The Korean Journal of Applied Statistics
2021, Vol. 34, No. 6, 923–936

DOI: https://doi.org/10.5351/KJAS.2021.34.6.923
Print ISSN1225-066X / Online ISSN 2383-5818

Bias corrected non-response estimation using nonparametric function estimation of super population model

Joo-Yong Sim[a], Key-Il Shin[1,a]

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

Abstract

A large number of non-responses are occurring in the sample survey, and various methods have been developed to deal with them appropriately. In particular, the bias caused by non-ignorable non-response greatly reduces the accuracy of estimation and makes non-response processing difficult. Recently, Chung and Shin (2017, 2020) proposed an estimator that improves the accuracy of estimation using parametric super-population model and response rate model. In this study, we suggested a bias corrected non-response mean estimator using a non-parametric function generalizing the form of a parametric super-population model. We confirmed the superiority of the proposed estimator through simulation studies.

Keywords: propensity score, sample distribution, population distribution, local linear regression

1. 서론

표본조사에서 발생한 무응답은 추정의 정확성을 떨어뜨리는 주요 요인이다. 특히 최근 조사에서 무응답 발생 확률은 높아지고 있으므로 적절한 무응답 처리는 매우 중요하다. 대표적으로 사용되는 무응답 처리 방법에는 가중치를 보정 또는 수정한 후 얻어진 최종 가중치를 사용하여 모수를 추정하는 방법인 가중치 보정 방법과 무응답으로 인해 발생한 결측치(missing value)에 타당하다고 판단되는 값을 대체하는 방법인 무응답 대체법이 있다. 두 방법과 관련된 다수의 방법이 이미 개발되었고, 실제 표본조사에서 사용되고 있다.

대부분의 무응답 처리법은 missing at random (MAR)를 가정한 후 이루어진다. 그러나 최근 무응답이 관심변수에 영향을 받는 경우가 있으며 이 경우 MAR 가정을 사용한 무응답 처리 방법은 효과적이지 않다. 따라서 무응답이 관심변수에 영향을 받는 경우는 관심변수의 응답률 함수를 설정하고 이를 통해 만들어진 응답률 모형을 이용하여 처리하는 것이 타당하다. 또한, 다수의 표본조사에서 관심변수와 보조변수 간에 관계가 형성되며 이 관계는 초모집단 모형으로 설명될 수 있다. 따라서 관심변수와 보조변수 간에 초모집단 모형이 형성되고, 타당한 관심변수의 응답률 모형을 사용한다면 정확한 추정값이 얻어질 수 있다. 초모집단 모형은 알려진 모형, 예를 들면 선형 모형, 로그-선형 모형 등이 사용된다. 그러나 실제 자료 분석에서 초모집단 모형이 알려진 형태에서 벗어날 수 있으며 또한 형태를 파악하기 어려울 수 있다. 이러한 상황에서 비모수적 함수 추정 기법을 사용하는 것은 매우 자연스럽다.

최근 정보적 표본설계 기법을 응용하여 무응답으로 인해 발생한 편향을 처리하는 다수의 연구가 진행되었다. 먼저 정보적 표본설계 기법은 Pfeffermann 등 (1998) 이후 다수의 논문이 발표되었으며 Chung과 Shin

This research was supported by Hankuk University of Foreign Studies research fund (2021)
[1] Corresponding Author : Department of Statistics, Hankuk University of Foreign Studies, 81 Oedae-ro, Yongin-si, Gyeonggi-do 17035, Korea. E-mail: keyshin@hufs.ac.kr

Published 31 December 2021 / journal homepage: http://kjas.or.kr

(2017, 2019, 2020), Min과 Shin (2018) 등은 이 방법을 응답률 모형에 적용한 편향 보정 평균추정량을 제안하였다. 또한, Bethlehem (2020)은 응답률을 추정할 수 있는 다양한 방법을 설명하였다.

본 연구에서는 Chung과 Shin (2017, 2019)에서 제안한 응답률 추정값과 Bethlehem (2020)의 성향점수를 이용한 응답률 추정값을 사용하여 얻어진 편향 추정량에 비모수적 함수 추정 기법을 적용한 편향 보정 평균추정량을 제안하였다.

본 논문의 구성은 다음과 같다. 먼저 2절에서는 기존의 연구에서 얻어진 응답률 모형과 편향 추정량을 간단히 살펴보았다. 3절에서는 본 연구에서 제안한 비모수적 함수 추정 기법을 이용하여 얻어진 결과를 편향 추정량에 적용하는 방법을 설명하였다. 4절에서는 최종적으로 얻어진 편향 보정 평균추정량을 설명하였다. 5절에서는 모의실험을 통해 본 연구에서 제안한 방법의 타당성과 우수성을 살펴보았으며 6절에 본 연구에서 제안한 방법과 관련된 결론을 수록하였다.

2. 편향 추정

2.1. 정보적 표본설계 기법을 이용한 표본 분포

본 연구에서 사용된 기본 개념은 응답률이 관심변수의 함수이고, 관심변수와 보조변수 간에 초모집단 모형이 형성되는 경우는 실제 얻어진 표본 자료의 표본 분포와 모집단 분포는 일치하지 않는다는 것이다. 이 내용은 정보적 표본설계를 수행할 때 기본적으로 사용되는 개념이다. Pfeffermann 등 (1998)은 정보적 표본설계를 사용할 때 $f_s(y_i|\theta^*, x_i) = f(y_i|i \in s, x_i) = Pr(i \in s|y_i, x_i) f_p(y_i|\theta, x_i)/Pr(i \in s|x_i)$이고 여기서 θ^*는 θ의 함수일 때 $Pr(i \in s|y_i x_i) = E_p(\pi_i|y_i, x_i)$이 되며, 또한 $Pr(i \in s|x_i) = E_p(\pi_i|x_i)$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, π_i는 표본 포함확률 또는 응답률, $E_p(\pi_i|y_i, x_i)$는 x_i, y_i가 주어졌을 때 자료가 표본에 포함될 포함확률 또는 최종적인 응답률이다. 만약 $E_p(\pi_i|y_i, x_i) = E_p(\pi_i|x_i)$이면 모집단 분포와 표본 분포는 같아진다. 이에 관한 내용은 Chung과 Shin (2017)을 참고하기 바란다. Lee (2019)를 살펴보면 매우 쉽게 이론을 이해할 수 있다.

