

연구논문

다변량 보조변수가 있는 경우에 갈퀴법을 이용한 나이브 성향점수보정 추정량에 관한 연구*

이유진** · 신기일***

무응답이 있는 표본조사에서 평균 추정량으로 성향점수보정 추정량이 흔히 사용된다. 그러나 성향점수보정 추정량은 흔히 무시할 수 없는 무응답이라 부르는 MNAR(missing not at random) 무응답을 적절히 처리하지 않으면 편향이 발생하는 것으로 알려져 있다. 최근 Chung & Shin(2022)은 무정보적 표본설계(non-informative sampling)에서 무시할 수 없는 무응답이 발생할 때 나이브 성향점수보정(naive propensity score adjusted: PSA) 추정량의 정확성을 향상하는 방법을 제안하였다. 특히 세부 층을 이용한 사후층화 방법을 기존의 나이브 성향점수보정 추정량에 적용하여 추정치의 정확성을 향상시켰다. 이 방법은 보조변수가 단변량인 경우 효과적인 것으로 나타났으며 이변량인 경우도 효과가 있는 것으로 나타났다. 또한 Yeo & Shin(2023)은 이중 사후층화보정 방법을 이용하여 나이브 성향점수보정 추정량의 정확성을 높였다. 본 연구에서는 보조변수가 다변량인 경우에 갈퀴법(raking ratio method)을 이용하여 기존의 사후층화 방법을 개선한 새로운 사후층화 방법을 제안하였다. 특히 초모집단모형이 로그-선형모형일 때 본 연구에서 제안한 방법이 기존의 방법에 비해 우수한 것을 모의실험을 통해 확인하였다.

주제어: 사후층화, 응답확률, 초모집단모형, 무정보적 표본설계

* 이 연구는 2024년 한국외국어대학교 교내연구비 지원을 받아 수행되었음.

** 한국외국어대학교 통계학과 박사과정, 국민건강보험공단 비급여관리실 주임연구원(leeeyj4612@gmail.com), 제1저자.

*** 한국외국어대학교 통계학과 교수(keyshin@hufs.ac.kr), 교신저자.

I. 서론

표본조사에서 발생하는 무응답은 추정의 정확성을 떨어뜨리므로 이를 적절히 처리해야 한다. 무응답은 결측 메커니즘(missing mechanism)에 따라 크게 세 가지로 나누어진다(Little & Rubin 2002). 먼저 무응답이 보조변수와 관심변수에 무관하게 완전 랜덤으로 발생하는 MCAR(Missing Completely At Random)과, 관심변수와 무관하고 보조변수에만 영향을 받는 경우인 MAR(Missing At Random)이 있다. 두 결측 메커니즘을 따르는 무응답은 처리에 어려움이 없어 흔히 무시할 수 있는 무응답(ignorable nonresponse)이라 부른다. 반면 MNAR(Missing Not At Random)은 무응답이 관심변수에 영향을 받는 경우로, 이를 적절히 처리하지 않으면 평균 추정량으로 사용하는 성향점수보정 추정량에 편향이 발생하는 어려움이 있다. 즉 MNAR 무응답이 발생하면 관심변수를 사용하여 응답확률을 추정해야 하지만, 얻어진 관심변수 자료에 무응답이 발생하여 관심변수를 사용한 응답확률 추정이 어렵다. 이에 MNAR을 따르는 무응답을 무시할 수 없는 무응답(Non-ignorable Nonresponse: NN)이라 부른다. 최근 무시할 수 없는 무응답에 관한 연구가 활발히 진행되고 있다(Morikawa et al. 2017; Berk et al. 2016; Morikawa & Kim 2018). 이에 따라 실질적으로 사용 가능한 추정량은 나이브 성향점수보정(naive PSA) 추정량이며 이 추정량은 편향 추정량이다. 나이브 성향점수보정 추정량의 정의는 다음과 같다.

$$\hat{Y} = \frac{1}{N} \sum_{i=1}^n w_i \frac{R_i}{\hat{p}_i} y_i \quad (1)$$

여기서 N 은 모집단 수, n 은 표본 수, w_i 는 표본 가중치, R_i 는 지시변수로 응답이면 1, 무응답이면 0을 갖는다. 또한 y_i 는 관심변수이고, $\hat{p}_i$ 는 보조변수를 이용하여 추정된 응답확률 또는 성향점수이다. 여기서 $\hat{p}_i$ 는 MAR 가정을 사용하여 추정한다. 나이브 성향점수보정 추정량에 관한 자세한 내용은 Kim & Riddle(2012)와 Riddles et al.(2016)을 살펴보기 바란다.

본 연구에서는 다변량 보조변수가 있는 경우, MAR 가정에서 추정된 $\hat{p}_i$ 의 정확성을 높여 최종적으로 나이브 성향점수보정 추정량의 정확성을 향상시키는 방법을 연구한다. 이를 위해 잘 알려진 갈퀴법을 사용하며, 갈퀴법 사용은 가중치와 응답확률 추정의 정확성 향상에 필요한 세부 층의 개수를 충분히 크게 할 수 있다는 장점이 있다. 또한, 최종 무응답 보정 가중치인 $w_i^F = w_i / \hat{p}_i$ 도 갈퀴법을 사용하여 보정함으로써 추정의 정확성을 향상시키는 방법을 제안한다. 물론 보조변수는 층화와 응답확률 추정에 동시에 사용될 수 있다.

본 논문의 구성은 다음과 같다. 2장에서 기존의 세부 층 구성 방법을 설명하고, 3장에서는 새롭게 제안한 세부 층 구성 방법과 제안한 레이킹 나이브 PSA 추정량을 설명한다. 4장에서는 모의실험을 통해 제안한 방법의 우수함을 확인한 후, 5장의 결론으로 이어진다.

II. 기존의 방법

1. 단변량에서 설계 가중치 보정을 위한 세부 층 구성 방법

이 장에서는 Chung & Shin(2022)과 Lee & Shin(2024, b)에서 사용한 세부 층 구성 방법을 간단히 설명한다. 먼저 정해진 세부 층의 개수를 L 이라 하고 모집단에 포함된 보조변수 x_i 의 분위수를 층 경계로 사용한다. 그러면 h 번째 세부 층의 모집단 수 N_h , 배정된 표본 수 n_h , 그리고 최종 응답자료 수 r_h 가 얻어지며 이때 h 번째 세부 층의 최종 가중치는 다음의 식으로 얻어진다.

$$w_h^D = \frac{N_h}{n_h} \times \frac{n_h}{r_h} = \frac{N_h}{r_h} \quad (2)$$

따라서 h 세부 층에 포함된 i 번째 가중치 w_{hi}^D 는 동일하게, 즉 $w_{hi}^D = w_h^D$ 가 되므로 세부 층 가중치의 합은 $\sum_{h=1}^L \sum_{i=1}^{r_h} w_{hi}^D = \sum_{h=1}^L (N_h / r_h) \times r_h = N$ 이 된다.

2. 단변량에서 응답확률 보정을 위한 세부 층 구성 방법

성향점수보정 추정량의 응답확률을 구할 때 가장 흔히 사용하는 방법은 로지스틱 회귀모형이다. 물론 비모수적 회귀방법으로 응답확률을 추정할 수 있다. Da Silva & Opsomer(2006; 2009)는 비모수적 함수추정 방법을 이용하여 응답확률을 추정하였으며 Da Silva & Opsomer(2009)는 로지스틱 추정 방법보다 비모수적 함수추정 방법이 우수하다는 것을 보였다. 다만 로지스틱 회귀모형이 자주 사용되므로 본 연구에서는 로지스틱 회귀모형을 사용한다. 먼저 식(1)의 확률변수 R_i 는 응답 유무의 지시함수로 다음과 같이 정의한다.

$$R_i = \begin{cases} 1, & \text{개체 } i \text{가 응답한 경우,} \\ 0, & \text{개체 } i \text{가 응답하지 않은 경우} \end{cases}$$

정의에 따라 확률변수 R_i 는 베르누이분포를 따르며 응답확률 $p_i = \Pr(R_i = 1)$ 가 추정된다. MNAR 무응답은 p_i 가 관심변수 y_i 의 함수이지만, y_i 를 사용할 수 없어 실질적으로 MAR 무응답 가정을 사용한다. 이에 p_i 의 추정은 보조변수인 x_i 에 의해 이루어지며, 따라서 $\hat{p}_i$ 는 x_i 의 함수로 구한다. 결국 $\hat{p}_i = \widehat{\Pr}(R_i = 1|x_i)$ 로 구하며 보조변수가 단변량인 경우에서 다음의 로지스틱 회귀모형으로 추정한다.

$$\ln\left(\frac{\hat{p}_i}{1 - \hat{p}_i}\right) = \hat{\alpha}_0 + \hat{\alpha}_1 x_i \quad (3)$$

Sim & Shin(2021)의 결과에 의하면 식(3)에서 얻어진 응답확률을 1절에서 설명한 세부 층 결과로 보정한 보정 응답확률이 식(2)을 사용하여 얻어진 응답확률보다 그 성능이 우수한 것을 보였다. 즉, 로지스틱 응답확률과 세부 층을 이용한 보정 응답확률이 우수한 결과를 보인다. 물론 Sim & Shin(2021)에서는 정보적 표본설계하에서가 아니라 무정보적 표본설계하에서 얻어진 결과이며 이 방법은 Chung & Shin(2022)에서도 사용되었다. 또한 정보적 표본설계하에서 표본 포함확률 또는 설계가중치를 위한 세부 층과 응답확률 보정을 위한 세부 층을 각각 사용한 이중 사후 층화 방법은 Yeo & Shin(2023)에서 연구되었다. 다만 Chung & Shin(2022)에서는

이변량 보조변수인 경우에 데카르트 곱(Cartesian Product)을 이용하여 세부 층을 구성하였는데, 이 경우 차원의 저주가 일어날 수 있어 충분히 큰 개수의 세부 층을 구성하는 것이 어렵다는 단점이 있다.

