The Korean Journal of Applied Statistics (2020) DOI: https://doi.org/10.5351/KJAS.2020.33.4.409
33(4), 409–420

A study on non-response bias adjusted estimation for take-all stratum

Hee Young Chung[a] · Key-Il Shin[a,1]

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

(Received May 8, 2020; Revised June 30, 2020; Accepted July 1, 2020)

Abstract

In business survey, modified cut-off sampling is commonly used to greatly increase the accuracy of the estimation while reducing the number of samples. However, non-response rate of take-all stratum has increased significantly and the sample substitution is not possible because the non-response in the take-all stratum affects the accuracy of the estimation. It is important to adjust the bias appropriately if non-response is affected by the variable of interest. In this study, a bias adjusted estimation is proposed as an appropriate method to deal with a non-response in the take-all stratum. In particular, the estimator proposed by Chung and Shin (2020) was applied to the bias adjustment for the take-all stratum; therefore, we suggest a new method to adjust properly for the take-all stratum. The superiority of the proposed estimator was examined through simulation studies and confirmed through actual data analysis.

Keywords: super population model, linear response rate model, power response rate model, gamma distribution, log-normal distribution

1. 서론

사업체조사에서는 흔히 수정절사법(modified cut-off sampling)이 사용된다. 수정절사법 또는 간단히 절사법은 적은 수의 표본 개수를 이용하여 상대적으로 우수한 정확성을 확보할 수 있다. 그러나 전수층(take-all stratum)에서 발생하는 다수의 무응답은 추정의 정확성을 크게 떨어뜨리고 있어 절사법 사용을 제한하고 있다. Hidroglou (1986)는 층화추출법의 특수한 경우로 표본층과 전수층으로 나누어 표본설계를 하는 수정절사표본설계법을 제안하였다. 이후 관련된 다수의 연구가 진행되었으며 Lavallee와 Hidiroglou (1988)는 여러 개의 표본층을 포함하는 수정 절사표본설계에 사용할 수 있는 LH 알고리즘을 제안하였다. 다수의 사업체 조사에서 절사법이 사용되고 있으며 이를 통해 추정의 정확성을 크게 향상시켰다. 그러나 최근 전수층에서 발생하는 다수의 무응답으로 인해 추정의 정확성이 떨어지는 문제점을 해결하기 위해 Lee와 Shin (2016)은 비용함수를 고려하여 전수층 규모를 축소하기 위한 최적 절사점 결정 방법을 제안하였다. 또한 조사업체에서도 전수층 단위무응답을 극복하기 위한 다양한 방법을 사용

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2018R1D1A1B07042736).
[1]Corresponding author: Professor, Department of Statistics, Hankuk University of Foreign Studies, 81 Oedae-ro, Yongin-si, Gyeonggi-do 17035, Korea. E-mail: keyshin@hufs.ac.kr

하고 있으나 실질적으로 무응답을 크게 줄이지 못하고 있다. 따라서 발생된 무응답의 영향력을 줄일 수 있는 적절한 통계적 처리 방법이 필요하다.

이를 해결할 수 있는 방법의 하나가 무응답으로 발생된 결측치(missing value)를 대체(imputation)하는 방법이다. 그러나 이 방법이 효과를 얻기 위해서는 충분한 보조정보가 존재해야 하는데 전수층에서 발생하는 무응답은 많은 경우 단위무응답(unit non-response)이므로 효과적인 결측치 대체가 쉽지 않다. 또한 전수층에는 표본대체(sample substitution)를 위한 예비표본을 준비할 수 없기 때문에 적절한 무응답 처리에 어려움이 가중되고 있다. 결국 현실적으로 전수층 무응답 처리에 사용 가능한 방법은 가중치 보정 방법(weight adjustment method)이다. 물론 이상점 또는 극단치가 있는 경우에는 이를 먼저 고려해야한다.

가중치 보정 방법에서 사용되는 가장 기본적인 가정은 무응답이 missing at random (MAR)를 따른다는 것이다. 주어진 자료가 이 가정을 만족한다면 흔히 사용하는 가중치 보정 방법을 사용할 때 편향을 발생시키지 않으며 효과적으로 무응답을 처리할 수 있다. 그러나 관심변수가 보조변수에 영향을 받는 초모집단모형(super population model)이 형성되고 응답이 관심변수에 영향을 받으면 흔히 사용하는 가중치 보정 방법은 큰 편향을 발생시킨다. 따라서 초모집단모형과 응답률 함수를 고려한 새로운 가중치 보정 방법이 사용되어야 한다.

최근 관심변수와 보조변수 간에 초모집단 모형이 성립되고 관심변수와 표본포함확률에 관계가 있는 경우에 사용할 수 있는 정보적 표본설계에 관한 연구가 Pfeffermann 등 (1998)과 Pfeffermann 등 (2006)에 의해 연구되었으며 Chung과 Shin (2017)은 표본포함확률 모형을 응답률 모형으로 바꾸어 편향을 추정하는 연구를 수행하였다. 이후 Chung과 Shin (2020)은 사업체조사에서 발생한 무응답을 적절히 처리하기 위한 편향보정 추정량을 제안하였다. 이 연구에서는 사업체조사 자료가 감마분포 또는 로그-정규분포를 따르고 응답률이 선형 또는 파워형 모형을 따를 경우의 편향보정 추정량을 제안하였다. 또한 Jeon과 Shin (2019)은 전수층에서 발생한 무응답을 적절히 처리하기 위한 방법을 연구하였으나 편향을 고려하지 않은 가중치 보정 방법을 제안하였다.

이에 새로운 세부 층 구성 방법을 이용한 전수층 편향보정 추정 방법을 제안하였다. 즉 Chung과 Shin (2020)에서 제안한 편향보정 추정량을 전수층 무응답 처리에 적용하였다. 그러나 전수층은 표본층과 달리 표본가중치(sample weight)가 '1'이고 또한 보조변수의 분포가 오른쪽으로 꼬리가 긴 경우가 실질적으로 많기 때문에 기존의 세부 층 구성 방법을 사용할 경우 효과가 떨어질 수 있다. 이에 본 연구에서는 전수층 편향보정을 위해 사용할 수 있는 새로운 방법을 제안하였다.

