The Korean Journal of Applied Statistics (2019) DOI: https://doi.org/10.5351/KJAS.2019.32.4.631
32(4), 631–642

Bias adjusted estimation in a sample survey with linear response rate

Hee Young Chung[a] · Key-Il Shin[a,1]

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

(Received May 23, 2019; Revised June 25, 2019; Accepted June 25, 2019)

Abstract

Many methods have been developed to solve problems found in sample surveys involving a large number of item non-responses that cause inaccuracies in estimation. However, the non-response adjustment method used under the assumption of random non-response generates a bias in cases where the response rate is affected by the variable of interest. Chung and Shin (2017) and Min and Shin (2018) proposed a method to improve the accuracy of estimation by appropriately adjusting a bias generated when the response rate is a function of the variables of interest. In this study, we studied a case where the response rate function is linear and the error of the super population model follows normal distribution. We also examined the effect of the number of stratum population on bias adjustment. The performance of the proposed estimator was examined through simulation studies and confirmed through actual data analysis.

Keywords: linear inclusion probability, sample distribution, regressive model, sample weight

1. 서론

표본조사에서 적절한 무응답 처리 방법의 사용은 그 중요성이 날로 증가하고 있다. 이는 무응답 발생 비율이 현저히 높아지고 있고 이로 인해 최종 조사 자료 수의 감소로 표본오차 뿐만 아니라 비표본오차 또한 증가하기 때문이다. 이러한 문제를 풀기 위한 다양한 방법이 연구되고 실무에 적용되고 있다. 그러나 최근 관심 변수 값에 따라 무응답 또는 응답 비율이 달라지는 경우가 있으며 이때 흔히 사용하는 방법으로 무응답을 처리할 경우 편향이 발생된다.

편향을 보정하기 위해서는 편향의 크기가 알려져 있어야하기 때문에 현실적으로 정확한 편향 보정은 쉽지 않다. 그러나 최근 응답률이 관심변수의 지수 함수이고, 모집단에서 관심변수와 보조변수 간에 선형관계가 있는 경우, 편향의 크기를 알 수 있으며 Chung과 Shin (2017)은 응답률이 지수형일 때 정보적 표본설계 기법에서 얻어진 결과를 이용하여 편향을 보정함으로써 추정의 정확성을 향상시키는 방법을 제안하였으며 Schouten 등 (2009)은 응답률과 변수와의 관계를 연구하였다. 또한 Min과 Shin (2018)은 편향 보정에 필요한 최적 세부 층 개수와 세부 층 경계점에 관하여 연구하였다.

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2018R1D1A1B07042736).
[1]Corresponding author: Department of Statistics, Hankuk University of Foreign Studies, 81 Oedae-ro, Yongin-si, Gyeonggi-do 17035, South Korea. E-mail: keyshin@hufs.ac.kr

본 연구에서는 응답률이 선형이고 초모집단 모형이 회귀모형인 경우를 연구하였다. 또한 모의실험 결과를 바탕으로 모집단 자료 수가 최적 세부 층 개수 결정에 영향을 주는지도 살펴보았다.

본 논문의 구성은 다음과 같다. 먼저 2절에서는 기존의 연구내용을 간단히 설명하였다. 3절에서는 본 연구의 핵심 내용인 응답률이 선형이며 초모집단 모형에서 오차 분포가 정규분포인 경우에 무응답으로 인해 발생한 편향의 크기와 이때 얻어진 표본 분포를 연구하였다. 4절에는 모의실험을 통하여 본 연구에서 제안한 편향 보정 추정량과 흔히 층화추출법에서 사용하는 추정량의 성능을 비교하였다. 5절에는 본 연구의 내용을 적용하여 실제자료를 분석하였으며 6절에 결론이 있다.

2. 정보적 표본설계 기법을 이용한 지수형 응답률 편향 추정

2.1. 정보적 표본설계 개요

정보적 표본설계는 표본 추출과정이 관심변수 자료 값에 영향을 받고 관심변수와 보조변수 간에 초모집단 모형이 존재하는 표본설계로 관련 내용은 Savitsky와 Toth (2016)를 살펴보기 바란다. Pfeffermann 등 (1998)은 정보적 표본설계 하에서 θ^*를 θ의 함수라 할 때

$$f_s(y_i|\theta^*, x_i) = f_s(y_i|i \in s, x_i) = \frac{\Pr(i \in s|y_i, x_i) f_p(y_i|\theta, x_i)}{\Pr(i \in s|x_i)}$$

이고 $\Pr(i \in s|y_i, x_i) = E_p(\pi_i|y_i, x_i)$, $\Pr(i \in s|x_i) = E_p(\pi_i|x_i)$가 되어 다음의 관계가 성립되는 것을 밝혔다.

$$f_s(y_i|x_i) = \frac{E_p(\pi_i|y_i, x_i) f_p(y_i|x_i)}{E_p(\pi_i|x_i)}, \tag{2.1}$$

여기서 $f_p(y_i|x_i)$는 모집단 분포, $f_s(y_i|x_i)$는 표본 분포이고 $E_p(\pi_i|y_i, x_i)$는 x_i, y_i가 주어졌을 때 포함확률 π_i의 조건부 기댓값이다. 만약 식 (2.1)에서 $E_p(\pi_i|y_i, x_i) = E_p(\pi_i|x_i)$이면 모집단 분포와 표본 분포는 일치한다.

