
Statistical Analysis with Missing Data

Single Imputation Methods

Accounting for Uncertainty from Missing Data

ESC 2024 Summer Special Session 3주차
최운형 민예빈 이준찬 옥승환



Contents

1. Single Imputation Methods

1.1 Introduction

1.2 Imputing Means from a Predictive Distribution

1.3 Imputing Draws from a Predictive Distribution

2. Accounting for Uncertainty

2.1 Background/Introduction

2.2 Valid SE Methods

2.3 Resampling Methods

2.4 Multiple Imputation

1

Single Imputation Methods

1.1 Introduction

1.2 Imputing Means from a Predictive Distribution

1.3 Imputing Draws from a Predictive Distribution

1.1 Introduction

[용어정리]

Imputation: 결측치 대체

Filled-in data : Imputation method을 이용하여 결측치들을 메운 data-set

Respondent: 모든 변수들의 값이 관찰되는 unit

Non-respondent: missing data가 존재하는 unit

1.1 Introduction

Single Imputation Method V.S. Multiple Imputation Method

- Single Imputation Method(4장) : 결측값을 단일값(ex. mean, draw)로 대체. Imputation uncertainty 고려 X
결측대체값이 관측값보다 더 높은 변동성을 가져야 하는 것 아니냐? 문제 제기됨
- Multiple Imputation Method(5장) : 여러 번 imputation을 수행. 각 대체본(imputed dataset)을 생성하고 이를 종합하여 imputation uncertainty 반영

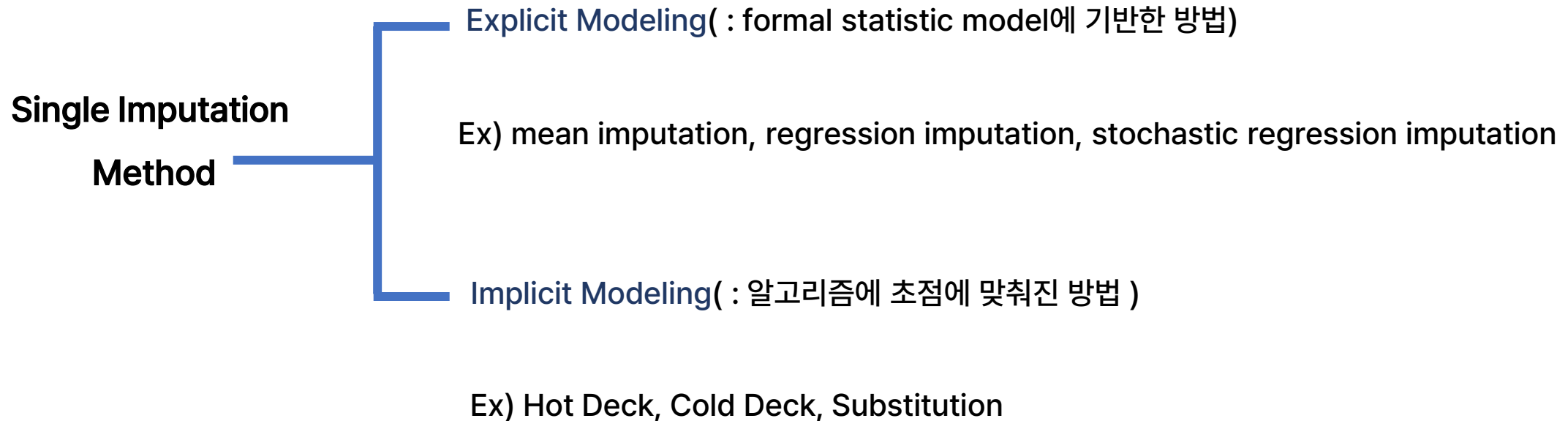
***imputation uncertainty?

imputation(결측치 대체)과정에서 발생하는 불확실성. 결측치를 대체할 때 사용된 값이 실제데이터와 얼마나 가까운지에 대한 확신이 없기 때문에 자연적으로 발생하는 불확실성.

1.1 Introduction

Single Imputation Method

Drawing predictive distribution (about missing value)



1.1 Introduction

Explicit Modeling

a) Mean imputation:

결측치가 존재하는 변수에서 결측되지 않은 나머지 값들의 평균을 내어 결측치(들)을 대체하는 방법

>> 뒤에서 Unconditional Mean Imputation(지양) VS Conditional Mean Imputation 논의

b) Regression imputation:

모든 변수들 값이 관찰되는 unit(=respondent)들을 이용하여 regression model을 만들고, 여기에 missing data가 존재하는 unit(=non-respondent)에서의 observed data를 regression model을 넣어 도출된 값 >> impute!

1.1 Introduction

c) Stochastic regression imputation : 앞에서 언급된 Regression imputation의 확장

Regression imputation에 의해 예측된 값 + 랜덤하게 선택된 residual(잔차) >> imputation!

Regression imputation에서 결측대체값에 변동이 전혀 없는 fitted value 를 넣다보니
계수추정치에 대한 신뢰도가 과대평가되는 경향이 있다. 이를 방지하기 위해 residual을 추가한 것

(참고) normal linear regression models 하에서 residual은 다음과 같이 된다


$e_i \sim N(0, (1 - h_{ii}))$ when $\varepsilon_i \sim N(0, \sigma^2)$ * h_{ii} : $H = X (X^T X)^{-1} X^T$ 의 i번째 diagonal element (X: Design Matrix)

1.1 Introduction

Implicit Modeling

a) Hot Deck Imputation:

연구중인 현재 데이터셋 내에서의 표본을 바탕으로 했을 때, 관측된 다른 변수에서 비슷한 값을 갖는 unit들중에서 하나를 랜덤 샘플링(random sampling) 하여 그 값을 복사해오는 방법



Row ID	Location	Income
1	A	850
2	A	800
3	B	1352
4	C	1450
5	A	NA
6	B	750

Row ID	Location	Income
1	A	850
2	A	800
3	B	1352
4	C	1450
5	A	800
6	B	750

1.1 Introduction

b) Cold deck imputation

Hot deck imputation과 유사하게, 다른 변수에서 비슷한 값을 갖는 데이터 중에서 하나를 골라 그 값으로 결측치를 대체하는 방식이다.

다만 cold deck imputation에서는 비슷한 양상의 데이터 중에서 하나를 랜덤 샘플링하는 것이 아니라 특정규칙 하에서 하나의 데이터를 선정하는 것이다.

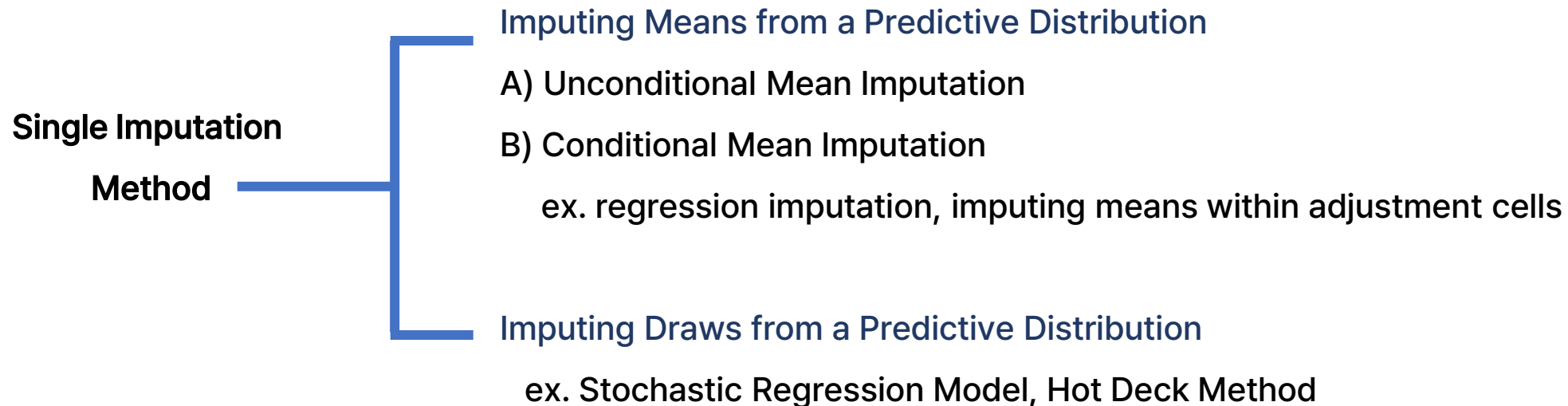
hot deck imputation 과정에서 부여되는 random variation이 제거된다.

c) **Substitution:** 결측치가 존재하는 unit에 sample로 아직 추출되지 않는 alternative unit을 넣는 것

1.1 Introduction

Single Imputation Method

-Drawing predictive distribution (about missing value)



1.2 Imputing Means from a Predictive Distribution

Imputing Means from a Predictive Distribution

Unconditional Mean Imputation

관측된 값들과 결측치들이 모인 filled-in data의 sample variance는 $s_{jj}^{(j)}(n^{(j)} - 1)/(n - 1)$

, 여기서 $s_{jj}^{(j)}$ 는 $n^{(j)}$ available units으로부터 추정된 분산

filled-in data의 두 개의 변수 Y_j 와 Y_k 의 sample covariance는 $\tilde{s}_{jk}^{(jk)}(n^{(jk)} - 1)/(n - 1)$

, 여기서 $\tilde{s}_{jk}^{(jk)} = \sum_{i \in (kl)} (y_{ij} - \bar{y}_j^{(j)})(y_{ik} - \bar{y}_k^{(k)})/(n^{(jk)} - 1)$

MCAR하에서, $s_{jj}^{(j)}$ 와 $\tilde{s}_{jk}^{(jk)}$ 는 각각 모분산과 모공분산의 일치추정량이 되기 때문에

위에서 언급된 filled-in data로부터 도출된 sample variance와 sample covariance는 각각 모분산과 모공분산을 과소평가하게 됨 >> 해당 방법은 지양!

