통계전문 교육프로그램 개발

- 무응답 자료처리 관련 -

2009. 8. 30.

고려대학교

연구책임자 | 송 주 원

공동연구자 | 안 형 진

통 계 교 육 원

통계전문 교육프로그램 개발

- 무응답 자료처리 관련 -

2009. 8. 30.

통 계 교 육 원

제 출 문

통계교육원장 귀하

본 보고서를 “무응답 자료처리를 위한 통계전문 교육프로그램 개발”에 관한 최종보고서로 제출합니다.

2009년 8월

고려대학교 산학협력단장

■ 연구기관·연구진 ■

◎ 책임연구원: 송 주 원(고려대학교 통계학과 부교수)

◎ 공동연구원: 안 형 진(고려대학교 의학통계학과 부교수)

목 차

표 목 차

그 림 목 차

제 1장. 서론

1.1 연구의 필요성과 목적

본 연구는 「통계교육원 무응답 자료처리를 위한 통계전문 교육프로그램 개발」 연구로 무응답 자료를 처리하기 위한 통계분석 기법을 설명하는 교재를 개발하고 통계전문가를 대상으로 한 교육 프로그램을 디자인하여 무응답 자료를 처리할 수 있는 전문가를 양성 및 교육하는 것을 목적으로 한다.

이와 같은 연구를 추진하게 된 배경 및 목적은 다음과 같다.

○ 우리가 직면하는 많은 자료는 여러 가지 원인으로 인하여 자료의 일부가 관찰되지 않는 무응답을 포함한다. 실험 자료에서도 무응답이 발생하지만 사회조사 자료에서는 응답 거부, 자료에 대한 추적 불가, 중도 이탈 등 다양한 원인으로 무응답의 비율이 더 높은 것이 일반적이다. 과거에는 자료의 분석을 실시할 때 무응답을 포함한 개체를 무시한 채 분석을 실시했는데 이런 단순한 분석은 편향된 결과를 줄 수 있다는 것이 알려져 왔다. 따라서 무응답을 포함한 자료에 대한 보다 정확한 분석의 필요성이 절실하게 대두되었고 미국 및 유럽 여러 나라에서는 이 분야에 대한 학술적 발전 뿐 아니라 대규모 사회조사 자료에 대하여 적절하게 무응답을 적용하는 일이 일반화되었으나 한국에서는 무응답 처리에 대한 연구 및 실제 자료에 대한 적용이 아직 활발하게 이루어지지 않고 있다. 이는 무응답 자료의 적절한 처리에 전문성이 필요하지만 이를 시행할 수 있는 통계전문가의 수가 한국에 그리 많지 않기 때문이다. 그러므로 무응답 자료처리를 위한 이론적 기초 및 적용 방법을 상세하게 설명할 수 있는 교재 및 전문 교육 프로그램의 개발은 매우 시급하다고 볼 수 있다.

○ 자료에서 발생하는 무응답을 무시한 채 통계 분석을 실시한다면 부정확한 결과를 얻을 수 있다. 무응답을 포함한 자료에 대한 연구 및 실제 자료에 대한 적용이 외국에서는 일반화되고 있으나 국내에서의 적용은 아직 미미하다고 볼 수 있다. 이는 한국어로 된 교재가 없고 무응답 처리를 담당할 수 있는 통계전문가의 수가 부족한 것이 주요 원인이다. 따라서 본 연구에서는 무응답을 포함한 자료의 분석을 위한 이론적 기초 및 대규모 자료에 이론을 실제로 적용하는 방법들을 설명하고 실습을 통해 무응답 자료 분석 전문가를 양성할 수 있는 교육 프로그램을 개발하였다. 이를 위하여 본 연구는 무응답 자료 분석의 이론 및 사례 연구를 포함한 교재를 개발하고 개발된 교재에 기초하여 통계청 통계교육원에서 실시할 수 있는 교육 프로그램을 제안하였다. 이 교육프로그램을 통하여 국내에 우수한 무응답 처리 전문가가 양성된다면 실제 자료에서 발생하는 무응답이 적절히 처리되고 정확한 결과를 얻을 수 있을 것으로 기대된다.

○ 한국에 무응답 자료를 적절히 처리할 수 있는 전문가가 부족한 주요 원인은 적절한 무응답 처리가 이론적 기초 및 실제 사례에의 적용 경험 모두를 필요로 하는데 반하여 이에 관련한 한국어로 된 교재조차 존재하지 않는 것이다. 이 분야에 관심이 있는 연구자들은 영어로 된 교재를 값비싼 가격으로 구입하여 독학해야 하는 것이 현실이므로 한국어로 된 교재 및 교육 프로그램은 더 많은 연구자들이 이 분야에 관심을 갖고 적절한 무응답 처리 및 분석이 일반화되는데 필수적 요소로 작용할 것이다.

1.2 연구의 내용

본 연구에서는 통계전문가를 대상으로 무응답을 포함한 자료의 분석에 관한 이론 및 적용 방법을 설명하고 실제자료에서 발생한 무응답에 대한 처리 사례를 기술한 한국어 교재를 개발하고자 한다. 또한, 개발된 교재에 근거하여 국내실정에 적합한 통계교육원의 전문 교육프로그램을 제안하고자 한다. 여기서, 통계전문가란 충분한 통계적 지식을 지닌 연구자를 의미하는데 통계학 또는 관련 분야 석사학위 소지자로서 조사연구 분야의 일정 경험이 있는 연구자나 통계학 또는 관련 분야의 박사학위 소지자 또는 이에 준하는 통계 분석 능력을 지닌 연구자를 의미한다. 또한, 무응답 자료의 분석을 위하여 통계프로그램이 사용되어야 하고 많은 무응답 자료의 분석을 위한 프로그램이 SAS를 사용하여 개발되어 왔으므로 SAS 프로그램을 사용할 수 있는 연구자를 대상으로 교재 및 교육 프로그램을 개발하고자 한다. 교재의 수준은 통계학 또는 관련 분야 석사 학위 소지자 또는 이에 준하는 연구자를 대상으로 하므로 수식을 통한 통계적 이론 설명을 포함하고 있으며 사례 연구에서는 통계적 내용 뿐 아니라 관련된 실제 조사연구의 내용을 포함하고 있다.

○ 교재 개발

교재는 무응답 자료 분석을 위한 이론 및 사례 연구의 두 가지 파트로 나누어 진행되었다. 이론 부분에서는 각 분야별로 무응답과 관련된 이론에 대한 설명 및 이를 적용하기 위한 통계 프로그램의 사용 방법을 설명하였다. 사례연구에서는 국내외의 대규모 사회조사에서 발생한 무응답을 처리하기 위하여 고려된 방법들을 소개하여 연구자들이 무응답 대체 방법을 적용할 때 고려해야 하는 사항 및 문제에 대한 해결책을 제시하고자 한다.

○ 교육 프로그램

무응답 대체에 관한 교육 프로그램은 선진 외국의 교육 프로그램을 벤치마킹하여 장단점을 분석한 후 개발된 교재에 근거하여 국내실정에 적합하게 수정 및 보완하여 제안되었다.

1.3 연구의 방법 및 절차

교재 개발은 무응답 자료 분석을 위한 이론 및 사례 연구의 두 가지 파트로 나누어 진행되었다. 이론 부분에서는 각 분야별로 무응답과 관련된 이론에 대한 설명 및 이를 적용하기 위한 통계 프로그램의 사용 방법을 설명하였다. 사례연구에서는 국내외의 대규모 사회조사에서 발생한 무응답을 처리하기 위하여 고려된 방법들을 소개하여 연구자들이 무응답 대체 방법을 적용할 때 고려해야 하는 사항 및 문제에 대한 해결책을 제시하고자 하였다.

교재 개발을 위하여 국내외 무응답 관련 교재 및 연구 논문들을 비교, 검토하는 작업이 우선적으로 진행되었다. 무응답 관련 연구논문 및 해외에서 발간된 무응답 관련서적을 검토하여 관련 이론이 충실하게 포함되도록 하였다. 또한 사례연구를 위하여 관련된 보고서와 연구논문들을 검토하고 사용된 자료를 구입하거나 홈페이지를 통하여 다운로드하여 분석을 진행하여 충실을 기하였다.

교육프로그램의 개발을 위하여 선진 외국의 교육프로그램들을 체계적으로 조사하고 프로그램의 대상, 내용, 구성 등을 알아보았다. 수집된 자료는 한국의 실정에 맞도록 보완되어 교육프로그램 개발에 반영되었다.

연구의 상세 내용은 다음과 같다.

1.3.1 교재의 개발

<Part-I> 이론 및 실습
무응답 분석을 위한 기초 이론을 서술한다. 본 연구에서는 실제 무응답 자료의 분석에 가장 많이 적용되는 무응답 대체(imputation) 방법에 중점을 두고 서술되었다. 포함된 세부 내용은 다음과 같다.

(1) 무응답이 있는 자료의 분석 이론과 관련된 정의 및 기본 가정 기술
자료에서의 무응답의 의미에 관하여 정의하고 무응답의 발생과 관련된 용어를 기술하고 분석을 실시하기 위한 기본 가정을 설명한다.

(2) 무응답을 포함한 자료에 대한 분석 기법 설명
무응답을 포함한 자료에 대한 여러 가지 분석 기법들을 분류하여 설명한다. 우선 자료가 무응답을 포함한 경우 흔히 사용되는 단순한 처리 방법들에 관하여 설명하고 각 방법의 장단점을 논의한다. 자료가 무응답을 포함하는 경우 흔히 사용되는 대체 방법에 대하여 설명하고 단일대체(single imputation) 및 다중대체(multiple imputation)에 대하여 비교 설명한다. 또한 가중치 접근 방법을 통하여 결측이 없는 자료 하에서의 모수에 대한 추정을 가능하게 하는 방법들의 원리 및 장단점을 논의한다. 마지막으로 통계 모형에 근거한 모수 추론 방법에 대하여 설명하고 장단점을 논의한다.

(3) 무응답을 포함한 자료에 대한 대체 방법
무응답을 포함한 자료에 대하여 최근에 많이 쓰이는 여러 가지 대체 방법에 대하

여 소개하고 통계 프로그램을 사용하여 각 대체를 시행하는 방법을 설명한다. 이는 모수적 모형에 근거한 대체 방법, 핫덱 대체 방법, 그리고 모수적 모형과 핫덱 대체의 혼합형인 준모수적 대체 방법에 대하여 소개하고 이 방법을 사용하여 대체를 실시할 수 있는 프로그램의 사용에 관하여 논의한다. 또한 다중 대체를 실시한 자료에 대한 분석 방법을 설명한다.

(4) 무응답을 포함한 패널 자료에 대한 분석 기법

앞에서는 횡단면 연구 (cross-sectional study)에 적용되는 대체법을 다루었는데 여기에서는 종단면 연구(longitudinal study)인 패널 자료(panel data)에 대하여 대체를 실시하는 방법을 소개한다.

(5) 연습 문제

무응답에서 발생하는 결측에 대하여 이해하고 실제 무응답을 포함한 자료에 대하여 분석을 실시할 수 있도록 연습문제를 통하여 훈련을 실시하고자 한다.

<Part-II> 사례연구

무응답을 포함한 자료에 대한 분석을 실시한 국내외 사례들을 연구하고 교재에 포함하여 실제 무응답 자료에 대한 대체 방법의 적용을 용이하게 하고자 한다.

(1) 사례 1. 인구주택 총조사 자료에 대한 무응답 대체 기법

- 연구자 : 최필근, 이현정
- 기고: 통계청 보고서 (이현정, 2009)
- 내용

  통계청에서 조사하는 인구주택 총조사는 1995년, 2000년, 2005년에 실시되었다.

한국 인구 규모, 인구 분포 및 구조와 주택에 관한 여러 가지 특성을 파악하여 각종 정책 입안의 기초자료가 되는 이 자료는 또한 각종 가구관련 경상조사의 기초 자료로 사용되어 왔다. 이 자료는 인구에 관한 항목 중 주요 항목(demographic information)에 대하여 전수조사를 실시하였고 거주 사항, 경제활동 사항 등은 표본에 대한 조사만을 실시하였다. 또한, 가구 및 주택에 관한 항목도 전수조사 항목과 표본에 대한 조사 항목으로 나뉘어 진다. 이 연구가 각종 정책 및 연구의 기초 자료로 사용되므로 결측을 무시하고 분석이 실시된 경우 편의된(biased) 결과를 줄 수 있다. 이에 본 연구진은 2005년 인구주택총조사 자료에서 발생하는 무응답에 대한 대체를 실시하는 방법을 소개함으로써 국가통계에서 발생하는 무응답에 대한 처리의 예시를 제공하고자 한다. 대체를 실시하기 위하여 2005년 자료 뿐 아니라 1995년과 2000년 인구주택 총조사 자료를 포함하여 대체를 실시함으로써 대체 모형의 정확도를 높이는 방법도 고려할 것이다. 또한, 이 대체에 사용할 수 있는 프로그램을 소개하고 적용하는 방법을 설명한다.

(2) 사례 2. 네덜란드 POLS 조사연구

- 연구자 : Bethlehem, J. G.
- 기고: Survey Nonresponse (Bethlehem, 2002)
- 내용

  1995년 이후 네덜란드 통계사무국은 POLS(Permanent Onderzoek Leefsituatie: 네덜란드 어)라고 불리는 하나의 통합된 사회조사 시스템을 개발하여 왔다. POLS는 매달 표본이 추출되며 목표 모집단은 12세 이상의 네덜란드 국민이다. 표본추출은 두 개의 스테이지로 구성된다. 첫 번째 스테이지에서 몇 개의 큰 지역을 층으로 하여 주민 수에 비례하는 추출확률로 지방자치구역을 추출한다. 두 번째 스테이지에서 추출된 각 지방자치구역에서 등확률(equal probability)로

개인표본을 추출한다. 이 사례 연구에서는 자원봉사에 관련된 한 변수에 관하여 가중방법의 영향을 조사한다. 이 사례 연구에서는 단지 Statistical Yearbook of Statistics Netherlands로부터 입수할 수 있는 정보만을 이용하여 가중치를 계산하였고 이 결과는 가중보정을 하지 않은 경우의 추정치와 비교하였다.

(3) 사례 3. 고령화연구패널조사 제 1차 기본조사 데이터에서 발생하는 무응답에 대한 대체 기법

- 연구자 : 송주원, 이수영, 윤초롱, 윤라헬, 송경화, 김병원, 장지연, 이혜정
- 기고: 노동연구원 보고서 (송주원 외, 2007a)
- 내용

  노동연구원에서 실시하는 고령화연구패널조사는 2006년에 처음 1차 기본조사가 실시되었고 2년에 한 번씩 추적 조사(follow-up study)가 실시되는 패널 조사로서 현재 2008년 2차 조사가 시행되어 자료 정리 단계에 있다. 이 조사는 45세 이상인 10,254명의 일반가구 거주자에 대한 여러 가지 체계적인 정보를 구축하여 곧 다가올 고령화 사회에 대비하는 것을 목적으로 한다. 이 자료는 8개 부문으로 나누어 인구학적 배경 및 가족 관계, 건강과 고용 상태, 소득과 자산 상황, 그리고 삶의 만족도 등 심리적 상태까지 1310개의 다양한 조사 항목을 포함하고 있다. 이 연구에서는 고령화연구패널 1차 조사에서 발생하는 무응답에 대한 대체 방법을 설명한다. 대부분의 사회 조사 자료와 마찬가지로 이 자료도 결측을 포함하고 있는데 대부분의 변수에서는 5% 미만의 결측이 발생하고 있지만 일부 소득 및 자산과 관련된 변수에서 결측 비율은 10% - 30%에 이르고 있다. 또한 소득 및 자산 관련 문항에서 결측의 비율을 낮추기 위하여 범주형 전환문항(Unfolding Bracket)을 사용하였다. 이 문항들은 정확한 응답을 거부하는 응답자에 대하여 구간 정보만을 제공하는 것을 가능하게 함으로써

결측의 비율을 낮추고 부분 정보를 얻는 것이 가능하게 하였다. 하지만 일부 응답자의 응답이 정확한 응답 대신 부분 정보만으로 이루어져 있으므로 이를 통합하여 분석하는 데 어려움이 발생하게 된다. 이 연구에서는 고령화연구패널 1차 조사 자료 중 소득과 자산 영역 변수들에 중점을 두어 다중대체를 실시하는 방법을 제안하였다. 이를 위하여 준모수적 모형인 수정된 평균에 근거한 핫덱 대체(Little, 1988)를 범주형 전환 문항에 적용할 수 있도록 확장하여 적용하였다. 핫덱 대체를 실시하는 SAS Macro 프로그램을 적용하여 이 방법을 이용한 다중 대체를 실시하는 방법을 설명한다. 또한, 다중대체된 자료는 통합하여 분석을 실시하는 방법을 설명한다. (송주원 외, 2007b)

(4) 사례 4. 미국 The Health and Retirement Study (HRS) 자료에 대한 무응답 대체 기법

- 연구자: Honggao Cao
- 기고: HRS 보고서 (Cao, 2001a)
- 내용

  미국에서 실시되고 있는 고령화 패널 조사인 HRS 자료는 1992년부터 2년마다 50세 이상인 22,000명에 대하여 건강, 경제, 근로 사항 및 은퇴 계획 등에 대하여 추적 조사를 실시하고 있다. 이 자료에서 발생하는 무응답에 대하여 다중대체가 실시되었고 대체된 자료가 홈페이지에 제공되고 있다. 이 연구에서는 HRS 자료에 대하여 다중대체를 실시하는 방법을 소개하고 대체 결과를 요약하고 있다. 이 자료도 결측의 비율을 낮추기 위하여 소득 및 자산 주요 문항에 대하여 범주형 전환 문항을 사용하고 있는데 이 정보를 통합하고 결측값을 대체하기 위하여 비모수적인 대체 모형인 핫덱 대체를 실시하였다. 또한, 핫덱 대체를 실시할 때 결측인 자료에 대한 기증자(donor)가 부족한 경우 모수적 대체 모형인 회귀모형(regression model)을 이용하여 대체를 실시하는 혼합된 대

체 방법을 제안하여 대체를 실시할 때 발생하는 어려움을 극복하고 있다. 또한, 부부 모두의 정보가 제공된 항목에 대하여 커플로부터 제공된 정보를 대체 모형에 포함시킴으로써 최대한의 정보를 포함시켜 대체를 실시하는 방법을 사용하였다. 대체를 실시하기 위하여 SAS Macro IMPUTE를 개발하였고 이 프로그램에 대한 사용 방법은 Cao, H. (2001b)에 설명되어 있다.

1.3.2 교육 프로그램 개발

연구에서 객관적이고 과학적인 결론을 도출하기 위하여 통계분석법의 중요성은 지속적으로 증가하였다. 이에 비례하여 국내에서도 통계전문가뿐만 아니라 비전문가를 상대로 한 통계 교육프로그램의 개발도 증가하였으며 개발된 프로그램의 시행도 많이 이루어지고 있다. 하지만 대부분의 통계 교육 프로그램의 학습내용은 일반적인 자료 분석에 초점이 맞추어져 있었고 결측자료 또는 무응답 자료 분석에 관한 교육은 전무한 실정이다. 물론 최근 들어 통계관련 학회나 산업계가 주최하는 단회성 결측자료 분석교육 비정기적으로 열려왔으나 (예: SPSS Open House, 학회의 tutorial 등) 짧은 교육시간(2-3시간)으로 대부분 결측자료 분석법의 기초적인 소개에 그치는 경우가 많고 통계 소프트웨어를 이용한 실제 분석방법 적용은 시간 및 공간의 제약으로 인해 거의 불가능했다. 또한, 국내에 결측자료 분석을 전공한 전문가 또한 외국에 비해 그 수가 많지 않은 점도 결측자료 분석 교육 프로그램의 개발을 지연시킨 한 요소이다. 이에 반해, 통계교육이 발달한 선진 외국에서는 2-3일 간의 결측자료처리 방법에 관한 정기교육을 실시하는 기관이 꽤 많으며 각종 학회에서도 1-day course등의 tutorial을 통해 무응답자료 처리를 교육하고 있다. 예를 들면 Statistical Horizons라는 미국의 사설기관에서 정기적으로 결측자료 분석에 관한 Short Course를 미국 전역을 순회하며 개최하고

있다. 이 과정에는 University of Pennsylvania의 사회통계방법론 전문가인 Paul Allison교수가 2일 동안 결측자료 분석에 관한 방법론 및 SAS 등 통계소프트웨어를 이용한 실제사례분석을 강의하고 있다. 본 연구에서는 해외기관에서 실시하고 있는 무응답자료 분석에 관한 교육프로그램을 벤치마킹하여 장단점을 분석하고 본 연구를 통해 개발된 교재를 이용하여 국내실정에 맞는 교육프로그램을 제안한다.

