The Korean Communications
in Statistics Vol. 10 No. 3, 2003
pp. 665-678

Recalibration Estimation for Unit Nonresponse at the Two Levels Auxiliary Information

Joon Keun Yum[1] Chang Kyoon Son[2] Young Mee Jeung[3]

Abstract

In this paper we suggest the new calibration estimator, which is called to the recalibration estimator, and its variance estimator using two-phase sampling technique according to the auxiliary information having strong correlation with the variable of interest under the unit nonresponse. In this unit nonresponse situation, an available information may exists at the level of whole population or the first-phase sample. The proposed recalibration estimator derives from the first and second phase weights respectively.

Keywords : Unit nonresponse, Auxiliary information, Two-phase sampling, Recalibration

1. 서론

일반적으로 조사에 있어서 무응답은 조사 항목에 발생하는 경우와 조사 단위에 대해 발생하는 경우로 나눌 수 있고, 이러한 두 가지 무응답 상황을 각각 항목 무응답(item nonresponse)과 단위 무응답(unit nonresponse)으로 정의하고 있다(Kalton과 Kasprzyk, 1986). 이러한 무응답이 존재하는 경우 무응답자를 제외한 응답자들만을 이용한 분석은 무응답 편향을 발생시키기 때문에 무응답 편향을 줄이기 위한 노력을 필요로 한다. 항목무응답과 단위무응답을 처리하기 위한 대표적인 방법으로 각각 대체법과 가중치 조정방법을 들 수 있다. 특히 조사단위(survey unit)에 발생하는 단위무응답에 대한 처리방법으로 가중치 조정방법중의 하나인 보정 추정방법을 적용한 방법들이 제안되었으며, 이러한 단위무응답 처리와 관련하여 Särndal과 Swensson(1987)은 층화이중추출 방법을 적용한 추정량을 제안하였고, Lundström과 Särndal(1999)(:L-S)은 관심변수와 강한 상관이 존재하는 보조변수의 기지의 모집단과 표본수준의 정보를 이용하여 관심변수의 총합과 그에 대한 분산 추정량을 도출하였다. 또한 손창균 등(2000)은 기지의 모집단 총합에 대한 정보뿐만 아니라 보조변수에 대한 기지의 분산 정보를 가지고 관심변수에 대한 분산 추정치를 구하였고, 염준근, 정영미(2002)는 단위 무응답이 존재할 때, 모집단 수준의 보조정보를 이용하여 이중추출접근방법으로 응답확률과 추출가중치를 보정(calibration)함으로서 관심 모수에 대한 추정량을 구하였다.

1) Professor, Department of Statistics, Dongguk University, Seoul, 100-715, Korea.
E-mail : joonkeun@dgu.edu
2 Full-time Lecturer, Department of Computer Science, Dongshin University, Jeonnam, 525-714, Korea.
E-mail : ckson85@dsu.ac.kr
3) Lecturer, Department of Statistics, Dongguk University, Seoul, 100-715, Korea
E-mail : jym007@orgio.net

Hidiroglou와 Deville(1995), Hidiroglou와 Särndal(1998)등은 완전응답 하에서 가중치 조정을 위해 이중추출(two-phase sampling) 기법을 적용하였다. 이들의 연구에 따르면 우선 1단계에서 모집단 보조정보를 이용하여 추출가중치를 조정한 다음 조정된 가중치를 2단계의 최종 가중치에 적용하여 총합 추정량을 도출하였다.

본 논문에서는 이들의 연구를 바탕으로 최종 조사단계인 2단계에서 단위 무응답이 발생한 경우 1단계에서는 추출가중치만을 조정하고, 조정된 1단계 가중치를 2단계에서 재보정(recalibration) 과정을 통해 최종가중치로서의 응답확률을 보정하는 방법을 제안하고자 한다. 특별히 추정과정에서 단순임의 추출설계에 비해 보다 현실적이고 시간과 비용 측면에서 효율적인 집락추출설계를 고려하였고, 보정 추정과정에서 이용 가능한 보조정보의 형태가 모집단과 표본에 대해 일치하지 않는 경우, 즉 이중보조정보를 사용한 경우에 대해서 효율성을 살펴보았다.

논문의 구성은 2절에서는 무응답이 존재하는 경우 L-S 추정량과 집락추출설계에서 이중추출방법으로 모집단 총합에 대한 추정량을 다루었으며, 3절에서는 각 단계별로 보조정보를 이용하여 추출가중치와 응답확률을 보정한 최종적인 보정추정량을 도출하였다. 4절과 5절에서는 분산추정량을 도출하고, 몬테칼로 모의 실험을 통해 제안된 추정량과 분산추정량의 정도를 비교하였다. 마지막으로 6절에서는 결론과 향후 연구과제를 다루었다.

2. 보정 추정량

2.1 Lundström과 Särndal의 보정 추정량

모집단으로부터 $p(s)$의 확률로 추출한 크기 n인 표본을 s라 하자. 그러면 모집단 단위가 표본에 포함될 확률들은 각각 $\pi_k = p(k \in s)$와 $\pi_{kl} = p(k \bigcap l \in s)$이다. π_k의 역수인 $\pi_k^{-1} = d_k$는 단위 k에 대한 설계가중치이며, 또한 $\pi_{kl}^{-1} = d_{kl}$이다. 조사과정에서 무응답이 발생하여, 추출된 표본에 대해 크기가 m인 응답자들의 집합 r을 얻었다고 하자. 여기서 $r \subseteq s$이고, $m \leq n$이다.

관심변수와 강한 상관을 가진 보조정보는 추출오차와 무응답 편향을 충분히 감소시킬 수 있다. 따라서 보조변수 벡터 $\boldsymbol{x}$를 가정하고, 이 벡터는 관심변수 y와 강한 상관을 가진다고 하자. k번째 단위에 대한 보조변수 벡터 값을 $\boldsymbol{x}_k$라 하자. Särndal, Swensson 그리고 Wretman(1992)은 다음과 같은 2가지 경우의 "정보의 차원"을 고려하여, 표본에 대한 정보를 "*Info-S*"와 모집단에 대한 정보를 "*Info-U*"라고 표현하였으며, 그에 따른 보조변수를 다음과 같이 정의하였다.

i) 표본에 대한 정보: $\boldsymbol{x}_k$는 모든 $k \in s$에 대해 기지이다.

ii) 모집단에 대한 정보: $\sum_U \boldsymbol{x}_k$가 기지이며, 더욱이 $\boldsymbol{x}_k$가 모든 $k \in s$에 대해 기지이다.

ii)의 경우에 대한 정보는 모집단 차원의 정보를 의미하며, 단순히 표본차원의 정보를 나타내는 i)의 경우를 확장한 것이다.

Särndal 등(1992)에 따르면, 응답 분포 $q(r|s)$를 $\Pr(k \in r|s) = \theta_k$와 $\Pr(k \& l \in r|s) = \theta_{kl}$인 기지의 응답확률을 가정하였다(이러한 확률들은 실현된 표본 s와는 독립인 확률로 가정한다). 이러한 가정하에서 표본과 모집단에 대한 보조변수의 정보에 따라 추정량은 각각 식(2.1)과 식(2.3)과 같다.