3. 응답확률 보정을 위한 이변량 구성 방법

Chung & Shin(2022)은 주어진 이변량 보조변수를 이용하여 세부 층을 나누기 위해 각 보조변수 x_1 과 x_2 의 분위수를 구한 후, 각각의 분위수를 이용하여 데카르트 곱을 만들어 이변량의 세부 층을 구성하였다. 예를 들어 y_i 의 보조변수 x_{1i} 가 h_1 층에 포함되고, x_{2i} 가 h_2 층에 포함되면 i 번째 자료는 (h_1, h_2) 세부 층에 포함된다.

이제 x_1 이 L_1 개의 세부 층으로 나누어지고 x_2 가 L_2 개의 세부 층으로 나누어지면, 총 $L = L_1 \times L_2$ 개의 세부 층이 구성되며, 구성된 (h_1, h_2) 세부 층의 모집단 수를 N_{h_1, h_2} , 최종 응답자료 수를 r_{h_1, h_2} 라 하면 1절에서 사용한 방법을 적용하여 최종 가중치가 구해진다. 즉, (h_1, h_2) 세부 층에 포함된 i 자료의 가중치는 $w_{i \in (h_1, h_2)}^{C-S} = w_{(h_1, h_2)}^{C-S} = N_{h_1, h_2} / r_{h_1, h_2}$ 이 된다. 이 가중치는 가중치의 합이 모집단 수가 되는 것을 쉽게 알 수 있다. 다만 이 방법은 L_1 과 L_2 를 크게 하면 $L = L_1 \times L_2$ 개의 층에 포함되는 자료 수가 없는 경우가 발생할 수 있고 따라서 가중치 합이 모집단 수와 일치하지 않을 수 있어 이를 적절히 처리해야 한다.

III. 제안된 세부 층 구성 방법과 레이킹 나이브 PSA 추정량

1. 갈퀴법(raking ratio method)

세부 층을 구성할 때 각 보조변수 세부 층의 개수는 차원의 저주로 인해 단변량인 경우처럼 세부 층 개수를 크게 할 수가 없다. 이는 각 보조변수 세부 층의 개수가 큰 경우, 일부 세부 층에 응답자료가 없거나 매우 소수의 응답자료가 포함되어 가중치의 변동이 클 가능성이 있기 때문이다. 이를 해결하는 방법으로 본 연구에서는 갈퀴법 사용을 제안한다.

갈퀴법은 반복비례가중법(iterative proportional weighting method)이라 부르기도 하며, 표본설계 당시 고려되지 않았던 모집단 정보를 위해 사용하는 방법으로 새로운 모집단 정보가 얻어졌으나 층별 정보는 얻어지지 않고 주변 정보만 있는 경우에 사용한다.

이 방법을 간단히 설명하면 다음과 같다. 먼저 알려진 행과 열 합(참값)을 N_{+j} , N_{i+} 이라 하고 $\widehat{N}_{ij}$ 를 i 열 j 행 자료에서 얻어진, 추정된 모집단 수라 하자. 또한, 행 합과 열 합을 $\widehat{N}_{+j}$, $\widehat{N}_{i+}$ 라 하자. 예를 들어 변수 A 의 층이 3개로 구성되어 있고, 변수 B 의 층이 4개로 구성되어 있을 때 자료의 구성은 다음과 같다.

구분	변수 A			합계	참값	
	A_1	A_2	A_3			
변수 B	B_1	$\widehat{N}_{11}$	$\widehat{N}_{21}$	$\widehat{N}_{31}$	$\widehat{N}_{+1}$	N_{+1}
	B_2	$\widehat{N}_{12}$	$\widehat{N}_{22}$	$\widehat{N}_{32}$	$\widehat{N}_{+2}$	N_{+2}
	B_3	$\widehat{N}_{13}$	$\widehat{N}_{23}$	$\widehat{N}_{33}$	$\widehat{N}_{+3}$	N_{+3}
	B_4	$\widehat{N}_{14}$	$\widehat{N}_{24}$	$\widehat{N}_{34}$	$\widehat{N}_{+4}$	N_{+4}
합계		$\widehat{N}_{1+}$	$\widehat{N}_{2+}$	$\widehat{N}_{3+}$	$\widehat{N}$	
참값		N_{1+}	N_{2+}	N_{3+}		N

세부 층의 개수가 작은 경우, 모든 층에 응답자료가 존재하고 $\widehat{N}_{ij}$ 가 구해지면 추정된 모집단 수와 참값은 일치하게 된다. 따라서 갈퀴법을 사용할 필요가 없다. 반면 세부 층의 개수가 많아서 일부 세부 층에 최종 응답자료가 없다면 해당 층에서 $\widehat{N}_{ij}$ 는 구해지지 않으며, 따라서 추정된 행과 열의 합인 $\widehat{N}_{+j}$, $\widehat{N}_{i+}$ 이 N_{+j} , N_{i+} 와 달라지므로 이를 보정해야 한다. 물론 i 열 j 행에 최종 응답자료가 없다면 초기치로 $\widehat{N}_{ij}=0$ 을 사용한다.

이제 초기치를 $N_{ij}^{(0)} = \widehat{N}_{ij}$ 이라 하고 다음의 식을 이용하여 반복적으로 행과 열을 갱신하면 두 주변 합의 값은 같아지게 된다.

$$N_{ij}^{(2t-1)} = N_{ij}^{(2t-2)} \times \frac{N_{+j}}{\widehat{N_{+j}}}$$

$$N_{ij}^{(2t)} = N_{ij}^{(2t-1)} \times \frac{N_{i+}}{\widehat{N_{i+}}}$$

이때 t 는 갈퀴법 시행 횟수를 의미한다. 위의 식에 따르면 $2t-2$ 번 시행 후의 N_{ij} 를 행의 합인 N_{+j} 의 값으로 맞춰 $2t-1$ 번 시행 후의 $N_{ij}^{(2t-1)}$ 를 구하고, 이를 이용하여 열의 합인 N_{i+} 의 값으로 맞춰 $2t$ 번 시행의 결과를 도출하는 것이다.

이 방법은 보조변수의 수가 3개 이상인 경우로도 쉽게 확장 가능하며 SAS 또는 R에서 이미 프로그램화되어 있다.

2. 제안된 세부 층 구성 방법

Bethlehem(2020)은 다양한 방법이 응답확률 추정의 정확성 향상에 사용될 수 있다고 설명하였으며, 다양한 방법 중에서 갈퀴법이 사용될 수 있다고 설명하였다. 본 연구에서는 Chung & Shin(2022)에서 사용한 이변량 보조변수의 세부 층 구성 방법과 Yeo & Shin(2023)의 이중 세부 층 구성 방법을 확장하기 위해 갈퀴법을 사용한 이중 사후보정 방법을 제안한다. 즉 다변량 보조변수에서 설계 가중치를 갈퀴법으로 보정하고, 또한 최종 다변량 응답자료를 이용하여 얻어진 응답확률도 갈퀴법을 이용하여 보정한다.

