The Korean Communications
in Statistics Vol. 9, No. 2, 2002
pp. 479-489

A Study for the Unit Nonresponse Calibration using Two-Phase Sampling Method

Joon Keun Yum[1], Young Mee Jung[2]

Abstract

The case which applies two-phase sampling to stratification and nonresponse problem, it is a powerful and effective technique. In this paper we study the calibration estimator and its variance estimator for the population total using two-phase sampling method according to the of auxiliary information for population and sample having strong correlation with an interested variable in unit nonresponse situation. The auxiliary information that available both at first-phase and second-phase sampling can be used to improve weights by the calibration procedure. A weight which corresponds to the product of sampling weights and response probability is calculated at each phase of sampling.

Keywords : Unit nonresponse, Auxiliary information, Two-phase sampling, Calibration estimator

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). 무응답을 제외한 응답자들만을 이용한 분석은 무응답 편향을 발생시키기 때문에 이러한 무응답 편향을 줄이기 위한 노력을 필요로 한다. 항목무응답과 단위무응답을 처리하기 위한 방법으로 각각 대체법과 가중치 조정방법을 들 수 있다. 본 논문에서는 조사 단위에 발생하는 단위무응답에 대해 가중치 조정방법중의 하나인 보정 추정방법을 적용하여 무응답 편향을 줄이고자 한다.

단위무응답 처리와 관련하여 Särndal과 Swensson(1987)은 층화이중추출 방법을 적용해서 추정량을 구했고, Lundström과 Särndal(1999)은 관심변수와 강한 상관이 존재하는 보조변수의 기지의 모집단 총합과 표본 총합을 이용하여 관심변수의 총합 추정량과 분산 추정량을 도출하였다. 또한 손창균, 홍기학 그리고 이기성(2000)은 기지의 모집단 총합에 대한 정보뿐만 아니라 보조변수에 대한 기지의 분산 정보를 가지고 관심변수에 대한 분산 추정치를 구하였고, 염준근, 손창균, 그리고 정영미(2002)는 단위무응답이 존재할 때, 2단계에 걸쳐서 응답확률과 추출가중치를 조정하였고, 관심 모수에 대한 추정량을 구하였다.

1) Professor, Department of Statistics, Dongguk University, Seoul, 100-715, Korea.
E-mail : joonkeun@dgu.edu
2) Lecturer, Department of Statistics, Dongguk University, Seoul, 100-715, Korea
E-mail : jym007@orgio.net

본 논문에서는 조사 단위에 무응답이 발생하는 경우에 대해 이중추출기법을 적용해서 미지의 응답확률을 보정 한 후 총합에 대한 추정량과 분산추정량을 구하고자 한다. 먼저 1단계에서는 관심변수와 강한 양의 상관이 있는 보조변수의 정보를 이용해서 추출 가중치를 조정하고, 2단계에서는 보정방정식을 이용해서 응답확률을 조정하였다.

논문의 구성은 우선 2장에서는 무응답이 발생했을 때의 일반적인 총합 추정량을 표현하고, 3장에서는 각 단계별로 보조정보를 이용해서 추출가중치와 응답확률을 조정 한 후 최종 보정 추정량을 도출하였다. 4장과 5장에서는 분산추정량을 도출하고, 몬테칼로 모의 실험을 통해 추정량과 분산추정량의 상대편향 뿐만 아니라 95%수준에서의 포함율을 구하였다. 마지막으로 6장에서는 5장의 모의실험 결과를 통하여 결론을 다루었다.

2. 모집단 총합에 대한 추정량

유한모집단 $U=\{1,\cdots,k,\cdots,N\}$에 대하여 모집단 총합 $Y=\sum_U y_k$을 추정하는 것이 연구의 목적이다. 이때 관심변수는 y이고, y_k는 k번째 단위에 대한 y의 값이다. 관심변수와 강한 상관을 가지는 보조변수 벡터 $\boldsymbol{x}$를 가정하고, k번째 단위에 대한 보조변수 벡터 값을 $\boldsymbol{x}_k=(x_{k1},\cdots,x_{kj},\cdots,x_{kJ})'$라 하자. 또한 모집단으로부터 $p(s)$의 확률로 추출한 크기 n인 표본을 s라 하자. 그러면 모집단 단위가 표본에 포함될 확률들은 각각 $\pi_k=\sum_{s\ni k}p(s)$와 $\pi_{kl}=\sum_{s\ni k,l}p(s)$이다. π_k의 역수인 $\pi_k^{-1}=d_k$는 단위 k에 대한 설계가중치이며, 또한 $\pi_{kl}^{-1}=d_{kl}$이다. 이때, 무응답이 발생하여 크기가 $m(m\le n)$인 응답집합 $r(r\subseteq s)$이 얻어졌다. 무응답이 발생했을 때 무응답을 조정하기 위해 관심변수와 강한 양의 상관을 가진 보조 정보를 이용한다면 추출오차와 무응답 편향을 충분히 감소시킬 수 있다. 이용가능한 보조정보의 수준은 표본정보와 모집단 정보로 구분된다. 이용 가능한 보조정보가 모집단에 대한 정보라는 것은 $\boldsymbol{x}_k$가 모든 단위 $k\in U$에 대하여 기지이거나 $\sum_U \boldsymbol{x}_k$가 기지이고, 모든 $k\in s$에 대하여 $\boldsymbol{x}_k$가 관측된다는 것을 의미한다. 또한 표본정보는 모든 모든 $k\in s$에 대해서만 $\boldsymbol{x}_k$가 관측된다는 것을 의미한다.

