The Korean Journal of Applied Statistics (2017) DOI: https://doi.org/10.5351/KJAS.2017.30.1.069
30(1), 69–82

A study on sensitivity of representativeness indicator in survey sampling

Yujin Lee[a] · Key-Il Shin[a,1]

[a]Department of Statistics, Hankuk University of Foreign Studies

(Received September 26, 2016; Revised November 14, 2016; Accepted December 12, 2016)

Abstract

R-indicator (representativeness indicator) is used to check the representativeness of samples when non-responses occur. The representativeness is related with the accuracy of parameter estimator and the accuracy is related with bias of the estimator. Hence, unbiased estimator generates high accuracy. Therefore, high value of *R*-indicator guarantees the accuracy of parameter estimation with a small bias. *R*-indicator is calculated through propensity scores obtained by logit or probit modeling. In this paper we investigate the degree of relation between *R*-indicator and different non-response rates in strata using simulation studies. We also analyze a modified Korea Economic Census data for real data analysis.

Keywords: bias, sample representativeness, propensity score, logit model

1. 서론

표본조사는 모집단 전체를 조사하지 않고 일부 표본을 추출하여 조사하기 때문에 필연적으로 오차가 발생하며 이를 표본오차(표집오차)라고 한다. 표본오차는 표본 추출틀이 갖고 있는 다양한 정보와 이에 따른 모수 추정 방법에 따라 달라지지만 실사 전에 이미 그 크기를 어느 정도 파악할 수 있다. 반면 조사 진행 중에 발생하는 비표본오차는 다양한 요인에 의해 발생되며 현실적으로 그 크기를 파악하는 것은 쉽지 않다. 이러한 비표본오차의 매우 큰 부분을 차지하는 것이 무응답에 의해 발생된 오차이다. 무응답은 표본수를 감소시켜 추정의 정밀성(precision)을 떨어뜨리며 무응답이 랜덤으로 발생하지 않을 경우에는 정확성(accuracy)도 떨어뜨리게 된다. 따라서 무응답으로 인해 발생한 오차를 줄여 추정의 정확성과 정밀성을 향상시키기 위한 다양한 방법이 제안되었으며 잘 알려진 것처럼 가중치 보정법과 무응답 대체법이 사용되고 있다.

응답이 없거나 적은 조사는 추정의 정밀도가 표본설계 당시에 주어진 기준을 만족하기 때문에 조사 품질을 무응답 비율 또는 응답률로 평가하고 있으며 이를 위해 많은 조사 보고서에 무응답률 또는 응답률을 수록하고 있다. 그러나 응답률이 조사 결과의 정확성을 결정하는 지표로 사용되는 것은 문제가 될 수 있다. 이와 관련된 내용은 Bethlehem 등 (2008)을 참조하기 바란다.

모수 추정의 우수성은 평균제곱오차(mean square error; MSE)를 기준으로 판단하며 평균제곱오차는

This research was supported by Hankuk University of Foreign Studies research fund (2016).
[1]Corresponding author: Department of Statistics, Hankuk University of Foreign Studies, 81, Oedae-ro, Mohyeon-myeon, Cheoin-gu, Yongin-si, Gyeonggi-do 17035, Korea. E-mail: keyshin@hufs.ac.kr

분산(variance)과 편향(bias)으로 구성되어 있다. 분산이 작으면 정밀도(precision)가 높다고 하고, 편향이 작으면 정확도(accuracy)가 높다고 한다. 정밀도의 경우 자료의 수가 늘어나면 일반적으로 높아진다. 따라서 무응답 비율이 작아지게 되면 분석에 사용된 자료의 수가 줄어들지 않기 때문에 정밀도도 낮아지지 않게 된다. 따라서 무응답 비율 또는 응답률은 정밀도의 지표로 사용될 수 있다.

그러나 정확도의 경우에는 자료의 수와 크게 관계가 없다. 정확도는 편향과 관계가 있고 표본이 모집단을 잘 대표하게 되면 편향은 없어진다. 따라서 정확도는 표본의 대표성을 이용하여 측정하게 된다. 최근 표본의 대표성(representativeness)을 진단하는 지수가 개발되었으며 이를 R-지수(representativeness indicator; R-indicator)라 부른다. Kruskal와 Mosteller (1979)는 대표성 개념에 관하여 연구하였다. 여러 논문에서 대표성의 영어 표현인 representativity와 representativeness가 혼용되어 사용되고 있으나 representativeness가 더 많이 사용되고 있다. 최근 Schouten 등 (2009, 2011)이 응답 대표성(response representativeness)의 정의를 구체적으로 제시하였으며 이후 여러 연구에서 이 개념이 사용되고 있다. 이들 논문에서는 R-지수의 기본 개념과 R-지수의 추정 방법이 연구되었다. R-지수의 신뢰구간을 추정하기 위해 사용된 붓스트랩 방법 대신에 선형 근사 분산 추정량이 Schouten 등 (2012)에서 제안되었다. 또한 Ouwehand와 Schouten (2014)과 Schouten 등 (2013)은 R-지수를 이용한 자료분석을 실시하였다.

본 연구에서는 R-지수가 편향을 얼마나 민감하게 표현하는지 연구하였다. 즉 편향과 R-지수와의 관계를 모의실험을 통하여 살펴보았다. 이를 위해 2절에서는 R-지수에 관한 기초 개념을 설명하였으며 3절에서는 분석에서 사용되는 통계 분석 방법을 설명하였다. 4절에서는 층별로 상이한 비율의 무응답이 발생하였을 때, 얻어진 R-지수와 그에 따른 편향, root mean squared error(RMSE) 등 비교통계량과의 관계를 살펴보았으며 실제 자료 분석도 실시하였다. 5절에 결론이 있다.

2. R-지수(representativeness indicator)

2.1. 모 R-지수(population R-indicator)

대표성 정의는 두 가지로 나누어진다. 먼저 강한 대표성(strong representativeness)의 정의는 다음과 같다.

$$\rho_k = \Pr(r_k = 1 | s_k = 1) = \rho, \qquad k = 1, \ldots, N.$$

즉, 자료 k가 표본으로 추출되었을 때 각 표본의 응답 확률이 모두 같은 경우를 의미한다. 이 경우는 missing completely at random(MCAR)에 해당되는 경우로 결측이 추정에 편향을 발생시키지 않는다. 여기서 s_k는 표본이면 '1'이고 표본이 아니면 '0'이며 r_k는 응답이면 '1'이고 무응답이면 '0'인 지시변수이다. 반면 약한 대표성(weak representativeness)은 모집단이 층으로 나누어졌다는 조건에서 정의된다. 즉 정의는 다음과 같다.

$$\bar{\rho}_h = \frac{1}{N_h} \sum_{k=1}^{N_h} \Pr(r_{hk} = 1 | s_{hk} = 1) = \rho, \qquad k = 1, \ldots, N_h,\ h = 1, \ldots, L.$$

이는 각 층별로 평균 응답확률이 같은 것을 의미하며 층별 응답 확률의 평균을 예상할 수 있기 때문에 쉽게 R-지수의 크기를 예상할 수 있다. 여기서 s_{hk}와 r_{hk}는 h층에서 계산되는 지시 변수이다.

