - 32	32 - 34	34이상
누락부분이 제외된 구성비					
구성비	**16.8%**	**30.0%**	**15.2%**	**18.0%**	**20.0%**
누락부분이 대체된 후의 구성비					
구성비	**15.7%**	**30.2%**	**13.6%**	**21.5%**	**19.0%**

누락부분이 제외된 경우의 모연령의 평균은 30.43(세)이며 대체가 된 후의 평균은 30.60(세)로 대략 0.5%정도가 증가하여 대체전후의 평균 차이는 크지 않는 것으로 보인다. 대체전후의 구성비를 살펴보면 26세 미만은 16.8%에서 15.7%, 26-30세는 30.0%에서 30.2%, 30-32세는 15.2%에서 13.6%, 32-34세는 18.0%에서 21.5%, 34세 이상은 20.0%에서 19.0%정도의 변화가 나타났다. 나이가 작은 범주에서 줄어드는 경향이 있으나 상대적으로 구성비가 크기 때문에 오히려 연령의 평균은 조금 높아진 것으로 판단된다. 하지만, 모의실험의 결과에서 보았듯이 2% 정도의 오차가 있다는 것을 감안하고 해석을 하여야 할 것이다.

라. 모연령(사산 자료)

모연령 항목의 총 자료의 수는 7,854명으로 전체 자료의 10%정도가 누락이 되어 있다. 이 항목의 모의실험 결과를 살펴보면 평균 추정에 대한 오차비율은 0.10%, 구성비 변화비율은 0.2%미만으로 나타났다. <표 3-31>에는 누락된 부분을 제외한 자료로 작성한 결과와 누락된 부분을 대체한 후의 전체 자료로 작성한 결과가 제시되어 있다.

누락부분이 제외된 경우의 모연령의 평균은 29.01(세)이며 대체가 된 후의 평균은 29.02(세)로 거의 차이를 보이지 않는다. 대체전후의 구성비 역시 최대 0.2% 정도만 차이가 있어 누락부분의 대체는 큰 영향을 주지 않는 것으로 생각된다. 이러한 결과는 영아사망 자료의 모연령에 비하여 자료의 수가 상당히 많고 또한 누락비율이 상대적으로 낮기 때문인 것으로 판단된다. 실제 모연령 항목의 경우에는 출생아체중이나 재태기간 항목에 비해 누락부분을 제외한 자료와 누락부분을 대체한 자료는 큰 차이를 보이지 않는 것으로 보인다.

〈표 3-31〉 모연령 항목에의 적용결과(누락비율 10%)

누락부분이 제외된 평균			**29.01(세)**		
누락부분이 대체된 후의 평균			**29.02(세)**		
범주(세)	26미만	26 - 30	30 - 32	32 - 34	34이상
누락부분이 제외된 구성비					
구성비	**27.3%**	**22.8%**	**13.4%**	**13.1%**	**23.4%**
누락부분이 대체된 후의 구성비					
구성비	**27.4%**	**22.6%**	**13.4%**	**13.1%**	**23.5%**

마. 사산아체중(사산 자료)

사산아체중 항목의 총 자료의 수는 95명으로 전체 자료의 99%가 누락이 되어 있다. 이 항목의 모의실험 결과를 살펴보면 평균 추정에 대한 오차비율은 5.95%, 구성비 변화비율은 2.1% 정도로 나타났지만 50%의 누락을 가정한 실험이므로 실제로는 이보다 오차가 더 클 것으로 판단된다. <표 3-32>에는 누락된 부분을 제외한 자료로 작성한 결과와 누락된 부분을 대체한 후의 전체 자료로 작성한 결과가 제시되어 있다.