1.2 Imputing Means from a Predictive Distribution

Imputing Means from a Predictive Distribution

Conditional Mean Imputation

a) imputing means within adjustment cells (condition: adjustment class내에서 따짐으로써)

1단계) non-respondent 와 respondent들을 J개의 adjustment-class로 분류하라

2단계) 각각의 adjustment-class 내에서 respondent의 평균값을 non-respondents에 impute!

>> 해당 imputation에 의해서 완성된 filled-data을 이용하여 평균의 추정치는 다음과 같다.

$$\frac{1}{n} \sum_{j=1}^J \left(\sum_{i=1}^{r_j} y_{ij} + \sum_{i=r_j+1}^{n_j} \overline{y_{jR}} \right) = \frac{1}{n} \sum_{j=1}^J n_j \overline{y_{jR}} = \overline{y_{wc}}$$

1.2 Imputing Means from a Predictive Distribution

Imputing Means from a Predictive Distribution

Conditional Mean Imputation

b) regression imputation

univariate nonresponse case에 대해서 따져보자

k개의 변수가 있고 여기서 Y_1, \dots, Y_{k-1} 은 fully-observed이고 Y_k 는 observed for r units and missing for the last n-r units 이라 하자.

이때 (모든 변수의 값이 관찰된 unit인) respondent r개만을 골라서 regression을 한다.

그리고 (결측치가 있는 unit인) non-respondent n-r개를 추정된 regression model에 넣어서 각각 구한 값들을 imputation

1.2 Imputing Means from a Predictive Distribution

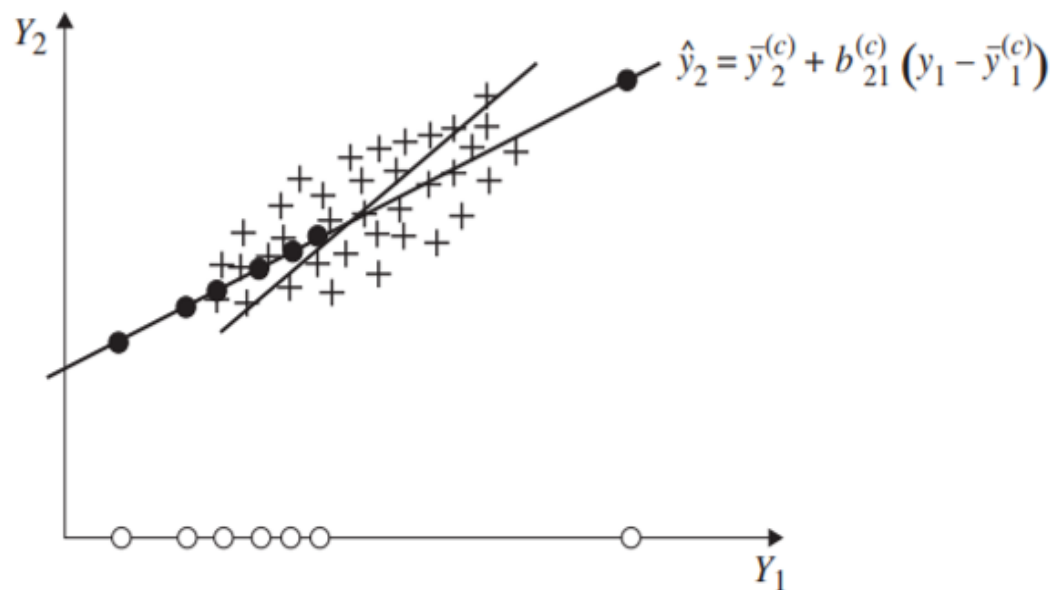


Figure 4.1 Example 4.2: regression imputation for $K = 2$ variables.

$$\widehat{y}_{ik} = \beta_{K012\dots K-1} + \sum_{i=1}^{K-1} \tilde{\beta}_{Kj12\dots K-1} y_{ij} + z_{iK}$$

해당 방법을 시각적으로 표현하여 이해해보자.

가령 $K=2$ (즉 관련된 변수가 Y_1, Y_2 2개)라고 가정하자.

그리고 Y_1 이 fully-observed된 변수, Y_2 이 결측값이 존재하는 변수라 하자.

+는 (Y_1, Y_2) 모두 관측된 unit들의 점들을 표시

이것들을 이용해 추정된 회귀선이 오른쪽에 표시되어 있다.

Y_1 축 위에 있는 하얀점들은 결측치가 존재하는 unit들을 표시

>> 까만점들의 Y_2 값들을 이용해 imputation 진행!

1.3 Imputing Draws from a Predictive Distribution

그전에 언급한 각종 mean-imputation 기법들을 이용하면...

- * 누락된 값을 평균값으로 대체함으로써 데이터 포인트들이 비슷해지고 데이터가 실제로보다 더 동질적인 것처럼 보이게 된다.

- * 평균이라는 특정된 값이 대체된다는 점에서 해당 데이터가 가질 수 있는 다양한 값의 가능성을 고려하지 않게 된다.

즉, imputation uncertainty를 반영하지 못함으로써 Sample variance가 낮게 구해진다

>> true population variance 과소평가하게 되는 문제 발생

>>>

Predictive Distribution으로부터의 mean 단일값을 impute하기보다

Predictive Distribution으로부터의 "plausible values", 즉 (random) draw를 impute하는 것이

imputation uncertainty를 고려한다는 점에서 더 나은 방법

1.3 Imputing Draws from a Predictive Distribution

Stochastic Regression Imputation

Regression Imputation에서는 단순히 mean imputation하는 것만으로
Sample variance를 너무 낮게 잡아서 실제 모분산을 과소평가하는 문제가 있어
imputation uncertainty를 고려한 새로운 방법인 **Stochastic Regression Imputation**

Regression imputation에서 예측된 값 + **랜덤추출된 잔차** > impute

$$\widehat{y}_{ik} = \beta_{K012\dots K-1} + \sum_{i=1}^{K-1} \tilde{\beta}_{Kj12\dots K-1} y_{ij} + \mathbf{Z}_{iK}$$

여기서 \mathbf{Z}_{iK} 는 (평균0과 respondent들의 값을 가지고 회귀했을 때의 잔차분산을 분산으로 갖는)
Random normal deviate이다

* Random normal deviate : 정규분포에서 무작위로 얻어진 값

1.3 Imputing Draws from a Predictive Distribution

Stochastic Regression Imputation

Bivariate monotone MCAR 데이터라면 Stochastic Regression Imputation을 통해 얻은 모수 추정치의 large sample bias는 다음과 같습니다.

Table 4.1 Example 4.5: bivariate normal monotone MCAR data: large sample bias of four imputation methods

Method	Parameter			
	μ_2	σ_{22}	$\beta_{21\cdot1}$	$\beta_{12\cdot2}$
Umean	0 ^a	$-\lambda\sigma_{22}$	$-\lambda\beta_{21\cdot1}$	0 ^a
Udraw	0	0	$-\lambda\beta_{21\cdot1}$	$-\lambda\beta_{12\cdot2}$
Cmean	0	$-\lambda(1-\rho^2)\sigma_{22}$	0 ^a	$\frac{\lambda(1-\rho^2)}{1-\lambda(1-\rho^2)}\beta_{12\cdot2}$
Cdraw	0	0	0	0

λ = fraction of missing data.

^aIndicates estimator is same as CC estimate.