2.2. 응답률 모형과 편향 추정

2.2.1. 응답률 모형

편향 추정을 위한 다수의 응답률 모형이 개발되었으며 기존의 연구에서 이미 편향 보정 평균추정량의 우수성이 확인되었다. 특히, Bethlehem (2020)에서는 로지스틱 응답률 모형과 선형 응답률 모형을 사용하여 자료를 분석하였으며 두 응답률 모형의 결과에 큰 차이를 보이지 않은 것을 확인하였다. 본 연구에서는 기본적으로 고려할 수 있는 모형이면서도 이론적으로 쉽게 편향이 추정되는 선형 응답률 모형을 고려하였다. 이는 선형 응답률 모형에서 효과가 있다면 다른 응답률 모형을 사용할 때도 효과가 있을 것으로 예상할 수 있기 때문이다. 선형 응답률 모형은 다음과 같이 정의된다.

$$E_p(\pi_i|y_i, x_i) = b_0 + b_1 y_i. \tag{2.2}$$

따라서 (2.1)의 분모는 다음과 같이 얻어진다.

$$E_p(\pi_i|x_i) = E(E_p(\pi_i|y_i, x_i)|x_i) = E(b_0 + b_1 y_i|x_i) = b_0 + b_1 E_p(y_i|x_i). \tag{2.3}$$

이제 (2.2)와 (2.3)을 이용하면 다음의 결과를 얻는다.

$$\frac{E_p(\pi_i|y_i, x_i)}{E_p(\pi_i|x_i)} = \frac{b_0 + b_1 y_i}{b_0 + b_1 E_p(y_i|x_i)}. \tag{2.4}$$

정보적 표본설계의 경우 알려진 b_0와 b_1을 사용하지만, 본 연구와 같이 응답률 모형을 사용할 경우에서는 모형에 포함된 모수를 추정하여야 한다. b_0와 b_1의 모수 추정에 관한 내용은 2.3절에서 설명하였다.

2.2.2. 편향 추정

초모집단 모형과 응답률 모형이 설정되면 이를 기반으로 편향이 추정된다. 이제 선형 응답률 모형에서 얻어진 결과인 (2.4)를 이용하면 표본 평균 $E_s(y_i|x_i) = \mu_i^{(s)}$는 다음과 같이 구해진다.

$$\mu_i^{(s)} = E_s(y_i|x_i) = \int y_i \left(\frac{b_0 + b_1 y_i}{b_0 + b_1 E_p(y_i|x_i)} \right) f_p(y_i|x_i) dy_i = \frac{b_0 \mu_i + b_1 E_p\left(y_i^2|x_i\right)}{b_0 + b_1 \mu_i}, \tag{2.5}$$

여기서 $\mu_i = E_p(y_i|x_i)$이다. 또한 $E_p(y_i^2|x_i) = Var_p(y_i|x_i) + \mu_i^2$이므로 이 결과를 (2.5) 대입하면,

$$\mu_i^{(s)} = \mu_i + \frac{b_1}{b_0 + b_1 \mu_i} \times Var_p(y_i|x_i)$$

를 얻게 된다. 따라서, 이론적으로 얻어진 편향은 $b_1/(b_0 + b_1\mu_i) \times \text{Var}_p(y_i|x_i)$이 된다. 이제 $\text{Var}_p(y_i|x_i) = \sigma^2$으로 i번째 자료에 무관하게 일정하다면 추정된 편향은 다음과 같다.

$$\frac{b_1}{b_0 + b_1 \mu_i} \sigma^2. \tag{2.6}$$

이 결과는 Chung과 Shin (2019)에서 확인할 수 있다. 이제 b_0, b_1이 응답률 모형에서 적절하게 추정되고 초모집단 모형에서 μ_i와 σ^2의 추정치를 이용하면 최종적으로 편향의 추정 결과를 얻을 수 있다.

2.3. 응답률 모형의 모수 b_0, b_1 추정

선형 응답률 모형인 (2.2)에 포함된 모수 b_0, b_1을 추정하기 위해서는 응답률 $E_p(\pi_i|y_i, x_i)$를 추정해야 한다. 응답률은 관심변수의 함수이므로 관심변수 y_i를 이용해서 추정해야 하지만 실질적으로 관심변수 y_i를 이용해 응답률을 추정할 수 없다. 이때 현실적으로 사용할 수 있는 방법은 모집단에 포함된 보조변수 x_i를 사용하는 것이다.

2.3.1. 성향점수를 이용한 응답률 추정

Bethlehem (2020)은 응답률을 계산하는 다양한 방법을 설명하였으며 특히 모형을 이용한 방법인 로짓 모형을 이용하여 응답률을 구하는 방법을 연구하였다. 이 방법은 이미 잘 알려진 방법으로 보조변수 x_i를 이용하여 응답률로 성향점수(propensity score)를 이용한다. 먼저 응답 여부를 나타내는 확률변수 R_i는 다음과 같이 정의된다.

$$R_i = \begin{cases} 1, & \text{개체 i 가 응답한 경우} \\ 0, & \text{개체 i 가 응답하지 않은 경우.} \end{cases}$$

정의에 따라 확률변수 R_i는 베르누이 분포를 따르며 응답 확률 $\pi_i = Pr(R_i = 1)$로 얻어진다. 이론적으로는 π_i가 관심변수 y_i의 함수이지만 실질적으로 π_i의 추정은 보조변수인 x_i에 의해 이루어지기 때문에 $\hat{\pi}_i$는 x_i의

함수로 구해야만 한다. 따라서 $\hat{E}_p(\pi_i|y_i, x_i) = \hat{\pi}_i \approx \hat{Pr}(R_i = 1|x_i)$로 구할 수 있으며 흔히 다음의 로짓 모형으로 추정한다. 즉

$$\ln\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = \hat{\alpha}_0 + \hat{\alpha_1} x_{1i} + \cdots + \hat{\alpha}_p x_{pi}, \tag{2.7}$$

에서 $\hat{\pi}_i$가 얻어진다. 이제 $\hat{w}_i = 1/\hat{\pi}_i$이라 하자. 그러면 얻어진 가중치 합 $\sum_{i=1}^{r} \hat{w}_i = N$이 되어야 하므로 모집단 보정인자 $f^{\text{pop}} = N/\sum_{i=1}^{r} \hat{w}_i$를 각각의 가중치에 곱하여 최종적으로 성향점수를 이용한 가중치를 얻는다. 즉 성향점수를 이용한 최종 가중치는 $\hat{w}_i^{(P)} = \hat{w}_i \times f^{\text{pop}}$으로 얻어진다.

2.3.2. 세부 층을 이용한 응답률 추정

이 방법은 Chung과 Shin (2017, 2019), Min과 Shin (2018)에서도 사용한 방법이다. 먼저 주어진 하나의 층을 L개의 세부 층으로 나눈다. 이때 모집단에 포함된 보조변수 x_i의 정보만 있으므로 보조변수 x_i를 기준으로 세부 층을 나누며 나누어진 세부 층에서 세부 층 가중치 $w_h, h = 1, \ldots, L$를 구한다. 세부 층을 나누는 다양한 기준이 연구되었으나 모집단의 보조변수 x_i를 정해진 세부 층 개수 L에 따라 분위수를 구한 후 분위수를 경계로 층을 나눈다. 나누어진 h번째 세부 층에서 모집단 수와 최종 응답 자료 수인 N_h, r_h를 이용하여 가중치 $\hat{w}_h = N_h/r_h$를 추정한다. 이제 i번째 자료가 h 세부 층에 속하면 가중치를 $\hat{w}_h$로 결정한다. 즉 $\hat{w}_i^{(D)} = \hat{w}_h \sum_{h=1}^{L} I(i \in s_h)$이고 s_h는 h 층의 표본집합이다.

