Journal of the Korean Data & Information Science Society
2015, **26**(2), 377–385

http://dx.doi.org/10.7465/jkdi.2015.26.2.377
한국데이터정보과학회지

재조사에서 효율 향상을 위한 추정법 연구[†]

박현아[1] · 나성룡[2]

[1]서울대학교 통계학과 · [2]연세대학교 정보통계학과

접수 2015년 2월 23일, 수정 2015년 3월 16일, 게재확정 2015년 3월 18일

요 약

표본조사에서 무응답이 발생한 개체에 대해 재조사 실시한 후 보조변수를 사용한 회귀추정의 형태를 가지는 추정량을 제시하고 복제치 기법을 이용한 분산추정량을 연구한다. 또한 응답여부에 따른 응답확률의 모수적 추론방법도 함께 제시한다. 재조사 후 모평균에 대하여 불편성을 만족하고 효율이 좋은 추정량과 일치성을 가지는 분산추정량을 이론적으로 연구하고 모의실험을 통하여 연구의 타당성을 입증한다.

주요용어: 보조변수, 분산추정, 응답확률, 재조사.

1. 서론

표본조사에서 발생하는 무응답의 증가는 자료의 품질 및 추정치에 영향을 주기 때문에 무응답을 보정하는 방안들이 필요하게 되는 데 그와 같은 방식으로는 재조사와 대체와 무응답 가중치 보정과 같은 것들이 있다. 이와 같이 무응답은 여러 방식으로 보정이 가능하지만 일차적으로 응답률을 높이기 위한 방안을 고려하자면 다시 방문하는 것과 같은 재조사 방식이 적절하다. 재조사에 대한 연구로는 여러 논문들이 있는 데 Deming (1953)은 부재자 가구를 i회 재조사한 다음 i회까지의 조사에서 얻는 응답을 가지고 모평균을 추정하는 문제를 연구하였고 재조사가 실시되는 개체에 대해 부차표본을 추출하는 방식의 연구로는 Hansen과 Hurwitz (1946), Elliott 등 (2000), Okafor 등 (2000), Ismail 등 (2011)이 있으며 재조사 후 보조정보를 사용한 연구로는 Park 등 (2008)과 Park과 Jeon (2010)이 있고 Han과 Byun (2014)는 재조사가 무응답의 편향을 감소시킬 수 있음을 연구하였다.

재조사가 필요한 개체는 조사시점에서 응답여부에 따라 달라지며 또한 재조사를 실시 했을 때 또 다시 무응답이 발생할 수 있다. 그러므로 본 연구에서는 첫번째 조사에서의 개체와 재조사를 실시하여 얻어지는 개체를 사용한 추정을 위하여 응답여부를 나타내는 확률변수와 응답확률을 사용하는 추정기법을 제안하고 동시에 추출틀에서 조사될 수 있는 보조변수 정보를 사용하는 추정기법을 연구한다. 제안된 추정량은 Park 등 (2008)에서 제시된 추정량의 확장된 기법으로써 본 연구에서는 모평균에 대하여 불편성을 만족하며 효율이 높아지는 것을 제시한다. 그리고 그 추정량의 산포에 대한 분산추정량으로 일치성을 만족하는 복제치 분산 추정기법을 연구한다. 이와 동시에 첫번째 조사와 재조사에서의 응답확률의 추정방법도 고려한다.

[†] 이 논문은 2012년도 정부 (교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NRF-2012R1A1A3003761)

[1] 교신저자: (151-742) 서울시 관악구 관악로 1, 서울대학교 통계학과, 연구원. E-mail: hapk@daum.net
[2] (220-710) 강원도 원주시 홍업면 매지리 234, 연세대학교 정보통계학과, 교수.

