

A Composite Estimator for Cut-off Sampling using Cost Function

Hyo-Seon Sim^a · Key-Il Shin^{a,1}

^aDepartment of Statistics, Hankuk University of Foreign Studies

(Received October 14, 2013; Revised December 12, 2013; Accepted December 12, 2013)

Abstract

Cut-off sampling has been widely used for a highly skewed population like a business survey by discarding a part of the population, so called a take-nothing stratum. For a more accurate estimate of the population total, Hwang and Shin (2013) suggested a composite estimator of a take-nothing stratum total that combined the survey results of a take-nothing stratum and a take-some sub-stratum (a part of take-some stratum). In this paper we propose a new cut-off sampling scheme by considering a cost function and a composite estimator based on the proposed sampling scheme. Small simulation studies compared the performances of known composite estimators and the new composite estimator suggested in this study. We also use Briquette Consumption Survey data for real data analysis.

Keywords: Bootstrap method, Ratio estimator, LH algorithm, BLUP estimator.

1. 서론

절사표본설계법은 모집단의 일부를 전수층으로 하고 또한 모집단의 일부를 조사에서 제외하는 표본추출 방법으로 특히 왜도가 심한 사업체 조사에서 자주 사용된다. 사업체 조사에서 대규모 사업체의 경우 그 수가 적음에도 불구하고 모집단 총합 추정에 미치는 영향이 매우 크기 때문에 흔히 전수 조사를 실시하여 총합 추정의 정도를 높이는 반면 소규모 사업체의 경우에는 표본 관리의 어려운 점과 조사의 편의를 감안하여 조사 대상에서 제외시키거나 매우 적은 수의 표본을 조사하게 된다. 이렇게 모집단을 전수층, 표본층 그리고 절사층으로 나누어 표본을 추출하는 표본설계 방법을 절사표본설계법이라 하며 최근 이에 관한 많은 연구가 진행되고 있다.

Hidiroglou (1986)는 층화추출법의 특수한 경우로 수정절사표본설계법(modified cut-off sampling)을 제안하였으며 이 방법은 절사층을 고려하지 않는 방법으로 모집단을 전수층과 표본층으로 나누는 방법이다. 이후 Lavallee와 Hidiroglou (1988)는 전수층과 절사층 그리고 여러 개의 표본층으로 구성된 절사표본설계법을 연구하였으며 흔히 LH 알고리즘이라 부르는 알고리즘을 제안하였다. 이후 Hidiroglou와 Srinath (1993)은 LH 알고리즘을 일반화한 General Lavallee-Hidiroglou 알고리즘을

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2012R1A1A2003919).

¹Corresponding author: Professor, Department of Statistics, Hankuk University of Foreign Studies, Yongin, Gyeonggi 449-791, Korea. E-mail: keyshin@hufs.ac.kr

제안하였다. 또한 Rivest (2002)는 수정된 LH 알고리즘을 발표하였으며 최근 Baillargeon과 Rivest (2011)는 R 패키지를 개발하여 질사표본설계법을 쉽게 사용할 수 있도록 하였다.

전술한 것처럼 Hidirogrou (1986)가 제안한 수정질사표본설계법은 국내에서는 에너지총조사 등 여러 표본설계에서 사용되고 있다. 그러나 질사층의 조사가 용이하지 않고 표본 관리에 어려움이 있는 경우에는 수정질사법을 사용하는 것은 비효율적일 수 있다. 물론 질사층이 총계 추정에 미치는 영향이 작다면 질사층을 사용하는 방법이 추정의 정도를 높일 수 있는 대안이 될 수 있다. 그러나 전체 총합에서 질사층이 차지하는 비중이 무시하지 못할 정도라면 조사에서 제외된 질사층에 대한 추정을 통하여 전체 모집단의 총합 추정을 실시하여야 한다. 이 경우 질사층의 총합 추정 정도가 전체 총합 추정의 정도를 결정하는 중요한 요소가 될 수도 있기 때문에 정확한 질사층 추정은 매우 중요하다.

현실적으로 질사층의 경우에는 조사에 어려움이 있는 것이 사실이다. 이러한 문제를 해결하기 위하여 질사층에서 조사를 전혀 하지 않고 기존에 있는 보조자료 또는 행정자료 등을 이용하여 질사층의 총합을 추정하는 방법이 연구되었다. 예를 들어 Elisson과 Elvers (2001)는 행정자료로부터 얻을 수 있는 보조변수를 이용하여 비추정량으로 질사층 총합을 추정하는 방법을 제안하였다. 또한 McDowney (2004)는 미국 에너지정보국(Energy Information Administration; EIA)의 월별 전기매출액 및 전기 발전 자료의 총매출액을 추정하는데 있어 모형기반이론(model-based theory)을 바탕으로 하는 질사층 추정방법을 소개하기도 하였다.

최근 질사층이 없는 수정질사법과 질사층에서 전혀 조사를 실시하지 않는 질사법의 장점을 결합한 방법이 제안되었다. 이 방법은 질사법처럼 질사층에서 전혀 조사를 실시하지 않는 것이 아니라 일부 표본 조사를 실시하지만, 수정질사법처럼 많은 수의 표본을 추출하지는 않는다. 즉 표본 관리의 어려움을 감안하고, 또한 질사층 정보를 확보하기 위하여 질사층에서도 소규모 표본을 대상으로 조사를 실시하는 것이다. 이러한 방법을 기초로 Kim과 Shin (2011), Hwang과 Shin (2012)은 효율성 면에서 상대적으로 우수한 복합추정량을 제안하였다.

이후 Hwang과 Shin (2013)은 표본층을 세분화한 후 질사층과 상관이 높은 층의 정보를 이용한 복합추정량을 제안하였다. 이를 위하여 LH 알고리즘을 기반으로 표본층을 H 개로 나눈 후 가장 상관이 높은 부차표본층의 정보를 이용한 복합추정량을 제안하였으며 기존의 복합추정량과의 모의실험 결과 비교를 통해 제안된 추정량의 우수성을 보였다. 그러나 Hwang과 Shin (2013)의 표본설계 및 모의실험은 질사층의 조사비용이 표본층의 조사비용에 비해 크다는 특성을 고려하지 않았다.

이에 본 연구에서는 비용을 고려한 복합추정량을 제안하였다. 즉 Hwang과 Shin (2013)이 제안한 표본설계 방법을 비용을 고려하여 현실에 맞도록 수정하였으며 Kim과 Shin (2011)에서 제안한 방법을 일반화한 새로운 복합추정량을 제안하였다. 제안된 추정량과 기존 추정량의 성능을 비교하기 위하여 선행 연구에서 사용한 동일한 모의실험 설계를 사용하였으며 또한 Hwang과 Shin (2013)에서 사용한 동일한 연탄소비실태조사 자료를 실제 자료분석에 사용하였다. 모의실험 결과 질사층의 조사비용이 매우 높은 일부의 경우를 제외하면 질사층에서 표본조사를 실시하고 제안된 복합추정량을 사용한 경우에 우수한 결과를 주는 것을 확인하였다.

본 논문의 구성은 다음과 같다. 먼저 2절에서는 Benedetti 등 (2010)에서 사용한 설정과 기호를 사용하여 질사표본추출법의 일반적 구조와 표현을 설명하고 기존에 제안된 여러 총합 추정량을 간단히 살펴보았다. 또한 본 논문에서 제안된 새로운 표본설계 방법 및 추정량을 설명하였다. 3절에서는 모의실험을 통하여 본 논문에서 제안한 복합추정량과 기존 추정량의 효율성을 비교하였으며 4절에서는 연탄소비실태조사 자료에 본 연구에서 제안한 방법을 적용하여 그 우수성을 확인하였다. 마지막으로 5절에 결론이 있다.

