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

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

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

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

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

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

$$M : n \times k \ missingness \ indicator \ matrix$$

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

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

1.2 Notation

(example)

세 명의 학생들이 시험을 봤고,

학생들이 작성한 답안에 따르면 왼쪽 표와 같이 성적이 부여되었어야 한다.

그런데 선생님이 OMR을 수거하던 중 1장을 잃어버렸고,

철수의 점수는 알 수 없는 상태가 되어버렸다. (Missing)

	점수
철수	70
영희	92
민수	85

	점수
철수	?
영희	92
민수	85

$$Y = \begin{bmatrix} \text{철수} & 70 \\ \text{영희} & 92 \\ \text{민수} & 85 \end{bmatrix} \qquad M = \begin{bmatrix} 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \qquad R = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}$$

1.3 Missing Data Patterns

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_3까지만 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, 다시말해 '몸무게 변화' column에서만이며,

이는 Missing Data Pattern 중 Univariate Nonresponse에 해당한다.

∴ ① ⇔ 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)

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

$$or$$

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

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

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 집중하냐의 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.

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

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

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.

2.3.1 Ignore

Missing Data를 처리하는 첫 번째 방법은,

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

앞선 예시에서 몸무게를 측정하기 전에 실종되어버린 피실험자 8의 데이터를 날려버리고,

남은 7명의 데이터로만 분석하는 것으로 이해할 수 있겠다.

	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.

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

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

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

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 (⇔ MAR)
합리적이라고 할 수 있다.

▼ 부연설명

만약 피실험자 8이 운동과 식단조절을 병행하면서 몸무게가 9kg 빠져버렸고,
이로 인한 급격한 건강 악화로 쓰러져서 몸무게를 재러 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 X에서 결측된 y값에 대응하는 row를 삭제한 후에도
(X가 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

$$rank(X_*^T X_*) = rank(X_*)$$

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_3값을 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_2값으로부터 어떻게 Y_3값을 추정해 낼 것인가?

당연하게도 Y_1, Y_2, Y_3가 모두 관측된 데이터들로부터

Y_1, Y_2 와 Y_3의 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

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

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

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.

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_k를 $y_{k(k)}$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

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

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

2.4 Filling in Least Squares Estimates

이제 전체 데이터 n개 중 missing data가 m개인 경우로 확장해 보자. (observed data는 n-m=r개)

LSE로 보완된 Complete Data를 가지고 추정한 β는 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 β를 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 n개가 아닌 r개의 data를 활용했으므로 σ^2의 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,

$\widehat{Cov}(\hat{\beta}) = \hat{\sigma}^2 (X^T X)^{-1}$ 의 모든 원소들의 절댓값이 작아진다는 것이다.

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

결측된 m개의 data를 LS로 채워넣고 회귀분석을 한다는 것은

마치 처음부터 Complete Data였던 것처럼 취급하겠다는 것,

즉 SSE를 r-p가 아닌 n-p로 나누겠다는 것이고, ($SSE = SSE_*$)

이때 $\hat{\sigma}^2$은 s^2으로,

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

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

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

위와 같이 y_{tb}에서 결측이 발생했다고 하자.

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

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

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

라고 정의할 때,

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

위 식의 값은, LS를 이용해 추정한 y_{tb}의 값과 일치한다.

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_2는 주행 속도를 나타내는 dummy variable,

X_3, X_4, X_5는 제조사를 나타내는 dummy variable이다.

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

$$\hat{\beta}_* = (X_*^T X_*)^{-1} X_*^T y_* = \begin{bmatrix} 607/240 \\ 23/40 \\ 91/80 \\ 2/3 \\ 17/60 \\ 13/30 \end{bmatrix}$$

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

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

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

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

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

훨씬 간단한 계산으로 같은 값을 도출해내는 것을 확인할 수 있다.