이제 각각의 $\rho_k, k = 1, \ldots, N$가 모든 원소에서 알려져 있다고 가정하자. 그러면 응답확률의 표준편차는 다음과 같이 구해진다.

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

이 값을 이용한 R-지수는 다음과 같이 정의된다.

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

여기서 $0 \le S(\rho) \le 0.5$이므로 $0 \le R(\rho) \le 1$이 된다.

2.2. 표본 R-지수(sample R-indicator)

실제 자료분석에서 모든 $\rho_k, k = 1, \ldots, N$가 알려진 경우는 없다. 따라서 자료로부터 ρ_k는 추정되어야 한다. 흔히 사용하는 방법이 일반화선형모형(generalized linear model)인 로짓(logit)모형과 프로빗(probit)모형이다 (Agresti, 2002). Bethlehem 등 (2008)과 Schouten 등 (2009)에서 사용한 기호를 사용하고 추정된 응답확률(response probability, response propensity)을 $\hat{\rho}_k$이라 하였을 때 응답확률의 평균 추정치는 다음과 같이 얻어진다.

$$\hat{\bar{\rho}} = \frac{1}{N}\sum_{k=1}^{N} \hat{\rho}_k \frac{s_k}{\pi_k} = \frac{1}{N}\sum_{k=1}^{n} \frac{\hat{\rho}_k}{\pi_k} = \frac{1}{N}\sum_{k=1}^{n} w_k \hat{\rho}_k, \tag{2.2}$$

여기서도 s_k는 표본인 경우는 '1'이고 표본이 아니면 '0'인 지시변수이다. π_k는 추출확률이고 $w_k = 1/\pi_k$로 설계가중치이다. 따라서 평균 추정치는 설계가중치를 이용하여 구해진 가중평균 값으로 얻어진다. 다음으로 R-지수의 추정값 $\hat{R}(\rho)$는 다음 식에 의해 구해진다.

$$\hat{R}(\rho) = 1 - 2\sqrt{\frac{1}{N-1}\sum_{k=1}^{n} w_k \left(\hat{\rho}_k - \hat{\bar{\rho}}\right)^2}. \tag{2.3}$$

2.3. R-지수를 이용한 최대 무응답 편향

무응답으로 인해 발생한 편향은 추정의 정확성에 매우 큰 영향을 준다. 따라서 무응답으로 인해 발생한 편향을 추정하는 것은 매우 중요하며 Horvitz-Thompson 추정량으로 얻어진 평균 추정량의 경우 무응답으로 인한 편향의 절대값은 다음과 같은 관계가 있는 것으로 알려져 있다 (Schouten 등, 2009).

$$\left|B\left(\hat{\bar{Y}}_{\mathrm{HT}}\right)\right| \le \frac{(1 - R(\rho))S(y)}{2\bar{\rho}}, \tag{2.4}$$

여기서 $S(y)$는 관심변수 y의 표준편차이다. 따라서 각각의 추정량을 구하여 대입하면 편향의 최대값을 구할 수 있다. 또한 식 (2.1)을 식 (2.4)에 대입하면 다음의 결과를 얻는다.

$$\left|B\left(\hat{\bar{Y}}_{\mathrm{HT}}\right)\right| \le \frac{(1 - R(\rho))\, S(y)}{2\bar{\rho}} = \frac{S(\rho)S(y)}{\bar{\rho}}.$$

결론적으로 표본조사에서 관심변수 y의 무응답과 관련된 최대 편향은 응답 확률 ρ_k의 변동계수와 관심변수의 표준편차에 의해 결정된다. 본 연구에서는 이 값이 편향의 최대값으로 그 의미가 크지 않고 Horvitz-Thompson 추정량일 경우에 얻어진 결과이기 때문에 모의실험에서는 살펴보지 않았다.

3. 분석에 사용될 통계 분석 방법

3.1. 일반화선형모형(generalized linear model)

응답률은 일반적으로 알려져 있지 않다. 따라서 자료에서 응답률을 추정해야 하며 이때 사용되는 모형이 일반화선형모형이다. 일반화선형모형에서 흔히 사용하는 모형은 로짓모형(logit model)과 프로빗모

형(probit model)이다. 이 두 모형을 자료에 적합하여 응답률을 추정하며 이때 얻어진 결과를 경향점수(성향점수, propensity score)라고 부른다. 본 연구에서 사용된 일반화선형모형 식은 다음과 같다.

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}.$$

이 선형모형식에서 흔히 사용하는 변환 방법은 로짓 변환과 프로빗 변환이다. 로짓 변환은 로지스틱회귀 모형을 가정할 경우에 사용되며 프로빗 변환은 정규분포의 누적확률함수(cumulative distribution function; cdf), Φ를 사용한다. 따라서 로짓 모형인 경우 $\eta = \log(\pi/(1-\pi))$를 사용하고 프로빗 모형인 경우는 $\eta = \Phi(\pi)^{-1}$를 사용한다. 이 분석은 SAS 등 기초적인 통계패키지를 이용하여 쉽게 수행할 수 있다. Bethlehem 등 (2008)과 Schouten 등 (2009)에 의하면 R-지수의 변동 상황을 확인하기 위해서는 기준이 되는 같은 모형과 같은 보조 변수가 사용되어 한다. 따라서 로짓 모형과 프로빗 모형 중에서 하나를 선택하고 자료를 잘 설명할 수 있는 보조 변수를 선택하여 응답률 또는 경향점수 $\hat{\rho}_k$을 계산하여야 한다.