2. 절사법 총합 추정량

절사법을 위한 추정 방법을 살펴보기 위하여 Benedetti 등 (2010)이 사용한 기호와 설정을 본 연구에서도 사용하였다. 먼저 크기 N 인 모집단 U 가 식 (2.1)과 같이 세 개의 층으로 분할되어 있다고 하자.

$$U = U_C \cup U_S \cup U_{SE}, \quad U_I = U_C \cup U_S \quad (2.1)$$

여기서,

U_C : 전수층 (take-all stratum) U_{SE} : 절사표본층 (take-nothing stratum, but surveyed)

U_S : 표본층 (take-some stratum) U_I : 조사포함층 (inclusion stratum, surveyed)

이며 표본층은 H 개의 세부층($U_S = \bigcup_{h=1}^H U_{S_h}$)으로 나누어진다. 본 논문에서는 표본층의 세부층을 부차표본층이라 부르겠다. 그러면 관심변수 y 에 대한 모총합 t_y 는 식 (2.2)와 같이 각 층의 총합을 더한 형태로 나타낼 수 있다.

$$t_y = t_{yC} + t_{yS} + t_{ySE}, \quad t_{yI} = t_{yC} + t_{yS}, \quad (2.2)$$

여기서 t_y 는 전체 모집단 총계를, t_{yC} , t_{yS} 그리고 t_{ySE} 는 각각 전수층, 표본층 그리고 절사표본층의 총합을 나타내며 조사포함층의 총합은 t_{yI} 이다. 따라서 모집단 총합 추정량 \hat{t}_y 는 각 층의 총합 추정량의 합으로 구해진다. 또한 조사포함층 총합 추정량 \hat{t}_{yI} 는 식 (2.3)으로 전수층과 표본층의 총합 추정량의 합으로 구해진다.

$$\hat{t}_y = \hat{t}_{yC} + \hat{t}_{yS} + \hat{t}_{ySE}, \quad \hat{t}_{yI} = \hat{t}_{yC} + \hat{t}_{yS}. \quad (2.3)$$

또한 보조변수 x 가 주어졌을 때 식 (2.1)에서 사용된 정의를 적용하여 x 에 대한 모집단 및 분할층의 총합을 t_x , t_{xC} , t_{xS} , t_{xSE} 로 표시하였으며 이에 대응되는 추정량도 같은 방법으로 표시한다. 또한 I , S , SE 는 해당 층별로 표본에 포함된 자료의 지시함수 집합을 의미하며 $S = \bigcup_{h=1}^H S_h$ 로 h 부차표본층에 포함된 표본의 지시함수 집합 S_h 의 합이 된다. 이상의 지시함수 집합은 복합추정량을 구할 때 사용된다.

2.1. 기존의 총합 추정량 및 층화 알고리즘

2.1.1. 기존의 총합 추정량 이 절에서는 기존에 제안된 총합 추정량 중에서 Sarndal 등 (1992)이 제안한 SSW 추정량과 Kim과 Shin (2011), Hwang과 Shin (2012) 그리고 Hwang과 Shin (2013)이 제안한 복합추정량을 간단히 살펴보았다.

(1) Sarndal-Swansson-Wretman(SSW) 추정량

Sarndal 등 (1992)은 모집단과 조사포함층의 두 변수, 관심변수 y 와 보조변수 x 사이의 총합비를 혹은 상대적 크기를 이용한 총합 추정량을 제안하였다. $R_{yx} = t_y/t_x$ 과 $R_{yXI} = t_{yI}/t_{xI}$ 로 정의할 때 이 두 값이 매우 근사한 경우, 즉 $R_{yXI} \approx R_{yx}$ 인 경우를 가정하여 모총합 추정량 \hat{t}_y^{SSW} 을 다음과 같이 제안하였다.

$$\hat{t}_y^{SSW} = \hat{R}_{yXI} t_x, \quad (2.4)$$

여기서 $\hat{R}_{yXI} = \hat{t}_{yI}/\hat{t}_{xI}$ 이며 $\hat{t}_{yI} = \hat{t}_{yC} + \hat{t}_{yS}$, $\hat{t}_{xI} = \hat{t}_{xC} + \hat{t}_{xS}$, $\hat{t}_{yS} = \sum_{k \in S} w_k y_k$, $\hat{t}_{xS} = \sum_{k \in S} w_k x_k$ 이며 w_k 는 표본에 주어진 가중치이다.

(2) 복합추정량

절사표본층에서 소규모 표본이 추출되었다고 가정하고 절사표본층에서 얻어진 결과와 조사포함층에서 얻어진 결과를 선형결합하여 만들어진 복합추정량을 간단히 살펴보면 다음과 같다.

먼저 Kim과 Shin (2011)은 절사표본층 총합 추정량으로 식 (2.4)의 SSW방법에 의한 절사표본층의 총합 추정량 $\hat{t}_{ySE}^{SSW} = (\hat{t}_{yI}/\hat{t}_{xI})t_{xSE}$ 과 조사에 의해 얻어진 절사표본층의 비추정량인 $\hat{t}_{ySE}^{Ratio} = (\hat{t}_{ySE}/\hat{t}_{xSE})t_{xSE}$ 을 선형결합한 복합추정량인 식 (2.5)를 제안하였다.

$$\hat{t}_{ySE}^{MODI-SSW} = \left(\alpha^{[1]} \frac{\hat{t}_{ySE}}{\hat{t}_{xSE}} + (1 - \alpha^{[1]}) \frac{\hat{t}_{yI}}{\hat{t}_{xI}} \right) t_{xSE}, \quad (2.5)$$

여기서 $\hat{t}_{ySE} = \sum_{k \in SE} w_k y_k$, $\hat{t}_{xSE} = \sum_{k \in SE} w_k x_k$ 이다.

또한 Hwang과 Shin (2012)은 조사포함층을 이용한 복합추정량인 식 (2.6)과 표본층만을 이용한 복합추정량인 식 (2.7)을 각각 제안하였다.

$$\hat{t}_{ySE}^{MODI-BLUP} = \hat{R}_{SE}^{MODI-BLUP} t_{xSE} = \left(\alpha^{[2]} \frac{\hat{t}_{ySE}}{\hat{t}_{xSE}} + (1 - \alpha^{[2]}) \frac{\hat{T}_{yS_I}}{\hat{T}_{xS_I}} \right) t_{xSE}, \quad (2.6)$$

$$\hat{t}_{ySE}^{MODI-BLUPA} = \hat{R}_{SE}^{MODI-BLUPA} t_{xSE} = \left(\alpha^{[3]} \frac{\hat{t}_{ySE}}{\hat{t}_{xSE}} + (1 - \alpha^{[3]}) \frac{\hat{T}_{yS}}{\hat{T}_{xS}} \right) t_{xSE}, \quad (2.7)$$

여기서 $\hat{T}_{yS_I} = \sum_{k \in I} y_k$, $\hat{T}_{xS_I} = \sum_{k \in I} x_k$ 이고 $\hat{T}_{yS} = \sum_{k \in S} y_k$, $\hat{T}_{xS} = \sum_{k \in S} x_k$ 이다.

