
Statistical Analysis with Missing Data:

Missing Data in Experiments,
Complete-Case Analysis & Available-Case Analysis
including Weight Methods

ESC 2024 Summer Session 2주차
김효은, 신재민, 지민구, 한지희



Contents

- 1. Backgrounds
- 2. Missing Data in Experiments
 - 2.1 Introduction
 - 2.2 The Exact Least Squares Solution with Complete Data
 - 2.3 The Correct Least Squares Analysis with Missing Data
 - 2.4 Filling in Least Squares Estimates
 - 2.5 Barlett's ANCOVA Method
 - 2.6 Least Squares Estimates of Missing Values by ANCOVA Using Only Complete-Data Methods
 - 2.7 Correct Least Square Estimates of Standard Errors and One & More Degree of Freedom Sum of Squares
- 3. Complete-Case & Available-Case Analysis including Weighting Methods
 - 3.1 Complete-Case Analysis
 - 3.2 Weighted Complete-Case Analysis
 - 3.3 Available-Case Analysis

1

Background

1. Definition of Missing Data
2. Notation
3. Missing Data Patterns
4. Missing Data Types

1.1 Definition of Missing Data

Missing Data

- ① unobserved values
- ② that would be meaningful for analysis if observed

(example)

여론조사 거절한 사람 (① O)

- 애초에 정치에 관심이 없어 투표하지 않을 계획이라서 (② X)

⇒ Not Missing Data ⇒ better to exclude

- 누구한테 투표할지 생각하고 있지만 정치성향 밝히기 싫어서 (② O)

⇒ Missing Data ⇒ meaningful to impute

1.2 Notation

$Y : n \times k$ complete data set matrix

$y_{ij} : \text{each component } (i = 1, \dots, n, j = 1, \dots, k)$

$y_i : i^{\text{th}} \text{ row } (i^{\text{th}} \text{ unit/sample})$

$y_{(0)i} : \text{components of } y_i \text{ that are observed}$

$y_{(1)i} : \text{components of } y_i \text{ that are missing}$

$M : n \times k$ missingness indicator matrix

$$m_{ij} = \begin{cases} 0 & \text{if } y_{ij} \text{ is not missing } (y_{ij} \text{ is observed}) \\ 1 & \text{if } y_{ij} \text{ is missing } \quad (y_{ij} \text{ is not observed}) \end{cases}$$

$m_i : \text{row (missingness indicator vector for } i^{\text{th}} \text{ unit } y_i)$

1.2 Notation

(example)

세 명의 학생들이 세 개의 과목에 대해서 시험을 봤고,

학생들이 작성한 답안에 따르면 왼쪽 표와 같이 성적이 부여되었어야 한다.

그런데 선생님이 OMR을 수거하던 중 3장을 잃어버렸고,

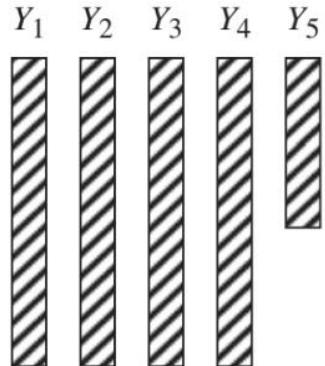
철수의 수학 점수, 영희의 국어 점수, 민수의 영어 점수는 알 수 없는 상태가 되어버렸다. (Missing)

	국어	수학	영어
철수	88	77	92
영희	85	92	88
민수	90	83	96

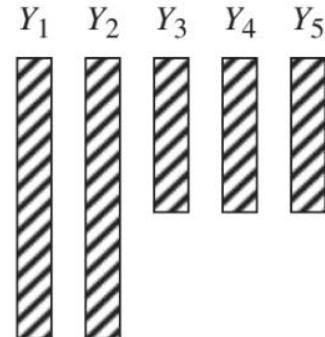
	국어	수학	영어
철수	88	?	92
영희	?	92	88
민수	90	83	?

$$Y = \begin{bmatrix} 88 & 77 & 92 \\ 85 & 92 & 88 \\ 90 & 83 & 96 \end{bmatrix} \quad M = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad R = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

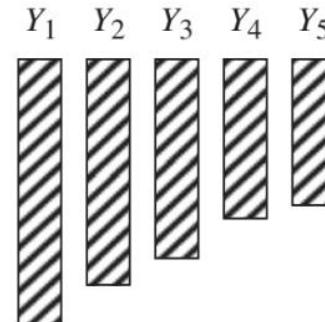
1.3 Missing Data Patterns



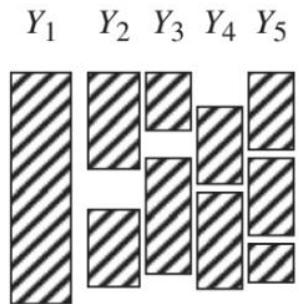
(a)



(b)



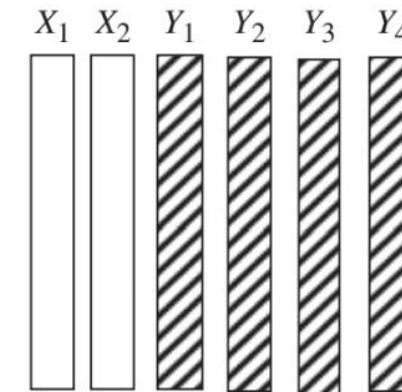
(c)



(d)



(e)



(f)

Figure 1.1 Examples of missingness patterns; rows correspond to units and columns to variables. (a) Univariate nonresponse, (b) multivariate with two patterns, (c) monotone, (d) general, (e) file matching, and (f) factor analysis, with two factors and four measured variables.

1.4 Missing Data Types

Y, M의 row들이 iid라 하자

conditional distribution of m_i given y_i (y값이 주어졌을 때 missing될 확률)를 기준으로

Missing Data의 종류를 구분할 수 있다

MCAR (Missing Completely At Random)

missingness가 데이터값에 의존하지 않음 (관측되었든, 결측이든)

$$f_{M|Y}(m_i|y_i, \phi) = f_{M|Y}(m_i|y_i^*, \phi)$$

MAR (Missing At Random)

missingness가 결측된 데이터값에 의존하지 않음 (관측된 데이터값에만 의존)

$$f_{M|Y}(m_i|y_{(0)i}, y_{(1)i}, \phi) = f_{M|Y}(m_i|y_{(0)i}, y_{(1)i}^*, \phi)$$

MNAR (Missing Not At Random)

missingness가 결측된 데이터값에 의존

2

Missing Data in Experiments

- 1. Introduction
- 2. The Exact Least Squares Solution with Complete Data
- 3. The Correct Least Squares Analysis with Missing Data
- 4. Filling in Least Squares Estimates
- 5. Barlett's ANCOVA Method
- 6. Least Squares Estimates of Missing Values by ANCOVA Using Only Complete-Data Methods
- 7. Correct Least Square Estimates of Standard Errors and One & More Degree of Freedom Sum of Squares

2.1 Introduction

전통적으로 '실험 상황'에서 Missing Data 문제가 중요하게 고려되어 왔다.

'실험 상황'에서는 실험자가 여러 조건을 원하는 대로 설계하고 결과를 관측하기 때문에

Missing이 일어나는 부분은 '관측'에 한정된다. → ①

다시 말하자면, '실험 상황'에서 Missing Data가 가지는 패턴은

0.3의 (a)에서 소개된 Univariate Nonresponse라는 것이다. → ②

(설명의 편의를 위해 Y_3 까지만 사용함)

왜 ①과 ②가 같은 말일까?

예를 들어서 보여주겠다.



2.1 Introduction

비슷한 키와 몸무게를 지닌 8명의 사람을 4개의 그룹으로 나누어서

운동 여부 및 식단조절 여부를 다르게 한 뒤 1주일이 지나 몸무게 변화를 관찰하는 실험을 진행했다

그런데 6일째 되는 날 피실험자 한 명이 실종되었고, 그의 몸무게를 측정할 수 없게 되었다.

나머지 7명에 대해서는 모두 데이터가 잘 수집되었고, 그 결과는 아래와 같다.

	식단조절 X	식단조절 O
운동 X	0, +3	-3, -2
운동 O	-1, +1	-5, ?

이때 '운동'과 '식단'이라는 각 factor에 대해 실행한 경우 1, 실행하지 않은 경우 0을 부여하고,

각 피실험자를 unit으로 하여 이 결과를 다시 정리해 보면

	운동	식단조절	몸무게 변화
피실험자 1	0	0	0
피실험자 2	0	0	3
피실험자 3	0	1	-3
피실험자 4	0	1	-2
피실험자 5	1	0	-1
피실험자 6	1	0	1
피실험자 7	1	1	-5
피실험자 8	1	1	?

$$Y_{obs} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 3 \\ 0 & 1 & -3 \\ 0 & 1 & -2 \\ 1 & 0 & -1 \\ 1 & 0 & 1 \\ 1 & 1 & -5 \\ 1 & 1 & ? \end{bmatrix}$$

여기서 잘 생각해 보면, '운동'과 '식단조절' column에 대해서는 결측이 일어날 수 없다.

왜냐하면 실험자가 처음부터 '설계'한 실험의 조건이기 때문이다.

즉, 결측이 일어나는 것은 실험의 결과, 다시 말해 '몸무게 변화' column에서만이며,

이는 Missing Data Pattern 중 Univariate Nonresponse에 해당한다.

$\therefore ① \Leftrightarrow ②$

2.2 The Exact Least Squares Solution with Complete Data

만약 위 예시에서 피실험자 8의 몸무게 변화가 결측되지 않았다면 (-9kg라 하자)
이는 Complete Data일 것이다.

Missing Data를 포함한 Incomplete Data를 분석하기 전에,
Complete Data를 Least Squares로 분석하는 방법에 대해 짚고 넘어가도록 하자.

운동 여부 및 식단 조절 여부를 독립변수, 몸무게 변화를 종속변수로 하는
multiple linear regression을 생각해 보자. (intercept 포함)

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$
$$\epsilon_i \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

or

$$Y = X\beta + \epsilon$$
$$\epsilon \sim \mathcal{N}_n(0, \sigma^2 I)$$

$$X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \quad y = \begin{bmatrix} 0 \\ 3 \\ -3 \\ -2 \\ -1 \\ 1 \\ -5 \\ -9 \end{bmatrix}$$

회귀분석에 대한 배경지식이 있다면, 위 수식들을 이해하는 데에 큰 무리는 없을 것이다.

