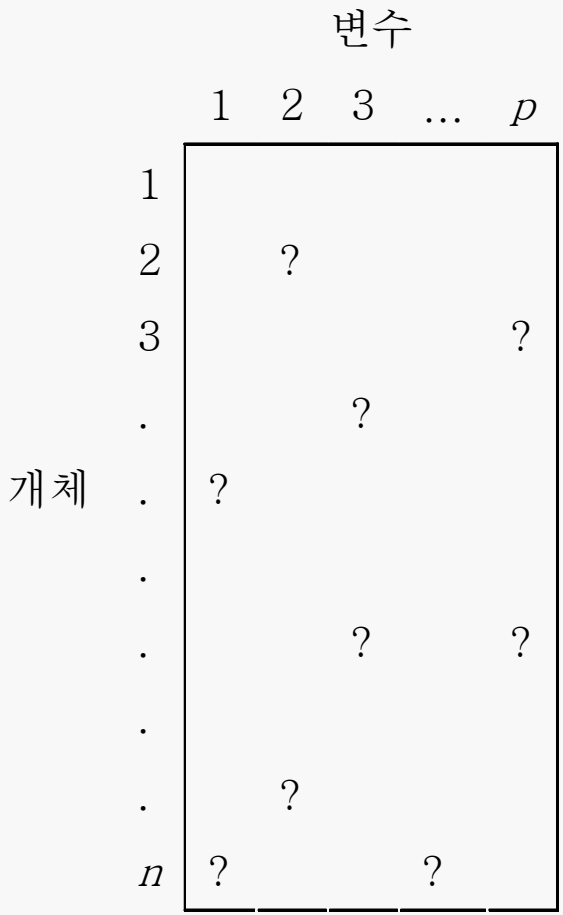
결측자료 분석을 위한 대체방법

송주원

고려대학교

목차

결측자료의 형태



결측자료의 예제

- 설문조사(questionnaire survey): 민감한 질문에 대한 응답 거부
  - 소득이나 지출 세부 사항에 대하여 무응답 발생
  - 주택마련시기에 대한 응답은 주택 소유하지 않은 경우 결측으로 남음

- 사회조사: 건너뛴 문항(skipped question)의 결측
  - 청소년의 흡연 정도를 조사하는 경우 비흡연자의 흡연량에 관한 문항에 대한 대답은 결측으로 남음

- 여론조사(public opinion survey): 응답 거부 또는 응답 불가능으로 인한 무응답
  - 대통령 선거 여론조사 시 응답 거부
  - 선호하는 후보자가 없어 응답 불가능

결측자료의 예제

- 화학실험
  - 시약을 잘못 투여하여 반응값이 나타나지 않는 경우 이 표본에 대한 반응값은 결측으로 남음

- 기업 자료: 제품의 소비자에 대한 여러 가지 기본 정보 중 일부 정보에서 결측이 발생
  - 제품의 구매 고객의 연령, 직업 또는 소득과 같은 개인 정보 일부 결측 발생

- 임상실험 자료
  - 중도에 참여를 포기하면 추후 경과는 결측
  - 약에 대한 심각한 부작용으로 인하여 연구에서 중도탈락

결측자료의 분석

- 모든 결측자료가 결측은 아니다.
  - 예 1) 사회조사: 비흡연자의 흡연량은 0
  - 예 2) 여론조사: 선호하는 후보자가 없는 경우는 다른 범주를 의미

- 결측 자료의 분석 대상
  - 문항에 대한 정확한 값을 얻을 수 있지만 여러 가지 원인으로 인하여 응답값이 정확히 측정되지 않은 경우만을 고려

패널조사의 결측

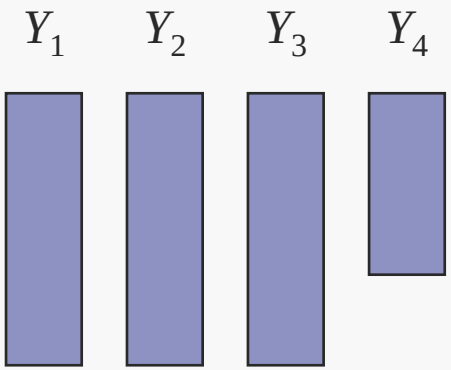
- 패널조사의 결측 분류
  - 단위무응답(Unit non-response): 특정조사 시점에서 한 개체(unit)가 응답거부나 여러 가지 사유로 조사에 응하지 않아서 생긴 결측
  - 항목무응답(Item non-response): 특정조사 시점에서 조사에는 응하였으나 몇 몇 항목의 값이 결측된 경우

