층화 추출에서 보정추정량에 대한 붓스트랩 분산 추정

Bootstrap Variance Estimation for Calibration Estimators in Stratified Sampling

염 준 근* · 정 영 미**

요 약

무응답 상황하에서 보정 추정량에 대해 관심변수와 강한 상관계수를 가진 보조정보의 수준에 따라 모집단 총합에 대한 추정량과 분산추정량을 붓스트랩 방법을 이용해서 구했다. 이때 존재하는 보조정보의 수준이 표본인 경우와 모집단인 경우로 나누어 모집단 총합에 대한 보정 추정량(calibration estimator)을 구하고, 그에 따른 붓스트랩 분산 추정량을 도출하였다. 또한 테일러 분산 추정량, 잭나이프 분산 추정량과 붓스트랩 분산 추정량의 효율성을 모의 실험을 통해 비교해 보았다.

Abstract

In this paper we study the calibration estimator and its variance estimator for the population total using a bootstrap method according to the levels of an auxiliary information having strong correlation with an interested variable in nonresponse situation. At this point, we find the calibration estimator in case of auxiliary information for population and sample, and then we drive the bootstrap variance estimator of it.

By simulation study we compare the efficiencies with the Taylor and Jackknife variance estimators.

*동국대학교 통계학과 교수

**동국대학교 대학원 통계학과 박사과정 수료

Ⅰ. 서 론

일반적으로 조사에 있어서 무응답은 조사 항목에 발생하는 경우와 조사 단위에 대해 발생하는 경우로 나눌 수 있다. 이러한 두 가지 무응답 상황을 각각 항목 무응답과 단위 무응답으로 정의하고, 무응답을 처리하기 위한 방법으로 표준적으로 대체(imputation) 방법과 가중치조정(weighting adjustment)방법이 있다.

무응답을 제외한 응답자들만의 분석은 무응답 편향으로 인하여 추정량에 심각한 영향을 줄 수 있기 때문에 조사과정에서 발생하는 무응답으로 인한 무응답 편향을 줄이기 위한 여러 학자들의 연구가 이루어져 왔다. 특히 Lundström과 Särndal(1999)은 단위 무응답이 존재하는 경우 관심변수와 강한 상관이 존재하는 보조변수의 기지의 모집단 총합과 표본 총합을 이용하여 관심변수의 총합 추정량과 분산 추정량을 도출하였다. 그러나 이와 같이 도출된 추정량은 선형이 아니기 때문에 직접적으로 정확한 분산 추정량을 구하기란 매우 어려우며, 분산추정에 대한 비 편향 추정방법이 명확하지 않기 때문에 근사적인 방법으로 테일러 전개나 잭나이프 방법, 붓스트랩 방법을 사용하게 된다.

전통적으로 잭나이프 추정방법은 추정량의 편향의 감소와 분산 추정량의 도출을 위해 사용되는 방법으로 잘 알려져 있다. 이에 대한 연구는 이미 많은 학자들에 의해 이루어 졌으며, 최근에 Stukel, Hidiroglou와 Särndal(1996)은 완전응답 하에서 여러 가지 거리함수에 따르는 보정 추정량의 분산 추정에 있어서 잭나이프방법과 테일러 전개 방법을 비교하였다.
Rao와 Wu(1988)는 붓스트랩표본의 크기를 조정한 후, 복합 표본설계하에서 모평균의 붓스트랩 분산추정량에 대한 연구를 하였고, Rao, Wu와 Yue(1992)는 층화다단추출하에서 조사단위에 대한 조사가중치를 조정하여 붓스트랩추정량을 재계산하고, 그에 따른 분산 추정량을 제안했다. 이외에도 여러 학자들이 근사적인 방법으로 분산추정량을 구하는 방법을 연구했으나 이들 모든 방법들은 완전응답하에서 추정량을 구하는 것으로 제한되었다.

손창균과 정훈조(2001)는 단위 무응답이 존재하는 경우에 대해 관심변수와 강한 양의 상관이 있는 보조정보의 수준에 따라 무응답 단위에 대해 가중치 보정을 실시하고, 그에 따르는 분산 추정량을 도출하고, 잭나이프 분산 추정량과의 비교를 통해 추정량의 무응답 편향 감소와 분산 추정량의 안정성에 대해 연구하였다.

본 논문에서는 단위 무응답의 발생시 붓스트랩 방법을 이용해서, 표본과 모집단에 존재하는 보조정보의 수준에 따라 무응답 단위에 대해 가중치 보정을 실시해서 총합에 대한 보정추정량을 구하고, 그에 따르는 붓스트랩 분산 추정량을 도출하였다. 또한 손창균외 1인(2001)의 연구에서 구한 테일러 분산추정량, 잭나이프 분산추정량과의 효율성을 비교해 보았다.