흔히 초모집단 모형이 단순 회귀 모형이고 응답률이 지수형인 경우에는 흔히 다음의 식 (2.2)와 (2.3)을 사용한다.

$$f_p(y_i|x_i) = N\left(\beta_0 + \beta_1 x_i, \sigma^2\right), \tag{2.2}$$

$$E_p(\pi_i|y_i, x_i) = \exp(a_0 + a_1 y_i), \tag{2.3}$$

여기서 $f_p(y_i|x_i)$는 초모집단 분포로 초모집단 모형을 따른다. 이제 식 (2.2)와 (2.3)을 식 (2.1)에 대입하면 다음의 모형을 갖는 표본 분포가 얻어진다.

$$f_s(y_i|x_i) = N\left(\beta_0 + a_1\sigma^2 + \beta_1 x_i, \sigma^2\right). \tag{2.4}$$

따라서 식 (2.2)와 (2.4)를 비교함으로써 편향이 $a_1\sigma^2$가 되는 것을 확인할 수 있다. 이와 관련된 내용은 Chung과 Shin (2017), Min과 Shin (2018), 그리고 Pfeffermann 등 (1998, 2006)을 참고하기 바란다.

2.2. 응답률 모형의 모수 추정

식 (2.4)에서 계산된 편향의 크기는 $a_1\sigma^2$이다. 정보적 표본설계에서는 설계 단계에서 알려진 a_0, a_1을 사용하기 때문에 표본 조사에서 얻어진 관심변수와 보조변수의 회귀모형에서 σ^2을 추정하면 편향의 크기를 계산할 수 있다. 그러나 본 연구에서 사용한 응답률 모형의 경우에서는 a_1을 추정해야 한다. 이

를 위해 Chung과 Shin (2017)은 주어진 하나의 특정 층을 세부 층으로 나누는 방법을 이용하였다. 즉, Pfeffermann과 Sverchkov (2003)에서 얻어진 결과인 $E_s(w_i|y_i,x_i) = 1/E_p(\pi_i|y_i,x_i)$와 $E_s(w_i|y_i) \approx w_i$를 적용하고 나누어진 세부 층에 의해 구해진 세부 층별 가중치 w_i를 사용하여 모형을 설정하였다.

$$\log\left(\frac{1}{w_i}\right) = a_0 + a_1 y_i + \eta_i. \tag{2.5}$$

이 논문에서는 식 (2.5)에서 최소제곱추정법을 이용하여 a_1을 추정하였으며 이때 η_i는 독립이고 $E(\eta_i) = 0$, $\mathrm{Var}(\eta_i) = \sigma_\eta^2$을 가정하였다.

3. 선형 응답률 모형의 편향 추정과 제안된 추정량

3.1. 표본 분포와 편향 추정

본 연구에서는 초모집단 모형의 오차가 정규분포를 따르는 경우를 고려하기 때문에 모집단 분포는 식 (2.2)를 사용한다. 그러나 응답률은 식 (3.1)의 선형 응답률 모형을 고려한다.

$$E_p(\pi_i|y_i,x_i) = b_0 + b_1 y_i. \tag{3.1}$$

따라서 선형 응답률 모형에 의해 다음의 결과가 얻어진다.

$$E_p(\pi_i|x_i) = E(E_p(\pi_i|y_i,x_i)) = E(b_0 + b_1 y_i|x_i) = b_0 + b_1 E_p(y_i|x_i). \tag{3.2}$$

이제 식 (3.1)과 (3.2)를 식 (2.1)에 대입하면 다음의 결과가 얻어진다.

$$f_s(y_i|x_i) = \frac{b_0}{b_0 + b_1 E_p(y_i|x_i)} f_p(y_i|x_i) + \frac{b_1}{b_0 + b_1 E_p(y_i|x_i)} y_i f_p(y_i|x_i). \tag{3.3}$$

다음으로 $f_p^*(y_i|x_i)$를 $y_i f_p(y_i|x_i)$의 분포라 하면 식 (3.3)은 다음의 형태가 된다.

$$f_s(y_i|x_i) = \frac{b_0}{b_0 + b_1\mu_i} f_p(y_i|x_i) + \frac{b_1\mu_i}{b_0 + b_1\mu_i} f_p^*(y_i|x_i),$$

여기서 $\mu_i = E_p(y_i|x_i) = \beta_0 + \beta_1 x_i$이다. 따라서 표본 분포 $f_s(y_i|x_i)$는 모집단 분포 $f_p(y_i|x_i)$와 $f_p^*(y_i|x_i)$의 선형결합 형태가 된다. 이제 표본 분포를 이용하여 기댓값을 구하면 다음 결과를 얻는다.

$$E_s(y_i|x_i) = \frac{b_0}{b_0 + b_1\mu_i}\mu_i + \frac{b_1\mu_i}{b_0 + b_1\mu_i}\frac{1}{\mu_i}(\mu_i^2 + \sigma_i^2) = \mu_i + \frac{b_1\sigma^2}{b_0 + b_1\mu_i}.$$

결국 계산된 편향의 크기는 $(b_1\sigma^2)/(b_0+b_1\mu_i)$가 된다. 이제 표본 자료에서 얻어진 기댓값을 $E_s(y_i|x_i) = \mu_i^{(s)} = \hat{\beta}_0 + \hat{\beta}_1 x_i$라 하고 계산을 간단하게 하기 위하여 $(b_1\sigma^2)/(b_0 + b_1\mu_i) \approx (b_1\sigma^2)/(b_0 + b_1\mu_i^{(s)})$를 사용하면 편향이 보정된 기댓값은 다음과 같이 된다.