본 논문의 구성은 다음과 같다. 먼저 2절에서는 Jeon과 Shin (2019)과 Chung과 Shin (2020)의 결과를 간단히 살펴보았다. 3절에서는 모의실험을 통하여 본 연구에서 제안한 방법의 우수성을 살펴보았다. 4절에서는 본 연구에서 제안한 방법을 적용하여 문화체육관광 사업체의 실제 자료를 분석하였으며 5절에는 결론이 있다.

2. 제안된 편향보정 추정법

2.1. 전수층 가중치 보정 방법

이 절에서는 Jeon과 Shin (2019)에서 제안한 가중치 보정 방법을 살펴보았다. 먼저 전수층의 보조변수는 오른쪽으로 꼬리가 긴 분포를 따르는 것을 고려하여 감마분포를 가정하였으며 설계가중치는 '1'이다. 또한 관심변수와 보조변수는 초모집단 모형으로 선형 회귀모형을 따른다고 가정한다. 이러한 가정 하에서 모집단에 포함된 보조변수의 분위수 기반으로 전수층을 L개의 세부 층(substratum)으로 나누고 이때 구성된 세부 층에서 조사 자료 수와 모집단 자료 수를 이용하여 보정 가중치를 구한다. 전수층에서

단위무응답이 발생하기 때문에 보정된 가중치는 '1'보다 크게 되며 가중치 합은 모집단 수가 된다. 이제 MAR 가정 하에서 사용하는 평균 추정량을 $\hat{\bar{Y}}_S$, 세부 층 이용 가중치 보정 평균 추정량을 $\hat{\bar{Y}}_{ST}$라 하면 사용된 추정량은 다음과 같다.

$$\hat{\bar{Y}}_S = \frac{1}{N}\sum_{i=1}^{r} w^F y_i, \tag{2.1}$$

$$\hat{\bar{Y}}_{ST} = \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{r_h} w_{hi}^F y_i, \tag{2.2}$$

여기서 N은 전수층 모집단 수, r은 최종 자료 수로 $w^F = N/r$이고 y_i는 응답 자료 값이다. 또한 L은 세부 층 개수, r_h는 h 세부 층의 최종 자료 수이다. 또한 N_h를 h 세부 층의 모집단 수라 하면 $w_{hi}^F = w_h^F = N_h/r_h$로 구해진다.

2.2. 편향보정 추정량

무응답이 관심변수에 영향을 받는 경우에 사용될 수 있는 편향보정 추정량은 Chung과 Shin (2017, 2020)과 Min과 Shin (2018)에서 연구되었다. 본 연구는 전수층에서 발생한 무응답을 적절히 처리할 수 있는 방법을 연구하는 것이 목적이므로 사업체조사와 관련된 내용을 연구한 Chung과 Shin (2020) 결과를 살펴보았다. 즉 초모집단 모형의 오차가 감마분포 또는 로그-정규분포를 따른다고 가정하고 응답률은 선형 또는 파워형을 따른다고 가정한다. 이러한 감마분포와 로그-정규분포 가정은 Lee와 Shin (2016)에서도 사용되었다. 먼저 감마분포와 로그-정규분포를 이용한 초모집단 모형은 다음과 같다.

• 감마분포 초모집단 모형:

$$y_i \overset{\text{iid}}{\sim} \text{Gamma}\left(\alpha, \frac{\mu_i}{\alpha}\right), \quad \mu_i = \exp\left(\beta_0 + \beta_1 x_i\right). \tag{2.3}$$

• 로그-정규분포 초모집단 모형:

$$\log\left(y_i\right) = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right), \quad \mu_i = \exp\left(\beta_0 + \beta_1 x_i + \frac{\sigma^2}{2}\right). \tag{2.4}$$

또한 선형과 파워형 응답률 모형은 각각 다음과 같다.

• 선형모형:

$$\pi_i = b_0 + b_1 y_i \quad \text{또는} \quad \frac{1}{w_{hi}^F} = b_0 + b_1 y_{hi} + \eta_i. \tag{2.5}$$

• 파워형 모형:

$$\log\left(\pi_i\right) = c_0 + c_1 \log\left(y_i\right) \quad \text{또는} \quad \log\left(\frac{1}{w_{hi}^F}\right) = c_0 + c_1 \log\left(y_{hi}\right) + \eta_i, \tag{2.6}$$

여기서 π_i는 응답률이며, η_i는 독립이고 $E(\eta_i) = 0$, $\text{Var}(\eta_i) = \sigma_\eta^2$을 따른다고 가정한다.

이러한 조건 하에서 Chung과 Shin (2020)은 다음의 편향보정 추정량을 얻었다.

$$\hat{\bar{Y}}_{LG} = \hat{\bar{Y}}_{ST} - \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{n_{hi}} w_{hi}^F \left(\frac{\hat{b}_1}{\hat{b}_0 + \hat{b}_1 \hat{\mu}_i^{(s)}} \times \frac{\hat{\mu}_i^{(s)}}{\hat{\alpha}}\right), \tag{2.7}$$

$$\hat{\bar{Y}}_{LL} = \hat{\bar{Y}}_{ST} - \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{n_{hi}} w_{hi}^{F}\left(\frac{\hat{b}_1\hat{\mu}_i^{(s)2}}{\hat{b}_0+\hat{b}_1\hat{\mu}_i^{(s)}} \times \left(\exp\left(\hat{\sigma}^2\right)-1\right)\right), \tag{2.8}$$

$$\hat{\bar{Y}}_{PG} = \hat{\bar{Y}}_{ST} - \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{n_{hi}} w_{hi}^{F}\left(\frac{\hat{c}_1}{\hat{\alpha}+\hat{c}_1}\hat{\mu}_i^{(s)}\right), \tag{2.9}$$

$$\hat{\bar{Y}}_{PL} = \frac{\hat{\bar{Y}}_{ST}}{\exp\left(\hat{c}_1\hat{\sigma}^2\right)}, \tag{2.10}$$

여기서 추정량의 앞 첨자인 L과 P는 선형과 파워형 응답률을 의미하고 뒤 첨자 G와 L은 각각 감마분포와 로그-정규 분포를 의미한다. 또한 $\hat{\alpha}$과 $\hat{\sigma^2}$은 초모집단 모형식인 (2.3)과 (2.4)를 이용하여 추정되며 b_0, b_1과 c_0, c_1은 응답률 모형인 (2.5)와 (2.6)을 이용하여 추정된다.

