The Korean Journal of Applied Statistics (2015) DOI: http://dx.doi.org/10.5351/KJAS.2015.28.3.417
28(3), 417–427

Usage and Estimation of R-indicator for Representative

Hyeonah Park[a,1] · Kee-Jae Lee[b]

[a]Department of Statistics, Seoul National University
[b]The Department of Information Statistics, Korea National Open University

(Received January 20, 2015; Revised February 27, 2015; Accepted March 12, 2015)

Abstract

Measures in response rate used to measure the representativeness of the sample (the more high response rate) better explain the representativeness of the sample. However, we cannot often explain the representativeness of the sample because there is nonresponse even in the high response rate. Therefore, Schouten *et al.* (2009) presented a new R-indicator measure that can be described as a representative of the sample. We research the new estimator of the R-indicator in this paper because there are parameters that require estimations. We describe the meanings as representative of the R-indicator; consequently, the bias and efficiency of the proposed estimator for R-indicator are compared to the existing estimator under various simulations. The representativeness of the sample is also explained by applying the proposed estimators in the actual data.

Keywords: R-indicator, representativeness, response rate, response probability

1. 서론

모집단에 대한 표본의 대표성은 포함확률이 없는 조사단위의 비율을 줄여서 편향이 없는 추정량을 제시하는 것을 말한다 (Kim, 2005). 실제조사에서 표본의 대표성을 높이기 위해서는 추출틀의 포괄범위율(coverage rate)을 높이는 노력과 함께 완벽한 응답이 이루어져야 하는 데 조사과정에서는 무응답이 발생하게 된다.

기존 연구에서는 대개 응답률을 파악함으로써 표본 대표성과 무응답으로 인한 편향을 살펴보고자 하였지만, 응답률을 높이는 것만으로 추정량의 편향을 줄일 수 없다 (Groves, 2006). 이와 같이 응답률이라는 단순한 척도가 표본의 대표성을 완벽하게 설명하는 것이라 할 수 없기 때문에 그것의 대안으로 응답확률을 이용한 산포, 절대편향의 상한값, 제곱근 MSE의 상한값, 부차그룹 응답률의 변동계수 등과 같은 다양한 척도(indicator)들이 연구되어 왔다. 특히 Schouten 등 (2009)는 R-indicator라는 새로운 보완적인 척도를 제시하여 표본의 대표성을 설명할 수 있게 하였다. 이후 Schouten 등 (2011)와 Schouten 등 (2012)와 Shlomo 등 (2012)와 Son 등 (2014) 등을 통해서 조사응답의 대표성 척도에 대한 연구가 활발하게 진행되었다.

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: Department of Statistics, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 151-742, Korea. E-mail: hapk@daum.net

R-indicator의 또 다른 유용성으로는 조사가 끝난 후에 표본의 대표성을 살펴 볼 수 있는 척도이기도 하지만 조사 과정에서 표본의 대표성을 높이기 위해서도 사용될 수도 있다는 것이 여러 논문들에 제시되어 있다 (Schouten 등, 2009, 2011).

본 논문에서는 Schouten 등 (2009)에서 제안된 R-indicator의 대표성에 관한 의미와 설명력을 연구하며 R-indicator의 새로운 추정방법을 제시하고자 한다. R-indicator는 본질적으로 응답확률에 기반을 두고 있으므로 그 응답확률 추정이 고려되어야 하며 또한 표본에 기반을 두어 그 값을 계산해야 하므로 모수에 대한 추정이 고려되어야 한다.

본 연구에서는 응답확률의 추정과 표본설계에서의 추출에 의한 추정으로 나누어서 새로운 추정방법을 연구한다. 여러 가지 응답확률 추정방법과 포함확률에 기반을 둔 새로운 R-indicator의 추정기법을 연구하고 다양한 모의실험 하에 R-indicator의 대표성으로써의 설명력과 제안된 추정량의 편향과 효율을 기존의 추정량과 비교분석한다. 또한 실제자료를 사용하여 제안된 R-indicator 접근도 연구한다.

2. 대표성을 위한 R-indicator

모집단에 대한 표본의 대표성을 표현하기 위한 척도로 응답률과 R-indicator를 정의하기에 앞서 추출여부와 응답 여부, 그리고 응답확률의 개념을 정의한다. 모집단의 크기를 N이라 하면 모집단은 $U = \{1, 2, \ldots, N\}$으로 정의될 수 있으며 거기서 표본설계에 의해 추출되는 표본의 크기를 n이라 할 때 지시변수 s_i를 정의하면 다음과 같다.

$$s_i = \begin{cases} 1, & i\text{번째 개체가 표본으로 추출됨 } (i = 1, \ldots, N), \\ 0, & \text{그외}, \end{cases}$$

여기서 추출된 표본은 $s = \{s_1, s_2, \ldots, s_n\}$로 정의할 수 있고 포함확률(Inclusion probability)은 $\pi_i = P(s_i = 1)$이다. 또한 응답여부를 나타내는 응답변수는

$$r_i = \begin{cases} 1, & i \in \{i : s_i = 1\}\ , \\ 0, & \text{그외} \end{cases}$$

이고, 응답확률(Response probability)은 $\rho_i = P(r_i = 1|s_i = 1)$이다.