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

여기서,

$$g_{sk\theta} = 1 + q_k(\sum_s d_k \boldsymbol{x}_k - \sum_r \boldsymbol{x}_k/\theta_k)'(\sum_r d_k q_k \boldsymbol{x}_k \boldsymbol{x}_k'/\theta_k)^{-1} \boldsymbol{x}_k . \tag{2.2}$$

그리고

$$\hat{Y}_{ssw,U\theta} = \sum_r d_k g_{Uk\theta} y_k/\theta_k \ . \tag{2.3}$$

여기서,

$$g_{Uk\theta} = 1 + q_k(\sum_U d_k \boldsymbol{x}_k - \sum_r \boldsymbol{x}_k/\theta_k)'(\sum_r d_k q_k \boldsymbol{x}_k \boldsymbol{x}_k'/\theta_k)^{-1} \boldsymbol{x}_k \ . \tag{2.4}$$

그러나 실제로는 응답 확률 θ_k는 미지이며, 따라서 임의의 값인 $\hat{\theta}_k$로 대체해야 한다.

이와 관련하여 Little(1986)과 Ekholm과 Laaksonen(1991)은 우선 관련된 응답 모형을 설정하고, 그 다음에 미지의 응답확률을 추정하는 방법을 제안하였다.

이러한 절차로부터 전형적인 모집단 총합 Y에 대한 추정량은 다음과 같다.

$$\hat{Y} = \sum_r d_k \nu_{1k} \nu_{2k} y_k \ . \tag{2.5}$$

여기서 $\nu_{1k} = 1/\hat{\theta}_k$이며, ν_{2k}는 식(2.2)나 식(2.4)의 g-가중치와 같고, g-가중치의 θ_k를 $\hat{\theta}_k$로 대치한다.

모집단 총합에 대한 일반적인 보정 추정과정은 보조정보를 정의한 후, 보정 가중치 w_k를 계산하고, Y의 w-가중추정량 $\hat{Y}_w = \sum_r w_k y_k$를 계산한다. w_k는 가능한 한 d_k에 근접하도록 하며, 이 값은 표본과 모집단에 대해 다음과 같은 보정 방정식을 만족한다.

$$\sum_r w_k \boldsymbol{x}_k = \sum_s d_k \boldsymbol{x}_k \ . \tag{2.6}$$

$$\sum_r w_k \boldsymbol{x}_k = \sum_U \boldsymbol{x}_k \ . \tag{2.7}$$

완전응답의 경우에 있어서 Deville과 Särndal(1992)은 보정 방정식인 $\sum_s w_k^0 \boldsymbol{x}_k = \sum_U \boldsymbol{x}_k$의 관점에서 가능한 한 d_k에 가까운 보정 가중치인 w_k^0를 구하여 $\hat{Y}_{DS} = \sum_s w_k^0 y_k$의 형태인 추정량을 도출하였다. Lundström과 Särndal(1999)는 동일한 보정 기법을 이용하지만, 응답집합을 고려한 식(2.6)과 식(2.7)을 만족하는 새로운 가중치 w_k를 구하였다. 이 가중치는 다음의 거리함수를 최소로 한다.

$$\sum_r (w_k - d_k)^2/d_k q_k \ . \tag{2.8}$$

여기서 q_k는 특정한 양의 인자로서 추정량의 형태를 결정한다.

완전응답, 즉 $r=s$인 경우 이 거리함수를 이용한 추정량은 일반화 회귀 추정량의 형태와 같다.

이들에 따르면, 모집단에 대해 식(2.7)의 조건하에서 거리함수의 식(2.8)을 최소로 하는 w-추정량은 다음과 같다.

$$\hat{Y}_{wU} = \sum_r d_k \nu_{Uk} y_k \ . \tag{2.9}$$

여기서 모든 $k \in r$에 대해,

$$\nu_{Uk} = 1 + q_k\Big(\sum_U \boldsymbol{x}_k - \sum_r d_k \boldsymbol{x}_k\Big)'\Big(\sum_r d_k q_k \boldsymbol{x}_k \boldsymbol{x}_k'\Big)^{-1} \boldsymbol{x}_k \ . \tag{2.10}$$

표본에 대해, 식(2.6)의 조건하에서 거리함수의 식(2.8)을 최소로 하는 w-추정량은 다음과 같다.

$$\hat{Y}_{us} = \sum_r d_k \nu_{sk} y_k \ . \tag{2.11}$$

여기서 모든 $k \in r$에 대해,

$$\nu_{sk} = 1 + q_k\Big(\sum_s \boldsymbol{x}_k - \sum_r d_k \boldsymbol{x}_k\Big)'\Big(\sum_r d_k q_k \boldsymbol{x}_k \boldsymbol{x}_k'\Big)^{-1} \boldsymbol{x}_k \ . \tag{2.12}$$

2.2 집락 추출 설계에서 무응답 보정 추정량

집락 추출설계 하에서 유한 모집단의 개체는 $U=\{1,2,\cdots,k,\cdots,N\}$으로 표시하자. 이때 모집단을 N_I개의 집락으로 분할하고, 이들을 각각 $U_1, U_2, \cdots, U_i, \cdots, U_{N_I}$ 라 하자. 그러면, 각 집락은 $U_I=\{1,2,\cdots,i,\cdots,M_I\}$으로 표현할 수 있다. i번째 집락 U_i의 크기를 M_i라 하면, $U=\bigcup_{i=1}^{N_I} U_i$와 $N=\sum_{i=1}^{N_I} M_i$이 된다. A가 원소들의 집합일 때 $\sum_{k\in U_I}$을 $\sum_A$으로 나타내고, B가 집락들의 집합일 때, $\sum_{i\in B}$을 $\sum_B$으로 간단하게 표현하자. 크기가 n_I인 확률표본 $s_I(\subset U_I)$는 U_I로 부터 추출설계 $p_I(\cdot)$에 의해서 추출된다. 이때, s는 추출된 집락내의 모집단 개체들로써 표본, 즉 $s=\bigcup_{i\in s_I} U_i$을 나타내고, s의 크기는 $n=\sum_{s_I} M_i$이다.

i번째 집락이 표본에 포함될 확률은 $\pi_{i(I)}=\sum_{s_I \ni i} p_I(s_I)$이고, 단위 i와 j가 동시에 표본 집락에 포함될 확률은 $\pi_{ij(I)}=\sum_{s_I \ni i,j} p_I(s_I)$이다. 이때, $\pi_{ii(I)}=\pi_{i(I)}$이고, 단위 i의 1단계 추출가중치는 $d_{i(I)}=1/\pi_{i(I)}$이 된다. 또한 각 집락내 개체들의 포함확률은 표본 s가 추출된 집락안의 모든 원소를 포함하기 때문에, U_i의 모든 k에 대하여 $\pi_k=\Pr(k\in s)=\Pr(i\in s_I)=\pi_{i(I)}$이고, k와 l이 같은 집락 U_i에 속하면 2차 포함확률은 $\pi_{kl}=\Pr(k\,\&\,l\in s)=\Pr(i\in s_i)=\pi_{i(I)}$이 된다. 또한 k와 l이 서로 다른 집락 U_i와 U_j에 속하면 $\pi_{kl}=\Pr(k\,\&\,l\in s)=\Pr(i\,\&\,j\in s_I)=\pi_{ij(I)}$으로 표현된다. 이와 같은 추출설계로부터 추정하고자 하는 관심 모수는 모집단 총합 $\hat{Y}=\sum_U y_k=\sum_{U_I} \hat{Y}_i$이고, $\hat{Y}_i=\sum_{U_i} y_k$ 로서 i번째 집락의 총합이다.

표본 집락의 원소중 조사과정에서 무응답이 발생하여, i번째 집락내에서 크기 $m_i(\le M_i)$인 응답집합 $r_i(i\subset s_i)$를 가정하자. 즉 $r=\bigcup_{i=1}^{s_I} r_i$과 $m=\sum_{i=1}^{s_I} m_i$이다. 결과적으로 $k\in r$단위에 대해서만 관심변수 y_k값을 관찰하게 된다.