1.4 연구의 효과

요즈음은 정보의 홍수 속에 산다고 해도 과언이 아닐 정도로 많은 정보들이 존재하고 이 정보들은 많은 경우 자료(data)의 형태로 모아지고 있다. 이 자료들로부터 객관적이고 과학적인 연구 결과를 도출하기 위하여 적절한 통계 분석의 중요성은 증대되어 왔다. 현실에서 다루게 되는 대부분의 자료가 무응답으로 인한 결측값을 포함하고 있으므로 외국에서는 무응답을 적절히 처리하여 분석의 정확성을 추구하는 노력들이 일반화되고 있다. 하지만 국내에서는 한국어로 된 무응답 처리에 관한 교재조차 전무한 상황에서 상당한 통계적 전문성 및 시간과 노력을 요하는 무응답 처리 방법은 일반 연구자들이 손쉽게 시행하기 어려운 단점이 있어 이 분야의 진보가 더디게 진행되어 왔다. 본 연구에서는 무응답 처리에 관한 한국어 교재를 집필하여 무응답 대체에 대한 이해를 향상시키고 이를 통한 통계자료의 수집 및 분석 수준에 향상을 도모하고자 한다. 본 연구에서 개발하는 교재는 단순히 무응답 대체에 관한 이론만을 기술하는 것이 아니라 사례 연구 및 적용 프로그램을 소개하고 연습 문제를 포함하여 연구자들이 무응답을 포함한 자료를 분석하는 데 필요한 적용 능력을 기를 수 있는 효과를 지닌다. 또한, 무응답 대체를 시행할 수 있는 통계전문가 집단을 양성할 수 있도록 무응답 대체를 위한

통계 교육프로그램을 제안함으로써 한국 내 자료 수집 및 분석의 수준을 향상시킬 수 있을 것으로 기대된다.

제 2장. 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.

2.1 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 관찰되지 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

암환자에 대한 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.

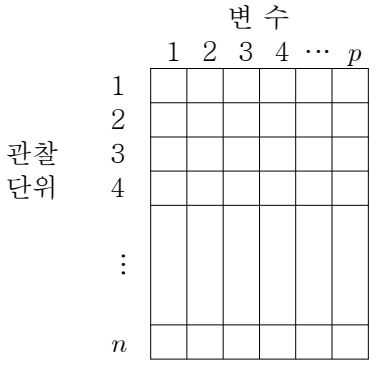
위의 예제 중 일부 무응답은 엄밀히 말하면 무응답이라 할 수 없다. 첫 번째로 일부 무응답은 실제로는 무응답이 아니라 어떤 특정한 값을 의미한다. 사회조사 자료의 예에서 흡연하지 않은 청소년의 흡연량은 무응답으로 남게 되는데 이는 흡연량이 0이기 때문에 질문 문항을 뛰어넘은 경우에 해당되므로 흡연량을 0으로 대입하여 분석하여야 한다. 이 때 무응답은 손쉽게 정확한 값을 알아 낼 수 있으므로 무응답이라 할 수 없다. 두 번째로 일부 무응답은 선택하도록 주어지지 않은 다른 항목에 대한 응답을 의미한다. 대통령 선거와 관련한 여론조사에서 후보자 중 누구에게 투표할 지 응답하지 않는 경우 중 선호하는 후보자가 없어 여론조사에 응답하지 않는 경우는 응답 문항에 선택할 문항이 없기 때문에 발생한다. 이 문항의 경우 응답할 수 있는 항목으로 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 응답

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

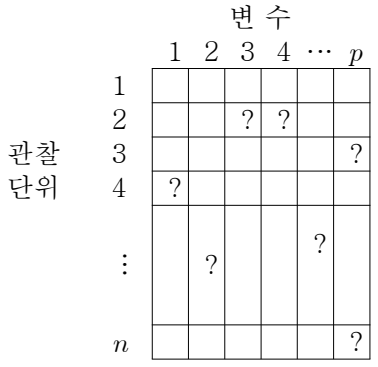
분석을 위하여 입력된 자료의 값들은 <그림 2.1(1)>과 같이 직사각형 형태로 행렬(matrix)을 이용하여 나타낸다. 이 자료행렬(data matrix)에서 각 행(row)은 관찰단위(observation unit)를 의미하고 각 열(column)은 변수(variable)를 의미한다. 무응답이 발생하지 않는다면 자료의 모든 원소(element)는 각 관찰단위에서의 각 변수에 대하여 응답된 값으로 채워지고 대부분의 통계분석 프로그램은 이와 같이 완전하게 관찰된 형태의 자료에 대한 분석을 시행하도록 개발되어 있다. 한편, 무응답이 발생한다면 이 직사각형의 모든 칸이 채워지지 못하고 응답되지 않은 값은 <그림 2.1(2)>와 같이 물음표를 사용하여 표현하는 것이 일반적이다. 이 때, 물음표로 표시하지 않은 칸은 응답된 값을 나타낸다.

<그림 2.1> 자료의 형태

(1) 무응답이 발생하지 않은 경우



(2) 무응답이 발생한 경우



직사각형의 행렬자료를 Y로 나타내자. 좀 더 자세히 n개의 행과 p개의 열을 가

지는 자료 $Y=(y_{ij})$, $i=1,...,n$, $j=1,...,p$로 표현할 수 있는데 여기서, i는 n개의 관찰값을 나타내고 j는 p개의 변수를 나타낸다. 무응답이 발생한다면 일부 원소인 y_{ij} 값은 관찰되지 않는다. 즉, 자료 Y값은 응답된 y_{ij} 원소들과 무응답인 y_{ij} 원소들로 나누어지는데 응답된 y_{ij}들을 관찰된(observed) Y라는 의미로 Y_{obs}로 나타내고 무응답인 y_{ij}들을 관찰되지 않고 빠진(missing) Y라는 의미로 Y_{mis}로 나타낸다. 이 때, 관찰단위에 따라 Y_{obs}와 Y_{mis}에 포함되는 변수들의 조합은 달라진다. 이를 명확하게 하기 위하여 관찰단위의 번호를 포함하여 $y_{i,obs}$와 $y_{i,mis}$, $i=1,...,n$, 로 나타내기도 한다. 예를 들어 <그림 2.1(2)>에서 첫 번째 관찰단위 $(i=1)$는 모두 응답되어 $y_{1,obs}$는 모든 변수를 포함하는데 반하여 $y_{1,mis}$는 존재하지 않는다. 두 번째 관찰단위$(i=2)$는 세 번째와 네 번째 변수에서 결측이 발생하므로 $y_{2,obs}$는 세 번째와 네 번째 변수에 대한 관찰값을 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 관찰단위마다 다른 길이의 벡터를 포함할 수 있다.

Y가 자료의 응답값을 나타내는 반면에 어느 관찰단위의 어느 변수에서 무응답이 발생하였는지를 나타내기 위하여 무응답 표시행렬(missing data indicator matrix) M을 사용한다. M은 Y와 동일하게 n개의 행과 p개의 열을 가지는 직사각형 행렬로서 $M=(m_{ij})$, $i=1,...,n$, $j=1,...,p$로 표현할 수 있는데 여기서, i는 n개의 관찰단위를 나타내고 j는 p개의 변수를 나타낸다. M은 무응답이 발생한 위치를 나타내므로 i번째 관찰단위의 j번째 변수에서 무응답이 발생하면 m_{ij}가 1의 값을 가지고 응답값이 존재한다면 m_{ij}가 0의 값을 가지도록 표현한다. 즉,

$$m_{ij}=\begin{cases} i\text{번째 관찰단위의 } j\text{번째 관찰값이 무응답이면 } 1 \\ i\text{번째 관찰단위의 } j\text{번째 관찰값이 응답이면 } 0 \end{cases}$$

와 같이 각 원소가 표시변수(indicator variable)로 표현될 수 있어 무응답 표시행렬이라 부른다. <그림 2.2>는 (1) 무응답을 포함한 가상의 자료행렬 Y와 (2) Y의 무응답 표시행렬 M의 예제이다.

<그림 2.2> 가상의 자료행렬 Y와 대응되는 무응답 표시행렬 M의 예제

(1) 자료행렬 Y

		변수					
		가구 번호	가구원 번호	성별	나이	…	교육 정도
관찰 단위	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 표시행렬 M

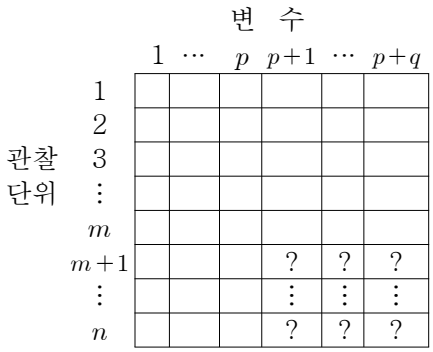
		변수					
		가구 번호	가구원 번호	성별	나이	…	교육 정도
관찰 단위	1	0	0	0	0		0
	2	0	0	0	1		1
	3	0	0	0	0		1
	4	0	0	1	0		1
	5	0	0	0	1		0
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	n	0	0	1	0		0

2.2 무응답 자료 패턴

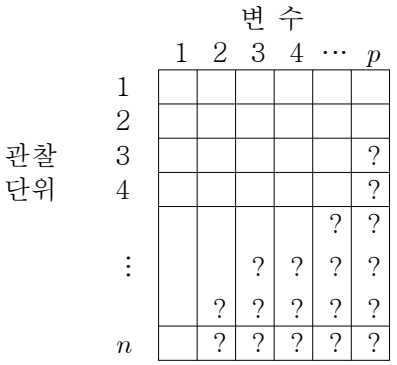
무응답의 발생 패턴은 다양하게 나타나는 데 대표적인 몇 가지 패턴이 <그림 2.3>에 나타난다(Little and Rubin, 2002). 일부 자료에서는 자료 자체가 정확히 이 패턴들을 따르지는 않지만 변수들 사이의 순서를 재정렬함으로써 이 패턴들로 표현이 가능하다. 각 패턴에 대한 자세한 설명은 다음과 같다.

<그림 2.3> 대표적인 무응답 발생 형태

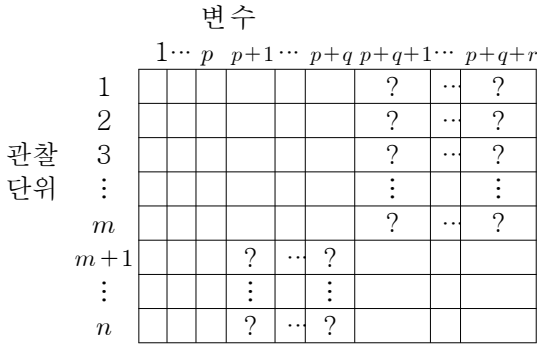
(1) 두 가지 패턴



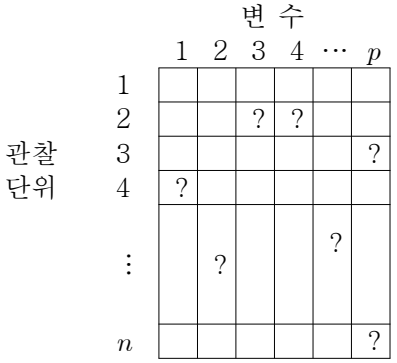
(2) 단조 패턴



(3) 자료 짝짓기



(4) 일반적인 패턴



(1) 두 가지 패턴(two pattern)

일부 관찰단위에 대하여 일부 변수에서만 무응답이 나타나는 경우를 의미한다. 즉, m개의 관찰단위에서는 모든 변수에 대하여 응답값이 존재하지만 $n-m$개의 관찰단위에서는 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 관찰단위(observation unit)에서 응답이 이루어지지만 두 번째 측정된 변수들에서부터 무응답이 발생하는데 한 번 무응답이 발생한 관찰단위에서는 이후 측정된 모든 변수에 대한 모든 응답값이 무응답으로 남게 된다. 즉, 일단 무응답이 발생한 관측단위의 추후 측정값은 모두 무응답으로 남게 되는 한편 이 후 변수들에서 무응답이 발생하는 관측단위가 지속적으로 추가되어 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 관찰단위에서 응답이 이루어지지만 이 후 시점에서 중도탈락이나 사망 등으로 인하여 관

찰되는 단위가 점차 줄어드는 경우에 흔히 나타난다. 엄밀하게 말하면 (1) 두 가지 패턴은 단조패턴의 가장 간단한 형태이다.

(3) 자료 짝짓기(file matching)

모든 관찰단위에서 기초 변수들에 대하여는 응답값이 모두 존재하지만 나머지 변수들에 대하여 응답되는 관찰단위가 서로 다르다. 즉, 처음 m개의 관찰단위에서는 처음 $p+q$개의 변수들에서의 응답값이 존재하는 반면 마지막 r개의 변수에 대한 값은 결측으로 되어 있고 뒤 $n-m$개의 관찰단위에서는 처음 p개의 변수 및 마지막 r개의 변수들에 대한 응답은 존재하지만 중간 q개의 변수들의 값은 무응답으로 나타난다. 더구나 처음 m개의 관찰값과 뒤 $n-m$개의 관찰값은 처음 p개의 공통변수들에 대하여 응답값이 모두 존재하지만 나머지 $q+r$ 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 어느 관찰단위의 어느 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 형태

를 요구하는 경우, 즉, 두 가지 패턴이나 단조 패턴의 무응답 형태를 보이는 경우 무응답을 대체할 때 손쉬운 방법을 사용하여 편의를 제거할 수 있지만 무응답 발생 형태에 대한 제약이 적은 일반패턴을 가지는 경우 더 복잡한 분석 기법이 요구된다. 자료 짝짓기 패턴을 가진 경우 동시에 측정되지 않는 변수들 사이의 연관성은 추정이 안 된다는 점을 유의하여야 한다.

2.3 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. 무응답 표시행렬 M의 자료행렬 Y에 대한 조건부 분포를 $f(M|Y,\phi)$, 여기서 ϕ는 M의 조건부 분포에 연관된 모수들(parameters),로 나타내면 완전임의결측은 모든 Y와 ϕ의 값에 대하여

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

로 표현할 수 있다. 즉, 무응답의 발생은 자료 Y의 값과 상관없이 발생한다는 것을 의미한다. 여기서 $Y=(Y_{obs},Y_{mis})$이므로 무응답의 발생은 관찰되지 않은 자료인 Y_{mis} 뿐 아니라 관찰된 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 분 밖

에 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 무응답 자료 메커니즘은 완전임의결측이라 할 수 있다.

(2) 임의결측(Missing At Random 또는 MAR)

자료행렬 Y에서 무응답이 발생할 확률은 자료의 관찰된 부분에는 연관되지만 자료의 관찰되지 않은 부분과는 연관이 없다. 즉, 무응답 표시행렬 M의 자료행렬 Y에 대한 조건부 분포 $f(M|Y,\phi)$는 모든 Y_{mis}와 ϕ의 값에 대하여

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

로 표현할 수 있다. 즉, 무응답의 발생은 자료 Y의 관찰된 값들에만 연관되고 관찰되지 않은 Y_{mis}와는 상관이 없다는 것을 의미한다. 예를 들어, 가계동향조사에서 조사한 가구별 수입에서 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 메커니즘은 임의결측을 따른다고 할 수 있다.

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

자료행렬 Y에서 무응답이 발생할 확률은 자료의 관찰된 부분인 Y_{obs} 뿐 아니라 관찰되지 않은 부분인 Y_{mis}와도 연관되어 있다. 즉, 무응답 표시행렬 M의 자료행

렬 Y에 대한 조건부 분포 $f(M|Y,\phi)$는

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

로 표현할 수 있다. 예를 들어, 가계동향조사에서 조사한 가구별 수입에서 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 무응답 표시행렬 M로 나타내므로 무응답 자료에 대한 분석은 이 두 개의 행렬의 결합분포(joint distribution)를 통해 시행해야 한다. 자료행렬 Y와 관련된 모수들을 θ로 표현하면 이 결합분포는

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

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 무응답 표시행렬 M에 대한 모형인 $f(M|Y,\phi)$도 포함해야 하는 것이다. 문제는 무응답을 포함한 자료에서 자료행렬 Y 전체가 아닌 측정된 부분인 Y_{obs}만을 실제로 측정할 수 있다는 점에서 더 복잡하다. 즉, 실제로 관찰된 자료에 의

한 결합분포는

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

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

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

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

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

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

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

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 $f(M|,\phi)$ 부분은 연관되어 있지 않으므로 $f(Y_{obs}|\theta)$만에 근거하여 분석을 실시할 수 있다. 즉, θ에 대한 추론이 $f(M|,\phi)$을 무시하고 시행할 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 무응답 표시행렬과 관련된 모수 ϕ가 서로 별개(distinct)라면 무응답 자료 메커니즘에 관한 $f(M|Y_{obs},\phi)$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 관찰된 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

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.

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 랜덤하게 발생하였으므로 응답자 집단과 무응답자 집단은 완전히 랜덤하게 나뉘어져 두 집단 사이의 분포가 동일해야 한다는 점에서 직관적이다. 물론 실제 자료에서는 관심인 변수가 소득 한 개가 아니라 여러 개로 구성되고 무응답 발생 형태도 <그림 2.3.(4)> 일반적인 형태를 따르는 경우가 많고 각 변수별 응답자 집단과 무응답자 집단의 직접적 비교는 너무 많은 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 무응답 표시행렬과 관련된 모수 ϕ가 서로 별개(distinct)라고 가정한 후 무응답 발생원인은 무시할 수 있는 무응답 자료 메커니즘이라 가정하여 무응답 표시행렬 M에 대한 모형을 포함하지 않고 관찰된 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, 지

출 등의 정보가 알려져 있다면 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.

<2장 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 3장. 여러 가지 무응답 분석 방법

<학습목표>

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

(2) 여러 가지 가중치 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.

3.1 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.

3.2 가중치 보정방법 (Weighting Adjustment)

이 방법은 완전히 응답한 개체에 가중치를 주어 편향을 보정하는 방법이다. 기본적인 아이디어는 유한한 모집단을 대상으로 한 조사(finite population survey)의 확률화 추론 (randomization inference)을 위해 가중치를 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 $Y_1, Y_2, ..., Y_N$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 가중치를 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 추정치를 호빗-톰슨 추정치(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 가중치 $w_i = n\pi_i^{-1} / \sum_{k=1}^{n} \pi_k^{-1}$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.

3.2.1 평균의 가중 클래스 추정법

먼저 표본을 응답자와 무응답자 모두 이용가능한 변수를 바탕으로 J개의 가중 클

래스로 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 가중치를 가지게 된다.

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

이 때 클래스 j 안의 개체 i의 응답확률 추정치 $\widehat{\phi_i}=r_j/n_j$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 추정치는 비편향성을 갖는다. Oh와 Scheuren (1983)은 위의 추정치의 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.

3.2.2 응답성향을 이용한 가중치 방법

X 를 응답자와 무응답자 모두 이용가능한 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 가중치가 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 방법(propensity score)은 두 군을 비교하는 관찰연구에서 다변량 교란변수들을 단변량 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

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

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 추정치를 구할 수 있다. 만일 X가 단변수라면 응답성향을 이용한 방법은 X를 이용한 가중 클래스 추정방법과 동일하다.

응답성향을 이용한 가중 방법에서 클래스를 이용하는 대신 응답개체 i의 가중치

를 추정된 응답성향점수의 역수인 $\hat{p}(x_i)^{-1}$로 직접 가중치를 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 가중치를 가지게 되며 이는 평균과 총합의 추정치에 과도한 영향을 미치게 되며 그 결과로 극도로 높은 분산을 가진 추정치를 제공할 수도 있다. 또한 응답성향점수의 역수인 $\hat{p}(x_i)^{-1}$로 직접 가중치를 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)

3.2.3 무응답 가중치 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.

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

가중클래스 방법을 이용하여 모수의 추정치를 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.

3.2.4.1 사후-층화 (Post-stratification)

모집단 비율 N_j/N가 외부의 출처로부터 알려져 있다고 가정하자. 이런 경우 가중 클래스를 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 $\bar{Y}$의 불편추정치이고 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 추정치는 위 식에서 모집단 분산인 S_{jR}^2를 클래스 j의 응답자들의 표본분산 s_{jR}^2로 대치하여 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.

3.2.4.2 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. $X_1 = j$, $X_2 = l$ $(j = 1,...,J,$ $l = 1,...,L)$을 가진 클래스 안에서 N_{jl} 모집단 개체 가운데 n_{jl} 개체가 표본추출

되었고 이 중에 r_{jl} 표본개체에서 Y변수의 값이 조사되었고 $(n_{jl}-r_{jl})$ 개체에서 무응답이 발생했다. 이 경우 사후층화 추정치와 가중 클래스 추정치는 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 추정치는 셀 안의 응답자 평균에 근거하여 구할 수 있다. X_1=성별 이고 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 추정치인 $\{N_{jl}^{*}\}$를 구하는 방법이다. 이 때, 레이킹 방법은 다음과 같은 주변 제약(marginal constraints)을 만족하여야 한다.

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

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

$\{N_{jl}^*\}$는 관찰된 클래스 계수(class counts)인 $\{n_{jl}\}$와는 다르다. 즉, 행의 상수 $\{a_j, j=1,...,J\}$ 와 열의 상수 $\{b_l, l=1,...,L\}$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 추정치를 주변 합 $\{N_{j+}\}$와 $\{N_{+l}\}$에 일치시키기 위하여 행 과 열 변수로 스케일링하게 된다. 즉, 첫 번째 단계에서 추정치는 다음처럼 행 주변 합 $\{N_{j+}\}$과 일치하여 구한다.

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

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

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

그리고 추정치는 다시 다음과 같이 수렴할 때 까지 계속해서 갱신한다.

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

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

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

추정치는 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 추정치는 정의되지 않는다. 이 경우는 그 클래스의 평균의 다른 추정치를 고려하여야 한다.

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

3.2.5.1 일반 회귀 추정

무응답이 없을 때, 적절한 보조 정보 (auxiliary information)를 이용하면 단순 표본 평균의 정밀도는 향상된다. p개의 보조변수가 있다고 가정하여 보자. i번째 표본의 보조변수의 값을 $X_{k1}, X_{k2}, ..., X_{kp}$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 $E_k = Y_k - X_k B$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 추정치를 이용한 일반 회귀 추정값은 다음과

같다.

