The Korean Journal of Applied Statistics
2023, Vol. 36, No. 6, 547–559

DOI: https://doi.org/10.5351/KJAS.2023.36.6.547
Print ISSN1225-066X / Online ISSN 2383-5818

A study to improve the accuracy of the naive propensity score adjusted estimator using double post-stratification method

Leesu Yeo[a], Key-Il Shin[1,a]

[a]Department of Statistics, Hankuk University of Foreign Studies

Abstract

Proper handling of nonresponse in sample survey improves the accuracy of the parameter estimation. Various studies have been conducted to properly handle MAR (missing at random) nonresponse or MCAR (missing completely at random) nonresponse. When nonresponse occurs, the PSA (propensity score adjusted) estimator is commonly used as a mean estimator. The PSA estimator is known to be unbiased when known sample weights and properly estimated response probabilities are used. However, for MNAR (missing not at random) nonresponse, which is affected by the value of the study variable, since it is very difficult to obtain accurate response probabilities, bias may occur in the PSA estimator. Chung and Shin (2017, 2022) proposed a post-stratification method to improve the accuracy of mean estimation when MNAR nonresponse occurs under a non-informative sample design. In this study, we propose a double post-stratification method to improve the accuracy of the naive PSA estimator for MNAR nonresponse under an informative sample design. In addition, we perform simulation studies to confirm the superiority of the proposed method.

Keywords: MNAR nonresponse, weight adjustment, informative sampling

1. 서론

다수의 무응답이 조사통계에서 발생하고 있다. 무응답의 결측 메카니즘은 크게 세 가지로 나누어진다. 무응답이 관심변수 또는 보조변수에 영향을 받지 않지 않아 완전히 랜덤으로 발생하면 MCAR (missing completely at random)이라 하고, 보조변수에 영향을 받을 수 있으나 관심변수에 영향을 받지 않으면 MAR (missing at random)이라 한다. 반면 관심변수에 영향을 받으면 MNAR (missing not at random) 메카니즘이 된다. 관심변수에 영향을 받는 MNAR 무응답은 이를 적절히 처리하지 않으면 편향이 발생하게 되고, 특히 타당한 응답확률을 구하는 것이 어려워 무시할 수 없는 무응답(non-ignorable nonresponse; NN)이라 한다. 본 연구에서는 무시할 수 없는 무응답을 다룬다.

흔히 표본설계에서는 단순랜덤추출법을 사용하여 표본을 추출한다. 그러나 일부 표본설계에서는 관심변수의 함수인 표본 포함확률 또는 표본 가중치를 사용한다. 단순임의추출법과 같이 관심변수의 함수가 아닌 표본 가중치를 사용한 표본설계를 무정보적 표본설계(non-informative sampling)라 하고, 관심변수의 함수인 표본 가중치를 사용한 표본설계를 정보적 표본설계(informative sampling)라 한다. 이와 관련된 내용은 Pfeffermann

This research was supported by Hankuk University of Foreign Studies research fund (2023).
[1]Corresponding author: Department of Statistics, Hankuk University of Foreign Studies, 81 Oedae-ro, Yongin-si, Gyeonggi-do 17035, Korea. E-mail: keyshin@hufs.ac.kr

Published 31 December 2023 / journal homepage: http://kjas.or.kr

등 (1998, 2006)에 설명되어 있다. 정보적 표본설계를 사용했음에도 설계 당시의 표본 가중치를 고려하지 않은 추정량을 사용하게 되면 편향이 발생하는 것으로 알려져 있다. 본 연구에서는 정보적 표본설계(informative sampling)를 사용하며, 추정 시에 설계 당시의 알려진 표본 가중치를 사용한다고 가정한다.

본 연구에서는 성향점수보정(propensity score adjusted; PSA) 추정량의 정확성을 향상하는 방법을 연구한다. 성향점수보정 추정량은 흔히 모집단 평균의 추정량으로 사용하는 호르비츠-톰슨 추정량을 무응답을 적절히 처리하도록 확장한 추정량이다. 정보적 표본설계가 사용되었지만 알려진 표본설계 가중치를 사용한 성향점수보정 추정량은 불편추정량이다 (Kim과 Riddle, 2012). 또한, MAR 가정하에서 응답확률이 적절히 추정되고, 이를 사용한 성향점수보정 추정량은 편향이 발생하지 않는다. 그러나 MNAR 무응답은 응답 자료만 있으므로 정확한 응답확률 추정치가 얻어지지 않는다. 따라서 MNAR 무응답에서는 흔히 편향이 발생한다.

발생한 편향을 추정하기 위해서는 정보적 표본설계에서 사용한 설계 가중치를 사용하고, 응답확률과 응답확률 모형이 알려져 있어야 한다. 또한, 초모집단 모형의 분포가 알려져 있어야 한다. 그러나 현실적으로 설계 가중치 모형, 응답확률 모형, 초모집단 모형이 알려진 경우는 거의 없다. 특히 MNAR 무응답에서는 무응답으로 인해 조사된 관심변수 값만이 있으므로 응답확률을 정확히 추정하는 것은 불가능하다. 따라서 현실적으로 사용할 수 있는 방법은 MNAR 무응답이지만 MAR 무응답을 가정하여 응답확률을 추정하는 것이다. 이렇게 부정확한 응답확률을 사용한 성향점수보정 추정량을 나이브 성향점수보정 추정량(naive propensity score adjusted estimator)이라 한다. 나이브 성향점수보정 추정량은 편향 추정량이지만 분산은 정확한 응답확률을 사용한 추정량과 큰 차이가 없다는 것이 알려져 있으며 이와 관련된 자세한 내용은 Riddle 등 (2016)에서 확인할 수 있다. 또한, Chung과 Shin (2022)은 무시할 수 없는 무응답에서 나이브 성향점수보정 추정량의 정확성을 향상하는 방법을 연구하였다. 해당 논문에서는 주어진 모집단을 다수의 세부 층으로 나누는 사후층화 방법을 제안하였으며 사후층화 방법이 추정의 정확성을 향상하는 것을 보였다.

본 연구에서는 Chung과 Shin (2022)에서 사용한 무정보적 표본설계에서 무시할 수 없는 무응답에 적용한 사후층화보정 방법을 확장하여 정보적 표본설계에서 무시할 수 없는 무응답이 발생한 경우의 사후층화보정 방법을 연구한다. 특히 사업체 조사에서는 종사자 수가 증가할수록 사업체 수가 줄어드는 경우가 흔히 발생한다. 이러한 모집단에 대해 정보적 표본설계에서 무시할 수 없는 무응답 처리를 위해 기존의 방법을 사용할 경우 세부 층에 포함된 최종 응답 자료 수가 적은 세부 층이 발생하여 추정의 정확성이 떨어질 수 있다. 이에 본 연구에서는 Chung과 Shin (2017)이 처음으로 제안한 사후층화보정 방법을 확장하여 설계 가중치 보정과 응답확률 보정을 위한 세부 층을 각각 구성하여 보정하는 새로운 방법을 제안하였다. 즉 1차로 모집단에 포함된 보조변수와 표본으로 추출된 자료 수를 이용하여 세부 층을 구하고, 2차로 최종적으로 응답한 자료를 이용하여 세부 층을 구한 후, 두 세부 층 구성 방법을 종합한 이중 사후층화 보정 방법을 제안하였다.