1) 설계 가중치 보정을 위한 다변량 세부 층 구성 방법

모집단 자료가 포함된 2개의 연속형 보조변수 x_{1i}, x_{2i} 의 분위수를 각각 구한다. 이때 x_{1i} 의 세부 층 개수를 L_1^P , x_{2i} 의 세부 층 개수를 L_2^P 라 할 때 L_1^P, L_2^P 는 $n/L_1^P \geq 10, n/L_2^P \geq 10$ 을 만족하도록 결정한다. 여기서 n 은 총 표본 수이다. 보조변수 x_{1i}, x_{2i} 가 범주형 변수인 경우에도 범주별로 층을 구성할 수 있다. 다만 범주형 보조변수의 세부 층 개수는 범주의 개수에 영향을 받게 되므로 충분한 개수의 세부 층 구성이 어려울 수 있다. 이제 $(h_1, h_2), h_1 = 1, \dots, L_1^P, h_2 = 1, \dots, L_2^P$ 층에

포함된 i 자료의 가중치 합을 $\hat{N}_{h_1, h_2}^{(0)}$ 라 하자. 이때 갈퀴법에 사용한 초기치 $\hat{N}_{h_1, h_2}^{(0)}$ 는 (h_1, h_2) 층에 포함된 i 자료의 가중치 합을 사용하기 때문에 정보적 표본설계, 추가 표본, 세부 층화 등으로 인해 (h_1, h_2) 층 내의 i 자료 가중치가 동일하지 않은 경우에도 사용이 가능하다. 그리고 h_1 층 자료 가중치의 총합을 $\hat{N}_{h_1+}^{(0)}$, h_2 층 자료 가중치의 총합을 $\hat{N}_{+h_2}^{(0)}$ 이라 하고 해당하는 모집단 수를 각각 N_{h_1+} , N_{+h_2} 라 하자. 갈퀴법을 이용하면 최종적으로 동일한 합계 모집단 수가 얻어지며 이때 (h_1, h_2) , $h_1 = 1, \dots, L_1^P$, $h_2 = 1, \dots, L_2^P$ 세부 층의 보정된 가중치 총합의 최종값 $\hat{N}_{h_1, h_2}^{(F)}$ 이 얻어진다. 이제 $\hat{N}_{h_1, h_2}^{(0)} > 0$ 인 세부 층에서 $f_{h_1, h_2} = \hat{N}_{h_1, h_2}^{(F)} / \hat{N}_{h_1, h_2}^{(0)}$ 라 하면 최종 보정 설계 가중치는 다음과 같이 구해진다.

$$w_{i \in (h_1, h_2)}^{AF} = w_{i \in (h_1, h_2)} \times f_{h_1, h_2} \quad (4)$$

당연히 갈퀴법을 이용한 최종 보정 가중치 합은 모집단 수와 같게 된다.

2) 응답확률의 사후보정을 위한 다변량 세부 층 구성 방법

모집단에 포함된 n 개의 표본자료를 이용하여 2개의 보조변수 x_{1i} , x_{2i} 의 분위수를 구한다. 이때 x_{1i} 의 세부 층의 수를 L_1^S , x_{2i} 의 세부 층의 수를 L_2^S 라 하자. 이때 $r/L_1^S \geq 5$, $r/L_2^S \geq 5$ 이 되도록 결정한다. 여기서 r 은 최종 응답자료 수이다. 이제 (h_1, h_2) , $h_1 = 1, \dots, L_1^S$, $h_2 = 1, \dots, L_2^S$ 층에서 얻어진 i 자료 성향점수 $\hat{p}_i^{(0)}$ 의 역수의 합, $\sum_{i \in (h_1, h_2)} 1/\hat{p}_i^{(0)} = \hat{n}_{h_1, h_2}^{(0)}$ 이라 하고, h_1 층에서 얻어진 성향점수 역수의 합을 $\hat{n}_{h_1+}^{(0)} = \sum_{h_2} \hat{n}_{h_1, h_2}^{(0)}$, h_2 층에서 얻어진 성향점수 역수의 합을 $\hat{n}_{+h_2}^{(0)} = \sum_{h_1} \hat{n}_{h_1, h_2}^{(0)}$ 이라 하며 해당하는 표본 수를 각각 n_{h_1+} , n_{+h_2} 라 하자. 2절 1항과 같이 갈퀴법을 이용하면 최종적으로 동일한 합계 표본 수가 얻어지며 이때 (h_1, h_2) , $h_1 = 1, \dots, L_1^S$, $h_2 = 1, \dots, L_2^S$, 세부 층의 최종 값 $\hat{n}_{h_1, h_2}^{(F)}$ 이 얻어진다.

이제 $f_{h_1, h_2}^S = \hat{n}_{h_1, h_2}^{(F)} / \hat{n}_{h_1, h_2}^{(0)}$ 라 하면 최종 보정 응답확률 $\hat{p}_{i \in (h_1, h_2)}^{AF}$ 은 다음의 식에

의해 구해진다.

$$\frac{1}{\hat{p}_{i \in (h_1, h_2)}^{AF}} = \frac{1}{\hat{p}_{i \in (h_1, h_2)}^{(0)}} \times f_{h_1, h_2}^S \quad (5)$$

당연히 $\sum_{i=1}^{r_{h_1, h_2}} 1/\hat{p}_{i \in (h_1, h_2)}^{AF} = \hat{n}_{h_1, h_2}^{(F)}$ 이 성립하며 여기서 r_{h_1, h_2} 는 (h_1, h_2) 세부 층의 최종 응답자료 개수이다.

3) 추정을 위한 최종 가중치와 레이킹 나이브 PSA 추정량

이제 $\hat{w}_i^{AF}$, $\hat{p}_i^{AF}$ 를 각각 식(4)와 식(5)에서 얻어진 (h_1, h_2) 세부 층의 보정 가중치와 보정 응답확률이라 하고 $w_i^{(*)} = \hat{w}_i^{AF} / \hat{p}_i^{AF}$ 라 하자. 얻어진 $w_i^{(*)}$ 는 2절 1항에서 설명한 방법으로, 즉 갈퀴법을 이용하여 최종적으로 행 합과 열 합이 일치하도록 보정한다. 즉 추정에 사용할 i 자료의 최종 가중치는

$$w_{i \in (h_1, h_2)}^{F(Raking)} = w_{i \in (h_1, h_2)}^{(*)} \times f_{h_1, h_2}^* \quad (6)$$

이 된다. 여기서 $f_{h_1, h_2}^* = \hat{N}_{h_1, h_2}^{(F(*))} / \hat{N}_{h_1, h_2}^{(0(*))}$ 이고 $\hat{N}_{h_1, h_2}^{(F(*))}$ 는 갈퀴법에서 얻어진 $w_i^{(*)}$ 의 최종 가중치의 합, $\hat{N}_{h_1, h_2}^{(0(*))}$ 는 초기 가중치의 합이다. 따라서 본 연구에서 제안한 레이킹 나이브 PSA 추정량은 다음과 같다.

$$\hat{Y}_{Raking} = \frac{1}{N} \sum_i^r w_i^{F(Raking)} y_i \quad (7)$$

3. 연구에 사용된 PSA 추정량

본 연구에서는 다음의 성향점수보정 추정량이 사용된다.

[1] $\hat{Y}_{COM}$: 무응답자료를 제거한 후 얻어진 최종 응답자료를 이용하여 평균을 추

- 정한다. 즉 r 개 완전자료의 단순 평균인 $\hat{Y}_{COM} = (1/r) \sum_{i=1}^r y_i$ 을 사용한다.
- [2] $\hat{Y}_{PSA}$: 식(3)의 로지스틱 회귀모형에서 얻어진 응답확률 $\hat{p}_i$ 을 사용한 성향점수보정 추정량을 사용한다. 정보적 표본설계에서 설계 가중치 w_i 가 알려진 경우는 알려진 w_i 와 추정된 응답확률 $\hat{p}_i$ 를 이용하여 무응답 보정 가중치 $w_i^{PSA} = w_i/\hat{p}_i$ 를 구한 후, 이를 보정한 최종 보정 가중치 $w_i^{F(PSA)} = w_i^{PSA} \times N / \sum_{i=1}^r w_i^{PSA}$ 를 사용한다. 따라서 $\sum_{i=1}^r w_i^{F(PSA)} = N$ 이 성립하며 최종적으로 $\hat{Y}_{PSA} = (1/N) \sum_{i=1}^r w_i^{F(PSA)} y_i$ 을 사용한다. 만약 단순임의추출법이 사용되었다면 $w_i = w = N/n$ 을 사용한다.
- [3] $\hat{Y}_{C-S}$: Chung & Shin(2022)에서 사용한 추정량으로 데카르트 곱을 이용하여 세부 층을 구성한 후 얻어진 최종 가중치를 이용하여 추정량을 계산한다. 즉, (h_1, h_2) 세부 층의 모집단 수를 N_{h_1, h_2} , 최종 응답 수를 r_{h_1, h_2} 라 하면 (h_1, h_2) 세부 층에 포함된 i 자료의 가중치는 $w_i^{C-S} = N_{h_1, h_2} / r_{h_1, h_2}$ 을 사용한다. 다만 최종 응답자료가 없는 세부 층이 있을 수 있으므로 최종 추정량으로 $\sum_i w_i^{C-S} = \sum_i (N_{h_1, h_2} / r_{h_1, h_2}) = N$ 이 되도록 보정한다. 최종적으로 얻어진 최종 보정 가중치 $w_i^{F(C-S)}$ 를 사용한 $\hat{Y}_{C-S} = (1/N) \sum_i w_i^{F(C-S)} y_i$ 를 사용한다.
- [4] $\hat{Y}_{Rakng}$: 본 연구에서 제안한 식(7)의 레이킹 PSA 추정량으로 $\hat{Y}_{Raking} = (1/N) \sum_i w_i^{F(Raking)} y_i$ 을 사용한다.