그러나 beta 추정량을 본격적으로 계산하기 전에, 한 가지 의문이 든다.

두 개의 독립변수, 운동(x_1)과 식단조절(x_2)가

모두 각각 0과 1만을 가지는 categorical (dummy) variable인데,

왜 two-way ANOVA를 하지 않고 regression을 하는가?

결론부터 말하자면 두 분석은 어디에 집중하나의 차이를 가지고 있을 뿐이다.

ANOVA는 독립변수가 종속변수에 유의미한 영향을 주느냐에 주안점을 두고,
(운동 여부와 식단조절 여부가 몸무게 변화에 유의미한 영향을 주는가)

Regression은 coefficient를 추정하는 데에 집중한다.

(운동 여부와 식단조절 여부가 몸무게 변화에 얼마만큼의 영향을 주는가)

2.2 The Exact Least Squares Solution with Complete Data

다시 수식으로 돌아와서 beta 추정량을 구해보자.

$$\begin{aligned}\hat{\beta} &= (X^T X)^{-1} X^T y \\ &= \begin{bmatrix} 9/4 \\ -3 \\ -11/2 \end{bmatrix} = \begin{bmatrix} 2.25 \\ -3 \\ -5.5 \end{bmatrix}\end{aligned}$$

$$\hat{y} = 2.25 - 3x_1 - 5.5x_2$$

즉, 식단조절 여부를 고정했을 때

운동을 하는 것이 하지 않는 것에 비해 몸무게 변화량 기댓값이 약 -3kg

운동 여부를 고정했을 때

식단조절을 하는 것이 하지 않는 것에 비해 몸무게 변화량 기댓값이 약 -5.5kg

라고 해석할 수 있겠다.

2.3 The Correct Least Squares Analysis with Missing Data

이제 본격적으로 Missing Data가 존재하는 자료를 분석하는 방법에 대해 알아보겠다.

앞선 예시에서 몸무게를 측정하기 전에 실종되어버린 피실험자 8의 데이터를 날려버리고, 남은 7명의 데이터로만 분석하는 것으로 이해할 수 있겠다.

2.3.1 Ignore

Missing Data를 처리하는 첫 번째 방법은,

전체 데이터에서 Missing Data를 포함하는 row(unit)를 무시하고 분석하는 것이다.

	운동	식단조절	몸무게 변화
피실험자 1	0	0	0
피실험자 2	0	0	3
피실험자 3	0	1	-3
피실험자 4	0	1	-2
피실험자 5	1	0	-1
피실험자 6	1	0	1
피실험자 7	1	1	-5

이제 이 데이터에는 결측이 존재하지 않는다는 점에서,
마치 2.2에서 논의했던 Complete Case와 같이 변해버렸다고 할 수 있다.
2.2에서와 같은 방법으로 Regression을 진행해 보자.

$$X_* = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \quad y_* = \begin{bmatrix} 0 \\ 3 \\ -3 \\ -2 \\ -1 \\ 1 \\ -5 \end{bmatrix}$$

$$\begin{aligned}\hat{\beta}_* &= (X_*^T X_*)^{-1} X_*^T y_* \\ &= \begin{bmatrix} 17/10 \\ -19/10 \\ -22/5 \end{bmatrix} = \begin{bmatrix} 1.7 \\ -1.9 \\ -4.4 \end{bmatrix}\end{aligned}$$

$$\hat{y} = 2.25 - 3x_1 - 5.5x_2$$

2.3 The Correct Least Squares Analysis with Missing Data

이렇게 Missing Data를 포함한 unit을 무시하고 분석하는 방법은
결측된 Y값(몸무게 변화량)이 그 참값과 관련이 없을 때 (\Leftrightarrow MAR)
합리적이라고 할 수 있다.

▼ 부연설명

만약 피실험자 8이 운동과 식단조절을 병행하면서 몸무게가 9kg 빠져버렸고,
이로 인한 급격한 건강 악화로 쓰러져서 몸무게를 재려 오지 못했다고 하자.

그렇다면 -9kg이라는 Y값 자체가 그 값이 missing되는 데에 기여했다고 볼 수 있지 않은가? (MAR X)
앞으로 같은 실험을 또 하더라도 운동과 식단조절을 모두 하는 사람 중 체중이 급격하게 빠진 사람은
비슷한 이유로 결측을 발생시킬 수 있지 않겠는가?
(즉, 앞으로도 $(X_1, X_2) = (1, 1)$ 에서 관측했을 때 Y로 절댓값이 큰 음수가 잘 관측되지 않을 수 있음)

그렇다면 이 실험에서 결측을 무시하는 분석은 $\hat{\beta}_0, \hat{\beta}_1$ 을
underestimate한다는 점에서 합리적이라고 할 수 없을 것이다.

실제로 2.2에서

운동을 하는 경우가 하지 않는 경우에 비해 몸무게 변화량 기댓값이 약 -3kg
식단 조절을 하는 경우가 하지 않는 경우에 비해 몸무게 변화량 기댓값이 약 -5.5kg
으로 추정되었었는데,

피실험자 8의 데이터를 무시한 결과 -1.9kg, -4.4kg으로 절댓값이 작아진 것을 확인할 수 있다.

그런데 MAR일 때도 결측을 무시하고 분석하는 것이 항상 가능하냐에 대한 의문을 던질 수 있다.
Normal equation $X^T X \hat{\beta} = X^T y$ 에서 unique solution $\hat{\beta}$ 이 존재하기 위한 조건은
 $X^T X$ 가 invertible하다는 것인데, 과연 X 에서 결측된 y 값에 대응하는 row를 삭제한 후에도
(X 가 X_* 이 된 후에도) 이것이 여전히 성립할 것인가?

책에서는 Dodge라는 학자가 이에 대해 자세히 서술했다고만 나오고,

그냥 nonsingular로 가정하고 가자고 나온다.

실제로 우리가 사용한 예시에서도

결측이 발생한 피실험자 8에 대응하는 마지막 행을 지웠음에도 문제없이 역행렬을 구할 수 있었다.

▼ 부연설명

$$\text{rank}(X_*^T X_*) = \text{rank}(X_*)$$

임을 고려했을 때 잘린 design matrix X_* 가 rank deficient만 아니면 될 것이다.

예를 들어 이전에 사용했던 데이터에서

$$X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

운동이나 식단 조절 중 한 쪽만 한 사람들의 몸무게 변화 데이터가 전부 결측이라면

(이정도는 결측이 많아야 rank deficient 만들 수 있음)

$$X_* = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix}$$

rank가 2가 되면서 문제가 발생하겠지만,

이렇게 결측이 많은 경우는 애초에 정보가 너무 없으니 일단 배제하고 생각하는 것이 맞겠다.

2.3 The Correct Least Squares Analysis with Missing Data

2.3.2 Impute

Missing Data를 처리하는 두 번째 방법은,
결측이 발생한 부분을 특정 값으로 대체하는 것이다.

잊지 말자.

우리는 실험 상황에서의 결측치를 다루고 있다.

즉, Y의 마지막 column (관측 column) 에서만
결측이 발생하는 경우를 가정하고 있는 것이다.



만약 Missing Type이 MCAR이라고 하자.

MCAR은 MAR과 달리 missingness가 관측된 값에도 의존하지 않는다.

즉, 특정 row에서 Y_3 값이 missing 되었을 때

그 row의 Y_1 , Y_2 값을 봐도 missing된 값을 추정하는 데 아무런 도움이 되지 않는다는 뜻이다.

이럴 때에는 관측된 Y_3 값들을 통해 결측된 Y_3 값을 추정할 수 있다.

대표적으로 평균(mean)으로 대체할 수 있을 것이고,

이상치가 존재한다면 중위값(median),

categorical data라면 최빈값(mode)으로 대체할 수 있을 것이다.

	운동	식단조절	몸무게 변화
피실험자 1	0	0	0
피실험자 2	0	0	3
피실험자 3	0	1	-3
피실험자 4	0	1	-2
피실험자 5	1	0	-1
피실험자 6	1	0	1
피실험자 7	1	1	-5
피실험자 8	1	1	?

예를 들어 몸무게 변화의 관측 여부가 운동이나 식단조절 여부와 관련이 없다면,

피실험자 8의 몸무게 변화를 피실험자 1 ~ 7의 몸무게 변화 평균으로 대체할 수 있을 것이다.

(R파프 강의 중 데이터프레임에서 결측치를 column별 평균으로 대체하는 코드를 배우는데,
이러한 내용과 맥락을 같이 한다고 보면 된다.)

그런데 만약 Missing Type이 MAR이라고 한다면

missingness가 관측된 값에 의존하기 때문에

특정 row에서 Y_3 값이 missing 되었을 때

그 row의 Y_1 , Y_2 값이 missing value를 추정하는 데 힌트를 제공한다.

2.4 Filling in Least Squares Estimates

그렇다면 Y_1, Y_2 값으로부터 어떻게 Y_3 값을 추정해 낼 것인가?

당연하게도 Y_1, Y_2, Y_3 가 모두 관측된 데이터들로부터

Y_1, Y_2 와 Y_3 의 상관관계를 밝혀내야 할 것이다.

2.3.1에서 결측을 포함한 행을 무시하고 Least Squares를 했던 것을 기억하는가?

$$\hat{y} = 2.25 - 3x_1 - 5.5x_2$$

이 추정회귀식을 통해 피실험자 8의 몸무게 변화를 추정할 수 있다.

운동을 했고 ($x_1=1$), 식단조절도 했으므로 ($x_2=1$)

추정된 몸무게 변화는 $2.25 - 3 - 5.5 = -6.25$ 가 된다.

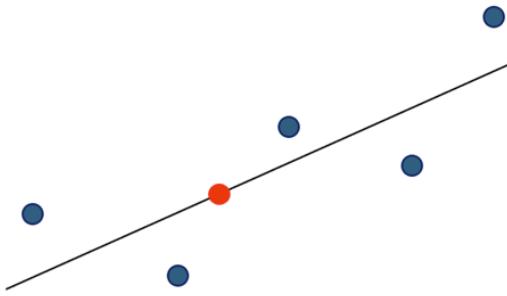
이 값으로 결측값을 대체할 수 있겠다.