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

무응답이 있는 경우에 수정 일반 회귀 추정값은

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

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

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

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 추정치의 편향은 0이다. 그러므로 무응답이 회귀계수에 영향을 미치지 않는 이상 회귀 추정치는 불편 추정치이다. 즉, 무응답이 메커니즘이 MAR이고 주변수와 관련성이 높은 보조변수가 회귀 모형에 포함된 경우 편향은 매우 작다.

3.2.5.2 범주형 보조변수를 이용한 선형 가중방법

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

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

여기서 응답자 i의 가중치는 $w_i = \nu^T X_i$ 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.

3.2.6 무응답 가중 추정치의 추론

가중 추정치는 일반적으로 계산하기가 쉬운 경우가 많다. 하지만 추정치의 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 추정치의 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 추정치의 편향을 줄여 주므로 비교적 쉬운 방법이다. 하지만 이 방법은 부분 정보를 가지고 있는 무응답 개체들을 분석에서 제외

하므로 추정치의 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.

3.3 이용가능한 개체 분석 (Available-case Analysis)

완전히 응답한 자료를 이용한 분석은 평균 또는 주변빈도분포의 추정 등 단변량 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 이용가능한 개체 분석법이 있다.

이 방법은 각 분석 단계에서 이용가능한 모든 자료를 사용한다. 완전히 이용가능한 자료 분석방법보다 더 많은 자료를 이용하므로 언 듯 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 일반패턴을 가진 결측이 있는 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 추정치는 쉽게 구할 수 있으나 그 추정치의 표본오차는 대표본적으로도 매우 복잡한 경우가 많다.

3.4 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 어떤 다른 값으로 채우는 것을 대체방법 (imputation methods)이라고 한다. 만일 무응답값을 하나의 값으로 채우는 방법을 단일 대체 (single imputation)라고 하고 대체로 발생하는 불확실성(uncertainty)을 고려하기 위해 여러 개의 값으로 채우는 방법을 다중 대체 (multiple imputation)라고 한다.

결측값을 채우고 나면 사용자들은 완전한 자료를 가졌다고 생각할 수 있으므로 대체방법은 매력적이나 잘못된 모형으로부터 무응답을 대체하게 되면 완전히 이용가능한 방법보다 더 큰 편향을 발생시킬 수도 있다.

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 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 회귀식으

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

함축적인 모형 방법 중 몇 가지는 다음과 같다.
(a) 핫덱 대체: 무응답을 "비슷한" 성향을 가진 응답자의 자료로 대체하는 방법이다. 핫덱방법은 표본조사에서 흔히 사용된다.
(b) 콜드덱 대체: 핫덱과 비슷하나 대체할 자료를 현재 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.

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

결측 자료의 분석에서 많은 경우 구체적인 모형을 가정하고 그 모형하의 우도함수를 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.

3.5.1 무응답이 없는 경우의 최대우도 추정방법 리뷰

Y를 데이터라고 하자. 이때 Y는 스칼라, 벡터, 또는 행렬이 될 수 있다. 또 데이

터는 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 3.5.1: *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 3.5.1: 단변량 정규 표본

n개의 독립적인 표본 $Y=(y_1,...,y_t)^T$가 평균 μ와 분산 σ^2을 갖는 정규분포로부터 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 3.5.2: 다변량 정규 표본

$Y=(y_{ij}), i=1,...,n, j=1,...,K$는 평균벡터 $\mu=(\mu_1,...,\mu_K)^T$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.

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

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

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

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 3.5.2: *모수 θ의 MLE는 우도함수 $L(\theta \mid Y)$ 또는 로그우도함수 $l(\theta \mid Y)$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.

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

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 3.5.3: 단변량 정규 표본

n개의 독립 정규표본의 로그우도함수는 다음과 같다.

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

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

예제 3.5.4: 다변량 정규 표본

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

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

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

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

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

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

μ와 Σ에 관하여 위의 로그우도함수를 최대화 하면 MLE는

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

이다. 여기서 $\overline{y} = (\overline{y_1}, ..., \overline{y_K})$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 3.5.5: 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 $y_i = (y_{i1}, y_{i2}), i = 1, ..., n$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

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

이고

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

이다. 여기서 $SSE = \sum_{i=1}^{n} \left[y_{i2} - \overline{y_2} - \widehat{\beta_{21.1}}(y_{i1} - \overline{y_1}) \right]^2$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 정보 행렬 (information matrix)라고 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 $D_g(\hat{\theta})\,CD_g(\hat{\theta})^T$와 형태를 갖는 정규분포로 근사한다.

즉,

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

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

3.5.2 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 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 성가신 모수(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 이용하는 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

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

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

예제 3.5.6: 무응답이 있는 지수 표본

무응답이 있는 단변량 지수 표본을 고려하자. 이 때, $Y_{obs} = (y_1, ..., y_r)^T$는 응답된 자료이고 $Y_{mis} = (y_{r+1}, ..., y_n)^T$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 $R=(R_1,...,R_n)^T$라고 하자. 여기서 $R_i=1, i=1,...,r$ 이고 $R_i=0, i=r+1,...,n$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

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

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

$$\frac{1}{\theta}{}^{i}\Big)\exp\Big(-\frac{(n-r)c}{\theta}\Big).$$

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.

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

3.5.3 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 $\phi_1, \phi_2, ..., \phi_J$ 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 $\phi_1, \phi_2, ..., \phi_J$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

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

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 3.5.7: 이변량 정규분포 자료에서 한 변수에만 무응답이 있는 자료의 ML 분석법

이변량 정규(bivariate normal) 자료에서 r개체에서는 Y_1과 Y_2 둘 다 응답이 있고

$(n-r)$에서는 Y_1만 응답이 있다고 하자. 즉, $\{y_i=(y_{i1},y_{i2}), i=1,...,r\}$이고 $\{y_{i1}, i=r+1,...,n\}$이다. 이 자료의 로그우도함수는

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

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 우도식의 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.

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

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

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

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

이다. 즉, n개의 표본 $\{y_{11}, y_{21}, ..., y_{n1}\}$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 추정치는 MLE의 불변의 속성을 이용하여 구할 수 있다. 즉,

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

이다. $\widehat{\mu_2}$의 식에서 $\overline{y_2}$ 부분은 (n-r)개의 무응답 표본은 버리고 r개의 응답한 표본을 이용하여 구한다. 그리고 $\hat{\beta}_{21.1}(\hat{\mu}_1 - \bar{y}_1)$ 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 $\hat{y}_{i2} = \bar{y}_2 + \hat{\beta}_{21.1}(y_{i1} - \bar{y}_1)$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.

3.5.4 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 추정치를 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.

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

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 추정치라고 하자. 예를 들면 이 초기값은 완전히 이용가능한 자료로부터 추정할 수 있다. 이제 $\theta^{(t)}$를 t번째 반복. 의 추정치라고 하자. 뉴튼-랩슨(Newton-Raphson) 알고리듬은 다음의 식으로 정의된다.

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

여기서 $I(\theta|Y_{obs})$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,

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

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

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

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 쉬

운 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.

3.5.5 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 추정치를 이용하여 무응답 자료를 추정한다. 좀 더 엄밀하게 말하면 응답된 자료와 현재 반복에서의 모수 추정치를 이용하여 완전한 자료의 기대 로그우도를 구한다. M-단계에서는 위에서 구한 완전한 자료의 기대 로그우도를 최대화 하는 모수 추정치를 갱신한다. 이제 E-단계와 M-단계를 모수의 추정치가 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 추정치 $\hat{\theta}$은 항상 수렴함을 보장한다. 예제를 통하여 EM을 좀 더 알아보자.

예제 3.5.1: 일변량 정규분포 자료

$y_i, i=1,...,n$ 는 정규분포 $N(\mu,\sigma^2)$에 추출된 임의표본이라고 하자. 이 때, $y_i, i=1,...,r$ 은 응답된 자료이고 $y_i, i=r+1,...,n$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($y_i,\ i=1,...,n$)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 모수추정치인 $\theta^{(t)}=(\mu^{(t)},\sigma^{2(t)})$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 추정치를 갱신한다. 즉,

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

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

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.

제 4장. 무응답을 포함한 자료에 대한 대체 방법 I

<학습목표>

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

(2) 깁스샘플러와 자료확대 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.

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

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

연속형 자료(continuous data)에 대하여 가장 흔히 고려되는 분포는 정규분포(continuous distribution)이다. p개의 변수 $Y_1, Y_2, ..., Y_p$를 확률변수 벡터 $(Y_1, Y_2, ..., Y_p)'$로 나타내고 이 확률 변수 벡터가 평균벡터(mean vector) μ와 분산공분산행렬(variance-covariance matrix) Σ를 가지는 다변량 정규분포(multivariate normal distribution)를 따른다고 가정하자. 여기서, $\mu_i, i = 1, ..., p$는 확률변수 Y_i의 평균을, $\sigma_{ij}, i = 1, ..., p, j = 1, ..., p,$ 는 Y_i와 Y_j의 공분산을 나타내

고 $\sigma_{ii}, i=1,...,p,$는 Y_i의 분산을 나타내면 평균벡터와 분산공분산행렬은

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

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

으로 나타낼 수 있다.

확률변수 벡터로부터 서로 독립적으로 (independently) 추출된 n개의 관찰단위 벡터들을 소문자 $y_1, y_2, ..., y_n$으로 나타내면

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

으로 나타낼 수 있다. 이 때, $y_i = (y_{i1}, y_{i1}, ..., y_{ip})$인 i번째 관측단위에서 측정된 p개의 변수값을 나타내는 $p \times 1$ 벡터가 되며 *iid*는 서로 독립이고 동일한 분포를 가진다는(independent and identically distributed) 의미이다.

4.1.1 완전자료(complete-data)의 최대우도 추정량

자료가 완전히 응답되었다면 완전자료의 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 우도함수의 최대화(maximizing the likelihood)를 통하여 모수 μ와 Σ의 최대우도추정량(maximum likelihood estimator)은 다음과 같이 구해진다.

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

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

4.1.2 무응답 패턴과 무응답 자료의 최대우도추정량

무응답 자료의 분석을 위하여 흔히 자료를 무응답 패턴(missingness patterns)에 따라 재정렬하게 된다. 무응답 패턴이란 자료에 존재하는 서로 다른 응답-무응답 형태를 의미한다. 간단한 예로서 2개의 변수 Y_1과 Y_2를 가진 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. 여기서 변수 Y_1과 Y_2가 모두 무응답이라면 그 관찰단위는 응답이 하나도 존재하지 않으므로 자료에 포함되지 않게 되어 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 관찰단위의 숫자는 일반적으로

서로 다르다.

일반 자료의 경우 2개의 변수 대신 p개의 변수가 있으므로 <그림 4.1>과 같이 최대 2^p-1개의 무응답 패턴이 존재 가능하다. 물론 변수의 수가 많아진다면 자료 내에 존재하지 않는 무응답 패턴도 많아져 실제 무응답 패턴의 개수는 2^p-1보다 훨씬 작게 되고 각 패턴마다 속하는 자료의 숫자도 일반적으로 각각 다르다. 자료 내에 실제로 존재하는 무응답 패턴의 숫자를 S개라고 가정하고 각 무응답 패턴에 속하는 관찰단위의 개수를 $n_s, s=1,2,...,S,$라 하자. 이 때, $\sum_{s=1}^{S} n_s = n$으로서 자료 전체 관찰단위의 숫자와 동일하다.

자료행렬 Y가 무응답을 포함하고 있는 경우 $y_i=(y_{i1},y_{i1},...,y_{ip})=(y_{i,obs},y_{i,mis})$ 중 $y_{i,obs}$만을 관찰할 수 있기 때문이다. 즉, 응답된 자료 부분만의 정보에 근거하여 추론을 해야 하고 이 관측된 자료의 우도함수를 관측자료 우도함수(observed-data likelihood)함수라 부른다. 관측자료 우도함수는

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

로 나타낼 수 있다. 여기서 n_s는 무응답 패턴 s를 가지는 관찰단위들의 개수를 의미하고 $\mu_{i,obs}$와 Σ_s는 평균벡터 μ와 분산공분산행렬 Σ에서 무응답 패턴 s하에서 관찰된 변수에 해당되는 부분 평균벡터와 부분 분산공분산행렬을 의미한다. 즉, 무응답 자료에서는 관찰자료의 우도함수를 최대화시키는 최대우도추정량을 구해야 하며 이 계산은 완전자료의 경우와 같이 폐쇄형(closed form)으로 표현될 수 없다.

<그림 4.1> 자료행렬에서 가능한 무응답 패턴의 예

무응답 패턴	관찰단위	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		?	?	…	?

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 자료를 회귀분석을 이용하여 분석하고 회귀계수가 유의한지가 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 관찰자료의 우도함수를 구하고 이 우도함수의 최대화를 위하여 EM algorithm 등의 방법을 이용하여 분석을 시행해야 한다. 문제는 각 모형 및 무응답 형태에 따라 우도함수가 달라지며 각 경우에 알맞은 상용 통계 프로그램이 모두 개발되어 있지 않으므로 연구자가 직접 본인의 모형에 적절한 프로그램을 개발하거나 다른 연구자가 개발한 프로그램을 구해서 사용해야 하는 번거로움이 있다. 이 번거로움을 피하는 한 가지 방법은 자료의 무응답 부분을 대체(imputation)하여 완전한 형태의 대체된 자료(imputed data)를 만드는 것이다. 대체된 자료는 무응답을 포함하지 않으므로 상용 프로그램을 사용하여 원하는 분석을 자유롭게 시행할 수 있다는 점에서 매우 유용하며 이 이유 때문에 무응답의 대체는 무응답을 다루는 인기 있는 방법이 되었다.

4.1.3 무응답 자료의 대체에 사용되는 기법

무응답 자료의 모수적 대체는 마르코프 체인 몬테칼로 방법(Markov Chain Monte Carlo 또는 줄여서 MCMC)을 사용한다. 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 $Z=(Z_1,Z_2,...,Z_J)$와 같이 Z가 J개의 부분벡터(subvector)로 나누어 질 수 있다면 깁스샘플러는 다음의 조건부 분포(conditional distribution)로부터의 반복 추출을 시행한다.

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

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

반복이 충분이 이루어지면 얻어진 $Z^{(t)}$값들은 우리가 추출하려고 하는 목표분포(target distribution)인 Z의 결합분포 $f(Z)$로부터 추출된 값으로 간주될 수 있다.

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

● 자료확충

확률 벡터 Z가 $Z=(U,V)$와 같이 두 개의 부분벡터(subvector)로 나누어 질

수 있고 Z의 결합분포(joint distribution) $f(Z)$로부터 난수를 생성하고자 하지만 $f(Z)$로부터 직접 난수를 생성하기 어려운 경우를 고려하자. 자료확충은 다음의 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 것을 가능하게 한다.

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

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

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

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

로부터 m개의 $v_i, i=1,...,m,$값들을 추출한다.

만약 $m=1$인 경우 자료확충은 깁스 샘플러의 $J=2$인 경우에 해당된다. Tanner and Wong(1987)은 자료확충을 이용하여 무응답 자료에서부터 대체를 실시하는 방법을 제안하였다.

● 무응답 자료에 대한 자료확충

무응답을 포함한 자료의 경우 관측자료 우도합수는 폐쇄형 해답(closed form solution)을 가지지 않기 때문에 분석이 용이하지 않다. 만약 관찰된 자료 Y_{obs}에 결측된 자료 부분인 Y_{mis}를 붙여서 확충(augment)시킬 수 있다면 자료는 완전자료 형태인 $Y=(Y_{obs}, Y_{mis})$가 되고 3.1.1.1에서 본 바와 같이 완전자료 우도함수를 사용하여 쉽게 분석할 수 있다.[2] 분석은 다음의 두 단계로 진행된다. 반복 시점을 $t, t=1,2,3,...$라 하자. 시점 t에서 모수의 값이 $\theta^{(t)}$라면,

(1) Imputation Step (I-step)

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) Posterior Step (P-step)

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

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

로부터 추출한다.

---

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

반복이 충분이 이루어지면 얻어진 $\theta^{(t)}, Y_{mis}^{(t)}$ 값들은 $f(\theta, Y_{mis}|Y_{obs})$로부터 추출된 값으로 간주될 수 있다. 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 따른 추출된 모수의 시계열 그림(time series plot of parameters)을 사용하게 된다. 각 모수에 대하여 시계열 그림을 그려 값들이 특정한 패턴을 보이지 않고 일정하게 변동하면 수렴상태를 의미한다. 또한, MCMC 방법으로 추출된 값들은 Markov Chain의 성질에 따라 연속된 시점들로부터 추출된 값들 사이에 연관성이 존재하지만 이 연관성은 시점이 멀어지면 줄어들어 없어지는데 이 사항을 파악하기 위하여 모수들 간 자기상관 그림(autocorrelation plot)도 사용된다.

4.1.4 다변량 정규분포(multivariate normal distribution)에서의 무응답 자료의 대체

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

(1) Imputation Step (I-step)

자료행렬 Y의 n개의 관찰단위는 $y_1, y_2, ..., y_n$으로 표현되고 각 관찰값은

응답된 변수 $y_{i,obs}$들과 무응답인 변수 $y_{i,mis}$들로 구성되어 함께 $y_i = (y_{i,obs}, y_{i,mis}), i = 1,...,n,$으로 표현 가능하다. 이 때 n개의 관찰단위는 서로 독립이므로 각 관찰단위 $i = 1,...,n,$에 대하여 $y_{i,mis}$는

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

로부터의 추출된다. 이 때, 자료행렬 Y가 다변량 정규분포를 따르고 다변량 정규분포 하에서 조건부 확률분포도 다변량 (또는 $y_{i,mis}$가 한 개의 변수만 포함하면 일변량) 정규분포가 되므로 $y_{i,obs}$와 $\theta^{(t)}$에 조건 지어진 $y_{i,mis}$의 분포도 정규분포가 된다. 정규분포를 따르는 $y_{i,mis}$의 평균은 i번째 자료 중 무응답인 변수들을 반응변수(response variables)로 놓고 응답인 변수들을 설명변수(explanatory variables)로 설정하여 회귀분석을 실시한 예측값(predictive value)이 되며 $y_{i,mis}$의 분산은 이 회귀분석의 잔차공분산 행렬(residual covariance matrix) (또는 $y_{i,mis}$가 한 개의 변수만 포함하면 잔차공분산)에 해당된다.

(2) Posterior Step (P-step)

다변량 정규분포의 모수 θ는 평균벡터 μ와 분산공분산행렬 Σ 두 가지이다. I-step에서 대체된 $y_{i,mis}^{(t+1)}, i = 1,...,n,$를 관찰된 자료 $y_{i,obs}, i = 1,...,n,$와 합하여 확충시키면 대체된 자료벡터 $y_i^{(t+1)} = \left(y_{i,obs}, y_{i,mis}^{(t+1)}\right), i = 1,...,n,$이 되고 n개의 관찰단위 전체로 표현하면 대체된 자료행렬인 $Y^{(t+1)} = \left(Y_{obs}, Y_{mis}^{(t+1)}\right)$이 된다. 대체된 자료값이 마치 주어진 것처럼 간주하면 분산공분산행렬 Σ의 분포와 Σ가 주어졌을 때 평균벡터 μ의 조건부분포는 각각

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

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

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 역 위샤트분포(inverted Wishart distribution)를 의미한다. 위의 식은 Σ의 분포와 Σ이 주어졌을 때 μ의 조건부 분포로 표현되는 데 이는 두 모수의 결합분포가 Σ의 분포와 Σ이 주어졌을 때 μ의 조건부 분포의 곱으로 표현할 수 있기 때문이다.

4.1.4.1 사전정보(prior information)를 이용한 대체

베이지안 분석(Bayesain 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 미리 알려진 정보에 근거하여 정해지며 $\tau > 0, m \geq p,$ 그리고 $\Lambda > 0$이다. 이 사전분포 하에서의 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, $\mu_1 = \left(\frac{n}{\tau+n}\right)\bar{y} + \left(\frac{\tau}{\tau+n}\right)\mu_0$,

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

을 의미한다.

공액사전분포를 사용하여 무응답을 포함한 자료에 대하여 대체를 실시하기 위해서는 (1) Imputation Step은 동일하지만 (2) Posterior Step에서 대체된 자료값이 마치 주어진 것처럼 간주하면 분산공분산행렬 Σ의 분포와 Σ가 주어졌을 때 평균벡터 μ의 조건부분포 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, 관찰단위의 수가 모수의 숫자보다 충분히 크지 않다면 위 모수를 안정적으로 추정하는 데 문제가 생길 수 있다. 또는 변수들이 서로 연관성이 높다면 $\hat{\Sigma}$이 비정칙행렬(singular matrix) 또는 비정칙행렬에 가까워지고 역함수(inverse matrix)의 계산에 문제가 생긴다. 이와 같이 Σ의 추정에 문제가 생기면 능형회귀(ridge regression)의 개념을 이용한 능형사전함수(ridge prior)의

사용이 종종 도움이 된다(Schafer, 1997). 능형사전함수는 공액사전함수에서 $\tau \to 0$ 인 극한분포를 의미한다. 능형사전함수하에서의 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 더하여 각각 m을 더하여 비정칙행렬의 문제를 해결하는 효과를 지닌다. 이 방법으로 대체를 실시하기 위해서는 위에서 주어진 분산공분산행렬 Σ의 사후분포와 Σ가 주어졌을 때 평균벡터 μ의 조건부 사후분포로부터 모수 μ와 Σ을 추출하면 된다.