4. 비교통계량(comparison statistics)

앞에서 얻어진 4개의 평균 추정량은 다음의 비교통계량: 편향(bias), 절대상대편향(Absolute Relative Bias: ARB) 그리고 제곱근 MSE(Root Mean Squared Error: RMSE)을 이용하여 결과의 성능을 비교한다. 각 통계량의 정의는 다음과 같다.

$$Bias = \frac{1}{R} \sum_{r=1}^R (\widehat{Y}_r - \overline{Y}_r)$$

$$Abias = \frac{1}{R} \sum_{r=1}^R |\widehat{Y}_r - \overline{Y}_r| / \overline{Y}_r \times 100$$

$$RMSE = \sqrt{\frac{1}{R} \sum_{r=1}^R (\widehat{Y}_r - \overline{Y}_r)^2}$$

여기서 $R = 1,000$ 을 사용하였으며 반복마다 새로운 모집단을 생성하여 통계량을 계산한다. 이는 생성된 특정 모집단의 영향을 줄이기 위함이다. 이에 r 번째 반복 모집단의 참값을 $\overline{Y}_r$ 로 표시한다.

IV. 모의실험

1. 모의실험 설계

1) 보조변수 생성

보조변수가 각각 감마분포와 균일분포에서 생성되었다.

(1) x_{1i} 생성: $x_{1i}^* \stackrel{iid}{\sim} Gamma(1, 50)$ 에서 추출한 후, 최종적인 보조자료는 $x_{1i} = 100 + x_{1i}^*$ 로 하고 $x_i \geq 300$ 인 자료는 삭제하였다. 따라서 보조변수 x_{1i} 는 100에서 300 사이의 값을 갖는다. 이때 모집단 자료 개수는 $N = 10,000$ 이다.

(2) x_{2i} 생성: $x_{2i} = 100 + \alpha \times x_{1i} + \epsilon_i$, $\epsilon_i \stackrel{iid}{\sim} Unif(100, 200)$, 여기서 $\alpha = 0$, $\alpha = 0.5$ 을 사용하였다. $\alpha = 0$ 인 경우 두 보조변수는 독립이고, $\alpha = 0.5$ 인 경우 두 자료는 상관관계가 존재하며 상관계수는 약 0.6이다.

2) 초모집단모형(Super-Population Model: SPM)

다음 네 개의 초모집단모형이 사용되었다.

(1) 선형 초모집단모형(SPM1): $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$, $\epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$,

여기서 $\beta_0 = 2$, $\beta_1 = 0.5$, $\beta_2 = 0.3$ 이고 $\sigma^2 = 1,600$ 이다.

(2) 2차 초모집단모형(SMP2): $y_i = \beta_0 + \beta_1 x_{1i} + \beta_{11} (x_{1i} - 105)^2 + \beta_2 x_{2i} + \epsilon_i$,

$\epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$, 여기서 $\beta_0 = 2$, $\beta_1 = 0.5$, $\beta_{11} = 0.03$, $\beta_2 = 0.3$ 이고 $\sigma^2 = 900$ 이다.

(3) 로그-정규분포(SPM3): $\ln(y_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$, $\epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$,

여기서 $\beta_0 = 0.01$, $\beta_1 = 0.03$, $\beta_2 = 0.01$, $\sigma^2 = 0.1$ 이다.

(4) 감마분포(SPM4): $y_i \stackrel{iid}{\sim} \text{Gamma}(\alpha_1, \mu/\alpha_1)$, 여기서 $\mu = \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 (x_{2i} - 200))$, $\alpha_1 = 10$, $\beta_0 = 0.5$, $\beta_1 = 0.02$, $\beta_2 = 0.01$ 이다.

3) 표본추출 과정

생성된 $N = 10,000$ 개의 모집단 자료에서 $n = 500$ 개의 표본을 추출하였다. 이 때 정보적 표본설계가 사용될 수 있으나 실무에서 사용하는 표본설계에서는 주로 무정보적 표본설계가 사용되므로 본 모의실험에서는 무정보적 표본설계인 단순임의추출법을 사용하였다.

4) 응답자료 생성 과정

본 모의실험에서는 응답확률 모형(Response Probability Model: RPM)으로 선형 응답확률 모형과 로지스틱 응답확률 모형을 고려하였으며, 사용된 모형은 다음과 같다.

(1) 선형 응답확률 모형: $p_i = a_0 + a_1 y_i$

(2) 로지스틱 응답확률 모형: $\ln\left(\frac{p_i}{1-p_i}\right) = c_0 + c_1 y_i$

추출된 $n = 500$ 개의 표본에서 선형 응답확률 모형인 $p_i = a_0 + a_1 y_i$, $p_i \in [0, 1]$ 을 이용하여 무응답을 생성하였다. 즉 y_i 의 최솟값에서의 응답확률을 $p_y^{\min}$, y_i 의 최댓값에서의 응답확률을 $p_y^{\max}$ 라 할 때, $(p_y^{\min}, p_y^{\max}) = (0.9, 0.3)$, $(0.3, 0.9)$ 을 사용하여 a_0 , a_1 을 계산하고, 얻어진 a_0 , a_1 값과 y_i 에 따라 응답확률을 계산하였다. 이는 선형 응답 모형에서 응답확률이 ‘0’과 ‘1’ 사이의 값을 갖도록 하기 위함이다. 계산된 응답확률에 따라 응답 및 무응답자료를 생성하였다. 같은 방법을 로지스틱 응답확률 모형인 $\ln(p_i/(1-p_i)) = c_0 + c_1 y_i$ 에도 적용하여 c_0 , c_1 을 구한 후, 응답 및 무응답자료를 생성하였다. 물론 로지스틱 응답확률은 ‘0’과 ‘1’ 사이이므로 최댓값, 최솟값을 정할 필요가 없지만, 사용된 최솟값과 최댓값인 0.3과 0.9는 현실적이라 판단하였다.

응답한 최종 응답자료는 r 개이다. 여기서 $(\pi_y^{\min}, \pi_y^{\max}) = (0.9, 0.3)$ 또는 $(0.3, 0.9)$ 인 경우, 전체 자료의 약 60%가 되어 주어진 자료 수 n 에 비해 약 40%가 감소하였다.

5) 층화

N 개의 모집단 자료에서 배정된 n 개의 자료를 $L_1^P \times L_2^P$ 개의 세부 층으로 나누었다. 이때 x_1 을 이용하여 L_1^P 개 세부 층을 나누고, x_2 를 이용하여 L_2^P 개의 세부 층을 나누었다. 본 연구에서는 모의실험을 간단히 하기 위해 $L_1^P = L_2^P$, $n/L_1^P \geq 10$ 를 사용하였다. 층 경계는 분위수를 사용하였고, 각 보조변수에 따라 분위수가 달라지므로 층 경계는 보조변수에 따라 달라진다. 본 모의실험에서는 $n = 500$ 을 사용하였고, $L_1^P = L_2^P$ 는 5와 8을 사용하였다. 이렇게 구성된 세부 층을 이용하여 보정가중치를 구하였다.

얻어진 최종 응답자료는 (x_{1i}, x_{2i}, y_i) , $i = 1, \dots, r$ 이므로 최종적으로 얻어진 r 개를 이용하여 세부 층 개수를 결정하였다. 모집단 가중치 보정을 위한 세부 층 구성 방법과 같은 방법으로 표본으로 배정된 자료 n 개의 보조변수 x_{1i} , x_{2i} 의 분위수를 이용하여 세부 층을 나누었다. 이때 2차 세부 층의 수 L_1^S , L_2^S 는 $r/L_1^S \geq 5$, $r/L_2^S \geq 5$ 이 되도록 결정하였다. 다양한 L_1^S , L_2^S 개수를 이용하여 갈퀴법의 영향을 살펴보았으며 $L_1^S = L_2^S$ 를 사용하였다.