조사단위에 대한 무응답을 보정하기 위해서 관심변수와 강한 상관이 있는 보조변수를 이용하며, 이때 이용 가능한 보조정보를 정의하면 $\boldsymbol{x}=(\boldsymbol{x}_1', \boldsymbol{x}_2')'$로 나타내고, k-번째 개체에 관한 보조정보는 $\boldsymbol{x}_k=(\boldsymbol{x}_{1k}', \boldsymbol{x}_{1k}')'$와 같다.

그러면, 무응답 단위에 대한 가중치를 조정하기 위해 이용 가능한 보조정보의 형태를 다음과

같이 정의하자.

[정의 1] 모집단 보조정보로 이용되는 보조변수는 $\boldsymbol{x}_{1k}$로서 모집단 총합 $\sum_U \boldsymbol{x}_{1k}$이 기지이고, 표본 s_I내의 모든 개체 값 $\boldsymbol{x}_k = (\boldsymbol{x}_{1k}', \boldsymbol{x}_{2k}')'$은 기지이다. 이때, $\boldsymbol{x}_{2k}$는 표본 보조정보로 이용되는 보조변수이다. 이러한 보조정보를 *Info-D*로 정의하자.

무응답 가정 하에서 응답분포 $q(r|s_I)$는 다음과 같은 기지의 응답 확률에 대응되는 기지인 분포로 가정하고, 또한 이들 확률은 표본 s_I내의 개체들에 대하여 독립이라고 가정하자.

$$\Pr(k\in r|s_I)=\theta_k \ , \ \Pr(k\&l\in r|s_I)=\theta_{kl} \ .$$

여기서, $k\in r(i\in s_I)$이다.

Hidiroglou와 Särndal(1995)에 따르면 보조정보 수준이 *Info-D*인 경우 무응답 총합 추정량은

$$\hat{Y}_{CL,D\theta}=\sum_{s_I} d_{i(I)} \sum_{r} g_{Dk\theta} y_k/\theta_k \tag{2.12}$$

으로 표현된다.

이때, 최종 가중치인 $g_{Dk\theta}$는 다음의 1단계 가중치 g_{1k}와 2단계 가중치 $g_{sk}^{D\theta}$ 의 결합인 $g_{Dk}=g_{1k}g_{sk}^{D}$ 으로서 Hidiroglou와 Särndal(1995)의 "승법식(multiplication formula)"을 적용하면

$$g_{Dk\theta}=1+q_k\Big(\sum_{s_I} d_{i(I)} \sum_{U_i} g_{1k}\, \boldsymbol{x}_k-\sum_{s_I} d_{i(I)} \sum_{r} \boldsymbol{x}_k/\theta_k\Big)'\Big(\sum_{s_I} d_{i(I)} \sum_{r} q_k\, \boldsymbol{x}_k\, \boldsymbol{x}_k'/\theta_k\Big)^{-1} \boldsymbol{x}_k \tag{2.14}$$

이고, 또한 1단계 추출 가중치는 다음과 같다.

$$g_{1k}=1+q_{1k}\Big(\sum_{U} \boldsymbol{x}_{1k}-\sum_{s_I} d_{i(I)} \sum_{U_i} \boldsymbol{x}_{1k}\Big)'\Big(\sum_{s_I} d_{i(I)} \sum_{U_i} q_{1k}\, \boldsymbol{x}_{1k}\, \boldsymbol{x}_{1k}'\Big)^{-1} \boldsymbol{x}_{1k} \ . \tag{2.15}$$

이때 2단계에서 조정된 가중치는

$$g_{sk}^{D\theta}=1+q_{2k}\Big(\sum_{s_I} d_{i(I)} \sum_{U_i} g_{1k}\, \boldsymbol{x}_k-\sum_{s_I} d_{i(I)} \sum_{r} q_{2k} g_{1k}\, \boldsymbol{x}_k/\theta_k\Big)\Big(\sum_{s_I} d_{i(I)} \sum_{r} q_{2k} g_{1k}\, \boldsymbol{x}_k\, \boldsymbol{x}_k'/\theta_k\Big)^{-1} \boldsymbol{x}_k \tag{2.16}$$

으로 표현할 수 있다.

3. 재보정(recalibration) 방법을 이용한 가중치 조정

2절에서 다룬 L-S추정량의 경우 이용 가능한 보조정보의 수준이 같은 경우에 적용가능하며, 또한 설계가중치와 응답확률을 동시에 보정하는 방법을 사용하였다. 그러나 보조정보의 수준이 서로 다른 경우 설계가중치와 응답확률을 동시에 보정하기는 사실상 불가능하며, 따라서 본 논문에서 제안한 단계적인 가중치 조정 방법인 재 보정 방법을 사용해야 한다. 이와 관련하여 3절에서는 재 보정 방법을 이용한 보정추정량을 구하기 위해서 필요한 보정 방정식을 정의하여 그에 따르는 보정 추정량을 구하고자 한다.

[정의 2] *Info-D*의 수준에서 보조정보를 이용한 보정 방정식은 다음과 같다.

$$\sum_{U} \boldsymbol{x}_{1k} = \sum_{s_I} d_{i(I)} \sum_{U_i} g_{1k}\, \boldsymbol{x}_{1k} = \sum_{s_I} \sum_{U_i} w_{1k}\, \boldsymbol{x}_{1k} \ , \tag{3.1a}$$

$$\sum_{s_I} d_{i(I)} \sum_{U_i} g_{1k}\, \boldsymbol{x}_k = \sum_{s_I} d_{i(I)} \sum_{r} g_{Dk}\, \boldsymbol{x}_k = \sum_{s_I}\sum_{U_i} w_{1k}\, \boldsymbol{x}_{1k} = \sum_{r} w_{Dk}\, \boldsymbol{x}_k \ . \tag{3.1b}$$

여기서, $\boldsymbol{x}_k = (\ \boldsymbol{x}_{1k}',\ \boldsymbol{x}_{2k}')'$이다.

식(2.12)로부터 모집단 총합을 구하기 위해서는 응답확률을 보정해야 하며, 만일 응답확률이 보정된다면, 이 값을 이용해서 최종 g-가중치를 구할 수 있을 것이다. 이때 응답확률과 g-가중치를 이용 가능한 보조정보의 수준에 따라 2단계의 가중치 조정과정을 통해 조정하고자 한다.

즉, 이용 가능한 보조정보가 모집단과 표본으로 각각 주어졌을 때(*Info-D*), 모집단 정보를 이용하는 1단계과정과 1단계에서 조정된 가중치와 표본 정보를 이용하는 2단계과정으로 구분하여 전개하고자 한다. 이때, 각 단계에서 존재하는 보조정보의 수준에 따라 보정방정식을 만들 수 있으며, 보정방정식의 조건하에서 일반화 최소제곱(generalized least square: GLS)거리함수를 최소로 하는 보정가중치를 구하는 절차를 단계별로 설명하면 다음과 같다.

[단계 1] s_I부터 U_I까지 집락내의 모든 표본단위들의 가중치를 조정한다.

식(3.1a)로부터 1단계 보정방정식 $\sum_{s_I} d_{i(I)} \sum_{U_i} g_{1k}\, \boldsymbol{x}_{1k} = \sum_{s_I}\sum_{U_i} w_{1k}\, \boldsymbol{x}_{1k} = \sum_{U} \boldsymbol{x}_{1k}$의 조건하에서 모든 $k \in U_i (i \in s_I)$에 대하여 GLS 거리함수

$$\sum_{s_I}\sum_{U_i} \frac{(\ w_{1k} - d_{i(I)})^2}{d_{i(I)} q_{1k}} \tag{3.2}$$

를 최소화하는 1단계 보정가중치 $w_{1k} = d_{i(I)} g_{1k}$를 계산하며, g-가중치는 식(2.3)과 같다.