4.1.4.2 다변량 정규분포를 따르는 무응답 자료의 대체 프로그램

다변량 정규분포를 따르는 무응답 자료의 대체를 실시하는 프로그램들은 상용화된 프로그램에 포함된 경우가 종종 있다. 예를 들면 SAS의 MI procedure, SPSS의 Missing Value Analysis(MVA), S-Plus의 missing data library를 포함한다.

다음은 SAS MI procedure에서 무응답 대체를 실시하는 방법을 설명한다. <그림 4.2>는 SAS MI procedure의 기본 신텍스(syntax)를 나타낸다.

<그림 4.2> 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	난수생성시 사용하는 0 또는 양수(positive number). (동일한 양수를 사용하면 동일한 결과를 얻음) (0이나 default는 컴퓨터 시간에 의해 결정되는 임의로 값)
ROUND=numbers	대체된 값을 반올림(round off)할 자리수를 지정
MAXIMUM=numbers	대체 가능한 최대값(maximum)
MINIMUM=numbers	대체 가능한 최소값(maximum) (최대값과 최소값을 지정해 줌으로써 정해진 구간 안에 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 대체를 실시하기 전에 MCMC를 시행하는 숫자로 목표함수로 수렴(converge)하기 이전의 반복을 버린다는 의미로 burn-in period라 불림

(default는 200)

NITER=number 단일연쇄(single chain)를 이용하여 다중대체를 실시할 때 반복이 어느 기간 지나면 다시 대체값을 선택할 지 지정

(default는 100)

PRIOR=name 사전분포를 사용한 대체시 지정

PRIOR=JEFFREYS (사전분포를 지정하지 않는 방법으로서 default)

PRIOR=RIDGE=number (능형사전함수의 m 지정)

PRIOR=INPUT=SAS-data-set (사전분포 정보를 포함한 SAS-Dataset 지정)

INITIAL=<options> MCMC를 실시할 때 모수의 초기값(initial value)을 지정

INITIAL=EM (EM algorithm을 사용하여 최대우도추정량이나 사후최빈값(posterior mode)을 구하여 이 값을 초기값으로 사용하는 방법

	으로서 default) INITIAL=INPUT=SAS-data-set (초기값으로 사용될 모수의 추정량을 포함하고 있는 SAS-Dataset 지정)
TIMEPLOT <options>	반복에 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>	반복에 따른 모수들의 자기상관 그림을 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 statement 에는 대체를 실시할 변수들을 포함시키며 이 문장이 사용되지 않으면 자료행렬의 연속형 변수 모두가 분석에 포함된다.

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

예제 4.1 기업활동실태조사의 무응답 대체

기업활동실태조사는 기업 활동의 다각화, 국제화, 계열화 등 기업의 다양한 경제활동을 포괄적으로 조사함으로써 기업의 경영전략이나 산업구조 변화를 파악하여 기업에 관한 각종 경제정책의 기초자료를 제공하기 위하여 통계청에서 실시된다. 매년 실시되는 이 조사는 각 연도별로 전국의 회사법인 중 종사자 50인 이상이며 자본금 3억원 이상인 11,650개 기업을 대상으로 실시하는 기업체 조사이다. 조사항목은 기업체명, 소재지, 자본금, 기업 내 조직 및 종사자수 관련, 자산·부채 및 자본 관련, 사업내용관련, 관계회사(자회사, 관련회사, 모회사)관련, 기업 간 거래 및 해외거래 관련, 기술소유 및 사용관련, 기업의 경영방향 관련 항목 등 7개 분야 111개 항목을 포함한다. 실제 조사에서 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 자료에서 일부항목에 대하여 임의로 무응답을 설정하여 분석을 실시한 결과를 보여준다. 본 예제의 목적은 무응답에 대한 가장 적절한 대체 방법을 제안하는 것이 아니라 다변량 정규분포를 가정한 대체 방법의 예제를 보여주기 위한 것임을 명시한다.

본 예제에서는 응답을 제공한 10,229개 기업에서 자본금(C5), 사업체수(C7), 상용종사자수(C8), 자산총계(C9), 유형자산 당기 취득액(C18), 매출액(C24), 그리고 영업비용(C41) 네 가지 변수 각각에서 무응답이 완전임의로 10% 발생하였다고 가정하고 대체를 실시하였다. 대체에는 SAS MI procedure를 사용하였고 <그림 4.3>는 프로그램 코드를 보여준다. 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 따른다

고 가정하는데 모든 변수의 값이 0보다 크므로 대체를 시행할 때 최소값이 0이 되도록 하였고 (MIN=0 사용) 이 중 두 변수 사업체수(C7)와 상용종사자수(C8)는 숫자로 응답되므로 정수가 되도록 반올림하였다. 단일대체를 시행한 (NIMPUTE=1) 후 <그림 4.4>에 나타난 것과 같은 출력문이 나오는데 이는 분석 정보 및 이 자료에서 나타난 무응답 자료 패턴을 보여준다. 그 외 EM algorithm을 통한 평균의 추정량에 대한 정보도 주어진다. <표 4.1>은 단일대체를 통한 평균의 추정값과 이용가능한 9206개 자료에 대한 분석방법을 시행한 경우 평균의 추정값을 비교한다.

<그림 4.3> 기업활동실태조사의 무응답 대체를 위한 SAS MI procedure 코드

```
PROC MI DATA=company OUT=micompany NIMPUTE=1
        seed=2340634 MIN=0;
  MCMC NITER=1000;
  VAR variables;
RUN;
```

<표 4.1> 단일대체를 통한 평균의 추정값과 이용가능한 자료 분석방법을 시행한 경우 평균의 추정값을 비교 (괄호안은 표준편차)

	단일대체	이용가능한 자료 분석
자본금(C5)	2.15 (0.16)	2.15 (0.16)
사업체수(C7)	1.00 (0.21)	0.76 (0.28)
상용종사자수(C8)	1.99 (0.10)	1.76 (0.14)
자산총계(C9)	9.56 (1.54)	9.96 (1.54)
유형자산 당기 취득액(C18)	6.22 (2.53)	6.22 (1.52)
매출액(C24)	10.16 (1.41)	10.17 (1.42)
영업비용(C41)	10.14 (1.37)	10.13 (1.37)

<그림 4.4> 기업활동실태조사의 무응답 대체를 위한 SAS MI procedure의 주요 출력문

The MI Procedure

Model Information

Data Set	WORK.COMPANY
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	4920	48.10
2	X	X	X	X	X	X	.	540	5.28
3	X	X	X	X	X	.	X	555	5.43
4	X	X	X	X	X	.	.	66	0.65
5	X	X	X	X	.	X	X	536	5.24
6	X	X	X	X	.	X	.	69	0.67
7	X	X	X	X	.	.	X	48	0.47
8	X	X	X	X	.	.	.	6	0.06
9	X	X	X	.	X	X	X	538	5.26
10	X	X	X	.	X	X	.	59	0.58
11	X	X	X	.	X	.	X	55	0.54
12	X	X	X	.	X	.	.	10	0.10
13	X	X	X	.	.	X	X	55	0.54
14	X	X	X	.	.	X	.	5	0.05
15	X	X	X	.	.	.	X	9	0.09
16	X	X	.	X	X	X	X	546	5.34
17	X	X	.	X	X	X	.	50	0.49
18	X	X	.	X	X	.	X	53	0.52
19	X	X	.	X	X	.	.	2	0.02
20	X	X	.	X	.	X	X	63	0.62
21	X	X	.	X	.	X	.	7	0.07
22	X	X	.	X	.	.	X	8	0.08
23	X	X	.	.	X	X	X	64	0.63
24	X	X	.	.	X	X	.	10	0.10
25	X	X	.	.	X	.	X	10	0.10
26	X	X	.	.	.	X	X	5	0.05
27	X	X	.	.	.	X	.	1	0.01
28	X	.	X	X	X	X	X	502	4.91
29	X	.	X	X	X	X	.	58	0.57
30	X	.	X	X	X	.	X	69	0.67
31	X	.	X	X	X	.	.	7	0.07
32	X	.	X	X	.	X	X	61	0.60
33	X	.	X	X	.	X	.	13	0.13
34	X	.	X	X	.	.	X	11	0.11
35	X	.	X	X	.	.	.	1	0.01
36	X	.	X	.	X	X	X	69	0.67
37	X	.	X	.	X	X	.	10	0.10
38	X	.	X	.	X	.	X	7	0.07
39	X	.	X	.	X	.	.	1	0.01
40	X	.	X	.	.	X	X	9	0.09
41	X	.	X	.	.	X	.	3	0.03
42	X	.	.	X	X	X	X	58	0.57
43	X	.	.	X	X	X	.	3	0.03
44	X	.	.	X	X	.	X	14	0.14
45	X	.	.	X	X	.	.	2	0.02
46	X	.	.	X	.	X	X	8	0.08
47	X	.	.	X	.	X	.	1	0.01
48	X	.	.	.	X	X	X	6	0.06
49	X	.	.	.	X	X	.	2	0.02
50	X	.	.	.	X	.	X	1	0.01
51	.	X	X	X	X	X	X	549	5.37
52	.	X	X	X	X	X	.	64	0.63
53	.	X	X	X	X	.	X	51	0.50

4.2 여러 가지 분포를 가진 변수들을 포함한 자료에 대한 대체 방법

대부분의 자료는 여러 형태(type)의 여러 가지 변수를 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 연속형 변수이나 음이 아닌 정수값만 가능하고 많은 사람들의 방문횟수가 0 또는 작은 숫자이지만 일부 사람들의 병원 방문횟수는 30번 이상으로 매우 크게 나타나 자료가 오른쪽으로 치우친 형태를 갖게 된다. 이렇게 여러 가지 다른 형태의 변수들을 포함한 자료의 경우 3.1.1의 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 어렵기 때문에 발생하게 된다.

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 관찰단위에 대한 k개의 설명변수들의 값을 행렬 X로 표현하다. 여기서, Y의 p개의 변수 $Y_1, Y_2, ..., Y_p$는 각각 다른 변수 타입을 가지고 k개의 설명변수들 또한 여러 가지 다른 변수 타입이 사용 가능하다. 이 자료에 대한 모수적 모형에 근거한 대체를 위하여 설명변수 X의 값이 주어졌을 때 p개의 변수 $Y_1, Y_2, ..., Y_p$들의 결합 조건

부 분포(joint conditional density)는 $f(Y_1, Y_2, ..., Y_p|X, \theta_1, \theta_2, ..., \theta_p)$인데 변수들의 타입이 다양하므로 이 분포로부터 표본을 직접 추출할 수 없다. 하지만 이 결합조건부 부포는

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

와 같이 여러 개의 조건부 분포(conditional density)의 곱으로 표현될 수 있다. 여기서, $\theta_i, i = 1, ..., p,$는 각 조건부 분포의 모수들의 벡터를 의미한다. p개의 변수 $Y_1, Y_2, ..., Y_p$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. 이와 같이 적절한 조건부 분포 모형을 가정하고 그 모형의 모수인 $\theta_i, i = 1, ..., p,$를 MCMC 기법을 이용하여 추출하고 관찰된 자료 Y_{obs}와 추출된 모수들 $\theta_i, i = 1, ..., p,$이 주어졌다는 가정 하에서 Y_{mis}의 예측분포(predictive distribution)로부터 Y_{mis}를 추출하는 과정을 반복 시행함으로써 대체를 실시할 수 있다.

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 주어졌다 가정한 후 $f\left(Y_2|X,\ Y_1^{(1)},\theta_2\right)$에 대한 적절한 모형을 사용하여 모수 θ_2를 추출하고 이 추출된 θ_2과 설명변수행렬 X, 대체된 Y_1, 관찰된 Y_2값이 주어졌다는 조건하에서 Y_2의 무응답값을 예측분포함수(또는 예측 사후분포함수)로부터 추출한다.

$\vdots$

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

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

(1) $f\left(Y_1|X,\ Y_2^{(c-1)},\ldots,Y_p^{(c-1)},\theta_1\right)$ 모형을 사용하여 모수 θ_1을 추출하고 이 추출된 θ_1과 설명변수행렬 X, 전 반복에서 대체된 $Y_2^{(c-1)},\ldots,Y_p^{(c-1)}$, 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 $f\left(Y_2|X,\ Y_1^{(c)},\ Y_3^{(c-1)},...,\ Y_p^{(c-1)},\theta_2\right)$에 대한 적절한 모형을 사용하여 모수 θ_2를 추출하고 이 추출된 θ_2과 설명변수행렬 X, 대체된 $Y_1^{(c)}, Y_3^{(c-1)},...,\ Y_p^{(c-1)}$, 관찰된 Y_2값이 주어졌다는 조건하에서 Y_2의 무응답값을 예측분포함수(또는 예측 사후분포함수)로부터 추출한다.

⋮

(p) 단계 (1)부터 단계 (p-1)에서 대체된 $Y_1^{(c)},...,\ Y_{p-1}^{(c)}$를 사용하여 $f\left(Y_p|X,\ Y_1^{(c)},...,\ Y_{p-1}^{(c)},\theta_p\right)$에 대한 적절한 모형을 사용하여 모수 θ_p를 추출하고 이 추출된 θ_p과 설명변수행렬 X, 대체된 $Y_1^{(c)},...,\ Y_{p-1}^{(c)}$, 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 $\hat{f}\left(Y_j|X,\ Y_1^{(c)},...,\ Y_{j-1}^{(c)},\ Y_{j+1}^{(c)},...,\ Y_p^{(c)},\theta_p\right)$을 가지고 $f\left(Y_j|X,\ Y_1^{(c)},...,\ Y_{j-1}^{(c)},\ Y_{j+1}^{(c)},...,\ Y_p^{(c)},\theta_p\right)$을 근사(approximate)하므로 SIR algorithm (Rubin, 1987b)이나 rejection algorithm(Gelman et. al., 2004)을 사용하여 근사를 개선하는 것이 바람직하다.

4.2.1 여러 가지 분포를 따르는 변수들을 포함한 자료에 대한 대체 프로그램

고려하는 자료가 여러 가지 다른 변수들의 타입을 포함하는 경우 대체를 실시하기 위하여 SAS 내부에서 실행되거나 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.

설문조사에는 많은 질문들이 해당되지 않는 사람들에 대하여 건너뛰도록(skip) 설계된다. 예를 들면, 흡연 관련 질문 중 흡연 기간에 관한 질문은 흡연자에게만 해당되는 질문이므로 비흡연자들에게는 건너뛰도록 만들어지고 이는 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 클 수 없다. 이와 같은 경계가 존재하는 변수에 대하여 정규분포 가정을 적용한다면 대체된 값이 가능한 경계 외부의 값이 될 수 있으므로 이를 막기 위하여 절단모형(truncated model)으로부터 대체를 실시하여야 한다.

IVEware 프로그램은 <표 4.2>와 같이 여러 가지 변수 타입에 따라 다른 회귀 모형을 다룰 수 있다. 혼합형 변수란 자료의 값이 0과 연속형 변수가 혼합되어 나타나는 형태로서 예를 들면 지난 일주일간 흡연량과 같은 변수에서 흔히 나타난다. 이 변수의 경우 비흡연자의 흡연량은 모두 0이 되고 흡연자의 흡연량은 연속형 값으로 나타나게 된다. 이와 같은 자료에 흔히 사용되는 두단계 모형(two-stage model)이란 흡연자인지 비흡연자인지를 로지스틱 회귀모형을 이용하여 적합시키고 흡연자의 흡연량은 선형회귀모형을 통하여 적합하는 두 단계로 모형을 세우는 방법을 의미한다(Schafer and Harel, 2002).

<표 4.2> 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 IVEware에서 SAS를 사용하여 무응답 대체를 실시하는 매크로 모듈 IMPUTE의 사용법을 설명한다. <그림 4.5>는 IVEware IMPUTE 모듈의 기본 신텍스(syntax)를 나타낸다.

매크로 %**IMPUTE**에서 사용되는 키워드(keyword) 및 설명은 다음과 같다.