6) 비교통계량

3장 3절에서 설명한 $\hat{Y}_{COM}$, $\hat{Y}_{PSA}$, $\hat{Y}_{C-S}$ 그리고 $\hat{Y}_{Raking}$ 이 계산되었으며 3장 4절의 비교통계량이 계산되었다.

2. 모의실험 결과

모의실험은 무정보적 표본설계, 즉 단순임의추출법을 사용하였으며 응답확률 모형으로 로지스틱 응답확률 모형과 선형 응답확률 모형이 사용되었다. 표에서 응답확률(Response Probability: RP)의 최소 응답확률과 최대 응답확률 ($p_y^{\min}, p_y^{\max}$) = (0.9, 0.3)인 경우를 RP I, (0.3, 0.9)인 경우를 RP II라 표시하였다.

1) 로지스틱 응답확률 모형

<표 1> ~ <표 4>는 로지스틱 응답확률 모형을 사용하였을 때의 결과이다. 3장 4절에서 설명한 세 가지 비교통계량을 살펴보면, 완전자료를 사용한 $\hat{Y}_{COM}$ 은 모든 비교통계량에서 매우 나쁜 결과를 보이므로 무응답 처리가 필요한 것을 확인할 수 있다. 또한 ($p_y^{\min}, p_y^{\max}$) = (0.9, 0.3)의 경우인 RP I의 결과를 보면 음의 편향이 발생하여 과소 추정이 되는 것을 알 수 있으며 ($p_y^{\min}, p_y^{\max}$) = (0.3, 0.9)의 경우인 RP II에서는 과대 추정되는 것을 확인할 수 있다.

<표 1>과 <표 2>는 각 초모집단모형에서 보조변수 간에 상관관계가 없는 경우, $\rho = 0$ 인 결과로 다양한 $L_1^S = L_2^S$ 를 사용하여 결과를 얻었으나 이 중에서 일부만을 표로 작성하였다. <표 1>의 RP I, SPM1의 결과를 살펴보면, 즉 관심변수 y_i 가 커질 때 응답확률이 작아지고, 초모집단모형이 선형인 경우의 결과를 살펴보면, $\hat{Y}_{PSA}$ 와 $\hat{Y}_{C-S}$ 의 최적값이 유사한 것을 확인할 수 있다. 따라서 이 경우 $\hat{Y}_{C-S}$ 은 최적의 세부 층 개수를 사용해야 하므로 현실적으로 최적의 세부 층 개수가 알려지지 않은 경우는 $\hat{Y}_{PSA}$ 가 우수한 결과를 주는 것으로 판단할 수 있다. 또한, 본 모의실험 결과에 수록하지는 않았지만, 사용된 분산 $\sigma^2 = 1,600$ 보다 작은 값의 분산을 사용한

경우에는 즉, $\sigma^2 = 900$ 에서는 $\hat{Y}_{PSA}$ 가 $\hat{Y}_{C-S}$ 의 최적값보다 더욱 우수하였다. 이는 초모집단모형이 선형이고 분산이 작은 경우에는 MAR 가정을 사용하여 얻은 응답확률 추정값 $\hat{p}_i$ 이 일정 수준 이상의 정확성을 보이기 때문이다. 극단적으로 초모집단모형이 선형이고, 오차의 분산이 매우 작다면 MNAR 무응답은 MAR 무응답 가정을 사용하여 $\hat{p}_i$ 을 추정하여도 정확한 응답확률 추정이 가능하다. 물론 분산이 커지면 $\hat{Y}_{C-S}$ 가 $\hat{Y}_{PSA}$ 에 비해 우수한 결과를 주게 된다. 본 논문에서는 제안한 $\hat{Y}_{Raking}$ 과 $\hat{Y}_{C-S}$ 의 결과를 비교하는 것이 주된 목적이다. 먼저 $\hat{Y}_{Raking}$ 은 세부 층 개수와 무관하게 최적의 결과를 준다. 즉 $\hat{Y}_{Raking}$ 은 모집단 세부 층 개수 (L_1^P, L_2^P)와 표본 세부 층 개수 (L_1^S, L_2^S)에 무관하게 편향, ARB, RMSE의 결과가 비슷하게 얻어진다. 반면, $\hat{Y}_{C-S}$ 결과는 최적의 층 개수보다 많아지면 세 비교통계량 값이 증가하는 것을 확인할 수 있다. 따라서 $\hat{Y}_{C-S}$ 의 최적 결과가 세부 층 개수에 무관한 $\hat{Y}_{Raking}$ 의 결과와 비슷하므로 $\hat{Y}_{Raking}$ 추정량을 사용하는 것이 더 효과적이라고 할 수 있다.

다음으로 <표 1>의 RP I, SPM2, 즉 초모집단모형이 2차 식인 경우의 결과를 살펴보면, $\hat{Y}_{PSA}$ 편향이 $\hat{Y}_{C-S}$ 편향보다 작지만, ARB와 RMSE를 기준으로 판단할 때 $\hat{Y}_{C-S}$ 가 우수한 결과를 준다. 특히 최적의 세부 층 개수를 사용한다면, 편향, ARB, RMSE 등 모든 비교통계량에서 $\hat{Y}_{C-S}$ 가 우수한 결과를 준다. 또한 $\hat{Y}_{C-S}$ 와 $\hat{Y}_{Raking}$ 의 결과를 비교하면 모든 세부 층 개수에서 $\hat{Y}_{Raking}$ 가 우수하며, 또한 충분한 세부 층 개수를 사용하면, 즉 (L_1^P, L_2^P)와 (L_1^S, L_2^S)의 수를 크게 하면, 매우 우수한 결과를 얻을 수 있다. 따라서 SPM1, SPM2 초모집단모형에서 모두 $\hat{Y}_{Raking}$ 가 $\hat{Y}_{C-S}$ 에 비해 우수한 결과를 준다.

또한 <표 1>의 RP II 결과를 살펴보면, <표 1>의 RP I 결과와 매우 유사한 패턴의 결과를 보여준다. 즉 응답확률이 관심변수 자료 값에 따라 증가하는 경우에서도 $\hat{Y}_{Raking}$ 가 우수한 결과를 준다. 결국, 관심변수 자료 값의 증가 또는 감소와 무관하게 $\hat{Y}_{Raking}$ 가 우수한 결과를 준다.

<표 2>는 로지스틱 응답확률 모형에서 로그-선형 초모집단모형인 SPM3와 SPM4의 결과로, <표 1>의 SPM2 결과와 유사한 패턴의 결과를 보인다. RP I의 결

과에서는 일부 너무 적거나 너무 많은 세부 층 개수를 제외하고, $\hat{Y}_{C-S}$ 가 $\hat{Y}_{PSA}$ 에 비해 우수한 결과를 준다. 물론 최적의 세부 층 개수를 사용하면 $\hat{Y}_{C-S}$ 가 $\hat{Y}_{PSA}$ 에 비해 매우 우수한 결과를 준다. 그러나 RP II, SPM3 결과에서는 최적의 $\hat{Y}_{C-S}$ 보다 $\hat{Y}_{PSA}$ 가 우수한 결과를 주는 반면 RP II, SPM4 결과에서는 최적의 $\hat{Y}_{C-S}$ 가 $\hat{Y}_{PSA}$ 보다 더욱 우수하다. 따라서 본 모의실험에서 생성한 자료에서는 어떤 추정량이 우수하다고 판단하기에는 어려움이 있다.

반면 $\hat{Y}_{Raking}$ 는 세부 층 개수에 민감하지 않고 안정적인 결과를 주는 것으로 확인되며, 특히 충분한 수의 세부 층 개수를 사용한다면 매우 우수한 결과를 준다. RP II의 $L_1^S = L_2^S = 20$ 인 경우에는 갈퀴법이 수렴하지 않아 $\hat{Y}_{Raking}$ 가 구해지지 않았고, 이에 따라 비교통계량 결과도 얻어지지 않았다. 결론적으로 너무 많은 수의 세부 층 개수를 사용하는 것은 타당하지 않다.

<표 3>과 <표 4>는 각 초모집단모형에 대해 보조변수 간에 상관관계가 약 $\rho = 0.6$ 인 경우의 결과이며 전체적으로 $\rho = 0$ 인 결과와 유사한 패턴의 결과가 얻어졌기 때문에 $L_1^P = L_2^P$ 에서 최소의 RMSE를 만족하는 결과만을 수록하였다. <표 3>의 SPM1 결과를 살펴보면 $\hat{Y}_{PSA}$, $\hat{Y}_{C-S}$ 그리고 $\hat{Y}_{Raking}$ 의 결과가 유사하며, <표 1> 결과처럼 $\hat{Y}_{Raking}$ 은 세부 층 개수와 무관하게 일정한 결과를 유지한다. 다른 초모집단, SPM2, SPM3, SPM4 결과를 살펴보면 본 연구에서 제안한 $\hat{Y}_{Raking}$ 의 결과가 모든 비교통계량을 기준으로 매우 우수한 것을 확인할 수 있다.

