The Korean Journal of Applied Statistics (2020) DOI: https://doi.org/10.5351/KJAS.2020.33.1.011
33(1), 11–23

A study on non-response bias adjusted estimation in business survey

Hee Young Chung[a] · Key-Il Shin[a,1]

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

(Received October 22, 2019; Revised December 2, 2019; Accepted December 10, 2019)

Abstract

Sampling design should provide statistics to meet a given accuracy while saving cost and time. However, a large number of non-responses are occurring due to the deterioration of survey circumstances, which significantly reduces the accuracy of the survey results. Non-responses occur for a variety of reasons. Chung and Shin (2017, 2019) and Min and Shin (2018) found that the accuracy of estimation is improved by removing the bias caused by non-response when the response rate is an exponential or linear function of variable of interests. For that case they assumed that the error of the super population model follows normal distribution. In this study, we proposed a non-response bias adjusted estimator in the case where the error of a super population model follows the gamma distribution or the log-normal distribution in a business survey. We confirmed the superiority of the proposed estimator through simulation studies.

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

1. 서론

표본조사에서 발생한 무응답은 최종 조사 자료 수의 감소로 인해 분산을 확대시키기고 또한 편향을 발생시킬 수도 있다. 이러한 문제를 적절히 처리하기 위한 다양한 연구가 진행되었고 실무에도 적용되고 있다. 최근 무응답이 관심변수 자료 값과 관련이 있는 경우가 다수 발생하고 있으며 이러한 관련성은 응답률 함수로 표현될 수 있다. 또한 많은 표본조사에서는 관심변수와 보조변수가 관계를 맺고 있으며 이러한 관계는 초모집단모형으로 표현될 수 있다. 흔히 자료가 표본에 포함될 확률이 관심변수와 관계가 있고 초모집단모형(super population model)을 갖는 자료의 표본설계를 정보적 표본설계라 한다. 정보적 표본설계와 관련된 다수의 논문이 발표되었으며 Pfefferman 등 (1998), Pfefferman 등 (2006) 그리고 Kim과 Skinner (2013)을 살펴보기 바란다. 이러한 정보적 표본설계(informative sampling)의 표본 포함확률 개념은 무응답이 발생한 경우에서의 응답률에 적용될 수 있다. 따라서 관심변수의 함수인 응답률 모형과 초모집단모형이 구성된 표본조사에서는 적절한 무응답 처리를 위해 정보적 표본설계 기법에서 얻어진 결과를 사용할 수 있다.

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: Department of Statistics, Hankuk University of Foreign Studies, 81 Oedae-ro, Yongin-si, Gyeonggi-do 17035, Korea. E-mail: keyshin@hufs.ac.kr

Chung과 Shin (2017)은 응답률 모형이 지수형이고 초모집단모형이 단순회귀모형을 따를 경우를 연구하였다. 이때 초모집단모형에 포함된 오차 분포는 정규분포를 가정하였다. Min과 Shin (2018)은 응답률 모형에 포함된 모수의 최적 추정에 필요한 최적 세부층 수 및 최적 세부층 표본 수를 제안하였다. 또한 Chung과 Shin (2019)은 응답률 모형이 선형이고 초모집단모형의 오차가 정규분포인 경우의 편향보정 추정량을 제안하였다.

반면 사업체조사에서는 초모집단모형의 오차가 비대칭 분포를 따를 가능성이 높고 실제 자료 분석에서는 흔히 감마분포 또는 로그-정규분포를 가정하여 분석한다. 이에 본 연구에서는 사업체조사에서 오차가 감마분포와 로그-정규분포인 경우에서 편향의 크기를 추정하고 편향보정 추정량을 제안하였다. 이때 감마 분포와 로그-정규분포에 적절히 적용될 수 있는 응답률 모형으로 선형 모형과 파워형 모형을 사용하였다.

본 논문의 구성은 다음과 같다. 먼저 2절에서는 기존에 얻어진 편향보정 추정량을 설명하였다. 3절에서는 본 연구의 핵심 내용으로 응답률 모형이 선형 또는 파워형이며 초모집단모형에서 오차의 분포가 감마분포와 로그-정규분포인 경우에서 편향의 크기를 추정하고 이 결과를 반영한 편향보정 추정량을 제안하였다. 4절에는 모의실험을 통하여 기존에 사용되는 추정량과 본 연구에서 제안한 추정량의 성능을 비교하였다. 5절에 결론이 있다.

2. 정보적 표본설계 기법을 이용한 응답률 편향 추정

2.1. 지수형 및 선형 응답률의 편향 추정

정보적 표본설계는 표본 추출과정이 관심변수 자료 값에 영향을 받고 관심변수와 보조변수 간의 모형인 초모집단모형이 존재하는 표본설계이다. Pfeffermann 등 (1998)은 정보적 표본설계 하에서 θ^*가 θ의 함수라 할 때 $f_s(y_i|\theta^*, x_i) = f_s(y_i|i \in s, x_i) = \Pr(i \in s|y_i, x_i) f_p(y_i|\theta, x_i)/\Pr(i \in s|x_i)$이고 $\Pr(i \in s|y_i, x_i) = E_p(\pi_i|y_i, x_i)$이 되며 또한 $\Pr(i \in s|x_i) = E_p(\pi_i|x_i)$가 되어 다음의 관계가 성립되는 것을 밝혔다.

$$f_s(y_i|x_i) = \frac{E_p(\pi_i|y_i, x_i) f_p(y_i|x_i)}{E_p(\pi_i|x_i)}, \tag{2.1}$$

여기서 $f_p(y_i|x_i)$는 모집단 분포, $f_s(y_i|x_i)$는 표본 분포 그리고 $E_p(\pi_i|y_i, x_i)$는 x_i, y_i가 주어졌을 때 i번째 자료가 표본에 포함될 포함확률이다.