3.2. 층별 최종 가중치

표본설계는 모집단을 잘 대표할 수 있는 최적의 표본을 추출하기 위해 사용되며 다양한 표본설계 방법이 연구되었고 또한 사용되고 있다. 가장 강력하면서도 흔히 사용되는 방법이 층화추출법(stratified sampling)이다. 이 방법은 모집단을 여러 개의 층으로 나눈 후 각 층에서 표본을 추출하는 방법이다. 이때 대표성을 만족하는 표본추출을 위해 표본추출틀(sampling frame)이 사용되는데 이 표본추출틀에는 층화에 필요한 정보가 포함되어 있다. 여기서 주어진 표본추출틀의 층별 자료 수, 즉 조사 모집단수(N_h)와 표본수(n_h)를 이용하면 가중치 $w_h = N_h/n_h$가 구해진다. 최종적으로 층별 가중치와 경향점수를 이용하여 R-지수를 구한다.

4. 모의실험

4.1. 모집단 생성 과정

이 절에서는 모의실험에서 사용된 모집단 생성과정을 설명하였다. 독립변수는 현실 자료와 유사하도록 지역(3개), 산업 분류(6개)의 층화 변수와 종사자수(연속형 변수 1개)를 공변량으로 설정하며 이때 관심변수는 연속형 자료인 매출을 가정한다. 응답과 관련된 로짓 모형은 다음과 같이 설정한다.

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha_i + \tau_j + \gamma x,$$

여기서 $i = 1, \ldots, 3$은 지역을 $j = 1, \ldots, 6$은 산업을 의미한다. 또한 x는 공변량인 종사자수이다. 이때 국내 현실에 맞도록 오른쪽으로 꼬리가 긴 자료를 생성하였으며 이를 위해 x는 감마분포, Gamma(1, 50)에서 자료를 생성한다.

모의실험을 간단히 하기 위해 무응답이 있는 모집단을 다음과 같은 방법으로 생성하였다.

모집단 1: 지역별, 산업분류별, 종사자수에 무관하게 무응답 비율을 5%, 10%, 15%로 모집단을 생성한다.

모집단 2: 지역에 따라 3%, 5%, 7%를 무응답 비율로 하여 모집단을 생성한다. 즉 α_i만 응답률에 영향을 준다.

모집단 3: 모집단 2에 추가하여 산업분류에 따라 2개 산업은 5%, 2개 산업은 10%, 나머지 2개 산업은 15%의 무응답을 고려하여 모집단을 생성한다. 예를 들면 지역 1의 산업분류 1과 2는

Table 4.1. Design weights for strata

Region	Industry					
	1	2	3	4	5	6
1	10.00	50.00	33.33	25.00	20.00	16.67
2	50.00	25.00	16.67	12.50	10.00	8.33
3	33.33	16.67	11.11	8.33	6.67	5.56

8%(3% + 5%) 무응답이 발생하고, 지역 1의 산업분류 3과 4는 13% 무응답이 발생하도록 만든다. 같은 방법으로 지역 3의 산업분류 5와 6은 22%의 무응답이 발생하도록 한다. 따라서 α_i와 τ_j가 모두 응답률에 영향을 준다.

모집단 4: 모집단 3에 추가하여 종사자수에 따라 두 가지 경우로 나누어 모집단을 생성한다. Case 1: 50명 이내에는 30%, 50–100명은 20%, 100명 이상은 10%의 추가 무응답을 만든다. Case 2: 50명 이내에는 10%, 50–100명은 20% 그리고 100명 이상은 30%의 추가 무응답을 만든다. 따라서 모든 독립변수가 응답률에 영향을 주며 다수의 무응답이 발생한다.

다음은 종속변수 생성을 위해 사용된 자료생성 모형이다.

$$y_{ijk} = \alpha_i + \tau_j + \epsilon_{ijk}, \qquad i = 1, \ldots, 3,\ j = 1, \ldots, 6,$$

여기서 종속변수의 오차 ϵ_{ijk}는 정규분포와 로그-정규 분포를 따른다고 가정하였다. 이는 현실자료 분석에서 매출과 같은 종속변수는 로그-정규 분포 가정을 많이 사용하기 때문이다. 실제 자료에서는 종속변수와 공변량 간에 관계가 있을 수 있다. 그러나 본 연구에서는 무응답으로 인해 발생한 결측의 영향을 줄이기 위해 무응답 보정인자를 설계 가중치에 곱해 최종 가중치를 구하는 무응답 가중치 보정법을 사용하기 때문에 만약 공변량이 관심변수와 관계가 있으면 이 가중치 보정법을 이용한 추정 결과는 매우 부정확해져 R-지수와의 관계를 파악하기 어렵게 된다. 이에 응답 모형에서 사용된 독립변수인 공변량은 자료와 무관하다고 가정하였다.

모형에 사용된 상수 값으로 $\alpha_1 = 20$, $\alpha_2 = 40$, $\alpha_3 = 60$, $\tau_1 = \tau_2 = 10$, $\tau_3 = \tau_4 = 50$, $\tau_5 = \tau_6 = 100$을 사용하였다. 또한 관심 변수 y_i의 오차는 $\epsilon_i \overset{\text{iid}}{\sim} N(0,4), N(0,16)$와 $\log(\epsilon_i) \overset{\text{iid}}{\sim} N(0,1), N(0,2)$에서 생성하였다. 이는 식 (2.4)에서 편향이 관심자료의 표준편차에 영향을 받는 것으로 알려져 있으므로 분산이 다른 경우에 편향의 차이를 확인하기 위함이다.

또한 현실적으로 설계가중치가 층별로 다르기 때문에 본 연구에서는 설계가중치를 층별로 다양하게 생성하였다. Table 4.1에 층별로 주어진 설계가중치를 수록하였다. 각 층별 모집단 수는 10,000을 사용하였기 때문에 설계가중치를 이용하면 표본수는 쉽게 파악할 수 있다.

관심변수의 총합은 층별 가중치 w_h와 자료 값 y_{hi}의 곱의 합으로 계산되며 추정된 총합 $\hat{t}$와 모든 자료가 조사된 경우의 총합 t^{TRUE}가 구해진다. R-지수의 크기에 따른 정확성과 정밀성 분석을 위해 본 연구에서는 비교 통계량으로 편향(bias)와 절대편향(absolute bias; AB) 그리고 제곱근 평균제곱오차(root mean squared error; RMSE)를 사용하였으며 정의는 다음과 같다.

$$\text{Bias} = \frac{1}{K}\sum_{k=1}^{K}\left(\hat{t}_k - t_k^{\text{TRUE}}\right),$$

$$\text{AB} = \frac{1}{K}\sum_{k=1}^{K}\left|\hat{t}_k - t_k^{\text{TRUE}}\right|,$$

Table 4.2. Results of R-indicator and comparison statistics for population 1 with $N(0, 4)$

N-R rate (%)	R-indicator	Bias	Absolute bias	RMSE
5	0.997	−198.0	3411.7	4320.7
10	0.996	−93.8	3548.7	4436.2
15	0.995	−108.8	3479.3	4299.9
20	0.994	35.0	3625.9	4499.5

R-indicator = representativeness indicator; RMSE = root mean squared error.

