The Korean Journal of Applied Statistics (2017) DOI: https://doi.org/10.5351/KJAS.2017.30.3.463
30(3), 463–473

Imputation method for missing data based on measure of property

Hyungju Kim[a] · Dongjae Kim[a,1]

[a]Department of Biomedicine·Health Science, The Catholic University of Korea

(Received April 5, 2017; Revised May 5, 2017; Accepted May 5, 2017)

Abstract

How to handle missing data is a main issue in clinical trials. We impute missing data based on missing data that follows a mechanism according to the intention-to-treat rule. However, using the right imputation method for missing data is very important because this supposition is unclear. We suggest a new imputation method for missing data using agreement and maintenance introduced by Kang and Kim (1997). We give an example and adapt a Monte Carlo simulation to compare the performance between the established method and the suggested method.

Keywords: missing data imputation, agreement, maintenance, clinical trials

1. 서론

임상시험이란 신약이나 식품, 의료기기, 새로운 시술법 등의 안정성과 유효성을 증명하기 위하여 시행하는 시험으로서, 국제의약품 규제조화 위원회(International Conference on Harmonization; ICH)는 통계분석 원칙에 대한 Guideline을 ICH E9에 제안하여 철저한 관리 하에 임상시험이 이루어지도록 하고 있다. 사람을 대상으로 진행하는 임상시험의 특성상 선정기준을 위반한 피험자의 시험참여, 임상시험 계획서를 위반하는 피험자 및 본인의 의사에 따라 임상시험 도중에 그만두는 피험자 등 다양한 요인으로 인하여 '결측치'가 발생한다고 알려져 있다 (Kang, 2013).

흔히 임상시험에서 주분석으로 Intention-To-Treat (ITT) 법칙을 이용하고 있으며 결측치에 의한 문제점이 지속적으로 제기되고 있다. 결측치가 발생하면 ITT법칙에 의해 어떠한 메커니즘을 따른다는 가정을 하고, 결측치를 대체하게 된다. 하지만 그 가정의 타당성을 밝히기 어렵다는 한계로 인해 결측치 문제는 해결되지 않는 문제 중 하나이다 (Kang, 2013). 따라서 임상시험에서 결측치를 다루는 방법은 매우 중요하며, 결측치 발생에 의한 문제점에 대처하기 위하여 완전제거법(list-wise deletion), 평균대체법(unconditional mean; UM), 핫덱 대체(hot-deck imputation by simple random sampling with replacement) 등이 사용되고 있다.

우선 완전제거법은 대부분의 통계프로그램에서 기본적으로 설정되어있는 보완 기법으로 하나라도 결측이 발생한 케이스를 분석대상에서 제외하고 분석하는 방법으로서, 완전임의결측일 때 모수추정치의 편

[1]Corresponding author: Department of Biomedicine · Health Science, The Catholic University of Korea, 222 Banpo-dero Seocho-gu, Seoul 06591, Korea. E-mail: djkim@catholic.ac.kr

Table 2.1. Structure of reapted measure data

N	Time			
	1	2	$\cdots$	m
1	y_{11}	y_{12}	$\cdots$	y_{1m}
2	y_{21}	y_{22}	$\cdots$	y_{2m}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n	y_{n1}	y_{n2}	$\cdots$	y_{nm}

의를 최소화 할 수 있으며, 사용이 용이하여 결측치의 비율이 낮을 때 선호되어진다. 하지만 현실적으로 완전임의결측일 가능성이 적으며, 많은 자료의 손실을 야기 할 수 있다 (Kim, 2009). 평균대체법은 결측치들을 관측된 자료의 평균값으로 대체하는 기법으로서, 사용이 편리하고 (Jo와 Kim, 2015), 관측수를 유지함으로서 자유도를 증가시키는 장점을 갖고 있다. 하지만 평균값이 반복적으로 대체되므로 통계량의 표준오차가 과소추정 되는 문제가 있다 (Jo와 Kim, 2015). 또한 핫덱 대체는 관측치 중 무작위 복원 추출하여 결측치를 대체하는 방법으로서 (Kang, 2013), 이 방법은 평균대체법의 추정량의 표준오차가 과소추정되는 문제를 해결할 수 있지만 표준오차를 구하기 어렵다는 문제점이 있다 (Lee 등, 2012).

