The Korean Journal of Applied Statistics (2018) DOI: http://dx.doi.org/10.5351/KJAS.2018.31.5.621
31(5), 621–636

A study on the determination of substrata using the information of exponential response rate by simulation studies

Joo-Won Min[a] · Key-Il Shin[a,1]

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

(Received July 23, 2018; Revised August 17, 2018; Accepted September 9, 2018)

Abstract

Research on the application of informative sampling technique has been conducted in order to reduce the influence of non-response. Chung and Shin (*Korean Journal of Applied Statistics*, **30**, 993–1004, 2017) showed that the estimation accuracy improved when using exponential response rate information for the parameter estimation if the distribution of errors included in the super population model follows normal distribution. However this method divides the stratum into equally spaced substrata to obtain the sample weight of the informative sampling technique and shows that the accuracy of the estimation improves as the number of substrata increases. In this study, with the given number of total sample size, the optimal substratum boundary points are calculated using equal space, quantile, and LH algorithm; consequently, the results using those methods are compared through simulation. We also studied the criteria to determine the number of substrata and substratum boundaries that can be used in practice with various types of auxiliary variable distributions.

Keywords: LH algorithm, stratified sampling, sample inclusion probability, super-population model

1. 서론

현재 국내에서는 다수의 표본조사가 실시되고 있으며 조사 현실의 악화로 거의 모든 조사에서 무응답이 발생하고 있다. 단위 무응답과 항목 무응답으로 분류되는 무응답은 편향을 발생시켜 추정의 정확성을 떨어뜨린다. 무응답으로 인해 발생되는 편향을 파악하기 위해 R-지수를 사용할 수 있으며 이 결과를 이용하면 편향을 발생시키는 주요 변수를 찾을 수 있다. R-지수에 관한 내용은 Schouten 등 (2009)과 Lee와 Shin (2017)을 살펴보면 된다.

표본조사 결과의 정확성에 영향을 주는 무응답을 줄이기 위한 많은 노력이 수행되고 있다. 이 중 대표적인 것은 실사에서 항목 무응답을 줄이는 것이며 파라 데이터를 이용하여 무응답을 줄이거나 표본 층에서 단위 무응답이 발생한 경우 대체 표본을 사용하는 것이다. 통계적 처리 방법으로는 무응답으로 인해 발

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, Mohyeon-eup, Cheoin-gu, Yongin-si, Gyeonggi-do 17035, Korea. E-mail: keyshin@hufs.ac.kr

생한 결측값에 대체값을 사용하여 대체하거나 가중치를 보정하는 방법을 사용한다. 그러나 조사자료에서 얻어진 응답률 정보를 이용하여 추정을 보정함으로써 편향을 줄이는 방법에 관한 연구는 미미하다.

최근 응답률이 관심변수의 지수함수이고 관심변수와 보조변수 간에 선형관계가 있을 때 편향을 줄일 수 있는 방법이 연구되고 있다. 특히 Chung과 Shin (2017)은 정보적 표본설계 기법에서 얻어진 결과를 이용하여 무응답으로 인해 발생된 편향을 줄이는 방법을 연구하였다.

관심변수와 보조변수 간에 관계가 있고, 표본 추출과정에서 관심변수 또는 보조변수 자료 값을 이용하는 표본설계를 정보적 표본설계(informative sampling)라 한다. 정보적 표본설계는 1990년대 후반부터 연구가 시작되어 2000년대에도 지속적으로 활발한 연구가 진행되고 있다. 정보적 표본설계는 두 과정으로 나누어진다. 첫 번째 과정은 관심변수의 자료 생성과정으로 유한모집단(finite population)에서 관심변수와 보조변수 간에 초모집단모형(super-population model)이 형성되고 이 관계를 통하여 관심변수가 생성된다. 두 번째 과정은 표본 추출과정(selection mechanism)으로 자료가 표본에 포함될 확률인 표본 포함확률(inclusion probability)은 관심변수와 보조변수 값의 함수가 되며 이 값을 기반으로 표본이 추출된다. 이러한 과정을 통해 얻어지는 정보적 표본설계는 현재 사용되고 있는 표본설계방법을 포함하고 있으며, 기존의 표본설계 방법에 비해 관심변수의 정보를 더욱 적극적으로 사용하는 표본설계라 할 수 있다.