2.3.3. 응답률 모형의 b_0, b_1 추정

응답률이 얻어지면 얻어진 응답률을 이용하여 모형을 만들고, 만들어진 모형에서 모수를 추정한다. 먼저 성향점수에서 추정된 응답률 $\hat{\pi}_i = \hat{E}_p(\pi_i|y_i, x_i)$를 이용하여 모수를 추정한다. 또한, Pfeffermann과 Sverchkov (2003)에서 얻어진 결과인 $E_s(w_i|y_i, x_i) \approx w_i$와 $E_s(w_i|y_i, x_i) = 1/E_p(\pi_i|y_i, x_i)$를 이용하여 세부 층에서 얻어진 가중치 $\hat{w}_i$로 응답률을 추정한다. 즉 $\hat{E}_p(\pi_i|y_i, x_i) = 1/\hat{w}_i$를 사용한다. 결국 (2.2)에 의해 다음의 모형이 완성된다.

$$\frac{1}{\hat{w}_i} = b_0 + b_1 y_i + \eta_i. \tag{2.8}$$

여기서 η_i는 독립이고, 같은 분포를 따르며 등분산성을 만족한다고 가정한다. 이제 세부 층에서 얻어진 가중치 $\hat{w}_i^{(D)}$ 또는 성향점수를 이용해 얻은 가중치 $\hat{w}_i^{(P)}$와 자료 y_i를 이용하여 각각 b_0, b_1가 추정된다. 얻어진 추정량 $\hat{b}_0, \hat{b}_1$을 이용하여 편향을 계산한다.

3. 초모집단 모형

3.1. 모수적 초모집단 모형

Chung과 Shin (2020) 그리고 Pfeffermann 등 (1998)에서 사용한 초모집단 모형은 선형 회귀 모형과 로그-선형 모형이다. 초모집단 모형의 오차가 정규분포인 경우에서는 선형 회귀 모형이 매우 효과적이며 초모집단 모형의 오차가 로그-정규분포 또는 감마분포인 경우는 로그-선형 모형이 효과적이다. 본 연구에서는 선형을 따르지 않거나 함수가 알려지지 않은 경우로 초모집단 모형을 확장하였으며 변환이 필요하지 않고 분산이 일정한 경우만을 고려하였다. 이에 본 연구에서 사용한 초모집단 모형의 형태는 다음과 같다.

1. 선형 회귀 모형

$$f_p(y_i|x_i) = N\left(\beta_0 + \beta_1 x_i, \sigma^2\right). \tag{3.1}$$

2. 2차 회귀 모형

$$f_p(y_i|x_i) = N\left(\beta_0 + \beta_1 x_i + \beta_2 x_i^2, \sigma^2\right). \tag{3.2}$$

3. Sine 함수

$$f_p(y_i|x_i) = N\left(\beta_0 + \beta_1 x_i + \beta_3 \sin(2\pi f x_i), \sigma^2\right). \tag{3.3}$$

흔히 보조변수가 증가할 때 관심변수는 증가하거나 감소하는 경향을 보이므로 본 연구에서도 위의 모형에서 적절한 $\beta_0, \beta_1, \beta_2, \beta_3$ 그리고 f를 사용하였다.

3.2. 초모집단 모형의 비모수적 함수 추정

본 연구에서 제안한 비모수적 형태의 초모집단 모형은 다음과 같다.

$$y_i = m(x_i) + \epsilon_i. \tag{3.4}$$

여기서 $m(x_i)$는 보조변수 x_i의 임의의 함수로 $\mu_i = m(x_i)$로 표시한다. 본 연구에서는 μ_i의 추정값인 $\hat{\mu}_i = \hat{m}(x_i)$는 국소 선형 회귀(local linear regression) 방법으로 추정한다. 잘 알려진 것처럼 국소 선형 회귀법은 선형 회귀를 국소적으로 커널 가중치를 적용하는 방법으로 주어진 점 x에서의 추정치를 구하기 위해 전체 관측치를 모두 사용하는 대신 x 근처의 관측치만을 사용하는 방법이다. 이제 커널이라 부르는 임의의 확률밀도함수 $K(x, x_i) = K(x_i - x/\lambda)$를 국소 선형 회귀법에 적합하면

$$\sum_{i=1}^{n} K(x, x_i) \{y_i - \beta_0 - \beta_1(x_i - x)\}^2 \tag{3.5}$$

와 같은 가중최소제곱 형태를 따른다. 본 연구에서는 최적의 평활 계수 λ를 추정하기 위해 일반으로 사용하는 일반화교차확인(generalized cross validation, GCV)을 최소화하는 방법인 (3.6)을 사용하였다.

$$\min \text{GCV}\ (\lambda) = \min \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{y_i - \hat{m}(x_i)}{1 - tr\,(H_\lambda)\,/n} \right\}. \tag{3.6}$$

여기서 $tr(H_\lambda)$는 평활 계수 λ가 특정한 값으로 주어지고 $m(x)$의 추정치를 $\hat{m}_\lambda(x)$라 할 때 $\hat{m}_\lambda(x) = H_\lambda y$로 표현되는 $n \times n$ 모자 행렬 H_λ의 대각원소의 합이다. 본 연구에서 사용한 국소 선형 회귀법은 이미 잘 알려진 방법으로 이를 통해 모수 μ_i 그리고 σ^2의 추정값인 $\hat{\mu}_i$와 $\hat{\sigma}^2$이 얻어지며 이 결과를 (2.6)의 편향 추정에 사용한다.

4. 편향 보정 평균추정량

본 연구에서 사용한 추정량은 다음과 같다.

M^P : 성향점수를 이용하여 얻어진 가중치를 사용한 가중평균을 이용한다. 즉 성향점수로 얻어진 가중치 w_i^P를 사용한 다음의 (4.1)을 사용한다.

$$\hat{\bar{Y}}^P = \frac{1}{N} \sum_{i=1}^{n} w_i^P y_{hi}. \tag{4.1}$$

M^D : 세부 층에서 얻어진 가중치를 사용하며 h 세부 층에 속한 모든 자료의 가중치가 같으므로 $w_{hi} = w_h = w_h^D$를 사용한다. 즉 다음의 (4.2)를 사용한다.

$$\hat{\bar{Y}}^D = \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{n_h} w_h^D y_{hi}. \tag{4.2}$$

M_N : 초모집단 모형에서 국소 선형 회귀법으로 기댓값의 추정치 $\hat{\mu}_i^{(s)}$와 분산 추정치 $\hat{\sigma}^2$을 구하고 또한 응답률 모형인 (2.8)에서 $\hat{b}_0, \hat{b}_1$을 구한 후 (4.3)을 사용한다.

$$\hat{\bar{Y}}_N = \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{n_h} w_h\left(\hat{\mu}_i^{(s)} - \frac{\hat{b}_1\hat{\sigma}^2}{\hat{b}_0 + \hat{b}_1\hat{\mu}_i^{(s)}}\right). \tag{4.3}$$

M_N 방법에 사용되는 가중치를 결정할 때, 세부 층을 사용한 경우는 $\hat{\bar{Y}}_N^D$ 그리고 성향점수를 사용한 경우는 $\hat{\bar{Y}}_N^P$라 표시한다.