특히 눈에 띄는 결과로 초모집단모형이 로그-선형모형인 <표 4>의 RP I 결과에서 $\hat{Y}_{Raking}$ 의 최적 결과는 $L_1^S = L_2^S = 50$ 에서 결정되었으며 이 값은 <표 1>부터 <표 3>에 얻어진 최적 세부 층 개수보다 매우 큰 개수에서 결정되었는데 이는 Lee & Shin(2004, b)의 최적 세부 층 개수 결과와도 일치한다.

최적 세부 층 개수는 초모집단모형의 형태 및 분포에 따라 달라지고, 보조변수의 분포에 따라서도 달라지며 최종 응답자료 수에 따라서도 달라지므로 이론적 최적 세부 층 개수 결정은 쉽지 않다. 다만 모의실험 결과에 의하면 $\hat{Y}_{Raking}$ 의 경우 보조변수별로 정해진 세부 층의 안에 포함되는 표본 수가 5개 이상이고, 갈퀴법이 수렴한다면 우수한 결과를 얻을 수 있을 것으로 판단한다.

<표 1> 로지스틱 응답확률이고 선형과 2차 초모집단인 경우($\alpha = 0$)

		초모집단모형											
RP	Est	L_1^P	L_2^P	SPM1					SPM2				
				L_1^S	L_2^S	BIAS	ARB(%)	RMSE	L_1^S	L_2^S	BIAS	ARB(%)	RMSE
I	$\hat{\bar{Y}}_{COM}$	-	-	-5.75	3.83	6.25	-	-19.68	7.78	21.38			
	$\hat{\bar{Y}}_{PSA}$	-	-	-4.36	2.93	4.96	-	-3.82	3.15	9.86			
	$\hat{\bar{Y}}_{C-s}$	5	5	4	4	-4.39	2.95	4.93	4	4	-4.70	2.44	7.58
				8	8	-4.41	2.97	4.97	8	8	-3.77	2.21	6.98
				10	10	-4.50	3.02	5.05	10	10	-3.96	2.30	7.15
				15	15	-4.81	3.23	5.37	15	15	-7.29	3.12	9.37
		8	8	6	6	-4.36	2.94	4.98	6	6	-4.19	2.25	7.00
				8	8	-4.38	2.96	5.02	8	8	-3.86	2.16	6.67
				12	12	-4.63	3.12	5.27	10	10	-4.08	2.25	6.99
				20	20	-5.10	3.42	5.69	20	20	-11.37	4.53	12.78
	$\hat{\bar{Y}}_{Raking}$	5	5	4	4	-4.36	2.92	4.87	4	4	-4.15	2.28	7.06
				8	8	-4.36	2.92	4.86	8	8	-3.26	2.10	6.60
				10	10	-4.36	2.92	4.87	10	10	-2.73	2.00	6.31
				15	15	-4.38	2.94	4.89	15	15	-2.01	1.90	6.02
		8	8	6	6	-4.35	2.92	4.88	6	6	-3.00	1.81	5.64
				10	10	-4.35	2.92	4.87	10	10	-2.59	1.72	5.44
				16	16	-4.37	2.92	4.88	16	16	-1.78	1.56	4.98
				20	20	-4.37	2.93	4.89	20	20	-1.40	1.50	4.78
	$\hat{\bar{Y}}_{COM}$	-	-	6.21	4.15	6.68	-	49.92	19.69	52.68			
	$\hat{\bar{Y}}_{PSA}$	-	-	4.77	3.21	5.32	-	2.50	3.00	9.59			
II	$\hat{\bar{Y}}_{C-s}$	5	5	4	4	4.75	3.20	5.26	4	4	5.93	3.00	9.51
				6	6	4.73	3.18	5.25	6	6	4.65	2.58	8.18
				10	10	4.80	3.23	5.35	8	8	6.19	3.05	9.39
				15	15	5.20	3.49	5.72	15	15	21.16	8.35	23.05
		8	8	6	6	4.77	3.20	5.35	6	6	5.06	2.74	8.75
				10	10	4.84	3.25	5.44	10	10	10.76	4.45	13.22
				16	16	5.29	3.55	5.84	12	12	15.34	6.08	17.33
				20	20	5.54	3.71	6.09	20	20	29.87	11.78	31.68
	$\hat{\bar{Y}}_{Raking}$	5	5	4	4	4.76	3.19	5.23	4	4	4.56	2.53	8.08
				8	8	4.76	3.19	5.24	8	8	2.99	2.23	7.14
				12	12	4.75	3.18	5.23	12	12	2.18	2.06	6.50
				15	15	4.75	3.18	5.24	15	15	1.80	1.92	6.15
		8	8	6	6	4.76	3.19	5.25	6	6	2.97	1.85	5.96
				10	10	4.77	3.20	5.25	10	10	2.51	1.79	5.77
				16	16	4.75	3.19	5.24	16	16	1.84	1.63	5.21
				20	20	4.77	3.20	5.27	20	20	1.63	1.61	5.16

<표 2> 로지스틱 응답확률이고 로그-선형 초모집단인 경우($\alpha = 0$)

RP Est		로그-선형 초모집단모형											
		L_1^P	L_2^P	SPM3					SPM4				
				L_1^S	L_2^S	BIAS	ARB(%)	RMSE	L_1^S	L_2^S	BIAS	ARB(%)	RMSE
I	$\hat{Y}_{COM}$	-	-	-387.28	14.93	453.66	-	-	-66.72	17.02	71.15		
	$\hat{Y}_{PSA}$	-	-	-203.63	10.03	324.23	-	-	-29.35	8.47	39.55		
	$\hat{Y}_{C-S}$	5	5	6	6	-210.42	9.30	297.57	6	6	-22.86	6.37	29.42
				10	10	-184.76	8.85	284.88	8	8	-21.50	6.16	28.78
				12	12	-188.03	8.94	288.41	10	10	-23.72	6.60	30.43
				15	15	-205.55	9.29	299.01	15	15	-32.26	8.45	37.62
		8	8	8	8	-193.13	8.97	283.58	6	6	-23.56	6.52	30.00
				10	10	-184.25	8.77	280.50	8	8	-22.46	6.35	29.53
				12	12	-188.13	8.94	284.07	10	10	-24.88	6.79	31.16
				16	16	-215.42	9.58	303.87	12	12	-28.02	7.54	34.17
	$\hat{Y}_{Raking}$	5	5	4	4	-217.84	9.43	300.24	4	4	-26.89	7.21	32.68
				8	8	-195.60	8.90	286.97	8	8	-23.60	6.51	30.01
				12	12	-165.29	8.26	270.07	12	12	-22.28	6.25	28.98
				15	15	-149.51	7.93	260.70	15	15	-21.75	6.12	28.46
		8	8	6	6	-191.56	8.51	270.19	6	6	-23.28	6.33	28.83
				10	10	-177.78	8.22	262.37	10	10	-22.37	6.14	28.12
				16	16	-144.19	7.53	243.06	16	16	-21.48	5.94	27.39
				20	20	-127.45	7.22	233.47	20	20	-21.03	5.84	26.91
$\hat{Y}_{COM}$	-	-	1,401.47	52.42	1,559.76	-	-	195.25	49.75	204.12			
$\hat{Y}_{PSA}$	-	-	355.83	15.19	507.15	-	-	32.56	9.56	45.52			
II	$\hat{Y}_{C-S}$	5	5	4	4	477.64	18.98	621.45	4	4	38.69	10.29	47.63
				6	6	428.90	17.58	581.46	6	6	31.50	8.60	40.28
				8	8	456.33	18.28	599.10	8	8	37.33	9.87	45.53
				12	12	626.72	24.01	744.87	12	12	69.11	17.70	75.43
		8	8	6	6	439.38	18.05	608.68	6	6	32.45	8.78	41.74
				8	8	476.86	19.01	629.81	8	8	39.53	10.40	47.59
				12	12	640.22	24.52	762.03	12	12	71.36	18.25	77.57
				16	16	817.63	30.82	937.12	16	16	104.79	26.71	110.54
	$\hat{Y}_{Raking}$	5	5	4	4	441.63	17.79	580.07	4	4	39.00	10.28	46.84
				8	8	350.85	14.98	501.68	8	8	31.73	8.66	40.37
				12	12	252.31	11.91	395.56	12	12	28.85	8.05	37.96
				15	15	214.06	10.64	355.80	15	15	28.24	7.87	37.24
		8	8	6	6	340.43	14.19	473.67	6	6	30.72	8.24	37.70
				10	10	300.74	13.09	447.84	10	10	29.30	7.96	36.76
				12	12	259.95	11.85	407.52	12	12	27.88	7.67	35.61
				16	16	214.12	10.32	352.68	16	16	27.32	7.59	35.32
				20	20	-	-	-	20	20	-	-	-

