머리말

통계학은 연구의 결론을 객관적이고 올바르게 도출하기 위하여 자료를 수집, 처리, 해석하는 학문이다. 연구의 모집단을 대표하는 표본을 추출하고 이 표본을 바탕으로 올바르게 적용된 통계분석방법은 모든 분야의 연구에서 결론의 타당성을 과학적으로 증명할 수 있도록 도와준다. 하지만 대다수의 연구에서는 여러 가지 이유로 인하여 자료에 무응답 또는 결측이 발생하게 된다. 일반적으로 무응답 자료는 응답 자료와는 그 성격이 다르고 따라서 응답 자료는 모집단을 대표하지 못한다. 따라서 무응답을 제거 또는 무시하고 완전히 관찰된 자료만을 이용하여 기존의 통계분석을 적용하게 되면 그 결과에 편향이 발생하며 또한 부분 정보를 분석에서 무시하므로 부적절한 추론을 야기한다.

연구자는 무응답을 줄이기 위해서 연구계획 및 설계 단계부터 노력을 기울여야 한다. 하지만 이러한 노력에도 불구하고 자료에 무응답이 발생하는 경우에는 무응답을 고려한 적절한 통계적 분석방법을 적용하여 연구결과의 신뢰성을 높이는 노력을 하여야 한다. 물론 무응답을 포함하는 자료의 분석방법은 완전히 응답된 자료만을 이용한 분석방법보다 어렵다. 또한 대다수의 무응답 자료 분석에 관한 교재들이 무응답 자료의 실제분석 예제보다는 이론 위주로 서술되어 많은 연구자들이 실제 자료 분석에 무응답 분석방법을 활용하기에 한계가 있다. 따라서 무응답 자료 분석법의 이론 설명과 더불어 실제 활용 방법을 적절하게 조화시킨 교재의 개발은 매우 중요하다. 이러한 필요성을 충족하기 위하여 본 교재를 개발하게 되

었다. 이 교재는 무응답 자료 분석을 실제로 시행하는 연구자들이 지침서로 사용할 수 있도록 이론적인 설명과 함께 실제 적용 방법에 대한 설명에 중점을 두고 있다. 이론도 가능한 한 수리적인 유도보다는 무응답 자료의 분석에 필요한 개념이나 기법을 설명하는 방식으로 접근하였으며 예제와 사례분석을 통해 설명된 분석기법을 실제 자료에 적용하는 방법을 소개하였다.

이 교재는 통계학 석사 수준의 통계학 지식을 지닌 연구자들이 이해할 수 있을 정도의 이론을 다루었다. 하지만 통계학을 전공하지 않은 연구자들도 이론 설명을 포함하지 않는 많은 부분들을 이해할 수 있을 것으로 기대된다. 따라서 이론 중심으로 서술된 절에는 * 표시를 해 두어 응용에 관심있는 독자들은 이 절들을 건너뛸 수 있도록 하였다. 이 교재의 내용은 크게 두 부분으로 나눌 수 있는데 제 1부는 무응답 자료의 분석을 위한 통계적 방법들을 서술하고 있고 제 2부에서는 4가지 자료에 대한 사례연구를 통해 실제 자료에서 적용방법을 예를 들어 설명하고자 하였다. 이론 기술 중에서 가중방법과 우도를 근거로한 분석방법은 무응답 자료분석의 대표 교재인 Little과 Rubin의 “Statistical Analysis with Missing Data, 2nd Ed."을 기초로 작성하였다.

이 교재를 순서대로 읽어 나가면 무응답 자료 처리를 위한 방법론에 대한 전체적인 흐름을 파악할 수 있을 것이다. 하지만, 특정 처리 방법에 관심이 있는 연구자는 어느 정도 통계적 지식이 있는 경우 해당 부분만 읽더라도 큰 어려움은 없을 것으로 예상된다. 하지만 무응답을 포함한 자료의 분석 방법에 대한 기초 개념을 설명하고 있는 제 1장을 읽고 다른 세부 방법론으로 진행할 것을 강력히 추천한다.

이 교재는 통계청 통계교육원의 무응답 자료처리 및 분석 강의 교재로서 개발되었고 이 책에서 논의하는 대부분의 예제들은 통계청에서 수집한 자료를 이용하여 설명하였다. 하지만 소개된 방법들은 다른 실험, 사회조사 및 의학 연구 자료에도 폭넓게 적용될 수 있을 것이다.

마지막으로 무응답 자료 처리를 위한 통계적 방법론들을 정리하여 교재를 집필할 수 있는 기회를 주신 통계교육원 변효섭 원장님, 본 교재 집필방향에 조언을 해 주고 교재를 통한 통계 교육 프로그램 개발에 도움을 주신 통계교육원 김정란 사무관님, 인구주택 총조사 자료의 사례분석을 위하여 상세한 정보를 제공해주신 통계개발원 최필근 박사님, 그리고 노동연구원 고령화패널조사를 사례분석에 포함시키는데 도움을 주신 노동연구원 장지연 박사님께 깊은 감사를 드린다. 또한 본 교재에서 설명한 통계방법에 대한 SAS 프로그램을 개발하고 예제 자료 준비 및 교정을 도와준 고려대학교 정경대학 통계학과 및 의과대학 의학통계학교실 대학원생들에게도 감사의 뜻을 전한다.

고려대학교 정경대학 통계학과 송주원
고려대학교 의과대학 의학통계학과 안형진

차 례

제 1부

무응답 자료의 분석을 위한 통계적 방법

제 1장 무응답의 발생

< 학습목표 >

(1) 무응답의 의미에 대하여 예를 들어 설명한다.

(2) 무응답 자료 분석을 위한 용어를 정의한다.

(3) 무응답 자료의 발생 패턴을 고려한다.

(4) 무응답 자료의 메커니즘을 정의하고 분석에 미치는 영향을 고려한다.

1.1 무응답의 의미

자료를 수집하는 과정에서 일부 항목이 측정되지 않으면 그 항목에 대한 응답이 발생하지 않았다는 의미로 무응답(nonresponse)이 발생했다고 한다. 또는 그 항목의 값이 관측되지 않고 빠져있다는 의미로 결측값(missing value)라고 부른다. 무응답은 자연과학 분야의 실험에서도 발생하지만 설문 조사, 정보 수집을 위한 자료, 의학 자료 등 거의 모든 분야의 자료에서 발생한다. 무응답이 발생하는 몇 가지 예는 다음과 같다.

- 실험에서의 무응답

  화학실험을 실시할 때 일부표본에 대하여 시약을 잘못 투여하여 반응값이 나타나지 않는 경우 이 표본들에 대한 반응값은 결측으로 남게 된다.

- 설문조사에서의 무응답

  통계청에서 실시하는 가계동향조사는 가구의 수입과 지출에 대한 조사 항목을

포함한다. 소득 액수나 지출의 세부 사항에 관한 질문에 대하여 응답 거부가 종종 발생한다. 주택마련 시기에 대한 항목의 경우 주택을 소유하지 않은 대상자들은 모두 응답이 불가능하여 결측으로 남게 된다.

● 사회조사에서의 무응답

청소년의 흡연 정도를 조사하는 경우 일정 기간 동안의 흡연량이 관심의 대상이다. 이 때 일반적으로 흡연 여부를 먼저 설문한 후 이 문항에 대하여 흡연으로 응답한 청소년에게는 흡연량에 대한 문항에 대하여 응답하도록 하고 흡연하지 않은 경우 흡연량에 대한 문항을 뛰어넘도록 질문지가 구성된다. 이 경우 흡연을 하지 않은 경우 흡연량에 관한 문항에 대한 대답은 결측으로 남게 된다.

● 여론조사에서의 무응답

대통령 선거에서 후보자 중 누구에게 투표할 지 여론조사를 실시하면 무응답이 발생하는데 이 중 일부는 본인의 선택을 알려주기를 꺼려하기 때문에 응답을 거부하는 반면 일부는 누구에게 투표할 지 결정하지 못하고 있기 때문에 응답하지 못하며 나머지는 선호하는 후보자가 없어 선택할 수 없는 경우이다.

● 정보 수집 자료

각 기업은 제품의 소비자에 대한 여러 가지 기본 정보를 수집하는데 일부 소비자의 경우 일부 정보에서 결측이 발생한다. 예를 들어, 제품의 구매 고객의 연령, 직업 또는 소득과 같은 개인 정보는 모든 고객에게서 응답되지 않고 이들에 대한 자료값은 결측으로 남게 된다.

● 임상실험 자료

암환자를 위하여 새로 개발된 약에 대한 임상실험을 실시하면 일부 환자들은 중도에 참여를 포기하며 이 경우 이들의 추후 경과는 결측으로 남게 된다. 이 중 일부는 약에 대한 심각한 부작용으로 인하여 연구에서 중도탈락하고 일부는 사망하여 이 후 자료를 제공하지 못한다.

● 의학 자료

병원에서 작성한 의무기록은 질병에 관한 소중한 정보를 포함한다. 하지만, 환자에 따라 다른 검사를 실시하거나 다른 항목의 정보를 포함하고 있다. 예를 들어, 대형 병원의 의무기록은 몸무게나 혈압 자료를 포함하는 경우가 많지만 1차 진료 기관에서 작성한 의무 기록은 몸무게나 혈압에 관한 정보를 포함하지 않는 경우가 발생하고 이 값들은 무응답으로 남게 된다.

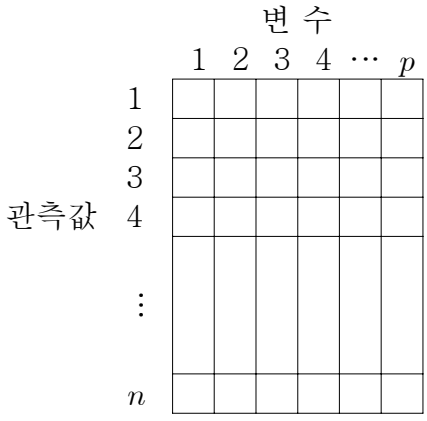
위의 예제 중 일부 경우의 무응답은 엄밀히 말하면 무응답이라 할 수 없다. 첫 번째로 일부 무응답은 실제로는 무응답이 아니라 어떤 특정한 값을 의미한다. 사회조사 자료의 예에서 흡연하지 않은 청소년의 흡연량은 무응답으로 남게 되는데 이는 흡연량이 0이기 때문에 질문 문항을 뛰어넘은 경우에 해당되므로 흡연량을 0으로 대입하여 분석하여야 한다. 이 때 무응답은 손쉽게 정확한 값을 알아 낼 수 있으므로 무응답이라 할 수 없다. 두 번째로 일부 무응답은 선택하도록 주어지지 않은 다른 항목에 대한 응답을 의미한다. 대통령 선거와 관련한 여론조사에서 후보자 중 누구에게 투표할 지 응답하지 않는 경우 중 만약 선호하는 후보자가 없어 여론조사에 응답하지 않았다면 무응답은 응답 문항에 선택할 문항이 없기 때문에 발생한다. 이 문항의 경우 각 후보들 외에 선호하는 후보 없음이라는 항목을 추가함으로써 이 집단을 분리해 내고 무응답 비율을 줄이는 방법을 선택하는 것이 적절하다. 일반적으로 무응답에 대한 연구는 위와 같이 문맥에서 응답값을

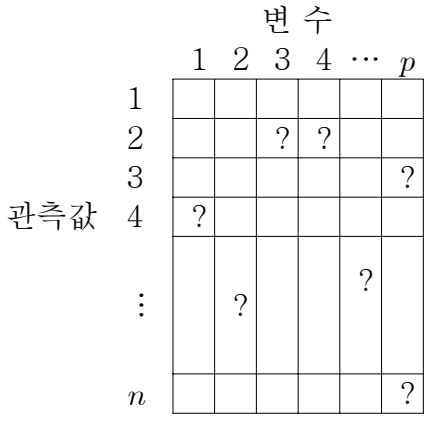
알 수 있거나 적절한 항목을 포함시킴으로써 제외시킬 수 있는 무응답은 고려하지 않는다. 즉, 본 교재에서 고려하는 무응답에 관한 분석 방법은 질문에 대하여 정확한 값을 얻을 수 있거나 항목에 대해 응답할 수 있지만 여러 가지 원인으로 인하여 응답이 무엇인지 알려지지 않은 경우만을 고려한다.

분석을 위하여 입력된 자료의 값들은 <그림 1.1(1)>과 같이 직사각형 형태로 행렬(matrix)을 이용하여 나타낸다. 이 자료행렬(data matrix)에서 각 행(row)은 관측값의 단위(observation unit)를 의미하고 각 열(column)은 변수(variable)를 의미한다. 무응답이 발생하지 않는다면 자료의 모든 원소(element)는 각 관측값에 대하여 변수 각각의 응답된 값으로 채워지고 대부분의 통계분석 프로그램은 이와 같이 완전하게 응답된 형태의 자료에 대한 분석을 시행하도록 개발되어 있다. 한편, 무응답이 발생한다면 이 직사각형 자료의 모든 칸이 채워지지 못하고 응답되지 않은 값은 <그림 1.1(2)>와 같이 물음표를 사용하여 표현하는 것이 일반적이다. 이 때, 물음표로 표시하지 않은 칸은 응답된 값을 나타낸다.

<그림 1.1> 자료의 형태

(1) 무응답이 발생하지 않은 경우





직사각형의 행렬자료를 Y로 나타내자. 좀 더 자세히 n개의 행과 p개의 열을 가지는 자료 $Y=(y_{ij})$, $i=1,...,n$, $j=1,...,p$로 표현할 수 있는데 여기서, i는 n개의 관측값을 나타내고 j는 p개의 변수를 나타낸다. 무응답이 발생한다면 일부 원소인 y_{ij} 값은 관측되지 않는다. 즉, 자료 Y값은 응답된 y_{ij} 원소들과 무응답인 y_{ij} 원소들로 나누어지는데 응답된 y_{ij}들을 관측된(observed) Y라는 의미로 Y_{obs}로 나타내고 무응답인 y_{ij}들을 관측되지 않고 빠진(missing) Y라는 의미로 Y_{mis}로 나타낸다. 이 때, 관측값에 따라 Y_{obs}와 Y_{mis}에 포함되는 변수들의 조합은 달라진다. 이를 명확하게 하기 위하여 관측값의 번호를 포함하여 $y_{i,obs}$와 $y_{i,mis}$, $i=1,...,n$, 로 나타내기도 한다. 예를 들어 <그림 1.1(2)>에서 첫 번째 관측값 $(i=1)$은 모두 응답되어 $y_{1,obs}$는 모든 변수를 포함하는데 반하여 $y_{1,mis}$는 존재하지 않는다. 두 번째 관측값$(i=2)$은 세 번째와 네 번째 변수에서 결측이 발생하므로 $y_{2,obs}$는 세 번째와 네 번째 변수에 대한 관측값을 제외한 $(p-2)\times 1$ 벡터(vector)로 나타내고 $y_{2,mis}$는 세 번째와 네 번째 변수를 나타내는 2×1 벡터(vector)로 나타내는데 이 값은 물론 무응답이므로 비어 있다. 따라서 Y_{obs}는 각 관측값마다 다른 길이의 벡터를 포함할 수 있다.

Y가 자료의 응답값을 나타내는 반면에 어느 관측값의 어느 변수에서 무응답이 발생하였는지를 나타내기 위하여 응답 지시행렬(missing data indicator matrix) R을 사용한다. R은 Y와 동일하게 n개의 행과 p개의 열을 가지는 직사각형 행렬로서 $R=(r_{ij})$, $i=1,...,n$, $j=1,...,p$로 표현할 수 있는데 여기서, i는 n개의 관측값을 나타내고 j는 p개의 변수를 나타낸다. R은 응답 및 무응답이 발생한 위치를 나타내므로 i번째 관측값의 j번째 변수에서 응답값이 존재한다면 r_{ij}가 1의 값을 가지고 무응답이 발생한다면 r_{ij}가 0의 값을 가지도록 표현한다. 즉,

$$r_{ij}=\begin{cases} i\text{번째 관측값의 } j\text{번째 변수가 응답이면 } 1 \\ i\text{번째 관측값의 } j\text{번째 변수가 무응답이면 } 0 \end{cases}$$

와 같이 각 원소가 지시변수(indicator variable)로 표현될 수 있어 응답 지시행렬이라 부른다. <그림 1.2>는 (1) 무응답을 포함한 가상의 자료행렬 Y와 (2) Y의 응답 지시행렬 R의 예제이다.

<그림 1.2> 가상의 자료행렬 Y와 대응되는 응답 지시행렬 R의 예제

(1) 자료행렬 Y

		변수					
		가구 번호	가구원 번호	성별	나이	…	교육 정도
관측값	1	10001	01	2	29		1
	2	10002	01	2	?		?
	3	10002	02	1	45		?
	4	10002	03	?	19		?
	5	10003	01	2	?		5
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	n	21084	02	?	50		3

(2) 자료행렬 Y에 대응되는 응답 지시행렬 R

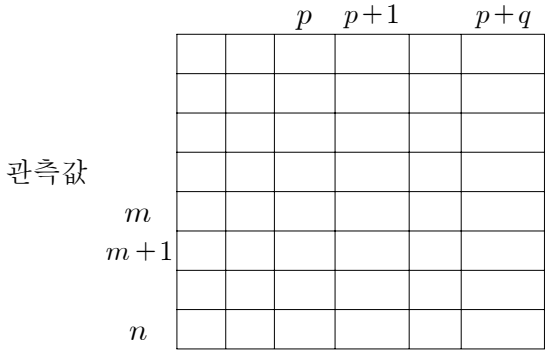
		변수					
		가구 번호	가구원 번호	성별	나이	…	교육 정도
관측값	1	1	1	1	1		1
	2	1	1	1	0		0
	3	1	1	1	1		0
	4	1	1	0	1		0
	5	1	1	1	0		1
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	n	1	1	0	1		1

1.2 무응답 자료 패턴

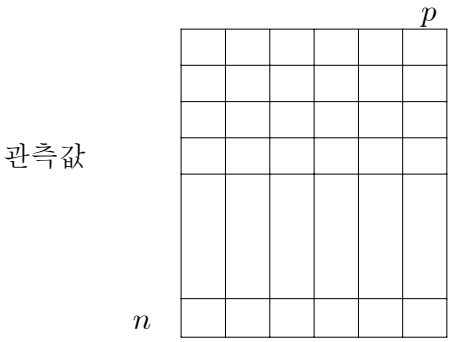
무응답의 발생 패턴은 다양하게 나타나는 데 대표적인 몇 가지 패턴이 <그림 1.3>에 나타난다(Little and Rubin, 2002). 일부 자료에서는 자료 자체가 정확히 이 패턴들을 따르지는 않지만 변수들 사이의 순서를 재정렬함으로써 이 패턴들로 표현이 가능하다. 각 패턴에 대한 자세한 설명은 다음과 같다.

<그림 1.3> 대표적인 무응답 발생 형태

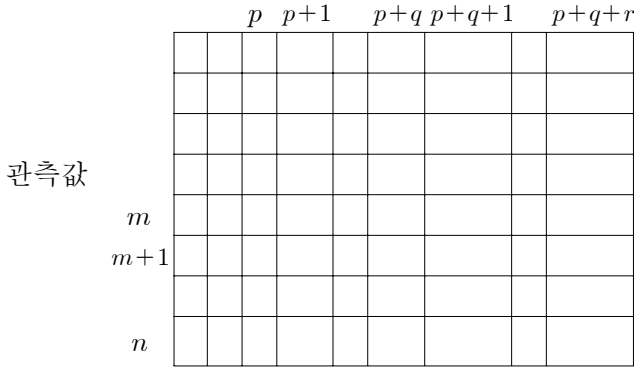
(1) 두 가지 패턴



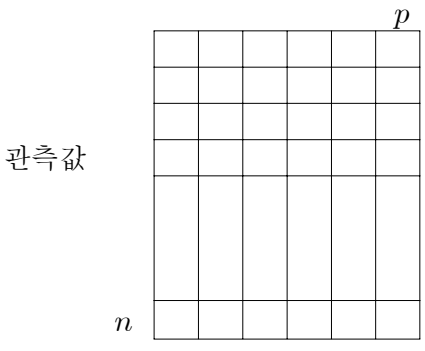
(2) 단조 패턴



(3) 자료 짝짓기



(4) 일반적인 패턴



(1) 두 가지 패턴(two pattern)

일부 관측값은 일부 변수에서만 무응답이 나타나는 경우를 의미한다. 즉, m개의 관측값에서는 모든 변수에 대하여 응답값이 존재하지만 나머지 $n-m$개의 관측값에서는 처음 p개의 변수에 대한 응답값만 존재하고 나머지 q개의 변수에 대한 응답은 결측이 된다. 이와 같은 무응답 패턴은 전체 표본에 대하여 조사를 시행한 후 일부 표본(m개)을 다시 추출하여 더 세부적인 조사를 시행하는 경우에 대표적으로 나타나는 무응답 발생 형태이다. 예를 들면, 통계청에서 인구주택총조사를 실시할 때 모든 사람들을 대상으로 전수조사(short form)를 실시하고 표본으로 뽑힌 사람들을 대상으로 자세한 조사(long form)를 통해 세부 정보를 얻어내는 경우를 들 수 있다. 이 무응답 발생 형태 중 가장 간단한 패턴은 변수 한 개에서만 무응답이 발생하는 형태로 일변량 무응답 패턴(univariate nonresponse pattern)이라고 부른다.

(2) 단조 패턴(monotone pattern)

처음 측정된 변수들에 대하여는 모든 관측값에서 응답이 이루어지지만 두 번째 측정된 변수들에서부터 무응답이 발생하는데 한 번 무응답이 발생한 관측값은 이후 측정된 모든 변수에 대한 모든 응답값이 무응답으로 남게 된다. 즉, 일단 무응답이 발생한 관측값의 추후 측정값은 모두 무응답으로 남게 되는 한편 이 후 변수들에서 무응답이 발생하는 관측값의 숫자가 지속적으로 증가되어 전체 무응답의 비율은 점차 증가하는 형태를 보이므로 무응답 비율이 점증적으로 증가한다는 의미로 단조 패턴이라 부른다. 이와 같은 형태의 무응답 패턴은 패널자료(panel data)에서 흔히 나타나는데 처음 시점에서는 모든 관측값에서 응답이 이루어지지만 이 후 시점에서 중도탈락이나 사망 등으로 인하여 관측되는 숫자가 점차 줄어드는 경우에 흔히 나타난다. 엄밀하게 말하면 (1) 두 가지 패턴은 단조패턴의 가장 간단한 형태이다.

(3) 자료 짝짓기(file matching)

모든 관측값에서 기초 변수들에 대하여는 응답값이 모두 존재하지만 나머지 변수들에 대하여 응답되는 관측값이 서로 다르다. 즉, 총 n개의 관측값 중에서 처음 m개의 관측값은 처음 $p+q$개의 변수들에 대한 응답값이 존재하는 반면 나머지 r개의 변수에 대한 값은 결측으로 되어 있고, 뒤 $n-m$개의 관측값에서는 처음 p개의 변수 및 마지막 r개의 변수들에 대한 응답은 존재하지만 중간 q개의 변수들의 값은 무응답으로 나타난다. 더구나 처음 m개의 관측값과 뒤 $n-m$개의 관측값은 처음 p개의 공통변수들에 대하여 응답값이 모두 존재하지만 나머지 $q+r$개의 변수들에서는 동시에 응답되지 않는다. 이와 같은 자료 형태는 인구통계학적 변수들(demographic variables)은 일치하지만 주요 관심 변수들이 각각 다른 두 자료를 병합(combine)할 때 주로 발생하는데 두 개의 다른 자료를 합하여 한 개의 자료로 합하는 과정에서 발생하는 무응답 패턴이라는 의미로 자료 짝짓기라 부른다.

(4) 일반적인 패턴(general pattern)

무응답은 어떤 관측값의 어떤 변수에서도 발생할 수 있으며 무응답의 비율도 변수별로 각각 다를 수 있다. 이 형태의 무응답은 어떤 특별한 형태를 지니지 않고 가장 일반적으로 나타날 수 있다는 의미로 일반적인 패턴이라 부른다. 즉, 이 무응답 발생 패턴은 앞의 세 가지 무응답 발생 패턴이 특별한 형태의 무응답 발생 형태를 요구하는 데 반하여 전혀 제약을 가지지 않는다는 의미로 가장 일반적인 패턴이라 말할 수 있다.

무응답의 발생 형태에 따라 분석 기법이 달라지는 데 특별한 형태의 무응답 형태를 요구하는 경우, 즉, 두 가지 패턴이나 단조 패턴의 무응답 형태를 보이는 경우 무응답을 대체할 때 손쉬운 방법을 사용하여 편향을 제거할 수 있지만 무응답 발생 형태에 대한 제약이 적은 일반적인 패턴을 가지는 경우 더 복잡한 분석 기법

이 요구된다. 또한, 자료 짝짓기 패턴을 가진 경우 동시에 측정되지 않는 변수들 사이의 연관성은 추정이 안 된다는 점을 유의하여야 한다.

1.3 무응답 자료 메커니즘

무응답 자료가 발생하는 메커니즘을 정확히 파악하는 것은 무응답 자료를 분석하기 위하여 매우 중요한 의미를 가진다. Little and Rubin(2002)에서는 무응답 자료의 메커니즘을 다음과 같이 세 가지로 분류하였다.

(1) 완전임의결측(Missing Completely At Random 또는 MCAR)
자료행렬 Y에서 무응답이 발생할 확률은 자료와 상관없이 완전히 무작위적이다. 응답 지시행렬 R의 자료행렬 Y에 대한 조건부 분포(conditional distribution)를 $f(R|Y,\phi)$, 여기서 ϕ는 R의 조건부 분포에 연관된 모수들(parameters)로 나타내면 완전임의결측은 모든 Y와 ϕ의 값에 대하여

$$f(R|Y,\phi)= f(R|\phi)$$

로 표현할 수 있다. 즉, 무응답의 발생은 자료 Y의 값과 상관없이 발생한다는 것을 의미한다. 여기서 $Y=(Y_{obs}, Y_{mis})$이므로 무응답의 발생은 관측되지 않은 자료인 Y_{mis} 뿐 아니라 관측된 자료인 Y_{obs}에도 의존하지 않는다. 예를 들어, 특정한 병을 가진 사람들을 상대로 유전자 검사를 실시하고자 한다. 이를 위하여 이 병을 가진 환자 1000명에 대한 의료 기록을 모았는데 유전자 검사 비용은 200명 분밖에 준비되지 않았다. 따라서 1000명 중 200명을 랜덤하게 선택하여 유전자 검사를 실시한다면 나머지 800명에 대한 의료기록은 존재하지만 유전자 검사 기록은 결측으로 남는다. 이 경우 1000명 중 200명을 랜덤하게 추출하였으므로 무응

답 자료 메커니즘은 완전임의결측이라 할 수 있다.

(2) 임의결측(Missing At Random 또는 MAR)

자료행렬 Y에서 무응답이 발생할 확률은 자료의 관측된 부분에는 연관되지만 자료의 관측되지 않은 부분과는 연관이 없다. 즉, 응답 지시행렬 R의 자료행렬 Y에 대한 조건부 분포 $f(R|Y,\phi)$는 모든 Y_{mis}와 ϕ의 값에 대하여

$$f(R|Y,\phi)= f(R|Y_{obs},\phi)$$

로 표현할 수 있다. 즉, 무응답의 발생은 자료 Y의 관측된 값들인 Y_{obs}에만 연관되고 관측되지 않은 Y_{mis}와는 상관이 없다는 것을 의미한다. 예를 들어, 가계동향조사에서 가구별 수입 변수에서 무응답이 발생하였고 수입에 관한 무응답은 수입이 높은 가구에서 많이 발생하는 경향이 있다고 가정하자. 하지만 우리가 수입과 매우 높은 연관이 있는 가구별 세금 정보를 구할 수 있고 가구의 수입에 대한 응답 여부는 세금 액수에 따라 다르게 나타나지만 동일한 액수의 세금을 납부한 가구들 중에서는 응답 여부가 완전히 임의로 결정된다면 분석에 세금 정보를 변수로 포함시킴으로써 가구수입에 관한 무응답 자료 메커니즘을 임의결측을 따르도록 할 수 있다.

(3) 비임의결측(Not Missing At Random 또는 NMAR)

자료행렬 Y에서 무응답이 발생할 확률은 자료의 관측된 부분인 Y_{obs} 뿐 아니라 관측되지 않은 부분인 Y_{mis}와도 연관되어 있다. 즉, 응답 지시행렬 R의 자료행렬 Y에 대한 조건부 분포 $f(R|Y,\phi)$는

$$f(R|Y,\phi)= f(R|Y_{obs}, Y_{mis},\phi)$$

로 표현할 수 있다. 예를 들어, 가계동향조사에서 가구별 수입 변수에서 무응답이 발생하였고 수입에 관한 무응답은 수입이 높은 가구에서 많이 발생하는 경향이 있다고 가정하자. 우리가 수입과 매우 높은 연관을 가지는 가구별 세금 정보를 구할 수 없고 대신 가구주의 학력 정보를 가지고 있는 경우를 생각하자. 가구주의 학력은 수입과 연관되지만 그 연관성이 매우 높지 않아 동일한 학력을 가진 가구주들 중에서도 여전히 가구 수입이 높은 경우 무응답이 많이 발생한다면 수입 응답 여부는 여전히 무응답인 수입액 자체에 의존하게 되므로 이 때 가구별 수입액의 무응답 자료 메커니즘은 비임의결측을 따른다고 볼 수 있다.

무응답을 포함한 자료에 대한 정보(information)는 두 개의 행렬인 자료행렬 Y와 응답 지시행렬 R로 나타내므로 무응답 자료에 대한 분석은 이 두 개의 행렬의 결합분포(joint distribution)를 통해 시행해야 한다. 자료행렬 Y와 관련된 모수들을 θ로 표현하면 이 결합분포는

$$f(Y, R|\theta,\phi)= f(Y|\theta)f(R|Y,\phi), \quad \text{여기서 } (\theta,\phi)\in\Omega_{\theta,\phi}$$

로 나타낼 수 있다. 이 때, $\Omega_{\theta,\phi}$은 (θ,ϕ)의 결합모수공간(joint parameter space)을 의미한다. 즉, 무응답을 포함한 자료에 대한 분석은 자료행렬에 대한 모형 $f(Y|\theta)$ 뿐 아니라 응답 지시행렬 R에 대한 모형인 $f(R|Y,\phi)$도 포함해야 하는 것이다. 문제는 무응답을 포함한 자료에서 자료행렬 Y 전체가 아닌 응답된 부분인 Y_{obs}만을 실제로 관측할 수 있다는 점에서 더 복잡하다. 즉, 실제로 응답된 자료에 근거한 결합분포는

$$f(Y_{obs}, R|\theta,\phi)$$

이며 이는 자료행렬 $Y=(Y_{obs}, Y_{mis})$와 응답 지시행렬 R의 결합 분포함수인 $f(Y, R|\theta,\phi)$에서 Y_{mis}를 적분(integrating out)하여 구할 수 있다. 즉,

$$\begin{aligned} f(Y_{obs}, R|\theta,\phi) &= \int f(Y, R|\theta,\phi) dY_{mis} \\ &= \int f(Y_{obs}, Y_{mis}, R|\theta,\phi) dY_{mis} \\ &= \int f(Y_{obs}, Y_{mis}|\theta) f(R|Y_{obs}, Y_{mis},\phi) dY_{mis} \end{aligned}$$

으로 표현할 수 있다. 무응답 자료 메커니즘이 임의결측을 따르는 경우 응답된 자료에 근거한 결합분포는

$$\begin{aligned} f(Y_{obs}, R|\theta,\phi) &= \int f(Y_{obs}, Y_{mis}|\theta) f(R|Y_{obs}, Y_{mis},\phi) dY_{mis} \\ &= \int f(Y_{obs}, Y_{mis}|\theta) f(R|Y_{obs},\phi) dY_{mis} \\ &= f(R|Y_{obs},\phi) \int f(Y_{obs}, Y_{mis}|\theta) dY_{mis} \\ &= f(R|Y_{obs},\phi) f(Y_{obs}|\theta) \end{aligned}$$

으로 표현할 수 있다. 즉, Y_{obs}와 M의 결합분포는 각각의 분포의 곱(product)으로 나타낼 수 있다. 무응답 자료 메커니즘이 완전임의결측을 따르는 경우에도 비슷한 방법으로

$$f(Y_{obs}, R|\theta,\phi) = f(R|\phi) f(Y_{obs}|\theta)$$

와 같이 나타낼 수 있다.

만약 (θ,ϕ)의 결합모수공간이 각각의 모수공간의 곱으로 표현될 수 있다면, 즉, $\Omega_{\theta,\phi}=\Omega_\theta\times\Omega_\phi$으로 나타낼 수 있다면 모수 θ와 ϕ가 별개(distinct)라고 부른다. 직관적 의미는 모수 θ의 값에 대한 정보가 주어져도 모수 ϕ에 대한 정보에 영향이 없으며 그 반대도 참이라는 뜻으로 이 가정은 많은 실제 자료에서 만족된다. 이 가정이 만족되고 무응답 자료 메커니즘이 완전임의결측 또는 임의결측이라면 관심 모수인 θ에 대한 추론(inference)을 실시하고자 할 때 $f(R|\phi)$ 부분은 연관되어 있지 않으므로 $f(Y_{obs}|\theta)$만에 근거하여 분석을 실시할 수 있다. 즉, θ에 대한 추론을 할 때 $f(R|,\phi)$을 무시할 수 있다는 점에서 무응답 자료 메커니즘은 무시할 수 있는 무응답 자료 메커니즘(ignorable missing data mechanism)이라 부른다. 요약하면, (1) 무응답 자료 메커니즘이 임의결측(완전임의결측을 포함)을 따르고 (2) 자료의 분포와 관련된 모수 θ와 응답 지시행렬과 관련된 모수 ϕ가 서로 별개(distinct)라면 무응답 자료 메커니즘에 관한 $f(R|Y_{obs},\phi)$은 관심 모수인 θ에 대한 추론을 시행할 때 무시할 수 있다는 의미이다.[1)]

무응답 자료 메커니즘이 완전임의결측이나 임의결측을 따르는 지 또는 비임의결측을 따르는 지 결정하는 것은 쉬운 일이 아니다. 무응답 자료 메커니즘이 완전임의결측을 따르는 지 알아 보기 위한 가장 쉬운 방법은 관측된 자료와 무응답 자료들 사이에서 완전히 응답된 다른 변수들에 대한 분포를 비교해 보는 것이다. 예를 들어, 소득에 대한 무응답이 발생한 경우 소득에 대한 응답자 집단과 무응답자 집단 사이의 성별, 연령, 교육 등 변수를 비교해 두 집단 사이에 유의한(significant) 차이가 없다면 무응답 자료 메커니즘은 완전임의결측을 따른다고 할 수 있다. 이 가정은 무응답이 완전히 무작위로 발생하였으므로 응답자 집단과 무응답자 집단은 완전히 랜덤하게 나뉘어져 두 집단 사이의 분포가 동일해야 한다

1) 무시할 수 있는 무응답 자료 메커니즘을 만족하기 위하여 필요한 두 개의 조건 중 두 번째 조건은 대부분의 예제에서 만족되므로 첫 번째 조건인 임의결측이나 완전임의결측이 만족되는지 여부가 더 중요시된다. 따라서 일부 문헌에서는 첫 번째 조건인 임의결측이 무시할 수 있는 무응답 자료 메커니즘과 동일한 것처럼 나타난다.

는 점에서 직관적이다. 물론 실제 자료에서는 관심인 변수가 소득 한 개가 아니라 여러 개로 구성되고 무응답 발생 형태도 <그림 1.3.(4)>의 일반적인 형태를 따르는 경우가 많고 각 변수별 응답자 집단과 무응답자 집단의 직접적 비교는 너무 많은 숫자의 비교를 요구한다. Little(1988a)은 일반적인 패턴의 무응답 자료에서 무응답 자료 메커니즘이 완전임의결측인지 검정하는 방법을 제안하였다. 한편, 무응답 자료 메커니즘이 임의결측을 따르는 지에 대하여는 일부 연구가 진행되었으나 자료의 모형의 잘못된 지정(misspecification)에 대한 민감성(sensitivity) 문제 때문에 일반적으로 사용되지 않는다.

대부분의 무응답 처리 기법은 무응답 자료 메커니즘이 임의결측이라 가정한다. 이는 무응답 자료 메커니즘이 임의결측이라면 자료의 분포와 관련된 모수 θ와 응답 지시행렬과 관련된 모수 ϕ가 서로 별개(distinct)라고 가정한 후 무응답 발생원인은 무시할 수 있는 무응답 자료 메커니즘이라 가정하여 응답 지시행렬 R에 대한 모형을 포함하지 않고 관측된 자료 $f(Y_{obs}|\theta)$에만 근거하여 관심 모수 θ에 대한 추론을 실시할 수 있기 때문이다. 따라서 이런 무응답 처리 기법을 적용하기 위해서는 무응답 자료 메커니즘이 임의결측이 되도록 무응답 발생과 연관된 변수들을 자료에 포함시켜 분석을 시행해야 한다는 점을 유의해야 한다. 하지만 무응답 자료 메커니즘을 결정하기 어려운 이유는 이 메커니즘이 자료가 어떤 변수를 포함하고 있는지에 따라 바뀔 수 있기 때문이다. 예를 들면 가계동향조사에서 가구의 수입에 대한 응답 여부가 수입이 매우 높거나 없기 때문에 발생하는 경우는 비임의결측에 해당되지만 수입과 밀접하게 관련되어 있는 변수인 세금, 자산, 지출 등의 정보가 알려져 있다면 이 정보를 포함시킴으로써 무응답 자료 메커니즘이 임의결측으로 바뀔 수 있는 것이다. 이런 이유 때문에 무응답 자료에 대한 분석에서는 무응답 자료 메커니즘이 비임의결측보다는 임의결측이 되도록 무응답의 발생과 연관된 변수들을 가능한 한 모두 분석에 포함하는 것이 매우 중요하다.

< 1장 연습문제 >

1. 다음의 각 항목에 대하여 참, 거짓을 판별하고 판별의 이유를 설명하시오.

(1) 무응답 자료 메커니즘에 따라 무응답을 포함한 자료의 분석 방법은 달라져야 한다.

(2) 무응답 자료 메커니즘이 임의결측이면 무응답 자료 메커니즘은 무시할 수 있는(ignorable)경우에 해당된다.

(3) 무응답 자료 메커니즘이 완전임의결측이면 무응답 자료 메커니즘은 무시할 수 있는(ignorable)경우에 해당된다.

(4) 한 변수에 대한 무응답 자료 메커니즘은 다른 변수와 상관없이 결정된다.

2. 다음의 예에 대하여 무응답 자료 메커니즘을 판정하고 판정의 이유를 제시하시오.

(1) 인구주택총조사에서 가구에 거주하는 모든 사람은 전수조사(short form) 대상이다. 이 중 10%의 표본을 뽑아 좀 더 자세한 문항에 대한 조사(long form)를 실시한다. 이 때 표본으로 뽑히지 않은 90%는 자세한 문항에 대한 응답을 실시한 적이 없으므로 이 문항들에 대하여 모두 무응답으로 표시된다.

(2) 20-25세 남자의 평균 키를 추정하고자 한다. 이를 위하여 1000명의 군인(일반병사)을 대상으로 키를 측정하였다.

(3) 어떤 연구에서 몸무게를 측정하였는데 남자 중 10%가, 여자 중 30%가 응답하지 않았다. 남자들 중에서 몸무게에 대한 응답 확률은 동일하고 여자 중에서도 응답 확률이 동일하다.

(4) 연소득에 대한 정확한 응답을 얻기 위하여 두 가지 질문이 사용되었다. 첫 번째는 월소득에 관한 질문이며 두 번째는 연소득에 관한 질문이었다.

월소득에 대한 응답에 12배를 하여 구한 값이 연소득과 같지 않은 경우에 연소득에 대한 확인 질문이 주어졌다. 연소득에 관한 확인 질문 항목에서 무응답이 나타난다.

(5) 새로 개발된 해열제의 효과를 연구하는 임상실험에서 기존 해열제 또는 새로 개발된 해열제를 투여 받은 두 고열환자 집단의 체온이 일주일간 매일 측정되었다. 하지만 일부 환자에서 체온 변화와는 상관없지만 위장장애 등 부작용이 나타났고 중간에 실험에서 제외되어 추후 체온은 결측으로 나타났다.

제 2장 여러 가지 무응답 분석 방법

<학습목표>

(1) 단순한 방법인 완전히 응답한 개체 분석법과 이용 가능한 개체 분석법에 관하여 설명한다.

(2) 여러 가지 가중값 방법에 관하여 고찰한다.

(3) 대체 방법의 기초에 관하여 설명한다.

(4) 우도방법의 기초와 무응답 자료에의 적용을 이해한다.

(5) 무응답이 있는 경우 MLE 추정법인 EM 알고리듬에 관하여 설명한다.

2.1 완전히 응답한 개체를 이용한 분석 (Complete-case Analysis)

무응답 자료 분석에서 가장 흔하게 사용되며 대부분의 통계프로그램에서 디폴트로 사용되는 방법은 완전히 응답한 개체를 이용한 방법이다. 이 방법은 모든 변수에 응답이 있는 자료만을 사용하는 방법으로 한 개체에서 어떤 한 변수만이라도 무응답이 있다면 그 개체는 분석에서 제외한다. 이 방법은 일반적인 통계방법을 사용할 수 있으므로 매우 쉬우며 공통적인 표본의 기저(sample base)를 이용하므로 단순 통계량들의 비교가 용이하다. 하지만 이 방법은 무응답의 메커니즘이 MCAR이 아닌 경우 결과에 편향(bias)이 발생하며 부분 정보를 담은 개체들도 분석에서 제외하므로 정보의 손실에 의하여 정밀도(precision)가 낮아져 검정력의 약화를 야기한다.

2.2 가중값 보정방법 (Weighting Adjustment)

이 방법은 완전히 응답한 개체에 가중값을 주어 편향을 보정하는 방법이다. 기본적인 아이디어는 유한한 모집단을 대상으로 한 조사(finite population survey)의 확률화 추론 (randomization inference)을 위해 가중값을 이용하는 방법과 매우 유사하다. 먼저 무응답이 없는 조사에서 추론을 하는 과정을 단순화 시켜서 보도록 하자.

유한 모집단 U는 N개의 개체를 포함하고 있다고 하자. 이 개체의 주변수 Y는 $Y_1, Y_2, \cdots, Y_N$으로 표시할 수 있다. 이제 표본조사의 목적을 Y의 모집단 평균인

$$\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

의 추정에 있다고 하자. 물론 모집단의 다른 모수에도 관심이 있을 수 있다.

크기가 n인 표본을 모집단으로부터 확률추출법(probability sampling)을 이용하여 추출하였다. 즉, 모집단으로부터 개체 i가 추출될 확률은 π_i이며 모집단의 π_i^{-1} 개체를 대표한다. (추출확률은 일반적으로 알려져 있다.) 그러므로 이 개체는 모집단의 모수 추정에 있어 π_i^{-1}의 가중값을 주어야 한다. 예를 들어 모집단의 Y의 합 T는 다음과 같은 가중합(weighted sum)으로 추정할 수 있다.

$$\widehat{T_{HT}} = \sum_{i=1}^{n} y_i \pi_i^{-1}.$$

위의 추정값을 호빗-톰슨 추정값(Horvitz-Thompson estimator: Horvitz and

Thomson, 1952)라고 부른다. 이 때 모집단의 Y의 평균은 다음과 같다.

$$\bar{y}_W = \frac{1}{n}\sum_{i=1}^{n} w_i y_i.$$

이 때 가중값 $w_i = n\pi_i^{-1} / \sum_{k=1}^{n} \pi_k^{-1}$이다. 무응답이 없는 경우 이 가중 평균 $\bar{y}_W$는 Y의 평균의 불편추정량이다. 만일 무응답이 있는 경우에는 이 가중평균을 확장할 수 있다. 만일 개체 i의 응답확률(또는 응답성향: response propensity)을 ϕ_i라고 하면

Pr(selection and response) = Pr(selection) × Pr(response|selection) = $\pi_i\phi_i$

이고 이 경우의 가중평균은 다음과 같다.

$$\bar{y}_W = \frac{1}{r}\sum_{i=1}^{r} w_i y_i$$

여기서 r은 응답한 개체의 수이며 $w_i = r(\pi_i\phi_i)^{-1} / \sum_{k=1}^{r} (\pi_k\phi_k)^{-1}$이다. 일반적으로 응답성향 ϕ_i는 알려져 있지 않으므로 응답자들과 무응답자들의 정보를 이용하여 추정하여야 한다.

2.2.1 평균의 가중 클래스 추정법

먼저 표본을 응답자와 무응답자 모두 이용 가능한 변수를 바탕으로 J개의 가중

클래스로 나누었다고 가정하자. 이 때, C를 이러한 가중 클래스 변수라고 하자. 만일 가중 클래스 변수 $C=j$ 인 경우에 n_j는 j 클래스의 표본수이고 r_j는 j 클래스의 응답자수이면 이 j 클래스안의 개체들의 응답확률은 단순히 r_j/n_j로 추정될 수 있다. 이런 경우, 가중 클래스 j 안의 응답자들은 다음과 같은 가중값을 가지게 된다.

$$w_i = r(\pi_i \hat{\phi_i})^{-1} / \sum_{k=1}^{r} (\pi_k \hat{\phi_k})^{-1}$$

이 때 클래스 j 안의 개체 i의 응답확률 추정값 $\hat{\phi_i} = r_j/n_j$이다. 임의추출 표본인 경우 (즉, π_i가 상수인 경우), 위의 가중평균은 좀 더 단순화된다.