이와같이 정의된 응답확률 개념을 바탕으로 표본자료의 대표성(representativity)은 두 가지 관점으로 정의할 수 있다. 첫째, 표본에 관한 strongly representative에 관한 개념이다. 이것은 모집단의 모든 개체에 대해 응답확률들이 모두 같고, 모든 응답들이 독립인 경우를 말하며 그것을 수식으로 표현하면

$$\rho_i = P(r_i = 1|s_i = 1) = \rho, \quad \text{for } i = 1, 2, \ldots, N$$

이다. 이것은 무응답에 대한 체계에서 모든 관심변수들에 대해 MCAR(missing completely at random)을 의미하며, 또한 무응답이 추정량에 대한 편향을 발생시키는 원인이 되지 않는다는 것을 의미한다. 그러나 이것은 각 개체별 응답확률들을 비교해야 하기 때문에 실제적으로 사용하는 데 제약사항이 있으므로 이것보다 약한 개념을 가지는 대표성 이론을 정의할 수 있다. 그래서 L개의 범주를 가지는 보조변수를 고려하며 그 보조변수는 표본에서 모두 응답함을 가정한다. 각 범주별 모집단의 크기는 $N_h\ (h = 1, \ldots, L)$이며 h층에서 i번째 개체에 대한 응답확률을 ρ_{hi}라 하여 다음과 같은 대표성 이론을 정의한다. 둘째, 보조변수에 대하여 표본에 관한 weakly representative에 관한 개념으로 평균 응답확률이 각 범주에서 똑같은 경우를 말한다. 그것을 수식으로 나타내면

$$\bar{\rho}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} \rho_{hi} = \rho, \quad \text{for } h = 1, 2, \ldots, L$$

이다.이것은 응답이 많은 보조변수들에 관하여 weakly representative이 있고 그 보조변수들이 관심변수와 강한 상관관계가 있는 경우에 추정량의 편향이 줄어들 수 있다는 것을 내포하고 있다.

모든 표본에 대해 포함확률이 다 존재한다는 가정하에 지금까지의 표본의 대표성은 전체 표본에 대한 응답률로 알 수 있었다. 그러나 응답률만으로는 표본의 대표성을 보장할 수 없는 경우가 발생하므로 위에서 제시된 이론들을 바탕으로 표본의 대표성을 알아보기 위한 척도로써 단순하게 전체 표본에 대한 응답률 외에 각 개체에 대한 응답확률들을 사용하는 것을 고려한다. 예를 들어 모든 응답확률이 동일하다면 응답은 strongly representative를 나타내고 응답의 구성과 표본사이에 어떤 체계적인 차이점이 없다는 것을 의미한다. 그러므로 만약 응답확률이 동일하지 않다면 다른 응답의 구성이 얼마나 영향을 미치는지를 알기 위해 응답확률들의 거리를 잴 수 있는 척도가 필요하다. 이와 같은 척도를 위해 응답확률들의 모표준편차를 고려하게 되는 데 수식은 다음과 같다.

$$S(\rho) = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\rho_i - \bar{\rho})^2}.$$

단 여기서 $\bar{\rho} = \sum_{i=1}^{N} \rho_i/N$이며 이 모표준편차의 범위는 0에서 0.5사이의 값을 나타낸다. 이것을 수정하여 상관계수가 0과 1 또는 -1에서 다른 개념을 주는 것처럼 다음과 같이 0에서 1사이의 값을 갖는 R-indicator를 정의한다.

$$R(\rho) = 1 - 2S(\rho). \tag{2.1}$$

이 값은 1에 가까이 갈수록 표본에 대해 strongly representative가 있음을 나타내며 반대로 0으로 갈수록 표본의 대표성이 없다는 것을 의미한다.

그러나 실제로는 응답확률을 알 수 있는 것이 아니기 때문에 R-indicator를 계산할 수 없다. 그러므로 각 개체에 대한 응답확률이 추정되어야 함을 알 수 있다. 이와 같이 응답확률에 대한 추정의 문제는 표본의 대표성을 위한 것 뿐만 아니라 가중치의 보정 및 대체법에서도 고려되어지고 있다. 응답확률 $\hat{\rho}_i\,(i = 1, 2, \ldots, n)$을 추정하기 위해서 표본에서 응답이 모두 이루어진 보조변수들이 존재하며 이를 이용하여 모수적 방법인 로지스틱회귀모형, 프로빗 모형 등을 이용한다. 또 다른 추가적인 방법으로 응답률이 비슷한 응답층을 구분하기 위해 위에서 제시된 모수적 방법들을 사용하고 응답층을 나누고 그 나누어진 층별 표본 응답률을 사용하여 응답확률을 추정하기도 한다. 응답확률의 추정을 연구한 논문들로는 Rosenbaum (1987)와 Ekholm과 Laaksonen (1991)와 Robins 등 (1994)와 Iannacchione (2003)와 Kim과 Park (2006) 등이 있다.

표본의 대표성을 위해서는 먼저 각 개체에 대하여 응답확률을 추정하고, 추정된 응답확률과 표본설계에서의 포함확률 π_i을 사용하여 R-indicator를 구한다.