논문의 구성은 제 2절에서 재조사 후 보조변수에 의한 회귀추정기법을 이용한 추정량의 근사적 불편성과 분산을 계산하고 응답확률의 추정을 연구한다. 또한 모평균의 추정량의 분산에 대한 복제치 분산 추정 기법으로 잭나이프기법을 연구한다. 제 3절에서는 제안된 추정량에 대한 근사불편성과 효율성을 기존의 추정량들과 비교한 것을 가상 모집단을 통하여 모의실험하고 제안된 잭나이프 분산추정량이 일치성을 만족하는 지를 여러 척도에 의해 살펴본다.

2. 재조사 후 추정과 분산추정방법

표본설계에 의해 추출된 표본을 조사할 때 무응답이 발생하고 그 개체에 대해 재조사를 실시하게 되는 데 모평균에 대한 추정방법을 연구하기 위해서는 모집단에서의 표본추출 여부과 응답여부와 응답확률을 정의할 필요가 있다. 모집단의 크기를 N이라 하고 관심변수를 y_i라 할 때 모집단은 $U = \{1, 2, ..., N\}$이고 관심모수인 모평균은 $\mu_y = N^{-1}\sum_{i=1}^{N} y_i$이다. 표본설계에 의해 크기 n의 표본을 추출하고 그때 추출된 표본은 $S = \{1, 2, ..., n\}$이고 설계가중치를 w_i라 할 때 모평균에 대한 추정량은 $\bar{y} = \sum_{i=1}^{n} w_i y_i$이다. 이 추정량은 표본설계에 의해

$$E(\bar{y}|U) = \mu_y. \tag{2.1}$$

이 만족되는 것을 가정한다.

첫번째 표본조사에서 응답여부에 대한 확률변수는

$$R_i = \begin{cases} 1, & i\text{번째 개체가 응답됨 } (i \in S) \\ 0, & \text{그외} \end{cases}$$

이고 이때 응답확률은 $\pi_i = Pr(R_i = 1|i \in S)$이다. 그리고, 첫번째 조사에서 무응답 개체의 집합 표본 $S_{NR} = \{i : R_i = 0, i \in S\}$에 대해 재조사를 실시할 때 응답여부에 대한 확률변수는

$$C_i = \begin{cases} 1, & i\text{번째 개체가 응답됨 } (i \in S_{NR}) \\ 0, & \text{그외} \end{cases}$$

이고 이때 응답확률은 $p_i = Pr(C_i = 1|i \in S_{NR})$이다. 여기서 R_i와 C_i에 관하여 MAR (Missing at random)을 가정한다.

재조사 후에 추정방법을 정의하기 위해 관심변수와 상관이 높고 n개의 표본에 대한 자료를 모두 얻을 수 있는 보조변수 x_i가 있다고 가정하고 필요한 추정량 $\bar{y}_T = \sum_{i=1}^{n} w_i(1-\pi_i)^{-1}p_i^{-1}(1-R_i)C_i y_i$, $\bar{x}_T = \sum_{i=1}^{n} w_i(1-\pi_i)^{-1}p_i^{-1}(1-R_i)C_i x_i$, $\bar{x}_R = \sum_{i=1}^{n} w_i(1-\pi_i)^{-1}(1-R_i)x_i$에 대하여 최종적인 추정량은

$$\bar{y}_C = \phi\bar{y}_f + (1-\phi)\bar{y}_S \tag{2.2}$$

이다. 여기에서 $\bar{y}_f = \sum_{i=1}^{n} w_i\pi_i^{-1}R_i y_i$, $\bar{y}_S = \bar{y}_T + \hat{\gamma}(\bar{x}_R - \bar{x}_T)$ 이고 $\bar{x}_y = \sum_{i=1}^{n} w_i(1-\pi_i)^{-1}p_i^{-1}(1-R_i)C_i x_i y_i$, $\bar{x}_2 = \sum_{i=1}^{n} w_i(1-\pi_i)^{-1}p_i^{-1}(1-R_i)C_i x_i^2$에 대하여 $\hat{\gamma} = \bar{x}_2^{-1}\bar{x}_y$이다. 추정량 $\bar{y}_C$는 첫번째 조사에서의 $\bar{y}_f$와 재조사에서의 $\bar{y}_S$을 ϕ로 가중합하는 형태를 지닌다. 여기서 가중치 ϕ를 결정하기 위해서는 $Var(\bar{y}_C)$를 최소화하면 되는 데 그때 ϕ는 $[Var(\bar{y}_f) - 2Cov(\bar{y}_f, \bar{y}_S) + Var(\bar{y}_S)]^{-1}[Var(\bar{y}_f) - Cov(\bar{y}_f, \bar{y}_S)]$이다.