더불어 분자를 살펴보면 T(B-1) + B(T-1) - (TB-1) = TB - T - B +1개의 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 값들을 채워 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 n개 있다면 3^n개의 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 것이므로,
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 찍음, H&W: 처음 한번 찍고 계속 update해 나감)

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

여기서는 처음 한 번만 찍고, 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.

1. 결측치를 임의의 값(initial value)으로 채워 넣는다.
2. 결측을 표현하는 dummy variable을 모델에 포함하여 회귀분석을 진행한다.
3. 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 채워넣자.

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

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_2는 주행 속도를 X_3, X_4, X_5는 제조사를 나타내는 dummy variable이다.

이때, X_6을 결측을 표현하는 dummy variable로 지정하자.

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

$$\begin{aligned} \hat{\beta}_* &= (X_*^T X_*)^{-1} X_*^T y_* \\ &= \begin{bmatrix} 607/240 \\ 23/40 \\ 91/80 \\ 2/3 \\ 17/60 \\ 13/30 \\ 1/20 \end{bmatrix} \end{aligned}$$

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 independent 를 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, $0\ 0\ \cdots\ 0$ 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	?

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

결측치 3개의 initial value는 모두 4.0으로 하자

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

2.4 Filling in Least Squares Estimates

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

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

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

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

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

2.5.2 Notations for ANCOVA

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

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

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

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

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

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

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

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

이를 바탕으로 residual sum of squares를 최소화시키는 β와 γ는 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_i가 observed value이면 $z_i\gamma = 0$이고, y_i가 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_i를 다음과 같이 설정하자.

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

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 $SS(\hat{\beta}_*, \hat{\gamma})$을 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.

$$\begin{aligned} &Correct\ least\ squares\ predicted\ value\ for\ ith\ missing\ value \\ &= Initial\ guess\ for\ ith\ missing\ value\ -\ Coefficient\ of\ ith\ missing\ value\ covariate \end{aligned}$$

결국 **missing value의 올바른 추정값**은 우리가 설정한 **초깃값에서 공변량 계수를 뺀 값**으로 구할 수 있다

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

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 MSE와 s_*^2의 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.

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

그렇다면 $(X^TX)^{-1}$는 어떤 방식으로 도출하면 될까?

우리는 위의 식에서 단순 X만이 아니라, Z도 고려해야 하므로 다음과 같은 행렬을 생각할 수 있다.

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

왜 (X, Z) 를 사용할까?

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

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

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

X^TX의 역행렬은 일반적으로 다음과 같이 구할 수 있고 이를 U라고 정의하면 다음과 같다.

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

z_i의 정의에 의해,

$X^TZ = \sum_{i=1}^{m} x_i^T z_i$ 이고

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

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

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

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

결국 최종적으로 $U = (\sum_{i=m+1}^{n} x_i^T x_i)^{-1}$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 앞서 보인 것 처럼 $\hat{\gamma}_i = \tilde{y}_i - x_i\hat{\beta}_*$에서 $\tilde{y}_i = x_i\hat{\beta}_* + \hat{\gamma}_i$ 라는 관계이다.

앞서 언급했던 이 방법은 incomplete data의 ANOVA와 complete data의 ANCOVA로 설명된다.

이번에는 missing value의 covariate와 관련해서

이를 complete data의 ANOVA만으로 도출하는 과정을 보이겠다.

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

$$\hat{\gamma} = B^{-1}\rho, \tag{2.16}$$

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 ρ를 우선 계산하고, B의 역행렬도 구해야 한다.

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 γ에 대해 편미분 하게 되면

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

$Z^TY - Z^TX\beta - Z^TZ\gamma = 0$이고,

$Z^TY = Z^TX\beta + Z^TZ\gamma$ 여기서 X와 Z가 adjusted 되었다고 생각하자. 즉, Z에 대한 효과를 분석하고 싶기 때문에, Z와 X가 uncorrelated되었다고 생각하자는 것이다. 앞서 언급한 X의 영향을 제거한다는 의미가 바로 이 것이다.