본 연구에서는 기존의 결측치 대체법이 각 개체의 특성을 충분히 고려하지 않는 문제점을 보완하기 위해 반복측정 된 자료에서 개체 및 현상의 특성을 나타내는 새로운 지속성 지수 (Kang과 Kim, 1997)의 일치도와 유지도의 개념을 이용한 결측치 대체법을 제안하였다. 일치도는 '한 개체 내에서 이분된 관찰치 중 상대적으로 어떠한 값을 더 많이 갖는 정도'를 의미하며, 유지도는 '한 개체 내에서 이분된 관찰치가 연속적으로 같은 값을 갖는 정도'를 의미한다. 일치도를 의미하는 일치 지수(index of agreement)를 정의하였으며, 시점 및 값의 크고 작음을 고려하지 않는다는 지속성 지수의 한계점을 보완하기 위해 '시점에 따른 유지도' 즉, '한 개체 내에서 이분된 관찰치 중 시점이 흐름에 따라 연속적으로 특정값을 더 많이 갖는 정도'를 의미하는 유지 지수(index of maintenance)를 정의하였다. 두 지수를 이용하여 개체의 상대적 특성을 나타내는 특성도(measure of property)를 정의하고, 비슷한 특성을 갖는 개체의 관측값만을 이용하여 결측치를 대체하였다.

또한 2.5절에서는 실제 결측 자료에 대한 예제를 제시하였으며, 3장에서는 다양한 분야로의 활용성을 높이기 위해 간헐적 결측(intermittent missing)에서의 모의실험을 통하여 기존 방법과 제안하는 방법의 대체성능을 비교하였다.

2. 제안하는 방법

n개의 개체와 m개의 반복수를 갖는 반복측정 자료 y_{ij} $(i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, m)$는 다음과 같다.

$$y_{ij} = \mu + \tau_j + \epsilon_{ij}, \quad (i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, m),$$

여기서 μ는 전체평균을 나타내고, τ_j는 j번째 시점의 효과, ϵ_{ij}는 오차항을 나타낸다.

2.1. 일치 지수

시점 j에서 관측치의 평균을 절단값 즉,

$$c_j^* = \frac{1}{n} \cdot \sum_{i=1}^{n} y_{ij}, \quad (i = 1, 2, \ldots, n) \tag{2.1}$$

Table 2.2. Divided data

N	Time				
	1	2	$\cdots$	m	x_i
1	x_{11}	x_{12}	$\cdots$	x_{1m}	x_1
2	x_{21}	x_{22}	$\cdots$	x_{2m}	x_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
n	x_{n1}	x_{n2}	$\cdots$	x_{nm}	x_n
Mean	c_1^*	c_2^*	$\cdots$	c_m^*	

라 할 때 x_{ij}는

$$x_{ij} = \begin{cases} 1, & \text{if } \ y_{ij} \geq c_j^*, \\ 0, & \text{if } \ y_{ij} < c_j^*, \end{cases} \quad (i = 1, 2, \ldots, n) \tag{2.2}$$

로 정의하였다. 이때, i번째 개체의 관측치 중 1의 값을 갖는 관측치의 수인 x_i를

$$x_i = \sum_{j=1}^{m} x_{ij}, \quad (i = 1, 2, \ldots, n, \ j = 1, 2, \ldots, m) \tag{2.3}$$

라 하면, 0의 값을 갖는 관측치의 수는 $m - x_i$가 된다.

각 개체의 x_{ij}가 같은 값을 갖는 정도를 일치도라 할 때, 각각의 이분된 값인 0과 1의 개수가 비슷할수록 일치도는 낮아지게 된다. 또한 1의 개수인 x_i와 0의 개수인 $(m - x_i)$가 같을 때 가장 일치하지 않으며 $x_i(m - x_i)$의 값이 최대가 된다. 여기서 i번째 개체의 불일치 지수(d_i : index of disagreement)와 일치 지수(a_i : index of agreement)를 각각

$$d_i = \frac{2}{m^2} x_i(m - x_i), \quad (i = 1, 2, \ldots, n), \tag{2.4}$$

$$a_i = 1 - d_i, \quad (i = 1, 2, \ldots, n) \tag{2.5}$$

와 같이 정의한다. 이때 $x_i(m - x_i)$의 최대값이 $m^2/4$이므로 d_i는 0과 0.5사이의 값을 가지며, 결론적으로 a_i는 m의 값에 상관없이 일치도가 낮을수록 0.5에 가까워지며, 일치도가 높을수록 1에 가까운 값을 가진다.