본 논문의 구성은 다음과 같다. 먼저 2절에서 기존의 성향점수보정 추정량을 설명하였다. 특히 관심변수의 함수인 설계 가중치와 관심변수의 함수인 응답확률 모형에서 편향이 발생하는 과정을 설명하였다. 또한, MNAR 무응답에서 사후층화 방법을 사용함으로써 추정의 정확성이 향상되는 이유를 설명하였다. 3절에서는 주어진 설계 가중치를 보정하여 추정의 정확성을 향상하기 위한 사후층화보정 방법과 응답확률의 사후층화보정 방법을 동시에 사용하는 이중 사후층화보정 방법을 제안하였다. 4절에서는 모의실험을 통하여 제안된 방법의 우수성을 확인하였으며 5절에 결론을 수록하였다.

2. 성향점수보정 추정량과 편향수정 성향점수보정 추정량

2.1. 성향점수보정 추정량

무응답이 있는 자료의 모평균 추정량으로 흔히 사용하는 성향점수보정 추정량의 정의는 다음과 같다.

$$\hat{\bar{Y}}_{PSA} = \frac{1}{N}\sum_{i=1}^{n} w_i \frac{R_i}{\hat{p}_i} y_i, \tag{2.1}$$

여기서 N은 모집단 수이고 n은 표본 수이다. w_i는 표본 가중치로 $w_i = 1/\pi_i$이며, π_i는 설계 당시의 알려진 표본 포함확률이다. R_i는 응답변수이며 응답이면 1, 무응답이면 0인 지시변수이다. $\hat{p}_i$는 응답확률 또는 성향점수의 추정값이다. 만약 w_i가 관심변수 y_i의 함수이면 정보적 표본설계이고, 관심변수의 함수가 아니면 무정보적 표본설계이다. 또한, 응답확률 p_i가 관심변수의 함수이면 MNAR 무응답이 되어 무응답은 무시할 수 없는 무응답이 된다. 특히 무시할 수 없는 무응답이 발생한 경우에서 MAR 가정에서 구해진 응답확률을 사용한 성향점수보정 추정량은 불편추정량이 아니다. 이와 같은 성향점수보정 추정량을 나이브 성향점수보정 추정량이라 한다. 성향점수보정 추정량에 관한 내용은 Kim과 Riddles (2012)과 Riddles 등 (2016)을 살펴보기 바란다. 따라서 편향을 제거한 편향수정 성향점수보정 추정량의 사용은 추정의 정확성을 향상할 수 있는 하나의 방법이다. 그러나 편향을 추정하기 위해서는 이론적인 응답확률 모형, 초모집단 모형과 오차 분포가 알려져 있어야 한다. 또한, 표본 포함확률 모형도 알려져 있어야 한다. 그러나 모형과 분포가 알려진 경우는 현실적으로 매우 드물며, 따라서 이론적으로 정확한 편향을 추정하는 것은 많은 경우 불가능하다. 편향 추정에 관한 다수의 연구가 진행되었으며 이와 관련한 내용은 Chung과 Shin (2022), Pfeffermann 등 (1998, 2006)을 살펴보기 바란다.

2.2. 사후층화의 타당성

층화는 층별로 평균이 다른 경우, 전체 평균 추정의 정확성 향상에 도움을 준다. MNAR 무응답의 적절한 처리에서는 표본가중치와 응답확률 그리고 이에 해당하는 타당한 모형이 사용된다. 이에 추가하여 초모집단 모형이 형성된다는 가정을 사용한다. 초모집단 모형은 모집단에서 관심변수 y와 보조변수 x로 만들어지는 모형을 의미한다. 실제 자료 분석에서 가장 흔히 볼 수 있는 초모집단 모형은 관심변수 y는 매출액, 보조변수는 종사자 수로 만들어지는 모형이다. 물론 정확한 모형의 형태는 일반적이지 않지만, 흔히 종사자 수가 증가하면 매출액은 증가한다. 따라서 보조변수로 세부 층을 나누게 되면 각 세부 층에서 구한 관심변수의 평균은 다르게 된다. 결국, 초모집단 모형이 형성되면 층화의 효과를 얻을 수 있다. Bethlehem (2020)은 다양한 사후층화 방법을 설명하였다. 따라서 모집단을 보조변수로 층화하는, 즉 세부 층을 나누는 방법을 사용한다면 추정의 정확성은 향상될 수 있다. 결국, 사후층화의 효과는 세부 층의 평균이 층별로 증가하거나 감소할 때 매우 효과적이다. 또한, 세부 층의 모집단 개수와 세부 층의 표본 개수 정보를 설계 가중치 보정에 사용한다면 세부 층별로 정확한 설계 가중치가 계산되기 때문에 모수 추정의 정확성이 향상할 것으로 판단한다

2.3. 사후층화를 이용한 응답확률 추정

나이브 성향점수보정 추정량에서는 MAR 무응답 가정을 사용하기 때문에 응답확률 추정방법으로 로지스틱 회귀모형이 흔히 사용된다. 로지스틱 회귀모형은 90년대 이후부터 사용하였으며 Bethlehem (2020)에서도 사용하였다. 결국, 응답확률 p_i의 추정치는 사용 가능한 보조변수 x를 이용하여 얻을 수 있으며 로지스틱 회귀모형은 다음과 같이 정의된다.

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha_0 + \alpha_1 x_{1i} + \cdots + \alpha_k x_{ki}. \tag{2.2}$$

다수의 독립변수 또는 보조변수가 사용될 수 있다. 그러나 본 연구에서는 보조변수가 하나인 로지스틱 회귀모형을 이용하여 응답확률을 추정하였다. 또한, (2.2)에서 얻어진 응답확률의 합은 배정된 표본 수와 같도록 보정한다. 이를 위해 먼저 $\tilde{p}_i$를 (2.2)에서 얻어진 응답확률이라 하자. 그러면 보정된 $\hat{p}_i$는 $1/\hat{p}_i = 1/\tilde{p}_i \times n/(\sum_{i=1}^{r} 1/\tilde{p}_i)$

으로 얻어진다. 여기서 r은 최종 응답 자료 수이다.