다음으로 Hwang과 Shin (2013)은 LH 알고리즘을 이용하여 표본층을 H 개로 분할한 후, 절사표본층과 가장 가까운 부차표본층(U_{S_1})이 가장 많은 정보를 갖고 있다는 가정 하에서 U_{S_1} 만을 이용한 절사층 총합 추정량인 식 (2.8)을 제안하였다.

$$\hat{t}_{ySE}^{MODI-BLUPN} = \left(\alpha^{[4]} \frac{\hat{t}_{ySE}}{\hat{t}_{xSE}} + (1 - \alpha^{[4]}) \frac{\hat{T}_{yS_1}}{\hat{T}_{xS_1}} \right) t_{xSE}, \quad (2.8)$$

여기서 $\hat{T}_{yS_1} = \sum_{k \in S_1} y_k$ 이고 $\hat{T}_{xS_1} = \sum_{k \in S_1} x_k$ 이다.

다음으로 식 (2.5)에서 식 (2.8)에 포함된 가중치 $\alpha^{[1]}, \dots, \alpha^{[4]}$ 에 대한 추정값은 복합추정량에서 흔히 사용하는 방법으로 각 추정량의 분산을 이용하여 추정할 수 있다. 예를 들어 $\alpha^{[1]}$ 의 경우에는 다음과 같이 구할 수 있다.

$$\hat{\alpha}^{[1]} = \frac{\text{MSE}(\hat{t}_{yI}/\hat{t}_{xI})}{\text{MSE}(\hat{t}_{ySE}/\hat{t}_{xSE}) + \text{MSE}(\hat{t}_{yI}/\hat{t}_{xI})} \approx \frac{\text{Var}(\hat{t}_{yI}/\hat{t}_{xI})}{\text{Var}(\hat{t}_{ySE}/\hat{t}_{xSE}) + \text{Var}(\hat{t}_{yI}/\hat{t}_{xI})}. \quad (2.9)$$

같은 방법을 이용하면 각각의 복합추정량 가중치를 추정할 수 있으며, 가중치 α 추정에 관한 자세한 내용은 Rao (2003)를 살펴보기 바란다. 최종적으로 모집단 총합 추정량은 식 (2.3)을 이용하면 된다. 본 연구에서는 조사포함층의 총합 추정량은 식 (2.4) 이후에서 정의한 추정량 \hat{t}_{yI} 을 동일하게 사용하고 절사표본층의 총합 추정량으로 식 (2.4), (2.5), (2.6) 그리고 (2.8)을 사용하였다.

2.1.2. 층화 알고리즘 왜도가 큰 모집단에 대하여 전수조사를 하는 전수층과 표본조사를 하는 표본층을 여러 층으로 나누기 위한 경계점과 층별 표본 수를 구하는 층화 알고리즘에는 LH(Lavallee-Hidioglou) algorithm, geometric stratification algorithm, random search algorithm 등이 있다. 이 방법들은 최종 표본 수가 최소가 되도록 최적의 층간 경계점을 계산하는 방법이다. 이 방법들에 관한 알고리즘과 R code는 Baillargeon과 Rivest (2011)을 참조하기 바란다. 또한 이론에 관한 자세한 내용은 Lavallee와 Hidioglou (1988)와 Rivest (2002)를 참조하기 바란다.

2.2. 제안된 총합 추정량

2.2.1. 제안된 복합추정량 2.1절에서 설명된 절사표본층 복합추정량을 살펴보면 절사표본층의 조사 결과를 이용한 추정량과 조사포함층에서 얻어진 추정량의 선형결합추정량으로 구성되어 있다. 이 경우 절사표본층 추정에 있어서, 전수층을 포함하는 조사포함층의 정보를 모두 이용할 수도 있고 절사표본층과 전수층의 특성이 다를 수도 있음을 고려하여 선택적으로 일부 표본층의 정보를 이용할 수도 있다. 일부 표본층 정보를 이용하기 위하여 Hwang과 Shin (2013)은 LH 방법을 이용하여 절사표본층과 근접한 부차표본층의 정보를 기반으로 BLUP을 이용한 복합추정량을 제안하였다.

본 논문에서는 최적 부차표본층을 찾기 위해 LH 방법을 동일하게 이용하였으나 가장 근접한 하나의 층을 사용하지 않고 층을 세분화 한 후 절사표본층에 가까운 부차표본층을 합쳐 나가면서 최적의 층을 구하는 방법을 사용하였다. 이는 최적 층을 찾기 위해 여러번의 표본설계를 실시해야 하는 Hwang과 Shin (2013)의 표본설계의 번거로움을 없애기 위한 방법으로 구체적인 설명은 다음과 같다.

먼저 표본층을 H 개의 부차표본층으로 나눈다. 표본층을 H 개로 나누는 방법으로 2.1.2절에서 설명한 LH 층화 알고리즘을 사용한다. 얻어진 H 개의 부차표본층을 $U_{S_1}, U_{S_2}, \dots, U_{S_H}$ 이라 하면 표본층 U_S 는 $U_S = \bigcup_{h=1}^H U_{S_h}$ 이 된다. 이제 $U_{T_k} = \bigcup_{h=1}^k U_{S_h}$ 라 하고 $U_{S_1}, U_{S_2}, \dots, U_{S_k}$ 에서 U_{S_1} 이 가장 절사표본층에 가깝고 U_{S_k} 가 가장 멀리 떨어진 부차표본층이라 하면 제안된 절사표본층의 복합추정량은 다음과 같다.

$$\hat{t}_{ySE}^{MODI-RATIO} = \left(\alpha^{[5]} \frac{\hat{t}_{ySE}}{\hat{t}_{xSE}} + (1 - \alpha^{[5]}) \frac{\hat{t}_{yT_k}}{\hat{t}_{xT_k}} \right) t_{xSE}, \quad (2.10)$$

여기서 $\hat{t}_{yT_k} = \sum_{i=1}^k \sum_{j \in S_i} w_i y_{ij}$ 이고 $\hat{t}_{xT_k} = \sum_{i=1}^k \sum_{j \in S_i} w_i x_{ij}$ 가 된다. 만약 $w_i = w$ 이면 식 (2.10)의 복합추정량은 Hwang과 Shin (2013)이 제안한 복합추정량과 동일하게 된다. 본 논문에서 제안한 복합추정량은 기존에 제안된 복합추정량의 일반식이라 할 수 있다. 또한 $\alpha^{[5]}$ 는 식 (2.9)에서 사용한 방법을 이용하여 추정한다. 최종적인 총합 추정량은 $\hat{t}_y^{MODI-RATIO} = \hat{t}_{yI} + \hat{t}_{ySE}^{MODI-RATIO}$ 이 된다.