본 논문의 구성은 우선 2장에서는 단위 무응답 상황하에서 보조정보의 수준에 따른 보정추정량을 도출하였고, 3장에서는 보정 추정량에 대한 붓스트랩 분산 추정량을 도출하였다. 4장에서는 몬테칼로 모의 실험을 통해 보정 추정량을 구하고, 본 논문에서 구한 추정량의 분

산과 잭나이프 분산, 테일러 분산에 대한 효율성을 비교하였다. 끝으로 5장에서는 이 연구를 통한 결론을 다루었다.

Ⅱ. 보정 추정량

$U=\{1,2,3,\cdots,N\}$를 크기 N인 유한 모집단이라 하자. 일반적으로 조사로부터 관심 모수에 대한 추정량으로는 표본 평균, 표본 총계, 비등이 있지만, 여기서는 모집단 총합에 대해 추정한다고 하자.

즉, $Y=\sum_U y_k$를 추정하고자 하며, y_k는 모집단의 단위 k에 대해 관심변수 y의 값이고, $\boldsymbol{x}_k=(\boldsymbol{x}_{k1}, \boldsymbol{x}_{k2}, \cdots, \boldsymbol{x}_{kj}, \cdots, \boldsymbol{x}_{kJ})$은 k번째 조사단위와 연관된 보조변수 벡터이다. 모집단으로부터 크기 n인 표본 s를 확률 $p(s)$로 추출한다고 하면, 단위 k가 표본 s에 포함될 확률은 $\pi_k=\sum_{k\in s} p(s)$이며, 단위 k와 l이 동시에 표본 s에 포함될 확률은 $\pi_{kl}=\sum_{k,l\in s} p(s)$이다. 이로부터 $d_k=1/\pi_k$ 와 $d_{kl}=1/\pi_{kl}$는 각각 단위 k에 대한 추출가중치와 단위 k와 l의 동시 추출 가중치로 정의된다. 그때 이용 가능한 보조정보의 수준은 Lundström외 1인(1999)에 의해서 각각 다음과 같이 정의되었다.

(1) 표본 정보 : 모든 $k\in s$에 대해 $\boldsymbol{x}_k$가 기지이다.

(2) 모집단 정보 : $\sum_U \boldsymbol{x}_k$가 기지이며, 또한 모든 $k\in s$에 대해 $\boldsymbol{x}_k$가 기지이다.

또한 뽑힌 표본단위에 대해, 두 가지 경우의 응답 모형을 가정할 수 있다. 즉, 모든 $k\in s$에 대해 $\Pr(k\in r|s)=\theta_k$ 과 모든 $k\neq l\in s$에 대해 $\Pr(k\&l\in r|s)=\theta_{kl}=\theta_k\theta_l$으로 나타낸다. 이때 응답집합을 $r(\subset s)$이라 하고, 이때 응답집합 r의 크기를 $m(\leq n)$이라 하자. Särndal, Swensson과 Wretman(1992)에 의해 앞에서 가정한 보조정보의 수준에 따른 모집단 총합 Y에 대한 추정량은 각각 다음과 같이 정의되었다.

$$\hat{Y}_{s\theta}=\sum_r d_k g_{sk\theta} y_k/\theta_k \tag{2.1}$$

여기서 $g_{sk\theta}=1+q_k(\sum_s d_k \boldsymbol{x}_k-\sum_r d_k \boldsymbol{x}_k/\theta_k)'(\sum_r d_k q_k \boldsymbol{x}_k \boldsymbol{x}_k'/\theta_k)^{-1}\boldsymbol{x}_k$ 이다

또한

$$\hat{Y}_{U\theta}=\sum_r d_k g_{Uk\theta} y_k/\theta_k \tag{2.2}$$

여기서 $g_{Uk\theta}=1+q_k(\sum_U \boldsymbol{x}_k-\sum_r d_k\,\boldsymbol{x}_k/\theta_k)'(\sum_r d_k q_k\,\boldsymbol{x}_k\,\boldsymbol{x}_k'/\theta_k)^{-1}\,\boldsymbol{x}_k$ 이다

그러나 실제로 응답확률 θ_k는 모르는 값이기 때문에 추정해야 한다. 일반적으로 이 값은 특정한 응답 모형의 가정하에서 추정치로 주어진다.

