The Korean Journal of Applied Statistics (2016) DOI: http://dx.doi.org/10.5351/KJAS.2016.29.7.1185
29(7), 1185–1199

An outlier weight adjustment using generalized ratio-cum-product method for two phase sampling

Jung-Taek Oh[a] · Key-Il Shin[a,1]

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

(Received June 22, 2016; Revised August 9, 2016; Accepted September 11, 2016)

Abstract

Two phase sampling (double sampling) is often used when there is inadequate population information for proper stratification. Many recent papers have been devoted to the estimation method to improve the precision of the estimator using first phase information. In this study we suggested outlier weight adjustment methods to improve estimation precision based on the weight of the generalized ratio-cum-product estimator. Small simulation studies are conducted to compare the suggested methods and the usual method. Real data analysis is also performed.

Keywords: outlier detection, ratio estimator, product estimator, MSE

1. 서론

정확한 표본 조사를 위해 많은 이론과 방법이 개발되었다. 그러나 우수한 이론과 방법에 우선하는 것이 표본설계에 필요한 정보의 양이다. 기본적으로 정확한 표본 조사를 위해서는 표본 추출틀이 갖고 있는 정보의 양이 충분해야 한다. 최근 산업구조의 빠른 변화와 산업의 융합으로 인해 대표적인 모집단 층화 정보인 정확한 산업분류를 갖고 있는 표본틀을 사용하는 경우는 흔치 않다. 이러한 정보부족 문제를 해결하여 추정의 정확성을 향상시킬 수 있는 표본설계 기법이 이중추출법(이상추출법)이다. 이중추출법에 관한 내용은 Cochran (1977)을 살펴보기 바라며 이중추출법에서 사용하는 대표적인 용어인 first phase sampling을 1차 조사, second phase sampling을 2차 조사로 사용하였다.

본 논문에서는 이중추출법의 2차 조사에서 발생한 이상점 처리 방법에 대하여 연구하였다. 흔히 1차 조사에서 관심변수와 관계가 높은 보조 변수가, 2차 조사에서 관심변수를 조사하게 된다. 따라서 2차 조사에서 얻어진 관심변수에서 발생한 이상점 처리는 모수 추정의 정확성 향상을 위해 매우 중요하다.

이중추출법에 관한 많은 연구가 매우 활발히 진행되고 있다. 먼저 Fuller (2000)는 이중추출에서의 회귀추정을 연구하였다. 또한 Hidiroglou (2001)는 내포 이중추출법(nested two phase sampling)과 비내포 이중추출법(non-nested two phase sampling)에서의 회귀추정량을 연구하였다.

이와 같이 보조변수를 이용한 추정량의 정밀성 향상과 관련된 여러 연구가 진행되었다. Hidiroglou와 Sandal (1998)은 이중추출법에서 보조 변수를 이용하여 추정량의 정확성을 높이는 방법으로 일반화최

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2014R1A1A2056857).
[1]Corresponding author: Department of statistics, Hankuk University of Foreign Studies, 81, Oedae-ro, Mohyeon-myeon, Cheoin-gu, Yongin-si, Gyeonggi-do 17035, Korea. E-mail: keyshin@hufs.ac.kr

소제곱거리를 이용한 캘리브레이션 방법을 제안하였다. 이 캘리브레이션 방법은 얻어진 보조정보를 이용하여 가중치를 보정함으로써 추정의 정밀성을 향상시키는 방법이다. 이후 Wu와 Sitter (2001)는 완전 보조정보를 이용하여 일반화회귀모형(generalized regressive model)을 기반으로 한 캘리브레이션 방법에 관하여 연구하였다. Wu와 Luan (2003)에서는 이중추출을 위한 최적 캘리브레이션 추정량이 연구되었고 이후 Koyuncu와 Kadilar (2009)에서는 두 개의 보조정보가 있을 때 비추정(ratio estimator), 곱추정(product estimator) 그리고 비추정과 곱추정의 곱으로 얻어지는 추정량을 제안하였으며 이 추정량의 특징을 연구하였다. 이후 Singh 등 (2010)은 이중추출법에서 사용할 수 있는 추정량을 제안하였으며 이 논문에서는 캘리브레이션 기법을 통하여 얻어지는 비추정과 회귀추정(regression estimator) 그리고 곱추정이 연구되었다. 이러한 비추정과 곱추정은 2차 조사에서 얻어진 보조변수의 총합 또는 평균이 되도록 만들어 주는 캘리브레이션 추정량(calibration estimator)이 된다. 최근 Tailor 등 (2014, 2015) 이중추출법에서 사용할 수 있는 비추정과 회귀추정 형태의 지수 추정량과 일반화 ratio-cum-product 형태의 추정량을 연구하였다.

이렇게 이중추출법에서 사용하는 추정량의 정확성 향상을 위한 많은 연구가 수행되었고 Singh와 Kumar (2010)과 Singh 등 (2010)은 이중추출법의 2차 표본조사에서 발생한 무응답 처리에 관한 연구를 수행했음에도 불구하고, 이중추출법에서 발생하는 이상점 처리에 관한 연구는 미미한 상태이다.

이에 본 연구에서는 이중추출법에 적용되는 이상점 처리에 관해 연구하였다. 이상점은 대부분의 표본조사에서 발생하게 되며 이중추출법을 사용할 경우에도 이상점은 발생하게 된다. 이상점을 탐지하기 위해 사용되는 방법은 여러 가지가 있으나 그 성능을 좌우하는 것은 관심변수와 관련된 보조변수 정보의 양이다. 따라서 이중추출과 같이 1차 조사에서 다양한 보조변수가 구해지고 많은 양의 보조 정보가 얻어진 경우에는 이 정보를 사용하는 방법에 따라 그 성능이 달라질 수 있다. 이상점 처리에 관한 내용은 먼저 Chamber과 Ren (2004)은 이상점인 경우 이상점을 탐지한 후 이상점을 신뢰구간의 상한 또는 하한으로 대체하는 방법과 랜덤으로 대체하는 방법을 제안하였다. 이후 Kim과 Shin (2013)에서는 이상점을 외표준화잔차를 이용하여 탐지한 후 이상점의 가중치를 보정하는 방법을 제안하였다. 최근 이상점 탐지를 위해 She와 Owen (2011)은 Θ-IPOD 방법을 제안하였으며 Kim과 Shin (2014)은 이 방법을 적용하여 이상점과 무응답이 동시에 있는 표본 조사에서의 무응답 대체법을 연구하였다. 본 연구에서는 이상점 탐지법을 연구하는 것이 아니라 탐지된 이상점을 처리하는 방법을 연구하는 것이 목적이므로 이 방법을 사용하지 않고 쉽게 사용할 수 있는 이상점 탐지법인 외표준화잔차법을 사용하여 이상점을 탐지하였다. 흔히 표본조사에서는 탐지된 이상점의 가중치를 “0” 또는 “1”로 주는데 본 연구에서는 이중추출법의 1차 조사에서 얻어진 정보를 캘리브레이션하여 가중치를 새롭게 보정하는 방법을 제안하였다. 결론적으로 본 연구에서는 이중추출법의 2차 조사에서 발생한 이상점 처리를 위해 2차 조사에서 탐지된 이상점의 가중치를 보정하여 ratio-cum-product 추정량의 성능을 향상시키는 방법을 제안하였다.