2.2.2. 비용함수를 고려한 표본 규모 배분 2.1절에서 설명한 기존의 복합추정량은 우수한 성능을 보이는 것으로 알려져 있으나 각 성능 비교를 위해 사용한 모의실험의 표본 규모는 표본층과 절사표본층에서 얻어진 표본의 조사비용이 동일하다는 조건에서 얻어진 결과이므로 비현실적일 수 있다. 따라서 본 연구에서 비용함수를 적용한 표본규모를 고려하였다. 먼저 흔히 사용하는 비용함수는 다음과 같다.

$$C = C_0 + C_{CNC} + C_1 n_1 + C_2 n_2 + \dots + C_{HNH} + C_{SENSE},$$

여기서 C_C 는 전수층의 조사 비용이고 n_C 는 전수층 표본 규모이다. C_{SE} 는 절사표본층의 조사비용이고 n_{SE} 는 절사표본층 표본 규모이다. 또한 C_h 와 n_h 는 표본층의 조사비용과 표본 규모이다. 본 연구에서는 모든 추정량에서 동일한 전수층의 크기를 사용하기 때문에 위의 비용함수는 다음과 같이 표시된다.

$$C = C_0^* + C_1 n_1 + C_2 n_2 + \dots + C_{HNH} + C_{SENSE}.$$

또한 표본층의 조사 비용은 모든 부차표본층에 무관하게 모두 같다고 가정한다면 최종 비용함수는 다음과 같이 표시된다.

$$C = C_0^* + C_S n_S + C_{SENSE}.$$

만약 $C_S = C_{SE}$ 라 가정한다면 기존에 제안된 복합추정량의 모의실험 결과는 타당할 것이다. 그러나 현실적으로는 $C_{SE} \geq C_S$ 이므로 절사표본층에서 표본을 조사하게 되면 그에 해당되는 표본 수보다 더 많은 수의 표본 수가 표본층에서 감소되어야 한다.

Table 3.1. Coefficients used for generating data

y_k 형태	a	b	c	d	g
비례형(ratio)	0	1.50	0.00	5.13	0.50
선형(regression)	20	1.50	0.00	13.79	0.25
볼록형(convex)	0	0.25	0.01	4.91	0.50
오목형(concave)	0	3.00	-0.01	5.60	0.50

Table 3.2. Summary of the sample design with 80% cut-off point

절사점 기준	N_C	N_{SE}	N_S	n_{SE}	n_S	적용한 목표 CV 값(%)	
80%	76	4028	5896	0	0배	424	0.170
				5	1배	419	0.171
					10배	374	0.182
					20배	324	0.197
					30배	274	0.215
				10	1배	414	0.172
					10배	324	0.197
					20배	224	0.239
					30배	124	0.325
				20	1배	404	0.175
					10배	224	0.239
					20배	24	0.751

3. 모의실험을 통한 모총합 추정량 비교

3.1. 모의실험 설계

2절에서 제안된 모총합 추정량의 효율성을 비교하기 위하여 모의실험을 실시하였다. 모의실험에서 사용된 모집단 자료 생성 방법은 Lee 등 (1995)과 Hwang과 Shin (2013)에서 사용한 방법과 동일하며 사용된 계수의 값도 동일하다. 이를 간단히 살펴보면 다음과 같다.

먼저 보조자료 x_k 를 평균 48, 분산 768을 갖는 감마분포로부터 생성시키고 각각에 대응하는 네 가지 형태의 관심변수 y_k 를 만들었다. 각 조사변수 y_k 는 평균이 $\mu(x) = a + bx + cx^2$ 이고 분산은 $\sigma^2(x) = d^2x^{2g}$ 을 갖는 감마분포를 따른다고 가정하였다. 각각의 모집단 크기는 $N = 10,000$ 이고 각 경우에 대하여 절사층을 만든다.

Table 3.1은 모집단 형태에 따른 상수 a, b, c, d, g 를 나타낸 것이다. 첫 번째 모집단 자료는 보조변수 x_k 와 관심변수 y_k 의 관계가 원점을 지나는 비례형(ratio)이고, 두 번째 모집단 자료는 양의 절편 값을 갖는 선형관계(regression), 세 번째 모집단 자료는 볼록한 형태(convex), 네 번째 모집단 자료는 오목한 형태(concave)가 된다.

다음으로 보조변수 x_k 값을 기준으로 모집단을 전수층, 표본층 그리고 절사표본층으로 분리하였다. 층간 경계점은 LH 알고리즘을 사용하여 계산하였다. 또한 절사층을 나누는 기준으로는 하위 80%와 90%의 두 개 값을 고려하였다. 전체 표본 규모는 $n = 500$ 이고 표본추출률은 $f = n/N = 0.05$ 이다. 또한 절사표본층의 표본 수는 $n_{SE} = 5, 10, 20$ 세 가지로 하였다.

LH 알고리즘을 사용하기 위해서는 사전에 층의 개수와 목표 CV를 입력하여야 하므로 미리 정한 총 표본 수와 층의 개수에 맞추어 역으로 CV를 계산하여야 한다. 예를 들어, 모집단을 전수층, 표본층 두 개($S = S_1 \cup S_2$), 절사표본층으로 나누고 절사표본층을 하위 80% 기준으로 구성하려 한다면 보조변수

Table 3.3. Population and sample sizes for sub-sample strata with 80% cut-off point

n_{SE}	비용	표본층										
			s10	s9	s8	s7	s6	s5	s4	s3	s2	s1
0	0배	N	858	834	798	752	727	591	510	377	287	162
		n	38	43	43	42	51	46	46	42	39	34
5	1배	N	860	832	796	755	726	592	509	377	287	162
		n	38	42	42	42	50	46	45	41	39	34
	10배	N	859	834	797	752	727	591	509	378	287	162
		n	34	38	38	37	45	41	40	37	34	30
	20배	N	858	835	797	752	727	590	510	378	287	162
		n	29	33	33	32	39	35	35	32	30	26
30배	N	862	834	789	758	726	592	508	378	287	162	
	n	25	28	27	28	33	30	29	27	25	22	
10	1배	N	860	833	797	752	727	592	508	378	287	162
		n	38	42	42	41	50	45	44	41	38	33
	10배	N	859	834	797	752	727	591	510	377	287	162
		n	29	33	33	32	39	35	35	32	30	26
	20배	N	857	836	787	762	727	590	510	378	288	161
		n	20	23	22	23	27	24	24	22	21	18
30배	N	854	839	782	767	731	586	518	373	284	162	
	n	11	13	12	13	15	13	14	12	11	10	
20	1배	N	860	833	797	752	727	591	509	378	287	162
		n	37	41	41	40	49	44	43	40	37	32
	10배	N	868	835	787	752	728	590	510	377	288	161
		n	21	23	22	22	27	24	24	22	21	18
	20배	N	858	799	765	715	732	629	539	413	282	164
		n	2	2	2	2	3	3	3	3	2	2

x_k 값 기준으로 하위 80% 미만의 값들을 모집단에서 제외시킨 뒤 LH 방법을 적용하였다. 따라서 층의 개수는 전수층을 포함하여 3개가 된다. 또한 총 표본 수가 500이므로 만약 절사표본층의 표본 수가 5라면 나머지 표본층 표본 수는 495가 되도록 목표 CV 값을 역으로 계산하였다.

한편 가중치 α 의 추정값을 얻기 위해 붓스트랩(Bootstrap) 방법을 이용하였으며 사용한 반복수는 $B = 200$ 이다. 또한 복합추정량의 비교를 위해 사용한 반복 횟수는 $R = 1,000$ 이며 비교를 위한 비교통계량으로 편향(bias), 상대편향(relative bias, rbias) 그리고 제곱근평균제곱오차(root mean square error; RMSE)를 사용하였다.

$$\text{bias} = \bar{\hat{t}}_y - t_y,$$

$$\text{rbias}(\%) = \frac{100(\bar{\hat{t}}_y - t_y)}{t_y},$$

$$\text{rmse} = \sqrt{\frac{1}{R} \sum_{r=1}^R [\hat{t}_y(r) - t_y]^2},$$

여기서 $\bar{\hat{t}}_y = \sum_{r=1}^R \hat{t}_y(r)/R$ 이다.

