이중 추출 방법을 이용한 단위 무응답의 가중치 조정방법에 관한 연구

염 준 근[1] · 손 창 균[2] · 정 영 미[3]

요 약

이중추출(two-phase)접근방법 이용의 주목적은 관심변수와 보조변수사이의 관계를 이용해서 더 좋은 추정을 하고자 하는 것이다. 특히 이 방법은 층화, 무응답 문제에 적용하는 경우 상당히 효과적이다. 본 논문에서는 무시할 수 있는 무응답이 발생했을 때 이중추출기법을 이용해서 g-가중치와 응답확률을 각 단계별로 조정해줌으로써 무응답 보정추정량과 분산추정량을 구했다.

주요용어 : 단위 무응답, 보조변수, 이중추출, 보정추정량, 분산추정량

1. 서 론

일반적으로 조사에 있어서 무응답은 조사 항목(item)에 발생하는 경우와 조사 단위(unit)에 대해 발생하는 경우로 나눌 수 있고, 이러한 두 가지 무응답 상황을 각각 항목 무응답(item nonresponse)과 단위 무응답(unit nonresponse)으로 정의한다. 무응답을 제외한 응답자들만의 분석은 무응답 편향으로 인하여 추정량에 심각한 영향을 줄 수 있기 때문에 조사과정에서 발생하는 무응답으로 인한 편향(bias)을 줄이는 것은 중요한 관점이다. 항목 무응답을 처리하기 위한 방법으로는 무응답된 항목을 다른 값으로 대체해주는 대체(imputation) 방법이 있고, 단위 무응답을 처리하기 위해서는 추정시 응답된 조사자료들의 가중치를 조정해줌으로써 무응답으로 인해서 발생하는 효과를 고려해주는 방법인 가중치 조정(weighting adjustment) 방법이 있다. 본 논문에서는 단위 무응답이 발생했을 때 여러 가지 가중치 조정방법 중에서 보정추정(calibration estimation)방법을 사용하여 무응답을 처리하고자 한다.

Särndal and Swensson(1987)은 단위무응답이 존재하는 경우 층화이중추출 방법을 적용해서 추정량을 구했고, Lundström and Särndal(1999)은 관심변수와 강한 상관이 존재하는 보조변수의 기지의 모집단 총합과 표본 총합을 이용하여 관심변수의 총합 추정량과 분산 추정량을 도출하였다. 손창균, 홍기학 그리고 이기성(2000)은 단위무응답이 존재할 때 기지의 모집단 총합에 대한 정보뿐만 아니라 보조변수에 대한 기지의 분산 정보를 가지고 관심변수에 대한 분산을 추정하는 연구를 실시했다.

본 논문에서는 조사 단위에 무응답이 발생하는 경우에 대해 이중추출기법을 적용해서 미지의 응답확률을 보정한 후 총합에 대한 추정량과 분산추정량을 구했다. 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, 3

1) (100-715) 서울시 중구 필동 3가 동국대학교 통계학과 교수
2) (520-714) 전남 나주시 대호동 동신대학교 컴퓨터학과 전임강사
3) (100-715) 서울시 중구 필동 3가 동국대학교 대학원 통계학과 박사과정 수료

장에서는 각 단계별로 보조정보를 이용해서 추출가중치와 응답확률을 조정 한 후 최종 보정 추정량을 도출하였다. 4장에서는 모집단 수준의 정보를 이용해서 분산추정량을 도출하였고, 5장과 6장에서는 몬테칼로 모의 실험을 통해 추정량의 상대편향을 구한 결과를 가지고 연구를 통한 결론을 다루었다.

2. 모집단 총합에 대한 추정량

유한모집단 $U=\{1,\cdots,k,\cdots,N\}$에 대하여 모집단 총합 $Y=\sum_U y_k$을 추정한다고 하자. 이때 관심변수는 y이고, y_k는 k번째 단위에 대한 y의 값이다. 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 $\pi_{kl} = p(k \text{and } l \in s)$이다. π_k의 역수인 $\pi_k^{-1} = d_k$는 단위 k에 대한 설계가중치이며, 또한 $\pi_{kl}^{-1} = d_{kl}$이다. 모집단 총합을 추정하기 위해 추출된 표본 s에 대해 크기가 $m(\leq n)$인 응답집합 $r(\subseteq s)$을 얻었다고 하자. 이러한 무응답을 조정하기 위해 관심변수와 강한 양의 상관을 가진 보조 정보를 이용할 수 있다면, 추출오차와 무응답 편향을 충분히 감소시킬 수 있다. 따라서, 보조변수 벡터 $\boldsymbol{x}$를 가정하고, 이 벡터는 관심변수와 강한 상관을 가진다고 하자. 이 때, k번째 단위에 대한 보조변수 벡터 값을 $\boldsymbol{x}_k$라 한다.

Särndal, Swensson and Wretman(1992)는 2가지 경우의 "정보의 차원"인 표본에 대한 정보와 모집단에 대한 정보로 나누어 보정 추정과정을 전개하였으나, 본 논문에서는 모집단 수준의 정보만을 고려했다.