■

[단계 2] 1단계에서 조정한 가중치를 이용해서 s_I내의 응답단위 r_i까지 조정함으로써 최종적으로 가중치를 보정한다.

2단계에서 이용되는 보정방정식 $\sum_{s_I}\sum_{r} w_{Dk}\, \boldsymbol{x}_k = \sum_{s_I}\sum_{U_i} w_{1k}\, \boldsymbol{x}_k$을 만족한다는 조건 하에서 GLS 거리함수 $\sum_{s_I}\sum_{r} \frac{(w_{Dk} - w_{1k})^2}{w_{1k} q_{2k}}$를 최소로 하는 최종 보정 가중치를 구하면 다음과 같다.

$$\begin{aligned} w_{Dk} &= d_{i(I)} g_{Dk} \\ &= d_{i(I)} \Big\{ 1 + q_k \Big(\sum_{s_I} d_{i(I)} \sum_{U_i} g_{1k}\, \boldsymbol{x}_k - \sum_{s_I} d_{i(I)} \sum_{r} \boldsymbol{x}_k \Big) \Big(\sum_{s_I} d_{i(I)} \sum_{r} q_k\, \boldsymbol{x}_k\, \boldsymbol{x}_k' \Big)^{-1} \boldsymbol{x}_k \Big\} \ . \end{aligned} \tag{3.3}$$

여기서, $q_k = q_{1k} q_{2k}$이고, g-가중치는 $g_{Dk} = g_{1k} g_{sk}^D$이며, 이때 g_{sk}^D는 다음과 같다.

$$g_{sk}^D = 1 + q_{2k} \Big(\sum_{s_I} d_{i(I)} \sum_{U_i} g_{1k}\, \boldsymbol{x}_k - \sum_{s_I} d_{i(I)} \sum_{r} g_{1k}\, \boldsymbol{x}_k \Big) \Big(\sum_{s_I} d_{i(I)} \sum_{r} q_{2k} g_{1k}\, \boldsymbol{x}_k\, \boldsymbol{x}_k' \Big)^{-1} \boldsymbol{x}_k \ . \tag{3.4}$$

결과적으로 최종적인 보정추정량은 다음과 같다.

$$\hat{Y}_{wD} = \sum_{k \in r} w_{Dk} y_k = \sum_{s_I} d_{i(I)} \sum_{r} g_{Dk} y_k \ . \tag{3.5}$$

■

4. 분산 추정

4.1 Lundström과 Särndal의 분산 추정량

Särndal 등(1992)은 각각의 수준에 따른 추정량 $\hat{Y}_{SSW,s\theta}$와 $\hat{Y}_{SSW,U\theta}$ 에 대한 분산추정량을 제안하였고, 이를 다음과 같은 각각의 단위에 대해 독립인 응답확률에 따라 분산추정량에 적용할 수 있다.

$$\Pr(k\&l|s) = \theta_{kl} = \theta_k\theta_l \text{ , 모든 } k \neq l \text{에 대해.} \tag{4.1}$$

이와 같은 가정으로부터 이용 가능한 보조정보의 수준에 따라 표본보조정보를 이용한 보정추정량의 분산 추정량은 다음과 같다.

$$\begin{aligned}\hat{V}(\hat{Y}_{SSW,s\theta}) = \sum_r\sum(d_kd_l - d_{kl})\left(\frac{y_k}{\theta_k}\right)\left(\frac{y_l}{\theta_l}\right) - \sum_r d_k(d_k-1)\left(\frac{y_k}{\theta_k}\right)^2(1-\theta_k) \\ + \sum_r d_k^2(1-\theta_k)\left(\frac{g_{sk\theta}e_{k\theta}}{\theta_k}\right)^2 .\end{aligned} \tag{4.2}$$

여기서 $g_{sk\theta}$는 식(2.2)이며, $e_{k\theta} = y_k - \boldsymbol{x}_k'\boldsymbol{B}_{r\theta}$ 이고, $\boldsymbol{B}_{r\theta} = (\sum_r d_kq_k\boldsymbol{x}_k\boldsymbol{x}_k'/\theta_k)^{-1}\sum_r d_kq_k\boldsymbol{x}_ky_k/\theta_k$이다.

또한 모집단 보조정보를 이용한 보정 추정량의 분산 추정량은 다음과 같다.

$$\begin{aligned}\hat{V}(\hat{Y}_{SSW,U\theta}) = \sum_r\sum(d_kd_l - d_{kl})\left(\frac{g_ke_{k\theta}}{\theta_k}\right)\left(\frac{g_le_{l\theta}}{\theta_l}\right) - \sum_r d_k(d_k-1)\left(\frac{g_ke_{k\theta}}{\theta_k}\right)^2(1-\theta_k) \\ + \sum_r d_k^2(1-\theta_k)\left(\frac{g_{sk\theta}e_{k\theta}}{\theta_k}\right)^2 .\end{aligned} \tag{4.3}$$

여기서 $g_k = 1 + q_k\left(\sum_U \boldsymbol{x}_k - \sum_s d_k\boldsymbol{x}_k\right)'\left(\sum_s d_k\boldsymbol{x}_k\boldsymbol{x}_k'\right)^{-1}\boldsymbol{x}_k$이며, $g_{sk\theta}$는 식(2.2)와 같다.

Lundström과 Särndal(1999)에 따르면, 추출설계 $p(s)$와 응답 분포 $q(r|s)$하에서 일반적인 w-가중 추정량 $\hat{Y}_w$의 MSE 가 다음과 같음을 보였다.

$$MSE_{pq}(\hat{Y}_w) = V_{SAM} + V_{NR} + 2Cov_p(\hat{Y}_s, B_{NR|s}) + E_p(B_{NR|s}^2) . \tag{4.4}$$

여기서, $V_{SAM} = V_p(\hat{Y}_s)$이며, $V_{NR} = E_pV_q(\hat{Y}_w|s)$는 무응답 오차의 분산이고, $B_{NR|s}$는 무응답 편향을 나타낸다. 또한 $Cov_p(\hat{Y}_s, B_{NR|s})$는 주어진 추출설계 하에서 $\hat{Y}_s$와 $B_{NR|s}$의 공분산이다.

만일 강한 보조정보를 가정하면, 무응답 편향은 $B_{NR|s_l} \approx 0$ 가 되며, 식(4.4)는 다음과 같이 축소 될 수 있다.

$$V_{pq}(\hat{Y}_w) = V_{SAM} + V_{NR} . \tag{4.5}$$

보조정보의 수준에 따라 각각의 추정량 $\hat{Y}_{ws}$와 $\hat{Y}_{wU}$에 대한 분산추정량은 각각 다음과 같다 (Lundström과 Särndal(1999)).

$$\begin{aligned}\hat{V}_{pq}(\hat{Y}_{ws}) = \sum_r\sum(d_kd_l - d_{kl})(\nu_{sk}y_k)(\nu_{sl}y_l) - \sum_r d_k(d_k-1)\nu_{sk}(\nu_{sk}-1)y_k^2 \\ + \sum_r d_k^2\nu_{sk}(\nu_{sk}-1)f_k^2e_k^2 ,\end{aligned} \tag{4.6}$$

그리고,

$$\hat{V}_{pq}(\hat{Y}_{wU}) = \sum_r\sum(d_k d_l - d_{kl})(g_k f_k \nu_{sk} e_k)(g_l f_l \nu_{sl} e_l) - \sum_r d_k(d_k-1)\nu_{sk}(\nu_{sk}-1)(g_k f_k e_k)^2 + \sum_r d_k^2 \nu_{sk}(\nu_{sk}-1) f_k^2 e_k^2 \ . \tag{4.7}$$

이다. 여기서 f_k는 자유도의 손실을 고려한 조정인자이다.