2.2. 유지 지수

i번째 개체가 시점에 따라 같은 값을 연속적으로 갖는 정도를 유지도라 할 때 특정값을 연속적으로 갖는 정도를 나타내는 지시함수 I와 시점의 흐름에 따른 유지도의 변화를 알기 위해 가중값 j를 곱한 유지 지수(m_i : index of maintenance)는

$$m_i = \frac{\sum_{j=1}^{m-1} I(x_{ij+1}, \ x_{ij}) \cdot j}{\binom{m}{2}}, \quad (i = 1, 2, \ldots, n), \ I(a, b) = \begin{cases} 1, & \text{if } \ a = b = 1, \\ -1, & \text{if } \ a = b = 0, \\ 0, & \text{if } \ a \neq b \end{cases} \tag{2.6}$$

와 같이 정의한다. 이때 m_i는 시점이 흐름에 따라 0의 값을 연속적으로 많이 가질수록 -1에 가까워 지며, 1의 값을 연속적으로 많이 가질수록 1에 가까운 값을 갖는다.

Table 2.3. Measure of osmotic pressure

N	Time			
	1	2	3	4
1	183.0	249.0	345.5	449.5
2	160.5	244.5	348.5	424.5
3	149.5	254.5	353.0	c
4	a	243.0	339.5	462.5
5	208.0	248.5	350.0	448.0
6	182.5	245.5	350.5	443.5
7	184.0	250.0	b	446.0
8	192.0	251.5	342.5	452.0
9	181.5	251.5	344.0	479.0
10	174.5	254.0	345.0	d
11	213.5	235.5	341.0	454.0

2.3. 특성도

일치도와 유지도가 개체 및 현상의 특성을 나타내는 지속성 지수를 설명 한다 (Kang과 Kim, 1997). 즉, 각 개체의 시점의 흐름에 따른 유지도를 나타내는 유지 지수와 일치도를 나타내는 일치 지수가 각 개체의 특성을 설명하고 있으므로 두 지수를 곱한 특성도(measure of property; p_i)는

$$p_i = a_i \cdot m_i = \left\{1 - \frac{2}{m^2} x_i(m - x_i)\right\} \frac{\sum_{j=1}^{m-1} I(x_{ij+1},\ x_{ij}) \cdot j}{\binom{m}{2}}, \quad (i = 1, 2, \ldots, n) \tag{2.7}$$

와 같이 정의한다. 이때 각 개체의 일치도가 클수록 특성도의 절대값이 1에 가까워 지며, 시점이 흐름에 따라 1의 값이 연속적으로 많이 나올수록 양의 값을, 0의 값이 연속적으로 많이 나올수록 음의 값을 가진다.

2.4. 대체 방법

특성도를 이용한 결측치 대체방법은 다음 단계를 따른다.

Step 1: 결측이 시점 3미만에서 발생한 경우 해당 시점의 관측값들의 평균을 이용하여 대체하는 평균대체법을 이용하여 대체한다.

Step 2: $t\,(t \geq 3)$ 시점에서 첫 결측이 발생하였을 경우 대체된 자료를 포함한 시점 $(t-1)$까지 자료를 이용, 이분화한 후 각 개체의 일치 지수, 유지 지수, 특성도를 구한다.

Step 3: $(t-1)$ 시점에서의 개체의 특성도를 이용하여 결측된 개체는 가장 가까운 특성도를 갖는 개체의 관측치를 이용하여 결측치를 대체한다.

Step 4: t 시점 이후 결측이 발생한 경우 대체된 자료를 포함하여 단계 2와 3을 반복하여 결측치를 대체한다. 만일 결측이 발생한 개체의 특성도와 가장 가까운 특성도를 갖는 개체가 2개 이상일 경우 해당 개체들의 관측치의 평균을 이용하여 대체한다.

Table 2.4. Value of index and measure of property, before time 3

N	Time		a_i	m_i	p_i
	1	2			
1	1	1	1	1	1
2	0	0	1	−1	−1
3	0	1	0.5	0	0
4	1	0	0.5	0	0
5	1	1	1	1	1
6	0	0	1	−1	−1
7	1	1	1	1	1
8	1	1	1	1	1
9	0	1	0.5	0	0
10	0	1	0.5	0	0
11	1	0	0.5	0	0

Table 2.5. Value of index and measure of property, before time 4

N	Time			a_i	m_i	p_i
	1	2	3			
1	1	1	0	0.55	0.3	0.165
2	0	0	1	0.55	−0.3	−0.165
3	0	1	1	0.55	0.6	0.33
4	1	0	0	0.55	−0.6	−0.33
5	1	1	1	1	1	1
6	0	0	1	0.55	−0.3	−0.165
7	1	1	1	1	1	1
8	1	1	0	0.55	0.3	0.165
9	0	1	0	0.55	0	0
10	0	1	0	0.55	0	0
11	1	0	0	0.55	−0.6	−0.33

2.5. 실제자료를 이용한 예제

투석기 이용시, 시간의 경과에 따른 삼투압의 변화를 11명의 피험자를 대상으로 4회 측정한 자료 (Kang과 Kim, 1997)에서 임의로 결측치를 발생시킨 후 제안한 방법을 적용하였다 (Table 2.3).