C Draw가 Stochastic Regression Imputation에 해당

Stochastic Regression Imputation을 통해 얻은 모수 추정치만 일치추정량

1.3 Imputing Draws from a Predictive Distribution

Hot Deck Method



Hot Deck Method은 랜덤샘플링을 통해 결측대체치에 대한 변동성을 가함으로써 sampling variance가 낮게 구해져 true population variance를 과소평가하게 되는 문제 완화 가능

1.3 Imputing Draws from a Predictive Distribution

Hot Deck Method

총 N 개의 unit >> 이중 n 개의 sample 따지기 >> 이중 $r(<n)$ 개는 respondents

Hot Deck에 의해 imputation된 값들과 기존에 관찰된 값들을 모은 filled-in data를 응용하여 평균을 구하면 다음과 같다.

$$\overline{y_{HD}} = \{r\overline{y_R} + (n - r)\overline{y_{NR}}\}/n$$

여기서 $\overline{y_R}$ 은 respondent unit들의 평균이고

$$\overline{y_{NR}} = \sum_{i=1}^r \frac{H_i y_i}{n - r}$$

여기서의 H_i 는 y_i 가 결측치변수인 Y 에 대해서 대체된 횟수를 가르킨다.

1.3 Imputing Draws from a Predictive Distribution

$$\overline{y_{HD}} = \frac{\{r\overline{y_R} + (n - r)\overline{y_{NR}}\}}{n} \quad \overline{y_{NR}} = \sum_{i=1}^r \frac{H_i y_i}{n - r}$$

$\overline{y_{HD}}$ 등 Hot-Deck에 의해 구해진 filled-in data에서의 parameter추정지 속성은 $\{H_1, H_2, \dots, H_r\}$ 에 따라 달라진다.

Hot Deck의 핵심은 $\{H_1, H_2, \dots, H_r\}$ 의 분포를 찾는 것

$\{H_1, H_2, \dots, H_r\}$ 의 분포를 찾는 가장 단순한 방법 중 하나는

Respondent unit에 존재하는 값들 중에서 **probability sampling(=random selection)**을 하는 것

$\overline{y_{HD}}$ 의 평균과 분산을 나타내면 다음과 같다.

$$E(\overline{y_{HD}}) = E(E(\overline{y_{HD}} | Y_{(0)}))$$

$$Var(\overline{y_{HD}}) = Var(E(\overline{y_{HD}} | Y_{(0)})) + E(Var(\overline{y_{HD}} | Y_{(0)}))$$

1.3 Imputing Draws from a Predictive Distribution

Hot Deck Method

Hot Deck by Simple Random Sampling "with Replacement"

$\{H_1, H_2, \dots, H_r\} \sim \text{multinomial}(n-r, 1/r, \dots, 1/r)$ 이다.

즉, sample-size= $n-r$ 이고 각각의 결측치변수에서 respondent-unit에 해당하는 값들 각각에서 값이 추출될 확률이 동일한 다항분포를 따르게 된다.

결측치변수에서 관측된 값들인 $Y_{(0)}$ 이 주어졌을 때, $\{H_1, H_2, \dots, H_r\}$ 와 관련된 moment들을 정리하면 다음과 같다.

$$E(H_i | Y_{(0)}) = \frac{n-r}{r}, \quad \text{Var}(H_i | Y_{(0)}) = (n-r)(1 - \frac{1}{r})/r$$

$$\text{Cov}(H_i, H_j | Y_{(0)}) = -(n-r)/r^2$$

* $\text{Multin}(n, p_1, p_2, \dots, p_{k-1})$ 인 경우 $E(X_i) = np_i, \quad \text{Var}(X_i) = np_i(1 - p_i), \quad \text{Cov}(X_i, X_j) = -np_i p_j (i \neq j)$

1.3 Imputing Draws from a Predictive Distribution

$$* \quad \overline{y_{HD}} = \{r\overline{y_R} + (n-r)\overline{y_{NR}}\}/n \quad \overline{y_{NR}} = \sum_{i=1}^r \frac{H_i y_i}{n-r}$$

$$E(H_i | Y_{(0)}) = \frac{n-r}{r}, \quad Var(H_i | Y_{(0)}) = (n-r)(1 - \frac{1}{r})/r$$

>> 위의 내용들을 모두 종합하여, H_i 에 대한 expectation과 summation을 진행하면 다음과 같은 결과 나온다.

$$E(\overline{y_{HD}} | Y_{(0)}) = \overline{y_R}, \quad Var(\overline{y_{HD}} | Y_{(0)}) = (1 - 1/r)(1 - r/n)s_R^2/n$$

여기서 s_R^2 은 respondent들의 결측치변수들 표본분산, $\overline{y_R}$ 은 respondent들의 결측치변수들 평균

$Var(\overline{y_{HD}} | Y_{(0)}) = (1 - 1/r)(1 - r/n)s_R^2/n$ 은 hot-deck method를 이용하면서 filled-in data의 분산이다.

Hot Deck을 복원추출이 아닌 비복원추출로 진행한다면 위의 값은 감소한다.

2

Accounting for Uncertainty from Missing Data

2.0 Background

2.1 Introduction for Dealing with Sampling Uncertainty

2.2 Valid SE Methods Approach

2.3 Resampling Approach

2.3.1 Bootstrap

2.3.2 Jackknife

2.4 Multiple Imputation Approach

2.5 Conclusion

2.0 Background - Review

5.0 Part1 Review

1단원 : Missing data 문제를 정의하고, Missingness pattern과 mechanisms을 제시한다. missing pattern 과 Missing mechanism에 대해 다루고 Missingness 처리 방법에 대해 논의했다.

2단원 : 실험 계획, experiment design 상황에서의 Missingness 처리를 다룬다. 실험 상황의 특징은, X 독립 variable을 내가 컨트롤 하기 때문에 Missing이 발생할 수 없고, Y 종속변수 하나에서만 Missing이 발생하는 Univariate 한 특수한 상황에 대한 적용 방법을 다루었다.

결측을 무시하는 Complete case Analysis 와 회귀, 평균 등으로 결측 값을 채워 넣는 regression Imputation 방법을 소개되었다.

3단원 : Weighting 방법, 즉 실험 계획 상황이 아닌 더 일반적인 상황, 즉 여러 변수에서 missing 이 일어날 수 있는 상황에서 ~missing pattern 상황에서 결측 상대 메커니즘을 다루었다. 결측 또한 weighting 디자인의 일부로 보는 weighting 방식이 소개되었다.

4단원 : Imputation (대체) 으로 빈 값을 대체함으로써 편향의 위험을 감수하나 완전한 형태의 데이터로 분석할 수 있는 방법이 소개되었다.

예측 분포가 일반적인 통계 모델(정규분포)를 명시적으로 따르는 것으로 상정하는 명시적 모델링 (회귀대체, 확률적 회귀 대체 등) 과 기저 모델을 암시할 수 있는 알고리즘에 중점을 두는 암시적 모델링(-> Hot deck 대체, cold deck 대체) 방법이 소개되었다.

Missing 을 가중 및 대체를 포함해 다양한 방법으로 처리할 수 있게 되었다. 하지만 이렇게 Missing을 처리했을 때 **문제가 발생한다.**

2.1 Introduction

5.1 Reason of dealing with Imputation Uncertainty

4장에서 다룬 Single value Imputation의 문제점은 Missing의 명시적, 암시적 모델이 올바르게 적용되었다 하더라도, single value imputation 이 단일 값을 Imputation 하는 특성상 **Imputation Uncertainty** 를 반영하지 못하고 Estimates의 True sampling variance **실제 샘플링 분산을 과소평가** 한다는 것이다.

SE 표준오차의 과소추정은 신뢰구간을 좁히며 유의미한 추론에 장애가 된다. $I_{\text{norm}}(\theta) = \hat{\theta} \pm z_{1-\alpha/2} * SE$

5장에서는 filled data를 기반으로 한 imputation 이 imputation uncertainty를 전혀 반영하지 못하는 4장의 문제점을 보완한다.

Missingness 를 Imputation 할 때 Imputation Uncertainty 를 반영해 유효한 Sample variance 와 Standard Error를 추정하고, 정확한 신뢰구간을 구해 유의한 추론을 하는 것을 목표로 이에 적용할 수 있는 4가지 접근론을 소개한다.

