



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

이 학 석 사 학 위 논 문

일반화추정방정식에서 결측대체 효율성
연구



2018년 8월

부 경 대 학 교 대 학 원

통 계 학 과

김 승 찬

이 학 석 사 학 위 논 문

일반화추정방정식에서 결측대체 효율성 연구

지도교수 노 맹 석

이 논문을 석사 학위논문으로 제출함.

2018년 8월

부 경 대 학 교 대 학 원

통 계 학 과

김 승 찬

김승찬의 이학석사 학위논문을 인준함.

2018년 8월 24일



위원장	이학박사	장대흥 (인)
위원	이학박사	하일도 (인)
위원	이학박사	노맹석 (인)

목 차

표 차례	iii
그림 차례	v
제 1장 서론	1
제 2장 연구방법	3
2.1 일반화추정방정식	3
2.1.1 일반화추정방정식 소개	3
2.1.2 독립추정방정식	6
2.1.3 일반화추정방정식	7
2.2 결측 자료	10
2.2.1 결측 구조	10
2.2.2 결측 패턴	12
2.2.3 결측 처리 방법론	13
제 3장 연구자료	20
3.1 자료 소개	20
3.2 결측 패턴	23
3.3 실제자료 분석	27
3.3.1 연속형 자료	28
3.3.2 순서형 자료	30
제 4장 모의실험	33
4.1 모의실험 설계	33
4.2 모의실험 결과	39
4.2.1 연속형 자료	39
4.2.2 순서형 자료	53
제 5장 결론	66

참고문헌	68
부록1 KOSCO 연구 중 본 논문 분석에 사용된 자료	72
부록2 결측 대체 방법론 별 GEE 분석을 위한 R코드	73



표 차례

[표 2.1] 일반적으로 GEE에서 사용되는 가상관행렬 구조	9
[표 2.2] 결측률에 따른 상대효율성	18
[표 3.1] KOSCO 등록/제외 기준	21
[표 3.2] FMA 구성지표	22
[표 3.3] FMA 각 영역별 등급 점수	22
[표 3.4] FAC 구성지표	23
[표 3.5] FMA 결측패턴	24
[표 3.6] FAC 결측패턴	24
[표 3.7] FMA 1년 시점 결측 여부와 연관성	25
[표 3.8] FMA 2년 시점 결측 여부와 연관성	26
[표 3.9] FAC 1년 시점 결측 여부와 연관성	26
[표 3.10] FAC 2년 시점 결측 여부와 연관성	27
[표 3.11] FMA 실제자료에서 회귀계수	29
[표 3.12] FMA 실제자료에서 Robust S.E.	29
[표 3.13] FMA 실제자료에서 P-value	30
[표 3.14] FAC 실제자료에서 회귀계수	31
[표 3.15] FAC 실제자료에서 Robust S.E.	31
[표 3.16] FAC 실제자료에서 P-value	32
[표 4.1] 종속변수 결측 생성 모형	33
[표 4.2] FMA 결측 생성 모형별 결측 형태에 따른	35
[표 4.3] FMA 결측 생성 모형별 결측 형태에 따른	36
[표 4.4] FAC 결측 생성 모형별 결측 형태에 따른	37
[표 4.5] FAC 결측 생성 모형별 결측 형태에 따른	38
[표 4.6] FMA 결측률 10%일 때, 회귀계수	41

[표 4.7] FMA 결측률 10%일 때, 편의	42
[표 4.8] FMA 결측률 30%일 때, 회귀계수	45
[표 4.9] FMA 결측률 30%일 때, 편의	46
[표 4.10] FMA 결측률 50%일 때, 회귀계수	49
[표 4.11] FMA 결측률 50%일 때, 편의	50
[표 4.12] FAC 결측률 10%일 때, 회귀계수	54
[표 4.13] FAC 결측률 10%일 때, 편의	55
[표 4.14] FAC 결측률 30%일 때, 회귀계수	58
[표 4.15] FAC 결측률 30%일 때, 편의	59
[표 4.16] FAC 결측률 50%일 때, 회귀계수	62
[표 4.17] FAC 결측률 50%일 때, 편의	63

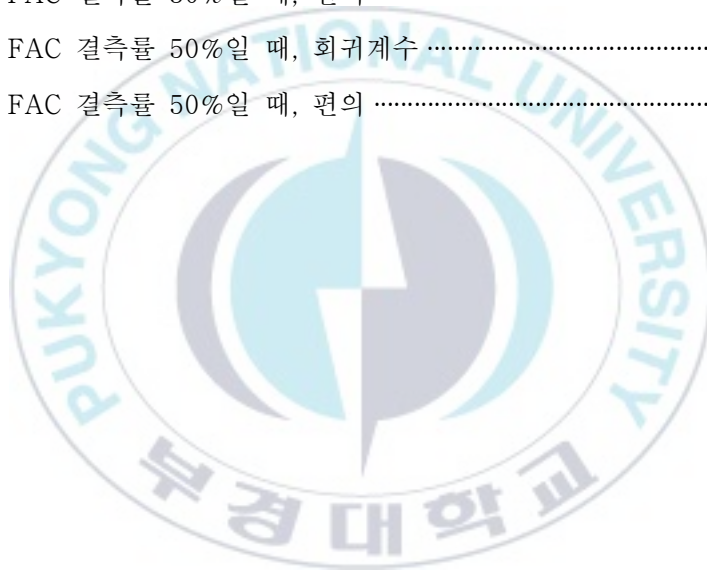


그림 차례

[그림 2.1] 결측 패턴	13
[그림 2.2] 다중대체 절차	16
[그림 4.1] FMA 결측률 10%일 때, MCAR에서의 상대편의	43
[그림 4.2] FMA 결측률 10%일 때, MAR에서의 상대편의	43
[그림 4.3] FMA 결측률 10%일 때, sNMAR에서의 상대편의	44
[그림 4.4] FMA 결측률 10%일 때, cNMAR에서의 상대편의	44
[그림 4.5] FMA 결측률 30%일 때, MCAR에서의 상대편의	47
[그림 4.6] FMA 결측률 30%일 때, MAR에서의 상대편의	47
[그림 4.7] FMA 결측률 30%일 때, sNMAR에서의 상대편의	48
[그림 4.8] FMA 결측률 30%일 때, cNMAR에서의 상대편의	48
[그림 4.9] FMA 결측률 50%일 때, MCAR에서의 상대편의	51
[그림 4.10] FMA 결측률 50%일 때, MAR에서의 상대편의	51
[그림 4.11] FMA 결측률 50%일 때, sNMAR에서의 상대편의	52
[그림 4.12] FMA 결측률 50%일 때, cNMAR에서의 상대편의	52
[그림 4.13] FAC 결측률 10%일 때, MCAR에서의 상대편의	56
[그림 4.14] FAC 결측률 10%일 때, MAR에서의 상대편의	56
[그림 4.15] FAC 결측률 10%일 때, sNMAR에서의 상대편의	57
[그림 4.16] FAC 결측률 10%일 때, cNMAR에서의 상대편의	57
[그림 4.17] FAC 결측률 30%일 때, MCAR에서의 상대편의	60
[그림 4.18] FAC 결측률 30%일 때, MAR에서의 상대편의	60
[그림 4.19] FAC 결측률 30%일 때, sNMAR에서의 상대편의	61
[그림 4.20] FAC 결측률 30%일 때, cNMAR에서의 상대편의	61
[그림 4.21] FAC 결측률 50%일 때, MCAR에서의 상대편의	64
[그림 4.22] FAC 결측률 50%일 때, MAR에서의 상대편의	64

[그림 4.23] FAC 결측률 50%일 때, sNMAR에서의 상대편의	65
[그림 4.24] FAC 결측률 50%일 때, cNMAR에서의 상대편의	65



Efficiency of Imputation Methods in Generalized Estimating Equations

Seung Chan Kim

Department of Statistics, The Graduate School,
Pukyong National University

Abstract

Comparing with using cross-sectional data, we can estimate the parameters more efficiently using panel data such as repeated measured data (Hsiao, 1985). In many medical research, we often try to collect repeatedly measured patient data over periods in order to learn their time trend if any. However, while collecting the data over time, there can be missing values at some time points for several reasons. Their missing mechanism can be also vary from MCAR(Missing Completely at Random), MAR(Missing at Random), and NMAR(Not Missing at Random) (Little and Rubin, 2002).

In this thesis, we want to investigate and compare the efficiency of several missing value imputation methods while we fit the generalized estimating equation models for panel data. We conducted simulation studies using several scenarios and analyzed real data using KOSCO (The Korean Stroke Cohort for functioning and rehabilitation) data.

We conducted the simulation study using simulation data for several scenarios and applied the idea to real data using KOSCO (The Korean Stroke Cohort for functioning and rehabilitation) data.

The results show that multiple imputation works best in terms of bias whether the missing mechanism is MCAR, MAR or NMAR.

keywords: Panel data, Missing data, MCAR, MAR, NMAR, GEE, KOSCO



제 1장 서론

패널자료(panel data)는 횡단면자료(cross sectional data) 분석에 비해 추정의 효율성 향상 등 장점이 있다(Hsiao, 1985). 패널자료를 다룸에 있어 주의할 점은 다시점에 관찰된 종속변수가 상관성을 가진다는 점이다. 이러한 상관성을 가지는 자료를 다루는 대표적인 방법으로는 일반화추정방정식(Generalized Estimating Equations)과 일반화선형혼합모형(Generalized Linear Mixed Model)이 있다. 두 방법론은 관심에 따라 적절하게 사용하게 되는데, 주된 관심이 '군집 간' 비교 일 때는 주변모형을 이용하는 일반화추정방정식의 사용이 적절하다(Agresti, 2007).

대상자를 추적 관찰하는 경우 시점에 따라 여러번 조사하게 되는데, 추적 도중 어떤 이유에서 한 차순 거절하거나, 중도 탈락, 또는 사망하게 되는 등 다양한 원인에 의해 결측치가 생기게 된다. 임상자료의 경우 기저(base line)에 조사된 독립변수는 결측이 잘 발생하지 않으나, 여러번 추적함에 따라 종속변수의 결측이 발생하기 쉽다. 이러한 결측 발생은 추정치의 편향을 야기할 수 있어, 그 처리에 주의를 기울여야 한다.

본 논문에서 다룰 일반화추정방정식의 경우, 관찰된 자료만을 이용한 분석은 완전임의결측인 경우에서만 비편향추정치들을 제공하는 문제가 있다. 완전임의결측이란, 결측치의 결측원인이 관찰된 자료에 영향을 받지 않는, 다시 말하면 통계적으로는 결측된 관측치와 결측되지 않은 관측치들이 서로 독립이라는 의미이다. 실제자료에서는 결측 발생이 관찰된 자료에 영향을 받는 경우가 자주 발생한다. 예를 들어, 어떤 환자의 '삶의 질'에 관해 조사한다면, 그 환자의 현재 삶의 질은 이전 시점에 관찰된 '삶의 질'이 낮은 경우, 결측 발생할 확률이 더 높을 수 있다. 이와 같은

경우, 결측치의 처리에 따라 결과가 크게 달라 질 수 있으므로 그처리에 주의를 요하게 된다.

본 논문에서는 질병관리본부 정책용역사업인 ‘뇌졸중 환자의 재활분야 장기적 기능 수준 관련 요인에 대한 10년 추적조사 연구(The Korean Stroke Cohort for functioning and rehabilitation, KOSCO)’ 자료를 이용해 분석하였다. 해당 자료는 종속변수가 다 시점에 걸쳐 측정되는 구조로, 연구 목적에 따라 GEE 방법을 썼으나 종속변수의 결측률이 높아 결측 처리에 따른 결과가 크게 차이날 것으로 예상되었다. 실제로 다양한 결측대체 대체 방법을 적용하여 분석해보니, 그 처리에 따른 결과가 크게 달랐다. 따라서 본 논문에서는 다 시점 자료에서 결측 처리에 따른 GEE 추정량의 효율성을 평가하고자 한다. 2장에서는 일반화추정방정식과 결측 처리 방법론에 대한 이론을 정리한다. 3장에서는 결측률이 높은 실제자료의 결측처리에 따른 GEE 분석결과에 대해 살펴본다. 4장에서는 모의실험을 통해 결측률과 결측구조에 따른 결측대체 방법론 별 GEE 추정량의 효율성을 평가하고, 5장에서는 그에 따른을 내고자 한다.

제 2장 연구방법

2.1 일반화추정방정식

2.1.1 일반화추정방정식 소개

패널자료에서는 다시점의 종속변수의 평균을 모형화해야 하는데, 이때 종속변수가 연속형이 아닌 경우 다변량 분포를 고려해 상관성을 모형화하기는 어렵다. 종속변수의 분포형태가 구체화 된다면 최우추정법을 통해 모형에서의 모수를 추정할 수 있다. 그러나 구체화되기 힘든 경우가 있는데, 이때 Wedderburn은 다음과 같은 유사우도함수(Quasi-likelihood function)를 제안했다.(Wedderburn, 1974)

$$\frac{\partial Q(\mu_i, y_i)}{\partial \mu_i} = \frac{y_i - \mu_i}{V(\mu_i)} \quad (1.1)$$

$$Q(\mu_i, y_i) = \int_{y_i}^{\mu_i} (y_i - t)^T V(\mu_i)^{-1} dt \quad (1.2)$$

y_i 들이 독립이라 가정하면, 유사우도함수는

$$Q(\mu, y) = \sum_i Q(\mu_i, y_i) \quad (1.3)$$

β 에 대한 유사우도추정량은 유사우도추정방정식 $U(\beta)$ 의 해이다.

$$U(\beta) = \sum_i D_i^T V_i^{-1} (y_i - \mu_i(\beta)), \text{ where } D_i = \frac{\partial \mu_i}{\partial \beta} \quad (1.4)$$

피셔 점수화(Fisher scoring) 방법을 이용, 다음 식을 만족하는 해를 구하면 β 에 대한 추정치를 얻을 수 있다.

$$U(\beta) = 0 \quad (1.5)$$

위 유사우도함수에 기반하여 추정법을 확장한 일반화추정방정식에 대해 알아보도록 하자.

일반화추정방정식은 군집자료 또는 패널자료에 대한 일반화선형모형의 확장으로, 전체 모집단의 평균적인 종속변수 증가분이 관심 대상일때 선호되는 방법이다.

패널자료에 대해 다음과 같이 기호를 표기하고자 한다. 개체 $i(i=1,2, \dots, N)$ 에 대해, j 번째 반응을 다음과 같이 쓰면 $j(j=1,2, \dots, n_i)$ Y_i 와 X_i 는 다음과 같다.

$$Y_i = (y_{i1}, \dots, y_{1n_i})^T, \quad (i=1, \dots, N; j=1, \dots, n_i), \quad n_i \times 1 \text{ 벡터} \quad (1.6)$$

$$X_i = (x_{i1}, \dots, x_{1n_i})^T, \quad (i=1, \dots, N; j=1, \dots, n_i), \quad n_i \times p \text{ 행렬} \quad (1.7)$$

이 방법은 반복측정된 Y_i 들 간의 결합확률분포 형태에 대한 가정 없이 각각의 주변확률분포에 대한 가정에 기초하여 추정이 가능하다. 대신 결합확률분포에 대한 약한 가정 하에서 β 에 대한 일치추정량을 얻는다. 즉, Y_i 들 간의 상관행렬의 구조만 가정하여 추정한다. 이러한 이유로 준모수적 모형(Semi-parametric model)으로 불린다. (Liang and Zeger, 1986; 질병관리본부, 2007) y_{ij} 에 대한 주변확률분포를 다음과 같이 가정하면,

$$f(y_{ij}) = \exp[\{y_{ij}\theta_{ij} - a(\theta_{ij}) + b(y_{ij})\}\phi], \text{ where } \theta_{ij} = g(\eta_{ij}), \eta_{ij} = x_{ij}'\beta \quad (1.8)$$

y_{ij} 에 대한 1, 2차 적률을 다음과 같이 얻을 수 있다.

$$E[y_{ij}|x_{ij}] = a'(\theta_{ij}) = \mu_{ij} \quad (1.9)$$

$$\text{var}(y_{ij}|x_{ij}) = a''(\theta_{ij})\phi = v(\mu_{ij})\phi = v_{ij} \quad (1.10)$$

y_{ij} 의 평균이 μ_{ij} 이고, 분산이 μ_{ij} 의 함수이다. 여기서 v 는 알려진 분산 함수이고, ϕ 는 산포모수이다.

Y_i 는 개체(또는 군집) 간에는 독립이라 가정하지만, 개체 내 종속성은 반복측정 등의 이유로 상관성에 대한 가정을 한다. Y_i 에 대한 분산-공분산행렬 V_i 는 다음과 같다.

$$V(\beta, \alpha, \phi) = V_i = A_i^{1/2} R_i(\alpha) A_i^{1/2} \phi \quad (1.11)$$

여기서 $A_i = \text{Diag}\{\nu(\mu_{i1}), \dots, \nu(\mu_{in_i})\}$ 이고, $R_i(\alpha)$ 는 가상관행렬로 $\text{Corr}(Y_i)$ 에 대한 모형이며, α 는 장애모수이다.

2.1.2 독립추정방정식

만약 개체 내 종속성을 고려하지 않는다고 가정해보자. 즉, 반복측정된 관찰값이 독립이라면 가상관행렬은 독립성 가정 $\text{Cov}(y_{ij}, y_{ij'}) = 0$,

if $j \neq j'$ 을 고려해야 한다. $y_i \sim N(\mu_i, V_i)$, where $\mu_i = (\mu_{i1}, \mu_{i2}, \dots, \mu_{in_i})'$ and $V_i = \text{diag}(v_{ij})$ 라 가정하자. 유사우도함수로부터 다음의 스코어함수를 얻을 수 있다.

$$U_I(\beta) = \sum_{i=1}^N D_i^T V_i^{-1} (y_i - \mu_i(\beta)) = \sum_{i=1}^N D_i^T \text{diag}(v_{ij}^{-1}) (y_i - \mu_i(\beta)) = 0 \quad (1.12)$$

여기서 $D_i = \frac{\partial \mu_i}{\partial \beta}$ 이다. (1.13)의 해로부터 $\hat{\beta}_I$ 를 얻을 수 있다.

추정한 $\hat{\beta}_I$ 에 대한 $\sqrt{N}(\hat{\beta}_I - \beta)$ 의 근사분포는 평균=0, (1.13)의 분산= V_I 인 다변량 정규분포를 따른다.

$$V_I = \lim_{N \rightarrow \infty} N \left(\sum_{i=1}^N D_i^T V_i^{-1} D_i \right)^{-1} \quad (1.13)$$

2.1.3 일반화추정방정식

위에서는 반복측정을 고려하지 않고 y_i 들이 독립일 때를 가정하여 유사우도함수를 정의했다. 반복측정을 고려하면 y_i 들이 상관성을 가질 수 있다. Liang and Zeger은 독립추정방정식의 한계 $Cov(y_{ij}, y_{ij'}) = 0$ 를 극복하고자 다음 방법을 제시하였다(Liang and Zeger, 1986). y_i 들의 상관성을 주변분포만을 이용해 종속변수의 평균 $E[y_i]$ 와 분산-공분산 행렬 V_i 를 모형화 하는 것이다. y_i 들의 상관성을 고려한 보다 일반적인 경우 (1.11)에 대해 유사우도함수로부터 다음의 스코어함수를 얻을 수 있다.

$$\begin{aligned} U_G(\beta) &= \sum_{i=1}^N D_i^T V_i^{-1} (y_i - \mu_i(\beta)) \\ &= \sum_{i=1}^N D_i^T A_i^{-1/2} R_i^{-1}(\alpha) A_i^{-1/2} \phi(y_i - \mu_i(\beta)) = 0 \end{aligned} \quad (1.14)$$

여기서 $D_i = \frac{\partial \mu_i}{\partial \beta}$ 이다. $\hat{\beta}_G$ 과 $\hat{\alpha}$ 는 다음의 두 단계를 거쳐 추정할 수 있다.