각 추정량과 $\hat{\bar{Y}}_{ST}$의 비교를 통해 편향의 크기를 계산할 수 있으며 일부 자료의 경우 극히 드물지만 매우 큰 편향이 얻어질 수 있다. 이는 전수층의 경우 초모집단 모형 구축에 사용된 자료 수가 적어 모수 추정이 정확하지 않을 수 있으며 또한 세부 층의 개수인 L이 작아 응답률 모형의 모수 추정이 정확하지 않을 수 있다. 특히 $\hat{\mu}_i^{(s)}$은 로그-선형 모형을 적합한 후 얻어진 값에 지수함수를 이용한 재변환(re-transformation)을 한 후 얻어지기 때문에 특이치가 얻어질 수 있다. 따라서 극히 일부의 경우이지만 매우 큰 편향추정값에 의해 $\hat{\bar{Y}}_{ST}$에 비해 매우 큰 차이를 보이는 편향보정 추정량이 얻어질 수 있다. 이는 과대편향추정값이 얻어진 경우이므로 편향보정 추정량을 사용하지 않는 것이 타당하다. 이에 본 연구에서는 $\hat{\bar{Y}}_{ST}$와 $\pm 30\%$ 이상 차이가 있는 경우에는 편향 추정이 잘못된 것으로 판단하여 편향을 '0'으로 하였다. 즉 편향보정 추정량 대신에 $\hat{\bar{Y}}_{ST}$를 사용하였다.

2.3. 세부 층 구성 방법

세부 층을 나누는 방법은 매우 다양하다. Min과 Shin (2018)은 세부 층을 나누는 방법으로 모집단에 포함된 보조 자료를 등간격으로 나누는 방법보다 분위수를 이용한 세부 층 구성 방법이 우수한 결과를 주는 것을 확인하였다. 그러나 전수층의 경우에는 모집단 수가 적고 많은 경우 응답률이 매우 낮기 때문에 자료의 형태에 따라서는 Min과 Shin (2018) 방법을 사용할 경우 편향보정 효과가 떨어질 수 있다. 이에 본 연구에서는 세부 층 구성 방법으로 모집단에 포함된 보조 자료의 분위수로 세부 층 경계를 구성하는 방법 이외에 조사된 최종 자료의 분위수를 이용하는 방법도 고려하였다. 특히 큰 관심변수 값에 해당되는 응답률은 낮고, 작은 관심변수 값에서 높은 응답률을 보이는 경우에는 기존의 세부층 구성 방법을 사용할 경우 효과가 떨어지는 것을 확인하였으며 이 경우에는 조사된 자료의 분위수를 사용하는 것이 더 효과적인 것을 확인하였다. 이에 본 연구에서는 응답률 함수의 기울기가 음수이면 모집단 보조 자료의 분위수를 이용하고 응답률 함수의 기울기가 양수이면 최종 조사된 보조 자료의 분위수를 이용하여 층을 구성하는 방법을 제안하였다. 물론 기울기가 알려지지 않은 경우에는 기울기값을 추정하여 이용하면 된다. 이제 전수층의 모집단 보조 자료 분위수를 이용하여 구한 세부 층 구성 방법을 M_1이라 하고 최종 조사된 보조 자료 분위수를 이용하여 세부 층을 구성하는 방법을 M_2라 하자. 또한 응답률 모형의 기울기를 기반으로 층을 구성하는 방법을 M이라 하자. 즉 방법 M은 응답률 함수의 기울기를 δ라 할 때 다음의 방법으로 얻어진다.

$$M = \begin{cases} M_1, & \text{if } \delta \le 0, \\ M_2, & \text{if } \delta > 0. \end{cases}$$

실제 자료분석에서는 기울기 δ의 부호를 알 수가 없기 때문에 이를 추정해야하며 식 (2.5)의 경우 $\delta = b_1$이고 식 (2.6)의 경우에는 $\delta = c_1$이므로 식 (2.5)와 (2.6)을 이용하여 δ의 부호를 추정한 후 사용한다.

3. 모의실험 설계 및 결과

3.1. 모의실험 설계

모의실험을 위한 자료생성 과정과 모수추정 방법은 Jeon과 Shin (2019) 그리고 Chung과 Shin (2020) 방법과 유사하다. 다만 본 모의실험에서는 전수층 분석에 타당하도록 자료생성 과정과 모수 추정 방법을 수정하여 사용하였다.

• Step 1: 모집단 생성과정

초모집단모형이 회귀모형이고 모형의 오차가 감마분포 또는 로그-정규분포인 경우의 정보적 표본설계를 위한 전수층 자료생성 과정은 다음과 같다.

(1) 보조변수 x_i 생성: $x_i = 300 + \gamma_i, i = 1, \ldots, N$.
여기서 $\gamma_i \sim \text{Gamma}(1, 50)$이다. 따라서 보조변수 x_i는 300 이상의 값을 갖는다.

(2) 초모집단 모형 : 감마분포의 경우는 식 (2.3)을 이용하며 초모집단 모형인 $\mu_i = \exp(\beta_0 + \beta_1 x_i)$에서 $\beta_0 = 0.01, \beta_1 = 0.02, \alpha = 10$을 사용한다. 또한 로그-정규분포의 경우는 식 (2.4)를 사용하며 초모집단 모형인 $\mu_i = \exp(\beta_0 + \beta_1 x_i + \sigma^2/2)$에서 $\beta_0 = 0.01$, $\beta_1 = 0.02$, 그리고 $\sigma^2 = 0.3$을 사용한다. 또한 두 분포 모두 전수층 자료 수 $N = 100$을 사용한다.