본 논문의 구성은 다음과 같다. 먼저 2절에서 이중추출법과 외표준화잔차를 이용한 이상점 탐지법에 대해 설명하였다. 다음으로 3절에서는 본 연구에서 제안한 이상점 처리법을 설명하였다. 4절에서는 모의실험이 수행되었으며 5절에서는 실제 자료 분석이 수행되었다. 6절에 결론이 있다.

2. 이중추출법과 이상점 탐지법

2.1. 이중추출법

조사 목적에 따른 관심변수 Y를 직접 조사하는 것은 비용이 많이 들지만 관심변수 Y와 상관이 높은 보조변수 X를 조사하는 것은 비용이 적게 드는 경우, 1차로 큰 규모의 표본을 추출하여 보조변수 X를 조사하고, 얻어진 보조변수의 정보를 기초로 층화한 후에 2차로 1차 자료의 각 층에서 표본을 추출하여 관

심변수 Y를 조사하는 것을 이중추출법(two phase sampling, double sampling)이라 한다. 이에 관한 내용은 Cochran (1977), Fuller (2000) 그리고 Hidiroglou (2001)을 살펴보기 바란다. 최근 이중추출법과 관련된 논문은 1차 조사에서 얻어진 정보를 이용하여 추정의 정확성을 향상시키는데 초점을 맞추고 있다. 먼저 Hidiroglou와 Sandal (1998)은 이중추출법에서 보조정보를 사용하여 캘리브레이션 방법으로 가중치를 보정하는 방법을 제안하였다. 이후 보조 정보를 활용하여 모형-캘리브레이션 방법을 사용하는 방법이 Wu와 Sitter (2001)에서 연구되었으며 이 방법은 이후 Wu와 Luan (2003)에서 이중추출법에 적용되었다. 최근 Singh 등 (2010), Tailor 등 (2014, 2015)은 이중 추출법에서 사용 가능한 방법인 비추정량과 곱추정량 그리고 이를 결합하여 만든 일반화 ratio-cum-product 추정량을 연구하였으며 ratio-cum-product 추정량의 정의는 다음과 같다.

$$\hat{\bar{Y}}_{ds}^{(\alpha,\beta)} = \bar{y}_2 \left(\frac{\bar{x}_1}{\bar{x}_2}\right)^{\alpha} \left(\frac{\bar{z}_2}{\bar{z}_1}\right)^{\beta},$$

여기서 $\bar{x}_1 = (1/n_{1h}) \sum_{i=1}^{n_{1h}} x_{1hi}$, $\bar{z}_{1h} = (1/n_{1h}) \sum_{i=1}^{n_{1h}} z_{1hi}$는 1차 조사 보조 변수들의 평균 추정값이며 $\bar{x}_2 = (1/n_{2h}) \sum_{i=1}^{n_{2h}} x_{2hi}$, $\bar{z}_{2h} = (1/n_{2h}) \sum_{i=1}^{n_{2h}} z_{2hi}$는 2차 조사 보조 변수들의 평균 추정값이다.

이 추정량의 편향과 분산은 Tailor 등 (2015)를 참조하기 바란다. 이제 $\alpha = 1$, $\beta = 0$인 경우, 즉 $\hat{\bar{Y}}_{ds}^{(1,0)} = \bar{y}_2(\bar{x}_1/\bar{x}_2)$인 경우가 흔히 이중추출법에서 사용하는 비추정량이 되고, $\alpha = 0$, $\beta = 1$인 $\hat{\bar{Y}}_{ds}^{(0,1)} = \bar{y}_2(\bar{z}_2/\bar{z}_1)$인 경우가 곱추정량이 된다. 만약 두 독립변수가 모두 있고, 종속변수와 독립변수의 관계가 비추정과 곱추정 사용에 적합한 경우에는 위의 식을 사용할 수 있다.

2.2. 이상점탐지법

본 연구에서는 회귀모형에서 사용되는 이상점 탐지법 중에서 외표준화잔차(studentized deleted residual)를 기준으로 한 방법을 이상점 탐지법으로 사용하였다. Kim과 Shin (2014)은 무응답 대체를 위해 Θ-IPOD 방법을 사용한 이상점 탐지법을 사용하였지만 본 연구는 이상점을 탐지하는 것이 주된 목적이 아니라 탐지된 이상점을 적절히 처리하는 것이 목적이므로 SAS에서 쉽게 사용할 수 있는 방법인 외표준화잔차를 이용하였다. 다음이 외표준화잔차(externally studentized residual, studentized deleted residual)의 정의이다.

$$t_i = \frac{r_i}{s(d_i)} = \frac{r_i}{s(i)\sqrt{(1-h_{ii})}},$$

여기서 $r_i = y_i - \hat{y}_{(i)}$, y_i는 관측값이고 $\hat{y}_{(i)}$는 i번째 관측값을 제거한 후에 얻어진 예측값을 의미한다. 또한 h_{ii}는 지렛값 또는 레버리지이다. t_i의 분포는 우리가 잘 알고 있는 자유도 $(n-p-1)$인 t-분포를 따르는 것으로 알려져 있다. 또한 외표준화잔차는 SAS/Proc REG의 출력결과(Rstudent)에서 쉽게 얻을 수 있는 결과이다.

3. 제안된 가중치 보정법

가중치 보정방법은 무응답과 이상점의 영향력을 줄이고 벤치마킹을 이용하기 위해 흔히 사용된다. 일반적인 방법은 표본설계 시에 정해진 설계 가중치에 각각의 요인에 해당되는 보정인자를 구한 후 이 보정인자를 곱하여 최종 가중치를 얻는다. 본 연구에서는 무응답 보정과 벤치마킹 보정을 고려하지 않고 다만 이상점 보정만을 고려한다. 따라서 최종 가중치 w^f는 설계 가중치를 w라 하고 이상점 보정인자를 f라 하였을 때 $w^f = w \times f$로 정해진다. 이제 n개의 자료에서 k개의 이상점이 존재할 경우에 흔히 사

용하는 가중치 보정방법을 설명하면 다음과 같다. 먼저 이상점인 경우 이상점 보정인자 $f = 0$으로 한다. 따라서 이상점의 최종 가중치 $w^f = 0$이 된다. 결국 이상점인 경우 $w^f = 0$이고 정상자료인 경우 $w^f = w(n/(n-k))$이 된다. 또한 이와 유사한 방법으로 이상점인 경우에는 $w^f = 1$을 사용하고 정상자료인 경우에는 $w^f = w(1 + \{k(w-1)\}/\{w(n-k)\})$을 사용한다. 이 방법들은 이미 여러 논문에서 사용되고 있다.