- 패널조사 결측에 대한 가장 흔한 처리 방법
  - 단위무응답: 대입(substitution) 및 가중값 분석
  - 항목무응답: 대체(imputation)

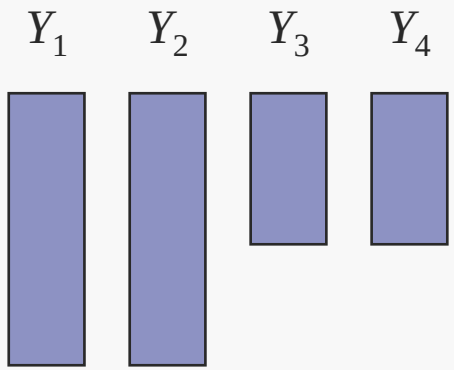
결측자료 패턴

- 예제

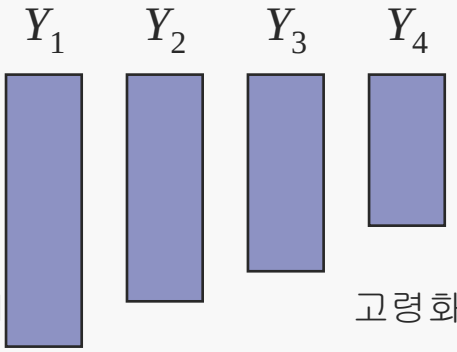
(a) 일변량 무응답



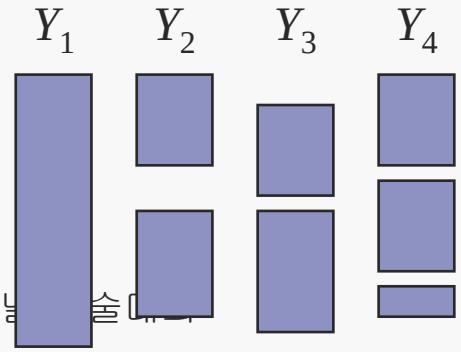
(b) 두 가지 패턴



(c) 단조패턴



(d) 일반적인 패턴



결측자료 메커니즘

- 결측자료 모형
  - $Y = (y_{ij})$: $(n \times p)$ 완전한 자료
    Y_{obs}: 자료 Y의 응답값
    Y_{mis}: 자료 Y의 결측값

  - 응답지시행렬 $R = (r_{ij})$

$$\begin{cases} r_{ij} = 1 & \text{만약 } y_{ij} \text{가 응답이면} \\ r_{ij} = 0 & \text{만약 } y_{ij} \text{가 결측이면} \end{cases}$$

  - 완전히 응답된 자료 모형: 자료 Y와 응답지시행렬 정보 모두 필요

  - 결측값을 포함한 자료의 모형: 응답자료 Y_{obs}와 응답지시행렬 정보 모두 필요

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

(1) 자료행렬 Y

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

(2) 자료행렬 Y에 대응되는 응답 지시행렬 R

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

결측자료 메커니즘

- 완전임의결측 (Missing Completely at random; MCAR)
  - 결측이 발생할 확률은 자료값에 의존하지 않는다.
  - 결측값이 발생한 자료는 완전한 자료의 임의의 부표본이다.
  - 예: 수입에서 결측 발생시 응답자와 무응답자 간에 어떤 본질적인 차이가 없다고 한다면 응답자의 수입에 관한 분포와 무응답자의 수입에 관한 분포가 같을 것이다.
  - 매우 강한 가정