Chung과 Shin (2022) 그리고 Sim과 Shin (2021)은 무정보적 표본설계에서 MNAR 무응답이 발생했을 때, 평균 추정의 정확성 향상을 위해 사후층화 방법을 사용하였다. 또한, Lee와 Shin (2022)에서는 MNAR 무응답이 발생하였을 때 대체 추정량(imputation estimator)의 정확성 향상을 위해 사후층화 방법을 사용하였다. 기존에 연구된 사후층화 방법은 모집단을 세부 층으로 나눈 후 이를 기초로 가중치를 보정하는 방법이다. 이를 간단히 설명하면 다음과 같다. 먼저 모집단에서 세부 층을 구성하기 위해 모집단에 포함된 보조변수 x_i를 이용하여 분위수를 구한다. 이때 세부 층의 층 개수 L은 $r/L \geq 10$이 되도록 결정한다. 구해진 보조변수의 분위수를 세부 층의 층 경계로 사용한다. 따라서 각 층에 최소 10개의 최종 응답 자료가 포함되도록 한다. 최적 층 개수에 관한 연구가 Min과 Shin (2018)에서 수행되었으며, 이후 Chung과 Shin (2022) 그리고 Sim과 Shin (2021)에서도 같은 방법이 사용되었다. 이에 본 연구에서도 같은 층 개수 결정 방법을 사용하였다. 결정된 층 경계를 이용하여 h층 안에 포함된 N_h, n_h, r_h를 각각 구한다. 여기서 N_h, n_h, r_h는 각각 h층에 포함된 모집단 수, 표본 수 그리고 최종 응답 자료 수이다.이제 잘 알려진 다음의 공식을 확장하여 사후층화보정 인자를 구한다. 흔히 세부 층을 사용하지 않는 경우는 잘 알려진 다음의 공식을 만족한다.

$$\sum_{i=1}^{n} w_i = N, \quad \sum_{i=1}^{r} \frac{1}{\hat{p}_i} = n. \tag{2.3}$$

따라서 세부 층이 구성된 경우에서도 보정된 가중치와 보정된 추정 응답확률이 세부 층별로 (2.3)과 동일한 수식이 만족되도록 한다. 즉 세부 층에서 (2.4)가 만족하도록 한다.

$$\sum_{i=1}^{n_h} w_{i\in h} = N_h, \quad \sum_{i=1}^{r_h} \frac{1}{\hat{p}_{i\in h}} = n_h, \quad h = 1, \ldots, L. \tag{2.4}$$

이제 사후층화보정 인자를 $f_h^N = N_h/(\sum_{i=1}^{n_h} w_{i\in h})$, $f_h^R = n_h/(\sum_{i=1}^{r_h} 1/(\hat{p}_{i\in h}))$이라 하면 보정 가중치와 보정 응답확률은

$$w_{i\in h}^S = w_{i\in h} \times f_h^N, \quad \hat{p}_{i\in h}^S = \hat{p}_{i\in h} \times \frac{1}{f_h^R}, \quad h = 1, \ldots, L \tag{2.5}$$

로 구해진다.

2.4. 단일 사후층화 성향점수보정 추정량

이제 $w_i^{A(S)} = w_i^S/\hat{p}_i^S$라 하자. 여기서 w_i^S와 $\hat{p}_i^S$는 (2.5)에서 얻어진 보정 가중치와 보정 응답확률이다. 최종적으로 세부 층별 가중치의 합이 모집단 수가 되어야 하므로 최종 단일 사후층화보정 가중치 $w_{i\in h}^{F(S)} = w_{i\in h}^{A(S)} \times N_h/(\sum_{i=1}^{r_h} w_{i\in h}^{A(S)})$가 구해지고 이 값을 추정에 사용한다. 즉, 최종적인 사후층화 나이브 성향점수보정 추정량은 다음과 같다.

$$\hat{\bar{Y}}_{PSA}^{(S)} = \frac{1}{N} \sum_{i=1}^{r} w_i^{F(S)} y_i. \tag{2.6}$$

(2.6)은 Chung과 Shin (2022) 그리고 Sim과 Shin (2021)에서 사용한 단일 사후층화 방법을 사용해 얻어진 사후층화 나이브 성형점수보정 추정량이다.

3. 이중 사후층화 방법을 이용한 성향점수보정 추정량

Chung과 Shin (2019)에서는 왜도가 큰 분포에 해당하는 경우인 사업체 조사의 전수층에 적용할 수 있는 방법을 제안하였다. 또한, Chung과 Shin (2020)에서 언급하였듯이, 모집단 분포의 왜도가 매우 큰 경우에서는 단일 사후층화 방법의 효율성이 떨어지는 것으로 알려져 있다. 이 절에서는 사업체 조사의 전수층 또는 이를 포함한 왜도가 큰 모집단에서 정보적 표본설계가 사용되고 MNAR 무응답이 발생한 경우에서 사용할 수 있는 이중 사후층화 방법을 제안한다.

3.1. 설계 가중치 보정을 위한 1차 사후층화

표본 포함확률 또는 설계 가중치 보정을 위한 1차 사후층화보정 방법은 배정된 표본 수를 이용하여 세부 층을 구성한다. 즉 모집단에 포함된 보조변수의 분위수를 층 경계로 하는 L_1개의 세부 층을 만든다. 이때 $n/L_1 \geq 10$이 되도록 L_1을 결정한다. 이제 $h^{(1)}, h^{(1)} = 1, \ldots, L_1$ 세부 층의 표본 가중치를 $w_{i\in h^{(1)}}$, 모집단 수를 $N_{h^{(1)}}$, 배정된 표본 수를 $n_{h^{(1)}}$라 하면

$$\sum_{i=1}^{n_{h^{(1)}}} w_{i\in h^{(1)}} = N_{h^{(1)}} \tag{3.1}$$

가 되도록 한다.

3.2. 응답확률 보정을 위한 2차 사후층화

표본으로 추출된 n개에서 r개의 최종 응답 자료를 얻었다고 하자. 2차 사후층화를 위해서는 1차 사후층화와 같은 방법으로 보조변수의 분위수를 구한 후, 이 분위수를 2차 세부 층의 경계로 한다. 2차 세부 층의 수 L_2는 $r/L_2 \geq 10$이 되도록 한다. 즉 이 방법은 기존의 단일 사후층화 방법을 그대로 사용한다. 다만 $n \geq r$이므로 $L_1 \geq L_2$가 되며 따라서 2차 사후층화에서 $h^{(2)} = 1, \ldots, L_2$층에 포함된 표본 수는 $n_{h^{(2)}}$, 최종 응답 표본 수는 $r_{h^{(2)}}$가 된다.

3.3. 이중 사후층화 최종 가중치

(2.1)의 성향점수보정 추정량에서 이중 사후층화를 이용할 경우의 추정량은 다음과 같이 정의된다.

$$\hat{\bar{Y}} = \frac{1}{N}\sum_{i=1}^{n} w_i^D \frac{R_i}{\hat{p}_i^D} y_i, \tag{3.2}$$

여기서 $w_i^D = w_i^D I(i \in h^{(1)}) = w_{i\in h^{(1)}}^D = w_{i\in h^{(1)}} \times f_{h^{(1)}}^D$, $f_{h^{(1)}}^D = N_{h^{(1)}}/(\sum_{i=1}^{n_{h^{(1)}}} w_{i\in h^{(1)}})$이고, $h^{(1)} = 1, \ldots, L_1$로 1차 세부 층 구성 방법으로 얻어진 세부 층을 사용한다. 또한,

$$\hat{p}_i^D = \hat{p}_i^D I\left(i \in h^{(2)}\right) = \hat{p}_{i\in h^{(2)}}^D = \hat{p}_{i\in h^{(2)}} \times \frac{1}{f_{h^{(2)}}^{DR}}, \quad f_{h^{(2)}}^{DR} = \frac{n_{h^{(2)}}}{\sum_{i=1}^{r_{h^{(2)}}} 1/(\hat{p}_{i\in h^{(2)}})}, \quad h^{(2)} = 1, \ldots, L_2 \tag{3.3}$$

이다. 만약 단일 사후층화 방법의 세부 층 개수와 세부 층 경계를 사용한다면 $n_{h^{(2)}} = n_h, r_{h^{(2)}} = r_h$이므로 $f_{h^{(2)}}^{DR} = f_h^R$이고 $\hat{p}_{i\in h^{(2)}}^D = \hat{p}_{i\in h}^D = \hat{p}_{i\in h}^S$로 보정된 응답확률은 동일하다.