Chung과 Shin (2017)은 표본 포함확률을 응답률에 적용하여 무응답에 의해 발생된 편향을 축소하는 편향보정 추정량을 제안하였다. 이 내용을 간단히 살펴보면 먼저 초모집단모형과 지수형 응답률 모형인 식 (2.2)와 (2.3)을 사용하였다.

$$f_p(y_i|x_i) = N\left(\beta_0 + \beta_1 x_i, \sigma^2\right), \tag{2.2}$$

$$E_p(\pi_i|y_i, x_i) = \exp(a_0 + a_1 y_i). \tag{2.3}$$

이제 식 (2.2)와 (2.3)을 식 (2.1)에 대입하면 $f_s(y_i|x_i) = N(\beta_0 + a_1\sigma^2 + \beta_1 x_i, \sigma^2)$인 표본 분포가 얻어지며 이를 통하여 편향의 크기가 $a_1\sigma^2$임을 확인하였다. 이후 Chung과 Shin (2019)은 초모집단모형으로 식 (2.2)를 사용하였으며 선형 응답률 모형인 식 (2.4)를 고려하였다.

$$E_p(\pi_i|y_i, x_i) = b_0 + b_1 y_i, \tag{2.4}$$

여기서 $E_p(\pi_i|x_i) = E(E_p(\pi_i|y_i, x_i)) = E(b_0 + b_1 y_i|x_i) = b_0 + b_1 E_p(y_i|x_i)$이 되므로 $E_p(y_i|x_i) = \mu_i$로

표시하면 표본 분포는 다음과 같이 얻어진다.

$$f_s(y_i|x_i) = \frac{b_0}{b_0 + b_1\mu_i} f_p(y_i|x_i) + \frac{b_1\mu_i}{b_0 + b_1\mu_i} f_p^*(y_i|x_i). \tag{2.5}$$

식 (2.5)의 $f_p^*(y_i|x_i)$는 $y_i f_p(y_i|x_i)$의 분포를 의미하며 따라서 표본 분포는 $f_p(y_i|x_i)$와 $f_p^*(y_i|x_i)$의 선형 결합 형태가 된다. 이제 초모집단모형의 오차가 정규 분포인 식 (2.2)를 따른다고 가정하면 식 (2.5)에 의해 표본 분포가 얻어지고 이를 통해 모집단 평균과 표본 평균과의 관계가 얻어진다. 즉 다음의 결과가 얻어진다.

$$\mu_i = \mu_i^{(s)} - \frac{b_1\sigma^2}{b_0 + b_1\mu_i},$$

여기서 μ_i는 모집단 평균이고 $\mu_i^{(s)}$는 표본 평균이다. 따라서 모집단 평균과 표본 평균의 차이인 편향 $b_1\sigma^2/(b_0 + b_1\mu_i)$가 얻어진다. 자세한 내용은 Chung과 Shin (2019)을 살펴보기 바란다.

2.2. 응답률 모수 추정

지수형 응답률 모형에서 계산된 편향의 크기는 $a_1\sigma^2$이다. 여기에서 σ^2의 경우에는 표본자료의 회귀모형을 이용하여 어렵지 않게 추정값을 얻을 수 있다. 반면에 a_1은 주어진 지수형 응답률 모형을 사용하여 추정한다. Min과 Shin (2018)은 주어진 층을 세부층으로 나누는 방법을 이용하여 a_1을 추정하였다. 세부층은 모집단에 포함되어진 보조변수의 분위수를 이용하여 구성되며 각 세부층에 포함된 모집단 수와 최종 자료 수를 이용하여 세부층의 가중치를 구한다. 또한 Pfeffermann과 Sverchkov (2003)에서 얻어진 결과인 $E_s(w_i|y_i, x_i) = 1/E_p(\pi_i|y_i, x_i)$와 $E_s(w_i|y_i, x_i) \approx w_i$를 적용하고 나누어진 세부층에 의해 구해진 w_i를 사용하면 지수형 응답률 모형의 모수 추정을 위한 회귀모형이 다음 식 (2.6)과 같이 얻어진다.

$$\log\left(\frac{1}{w_i}\right) = a_0 + a_1 y_i + \eta_i. \tag{2.6}$$

따라서 응답률 모형의 모수 a_1은 식 (2.6)을 이용하여 추정된다. 또한 유사한 방법을 사용하여 Chung과 Shin (2019)은 선형 응답률 모형의 모수 추정을 위해 다음의 모형을 사용하였다.

$$\frac{1}{w_i} = b_0 + b_1 y_i + \eta_i. \tag{2.7}$$

따라서 선형 응답률 모형에서도 세부층에서 얻어진 세부층 가중치 w_i와 자료 y_i를 이용하여 b_0, b_1을 추정하게 된다. 이때 η_i는 독립이고, 같은 분포를 따르며 등분산성이 만족된다고 가정하였다.

본 연구에서는 선형 응답률 모형과 파워형 응답률 모형을 이용하기 때문에 선형 응답률 모형인 (2.7)에서 얻어진 모수 추정 결과를 편향 추정에 사용한다.

3. 응답률 모형에 따른 편향 추정 및 편향보정 추정량

이 절에서는 사업체조사에서 흔히 발생할 수 있는 초모집단모형인 로그-선형 모형을 살펴보았다. 로그-선형 모형은 로그변환을 했을 때 선형모형이 되기 때문에 이를 반영하기 위해 모형의 오차는 감마분포와 로그-정규 분포를 따른다고 가정하였다. 또한 본 연구에서 사용한 응답률 모형은 선형 응답률 모형과 파워형 응답률 모형이다. 이 절에서는 주어진 가정 하에서 각 응답률 모형과 분포에 따른 편향을 구하였으며 이를 이용하여 편향보정 추정량을 제안하였다.

3.1. 선형 응답률 모형에서의 편향 추정

3.1.1. 오차가 감마분포를 따를 때 Pfeffermann 등 (1998)의 내용처럼 먼저 모집단 분포 $f_p(y_i|x_i)$가 감마 분포, Gamma(α, β_i)를 따른다고 가정하고 $\mu_i = E_p(y_i|x_i) = \alpha\beta_i$라 하자. 그러면 모집단 분포는 다음과 같이 정의된다.

$$f_p(y_i|\alpha, \mu_i) \propto y_i^{\alpha-1} \exp\left(-\alpha \frac{y_i}{\mu_i}\right). \tag{3.1}$$

이제 관심변수 y_i와 보조변수 x_i의 관계는 로그변환 후 선형모형을 가정할 수 있으므로 초모집단모형은 다음의 로그-선형 모형을 사용한다.