결측자료 메커니즘

- 임의결측 (Missing at Random; MAR)
  - 관측값이 응답일 확률은 응답값에는 의존하지만 결측값에는 의존하지 않는다.
  - 예: 수입은 일반적으로 높은 소득자가 무응답률이 높을 것이다. 하지만, 모든 표본의 세금에 관한 정보가 있고 이 정보가 주어졌다면 소득에 관한 무응답은 임의적(random)이라고 가정하자. 즉, 같은 세금을 내는 사람들이 소득에 관해 무응답을 할 확률은 서로 같다. 세금에 관한 정보가 주어진 경우, 소득에 관한 무응답은 소득과는 무관하므로 이런 경우를 임의 결측이라고 한다.
  - 완전 임의 결측보다는 덜 강한 가정

결측자료 메커니즘

- 비임의결측 (Not Missing at Random; NMAR)
  - 관측값이 결측일 확률은 응답값 뿐 아니라 결측값에도 의존한다.
  - 결측 메커니즘이 MCAR도 아니고 MAR도 아닌 경우를 비임의 결측이라고 한다.
  - 비임의결측의 발생 확률은 결측값 그 자체와 관련이 있다.
  - 예: 소득에 관한 무응답이 소득 자체와 관련이 있다. 즉, 세금에 관한 정보가 주어졌더라도 소득이 높은 사람이 더 높은 무응답률을 보이는 경우 비임의결측이 된다.
  - 가장 약한 가정
  - 유효한 통계적 추론을 위해서는 자료와 응답지시행렬 정보가 모두 필요하다.

결측자료 메커니즘

- 무시할 수 있는 결측자료 메커니즘
  - 결측자료 메커니즘이 완전임의결측(MCAR)이거나 임의결측(MAR)이며 자료의 모수가 결측자료 메커니즘의 모수와 별개인 경우 자료의 모수 추정은 결측자료 메커니즘을 무시한 채 실시될 수 있다.

- 대부분의 결측자료 분석 기법은 결측자료 메커니즘이 무시할 수 있는 결측자료 메커니즘을 따른다고 가정

완전히 응답한 개체를 이용한 분석 (Complete Case Analysis)

- 모든 변수들이 관측된 개체들만 이용하여 분석
- 단 하나의 변수에서 결측값이 있어도 그 개체는 분석에서 제외
- 대부분의 통계프로그램에서 이 방법을 사용

- 장점
  - 간편성
  - 일변량 통계량들의 비교가능
  - MCAR 가정하에서 모수 추정치에 편향(bias)이 거의 발생하지 않음.

- 단점
  - 많은 표본수의 감소 → 정보의 손실 → 검정력의 약화
  - MCAR이 아닌 경우 편의 발생 가능성

이용 가능한 개체분석 (Available Case Analysis)

- 각 각의 분석 단계에서 사용 가능한 자료를 이용

- 장점
  - 일반적으로 표본수는 완전히 응답한 개체를 이용한 분석 (complete case analysis)보다 많음
  - MCAR 가정하에서 모수 추정치에 편향이 거의 발생하지 않음

- 단점
  - 표본의 기저(base)가 분석마다 변한다.
  - 모수 추정시 수학적 문제가 발생하기도 한다.

결측값의 대체

- 대체(Imputation): 결측값을 그럴듯한 값을 가지고 대체하는 통계적 기법
  - 대체된 자료는 결측값이 없이 완전한 형태를 지님

- 명시적 모형(Explicit modeling)에 의한 대체
  - 각 변수들이 특정한 확률분포를 따른다고 가정하고 분포의 모수들을 추정하여 대체를 실시하는 방법
  - 가정이 명시적이다.
  - 예) 평균대체, 중앙값 대체
    확률대체
    비율대체
    회귀대체, 확률적 회귀대체
    분포를 가정한 대체

결측값의 대체

- 내재적 모형(Implicit modeling)에 의한 대체
  - 각 변수들이 특정한 확률분포를 따른다고 가정하는 대신 가능한 한 정확한 값을 가지고 대체하기 위한 알고리듬에 중점을 둔 방식
  - 대체를 위한 분포 가정이 명시적이지 않고 내재적이다.
  - 예) 핫덱대체(Hotdeck imputation)
    콜드덱대체(Colddeck imputation)
    대입(substitution)

- 명시적 모형과 내재적 모형이 혼합된 방식(Composite Methods)의 대체도 가능

대체군을 사용한 대체 (Imputation Within Adjustment Cells)