Table 4.3. Results of R-indicator and comparison statistics for population 1 with $N(0, 16)$

N-R rate (%)	R-indicator	Bias	Absolute bias	RMSE
5	0.997	−396.0	6823.4	8641.3
10	0.996	−187.6	7097.3	8872.3
15	0.995	−217.6	6958.6	8599.9
20	0.994	69.9	7251.8	8999.0

R-indicator = representativeness indicator; RMSE = root mean squared error.

$$\text{RMSE} = \left\{ \frac{1}{K} \sum_{k=1}^{K} \left(\hat{t}_k - t_k^{\text{TRUE}} \right)^2 \right\}^{\frac{1}{2}}.$$

이때 반복수는 $K = 1{,}000$을 사용하였다.

4.2. 모의실험 결과

다음의 Table 4.2에서 Table 4.22까지 모의실험 결과를 수록하였다. Table 4.2에서 Table 4.12는 오차 생성 시 정규분포를 이용한 결과이며 Table 4.13에서 Table 4.22는 로그-정규 분포를 이용한 결과이다.

4.2.1. 정규분포를 이용한 결과 정규분포 결과는 오차에서 $\sigma = 2$와 4를 이용한 결과가 수록되어 있다. R-지수의 경우 관심변수 y값에 무관하기 때문에 어떤 분산을 사용하여도 결과는 같다. 또한 편향, 절대편향, RMSE는 표준편차의 크기와 비례하기 때문에 $\sigma = 2$인 결과에 비해 $\sigma = 4$의 결과는 비교통계량이 모두 '2'배의 결과를 주고있다. 따라서 $N(0, 16)$에서 얻어진 결과는 추가적인 정보를 주고 있지 않다. 구체적으로 얻어진 결과는 다음과 같다. 먼저 Table 4.2와 Table 4.3은 모집단 1의 결과로 층과 무관하게 모든 층에서 고루 무응답이 발생한다. 따라서 전체 모집단에서 MCAR으로 무응답이 발생하여 무응답 비율에 무관하게 높은 R-지수를 보이고 있다. 또한 층별로 가중치를 보정하여 추정하기 때문에 편향은 크게 발생하지 않고 있다. 다만 무응답으로 표본수가 줄어들기 때문에 RMSE가 약간 증가하는 것 같으나 그 차이는 미미하다. 다음으로 Table 4.4와 Table 4.7은 모집단 2와 모집단 3의 결과로 층별로 다른 크기의 무응답이 발생한다. 따라서 층별 무응답률이 다르기 때문에 R-지수는 낮아진다. 그러나 모수 추정 시 층별 보정가중치를 사용하기 때문에 편향은 작게 나온다. 또한 Table 4.2에서 Table 4.7의 절대 편향과 RMSE를 기준으로 살펴보아도 큰 차이를 보이고 있지 않다. 따라서 층별로 무응답률이 상이하여 상대적으로 낮은 R-지수가 얻어지더라도 층 내의 무응답률이 비슷하다면 전체 모수 추정의 정확성에는 크게 영향을 미치지 않는다. 이러한 결과는 Table 4.8에서 Table 4.11에서도 얻어진다. Table 4.8에서 Table 4.11은 모집단 4의 결과로 이 표에서는 종사자수에 따라 추가로 무응답이 발생하고 있다. 따라서 층별로 응답률이 더욱 달라지기 때문에 R-지수는 더욱 작아지게 된다. 그러나 편향과 절대편향 그리고 RMSE는 크게 증가하지 않는다.

Table 4.4. Results of R-indicator and comparison statistics for population 2 with $N(0, 4)$

N-R rate (%) in α_i	R-indicator	Bias	Absolute bias	RMSE
(3, 5, 7)	0.967	219.9	3303.2	4105.6
(3, 5, 15)	0.895	102.9	3438.9	4353.3
(3, 7, 15)	0.900	−58.7	3444.1	4321.9
(5, 10, 15)	0.918	−29.2	3442.7	4338.0

R-indicator = representativeness indicator; RMSE = root mean squared error.

Table 4.5. Results of R-indicator and comparison statistics for population 2 with $N(0, 16)$

N-R rate (%) in α_i	R-indicator	Bias	Absolute bias	RMSE
(3, 5, 7)	0.967	439.8	6606.4	8211.1
(3, 5, 15)	0.895	205.9	6877.8	8706.6
(3, 7, 15)	0.900	−117.4	6888.3	8643.7
(5, 10, 15)	0.918	−58.4	6885.3	8676.1

R-indicator = representativeness indicator; RMSE = root mean squared error.

Table 4.6. Results of R-indicator and comparison statistics for population 3 with $N(0, 4)$

N-R rate (%) in α_i	N-R rate (%) in β_i	R-indicator	Bias	Absolute bias	RMSE
(3, 5, 7)	(3, 5, 7)	0.956	−68.9	3542.8	4456.9
	(2, 5, 10)	0.930	−260.2	3498.8	4412.2
(3, 5, 15)	(3, 5, 7)	0.895	−240.7	3506.5	4366.3
	(2, 5, 10)	0.885	26.0	3528.4	4366.9
(3, 7, 15)	(3, 5, 7)	0.899	183.9	3505.6	4296.7
	(2, 5, 10)	0.889	−278.0	3535.4	4423.2
(5, 10, 15)	(3, 5, 7)	0.915	141.9	3670.5	4596.1
	(2, 5, 10)	0.902	179.3	3545.3	4469.9

R-indicator = representativeness indicator; RMSE = root mean squared error.

Table 4.7. Results of R-indicator and comparison statistics for population 3 with $N(0, 16)$

N-R rate (%) in α_i	N-R rate (%) in β_i	R-indicator	Bias	Absolute bias	RMSE
(3, 5, 7)	(3, 5, 7)	0.956	−137.8	7085.5	8913.8
	(2, 5, 10)	0.930	−520.4	6997.7	8824.3
(3, 5, 15)	(3, 5, 7)	0.895	−481.5	7012.9	8732.5
	(2, 5, 10)	0.885	52.0	7056.8	8733.8
(3, 7, 15)	(3, 5, 7)	0.899	367.7	7011.3	8593.5
	(2, 5, 10)	0.889	−555.9	7070.8	8846.4
(5, 10, 15)	(3, 5, 7)	0.915	283.7	7341.0	9192.1
	(2, 5, 10)	0.902	358.5	7090.6	8939.9

R-indicator = representativeness indicator; RMSE = root mean squared error.