$$\log(\mu_i) = \beta_0 + \beta_1 x_i. \tag{3.2}$$

물론 식 (3.1)과 (3.2)는 정보적 표본설계에서 흔히 사용하는 가정이다. 이제 선형 응답률 모형인 식 (2.4)의 결과에서 $E_p(\pi_i|x_i) = b_0 + b_1 E_p(y_i|x_i)$이 얻어지고 또한 식 (3.1)에 의해 $y_i f_p(y_i|x_i) \propto y_i y_i^{\alpha-1} \exp(-y_i/(\mu_i/\alpha))$이 되어 $y_i f_p(y_i|x_i)$의 분포인 $f_p^*(y_i|x_i)$은 다음과 같이 얻어진다.

$$f_p^*(y_i|x_i) = \text{Gamma}\left(\alpha + 1, \frac{\mu_i}{\alpha}\right).$$

따라서 식 (3.1)과 $f_p^*(y_i|x_i)$ 결과를 이용하면 식 (2.5)에 의한 표본 분포는 다음과 같이 얻어진다.

$$f_s^*(y_i|x_i) = \frac{b_0}{b_0 + b_1\mu_i}\Gamma\left(\alpha, \frac{\mu_i}{\alpha}\right) + \frac{b_1\mu_i}{b_0 + b_1\mu_i}\Gamma\left(\alpha + 1, \frac{\mu_i}{\alpha}\right).$$

이 결과를 이용하여 표본 분포의 표본 평균을 구하면

$$\begin{aligned}\mu_i^{(s)} &= E_s(y_i|x_i) \\ &= \frac{b_0}{b_0 + b_1\mu_i}\mu_i + \frac{b_1\mu_i}{b_0 + b_1\mu_i}\left(\frac{\alpha+1}{\alpha}\right)\mu_i\end{aligned}$$

이 되므로 최종적으로

$$\mu_i^{(s)} = \mu_i + \frac{b_1}{b_0 + b_1\mu_i}\frac{\mu_i^2}{\alpha} \tag{3.3}$$

이 얻어진다. 또한 식 (3.3)은 다음과 같이 표현될 수 있다.

$$\mu_i = \mu_i^{(s)} / \left(1 + \frac{b_1}{b_0 + b_1\mu_i} \times \frac{\mu_i}{\alpha}\right).$$

이제 식 (3.3)에서 계산을 용이하게 하기 위하여 $\{b_1/(b_0+b_1\mu_i)\}(\mu_i^2/\alpha) \approx \{b_1/(b_0+b_1\mu_i^{(s)})\}(\mu_i^{(s)2}/\alpha)$을 사용하게 되면 최종적으로 다음의 결과를 얻는다.

$$\mu_i = \mu_i^{(s)} - \frac{b_1}{b_0 + b_1\mu_i^{(s)}}\frac{\mu_i^{(s)2}}{\alpha}, \tag{3.4}$$

여기서 $\mu_i^{(s)}$는 i번째 표본 자료의 기댓값이며 따라서 추정된 편향은 근사적으로 다음과 같이 얻어진다.

$$\frac{b_1}{b_0 + b_1\mu_i^{(s)}}\frac{\mu_i^{(s)2}}{\alpha}.$$

3.1.2. 오차가 로그-정규분포를 따를 때 이 절에서는 모집단 분포 $f_p(y_i|x_i)$가 로그-정규 분포, log-normal(μ_i^*, σ^2)를 따른다고 가정한다. 즉

$$f_p(y_i|\mu_i^*) = \frac{1}{\sqrt{2\pi}\sigma}\frac{1}{y_i}\exp\left(\frac{(\log(y_i)-\mu_i^*)^2}{2\sigma^2}\right)$$

을 가정한다. 여기서 $\mu_i^* = \beta_0 + \beta_1 x_i$이고 또한 모집단 분포의 평균은

$$E_p(y_i|x_i) = \mu_i = \exp\left(\beta_0 + \beta_1 x_i + \frac{\sigma^2}{2}\right) = \exp\left(\mu_i^* + \frac{\sigma}{2}\right) \tag{3.5}$$

이 된다. 이제 $y_i f_p(y_i|x_i)$의 분포를 $f_p^*(y_i|x_i)$라 하면

$$f_p^*(y_i|x_i) = y_i f_p(y_i|x_i) \propto y_i \frac{1}{y_i}\exp\left(\frac{(\log(y_i)-\mu_i^*)^2}{2\sigma^2}\right) = \exp\left(\frac{(\log(y_i)-\mu_i^*)^2}{2\sigma^2}\right)$$

이므로 최종적으로 $f_p^*(y_i|x_i) = (1/\sqrt{2\pi}\sigma)\exp(-(\mu_i^* + \sigma^2/2))\exp((\log(y_i)-\mu_i^*)^2/2\sigma^2)$가 된다. 다음으로 $f_p^*(y_i|x_i)$의 기댓값을 구하면

$$\exp\left(\mu_i^* + \frac{3}{2}\sigma^2\right) = \exp\left(\mu_i^* + \frac{1}{2}\sigma^2\right) \times \exp\left(\sigma^2\right) = \mu_i \times \exp\left(\sigma^2\right)$$

이 된다. 이제 선형 응답률 모형인 $E_p(\pi_i|y_i, x_i) = b_0 + b_1 y_i$을 가정하였으므로 식 (2.5)에 의해 표본 분포가 구해지며 이를 이용하여 기댓값을 구하면 다음의 결과를 얻는다.

$$\mu_i^{(s)} = E_p(y_i|x_i) = \frac{b_0}{b_0 + b_1\mu_i}\mu_i + \frac{b_1\mu_i}{b_0 + b_1\mu_i}\left(\mu_i \exp\left(\sigma^2\right)\right) = \mu_i + \frac{b_1\mu_i^2}{b_0 + b_1\mu_i}\left(\exp\left(\sigma^2\right) - 1\right).$$

또한 위의 식을 정리하면 다음의 결과를 얻는다.

$$\mu_i = \mu_i^{(s)} / \left(1 + \frac{b_1\mu_i}{b_0 + b_1\mu_i}\left(\exp\left(\sigma^2\right) - 1\right)\right). \tag{3.6}$$

이제 계산을 간단하게하기 위해 $\{b_1\mu_i^2/(b_0 + b_1\mu_i)\}(\exp(\sigma^2) - 1) \approx \{b_1\mu_i^{(s)2}/(b_0 + b_1\mu_i^{(s)})\}(\exp(\sigma^2) - 1)$라 하면 최종적으로 다음의 결과를 얻는다.

$$\mu_i = \mu_i^{(s)} - \frac{b_1\mu_i^{(s)2}}{b_0 + b_1\mu_i^{(s)}}\left(\exp\left(\sigma^2\right) - 1\right). \tag{3.7}$$

따라서 추정된 근사적 편향은 $\{b_1\mu_i^{(s)2}/(b_0 + b_1\mu_i^{(s)})\}(\exp(\sigma^2) - 1)$으로 얻어진다.