2.4 Filling in Least Squares Estimates

2.4.1 Yates' Method

$$\text{Let } \tilde{y}_i = \begin{cases} \hat{y}_{k(k)} & i = k \\ y_i & i \neq k \end{cases}$$

$$\begin{aligned} b_{(k)} &= \underset{\beta}{\operatorname{argmin}} \sum_{i \neq k}^n (y_i - x_i^T \beta)^2 \\ &= \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^n (\tilde{y}_i - x_i^T \beta)^2 \end{aligned}$$



갑자기 이 이야기를 왜 했을까?

k번째 data가 Missing Data라고 생각해 보자.

(그러니까 바로 피실험자 8이라는 것이다.)

우리는 앞서 Missing Data를 제외하고 Regression하여 추정회귀선을 얻은 후, 이를 통해 Missing Data를 추정하고 집어넣었다.

(y_k 를 $\tilde{y}_{k(k)}$ 로 대체, 피실험자 8의 몸무게 변화를 -6.25로 대입)

이렇게 보완된 Complete Data를 가지고 다시 추정회귀선을 구하면
Observed Data만 사용해서 구한 추정회귀선과 일치한다는 소리다.

▼ 간단히 확인해 보자

OLS를 이용한 회귀분석에서

n개의 data 중 k번째 data를 빼고 fitting한 regression line은

n개의 data를 모두 사용하되

'k번째 data를 앞서 fitting한 line으로 추정한 값으로 사용하여' fitting한 regression line과 일치한다.

이는 수식으로도 보일 수 있지만 직관적으로 당연한 것이,

Least Squares 자체가 회귀선까지의 수직거리의 제곱합을 최소화하는 것이므로

회귀선 위에 새로운 data point가 찍힌다면

해당 point로부터 회귀선까지의 수직거리는 이미 0이므로

잔차제곱합을 최소화하기 위해 회귀선을 움직일 필요 자체가 없다.

$$X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \quad y_{imp} = \begin{bmatrix} 0 \\ 3 \\ -3 \\ -2 \\ -1 \\ 1 \\ -5 \\ -6.25 \end{bmatrix}$$

$$\begin{aligned} \hat{\beta}_{imp} &= (X^T X)^{-1} X^T y_{imp} \\ &= \begin{bmatrix} 1.7 \\ -1.9 \\ -4.4 \end{bmatrix} \\ &= \hat{\beta}_* \end{aligned}$$

2.4 Filling in Least Squares Estimates

이제 전체 데이터 n 개 중 missing data가 m 개인 경우로 확장해 보자. (observed data는 $n-m=r$ 개)

LSE로 보완된 Complete Data를 가지고 추정한 β 는 다음과 같다.

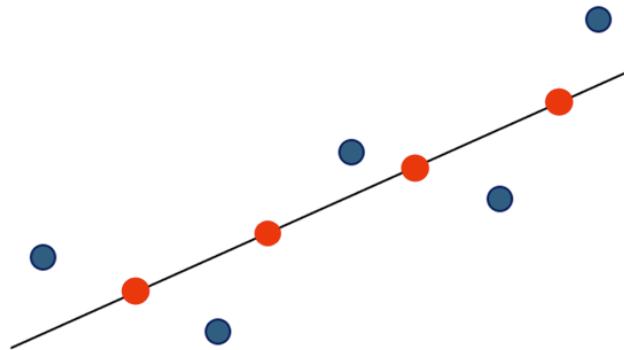
$$\hat{\beta}_* = \underset{\beta}{\operatorname{argmin}} \left[\sum_{i=1}^m (\hat{y}_i - x_i^T \beta)^2 + \sum_{i=m+1}^n (y_i - x_i^T \beta)^2 \right]$$

그런데 Observed Data만 가지고 β 를 추정했을 때,

앞의 term은 0이 되고, LSE의 정의에 의해 뒤의 term이 최소화되므로,

여전히 두 추정회귀선은 같다.

▼ 쉽게 말해서, 회귀선 위에 점을 아무리 추가해 봤자다



2.4 Filling in Least Squares Estimates

이때 n 개가 아닌 r 개의 data를 활용했으므로 σ^2 의 추정량은 다음과 같다.

$$\begin{aligned}s_*^2 &= \frac{1}{r-p} \sum_{i=m+1}^n (y_i - x_i^T \hat{\beta}_*)^2 \\&= \frac{n-p}{r-p} \times s^2 \\&= \frac{n-p}{r-p} \times \frac{1}{n-p} \left[\sum_{i=1}^m (\hat{y}_i - x_i^T \hat{\beta}_*)^2 + \sum_{i=m+1}^n (y_i - x_i^T \hat{\beta}_*)^2 \right]\end{aligned}$$

여기서 Missing Data를 LS로 추정하여 대체하고 분석하는 것의 단점을 생각해볼 수 있는데,

$$\widehat{Cov}(\hat{\beta}) = \hat{\sigma}^2 (X^T X)^{-1} \text{ 의 모든 원소들이 shrunk된다는 것이다.}$$

왜냐하면 우리는 관측된 r 개의 data만 활용해 회귀분석을 했고 이때 $\hat{\sigma}^2$ 은 s_*^2 였는데,

결측된 m 개의 data를 LS로 채워넣고 회귀분석을 한다는 것은

마치 처음부터 Complete Data였던 것처럼 취급하겠다는 것,

즉 SSE를 $r-p$ 가 아닌 $n-p$ 로 나누겠다는 것이고, ($SSE = SSE_*$)

이때 $\hat{\sigma}^2$ 은 s^2 으로

s_*^2 에 비해 $\frac{r-p}{n-p}$ 만큼 shrunk되기 때문이다.

그러나 전체 Data에서 Missing Data가 차지하는 비율이 크지 않다면,

다시말해 어떤 c 에 대해 $0 < c < \frac{r-p}{n-p} < 1$ 이라면,

이로 인한 편향은 미미한 수준일 것이다.

2.4 Filling in Least Squares Estimates

2.4.2 Using a Formula for the Missing Values

앞서 우리는 Missing Data에 집어넣을 LSE를 구하기 위해

Observed Data만을 이용해 추정회귀식을 도출하고,

조건에 맞는 값을 여기에 대입해야 했다.

2.4.2의 내용은,

Randomized Block Design에서 Missing Data가 1개만 존재하는 경우

특정 공식을 통해 결측치를 대체할 LSE를 빠르게 구할 수 있다는 것이다.

*특수한 상황에 대해 논하고 있음을 인지하자

위와 같이 y_{tb} 에서 결측이 발생했다고 하자.

$$y_+^{(t)} = \sum_{j=1}^B y_{tj}$$

$$y_+^{(b)} = \sum_{i=1}^T y_{ib}$$

$$y_+ = \sum_{(i,j) \neq (t,b)} y_{ij}$$

라고 정의할 때,

$$\frac{Ty_+^{(t)} + By_+^{(b)} - y_+}{(T-1)(B-1)}$$

위 식의 값은, LS를 이용해 추정한 y_{tb} 의 값과 일치한다.

Treatment \ Block	1	2	...	b	...	B
1	y_{11}	y_{12}	...	y_{1b}	...	y_{1B}
2	y_{21}	y_{22}	...	y_{2b}	...	y_{2B}
:	:	:	..	:		:
t	y_{t1}	y_{t2}	...	?	...	y_{tB}
:	:	:		:	..	:
T	y_{T1}	y_{T2}	...	y_{Tb}	...	y_{TB}

2.4 Filling in Least Squares Estimates

(example)

자동차의 주행 속도 (느림, 중간, 빠름)에 따른 타이어 마모 정도를 측정하기 위한 실험을 설계했다.

이때 4개의 타이어 제조사에서 각각 3개씩의 타이어를 선정해 각 속도에서 실험을 진행했다고 하자.

즉, Independent Variable (Treatment)은 주행 속도($T=3$), Blocking Variable은 제조사($B=4$)가 된다.

실험 결과는 다음과 같다.

Treatment \ Block	Brand A (1 0 0)	Brand B (0 1 0)	Brand C (0 0 1)	Brand D (0 0 0)
Slow (1 0)	3.7	3.4	3.5	3.2
Medium (0 1)	4.5	?	4.1	3.5
Fast (0 0)	3.1	2.8	3.0	2.6

결측값을 LS로 메꿔 보자.

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon$$

X_1, X_2 는 주행 속도를 나타내는 dummy variable,

X_3, X_4, X_5 는 제조사를 나타내는 dummy variable이다.

$$X_* = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad y_* = \begin{bmatrix} 3.7 \\ 3.4 \\ 3.5 \\ 3.2 \\ 4.5 \\ 4.1 \\ 3.5 \\ 3.1 \\ 2.8 \\ 3.0 \\ 2.6 \end{bmatrix}$$

$$\hat{\beta}_* = (X_*^T X_*)^{-1} X_*^T y_*$$

$$= \begin{bmatrix} 607/240 \\ 23/40 \\ 91/80 \\ 2/3 \\ 17/60 \\ 13/30 \end{bmatrix}$$

$$\hat{y}_{22}^{LS} = \frac{607}{240} + \frac{91}{80} \times 1 + \frac{17}{60} \times 1 = \frac{79}{20}$$

2.4 Filling in Least Squares Estimates

이제 위에서 소개된 공식을 이용해서 결측값을 메꾸어 보자.

3.7	3.4	3.5	3.2
4.5	?	4.1	3.5
3.1	2.8	3.0	2.6

4.5	3.4	4.1	3.5
4.5	?	4.1	3.5
4.5	2.8	4.1	3.5

3.7	3.4	3.5	3.2
4.5	?	4.1	3.5
3.1	2.8	3.0	2.6

3.4	3.4	3.4	3.4
4.5	?	4.1	3.5
2.8	2.8	2.8	2.8

3.7	3.4	3.5	3.2
4.5	?	4.1	3.5
3.1	2.8	3.0	2.6

$$\frac{Ty_+^{(t)} + By_+^{(b)} - y_+}{(T-1)(B-1)} = \frac{3(4.5 + 4.1 + 3.5) + 4(3.4 + 2.8) - (3.7 + \dots + 2.6)}{(3-1)(4-1)}$$

$$= \frac{237/10}{6}$$

$$= \frac{79}{20}$$

훨씬 간단한 계산으로 같은 값을 도출해내는 것을 확인할 수 있다.

더불어 분자를 살펴보면 $T(B-1) + B(T-1) - (TB-1) = TB - T - B + 1$ 개의 항이 더해지므로, 분모에 $(T-1)(B-1)$ 이 자리하는 것이 scaling의 관점에서 합리적이라는 생각이 들지 않는가?

