The Korean Journal of Applied Statistics (2017) DOI: http://dx.doi.org/10.5351/KJAS.2017.30.6.993
30(6), 993–1004

Estimation using informative sampling technique when response rate follows exponential function of variable of interest

Hee Young Chung[a] · Key-Il Shin[a,1]

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

(Received October 17, 2017; Revised November 24, 2017; Accepted December 4, 2017)

Abstract

A stratified sampling method is generally used with a sample selected using the same sample weight in each stratum in order to improve the accuracy of the sampling survey estimation. However, the weight should be adjusted to reflect the response rate if the response rate is affected by the value of the variable of interest. It may be also more effective to adjust the weights by subdividing the stratum rather than using the same weight if the variable of interest has a linear relationship with the continuous auxiliary variables. In this study, we propose a method to increase the accuracy of estimation using an informative sampling design technique when the response rate is an exponential function of the variable of interest and the variable of interest has a linear relationship with the auxiliary variable. Simulation results show the superiority of the proposed method.

Keywords: linear inclusion probability, sample distribution, regression model, sample weight

1. 서론

국내에서는 다수의 표본조사가 실시되고 있으며 흔히 층화추출법이 사용된다. 층화추출법을 사용하기 위해서는 층화변수 또는 층화에 사용되는 보조변수가 필요하며 보조변수 및 층화변수는 표본추출을 위해 준비된 표본틀에 포함되어 있다. 대부분의 층화 표본설계에서는 같은 층에 포함된 자료에 동일한 표본 가중치를 적용한다. 따라서 층화변수를 기준으로 모집단을 층으로 나눈 후 층 내에서 랜덤으로 또는 계통추출로 표본을 추출함으로써 동일한 추출률 또는 표본 포함확률을 갖도록 한다. 최근에는 동일한 추출률을 사용하지 않는 표본설계법인 정보적 표본설계법(informative sampling)이 제안되었으며 이 표본설계법은 관심변수와 연속형 보조변수 간에 관계가 있고, 표본추출과정에서 표본 포함확률이 관심변수의 함수가 되는 표본설계를 말한다. 정보적 표본설계는 대부분의 층화추출법에 적용되고 있으며 정보적 표본설계의 예는 Savitsky와 Toth (2016)를 참조하면 된다.

정보적 표본설계는 1990년대 후반부터 연구가 시작되어 2000년대에도 지속적으로 활발한 연구가 진행되고 있다. 정보적 표본설계는 두 과정으로 나누어진다. 첫 번째 과정은 모집단 관심변수의 자료 생성

This research was supported by Hankuk University of Foreign Studies research fund (2017).
[1]Corresponding author: Department of Statistics, Hankuk University of Foreign Studies, 50 Oedae-ro 54beon-gil, Mohyeon-myeon, Cheoin-gu, Yongin 17035, Korea. E-mail: keyshin@hufs.ac.kr

과정이다. 먼저 모집단은 표본 포함확률 또는 가중치를 구할 수 있어야하기 때문에 유한 모집단(finite population)을 사용한다. 모집단에 포함된 관심변수와 보조변수는 많은 경우 관계가 있으며 이 관계는 모형으로 표현될 수 있는데 이때의 모형을 초모집단모형(super-population model)이라 한다. 결국 정보적 표본설계에서 관심변수는 초모집단모형을 기반으로 생성된다. 두 번째 과정은 표본추출(selection mechanism)과정이다. 정보적 표본설계의 표본추출과정에서 자료가 표본에 포함될 확률인 표본 포함확률(inclusion probability)은 관심변수 자료 값의 함수에 의해 결정된다. 이러한 정보적 표본설계는 현재 사용되고 있는 표본추출방법의 특수한 경우이며, 기존의 표본설계 방법에 비해 관심변수와 보조변수의 정보를 더욱 적극적으로 사용하는 표본설계라 할 수 있다. 정보적 표본설계와 관련된 내용은 Pfeffermann과 Sverchkov (2003), Pfeffermann 등 (1998, 2006)을 참조하기 바란다.

관심변수의 정보를 사용하는 정보적 표본설계가 사용되면 모집단분포와 표본분포는 일치하지 않는 것으로 알려져 있으며 이에 관한 연구로는 Pfeffermann 등 (1998)과 Kim과 Skinner (2013)가 있다. 따라서 모집단 자료에서 만들어진 초모집단모형은 표본 자료에서 관심변수와 보조변수에 의해 만들어지는 모형과 다르게 된다는 단점이 있다. 반면 표본추출과정에서 관심변수 자료 값의 크기에 관계가 있는 표본 포함확률을 사용하게 됨으로써 추정의 정확성은 향상될 수 있다. 그러나 각 관심변수 자료의 표본 포함확률을 사용하기 위해서는 표본틀에 모든 관심변수 자료 값이 있어야 하기 때문에 기존에 사용하고 있는 표본틀 정보를 바탕으로 정보적 표본설계법을 사용하는 것은 실질적으로 거의 불가능하다.

최근 많은 표본조사에서 무응답이 발생하고 있으며 응답률이 층별로 다르거나 관심변수 또는 보조변수의 크기와 관련이 있는 것으로 나타나고 있다. 예를 들면 관심변수인 부채 자료의 경우 부채가 많은 응답자는 응답을 거절할 확률이 높게 나타나며 보조변수인 연령의 경우 30대, 40대에 비해 20대의 응답률이 떨어진다. 또한 Lee 등 (2008)은 종사자 수가 응답률과 관계가 있다는 연구 결과를 발표하였다. 이렇게 관심변수 또는 보조변수 자료 값의 크기가 응답률에 영향을 줄 때 이를 고려하지 않으면 모수추정 결과는 과대 또는 과소 추정될 수 있다. 최근 무응답이 발생한 표본자료의 대표성을 분석하는 기법으로 R-지수 분석이 사용되고 있으며 이 분석을 실시하게 되면 응답률과 관련된 항목 또는 관심변수를 찾을 수 있다. R-지수 분석과 관련된 내용은 Schouten 등 (2009)과 Lee와 Shin (2017)을 참조하기 바란다. 결론적으로 비록 표본 포함확률을 적용하는 정보적 표본설계법을 직접 사용할 수는 없지만 응답률이 관심변수 자료 값과 관계가 있다면 정보적 표본설계 방법에서 얻어진 이론 및 분석 기법을 모수추정에 적용할 수 있으며 이를 통해 과대 또는 과소 추정의 영향을 줄일 수 있게 된다.