• Step 2: 무응답 생성과정

(3) 전수층이므로 $N = n = 100$개의 표본에서 선형 응답률 모형인 $\pi_i = b_0 + b_1 y_i$, $\pi_i \in [0, 1]$을 이용하여 무응답을 생성한다. 즉 y_i의 최솟값에서의 응답률을 $\pi_y^{\min}$, y_i의 최댓값에서의 응답률을 $\pi_y^{\max}$라 할 때 $(\pi_y^{\min}, \pi_y^{\max}) = (0.9, 0.1), (0.5, 0.2), (0.2, 0.5), (0.1, 0.9)$를 사용하고 식 (2.5)인 선형 응답률 모형을 이용하여 b_0, b_1을 구한 후 구해진 값을 이용하여 y_i에 따라 응답률을 계산한다. 계산된 응답률에 따라 무응답을 생성한다. 같은 방법을 파워형 응답률 모형인 식 (2.6) 또는 $\pi_i = c_0 y_i^{c_1}$, $\pi_i \in [0, 1]$에 적용하여 무응답을 생성한다.

(4) 응답한 최종 조사 자료는 r개이다. 예를 들어 선형 응답률이고 오차가 감마분포를 따를 경우를 살펴보면 $(\pi_y^{\min}, \pi_y^{\max}) = (0.9, 0.1)$인 경우는 전체 자료의 약 75%가 조사되어 주어진 자료 수에 비해 약 25%가 감소하고 $(\pi_y^{\min}, \pi_y^{\max}) = (0.5, 0.2)$인 경우는 전체 자료의 약 45%가 조사되어 약 55%가 감소된다. 반면 $(\pi_y^{\min}, \pi_y^{\max}) = (0.2, 0.5)$ 또는 $(\pi_y^{\min}, \pi_y^{\max}) = (0.1, 0.9)$인 경우는 전체 자료의 약 25%가 조사되어 주어진 자료 수에 비해 약 75%가 감소한다.

• Step 3: 층화

얻어진 표본 자료는 (x_i, y_i), $i = 1, \ldots, r$이고 무응답에 의해 각 자료의 가중치는 달라진다. 이를 반영하기 위해 주어진 하나의 모집단 층을 L개의 세부 층으로 나눈다. 세부 층을 구성하는 방법은 모집단의 보조변수 x_i의 분위수를 이용한 방법인 M_1, 조사 결과 자료의 보조변수 x_i의 분위수를 이용한 방법인 M_2를 사용한다. 이후 추정된 기울기를 이용하여 방법 M을 사용한다.

(5) 제안된 두 가지 방법으로 전수층을 L개의 세부 층으로 나눈다. 세부 층의 개수가 많으면 각각의 세부 층에 포함된 자료 수가 적어 계산된 가중치가 불안정하며 반면 세부 층의 수가 적으면 응답률 모형의 모수 추정이 정확하지 않을 수 있다. 이에 본 연구에서는 약 30개에서 70개 정도의 최종 자료 수가 얻어지므로 $L = 5$를 사용하였으며 이에 관한 자세한 내용은 Min과 Shin (2018)을 참조하기 바란다.

• Step 4: 모수추정

(6) 나누어진 세부 층의 모집단 수와 조사된 자료 수 (N_h, r_h)를 이용하여 세부 층 보정 가중치 $w_h = N_h/r_h$를 계산한다. 이때 $w_i = w_{(i\in h)} = w_h$가 된다. 즉 세부 층에 포함된 자료의 가중치는 동일하다.

(7) 선형 응답률 모형은 식 (2.5)를 사용하고 파워형 응답률 모형은 식 (2.6)을 사용한 모형으로 단순 회귀분석을 이용하여 모수 b_0, b_1과 c_0, c_1을 추정한다.

(8) 추출된 자료 (y_i, x_i)와 초모집단모형인 식 (2.3)과 (2.4)를 이용하고 회귀분석을 실시하여 $\mu_i^{(s)}$와 σ^2을 추정한다. 여기서 $\mu_i^{(s)}$는 표본에서 얻어진 값을 의미한다. 또한 감마분포에서는 적률추정법으로 α를 추정한다.

(9) 계산된 결과를 이용하여 식 (2.1), (2.2), 그리고 식 (2.7)에서 식 (2.10)인 $\hat{\bar{Y}}_S, \hat{\bar{Y}}_{ST}, \hat{\bar{Y}}_{LG}, \hat{\bar{Y}}_{LL}, \hat{\bar{Y}}_{PG}$, $\hat{\bar{Y}}_{PL}$을 계산한다.

(10) 식 (2.5)와 (2.6)에서 추정된 b_1 또는 c_1값을 이용하여 M_1, M_2 그리고 추정된 기울기를 이용한 방법인 M을 사용하여 추정값을 계산한다.

이제 얻어진 평균 추정값은 다음의 비교통계량, 편향(bias), 절대편향(absolute bias; Abias) 그리고 root mean squared error (RMSE)를 이용하여 결과의 성능이 비교되었다. 각 통계량의 정의는 다음과 같다.

$$\text{Bias} = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\bar{Y}}_r - \bar{Y}_r\right),$$

$$\text{Abias} = \frac{1}{R}\sum_{r=1}^{R}\left|\hat{\bar{Y}}_r - \bar{Y}_r\right|,$$

$$\text{RMSE} = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\bar{Y}}_r - \bar{Y}_r\right)^2},$$

여기서 $R = 1{,}000$을 사용하였으며 각 반복마다 새로운 모집단을 생성하여 통계량을 계산하였다. 이는 생성된 특정 모집단의 영향을 줄이기 위함이며 이에 r번째 반복 모집단의 참값을 $\bar{Y}_r$로 표시하였다.