M_{INF} : 초모집단 모형의 형태가 알려진 경우에서 얻어진 기댓값의 추정치 $\hat{\mu}_i^{(s)}$와 분산 추정치 $\hat{\sigma}^2$를 (4.3)에 대입하여 추정한다. 이때 $\hat{b}_0, \hat{b}_1$은 M_N 방법에서 얻어진 같은 추정값을 사용한다. 또한, 성향점수를 사용해서 얻어진 추정량을 $\hat{\bar{Y}}_{\text{INF}}^P$, 그리고 세부 층을 사용해서 얻어진 추정량을 $\hat{\bar{Y}}_{\text{INF}}^D$라 표시한다.

결론적으로 M^P와 M^D방법은 편향을 보정하지 않는 가중평균추정량이다. 여기서 P는 성향점수에서 얻어진 가중치 사용을 의미하고 D는 세부 층을 사용하여 얻어진 가중치 사용을 의미한다. 또한, M_{INF}는 초모집단 모형의 형태가 알려진 경우의 결과이고, M_N은 초모집단 모형의 형태를 사용하지 않고 국소 선형 회귀법을 사용한 경우의 결과이다.

5. 모의실험

5.1. 모의실험 설계

흔히 표본설계에서는 층화추출법을 사용하고 전체 모집단 평균을 추정한다. 본 연구에서는 여러 개의 층 중에서 주어진 한 개의 특정 층의 추정을 고려하였다. 이는 층화추출법에서는 층별로 모수 추정이 이루어지기 때문에 하나의 층을 고려하여도 일반성을 잃지 않기 때문이다. 다음이 모의실험을 위한 자료생성 과정과 모수 추정 방법이다. 전체적인 모의실험 방법은 Chung과 Shin (2017, 2019) 방법과 유사한 방법을 사용하였다.

Step 1 **모집단 생성과정**

초모집단 모형의 오차가 정규분포인 경우의 정보적 표본설계를 위한 모집단 자료생성 과정은 다음과 같다.

1. 보조변수 자료 x_i 생성:
   $x_i = 200 + \gamma_i, i = 1, \ldots, N$
   여기서 $\gamma_i \overset{iid}{\sim} \text{Unif}(0, 100)$이다. 따라서 보조변수 x_i는 200에서 300 사이의 값을 갖는다.
2. 초모집단 모형
   주어진 보조변수 자료를 이용하여 관심변수 자료를 생성한다.
   (1) 1차 식 : $y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$
   (2) 2차 식 : $y_i = \beta_0 + \beta_1 x_i + \beta_2 (x_i - 230)^2 + \epsilon_i, \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$
   (3) Sine 함수 : $y_i = \beta_0 + \beta_1 x_i + \beta_3 sin(2\pi x/100) + \epsilon_i, \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$
   (4) logistic 함수 : $z_i = -25 + 0.1 x_i, y_i = 200 + 400 \times e^{z_i}/(1 + e^{z_i}) + \epsilon_i, \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

여기서 $\beta_0 = 10$, $\beta_1 = 5$, $\beta_2 = 0.1$, $\beta_3 = 100$, $\sigma^2 = 900$이고, 모집단 자료 수 $N = 10,000$을 사용한다. (1), (2), (3)의 경우는 초모집단 모형의 형태를 파악할 수 있는 경우이고, (4)는 초모집단 모형을 파악할 수 없는 경우를 위해 관심변수를 생성하였다.

Step 2 **표본추출과정**

생성된 N개의 모집단 자료에서 n개의 표본을 추출한다. 추출된 n개의 자료에서 주어진 응답률 모형에 따라 랜덤으로 무응답을 만든다.

3. N개의 모집단 자료에서 단순임의추출(simple random sample)로 n개의 표본을 추출한다. 이때 $n = 200, 500$을 사용한다.
4. 추출된 n개의 표본에서 선형 모형인 $\pi_i = b_0 + b_1 y_i, \pi_i \in [0, 1]$을 이용하여 무응답을 생성한다. 즉 y_i의 최솟값에서의 응답률을 $\pi_y^{\min}$, y_i의 최댓값에서의 응답률을 $\pi_y^{\max}$라 할 때, $(\pi_y^{\min}, \pi_y^{\max})$=(0.9,0.3), (0.9,0.5), (0.5,0.9), (0.3,0.9)을 사용하여 b_0, b_1을 구하고 y_i에 따라 응답률을 계산한다. 계산된 응답률에 따라 무응답을 생성한다.
5. 응답한 최종 조사 자료는 r개이다. 여기서 $(\pi_y^{\min}, \pi_y^{\max})$=(0.9,0.3) 또는 $(\pi_y^{\min}, \pi_y^{\max})$=(0.3,0.9)에서는 전체 자료의 약 60%가 되어 주어진 자료 수 n에 비해 약 40%가 감소한다. $(\pi_y^{\min}, \pi_y^{\max})$=(0.9,0.5) 또는 $(\pi_y^{\min}, \pi_y^{\max})$=(0.5,0.9)인 경우는 전체 자료의 70%가 되어 주어진 자료 수 n에 비해 약 30%가 감소한다. 물론 초모집단 모형의 형태에 따라 비율은 달라진다.

Step 3 **층화**

얻어진 표본 자료는 $(x_i, y_i), i = 1, \ldots, r$ 이고 무응답에 의해 각 자료의 가중치는 달라진다. 이를 반영하기 위해 두 방법이 사용되었다.

6.1 세부 층 방법: 먼저 주어진 하나의 모집단 층을 L개의 세부 층으로 나눈다. 실제 자료 분석에서는 모집단에 보조변수 x_i의 정보만 있으므로 보조변수를 기준으로 세부 층을 나눈다. 보조변수 x_i를 기준으로 분위수를 구한 후, 분위수를 이용하여 모집단을 L개의 세부 층으로 나눈다. 여기서 L은 표본 수에 따라 $n = 200$일 때 $L = 10$ 그리고 $n = 500$일 때 $L = 25$를 사용한다.