- 완전하게 응답된 범주형 변수를 가지고 대체군을 형성한 후 각 대체군내에서 결측값을 동일한 대체군의 응답된 값들로 대체하는 방법
  - 평균대체, 확률대체, 핫덱대체 등에서 흔히 사용

- 대체군을 이용한 대체의 성능은 대체군의 형성에 의존
  - 대체군 내에서 응답값과 결측값의 분포가 동일하도록 대체군 형성
  - 대체군을 형성한 변수들이 주어졌을 때 결측자료 메커니즘이 완전임의(MCAR)가 되도록 대체군 형성
  - 대체군을 형성하기 위하여 결측값이 발생한 변수와 연관되어 있는 변수들 포함

분포(distribution)를 가정한 대체 방법

- 다변량 정규분포를 가정한 대체 방법
  - 개념적으로 단순하고 적용이 쉬워 인기
  - 장점

    (1) 상용 통계프로그램을 사용하여 대체 시행 가능

    (2) 자료가 다변량 정규분포를 따르는 경우 검정력(power)이 높음.
  - 단점

    (1) 대부분의 자료는 여러 가지 타입의 변수들을 포함하므로 이 변수들에 대하여 다변량 정규분포를 가정하기 어려움

    (2) 변수들의 분포가 다변량 정규분포와 멀다면 추정량 편의 발생

- 비정규 분포 하에서의 대체방법
  - 범주형 자료에 대한 대체 모형, 연속형과 범주형 혼합자료의 대체 모형 개발
  - 자료에 따라 적용 어려움

여러 가지 분포를 가진 변수들을 포함한 자료에 대한 대체 방법

- 대부분의 자료는 여러 형태(type)의 변수들을 포함
- 순차회귀 대체(imputation using a sequence of regression models)
  - 각 변수에 알맞은 여러 가지 분포를 가정
  - 여러 가지 타입의 변수들에 단순회귀모형을 순차적으로 적용

변수의 타입	모형
연속형	선형회귀모형
이산형	로지스틱 회귀모형
범주형	다항 또는 일반화로짓 회귀모형
가산형	포아송 회귀모형
혼합형	두단계 모형

핫덱대체

- 핫덱대체: 자료내의 응답값을 사용하여 결측값을 대체하는 기법
  - 기증자를 선택하는 방법에 따라 분류

- 기증자와 수증자
  - 기증자(donor): 결측값의 대체에 사용된 응답 개체
  - 수증자(donee): 결측값이 발생하여 응답값의 기증을 받은 개체

- 주로 표본조사의 결측값을 대체하는 데 적용

- 신중하게 선택된 방법에 근거한 핫덱대체는 정확성 높은 대체 가능

핫덱대체

- 핫덱대체는 명시적 형태의 모형(explicit model)을 정의하지 않고 대체를 실시한다는 의미로 내재적 모형(implicit model)하에서의 대체 방법

- 문제점
  - 가정이 명시적이지 않으므로 추정값의 편향을 수리적으로 계산하기 어려움
  - 복잡한 함수들(complex functions)에 근거하여 결측값에 대한 대체가 실시되면 대체된 자료에 근거한 추정량의 성질을 평가하기 어려움
  - 핫덱대체의 성능(performance)은 대부분 비슷한 상황 하에서의 모의실험을 통해 평가

대체군을 사용한 핫덱대체 방법 (Hotdeck Within Adjustment Cells)

- 가능한 한 많은 변수를 고려하여 대체군을 형성할수록 대체군 내에서 완전임의 결측자료 메커니즘이 달성될 가능성이 높음
  - 대체군을 형성하기 위하여 변수가 추가될수록 대체군의 숫자가 기하급수적으로 늘어나는 문제
  - 대체군의 숫자가 늘어나면 일부 대체군에 속하는 응답값을 가진 관측값의 수가 적거나 없어 결측값에 대한 기증자를 찾지 못하는 문제점 발생 가능