2.4 Filling in Least Squares Estimates

2.4.3 Iterating to Find the Missing Value

앞서 우리는 Least Squares를 이용해 결측치를 어떤 값으로 대체해야

SSE를 가장 작게 만들 수 있는지 한번에 계산해 냈다.

2.4.3에서는 결측치에 여러가지 값을 넣어 보면서

SSE를 작게 하는 값을 선택해 나가는 방법을 이야기한다.

먼저 Hartley의 방법을 소개하자면, 하나의 결측치가 있을 때 3개의 값(trial values)으로 대체해 보고, 그 중 SSE가 가장 작은 값을 골라내자는 것이다.

그러나 이러한 방법은 처음 선정한 trial value들 중에서만 최적값을 골라낸다는 점에서, 처음 선정한 값들이 최적해와 거리가 멀다면 도토리 키 재기를 하는 꼴이 된다.

게다가 Hartley는 결측값이 여러 개 존재하는 경우 이 방법을 반복해서 사용할 것을 제안하는데, 결측값이 n 개 있다면 3^n 개의 조합을 테스트해 보아야 한다는 점에서, 결측치의 개수가 많아짐에 따라 계산량이 기하급수적으로 증가할 것이다.

이에 Healy and Westmacott은 더 나은 Iterative technique을 제시하는데, 그 절차는 다음과 같다.

1. 결측치에 trial value를 채워넣는다
2. Complete Data에 대해 회귀분석을 수행한다
3. 추정회귀선으로부터 결측치가 있던 자리의 predicted value를 계산한다.
4. 해당 predicted value로 결측치를 대체한다
5. 이 Complete Data에 대해 다시 회귀분석을 수행하고, 결측치를 predicted value로 대체한다.
6. missing value 자리의 값이 크게 달라지지 않거나, SSE의 감소분이 충분히 미미해질 때까지 반복한다.

2.4.1에서 본 것과 달리,

관측된 값들로 LS를 해서 결측치를 채워넣는 것이 아니라 임의의 trial value를 시도하는 것으로, 결측치를 채워넣은 뒤 LS를 하면 추정회귀선이 바뀐다.

즉, 새로 찍는 점이 직선 위에 있지 않다는 뜻이다.

이러한 방식은 LS로 최적해(SSE를 최소화하는 값)를 한번에 구할 수 없는 상황에서, 마치 Gradient Descent처럼 임의의 값에서 시작해 SSE를 감소시키는 방향으로 조금씩 나아가면서 최적해에 가까워질 수 있다는 점에서 의의가 있다고 생각한다.

2.4 Filling in Least Squares Estimates

2.4.4 ANCOVA with Missing Value Covariates

2.4.3에서 살펴본 방법들은

새로운 값을 결측치에 채워 넣고 추정회귀선을 구하는 행위를 계속 반복(Iteration)하는 방법이었다.

(Hartley : 계속 짹음, H&W: 처음 한번 짹고 계속 update해 나감)

2.4.4에서는 이러한 반복적인(Iterative) 방법 대신, 비반복적인(Noniterative) 방법을 소개한다.

여기서는 처음 한 번만 짹고, 단 한 번의 update를 통해 LS 추정값과 동일한 값을 산출해낼 것이다.

여기에서의 핵심 아이디어는 똑같이 Regression을 하되,

결측이라는 하나의 속성에도 dummy variable을 부여하겠다는 것이다.

(관측된 데이터에는 0, 결측치에는 1을 할당하는 것이다.)

▼ ANCOVA?

책에서는 해당 내용을 'ANCOVA'라고 설명하는데, 이것이 무엇인지 알아보자

ANCOVA(공분산분석)는 ANalysis of COVariance의 약자로,

독립변수(IV)가 종속변수(DV)에 미치는 영향을 보고자 할 때,

IV 외에 DV에 영향을 미치는 다른 변수(Covariate, CV, 공변량)을 고려한다.

우리는 여기서 'Missing'이라는 변수를 공변량으로 고려하고자 하는 것이다.

2.4 Filling in Least Squares Estimates

이제 구체적인 절차를 살펴보자.

- 결측치를 임의의 값(initial value)으로 채워 넣는다.
- 결측을 표현하는 dummy variable을 모델에 포함하여 회귀분석을 진행한다.
- 결측치의 LS 추정값 = (initial guess) - (결측을 나타내는 dummy variable의 계수 추정값)

(example)

Treatment \ Block	Brand A (1 0 0)	Brand B (0 1 0)	Brand C (0 0 1)	Brand D (0 0 0)
Slow (1 0)	3.7	3.4	3.5	3.2
Medium (0 1)	4.5	?	4.1	3.5
Fast (0 0)	3.1	2.8	3.0	2.6

첫째, 결측치를 임의의 값 4.0으로 채워 넣자.

$$y = \begin{bmatrix} 3.7 \\ 3.4 \\ 3.5 \\ 3.2 \\ 4.5 \\ 4.0 \\ 4.1 \\ 3.5 \\ 3.1 \\ 2.8 \\ 3.0 \\ 2.6 \end{bmatrix}$$

$$X = \left[\begin{array}{cccccc|c} 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{array} \right]$$

$$\hat{\beta}_* = (X_*^T X_*)^{-1} X_*^T y_*$$

$$= \begin{bmatrix} 607/240 \\ 23/40 \\ 91/80 \\ 2/3 \\ 17/60 \\ 13/30 \\ 1/20 \end{bmatrix}$$

둘째, 결측을 표현하는 dummy variable을 모델에 포함하여 회귀분석을 진행하자.

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \epsilon$$

2.4.2에서와 같이, X_1, X_2 는 주행 속도를 X_3, X_4, X_5 는 제조사를 나타내는 dummy variable이다.

이때, X_6 을 결측을 표현하는 dummy variable로 지정하자.

셋째, 결측치의 LS 추정값 = (initial guess) - (결측을 나타내는 dummy variable의 계수 추정값)

$$\hat{y}_{22}^{LS} = 4.0 - \frac{1}{20} = \frac{79}{20}$$

2.4 Filling in Least Squares Estimates

만약 결측치가 여러 개(m개)일 경우 어떻게 하면 될까?

결측이라는 속성을 나타내는 dummy variable을 m개 만들면 된다.

▼ 부연설명

Categorical variable을 모델에 포함하고자 할 때,

design matrix X의 column들 간 linearly independent를 보장하기 위해

($X^T X$ 가 invertible해야 unique solution $\hat{\beta}$ 를 구할 수 있으므로)

카테고리 개수 - 1 만큼의 dummy variable을 활용하고

마지막 카테고리는 해당 dummy variable들을 전부 0으로 설정하는 것으로 표현한다.

그렇다면 결측치 m개에 대해 왜 m-1개가 아닌 m개의 dummy variable을 부여하는가?

결측치라는 기준으로 데이터를 분류할 때에는 1번째 결측치, ..., m번째 결측치뿐 아니라

'관측된 데이터'라는, 0 0 ... 0 으로 표현되는 마지막 카테고리가 있기 때문이다.

(example)

Treatment \ Block	Brand A (1 0 0)	Brand B (0 1 0)	Brand C (0 0 1)	Brand D (0 0 0)
Slow (1 0)	3.7	3.4	?	3.2
Medium (0 1)	4.5	?	4.1	3.5
Fast (0 0)	3.1	2.8	3.0	?

$$\left[\begin{array}{cccccc|ccc} 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array} \right] \rightarrow \left[\begin{array}{cccccc|ccc} 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{array} \right] = X$$

결측치 3개의 initial value는 모두 4.0으로 하자

$$\left[\begin{array}{c} 3.7 \\ 3.4 \\ 4.0 \\ 3.2 \\ 4.5 \\ 4.0 \\ 4.1 \end{array} \right] \rightarrow \left[\begin{array}{c} 4.0 \\ 4.0 \\ 4.0 \\ 3.7 \\ 3.4 \\ 3.2 \\ 4.5 \\ 4.1 \\ 3.5 \\ 3.1 \\ 3.5 \\ 2.8 \\ 3.1 \\ 3.0 \\ 2.8 \\ 4.0 \end{array} \right] = y$$

2.4 Filling in Least Squares Estimates

$$X\beta = \left[\begin{array}{cccccc|ccc} 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{array} \right] \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \gamma_0 \\ \gamma_1 \\ \gamma_2 \end{bmatrix}$$

마찬가지로 $\hat{\gamma}_0, \hat{\gamma}_1, \hat{\gamma}_2$ 를 구하면,

각 결측치에서 (initial value) - (estimated coefficient)를 통해 최소제곱추정값을 구할 수 있다.

$$= \left(\begin{array}{cccccc|ccc} 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{array} \right) \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \gamma_0 \\ \gamma_1 \\ \gamma_2 \end{bmatrix} = X\beta + Z\gamma$$

$$\hat{y}_{13}^{LS} = 4.0 - \hat{\gamma}_0, \quad \hat{y}_{22}^{LS} = 4.0 - \hat{\gamma}_1, \quad \hat{y}_{34}^{LS} = 4.0 - \hat{\gamma}_2$$

2.5.2 Notations for ANCOVA

Y : Missing Value들을 모두 초기값을 넣어 채운 $n \times 1$ Complete Vector.

\tilde{y}_i ($i = 1, \dots, m$) = Initial Guesses, 초기값

$Z = n \times m$ matrix of m Missing Value Covariates

1^{st} row: $(1, 0, \dots, 0)$

m^{th} row : $(0, \dots, 0, 1)$

last r rows : $(0, 0, \dots, 0)$

Model : $Y = X\beta + Z\gamma + e$

$\gamma = m \times 1$ column vector of m regression coefficients for the Missing Value Covariates.

이를 바탕으로 residual sum of squares를 최소화시키는 β 와 γ 는 다음과 같은 식으로 유도된다.

$$SS(\beta, \gamma) = \sum_{i=1}^m (\tilde{y}_i - x_i\beta - z_i\gamma)^2 + \sum_{i=m+1}^n (y_i - x_i\beta - z_i\gamma)^2$$

y_i 가 observed value이면 $z_i\gamma = 0$ 이고, y_i 가 missing value이면 $z_i\gamma_i = \gamma_i$ 이므로

다음과 같이 정리 가능하다.

$$SS(\beta, \gamma) = \sum_{i=1}^m (\tilde{y}_i - x_i\beta - \gamma_i)^2 + \sum_{i=m+1}^n (y_i - x_i\beta)^2$$

결국 이 식을 최소화시키기 위 앞 부분은 $\hat{\gamma}_i$, 뒷 부분은 $\hat{\beta}_i$ 같은 적절한 estimators가 필요하다.