3.2. 모의실험 결과

응답률 모형 및 초모집단 오차 분포별 모의실험 결과가 Table 3.1에서 Table 3.4에 수록되었다. $\hat{\bar{Y}}_S, \hat{\bar{Y}}_{ST}$의 비교 결과에서는 예상대로 Jeon과 Shin (2019)에서 제안된 $\hat{\bar{Y}}_{ST}$가 모든 응답률 모형과 오차 분포 조합에서 우수한 결과를 주고 있다. 따라서 본 연구의 모의실험 결과는 $\hat{\bar{Y}}_{ST}$와 제안된 편향보정 방법의 비교 결과를 중점적으로 설명하였다. 먼저 선형 응답률 모형과 오차가 감마분포를 따르는 경우의 결과인 Table 3.1을 살펴보면 본 연구에서 제안한 편향보정 추정량인 $\hat{\bar{Y}}_{LG}$가 편향, 절대편향, 그리고 RMSE 기준에서 매우 우수한 결과를 준다. 예를 들면 $(\pi_y^{\min}, \pi_y^{\max}) = (0.1, 0.9)$인 경우 매우 큰 양의 편향이 발생하나 본 연구에서 제안한 방법을 사용함으로써 편향의 영향을 크게 줄일 수 있음을 확인하였다. 다만 응답률의 기울기가 음수인 $(\pi_y^{\min}, \pi_y^{\max}) = (0.9, 0.1), (0.5, 0.2)$인 경우 M_1의 결과가 우수하고, 응답률의 기울기가 양수인 $(\pi_y^{\min}, \pi_y^{\max}) = (0.2, 0.5), (0.1, 0.9)$인 경우 M_2의 결과가 우수하기 때문에 기울기에 따라 세부 층 구성 방법을 선택하여 사용할 필요가 있다. 다만 실제자료 분석에서는 응답률의 기울기가 알려져 있지 않기 때문에 먼저 응답률 모형을 적합한 후 이때 추정된 기울기에 따라 세부 층 구성 방법을 결정하는 방법인 M을 선택하는 것이 타당하다. 모의실험 결과를 살펴보면 추정된 기

Table 3.1. Results of linear response model with gamma distribution

$\pi_y^{\min}$	$\pi_y^{\max}$	r	Method	Bias			Abias			RMSE		
				$\hat{\bar{Y}}_S$	$\hat{\bar{Y}}_{ST}$	$\hat{\bar{Y}}_{LG}$	$\hat{\bar{Y}}_S$	$\hat{\bar{Y}}_{ST}$	$\hat{\bar{Y}}_{LG}$	$\hat{\bar{Y}}_S$	$\hat{\bar{Y}}_{ST}$	$\hat{\bar{Y}}_{LG}$
0.9	0.1	76	M_1	−353.6	−214.7	−167.7	353.7	215.0	173.4	418.3	278.2	237.2
			M_2	−353.6	−243.8	−202.0	353.7	244.1	204.5	418.3	311.7	273.4
			M	−353.6	−213.4	−165.7	353.7	213.8	172.0	418.3	277.5	236.2
0.5	0.2	45	M_1	−257.0	−149.7	−115.1	265.9	165.6	152.9	343.3	232.0	216.4
			M_2	−257.0	−172.1	−138.7	265.9	183.0	164.7	343.3	256.8	233.7
			M	−257.0	−145.4	−107.5	265.9	162.7	150.7	343.3	229.3	213.3
0.2	0.5	26	M_1	485.1	485.1	122.3	500.3	166.3	155.7	664.5	234.1	222.0
			M_2	485.1	91.0	63.2	500.3	120.9	111.3	664.5	169.3	161.7
			M	485.1	96.3	69.9	500.3	124.0	115.4	664.5	173.8	168.0
0.1	0.9	24	M_1	1086.2	199.3	153.5	1086.3	204.8	169.5	1226.1	244.2	210.8
			M_2	1086.2	134.3	70.8	1086.3	138.6	95.1	1226.1	166.3	122.3
			M	1086.2	135.2	71.9	1086.3	139.4	96.1	1226.1	167.4	124.1

Abias = absolute bias; RMSE = root mean squared error.

Table 3.2. Results of linear response model with log-normal distribution

$\pi_y^{\min}$	$\pi_y^{\max}$	r	Method	Bias			Abias			RMSE		
				$\hat{\bar{Y}}_S$	$\hat{\bar{Y}}_{ST}$	$\hat{\bar{Y}}_{LL}$	$\hat{\bar{Y}}_S$	$\hat{\bar{Y}}_{ST}$	$\hat{\bar{Y}}_{LL}$	$\hat{\bar{Y}}_S$	$\hat{\bar{Y}}_{ST}$	$\hat{\bar{Y}}_{LL}$
0.9	0.1	72	M_1	−458.9	−352.0	−289.9	458.9	352.2	291.9	530.8	420.0	358.7
			M_2	−458.9	−371.3	−311.2	458.9	371.4	312.0	530.8	439.3	377.3
			M	−458.9	−351.2	−288.6	458.9	351.4	290.6	530.8	419.5	357.8
0.5	0.2	43	M_1	−355.4	−256.5	−200.4	361.5	266.9	224.8	438.5	336.5	290.2
			M_2	−355.4	−274.7	−220.7	361.5	282.6	237.3	438.5	352.3	302.3
			M	−355.4	−253.5	−195.3	361.5	264.7	221.1	438.5	333.9	285.6
0.2	0.5	27	M_1	633.7	232.4	158.1	642.0	247.6	187.9	804.0	311.4	248.9
			M_2	633.7	187.8	105.7	642.0	205.0	152.3	804.0	257.8	207.0
			M	633.7	190.4	110.2	642.0	207.1	155.5	804.0	260.5	212.3
0.1	0.9	28	M_1	1136.5	337.6	218.4	1136.5	339.0	225.0	1283.5	378.2	267.5
			M_2	1136.5	302.3	175.1	1136.5	302.8	182.0	1283.5	330.2	217.1
			M	1136.5	302.8	175.9	1136.5	303.2	182.6	1283.5	330.6	218.3

Abias = absolute bias; RMSE = root mean squared error.