즉,

$$\overline{y_{wc}} = \frac{1}{n} \sum_{j=1}^{J} n_j \overline{y_{jR}}.$$

여기서 $\bar{y}_{jR}$는 클래스 j 안의 응답자들의 평균이고 $n = \sum_{j=1}^{J} n_j$는 총 표본수이다. 가중 클래스 j 안의 응답자들이 표본으로 선택된 개체의 임의표본(random sample)인 경우 (즉, MAR 가정)에 위의 추정값은 불편성을 갖는다. Oh와 Scheuren (1983)은 위의 추정값의 분산을 다음과 같이 구하였다.

$$Var(\overline{y_{wc}}) = \sum_{j=1}^{J} \left(\frac{n_j}{n} \right)^2 f_j S_j^2$$

여기서 S_j^2은 클래스 j의 Y의 모분산이고 N_j는 클래스 j의 모집단의 수이고 $f_j = 1/r_j - 1/N_j$는 유한모집단 보정(finite population correction)에 해당한다.

2.2.2 응답성향을 이용한 가중값 방법

X 를 응답자와 무응답자 모두 이용 가능한 변수의 집합이라고 하자. 만일 이러한 X 안의 변수의 수가 적은 경우는 위에서 고려한 가중 클래스 추정방법을 사용할 수 있다. 하지만 패널표본조사와 같은 경우에는 현 시점 조사에서 발생한 무응답자들에 대한 과거 시점의 많은 정보를 이용할 수도 있다. 이런 경우에 가중 클래스 추정방법을 사용하면 모든 관측된 변수들을 이용하여 클래스를 만들어야 하는데 그 조합을 모두 고려하게 되면 클래스의 수가 너무 커져서 각 클래스 안의 무응답 가중값이 무한해지는 현실적인 문제가 발생할 수 있다. 이런 경우는 모든 기록된 변수들을 다 이용하는 것이 아니라 이런 다변량 변수들을 응답성향(response propensity)이라는 하나의 변수로 차원을 축약하여 가중 클래스 변수로 사용할 수 있다. 원래 성향점수(propensity score) 방법은 두 군을 비교하는 관찰연구에서 다변량 교란변수들을 일변량 성향점수로 축소하여 짝짓기 등의 방법을 통하여 교란효과(confounding effects)를 보정하는 인과추론(causal inference) 방법으로 Rosenbaum과 Rubin (1983, 1985)이 소개하였다. 이 때 성향점수는 두 군 중 한 군에 할당 또는 포함될 확률로 정의된다. 이 성향점수 방법은 무응답 자료분석에서도 적용할 수 있다.

먼저 R을 응답 지시 변수라고 하자. 이 때, 무응답이 MAR이라고 가정하면 $\Pr(R|X, Y) = \Pr(R|X)$이다. 왜냐하면, 무응답은 단지 Y에서만 발생하였기 때문이다. 이제 개체 i에 대하여 응답성향을 다음과 같이 정의한다.

$$p(x_i) = \Pr(r_i = 1|x_i)$$

그러면 모든 x_i에 대하여

$$\begin{aligned}
\Pr(r_i = 1|y_i, p(x_i)) &= E\big[\Pr(r_i = 1 \,|\, y_i, x_i) \,|\, y_i, p(x_i)\big] \\
&= E\big[\Pr(r_i = 1 \,|\, x_i) \,|\, y_i, p(x_i)\big] \;\; by\, MAR \\
&= E\big[p(x_i) \,|\, y_i, p(x_i)\big] \\
&= p(x_i)
\end{aligned}$$

이다. 그러므로 $\Pr[R \,|\, p(X), Y] = \Pr[R \,|\, p(X)]$. 즉, 응답성향점수인 $P(X)$로 정의된 층들(strata) 안에서는 응답자들은 임의 부표본(random subsample)이 된다.

이 응답성향을 이용한 가중 방법은 다음과 같은 순서로 요약할 수 있다.

1) 먼저 변환된 응답성향 변수 $p(X)$는 미지의 값이므로 표본으로부터 추정한다. 많이 사용되는 추정방법으로 응답 지시변수를 종속변수로 하고 X를 독립변수로 하여 로지스틱(logistic) 또는 프로빗(probit) 회귀분석을 이용한다.
2) 다음으로 추정된 $p(X)$를 순서대로 5개나 6개의 값으로 묶은 하나의 집단변수를 생성한다.
3) 이제 가중 클래스 변수인 C를 이 집단변수라고 하면 가중 클래스 추정방법을 이용하여 추정값을 구할 수 있다. 만일 X가 단변수라면 응답성향을 이용한 방법은 X를 이용한 가중 클래스 추정방법과 동일하다.

응답성향을 이용한 가중 방법에서 클래스를 이용하는 대신 응답개체 i의 가중값

을 추정된 응답성향점수의 역수인 $\hat{p}(x_i)^{-1}$로 직접 가중값을 줄 수 있는 방법도 있다. 또한, 응답성향은 대체 방법에도 사용될 수 있다. 무응답 자료가 MAR을 가정하는 경우, Y와 $p(x_i)$와의 관계를 회귀모형으로 하여 응답자들의 자료로 적합한 후 무응답자들의 $p(x_i)$를 적합식에 대입하여 Y의 예측값을 구한 후 무응답을 이 예측값으로 대체한다. (Little and An, 2004, 2008)

응답성향을 이용한 가중 방법의 문제점으로는 매우 작은 응답성향 추정값을 갖는 응답개체는 매우 큰 가중값을 가지게 되며 이는 평균과 총합의 추정값에 과도한 영향을 미치게 되며 그 결과로 극도로 높은 분산을 가진 추정값을 제공할 수도 있다. 또한 응답성향점수의 역수인 $\hat{p}(x_i)^{-1}$로 직접 가중값을 주는 방법을 사용하는 경우는 응답성향 가중 클래스 방법보다 응답성향을 추정하는 모형에 좀 더 민감하다. 즉, 응답성향을 추정하는 모형이 맞지 않는 경우 큰 편향(bias)을 초래할 수 있다.

응답성향을 이용한 가중방법의 대안으로 응답성향을 이용한 대체방법을 고려할 수 있다. 이 방법은 응답성향과 결과변수 Y와의 관계를 회귀모형화한 후 응답자들의 자료를 이용하여 회귀식을 추정하고 무응답자들의 응답성향을 이용하여 Y의 예측값을 얻은 후 무응답을 대체하게 된다. (Little and An 2004, An and Little 2008)

2.2.3 무응답 가중값 방법에서 분산의 증가

가중 클래스 방법은 조사 응답 변수 Y에 관련없이 같은 가중값을 얻을 수 있으므로 매우 쉽게 적용될 수 있다. 그러므로 매우 큰 표본조사에서 무응답이 MAR

이고 응답변수의 수가 많은 경우 하나의 가중 집합으로 편향을 다룰 수 있다. 하지만 가중 클래스 방법은 적용이 쉬운 반면 분산이 증가하게 되는 단점이 있다. 가중 클래스 안에서 임의 표본을 가정하고 가중값에 표본변동이 없고 반응변수 Y의 분산이 σ^2으로 상수인 경우 표본 평균의 분산의 증가 정도는 다음과 같다.

$$Var\left(\frac{1}{r}\sum_{i=1}^{r} w_i y_i\right) = \frac{\sigma^2}{r^2}\left(\sum_{i=1}^{r} w_i^2\right) = \frac{\sigma^2}{r}\left[1 + cv(w_i)^2\right]$$

여기서 가중값들은 합이 1이 되도록 척도화 되었고 $cv(w_i)$는 가중값의 변동계수(coefficient of variation)이다.

이 식에서 변동계수의 제곱부분은 가중방법으로 발생하는 분산의 증가비율을 반영한다. 분산의 증가는 무응답과 매우 관련이 있는 변수들의 편향을 줄이는 비용으로 발생한다고 그 정당성을 이야기 할 수 있으나 무응답과 관련이 없거나 약한 관련성을 가지는 변수들은 편향을 줄이는 데 큰 도움이 되지 않으면서 분산만 증가시키게 된다.

2.2.4 알려진 주변(margins)에 대한 사후-층화(post stratification)와 레이크(rake) 방법

가중클래스 방법을 이용하여 모수의 추정값을 계산할 때 가중 클래스 j안의 모집단 비율 N_j/N은 표본 비율 n_j/n으로 추정된다. 하지만 어떤 경우에는 모집단 비율 N_j/N가 외부의 출처로부터 이용 가능할 수도 있다. 이런 경우에 가중을 이용한 무응답 분석방법에 관하여 알아보자.

2.2.4.1 사후-층화 (Post-stratification)

모집단 비율 N_j/N가 외부의 출처로부터 알려져 있다고 가정하자. 이런 경우 가중클래스를 이용한 평균추정의 대안은 다음과 같은 사후-층화 평균이다.

$$\bar{y}_{ps} = \frac{1}{N}\sum_{j=1}^{J} N_j \bar{y}_{jR}$$

무응답의 메커니즘이 MAR인 경우 $\bar{y}_{ps}$는 $\overline{Y}$의 불편추정값이고 그 분산은 다음과 같다.

$$Var(\bar{y}_{ps}) = \frac{1}{N^2}\sum N_j^2 \left(1 - \frac{r_j}{N_j}\right)\frac{S_{jR}^2}{r_j}$$

이 분산의 추정값은 위 식에서 모집단 분산인 S_{jR}^2를 클래스 j의 응답자들의 표본분산 s_{jR}^2로 대체하여 구할 수 있다. 대부분의 경우 $\bar{y}_{ps}$는 $\bar{y}_{wc}$보다 분산이 작다. 하지만 클래스 j안의 응답자의 표본수 r_j 와 Y의 클래스 간 분산이 작은 경우는 $\bar{y}_{wc}$의 분산이 $\bar{y}_{ps}$의 분산보다 클 수 있다.

2.2.4.2 레이킹 비율 추정방법 (Raking Ratio Estimation)

두 개의 교차-분류 인자(cross-classifying factors) X_1과 X_2의 결합 수준(joint levels)을 이용하여 가중 클래스를 정한다고 가정하자. $X_1 = j,\ \ X_2 = l$ $(j = 1,\cdots,J,\ \ l = 1,\cdots,L)$을 가진 클래스 안에서 N_{jl} 모집단 개체 가운데 n_{jl} 개체

가 표본추출 되었고 이 중에 r_{jl} 표본개체에서 Y변수의 값이 조사되었고 $(n_{jl}-r_{jl})$ 개체에서 무응답이 발생했다. 이 경우 사후층화 추정값과 가중 클래스 추정값은 다음과 같다.

$$\bar{y}_{ps} = \frac{1}{N}\sum_{j=1}^{J}\sum_{l=1}^{L} N_{jl}\bar{y}_{jlR}$$

$$\bar{y}_{wc} = \frac{1}{n}\sum_{j=1}^{J}\sum_{l=1}^{L} n_{jl}\bar{y}_{jlR}$$

여기서 $\bar{y}_{jlR}$은 $X_1 = j,\ X_2 = l$을 가진 클래스 안에서 응답한 개체들의 평균이다. X_1과 X_2의 주변계수(marginal counts)인 $N_{j+} = \sum_{l=1}^{L} N_{jl}$과 $N_{+l} = \sum_{j=1}^{J} N_{jl}$가 외부의 출처로부터 알려져 있을 때 추정값은 셀 안의 응답자 평균에 근거하여 구할 수 있다. X_1=성별이고 X_2=인종이라고 할 때 성별과 인종의 주변분포는 알려져 있으나 인종과 성별의 결합분포는 알려져 있지 않은 경우가 하나의 예가 될 수 있다.

행과 열 변수에서 관찰된 클래스 계수(class counts)인 $\{n_{jl}\}$에 적용되는 레이킹 방법은 $\{N_{jl}\}$의 추정값인 $\{N_{jl}^{*}\}$를 구하는 방법이다. 이 때, 레이킹 방법은 다음과 같은 주변 제약(marginal constraints)을 만족하여야 한다.

$$N_{j+}^{*} = \sum_{l=1}^{L} N_{jl}^{*} = N_{j+}, \qquad j = 1,\cdots,J:$$

$$N_{+l}^{*} = \sum_{j=1}^{J} N_{jl}^{*} = N_{+l}, \qquad l = 1,\cdots,L.$$

$\{N_{jl}^*\}$는 관찰된 클래스 계수(class counts)인 $\{n_{jl}\}$와는 다르다. 즉, 행의 상수 $\{a_j, j=1,\cdots,J\}$와 열의 상수 $\{b_l, l=1,\cdots,L\}$에 대해 $N_{jl}^* = a_j b_l n_{jl}$이다. $\{N_{jl}^*\}$ 분할표의 주변은 알려진 주변 합인 $\{N_{j+}\}$와 $\{N_{+l}\}$와 같다. 하지만 $\{N_{jl}^*\}$ 분할표 안의 행과 열의 관계는 $\{n_{jl}\}$ 분할표 안에서의 행과 열의 관계와 같다. 레이크 표본 계수 $\{N_{jl}^*\}$는 반복적 비율 적합 절차(iterative proportional fitting procedure)를 이용하여 구할 수 있다. 이 방법은 현재의 추정값을 주변 합 $\{N_{j+}\}$와 $\{N_{+l}\}$에 일치시키기 위하여 행과 열 변수로 스케일링하게 된다. 즉, 첫 번째 단계에서 추정값은 다음처럼 행 주변 합 $\{N_{j+}\}$과 일치하여 구한다.

$$N_{jl}^{(1)} = n_{jl}(N_{j+}/n_{j+}).$$

다음으로 추정값은 열 주변 합 $\{N_{+l}\}$와 일치시켜 구한다.

$$N_{jl}^{(2)} = N_{jl}^{(1)}\left(N_{+l}/N_{+l}^{(1)}\right)$$

그리고 추정값은 다시 다음과 같이 수렴할 때 까지 계속해서 갱신한다.

$$N_{jl}^{(3)} = N_{jl}^{(2)}\left(N_{j+}/N_{j+}^{(2)}\right) \ \cdots$$

Ireland 와 Kullback(1968)은 모집단 클래스 비율의 레이크 추정값 $\{N_{jl}^*/N\}$는 클래스 계수인 $\{n_{jl}\}$이 다항분포를 따른다고 가정할 경우 점근적으로 정규분포를 따르게 되며 또한 다항분포 모형 하에서의 최대우도 추정값과 점근적으로 동일함을 보였다.

레이크 표본 계수들인 $\{N_{jl}^*\}$와 응답자의 평균들인 $\{\overline{y}_{jl}\}$를 결합하면 $\overline{Y}$의 레이크

추정값은 아래와 같이 표현된다.

$$\bar{y}_{rake} = \frac{1}{N}\sum_{j=1}^{J}\sum_{l=1}^{L} N_{jl}^{*}\,\overline{y_{jR}}$$

여기서 $r_{jl}=0$이고 $n_{jl}\neq 0$이면 이 추정값은 정의되지 않는다. 이 경우는 그 클래스의 평균의 다른 추정값을 고려하여야 한다.

2.2.5 알려진 주변(margins)에 대한 선형 가중방법(linear weighting)

2.2.5.1 일반 회귀 추정

무응답이 없을 때, 적절한 보조 정보(auxiliary information)를 이용하면 단순 표본평균의 정밀도는 향상된다. p개의 보조변수가 있다고 가정하여 보자. i번째 표본의 보조변수의 값을 $X_{i1}, X_{i2}, ..., X_{ip}$라고 하고 모평균의 벡터는 $\overline{X}$라고 하자. 만일 보조변수들이 주변수와 상관이 있으면 Y를 X에 회귀시켜 얻은 회귀계수 $B=(B_1, ..., B_k)^T$에 대하여 잔차 $E_i = Y_i - X_iB$는 주변수 그 자체값보다도 변동이 더 작다. 무응답이 없는 경우, 회귀계수 B는 보통 최소제곱법으로 다음과 같이 추정할 수 있다.

$$b = \left(\sum_{i=1}^{n} x_i x_i^T\right)^{-1}\left(\sum_{i=1}^{n} x_i y_i\right)$$

여기서 x_i와 y_i는 각 각 보조변수와 주변수의 표본값이다.
무응답이 없는 경우 위의 회귀계수 추정값을 이용한 일반 회귀 추정값은 다음과

같다.

$$\bar{y}_{REG} = \overline{X}^T b$$

무응답이 있는 경우에 수정 일반 회귀 추정값은

$$\bar{y}^*_{REG} = \overline{X}^T b^*$$

이다. 여기서 b^*는 응답표본을 이용하여 추정한 회귀계수이다. 추정값 $\bar{y}^*_{REG}$의 편향(bias)은

$$bias = \overline{X}^T B^* - \overline{Y}$$

이다. 여기서

$$B^* = \left(\sum_{i=1}^{N} p_i X_i X_i^T \right)^{-1} \left(\sum_{i=1}^{N} p_i X_i Y_i \right)$$

이고 p_i는 무응답 확률이다. 만일 $B^* = B$이면 회귀 추정값의 편향은 0이다. 그러므로 무응답이 회귀계수에 영향을 미치지 않는 이상 회귀 추정값은 불편 추정값이다. 즉, 무응답이 메커니즘이 MAR이고 주변수와 관련성이 높은 보조변수가 회귀 모형에 포함된 경우 편향은 매우 작다.

2.2.5.2 범주형 보조변수를 이용한 선형 가중방법

Bethlehem 와 Keller (1987)는 일반 회귀 추정값이 다음과 같은 가중 추정값의 형태로 표현될 수 있음을 보였다.

$$\bar{y}_w = r^{-1}\sum_{i=1}^{r} w_i y_i$$

여기서 응답자 i의 가중값은 $w_i = \nu^T X_i$ 이고 ν는 다음과 같은 가중계수의 벡터이다.

$$\nu = r\left(\sum_{i=1}^{r} x_i x_i^T\right)^{-1}.$$

사후-층화 방법은 선형 가중방법의 특별한 경우이다. 먼저 범주형 보조변수를 가변수(dummy variables)로 만든다. L수준을 가진 하나의 범주형 보조변수가 있다고 가정하자. 먼저 L개의 가변수 $X_1, X_2, ..., X_L$을 만든다. 보조변수의 값이 l번째 $(l = 1, ..., L)$ 층(stratum)에 속하면 $X_l = 1$로 그렇지 않으면 $X_l = 0$으로 정의한다. 이 가변수의 모평균 벡터는

$$\bar{X} = \left(\frac{N_1}{N}, \frac{N_2}{N}, ..., \frac{N_L}{N}\right)^T$$

이고

$$\nu = \left(\frac{n}{N}\right)\left(\frac{N_1}{N}, \frac{N_2}{N}, ..., \frac{N_L}{N}\right)^T$$

이다.

2.2.6 무응답 가중 추정값의 추론

가중 추정값은 일반적으로 계산하기가 쉬운 경우가 많다. 하지만 추정값의 적절한 분산의 계산은 점근적(asymptotic)으로도 어려운 경우가 많다. 단순임의추출인 경우의 분산계산 공식은 위에서 본 것과 같이 닫힌 형태로 제시되어 있다. 좀 더 복잡한 표본추출을 사용한 경우에는 테일러 급수 확장 (Taylor series expansions: Robins, Rotnitsky and Zhao, 1995), 붓스트랩(bootstrap) 또는 잭나이프(jackknife) 등의 방법을 이용할 수 있다.

몇 몇 상용화 프로그램(예를 들면 SUDDAN)에서는 복잡한 표본추출설계에서의 추정값의 표준오차를 계산하여 주기도 하지만 이런 프로그램들은 가중값은 고정되고 알려져 있다고 가정하는 경우가 일반적이다. 하지만 무응답 가중값은 관찰된 자료로부터 추정되어지므로 추정된 가중값은 표본 불확실성 (sampling uncertainty)을 가지고 있다. 이런 불확실성으로 생기는 추가적인 변동이 표준오차의 계산에서 고려되어야 한다. 이런 추가적인 변동은 표본을 이용한 반복적인 재표본 추출방법(repeated resampling method)인 붓스트랩이나 잭나이프 방법 등을 이용하여 고려될 수 있다. 이런 재표본 추출방법은 집중적인 계산을 요구한다.

가중방법은 각 개체에서 얻어진 모든 변수에서 같은 가중값을 이용하고 완전히 이용할 수 있는 자료를 사용하여 추정값의 편향을 줄여 주므로 비교적 쉬운 방법이다. 하지만 이 방법은 부분 정보를 가지고 있는 무응답 개체들을 분석에서 제외

하므로 추정값의 분산이 증가하게 되며 이 분산의 조정이 쉽지 않다. 그러므로 가중방법은 무응답 보정을 위한 공변수의 수가 작고 표본의 수가 커서 분산 보다는 편향의 보정이 중요한 경우에 가장 유용하게 사용될 수 있는 방법이다.

2.3 이용 가능한 개체 분석 (Available-case Analysis)

완전히 응답한 자료를 이용한 분석은 평균 또는 주변빈도분포의 추정 등 일변량 분석에서는 정보의 손실이 너무 커서 좋지 않다. 특히 많은 변수들을 가진 자료에서 결측이 일어나게 되면 정보의 손실이 매우 크다. 예를 들면, 만일 한 자료에 20개의 변수가 있고 각 변수에서 독립적으로 10%의 확률로 결측이 발생하게 되면, 완전히 응답한 개체의 기대 비율은 $0.90^{20} \simeq 0.12$이 된다. 이러한 정보의 손실을 보완하는 방법으로 이용 가능한 개체 분석법이 있다.

이 방법은 각 분석 단계에서 이용가능한 모든 자료를 사용한다. 완전히 이용 가능한 자료 분석방법보다 더 많은 자료를 이용하므로 언뜻 보기에 매력적인 방법으로 생각될 수 있으나 이 방법은 장점보다 단점이 더 많은 방법으로 현실적으로는 추천되지 않는다. 가장 큰 단점으로는 표본의 기저(sample base)가 결측의 패턴에 따라 변수별로 달라진다. 예를 들어 일반적인 결측 패턴을 가진 세 변수, Y_1, Y_2, Y_3를 이용하여 변수 간의 상관계수를 구한다고 하자. 이런 경우 Y_1과 Y_2, Y_1과 Y_3, Y_2와 Y_3의 상관계수를 구할 때 마다 표본의 기저는 다르게 된다. 이렇게 구한 상관행렬은 양정치행렬(positive definite matrix)이 아닐 수도 있게 된다. 또한 추정값은 쉽게 구할 수 있으나 그 추정값의 표본오차는 대표본적으로도 매우 복잡한 경우가 많다.

2.4 대체방법 (Imputation Methods)

Y_j의 주변분포를 추정하거나 Y_j와 다른 변수간의 상관계수를 구할 때 완전히 이용가능한 개체 분석법이나 이용가능한 개체 분석법 모두 Y_j 내의 무응답 개체는 분석에서 제외한다. 하지만 만일 무응답이 있는 변수 Y_j와 다른 변수 Y_k가 서로 높은 상관이 있는 경우에는 Y_k의 정보를 이용하여 Y_j의 무응답을 예측하고 Y_j의 무응답을 그 예측값으로 대체하는 방법도 고려할 수 있다. 이렇게 무응답값을 통계적 모형을 통하여 어떤 다른 값으로 채우는 것을 대체방법(imputation methods)이라고 한다. 만일 무응답값을 하나의 값으로 채워 하나의 대체된 데이터셋을 만드는 방법을 단일 대체(single imputation)라고 한다. 단일 대체방법은 결측된 값을 마치 실제로 관찰된 값으로 생각하고 자료를 분석하게 되므로 대체로 발생하는 불확실성(uncertainty), 즉 결측으로 발생하는 불확실성을 고려하지 못하므로 표준오차의 추정값이 과소추정(underestimate)되어 p-값이 실제보다 작아지고 신뢰구간이 실제보다 좁아지는 문제 등이 발생한다. 이러한 결측으로 발생하는 불확실성을 고려하여 편향이 없는 표준오차의 추정값을 구하는 방법으로는 반복적인 재표본 방법(resampling method)과 다중대체(multiple imputation) 방법이 있다. 이 절에서는 몇 가지 명백한 모형을 바탕으로 하는 단일 대체방법과 결측으로 발생하는 불확실성을 고려하는 방법인 재표본 방법과 다중대체 방법에 관하여 간략하게 소개한다. 대체방법에 관한 좀 더 자세한 소개는 3장과 4장을 참조하기 바란다.

2.4.1 단일 대체방법

대체방법을 이용하여 자료의 결측값을 채우고 나면 사용자들은 완전한 자료를 가졌다고 생각할 수 있으므로 매우 매력적인 방법으로 고려된다. 하지만 적절하지 못한 통계 모형으로부터 무응답을 대체하게 되면 오히려 완전히 이용가능한 방법보다 더 큰 편향을 발생시킬 수도 있다는 것을 항상 염두에 두어야한다.

대체는 결측값의 예측분포(predictive distribution)의 평균 또는 추출값(draw)을 사용한다. 그러므로 응답자료로부터 무응답의 예측분포를 만드는 방법이 필요하다. 일반적으로 무응답의 예측분포는 명백한 모형 또는 함축적인 모형을 통하여 만든다.

명백한 모형을 근거로 한 대체방법은 무응답의 예측분포를 다변량 정규분포와 같은 통계적 모형을 근거로 구한다. 즉, 이 경우는 모형구축에 사용된 가정이 명백하다. 이런 명백한 모형 방법 중 몇 가지는 다음과 같다.

(a) 평균 대체: 응답자의 평균을 이용하여 무응답값을 대체한다.

(b) 회귀 대체: 회귀식을 응답자의 자료로 추정한 후 무응답값을 적합된 회귀식으로부터 예측하여 대체하는 방법이다.

(c) 확률적 회귀 대체: 위의 회귀대체에서 설명한 회귀예측값에 예측값의 불확실성을 고려하는 잔차를 더하여 무응답을 대체하는 방법이다.

함축적인 모형을 근거로 한 대체방법은 기저의 모형을 내포하는 하나의 알고리듬을 이용하여 결측값을 대체한다. 이 방법에서는 가정이 매우 함축적이며 가정의 검증이 쉽지 않다. 하지만 이런 함축적인 모형 방법에서도 사용된 방법이 합당함을 조심스럽게 평가하여야 한다. 이러한 함축적인 모형 방법 중 몇 가지는 다음과 같다.

(a) 핫덱 대체: 무응답을 현재 진행 중인 연구에서 “비슷한” 성향을 가진 응답자의 자료로 대체하는 방법이다. 핫덱방법은 표본조사에서 흔히 사용된다.
(b) 콜드덱 대체: 핫덱과 비슷하나 대체할 자료를 현재 진행 중인 연구에서 얻는 것이 아니라 외부출처 또는 이전의 비슷한 연구에서 가져오는 방법이다.
(c) 혼합방법: 몇 가지 다른 방법을 혼합하는 방법이다. 예를 들어, 회귀대체를 이용하여 예측값을 얻고 핫덱방법을 이용하여 잔차를 얻어 두 값을 더하는 경우를 생각할 수 있다.

대체방법은 무응답으로 발생하는 편향을 줄이기 위해, 정확도(precision)를 높이기 위해, 또 무응답변수와 관측된 변수와의 관계를 보전하기 위해, 대체 모형에 관측된 변수들을 조건으로 이용하는 것이 바람직하다. 또한 무응답 변수들간의 관계도 보전하기 위하여 대체모형은 결합분포모형을 고려하는 것이 좋다.

다음 절에서는 몇 가지 명백한 모형을 바탕으로 하는 대체방법을 소개한다.

2.4.1.1 비조건부 평균 대체법 (unconditional mean imputation)

y_{ij}를 i번째 개체의 Y_j변수의 값이라고 하자. 가장 간단한 형태의 대체는 결측값 y_{ij}를 변수 Y_j에서 관측된 값만을 이용하여 구한 평균, $\bar{y}_j^{(j)}$로 대체하는 것이다. 이 값으로 대체한 후 관찰된 값과 대체된 값을 이용하여 Y_j의 평균을 구하면 $\bar{y}_j^{(j)}$가 된다. 관찰된 값과 대체된 값을 이용하여 구한 표본분산은 $s_{jj}^{(j)}\left(n^{(j)}-1\right)/(n-1)$이고, 여기서 $s_{jj}^{(j)}$는 Y_j에서 관측된 값만을 이용하여 구한 표본분산 추정값이다. 완전임의 결측의 가정에서 $s_{jj}^{(j)}$는 실제 분산의 일치 추정값이다. 그러므로 위의 대체된 자료로부터 구한 표본분산은 실제분산보다

$(n^{(j)}-1)/(n-1)$만큼 과소추정된다. 또한, 이 방법은 결측이 완전임의가 아닌 경우에는 추정값에 편향이 발생한다. 이 방법은 추천되지 않는 방법이다.

2.4.1.2 조건부 평균 대체법 (conditional mean imputation)

비조건부 평균 대체법을 보완한 방법으로 관측된 자료를 사용하여 조건부 평균을 이용하여 대체하는 방법이 있다.

예제 2.1 보정 클래스를 이용한 평균대체법

무응답자와 응답자를 관찰된 변수들을 바탕으로 J개의 보정클래스(또는 가중 클래스)로 분류한다. 각 클래스 안에서 응답자의 평균을 구하여 무응답을 대체한다. 동일 확률 표본추출을 이용한 표본조사에서 클래스 j안의 응답자를 이용하여 구한 Y변수의 평균을 $\bar{y}_{jR}$ 이라고 하자. 이 경우 대체된 자료를 이용한 평균은 다음과 같다.

$$\frac{1}{n}\sum_{j=1}^{J}\left(\sum_{i=1}^{r_j} y_{ij} + \sum_{i=r_j+1}^{n_j} \bar{y}_{jR}\right) = \frac{1}{n}\sum_{j=1}^{J} n_j \bar{y}_{jR}.$$

이렇게 구한 평균은 각 클래스 안의 응답자의 비율의 역수를 가중치로 이용하여 구한 가중평균 $\bar{y}_{wc}$와 같다 (2.2.1절 참조).

예제 2.2 회귀 대체법 (regression mean imputation)

이 예제에서는 일변량 패턴 무응답 자료를 가정하자. 즉, $Y_1, Y_2, ..., Y_{K-1}$의 변수에는 무응답이 없고 오직 Y_K 변수에서 처음 r개체는 관측이 되었고 다음 $(n-r)$개체에서는 무응답이 발생하였다. 회귀 대체방법은 먼저 r개의 완전히 관측된 자료를 이용하여 Y_K를 종속변수로 $Y_1, Y_2, ..., Y_{K-1}$를 독립변수로 하여 회귀분석을 하여 다음과 같은 회귀식을 적합한다.

$$\hat{y}_{iK} = \hat{\beta}_0 + \hat{\beta}_1 y_{i1} + \hat{\beta}_2 y_{i2} + ... + \widehat{\beta_{K-1}} y_{iK-1}$$

이 적합된 회귀식에 $(n-r)$개의 무응답 개체의 $y_1, y_2, ..., y_{K-1}$을 이용하여 y_K의 예측값을 구하고 그 값으로 무응답값을 대체한다. 만일 모두 관찰된 변수가 범주형인 경우는 예측값이 그 변수의 클래스 안의 응답자들의 평균이므로 이 방법은 위의 예제 2.1과 같아진다. 이 회귀대체법은 회귀식에 연속형 변수, 범주형 변수, 상호작용, 그리고 이차항 또는 스플라인과 같은 비선형 회귀식을 고려함으로써 예측력을 향상시킬 수 있다.

예제 2.3 확률적 회귀 대체법 (stochastic regression imputation)

위의 예제 2.2에서 본 대체법은 추정된 회귀식을 실제 회귀식으로 간주하고 대체값을 생성하였다. 이런 경우 회귀식의 추정에서 발생되는 불확실성, 즉 표본 변이를 고려하지 않아서 추정값의 분산이 과소 추정되는 경향이 발생한다. 이런 표본변이를 고려해 주는 방법으로는 예측분포로부터 대체값을 추출(draw)하는 방법을 이용할 수 있다. 물론 대체값을 추출하는 경우 표본변이는 고려할 수는 있으나 주

의할 점은 단일대체로 발생하는 결측으로 인한 불확실성은 아직 문제점으로 남아 있다는 것이다.

이렇게 회귀식을 적합하는데서 발생하는 표본변이를 고려하기 위해서는 확률적 회귀 대체법을 사용할 수 있다. 이 방법은 먼저 예제 2.2에서 같이 회귀식을 적합하여 예측값을 구한 후 잔차의 분포, 즉 평균이 0이고 잔차분산이 $\hat{\sigma}_{\epsilon}^{2}$인 정규분포에서 하나의 값을 추출하여 위에서 구한 예측값에 더하여 대체값을 생성하는 방법이다.

예제 2.4 기업활동실태조사에서의 무응답 대체

기업활동실태조사는 기업 활동의 다각화, 국제화, 계열화 등 기업의 다양한 경제활동을 포괄적으로 조사함으로써 기업의 경영전략이나 산업구조 변화를 파악하고 기업에 관한 각종 경제정책의 기초자료를 제공하기 위하여 통계청에서 실시되고 있다. 매년 실시되는 이 조사는 각 연도별로 전국의 회사법인 중 종사자 50인 이상이며 자본금 3억원 이상인 기업을 대상으로 실시하는 기업체 조사이다. 조사항목은 기업체명, 소재지, 자본금, 기업 내 조직 및 종사자수 관련, 자산·부채 및 자본 관련, 사업내용 관련, 관계회사(자회사, 관련회사, 모회사) 관련, 기업 간 거래 및 해외거래 관련, 기술소유 및 사용 관련, 기업의 경영방향 관련 항목 등 7개 분야 111개 항목을 포함한다. 2007년 자료에서는 11,650개 해당 기업 중 13%의 기업이 응답을 거부하였고 이 기업들은 행정구역, 산업분류 등 계획변수(design variables)들에 대한 정보만 존재한다. 본 예제에서는 2007년 기업활동실태조사에서 정보를 제공한 87% 기업의 자료에 대하여 일부항목에 임의로 무응답을 설정하고 무응답 대체를 실시한 결과를 보여준다. 본 예제의 목적은 기업활동실태조사

에서 발생하는 무응답에 대한 가장 적절한 대체 방법을 제시하는 것이 아니라 다변량 정규분포를 가정한 대체 방법의 예제를 보여주기 위한 것임을 명시한다.

본 예제에서는 응답을 제공한 10,229개 기업에 대한 매출액(C24)에서 무응답이 30% 발생하였다고 가정하였다. 특히 매출액은 자산총계가 높을수록 무응답이 많이 발생한다고 가정하여 무응답 자료를 생성하였다. 즉, 무응답의 절반인 15%가 자산총계가 상위 20% 이내인 기업들 중에서 임의로 발생한다고 가정하였다. 또한, 무응답의 1/3인 10%가 자산총계가 차상위인 20% - 40%에서 임의로 발생하고 나머지 5%의 무응답은 자산총계가 상위 40% - 60% 사이의 그룹에서 임의로 발생한다고 가정하였다. 즉, 무응답의 발생은 자산총계에 의존하여 발생하므로 대체 모형에 자산총계를 포함시키면 무응답자료 메커니즘은 임의결측이 된다. <표 2.1>은 무응답이 발생한 매출액 변수에 대하여 평균대체, 회귀대체, 확률적 회귀대체를 실시한 후 평균 및 표준오차를 무응답이 발생하기 이전의 완전한 자료의 평균 및 표준오차, 그리고 무응답을 제외한 채 분석을 시행하는 완전히 응답된 개체를 이용한 분석 결과에 비교한다.

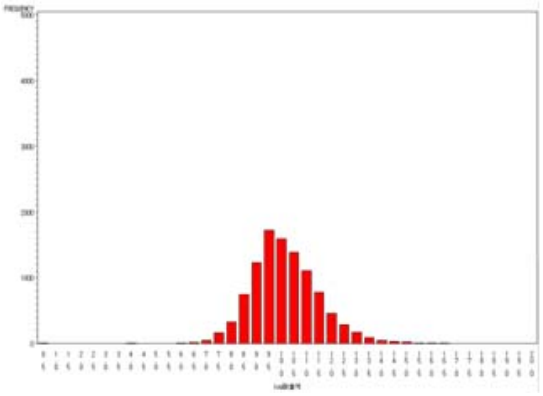
<표 2.1> 평균대체, 회귀대체, 확률적 회귀대체를 통한 평균의 추정, 완전한 자료의 평균, 그리고 이용가능한 자료 분석방법을 시행한 경우 평균의 추정값을 비교 (괄호안은 표준편차)

분석방법	매출액(log_C24)	
	평균	표준오차
완전한 자료	10.16	0.0139
이용가능한 자료 분석	9.71	0.0140
단순 평균대체	9.71	0.0098
조건부 평균대체	10.16	0.0127
회귀대체	10.14	0.0127
확률적 회귀대체	10.14	0.0129

<그림 2.1> 평균대체, 회귀대체, 확률적 회귀대체를 통하여 대체된 자료, 무응답이 발생하기 전 완전한 자료, 그리고 이용가능한 자료의 히스토그램

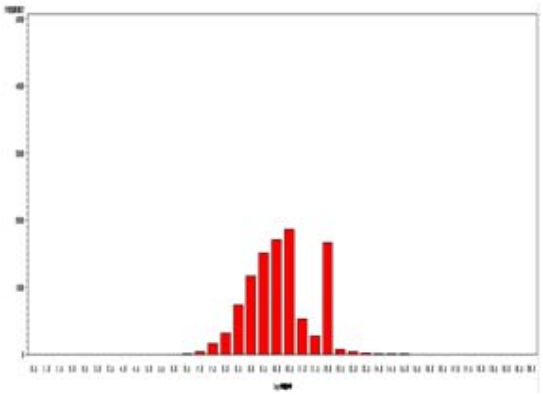
(a) 완전한 자료

(b) 이용가능한 자료

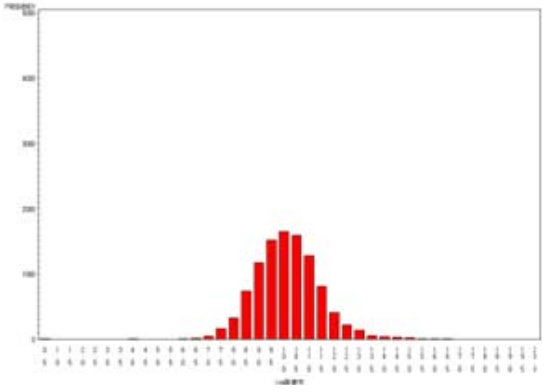


(c) 비조건부 평균대체

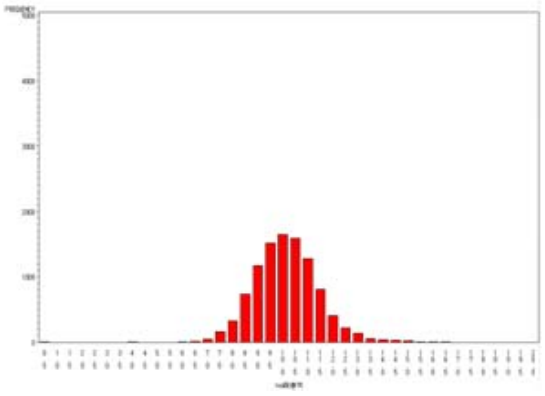
(d) 조건부 평균대체



(e) 회귀대체



(f) 확률적 회귀대체



<표 2.1>에서 나타난 바와 같이 이용가능한 자료분석이나 비조건부 평균대체는 평균의 추정값이 완전한 자료보다 작게 나타나지만 조건부 평균대체, 회귀대체, 그리고 확률적 회귀대체에서는 완전한 자료의 평균과 비슷하게 추정되고 있다. <그림 2.1>을 보면 비조건부 또는 조건부 평균대체에서는 분포의 형태가 왜곡되는 것을 알 수 있다. 이는 평균대체가 모든 무응답값을 평균으로 대체하므로 대체된 자료에서 평균에 해당하는 값의 비율이 비정상적으로 많아져 발생하게 된다. <표 2.1>의 비조건부 평균대체의 평균값의 표준오차는 0.0098로 완전한 자료의 표준오차보다 작은데 이유는 대체된 값이 모두 평균과 동일하여 표준오차 계산시 잔차의 제곱합 계산에는 기여하지 않는데 반하여 관측값의 숫자만 늘어나기 때문이다. 즉, 평균대체 시 표준오차의 추정에 편향이 나타남을 볼 수 있다.

2.4.2 대체로 인한 불확실성을 고려하는 분석방법

2.4.1에서 고려한 분석방법들은 무응답이 있는 경우의 모수의 점추정값에 초점을 맞추었다. 대체를 한 번만 하고 대체된 값을 마치 관측된 값처럼 여겨 분석을 하게 되면 올바른 대체모형과 결측 메커니즘의 가정 하에서는 점추정값에 편향이 발생하지 않는다. 하지만 편향을 발생하지 않았을 지라도 추정값의 분산을 추정하는 데 있어서는 대체로 인한 불확실성을 고려하지 않아서 실제 분산보다 더 작게 된다. 즉, 추정값의 분산이 과소 추정된다. 그러므로 좀 더 나은 추정을 위해서는 대체로 인한 불확실성을 고려한 분석법이 필요하다.

이런 대체로 인한 불확실성을 고려하는 방법으로는 1) 무응답의 불확실성을 고려하는 명백한 분산 공식을 유도하는 방법, 2) 단순 대체방법을 수정하여 단순 대체된 자료에서도 유효한 분산을 구할 수 있도록 하는 방법, 3) 붓스트랩이나 잭나이

프 같은 재표본 방법을 이용하여 불확실성을 추정하고 분산추정에 더해주는 방법, 4) 한 번만 대체 자료를 얻는 것이 아니라 여러 개의 대체 자료를 구하는 다중대체방법이 있다. 1)번 방법은 정확한 분산의 공식을 수리적으로 구하기 어려운 경우가 많고 2)번 방법은 대체방법의 수정으로 인한 추정치 자체의 질이 낮아지는 문제가 있을 수 있다. 그러므로 현실적으로 많이 사용되는 방법은 3)과 4)이고 이 절에서도 이 두 가지 방법을 간략히 소개한다.

2.4.2.1 붓스트랩 방법

예제 2.5 결측이 없는 자료에서의 단순 붓스트랩 방법

독립 표본 $S=\{x_i; i=1,..,n\}$로부터 구한 추정치 $\hat{\theta}$이 모수 θ의 일치추정값이라고 하자. 이제 원 표본 S로부터 복원추출(sampling with replacement)된 n개의 표본을 $S^{(b)}$라고 하고 이 $S^{(b)}$로부터 $\hat{\theta}$을 구할 때 사용한 방법 그대로 이용하여 구한 추정치를 $\hat{\theta}^{(b)}$라고 하자. 여기서 $b=1,...,B$이고 복원추출된 횟수를 나타낸다. 이제 총 B개의 붓스트랩 표본으로부터 구한 추정치를 $(\hat{\theta}^{(1)},...,\hat{\theta}^{(B)})$라고 하자. 이 때, θ의 붓스트랩 추정치는 다음과 같은 $(\hat{\theta}^{(1)},...,\hat{\theta}^{(B)})$의 평균이다.

$$\hat{\theta}_{boot}=\frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{(b)}.$$

$\hat{\theta}$ 또는 $\hat{\theta}_{boot}$의 분산은 다음과 같이 추정된다.

$$\widehat{VAR}_{boot}(\hat{\theta})=\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}^{(b)}-\hat{\theta}_{boot}\right)^2.$$

몇 가지 조건을 만족하는 경우, $\hat{\theta}_{boot}$은 원 추정치 $\hat{\theta}$보다 편향이 작고, n과 B가 무한대로 가는 경우 $\widehat{VAR}_{boot}(\hat{\theta})$가 $\hat{\theta}$ 또는 $\hat{\theta}_{boot}$의 분산의 일치 추정량이다. 또한, 붓스트랩 분포가 근사적으로 정규분포인 경우 θ에 관한 $100(1-\alpha)\%$ 붓스트랩 신뢰구간은 다음과 같다.

$$\hat{\theta} \pm z_{1-\alpha/2}\sqrt{\widehat{VAR}_{boot}(\hat{\theta})}\,.$$

여기서 $z_{1-\alpha/2}$는 정규분포의 $100(1-\alpha/2)\%$에 해당되는 분위수이다. 만일 붓스트랩 분포가 정규분포가 아닌 경우는 θ에 관한 $100(1-\alpha)\%$ 붓스트랩 신뢰구간은 다음과 같이 경험적으로(empirically) 구할 수 있다.

$$\left(\hat{\theta}^{(b)}_{low},\, \theta^{(b)}_{upp}\right).$$

여기서 $\hat{\theta}^{(b)}_{low}$는 붓스트랩 분포의 $(\alpha/2)$에 해당되는 경험적 분위수이고 $\hat{\theta}^{(b)}_{upp}$는 $(1-\alpha/2)$에 해당하는 경험적 분위수이다.

예제 2.6 단일 대체된 자료에 적용된 단순한 붓스트랩 방법

위의 예제와 마찬가지로 독립 표본 $S=\{x_i; i=1,..,n\}$을 고려하자. 이때 S안에는 결측자료가 포함되어 있다. 이 경우 붓스트랩 방법은 다음과 적용될 수 있다.

$b=1,...,B$에 대해

(a) 먼저 결측이 있는 원자료 S로부터 붓스트랩 자료 $S^{(b)}$를 만든다.

(b) 이 붓스트랩 자료 $S^{(b)}$에 대체방법을 적용하여 $S^{(b)}$안의 결측값을 모두 대체

한다. 이 대체된 자료를 $S_{imp}^{(b)}$라고 하자.

(c) 이 대체된 붓스트랩 자료 $S_{imp}^{(b)}$로부터 $\hat{\theta}^{(b)}$를 구한다.

(d) 위 예제 2.5 방법을 이용하여 분산과 신뢰구간을 구한다.

이 방법은 대체방법이 B번 적용되어야 하므로 계산에 시간이 많이 걸린다. 또한 이 방법은 표본수가 큰 경우를 가정한다. 만일 표본수가 작은 경우에는 붓스트랩 자료가 모두 결측만 가지고 있을 수도 있게 된다. 이런 경우를 방지하기 위해서는 붓스트랩 표본을 만들 때 결측개체와 관찰개체를 나누어서 붓스트랩한 후 다시 합치는 방법을 고려할 수 있다.

2.4.2.2 잭나이프 방법

잭나이프 방법은 붓스트랩보다 훨씬 오래전부터 표본조사에서 사용되어왔다. 이 방법은 표본으로부터 한 개체씩 제외하고 추정치를 구하는 방법으로 이제 이 방법을 예제를 통하여 보도록 하자.