3.4. 이중 사후층화 성향점수보정 추정량

이제 $w_i^{A(D)} = w_i^D/\hat{p}_i^D$라 하자. 여기서 w_i^D와 $\hat{p}_i^D$는 (3.2)와 (3.3)에서 얻어진 이중 보정 설계 가중치와 이중 보정 응답확률이다. 최종적으로 세부 층별 가중치의 합이 모집단 수가 되어야 하므로 $w_{i\in h^{(2)}}^{F(D)} = w_{i\in h^{(2)}}^{A(D)} \times (N_{h^{(2)}}/(\sum_{i=1}^{r_{h^{(2)}}} w_{i\in h^{(2)}}^{A(D)}))$

를 추정에 사용한다. 즉, 제안된 이중 사후층화 나이브 성향점수보정 추정량은 다음과 같다.

$$\hat{\bar{Y}}_{PSA}^{(D)} = \frac{1}{N}\sum_{i=1}^{r} w_i^{F(D)} y_i. \tag{3.4}$$

4. 모의실험

4.1. 모의실험 개요

자료 생성 과정과 모수 추정 과정은 다음의 단계별로 이루어졌다.

Step 1: 모집단 생성과정

[1] 보조 자료 x_i 생성

먼저 $x_i^* \overset{\text{iid}}{\sim} \text{Gamma}(1, 50)$를 따르는 $N = 10,000$개를 생성하였다. 다음으로 보조자료는 $x_i = 100 + x_i^*$을 이용하여 생성하였고, 또한 $x_i >= 300$인 자료는 삭제하였다. 따라서 보조변수 x_i는 100에서 300 사이의 값을 갖는다.

[2] 초모집단 모형

초모집단 모형은 정규분포를 따르는 단순 회귀모형과 로그-선형 모형을 고려하였다.

1) 단순 회귀모형

$$\mathrm{y}_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \sim N\left(0, \sigma^2\right), \tag{4.1}$$

여기서 $\beta_0 = 10, \beta_1 = 5, \sigma^2 = 400$을 사용하였다.

2) 로그-선형모형

$$\log(y_i) = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \sim N\left(0, \sigma^2\right), \tag{4.2}$$

여기서 $\beta_0 = 0.01, \beta_1 = 0.03, \sigma^2 = 0.3$을 사용하였다.

Step 2 : 표본추출과정

생성된 N개의 모집단 자료에서 $n = 500$개의 표본을 추출하였다. 이때 정보적 표본설계 방법과 무정보적 표본설계 방법을 각각 사용하였다. 정보적 표본설계는 표본 포함확률이 관심변수 y_i값에 선형관계가 있는 선형 표본 포함확률 모형을 사용하였고, 무정보적 표본설계는 관심변수와 무관한 단순랜덤추출법을 사용하여 추출하였다.

[3-1] 무정보적 표본설계를 위하여 N개의 모집단 자료에서 단순임의추출(simple random sample)로 $n = 500$개의 표본을 추출하였다.

[3-2] 정보적 표본설계를 위해 선형 표본포함확률 모형인 $\pi_i^* = b_0 + b_1 y_i$를 이용하였다. 이미 생성된 관심변수 자료 y_i의 최솟값에서의 표본 포함확률을 π_y^{*min}, y_i의 최댓값에서의 표본 포함확률을 π_y^{*max}라 할 때, $(\pi_y^{*min}, \pi_y^{*max}) = (0.9, 0.5)$를 만족하는 b_0, b_1을 구한 후, 이 값과 포아송 표본추출방법(Poisson sampling)을 이용하여 $n = 500$

개의 표본을 추출하였다. 또한, 최종 표본 포함확률은 $\pi_i = \pi_i^* \times (500/\sum \pi_i^*)$으로 구하였다. 물론 SAS와 R에서 최종 표본 포함확률을 계산해준다.

Step 3 : 응답 자료 생성 과정

본 모의실험에서는 응답확률 모형으로 선형 응답확률 모형과 로지스틱 응답확률 모형을 고려하였다.

1) 선형 응답확률 모형

선형 응답확률 모형은 응답이 관심변수 y_i와 관계가 있으며 그 관계가 선형인 경우를 의미한다. 이를 반영하기 위해 다음의 모형을 사용하였다.

$$\text{선형 응답 확률 모형 : } \; p_i = a_0 + a_1 y_i, \tag{4.3}$$

여기서 p_i는 i 번째 개체가 응답할 확률이며 a_1의 부호에 따라 관심변수가 커짐에 따라 응답률이 증가할 수도 감소할 수도 있다.

2) 로지스틱 응답확률 모형

로지스틱 응답확률은 관심변수와의 관계가 다음의 로지스틱 함수로 표현된다.

$$\text{로지스틱 응답 확률 모형 : } \; \log\left(\frac{p_i}{1-p_i}\right) = c_0 + c_1 y_i. \tag{4.4}$$

[4] 추출된 $n = 500$개의 표본에서 선형 응답확률 모형인 (4.3)을 이용하여 무응답을 생성하였다. 즉 자료 y_i의 최솟값에서의 응답확률을 p_y^{min}, y_i의 최댓값에서의 응답확률을 p_y^{max}라 할 때, $(p_y^{min}, p_y^{max}) = (0.9, 0.5), (0.5, 0.9), (0.9, 0.7), (0.7, 0.9)$를 사용하여 a_0, a_1을 계산하고 주어진 y_i에 따라 응답확률을 계산하였다. 얻어진 응답확률에 따라 응답 및 무응답 자료를 생성하였다. 같은 방법을 로지스틱 응답확률 모형인 (4.4)에도 적용하여 응답과 무응답 자료를 생성하였다.

[5] 응답한 최종 조사 자료를 r개라 하자. $(p_y^{min}, p_y^{max}) = (0.9, 0.7)$ 또는 $(p_y^{min}, p_y^{max}) = (0.9, 0.5)$인 경우는 전체 자료의 약 80% 이상의 응답률을 보였으나 $(p_y^{min}, p_y^{max}) = (0.7, 0.9)$ 또는 $(p_y^{min}, p_y^{max}) = (0.5, 0.9)$인 경우는 약 50%에서 70% 사이의 응답율을 보이고 있다. 이는 보조변수가 감마 분포를 따르기 때문으로 풀이된다.