2.5.3 The ANCOVA Estimators of Parameters and Missing Y-Value

$$SS(\beta, \gamma) = \sum_{i=1}^m (\tilde{y}_i - x_i\beta - \gamma_i)^2 + \sum_{i=m+1}^n (y_i - x_i\beta)^2$$

앞서 observed value인 $\beta = \hat{\beta}_*$ 을 대입하면 두 번째 term이 고정이 된다.

그렇다면 최소화를 위해 첫 번째 term의 γ_i 를 다음과 같이 설정하자.

$$\hat{\gamma}_i = \tilde{y}_i - x_i\hat{\beta}_*, \quad i = 1, \dots, m$$

이렇게 되면 첫 번째 term이 0이 되고 남은 식을 정리하면 다음과 같게 된다.

$$SS(\hat{\beta}_*, \hat{\gamma}) = \sum_{i=m+1}^n (y_i - x_i\hat{\beta}_*)^2$$

따라서 $(\hat{\beta}_*, \hat{\gamma})$ 가 $SS(\hat{\beta}_*, \hat{\gamma})$ 를 최소화하고, 이는 곧 Model의 최소제곱추정량임을 알 수 있다.

또한, $\hat{\gamma}_i = \tilde{y}_i - x_i\hat{\beta}_*$ 에서 $\tilde{y}_i = x_i\hat{\beta}_* + \hat{\gamma}_i$ 임을 알 수 있고 정리하면 다음과 같다.

Correct least squares predicted value for ith missing value

Initial guess for ith missing value – Coefficient of ith missing value covariate

결국 missing value의 올바른 추정값은 우리가 설정한 초기값에서 공변량 계수를 뺀 값으로 구할 수 있다

2.5.4 The ANCOVA Estimates of Residual Sums of Squares and Covariance Matrix of $\hat{\beta}$

우리가 도출한 최종적인 식 $SS(\hat{\beta}_*, \hat{\gamma}) = \sum_{i=m+1}^n (y_i - x_i \hat{\beta}_*)^2$ 의 자유도는

$n - m - p = r - p$ 일 것이다.

우리가 구한 저 식이 맞다면, $\frac{SS(\hat{\beta}_*, \hat{\gamma})}{n-m-p}$ 라는 MSE 와 s_*^2 의 값도 같을 것이다.

나아가 $Cov(\hat{\beta}_*)$ 가 r개의 observed value를 통해 구한 V_* 와도 같다면, 나머지 모든 값들 standard errors, Sum of Squares, 모든 검정들도 같은 결과를 도출할 것이다.

$Cov(\hat{\beta}_*) = \sigma^2(X^T X)^{-1}$ 인데, 앞의 σ^2 는 s_*^2 즉, MSE 로 추정할 수 있다.

그렇다면 $(X^T X)^{-1}$ 는 어떤 방식으로 도출하면 될까?

우리는 위의 식에서 단순 X 만이 아니라, Z 도 고려해야 하므로 다음과 같은 행렬을 생각할 수 있다.

$$(X, Z)^T (X, Z) = \begin{pmatrix} X^T X & Z^T X \\ X^T Z & Z^T Z \end{pmatrix}$$

왜 (X, Z) 를 사용할까?

β 와 γ 를 동시에 추정하려면 X 와 Z 를 결합한 행렬 (X, Z) 를 사용해야 한다. 이 결합 행렬은 model에 대한 전체 예측 변수 공간을 포함하고 있다. model에 대한 정확한 예측을 위해서는 X 와 Z 를 모두 결합한 행렬을 사용해야 한다.

2.5.4 The ANCOVA Estimates of Residual Sums of Squares and Covariance Matrix of $\hat{\beta}$

우리가 이 행렬 중 이용해야 하는 것은 1행 1열의 $X^T X$ 부분이고, 이는 역행렬을 통해 구할 수 있다.

$X^T X$ 의 역행렬은 일반적으로 다음과 같이 구할 수 있고 이를 U 라고 정의하면 다음과 같다.

$$U = [X^T X - (X^T Z)(Z^T Z)^{-1}(Z^T X)]^{-1}$$

z_i 의 정의에 의해,

$$X^T Z = \sum_{i=1}^m x_i^T z_i \text{이고}$$

$$Z^T Z = \sum_{i=1}^m z_i^T z_i = I_m \text{이므로}$$

$$(X^T Z)(Z^T Z)^{-1}(Z^T X) = (\sum_{i=1}^m x_i^T z_i)(\sum_{j=1}^m z_j^T x_j) \text{이다.}$$

또한 $z_i z_j^T = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$ 이므로

$$X^T X = \sum_{i=1}^m x_i^T x_i \text{만 남게 된다.}$$

$$\text{결국 최종적으로 } U = (\sum_{i=m+1}^n x_i^T x_i)^{-1} \text{이 된다.}$$

따라서 우리가 최종적으로 구하고 싶었던 $s_*^2 U = V_* = Cov(\hat{\beta}_*)$ 임이 증명 가능하다.

이는 곧 Bartlett's ANCOVA가 모든 요약값에 대해 least squares values를 도출한다는 말과 동치이다.

2.6 Least Squares Estimates of Missing Values by ANCOVA Using Only Complete-Data Methods

m개의 missing value에 대한 최소제곱추정량(least squares estimates)은 어떤 방법으로 구할 수 있을까?

앞서 언급했던 complete data를 이용한 ANOVA를 이용해서 ANCOVA를 적용할 수 있다.

말이 거창해서 그렇지, 별 것이 아니다. 결국 우리에게 주어진 목표는 missing value를 구하는 것이고,

이미 주어진 data들을 이용해 이 추정값을 구해보겠다는 거다.

The argument here will appeal to the ANCOVA results; a direct algebraic proof appears in Rubin (1972).

By ANCOVA theory, the vector $\hat{\gamma}$ can be written as

$$\hat{\gamma} = B^{-1}\rho, \quad (2.16)$$



Textbook에서는 이처럼 특별한 증명이나 설명 없이 그저 Rubin의 증명결과를 따라간다고 한다. 설명이 부족하고 직관적인 이해에 대한 한계도 존재하므로 어느 정도의 증명은 거치려고 한다.

B: $m \times m$ cross-products matrix for the residuals of the m missing value covariates after adjusting for the matrix X,

ρ : the $m \times 1$ vector of cross products of Y and the residuals of the missing-value covariates after adjusting for the X matrix.

adjusted는 그냥 X의 영향을 제거했다고 생각하면 된다.

결국 γ 를 구하기 위해선 B와 ρ 를 우선 계산하고, B의 역행렬도 구해야한다.

참고로 B가 singular한 경우(가역이지 않은 경우)는 특정 매개변수를 추정하는 것이 불가능하다는 의미이다. 예를 들어, 특정 treatment의 효과를 추정하고 싶은데, 그 treatment를 받는 모든 단위 (unit, case)의 데이터가 결측임을 의미한다.

Rubin's Proof

우선 Rubin의 내용을 살펴보자.

$Y = X\beta + Z\gamma + e$ 에서 우선 최소제곱법을 진행하면,

$SSR = \sum_{i=1}^n (y_i - X_i\beta - Z_i\gamma)^2$ 이고, 이를 행렬형태로 표현하면 다음과 같다.

$SSR = (Y - X\beta - Z\gamma)^T (Y - X\beta - Z\gamma)$. 이제 이걸 γ 에 대해 편미분 하게 되면

$$\frac{\partial SSR}{\partial \gamma} = -2Z^T(Y - X\beta - Z\gamma) = 0 \text{ 이다.}$$

$Z^T Y - Z^T X\beta - Z^T Z\gamma = 0$ 이고,

$Z^T Y = Z^T X\beta + Z^T Z\gamma$ 여기서 X 와 Z 가 adjusted 되었다고 생각하자. 즉, Z 에 대한 효과를 분석하고 싶기 때문에, Z 와 X 가 uncorrelated되었다고 생각하자는 것이다. 앞서 언급한 X 의 영향을 제거한다는 의미가 바로 이 것이다.

그러면 $Z^T Y = Z^T Z\gamma$ 이므로 최종적으로

$$\hat{\gamma} = (Z^T Z)^{-1} Z^T Y \text{이 도출된다.}$$

앞서 언급한 것처럼 $\hat{\gamma} = B^{-1}\rho$ 이므로 우리는 $Z^T Z = B$, $Z^T Y = \rho$ 임을 예측할 수 있다.

Proof and Difference With Notations In Textbook

textbook에서는 B 라는 행렬을 잔차행렬의 곱으로 설명한다. 따라서 위에서 유도한 Rubin의 방법과는 상이하지만, 이러한 이유에 대해서는 후술하도록 하겠다.

우선 B 를 계산하기 위해 Z 의 각 열(1~m) Z_j 에 대해 X 를 독립 변수로 하는 회귀 모델을 적합하자.

즉, $Z_j = X\beta_j + \epsilon_{Z_j}$ 라는 새로운 model을 생각해보자.

그렇다면 $\epsilon_{Z_j} = Z_j - \hat{Z}_j$ 일 것이고 $B = \epsilon_Z^T \epsilon_Z$ 이므로 행렬 B 의 j행 k열의 원소 b_{jk} 는

$$b_{jk} = \sum_{i=1}^n (z_{ij} - \hat{z}_{ij})(z_{ik} - \hat{z}_{ik})$$

그런데 최소제곱추정량의 성질에 의해 $\sum_{i=1}^n x_{il}(z_{ik} - \hat{z}_{ik}) = 0$ 이고

\hat{z}_{ik} 는 X 의 선형결합으로 표현 가능하므로 식은 다음과 같이 축소된다.

$$\sum_{i=1}^n z_{ij}(z_{ik} - \hat{z}_{ik}) = z_{jk} - \hat{z}_{jk}$$

$$\text{이와 비슷한 방법으로 } \rho_j = \sum_{i=1}^n y_i(z_{ij} - \hat{z}_{ij})$$

참고로 이 때 Y 는 initial guess를 추가한 상태이다.

Reason Why We Can Approve The Difference of Matrix B : Z vs ϵ_Z

B 를 구하는 과정에서 Z 와 잔차 ϵ_Z 를 혼용하는 것은 이론적으로 약간의 차이를 초래할 수 있지만, 실제 분석에서는 이 차이가 매우 작아 무시할 수 있는 수준일 수 있다. Z 가 0과 1로 구성되어 있으므로, 잔차 ϵ_Z 도 유사한 형태를 가질 가능성이 높다. 따라서 Z 를 사용한 분석과 ϵ_Z 를 사용한 분석은 실질적으로 큰 차이가 나지 않을 수 있다.