3.2. 파워형 응답률 모형에서의 편향 추정

3.2.1. 파워형 응답률 모형의 모수 추정 이 절에서는 감마분포와 로그-정규분포에 사용될 수 있는 파워형 응답률 모형을 가정하였으며 사용된 파워형 응답률 모형의 정의는 다음과 같다.

$$E_p(\pi_i|y_i, x_i) = c_0 y_i^{c_1}. \tag{3.8}$$

파워형 응답률 모형을 이용하면 표본 분포인 식 (2.1)을 계산에 필요한 다음의 결과를 얻는다.

$$\frac{E_p(\pi_i|y_i, x_i)}{E_p(\pi_i|x_i)} = \frac{c_0 y_i^{c_1}}{c_0 E_p(y_i^{c_1}|x_i)} = \frac{y_i^{c_1}}{E_p(y_i^{c_1}|x_i)}. \tag{3.9}$$

식 (3.9)를 살펴보면 모수 c_0는 표본 분포에 영향을 주지 않기 때문에 편향에도 영향을 주지 않는다. 이제 선형 응답률 모형의 경우처럼 파워형 응답률 모형에서도 모형에 포함된 모수를의 추정이 필요하다. 이를 위해 선형 응답률 모형에서 사용한 방법을 적용하여 얻어진 다음 모형을 이용해 모수를 추정한다.

$$\log\left(\frac{1}{w_i}\right) = \log(c_0) + c_1 \log(y_i) + \eta_i. \tag{3.10}$$

따라서 세부층에서 얻어진 세부층 가중치 w_i와 자료 y_i를 이용하여 c_0, c_1을 추정하게 된다. 이때 η_i는 독립이고, 같은 분포를 따르며 등분산성이 만족된다고 가정한다.

3.2.2. 초모집단모형의 오차가 감마분포인 경우의 편향 추정 이 절에서는 초모집단모형의 오차가 감마분포를 따르는 경우를 살펴보았다. 이제 식 (2.1)에 식 (3.9)를 대입하면 다음의 표본 분포 결과를 얻는다.

$$f_s(y_i|x_i) = \frac{y_i^{c_1}}{E_p(y_i^{c_1}|x_i)} f_p(y_i|x_i). \tag{3.11}$$

또한 감마분포를 따르기 때문에 $E_p(y_i^{c_1}|x_i) = \Gamma(\alpha + c_1)\beta_1^c/\Gamma(\alpha)$가 얻어진다

따라서 $f_s(y_i|x_i)$는 쉽게 $\Gamma(\alpha + c_1, \beta) = \Gamma(\alpha + c_1, \mu_i/\alpha)$가 되는 것을 알 수 있다. 결국 표본 분포의 기댓값은 다음과 같이 얻어진다.

$$\mu_i^{(s)} = \frac{(\alpha + c_1)}{\alpha} \times \mu_i. \tag{3.12}$$

따라서 최종적으로 $\mu_i = \{\alpha/(\alpha + c_1)\} \times \mu_i^{(s)} = \mu_i^{(s)} - \{c_1/(\alpha + c_1)\} \times \mu_i^{(s)}$이 되며 추정된 편향은 $\{c_1/(\alpha + c_1)\} \times \mu_i^{(s)}$가 된다.

3.2.3. 초모집단모형의 오차가 로그-정규분포인 경우의 편향 추정 이 절에서는 초모집단모형의 오차가 로그-정규분포를 따르고 응답률 모형은 파워형 응답률을 따른다고 가정한다. 즉 모집단 분포 $f_p(y_i|x_i)$는 다음과 같다.

$$f_p(y_i|\mu_i^*) = \frac{1}{\sqrt{2\pi}\sigma} \frac{1}{y_i} \exp\left(\frac{(\log(y_i) - \mu_i^*)^2}{2\sigma^2}\right)$$

따라서 표본 분포인 식 (3.11)의 $f_s(y_i|x_i) = \{y_i^{c_1}/E_p(y_i^{c_1}|x_i)\} f_p(y_i|x_i)$에서 $E_p(y_i^{c_1}|x_i) = \exp(c_1\mu_i^* + (1/2)c_1^2\sigma^2)$로 알려져 있으므로 표본 분포 $f_s(y_i|x_i)$의 기댓값은 쉽게 다음의 값으로 얻어 진다.

$$\begin{aligned} \mu_i^{(s)} &= \exp\left[(c_1 + 1)\mu_i^* + \frac{1}{2}(c_1 + 1)^2\sigma^2 - c_1\mu_i^* + \frac{1}{2}c_1^2\sigma^2\right] \\ &= \exp\left(\mu_i^* + \frac{\sigma^2}{2} + c_1\sigma^2\right) = \exp\left(\mu_i^* + \frac{\sigma^2}{2}\right)\exp(c_1\sigma^2) \\ &= \mu_i \times \exp(c_1{\sigma^2}) \end{aligned} \tag{3.13}$$

따라서 최종적으로 $\mu_i = \mu_i^{(s)}/\exp(c_1\sigma^2)$이 얻어진다.

3.3. 편향보정 모형균 추정량

본 연구에서 제안한 편향보정 추정량은 4개로 각각 $M_{LG}, M_{LL}, M_{PG}, M_{PL}$로 표시하였다. 그리고 기존에 흔히 사용하는 추정량은 M_S로 표시하였고 편향을 보정하지는 않지만 세부층 정보를 이용한 추정

량을 M_{ST}로 표시하였다. 제안된 편향보정 추정량의 첨자에서 앞 첨자 L과 P는 각각 선형과 파워형 응답률 모형을 그리고 뒤 첨자 G와 L은 각각 감마 분포와 로그-정규 분포를 의미한다. 연구에 사용된 추정량의 정의는 다음과 같다.