Schouten 등 (2009)에 제시된 평균응답확률 추정량은

$$\hat{\bar{\rho}} = \frac{1}{N}\sum_{i=1}^{n}\frac{\hat{\rho}_i}{\pi_i}$$

이다. 이와 같은 추정의 형태는 H-T 추정량 (Horvitz와 Thompson, 1952)의 형태를 형성하고 있음을 알 수 있다. 그리고 R-indicator의 추정량은

$$\bar{R}(\rho) = 1 - 2\sqrt{\frac{1}{N-1}\sum_{i=1}^{n}\frac{(\hat{\rho}_i - \hat{\bar{\rho}})^2}{\pi_i}} \tag{2.2}$$

이다.

좋은 추정량이란 기본적으로 불편을 만족하고 효율이 좋은 것을 선호하는 데 본 연구에서는 근사적인 불편을 만족하고 기존의 추정량보다 효율의 증대를 가져올 수 있는 추정량을 제시하고자 한다. 본 논문에서는 평균응답확률 추정량과 R-indicator 추정량을 위해 Hájeck 추정량의 형태를 적용하는 것을 연구한다. 모평균을 추정할때 Hájeck 추정량은 모집단의 크기 N을 사용한 추정량보다 종종 효율이 더 좋다는 것이 알려져 있다. 즉, 모평균을 추정할 때 Hájeck 추정량이 N으로 나누는 추정량보다 분산이 작아지는 경향이 있는 것인데 그 이유는 Hájeck 추정량은 비추정량에서 $x_i = 1$를 사용하는 것이기 때문이다 (Kim, 2008). 결론적으로 본 논문에서 사용한 평균응답확률 추정량의 형태는

$$\hat{\rho} = \frac{\sum_{i=1}^{n} {\pi_i}^{-1} \hat{\rho}_i}{\sum_{i=1}^{n} {\pi_i}^{-1}} \tag{2.3}$$

이며 이것은 비추정의 형태를 가지고 있으며 추정된 응답확률이 응답확률과 거의 비슷하다는 가정을 한다면 근사적인 불편성을 만족함을 알 수 있다. 평균응답확률을 추정하기 위해 모집단의 크기 N이 적용되는 곳에 $\sum_{i=1}^{n} {\pi_i}^{-1}$을 사용한 것은 R-indicator추정량에도 적용하여 다음과 같은 형태를 제시한다.

$$\hat{R}(\rho) = 1 - 2\sqrt{\frac{1}{\left(\sum_{i=1}^{n} {\pi_i}^{-1}\right) - 1} \sum_{i=1}^{n} \frac{\left(\hat{\rho}_i - \hat{\rho}\right)^2}{\pi_i}}. \tag{2.4}$$

그리고 위에서 제시된 R-indicator 추정량의 통계적 성질을 다루고자 다음과 같은 조건을 가정한다.

$$\hat{\rho}_i - \rho_i = O_p\left(n^{-\frac{1}{2}}\right) \tag{2.5}$$

과

$$\sum_{i=1}^{n} \pi_i z_i - \sum_{i=1}^{N} z_i = O_p\left(N n^{-\frac{1}{2}}\right) \tag{2.6}$$

이며 여기서 z_i는 ρ_i, x_i 또는 ρ_i의 함수를 뜻한다.

정리 2.1 식 (2.5), (2.6)의 가정하에

$$\begin{aligned} \frac{\sum_{i=1}^{n} {\pi_i}^{-1}(\hat{\rho}_i - \hat{\rho})^2}{\left(\sum_{i=1}^{n} {\pi_i}^{-1}\right) - 1} = \frac{\sum_{i=1}^{N}(\rho_i - \bar{\rho})^2}{N-1} + \frac{1}{N-1}\left[\left(\sum_{i=1}^{n} \frac{(\rho_i - \bar{\rho})^2}{\pi_i} - \sum_{i=1}^{N}(\rho_i - \bar{\rho})^2\right)\right. \\ \left. - \frac{\sum_{i=1}^{N}(\rho_i - \bar{\rho})^2}{N-1}\left(\sum_{i=1}^{n} \pi_i^{-1} - N\right)\right] + O_p\left(n^{-1}\right) \end{aligned} \tag{2.7}$$

을 증명한다.

증명: 먼저 식 (2.5)의 가정하에 $\hat{\rho}$의 테일러 전개를 이용하면

$$\hat{\rho} = \frac{\sum_{i=1}^{N} \rho_i}{N} + \frac{1}{N}\left[\left(\sum_{i=1}^{n} \pi_i^{-1}\rho_i - \sum_{i=1}^{N} \rho_i\right) - \frac{\sum_{i=1}^{N} \rho_i}{N}\left(\sum_{i=1}^{n} \pi_i^{-1} - N\right)\right] + O_p\left(n^{-\frac{1}{2}}\right)$$

이 되고 이와 더불어서 식 (2.6)의 가정하고 $[\sum_{i=1}^{n} {\pi_i}^{-1} - 1]^{-1} \sum_{i=1}^{n} {\pi_i}^{-1}(\hat{\rho}_i - \hat{\rho})^2$의 테일러 전개를 사용하면 식 (2.7)를 유도할 수 있다. □

식 (2.7)에 기대값을 취하면

$$E\left[\frac{\sum_{i=1}^{n}\pi_i^{-1}(\hat{\rho}_i-\hat{\rho})^2}{(\sum_{i=1}^{n}\pi_i^{-1})-1}\right]=\frac{\sum_{i=1}^{N}(\rho_i-\bar{\rho})^2}{N-1}+O\left(n^{-1}\right)$$

이 유도되며 만약 식 (2.7)의 $(N-1)^{-1}\sum_{i=1}^{N}(\rho_i-\bar{\rho})^2$의 뒷부분이 무시될 정도로 작다면 $E(\hat{R}(\rho))\doteq R(\rho)$로 근사할 수 있다.