따라서 Z 를 사용하여 모델을 적합하고 분석하는 것은 실질적으로 유효하며, 잔차 ϵ_Z 를 사용한 분석과 거의 동일한 결과를 제공할 수 있다. 이는 결측값이 있는 데이터에서도 정확한 추정치를 얻는데 충분히 신뢰할 수 있는 방법이다.

$$X = \begin{bmatrix} 1 & 0.5 \\ 1 & 1.5 \\ 1 & 2.5 \\ 1 & 3.5 \end{bmatrix} \quad Y = \begin{bmatrix} 1.2 & 2.3 & \text{NA} \\ 3.4 & \text{NA} & 4.5 \\ 5.6 & 6.7 & 7.8 \\ 8.9 & 9.0 & 1.2 \end{bmatrix} \quad \hat{\gamma} = \begin{bmatrix} -0.1632 \\ -0.1632 \end{bmatrix} \quad \text{잔차를 사용한 경우}$$

$$Z = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \quad Y = \begin{bmatrix} 1.2 & 2.3 & 4.5 \\ 3.4 & 6.0 & 4.5 \\ 5.6 & 6.7 & 7.8 \\ 8.9 & 9.0 & 1.2 \end{bmatrix} \quad \hat{\gamma} = \begin{bmatrix} -0.1 \\ -0.1 \end{bmatrix} \quad \text{그냥 } Z \text{를 사용한 경우}$$

Some Problems Related To Other Sum of Squares and Standard Errors

일련의 과정을 거쳐 알맞은 estimate를 구하게 되면 나아가 올바른 Residual Sum of Square(SSE)와 Residual Mean Square(s^*)도 구할 수 있을 것이다. 이 과정에서 자유도는 missing data의 수인 m 만큼 감소할 것이긴 하지만. 그러나 SSE가 아닌 다른 Sum of Squares의 값은 비교적 크게 나올 수 있고, 이에 따라 표준 오차(Standard errors)도 작아질 수 있다. 풀어 설명하면, 관측된 data만을 기반으로 추정값을 구하기에, SSE는 정확하게 추정할 수 있어도 다른 제곱합의 정확한 추정은 어려울 수 있다. 그룹 간 변동성이나 treatment effect가 정확히 추정되지 않는 경우도 존재할텐데, 예를 들어 변동성에 대한 과소평가가 나타나게 되면, 그만큼 표준오차가 작아질 수 있다는 것이다.

2.7 Correct Least Squares Estimates of Standard Errors and One & More Degree of Freedom Sums of Squares

$\lambda = C^T \beta$ 라고 하자.

여기서 C 는 $p \times 1$ 크기의 constant vector이고, 결국 λ 가 의미하는 건 β 의 linear combination이다. 또한 이는 LSE를 이용해 missing value를 채워 넣은 $\hat{\lambda} = C^T \hat{\beta}$ 의 estimate이다.

LSE를 통해 얻은 $\hat{\beta} = \hat{\beta}_*$ 임을 알 수 있고, 같은 논리로 $\hat{\lambda} = \hat{\lambda}_*$ 이다.

가설 검정을 통해 SS와 SE의 값을 차례대로 구해보자.

$$H_0 : C^T \beta = 0 \text{ vs } H_1 : \text{not } H_0$$

where C is a $p \times 1$ vector.

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

$$\lambda = C^T \beta, \text{ so that } \hat{\lambda} = C^T \hat{\beta}$$

$$Var(\hat{\lambda}) = Var(C^T \hat{\beta}) = C^T Var(\hat{\beta}) C = \sigma^2 C^T (X^T X)^{-1} C$$

$$SE(\hat{\lambda}) = \sqrt{Var(\hat{\lambda})} \text{이므로}$$

$$SE = s \sqrt{(C^T (X^T X)^{-1} C)}$$

$$\begin{aligned} SS &= (\hat{\lambda} - 0)^T (Var(\hat{\lambda}))^{-1} (\hat{\lambda} - 0) \\ &= (C^T \hat{\beta})^T (Var(C^T \hat{\beta}))^{-1} (C^T \hat{\beta}) \\ &= \frac{1}{\sigma^2} (C^T \hat{\beta})^T ((C^T (X^T X)^{-1} C)^{-1}) (C^T \hat{\beta}) \end{aligned}$$

textbook에서 나온 SS는 그냥 여기서 σ^2 을 제거했다고 이해하면 된다.

2.7 Correct Least Squares Estimates of Standard Errors and One & More Degree of Freedom Sums of Squares

정리하면

$$SE = s\sqrt{(C^T(X^T X)^{-1} C)}$$

$$SS = \hat{\lambda}^2 / C^T(X^T X)^{-1} C$$
이고

올바른 제곱합(sum of squares)과 표준오차(standard error)를 구하면

$$SE_* = s_* \sqrt{C^T(X^T X)^{-1} C}$$
이고

$$SS_* = \hat{\lambda}_*^2 / C^T U C$$
이다. ($\because U = (\sum_{i=m+1}^n x_i^T x_i)^{-1}$)이 된다.

H 라는 새로운 vector를 생각해보자. H 는 complete data가 주어졌을 때 λ 의 estimate로, m개의 missing value의 covariates(Z)를 dependent variable로 regression을 통해 나타낼 수 있다.

위에서 사용한 model을 가지고 그대로 이를 유도해보면,

$$Z = X\beta + \epsilon$$

$$\hat{\beta} = (X^T X)^{-1} X^T Z$$
이므로

$$\hat{\lambda} = C^T \hat{\beta} = C^T (X^T X)^{-1} X^T Z$$

$$H^T = C^T (X^T X)^{-1} X^T Z$$
 (transpose는 행렬곱을 위한 차원조정이라고 생각)

여기서 주목할 부분이 있는데, 지금 우리가 구한 H 와 위에서 구했던 B 가 동시에 계산이 가능하다는 것이다. H 와 B 모두 complete data의 ANOVA와 missing data의 covariate를 통해 구해진다.

즉,

H 의 i번째 component와 B 의 ith row는 같은 방식으로 계산되었다는 뜻이다.

그러면 $C^T U C = C^T (X^T X)^{-1} C + H^T B^{-1} H$ 라고 할 수 있다.

전자는 앞서 정리한 complete에 대한 부분, 뒷부분은 missing 부분으로 이해할 수 있다.

$$s_*^2 = s^2(n - p) / (r - p)$$
 이므로 지금껏 계산한 것들을 총 정리하면

$$SE_* = \sqrt{\frac{n-p}{r-p} (SE^2 + s^2 H^T B^{-1} H)}$$
이고,

$$SS_* = SS / (1 + (SS / \hat{\lambda}^2) H^T B^{-1} H)$$
이다.

결국 자유도를 1로 갖는 SE와 SS 모두 complete data의 LSE로 도출이 가능하다.

2.7 Correct Least Squares Estimates of Standard Errors and One & More Degree of Freedom Sums of Squares

C 를 $p \times w$ 라는 matrix로 보면서 자유도가 더 큰 경우를 고려하자.

2.7에서는 한 개의 linear combination에 대한 것이라면, 2.8은 여러 개의 경우를 고려하는 것이고,
여기서

$\lambda = C^T \beta$ 는 더이상 스칼라가 아니라 하나의 벡터이다.

2.7처럼 $\hat{\lambda} = \hat{\lambda}_*$ 나 $\hat{\beta} = \hat{\beta}_*$ 라는 같은 가정을 이용할 것인데,
단순성을 위해 w 개의 linear combination에 대해 orthonormal을 가정하자. 즉,

$C^T(X^T X)^{-1}X = I_w$ 라는 뜻이다.

이렇게 되면 $Cov(\hat{\lambda}) = \sigma^2 I_w$ 이고 $SS = \hat{\lambda}^T \hat{\lambda}$ 가 될 것이다.

올바른 sum of square인 $SS_* = \hat{\lambda}_*^T (C^T U C)^{-1} \hat{\lambda}_*$ 가 될 것이다.

2.7과 마찬가지로, 적당한 H 라는 matrix를 설정하면

$SS_* = \hat{\lambda}^T (I + H^T B^{-1} H)^{-1} \hat{\lambda}$ 라고 표현 가능하고

$SS_* = SS - (H\hat{\lambda})^T (HH^T + B)^{-1} (H\hat{\lambda})$ 라고도 표현 가능하다.

첫 번째의 경우 $w \times w$ 인 symmetric matrix의 inverse를 다루고 있고, 두 번째는 $m \times m$ 의 matrix의
inverse를 다룬다. 일반적으로 $w < m$ 인 경우 첫 번째를 선호한다. 계산에서의 이점이 있기 때문이다.

3

Complete-Case & Available-Case Analysis

1. Complete-Case Analysis
2. Weighted Complete-Case Analysis
3. Available-Case Anlaysis

Complete-Case Analysis

Complete-Case Analysis (Listwise Deletion): 한 개라도 결측이 있는 자료들은 모두 제거한 후 모든 변수들이 정확한 값으로 다 채워져있는 관측치만을 대상으로 분석을 진행하는 방법 (어떤 결측치가 하나라도 존재하는 행 삭제)

- MCAR를 전제로 함: MCAR일 경우 행을 지워도 데이터 분포에 영향이 없기 때문에 해당 방법을 쓸 수 있다.
- 예시를 통해 살펴보자. 아래 표와 같은 Data가 있다고 할 때, CCA를 진행하면 다음과 같다. 변수에 대하여 N/A가 하나라도 있으면 해당 행을 삭제한 후 분석한다.

- 장점: (1) 간단함 (simplicity)
(2) 단변량 통계의 비교가 용이함 (comparability of univariate statistics)
- 단점: (1) 정보의 손실 (potential **loss of information: loss of precision & bias**)
(2) sample size가 줄어들어 통계의 검정력이 줄어들게 됨.
(3) MCAR이 아닌 이상 complete cases는 original sample의 random subsample이 아니다.
>> complete cases가 random sample이 아니면 평균, 표준편차 등과 같은 값을 구할 때 bias가 발생한다.