첫 번째 단계로 결측치 a의 경우 시점 3미만에서 발생한 결측치이므로 평균대체법을 이용하여($a = 182.9$) 대체를 한다. 두 번째 단계로 시점 3에서 발생한 결측치 b를 대체하기 위해 시점 2까지의 자료를 이용하여 각 시점의 절단값인 $c_1^*(= 182.9)$, $c_2^*(= 247.9)$를 구하고 자료를 이분화 한 후 일치 지수와 유지 지수, 특성도를 다음과 같이 구한다 (Table 2.4). 세 번째 단계로 p_7의 특성도가 특성도 p_1, p_5, p_8과 같은 1의 값을 갖고 있으므로 결측치 $b(= y_{73})$는 $(y_{13} + y_{53} + y_{83})/3 = 346.0$으로 대체를 한다. 네 번째 단계로 시점 4에서 발생한 결측치 c, d를 대체하기 위해 시점 3까지의 자료를 이용하여 각 시점의 절단값인 $c_1^*(= 182.9), c_2^*(= 247.9), c_3^*(= 345.9)$를 구하고 위와 같은 방법으로 일치 지수와 유지 지수, 특성도를 구한다 (Table 2.5).

특성도 p_3는 0.33으로 첫 번째와 여덟 번째 개체의 특성도 0.165와 가장 근사한 값을 갖고 있으므로 결측치 $c(= y_{34})$는 $(y_{14} + y_{84})/2 = 450.75$으로 대체하고, 특성도 p_9와 p_{10}은 0으로 같은 값을 갖고 있으므로 결측치 $d(y_{104})$는 $y_{94} = 479.0$으로 대체를 한다 (Table 2.6).

Table 2.6. After all missing values are imputed

N	Time			
	1	2	3	4
1	183.0	249.0	345.5	449.5
2	160.5	244.5	348.5	424.5
3	149.5	254.5	353.0	$c = 450.75$
4	$a = 182.9$	243.0	339.5	462.5
5	208.0	248.5	350.0	448.0
6	182.5	245.5	350.5	443.5
7	184.0	250.0	$b = 346.0$	446.0
8	192.0	251.5	342.5	452.0
9	181.5	251.5	344.0	479.0
10	174.5	254.0	345.0	$d = 479.0$
11	213.5	235.5	341.0	454.0

Table 3.1. Basic Statistics about original data

Data	Mean ± STD	Min ∼ Max	p-value
A	20.42 ± 10.35	7.55 ∼ 39.25	0.0035
B	0.78 ± 0.24	0.41 ∼ 1.22	0.0347
C	25.85 ± 22.66	2.55 ∼ 71.00	0.0762
D	0.75 ± 0.22	0.11 ∼ 1.20	0.1376

3. 모의실험 및 결과

제안하는 방법의 대체성능을 기존 방법인 평균대체법, 핫덱 대체와 비교하기 위하여 Monte Carlo 모의실험을 시행하였다.

11개의 개체($n = 11$)에 대해 7회 반복측정($m = 7$)한 자료를 통해 4개의 각각 다른 p-값을 갖는 데이터 셋을 임의로 구축하였다. 이때 데이터의 기초통계량과 p-값은 Table 3.1과 같다.

4개의 자료는 p-값이 각각 0.0035, 0.0347, 0.0762, 0.1376으로 α가 0.05일 때 기각 및 채택 여부를 비교할 수 있도록 하였다. 자료 B와 D는 평균과 표준편차가 각각 0.78, 0.24, 0.75, 0.22로 값이 작으며, 자료 A와 C는 평균과 표준편차가 각각 20.42, 10.35, 25.85, 22.66으로 큰 값을 가졌다.