단계1: 초기값 α 를 설정준 후, β 의 해를 (1.14)으로부터 추정한다. 이때, 초기값은 상관행렬구조의 특별한 경우인 단위행렬(Identity matrix)이 편리하게 이용될 수 있다.

단계2: 단계1에서 추정된 $\widehat{\beta}_G$ 를 이용해 얻어진 피어슨잔차 $r_{ij} = \frac{y_{ij} - \widehat{\mu}_{ij}}{\sqrt{v(\widehat{\mu}_{ij})}}$ 로부터 α 와 ϕ 를 적률 추정법으로 구하고, 피셔-점수화 방법을 이용해 $\widehat{\beta}^{(m+1)} \approx \widehat{\beta}^{(m)}$ 이 될 때까지 반복하여 수렴하는 해를 얻을 수 있다.

$$\begin{aligned} \widehat{\beta}^{(m+1)} = \widehat{\beta}^{(m)} + [\sum_{i=1}^N D_i^T(\widehat{\beta}^{(m)}) \{ V_i^{(m)} \}^{-1} D_i(\widehat{\beta}^{(m)})]^{-1} \\ [\sum_{i=1}^N D_i^T(\widehat{\beta}^{(m)}) \{ V_i^{(m)} \}^{-1} \{ Y_i - \mu_i(\widehat{\beta}^{(m)}) \}] \end{aligned} \quad (1.15)$$

적률 추정법으로 ϕ 는 다음과 같이 추정할 수 있고,

$$\widehat{\phi} = \frac{1}{Nn_i - p} \sum_{i=1}^N \sum_{j=1}^{n_i} \frac{(y_{ij} - \widehat{\mu}_{ij})^2}{v(\widehat{\mu}_{ij})} \quad (1.16)$$

α 는 가상관행렬 구조에 따라 다음과 같이 추정할 수 있다.

표 2.1 일반적으로 GEE에서 사용되는 가상관행렬 구조

Correlation Structure	Corr($Corr(Y_{ij}, Y_{ij'})$)	Sample matrix	Estimator
Independent	$Corr(Y_{ij}, Y_{ij'}) = \begin{cases} 1, & j = j' \\ 0, & j \neq j' \end{cases}$	$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$	NA
Exchangeable	$Corr(Y_{ij}, Y_{ij'}) = \begin{cases} 1, & j = j' \\ \alpha, & j \neq j' \end{cases}$	$\begin{pmatrix} 1 & \alpha & \alpha \\ \alpha & 1 & \alpha \\ \alpha & \alpha & 1 \end{pmatrix}$	$\hat{\alpha} = \frac{1}{(K'-p)\phi} \sum_{i=1}^N \sum_{j \neq j'} e_{ij} e_{ij'}$ $N' = \sum_{i=1}^N n_i(n_i - 1)$
AR(1)	$Corr(Y_{ij}, Y_{ij'+m}) = \alpha^m, m = 0, 1, 2, \dots, n_i - j$	$\begin{pmatrix} 1 & \alpha & \alpha^2 \\ \alpha & 1 & \alpha \\ \alpha^2 & \alpha & 1 \end{pmatrix}$	$\hat{\alpha} = \frac{1}{(N_1-p)\phi} \sum_{i=1}^N \sum_{j \neq j'} e_{ij} e_{i,j+1}$ $N_1 = \sum_{i=1}^N (n_i - 1)$
Unstructured	$Corr(Y_{ij}, Y_{ij'}) = \begin{cases} 1, & j = j' \\ \alpha_{jj'}, & j \neq j' \end{cases}$	$\begin{pmatrix} 1 & \alpha_{12} & \alpha_{13} \\ \alpha_{21} & 1 & \alpha_{23} \\ \alpha_{31} & \alpha_{32} & 1 \end{pmatrix}$	$\hat{\alpha}_{jj'} = \frac{1}{(N-p)\phi} \sum_{i=1}^N e_{ij} e_{ij'}$

(Wang, 2014)

위 방법을 이용해 추정한 $\hat{\beta}_G$ 에 대해 $\sqrt{N}(\hat{\beta}_G - \beta)$ 의 근사분포는 평균 = 0, (1.14)의 분산 = V_G 인 다변량 정규분포를 따른다.

$$V_G = \lim_{N \rightarrow \infty} N(\hat{B}^{-1} \hat{M} \hat{B}^{-1}) \quad (1.17)$$

여기서, $\hat{B} = (\sum_{i=1}^N D_i^T V_i^{-1} D_i)$, $\hat{M} = \sum_{i=1}^N D_i^T V_i^{-1} Cov(Y_i) V_i^{-1} D_i$ 이다.

GEE 추정량의 특징은 어떠한 가상관행렬을 선택하더라도 평균모형이

옳게 가정된다면 $\hat{\beta}_G$ 은 β 의 일치추정량이 되며, 근사적으로 정규분포를 따르게 된다는 점이다.

2.2 결측 자료

2.2.1 결측 구조

결측구조는 MCAR(Missing Completely at Random), MAR(Missing at Random), NMAR(Not Missing at Random) 세가지로 나눠 정의할 수 있다(Little and Rubin, 2002). 종속변수 Y 는 관측된 경우를 Y_{obs} (관찰된 자료), 결측된 경우를 Y_{miss} (관찰되지 않은 자료)로 나타낸다면 $Y = (Y_{obs}, Y_{miss})$ 로 분할 할 수 있다. 또한 결측 여부에대한 지시변수를 R 이라 하자. 그러면 결측여부를 다음과 같이 나타낼 수 있다.

$$R_i = \begin{cases} 1 & \text{if } Y_i \text{ is missing} \\ 0 & \text{if } Y_i \text{ is observed} \end{cases} \quad (2.1)$$

가. 완전임의결측(MCAR: Missing Completely at Random)

MCAR은 결측 발생이 무작위로 나타나는 경우이다. Y 가 주어졌을 때 결측이 발생할 확률은 Y 내의 어떤 자료와도 연관되지 않음을 의미한다.

예를 들어, 자료 기입 도중에 실수로 해당 값을 누락시키는 경우 완전임의결측에 해당된다.

$$P(R|Y_{obs}, Y_{miss}) = P(R) \quad (2.2)$$

나. 임의결측(MAR: Missing at Random)

MAR은 결측 발생이 다른 변수에 영향 받는 경우로, 결측이 일어날 확률이 관측된 자료 Y_{obs} 에만 연관되어 있고, 결측자료 Y_{miss} 와는 관련이 없는 경우를 의미한다. 예를 들어, 삶의 질에 대한 변수에 결측이 발생했다고 가정하면, 임의결측은 결측여부가 삶의 질 외의 다른 요인인 나이, 성별, 학력 등과 같은 변수와 연관이 있는 경우를 말한다. 패널자료에선 종속변수를 반복측정한다고 할 때, t시점의 삶의 질이 t이전 시점의 삶의 질과 연관되는 경우를 말한다. 즉, 종속변수의 결측여부 R는 관측된 자료 Y_{obs} 와는 연관이 있을 수 있으나, Y_{miss} 와는 연관이 없음을 의미한다.

$$P(R|Y_{obs}, Y_{miss}) = P(R|Y_{obs}) \quad (2.3)$$

다. 비임의결측 (NMAR: Not Missing at Random)

NMAR은 결측 발생이 자기자신에 영향 받는 경우. 즉, 결측발생이 그

변수 자체와 연관 있거나 주어진 자료 외의 다른 변수들과 연관있는 경우를 말한다. 예를들어, 실제 삶의 질이 낮을수록 삶의 질에 대한 응답률이 저조해 지는 경우가 해당된다. 패널자료의 경우 t 이전 시점이 아닌 t 시점에서의 삶의 질이 결측률과 연관되어 있는 것을 의미한다.

$$P(R|Y_{obs}, Y_{miss}) \neq P(R|Y_{obs}) \quad (2.4)$$

2.2.2 결측 패턴

결측패턴은 대표적으로 일변량패턴(Univariate Nonresponse Pattern), 단조패턴(Monotone Pattern), 임의패턴(Arbitrary Pattern)이 있다. 일변량패턴은 다른 변수들에서는 결측이 발생하지 않으나, 단일 변수에서 결측이 생기는 경우이다. 단조패턴은 해당 시점에 결측이 생기면 그 이후 추적시점에도 계속 결측으로 볼 수 있다. 임의패턴은 무계획적인 결측 패턴으로, 결측 발생의 가장 일반적인 형태이다.

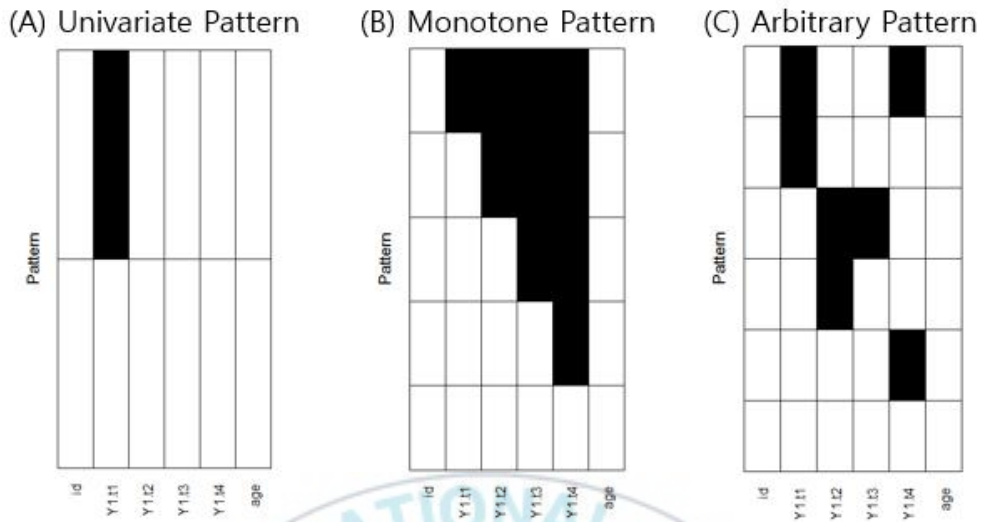


그림 2.1 결측 패턴

2.2.3 결측 처리 방법론

결측 처리 방법론으로는 대표적으로 삭제법, 단일대체법, 다중대체법을 고려할 수 있다.

가. 삭제법

삭제법으로는 완전제거법 (Listwise Deletion or Complete Case Analysis)이 있다. 완전제거법은 한 개체에 적어도 하나의 결측이 있는 경우 그 개체를 삭제하고 분석하는 방법이다. 표본수가 줄어들기에 검정력이

줄어든다는 단점이 있다. 그리고 MCAR가정을 벗어나면 편의가 생길 수 있다(Baraldi and Enders, 2010).

나. 단일대체법

패널자료에서 대표적인 단일대체법은 이월대체 (Last Observation Carried Forward), 핫덱대체 (Hotdeck Imputation), 행평균대체 (Row Mean Imputation), 행중위수대체 (Row Median Imputation), 회귀적대체 (Regression Imputation)가 있다. 이월대체는 한 개체를 추적조사 하는 중 결측이 발생하면 결측이 발생하기 이전 시점의 관찰값으로 해당 시점값을 채워넣는 방법이다. 이전 시점의 측정 이후로 값이 변하지 않는다는 가정이 필요하다. 임상자료에서 자주 사용되는 방법인데, MCAR가정하에서도 추정값이 왜곡될 수 있어 주의가 필요하다(Enders, 2010).

핫덱대체는 추적 관찰된 시점의 값을 임의로 선택하여 결측값을 채워넣는 방법이다. 예를들어 3개 시점중 마지막 시점의 값이 결측된다면 처음 시점과 두 번째 시점 관찰 값을 임의로 선택해, 그 값으로 해당 시점의 결측을 채워넣는 방법이다.

행평균대체는 개체의 추적 관찰된 시점의 값만으로 평균을 낸 뒤, 결측값을 채워넣는 방법이다.

$$\hat{y}_{it} = \frac{\sum_{t=1}^T r_{it} y_{it}}{\sum_{t=1}^T r_{it}}, \text{ where } r_i = \begin{cases} 1 & \text{if } y_i \text{ observed} \\ 0 & \text{if } y_i \text{ missing} \end{cases} \quad (2.5)$$

행중위수대체는 행평균대체와 유사하게 개체의 개체의 추적 관찰된 시점의 값만으로 중위수를 낸 뒤, 결측값을 채워넣는 방법이다.

회귀적대체는 결측이 발생한 변수를 종속변수로 그 외 변수들을 독립변수로 설정해 모형에 적합시킨 후, 그 모형으로 결측값을 예측하는 방법이다. 평균대체보다 뛰어난 방법이지만 예측편의가 발생할 수 있다. 또한 상관관계를 과대추정하는 단점이 있다(Enders, 2010).

다. 다중대체법

단일대체는 분산을 과소 추정하는 문제점이 있다. 한계점을 극복하고자 Rubin(1987)이 제안한 다중대체 방법은 다음과 같은 절차를 따른다.

- 1) 대체단계(Imputation phase): 지정한 결측대체 대체 방법을 m ($m \geq 2$)번 시행하여 m 개의 완전한 자료를 만든다.
- 2) 분석단계(Analysis phase): m 개의 완전한 자료 각각에 일반적인 분석 방법론대로 분석한다.
- 3) 통합단계(Pooling phase): 각각의 m 개의 모수 추정치를 하나의 결과로 통합추론한다.

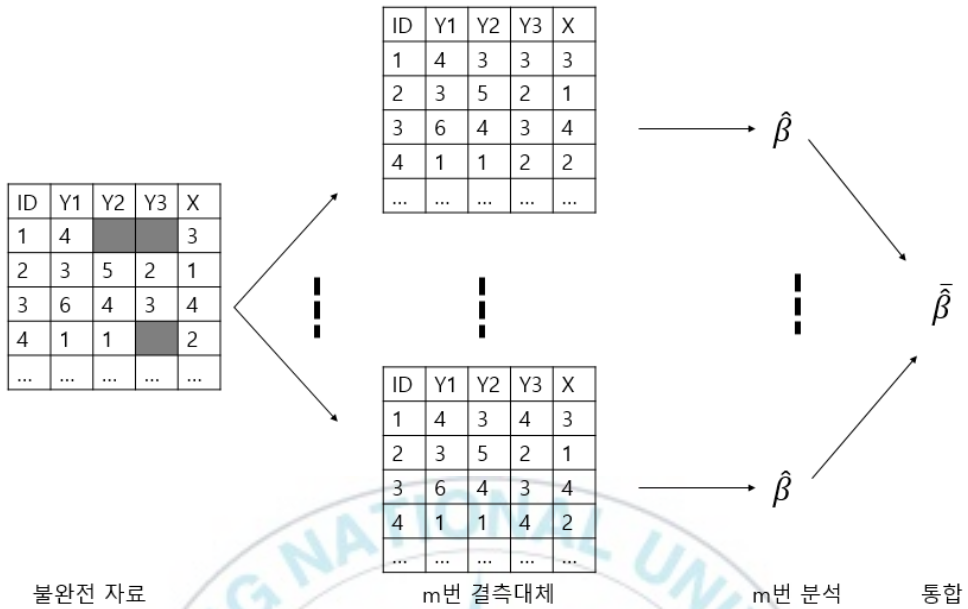


그림 2.2 다중대체 절차

대체단계에서 m번의 결측대체에 대한 대체방법으로 데이터증대(Data Augmentation)가 이용될 수 있다. 데이터증대는 마르코프 체인(Markov Chain Monte Carlo) 알고리즘의 종류 중 하나로, 사후분포로부터 모수를 재추정해 대체값을 구하는 방법이다. 이 방법은 다음의 절차를 따른다. (Allison, 2001)

- 1) 모수에 대한 초기 값을 설정해준다. 다변량 정규분포인 경우 평균과 분산을 고려해야한다. 완전사례분석 또는 EM알고리즘을 이용해 초기 값을 설정할 수 있다.
- 2) 현재의 평균과 분산을 이용해 결측치를 종속변수로, 관찰값을 독립

변수로하는 회귀모형을 적합시킨후 회귀계수를 얻는다.

3) 모든 결측치에대해 회귀계수를 이용해 예측값을 생성한다. 각 예측값에 잔차를 랜덤하게 뽑아 더해준다.

4) 완전한자료를 이용해 평균과 분산을 다시 계산한다.

5) 4)에서 다시 계산된 평균과 분산을 이용해 사후분포로부터 사후평균과 사후분산을 랜덤으로 뽑는다.

6) 값이 수렴할때까지 2~5과정을 반복수행한다.

예측값을 생성하는 방법은 다양한데, 그 중 R의 패키지 ‘MICE’에서 사용되는 방법 PMM(Predictive Mean Matching)을 소개한다. PMM은 다음 절차로 결측대체를 수행한다. (허명희, 2014; van Buuren, 2012)

사후분포로부터 모의생성된 모수를 이용해 예측값 $\hat{y}$ 을 구한다. $\hat{y}_i$ 를 관측치 y_i , where $i = 1, \dots, n$ 에 대한 예측값이라 하고, $\hat{y}_j$ 를 결측치 y_j , where $j = 1, \dots, m$ 에 대한 예측값이라 하자.

1) 한계값 η 을 정한 후, 결측치 y_j 에 대한 대체를 위해 $|\hat{y}_i - \hat{y}_j| < \eta$ 에 해당하는 기증자를 찾는다. 기증자들 중 랜덤하게 뽑은 값에 대응하는 y_i 를 대체값으로 정한다.

2) $|\hat{y}_i - \hat{y}_j| < \eta$ 에 해당하는 가장 가까운 후보자를 기증자로 뽑는다. “Nearest Neighbor Hot Deck”과 유사한 방법으로 찾을 수 있다.

3) $|\hat{y}_i - \hat{y}_j|$ 을 최소로 하는 후보자 d 를 여러명 뽑는다. d 는 3, 5, 10 이 주로 쓰인다.

4) $|\hat{y}_i - \hat{y}_j|$ 에 따른 확률로 기증자를 뽑는다.

또 다른 방법으로 회귀적 방법을 이용할 수 있다. 순서형 자료에 대해서 서열로지스틱 법을 이용하면 다음 절차를 따른다.

- 1) 서열로지스틱 회귀모형을 적합한 후, 회귀계수를 얻는다.
- 2) 회귀계수의 베이즈 사후분포로부터 β^* 가 생성된다.
- 3) 예측값이 모의 생성되어 결측치를 대체한다.

대체단계에서 결측대체 수 m 은 이상적 크면 클수록 좋으나, 그만큼 계산량이 많아지기에 적절한 대체 수 m 이 몇인지에 대해 생각해볼 필요가 있다. Rubin이 제안한 근사적 상대효율성은 결측률 λ , 결측대체수 m 에 대해 아래와 같이 계산될 수 있다(Rubin, 1987; Schafer, 1997).

$$Relative\ Efficiency = \left(1 + \frac{\lambda}{m}\right)^{-1} \quad (2.6)$$

표 2.2 결측률 λ 에 따른 상대효율성

λ					
m	0.1	0.3	0.5	0.7	0.9
2	0.95	0.87	0.80	0.74	0.69
3	0.97	0.91	0.86	0.81	0.77
5	0.98	0.94	0.91	0.88	0.85
10	0.99	0.97	0.95	0.93	0.92
∞	1.00	1.00	1.00	1.00	1.00

결측대체 수 m 에 대하여, Schafer는 $m=5\sim 10$ 이상 사용하는 것은 별다른 소득을 얻기 힘든 경향이 있다고 했다(Schafer, 1999). 그러나, Enders(2011)와 Graham, Olchowski, Gilreath(2007)은 시뮬레이션 연구로 $m=20$ 이 많은 경우 경험적으로 효율적이라고 했다(Enders, 2011; Graham, Olchowski, Gilreath, 2007).