응답확률이

$$pr(k \in r|s) = \theta_k, \quad pr(k \& l \in r|s) = \theta_{kl}$$

으로 알려진 특정한 응답분포 $q(r|s)$하에서 모집단에 대한 보조정보에 따라 구한 Y의 추정량은 식 (2.1)과 같다.

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

여기서, g-가중치는

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

이며, q_k는 특정한 양의 인자로서 추정량의 형태를 결정한다. 또한 이용 가능한 보조정보가 모집단에 대한 정보라는 것은 $\boldsymbol{x}_k$는 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 것이다. 즉, $\boldsymbol{x}_k$는 정보가 완전한 모집단 수준까지 이용가능한 벡터이다. y_k는 모든 $k \in r$에 대해서만 관측된다.

식(2.2)에서 실제로 응답확률 θ_k는 미지이며, 따라서 임의의 값인 $\hat{\theta}_k$로 대체해야 한다. 우선 관련된 응답모형을 세우고, 그 다음에 미지의 응답확률을 추정하는 전형적인 절차로부터 구한 Y의 추정량은 다음과 같다.

$$\hat{Y} = \sum_r d_k \nu_{1k} \nu_{2k} y_k \tag{2.3}$$

이때, $\nu_{1k} = \hat{\theta}_k^{-1}$이며, ν_{2k}는 식(2.2)의 g-가중치와 같고, g-가중치의 θ_k를 $\hat{\theta}_k$로 대체한다. 식(2.3)을 구하기 위해서 LS(1999)는 두 가중치를 동시에 보정한 가중치를 이용했으나 다음

장에서는 가중치와 응답확률을 각 단계별로 조정한 새로운 보정가중치를 이중추출기법을 이용해서 구해보았다.

3. 무응답에 대한 보정추정량

무응답이 있는 경우와 이중추출간의 유사한 점은 처음에 추출된 표본은 s이나 실제로 측정된 연구변수를 나타내는 최종표본은 s의 부분집합인 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 이용되고, 초기가중치와 가능한 한 밀접한 새로운 보정가중치를 구하기 위해서는 거리함수가 이용된다. Devile and Särndal(92)이 여러 가지 거리함수를 제안했고, 여기에서는 임의의 단위 집합 $k\in r$ 에 대하여, 일반화최소제곱 거리함수(GLS)를 이용하고자 한다. 그러나 GLS 거리함수의 단점은 음의 가중치가 발생할 수 있다는 것이다. 음의 가중치는 결과의 의미있는 해석을 어렵게 하므로 단점을 보완해주는 방법으로 주어진 제약조건을 만족할 뿐만 아니라 g-가중치가 임의의 범위제한(Range Restrictions(RR))을 만족하도록 제한성이 부여된 제한된(Restricted) 일반화최소제곱(RGLS) 거리함수가 Devile and Särndal(92)에 의해서 제안되었다.

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 부차표본(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. $d_{2k}=1/\pi_{k|s}$ 는 단위 k의 2단계 추출 가중치이다.

각 단계마다 주어진 제약 조건 즉, 보정방정식을 만족한다는 조건하에서 GLS 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 가중치로 1로 놓는다. $\sum_s w_{1k}\,\boldsymbol{x}_k=\sum_U \boldsymbol{x}_k$을 만족한다는 조건하에서 다음과 같은 RGLS 거리함수를 최소화하는 1단계 보정 추출 가중치 w_{1k}을 결정할 수 있다.

$$\sum_s \frac{(w_{1k}-d_{1k})^2}{d_{1k}c_{1k}} \qquad \text{if} \quad L<w_{1k}/d_{1k}<U \tag{3.1}$$

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

$$g_{1k}=1+c_{1k}\Big(\sum_U \boldsymbol{x}_k-\sum_s d_{1k}\,\boldsymbol{x}_k\Big)'\,T_1^{-1}\,\boldsymbol{x}_k \tag{3.2}$$

이며, $T_1=\sum_s d_{1k}c_{1k}\,\boldsymbol{x}_k\,\boldsymbol{x}_k'$ 이다.

식(3.1)에서 $L<1<U$를 만족하는 하한과 상한을 조건으로 지정할 수 있고, 또한 양의 가중치를

보장하기 위해서 $L>0$을 선택한다.

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.(Hidiroglou, and Deville(1995)). 수정된 거리함수인 RGLS를 이용한 추정결과가 원래의 GLS 거리함수를 이용한 결과와 점근적으로 동등하기 때문에 2단계에서는 GLS 거리함수를 이용해서 응답확률에 대한 가중치를 보정하고자 한다.

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.

$$\sum_r \frac{(w_k^*-w_{1k}d_{2k})^2}{w_{1k}d_{2k}c_{2k}} \tag{3.3}$$

여기서, $\{c_{2k}: k\in r\}$는 미리 조건으로서 주어진 기지의 양의 가중치이다.