6.2 성향점수 방법: 응답인 경우 $R_i = 1$, 무응답인 경우 $R_i = 0$으로 하고, 독립변수를 x_i로 하는 로짓 모형을 이용하여 성향점수인 응답률 $\hat{\pi}_i$를 구한다.

Step 4 **모수 추정**

7.1 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 $w_i = w_{(i \in s_h)} = w_h$가 된다. 즉 같은 세부 층에 포함된 자료의 가중치는 같다.

7.2 로짓 모형에서 얻어진 응답률의 역수를 모집단 수로 보정하여 얻은 최종 가중치를 사용한다. 즉 $\hat{w}_i^{(P)} = \hat{w}_i \times f^{\text{pop}}$이고 $\hat{w}_i = 1/\hat{\pi}_i$, $f^{\text{pop}} = N/\sum_{i=1}^{r} \hat{w}_i$이다.

8. 선형 응답률 모형인 (2.8)을 사용하여 모수 b_0, b_1을 추정한다.
9. 추출된 자료 (x_i, y_i)와 알려진 초모집단 모형으로 회귀분석을 수행하여 $\mu_i^{(s)}, \sigma^2$을 추정한다. 또한, 본 연구에서 제안한 방법인 국소 선형 회귀법을 적용하여 $\mu_i^{(s)}, \sigma^2$을 추정한다. 다만 초모집단 모형의 형태가 로지스틱 함수인 경우는 모형의 형태를 파악하기 어렵다고 가정하여 초모집단 모형으로 1차 식을 가정한 후 회귀분석을 수행한다.
10. 계산된 결과를 이용하여 제시된 평균추정값을 계산한다.

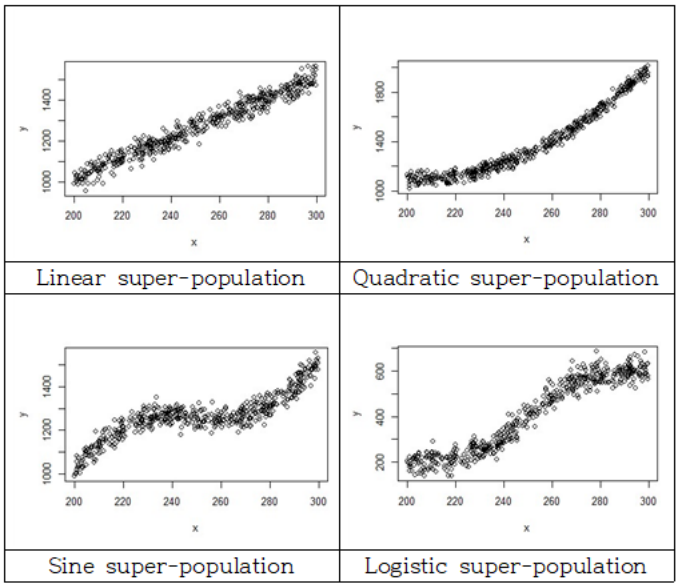

Figure 1: *Sample plot of generated super-population.*

이제 얻어진 평균추정값은 다음의 비교통계량, 편향(bias), 상대절대편향(relative absolute bias, Rabias) 그리고 제곱근 mean squared error (root mean squared error, RMSE)를 이용하여 결과의 성능이 비교되었다. 각 비교 통계량의 정의는 다음과 같다.

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

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

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

여기서 $R = 1,000$을 사용하였으며 반복마다 새로운 모집단을 생성하여 비교 통계량을 계산하였다. 이는 생성된 특정 모집단의 영향을 줄이기 위함이다. 이에 r번째 반복 모집단의 참값을 $\bar{Y}_r$로 표시하였다.

5.2. 모의실험 결과

응답률 모형이 선형이고, 초모집단 모형의 오차가 정규분포를 따르는 경우에서 얻어진 편향 보정 평균추정 결과가 초모집단 모형의 형태에 따라 분석되었다. 먼저 Figure 1은 초모집단 모형에 따라 생성된 모집단에서 추출된 $n = 500$개의 표본 자료를 이용하여 그린 그림이다. 1차 식의 경우 그 형태를 파악할 수 있으나 2차 식 또는 Sine 함수의 경우 모형 형태 파악이 쉽지 않으며 특히 Sine 함수를 가정하더라도 주기 등을 파악하여 독립변수를 생성하기가 쉽지 않다. 다만 본 모의실험에서는 독립변수를 알고 있으므로 이를 사용하였다. 반면 로지스틱 모형의 경우는 독립변수를 파악하기가 매우 어렵다.

Table 1: Comparison result of linear super-population model with $n = 200$

$\pi_y^{\min}$	$\pi_y^{\max}$	r	β_0	β_1	Comparison Statistics	Estimator					
						$\hat{\bar{Y}}^P$	$\hat{\bar{Y}}^D$	$\hat{\bar{Y}}^P_{\text{INF}}$	$\hat{\bar{Y}}^D_{\text{INF}}$	$\hat{\bar{Y}}^P_N$	$\hat{\bar{Y}}^D_N$
0.9	0.3	120	10	5	Bias	−0.981	−1.807	0.385	−0.539	0.371	−0.552
					Rabias	0.007	0.002	0.007	0.002	0.007	0.002
					RMSE	10.365	3.837	10.352	3.532	10.351	3.533
0.9	0.5	140	10	5	Bias	−0.455	−0.945	0.312	−0.230	0.305	−0.237
					Rabias	0.006	0.002	0.006	0.002	0.006	0.002
					RMSE	10.147	3.078	10.142	3.006	10.142	3.005
0.5	0.9	140	10	5	Bias	0.778	1.049	0.035	0.338	0.041	0.344
					Rabias	0.006	0.002	0.006	0.002	0.006	0.002
					RMSE	10.144	3.087	10.129	2.977	10.129	2.978
0.3	0.9	120	10	5	Bias	1.151	1.830	−0.182	0.570	−0.167	0.583
					Rabias	0.007	0.002	0.007	0.002	0.007	0.002
					RMSE	10.371	3.710	10.349	3.352	10.346	3.354

Table 2: Comparison result of linear super-population model with $n = 500$

$\pi_y^{\min}$	$\pi_y^{\max}$	r	β_0	β_1	Comparison Statistics	Estimator					
						$\hat{\bar{Y}}^P$	$\hat{\bar{Y}}^D$	$\hat{\bar{Y}}^P_{\text{INF}}$	$\hat{\bar{Y}}^D_{\text{INF}}$	$\hat{\bar{Y}}^P_N$	$\hat{\bar{Y}}^D_N$
0.9	0.3	300	10	5	Bias	−1.280	−1.467	0.067	−0.212	0.061	−0.217
					Rabias	0.004	0.001	0.004	0.001	0.004	0.001
					RMSE	6.676	2.335	6.553	1.893	6.553	1.894
0.9	0.5	350	10	5	Bias	−0.738	−0.748	0.011	−0.036	0.009	−0.039
					Rabias	0.004	0.001	0.004	0.001	0.004	0.001
					RMSE	6.446	1.800	6.406	1.683	6.406	1.683
0.5	0.9	350	10	5	Bias	0.542	0.872	−0.205	0.173	−0.203	0.175
					Rabias	0.004	0.001	0.004	0.001	0.004	0.001
					RMSE	6.395	1.879	6.379	1.721	6.379	1.721
0.3	0.9	300	10	5	Bias	1.055	1.563	−0.296	0.311	−0.292	0.315
					Rabias	0.004	0.002	0.004	0.001	0.004	0.001
					RMSE	6.666	2.415	6.606	1.920	6.605	1.921