4.2 집락 추출설계에서 $\hat{Y}_{CL,D\theta}$의 분산 추정량

식(4.4)의 재보정 추정량 $\hat{Y}_{wD}$의 평균제곱오차(Mean Square Error: MSE)는 다음과 같다.

$$\begin{aligned} MSE_{pq}(\hat{Y}_{wD}) &= E_{pq}[\hat{Y}_{wD} - E_{pq}(\hat{Y}_{wD})]^2 \\ &= V_{pq}(\hat{Y}_{wD}|s_I) + (B_{pq}(\hat{Y}_{wD}))^2 \\ &= E_p V_q(\hat{Y}_{wD}|s_I) + V_p(E_q(\hat{Y}_{wD}|s_I)) + (B_{pq}(\hat{Y}_{wD}))^2 \\ &= E_p V_q(\hat{Y}_{wD}|s_I) + V_p(B_{NR|s_I}) + V_p(\hat{Y}_{s_I}) + 2Cov_p(\hat{Y}_{s_I}, B_{NR|s_I}) + (B_{pq}(\hat{Y}_{wD}))^2 \ . \end{aligned} \tag{4.8}$$

식(4.8)에서 오른쪽 두 번째 항은

$$\begin{aligned} V_p(B_{NR|s_I}) &= E_p(B_{NR|s_I}^2) - [E_p(B_{NR|s_I})]^2 \\ &= E_p(B_{NR|s_I}^2) - [E_p E_q(\hat{Y}_{wD} - \hat{Y}_{s_I}|s_I)]^2 \\ &= E_p(B_{NR|s_I}^2) - (B_{pq}(\hat{Y}_{wD}))^2 \ , \end{aligned} \tag{4.9}$$

으로 표현된다. 따라서 식(4.9)를 식(4.8)에 대입하여 정리하면 다음과 같다.

$$MSE_{pq}(\hat{Y}_{wD}) = V_p(\hat{Y}_{s_I}) + E_p V_q(\hat{Y}_{wD}|s_I) + E_p(B_{NR|s_I}^2) + 2Cov_p(\hat{Y}_{s_I}, B_{NR|s_I}) \ . \tag{4.10}$$

여기서, $V_{SAM} = V_p(\hat{Y}_{s_I})$은 표본 분산이고, $V_{NR} = E_p V_q(\hat{Y}_{wD}|s_I)$은 무응답 오차 분산이다. 또한 $B_{NR|s_I} = E_q(\hat{Y}_{wD} - \hat{Y}_{s_I}|s_I)$은 무응답 편향이며, $Cov_p(\hat{Y}_{s_I}, B_{NR|s_I})$은 표본추출 설계 하에서 $\hat{Y}_{s_I}$과 $B_{NR|s_I}$의 공분산을 나타낸다.

무응답이 존재할 때, 관심변수와 강한 상관관계가 존재하는 보조정보를 이용할 경우 조건부 무응답 편향은 $k \in r(i \in s_I)$에 대하여 $B_{NR|s_I} \approx 0$ 이므로 식(4.10)은 식(4.5)와 유사하게 다음의 식으로 축소된다.

$$V_{pq}(\hat{Y}_w) \approx V_{pq}^0(\hat{Y}_{wD}) = V_{SAM} + V_{NR} \ . \tag{4.11}$$

식(4.11)의 $V_{pq}^0(\hat{Y}_{wD})$의 두 개 성분인 V_{SAM}과 V_{NR}을 추정한 후 $V_{pq}(\hat{Y}_{wD})$의 추정량으로써 $\hat{V}_{pq}^0(\hat{Y}_{wD})$을 두 성분의 추정치의 합으로서 이용한다. 실제로 보정추정량 $\hat{Y}_{wD}$에 대해 복잡한 수식 표현은 추정량의 분산에 대하여 근사적인 표현을 구하는 것조차도 어려울 뿐만 아니라, 추정 분산식에 대한 유도과정이 매우 복잡하다.

따라서 $\hat{Y}_{wD}$에 대한 분산 추정량을 얻기 위해서 $\hat{Y}_{CL,D\theta}$의 분산 추정량을 먼저 유도하고, $\hat{Y}_{wD}$와 $\hat{Y}_{CL,D\theta}$가 점근적으로 같아짐을 이용하여 $\hat{Y}_{CL,D\theta}$의 분산추정량으로 부터 $\hat{Y}_{wD}$의 분산 추정량을 대체하고자 한다.

$\hat{Y}_{CL,D\theta}$의 분산 추정량을 구하기 위해 먼저 $\hat{Y}_{CL,D\theta}$과 Y의 차는 다음과 같다.

$$\hat{Y}_{CL,D\theta}-Y=\sum_{s_I}d_{i(I)}\sum_{U_i}(y_k-\boldsymbol{x}_{1k}'\,\boldsymbol{C}_\theta)-\sum_{U_I}\sum_{U_i}(y_k-\boldsymbol{x}_{1k}'\,\boldsymbol{C}_\theta)-\sum_{s_I}d_{i(I)}\sum_{U_i}(y_k-\boldsymbol{x}_k'\,\boldsymbol{B}_{r\theta})$$
$$+\sum_{s_I}d_{i(I)}\sum_{r}(y_k-\boldsymbol{x}_k'\,\boldsymbol{B}_{r\theta})/\theta_k$$
$$=(\sum_{s_I}d_{i(I)}\sum_{U_i}e_{1k\theta}-\sum_{U_I}\sum_{U_i}e_{1k\theta})+(\sum_{s_I}d_{i(I)}\sum_{r}e_{k\theta}/\theta_k-\sum_{s_I}d_{i(I)}\sum_{U_i}e_{k\theta})\ .$$

여기서, $\boldsymbol{C}_\theta=\boldsymbol{B}_{1r\theta}+(\sum_{s_I}d_{i(I)}\sum_{U_i}\boldsymbol{x}_{1k}\,\boldsymbol{x}_{1k}')^{-1}(\sum_{s_I}d_{i(I)}\sum_{U_i}\boldsymbol{x}_{1k}\,\boldsymbol{x}_{2k}')\,\boldsymbol{B}_{2r\theta}$과 $\boldsymbol{B}_{r\theta}=(\boldsymbol{B}_{1r\theta},\ \boldsymbol{B}_{2r\theta})'$
$=(\sum_{s_I}d_{i(I)}\sum_{r}q_k\,\boldsymbol{x}_k\,\boldsymbol{x}_k'/\theta_k)^{-1}(\sum_{s_I}d_{i(I)}\sum_{r}q_k\,\boldsymbol{x}_k\,y_k/\theta_k)$이다.