제 3장 연구자료

3.1 자료 소개

질병관리본부 정책용역사업인 ‘뇌졸중 환자의 재활분야 장기적 기능 수준 관련 요인에 대한 10년 추적조사 연구(The Korean Stroke Cohort for functioning and rehabilitation, KOSCO)’ 데이터를 활용해 분석하였다. 해당 연구는 연구윤리심의위원회(Institutional Review Board, IRB)로부터 승인을 받아 2012년 8월 1일부터 6개 세부기관(제1~6세부기관)에서 연구 대상자 모집을 시작하였고, 3개 세부기관(제7~9세부기관)에서 2013년 6월 1일부터 연구대상자를 모집한 추적조사 연구이다. KOSCO연구는 다음의 조건을 만족하는 환자를 대상으로 모집조사 하였다(Chang, W. H. et al., 2015).

표 3.1 KOSCO 등록/제외 기준

KOSCO 등록/제외 기준
<u>등록기준</u>
19세 이상
급성 뇌졸중인 자로, 신경과 또는 신경외과 전문의 진단
급성 뇌졸중인 자로, 영상의학적 확진
급성 뇌졸중인 자로, 발병 7일 이내
연구 동의 함
<u>제외기준</u>
일과성 허혈발작(TIA)
기존의 뇌졸중 과거력
연구 자발적 동의 불가, 사망 등

본 논문에서 종속변수는 FMA(Fugl-Meyer Assessment)와 FAC(Functional Ambulatory Category)가 고려되었다. FMA는 Fugl-Meyer (1975)를 주축으로 뇌졸중 후 편마비 환자의 운동 기능을 평가하고자 개발한 척도이다. 초기에는 건측(Unaffected)을 평가한 후, 환측(Affected - infected or attacked, as by disease)을 평가하는 것으로 개발되었다. 본 연구에서는 양쪽(bilateral)마비 환자도 있어, 적용하기에는 아주 좋은 것은 아니다. 그러나 FMA는 운동기능 회복에 대해 섬세한 평가가 가능하다는 장점이 있어 사용되고 있다. FMA는 다섯가지 영역(관절의 움직임과 통증, 감각, 균형, 상지의 운동 기능, 하지의 운동 기능)으로 구성되어 있으며, 각 영역별 3등급으로 점수가 나뉜다. 각 영역의 검사는 3회 실시, 가장 높은 점수를 채택한다. 평가점수의 범위는 0~10

0점이고 해당 점수가 높을수록 운동기능 수준이 좋음을 나타낸다.

표 3.2 FMA 구성지표

FMA 구성지표
관절의 움직임과 통증
감각
균형
상지의 운동 기능
하지의 운동 기능

표 3.3 FMA 각 영역별 등급 점수

Characteristics	Score
과제 수행 불가	0
부분적으로 수행 가능	1
완전한 수행 가능	2

FAC(Functional Ambulatory Category)는 이동능력에 대한 평가이다. 독립적으로 이동 가능한지 또는 불가능한지에 대해 0~5점으로 분류한다. 해당 척도는 적어도 3미터 이상 걸을수 있는 환경에서 평가하며, 평행봉에서는 평가하지 않는다. 또한, 바닥의 상태에 따라 평가기준이 세부적으로 나뉜다. Level surface는 타일, 양탄자, 보도블럭과 같은 바닥을 의미하며, Non-level surface는 잔디, 자갈, 흙, 눈과 같은 바닥을 의미한다. 독립적으로 이동불가능한 경우(Dependent), 한 사람만의 도움만으로 평가하게 된다. 독립적으로 이동 가능(Independent)이란, 보

호자와 같은 다른 사람으로부터 관리를 받지 않는 상태 또는 물리적 도움을 받지 않는 상태를 의미한다. 평행봉을 제외하고는 보조기구를 이용해도 괜찮다. 해당 척도는 값이 클수록 정상임을 나타낸다.

표 3.4 FAC 구성지표

Characteristics	Score
Nonfunctional ambulation	0
Dependent for physical assistance - level I	1
Dependent for physical assistance - level II	2
Dependent for supervision	3
Independent, level surfaces only	4
Independent, all surfaces	5

3.2 결측 패턴

본 논문에서 사용된 데이터는 2018년도 2월 9일까지 입력 마감한 30개월 추적완료 데이터를 사용하였다. 고려된 시점은 사전평가(발병 7일, Baseline Function), 1년, 2년 이다. 아래 표는 시점별 결측패턴을 표로 나타낸 것으로, 결측이 발생한 경우 "M", 그렇지 않은 경우 "O"으로 나타냈다. FMA와 FAC는 다음과 같은 결측패턴을 보인다. 세 시점 모두 결측이 발생하지 않은 경우 "OOO"는 3,981명, 한 시점만 결측이 발생한 경우는 각각 1년 결측 발생 "OMO" 518명, 2년 결측 발생 "OOM" 7

12명이다. 두 시점 모두 결측이 발생한 경우 "OMM"는 1,999명이다.

표 3.5 FMA 결측패턴

FMATOTA_7D	FMATOTA_12M	FMATOTA_24M	빈도	퍼센트
O	O	O	3,981	55%
O	M	O	518	7%
O	O	M	712	10%
O	M	M	1,999	28%

표 3.6 FAC 결측패턴

FAC_7D	FAC_12M	FAC_24M	빈도	퍼센트
O	O	O	3,981	55%
O	M	O	518	7%
O	O	M	712	10%
O	M	M	1,999	28%

종속변수의 결측패턴이 MCAR인지에 대한 검정 방법으로 간단하게 결측여부와 연관성 여부를 살펴 볼 수 있다. 독립변수가 연속형인 경우 $t-test$, 범주형인 경우 χ^2-test 를 사용해 유의성 여부를 확인할 수 있다. 독립변수가 해당시점의 결측여부와 연관되어 있으면, MCAR로 보기 어렵다. 아래는 FMA와 FAC에 대한 시점별 검정 결과이다. FMA와 FAC의 1년 시점 결측은 나이가 많을 수록, 남자 대비 여자일 때 결측이 발생할 가능성이 높았으며, 뇌병변 위치가 양쪽일 때 결측이 발생할 가능성이 더 높았다. 그리고 Baseline의 FMA점수가 낮을수록(상태가 안

좋을수록) 결측일 발생할 가능성이 더 높았다. FAC 또한 비슷한 양상을 보였다. 따라서, 데이터의 결측구조는 MCAR라고 볼 수 없다.

표 3.7 FMA 1년 시점 결측 여부와 연관성

변수	관측 (N=4,693)	결측 (N=2,517)	P-value
AGE	64.3 ± 12.8	67.6 ± 13.7	<.001*
SC_SEX			0.002*
- 남자	2,851 (66.5%)	1,436 (33.5%)	
- 여자	1,842 (63.0%)	1,081 (37.0%)	
XDX_DIAG			0.947
- 뇌경색	4,078 (65.1%)	2,185 (34.9%)	
- 뇌출혈	615 (64.9%)	332 (35.1%)	
LOC			0.009*
- 양쪽	315 (59.1%)	218 (40.9%)	
- 오른쪽	2,125 (65.3%)	1,131 (34.7%)	
- 왼쪽	2,253 (65.9%)	1,168 (34.1%)	
FMA_TOTA_7D	78.5 ± 32.4	68.0 ± 37.9	<.001*

표 3.8 FMA 2년 시점 결측 여부와 연관성

변수	관측 (N=4,499)	결측 (N=2,711)	P-value
AGE	64.0 ± 12.6	67.7 ± 13.9	<.001*
SC_SEX			0.002*
- 남자	2,746 (64.1%)	1,541 (35.9%)	
- 여자	1,753 (60.0%)	1,170 (37.0%)	
XDX_DIAG			0.947
- 뇌경색	3,909 (62.4%)	2,354 (37.6%)	
- 뇌출혈	590 (62.3%)	357 (37.7%)	
LOC			0.003*
- 양쪽	296 (55.5%)	237 (44.5%)	
- 오른쪽	2,048 (62.9%)	1,208 (37.1%)	
- 왼쪽	2,155 (63.0%)	1,266 (37.0%)	
FMATOTA_7D	79.4 ± 31.8	67.3 ± 38.1	<.001*

표 3.9 FAC 1년 시점 결측 여부와 연관성

변수	관측 (N=4,693)	결측 (N=2,517)	P-value
AGE	64.3 ± 12.8	67.6 ± 13.7	<.001*
SC_SEX			0.002*
- 남자	2,851 (66.5%)	1,436 (33.5%)	
- 여자	1,842 (63.0%)	1,081 (37.0%)	
XDX_DIAG			0.947
- 뇌경색	4,078 (65.1%)	2,185 (34.9%)	
- 뇌출혈	615 (64.9%)	332 (35.1%)	
LOC			0.009*
- 양쪽	315 (59.1%)	218 (40.9%)	
- 오른쪽	2,125 (65.3%)	1,131 (34.7%)	
- 왼쪽	2,253 (65.9%)	1,168 (34.1%)	
FAC_7D	2.9 ± 2.4	2.4 ± 2.0	<.001*

표 3.10 FAC 2년 시점 결측 여부와 연관성

변수	관측 (N=4,499)	결측 (N=2,711)	P-value
AGE	64.0 ± 12.6	67.7 ± 13.9	<.001*
SC_SEX			0.002*
- 남자	2,746 (64.1%)	1,541 (35.9%)	
- 여자	1,753 (60.0%)	1,170 (37.0%)	
XDX_DIAG			0.947
- 뇌경색	3,909 (62.4%)	2,354 (37.6%)	
- 뇌출혈	590 (62.3%)	357 (37.7%)	
LOC			0.003*
- 양쪽	296 (55.5%)	237 (44.5%)	
- 오른쪽	2,048 (62.9%)	1,208 (37.1%)	
- 왼쪽	2,155 (63.0%)	1,266 (37.0%)	
FAC_7D	3.0 ± 1.9	2.3 ± 2.0	<.001*

데이터의 결측구조가 MCAR이 아니면 MAR이거나 NMAR이다. 패널 자료에서는 종속변수의 상관성 및 종속변수와 독립변수의 연관성으로 인해, 결측구조가 MAR인지 NMAR인지 구분하기 명확하지 않다. 독립변수는 종종 관찰된 종속변수에 영향을 받을 뿐 아니라, 관찰되지 않은 종속변수에 영향을 받기 때문이다(Laird, 1988).

3.3 실제자료 분석

해당 자료에서 연속형 변수인 FMA, 순서형 변수인 FAC 각각에 대해 GEE분석을 수행하였다. 변수는 각각 ‘나이=AGE’, ‘성별=SC_SE

X' , '진단명=XDX_DIAG' , '병변위치=LOC' , '발병후 추적 기간=TIME' 을 나타낸다. (표에서 표기는 다음과 같이 하였다. ' x_1 =AGE' , ' x_2 =SC_SEX(여자)' , ' x_3 =XDX_DIAG(뇌출혈)' , ' x_4 =LOC(오른쪽)' , ' x_5 =LOC(왼쪽)' , ' x_6 =TIME') 고려된 결측치 처리 방법론은 'OBS=관찰자료분석' , 'COMP=완전사례분석' , 'RM=행평균대체' , 'RMD=행중위수대체' , 'HD=핫덱대체' , 'LOCF=이월대체' , 'REG=회귀대체' , 'MI=다중대체' 이다. 가상관행렬은 교환가능(Exchangeable)으로 지정했다. 분석에 쓰인 통계 툴은 'R version 3.4.2' 이며 , GEE분석을 위한 라이브러리 'gee version 4.13-19' 와 Multiple Imputation을 위한 라이브러리 'MICE' 를 사용하여 분석했다.

3.3.1 연속형 자료

적어도 모든 군집에 가능한 자료 전체를 사용하여 분석한 결과와 전혀 결측치가 없는 군집들의 자료만을 사용하여 분석한 결과를 비교해야 한다. 만약 두 결과가 크게 다를 경우에는 결측치의 원인을 조사해야 한다 (Agresti, 2007).

연속형 자료의 분석 결과인 표3.11의 회귀계수를 살펴보면, OBS와 COMP간 결과가 크게 달라 보인다. 그리고 표3.7, 표3.8에서 종속변수와 결측여부의 연관성에 관해 살펴보면, 종속변수의 결측이 독립변수에 영향을 받고 있는 것을 알 수 있다. 또한 표 3.13의 P-value를 살펴보

면, x_5 에서 결측처리 방법론에 따라 유의성이 달라지는 것이 보인다. 따라서 결측이 분석결과에 영향을 줄 것이라 판단되며, 그에 따른 결측처리 방법론 선택이 중요할 것으로 보인다.

표 3.11 FMA 실제자료에서 회귀계수

Variables	Observed		Imputed					
	OBS	COMP	RM	RMD	HD	LOCF	REG	MI
x_1	-0.39	-0.29	-0.49	-0.49	-0.49	-0.49	-0.38	-0.43
x_2	-3.69	-3.04	-4.49	-4.49	-4.41	-4.50	-3.58	-3.95
x_3	-18.39	-18.14	-19.12	-19.12	-19.06	-19.13	-17.44	-16.64
x_4	-0.38	-0.71	1.03	1.03	1.07	1.03	-0.75	-0.16
x_5	0.30	-0.51	1.79	1.79	1.80	1.80	-0.02	0.70
x_6	0.53	0.34	0.22	0.22	0.22	0.23	0.50	0.40

표 3.12 FMA 실제자료에서 Robust S.E.

Variables	Observed		Imputed					
	OBS	COMP	RM	RMD	HD	LOCF	REG	MI
x_1	0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.03
x_2	0.49	0.52	0.46	0.46	0.46	0.46	0.37	0.81
x_3	0.83	0.95	0.75	0.75	0.76	0.76	0.63	1.16
x_4	0.95	0.98	0.90	0.90	0.90	0.90	0.69	1.47
x_5	0.95	0.98	0.89	0.89	0.90	0.89	0.69	1.42
x_6	0.02	0.03	0.02	0.02	0.02	0.02	0.02	0.02

표 3.13 FMA 실제자료에서 P-value

Variables	Observed		Imputed					
	OBS	COMP	RM	RMD	HD	LOCF	REG	MI
x_1	<.01*	<.01*	<.01*	<.01*	<.01*	<.01*	<.01*	<.01*
x_2	<.01*	<.01*	<.01*	<.01*	<.01*	<.01*	<.01*	<.01*
x_3	<.01*	<.01*	<.01*	<.01*	<.01*	<.01*	<.01*	<.01*
x_4	0.69	0.47	0.25	0.25	0.23	0.25	0.28	0.91
x_5	0.76	0.60	0.05*	0.05*	0.04*	0.04*	0.98	0.62
x_6	<.01*	<.01*	<.01*	<.01*	<.01*	<.01*	<.01*	<.01*

3.3.2 순서형 자료

순서형 자료의 분석 결과인 표 3.14의 회귀계수를 보면 연속형 자료의 분석 결과만큼 OBS와 COMP간의 차이가 크진 않았다. 그러나 결측 대체 방법론 별 분석결과는 다소 차이가 나는 것으로 보인다. 그리고 표 3.9, 표3.10에서 종속변수와 결측여부의 연관성에 관해 살펴보면 종속변수의 결측이 독립변수에 영향을 받고 있는 것을 알 수 있다. 표 3.16의 P-value는 연속형 자료 분석 결과와 달리 결측처리 방법론에 따른 유의성이 달라지진 않았다. 그러나 결측대체 방법론에 따른 분석 결과가 차이가 난 만큼 결측이 분석 결과에 어느정도 영향을 줄 것이라 판단되며, 순서형 자료 또한 결측처리 방법론 선택이 중요할 것으로 보인다.

표 3.14 FAC 실제자료에서 회귀계수

Variables	Observed		Imputed			
	OBS	COMP	HD	LOCF	REG	MI
x_1	0.04	0.04	0.04	0.04	0.03	0.05
x_2	0.38	0.38	0.36	0.36	0.34	0.40
x_3	1.24	1.21	1.10	1.12	1.19	1.10
x_4	-0.25	-0.19	-0.31	-0.31	-0.15	-0.22
x_5	-0.3	-0.24	-0.35	-0.36	-0.20	-0.26
x_6	-0.08	-0.08	-0.04	-0.04	-0.11	-0.07

표 3.15 FAC 실제자료에서 Robust S.E.

Variables	Observed		Imputed			
	OBS	COMP	HD	LOCF	REG	MI
x_1	0.04	0.04	0.04	0.04	0.03	0.05
x_2	0.38	0.38	0.36	0.36	0.34	0.40
x_3	1.24	1.21	1.10	1.12	1.19	1.10
x_4	-0.25	-0.19	-0.31	-0.31	-0.15	-0.22
x_5	-0.3	-0.24	-0.35	-0.36	-0.20	-0.26
x_6	-0.08	-0.08	-0.04	-0.04	-0.11	-0.07

표 3.16 FAC 실제자료에서 P-value

Variables	Observed		Imputed			
	OBS	COMP	HD	LOCF	REG	MI
x_1	<.01*	<.01*	<.01*	<.01*	<.01*	<.01*
x_2	<.01*	<.01*	<.01*	<.01*	<.01*	<.01*
x_3	<.01*	<.01*	<.01*	<.01*	<.01*	<.01*
x_4	<.01*	0.05*	<.01*	<.01*	0.03*	0.01*
x_5	<.01*	<.01*	<.01*	<.01*	<.01*	0.01*
x_6	<.01*	<.01*	<.01*	<.01*	<.01*	<.01*



제 4장 모의실험

4.1 모의실험 설계

완전하게 응답된 값들을 모집단으로 가정하고, 이 모집단에서 임의로 표본을 뽑은 후 선택된 표본에 결측을 발생시켰다. 단순임의추출로 표본 추출하였으며, 각 방법별 추정값을 500회 반복하여 구하였다.

결측 생성은 결측 발생확률과 대체군 형성 변수들 간의 연관성에 대한 로짓모형 고려하였다. 실제 자료 중 응답된 자료에 근거하여 모의실험을 실시함으로써 실제 자료의 상황과 가능한 비슷한 조건 하에서 모의실험을 진행하였다.

표 4.1 종속변수 결측 생성 모형

종속변수 결측 생성 모형
1) $\logit[P(R_{i2})] = \beta_0 + \beta_1 y_{i1} + \beta_2 y_{i2}$
2) $\logit[P(R_{i3})] = \beta_0 + \beta_1 y_{i1} + \beta_2 y_{i3}$

결측을 포함한 전체자료에 대해 종속변수를 결측 여부로, 독립변수는 다음을 포함하는 로짓모형을 설정하여 결측 발생 확률을 계산하였다. 완전히 응답된 자료에 대해, 이 확률에 근거하여 결측 여부를 임의 추출한 후 절편을 조정하여 설정한 결측률에 따라 결측이 생기도록 자료를 생성

하였다. 이와 같이 생성된 결측 자료는 원래 자료의 결측과 유사한 특색을 가질 것으로 기대된다(강현철, 2010).

β_0 는 β_1 과 β_2 가 주어졌을때 표본추출로 인해 데이터가 달라짐에 따라 목표 결측률 $P(Missing)$ 에 맞춰 결정되므로 수치적으로 해를 찾았다.

$P(Missing) = \frac{1}{n} \sum_{i=1}^n \frac{1}{1 + \exp(-\beta_0 - x_i^T \beta)}$ 이 되도록 수치적 해 β_0 를 구하였다.

FMA에대한 결측구조의 설정은 종속변수 결측 생성모형인 로짓 모형에 다른 변수와 연관성을 고려하여 각각 설정해주었다. 고려된 결측구조는 MCAR, MAR sNMAR(simple NMAR), cNMAR(complicated NMAR)이다.