NAME=filename　　setup file의 이름을 지정

DIR=directory　　setup file과 output file이 저장되는 directory 지정

<그림 4.5> 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;
```

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 경계식을 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 난수생성시 사용하는 0 또는 양수(positive number)값을 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 4.2 기업활동실태조사의 무응답 대체

예제 4.1의 기업활동실태조사의 무응답을 대체하기 위하여 IVEware를 사용하였다. 이 프로그램은 여러 가지 타입의 변수를 포함하는 것이 가능하기 때문에 숫자 변수인 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 4.6>는 프로그램 코드를 보여주고 <표 4.3>은 단일대체된 평균의 추정값을 보여준다. 결과는 <표 4.1>의 단일대체와 비슷하지만 이 분석에서는 예제 3.1과 달리 변환되지 않은 사업체수(C7)와 상용종사자수(C8)에 대한 평균의 추정값을 제공한다.

<그림 4.6> 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 4.3> 여러 가지 분포를 가정하여 단일대체된 평균의 추정값(괄호안은 표준편차)

	단일대체	
자본금(C5)	2.15 (0.16)	
사업체수(C7)	5.34 (0.22)	
상용종사자수(C8)	262.8 (1059)	
자산총계(C9)	9.97 (1.54)	
유형자산 당기 취득액(C18)	6.21 (2.53)	
매출액(C24)	10.17 (1.41)	
영업비용(C41)	10.14 (1.37)	

<4장 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 편의를 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?

(나) MCMC 기법을 사용하여 대체를 실시할 때 추출된 모수값들의 수렵여부를 평가하는 방법을 설명하시오.

3. 다음의 자료는 2005년 인구주택총조사 자료의 일부분이다. 제공된 자료에서 무응답은 연습문제를 위하여 임의로 생성되었다. 3장에서 학습한 대체방법들을 사용하여 대체를 실시하시오. 사용된 대체 방법들을 변수별 평균 및 표준오차를 사용하여 비교하고 이 자료에 대하여 한 가지 대체 방법을 추천하시오.

제 5장. 무응답을 포함한 자료에 대한 대체 방법 II

<학습목표>

(1) 무응답을 포함한 자료에 대한 핫덱대체 방법을 소개한다.

(2) 무응답을 포함한 자료에 대한 준모수적 모형에 근거한 대체 방법을 소개한다.

(3) 다중대체된 자료에 대한 분석 및 추론 방법을 소개한다.

5.1 핫덱대체 방법

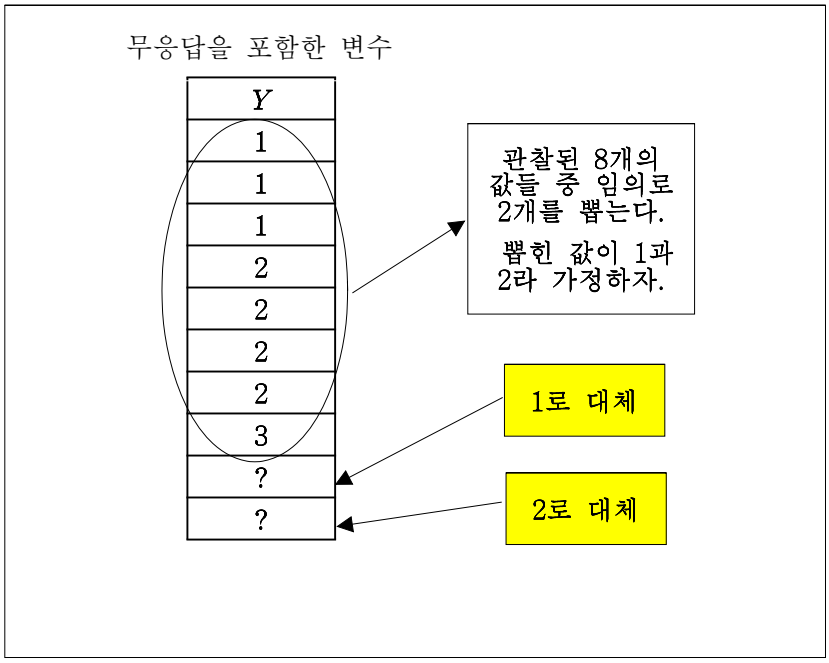
자료의 분포를 가정하지 않고 대체를 실시하는 방법 중 가장 흔히 사용되는 대체 방법이 핫덱대체(hotdeck imputation) 방법이다. 핫덱 대체방법은 무응답값의 대체를 위하여 자료내의 응답된 관찰단위의 값을 사용하는 방법이다. 카드 게임에서 핫덱(hot deck)은 현재 게임에서 다루는 카드 세트를 의미하는데 대체에서는 컴퓨터 개발 초기에 각 관찰단위가 한 개의 펀치카드(punched card)에 기록되었으므로 현재 모아진 자료를 펀치카드들의 핫덱이라 생각하여 이름 지어졌다. 핫덱 대체에서 무응답값은 자료내의 응답된 값을 가지고 대체되므로 응답값이 무응답값에 기증되었다는 의미로 대체에 사용된 응답값을 기증자(donor)로 무응답을 기증을 받은 사람(donee)로 부른다. 핫덱은 기증자를 선택하는 방법에 따라 여러 가지로 분류되는데 흔히 사용되는 방법들을 다루고자 한다.

5.1.1 단순임의 핫덱대체 방법(Hotdeck by Simple Random Sampling)

자료내의 각 무응답은 응답값 중에 한 개의 값을 임의로 선택하여 대체하는 방법

이다. <그림 5.1>은 간단한 예로서 한 개의 변수에서 무응답이 발생할 때 단순임의 핫덱대체를 시행하는 방법을 설명한다. 무응답을 포함한 자료 Y가 한 개의 변수를 포함하고 이 변수에 무응답이 발생하는 경우로서 전체 관찰단위의 숫자는 10개인데 그 중 2개의 관찰단위에서 무응답이 발생하는 경우에 응답된 8개의 관찰값이 2개의 무응답 개체를 대체하기 위하여 기증자로 사용된다. 8개의 기증자 중에서 임의로 2개의 관찰값이 복원(with replacement) 또는 비복원(without replacement)으로 추출된다. 실제 적용 시에는 비복원 추출이 선호되는 경향이 있다. 예를 들어 8개의 응답값 중에서 추출된 값이 1과 2라면 첫 번째 무응답 관찰단위에 1의 값을 두 번째 무응답 관찰단위에 2를 대체한다.

<그림 5.1> 단순임의 핫덱대체 방법



자료 Y가 여러 개의 변수를 포함한 경우에도 이 방법이 확장되어 사용되어 질

수 있다. 즉, 각 변수별로 단순임의로 응답값을 그 변수의 무응답값에 대한 기증자로 추출하고 이 값들을 가지고 무응답에 대한 대체를 실시할 수 있다. 이 방법의 문제는 이 방법으로 대체된 값에 근거한 추정량은 무응답 자료 메커니즘이 완전임의결측이 아니라면 편의가 발생한다는 데 있다. 따라서 이 방법의 적용은 한정적일 수밖에 없다.

예제 5.1.1 2008년 사회조사의 무응답 대체

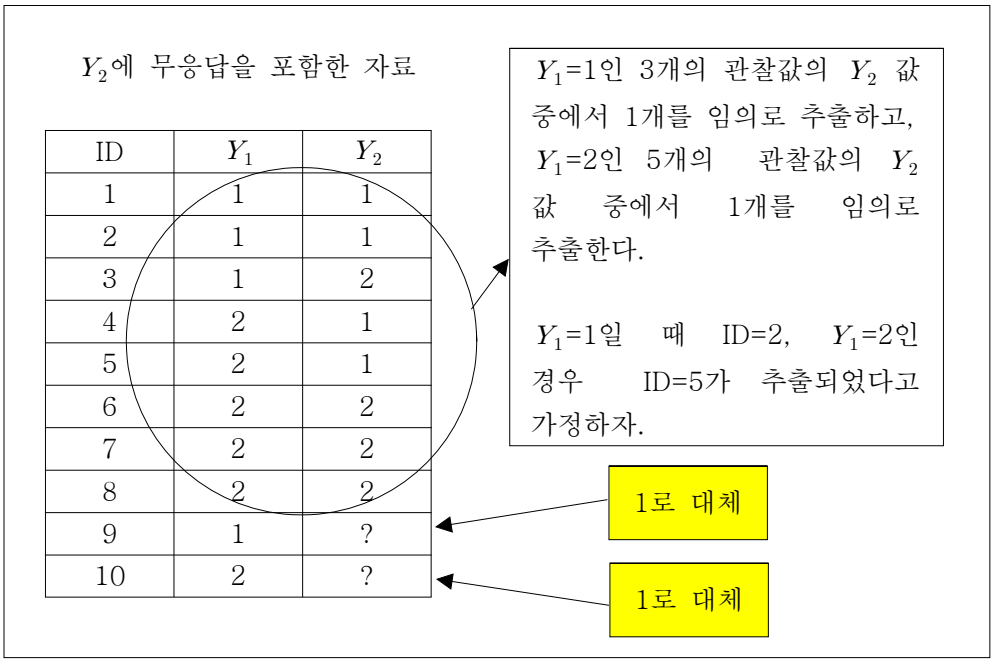
2008년에 통계청에서 시행된 사회조사는 국민의 일상생활과 관련하여 현재 처한 상황들을 조사하여 앞으로 나아가야 할 방향을 모색하고자 시행되고 있는데 매년 12개 부분 중 3-4개 부문에 대하여 조사를 실시하며 대상은 약 20,000가구의 만 15세 이상 상주 가구원이다. 2008년 사회조사는 인적사항, 교육부문, 안전부문, 환경부문의 44개 조사 항목에 대한 문항들을 포함하고 있다. 본 예제에서는 가구주 대상 설문지 문항 중 학생인 자녀가 있는 가구주에게 자녀교육비의 부담 정도를 질문한 문항에 대한 대체를 고려한다. 이 문항은 “매우 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 약 20%에 대하여 무응답을 임의로 생성하였다. <표 5.1>은 무응답 발생 전 완전자료, 무응답을 무시한 채 분석을 실시한 완전자료 분석(complete-case analysis), 그리고 단순임의 핫덱대체를 실시한 후 대체된 자료의 자녀 교육비 부담 정도에 대한 응답 비율을 비교한다.

<표 5.1> 자녀 교육비 부담 정도에 대한 응답 비율에 대한 비교

5.1.2 대체군을 이용한 핫덱대체 방법(Hotdeck by Simple Random Sampling)

무응답의 대체를 실시할 때 관찰된 다른 변수들이 동일하거나 비슷한 개체가 기증자로 선정된다면 완전임의로 대체를 실시하는 것보다 더 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 기증을 위한 대체군을 형성하고 이 대체군 내에서 임의로 한 응답자를 추출하여 이 응답자의 소득을 가지고 무응답자의 소득을 대체한다면 더 정확한 대체를 실시할 수 있다. <그림 5.2>은 대체군을 이용한 핫덱대체의 간단한 예를 보여준다. 자료가 관찰자 식별번호인 ID 변수와 두 개의 변수(Y_1과 Y_2)를 포함하고 두 변수 중 Y_2에서만 무응답이 발생할 때 완전히 응답된 Y_1의 값에 의존하여 대체군을 형성하고 이 대체군을 이용한 핫덱대체를 시행하는 방법을 설명한다. 무응답은 9번째와 10번째 관찰단위(ID = 9와 10)에서 발생하였다. 즉, ID = 9는 $Y_1 = 1$의 값을 ID = 10은 $Y_1 = 2$가 관찰되었으나 Y_2 변수의 값은 결측이었다. 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 = 4 ~ 8 다섯 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.

<그림 5.2> 대체군을 이용한 핫덱대체 방법



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 대체의 편의(bias)가 발생하지 않을 것이다. 이것은 대체군을 형성하기 위하여 무응답이 발생한 변수와 연관되어 있는 여러 가지 변수를 포함함으로써 달성되기 쉬울 것이다. 즉, 대체군을 형성하는 변수를 더 많이 고려할 수록 이 가정은 달성이 될 가능성이 높다. 하지만 문제는 대체군을 형성하기 위하여 변수가 추가될수록 대체군의 숫자가 기하급수적으로 늘어나는 데 있다. 예를 들어 대체군 형성 변수로 성별(“남”, “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 늘어나

며 거주지 구분(“시,” “도,” “군”) 및 자산정도(“1천만원 미만,” “1천 초과 ~ 1억,” “1억 초과 ~ 5억,” “5억 초과 ~ 10억,” “10억 초과”)를 추가하면 $2 \times 8 \times 3 \times 5 = 240$개로 크게 늘어난다. 대체군의 숫자가 늘어나면 일부 대체군에 속하는 응답값을 가진 개체의 수가 적어지거나 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 대체군에 속하는 응답이 하나도 없는 경우가 발생하는 것이다. 또한 특정 대체군 내에서 응답자의거나 무응답자의 숫자보다 적은 응답자가 있는 경우도 가능하며 이 경우 이 대체군에 속하는 무응답자는 기증자를 발견할 수 없어 대체될 수 없거나 복원추출을 사용하는 경우 한 명의 응답자가 여러 무응답자에 대하여 대체되는 등의 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 기법의 문제점은 대체군을 형성하는 변수를 어떻게 설정해야 하는가와 기증자를 찾기 위하여 포기해야 하는 대체군 형성 변수의 순서

를 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 무응답에 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 포기도 이와 마찬가지다.

예제 5.1.2 2008년 사회조사의 무응답 대체

4.1절에서 다룬 2008년 사회조사의 예를 다시 고려하자. 본 예에서는 가구주의 성별, 연령, 학력 및 결혼 상태에 근거하여 대체군을 형성하고 그 대체군 내에서 무응답의 대체를 실시하였다. 가구주의 성별은 2개, 연령은 “30대 또는 미만,” “40대”, 그리고 “50대 또는 이상”의 3개 항목으로 나누었는데 이는 대부분의 학생 자녀를 가진 가구주의 연령을 고려하여 결정되었다. 학력은 “고졸 이하”와 “대학 이

상"으로 나누고 결혼 상태는 "혼인 중"과 "그 외"로 구분하여 전체 대체군의 숫자는 2 × 3 × 2 × 2 = 24개였다. <표 5.2>은 무응답 발생 전 완전자료, 무응답을 무시한 채 분석을 실시한 완전자료 분석(complete-case analysis), 그리고 두 변수에 근거한 대체군을 이용한 핫덱대체를 실시한 후 대체된 자료의 변수별 추정된 평균을 비교한다.

<표 5.2> 변수별 평균의 비교

5.1.3 최근접 이웃 핫덱대체 방법(Nearest Neighbor Hotdeck)

3.2.2절에서 다룬 대체군을 이용한 핫덱대체를 실시할 때 대체군을 형성하는 변수는 범주형 변수이다. 연속형 변수가 대체군을 형성하는데 포함되기 위해서는 연속형 변수를 범주형 변수로 범주화하여야 한다. 연속형 변수는 범주형 변수보다 더 정확한 정보를 포함하고 연속형 범주를 어떻게 범주화하느냐에 따라서 분석의 결과가 달라지는 등 민감한 문제를 포함하고 있다. 정확한 대체를 실시하기 위하여 고려하고자 하는 변수가 연속형이라면 최근접 이웃 핫덱대체 방법을 실시할 수 있다. 이 방법은 대체에 고려하고자 하는 변수들의 값에 근거하여 관찰단위 개체들 사이의 거리를 측정하는 거리(distance metric)를 정의하고 무응답이 발생한 개체와 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 근접한 연령이 될 수도 있다.

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

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

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

(1) 거리의 최대 편차(maximum deviation): i번째 응답자와 j번째 응답자의 각 변수별 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 $X_1,...,X_k$들 간의 추정된 분산공분산행렬(the estimate of the covariance matrix)을 의미한다.

(3) 예측평균 (predictive mean): 무응답을 포함한 변수 Y에 대한 변수 $X_1,...,X_k$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 $X_1,...,X_k$를 설명변수로 고려한 Y의 회귀분석에서 i번째 응답자의 Y의 예측값(predictive value)을 의미한다.

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

최근접 이웃 핫덱 방법은 고려하는 변수들이 범주형인 경우 적용될 수 없다. 하지만 3.2.2절의 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 최근접 이웃인 j번째 응답자의 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 획득할 수 있다.

핫덱 대체는 명시적 형태의 모형(explicit model)을 사용하지 않고 내재적 모형(implicit model)을 사용한다. 더구나 단순임의 핫덱 대체보다는 대체군에 근거한 핫덱대체나 최근접 이웃 핫덱대체, 또는 혼합 등 고려된 변수들의 복잡한 함수들(complex functions)에 근거하여 무응답값에 대한 대체가 발생하므로 대체된 자료에 근거한 추정량의 성질을 평가하기 힘든 문제점이 있다. 하지만 많은 모의실험은 신중하게 선택된 방법에 근거한 핫덱대체는 정확한 대체를 실시할 수 있고 대체된 자료에 근거한 추정량에서 편의가 발생하지 않는다고 보고하고 있다.

5.2 준모수적 모형에 근거한 대체 방법

모수적 모형에 근거한 대체 방법은 자료가 모수적 모형의 가정을 만족시키는 경우 우수한 성능을 보이는 것으로 나타났다. 자료가 모형의 가정을 만족시키지 못하는 경우에도 이 모형은 많은 경우 강건한데(robust) 그 이유는 대체가 전체 자료에 대하여 이루어지는 것이 아니라 무응답값에 대하여만 실행되므로 무응답의 비율이 높지 않다면 잘못된 대체로 인한 추정량의 편의는 크지 않다는 것이다. 하지만 분포 가정이 심하게 위배되는 경우 무응답의 비율이 높지 않아도 편의가 발생할 수 있다고 보고되고 있다 (Tang et. al., 2005).

핫덱대체 방법은 자료에 대한 분포 가정을 고려하지 않으므로 여러 가지 다른 타입의 변수에 유연성있게 적용할 수 있고 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 조

정 등 많은 노력이 필요한 기법이다.

모수적 모형에 근거한 대체 방법과 핫덱 대체 방법들의 장점을 유지하면서 단점을 보완하기 위한 여러 가지 모형들이 제시되어 왔다. 본 절에서는 핫덱 대체에 모수적 모형 기법을 접목시킨 대체 방법과 모수적 모형의 예측력을 높이기 위한 비선형 회귀모형(semiparametric model)에 근거한 대체방법과 에 관하여 논의한다.

5.2.1 응답성향점수(response propensity score)를 이용한 핫덱대체 방법

성향점수(propensity score)는 두 개의 처리(treatment)가 존재할 때 공변량이 주어졌을 때의 한 개의 처리에 할당할 조건부 확률을 의미한다(Rosenbaum and Rubin, 1983, 1984). 무응답의 대체를 위하여 두 개의 처리 대신 응답이 발생했는지 여부를 나타내는 응답성향점수는 공변량이 주어졌을 때 응답의 확률을 의미한다. 무응답을 포함한 변수벡터 Y와 이와 연관된 p개의 변수, $X_1, X_2, ..., X_p$가 존재할 때 응답 여부를 나타내는 변수벡터 R의 i번째 원소는

$$r_i = \begin{cases} i\text{번째 관찰단위의 } Y\text{값이 응답이면 } 1 \\ i\text{번째 관찰단위의 } Y\text{값이 무응답이면 } 0 \end{cases}$$

으로 정의된다. 이 때 응답성향점수는

$$P(R = 1 \mid X_1, ..., X_p)$$

로 정의된다. 즉, 응답성향점수는 공변량 $X_1, X_2, ..., X_p$의 값이 주어졌을 때 변수

Y에서 응답이 발생할 확률을 의미한다. 실제로 이 응답성향점수를 계산하기 위해서는 응답여부가 이산형(binary)으로 나타나므로 로지스틱 회귀분석이나 프로빗 회귀분석(probit regression)을 적용한다.

성향점수는 p개의 변수, $X_1, X_2, ..., X_p$ 변수에 근거한 p차원 공간을 1차원 성향점수값으로 차원축소(dimension reduction)시키는 기법이다. 즉, 대체군을 형성하기 위하여 p개의 변수를 사용한다면 대체군의 개수는 p가 증가함에 따라 기하급수적으로 증가하지만 성향점수는 p의 개수와 상관없이 0과 1의 값을 가진다. 더구나, 로지스틱 분석이나 프로빗 분석에서 설명변수 $X_1, X_2, ..., X_p$들은 연속형이거나 범주형 모두가 사용 가능함으로써 변수 타입에 관계없이 적용이 가능하다.

무응답 대체를 위하여 응답성향점수가 비슷한 개체들을 짝짓는 것이 필요한데 Little(1988b)은 예측평균에 근거한 짝짓기 방법(predictive mean matching method)을 제안하였고 Bell(1999)은 수정된 예측평균에 근거한 대체 방법을 제안하였다. 이 방법의 아이디어는 예측평균(predictive mean)이 가장 가까운 응답자와 무응답자들을 짝지어(match) 대체를 실시하는 것에서 시작된다. 즉, 무응답 예측평균이 가까운 응답자와 무응답자들을 짝짓는데 이를 위하여 비슷한 응답성향점수를 가진 자료에 근거하여 몇 개의 대체군을 형성하고 각 대체군 내에서 무응답자의 값을 같은 대체군 내의 응답자의 값으로 대체시키는 방법을 제안하였다. 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 대체방법을 시행하는 SAS Macro의 사용방법은 <그림 5.3>에 나타난다.

<그림 5.3> SAS Macro %hotdwr

```
%hotdwr (y, my, xlist, data, id, seed, replace=yes, iy= _iy, by=, class=,
         mincell=10, out=_outhot, weight=_one, where=1, maxvar=999,
         keep= ,print=n, keepdonr=n,lower=_mis,upper=_mis)
```

%hotdwr를 사용하기는 데 필수기입사항은 다음과 같다

```
y = variable to be imputed
my = missing indictor created by the macro
xlist = list of variable used to impute y (may include y)
data = name of SAS data set
id = unique ID variable
seed = seed for random number generator
```

그 외의 선택모수들은 다음과 같다.

```
replace = whether y is replaced (yes)
iy = name of variable that contains imputations (if replace=no)
by = by variable for regressions (none)
class = variable defining strata for imputation cells (none)
mincell = minimum number of donors per cell before printing
          caution (10)
out = output data set, contains old variables, imputed values, and
      my   (_outhot)   weight = weights to allow unequal sampling
      probabilities within cell, useful for multiple imputation, must be
      a variable (constant)
```

```
where = expression for subsetting observations (all are used)
maxvar = maximum number of nonmissing variables used from xlist
         starting at the beginning (999)