일반적으로 무응답 상황하에서 가중치 조정과 관련하여 모집단 총합에 대한 보정 추정과정은 우선, 이용 가능한 보조정보를 정의하고, 그 다음으로 조정된 새로운 가중치와 그에 따르는 추정량을 계산한다. Deville과 Särndal(1992)의 연구와 Deville, Särndal과 Sautory(1993)의 연구에서는 완전응답 하에서 특정한 거리함수에 따르는 원래의 추출설계 가중치와 가능한 한 근접한 새로운 가중치 w_k를 구하는 방법을 제안하였다.

이들의 연구를 무응답 상황에 적용하여, 새로운 가중치를 w_k라 하면, w_k는 원래의 추출가중치인 d_k에 가능한 한 가장 근접한 값이 되도록 하며, 다음과 같은 보조정보의 수준에 따른 각각의 보정 방정식을 만족한다.

$$\sum_r w_k\,\boldsymbol{x}_k=\sum_U \boldsymbol{x}_k \tag{2.3}$$

$$\sum_r w_k\,\boldsymbol{x}_k=\sum_s d_k\,\boldsymbol{x}_k \tag{2.4}$$

식(2.3)과 (2.4)는 각각 보조정보의 수준이 모집단과 표본에 존재하는 것으로서, 이러한 보정 방정식을 만족하는 새로운 가중치 w_k는 다음과 같은 거리함수를 최소로 한다.

$$G(w_k,d_k)=\sum_r (w_k-d_k)^2/d_k q_k \tag{2.5}$$

보조정보의 수준에 따르는 보정 방정식을 만족한다는 조건하에서, 식(2.5)를 최소로 하는 새로운 가중치는 다음과 같다.

$$w_k=d_k g_k=d_k\left[1+q_k(T-\sum_r d_k\,\boldsymbol{x}_k)'(\sum_r q_k d_k\,\boldsymbol{x}_k\,\boldsymbol{x}_k')^{-1}\,\boldsymbol{x}_k\right] \tag{2.6}$$

여기서, T는 보조변수의 모집단 정보를 이용하는 경우 $T=\sum_U \boldsymbol{x}_k$이고, 보조변수의 표본정보를 이용하는 $T=\sum_s d_k\,\boldsymbol{x}_k$이 된다. 이용하는 보조정보의 수준에 따른 첨자구분은 U와 s를 각각 사용한다. 예를 들어서 모집단 보조정보를 이용하는 경우의 첨자 표현방법은 w_{Uk}와 g_{Uk}을 사용하는 것이다.

따라서 식(2.6)을 총합 추정량에 대입하면 다음과 같은 보정 추정량을 구할 수 있다.

$$\hat{Y}_w=\sum_{k\in r} w_k y_k=\sum_{k\in r} d_k g_k y_k \tag{2.7}$$

보조정보의 수준에 따른 추정량의 첨자는 w대신 각각 wU와 ws을 사용하고, 각각의 보조정보를 이용해서 구한 보정가중치를 적용한다.

Ⅲ. 보정 추정량의 붓스트랩 분산 추정

일반적으로 단순 설계 또는 복합 표본설계를 이용한 표본 조사에서 비선형인 추정량을 다루는 경우가 있다. 비선형 추정량의 분산 추정량을 구하는데 있어서 유용한 방법들로 많이 적용되는 것으로서 잭나이프 추정 방법, 붓스트랩 추정방법이 있다.

따라서, 여기서는 먼저 손창균외 1인(2001)의 연구에서 보인 테일러 분산추정량과 잭나이프 분산 추정을 간단하게 정리하고, 무응답 상황하에서 가중치 조정 과정을 적용한 새로운 보정 추정량에 대해 붓스트랩 분산 추정량을 유도 했다.

1. 테일러 전개에 의한 분산 추정량

특별히 층화 다단 설계와 같은 복합 표본설계를 고려할 경우 테일러 분산 추정량은 다음과 같이 다시 표현할 수 있다.

$$\hat{V}_T(\hat{Y}_w) = \sum_{h=1}^{H} \frac{m_h}{m_h-1} \sum_{i=1}^{m_h}\left[\sum_{k\in r_{hi}} w_{hik}e_{hik} - \frac{1}{m_h}\sum_{i=1}^{m_h}\sum_{k\in r_{hi}} w_{hik}e_{hik}\right] \qquad (3.1)$$

여기서 r_{hi}는 i번째 1차 추출단위 중 h층의 응답 표본이며, w_{hik} 는 1차추출단위 i와 층 h에 있는 응답표본단위 k에 대하여 보조정보를 이용하여 무응답이 보정된 가중치이며, m_h는 층 h에 있는 1차 추출단위의 수이다.