완전임의결측 가정 하에 5%, 10%의 결측을 각각 발생 시킨 후 평균대체법, 핫덱 대체, 제안하는 방법을 통해 결측치를 대체하였다. 또한 간헐적 결측의 경우 대표적으로 사용하는 Markov Chain Monte Carlo 방법을 추가적으로 이용하였다. Markov Chain Monte Carlo 방법은 정상분포에 안정화하기 위해 충분히 긴 마코브 연쇄를 만든 후 자료가 다변량 정규분포를 따른다는 가정하에 베이지안 추론에 자료증대 방법을 통하여 l개의 완전한 자료집합을 만든 후 완전한 자료집합으로부터 구한 추정치들의 평균을 이용하여 결측치를 보정하는 방법이다 (Lee, 2008). 결측 대체한 자료를 이용하여 repeated ANOVA 검정을 실시하는 과정을 1,000회 반복하였다.

대체방법을 비교하기 위한 지수인 원자료의 검정결과의 기각 여부에 따른 대체된 자료의 검정결과가 원자료와 같은 비율(rate of rejected or accepted; RRA), 대체된 자료의 p-값과 원자료의 p-값 간 차이의 제곱합(sum of square of p-value's difference; SSP)을 다음과 같이 정의하였으며, 추가적으로 정규화 제곱근 평균오차(normalized root mean squared error; NRMSE)을 이용하였다. 결측치 대체의 유의성을 알기 위해 대체하지 않은 결측자료의 검정결과를 추가하여 모의실험 결과를 각각 Tables 3.2–

Table 3.2. Rate of rejected or accepted (RRA, $\alpha = 0.05$)

Data	Missing	Mean[a]	Hotdeck[b]	MCMC[c]	Subject[d]	Missing[e]
A	5%	1.000	0.991	0.898	0.998	0.777
	10%	1.000	0.912	0.743	0.988	0.471
B	5%	0.897	0.620	0.703	0.813	0.539
	10%	0.872	0.427	0.541	0.819	0.284
C	5%	0.699	0.835	0.747	0.733	0.740
	10%	0.476	0.778	0.678	0.515	0.743
D	5%	0.931	0.965	0.903	0.881	0.829
	10%	0.805	0.952	0.812	0.736	0.856

a: unconditional mean, *b*: Hot-Deck, *c*: Markov Chain Monte Carlo, *d*: suggest method, *e*: missing data.

3.4로 정리하였다.

$$\text{RRA} = \frac{[\text{원자료가 기각일 때 대체된 자료가 기각한 횟수(반대의 경우도 포함)}]}{1000}, \tag{3.1}$$

$$\text{SSP} = \sum_{k=1}^{1000} (p_k - p)^2, \tag{3.2}$$

$$\text{NRMSE} = \frac{1}{x'_{max} - x'_{min}} \left\{ \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(x_{ij} - x'_{ij})^2}{M} \right\}^2 \tag{3.3}$$

여기서, p_k는 대체된 자료의 p-값, p는 원자료의 p-값, x_{ij}는 실제값, x'_{ij}는 추정값, M은 결측치 수, $x'_{\max}$는 추정치 중 최대값, $x'_{\min}$는 추정치 중 최솟값이다. RRA는 그 값이 클수록 옳은 검정 결과를 도출할 확률이 높음을 의미하며, SSP는 그 값이 작을수록 대체된 자료와 원자료의 통계량이 비슷함을 의미한다. 또한 NRMSE는 그 값이 작을수록 원자료와 대체된 자료의 값이 비슷함을 의미한다.

Table 3.2는 대체법의 자료에 따른 RRA를 나타낸 표이다. 우선 자료 A에서 5%결측일 때 Markov Chain Monte Carlo가 0.898로 가장 작은 값을 가졌으며, 핫덱 대체는 0.912로 그 뒤를 이었다. 또한 제안하는 방법이 0.998, 평균대체법이 1.000으로 대체로 우수한 대체성능을 보였으며, 특히 평균대체법이 매우 우수한 성능을 보였다. 10%결측일 때 또한 같은 결과를 얻을 수 있었다. 자료 B에서 5%결측일 때 핫덱 대체와 Markov Chain Monte Carlo가 각각 0.620, 0.703으로 작은 값을 가졌으며, 제안하는 방법과 평균대체법이 각각 0.813, 0.897로 큰 값을 가졌다. 즉, 이 경우 또한 평균대체가 가장 우수한 성능을 보였지만, p-값이 0.0035인 경우와는 달리 핫덱 대체가 가장 나쁜 결과를 보였다. 10%결측일 때 또한 같은 결과를 얻을 수 있었다.