본 연구에서는 비추정 캘리브레이션을 이용한 이상점 처리 방법을 제안하였다. 먼저 Kim과 Shin (2014)에서는 이상점으로 탐지된 경우, 이상점을 제거하는 대신 이상점의 가중치를 "1"로 보정하여 사용할 것을 제안하였다. 본 연구에서는 이중추출법이 사용되기 때문에 이 방법을 확장할 수 있다. 1차 조사에서 두 보조변수의 평균 $\bar{x}_1$과 $\bar{z}_1$이 구해지고, 2차 조사에서 $\bar{x}_2$와 $\bar{z}_2$가 구해지면 ratio-cum-product 추정량은 이중추출 추정량을 1차 조사 추정값으로 캘리브레이션 해주는 방법이다. 따라서 ratio-cum-product 추정량의 가중치를 기본으로 이상점인 경우에는 가중치를 "1"로 하고 정상 자료인 경우에는 이상점이 갖고 있는 나머지 가중치를 분배하는 방법을 이용하여 가중치를 보정할 수 있다. 본 연구에서는 실제 표본설계에서 사용하는 층화이중추출법을 고려하였고 따라서 가중치 보정도 층별로 이루어진다.

3.1. 이상점 처리전 가중치

이상점을 처리하지 않고, 기존의 방법인 ratio-cum-product 추정량의 가중치를 사용한다.

방법 0:

다음의 층별 가중치는 ratio-cum-product 추정량의 가중치이다.

$$w_h^{f(0)} = w_{1h} w_{2h} \left(\frac{\bar{x}_{1h}}{\bar{x}_{2h}}\right)^{\alpha} \left(\frac{\bar{z}_{2h}}{\bar{z}_{1h}}\right)^{\beta},$$

여기서 $w_{1h} = N_{1h}/n_{1h}$, $w_{2h} = N_{1h}/n_{2h}$는 h층의 1차 설계 가중치와 2차 설계 가중치이며 $\bar{x}_{1h}, \bar{x}_{2h}$, $\bar{z}_{1h}, \bar{z}_{2h}$는 각각 1차 조사와 2차 조사의 h층의 평균 추정량이다. 따라서 ratio-cum-product 추정량은 $\hat{\bar{Y}}_{ds}^{\alpha,\beta(0)} = (1/N_{1h}) \sum_{i=1}^{n_{2h}} w_h^{f(0)} y_{2hi}$이 된다. 모의실험에서는 이 결과를 M_0로 표시하였다.

3.2. 제안된 보정 가중치

다음의 방법 1에서 방법 4는 본 연구에서 제안한 방법이다. 이 방법은 이중추출에서 얻을 수 있는 보조정보를 이용하여 가중치를 캘리브레이션 방법으로 보정하는 방법이다.

방법 1:

2차 조사 자료에서 탐지된 이상점의 가중치를 "1"로 한다. 이상점의 나머지 가중치는 정상 자료에 나누어 주는 보정 가중치를 사용한다.

- 이상점인 경우:

$$w_h^{f(1)} = w_{1h} \times w_{2h} \times \frac{1}{w_{2h}} = w_{1h}.$$

- 정상자료인 경우:

$$w_h^{f(1)} = w_{1h} \times w_{2h} \left(1 + \frac{k_{2h}(w_h - 1)}{w_h(n_{2h} - k_{2h})}\right) \left(\frac{\bar{x}_{1h}}{\bar{x}_{2h}}\right)^{\alpha} \left(\frac{\bar{z}_{2h}}{\bar{z}_{1h}}\right)^{\beta}.$$

따라서 ratio-cum-product 추정량은 $\hat{\bar{Y}}_{ds}^{\alpha,\beta(1)} = (1/N_{1h})\sum_{i=1}^{n_{2h}} w_h^{f(1)} y_{2hi}$이 된다. 모의실험에서는 이 결과를 M_1로 표시하였다.

방법 2:

2차 조사 자료에서 탐지된 이상점의 가중치를 "1"로 하면서 캘리브레이션 방법을 적용한다. 이상점의 나머지 가중치는 정상 자료에 나누어 주는 보정 가중치를 사용한다.

- 이상점인 경우:

$$w_h^{f(2)} = w_{1h} \times w_{2h} \times \frac{1}{w_{2h}} \times \left(\frac{\bar{x}_{1h}}{\bar{x}_{2h}}\right)^{\alpha} \left(\frac{\bar{z}_{2h}}{\bar{z}_{1h}}\right)^{\beta} = w_{1h} \times \left(\frac{\bar{x}_{1h}}{\bar{x}_{2h}}\right)^{\alpha} \left(\frac{\bar{z}_{2h}}{\bar{z}_{1h}}\right)^{\beta}.$$

- 정상자료인 경우:

$$w_h^{f(2)} = w_{1h} \times w_{2h} \left(1 + \frac{k_{2h}(w_h - 1)}{w_h(n_{2h} - k_{2h})}\right) \left(\frac{\bar{x}_{1h}}{\bar{x}_{2h}}\right)^{\alpha} \left(\frac{\bar{z}_{2h}}{\bar{z}_{1h}}\right)^{\beta}.$$

따라서 ratio-cum-product 추정량은 $\hat{\bar{Y}}_{ds}^{\alpha,\beta(2)} = (1/N_{1h})\sum_{i=1}^{n_{2h}} w_h^{f(2)} y_{2hi}$이 된다. 모의실험에서는 이 결과를 M_2로 표시하였다.

방법 3:

2차 조사 자료에서 탐지된 이상점의 가중치를 "1"로 한다. 이상점의 나머지 가중치는 정상 자료에 나누어 주는 보정 가중치를 사용한다. 다만 이상점으로 탐지된 자료인 x_{2hi}^{out}, z_{2hi}^{out}을 평균 추정에서 제외시킨다.