- 대체군을 이용한 핫덱대체 기법에서 기증자를 찾기 어려운 경우
  - 대체군을 형성하는 변수 일부를 생략하고 기증자를 찾는 방식이 선호됨

혼합적 모형에 근거한 대체 방법

- 명시적 모형에 근거한 대체 방법들은 자료가 모형의 가정을 만족시키는 경우 우수한 성능을 보임

- 핫덱대체 방법은 자료에 대한 분포 가정을 포함하지 않으므로 여러 가지 다른 형태의 변수에 유연성 있게 적용 가능
  - 대체군을 잘 형성한다면 상당히 정확한 대체 실시 가능

- 명시적 모형에 근거한 대체 방법과 핫덱대체 방법들의 장점을 유지하면서 위에서 언급한 단점을 보완하기 위한 모형들이 제시
  - 예측평균에 근거한 짝짓기 방법(predictive mean matching method)은 명시적 모형으로 예측평균값(predictive mean)을 계산한 후 결측값의 예측값과 가까운 예측값을 갖는 응답값들로 핫덱대체를 실시

단일대체(Single Imputation)

- 단일대체(Single imputation): 한 개의 결측값을 한 개의 그럴듯한 값을 가지고 대체하는 방법

- 단일대체의 장점
  - 대체된 자료는 더 이상 결측값을 포함하고 있지 않으므로 연구자가 원하는 분석 시행 가능

- 단일대체의 문제점
  - 대체된 자료값 들 중 어느 값이 응답값이고 어느 값이 대체된 값인지 구별 불가
  - 추정량의 분산이 과소추정되므로 분산에 대한 보정(adjustments) 필요.

단일대체(Single Imputation)

- 응답값과 대체된 값 구별해야 하는 이유
  - 응답값은 참값에 대하여 정확하게 측정한 값이지만 대체된 값은 원래의 응답값과 동일하지 않을 가능성이 높음
  - 상당히 정확도가 높은 대체모형을 사용해도 대체된 값 모두가 원래의 값과 동일할 가능성 희박
  - 모수에 대한 추론은 응답된 자료만의 정보(information)에 근거하여 실시해야 함
  - 단일대체된 자료는 응답된 자료 뿐 아니라 대체된 자료로부터의 정보의 양도 추가되어 정보의 양을 과다추정