Table 3.2-Table 3.3은 본 모의실험에서 사용한 층별 모집단 수, 층별 표본 수 및 LH 방법에 사용한 목

Table 3.4. Simulation results with 80% cut-off point and $n_{SE} = 5$

형태		S_I			최적 층 결과			
		SSW	$M-B$	$M-R$	층	$M-B_{OP}$	$M-R_{OP}$	
비례	1배	bias	-4582	-3996	-4777		-3777	-3894
		rbias(%)	-0.64	-0.56	-0.66	10	-0.53	-0.54
		rmse	14557	13895	13325		13540	13936
	10배	bias	-4582	-3773	-4595		-3527	-3674
		rbias(%)	-0.64	-0.52	-0.64	10	-0.49	-0.51
		rmse	14557	14917	14310		14663	14966
선형	1배	bias	-47983	-29312	-30552		-21197	-21197
		rbias(%)	-5.19	-3.17	-3.31	1	-2.29	-2.29
		rmse	49918	34820	37472		27555	27555
	10배	bias	-47983	-29606	-30921		-21017	-21017
		rbias(%)	-5.19	-3.20	-3.35	1	-2.27	-2.27
		rmse	49918	35516	38232		28634	28634
	20배	bias	-47983	-29226	-30722		-20236	-20236
		rbias(%)	-5.19	-3.16	-3.33	1	-2.19	-2.19
		rmse	49918	35077	38022		28108	28108
	30배	bias	-47983	-29149	-30845		-20361	-20361
		rbias(%)	-5.19	-3.15	-3.34	1	-2.20	-2.20
		rmse	49918	35467	38575		28903	28903
블록	1배	bias	46290	7402	5358		3554	3496
		rbias(%)	10.75	1.72	1.24	2	0.83	0.81
		rmse	48338	25884	33927		16836	16805
	10배	bias	46290	7840	5794		3727	3669
		rbias(%)	10.75	1.82	1.35	2	0.87	0.85
		rmse	48338	25921	34189		17264	17239
	20배	bias	46290	8729	6534		3942	3882
		rbias(%)	10.75	2.03	1.52	2	0.92	0.90
		rmse	48338	26169	34829		17670	17679
	30배	bias	46290	8512	6436		4033	3973
		rbias(%)	10.75	1.98	1.50	3	0.94	0.92
		rmse	48338	26075	34937		19121	19129
오목	1배	bias	-42386	-21516	-21978		-9929	-9853
		rbias(%)	-3.73	-1.89	-1.94	2	-0.87	-0.87
		rmse	45296	29698	36446		20118	20102
	10배	bias	-42386	-21943	-22259		-10059	-9970
		rbias(%)	-3.73	-1.93	-1.96	2	-0.89	-0.88
		rmse	45296	30603	37689		20957	20945
	20배	bias	-42386	-21704	-21988		-9945	-9834
		rbias(%)	-3.73	-1.91	-1.94	2	-0.88	-0.87
		rmse	45296	30703	38058		21228	21193
	30배	bias	-42386	-22678	-23278		-10431	-10301
		rbias(%)	-3.73	-2.00	-2.05	2	-0.92	-0.91
		rmse	45296	31992	39547		23363	23338

표 CV를 정리한 표이다. 참고로 90% 기준 질사표본층에 관한 모의실험도 실시하였으나 80% 결과와 매우 유사하여 본 논문에는 수록하지 않았다.

Table 3.5. Simulation results with 80% cut-off point and $n_{SE} = 10$

형태	S_I			최적 층 결과				
	SSW	$M-B$	$M-R$	층	$M-B_{OP}$	$M-R_{OP}$		
비례	1배	bias	-4582	-3771	-4510		-3571	-3677
		rbias(%)	-0.64	-0.52	-0.63	10	-0.50	-0.51
		rmse	14557	13609	13078		13301	13648
	10배	bias	-4582	-3102	-4069		-2936	-3005
		rbias(%)	-0.64	-0.43	-0.57	10	-0.41	-0.42
		rmse	14557	14733	14074		14508	14786
선형	1배	bias	-47983	-21828	-21482		-16585	-16585
		rbias(%)	-5.19	-2.36	-2.33	1	-1.80	-1.80
		rmse	49918	28356	29715		23031	23031
	10배	bias	-47983	-21911	-21641		-15623	-15623
		rbias(%)	-5.19	-2.37	-2.34	1	-1.69	-1.69
		rmse	49918	28713	30200		23908	23908
	20배	bias	-47983	-22188	-22314		-15131	-15131
		rbias(%)	-5.19	-2.40	-2.42	1	-1.64	-1.64
		rmse	49918	29430	31167		25821	25821
	30배	bias	-47983	-23432	-24684		-17204	-17091
		rbias(%)	-5.19	-2.54	-2.67	2	-1.86	-1.85
		rmse	49918	33195	35238		30838	30858
블록	1배	bias	46290	9308	7892		5271	5216
		rbias(%)	10.75	2.16	1.83	2	1.22	1.21
		rmse	48338	24871	30409		17257	17258
	10배	bias	46290	9685	8085		4483	4419
		rbias(%)	10.75	2.25	1.88	2	1.04	1.03
		rmse	48338	25143	31040		17882	17846
	20배	bias	46290	10596	8454		5299	5227
		rbias(%)	10.75	2.46	1.96	3	1.23	1.21
		rmse	48338	25860	32575		20101	20117
	30배	bias	46290	12612	10742		6843	6696
		rbias(%)	10.75	2.93	2.50	4	1.59	1.56
		rmse	48338	28007	34796		24815	24840
오목	1배	bias	-42386	-17225	-15610		-8279	-8208
		rbias(%)	-3.73	-1.52	-1.37	2	-0.73	-0.72
		rmse	45296	26317	30844		19230	19212
	10배	bias	-42386	-17342	-15769		-8047	-7960
		rbias(%)	-3.73	-1.53	-1.39	2	-0.71	-0.70
		rmse	45296	26607	31377		20094	20075
	20배	bias	-42386	-18615	-16911		-8158	-8063
		rbias(%)	-3.73	-1.64	-1.49	2	-0.72	-0.71
		rmse	45296	29118	34209		24190	24222
	30배	bias	-42386	-18121	-17881		-10240	-10008
		rbias(%)	-3.73	-1.60	-1.57	4	-0.90	-0.88
		rmse	45296	31652	36625		28834	28825

3.2. 모의실험 결과

모의실험 결과는 Table 3.4-Table 3.6에 수록하였다. 결과표에서 SSW는 식 (2.4)의 \hat{t}_y^{SSW} 를, $M-B$ 는