이때 보조정보수준이 *Info-D*인 경우 식(4.5)의 가정 하에서 $\hat{Y}_{CL,D\theta}$의 분산추정량을

$$\begin{aligned}\hat{V}(\hat{Y}_{CL,D\theta})&=\sum\sum_{s_I}(d_{i(I)}d_{j(I)}-d_{ij(I)})\,\hat{t}_{E1\theta_i}\,\hat{t}_{E1\theta_j}\\&\quad-\sum_{s_I}d_{i(I)}(d_{i(I)}-1)\Big(\sum\sum_{r}\frac{(\theta_{kl}-\theta_k\theta_l)}{\theta_{kl}}\,\frac{g_{1k}e_{1k\theta}}{\theta_k}\,\frac{g_{1l}e_{1l\theta}}{\theta_l}\Big)\\&\quad+\sum_{s_I}d_{i(I)}^2\Big(\sum\sum_{r}\frac{(\theta_{kl}-\theta_k\theta_l)}{\theta_{kl}}\,\frac{g_{sk}^{D\theta}e_{k\theta}}{\theta_k}\,\frac{g_{sl}^{D\theta}e_{l\theta}}{\theta_l}\Big)\\&=\sum\sum_{s_I}(d_{i(I)}d_{j(I)}-d_{ij(I)})\,\hat{t}_{E1\theta_i}\,\hat{t}_{E1\theta_j}-\sum_{s_I}d_{i(I)}(d_{i(I)}-1)\Big\{\sum_{r}\frac{1}{\theta_k}\Big(\frac{1}{\theta_k}-1\Big)(g_{1k}e_{1k\theta})^2\Big\}\\&\quad+\sum_{s_I}d_{i(I)}\Big\{\sum_{r}\frac{1}{\theta_k}\Big(\frac{1}{\theta_k}-1\Big)(g_{sk}^{D\theta}e_{k\theta})^2\Big\}\end{aligned}\qquad(4.12)$$

으로 나타낼 수 있고, 이때 $\hat{t}_{E1\theta_i}=\sum_{r}g_{1k}e_{1k\theta}/\theta_k$이고, $e_{1k\theta}=y_k-\boldsymbol{x}_{1k}'\,\boldsymbol{C}_\theta$이다. 또한

$$e_{k\theta}=y_k-\boldsymbol{x}_k'\,\boldsymbol{B}_{r\theta}\qquad(4.13)$$

으로 표현되며, 여기서 $\boldsymbol{B}_{r\theta}=(\sum_{s_I}d_{i(I)}\sum_{r}q_k\,\boldsymbol{x}_k\,\boldsymbol{x}_k'/\theta_k)^{-1}\sum_{s_I}d_{i(I)}\sum_{r}q_k\,\boldsymbol{x}_k y_k/\theta_k$이다.

일반적으로 $\hat{Y}_{CL,D\theta}$의 분산은 다음과 같이 표현되며,

$$V(\hat{Y}_{CL,D\theta})=V_{SAM}+E_pV_q(\hat{Y}_{CL,D\theta}|s_I)\ .\qquad(4.14)$$

만일 완전응답을 가정한다면 추정량 $\hat{Y}_{CL,D\theta}$과 $\hat{Y}_{Dw}$는 모두 $\sum_{s_I}d_{i(I)}\sum_{U_i}g_{Dks}y_k$으로 동일한 형태가 된다. 여기서 $g_{Dks}=1+q_k\Big(\sum_{s_I}d_{i(I)}\sum_{U_i}g_{1k}\,\boldsymbol{x}_k-\sum_{s_I}d_{i(I)}\sum_{U_i}\boldsymbol{x}_k\Big)\Big(\sum_{s_I}d_{i(I)}\sum_{U_i}q_k\,\boldsymbol{x}_k\,\boldsymbol{x}_k'\Big)^{-1}\boldsymbol{x}_k$이다.

또한 완전응답 하에서는 표본분산 V_{SAM}과 같고, 이때 $V_{SAM}=V_p(\sum_{s_I}d_{i(I)}\sum_{U_i}g_{Dks}y_k)$이다. 그러나 무응답이 존재할 때는 식(4.4)와 식(4.8)의 두 번째 항은 서로 다른 값을 갖는다. 만일 $\hat{Y}_{CL,D\theta}$과 $\hat{Y}_{wD}$이 동일하면 식(4.4)의 무응답 분산인 V_{NR}의 추정은 간단하게 될 것이다.

따라서 이중추출기법의 무응답 총합 추정량 $\hat{Y}_{CL,D\theta}$과 보정추정량 $\hat{Y}_{wD}$이 일치하도록 응답확률이 보정된다면 유도된 $\hat{Y}_{CL,D\theta}$의 분산추정량을 $\hat{Y}_{wD}$의 분산추정량으로 대체할 수 있다.

4.3 응답확률의 보정

$\hat{Y}_{CL,D\theta}$의 분산추정량을 구하기 위해서 3절의 보정가중치를 이용하여 응답확률을 보정하고자

한다. 즉, *Info-D*인 경우 추정량 $\hat{Y}_{CL,D\theta}$과 추정량 $\hat{Y}_{wD}$이 동일하게 되도록 하는 응답확률 θ_k를 보정하고자 한다.

총합 추정량 $\hat{Y}_{CL,D\theta}=\sum_{s_I} d_{i(I)}\sum_{r} g_{1k}g_{sk}^{D\theta}y_k/\theta_k$과 $\hat{Y}_{wD}=\sum_{s_I} d_{i(I)}\sum_{r} g_{1k}g_{sk}^{D}y_k$이 동일하기 위해서는 $g_{sk}^{D}=g_{sk}^{D\theta}/\theta_k$이 성립되어야 한다. 이때의 응답확률은

$$\theta_k=\frac{1+q_k(\sum_{s_I} d_{i(I)}\sum_{U_i} g_{1k}\,\boldsymbol{x}_k-\sum_{s_I} d_{i(I)}\sum_{r} g_{1k}\,\boldsymbol{x}_k/\theta_k)'(\sum_{s_I} d_{i(I)}\sum_{r} g_{1k}\,\boldsymbol{x}_k\,\boldsymbol{x}_k'/\theta_k)^{-1}\boldsymbol{x}_k}{1+q_k(\sum_{s_I} d_{i(I)}\sum_{U_i} g_{1k}\,\boldsymbol{x}_k-\sum_{s_I} d_{i(I)}\sum_{r} g_{1k}\,\boldsymbol{x}_k)'(\sum_{s_I} d_{i(I)}\sum_{r} g_{1k}\,\boldsymbol{x}_k\,\boldsymbol{x}_k')^{-1}\boldsymbol{x}_k} \quad , \tag{4.15}$$

으로 나타낼 수 있다. 위 식을 만족하는 무수히 많은 해들 중에서 아래 조건 식

$$E_q(\sum_{s_I} d_{i(I)}\sum_{r} g_{1k}\,\boldsymbol{x}_k/\theta_k|s_I,U_i)=\sum_{s_I} d_{i(I)}\sum_{U_i} E_q(I_{k(i)})g_{1k}\,\boldsymbol{x}_k/\theta_k=\sum_{s_I} d_{i(I)}\sum_{U_i} g_{1k}\,\boldsymbol{x}_k$$

을 만족하는 θ_k는 다음과 같다.

$$\tilde{\theta}_k^{-1}=g_{sk}^{D}\ . \tag{4.16}$$

따라서 보정된 응답확률이 *Info-D*의 보조정보수준의 보정방정식을 만족한다는 조건을 다음과 같이 보일 수 있다.

$$\begin{aligned}\sum_{s_I} d_{i(I)}\sum_{r} g_{1k}\,\boldsymbol{x}_k/\,\tilde{\theta}_k &= \sum_{s_I} d_{i(I)}\sum_{r} g_{1k}\,\boldsymbol{x}_k\Big(1+q_{2k}(\sum_{s_I} d_{i(I)}\sum_{U_i} g_{1k}\,\boldsymbol{x}_k-\sum_{s_I} d_{i(I)}\sum_{r} g_{1k}\,\boldsymbol{x}_k)'\ \boldsymbol{T}_D^{-1}\,\boldsymbol{x}_k\Big)\\ &= \sum_{s_I} d_{i(I)}\sum_{r} g_{1k}\,\boldsymbol{x}_k+(\sum_{s_I} d_{i(I)}\sum_{U_i} g_{1k}\,\boldsymbol{x}_k-\sum_{s_I} d_{i(I)}\sum_{r} g_{1k}\,\boldsymbol{x}_k)\ \boldsymbol{T}_D^{-1}\,\boldsymbol{T}_D\\ &= \sum_{s_I} d_{i(I)}\sum_{U_i} g_{1k}\,\boldsymbol{x}_k\ .\end{aligned}$$

여기서, $\boldsymbol{T}_D^{-1}=\sum_{s_I} d_{i(I)}\sum_{r} g_{1k}q_{2k}\,\boldsymbol{x}_k\,\boldsymbol{x}_k'$이다.