제안된 추정량 (2.2)의 성질인 불편성과 효율을 살펴보기 위해 다음과 같은 가정이 필요하다.

(A1) Isaki와 Fuller (1982)에서 제시된 유한 모집단과 표본에 대한 배열을 고려한다. 유한모집단은 어떤 $\tau > 0$에 대하여

$$N^{-1}\sum_{i=1}^{N} z_i^{2+\tau} = O(1),$$

을 가정하며 $z_i' = (x_i, y_i)$이다.

(A2) 음수가 아닌 상수 C_1, C_2, C_3에 대하여

$$C_1 < \pi_i < C_2, \quad p_i > C_3.$$

을 가정한다.

(A3) 응답변수 R_i, C_i는 각각 서로 독립을 가정한다.

(A4) 음수가 아닌 상수 D_1, D_2, D_3, D_4에 대해

$$D_1 < \max_i(nw_i) < D_2$$

과

$$D_3 < nVar(\bar{y}|U) < D_4,$$

을 가정한다. 여기서 $Var(\bar{y}|U)$은 표본설계에 의한 분산이다.

(A5) 표본적률이 모집단적률로 수렴하는 것을 가정한다.

$$\sum_{i=1}^{n} w_i a_i z_i z_i' - N^{-1}\sum_{i=1}^{N} z_i z_i' = O_p(n^{-1/2})$$

여기서 $a_i = (1-\pi_i)^{-1}p_i^{-1}(1-R_i)C_i$이다.

정리 2.1 (2.1)과 (A1)∼(A5)가정하에

$$E(\bar{y}_C) = \mu_y + o(n^{-1/2}) \tag{2.3}$$

의 불편성과

$$Var(\bar{y}_C) = Var(\bar{y}|U) + \frac{E_1E_2 - E_3^2}{E_1 + E_2 + 2E_3} + o(n^{-1}) \tag{2.4}$$

식이 성립하며, 여기에서

$$\begin{aligned} E_1 &= E(\sum_{i=1}^{n} w_i^2(\pi_i^{-1} - 1)y_i^2|U) \\ E_2 &= E(\sum_{i=1}^{n} w_i^2(1-\pi_i)^{-1}[(p_i^{-1} - 1)(y_i - \gamma x_i)^2 + \pi_i y_i^2]|U) \\ E_3 &= E(\sum_{i=1}^{n} w_i^2 y_i^2|U) \end{aligned}$$

이고 $\gamma = (\sum_{i=1}^{N} x_i^2)^{-1}(\sum_{i=1}^{N} x_i y_i)$이다.

증명: 먼저 $\bar{y}_S$를 전개하면

$$\bar{y}_S = \bar{y}'_S + (\hat{\gamma} - \gamma)(\bar{x}_R - \bar{x}_T) \tag{2.5}$$

와 같고 여기서 $\bar{y}'_S = \bar{y}_T + \gamma(\bar{x}_R - \bar{x}_T)$이다. (A5)와 $\hat{\gamma}$의 테일러 전개를 사용하여

$$\hat{\gamma} - \gamma = (\sum_{i=1}^{N} x_i^2/N)^{-1}[(\bar{x}_y - \sum_{i=1}^{N} x_i y_i/N) - \gamma(\bar{x}_2 - \sum_{i=1}^{N} x_i^2/N)] + o_p(n^{-1/2})$$