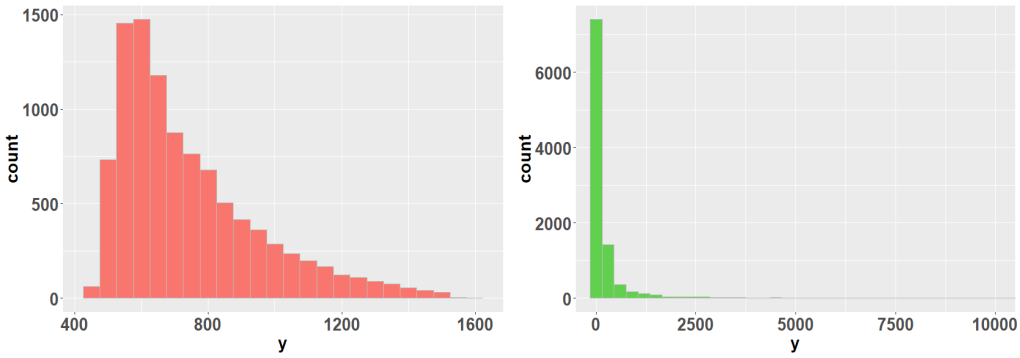
Step 4 : 층화

배정된 표본 자료 $(x_i, y_i), i = 1, \ldots, r$에서 r개의 최종 응답 자료의 가중치는 보정된다. 배정된 n개 표본의 표본포함확률을 보정하기 위해 L_1개의 세부 층으로 나누었다. 여기서 $n/L_1 \geq 10$을 기준으로 사용하였다. 또한, 추출된 n개에서 r개의 최종 응답 자료가 얻어졌으며 L_2개의 세부 층으로 나누었다. 여기서 2차 세부 층의 수 L_2는 $r/L_2 >= 10$이 되도록 하였다.

[6] $N = 10,000$이고 $n = 500$이며 $L_1 = 40$을 사용하였고, $L_2 = L = 20$을 사용하였다. 본 연구는 기존의 방법에 비해 이중 사후층화 보정 방법의 우수함을 보이는 것이 목적이므로 단일 사후보정 방법에 추가로 설계 가중치 보정을 사용할 때 효과가 있는지를 확인할 필요가 있다. 따라서 동일한 응답확률 보정 방법을 사용하는 것이 타당하므로 $L_2 = L = 20$을 사용하였다.

Step 5 : 최종 가중치 및 성향점수보정 추정량

[7] 나누어진 세부 층의 모집단 수와 조사된 자료 수 $(N_{h^{(1)}}, n_{h^{(1)}}, r_{h^{(1)}}), (N_{h^{(2)}}, n_{h^{(2)}}, r_{h^{(2)}})$를 이용하여 보정된 세부 층 가중치 w_i^S와 w_i^D를 구하였다.



(a) Linear super-population (b) Log-linear super-population

Figure 1: *Histogram of generated super-population.*

[8] 로지스틱 회귀모형을 이용하여 $\hat{p}_i$를 추정하고 (2.5)와 (3.3)을 이용하여 $\hat{p}_i^S, \hat{p}_i^D$를 계산하였다. 물론 $\hat{p}_i^S = \hat{p}_i^D$이다.
[9] 구해진 w_i^S와 w_i^D 그리고 $\hat{p}_i^S, \hat{p}_i^D$ 를 이용하여 최종 가중치 $w_i^{F(D)}, w_i^{F(S)}$를 구한 후 (2.6)과 (3.4)의 사후층화 성향점수보정 추정량을 구하였다. 또한, (2.1)의 사후층화를 하지 않은 성향점수보정 추정량 $\hat{\bar{Y}}_{PSA}$도 구하였다. 물론 (2.1)에서 $w_i^F = w_i(R_i/\hat{p}_i)$라 할 때 $\sum_{i=1}^n w_i^F = N$이 만족하도록 보정하였다. 이때 w_i는 표본설계 당시에 주어진 설계 가중치를 사용하고, $\hat{p}_i$는 로지스틱 회귀모형을 이용하여 얻어진 성향점수를 사용하였다.
이제 얻어진 평균 추정값은 다음의 비교통계량, 편향(bias), 절대상대편향(absolute relative bias; ARB) 그리고 제곱근 MSE (root mean squared error; RMSE)을 이용하여 결과의 성능이 비교되었다. 각 통계량의 정의는 다음과 같다.

$$\text{Bias} = \frac{1}{K}\sum_{k=1}^{K}\left(\hat{\bar{Y}}_k - \bar{Y}_k\right),$$

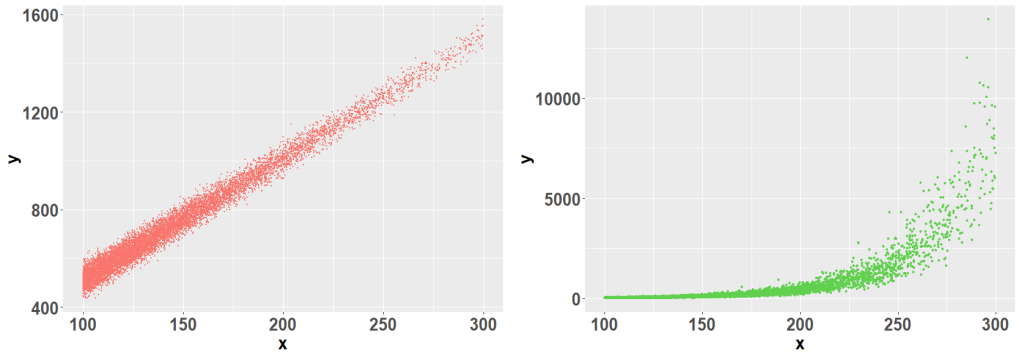
$$\text{ARB} = \frac{1}{K}\sum_{k=1}^{K}\frac{\left|\hat{\bar{Y}}_k - \bar{Y}_k\right|}{\bar{Y}_k},$$

$$\text{RMSE} = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(\hat{\bar{Y}}_k - \bar{Y}_k\right)^2},$$

여기서 K = 1,000을 사용하였으며, 반복마다 새로운 모집단을 생성하여 통계량을 계산하였다. 이는 생성된 특정 모집단의 영향을 줄이기 위함이다. 이에 k 번째 반복 모집단의 참값을 $\bar{Y}_k$로 표시하였다.