예제 2.7 결측이 없는 자료에서의 단순 잭나이프 방법

독립 표본 $S=\{x_i; i=1,..,n\}$로부터 구한 추정치 $\hat{\theta}$이 모수 θ의 일치추정값이라고 하자. $S^{(-j)}$를 원표본에서 j번째 개체를 빼고 구한 크기가 $(n-1)$인 표본이라고 하고 이 표본 $S^{(-j)}$로부터 구한 θ의 추정값을 $\hat{\theta}^{(-j)}$라고 하자. 이 경우 다음의 값을 유사값(pseudovalue)이라고 한다.

$$\tilde{\theta}_j = n\hat{\theta} - (n-1)\hat{\theta}^{(-j)}.$$

θ의 잭나이프 추정치는 위의 유사값들의 평균으로 다음과 같이 구한다.

$$\hat{\theta}_{jack} = \frac{1}{n}\sum_{j=1}^{n}\widetilde{\theta}_j = \hat{\theta} + (n-1)(\hat{\theta} - \overline{\theta}).$$

여기서 $\overline{\theta} = \sum_{j=1}^{n}\hat{\theta}^{(-j)}/n$. $\hat{\theta}$ 또는 $\hat{\theta}_{jack}$의 잭나이프 분산 추정치는 다음과 같다.

$$\widehat{VAR}_{jack}(\hat{\theta}) = \frac{1}{n(n-1)}\sum_{j=1}^{n}\left(\tilde{\theta}_j - \hat{\theta}_{jack}\right)^2 = \frac{n-1}{n}\sum_{j=1}^{n}(\hat{\theta}^{(-j)} - \overline{\theta})^2.$$

만일 잭나이프 분포가 근사적으로 정규분포인 경우, $100(1-\alpha)\%$ 신뢰구간은 다음과 같다.

$$\hat{\theta} \pm z_{1-\alpha/2}\sqrt{\widehat{VAR}_{jack}(\hat{\theta})}\,.$$

예제 2.8 단일 대체된 자료에 적용된 단순한 잭나이프 방법

독립 표본 $S = \{x_i; i = 1, .., n\}$을 고려하자. 이때 S안에는 결측자료가 포함되어 있다. 이 경우 잭나이프 방법은 다음과 적용될 수 있다.

$j = 1, ..., n$에 대해

(a) S에서 j번째 개체를 제외한 표본 $S^{(-j)}$를 만든다.

(b) 대체 방법을 적용하여 잭나이프 표본 $S^{(-j)}$의 결측값을 대체하여 대체된 자료 $S_{imp}^{(-j)}$를 만든다.

(c) 대체된 자료 $S_{imp}^{(-j)}$를 이용하여 $\hat{\theta}^{(-j)}$를 구한다.

2.4.2.3 다중 대체법

다중대체 방법은 원표본의 결측값을 한 번 이상 대체하여 여러 개 ($D \geq 2$)의 대체된 표본을 구하는 방법이다. 다중대체 방법은 Rubin (1978)에 의해 처음 소개되었고 이후 Rubin (1987a)의 책에서 자세하게 다루어졌다. 현재는 단일 대체방법에서 발생하는 대체에 의한 불확실성을 고려하는 방법으로 여러 분야에서 사용되고 있다. 다중 대체법은 D개의 대체된 표본을 만들어야 하므로 항상 같은 값으로 결측자료를 대체할 수 없다. 그러므로 각 대체표본은 결측자료의 예측분포 또는 사후분포에서 추출된 값으로 결측값을 대체하는 방법이 자연스럽다. 이런 이유로 다중대체방법은 베이지안 방법을 이용하는 것이 적절하다고 알려져 있다. 이렇게 같은 예측분포로부터 대체값을 구하여 D개의 대체 표본을 구하게 되면 이 D개의 대체 표본으로부터 원하는 분석을 각 각 수행하여 모수 θ의 점추정치와 표준오차의 추정치를 D개 구한 후 이 들을 Rubin이 제시한 결합공식을 이용하여 결합(pooling)한 후 하나의 결과를 제시하게 된다 (4.3.2절 참조).

이 다중대체법은 여러 번의 대체표본으로 대체-내 분산 (within-imputation variance)과 대체-간 분산 (between-imputation variance)을 구하여 추정치의 총 분산을 추정하는 방법이다. 이 때 대체로 발생하는 불확실성은 대체-간 분산 부분에서 고려함으로써 과소 추정된 분산 추정치가 원 분산에 가까워지도록 하는 것이 다중대체법의 주안점이다.

2.5 우도함수(likelihood function)를 근거로 한 무응답 자료 분석법

결측 자료의 분석에서 많은 경우 구체적인 모형을 가정하고 그 모형하의 우도함

수를 근거로 모수에 관한 추론을 하게 된다. 이 장에서는 먼저 결측이 없는 자료에서의 우도함수에 의한 추론방법에 관하여 리뷰를 하고 이 후 MAR 결측을 가정하는 경우의 방법인 분해우도방법(factored likelihood method)에 관하여 소개한다.

2.5.1 무응답이 없는 경우의 최대우도 추정방법 리뷰

Y를 데이터라고 하자. 이 때 Y는 스칼라, 벡터, 또는 행렬이 될 수 있다. 또 데이터는 확률밀도(또는 질량)함수 $f(Y|\theta)$를 가지는 모형에서 얻어졌다고 가정하자. 여기서 θ는 모수 스칼라 또는 벡터이고 모수공간 Ω_θ에 속한다. 예를 들면 평균의 모수공간은 실수 공간이고 분산의 모수공간은 양의 실수 공간이다.

정의 2.5.1: *데이터 Y가 주어졌을 때, 우도함수 (likelihood function) $L(\theta|Y)$는 $f(Y|\theta)$에 비례하는 $\theta \in \Omega_\theta$의 모든 함수이다. 만일 $\theta \notin \Omega_\theta$이면 $L(\theta|Y)=0$이다.*

우도함수는 Y가 주어진 경우 모수 θ의 함수이고 확률함수는 θ가 주어진 경우 Y의 함수이다. 많은 경우 우도함수를 직접 이용하는 것보다 우도함수에 로그를 취하여 이용하는 것이 훨씬 이용하기 쉽다. 우도함수에 자연 로그를 취한 경우를 로그우도라고 하고 $l(\theta|Y)$로 나타낸다.

예제 2.9 일변량 정규 표본

n개의 독립적인 표본 $Y=(y_1,\cdots,y_t)^T$가 평균 μ와 분산 σ^2을 갖는 정규분포로부

터 동일하게(identically) 추출된 경우 그 결합확률은 다음과 같다.

$$f(Y|\mu,\sigma^2)=\frac{1}{(\sqrt{2\pi}\,\sigma)^n}exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{\sigma^2}\right).$$

Y가 주어진 경우, 로그 우도 함수는 $l(\mu,\sigma^2|\,Y)=\ln[f(Y|\mu,\sigma^2)]$로 표현된다. 또는 상수를 무시하면 로그 우도 함수는 다음과 같다.

$$l(\mu,\sigma^2|\,Y)\propto-\frac{n}{2}\ln\sigma^2-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{\sigma^2}$$

예제 2.10 다변량 정규 표본

$Y=(y_{ij}), i=1,...,n, j=1,...,K$는 평균벡터 $\mu=(\mu_1,\cdots,\mu_K)^T$와 공분산 행렬 $\Sigma=\{\sigma_{jk}, j=1,...,K; k=1,...,K\}$를 갖는 다변량 정규분포로부터 독립적이고 동일하게 얻어진 n개의 표본을 나타내는 표본행렬이라고 하자. y_{ij}는 표본의 i번째 개체의 j번째 변수의 값을 나타낸다. Y의 확률함수는 다음과 같다.

$$f(Y|\mu,\Sigma)=\frac{1}{\sqrt{(2\pi)^{nK}}}|\Sigma|^{-n/2}\exp\left(-\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^T\Sigma^{-1}(y_i-\mu)\right)$$

여기서 $|\Sigma|$는 Σ의 행렬식(determinant)을 나타내고 y_i는 i번째 개체의 행벡터를 나타낸다. 로그우도함수는 다음과 같다.

$$l(\mu, \Sigma) = -\frac{n}{2}\ln|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)^T\Sigma^{-1}(y_i - \mu).$$

우도 함수를 최대화시키는 방법은 모수 θ에 관한 추론의 기본적인 수단이다. 주어진 데이터 Y에 대해 두 개의 가능한 θ값 (θ'과 θ'')을 고려한다고 생각해보자. 또한 $L(\theta' | Y) = 2L(\theta'' | Y)$라고 하자. 그러면 $\theta = \theta''$일 때보다 $\theta = \theta'$일 때 관찰된 Y값이 두 배 더 일어날 수 있을 수 있다고 할 수 있다. 좀 더 일반적으로 어떤 특정한 θ값 ($\hat{\theta}$이라고 하자)이 모든 가능한 다른 θ값에 대하여 $L(\hat{\theta} | Y) \geq L(\theta | Y)$라고 하자. 그러면 관찰된 값 Y는 $\theta = \hat{\theta}$일 때 다른 θ에 비해 일어날 가능성이 적어도 같다고 할 수 있다. 이런 논리를 바탕으로 우도함수를 최대화 시키는 θ값을 추정하게 되고 그 때 추정된 값을 최대우도 추정량(maximum likelihood estimator: MLE)이라고 한다.

정의 2.5.2: *모수 θ의 MLE는 우도함수 $L(\theta | Y)$ 또는 로그우도함수 $l(\theta | Y)$를 최대화 시키는 θ의 값이다.*

MLE는 하나보다 많을 수도 있으나 통계학에서 사용되는 많은 모형에서 MLE는 유일하다. (예: 지수족 exponential family) 만일 우도함수가 미분가능하고 위로 유계한(bounded) 경우, MLE는 θ에 관하여 우도 또는 로그우도 함수를 미분하여 그 결과를 0으로 놓고 θ에 관하여 풀어서 구할 수 있다. 그 결과식은 다음과 같다.

$$D_l(\theta) = \frac{\partial l(\theta | Y)}{\partial \theta} = 0.$$

여기서 로그우도 함수를 미분한 식인 $D_l(\theta)$를 점수함수(score function)라고 하고 이 점수함수를 0으로 놓은 위의 식을 우도식(likelihood equation)이라고 한다.

예제 2.11 일변량 정규 표본

n개의 독립 정규표본의 로그우도함수는 다음과 같다.

$$\begin{aligned} l(\mu, \sigma^2 | Y) &= -\frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2 \\ &= -\frac{n}{2}\ln\sigma^2 - \frac{n(\bar{y}-\mu^2)}{2\sigma^2} - \frac{(n-1)s^2}{2\sigma^2} \end{aligned}$$

여기서 $s^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2/(n-1)$로 표본분산이다.

먼저 위의 로그우도함수를 μ에 관하여 미분하고 그 결과를 $\mu = \hat{\mu}$에 $\sigma^2 = \hat{\sigma}^2$에서 0으로 놓으면 $(\bar{y}-\hat{\mu})^2/\hat{\sigma}^2 = 0$ 이고 이 식을 $\hat{\mu}$에 관하여 풀면 μ의 MLE는 $\hat{\mu} = \bar{y}$ 이다. 다음으로 로그우도함수를 σ^2에 관하여 미분하고 그 결과를 $\mu = \hat{\mu}$에 $\sigma^2 = \hat{\sigma}^2$에서 0으로 놓으면

$$-\frac{n}{2\hat{\sigma}^2} + \frac{n(\bar{y}-\hat{\mu})^2}{2\hat{\sigma}^4} + \frac{(n-1)s^2}{2\sigma^4} = 0$$

이고 이 식을 $\hat{\sigma}^2$에 관하여 풀게 되면 $\hat{\mu} = \bar{y}$이므로 σ^2의 MLE는 $\hat{\sigma}^2 = (n-1)s^2/n$ 이다.

예제 2.12 다변량 정규 표본

$Y = (y_{ij}), i = 1, \cdots, n, j = 1, \cdots, K$는 평균 $\mu = (\mu_1, \cdots, \mu_K)^T$와 공분산 행렬

$$\Sigma = \begin{bmatrix} \sigma_{11}\ \sigma_{12} \cdots\ \sigma_{1K} \\ \sigma_{12}\ \sigma_{22} \\ \vdots \qquad \ddots \quad \vdots \\ \sigma_{1K} \qquad \cdots\ \sigma_{KK} \end{bmatrix}$$

를 가지는 다변량 정규분포에서 추출된 n개의 독립표본 행렬이라고 하자. 즉, y_{ij}는 i번째 표본의 j번째 변수의 값이다. 이 때 Y의 확률함수는

$$f(Y|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^{nK}}} \frac{1}{\sqrt{|\Sigma|^n}} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} (y_i - \mu)^T \Sigma^{-1} (y_i - \mu) \right)$$

이다. 여기서 $|\Sigma|$는 Σ의 행렬식(determinant)이고 y_i는 표본 i의 행벡터 값을 나타낸다. $\theta = (\mu, \Sigma)$의 로그우도함수는

$$l(\mu, \Sigma | Y) = -(n/2) \ln|\Sigma| - \frac{1}{2} \sum_{i=1}^{n} (y_i - \mu)^T \Sigma^{-1} (y_i - \mu).$$

μ와 Σ에 관하여 위의 로그우도함수를 최대화 하면 MLE는

$$\hat{\mu} = \overline{y}, \quad \hat{\Sigma} = S$$

이다. 여기서 $\overline{y} = (\overline{y_1}, \cdots, \overline{y_K})$는 K 변수의 표본 평균들의 행벡터이고 $S = \left(s_{jk} = n^{-1} \sum_{i=1}^{n} (y_{ij} - \overline{y_j})(y_{ik} - \overline{y_k}) \right)$는 $(K \times K)$ 제곱합(sum of squares)과 교차곱(cross-product)을 n으로 나눈 값의 행렬이다.

MLE의 속성 중에 불변의 속성 (invariant property)이 있다. 이 속성은 모수 θ의

어떤 함수 $g(\theta)$가 있을 때, $\hat{\theta}$이 θ의 MLE라고 하면 $g(\hat{\theta})$도 $g(\theta)$의 MLE가 된다.

예제 2.13 이변량 정규분포에서 유도된 조건부 분포

평균이 $\mu = (\mu_1, \mu_2)^T$이고 공분산 행렬이

$$\Sigma = \begin{bmatrix} \sigma_{11} \sigma_{12} \\ \sigma_{12} \sigma_{22} \end{bmatrix}$$

를 갖는 이변량 정규분포로부터 n개의 표본 $y_i = (y_{i1}, y_{i2}), i = 1, \cdots, n$을 추출하였다. 이 경우 MLE는 위의 예제에서 본 바와 같이

$$\widehat{\mu_j} = \overline{y_j}, \quad j = 1, 2$$
$$\widehat{\sigma_{jk}} = s_{jk}/n, \quad j, k = 1, 2$$

이변량 정규분포의 특성에 의하여, y_{i1}이 주어진 경우 y_{i2}의 조건부 분포는 평균 $\mu_2 + \beta_{21.1}(y_{i1} - \mu_1)$과 분산 $\sigma_{22.1}$을 가진 정규분포이다. 여기서

$$\beta_{21.1} = \sigma_{12}/\sigma_{11} \text{ 이고 } \sigma_{22.1} = \sigma_{22} - \sigma_{12}^2/\sigma_{11}$$

이다. 이 때 $\beta_{21.1}$과 $\sigma_{22.1}$은 각 각 Y_1을 독립변수로 Y_2를 종속변수로 한 선형회귀식의 기울기와 오차의 분산이다. MLE의 불변의 속성을 이용하여 이 회귀모수들의 MLE를 구하면

$$\widehat{\beta_{21.1}} = \widehat{\sigma_{12}}/\widehat{\sigma_{11}} = s_{12}/s_{11},$$ (기울기의 최소제곱 추정값)

이고

$$\widehat{\sigma_{22.1}} = \widehat{\sigma_{22}} - \widehat{\sigma_{12}^2}/\widehat{\sigma_{11}} = SSE/n$$

이다. 여기서 $SSE = \sum_{i=1}^{n} \left[y_{i2} - \overline{y_2} - \widehat{\beta_{21.1}} \left(y_{i1} - \overline{y_1} \right) \right]^2$는 회귀분석에서 오차제곱합(sum of squares of errors)에 해당한다.

다음은 MLE의 대표본(large sample) 속성이다. 위와 같이 $\hat{\theta}$이 θ의 MLE라고 하면 몇 가지 전형적인 제약 하에서 $(\hat{\theta} - \theta)$의 분포는 표본이 큰 경우 근사적으로 평균은 0이고 공분산은 C인 정규분포로 근사한다. 즉,

$$(\hat{\theta} - \theta) \dot{\sim} N(0, C).$$

여기서 C는 $\hat{\theta}$의 공분산 행렬로 $C = I^{-1}(\theta | Y)$이고 $I(\theta | Y) = -\frac{\partial^2 l(\theta | Y)}{\partial \theta \, \partial \theta}$이다. $I(\theta | Y)$는 정보행렬(information matrix)이라고 한다. 이제 $g(\theta)$를 θ의 단조 미분 가능한 함수라고 하고 $(\hat{\theta} - \theta)$의 대표본 공분산 행렬을 C 라고 하면 $g(\hat{\theta}) - g(\theta)$의 분포는 평균이 0이고 공분산 행렬은 $D_g(\hat{\theta}) \, C D_g(\hat{\theta})^T$와 형태를 갖는 정규분포로 근사한다.

즉,

$$g(\hat{\theta}) - g(\theta) \dot{\sim} N\left[0, D_g(\hat{\theta}) \, C D_g(\hat{\theta})^T \right].$$

여기서 $D_g(\hat{\theta}) = \partial g(\theta)/\partial\theta$는 θ에 관한 g함수의 부분 미분이다.

2.5.2 무응답이 있는 경우 우도에 근거한 추론 방법

무응답이 있으나 없으나 최대우도 방법의 원리는 같다. 단지 무응답이 있는 경우는 무응답을 가진 자료를 근거로 모수에 관한 우도함수를 유도한 후 우도 식을 풀어서 MLE를 구한다. 이제 그 과정을 보도록 하자.

이전과 마찬가지로 Y를 무응답이 없는 경우의 자료 행렬이라고 하자. 이제 이 Y는 응답된 자료 Y_{obs}와 무응답된 자료 Y_{mis}로 구성된다. 즉, $Y = (Y_{obs}, Y_{mis})$이다. $f(Y|\theta) \equiv f(Y_{obs}, Y_{mis}|\theta)$는 Y_{obs}와 Y_{mis}의 결합분포의 확률함수이다. 이 경우 Y_{obs}의 주변확률함수는 무응답 자료인 Y_{mis}에 관하여 다음과 같이 적분하면 구할 수 있다.

$$f(Y_{obs}|\theta) = \int f(Y_{obs}, Y_{mis}|\theta)\, dY_{mis}.$$

무응답 메커니즘을 무시하면서 (ignoring missing mechanism) 응답된 자료 Y_{obs}에 근거한 모수 θ의 우도를 $f(Y_{obs}|\theta)$에 비례하는 θ에 관한 함수라고 정의하자. 즉, $L_{ign}(\theta|Y_{obs}) \propto f(Y_{obs}|\theta)$로 정의된다. 만일 무응답에 관한 메커니즘을 무시할 수 있으면 위의 우도 $L_{ign}(\theta|Y_{obs})$를 이용하여 θ에 관한 추론을 할 수 있다.

무응답에 관한 추론을 하는 경우에는 모형에 응답 지시 변수의 분포를 함께 고려하여야 한다. 전과 마찬가지로 응답 지시변수는 다음과 같이 정의 된다.

$$R_{ij} = \begin{cases} 1, & y_{ij}\text{가 응답인 경우} \\ 0, & y_{ij}\text{가 무응답인 경우} \end{cases}$$

이 경우 R과 Y의 결합분포는 다음과 같이 표현할 수 있다.

$$f(Y, R|\theta, \psi) = f(Y|\theta) f(R|Y, \psi), \quad (\theta, \psi) \in \Omega_{\theta, \psi}$$

즉, Y의 주변분포의 모수는 θ이고 Y가 주어진 경우 R의 조건분포의 모수는 ψ이고 $\Omega_{\theta, \psi}$는 (θ, ψ)의 결합모수공간이다. 우리의 관심은 ψ에 관한 추론이 아니고 θ에 관한 추론이지만 특별한 경우가 아니면 둘을 떼어놓고 θ에 관한 추론을 할 수 없다. 어떤 의미에서 ψ는 장애모수(nuisance parameter)이다.

실제 응답된 자료는 변수 (Y_{obs}, R)의 값들로 구성된다. 응답된 자료의 확률함수는 $Y = (Y_{obs}, Y_{mis})$와 R의 결합 확률함수를 Y_{mis}에 관하여 적분함으로써 구할 수 있다. 즉,

$$f(Y_{obs}, R|\theta, \psi) = \int f(Y_{obs}, Y_{mis}|\theta) f(R|Y_{obs}, Y_{mis}, \psi) dY_{mis}$$

θ와 ψ의 완전 우도는 $f(Y_{obs}, R|\theta, \psi)$ 에 비례하는 θ와 ψ의 함수이다:

$$L_{full}(\theta, \psi|Y_{obs}, R) \propto f(Y_{obs}, R|\theta, \psi).$$

이제 무응답 자료에서 우도를 이용하여 모수에 관한 추론을 할 때 언제 완전우도 $L_{full}(\theta, \psi|Y_{obs}, R)$를 이용하고 언제 메커니즘을 무시한 우도 $L_{ign}(\theta|Y_{obs})$를 이용

하는 지가 문제이다. 메커니즘을 무시한 우도 $L_{ign}(\theta|\, Y_{obs})$는 R의 분포와 관련이 없으므로 훨씬 단순하고 쉽다. 만일 무응답이 MAR 메커니즘이면, 즉

$$f(R|\, Y_{obs},\, Y_{mis}, \psi) = f(R|\, Y_{obs}, \psi)$$

이런 경우

$$f(Y_{obs}, R|\theta, \psi) = f(R|\, Y_{obs}, \psi) \times \int f(Y_{obs},\, Y_{mis}|\theta)\, dY_{mis}$$
$$= f(R|\, Y_{obs}, \psi) \times f(Y_{obs}|\theta)$$

만일 무응답 메커니즘이 MAR이고 모수 θ와 ψ가 서로 별개(distinct)인 경우, 즉 θ와 ψ의 결합모수공간 $\Omega_{\theta,\psi}$이 θ의 모수공간 Ω_θ과 ψ의 모수공간 Ω_ψ의 곱$(\Omega_{\theta,\psi} = \Omega_\theta \times \Omega_\psi)$인 경우, $L_{full}(\theta, \psi|\, Y_{obs}, R)$을 이용한 우도 바탕의 추론은 $L_{ign}(\theta|\, Y_{obs})$를 이용한 추론과 같다. 지금까지의 설명은 다음과 같이 정의된다.

정의 2.5.3: *우도를 이용한 모수의 추론에서 만일 아래의 두 조건을 만족하면 무응답 메커니즘은 무시할 만 하다라고 한다.*

(가) MAR: 무응답 메커니즘은 임의결측이다.

(나) 별개성(distinctness): 모수 θ와 ψ가 서로 별개인 경우, 즉 θ와 ψ의 결합모수공간 $\Omega_{\theta,\psi}$이 θ의 모수공간 Ω_θ과 ψ의 모수공간 Ω_ψ의 곱$(\Omega_{\theta,\psi} = \Omega_\theta \times \Omega_\psi)$인 경우

별개성이 보장이 되지 않는 경우 $L_{ign}(\theta|\, Y_{obs})$를 이용하여 추론을 하더라도 그에

따른 결과는 비록 효율성은 떨어질지라도 편향 측면에서는 타당하다. 이런 이유로 메커니즘을 무시하기 위해서는 위의 두 조건에서 별개성보다는 MAR이 더 중요한 가정이다.

예제 2.14 무응답이 있는 지수 표본

무응답이 있는 일변량 지수 표본을 고려하자. 이 때, $Y_{obs}=(y_1,\cdots,y_r)^T$는 응답된 자료이고 $Y_{mis}=(y_{r+1},\cdots,y_n)^T$는 무응답 자료이다. 무응답이 없는 경우의 지수 확률함수는

$$f(Y|\theta)=\frac{1}{\theta}\exp\left(-\sum_{i=1}^{n}\frac{y_i}{\theta}\right)$$

이다. 무응답 메커니즘을 무시한 우도는 θ가 주어진 경우 Y_{obs}의 확률함수에 비례한다. 즉,

$$f(Y_{obs}|\theta)=\frac{1}{\theta^r}\exp\left(-\sum_{i=1}^{r}\frac{y_i}{\theta}\right).$$

이제 $R=(R_1,\cdots,R_n)^T$라고 하자. 여기서 $R_i=1, i=1,\cdots,r$ 이고 $R_i=0, i=r+1,\cdots,n$이다.

각 개체는 Y와 독립적으로 확률 ψ를 가지고 응답하였다고 하자. 그러면

$$f(R|\,Y,\psi)=\frac{n!}{r!\,(n-r)!}\psi^r\,(1-\psi)^{n-r}$$

이고

$$f(Y_{obs}, R|\theta) = \frac{n!}{r!(n-r)!}\psi^r(1-\psi)^{n-r}\theta^r \exp\left(-\sum_{i=1}^{r}\frac{y_i}{\theta}\right)$$

이다.

무응답이 MAR이므로 만일 ψ와 θ가 별개이면 모수 θ에 관한 우도 추론은 $f(Y_{obs}|\theta)$에 비례하는 메커니즘을 무시한 우도를 이용할 수 있다. 이 경우 MLE는 단순히 Y_{obs}의 평균인 $\sum_{i=1}^{r} y_i/r$이다.

이제 Y가 알려진 절단값 (censoring point) c보다 큰 경우에 무응답이 발생한다고 가정하자. 그러면

$$f(R|Y,\psi) = \prod_{i=1}^{n} f(R_i|y_i,\psi)$$

이다. 여기서

$$f(R_i|y_i,\psi) = \begin{cases} 1, & \text{if } R_i = 0 \text{ and } y_i \geq c, \text{ or } R_i = 1 \text{ and } y_i < c \\ 0, & otherwise. \end{cases}$$

그래서

$$L_{full}(\theta|\, Y_{obs}, R) = f(\, Y_{obs}, R|\theta) = \prod_{i=1}^{r} f(y_i, R_i|\theta) \prod_{i=1+r}^{n} f(R_i|\theta)$$

$$= \prod_{i=1}^{r} f(y_i|\theta)\, P\,(y_i < c|\, y_i) \prod_{i=1+r}^{n} P\,(y_i \geq c|\theta)$$

$$= \theta^{-r} \exp\left(-\sum_{i=1}^{r} \frac{y_i}{\theta}\right) \exp\left(-\frac{(n-r)c}{\theta}\right).$$

왜냐하면 지수분포에서 무응답의 경우는 $\Pr(y_i < c|\, y_i) = 1$이고 응답자의 경우는 $\Pr(y_i \geq c|\theta) = \exp(-c/\theta)$이기 때문이다. 이런 경우 무응답 메커니즘을 무시할 수 없다. 위의 우도를 최대화하는 MLE는 다음과 같다.

$$\hat{\theta} = \frac{\sum_{i=1}^{r} y_i + (n-1)\, c}{r}$$

2.5.3 분해우도방법

분해우도방법을 이용하려면 무응답 메커니즘은 무시할 수 있다고 가정한다. 무응답 메커니즘을 무시한 우도인 $l_{ign}(\theta|\, Y_{obs})$는 명백한 최대값도 없고 그에 따른 정보행렬도 매우 복잡하다. 하지만 어떤 모형과 무응답 패턴에서는 $l_{ign}(\theta|\, Y_{obs})$에 근거한 분석을 하는 경우 흔히 사용되는 완전한 자료 분석 방법을 이용할 수도 있다.

무응답 분석에서 많은 경우 우리의 관심모수인 θ를 θ의 일대일 함수인 $\phi(\,)$를 이용하여 아래와 같이 로그우도를 분해하는 재모수(alternative parameter) $\phi = \phi(\theta)$를 고려할 수 있다.

$$l(\phi|\, Y_{obs}) = l_1(\phi_1|\, Y_{obs}) + l_2(\phi_2|\, Y_{obs}) + \cdots + l_J(\phi_J|\, Y_{obs})$$

여기서 $\phi_1, \phi_2, \cdots, \phi_J$ 는 별개(distinct)의 모수이고 구성요소 $l_j(\phi_j|\, Y_{obs})$는 완전한 자료의 로그우도와 상응한다. 만일 이러한 속성을 가진 분해 로그우도를 찾을 수 있으면 각 각의 $l_j(\phi_j|\, Y_{obs})$를 최대화함으로써 $l(\phi|\, Y_{obs})$ 를 최대화할 수 있다. 만일 $\hat{\phi}$가 위의 분해 우도를 최대화하는 ϕ의 MLE라고 하면 ϕ에 관한 어떤 함수 $\theta(\phi)$의 MLE는 $\theta(\phi)$에 $\hat{\phi}$를 대체하여 구할 수 있다. 즉, $\hat{\theta} = \theta(\hat{\phi})$이다.

분해 우도는 MLE의 공분산 행렬을 구하는 데 사용될 수 있다. 위의 분해우도 식을 $\phi_1, \phi_2, \cdots, \phi_J$에 관하여 두 번 미분하면 다음과 같은 블록 대각 정보행렬을 구할 수 있다.

$$I(\phi|\, Y_{obs}) = \begin{bmatrix} I(\phi_1|\, Y_{obs}) & & 0 \\ & I(\phi_2|\, Y_{obs}) & \\ & & \ddots \\ 0 & & I(\phi_J|\, Y_{obs}) \end{bmatrix}$$

그러므로 $\hat{\phi} - \phi$의 대표본 공분산 행렬은 다음과 같다.

$$C(\hat{\phi} - \phi|\, Y_{obs}) = \begin{bmatrix} I^{-1}(\widehat{\phi_1}|\, Y_{obs}) & & 0 \\ & I^{-1}(\widehat{\phi_2}|\, Y_{obs}) & \\ & & \ddots \\ 0 & & I^{-1}(\widehat{\phi_J}|\, Y_{obs}) \end{bmatrix}$$

위 행렬의 원소들은 완전히 응답한 자료의 분석과 상응하므로 비교적 쉽게 구할 수 있다. MLE의 불변의 속성을 이용하면 $\theta = \theta(\phi)$의 MLE의 근사 공분산 행렬은

$$C(\hat{\theta}-\theta \mid Y_{obs}) = D(\hat{\theta})\, C(\hat{\phi}-\phi \mid Y_{obs}) D^{T}(\hat{\theta})$$

이다. 여기서 $D(\bullet)$은 ϕ에 관하여 $\theta=\theta(\phi)$를 부분 미분한 행렬로

$$D(\theta) = \{d_{jk}(\theta)\}, \quad d_{jk}(\theta) = \frac{\partial \theta_j}{\partial \phi_k},$$

이고 θ는 열벡터로 표현된다.

예제 2.15 이변량 정규분포 자료에서 한 변수에만 무응답이 있는 자료의 ML 분석법

이변량 정규(bivariate normal) 자료에서 r개체에서는 Y_1과 Y_2 둘 다 응답이 있고 $(n-r)$개체에서는 Y_1만 응답이 있다고 하자. 즉, $\{y_i=(y_{i1}, y_{i2}), i=1,\cdots,r\}$이고 $\{y_{i1}, i=r+1, ..., n\}$이다. 이 자료의 로그우도함수는

$$l_{ign}(\mu, \Sigma \mid Y_{obs}) = \ln\left[l_{ign}(\mu, \Sigma \mid Y_{obs})\right] = -\frac{1}{2} r \ln|\Sigma| - \frac{1}{2}\sum_{i=1}^{r}(y_i-\mu)\Sigma^{-1}(y_i-\mu)^T$$
$$-\frac{1}{2}(n-r)\ln\sigma_{11} - \frac{1}{2}\sum_{i=r+1}^{n}\frac{(y_{i1}-\mu_1)^2}{\sigma_{11}}$$

이다. 여기서 $\Sigma = \begin{bmatrix}\sigma_{11}\,\sigma_{12}\\ \sigma_{12}\,\sigma_{22}\end{bmatrix}$이고 σ_{11}은 Y_1의 분산이고 σ_{22}는 Y_2의 분산이고 σ_{12}는 Y_1과 Y_2의 공분산이다. 평균 벡터 $\mu=(\mu_1, \mu_2)^T$와 공분산 Σ의 MLE는 위의 로그 우도함수를 μ와 Σ에 관하여 최대화함으로써 구할 수 있다. 하지만 이 우도

식의 명백한 해를 구하기 어렵다. Anderson (1957)은 y_{i1}과 y_{i2}의 결합분포를 y_{i1}의 주변분포와 y_{i1}이 주어진 경우의 y_{i2}의 조건부 분포로 분해하는 방법을 제시하였다.

$$f(y_{i1}, y_{i2}|\mu, \Sigma) = f(y_{i1}|\mu_1, \sigma_{11}) f(y_{i2}|y_{i1}, \beta_{20.1}, \beta_{21.1}, \sigma_{22.1})$$

여기서 $f(y_{i1}|\mu_1, \sigma_{11})$은 평균 μ_1과 분산 σ_{11}을 가진 정규분포이고 $f(y_{i2}|y_{i1}, \beta_{20.1}, \beta_{21.1}, \sigma_{22.1})$은 평균 $\beta_{20.1} + \beta_{21.1} y_{i1}$과 분산 $\sigma_{22.1}$을 가진 정규분포이다.

변환 모수 $\phi = (\mu_1, \sigma_{11}, \beta_{20.1}, \beta_{21.1}, \sigma_{22.1})^T$ 은 원 모수인 $\theta = (\mu_1, \mu_2, \sigma_{11}, \sigma_{12}, \sigma_{22})^T$의 일대일 함수이다. μ_1과 σ_{11}은 원 모수와 변환 모수 둘 다 포함되어 있고 ϕ의 다른 구성요소는 다음과 같이 θ의 구성요소들의 함수로 표현할 수 있다.

$$\beta_{21.1} = \sigma_{12}/\sigma_{11},$$
$$\beta_{20.1} = \mu_2 - \beta_{21.1}\mu_1,$$
$$\sigma_{22.1} = \sigma_{22} - \sigma_{12}^2/\sigma_{11}$$

비슷하게, μ_1과 σ_{11}을 제외한 θ의 구성요소는 ϕ의 구성요소들의 함수로 표현할 수 있다.

$$\mu_2 = \beta_{20.1} + \beta_{21.1}\mu_1,$$
$$\sigma_{12} = \beta_{21.1}\sigma_{11},$$
$$\sigma_{22} = \sigma_{22.1} + \beta_{21.1}^2\sigma_{11}$$

이제 Y_{obs}의 확률함수는 다음과 같이 분해된다.

$$
\begin{aligned}
f(Y_{obs}|\theta) &= \prod_{i=1}^{r} f(y_{i1}, y_{i2}|\theta) \prod_{i=r+1}^{n} f(y_{i1}|\theta) \\
&= \left[\prod_{i=1}^{r} f(y_{i1}|\theta) f(y_{i2}|y_{i1}, \theta)\right] \left[\prod_{i=r+1}^{n} f(y_{i1}|\theta)\right] \\
&= \left[\prod_{i=1}^{n} f(y_{i1}|\ \mu_1, \sigma_{11})\right] \left[\prod_{i=1}^{r} f(y_{i2}|y_{i1}, \beta_{20.1}, \beta_{21.1}, \sigma_{22.1})\right]
\end{aligned}
$$

위 식에서 $\left[\prod_{i=1}^{n} f(y_{i1}|\mu_1, \sigma_{11})\right]$는 평균 μ_1과 분산 σ_{11}을 가진 정규분포에서 추출된 n개의 독립표본의 확률함수이다. $\left[\prod_{i=1}^{r} f(y_{i2}|y_{i1}, \beta_{20.1}, \beta_{21.1}, \sigma_{22.1})\right]$는 평균 $\beta_{20.1} + \beta_{21.1} y_{i1}$과 분산 $\sigma_{22.1}$을 가진 조건부 정규분포의 r개의 응답표본의 확률함수이다. (μ_1, σ_{11})은 $(\beta_{20.1}, \beta_{21.1}, \sigma_{22.1})$의 어떤 정보도 가지고 있지 않으므로 (μ_1, σ_{11})과 $(\beta_{20.1}, \beta_{21.1}, \sigma_{22.1})$은 서로 개별적이다. 그러므로 ϕ의 MLE는 위의 두 구성 요소들에 대응되는 로그우도를 각 각 최대화함으로써 구할 수 있다.

$\left[\prod_{i=1}^{n} f(y_{i1}|\mu_1, \sigma_{11})\right]$를 최대화하는 MLE는

$$
\widehat{\mu_1} = \frac{\sum_{i=1}^{n} y_{i1}}{n},
$$

$$
\widehat{\sigma_{11}} = \frac{\sum_{i=1}^{n} \left(y_{i1} - \widehat{\mu_1}\right)^2}{n}
$$

이다. 즉, n개의 표본 $\{y_{11}, y_{21}, \cdots, y_{n1}\}$의 표본 평균과 표본 분산이다.

$\left[\prod_{i=1}^{r} f(y_{i2}|y_{i1},\beta_{20.1},\beta_{21.1},\sigma_{22.1})\right]$를 최대화하는 방법은 이전에 본 바와 같이 표준 회귀분석 방법을 사용할 수 있다.

$$\widehat{\beta_{21.1}} = s_{12}/s_{11},$$
$$\widehat{\beta_{20.1}} = \overline{y_2} - \widehat{\beta_{21.1}}\,\overline{y_1},$$
$$\widehat{\sigma_{22.1}} = s_{22.1}$$

여기서 $\overline{y_j} = r^{-1}\sum_{i=1}^{r} y_{ij}$, $s_{jk} = r^{-1}\sum_{i=1}^{r}(y_{ij}-\overline{y_j})(y_{ik}-\overline{y_k})$, $j=1,2,k=1,2,$ $s_{22.1} = s_{22} - s_{12}^2/s_{11}$이다. 이제 다른 모수들의 추정값은 MLE의 불변의 속성을 이용하여 구할 수 있다. 즉,

$$\hat{\mu}_2 = \hat{\beta}_{20.1} + \hat{\beta}_{21.1}\hat{\mu}_1 = \overline{y}_2 + \hat{\beta}_{21.1}(\hat{\mu}_1 - \overline{y}_1),$$
$$\hat{\sigma}_{22} = \hat{\sigma}_{22.1} + \hat{\beta}_{21.1}^2\hat{\sigma}_{11} = s_{22} + \hat{\beta}_{21.1}^2(\hat{\sigma}_{11} - s_{11})$$

이다. $\widehat{\mu_2}$의 식에서 $\overline{y}_2$ 부분은 $(n-r)$개의 무응답 표본은 버리고 r개의 응답한 표본을 이용하여 구한다. 그리고 $\hat{\beta}_{21.1}(\hat{\mu}_1 - \overline{y}_1)$ 부분은 $(n-r)$개의 무응답에서 얻을 수 있는 y_{i1}의 추가적인 정보를 이용하여 보정하는 부분이다. $\widehat{\mu_2}$을 다시 표현하면

$$\widehat{\mu_2} = \frac{1}{n}\left(\sum_{i=1}^{r} y_{i2} + \sum_{i=r+1}^{n} \hat{y}_{i2}\right)$$

이고 여기서 $\hat{y}_{i2} = \overline{y}_2 + \hat{\beta}_{21.1}(y_{i1} - \overline{y}_1)$이다. 이는 응답한 자료를 이용하여 Y_2를

종속변수로 하고 Y_1을 독립변수로 하여 회귀식을 구하고 무응답 표본의 y_{i1}을 회귀식에 대입하여 무응답 값 y_{i2}를 예측하여 이 값으로 무응답값을 대체하는 방법과 궁극적으로는 같다.

2.5.4 무응답 패턴이 일반적인 경우의 최대우도 방법

무응답 패턴이 일반적인 경우는 분해우도를 사용하여 MLE를 구하는 것이 쉽지 않다. 어떤 모형에서는 분해우도가 존재할 수 있으나 분해우도 안의 모수 ϕ_j가 서로 별개(distinct)가 아닌 경우에는 각 분해된 우도를 따로 최대화하는 것이 전체 우도를 최대화 하는 것은 아닐 수 있다. 이런 경우는 MLE를 구하기 위하여 반복법(iteration method)을 사용하여야 한다.

이전과 마찬가지로 Y를 무응답이 없는 경우의 자료 행렬이라고 하자. 이제 이 Y는 응답된 자료 Y_{obs}와 무응답된 자료 Y_{mis}로 구성된다. 즉, $Y=(Y_{obs}, Y_{mis})$이다. $f(Y|\theta) \equiv f(Y_{obs}, Y_{mis}|\theta)$는 Y_{obs}와 Y_{mis}의 결합분포의 확률함수이다. 무응답 메커니즘은 MAR이라고 가정하면 다음의 우도를 최대화하는 모수 θ의 추정값을 구하는 것이 목적이다.

$$L(\theta | Y_{obs}) = \int f(Y_{obs}, Y_{mis}|\theta)\, dY_{mis}$$

우도가 미분가능하고 단봉(unimode)형태의 함수라면 MLE는 다음의 우도함수를 θ에 관하여 풀어서 구할 수 있다.

$$D_l(\theta \mid Y_{obs}) \equiv \frac{\partial l(\theta \mid Y_{obs})}{\partial \theta} = 0$$

위 식의 폐쇄형 해(closed-form solution)를 구할 수 없으면, 뉴튼-랩슨(Newton-Raphson) 알고리듬과 같은 반복법을 사용할 수 있다. 먼저 $\theta^{(0)}$가 θ의 초기 추정값이라고 하자. 예를 들면 이 초기값은 완전히 이용가능한 자료로부터 추정할 수 있다. 이제 $\theta^{(t)}$를 t번째 반복의 추정값이라고 하자. 뉴튼-랩슨(Newton-Raphson) 알고리듬은 다음의 식으로 정의된다.

$$\theta^{(t+1)} = \theta^{(t)} + I^{-1}\left(\theta^{(t)} \mid Y_{obs}\right) D_l\left(\theta^{(t)} \mid Y_{obs}\right)$$

여기서 $I(\theta \mid Y_{obs})$는 아래와 같이 정의된 관측정보행렬(observed information matrix)이다.

$$I(\theta \mid Y_{obs}) = \frac{\partial^2 l(\theta \mid Y_{obs})}{\partial\theta\,\partial\theta}$$

만일 로그우도함수가 오목하고(concave) 단봉형태이면 뉴튼-랩슨 알고리듬은 MLE $\hat{\theta}$로 수렴한다. 뉴튼-랩슨과 비슷한 방법으로 점수법(scoring method)이 있다. 점수법에서는 관측 정보행렬을 기대 정보행렬(expected information matrix)로 대신한다. 즉,

$$\theta^{(t+1)} = \theta^{(t)} + J^{-1}\left(\theta^{(t)} \mid Y_{obs}\right) D_l\left(\theta^{(t)} \mid Y_{obs}\right)$$

여기서 $J(\theta \mid Y_{obs})$는 아래와 같이 정의된다.

$$\mathcal{J}(\theta) = E\left[\mathcal{I}(\theta \mid Y_{obs}) \mid \theta\right] = -\int \frac{\partial^2 l(\theta \mid Y_{obs})}{\partial\theta\,\partial\theta} f(Y_{obs} \mid \theta)\, dY_{obs}$$

뉴튼-랩슨 알고리듬과 점수법, 두 방법 다 로그우도의 이차미분 행렬을 계산하여야 한다. 무응답의 패턴이 복잡할수록 이 행렬 안의 원소는 θ의 매우 복잡한 형태의 함수가 된다. 또한 θ안의 모수의 수가 많을수록 이차미분 행렬의 크기도 커진다. 그러므로 실제에서는 조심스러운 수학적 접근과 효율적인 프로그램을 만드는 것이 매우 중요하다.

위의 두 방법의 대안으로 EM (expectation and maximization: EM) 알고리듬이 있다. EM 알고리듬에서는 로그 우도함수의 이차미분을 필요로 하지 않는다. 이 방법은 $l(\theta \mid Y_{obs})$을 이용한 θ의 추정을 완전한 자료의 로그우도인 $l(\theta \mid Y)$를 이용한 θ의 추정으로 연관시킨다. EM 알고리듬은 개념적으로나 계산적으로 매우 쉬운 경우가 많다. 하지만 EM 알고리듬은 두 가지 주요한 단점도 있다. 자료에서 무응답의 비율이 높은 경우 수렴의 속도가 매우 느리다. 물론 컴퓨팅 환경의 빠른 개선으로 이런 문제는 많이 해소되었다. 또한 어떤 경우에 있어서는 최대화 단계(maximization step)에서 폐쇄형 해를 구하지 못하는 경우가 발생하기도 한다. 이런 어려움을 극복하기 위해 많은 EM의 변종이 개발되었다. 본 교재에서는 표준 EM에 관해서만 언급하도록 하겠다.

2.5.5 EM 알고리듬 소개

EM 알고리듬은 1977년 Dempster, Laird와 Rubin에 의하여 소개되었다. 물론 이전에도 비슷한 방법이 특별한 상황에서 제안되었으나 Dempster, Laird와 Rubin이 그 방법들을 일반화하고 EM이라는 이름을 붙였다. EM 알고리듬은 모수의

MLE를 구하는 반복법으로 각 반복에서 E(expectation)-단계와 M(maximization)-단계로 구성된다. E-단계에서는 응답된 자료와 현재 반복에서의 모수 추정값을 이용하여 무응답 자료를 추정한다. 좀 더 엄밀하게 말하면 응답된 자료와 현재 반복에서의 모수 추정값을 이용하여 완전한 자료의 기대 로그우도를 구한다. M-단계에서는 위에서 구한 완전한 자료의 기대 로그우도를 최대화 하는 모수 추정값을 갱신한다. 이제 E-단계와 M-단계를 모수의 추정값이 수렴할 때까지 반복한다. 이 알고리듬은 만일 $l(\theta|Y_{obs})$가 유계(bounded)하면 로그 우도함수 $l(\theta|Y_{obs})$는 각 반복에서 계속 증가한다. 이러한 속성은 모수의 추정값 $\hat{\theta}$은 항상 수렴함을 보장한다. 예제를 통하여 EM을 좀 더 알아보자.