$$E_p(y_i|x_i) = \mu_i = \mu_i^{(s)} - \frac{b_1\sigma^2}{b_0 + b_1\mu_i^{(s)}}. \tag{3.4}$$

3.2. 선형 응답률 모형의 모수 추정

2.2절에서 설명한 바와 같이 선형 응답률 모형에서도 모형에 포함된 모수를 추정하여야 한다. 이때 지수형 응답률 모수 추정에서 사용한 동일한 이론과 가정을 사용하면 모형 (3.5)가 얻어진다.

$$\frac{1}{w_i} = b_0 + b_1 y_i + \eta_i. \tag{3.5}$$

결국 세부 층에서 얻어진 가중치 w_i와 자료 y_i를 이용하여 b_0, b_1을 추정하게 된다. 물론 η_i에 주어진 가정과 추정법은 2.2절과 동일하다.

3.3. 모평균 추정을 위해 제안된 편향 보정 추정량

본 연구에서는 기존의 연구에서 사용한 세 개의 추정량을 비교하였다. 즉 주어진 층의 모평균 추정량으로 층 내 가중치가 같다고 가정한 단순 평균 추정량과 세부 층의 가중치를 사용한 층화추출 추정량이 사용될 수 있다. 또한 본 연구에서 제안한 편향 보정 추정량으로 식 (3.4)에 기초한 추정량을 사용할 수 있다. 결론적으로 식 (3.6)–(3.8)이 본 모의실험에서 사용된 추정량이다. 여기서 L은 세부 층의 층 개수이고 $n_h, h = 1, \ldots, L$은 h 세부 층의 자료 수, w_h는 h 세부 층의 가중치이다. 또한 y_{hi}는 h 세부 층의 i번째 자료이다.

(1) 단순 평균 추정량
주어진 층의 가중치가 세부 층에 무관하게 일정하기 때문에 $w_h = w = N/n$이 되어 다음의 수식이 얻어진다.

$$\hat{\bar{Y}}_s = \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{n_h} w y_{hi} = \frac{1}{n}\sum_{h=1}^{L}\sum_{i=1}^{n_h} y_{hi} = \bar{y}. \tag{3.6}$$

(2) 층화 추출 추정량
세부 층의 가중치가 다르기 때문에 다음의 평균 추정량을 사용한다.

$$\hat{\bar{Y}}_{st} = \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{n_h} w_h y_{hi}. \tag{3.7}$$

(3) 제안된 편향 보정 추정량
표본 자료에서 얻어진 기댓값 $\mu_i^{(s)}$와 식 (3.5)의 응답률 모형에서 얻어진 $\hat{b}_0, \hat{b}_1$을 이용하여 다음 식 (3.8)의 추정량을 사용한다.

$$\hat{\bar{Y}}_{\text{inf}}^{L} = \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{n_h} w_h \left(\mu_i^{(s)} - \frac{\hat{b}_1\hat{\sigma}^2}{\hat{b}_0 + \hat{b}_1\mu_i^{(s)}}\right), \tag{3.8}$$

여기서 $\hat{\sigma}^2$은 초모집단 모형을 기초로 하여 회귀분석을 한 후 얻어진 분산 추정값이다.

4. 모의실험 및 실제 자료 분석

4.1. 모의실험 설계

본 모의실험에서는 층화추출법을 사용하게 되면 다수의 층이 존재하지만 여러 개의 층중에서 주어진 한 개의 특정 층의 추정을 고려하였다. 이는 층화추출법에서는 각 층별로 모수추정이 이루어지기 때문에 하나의 층을 고려하여도 일반성을 잃지 않기 때문이다. 다음이 모의실험을 위한 자료생성 과정과 모수추정 방법이다. 전체적인 모의실험 방법은 Chung과 Shin (2017)과 Min과 Shin (2018)에서 사용한 방법을 사용하였다.

- Step 1: 모집단 생성과정
초모집단 모형이 회귀모형이고 모형의 오차가 정규분포인 경우의 정보적 표본설계를 위한 모집단 자료생성 과정은 다음과 같다.

(1) 보조변수 x_i 생성: $x_i = 100 + \gamma_i$, $i = 1, \ldots, N$
여기서 $\gamma_i \overset{\text{iid}}{\sim} \text{Unif}(0, 100)$과 TGamma$(1, 100)$을 사용하고 TGamma$(1, 100)$은 절단 감마분포(truncated gamma distribution)로 0과 100 사이의 값을 갖기 위해 100 이상인 값은 버린다. 따라서 보조변수 x_i는 100에서 200 사이의 값을 갖는다.

(2) 초모집단 모형: $y_i = \beta_0 + \beta_1 x_i + \epsilon$, $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$
여기서 $\beta_0 = 10$, $\beta_1 = 5$, $\sigma^2 = 400$과 모집단 자료 수 $N = 5{,}000$을 사용하며 모집단의 영향을 살펴보기 위해 추가로 $N = 200, 1000, 30000$과 $N = 1000, 10000, 50000$을 살펴보았다.

- Step 2: 표본추출과정

생성된 모집단에서 N보다 작은 n개의 표본을 추출한다. 추출된 자료에서 랜덤으로 무응답을 만든다.