Table 3.6. Simulation results with 80% cut-off point and $n_{SE} = 20$

형태		S_I			최적 층 결과			
		SSW	$M-B$	$M-R$	층	$M-B_{OP}$	$M-R_{OP}$	
비례	1배	bias	-4582	-3091	-3804		-2930	-3001
		rbias(%)	-0.64	-0.43	-0.53	10	-0.41	-0.42
		rmse	14557	13233	12688		12951	13274
	10배	bias	-4582	-3256	-4323		-3092	-3173
		rbias(%)	-0.64	-0.45	-0.60	10	-0.43	-0.44
		rmse	14557	16963	16018		16675	17025
선형	1배	bias	-47983	-16320	-15265		-13089	-13089
		rbias(%)	-5.19	-1.77	-1.65	1	-1.42	-1.42
		rmse	49918	23837	24455		20832	20832
	10배	bias	-47983	-16865	-16081		-13596	-13545
		rbias(%)	-5.19	-1.83	-1.74	2	-1.47	-1.47
		rmse	49918	25060	25857		23134	23143
20배	bias	-47983	-18074	-26102		-19705	-17736	
	rbias(%)	-5.19	-1.96	-2.83	10	-2.13	-1.92	
	rmse	49918	50781	47327		49860	50971	
블록	1배	bias	46290	7449	5301		5511	5475
		rbias(%)	10.75	1.73	1.23	3	1.28	1.27
		rmse	48338	20895	24035		15710	15711
	10배	bias	46290	9013	6138		5323	5279
		rbias(%)	10.75	2.09	1.43	3	1.24	1.23
		rmse	48338	22798	26354		19473	19469
20배	bias	46290	14670	21767		17474	13746	
	rbias(%)	10.75	3.41	5.06	10	4.06	3.19	
	rmse	48338	50677	49158		49738	51149	
오목	1배	bias	-42386	-12907	-10146		-7176	-7111
		rbias(%)	-3.73	-1.14	-0.89	2	-0.63	-0.63
		rmse	45296	23048	25705		18010	18007
	10배	bias	-42386	-13834	-10947		-7705	-7629
		rbias(%)	-3.73	-1.22	-0.96	3	-0.68	-0.67
		rmse	45296	25321	28202		22669	22673
20배	bias	-42386	-15156	-28739		-18254	-14305	
	rbias(%)	-3.73	-1.33	-2.53	10	-1.61	-1.26	
	rmse	45296	60566	54181		59259	61011	

식 (2.6)을 이용한 $\hat{t}_y^{MODI-BLUP}$ 을, 그리고 $M-R$ 은 식 (2.5)를 이용한 $\hat{t}_y^{MODI-SSW}$ 를 나타낸다. 또한 $M-B_{OP}$ 는 식 (2.8)의 Hwang과 Shin (2013)이 제안한 방법이고 $M-R_{OP}$ 는 식 (2.10)을 이용한 총합추정량으로 본 논문에서 제안한 방법에서 얻어진 결과이다.

Table 3.4의 비례형 결과에서 비용이 같을 경우를 살펴보면 최적층을 사용한 결과는 층 10에서 결정되고 있어 모든 표본층을 사용하였을 때 가장 우수한 결과를 준다. 그러나 최종적으로는 모든 표본층 자료와 전수층 자료를 모두 사용한 $M-R$ 이 가장 우수한 결과를 준다. 만약 질사표본층의 비용이 표본층에 비해 10배가 된다면 가장 우수한 결과를 주는 $M-R$ 만이 사용가능하나 이 또한 SSW와 큰 차이를 보이고 있지 않다. 따라서 모집단 형태가 비례형이고 질사표본층의 조사비용이 많이 드는 경우에는 그 사용에 주의를 하여야 한다.

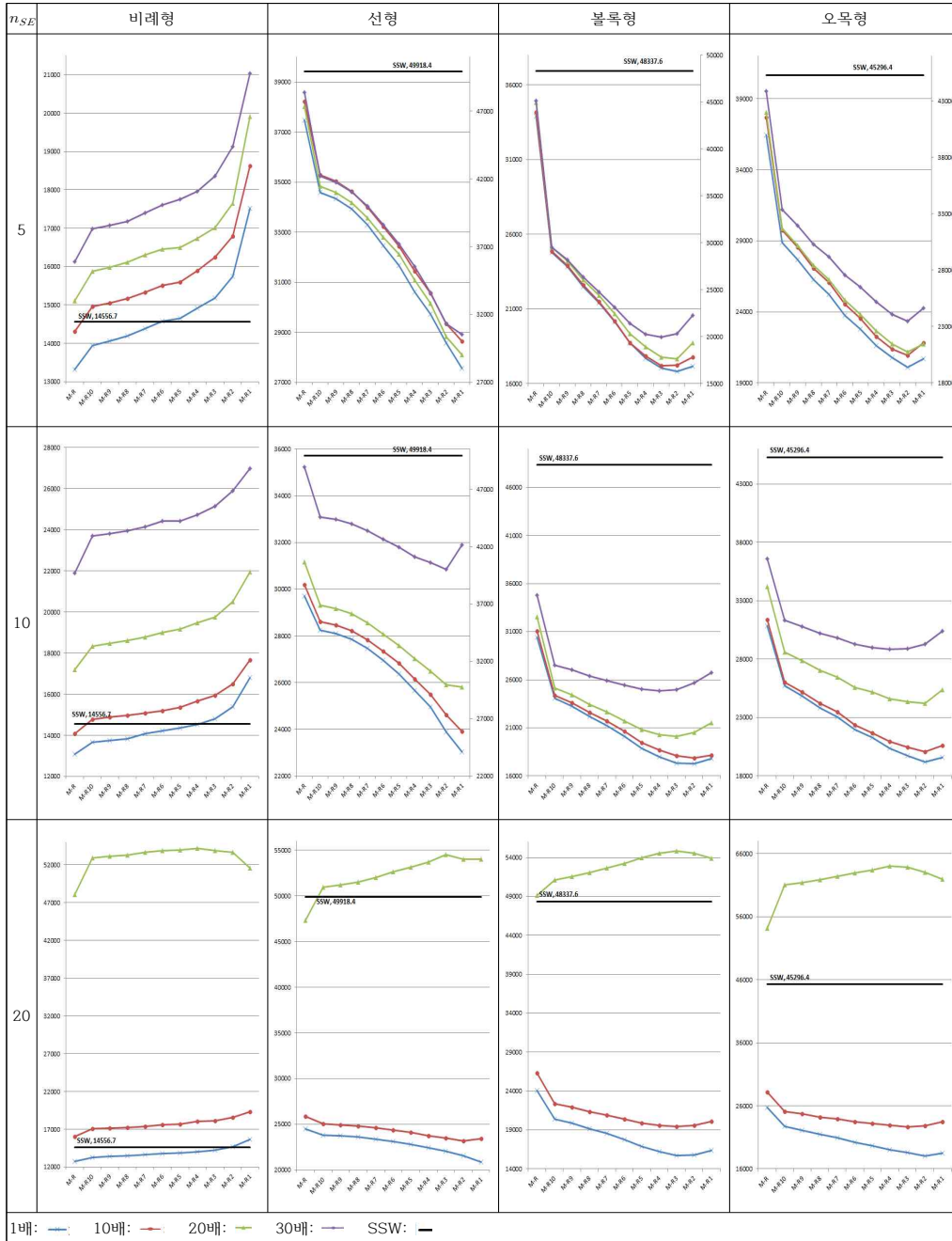


Figure 3.1. Simulation results with 80% cut-off point and $n_{SE} = 5, 10$ and 20

다음으로 모집단이 선형인 경우를 살펴보면 최적층을 사용한 결과가 매우 우수한 것을 확인할 수 있다. 특히 최적층이 1층에서 결정되기 때문에 절사표본층과 가장 가까운 표본층을 사용하는 것이 가장 우수