본 논문에서는 무응답을 처리하기 위해서 2단계에서 응답메카니즘이 고려되는 이중추출기법을 적용하고자 한다. $pr(k\in r|s)=\theta_k$, $pr(k\&l\in r|s)=\theta_{kl}$(표본 s에 독립)으로 나타내는 기지의 응답확률과 대응하는 응답분포 $q(r|s)$이 기지라고 가정하면, Särndal, Swensson 그리고 Wretman(1992)이 제안한 관심변수 y의 모집단 총합 Y의 추정량은 모집단 정보를 이용할 때 다음과 같이 정의된다.

$$\hat{Y}_{ssw,U\theta}=\sum_r d_k g_{Uk\theta} y_k/\theta_k \tag{2.1}$$

여기서, g-가중치는

$$g_{Uk\theta}=1+c_k\left(\sum_U \boldsymbol{x}_k-\sum_r d_k\,\boldsymbol{x}_k/\theta_k\right)'\left(\sum_r d_k c_k\,\boldsymbol{x}_k\,\boldsymbol{x}_k'/\theta_k\right)^{-1}\boldsymbol{x}_k \tag{2.2}$$

이다.

표본정보를 이용할 때 추정량은 다음과 같다.

$$\hat{Y}_{ssw,s\theta} = \sum_{r} d_k g_{sk\theta} y_k / \theta_k \tag{2.3}$$

여기서, g-가중치는

$$g_{sk\theta} = 1 + c_k(\sum_{s} d_k \boldsymbol{x}_k - \sum_{r} d_k \boldsymbol{x}_k/\theta_k)'(\sum_{r} d_k c_k \boldsymbol{x}_k \boldsymbol{x}_k'/\theta_k)^{-1} \boldsymbol{x}_k \tag{2.4}$$

이다.

위의 식에서 응답확률 θ_k는 실제로 미지이며, 따라서 임의의 값인 $\hat{\theta}_k$로 대체해야 한다. 우선 관련된 응답모형을 세우고, 그 다음에 미지의 응답확률을 추정하는 전형적인 절차로부터 구한 모집단 총합 Y의 추정량은 다음과 같다.

$$\hat{Y} = \sum_{r} d_k \nu_{1k} \nu_{2k} y_k \tag{2.5}$$

이때, $\nu_{1k} = \hat{\theta}_k^{-1}$이며, ν_{2k}는 식(2.2), (2.4)의 g-가중치와 같고, g-가중치의 θ_k 를 $\hat{\theta}_k$로 대체한다. 식(2.5)를 구하기 위해서 Lundström 등(1999)은 가중치를 동시에 조정하는 방법을 이용했으나, 다음 절에서는 가중치와 응답확률을 각 단계별로 조정한 새로운 보정가중치를 이중추출기법을 이용해서 구해보았다. 여기에서 중요한 것은 두 방법 모두 응답모형을 따로 가정할 필요가 없다는 것이다.

3. 무응답에 대한 보정추정량

무응답이 있는 경우와 이중추출간의 유사한 점은 처음에 추출된 표본은 s이나 실제로 측정된 연구변수를 나타내는 최종표본은 s의 부분집합인 응답집합 r이 된다는 점이다. 이 절에서는 이중추출기법을 적용해서 설계 가중치를 조정하고, 응답확률을 보정해서 무응답 보정 추정량을 구하고자 한다. 추출의 각 단계에서 유용하게 이용되는 보조정보는 보정가중치를 구하기 위해서 이용되었고, 초기가중치와 가능한 한 밀접한 새로운 보정가중치를 구하기 위해서는 거리함수를 이용하게 된다. Devile과 Särndal(1992)은 여러 가지 거리함수를 제안했으며, Lundström 등(1999)은 임의의 단위 집합 $k \in r$에 대하여, 일반화최소제곱 거리함수(GLS)를 이용해서 보정추정량을 구했다. 그러나 GLS 거리함수의 단점은 음의 가중치가 발생할 수 있다는 것이다. 음의 가중치는 결과의 의미있는 해석을 어렵게 하므로 단점을 보완해주는 제한적인 방법으로 보정방정식을 만족하면서 가중치의 범위를 제한 해주는 방법을 사용하기는 하나 본 논문에서는 일반적인 방법으로 Devile 등(1992)에 의해서 제안된 거리함수 중에서 최소 엔트로피 거리측도를 이용해서 가중치를 보정함으로써 음의 가중치가 발생되지 않도록 하였다.

이론전개를 위해 기호를 정리해 보면 먼저 크기가 n인 1단계 확률 표본 $s(s \subset U)$는 추출설계 $p(\cdot)$에 의해서 추출되고, k번째 단위가 표본에 포함될 확률은 $\pi_{1k} = \sum_{s \ni k} p(s)$이고, k, l이 동시에 표본에 포함될 확률은 $\pi_{1kl} = \sum_{s \ni k, l} p(s)$이다. 따라서 단위 k의 1단계 추출가중치는 $d_{1k} = 1/\pi_{1k}$이 된다. 또한 크기가 m인 2단계 응답표본 $r(r \subset s)$은 이상(two-phase) 추출설계 $q(\cdot|s)$에 의해서 추출된다. 즉, s표본에서 얻은 보조정보를 이용하여 설계 $q(\cdot|s)$에 의해서 부

차표본(subsample) r를 추출하고, r에서만 관심변수 y를 조사할 수 있다. s가 주어지면 k번째 단위가 r에 포함될 포함확률은 $\pi_{k|s} = \sum_{r \ni k} q(r|s)$이 된다.