4.2. 기초 자료 분석

K = 1,000번 반복에서 하나의 모집단을 선택한 후 기초자료 분석을 수행하였다. 먼저 보조변수 x는 100에서 300 사이이며 평균과 분산은 각각 143.6과 1737.99로 얻어졌다. 선형 초모집단 모형에서, 관심변수 y는 454.8에서 1528.2 사이가 되고 평균과 분산은 각각 741.62와 43848.6로 얻어졌다. 또한, 로그-선형 모형을 초모집단 모형으로 사용한 경우, 관심변수 y는 8.168에서 13932.54가 되고 평균과 분산은 각각 272.85와 648936.1로 얻어졌다. 관심변수 자료의 히스토그램과 산점도가 Figure 1과 Figure 2에 나와 있다. Figure 1에서 관심변수 y는 꼬리가 긴 분포를 따르며 특히 초모집단 모형이 로그-선형 모형인 경우는 매우 긴 꼬리를 갖고 있다. Figure



(a) Linear super-population (b) Log-linear super-population

Figure 2: *Scatter plot of generated super-population.*

2를 살펴보면 보조변수가 증가할수록 관심변수도 증가하며 특히 보조변수가 큰 경우, 또는 관심변수가 큰 경우 자료 수가 매우 적은 것을 확인할 수 있다. 따라서 관심변수 값이 큰 경우에서 응답확률이 작은 경우 최종 얻어진 자료의 수는 매우 작게 된다.

4.3. 모의실험 결과

4.3.1. 정보적 표본설계 결과

Tables 1–4는 선형 표본포함확률을 사용한 정보적 표본설계 결과이다. 정보적 표본설계에서는 많은 경우 선형 표본 포함확률 모형을 사용하기 때문에 본 모의실험에서도 선형 표본 포함확률 모형을 사용하였다. 500개의 표본이 추출되었으며 응답확률 모형에 따라 최종적으로 약 250에서 450개의 최종 응답 자료가 얻어졌다. 정보적 표본설계에서 선형 초모집단 모형이고 선형 응답확률 모형이 사용된 결과인 Table 1을 살펴보면 모든 평균 추정량에서 편향이 발생하는 것을 확인할 수 있다. 이 결과는 나이브 성향점수보정 추정량이 불편 추정량이 아니라는 결과와 일치한다. 다음으로 기존의 단일 사후층화 성향점수보정 추정량인 $\hat{\bar{Y}}_{PSA}^{(S)}$는 ARB와 RMSE 측면에서 $\hat{\bar{Y}}_{PSA}$보다 우수한 결과를 주는 것을 확인할 수 있다. 물론 본 연구에서 제안한 이중 사후층화 성향점수보정 추정량 $\hat{\bar{Y}}_{PSA}^{(D)}$은 ARB와 RMSE 측면에서 가장 우수한 결과를 준다. 이러한 결과 추세는 정보적 표본설계에서 선형 초모집단 모형이고 로지스틱 응답확률 모형인 Table 2에서도 확인할 수 있다. 즉 본 연구에서 제안한 방법이 ARB와 RMSE 측면에서 가장 우수한 것을 확인할 수 있다. 따라서 응답확률 모형으로 가장 타당하다고 알려진 로지스틱 응답확률 모형에서도 제안한 방법이 가장 우수한 것을 확인할 수 있다. 다음으로 정보적 표본설계에서 선형 응답확률모형을 사용하고, 사업체 조사에서 주로 발생하는 로그-선형 초모집단 모형을 사용한 결과인 Table 3을 살펴보면 $\hat{\bar{Y}}_{PSA}^{(S)}$과 $\hat{\bar{Y}}_{PSA}^{(D)}$의 편향 결과는 응답확률의 형태에 따라 우수성이 달라진다. 즉 응답확률이 감소하는 경우에서는 $\hat{\bar{Y}}_{PSA}^{(D)}$의 결과가 우수하고, 응답확률이 증가하는 경우에서는 $\hat{\bar{Y}}_{PSA}^{(S)}$의 결과가 우수하다. 다만 사업체 조사에서는 사업체 규모가 커질 때 응답확률이 감소하는 경우가 흔히 일어나기 때문에 $\hat{\bar{Y}}_{PSA}^{(D)}$가 편향을 기준으로 할 때도 우수한 결과를 준다고 판단할 수 있다. 물론 ARB와 RMSE 측면을 살펴보면 본 연구에서 제안한 $\hat{\bar{Y}}_{PSA}^{(D)}$가 가장 우수한 결과를 준다. 또한, 정보적 표본설계에서 가장 타당하다고 알려진 로지스틱 응답확률 모형과 사업체 조사에서 가장 흔히 사용하는 로그-선형 초모집단 모형 결과인 Table 4 결과를 살펴보면 Table 3과 매우 유사한 패턴의 결과를 주는 것을 확인할 수 있다. 따라서 전체적으로 본 연구에서 제안한 $\hat{\bar{Y}}_{PSA}^{(D)}$가 가장 우수한 결과를 주는 것을 확인할 수 있다.

Table 1: Linear response probability model with linear super-population model (informative sampling)

$p_y^{\min}$	$p_y^{\max}$	r	Bias			ARB			RMSE		
			$\hat{\bar{Y}}_{PSA}$	$\hat{\bar{Y}}_{PSA}^{(S)}$	$\hat{\bar{Y}}_{PSA}^{(D)}$	$\hat{\bar{Y}}_{PSA}$	$\hat{\bar{Y}}_{PSA}^{(S)}$	$\hat{\bar{Y}}_{PSA}^{(D)}$	$\hat{\bar{Y}}_{PSA}$	$\hat{\bar{Y}}_{PSA}^{(S)}$	$\hat{\bar{Y}}_{PSA}^{(D)}$
0.9	0.7	425	−5.4952	−0.4376	−0.1617	0.0122	0.0026	0.0020	11.2517	2.4265	1.8952
0.9	0.5	401	−4.8984	−0.3146	−0.1038	0.0122	0.0030	0.0025	11.2114	2.8735	2.3959
0.5	0.9	299	−5.4189	−0.0090	0.1796	0.0121	0.0023	0.0020	11.1586	2.1668	1.8368
0.7	0.9	375	−5.5093	−0.2566	0.0105	0.0121	0.0022	0.0018	11.1842	2.0985	1.6756

Table 2: Logistic response probability model with linear super-population model(informative sampling)

$p_y^{\min}$	$p_y^{\max}$	r	Bias			ARB			RMSE		
			$\hat{\bar{Y}}_{PSA}$	$\hat{\bar{Y}}_{PSA}^{(S)}$	$\hat{\bar{Y}}_{PSA}^{(D)}$	$\hat{\bar{Y}}_{PSA}$	$\hat{\bar{Y}}_{PSA}^{(S)}$	$\hat{\bar{Y}}_{PSA}^{(D)}$	$\hat{\bar{Y}}_{PSA}$	$\hat{\bar{Y}}_{PSA}^{(S)}$	$\hat{\bar{Y}}_{PSA}^{(D)}$
0.9	0.7	432	−5.7052	−0.4798	−0.2069	0.0123	0.0025	0.0019	11.3420	2.3860	1.8449
0.9	0.5	416	−5.7515	−0.5566	−0.3745	0.0124	0.0029	0.0024	11.4714	2.7669	2.2742
0.5	0.9	313	−5.3812	−0.0724	0.1257	0.0121	0.0021	0.0018	11.1315	2.0158	1.7133
0.7	0.9	381	−5.5477	−0.2767	−0.0203	0.0122	0.0022	0.0017	11.1935	2.0335	1.6279

Table 3: Linear response probability model with log-linear super-population model (informative sampling)