식(2.4)에 주어진 가중치 $g_{sk}^{D\theta}$안에서 θ_k가 $\tilde{\theta}_k$로 대체될 때, $g_{sk}^{D\theta}$는 모든 k에 대하여 1이 된다.

4.4 제안된 분산 추정량

모든 s_I에 대하여 $B_{NR|s_I}\approx 0$ 라 할때, $\hat{Y}_{wD}$과 $\hat{Y}_{CL,D\theta}$은 근사적으로 같고, 만일 $\hat{Y}_{CL,D\theta}$과 $\hat{Y}_{wD}$가 동일하다면(즉, 모든 표본 s에 대하여 동등) 각각의 분산 $V_{pq}(\hat{Y}_{wD})$과 $V_{pq}(\hat{Y}_{CL,D\theta})$이 동일함을 뜻하므로 식(4.12)에서 응답확률 $\tilde{\theta}_k^{-1}=g_{sk}^{D}$을 대입하면 다음과 같은 분산추정량을 얻을 수 있다.

$$\hat{V}(\hat{Y}_{wD})=\sum\sum_{s_I}(d_{i(I)}-d_{i(I)}d_{j(I)})\,\hat{t}_{E1_i}\,\hat{t}_{E1_i}-\sum_{s_I} d_{i(I)}(d_{i(I)}-1)\Big(\sum_{r_i} g_{sk}^{D}(g_{sk}^{D}-1)(g_{1k}f_k e_{1k})^2\Big) \tag{4.17}$$

여기서, $\hat{t}_{E1_i}=\sum_{r} g_{1k}g_{sk}^{D}f_k e_{1k}$이고, g_{1k}는 식(2.15)와 같다.

또한 $e_{1k}=y_k-\boldsymbol{x}_{1k}'\{(\sum_{s_I} d_{i(I)}\sum_{r} q_k g_{sk}^{D}\,\boldsymbol{x}_{1k}\,\boldsymbol{x}_{1k}')^{-1}(\sum_{s_I} d_{i(I)}\sum_{r} q_k g_{sk}^{D}\,\boldsymbol{x}_{1k}y_k)\}$이고, $e_k=y_k-\boldsymbol{x}_k'\,\boldsymbol{B}_{rv}$이고, 이때 $\boldsymbol{B}_{rv}=(\sum_{s_I} d_{i(I)}\sum_{r} q_k g_{sk}^{D}\,\boldsymbol{x}_k\,\boldsymbol{x}_k')^{-1}\sum_{s_I} d_{i(I)}\sum_{r} q_k g_{sk}^{D}\,\boldsymbol{x}_k y_k$이다.

식(4.17)의 분산추정량은 분산 추정시 모집단 잔차 대신 표본 잔차를 이용하게 되므로 자유도 손실이 발생하며, 따라서 분산을 과소 추정하게 됨으로 분산의 과소추정을 조정하기 위한 인자

f_k를 다음과 같이 정의하였다. 모든 $k \in r$에 대하여 $f_k = \sqrt{(m-1)/(m-3)}$ 로서 즉, $e_{k\theta}$ 대신 $f_k e_{k\theta}$을 사용함으로서 자유도의 손실 분을 감안하였다.

추정량에 대한 무응답 편향을 수식으로 전개하기는 너무 복잡하므로 5절의 모의 실험을 통하여 분산추정량의 상대편향을 이용하여 효율성을 비교하였다.

5. 모의 실험

모의 실험을 위해서 Estevao와 Särndal(2002)이 제시한 방법에 따라 인위적으로 크기가 $N=12000$인 가상 모집단을 만들었다. 여기서 보조변수는 $\boldsymbol{x}=(x_1', x_2')'$로 표현하였다.

모집단에 대한 구성 요소를 살펴보면 각 단위 $k=1,\cdots,12000$에 대하여 $x_{1k} \sim G(9,10)$, $x_{2k} \sim G(9,10)$, $\varepsilon_k \sim N(0, 25^2)$이 독립적으로 생성하였다. 이때 관심변수 y_k는 $y_k = x_{1k} + x_{2k} + \varepsilon_k$와 같은 구조를 갖는다고 가정하였다. 생성된 가상의 모집단으로부터 관심 모수의 총합은 $Y = \Sigma_k y_k = 2.1581145 \times 10^6$이며, 또한 관심변수와 보조변수의 상관계수는 각각 $\rho(x_1, y) = 0.6$과 $\rho(x_2, y) = 0.59$이고, $\rho(x_1, x_2) \approx 0$이다.

모의 실험은 인위적으로 생성된 12000개의 단위로부터 각각 $N_I = 50$개의 집락을 구성하였다. 계산을 간단히 하기 위해 각 집락의 크기는 동일하게 $M_i = 240$단위가 포함되도록 하였으며, 50개의 1차 추출단위(집락)로부터 단순임의 추출로 크기가 $n_I = 20$인 5000개의 집락표본을 추출하였다. 1차 추출단위에서 선정된 표본 집락내 전체 개체에 대한 조사를 실시하고자 하였으나 조사과정에서 단위 무응답이 발생하였다고 가정하였다.

각각의 응답률은 90%, 80%, 70%, 60%로 균등분포 $U(0,1)$를 이용하여 발생시켰으며, 각 응답률에 따라 표본 집락의 조사단위에 대한 가중치 조정과정을 적용하여 집락추출하에서 무응답 보정추정량과 분산 추정량을 구해 보았다.

모의실험을 통해 모집단 총합에 대한 보정 추정량과 그에 따르는 분산 추정량의 정도를 파악하기 위해 집락 추출설계의 경우 L-S 방법과 제안된 방법에 대해 각각 다음과 같이 정의되는 상대편향을 구하여 이들간의 효율성을 비교하였다. 즉, 주어진 방법에 대해 추정량의 상대편향이 0에 가까울수록 무응답 편향을 제거한 근사적인 비편향 추정량이 됨을 나타내며, 제안된 방법 즉, 재보정 추정량이 L-S 추정량에 비해 상대편향이 작기를 기대한다.

집락 총합에 대한 추정량의 상대 편향(Relative Bias)의 백분율은

$$RB(\hat{Y}_{wD}) = \left(\frac{E(\hat{Y}_{wD}) - Y}{Y} \right) \times 100(\%) \tag{5.1}$$

이고, 여기서 $E(\hat{Y}_{wD}) = \frac{1}{K} \sum_{k=1}^{K} \hat{Y}_{wD_k}$은 $K=5000$개 표본에 대하여 취해진 보정추정량 $\hat{Y}_{wD}$의 몬테 카를로 기대치이고, $\hat{Y}_{wD_k}$는 표본 $k=1,\cdots,5000$에 대한 무응답 보정추정량 $\hat{Y}_{wD}$의 값이다.