    -> 추정량의 분산 과소 추정

    -> 추론에서 기각하지 않아야 하는 모수 기각 가능

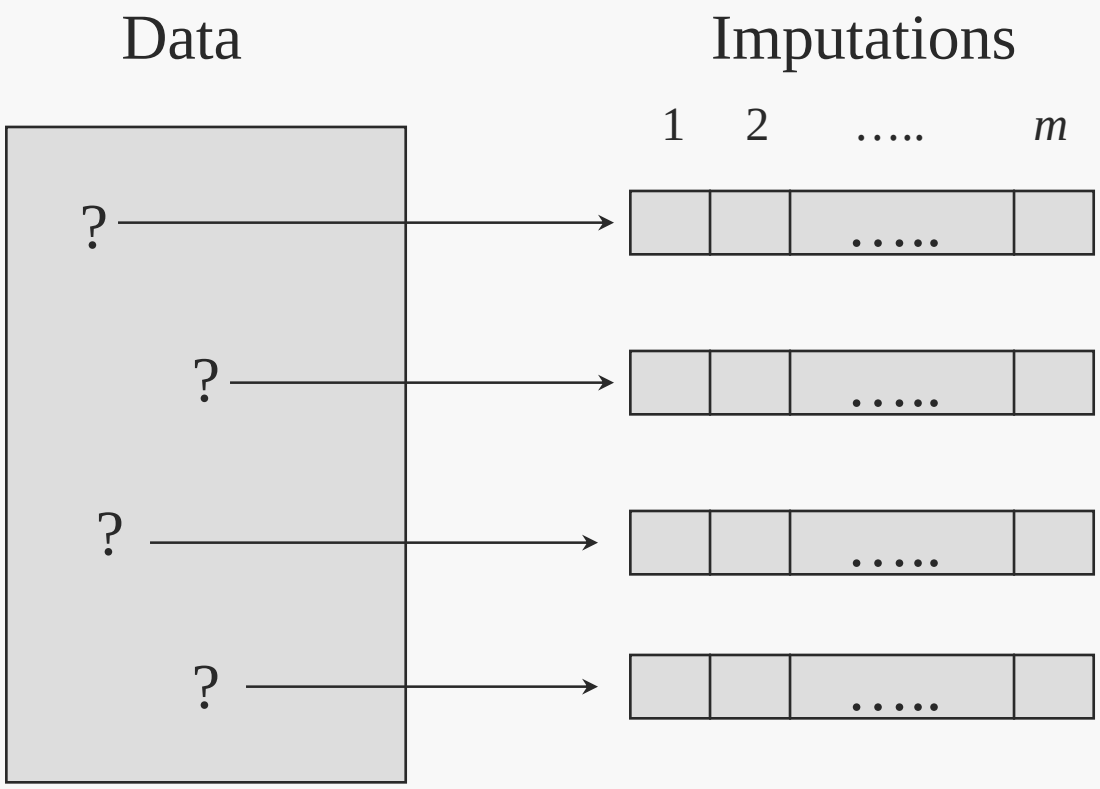
단일대체 실시 후 분산보정

- 단일대체를 실시한 후 모수의 분산 추정 시 발생하는 편향을 제거하는 방법들
  - 관심 모수에 대하여 편향이 없는 분산을 이론적으로 계산하여 보정
  - 반복 방법(replication methods) 사용: 여러 개의 반복 자료들(replicated datasets 또는 pseudo-replicates)을 형성한 후 반복 자료들로부터 추정된 모수의 분산들을 이용하여 보정

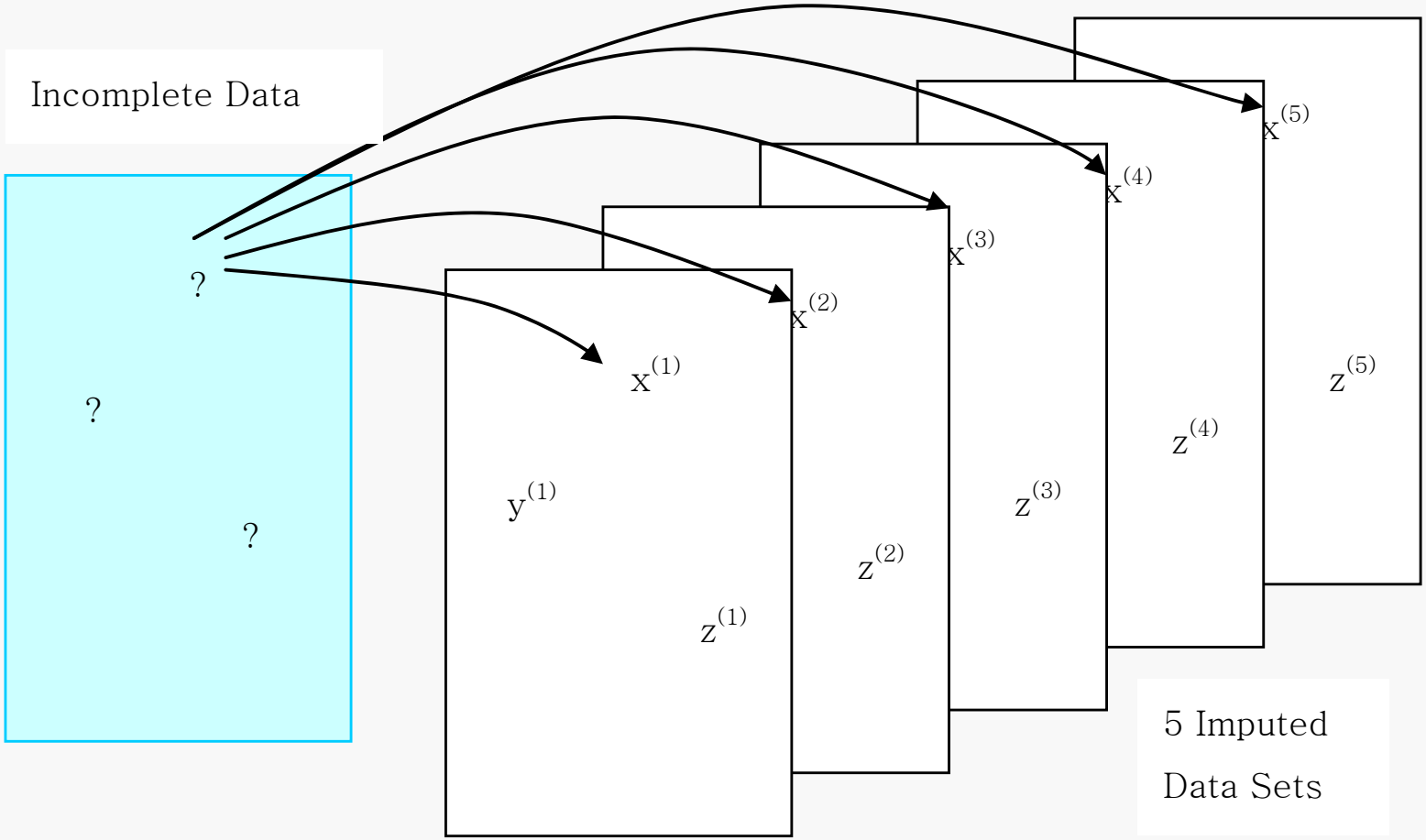
다중대체(Multiple Imputation)