이고 (A1), (A3), (A4)를 사용하면

$$E(\bar{x}_T - N^{-1}\sum_{i=1}^{N} x_i)^2 = Var(\sum_{i=1}^{n} w_i x_i|U) + E[\sum_{i=1}^{n} w_i^2(1-\pi_i)^{-1}(\pi_i + p_i^{-1} - 1)x_i^2|U] = O(n^{-1})$$

이며

$$E(\bar{x}_R - N^{-1}\sum_{i=1}^{N} x_i)^2 = Var(\sum_{i=1}^{n} w_i x_i|U) + E[\sum_{i=1}^{n} w_i^2\pi_i(1-\pi_i)^{-1}x_i^2|U] = O(n^{-1})$$

이다. 위의 식들을 사용하면 식 (2.5)는

$$\bar{y}_S = \bar{y}'_S + o_p(n^{-1/2})$$

이다. 한편 $\phi' = [Var(\bar{y}_f) - 2Cov(\bar{y}_f, \bar{y}'_S) + Var(\bar{y}'_S)]^{-1}[Var(\bar{y}_f) - Cov(\bar{y}_f, \bar{y}'_S)]$에 대하여 $\phi - \phi' = o(1)$이며 이를 사용하여

$$\bar{y}_C = \phi'\bar{y}_f + (1-\phi')\bar{y}'_S + o_p(n^{-1/2})$$

을 전개하며 (A1)에 의해 식 (2.3)이 성립한다.

ϕ의 정의에 의해

$$Var(\bar{y}_C) = [Var(\bar{y}_f) + Var(\bar{y}'_S) - 2Cov(\bar{y}_f, \bar{y}'_S)]^{-1}[Var(\bar{y}_f)Var(\bar{y}'_S) - 2Cov(\bar{y}_f, \bar{y}'_S)^2] + o(n^{-1})$$

식이 성립한다. 또한 (A3)에 의해

$$\begin{aligned} Var(\bar{y}_f) &= Var(\bar{y}|U) + E_1 \\ Var(\bar{y}'_S) &= Var(\bar{y}|U) + E_2 \\ Cov(\bar{y}_f, \bar{y}'_S) &= Var(\bar{y}|U) - E_3 \end{aligned}$$

가 계산되며 결과 (2.4)식이 입증된다. □

위 정리 2.1의 (2.4) 결과와 $E_3^2 \le E_1E_2$의 사실에서 $\bar{y}_C$의 분산이 $\bar{y}$의 분산보다 $(E_1 + E_2 + 2E_3)^{-1}(E_1E_2 - E_3^2)$만큼 커짐을 알 수 있다. 한편 분산에는 여러 모수들이 존재하기 때문에 표본자료에 의해 추정량의 분산을 계산하기 위해서는 분산의 추정이 필요하다. 일치성과 같은 성질을 만족하는 분산추정을 위하여 복제치 기법으로 잭나이프 (Jackknife) 분산추정량을 제시한다. 분산추정량을 제시하기 위해 여러 기호를 먼저 정의하면 c_k는 k번째 복제(반복)에서 결정되어지는 요소로써 단순임의표집하에 $n^{-1}(n-1)$로 결정되어 지는 것과 L은 복제횟수로써 일반적으로 표본크기로 결정되어지는 것이 있다. 잭나이프 분산추정량에서의 k번째 복제라는 것은 k번째 자료를 제거하고 나머지 표본자료를 가지고 추정하는 것을 말한다. $\bar{y}_C$의 k번째 복제(반복)에서의 추정량이 $\bar{y}_C^{(k)} = \phi\bar{y}_f^{(k)} + (1-\phi)\bar{y}_S^{(k)}$일때 여기서