자료 C에서 5%결측일 때 평균대체법과 제안하는 방법이 각각 0.699, 0.733으로 매우 작은 값을 가졌다. 또한 Markov Chain Monte Carlo는 0.747을 가짐으로서 이전과 큰 차이가 없이 작은 값을 가졌다. 하지만 핫덱 대체가 0.835로 가장 우수한 대체성능을 보임으로써 p-값이 0.05보다 작은 경우와 많이 상이한 결과를 보였다. 10%결측일 때 또한 같은 결과를 얻을 수 있었다. 자료 D에서 5%결측일 때 제안하는 방법이 0.881로 가장 작은 값을 가졌으며, Markov Chain Monte Carlo가 0.903으로 그 뒤를 이었다. 또한 평균대체법과 핫덱 대체가 각각 0.931, 0.965로 전반적으로 모든 대체법이 우수한 성능을 보였다. 10%결측일 때 핫덱 대체가 0.952로 가장 큰 값을 가졌으며, 제안하는 방법이 0.736으로 가장 작은 값을 가져 5%결측일 때와 같은 결과를 가졌지만, Markov Chain Monte Carlo와 평균대체법이 각각 0.812, 0.805로 다른 결과를 보였다.

Table 3.3은 대체법의 자료에 따른 SSP를 나타낸 표이다. 우선 자료 A에서 5%결측일 때 평균대체법

Table 3.3. Sum of square of p-value's difference (SSP)

Data	Missing	Mean[a]	Hotdeck[b]	MCMC[c]	Subject[d]	Missing[e]
A	5%	0.0030	0.1347	8.5929	0.0314	7.6975
	10%	0.0060	1.4048	24.2627	0.5889	56.8484
B	5%	0.2401	2.6608	7.8780	1.0075	22.0930
	10%	0.4628	11.3326	15.4880	1.5115	78.7400
C	5%	0.9926	4.4582	28.1816	4.7000	51.5374
	10%	1.9639	12.5369	47.9938	4.9482	133.2092
D	5%	2.2311	9.4418	14.6309	4.5542	39.5102
	10%	5.0762	25.3667	24.7022	8.4723	92.7700

a: unconditional mean, *b*: Hot-Deck, *c*: Markov Chain Monte Carlo, *d*: suggest method, *e*: missing data.

Table 3.4. Normalized root mean squared error (NRMSE)

Data	Missing	Mean[a]	Hotdeck[b]	MCMC[c]	Subject[d]
A	5%	3.60	3.97	1.25	2.14
	10%	1.14	1.02	0.50	0.78
B	5%	5.19	1.58	4.57	0.01
	10%	2.95	0.68	0.53	0.93
C	5%	10.56	4.35	1.36	2.44
	10%	1.47	1.08	0.47	1.01
D	5%	6.88	1.41	1.15	3.12
	10%	1.90	0.66	0.54	0.75

a: unconditional mean, *b*: Hot-Deck, *c*: Markov Chain Monte Carlo, *d*: suggest method.

이 0.003으로 가장 작은 값을 가졌다. 또한 제안하는 방법, 핫덱 대체가 각각 0.0314, 0.1347로 작은 값을 가졌다. 하지만 Markov Chain Monte Carlo는 8.5929로 매우 큰 값을 가졌다. 즉, Markov Chain Monte Carlo가 가장 나쁜 대체성능을 보였으며, 평균대체법이 가장 우수한 대체성능을 가짐을 알 수 있다. 10%결측일 때 또한 같은 결과를 얻을 수 있었다. 자료 B에서 평균대체법이 0.2401로 가장 작은 값을 가졌으며, 제안하는 방법 또한 1.0075로 작은 값을 가졌다. 하지만 핫덱 대체와 Markov Chain Monte Carlo가 각각 2.6608, 7.8780으로 자료 A에 비해 핫댁 대체의 값이 급격히 커짐을 알 수 있다. 10%결측일 때 또한 같은 결과를 얻을 수 있었다.

자료 C에서 평균대체법이 0.9926으로 가장 작은 값을 가졌으며, 제안하는 방법이 4.7000으로 그 뒤를 이었다. 또한 핫덱 대체와 Markov Chain Monte Carlo가 각각 4.4582, 28.1816으로 자료 A, B와 같은 양상을 보였다. 10%결측일 때 평균대체법이 1.9639로 가장 작았으며, Markov Chain Monte Carlo가 47.9938로 가장 큰 값을 가져 5%결측과 같은 결과를 보였지만, 제안하는 방법과 핫덱 대체가 각각 4.9482, 12.5369로 다른 결과를 보였다. 자료 D에서 5%결측일 때 평균대체법이 2.2311으로 가장 작은 값을 가졌으며, 제안하는 방법이 4.5542로 그 뒤를 이었다. 또한 핫덱 대체와 Markov Chain Monte Carlo가 각각 9.4418, 14.6309로 매우 큰 값을 가졌다. 10%결측일 때 평균대체법이 5.0762로 가장 작은 값을 가졌으며, 제안하는 방법이 8.4723으로 그 뒤를 이어, 5%결측일 때와 같은 결과를 보였지만, Markov Chain Monte Carlo와 핫덱 대체가 각각 24.7022, 25.3667로 다른 결과를 보였다.