본 연구에서는 응답률이 관심변수 자료 값의 지수함수인 경우를 연구하였다. 이는 정보적 표본설계에서 표본 포함확률이 지수형 모형을 따른다고 가정하는 것과 동일하다. 또한 관심변수와 독립변수와의 관계인 초모집단모형은 회귀모형을 따르고 이때 발생하는 오차는 정규분포를 따르는 경우를 연구하였다. 이러한 가정은 정보적 표본설계법의 가장 기초적인 가정이며 이와 관련된 내용은 Pfeffermann 등 (1998)의 논문을 살펴보기 바란다. 또한 층화추출법에서는 다수의 층이 있지만 본 연구에서는 특정된 하나의 층에서의 모수 추정을 연구하였다. 이는 층화추출의 경우 층별로 추정이 이루어지기 때문에 하나의 층을 고려하여도 일반성을 잃지 않기 때문이다.

본 논문의 구성은 다음과 같다. 먼저 2절에서는 기본적인 정보적 표본설계를 설명하였다. 특히 본 연구에서는 표본 포함확률이 지수형이며 초모집단모형인 회귀모형의 오차 분포가 정규분포인 경우의 정보적 표본설계를 설명하였다. 3절에서는 층별 응답률이 지수함수를 따를 때 모수추정의 정확성 및 정밀성을 향상시킬 수 있는 새로운 방법을 제안하였다. 4절에는 모의실험을 통하여 정보적 표본설계 기법을 적용한 경우와 층화추출법에서 사용하는 추정량을 이용한 경우의 성능을 비교하였다. 마지막으로 5절에 결론이 있다.

2. 초모집단모형의 오차가 정규분포인 경우의 정보적 표본설계

2.1. 정보적 표본설계 개요

대부분의 표본설계는 층화추출법이 사용되고 층을 나누기 위한 범주형 층화변수와 관심변수와 관계가 높은 연속형 보조변수가 층화변수로 사용된다. 사업체 조사에서 대표적인 범주형 층화변수로는 지역, 산업분류 등이 있으며 관심변수와 관계가 높은 보조변수이면서 층화변수로 사용되는 변수로는 종사자 수가 있다. 이와 같이 흔히 표본틀에는 층을 나누는 층화변수 및 연속형 보조변수가 존재한다. 정보적 표본설계는 관심변수가 보조변수의 함수이고 표본추출과정이 관심변수 자료 값에 영향을 받는 표본설계이며 이와 관련된 내용은 Pfeffermann 등 (1998)에 자세히 설명되어 있다. 또한 본 논문에서 사용된 모든 기호는 Pfeffermann 등 (1998)에서 사용한 기호를 사용하였다.

먼저 정보적 표본설계는 유한 모집단에서 관심변수와 보조변수로 만들어진 초모집단모형으로 흔히 선형관계 또는 회귀모형을 가정한다. 물론 모형의 오차 분포에 따라 다양한 모형이 사용되기도 한다. 또한 표본틀에서 표본을 추출하는 과정인 표본추출과정은 자료의 표본 포함확률이 관심변수 y_i의 함수가 되도록 하는 과정이다. 결론적으로 s를 표본 집합의 index라 할 때 i번째 자료가 표본으로 추출될 확률이 $P(i \in s|y_i) = P(i \in s) = \pi_i$가 되면, 즉 y_i가 어떤 값을 갖더라도 표본으로 추출될 확률이 모두 같게 되면 비정보적 표본설계(non-informative sampling)가 되고 만약 다르다면 정보적 표본설계가 된다. Pfeffermann 등 (1998)은 정보적 표본설계 하에서 x_i를 보조변수라 하고, θ^*를 모수 θ의 함수라 할 때 표본분포는

$$f_s(y_i|\theta^*, x_i) = f(y_i|i \in s, x_i) = \frac{\Pr(i \in s|y_i, x_i) f_p(y_i|\theta, x_i)}{\Pr(i \in s|x_i)}$$

가 되고 또한 $\Pr(i \in s|y_i, x_i) = E_p(\pi_i|y_i, x_i)$, $\Pr(i \in s|x_i) = E_p(\pi_i|x_i)$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 모집단분포이고 $f_s(y_i|x_i)$는 표본분포이며 $E_p(\pi_i|y_i, x_i)$는 x_i, y_i가 주어졌을 때 자료가 표본에 포함될 포함확률이다. 만약 $E_p(\pi_i|y_i, x_i) = E_p(\pi_i|x_i)$이면 모집단 분포와 표본분포는 같아진다.

2.2. 지수형 표본 포함확률에서의 표본분포

이 절에서는 표본 포함확률이 지수형을 따르고 초모집단모형인 회귀모형의 오차가 정규 분포를 따르는 경우의 표본분포를 살펴보았다. 먼저 사용된 회귀모형은 다음과 같다.

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

따라서 모집단분포는 다음과 같이 표현된다.

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

다음으로 표본 포함확률이 관심변수와 보조변수의 함수이고 식 (2.3)의 지수 형태를 취한다고 가정하자. 이는 초모집단모형의 오차가 정규분포일 때 흔히 사용하는 가정이다.