$\bar{y}_f^{(k)} = \sum_{i=1}^n w_i^{(k)} \pi_i^{-1} R_i y_i$이고 $\bar{y}_S^{(k)} = \bar{y}_T^{(k)} + \hat{\gamma}^{(k)}(\bar{x}_R^{(k)} - \bar{x}_T^{(k)})$이고 $w_i^{(k)}$는 k번째 복제에서 i번째 단위의 복제가중치이다. 잭나이프 분산추정량을 정의하면

$$\hat{V}(\bar{y}_C) = \sum_{k=1}^{L} c_k (\bar{y}_C^{(k)} - \bar{y}_C)^2$$

이고 여기서 $\bar{y}_T^{(k)} = \sum_{i=1}^n w_i^{(k)} (1-\pi_i)^{-1} p_i^{-1} (1-R_i) C_i y_i$, $\bar{x}_T^{(k)} = \sum_{i=1}^n w_i^{(k)} (1-\pi_i)^{-1} p_i^{-1} (1-R_i) C_i x_i$, $\bar{x}_R^{(k)} = \sum_{i=1}^n w_i^{(k)} (1-\pi_i)^{-1} (1-R_i) x_i$이고 $\bar{x}_y^{(k)} = \sum_{i=1}^n w_i^{(k)} (1-\pi_i)^{-1} p_i^{-1} (1-R_i) C_i x_i y_i$, $\bar{x}_2^{(k)} = \sum_{i=1}^n w_i^{(k)} (1-\pi_i)^{-1} p_i^{-1} (1-R_i) C_i x_i^2$에 대하여 $\hat{\gamma}^{(k)} = \bar{x}_2^{(k)-1} \bar{x}_y^{(k)}$이다.

응답확률 π_i와 p_i는 사전에 알 수 없기 때문에 추정되어야 한다. 또한 첫번째 조사에서의 추정량과 재조사에서의 추정량의 가중치 ϕ가 0에서 1사이의 값을 가지면 되지만 추정량의 분산을 최소화하는 것이 더 효율이 좋아지기 때문에 여러 추정량의 분산으로 표현되는 것을 사용해야 한다. 그러나 추정량의 분산들도 각각 모수가 되기 때문에 추정되어야 한다. 즉 응답확률 π_i와 p_i과 가중치 ϕ가 추정되어서 모평균에 의한 추정량이 결정이 되어야 한다. 그러므로 추정을 단계별로 살펴보면 첫째, π_i와 p_i를 추정하기 위하여 모수적 추정 방법인 로지스틱회귀모형을 가정하고 회귀모수를 추정하는 것 (Park 등, 2008, Park과 Jeon, 2010)과 비모수적 추정 방법인 응답률로 추정하거나 응답층별 표본 응답률로 추정하는 것 (Kim 등, 2005, Park과 Park, 2013) 등이 있다. 재조사 후 추정에 대해서는 추정량 $\bar{y}_C$에 있는 응답확률들을 추정된 확률 $\hat{\pi}_i, \hat{p}_i$로 변경하여 사용하면 된다. 둘째, 가중치 ϕ의 추정을 위해 $\hat{V}(\bar{y}_f) = \sum_{k=1}^L c_k (\bar{y}_f^{(k)} - \bar{y}_f)^2$과 $\hat{V}(\bar{y}_S) = \sum_{k=1}^L c_k (\bar{y}_S^{(k)} - \bar{y}_S)^2$과 $\hat{C}(\bar{y}_f, \bar{y}_S) = \sum_{k=1}^L c_k (\bar{y}_f^{(k)} - \bar{y}_f)(\bar{y}_S^{(k)} - \bar{y}_S)$을 정의하면 추정된 $\hat{\phi}$은

$$\hat{\phi} = [\hat{V}(\bar{y}_f) - 2\hat{C}(\bar{y}_f, \bar{y}_S) + \hat{V}(\bar{y}_S)]^{-1} [\hat{V}(\bar{y}_f) - \hat{C}(\bar{y}_f, \bar{y}_S)]$$

이다. 그리고 추정된 응답확률 $\hat{\pi}_i, \hat{p}_i$을 잭나이프분산추정량에 사용하기 위해서는 약간의 변경이 있어야만 일치성을 만족하는 추정이 된다. 그것은 k번째 복제(반복)가 실시되는 곳에서는 k번째 표본자료를 제거하고 나머지 표본자료를 가지고 추정하는 방법 $\hat{\pi}_i^{(k)}, \hat{p}_i^{(k)}$을 사용하는 것이다 (Kim과 Park, 2006).