3.1. 이중추출기법을 이용해서 조정한 보정추정량

최소엔트로피 측도의 일반식은 다음과 같다.

$$\sum_r (-d_k \log(w_k/d_k) + w_k - d_k)/c_k$$

위의 일반식을 이용해서 본 논문에서는 2단계로 구분하여 각 단계마다 주어진 제약 조건 즉, 보정방정식을 만족한다는 조건 하에서 최소엔트로피 거리함수를 최소화하는 최종 보정 가중치를 다음과 같은 과정에 의해서 얻었다.

[step1] 1단계 보정 (s부터 U까지) : 1단계 추출 가중치 $\{d_{1k} : k \in s\}$는 초기 가중치이고, $\{c_{1k} : k \in s\}$는 기지인 양의 가중치이다. 1단계 보정방정식 $\sum_s w_{1k}\, \boldsymbol{x}_k = \sum_U \boldsymbol{x}_k$을 만족한다는 조건 하에서 다음과 같은 거리함수를 최소화하는 1단계 보정 추출 가중치 w_{1k}을 결정할 수 있다.

$$\sum_s (-d_{1k} \log(w_{1k}/d_{1k}) + w_{1k} - d_{1k})/c_{1k} \tag{3.1.1}$$

여기에서, 모든 $k \in s$에 대하여 1단계 보정된 가중치 $w_{1k} = d_{1k} g_{1k}$를 얻고, g-가중치는

$$g_{1k} = (1 - c_{1k}\, \boldsymbol{x}_k{}'\, \boldsymbol{\lambda}_1)^{-1} \tag{3.1.2}$$

이며, 1단계 보정방정식 $\sum_U \boldsymbol{x}_k = \sum_s d_{1k} g_{1k}\, \boldsymbol{x}_k$으로부터 $\boldsymbol{\lambda}_1$를 구하기 위해서 $\phi_s(\boldsymbol{\lambda}_1)$를

$$\phi_s(\boldsymbol{\lambda}_1) = \sum_s d_{1k}\{(1 - c_{1k}\, \boldsymbol{x}_k{}'\, \boldsymbol{\lambda}_1)^{-1} - 1\}\, \boldsymbol{x}_k = \sum_U \boldsymbol{x}_k - \sum_s d_{1k}\, \boldsymbol{x}_k$$

으로 정의하면 반복수 $\nu = 1, 2, \cdots$에 대하여 1단계 반복값 $\boldsymbol{\lambda}_{1(\nu)}$는 다음과 같이 얻어진다.

$$\boldsymbol{\lambda}_{1(\nu+1)} = \boldsymbol{\lambda}_{1(\nu)} + \{\phi_s{}'(\boldsymbol{\lambda}_{1(\nu)})\}^{-1} \left\{ \sum_U \boldsymbol{x}_k - \sum_s d_{1k}\, \boldsymbol{x}_k - \phi_s(\boldsymbol{\lambda}_{1(\nu)}) \right\}$$

여기서, 초기치는 $\boldsymbol{\lambda}_{1(0)} = 0$로 시작하고, $\phi_s(0) = 0$이다. $\phi_s{}'(\boldsymbol{\lambda}_1) = \partial \phi_s(\boldsymbol{\lambda}_1)/\partial \boldsymbol{\lambda}_1$라 하면 $\phi_s{}'(0) = \sum_s d_{1k} c_{1k}\, \boldsymbol{x}_k\, \boldsymbol{x}_k{}'$이고, 첫 번째 결과는 $\boldsymbol{\lambda}_{1(1)} = (\sum_r d_{1k} c_{1k}\, \boldsymbol{x}_k\, \boldsymbol{x}_k{}')^{-1} \left(\sum_U \boldsymbol{x}_k - \sum_s d_{1k}\, \boldsymbol{x}_k \right)$이고, 수렴할 때 까지 반복을 계속한다.

[step2] 2단계 보정 (r부터 s까지) : 초기 가중치는 $\{ w_{1k} d_{2k} : k \in r\}$으로 주어진다. 여기서, $w_{1k} d_{2k} = d_k^* g_{1k}$으로 표현할 수 있고, $d_k^* = d_{1k} d_{2k}$이다. 최종적인 g-가중치를 구하기 위해 1단계 가중치와 2단계 가중치를 곱으로 표현한 승법(Multiplicative)방법을 적용하였다(Hidiroglou와 Deville(1995)). 가중치를 구하기 위해서 먼저 2단계 보정 방정식 $\sum_r w_k{}^*\, \boldsymbol{x}_k = \sum_s w_{1k}\, \boldsymbol{x}_k$을 만족한다는 조건하에서 다음 식을 최소화하는 보정 가중치 $w_k{}^*$를 결정할 수 있다.