울기를 사용한 방법인 M을 사용하면 응답률의 형태 및 기울기에 무관하게 우수한 추정 결과를 얻을 수 있음을 확인할 수 있다. 또한 Table 3.2 결과는 선형 응답률과 오차가 로그-정규분포를 따르는 경우로 Table 3.1과 유사한 결과를 확인할 수 있다. 특히 방법 M을 사용한 경우의 결과도 매우 우수한 것을 확인할 수 있다.

다음으로 파워형 응답률과 오차가 감마분포를 따르는 경우인 Table 3.3 결과는 Table 3.1 결과와 유사한 것을 확인할 수 있다. 즉 본 연구에서 제안한 $\hat{\bar{Y}}_{PG}$가 $\hat{\bar{Y}}_{ST}$에 비해 모든 비교 통계량을 기준으로 우수한 결과를 준다. 다만 감마분포에 비해 로그-정규분포 결과는 상대적으로 편향보정결과가 작은 것을 확인할 수 있다. 특히 $(\pi_y^{\min}, \pi_y^{\max}) = (0.2, 0.5), (0.5, 0.2)$에서 RMSE 결과를 비교하면 $\hat{\bar{Y}}_{ST}$와 큰 차이를 보이고 있지 않다. 그러나 편향을 살펴보면 $\hat{\bar{Y}}_{ST}$에 비해 매우 우수한 결과를 주고 있어 본 연구에서 제안한 추정량이 편향을 잘 보정한다는 것을 확인할 수 있다. 또한 응답률의 기울기가 음수인 경우에는 M_1 방법을 반대로 응답률의 기울기가 양수인 경우에는 M_2 방법을 사용하는 것이 효과적이며 기울기가

Table 3.3. Results of power response model with gamma distribution

$\pi_y^{\min}$	$\pi_y^{\max}$	r	Method	Bias			Abias			RMSE		
				$\hat{\bar{Y}}_S$	$\hat{\bar{Y}}_{ST}$	$\hat{\bar{Y}}_{PG}$	$\hat{\bar{Y}}_S$	$\hat{\bar{Y}}_{ST}$	$\hat{\bar{Y}}_{PG}$	$\hat{\bar{Y}}_S$	$\hat{\bar{Y}}_{ST}$	$\hat{\bar{Y}}_{PG}$
0.9	0.1	42	M_1	−490.2	−205.8	−187.8	490.2	243.5	232.4	541.6	327.4	318.7
			M_2	−490.2	−280.0	−258.0	490.2	289.5	271.6	541.6	360.7	344.2
			M	−490.2	−197.0	−178.0	490.2	235.2	223.8	541.6	311.2	301.7
0.5	0.2	35	M_1	−252.7	−87.1	−79.5	268.7	159.9	157.6	338.7	225.6	224.0
			M_2	−252.7	−122.2	−113.3	268.7	170.5	166.6	338.7	235.3	231.1
			M	−252.7	−82.7	−74.5	268.7	159.4	157.2	338.7	223.7	222.2
0.2	0.5	30	M_1	355.5	86.9	79.9	370.6	119.8	116.4	469.3	165.6	162.2
			M_2	355.5	64.6	55.1	370.6	103.0	99.6	469.3	148.1	145.1
			M	355.5	65.5	56.2	370.6	102.8	99.4	469.3	142.4	139.2
0.1	0.9	24	M_1	995.0	187.5	172.0	995.2	196.0	184.1	1144.4	243.3	231.7
			M_2	995.0	133.9	112.1	995.2	139.9	122.4	1144.4	172.4	155.5
			M	995.0	134.6	112.8	995.2	140.5	123.0	1144.4	173.3	156.6

Abias = absolute bias; RMSE = root mean squared error.

Table 3.4. Results of power response model with log-normal distribution

$\pi_y^{\min}$	$\pi_y^{\max}$	r	Method	Bias			Abias			RMSE		
				$\hat{\bar{Y}}_S$	$\hat{\bar{Y}}_{ST}$	$\hat{\bar{Y}}_{PL}$	$\hat{\bar{Y}}_S$	$\hat{\bar{Y}}_{ST}$	$\hat{\bar{Y}}_{PL}$	$\hat{\bar{Y}}_S$	$\hat{\bar{Y}}_{ST}$	$\hat{\bar{Y}}_{PL}$
0.9	0.1	38	M_1	−580.6	−346.2	−317.4	580.6	376.6	356.0	642.9	457.5	439.0
			M_2	−580.6	−396.5	−362.7	580.6	405.3	376.1	642.9	471.6	442.2
			M	−580.6	−336.6	−305.9	580.6	367.5	345.9	642.9	443.8	423.5
0.5	0.2	34	M_1	−299.4	−149.7	−133.5	314.9	214.1	208.1	383.1	281.9	276.9
			M_2	−299.4	−175.6	−156.8	314.9	218.8	208.8	383.1	280.6	269.6
			M	−299.4	−143.2	−125.1	314.9	210.0	203.1	383.1	274.3	268.3
0.2	0.5	32	M_1	406.4	154.0	131.8	416.4	176.1	161.1	526.7	227.6	212.6
			M_2	406.4	136.5	109.7	416.4	161.7	145.1	526.7	207.6	192.1
			M	406.4	138.1	111.5	416.4	162.8	146.5	526.7	208.7	193.3
0.1	0.9	33	M_1	967.7	312.6	248.4	967.7	312.9	250.0	1089.6	345.1	283.6
			M_2	967.7	273.5	198.0	967.7	273.8	200.9	1089.6	297.6	230.8
			M	967.7	274.0	198.5	967.7	274.3	201.4	1089.6	298.0	231.4

Abias = absolute bias; RMSE = root mean squared error.

알려져 있지 않는 경우에는 방법 M을 사용하는 것이 타당하다. 이러한 결과는 파워형 응답률과 오차가 로그-정규분포를 따르는 경우인 Table 3.4에서도 확인되었다.