관심변수의 정보를 사용하는 정보적 표본설계가 사용되면 모집단 분포와 표본 분포는 일치하지 않는 것으로 알려져 있으며 따라서 모집단 자료에서 만들어진 관심변수와 보조변수의 초모집단모형은 표본 자료에서 만들어지는 관심변수와 보조변수와의 모형과 다르게 된다. 이와 같은 결과는 표본조사에서 얻어진 응답률을 표본 포함확률에 적용하였을 때 편향의 크기를 파악할 수 있는 이론적 근거를 제시하는 것으로 이를 통하여 편향을 보정할 수 있다. 본 연구에서는 관심변수가 보조변수와 선형관계가 있고 응답률이 지수형인 경우를 연구하였다. 특히 사업체 조사와 같이 층 내의 종사자 수가 커짐에 따라 사업체 수가 급격히 감소하는 경우에는 세부 층에 매우 작은 수의 사업체가 존재하기 때문에 등간격으로 세부층을 나눌 수 없는 경우도 발생한다. 이에 본 연구에서는 Chung과 Shin (2017)의 결과를 확장하여 등간격, 분위수, 그리고 LH 알고리즘을 사용하여 최적 세부 층 경계점과 최적 세부 층 개수를 구하였으며 그 결과를 모의실험을 통해 살펴보았다. 따라서 비록 본 논문에서 얻어진 결과는 이론적으로 최적값을 구한 것이 아니기 때문에 그 결과를 일반화할 수는 없지만 실무에서 사용할 수 있는 세부 층 구성 기준은 마련할 수 있을 것으로 판단된다.

본 논문의 구성은 다음과 같다. 먼저 2절에서는 기본적인 정보적 표본 설계를 설명하였다. 3절에서는 응답률이 지수형이고 초모집단 모형의 분포가 정규분포인 경우에서 세부 층을 나눌 때 등간격, 분위수 그리고 LH 알고리즘을 사용하는 방법을 설명하였다. 4절에는 모의실험을 통하여 정보적 표본설계 기법을 사용한 경우와 일반 층화추출법을 이용한 경우에서 얻어진 추정량의 성능을 비교하였다. 5절에 결론이 있다.

2. 초모집단 모형의 오차가 정규분포인 경우의 정보적 표본설계

2절의 정보적 표본설계에 관한 자세한 내용은 Pfeffermann 등 (1998, 2003, 2006)을 살펴보면 되고 본 연구에서는 Chung과 Shin (2017)의 내용을 수록하였다.

2.1. 정보적 표본설계

정보적 표본설계는 표본 추출과정이 관심변수 자료 값에 영향을 받는 표본설계이다. 이 표본설계에서 조사 자료는 두 과정으로 나누어 생성된다. 먼저 자료생성 과정이다. 표본 틀(sampling frame)에는 관

심변수 자료와 보조변수 자료가 있으며 관심변수는 보조변수와 통계모형을 이루면서 생성된다. 이러한 모형을 초모집단 모형이라 부르며 많은 경우 초모집단 모형은 선형 관계 또는 회귀모형을 가정할 수 있다. 물론 모형의 오차 분포에 따라 다양한 모형이 사용되기도 한다. 다음은 모집단을 대표하는 표본틀에서 표본이 추출되는 과정으로 이를 표본 추출과정이라 부른다. 여기서 가중치 또는 표본 추출확률이 계산되어야 하기 때문에 유한모집단을 가정한다. 표본 추출과정에서 관심변수 자료가 추출확률 또는 표본 포함확률에 영향을 주게되며 표본 포함확률은 y_i값의 함수로 표현된다. 결국 s를 표본 집합의 index라 하고 i번째 자료가 표본으로 추출될 확률이 $P(i \in s|y_i) = P(i \in s) = \pi_i$이면 즉 y_i가 어떤 값을 갖더라도 표본으로 추출될 확률이 모두 같은 경우에는 비정보적 표본설계(non-informative sampling)가 되고 만약 다르게 되면 정보적 표본설계가 된다. Pfeffermann 등 (1998)은 정보적 표본설계 하에서 θ^*를 θ의 함수라 할 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 $\Pr(i \in s|y_i, x_i) = E_p(\pi_i|y_i, x_i)$, $\Pr(i \in s|x_i) = E_p(\pi_i|x_i)$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 자료가 표본에 포함될 포함확률이다. 따라서 $E_p(\pi_i|y_i, x_i) = E_p(\pi_i|x_i)$이면 모집단 분포와 표본 분포는 같아진다. 만약 모집단 분포가 독립변수가 하나인 회귀모형을 따르며 오차가 정규분포를 따른다고 가정하면 모집단 분포는 다음과 같이 표현된다.

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

다음으로 표본 포함확률이 관심변수 y_i와 보조변수 x_i의 함수이고 식 (2.3)의 지수 형태를 따른다고 가정하자. 이는 초모집단 모형의 오차가 정규분포일 때 흔히 사용하는 가정이다.

$$E_p(\pi_i|y_i, x_i) = \exp\left\{a_0 + a_1 y_i + g(x)\right\}. \tag{2.3}$$

이제 식 (2.2)와 (2.3)을 식 (2.1)에 적용하면 최종적으로 다음의 표본 분포를 얻게 된다.