$$\sum_r (-w_{1k} d_{2k} \log(w_k^*/w_{1k} d_{2k}) + w_k^* - w_{1k} d_{2k})/c_{2k} \tag{3.1.3}$$

여기서, $\{c_{2k} : k \in r\}$는 미리 조건으로서 주어진 기지의 양의 가중치이다. 이러한 보정을 실시한 결과 가중치는 다음과 같이 최종 보정된 가중치로 정의된다.

$$w_k^* = d_k^*(g_{1k} g_k^M) \tag{3.1.4}$$

여기서, $k \in r$에 대하여

$$g_k^M = (1 - c_{2k}\, \boldsymbol{x}_k{}'\, \boldsymbol{\lambda}_2)^{-1} \tag{3.1.5}$$

이며, 2단계의 보정 방정식으로부터 $\phi_s(\boldsymbol{\lambda}) = \sum_r w_{1k} d_{2k}\{(1 - c_{2k}\, \boldsymbol{x}_k{}'\, \boldsymbol{\lambda}_2)^{-2} - 1\}\, \boldsymbol{x}_k$ 으로 놓으면, 뉴튼-랩슨 방법을 이용해서 반복수 $\nu = 1, 2, \cdots$에 대하여 2단계 반복값 $\boldsymbol{\lambda}_{2(\nu)}$는 다음과 같다.

$$\boldsymbol{\lambda}_{2(\nu+1)} = \boldsymbol{\lambda}_{2(\nu)} + \{\phi_s{}'(\boldsymbol{\lambda}_{2(\nu)})\}^{-1}\Big\{\sum_s w_{1k}\, \boldsymbol{x}_k - \sum_r w_{1k} d_{2k}\, \boldsymbol{x}_k - \phi_s(\boldsymbol{\lambda}_{2(\nu)})\Big\}$$

여기서, 초기치는 $\boldsymbol{\lambda}_{2(0)} = 0$로 시작하고, $\phi_s(0) = 0$이다. $\phi_s{}'(\boldsymbol{\lambda}_2) = \partial \phi_s(\boldsymbol{\lambda}_2)/\partial \boldsymbol{\lambda}_2$라 하면 $\phi_s{}'(0) = \sum_r w_{1k} d_{2k} c_{2k}\, \boldsymbol{x}_k\, \boldsymbol{x}_k{}'$이고, $\boldsymbol{\lambda}_{2(1)} = (\sum_r w_{1k} d_{2k} c_{2k}\, \boldsymbol{x}_k\, \boldsymbol{x}_k{}')^{-1}(\sum_s w_{1k}\, \boldsymbol{x}_k - \sum_r w_{1k} d_{2k}\, \boldsymbol{x}_k)$이다. 수렴할 때까지 반복한다. 식(2.5)에서 응답확률과 g-가중치 즉, ν_{2k}와 ν_{1k}를 각 단계별로 조정한 최종 g-가중치 $g_k^* = g_{1k} g_k^M$를 이용해서 총합에 대한 무응답 보정 추정량을 다음과 같이 구할 수 있다.

$$\hat{Y}_{Dw} = \sum_r w_k^* y_k \tag{3.1.6}$$

Lundström 등(1999)은 보정방정식 $\sum_U \boldsymbol{x}_k = \sum_r w_k\, \boldsymbol{x}_k$를 만족하면서 거리함수를 최소화하는 보정가중치 w_k로 응답확률과 설계가중치의 곱의 형태인 $\nu_{1k}\nu_{2k}$를 동시에 조정하는 가중치를 이용했다. 동시적 방법을 이용해서 가중치를 조정하는 방법과 이중추출기법을 이용해서 가중치와 응답확률을 조정하는 방법에 대한 추정량의 효율성 비교는 모의실험을 통해서 보였다.

3.2. 추정량의 무응답 편향

Särndal 등(1992)의 방법을 수정해서 2단계에서 표본정보에 대한 추정량을 구하기 위해서 먼저 최소 엔트로피 거리측도를 이용해서 구한 비선형 g-가중치인 g_k^M을 테일러 전개를 이용해서 근사적으로 선형화시킨 후 일반화 회귀 추정량에 대한 이중추출 기법을 적용하면 다음과 같다.

$$\hat{Y}_{SSW,\, sw\theta} = \sum_r d_k g_{skw\theta} y_k / \theta_k \tag{3.2.1}$$

여기서, $g_{skw\theta} = 1 + c_{2k}(\sum_s w_{1k}\, \boldsymbol{x}_k - \sum_r w_{1k} d_{2k}\, \boldsymbol{x}_k/\theta_k)'(\sum_r w_{1k} d_{2k} c_{2k}\, \boldsymbol{x}_k\, \boldsymbol{x}_k{}'/\theta_k)^{-1}\, \boldsymbol{x}_k$이다. 또한 보정추정량은 $\hat{Y}_{ws} = \sum_r d_k g_k^M y_k$이고, 모든 $k \in r$에 대하여 $\theta_k = g_{skw\theta}/g_k^M$을 만족할 때 이 보정추정량은 식(3.2.1)과 동일하게 된다. 따라서 조건식 $\sum_r \frac{w_{1k} d_{2k}\, \boldsymbol{x}_k}{\theta_k} = \sum_s w_{1k}\, \boldsymbol{x}_k$을 만족하는 응답확률을 구하면 $\hat{\theta}_k{}^{-1} = g_k^M$이고, $c_{2k} = 1$로 놓으면 추정된 응답확률은 조건식을 만족하게 된다.