4. 실제 자료 분석

4.1. 자료 설명

본 연구에서는 2018년 기준 문화체육관광산업 사업체 자료 중 종사자 수 400인 이상 자료로 삼성물산(주)에버랜드와 같이 매출액이 10조 이상인 자료를 제외한 129개의 자료가 실제 자료 분석에 사용되었다. 한국문화관광연구원에서는 문화체육관광산업과 관련된 통계를 작성하고 있으나 본 연구의 실제 자료 분석에 필요한 전수층 사업체 자료를 얻을 수 없기 때문에 본 분석에서는 각 사업체의 홈페이지에 실린 2018년 공시자료를 사용하였으며 관심변수와 보조변수로 각 사업체의 종사자 수와 매출액을 사용

Table 4.1. Normality test results

Test	Test statistics	p-value
Shapiro-Wilk	$W = 0.9782$	0.0357
Kolmogorov-Smirnov	$D = 0.0586$	> 0.1500
Cramer-von Mises	W-sq $= 0.0749$	0.2426
Anderson-Darling	A-sq $= 0.5257$	0.1854

Table 4.2. Data of linear response model with log-normal distribution ($\times 10^8$)

$\pi_y^{\min}$	$\pi_y^{\max}$	r	Method	Bias			Abias			RMSE		
				$\hat{\bar{Y}}_S$	$\hat{\bar{Y}}_{ST}$	$\hat{\bar{Y}}_{LL}$	$\hat{\bar{Y}}_S$	$\hat{\bar{Y}}_{ST}$	$\hat{\bar{Y}}_{LL}$	$\hat{\bar{Y}}_S$	$\hat{\bar{Y}}_{ST}$	$\hat{\bar{Y}}_{LL}$
0.9	0.1	95	M_1	−3.618	−3.312	−2.792	3.618	3.312	2.792	3.627	3.323	2.817
			M_2	−3.618	−3.299	−2.771	3.618	3.299	2.771	3.627	3.310	2.793
			M	−3.618	−3.309	−2.789	3.618	3.309	2.789	3.627	3.320	2.815
0.5	0.2	56	M_1	−3.089	−2.774	−2.277	3.089	2.777	2.295	3.146	2.851	2.452
			M_2	−3.089	−2.753	−2.217	3.089	2.757	2.236	3.146	2.829	2.387
			M	−3.089	−2.747	−2.218	3.089	2.750	2.239	3.146	2.825	2.398
0.2	0.5	35	M_1	5.686	3.027	1.678	5.687	3.036	1.769	5.982	3.302	2.352
			M_2	5.686	2.750	1.295	5.687	2.762	1.497	5.982	3.062	2.192
			M	5.686	2.763	1.355	5.687	2.776	1.556	5.982	3.073	2.278
0.1	0.9	34	M_1	10.152	4.706	2.416	10.152	4.706	2.416	10.282	4.829	2.592
			M_2	10.152	4.238	1.746	10.152	4.238	1.752	10.282	4.367	1.934
			M	10.152	4.238	1.746	10.152	4.238	1.752	10.282	4.367	1.934

Abias = absolute bias; RMSE = root mean squared error.

하였다. 따라서 본 자료는 기존에 공표되고 있는 문화체육관광산업 사업체의 전체 모집단을 대표하지 못한다는 단점이 있으나 본 분석의 목적인 편향보정 추정량의 우수성을 확인하기에는 충분하다고 판단된다. 사용된 모집단 129개 자료의 평균 종사자 수는 1250.46명이고, 평균 매출액은 729,087백만 원이다.

초모집단 모형 분석을 위해 로그 변환 후의 회귀분석 결과를 살펴보면 각 모수 추정값으로 $\hat{\beta}_0 = 11.8578$, $\hat{\beta}_1 = 0.000574$, $\hat{\sigma} = 1.1463$, $R^2 = 0.3148$가 얻어졌으며 잔차의 정규성 검정 결과는 Table 4.1에서 확인할 수 있다. Table 4.1 결과를 살펴보면 잔차 자료의 정확한 분포는 알 수 없으나 정규분포와 유사한 분포라고 판단하는 것은 큰 무리가 없다고 판단된다. 이에 따라 본 연구에서 얻어진 결과 중에서 오차가 로그-정규분포를 따르는 경우의 추정량을 사용하였다. 주어진 모집단 자료는 전수층 자료이므로 $N = n = 129$가 되며 또한 $L = 5$를 사용하였다. 실제조사에서 전수층의 경우 조사 결과에 따라 다르지만 약 50% 정도 응답률을 보이는 경우가 많이 있다. 이에 본 논문에서는 모의실험에서 사용한 것처럼 응답률이 $30 \sim 50\%$인 경우를 살펴보았으며 모의실험에서 사용한 선형과 파워형 응답률 모형을 이용하여 무응답을 생성하였다. 따라서 초모집단 모형은 식 (2.4)를 사용하였고 응답률 모형은 식 (2.5)와 (2.6)을 사용하였으며 사용된 편향보정 추정량은 식 (2.8)과 (2.10)인 $\hat{\bar{Y}}_{LL}$과 $\hat{\bar{Y}}_{PL}$을 사용하였다.

자료분석 결과는 Table 4.2와 Table 4.3에 수록되었다. 결과를 살펴보면 모의실험 결과와 유사하게 본 연구에서 제안한 편향보정 추정량이 모든 비교통계량 기준으로 매우 우수한 결과를 주고 있다. 특히 편향결과를 살펴보면 편향의 크기가 크게 줄어든 것을 확인할 수 있다. 다만 응답률 모형의 기울기가 음수 또는 양수 두 경우 모두 M_2가 우수한 결과를 보이고 있으나 방법 M을 사용한 경우의 결과도 매우 우수한 결과를 주고 있음을 확인할 수 있다.