- 이상점인 경우:

$$w_h^{f(3)} = w_{1h} \times w_{2h} \times \frac{1}{w_{2h}} = w_{1h}.$$

- 정상자료인 경우:

$$w_h^{f(3)} = w_{1h} \times w_{2h} \left(1 + \frac{k_{2h}(w_h - 1)}{w_h(n_{2h} - k_{2h})}\right) \left(\frac{\bar{x}_{1h}^*}{\bar{x}_{2h}^*}\right)^{\alpha} \left(\frac{\bar{z}_{2h}^*}{\bar{z}_{1h}^*}\right)^{\beta},$$

여기서 $\bar{x}_{1h}^* = (\sum_{i=1}^{n_{1h}} x_{1hi} - \sum_{i=1}^{k_{2h}} x_{2hi}^{out})/(n_{1h} - k_{2h})$, $\bar{x}_{2h}^* = (\sum_{i=1}^{n_{2h}} x_{2hi} - \sum_{i=1}^{k_{2h}} x_{2hi}^{out})/(n_{2h} - k_{2h})$이고 x_{2hi}^{out}은 2차 조사에서 이상점으로 탐지된 자료이다. 같은 방법으로 $\bar{z}_{1h}^* = (\sum_{i=1}^{n_{1h}} z_{1hi} - \sum_{i=1}^{k_{2h}} z_{2hi}^{out})/(n_{1h} - k_{2h})$, $\bar{z}_{2h}^* = (\sum_{i=1}^{n_{2h}} z_{2hi} - \sum_{i=1}^{k_{2h}} z_{2hi}^{out})/(n_{2h} - k_{2h})$이다. 따라서 ratio-cum-product 추정량은 $\hat{\bar{Y}}_{ds}^{\alpha,\beta(3)} = (1/N_{1h})\sum_{i=1}^{n_{2h}} w_h^{f(3)} y_{2hi}$이 된다. 모의실험에서는 이 결과를 M_3로 표시하였다.

방법 4:

2차 조사 자료에서 탐지된 이상점의 가중치를 "1"로 하면서 캘리브레이션 방법을 적용한다. 이상점의 나머지 가중치는 정상 자료에 나누어 주는 보정 가중치를 사용한다.

- 이상점인 경우:

$$w_h^{f(4)} = w_{1h} \times \left(\frac{\bar{x}_{1h}^*}{\bar{x}_{2h}^*}\right)^{\alpha} \left(\frac{\bar{z}_{2h}^*}{\bar{z}_{1h}^*}\right)^{\beta}.$$

Table 4.1. Coefficients for the simulation

Population type	a	b	c	d
Ratio	0	1.50	−2.00	0.25
	0	1.50	−2.00	0.50
Linear	20	1.50	−2.00	0.25
	20	1.50	−2.00	0.50

- 정상자료인 경우:

$$w_h^{f(4)} = w_{1h} \times w_{2h} \left(1 + \frac{k_{2h}(w_h - 1)}{w_h(n_{2h} - k_{2h})}\right) \left(\frac{\bar{x}_{1h}^*}{\bar{x}_{2h}^*}\right)^{\alpha} \left(\frac{\bar{z}_{2h}^*}{\bar{z}_{1h}^*}\right)^{\beta},$$

여기서 $\bar{x}_{1h}^*, \bar{x}_{2h}^*, \bar{z}_{1h}^*, \bar{z}_{2h}^*$은 방법 3과 같은 값을 사용한다. 따라서 ratio-cum-product 추정량은 $\hat{\bar{Y}}_{ds}^{\alpha,\beta(4)} = (1/N_{1h}) \sum_{i=1}^{n_{2h}} w_h^{f(4)} y_{2hi}$이 된다. 모의실험에서는 이 결과를 M_4로 표시하였다.

4. 모의실험

4.1. 모의실험 세팅

이상점의 영향력을 줄이기 위한 가중치 보정법의 성능을 살펴보기 위해 모의실험이 수행되었다. 모의실험을 위한 자료의 생성과정은 Lee 등 (1995)에서 사용한 방법과 유사한 방법을 사용하였다. 먼저 크기 N = 100,000인 모집단을 다음과 같이 생성하였다. 종속변수와 독립변수 간에 선형 및 비선형 관계를 만들기 위해 다음의 모형이 사용되었다.

$$y_i = a + bx_i + cz_i + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} \left(0, x_i^d \sigma^2\right),$$

여기서 보조자료 x_i는 Gamma(α^*, β^*), $(\alpha^* = 2, \beta^* = 10)$에서, z_i는 Gamma(α^*, β^*), $(\alpha^* = 2, \beta^* = 2)$에서 생성하였다. 이는 현실 자료에서는 꼬리가 오른쪽으로 긴 분포에서 생성된 자료가 많기 때문이다. 다음으로 오차 ϵ_i의 경우에는 감마분포$(\alpha^* = 1, \beta^* = 1)$와 표준정규분포를 이용하여 난수를 발생한 후 발생된 수에 $x^{d/2}$를 곱하여 오차를 생성하였다. Table 4.1은 선택된 상수 a, b, c, d의 값을 나타낸다. 상수 a, b, d는 Lee 등 (1995)에서 사용한 숫자이고 c는 곱추정을 위해 선택하였다. 첫 번째로 생성된 자료는 관심변수와 보조변수의 관계가 원점을 지나는 비례적 형태(ratio)이고, 두 번째 자료는 양의 절편값을 갖는 선형관계(regression)를 갖도록 하였다.

이제 1차 조사의 표본수, n_1을 20,000, 30,000으로 하고, 2차 조사의 표본 수, n_2를 300, 500, 700으로 하였다. 층화추출의 특성상 하나의 층에서 우수한 결과가 나오면 이를 전체 모집단 결과로 확장할 수 있기 때문에 모의실험을 간단히 하기 위하여 하나의 층만 있는 경우를 살펴보았다. 본 연구에서 사용한 비교 통계량은 편향(bias)와 절대편향(absolute bias; AB), 제곱근평균제곱오차(root mean squared error; RMSE)이고 다음과 같이 정의한다.

$$\text{Bias} = \frac{1}{R} \sum_{r=1}^{R} \left(\hat{Y}_r - Y_r\right), \tag{4.1}$$

$$\text{AB} = \frac{1}{R} \sum_{r=1}^{R} \left|\hat{Y}_r - Y_r\right|, \tag{4.2}$$

$$\text{RMSE} = \left(\frac{1}{R} \sum_{r=1}^{R} \left(\hat{Y}_r - Y_r\right)^2\right)^{\frac{1}{2}}, \tag{4.3}$$

Table 4.2. Ratio estimator results for bias with $n_1 = 30{,}000$, $d = 0.25$ (Normal dist)

n_2	Pop type	Bias				
		M_0	M_1	M_2	M_3	M_4
300	Ratio	7477	3962	3964	7761	7765
	Linear	14278	10783	10787	17102	17107
500	Ratio	7854	4424	4428	8102	8108
	Linear	12376	8964	8970	15099	15107
700	Ratio	8677	5276	5282	8925	8933
	Linear	13626	10181	10189	16368	16380