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

그러면 식 (2.3)에 의해 식 (2.1)의 $E_p(\pi_i|x_i)$는 다음과 같이 구해진다.

$$E_p(\pi_i|x_i) = E[E_p(\pi_i|y_i, x_i)] = E(\exp(a_0 + a_1 y_i + g(x))) = \exp(a_0 + g(x))E(\exp(a_i y_i)). \quad (2.4)$$

이제 식 (2.3)과 (2.4)를 식 (2.1)에 대입하면 표본분포는 다음과 같이 얻어진다.

$$f_s(y_i|x_i) = \frac{E_p(\pi_i|y_i, x_i)f_p(y_i|x_i)}{E_p(\pi_i|x_i)} = \frac{\exp(a_1 y_i)}{E(\exp(a_1 y_i))} f_p(y_i|x_i). \quad (2.5)$$

최종적으로 식 (2.5)를 정리하면 다음의 표본분포 결과를 얻는다.

$$f_s(y_i|x_i) = N\left(\beta_0 + a_1\sigma^2 + \beta_1 x_i, \sigma^2\right). \quad (2.6)$$

결국 모집단분포인 식 (2.2)와 표본분포인 식 (2.6)을 비교하면 회귀모형의 절편이 β_0에서 $\beta_0 + a_1\sigma^2$으로 변한 것을 확인할 수 있으며 이 내용은 정보적 표본설계 관련 논문에서 쉽게 찾을 수 있다.

3. 제안된 모수추정 방법

3.1. 응답률이 관심변수의 함수인 경우의 추정

무응답은 단위 무응답과 항목 무응답으로 분류된다. 단위 무응답의 경우 조사 자체를 거부하여 모든 항목이 무응답인 경우를 의미하고 항목 무응답은 특정 항목에 응답을 하지 않은 것을 의미한다. 무응답이 랜덤으로 발생한 경우에는 다양한 방법으로 이를 처리할 수 있으며 또한 기본적으로 층화추출에서 무응답 처리는 층별로 이루어진다. 예를 들면 단위 무응답이 발생한 경우에는 층별로 표본 대체(sample substitution)를 이용하여 단위 무응답을 해결할 수 있으며 항목 무응답이 발생하면 층별로 무응답 대체법을 이용하여 이를 해결할 수 있다.

그러나 서론에서 설명한 부채자료와 같이 단위 무응답은 관심변수 y_i 값의 크기와 관련이 있을 수 있다. 즉 응답률이 관심변수 y_i의 함수가 될 수 있다. 그러나 현실적으로는 특정 관심변수 y_i가 큰 경우 응답률이 높거나 반대로 낮을 수 있음에도 불구하고 관심변수 y_i 값을 알 수 없기 때문에 관심변수 y_i와 무관하게 같은 층에 포함된 다른 표본으로 표본을 대체하게 된다. 당연히 이렇게 대체된 대체 표본의 응답률 또한 관심변수의 함수가 되기 때문에 최종적으로 조사된 자료의 응답률도 관심변수 y_i 값에 영향을 받게 된다. 항목 무응답의 경우에도 응답률이 관심변수 y_i 값의 함수라면 얻어진 응답 결과를 바탕으로 무응답 대체가 이루어지기 때문에 최종적으로 얻어진 자료의 응답률은 관심변수 y_i 값에 영향을 받게 된다.

결론적으로 층별로 최종 조사된 자료의 응답률이 관심변수 y_i 값에 영향을 받음에도 불구하고 랜덤으로 무응답이 발생한 경우에서 사용하는 방법으로 무응답을 처리하게 되면 추정 결과는 과대 또는 과소 추정될 수 있다.

3.2. 층별 지수형 응답률의 모수 추정

식 (2.1)에서 정보적 표본설계의 핵심은 표본 포함확률이 관심변수 y_i의 함수라는 것이다. 대표적으로 Savitsky와 Toth (2016)처럼 층화추출법에서 관심변수로 층을 나눌 경우에는 정보적 표본설계가 된다. 그러나 현실적으로 모든 관심변수 y_i의 값은 알 수 없기 때문에 실질적인 표본설계에서 정보적 표본설계를 사용하는 것은 거의 불가능하다. 반면에 실사에서 단위 또는 항목 무응답으로 인해 응답률이 관심변수 y_i 값의 함수가 되는 경우는 응답률이 표본 포함확률과 같은 개념이 된다. 따라서 특정 층의 응답률이 관심변수 y_i의 함수가 되는 경우에는 식 (2.1)의 결과를 이용할 수 있다. 다만 표본 포함확률은 표본

설계 당시에 결정되지만 응답률은 설계 당시 얻어지지 않기 때문에 응답률에 포함된 모수는 추정되어야 한다. 따라서 특정 층의 응답률이 지수형을 따른다면 식 (2.3)을 사용할 수 있으며 이때 $g(x)$가 결과에 영향을 주지 않기 때문에 $E_p(\pi_i|y_i, x_i) = \exp(a_0 + a_1 y_i)$가 되고 모형에 포함된 모수 a_0, a_1은 추정되어야 한다. 모수를 추정하기 위해서는 자료에서 얻어진 y_i와 표본 포함확률이 필요하다. 이를 위해 본 연구에서는 주어진 하나의 특정 층을 여러 개의 세부 층으로 나누는 방법을 제안하였다. 이는 초모집단모형에서 관심변수가 보조변수와 선형 관계가 있기 때문에 층을 세분화하여 응답률을 구한 후 이 결과를 사용하게 되면 정확성 및 정밀성이 높은 추정량을 얻을 수 있기 때문이기도 하다. 이제 주어진 하나의 층을 L개의 세부 층으로 나눈다고 하자. 이때 관심변수의 모집단 정보는 알 수 없지만 보조변수의 모집단 정보는 알 수 있다. 따라서 층을 나누는 기준은 보조변수 x_i에 의해 이루어졌으며 보조변수 x_i를 이용하여 등간격으로 L개의 세부 층을 구성하였다. 물론 구성된 L개의 세부 층을 이용하더라도 개별 표본 포함확률 π_i를 구할 수는 없다. 그러나 세부 층 h에 포함된 π_i를 $\pi_{i\in h} = \pi_h = n_h/N_h$로 동일하게 줄 수 있게 된다.