예제 2.16 일변량 정규분포 자료

$y_i, i=1,\cdots,n$ 는 정규분포 $N(\mu,\sigma^2)$에 추출된 임의표본이라고 하자. 이 때, $y_i, i=1,\cdots,r$ 은 응답된 자료이고 $y_i, i=r+1,\cdots,n$은 무응답 자료이다. 또한 무응답 메커니즘은 MAR이라고 하자. 무응답 자료에 대한 조건부 기댓값, $E\left[y_i|Y_{obs},\theta=(\mu,\sigma^2)\right], i=r+1,...,n$는 μ이다. 이제 모든 자료($y_i, i=1,\cdots,n$)의 로그우도함수는 아래와 같다.

$$l(\mu,\sigma^2|y_1,...,y_n)\propto -\frac{n}{2}\ln\sigma^2-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{\sigma^2}$$

$$=-\frac{n}{2}\ln\sigma^2-\frac{1}{2}\sum_{i=1}^{n}\frac{y_i^2-2\mu y_i+\mu^2}{\sigma^2}$$

여기서 우도함수 $l(\mu,\sigma^2|y_1,...,y_n)$는 충분통계량 $\sum_{i=1}^{n}y_i$와 $\sum_{i=1}^{n}y_i^2$에 대해 선형이다.

EM 알고리듬의 E-단계에서는 현재 반복의 모수추정값인 $\theta^{(t)} = (\mu^{(t)}, \sigma^{2(t)})$가 주어진 상태에서 충분 통계량의 조건부 기댓값을 다음과 같이 구한다.

$$E\left(\sum_{i=1}^{n} y_i | \theta^{(t)},\ Y_{obs}\right) = \sum_{i=1}^{r} y_i + (n-r)\,\mu^{(t)}$$

$$E\left(\sum_{i=1}^{n} y_i^2 | \theta^{(t)},\ Y_{obs}\right) = \sum_{i=1}^{r} y_i^2 + (n-r)\left[(\mu^{(t)})^2 + \sigma^{2(t)}\right]$$

만일 무응답이 없다면 μ의 MLE는 $\sum_{i=1}^{n} y_i/n$이고 σ^2의 MLE는

$$\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n} = \frac{\sum_{i=1}^{n} y_i^2}{n} - \left(\frac{\sum_{i=1}^{n} y_i}{n}\right)^2$$

이다. 이제 M-단계에서는 E-단계에서 구한 충분 통계량의 기댓값을 위의 무응답이 없는 경우의 MLE식에 대입하여 추정값을 갱신한다. 즉,

$$\mu^{(t+1)} = E\left(\sum_{i=1}^{n} y_i | \theta^{(t)},\ Y_{obs}\right)/n$$

$$\sigma^{2(t+1)} = E\left(\sum_{i=1}^{n} y_i^2 | \theta^{(t)},\ Y_{obs}\right)/n - (\mu^{(t+1)})^2$$

여기서 $\mu^{(t)} = \hat{\mu} = \sum_{i=1}^{r} y_i/r$이라고 하면

$$\mu^{(t+1)} = E\left(\sum_{i=1}^{n} y_i | \theta^{(t)}, Y_{obs}\right)/n = \frac{\sum_{i=1}^{r} y_i}{n} + \frac{(n-r)\mu^{(t)}}{n} = \frac{r}{n}\hat{\mu} + \frac{(n-r)}{n}\hat{\mu}$$

$$= \hat{\mu}$$

이다. 즉, $\mu^{(t)} = \mu^{(t+1)} = \hat{\mu}$이다.

비슷하게 $\hat{\sigma}^2 = \sum_{i=1}^{r} y_i^2/r - \hat{\mu}^2$이면 $\sigma^{2(t)} = \sigma^{2(t+1)} = \hat{\sigma}^2$이다. 물론 이 예제에서 폐쇄형 MLE가 존재하므로 EM 알고리듬은 불필요하다.

< 2장 연습문제 >

1. 단일대체 방법의 장, 단점을 서술하고 단점을 극복하는 대안을 제시해 보아라.

2. 한 지방 자치단체에서 주민의 콜레스테롤 레벨에 관하여 알아보기 위해서 단순임의추출(simple random sampling)방법으로 1,000명의 표본을 추출하여 다음 <표 2.2>와 같은 자료를 얻었다.

<표 2.2> 1,000명의 표본에서 얻어진 자료

나이군	표본수	응답자수	콜레스테롤 레벨	
			평균	표준편차
20대	250	220	220	30
30대	350	270	225	35
40대	280	160	250	44
50대	120	50	270	41

(가) 응답한 개체를 이용하여 평균과 표준오차를 구하라. 정규분포를 가정하고 응답자의 평균 콜레스테롤의 95% 신뢰구간을 구하라.

(나) 가중평균 콜레스테롤과 그것의 평균제곱오차(mean squared error)를 구하고 그에 따른 95% 신뢰구간을 구하라.

(다) (가)와 (나)에서 구한 값들을 비교하고 설명하라.

(라) 이제 인구주택총조사의 결과로부터 이 지방자치단체의 연경구성이 다음과 같음을 알 수 있다. 20대: 20%, 30대: 40%, 40대: 30%, 50대: 10%. 콜레스테롤 평균의 사후-층화 추정값, 표준오차, 그리고 95% 신뢰구간을 구하라.

3. 두 변수 Y_1, Y_2을 300 개체로부터 관측하였다. 자료는 CD의 exercise2_3.sas에 포함되었다. 이 때 변수 Y_1에는 결측이 없고 Y_2에는 결측이 있다. 결측 메커니즘은 임의결측이다. 다음의 방법을 이용하여 Y_2의 평균과 분산을 추정하여라.

(가) 완전히 응답한 개체를 이용한 분석
(나) 회귀 단일대체법
(다) 확률적 회귀 단일대체법

제 3장 무응답을 포함한 자료에 대한 대체 방법 I

< 학습목표 >

(1) 무응답을 포함한 자료에 대한 명시적 모형에 근거한 대체 방법을 소개한다.

(2) 깁스샘플러와 자료확충 기법을 소개한다.

(3) 무응답을 포함한 자료에 대한 다변량 정규분포를 가정한 대체 방법을 소개한다.

(4) 여러 가지 분포를 가진 변수들을 포함한 무응답 자료에 대한 대체 방법을 소개한다.

3.1 다변량 정규분포(multivariate normal distribution)를 가정한 대체 방법

자료 행렬 Y의 p개의 변수들을 $Y_1, Y_2, \cdots, Y_p$로 나타내자. 각 변수들이 특정한 확률분포(probability distribution)를 따른다고 가정하고 분포의 모수들(parameters)을 추정하여 대체를 실시하는 방법을 명시적 모형(explicit model)에 근거한 대체 방법 또는 모수적 모형(parametric model)에 근거한 대체 방법이라 부른다. 명시적 모형에 근거한 대체 방법은 변수들에 대하여 어떤 확률 분포를 가정하느냐에 따라 달라진다.

연속형 자료(continuous data)에 대하여 가장 흔히 가정하는 분포는 정규분포(normal distribution)이다. 자료에서 측정한 p개의 확률변수 $Y_1, Y_2, \cdots, Y_p$를 확률변수 벡터 $(Y_1, Y_2, \cdots, Y_p)'$로 나타내고 이 확률변수 벡터가 평균벡터(mean vector) μ와 분산공분산행렬(variance-covariance matrix) Σ를 가지는 다변량 정

규분포(multivariate normal distribution)를 따른다고 가정하자. 즉, 평균벡터와 분산공분산행렬은 각각

$$\mu = (\mu_1, \mu_2, \cdots, \mu_p)',$$

$$\Sigma = \begin{pmatrix} \sigma_{11}\, \sigma_{12} \cdots \sigma_{1p} \\ \sigma_{21}\, \sigma_{22} \cdots \sigma_{2p} \\ \vdots \quad \vdots \quad \ddots \quad \vdots \\ \sigma_{p1}\, \sigma_{p2} \cdots \sigma_{pp} \end{pmatrix}$$

으로 나타낼 수 있다. 여기서, $\mu_i, i = 1,...,p$,는 확률변수 Y_i의 평균을, $\sigma_{ij}, i = 1, \cdots, p, j = 1, \cdots, p,$ 는 Y_i와 Y_j의 공분산을 나타내고 $\sigma_{ii}, i = 1, \cdots, p,$는 Y_i의 분산을 나타낸다.

확률변수 벡터로부터 서로 독립적으로 (independently) 추출된 n개의 관측값이 존재할 때 각각의 관측값을 소문자 $y_1, y_2, \cdots, y_n$으로 나타내면 y_i의 분포는

$$y_i | \mu, \Sigma \sim iid\, N(\mu, \Sigma), i = 1, \cdots, p$$

와 같다. 이 때, $y_i = (y_{i1}, y_{i2}, \cdots, y_{ip})'$는 i번째 관측값에서 측정된 p개의 변수들의 값을 나타내는 $p \times 1$ 벡터가 되며 *iid*는 서로 독립이고 동일한 분포를 가진다는 (independent and identically distributed) 의미이다.

3.1.1 완전한 자료(complete-data)의 최대우도 추정량

자료가 무응답이 없이 완전하게 응답되었다면 완전한 자료의 우도함수

(complete-data likelihood)는

$$L(\mu, \Sigma | Y) \propto |\Sigma|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)'\Sigma^{-1}(y_i-\mu)\right\}$$

로 나타낼 수 있고 이 우도함수를 최대화(maximizing the likelihood)하는 모수 μ와 Σ의 최대우도추정량(maximum likelihood estimator)은 다음과 같이 구해진다.

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{\mu})'(y_i-\hat{\mu})$$

3.1.2 무응답 패턴과 무응답 자료의 최대우도추정량

무응답 자료의 설명 및 분석을 용이하게 하기 위하여 자료를 무응답 패턴(missingness pattern)에 따라 재정렬한다. 무응답 패턴이란 자료에서 발생하는 서로 다른 응답-무응답 패턴을 의미한다. 간단한 예를 들면, 2개의 변수 Y_1과 Y_2를 측정한 자료에서 발생할 수 있는 무응답 패턴의 종류는

(1) 변수 Y_1과 Y_2 모두 응답함

(2) 변수 Y_1은 응답하지만 Y_2는 응답하지 않음

(3) 변수 Y_1은 응답하지 않았지만 Y_2는 응답함

의 세 가지이다. 만약 변수 Y_1과 Y_2가 모두 무응답이라면 그 관측값은 응답이 하나도 존재하지 않으므로 자료에 포함되지 않아 무응답 패턴 중 하나로 고려할

필요가 없다. 또한, 자료에 따라 위의 세 가지 패턴이 모두 나타나지 않을 수도 있다. 예를 들어, 1장에서 논의한 두 가지 패턴의 자료에서는 (1)과 (2), 또는 (1)과 (3)의 2개의 무응답 패턴만 존재한다. 물론 2개의 패턴에 속하는 관측값의 숫자는 동일할 필요가 없다.

일반적으로 자료는 2개의 변수 대신 p개의 변수가 있으므로 <그림 3.1>과 같이 최대 2^p-1개의 무응답 패턴이 존재 가능하다. 물론 변수의 수가 많아진다면 자료에 나타나지 않는 무응답 패턴도 많아져 실제 무응답 패턴의 개수는 2^p-1보다 훨씬 작게 되고 각 패턴마다 속하는 관측값의 숫자도 일반적으로 각각 다르다. 자료 내에 실제로 존재하는 무응답 패턴의 숫자를 S개라고 가정하고 각 무응답 패턴에 속하는 관측값의 개수를 $n_s, s=1,2,...,S$,라 하자. 이 때, $\sum_{s=1}^{S} n_s = n$으로서 자료전체 관측값의 숫자와 동일하다.

자료행렬 Y가 무응답을 포함하고 있는 경우 $y_i=(y_{i1},y_{i2},\cdots,y_{ip})=(y_{i,obs},y_{i,mis})$ 중 $y_{i,obs}$만을 관측할 수 있다. 즉, 응답된 자료만의 정보에 근거하여 추론을 실시해야 하고 이 때 사용하는 우도함수를 응답된 자료에 근거한 우도함수(observed-data likelihood)라 부른다. 응답된 자료에 근거한 우도함수는

$$L(\mu,\Sigma|\,Y_{obs}) \propto \prod_{s=1}^{S}\prod_{i=1}^{n_s}|\Sigma_s|^{-\frac{n}{2}}\exp\left\{-\frac{1}{2}(y_{i,obs}-\mu_{i,obs})'\Sigma_s^{-1}(y_{i,obs}-\mu_{i,obs})\right\}$$

와 같이 나타낼 수 있다. 여기서 n_s는 무응답 패턴 s를 가지는 관측값의 숫자를 의미하고 $\mu_{i,obs}$와 Σ_s는 평균벡터 μ와 분산공분산행렬 Σ에서 무응답 패턴 s하에서 응답된 변수에 해당되는 평균들과 분산, 그리고 공분산들로 형성된 부분 평균벡터 및 부분 분산공분산행렬을 의미한다. 즉, 무응답 자료에서는 응답된 자료에

근거한 우도함수를 최대화시키는 최대우도추정량을 구해야 하며 이 계산은 완전한 자료의 우도함수를 최대화시키는 경우와 같이 폐쇄형(closed form)으로 표현될 수 없다.

자료가 무응답을 포함하는 경우 자료의 평균 및 분산을 추정하는 문제에서조차 모수를 추정하는 것이 쉽지 않다. 대부분의 분석은 평균과 분산 추정보다는 특정한 모형을 고려하고 그 모형 하에서의 모수의 추정 및 검정에 관심이 있다. 예를 들면 회귀분석을 실시하고 회귀계수를 구한 뒤 이 회귀계수가 유의한지에 관심이 있는 것이다. 자료가 무응답을 포함하는 경우 위에서 보인 바와 마찬가지로 모수의 추정량은 폐쇄형으로 표현되어 질 수 없으므로 2장에서 언급한 바와 같이 각 모형에 맞는 응답된 자료에 근거한 우도함수를 구하고 이 우도함수를 최대화하기 위하여 EM 알고리듬 등의 방법을 이용하여 분석을 시행해야 한다. 문제는 각 모형 및 무응답 형태에 따라 우도함수가 달라지며 각 경우에 알맞은 상용 통계프로그램이 모두 개발되어 있지 않으므로 많은 경우 연구자가 직접 본인의 모형에 적절한 프로그램을 개발하거나 다른 연구자가 개발한 프로그램을 구해서 사용해야 하는 번거로움이 있다. 이 번거로움을 피하는 한 가지 해법은 자료의 무응답을 대체(imputation)하여 완전한 형태의 대체된 자료(imputed data)로 만드는 것이다. 대체된 자료는 무응답을 포함하지 않으므로 상용 통계프로그램을 사용하여 원하는 분석을 자유롭게 시행할 수 있다는 점에서 매우 유용하며 이 이유 때문에 무응답의 대체는 무응답을 포함한 자료에 대한 인기 있는 대처 방법으로 자리 잡게 되었다.

<그림 3.1> 자료행렬 Y에서 관측 가능한 무응답 패턴의 예

무응답 패턴	관측개체	변수 1	2	3	…	p
1	1					
	⋮					
	n_1					
2	1	?				
	⋮	⋮				
	n_2	?				
3	1		?			
	⋮		⋮			
	n_3		?			
4	1			?		
	⋮			⋮		
	n_4			?		
⋮					?	
					⋮	
					?	
$p+1$	1					?
	⋮					⋮
	n_{p+1}					?
$p+2$	1	?	?			
	⋮	⋮	⋮			
	n_{p+2}	?	?			
$p+3$	1	?		?		
	⋮	⋮		⋮		
	n_{p+3}	?		?		
⋮		⋮	⋮	⋮	⋮	⋮
S	1		?	?		?
	⋮		⋮	⋮	…	⋮
	n_S		?	?		?

3.1.3 무응답 자료의 대체에 사용되는 기법

무응답 자료를 명시적 모형 하에서 대체하기 위하여 주로 사용하는 기법이 마르코프 체인 몬테칼로 방법(Markov Chain Monte Carlo 또는 줄여서 MCMC)이다. 마르코프 체인 몬테칼로 방법이란 확률분포들로부터 유사난수(pseudo random number)를 생성하는 기법을 총괄적으로 의미하는데 그 중 대표적인 깁스샘플러(gibbs sampler)(Geman and Geman, 1984)는 다음과 같은 방법으로 유사난수를 생성한다.

- 깁스샘플러

  확률벡터 Z에 대하여 Z의 결합분포(joint distribution)인 $f(Z)$로부터 난수를 생성하고자 하지만 $f(Z)$로부터 직접 난수를 생성하기 어려운 경우를 고려하자. 만약 $Z=(Z_1, Z_2, \cdots, Z_J)$와 같이 Z가 J개의 부분벡터(subvector)로 나누어 질 수 있다면 깁스샘플러는 다음의 조건부 분포(conditional distribution)로부터의 반복 추출을 시행한다.

  반복 시점 $t, t=1,2,3,\cdots,$에 대하여 시점 t에서 Z의 추출된 값을 $Z^{(t)}=\left(Z_1^{(t)}, Z_2^{(t)}, \cdots, Z_J^{(t)}\right)$라 하면 다음 시점인 $t+1$ 시점에서의 Z의 값은 다음과 같은 조건부 분포로부터의 연속적인 추출로 얻어진다.

$$
\begin{aligned}
Z_1^{(t+1)} &\sim f\left(Z_1 | Z_2^{(t)}, Z_3^{(t)}, \cdots, Z_J^{(t)}\right) \\
Z_2^{(t+1)} &\sim f\left(Z_2 | Z_1^{(t+1)}, Z_3^{(t)}, \cdots, Z_J^{(t)}\right) \\
&\vdots \\
Z_J^{(t+1)} &\sim f\left(Z_J | Z_1^{(t+1)}, Z_2^{(t+1)}, \cdots, Z_{J-1}^{(t+1)}\right)
\end{aligned}
$$

위의 반복이 충분이 이루어지면 얻어진 $Z^{(t+1)}$값들은 우리가 추출하려고 하

는 목표분포(target distribution)인 Z의 결합분포 $f(Z)$로부터 추출된 값으로 간주할 수 있다.

깁스샘플러와 밀접하게 연관된 알고리듬이 Tanner and Wong(1987)이 제안한 자료확충(data augmentation)이다.

● 자료확충

확률 벡터 Z가 $Z=(U,V)$와 같이 두 개의 부분벡터(subvector)로 나누어질 수 있고 Z의 결합분포(joint distribution) $f(Z)$로부터 난수를 생성하고자 하지만 $f(Z)$로부터 직접 난수를 생성하기 어려운 경우를 고려하자. 자료확충은 두 개의 조건부 분포 $f(U|V)$와 $f(V|U)$로부터의 연속적인 추출을 통하여 추출하고자 하는 목표분포(target distribution)인 Z의 결합분포 $f(Z)$로부터 유사난수를 생성하는 방법을 의미한다.

반복 시점 $t, t=1,2,3,\cdots,$에 대하여 시점 t에서 $f(Z)$로부터 m개의 값을 추출하고 이를 $Z^{(t)}=\left(\left(u_1^{(t)},v_1^{(t)}\right),\left(u_2^{(t)},v_2^{(t)}\right),\cdots,\left(u_m^{(t)},v_m^{(t)}\right)\right)$라 하면 다음 시점인 $t+1$ 시점에서의 Z의 값은

$$U_i^{(t+1)} \sim f\left(U|v_i^{(t)}\right), i=1,\cdots,m$$

으로부터 m개의 $u_i, i=1,\cdots,m,$값을 추출하고 추출된 u_i 값들에 근거한 조건부 분포인 $f\left(V|u_i^{(t+1)}\right), i=1,\cdots,m,$들의 혼합분포(mixture distribution) $\overline{f}(V|U^{(t+1)})=\frac{1}{m}\sum_{i=1}^{m} f\left(V|u_i^{(t+1)}\right)$를 사용하여

$$V_i^{(t+1)} \sim \overline{f}\left(V|u_i^{(t+1)}\right), i=1,\cdots,m$$

로부터 m개의 $v_i, i=1,\cdots,m,$값들을 추출한다.

추출의 개수 $m=1$인 경우 자료확충은 깁스 샘플러에서 부분벡터의 숫자 $J=2$인 경우에 해당된다. Tanner and Wong(1987)은 자료확충을 이용하여 무응답 자료에서 대한 추론을 실시하는 방법을 제안하였다.

● 무응답 자료에 대한 자료확충

무응답을 포함한 자료의 경우 응답된 자료에 근거한 우도함수는 폐쇄형 해답(closed form solution)을 가지지 않기 때문에 분석이 용이하지 않다. 만약 응답된 자료 Y_{obs}에 결측인 자료 Y_{mis}를 붙여서 확충(augment)시킬 수 있다면 자료는 완전한 자료 $Y=(Y_{obs}, Y_{mis})$의 형태를 가지며 3.1.1에서 나타난 바와 같이 완전한 자료의 우도함수(complete-data likelihood)를 사용하여 쉽게 분석할 수 있다.[2] 즉, 무응답 자료의 분석은 다음의 두 단계로 진행될 수 있다. 반복 시점을 $t, t=1,2,3,\cdots,$라 하고 시점 t에서 모수의 값이 $\theta^{(t)}$라 하면,

(1) Imputation 단계 (I-단계)

시점 $(t+1)$에서 무응답 자료 Y_{mis}는 Y_{obs}와 $\theta^{(t)}$값이 주어졌다고 가정한 Y_{mis}의 조건부 예측분포(conditional predictive distribution)

$$Y_{mis}^{(t+1)} \sim f\left(Y_{mis} | Y_{obs}, \theta^{(t)}\right)$$

로부터 추출한다.

2) 자료확충(data augmentation)이란 Y_{obs}에 결측인 자료 Y_{mis}를 붙여서 자료를 확충한다는 의미로 명명되었다.

(2) Posterior 단계 (P-단계)

시점 $(t+1)$에서 모수의 값 θ는 Y_{obs}와 $Y_{mis}^{(t+1)}$의 값이 자료의 주어진 것으로 간주한 후 완전한 자료의 분포함수

$$\theta^{(t+1)} \sim f\left(\theta | Y_{obs}, Y_{mis}^{(t+1)}\right)$$

로부터 추출한다.

반복이 충분이 이루어지면 얻어진 $\theta^{(t)}, Y_{mis}^{(t)}$ 값들은 $f\left(\theta, Y_{mis} | Y_{obs}\right)$로부터 추출된 값으로 간주될 수 있다. 또한, 이 때 얻어진 $Y_{mis}^{(t)}$ 값들은 무응답 대체를 위하여 사용될 수 있다.

실제로 대체를 실시할 때 모수의 분포가 목표분포(target distribution)로 수렴하도록 시점 t까지 충분히 반복을 실시한 후에 시점 $t+1$에서 얻어진 $Y_{mis}^{(t+1)}$값을 대체값으로 선정한다. 목표분포에 수렴한 시점 t를 선정하는 방법은 시점에 따라 추출된 모수의 시계열 그림(time series plot of parameters)을 사용하게 된다. 이 때, 모든 모수에 대하여 시계열 그림을 그려 모수값들이 시점에 따라 특정한 패턴을 보이지 않고 랜덤하게 변동하면 목표변수로 수렴한 상태를 의미한다. 또한, MCMC 방법으로 추출된 값들은 마르코프 체인(Markov Chain)의 성질에 따라 근접한 시점들에서 추출된 값들 사이에 연관성이 존재하지만 이 연관성은 시점이 멀어지면 줄어들어 없어지는데 이 연관성을 파악하기 위하여 모수들의 시점 간 자기상관 그림(autocorrelation plot)도 사용된다. 자기상관 그림에서 자기상관이 빨리 줄어들수록 안정적이라 할 수 있다.

3.1.4 다변량 정규분포(multivariate normal distribution)를 따르는 무응답 자료의 대체

다변량 정규분포를 따르는 자료행렬 Y가 무응답을 포함하는 경우 무응답 자료의 대체는 다음과 같이 실행된다. 반복 시점을 $t, t=1,2,3,\cdots,$라 하면 시점 t에서 다음의 I-단계와 P-단계를 반복적으로 시행한다.

(1) I-단계 (Imputation 단계)

자료행렬 Y의 n개의 관측값들은 $y_1, y_2, \cdots, y_n$으로 표현되고 각 관측값 $y_i, i=1,\cdots,n,$는 응답된 변수 $y_{i,obs}$와 무응답인 변수 $y_{i,mis}$로 구성되어 함께 $y_i=(y_{i,obs}, y_{i,mis})$로 표현 가능하다. 이 때 n개의 관측값은 서로 독립이므로 $i=1,\cdots,n,$에 대하여 $y_{i,mis}$는

$$y_{i,mis}^{(t+1)} \sim f\left(y_{i,mis}|y_{i,obs}, \theta^{(t)}\right)$$

와 같이 서로 독립적으로 추출된다. 이 때, 자료행렬 Y가 다변량 정규분포를 따르므로 다변량 정규분포 하에서 조건부 확률분포도 다변량 (또는 $y_{i,mis}$가 한 개의 변수만 포함하면 일변량) 정규분포가 되므로 $y_{i,obs}$와 $\theta^{(t)}$이 주어졌을 때 $y_{i,mis}$의 조건부 분포도 정규분포를 따른다. 정규분포를 따르는 $y_{i,mis}$의 평균은 i번째 관측개체에 대하여 무응답인 변수들을 반응변수(response variables)로 놓고 응답인 변수들을 설명변수(explanatory variables)로 설정하여 실시한 회귀분석에서의 예측값(predictive value)이 되며 $y_{i,mis}$의 분산은 이 회귀분석의 잔차공분산 행렬(residual covariance matrix) (또는 $y_{i,mis}$가 한 개의 변수만으로 구성되면 잔차분산)이 된다.

(2) P-단계 (Posterior 단계)

다변량 정규분포의 모수 θ는 평균벡터 μ와 분산공분산행렬 Σ 두 가지이다. I-단계에서 대체된 $y_{i,mis}^{(t+1)}, i=1,\cdots,n,$를 응답된 변수들인 $y_{i,obs}, i=1,\cdots,n,$와 합하여 확충시키면 대체된 자료벡터 $y_i^{(t+1)}=\left(y_{i,obs}, y_{i,mis}^{(t+1)}\right), i=1,\cdots,n,$이 되고 이를 n개의 관측값 전체에 관하여 표현하면 대체된 자료행렬인 $Y^{(t+1)}=\left(Y_{obs}, Y_{mis}^{(t+1)}\right)$이 된다. 대체된 자료의 값을 마치 주어진 값인 것처럼 간주하면 분산공분산행렬 Σ의 분포와 Σ가 주어졌을 때 평균벡터 μ의 조건부분포는 각각

$$\Sigma | Y_{obs}, Y_{mis}^{(t+1)} \sim W^{-1}\left(n-1, \frac{1}{n}\widehat{\Sigma^{-1}}\right),$$

$$\mu | \Sigma, Y_{obs}, Y_{mis}^{(t+1)} \sim N\left(\overline{y}, \frac{1}{n}\Sigma\right)$$

을 따르므로 이 분포들로부터 모수 μ와 Σ을 추출한다. 여기서, $W^{-1}\left(n-1, \frac{1}{n}\widehat{\Sigma^{-1}}\right)$은 자유도(degree of freedom) $n-1$이고 척도모수(scale parameter) $\frac{1}{n}\widehat{\Sigma^{-1}}$을 가지는 역위샤트분포(inverted Wishart distribution)를 의미한다. 위의 식은 Σ의 분포와 Σ이 주어졌을 때 μ의 조건부 분포로 표현되는데 이 두 분포를 곱하면 두 모수의 결합분포를 표현할 수 있기 때문이다.

I-단계와 P-단계를 반복적으로 충분히 시행한 후 마지막으로 추출된 $y_{i,mis}^{(t+1)}$값을 가지고 무응답의 대체를 시행한다. 이 때, 반복을 충분히 시행한다는 것은 P-단계에서 추출된 모수의 값들이 목표분포에서 추출되는 것을 의미한다. 목표분포에

수렴한 시점 t를 선정하는 방법은 3.1.3에서 언급한 추출된 모수의 시계열 그림(time series plot of parameters)을 사용하면 된다.

3.1.4.1 사전정보(prior information)를 이용한 대체

베이지안 자료분석(Bayesian data analysis)에서는 모수에 관한 사전정보(prior information)가 존재한다고 가정하고 이 사전정보를 포함하여 분석을 실시한다. 사전정보는 모수에 관한 사전분포(prior distribution)의 형태로 나타내는데 다변량 정규분포를 따르는 자료에 대한 공액사전분포(conjugate prior distribution)는

$$\Sigma \sim W^{-1}(m, \Lambda),$$
$$\mu|\Sigma \sim N\left(\mu_0, \frac{1}{\tau}\Sigma\right)$$

으로 표현된다. 여기서, 공액사전분포의 모수 $(m, \Lambda, \mu_0, \tau)$는 사전정보에 근거하여 정해지며 $\tau > 0, m \geq p,$ 그리고 $\Lambda > 0$이다. 이 사전분포 하에서의 다변량 정규분포 모수들의 사후분포(posterior distribution)는

$$\Sigma | Y_{obs}, Y_{mis}^{(t+1)} \sim W^{-1}(m+n, \Lambda_1),$$
$$\mu|\Sigma, Y_{obs}, Y_{mis}^{(t+1)} \sim N\left(\mu_1, \frac{1}{\tau+n}\Sigma\right),$$

여기서, $\mu_1 = \left(\frac{n}{\tau+n}\right)\bar{y} + \left(\frac{\tau}{\tau+n}\right)\mu_0,$

$$\Lambda_1 = \left[\Lambda^{-1} + n\hat{\Sigma} + \left(\frac{\tau n}{\tau+n}\right)(\bar{y}-\mu_0)(\bar{y}-\mu_0)'\right]^{-1},$$

으로 나타낼 수 있다.

공액사전분포를 사용하여 무응답을 포함한 자료에 대하여 대체를 실시하기 위해서는 (1) I-단계는 동일하고 (2) P-단계에서 대체된 자료값이 마치 주어진 것처럼 간주한 후 분산공분산행렬 Σ의 분포와 Σ가 주어졌을 때 평균벡터 μ의 조건부 분포 대신 분산공분산행렬 Σ의 사후분포와 Σ가 주어졌을 때 평균벡터 μ의 조건부 사후분포로부터 모수 μ와 Σ을 추출하면 된다.

변수의 숫자가 커짐에 따라 다변량 정규분포의 모수 숫자는 기하급수적으로 늘어나게 된다. 평균벡터 μ는 p개의 추정할 모수를 포함하고 분산공분산 행렬 Σ는 $\frac{p(p+1)}{2}$개의 추정할 모수를 포함하므로 전체 추정해야 하는 모수의 숫자는 $\frac{p(p+3)}{2}$개가 된다. 만약, 관측값의 개수가 모수의 숫자보다 충분히 크지 않다면 모수들을 모두 안정적으로 추정하는 데 문제가 생길 수 있다. 또한 일부 변수들 간에 연관성이 매우 높은 경우 $\hat{\Sigma}$이 비정칙행렬(singular matrix) 또는 비정칙행렬에 가까워지고 $\hat{\Sigma}$의 역함수(inverse matrix)를 계산할 때 문제가 발생할 수 있다. 이와 같이 Σ의 추정에 문제가 생기면 능형회귀(ridge regression)의 개념을 이용한 능형사전함수(ridge prior)의 사용이 도움이 된다(Schafer, 1997). 능형사전함수는 공액사전함수에서 $\tau \to 0$인 극한분포를 의미한다. 능형사전함수 하에서 사후분포(posterior distribution)는

$$\Sigma | Y_{obs}, Y_{mis}^{(t+1)} \sim W^{-1}\left(m+n, [\Lambda^{-1}+n\hat{\Sigma}]^{-1}\right),$$

$$\mu | \Sigma, Y_{obs}, Y_{mis}^{(t+1)} \sim N\left(\overline{y}, \frac{1}{n}\Sigma\right)$$

이 되며 이 사후분포는 $m+n \geq p$, 그리고 $(\Lambda^{-1}+n\hat{\Sigma})>0$이면 적절분포(proper distribution)가 된다. 이 사후분포는 사전분포를 사용하지 않은 경우와 비교해 보면 $\widehat{\Sigma^{-1}}$ 대신 $[\Lambda^{-1}+n\hat{\Sigma}]^{-1}$을 사용하므로 $\Lambda^{-1}=m\times diag(\hat{\Sigma})$ 라고 놓으면 $\hat{\Sigma}$의 대각원소(diagonal element)에 각각 m을 더하여 비정칙행렬의 문제를 해결하는 효과를 지닌다. 이 방법으로 대체를 실시하기 위해서는 (1) I-단계는 동일하고 (2) P-단계에서 분산공분산행렬 Σ의 사후분포, $W^{-1}\left(m+n,[\Lambda^{-1}+n\hat{\Sigma}]^{-1}\right)$, 와 Σ가 주어졌을 때 평균벡터 μ의 조건부 사후분포, $N\left(\overline{y},\frac{1}{n}\Sigma\right)$,로부터 모수 μ와 Σ을 추출하면 된다.

3.1.4.2 다변량 정규분포를 따르는 무응답 자료의 대체 프로그램

무응답 자료에 대하여 다변량 정규분포를 가정하여 대체를 실시하는 프로그램들은 대부분의 상용화된 통계프로그램에 포함되어 있다. 예를 들면 SAS의 MI procedure, SPSS의 Missing Value Analysis(MVA) 모듈, S-Plus의 missing data library를 포함한다. 다음은 SAS MI procedure에서 무응답 대체를 실시하는 방법을 설명한다. <그림 3.2>는 SAS MI procedure의 기본 신텍스(syntax)를 나타낸다.

<그림 3.2> SAS MI procedure

```
PROC MI <options>;
   MCMC <options>;
   VAR variables;
 RUN;
```

PROC MI에서 흔히 사용되는 옵션(option) 및 설명은 다음과 같다.

DATA=SAS-Dataset	무응답 대체를 실시할 자료의 이름을 지정
OUT=SAS-Dataset	대체를 실시한 후 대체된 자료의 이름을 지정
NIMPUTE=number	대체될 자료의 숫자 (단일대체는 1을 다중대체는 대체숫자를 지정) (default는 다중대체의 $m=5$)
SEED=number	난수생성 시 사용하는 0 또는 양수. (동일한 난수를 사용하면 동일한 결과를 얻음; 0이나 default는 컴퓨터 시간에 의해 결정되는 임의의 값)
ROUND=numbers	대체된 값을 지정하는 자리수에서 반올림(round off) (반올림값을 지정하여 대체된 값이 관측값과 같은 자리수로 표현 가능) (default는 반올림없음)
MAXIMUM=numbers	대체 가능한 최대값(maximum)
MINIMUM=numbers	대체 가능한 최소값(maximum) (최대값과 최소값을 지정해 줌으로써 정해진 구간에 속하는 값으로 대체 가능)

MCMC에서 흔히 사용되는 옵션(option) 및 설명은 다음과 같다.

CHAIN=SINGLE/MULTIPLE	단일대체를 실시할 지 다중대체를 실시할 지 지정 (default는 단일대체)

NBITER=number 목표함수로 수렴할 때까지의 반복 숫자. 목표함수로 수렴(converge)하기 이전의 반복을 버린다는 의미로 burn-in period라 불림
(default는 200)

NITER=number 단일연쇄(single chain)를 이용하여 다중대체를 실시할 때 몇 번의 반복 이후 대체값을 선택할 지 지정
(default는 100)

PRIOR=name 사전분포를 사용한 대체를 실시할 때 지정
PRIOR=JEFFREYS (사전분포를 지정하지 않는 방법으로서 default)
PRIOR=RIDGE=number (능형사전함수의 m 지정)
PRIOR=INPUT=SAS-data-set (사전분포의 정보를 포함하는 SAS-Dataset 지정)

INITIAL=<options> MCMC를 실시할 때 모수의 초기값(initial value)을 지정
INITIAL=EM (EM algorithm을 사용하여 최대우도추정량이나 사후최빈값(posterior mode)을 구하고 이 값을 초기값으로 사용하는 방법으로서 default)
INITIAL=INPUT=SAS-data-set (초기값으로 사용될 모수의 추정량을 포함하고 있는 SAS-Dataset 지정)

TIMEPLOT <options> 반복시점에 따른 모수들의 시계열 그림을 출력

COV (분산공분산들의 시계열 그림 출력)

MEAN (평균들의 시계열 그림 출력)

WLF (모수들의 선형 함수 중 가장 늦게 수렴하는 함수(worst liner function)의 시계열 그림 출력)

ACFPLOT <options> 반복시점에 따른 모수들의 자기상관 그림 출력

COV (분산공분산들의 자기상관 그림 출력)

MEAN (평균들의 자기상관 그림 출력)

WLF (모수들의 선형 함수 중 가장 늦게 수렴하는 함수(worst liner function)의 자기상관 그림 출력)

VAR statement에는 대체를 실시할 변수들을 포함시키며 이 문장이 사용되지 않으면 자료행렬의 숫자형 변수 모두가 분석에 포함된다.

이 외에 TRANSFORM statement을 이용하여 변수의 변환을 실시한 후 분석을 시행하거나 MONOTONE statement을 사용하여 단조 무응답 패턴을 지닌 자료에 대한 대체를 실시할 수 있다.

예제 3.1 기업활동실태조사에서의 무응답 대체

예제 2.4에서 고려한 기업활동실태조사를 고려하자. 본 예제에서는 응답을 제공한 10,229개 기업에 대한 자본금(C5), 사업체수(C7), 상용종사자수(C8), 자산총계

(C9), 유형자산 당기 취득액(C18), 매출액(C24), 그리고 영업비용(C41)의 7개 변수 중 자산총계(C9)를 제외한 나머지 6개 변수에서 각각 무응답이 30% 발생하였다고 가정하였다. 이 중 자본금, 사업체수, 상용종사자수 세 가지 변수에서는 무응답이 완전임의로 발생하였고 유형자산 당기 취득액, 매출액, 그리고 영업비용 세 변수에 대해서는 자산총계가 높을수록 무응답이 많이 발생한다고 가정하여 무응답 자료를 생성하였다. SAS MI procedure를 사용하여 생성된 무응답 자료에 대하여 다변량 정규분포에 근거한 대체를 실시하기 위한 프로그램이 <그림 3.3>에 나타난다.

이 자료의 경우 변수들이 오른쪽으로 기운(skewed to the right) 분포를 보여주고 있기 때문에 정규분포 가정에 적합하도록 각 변수들에 대하여 먼저 log 또는 log-log 변환을 실시하고 변환된 자료에 대하여 대체를 실시하였다. 처음 세 변수들은 log-log 변환되어 loglog_C5, loglog_C7, loglog_C8로 이름 지어졌고 나머지 네 변수들은 log 변환되어 log_C9, log_C18, log_C24, 그리고 log_C41로 이름 지어졌다. 이 방법은 자료가 정규분포를 따른다고 가정하는데 모든 변수의 값이 0보다 크므로 대체를 시행할 때 최소값이 0이 되도록 하였다(옵션 MIN = 0 사용). 다변량정규분포를 가정한 후 단일대체를 시행하면 (NIMPUTE=1) <그림 3.4>에 나타난 것과 같은 출력문이 나오는데 이는 분석 정보 및 이 자료에서 나타난 무응답 자료의 패턴을 보여준다. 그 외 EM 알고리듬을 통한 평균의 추정량에 대한 정보도 주어진다.

<그림 3.3> 기업활동실태조사의 무응답 대체를 위한 SAS MI procedure 코드

```
PROC MI DATA=company OUT=micompany NIMPUTE=1
        seed=2340634 MIN=0;
   MCMC NITER=1000 TIMEPLOT ACFPLOT;
   VAR loglog_C5 loglog_C7 loglog_C8 log_C9 log_C18 log_C24 log_C41;
RUN;
```

<그림 3.4> 기업활동실태조사의 무응답 대체를 위한 SAS MI procedure의 주요 출력문 - 무응답 패턴

The MI Procedure

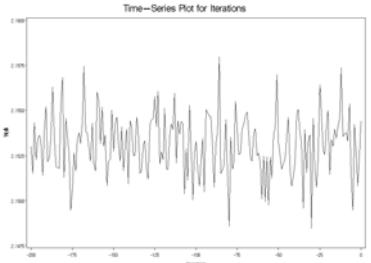
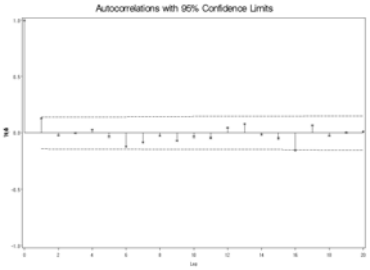
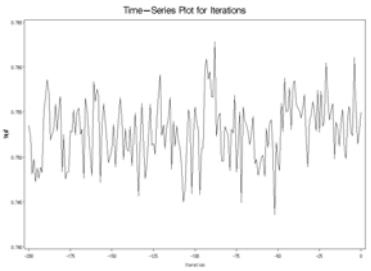
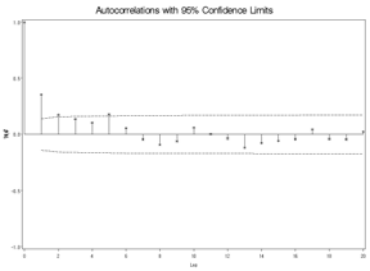
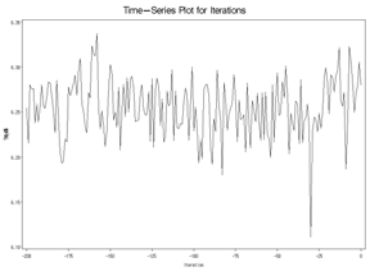
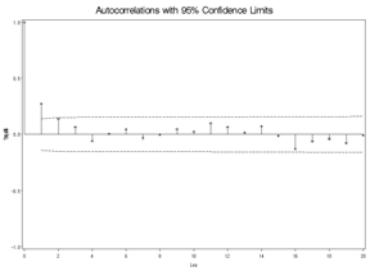
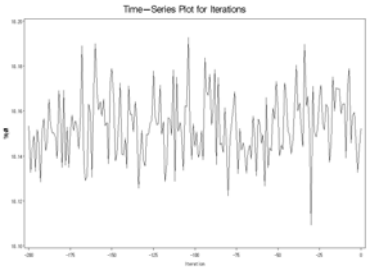
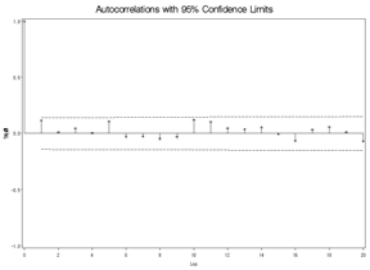
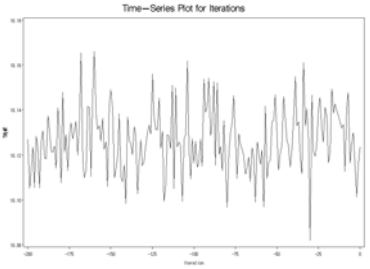
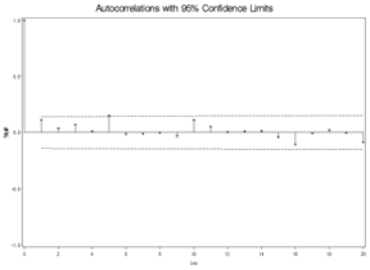
Model Information

Data Set	WORK.A
Method	MCMC
Multiple Imputation Chain	Single Chain
Initial Estimates for MCMC	EM Posterior Mode
Start	Starting Value
Prior	Jeffreys
Number of Imputations	1
Number of Burn-in Iterations	200
Number of Iterations	1000
Seed for random number generator	2340634

Missing Data Patterns

Group	loglog_C5	loglog_C7	loglog_C8	log_C9	log_C18	log_C24	log_C41	Freq	Percent
1	X	X	X	X	X	X	X	1816	17.75
2	X	X	X	X	X	X	.	220	2.15
3	X	X	X	X	X	.	X	204	1.99
4	X	X	X	X	X	.	.	233	2.28
5	X	X	X	X	.	X	X	220	2.15
6	X	X	X	X	.	X	.	212	2.07
7	X	X	X	X	.	.	X	217	2.12
8	X	X	X	X	.	.	.	384	3.75
9	X	X	.	X	X	X	X	757	7.40
10	X	X	.	X	X	X	.	96	0.94
11	X	X	.	X	X	.	X	104	1.02
12	X	X	.	X	X	.	.	80	0.78
13	X	X	.	X	.	X	X	99	0.97
14	X	X	.	X	.	X	.	99	0.97
15	X	X	.	X	.	.	X	108	1.06
16	X	X	.	X	.	.	.	185	1.81
17	X	.	X	X	X	X	X	778	7.61
18	X	.	X	X	X	X	.	91	0.89
19	X	.	X	X	X	.	X	88	0.86
20	X	.	X	X	X	.	.	86	0.84
21	X	.	X	X	.	X	X	94	0.92
22	X	.	X	X	.	X	.	78	0.76
23	X	.	X	X	.	.	X	95	0.93
24	X	.	X	X	.	.	.	181	1.77
25	X	.	.	X	X	X	X	337	3.29
26	X	.	.	X	X	X	.	37	0.36
27	X	.	.	X	X	.	X	38	0.37
28	X	.	.	X	X	.	.	32	0.31
29	X	.	.	X	.	X	X	39	0.38
30	X	.	.	X	.	X	.	35	0.34
31	X	.	.	X	.	.	X	40	0.39
32	X	.	.	X	.	.	.	77	0.75
33	.	X	X	X	X	X	X	752	7.35
34	.	X	X	X	X	X	.	115	1.12
35	.	X	X	X	X	.	X	61	0.60
36	.	X	X	X	X	.	.	106	1.04
37	.	X	X	X	.	X	X	89	0.87
38	.	X	X	X	.	X	.	95	0.93
39	.	X	X	X	.	.	X	103	1.01
40	.	X	X	X	.	.	.	164	1.60
41	.	X	.	X	X	X	X	329	3.22
42	.	X	.	X	X	X	.	36	0.35
43	.	X	.	X	X	.	X	38	0.37
44	.	X	.	X	X	.	.	43	0.42
45	.	X	.	X	.	X	X	44	0.43
46	.	X	.	X	.	X	.	35	0.34
47	.	X	.	X	.	.	X	40	0.39
48	.	X	.	X	.	.	.	76	0.74
49	.	.	X	X	X	X	X	341	3.33
50	.	.	X	X	X	X	.	37	0.36
51	.	.	X	X	X	.	X	56	0.55
52	.	.	X	X	X	.	.	48	0.47
53	.	.	X	X	.	X	X	44	0.43
54	.	.	X	X	.	X	.	35	0.34
55	.	.	X	X	.	.	X	37	0.36
56	.	.	X	X	.	.	.	80	0.78
57	.	.	.	X	X	X	X	143	1.40
58	.	.	.	X	X	X	.	22	0.22
59	.	.	.	X	X	.	X	16	0.16
60	.	.	.	X	X	.	.	19	0.19
61	.	.	.	X	.	X	X	20	0.20
62	.	.	.	X	.	X	.	14	0.14
63	.	.	.	X	.	.	X	12	0.12
64	.	.	.	X	.	.	.	19	0.19

<그림 3.5> 일부 변수들의 평균 추정값들의 반복에 따른 시계열 그림과 자기상관 그림[a)]

변수명	(a) 평균의 시계열그림	(b) 평균의 자기상관 그림
업에 대한 자본금		
glog_C5)		
·업체수(loglog_C7)		
g_C18)		
		
		

[a)]상용종사자수(loglog_C8)와 자산총계(log_C9)의 그림들도 유사하게 나타남

<표 3.1> 단일대체를 통한 평균의 추정값과 이용가능한 자료 분석방법을 시행한 경우 평균의 추정값을 비교 (괄호안은 표준편차)

	완전한 자료	다변량 정규분포하 단일대체	이용가능한 자료 분석
자본금(loglog_C5)	2.15 (0.16)	2.15 (0.16)	2.15 (0.16)
사업체수(loglog_C7)	0.75 (0.28)	0.76 (0.28)	0.76 (0.28)
상용종사자수(loglog_C8)	1.76 (0.14)	1.76 (0.14)	1.76 (0.14)
자산총계(log_C9)	9.96 (1.54)	9.96 (1.54)	9.96 (1.54)
유형자산 당기 취득액(log_C18)	6.21 (2.53)	6.25 (2.66)	5.60 (2.45)
매출액(log_C24)	10.16 (1.41)	10.15 (1.38)	9.71 (1.19)
영업비용(log_C41)	10.14 (1.37)	10.12 (1.34)	9.70 (1.15)

<그림 3.5>는 변수들의 각 시점에서 추출된 평균에 대한 추정값들의 시계열 그림과 자기 상관 그림을 보여준다. 평균의 시계열 그림들은 반복에 따라 특정한 패턴을 보이지 않고 무작위적으로 변동하므로 목표함수(target distribution)에 빨리 수렴하였음을 알 수 있다. 또한, 평균의 자기상관 그림은 시차가 커질때 급격히 줄어들어 반복 시점간 자기상관(autocorrelation)이 크지 않다는 것을 알 수 있다.