2.1 Introduction

5.1 Four general approaches to accounting for uncertainty from missing data

❖ Explicit sampling variance formulas

- ❖ Missingness를 허용하는 sampling variance 공식 적용

Missingness를 허용하는 명시적 Sampling variance 공식 적용

❖ Imputation Methods that Provide Valid Standard Errors

- ❖ Valid SE를 구할 수 있도록 단일 데이터에서 imputation 수정

유효한 SE를 계산할 수 있도록 Imputation 상황을 관리

❖ **Resampling**

- Missing data의 Resampling에 Imputation 반복 적용

Uncertainty를 원래 표본에서 Resampling으로 추출한 적절한 표본 세트의 point estimates 의 변동성으로 추정

❖ **Multiple imputation**

- Valid SE를 구할 수 있도록 단일 데이터에서 imputation 수정

Imputation으로 인한 추가적인 불확실성을 평가할 수 있는 다중 대치 데이터 세트를 생성

2.2 Imputation Methods that Provide Valid Standard Errors

5.2 explicit sampling variance formulas that allow for missingness

weighting class estimator/ Quasirandomization

Missingness를 허용하는 sampling variance 공식 적용

- Example 4.1에서 weighting class estimator가 adjustment cell 내의 평균을 대체하는 경우
- 선택이 단순 무작위 샘플링에 의해 이루어지고, Missing 이 MCAR 완전 무작위 방식인 경우, Quasirandomization의 명시적 공식을 사용하여 sampling variance 를 구할 수 있다.
- 따라서 명시적 – explicit 공식을 적용해 missing 상황에서 올바른 신뢰구간을 추정할 수 있다.
- 명시적 공식 적용이 가능한 특수한 상황에 대해서는 5장에서 추가로 논의하지 않는다.

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

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

$$\bar{y}_{wc} \pm z_{1-\alpha/2} \{ \hat{\text{mse}}(\bar{y}_{wc}) \}^{1/2}$$

2.2 Imputation Methods that Provide Valid Standard Errors

5.2 From a Single Filled-in Data Set

Standard Errors from Cluster Samples with Imputed Data

Valid SE를 구할 수 있도록 단일 데이터에서 imputation 수정

- 단일 Imputation 으로서도 유효한 Standard Error를 보장하는 방법
- Multistage Cluster Sampling 의 상황에서, K개의 Ultimate Cluster가 단순 무작위 복원 추출 되었을 때 두 가지 조건을 만족하면 \hat{t}_{HT} 가 T 의 Unbiased Estimates 가 되어, Variance 가 구해진다.
- Condition 2 의 경우, Ultimate Cluster 간에 part를 공유 시 조건을 만족하지 못하는데 valid estimate 추정과 작은 편향의 point estimator 추정이 서로 충돌하게 된다.
- 실제로 UC는 복원추출되기 어려워 correlation을 가질 수 밖에 없다.
- -> 제한된 상황에서 적용되어, 일반화된 방법론이 필요하다.

population total for Y

$$T = \sum_{j=1}^K t_j,$$

Horvitz-Thompson estimate

$$\hat{t}_{HT} = \sum_{j=1}^k \frac{\hat{t}_j}{\pi_j},$$

$$\hat{V}(\hat{t}_{HT}) = \sum_{j=1}^k \frac{(k\hat{t}_j/\pi_j - \hat{t})^2}{k(k-1)}$$

An unbiased estimator of the variance of \hat{t} .

Condition 5.1 The estimates \hat{t}_j are unbiased for t_j .

-> Imputation or weighting 이 UCj 안에서 unbiased estimates가 되어야 함.

Condition 5.2 The imputations or weighting adjustments are conducted independently within each UC.

-> Imputation 이나 weighting 이 각 UCj에서 독립적으로 실행되어야 함.

2.3 Standard Errors for Imputed Data by Resampling

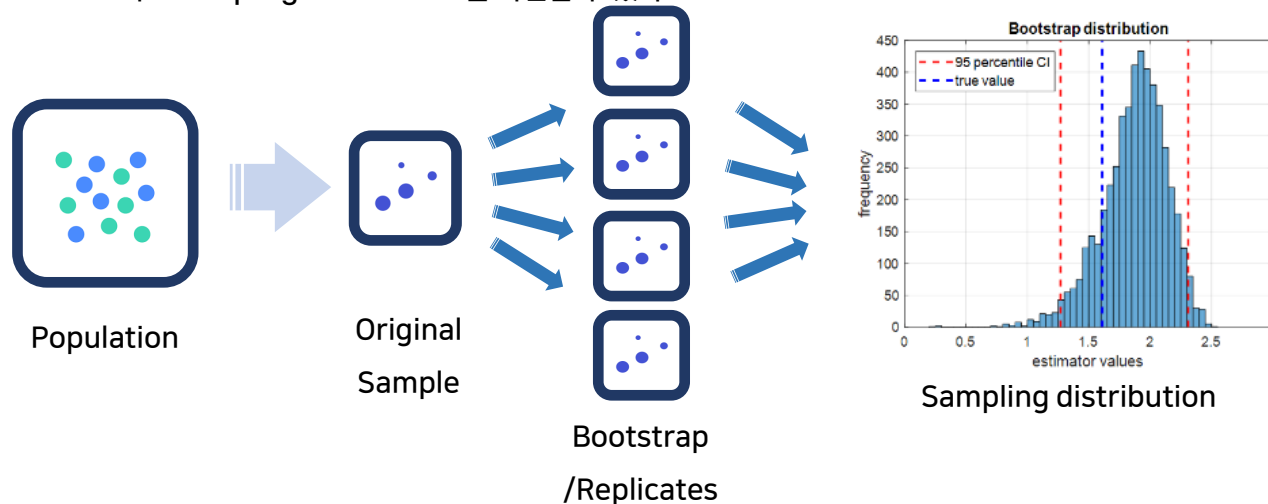
5.3 Resampling Approach

Bootstrap resampling & Jackknife resampling

Missing data의 Resampling에 Imputation 반복 적용

Resampling

- 재표집(Resampling)은 Population에서 추출한 원본샘플 (Original Sample)에서 반복적으로 sample을 뽑는 과정을 말한다.
- sample을 뽑은 후 각 sample 별로 estimates를 계산해 **sampling variance**와 **SE**를 계산하고 sampling distribution을 확인할 수 있다.



Bootstrap

- Original sample에서 크기가 동일한 Bootstrap sample을 단순 무작위 복원 추출로 구한다.
- 몇 개의 Sample을 구할지 설정해야 하며, 통상적으로 많은 Sample을 뽑아야 유의한 추론을 할 수 있다.

Jackknife

- Original Sample에서 한 개의 unit을 drop한 sample을 모아, 원본 샘플 unit의 수 만큼의 Replicates를 구한다.
- 독립적인 추출이 아니기 때문에 consistency에 제약이 있을 수 있다.
- Jackknife를 선형 근사하여 Bootstrap을 얻을 수 있다.

2.3 Standard Errors for Imputed Data by Resampling

5.3.1 Bootstrap Standard Errors

The Simple Bootstrap by Complete data

- 부트스트랩 절차는 하나의 표본에 의해 정의된 경험적 분포로부터 재표집(Resampling) 하는 것을 말한다.
- 단일 표본에서 평균에 대한 estimates는 하나만 얻을 수 있으므로 Variability를 관찰할 수 없다. 따라서 충분히 큰 B에 대하여 Original sample에서 Original sample과 동일하지 않은 B개의 bootstrap sample을 추출함으로써 변동성, Uncertainty를 시뮬레이션 한다.
- Original sample S에서 Bootstrap 표본을 Resampling 할 때 각 Unit들에 대해 단순 무작위 복원 추출이 이루어지며 Bootstrap 표본의 크기는 원본샘플의 크기 n과 같다.
- θ 를 추정하고자 하는 모집단의 parameter라 하면 $\hat{\theta}$ 은 θ 의 일치 추정량이 된다. B개의 Bootstrap sample $S^{(b)}$ 를 추출한 후 각 Bootstrap sample 에 대해 추정한 $\hat{\theta}^{(b)}$ 의 distribution으로 변동성을 나타낼 수 있다.
- 평균(mean)같이 표준 오차가 잘 알려진 estimator들은 이런 부트스트랩 같은 방법을 쓸 이유가 없지만 표준 오차를 계산하는 방법이 잘 알려져 있지 않은 estimator의 경우 의미가 있다.