Table 4.1. Summary of the sample design for real data analysis

절사점 기준	N_C	N_{SE}	N_S	n_{SE}	n_S	적용한 목표 CV 값(%)						
80%	80	960	572	0	0배	110	0.458					
				5	1배	105	0.472					
					3배	95	0.503					
					5배	85	0.539					
					7배	75	0.581					
				10	1배	100	0.487					
					3배	80	0.559					
					5배	60	0.662					
					90%	80	748	784	0	0배	110	0.629
									5	1배	105	0.646
3배	95	0.685										
5배	85	0.729										
7배	75	0.783										
10	1배	100	0.665									
	3배	80	0.755									
	5배	60	0.885									

한 결과를 주고 있다. 또한 비용을 고려한 경우를 살펴보면 비용이 30배가 되는 경우까지도 복합추정량이 우수한 결과를 준다. 참고로 최적층이 1층이 되면 $M-B_{OP}$ 와 $M-R_{OP}$ 는 같은 결과를 준다. 모집단이 블록인 경우와 오목인 형태의 결과는 모두 최적층을 사용하였을 때 가장 좋은 결과를 주고 있으며 이 또한 비용이 30배가 되는 경우에도 복합추정량이 우수한 결과를 주고 있다. 전체적으로 $M-B_{OP}$ 와 $M-R_{OP}$ 는 유사한 성능을 보이고 있다.

Table 3.5의 결과를 살펴보면 전체적으로 Table 3.4와 유사한 결과를 주고 있다. 다만 절사표본층의 표본 규모가 증가하였기 때문에 전체적으로 표본층의 표본 규모가 더 많이 감소하여 최적층이 좀 더 큰 쪽으로 결정되었다. 이는 복합추정량의 성능을 향상시키기 위해서는 어느 수준 이상의 부차표본층 규모가 필요한 것을 말해주는 것이다. 다음으로 Table 3.6의 결과는 절사표본층의 규모가 매우 크기 때문에 표본층의 규모는 비용에 따라 매우 빠르게 감소하게 된다. 결과를 살펴보면 비용이 20배 정도 되면 복합추정량의 결과가 SSW에 비해 나쁘게 된다. 따라서 많은 수의 표본을 절사표본층에 배정하면 추정의 효과가 떨어지는 것을 확인 할 수 있다.

Figure 3.1은 각 모집단 형태별, 절사표본층 표본 규모별로 $M-R_{OP}$ 의 RMSE를 그림으로 나타낸 것이다. 비례형 그림을 살펴보면 기준선인 SSW 결과에 비해 작은 부차표본층 정보를 사용할 경우 급속도로 결과가 나빠지는 것을 확인 할 수 있다. 또한 일부 비용이 낮은 경우를 제외하면 복합추정량을 사용하는 것은 올바른 결정이 아닌 것을 확인 할 수 있다. 반면 모집단의 형태가 선형, 블록형, 오목형인 경우에는 n_{SE} 가 20이고 비용이 20배인 경우를 제외하면 모든 경우에서 복합추정량이 우수한 것을 확인 할 수 있다. 또한 최적의 부차표본층은 절사표본층과 가까운 쪽에서 결정 되고 있음도 확인할 수 있다.

4. 자료 분석

자료 분석은 2012년 연탄소비실태조사의 건별 판매량 자료를 이용하였다. 연탄소비실태조사는 약 1600여개의 배달 사업자 자료로, 각 자료는 배달 사업자별로 용도별 판매량과 용도별 판매건수로 구성되어 있으며 전체 판매량의 총합 추정이 본 조사의 목적이다. 이 자료는 Hwang과 Shin (2013)에서도 실제 자료분석에 사용되었다.

Table 4.2. Population and sample sizes for sub-sample strata

절사점 기준	n_{SE}	비용	표본층						
			$s6$	$s5$	$s4$	$s3$	$s2$	$s1$	
80%	0	0배	N	95	86	111	108	84	88
			n	10	10	16	20	23	31
	5	1배	N	98	83	111	108	84	88
			n	10	9	15	19	22	30
		3배	N	98	83	111	108	84	88
			n	9	8	14	17	20	27
	5배	N	98	84	110	108	84	88	
		n	8	8	12	15	18	24	
	7배	N	95	86	111	109	83	88	
		n	7	7	11	14	15	21	
	10	1배	N	98	84	110	108	84	88
			n	9	9	14	18	21	29
3배		N	98	84	110	109	83	88	
		n	8	7	11	15	16	23	
5배	N	98	83	111	109	83	88		
	n	6	5	9	11	12	17		
90%	0	0배	N	195	178	129	111	83	88
			n	23	23	15	14	15	20
	5	1배	N	195	178	129	111	83	88
			n	22	22	14	13	14	20
		3배	N	195	179	128	111	83	88
			n	20	20	12	12	13	18
	5배	N	196	178	128	112	82	88	
		n	18	18	11	11	11	16	
	7배	N	195	179	125	114	83	88	
		n	16	16	9	10	10	14	
	10	1배	N	195	178	129	110	84	88
			n	21	21	13	12	14	19
3배		N	196	178	128	111	83	88	
		n	17	17	10	10	11	15	
5배	N	196	178	130	109	83	88		
	n	13	13	8	7	8	11		

총합 추정을 위한 절사표본설계는 다음과 같다. 먼저 전수층은 판매량이 큰 80개 자료로 구성하였다. 절사표본층의 절사점은 전체 자료의 80%와 90%로 구성하였으며 절사표본층에서는 5개와 10개의 표본을 추출하였다. 따라서 표본층에서는 각각 105, 100개의 표본을 추출하였다. 다음으로 표본층은 6개의 부차표본층을 만들어 부차표본층의 크기에 따른 비교가 가능하도록 구성하였으며 각 층의 경계점은 LH 방법을 사용하여 구하였다. 표본설계 요약은 Table 4.1과 Table 4.2에 수록하였다. 또한 Table 4.3에 분석 결과를 수록하였다.

자료 분석 결과인 Table 4.3을 살펴보면 80%절사점을 사용하고 $n_{SE} = 5$ 인 경우에 비용이 7배가 되어도 복합추정량이 우수한 결과를 주고 있으며, 이 중에서도 본 논문에서 제시한 $M-ROP$ 가 가장 우수한 결과를 주고 있다. 또한 $n_{SE} = 10$ 인 경우를 살펴보면 $n_{SE} = 5$ 인 경우와 유사한 결과를 주고 있다. 최적층은 2 또는 3에서 결정되었다. 90%절사점을 사용하고 $n_{SE} = 5$ 인 경우에는 최적층이 2에서 결정되