또한 $e_{hik} = y_{hik} - x_{hik}\hat{\beta}_r$는 $\hat{\beta}_r = (\sum_{hik\in r} q_{hik}w_{hik}x_{hik}x_{hik}')^{-1}\sum_{hik\in r} q_{hik}w_{hik}x_{hik}'y_{hik}$인 회귀 추정량과 연관된 추정된 잔차이다. 보조 정보의 수준에 따라 각각의 분산추정량이 도출될 수 있다.

2. 잭나이프 분산 추정량

무응답 상황하에서 가중치 조정과정으로부터 구해진 최종적인 가중치를 적용한 보정 추정량의 잭나이프 분산 추정량은 층화 다단 설계 하에서 다음과 같이 정의되었다.

$$\hat{V}_J(\hat{Y}_w) = \sum_{h=1}^{H} \frac{m_h-1}{m_h} \sum_{i=1}^{m_h}(\hat{Y}_w(hi) - \hat{Y}_w)^2 \qquad (3.3)$$

보조정보의 수준에 따라 추정량의 첨자는 w대신 각각 wU와 ws에 해당되는 추정량으로 대치하면 된다. 식(3.3)에서 $\hat{Y}_w(hi)$ 는 반복 추정량으로서 h 층의 i번째 1차 추출단위를 제거한 후 나머지 추출단위로부터 구한 추정량이다. 즉, $\hat{Y}_w(hi)$ 는 $h=1,2,\cdots,H$, $i=1,2,\cdots,m_h$에 대해 h번째 층으로부터 i번째 1차 추출단위를 제거한 후, 보정추정량 $\hat{Y}_w$를 재 계산한 값이다.

3. 붓스트랩 분산 추정량

여기서는 무응답 상황하에서 가중치 조정과정으로부터 구해진 붓스트랩 보정 가중치를 적용한 보정추정량을 구하고, 층화다단설계하에서 분산추정량을 계산하는 과정을 다음과 같이 4단계에 걸쳐서 실시했다.

[1단계] 각 층 h에 대하여, 일차추출단위 m_h로부터 c_h개의 1차 추출단위를 독립적으로 복원추출 한다. c^*_{hi}는 h층의 i번째 일차추출단위가 붓스트랩 표본 안에 반복 추출되는 횟수라고 하고, $\sum_i c^*_{hi} = c_h$을 만족한다.

[2단계] 각각의 붓스트랩 응답표본 r ($r=1,2,\cdots,R$)에 대하여, 붓스트랩 보정 가중치를 다음과 같이 정의했다.

$$w^*_{hik}(r) = w_{hik}\left[\{1-(c_h/(m_h-1))^{1/2}\}+(c_h/(m_h-1))^{1/2}(m_h/c_h)\, c^*_{hi}(r)\right] \quad (3.4)$$

여기서, w_{hik}는 h층의 i번째 일차 추출단위내의 k번째 응답표본에 부여되는 보정 가중치이다.

[3단계] r번째 붓스트랩 응답표본에 대한 붓스트랩 보정 가중치를 이용하면, 총합에 대한 r번째 붓스트랩 보정추정량은 다음과 같이 얻을 수 있다.

$$\hat{Y}^*_{bw}(r) = \sum_{h=1}^{m_h} w^*_{hik}(r)\; y_{hik} \quad (3.5)$$

여기서, y_{hik}는 h층의 i번째 일차 추출단위의 k번째 응답표본의 값이다.

[4단계] R개의 붓스트랩 응답표본에 대하여 얻어진 보정추정량들을 이용하여 다음과 같은 붓스트랩 분산추정량을 근사적으로 얻는다.

$$\hat{V}_B(\hat{Y}^*_w) = \frac{1}{R}\sum_{r=1}^{R}(\hat{Y}^*_{bw}(r)-\hat{Y}^*_{bw})^2 \quad (3.6)$$

여기서, $\hat{Y}^*_{bw} = \sum_{r=1}^{R}\hat{Y}^*_{bw}(r)/R$이다. 위의 두가지 분산추정량의 경우와 마찬가지로 보조정보의 수준에 따라 추정량의 첨자는 w대신 각각 모집단인 경우에는 wU와 표본인 경우에는 ws에 해당되는 추정량으로 대체하면 된다.