표 4.2 FMA 결측 생성 모형별 결측 형태에 따른 $\beta_0, \beta_1, \beta_2$

결측구조	결측률	$\beta = (\beta_0, \beta_1, \beta_2)$
1) $\text{logit}[P(R_{i2})] = \beta_0 + \beta_1 y_{i1} + \beta_2 y_{i2}$		
MCAR	10%	$\beta = (-2.197, 0.000, 0.000)$
MAR	10%	$\beta = (-1.586, -0.008, 0.000)$
sNMAR	10%	$\beta = (-1.512, 0.000, -0.008)$
cNMAR	10%	$\beta = (-0.936, -0.008, -0.008)$
MCAR	30%	$\beta = (-0.847, 0.000, 0.000)$
MAR	30%	$\beta = (-0.220, -0.008, 0.000)$
sNMAR	30%	$\beta = (-0.152, 0.000, -0.008)$
cNMAR	30%	$\beta = (0.466, -0.008, -0.008)$
MCAR	50%	$\beta = (0.000, 0.000, 0.000)$
MAR	50%	$\beta = (0.640, -0.008, 0.000)$
sNMAR	50%	$\beta = (0.704, 0.000, -0.008)$
cNMAR	50%	$\beta = (1.351, -0.008, -0.008)$

표 4.3 FMA 결측 생성 모형별 결측 형태에 따른 $\beta_0, \beta_1, \beta_2$

결측구조	결측률	$\beta = (\beta_0, \beta_1, \beta_2)$
2) $\text{logit}[P(R_{i3})] = \beta_0 + \beta_1 y_{i1} + \beta_2 y_{i3}$		
MCAR	10%	$\beta = (-2.197, 0.000, 0.000)$
MAR	10%	$\beta = (-1.442, -0.010, 0.000)$
sNMAR	10%	$\beta = (-1.345, 0.000, -0.010)$
cNMAR	10%	$\beta = (-0.644, -0.010, -0.010)$
MCAR	30%	$\beta = (-0.847, 0.000, 0.000)$
MAR	30%	$\beta = (-0.067, -0.010, 0.000)$
sNMAR	30%	$\beta = (0.022, 0.000, -0.010)$
cNMAR	30%	$\beta = (0.792, -0.010, -0.010)$
MCAR	50%	$\beta = (0.000, 0.000, 0.000)$
MAR	50%	$\beta = (0.802, -0.010, 0.000)$
sNMAR	50%	$\beta = (0.883, 0.000, -0.010)$
cNMAR	50%	$\beta = (1.697, -0.010, -0.010)$

표 4.4 FAC 결측 생성 모형별 결측 형태에 따른 $\beta_0, \beta_1, \beta_2$

결측구조	결측률	$\beta = (\beta_0, \beta_1, \beta_2)$
1) $\text{logit}[P(R_{i2})] = \beta_0 + \beta_1 y_{i1} + \beta_2 y_{i2}$		
MCAR	10%	$\beta = (-2.197, 0.000, 0.000)$
MAR	10%	$\beta = (-1.789, -0.145, 0.000)$
sNMAR	10%	$\beta = (-1.598, 0.000, -0.145)$
cNMAR	10%	$\beta = (-1.218, -0.145, -0.145)$
MCAR	30%	$\beta = (-0.847, 0.000, 0.000)$
MAR	30%	$\beta = (-0.423, -0.145, 0.000)$
sNMAR	30%	$\beta = (0.238, 0.000, -0.145)$
cNMAR	30%	$\beta = (0.178, -0.145, -0.145)$
MCAR	50%	$\beta = (0.000, 0.000, 0.000)$
MAR	50%	$\beta = (0.439, -0.145, 0.000)$
sNMAR	50%	$\beta = (0.619, 0.000, -0.145)$
cNMAR	50%	$\beta = (1.062, -0.145, -0.145)$

표 4.5 FAC 결측 생성 모형별 결측 형태에 따른 $\beta_0, \beta_1, \beta_2$

결측구조	결측률	$\beta = (\beta_0, \beta_1, \beta_2)$
2) $\text{logit}[P(R_{i3})] = \beta_0 + \beta_1 y_{i1} + \beta_2 y_{i3}$		
MCAR	10%	$\beta = (-2.197, 0.000, 0.000)$
MAR	10%	$\beta = (-1.695, -0.182, 0.000)$
sNMAR	10%	$\beta = (-1.447, 0.000, -0.182)$
cNMAR	10%	$\beta = (-0.986, -0.182, -0.182)$
MCAR	30%	$\beta = (-0.847, 0.000, 0.000)$
MAR	30%	$\beta = (-0.319, -0.182, 0.000)$
sNMAR	30%	$\beta = (-0.079, 0.000, -0.182)$
cNMAR	30%	$\beta = (0.438, 0.000, 0.000)$
MCAR	50%	$\beta = (0.000, -0.182, 0.000)$
MAR	50%	$\beta = (0.552, 0.000, -0.182)$
sNMAR	50%	$\beta = (0.783, -0.182, 0.000)$
cNMAR	50%	$\beta = (1.343, -0.182, -0.182)$

결측 처리 방법별로 GEE모형의 모수 추정값 $\hat{\beta}_s$ 을 구한 모의실험 결과에 대해 효율성 평가 기준으로 편의(Bias)와 상대편의(Relative Bias)를 이용했다. 반복수 S=500회에 대해 편의는 다음과 같다.

$$Bias = \bar{\hat{\beta}} - \beta, \text{ where } \bar{\hat{\beta}} = \sum_{s=1}^S \hat{\beta}_s / S \quad (4.1)$$

상대편의는 편의의 상대적 수치로, 1에 가까울수록 모수 추정에 대한 추정량의 편의가 작다고 판단한다.

$$Relative \ Bias = \hat{\beta} / \beta \quad (4.2)$$

4.2 모의실험 결과

4.2.1 연속형 자료

아래는 연속형 자료의 모의실험 결과이다. 결측구조 MCAR일때, 결측률이 높아질수록 HD, RM, RMD, LOCF에서 상대편의가 1과 비교했을 때 크게 차이 나는 것으로 보였다. 그 외, 다른 결측처리 방법론은 MCAR에서 편의 정도가 크지 않은 것으로 보였다.

결측구조 MAR 일 때 결측률이 높을수록 결측처리 방법론에 따라 상대편의가 크게 차이 났었는데, MCAR에서와 마찬가지로 HD, RM, RMD, LOCF에서 상대편의가 1과 비교했을 때 크게 차이 남을 알 수 있었다. MAR구조일 때 모수추정에 대한 편의가 작은 결측처리 방법은 COMP와 MI인 것으로 보인다. COMP의 경우 해당 데이터 특성상 종속변수에만 결측치가 있어, MCAR 뿐 아니라 MAR일때도 회귀계수의 편의 정도가 크지 않은 것으로 보인다. 이는 Allison (2001), Schafer and Graham (2002)의 연구와 비슷한 결과를 나타내는 것으로 보인다.

sNMAR, cNMAR에서는 COMP와 MI가 상대편의가 1과 비교했을 때 차이가 크지 않은 것으로 보였다. MAR에서 COMP의 회귀계수 편의 정도가 크지 않다는 것은 Allison (2001), Schafer and Graham (2002)의 연구로 알 수 있었지만, NMAR에서 회귀계수의 편의 정도가 크지 않

다는 것은 기존 연구결과와 달라 향후 조사가 필요해 보인다. 라이브러리 ‘MICE’의 MI또한 MAR 가정을 만족해야 잘 작동함에도 불구하고 NMAR에서 편의 정도가 크지 않아 향후 조사가 필요해 보인다.

그 외, 더미변수에서는 결측처리 방법론 별로 편의가 작은 결과가 뚜렷하게 나타나지 않았는데, 이는 Harrell (2015)이 언급한대로, 더미변수의 경우 결측대체가 종속변수의 정보를 손실시켜 회귀계수의 편의를 야기하는 것으로 보인다.



표 4.6 FMA 결측률 10%일 때, 회귀계수

Mechanism	Variables	No miss	Observed		Imputed					
		FULL	OBS	COMP	RM	RMD	HD	LOCF	REG	MI
MCAR	x_1	-0.29	-0.28	-0.28	-0.27	-0.27	-0.27	-0.27	-0.28	-0.27
	x_2	-3.04	-3.17	-3.20	-3.13	-3.13	-3.13	-3.13	-3.19	-3.09
	x_3	-18.14	-18.23	-18.13	-18.39	-18.39	-18.39	-18.39	-18.00	-17.61
	x_4	-0.71	-0.82	-0.76	-1.03	-1.03	-1.03	-1.03	-0.80	-0.76
	x_5	-0.51	-0.45	-0.32	-0.81	-0.81	-0.81	-0.81	-0.43	-0.47
	x_6	0.34	0.35	0.35	0.31	0.31	0.31	0.31	0.35	0.35
MAR	x_1	-0.29	-0.27	-0.27	-0.27	-0.27	-0.27	-0.27	-0.27	-0.27
	x_2	-3.04	-3.11	-3.10	-3.14	-3.14	-3.14	-3.14	-3.13	-3.07
	x_3	-18.14	-18.02	-17.55	-18.74	-18.74	-18.74	-18.72	-17.73	-17.54
	x_4	-0.71	-0.69	-0.56	-1.00	-1.00	-1.00	-1.01	-0.66	-0.68
	x_5	-0.51	-0.34	-0.12	-0.82	-0.82	-0.82	-0.82	-0.30	-0.42
	x_6	0.34	0.38	0.33	0.30	0.30	0.30	0.30	0.37	0.36
sNMA R	x_1	-0.29	-0.26	-0.26	-0.27	-0.27	-0.27	-0.27	-0.27	-0.26
	x_2	-3.04	-3.01	-2.98	-3.09	-3.09	-3.09	-3.09	-3.03	-2.99
	x_3	-18.14	-17.84	-17.52	-18.34	-18.34	-18.34	-18.35	-17.56	-17.32
	x_4	-0.71	-0.62	-0.44	-0.99	-0.99	-0.99	-1.00	-0.60	-0.64
	x_5	-0.51	-0.31	-0.06	-0.81	-0.81	-0.81	-0.81	-0.28	-0.40
	x_6	0.34	0.38	0.35	0.32	0.32	0.32	0.32	0.38	0.37
cNMA R	x_1	-0.29	-0.25	-0.24	-0.26	-0.26	-0.26	-0.26	-0.25	-0.25
	x_2	-3.04	-2.94	-2.86	-3.07	-3.07	-3.08	-3.07	-2.96	-2.99
	x_3	-18.14	-17.28	-16.34	-18.71	-18.71	-18.71	-18.69	-16.89	-17.03
	x_4	-0.71	-0.27	0.20	-0.97	-0.97	-0.96	-0.97	-0.22	-0.46
	x_5	-0.51	0.07	0.64	-0.81	-0.81	-0.81	-0.81	0.13	-0.20
	x_6	0.34	0.42	0.33	0.30	0.30	0.30	0.30	0.42	0.38

Missing Rate=10%

표 4.7 FMA 결측률 10%일 때, 편의

Mechanism	Variables	Observed		Imputed					
		OBS	COMP	RM	RMD	HD	LOCF	REG	MI
MCAR	x_1	0.01	0.01	0.02	0.02	0.02	0.02	0.01	0.02
	x_2	-0.13	-0.15	-0.09	-0.09	-0.09	-0.09	-0.14	-0.05
	x_3	-0.09	0.01	-0.25	-0.25	-0.25	-0.25	0.14	0.53
	x_4	-0.11	-0.05	-0.32	-0.32	-0.32	-0.32	-0.09	-0.05
	x_5	0.06	0.19	-0.31	-0.31	-0.31	-0.31	0.08	0.04
	x_6	0.01	0.00	-0.03	-0.03	-0.03	-0.03	0.01	0.01
MAR	x_1	0.02	0.01	0.02	0.02	0.02	0.02	0.01	0.02
	x_2	-0.07	-0.06	-0.10	-0.10	-0.10	-0.10	-0.08	-0.03
	x_3	0.13	0.59	-0.60	-0.60	-0.60	-0.58	0.41	0.60
	x_4	0.02	0.15	-0.29	-0.29	-0.29	-0.30	0.05	0.03
	x_5	0.17	0.39	-0.31	-0.31	-0.31	-0.31	0.20	0.09
	x_6	0.04	-0.01	-0.05	-0.05	-0.05	-0.04	0.03	0.02
sNMA R	x_1	0.02	0.03	0.02	0.02	0.02	0.02	0.02	0.03
	x_2	0.03	0.07	-0.05	-0.05	-0.05	-0.05	0.02	0.05
	x_3	0.30	0.62	-0.20	-0.20	-0.20	-0.21	0.58	0.82
	x_4	0.09	0.27	-0.28	-0.28	-0.28	-0.29	0.11	0.07
	x_5	0.20	0.44	-0.30	-0.30	-0.30	-0.30	0.23	0.11
	x_6	0.04	0.01	-0.02	-0.02	-0.02	-0.02	0.04	0.03
cNMA R	x_1	0.04	0.05	0.02	0.02	0.02	0.02	0.03	0.03
	x_2	0.10	0.19	-0.03	-0.03	-0.03	-0.03	0.08	0.05
	x_3	0.86	1.81	-0.57	-0.57	-0.56	-0.55	1.25	1.12
	x_4	0.44	0.90	-0.26	-0.26	-0.26	-0.26	0.49	0.25
	x_5	0.58	1.15	-0.30	-0.30	-0.30	-0.30	0.63	0.31
	x_6	0.08	-0.01	-0.04	-0.04	-0.04	-0.04	0.08	0.04