- 모집단에서 크기 n의 Original Sample을 뽑는다.
- 단순 무작위 복원 추출로 크기가 n인 Bootstrap sample B개를 만든다.
- 각 Bootstrap sample의 estimates를 계산해 평균을 계산한다
- Bootstrap 평균과 분산(Sampling variance) $\hat{\theta}_{boot}$ 과 \hat{V}_{boot} 를 아래 식과 같이 구할 수 있으며, 신뢰구간을 추정할 수 있게 된다.

$$\begin{array}{l} S \Rightarrow S^{(1)} \Rightarrow \hat{\theta}^{(1)} \\ \{i : i = 1, \dots, n\} \quad \quad \quad S^{(b)} \Rightarrow \hat{\theta}^{(b)} \\ \text{단순무작위복원추출} \quad \quad \quad S^{(B)} \Rightarrow \hat{\theta}^{(B)} \end{array} \Rightarrow \begin{array}{l} \hat{\theta}_{boot} = \frac{1}{B} \sum_{b=1}^B \hat{\theta}^{(b)} \\ \hat{V}_{boot} = \frac{1}{B-1} \sum_{b=1}^B (\hat{\theta}^{(b)} - \hat{\theta}_{boot})^2. \\ I_{norm}(\theta) = \hat{\theta} \pm z_{1-\alpha/2} \sqrt{\hat{V}_{boot}} \end{array}$$

2.3 Standard Errors for Imputed Data by Resampling

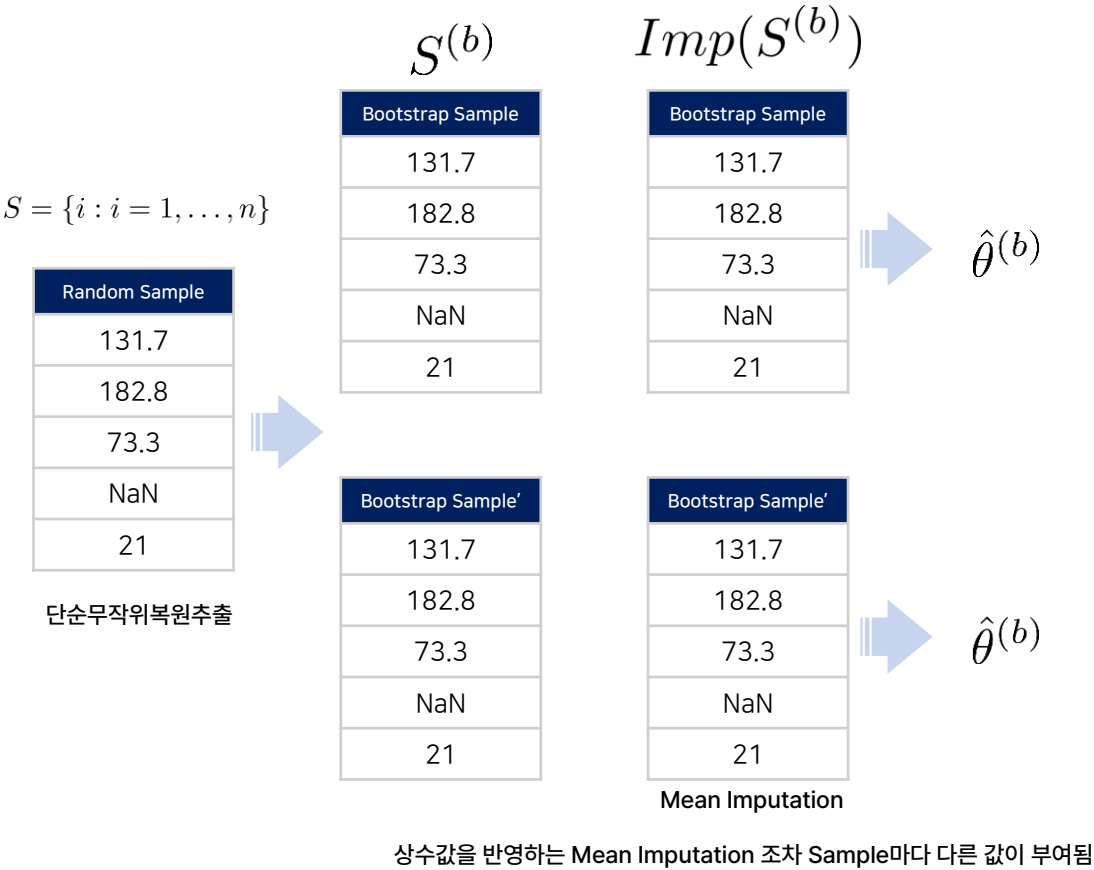
5.3.1 Bootstrap Standard Errors

The Simple Bootstrap Applied to Data Completed by Imputation

- Variability를 시뮬레이션하는 Resampling 방법의 하나인 Bootstrap을 Missing Data에 적용하는 상황
- Single Value Imputation의 Uncertainty 문제를 모든 **Bootstrap Sample의 Missing을 각각 Imputation 하는 방식**으로 해결한다.

1. Missing이 존재하는 원본 S 에서 Bootstrap Sample $S^{(b)}$ 를 생성한다.
 2. Missing이 존재하는 모든 $S^{(b)}$ 에 **각각** Imputation을 적용해 filled-in sample로 만든다.
 3. 각 $S^{(b)}$ 에 대해 동일하게 $\hat{\theta}^{(b)}$ 를 구하고, 이를 활용해 $\hat{\theta}_{boot}$ 과 \hat{V}_{boot} 를 구해 변동성과 신뢰구간을 추정한다.

- 이렇게 표집 단계의 변동성을 Bootstrap Resampling 방식으로 반영하였을 때 표집의 변동성과 함께 Imputation의 Uncertainty를 반영하여 적절한 Standard Error를 계산하고 올바른 신뢰구간을 추론할 수 있다. (*먼저 원본샘플에서 Imputation 후 Bootstrap하면 Imputation Uncertainty를 반영하지 못함)
- Bootstrap sample에 대해 Imputation을 적용해야 해 계산량이 많은 단점이 있다.
- B를 2000 이상 충분히 크게 가져갈 경우 정규성 가정 없이 추론도 가능하다.

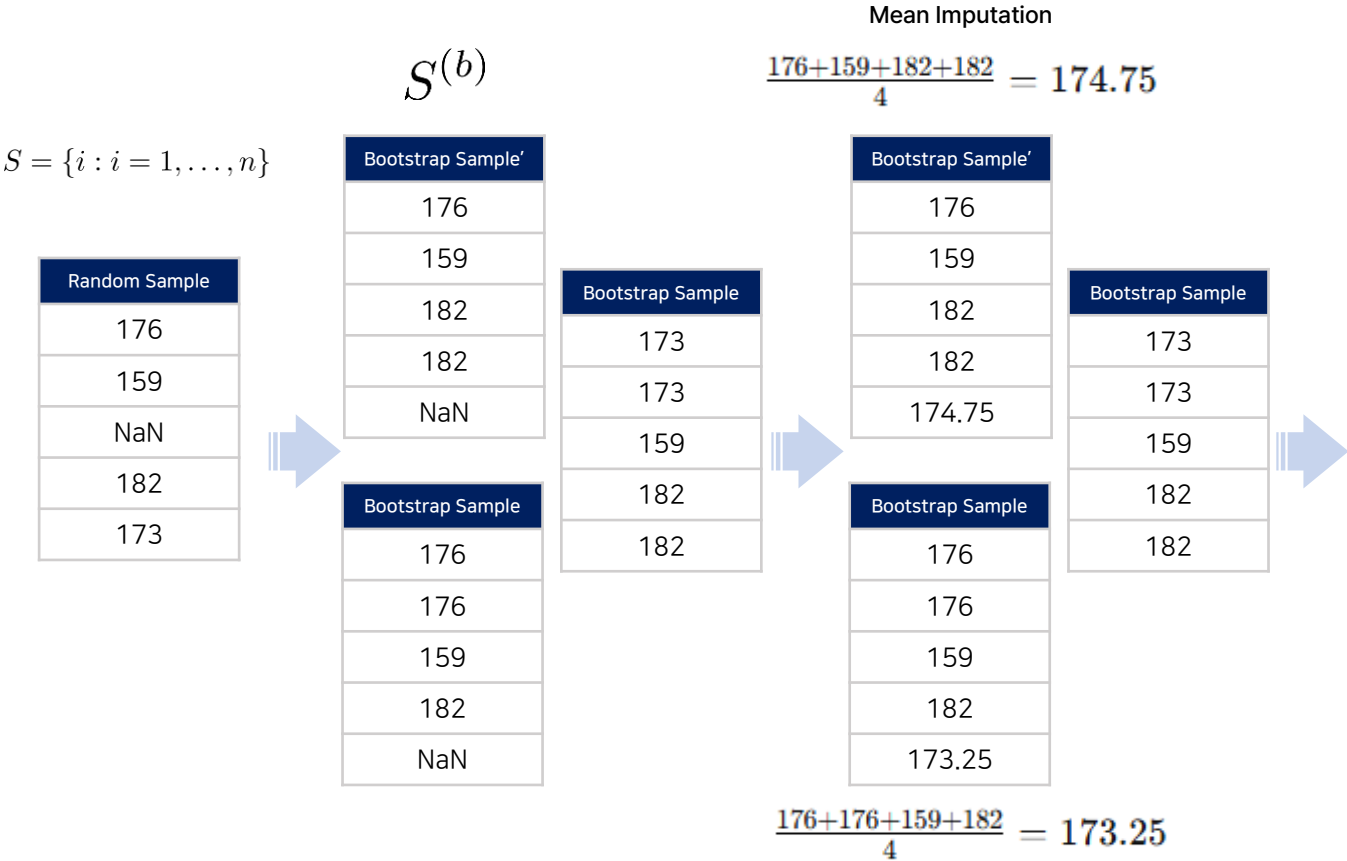


2.3 Standard Errors for Imputed Data by Resampling

5.3.1 Bootstrap Standard Errors : Example

The Simple Bootstrap Applied to Data Completed by Imputation

- n=5인 original sample 에 대해 B=3의 Bootstrap Sample을 뽑아 추정



$$\bar{\theta}_{\text{boot}} = \frac{1}{3}(\hat{\theta}_1 + \hat{\theta}_2 + \hat{\theta}_3) = \frac{1}{3}(174.75 + 173.8 + 173.65) = 174.07$$

$$\hat{V}_{\text{boot}} = \frac{1}{3-1} \sum_{i=1}^3 (\hat{\theta}_i - \bar{\theta}_{\text{boot}})^2$$

$$= \frac{1}{2} ((174.75 - 174.07)^2 + (173.8 - 174.07)^2 + (173.65 - 174.07)^2)$$

$$SE = \sqrt{0.35585} \approx 0.5965$$

$$I_{\text{norm}}(\theta) = 174.07 \pm 1.96 \times 0.5965 \approx 174.07 \pm 1.17$$