그러면 $Z^TY = Z^TZ\gamma$ 이므로 최종적으로

$\hat{\gamma} = (Z^TZ)^{-1}Z^TY$이 도출된다.

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

Proof and Difference With Notations In Textbook

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

우선 B를 계산하기 위해 Z의 각 열(1~m) Z_j에 대해 X를 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

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

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

$\hat{z_{ik}}$는 X의 선형결합으로 표현 가능하므로 식은 다음과 같이 축소된다.

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

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

참고로 이 때 Y는 initial guess를 추가한 상태이다.

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

B를 구하는 과정에서 Z와 잔차 ϵ_Z를 혼용하는 것은 이론적으로 약간의 차이를 초래할 수 있지만, 실제 분석에서는 이 차이가 매우 작아 무시할 수 있는 수준일 수 있다. Z가 0과 1로 구성되어 있으므로, 잔차 ϵ_Z도 유사한 형태를 가질 가능성이 높다. 따라서 Z를 사용한 분석과 ϵ_Z를 사용한 분석은 실질적으로 큰 차이가 나지 않을 수 있다.

따라서 Z를 사용하여 모델을 적합하고 분석하는 것은 실질적으로 유효하며, 잔차 ϵ_Z를 사용한 분석과 거의 동일한 결과를 제공할 수 있다. 이는 결측값이 있는 데이터에서도 정확한 추정치를 얻는 데 충분히 신뢰할 수 있는 방법이다.

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

잔차를 사용한 경우

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

그냥 Z를 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 m만큼 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 C는 $p \times 1$ 크기의 constant vector이고, 결국 λ가 의미하는 건 β의 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.

$$H_0 : C^T\beta = 0 \;\; vs \; H_1 : not \; H_0$$
$$where \; C \; is \; a \; p \times 1 \; vector.$$

$$\hat{\beta} = (X^TX)^{-1}X^TY$$

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

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

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

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

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

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

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

정리하면

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

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

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

$SE_* = s_*\sqrt{C^T(X^TX)^{-1}C}$ 이고

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

H라는 새로운 vector를 생각해보자. H는 complete data가 주어졌을 때 λ의 estimate로, m개의 misssing value의 covariates(Z)를 dependent varaialbe는 regression을 통해 나타낼 수 있다.

위에서 사용한 model을 가지고 그대로 이를 유도해보면,

$Z = X\beta + \epsilon$

$\hat{\beta} = (X^TX)^{-1}X^TZ$이므로

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

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

여기서 주목할 부분이 있는데, 방금 우리가 구한 H와 위에서 구했던 B가 동시에 계산이 가능하다는 것이다. H와 B모두 complete data의 ANOVA와 missing data의 covariate을 통해 구해진다.
즉,
H의 i번째 component와 B의 ith row는 같은 방식으로 계산되었다는 뜻이다.

그러면 $C^TUC = C^T(X^TX)^{-1}C + H^TB^{-1}H$ 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

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

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

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

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

C를 $p \times w$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 w개의 linear combination에 대해 orthonormal을 가정하자. 즉,
$C^T(X^TX)^{-1}X = I_w$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 $SS_* = \hat{\lambda_*}^T(C^TUC)^{-1}\hat{\lambda_*}$가 될 것이다.

2.7과 마찬가지로, 적당한 H라는 matrix를 설정하면

$SS_* = \hat{\lambda}^T(I + H^TB^{-1}H)^{-1}\hat{\lambda}$ 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 Anlalysis

Complete-Case Analysis

Complete-Case Analysis (Listwise Deletion Analysis)

모든 변수들이 관측된 개체들만 이용하여 분석을 진행한다.
단 하나의 변수에서 결측치가 있어도 그 개체는 분석에서 제외된다.
결측치를 가지는 case가 작은 경우에 일반적으로 사용한다.