〈표 3-32〉 사산아체중 항목에의 적용결과(누락비율 99%)

누락부분이 제외된 평균			**687.62(g)**		
누락부분이 대체된 후의 평균			**503.65(g)**		
범주(g)	250미만	250 - 500	500 - 750	750 - 1000	1000이상
누락부분이 제외된 구성비					
구성비	**28.4%**	**23.2%**	**17.9%**	**9.5%**	**21.1%**
누락부분이 대체된 후의 구성비					
구성비	**36.9%**	**28.2%**	**16.0%**	**6.9%**	**12.1%**

누락부분이 제외된 경우의 사산아체중의 평균은 687.62(g)이었으나, 대체가 된 후의 평균은 503.65(g)로 약 26.8% 정도가 감소된 것을 볼 수 있다. 따라서 체중이 작은 사산아 자료가 많은 부분 누락이 되었음을 알 수 있다. 이와 같은 결과는 누락된 부분이 대체된 후의 구성비 변화에서도 확인할 수가 있는데 250g 미만은 28.4%에서 36.9%, 250-500g은 23.2%에서 28.2%로 높은 비율로 늘어났음을 볼 수 있다. 반면에 750-1000g은 9.5%에서 6.9%, 1000g 이상은 21.1%에서 12.1%로 많은 감소가 된 것을 알 수 있다. 이는 대체군으로 사용되었던 재태기간의 값이 작은 사산아의 체중이 많이 누락이 되었기 때문인 것으로 판단된다.

2. 2008년 출생전후기 사망통계 자료

2008년 출생전후기 사망통계 자료는 영아사망 1,580명, 사산 8,276명의 자료로 구성되어 있으며 각 항목에 따라서 5%~99%까지의 누락 부분을 포함하고 있다. 2008년 자료는 2007년의 자료와 거의 대부분이 유사하여 2007년 자료로부터 얻어진 모의실험의 결과를 그대로 받아들여도 큰 문제가 없을 것으로 판단된다.

가. 출생아체중(영아사망 자료)

출생아체중 항목의 총 자료의 수는 1,071명으로 전체 자료의 30% 정도가 누락이 되어 있다. <표 3-33>에 제시된 결과는 2007년 자료의 출생아체중 항목과 거의 유사함을 알 수 있다.

〈표 3-33〉 출생아체중 항목에의 적용결과(누락비율 30%)

누락부분이 제외된 평균			**2.20(kg)**		
누락부분이 대체된 후의 평균			**2.02(kg)**		
범주(kg)	1미만	1 - 2	2 - 3	3 - 3.5	3.5이상
누락부분이 제외된 구성비					
구성비	**25.3%**	**16.3%**	**23.9%**	**23.2%**	**11.3%**
누락부분이 대체된 후의 구성비					
구성비	**31.8%**	**17.5%**	**20.8%**	**19.7%**	**10.2%**

누락부분이 제외된 경우의 출생아체중의 평균은 2.20(kg)이었으나, 대체가 된 후의 평균은 2.02(kg)으로 약 8.2% 정도가 감소된 것을 볼 수 있으며, 누락된 부분이 대체된 후의 구성비 변화에서도 체중이 낮은 범주는 비율이 증가하고 높은 범주는 감소한 것을 알 수 있다.

나. 재태기간(영아사망 자료)

재태기간 항목의 총 자료의 수는 999명으로 전체 자료의 40% 정도가 누락이 되어 있다. <표 3-34>에는 누락된 부분을 제외한 자료로 작성한 결과와 누락된 부분을 대체한 후의 전체 자료로 작성한 결과가 제시되어 있다.

〈표 3-34〉 재태기간 항목에의 적용결과(누락비율 40%)

누락부분이 제외된 평균			**34.19(주)**		
누락부분이 대체된 후의 평균			**32.90(주)**		
범주(주)	26미만	26 - 30	30 - 35	35 - 40	40이상
누락부분이 제외된 구성비					
구성비	**13.6%**	**14.7%**	**10.2%**	**43.7%**	**17.8%**
누락부분이 대체된 후의 구성비					
구성비	**20.0%**	**18.4%**	**10.6%**	**36.4%**	**14.6%**

누락부분이 제외된 경우의 재태기간의 평균은 34.19(주)이었으나, 대체가 된 후의 평균은 32.90(주)로 약 3.8%정도가 감소된 것을 볼 수 있으며, 누락된 부분이 대체된 후의 구성비 변화에서도 2007년 자료와 유사하게 재태기간이 낮은 범주는 비율이 증가하고 높은 범주는 감소한 것을 알 수 있다.