- M_S: 세부층 가중치를 고려하지 않고 층 내의 모든 자료에 동일한 가중치를 적용한다.

$$\hat{\bar{Y}}_S = \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{n_{hi}} w y_{hi}. \tag{3.14}$$

- M_S: 층화추출법의 층화추정량을 사용한다. 즉 h 세부층 가중치인 $w_{hi}=w_h$와 자료 y_{hi}를 이용한다.

$$\hat{\bar{Y}}_{ST} = \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{n_{hi}} w_h y_{hi}. \tag{3.15}$$

- M_{LG}: 응답률 모형은 선형모형을 따르고 초모집단모형의 오차가 감마 분포를 따를 경우의 편향보정 추정량은 수식 (3.4)를 이용한다. 여기서 추정된 편향은 $\{b_1/(b_0+b_1\mu_i^{(s)})\}\{\mu_i^{(s)2}/\alpha\}$이므로 식 (3.15)에서 얻어진 추정값에서 편향을 제거한 후 얻어진다. 즉 식 (3.16)을 사용한다.

$$\hat{\bar{Y}}_{LG} = \hat{\bar{Y}}_{ST} - \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{n_{hi}} w_{hi}\left(\frac{\hat{b}_1}{\hat{b}_0+\hat{b}_1\hat{\mu}_i^{(s)}}\frac{\hat{\mu}_i^{(s)}}{\hat{\alpha}}\right), \tag{3.16}$$

  여기서 $\hat{b}_0, \hat{b}_1$은 식 (2.7)을 이용하여 얻어지며 $\hat{\mu}_i^{(s)}$는 식 (3.2)를 이용하여 얻어진다. 또한 $\hat{\alpha}$은 관심변수 y_i의 평균과 분산을 이용한 적률추정법으로 추정한다.

- M_{LL}: 응답률 모형은 선형모형을 따르고 초모집단모형의 오차가 로그-정규 분포를 따를 경우의 편향보정 추정량으로 수식 (3.7)을 이용한다. 추정된 편향은 $\{b_1\mu_i^{(s)2}/(b_0+b_1\mu_i^{(s)})\}(\exp(\sigma^2)-1)$이므로 식 (3.15)에서 얻어진 값에서 편향을 제거한 후 얻어진다. 즉 식 (3.17)을 사용한다.

$$\hat{\bar{Y}}_{LL} = \hat{\bar{Y}}_{ST} - \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{n_{hi}} w_{hi}\left(\frac{\hat{b}_1\hat{\mu}_i^{(s)2}}{\hat{b}_0+\hat{b}_1\hat{\mu}_i^{(s)}}\left(\exp\left(\hat{\sigma}^2\right)-1\right)\right), \tag{3.17}$$

  여기서 $\hat{b}_0, \hat{b}_1$은 식 (2.7)을 이용하여 얻어지고 $\hat{\mu}_i^{(s)}$는 관심변수의 로그변환 후 회귀분석 결과에서 얻어진 추정값이며 이때 $\hat{\sigma}^2$도 얻어진다.

- M_{PG}: 응답률 모형은 파워형 모형을 따르고 초모집단모형의 오차가 감마 분포를 따를 경우의 편향보정 추정량은 수식 (3.12)를 이용한다. 구해진 편향은 $\{c_1/(\alpha+c_1)\}\mu_i^{(s)}$이므로 식 (3.15)에서 얻어진 값에서 편향을 제거한 후 얻어진다. 즉 식 (3.18)을 사용한다.

$$\hat{\bar{Y}}_{PG} = \hat{\bar{Y}}_{ST} - \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{n_{hi}} w_{hi}\left(\frac{\hat{c}_1}{\hat{\alpha}+\hat{c}_1}\hat{\mu}_i^{(s)}\right), \tag{3.18}$$

  여기서 $\hat{c}_1$은 식 (3.10)을 이용하여 얻어지며 $\hat{\mu}_i^{(s)}$는 식 (3.2)를 이용하여 얻어진다. 또한 $\hat{\alpha}$는 관심변수 y_i의 평균과 분산을 이용한 적률추정법으로 추정한다.

- M_{PL}: 응답률 모형은 파워형 모형을 따르고 초모집단모형의 오차는 로그-정규 분포를 따를 경우의 편향보정 추정량은 수식 (3.13)을 이용한다. 구해진 편향은 $\exp(c_1\sigma^2)$이므로 식 (3.15)에서 얻어진 값에서 편향을 제거한 후 얻어진다. 즉 식 (3.19)를 사용한다.

$$\hat{\bar{Y}}_{PL} = \frac{\hat{\bar{Y}}_{ST}}{\exp\left(\hat{c}_1\hat{\sigma}^2\right)}, \tag{3.19}$$

  여기서 $\hat{c}_1$은 식 (3.10)을 이용하여 얻어지며 관심변수의 로그변환 후 회귀분석 결과에서얻어진 $\hat{\sigma}^2$을 사용한다.

4. 모의실험 설계 및 결과

4.1. 모의실험 설계

본 모의실험에서는 층화추출법을 고려하지만 층으로 나누어진 여러 개의 층 중에서 주어진 한 개의 특정 층의 추정을 고려하였다. 이는 층화추출법에서는 각 층별로 모수추정이 이루어지기 때문에 하나의 층을 고려하여도 일반성을 잃지 않기 때문이다. 다음은 모의실험을 위한 자료생성 과정과 모수추정 방법이다. 전체적인 모의실험 방법은 Chung과 Shin (2017, 2019) 방법을 사용하였다.

- Step 1: 모집단 생성과정

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

  (1) 보조변수 x_i 생성: $x_i = 100 + \gamma_i$, $i = 1, \ldots, N$, 여기서 $\gamma_i \overset{\text{iid}}{\sim} \text{Unif}(0, 100)$이다.
  따라서 보조변수 x_i는 100에서 200사이의 값을 갖는다.

  (2) 초모집단 모형

  ① 감마분포: $y_i \overset{\text{iid}}{\sim} \text{Gamma}(\alpha, \mu_i/\alpha)$,
  여기서 $\mu_i = \beta_0 + \beta_1 x_i$이고 $\beta_0 = 0.01$, $\beta_1 = 0.03$, $\alpha = 10$을 사용한다.

  ② 로그-정규분포: $\log(y_i) = \beta_0 + \beta_1 x_i + \epsilon_i$, $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$,
  여기서 $\beta_0 = 0.01$, $\beta_1 = 0.03$, $\sigma^2 = 1$을 사용한다.
  또한 두 분포 모두 모집단 자료 수 $N = 10{,}000$을 사용한다.