<표 3.1>은 다변량정규분포를 가정한 대체를 통한 평균의 추정값과 이용가능한 자료 7160개(유형자산 당기 취득액, 매출액, 그리고 영업비용의 경우 7159개)에 대하여 분석을 시행한 결과, 그리고 무응답이 발생하기 전 완전한 자료의 평균의 추정값을 비교한다. 완전한 자료의 값들이 평균 및 표준편차의 참값을 의미하는데 자본금, 사업체수, 그리고 상용종사자수에서는 다변량 정규분포를 가정한 단일대체와 이용가능한 자료의 분석 결과 얻어진 평균들이 모두 완전한 자료의 평균들과 소수점 2째 자리까지 동일하게 나타나 두 분석 방법 모두 적절한 것으로 나타난다. 그 이유는 이 변수들에서 무응답은 완전임의로 발생하여 평균 추정시 편향이 발생하지 않았기 때문이다. 한편, 유형자산 당기 취득액, 매출액, 그리고 영업

비용의 경우 다변량 정규분포를 가정한 단일대체를 실시한 후 대체된 자료의 평균값은 완전한 자료의 평균값과 비슷한 데 반하여 이용가능한 자료에 근거한 분석방법은 평균이 훨씬 작게 추정되고 있다. 이와 같이 편향이 발생하는 이유는 무응답이 발생할 확률이 자산총계에 의존하기 때문이다. 즉, 자산총계가 클수록 무응답의 확률이 커지므로 이용가능한 자료만에 근거하여 분석을 시행한다면 평균이 과소추정되는 것이다. 마지막으로, 자산총계는 무응답이 발생하지 않는다고 가정하였으므로 세 가지 결과가 모두 동일하게 나타난다.

3.2 여러 가지 분포를 가진 변수들을 포함한 자료에 대한 대체 방법

대부분의 자료는 여러 형태(type)의 변수들을 포함한다. 예를 들면 성별 변수는 “남,” “여” 두 가지 항목의 응답이 가능하고, 몸무게는 0 이상의 숫자로 응답되며 지난 3개월 간 병원 방문횟수는 0 이상의 정수로 응답된다. 즉, 성별 변수는 이산형 변수(binary variable)이고 몸무게는 연속형 변수(continuous variable)로 측정된다. 한편, 병원 방문횟수는 음이 아닌 정수값만 가능하고 많은 사람들의 방문횟수가 0 또는 작은 숫자이지만 일부 사람들의 병원 방문 횟수는 30번 이상으로 매우 크게 나타나 자료가 오른쪽으로 치우친 형태를 갖게 된다. 이렇게 여러 가지 다른 형태의 변수들을 포함한 자료의 경우 3.1의 다변량 정규분포를 가정할 수 없다. 즉, 성별은 이산형 변수이므로 이항 분포를, 몸무게는 연속형 변수이므로 정규분포를, 그리고 병원 방문 횟수는 가산변수(count variable)이므로 포아송 분포를 가정하는 것이 적절할 것이다. 문제는 여러 가지 형태의 변수들을 한꺼번에 다변량 분포로 표현하여 모형을 세우기가 어렵다는 점이다.

Raghunathan, et. al. (2001)은 여러 가지 형태의 변수들을 가진 자료에 대한 대체 방법인 순차회귀 다중대체법(sequential regression multivariate imputation)을

제안하였다. 무응답을 포함한 자료를 자료행렬 Y로 나타내고 n개의 관측값에 대한 k개의 설명변수들의 값을 행렬 X로 표현하다. 여기서, Y의 p개의 변수 $Y_1, Y_2, \cdots, Y_p$는 각각 다른 변수 타입을 가질 수 있고 k개의 설명변수들 또한 여러 가지 다른 변수 타입이 사용 가능하다. 이 자료에 대한 모수적 모형에 근거한 대체를 실시하기 위하여 설명변수 X의 값이 주어졌을 때 p개의 변수 $Y_1, Y_2, \cdots, Y_p$들의 결합 조건부 분포(joint conditional density)는 $f(Y_1, Y_2, \cdots, Y_p | X, \theta_1, \theta_2, \cdots, \theta_p)$인데 변수들의 형태가 다양하므로 이 분포로부터 표본을 직접 추출할 수 없다. 하지만 이 결합 조건부 분포는

$$\begin{aligned} & f(Y_1, Y_2, \cdots, Y_p | X, \theta_1, \theta_2, \cdots, \theta_p) \\ & = f(Y_1 | X, \theta_1) f(Y_2 | X, Y_1, \theta_2) \cdots f(Y_p | X, Y_1, \cdots, Y_{p-1}, \theta_p) \end{aligned}$$

와 같이 여러 개의 조건부 분포(conditional density)들의 곱으로 표현될 수 있다. 여기서, $\theta_i, i = 1, \cdots, p,$는 각 조건부 분포의 모수들의 벡터를 의미한다. p개의 변수 $Y_1, Y_2, \cdots, Y_p$는 각각 다른 변수 타입을 가지므로 각 조건부 분포는 적절한 모형을 가지도록 선택한다. 예를 들어 첫 번째 변수 Y_1이 이산형 변수라면 로짓회귀분석 모형(logistic regression model)을, 두 번째 변수 Y_2가 정규 분포를 따르는 연속형 변수라면 일반회귀모형(regression model)을, 세 번째 변수 Y_3가 가산변수라면 포아송 회귀모형(Poisson regression model)을 가지고 적합할 수 있다. Raghunathan, et. al. (2001)은 위 식의 조건부 분포를 대신하여 $f(Y_j | X, Y_1^{(c)}, \cdots, Y_{j-1}^{(c)}, Y_{j+1}^{(c)}, \cdots, Y_p^{(c)}, \theta_p), j = 1, \cdots, p,$에서부터 모수인 $\theta_i, i = 1, \cdots, p,$를 마르코프 체인 몬테칼로 기법을 이용하여 추출하고 관찰된 자료 Y_{obs}와 추출된 모수들 $\theta_i, i = 1, \cdots, p,$이 주어졌다는 가정 하에서 Y_{mis}의 예측분포(predictive distribution)로부터 Y_{mis}를 추출하는 과정을 반복 시행함으로써 대체를 실시하는 방법을 제안하였다.

대체는 다음과 같은 순서로 이루어진다. 전체 C번의 반복으로 이루어지는데 첫 번째 반복 $c=1$에 대하여 다음의 (1) - (p)를 반복한다.

(1) $f(Y_1|X,\theta_1)$ 모형을 사용하여 모수 θ_1을 추출하고 이 추출된 θ_1과 설명변수행렬 X, 그리고 관찰된 Y_1값이 주어졌다는 조건하에서 Y_1의 무응답 값을 예측분포함수(또는 예측 사후분포함수)로부터 추출한다.

(2) 단계 (1)에서 추출된 Y_1의 무응답값을 Y_1의 응답값과 합하면 변수 Y_1은 무응답이 없도록 대체된다. 이 값을 $Y_1^{(1)}$이라 하면 이 값이 주어졌다고 가정한 후 $f(Y_2|X,\ Y_1^{(1)},\theta_2)$에 대한 적절한 모형을 사용하여 모수 θ_2를 추출하고 이 추출된 θ_2과 설명변수행렬 X, 무응답이 대체된 Y_1, 관찰된 Y_2값이 주어졌다는 조건하에서 Y_2의 무응답값을 예측분포함수(또는 예측 사후분포함수)로부터 추출한다.

⋮

(p) 단계 (1)부터 단계 (p-1)에서 대체된 $Y_1^{(1)},\cdots,Y_{p-1}^{(1)}$을 사용하여 $f(Y_p|X,\ Y_1^{(1)},\cdots,Y_{p-1}^{(1)},\theta_p)$에 대한 적절한 모형을 사용하여 모수 θ_p를 추출하고 이 추출된 θ_p과 설명변수행렬 X, $Y_1^{(1)},\cdots,Y_{p-1}^{(1)}$, 관찰된 Y_p값이 주어졌다는 조건하에서 Y_p의 무응답값을 예측분포함수(또는 예측 사후분포함수)로부터 추출한다.

두 번째 이후의 반복 $c=2,\cdots,C$까지는 위의 (1)-(p) 단계에서 설명변수로 다른 모든 변수들을 수정하도록 변경된다. 즉,

(1) $f\left(Y_1|X,\, Y_2^{(c-1)},\cdots,Y_p^{(c-1)},\theta_1\right)$ 모형을 사용하여 모수 θ_1을 추출하고 이 추출된 θ_1과 설명변수행렬 X, 전 반복에서 대체된 $Y_2^{(c-1)},\cdots,Y_p^{(c-1)}$, 그리고 관찰된 Y_1값이 주어졌다는 조건하에서 Y_1의 무응답값을 예측분포함수(또는 예측 사후분포함수)로부터 추출한다.

(2) 단계 (1)에서 추출된 Y_1의 무응답값을 Y_1의 응답값과 합하면 변수 Y_1은 무응답이 없도록 대체된다. 이 값을 $Y_1^{(c)}$이라 가정한 후 $f\left(Y_2|X,\, Y_1^{(c)},\, Y_3^{(c-1)},\cdots,Y_p^{(c-1)},\theta_2\right)$에 대한 적절한 모형을 사용하여 모수 θ_2를 추출하고 이 추출된 θ_2과 설명변수행렬 X, 대체된 $Y_1^{(c)},\, Y_3^{(c-1)},\cdots,Y_p^{(c-1)}$, 관찰된 Y_2값이 주어졌다는 조건하에서 Y_2의 무응답값을 예측분포함수(또는 예측 사후분포함수)로부터 추출한다.

⋮

(p) 단계 (1)부터 단계 (p-1)에서 대체된 $Y_1^{(c)},\cdots,Y_{p-1}^{(c)}$를 사용하여 $f\left(Y_p|X,\, Y_1^{(c)},\cdots,Y_{p-1}^{(c)},\theta_p\right)$에 대한 적절한 모형을 사용하여 모수 θ_p를 추출하고 이 추출된 θ_p과 설명변수행렬 X, 대체된 $Y_1^{(c)},\cdots,Y_{p-1}^{(c)}$, 관찰된 Y_p값이 주어졌다는 조건하에서 Y_p의 무응답값을 예측분포함수(또는 예측 사후분포함수)로부터 추출한다.

이 때 반복의 수 C는 안정된 대체값을 얻을 수 있도록 결정되는데 대부분의 자료에서 10번 정도면 충분하다는 것이 경험적으로(empirically) 알려져 있다. 이 방법은 추정된 함수 $\hat{f}\left(Y_j|X,\, Y_1^{(c)},\cdots,Y_{j-1}^{(c)},\, Y_{j+1}^{(c)},\cdots,Y_p^{(c)},\theta_p\right)$을 가지고

$f\left(Y_j|X, Y_1^{(c)},\cdots, Y_{j-1}^{(c)}, Y_{j+1}^{(c)},\cdots, Y_p^{(c)},\theta_p\right)$을 근사(approximate)하므로 SIR algorithm (Rubin, 1987b)이나 rejection algorithm(Gelman et. al., 2004)을 사용하여 근사를 개선하는 것이 바람직하다.

3.2.1 여러 가지 분포를 따르는 변수들을 포함한 자료에 대한 대체 프로그램

무응답 자료가 여러 가지 다른 타입의 변수들을 포함하는 경우 대체를 실시하기 위하여 SAS 내에서 Macro를 불러 시행하거나 Windows나 Linux 하에서 독립적으로 실행될 수 있는 프로그램 IVEware를 사용할 수 있다(Raghunathan, Solenberger, and Hoewyk, 2002). 이 프로그램은 여러 가지 분포를 따르는 변수들에 대하여 적절한 조건부 분포를 설정하여 대체를 가능하게 할 뿐 아니라 설문조사(survey)에서 종종 발생하는 자료의 특성을 고려하여 대체가 가능하게 한다.

설문조사에서는 문항이 해당되지 않는 사람들에 대하여 건너뛰도록(skip) 설계된 질문지를 흔히 사용한다. 예를 들면, 흡연 관련 질문 중 흡연 기간에 관한 질문은 흡연자에게만 해당되는 질문이므로 비흡연자들에게는 건너뛰도록 만들어지고 이는 모든 비흡연자의 흡연 기간이 무응답으로 남도록 만든다. 1.1에서 언급한 바와 같이 이 문항은 모집단이 흡연자이므로 비흡연자에게는 해당되는 문항이 아니고 비흡연자에 대한 무응답은 실제로는 무응답이 아니므로 대체를 실시하지 않아야 한다. 이런 문항에 대한 대체는 흡연자들만 대체를 실시하도록 제약(restriction)하에서 실시되어야 한다.

가능한 값에 경계(bound)가 존재하는 변수들의 경우 이 경계 내부의 값으로 대체가 실시되어야 한다. 예를 들어 나이, 키, 몸무게 등은 모두 음수가 될 수 없으며

흡연 기간은 자신의 나이보다 길 수 없다. 이와 같은 경계가 존재하는 변수에 대하여 정규분포 가정을 적용한다면 무응답 값이 가능한 경계 밖의 값으로 대체될 수 있으므로 이와 같은 모순을 막기 위하여 절단모형(truncated model)으로부터 대체를 실시하여야 한다.

IVEware 프로그램은 <표 3.2>와 같이 변수들의 타입에 따라 다른 회귀 모형을 다룰 수 있다. 변수의 타입 중 혼합형 변수란 자료의 값이 0과 연속형 변수가 혼합되어 나타나는 형태로서 예를 들면 지난 일주일간 흡연량과 같은 변수에서 흔히 나타난다. 즉, 이 변수의 경우 비흡연자의 흡연량은 모두 0이 되고 흡연자의 흡연량은 연속형 값으로 나타나게 된다. 혼합형 자료에 흔히 사용되는 두단계 모형(two-stage model)이란 우선 첫 번째 단계에서 흡연자인지 비흡연자인지를 로지스틱 회귀모형을 이용하여 적합시키고 두 번째 단계에서는 흡연자의 흡연량은 선형회귀모형을 통하여 적합하는 두 단계로 모형을 세우는 방법을 의미한다 (Schafer and Harel, 2002).

<표 3.2> IVEware에서 변수 타입에 따라 대체를 위해 고려할 수 있는 회귀모형

변수의 타입	회귀모형
연속형	정규분포 가정한 선형회귀모형
이산형	로지스틱 회귀모형
범주형	범주형 자료를 위한 로짓회귀모형
가산형	포아송 로그선형 모형
혼합형	두단계 모형(two-stage model)

다음은 SAS내에서 IVEware를 사용하여 무응답 대체를 실시하는 매크로 모듈 IMPUTE의 사용법을 설명한다. <그림 3.6>는 IVEware IMPUTE 모듈의 기본 신텍스(syntax)를 나타낸다.

<그림 3.6> IVEware IMPUTE 모듈

```
%IMPUTE (NAME=filename, DIR=);
    DATAIN filename;
    DATAOUT filename;
    DEFAULT variable type;
    CATEGORICAL variables;
    RESTRICT variable(logical expression);
    BOUNDS variable(logical expression);
    ITERATIONS number;
    MULTIPLES number;
    SEED number;
 RUN;
```

매크로 %IMPUTE에서 사용되는 키워드(keyword) 및 설명은 다음과 같다.

NAME=filename	setup file의 이름을 지정
DIR=directory	setup file과 output file이 저장되는 컴퓨터 내의 directory 지정

DATAIN은 대체될 무응답 자료를 지정한다.

DATAOUT은 대체가 실시된 후 대체된 자료의 파일명을 지정한다.

DEFAULT는 default로 생각될 변수의 타입을 지정하는데 일반적으로 가장 많은 변수 타입을 선택한다. 그 외에 CONTINUOUS, CATEGORICAL, COUNT, MIXED 문은 각각 연속형, 범주형, 가산형, 혼합형 변수들을 지정할 수 있다.

RESTRICT 문에서는 제약을 둘 변수들 및 각 변수별 제약식을 지정한다.

BOUNDS 문에서는 경계값을 가지는 변수들 및 각 변수별 경계를 식으로 지정한다.

ITERATIONS 문에서는 결측값을 추출하는 깁스샘플러의 반복의 수 C를 지정한다. 2 이상의 값을 지정할 수 있다.
MULTIPLES 문에서는 대체의 숫자를 지정하는데 default는 단일대체인 1이 된다.

SEED 문은 난수생성 시 사용하는 0 또는 양수값을 지정하는 데 동일한 양수를 사용하면 동일한 결과를 얻을 수 있다. 0을 지정하면 예측값이나 회귀계수를 분포로부터 추출하는 대신 값 자체를 사용하고 default는 컴퓨터 시간에 의해 결정되는 임의로 값이 된다.

예제 3.2 기업활동실태조사의 무응답 대체

예제 3.1의 기업활동실태조사의 무응답을 여러 가지 분포를 가정하여 대체하기 위하여 IVEware를 사용하였다. 이 프로그램은 여러 가지 타입의 변수를 포함하는 것이 가능하기 때문에 가산변수인 두 변수 사업체수(C7)와 상용종사자수(C8)는 포아송 분포를 따르도록 설정하였다. 나머지 5개의 변수 자본금(C5), 자산총계(C9), 유형자산 당기 취득액(C18), 매출액(C24), 그리고 영업비용(C41)은 예제 3.1에서와 마찬가지로 log 또는 log-log 변환을 실시하고 변환된 자료에 대하여 대체를 실시하였다. 그 외에 설명변수로 각각 16개 범주를 가지는 행정구역(C2)과 산업분류(대)(C3)를 범주형 변수로 포함시켜 추정의 설명력을 높였다. <그림 3.7>는 프로그램 코드를 보여주고 <표 3.3>은 여러 가지 분포를 가정하여 단일대체된 평균의 추정값, 이용가능한 56650개 관측값에 대한 평균의 추정값, 그리고 무응답이 발생하기 전 완전한 자료의 평균의 추정값을 비교한다.

<그림 3.7> 기업활동실태조사의 무응답 대체를 위한 IVEware IMPUTE 모듈 코드

```
%IMPUTE (NAME=impute, DIR=c:);
    DATAIN company;
    DATAOUT micompany;
    DEFAULT continuous;
    CATEGORICAL c2 c3;
    COUNT c7 c8;
    SEED 100;
 RUN;
```

<표 3.3> 단일대체를 통한 평균의 추정값과 이용가능한 자료 분석방법을 시행한 경우 평균의 추정값을 비교 (괄호안은 표준편차)

	완전한 자료	여러 가지 분포를 가정한 단일대체	이용가능한 자료 분석
자본금(loglog_C5)	2.15 (0.16)	2.15 (0.16)	2.15 (0.16)
사업체수(C7)	5.39 (23.88)	5.40 (19.45)	5.31 (22.53)
상용종사자수(C8)	269.18 (1359.34)	256.19 (860.95)	247.68 (885.07)
자산총계(log_C9)	9.96 (1.54)	9.96 (1.54)	9.96 (1.54)
유형자산 당기 취득액 (log_C18)	6.21 (2.53)	6.25 (2.63)	5.60 (2.45)
매출액(log_C24)	10.16 (1.41)	10.13 (1.36)	9.71 (1.19)
영업비용(log_C41)	10.14 (1.37)	10.13 (1.35)	9.70 (1.15)

결과는 <표 3.1>의 대체결과와 비슷하다. 자본금, 사업체수, 그리고 상용종사자수에서는 여러 가지 분포를 가정한 단일대체와 이용가능한 자료의 분석 결과 얻어진 평균들이 모두 완전한 자료의 평균들과 비슷하게 나타나지만 여러 가지 분포

를 가정한 단일대체의 결과가 완전자료에 더 근접하게 나타난다. 한편, 유형자산 당기 취득액, 매출액, 그리고 영업비용의 경우 여러 가지 분포를 가정한 단일대체를 실시한 후 대체된 자료의 평균값은 완전한 자료의 평균값과 비슷한 데 반하여 이용가능한 자료에 근거한 분석방법은 평균이 훨씬 작게 추정되고 있다. 이 분석에서는 예제 3.1과 달리 변환되지 않은 사업체수(C7)와 상용종사자수(C8)에 대한 변환하지 않은 원변수의 평균의 추정값을 제공하므로 원변수에 대한 재변환없이 추정이 가능하여 편리하다.

< 3장 연습문제 >

1. 다음의 각 항목에 대하여 참, 거짓을 판별하고 판별의 이유를 설명하시오.
   (가) 다변량 정규분포 가정 하에서 대체를 실시할 때 일반 선형 회귀를 통하여 최적의 대체값을 구할 수 있다.
   (나) 자료가 연속형 변수들 뿐 아니라 이산형 분포를 가진 변수도 포함하고 있을 때 다변량 정규분포 하에서의 대체는 항상 추정량의 편향을 가져온다.

2. 다음의 질문에 답하시오.
   (가) 대체를 실시하는 모형을 결정할 때 가장 중요하게 고려해야 하는 요소는 무엇인가?
   (나) 무응답 자료에 대하여 대체를 실시하였다. 동일한 대체를 다시 얻을 수 있는가? 얻을 수 있다면 어떤 방법을 사용해야 하는가?
   (다) MCMC 기법을 사용하여 대체를 실시할 때 추출된 모수값들의 목표분포로의 수렴여부를 평가하는 방법을 설명하시오.

3. 다음의 자료는 2007년 기업활동실태조사 자료의 일부분이다. 자료에 대한 설명은 <표 3.4>에 나타난다. 제공된 자료의 무응답은 본 장에서 배운 방법을 복습하기 위하여 임의로 생성되었다. 자료는 CD의 exercise3_3.sas에 포함되어 있다. 3장에서 배운 대체방법들을 사용하여 대체를 실시하시오. 사용된 대체 방법들을 변수별 평균 및 표준오차를 사용하여 비교하시오. 이 자료에 대하여 적합한 것으로 생각되는 한 가지 대체 방법을 추천하시오.

<표 3.4> 기업활동실태조사 자료 일부 변수의 명칭 및 설명

변수명	변수 설명
C2	행정구역
C4	산업분류(대)
C5	기업의 자본금
C6	외국자본금 비율
C7	사업체수
C15	부채의 총계
C16	자본총계

제 4장 무응답을 포함한 자료에 대한 대체 방법 II

< 학습목표 >

(1) 무응답을 포함한 자료에 대한 핫덱대체 방법을 소개한다.

(2) 무응답을 포함한 자료에 대한 명시적 모형과 내재적 모형의 혼합 기법에 근거한 대체 방법을 소개한다.

(3) 다중대체된 자료에 대한 분석 및 추론 방법을 소개한다.

4.1 핫덱대체 방법

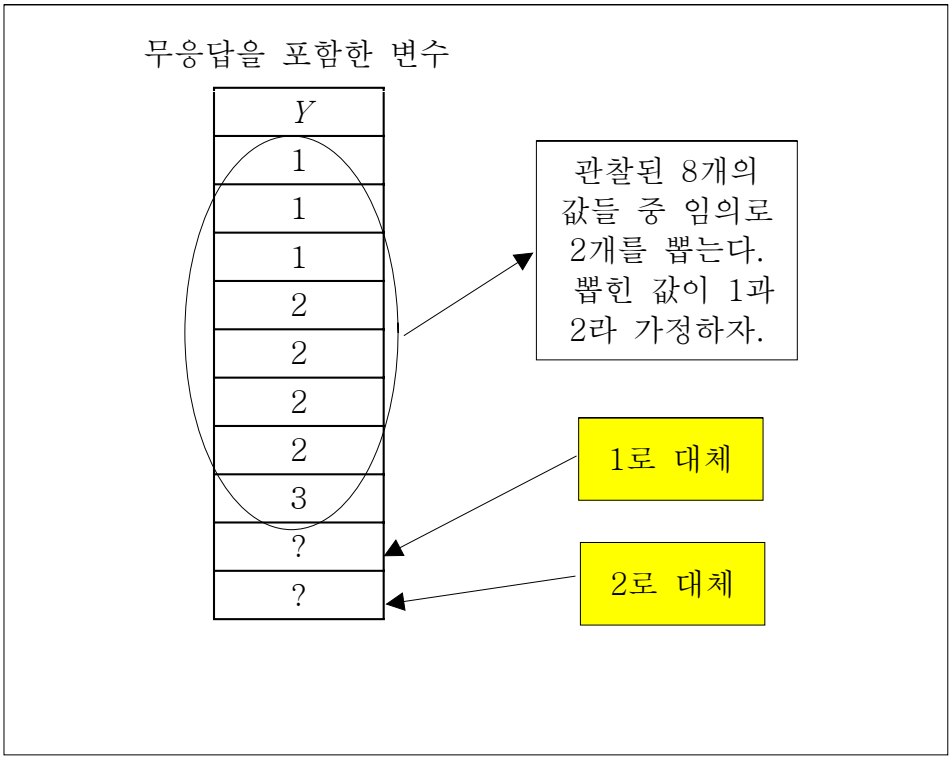
자료의 분포를 가정하지 않고 대체를 실시하는 방법 중 가장 흔히 사용되는 대체 방법이 핫덱대체(hotdeck imputation) 방법이다. 핫덱대체 방법은 무응답값의 대체를 위하여 자료내의 응답값을 사용하는 방법이다. 핫덱대체에서 무응답값은 자료내의 응답된 값을 가지고 대체되므로 응답값이 무응답값에 기증되었다는 의미로 대체에 사용된 응답 개체를 기증자(donor)로, 무응답이 발생하여 응답값의 기증을 받은 개체를 수증자(donee)라 부른다. 핫덱은 기증자를 선택하는 방법에 따라 여러 가지 방법으로 분류되는데 여기서는 흔히 사용되는 몇 가지 핫덱대체 방법들을 소개하고자 한다.

4.1.1 단순임의 핫덱대체 방법(Hotdeck by Simple Random Sampling)

단순임의 핫덱대체 방법은 자료내 각각의 무응답값에 대하여 한 개의 응답값을

임의로 선택하여 대체하는 방법이다. <그림 4.1>는 한 개의 변수에서 무응답이 발생할 때 단순임의 핫덱대체를 시행하는 간단한 예를 설명한다. 무응답을 포함한 자료 Y가 한 개의 변수를 포함하고 이 변수에 무응답이 발생하는 경우를 고려하자. 전체 관측값의 숫자는 10개인데 그 중 2개의 관측값에서 무응답이 발생한다면 응답된 8개의 응답값이 2개의 무응답 개체를 대체하기 위하여 기증자로 사용된다. 8개의 기증자 중에서 임의로 2개의 응답값이 복원(with replacement) 또는 비복원(without replacement)으로 추출된다. 핫덱대체를 실제로 적용할 때에는 동일한 응답값이 여러 결측값에 대하여 기증자로 사용되는 것을 방지하기 위하여 복원추출보다는 비복원추출이 선호되는 경향이 있다. 예를 들어, 8개의 응답값 중에서 2개를 추출할 때 추출된 값이 1과 2라면 첫 번째 무응답값에 1의 값을 두 번째 무응답값에 2를 대체한다.

<그림 4.1> 단순임의 핫덱대체 방법



자료 Y가 여러 개의 변수를 포함한 경우에도 이 방법을 확장하여 사용할 수 있다. 각 변수별로 단순임의로 무응답값의 개수만큼 응답값을 추출한 후 그 변수의 무응답값에 대한 기증자로 사용하여 대체를 실시할 수 있다. 이 방법의 문제점은 이 방법으로 대체된 값에 근거한 추정량은 무응답 자료 메커니즘이 완전임의결측이 아니라면 편향이 발생한다는 데 있다. 따라서 이 방법의 적용은 한정적일 수 밖에 없다.

예제 4.1 2008년 사회조사의 무응답 대체

2008년에 통계청에서 시행된 사회조사는 국민의 일상생활과 관련하여 현재 처한 상황들을 조사하여 앞으로 나아가야 할 방향을 모색하고자 시행되고 있는데 매년 12개 부분 중 3 - 4개 부문에 대하여 조사를 실시하며 대상은 약 20,000가구의 만 15세 이상 상주 가구원이다. 2008년 사회조사는 인적사항, 교육부문, 안전부문, 환경부문의 44개 조사 항목에 대한 문항들을 포함하고 있다. 본 예제에서는 가구주 대상 설문지 문항 중 학생인 자녀가 있는 가구주에게 자녀교육비의 부담정도(C44)를 질문한 문항에 대한 대체를 고려한다. 이 문항은 “매우 부담스럽다”(1), “약간 부담스럽다”(2), “보통이다”(3), “별로 부담스럽지 않다”(4), 그리고 “전혀 부담스럽지 않다”(5)의 5개 범주로 나뉘어져 있고 대부분의 응답자가 “매우 부담스럽다” 또는 “약간 부담스럽다”로 응답하여 범주 별 빈도수가 크게 다른 경우였다. 응답자 8,115명 중 30%에 대하여 무응답을 생성하였는데 학력이 낮고 배우자가 있는 경우에 무응답이 더 많이 발생한다고 가정하였다. 본 예제의 목적은 사회조사에서 발생하는 무응답에 대한 가장 적절한 대체 방법을 제시하는 것이 아니라 핫덱대체 방법의 예제를 보여주기 위한 것임을 명시한다.

<표 4.1>은 무응답 발생 전 완전한 자료, 무응답을 무시한 채 분석을 실시한 완전히 응답한 개체를 이용한 분석(complete-case analysis), 그리고 단순임의 핫덱대체를 실시한 후 대체된 자료의 자녀 교육비 부담 정도에 대한 응답 비율을 비교한다. 완전임의 핫덱대체를 실시한 후 응답항목의 비율은 완전히 응답한 개체를 이용한 분석에서 구한 비율과 거의 비슷한 반면 무응답이 발생하기 전 완전한 자료의 비율과는 다르게 나타난다. 이 차이는 대체한 값이 응답한 값들 중 완전히 임의로 선택되어 발생하게 된다.

<표 4.1> 완전임의 핫덱대체를 통한 자녀 교육비 부담 정도에 대한 응답 비율을 완전한 자료 및 완전히 응답한 개체를 이용한 분석 방법과 비교

응답항목	완전한 자료	완전임의 핫덱대체	완전히 응답한 개체를 이용한 분석
매우 부담스럽다 (1)	39.57	37.12	37.11
약간 부담스럽다 (2)	40.17	41.18	41.25
보통이다 (3)	15.76	17.19	17.14
별로 부담스럽지 않다 (4)	3.89	3.94	3.90
전혀 부담스럽지 않다 (5)	0.60	0.57	0.60

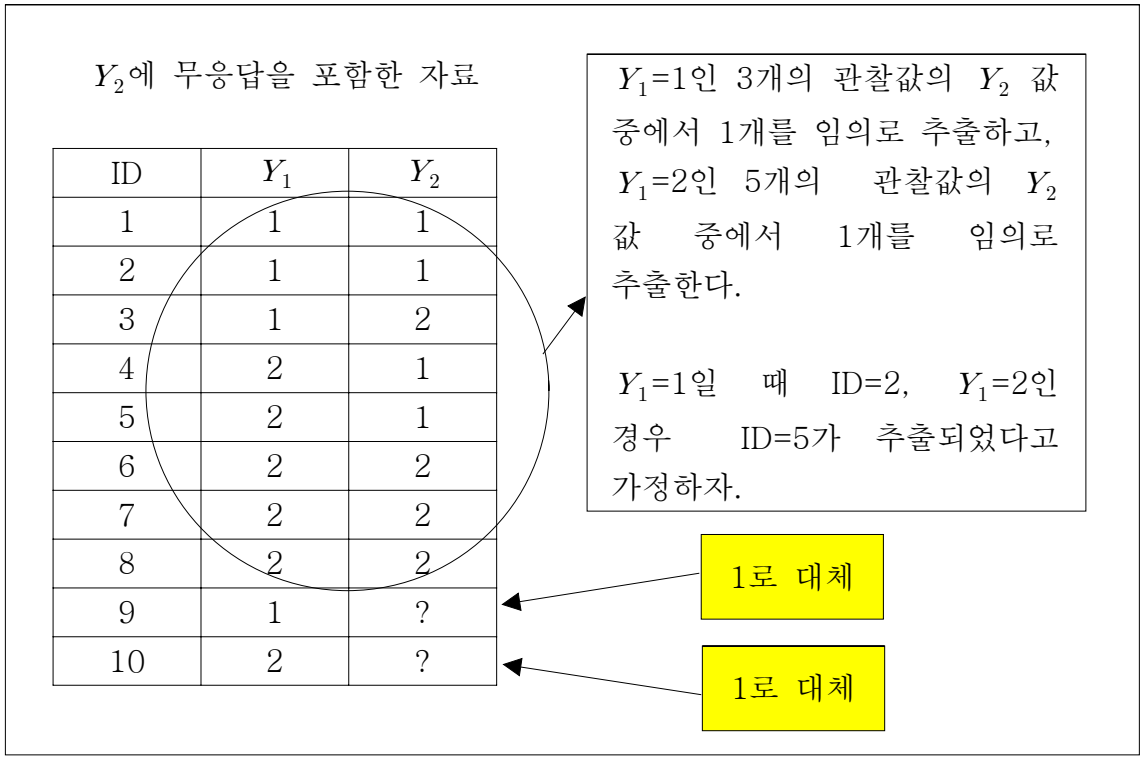
4.1.2 대체군을 이용한 핫덱대체 방법(Hotdeck Within Adjustment Cells)

무응답 자료에 대한 대체를 실시할 때 관찰된 다른 변수들이 동일하거나 비슷한 응답값이 기증자로 선정된다면 완전임의로 응답값을 선정하여 대체를 실시하는 것보다 정확한 대체를 실시할 수 있다. 예를 들어 소득 변수에서 무응답이 발생한 경우 응답자 중 완전 임의로 기증자를 선정하는 것 보다 소득에 대하여 응답하지 않은 사람과 동일한 성별, 나이, 자산 수준을 가진 응답자들로 대체군을 형성하고

이 대체군 내에서 임의로 한 응답자를 추출하여 이 응답자의 소득을 가지고 무응답자의 소득을 대체한다면 더 정확한 대체를 실시할 수 있을 것으로 예상된다. <그림 4.2>는 대체군을 이용한 핫덱대체의 간단한 예를 보여준다. 자료가 관찰값 식별번호인 ID 변수와 두 개의 변수(Y_1과 Y_2)를 포함하고 두 변수 중 Y_2에서만 무응답이 발생할 때 완전히 응답된 Y_1의 값을 가지고 대체군을 형성하고 이 대체군을 이용한 핫덱대체를 시행하는 방법을 설명한다. 무응답은 9번째와 10번째 관찰값(ID = 9와 10)에서 발생하였다. 즉, ID = 9는 $Y_1 = 1$의 값을, ID = 10은 $Y_1 = 2$을 가지는 것으로 관측되었으나 Y_2 변수의 값은 무응답으로 관측되지 않았다. 이 경우 ID = 9는 $Y_1 = 1$이므로 $Y_1 = 1$인 관찰단위 ID = 1, 2, 3 세 응답자의 값으로 대체군을 형성하고 이 대체군 안에서 임의로 한 개의 ID를 추출한다. 예를 들어 ID = 2가 추출되었다면 ID = 2의 Y_2값인 1을 가지고 ID = 9의 Y_2값을 대체한다. 마찬가지로 ID = 10은 $Y_1 = 2$이므로 $Y_1 = 2$인 관찰단위 ID = 4부터 ID = 8의 다섯 명의 응답자의 값으로 대체군을 형성하고 이 대체군 안에서 임의로 한 개의 ID를 추출한다. 예를 들어 ID = 5가 추출되었다면 ID = 5의 Y_2값인 1을 가지고 ID = 10의 Y_2값을 대체한다.

대체군을 이용한 핫덱대체의 성능은 대체군의 형성에 의존한다. 즉, 대체군 내에서 응답값과 무응답값의 분포가 동일하도록, 즉 대체군을 형성한 변수들이 주어졌을 때 무응답 발생 메커니즘이 무시할 수 있는 메커니즘이 되도록 대체군을 형성한다면 대체로 인한 편향(bias)이 발생하지 않을 것이다. 이것은 대체군을 형성하기 위하여 무응답이 발생한 변수와 연관되어 있는 변수들을 포함함으로써 달성할 수 있을 것이다. 이 목적은 가능한 한 많은 변수를 고려하여 대체군을 형성할수록 달성될 가능성이 높다. 하지만 대체군을 형성하기 위하여 변수가 추가될수록 대체군의 숫자가 기하급수적으로 늘어나는 데 문제가 있다. 예를 들어 소득변수에서 발생하는 무응답을 대체하기 위하여 대체군을 형성 하는 변수로 성별(“남”,

<그림 4.2> 대체군을 이용한 핫덱대체 방법



“여”로 구분)만을 사용하면 대체군의 숫자는 2개뿐이지만 연령(“0-10세,” “11-20세,” “21-30세,” “31-40세,” “41-50세”, “51-60세”, “61-70세”,“71세 이상”으로 구분)도 포함시키면 대체군의 숫자는 2 × 8 = 16개로 늘어나며 거주지 구분(“시,” “도,” “군”) 및 자산정도(“1천만원 미만,” “1천 초과 ~ 1억,” “1억 초과 ~ 5억,” “5억 초과 ~ 10억,” “10억 초과”)를 추가하면 2 × 8 × 3 × 5 = 240개로 크게 늘어난다. 대체군의 숫자가 늘어나면 일부 대체군에 속하는 응답값을 가진 관측값의 수가 적거나 일부 대체군에는 무응답은 있으나 관측값이 없어 무응답값에 대한 기증자를 찾지 못하는 문제점이 발생할 수 있다. 예를 들어 소득에 대한 무응답은 소득이 많은 사람들 중에서 많이 발생하는 경향이 있다. 성별, 연령, 거주지 구분, 그리고 자산정도를 가지고 대체군을 형성하였는데 “여,” “21-30세,”, “도”에 거주하고 자산이 “10억 초과”인 경우 무응답이 발생하였는데 동일한 대체군에 속하는 응답자이 하나도 없는 경우가 발생하는 것이다. 또한 특

정 대체군 내에서 응답자의 숫자가 무응답자의 숫자보다 적은 경우도 발생할 수 있으며 이 경우 이 대체군에 속하는 일부 무응답자는 기증자를 발견할 수 없어 대체될 수 없다. 물론, 비복원추출 대신 복원추출을 사용하는 경우 이 문제는 덜 발생하지만 한 명의 응답자가 여러 무응답자에 대하여 대체되어 대체된 자료의 분산이 부정확하게 추정되는 문제점이 발생할 수 있다.

대체군을 이용한 핫덱대체 기법에서 이와 같이 대체군을 만드는 변수들이 많아져 기증자를 찾기 어려운 경우에 흔히 사용되는 방법은 대체군을 형성하는 변수 일부를 생략하고 기증자를 찾는 방식이다. 예를 들어 “여,” “21-30세,”, “도”에 거주하고 자산이 “10억 초과”인 경우 무응답이 발생하였는데 대체군에 속하는 응답이 하나도 없는 경우 연령을 대체군 형성에서 제외시키고 나머지 세 변수들로만 다시 대체군을 만든 다음 기증자를 찾는다. 이 단계에서도 기증자를 찾을 수 없다면 다시 거주지 구분 변수를 대체군 형성 변수에서 제외시키고 성별과 자산 정도만을 가지고 다시 대체군을 형성한 다음 기증자를 찾는 방식을 취한다. 즉, 기증자를 찾을 때까지 대체군 형성 변수의 숫자를 줄여가는 것이다.

대체군을 이용한 핫덱대체 방법의 문제점은 대체군을 형성하는 변수를 어떻게 설정해야 하는가와 기증자를 찾기 위하여 포기해야 하는 대체군 형성 변수의 순서를 결정하는 것이다. 우선 대체군을 형성하는 변수들은 무응답 발생 메커니즘이 무시할 수 있는 메커니즘이 되도록 만들어주는 변수들이 되어야 한다. Collins, Schafer, and Kam (2001)은 대체를 실시할 때 무응답이 발생한 변수와 밀접한 연관성을 가지는 변수들이 포함되어야 편향이 발생하지 않는다는 것을 모의실험을 통해 보였다. 따라서 무응답이 발생한 변수와 연관성을 가지는 것으로 생각되는 변수들을 가능한 한 모두 포함하도록 대체군을 형성해야 한다. 문제는 이와 같이 대체군을 형성한 경우 대체군의 숫자가 기하급수적으로 증가하여 일부 무응답

값에 대한 기증자를 찾을 수 없어 대체군 형성 변수 일부를 포기해야 하는 경우에 발생한다. 연관성 정도가 약한 변수를 포기하는 방법을 고려할 수도 있지만 특정 변수 때문에 기증자를 구할 수 없는 경우도 종종 발생한다. 즉, 특정 변수에서 동일한 항목에 속하는 응답자를 발견할 수 없다면 그 변수를 포기해야 기증자를 발견할 수 있다. 예를 들어 앞의 소득에 대한 대체에서 연령이 동일한 응답자를 발견할 수 없다면 다른 대체 형성 변수를 생략한다고 하여 기증자를 찾아낼 수 있지 않다는 것이다. 어느 변수를 먼저 포기해야 하는지에 대한 통일된 의견이 제시되기 어려운 이유이다. 따라서 대체군 형성은 자료에 대한 풍부한 정보와 경험에 근거하여 조심스럽게 선택되어야 하며 대체군 형성 변수의 포기 순서를 결정하는 것도 이와 마찬가지로 조심스럽게 결정되어야 한다.

예제 4.2 2008년 사회조사의 무응답 대체

예제 4.1에서 다룬 2008년 사회조사의 예를 다시 고려하자. 본 예에서는 가구주의 학력 및 결혼 상태에 근거하여 대체군을 형성하고 그 대체군 내에서 무응답의 대체를 실시하였다. 즉, 자녀 교육비 부담 정도를 측정한 변수에서 무응답이 발생하는 데 연관된 변수들인 학력과 결혼상태 변수들로 대체군을 형성하였다. 학력은 “고졸 이하”와 “대학 이상”으로 나누고 결혼 상태는 “혼인 중”과 “그 외”로 구분하여 전체 대체군의 숫자는 $2 \times 2 = 4$개였다. <표 4.2>은 무응답 발생 전 완전한 자료, 무응답을 무시한 채 분석을 실시한 완전히 응답한 개체를 이용한 분석(complete-case analysis), 그리고 두 변수에 근거한 대체군을 이용한 핫덱대체를 실시한 후 대체된 자녀교육비 부담정도의 항목별 응답 비율을 비교한다. 대체군을 이용한 핫덱대체를 실시한 경우 자녀 교육비 부담 정도에 대한 응답비율은 무응답이 발생하기 전의 완전한 자료의 응답 비율과 비슷하게 나타났다. 반면, 완

전히 응답한 개체를 이용한 분석의 비율은 완전한 자료의 응답 비율과 다르게 나타났다. 대체군을 형성한 변수들은 무응답 자료를 생성하는데 사용된 변수들로서 이 변수들로 대체군을 형성함으로써 무응답 자료 메커니즘이 임의결측이 되어 대체의 정확성을 향상시킬 수 있었다.