다. 모연령(영아사망 자료)

모연령 항목의 총 자료의 수는 798명으로 전체 자료의 50%정도가 누락이 되어 있다. <표 3-35>에 제시된 결과는 2007년 자료의 모연령 항목과 거의 유사함을 알 수 있다.

〈표 3-35〉 모연령 항목에의 적용결과(누락비율 50%)

누락부분이 제외된 평균			**30.57(세)**		
누락부분이 대체된 후의 평균			**31.07(세)**		
범주(세)	26미만	26 - 30	30 - 32	32 - 34	34이상
누락부분이 제외된 구성비					
구성비	**16.5%**	**29.3%**	**17.0%**	**13.2%**	**24.0%**
누락부분이 대체된 후의 구성비					
구성비	**13.5%**	**28.8%**	**14.2%**	**18.0%**	**25.5%**

라. 모연령(사산 자료)

모연령 항목의 총 자료의 수는 7,906명으로 전체 자료의 5%정도가 누락이 되어 있다. <표 3-36>에는 누락된 부분을 제외한 자료로 작성한 결과와 누락된 부분을 대체한 후의 전체 자료로 작성한 결과가 제시되어 있다.

〈표 3-36〉 모연령 항목에의 적용결과(누락비율 5%)

누락부분이 제외된 평균			**29.23(세)**		
누락부분이 대체된 후의 평균			**29.21(세)**		
범주(세)	26미만	26 - 30	30 - 32	32 - 34	34이상
누락부분이 제외된 구성비					
구성비	**26.0%**	**22.7%**	**13.4%**	**12.2%**	**25.7%**
누락부분이 대체된 후의 구성비					
구성비	**26.1%**	**22.6%**	**13.5%**	**12.2%**	**25.6%**

누락부분이 제외된 경우의 모연령의 평균은 29.23(세)이며 대체가 된 후의 평균은 29.21(세)로 거의 차이를 보이지 않는다. 또한 대체전후의 구성비 역시 변화를 보이지 않아 누락부분의 대체는 큰 영향을 주지 않는 것으로 생각된다.

마. 사산아체중(사산 자료)

사산아체중 항목의 총 자료의 수는 109명으로 전체 자료의 99%가 누락이 되어 있다. <표 3-37>에는 누락된 부분을 제외한 자료로 작성한 결과와 누락된 부분을 대체한 후의 전체 자료로 작성한 결과가 제시되어 있다.

〈표 3-37〉 사산아체중 항목에의 적용결과(누락비율 99%)

누락부분이 제외된 평균			**879.89(g)**		
누락부분이 대체된 후의 평균			**569.86(g)**		
범주(g)	250미만	250 - 500	500 - 750	750 - 1000	1000이상
누락부분이 제외된 구성비					
구성비	**11.9%**	**33.0%**	**20.3%**	**6.4%**	**28.4%**
누락부분이 대체된 후의 구성비					
구성비	**29.9%**	**35.9%**	**15.1%**	**4.9%**	**14.2%**

누락부분이 제외된 경우의 사산아체중의 평균은 879.89(g)이었으나, 대체가 된 후의 평균은 569.86(g)로 약 35.2%정도가 감소된 것을 볼 수 있다. 따라서 체중이 작은 사산아 자료가 많은 부분 누락이 되었음을 알 수 있다. 이와 같은 결과는 누락된 부분이 대체된 후의 구성비 변화에서도 확인할 수가 있다. 본 결과는 2007년의 결과와 조금의 차이가 보인다. 범주가 250g 미만과 250 이상-500 미만에서의 구성비가 상반되게 나오는 것을 볼 수 있다. 이는 2008년 자료의 실제 구성비에 의하여 이러한 결과를 가져올 수도 있을 것이고, 다른 하나의 이유는 누락되지 않은 자료의 특성 때문일 수가 있을 것이다. 연구

자는 후자의 가능성이 더 클 것으로 생각된다. 즉, 사용할 수 있는 자료의 양(약 1%)이 너무나도 작기 때문에 누락부분이 제외된 구성비가 전체의 구성비 추정에 많은 영향을 주게 될 것이다. 따라서 이 항목의 경우에는 누락되지 않은 자료의 특성에 의하여 최종 대체결과가 상당히 바뀔 수도 있음을 명심해야 할 것이다.