<표 3> 로지스틱 응답확률이고 선형과 2차 초모집단인 경우($\alpha = 0.5$)

		초모집단모형											
RP	Est	L_1^P	L_2^P	SPM1				SPM2					
				L_1^S	L_2^S	BIAS	ARB(%)	RMSE	L_1^S	L_2^S	BIAS	ARB(%)	RMSE
I	$\hat{Y}_{COM}$	-	-			-6.01	3.50	6.55	-	-	-20.40	7.42	22.15
	$\hat{Y}_{PSA}$	-	-			-3.98	2.36	4.69	-	-	-3.89	2.97	10.06
	$\hat{Y}_{C-S}$	5	5	4	4	-3.99	2.35	4.57	10	10	-2.85	1.81	6.15
		8	8	6	6	-4.01	2.39	4.67	10	10	-2.80	1.66	5.78
	$\hat{Y}_{Raking}$	5	5	15	15	-3.98	2.34	4.54	15	15	-1.94	1.71	5.85
		8	8	20	20	-3.97	2.34	4.55	20	20	-1.32	1.36	4.75
	$\hat{Y}_{COM}$	-	-			6.84	3.98	7.34	-	-	51.67	18.75	54.46
	$\hat{Y}_{PSA}$	-	-			4.64	2.74	5.28	-	-	2.29	2.78	9.69
II	$\hat{Y}_{C-S}$	5	5	4	4	4.65	2.73	5.19	6	6	4.12	2.17	7.55
		8	8	6	6	4.66	2.74	5.26	6	6	4.40	2.21	7.71
	$\hat{Y}_{Raking}$	5	5	15	15	4.66	2.73	5.19	15	15	1.89	1.75	6.06
		8	8	20	20	4.65	2.73	5.19	20	20	1.71	1.46	5.05

<표 4> 로지스틱 응답확률이고 로그-선형 초모집단인 경우($\alpha = 0.5$)

		로그-선형 초모집단모형											
RP	Est	L_1^P	L_2^P	SPM3					SPM4				
				L_1^S	L_2^S	BIAS	ARB(%)	RMSE	L_1^S	L_2^S	BIAS	ARB(%)	RMSE
I	$\hat{Y}_{COM}$	-	-	-	-	-1,309.94	17.03	1,563.03	-	-	-443.63	19.21	478.27
	$\hat{Y}_{PSA}$	-	-	-	-	-763.84	12.34	1,190.37	-	-	-225.27	10.83	296.33
	$\hat{Y}_{C-S}$	5	5	20	20	-412.16	9.16	900.56	12	12	-113.54	6.25	177.98
		8	8	20	20	-403.69	8.37	819.29	12	12	-119.02	6.07	170.52
	$\hat{Y}_{Raking}$	5	5	40	40	-328.90	8.86	874.97	40	40	-103.50	6.06	173.07
		8	8	50	50	-289.30	7.85	774.74	50	50	-103.02	5.67	158.71
	$\hat{Y}_{COM}$	-	-	-	-	4,836.31	60.63	5,507.73	-	-	1,404.11	60.66	1,481.62
	$\hat{Y}_{PSA}$	-	-	-	-	1,522.11	21.23	2,098.98	-	-	310.06	14.43	398.33
II	$\hat{Y}_{C-S}$	5	5	8	8	1,525.40	20.67	2,035.00	4	4	233.01	10.78	299.46
		8	8	6	6	1,512.59	20.60	2,033.87	6	6	234.44	10.70	301.18
	$\hat{Y}_{Raking}$	5	5	15	15	795.39	13.87	1,401.90	15	15	153.72	8.19	236.22
		8	8	16	16	763.20	12.88	1,310.33	16	16	142.77	7.34	212.02

2) 선형 응답확률 모형

<표 5>~<표 8>은 선형 응답확률 모형을 사용하였을 때의 결과이며 RMSE가 최소인 세부 층의 모의실험 결과를 수록하였다. 로지스틱 응답확률 모형의 결과와 동일하게 완전자료를 이용하여 구한 $\hat{Y}_{COM}$ 은 모든 비교통계량에서 매우 나쁜 결과를 보이며, $(p_y^{\min}, p_y^{\max}) = (0.9, 0.3)$ 인 RP I의 결과를 보면 과소 추정이 되고, $(p_y^{\min}, p_y^{\max}) = (0.3, 0.9)$ 인 RP II 경우에는 과대 추정되는 것을 알 수 있다.

로지스틱 응답확률 모형의 결과와 동일하게 초모집단모형이 SPM1인 경우, $\hat{Y}_{PSA}$ 과 최적의 $\hat{Y}_{C-S}$ 그리고 $\hat{Y}_{Raking}$ 이 유사한 결과를 주며 선형 응답확률 모형에서도 $\hat{Y}_{Raking}$ 은 세부 층 개수와 무관하게 안정적인 결과를 보여준다. 초모집단모형이 SPM2, SPM3 그리고 SPM4인 경우는 $\hat{Y}_{Raking}$ 이 매우 우수한 결과를 준다. 결론적으로 $\hat{Y}_{Raking}$ 은 충분한 세부 층 개수를 사용한다면 $\hat{Y}_{C-S}$ 에 비해 우수한 결과를 얻을 수 있다.

<표 5> 선형 응답확률이고 선형과 2차 초모집단인 경우($\alpha = 0$)

RP Est		L_1^P L_2^P		초모집단모형									
				SPM1					SPM2				
									L_1^S L_2^S	BIAS	ARB(%)	RMSE	L_1^S L_2^S
I	$\hat{Y}_{COM}$	-	-			-5.80	3.88	6.35	-	-	-22.62	8.93	24.14
	$\hat{Y}_{PSA}$	-	-			-4.42	2.97	5.06	-	-	-3.32	3.18	10.01
	$\hat{Y}_{C-S}$	5	5	4	4	-4.45	2.99	5.04	8	8	-3.46	2.24	7.09
		8	8	6	6	-4.43	2.99	5.10	8	8	-3.54	2.17	6.85
	$\hat{Y}_{Raking}$	5	5	12	12	-4.42	2.97	4.97	15	15	-1.72	1.92	6.08
		8	8	20	20	-4.42	2.97	5.00	20	20	-1.25	1.52	4.83
II	$\hat{Y}_{COM}$	-	-			6.04	4.04	6.58	-	-	44.68	17.62	47.79
	$\hat{Y}_{PSA}$	-	-			4.60	3.11	5.22	-	-	2.53	3.02	9.64
	$\hat{Y}_{C-S}$	5	5	4	4	4.59	3.10	5.18	6	6	4.73	2.69	8.54
		8	8	6	6	4.63	3.13	5.31	6	6	5.19	2.84	9.41
	$\hat{Y}_{Raking}$	5	5	15	15	4.58	3.08	5.14	15	15	1.94	2.00	6.38
		8	8	16	16	4.59	3.09	5.17	20	20	1.73	1.65	5.32

<표 6> 선형 응답확률이고 로그-선형 초모집단인 경우($\alpha = 0$)

		로그-선형 초모집단모형											
RP	Est	L_1^P	L_2^P	SPM3				SPM4					
				L_1^S	L_2^S	BIAS	ARB(%)	RMSE	L_1^S	L_2^S	BIAS	ARB(%)	RMSE
I	$\hat{Y}_{COM}$	-		-		-463.85	17.55	518.65	-		-77.16	19.67	80.98
	$\hat{Y}_{PSA}$	-		-		-214.66	10.45	334.86	-		-29.17	8.50	39.93
	$\hat{Y}_{C-S}$	5	5	10	10	-191.46	9.18	295.22	6	6	-21.82	6.28	29.16
		8	8	10	10	-191.84	9.08	290.55	6	6	-22.53	6.43	29.83
	$\hat{Y}_{Raking}$	5	5	15	15	-148.41	8.05	264.63	15	15	-21.33	6.10	28.50
		8	8	20	20	-126.22	7.25	235.85	20	20	-20.74	5.87	27.12
	$\hat{Y}_{COM}$	-		-		1,245.24	46.80	1,421.08	-		175.93	44.83	185.60
	$\hat{Y}_{PSA}$	-		-		336.46	14.90	501.04	-		32.33	9.59	45.76
II	$\hat{Y}_{C-S}$	5	5	6	6	397.33	16.89	564.79	6	6	31.11	8.57	40.50
		8	8	6	6	407.94	17.39	593.57	6	6	31.95	8.82	42.22
	$\hat{Y}_{Raking}$	5	5	15	15	207.95	10.69	359.74	15	15	27.12	7.74	36.85
		8	8	16	16	208.31	10.37	355.47	16	16	26.34	7.52	35.23