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Lite	99	70%
3	Fast+	167	N/A
4	Fast+	N/A	80%
5	Lite	76	70%
6	Fast+	155	10%
7	N/A	N/A	95%



Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Lite	99	70%
5	Lite	76	70%
6	Fast+	155	10%

Complete-Case Analysis

- loss of precision = larger variance
 - $\hat{\theta}_{CC}$ 를 complete units로부터 추정한 θ 의 estimate, $\hat{\theta}_{EFF}$ 를 available data를 바탕으로 추정한 θ 의 efficient estimate라 할 때, $\hat{\theta}_{CC}$ 의 loss of efficiency는 다음 식을 만족하는 Δ_{CC} 로 표현된다.

$$Var(\hat{\theta}_{CC}) = Var(\hat{\theta}_{EFF})(1 + \Delta_{CC})$$

- 자료가 MCAR인 것을 가정하고 진행하기 때문에 이 가정이 흔들리는 정도(데이터의 누락이 random이 아닐 때)에 따라 bias가 발생한다.
- 제거한 결측이 있는 자료들(incomplete units)은 MCAR 가정이 합리적인 가정이었는지 검정하기 위해 사용된다.
- 방법: Complete units의 특정한 변수 Y_j 와 incomplete units의 Y_j (여기서 Y_j 는 관측된 특정한 변수를 의미)의 분포를 비교한다.
 - 차이가 클수록 MCAR 가정이 invalid하고, CC analysis로부터 추정한 estimate가 biased일 가능성성이 높음을 의미한다.
- Weights를 사용하여 CC analysis에서 발생한 bias를 조정할 수 있다.

Weighted Complete-Case Analysis

CC analysis의 경우 MCAR인 경우를 제외하고는 bias가 발생할 수 있기 때문에, Weighted Complete-Case Analysis에서는 complete units에 가중치 (weight)를 부여하여 bias를 조정하는 방법을 활용한다. (CC Analysis에 weight을 추가한 방법)

여기서 사용되는 weight의 개념이 Randomization Inference 과정에서의 weighting과 유사하기 때문에, 이해를 위해 이를 먼저 살펴보자.

[Notations]

- Population size: N , sample size: n
- Number of variables (items): K
- Data: $Y = (y_{ij})$ where $i = 1, \dots, N$ and $j = 1, \dots, K$
- Design Information (about sampling or missingness): Z
- Sample Indicator: $I = (I_1, \dots, I_N)'$ for unit i

$$I_i = \begin{cases} 1, & \text{unit } i \text{ included in the sample} \\ 0, & \text{unit } i \text{ not included in the sample} \end{cases}$$

- Sample selection processes can be characterized by a distribution for I given Y and Z

Weighted Complete-Case Analysis

Randomization Inference에서 unit들은 확률적 표본추출 (Probability Sampling)을 따르는데, 이는 다음과 같은 두 가지 요건을 가진다.

- 1) Unconfounded
- 2) 각 unit이 표본으로 선택될 확률 (probability of selection)은 모두 0보다 크고 알려져 있다. Simple Random Sampling과 같은 equal probability sample design에서는 모든 unit에 대하여 π_i 가 동일하다.

$$\pi_i = E(I_i | Y, Z) = \Pr(I_i = 1 | Y, Z)$$

$$\pi_i = P(I_i = 1 | Z) > 0, \quad \text{for all } i$$

Complete Response survey를 가정하고 Z 가 strata를 정의하는 변수라면, 해당 표본 추출 과정은 층화 무작위 표본추출 (Stratified Random Sampling)이 된다.

- $Z = j, j = 1, \dots, J$
- N_j population units
- In stratum j , stratified random sampling takes SRS of n_j units
- Stratified random sampling 아래 I (sampling indicator)의 분포는 다음과 같다.

$$f(I | Y, Z) = f(I | Z) = \begin{cases} \prod_{j=1}^J \binom{N_j}{n_j}^{-1}, & \text{if } \sum_{i:z_i=j} I_i = n_j \text{ for all } j, \\ 0, & \text{otherwise} \end{cases}$$

Weighted Complete-Case Analysis

모평균 \bar{Y} 추정 비교: 일반적 방법 vs Weighting Methods

모평균을 추정해보자. Stratified Random Sampling이기 때문에 다음과 같이 구할 수 있다.

$$t = \bar{y}_{st} = \frac{1}{N} \sum_{j=1}^J N_j \bar{y}_j$$

$$\text{Var}(\bar{y}_{st}) = \frac{1}{N^2} \sum_{j=1}^J N_j^2 \left(\frac{1}{n_j} - \frac{1}{N_j} \right) S_j^2$$

$$v(Y_{\text{inc}}) = \frac{1}{N^2} \sum_{j=1}^J N_j^2 \left(\frac{1}{n_j} - \frac{1}{N_j} \right) s_j^2$$

모평균을 추정할 때 weighting methods를 사용하면 어떻게 표현할 수 있을까?

기본적인 아이디어는 π_i 의 확률로 선택되는 unit i가 전체 모집단의 π_i^{-1} 만큼을 상징하기 때문에 모집단의 평균, 분산 등을 추정할 때 가중치로 π_i^{-1} 만큼을 부여해야 한다는 점이다. 그러면 표본 평균은 다음과 같이 나타낼 수 있다.

$$t_{\text{HT}} = \sum_{i=1}^n y_i \pi_i^{-1} \quad : \text{Horvitz-Thompson estimate}$$

$$\bar{y}_{\text{st}} \equiv \bar{y}_w = \frac{1}{n} \sum_{i=1}^n w_i y_i \quad \pi_i = \frac{n_j}{N_j}, \quad w_i = n \cdot \frac{\pi_i^{-1}}{\sum_k \pi_k^{-1}}$$

Weighted Complete-Case Analysis

- Weighting with Nonresponses

weighting class estimator를 살펴보기 전, unit i 에 대하여 응답 확률(response probability) ϕ_i 를 알고 있을 때 어떻게 추정하는지 알아보자.

$$Pr(selection \text{ and } response) = Pr(selection) \times Pr(response|selection) = \pi_i \phi_i$$

위 식에 따라, \bar{y}_w 식은 다음과 같이 표현된다. 여기서 r 은 관측된 units이다 (응답자들). 앞서 complete response일 때의 식과 비교해보면, π_i 에 ϕ_i 가 곱해진 형태임을 확인할 수 있다.

$$\bar{y}_w = \frac{1}{r} \sum_{i=1}^r w_i y_i$$
$$w_i = r \cdot \frac{(\pi_i \phi_i)^{-1}}{\sum_k (\pi_k \phi_k)^{-1}}$$

하지만 실제 상황에서는 ϕ_i 를 모르는 경우가 더 많다. ϕ_i 를 모르는 경우에는 응답자와 미응답자로부터 얻은 데이터를 이용하여 ϕ_i 를 추정한 후, 추정된 ϕ_i 이용하여 추정을 진행한다. ϕ_i 를 추정하여 weight를 조정하는 방법에는 두 가지 방법이 있다

Weighted Complete-Case Analysis

1. Weighting Class Estimation

Sample을 J개의 “Weighting Classes”로 나누는 방법.

- n_j : sample size
- r_j : number of observed samples

$$r = \sum_{j=1}^J r_j$$

- simple estimator for ϕ_i : $\hat{\phi}_i = r_j/n_j$
- C : weighting class variable

ϕ_i 를 $\hat{\phi}_i$ 로 추정했기 때문에, weighting class j 의 responding units은 다음과 같은 weight를 가지게 된다.

$$w_i = r(\pi_i \hat{\phi}_i)^{-1} / \sum_{k=1}^r (\pi_k \hat{\phi}_k)^{-1}$$

where $\hat{\phi}_i = r_j/n_j$ for unit i in class j

Class내에서 sampling weight이 일정하지 않은 경우:

$$w_i = r(\pi_i \hat{\phi}_i)^{-1} / \sum_{k=1}^r (\pi_k \hat{\phi}_k)^{-1}$$
$$\bar{y}_W = \frac{1}{r} \sum_{i=1}^r w_i y_i$$

Equal Probability Design의 경우:

$$\bar{y}_{WC} = n^{-1} \sum_{j=1}^J n_j \bar{y}_{jR}$$

Weighted Complete-Case Analysis

1. Weighting Class Estimation

- 이 추정값은 MAR 가정 (quasirandomization) 하에 unbiased 하다.
 - **Quasirandomization (Assumption 3.1):** weighting class j 내의 데이터는 MCAR을 따른다.

그러면 Weighting class는 어떻게 설정하는 것이 좋을까?

(1) 가정 3.1을 만족하도록, (2) 가정 3.1 하에 추정치의 MSE (mean squared error)이 작아지도록 한다.

또한, bias와 variance에 주는 효과를 고려하여 설정한다.

		Association with outcome	
		Low (L)	High (H)
Association with nonresponse	Low (L)	Bias: —	Bias: —
	High (H)	Var: —	Var: ↓
	Low (L)	Bias: —	Bias: ↓
	High (H)	Var: ↑	Var: ↓

C와 Y가 약하게 관련되어 있을 때 (LL, HL): weighting은 bias를 감소하는데 큰 효과가 없다. 오히려 HL의 경우 sampling variance를 증가시켜 비효율적일 수 있다.

C와 Y가 강하게 관련되어 있을 때: sampling variance를 감소시키고, bias도 감소시키는 경향을 보인다.

Weighted Complete-Case Analysis

2. Propensity Weighting

조금 더 일반화된 방법으로, propensity score theory를 사용한다.

- X: set of variables observed for respondents and nonrespondents
- Data는 MAR이라고 가정, ϕ 는 unknown parameter

$$P(M|X, Y, \phi) = P(M|X, \phi)$$

- C가 X가 되면 Quasirandomization을 만족한다.

Response Propensity Stratification

Response Propensity for unit i: $\rho(x_i, \phi) = P(m_i = 0 | \rho(x_i, \phi), \phi)$

모든 x값에 대하여 response propensity는 0보다고 크다고 가정하면,

$$\begin{aligned} & Pr(m_i = 0 | y_i, \rho(x_i, \phi), \phi) \\ &= E(Pr(m_i = 0 | y_i, x_i, \phi) | y_i, \rho(x_i, \phi), \phi) \\ &= E(Pr(m_i = 0 | x_i, \phi) | y_i, \rho(x_i, \phi), \phi) \\ &= E(\rho(x_i, \phi) | y_i, \rho(x_i, \phi), \phi) \\ &= \rho(x_i, \phi), \text{ for all } x_i \end{aligned}$$

$Pr(M|\rho(X, \phi), Y, \phi) = Pr(M|\rho(X, \phi), \phi)$ 즉, propensity score $\rho(X, \phi)$ 를 갖는 strata 내에서 응답자들은 random subsample임을 의미한다.