Table 4.8. Results of R-indicator and comparison statistics for population 4 with $N(0, 4)$ and non-response rate of employee (30%, 20%, 10%)

N-R rate (%) in α_i	N-R rate (%) in β_i	R-indicator	Bias	Absolute bias	RMSE
(3, 5, 7)	(3, 5, 7)	0.873	249.4	4005.9	4952.8
	(2, 5, 10)	0.870	90.3	4005.9	5041.5
(3, 5, 15)	(3, 5, 7)	0.860	−265.2	4131.1	5124.2
	(2, 5, 10)	0.856	−132.5	4061.9	5172.2
(3, 7, 15)	(3, 5, 7)	0.862	222.0	4097.6	5108.5
	(2, 5, 10)	0.859	199.8	3957.6	5013.0
(5, 10, 15)	(3, 5, 7)	0.870	−65.7	4178.0	5172.7
	(2, 5, 10)	0.867	35.6	4011.2	4969.4

R-indicator = representativeness indicator; RMSE = root mean squared error.

Table 4.9. Results of R-indicator and comparison statistics for population 4 with $N(0, 16)$ and non-response rate of employee (30%, 20%, 10%)

N-R rate (%) in α_i	N-R rate (%) in β_i	R-indicator	Bias	Absolute bias	RMSE
(3, 5, 7)	(3, 5, 7)	0.873	498.9	8011.7	9905.6
	(2, 5, 10)	0.870	180.6	8011.7	10083.1
(3, 5, 15)	(3, 5, 7)	0.860	−530.4	8262.2	10248.4
	(2, 5, 10)	0.856	−264.9	8123.8	10344.3
(3, 7, 15)	(3, 5, 7)	0.862	444.0	8195.1	10217.0
	(2, 5, 10)	0.859	399.6	7915.1	10026.0
(5, 10, 15)	(3, 5,7)	0.870	−131.4	8356.1	10345.3
	(2, 5, 10)	0.867	71.2	8022.4	9938.8

R-indicator = representativeness indicator; RMSE = root mean squared error.

Table 4.10. Results of R-indicator and comparison statistics for population 4 with $N(0, 4)$ and non-response rate of employee (10%, 20%, 30%)

N-R rate (%) in α_i	N-R rate (%) in β_i	R-indicator	Bias	Absolute bias	RMSE
(3, 5, 7)	(3, 5, 7)	0.870	−35.1	3705.0	4651.9
	(2, 5, 10)	0.865	−114.6	3802.8	4757.8
(3, 5, 15)	(3, 5, 7)	0.853	27.5	3585.2	4532.2
	(2, 5, 10)	0.849	−184.7	3720.8	4708.7
(3, 7, 15)	(3, 5, 7)	0.856	−75.9	4023.2	5076.4
	(2, 5, 10)	0.852	202.9	3695.4	4687.1
(5, 10, 15)	(3, 5, 7)	0.865	5.9	3919.0	4863.4
	(2, 5, 10)	0.860	−63.8	3984.3	4977.3

R-indicator = representativeness indicator; RMSE = root mean squared error.

Table 4.11. Results of R-indicator and comparison statistics for population 4 with $N(0, 16)$ and non-response rate of employee (10%, 20%, 30%)

N-R rate (%) in α_i	N-R rate (%) in β_i	R-indicator	Bias	Absolute bias	RMSE
(3, 5, 7)	(3, 5, 7)	0.870	−70.3	7410.0	9303.7
	(2, 5, 10)	0.865	−229.1	7605.7	9515.5
(3, 5, 15)	(3, 5, 7)	0.853	−54.9	7170.4	9064.4
	(2, 5, 10)	0.849	−369.4	7441.6	9417.5
(3, 7, 15)	(3, 5, 7)	0.856	−151.7	8046.3	10152.7
	(2, 5, 10)	0.852	405.8	7390.8	9374.2
(5, 10, 15)	(3, 5, 7)	0.865	11.8	7837.9	9726.8
	(2, 5, 10)	0.860	−127.7	7968.7	9954.5

R-indicator = representativeness indicator; RMSE = root mean squared error.

Table 4.12. Results of R-indicator for population 4 with $N(0, 4)$

N-R rate (%) in α_i	N-R rate (%) in β_i	R-indicator			
		Case 1	Case 1*	Case 2	Case 2*
(3, 5, 7)	(3, 5, 7)	0.873	0.971	0.870	0.968
	(2, 5, 10)	0.870	0.950	0.865	0.945
(3, 5, 15)	(3, 5, 7)	0.860	0.921	0.853	0.913
	(2, 5, 10)	0.856	0.912	0.849	0.904
(3, 7, 15)	(3, 5, 7)	0.862	0.925	0.856	0.918
	(2, 5, 10)	0.859	0.916	0.852	0.908
(5, 10, 15)	(3, 5, 7)	0.870	0.938	0.865	0.932
	(2, 5, 10)	0.867	0.927	0.860	0.920

Table 4.13. Results of R-indicator and comparison statistics for population 1 with $\log -N(0,1)$

N-R rate (%)	R-indicator	Bias	Absolute bias	RMSE
5	0.997	−237.6	3735.3	4677.5
10	0.996	−236.9	3777.5	4724.8
15	0.995	−41.3	3747.8	4691.5
20	0.994	188.4	3906.8	4906.2

R-indicator = representativeness indicator; RMSE = root mean squared error.

Table 4.14. Results of R-indicator and comparison statistics for population 1 with $\log -N(0,2)$

N-R rate (%)	R-indicator	Bias	Absolute bias	RMSE
5	0.997	−608.6	11601.4	14711.5
10	0.996	−839.6	11578.4	14703.8
15	0.995	162.0	11958.8	15086.8
20	0.994	635.1	12505.2	16071.7

R-indicator = representativeness indicator; RMSE = root mean squared error.

Table 4.15. Results of R-indicator and comparison statistics for population 2 with $\log -N(0,1)$

N-R rate (%) in α_i	R-indicator	Bias	Absolute bias	RMSE
(3, 5, 7)	0.967	191.1	3520.3	4462.5
(3, 5, 15)	0.895	102.3	3635.0	4587.5
(3, 7, 15)	0.900	34.9	3793.3	4699.9
(5, 10, 15)	0.918	44.3	3783.8	4747.9

R-indicator = representativeness indicator; RMSE = root mean squared error.