제6절 결 론

본 연구에서는 출생전후기 사망통계의 주요 항목(출생아체중, 재태기간, 모연령(영아사망), 모연령(사산), 사산아체중)에 대하여 연관성 분석을 실시하여 대체군을 제시하였고, 여러 대체방법을 이용한 모의실험을 실시하여 대체의 정확성을 검토하였다. 또한 연구의 결과를 바탕으로 2007년 및 2008년의 출생전후기 사망통계 자료의 실제 누락부분에 대해서 대체를 실시하여 대체전후의 평균 및 구성비의 변화를 살펴보았다.

출생전후기 사망통계 자료는 영아사망 자료와 사산 자료로 나누어지는데, 이들 자료는 사망신고, 화장신고, 모자보건신고와 같은 행정자료를 통하여 획득되어진다. 그러나 신고서식의 상이하여 자료의 수집과정에서 주요 항목에 대한 누락비율이 높으며 연관성이 높은 항목들이 동시에 누락이 되어 대체를 위한 과정에서 많은 어려움이 발생하고 있다. 연관성 분석의 결과로부터 대체군으로 사용되어야 하는 항목들이 동시에 누락이 되어 연관성이 낮은 항목들을 대체군으로 이용할 수밖에 없으며, 누락비율이 매우 높은 항목들은 대체결과의 정확성이 다소 떨어짐을 볼 수 있다. 따라서 이러한 문제점을 해결하기 위해 자료의 수집과정에서의 노력도 계속되어야 할 것이다.

본 연구의 모의실험에서는 2007년 출생전후기 사망통계 자료(영아사망 1,703명, 사산 8,572명) 중에서 각 항목별로 누락이 되지 않은 자료를 이용하였고, 실험을 위해 각각의 목표변수에 대해서 다양한 비율로 무응답을 발생시켰다. 그리고 평균대체, 회귀대체, 응용 핫덱 대체방법을 이용하여 항목별로 무응답을 모두 대체하고 하고 난 후에 평균의 차이정도와 임의로 범주화된 분포변화비율을 살펴보았다. 모의실험을 실시한 항목들은 모두 비슷한 결과를 보여주고 있다. 평균 추정에 대해서는 평균 및 회귀 대체방법이 응용 핫덱 대체방법에 비해 오차비율이 조금 더 낮은 것을 볼 수 있다. 반면에 구성비 변화 측면에서는 상대적으로 응용 핫덱 대체방법이 훨씬 더 정확한 추정을 하고 있음을 알 수 있다. 따라서 전체적인 오차를 감안하면 응용 핫덱 대체방법이 가장 적절할 것으로 판단된다. 또한 구성비 변화를 줄이기 위해서 오차변동을 감안한 랜덤 회귀 대체방법으로 대체를 실시한 경우 기존의 회귀 대체방법에 비해서 구성비 변화의 폭은 줄어들지만 평균 추정에 대한 오차비율이 커짐을 확인하였다. 마지막으로 사산아체중 항목의 경우 누락부분이 99%이므로 이를 대체하여 사용하기에는 신중한 결정을 내려야 할 것이

다. 모의실험의 결과 다른 항목들에 비해 추정 오차가 훨씬 크다는 것을 알 수 있으며 누락되지 않은 1%의 자료가 전체자료와 얼마나 근사한 분포를 가지는지에 따라서 대체결과가 큰 폭으로 변할 수도 있음을 명심해야 할 것이다. 또한 본 연구에서의 모의실험은 완전한 자료에 대해서 임의로 무응답을 발생시킨 것으로 실제 누락부분을 대체하였을 경우에는 이 결과보다는 다소 추정오차가 커질 가능성이 있을 것으로 판단된다. 하지만 본 연구에서 제시된 대체군과 대체방법은 현재의 출생전후기 사망통계 자료를 적절하게 대체하기 위한 하나의 방법임에 틀림없을 것으로 확신한다.