5.2.1. 초모집단 모형이 선형인 경우

이 절에서는 선형 초모집단 모형에서 모수적 초모집단 모형과 국소 선형 회귀 추정을 사용한 결과를 수록하였다. 먼저 Table 1은 자료의 수가 200인 경우이고, Table 2는 자료 수가 500인 경우이다. 두 표에서 성향점수를 사용한 결과인 $\hat{\bar{Y}}^P$와 세부 층을 사용한 결과인 $\hat{\bar{Y}}^D$를 비교하면 편향의 경우 $\hat{\bar{Y}}^P$의 결과가 모든 응답률에서 우수한 것을 확인할 수 있으나 다른 비교 통계량에서는 $\hat{\bar{Y}}^D$가 매우 큰 차이로 우수한 것을 확인할 수 있다. 또한, 편향을 보정한 결과를 살펴보면 성향점수를 사용한 결과인 $\hat{\bar{Y}}^P_{\text{INF}}$보다 세부 층을 사용한 결과인 $\hat{\bar{Y}}^D_{\text{INF}}$가 편향을 제외한 다른 비교 통계량에서 매우 우수한 결과를 주는 것을 확인할 수 있다. 특히 RMSE를 살펴보면 $\hat{\bar{Y}}^P$와 $\hat{\bar{Y}}^P_{\text{INF}}$의 비교에서 편향 보정 효과가 크지 않지만 $\hat{\bar{Y}}^D$와 $\hat{\bar{Y}}^D_{\text{INF}}$의 비교에서는 편향 보정 효과가 매우 큰 것을 확인할 수 있다. 따라서 비교된 4개의 추정량에서는 $\hat{\bar{Y}}^D_{\text{INF}}$가 가장 우수한 결과를 주고 있다. 다음으로 모수적 초모집단 모형을 사용한 결과와 국소 선형 회귀 추정을 사용한 결과를 살펴보면 $\hat{\bar{Y}}^P_{\text{INF}}$와 $\hat{\bar{Y}}^P_N$의 결과가 유사하고, $\hat{\bar{Y}}^D_{\text{INF}}$의 결과와 $\hat{\bar{Y}}^D_N$의 결과가 매우 유사하다. 따라서 모수적 초모집단 모형 사용과 국소 선형 회귀 모형 사용이 유사한

Table 3: Comparison result of quadratic super-population model with $n = 200$

$\pi_y^{\min}$	$\pi_y^{\max}$	r	β_0	β_1	β_2	Comparison Statistics	Estimator					
							$\hat{\bar{Y}}^P$	$\hat{\bar{Y}}^D$	$\hat{\bar{Y}}^P_{\mathrm{INF}}$	$\hat{\bar{Y}}^D_{\mathrm{INF}}$	$\hat{\bar{Y}}^P_N$	$\hat{\bar{Y}}^D_N$
						Bias	−5.412	−1.667	−4.635	−0.914	−4.634	−0.893
0.9	0.3	136	10	5	0.1	Rabias	0.012	0.003	0.012	0.003	0.012	0.003
						RMSE	20.881	4.585	20.702	4.390	20.703	4.420
						Bias	−2.156	−0.917	−1.713	−0.483	−1.712	−0.483
0.9	0.5	151	10	5	0.1	Rabias	0.011	0.002	0.011	0.002	0.011	0.002
						RMSE	19.408	3.923	19.371	3.853	19.371	3.853
						Bias	3.057	1.148	2.598	0.665	2.599	0.666
0.5	0.9	129	10	5	0.1	Rabias	0.011	0.002	0.011	0.002	0.011	0.002
						RMSE	19.070	4.092	19.013	4.005	19.012	4.005
						Bias	1.889	2.178	0.997	1.235	1.000	1.239
0.3	0.9	103	10	5	0.1	Rabias	0.011	0.003	0.011	0.003	0.011	0.003
						RMSE	18.795	5.118	18.760	4.821	18.759	4.821

Table 4: Comparison result of quadratic super-population model with $n = 500$

$\pi_y^{\min}$	$\pi_y^{\max}$	r	β_0	β_1	β_2	Comparison Statistics	Estimator					
							$\hat{\bar{Y}}^P$	$\hat{\bar{Y}}^D$	$\hat{\bar{Y}}^P_{\mathrm{INF}}$	$\hat{\bar{Y}}^D_{\mathrm{INF}}$	$\hat{\bar{Y}}^P_N$	$\hat{\bar{Y}}^D_N$
						Bias	−6.390	−0.998	−5.626	−0.272	−5.624	−0.271
0.9	0.3	342	10	5	0.1	Rabias	0.008	0.001	0.008	0.001	0.008	0.001
						RMSE	14.194	2.112	13.872	1.895	13.871	1.895
						Bias	−3.103	−0.591	−2.664	−0.160	−2.664	−0.160
0.9	0.5	378	10	5	0.1	Rabias	0.007	0.001	0.007	0.001	0.007	0.001
						RMSE	12.674	1.802	12.573	1.725	12.573	1.725
						Bias	2.180	0.601	1.718	0.120	1.718	0.121
0.5	0.9	322	10	5	0.1	Rabias	0.007	0.001	0.007	0.001	0.007	0.001
						RMSE	12.078	1.940	12.007	1.868	12.006	1.868
						Bias	1.125	1.285	0.234	0.356	0.235	0.358
0.3	0.9	258	10	5	0.1	Rabias	0.007	0.001	0.007	0.001	0.007	0.001
						RMSE	11.808	2.622	11.778	2.335	11.777	2.335

결과를 주는 것을 확인하였다. 표본 수 $n = 200$인 Table 1과 $n = 500$인 Table 2에서 얻어진 추정량 비교 결과는 매우 유사하여 표본 수의 차이에 의한 결과의 차이는 없다고 판단되며 $\hat{\bar{Y}}^D_{\mathrm{INF}}$와 $\hat{\bar{Y}}^D_N$의 상대절대편향 결과인 Rabias를 살펴보면 매우 안정적인 결과를 준다.