이 결과를 이용해서 다음과 같은 결론을 유도해 낼 수 있다.

[정리1] 2단계로 구분한 방법을 이용한 보정추정량(식(3.1.6))의 무응답 편향은 근사적으로 0이 된다.

증명) 일반적인 경우에 추정량의 무응답 편향을 구하기 위해서 식(3.1.6)을 다음과 같이 재 표현했다.

$$\hat{Y}_{Dw} = \sum_r d_k^* y_k + (\sum_U \boldsymbol{x}_k - \sum_r d_k^* \boldsymbol{x}_k)\ \boldsymbol{B}_r \tag{3.2.2}$$

여기서, $\boldsymbol{B}_r = \sum_r d_k^* c_k \boldsymbol{x}_k y_k / \sum_r d_k^* c_k \boldsymbol{x}_k \boldsymbol{x}_k'$이고, 모든 $k \in U$에 대하여 $c_k = 1/\boldsymbol{\mu}' \boldsymbol{x}_k$을 만족한다는 가정을 제시했다. 또한 $\boldsymbol{\mu}$는 $\boldsymbol{x}_k$와 같은 차원의 열벡터이고, k에 의존하지 않는다. Lundström 등(1999)이 제시한 결과를 적용하면 큰 응답집합에 대하여 무응답 편향에 대한 식은

$$\begin{aligned} B_{pq}(\hat{Y}_{Dw}) &= \hat{Y}_{Dw} - \sum_U y_k \approx E(\hat{Y}_{Dw} - Y) \\ &\approx \sum_U \boldsymbol{x}_k' \boldsymbol{B}_U^{\theta} - \sum_U \theta_k \boldsymbol{x}_k' \boldsymbol{B}_U^{\theta} - \sum_U (1-\theta_k) y_k \end{aligned} \tag{3.2.3}$$

이고, 여기서 $\boldsymbol{B}_U^{\theta} = (\sum_U \theta_k c_k \boldsymbol{x}_k \boldsymbol{x}_k')^{-1} \sum_U \theta_k c_k \boldsymbol{x}_k y_k$이 된다. 본 논문에서 제안한 2단계에서 조정한 가중치를 이용해서 응답확률을 $\hat{\theta}^{-1} = g_k^M = (1 - c_k \boldsymbol{\lambda}' \boldsymbol{x}_k)^{-1}$로 놓고, 다음 절차에 따라 테일러 전개를 이용해서 g_k^M을 1차 근사 시킨다.

[1단계] $F(z) = (1 - c_k \boldsymbol{\lambda}' \boldsymbol{x}_k)^{-1}$으로 놓고, $z = \boldsymbol{\lambda}' \boldsymbol{x}_k$에 관해서 미분한다.

[2단계] $F'(z) = \dfrac{c_k}{(1 - c_k \boldsymbol{\lambda}' \boldsymbol{x}_k)^2}$, $F'(0) = c_k$, $F(0) = 1$이다.

[3단계] $F(z) = F(0) + F'(0)(z - 0) +$ 나머지항 $\approx 1 + c_k(\boldsymbol{\lambda}' \boldsymbol{x}_k)$

「단계3」에서 구한 응답확률 $\hat{\theta}^{-1} \approx 1 + c_k \boldsymbol{\lambda}' \boldsymbol{x}_k$을 식(3.2.3)에 대입하면 $B_{pq}(\hat{Y}_{Dw}) \approx 0$을 만족한다. 따라서 무응답 편향이 근사적으로 0이 된다.

[Remark] 대부분의 경우에는 일반화 회귀 추정량을 유도하기 위해서 가중치 $c_k = 1$로 두고 계산하지만, 균등한 가중치를 제시하지 않는 특별한 경우도 있다. 즉 식(3.2.2)에서 $\boldsymbol{x}_k = x_k$를 취하고, $c_k = 1/x_k$라고 놓으면, $\hat{Y}_{Rw} = \sum_U x_k \sum_r d_k y_k / \sum_r d_k x_k$과 같이 비추정량의 형태가 된다. 따라서 $1/\hat{\theta}_k = g_k^M = \sum_s d_k x_k / \sum_r d_k x_k$으로 응답확률을 조정할 수 있으며, 단변량 보조정보를 이용할 경우에는 비추정량의 형태로 보정추정량을 구하는 것이 회귀추정량을 이용하는 것보다 더 효율적이며, 총합에 대한 비추정량의 편향이 근사적으로 0이 된다.