이러한 보정을 실시한 결과 가중치는 다음과 같이 최종 보정된 가중치로 정의된다.

$$w_k^*=d_k^*(g_{1k}g_k^M) \tag{3.4}$$

여기서, $k\in r$에 대해

$$g_k^M=1+c_{2k}(\sum_s w_{1k}\boldsymbol{x}_k-\sum_r w_{1k}d_{2k}\boldsymbol{x}_k)'(T^M)^{-1}\boldsymbol{x}_k \tag{3.5}$$

이며, $T^M=\sum_r d_k^*g_{1k}c_{2k}\boldsymbol{x}_k\boldsymbol{x}_k'$ 이다.

따라서, 식(2.3)에서 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.

$$\hat{Y}_{Dw}=\sum_r w_k^*y_k \tag{3.6}$$

4. 분산추정량

Lundström and Särndal(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 도출하였다. 만일 강한 보조정보를 이용할 수 있다면, 무응답 편향은 근사적으로 0이 되며, MSE는 추정량의 분산으로 표현된다. 분산추정 결과는 Särndal, Swensson, and Wretman(1992)의 연구를 바탕으로 전개되었다. 분산추정량을 구하기 위해서 표본개체가 독립적으로 응답한다는 다음과 가정을 제시한다.

$$pr(k\&l\in r|s)=\theta_{kl}=\theta_k\theta_l \quad \text{for all } k\neq l \tag{4.1}$$

이러한 가정 하에서 Särndal 등(1992)이 제안한 방정식 (9.7.22)로 부터 유도된 $\hat{Y}_{ssw,U\theta}$에 대한 분산추정량은 다음과 같이 얻어진다.

$$\hat{V}(\hat{Y}_{ssw,U\theta})=\sum\sum_r(d_kd_l-d_{kl})(g_{1k}e_{k\theta}/\theta_k)(g_{1l}e_{l\theta}/\theta_l)$$
$$-\sum_r d_k(d_k-1)(1-\theta_k)(g_{1k}e_{k\theta}/\theta_k)^2+\sum_r d_k^2(1-\theta_k)(g_{sk\theta}e_{k\theta}/\theta_k)^2$$

여기에서, $g_{sk\theta} = 1 + c_k(\sum_s w_{1k}\ \boldsymbol{x}_k - \sum_r w_{1k} d_{2k}\ \boldsymbol{x}_k/\theta_k)'(\sum_r d_k c_k\ \boldsymbol{x}_k\ \boldsymbol{x}_k'/\theta_k)^{-1}\ \boldsymbol{x}_k$ 이다. 따라서, 3장에서 구한 조정가중치와 보정된 응답확률을 다음과 같이 분산추정량에 적용할 수 있다.

$$\begin{aligned}\hat{V}(\ \hat{Y}_{Dw}) = &\sum\sum_r (d_k^* d_l^* - d_{kl}^*)(g_k^* f_k e_k)(g_l^* f_l e_l) \\ &- \sum_r d_k^*(d_k^* - 1) g_k^M (g_k^M - 1)(g_{1k} f_k e_k)^2 + \sum_r d_k^{*2} g_k^M (g_k^M - 1) f_k^2 e_k^2\end{aligned} \tag{4.2}$$

여기서, f_k는 모수를 추정함으로써 발생하는 자유도의 상실을 수정한 인자이다. 그리고, $e_k = y_k - \boldsymbol{x}_k'\ \hat{\boldsymbol{B}}_{rv}$ 이며, $\hat{\boldsymbol{B}}_{rv} = (\sum_r d_k^* g_k^M c_k\ \boldsymbol{x}_k\ \boldsymbol{x}_k)^{-1} \sum_r d_k^* g_k^M c_k\ \boldsymbol{x}_k y_k$이다.

5. 모의 실험(simulation)

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

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

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

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

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

$\hat{V}_k(\ \hat{Y}_{Dw})$은 응답표본 k에 대한 분산추정량 값이다.

이와 더불어 95% 신뢰구간에 대하여 다음과 같은 포함율(coverage rate)도 구해보고자 한다.

$$CR[E(\ \hat{V}(\ \hat{Y}_{Dw}))] = \sum_{k=1}^{K} I_{(k)}/100 \tag{5.3}$$

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

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

6. 결 론

일반적으로 단위 무응답이 발생했을 때, 무응답 편향을 감소시키기 위한 방법으로 가중치 조정법을 사용한다. 가중치 조정방법에는 표준가중값 조정법, raking ratio, 보정추정법 등이 있다. 그 중에서도 본 논문에서는 단위 무응답이 존재하는 경우 이중추출방법에 보정추정법을 적용하여 단계별로 무응답 단위에 대해 추출가중치와 응답확률에 대한 보정을 실시한 후, 총합에 대한 보정추정량을 구하고 그에 따르는 분산추정량을 도출하였다. 모의 실험을 통해 총합추정량과 분산추정량의 상대편향을 구해본 결과 이중추출기법으로 구한 보정가중치를 사용함으로써 추정량의 무응답 편향이 무시할 만큼 작게 나왔으며, 분산추정량이 상당히 안정적임을 살펴보았다.

참고 문헌

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

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

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

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

6. Särndal, C. E., 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,* pp 279-294