- Step 2: 표본추출과정

  생성된 N개의 모집단 자료에서 n개의 표본을 추출한다. 추출된 자료에서 주어진 응답률 모형에 따라 랜덤으로 무응답을 만든다.

  (3) N개의 모집단 자료에서 단순임의추출(simple random sample)로 n개의 표본을 추출한다. 이때 $n = 100, 200, 500$을 사용한다.

  (4) 추출된 n개의 표본에서 선형 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, $(\pi_y^{\min}, \pi_y^{\max}) = (0.9, 0.5), (0.5, 0.9), (0.8, 0.4), (0.4, 0.8)$을 사용하여 b_0, b_1을 구하고 y_i에 따라 응답률 π_i를 계산한다. 계산된 응답률에 따라 랜덤으로 무응답을 생성한다. 같은 방법을 파워형 응답률 모형인 $\pi_i = c_0 y_i^{c_1}$, $\pi_i \in [0, 1]$에도 적용한다.

  (5) 응답한 최종 조사 자료 수는 r개이다. 여기서 $(\pi_y^{\min}, \pi_y^{\max}) = (0.9, 0.5)$ 또는 $(\pi_y^{\min}, \pi_y^{\max}) = (0.5, 0.9)$인 경우는 전체 자료의 약 80%가 응답하게 되어 주어진 자료 수 n에 비해 약 20%가 감소한다. $(\pi_y^{\min}, \pi_y^{\max}) = (0.8, 0.4)$ 또는 $(\pi_y^{\min}, \pi_y^{\max}) = (0.4, 0.8)$인 경우는 전체 자료의 약 70%가 응답하게 되어 주어진 자료 수 n에 비해 약 30%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 층을 L개의 세부층으로 나눈다. 실제 자료 분석에서는 모집단에 보조변수 x_i의 정보만 있으므로 보조변수를 기준으로 세부층을 나눈다.

  (6) 보조변수 x_i를 기준으로 분위수를 구한 후, 얻어진 분위수를 이용하여 모집단을 L개의 세부층으로 나눈다. 여기서 L은 표본 수 n에 따라 $L = 4$ 또는 5 그리고 15, 30을 사용한다.

- Step 4: 모수추정

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

  (8) 선형 응답률 모형은 식 (2.7)을 사용하고 파워형 응답률 모형은 식 (3.10)을 사용한 모형으로 단순 회귀분석을 이용하여 모수 b_0, b_1과 c_0, c_1을 추정한다.

  (9) 추출된 자료 (y_i, x_i)와 초모집단모형에 기초한 회귀분석을 실시하여 $\mu_i^{(s)}$ 및 σ^2을 추정한다. 또한 감마분포에서는 적률추정법으로 α를 추정한다.

  (10) 계산된 결과를 이용하여 식 (3.14)에서 식 (3.19)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.

이제 얻어진 평균 추정값은 다음의 비교통계량, 편향(bias), 절대편향(absolute bias; Abias) 그리고 제곱근 MSE(root mean squared error; RMSE)을 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 위함이다. 이에 r번째 반복 모집단의 참값을 $\bar{Y}_r$로 표시하였다.

4.2. 모의실험 결과

선형 응답률 모형 및 파워형 응답률 모형인 경우에서 $N = 10{,}000$과 $n = 100, 200, 500$을 이용한 모의실험을 수행하였다.

4.2.1. 선형 응답률 모형인 경우 Table 4.1과 Table 4.2에 선형 응답률 모형을 적용한 결과를 수록하였다. 선형 응답률과 감마분포를 가정한 후 얻어진 편향보정 추정량은 $\hat{\bar{Y}}_{LG}$이고 로그-정규분포를 가정한 후 얻어진 편향보정 추정량은 $\hat{\bar{Y}}_{LL}$이다.

먼저 Table 4.1에서 기존의 방법인 $\hat{\bar{Y}}_S$의 편향 결과를 살펴보면 $\pi_y^{\min}$이 크고 $\pi_y^{\max}$이 작은 경우에는 큰 음수 값을 보이고 있어 과소추정이 되는 것을 확인할 수 있다. 반대로 $\pi_y^{\min}$이 작고 $\pi_y^{\max}$가 큰 경우의 편향을 살펴보면 편향이 매우 큰 양수를 보이고 있어 과대추정 되는 것을 알 수 있다. 물론 이는 예상된 결과이다. 반면 $\hat{\bar{Y}}_{ST}$는 $\hat{\bar{Y}}_S$에 비해 크게 편향이 줄어든 것을 확인할 수 있으며 $\hat{\bar{Y}}_{LG}$가 가장 작은 편향을 보이고 있다.

$\hat{\bar{Y}}_S$의 절대 편향과 RMSE의 경우 편향의 영향으로 자료 수가 100에서 500으로 증가하였음에도 그 크기가 줄어들지 않았다. 반면 $\hat{\bar{Y}}_{ST}$는 자료의 크기가 증가하면서 절대 편향과 RMSE가 줄어들었으며 $\hat{\bar{Y}}_{LG}$의 경우 매우 크게 줄어들었다. 따라서 본 연구에서 제안한 편향보정 추정량이 매우 우수한 결과를 주고 있음을 확인할 수 있다.

Table 4.1. Results for gamma distribution with linear response function

$\pi_y^{\min}$	$\pi_y^{\max}$	n	r	L	Bias			Abias			RMSE		
					$\hat{\bar{Y}}_S$	$\hat{\bar{Y}}_{ST}$	$\hat{\bar{Y}}_{LG}$	$\hat{\bar{Y}}_S$	$\hat{\bar{Y}}_{ST}$	$\hat{\bar{Y}}_{LG}$	$\hat{\bar{Y}}_S$	$\hat{\bar{Y}}_{ST}$	$\hat{\bar{Y}}_{LG}$
		100	78	5	−17.8	−5.7	−2.1	18.5	7.4	6.6	21.1	9.0	8.3
0.9	0.5	200	156	15	−17.9	−4.7	−2.4	18.0	5.6	4.6	19.7	6.7	5.7
		500	390	30	−17.7	−4.6	−2.3	17.7	4.8	3.2	18.5	5.5	3.9
		100	70	5	−20.4	−6.7	−2.6	20.9	8.5	7.7	23.5	10.2	9.6
0.8	0.4	200	140	15	−20.6	−5.6	−3.0	20.7	6.5	5.3	22.2	7.8	6.6
		500	350	30	−20.3	−5.4	−2.9	20.3	5.5	3.7	21.0	6.3	4.5
		100	52	5	26.1	5.4	2.9	27.0	7.7	6.9	31.8	9.5	8.6
0.4	0.8	200	104	15	26.2	4.4	2.8	26.3	5.5	4.8	28.9	6.8	6.0
		500	260	30	26.3	4.2	2.4	26.3	4.4	3.2	27.5	5.2	3.9
		100	62	5	21.7	4.5	2.3	22.9	6.9	6.3	27.1	8.5	7.9
0.5	0.9	200	124	15	21.3	3.5	2.0	21.5	4.7	4.1	24.1	5.8	5.1
		500	310	30	21.7	3.5	1.7	21.7	3.7	2.6	22.8	4.4	3.3