(3) N개의 모집단 자료에서 단순임의추출(simple random sample)로 n개의 표본을 추출한다. 이때 $n = 50, 100, 150, 200, 250, 300, 400, 500$을 사용한다.

(4) 추출된 n개의 표본에서 $\pi_i = b_0 + b_1 y_i$, $\pi_i \in [0, 1]$를 계산한다. y_i의 최솟값에서의 응답률을 $\pi_y^{\min}$, y_i의 최댓값에서의 응답률을 $\pi_y^{\max}$라 할 때, $(\pi_y^{\min}, \pi_y^{\max}) = (0.9, 0.7)$ 그리고 (0.7, 0.9)를 사용하여 b_0, b_1을 구하고 y_i에 따라 응답률을 계산한다. 또한 y_i의 응답률이 모두 같은 $\pi_i = 1$인 경우도 고려한다. 이와 관련된 내용은 Chung과 Shin (2017)을 살펴보기 바란다.

(5) 응답한 최종 조사 자료 수는 r개이다. 여기서 $(\pi_y^{\min}, \pi_y^{\max}) = (0.9, 0.7)$ 또는 $(\pi_y^{\min}, \pi_y^{\max}) = (0.7, 0.9)$인 경우는 전체 자료의 약 80%가 응답하게 되어 주어진 자료 수 n에 비해 약 20%가 감소한다.

- Step 3: 층화

얻어진 표본 자료는 (x_i, y_i), $i = 1, \ldots, r$이고 무응답에 의해 각 자료의 가중치는 달라진다. 이를 반영하기 위해 주어진 하나의 모집단 층을 L개의 세부 층으로 나눈다. 실제 자료 분석에서는 모집단에 보조변수 x_i의 정보만 있으므로 보조변수를 기준으로 층을 나눈다.

(6) 보조변수 x_i를 기준으로 분위수를 이용하여 모집단을 L개의 세부 층으로 나눈다. 여기서 $L = 4$에서 100까지 다양한 세부 층 개수를 적용한다.

- Step 4: 모수추정

(7) 나누어진 세부 층의 모집단 수와 조사된 자료 수 (N_h, r_h)를 이용하여 세부 층 가중치 $w_h = N_h / r_h$를 계산한다. 이때 $w_i = w_{(i \in h)} = w_h$가 된다. 즉 세부 층에 포함된 자료의 가중치는 동일하다.

(8) $1/w_i = b_0 + b_1 y_i + \eta_i$를 설정하고 단순 회귀모형을 이용하여 모수 b_0, b_1을 추정한다.

(9) 추출된 자료 (y_i, x_i)를 이용해서 단순 회귀분석을 실시하고 $\beta_0, \beta_1, \sigma^2$을 추정한다.

(10) 계산된 결과를 이용하여 식 (3.6)–(3.8)인 $\hat{\bar{Y}}_s, \hat{\bar{Y}}_{st}, \hat{\bar{Y}}_{\text{inf}}^L$을 계산한다.

이제 얻어진 평균 추정값은 다음의 비교통계량, 편향(bias), 절대편향(absolute bias; Abias), 그리고 제곱근 MSE(root mean squared error; RMSE)를 이용하여 결과의 성능이 비교되었다.

$$\text{Bias} = \frac{1}{R} \sum_{r=1}^{R} \left(\hat{\bar{Y}}_r - \bar{Y}_r \right),$$

$$\text{Abias} = \frac{1}{R} \sum_{r=1}^{R} \left| \hat{\bar{Y}}_r - \bar{Y}_r \right|,$$

$$\text{RMSE} = \sqrt{\frac{1}{R} \sum_{r=1}^{R} \left(\hat{\bar{Y}}_r - \bar{Y}_r \right)^2}.$$

Table 4.1. Comparison results of $U(0, 100)$ with $n = 50$

$\pi_y^{\min}$	$\pi_y^{\max}$	r	L	Bias			Abias			RMSE		
				$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{\text{inf}}^L$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{\text{inf}}^L$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{\text{inf}}^L$
0.9	0.7	40	4	−8.178	−0.804	−0.661	19.309	5.645	5.603	24.256	8.059	8.019
			5	−8.178	−0.730	−0.608	19.309	5.091	5.012	24.256	9.337	9.283
			6	−8.178	−0.984	−0.880	19.309	5.083	4.975	24.256	12.282	12.211
1.0	1.0	50	4	0.167	−0.043	−0.038	16.114	4.870	4.858	20.300	6.089	6.084
			5	0.167	−0.004	−0.017	16.114	4.263	4.207	20.300	6.610	6.558
			6	0.167	0.139	0.118	16.114	3.804	3.721	20.300	5.428	5.323
0.7	0.9	40	4	8.804	0.541	0.382	19.467	5.620	5.594	24.358	8.010	7.981
			5	8.804	0.452	0.276	19.467	5.012	4.929	24.358	8.217	8.136
			6	8.804	0.012	−0.164	19.467	4.826	4.723	24.358	9.774	9.715

Abias = absolute bias; RMSE = root mean squared error.