<표 4.2> 자녀 교육비 부담 정도에 대한 각 항목의 비율에 대한 대체군을 이용한 핫덱대체, 완전한 자료, 그리고 완전히 응답한 개체를 이용한 분석 결과 비교

응답항목	완전한 자료	대체군을 이용한 핫덱대체[a)]	완전히 응답한 개체를 이용한 분석
매우 부담스럽다 (1)	39.57	38.99	37.11
약간 부담스럽다 (2)	40.17	40.19	41.25
보통이다 (3)	15.76	16.45	17.14
별로 부담스럽지 않다 (4)	3.89	3.76	3.90
전혀 부담스럽지 않다 (5)	0.60	0.60	0.60

a) 대체군을 이용한 핫덱대체를 실시할 때 342개의 관측값은 기증자를 발견하지 못하여 대체가 실시되지 못하였고 이 결과는 8,115명 대신 7,773명에 근거한 결과임

한편, 핫덱대체의 문제점이 기증자를 발견하지 못하는 경우에 발생하는데 이 경우 342개의 무응답값이 기증자를 구하지 못하여 대체를 실시하지 못했다. 기증자를 발견하지 못하는 경우에 대체군을 형성하는 변수들을 포기하여 대체를 실시하는 방법으로 대체를 실시할 수 있다. 즉, 기증자를 찾지 못한 경우 대체군을 형성하는 변수 중 결혼상태 변수를 포기하여 모든 무응답값에 대하여 대체를 실시한 결과가 <표 4.3>에 나타난다. 대체군을 변경하면서 핫덱대체를 실시한 결과 자녀 교육비 부담 정도에 대한 응답비율은 무응답이 발생하기 전의 완전한 자료의 응답 비율과 비슷하게 나타나 이 방법의 유용성을 지지하고 있다.

<표 4.3> 자녀 교육비 부담 정도에 대한 각 항목의 비율에 대한 대체군(필요한 경우 일부 대체 변수 포기 가능)을 이용한 핫덱대체, 완전한 자료, 그리고 완전히 응답한 개체를 이용한 분석 결과 비교

응답항목	완전자료	대체군을 이용한 핫덱대체	완전히 응답한 개체를 이용한 분석
매우 부담스럽다 (1)	39.57	39.10	37.11
약간 부담스럽다 (2)	40.17	41.01	41.25
보통이다 (3)	15.76	16.75	17.14
별로 부담스럽지 않다 (4)	3.89	2.72	3.90
전혀 부담스럽지 않다 (5)	0.60	0.42	0.60

4.1.3 최근접이웃 핫덱대체 방법(Nearest Neighbor Hotdeck)

4.1.2절에서 다룬 대체군을 이용한 핫덱대체를 실시할 때 대체군을 형성하는 변수는 범주형 변수이다. 연속형 변수가 대체군을 형성하는데 포함되기 위해서는 연속형 변수를 범주형 변수로 범주화(categorization)하여야 한다. 연속형 변수는 범주형 변수보다 더 정확한 정보를 포함하고 연속형 범주를 어떻게 범주화하느냐에 따라서 분석의 결과가 달라지는 등 민감한 문제를 포함하고 있다. 대체를 실시할 때 정확도를 높이기 위하여 고려하고자 하는 변수가 연속형이라면 최근접이웃 핫덱대체 방법을 실시할 수 있다. 이 방법은 대체를 실시할 때 고려하고자 하는 변수들의 값에 근거하여 관찰값 사이의 거리(distance metric)를 정의하고 무응답값과 이 거리가 가장 가까운 응답자를 선택하여 이 응답자의 값을 대체에 사용하는 방법을 의미한다. 가장 간단한 예로 소득에 있어서 무응답이 발생한 경우 동일한 연령을 가진 응답자들 중에서 한 명을 임의로 선택하고 그 기증자의 소득을 가지고 대체를 실시하는 것이다. 물론 여기서, 동일한 연령은 정확한 연령이 될 수도

있고 무응답자와 비슷한 연령이 될 수도 있다.

대체에 고려하고자 하는 k개의 변수가 있다고 가정하자. 이 변수들을 $X=(X_1,\cdots,X_k)$라 하면 i번째 관찰값의 X값인 $x_i=(x_{i1},\cdots,x_{ik})'$, $i=1,\cdots,n$,는 i번째 응답자에 대한 k개의 변수의 측정값들로 구성된 공변량 벡터를 표현한다. 이 k개의 변수들을 사용하여 대체를 시행하기 위하여 i번째 응답자와 j번째 응답자간의 거리는

$$d(i,j)=\begin{cases} i\text{번째 응답자와 } j\text{번째 응답자가 동일한 대체군에 속하면 } 0 \\ i\text{번째 응답자와 } j\text{번째 응답자가 동일한 대체군에 속하지 않으면 } 1 \end{cases}$$

와 같이 정의한다. 이 때 거리$d(i,j)$는 여러 가지 방법으로 정해질 수 있는데 흔히 선택되는 방법들은 다음을 포함한다.

(1) 변수값들의 최대 편차(maximum deviation): i번째 응답자와 j번째 응답자의 각 변수별 측정값의 차이 중 최대 차이로서

$$d(i,j)=\max_k|x_{ik}-x_{jk}|$$

로 나타난다.

(2) 마할라노비스 거리(Mahalanobis distance): i번째 응답자와 j번째 응답자 사이의 마할라노비스 거리

$$d(i,j)=(x_i-x_j)'S_{xx}^{-1}(x_i-x_j)$$

를 사용한다. 여기서, S_{xx}는 변수 $X_1,\cdots,X_k$들 간의 추정된 분산공분산행렬(the estimate of the covariance matrix)을 의미한다.

(3) 예측평균 (predictive mean): 무응답을 포함한 변수 Y를 반응변수로, 변수 $X_1,\cdots,X_k$를 설명변수로 고려한 회귀분석에서 Y의 예측값(predictive value)에 근거하여 i번째 응답자와 j번째 응답자간 예측값의 차이의 제곱인

$$d(i,j)=\left[\hat{y}(x_i)-\hat{y}(x_j)\right]^2$$

을 사용한다. 여기서, $\hat{y}(x_i)$는 $X_1,\cdots,X_k$를 설명변수로 고려한 Y의 회귀분석에서 i번째 응답자의 Y의 예측값(predictive value)을 의미한다.

i번째 관찰값의 Y변수의 값이 무응답이라면 응답자인 j번째 관찰값들 중에서 $d(i,j)< d_0$인 응답값들 중에서 한 개의 관측값을 선택하여 그 개체의 Y변수의 값을 가지고 대체한다. 여기서 d_0을 크게 잡으면 잡을수록 더 많은 기증 가능한 응답자 대체군이 형성되게 된다.

최근접이웃 핫덱대체 방법 중 처음 두 방법인 (1) 변수값들의 최대편차를 이용하는 방법과 (2) 마할라노비스 거리를 사용하는 방법은 고려하는 변수들이 범주형인 경우 적용될 수 없다. 하지만 이 방법들은 4.1.2절의 대체군을 사용한 핫덱대체와 혼합하여 사용하는 것이 가능하다. 예를 들어 범주형 변수들을 사용하여 대체군을 형성하고 각 대체군 내에서 연속형 변수들을 공변량 벡터로 사용하여 거리 $d(i,j)$를 정의하고 이 거리에 근거하여 i번째 무응답에 대하여 최근접이웃인 j번째 응

답자의 값을 가지고 대체를 실시할 수 있다. 이렇게 대체된 값은 무응답인 개체와 범주형 변수들에 대하여 동일한 값을 가지면서 연속형 변수들 간의 거리도 최근접인 응답값이므로 대체의 정확성을 추구할 수 있다. 한편, 예측평균을 이용한 핫덱대체 방법의 경우 범주형 변수를 지시변수(indicator variable)로 만들어 설명변수에 포함시키면 범주형 변수 및 연속형 변수를 모두 고려하여 대체를 실시할 수 있다.

핫덱대체는 명시적 형태의 모형(explicit model)을 정의하지 않고 대체를 실시한다는 의미로 내재적 모형(implicit model)하에서의 대체 방법이라 할 수 있다. 내재적 모형에는 가정도 명시적이지 않으므로 추정값의 편향을 수리적으로 계산하기 어렵다. 더구나 단순임의 핫덱대체보다는 대체군에 근거한 핫덱대체나 최근접이웃 핫덱대체, 또는 혼합 등 고려된 변수들의 복잡한 함수들(complex functions)에 근거하여 무응답에 대한 대체가 실시되므로 대체된 자료에 근거한 추정량의 성질을 평가하기 힘든 어려움이 존재한다. 따라서 핫덱대체의 성능(performance)은 대부분 비슷한 상황 하에서의 모의실험을 통해 평가되는데 상당수 모의실험에서 신중하게 선택된 방법에 근거한 핫덱대체는 정확성 높은 대체를 가능하게 하고 대체된 자료에 근거한 추정량에서 편향이 발생하지 않는다고 보고하고 있다.

4.2 혼합적 모형에 근거한 대체 방법

3절에서 다룬 명시적 모형에 근거한 대체 방법들은 자료가 모형의 가정을 만족시키는 경우 우수한 성능을 보이는 것으로 나타나고 있다. 자료가 모형의 가정을 만족시키지 못하는 경우에도 이 모형들은 많은 경우 강건한(robust) 것으로 나타나는데 그 이유는 대체가 전체 자료에 대하여 이루어지는 것이 아니라 무응답값에

대하여만 발생하므로 무응답의 비율이 높지 않다면 잘못된 대체로 인한 추정량의 영향이 크지 않아 모수의 편향은 심각하게 발생하지 않다는 것이다. 하지만 모형의 분포 가정이 심하게 위배되는 경우 무응답의 비율이 높지 않아도 편향이 발생할 수 있다고 보고되고 있다 (Tang et. al., 2005).

핫덱대체 방법은 자료에 대한 분포 가정을 포함하지 않으므로 여러 가지 다른 형태의 변수에 유연성있게 적용할 수 있고 대체군을 잘 형성한다면 상당히 정확한 대체를 실시할 수 있다. 하지만 대체군을 이용한 대체의 경우 대체군을 형성하는 변수들의 숫자가 늘어남에 따라 대체군의 숫자가 기하급수적으로 늘어나 기증자를 찾을 수 없는 무응답값들이 생기게 되고 이 무응답값들에 대하여 대체 변수의 숫자를 줄여 대체를 실시하면 무응답값들마다 대체 기준이 달라지는 문제점을 가진다. 더구나 대체군 선정 및 조정 등 많은 노력이 필요한 기법이다.

모수적 모형에 근거한 대체 방법과 핫덱대체 방법들의 장점을 유지하면서 위에서 언급한 단점을 보완하기 위한 모형들이 제시되어 왔다. 본 절에서는 핫덱대체에 모수적 모형의 기법을 접목시킨 대체 방법과 모수적 모형의 예측력을 높이기 위한 비선형 회귀모형(nonlinear model)에 근거한 대체방법에 관하여 논의한다.

4.2.1 예측평균값(predictive mean value)에 근거한 핫덱대체 방법

무응답 대체를 위하여 무응답이 발생한 변수의 값과 비슷한 응답값을 갖는 개체들을 찾아내서 무응답을 대체하는 것이 도움이 되는데 Little(1988b)은 예측평균에 근거한 짝짓기 방법(predictive mean matching method)을 제안하였고 Bell(1999)은 수정된 예측평균에 근거한 대체 방법을 제안하였다. 이 방법의 아이

디어는 예측평균의 값(predictive mean)을 계산한 후 무응답값의 예측평균과 가까운 예측평균을 갖는 응답값들을 짝지어(match) 대체를 실시하는 것에서 비롯되었다. 무응답을 포함한 변수벡터 Y와 이와 연관된 p개의 공변량, $X_1, X_2, \cdots, X_p$가 존재한다고 가정하자. 무응답이 발생한 Y 변수를 반응변수로, 공변량 $X_1, X_2, \cdots, X_p$를 설명변수로 설정하여 회귀분석을 실시한 후 Y 변수의 예측값을 계산한다. 이 예측값에 근거하여 자료를 몇 개의 대체군으로 나누고 각 대체군 내에서 무응답값을 같은 대체군 내의 응답자의 값으로 대체시키는 방법이다. 응답성향점수에 근거한 대체군의 숫자는 Rosenbaum and Rubin(1984)에서 제안한 바와 같이 전체 자료의 숫자에 따라 4개 또는 그 이상으로 결정될 수 있다. 이 방법은 예측평균을 구하기 위하여 모수적 모형을 적합시키지만 이 모형은 핫덱대체를 실시하기 위한 대체군의 형성에만 사용되므로 모형의 오지정(misspecification)에 덜 영향을 받는 장점을 지닌다.

이 방법은 공변량 $X_1, X_2, \cdots, X_p$에도 무응답이 발생하는 경우에도 사용이 가능하도록 확장이 가능하다. 즉, 공변량 $X_1, X_2, \cdots, X_p$에 무응답이 발생하는 경우 응답된 변수들만에 근거하여 회귀모형을 적합하고 모형의 예측값에 근거하여 대체군을 형성할 수 있다.

예제 4.3 2008년 사회조사의 무응답 대체

예제 4.1과 4.2에서 다룬 2008년 사회조사의 예를 다시 고려하자. 본 예에서는 가구주의 학력 및 결혼 상태에 근거하여 자녀 교육비 부담 정도에 대한 예측값을 계산하고 이 예측값에 근거하여 형성된 대체군 내에서 무응답에 대한 대체를 실시하였다. 즉, 학력 및 결혼상태를 설명변수로 포함하여 자녀 교육비 부담정도에

대한 회귀분석을 실시하였고 예측값에 근거하여 대체군을 형성한 뒤 무응답에 대한 대체를 실시하였다. <표 4.5>은 무응답 발생 전 완전한 자료, 무응답을 무시한 채 분석을 실시한 완전히 응답한 개체를 이용한 분석(complete-case analysis), 그리고 두 변수를 설명변수로 포함한 예측평균값에 근거한 대체군에서의 핫덱대체 결과를 비교한다. 예측평균값에 근거한 핫덱대체를 실시한 경우 자녀 교육비 부담 정도에 대한 응답비율은 무응답이 발생하기 전의 완전한 자료의 응답 비율과 비슷하게 나타나 이 모형이 대체군에 근거한 대체 방법과 비슷하게 좋은 결과를 보인다는 것을 알 수 있다.

<표 4.4> 예측값에 근거한 핫덱대체 후 자녀 교육비 부담 정도에 대한 응답 비율을 완전한 자료 및 완전히 응답한 개체를 이용한 분석 방법과 비교

응답항목	완전자료	대체군을 이용한 핫덱대체	완전히 응답한 개체를 이용한 분석
매우 부담스럽다 (1)	39.57	39.72	37.11
약간 부담스럽다 (2)	40.17	39.61	41.25
보통이다 (3)	15.76	16.46	17.14
별로 부담스럽지 않다 (4)	3.89	3.68	3.90
전혀 부담스럽지 않다 (5)	0.60	0.53	0.60

4.2.2 비선형 회귀모형에 근거한 대체 방법

3절에서 고려한 명시적 모형인 모수적 선형모형(parametric linear model)을 가정하는 대체법의 단점을 보완하는 비선형 모형(nonlinear model) 하에서의 대체법을 고려할 수 있다. 무응답을 포함한 변수 Y와 설명 변수 행렬 X의 비선형적 관계(nonlinear relationship)는 X의 이차식(quadratic equation)이나 삼차식(cubic

equation) 등 다항식(polynomial equation)을 통해 회귀모형에서 고려할 수도 있지만 좀 더 강건한(robust) 방법으로 비모수적 회귀식을 고려할 수도 있다. 이 방법은 회귀식 $Y=g(X)+\epsilon$에서 함수 $g(\bullet)$를 스플라인(spline)이나 커넬(kernel)을 이용하여 적합하는 방법이다(Cheng, 1994; Little and An, 2004). 이 방법 또한 만일 공변수의 개수가 늘어나면, 즉 공변수의 차원이 늘어나면 자료의 수가 무한대로 필요한 "차원의 저주(curse of dimensionality)"의 문제가 생기게 된다(Bellman, 1957). 이러한 다차원의 문제를 해결하기 위한 대체모형으로 일반가법모형(generalized additive model)을 대체모형으로 이용할 수 있다(Hastie and Tibshirani, 1990). 무응답을 포함한 변수 Y에 대하여 p개의 결측이 없는 공변수 $X_1, X_2, \cdots, X_p$가 있다면

$$Y=g_1(X_1)+g_2(X_2)+\cdots+g_p(X_p)+\epsilon,\ \ \epsilon \sim N(0,\sigma^2)$$

의 회귀식을 고려하고 함수 $g_i(\bullet),\ i=1,2,\cdots,p$를 스플라인(spline)이나 커넬(kernel)을 이용하여 완전히 관측된 자료를 사용하여 예측값을 추정하는 방법이다.

다차원의 문제를 해결하기 위한 다른 방법으로 Little and An(2004)은 응답성향(response propensity)을 이용한 차원의 축소를 제안하고 이를 벌점스플라인 대체(penalized spline imputation) 방법이라고 하였다. 자료에 결측이 없는 p개의 변수, $X_1, X_2, \cdots, X_p$가 존재하고 1개의 변수 Y에서 결측이 발생한다고 하면 결측자료 메커니즘이 임의결측일 때 응답성향은 $P(R=1|X_1,...,X_p)=X_1^*$로 정의되고 이 응답성향이 주어진 경우 Y_{obs}와 Y_{mis}의 분포는 같다. 즉, $f(Y_{obs}|X_1^*)=f(Y_{mis}|X_1^*)=f(Y|X_1^*)$이다. 그러므로 Y를 $X_1, X_2, \cdots, X_p$로 회귀하는 대신 Y를 X_1^*로 회귀하여도 편향이 없는 추정치를 얻을 수 있게 된다. 즉 p차원을 1차원으로 축소한 후 비모수적인 회귀식을 적합하는 방법이다. 이 방법은

강건한 모형을 바탕으로 대체값을 예측하기 때문에 변수 사이의 관계가 선형이 아니더라도 또는 변수의 차원이 높더라도 응답성향의 모형이 올바르게 선택되고 결측 메커니즘이 무시할 수 있는 메커니즘이라면 편향이 없는 추정치를 구할 수 있다. 이 방법은 현재까지는 상용화된 프로그램에는 포함되어 있지 않다.

4.3 다중대체

3장 및 4.2절까지는 각각의 무응답에 한 개의 그럴듯한 값을 대체하는 방법을 고려하였다. 이와 같이 한 개의 무응답에 한 개의 값을 대체하는 방법을 2.4절에서 소개한 바와 같이 단일대체(single imputation)라고 한다. 이 방법의 장점은 대체된 자료는 더 이상 무응답을 포함하고 있지 않으므로 연구자가 원하는 분석을 마음대로 시행할 수 있다는 점이다. 하지만 단일대체를 시행하면 대체된 자료값들 중 어느 값이 응답값이고 어느 값이 대체된 값인지 구별할 수 없어 추정량의 분산이 과소추정되는 결과를 가져온다. 응답값과 대체된 값을 구별해야 하는 이유는 응답값은 참값에 대하여 정확하게 측정한 값[3]이지만 대체된 값은 원래의 응답값과 동일하지 않을 가능성이 높기 때문이다. 즉, 상당히 정확도가 높은 대체모형을 사용하여 대체를 실시하더라도 대체된 값 모두가 무응답이 발생한 원래의 값과 동일할 가능성은 희박하기 때문이다. 따라서 모수에 대한 추론은 응답된 자료인 Y_{obs}만의 정보(information)에 근거하여 실시되어야 하는데 단일대체를 실시한 후 대체된 자료는 응답된 자료 뿐 아니라 대체된 자료로부터의 정보의 양도 추가되어 정보의 양을 과다추정하고 이는 추정량의 분산이 작게 추정되도록 하여 추론을 실시할 때 기각하지 않아야 하는 모수를 기각으로 이끄는 문제점을 야기할 수

3) 응답값이 참값을 정확하게 측정한다는 것은 측정오차가 발생하지 않았다고 가정하는 것이다. 실제 자료에서는 측정오차가 발생할 수도 있지만 이 경우 문제가 복잡해지므로 기본적인 대체모형은 측정오차가 발생하지 않았다는 가정 하에 시행된다.

있다는 점에서 주의가 필요하다.

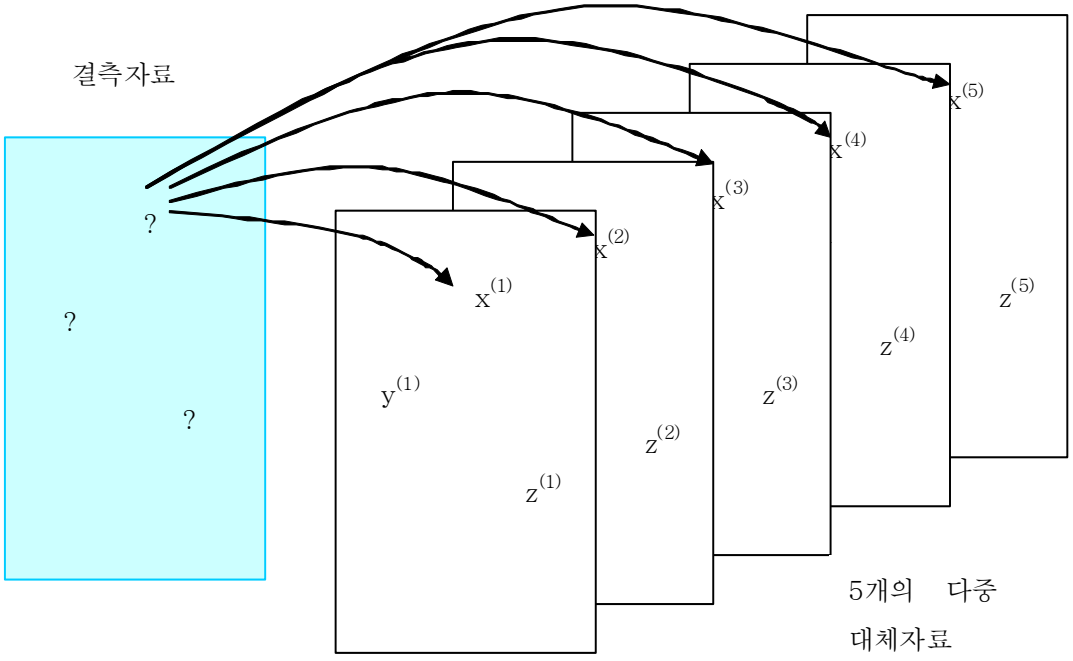
다중대체는 단일대체의 문제점을 해결하기 위하여 한 개의 무응답을 두 개 이상의 값을 가지고 대체를 실시하는 방법을 통틀어 칭한다. 다중대체에서는 한 개의 무응답을 그럴듯한(plausible) 여러 개의 값으로 대체하고 여러 개의 대체된 값들은 무응답이 발생한 값에 대한 불확실성(uncertainty)을 반영하게 된다. 즉, 무응답값을 정확하게 아는 것이 거의 불가능하므로 무응답값을 여러 개의 그럴듯한 값으로 표현함으로써 무응답값에 대한 우리의 불확실성을 모형에 포함시키는 통계적 기법이다. 다중대체를 실시하면 무응답값들이 여러 개의 값들로 대체되고 이 값들간 차이에서 오는 분산이 추정량의 분산을 계산할 때 추가되어 분산이 과소추정되는 문제점을 방지하게 된다. Rubin(1987a)은 한 개의 무응답을 무한개의 값으로 대체한다면 추정량의 분산이 정확하게 추정되는 점을 이론적으로 보였다. 물론 한 개의 무응답을 무한개의 값으로 대체하는 것은 현실적으로 불가능하므로 실제로는 유한개의 대체를 실시하게 된다. 특히 무응답으로 인하여 손실된 모수에 대한 정보량(missing information)이 아주 크지 않다면 작은 숫자의 대체를 통해서도 모수의 분산이 거의 비슷하게 추정된다는 것을 보였다. 여기서, 무응답으로 인하여 손실된 모수에 대한 정보량이란 무응답이 발생하지 않은 완전한 자료(complete data)와 비교하여 무응답이 발생함으로 인해서 발생한 모수의 정밀도(precision)의 감소분(reduction)을 의미한다. 한 개의 모수(scalar parameter)를 고려할 때, 무응답으로 인하여 손실된 모수에 대한 정보량을 γ라고 한다면 무한개의 대체를 실시했을 때와 비교하여 유한한 m개의 대체를 실시할 때 모수의 분산에 관한 상대적 효율(relative efficiency)은 대략적으로 $\left(1+\frac{\lambda}{m}\right)^{-1}$으로 표현할 수 있다. 예를 들면, 손실된 모수에 대한 정보량이 50%라 하더라도 5번의 다중대체를 실시하면 무한개의 대체를 실시한 경우와 비교하여 모수에 대한 추정량의

표준오차는 $\sqrt{1+\frac{0.5}{5}}=1.049$로서 나타나 약 5%만 늘어나 거의 차이가 없다. 손실된 모수의 정보량을 구하는 방법은 4.3.2절에서 좀 더 자세하게 설명되지만 종종 무응답의 비율과 연관되어 생각되어지고 있다.

무응답을 포함한 자료에 대하여 다중대체를 실시하면 결과로서 단일대체와는 달리 한 개의 대체된 자료 대신 여러 개의 대체된 자료가 생성되게 된다. 즉, 무응답을 포함한 자료에 대하여 다중대체를 시행한 후 생성된 대체된 자료는 다음의 <그림 4.3>에 나타난 것과 같이 여러 개가 존재하게 된다. 이 여러 개의 대체된 자료들은 관찰된 값들은 모두 동일하지만 대체된 무응답값은 같기도 하고 다르기도 한 형태를 지닌다.

다중대체에서 대체의 숫자가 m이면 연구자는 m개의 대체된 자료 각각에 대하여 원하는 분석을 반복적으로 시행할 수 있다. 각 자료에 대하여 독립적으로 동일한 분석이 시행된 후 분석 결과는 유사하지만 동일하지 않은 m개의 통계량 및 관련 분산(또는 표준 오차)으로 나타나는데 연구자는 m개의 각각 다른 통계량이 아닌 하나의 통합된 통계량을 구하거나 통합된 추론을 실시하는 데 목적이 있다. 이 목적은 (1) 다중대체된 각 자료의 분석, 그리고 (2) 분석된 자료를 통합한 결과 도출의 두 단계를 통하여 달성된다.

<그림 4.3> 무응답을 포함한 결측 자료에 대하여 5개의 다중대체를 실시한 경우의 예



4.3.1 다중대체(multiple imputation)된 자료의 분석

다중대체된 자료 각각은 무응답이 대체되어 무응답이 없는 완전한 형태의 자료를 가지게 되므로 자료 각각에 대하여 연구목적에 알맞은 분석을 시행하면 된다. 예를 들어, 회귀분석을 시행하고자 한다면 관심있는 반응변수에 대하여 관심있는 설명변수를 포함하여 m개의 대체된 자료 각각에 대하여 회귀분석을 실시하면 된다. 이렇게 m번의 분석을 실시하면 추정된 회귀계수(regression coefficients), 표준오차(standard errors), 그리고 검정통계량(test statistics)은 m개 자료 각각으로부터 약간씩 다르게 나타나는데 이는 관심 변수가 완전하게 측정되지 못하고 무응답을 포함하므로 무응답값에 대한 불확실성에 근거한 차이를 나타내는 것이다.

하지만 연구자의 분석 목적은 관심 자료에 대한 m개의 서로 다른 결론이 아니라 한 개의 통합된 결론을 내리는 것이므로 m개 분석의 결과를 통합하여 한 개의 결론을 도출하기 위하여 다음의 통합 과정을 거쳐야 한다.

4.3.2 다중대체 자료를 분석한 후 결과의 통합

다중대체를 m번 시행한 후 대체된 자료 각각에 대하여 알맞은 분석을 시행한 후 얻어진 모수의 추정값들을 $\widehat{\theta_1},\widehat{\theta_2},\cdots,\widehat{\theta_m}$ 이라 정의하고 이 모수들의 추정된 분산을 각각 $W_1, W_2,\cdots, W_m$이라 정의하자. 예를 들어, 회귀분석을 실시한 경우 i번째 대체자료에 근거한 회귀 분석에서 관심 설명 변수의 회귀계수의 추정값 벡터는 $\widehat{\theta_i}$이 되고 그 회귀계수 벡터의 표준오차의 추정값의 제곱이 W_i 행렬로 표현된다. 이 경우 통합된 모수의 추정값은

$$\bar{\theta} = \frac{1}{m}\sum_{i=1}^{m}\widehat{\theta_i}$$

으로 표현될 수 있다. 즉, 모수 추정값들의 평균값이 통합된 모수의 추정값이 된다.

통합된 모수의 분산의 추정값은 다음의 두 개의 분산 성분의 합으로 표현된다. 첫 번째 분산 성분은

$$\overline{W_m} = \frac{1}{m}\sum_{i=1}^{m} W_i$$

로서 대체된 자료들로부터 추정된 m개의 모수의 분산들의 평균이다. 이 분산 성분은 대체내분산(within-imputation variance)이라 부른다. 두 번째 분산 성분은

$$B_m = \frac{1}{m-1}\sum_{i=1}^{m}\left(\hat{\theta}_i - \overline{\theta}\right)^2$$

으로 표현되는데 이는 m개의 대체된 자료로부터의 m개의 모수의 추정값들 사이의 분산을 나타내므로 대체간분산(between-imputation variance)이라 부른다. 이 두 개의 분산 성분을 종합하여 계산한 모수의 분산에 대한 통합된 추정값은

$$T_m = \overline{W_m} + \frac{m+1}{m}B_m$$

으로 구할 수 있다.

관측값의 개수가 충분히 큰 경우, 일변량 모수(scalar parameter)에 대한 분포는 다음의 t-분포를 따른다.

$$(\theta - \overline{\theta})\,T_m^{-\frac{1}{2}} \sim t_v$$

여기서, t-분포의 자유도 v는 $\nu = \left(\nu_0^{-1} + \widehat{\nu_{obs}^{-1}}\right)^{-1}$로 계산되는데 이 중 ν_0^{-1}는 $v_0 = (m-1)\left(1 + \frac{1}{m+1}\frac{\overline{W_m}}{B_m}\right)^2$으로 계산되고 $\widehat{\nu_{obs}} = \left(1 - \widehat{\gamma_m}\right)\left(\frac{\nu_{com}+1}{\nu_{com}+3}\right)\nu_{com}$으로 나타나는데 이 때 v_{com}은 결측값이 없을 때 모수 θ에 대한 추정의 자유도(degree of freedom)를 나타내고 $\widehat{\gamma_m} = \left(1 + \frac{1}{m}\right)\frac{B_m}{T_m}$으로서 무응답으로 인하여 손실된 모수

θ에 대한 정보량(fraction of missing information about θ due to nonresponse)의 추정값이다. 모수의 분포가 t-분포를 따르므로 모수 θ에 관한 t-검정을 시행하거나 모수의 신뢰구간(confidence interval)을 구할 수 있다. 이 통합 방법은 관심 모수들이 벡터로 표현될 때 다변량 검정 및 신뢰구간의 계산 등으로의 확장도 가능하다 (Rubin, 1987a).

예제 4.4 기업활동실태조사의 무응답 대체

다중대체를 시행하여 만들어진 m개의 자료들 각각에 대하여 m개의 분석을 시행하고 그 결과를 통합하는 과정은 여러 가지 통계 프로그램에서 프로시저로 만들어 제공하고 있다. 예를 들어, SAS의 PROC MIANALYZE 프로시저는 위와 같이 분석된 자료의 모수들을 통합한 결과를 제공해 준다. 그 외에 SPSS, S-PLUS와 무료 통계 프로그램인 R도 다중대체된 자료를 분석한 후 통합하는 모듈을 제공하고 있다. 또한 J. L. Schafer가 개발한 Windows에서 독립으로 시행되는 프로그램 NORM은 위의 단계를 수행하고 통합된 결과를 제공하는데 사이즈가 작고 쉽게 설치할 수 있고 이 프로그램은 http://www.stat.psu.edu/~jls/misoftwa.html에서 다운받을 수 있다. 다음은 SAS에서 다중대체된 5개의 대체자료를 이용한 단순 평균의 계산을 위한 단계를 보여준다.

우선 SAS에서는 대체된 자료에 대체자료의 순서를 의미하는 변수를 제공하는데 그 변수의 이름은 _IMPUTATION_이다. 즉, m개의 대체를 실시하면 다중대체된 자료의 관찰단위는 $m \times n$개로 원래 관찰단위 n의 m배가 되는데 이는 m개의 다중대체된 자료를 의미한다. 즉, 각 관찰단위가 m개 중복되어 나타나는데 이 m개의 중복은 관찰된 변수에 관해서는 동일하지만 대체된 변수값들은 서로 다르다.

이 m개의 대체 자료를 구분해 주는 변수가 _IMPUTATION_이 된다. <그림 4.4>는 기업활동실태조사 자료를 다변량 정규분포를 가정하여 다중대체한 후 로그 변환한 매출액(log_C24)의 단순 평균의 추정량을 계산하기 위한 SAS 프로그램을 보여준다.

<그림 4.4> 다중대체된 매출액(log_C24)의 단순 평균을 추정하기 위한 SAS 프로그램

```
* 각 대체 자료별 단순 평균 계산;
PROC SURVEYMEANS DATA = total;
   VAR log_C24;
   BY _imputation_;
   ODS OUTPUT STATISTICS = stat1;
RUN;

* 각 대체된 자료별로 계산된 단순 평균을 통합하여 원 자료의 단순 평균
  추정;
PROC MIANALYZE DATA = stat1;
   MODELEFFECTS mean;
   STDERR stderr;
RUN;
```

SAS Procedure SURVEYMEANS에서는 대체된 자료 각각에 대한 분석을 시행하기 위하여 BY변수를 사용하였고 단순평균의 추정값 및 표준 오차가 SAS 데이터 셋인 stat1에 저장되었다. 이 저장된 통계량들은 Procedure MIANALYZE를 사용하여 통합되었는데 이 때 MODELEFFECTS 문에는 통합할 통계량 $\hat{\theta}_i$(여기서는, 평균의 추정값)을 나타내는 변수 mean을 써 주고 STDERR문에는 W_i 의 제곱근인 평균의 표준 오차를 나타내는 stderr 변수를 써 주면 된다. Procedure MIANLYZE를 실시한 결과는 <그림 4.5>에 나타난다. 매출량의 단순 평균은 10.15로 추정되고 표준편차는 0.014이다. 매출량의 평균에 대한 95% 신뢰구간은 (10.13, 10.18)

로 계산되며 평균이 0이라는 귀무가설은 t-통계량이 723.86, p-value가 <.0001로 5% 유의수준 하에서 유의하게 나타난다.

<그림 4.5> 다중대체된 매출액(log_C24)의 단순 평균 추정량의 계산을 위한 SAS 결과

The MIANALYZE Procedure

Model Information

Data Set	WORK.STAT1
Number of Imputations	5

Multiple Imputation Variance Information

Parameter	Variance Between	Variance Within	Variance Total	DF
mean	0.000008899	0.000186	0.000197	1357.4

Multiple Imputation Variance Information

Parameter	Relative Increase in Variance	Fraction Missing Information	Relative Efficiency
mean	0.057400	0.055675	0.988988

Multiple Imputation Parameter Estimates

Parameter	Estimate	Std Error	95% Confidence Limits		DF
mean	10.152675	0.014026	10.12516	10.18019	1357.4

Multiple Imputation Parameter Estimates

Parameter	Minimum	Maximum	ThetaO	t for H0: Parameter=ThetaO	Pr > |t|
mean	10.147922	10.155054	0	723.86	<.0001

< 4장 연습문제 >

1. 다음의 각 항목에 대하여 참, 거짓을 판별하고 판별의 이유를 설명하시오.

   (가) 완전임의 핫덱대체는 무응답 메커니즘이 임의결측인 경우 추정량에 편향이 발생하지 않는다.

(나) 핫덱대체를 실시할 때 정확한 대체를 실시하기 위하여 무응답이 발생한 변수와 연관된 변수를 모두 포함하여 대체군을 구성해야 한다.

(다) 다중대체를 실시해야 하는 이유는 모수의 추정량의 편향이 발생하지 않도록 하기 위함이다.

2. 다음의 자료는 2008년 사회조사 자료의 일부분이다. 주요 관심이 되는 변수는 학교 교육의 효과를 측정하는 변수로서 지식·기술 습득(C33), 인격형성(C34), 국가관 및 사회관 정립(C35), 그리고 생활·직업·취업에의 활용(C36)의 네 가지 문항에 대하여 해당자 38,991명이 응답하였다. 본 예제에서는 생활·직업·취업에의 활용(C36) 항목의 응답자 중 30%에 대하여 무응답을 생성하였다. 4장에서 학습한 대체방법들을 <표 4.5>에 나타난 변수들을 고려하여 대체를 실시하시오. 각 방법으로 대체된 자료들을 항목별 비율을 통해 비교하시오. 이 자료에 대하여 적합한 것으로 생각되는 한 가지 대체 방법을 추천하시오.

<표 4.5> 기업활동실태조사 자료 일부 변수의 명칭 및 설명

변수명	변수 설명
C33	지식·기술 습득
C34	인격형성
C35	국가관 및 사회관 정립
C36	생활·직업·취업에의 활용

3. 예제 4.4의 자료 중 기업활동실태조사 자료를 다변량 정규분포를 가정하여 다중대체한 후 로그 변환한 유형자산 당기 취득액(log_C18)의 단순 평균의 추정량을 계산하시오. 이 결과를 단일대체의 결과와 비교하시오.

제 5장 무응답이 있는 경시적 자료 또는 패널자료 분석방법

<학습목표>

(1) 경시적 연구 또는 패널조사에서 무응답의 특성을 알아본다.

(2) 무응답이 있는 경시적 자료에 대한 가중법을 소개한다.

(3) 여러 가지 실제 패널조사 예제를 통하여 분석법을 알아본다.

5.1 개요

조사연구는 한 시점에서 한 번만 행해지는 횡단면적(cross-sectional)으로 이루어지는 경우가 많다. 하지만 횡단면적 조사는 한 시점에서의 조사모집단의 특성만을 파악할 수 있을 뿐, 동적인 측면에서 개인의 패턴이나 변화 등을 파악하는데 한계가 있다. 즉, 횡단면 조사 자료는 age effect와 cohort effect의 분리가 불가능하므로 정책적 시사점을 얻는데 한계를 지닌다. 이러한 횡단면 조사의 단점을 보완하고, 동적인 차원에서 조사 참여자의 장기간에 걸친 변화와 상태 간 이동 과정을 보여주는 방법으로 패널조사가 있다. 패널조사란, 조사대상을 고정시키고 동일한 조사대상자에 대하여 동일 질문을 일정 기간 동안 반복적으로 실시하여 조사하는 방법이다. 이 고정된 조사 대상 전체를 패널이라 하며 종적(longitudinal)자료의 성격과 횡적(cross-sectional)자료의 성격을 동시에 가지고 있다.

일반적인 횡단면 조사를 반복 실시하여 (이 경우 조사 대상자는 조사마다 달라진다.) 모집단의 변화추이를 제한적인 수준에서 파악할 수 있지만, 매번 조사 대상자가 상이하기 때문에 개인의 변동 수준은 파악할 수 없다. 하지만 패널조사는

매 조사마다 동일한 표본을 가지고 지속적으로 추적 관찰함으로써 조사관심 효과나 개인 상태의 동적인 변화를 직접적으로 평가할 수 있으므로 횡단면 자료만 있을 때 없었던 심도 있는 결과의 도출이 가능한 장점을 가지고 있다. 현재 우리나라에서는 패널조사 기법을 활용하여 '한국복지패널조사', '한국노동패널조사', '한국교육고용패널조사', '고령화연구패널조사' 등 다양한 분야에서 패널조사가 이루어지고 있다.

이러한 많은 장점에도 불구하고, 패널조사는 장기간 수행되므로 패널의 중도탈락(drop-out)이 큰 문제점이다. 그러므로 패널조사는 초기표본을 장기적으로 유지할 수 있는 체계적인 패널 관리와 초기 표본 추출 시 모집단을 대표할 수 있는 표본추출 설계가 매우 중요하다.

이 번 장에서는 패널조사에서 결측이 있는 경우의 자료 분석법을 소개한다. 물론 이 장에서 소개하는 방법은 표본조사가 아닌 다른 경시적 연구(예를 들면, 코호트 연구)에서 발생한 결측자료 분석에도 적용이 될 수 있다.

5.2 웨이브 무응답

웨이브 무응답은 패널조사에서 흔하게 일어나는 부분 무응답의 한 종류이다. 표본개체가 하나 이상의 조사 웨이브에 조사 참여를 하지 않는 경우가 매우 흔하다. 어떤 표본 개체는 어느 한 시점의 웨이브에서 중도탈락하여 그 이후의 모든 웨이브에서 무응답이 일어나는 경우가 있고 어떤 표본 개체는 어떤 한 시점의 웨이브에서 무응답을 했으나 이 후의 웨이브는 다시 조사에 참여하는 경우가 있다. 전자를 감소(attrition)라고 하고 후자를 간헐적 중도탈락(intermittent dropout)이라고

한다. 어떤 한 웨이브에서 조사에 적합하지 않은 표본개체는 비록 그 웨이브에서 자료를 제공하지 않더라도 무응답으로 간주하지 않는다. 즉, 무응답은 조사에 적합한 표본이 응답을 하지 않은 경우를 말한다. <표 5.1>은 다섯 개의 웨이브로 구성된 패널 조사에서 웨이브 응답(X)과 무응답(0)의 몇 몇 전형적인 패턴을 보여준다.

<표 5.1> 5개의 웨이브로 구성된 패널조사의 무응답 패턴

패턴	응답 상태	웨이브				
		1	2	3	4	5
1	완전응답	X	X	X	X	X
2	감소	X	X	X	X	0
3		X	X	X	0	0
4		X	X	0	0	0
5		X	0	0	0	0
6	간헐적 중도탈락	X	X	0	X	X
7		X	0	0	X	X
8		X	0	0	0	X

X는 무응답을 0은 응답을 나타낸다.

위의 <표 5.1>에 나타나지 않은 한 가지 무응답 패턴은 모든 웨이브에서 무응답이 일어난 경우이다. 많은 패널조사에서 첫 번째 웨이브에 무응답을 한 표본 개체들은 더 이상 추적조사하지 않는 경우가 흔하다. 대부분의 패널조사는 웨이브 무

응답자를 언제 그만 추적할 것인지에 대한 규정이 있다. 예를 들면, 미국의 패널 조사인 Medicare Current Beneficiary Survey(MCBS)와 U.S. Census Bureau's Survey of Income and Program Participation(SIPP)에서는 한 웨이브의 무응답자는 그 다음 웨이브에 조사를 시도한다. 만일 그 다음 웨이브에서도 무응답을 하는 경우는 이 후의 웨이브에서 더 이상 추적조사를 하지 않는다. 즉, 연속적인 두 웨이브에서 모두 무응답인 경우에는 이 후의 웨이브에서 모두 무응답이 된다. 그러므로 SIPP와 MCBS의 규정에 의하면 <표 5.1>에서 패턴 6은 가능하지만 패턴 7과 8은 일어날 수 없다.

<표 5.1>은 분석을 위하여 가능한 응답자의 형태를 보여준다. 먼저 각 웨이브의 횡단면 분석을 고려하여 보자. 예를 들어, 웨이브 1의 분석은 패턴 1에서 8까지의 모든 조사대상자들을 포함한다. 이 때 초기 무응답자에 대하여 보정된 가중값을 이용할 수 있다. 이 경우 제 2장에서 설명된 횡단면 표본조사의 무응답에 대한 가중방법을 이용할 수 있다. 비슷하게, 웨이브 2의 횡단면 분석은 <표 5.1>의 패턴 1-4와 6에 관련된 조사대상자들을 이용한다. 이 때 패턴 5, 7, 8에 속하는 조사대상자는 무응답자로 간주하여 가중방법 또는 대체방법을 고려한다. 또한 웨이브 2의 무응답 보정에는 웨이브 2에서의 무응답이나 웨이브 1에서 응답한 조사대상자의 정보를 이용할 수 있다. 이제 웨이브 5의 분석을 고려하여 보자. 이 분석에는 패턴 1, 6, 7, 8에 속하는 조사대상자들을 이용할 수 있다. 하지만 웨이브 5에서의 무응답자들은 무응답 패턴에 따라 이전 웨이브로부터 얻을 수 있는 정보의 양이 다양하다. 이러한 점이 분석을 좀 더 복잡하게 만든다.

이제 패널자료의 경시적 분석을 고려하여 보자. 웨이브 1과 웨이브 5의 분석은 패턴 1, 6, 7, 8에 속하는 응답자를 이용할 수 있다. 웨이브 2와 웨이브 5의 분석은 패턴 1과 6의 자료를 사용할 수 있다. 이 경우 무응답 보정은 모든 다른 패턴

을 위해서는 웨이브 1의 자료를 이용하고 패턴 2, 3, 4를 위해서는 웨이브 2의 자료를 이용할 수 있다. 마지막으로 5개의 모든 웨이브를 고려한 경시적 자료분석에서는 단지 패턴 1에서만 모든 조사대상자가 이용 가능하다. 이 경우 다른 패턴은 모두 무응답 보정이 필요하다. 이 때 많이 사용되는 분석방법은 선형혼합모형 (linear mixed model)이다. 선형혼합모형은 무응답이 MAR인 경우 모든 이용가능한 자료를 사용하여 우도방법으로 무응답 및 개체 내 상관을 보정한 후 유효한 결과를 도출한다.