본 연구의 결과를 바탕으로 2007년 및 2008년 출생전후기 사망통계 자료의 누락부분을 대체한 결과 출생아체중, 재태기간, 사산아체중이 낮은 자료들이 많은 부분 누락이 되었음을 알 수 있다. 그러므로 이 누락부분을 제외한 결과를 사용하는 것은 상당히 왜곡된 결과를 가져올 수도 있다는 것을 보여준다. 모든 항목들에 대한 자세한 적용결과는 제5절을 참조하기 바란다.

출생전후기 사망통계 자료는 일반 조사통계들에 비하여 무응답(누락) 비율이 너무나도 높다. 무응답률의 정도와 대체 가능성의 규칙이 일괄적으로 정해지지는 않았지만 대체를 적용하기 위해서는 많은 검토가 필요할 것으로 판단된다. 그리고 무응답률이 높다고 할지라도 획득된 자료의 상태 및 연관성 높은 대체군의 존재 등 다른 여건에 따라서 대체가 가능할 수도 있을 것이다. 다른 나라의 경우 30~50% 정도의 무응답을 대체하여 사용한 사례도 있다. 본 연구에서 다루었던 5개 항목 중에서 출생아체중(30%누락), 재태기간(40%누락), 모연령(영아사망, 50%누락), 모연령(사산, 10%누락)의 4개 항목은 누락의 비율은 높지만 모의실험의 결과 대체하여 사용하는 것이 가능하다고 판단되며 적용결과에서도 일괄성이 있는 것으로 나타났다. 하지만 사산아체중(99%누락)의 경우 누락되지 않은 자료가 너무 작기 때문에 이 자료에 의하여 대체결과의 변동이 매우 클 것으로 보인다. 따라서 국가통계를 생산한다는 의미에서 볼 때 사산아체중 항목은 대체를 하여 사용하기에는 위험 부담이 크다고 생각되며 실제 적용 단계에서는 제외를 시키는 것이 합당할 것으로 판단된다.

본 연구로써 출생전후기 사망통계 자료의 무응답 대체 연구는 마무리하고자 한다. 하지만 자료수집시 새로운 항목이 추가되거나 현재의 자료에서 변화가 생겼을 때에는 추가적인 보완 연구가 수행되어야 할 것이다. 최근 들어 통계조사의 환경이 계속적으로 악화되고 있어 각종 조사의 무응답률이 높아지고 있다. 이와 같은 현실에서 행정자료 활용에 관한 노력은 상당히 중요하다고 할 수 있다. 따라서 행정자료를 활용하여 수집된 자료를 바탕으로 생산되는 출생전후기 사망통계가 더 정확한 통계가 되기 위해서는 다양한 정보와 정확한 자료를 수집할 수 있도록 지속적인 노력과 연구가 진행되어야 할 것이다. 본 연구도 이러한 과정 중에 하나이며 향후 출생전후기 사망통계의 품질을 향상시키는데 많은 부분 도움이 되기를 기대한다.

참고문헌

김규성(2000), “무응답 대체 방법과 대체 효과”, 「조사연구」, 제1권 2호, pp.1-14.
김규성(2000), “표본 대체 방법과 대체자료의 합리적 이용”, 한국은행 지원논문
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.
김재광・한근식・윤연옥(2004), “가계조사 무응답 처리기법 연구”, 통계청,「통계연구」, 제9권 1호, pp.79-102.
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.
이진희・김진・이기재(2006), “표본조사에서 공간 변수를 이용한 결측 대체의 효율성 비교”, 「응용통계연구」, 제19권 1호, pp.57-67.
이현정(2008), “인구주택총조사 무응답 처리기법 연구(Ⅰ)”, 연구보고서, 통계개발원.
이현정・최필근(2009), “인구주택총조사 무응답 처리기법 연구(Ⅱ)”, 연구보고서, 통계개발원.
최통진(2006), “농림어업총조사를 위한 무응답 보정에 관한 연구”, 석사학위논문.
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.
통계교육원(2005), 「무응답처리 실무론」.
통계청(2005), 「2005 농림어업총조사 조사지침서」.
(2007), 「농림어업총조사 조사항목 변천 자료집」.

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.
Agresti, A.(1990), Categorical Data Analysis, A Wiley-Interscience Publication.
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.