Remark 2.1: R-indicator의 값을 구하기 위해서는 첫째 응답확률이 추정되어야 하며 둘째 표본자료에 의한 포함확률을 사용하는 추정이 이루어져야 한다. 포함확률을 사용하는 것은 모집단의 분포와 상관없이 결정되기 때문에 분포의 가정에 영향을 받지 않지만 응답확률의 추정에서 모수적 방법인 로지스틱 회귀모형을 사용하는 것은 분포의 가정에 영향을 받음을 알 수 있다. 즉 응답확률모형이 틀리다면 그와 같은 모형에 의해 추정된 응답확률은 잘 맞지 않을 것이다. 또한 응답확률 추정에 비모수적인 방법을 이용하면 분포에 영향을 받지는 않지만 효율면에서 모수적 방법보다 감소된다는 것이 알려져 있다. 그러나 기존의 연구로의 비모수적 방법은 가중치를 사용한 것이 아니기 때문에 더 연구가 진행되어야 한다. 결론적으로 R-indicator의 계산을 위해서 응답확률의 추정을 조금 더 정확하게 할 필요가 있다. 예를 들어 응답확률을 추정하기 위한 보조변수의 선택도 신중해야 하며 가중치 조정과정에서의 무응답 조정층의 구성과 R-indicator 계산에서의 조정층이 상이함으로 발생하는 여러 문제점도 신중하게 접근해야 한다.

Remark 2.2: 식 (2.3)은 추정된 응답확률들의 평균으로써 응답률의 값으로 사용될 수 있다. 그러므로 식 (2.3)을 대신할 수 있는 추정량으로

$$\tilde{\rho}=\sum_{h=1}^{L}\frac{N_h}{N}\frac{\sum_{i=1}^{n_h}\pi_{hi}{}^{-1}\hat{\rho}_{hi}}{\sum_{i=1}^{n_h}\pi_{hi}{}^{-1}},\qquad \tilde{\rho}'=\frac{\sum_{i=1}^{n}\pi_i{}^{-1}r_i}{\sum_{i=1}^{n}\pi_i{}^{-1}},\qquad \tilde{\rho}''=\sum_{h=1}^{L}\frac{N_h}{N}\frac{\sum_{i=1}^{n_h}\pi_{hi}{}^{-1}r_{hi}}{\sum_{i=1}^{n_h}\pi_{hi}{}^{-1}}$$

들을 고려할 수 있으며 식 (2.4)에 $\hat{\rho}$ 대신에 삽입하여 R-indicator의 추정량으로 사용한다. 또한 표본의 대표성을 알 수 있는 다른 척도로는 K개의 부차 그룹으로부터 계산된 응답률에 의해 계산되는 변동계수로 부차그룹별 응답률의 수준 차이가 낮을 수록 표본의 대표성이 성립됨을 알 수 있는 것이다. 그리고 응답률의 역수를 조사단위의 가중치로 삼아 그 가중치들에 대한 분산을 구하는 것인데 가중치들의 변동이 클 수록 무응답 편향의 위험도가 증가된다고 판단한다. 마지막으로 무응답의 가중치와 조사변수간의 상관계수를 구하는 것이 있다. 그 외에도 사후층화 가중치의 분산과 추정량에 대한 절대편향 또는 제곱근 MSE의 상한선과 부분적 R-indicator 등이 있다 (Schouten 등, 2009, 2011; Son 등, 2014).

3. 모의실험과 실제 조사 적용 사례

3.1. 가상모집단에서 표본자료의 대표성

먼저 가상인 유한모집단을 구현하기 위해 오차의 분포는 $\epsilon_i\sim N(0,1)$을 사용하고 보조변수의 분포는 $x_i\sim \text{uni}(0,20)$을 사용한다. 그리고 관심변수의 분포는 회귀모형

$$y_i=4x_i+\epsilon_i$$

을 가정하여 10,000개의 자료를 생성한다. 즉 $N=10{,}000$개의 자료를 유한모집단으로 가정한다. 또한 각 모집단의 자료별로 응답확률을 안다는 가정하에 (α_0,α_1)의 회귀계수를 가지는 로지스틱회귀모형을

Table 3.1. Population R-indicator(0.4)

(α_0, α_1)	mr	R-indi	bias	rbias
$(1, -0.1)$	0.4774	0.7258	−6.3658	−0.1449
$(5, -0.5)$	0.4547	0.2291	−19.1526	−0.4367
$(-1.2, 0.1)$	0.4759	0.7255	6.6229	0.1525
$(10, -1.0)$	0.4452	0.1134	−21.5020	−0.4863
$(20, -2.0)$	0.4525	0.0534	−21.9501	−0.4987
$(50, -5.0)$	0.4518	0.0243	−21.8797	−0.4983