Table 3.4는 대체법의 자료에 따른 NRMSE를 나타내며, 우선 자료 A에서 5%결측일 때 Markov Chain Monte Carlo가 1.25로 가장 작은 값을 가졌으며, 제안하는 방법이 2.14로 그 뒤를 이으며 작은 값을 가졌다. 평균대체법과 핫덱 대체는 각각 3.60, 3.97로 약간 큰 값을 가졌다. 10%결측일 때 Markov Chain Monte Carlo가 0.50으로 가장 작은 값을 가졌으며, 제안하는 방법이 0.78로 그 뒤를 이어 5%결

측일 때와 같은 결과를 보였지만, 핫덱 대체와 평균대체법이 각각 1.02, 1.14로 다른 결과를 보였다. 즉, Markov Chain Monte Carlo가 원자료와 가장 근사한 값을 가졌다. 자료 B에서 5%결측일 때 제안하는 방법이 0.01로 매우 작은 값을 가졌으며, 핫덱 대체가 1.58로 작은 값을 가졌다. 자료 A와는 달리 Markov Chain Monte Carlo가 4.57로 큰 값을 가졌으며, 평균대체법이 5.19로 가장 큰 값을 가졌다. 10%결측일 때 Markov Chain Monte Carlo가 0.53으로 가장 작은 값을 가졌으며, 핫덱 대체와 제안하는 방법이 각각 0.68, 0.93으로 뒤를 이어 5%결측일 때와 매우 다른 결과를 보였지만 평균대체법은 2.95로 가장 큰 값을 가져 같은 결과를 보였다. 즉, 평균대체법이 원자료와 가장 상이한 값을 가졌다.

자료 C에서 5%결측일 때 Markov Chain Monte Carlo가 1.36으로 가장 작은 값을 가졌으며, 제안하는 방법과 핫덱 대체가 각각 2.44, 4.35로 그 뒤를 이었다. 평균대체법은 10.56으로 원자료와 매우 상이한 값을 가졌다. 10%결측일 때 또한 같은 결과를 얻을 수 있었다. 자료 D에서 5%결측일 때 Markov Chain Monte Carlo와 핫덱 대체가 각각 1.15, 1.41로 매우 작은 값을 가졌으며, 제안하는 방법이 3.12로 그 뒤를 이었다. 평균대체법은 6.88로 큰 값을 가졌다. 10%결측일 때 또한 같은 결과를 얻을 수 있었다.

4. 결론 및 고찰

임상시험이 진행되며 발생한 결측치를 대체하는 방법에 따라 임상결과가 다를 수 있기 때문에 대체방법에 대한 연구가 활발히 이루어지고 있다. 본 연구에서는 특성도를 정의하고, 이를 통하여 결측치를 대체하는 새로운 방법을 제안하였다. 또한 모의실험을 통해 결측 자료를 대체한 후 RRA, SSP, NRMSE를 이용하여 제안하는 방법과 기존 방법들 간 대체 성능을 비교하였다.

모의실험 결과 p-값이 0.05보다 작은 자료에서 5%결측일 때 평균대체법은 다른 대체법에 비해 NRMSE가 컸지만, 대체로 RRA와 SSP가 매우 좋은 결과를 가졌다. 이는 모의실험에서 좋은 검정결과를 보였지만 원자료와 상이한 값을 추정함으로서 검정방법에 따라 다른 결과를 초례할 수 있다. 핫덱 대체와 Markov Chain Monte Carlo는 상대적으로 RRA가 작았으며, SSP가 큰 값을 가졌다. 특히 핫덱 대체는 SSP와 NRMSE에 비해 RRA가 급격히 나빠져 p-값을 증가시키는 성향이 있음을 알 수 있었다. 이는 임상실험에서 약효가 있는 시약에 대해 잘못된 결과를 초례할 수 있다. 제안한 방법은 RRA, SSP 그리고 NRMSE 모두 충분히 제어함으로서 우수한 대체성능을 보였다. 또한 10%결측일 때 역시 같은 결과를 얻음으로서 결론적으로 p-값이 0.05보다 작은 경우 제안한 방법이 우수한 대체성능을 가졌다.