Table 4.3. Ratio estimator results for absolute bias (AB) with $n_1 = 30{,}000$, $d = 0.25$ (Normal dist)

n_2	Pop type	AB				
		M_0	M_1	M_2	M_3	M_4
300	Ratio	7514	4082	4085	7797	7801
	Linear	14910	11826	11830	17593	17599
500	Ratio	7858	4438	4442	8106	8112
	Linear	12710	9660	9667	15314	15323
700	Ratio	8677	5277	5283	8925	8933
	Linear	13721	10430	10439	16427	16438

Table 4.4. Ratio estimator results for root mean squared error (RMSE) with $n_1 = 30{,}000$, $d = 0.25$ (Normal dist)

n_2	Pop type	RMSE				
		M_0	M_1	M_2	M_3	M_4
300	Ratio	8515	4734	4737	8874	8879
	Linear	17568	14200	14206	20608	20616
500	Ratio	8472	4883	4887	8762	8769
	Linear	14816	11618	11627	17613	17624
700	Ratio	9113	5612	5618	9390	9399
	Linear	15217	11889	11900	18030	18044

여기서 반복수 $R = 2{,}000$을 사용하였다. 또한 외표준화잔차의 절대값이 "3" 이상인 경우를 이상점으로 탐지하였다.

4.2. 모의실험 결과

모의실험 결과는 비추정량을 사용하면서 오차의 분포가 정규분포 그리고 감마분포인 경우를 정리하였으며 다음으로 ratio-cum-product 추정량을 사용하면서 오차의 분포가 정규분포 그리고 감마분포인 경우를 정리하였다.

4.2.1. 오차가 정규분포를 따르고, 비추정량을 사용한 결과 Tables 4.2–4.9에 비추정량($\alpha = 1, \beta = 0$)을 사용하면서 오차가 정규분포인 결과를 수록하였다. 여기서 1차 조사의 표본 수 n_1은 30,000이고 2차 조사의 표본 수 n_2는 300, 500, 700이다.

결과를 살펴보면 M_3, M_4의 경우는 이상점을 처리하지 않은 M_0에 비해 성능이 떨어지는 것을 확인 할 수 있다. 반면 M_1, M_2의 경우는 M_0에 비해 우수한 결과를 주고 있다. 특히 본 추정량이 비추정량이기 때문에 모집단 형태가 Ratio 형태인 경우에 매우 우수한 결과를 주고 있다. M_1과 M_2를 비교하면 미미

Table 4.5. Ratio estimator results for bias with $n_1 = 30{,}000$, $d = 0.5$ (Normal dist)

n_2	Pop type	Bias				
		M_0	M_1	M_2	M_3	M_4
300	Ratio	15366	7651	7658	16212	16224
	Linear	25088	17245	17255	31875	31891
500	Ratio	13709	6019	6031	14531	14551
	Linear	24063	16337	16353	30696	30721
700	Ratio	13601	5952	5969	14378	14405
	Linear	24272	16640	16662	30819	30854

Table 4.6. Ratio estimator results for absolute bias (AB) with $n_1 = 30{,}000$, $d = 0.5$ (Normal dist)

n_2	Pop type	AB				
		M_0	M_1	M_2	M_3	M_4
300	Ratio	15370	7680	7687	16216	16229
	Linear	25168	17502	17512	31926	31943
500	Ratio	13709	6046	6058	14531	14551
	Linear	24074	16419	16435	30698	30723
700	Ratio	13601	5967	5983	14378	14405
	Linear	24273	16668	16690	30819	30854

Table 4.7. Ratio estimator results for root mean squared error (RMSE) with $n_1 = 30{,}000$, $d = 0.5$ (Normal dist)

n_2	Pop type	RMSE				
		M_0	M_1	M_2	M_3	M_4
300	Ratio	16352	8480	8488	17335	17348
	Linear	27754	19998	20012	34942	34962
500	Ratio	14411	6721	6733	15324	15345
	Linear	25839	18207	18227	32722	32752
700	Ratio	14126	6502	6519	14968	14996
	Linear	25561	17995	18021	32273	32312

Table 4.8. Ratio estimator results for bias with $n_1 = 30{,}000$, $d = 0.25$ (Gamma dist)

n_2	Pop type	Bias				
		M_0	M_1	M_2	M_3	M_4
300	Ratio	13421	8451	8457	11796	11804
	Linear	21295	16199	16208	21863	21874
500	Ratio	13600	8641	8660	11983	11995
	Linear	20442	15497	15511	20910	20928
700	Ratio	13003	8132	8146	11381	11397
	Linear	19433	14515	14534	19897	19921

하지만 M_1이 모든 통계량을 기준으로 우수한 것을 확인할 수 있다. 이러한 결과는 $d = 0.25, 0.5$에서 모두 확인된다.

4.2.2. 오차가 감마분포를 따르고, 비추정량을 사용한 결과 다음으로 오차가 감마분포를 따르고, 비추정량($\alpha = 1, \beta = 0$)을 사용한 결과를 Tables 4.8–4.13에 수록하였다. Tables 4.8–4.13의 결과를 살펴보면 정규분포를 사용한 결과인 Tables 4.2–4.7의 결과와 매우 유사하다. 따라서 종속변수의 분포가 정규분포 또는 감마분포를 따르더라도 M_1방법이 가장 우수한 것을 확인할 수 있다.

Table 4.9. Ratio estimator results for absolute bias (AB) with $n_1 = 30{,}000$, $d = 0.25$ (Gamma dist)

n_2	Pop type	AB				
		M_0	M_1	M_2	M_3	M_4
300	Ratio	13421	8451	8457	11796	11804
	Linear	21452	16564	16573	22069	22081
500	Ratio	13600	8651	8660	11983	11995
	Linear	20467	15580	15594	20940	20957
700	Ratio	13003	8132	8146	11381	11397
	Linear	19438	14566	14586	19906	19930

Table 4.10. Ratio estimator results for root mean squared error (RMSE) with $n_1 = 30{,}000$, $d = 0.25$ (Gamma dist)

n_2	Pop type	RMSE				
		M_0	M_1	M_2	M_3	M_4
300	Ratio	13976	8807	8814	12434	12443
	Linear	23910	18997	19009	24816	24830
500	Ratio	13971	8898	8909	12406	12419
	Linear	22074	17284	17301	22743	22764
700	Ratio	13294	8359	8373	11713	11731
	Linear	20753	15964	15987	21386	21414

Table 4.11. Ratio estimator results for bias with $n_1 = 30{,}000$, $d = 0.5$ (Gamma dist)

n_2	Pop type	Bias				
		M_0	M_1	M_2	M_3	M_4
300	Ratio	16009	5513	5523	12091	12105
	Linear	26810	16139	16152	27001	27020
500	Ratio	15294	4903	4919	11362	11384
	Linear	23788	13384	13405	23772	23801
700	Ratio	15827	5449	5471	11915	11947
	Linear	24081	13762	13790	24072	24112