증명하기 위해서 먼저 비선형인 비추정량을 선형화로 근사시키는 1차 테일러 전개를 실시하고, Y와의 차를 다음과 같이 구한다.

$$\hat{Y}_{Rw} - Y \approx \sum_r d_k y_k + (\sum_U x_k - \sum_r d_k x_k)' B_r - \sum_U y_k$$

여기서 $B_r = (\sum_r d_k x_k)^{-1} \sum_r d_k y_k$이 된다. 양변에 기대치를 취하면

$$E_{pq}(\hat{Y}_{Rw} - Y) \approx \sum_U \theta_k y_k + (\sum_U x_k - \sum_U \theta_k x_k) B_U^\theta - \sum_U y_k$$
$$= -\sum_U (1-\theta_k) y_k + \sum_U (1-\theta_k) x_k' B_U^{\theta_k} = -\sum_U (1-\theta_k) E_k^\theta$$
$$= \sum_U x_k' B_U^\theta - \sum_U \theta_k x_k' B_U^\theta - \sum_U (1-\theta_k) y_k$$

이고, 여기서 $E_k^\theta = y_k - x_k' B_U^\theta$이다. [정리1]과 같이 모든 k에 대하여 $c_k \mu' x_k = 1$이라는 가정을 만족하면, $\sum_U E_k^\theta = 0$식이 성립하므로 $\sum_U \theta_k x_k' B_U^\theta = \sum_U \theta_k y_k$에 의해서 다음과 같은 식이 유도된다.

$$B_{pq}(\hat{Y}_{Rw}) \approx \sum_U x_k' B_U^\theta - \sum_U y_k = -\sum_U E_k^\theta$$
$$= \sum_U x_k' B_U + \sum_U x_k' B_{UE}^\theta - \sum_U y_k$$
$$= \sum_U x_k' B_{UE}^\theta - \sum_U E_k$$

여기에서 응답확률 θ_k에 대한 방정식을 $1 = \theta_k(1+\lambda)$으로 두면,

$$\sum_U x_k' B_U^\theta = \sum_U \theta_k x_k' B_U^\theta + \lambda' \sum_U \theta_k x_k' B_U^\theta$$
$$= \sum_U \theta_k x_k' B_U^\theta + \sum_U \theta_k \lambda' y_k = \sum_U \theta_k x_k' B_U^\theta + \sum_U (1-\theta_k) y_k$$

이 만족된다. 여기에 $B_U^\theta = \sum_U \theta_k y_k / \sum_U \theta_k x_k$를 대입하면 $B_{pq}(\hat{Y}_{Rw}) \approx 0$을 만족한다.

4. 분산추정량

Lundström 등(1999)은 추출설계 $p(s)$와 응답분포 $q(r|s)$하에서 일반적인 w-가중 추정량 $\hat{Y}_w$의 MSE(평균제곱오차)를 구하였으며, 무응답 편향에 대한 식을 다음과 같이 도출하였다.

$$MSE_{pq}(\hat{Y}_w) = V_{SAM} + V_{NR} + 2Cov_p(\hat{Y}_s, B_{NR|s}) + E_p(B_{NR|s}^2) \qquad (4.1)$$

여기서, $V_{SAM} = V_p(\hat{Y}_s)$은 표본분산이고, $V_{NR} = E_p V_q(\hat{Y}_w | s)$은 무응답으로 인해 발생하는 오차 분산, $B_{NR|s} = E_q(\hat{Y}_w - \hat{Y}_s | s)$은 s에 관한 조건부 무응답 편향, $Cov_p(\hat{Y}_s, B_{NR|s})$은 추출설계하에서 $\hat{Y}_s$과 $B_{NR|s}$의 공분산이다. 그렇지만 강한 보조정보를 이용할 수 있다면, 모든 s에 대하여 무응답 편향은 근사적으로 0이 된다는 다음과 같은 조건이 성립된다.

$$B_{NR|s} = 0 \qquad (4.2)$$

따라서 MSE는 추정량의 분산으로 다음과 같이 다시 표현된다.

$$MSE_{pq}(\hat{Y}_w) \approx V_{pq}(\hat{Y}_w) = V_{SAM} + V_{NR}$$

실제로 보정 추정량에 대한 복잡한 표현은 추정량의 분산에 대한 표현식을 근사적으로 얻는 것조차도 어렵게 만든다. 따라서 이중추출기법을 적용해서 Särndal 등(1992)이 구한 점 추정량의 분산추정량의 형태를 단위무응답이 존재할 때 제시한 보정추정량의 분산추정량에 적용해서 사용했

다(Lundström 등(1999)). 무응답 상황을 고려하는 경우에 간단히 하기 위해서 표본개체가 각각 독립적으로 응답한다고 가정한다. 따라서 식(4.3)을 만족하게 된다.

$$pr(k\&l \in r|s) = \theta_{kl} = \theta_k \theta_l \quad \text{for } all\ k \neq l \tag{4.3}$$

최소 엔트로피 거리측도를 이용해서 구한 추정량은 비선형 추정량이지만 추정결과가 일반화 회귀추정량과 근사적으로 동일함으로 일반화 회귀추정량의 이중추출이론의 분산추정량에 근사 시켜서 정의한다.

$$\begin{aligned}\hat{V}(\hat{Y}_{Dw}) = & \sum\sum_r (d_k^* d_l^* - d_{kl}^*)(g_k^* f_k e_k)(g_l^* f_l e_l) \\ & - \sum_r d_k^*(d_k^* - 1) g_k^M (g_k^M - 1)(g_{1k} f_k e_k)^2 + \sum_r d_k^{*2} g_k^M (g_k^M - 1) f_k^2 e_k^2\end{aligned} \tag{4.4}$$