Bootstrap 분포가 정규분포에 가깝다고 가정할 때 95%신뢰수준에서 **[172.90, 175.24]**

2.3 Standard Errors for Imputed Data by Resampling

5.3.2 Jackknife Standard Errors

The Simple Jackknife for Complete Data

- Jackknife 방법은 Original sample에서 Unit을 하나 제거함으로써 재표집 (Resampling) 하는 것을 말한다.
- 여러 Resample을 통해 Variability를 반영한다는 점에서 Bootstrap과 동일하지만, 단순 무작위 복원 추출로 Resample하는 대신 각 Unit을 제거해 Resample하는 점에서 Bootstrap과 다르다. 중복을 허용하지 않기 때문에 Jackknife replicates (resampled) 들은 상호간에 독립적이지 않다. 따라서 평균이 아닌 추정량의 경우 Unbiasness 을 보장할 수 없다.

- 모집단에서 크기 n 의 Original Sample을 뽑는다.
- 원본샘플 n 개의 Unit을 하나씩 drop하며 n 개의 Replicates를 만든다.
- 각 Replicates에 대해 pseudo-value를 구하고, 의사도의 평균으로 Jackknife estimates를 구한다.
- Jackknife variance** 를 통해 Standard Error 를 구하고 신뢰구간을 추론한다.

전체 데이터를 모두 이용해 얻은 추정값 i 번째 데이터 제외 추정값

$$\tilde{\theta}_j = \overline{n\hat{\theta}} - (n-1)\hat{\theta}^{(\setminus i)}$$

의사도 : 전체 데이터를 모두 이용해 얻은 추정값에서 i 번째 데이터가 미치는 영향력

-> 영향력이 크다 - Variability가 크다

Jackknife 추정치는 의사도의 평균이다. 영향력의 분산 -> Sampling variance 이다.

$$\begin{aligned} \hat{\theta}_{\text{jack}} &= \frac{1}{n} \sum_{j=1}^n \tilde{\theta}_j = \hat{\theta} + (n-1)(\hat{\theta} - \bar{\theta}), \\ \hat{V}_{\text{jack}} &= \frac{1}{n(n-1)} \sum_{j=1}^n (\tilde{\theta}_j - \hat{\theta}_{\text{jack}})^2 \\ &= \frac{n-1}{n} \sum_{j=1}^n (\hat{\theta}^{(\setminus j)} - \bar{\theta})^2. \\ I_{\text{norm}}(\theta) &= \hat{\theta} \pm z_{1-\alpha/2} \sqrt{\hat{V}_{\text{jack}}} \end{aligned}$$

$S \Rightarrow S^{(\setminus j)} \Rightarrow \hat{\theta}^{(\setminus j)}$ $S^{(\setminus 1)} \Rightarrow \hat{\theta}^{(\setminus 1)}$ $S^{(\setminus n)} \Rightarrow \hat{\theta}^{(\setminus n)}$

$\{i : i = 1, \dots, n\}$ 단순무작위복원추출

2.3 Standard Errors for Imputed Data by Resampling

5.3.2 Jackknife Standard Errors

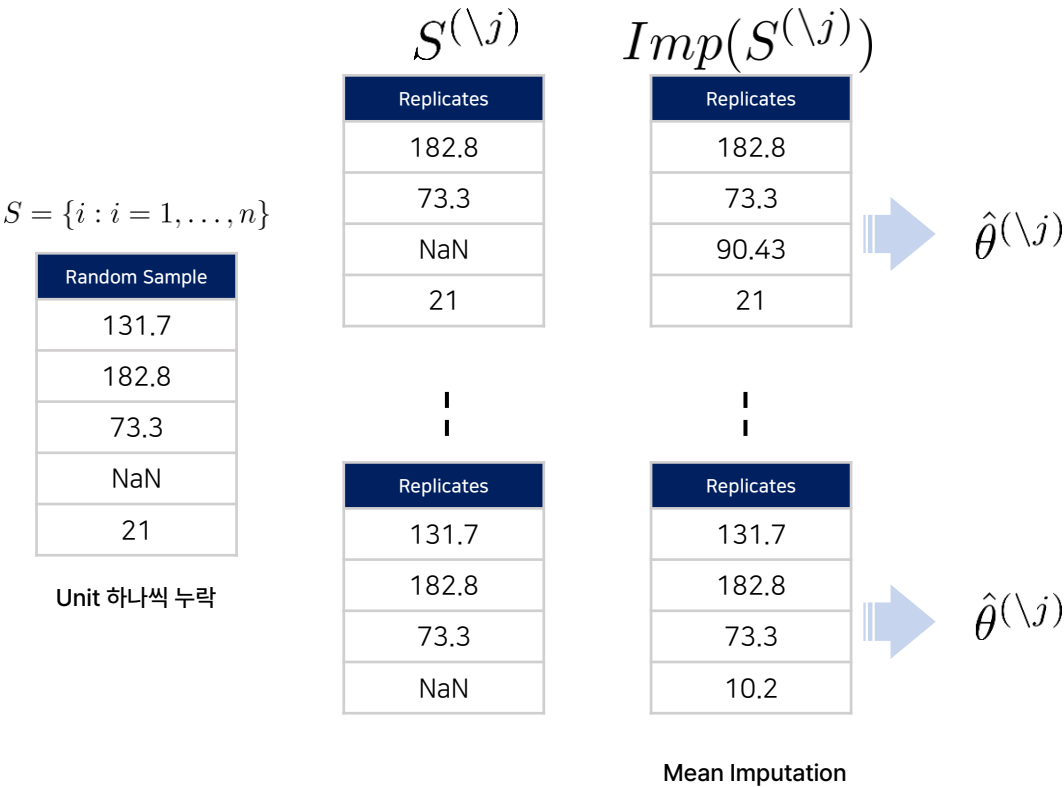
The Simple Jackknife Applied to Data Completed by Imputation.

- Jackknife로 Missing Data를 포함해 Variability 를 시뮬레이션 하는 상황
- Bootstrap과 동일한 절차로 진행된다.

- Single Value Imputation의 Uncertainty 문제를 모든 Jackknife Sample의 Missing을 각각 Imputation 하는 방식으로 해결한다.

1. Missing이 존재하는 원본 S 에서 Jackknife Replicates $S^{(\setminus j)}$ 를 생성한다.
 2. Missing이 존재하는 모든 $S^{(\setminus j)}$ 에 각각 Imputation 된 filled-in sample 로 만든다.
 3. 각 $S^{(\setminus j)}$ 에 대해 동일하게 $\hat{\theta}^{(\setminus j)}$ 와 의사도 $\tilde{\theta}_j$ 를 구하고, 이를 활용해 $\hat{\theta}_{jack}$ 과 \hat{V}_{jack} 를 구해 변동성과 신뢰구간을 추정한다.