p-값이 0.05보다 큰 자료에서 5%결측일 때 평균대체법과 Markov Chain Monte Carlo는 p-값이 0.05보다 작을 때와 같은 모습을 보였으며, 특히 p-값이 0.05보다 작은 자료에 비해 평균대체법의 RRA의 값이 매우 작아지는 모습을 보였다. 또한 핫덱 대체는 RRA가 매우 좋았지만 그에 반해 SSP와 NRMSE가 충분히 제어되지 않아 p-값을 증가시키는 성향이 뚜렷하였다. 제안한 방법은 평균대체와 마찬가지로 RRA가 나빠짐을 알 수 있었다. 또한 SSP의 값이 평균대체법과 비슷한 양상을 보여 아쉬운 결과를 보였다. 하지만 NRMSE가 최대 3.12로 다른 대체법에 비해 충분히 제어됨으로서 개체의 특성을 충분히 반영하고 있음을 보였다. 10%결측일 때 핫덱 대체와 Markov Chain Monte Carlo는 SSP가 매우 크며, 모든 대체법의 RRA가 이전의 다른 자료에 비해 낮았다. 특히 평균대체법과 제안한 방법의 RRA가 매우 낮았다.

결론적으로 제안하는 방법이 p-값이 0.05보다 작을 때 가장 좋은 모습을 보였으며, 이는 임상시험에서 많이 발생하는 중도탈락(dropout missing)의 경우에서도 유용할 것으로 생각된다. 또한 NRMSE를 통해 다른 대체법에 비해 개체의 특성을 고려함을 알 수 있었지만, p-값이 0.05보다 큰 경우 RRA와 SSP를 충분히 제어하지 못해 아쉬운 모습을 보였다. 이는 각 시점의 평균값을 절단값으로 이용함으로

서 이분된 자료가 개체의 특성을 충분히 설명하는데에 한계점이 있는 것으로 생각된다. 또한 시점 3미만의 결측에 대해 평균대체법을 이용함으로서 평균대체법과 RRA와 SSP가 비슷한 추세를 보인 것으로 생각된다. 개체의 특성을 조금 더 명확하게 나타내기 위해 관측치의 절단값이 되는 c_j^*와 최초 결측이 발생한 시점 t가 3이상일 때 제안하는 방법의 대체성능에 대한 연구가 필요할 것으로 생각된다.

References

Jo, B. and Kim, D. (2015). Comparison of single imputation methods in 2×2 cross-over design with missing observations, *Korean Journal of Applied Statistics*, **28**, 529–540.

Kang, H. and Kim, B. (1997). A new measure of tracking in repeated measurement data, *The Korean Communications in Statistics*, **10**, 189–201.

Kang, S. (2013). *Medical Statistics for New Medicine Development*, Freeacademy, Seoul.

Kim, D. (2009). Quantitative methods for missing value in the study of public administration - focused on the application of SPSS/MVA, *Korean Comparative Government Review*, **13**, 177–196.

Lee, S. (2008). Conjugation plan of Proc MI, *Industrial Science Research*, **26**, 35–41.

Lee, S., An, J., and Kim, S. (2012). Comparison of imputation methods for GARCH model, *Bulletin of the Natural Sciences*, **26**, 35–42.

특성도를 이용한 결측치 대체방법

김형주[a] · 김동재[a,1]

[a]가톨릭대학교 의생명 · 건강과학과

(2017년 4월 5일 접수, 2017년 5월 5일 수정, 2017년 5월 5일 채택)

요 약

임상시험에서 어떻게 결측치를 다룰 것인가 하는 것은 큰 문제이다. 주로 주분석에서 사용하는 ITT원칙은 결측치가 어떠한 메커니즘을 따른다는 가정 하에 결측치를 대체 하지만 가정에 대한 타당성이 불확실한 문제가 있다. 즉, 올바른 결측치 대체방법은 매우 중요하다. 본 연구에서는 Kang과 Kim (1997)이 제안한 일치도와 유지도의 개념을 이용하여 새로운 결측치 대체방법을 제안하였다. 또한 실제자료를 이용하여 예제를 제시하고 Monte Carlo 모의실험을 통하여 기존방법과 대체 성능을 비교하였다.

주요용어: 결측치 대체, 일치도, 유지도, 임상시험

[1]교신저자: (06591) 서울 서초구 반포대로 222, 가톨릭대학교 의생명 · 건강과학과.
E-mail: djkim@catholic.ac.kr