Table 4.16. Results of R-indicator and comparison statistics for population 2 with $\log -N(0,2)$

N-R rate (%) in α_i	R-indicator	Bias	Absolute bias	RMSE
(3, 5, 7)	0.967	519.6	11056.5	15317.6
(3, 5, 15)	0.895	356.5	11589.4	14553.6
(3, 7, 15)	0.900	15.8	11732.0	14731.9
(5, 10, 15)	0.918	368.9	11920.6	15306.1

R-indicator = representativeness indicator; RMSE = root mean squared error.

Table 4.17. Results of R-indicator and comparison statistics for population 3 with $\log -N(0,1)$

N-R rate (%) in α_i	N-R rate (%) in β_i	R-indicator	Bias	Absolute bias	RMSE
(3, 5, 7)	(3, 5, 7)	0.956	−323.1	3787.6	4737.6
	(2, 5, 10)	0.930	−31.6	3914.4	4882.1
(3, 5, 15)	(3, 5, 7)	0.895	−213.8	3804.3	4789.6
	(2, 5, 10)	0.885	−120.7	3854.0	4880.2
(3, 7, 15)	(3, 5, 7)	0.899	39.2	3788.2	4709.7
	(2, 5, 10)	0.889	−222.1	3829.9	4735.8
(5, 10, 15)	(3, 5, 7)	0.915	63.6	3853.6	4869.9
	(2, 5, 10)	0.902	99.8	3902.9	4844.7

R-indicator = representativeness indicator; RMSE = root mean squared error.

Table 4.18. Results of R-indicator and comparison statistics for population 3 with $\log -N(0,2)$

N-R rate (%) in α_i	N-R rate (%) in β_i	R-indicator	Bias	Absolute bias	RMSE
(3, 5, 7)	(3, 5, 7)	0.956	−1263.0	11608.7	14565.6
	(2, 5, 10)	0.930	130.9	12211.3	15421.8
(3, 5, 15)	(3, 5, 7)	0.895	−404.5	12049.0	15509.1
	(2, 5, 10)	0.885	−453.1	12328.7	15601.6
(3, 7, 15)	(3, 5, 7)	0.899	−17.4	11818.4	15033.1
	(2, 5, 10)	0.889	−478.8	11587.7	14930.3
(5, 10, 15)	(3, 5, 7)	0.915	164.5	12058.8	15450.6
	(2, 5, 10)	0.902	218.4	12256.6	15448.3

R-indicator = representativeness indicator; RMSE = root mean squared error.

Table 4.19. Results of R-indicator and comparison statistics for population 4 with $\log -N(0,1)$ and non-response rate of employee (30%, 20%, 10%)

N-R rate (%) in α_i	N-R rate (%) in β_i	R-indicator	Bias	Absolute bias	RMSE
(3, 5, 7)	(3, 5, 7)	0.873	362.6	4349.4	5433.9
	(2, 5, 10)	0.870	21.3	4207.9	5306.8
(3, 5, 15)	(3, 5, 7)	0.860	−135.3	4108.1	5182.9
	(2, 5, 10)	0.856	−169.8	4425.3	5636.3
(3, 7, 15)	(3, 5, 7)	0.862	277.2	4456.7	5649.1
	(2, 5, 10)	0.859	250.6	4482.6	5612.2
(5, 10, 15)	(3, 5, 7)	0.870	89.3	4301.5	5418.1
	(2, 5, 10)	0.867	158.3	4384.4	5470.8

R-indicator = representativeness indicator; RMSE = root mean squared error.

Table 4.20. Results of R-indicator and comparison statistics for population 4 with $\log -N(0,2)$ and non-response rate of employee (30%, 20%, 10%)

N-R rate (%) in α_i	N-R rate (%) in β_i	R-indicator	Bias	Absolute bias	RMSE
(3, 5, 7)	(3, 5, 7)	0.873	738.8	13550.4	17277.4
	(2, 5, 10)	0.870	−108.8	12789.1	16443.2
(3, 5, 15)	(3, 5, 7)	0.860	−362.0	13010.8	16505.5
	(2, 5, 10)	0.856	−647.2	13682.4	17568.1
(3, 7, 15)	(3, 5, 7)	0.862	602.2	13935.7	18057.0
	(2, 5, 10)	0.859	1024.0	14261.9	18164.4
(5, 10, 15)	(3, 5, 7)	0.870	375.1	13226.9	16810.4
	(2, 5, 10)	0.867	510.9	13795.4	17287.7

R-indicator = representativeness indicator; RMSE = root mean squared error.

Table 4.21. Results of R-indicator and comparison statistics for population 4 with $\log -N(0,1)$ and non-response rate of employee (10%, 20%, 30%)

N-R rate (%) in α_i	N-R rate (%) in β_i	R-indicator	Bias	Absolute bias	RMSE
(3, 5, 7)	(3, 5, 7)	0.870	−105.2	4138.5	5212.1
	(2, 5, 10)	0.865	−277.0	4055.3	5106.1
(3, 5, 15)	(3, 5, 7)	0.853	−0.8	3918.2	4922.4
	(2, 5, 10)	0.849	65.6	4126.8	5153.3
(3, 7, 15)	(3, 5, 7)	0.856	−78.8	4299.9	5422.1
	(2, 5, 10)	0.852	223.7	4002.8	5035.8
(5, 10, 15)	(3, 5, 7)	0.865	211.6	4209.8	5225.0
	(2, 5, 10)	0.860	−175.8	4170.4	5260.0

R-indicator = representativeness indicator; RMSE = root mean squared error.

Table 4.22. Results of R-indicator and comparison statistics for population 4 with $\log -N(0,2)$ and non-response rate of employee (10%, 20%, 30%)

N-R rate (%) in α_i	N-R rate (%) in β_i	R-indicator	Bias	Absolute bias	RMSE
(3, 5, 7)	(3, 5, 7)	0.870	−119.6	13027.3	16515.7
	(2, 5, 10)	0.865	−761.0	12737.0	16611.8
(3, 5, 15)	(3, 5, 7)	0.853	−55.4	12723.2	15978.3
	(2, 5, 10)	0.849	473.9	13179.0	16674.7
(3, 7, 15)	(3, 5, 7)	0.856	−134.8	13202.1	17072.7
	(2, 5, 10)	0.852	447.3	12272.3	17072.7
(5, 10, 15)	(3, 5, 7)	0.865	684.7	13084.8	16384.8
	(2, 5, 10)	0.860	−709.4	12930.4	16530.9

R-indicator = representativeness indicator; RMSE = root mean squared error.