Table 3.2. Population R-indicator(0.5)

(α_0, α_1)	mr	R-indi	bias	rbias
$(1.3, -0.1)$	0.5463	0.726540	−5.8705	−0.13270
$(1.8, -0.15)$	0.5278	0.612440	−8.2442	−0.18781
$(5, -0.4)$	0.5714	0.294610	−13.6600	−0.30968
$(-2.0, 0.2)$	0.5361	0.513030	10.4168	0.23727
$(14, -1.2)$	0.5338	0.089924	−18.2269	−0.41340
$(20, -1.8)$	0.5085	0.056934	−19.5129	−0.44495

Table 3.3. Population R-indicator(0.6)

(α_0, α_1)	mr	R-indi	bias	rbias
$(1.8, -0.1)$	0.6520	0.7521	−4.3436	−0.0984
$(5, -0.35)$	0.6475	0.3757	−10.7212	−0.2420
$(-1.2, 0.2)$	0.6861	0.5781	7.0606	0.1605
$(14, -1.0)$	0.6512	0.1571	−13.3841	−0.3040
$(24, -1.8)$	0.6227	0.0910	−14.9549	−0.3433
$(58, -4.0)$	0.6839	0.0989	−12.5913	−0.2888

통하여 그 확률을 생성한다.

$$\rho_i = \left[1 + \exp(\alpha_0 + \alpha_1 x_i)^{-1}\right]^{-1}.$$

그리고 x_i변수가 7, 14의 값을 가지는 곳을 층의 경계값으로 삼아 3개의 층을 가지는 층화변수 h_i를 생성하고 pps추출에 사용되는 크기척도로 $s_i \sim \text{uni}(0, 5)$를 생성하였다. 즉 모집단의 자료로 $(y_i, x_i, \rho_i, h_i, s_i)$ for $i = 1, 2, \ldots, 10000$가 구현된다. Table 3.1∼3.5에서 응답률에 따른 R-indicator의 값과 편차과 상대편차을 계산한다. 각 Table 별로 $\mu = \sum_{i=1}^{N} y_i/N$일때 편차와 상대편차인

$$\text{bias} = \frac{\sum_{i=1}^{N} r_i y_i}{\sum_{i=1}^{N} r_i} - \mu, \qquad \text{rbias} = \frac{\text{bias}}{\mu}$$

과 응답률인 mr와 식 (2.1)의 R-indicator값 R-indi을 구현한다. 각 Table은 응답률이 각각 0.4, 0.5, 0.6, 0.8, 0.9 정도에서 R-indicator의 값을 계산하여 그 값이 작은 곳과 큰 곳에서 편차와 상대편차의 값을 살펴본다. Table 3.1을 살펴보면 R-indi의 값이 0.7258인 곳에서는 rbias의 값이 −0.1449를 나타내고 이보다 더 작은 R-indi의 값이 0.2291인 곳에서는 rbias의 값이 −0.4367를 나타냄으로서 상대편차가 커짐을 알 수 있다. 즉 Table 3.1은 응답률이 0.4정도인 것으로써 R-indicator가 클수록 편차와 상대편차가 작아짐을 알 수 있으며 R-indicator가 작을 수록 편차와 상대편차가 커짐을 알 수 있다. 이와 같은 특징들은 Table 3.2∼3.5에서 나타남을 알 수 있다. 그러므로 표본자료의 대표성을 살펴보기 위해서 먼저 응답률을 계산하고 그리고 R-indicator의 값을 계산하여서 그 값들이 큰 여부를 알아보면 된다.

Table 3.4. Population R-indicator(0.8)

(α_0, α_1)	mr	R-indi	bias	rbias
$(3, -0.1)$	0.8578	0.8636	−1.5112	−0.0341
$(5, -0.25)$	0.8423	0.6745	−4.0066	−0.0913
$(-1.2, 0.5)$	0.8906	0.6492	3.6345	0.0823
$(14, -0.8)$	0.8211	0.4044	−6.6553	−0.1512
$(20, -1.1)$	0.8560	0.4347	−5.5007	−0.1250
$(58, -3.1)$	0.8867	0.4140	−4.4849	−0.1018

Table 3.5. Population R-indicator(0.9)

(α_0, α_1)	mr	$R - indi$	$bias$	$rbias$
$(5, -0.1)$	0.9740	0.9746	−0.3095	−0.0070
$(9, -0.25)$	0.9951	0.9885	−0.1218	−0.0028
$(-1.2, 0.8)$	0.9452	0.7455	1.9600	0.0446
$(14, -0.7)$	0.9222	0.6860	−2.7423	−0.0621
$(20, -1.0)$	0.9341	0.6808	−2.4175	−0.0548
$(55, -2.5)$	0.9993	0.9848	−0.0273	−0.0006