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

Table 4.2. Results for log-normal distribution with linear response function

$\pi_y^{\min}$	$\pi_y^{\max}$	n	r	L	Bias			Abias			RMSE		
					$\hat{\bar{Y}}_S$	$\hat{\bar{Y}}_{ST}$	$\hat{\bar{Y}}_{LL}$	$\hat{\bar{Y}}_S$	$\hat{\bar{Y}}_{ST}$	$\hat{\bar{Y}}_{LL}$	$\hat{\bar{Y}}_S$	$\hat{\bar{Y}}_{ST}$	$\hat{\bar{Y}}_{LL}$
		100	80	5	−65.2	−56.9	−44.4	65.2	57.0	49.1	68.5	60.4	56.5
0.9	0.5	200	160	15	−66.0	−56.1	−44.6	66.0	56.1	45.5	67.7	58.1	49.5
		500	400	30	−65.4	−55.8	−46.2	65.4	55.8	46.2	66.1	56.5	47.8
		100	69	5	−70.8	−62.0	−47.4	70.8	62.1	53.7	74.0	65.4	59.9
0.8	0.4	200	138	15	−71.2	−60.8	−48.1	71.2	60.8	49.4	72.8	62.8	53.5
		500	345	30	−70.8	−60.7	−50.2	70.8	60.7	50.2	71.5	61.4	51.9
		100	48	5	89.2	56.6	42.9	92.9	62.1	54.5	114.9	81.0	76.1
0.4	0.8	200	96	15	89.0	57.3	43.3	89.4	58.9	48.3	101.3	70.8	63.2
		500	240	30	88.4	55.2	39.0	88.4	55.2	40.0	93.7	60.5	47.3
		100	58	5	66.6	42.2	32.2	71.2	49.2	45.3	90.1	65.1	63.2
0.5	0.9	200	116	15	66.0	43.1	32.7	66.9	45.7	39.2	78.2	56.1	51.4
		500	290	30	66.0	43.1	32.7	66.9	45.7	39.2	78.2	56.1	51.4

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

다음으로 선형 응답률 모형과 로그-정규 분포를 사용하여 얻은 결과인 Table 4.2의 결과를 살펴보면 모든 비교 통계량에서 단순 평균을 사용한 $\hat{\bar{Y}}_S$에 비해 $\hat{\bar{Y}}_{ST}$가 우수한 결과를 주며 본 연구에서 제안한 추정량인 $\hat{\bar{Y}}_{LL}$이 가장 우수한 결과를 주는 것을 확인할 수 있다. 전체적으로 감마분포를 가정한 Table 4.1의 결과와 로그-정규분포 결과인 Table 4.2의 결과가 매우 유사하다. 결론적으로 응답률이 관심변수와 선형 관계가 있는 사업체조사에서는 본 연구에서 제안한 추정량이 매우 효과적으로 사용될 수 있다고 판단된다.

4.2.2. 파워형 응답률 모형인 경우 Table 4.3과 Table 4.4에 파워형 응답률을 적용한 결과가 수록되어 있다. 파워형 응답률은 선형 응답률에 비해 전체적으로 응답 수인 r이 작은 것을 확인할 수 있다. 먼저 Table 4.3의 편향 결과를 살펴보면 $\pi_y^{\min}$이 크고 $\pi_y^{\max}$이 작은 경우에 $\hat{\bar{Y}}_S$가 과소 추정되며 반대의 경우 과대 추정되는 결과는 선형 응답률과 같은 결과를 준다. 또한 $\hat{\bar{Y}}_S$에 비해 $\hat{\bar{Y}}_{ST}$가 우수하고 최종적으

Table 4.3. Results for gamma distribution with power response function

$\pi_y^{\min}$	$\pi_y^{\max}$	n	r	L	Bias			Abias			RMSE		
					$\hat{\bar{Y}}_S$	$\hat{\bar{Y}}_{ST}$	$\hat{\bar{Y}}_{PG}$	$\hat{\bar{Y}}_S$	$\hat{\bar{Y}}_{ST}$	$\hat{\bar{Y}}_{PG}$	$\hat{\bar{Y}}_S$	$\hat{\bar{Y}}_{ST}$	$\hat{\bar{Y}}_{PG}$
		100	59	5	−12.2	−2.2	−0.8	15.5	6.8	6.8	18.7	8.6	8.5
0.9	0.5	200	118	15	−12.5	−1.9	−1.0	13.6	4.7	4.6	15.9	6.0	5.8
		500	295	30	−12.6	−1.7	−0.7	12.8	3.1	2.8	14.1	3.8	3.5
		100	50	5	−14.2	−2.7	−1.1	17.3	7.8	7.7	20.8	9.8	9.7
0.8	0.4	200	100	15	−14.6	−2.3	−1.4	15.6	5.5	5.3	18.1	7.0	6.8
		500	250	30	−14.9	−2.1	−1.1	15.0	3.5	3.3	16.3	4.4	4.0
		100	62	5	16.2	2.7	0.9	18.3	6.2	5.9	22.4	7.7	7.4
0.4	0.8	200	124	15	16.4	2.3	1.1	17.0	4.1	3.8	19.6	5.1	4.8
		500	310	30	16.0	2.1	1.0	16.1	2.9	2.4	17.4	3.5	3.0
		100	72	5	14.0	2.4	0.8	16.1	5.6	5.5	19.9	7.1	6.9
0.5	0.9	200	144	15	13.7	1.9	0.9	14.3	3.7	3.6	16.8	4.7	4.5
		500	360	30	13.7	1.8	0.7	13.7	2.6	2.2	14.9	3.2	2.7