- 장점: (1) 간편성
  (2) 일변량 통계량들의 비교 가능
  (3) MCAR 가정 하에서 모수 추정치에 bias가 거의 발생하지 않음
- 단점: (1) 많은 표본 수의 감소 >> 정보의 손실 >> 검정력 약화
  (2) MCAR이 아닌 경우 bias 발생

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 Information: loss of precision & bias

- Loss of precision = 분산 증가
  - $\hat{\theta}_{CC}$를 complete units로부터 추정한 θ의 추정값, $\hat{\theta}_{EFF}$를 사용 가능한 데이터를 바탕으로 추정한 θ의 efficient estimate이라고 할 때,

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

- Bias
- 자료가 MCAR인 것을 가정하고 진행하기 때문에 이 가정이 흔들리는 정도(데이터의 누락이 random이 아닐 때)에 따라 bias가 발생한다.

- 제거한 결측이 있는 자료들은 MCAR 가정이 합리적인 가정이었는지 검정하기 위해 사용된다.
  - Complete units의 특정한 변수 Y_j와 Incomplete units의 Y_j의 분포를 비교한다.
  - 차이가 클수록 MCAR 가정이 합리적이지 않고(invalid) CC analysis로부터터 추정한 추정값이 불편추정량이 아닐 가능성이 높음을 의미한다.

Weighted Complete-Case Analysis

CC analysis의 경우 MCAR인 경우를 제외하고는 bias가 발생할 수 있다. CC analysis에 가중치(weight)를 부여하여 bias를 조정하는 방법이 **Weighted Complete-Case Analysis**이다.

여기서 가중치를 설정하는데 사용되는 기본적인 아이디어가 표본 추출 과정(survey sampling)과 유사하기 때문에, 이에 대해 먼저 알아보자.

[Notations]

- Population Size: N, Sample Size: n
- Number of variables (items): K
- Data: $Y = \left(y_{ij}\right) \; where \; i = 1, \dots, N \; and \; j = 1, \dots, K$
- Design Information (about sampling or missingness): Z
- Sample Indicator: $I = (I_1, \dots, I_N)' \; for \; unit \; i$
- $I = \begin{cases} 1, & unit \; i \; included \; in \; the \; sample \\ 0, & unit \; i \; not \; included \; in \; the \; sample \end{cases}$

Weighted Complete-Case Analysis

표본 추출 과정에서 unit들은 **확률적 표본추출 (Probability Sampling)**을 따르는데, 이는 다음과 같은 두 가지 요건을 가진다.

1) Unconfounded
$f(I \mid Y, Z) = f(I \mid Z)$

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.

$$\pi_i = E\left(I_i \mid Y, Z\right) = \Pr\left(I_i = 1 \mid Y, Z\right)$$
$$\pi_i = P\left(I_i = 1 \mid Z\right) > 0, \quad \text{for all } i$$

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

- $\mathrm{Z} = j, \quad j = 1, \ldots, J$
- N_j population units
- Stratum j에서 n_j units의 SRS 진행
- Stratified random sampling 아래, I (sampling indicator)의 분포는 다음과 같다.

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

Weighted Complete-Case Analysis

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

Stratified Random Sampling이기 때문에 다음과 같이 구할 수 있다.

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

$$\mathrm{Var}\left(\bar{y}_{st}\right) = \frac{1}{N^2}\sum_{j=1}^{J} N_j^2 \left(\frac{1}{n_j} - \frac{1}{N_j}\right) S_j^2 \quad , \quad v\left(Y_{\mathrm{inc}}\right) = \frac{1}{N^2}\sum_{j=1}^{J} N_j^2 \left(\frac{1}{n_j} - \frac{1}{N_j}\right) s_j^2$$

모평균을 추정할 때 weighting methods를 사용하면 어떻게 표현할 수 있을까?