keep = additional variables need for where or other reasons (none)
print = execute the proc print statements y/n (n)
keepdonr = keep the variable _donor in final dataset y/n (n)
lower = lower bound of imputed values with regression imputation
        of observations without donors  (none)
upper = upper bound of imputed values with regression imputation
        of observations without donors  (none)
```

5.2.2 비선형 회귀모형에 근거한 대체 방법

4.1절에서 고려한 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 회귀모형에서 고려할 수도 있으나 좀 더 강건한(robust) 방법으로는 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 $X_1, X_2, \ldots, X_p$가 있다

면

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

의 회귀식을 고려하고 함수 $g_i(\bullet)$, $i = 1, 2, ..., p$를 스플라인이(spline)나 커넬(kernel)을 이용하여 완전히 관찰된 자료를 이용하여 추정하는 방법이다.

다차원의 문제를 해결하기 위한 다른 방법으로 Little and An(2004)은 응답성향(response propensity)을 이용한 차원의 축소를 제안하고 이를 penalized spline imputation 방법이라고 하였다. 자료에 결측이 없는 p개의 변수, $X_1, X_2, ..., X_p$가 존재하고 1개의 변수 Y에서 결측이 발생한다고 하면 결측자료 메커니즘이 임의결측일 때 응답성향은 $P(R = 1 | X_1, ..., X_p) = X_1^*$ 로 정의 되고 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 $X_1, X_2, ..., X_p$로 회귀하는 대신 Y를 X_1^*로 회귀하여도 편향이 없는 추정치를 얻을 수 있게 된다. 즉 p차원을 1차원으로 축소한 후 비모적인 회귀식을 적합하는 방법이다. 이 방법은 강건한 모형을 바탕으로 대체값을 예측하기 때문에 변수 사이의 관계가 선형이 아니더라도 또한 변수의 차원이 높더라도 응답성향의 모형이 올바르게 선택되고 결측 메커니즘이 무시할 수 있는 메커니즘이라면 편향이 없는 추정치를 구할 수 있다. 이 방법은 현재까지는 상용화된 프로그램에 포함되어 있지 않다.

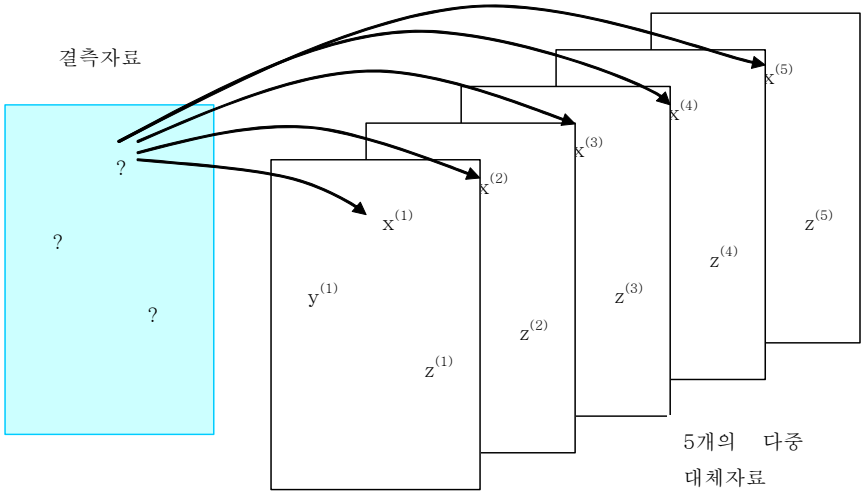
5.3 다중대체된 자료에 대한 분석 및 추론

3장에서 단일대체보다 다중대체를 실시하면 한 개의 결측값에 대하여 한 개의 값

으로 대체하는 대신 타당한 여러 개의 값을 가지고 대체하므로 대체된 값들은 각각 다르고 이 점은 대체된 값이 알려져 있지 않은 불확실성(uncertainty)을 모형에 반영시켜 한 개의 값을 대체하는 방법인 단일대체(single imputation)에서 발생하는 추정량의 편의를 보정하는 것을 가능하게 한다.

무응답을 포함한 자료에 대하여 다중대체를 실시하면 결과로서 여러 개의 결측값이 없는 대체된 자료가 만들어지게 된다. 여러 개의 자료들은 관찰된 값들은 모두 동일하지만 대체된 무응답값은 같기도 하고 다르기도 한 형태를 가지고 있다. 무응답을 포함한 자료에 대하여 다중대체를 시행한 후 생성된 대체된 자료는 다음의 <그림 5.4>에 나타난 것과 같이 여러 개가 존재하게 된다.

<그림 5.4> 무응답을 포함한 결측 자료에 대하여 5개의 다중대체를 실시한 경우의 예



다중대체에서 대체의 숫자가 m이면 연구자는 m개의 대체된 자료 각각에 대하여

원하는 분석을 반복적으로 시행할 수 있다. 각 자료에 대하여 독립적으로 분석이 시행된 후 분석 결과는 일반적으로 개의 통계량 및 관련 분산(또는 표준 오차)으로 나타나는데 연구자는 m개의 각각 다른 통계량이 아닌 하나의 통합된 통계량을 구하는 데 목적이 있다. 이 목적은 (1) 다중대체된 각 자료의 분석, 그리고 (2) 분석된 자료를 통합한 결과 도출의 두 단계를 통하여 달성된다.

5.3.1 다중대체(multiple imputation)된 자료의 분석

다중대체된 자료 각각은 무응답이 대체되어 결측값이 없는 완전한 자료 형태를 가지고 있으므로 자료 각각에 대하여 연구목적에 알맞은 분석을 시행하면 된다. 예를 들어, 회귀분석을 시행하고자 한다면 동일 관심변수에 대하여 동일 설명변수를 가지고 m개 자료 각각에 대하여 회귀분석을 실시하면 된다. 이렇게 분석을 실시하는 경우 추정된 회귀계수(regression coefficients), 표준오차(standard errors), 그리고 검정통계량(test statistics)은 m개 자료 각각으로부터 약간씩 다르게 나타나는데 이는 관심 변수가 결측되어 참값을 알지 못하는 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.

5.3.2 분석된 자료를 통합한 결과 도출

다중대체를 m번 시행한 자료 각각에 대하여 분석을 시행한 후 얻어진 모수의 추정값들을 $\widehat{\theta_1},\widehat{\theta_2},...,\widehat{\theta_m}$ 이라 정의하고 이 모수들의 추정된 분산을 각각

$W_1, W_2, ..., W_m$이라 정의하자. 예를 들어, 회귀분석을 실시하면 i번째 자료에 근거한 회귀 분석에서 관심 설명 변수의 회귀계수의 추정값 벡터는 $\hat{\theta}_i$이 되고 그 회귀계수의 표준오차의 추정값의 제곱이 W_i 행렬로 표현된다. 이 경우 통합된 모수의 추정값은

$$\bar{\theta} = \frac{1}{m}\sum_{i=1}^{m}\hat{\theta}_i$$

으로 표현될 수 있다. 즉, 추정된 모수들의 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 각 모수의 추정된 분산들의 평균이다. 이 분산 성분은 대체내분산 (within imputation variance)으로 부른다. 두 번째 분산 성분은

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

으로 표현되는데 이는 각 대체된 자료의 모수의 추정값들 사이의 분산을 나타내므로 대체간분산(between imputation variance)이라 부른다. 통합된 모수의 분산의 추정값은

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

으로 구할 수 있다.

자료가 충분히 큰 경우, 이 모수에 대한 분포는 다음의 t-분포를 따른다.

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

여기서, t-분포의 자유도 v는 $\nu = \left(\nu_0^{-1} + \widehat{\nu_{obs}^{-1}}\right)^{-1}$로 계산되는데 $v_0 = (m-1)\left(1 + \frac{1}{m+1}\frac{\overline{W_m}}{B_m}\right)^2$을 의미하고 $\widehat{\nu_{obs}} = \left(1 - \widehat{\gamma_m}\right)\left(\frac{\nu_{com}+1}{\nu_{com}+3}\right)\nu_{com}$으로 나타나는데 여기서 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 결측에 의하여 손실된 모수 θ에 대한 정보의 부분(fraction of information about θmissing due to nonresponse)이라 불린다. 모수의 분포가 t-분포를 따르므로 모수 θ에 관한 t-검정을 시행하거나 모수의 신뢰구간(confidence interval)을 구할 수 있다. 이 통합 방법은 관심 모수들에 대한 다변량 검정 및 신뢰구간의 계산 등으로의 확장도 가능하다 (Rubin, 1987a).

예제 5.3 고령화연구패널 제 1차 조사에서 금융자산액의 평균

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, SPLUS와 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 있다 (프로그램은 http://www.stat.psu.edu/~jls/misoftwa.html에서 무료로 다운받을 수 있다. 다음은 SAS에서 다중대체된 5개의 대체자료를 이용한 단순평균의 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 5.5>는 다중대체된 금융자산(w01f085)의 단순 평균 계산을 위한 SAS 프로그램을 보여준다.

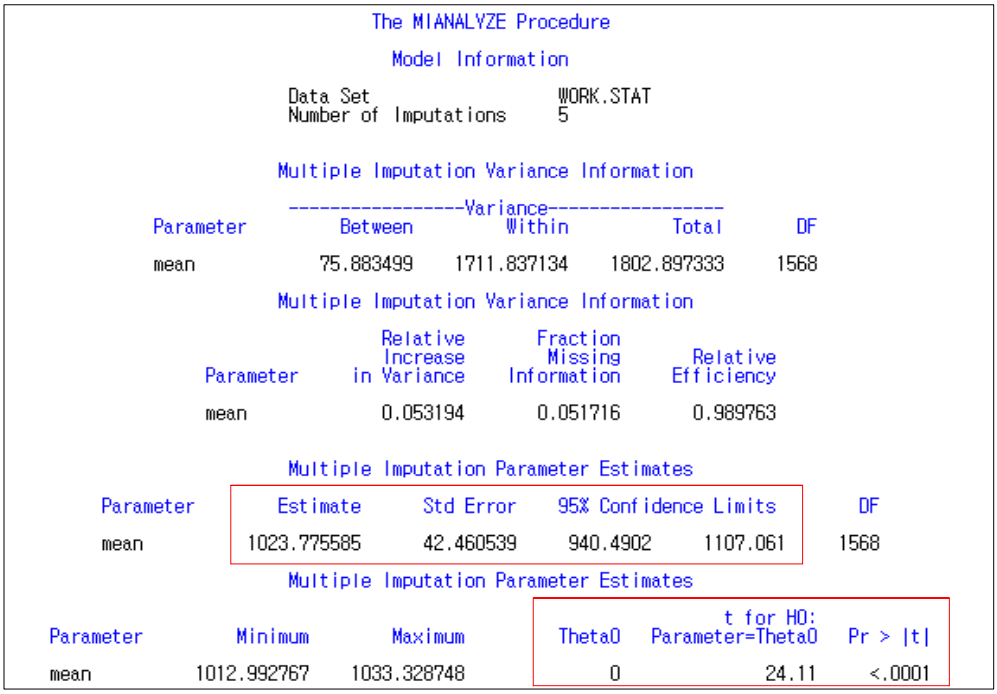
<그림 5.5> 다중대체된 금융자산(w01f085)의 단순 평균 계산을 위한 SAS 프로그램

```
* 각 대체 자료별 단순 평균 계산;
PROC SURVEYMEANS DATA = total;
   VAR w01f085;
   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 SURVEYMEANS에서 BY변수를 사용하여 각 대된 자료별로 금융자산의 단순평균 및 표준 오차를 계산한 후 이들 통계량들을 자료(data set)명 stat1에 저장하였다. 이 저장된 통계량들은 Procedure MIANALYZE를 사용하여 통합되었다. 이 때 MODELEFFECTS 문에는 통합할 통계량 $\hat{\theta}_i$(여기서는, 평균)을 나타내는 변수 mean을 써 주고 STDERR문에는 W_i 의 제곱근인 평균의 표준 오차를 나타내는 stderr 변수를 써 주면 된다. Procedure MIANLYZE는 <그림 5.6>과 같은 결과를 제공한다.

<그림 5.6> 다중대체된 금융자산(w01f085)의 단순 평균 계산을 위한SAS 결과



The MIANALYZE Procedure

Model Information

Data Set	WORK.STAT
Number of Imputations	5

Multiple Imputation Variance Information

Parameter	Between	Within	Total	DF
mean	75.883499	1711.837134	1802.897333	1568

Multiple Imputation Variance Information

Parameter	Relative Increase in Variance	Fraction Missing Information	Relative Efficiency
mean	0.053194	0.051716	0.989763

Multiple Imputation Parameter Estimates

Parameter	Estimate	Std Error	95% Confidence Limits		DF
mean	1023.775585	42.460539	940.4902	1107.061	1568

Multiple Imputation Parameter Estimates

Parameter	Minimum	Maximum	Theta0	t for H0: Parameter=Theta0	Pr > |t|
mean	1012.992767	1033.328748	0	24.11	<.0001

<그림 5.6>에서 보는 바와 같이 금융자산의 단순 평균은 1023.78 MW으로 타나나고 표준편차는 42.46이다. 금융자산의 평균에 대한 95% 신뢰구간은 (940.49, 1107.06)으로 계산되며 금융자산이 0이라는 귀무가설은 t-통계량이 24.11, p-value가 <.0001로 5% 유의수준 하에서 유의하게 나타난다.

<5장 연습문제>

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

(가) 완전임의 핫덱대체는 무응답 메커니즘이 임의결측인 경우 추정량에 편의가 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 편의가 발생하지 않도록 하기 위함이다.

2. 다음의 자료는 2005년 인구주택총조사 자료의 일부분이다. 제공된 자료에서 무응답은 연습문제를 위하여 임의로 생성되었다. 4장에서 학습한 대체방법들을 사용하여 대체를 실시하시오. 사용된 대체 방법들을 변수별 평균 및 표준오차를 사용하여 비교하고 이 자료에 대하여 한 가지 대체 방법을 추천하시오.

제 6장. 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) 무응답이 있는 경시적 자료에 대한 우도에 근거한 분석법을 소개한다.

(4) 무응답이 있는 경시적 자료에 대한 대체법 분석을 소개한다.

(5) 여러 가지 실제 패널조사 예제를 통하여 분석법을 알아본다.

6.1 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 이동과정을 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 달라진

다.) 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.

6.2 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 웨이브

에서 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 6.1>은 다섯 개의 웨이브로 구성된 패널 조사에서 웨이브 응답 (X)과 무응답 (0)의 몇 몇 전형적인 패턴을 보여준다.

<표 6.1> 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 6.1>에 나타나지 않은 한 가지 무응답 패턴은 모든 웨이브에서 무응답

이 일어난 경우이다. 많은 패널조사에서 첫 번째 웨이브에 무응답을 한 표본 개체들은 더 이상 추적조사하지 않는 경우가 흔하다. 대부분의 패널조사는 웨이브 무응답자를 언제 그만 추적할 것인 지에 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 6.1>에서 패턴 6은 가능하지만 패턴 7과 8은 일어날 수 없다.

<표 6.1>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 6.1>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 2에서는 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 패

턴 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 패턴을 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 6.1>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.

6.3 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 말한바와 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 4

개월간의 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 6.3.1 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 평균

보다 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 Verbeek 와 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 6.3.2 National Longitudinal Survey of Youth (NLSY79)

NLSY79는 미국의 14-22세 청소년을 모집단으로 1979년에 처음 표본조사를 실시한 패널 조사이다. 이 조사는 미국 노통 통계국(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 연방법에 의해 청소년들에게 고용과 교육의 기회를 확대됨에 따라 NLSY79는 미국

인들의 노동 시장 경험에 관한 이전 연구들의 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 횟수, 부모의 결혼 유지 여부, 추적(follow-up)시의 어머니의 나이를 이용하였다. 무응답이 분석의 결과변수와 스틱이 ▯ 여러 가지 가정하여 (즉, MAR 무응답을 가정) 다중대체법과 완전히 응답한 개체를 이용한 분석법을 비교하였다.

예제 6.3.3 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 건강 기능상태와 장애의 변화 및 이러한 변화의 위험요소를 조사하는 것이다. 연구가

진행되는 동안 많은 수의 참여자들이 하나 이상의 웨이브에서 무응답하였으나 이후의 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 추정치의 표준오차를 추정함으로써 복잡한 디자인의 특성을 고려하였다.

제 7장. 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.

7.1 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 전국 전인구를 대상으로 통계청에서 실시하는 조사로서 대한민국 영토 중 행정권이 미치는 전 지역을 대상으로 조사기준 시점 현재 조사지역 내에 상주하는 내, 외국인 및 이들이 살고 있는 모든 거처에서 조사가 실시된다. 본 연구에서는 2005년 인구주택총조사를 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. 이외에 추가로 16개 시, 도별로 각각 서로 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 구분이 된다. 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)로 나누어진다. 단위무응답은 가중값을 사용하여 처리하였고 항목무응답은 통계청에서 무응답대체를 통하여 완전한 형태의 자료를 만들어 제공해 왔다. 본 사례에서는 2005년 인구주택 총조사에 사용된 대체 방법을 무응답 대체를 위한 실제 사례로서 설명하고자 한다.

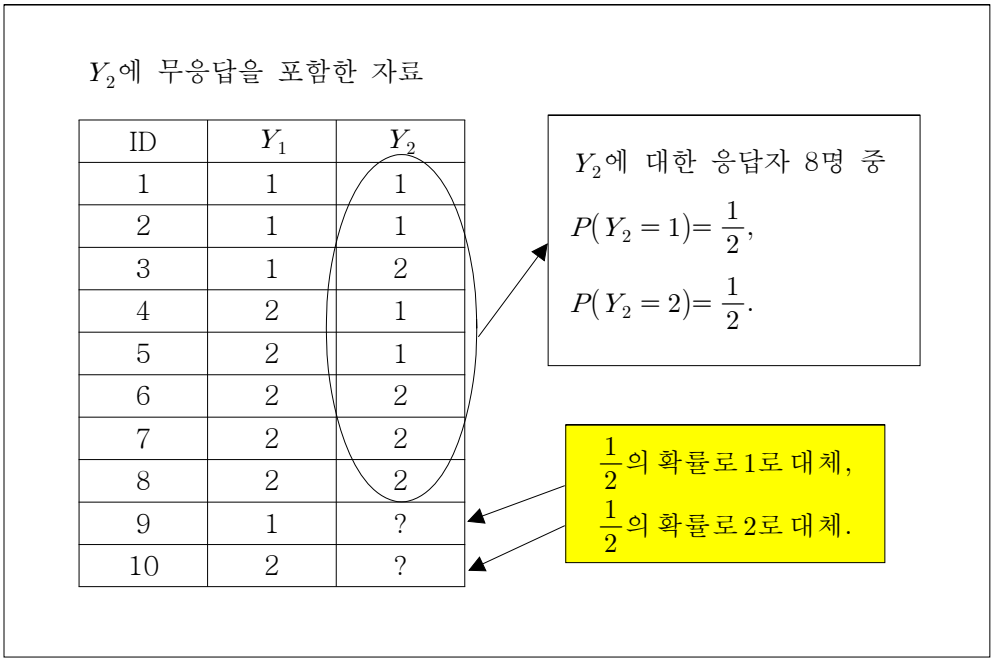
7.2 2005년 인구주택총조사에서 사용된 무응답 처리 기법

거의 대부분의 항목에서 무응답이 발생하고 있으며 다음의 세 가지 대체방법이 적용되었다.

7.2.1 확률에 근거한 대체(Probability Imputation)

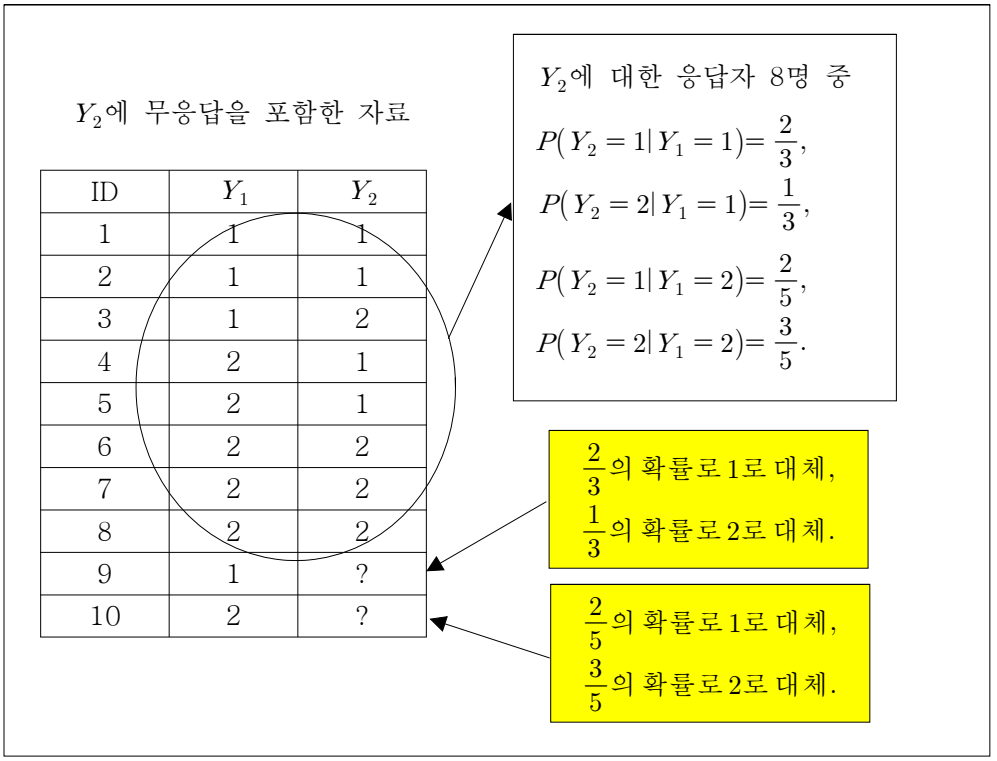
이 대체법은 무응답을 포함한 범주형 변수에서 각 범주별 분포를 이용한 방법으로 범주별로 응답자들의 분포에 따라 무응답자를 확률적으로 배분하는 방법이다. 이 방법은 응답자의 범주별 비율을 대체에 사용하므로 변수가 범주형으로 측정된 경우에만 사용 가능하다. <그림 7.1>은 무응답이 1개의 변수 에서만 발생하는 경우 확률에 근거한 대체를 실시하는 방법을 설명한다. 우선, 무응답이 발생한 Y_2 변수의 응답값들 중 각 항목의 비율의 계산한다. 전체 10명의 관찰자료 중 8개의 관찰자료에서 응답이 발생하였고 그 중 4개가 "1"로, 4개가 "2"로 응답되었다. 즉, 응답자 중 "1"이 발생할 확률 $P(Y_2=1)=\frac{4}{8}=\frac{1}{2}$이고 응답자 중 "2"가 발생할 확률 $P(Y_2=2)=\frac{4}{8}=\frac{1}{2}$로 동일하게 나타난다. 이 확률들에 비례하게 무응답에 대한 대체가 발생한다는 의미로 확률에 근거한 대체라 부른다. 즉, 이 예에서는 "1"과 "2"의 발생확률이 동일하므로 각각 $\frac{1}{2}$의 확률로 "1"과 "2"로 대체하게 된다. 실제로 대체를 실행할 때는 0과 1사이의 난수(random number)를 생성하고 이 값이 $\frac{1}{2}$보다 작으면 "1"로, 이 값이 $\frac{1}{2}$보다 크면 "2"로 대체값을 할당한다.

<그림 7.1> 확률에 근거한 대체방법



<그림 7.1>에서 설명한 대체는 무응답 자료 메커니즘이 완전임의결측인 경우에만 추정량에 편의가 발생하지 않는다. 물론 완전임의결측인 무응답 메커니즘은 현실에서 만족되는 자료의 숫자가 많지 않다. 따라서 대체군 내에서 응답값과 무응답값의 분포가 동일하도록 대체군을 형성하여 대체를 실시하는 것이 바람직하다. 즉, 3.2.2절에서 언급한 바와 같이 대체군을 이용한다면 더 정확한 대체를 실시할 수 있을 것이다. 이 경우 대체군에 포함된 변수들이 주어졌을 때 무응답 자료 메커니즘이 임의결측인 경우에도 추정량에 편의가 발생하지 않는다. <그림 7.2>는 대체군을 사용한 확률에 근거한 대체방법에 대한 예를 설명한다.

<그림 7.2> 대체군을 사용한 확률에 근거한 대체방법



Y_2 변수에 무응답이 발생하였고 Y_1 변수로 대체군을 형성한 경우를 고려하자. 우선, 무응답이 발생한 Y_2 변수의 응답값들 중 대체군을 내에서 각 항목의 비율의 계산한다. 전체 10명의 관찰자료 중 8개의 관찰자료에서 응답이 발생하였고 그 중 $Y_1=1$인 대체군을 고려하면 대체군내에는 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 존재하는 데 그 중 2개가 “1”로, 나머지 3개가 “2”로 응답되었다. 즉, $Y_1=2$인 대

체군 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. 즉, 이 예에서는 $Y_1=1$인 대체군 내에서 “1”과 “2”의 발생확률이 각각 $\frac{2}{3}$와 $\frac{1}{3}$이므로 $\frac{2}{3}$의 확률로 “1”로, 그리고 $\frac{1}{3}$의 확률로 “2”로 대체하게 된다. 마찬가지고 $Y_1=2$인 대체군 내에서 “1”과 “2”의 발생확률이 각각 $\frac{2}{5}$와 $\frac{3}{5}$이므로 $\frac{2}{5}$의 확률로 “1”로, 그리고 $\frac{3}{5}$의 확률로 “2”로 대체하게 된다.