3. 모의실험

모의실험을 통하여 이론적 근거를 살펴보기 위해 모의 실험의 횟수는 $B = 1000$이고 표본크기는 $n = 100$인 자료를 아래의 회귀 모형에서 발생시킨다. 표본추출방법으로는 단순랜덤복원추출을 가정한다.

$$y_i = 20 + 4x_i + \sqrt{x_i}\epsilon_i, \text{ for } i = 1, 2, ..., n$$

단 여기서 $\epsilon_i \sim N(0,1)$이고 $x_i \sim uniform(0, 20)$이며 ϵ_i와 x_i는 서로 독립이다.

첫번째 조사에서의 응답확률은 모수적 모형인

$$\pi(x_i') = [1 + exp(\alpha_0 + \alpha_1 x_i')]^{-1} exp(\alpha_0 + \alpha_1 x_i')$$

을 가정하고 (α_0, α_1)를 추정하기 위해 MLE를 사용하는 데 뉴튼-랩슨 (Newton-Raphson)방법을 사용하여 반복적으로 계산하여 수렴하는 값을 찾는다. 단 여기서 $x_i' \sim uniform(0, 1)$이고 (α_0, α_1)의 값은 모의실험에서 $(0.0, -1.0)$, $(-0.5, 1.0)$, $(1.5, -1.0)$을 사용하며 각각 평균 응답률은 0.38, 0.5, 0.73이다. 재조사에서 응답확률 p_i는 비모수적 방법으로

$$\hat{p}_i = \left(\sum_{i=1}^n w_i \right)^{-1} \sum_{i=1}^n w_i R_i$$

으로 추정한다. 모의 실험을 위하여 재조사에서의 응답확률 p_i는 0.35, 0.5, 0.7을 가정한다. 가중치 $w_i = n^{-1}$를 사용하고 잭나이프 분산추정량을 위해서 $c_k = n^{-1}(n-1)$이고

$$w_i^{(k)} = \begin{cases} (n-1)^1 n w_i, \text{ for } i \neq k \\ 0, \text{ for } i = k \end{cases}$$

을 사용한다.

표본크기 $n = 100$인 표본 $(y_i, x_i, \epsilon_i, R_i, C_i)$을 B번 생성시켜 실험적 값으로 MSE와 편향을 계산한다. Table 3.1에서 괄호안의 값이 편향을 나타낸다. Table 3.1의 각 칸에서의 세가지의 MSE와 괄호안의 편향 값은 세개의 추정량에 대한 것을 나타내는 데 첫번째 값은 Deming (1953)이 제시한 추정량을 사용한 것으로 재조사까지 조사된 표본자료를 가지고 표본평균을 낸 것이며 두번째 값은 Park 등 (2008)이 연구한 추정량으로

$$\bar{y}_P = \tilde{\phi}\bar{y}_f + (1-\tilde{\phi})(\bar{y}_T/\bar{x}_T)\bar{x}_R \tag{3.1}$$

이며 여기서 $\tilde{\phi}$는 $\bar{y}_P$의 분산을 최소화 하는 방향으로 결정된다. 세번째 값은 본 논문에서 제시한 추정량을 사용한 것으로 식 (2.2)의 실험적 MSE값과 편향이다. 여기서 사용되는 응답확률들과 ϕ값은 실제 표본조사에서 사용될 수 있는 것을 고려하여 모두 추정된 확률 $\hat{\pi}_i, \hat{p}_i$과 추정된 $\hat{\phi}$를 사용한다. Table 3.1의 값을 살펴보면 모든 π_i와 p_i에서 제안된 추정량 (2.2)의 편향이 거의 작아짐을 알 수 있고 MSE가 다른 추정량과 비교하여 가장 작음을 알 수 있다. 식 (3.1)의 추정량의 MSE와 편향도 Deming (1953)의 것보다 작아짐을 알 수 있다. 또한 제안된 추정량은 π_i가 작아질 때 MSE와 편향이 작아짐을 알 수 있고 p_i가 작아질 때도 MSE와 편향이 줄어듬을 알 수 있다.