다음으로 본 논문에서 사용할 지수형 포함확률을 고려하자.

$$E_p(\pi_i|y_i, x_i) = E_p\left(\pi_i, i \in h|y_i, x_i\right) = \exp(a_0 + a_1 y_i). \tag{3.1}$$

이제 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 식 (3.1)에 적용하게 되면 $1/w_i \approx \exp(a_0 + a_1 y_i)$이 얻어진다. 결국 다음의 모형을 이용하여 a_0, a_1을 추정하게 된다.

$$\log\left(\frac{1}{w_i}\right) = a_0 + a_1 y_i + \eta_i, \tag{3.2}$$

여기서 $E(\eta_i) = 0$이고 $\mathrm{Var}(\eta_i) = \sigma_\eta^2$이다. 참고로 정보적 표본설계에서는 설계자가 미리 정해진 a_0, a_1을 이용하여 표본 포함확률을 구하고 이를 기반으로 표본을 추출하게 된다. 반면에 응답률을 기반으로 얻어진 π_i는 관심변수 y_i를 이용하여 얻어진 것이 아니라 보조변수 x_i를 이용하여 얻어지게 된다. 결국 본 연구에서 제안한 추정량은 식 (3.2)를 사용하여 추정된 $\hat{a}_0, \hat{a}_1$을 이용하기 때문에 정확성 면에서 떨어질 수 있다. 물론 식 (3.2)에서 얻어진 추정된 $\hat{a}_0, \hat{a}_1$ 값이 표본분포에 적용된다.

3.3. 제안된 추정량

만약 응답률이 관심변수와 무관하다고 판단된다면 층 내에 속한 모든 자료의 가중치를 동일하게 적용할 수 있다. 따라서 주어진 하나의 특정 층 모평균은 단순평균으로 추정할 수 있다. 만약 응답률이 자료에 따라 다르다면 층 내의 가중치를 달리하여 추정하는 것이 타당하다. 이때 현실적으로 사용할 수 있는 방법이 주어진 하나의 특정 층을 여러 개의 세부 층으로 나눈 후 층화추출법을 이용하는 것이다. 즉 세부 층의 가중치에 기초한 가중 평균을 사용하여 모수를 추정하는 것이 타당하다. 특히 관심변수가 보조변수의 선형함수이면 가중 평균을 사용함으로써 추정의 효율을 높일 수 있다. 또한 세부 층을 이용하여 얻어진 가중치와 정보적 표본설계에서 얻어진 결과를 결합한 추정량을 사용할 수 있다. 이에 본 연구에서는 세부 층으로 나눈 후 얻어진 가중치를 적용한 추정량과 정보적 표본설계 기법을 적용한 새로운 추정량을 제안하였다.

E1: 주어진 특정 층 내의 모든 자료에 동일한 가중치가 사용된 단순 평균 추정량을 사용한다.

$$\hat{\bar{Y}}_s = \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{n_h} w_{hi} y_{hi} = \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{n_h} w y_{hi} = \bar{y}. \tag{3.3}$$

E2: 주어진 특정 층을 세부 층으로 층화한 후 층화추출법의 층화추정량을 사용한다. 여기서 세부 층의 가중치를 동일하게 적용한 가중치 $w_{hi} = w_h$와 자료 y_{hi}를 사용한다.

$$\hat{\bar{Y}}_{st} = \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{n_h} w_h y_{hi}. \tag{3.4}$$

E3: 표본 자료를 이용한 단순 회귀추정량 $\hat{\beta}_0, \hat{\beta}_1$과 가중치 $w_{hi} = w_h$를 이용하여 얻어진 예측값의 편향을 보정한 후 얻어진 추정량인 식 (3.5)를 사용한다.

$$\hat{\bar{Y}}_{\text{inf}} = \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{n_h} w_h \left(\hat{\beta}_0 + \hat{\beta}_1 x_{hi} - \hat{a}_1\hat{\sigma}^2\right). \tag{3.5}$$

4. 모의실험

4.1. 모의실험 설계 및 모수추정 방법

본 모의실험에서는 층으로 나누어진 여러 개의 층중에서 주어진 하나의 특정 층의 추정을 고려하였다. 이는 층화추출법에서는 각 층별로 모수추정이 이루어지기 때문에 한 개 층을 고려하여도 일반성을 잃지 않기 때문이다. 다음은 모의실험을 위한 자료생성 과정과 모수추정 방법이다.

Step 1: 모집단 생성과정

초모집단모형이 회귀모형이고 모형의 오차가 정규분포인 경우의 정보적 표본설계를 위한 모집단 자료생성 과정은 다음과 같다.

(1) 보조 자료 x_i 생성: $x_i = 100 + \gamma_i$, $i = 1, \ldots, N$.
여기서 $\gamma_i \overset{\text{iid}}{\sim} \text{Unif}(100, 200)$과 $T\text{-Gam}(1, 100)$을 사용하고 $T\text{-Gam}(1, 100)$은 절단 감마분포로 0과 100 사이의 값을 갖기 위해 100 이상인 값은 버린다. 따라서 보조변수 x_i는 100에서 200 사이의 값을 갖는다.

(2) 초모집단모형 : $y_i = \beta_0 + \beta_1 x_i + \epsilon$, $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$.
여기서 $\beta_0 = 10$, $\beta_1 = 5$, $\sigma^2 = 400,900$과 모집단 자료 수 $N = 10{,}000$을 사용한다.

Step 2: 표본추출과정

생성된 모집단에서 n개의 표본을 추출한다. 추출된 표본 자료에서 랜덤으로 무응답을 만들었으며 응답률은 식 (3.1)을 이용하여 관심변수 y_i에 지수형 관계가 되도록 한다.

(3) N개의 모집단 자료에서 단순임의추출(simple random sample)로 n개의 표본을 추출한다. 이때 $n = 500, 1000$을 사용한다.