Missing Rate=10%

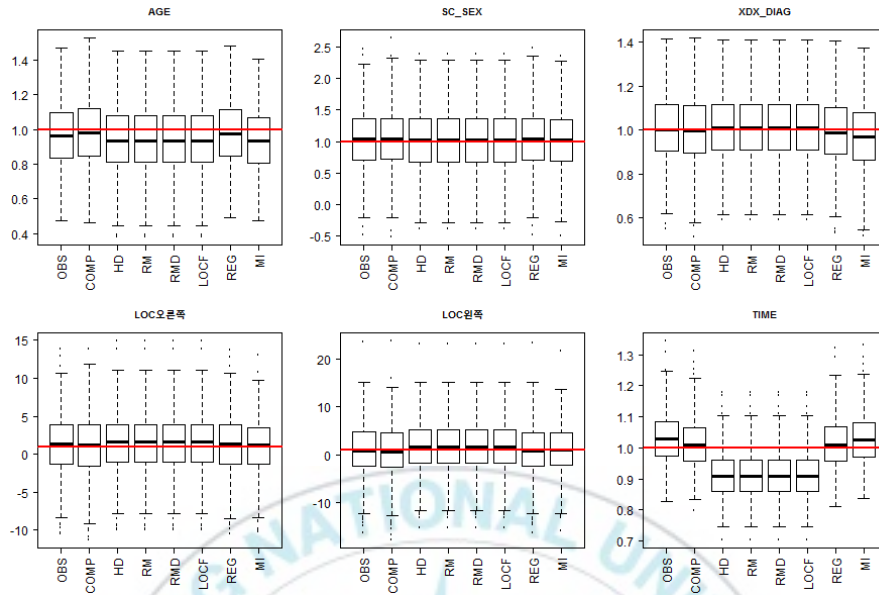


그림 4.1 FMA 결측률 10%일 때, MCAR에서의 상대편의

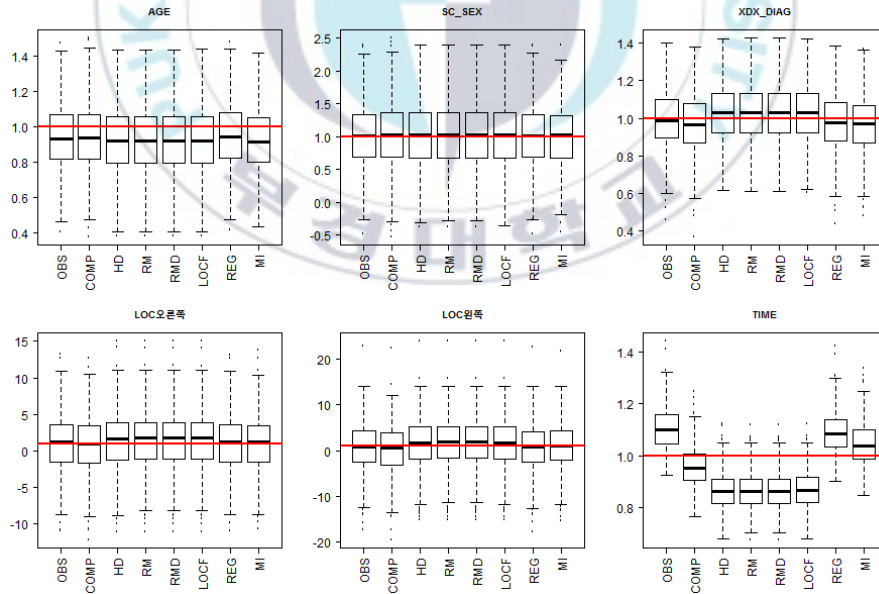


그림 4.2 FMA 결측률 10%일 때, MAR에서의 상대편의

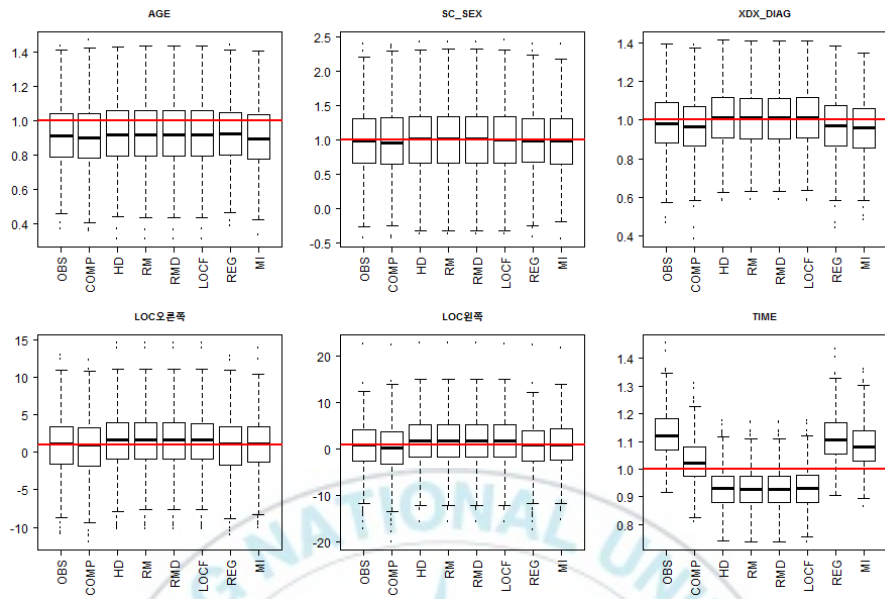


그림 4.3 FMA 결측률 10%일 때, sNMAR에서의 상대편의

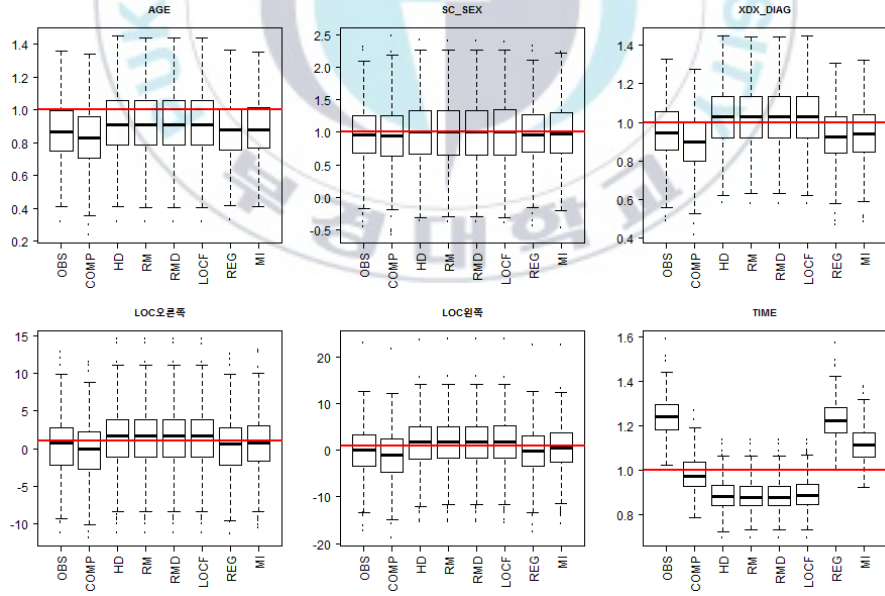


그림 4.4 FMA 결측률 10%일 때, cNMAR에서의 상대편의

표 4.8 FMA 결측률 30%일 때, 회귀계수

Mechanism	Variables	No miss	Observed		Imputed					
		FULL	OBS	COMP	RM	RMD	HD	LOCF	REG	MI
MCAR	x_1	-0.29	-0.25	-0.27	-0.24	-0.24	-0.24	-0.24	-0.27	-0.24
	x_2	-3.04	-3.15	-3.27	-2.93	-2.93	-2.93	-2.93	-3.21	-2.95
	x_3	-18.14	-18.07	-17.26	-19.12	-19.12	-19.12	-19.12	-17.21	-16.52
	x_4	-0.71	-0.51	-0.49	-0.73	-0.73	-0.73	-0.73	-0.40	-0.22
	x_5	-0.51	-0.10	0.23	-0.75	-0.75	-0.75	-0.75	0.06	-0.07
	x_6	0.34	0.38	0.35	0.24	0.24	0.24	0.24	0.36	0.37
MAR	x_1	-0.29	-0.24	-0.25	-0.23	-0.23	-0.23	-0.24	-0.25	-0.23
	x_2	-3.04	-2.80	-2.60	-2.98	-2.98	-2.98	-2.98	-2.81	-2.83
	x_3	-18.14	-17.56	-15.82	-19.75	-19.75	-19.76	-19.70	-16.43	-16.40
	x_4	-0.71	-0.11	0.31	-0.71	-0.71	-0.71	-0.69	0.07	-0.02
	x_5	-0.51	0.41	1.29	-0.82	-0.82	-0.82	-0.79	0.66	0.17
	x_6	0.34	0.46	0.29	0.20	0.20	0.20	0.21	0.44	0.38
sNMA R	x_1	-0.29	-0.24	-0.23	-0.24	-0.24	-0.24	-0.24	-0.25	-0.24
	x_2	-3.04	-2.68	-2.46	-2.84	-2.84	-2.84	-2.85	-2.66	-2.66
	x_3	-18.14	-17.33	-15.65	-19.33	-19.33	-19.34	-19.29	-16.27	-16.17
	x_4	-0.71	-0.32	0.28	-0.86	-0.86	-0.86	-0.83	-0.16	-0.32
	x_5	-0.51	0.23	1.02	-0.85	-0.85	-0.86	-0.84	0.46	0.06
	x_6	0.34	0.45	0.35	0.24	0.24	0.24	0.25	0.42	0.40
cNMA R	x_1	-0.29	-0.22	-0.19	-0.24	-0.24	-0.24	-0.24	-0.22	-0.22
	x_2	-3.04	-2.69	-2.10	-3.18	-3.18	-3.18	-3.09	-2.65	-2.72
	x_3	-18.14	-16.12	-12.16	-19.84	-19.84	-19.85	-19.80	-14.64	-15.36
	x_4	-0.71	-2.18	-3.69	-1.07	-1.07	-1.07	-1.06	-2.39	-1.50
	x_5	-0.51	-1.94	-3.19	-1.22	-1.22	-1.22	-1.23	-2.11	-1.46
	x_6	0.34	0.51	0.28	0.22	0.22	0.22	0.24	0.49	0.42

Missing Rate=30%

표 4.9 FMA 결측률 30%일 때, 편의

Mechanism	Variables	Observed		Imputed					
		OBS	COMP	RM	RMD	HD	LOCF	REG	MI
MCAR	x_1	0.03	0.02	0.05	0.05	0.05	0.05	0.02	0.05
	x_2	-0.11	-0.23	0.11	0.11	0.11	0.11	-0.16	0.09
	x_3	0.07	0.88	-0.98	-0.98	-0.98	-0.98	0.93	1.63
	x_4	0.19	0.22	-0.02	-0.02	-0.02	-0.02	0.31	0.48
	x_5	0.41	0.73	-0.24	-0.24	-0.24	-0.24	0.57	0.44
	x_6	0.04	0.01	-0.10	-0.10	-0.10	-0.10	0.02	0.03
MAR	x_1	0.05	0.04	0.05	0.05	0.05	0.05	0.04	0.05
	x_2	0.24	0.44	0.06	0.06	0.07	0.06	0.23	0.21
	x_3	0.58	2.32	-1.61	-1.61	-1.62	-1.56	1.71	1.74
	x_4	0.60	1.01	0.00	0.00	0.00	0.02	0.78	0.68
	x_5	0.92	1.80	-0.31	-0.31	-0.32	-0.28	1.17	0.68
	x_6	0.12	-0.05	-0.14	-0.14	-0.14	-0.13	0.10	0.03
sNMA R	x_1	0.04	0.06	0.05	0.05	0.05	0.05	0.04	0.05
	x_2	0.37	0.58	0.20	0.20	0.20	0.19	0.38	0.38
	x_3	0.81	2.50	-1.19	-1.19	-1.20	-1.15	1.87	1.97
	x_4	0.39	0.99	-0.15	-0.15	-0.16	-0.12	0.55	0.38
	x_5	0.74	1.53	-0.34	-0.34	-0.35	-0.33	0.97	0.57
	x_6	0.11	0.01	-0.10	-0.10	-0.10	-0.10	0.08	0.06
cNMA R	x_1	0.07	0.10	0.05	0.05	0.05	0.05	0.07	0.07
	x_2	0.35	0.94	-0.14	-0.14	-0.14	-0.05	0.39	0.32
	x_3	2.02	5.98	-1.70	-1.70	-1.71	-1.66	3.51	2.78
	x_4	-1.47	-2.98	-0.36	-0.36	-0.36	-0.35	-1.68	-0.79
	x_5	-1.43	-2.68	-0.71	-0.71	-0.71	-0.72	-1.60	-0.95
	x_6	0.17	-0.06	-0.12	-0.12	-0.12	-0.10	0.15	0.08

Missing Rate=30%

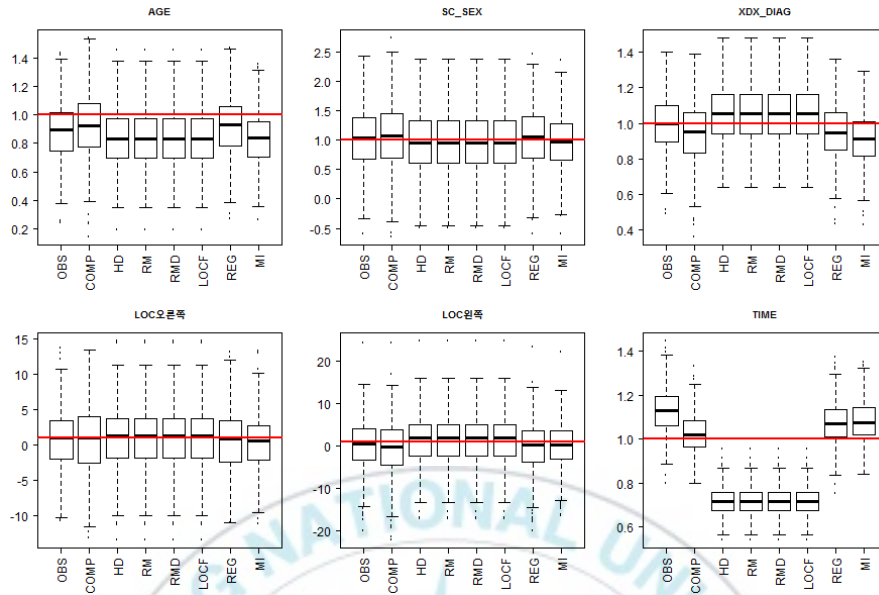


그림 4.5 FMA 결측률 30%일 때, MCAR에서의 상대편의

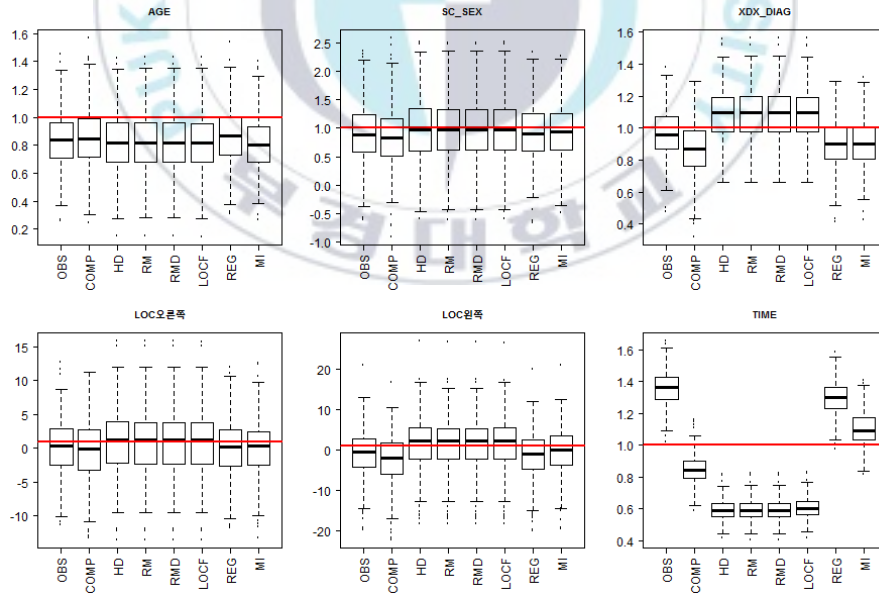


그림 4.6 FMA 결측률 30%일 때, MAR에서의 상대편의

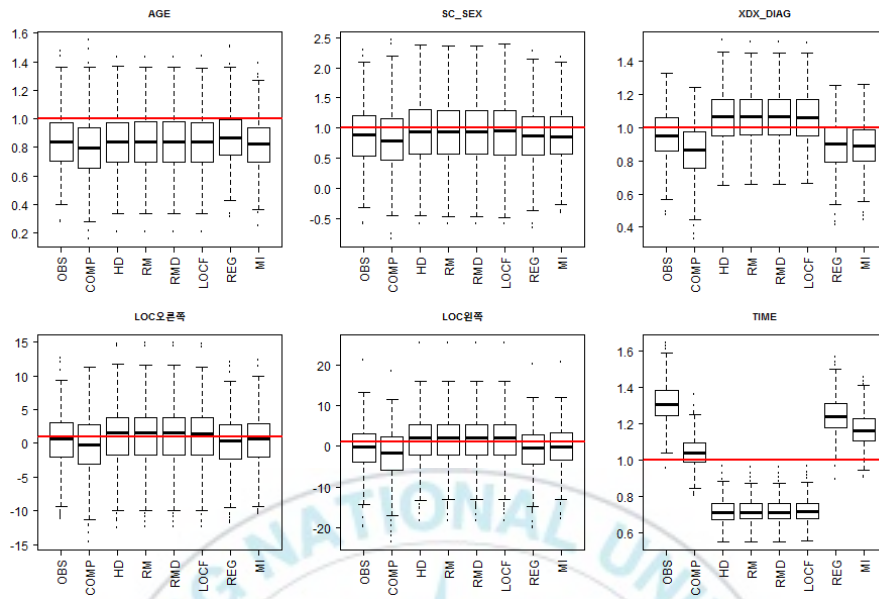


그림 4.7 FMA 결측률 30%일 때, sNMAR에서의 상대편의

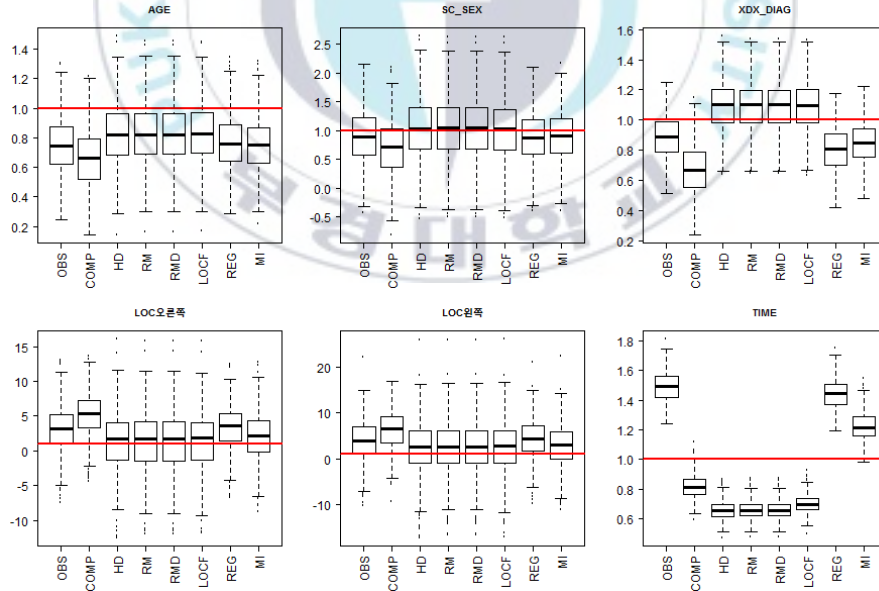


그림 4.8 FMA 결측률 30%일 때, cNMAR에서의 상대편의

표 4.10 FMA 결측률 50%일 때, 회귀계수

Mechanism	Variables	No miss	Observed		Imputed					
		FULL	OBS	COMP	RM	RMD	HD	LOCF	REG	MI
MCAR	x_1	-0.29	-0.26	-0.30	-0.23	-0.23	-0.23	-0.23	-0.28	-0.23
	x_2	-3.04	-2.80	-2.77	-2.84	-2.84	-2.84	-2.84	-2.81	-2.58
	x_3	-18.14	-19.54	-18.49	-20.53	-20.53	-20.53	-20.53	-17.95	-16.42
	x_4	-0.71	-0.14	0.22	-0.60	-0.60	-0.60	-0.60	0.23	0.21
	x_5	-0.51	0.52	1.68	-0.73	-0.73	-0.73	-0.73	1.10	0.40
	x_6	0.34	0.42	0.36	0.18	0.18	0.18	0.18	0.38	0.38
MAR	x_1	-0.29	-0.24	-0.25	-0.22	-0.22	-0.22	-0.22	-0.25	-0.21
	x_2	-3.04	-2.81	-2.33	-3.17	-3.17	-3.17	-3.17	-2.75	-2.71
	x_3	-18.14	-18.73	-14.04	-22.99	-22.99	-22.99	-22.94	-16.31	-17.03
	x_4	-0.71	-2.21	-3.37	-1.06	-1.06	-1.06	-1.06	-2.71	-1.52
	x_5	-0.51	-2.23	-3.50	-0.88	-0.88	-0.88	-0.88	-2.75	-1.57
	x_6	0.34	0.47	0.27	0.14	0.14	0.14	0.14	0.44	0.38
sNMA R	x_1	-0.29	-0.24	-0.25	-0.23	-0.23	-0.23	-0.23	-0.25	-0.21
	x_2	-3.04	-2.75	-2.08	-3.38	-3.38	-3.39	-3.31	-2.67	-2.72
	x_3	-18.14	-17.70	-14.54	-21.06	-21.06	-21.06	-21.17	-15.40	-15.35
	x_4	-0.71	-1.86	-3.01	-0.72	-0.72	-0.72	-0.64	-2.19	-1.12
	x_5	-0.51	-1.84	-2.99	-0.60	-0.60	-0.61	-0.58	-2.16	-1.13
	x_6	0.34	0.47	0.31	0.18	0.18	0.18	0.20	0.43	0.39
cNMA R	x_1	-0.29	-0.21	-0.18	-0.23	-0.23	-0.23	-0.23	-0.22	-0.20
	x_2	-3.04	-2.43	-1.71	-3.16	-3.16	-3.16	-3.17	-2.25	-2.48
	x_3	-18.14	-17.34	-10.65	-23.13	-23.13	-23.13	-23.09	-14.21	-15.78
	x_4	-0.71	-1.35	-2.06	-0.73	-0.73	-0.73	-0.73	-1.55	-0.95
	x_5	-0.51	-1.19	-1.87	-0.49	-0.49	-0.49	-0.46	-1.35	-0.81
	x_6	0.34	0.58	0.26	0.13	0.13	0.13	0.13	0.55	0.43

Missing Rate=50%

표 4.11 FMA 결측률 50%일 때, 편의

Mechanism	Variables	Observed		Imputed					
		OBS	COMP	RM	RMD	HD	LOCF	REG	MI
MCAR	x_1	0.03	-0.01	0.06	0.06	0.06	0.06	0.00	0.06
	x_2	0.24	0.27	0.20	0.20	0.20	0.20	0.23	0.47
	x_3	-1.40	-0.35	-2.39	-2.39	-2.39	-2.39	0.19	1.72
	x_4	0.57	0.92	0.11	0.11	0.11	0.11	0.94	0.92
	x_5	1.03	2.18	-0.22	-0.22	-0.22	-0.22	1.61	0.91
	x_6	0.08	0.02	-0.16	-0.16	-0.16	-0.16	0.04	0.04
MAR	x_1	0.05	0.03	0.07	0.07	0.07	0.07	0.03	0.08
	x_2	0.23	0.71	-0.13	-0.13	-0.13	-0.13	0.29	0.33
	x_3	-0.59	4.10	-4.85	-4.85	-4.85	-4.80	1.83	1.11
	x_4	-1.50	-2.66	-0.35	-0.35	-0.35	-0.35	-2.00	-0.82
	x_5	-1.72	-2.99	-0.37	-0.37	-0.37	-0.37	-2.25	-1.06
	x_6	0.13	-0.07	-0.20	-0.20	-0.20	-0.20	0.10	0.04
sNMA R	x_1	0.05	0.03	0.06	0.06	0.06	0.05	0.04	0.07
	x_2	0.30	0.96	-0.34	-0.34	-0.35	-0.27	0.37	0.33
	x_3	0.45	3.60	-2.92	-2.92	-2.92	-3.03	2.74	2.79
	x_4	-1.15	-2.30	-0.01	-0.01	-0.02	0.07	-1.48	-0.41
	x_5	-1.33	-2.48	-0.09	-0.09	-0.10	-0.07	-1.66	-0.62
	x_6	0.13	-0.03	-0.16	-0.16	-0.16	-0.14	0.09	0.05
cNMA R	x_1	0.08	0.11	0.06	0.06	0.06	0.06	0.07	0.08
	x_2	0.62	1.33	-0.12	-0.12	-0.11	-0.12	0.80	0.56
	x_3	0.80	7.49	-4.98	-4.98	-4.99	-4.95	3.93	2.36
	x_4	-0.64	-1.35	-0.02	-0.02	-0.02	-0.02	-0.84	-0.24
	x_5	-0.69	-1.37	0.02	0.02	0.02	0.05	-0.85	-0.30
	x_6	0.24	-0.08	-0.21	-0.21	-0.21	-0.21	0.20	0.09