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

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

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

Weighted Complete-Case Analysis

Weighting with Nonresponses

- 무응답이 존재하는 경우에는 모집단에서 표본으로 추출될 확률과 함께 추출된 표본에서의 **응답확률** (response probability, $\boldsymbol{\phi_i}$) 도 함께 고려

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

- survey sampling에서 표본으로 추출될 확률의 역수를 가중치로 사용했던 것처럼, 여기서도 응답자들에게 응답확률의 역수를 가중치로 부여함
- 무응답으로 인하여 조사 목표 표본 크기와 실제 조사 성공 표본 크기의 차이를 보정

  Ex. 100명의 남자, 100명의 여자로 이루어진 표본에서 남자 중 80명, 여자 중 75명이 응답했다면,

  남자 응답자의 가중치: 1/응답확률 = 100/80 = 1.25

  여자 응답자의 가중치: 1/응답확률 = 100/75 = 1.33

- 이렇게 weighting을 해서 모평균에 대한 추정치를 구하면 다음과 같다. 여기서 r은 respondents 수를 의미한다.

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

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

Weighted Complete-Case Analysis

실제 상황에서는 ϕ_i를 모르는 경우가 대부분이다. 따라서 우리는 ϕ_i를 추정하여 사용하게 된다.

ϕ_i를 추정하여 사용하는 방법에는 두 가지가 있다: 1) Weighting Class Adjustment 2) Propensity Weighting

1. Weighting Class Adjustment

응답, 무응답을 포함한 표본을 J개의 Weighting Classes로 나누는 방법

- Sample size: n_j
- Number of observed units (respondents): r_j $\quad (\mathrm{r} = \sum_{j=1}^{J} r_j)$
- Estimator of the response probability (ϕ_i) : $\widehat{\phi_i} = r_j/n_j$
- Weighting class variable: C
- Weighting class j의 responding unit은 다음과 같은 가중치를 가지게 된다:

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

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

이를 이용하여 모평균을 추정하면 다음과 같다.

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

만약 equal probability design처럼 sampling weight(π_i^{-1})이 동일한 경우에는 위 식이 다음과 같이 간략하게 표현될 수 있다.

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

Weighted Complete-Case Analysis

1. Weighting Class Adjustment

이 추정값은 MAR 가정 (quasirandomization) 하에 unbiased하다.

- **Quasirandomization**: weighting class j내의 데이터는 MCAR을 따른다.

  >> class j내의 응답자들은 sample의 random subsample이 된다.

그러면 Weighting class는 어떻게 설정하는 것이 좋을까?

(1) Quasirandomization 가정을 만족하도록, (2) 그 가정 하에 추정치의 MSE (mean squared error)가 작아지도록 설정한다.

Respondent sample size가 너무 작아지지 않도록 주의한다. (>> 큰 가중치)

Bias와 분산에 주는 효과도 고려한다.

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

- **C와 Y가 약하게 관련되어 있을 때 :**

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: 응답자, 무응답자 모두에 대해 관찰된 변수
- 데이터는 MAR이라고 가정, ϕ는 모르는 모수

  $$P(M|X,Y,\phi) = P(M|X,\phi)$$
- C가 X가 되면 Quasirandomization을 만족한다.

Response Propensity Stratification

Unit i에 대한 nonresponse propensity: $\rho(x_i, \phi) = \Pr(m_i = 1 \mid \phi)$

모든 x값에 대하여 response propensity는 0보다고 크다고 가정하면,

$$\begin{aligned}
\Pr(m_i = 0 \mid y_i, \rho(x_i, \phi), \phi) &= E(\Pr(m_i = 0 \mid y_i, x_i, \phi) \mid y_i, \rho(x_i, \phi), \phi) \\
&= E(\Pr(m_i = 0 \mid x_i, \phi) \mid y_i, \rho(x_i, \phi), \phi), \quad \text{by Eq. } (3.9) \\
&= E(\rho(x_i, \phi) \mid y_i, \rho(x_i, \phi), \phi), \quad \text{by definition of } \rho(x_i, \phi) \\
&= \rho(x_i, \phi), \quad \text{for all } x_i.
\end{aligned}$$

$$Pr(M|\rho(X,\phi),Y,\phi) = Pr(M|\rho(X,\phi),\phi)$$

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

Weighted Complete-Case Analysis

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

(1) 응답자, 무응답자 데이터를 기반으로 X에 대한 M의 binary regression (logistic, probit, robit)으로부터 $\hat{\phi}$ 를 추정한다.