(4) 추출된 n개의 표본에서 $\pi_i = \exp(a_0 + a_1 y_i)$, $\pi_i \in [0, 1]$를 계산한다. y_i의 최솟값에서의 응답률을 $\pi_y^{\min}$, y_i의 최댓값에서의 응답률을 $\pi_y^{\max}$라 할 때, $(\pi_y^{\min}, \pi_y^{\max}) = (0.9, 0.7), (1, 1)$, 그리고 $(0.7, 0.9)$를 적용하여 지수 관계식에 포함된 모수 (a_0, a_1)을 계산한다. 즉 $(\pi_y^{\min}, \pi_y^{\max}) = (0.9, 0.7)$인 경우는 y_i 값이 최소인 경우 응답률이 0.9이고 이후 지수적으로 감소하여 최대의 y_i 값에 해당되는 응답률이 0.7이 되도록 한다. $(\pi_y^{\min}, \pi_y^{\max}) = (1, 1)$인 경우는 y_i 값에 무관하게 응답률이 모두 '1'이며 $(\pi_y^{\min}, \pi_y^{\max}) = (0.7, 0.9)$인 경우는 최소 y_i 값의 응답률이 0.7에서 시작한 후 지수적으로 증가하여 최대의 y_i 값에서는 0.9가 되도록 한다.

(5) 응답한 최종 조사 자료는 r개이다. 여기서 $(\pi_y^{\min}, \pi_y^{\max}) = (0.9, 0.7)$ 또는 $(\pi_y^{\min}, \pi_y^{\max}) = (0.7, 0.9)$인 경우는 약 $r = 400$ 또는 800이며 무응답이 없는 경우는 $r = n = 500$ 또는 1,000이 된다.

Step 3: 층화

얻어진 표본 자료는 (x_i, y_i), $i = 1, \ldots, r$이고 응답률에 따라 자료의 가중치는 달라진다. 이를 반영하기 위해 주어진 층을 L개의 세부 층으로 나눈다. 실제 자료 분석에서는 모집단에 보조변수 x_i의 정보만 있으므로 보조변수를 기준으로 층을 나눈다.

(6) 보조변수 x_i를 기준으로 등간격으로 모집단을 L개의 세부 층으로 나눈다. 여기서 $L = 10, 20, 30$을 사용한다.

Step 4: 모수추정

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

(8) $\log(1/w_i) = a_0 + a_1 y_i + \eta_i$를 설정하고 단순 회귀모형을 이용하여 모수 a_0, a_1을 추정한다.

(9) 추출된 자료 (y_i, x_i)에 단순 회귀분석을 실시하여 $\beta_0, \beta_1, \sigma^2$을 추정한다.

(10) 계산된 결과를 이용하여 식 (3.3)에서 식 (3.5)인 $\hat{\bar{Y}}_s, \hat{\bar{Y}}_{st}, \hat{\bar{Y}}_{\inf}$를 계산한다.

4.2. 모의실험 결과

3.3절에서 설명한 세 추정량의 성능을 다음의 비교통계량을 이용하여 비교하였다. 이때 사용된 반복수 $R = 3{,}000$이다.

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

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

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

각 반복에서 모집단 자료를 랜덤으로 새롭게 생성하여 특정 모집단에 의한 영향을 제거하였다. 또한 주어진 모집단과 표본에서 세부 층의 수를 10, 20, 그리고 30개로 나눈 후 각각 추정량을 계산하였다. 이는 세부 층의 수가 추정에 어떤 영향을 주는지 살펴보기 위해서이다. 따라서 세부 층 개수와는 무관하게 $\hat{\bar{Y}}_s$는 동일한 결과를 준다. 생성된 자료와 비교 추정량을 이용하여 다음의 결과가 얻어졌다.

4.2.1. 보조변수 x_i가 균등 분포를 따를 경우 Table 4.1에서 Table 4.4에 결과를 수록하였다. Table 4.1의 결과를 살펴보면 모든 비교 통계량에서 단순 평균을 사용한 $\hat{\bar{Y}}_s$에 비해 $\hat{\bar{Y}}_{st}$가 우수한 결과를 준다. 이는 세부 층의 평균이 관심변수 y_i에 따라 달라지기 때문에 하나의 층을 세부 층으로 나누어 추정하는 것이 타당하기 때문이다.

다음으로 본 연구에서 제안한 추정량인 $\hat{\bar{Y}}_{\inf}$와 층화추출법을 이용하여 얻어진 $\hat{\bar{Y}}_{st}$ 결과를 살펴보자. 먼저 Table 4.1 결과에서 편향(bias)을 살펴보면 무응답이 발생하지 않은 경우는 $\hat{\bar{Y}}_{st}$가 우수하다. 이는 잘

Table 4.1. Comparison results of $U(100, 200)$ with $n = 500$ and $\sigma^2 = 400$

$\pi_y^{\min}$	$\pi_y^{\max}$	r	L	Bias			Abias			RMSE		
				$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{\inf}$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{\inf}$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{\inf}$
0.9	0.7	400	10	−7.432	−0.225	−0.092	8.551	1.010	0.991	10.319	1.273	1.251
			20	−7.432	−0.159	−0.028	8.551	0.870	0.846	10.319	1.098	1.071
			30	−7.432	−0.147	−0.018	8.551	0.861	0.826	10.319	1.088	1.042
1.0	1.0	500	10	0.134	0.006	0.006	5.051	0.873	0.868	6.293	1.101	1.096
			20	0.134	0.016	0.019	5.051	0.754	0.742	6.293	0.950	0.937
			30	0.134	0.016	0.019	5.051	0.736	0.719	6.293	0.928	0.904
0.7	0.9	400	10	7.796	0.213	0.070	8.797	1.009	0.993	10.565	1.271	1.249
			20	7.796	0.162	0.021	8.797	0.869	0.847	10.565	1.097	1.068
			30	7.796	0.154	0.016	8.797	0.864	0.824	10.565	1.089	1.039

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

Table 4.2. Comparison results of $U(100, 200)$ with $n = 500$ and $\sigma^2 = 900$