다음으로 Table 4.12에는 모집단 4의 case 1과 case 2의 R-지수 계산 과정에서 공변량 x를 일반화선형모형에 추가한 경우와 추가하지 않은 결과를 수록하였다. 즉 case 1*, case 2*는 일반화선형모형에 공변량 x가 없는 경우이다. 이 결과를 살펴보면 R-지수가 매우 상승한 것을 확인할 수 있다. 따라서 일반화선형모형 선택 시 유의한 변수의 존재 유무가 R-지수에 매우 중요한 요인으로 작용하며 이로 인해 결과가 매우 민감하게 반응하는 것을 확인할 수 있다. 반면 모집단 1에서 모집단 3의 결과인 Table 4.2에서 Table 4.7의 경우에는 공변량과 무관하게 무응답이 발생하고 있음에도 공변량을 모형에 추가하여 구한 결과이다. 당연한 결과이지만 이러한 경우에는 R-지수에 크게 영향을 주지 않는다.

4.2.2. 로그-정규분포를 이용한 결과

Table 4.13에서 Table 4.22는 오차가 로그-정규분포를 따른다는 가정 하에서 얻어진 결과이다. 전체적으로 정규분포를 이용한 결과와 유사한 경향을 보이고 있다. 특히 R-지수는 경우 관심변수 y값과 무관하게 응답 유무만이 반영되므로 동일한 R-지수를 주고 있다. 여기서 로그-정규 분포의 분산은 $(e^{\sigma^2}-1)e^{2\mu+\sigma^2}$이므로 $\mu=0, \sigma^2=1$인 경우에는 분산이 약 4.671 또는 표준편차가 $\sqrt{4.671}\approx 2.161$이 된다. 정규분포의 표준편차가 2인 결과와 비교하면, 즉 Table 4.2와 Table 4.13의 결과를 비교하면 약 1.0805배의 편향, 절대편향, RMSE가 예상된다. 결과를 비교하면 절대편향과 RMSE의 경우 유사한 결과 값을 주고 있어 예상과 일치한다. 다만 편향의 경우는 이와 다른 결과를 주고 있다. 다음으로 Table 4.13과 Table 4.14의 결과를 비교하면 다음의 결과를 얻을 수 있다. 먼저 $\mu=0, \sigma^2=2$인 경우에는 분산이 약 47.2이므로 표준편차는 약 6.870이 된다. 따라서 Table 4.13의 표준편차와 비교하면 약 3.18배 증가할 것으로 예상된다. 결과를 살펴보면 절대편향과 RMSE는 예상된 결과와 유사한 값을 주고 있다. 반면 편향은 다른 결과를 주고 있다. 이러한 현상은 다른 모든 표에서도 모두 얻어지고 있다. 2.3절의 수식 (2.4)의 결과를 보면 편향의 최대값은 응답률의 변동계수(coefficient of variation; cc)와 자료의 표준편차의 곱보다 작은 것으로 나타났으나 실제 편향과 관련된 결과는 아직 정립되지 않았다. 그러나 모의실험 결과를 살펴보면 같은 R-지수라도 편향은 자료의 표준편차만 영향을 받는 것이 아니라 응답률 그리고 자료의 분포 형태도 영향을 받을 수 있는다는 것을 확인할 수 있다.

4.3. 실제 자료 분석

이 절에서는 변형된 2010년 경제총조사 자료 중 중소기업자료를 이용하여 자료 분석을 실시하였다. 본 연구에서 사용된 자료는 산업 대분류 C, D, F 등 3개 대분류와 6개 지역 자료이다. 이에 추가하여 사업체 규모(종사자수 기준)와 종사자수 그리고 매출이 있다. 사업체 규모는 4인 이하, 19인 이하, 그리고 20인 이상으로 구분하였다. 모의실험에서와 같이 관심변수는 사업체의 매출이고, 로짓 모형이 사용되었으며 독립변수로 층화 변수는 3개 대분류, 6개 지역 그리고 공변량으로 종사자수를 사용하였다. 자료에는 무응답이 없어 3개 산업분류에 무응답 비율 (3, 5, 7), (3, 5, 10) 그리고 (3, 10, 15)%을 이용하여 무응답을 생성하였다. 또한 6개 지역의 무응답 비율은 3%에서 30% 이내로 다양하게 선택하였다. 또한 흔히 종사자수가 매출에 영향을 미치고 또한 종사자수가 무응답 비율에 영향을 주기 때문에 종사자 기준으로 얻어진 사업체 규모별로도 무응답을 추가하였다. 여기서 무응답은 5%에서 30%이내를 사용하였다. 모수 추정을 위해 사용된 가중치는 지역별, 산업분류별로 구하였으며 이는 모의실험과 같은 조건을 만들기 위해서이다. 모의실험과 같은 비교통계량이 사용되었으며 반복수는 $K=1{,}000$이다.

다음의 Table 4.23과 Table 4.24에 결과를 수록하였다. Table 4.23은 무응답 생성 시에 사용된 층별 무응답 비율을 작성한 표이며 Table 4.24는 R-지수와 비교통계량 결과이다. Table 4.24에서 rate 1, rate 5 그리고 rate 6의 결과는 지역과 산업분류의 무응답이 영향을 주지만 층별로 무응답의 영향이 크지 않고 종사자수와 관련된 무응답은 관련이 없다. 따라서 추정 시 지역과 산업분류별로 층별 가중치를 사용

Table 4.23. Non-response rates for real data analysis

N-R rate (%) in region	N-R rate (%) in company size	N-R rate (%) in number of employees	N-R rate
(3, 5, 7)	(10, 3, 5, 5, 10, 10)	(10, 10, 10)	Rate 1
		(20, 10, 30)	Rate 2
	(10, 3, 5, 5, 10, 30)	(30, 10, 15)	Rate 3
		(5, 10, 20)	Rate 4
(3, 5, 10)	(10, 3, 5, 5, 15, 15)	(20, 20, 20)	Rate 5
		(10, 30, 10)	Rate 6
(3, 10, 15)	(30, 30, 5, 5, 30, 30)	(30, 10, 15)	Rate 7
		(5, 10, 20)	Rate 8

Table 4.24. Results of R-indicator and comparison statistics for real data analysis

N-R rate	R-indicator	Bias ($\times 10^4$)	Absolute bias ($\times 10^4$)	RMSE ($\times 10^4$)
Rate 1	0.953	−134	1038	1304
Rate 2	0.938	−4266	4266	4444
Rate 3	0.806	7418	7418	7547
Rate 4	0.794	−6088	6088	6208
Rate 5	0.923	−79	1073	1340
Rate 6	0.924	−288	1053	1324
Rate 7	0.831	7399	7399	7575
Rate 8	0.847	−6176	6176	6327

하기 때문에 상대적으로 높은 R-지수와 우수한 비교통계량 결과를 주고 있다. 즉 상대적으로 무응답이 크게 영향을 주지 않는다.