<표 7> 선형 응답확률이고 선형과 2차 초모집단인 경우($\alpha = 0.5$)

		초모집단모형													
RP	Est	SPM1						SPM2							
		L_1^P	L_2^P	L_1^S	L_2^S	BIAS	ARB(%)	RMSE	L_1^P	L_2^P	L_1^S	L_2^S	BIAS	ARB(%)	RMSE
I	$\hat{\bar{Y}}_{COM}$	-		-		-6.15	3.58	6.72	-		-		-23.45	8.52	25.02
	$\hat{\bar{Y}}_{PSA}$	-		-		-4.09	2.43	4.84	-		-		-3.31	2.98	10.14
	$\hat{\bar{Y}}_{C-S}$	5	5	4	4	-4.09	2.42	4.71	8	8	8	8	-2.48	1.71	5.86
	$\hat{\bar{Y}}_{Raking}$	8	8	16	16	-4.06	2.40	4.68	8	8	20	20	-1.16	1.39	4.83
II	$\hat{\bar{Y}}_{COM}$	-		-		6.59	3.84	7.16	-		-		46.29	16.80	49.47
	$\hat{\bar{Y}}_{PSA}$	-		-		4.41	2.62	5.13	-		-		2.34	2.80	9.73
	$\hat{\bar{Y}}_{C-S}$	5	5	4	4	4.42	2.61	5.06	5	5	6	6	4.18	2.25	7.87
	$\hat{\bar{Y}}_{Raking}$	5	5	15	15	4.42	2.61	5.05	8	8	20	20	1.83	1.49	5.21

<표 8> 선형 응답확률이고 로그-선형 초모집단인 경우($\alpha = 0.5$)

		로그-선형 초모집단모형													
RP	Est	SPM3						SPM4							
		L_1^P	L_2^P	L_1^S	L_2^S	BIAS	ARB(%)	RMSE	L_1^P	L_2^P	L_1^S	L_2^S	BIAS	ARB(%)	RMSE
I	$\hat{Y}_{COM}$	-		-		-1,560.75	19.85	1,773.67	-		-		-515.03	22.26	544.64
	$\hat{Y}_{PSA}$	-		-		-826.03	13.04	1,248.16	-		-		-230.94	11.08	303.45
	$\hat{Y}_{C-S}$	8	8	20	20	-433.28	8.88	865.09	8	8	12	12	-121.97	6.39	180.16
	$\hat{Y}_{Raking}$	8	8	50	50	-292.45	8.08	799.55	8	8	30	30	-105.63	5.81	162.76
II	$\hat{Y}_{COM}$	-		-		4,307.86	54.34	5,027.40	-		-		1,259.90	54.45	1,344.62
	$\hat{Y}_{PSA}$	-		-		1,434.25	20.64	2,056.76	-		-		300.30	14.20	394.02
	$\hat{Y}_{C-S}$	8	8	6	6	1,428.18	20.08	1,999.58	5	5	6	6	229.92	10.77	301.94
	$\hat{Y}_{Raking}$	8	8	16	16	751.64	13.07	1,325.85	8	8	16	16	142.52	7.42	215.04

V. 결론

무응답을 적절히 처리하는 방법의 하나로 가중치 보정법과 무응답 대체법이 흔히 사용된다. 표본조사에는 관심변수에 영향을 주는 다변량 보조변수가 있을 수 있으며, 따라서 다변량 보조변수를 이용하여 가중치를 보정하거나 무응답 대체를 진행하면 더욱 정확한 추정결과를 얻을 수 있다. 본 연구에서는 기존의 데카르트 곱을 이용한 세부 층 구성 방법의 단점을 보완하기 위해 갈퀴법 사용을 제안하였으며 갈퀴법의 경우 이변량뿐 아니라 다변량에서도 적용이 가능하므로 보조변수의 차원과 무관하게 사용할 수 있다. 모의실험을 통하여 제안된 방법의 성능이 우수함을 확인하였으며 Lee & Shin(2024, a)에서 제안한 편향 보정 무응답 대체법에서 다변량 보조변수가 있는 경우로 확장이 가능할 것으로 판단한다. 다만 기존의 단변량 세부 층 구성 방법에서도 최소 응답확률이 ‘0’에 가깝고 각 보조변수의 세부 층 개수가 큰 경우에는 추정의 정확성이 떨어지므로 이를 해결하는 방법을 연구할 필요가 있다.

참고문헌

- Little, R.J.A. and D.B. Rubin. 2002. *Statistical Analysis with Missing Data* (2nd ed.). John Wiley & Sons.
- Bethlehem, J. 2020. "Working with Response Probabilities." *Journal of Official Statistics* 36(3): 647-674.
- Berg, E., J.K. Kim, and C. Skinner. 2016. "Imputation under Informative Sampling." *Journal of Survey Statistics and Methodology* 4(4): 436-462.
- Chung, H.Y. and K.I. Shin. 2022. "A Response Probability Estimation for Non-ignorable Non-response." *Communications for Statistical Applications and Methods* 29(2): 263-275.
- Da Silva, D.N. and J.D. Opsomer. 2006. "A Kernel Smoothing Method of Adjusting for Unit Non-response in Sample Surveys." *Canadian Journal of Statistics* 34 (4): 563-579.
- Da Silva, D.N. and J.D. Opsomer. 2009. "Nonparametric Propensity Weighting for Survey Nonresponse through Local Polynomial Regression." *Survey Methodology* 35(2): 165-176.
- Kim, J.K. and M. Riddles. 2012. "Some Theory for Propensity-score-adjustment Estimators in Survey Sampling." *Survey Methodology* 38(2): 157-165.
- Lee, Y. and K.-I. Shin. 2024.(a) "Bias-corrected Imputation Method for Non-ignorable Nonresponse with Heteroscedasticity in Super-population Model." *The Korean Journal of Applied Statistics* 37: 283-295.
- Lee, Y. and K.-I. Shin. 2024.(b) "A Study on Determining the Optimal Number of Strata for Propensity Score Estimation by Simulation Studies." *Journal of The Korean Official Statistics* 29(2): 1-23.
- Morikawa, K., J.K. Kim, and Y. Kano. 2017. "Semiparametric Maximum Likelihood Estimation with Data Missing Not at Random." *Canadian Journal of Statistics* 45(4): 393-409.
- Morikawa, K. and J.K. Kim. 2018. "A Note on the Equivalence of Two Semiparametric Estimation Methods for Nonignorable Nonresponse." *Statistics & Probability Letters* 140: 1-6.
- Riddles, M.K., J.K. Kim, and J. Im. 2016. "A Propensity-score-adjustment Method for Nonignorable Nonresponse." *Journal of Survey Statistics and Methodology*

4(2): 215-245.

Sim, J.Y. and K.I. Shin. 2021. "Bias Corrected Non-response Estimation Using Nonparametric Function Estimation of Super Population Model." *Korean Journal of Applied Statistics* 34(6): 923-936.

Yeo, L. and K.-I. Shin. 2023. "A Study to Improve the Accuracy of the Naive Propensity Score Adjusted Estimator Using Double Post-stratification Method." *The Korean Journal of Applied Statistics* 36(6): 547-559.

<접수 2024.07.25; 수정 2024.08.17; 게재확정 2024.08.17>

A Study on the Naive Propensity Score Adjusted Estimator Using the Raking Ratio Method for Multivariate Auxiliary Variables

Yujin Lee

(Hankuk University of Foreign Studies)

Key-Il Shin

(Hankuk University of Foreign Studies)

In sample surveys with nonresponse, propensity score adjusted(PSA) estimator is commonly used as population mean estimator. The PSA estimator is known to produce bias under MNAR nonresponse, so called, non-ignorable nonresponse. Recently, Chung & Shin (2022) proposed a method to improve the accuracy of the naive PSA estimator when non-ignorable nonresponse occurs in non-informative sampling. In that paper, the accuracy of estimation was improved by applying the post-stratification method using sub-strata to the naive PSA estimator. That method was effective for univariate auxiliary variable, and also effective for bivariate auxiliary variables. Additionally, Yeo & Shin(2023) improved the accuracy of the naive PSA estimator by using a double post-stratification adjusted method.

In this study, we propose a new method to improve the naive PSA estimator by using the raking ratio method for multivariate auxiliary variables. In particular, when the super-population model is a log-linear model, it is confirmed through simulation that the method proposed in this study is superior to the existing method.

Key words: post-stratification, response probability, super-population, non-informative sampling