$\pi_y^{\min}$	$\pi_y^{\max}$	r	L	Bias			Abias			RMSE		
				$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{\inf}$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{\inf}$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{\inf}$
0.9	0.7	400	10	−7.629	−0.373	−0.073	8.642	1.339	1.320	10.449	1.685	1.663
			20	−7.629	−0.307	−0.014	8.642	1.248	1.230	10.449	1.568	1.541
			30	−7.629	−0.322	−0.030	8.642	1.259	1.219	10.449	1.720	1.676
1.0	1.0	500	10	0.139	0.035	0.033	5.109	1.171	1.183	6.365	1.459	1.476
			20	0.139	0.036	0.033	5.109	1.091	1.103	6.365	1.357	1.370
			30	0.139	0.023	0.026	5.109	1.081	1.084	6.365	1.345	1.340
0.7	0.9	400	10	8.019	0.412	0.100	8.964	1.349	1.321	10.712	1.695	1.661
			20	8.019	0.370	0.055	8.964	1.258	1.231	10.712	1.581	1.540
			30	8.019	0.340	0.038	8.964	1.247	1.207	10.712	1.566	1.508

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

알려진 것처럼 $\hat{\bar{Y}}_{st}$가 불편추정량이기 때문이다. 반면 $\hat{\bar{Y}}_{\inf}$의 경우는 지수형 표본 포함확률에 포함된 모수 a_1의 추정값을 이용하여 얻어지기 때문에 두 결과에는 차이가 난다. 그러나 $\hat{\bar{Y}}_{\inf}$과 $\hat{\bar{Y}}_{st}$의 차이는 매우 미미하고 편향의 절대값이 모두 '0'에 가깝기 때문에 두 추정량 모두 불편 추정량이라 판단할 수 있다. 반면 무응답이 발생한 경우에는 편향이 발생하게 되며 이 경우에는 $\hat{\bar{Y}}_{\inf}$가 $\hat{\bar{Y}}_{st}$에 비해 우수한 결과를 주고 있다.

결론적으로 큰 y_i값에서 무응답이 많이 발생하게 되면 과소 추정이 일어나고 반대로 작은 y_i값에서 무응답이 많이 발생하게 되면 과대 추정이 일어나지만 이 문제는 주어진 층을 세부 층으로 나누어 줌으로써 일정 수준까지는 해결이 가능한 것으로 판단된다. 다만 본 연구에서 제안한 $\hat{\bar{Y}}_{\inf}$를 사용하게 되면 편향을 더욱 줄일 수 있다.

절대편향과 RMSE 결과를 비교해 보면 모든 경우에서 정보적 표본설계 기법을 사용한 $\hat{\bar{Y}}_{\inf}$가 우수한 결과를 준다. 이는 과소 추정이 된 경우에는 추정 값을 크게 하고 과대 추정이 된 경우에는 추정 값을 작게 해 주는 정보적 표본설계의 특성 상 당연한 결과라 판단된다.

다음으로 당연한 결과이지만 무응답 발생으로 인해 표본 수가 줄어들고 편향도 발생하기 때문에 $(\pi_y^{\min}, \pi_y^{\max}) = (1, 1)$ 결과가 가장 우수하다.

또한 세부 층의 수에 따른 절대편향과 RMSE의 결과를 살펴보면 세부 층의 개수가 증가할수록 추정의 정확성 및 정밀성이 향상되는 것을 확인할 수 있다. 물론 세부 층 정보를 사용하지 않는 $\hat{\bar{Y}}_s$는 일정한 값

Table 4.3. Comparison results of $U(100, 200)$ with n = 1,000 and $\sigma^2 = 400$

$\pi_y^{\min}$	$\pi_y^{\max}$	r	L	Bias			Abias			RMSE		
				$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{\text{inf}}$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{\text{inf}}$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{\text{inf}}$
0.9	0.7	800	10	−7.534	−0.214	−0.080	7.817	0.692	0.675	8.997	0.874	0.851
			20	−7.534	−0.153	−0.019	7.817	0.595	0.577	8.997	0.747	0.725
			30	−7.534	−0.137	−0.003	7.817	0.584	0.565	8.997	0.735	0.710
1	1	1000	10	0.116	0.001	−0.003	3.496	0.592	0.592	4.346	0.742	0.741
			20	0.116	0.004	−0.001	3.496	0.511	0.510	4.346	0.641	0.638
			30	0.116	0.010	0.007	3.496	0.502	0.499	4.346	0.630	0.624
0.7	0.9	800	10	7.786	0.222	0.077	8.012	0.695	0.673	9.200	0.869	0.843
			20	7.786	0.171	0.025	8.012	0.598	0.575	9.200	0.747	0.722
			30	7.786	0.168	0.026	8.012	0.592	0.566	9.200	0.738	0.708

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

Table 4.4. Comparison results of $U(100, 200)$ with n = 1,000 and $\sigma^2 = 900$

$\pi_y^{\min}$	$\pi_y^{\max}$	r	L	Bias			Abias			RMSE		
				$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{\text{inf}}$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{\text{inf}}$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{\text{inf}}$
0.9	0.7	800	10	−7.698	−0.398	−0.100	7.991	0.962	0.921	9.175	1.203	1.150
			20	−7.698	−0.347	−0.049	7.991	0.891	0.849	9.175	1.116	1.067
			30	−7.698	−0.338	−0.039	7.991	0.874	0.829	9.175	1.103	1.046
1.0	1.0	1000	10	0.119	0.009	0.002	3.538	0.795	0.808	4.398	0.992	1.006
			20	0.119	0.005	−0.001	3.538	0.745	0.754	4.398	0.929	0.939
			30	0.119	0.003	−0.003	3.538	0.734	0.738	4.398	0.919	0.923
0.7	0.9	800	10	7.976	0.409	0.094	8.202	0.980	0.933	9.440	1.218	1.159
			20	7.976	0.351	0.035	8.202	0.895	0.859	9.440	1.126	1.076
			30	7.976	0.338	0.024	8.202	0.888	0.839	9.440	1.112	1.053