Table 3.6. Estimation of R-indicator using response probability

(α_0, α_1)	r1	sr1	r2	sr2	r1/r2	sr1/sr2
$(-1.2, 0.1)$	0.0198	0.0960	0.0145	0.0943	1.3623	1.0180
	(0.0036)	(0.0058)	(0.0034)	(0.0058)		
$(5, -0.5)$	0.0499	0.0698	0.0310	0.0581	1.6091	1.2003
	(0.0086)	(0.0144)	(0.0070)	(0.0127)		
$(1.8, -0.1)$	0.0193	0.0917	0.0135	0.0909	1.4284	1.0087
	(0.0028)	(0.0045)	(0.0021)	(0.0038)		
$(5, -0.35)$	0.0489	0.0832	0.0327	0.0743	1.4961	1.1195
	(0.0089)	(0.0045)	(0.0049)	(0.0006)		
$(3, -0.1)$	0.0109	0.0724	0.0089	0.0709	1.2238	1.0210
	(0.0019)	(0.0023)	(0.0013)	(0.0021)		
$(-1.2, 0.2)$	0.0341	0.0917	0.0236	0.0872	1.4445	1.0514
	(0.0045)	(0.0013)	(0.0027)	(−0.0002)		
$(-1.2, 0.8)$	0.0527	0.1037	0.0453	0.0994	1.1636	1.0430
	(0.0090)	(−0.0033)	(0.0085)	(−0.0038)		
$(14, -0.7)$	0.0488	0.0956	0.0469	0.0946	1.0420	1.0097
	(0.0076)	(0.0060)	(0.0065)	(0.0050)		

그러나 R-indicator를 알기 위해서는 응답확률뿐만 아니라 표본설계에 관련된 추정이 이루어져야 한다. Table 3.6에서 여러개의 응답확률별 표본의 크기 100개를 층화pps추출을 사용하고 배정방법으로는 네이만배정을 사용한다. 그리고 이와 같은 추출을 1000번의 모의실험을 통하여 R-indicator의 추정량의 편향 및 MSE를 계산한다.

Table 3.6에 비교를 위해 사용된 추정량은 4가지로써

r1: 식 (2.2)에서 $\hat{\rho}_i$대신 ρ_i를 사용한 R-indicator의 MSE와 편향.

sr1: 식 (2.2)를 사용한 R-indicator의 MSE와 편향.

r2: 식 (2.4)에서 $\hat{\rho}_i$대신 ρ_i를 사용한 R-indicator의 MSE와 편향.

Table 3.7. Results of logistic model for second, third and fourth sample surveys

Effect	DF	2009년 조사 응답 여부		2010년 조사 응답 여부		2011년 조사 응답 여부	
		Wald	P-value	Wald	P-value	Wald	P-value
지역	14	16.3583	0.2920	27.5849	0.0161	31.3341	0.005
성별	1	7.0241	0.0080	13.6601	0.0002	17.5229	<.0001
장애유형	14	20.2501	0.1225	31.9744	0.0040	36.8971	0.0008
장애등급	5	4.6085	0.4655	5.9529	0.3108	7.2179	0.2049
경제활동상태	2	5.5097	0.0636	5.4840	0.0644	4.0406	0.1326
연령대	3	7.9881	0.0463	8.0717	0.0446	9.7908	0.0204

sr2: 식 (2.4)를 사용한 R-indicator의 MSE와 편향.

이다.

Table 3.6의 가운데 각각의 값은 MSE와 괄호안에 편향을 나타내며 r1/r2와 sr1/sr2는 논문에서 제시된 추정량의 MSE에 대한 기존의 추정량의 MSE의 비를 나타낸다. 아래 Table에 나타난 값을 살펴보면 r1/r2값과 sr1/sr2값들이 모두 1보다 큼을 알 수 있고 편향도 r1과 sr1보다 r2와 sr2가 작음을 알 수 있다. 결론적으로 본 논문에서 제시되는 R-indicator의 추정량의 효율이 더 좋음을 알 수 있다.

3.2. 장애인고용패널조사에서 표본자료의 대표성

장애인고용패널조사는 급변하는 장애인의 경제활동상태와 관련된 동태적 기초통계를 생산하고, 예를 들어 장애인의 노동시장 참여에 대한 기초자료로 취업자수,실업자수, 비경제활동 인구수 등을 추계하고, 경제활동상태에 영향을 주는 개인적, 환경적 요인을 규명하여 장애인 고용정책 수립 및 평가에 유용한 자료를 제공하는 것을 목적으로 한다. 1차년도인 2008년도 조사에서는 장애인복지법에서 규정하고 있는 15개 유형의 장애를 지니고 있는 등록장애인 5,092명을 대상으로 진행되었고 연도별 패널조사가 계속해서 진행되고 있다.

본 연구에서는 연도별 종단면 가중치를 사용하여 R-indicator 값을 계산함으로써 표본의 대표성을 살펴본다. 우선 각 개체별 응답확률 추정값은 로지스틱회귀모형을 적용하고자 응답확률에 영향을 미치는 보조변수들을 선별하는 작업을 실시한다. Table 3.7은 2–4차 조사에 대한 로지스틱회귀모형 적합 결과로써 응답여부에 거의 공통적으로 영향을 미치는 변수는 지역, 성별, 장애유형, 연령대 변수임을 알 수 있다. 여기서 연령대는 15–34세, 35–49세, 50–64세, 65세 이상으로 나눈다. 각 차수에서 각 개체의 응답확률의 추정은 가중치 조정과정에서 무응답 조정층을 위한 변수를 고려하여 지역, 성별, 연령대의 보조변수를 이용한 로지스틱모형에 의해 계산된다.