Table 4.2. Comparison results of $U(0, 100)$ with $n = 250$

$\pi_y^{\min}$	$\pi_y^{\max}$	r	L	Bias			Abias			RMSE		
				$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{\text{inf}}^L$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{\text{inf}}^L$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{\text{inf}}^L$
0.9	0.7	200	18	−8.679	−0.191	−0.033	11.025	1.258	1.203	13.574	1.588	1.514
			20	−8.679	−0.228	−0.070	11.025	1.282	1.222	13.574	1.862	1.794
			21	−8.679	−0.245	−0.090	11.025	1.317	1.249	13.574	2.242	2.177
1	1	250	18	−0.119	−0.007	−0.001	7.381	1.099	1.066	9.236	1.391	1.348
			20	−0.119	−0.017	−0.007	7.381	1.089	1.061	9.236	1.365	1.326
			21	−0.119	−0.008	0.004	7.381	1.087	1.051	9.236	1.369	1.323
0.7	0.9	200	18	8.394	0.177	0.043	10.952	1.250	1.197	13.451	1.581	1.510
			20	8.394	0.141	0.009	10.952	1.264	1.208	13.451	1.753	1.683
			21	8.394	0.136	−0.001	10.952	1.309	1.238	13.451	2.026	1.964

Abias = absolute bias; RMSE = root mean squared error.

4.2. 모의실험 결과

보조변수의 분포가 균일분포 및 절단 감마분포인 경우의 모의실험이 수행되었다. n = 50, 100, 150, 200, 250, 300, 400, 500을 이용하여 모의실험을 수행하였으나 결과의 특징이 매우 유사하여 이 중에서 n = 50, 250, 500 결과만을 수록하였다.

4.2.1. 보조변수가 균일분포일 경우

표본 수가 50인 경우의 결과인 Table 4.1을 살펴보면 층내 가중치를 동일하게 사용하는 $\hat{\bar{Y}}_s$에 비해 층화추출 추정량을 사용한 $\hat{\bar{Y}}_{st}$의 편향이 크게 줄어든 것을 확인할 수 있다. 또한 절대 편향과 RMSE도 매우 작아진 것을 확인할 수 있다. 또한 $\hat{\bar{Y}}_{\text{inf}}^L$은 다른 두 추정량 $\hat{\bar{Y}}_s, \hat{\bar{Y}}_{st}$에 비해 모든 비교통계량을 기준으로 비교하였을 때 가장 우수한 결과를 준다. 이러한 경향은 표본의 크기가 커진 결과인 Table 4.2와 Table 4.3에서 모두 나타나고 있다. 결론적으로 정보적 표본설계 기법을 이용한 편향 보정 방법이 매우 우수한 결과를 주고 있음을 확인할 수 있다.

4.2.2. 보조변수가 절단 감마분포일 경우 보조변수가 절단 감마분포인 경우의 결과가 Table 4.4에서 Table 4.6에 수록되어 있다. 결과를 살펴보면 4.2.1절의 균일분포 결과와 유사하게 모든 비교통계량 결과에서 $\hat{\bar{Y}}_{\text{inf}}^L$가 가장 우수한 것을 확인할 수 있다.

Table 4.3. Comparison results of $U(0, 100)$ with $n = 500$

$\pi_y^{\min}$	$\pi_y^{\max}$	r	L	Bias			Abias			RMSE		
				$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{\text{inf}}^L$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{\text{inf}}^L$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{\text{inf}}^L$
0.9	0.7	400	28	−8.485	−0.147	−0.006	9.322	0.848	0.811	11.065	1.059	1.013
			30	−8.485	−0.150	−0.006	9.322	0.849	0.817	11.065	1.057	1.013
			40	−8.485	−0.165	−0.033	9.322	0.863	0.820	11.065	1.305	1.250
1.0	1.0	500	28	0.066	0.014	0.009	5.061	0.733	0.714	6.303	0.913	0.892
			30	0.066	0.009	0.002	5.061	0.732	0.716	6.303	0.912	0.892
			40	0.066	0.017	0.008	5.061	0.731	0.700	6.303	0.908	0.870
0.7	0.9	400	28	8.609	0.192	0.033	9.380	0.857	0.805	11.194	1.067	1.011
			30	8.609	0.180	0.027	9.380	0.852	0.805	11.194	1.059	1.007
			40	8.609	0.158	0.002	9.380	0.867	0.810	11.194	1.262	1.208

Abias = absolute bias; RMSE = root mean squared error.

Table 4.4. Comparison results of $TGamma(1, 100)$ with $n = 50$

$\pi_y^{\min}$	$\pi_y^{\max}$	r	L	Bias			Abias			RMSE		
				$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{\text{inf}}^L$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{\text{inf}}^L$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{\text{inf}}^L$
0.9	0.7	40	4	−7.654	−0.599	−0.449	19.020	5.722	5.709	23.688	7.940	7.934
			5	−7.654	−0.626	−0.466	19.020	4.996	4.985	23.688	7.985	7.963
			6	−7.654	−1.121	−0.953	19.020	5.096	5.021	23.688	11.353	11.282
1.0	1.0	50	4	0.258	0.166	0.172	16.402	5.048	5.064	20.476	7.205	7.208
			5	0.258	0.066	0.064	16.402	4.291	4.284	20.476	5.340	5.317
			6	0.258	−0.069	−0.075	16.402	3.907	3.847	20.476	6.367	6.315
0.7	0.9	40	4	8.131	0.828	0.699	19.614	5.744	5.719	24.498	7.939	7.930
			5	8.131	0.286	0.167	19.614	5.024	5.002	24.498	8.522	8.487
			6	8.131	−0.497	−0.626	19.614	5.241	5.151	24.498	12.161	12.094

Abias = absolute bias; RMSE = root mean squared error.