위에서 본 바와 같이 무응답의 패턴에 따라 각 각의 분석은 서로 다른 가중값의 집합이 필요하다. t개의 웨이브가 있는 패널조사에서 2^t-1개의 웨이브들의 조합이 잠재적인 관심 분석이 될 수 있다. 이 조합의 숫자는 웨이브의 수가 증가할수록 급격하게 증가한다. 모든 가능한 분석에 대하여 서로 다른 가중값의 집합을 고려하는 것은 현실적으로 어려우며 같은 자료를 가지고 서로 다른 가중값 집합을 사용하는 것도 좋은 방법은 아니다.

이 문제에 대한 한 가지 대안으로는 처음으로 무응답이 발생한 웨이브 이후에 응답된 웨이브를 무응답으로 간주하여 간헐적 중도탈락을 모두 감소패턴으로 바꾸는 방법이 있다. 예를 들면 <표 5.1>의 패턴 6과 7에서 웨이브 4와 5의 자료를 무시하고 패턴 8에서는 웨이브 5의 자료를 무시한다. 이 방법은 잠재적인 가중값 집합의 수를 2^t-1에서 t로 줄여준다. 또한 분석방법도 훨씬 단순해진다. 하지만 패널조사의 웨이브의 수가 많을수록 버려야 할 자료의 양이 매우 클 수 있다. 다른 대안으로는 모형을 바탕으로 한 대체방법이 있다.

5.3 감소(attrition) 패널자료에서 무응답 보정방법

감소 패널자료는 무응답이 단조패턴(monotone pattern)이므로 무응답 자료의 보정은 일반적으로 좀 더 쉽다. 현 웨이브에서의 무응답자는 이전 웨이브에서 중도탈락한 무응답자와 더불어 추가적인 무응답자들로 구성된다. 가중을 이용한 무응답 보정방법은 각 웨이브별로 가중값을 구하게 되는데 현재 웨이브의 가중값은 이전 웨이브의 가중값을 이용하여 갱신한다. 웨이브 t에서의 종합적인 가중값은 다음과 같이 이전 웨이브의 가중값과 현재 웨이브의 가중값의 곱으로 표현할 수 있다.

$$w = w_i \times r_{i1}^{-1} \times r_{i2}^{-1} \times \cdots \times r_{it}^{-1}$$

여기서 w_i는 개체 i의 기저 가중값이고 r_{it}는 웨이브 t에서 개체 i가 속한 클래스 안의 응답비율이다. 만일 $t \geq 2$이면 r_{it}는 웨이브 $t-1$에서 응답한 조사대상자들 가운데 웨이브 t에서 응답한 "조건부" 응답비율이다. r_{it}는 다음과 같이 계산된다.

$$f_{it} = \frac{\sum_{j \in c_i(t)} I_{jt} w_{j(t-1)}}{\sum_{j \in c_i(t)} w_{j(t-1)}}$$

여기서 $c_i(t)$는 웨이브 t에서 개체 i를 포함하는 클래스이고 I_{jt}는 웨이브 t에서 응답을 나타내는 지시변수이다. 즉, 웨이브 t에서 개체 j가 응답한 경우에 $I_{jt} = 1$이고 그렇지 않은 경우에 $I_{jt} = 0$이다. 그리고 $w_{j(t-1)}$는 개체 j의 웨이브 $t-1$까지의 감소에 관한 가중값이다. 만일 웨이브와 웨이브 사이의 감소가 매우 작은 경우에는 이 r_{it}값이 1에 가깝다.

편향을 감소시키기 위한 효율적인 무응답 보정방법을 개발하기 위하여 각 클래스 내의 표본개체들이 비슷한 응답성향을 가질 수 있도록 클래스를 정하는 것이 매우 중요하다. 첫 웨이브에서는 무응답에 관한 정보가 한정적이나 후반 웨이브에서는 이전 웨이브로부터 얻은 무응답에 관한 정보가 더 많다. 클래스를 정하기 위한 여러 가지 통계적 방법 중에 로지스틱 회귀분석 방법과 분류나무(classification tree)에 근거한 방법이 있다. 로지스틱 회귀분석의 틀에서는 응답성향 ϕ_{it}을 다음과 같은 식으로 나타낸다.

$$\log\left(\frac{\phi_{it}}{1-\phi_{it}}\right)= x_{it}\beta_t$$

여기서 x_{it}는 표본 개체 i의 특성변수들의 벡터이고 β_t는 그에 따른 회귀계수의 벡터이다. 여기서 β_t와 ϕ_{it}는 I_{it}를 종속변수로 x_{it}를 독립변수로 $w_{i(t-1)}$을 가중값으로 하여 가중 로지스틱 회귀를 적합하여 추정한다.

위에서 말한 바와 같이 후반 웨이브의 무응답을 위한 클래스를 정하는 모형에서는 잠재적인 예측변수의 수가 매우 크게 되는 경향이 있다. 이런 경우에는 응답변수와 관련이 높은 예측변수만을 골라 예측변수의 차원을 줄이기 위하여 로지스틱 회귀분석에서 후진제거(backward elimination), 전진선택(forward selection), 단계선택(stepwise selection) 등의 변수선택 방법을 사용할 수 있다(Rizzo, Kalton, and Brick, 1996). 예측변수의 수가 많은 경우 로지스틱 회귀모형으로 구해진 클래스를 전부 이용하고 각 클래스의 응답성향을 구하게 되면 매우 작은 응답성향을 가진 개체들의 가중값은 매우 커지게 되는 문제점을 야기한다. Rizzo 등(1996)은 이런 문제점의 해답으로 비슷한 응답성향점수를 가진 표본 개체들을 서로 합침으로써 로지스틱 회귀모형으로부터 마지막 클래스를 정하는 방법을 논의하였다.

클래스를 정하기 위한 방법으로 CHi-squared Automatic Interaction Detector (CHAID, Magidson 1993) 또는 Classification And Regression Tree (CART, Breiman et al. 1993)와 같은 분류나무에 근거한 방법을 고려할 수 있다. 이러한 방법들은 중요한 상호작용들을 모형에서 고려하고 클래스를 정하는데 반영할 수 있다. 예를 들면, 전국에서는 남자와 여자의 응답률이 유의하게 다르지 않은데 경상도에서는 남자와 여자 사이의 응답률이 다른 경우에 로지스틱 회귀분석에서 변수선택을 이용하면 성별이 클래스를 정하기 위한 예측변수에서 제외될 수 있다. 하지만 분류나무 방법은 성별로 경상도 표본을 골라낼 수 있다. 이런 분류나무 방법의 단점 중 하나는 이 방법이 강건(robust)하지 않은 경향이 있다는 것이다. 예를 들면 자료에서 작은 변화가 선택된 나무에서는 큰 변화를 일으킬 수도 있다. 이외에도 클래스를 정하기 위하여 순위를 이용한 방법 (Kalton and Flores-Cervantes, 2003)등이 있다.

이전에 언급한 바와 같이 패널조사에서 웨이브에 따른 무응답의 패턴이 단조(monotone)가 아니고 일반패턴인 경우는 가중을 이용한 무응답 보정방법은 매우 제한적이다. 이런 경우는 대체(imputation)방법이 대안이 될 수 있다. 물론 단조 패턴의 무응답 패널자료도 대체를 이용하여 분석할 수 있다.

예를 들어, SIPP 패널조사는 전체 웨이브에서 발생한 무응답을 대체하기 위하여 경시적 대체방법을 이용하였다. 경시적 대체방법이란 한 시점의 웨이브의 무응답을 대체하기 위하여 현 웨이브 이전과 이후의 웨이브에서 응답한 자료들을 이용하는 방법이다. SIPP 패널조사에서 1991년, 1992년, 1993년 웨이브에 이 방법을 적용하였을 때 5~8% 정도의 추가적인 패널표본을 유지하는 결과를 보였다. SIPP 패널조사에서 1996년 웨이브부터는 두 개의 연속적인 웨이브에서 무응답을

보이는 표본개체를 대체하기 위하여 이전의 경시적 대체방법을 확장하였다. SIPP는 4개월간의 회상기간(recall-period)을 이용한다. 이 기간 동안 매달 인터뷰가 행해진다. 각 웨이브에서 사용된 경시적 대체방법은 다음과 같다. 웨이브 t의 무응답을 대체하기 위해 한 항목에서 웨이브$(t-1)$의 마지막 달과 웨이브$(t+1)$의 첫 번째 달의 값이 같으면 웨이브 t의 그 항목의 모든 달의 무응답은 그 값으로 대체된다. 만일 값이 다른 경우에는 달을 바꾸기 위해 확률화 과정을 이용한다. 무응답 웨이브 t의 4달 가운데 한 달이 각 가정에서 동일 확률로 선택된다. 웨이브 t에서 선택된 달까지의 무응답은 웨이브 $(t-1)$의 마지막 달의 값으로 대체되고 선택된 달 이후의 무응답은 웨이브 $(t+1)$의 첫 번째 달의 값으로 대체된다. SIPP Quality Profile (U.S. Bureau of the Census, 1998: http://www.census.gov/sipp/effects.html)은 이러한 경시적 대체방법에 관한 연구결과를 기술하고 있다.

MCBS 패널조사에서도 SIPP와 비슷한 대체방법을 이용하여 무응답 웨이브를 대체하고 있다. MCBS 패널조사에서는 매년 연간 비용과 지출에 관한 통계가 작성된다. 각 해는 총 세 개의 웨이브로 구성된다. 만일 조사참여자가 세 개의 모든 웨이브에서 비용과 지출에 관하여 응답하지 않으면, 이용가능한 보고된 자료로부터 항목별 평균을 구하여 무응답을 대체한다. MCBS 패널조사에서 사용된 대체방법에 관한 추가적인 정보는 http://www.cms.hhs.gov 에서 얻을 수 있다.

예제 5.3.1 National Educational Longitudinal Survey (NELS -88)

NELS는 1988년에 미국의 8학년(중학교 2학년) 학생을 모집단으로 multi-stage 확률추출법을 이용하여 표본추출한 후 매 2년마다 추적 조사한 패널조사이다.

1998년 원년의 표본설계는 사립학교와 히스패닉과 아시안계 학생의 등록비율이 평균보다 높은 학교들을 과추출(oversample)하였다. 다른 패널조사와 마찬가지로 NELS-88도 감소에 의한 무응답이 매우 많았다. 첫 번째 베이스라인 웨이브와 이후 첫 추적 웨이브 (즉, 8학년과 10학년 사이)에서 특히 감소가 많았다. 이 때 중도탈락이 특히 많았던 이유는 조사비용의 제약 때문이었다. 설계에 의해서 무응답이 발생했으므로 다른 연구와는 달리 무응답이 조사에서 측정된 값에 관련은 작을 것이고 그래서 무응답에 의한 편향도 작을 것으로 여겨졌다. 그럼에도 불구하고, 이 패널연구에 관련되어 출판된 논문들(Lee and Smith 1995, Kao and Tienda 1998, Rojewski and Yang 1997)에서는 각 웨이브에 대해 경시적 가중방법이 사용되었다. 또한 Baltagi(1998)와 Verbeek와 Nijman(1992) 등은 무응답을 고려한 경시적 모형을 이용하여 자료를 분석하였다. 이들은 혼합선형 회귀모형에서 다음과 같은 세 개의 단순한 보조변수의 사용을 제안하였다: (1) 패널에서 개체 i가 참여한 웨이브의 수, (2) 개체 i가 전체 패널조사 기간 동안 모두 응답하였으면 1이고 그렇지 않으면 0인 이분형 변수, (3) 개체 i가 조사의 마지막 웨이브에서 응답하였으면 1이고 그렇지 않으면 0인 이분형 변수.
만일 이들 보조변수가 회귀모형에서 유의하면 분석에서 선택편향(selection bias)의 효과를 무시할 수 없다.

예제 5.3.2 National Longitudinal Survey of Youth (NLSY79)

NLSY79는 미국의 14-22세 청소년을 모집단으로 1979년에 처음 표본조사를 실시한 패널 조사이다. 이 조사는 미국 노동통계국(Bureau of Labor Statistics)의 지원을 받은 전국 경시적 표본조사의 한 부분이다. 전국 경시적 표본조사는 원 코호트로부터 1966년에 표본을 추출하였는데 이 표본들이 나이가 들어가고 새로운 연

방노동법으로 인해 청소년들의 고용과 교육의 기회가 확대됨에 따라 NLSY79는 미국인들의 노동 시장 경험에 관한 이전 연구들과의 비교를 위한 자료를 제공하기 위해 1979년에 시작되었다. 이 표본조사는 교육정도, 직업훈련 투자, 고용 여력, 수입과 자산, 복지수혜정도, 탁아 비용, 보험비용, 건강수준, 작업장내의 사고, 음주와 마약사용여부, 성적 활동, 결혼과 출생에 관한 항목들을 조사하였다. NLSY79년 1979년 첫 조사 이래 매년 조사가 진행되었으며 1994년 이후는 매 2년마다 조사가 진행되었다.

Davey, Shanahan 과 Schafer(2001)는 이 연구의 자료를 이용하여 다중대치 방법을 이용하여 무응답을 보정하는 방법을 기술하였다. 먼저 여러 가지 무응답의 패턴에서 무응답 여부를 예측하는 로지스틱 회귀분석을 적합하였다. 무응답의 예측변수로는 출생 시 어머니의 나이, 가정이 빈곤했던 기간, 빈곤으로 전환을 한 횟수, 부모의 결혼 유지 여부, 추적(follow-up)시의 어머니의 나이를 이용하였다.

예제 5.3.3 Established Populations for the Epidemiologic Study of the Elderly (EPESE)

EPESE는 미국의 National Institute on Aging of the National Institutes of Health에 의해 지원을 받은 연구이다. 이 연구는 네 개의 지역 사회 안의 65세 이상의 조사 참여자들을 대상으로 1981년부터 1988년까지 매년 인터뷰를 통한 조사를 하였다. 이 연구의 목적은 65세 이상 조사 참여자들의 시간에 따른 인구학적 및 건강관련 특성의 변화를 보는 것이다. EPESE 연구에서 강조하는 목적은 건강기능상태와 장애의 변화 및 이러한 변화의 위험요소를 조사하는 것이다. 연구가 진행되는 동안 많은 수의 참여자들이 하나 이상의 웨이브에서 무응답하였으나

이 후의 웨이브에서는 인터뷰에 다시 응하였다. 자료 안의 이런 무응답을 보정하기 위하여 Beckett Brock 등 (1993, 1996)은 마코프 전이모형 (Markov transition model)을 이용하였다.

로지스틱 회귀분석을 고려한 마코프 모형은 연구 참여자가 간헐적 응답을 한 경우 무응답 웨이브의 모든 가능한 경로를 통한 전이의 우도를 구할 수 있다. 또한 마코프 모형에서 테일러 급수 근사(Taylor series approximation)를 통하여 모수 추정값의 표준오차를 추정함으로써 복잡한 디자인의 특성을 고려하였다.

< 5장 연습문제 >

1. 경시적 연구 또는 패널 조사의 장점과 단점을 기술하시오.

2. 다음의 각 항목에 대하여 참, 거짓을 판별하고 판별의 이유를 설명하시오.
   (가) 횡단면 조사를 조사 대상자를 달리하여 반복 실시하는 경우 개인의 변동 수준을 잘 파악할 수 있는 장점이 있다.
   (나) 경시적 연구 자료의 무응답을 고려한 분석은 무응답 및 개체 내 상관을 보정하는 방법을 사용하여야 유효한 결과를 도출할 수 있다.

제 2부

무응답 자료 분석

사례연구

제 6장 사례연구 I

- 2005년 인구주택총조사 자료에 대한 무응답 대체기법 -

< 학습목표 >

(1) 인구주택총조사에 대하여 설명한다.

(2) 인구주택총조사에서 발생하는 무응답을 처리하는 기법을 설명한다.

(3) 인구주택총조사에서 조사된 변수들 및 무응답이 대체된 변수들을 소개한다.

6.1 인구주택총조사 개요

인구주택총조사는 우리나라의 인구, 가구, 주택에 관한 정보를 파악하여 각종 정책 입안을 위한 기초 자료를 제공할 뿐 아니라 여러 가지 가구와 관련된 경상조사를 위한 표본틀(Sampling Frame)을 만드는 데 있어 기초자료로 활용하기 위하여 실시된다. 0과 5로 끝나는 연도를 기준으로 매 5년마다 통계청에서 실시하는 이 조사는 대한민국 영토 중 행정권이 미치는 전 지역을 대상으로 조사기준 시점 현재 조사지역 내에 상주하는 내, 외국인 및 이들이 살고 있는 모든 거처에서 실시된다. 본 연구에서는 2005년 인구주택총조사 자료를 고려하는데 조사표는 전수조사표와 표본조사표로 구분되어 있고 전수조사표는 기본적인 특성을 파악하기 위해 21개 항목으로 구성되어 있으며, 표본조사표는 전수조사항목 이외에 보다 세부적인 특성을 파악하기 위한 20개 항목을 추가하여 총 41개 항목으로 구성되어 있다. 이 항목들 외에 추가로 16개 시, 도별로 각각 다른 조사항목 3개가 포함되어 전체적으로는 44개 조사항목으로 구성되어 있다.

현행 인구주택총조사에서는 60-80 가구를 하나의 조사구로 설정한 후, 이 중 10%를 표본조사구로 추출하고 표본조사구내 모든 가구는 표본조사표를 작성하도록 하고 있다. 여기서 가구란 1인 또는 2인 이상이 모여서 취사, 취침 등 생계를 같이 하는 생활단위를 말하는데, 크게 일반가구와 집단가구, 외국인가구로 구분된다. 또한 조사구란 전국의 모든 지역에 대하여 식별이 명확한 지형지물을 기준으로 지도상에서 일정한 가구수가 포함되도록 분할한 조사담당 구역을 말한다. 조사구는 아파트조사구, 보통조사구, 섬조사구, 기숙시설조사구, 특수사회시설조사구, 관광호텔 및 외국인 거주 지역 조사구 등 6개로 구분된다.

인구주택총조사에서 발생하는 무응답에는 두 가지 종류가 있는데 (1) 전체 문항에 대하여 응답을 하지 않는 단위무응답(unit nonresponse)과 (2) 일부 항목에 대한 무응답인 항목무응답(item nonresponse)으로 나누어진다. 단위무응답은 가중값을 주어 처리하였고 항목무응답은 통계청에서 무응답대체를 실시한 후 대체 자료를 제공해 왔다. 본 사례에서는 2005년 인구주택총조사에 사용된 대체 방법을 설명하고자 한다.

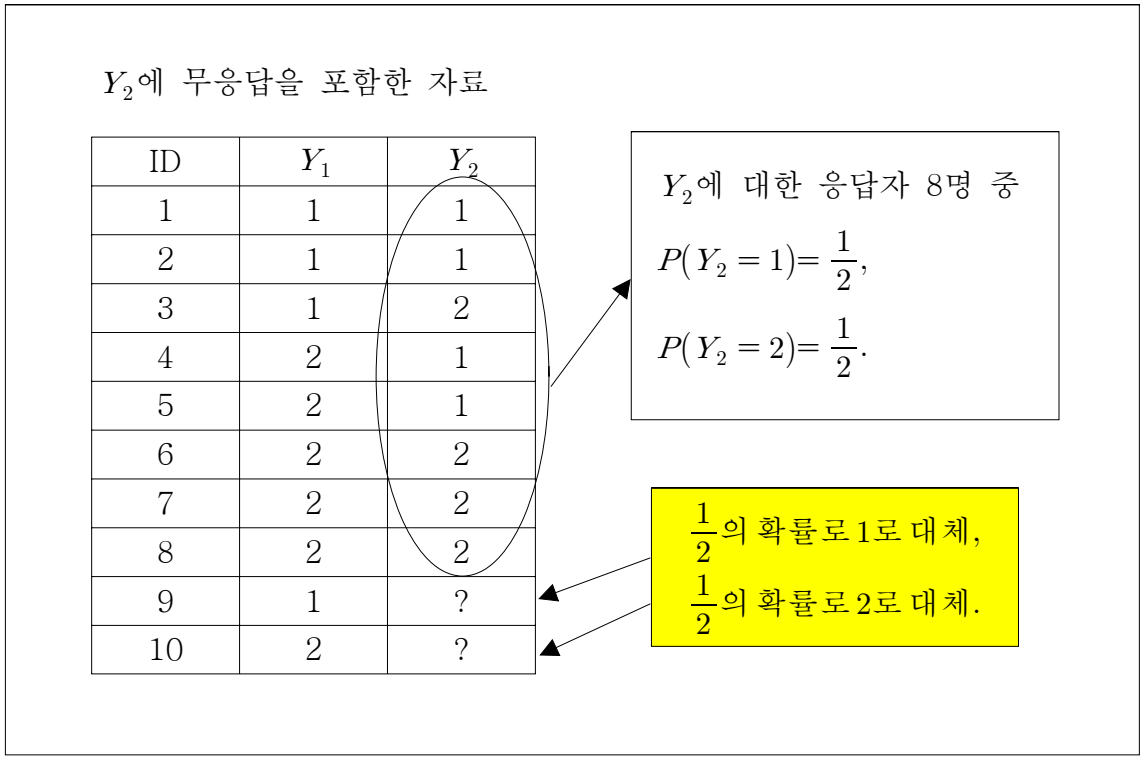
6.2 2005년 인구주택총조사에서 사용된 무응답 처리 기법

2005년 인구주택총조사 자료의 거의 대부분의 항목에서 무응답이 발생하였으므로 다음의 세 가지 대체방법이 단일대체를 위하여 적용되었다(통계청, 2008).

6.2.1 확률에 근거한 대체(Probability Imputation)

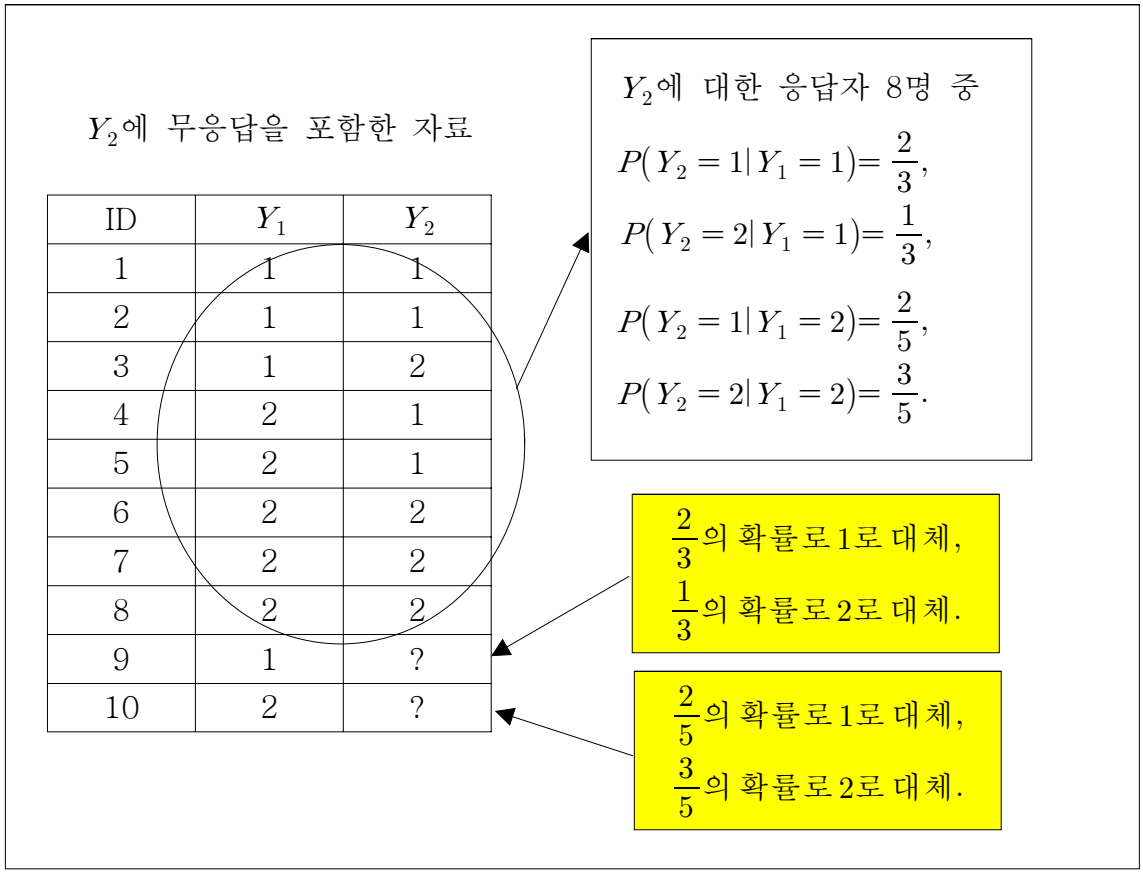
이 대체법은 무응답을 포함한 범주형 변수에서 각 범주별 비율을 대체에 이용하는 방법으로 범주별로 응답자들의 분포에 따라 무응답자를 확률적으로 배분하는 방법이다. 물론 이 방법은 응답자의 범주별 비율을 대체에 사용하므로 변수가 범주형으로 측정된 경우에만 사용 가능하다. <그림 6.1>은 간단한 예로서 두 개의 변수 Y_1과 Y_2가 측정되고 무응답이 1개의 변수 Y_2에서만 발생하는 경우 Y_2의 무응답에 대하여 확률에 근거한 대체를 실시하는 방법을 설명한다. 우선, 무응답이 발생한 Y_2 변수의 응답값들 중 각 항목의 비율의 계산한다. 전체 10명의 관측값 중 8개의 관측값에서는 응답이 되었으며 그 중 4개가 "1"로, 4개가 "2"로 응답되었다. 즉, 응답자 중 "1"이 발생할 확률은 $P(Y_2=1)=\frac{4}{8}=\frac{1}{2}$이고 "2"가 발생할 확률은 $P(Y_2=2)=\frac{4}{8}=\frac{1}{2}$로 동일하게 나타난다. 이 응답의 항목별 확률들에 비례하게 무응답에 대한 대체가 발생한다는 의미로 확률에 근거한 대체라 부른다. 즉, 이 예에서는 "1"과 "2"의 발생확률이 동일하므로 Y_2 변수의 무응답에 대하여 각각 $\frac{1}{2}$의 확률로 "1"과 "2"의 값을 가지고 대체를 실시하게 된다. 실제로 대체를 실행할 때는 0과 1사이의 난수(random number)를 생성하고 이 값이 $\frac{1}{2}$작으면 "1"로, 이 값이 $\frac{1}{2}$보다 크거나 같으면 "2"로 대체값을 할당한다.

<그림 6.1> 확률에 근거한 대체방법



<그림 6.1>에서 설명한 대체는 무응답 자료 메커니즘이 완전임의결측인 경우에만 추정량에 편향이 발생하지 않는다. 물론 완전임의결측인 무응답 메커니즘은 현실에서 만족되는 경우가 그리 많지 않다. 따라서 대체군 내에서 응답값과 무응답값의 분포가 동일하도록 대체군을 형성하여 대체를 실시하는 것이 바람직하다. 즉, 4.1.2절에서 언급한 바와 같이 연관된 변수를 사용하여 대체군을 형성하고 이를 이용한다면 더 정확한 대체를 실시할 수 있을 것이다. 이 경우 대체군에 포함된 변수들이 주어졌을 때 무응답 자료 메커니즘이 임의결측인 경우에도 추정량에 편향이 발생하지 않기 때문이다. <그림 6.2>는 변수 Y_1을 사용하여 대체군을 형성하고 대체군내의 항목별 확률에 근거한 대체방법을 간단한 예를 통해 설명한다.

<그림 6.2> 대체군을 사용한 확률에 근거한 대체방법



ID	Y_1	Y_2
1	1	1
2	1	1
3	1	2
4	2	1
5	2	1
6	2	2
7	2	2
8	2	2
9	1	?
10	2	?

<그림 6.2>에서는 Y_2 변수에서 무응답이 발생하였고 Y_1 변수로 대체군을 형성한 경우를 고려하자. 우선, 무응답이 발생한 Y_2 변수의 응답값들 중 Y_1 변수의 값에 따라 형성된 대체군 내에서 각 항목의 비율의 계산한다. 전체 10명의 관측값 중 8개의 관측값에서 응답이 발생하였고 그 중 $Y_1=1$인 대체군을 고려하면 대체군 내에는 3명의 응답자가 있는 데 이 중 2개가 "1"로, 그리고 1개가 "2"로 응답되었다. 즉, $Y_1=1$인 대체군 내에서 응답자 중 "1"이 발생할 확률은 $P(Y_2=1|Y_1=1)=\frac{2}{3}$이고 "2"가 발생할 확률은 $P(Y_2=2|Y_1=1)=\frac{1}{3}$로 나타난다. 한편, $Y_1=2$인 대체군에는 5명의 응답자가 존재하는데 그 중 2개가 "1"로, 나머지 3개가 "2"로 응답되었다. 즉, $Y_1=2$인 대체군 내에서 응답자 중 "1"

이 발생할 확률은 $P(Y_2=1|Y_1=2)=\frac{2}{5}$이고 응답자 중 "2"가 발생할 확률 $P(Y_2=2|Y_1=2)=\frac{3}{5}$로 나타난다. 각 대체군 내에서 이 확률들에 비례하게 무응답에 대한 대체가 발생한다는 의미로 대체군을 사용한 확률에 근거한 대체라 부른다. 이 예에서는 $Y_1=1$인 대체군 내에서 "1"과 "2"의 발생확률이 각각 $\frac{2}{3}$와 $\frac{1}{3}$이므로 $Y_1=1$이고 Y_2 변수는 무응답인 개체의 Y_2값을 $\frac{2}{3}$의 확률로 "1"로, 그리고 $\frac{1}{3}$의 확률로 "2"로 대체하게 된다. 마찬가지로 $Y_1=2$인 대체군 내에서 "1"과 "2"의 발생확률이 각각 $\frac{2}{5}$와 $\frac{3}{5}$이므로 $Y_1=2$이고 Y_2 변수는 무응답인 개체의 Y_2값을 $\frac{2}{5}$의 확률로 "1"로, 그리고 $\frac{3}{5}$의 확률로 "2"로 대체하게 된다.

인구주택총조사에서는 각 항목별 무응답 발생 확률을 추정하기 위하여 전체조사 자료와 유사한 환경을 구축하여 분포도를 산출한다. 즉, 조사 단위들을 동질적인 특성이 있는 대체군으로 분해하여 대체 확률을 구하도록 하는 것이다. 대체에 사용된 변수들은 주로 무응답을 포함한 변수 Y를 잘 설명해 줄 수 있는 연관된 변수들로 선택하여 각 대체군 내에서 대체 확률을 산출하는데 이 변수들은 각 대체군내에서 무응답을 포함한 변수 Y의 값이 동질성(homogeneity)있게 자료를 분류할 수 있는 변수가 바람직하다. 관심변수의 분포도는 응답자들을 이용하여 구하여지는데 전체 조사대상자에 비하여 무응답자가 무시하기에는 너무 높은 비율을 차지하는 경우에는 과거 조사결과 혹은 유사한 다른 조사의 결과를 이용하여 대체 확률을 구하여 이용할 수도 있다. 2005년 인구주택총조사 자료의 경우 각 문항에서의 항목무응답의 비율이 높지 않으므로 금번 조사 응답자로부터 대체 확률을 산출하였다.

6.2.2 핫덱대체

4.1.2절에서 설명한 대체군을 이용한 핫덱대체 방법은 가구관련 통계조사의 무응답 처리기법으로 주로 사용하는 대체방법으로 연속변수 및 불연속변수에 범용적으로 쓰인다. 이 방법은 이용하는 통계기관마다 무응답이 발생하는 변수의 특성에 따라 약간씩 다른 기법을 적용하여 대체를 실시하기도 한다.

인구주택총조사의 무응답 대체는 무응답을 포함한 변수 Y에 대한 응답자들을 모아 기증자 풀(donor pool)을 만들고 그 가운데 무응답자의 숫자만큼을 무작위로 추출하여 각 무응답자에게 무작위로 한 명의 응답자를 기증자로 할당하여 응답자의 Y값을 대체하는 단일대체 방법이다. 인구주택총조사에 대한 핫덱대체는 변수 Y를 잘 설명할 수 있는 변수들로 구성된 대체군에 따라 나누어 각 대체군 내에서 응답자들에게 균등확률분포(uniform distribution)에서 생성된 무작위 확률값(0과 1 사이의 값)을 할당하고, 동일 대체군 내의 무응답자들에게도 균등확률분포에서 생성된 무작위 확률값을 할당하였다. 응답자와 무응답자들을 각각 할당 받은 확률값을 순서로 나열한 다음 무응답자의 숫자에 해당하는 숫자의 응답자를 할당받은 확률값의 상위부터 차례로 선택하고 이들을 무응답자들의 배열된 순서에 따라 차례차례 할당하는 방법을 사용하였다. 예를 들어, 가구원수 변수를 대체하고자 하는 경우 거주지역과 가구구분 두 개의 변수가 가구원수 변수의 값을 잘 설명할 수 있는 대체군을 형성하는 변수라고 가정하자. 거주지역이 “서울 성북”이며 가구구분이 “가족가구”인 대체군내에 응답자가 500명 존재하고 무응답자가 3명 존재한다면 응답자 500명에 대하여 균등분포에서 생성된 확률값을 할당하여 크기순으로 정렬(sort)하고, 무응답자 3명도 균등분포에서 생성된 확률값을 할당한 후 크기순으로 정렬하여 <표 6.1>과 같이 응답자와 무응답자의 그룹이 할당된 무작위 확률값에 따라 정렬되었다고 하자. 핫덱대체는 무작위 확률값이 가장 작은 3

명의 응답자를 무응답자에게 차례로 할당하는 방식으로 진행되었다. 즉, 무응답자 ID “512”의 가구원수에 응답자 ID “101”의 가구원수 “2”를 대체하였고, 무응답자 ID “130”의 가구원수는 응답자 ID “235”의 가구원수인 “4”를 가지고, 마지막으로 무응답자 ID “247”의 가구원수는 응답자 ID “176”의 가구원수인 “5”를 가지고 대체한다.

<표 6.1> 가구원수를 핫덱대체하기 위한 서울시 성북구 1인가구 대체군내 응답자 그룹과 무응답자 그룹에 할당된 무작위 확률값

		응답자			무응답자		
거주 지역	가구구분	ID	가구원수	무작위 확률값	ID	가구원수	무작위 확률값
서울 성북	가족가구	101	2	0.001	512	?	0.374
서울 성북	가족가구	235	4	0.020	130	?	0.512
서울 성북	가족가구	176	5	0.042	247	?	0.689
서울 성북	가족가구	305	3	0.112			
서울 성북	가족가구	222	4	0.150			
⋮	⋮	⋮	⋮	⋮			

여기서 사용된 핫덱대체는 응답자를 비복원으로(without replacement) 추출하기 때문에 각 대체군에 속하는 응답자의 숫자는 항시 무응답자수 보다 커야한다. 따라서 대체군을 형성하기 위하여 사용될 변수를 선정할 때 미리 각 대체군별로 무응답자 숫자에 비하여 응답자수가 충분한지 검토되어야 한다.

6.2.3 계층적 핫덱대체 (Hierarchical Hotdeck)

이 방법은 4.1.2절에서 언급한 바와 같이 대체군을 이용한 핫덱대체 기법에서 대체군을 형성하는 변수들의 개수가 많아져 일부 대체군내의 무응답자에 대하여 기증자를 찾기 어려운 경우에 흔히 사용되는 방법을 의미하는데 미국 통계청에서 Current Population Survey(CPS)의 소득영역 변수들을 대체하는데 사용했기 때문에 CPS 핫덱이라고도 부르고(David, et. al, 1986) 융통성있는 짝짓기대체 방법(Flexible Matching Imputation method)이라고도 한다. 핫덱대체를 시행할 때 응답자 중에서 기증자를 비복원으로 추출하는데 모든 무응답자에 대하여 대체를 실시하려면 모든 대체군 내에서 응답자의 숫자가 무응답의 숫자보다 훨씬 커야 하지만 현실 자료의 경우 자료의 특성상 일부 대체군 내에서 무응답의 숫자가 응답자의 숫자보다 많아 일부 무응답의 경우 대체할 기증자를 발견할 수 없는 경우에 기증자를 얻기 위해 고안된 방법이다.

6.2.2.에서 고려한 핫덱대체에서는 고려된 모든 공변량들의 항목의 값들을 조합하여 대체군을 형성한다. 이 대체군을 수준-1(level-1) 대체군이라 부른다. 이 대체군 내에서 무응답값은 응답자들 중에서 비복원 무작위 추출로 기증자로 선택된 응답자의 Y값을 가지고 대체된다. 만약에 수준-1 대체군의 일부에서 무응답의 숫자가 응답자의 숫자보다 많은 경우 무응답자 중 일부는 응답자들 중에서 기증자를 구할 수 없어 대체값을 할당 받지 못하게 된다. 이 경우 대체군을 형성하기 위하여 고려한 공변량 중 한 개의 공변량을 제외시키고 나머지 공변량들로 다시 대체군을 형성하면 이를 수준-2(level-2) 대체군이라 부른다. 이렇게 형성된 수준-2 대체군 내에서 수준-1 대체시 대체되지 못한 무응답자들은 대체에 기증자로 사용되지 않은 응답자들 중에서 비복원 무작위 추출로 기증자들을 선택하고 그 기증자들의 Y값을 가지고 대체된다. 이 때 수준-1 대체에서 무응답자들에게 그

들의 값을 기증한 응답자는 제외하고 대체군을 구성하므로 한 명의 응답자가 여러 번 그의 값을 기증하는 일은 발생하지 않는다. 만약 수준-2 대체군 내에서도 기증자를 발견하지 못한 무응답자가 발생한다면 다시 대체군을 형성하기 위하여 고려한 공변량 중 한 개의 공변량을 제외시키고 나머지 공변량들로 수준-3(level-3) 대체군을 형성하여 대체를 진행한다. 이와 같이 대체를 하지 못한 무응답자가 존재하는 경우 대체의 수준(level)을 증가시키는 절차를 모든 무응답자에 대하여 대체가 이루어질 때까지 반복한다. 물론 이 방법도 만약 전체 무응답자의 수가 전체 응답자의 수보다 큰 경우에는 적용하는 데 한계가 있다.

거처하는 주택의 연건평을 측정한 변수에서 무응답이 발생하는 경우 대체군을 형성하기 위하여 고려된 변수들이 거처의 종류, 방의 숫자, 그리고 화장실의 숫자라고 가정하자. 거처의 종류가 “단독주택,” “아파트,” “연립/다세대 주택,” “기타”의 4개의 범주를 가지고, 방의 숫자는 “1개,” “2개,” “3개,” “4개 이상”의 4개의 범주를, 화장실의 숫자는 “1개,” “2개,” “3개 이상”의 3개의 범주를 가진다면 대체군의 개수는 4 × 4 × 3 = 48개가 된다. 즉, 수준-1 대체에서 대체군의 숫자는 최대한 48개가 가능한데 일부 대체군 내에서 무응답자가 기증자를 찾지 못해 대체할 값을 할당받지 못하면 다음 단계에서는 최단 변수인 화장실의 숫자 변수를 제거하고 수준-2 대체군을 형성한다. 이 때 형성된 대체군의 숫자는 4 × 4 = 16개가 되며 수준-1의 대체에서 대체되지 못한 무응답자들은 새롭게 형성된 수준-2 대체군 내에서 대체할 값을 기증받게 된다. 이러한 작업을 모든 무응답자의 값이 대체될 때까지 반복적으로 시행한다. <표 6.2>는 이 방법을 예를 들어 설명한다.

<표 6.2> 계층적 핫덱대체

수준-4 에서 제외	수준-3 에서 제외	수준-2 에서 제외	응답가구			무응답가구		
거처의 종류	방의 숫자	화장실의 숫자	ID	연건평	무작위 확률값	ID	연건평	무작위 확률값
단독주택	4개 이상	3개이상	101	150	0.78	512	?	0.30.
단독주택	4개 이상	3개이상				130	?	0.512
단독주택	4개 이상	3개이상				247	?	0.689
단독주택	4개 이상	2개	305	50	0.002	345	?	0.381
단독주택	4개 이상	2개	222	76	0.070	375	?	0.556
단독주택	4개 이상	2개	454	90	0.099			
단독주택	4개 이상	2개	448	45	0.115			
단독주택	4개 이상	⋮	⋮	⋮	⋮			
단독주택	4개 이상	1개	170	40	0.032	289	?	0.851
단독주택	4개 이상	1개	165	38	0.092			
단독주택	4개 이상	1개	530	50	0.159			
⋮	⋮	⋮	⋮	⋮	⋮			

(1) 수준-1 대체

우선, 수준-1 대체군을 형성한 48개 군 각각에서 핫덱대체를 실시한다. 거처의 종류가 “단독주택”이고 방의 숫자가 “4개 이상”이며 화장실의 숫자가 “3개 이상”인 대체군 내에는 3명의 무응답 가구가 존재하지만 응답가구는 1 가구밖에 없다. 따라서 3명의 무응답 가구 중 ID “512”는 응답가구인 ID “101”의 연건평 “150”을 가지고 연건평의 값에 대한 대체가 실시되지만 나머지 2 가구인 ID “130”과 “247”은 기증자를 발견하지 못하여 대체되지 못하였다. 한편, 다른 대체군인 거처의 종류가 “단독주택”이고 방의 숫자가 “4개 이상”이며 화장실의 숫자가 “2개”인 대체군 내에는 무응답 가구가 한 2가구인데 응답가구는 2개보다 많아 모든 무응답 가구가 대체가 가능하다. 즉, 무응답 가구 ID “345”는 응답가구인 ID “305”의 연건평 “50”을 가지고 연건평의 값에 대한 대체를 실시하고 무응답 가구 ID “375”는 응답가구인

ID “222”의 연건평 “76”을 가지고 연건평의 값에 대한 대체가 실시된다. 마찬가지로 또 다른 대체군인 거처의 종류가 “단독주택”이고 방의 숫자가 “4개 이상”이며 화장실의 숫자가 “1개”인 대체군 내에는 무응답 가구가 1 가구인데 응답가구는 1개보다 많아 이 무응답 가구에 대한 대체가 가능하다. 즉, 무응답 가구 ID “289”는 응답가구인 ID “170”의 연건평 “40”을 가지고 연건평의 값에 대한 대체를 실시한다.

(2) 수준-2 대체

수준-1 대체군에서 6개의 무응답 가구 중 2개의 가구인 ID “130”과 “247”만이 기증자를 발견하지 못하여 연건평값이 대체되지 못하고 무응답으로 남는다. 따라서 대체군을 형성하는 변수 중 화장실의 숫자를 나타내는 변수를 제거하고 수준-2 대체군을 형성한다. 이 대체군 내에서는 예를 들면 <표 6.2>의 거처의 형태가 “단독주택”이며 방의 숫자가 “4개 이상”인 가구들이 화장실의 숫자와 상관없이 한 개의 대체군으로 묶이게 된다. 이 대체군 내에는 화장실의 숫자가 “3개 이상”이며 연건평이 무응답인 ID “130”과 “247”이 대체를 기다리고 있으므로 이 두 가구에 무작위 확률값을 할당한다. 한편, 이 대체군 내에 화장실의 숫자가 “3개 이상”이면서 기증자로 사용될 수 있는 응답가구는 더 이상 남아있지 않지만 화장실의 숫자가 “2개” 또는 “1개”인 응답자 중 기증자로 사용될 수 있는 응답가구는 ID “454,” “448,” “165,” “530” 외 여러 가구가 남아있다. 이 기증자로 사용가능한 응답가구에 다시 무작위 확률값을 할당하고 그 값에 의하여 무응답 가구 ID “130”과 “247”을 대체할 가구를 선정한다.

인구주택총조사 자료에 대한 계층적 핫덱대체를 시행하기 위하여 공변수들은 무응답을 포함한 변수 Y와 상관관계가 높은 범주형 변수들을 고려하였고 상관관계가 가장 높은 변수를 가장 마지막에 제거되도록 지정하였다. 이는 대체군을 형성하는 수준이 높아질 때 무응답이 발생한 변수와 가장 상관관계가 작은 변수를 제거함으로써 연관성이 강한 변수는 최후까지 대체군의 형성에 기여하도록 하였다.

변수의 제거 순서의 결정은 전문가와 상의하여 결정되었다. 이현정(2009)은 이 순서의 결정을 위한 통계적 기법을 제안하였다.

6.2.4 특이점(outlier)의 제거

인구주택총조사 자료 무응답에 대한 핫덱대체에서는 대체군을 설계하기 전에 특이점에 대한 제거작업을 수행하였다. 하지만 수준-1 대체군에서는 대체군의 숫자가 많아 특이점을 제거한다면 응답자 수가 충분하지 않아 일부 대체군 내에서 특이점을 결정하기 어려울 수 있다. 따라서 특이점은 응답자 전체 중에서 특이점 클래스(outlier class)를 형성하고 그 클래스 내에서 이상치를 제거하도록 설계하였다. 특이점 제거를 목적으로 구성하는 특이점 클래스의 형성은 주로 분석하고자 하는 관심 있는 교차표(cross table)에서 사용하는 주요 보조변수를 사용함으로써 대체가 이루어진 이후에 교차표에 예상치 않았던 특이점이 나타날 가능성을 사전에 예방하고자 하는 의도에서 비롯된 것이다.

6.3 2005년 인구주택총조사 변수들 및 무응답 대체 방법

인구주택총조사 변수들은 (1) 가구원 관련 사항, (2) 가구에 관한 사항, 그리고 (3) 주택에 관한 사항으로 구분된다. 이현정(2009)은 (1) 가구원 관련 사항, (2) 가구에 관한 사항, 그리고 (3) 주택에 관한 사항 각각에 대하여 측정된 변수들의 무응답 처리 여부, 무응답 대체를 실시한 경우 6.2절에서 설명한 대체 방법 중 사용된 대체 방법, 그리고 대체군을 형성하는 데 사용된 변수명을 표로 나타내고 있다(이현정, 2009, 표 4-1부터 표 4-3까지).

제 7장 사례연구 II

- 네덜란드 POLS 조사연구 -

< 학습목표 >

(1) 네덜란드 POLS 조사연구에 대하여 설명한다.

(2) POLS조사에서 발생하는 무응답을 처리하는 기법을 설명한다.

(3) 여러 가지 가중방법을 비교한다.