Table 4.12. Ratio estimator results for absolute bias (AB) with $n_1 = 30{,}000$, $d = 0.5$ (Gamma dist)

n_2	Pop type	AB				
		M_0	M_1	M_2	M_3	M_4
300	Ratio	16009	5644	5654	12105	12119
	Linear	26847	16575	16589	27090	27109
500	Ratio	15294	4957	4972	11362	11384
	Linear	23836	13686	13707	23839	23869
700	Ratio	15827	5467	5489	11915	11947
	Linear	24086	13873	13901	24081	24121

4.2.3. 오차가 정규분포를 따르고, ratio-cum-product 추정량을 사용한 결과 오차가 정규분포를 따르고, ratio-cum-product 추정량을 사용한 결과를 Tables 4.14–4.19에 수록하였다. 여기서 M_1과 M_2 결과는 이상점 처리를 하지 않은 M_0에 비해 매우 나쁜 결과를 주기 때문에 결과표에 수록하지 않았다. 반면 M_3와 M_4의 결과에서는 $\beta = 1$과 $\beta = 0.5$를 사용하였다. 아직 β의 추정에 관한 연구가 실시되지 않았기 때문에 여기서 여러 β값을 사용하여 얻은 결과 중에서 $\beta = 0.5$인 결과를 수록하였다. 표에서 $M_{3\text{-}(1)}$, $M_{4\text{-}(1)}$이 $\beta = 1$을 사용한 결과이며 $M_{3\text{-}(0.5)}$, $M_{4\text{-}(0.5)}$는 $\beta = 0.5$를 사용한 결과이다.

Table 4.13. Ratio estimator results for root mean squared error (RMSE) with n_1 = 30,000, d = 0.5 (Gamma dist)

n_2	Pop type	RMSE				
		M_0	M_1	M_2	M_3	M_4
300	Ratio	16909	6426	6437	13233	13249
	Linear	29530	19329	19347	30348	30371
500	Ratio	15885	5589	5605	12119	12143
	Linear	25605	15728	15754	25949	25983
700	Ratio	16283	5957	5979	12497	12530
	Linear	25372	15367	15400	25651	25695

Table 4.14. Simulation results for bias with n_1 = 30,000, d = 0.25 (Normal dist)

n_2	Pop type	Bias				
		M_0	$M_{3\text{-}(1)}$	$M_{4\text{-}(1)}$	$M_{3\text{-}(0.5)}$	$M_{4\text{-}(0.5)}$
300	Ratio	2169	−703	−702	1313	1314
	Linear	4579	1148	1149	4311	4313
500	Ratio	2248	−671	−670	1441	1443
	Linear	5449	2067	2069	5489	5493
700	Ratio	2027	−881	−879	1243	1246
	Linear	4135	811	814	4206	4211

Table 4.15. Simulation results for absolute bias (AB) with n_1 = 30,000, d = 0.25 (Normal dist)

n_2	Pop type	AB				
		M_0	$M_{3\text{-}(1)}$	$M_{4\text{-}(1)}$	$M_{3\text{-}(0.5)}$	$M_{4\text{-}(0.5)}$
300	Ratio	6840	6502	6503	3382	3382
	Linear	12394	11567	11571	6860	6864
500	Ratio	5494	5106	5107	2834	2834
	Linear	10490	9299	9304	6711	6716
700	Ratio	4726	4365	4366	2473	2473
	Linear	8489	7645	7652	5369	5376

Table 4.16. Simulation results for root mean squared error (RMSE) n_1 = 30,000, d = 0.25 (Normal dist)

n_2	Pop type	RMSE				
		M_0	$M_{3\text{-}(1)}$	$M_{4\text{-}(1)}$	$M_{3\text{-}(0.5)}$	$M_{4\text{-}(0.5)}$
300	Ratio	8631	8117	8119	4267	4267
	Linear	15624	14494	14499	8638	8643
500	Ratio	6965	6429	6430	3553	3552
	Linear	13043	11618	11625	8189	8195
700	Ratio	5932	5492	5494	3100	3099
	Linear	10694	9556	9564	6597	6605

먼저 bias를 기준으로 결과를 살펴보면 제안된 방법인 $M_{3\text{-}(1)}, M_{3\text{-}(0.5)}, M_{4\text{-}(1)}, M_{4\text{-}(0.5)}$가 M_0보다 우수한 것을 확인할 수 있다. 이에 추가하여 $\beta = 1$을 사용한 경우인 $M_{3\text{-}(1)}, M_{4\text{-}(1)}$가 약간 우수한 것을 확인할 수 있다. 그러나 $d = 0.5$이고 비례형인 경우에는 $\beta = 0.5$인 결과가 우수하다. 다음으로 Absolute bias와 RMSE를 기준으로 하면 $\beta = 0.5$인 $M_{3\text{-}(0.5)}, M_{4\text{-}(0.5)}$결과가 매우 우수한 것을 확인할 수 있다. 이를 종합해 보면 미미한 차이이기는 하지만 $M_{3\text{-}(0.5)}$가 가장 우수한 결과를 준다고 판단된다.

Table 4.17. Simulation results for bias with $n_1 = 30{,}000$, $d = 0.5$ (Normal dist)

n_2	Pop type	Bias				
		M_0	$M_{3\text{-}(1)}$	$M_{4\text{-}(1)}$	$M_{3\text{-}(0.5)}$	$M_{4\text{-}(0.5)}$
300	Ratio	3460	−3280	−3277	−56	−52
	Linear	11785	4120	4126	8839	8846
500	Ratio	3723	−2869	−2865	84	91
	Linear	9893	2388	2396	7371	7382
700	Ratio	5002	−1599	−1594	1596	1604
	Linear	10846	3540	3550	8472	8487

Table 4.18. Simulation results for absolute bias (AB) with $n_1 = 30{,}000$, $d = 0.5$ (Normal dist)

n_2	Pop type	AB				
		M_0	$M_{3\text{-}(1)}$	$M_{4\text{-}(1)}$	$M_{3\text{-}(0.5)}$	$M_{4\text{-}(0.5)}$
300	Ratio	7845	7383	7384	3850	3849
	Linear	15818	12370	12378	9894	9903
500	Ratio	6723	6111	6112	3184	3183
	Linear	12637	9313	9323	8066	8078
700	Ratio	6411	4859	4860	2972	2974
	Linear	12523	8509	8522	8789	8804

Table 4.19. Simulation results for root mean squared error (RMSE) with $n_1 = 30{,}000$, $d = 0.5$ (Normal dist)

n_2	Pop type	RMSE				
		M_0	$M_{3\text{-}(1)}$	$M_{4\text{-}(1)}$	$M_{3\text{-}(0.5)}$	$M_{4\text{-}(0.5)}$
300	Ratio	9777	9219	9221	4785	4784
	Linear	19888	15700	15709	12070	12080
500	Ratio	8314	7601	7602	3965	3964
	Linear	15731	11733	11746	9719	9733
700	Ratio	7913	6023	6025	3719	3721
	Linear	15162	10625	10641	10166	10184