- 다중대체(Multiple Imputation): 각각의 결측값을 2개 이상의 그럴듯한 값으로 대체
  - $m \geq 2$ 개의 다중대체가 실행되면 m개의 대체된 자료 생성
  - 각 결측값이 여러 개의 값들로 대체되고 대체된 여러 개의 값들간 차이 존재
  - 여러 개의 대체된 값들은 결측값에 대한 불확실성(uncertainty)을 반영
  - 대체된 값들간 차이에서 오는 분산이 추정량의 분산을 계산할 때 추가되어 분산이 과소추정 방지
  - 결측값의 참값을 정확하게 아는 것이 거의 불가능하므로 결측값을 여러 개의 그럴듯한 값으로 표현함으로써 결측값에 대한 우리의 불확실성을 모형에 포함
  - 대체된 m개 각각의 자료에 대하여 원하는 분석 시행 가능
  - 분석 후 추론을 위하여 결과를 결합하여 한 개의 추론 실시

다중대체(Multiple Imputation)



예제: 5개의 다중대체 자료



다중대체(Multiple Imputation)

- 다중대체의 단점: 단일대체보다 복잡
  - 대체를 여러 번 실시해야 함
  - 대체된 자료를 여러 번 분석해야 함
  - 각 분석된 자료를 통합하여 추론 필요

- 다중대체의 개수
  - 결측으로 인하여 손실된 모수에 대한 정보량(missing information)이 아주 크지 않다면 작은 숫자의 대체 가능
  - 결측으로 인하여 손실된 모수에 대한 정보량: 결측이 발생하지 않은 완전한 자료와 비교하여 결측이 발생함으로 인해서 발생한 모수의 정밀도(precision)의 감소분
  - 예: 손실된 모수에 대한 정보량이 50%라 하더라도 5번의 다중대체를 실시하면 무한개의 대체를 실시한 경우와 비교하여 모수에 대한 추정량의 표준오차는 약 5%만 늘어남
  - 손실된 모수의 정보량은 종종 결측값의 비율과 연관

다중대체 자료 분석 결과의 통합

- 우선 대체된 m개 각각의 자료에 대하여 원하는 동일한 분석 시행
  - m개의 관심있는 모수 및 모수의 분산 추정량 구함
  - $i = 1, \cdots, m$,에 대하여

    $\hat{\theta}_i$: i번째 대체된 자료에서 구한 모수의 추정량

    W_i : i번째 대체된 자료에서 구한 모수 θ의 분산의 추정량.

- 통합된 모수 θ 의 추정

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

  - m개의 대체된 자료들의 평균은 단일대체로 얻어진 추정량보다 효율적

다중대체 자료 분석 결과의 통합

- 통합된 모수 θ의 분산의 추정

  - 대체내 분산: $\bar{W}_m = \frac{1}{m}\sum_{i=1}^{m} W_i$

  - 대체간 분산: $B_m = \frac{1}{m-1}\sum_{i=1}^{m}\left(\hat{\theta}_i - \bar{\theta}_m\right)^2$