그러나 추가로 종사자수가 무응답에 크게 영향을 주는 경우는 R-지수 뿐만 아니라 비교통계량의 결과에도 큰 영향을 주고 있다. 먼저 소규모 사업체의 무응답 비율이 30%인 rate 3과 rate 7의 결과를 살펴보면 편향이 큰 양수로 나와 과대추정되고 있다. 즉 종사자수는 매출과 관계가 있고, 종사자수가 작아 매출이 작은 사업체에 다수의 무응답이 발생하지만 매출이 큰 사업체의 가중치를 증가시키는 보정가중치를 사용하였기 때문에 과대추정이 된다. 또한 편향의 절대값과 절대편향이 일치하고 RMSE와 차이가 작기 때문에 RMSE의 대부분을 편향이 차지하고 있음도 확인할 수 있다. 반면 대규모 사업체에 무응답 비율이 상대적으로 큰 경우인 rate 2, rate 4 그리고 rate 8의 경우에는 반대의 현상이 일어나고 있다. 즉 이 경우는 과소추정이 발생하여 편향이 매우 큰 음수값을 갖게 된다.

결론적으로 높은 R-지수는 정확성을 보장하지만 이때 얻어진 R-지수는 무응답에 영향을 주는 모든 독립변수를 이용하여 구해야만 얻어진 결과가 의미가 있다. 또한 편향은 추정방법 특히 유의한 독립변수의 사용 유무에 따라 많은 영향을 받는 다는 것도 확인할 수 있다.

5. 결론

본 논문에서는 최근 개발된 정확성 측도인 R-지수에 관하여 연구하였다. 대부분의 무응답 분석은 정밀도, 즉 분산과 관련된 내용이다. 그러나 최근 R-지수의 개발로 정확성과 관련된 내용도 함께 고려할 필요가 있다. 본 연구에서는 R-지수와 편향과의 관계를 모의실험을 통하여 규명하고자 하였다. 아직 R-지수가 몇 이상이면 정확성이 보장되는지에 관한 정확한 기준이 마련되어 있지 않다. 이는 본 연구에서도 확인되었지만 추정 방법, 무응답 패턴과 독립변수와의 관계, 관심변수와 독립변수와의 관계 그리고

관심 변수의 분포와 분산 등이 모두 연관되어 있기 때문으로 판단된다. 다만 본 연구의 모의실험과 실제 자료 분석을 통하여 얻은 결과를 바탕으로 다음의 결론을 얻을 수 있다. 먼저 무응답에 영향을 주는 모든 변수를 사용하여 R-지수를 구하여야 한다. 이는 무응답 패턴과 관계가 높은 독립변수가 모형에 포함되는 것과 포함되지 않은 것은 R-지수 값에 큰 영향을 주기 때문이다. 물론 무응답과 관련이 없거나 적은 독립변수가 모형에 포함이 되는 경우에는 R-지수에 크게 영향을 크게 주지 않는다. 따라서 중요 독립변수를 모형에 포함하지 않는 다면 R-지수값은 의미가 없다. R-지수는 정확성과 관련이 있고 이 개념은 편향과 관련이 있다. 관심 변수의 모수 추정 시 가중치 보정법, 대체법 등을 이용한다면 무응답의 영향력을 줄일 수 있다. 이때 당연한 결론이지만 충분한 정보가 있는 보조 자료가 있다면 이 정보를 이용한 무응답 대체법을 사용하여야 효과적으로 편향을 줄일 수 있을 것으로 판단된다.

References

Agresti, A. (2002). *Categorical Data Analysis, Wiley Series in Probability and Statistics*, John Wiley and Sons, New York.

Bethlehem, J., Cobben, F., and Schouten, B. (2008). Indicator for the representativeness of survey response. In *Proceedings of Statistics Canada Symposium, Data Collection: Challenges, Achievements and New Directions*.

Kruskal, W. and Mosteller, F. (1979). Representative sampling III: current statistical literature, *International Statistical Review*, **47**, 245–265.

Ouwehand, P. and Schouten, B. (2014). Measuring representativeness of short-term business statistics, *Journal of Official Statistics*, **30**, 623–649.

Schouten, B., Bethlehem, J., Beullens, K., Kleven, O., Loosveldt, G., Luiten, A., Rutar, K., Shlomo, N., 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., Calinescu, M., and Luiten, A. (2013). Optimizing quality of response through adaptive survey designs, *Survey Methodology*, **39**, 29–58.

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 survey response, *Journal of Official Statistics*, **27**, 231–253.

표본 추출법에서 R-지수의 민감도에 관한 연구

이유진[a] · 신기일[a,1]

[a]한국외국어대학교 통계학과

(2016년 9월 26일 접수, 2016년 11월 14일 수정, 2016년 12월 12일 채택)

요 약

R-지수(representativeness indicator)는 무응답이 발생했을 때 표본의 대표성을 나타내주는 지표이다. 표본의 대표성은 모수 추정의 정확성(accuracy)과 관계가 있으며 정확성은 편향(bias)와 관계가 있다. 따라서 표본의 대표성을 나타내는 R-지수가 높으면 대표성이 높아 편향이 없고 정확성이 높은 결과를 얻을 수 있다. R-지수는 일반화선형모형의 로짓 또는 프로빗 모형을 적합한 후 얻어진 경향 점수(propensity score)에 의해 계산된다. 본 논문에서는 R-지수와 이질적인 층별 응답률과의 관련성을 연구하였으며 편향, 제곱근 RMSE 등과 같은 비교통계량이 무응답에 얼마나 민감한지 등을 모의실험을 통하여 살펴보았다. 또한 변형된 2010년 경제총조사 자료를 이용하여 실제 자료 분석도 실시하였다.

주요용어: 편향, 표본 대표성, 경향점수, 로짓모형

이 연구는 2016년 한국외국어대학교 교내연구비 지원을 받아 수행되었음.

[1]교신저자: (17035) 경기도 용인시 처인구 모현면 외대로 81, 한국외국어대학교 통계학과.
E-mail: keyshin@hufs.ac.kr