Table 4.5. Comparison results of $TGamma(1, 100)$ with $n = 250$

$\pi_y^{\min}$	$\pi_y^{\max}$	r	L	Bias			Abias			RMSE		
				$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{\text{inf}}^L$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{\text{inf}}^L$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{\text{inf}}^L$
0.9	0.7	200	15	−7.846	−0.180	−0.035	10.255	1.312	1.275	12.649	1.650	1.594
			16	−7.846	−0.179	−0.036	10.255	1.273	1.231	12.649	1.602	1.541
			18	−7.846	−0.158	−0.020	10.255	1.264	1.205	12.649	1.582	1.512
1.0	1.0	250	15	−0.094	0.009	0.014	7.202	1.166	1.139	9.025	1.453	1.422
			16	−0.094	−0.009	−0.005	7.202	1.149	1.119	9.025	1.432	1.395
			18	−0.094	0.001	−0.002	7.202	1.128	1.093	9.025	1.405	1.370
0.7	0.9	200	15	8.007	0.204	0.060	10.600	1.313	1.269	13.019	1.652	1.594
			16	8.007	0.179	0.032	10.600	1.296	1.243	13.019	1.723	1.662
			18	8.007	0.181	0.028	10.600	1.296	1.234	13.019	1.847	1.780

Abias = absolute bias; RMSE = root mean squared error.

4.2.3. 모집단 자료 수가 다른 경우 Min과 Shin (2018)은 얻어진 층의 최종 자료 수가 최적 세부 층 개수에 영향을 주는 것을 보였다. 본 논문에서는 층 내 모집단 수가 최적 세부 층 개수에 영향을 주는지 모의실험을 통해 살펴보았다. Table 4.7에서 Table 4.10에 결과가 수록되었다. Table 4.7과 Table 4.8의 결과는 보조변수가 정규분포인 경우의 결과이며 Table 4.7은 표본 수가 50인 경우의 결과이다.

Table 4.6. Comparison results of $TGamma(1, 100)$ with $n = 500$

$\pi_y^{\min}$	$\pi_y^{\max}$	r	L	Bias			Abias			RMSE		
				$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{\text{inf}}^L$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{\text{inf}}^L$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{\text{inf}}^L$
0.9	0.7	400	28	−7.803	−0.179	−0.046	8.617	0.856	0.820	10.386	1.076	1.035
			30	−7.803	−0.170	−0.037	8.617	0.851	0.818	10.386	1.070	1.025
			40	−7.803	−0.216	−0.089	8.617	0.903	0.845	10.386	1.383	1.326
1.0	1.0	500	28	−0.090	−0.022	−0.025	5.005	0.760	0.747	6.249	0.950	0.935
			30	−0.090	−0.019	−0.022	5.005	0.755	0.740	6.249	0.945	0.928
			40	−0.090	−0.042	−0.047	5.005	0.770	0.741	6.249	1.027	0.993
0.7	0.9	400	28	7.946	0.147	−0.004	8.977	0.868	0.839	10.722	1.084	1.048
			30	7.946	0.157	0.004	8.977	0.862	0.829	10.722	1.081	1.039
			40	7.946	0.106	−0.046	8.977	0.908	0.852	10.722	1.309	1.261

Abias = absolute bias; RMSE = root mean squared error.

Table 4.7. Comparison results of $U(1, 100)$ with $n = 50$

$\pi_y^{\min}$	$\pi_y^{\max}$	r	N	RMSE	L
0.9	0.7	40	200	6.167	4
			1000	6.710	4
			30000	7.029	4
1.0	1.0	50	200	4.497	5
			1000	4.955	5
			30000	5.131	5
0.7	0.9	40	200	5.687	5
			1000	6.715	4
			30000	6.872	4

RMSE = root mean squared error.

Table 4.8. Comparison results of $U(1, 100)$ with $n = 300$

$\pi_y^{\min}$	$\pi_y^{\max}$	r	N	RMSE	L
0.9	0.7	240	1000	1.216	20
			10000	1.375	20
			50000	1.401	21
1.0	1.0	300	1000	1.027	30
			10000	1.185	27
			50000	1.218	24
0.7	0.9	240	1000	1.217	20
			10000	1.362	20
			50000	1.395	21

RMSE = root mean squared error.

결과를 살펴보면 모집단이 200에서 30,000으로 매우 크게 증가하였음에도 세부 층의 수에는 큰 영향을 주지 않는 것을 확인할 수 있다. 이러한 현상은 표본 수가 300인 경우에도 유사한 결과를 준다. 따라서 모집단 수에 비해 표본 수가 최적 세부 층 결정에 큰 영향을 주는 것을 확인할 수 있다. Table 4.9와 Table 4.10의 결과는 보조변수가 절단 감마분포인 경우로 균일분포처럼 모집단 수는 최적 세부 층 개수에 크게 영향을 주지 않는 것을 확인할 수 있다. 다만 균일분포 보다는 모집단 수에 따라 약간의 차이를 보이고 있다.