5.2.2. 초모집단 모형이 2차 식인 경우

초모집단 모형이 2차 식인 경우의 결과인 Table 3과 Table 4는 선형 모형 결과인 Table 1과 Table 2 결과와 매우 유사하다. 즉 $\hat{\bar{Y}}^P$에 비해 $\hat{\bar{Y}}^D$의 결과가 매우 우수하며 편향을 보정한 $\hat{\bar{Y}}^P_{\mathrm{INF}}$와 $\hat{\bar{Y}}^D_{\mathrm{INF}}$의 결과가 편향을 보정하지

Table 5: Comparison result of sine super-population model with $n = 200$

$\pi_y^{\min}$	$\pi_y^{\max}$	r	β_0	β_1	β_3	Comparison Statistics	Estimator					
							$\hat{\bar{Y}}^P$	$\hat{\bar{Y}}^D$	$\hat{\bar{Y}}^P_{\text{INF}}$	$\hat{\bar{Y}}^D_{\text{INF}}$	$\hat{\bar{Y}}^P_N$	$\hat{\bar{Y}}^D_N$
						Bias	−4.390	−1.782	−3.386	−0.632	−3.384	−0.644
0.9	0.3	120	10	5	100	Rabias	0.006	0.002	0.005	0.002	0.005	0.002
						RMSE	8.960	3.762	8.539	3.496	8.538	3.496
						Bias	−2.294	−0.878	−1.731	−0.235	−1.730	−0.235
0.9	0.5	140	10	5	100	Rabias	0.005	0.002	0.005	0.002	0.005	0.002
						RMSE	7.949	3.122	7.822	3.086	7.822	3.088
						Bias	2.885	1.409	2.296	0.733	2.295	0.732
0.5	0.9	140	10	5	100	Rabias	0.005	0.002	0.005	0.002	0.005	0.002
						RMSE	8.063	3.323	7.902	3.179	7.902	3.178
						Bias	4.919	2.284	3.888	1.086	3.886	1.085
0.3	0.9	120	10	5	100	Rabias	0.006	0.003	0.005	0.002	0.005	0.002
						RMSE	9.196	3.993	8.723	3.586	8.722	3.574

Table 6: Comparison result of sine super-population model with $n = 500$

$\pi_y^{\min}$	$\pi_y^{\max}$	r	β_0	β_1	β_3	Comparison Statistics	Estimator					
							$\hat{\bar{Y}}^P$	$\hat{\bar{Y}}^D$	$\hat{\bar{Y}}^P_{\text{INF}}$	$\hat{\bar{Y}}^D_{\text{INF}}$	$\hat{\bar{Y}}^P_N$	$\hat{\bar{Y}}^D_N$
						Bias	−4.822	−1.526	−3.816	−0.336	−3.814	−0.335
0.9	0.3	301	10	5	100	Rabias	0.004	0.002	0.004	0.001	0.004	0.001
						RMSE	6.864	2.445	6.220	2.031	6.219	2.030
						Bias	−2.707	−0.831	−2.137	−0.152	−2.136	−0.151
0.9	0.5	351	10	5	100	Rabias	0.003	0.001	0.003	0.001	0.003	0.001
						RMSE	5.394	1.894	5.143	1.770	5.143	1.770
						Bias	2.432	0.898	1.845	0.233	1.844	0.232
0.5	0.9	350	10	5	100	Rabias	0.003	0.001	0.003	0.001	0.003	0.001
						RMSE	5.203	1.924	4.964	1.782	4.964	1.783
						Bias	4.473	1.527	3.450	0.356	3.448	0.353
0.3	0.9	300	10	5	100	Rabias	0.004	0.002	0.004	0.001	0.004	0.001
						RMSE	6.623	2.450	5.987	2.028	5.985	2.029

않은 $\hat{\bar{Y}}^P$와 $\hat{\bar{Y}}^D$에 비해 우수한 결과를 준다. 특히 $\hat{\bar{Y}}^D_{\text{INF}}$는 $\hat{\bar{Y}}^P_{\text{INF}}$에 비해 매우 우수한 결과를 준다. 또한, 국소 선형 회귀 추정을 사용한 $\hat{\bar{Y}}^P_N$와 $\hat{\bar{Y}}^D_N$ 결과는 모수적 초모집단 모형을 이용한 결과와 매우 유사하다. 또한 표본 수의 차이는 결과에 영향을 주지 않는다.

5.2.3. 초모집단 모형이 Sine 함수인 경우

초모집단 모형이 sine 함수인 경우의 결과인 Table 5와 Table 6은 선형과 2차 식 결과인 Table 1에서 Table 4 결과와 매우 유사하다. 즉 $\hat{\bar{Y}}^P$에 비해 $\hat{\bar{Y}}^D$의 결과가 매우 우수하며 편향을 보정한 $\hat{\bar{Y}}^P_{\text{INF}}$와 $\hat{\bar{Y}}^D_{\text{INF}}$의 결과가 편향을

Table 7: Comparison result of logistic super-population model with $n = 200$

$\pi_y^{\min}$	$\pi_y^{\max}$	r	Comparison Statistics	Estimator					
				$\hat{\bar{Y}}^P$	$\hat{\bar{Y}}^D$	$\hat{\bar{Y}}^P_{\text{INF}}$	$\hat{\bar{Y}}^D_{\text{INF}}$	$\hat{\bar{Y}}^P_N$	$\hat{\bar{Y}}^D_N$
0.9	0.3	120	Bias	−2.297	−2.167	1.244	1.272	−0.827	−0.730
			Rabias	0.023	0.008	0.023	0.008	0.022	0.007
			RMSE	11.404	4.033	11.402	3.975	11.249	3.559
0.9	0.5	140	Bias	−1.450	−1.173	0.469	0.709	−0.651	−0.387
			Rabias	0.022	0.006	0.022	0.006	0.022	0.006
			RMSE	10.919	3.140	10.883	3.271	10.858	3.001
0.5	0.9	140	Bias	1.835	1.033	−0.157	−0.979	1.006	0.194
			Rabias	0.022	0.006	0.022	0.007	0.022	0.006
			RMSE	10.946	3.073	10.862	3.356	10.861	2.976
0.3	0.9	120	Bias	2.758	1.949	−0.826	−1.606	1.261	0.457
			Rabias	0.023	0.008	0.022	0.008	0.022	0.007
			RMSE	11.296	3.861	11.170	4.096	11.092	3.472