(2) $\rho(X, \hat{\phi})$ 을 5개 혹은 10개 값으로 나누는 grouped variable C를 설정한다.

(3) 같은 adjustment class j내에서 응답자와 무응답자가 같은 값의 grouped propensity score을 갖도록 한다.

이 과정의 또 다른 방식은 propensity score의 역수를 가중치로 직접 사용하는 것이다.

가중치 = $\rho(X, \hat{\phi})^{-1}$

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

Weighted Complete-Case Analysis

Post Stratification and Ranking to Known Margins

앞서 weighting class estimate에서는 weighting class j에 속하는 모집단의 비율 N_j/N을 표본의 비율 n_j/n으로 추정하여 사용하였다.

이 비율을 외부의 정보 등으로부터 알고 있다면, 평균을 어떻게 추정할 수 있을지 알아보자.

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.

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

$$\mathrm{Var}(\bar{y}_{\mathrm{ps}}) = \frac{1}{N^2}\sum_{j=1}^{J} N_j^2 \left(1 - \frac{r_j}{N_j}\right)\frac{S_{j\mathrm{R}}^2}{r_j}$$

Weighted Complete-Case Analysis

2. Raking Ratio Estimation

- Raking: 다수의 주요 변수의 알려진 주변 분포를 이용하여 한 변수씩 표본의 특성을 모집단의 특성에 일치시키는 과정을 반복하여 모집단 분포를 추정하는 방법
- 표본의 특성이 모집단의 특성에 일치되도록 레이킹 가중값을 산출한다.
- Post-stratification의 한 방법으로, 한 번에 한 변수씩 고려해서 모집단 특성에 일치시키므로 레이킹할 변수가 다소 많더라도 안정적인 가중값을 계산할 수 있다.

예시를 통해 과정을 살펴보자.

우리가 모집단에서 성과 연령의 각 구성비율을 알고 있고, 결합구성비율은 모른다고 하자. 각 구성비율에 맞춰 우리가 목표로 하는 크기로 주변합계를 구성하면 아래와 같다.

		모집단 구성비율 (%)
성	남자	50%
	여자	50%
연령	20대	18%
	30대	20%
	40대	22%
	50대	18%
	60세 이상	22%

	20대	30대	40대	50대	60세 이상	합계
남자						250
여자						250
합계	90	100	110	90	110	500

Weighted Complete-Case Analysis

(1) 초기값 구성

조사 표본으로부터 교차표를 만든다.

	20대	30대	40대	50대	60세 이상	합계
남자	20	70	80	70	60	300
여자	30	50	50	30	40	200
합계	50	120	130	100	100	500

(2) 행의 레이킹 비 조정

	20대	30대	40대	50대	60세 이상	합계
남자	16.7	58.3	66.7	58.3	50.0	250.0
여자	37.5	62.5	62.5	37.5	50.0	250.0
합계	54.2	120.8	129.2	95.8	100.0	500.0

(3) 열의 레이킹 비 조정

	20대	30대	40대	50대	60세 이상	합계
남자	27.7	48.3	56.8	54.8	55.0	242.5
여자	62.3	51.7	53.2	35.2	55.0	257.5
합계	90.0	100.0	110.0	90.0	110.0	500.0