Table 4.9. Comparison results of TGamma(1, 100) with $n = 50$

$\pi_y^{\min}$	$\pi_y^{\max}$	r	N	RMSE	L
0.9	0.7	240	200	6.388	4
			1000	7.059	4
			30000	7.098	4
1.0	1.0	300	200	4.172	6
			1000	5.232	5
			30000	5.398	5
0.7	0.9	240	200	6.250	4
			1000	6.969	4
			30000	6.566	5

RMSE = root mean squared error.

Table 4.10. Comparison results of TGamma(1, 100) with $n = 300$

$\pi_y^{\min}$	$\pi_y^{\max}$	r	N	RMSE	L
0.9	0.7	240	1000	1.208	25
			10000	1.384	18
			50000	1.372	21
1.0	1.0	300	1000	1.026	270
			10000	1.227	24
			50000	1.204	25
0.7	0.9	240	1000	1.250	24
			10000	1.400	20
			50000	1.389	21

RMSE = root mean squared error.

Table 4.11. Optimal number of substrata with $N = 5{,}000$

Distribution	r								
	40	50	80	100	160	200	240	300	400
$U(0, 100)$	4	6	9	9	12	18	20	21	30
TGamma(1, 100)	5	5	9	9	12	18	21	28	30
Average number of substrata	4.5	5.5	9	9	12	18	20.5	24.5	30

Table 4.12. Optimal number of substrata with $N = 10{,}000$

Distribution	r								
	40	50	80	100	160	200	240	300	400
$U(0, 100)$	4	6	9	10	12	15	20	27	28
TGamma(1, 100)	4	5	8	10	12	18	18	24	30
Average number of substrata	4	5.5	8.5	10	12	16.5	19	25.5	29

4.3. 최적 표본 수

Table 4.11에서 Table 4.13은 최적 세부 층 개수를 모집단 수와 최종 조사 자료 수에 따라 정리한 결과이다. 전체적으로 모집단 크기에 영향을 크게 받지 않으며 조사 자료 수가 100 이하인 경우 각각의 세부 층에 10개 정도 표본이 배분되도록 세부 층 개수를 정하면 되며 r이 100보다 큰 경우에는 11–13개 정도의 표본이 배분되도록 세부 층의 개수를 정하면 실무에서 본 연구 결과를 사용하는데 무리가 없다고 판단된다.

Table 4.13. Optimal number of substrata with $N = 50{,}000$

Distribution	r								
	40	50	80	100	160	200	240	300	400
$U(0, 100)$	4	5	9	10	12	20	20	27	28
TGamma$(1, 100)$	4	5	8	10	10	18	20	27	30
Average number of substrata	4	5	8.5	10	11	19	20	27	29

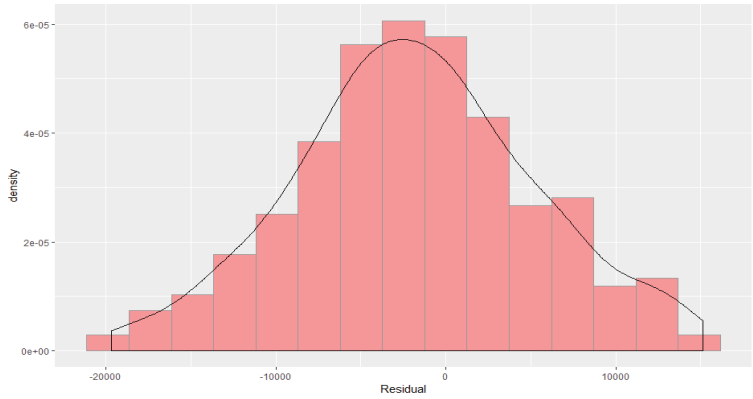

Figure 5.1. Distribution of residual.

Table 5.1. Comparison results of Data and $(\pi_y^{\min}, \pi_y^{\max}) = (0.8, 0.4)$

n	L	Bias			Abias			RMSE		
		$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{\text{inf}}^L$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{\text{inf}}^L$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{\text{inf}}^L$
100	5	−983.9	−772.5	−639.3	1071.3	900.0	846.8	1258.8	1076.9	1020.9
	10	−983.9	−768.5	−637.7	1071.3	920.6	845.6	1258.8	1094.9	1023.7
150	5	−1013.1	−789.8	−645.2	1046.8	844.1	749.5	1193.4	986.4	895.1
	10	−1013.1	−772.7	−630.4	1046.8	838.1	742.5	1193.4	983.1	886.7

Abias = absolute bias; RMSE = root mean squared error.

5. 실제 자료 분석

실제 자료 분석에 사용된 자료는 2015년도 외감 기업 자료 중에서 V6, V7 자료가 사용되었으며 전체 자료 중에서 이상점을 제거한 272개의 자료가 분석에 사용되었다. 이 자료는 한국은행에서 발간하는 국민계정리뷰 (2017, 3호)의 무응답 대체법에 관한 연구를 위해 사용되었던 자료의 일부이다. 다음은 초모집단 모형의 회귀분석 결과이다. 회귀분석 결과 $\hat{\beta}_0 = 211.057, \hat{\beta}_1 = 0.051, \hat{\sigma} = 6866.159, R^2 = 0.32$가 얻어졌으며 잔차의 정규성 검정결과 Shapiro-Wilk 통계량 W의 p-value는 0.4275, Kolmogorov-Smirnow 통계량 D의 p-value는 0.1500보다 큰 것으로 나타났다. 또한 Cramer-von Mises 통계량 $W - Sq$의 p-value와 Anderson-Darling 통계량 $A - Sq$의 p-value는 0.2500보다 큰 값으로 나타났다.