Table 4.20. Simulation results for bias with $n_1 = 30{,}000$, $d = 0.25$ (Gamma dist)

n_2	Pop type	Bias				
		M_0	$M_{3\text{-}(1)}$	$M_{4\text{-}(1)}$	$M_{3\text{-}(0.5)}$	$M_{4\text{-}(0.5)}$
300	Ratio	882	−3535	−3534	469	472
	Linear	3187	−1691	−1689	5009	5013
500	Ratio	1999	−2389	−2387	1869	1873
	Linear	2319	−2526	−2523	3973	3980
700	Ratio	594	−3747	−3745	477	483
	Linear	3567	−1141	−1137	5348	5358

4.2.4. 오차가 감마분포를 따르고, ratio-cum-product 추정량을 사용한 결과 마지막으로 오차가 감마분포를 따르고 ratio-cum-product 추정량을 사용한 결과를 Tables 4.20–4.25에 수록하였다. 결과를 살펴보면 정규분포 결과와 같이 감마분포에서도 AB와 RMSE를 기준으로 하면 $M_{3\text{-}(0.5)}$와 $M_{4\text{-}(0.5)}$가 가장 우수한 결과를 준다. 그러나 bias를 기준으로 하면 선형이고 $d = 0.25$인 경우 오히려 M_0에 비해 성능이 떨어지는 경우도 발생할 수 있음을 확인할 수 있다. 결과를 종합해 보면 ratio-cum-product 추정량에서는 $M_{3\text{-}(0.5)}$가 매우 우수한 결과를 주고 있다.

Table 4.21. Simulation results for absolute bias (AB) with $n_1 = 30{,}000$, $d = 0.25$ (Gamma dist)

n_2	Pop type	AB				
		M_0	$M_{3\text{-}(1)}$	$M_{4\text{-}(1)}$	$M_{3\text{-}(0.5)}$	$M_{4\text{-}(0.5)}$
300	Ratio	7153	7395	7397	3196	3195
	Linear	13225	12574	12582	7733	7740
500	Ratio	5673	5589	5591	2920	2921
	Linear	10221	9944	9953	6122	6132
700	Ratio	4722	5575	5577	2240	2239
	Linear	8795	7994	8004	6294	6306

Table 4.22. Simulation results for root mean squared error (RMSE) with $n_1 = 30{,}000$, $d = 0.25$ (Gamma dist)

n_2	Pop type	RMSE				
		M_0	$M_{3\text{-}(1)}$	$M_{4\text{-}(1)}$	$M_{3\text{-}(0.5)}$	$M_{4\text{-}(0.5)}$
300	Ratio	9011	9268	9270	4015	4014
	Linear	16486	15670	15679	9656	9665
500	Ratio	7154	6985	6988	3648	3649
	Linear	12860	12450	12462	7656	7668
700	Ratio	5910	6810	6812	2820	2819
	Linear	10939	10043	10056	7566	7581

Table 4.23. Simulation results for bias with $n_1 = 30{,}000$, $d = 0.5$ (Gamma dist)

n_2	Pop type	Bias				
		M_0	$M_{3\text{-}(1)}$	$M_{4\text{-}(1)}$	$M_{3\text{-}(0.5)}$	$M_{4\text{-}(0.5)}$
300	Ratio	4138	−5664	−5662	−1142	−1136
	Linear	7804	−2413	−2409	4502	4510
500	Ratio	2381	−7261	−7257	−2739	−2731
	Linear	9178	−968	−961	5722	5735
700	Ratio	4481	−4978	−4973	−805	−793
	Linear	7436	−2523	−2514	4309	4327

Table 4.24. Simulation results for absolute bias (AB) with $n_1 = 30{,}000$, $d = 0.5$ (Gamma dist)

n_2	Pop type	AB				
		M_0	$M_{3\text{-}(1)}$	$M_{4\text{-}(1)}$	$M_{3\text{-}(0.5)}$	$M_{4\text{-}(0.5)}$
300	Ratio	8593	8828	8830	3797	3795
	Linear	15255	13484	13493	8081	8091
500	Ratio	6573	8439	8439	3647	3641
	Linear	13098	10082	10093	7215	7229
700	Ratio	6456	6529	6530	2593	2589
	Linear	10840	8829	8842	5880	5898

5. 실제 자료 분석

2012년 연탄소비실태조사의 건별 판매량 자료 중 각 배달업자의 판매건수 및 판매량 자료 1,857개가 분석에 사용되었으며 이때 관심변수는 판매량이고 보조변수는 판매건수이다. 유사모집단을 생성하기 위해 20번의 복원추출이 사용되었다. 이렇게 만들어진 유사모집단에서 1차 조사를 위해 2,000개의 표본을 추출하고 2차 조사에서는 각각 200, 400, 600개의 표본을 추출하였다. 3.1절에서 설명한 외표준화잔차를 이용하여 이상점을 탐지하였으며 이상점 탐지 기준으로 외표준화잔차의 절대값이 2.54 이상을 사

Table 4.25. Simulation results for root mean squared error (RMSE) with $n_1 = 30{,}000$, $d = 0.5$ (Gamma dist)

n_2	Pop type	RMSE				
		M_0	$M_{3\text{-}(1)}$	$M_{4\text{-}(1)}$	$M_{3\text{-}(0.5)}$	$M_{4\text{-}(0.5)}$
300	Ratio	10732	10882	10885	4711	4709
	Linear	19397	16897	16909	10155	10167
500	Ratio	8128	10194	10195	4545	4539
	Linear	16292	12677	12692	8910	8928
700	Ratio	8008	7996	7998	3239	3235
	Linear	13542	10912	10928	7280	7303

Table 5.1. Bias results for real data analysis

n_1	n_2	Bias				
		M_0	M_1	M_2	M_3	M_4
2000	200	106935	65784	67395	77098	76244
	400	57105	50736	50869	54292	54490
	600	20656	4446	5528	21619	23715

Table 5.2. Absolute bias (AB) results for real data analysis

n_1	n_2	AB				
		M_0	M_1	M_2	M_3	M_4
2000	200	107490	71102	71511	153223	162179
	400	58573	52205	52337	55761	55959
	600	48506	36863	37723	51182	53000

Table 5.3. Root mean squared error (RMSE) results for real data analysis

n_1	n_2	RMSE				
		M_0	M_1	M_2	M_3	M_4
2000	200	120550	84283	84981	471337	528501
	400	74028	65998	66182	71581	71888
	600	57570	41039	42175	61080	64147

용하였다. 여기서 사용한 반복수는 $R = 2{,}000$이다. 독립변수가 하나뿐이고 양의 관계가 있기 때문에 비추정량이 사용되었다. 실제 자료 분석 결과를 Tables 5.1–5.3에 수록하였다. 이때 2차 표본 수 200, 400 그리고 600에 해당되는 이상점은 평균적으로 각각 9.7개, 13.4개 그리고 20.1개로 나타났으며 비율로 보면 각각 0.048%, 0.033% 그리고 0.033%가 된다.