Table 3.1 MSE and bias of estimators under callback

$(\alpha_0,\ \alpha_1)$	p		
	0.35	0.5	0.7
(0.0, -1.0)	9.491 (-0.040)	8.328 (-0.046)	6.779 (0.021)
	7.410 (0.105)	6.688 (0.088)	6.227 (0.025)
	6.811 (0.053)	6.346 (0.032)	6.091 (-0.001)
(-0.5, 1.0)	8.337 (-0.012)	7.539 (-0.010)	6.434 (0.045)
	6.838 (0.089)	6.238 (0.081)	5.984 (0.015)
	6.409 (0.048)	6.001 (0.030)	5.887 (-0.011)
(1.5, -1.0)	6.809 (-0.014)	6.529 (-0.014)	5.939 (0.015)
	6.323 (0.065)	6.032 (0.074)	5.870 (0.021)
	6.103 (0.048)	5.876 (0.038)	5.812 (-0.004)

분산추정의 일치성을 알아보기 위해 상대비와 t-검정통계량을 계산한다. 상대비는 분산추정량의 실험적 평균을 추정량의 실험적 MSE로 나눈 것을 나타내고 t-검정통계량은 kim (2004)에 제시된 것으로써 모의실험적 t-검정통계량으로 분산추정량의 실험적 편향을 실험적 편향의 실험적 표준오차로 나눈 것을 말한다. 그리고 이것의 절대값이 2보다 크면 분산추정량이 일치성을 만족하지 못한다고 판단한다. Table 3.2의 각 칸은 식 (2.2)의 분산추정량에 대해 첫번째 줄은 상대비를 나타내고 두번째 줄은 t-검정통계량을 나타내는 데 상대비를 살펴보면 0.966에서 1.024의 값으로 거의 1에 가깝고 t-검정통계량의 절대값은 0.014에서 0.799의 값을 가지므로 2보다 작음을 알 수 있다. 그러므로 제안된 분산추정량이 일치성의 성질을 가지고 있음을 알 수 있다.

Table 3.2 Relative ratio and t-statistic of variance estimator

$(\alpha_0,\ \alpha_1)$	p		
	0.35	0.5	0.7
(0.0, -1.0)	0.986	0.981	0.966
	-0.339	-0.472	-0.799
(-0.5, 1.0)	1.022	1.024	0.998
	0.506	0.551	-0.057
(1.5, -1.0)	0.999	1.001	0.985
	-0.033	0.014	-0.357

4. 결론

보조변수의 정보를 추출틀과 같은 곳에서 얻을 수 있다는 가정하에 재조사 후 추정을 위하여 그 정보를 사용하여 효율을 증대시킬 수 있는 방안을 연구하였으며 이것과 함께 재조사가 되는 개체가 조사할 때 마다 다르다라는 것에 착안하여 첫번째 조사와 재조사에서의 응답확률들을 이용하는 추정방안을 고려하였다. 보조변수를 사용하는 추정의 형태는 회귀추정의 방안을 적용해 보았다. 재조사하에 제안된 추정량의 근사 불편성과 분산의 크기를 이론적으로 증명해 보였고 모의실험을 통하여 기존의 추정량보다 제안된 추정량의 편향이 작고 효율이 증대한다는 것을 보였다. 또한 제안된 분산추정량도 일치성을 만족하고 있음을 모의실험을 통해 입증하였다. 이와 같은 연구는 재방문과 응답확률과 보조정보 등이 융합하여 적용되고 있으므로 재조사 분야에서 여러 방향으로 확장하여 적용할 수 있다. 첫째, 본 연구에서는 모수적 응답확률 추정 문제를 다루었지만 다른 방향으로 비모수적 추정 방법인 데이터마이닝 기법으로 확장하여 연구할 수 있다. 데이터마이닝 기법으로는 신경망 모형 또는 의사결정 나무 등이 있다. 또한 첫번째 조사에서의 응답확률과 재조사에서의 응답확률을 본 연구에서 가정한 것 외에 여러가지 모형을 가정하여 적용해 볼 수도 있다. 둘째, 보조정보의 사용도 다변량의 보조변수를 사용한 비추정, 회귀추정 방법을 고려할 수도 있다. 셋째, 여기에서는 한번의 재조사만을 고려하였는 데 실제 조사에서는 여러번의 재조사가 이용될 수 있으므로 본 연구에서도 두번이상의 재조사를 고려하는 연구로 확장할 수 있다.