$p_y^{\min}$	$p_y^{\max}$	r	Bias			ARB			RMSE		
			$\hat{\bar{Y}}_{PSA}$	$\hat{\bar{Y}}_{PSA}^{(S)}$	$\hat{\bar{Y}}_{PSA}^{(D)}$	$\hat{\bar{Y}}_{PSA}$	$\hat{\bar{Y}}_{PSA}^{(S)}$	$\hat{\bar{Y}}_{PSA}^{(D)}$	$\hat{\bar{Y}}_{PSA}$	$\hat{\bar{Y}}_{PSA}^{(S)}$	$\hat{\bar{Y}}_{PSA}^{(D)}$
0.9	0.7	449	−7.9132	−8.311	−6.575	0.141	0.107	0.092	56.459	42.239	37.611
0.9	0.5	448	−13.129	−11.701	−9.938	0.141	0.109	0.095	56.126	42.789	38.356
0.5	0.9	252	19.582	9.853	11.183	0.190	0.140	0.131	86.665	60.045	56.757
0.7	0.9	351	6.053	1.862	3.722	0.159	0.119	0.108	67.465	48.966	44.939

Table 4: Logistic response probability model with log-linear super-population model (informative sampling)

$p_y^{\min}$	$p_y^{\max}$	r	Bias			ARB			RMSE		
			$\hat{\bar{Y}}_{PSA}$	$\hat{\bar{Y}}_{PSA}^{(S)}$	$\hat{\bar{Y}}_{PSA}^{(D)}$	$\hat{\bar{Y}}_{PSA}$	$\hat{\bar{Y}}_{PSA}^{(S)}$	$\hat{\bar{Y}}_{PSA}^{(D)}$	$\hat{\bar{Y}}_{PSA}$	$\hat{\bar{Y}}_{PSA}^{(S)}$	$\hat{\bar{Y}}_{PSA}^{(D)}$
0.9	0.7	449	−6.659	−7.376	−5.675	0.140	0.106	0.091	56.177	42.009	37.238
0.9	0.5	449	−10.381	−10.217	−8.418	0.139	0.107	0.093	55.513	41.890	37.535
0.5	0.9	253	23.144	11.516	13.059	0.191	0.138	0.130	86.688	58.470	55.623
0.7	0.9	352	7.511	2.595	4.348	0.160	0.119	0.107	67.800	49.111	44.850

4.3.2. 무정보적 표본설계 결과

많은 표본설계에서는 무정보적 표본설계가 사용된다. 무정보적 표본설계에서 흔히 사용하는 표본추출방법은 단순임의추출법이므로 본 연구에서도 단순임의추출법으로 표본을 추출하였다. 무정보적 표본설계인 단순임의추출법에서 얻어진 Tables 5–8의 편향을 살펴보면 전체적으로 모든 추정량에서 편향이 발생하며 일부 결과에서는 단일 사후층화 또는 이중 사후층화로 인해 편향이 커지는 결과가 얻어졌다. 그러나 Tables 5– 8의 ARB와 RMSE를 기준으로 살펴보았을 때 본 연구에서 제안한 $\hat{\bar{Y}}_{PSA}^{(D)}$가 모든 표에서 가장 우수한 결과를 준다. 특히 이중 사후층화보정 방법은 응답확률이 감소하는 경우에서는 매우 큰 폭으로 정확성이 향상되는 것을 확인할 수 있다. 본 연구에서 제안한 이중 사후층화보정 방법은 기존의 단일 사후층화 방법에 설계 가중치의 사후보정 방법을 추가로 사용하기 때문에 정보적 표본설계에서 효과적일 것으로 예상되었으며, 이에 추가하여 무정보적 표본설계에서도 모의실험을 통하여 본 연구에서 제안한 방법이 효과적인 것을 확인하였다.

Table 5: Linear response probability model with linear super-population model (non-informative sampling)

$p_y^{\min}$	$p_y^{\max}$	r	Bias			ARB			RMSE		
			$\hat{\bar{Y}}_{PSA}$	$\hat{\bar{Y}}_{PSA}^{(S)}$	$\hat{\bar{Y}}_{PSA}^{(D)}$	$\hat{\bar{Y}}_{PSA}$	$\hat{\bar{Y}}_{PSA}^{(S)}$	$\hat{\bar{Y}}_{PSA}^{(D)}$	$\hat{\bar{Y}}_{PSA}$	$\hat{\bar{Y}}_{PSA}^{(S)}$	$\hat{\bar{Y}}_{PSA}^{(D)}$
0.9	0.7	423	−0.1194	−0.1964	−0.0859	0.0099	0.0018	0.0015	9.1126	1.7117	1.4591
0.9	0.5	397	0.2156	−0.1729	−0.0959	0.0101	0.0022	0.0019	9.2343	2.0963	1.7904
0.5	0.9	383	−0.0922	0.1445	0.2313	0.0099	0.0018	0.0018	9.0881	1.7072	1.6417
0.7	0.9	377	−0.0776	0.0121	0.1100	0.0099	0.0017	0.0015	9.0590	1.5527	1.3687

Table 6: Logistic response probability model with linear super-population model (non-informative sampling)

$p_y^{\min}$	$p_y^{\max}$	r	Bias			ARB			RMSE		
			$\hat{\bar{Y}}_{PSA}$	$\hat{\bar{Y}}_{PSA}^{(S)}$	$\hat{\bar{Y}}_{PSA}^{(D)}$	$\hat{\bar{Y}}_{PSA}$	$\hat{\bar{Y}}_{PSA}^{(S)}$	$\hat{\bar{Y}}_{PSA}^{(D)}$	$\hat{\bar{Y}}_{PSA}$	$\hat{\bar{Y}}_{PSA}^{(S)}$	$\hat{\bar{Y}}_{PSA}^{(D)}$
0.9	0.7	430	−0.1897	−0.2108	−0.1014	0.0099	0.0018	0.0016	9.1066	1.7014	1.4486
0.9	0.5	413	−0.2386	−0.2898	−0.2218	0.0100	0.0021	0.0018	9.1570	2.0382	1.7114
0.5	0.9	318	0.1650	0.1882	0.2669	0.0099	0.0017	0.0017	9.1168	1.6324	1.5562
0.7	0.9	383	−0.0333	−0.0147	0.0823	0.0099	0.0016	0.0015	9.0620	1.5052	1.3394

Table 7: Linear response probability model with log-linear super-population model (non-informative sampling)

$p_y^{\min}$	$p_y^{\max}$	r	Bias			ARB			RMSE		
			$\hat{\bar{Y}}_{PSA}$	$\hat{\bar{Y}}_{PSA}^{(S)}$	$\hat{\bar{Y}}_{PSA}^{(D)}$	$\hat{\bar{Y}}_{PSA}$	$\hat{\bar{Y}}_{PSA}^{(S)}$	$\hat{\bar{Y}}_{PSA}^{(D)}$	$\hat{\bar{Y}}_{PSA}$	$\hat{\bar{Y}}_{PSA}^{(S)}$	$\hat{\bar{Y}}_{PSA}^{(D)}$
0.9	0.7	449	−5.995	−7.160	−5.658	0.122	0.086	0.078	46.854	33.444	30.454
0.9	0.5	448	−11.309	−11.175	−9.840	0.124	0.091	0.080	47.316	34.208	31.056
0.5	0.9	252	19.339	10.447	12.142	0.166	0.117	0.114	66.695	48.548	48.183
0.7	0.9	351	7.407	1.886	3.521	0.139	0.097	0.090	54.674	38.237	36.446