인구주택총조사에서는 각 항목별 발생 확률을 추정하기 위하여 전체조사 자료와 유사한 환경을 구축하여 분포도를 산출한다. 즉, 조사 단위들이 동질적인 특성이 있는 대체군으로 분해하여 대체확률을 구하도록 하는 것이다. 대체에 사용된 변수들은 주로 무응답을 포함한 변수 Y를 잘 설명해 줄 수 있는 변수들로 선택하여 각 대체군 내에서 대체확률을 산출하는데 이 변수들은 주로 무응답을 포함한 변수 Y를 동질성(homogeneity) 있게 분류할 수 있는 변수가 바람직하다. 관심변수의 분포도는 응답자들을 이용하여 주로 구하여 지는데 전체 조사대상자에 비하여 무응답자가 무시하기에는 너무 높은 비율을 차지하는 경우에는 과거 조사결과 혹은 유사한 다른 조사의 결과를 이용하여 대체확률을 구하여 이용할 수도 있다. 예를 들면, 인구센서스는 과거의 경험으로 보아 각 문항에서의 항목무응답의 비율이 높지 않으므로 금번 조사 응답자로부터 대체확률을 산출하였다.

7.2.2 핫덱대체

5.2절에서 설명한 핫덱대체 방법은 연속변수 및 불연속변수에 범용적으로 쓰이는 가구관련 통계조사의 무응답 처리기법으로 주로 사용하는 대체방법인데 이 방법을 이용하는 통계기관마다 사용하고자 하는 목적에 따라 다른 기법을 적용한다.

무응답을 포함한 변수(imputing variable) Y에 대한 응답자들을 모아 기증자 풀(donor pool)을 만들고 그 가운데 무응답자 수만큼 무작위로 추출하여 각 무응답자에게 무작위로 한 명의 응답자(donor)를 할당하여 응답자의 해당변수의 값을 일대일로 대체하는 방법이다. 인구주택총조사에 대한 핫덱대체에는 응답자와 무응답자를 Y를 잘 설명할 수 있는 변수들로 구성된 대체군별로 그룹화하여 각 대체군 내에서 각 응답자에게 균등확률분포(uniform distribution)에서 생성된 무작위 확률값(0과 1 사이의 값)을 각각 할당하고, 동일 대체군 내의 각 무응답자에게 균등확률분포에서 생성된 무작위 확률값을 각각 할당하여 이들을 각각 할당 받은 확률값을 순서로 나열한 다음 무응답자수에 해당하는 개수를 응답자의 상위부터 배열된 순서로 하나하나 할당하는 방법을 사용하였다. 예를 들어, '교육상태' 변수를 대체하고자 하는 경우 '성별', '연령'이 '교육상태'를 설명할 수 있는 대체군을 형성하는 변수라고 가정하자. 남자(1)이며 연령이 15~19세 대체군의 응답자가 100명, 무응답자가 3명이라면 응답자 100명에 대하여 균등분포에서 생성된 확률값을 할당하여 크기순으로 정렬(sort)하고, 3명도 균등분포에서 생성된 확률값을 할당하고 크기순으로 정렬하여 <표 7.1>과 같이 응답자와 무응답자의 그룹이 확률값에 따라 정렬되었다고 하자. 응답자를 처음부터 3명을 무응답자에게 차례로 각각 한명씩 할당하는 방법이다. 즉, 응답자 'A11'의 교육상태 '1'은 무응답자 'A51'에게 할당하고, 응답자 'A31'의 학력상태 '3'은 무응답자 'A19'에게 할당하고 마지막으로 응답자 'A22'의 교육상태 '2'는 무응답자 'A33'에게 할당한다.

<표 7.1>대체군내 응답자 그룹과 무응답자 그룹에 할당된 확률값 (최필근, 2008)

응답자 그룹					무응답자 그룹				
ID	Imputation class		교육 상태	확률값	ID	Imputation class		교육 상태	확률값
	성별	연령				성별	연령		
A11	1	15-19	1	0.002	A51	1	15-19	·	0.541
A31	1	15-19	3	0.012	A19	1	15-19	·	0.620
A22	1	15-19	2	0.035	A33	1	15-19	·	0.655
A54	1	15-19	5	0.120					
A44	1	15-19	2	0.223					
…	…	…	…	…					

여기서 사용된 핫덱대체는 응답자를 비복원으로(without replacement) 추출하기 때문에 각 대체군에 속하는 응답자수는 항시 무응답자수 보다 커야한다. 따라서 대체군에 사용될 변수를 선정할 때 미리 무응답자 숫자에 비교하여 응답자수 가 충분한지 검토되어야 한다.

7.2.3 계층적 핫덱대체 (Hierarchical Hotdeck)

이 방법은 3.2.2절에서 고려한 대체군을 이용한 핫덱대체 기법에서 대체군을 만드는 변수들이 많아져 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) Flexible Matching Imputation method라고도 한다. 핫덱대체 시 응답자 중에서 기증자를

비복원으로 추출하기 위하여 대체군 내의 응답자수가 무응답자수보다 훨씬 커야 하는데 자료의 구조와 특성상 일부 대체군에서 무응답자의 수가 응답자보다 많은 경우 사용하기 위하여 고안된 방법이다.

핫덱대체에서는 모든 고려된 공변량들의 항목값들을 조합하여 대체군을 형성한다. 이 대체군을 대체 레벨-1 (level-1) 대체군이라 부른다. 이 대체군내에서 무응답자는 응답자의 그룹으로부터 비복원 무작위 추출로 선택된 응답자로부터 Y의 값을 기증받는다. 만약에 레벨 1 대체군의 일부에서 무응답자수가 응답자수보다 많은 경우는 일부 무응답자는 응답자로부터 값을 할당 받지 못하게 되므로 이 경우 공변량을 한 개 제거하고 나머지 공변량들로 레벨-1 (level-1) 대체군을 형성한다. 이렇게 형성된 레벨-2 대체군내에서 무응답자는 응답자의 값을 비복원 무작위 추출로 할당받는다. 이때 레벨-1 대체에서 무응답자에게 그들의 값을 기증한 응답자는 제외하고 대체군을 구성하여 한 명의 응답자가 여러 번 그의 값을 기증하는 것을 막을 수 있다. 이러한 절차를 모든 무응답자에 대하여 대체 가 이루어질 때까지 반복한다. 이 방법도 전체 무응답자의 수가 전체 응답자의 수보다 큰 경우에는 적용하는 데 한계가 있다.

대체군을 형성하기 위하여 고려된 변수들이 성별(SEX), 교육(EDUCATION), 그리고 결혼상태(MARRIAGE) 3개인 경우를 고려하자. 성별이 "남"(1로 표현), "여"(2로 표현) 2개의 범주를 가지고 교육은 5개의 범주를, 결혼상태는 4개의 범주를 가진다면 대체군의 개수는 $2 \times 5 \times 4 = 40$개가 된다. 즉, 레벨-1 대체군의 수는 최대한 40개가 가능한데 일부 대체군 내에서 무응답자가 값을 할당 받지 못하면 다음 단계에서는 최단 변수인 결혼상태를 제거하고 레벨-2 대체군을 형성한다. 이 때 형성된 대체군의 숫자는 $2 \times 5 = 10$개가 되며 앞 레벨에서 대체되지 못한 무응답자가 새롭게 형성된 레벨-2 대체군 내에서 대체할 값을 기증받게 된다. 이

러한 작업을 모든 무응답자가 값을 할당 받을 때까지 반복한다. <표 7.2>는 이 방법을 구체적인 예를 들어 설명한다.

<표 7.2> 계층적 핫덱 대체의 예

항상 유지	Level-3에서 제거	Level-2에서 제거	응답자수	무응답자수
SEX	EDUCATION	MARRIAGE		
1	대졸	미혼	2	4
1	대졸	기혼	8	2

(1) 레벨-1 대체

우선, 레벨-1 대체군에서 성별이 “남”(1)이고 교육이 “대졸”이며 결혼상태가 “미혼”인 대체군 내에는 4명의 무응답자가 존재하지만 응답자는 2명밖에 없다. 따라서 4명의 무응답자 중 2명은 2명의 응답자로부터 무작위로 응답된 Y의 값을 기증받아 대체한다. 2개의 무응답자는 응답자의 값을 할당받지 못하였다. 한편, 다른 대체군인 성별이 “남”(1)이고 교육이 “대졸”이며 결혼상태가 “기혼”인 대체군 내에는 무응답 자 숫자는 2명이고 응답자 숫자는 8명이므로 8명의 응답자로부터 무작위로 2명을 추출하여 그들의 Y의 값을 기증받는다.

(2) 레벨-2 대체

레벨-1에서 성별이 “남”(1)이고 교육이 “대졸”이며 결혼상태가 “미혼”인 대체군에서 2개의 무응답자는 대체가 실시되지 않았다. 따라서 결혼상태 변수를 제거하고 레벨-2 대체군을 형성한다. 이 대체군 내에서는 <표 7.3>의 두 대체군이 성별이 “남”(1)이고 교육이 “대졸”인 한 개의 대체군으로 묶이게 된다. 이 대체 내에는 성별이 “남”(1)이고 교육이 “대졸”이며 결혼상태가 “미혼”인 2명의 무응답자가 대체

를 기다리고 있고 성별이 “남”(1)이고 교육이 “대졸”이며 결혼상태가 “기혼”인 6명의 6명의 응답자가 대체값을 기증할 수 있게 된다. 즉, 2개의 무응답자는 이 6명의 응답자로부터 무작위로 2명을 추출하여 대체할 값을 기증받는다.

인구주택총조사 자료에 대한 계층적 핫덱대체를 시행하기 위하여 공변수들은 무응답을 포함한 변수 Y와 상관관계가 높은 범주형 변수들을 고려하였고 상관이 가장 높은 변수를 가장 마지막에 제거되도록 지정하여 대체군의 레벨이 높아질 때 상관이 가장 약한 변수를 제거대상이 되게 함으로써 가장 강한 상관관계의 변수는 최후까지 대체군의 형성에 기여하도록 하였다. 변수의 제거 순서의 결정은 전문가와 상의하여 결정되었다. 이현정(2009)은 이 순서의 결정을 위한 방법을 제안하고 모의실험을 통해 그 성능을 비교하였다.

7.2.3.1 특이점(outlier)의 제거

핫덱대체에서 대체군을 설계하기 전에 특이점에 대한 제거작업을 수행하도록 설계되었다. 레벨-1 대체군에서 특이점을 제거한다면 대체군의 숫자가 많아 응답자 수가 충분하지 않아 특이점을 결정하기 어려울 수 있다. 따라서 특이점은 응답자 전체 중에서 새롭게 특이점 군집(outlier class)을 형성하여 그 군집내에서 이상치를 제거하도록 설계하였다. 특이점 제거를 목적으로 구성하는 특이점 군집의 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.

7.3 2005년 인구주택총조사 변수들 및 무응답 대체 방법

인구주택총조사 변수들은 크게 (1) 가구원 관련 사항, (2) 가구에 관한 사항, 그리고 (3) 주택에 관한 사항으로 나눠진다. 이현정(2009)은 (2) 가구에 관한 사항, 그리고 (3) 주택에 관한 사항 별로 각 변수에 대하여 무응답 처리 여부, 무응답 처리를 실시한 경우 사용된 대체 방법, 그리고 대체시 사용된 대체군을 표로 나타낸다.

제 8장. 사례연구 II

- 네덜란드 POLS 조사연구 -

<학습목표>

(1) 네덜란드 POLS 조사연구에 대하여 설명한다.

(2) POLS조사에서 발생하는 무응답을 처리하는 기법을 설명한다.

(3) 여러 가지 가중방법을 비교한다.

8.1 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.

8.2 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 8.1>은 성별, 나이, 결혼유무에 관한 모집단 결합분포이다.

<표 8.1> 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 8.2>와 <표 8.3>)

<표 8.2> 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 8.3> 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 8.2>와 <표 8.3>에서는 나이가 5수준으로 구성되어 있으나 <표 8.1>은 나이가 8수준으로 구성되어 있다. 선형 가중방법에서는 같은 변수의 다른 범주화가 문제가 되지 않는다. 두 개의 나이 범주 변수를 동시에 고려할 수 있다. 사후-층화 방법은 이 표 중에서 단지 하나만을 이용할 수 있다. 또한 <표 8.1>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 8.4>)

<표 8.4> 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 8.5>는 이 사례 연구에서 적용된 여러 가지 가중 모형으로부터 얻은 결과를 보여준다.

<표 8.5> 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 8.5>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 가중치는 불안정하다. 그래서 Sex × Marital Status × Age8을 이용한 사후층화분석 대신에 선형 가중 모형 (Sex×Age8)+(Sex×Marital)을 이용하였다.

<표 8.5>의 모형 11은 최대로 가능한 보조변수의 조합을 고려한 모형이다. (Sex×Age8)+(Sex×Marital)에 해당하는 모집단의 정보는 <표 8.1>로부터 (Age5×Urbanization)에 해당하는 모집단의 정보는 <표 8.3>으로부터 얻을 수 있다. 한 가지 주목할 점은 Age5 × Province의 몇 몇 칸의 빈도가 너무 작아서 모형에서 Age5 × Province 대신에 Province의 정보만 <표 8.4>로부터 사용하였다. 이 모형 11을 적용한 가중치는 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 9장. 사례연구 III

- 2006년 고령화연구패널 제 1차자료에 대한 무응답 대체기법 -

<학습목표>

(1) 고령화연구패널조사에 대하여 설명한다.

(2) 고령화연구패널조사에서 발생하는 무응답을 처리하는 기법을 설명한다.

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

9.1 고령화연구패널조사 개요

고령화연구패널조사(KLoSA)는 고령화 시대에 대비하기 위하여 한국의 고령자 모집단에 대한 실태를 파악하고 기초 자료를 수집하기 위하여 계획되었고 1차 조사가 2006년에 실시되었다. 고령화연구패널 1차 조사는 45세 이상인 10,254명의 일반가구 거주자에 시행되었고 동일 참여자에 대한 매 2년마다 추적조사도 계획되어 고령인구집단의 변화 추세에 대한 우수한 정보를 축적할 수 있을 것으로 기대된다. 이 자료는 8개 부문으로 나누어 인구학적 배경 및 가족 관계, 건강과 고용상태, 소득과 자산 상황, 그리고 삶의 만족도 등 심리적 상태까지 1310개의 다양한 조사 항목을 포함하고 있다. 다음은 각 부문별 조사항목에 관한 개요이다.

CV. 커버스크린

조사대상으로 적합한 지 확인하기 위한 변수들로 구성되어 있으며 가족별로 대표자에 대해서만 조사되었다.

A. 인구학적 배경

인구학적 배경 관련 변수들은 생년월일, 학력, 혼인상태, 종교 등으로 구성되어 있다.

B. 가족

가족 관련 영역에서는 자녀에 대한 자세한 항목으로 구성되어 있다. 가족 영역의 경우 가족별 세대 대표자가 선정되어 그 대표자에 대해서만 조사되었고, 나머지 사람들은 대표자의 값으로 대체되었다.

C. 건강

건강영역은 건강상태(Ca), ADL과 간병인(Cb), 의료보장과 시설 이용(Cc), 인지력(Cd), 신체기능(Ce)으로 구성되어 있다.

D. 고용

고용영역에서는 응답자의 경제 활동 상태에 따라서 임금근로자, 자영업자, 무급가족종사자, 은퇴자, 구직자로 나누어 해당하는 고용 상태에 따라 고용 상태나 임금, 퇴직금 등에 관한 설문이 진행되도록 구성되었다.

E. 소득

소득 관련 영역은 다양한 분야의 소득 여부 및 소득 액수 및 실업급여 등에 관한 질문들로 구성되어 있다.

F. 자산

자산영역은 부동산 총자산, 사업체/농장 총자산, 금융 총자산, 기타 총자산, 개인 총부채, 소득 이렇게 여섯 개의 하위영역(subsection)으로 나눌 수 있다. 각

영역별로 자산 및 부채 액수와 관련된 질문들로 구성되어 있다.

G. 주관적 기대감 및 삶의 만족도

유산, 고용, 기대 수명 등에 관한 기대감 및 건강, 경제, 가족 및 삶에 대한 만족도에 관한 질문들로 구성되어 있다.

9.2 2006년 인구주택총조사에서 사용된 무응답 처리 기법

결측값의 대체를 위하여 수정된 예측 평균에 근거한 핫덱 방법(hotdeck based on a modified predictive mean matching)이 사용되었다(Little, 1988; Bell 1999). 이 방법은 여러 가지 조사 연구에서 우수한 결과를 보여 온 대체방법으로서 결측이 발생한 자료값을 자료 내 관찰된 값 들 중 하나 또는 여러 개의 값으로 대체시키는 일종의 핫덱대체법이지만 관찰값 중 하나 또는 여러 개의 값을 임의로 선택하는 완전임의 핫덱 대신 자료를 비슷한 특성을 가진 여러 개의 하위 그룹(subclass)으로 나누어 같은 그룹 내에서 핫덱 대체를 실시한다. 이 때 하위그룹은 결측이 발생한 변수에 대하여 관찰된 자료만을 대상으로 회귀모형을 적합하여 결측이 포함된 모든 자료에 대한 예측값을 구한 후 예측값에 근거하여 층화(stratification)를 하여 구성한다. 각 층 내에서 결측값은 같은 층의 관찰자 중에서 기증자를 선택하여 기증자의 값으로 대체를 실시한다. 이 방법은 기증자를 선택하는 데 있어서 임의로 한 명 또는 여러 명의 기증자를 선택하는 단순임의 핫덱방법보다 회귀모형의 예측력이 클수록 좋은 결과를 기대할 수 있다.

고령화연구패널조사의 대체를 실시할 때 몇 가지 특징이 발견되었고 그에 따라 대체 방법이 적절히 변형되었다. 첫 번째로 일부 소득 및 자산 항목은 응답이 거

절되거나 응답 문항들 사이에 불일치가 나타나는 경우 대괄호 질문들(unfolding brackets)을 이용하여 얻어진 부분 정보를 포함하고 있다. 두 번째로 응답자가 많지 않은 일부 문항의 경우 대괄호 질문으로부터 얻어진 부차 정보에 근거한 하위그룹(subclass)에서 기증자를 발견하지 못 한 경우가 발생하였다. 세 번째로 일부 항목의 경우 한 사람이 여러 개의 답을 제시하는 것이 가능하였다. 예를 들어 보험의 경우 개인당 여러 개의 보험을 가지고 있을 수 있고 각 보험에 대한 응답이 요구되었다. 이 경우 각각의 보험 액수에 대한 대체에서 동일인에 의한 여러 가지 보험의 액수는 서로 연관되어 나타날 수 있으므로 연관성을 고려하여 예측이 실시되어야 한다. 네 번째로 같은 영역 내의 연관된 문항들 사이에 일치성(consistency)을 만족시키도록 대체가 실시될 필요가 있었다. 각 특징별로 대체를 시행하기 위하여 사용한 방법은 아래와 같다.

고령화연구패널조사의 항목 중 소득과 자산 항목의 경우 결측값으로 인한 정보의 손실을 줄이기 위하여 대괄호 질문들(bracket questions)을 사용하였다. 예를 들어 임금소득 액수에 관한 응답이 거절되거나 부정확하게 응답되는 경우 1년 총 임금소득을 정확한 액수 대신 서로 겹쳐지지 않는(unfolding) 정해진 구간들(brackets) 중에서 선택하도록 함으로써 무응답자에 대한 정확한 임금소득은 측정하지 못하였지만 대신 구간으로 응답된 부차적 임금 소득 정보를 얻을 수 있도록 하였다. 하지만 이렇게 얻어진 정보는 정확한 임금 소득이 아니고 구간으로 표현되므로 연구자가 임금소득에 대한 분석을 시행할 때 정확한 임금소득과 단순히 통합하여 사용하기 힘들다. 이에 고령화연구패널 자료의 대체 시 대괄호 질문들을 포함한 항목들의 정보를 이용하여 정확한 임금 소득에 대한 대체를 실시하였다. 즉, 소득 및 자산 항목에 대한 대괄호 질문 응답자는 전체 구간이 아니라 일정 구간에 소득이 존재한다는 정보를 제공하였으므로 대괄호 질문에서 응답된 구간내의 관찰된 자료들만을 기증 대상으로 선택하고 그 관찰값 중 동일한 하위그룹(subclass)

에 속하는 관찰값을 기증자로 선택함으로써 대체된 값들이 응답된 구간 안에 존재하도록 하여 대체된 자료의 일치성(consistency)을 만족시키는 동시에 동일 하위 그룹에 속한 비슷한 예측값을 가진 관찰값으로 대체하여 대체의 정확도를 추구하였다. 물론, 대괄호 질문에도 응답하지 않은 응답자가 문항에 따라 상당수 존재하였으며 이 경우 전체 관찰값 중 회귀 모형을 통하여 동일한 하위그룹에 속하는 관찰값들 중 기증자가 선택되었다.

핫덱 방법의 문제점은 결측값을 가진 자료가 많거나 특이한 값 (예를 들어, 소득이 아주 많은 사람들의 대부분이 소득 항목에 대하여 무응답인 경우)이라면 기증자를 발견하기 힘든 경우가 발생할 수 있다는 것이다. 이와 같은 핫덱 방법의 문제점을 극복하기 위하여 미국 Health and Retirement Study(HRS) 자료의 경우 다중대체를 시행할 때 기증자가 없는 극단 관찰값 구간에 대하여 회귀모형(regression model)에 근거한 예측값(predicted value)에 근거한 대체를 가능하도록 하였는데 (Cao, 2001) 이런 문제점은 대괄호 질문을 포함한 고령화연구패널조사의 일부 항목의 극단 대괄호 구간에서도 발생하였다. 따라서 일부 대괄호 질문을 포함한 문항에서 결측값에 대한 기증자가 존재하지 않는 경우 수정된 예측 평균에 근거한 짝짓기(modified predictive mean matching) 방법을 HRS에서 사용한 혼합된 대체법(mixed method)에서처럼 회귀모형(regression model)에 근거한 예측값(predicted mean)에 근거한 대체와 혼합해 실시하도록 확장하였다.