Missing Rate=50%

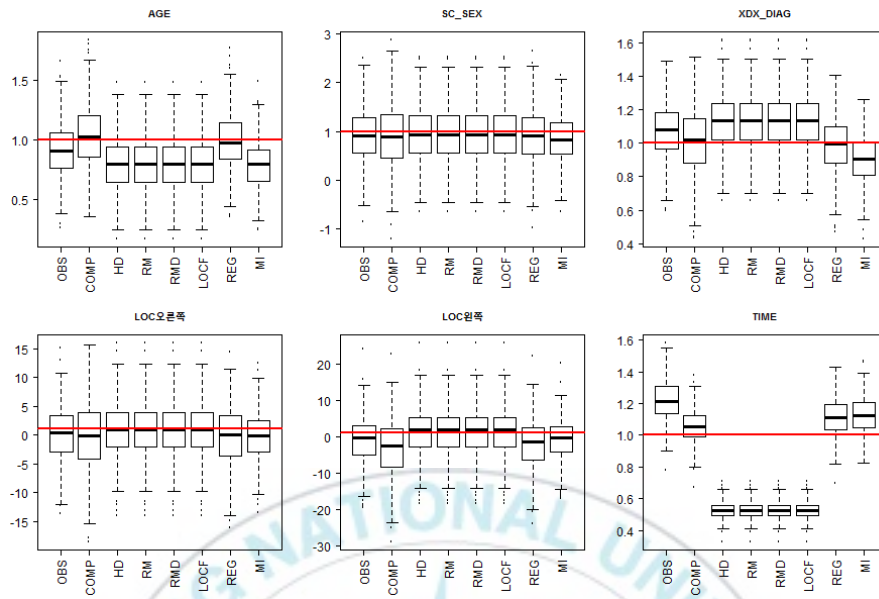


그림 4.9 FMA 결측률 50%일 때, MCAR에서의 상대편의

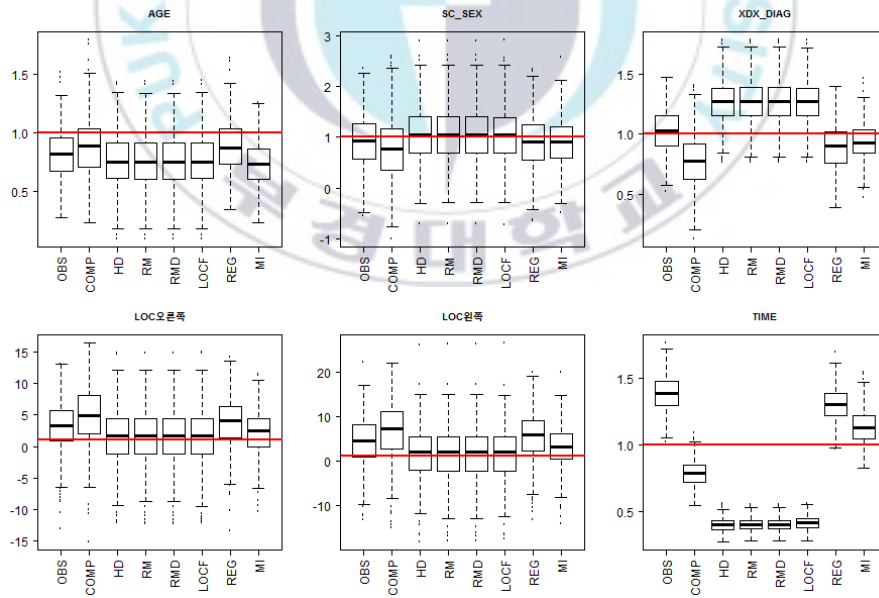


그림 4.10 FMA 결측률 50%일 때, MAR에서의 상대편의

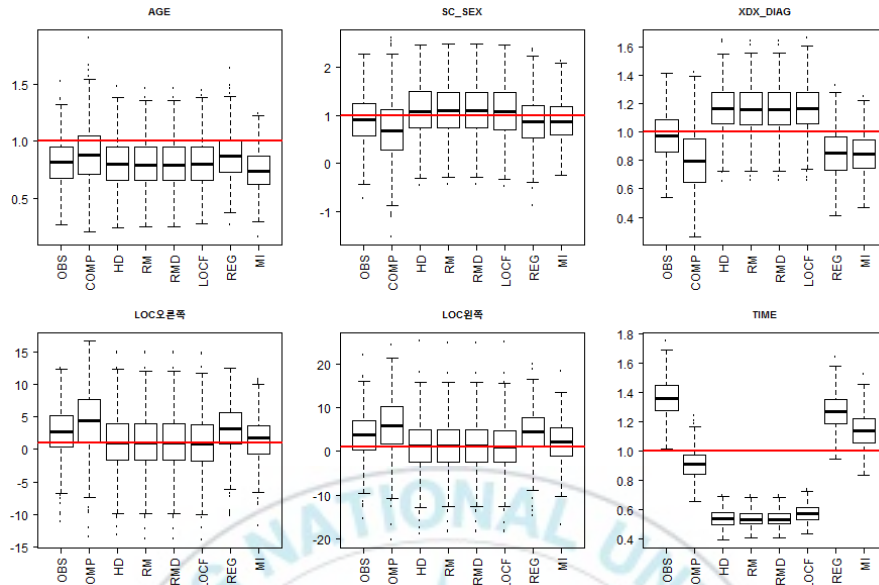


그림 4.11 FMA 결측률 50%일 때, sNMAR에서의 상대편의

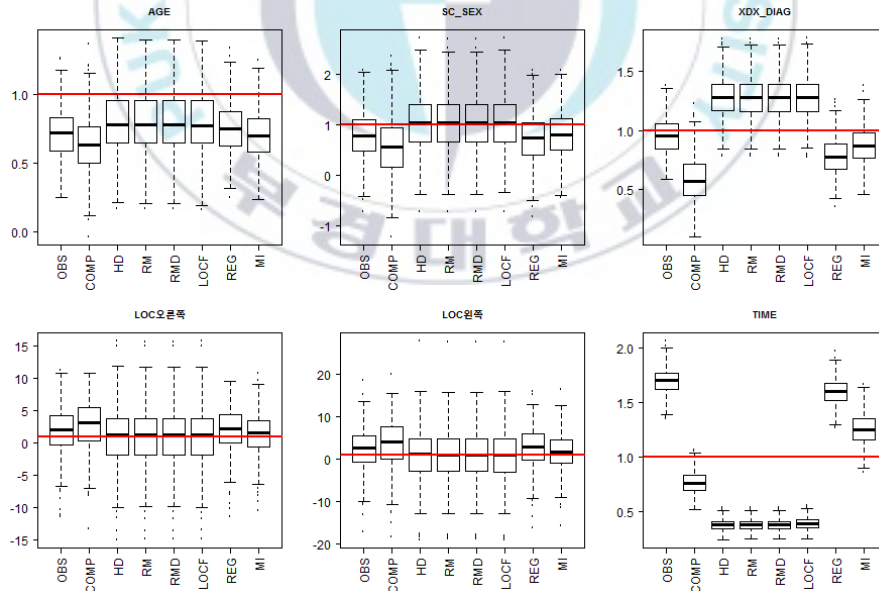


그림 4.12 FMA 결측률 50%일 때, cNMAR에서의 상대편의

4.2.2 순서형 자료

아래는 순서형 자료의 모의실험 결과이다. 결측구조 MCAR인 경우, 결측률이 낮을 때는 HD와 LOCF의 편의 정도가 낮으나, 결측률이 높아질수록 HD와 LOCF에서 상대편의가 1과 비교했을 때 크게 차이 남을 알 수 있었다.

결측구조 MAR인 경우, 연속형 자료와 마찬가지로, MAR구조 일 때 모수추정에 대한 편의가 작은 결측처리 방법은 COMP와 MI인 것으로 보인다. 그러나 연속형자료와는 다르게, 결측률이 높을수록 HD와 LOCF 뿐 아니라, REG의 상대편의정도가 크게 증가했다. 해당 부분은 향후 조사가 필요한 것으로 보인다.

결측구조 sNMAR, cNMAR인 경우, 결측률이 낮을 때는 HD와 LOCF의 편의 정도가 낮으나, 결측률이 높아질수록 HD와 LOCF에서 상대편의가 1과 비교했을 때 크게 차이 남을 알 수 있었다. sNMAR, cNMAR 구조 일 때 모수추정에 대한 편의가 작은 결측처리 방법은 MI인 것으로 보인다. 라이브러리 ‘MICE’의 MI또한 MAR 가정을 만족해야 잘 작동함에도 불구하고 NMAR에서 편의 정도가 크지 않아, 향후 조사가 필요해 보인다.

그 외, 더미변수에서는 결측처리 방법론 별로 편의가 작은 결과가 뚜렷하게 나타나지 않았는데, 이는 Harrell (2015)이 언급한대로 더미변수의 경우, 결측대체가 종속변수의 정보를 손실시켜 회귀계수의 편의를 야기하는 것으로 보인다.

표 4.12 FAC 결측률 10%일 때, 회귀계수

Mechanism	Variables	No miss	Observed		Imputed			
		FULL	OBS	COMP	HD	LOCF	REG	MI
MCAR	x_1	0.04	0.04	0.04	0.03	0.03	0.03	0.04
	x_2	0.36	0.37	0.38	0.28	0.28	0.36	0.37
	x_3	1.26	1.22	1.21	0.57	0.57	1.21	1.20
	x_4	-0.19	-0.19	-0.19	-0.12	-0.12	-0.21	-0.18
	x_5	-0.24	-0.23	-0.25	-0.16	-0.16	-0.27	-0.23
	x_6	-0.06	-0.08	-0.08	-0.06	-0.06	-0.09	-0.08
MAR	x_1	0.04	0.04	0.04	0.03	0.03	0.03	0.04
	x_2	0.36	0.37	0.39	0.28	0.28	0.36	0.38
	x_3	1.26	1.23	1.20	0.58	0.58	1.20	1.19
	x_4	-0.19	-0.18	-0.20	-0.12	-0.12	-0.21	-0.18
	x_5	-0.24	-0.23	-0.26	-0.17	-0.17	-0.27	-0.23
	x_6	-0.06	-0.09	-0.08	-0.06	-0.06	-0.09	-0.08
sNMA R	x_1	0.04	0.03	0.04	0.03	0.03	0.03	0.04
	x_2	0.36	0.37	0.38	0.28	0.28	0.36	0.37
	x_3	1.26	1.23	1.21	0.58	0.58	1.20	1.20
	x_4	-0.19	-0.19	-0.21	-0.12	-0.12	-0.22	-0.19
	x_5	-0.24	-0.24	-0.27	-0.17	-0.17	-0.28	-0.23
	x_6	-0.06	-0.09	-0.09	-0.06	-0.06	-0.10	-0.09
cNMA R	x_1	0.04	0.04	0.03	0.03	0.03	0.03	0.04
	x_2	0.36	0.37	0.39	0.28	0.28	0.36	0.38
	x_3	1.26	1.23	1.18	0.59	0.59	1.18	1.19
	x_4	-0.19	-0.19	-0.21	-0.12	-0.12	-0.22	-0.19
	x_5	-0.24	-0.24	-0.28	-0.17	-0.17	-0.28	-0.24
	x_6	-0.06	-0.09	-0.09	-0.07	-0.07	-0.10	-0.09

Missing Rate=10%

표 4.13 FAC 결측률 10%일 때, 편의

Mechanism	Variables	Observed		Imputed			
		OBS	COMP	HD	LOCF	REG	MI
MCAR	x_1	0.00	0.00	-0.01	-0.01	0.00	0.00
	x_2	0.01	0.02	-0.09	-0.09	0.00	0.01
	x_3	-0.04	-0.05	-0.69	-0.69	-0.05	-0.07
	x_4	0.01	0.00	0.08	0.08	-0.02	0.01
	x_5	0.01	-0.01	0.07	0.07	-0.03	0.01
	x_6	-0.02	-0.02	0.00	0.00	-0.03	-0.02
MAR	x_1	0.00	0.00	-0.01	-0.01	0.00	0.00
	x_2	0.01	0.03	-0.08	-0.08	0.00	0.01
	x_3	-0.03	-0.06	-0.69	-0.69	-0.07	-0.07
	x_4	0.01	-0.01	0.08	0.08	-0.02	0.01
	x_5	0.01	-0.02	0.07	0.07	-0.03	0.01
	x_6	-0.02	-0.02	0.00	0.00	-0.03	-0.02
sNMA R	x_1	0.00	0.00	-0.01	-0.01	0.00	0.00
	x_2	0.01	0.02	-0.08	-0.08	0.00	0.01
	x_3	-0.03	-0.06	-0.68	-0.69	-0.06	-0.07
	x_4	0.00	-0.01	0.07	0.07	-0.03	0.00
	x_5	0.00	-0.03	0.07	0.07	-0.04	0.01
	x_6	-0.03	-0.03	0.00	0.00	-0.03	-0.02
cNMA R	x_1	0.00	0.00	-0.01	-0.01	0.00	0.00
	x_2	0.01	0.03	-0.08	-0.08	0.00	0.02
	x_3	-0.03	-0.08	-0.67	-0.68	-0.09	-0.07
	x_4	0.00	-0.02	0.07	0.07	-0.03	0.00
	x_5	0.00	-0.04	0.07	0.07	-0.04	0.00
	x_6	-0.03	-0.03	-0.01	-0.01	-0.04	-0.03