- Bootstrap sample에 비해 Imputation을 적용해야 해 계산량이 적다.
- Bootstrap 방식에 비해 정확도가 부족하다



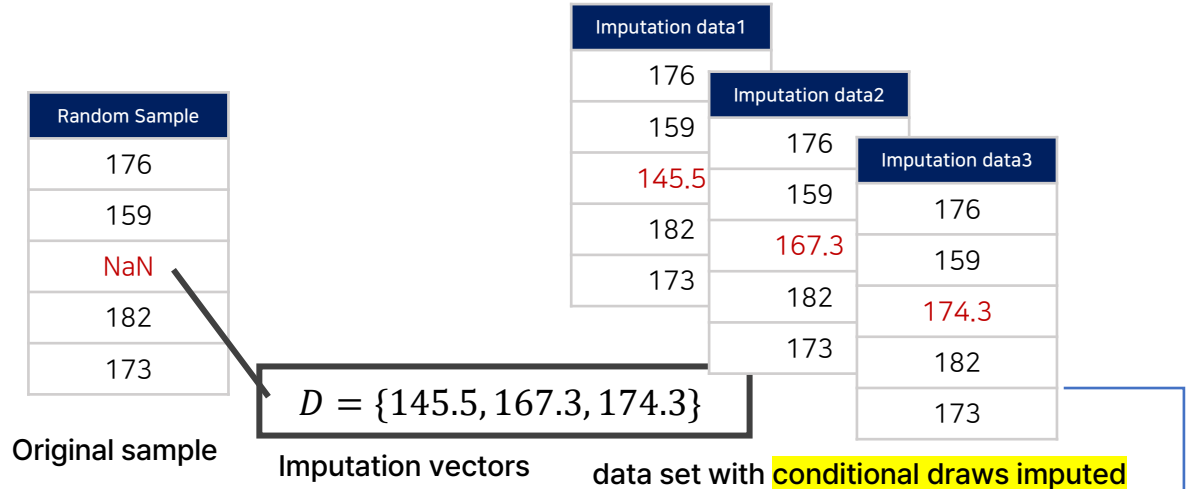
2.4 Introduction to Multiple Imputation

5.4.1 Multiple Imputation Standard Errors

Introductions

Valid SE를 구할 수 있도록 단일 데이터에서 imputation 수정

- Multiple Imputation은 Missing 값을 single value가 아닌 2 이상의 값 D개의 Imputation 값을 가지는 벡터 {Imputation1, Imputation2,..., Imputation D}로 대체하는 것을 말한다.
- 각 Missing에 대해 벡터의 각 값을 대입해 D개의 data set을 생성해 Missing 이 없는 표준적인 방식으로 분석할 수 있다.
- D 개의 Imputation 값은 Missing의 posterior distribution 혹은 hot deck imputation 에 의해 매번 다른 값으로 대체하게 된다.
- Multiple Imputation은 여러 번의 Imputation을 통해 각 샘플의 Uncertainty를 반영하며, 각 Imputation 간의 차이로 Uncertainty를 측정한다.



$$\hat{\theta}^{(d)} \leftarrow$$
$$\bar{\theta}_D = \frac{1}{D} \sum_{d=1}^D \hat{\theta}_d$$

combined estimates

$$W_d = \frac{1}{n_d} \sum_{i=1}^{n_d} (Y_{di} - \bar{Y}_d)^2$$

within-imputation variance

2.4 Introduction to Multiple Imputation

5.4.1 Multiple Imputation Standard Errors

Multiple Imputation methodology

- combined estimate (Imputation data estimate의 추정치)

$$\bar{\theta}_D = \frac{1}{D} \sum_{d=1}^D \hat{\theta}_d$$

- the average within-imputation variance (Imputation data 내 분산의 평균)

$$\bar{W}_D = \frac{1}{D} \sum_{d=1}^D W_d,$$

- between-imputation component (Imputation data 간 분산)

$$B_D = \frac{1}{D-1} \sum_{d=1}^D (\hat{\theta}_d - \bar{\theta}_D)^2.$$

- Total variance** (Imputation 내 분산과 Imputation 간 분산 모두 반영)

$$T_D = \bar{W}_D + \left(1 + \frac{1}{D}\right) B_D$$

- estimates of fraction of information loss due to nonresponse

$$\hat{\gamma}_D = \left(1 + \frac{1}{D}\right) \frac{B_D}{T_D}$$

- \bar{W}_D : *Average Within Imputation variance* 는 각 Dataset 내의 분산들의 평균으로 원래 존재하던 Variability를 의미한다.
- B_D : *Between Imputation Component* 는 Imputation 간에 얼마나 다른 결과가 나왔는지를 의미하며, 이는 Imputation Uncertainty가 크다는 것을 의미한다.
-> Missing으로 생기는 데이터를 채울 때의 Uncertainty를 반영한다.
- T_D : *Total Variance* 는 각 Imputed data 안의 분산과, 각각의 분산을 모두 고려하여 전체 샘플링 Uncertainty 를 나타낸다.
- Adjustment for finite due : D가 작아 Uncertainty를 충분히 반영하지 못하는 문제를 위해 B_D 에 조정 항 $\left(1 + \frac{1}{D}\right)$ 을 곱해준다. D가 무한대로 가면 1로 수렴한다.
- Missing으로 손실된 정보는 Missing으로 인해 추가된 Uncertainty를 말한다.

2.4 Introduction to Multiple Imputation

5.4.1 Multiple Imputation Standard Errors

Multiple Imputation methodology

- 보다 큰 Sample size n 과 상수 θ 에 대해서 interval estimates의 reference distribution 은 t 분포를 따르며, 이때 자유도 ν 는 다음과 같다.
- 이 분포는 Satterthwaite approximation 에 의해 유도된다.
- 위의 Bootstrap, Jackknife 예시처럼 T_D 로 Standard Error를 구해 Variability와 Imputation Uncertainty를 반영하여 신뢰구간을 추정한다.
- 작은 데이터셋에서 더 정확히 자유도를 추정하기 위해 개선된 자유도 공식을 적용하는 경우가 있다.
- 완전한 데이터셋의 자유도와 관찰된 데이터셋의 자유도의 조화평균을 활용해 작은 데이터셋에서 더 정확한 자유도를 추정한다. (Ch.10에서 backup)

- interval estimates $(\theta - \bar{\theta}_D)T_D^{-1/2} \sim t_\nu,$
- degree of freedom ν $\nu = (D - 1) \left(1 + \frac{1}{D + 1} \frac{\bar{W}_D}{B_D} \right)^2,$
- I Intervals : $\bar{\theta}_D \pm t_{\nu, 0.975} \times T_D^{1/2}$
- ν^* : improved expression for df $\nu^* = (\nu_{\text{obs}}^{-1} + \nu_{\text{com}}^{-1})^{-1}$
- ν_{com} : 완전한 데이터셋의 자유도 $\hat{\nu}_{\text{obs}} = (1 - \hat{\gamma}_D) \left(\frac{\nu_{\text{com}} + 1}{\nu_{\text{com}} + 3} \right) \nu_{\text{com}}$

2.4 Introduction to Multiple Imputation

5.4.1 Multiple Imputation Standard Errors : Example

Problem 5.1~5.5 Solving

- 100개의 Unit을 갖는 이변량 데이터 $\{(y_{i1}, y_{i2}), i = 1, \dots, 100\}$ on $(Y1, Y2)$ 는 다음 z_{i1}, z_{i2}, z_{i3} 라는 independent standard normal deviates에 의해 결정된다.
- Missing mechanism 은 관측되지 않는 Latent variable u_i 에 의해 결정된다. 이때 u_i 가 0보다 크면 y_{i2} 에 결측이 생긴다.
- 따라서 y_{i1} 은 missing이 없고, y_{i2} 만 y_{i1} 에 의해 missing이 발생하는 **MAR(Missing at Random) case**이다.
- Python numpy 를 사용해 데이터를 구현하고 Missing을 Resampling과 Multiple imputation으로 처리해 **Sampling Distribution** 을 관찰한다.

Observed data mechanism

$$y_{i1} = 1 + z_{i1},$$

$$y_{i2} = 5 + 2 \times z_{i1} + z_{i2},$$

$$u_i = 2 * (y_{i1} - 1) + z_{i3},$$

$$\{(z_{i1}, z_{i2}, z_{i3}), i = 1, \dots, 100\} \sim N(0, 1)$$

Missing mechanism

$$\Pr(m_{i2} = 1 \mid y_{i1}, y_{i2}, u_i, \phi) = \begin{cases} 1, & \text{if } u_i < 0, \\ 0, & \text{if } u_i \geq 0. \end{cases}$$

2.4 Introduction to Multiple Imputation

5.4.1 Multiple Imputation Standard Errors : Example

Multiple Imputation methodology

```
import numpy as np
import pandas as pd

# 랜덤 시드 설정
np.random.seed(42)

# z_i1, z_i2, z_i3 생성 (표준 정규 분포)
z_i1 = np.random.normal(0, 1, 100)
z_i2 = np.random.normal(0, 1, 100)
z_i3 = np.random.normal(0, 1, 100)

# y_i1, y_i2 생성
y_i1 = 1 + z_i1
y_i2 = 5 + 2 * z_i1 + z_i2

# 잠재 변수 u 생성 및 결측값 설정
u = 2 * (y_i1 - 1) + z_i3
y_i2[u < 0] = np.nan

# 데이터프레임 생성
data = pd.DataFrame({'y_i1': y_i1, 'y_i2': y_i2, 'u': u})

data
```