Table 8: Logistic response probability model with log-linear super-population model(non-informative sampling)

$p_y^{\min}$	$p_y^{\max}$	r	Bias			ARB			RMSE		
			$\hat{\bar{Y}}_{PSA}$	$\hat{\bar{Y}}_{PSA}^{(S)}$	$\hat{\bar{Y}}_{PSA}^{(D)}$	$\hat{\bar{Y}}_{PSA}$	$\hat{\bar{Y}}_{PSA}^{(S)}$	$\hat{\bar{Y}}_{PSA}^{(D)}$	$\hat{\bar{Y}}_{PSA}$	$\hat{\bar{Y}}_{PSA}^{(S)}$	$\hat{\bar{Y}}_{PSA}^{(D)}$
0.9	0.7	449	−4.948	−6.472	−4.895	0.121	0.086	0.077	46.602	33.189	30.182
0.9	0.5	449	−8.920	−9.769	−8.324	0.122	0.087	0.078	46.598	33.317	30.271
0.5	0.9	253	23.144	11.946	13.701	0.169	0.114	0.112	67.536	47.556	47.363
0.7	0.9	352	9.109	2.939	4.593	0.140	0.097	0.090	54.981	38.500	36.694

5. 결론

본 논문에서는 정보적 표본설계에서 MNAR 무응답이 발생하였을 때, 나이브 PSA 추정량의 정확성을 향상하는 방법을 제안하였다. MNAR 무응답이 발생하였을 때 나이브 PSA 추정량은 편향 추정량이므로 정확성이 떨어진다. 기존의 연구에서 정확성을 향상하는 방법으로 단일 사후층화 방법이 제안되었으며, 본 연구에서는 이를 확장한 이중 사후층화 방법을 제안하였다. 모의실험 결과에 의하면 제안한 이중 사후층화 방법은 기존의 단일 사후층화 방법에 비해 우수한 결과를 주는 것으로 확인되었으며 특히 국내 현실과 같이 기업의 크기가 커질수록 응답확률이 떨어지는 경우에서는 매우 효과적인 것으로 확인되었다. 또한, 제안된 방법은 사후층화 방법을 두 번 사용하는 것이므로 기존의 사후층화 방법에 적용하기 쉽다는 장점이 있다. 다만 사후층화 방법에서 세부 층 개수를 최적으로 결정하는 방법이 연구될 필요가 있으며 실제 자료 분석에서 세부 층에 포함된 최종 자료 수가 적을 경우는 세부 층을 병합하는 방법을 사용하는 것이 필요하다고 판단된다. 또한, 본 연구에

서는 알려진 표본 가중치를 사용하였는데 표본 가중치가 알려지지 않아 추정된 가중치를 사용한 경우에서도 본 논문에서 제안한 방법이 효과적인지도 살펴볼 필요가 있다.

References

Bethlehem J (2020). Working with response probabilities, *Journal of Official Statistics*, **36**, 647–674.

Chung HY and Shin KI (2017). Estimation using informative sampling technique when response rate follows exponential function of variable of interest, *Korean Journal of Applied Statistics*, **30**, 993–1004.

Chung HY and Shin KI (2019). Bias adjusted estimation in a sample survey with linear response rate, *Korean Journal of Applied Statistics*, **32**, 631–642.

Chung HY and Shin KI (2020). A study on non-response bias adjusted estimation in business survey, *Korean Journal of Applied Statistics*, **33**, 11–23.

Chung HY and Shin KI (2022). A response probability estimation for non-ignorable non-response, *Communications for Statistical Application and Methods*, **29**, 263–275.

Kim JK and Riddles MK (2012). Some theory for propensity-score-adjustment estimators in survey sampling, *Survey Methodology*, **38**, 157–165.

Lee MH and Shin KI (2022). Bias corrected imputation method for non-ignorable non-response, *Korean Journal of Applied Statistics*, **35**, 485–499.

Min JW and Shin KI (2018). A study on the determination of substrata using the information of exponential response rate by simulation studies, *Korean Journal of Applied Statistics*, **31**, 621–636.

Pfeffermann D, Krieger AM, and Rinott Y (1998). Parametric distributions of complex survey data under informative probability sampling, *Statistica Sinica*, **8**, 1087–1114.

Pfeffermann D, Mour F, and Silva PN (2006). Multi-level modelling under informative sampling, *Biometika*, **93**, 943–959.

Riddles MK, Kim JK, and Im J (2016). A propensity-score-adjustment method for nonignorable nonresponse, *Journal of Survey Statistics and Methodology*, **4**, 215–245.

Sim JY and Shin KI (2021). Bias corrected non-response estimation using nonparametric function estimation of super population model, *Korean Journal of Applied Statistics*, **34**, 923–936.

Received April 28, 2023; Revised July 4, 2023; Accepted July 10, 2023

나이브 성향점수보정 추정량의 정확성 향상을 위한 이중 사후층화 방법 연구

여이수[a], 신기일[1,a]

[a]한국외국어대학교 통계학과

요 약

표본조사에서 무응답의 적절한 처리는 추정의 정확성을 향상한다. 결측 메카니즘이 MCAR (missing completely at random) 또는 MAR (missing at random)인 경우에서는 이를 적절히 처리할 수 있는 다양한 방법이 연구되었다. 무응답이 발생하였을 때 사용하는 평균 추정량으로 흔히 성향점수보정 추정량이 사용되며 MAR 또는 MCAR 무응답인 경우, 알려진 표본 가중치와 타당한 방법으로 추정된 응답확률을 사용할 수 있으므로 성향점수보정 추정량은 불편추정량이 된다. 그러나 관심변수 값에 영향을 받는 무응답인 MNAR (missing not at random) 무응답에서는 정확한 응답확률을 구하는 것이 어려워 성향점수보정 추정량에 편향이 발생할 수 있다. Chung과 Shin (2017, 2022)은 무정보적 표본설계에서 MNAR 무응답이 발생하였을 때 평균 추정의 정확성을 향상하는 방법으로 단일 사후층화 방법을 제안하였다. 본 연구에서는 정보적 표본설계를 사용하고, MNAR 무응답이 발생한 경우에서 나이브 성향점수보정 추정량의 정확성 향상을 위한 이중 사후층화 방법을 제안하였다. 또한, 모의실험을 통해 제안된 방법의 우수성을 확인하였다.

주요용어: MNAR 무응답, 가중치 보정, 정보적 표본설계

이 연구는 2023년 한국외국어대학교 교내연구비 지원을 받아 수행되었음.

[1]교신저자 : (17035) 경기도 용인시 처인구 모현읍, 한국외국어대학교 통계학과. E-mail : keyshin@hufs.ac.kr