Ⅳ. 모의실험

1. 분산 추정량들의 효율성 비교

관심변수와 보조변수의 모집단을 적당한 크기로 발생시킨 후, 이 모집단을 보조변수의 크

기에 따라 10개의 층으로 층화하여 2단계 추출을 한 후, 몬테칼로 모의실험을 실시하였다. 보조정보의 수준에 따라 보정 추정량을 구하기 위해 우선 보조변수의 모집단 정보는 처음 발생된 보조변수의 값의 총합을 이용하고, 다음으로 표본정보는 1단계 추출단위 중 표본으로 선정된 단위들과 연관된 보조변수를 이용하였다. 관심모수는 모집단 총합이다. 보조 정보의 수준에 따른 각각의 보정추정량 $\hat{Y}_w$ 과 각 방법에 대한 분산 추정량을 계산하고, 각각에 대한 효율성을 비교하였다. 또한 추정에 이용한 보조정보의 수준에 따라 추정량의 첨자에서 각각 w 대신 표본인 경우 ws와 모집단인 경우 wU를 적용했다.

(1) 보정 추정량의 상대 편향(Relative Bias)의 백분율을 다음과 같이 정의하였다.

$$RB(\hat{Y}_w) = \left(\frac{E(\hat{Y}_w) - Y}{Y} \right) \times 100(\%) \qquad (4.1)$$

여기서 $E(\hat{Y}_w) = \frac{1}{K} \sum_{k=1}^{K} \hat{Y}_{w_k}$은 K개 응답 표본에 대하여 구한 보정추정량들의 기대값이다. $\hat{Y}_{w_k}$는 응답표본 k에 대한 보정추정량 $\hat{Y}_w$을 나타낸다.

(2) 분산추정량에 대해 상대 편향의 백분율을 다음과 같이 정의하였다.

$$RB(\hat{V}) = \frac{[E(\hat{V}(\hat{Y}_w)) - V_{true}]}{V_{true}} \times 100(\%) \qquad (4.2)$$

여기서 $E(\hat{V}(\hat{Y}_w)) = \frac{1}{K} \sum_{k=1}^{K} \hat{V}_k(\hat{Y}_w)$, $V_{ture} = \frac{1}{K} \sum_{k=1}^{K} (\hat{Y}_{w_k} - E(\hat{Y}_w))^2$이다.

$\hat{V}_k(\hat{Y}_w)$은 응답표본 k에 대하여, 세 가지 방법 각각에 대한 분산추정량 값이다.

(3) 분산 추정량 자체의 변동을 알아보기 위해서, 보조정보의 수준에 따른 3가지 분산 추정량의 변이계수(Coefficient of Variation)의 백분율을 다음과 같이 정의한다.

$$CV(\hat{V}) = \frac{\sqrt{\frac{1}{K} \sum_{k=1}^{K} (\hat{V}_k(\hat{Y}_w) - V_{true})^2}}{V_{true}} \times 100(\%) \qquad (4.3)$$

V. 결 론

단위 무응답이 존재하는 경우 보조정보의 수준에 따라 무응답 단위에 대해 가중치 보정을 실시한 후, 총합에 대한 보정추정량을 구하고 그에 따르는 분산추정량을 붓스트랩 방법을 이용한 근사적인 방법으로 도출하였다. 추정 단계에서 이용된 보조정보의 수준은 모든 $k \in s$에

대해 x_k가 기지인 표본에 대한 보조정보와 $\sum_U x_k$가 기지이며, 또한 모든 $k \in s$에 대해 x_k가 기지인 모집단 전체에 대한 보조정보이다.

보정가중치를 사용함으로써 보정추정량의 편향이 무시할 만큼 작게 나왔으며, 표본수가 적은 경우에 분산을 추정하는데 있어서 편향을 줄여주는 효과가 있음을 모의실험 결과를 통해서 살펴보았다.

참고 문헌

1. 손창균. 정훈조(2001). 무응답 상황하에서 테일러 분산 추정량과 잭나이프 분산 추정량에 대한 연구, *품질경영학회 춘계발표 논문집*.
2. Deville, J. C., and Särndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, pp.376-382.
3. Deville, J. C., Särndal, C. E., and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, pp. 1013-1020.
4. Lundström, S., and Särndal, C. E. (1999). Calibration as a Standard Method for Treatment of Npnresponse. *Journal of Official Statistics*, 15, pp.305-327.
5. Rao, J. N. K., and Wu, C. F. J. (1988). Resampling inference With Complex Survey Data. *Journal of the American Statistical Association*, 83, pp.231-241.
6. Rao, J. N. K., and Wu, C. F. J., and Yue, K. (1992). Some Recent Work on Resampling Methods for Complex Surveys. *Survey Methodology*, 18, pp.209-217.
7. Särndal, C. E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
8. Stukel, D. M., Hidiroglou, M. A., and Särndal, C. E. (1996). Variance Estimation for Calibration Estimators: A Comparison of Jackknifing versus Taylor Linearization. *Survey Methodology*, 22, pp.117-125.