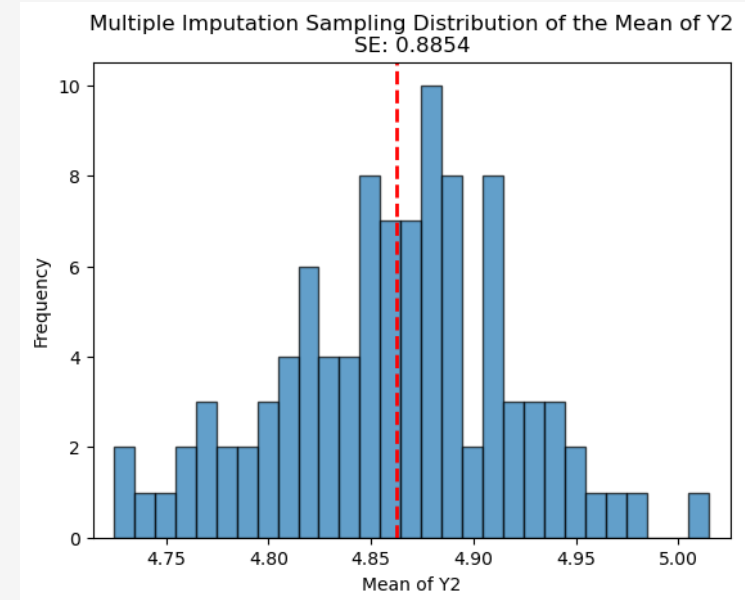
	y_i1	y_i2	u
0	1.496714	4.578058	1.351216
1	0.861736	4.302826	0.284256
2	1.647689	5.952663	2.378428
3	2.523030	7.243782	4.099862
4	0.765847	NaN	-1.845976
...
95	-0.463515	NaN	-3.619939
96	1.296120	4.708383	1.491840
97	1.261055	5.675836	0.829410
98	1.005113	5.068436	0.823089
99	0.765413	3.387855	0.160455

100 rows x 3 columns

	y_i1	y_i2
count	100.000000	55.000000
mean	0.896153	5.845960
std	0.908168	1.682302
min	-1.619745	2.966131
25%	0.399094	4.577171
50%	0.873044	5.678040
75%	1.405952	7.252023
max	2.852278	10.044612

yi1, yi2 데이터 생성 및 Missing 반영

ui < 0 인 row의 yi2 Missing 으로 45개 결측 발생

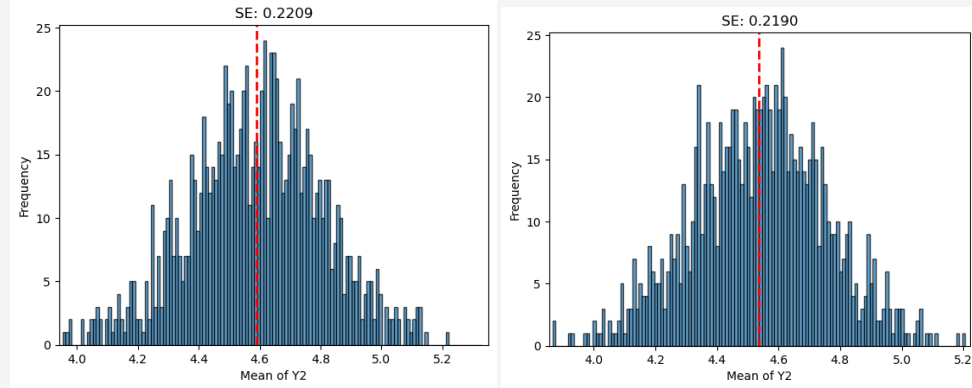


- D = 100, impute method 는 stochastic regression 으로 Multiple Imputation으로 구한 sampling distribution 확인.

2.4 Introduction to Multiple Imputation

5.4.1 Multiple Imputation Standard Errors : Example

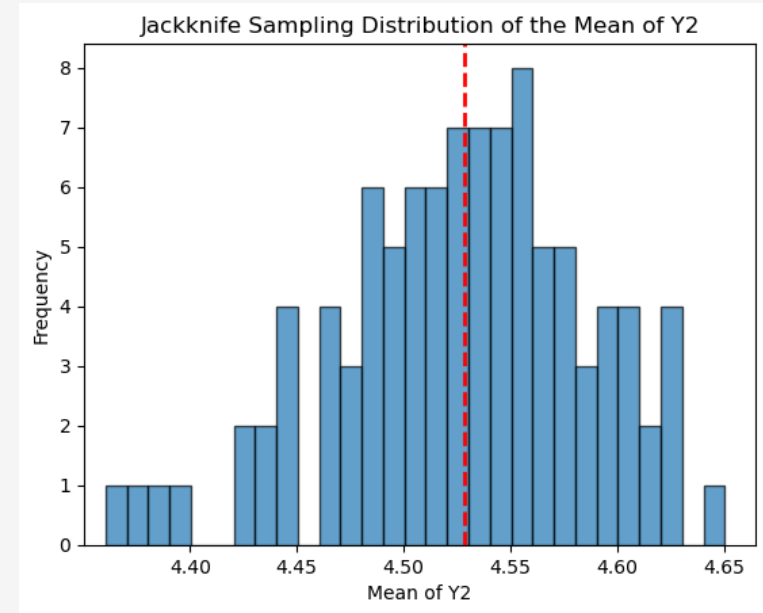
Multiple Imputation methodology



yi2 의 mean estimates 에 대한 Bootstrap sampling distribution.

왼쪽이 bootstrap sampling 후 impute한 결과, 오른쪽이 먼저 impute 후 bootstrap sampling 한 결과.

올바르게 추정된 SE가 Uncertainty를 반영해 더 큰 값을 보인다.



yi2 의 mean estimates 에 대한 Jackknife (n =100) Sampling distribution

Bootstrap 보다 resample의 수가 적음을 확인할 수 있다. 100개의 replicates의 mean분포

2.5 Conclusion

5.5 Comparison of Resampling Methods and Multiple Imputation

Dealing with Missingness uncertainty : Resampling vs Multiple Imputation

Resampling

- 두 방식 모두 Missing value의 predictive distribution을 가정하는 model based 접근이다.
- 두 방식 모두 Imputation을 통해 Missingness를 해결한다.

- Resampling을 통한 추출로 Imputation Uncertainty 반영
- 200개 이상의 Imputed dataset(bootstrap sample, jackknife replicates)가 필요하다. – 계산 집약적이고 많은 storage를 요구한다.
- 최소한의 Missingness model assumption 만으로 Sampling variance에 대한 Consistence estimates를 추정한다.
- Model의 적절성에 영향을 덜 받으며 더 일반적이나, 적절한 model에 대해서는 부정확한 편이다.

Multiple Imputation

- Model에 의한 Imputation Uncertainty를 여러 값을 받아 반영
- Model이 잘 정의되어 있다면 유용하다.
- Model로 Imputation 하는 값 자체가 Sampling Variance 추정의 바탕이 되기 때문에 data의 모델 및 missing mechanism과 더 관련된다
- Model이 적절할 때 더 효과적이지만, 반대로 Model이 부적절한 상황에 더 취약하다.

잘못된 Model에 의한 표준오류 문제는 6장에서 다뤄진다.

2.5 Conclusion

5.5 Review

End of Part1

Part1

- Part1에서는 보다 basic한 approach로서 Complete case analysis, weighting methods, imputation methods를 다루고, Imputation 상황에서의 Uncertainty 를 해결하는 방법을 소개한다.



Part 2, 3

- Part2 에서는 더 이론적인 접근으로, statistical model과 likelihood function 에 의한 model 추정을 다룬다.
 - Part3에서는 Part2에서 다른 방법론들을 적용하는 과정을 다룬다.
-

Assignment

총 30개의 unit >> 이중 11개의 sample 따지기 >> 이중 4개는 respondents이 있다.

결측치 변수 Y에 대하여 4개의 respondent 값은 다음과 같다.

$y_1=10, y_2=12, y_3=10, y_4=18$

그리고 Hot-Deck Method에 의해 $y_1 \sim y_4$ 각각의 값들이 랜덤 샘플링된 횟수 $H_i (i=1,2,3,4)$ 는 다음과 같다.

$H_1=2, H_2=1, H_3=3, H_4=5$

기존의 값 7개는 {11, 12, 17, 19, 23, 13, 17}일 때, Hot Deck에 의해 imputation 된 값들과 기존에 관찰된 값들을 모은 filled-in data를 응용하여 평균을 구하라.

$$\overline{y_{HD}} = \{r\overline{y_R} + (n - r)\overline{y_{NR}}\}/n$$

$$\overline{y_{NR}} = \sum_{i=1}^r \frac{H_i y_i}{n - r}$$

감사합니다