Table 8: Comparison result of logistic super-population model with $n = 500$

$\pi_y^{\min}$	$\pi_y^{\max}$	r	Comparison Statistics	Estimator					
				$\hat{\bar{Y}}^P$	$\hat{\bar{Y}}^D$	$\hat{\bar{Y}}^P_{\text{INF}}$	$\hat{\bar{Y}}^D_{\text{INF}}$	$\hat{\bar{Y}}^P_N$	$\hat{\bar{Y}}^D_N$
0.9	0.3	300	Bias	−2.543	−1.721	0.96	1.685	−1.073	−0.285
			Rabias	0.015	0.005	0.014	0.006	0.014	0.004
			RMSE	7.547	2.596	7.238	2.72	7.206	1.994
0.9	0.5	350	Bias	−1.653	−0.916	0.272	0.994	−0.848	−0.116
			Rabias	0.014	0.004	0.014	0.004	0.014	0.003
			RMSE	7.162	1.937	6.998	2.109	7.026	1.739
0.5	0.9	350	Bias	1.791	0.96	−0.131	−0.937	0.986	0.164
			Rabias	0.014	0.004	0.014	0.004	0.014	0.003
			RMSE	7.150	1.932	6.929	2.062	6.992	1.710
0.3	0.9	299	Bias	2.644	1.723	−0.839	−1.655	1.181	0.298
			Rabias	0.015	0.005	0.015	0.005	0.015	0.004
			RMSE	7.605	2.611	7.255	2.722	7.251	2.015

보정하지 않은 $\hat{\bar{Y}}^P$와 $\hat{\bar{Y}}^D$에 비해 우수한 결과를 준다. 특히 $\hat{\bar{Y}}^D_{\text{INF}}$는 $\hat{\bar{Y}}^P_{\text{INF}}$에 비해 매우 우수한 결과를 준다. 또한, 국소 선형 회귀 추정을 사용한 $\hat{\bar{Y}}^P_N$와 $\hat{\bar{Y}}^D_N$ 결과는 모수적 초모집단 모형을 이용한 결과와 매우 유사하다.

5.2.4. 초모집단 모형이 로지스틱 함수인 경우

초모집단 모형이 로지스틱 함수인 경우의 결과인 Table 7과 Table 8은 선형과 2차식 그리고 Sine 함수인 경우의 결과와 유사하다. 다만 $\hat{\bar{Y}}^P_{\text{INF}}$와 $\hat{\bar{Y}}^D_{\text{INF}}$를 사용할 때 모수적 초모집단 모형으로 로지스틱 모형을 사용해야 하지만 선형을 가정하여 평균을 추정하였다. 따라서 잘못된 모수적 초모집단 모형을 사용하였기 때문에 편향을 보정

한 효과가 없을 수 있다. 특히 $\hat{\bar{Y}}^{D}_{\mathrm{INF}}$ 결과는 매우 나빠지는 것을 확인할 수 있다. 따라서 세부 층 가중치를 사용할 때 초모집단 모형의 형태가 알려지지 않은 경우는 사용에 주의할 필요가 있다. 반면 $\hat{\bar{Y}}^{P}_{\mathrm{INF}}$는 편향 보정 효과가 있는 것으로 나타났지만 그 효과는 다른 모의실험 결과와 유사하게 미미하게 좋아지고 있다. 이는 초모집단 형태가 정확히 선형 형태는 아니지만, 선형과 유사한 형태를 따르기 때문으로 판단된다. 그러나 $\hat{\bar{Y}}^{P}_{\mathrm{INF}}$의 RMSE 결과보다 $\hat{\bar{Y}}^{D}$의 RMSE 결과가 우수하기 때문에 $\hat{\bar{Y}}^{D}$를 사용하는 것이 타당하다. 이제 초모집단 형태를 가정하지 않는 결과인 $\hat{\bar{Y}}^{P}_{N}$와 $\hat{\bar{Y}}^{D}_{N}$을 살펴보면 먼저 $\hat{\bar{Y}}^{P}_{N}$은 $\hat{\bar{Y}}^{P}$에 비해 우수한 결과를 주지만 큰 차이를 보이지 않는다. 반면 $\hat{\bar{Y}}^{D}_{N}$은 $\hat{\bar{Y}}^{D}$에 비해 매우 우수한 결과를 주며 모든 추정량 중에서 가장 우수한 결과를 주고 있다. 따라서 초모집단 모형의 형태가 알려지지 않은 경우는 국소 선형 회귀 추정을 사용하는 것이 타당하다.

6. 결론

본 연구에서는 초모집단 모형이 선형이 아닌 일반 형태일 때 비모수적 회귀 모형을 이용하여 μ_i와 σ^2을 추정하여 편향을 보정 하는 방법을 제안하였다. 이론적으로 추정된 편향을 살펴보면 선형 응답률 모형에서 추정된 계수 b_0, b_1과 초모집단 모형에서 추정된 μ_i와 σ^2의 추정치를 이용하면 편향이 추정될 수 있다. 따라서 어떠한 초모집단 모형이라도 평균과 분산이 추정된다면 편향이 추정될 수 있고 이를 통하여 편향 보정 평균추정량이 얻어지게 된다. 따라서 초모집단 모형의 형태가 알려지지 않은 경우에서는 제안된 편향 보정 평균추정량을 사용함으로써 추정의 정확성이 향상될 수 있다고 판단된다. 다만 본 연구의 결과는 선형 응답률 모형과 분산이 일정한 임의의 초모집단 모형에 국한된 결과이므로 이를 확장한 연구가 필요하다.

References

Bethlehem J (2020). Working with response probabilities, *Journal of Official Statistics*, **36**, 647–674.

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

Chung HY and Shin KI (2019). Bias adjusted estimation in a sample survey with linear response rate, *Korean Journal of Applied Statistics*, **32**, 631–642.

Chung HY and Shin KI (2020). A study on non-response bias adjusted estimation in business survey, *Korean Journal of Applied Statistics*, **33**, 11–23.

Lee SE (2019). Application of informative sampling on big data, *Journal of The Korean Official Statistics*, **24**, 33–49.

Min JW and Shin KI (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 AM, and Rinott Y (1998). Parametric distributions of complex survey data under informative probability sampling, *Statistica Sinica*, **8**, 1087–1114.

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

Received August 10, 2021; Revised September 10, 2021; Accepted September 11, 2021

선형 응답률 모형에서 초모집단 모형의 비모수적 함수 추정을 이용한 무응답 편향 보정 추정

심주용[a] 신기일[1,a]

[a]한국외국어대학교 통계학과

요 약

표본조사에서는 다수의 무응답이 발생하며 이를 적절히 처리하는 다양한 방법이 개발되었다. 특히 무응답이 관심변수에 영향을 받고 이로 인해 발생한 편향은 추정의 정확성을 크게 떨어뜨리며 무응답 처리를 어렵게 한다. 최근 Chung과 Shin (2017, 2020)은 알려진 모수적 초모집단 모형과 응답률 모형을 이용하여 추정의 정확성을 향상한 추정량을 제안하였다. 본 연구에서는 초모집단 모형의 형태를 일반화하여 비모수적 함수 형태를 설정한 후 이를 기반으로 얻어진 편향을 적절히 처리한 편향 보정 평균추정량을 제안하였다. 모의실험을 통해 본 연구에서 제안한 방법의 우수성을 확인하였다.

주요용어: 성향점수, 표본 분포, 모집단 분포, 국소 선형 회귀

이 연구는 2021년 한국외국어대학교 교내연구비 지원을 받아 수행되었음.

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