Missing Rate=10%

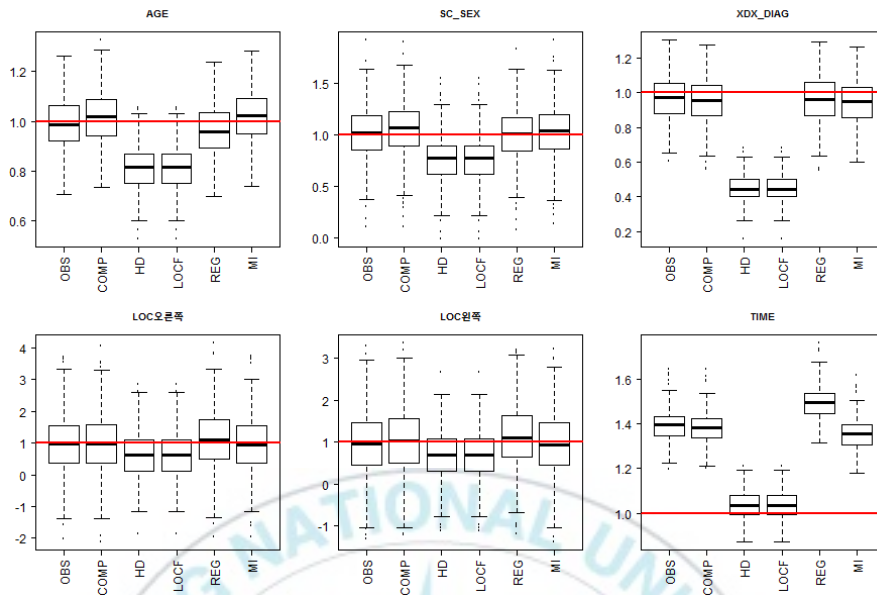


그림 4.13 FAC 결측률 10%일 때, MCAR에서의 상대편의

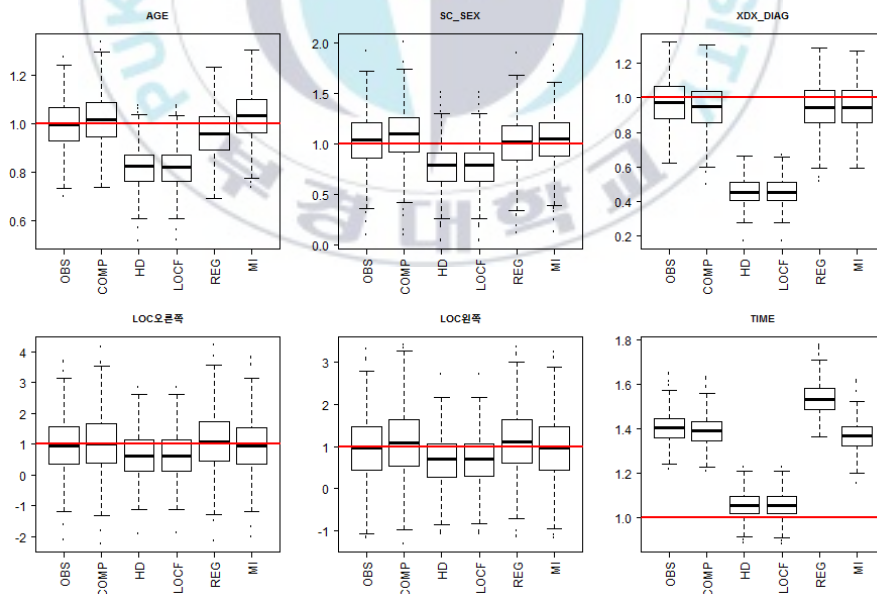


그림 4.14 FAC 결측률 10%일 때, MAR에서의 상대편의

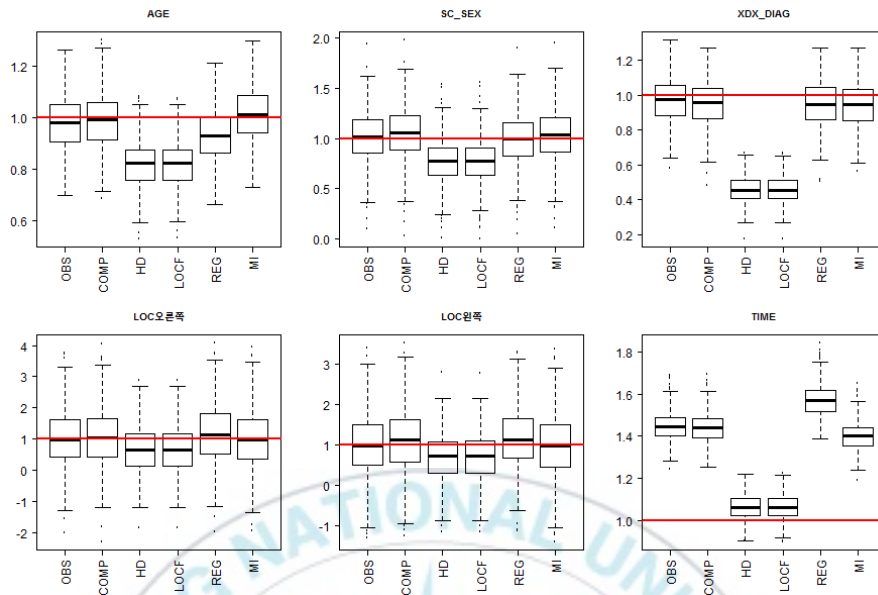


그림 4.15 FAC 결측률 10%일 때, sNMAR에서의 상대편의

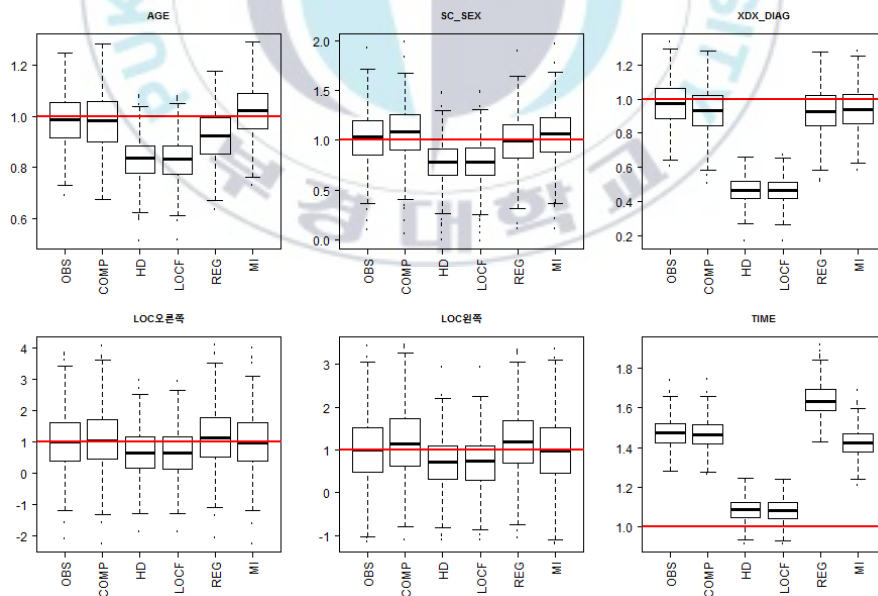


그림 4.16 FAC 결측률 10%일 때, cNMAR에서의 상대편의

표 4.14 FAC 결측률 30%일 때, 회귀계수

Mechanism	Variables	No miss	Observed		Imputed			
		FULL	OBS	COMP	HD	LOCF	REG	MI
MCAR	x_1	0.04	0.03	0.03	0.02	0.02	0.03	0.04
	x_2	0.36	0.36	0.41	0.24	0.24	0.36	0.37
	x_3	1.26	1.26	1.18	0.48	0.48	1.21	1.15
	x_4	-0.19	-0.19	-0.21	-0.09	-0.09	-0.24	-0.19
	x_5	-0.24	-0.23	-0.30	-0.13	-0.13	-0.30	-0.23
	x_6	-0.06	-0.09	-0.09	-0.05	-0.05	-0.12	-0.08
MAR	x_1	0.04	0.03	0.03	0.02	0.02	0.03	0.04
	x_2	0.36	0.36	0.42	0.24	0.24	0.35	0.38
	x_3	1.26	1.28	1.20	0.50	0.50	1.18	1.15
	x_4	-0.19	-0.19	-0.21	-0.09	-0.09	-0.23	-0.19
	x_5	-0.24	-0.23	-0.32	-0.13	-0.13	-0.29	-0.23
	x_6	-0.06	-0.10	-0.09	-0.05	-0.05	-0.13	-0.09
sNMA R	x_1	0.04	0.03	0.03	0.02	0.02	0.03	0.04
	x_2	0.36	0.36	0.40	0.24	0.24	0.33	0.37
	x_3	1.26	1.28	1.19	0.49	0.49	1.19	1.17
	x_4	-0.19	-0.20	-0.22	-0.09	-0.09	-0.23	-0.19
	x_5	-0.24	-0.24	-0.32	-0.13	-0.13	-0.29	-0.23
	x_6	-0.06	-0.10	-0.10	-0.05	-0.05	-0.13	-0.09
cNMA R	x_1	0.04	0.03	0.03	0.03	0.03	0.03	0.04
	x_2	0.36	0.35	0.39	0.24	0.24	0.32	0.37
	x_3	1.26	1.26	1.10	0.52	0.51	1.13	1.15
	x_4	-0.19	-0.15	-0.13	-0.08	-0.08	-0.14	-0.14
	x_5	-0.24	-0.20	-0.22	-0.12	-0.12	-0.20	-0.18
	x_6	-0.06	-0.10	-0.10	-0.05	-0.05	-0.14	-0.09

Missing Rate=30%

표 4.15 FAC 결측률 30%일 때, 편의

Mechanism	Variables	Observed		Imputed			
		OBS	COMP	HD	LOCF	REG	MI
MCAR	x_1	0.00	0.00	-0.01	-0.01	-0.01	0.00
	x_2	0.00	0.05	-0.13	-0.13	0.00	0.01
	x_3	0.00	-0.08	-0.78	-0.78	-0.05	-0.11
	x_4	0.00	-0.02	0.10	0.10	-0.04	0.00
	x_5	0.01	-0.06	0.11	0.11	-0.06	0.01
	x_6	-0.03	-0.03	0.02	0.02	-0.06	-0.02
MAR	x_1	0.00	0.00	-0.01	-0.01	-0.01	0.00
	x_2	0.00	0.06	-0.12	-0.12	-0.02	0.02
	x_3	0.02	-0.06	-0.76	-0.76	-0.09	-0.11
	x_4	0.00	-0.02	0.10	0.10	-0.03	0.00
	x_5	0.01	-0.08	0.11	0.11	-0.05	0.01
	x_6	-0.04	-0.03	0.01	0.01	-0.07	-0.03
sNMA R	x_1	0.00	0.00	-0.01	-0.01	-0.01	0.00
	x_2	-0.01	0.04	-0.12	-0.12	-0.03	0.01
	x_3	0.02	-0.08	-0.77	-0.77	-0.08	-0.10
	x_4	0.00	-0.02	0.11	0.11	-0.04	0.00
	x_5	0.00	-0.08	0.11	0.11	-0.05	0.01
	x_6	-0.04	-0.04	0.01	0.01	-0.07	-0.03
cNMA R	x_1	0.00	0.00	-0.01	-0.01	-0.01	0.00
	x_2	-0.01	0.03	-0.12	-0.12	-0.04	0.01
	x_3	0.00	-0.16	-0.75	-0.75	-0.13	-0.11
	x_4	0.04	0.06	0.11	0.11	0.05	0.06
	x_5	0.04	0.02	0.12	0.12	0.04	0.06
	x_6	-0.04	-0.04	0.01	0.01	-0.08	-0.03

Missing Rate=30%

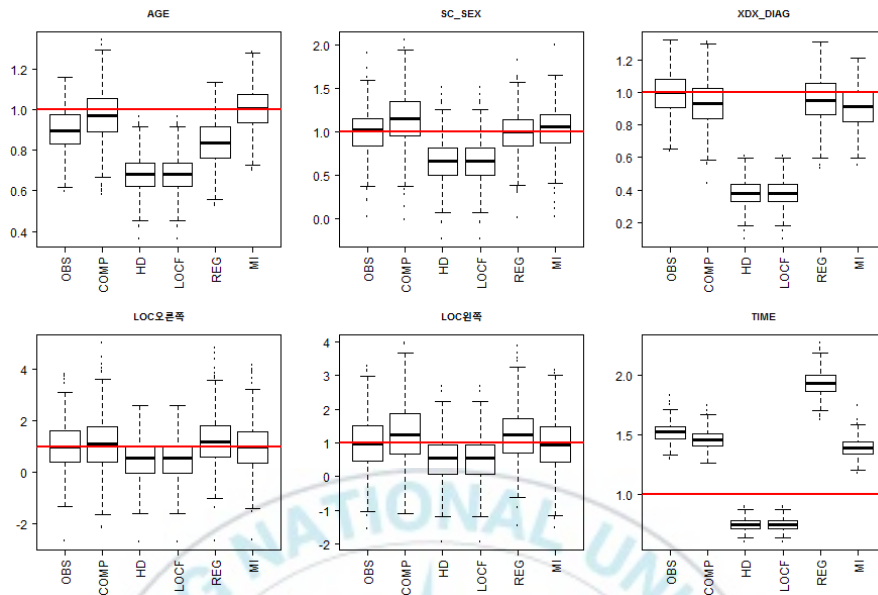


그림 4.17 FAC 결측률 30%일 때, MCAR에서의 상대편의

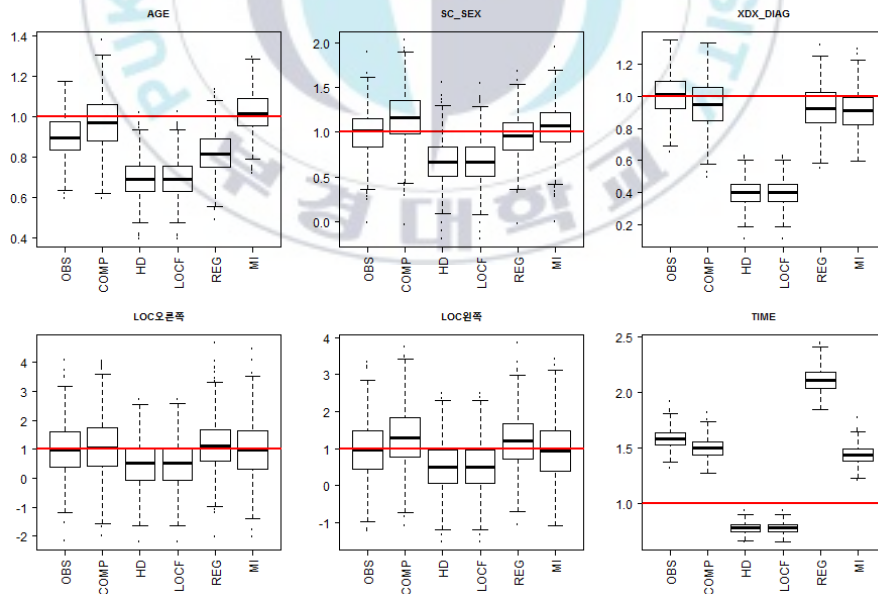


그림 4.18 FAC 결측률 30%일 때, MAR에서의 상대편의

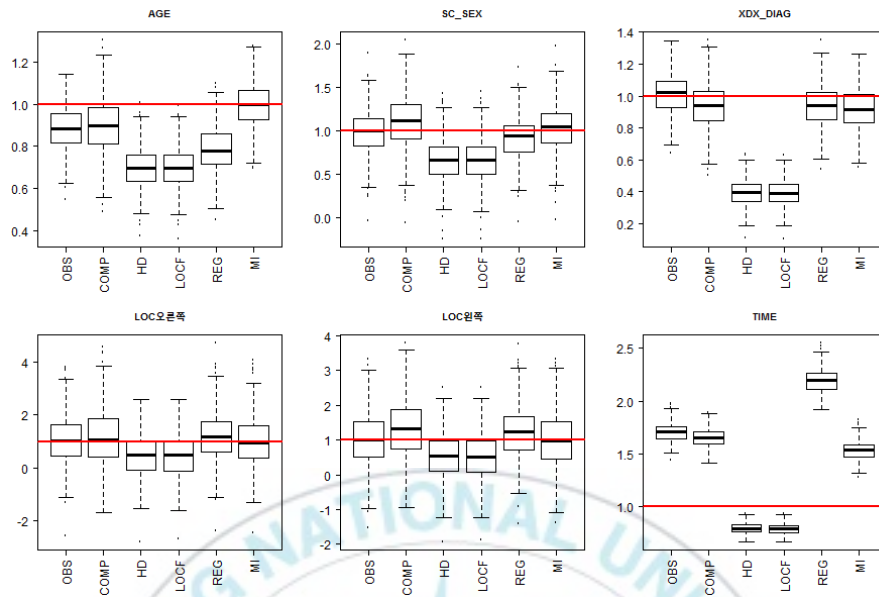


그림 4.19 FAC 결측률 30%일 때, sNMAR에서의 상대편의

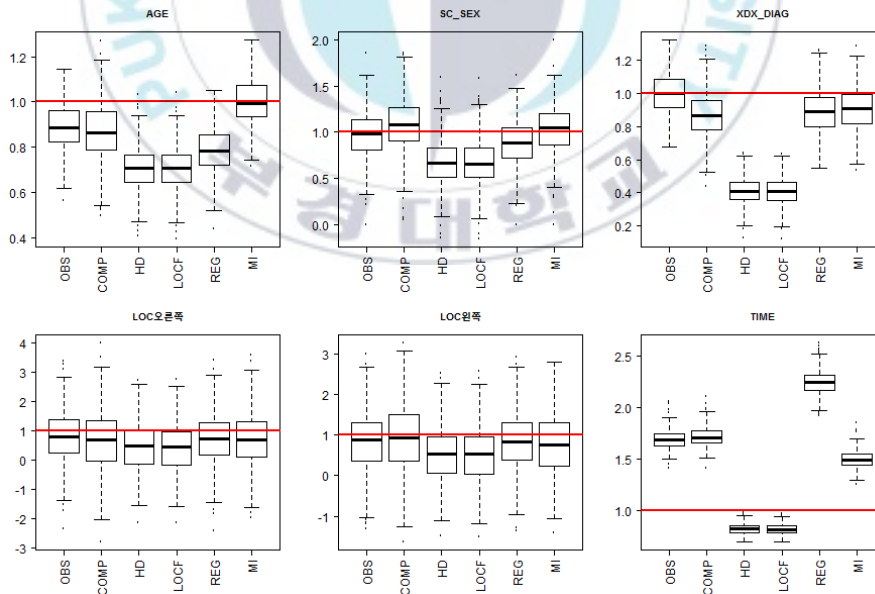


그림 4.20 FAC 결측률 30%일 때, cNMAR에서의 상대편의

표 4.16 FAC 결측률 50%일 때, 회귀계수

Mechanism	Variables	No miss	Observed		Imputed			
		FULL	OBS	COMP	HD	LOCF	REG	MI
MCAR	x_1	0.04	0.03	0.03	0.02	0.02	0.03	0.04
	x_2	0.36	0.34	0.37	0.25	0.25	0.32	0.36
	x_3	1.26	1.36	1.28	0.47	0.47	1.27	1.17
	x_4	-0.19	-0.19	-0.19	-0.13	-0.13	-0.23	-0.16
	x_5	-0.24	-0.24	-0.34	-0.19	-0.19	-0.29	-0.22
	x_6	-0.06	-0.11	-0.10	-0.03	-0.03	-0.16	-0.09
MAR	x_1	0.04	0.03	0.04	0.02	0.02	0.03	0.04
	x_2	0.36	0.34	0.34	0.26	0.26	0.29	0.36
	x_3	1.26	1.44	1.18	0.48	0.48	1.23	1.26
	x_4	-0.19	-0.25	-0.18	-0.15	-0.15	-0.21	-0.24
	x_5	-0.24	-0.28	-0.18	-0.20	-0.20	-0.21	-0.27
	x_6	-0.06	-0.08	-0.07	-0.03	-0.03	-0.15	-0.07
sNMA R	x_1	0.04	0.03	0.03	0.02	0.02	0.02	0.03
	x_2	0.36	0.34	0.31	0.25	0.25	0.28	0.28
	x_3	1.26	1.25	1.14	0.42	0.41	1.06	0.87
	x_4	-0.19	-0.16	-0.08	-0.08	-0.11	-0.14	-0.11
	x_5	-0.24	-0.19	-0.09	-0.13	-0.16	-0.18	-0.16
	x_6	-0.06	-0.10	-0.10	-0.03	-0.03	-0.15	-0.08
cNMA R	x_1	0.04	0.03	0.03	0.02	0.02	0.02	0.04
	x_2	0.36	0.33	0.31	0.27	0.27	0.25	0.34
	x_3	1.26	1.44	1.10	0.50	0.49	1.11	1.20
	x_4	-0.19	-0.25	-0.19	-0.13	-0.14	-0.18	-0.24
	x_5	-0.24	-0.29	-0.21	-0.18	-0.19	-0.20	-0.27
	x_6	-0.06	-0.09	-0.09	-0.03	-0.03	-0.17	-0.08

Missing Rate=50%

표 4.17 FAC 결측률 50%일 때, 편의

Mechanism	Variables	Observed		Imputed			
		OBS	COMP	HD	LOCF	REG	MI
MCAR	x_1	-0.01	0.00	-0.01	-0.01	-0.01	0.00
	x_2	-0.02	0.01	-0.11	-0.11	-0.04	0.00
	x_3	0.09	0.02	-0.79	-0.79	0.01	-0.09
	x_4	0.00	0.00	0.06	0.06	-0.03	0.03
	x_5	0.00	-0.10	0.05	0.05	-0.05	0.02
	x_6	-0.05	-0.04	0.03	0.03	-0.10	-0.03
MAR	x_1	-0.01	0.00	-0.02	-0.02	-0.01	0.00
	x_2	-0.02	-0.02	-0.10	-0.10	-0.08	0.00
	x_3	0.18	-0.08	-0.78	-0.78	-0.03	0.00
	x_4	-0.06	0.01	0.05	0.04	-0.02	-0.04
	x_5	-0.04	0.06	0.04	0.04	0.03	-0.03
	x_6	-0.02	-0.01	0.03	0.03	-0.08	-0.01
sNMA R	x_1	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
	x_2	-0.02	-0.05	-0.11	-0.11	-0.08	-0.09
	x_3	-0.01	-0.12	-0.84	-0.85	-0.20	-0.40
	x_4	0.04	0.12	0.11	0.08	0.05	0.08
	x_5	0.05	0.15	0.11	0.08	0.06	0.08
	x_6	-0.04	-0.04	0.03	0.03	-0.08	-0.02
cNMA R	x_1	-0.01	-0.01	-0.02	-0.02	-0.01	0.00
	x_2	-0.03	-0.05	-0.09	-0.09	-0.11	-0.03
	x_3	0.18	-0.17	-0.77	-0.77	-0.16	-0.06
	x_4	-0.06	0.00	0.06	0.05	0.01	-0.05
	x_5	-0.05	0.03	0.06	0.05	0.04	-0.03
	x_6	-0.03	-0.03	0.03	0.03	-0.10	-0.02