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

Table 4.4. Results for log-normal distribution with power response function

$\pi_y^{\min}$	$\pi_y^{\max}$	n	r	L	Bias			Abias			RMSE		
					$\hat{\bar{Y}}_S$	$\hat{\bar{Y}}_{ST}$	$\hat{\bar{Y}}_{PL}$	$\hat{\bar{Y}}_S$	$\hat{\bar{Y}}_{ST}$	$\hat{\bar{Y}}_{PL}$	$\hat{\bar{Y}}_S$	$\hat{\bar{Y}}_{ST}$	$\hat{\bar{Y}}_{PL}$
		100	64	4	−36.3	−25.0	−17.1	45.9	38.7	37.7	53.4	46.3	46.3
0.9	0.5	200	128	15	−36.8	−22.3	−12.7	40.0	31.6	29.6	46.0	37.7	36.5
		500	320	30	−35.1	−21.2	−12.0	35.8	24.5	19.9	39.5	28.4	24.1
		100	53	4	−41.6	−28.8	−20.1	51.8	43.3	41.3	59.5	51.4	51.0
0.8	0.4	200	106	15	−42.5	−26.3	−15.5	45.4	35.7	33.3	51.7	42.4	40.8
		500	265	30	−40.9	−24.9	−14.4	41.3	28.7	23.4	45.2	32.9	28.2
		100	60	4	53.2	30.8	17.9	60.1	41.6	36.6	78.1	56.3	50.7
0.4	0.8	200	120	15	53.1	31.7	19.9	55.0	36.7	30.7	66.2	46.9	40.9
		500	300	30	52.9	30.4	17.4	53.0	31.0	21.3	58.7	36.8	27.2
		100	70	4	42.6	24.1	13.3	50.9	36.6	33.6	66.6	49.1	45.1
0.5	0.9	200	140	15	42.6	25.4	15.5	45.1	30.8	26.2	55.3	39.7	35.0
		500	350	30	42.8	24.4	13.5	43.0	25.5	18.2	48.3	30.6	23.1

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

로 편향을 보정한 추정량인 $\hat{\bar{Y}}_{PG}$가 가장 우수한 것을 확인할 수 있으며 편향이 크게 줄어든 것도 확인할 수 있다. 전체적으로 선형 응답률 모형과 유사한 결과를 준다. 다만 본 연구에서 제안한 추정량이 가장 우수한 결과를 주지만 $\hat{\bar{Y}}_{ST}$와 $\hat{\bar{Y}}_{PG}$의 절대 편향과 RMSE를 비교하면 상대적으로 그 값이 크게 줄어들지 않는다. 즉 선형 응답률에 비해 절대 편향과 RMSE의 감소 폭이 상대적으로 작은 것을 확인할 수 있다. 이러한 결과는 Table 4.4에서도 확인된다. 결론적으로 파워형 응답률에서도 본 연구에서 제안한 편향보정 추정량을 사용하면 편향을 크게 줄일 수 있을 것으로 판단되며 이를 통해 절대 편향과 RMSE도 크게 축소시킬 수 있을 것으로 판단된다.

5. 결론

본 논문에서는 사업체조사에서 발생한 무응답을 적절히 처리하는 방법을 연구하였다. 특히 사업체 조사에서는 사업체 규모가 커질수록 관심변수의 분산이 커지는 자료가 흔히 얻어지고 있으며 이를 해결하기

위한 방법의 하나로 로그 변환 후 회귀분석을 실시한다. 따라서 초모집단모형으로 로그선형 모형을 고려할 수 있으며 이에 적합한 분포로 본 연구에서는 감마 분포와 로그-정규 분포를 고려하였다. 또한 선형 응답률 모형과 파워 응답률 모형을 사용하여 무응답으로 인해 발생된 편향의 크기를 추정하고 이를 축소할 수 있는 편향보정 추정량을 제안하였다. 모의실험 결과 본 논문에서 제안한 편향보정 추정량 사용으로 무응답으로 인해 발생된 편향을 크게 축소할 수 있으며 절대편향 및 RMSE도 모두 줄일 수 있음을 확인하였다. 다만 본 연구에서 얻어진 결과는 초모집단모형이 로그-선형 모형을 따르고 응답률 모형이 선형, 파워형 등 관심변수 분포와 응답률 모형이 알려진 경우에만 사용이 가능하다는 단점이 있다.

결론적으로 본 연구에서 제안한 편향보정 추정량을 관심변수에 영향을 받은 무응답이 발생한 사업체 조사에 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, *Korean Journal of Applied Statistics*, **30**, 993–1004.

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

Kim, J. K. and Skinner, C. J. (2013), Weighting in survey analysis under informative sampling, *Biometrika*, **100**, 385–398.

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, *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.

Pfeffermann, D. and Sverchkov, M. (2003). Small area estimation under informative sampling. In *2003 Joint Statistical Meeting - Section on Survey Research Methods*, 3284–3295.

사업체조사에서의 무응답 편향보정 추정에 관한 연구

정희영[a] · 신기일[a,1]

[a]한국외국어대학교 통계학과

(2019년 10월 22일 접수, 2019년 12월 2일 수정, 2019년 12월 10일 채택)

요 약

표본조사는 비용과 시간을 절약하면서도 주어진 정확성을 만족하는 통계를 얻을 수 있다. 그러나 최근에는 다수의 무응답 발생으로 인해 조사의 정확성이 크게 떨어지고 있다. 무응답은 다양한 이유로 발생하고 있으나 무응답이 관심변수와 함수 관계가 있는 경우에는 이 정보를 이용하여 무응답을 적절히 처리해야 추정의 정확성이 유지될 수 있다. 최근 Chung과 Shin (2017, 2019), Min과 Shin (2018)은 응답률이 관심변수의 지수 또는 선형함수이고 초모집단모형의 오차가 정규분포를 따를 때 무응답으로 인해 발생한 편향을 제거함으로써 추정의 정확성이 향상되는 것을 확인하였다. 이에 본 연구에서는 사업체조사에서 초모집단모형의 오차가 감마분포 또는 로그-정규분포를 따르는 경우에서의 무응답 편향보정 추정량을 제안하였다. 또한 모의실험을 통하여 제안된 추정량의 우수성을 확인하였다.

주요용어: 선형 응답률 모형, 파워형 응답률 모형, 감마분포, 로그-정규 분포, 초모집단모형

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