여기서, f_k는 모수 추정시 발생하는 자유도의 상실을 수정한 인자를 나타낸다. $g_k^* = g_{1k} g_k^M$이고, $e_k = y_k - \boldsymbol{x}_k' \hat{\boldsymbol{B}}_{rv}$으로 정의된다. 여기서 $\hat{\boldsymbol{B}}_{rv} = (\sum_r d_k^* g_k^M c_k \boldsymbol{x}_k \boldsymbol{x}_k')^{-1} \sum_r d_k^* g_k^M c_k \boldsymbol{x}_k y_k$이고, c_k는 기지의 양의 가중치로 여기서는 1로 놓는다. 식(4.4)에서 이용한 조정가중치 $g_k^* = g_{1k} g_k^M$에서 g_{1k}는 Lundström 등(1999)이 제안한 가중치와 동일하나 응답확률을 조정한 가중치 g_k^M는 동시적 방법을 이용한 경우와 2단계로 구분한 방법을 이용한 경우가 서로 다르다. 간편하게 두 가중치의 형태를 비교하기 위해서 1차 수렴한 $\boldsymbol{\lambda}$를 이용해서 얻은 결과를 테일러 전개를 이용해서 1차항으로 근사시켜 다음과 같은 결과를 얻었다. 먼저 동시적 방법을 적용한 경우의 응답확률 조정 가중치는

$$g_k^M \approx 1 + c_k(\sum_s d_k \boldsymbol{x}_k - \sum_r d_k \boldsymbol{x}_k)'(\sum_r d_k c_k \boldsymbol{x}_k \boldsymbol{x}_k')^{-1} \boldsymbol{x}_k \tag{4.5}$$

이고, 본 논문에서 제안한 2단계로 구분한 방법을 이용해서 구한 가중치는

$$g_k^M \approx 1 + c_{2k}(\sum_s w_{1k} \boldsymbol{x}_k - \sum_r w_{1k} d_{2k} \boldsymbol{x}_k)'(\sum_r d_k^* g_{1k} c_{2k} \boldsymbol{x}_k \boldsymbol{x}_k')^{-1} \boldsymbol{x}_k \tag{4.6}$$

이다. 두 식 모두에서 c-가중치는 1로 놓는다. 식(4.5)와 식(4.6)은 표본에 대한 보조정보가 주어진 경우에 대하여 보정방정식을 만족하는 g-가중치를 구한 식인데, 식(4.6)에서는 $\sum_s w_{1k} \boldsymbol{x}_k = \sum_U \boldsymbol{x}_k$를 만족하므로 엄밀한 의미에서 식(4.5)가 가지는 정보보다 더 많은 정보를 함축하고 있다. 제시한 분산 추정량들에 대한 효율성을 이론적으로 수식화하는 것은 매우 복잡하고 어렵기 때문에 모의실험을 통해서 동시적 방법을 이용한 경우와 제안한 방법의 효율성을 비교 판단해 보았다.

5. 모의 실험(simulation)

모의실험을 통해서 살펴보고자 하는 것은 다음과 같다. 먼저 보정 추정량의 상대 편향(Relative Bias)의 백분율을 다음과 같이 정의하였다.

$$RB_{sim}(\hat{Y}_w) = \left(\frac{E(\hat{Y}_w) - Y}{Y}\right) \times 100(\%) \tag{5.1}$$

여기서 $E(\hat{Y}_w)=\frac{1}{K}\sum_{k=1}^{K}\hat{Y}_{w(k)}$은 K개 표본에 대하여 구한 보정추정량들의 기대값이다. 또한 분산추정량에 대해 상대 편향의 백분율은 다음과 같다.

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

여기서 $E(\hat{V}(\hat{Y}_w))=\frac{1}{K}\sum_{k=1}^{K}\hat{V}_k(\hat{Y}_w)$, $V_{sim}=\frac{1}{K}\sum_{k=1}^{K}(\hat{Y}_{w(k)}-E(\hat{Y}_w))^2$이고,

$\hat{V}_k(\hat{Y}_w)$은 표본 k에 대한 분산추정량 값이다. 이와 더불어 95% 수준에서 근사적으로 타당하게 적용되는 신뢰구간에 대하여 다음과 같은 포함율(coverage rate)도 구해보고자 한다.

$$CR_{sim}[(\hat{V}(\hat{Y}_w))]=\sum_{k=1}^{K}I_{(k)}/100 \tag{5.3}$$

여기서, $I_{(k)}=\begin{cases}1 & , [a_{1k},\ a_{2k}]\ni Y\\ 0 & , \text{그외}\end{cases}$

이고, $a_{1k}=\hat{Y}_{w(k)}-1.96[\hat{V}_k(\hat{Y}_w)^{1/2}]$, $a_{2k}=\hat{Y}_{w(k)}+1.96[\hat{V}_k(\hat{Y}_w)^{1/2}]$이다.