7.1 네덜란드 POLS 조사 개요

1995년 이후로 네덜란드 통계사무국은 POLS(Permanent Onderzoek Leefsituatie: 네덜란드 어)라고 불리는 하나의 통합된 사회조사 시스템을 개발하여 왔다. Statistics Netherlands(1998)에 이 조사에 대한 자세한 사항을 기술하고 있다. POLS는 매달 표본이 추출된다. 목표 모집단은 12세 이상의 네덜란드 국민이다. 표본추출은 두 개의 스테이지로 구성된다. 첫 번째 스테이지에서 몇 개의 큰 지역을 층으로 하여 주민수에 비례하는 추출확률로 지방자치구역을 추출한다. 두 번째 스테이지에서 추출된 각 지방자치구역에서 등확률(equal probability)로 개인표본을 추출한다.

7.2 가중방법

이 사례 연구에서는 자원봉사에 관련된 한 변수에 관하여 가중방법의 영향을 조사한다. 주변수는 한 개인의 자원봉사 여부이다. 사회활동에 참여하는 사람이 좀 더 조사에 적극적으로 참여하는 경향이 있으므로 자원봉사 여부와 조사에 응답여부는 관련이 있을 것으로 예상된다. 1997년 조사에서 6672명의 표본에서 56.6%가 조사에 응답하였다.

이 사례 연구에서는 단지 Statistical Yearbook of Statistics Netherlands로부터 입수할 수 있는 정보만을 이용하였다. 즉, 성별, 나이, 결혼여부, 지역, 지역의 도시화 정도에 관한 빈도분포를 이용하였다. 이 다섯 변수의 완전한 결합분포를 알면 가장 이상적이겠으나 Statistical Yearbook은 단지 변수들의 부분 결합분포에 관한 정보만 제공한다. <표 7.1>은 성별, 나이, 결혼유무에 관한 모집단 결합분포이다.

<표 7.1> Population distribution of age × sex × marital status (×1000)

	Male				Female			
Age	Unmarried	Married	Widowed	Divorced	Unmarried	Married	Widowed	Divorced
12-19	752.4	0.4	0.0	0.0	716.5	3.5	0.0	0.0
20-29	981.5	185.7	0.2	10.2	785.0	330.6	0.7	22.7
30-39	445.4	795.1	1.9	72.1	283.5	879.3	5.7	93.8
40-49	164.7	899.0	6.9	113.9	103.1	882.9	21.5	138.4
50-59	67.3	732.9	15.8	86.3	44.4	647.9	56.1	98.8
60-69	42.0	519.2	31.7	42.6	41.4	458.9	140.0	51.5
70-79	21.4	308.4	52.5	16.6	43.0	239.9	254.3	27.9
80+	8.0	84.0	50.4	4.0	35.0	49.6	243.9	12.4

지역(province)과 도시화정도에 관해서는 나이와의 결합분포만이 알려져 있다. (<표 7.2>와 <표 7.3>)

<표 7.2> Population distribution of province × age (×1000)

	Age				
Province	12-19	20-44	45-64	65-79	80+
Groningen	49.3	222.1	127.8	48.4	20.8
Friesland	61.5	225.7	144.0	65.7	21.6
Drenthe	43.7	165.9	114.3	53.9	15.3
Overijssel	106.2	404.2	240.2	110.8	31.9
Flevoland	33.8	115.7	53.0	21.8	4.2
Gelderland	183.4	720.6	443.4	195.7	56.9
Utrecht	103.7	439.4	238.6	102.2	32.5
Noord-Holland	220.5	999.9	574.4	254.3	82.1
Zuid-Holland	316.2	1307.9	759.5	347.1	117.7
Zeeland	35.0	130.1	89.2	44.1	15.6
Noord-Brabant	218.7	898.7	562.4	225.2	57.9
Limburg	100.8	429.5	289.8	127.1	30.8

<표 7.3> Population distribution of degree of urbanization × age (×1000)

	Age				
Province	12-19	20-44	45-64	65-79	80+
Very Strong	223.2	1196.6	565.3	293.4	113.0
Strong	336.5	1468.4	839.0	389.8	114.6
Moderate	317.1	1223.7	766.3	321.7	90.5
Little	333.7	1226.3	820.6	328.8	93.4
None	232.3	944.7	645.4	262.6	75.8

<표 7.2>와 <표 7.3>에서는 나이가 5수준으로 구성되어 있으나 <표 7.1>은 나이

가 8수준으로 구성되어 있다. 선형 가중방법에서는 같은 변수의 다른 범주화가 문제가 되지 않는다. 두 개의 나이 범주 변수를 동시에 고려할 수 있다. 사후-층화방법은 이 표 중에서 단지 하나만을 이용할 수 있다. 또한 <표 7.1>은 네 개의 빈 칸이 있으므로 사후-층화방법에서 그대로 사용할 수 없다. 이 문제를 해결하기 위해서는 빈 칸이 있는 층을 다른 층과 합치는 방법을 생각해 볼 수 있다. 예를 들면 나이 수준 12-19와 20-29를 합쳐서 12-29의 새로운 수준으로 만들 수 있다.

가중방법을 적용하기 위해서 각 보조변수의 응답자에서의 비율과 모집단에서의 비율을 비교하였다. (<표 7.4>)

<표 7.4> Comparing population and response distributions of the auxiliary variables (%)

Variable	Response	Population	Difference
Age			
12-19	12.8	11.1	1.7
20-29	15.9	17.5	-1.6
30-39	20.5	19.4	1.1
40-49	17.9	17.6	0.3
50-59	14.0	13.4	0.6
60-69	10.0	10.0	0.0
70-79	6.5	7.3	-0.8
80+	2.5	3.7	-1.2
Marriage Status			
Unmarried	32.7	34.2	-1.5
Married	57.2	53.2	4.0
Widowed	5.2	6.0	-0.8
Divorced	4.9	6.7	-2.8
Province			
Groningen	2.7	3.5	-0.8
Friesland	4.3	3.9	0.4
Drenthe	2.3	3.0	-0.7

Overijssel	6.8	6.7	0.1
Flevoland	1.8	1.7	0.1
Gelderland	15.4	12.1	3.3
Utrecht	5.4	6.9	-1.5
Noord-Holland	14.0	16.1	-2.1
Zuid-Holland	18.0	21.5	-3.5
Zeeland	2.7	2.4	0.3
Noord-Brabant	17.6	14.8	2.8
Limburg	9.1	7.4	1.7
Sex			
Male	48.6	49.1	-0.5
Female	51.4	50.9	0.5
Urbanization			
Very strong	11.8	18.0	-6.2
Strong	24.0	23.8	0.2
Moderate	23.2	20.5	2.7
Little	23.3	21.1	2.2
None	17.7	16.5	1.2

나이 변수에서 무응답은 20대와 30대에서 가장 높았다. (집에 없는 경우가 많음) 또한 나이가 많은 군에서도 무응답이 높았다. (응답거절이 많음) 결혼한 사람들의 응답률이 비교적 높았고 Gelderland와 Noord-Brabant 지역에 사는 사람들의 응답이 비교적 높았다. Noord-Holland와 Zuid-Holland와 같이 산업화된 지역의 응답률은 비교적 낮았다. 이런 무응답율에 관한 분석은 적어도 “결혼여부”, “지역”, “도시화 정도”는 가중 모형에 포함시켜야함을 보여준다. 여기서 “지역”과 “도시화 정도”는 부분적으로 교란(confound)되어 있다.

<표 7.5>는 이 사례 연구에서 적용된 여러 가지 가중 모형으로부터 얻은 결과를 보여준다.

<표 7.5> Estimates of the percentage of people doing volunteer work, based on various weighting models

Weighting Model		Number of parameters	Estimate	Standard error
1	No weighting	0	43.4	1.2
2	Sex	2	43.4	1.2
3	Province	12	43.3	1.2
4	Marital Status	4	42.9	1.2
5	Urbanization	5	42.9	1.0
6	Age8*	8	42.8	1.2
7	Age5**×Province	60	42.9	1.2
8	(Sex×Age8)+(Sex×Marital)	22	42.3	1.1
9	Age5×Urbanization	25	42.5	1.0
10	Sex+Age8+Marital+Urban+Province	23	42.1	1.0
11	(Sex×Age8)+(Sex×Marital)+(Age5×Urbanization)+Province	53	42.0	0.9

* Age8: Age variable with 8 levels

** Age5: Age variable with 5 levels

<표 7.5>를 보면 보조변수가 많이 고려될수록 주모수인 자원봉사에 참여하는 사람의 비율의 추정치가 점점 작아지고 있음을 알 수 있다. 물론 가중모형의 유효성을 단지 보정되지 않은 추정치와 단순비교로 판단할 수는 없다. 하지만 좀 더 많은 보조변수의 정보를 이용할수록 표준오차가 감소함을 볼 수 있다. 이는 가중모형이 잘 적합되었음을 간접적으로 보여주고 있다.

성별, 결혼여부와 나이를 고려한 사후-층화는 빈칸 때문에 불가능하다. 나이 수준 12-19와 20-29를 합치더라도 그 칸에서의 빈도는 5보다 작으므로 가중값은 불안정하다. 그래서 Sex × Marital Status × Age8을 이용한 사후층화분석 대신에 선형 가중 모형 (Sex×Age8)+(Sex×Marital)을 이용하였다.

<표 7.5>의 모형 11은 최대로 가능한 보조변수의 조합을 고려한 모형이다. (Sex×Age8)+(Sex×Marital)에 해당하는 모집단의 정보는 <표 7.1>로부터 (Age5×Urbanization)에 해당하는 모집단의 정보는 <표 7.3>으로부터 얻을 수 있다. 한 가지 주목할 점은 Age5 × Province의 몇 몇 칸의 빈도가 너무 작아서 모형에서 Age5 × Province 대신에 Province의 정보만 <표 7.4>로부터 사용하였다. 이 모형 11을 적용한 가중값은 가중보정을 하지 않은 경우의 추정치와 비교하였을 때 가장 큰 감소를 보였다. (43.4에서 42.0으로) 좀 더 단순한 모형인 모형 10도 모형 11과 거의 비슷한 결과를 제시하였다. 이는 추정치의 편향을 줄이는데 보조변수의 주효과 (main effects)가 보조변수간의 상호작용효과 (interaction effects) 보다 더 중요한 역할을 하고 있음을 보여준다.

제 8장 사례연구 III

- 2006년 고령화연구패널 제 1차 자료에 대한 무응답 대체기법 -

< 학습목표 >

(1) 고령화연구패널조사에 대하여 설명한다.

(2) 고령화연구패널조사에서 발생하는 무응답을 처리하는 기법을 설명한다.

(3) 고령화연구패널조사에서 무응답 현황 및 무응답이 대체된 변수들을 소개한다.

8.1 고령화연구패널조사 개요

한국 노동연구원에서 실시하는 고령화연구패널조사(KLoSA)는 한국 내 고령자 모집단에 대한 실태를 파악하고 기초 자료를 수집하여 다가오는 고령화 시대에 대비하기 위하여 시행되고 있다. 2006년 제 1차 조사(baseline survey)가 실시된 이후 매 2년마다 추적조사를 시행하고 추적조사를 시행하지 않는 연도에는 부가조사가 시행되고 있다. 고령화연구패널은 45세 이상인 일반가구 거주자를 대상으로 하는데 제 1차 조사는 10,254명의 연구 참여자(participant)에 대한 자료를 축적하고 있다. 고령화연구패널조사 자료는 커버스크린, 인구학적 배경, 가족 관계, 건강과 고용 상태, 소득과 자산 상황, 그리고 삶의 만족도 등 심리적 상태까지 총 8개 부문으로 나누어 1,310개의 다양한 조사 항목을 포함하고 있다. 본 사례연구는 고령화연구패널 제 1차 조사에서 발생하는 무응답에 대한 대체 방법을 고찰한다.

고령화연구패널 제 1차 조사 자료는 대부분의 변수들에서 무응답의 비율이 5% 미만으로 작게 나타났다. 하지만 주요 관심영역인 소득 및 자산 영역의 경우 변수 중 일부에서 결측값의 비율이 10~20%로 상대적으로 높게 나타났다. 특히 자산 영역의 일부 변수의 경우 무응답의 비율이 거의 30%까지 나타나 낮지 않은 비율을 차지하고 있다. 한편, 표본을 추출할 때 사용한 설계변수(design variable)들은 완전하게 관측되어 무응답이 존재하지 않았다.

고령화연구패널 제 1차 조사 자료에 대하여 다중대체가 실시되었다. 대체의 숫자는 무응답으로 인하여 손실된 모수에 대한 정보량을 고려하여 충분할 것으로 생각되는 5번으로 결정되었다. 즉, 5개의 다중대체된 자료가 생성되었고 노동연구원 홈페이지를 통해 연구자들에게 제공되어 왔다.

8.2 2006년 고령화연구패널 제 1차 조사에서 사용된 무응답 처리 기법

고령화연구패널 제 1차 조사에서 발생하는 무응답을 대체하는 작업은 8개 부문에 대하여 순차적으로 실시되었다. 그 이유는 설문 조사를 통해 측정된 이 자료는 변수의 개수가 많고 변수마다 응답이 가능한 참여자의 숫자가 다른 경우도 많았기 때문이다. 예를 들면, 임금 총소득은 응답 대상이 임금을 받는 근로자에 한정되고 농어업 소득은 그 대상이 농어업 종사자이므로 각 변수마다 응답 대상자가 다르기 때문이다. 이 때 응답 대상자가 아닌 연구 참여자에 대한 응답값은 입력된 자료상에는 무응답으로 표시되지만 1.1절에서 설명한 바와 같이 그 변수에 대한 해당자가 아니므로 무응답이라 할 수 없고 따라서 대체가 실시될 필요도 없다. 대체는 결측값을 영역에 따라 몇 개의 단계로 나누고 단계에 따라 순차적으로 다음과 같이 적용하였다. 첫 번째로 인구영역의 변수들에서 발생하는 무응답이 대체되었

는데 이 변수들은 배경변수(background variables)에 대하여 조사되었고 결측값의 비율도 매우 낮아 우선적으로 대체가 실시되었다. 대체모형에는 설계변수 및 다른 연관된 인구영역 변수들이 설명변수로 포함되었다. 두 번째로 건강영역의 변수들에서 발생하는 무응답이 대체되었는데 인구영역 및 설계변수들을 설명변수로 대체모형에 포함시켜 진행되었다. 고용영역 변수들에서 발생한 무응답이 그 다음 순서로 대체되었고 대체모형은 앞 영역에서 대체된 변수들을 설명변수로 포함하였다. 네 번째로 소득영역의 주요 변수들에서 발생한 무응답이 대체되었는데 대체모형은 인구영역, 설계변수, 건강영역, 고용영역, 그리고 주요 자산 영역의 대체된 변수들을 설명변수로 포함하였다. 다섯 번째로 자산영역의 주요 변수들에 대한 무응답을 앞 단계에서 대체된 연관된 변수들을 설명변수로 포함하여 대체하였다. 마지막으로 가족영역의 금전관계 변수들이 대체되었다. 이 영역이 마지막으로 대체된 이유는 가족영역의 경우 가족 대표자 한 명만이 응답하였고 여러 자녀에 대한 자료를 포함하므로 다른 영역과 자료의 형태가 다르기 때문이었다.

고령화연구패널 제 1차 조사에서 선택한 무응답 대체 모형은 4.2.1절에서 설명한 예측 평균값에 근거한 핫덱 방법(hotdeck based on a predictive mean matching)이다. 이 방법은 모의실험에서 우수한 결과를 보여 온 대체방법으로서 무응답을 자료 내 응답값 들 중 하나 또는 여러 개의 값으로 대체시키는 대체군을 이용한 핫덱 방법이다. 이 때 대체군은 무응답이 발생한 변수에 대하여 관측된 자료들만을 대상으로 회귀모형을 적합한 후 무응답값을 포함하여 모든 개체에 대하여 예측값을 구한 후 그 값에 근거하여 층화(stratification)하여 구성한다. 각 대체군 내에서 무응답은 동일 대체군 내 응답자 중에서 기증자를 선택하여 기증자의 값으로 대체하므로 단순임의 핫덱방법보다 회귀모형의 예측력이 클수록 좋은 결과를 기대할 수 있다. 또한, 예측값에 근거하여 대체군을 형성하기 때문에 설명변수의 숫자나 형태에 의존하지 않는 장점을 지닌다.

고령화연구패널조사 자료의 경우 변수별 특징에 따라 예측 평균값에 근거한 핫덱 대체 방법이 적절히 변형되었다.

8.2.1 예측 평균값에 근거한 핫덱대체

변수가 특별한 특징을 가지지 않는 경우 무응답에 대한 대체는 예측 평균값에 근거한 핫덱대체 모형을 사용하였다. 우선, 무응답을 포함한 각 변수에 대하여 응답자의 자료만을 대상으로 회귀모형(regression model)을 적합한 후 무응답을 포함한 모든 자료에 대하여 예측값을 구하고 그 값에 근거하여 층화(stratification)하여 대체군을 형성한다. 이 때 각 대체군은 가능한 한 10명 이상의 구성원을 포함하도록 구성하여 대체군 내에서 기증자를 선택하기 용이하도록 하였다. 구성된 각 대체군 내에서 무응답은 같은 대체군의 응답자 중에서 기증자를 임의로 선택한 후 기증자의 값으로 대체하였다.

8.2.2 범주형 전환문장(unfolding bracket question)을 포함한 변수에 대한 예측 평균값에 근거한 핫덱대체

소득 및 자산 부문의 주요 변수들에 대하여 가능한 한 많은 정보를 얻기 위하여 범주형 전환문장이 사용되었다. 즉, 주요 변수에서 응답을 거절하거나 응답 문항들 사이에 불일치가 나타나는 경우 범주형 전환문장들을 사용하여 부분 정보를 제공할 수 있도록 하였다. 예를 들어, 임금 총소득에 관한 응답이 거절되거나 부정확하게 응답되는 경우 임금 총소득에 관한 서로 겹쳐지지 않는(unfolding) 선정

된 구간들(brackets)을 제시하고 그 구간들 중에서 선택하도록 함으로써 무응답자에 대한 정확한 임금 총소득 대신 구간으로 응답된 부차적 임금 소득 정보를 얻을 수 있도록 하였다. 문제는 이렇게 얻어진 정보가 정확한 임금 소득이 아니라 구간으로 표현되므로 임금 총소득에 대한 분석을 시행할 때 정확한 임금 총소득액과 간단히 통합할 수 없다는 점이다. 따라서 이 변수에 대한 무응답 대체는 범주형 전환문장들에 포함된 정보를 이용하여 실시되었다. 즉, 범주형 전환문장에 대한 응답자는 소득이 일정 구간에 속한다는 정보를 제공하였으므로 동일한 구간 내의 응답된 자료들만을 기증 대상으로 선택하고 동일한 대체군에 속하는 관측값을 기증자로 선택함으로써 대체된 값들이 응답된 구간 안에 존재하도록 하여 대체된 자료의 일치성(consistency)을 만족시켰다. 또한, 이 대체방법은 동일 대체군에 속한 기증자들의 예측값이 무응답에 대한 예측값이 비슷하므로 대체된 값을 가능한 한 잘 예측할 수 있을 것으로 기대된다. 물론, 문항에 따라 범주형 전환문장에도 응답하지 않은 응답자가 상당수 존재하였으며 이 경우 전체 응답자 중 회귀 모형을 통하여 동일한 대체군에 속하는 기증자를 선택하여 대체하였다.

고령화연구패널조사 설문에는 서로 겹쳐지지 않는 5개의 범주형 전환문장들이 시행되었으므로 각 응답값은 6개의 구간 중 하나의 구간으로 표현될 수 있다. 정확한 값 대신 범주형 전환문장에 응답한 경우 참값이 어느 구간 안에 존재하는 지에 관한 정보가 주어지므로 이 정보를 사용하여 대체가 실시되었다. 예를 들어 한 무응답자의 범주형 전환문장에 대한 응답이 “2400 MW 이상 6000 MW 미만“이라면 이 사람의 대체된 값은 이 구간 안에 속해야 제공된 정보와 일치하는 대체가 이루어지는 것이다. 따라서 우선 응답자들의 관측값들을 범주형 전환문장에서 선택한 6개의 구간으로 분리하였다. 범주형 전환문장의 동일 구간에 속하는 응답자들은 무응답자에 대한 대체를 위한 기증자 후보 집단(pool)으로 사용되었다. 각 구간별로 다수의 응답자가 있는 경우 회귀 모형에 의한 예측값을 사용하여 대체

군을 형성하고 동일한 대체군 내에서 기증자를 선택하여 기증자의 값을 가지고 대체를 실시하였다.

8.2.3 기증자를 발견하지 못한 경우

응답자의 숫자가 많지 않은 일부 문항의 경우 범주형 전환문장을 사용하여 얻어진 부차 정보에 근거한 대체군 내에서 기증자를 발견하지 못하는 경우가 발생하였다. 예를 들어, 소득이 아주 많은 사람들의 대부분이 소득 항목에 대하여 무응답인 경우 기증자를 발견하기 힘든 경우가 발생하였다. 대체군 내에서 기증자가 존재하지 않는 경우 예측 평균값에 근거한 핫덱대체 방법을 9.2.4절에서 토의하는 혼합 모형 대체(mixed imputation method)에서처럼 회귀모형(regression model)의 예측값(predicted mean)에 근거한 대체와 혼합해 실시하도록 확장하였다.

8.2.4 선다형 문항에 대한 무응답 대체

고령화연구패널조사 일부 항목은 한 사람이 여러 개의 답을 제시하는 선다형 문항(multiple choice question)이다. 예를 들어 보험의 경우 개인당 여러 개의 보험을 가지고 있을 수 있고 이 경우 각각의 보험에 대하여 응답이 요구되었다. 따라서 각각의 보험 액수에 대하여 대체를 실시할 때 동일인에 의한 여러 가지 보험의 액수는 서로 연관되어 있으므로 연관성을 고려하여 예측이 실시되어야 한다. 또한, 가족 영역 지원금 관련 문항의 경우 여러 자녀로부터 또는 여러 자녀에게 동시에 지원을 받거나 지원을 하는 경우가 이에 해당된다. 하지만 회귀모형을 적합할 때 각 관측값은 서로 독립이라 가정하여 모형이 적합된다. 선다형 문항의 경

우 복수의 응답들은 서로 독립이 아니므로 각 개인당 여러 개의 관찰값의 연관성을 포함하도록 회귀모형의 적합 방법을 변형시켰다. 즉, 대체에 사용된 회귀모형이 각 응답과 관련된 특성 변수를 설명변수로 포함하는 동시에 관측값들 사이의 연관을 분산공분산 행렬을 통해 고려하고 Generalized Estimating Equations(GEE) 방법으로 모수를 추정한 후 예측값을 계산하도록 확장하였다. 이 방법은 회귀 모형을 적합할 때 회귀 모수가 관측값 간의 연관성을 포함하여 GEE 방법으로 추정되도록 하기 위하여 SAS Macro에서 REG procedure 대신 GENMOD procedure를 사용하도록 수정함으로써 적용되었다.

8.2.5 연관된 문항들 사이의 일치성 만족

설문 문항 중 동일 영역의 일부 변수들은 연관되어 있으므로 연관된 변수들 사이에 일치성(consistency)을 만족시키도록 대체가 실시될 필요가 있다. 예를 들어 임금 총소득을 측정할 때 그 해 일을 한 달들도 측정되었고 임금 총소득은 일을 한 달의 숫자에 따라 달라진다. 이와 같이 연관이 있는 것으로 생각되는 변수들의 경우 n-분할 대체(n-partition imputation)를 사용하여 한꺼번에 무응답에 대한 대체를 실시하였다(Marker et. al., 2002). 즉, 연관된 변수들 중 주요 관심 변수에 대하여 수정된 예측 평균에 근거한 핫덱대체를 실시하고 대체를 위해 선택된 기증자의 다른 변수값들을 가지고 결측값이 발생하는 연관된 변수들을 그룹으로 한꺼번에 대체하였다. 예를 들어 임금 총소득을 대체한 후에 일을 한 달의 숫자도 무응답인 경우 대체에 사용된 기증자의 일을 한 달 수를 가지고 무응답인 일을 한 달 숫자를 대체하는 방식을 채택하였다.

8.2.1절부터 8.2.5절에서 논의한 대체방법들은 혼합적으로 적용되는 게 가능했다.

예를 들어 임금 총소득의 대체를 실시할 때 일부 참여자는 정확한 임금 소득 대신 범주형 전환문장을 사용한 부분적인 임금 정보만을 제공하는데 반하여 일부는 범주형 전환문장에 대한 응답도 거부하였다. 이 경우 범주형 전환문장에 대하여 응답한 사람의 임금 총소득은 범주형 전환문장에 대한 응답 구간에 따라 8.2.2절의 대체 방법으로 대체를 실시하고 범주형 전환문장에 대한 응답조차 거부한 경우 8.2.1절의 대체모형을 사용하여 대체하였다. 또한 기증자를 찾지 못한 극한 구간의 무응답은 8.2.3절의 대체모형을 적용하여 대체를 실시하고 나머지는 8.2.1절이나 8.2.2절의 모형을 이용하여 대체되었다.

대체는 남자와 여자에 대하여 독립적으로 시행되었다. 그 이유는 고령화연구패널조사가 45세 이상 남녀를 대상으로 하고 있는데 직업에 종사하는 비율, 임금 및 자산의 분포가 성별에 따라 크게 다르게 나타났기 때문이다.

8.3 고령화연구패널조사에서의 변수별 무응답 현황 및 무응답 대체 방법

송주원 외(2007)는 영역별로 주요 변수들 각각에 대하여 변수별 해당되는 관측값의 숫자, 무응답의 숫자 및 해당자 중 무응답 비율을 보고하고 있다. 또한 각 영역별 그리고 변수별로 8.2절에서 설명한 모형 중 어느 대체모형이 선택되었는지를 표로 나타내고 대체를 시행할 때 설명변수로 사용한 변수들에 관한 세부 정보도 포함하고 있다.

무응답에 대한 대체를 시행한다고 하더라도 대체된 자료를 사용하여 분석을 실시할 지 또는 응답된 자료만을 사용하여 분석을 실시할 지에 관한 판단은 연구자의 몫으로 남겨두는 것이 바람직하다. 대체된 자료안의 응답값들은 정확하게 측정된

값이지만 대체된 값은 응답되었다면 관측했을 값이라고 100% 확신할 수 없기 때문이다. 하지만 무응답이 대체된 자료만 제공되는 경우 이 자료는 무응답이 없이 완전한 형태를 가지게 된다. 따라서 대체된 자료에서 어느 값이 원래 응답값이며 어느 값이 대체된 값인지를 구분할 수 있는 정보는 매우 유용하게 쓰일 수 있다. 이 구분이 존재한다면 응답된 관측값 만에 근거하여 분석을 실시하길 원하는 연구자는 원하는 분석을 시행하는 것이 가능할 것이고 대체된 자료값들과 응답된 자료값들의 비교도 가능하다. 고령화연구패널 제 1차 조사 자료의 대체를 실시할 때 대체된 각 변수에 대하여 대체 여부를 나타내는 플래그 변수(flag variable)를 추가함으로써 응답값과 대체된 값의 구분이 가능하도록 하였다. 플래그 변수는 원 변수의 이름에 밑줄(underline) 기호 "_"를 추가시킨 변수명을 취한다. 예를 들어, 임금 소득액을 나타내는 변수 w01E003의 경우 w01E003_라는 플래그 변수가 새롭게 생성되고 이 변수는 다음의 값들로 구분된다.

$$\begin{cases} 0 : \text{응답한 관측값임} \\ 1 : \text{대체된 값임} \\ 2 : \text{범주형 전환문장에 응답} \\ 3 : \text{가족대표자의 응답을 가지고 대체} \\ \bullet : \text{이 문항에 대한 응답 대상자가 아님} \end{cases}$$

즉, 플래그 변수를 사용하여 자료를 응답된 관측값, 대체된 값, 또는 문항에 대한 응답대상자가 아니므로 분석에 고려하지 말아야 할 개체들을 구분하는 것이 가능하도록 하였다.

제 9장 사례연구 IV

- 미국 Health and Retirement Survey 자료에 대한 무응답 대체기법 -

< 학습목표 >

(1) 미국 Health and Retirement Survey(HRS)에 대하여 설명한다.

(2) 미국 HRS에서 발생하는 무응답을 처리하는 기법을 설명한다.

(3) 미국 HRS 자료의 무응답이 대체된 변수들을 소개한다.

9.1 미국 Health and Retirement Survey (HRS) 개요

미국 Health and Retirement Survey(HRS)는 미국 전역에 걸쳐 노인들의 경제, 건강, 가족 상태를 조사하는 패널자료(panel data)이다. 1992년부터 매 2년마다 실시되는 이 조사는 22,000명 이상의 패널에 대한 신체적, 정신적 건강 상태, 경제 및 고용 상태, 그리고 가족 상태를 연구한다. 이 자료는 무응답의 비율이 높은 변수들을 포함한다는 점에서 무응답 자료의 연구에 중요한 가치를 지닌다. 또한, 이 자료는 제 8절에서 고려한 고령화패널조사 자료와 측정한 변수 및 대상 집단에서 유사한 면이 많아 대체 방법의 다양성을 보여주는 사례이다.

HRS 자료에서 발생하는 무응답에 대하여 다중대체가 실시되었고 대체된 자료를 홈페이지를 통해 사용자에게 제공해 왔다. 또한, 무응답에 대한 대체를 실시하는데 사용한 통계 프로그램 모듈(SAS Macro IMPUTE)도 개발하여 연구자들에게 제공하고 있으므로 연구자들이 무응답에 관한 대체를 시행할 때 좋은 참조가 될 수 있다.

9.2 미국 HRS에서 발생하는 무응답을 처리하는 기법

미국 HRS 자료의 특징은 제 8장의 사례연구와 마찬가지로 무응답으로 인한 정보의 손실을 줄이기 위하여 범주형 전환문장들을 사용하여 부차적 정보를 얻었다. 하지만 모든 변수에 대하여 범주형 전환문장들이 사용된 것은 아니므로 변수의 특성에 따라 적절한 대체 방법을 사용하였다. 또한 연금 항목의 경우 부부 모두로부터 정보를 얻었고 부부간 정보가 연관되어 있으므로 부부의 정보를 함께 포함하여 대체를 실시할 수 있도록 프로그램을 구현하였다. 적용된 대체 방법은 다음의 네 가지로 구분될 수 있다.

9.2.1. 중위수 대체

일부 변수의 무응답값은 무응답이 발생한 변수에 대한 응답값의 중위수(median)를 이용하여 대체를 실시하였다. 이 대체 방법은 한 변수의 응답된 값들 중 중위수를 가지고 그 변수의 모든 무응답값을 대체한다는 점에서 중위수 대체라 부른다. 이 방법은 자료의 분포가 비대칭(asymmetric)이거나 특이점이 발생할 때 평균대체보다 좋은 결과를 보이는 것으로 알려져 있다. 평균대체와 비슷하게 대체군을 형성한 뒤 대체군 내에서 중위수 대체를 실시하는 것도 가능하다. HRS 자료의 경우 범주형 전환문장을 사용한 경우 정확한 응답값 대신 값에 대한 부분적인 정보가 구간으로 주어지므로 주어진 구간 내의 응답값들 중 중위수를 구하여 대체를 실시하였다.

9.2.2 핫덱대체

일부 변수의 무응답값은 4.1절에서 언급한 완전임의 핫덱 방법을 사용하여 대체를 실시하였다. 이를 위하여 우선 응답자와 무응답자 모두에게 임의로 순서화된 값들을 할당하였다. 예를 들어 전체 자료에 관측개체가 10개 존재하는 경우 (이 중 일부는 응답값이고 일부는 무응답값이다) 이 10개의 값들은 1부터 10까지의 숫자 중 한 개의 숫자를 임의로 할당받았다. 이 할당받은 숫자들을 가지고 자료를 증가하는 순서(ascending order)로 정렬(sort)한 후 각 무응답값을 자신의 할당된 숫자 바로 직전의 숫자를 할당받은 응답값을 가지고 대체하였다. 무응답값이 연속인 두 개 이상의 숫자를 할당받은 경우 이 값들의 직전 숫자를 할당받은 응답값을 가지고 이 무응답값 모두를 대체하였다. 범주형 전환문장이 포함된 변수의 경우 값에 대한 부분 정보인 구간 정보가 주어진다면 그 주어진 구간 내의 응답 또는 무응답값들에 대하여 임의로 순서화된 값들을 할당한 후 무응답자의 할당된 숫자 직전의 값을 할당받은 응답자를 기증자로 정하여 대체를 실시하였다. 이 방법은 4.1절에서 다룬 핫덱방법을 시행하는 다른 알고리즘을 설명한다.

9.2.3 점수(score) 대체

무응답이 발생한 변수를 반응변수로 놓고 이 변수와 연관된 변수들을 설명변수로 사용하여 응답값들만에 근거하여 회귀분석을 실시한다. 회귀모형의 계수가 추정되면 이 계수들을 사용하여 응답값과 무응답값 모두에 대하여 예측값을 계산한다. 예측값에 근거하여 응답자와 무응답자를 증가하는 순서로 정렬한 후 순서에 따라 차례로 증가하는 숫자를 할당한다. 무응답값을 자신의 할당된 숫자 직전의 숫자를 할당받은 응답값을 가지고 대체하였다. 무응답값들이 연속적으로 두 개 이상의 숫자를 할당받은 경우 이 값들의 직전 숫자를 할당받은 응답값을 가지고 이 무응답값

모두를 대체하였다. 마찬가지로 범주형 전환문장이 포함된 변수의 경우 값에 대한 부분적인 정보인 구간 정보가 주어진다면 그 주어진 구간내의 응답값들 중 가장 가까운 예측값을 가진 응답자를 기증자로 정하여 무응답에 대한 대체를 실시하였다.

9.2.4 혼합 모형 대체

이 방법은 핫덱대체와 점수 대체를 혼합하여 사용하는 방법이다. 혼합 모형은 기본적으로 무응답에 대하여 핫덱대체를 실시하는 데 범주형 전환문장의 응답 정보에 따라 핫덱대체 또는 점수 대체를 섞어 사용하는 방법을 의미한다. 즉, 소득에 관한 범주형 전환문장에 대한 응답이 하한구간(bottom-open bracket; 예를 들면 소득 600만원 미만) 또는 막힌 구간(closed bracket; 예를 들면 소득 600-1200만원 사이)인 경우 동일한 구간 내의 응답값을 가지고 무응답에 대한 핫덱대체를 실시하고 상한구간(top-open bracket; 예를 들면 소득 12000만원 이상)이나 범주형 전환문장에 대하여 응답하지 않은 경우 점수 대체를 실시하였다. 이 자료는 응답여부 질문(ownership question)도 포함하는데 위의 네 가지 대체 방법이 문항에 따라 <표 9.1>과 같이 혼합되어 적용되었다. Cao(2001a)은 이 대체를 위해 사용된 IMPUTE 프로그램에 대한 사용법을 상세히 설명하고 있다.

<표 9.1> 여러 가지 대체 방법의 실제 적용

대체방법	응답여부 질문	하한 구간과 막힌 구간	상한구간과 범주형 전환문장 무응답
중위수 대체	핫덱	중위수	중위수
핫덱대체	핫덱	핫덱	핫덱
점수 대체	회귀	회귀	회귀
혼합 모형 대체	회귀	핫덱	회귀

9.3 미국 HRS 자료의 무응답이 대체된 변수들 및 대체 방법

HRS 자료에서 발생하는 무응답 대체 모형은 세 가지 특징을 가진다. 첫 번째는 변수에 따라 필요하다고 판단되면 9.2.4절에서 소개한 혼합 모형에 의하여 대체가 실시되었다는 점이고, 두 번째는 회귀모형이 적합되는 점수 대체나 혼합 모형 대체는 네 개의 인구학적 변수들인 성별, 연령, 교육수준, 그리고 결혼 상태를 항상 설명변수로 포함시켰다는 것이다. 세 번째는 부부간 동일한 정보를 가진 변수들에 대하여 결합된 기증자 풀(joint donor pool)을 형성하여 기증자의 숫자를 증가시킴으로써 대체의 신뢰성(reliability)을 추구하였다는 점이다.

Cao(2001b)는 위의 방법으로 대체된 변수들 및 각 변수별 대체 방법에 대한 자세한 설명을 포함하고 있다. 대체된 자료는 새로운 변수명으로 생성되어 무응답을 포함한 원래 변수와 구별되도록 하였다. 대체된 변수의 이름은 무응답을 포함한 변수의 이름에 "x"를 덧붙여 생성되었다. 예를 들면, 무응답을 포함한 원 변수의 이름이 Q1234인 경우 대체된 자료를 포함한 변수의 이름은 Q1234x로 주어져 원래 변수와 비교가 가능하도록 하였다.

< 참고문헌 >

이현정 (2009) *인구주택총조사 무응답 처리기법 연구*, 통계청

최필근 (2008) *인구주택 총조사 무응답 처리기법 연구*, 통계청.

An, H., and Little, R. J. A. (2008) "Robust model-based inference for incomplete data via penalized spline propensity prediction," *Communications in Statistics - Simulation and Computation*, 37, 1718-1731.

Baltagi, B. H. (1998) *Panel data methods in Handbook of Applied Economic Statistics*, 291-323. New York: Marcel Dekker.

Beckett, L. A., Brock, D. B., Scherr P. A. and Mendes de Leon, C.F. (1993) "Markov models for longitudinal data from complex samples." *Proceedings of the American Statistical Association, Section on Survey Research Methods,* 921-925.

Beckett, L. A., Brock, D. B., Lemke, J. H., Mendes de Leon, C. F., Guralnik, J. M., Fillenbaum, G. G., Branch, L. G., Wetle, T. T., and Evans, D. A. (1996) "Analysis of change in self-reported physical function among older persons in four population studies," *American Journal of Epidemiology*, 143, 766-778.Bell R. (1999) Depression PORT Methods Workshop (I). RAND: Santa Monica, CA.

Bell R. (1999) Depression PORT Methods Workshop (I). RAND: Santa Monica, CA.

Bellman, R. (1957) *Dynamic Programming*, Princeton University Press.

Bethlehem, J. G., and Keller, W. J. (1987) "Linear weighting of sample survey data," *Journal of Official Statistics*, 3, 141-153.

Breiman, L., Friedman, J. H., Olschen, R. A., and Stone, C. J. (1993) *Classification*

and Regression Trees. New York: Chapman & Hall

Cao, H. (2001a) *IMPUTE: A SAS Application System from Missing Value Imputations --- With Special Reference to HRS Income/Assets Imputations*, Institute for Social Research, University of Michigan: Ann Arbor.

Cao, H. (2001b) *HRS 1996 Imputations: Documentation*, Institute for Social Research, University of Michigan: Ann Arbor.

Cheng, P. E. (1994) "Nonparametric estimation of mean functionals with data missing at random," *Journal of the American Statistical Association*, 89, 81-87.

Collins, L. M., Schafer, J. L., Kam, C. M. (2001) "A comparison of inclusive and restrictive strategies in modern missing-data procedures," *Psychological Methods*, 6:330 -.351.

Davey, A., Shanahan, M. J., and Schafer, J. L. (2001) "Correcting for selective nonresponse in the National Longitudinal Survey of Youth using multiple imputation," *Journal of Human Resources*, 36, 500-519.

David, M., Little, R. J. A., Samuhel, M. E., and Triest, R. K. (1986) "Alternative Methods for CPS Income Imputation," *Journal of the American Statistical Association*, 81, 29-41.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977) "Maximum likelihood from incomplete data via the EM algorithm (with discussion)," *Journal of Royal Statistical Society, Series B*, 39, 1-38.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004) *Bayesian Data Analysis*, Chapman and Hall.

Geman, D. and Geman, S (1984) "Stochastic relaxation, Gibbs distributions, and the Bayesian reconstruction of images," *IEEE Transactions on Pattern*

Analysis and Machine Intelligence, 6, 721-741.

Hastie, T. J. and Tibshirani, R. J. (1990) *Generalized Additive Models*, Chapman and Hall.

Horvitz, D. G. and Thompson D. J. (1952) "A generalization of sampling without replacement from a finite population," *Journal of American Statistical Association*, 47, 663-685.

Kalton, G. and Flores-Cervantes, I. (2003) "Weighting methods," *Journal of Official Statistics*, 19, 81-97.

Kao, G. and Tienda, M. (1998) "Educational aspirations of minority youth," *American Journal of Education*, 106, 349-384.

Ireland, C. T., and Kullback, S. (1968) "Contingency tables with given marginals," *Biometrika*, 55, 179-188.

Lee, V. E., and Smith, J. B. (1995) "Effects of high school restructuring and size on early gains in achievement and engagement," *Sociology of Education*, 68, 241-270.

Little, R. J. A. (1988a) "A Test of Missing Completely at Random for Multivariate Data with Missing values," *Journal of the American Statistical Association*, 83, 1198-1202.

Little R. J. A. (1988b) "Missing data adjustments in large surveys," *Journal of Business and Economic Statistics*, 6, 287-301.

Little, R. J. A., and An, H. (2004) "Robust likelihood-based analysis of multivariate data with missing values," *Statistica Sinica*, 14, 949-968.

Little, R. J. A. and Rubin, D. B. (2002) *Statistical Analysis with Missing Data*, Wiley: New York.

Magidson, J. (1993) *SPSS for Windows CHAID Release 6.0*, Belmont, MA: Statistical

Innovations Inc.

Marker, D. A., Judkins, D. R., and Winglee, M. (2002) "Large-Scale Imputation for Complex Surveys," *Survey Nonresponse*, Groves, R. M., Dillman, D. A., Eltinge, J. L., and Little, R. A. J. (eds.), 329-341.

Oh, H. L., and Scheuren, F. S. (1983) "Weighting adjustment for unit nonresponse" in *Incomplete Data in Sample Surveys, Vol. II: Theory and Annotated Bibliography*, New York: Academic Press.

Raghunathan, T. E., Lepkowski, J. M., Hoewyk, J. V., and Solenberger, P. (2001) "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models," *Survey Methodology*, 85, 85-95.

Raghunathan, T. E., Solenberger, P., and Hoewyk, J. V., and (2002) *IVEware: Imputation and Variance Estimation Software User Guide*, Survey Research Center, Institute for Social Research, University of Michigan, available at http://www.isr.umich.edu/src/smp/ive/.

Rizzo, L., Kalton, G., and Brick J. M. (1996) "A comparison of some weighting adjustment methods for panel nonresponse," *Survey Methodology*, 22, 44-53.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995) "analysis of semiparametric regression models for repeated outcomes in the presence of missing data," *Journal of American Statistical Associatio*n, 90, 106-121.

Rojewski, J. W. and Yang, B. (1998) "Longitudinal analysis of select influences on adolescents' occupational aspirations," *Journal of Vocational Behavior*, 51, 375-410.

Rosenbaum, P. R., and Rubin, D. B. (1983) "The central role of the propensity scores in observational studies for causal effects," *Biometrika*, 70, 41-55.

Rosenbaum, P. R. and Rubin, D. B. (1984) "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score," *Journal of American Statistical Association*, 516-524.

Rosenbaum, P. R., and Rubin, D. B. (1985) "Constructing a control group using multivariate matched sampling incorporating the propensity score", *Annals of Statistics*, 21, 136-141.

Rubin D. B. (1978) Multiple imputation in sample surveys, *Proceedings in Survey Research Methodology, American Statistical Association,* 20-34.

Rubin D. B. (1987a) *Multiple Imputation for Nonresponse in Surveys*, John Wiley: New York.

Rubin, D. B. (1987b) "A Noniterative Sampling/Importance Resampling Alternative to the Data Augmentation Algorithm for Creating a Few Imputations When Fractions of Missing Information are Modest: The SIR-algorithm," A discussion of Tanner and Wong's "The Calculation of Posterior Distributions by Data Augmentation," *Journal of American Statistical Association*, 82, 543-546.

Schafer, J. L. (1997) *Analysis of Incomplete Multivariate Data*, Chapman and Hall.

Schafer, J. L. and Harel, O. (2002) "Multiple imputation in two stages," *ASA Proceedings of the Joint Statistical Meetings*, 1359-1363.

Statistics Netherlands (1998) Integration of Household Surveys: Design, Advantages, Methods, Netherlands Official Statistics, Vol. 13, Special Issue, Statistics Netherlands, Voorburg, The Netherlands.

Tanner, M. A. and Wong, W. H. (1987) "The Calculation of Posterior Distributions by Data Augmentation (with discussion)," *Journal of the American Statistical Association*, 82, 528-550.

Tang, L., Song. J, Belin, T. R. and Unützer, J. (2004) “A Comparison of Imputation Methods in a Longitudinal Randomized Clinical Trial,” *Statistics in Medicine*, 24, 2111-2128.

Verbeek, M. and Nijiman, T. E. (1992) "Testing for selectivity bias in panel data models," *International Economic Review*, 33, 681-703.

저자 소개

송주원

고려대학교 통계학과 졸업

고려대학교 대학원 통계학과(이학석사)

미 UCLA 대학원 의학통계학과(통계학박사)

미 University of Texas MD Anderson Cancer Center 의학통계학과 조교수

현 고려대학교 통계학과 부교수

e-mail: jsong@korea.ac.kr

안형진

고려대학교 통계학과 졸업

미 University of Chicago 통계학과(이학석사)

미 University of Michigan - Ann Arbor 대학원 의학통계학과(통계학박사)

미 University of Iowa 의학통계학과 조교수

현 고려대학교 의학통계학과 부교수

e-mail: hyonggin@korea.ac.kr

무응답 자료 처리 및 분석

저자	송주원 • 안형진
발행인	변효섭
기획	허남거 • 김정란 • 강태경
펴낸곳	통계교육원
주소	대전광역시 서구 월평 2동 282-1번지 통계센터 5F
전화	042-366-6232
팩스	042-366-6499
이메일	stimaster@korea.kr
홈페이지	http://sti.kostat.go.kr/
등록번호	통계교재 2009-6
발간번호	11-1240162-000016-01
발행일	2009년 8월 30일

ISBN 978-89-5801-186-6 93310