또한 분산 추정량에 대해 상대 편향의 백분율은

$$RB(\hat{V}) = \frac{[E(\hat{V}(\hat{Y}_{wD})) - V_{sim}]}{V_{sim}} \times 100(\%) \tag{5.2}$$

이다. 여기서 $E(\hat{V}(\hat{Y}_{wD}))=\frac{1}{K}\sum_{k=1}^{K}\hat{V}_k(\hat{Y}_{wD})$이며 $\hat{V}_k(\hat{Y}_{wD})$는 각각의 표본 $k=1,\cdots,5000$에 대한 분산추정량 $\hat{V}(\hat{Y}_{wD})$의 값을 의미한다. 또한 $V_{sim}=\frac{1}{K-1}\sum_{k=1}^{K}(\hat{Y}_{wD_k}-E(\hat{Y}_{wD}))^2$ 으로 모의실험 분산을 의미한다.

95% 수준에서 근사적으로 신뢰구간의 포함률을 평가하는 측도는 다음과 같이 정의하였다.

$$\hat{Y}_{wD}\pm 1.96\sqrt{\hat{V}(\hat{Y}_{wD})} \tag{5.3}$$

참고로 다음의 표에서 나타낸 L-S 추정량이 의미하는 바는 Lundström과 Särndal(1999)이 제안한 방법을 나타내며, 이때 보조정보의 수준은 모집단 보조정보만을 이용한 경우이다.

[표 5.1]은 응답률과 추정방법에 따른 보정 추정량과 상대편향, 그리고 95% 포함률을 모의실험을 통해 얻은 결과이다.

[표 5.1] 보정 추정량과 분산 추정량의 상대편향(%), 그리고 95% 포함률

통계량 / 방 법 / 응답률(%)	보정추정량의 상대편향		분산추정량의 상대편향		95% 포함률	
	L-S추정량	재보정추정량	L-S추정량	재보정추정량	L-S추정량	재보정추정량
90	0.003	0.004	-2.019	-0.234	93.1	93.3
80	0.004	0.005	-2.412	0.619	93.8	92.9
70	0.002	0.003	-2.014	0.182	94.0	92.7
60	-0.000	0.001	-2.581	-0.439	94.7	92.6

모의 실험 결과 응답률 변화에 대해, 두 방법 모두 모집단 총합에 대해 보정 추정량의 상대편향이 거의 0인 것으로 나타나고 있으며, 이와 같은 사실은 L-S 추정량이나, 제안된 추정량 모두 무응답 상황 하에서 무응답에 기인한 편향이 거의 없도록 추정이 이루어진 것으로 사료된다.

또한 보정추정량에 대한 분산 추정량의 상대편향으로부터 L-S 방법으로부터 구한 분산추정량의 상대편향은 약 -2%정도로서 집락 추출설계 하에서 분산 추정량이 음의 편향이 나타나 모집단 분산을 과소 추정하고 있는 반면, 본 논문에서 제안한 방법으로 추정한 분산추정량의 상대편향은 거의 0%로 나타난 것으로 볼 때, 재보정 추정량의 분산추정량이 모집단 분산인 V_{sim}에 점근적으로 근사함을 알 수 있으며, 따라서 L-S 추정량에 비해 무응답에 기인한 편향을 감소시키는 것으로 나타나고 있다. 마지막으로 추정량들의 95%신뢰수준에서의 포함률을 비교해보면 두 방법 모두 95%신뢰 수준에 적합함을 보여주고 있다. 이와 같은 모의 실험의 결과를 요약하면 다음과 같다.

첫째, 집락 추출설계하에서 단위 무응답이 발생한 경우 단위 무응답을 보정하기 위한 L-S 방법과 본 논문에서 제안한 재보정 방법은 모두 모집단 총합에 대해 무응답편향을 상당히 제거할 수 있음을 할 수 있다.

둘째, 분산 추정량과 관련하여 L-S 방법의 경우 재보정 방법에 비해 상대적으로 모집단 분산을 과소 추정하는 것으로 나타나며, 따라서 제안된 방법이 보다 효과적으로 추정함을 알 수 있다.

6. 결 론

일반적으로 단위 무응답이 발생했을 때, 무응답 편향을 감소시키기 위한 방법으로 가중치 조정법을 사용하게 된다. 본 논문에서는 이와 같이 단위 무응답이 존재하는 경우 가중치 조정방법 중에서 이중추출방법을 통한 재보정 추정법을 적용하여 추출가중치와 무응답 단위에 대하여 최종 가중치인 응답확률을 보정하고, 그에 대한 보정 추정량과 그에 따르는 분산추정량을 도출하였다.

특히 본 논문에서는 보정에 이용되는 보조정보의 수준이 단계별로 다른 경우를 고려하였다. 즉, 1단계에서는 $\boldsymbol{x}_{1k}'$을 모집단 보조정보 수준으로 이용하였고, 2단계에서는 표본보조정보 수준으로 $\boldsymbol{x}_k = (\boldsymbol{x}_{1k}', \boldsymbol{x}_{2k}')'$을 이용하였다. 이와 같은 보조정보를 이용하여 1단계에서는 추출가중치를 조정하고, 조정된 추출가중치를 이용하여 2단계에서 최종 가중치와 응답확률을 보정하였다.

모의 실험결과 보정추정량의 상대편향은 L-S 방법과 제안한 방법 모두 거의 0으로 나타났고, 95% 신뢰수준에서 포함률은 두 방법 모두 거의 비슷하게 나타났다. 그러나 분산추정량의 정도를 살펴보면 모든 보조변수들을 모집단 정보로 이용한 L-S의 추정 결과는 상대편향이 약 -2% 정도로 과소 추정하고 있지만, 보조변수의 형태가 제한된 본 논문의 분산추정량의 상대편향은 거의 0%로서 L-S 방법보다 더 정확한 추정방법임을 알 수 있었다.

본 논문에서는 보조변수들간에 상관이 낮은 경우에 대해서만 고려하였지만, 향후 보조변수들간에 상관이 높은 경우에는 어떤 효과가 나타나는지는 연구과제로 남긴다.

참고 문헌

[1] 손창균. 홍기학, 이기성(2000). 무응답 상황하에서 보조정보의 수준에 따른 분산추정량에 관한 연구, *한국통계학회 춘계발표 논문집*. pp.239-244.

[2] 염준근, 정영미(2002). A Study for the Unit Nonresponse Calibration using Two-Phase Sampling method, *한국통계학회 논문집* 9. pp.479-489.

[3] Deville, J. C., and Särndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, pp.376-382.

[4] Ekholm, A. and Laaksonen, S. (1991). Weighting via Response Modeling in the Finnish Household Budget Survey. *Journal of Official Statistics*, 3, pp.325-337.

[5] Estevao, V. M., and Särndal, C. E. (2002). The Ten Cases of Auxiliary Information for Calibration in Two-Phase Sampling. *Journal of Official Statistics*, 18, pp.233-255.

[6] Hidiroglou, M. A. and Särndal, C. E. (1995). Use of Auxiliary Information for Two-phase sampling. pp.873-878.

[7] Kalton, G. and Kasprzyk, D. (1986). The Treatment of Missing Survey Data. *Survey Methodology*, 12, pp.1-16.

[8] Little, R. J. A. (1986). Survey Nonresponse Adjustment for Estimates of Means. *International Statistical Review*, 54, pp.139-157.

[9] Lundström, S., and Särndal, C. E. (1999). Calibration as a Standard Method for Treatment of Nonresponse. *Journal of Official Statistics*, 15, pp.305-327.

[10] Särndal, C. E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling.* New York: Springer-Verlag.

[2003년 5월 접수, 2003년 8월 채택]