총합 추정량과 분산추정량의 상대편향 백분율, 그리고 포함율은 동시적 방법과 제안한 방법 각각에 대하여 위의 식들을 이용해서 구했다. 위에서 제시한 값들을 구하기 위한 모의실험을 위해서 다음과 같이 구성된 모집단을 발생하였다. 모집단 크기 $N=1000$로부터 표본크기 $n=200$개를 단순임의추출 하고, 응답표본은 제시된 응답률에 따라 베르누이 추출에 의해서 얻었다. 다음 식에 의해서 관심변수에 대한 자료를 난수 발생시켰다.

$$y=3+\sqrt{S_y^2(1-\rho^2)}\times seed(8987878)+\rho S_y\times seed(2348789) \tag{5.4}$$

또한 보조변수에 대한 자료는 식

$$x=4+S_x\times seed(2348789) \tag{5.5}$$

을 이용해서 얻었다. 식(5.4)와 식(5.5)에서 ρ는 관심변수와 보조변수간의 상관계수이고, $S_x^2=50$, $S_y^2=50$으로 놓는다. 모의실험은 총 $K=5000$번 반복 실험을 실시했다. 또한 여러가지의 응답율(0.7, 0.8, 0.9)에 따라 최소엔트로피 거리측도를 이용해서 총합에 대한 보정추정량의 상대편향과 분산추정량의 상대편향을 구하였고, 95%수준에서의 포함율을 구하였다. 모의실험에 대한 분석 결과는 아래 [표1], [표2], 그리고 [표3]에서 보였다.

[표1] 응답율과 상관계수에 따른 총합 추정량의 상대편향(%)

응답율 / ρ / 방 법	0.7			0.8			0.9		
	0.7	0.8	0.9	0.7	0.8	0.9	0.7	0.8	0.9
동시적 방법	-0.65	-0.48	-0.27	-0.37	-0.27	-0.15	-0.24	-0.16	-0.09
2단계로 구분한 방법	-0.11	-0.09	-0.06	-0.09	-0.07	-0.05	-0.05	-0.05	-0.03

[표2] 응답율과 상관계수에 따른 분산 추정량의 상대편향(%)

응답율 / ρ / 방 법	0.7			0.8			0.9		
	0.7	0.8	0.9	0.7	0.8	0.9	0.7	0.8	0.9
동시적 방법	-2.44	-2.09	-1.45	-1.67	-1.57	-1.38	-2.20	-2.12	-1.95
2단계로 구분한 방법	-1.33	-1.11	-0.68	-1.04	-0.99	-0.89	-1.79	-1.75	-1.65

[표3] 응답율과 상관계수에 따른 95%수준에서의 포함율

응답율 / ρ / 방 법	0.7			0.8			0.9		
	0.7	0.8	0.9	0.7	0.8	0.9	0.7	0.8	0.9
동시적 방법	92	92.5	93.4	93.9	94.1	94.5	94.3	94.4	94.5
2단계로 구분한 방법	94.9	94.8	94.9	95	94.9	94.7	94.6	94.5	94.5

모의 실험을 통해 총합추정량과 분산추정량의 상대편향을 구해본 결과를 살펴보면 [표1]의 총합추정량에 대한 상대편향은 동시적 방법을 이용한 경우보다 제시한 이중추출기법으로 구한 보정가중치를 사용한 경우가 추정량에 대한 무응답 편향이 더 작게 나왔으며, 특히 낮은 응답율 일수록 제안된 방법의 효율성이 상당히 높게 나타났다. 또한 보조변수와 관심변수간의 상관관계가 높을수록 추정량의 상대편향이 약간 더 작게 나타났으며, 일반적으로 상관관계의 크기가 0.6이상인 경우에는 보조변수의 사용이 효율적이다. [표2]의 분산추정량에 대한 상대편향 역시 동시적으로 조정가중치를 구하는 경우보다 2단계로 구분한 방법을 이용한 경우가 더 효율적인 것으로 나타났다. 또한 [표3]의 95% 수준에서 포함율도 상당히 적합함을 알 수 있었다.

6. 결론

일반적으로 단위 무응답이 발생했을 때, 무응답 편향을 감소시키기 위한 방법으로 가중치 조정법을 사용한다. 본 논문에서는 단위 무응답이 존재하는 경우 가중치 조정방법중에서 보정추정법을 적용하여 2단계로 구분하여 추출가중치 조정하고 무응답 단위에 대하여 응답확률을 보정한 후, 총합에 대한 추정량과 그에 따르는 분산추정량을 도출하였다. 총합에 대한 무응답 보정추정량이 제

시된 응답확률을 사용함으로서 불편추정량이 됨을 보였으며, 모의실험을 통해서 총합추정량뿐만 아니라 분산추정량도 동시적 방법보다 좀 더 효율적임을 보였다. 또한 본 논문에서는 최소엔트로피 거리함수를 사용함으로써 음의 가중치가 발생되지 않게 하였다.

차후 연구과제로는 Singh과 Mohl(1996)의 논문에서 제시하고 있는 실제적인 거리함수를 이용해서 보정가중치를 구해보고, 총합 추정량 이외에 총계비율에 대한 추정량, 모평균 추정량 등과 같은 형태를 다뤄보고자 한다.

참고문헌

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

[2] 염준근, 손창균. 정영미(2002). 이중 추출 방법을 이용한 단위 무응답의 가중치 조정방법에 관한 연구, 「한국통계학회 춘계발표 논문집」. 13-18

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

[4] Hidiroglou, M. A. and Deville, J. C. (1995). Use of Auxiliary Information for Two-phase sampling. 873-878.

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

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

[7] Särndal, C. E., and Swensson, B. (1987). A General View of Estimation for Two Phases of Selection with Applications to Two-Phase Sampling and Nonresponse. *International Statistical Review. 55,* 279-294

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

[9] Singh, A. C., and Mohl, C. A. (1996). Understanding Calibration Estimators in Survey Sampling. *Survey Methodology,* 22, 107-115

[2002년 5월 접수, 2002년 8월 채택]