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

을 유지한다. 이러한 결과는 오차의 분산인 σ^2을 400에서 900으로 크게 한 경우에도 같은 결과를 주며 자료 수 n = 500에서 n = 1,000으로 크게 했을 경우의 결과인 Table 4.3과 Table 4.4에서도 확인할 수 있다. 반면 오차의 분산인 σ^2이 900인 Table 4.2와 Table 4.4의 결과를 보면 무응답이 발생하지 않는 경우 절대편향과 RMSE에서 $\hat{\bar{Y}}_{\text{inf}}$가 $\hat{\bar{Y}}_{st}$에 비해 우수하지 않다. 그러나 무응답이 없는 경우에는 $\hat{\bar{Y}}_{\text{inf}}$를 사용하지 않아도 되므로 큰 문제는 없다. 결론적으로 무응답이 발생한 경우 충분한 개수의 세부 층으로 층을 나눈 후 본 연구에서 제안한 방법을 사용하게 되면 매우 우수한 결과를 얻을 수 있다고 판단된다.

4.2.2. 보조변수 x_i가 절단 감마분포를 따를 경우 사업체 조사에서 종사자 수가 층화변수로 사용될 경우 사업체 규모로 층이 나누어지며 각 층은 대부분 종사자 수가 커질수록 사업체 수가 감소한다. 따라서 일부 층에서는 절단 감마분포를 따르는 모집단 형태가 현실적으로 타당할 수 있다. 이를 반영하기 위해 절단 감마분포를 사용한 모의실험이 수행되었으며 결과는 Table 4.5에서 Table 4.8에 수록되었다.

Table 4.5 결과를 살펴보면 전체적으로 보조변수가 균등분포일 때와 유사한 경향을 보이고 있다. 편향(bias)을 살펴보면 무응답이 없는 경우와 무응답이 발생한 경우 모두 $\hat{\bar{Y}}_{\text{inf}}$가 우수한 결과를 준다. 또한 절대편향과 RMSE 결과를 비교해 보면 모든 경우에서 정보적 표본설계 기법을 사용한 $\hat{\bar{Y}}_{\text{inf}}$가 가장 우수한 결과를 준다. 그러나 전체적으로 Table 4.6에서 Table 4.8의 결과는 Table 4.1에서 Table 4.4와 유사하다.

Table 4.5. Comparison results of T-Gam$(1, 100)$ with $n = 500$ and $\sigma^2 = 400$

$\pi_y^{\min}$	$\pi_y^{\max}$	r	L	Bias			Abias			RMSE		
				$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{\text{inf}}$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{\text{inf}}$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{\text{inf}}$
0.9	0.7	400	10	−7.109	−0.202	−0.074	8.200	1.004	0.994	9.851	1.256	1.241
			20	−7.109	−0.154	−0.037	8.200	0.857	0.842	9.851	1.078	1.055
			30	−7.109	−0.168	−0.060	8.200	0.875	0.841	9.851	1.414	1.371
1.0	1.0	500	10	0.009	0.012	0.005	4.908	0.894	0.892	6.131	1.114	1.111
			20	0.009	0.009	−0.008	4.908	0.765	0.759	6.131	0.964	0.951
			30	0.009	0.014	−0.009	4.908	0.760	0.737	6.131	0.943	0.917
0.7	0.9	400	10	7.413	0.224	0.077	8.513	1.025	1.013	10.299	1.283	1.265
			20	7.413	0.165	0.006	8.513	0.876	0.858	10.299	1.099	1.074
			30	7.413	0.155	− 0.012	8.513	0.880	0.847	10.299	1.251	1.213

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

Table 4.6. Comparison results of T-Gam$(1, 100)$ with $n = 500$ and $\sigma^2 = 900$

$\pi_y^{\min}$	$\pi_y^{\max}$	r	L	Bias			Abias			RMSE		
				$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{\text{inf}}$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{\text{inf}}$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{\text{inf}}$
0.9	0.7	400	10	−7.349	−0.411	−0.122	8.370	1.364	1.342	10.087	1.720	1.689
			20	−7.349	−0.325	−0.061	8.370	1.274	1.246	10.087	1.600	1.565
			30	−7.349	−0.352	−0.100	8.370	1.306	1.247	10.087	1.777	1.723
1.0	1.0	500	10	0.011	−0.018	−0.036	4.970	1.184	1.197	6.209	1.484	1.501
			20	0.011	0.007	−0.030	4.970	1.108	1.114	6.209	1.389	1.393
			30	0.011	−0.010	−0.051	4.970	1.110	1.101	6.209	1.385	1.372
0.7	0.9	400	10	7.541	0.387	0.369	8.565	1.364	1.343	10.373	1.715	1.689
			20	7.541	0.356	0.017	8.565	1.287	1.255	10.373	1.609	1.567
			30	7.541	0.325	−0.019	8.565	1.296	1.243	10.373	1.656	1.590

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

Table 4.7. Comparison results of T-Gam$(1, 100)$ with $n = 1{,}000$ and $\sigma^2 = 400$

$\pi_y^{\min}$	$\pi_y^{\max}$	r	L	Bias			Abias			RMSE		
				$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{\text{inf}}$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{\text{inf}}$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{\text{inf}}$
0.9	0.7	800	10	−7.088	−0.206	−0.072	7.363	0.683	0.664	8.536	0.856	0.833
			20	−7.088	−0.159	−0.030	7.363	0.593	0.574	8.536	0.739	0.718
			30	−7.088	−0.152	−0.027	7.363	0.574	0.557	8.536	0.720	0.699
1.0	1.0	1000	10	0.088	0.004	0.001	3.373	0.590	0.587	4.195	0.737	0.735
			20	0.088	−0.004	−0.012	3.373	0.512	0.507	4.195	0.639	0.635
			30	0.088	−0.007	−0.018	3.373	0.497	0.495	4.195	0.626	0.621
0.7	0.9	800	10	7.489	0.236	0.091	7.750	0.699	0.678	8.949	0.882	0.853
			20	7.489	0.167	0.017	7.750	0.604	0.588	8.949	0.761	0.737
			30	7.489	0.153	−0.001	7.750	0.586	0.570	8.949	0.737	0.715

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

여기서 세부 층의 개수에 따른 RMSE 결과를 살펴보면 무응답이 발생한 경우에서 $n = 500$일 때 $L = 20$에서 가장 우수한 결과를 주고 있다. 반면 $n = 1{,}000$인 경우에는 $L = 30$에서 가장 우수한 결과를 준다. 따라서 자료 수에 따라 또는 보조변수의 분포 형태에 따라 최적의 세부 층 개수가 결정될 수 있음을 확인할 수 있다.