  - 통합된 분산

$$T_m = \bar{W}_m + \frac{m+1}{m} B_m$$

여기서, $\frac{m+1}{m}$ 은 유한한 숫자의 대체에 대한 보정

  - 결측으로 인하여 손실된 모수에 대한 정보량

$$\hat{\gamma}_m = \left(1+1/m\right) B_m / T_m$$

다중대체 자료 분석 결과의 통합

- *m*개의 대체된 자료의 분석 결과를 통합해 주는 통계 프로그램
  - SAS PROC MIANALYZE
  - R 함수
  - Schafer의 프로그램 NORM (http://www.stat.psu.edu/~jls/misoftwa.html)

고령화연구패널 제 1차 조사의 결측값 대체방법

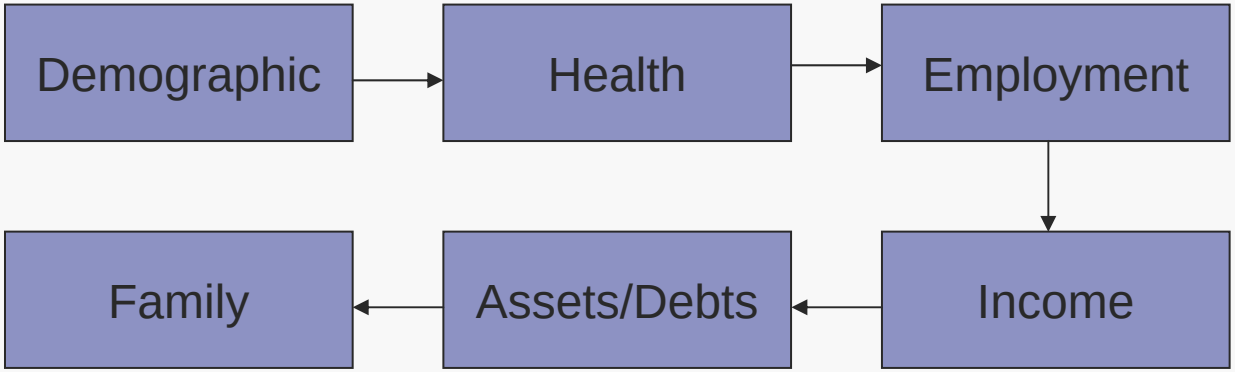
- 고령화연구패널(KLoSA)
  - 대상자: 6,171 표본 가구의 45세 이상 가구원 10,254명
  - 경시적 자료(Longitudinal study): Baseline in 2006
    1st follow-up in 2008

- 대부분의 조사자료와 마찬가지로 결측값 포함

- 예측 평균값에 근거한 핫덱대체 방법으로 다중대체(Multiple imputation based on hotdeck imputation using predictive mean matching) 실시

고령화연구패널 제 1차 조사의 결측값 대체방법

- 다중대체
  - 소득 및 자산 변수들의 대체에 중점을 둠
  - 영역별로 차례로 대체 실시



  - 다중대체의 수: 5 (결측값의 불확실성 표현)
  - 대체 방법: 예측 평균값에 근거한 핫덱대체 (Hotdeck based on a predictive mean matching)

사업체패널 조사 무응답에 대한 대체

- 사업체패널(WPS)
  - 대상: 농림어업 및 광업을 제외한 전 산업에서 상용근로자가 30인 이상 규모의 사업장 중 표본 추출된 1,905개의 사업장
  - 경시적 자료(Longitudinal study)

- 설문문항: 근로자/재무현황 설문
  인사담당자 설문
  노사관계 담당자 설문
  근로자 대표 설문

- 주요업종 및 근로자 수로 대체군을 형성한 후 대체군 내에서 대체 실시

사업체패널 조사 무응답에 대한 대체

- 변수의 타입을 고려하여 대체법을 선택한 후 대체 실시

- 사용된 대체 방법들
  - 핫덱대체
  - 비율대체
  - 중앙값 대체
  - 확률적 대체

토의

- 결측자료 분석을 실시할 때 결측자료 메커니즘을 고려해야 함
  - 결측값의 발생원인을 고려하여 적절한 가정 필요
- 결측값의 비율이 높은 경우 조심스러운 자료분석이 필요함
- 한 가지 대체 방법만 사용하는 대신 여러 가지 방법을 시행해 보고 결과 비교
  - 민감도 분석