Table 3.8은 각 연도별 조사에 대한 R-indicator의 추정값과 다른 척도들의 값을 비교한 것이다. 본 연구에서는 모집단의 정확한 크기를 추정해야 하였으므로 R-indicator 계산에서 연도별 모집단의 크기는 가중치의 합을 사용한다. 즉 본 연구에서 제시된 R-indicator 추정량을 사용한다. Table 3.8의 결과를 살펴보면 시점이 증가할수록 응답률은 낮아지고 R-indicator의 값 R-indi도 낮아진다. 그러나 각국의 가구패널조사의 표본유지율의 예를 살펴보면 5차년도를 기준으로 미국 PSID는 81%, 독일 GSEP는 81%, BHPS는 75% 등을 나타내며 한국의 가구패널조사의 경우 대우패널의 5년 유지율은 60%, 노동패널의 경우 76%를 나타낸다 (Kang과 Bang, 2011). 본조사는 4차년도를 나타내지만 위의 예와 비교해보면 그리 낮은 응답률이 아님을 알 수 있으며 R-indicator의 값도 가장 작은 값이 0.91907정도이므로 0보다는 1에 가까운 값을 나타냄을 알 수 있다. 즉 종단적 성격을 가지는 표본이 모집단에 대한 대표성을 가진다는 것을 알 수 있다.

Table 3.8. Comparison of R-indicator and other representative scales

조사연도	조사대상자	응답	무응답	응답률	R-indi	CV	Var	Corr
2008년	5,092	5,092						
2009년	5,092	4,677	415	0.9185	0.95585	0.14279	0.00533	0.01424
2010년	4,677	4,450	227	0.8739	0.93236	0.15126	0.00798	0.02416
2011년	4,450	4,233	217	0.8313	0.91905	0.17765	0.02605	0.02316

하지만 R-indicator로 표본의 대표성을 다 알 수 있기보다는 응답률과 함께 대표성을 나타내는 데 보완적인 성격을 가지게 된 것이라 할 수 있다. 그래서 다른 여러 척도들을 같이 비교함으로써 표본의 대표성 확보가 더 견고해 지도록 한다. 대표성을 위한 다른 여러 척도는 Remark 2.2에 설명된 것으로 Son 등 (2014)에 제시된 것이다. 부차그룹의 응답률의 변동계수를 나타내는 CV는 2009년도에 0.143에서 2011년도에 0.178로 증가하며 응답률의 역수를 무응답 가중치로 삼아 분산을 구한것이 Var이며 2009년도에 0.0053에서 2011년도에 0.0261로 증가하는 것을 볼 수 있다. 또한 가중치와 관심변수간의 상관계수인 Corr을 살펴보면 2009년도에 0.0142에서 2011년도 0.02316으로 증가하는 것을 알 수 있다. 시점이 증가할 수록 응답률이 낮아지기 때문에 전반적으로 여러 척도들의 값이 증가함을 알 수 있지만 증가량이 그리 크지 않고 Son 등 (2014)의 복지패널자료에 대한 척도들의 값보다 작음을 알 수 있다. 결과적으로 응답률과 R-indicator 외 여러 척도들을 비교해 본 결과 장애인 고용 패널자료는 모집단에 대한 대표성을 가지고 있다고 할 수 있다.

본 연구에서 제시되는 다양한 척도들이 모든 표본의 대표성을 다 알 수 있다고 할 수는 없기 때문에 향후 연구방향으로 대표성을 나타낼 수 있는 다양한 지표로의 연구가 진행되어야 한다. 그리고 비모수적 방법인 CHAID 분류나무모형 또는 커널(kernel)모형 등에 가중치를 사용한 응답확률 추정방법에 대한 연구와 R-indicator추정을 위하여 응답확률의 추정문제를 비모수적 방법과 모수적 방법의 비교에 대한 연구가 향후 진행될 수 있다. 또한 응답확률 추정을 위한 보조변수의 선택방법도 향후 연구 주제로 고려할 수 있으며 재조사의 결정방법에 적용하여 최적의 횟수를 정하는 데 R-indicator의 적용을 논의할 수도 있을 것이다.

4. 결론

표본을 추출할 때 모집단을 얼마나 대표할 수 있는 가의 문제는 중요한 부분이다. 그러나 엄밀한 확률추출임에도 불구하고 조사과정에서 무응답이 발생한다면 모집단에 대한 표본의 대표성은 저하되고 만다. 그러므로 조사과정에서 응답률을 높이기 위한 작업이 시행되고 있으며 높은 응답률로 모집단에 대한 표본의 대표성을 말해오고 있었다. 그러나 높은 응답률만으로는 표본의 대표성을 설명할 수 없다는 것을 본 연구를 통해 제시하고자 하였으며 그것에 관한 척도로 Schouten 등 (2009)가 제시한 R-indicator를 사용하였다. 모의실험을 통하여 높은 응답률에도 표본의 대표성이 떨어질수 있다는 것을 보였으며 그때 R-indicator의 값도 그것을 설명할 수 있다는 것을 보였다.