References

Deming, W. E. (1953). On a Probability Mechanism to Attain an Economic Balance Between the Resultant Error of Response and the Bias of Nonresponse. *Journal of the American Statistical Association*, **48**, 743-772.

Elliott, M. R., Little, R. J. and Lewitzky, S. (2000). Subsampling callbacks to improve survey efficiency. *Journal of the American Statistical Association*, **95**, 730-738.

Han, H. and Byun, J. (2014). Callbacks effects on nonresponse bias. *Survey Research*, **15**, 21-45.

Hansen, M. H. and Hurwitz, W. N. (1946). The Problem of Nonresponse in Sample Surveys. *Journal of the American Statistical Association*, **41**, 517-529.

Isaki, C. T. and Fuller, W. A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, **77**, 89-96.

Ismail, M., Shahbaz, M. Q. and Hanif, M. (2000). A general class of estimator of population mean in presence of non-response. *Pakistan Journal of Statistics*, **27**, 467-476.

Kim, S., Lee, S. and Yoon, Y. (2005). Formation of unit nonresponse weighting adjustment cell using multivariate regression trees. *Journal of The Korean Official Statistics*, **10**, 172-190.

Kim, J. K. and Park, H. (2006). Imputation using response probability. *The Canadian Journal of Statistics*, **34**, 171-182.

Okafor, F. C. and Lee, H. (2000). Double sampling for ratio and regression estimation with sub-sampling the non-respondents. *Survey Methodology*, **26**, 183-188.

Park, H., Na, S. and Jeon, J. (2008). Estimation using response probability under callbacks. *Statistics and Probability Letters*, **78**, 1735-1741.

Park, H. and Jeon, J. (2010). Using response probability and ratio imputation in the estimation under callbacks. *Journal of the Korean Statistical Society*, **39**, 511-521.

Park, H. and Park, W. (2013). Usage of auxiliary variable and neural network in doubly robust estimation. *Journal of theKorean Data & Information Science Society*, **24**, 659-667.

Journal of the Korean Data & Information Science Society
2015, **26**(2), 377–385

http://dx.doi.org/10.7465/jkdi.2015.26.2.377
한국데이터정보과학회지

Estimation to improve survey efficiency in callback[†]

Hyeonah Park[1] · Seongryong Na[2]

[1]Department of Statistics, Seoul National University
[2]Department of Information and Statistics, Yonsei University

Received 23 February 2015, revised 16 March 2015, accepted 18 March 2015

Abstract

After performing callback for nonresponses in sample survey, we present an estimator of regression form using an auxiliary variable and a variance estimator using replicate method. Parametric inference method of the response probability is also presented. We research an unbiased estimator of high efficiency for the population mean and a variance estimator with consistency under callback. We also prove the validity of the theory through the simulation.

Keywords: Auxiliary variable, callback, response probability, variance estimation.

[†] 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 (NRF-2012R1A1A3003761).
[1] Corresponding author: Postdoc, Department of Statistics, Seoul 151-742, Korea.
E-mail: hapk@daum.net
[2] Professor, Department of Information and Statistics, Gangwon 220-710, Korea.