Table 4.8. Comparison results of T-Gam$(1, 100)$ with $n = 1{,}000$ and $\sigma^2 = 900$

$\pi_y^{\min}$	$\pi_y^{\max}$	r	L	Bias			Abias			RMSE		
				$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{\text{inf}}$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{\text{inf}}$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{\text{inf}}$
0.9	0.7	800	10	−7.290	−0.395	−0.105	7.537	0.967	0.915	8.705	1.206	1.145
			20	−7.290	−0.337	−0.053	7.537	0.881	0.837	8.705	1.106	1.052
			30	−7.290	−0.319	−0.050	7.537	0.873	0.827	8.705	1.094	1.040
1.0	1.0	1000	10	0.089	0.002	−0.009	3.415	0.808	0.808	4.246	1.006	1.007
			20	0.089	0.001	−0.015	3.415	0.748	0.749	4.246	0.938	0.935
			30	0.089	0.007	−0.019	3.415	0.743	0.742	4.246	0.931	0.927
0.7	0.9	800	10	7.656	0.405	0.084	7.916	0.981	0.933	9.115	1.235	1.169
			20	7.656	0.349	0.021	7.916	0.903	0.856	9.115	1.133	1.076
			30	7.656	0.348	0.009	7.916	0.897	0.847	9.115	1.125	1.064

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

5. 결론

본 논문에서는 무응답이 있는 표본조사에서 모수추정의 정확성 및 정밀성 향상을 위해 주어진 하나의 층을 세부 층으로 나누어 가중치를 새롭게 정의한 후 추정하는 방법과 정보적 표본설계 기법을 적용하는 방법을 제안하였다. 특히 무응답으로 인해 자료의 응답률이 관심변수 자료 값의 지수함수가 될 경우에는 모수추정 결과가 과대 또는 과소 추정되며 이때 본 연구에서 제안한 방법을 사용하게 되면 과대 또는 과소 추정의 영향을 줄일 수 있게 된다.

모의실험을 통해 각 추정량의 성능을 비교한 결과 정보적 표본설계 기법을 적용할 경우 모든 비교 추정량에서 우수한 결과가 얻어졌다. 다만 본 연구에서 사용한 가중치는 관심변수를 직접 이용하여 얻지 않고 보조변수를 이용하여 얻은 결과를 사용하였다. 따라서 정확히 정보적 표본설계 방법을 적용한 것은 아니다. 그러나 본 연구에서 사용한 방법이 현실적으로 사용 가능한 방법이며 응답률이 관심변수 자료 값의 지수함수를 따르는 경우에는 본 연구에서 제안한 방법을 적용하여 모수를 추정하는 것이 타당하다. 다만 세부 층을 나누는 방법과 세부 층의 수를 최적으로 정하는 방법에 관한 추가적인 연구가 필요하고 판단된다.

References

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

Lee, K. J., Kim, H. W., Kim, S. J., Kim, K. M., and Lee, Y. H. (2008). Survey design of the workplace panel survey in Korea, *Survey Research*, **9**, 71–91.

Lee, Y. and Shin, K.-I. (2017). A study on sensitivity of representativeness indicator in survey sampling, *The Korean Journal of Applied Statistics*, **30**, 69–82.

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, *2003 Joint Statistical Meeting-Section on Survey Research Methods*, 3284–3295.

Savitsky, T. D. and Toth, D. (2016). Bayesian estimation under informative sampling, *Electronic Journal of Statistics*, **30**, 1677–1708.

Schouten, B., Cobben, F., and Bethlehem, J. (2009). Indicators for the representativeness of survey response, *Survey Methodology*, **35**, 101–113.

응답률이 관심변수의 지수함수를 따를 경우 정보적 표본설계 기법을 이용한 모수추정

정희영[a] · 신기일[a,1]

[a]한국외국어대학교 통계학과

(2017년 10월 17일 접수, 2017년 11월 24일 수정, 2017년 12월 4일 채택)

요 약

표본조사에서는 추정의 정확성 및 정밀성 향상을 위해 흔히 층화추출법을 사용하며 층 내에서는 동일한 표본 가중치를 이용하여 표본을 추출한다. 그러나 실제 응답률은 관심변수 값에 영향을 받을 수 있기 때문에 주어진 동일한 가중치는 응답률을 반영하여 보정되어야 한다. 또한 관심변수가 연속형 보조변수와 선형 관계가 있고 보조변수를 기준으로 층이 나누어진 경우에는 층 내에서 동일한 가중치를 사용하는 것 보다 층을 세분화한 후 얻어진 가중치를 사용하는 것이 효과적일 수 있다. 본 연구에서는 응답률이 관심변수 자료 값의 지수함수이고, 관심변수가 보조변수와 선형 관계가 있을 때 정보적 표본설계 기법을 이용하여 추정의 정확성과 정밀성을 높이는 방법을 제안하였다. 또한 모의실험을 통하여 제안된 방법의 우수성을 확인하였다.

주요용어: 층화추출법, 가중치 보정, 지수 포함확률, 초모집단모형

이 연구는 2017년 한국외국어대학교 교내연구비 지원을 받아 수행되었음.

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