일부 문항의 경우 참여자 각각에 대하여 복수의 응답이 가능하였다. 예를 들어 개인당 복수의 보험을 가지고 있을 수 있거나 가족 영역에서 여러 자녀로부터 지원을 받거나 지원을 하는 경우 등이 이에 해당된다. 고령화연구패널조사의 경우 개인이 여러 개의 보험을 가지고 있는 경우 각각의 보험에 대하여 보험 액수를 조사하였는데 일부 보험 액수에 결측이 발생한다면 동일인이 가지고 있는 다른 응

답한 보험 액수는 결측된 액수에 대하여 정보를 가지고 있을 것이므로 대체에 포함하는 것이 적절하다. 하지만 회귀모형을 적합할 때 각 관찰값은 서로 독립이라 가정하지만 이 경우 복수의 응답들은 서로 독립이 아니다. 따라서 이런 문항의 경우 각 개인당 복수의 관찰값의 연관성 및 각 관찰값의 특성들을 포함하도록 회귀모형의 추정 시 Generalized Estimating Equations(GEE) 방법으로 모수를 추정한 후 예측값을 계산하도록 변형하였다.

설문 문항 중 동일 세션의 일부 변수들은 연관되어 있다. 예를 들어 1년 총 임금 소득을 측정할 때 그 해 일을 한 달들도 측정되었는데 총 임금 소득은 일을 한 달의 숫자에 따라 달라질 것으로 생각된다. 이와 같이 연관이 있는 것으로 생각되는 변수들의 경우 n-partition imputation을 사용하여 대체를 실시하였다(Marker et. al., 2002). 즉, 연관된 변수들 중 주요 관심 변수에 대한 수정된 예측평균에 근거한 핫덱 대체를 실시하고 대체를 위해 선택된 기증자의 다른 변수값들을 가지고 결측값이 발생하는 다른 연관된 변수들을 그룹으로 대체하였다. 예를 들어 1년 총 임금 소득을 대체하고 만약 일을 한 달 수가 결측이라면 대체에 사용된 기증자의 일을 한 달 수를 가지고 대체하는 방식을 채택하였다.

소득 및 자산 일부 항목의 경우 응답은 가족 전체가 동일해야 하지만 설문은 각 개인별로 진행되었다. 예를 들면 소득 영역의 가구 총소득은 가구원 별로 동일해야 하고 자산 중 집 소유 여부나 집의 가격의 경우 가구마다 동일한 값을 가져야 한다. 하지만 이 문항들의 응답이 개인별로 이루어짐에 따라 가구원들 간 응답의 차이가 존재하거나 일부 가구원의 경우 무응답으로 인하여 결측이 발생하였다. 이런 문항의 경우 응답자의 값이 결측인 경우 커버스크린(Coverscreeen) 영역에서 선택된 가구별 대표응답자의 응답값을 가지고 대체하고 대표응답자가 존재하지 않는 경우 가구 내 소득이 가장 많은 응답자의 값을 가지고 대체하였으며 이 값

들이 모두 결측인 경우 개별 대체를 실시하였다. 가족 영역의 경우 가구별로 각 세대 대표응답자만 응답하였으므로 이 값을 가지고 나머지 가구원의 값을 대체하였다.

다중대체는 남자와 여자에 대하여 독립적으로 시행되었다. 위 방법에 의하여 핫덱 대체를 시행하기 위하여 4가지 모형이 고려되었다. 이 모형들은 차례로 모형 1(Hotdeck), 모형 2(Bounded Hotdeck), 모형 3(Hotdeck - GEE), 그리고 모형 4(Hotdeck & Regression)이라 부른다.

(1) 모형 1 (Hotdeck)

이 모형은 원래 수정된 예측 평균에 근거한 핫덱 모형이다. 우선, 결측이 발생한 각 변수에 대하여 관찰된 자료만을 대상으로 회귀모형(regression model)을 적합한 후 결측이 포함된 모든 자료에 대한 예측값을 구한 후 예측값에 근거하여 층화(stratification)를 하여 가능한 한 10명 이상의 구성원을 포함하는 하위 그룹(subclass)을 구성한다. 구성된 각 하위그룹(subclass) 내에서 각 결측값은 같은 하위그룹의 관찰자 중에서 기증자를 선택하여 기증자의 값으로 대체를 실시하였다.

(2) 모형 2 (Bounded Hotdeck)

이 모형은 위의 모형을 대괄호 질문(bracket questions)들에서 얻어진 정보를 포함하도록 확장한 핫덱 모형이다. 고령화연구패널조사에서는 서로 겹쳐지지 않는 5개의 대괄호 질문들이 시행되었으므로 각 응답값은 6개의 구간 중 하나의 구간으로 나타내 질 수 있다. 정확한 값 대신 대괄호 질문에 응답한 경우 값이 어느 구간 안에 존재하는 지에 관한 정보가 주어지므로 이 정보를 사용하여 대체가 실시되어야 한다. 예를 들어 한 참여자의 임금소득에 대한 대괄호

질문의 응답이 "2400 MW 이상 6000 MW 미만"이라면 이 사람의 대체된 값은 이 구간 안에 속해야 제공된 정보와 일치하는 대체가 이루어지는 것이다. 따라서 우선 정확한 값이 응답된 참여자들의 관찰값들을 대괄호 질문에서 선택한 6개의 구간으로 분리하였다. 각 구간 안에서 이 관찰값들은 대괄호 질문에서 동일 구간에 속하는 것으로 응답된 결측값에 대한 기증자의 후보 집단(pool)으로 사용된다. 각 구간별로 다수의 응답자가 있는 경우 회귀 모형에 의한 예측값을 사용하여 하위그룹(subclass)을 나누고 동일한 하위 그룹 내에서 결측값에 대한 기증자를 선택하여 기증자의 값을 가지고 대체를 실시하였다.

(3) 모형 3 (Hotdeck - GEE)

문항에서 참여자 각각에 대하여 복수의 응답이 가능한 경우, 회귀모형을 적합할 때 각 관찰값은 서로 독립이라 가정하지만 한 참여자로부터의 복수의 응답들은 서로 독립이 아니다. 따라서 회귀모형에서 각 응답과 관련된 특성 변수를 포함하는 동시에 복수의 관찰값들 간의 연관성을 포함하도록 하는 분산-공분산 행렬을 고려하고 Generalized Estimating Equations(GEE) 방법으로 모수를 추정한 후 예측값을 계산하도록 변형하였다. 이 방법은 모형 1에서 회귀 모형을 적합할 때 회귀 모수가 관찰값의 연관성을 포함하고 GEE 방법으로 추정되도록 하기 위하여 SAS Macro에서 REG procedure 대신 GENMOD procedure를 사용하도록 수정함으로써 적용되었다. 또한, 대괄호 질문을 포함한 항목에 대하여 모형 2에서와 같이 구간 정보를 이용하여 대체를 시행하도록 하였다.

(4) 모형 4 (Regression)

괄호 질문을 포함한 일부 문항의 극한 (특히 액수가 많은 방향으로) 구간에서 결측값이 발생하였으나 응답자가 존재하지 않는 경우가 발생하였다. 예를 들면 남성의 경우 임금 소득의 가장 상위 구간으로 대괄호 질문에 대하여 구분된

결측값은 존재하였으나 그 구간에 해당하는 관찰값이 존재하지 않았다. 따라서 이 결측값에 대한 대체를 실시하고자 할 때 동일 구간 내에 기증자가 존재하지 않으므로 결측값에 대한 핫덱 대체가 불가능하였다. 이 경우 핫덱 대체 대신에 회귀 모형(regression model)에서 계산되어진 예측값을 가지고 결측값을 대체하였다. 수정된 예측 평균에 근거한 핫덱 대체의 경우 모든 관찰값 및 결측값에 대한 회귀 모형의 예측값이 하위그룹(subclass)을 나누기 위하여 계산되므로 이 계산된 예측값을 가지고 대체를 실시하면 추가적으로 모형을 적합하는 부담 없이 모형 1에 대한 약간의 수정을 통하여 쉽게 적용할 수 있다.

모형 1부터 4는 서로 독립적으로 적용되지 않고 혼합적으로 적용되는 경우가 많았다. 예를 들어 임금 소득의 대체 시에 일부 참여자는 정확한 임금 소득 대신 대괄호 질문에 의한 부분 정보를 제공하지만 일부 참여자는 대괄호 질문에도 응답을 거부하였다. 이 경우 대괄호 질문에 응답한 사람의 임금 소득은 응답 구간에 따라 모형 2를 이용하여 대체되었고 대괄호 질문에 대한 응답조차 거부한 경우 구간이 없이 모형 1을 이용하여 대체를 실시하였다. 또한 일부 문항에서 기증자를 찾지 못한 극한 관찰값은 모형 4를 이용하여 대체를 실시하고 나머지는 모형 1이나 2를 통하여 대체가 실시되었다.

9.3 고령화연구패널조사에서의 변수별 무응답 현황 및 무응답 대체 방법

송주원 외 (2007)는 각 영역에 속하는 주요 변수들에 대하여 전체 해당되는 관찰단위의 숫자, 결측의 숫자 및 결측 비율을 보고하고 있다. 또한 각 영역별 그리고 변수별로 위에서 언급한 모형 중 어느 대체모형이 선택되었는지를 표로 정리하였고 대체를 시행할 때 사용한 변수들에 관한 세부 정보도 나타나 있다.

제 10장. 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.

10.1 미국 Health and Retirement Survey (HRS) 개요

미국 Health and Retirement Survey (HRS)는 미국 전역에 걸쳐 노인들의 경제, 건강, 가족 상태를 조사하는 패널자료이다. 1992년 이래 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. 이 자료는 무응답을 포함한 자료에 대하여 다중대체를 실시하고 대체된 자료를 홈페이지를 통해 사용자에게 제공해 왔다. 또한, 대체를 실시하기 위한 프로그램 (SAS Macro IMPUTE)을 개발하여 연구자들에게 제공하고 있다.

10.2 미국 HRS에서 발생하는 무응답을 처리하는 기법

이 자료의 특징은 사례 3과 비슷하게 일부 항목에서 발생하는 결측값으로 인한 정보의 손실을 줄이기 위하여 대괄호 질문들(bracket questions)을 사용하였다는 점이다. 하지만 모든 항목에서 대괄호 질문들이 사용된 것은 아니므로 항목의 특성에 따라 적절한 대체 방법을 사용하였다. 또한 연금 항목의 경우 부부로부터 정보를 얻었는데 부부간 정보가 연관되어 있으므로 부부의 정보를 포함하여 대체를 실시할 수 있도록 프로그램을 구현하였다. 전체적으로 다음의 네 가지 대체 방법이 적용되었다.

(1) 중위수 대체

무응답인 관찰단위의 변수값은 그 변수에 대한 응답값의 중위수를 이용하여 대체를 실시한다. 대괄호 질문이 포함된 변수의 경우 값에 대한 부분적인 정보인 구간 정보가 주어진다면 그 주어진 구간내의 응답값들 중 중위수를 구하여 대체를 실시한다.

(2) 핫덱 대체

무응답인 관찰단위의 변수값은 4.1절에서 언급한 완전임의 핫덱 방법을 사용하여 대체를 실시한다. 우선 응답자와 무응답자 모두에게 임의로 순서화된 값들을 할당한다. 예를 들어 관찰단위가 10개가 있다면 이 10개의 값들은 1부터 10까지의 숫자 중 한 개의 숫자를 할당받는다. 그 후에 이 할당받은 숫자들을 가지고 자료를 정렬(sort)한 후 무응답값에 해당하는 할당된 숫자와 가장 가까운 숫자를 할당받은 응답값을 가지고 무응답을 대체한다. 대괄호 질문이 포함된 변수의 경우 값에 대한 부분적인 정보인 구간 정보가 주어진다면 그 주어진 구간내의 응답값들 중 임의로 순서화된 값들을 할당한 후 무응답자와 가장 가까운 값을 할당받은 응답

자를 기증자로 정하여 대체를 실시한다.

(3) 점수(score) 대체

무응답이 발생한 변수를 반응변수로 이 변수와 연관된 공변량들을 설명변수로 사용하여 회귀분석을 실시하고 각 관찰단위에 대하여 예측값을 계산한다. 예측값에 근거하여 기증자 풀(pool)을 만들고 기증자 풀 내에서 응답값과 가장 비슷한 예측값을 주는 기증자를 선택하여 그 기증자의 값을 가지고 핫덱 대체를 실시한다. 마찬가지로 대괄호 질문이 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 실시한다.

(4) 혼합 모형 대체

이 방법은 핫덱 대체와 회귀분석에 기초한 대체를 혼합하여 사용하는 방법이다. 혼합 모형은 기본적으로 결측된 소득에 대하여 핫덱 대체를 실시하는 데 대괄호 질문들(bracket questions)에 따라 회귀모형 또는 일반 핫덱을 섞어 사용하는 혼합 모형에 기초한다. 즉, 소득에 관한 대괄호 (bracket) 중 하한 구간 (bottom-open bracket; 예를 들면 소득 600만원 미만) 또는 막힌 구간 (closed bracket; 예를 들면 소득 600-1200만원 사이)에 대하여 같은 구간내의 관찰된 소득으로 핫덱 대체를 실시하고 상한구간 (top-open bracket; 예를 들면 소득 12000만원 이상)에 대하여 회귀분석을 실시하여 예측값을 구하고 이에 근거한 대체를 실시하는 방법을 의미한다. 이 방법을 사용한 다중대체를 시행하는 SAS Macro 프로그램은 홈페이지(http://hrsonline.isr.umich.edu/)에서 무료로 제공되고 있다.

이 자료는 응답 여부를 결정짓는 질문(ownership question)은 포함하는데 위의 네 가지 방법이 문항에 따라 <표 10.1>과 같이 혼합되어 적용되었다.

<표 10.1> 여러 가지 대체 방법의 실제 적용

	Ownership Question	맨 아래 대괄호 중간 대괄호	맨 위 대괄호 대괄호 정보 무응답
중위수	핫덱	중위수	중위수
핫덱	핫덱	핫덱	핫덱
점수	회귀	회귀	회귀
혼합	회귀	핫덱	회귀

Cao(2001a)은 이 대체를 위해 사용된 IMPUTE 프로그램에 대한 사용법을 상세히 설명하고 있다.

10.3 미국 HRS 자료의 무응답이 대체된 변수

Cao(2001b)는 위의 방법으로 대체된 변수들 및 각 변수별 대체 방법에 대한 자세한 설명을 포함하고 있다.

부록 A. 무응답 대체를 위한 교육 프로그램

교수요목(syllabus)

A1. 교육 프로그램의 목적

통계전문가를 대상으로 무응답을 포함한 자료의 문제점을 이해하고 어려 가지 무응답 자료의 분석 방법을 설명하며 실제 무응답 자료에 적절한 분석 방법을 선택할 수 있도록 하는 것을 목적으로 한다. 이를 위하여 무응답 관련 이론 및 분석 방법을 습득하고 무응답 처리 프로그램들을 이용하여 실제자료에 적용할 수 있도록 돕는다.

A2. 교육 프로그램의 대상

이 교육 프로그램은 충분한 통계적 지식을 지닌 연구자를 대상으로 한다. 즉, 통계학 또는 관련분야 박사학위 소지자나 통계학 또는 관련 분야의 석사학위 소지자로서 조사 연구 분야에 대한 일정 경험이 있는 연구자, 또는 이에 준하는 통계전문가를 대상으로 한다. 본 교육 프로그램이 무응답 자료를 다루기 위한 통계 프로그램들에 대한 소개 및 실습을 포함하므로 대상자는 통계프로그램 SAS를 사용할 수 있기를 기대한다.

A3. 교육 프로그램의 내용

본 교육 프로그램은 무응답을 포함한 자료의 분석에 대한 다양한 수요를 충족시키기 위하여 다음의 두 가지 과정으로 나누어진다.

A3.1. 중급 과정 (5일간)

이 과정은 무응답을 포함한 자료의 분석에 대한 기초 및 실제 무응답 자료의 분석에서 흔히 사용하는 기법 및 대체 방법들에 대하여 이해하고 연관 프로그램의 사용법을 익혀 실제 자료에 적용 가능하게 하는 것을 목적으로 한다. 이 과정의 제안된 예시 일정은 다음 <표 A.1>과 같다.

<표 A.1> 중급과정 예시일정

< 1일차 >

시간	내용
9:10 - 10:00	교육개관
10:10 - 11:00	무응답을 포함한 자료에 대한 소개
11:10 - 12:00	무응답 자료의 특성 이해 - 무응답 발생 패턴, 발생 메커니즘 정의
12:00 - 13:10	점심시간
13:10 - 14:00	단순한 결측 자료 분석 방법 단순한 대체방법 소개 I
14:10 - 15:00	단순한 대체방법 소개 II
15:10 - 16:00	실습 I - 단순한 대체방법 I
16:10 - 17:00	실습 II - 단순한 대체방법 II

< 2일차 >

시간	내용
9:10 - 10:00	가중값 접근 방법 I
10:10 - 11:00	가중값 접근 방법 II
11:10 - 12:00	가중값 접근 방법 III
12:00 - 13:10	점심시간
13:10 - 14:00	단일대체 방법 I
14:10 - 15:00	실습 I - 가중값 접근 방법 I, II
15:10 - 16:00	실습 II - 가중값 접근 방법 III
16:10 - 17:00	실습 III - 단일대체 방법 I

< 3일차 >

시간	내용
9:10 - 10:00	단일대체 방법 II
10:10 - 11:00	다중대체 방법 I
11:10 - 12:00	다중대체 방법 II
12:00 - 13:10	점심시간
13:10 - 14:00	실습 I - 단일대체 방법 II
14:10 - 15:00	실습 II - 다중대체 방법 I
15:10 - 16:00	실습 III - 다중대체 방법 II
16:10 - 17:00	소양체력단련

< 4일차 >

시간	내용
9:10 - 10:00	핫덱대체 방법 I
10:10 - 11:00	핫덱대체 방법 II
11:10 - 12:00	핫덱대체 방법 III
12:00 - 13:10	점심시간
13:10 - 14:00	실습 I - 핫덱방법 I
14:10 - 15:00	실습 II - 핫덱방법 II
15:10 - 16:00	실습 III - 핫덱방법 III
16:10 - 17:00	실습 IV - 핫덱방법 III

< 5일차 >

시간	내용
9:10 - 10:00	실습 I - 실제 자료 사용
10:10 - 11:00	실습 II - 실제 자료 사용
11:10 - 12:00	실습 III - 실제 자료 사용
12:00 - 13:10	점심시간
13:10 - 14:00	요약
14:10 - 15:00	토론 및 정리
15:10 - 16:00	설문조사 및 수료

A3.2. 고급 과정 (3시간씩 15주간)

이 과정은 무응답을 포함한 자료의 분석과 연관된 이론의 설명 및 폭 넓은 무응답 자료 분석 기법들에 대하여 자세히 설명한다. 또한 여러 가지 무응답 자료 분석 기법을 실제 자료에 적용하는 방법에 대하여 배우고 실제 자료가 주어졌을 때 적절한 분석 기법을 선택할 수 있도록 이론과 실습이 조화를 이루도록 한다. 이 과정은 중급 과정보다 이론을 더 심화로 다루며 실제로 무응답 자료 분석팀을 리드할 수 있는 무응답 전문가를 양성하는 것을 목적으로 한다.

이 과정의 제 1주는 3시간 모두 강의로 구성되고 제 2주부터는 강의 2시간, 실습 1시간으로 구성된다. 이 과정의 제안된 예시 일정은 다음 <표 A.2>와 같다.

<표 A.2> 고급과정 예시일정

주차	내용
1주	무응답을 포함한 자료에 대한 소개 무응답 자료의 특성 이해 - 무응답 발생원인, 패턴, 발생 메커니즘 정의
2주	단순한 결측 자료 분석 방법 및 대체방법 소개 - 완전히 응답한 자료 분석방법(complete-case analysis) - 이용가능한 자료 분석방법(available-case analysis) - 단순한 대체방법들 I
3주	단순한 결측 자료 분석 방법 및 대체방법 소개 - 단순한 대체방법들 II
4주	가중값 접근 방법 - 가중값 접근 방법 I
5주	가중값 접근 방법 - 가중값 접근 방법 II - Resampling 방법
6주	우도에 기초한 방법들 - EM algorithm
7주	베이지안 분석 기법 - MCMC 기법들
8주	단일대체와 다중대체 - 단일대체 방법 - 다중대체 방법
9주	대체 방법 I - 핫덱대체
10주	대체 방법 II - 핫덱대체
11주	대체 방법 I - 모수적 대체
12주	대체 방법 I - 모수적 대체
13주	패널자료 분석 - 가중값 접근 방법
14주	패널자료 분석 - 대체 방법
15주	토론 및 정리

<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 Association*, 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. (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.

통계전문 교육프로그램 개발
- 무응답 자료처리 관련 -

발행일 2009년 8월 30일

발행처 고려대학교
서울시 성북구 안암동 5가
전화: 02) 3290 - 2241
팩스: 02) 3290 - 2241

통계전문 교육프로그램 개발
- 무응답 자료처리 관련 -

통 계 교 육 원