주어진 모집단 자료 중에서 $n = 100, 150$개의 표본을 추출하였으며 $L = 5, 10$을 각각 사용하였다. 실제 표본조사에서는 표본 층의 경우 3배수 이상의 예비표본을 사용하거나 전수 층의 경우 조사에 따르지만 약 50%정도 응답률을 보이는 경우가 많이 있다. 이에 본 논문에서는 응답률이 약 60%가 되도록 결정하였다. 또한 무응답은 모의실험에서 사용한 선형 응답률 모형을 이용하여 생성하였다. 자료분석 결과는 Table 5.1과 Table 5.2에 수록되어 있다.

Table 5.2. Comparison results of Data and $(\pi_y^{\min}, \pi_y^{\max}) = (0.4, 0.8)$

n	L	Bias			Abias			RMSE		
		$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{\text{inf}}^L$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{\text{inf}}^L$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{\text{inf}}^L$
100	5	1366.1	938.9	685.3	1496.8	1096.0	972.9	1778.3	1333.2	1230.2
	10	1366.1	881.9	630.9	1496.8	1104.4	975.6	1778.3	1359.3	1227.1
150	5	1351.1	911.9	626.6	1388.5	976.7	811.7	1592.7	1151.1	1002.9
	10	1351.1	900.3	607.9	1388.5	977.8	805.1	1592.7	1171.5	1006.3

Abias = absolute bias; RMSE = root mean squared error.

결과를 살펴보면 $L = 5, 10$인 두 경우 모두 결과에 큰 차이를 보이지 않고 있다. 다만 모의실험 결과와 유사하게 편향 보정 추정량이 모든 비교통계량을 기준으로 매우 우수한 결과를 주고 있다. 특히 편향을 살펴보면 편향의 크기가 크게 줄어든 것을 확인할 수 있다.

6. 결론

응답률이 관심변수의 함수인 경우에는 함수관계 정보를 기반으로 무응답 편향의 크기를 파악할 수 있다. 응답률 모형에 사용될 수 있는 함수는 다양하지만 초모집단 모형의 오차 분포가 정규분포인 경우에는 지수형과 선형 모형이 흔히 사용된다. 본 연구에서는 선형 응답률 모형을 이용하여 무응답으로 인해 발생된 편향의 크기를 파악하였으며 파악된 편향을 모집단 평균 추정 시에 반영함으로써 편향이 제거된 우수한 추정 결과를 얻을 수 있었다. 또한 모의실험을 통하여 최적의 세부 층 개수와 세부 층 내 표본 수를 살펴보았다. 전체적으로 최적의 세부 층 개수와 세부 층 내 표본 수는 모집단 수에는 크게 영향을 받지 않는 것으로 나타났으나 표본 수에는 영향을 받는 것으로 확인되었다.

References

Chung, H. Y. and Shin, K. I. (2017), Estimation using informative sampling technique when response rate follows exponential function of variable of interest, *Korean Journal of Applied Statistics*, **30**, 993–1004.

Min, J. W. and Shin, K. I. (2018). A study on the determination of substrata using the information of exponential response rate by simulation studies, *Korean Journal of Applied Statistics*, **31**, 621–636.

Pfeffermann, D., Krieger, A. M., and Rinott, Y. (1998). Parametric distributions of complex survey data under informative probability sampling, *Statistica Sinica*, **8**, 1087–1114.

Pfeffermann, D., Moura, F. A. D. S., and Silva, P. L. D. N. (2006). Multi-level modelling under informative sampling, *Biometrika*, **93**, 943–959.

Pfeffermann, D. and Sverchkov, M. (2003), Small area estimation under informative sampling, *2003 Joint Statistical Meeting-Section on Survey Research Methods*, 3284–3295.

Savitsky, T. D. and Toth, D. (2016). Bayesian estimation under informative sampling, *Electronic Journal of Statistics*, **30**, 1677–1708.

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

응답률이 선형인 표본조사에서 편향 보정 추정

정희영[a] · 신기일[a,1]

[a]한국외국어대학교 통계학과

(Received May 23, 2019; Revised June 25, 2019; Accepted June 25, 2019)

요 약

다수의 항목무응답이 발생한 표본조사에서는 추정의 정확성이 떨어진다. 이를 해결하기 위한 많은 방법이 개발되었으나 응답률이 관심변수에 의해 영향을 받는 경우임에도 이를 고려하지 않고 랜덤으로 무응답이 발생한다는 가정 하에서 사용하는 무응답 처리 방법을 사용하게 되면 편향이 발생하는 것으로 알려져 있다. Chung과 Shin (2017)과 Min과 Shin (2018)은 응답률이 관심변수의 함수인 경우에서 발생된 편향을 적절히 처리하여 추정의 정확성을 향상시키는 방법을 제안하였다. 본 연구에서는 응답률 함수가 선형(linear)이면서 초모집단 모형의 오차가 정규분포를 따르는 경우를 살펴보았으며 층별 모집단 수가 편향 보정에 영향을 주는지도 살펴보았다. 모의실험을 통하여 제안된 추정량의 성능을 살펴보았으며 실제 자료 분석을 통해 이를 확인하였다.

주요용어: 선형 표본 포함확률, 표본 분포, 회귀모형, 표본 가중치

이 논문은 2018년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NRF-2018R1D1A1B07042736).

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