Table 4.3. Briquette Consumption Survey results

절사점	n_{SE}	형태	S_I			최적 층 결과		
			SSW	$M-B$	$M-R$	층	$M-B_{OP}$	$M-R_{OP}$
80%	5	bias	3325	1092	1310		359	354
		1배 rbias(%)	6.59	2.16	2.60	3	0.71	0.70
		rmse	3570	2324	2812		1598	1594
		bias	3325	1151	1375		430	423
		3배 rbias(%)	6.59	2.28	2.73	3	0.85	0.84
		rmse	3570	2399	2909		1659	1662
		bias	3325	1159	1398		446	440
		5배 rbias(%)	6.59	2.30	2.77	3	0.88	0.87
		rmse	3570	2412	2937		1683	1680
	bias	3325	1153	1400		392	388	
	7배 rbias(%)	6.59	2.28	2.77	3	0.78	0.77	
	rmse	3570	2392	2941		1700	1697	
	bias	3325	806	888		192	200	
	1배 rbias(%)	6.59	1.60	1.76	2	0.38	0.40	
	rmse	3570	1945	2228		1421	1421	
	bias	3325	894	978		371	363	
	3배 rbias(%)	6.59	1.77	1.94	3	0.74	0.72	
	rmse	3570	2110	2402		1661	1658	
	bias	3325	960	1054		393	385	
	5배 rbias(%)	6.59	1.90	2.09	3	0.78	0.76	
	rmse	3570	2141	2450		1778	1785	
	bias	1522	688	890		274	271	
	1배 rbias(%)	3.02	1.36	1.76	2	0.54	0.54	
	rmse	2030	1695	1933		1457	1457	
bias	1522	711	923		293	291		
3배 rbias(%)	3.02	1.41	1.83	2	0.58	0.58		
rmse	2030	1737	1982		1499	1499		
bias	1522	724	956		286	283		
5배 rbias(%)	3.02	1.43	1.89	2	0.57	0.56		
rmse	2030	1793	2036		1585	1585		
bias	1522	683	941		250	248		
7배 rbias(%)	3.02	1.35	1.86	2	0.50	0.49		
rmse	2030	1836	2054		1658	1658		
bias	1522	613	754		263	260		
1배 rbias(%)	3.02	1.21	1.49	2	0.52	0.52		
rmse	2030	1600	1761		1435	1435		
bias	1522	595	760		151	151		
3배 rbias(%)	3.02	1.18	1.51	1	0.30	0.30		
rmse	2030	1683	1846		1533	1533		
bias	1522	674	908		298	295		
5배 rbias(%)	3.02	1.34	1.80	2	0.59	0.59		
rmse	2030	2044	2156		1932	1932		

(* bias와 rmse의 단위는 10^4 임)

있으며 $M-ROP$ 와 $M-BOP$ 가 거의 같은 결과를 주고 있으나 편향을 기준으로 한다면 $M-ROP$ 가 더 좋은 결과를 주고 있다. 그러나 $n_{SE} = 10$ 이고 비용이 5배가 되는 경우에는 복합추정량의 결과가 SSW 결과와 유사하게 나온다. 따라서 비용이 5배 이상이 된다면 복합추정량을 사용하는 것은 좋은 방법이 아니다. 결론적으로 매우 비용이 큰 경우를 제외한다면 본 연구에서 제안된 복합추정량이 매우 우수한 것을 확인할 수 있다.

5. 결론

본 논문에서는 절사표본층과 표본층의 조사비용을 고려한 복합추정량을 제안하였다. 이를 위하여 Hwang과 Shin (2013)이 제안한 표본설계법의 번거로움을 제거하기 위한 새로운 표본설계법을 제안하였으며, LH 알고리즘을 이용하여 비용을 고려한 층별 표본 수 및 층 경계점을 결정하였다. 또한 Hwang과 Shin (2013)이 제안한 추정량의 일반식을 제안하였으며 비용을 고려한 모의실험을 이용하여 기존에 제안된 복합추정량과 본 연구에서 제안된 복합추정량의 성능을 비교하였다.

모의실험 결과를 살펴보면 먼저 복합추정량을 사용하지 않는 방법인 SSW추정량과의 비교 결과, 모집단 자료가 비례형인 경우에는 절사표본층의 조사비용이 많이 들거나 표본 규모가 큰 경우에는 기존 및 제안된 복합추정량의 성능이 떨어지는 것을 확인하였다. 그러나 비례형을 제외한 다른 모집단 형태에서는 절사표본층 비용이 매우 큰 경우에도 본 연구에서 제안한 복합추정량이 매우 우수한 결과를 주고 있음을 확인하였다.

References

- Baillargeon, S. and Rivest, L.-P. (2011). The construction of stratified designs in R with the package stratification, *Survey Methodology*, **37**, 53–65.
- Benedetti, R., Bee, M. and Espa, G. (2010). A framework for cut-off sampling in business survey design, *Journal of Official Statistics*, **26**, 651–671.
- Elisson, H. and Elvers, E. (2001). Cut-off sampling and estimation, In *Proceedings of Statistics Canada Symposium on Achieving Data Quality in a Statistical Agency: A Methodological Perspective*, 2001.
- Hidirolou, M. A. (1986). The construction of a self-representing stratum of large units in survey design, *The American Statistician*, **4**, 27–31.
- Hidirolou, M. and Srinath, K. P. (1993). Problems associated with designing subannual business surveys, *Journal of Business and Economic Statistics*, **11**, 397–405.
- Hwang, H.-J. and Shin, K.-I. (2013). A composite estimator for cut-off sampling, *Communications for Statistical Applications and Methods*, To appear.
- Hwang, J. M. and Shin, K.-I. (2012). An alternative composite estimator for the take-nothing stratum of the cut-off Sampling, *Communications for Statistical Applications and Methods*, **19**, 13–22.
- Kim, J.-H. and Shin, K.-I. (2011). A composite estimator for the take-nothing stratum of cut-off Sampling, *The Korean Journal of Applied Statistics*, **24**, 1115–1128.
- Lavallee, P. and Hidirolou, M. (1988). On the stratification of skewed populations, *Survey Methodology*, **14**, 33–43.
- 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.
- McDowney, P. (2004). Simulation Result of Probability Proportional Size Sampling for EIA's Monthly Natural Gas Production Survey, a summary note of the Fall 2004 meeting of the American Statistical Association committee on Energy Statistics. <http://circa.europa.eu/irc/dsis/nacepacon/info/data/en/handbook>
- Rao, J. N. K. (2003). *Small Area Estimation*, Wiley Interscience, John Wiley & Sons, New York.

- Rivest, L. P. (2002). A generalization of Lavallee and Hidioglou algorithm for stratifications in business survey, *Survey Methodology*, **28**, 191–198.
- Sarndal, C. E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*, Springer-Verlag, New York.

절사표본 설계에서 비용함수를 고려한 복합추정량

심효선^a · 신기일^{a,1}

^a한국의국어대학교 통계학과

(2013년 10월 14일 접수, 2013년 12월 12일 수정, 2013년 12월 12일 채택)

요약

왜도가 심한 사업체 조사에서는 모집단의 일부를 제외하고 표본을 추출하는 방법인 절사표본추출법이 자주 사용된다. 절사층의 경우 표본 관리가 어렵고 조사비용이 많이 들기 때문에 이를 제외한 조사를 실시함으로써 조사의 효율을 높일 수 있다. 그러나 전체 모집단 추정을 위해서는 절사층의 정확한 총합 추정이 매우 중요하다. 최근 Hwang과 Shin (2013)은 Lavalley와 Hidioglou (1988)가 제안한 LH 알고리즘을 이용하여 표본층을 층화한 후 표본층에서 얻어진 정보와 절사층에서 얻어진 정보를 결합한 복합추정량을 제안하였다. 본 논문에서는 비용함수를 고려한 새로운 표본 설계를 제안하고, 이를 위한 새로운 복합추정량을 제안하였다. 모의실험과 실제 자료 분석을 통하여 본 논문에서 제안한 복합추정량의 우수성을 확인하였다.

주요용어: 붓스트랩 방법, 비추정량, LH 알고리즘, BLUP 추정량.

이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (2012R1A1A2003919).

¹교신저자: (449-791) 한국의국어대학교 통계학과, 경기도 용인시 모현면 산 89, 교수.

E-mail: keyshin@hufs.ac.kr