일반적으로 척도라는 것은 표본으로 그 값을 도출해야 하기 때문에 추정이라는 과정을 거쳐야 한다. 그래서 본 연구에서는 R-indicator에 대한 기존의 추정방법보다 편향이 줄어들고 효율이 좋아지는 방법을 제시하였으며 모의실험을 통하여 그와 같은 성질들을 규명하였다.

여기에서의 추정은 표본설계에 의한 추정과 응답확률의 추정으로 두가지 방향으로 전개 되었는데 응답확률의 추정문제는 표본조사 분야에서는 무응답 가중치 보정 및 대체에서 많이 연구되어져 왔으며 본 연구에서는 표본의 대표성 척도분야에 그 확률의 추정문제를 확장하여 적용하고자 하였다.

또한 응답율과 R-indicator외에 표본의 대표성을 알기위한 척도의 비교를 위해 Son 등 (2014)에 제안된 부차그룹 응답률의 변동계수와 무응답 가중치의 분산 등을 장애인고용패널자료에 적용하여 유용성을 살펴보았다.

종합해 보면 본 연구는 표본의 대표성을 구현하기 위한 보조 수단으로의 척도의 제안과 추정을 연구하고 응답확률의 추정방안을 제안하였다 할 수 있으며 연구의 발전된 방향으로는 다른 여러 대표성을 위한 척도들의 개발과 모형의 가정에 민감하지 않은 비모수적 방법들로 응답확률을 추정하는 것과 각각의 표본조사의 특색별로 재조사의 횟수를 응답률과 R-indicator와 같은 척도들로 결정하는 것을 들 수 있을 것이다.

References

Ekholm, A. and Laaksonen, S. (1991). Weighting via response modeling in the Finnish household budget survey, *Journal of Official Statistics*, **7**, 325–337.

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys, *Public Opinion Quarterly*, **70**, 646–675.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association*, **47**, 663–685.

Iannacchione, V. G. (2003). Sequential weight adjustment for location and cooperation propensity for the 1995 national survey of family growth, *Journal of Official Statistics*, **19**, 31–43.

Kang, S. and Bang, T. K. (2011). Constructing panel data using repeated cross-sectional survey data: A case of farm household survey and its analysis, *Survey Research*, **12**, 89–112.

Kim, J. K. (2008). *The Sampling Survey*, Free Academy, Gyeonggido.

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

Kim, K. S. (2005). Representative of sample and efficiency of estimation, *Survey Research*, **6**, 39–62.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed, *Journal of the American Statistical Association*, **89**, 846–866.

Rosenbaum, P. R. (1987). Model-based direct adjustment, *Journal of the American Statistical Association*, **82**, 387–394.

Schouten, B., Bethlehem, J., Beullens, K., Kleven, Ø., Loosveldt, G., Luiten, A., Rutar, K., Natalie, S. and Skinner, C. (2012). Evaluating, comparing, monitoring and improving representativeness of survey response through R-indicators and partial R-indicators, *International Statistical Review*, **80**, 382–399.

Schouten, B., Cobben, F. and Bethlehem, J. (2009). Indicators for the representativeness of survey response, *Survey Methodology*, **35**, 101–113.

Schouten, B., Shlomo, N. and Skinner, C. (2011). Indicators for monitoring and improving representativeness of response, *Journal of Official Statistics*, **27**, 1–24.

Shlomo, N., Skinner, C. J. and Schouten, B. (2012). Estimation of an indicator of the representativeness of survey response, *Journal of Statistical Planning and Inference*, **142**, 201–211.

Son, C. K., Kim, H., Gang, H. and Oh, H. (2014). An evaluation analysis of nonresponse effect for the panel survey, *Proceeding of the Korean Association for Survey Research in Fall, 2014*, 39–53.

대표성을 위한 R-indicator의 사용과 추정법 연구

박현아[a,1] · 이기재[b]

[a]서울대학교 통계학과, [b]한국방송통신대학교 정보통계학과

(2015년 1월 20일 접수, 2015년 2월 27일 수정, 2015년 3월 12일 채택)

요 약

표본의 대표성을 측정하기 위한 척도로 응답률이 사용된다.즉 높은 응답률일수록 표본의 대표성을 더 잘 나타낸다고 할 수 있다. 그러나 높은 응답률이라 할지라도 무응답이 존재하는 것이므로 표본의 대표성을 설명하기에는 한계가 있는 경우가 발생한다. 그래서 Schouten 등 (2009)에서는 R-indicator라는 새로운 척도를 제시하여 표본의 대표성을 더 설명할 수 있게 하였다. 본 논문에서는 R-indicator도 표본에 의해 추정되어야 한다는 것에 착안하여 그것에 관한 새로운 추정량을 제시한다. 또한 여러 모의실험하에 R-indicator의 대표성으로써의 설명력과 제안된 추정량의 편향과 효율을 기존의 추정량과 비교분석하며 실제자료에도 제안한 추정량을 적용하여 표본의 대표성을 설명한다.

주요용어: R-indicator, 대표성, 응답률, 응답확률

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

[1]교신저자: (151-742) 서울시 관악구 관악로 1, 서울대학교 통계학과. E-mail: hapk@daum.net