Table 4.3. Data of power response model with log-normal distribution ($\times 10^8$)

$\pi_y^{\min}$	$\pi_y^{\max}$	r	Method	Bias			Abias			RMSE		
				$\hat{\bar{Y}}_S$	$\hat{\bar{Y}}_{ST}$	$\hat{\bar{Y}}_{PL}$	$\hat{\bar{Y}}_S$	$\hat{\bar{Y}}_{ST}$	$\hat{\bar{Y}}_{PL}$	$\hat{\bar{Y}}_S$	$\hat{\bar{Y}}_{ST}$	$\hat{\bar{Y}}_{PL}$
0.9	0.1	95	M_1	−3.695	−3.020	−2.910	3.701	3.086	3.002	3.873	3.385	3.317
			M_2	−3.695	−2.959	−2.827	3.701	3.011	2.908	3.873	3.302	3.215
			M	−3.695	−2.945	−2.822	3.701	3.025	2.933	3.873	3.323	3.246
0.5	0.2	56	M_1	−1.871	−1.307	−1.210	2.028	1.717	1.700	2.335	2.052	2.038
			M_2	−1.871	−1.285	−1.171	2.028	1.670	1.647	2.335	1.993	1.973
			M	−1.871	−1.233	−1.120	2.028	1.678	1.659	2.335	2.009	1.995
0.2	0.5	35	M_1	2.324	1.354	1.163	2.427	1.496	1.355	2.807	1.799	1.676
			M_2	2.324	1.325	1.120	2.427	1.462	1.314	2.807	1.752	1.636
			M	2.324	1.338	1.136	2.427	1.472	1.327	2.807	1.770	1.661
0.1	0.9	34	M_1	5.680	2.945	2.287	5.680	2.945	2.287	5.748	3.028	2.421
			M_2	5.680	2.571	1.760	5.680	2.571	1.763	5.748	2.659	1.939
			M	5.680	2.572	1.762	5.680	2.572	1.765	5.748	2.662	1.944

Abias = absolute bias; RMSE = root mean squared error.

5. 결론

본 연구에서는 전수층에서 발생한 무응답을 적절히 처리하는 방법을 연구하였다. 최근에는 단위무응답이 MAR과 같이 랜덤으로 발생하지 않고 관심변수에 영향을 받는 경우가 다수 발생하고 있으며 특히 예비표본이 없는 전수층에서는 단위무응답의 적절한 처리가 매우 중요하다. 이러한 경우에 MAR 가정 하에서 사용하는 가중치보정방법은 편향을 발생시키기 때문에 편향을 제거하여 추정의 정확성을 향상시키는 방법인 편향보정 방법을 사용하는 것이 타당하기 때문에 본 연구에서도 이미 개발된 편향보정 추정량을 적용하였다.

특히 본 논문의 연구대상인 전수층은 표본층과 달리 모집단 수 및 표본 수가 적어 기존의 세부 층 구성방법을 적용할 경우 효율이 떨어질 수 있어 이를 해결하기 위한 새로운 세부 층 구성방법을 제안하였다. 모의실험 결과 본 논문에서 제안한 편향보정 방법을 사용함으로써 무응답으로 인해 발생된 편향을 크게 축소할 수 있었으며 절대편향 및 RMSE도 모두 줄일 수 있음을 확인하였다. 실제 자료 분석 결과도 본 연구에서 제안한 방법이 매우 효과적임을 확인하였다. 특히 실제 자료가 정확한 정규분포가 아님에도 불구하고 제안된 방법을 사용하게 되면 편향이 제거된 우수한 추정 결과를 얻을 수 있었다. 결론적으로 본 연구에서 제안한 편향보정 방법을 무응답이 발생한 전수층에 적용한다면 정확한 추정 결과가 얻어질 것으로 판단된다.

References

Chung, H. Y. and Shin, K.-I. (2017). Estimation using informative sampling technique when response rate follows exponential function of variable of interest, *The Korean Journal of Applied Statistics*, **30**, 933–1004.

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

Hidiroglou, M. A. (1986). The construction of a self-representing stratum of large units in survey design, *The American Statistician*, **4**, 27–31.

Jeon, S. S. and Shin, K.-I. (2019). A study on weight adjustment method for non-response in take-all stratum, *Journal of The Korean Official Statistics*, **24**, 28–47.

Lee, S. E. and Shin, K.-I. (2016). The cut-off point based on underlying distribution and cost function, *Journal of Applied Statistics*, **43**, 1061–1073.

Lavallee, P. and Hidiroglou, M. (1988). On the stratification of skewed populations, *Survey Methodology*, **14**, 33–43.

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

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

Pfeffermann, D., Moura, F. A. D. S., and Silva, P. L. D. N. (2006). Multi-level modelling under informative sampling, *Biometrika*, **93**, 943–959.

전수층 무응답 편향보정 추정법에 관한 연구

정희영[a] · 신기일[a,1]

[a]한국외국어대학교 통계학과

(2020년 5월 8일 접수, 2020년 6월 30일 수정, 2020년 7월 1일 채택)

요 약

사업체조사에서는 흔히 수정절사법이 사용되며 이 방법을 사용함으로써 표본의 수를 줄이면서도 추정의 정확성을 향상 시킬 수 있다. 그러나 전수층의 무응답률은 크게 높아지고 있으며 예비표본을 이용한 표본대체가 불가능하기 때문에 전수층에서 발생한 무응답은 추정의 정확성을 크게 떨어뜨리고 있다. 특히 무응답이 관심변수에 영향을 받는 경우에는 편향이 발생할 가능성이 매우 높기 때문에 이를 적절히 처리하는 것은 매우 중요하다. 본 연구에서는 전수층에서 발생한 무응답을 적절히 처리하는 방법의 하나로 편향보정 추정법을 제안하였다. 특히 Chung과 Shin (2020)에서 제안한 편향보정 추정량을 전수층 편향보정에 적용하였으며 전수층이라는 특수한 경우에 맞는 새로운 추정 방법을 제안하였다. 또한 모의실험을 통해 제안된 방법의 우수성을 살펴보았으며 실제 자료 분석을 실시하여 본 논문에서 제안한 방법의 우수성을 확인하였다.

주요용어: 초모집단 모형, 선형 응답률 모형, 파워형 응답률 모형, 감마분포, 로그-정규분포

이 논문은 2018년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NRF-2018R1D1A1B07042736).

[1]교신저자: (17035) 경기도 용인시 처인구 모현면 외대로 81, 한국외국어대학교 통계학과.
E-mail: keyshin@hufs.ac.kr