Weighted Complete-Case Analysis

실제 상황에서는 ϕ 을 모르기 때문에 표본 데이터를 사용하여 이를 추정한다. 과정은 다음과 같다:

1. 응답자, 무응답자 데이터를 기반으로 X 에 대한 M 의 binary regression (logistic, probit, robit)으로부터 $\hat{\phi}$ 를 추정한다.
2. $\rho(X, \hat{\phi})$ 를 5개 혹은 10개의 값으로 나누는 grouped variable C 를 설정한다. (coarsening $\rho(X, \hat{\phi})$)
3. 같은 adjustment class j 내에서 응답자와 무응답자가 같은 값의 grouped propensity score을 갖도록 한다.

이 방법의 또 다른 방식은 propensity score의 역수 ($\rho(X, \hat{\phi})^{-1}$) 를 weight로 직접 사용하는 것이다.

이렇게 하면 nonresponse bias를 없앨 수 있다는 장점이 있다. 하지만, sampling variance가 급격하게 커질 수 있다는 단점이 발생한다.

Weighted Complete-Case Analysis

Inverse Probability Weighted for Generalized Estimating Equations (IPWGEE)

- $y_i = (y_{i1}, \dots, y_{iK})$: fully observed for $i = 1, \dots, r$, missing/partially observed for $i = r + 1, \dots, n$
- $\begin{cases} m_i = 1, & \text{if } y_i \text{ is incomplete} \\ m_i = 0, & \text{if } y_i \text{ is complete} \end{cases}$
- $x_i = (x_{i1}, \dots, x_{ip})^T$: vector of covariates (fully observed)
- $z_i = (z_{i1}, \dots, z_{iq})^T$: vector that can predict missing mechanism (auxiliary variables, fully observed)
- $g(x_i, \beta)$: regression function (unknown parameter β , 차원: d)

결측치가 없는 경우, GEE의 solution은 다음과 같다.

$$\sum_{i=1}^n D_i(x_i, \beta)(y_i - g(x_i, \beta)) = 0$$

결측치가 존재하는 경우, CC analysis에 의해 위 식은 아래와 같이 표현된다.

$$\sum_{i=1}^r D_i(x_i, \beta)(y_i - g(x_i, \beta)) = 0$$

$$P(m_i = 1|x_i, y_i, z_i, \phi) = P(m_i = 1|x_i, \phi)$$

Inverse-probability weighted generalized estimating equation (IPWGEE)에 의해, GEE solution은 아래와 같이 대체된다.

$$\sum_{i=1}^r w_i(\hat{\alpha}) D_i(x_i, \beta)(y_i - g(x_i, \beta)) = 0$$
$$w_i(\hat{\alpha}) = 1/p(x_i, z_i | \hat{\alpha})$$

이때 $p(x_i, z_i | \hat{\alpha})$ 는 complete unit일 확률의 추정값을 의미한다. 예를 들어 logistic regression을 이용한다고 할 때, $p(x_i, z_i | \hat{\alpha})$ 는 x_i, y_i 에 대한 m_i 의 logistic regression을 통해 얻을 수 있으며, α 는 logistic regression의 vector parameter이다.

$$P(m_i = 1|x_i, y_i, z_i, \phi) = P(m_i = 1|x_i, z_i, \phi)$$

위 식을 살펴보면, missingness가 z_i 에도 의존할 수 있다는 점에 CC analysis에 의해 표현된 식보다 덜 제한적이라고 할 수 있다. 따라서, IPWGEE는 z_i 에 대한 missingness mechanism의 의존성으로 인해 unweighted GEE의 bias를 보정할 수 있다.

Weighted Complete-Case Analysis

Post Stratification and Ranking to Known Margins

(1) Post-stratification

$\frac{N_j}{N}$ 이 알려진 경우, post-stratified mean \bar{y}_{ps} 를 통해 앞서 구한 \bar{y}_{wc} 를 대체할 수 있다.
Quasirandomization 가정 하에서 \bar{y}_{ps} 는 \bar{Y} 에 대한 불편추정량이다.

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

(2) Raking Ratio Estimation

weighting class가 각각 J, L 개의 수준을 갖는 두 요인 X_1, X_2 에 의해 joint하게 정의되는 경우 post-stratified mean은 다음과 같다.

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

이를 계산하려면 두 요인의 joint count인 N_{jl} 을 구해야 한다.

먼저 각 요인에 대한 marginal count인 $N_{j+} = \sum_{l=1}^L N_{jl}, N_{+l} = \sum_{j=1}^J N_{jl}$ 를 모든 $j = 1, \dots, J, l = 1, \dots, L$ 에 대해 알고 있다고 가정하자.

$\{N_{jl}\}$ 의 raking estimates인 $\{N_{jl}^*\}$ 는 다음과 같은 제한 조건을 만족한다.

(1) marginal constraints:

$$\textcircled{1} \quad N_{j+}^* = \sum_{l=1}^L N_{jl}^* = N_{j+}$$

$$\textcircled{2} \quad N_{+l}^* = \sum_{j=1}^J N_{jl}^* = N_{+l}$$

(2) $N_{jl}^* = a_j b_l n_{jl}$ for certain row constraints $\{a_j, j = 1, \dots, J\}$ and certain column constraints $\{b_l, l = 1, \dots, L\}$ 의 형태를 가진다.

이 때 아래와 같은 iterative proportional fitting procedure의 수렴값을 $\{N_{jl}^*\}$ 로 사용한다.

$$\begin{aligned} N_{jl}^{(1)} &= n_{jl}(N_{j+}/n_{j+}) \\ N_{jl}^{(2)} &= N_{jl}^{(1)}(N_{+l}/N_{+l}^{(1)}) \\ N_{jl}^{(3)} &= N_{jl}^{(2)}(N_{j+}/N_{j+}^{(2)}) \\ &\vdots \end{aligned}$$

이렇게 구한 $\{N_{jl}^*\}$ 를 위 식에 대입하면 \bar{Y} 에 대한 raked estimate \bar{y}_{rake} 를 얻을 수 있다.

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

Weighted Complete-Case Analysis

Inference from Weighting Data

- 장점: 계산하기 쉽다.
- 단점:
 - **standard error** 계산에 어려움이 있다: 간단한 상황이라면 계산 가능하지만, 복잡 하다면 Taylor series expansions, balanced repeated replication, Jackknifing과 같은 다른 방법들을 적용해야 한다.
 - asymptotic standard error를 계산하기 위한 통계 패키지가 있지만 weights를 고정된 값으로 처리해 sampling variability를 무시하는데, 이는 문제가 될 수 있다.
 - 유효한 asymptotic inference를 하기 위해서는 잭나이핑이나 부트스트래핑같은 sample reuse method를 써야 하는데 계산량이 커진다.

Summary of Weighting Methods

⇒ (1) covariate information이 제한적이고 (2) sample size가 충분히 커서 sampling variance보다 bias가 더 큰 문제일 때 사용하는 것이 바람직하다.

Available-Case Analysis

앞서 다룬 CC analysis는 결측이 하나라도 있는 자료를 모두 분석에서 제외해야하기 때문에 분석에 가용할 수 있는 데이터의 낭비가 크다는 한계가 있다. 만약 특정 변수 하나에만 관심이 있어 univariate analyses를 수행해야 한다면, 관심이 있는 변수에 결측이 있는 자료만 제외하는 방법을 고려할 수 있다. complete-case가 아니더라도 분석에 사용할 수 있는 자료는 모두 사용한다는 뜻에서 이를 **available-case analysis**라 한다. 변수마다 분석에 사용되는 자료가 상이하다는 문제가 있지만, MCAR 하에서 특정 변수 하나의 통계량이 필요한 경우 AC analysis를 통해 구할 수 있다.

covariation을 구해야 하는 경우, 변수 두 개가 필요하므로 AC analysis를 확장한 **pairwise AC methods**를 적용한다. pairwise AC methods를 통해 변수 Y_j 와 Y_k 의 pairwise covariance $s_{jk}^{(jk)}$ 를 다음과 같이 계산할 수 있다.

$$s_{jk}^{(jk)} = \sum_{i \in I_{jk}} (y_{ij} - \bar{y}_j^{(jk)})(y_{ik} - \bar{y}_k^{(jk)}) / (n^{(jk)} - 1)$$

이 때 I_{jk} 는 변수 Y_j 와 Y_k 에서 모두 관측되는 자료의 집합, $n^{(jk)}$ 는 그 수이다.

이를 바탕으로 correlation의 추정치를 계산하면 다음과 같다.

$$r_{jk}^* = s_{jk}^{(jk)} / \sqrt{s_{jj}^{(j)} s_{kk}^{(k)}}$$

Available-Case Analysis

$$r_{jk}^* = s_{jk}^{(jk)} / \sqrt{s_{jj}^{(j)} s_{kk}^{(k)}}$$

그러나 위 추정치는 경우에 따라 (-1, 1)의 범위를 벗어난 값을 가질 수 있는데, 이는 각 변수의 sample variance인 $s_{jj}^{(j)}$ 와 $s_{kk}^{(k)}$ 가 서로 다른 자료를 바탕으로 계산되었기 때문이다. 이를 보완하기 위해 두 변수에서 모두 존재하는 자료를 바탕으로 sample variance를 계산해 다음과 같은 correlation의 추정치를 구할 수 있다.

$$r_{jk}^{(jk)} = s_{jk}^{(jk)} / \sqrt{s_{jj}^{(jk)} s_{kk}^{(jk)}}$$

추정치가 (-1, 1)의 범위 내에 존재한다는 점에서 두 번째 추정치인 $r_{jk}^{(jk)}$ 이 조금 더 좋지만, 두 추정치 모두 4개 이상의 변수를 사용하는 경우 positive definite하지 않은 correlation matrix가 계산될 수 있다는 문제가 있다.

정리하면, (1) 데이터가 MCAR을 따르고 (2) 변수 간 correlation이 크지 않은(modest) 경우 AC가 CC보다 분석에 사용할 수 있는 자료가 더 많으므로 더 효과적인 분석 방법이라고 볼 수 있다. 만약 변수 간 correlation이 크다면 AC보다 CC가 더 낫다. 그러나 일반적인 상황에서 두 방법 모두 충분치 않다.

감사합니다