(4) 행 합계와 열 합계가 목표 크기에 같아질 때까지 (2), (3) 반복

(5) 레이킹 가중값 산출

셀	성별	연령대	**표본에서 셀의 크기** (A)	**Raking으로 추정한 셀의 크기** (B)	**레이킹 가중값** (B/A)
1	남자	20대	20	28.9	w1= 1.445
2	남자	30대	70	49.8	w2= 0.712
3	남자	40대	80	58.5	w3= 0.731
4	남자	50대	70	56.1	w4= 0.801
5	남자	60세 이상	60	56.7	w5= 0.945
6	여자	20대	30	61.1	w6= 2.037
7	여자	30대	50	50.2	w7= 1.004
8	여자	40대	50	51.5	w8= 1.031
9	여자	50대	30	33.9	w9= 1.130
10	여자	60세 이상	40	53.3	w10= 1.332
합계			500	500	

Weighted Complete-Case Analysis

Summary of Weighted CC Analysis

- Weighted CC estimate의 경우, 계산이 간편하다는 장점이 있으나, 적절한 표준 오차를 계산하기 어렵다.

- Weighting methods는 CC analysis의 bias를 조정해주는 비교적 간단한 방법이다.
- 각 unit에 대해 측정된 모든 변수에 대해 동일한 가중치를 산출하기 때문에 과정이 간편하다.
  (1) covariate information이 제한적이고
  (2) 표본의 크기가 충분히 커서 sampling variance보다 bias가 더 큰 문제일 때 사용하는 것이 바람직하다.

Available-Case Analysis

Available-Case Analysis (Pair-wise Deletion)

각각의 분석 단계에서 사용 가능한 자료를 이용한다.

Univariate analyses를 수행할 때 유용하다.

- 장점: (1) 일반적으로 CC analysis보다 표본수가 많음
  (2) MCAR 가정 하에서 모수 추정치에 bias가 거의 발생하지 않음
- 단점: (1) 표본의 수가 분석에서 사용하는 변수마다 달라짐
  (2) 모수 추정시 수학적 문제 발생 가능
  (3) 실용적이지 못함

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%
3	Fast+	167	
4	Fast+		80%
5	Lite	76	70%
6	Fast+	155	10%
7			95%

Available-Case Analysis

Covariation을 계산하기 위해서는 pairwise AC method를 적용한다.

관측된 y_{ij}와 y_{ik}에 대하여 Y_j와 Y_k의 pairwise 공분산은 다음과 같다.

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

상관계수는 다음과 같이 추정된다.

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

각 변수의 표본 분산이 서로 다른 바탕으로 계산되었기 때문에 위 추정치는 경우에 따라 (-1, 1)의 범위를 벗어난 값을 갖기도 한다. 이를 보완하기 위해 두 변수에서 모두 존재하는 자료를 바탕으로 표본 분산을 계산해 다음과 같이 구할 수 있다.

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

Summary

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 않아 적용 가능한 상황이 제한적이다. CC나 AC를 적용하기 좋은 대표적인 상황은 설문조사에서 unit-nonresponse를 처리하는 경우이다.

Assignment

Show $SE_* = \sqrt{\frac{n-p}{r-p}(SE^2 + s^2 H^T B^{-1} H)}$ and $SS_* = \frac{SS}{(1+(SS/\hat{\lambda}^2)H^T B^{-1} H)}$

where

C is the vector of p constants, B is the $m \times m$ cross-products matrix for the residuals of the m missing value covariates after adjusting the matrix X, and H be the $m \times 1$ vector of complete-data ANOVA estiamtes of λ taking each of the m missing data covariates as the dependent variable rather than Y. All other notations follow the same meaning with textbook(Ch 2) and session. You can use the values of SS, SS_*, SE, SE_* or other values computed in the session without specific justification of proof.

Additionally, describe the roles of

B and H, and provide your own opinion about the implications of the derived values.

2.7에서 마지막에 도출했던 값들입니다

모든 notation은 textbook과 세션에서 진행한 것과 같고, 세션에서 다룬 값들은 따로 증명하실 필요 없습니다.

해당 값들에 대한 증명과 함께 이들의 역할이 어떤 것인지, B와 H라는 행렬은 어떤 의미를 갖는지 한 번 씩 자유롭게 생각해보면 좋을 것 같습니다!

감사합니다