Missing Rate=50%

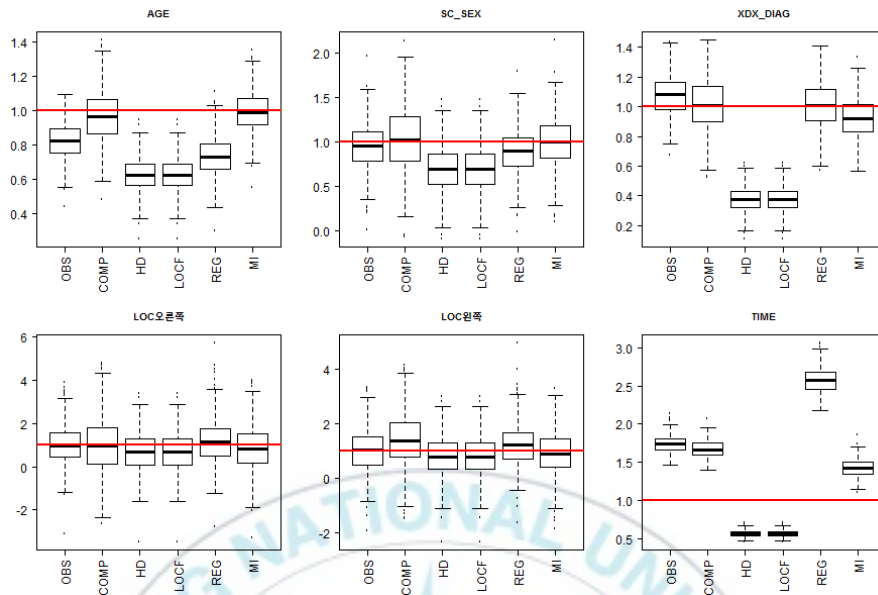


그림 4.21 FAC 결측률 50%일 때, MCAR에서의 상대편의

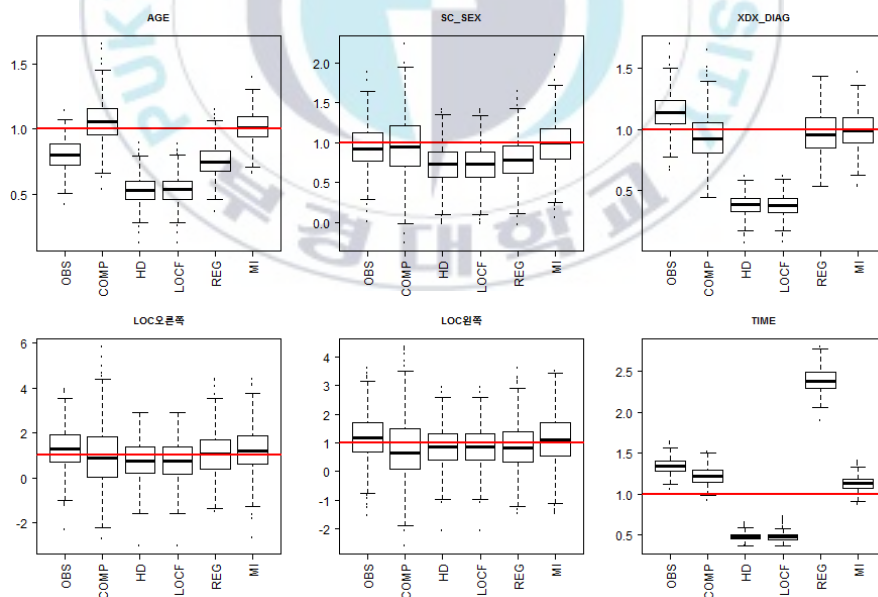


그림 4.22 FAC 결측률 50%일 때, MAR에서의 상대편의

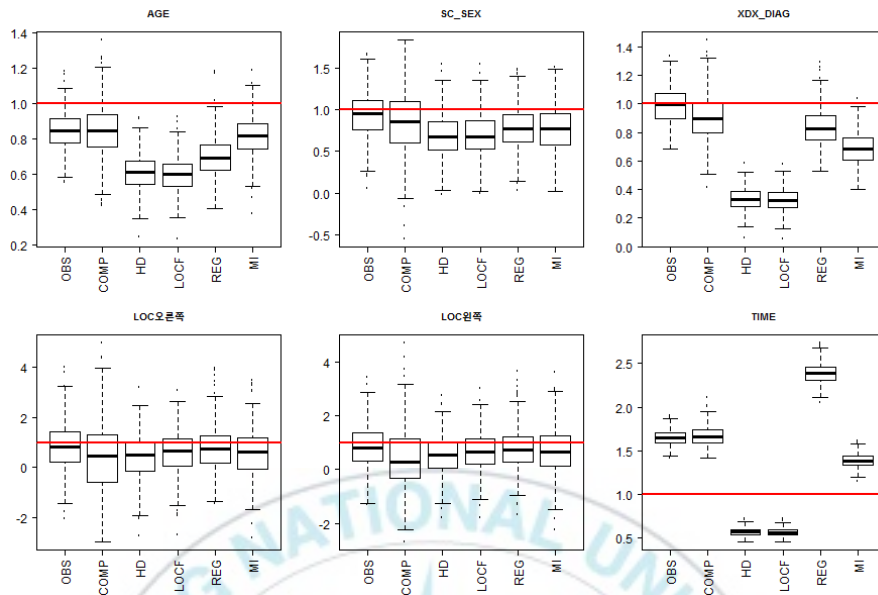


그림 4.23 FAC 결측률 50%일 때, sNMAR에서의 상대편의

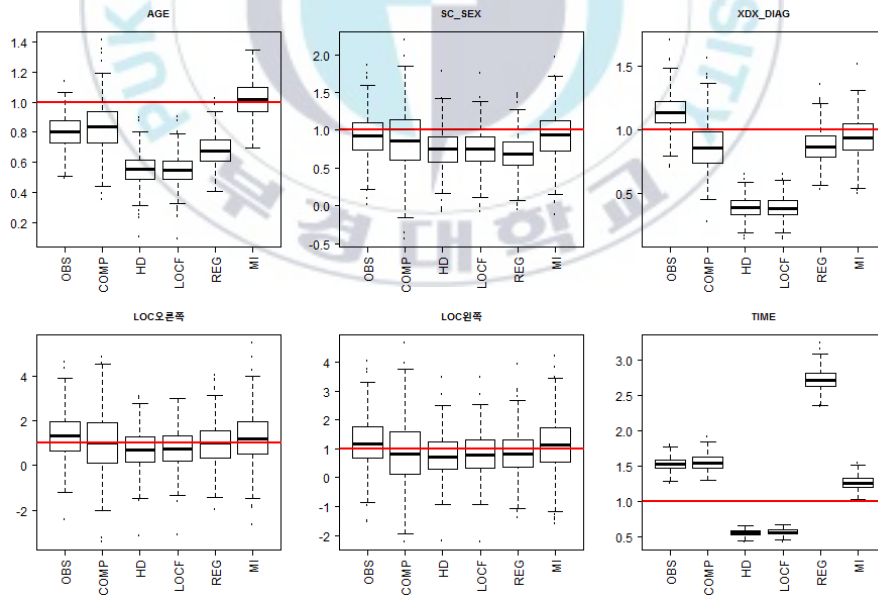


그림 4.24 FAC 결측률 50%일 때, cNMAR에서의 상대편의

제 5장 결론

본 논문에서는 결측이 있는 반복측정을 고려한 임상자료에서 흔히 관찰되는 임의패턴 결측자료에 대한 결측 처리 방법론별 종속변수 형태에 따른 GEE추정량의 효율성을 알아보았다.

실제자료 분석 결과 결측처리 방법론 별로 결과에 큰 차이가 났다. 결측이 있는 반복측정된 임상자료에서 흔히 사용되는 완전사례분석법, 관찰자료분석법, 행평균대체, 이월대체 등 결측처리 방법론에 따라 결과가 달라지는 문제가 있어, 모의실험을 통해 그에 따른 편이의 정도를 알아보았다.

모의실험 결과 MCAR, MAR 뿐 아니라, NMAR에서도 R의 라이브러리 ‘MICE’를 이용한 다중대체 방법이 고려한 결측처리 방법 중에서 가장 작은 편이추정량을 제공하는 것으로 보였다. 그러나 일반적인 경우, 다른 연구에서는 MAR가정 하에서 전개한 다중대체 방법은 NMAR구조에서는 잘 작동하지 않는 것으로 보였다. NMAR구조에서도 잘 작동한 이유에 대해서는 추후 연구가 더 필요해 보인다. 또한 김동욱 그리고 노영화 (2003)의 연구에서는 행평균 대체의 효율성이 좋았으나, 본 연구에서는 상대적으로 나빠 보였다. 이는 독립변수에 결측이 존재하는지, 아니면 종속변수에 결측이 존재하는지에 따른 차이로 보인다. 본 연구의 결과로는 종속변수의 결측이 있는 경우, 핫덱 대체, 행평균 대체, 행중위수 대체, 이월대체를 적용하는 것은 별로 권장할 만한 결과가 아니었다. 따라서 결측이 있는 반복측정된 임상자료에서 GEE추정량을 구하고자

한다면, 어떤 변수에 결측이 있는지에 따라 결측처리 방법론의 선택이 중요할 것이라 생각되어진다.



참고문헌

- [1] 강현철 외 (2010). 인구주택총조사 자료분석 및 활용제고방안 연구. 한국통계학회.
- [2] 김덕준 (2011). 범주형 자료의 결측치 처리 방법. 한국비교정부학보, 15 권 2 호, 319-342.
- [3] 김동욱, 김재직 (2002). 범주형 반복측정자료를 위한 일반화 추정 방정식의 소표본 특성. 응용통계연구, 15권 2호, 297-310.
- [4] 김동욱, 노영화 (2003). 대체방법별 GEE추정량 비교. 응용통계연구, 16권 2호, 407-426.
- [5] 김태훈 (2004). GEE모형에서 종속변수에 결측이 있는 경우 처리방법. 성균관대학교학위논문.
- [6] 민인식, 최필선 (2012). STATA 고급 패널데이터 분석. 한국STATATA학회.
- [7] 서울대학교산학협력단 (2007). 반복측정자료를 이용한 질환연관성 분석연구. 질병관리본부 학술연구용역사업 최종결과보고서.
- [8] 조현진 (2012). 상호작용 분석시 다중대체방법의 비교. 고려대학교 학위논문.
- [9] 허명희 (2014). 응용데이터분석 결측값 대체. 자유아카데미.
- [10] Agresti, Alan (2007). An Introduction to Categorical Data Anal

- ysis. Wiley.
- [11] Allison, Paul D. (2001). Missing Data. SAGE.
- [12] Baraldi, A. N. and Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology*, 48(1), 5-37.
- [13] Chang, W. H. et al (2015). Korean Stroke Cohort for functioning and rehabilitation (KOSCO): study rationale and protocol of a multi-centre prospective cohort study. *BMC Neurology*, 15(42).
- [14] Enders, Craig K. (2010). *Applied Missing Data Analysis*. The Guilford Press.
- [15] Fitzmaurice, G. M. (2008). *Longitudinal Data Analysis*. CRC Press.
- [16] Graham, J. W., Olchowski, A. E., Gilreath, T. D. (2007). How Many Imputations are Really Needed? Some Practical Clarifications of Multiple Imputation Theory. *Prevention Science*, 8(3), 206-213.
- [17] Hardin, James W. and Hilbe, Joseph M. (2012). *Generalized Estimating Equations*, Second Edition. CRC Press.
- [18] Harrell, Frank E., Jr. (2015). *Regression Modeling Strategies*. Springer.

- [19] Hsiao, C. (1985). Benefits and limitations of panel data. *Econometric Reviews*, 4(1), 121-174.
- [20] Laird, Nan M. (1988). Missing data in longitudinal studies. *Statistics in Medicine*, 7, 305-15.
- [21] Liang, Kung-Yee and Zeger, Scott L. (1986). Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika*, 73(1), 13-22.
- [22] Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley-Interscience.
- [23] Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*, 63(3), 581-592.
- [24] Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley.
- [25] Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Analysis*. Chapman and Hall.
- [26] Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research*, 8, 3-15.
- [27] Schafer, J. L., Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7(2), 147-177.
- [28] Shults, Justine and Hilbe, Joseph M. (2014). *Quasi Least Squares Regression*. CRC Press.
- [29] van Buuren, Stef (2012). *Flexible Imputation of Missing Data*. Chapman and Hall/CRC.

- [30] Wang, Ming (2014). Generalized Estimating Equations in Longitudinal Data Analysis: A Review and Recent Developments. *Advances in Statistics*, 2014, 1-11.
- [31] Wedderburn, R. W. M. (1974). Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method. *Biometrika*, 61 (3), 439.
- [32] Yang, Sookyeong (2009). A Comparison of Imputation Methods for Incomplete Longitudinal Data. Sungkyunkwan University.
- [33] Ziegler, Andreas (2011). Generalized Estimating Equations. Springer.



부 록

부록1. KOSCO 연구 중 본 논문 분석에 사용된 자료

변수	변수설명
SUBJNO	Screening 번호
SUBJFUNO	환자 추적 번호
AGE	발병기준 나이
SC_SEX	성별
XDX_DIAG	진단명
LOC	뇌병변 위치
FMATOTA	FMA Affected
FMATOTAYN	FMA Affected 결측 여부
FAC	Functional Ambulatory Category
FACYN	Functional Ambulatory Category 결측 여부
TIME	발병 후 추적기간

부록2. 결측 대체 방법론 별 GEE 분석을 위한 R코드

```
library(gee)
library(mice)
library(longitudinalData)
library(reshape)

dat <- read.csv("C:\\Users\\Thinkpad\\Desktop\\18-05-02-3-MU.csv")
dat2 <- subset(dat, subset = !(SUBJFUNO %in% ""))
dat3 <- subset(dat2, subset = !(is.na(FMATOTA_7D) ))

# select variables
nd.X <- names(dat3)[!names(dat3) %in% c("SUBJNO", "SUBJFUNO", "ALL",
names(dat3)[grep("FMATOTA_", names(dat3))], names(dat3)[grep("FMATOTAYN_",
names(dat3))])]
nd.Y <- names(dat3)[names(dat3) %in% c(names(dat3)[grep("FMATOTA_",
names(dat3))])]

dat.anal <- dat3[,c("SUBJNO", nd.X, nd.Y)]

① Observed Analysis
# long form data
VARY <- which(names(dat.anal) %in% c(names(dat.anal)[grep("FMATOTA_",
names(dat.anal))]))
dat.l <- reshape(dat.anal, varying = VARY, v.names="FMATOTA",
idvar="SUBJNO", timevar="TIME", times=c(0,12,24), direction = "long")
dat.l <- completeFun(dat.l, "FMATOTA")
cat("Number of Obs: ", nrow(dat.l), "\n")
cat("Number of IDs: ", sum(count(unique(dat.l$SUBJNO))$freq), "\n")

# GEE
fit <- gee(FMATOTA ~ AGE + SC_SEX + XDX_DIAG + LOC +
TIME,id=SUBJNO,family=gaussian,corstr="exchangeable",data=dat.l)
```

② Complete Case Analysis

```
# select variables
dat.anal <- na.omit(dat3[,c("SUBJNO", nd.X, nd.Y)])

# long form data
VARY <- which(names(dat.anal) %in% c(names(dat.anal)[grep("FMATOTA_",
names(dat.anal))]))
dat.l <- reshape(dat.anal, varying = VARY, v.names="FMATOTA",
                 idvar="SUBJNO", timevar="TIME", times=c(0,12,24), direction = "long")
cat("Number of Obs: ", nrow(dat.l), "\n")
cat("Number of IDs: ", sum(count(unique(dat.l$SUBJNO))$freq), "\n")

# GEE
fit <- gee(FMATOTA ~ AGE + SC_SEX + XDX_DIAG + LOC +
TIME,id=SUBJNO,family=gaussian,corstr="exchangeable",data=dat.l)
```

③ Hot Deck Imputation

```
# select variables
set.seed(123)
imp.Y <- imputation(as.matrix(dat3[,nd.Y]), "trajHotDeck")
dat.anal <- data.frame(dat3[,c("SUBJNO", nd.X)], imp.Y)

# long form data
VARY <- which(names(dat.anal) %in% c(names(dat.anal)[grep("FMATOTA_",
names(dat.anal))]))
dat.l <- reshape(dat.anal[,c("SUBJNO", nd.X, nd.Y)], varying = VARY,
v.names="FMATOTA",
                 idvar="SUBJNO", timevar="TIME", times=c(0,12,24), direction = "long")
```

```
# GEE
fit <- gee(FMATOTA ~ AGE + SC_SEX + XDX_DIAG + LOC +
TIME,id=SUBJNO,family=gaussian,corstr="exchangeable",data=dat.l)
```

④ Row Mean Imputation

```
# select variables
imp.Y <- imputation(as.matrix(dat3[,nd.Y]), "trajMean")
dat.anal <- data.frame(dat3[,c("SUBJNO", nd.X)], imp.Y)

# long form data
VARY <- which(names(dat.anal) %in% c(names(dat.anal)[grep("FMATOTA_",
names(dat.anal))]))
dat.l <- reshape(dat.anal[,c("SUBJNO", nd.X, nd.Y)], varying = VARY,
v.names="FMATOTA",
idvar="SUBJNO", timevar="TIME", times=c(0,12,24), direction = "long")
```

```
# GEE
fit <- gee(FMATOTA ~ AGE + SC_SEX + XDX_DIAG + LOC +
TIME,id=SUBJNO,family=gaussian,corstr="exchangeable",data=dat.l)
```

⑤ Row Median Imputation

```
# select variables
imp.Y <- imputation(as.matrix(dat3[,nd.Y]), "trajMedian")
dat.anal <- data.frame(dat3[,c("SUBJNO", nd.X)], imp.Y)

# long form data
VARY <- which(names(dat.anal) %in% c(names(dat.anal)[grep("FMATOTA_",
names(dat.anal))]))
dat.l <- reshape(dat.anal[,c("SUBJNO", nd.X, nd.Y)], varying = VARY,
v.names="FMATOTA",
idvar="SUBJNO", timevar="TIME", times=c(0,12,24), direction = "long")
```

```

# GEE
fit <- gee(FMATOTA ~ AGE + SC_SEX + XDX_DIAG + LOC +
TIME,id=SUBJNO,family=gaussian,corstr="exchangeable",data=dat.l)

⑥ LOCF
# select variables
imp.Y <- imputation(as.matrix(dat3[,nd.Y]), "locf")
colnames(imp.Y) <- nd.Y
dat.anal <- data.frame(dat3[,c("SUBJNO", nd.X)], imp.Y)

# long form data
VARY <- which(names(dat.anal) %in% c(names(dat.anal)[grep("FMATOTA_",
names(dat.anal))]))
dat.l <- reshape(dat.anal[,c("SUBJNO", nd.X, nd.Y)], varying = VARY,
v.names="FMATOTA",
idvar="SUBJNO", timevar="TIME", times=c(0,12,24), direction = "long")

# GEE
fit <- gee(FMATOTA ~ AGE + SC_SEX + XDX_DIAG + LOC +
TIME,id=SUBJNO,family=gaussian,corstr="exchangeable",data=dat.l)

⑦ Regression Imputation
fit <- lm(FMATOTA_12M ~ AGE + SC_SEX + XDX_DIAG + LOC, data = dat3)
FMATOTA_12M.regress <- round(ifelse(is.na(dat3$FMATOTA_12M),
predict(fit,newdata=dat3), dat3$FMATOTA_12M),1)
fit <- lm(FMATOTA_24M ~ AGE + SC_SEX + XDX_DIAG + LOC, data = dat3)
FMATOTA_24M.regress <- round(ifelse(is.na(dat3$FMATOTA_24M),
predict(fit,newdata=dat3), dat3$FMATOTA_24M),1)
dat.anal <- data.frame(dat3[,c("SUBJNO", nd.X)], FMATOTA_7D = dat3$FMATOTA_7D
,FMATOTA_12M = FMATOTA_12M.regress, FMATOTA_24M = FMATOTA_24M.regress)

```

```

# long form data
VARY <- which(names(dat.anal) %in% c(names(dat.anal)[grep("FMATOTA_",
names(dat.anal))]))
dat.l <- reshape(dat.anal[,c("SUBJNO", nd.X, nd.Y)], varying = VARY,
v.names="FMATOTA",
idvar="SUBJNO", timevar="TIME", times=c(0,12,24), direction = "long")

# GEE
fit <- gee(FMATOTA ~ AGE + SC_SEX + XDX_DIAG + LOC +
TIME,id=SUBJNO,family=gaussian,corstr="exchangeable",data=dat.l)

```

⑦ Multiple Imputation

```

M <- 5
hyper.imp <- mice(dat3[,c(nd.X, nd.Y)], m = M, method = "pmm", seed = 1)
all.imp <- complete(hyper.imp, action="long")

hyper.imp.data <- as.list(1:M)
for (i in 1:M){
  hyper.imp.data[[i]] <- data.frame(SUBJNO = dat3$SUBJNO,
complete(hyper.imp,action=i))
}

VARY <- which(names(hyper.imp.data[[1]]) %in%
c(names(hyper.imp.data[[1]])[grep("FMATOTA_", names(hyper.imp.data[[1]]))]))
hyper.imp.data2 <-
lapply(hyper.imp.data,reshape,varying=VARY,v.names="FMATOTA",idvar="SUBJNO",timeva
r="TIME",
times=c(0,12,24),direction="long")
hyper.imp.data2 <- lapply(hyper.imp.data2,FUN=function(u){ u[order(u$SUBJNO),]})

```

```

hyper.imp.gee <- lapply(hyper.imp.data2, FUN=function(u){
  gee(FMATOTA ~ AGE + SC_SEX + XDX_DIAG + LOC +
TIME,id=SUBJNO,family=gaussian,corstr="exchangeable",data=u)
})

hyper.imp.parmmat <- sapply(hyper.imp.gee,coefficients)
hyper.imp.parmmat

minparm <- apply(hyper.imp.parmmat,1,min)
maxparm <- apply(hyper.imp.parmmat,1,max)

hyper.imp.parms <- lapply(hyper.imp.gee, coefficients)
hyper.imp.nse <- lapply(hyper.imp.gee, function(y) sqrt(diag(y$naive.variance)))
hyper.imp.rse <- lapply(hyper.imp.gee, function(y) sqrt(diag(y$robust.variance)))

mi.parms.nse <- mi.inference(hyper.imp.parms,hyper.imp.nse,confidence=0.95)
mi.parms.rse <- mi.inference(hyper.imp.parms, hyper.imp.rse,confidence=0.95)

```