Tables 5.1–5.3의 실제 자료 분석 결과를 살펴보면 모의실험 결과와 일치하는 것을 알 수 있다. 즉 M_3, M_4의 경우 비추정량에서는 이상점 처리를 하지 않은 M_0에 비해 우수한 결과를 주지 못하고 있다. 반면 M_1과 M_2는 모든 통계량을 기준으로 하였을 때 매우 우수한 결과를 주고 있다.

6. 결론

본 논문에서는 원할한 모의실험을 위해 이상점 탐지법으로 외표준화잔차를 이용하였다. 그러나 모의실험이 아닌 실제 자료 분석에서는 Θ-IPOD를 사용하면 더욱 우수한 이상점 탐지 결과를 얻을 수 있을 것으로 판단되며 현재 Θ-IPOD는 R-code로 되어있어 R에 익숙한 사람은 쉽게 사용할 수 있다.

흔히 이상점에 가중치를 "1"로 주는 이상점 가중치 보정 방법이 표본 조사에서 사용된다. 본 논문에 결

과를 수록하지 않았지만 이중추출의 경우 이상점으로 식별된 자료에 단순히 가중치를 "1"로 주는 방법은 매우 좋지 않은 결과를 주기 때문에 1차 조사 정보를 반드시 사용해야 한다. 이 방법이 M_1과 M_3이다.

이상점 탐지 방법의 선택은 매우 중요하다. 방법에 따라 탐지된 이상점 수가 달라지게 되고, 이상점이 아님에도 이상점이라 판단하게 되면 그 결과는 추정의 정밀성에 매우 큰 영향을 미치게 된다. 같은 이유로 이상점 판단 기준점도 이상점 탐지에 중요한 요인이 될 수 있다. 본 연구에서의 모의실험에서는 잔차의 크기가 3 이상인 경우 이상점이라 판단하였으며 실제자료 분석에서는 2.54 이상인 경우를 이상점이라 판단하였다.

ratio-cum-product 추정량의 경우 β값을 결정하는 것이 매우 중요하다. 흔히 $\alpha = 1, \beta = 1$을 사용하고 있는데, 본 모의실험 결과 $\beta = 0.5$인 경우의 결과가 매우 우수한 것을 확인하였다. 따라서 향후 α와 β를 적절히 추정하여 사용한다면 더욱 좋은 결과를 얻을 수 있을 것으로 판단된다.

References

Chamber, R. L. and Ren, R. (2004). Outlier robust imputation of survey data, ASA Section on Survey Research Methods.

Cochran, W. G. (1977). *Sampling Techniques* (3rd ed.), John Wiley & Sons, New York.

Fuller, W. A. (2000). Two-phase sampling, SSC Annual meeting. In *Proceedings of the Survey Methods Section*, 23–30, Ottawa, Canada.

Hidiroglou, M. A. (2001). Double sampling, *Survey Methodology*, **27**, 143–154.

Hidiroglou, M. A. and Sandal, C. E. (1998). Use of auxiliary information for two-phase sampling, *Survey Methodology*, Amstat proceeding.

Kim, J.-Y. and Shin, K.-I. (2013). Multiple imputation reducing outlier effect using weight adjustment methods.

Kim, M.-K. and Shin, K.-I. (2014). A multiple imputation for reducing outlier effect, *The Korean Journal of Applied statistics*, **27**, 1229–1241.

Koyuncu, N., and Kadilar, C. (2009). Family of estimators of population mean using two auxiliary variable in stratified random sampling, *Communications in Statistics-Theory and Methods*, **38**, 2938–2417.

Lee, H., Rancourt, E., and Sarndal, C.-E. (1995). Experiment with variance estimation from survey data with imputed value, *Journal of Official Statistics*, **10**, 231–243.

She, Y. and Owen, A. B. (2011). Outlier detection using nonconvex penalized regression, *Journal of the American Statistical Association*, **106**, 626–639.

Singh, H. P. and Kumar, S. (2010). Estimation of mean in presence of non-response using two phase sampling scheme, *Statistical Papers*, **51**, 559–582.

Singh, H. P., Kumar, S., and Kozak, M. (2010). Improved estimation of finite-population mean using subsampling to deal with non response in two-phase sampling scheme, *Communications in Statistics-Theory and Methods*, **39**, 791–802.

Tailor, R., Chouhan, S., and Kim, J.-M. (2014). Ratio and product type exponential estimators of population mean in double sampling for stratification, *Communications for Statistical Application and Methods*, **21**, 1–9.

Tailor, R., Lone, H. A., and Pandey, R. (2015). Generalized ratio-cum-product type estimator of finite population mean in double sampling for stratification, *Communications for Statistical Application and Methods*, **22**, 255–264.

Wu, C. and Luan, Y. (2003). Optimal calibration estimator under two-phase sampling, *Journal of Official Statistics*, **19**, 119–131.

Wu, C. and Sitter, R. R. (2001). A Model-calibration approach to using complete auxiliary information from survey data, *Journal of the American Statistical Association*, **96**, 185–193.

이중추출법에서 일반화 ratio-cum-product 방법을 이용한 이상점 가중치 보정법

오정택[a] · 신기일[a,1]

[a]한국외국어대학교 통계학과

(2016년 6월 22일 접수, 2016년 8월 9일 수정, 2016년 9월 11일 채택)

요 약

이중추출법은 모집단 정보가 충분하지 않아 층화 추출법을 사용할 때 정확한 층화 정보가 없는 경우에 흔히 사용하는 표본추출법이다. 특히 최근에는 이중추출법을 위해 1차 조사에서 얻어진 보조 정보를 이용하여 추정의 정확성을 향상시키는 방법들이 제안되었다. 본 연구에서는 최근 제안된 일반화 ratio-cum-product 추정량에서 사용하는 가중치를 이상점 처리를 위한 가중치 보정에 맞도록 보정하여 추정의 정밀성을 향상시키는 방법을 제안하였다. 모의실험을 통하여 본 연구에서 제안한 방법과 기존의 이상점 가중치 보정법의 성능을 비교하였으며 사례 분석을 통하여 제안된 방법의 우수성을 확인하였다.

주요용어: 무응답 탐지, 비추정, 곱추정, 최소제곱오차, 편향

이 논문은 2014년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NRF-2014R1A1A2056857).

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