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

결국 모집단 분포인 식 (2.2)와 표본 분포 식 (2.4)를 비교하면 회귀모형의 절편이 β_0에서 $\beta_0 + a_1\sigma^2$으로 변한 것을 확인할 수 있다. 자세한 내용은 Chung과 Shin (2017)을 살펴보기 바란다.

2.2. 응답률 함수의 모수 추정

식 (2.1)을 살펴보면 정보적 표본설계의 핵심 내용은 표본 포함확률이 관심변수 y_i의 함수라는 것이며 정보적 표본설계 기법을 사용하게 되면 편향이 발생하게 되고 결과적으로 편향의 크기가 식 (2.4)에서 파악될 수 있다는 것이다. Chung과 Shin (2017)은 정보적 표본설계 기법의 결과를 응답률에 적용하여 편향을 보정함으로써 추정의 정확성을 향상시켰다. 여기서 정보적 표본설계 기법을 지수형 응답률 모형에 적용하기 위해 $E_p(\pi_i|y_i, x_i) = \exp(a_0 + a_1 y_i)$ 모형에 포함된 모수 a_0, a_1을 추정하였다. 이때 모수 추정을 위해서는 자료에서 얻어진 y_i와 응답률이 필요하기 때문에 주어진 하나의 층을 여러 개의 세부 층으로 나누는 방법이 사용되었다.

이제 주어진 하나의 층을 L개의 세부 층으로 나눈다고 가정하자. 여기서 관심변수의 모집단 정보는 알 수 없으나 보조변수의 모집단 정보는 알 수 있으므로 층을 나누는 기준은 보조변수 x_i에 의해 이루어진

다. 층을 나누는 방법으로 Chung과 Shin (2017)은 보조변수를 등간격으로 나누는 방법을 사용하였다. 그러나 사업체조사와 같이 보조변수 값에 따라 한쪽으로 치우친 분포인 경우에는 등간격보다 분포를 고려하여 세부 층을 나누는 것이 더욱 타당할 수 있다. 이에 본 연구에서는 분위수를 이용하는 방법과 절사표본설계에서 사용하는 LH 방법을 이용하여 층을 나누는 방법을 추가로 고려하였다.

이제 보조변수 x_i를 이용하여 앞에서 언급한 세가지 방법으로 L개의 세부 층을 구성한다. 구성된 L개의 세부 층을 이용하면 개별 응답률 π_i를 구할 수는 없지만 세부 층에 포함된 π_i를 $\pi_{i\in h} = \pi_h = n_h/N_h$로 동일하게 구할 수 있게 된다. 여기서 n_h와 N_n는 각각 h세부 층의 표본 수와 모집단 수이다. 본 논문에서는 지수형 응답률 모형인 $E_p(\pi_i|y_i, x_i) = \exp(a_0 + a_1 y_i)$을 사용한다. 결국 이 식에 $E_s(w_i|y_i) = 1/E_p(\pi_i|y_i)$와 $E_s(w_i|y_i) \approx w_i$를 적용하게 되면 $\pi_{h,i\in h} = \exp(a_0 + a_1 y_i)$이고 따라서 $1/w_i = \pi_{h,i\in h}$라 하면 $1/w_i = \exp(a_0 + a_1 y_i)$이 얻어진다. 따라서 식 (2.5)인 선형회귀모형을 이용하여 모수 a_0, a_1을 추정한다.

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

여기서 η_i는 $E(\eta_i) = 0$, $\mathrm{Var}(\eta_i) = \sigma_\eta^2$을 가정한다. 또한 식 (2.4)를 이용하여 σ^2을 추정한다. 식 (2.4)에서 얻어진 편향의 크기는 $a_1\sigma^2$이므로 모수 a_1의 정확한 추정은 매우 중요하다. 정확한 a_1 추정을 위해서는 가중치 w_i가 정확히 계산되어야 하며 충분한 세부 층 개수가 있어야 한다. 주어진 총 표본수 n을 다수의 세부 층으로 나누게 되면 세부 층 내의 표본 개수가 줄어들어 가중치 w_i가 정확히 계산되지 않는다. 반면 적은 개수로 세부 층을 나누면 식 (2.5)의 모형을 이용한 모수 추정 시에 자료의 개수가 줄어들게 되어 정확한 모수 추정이 불가능하게 된다. 따라서 이 내용을 모두 고려한 최적 세부 층 개수를 정하는 것이 매우 중요하다.

3. 비교된 모수 추정 방법

층화추출법은 표본조사에서 대표적으로 사용되는 표본 추출법이다. 층화추출법에서는 층별로 모수를 추정한 후 이를 결합하여 전체 모집단의 모수를 추정하기 때문에 본 연구에서는 주어진 한 개 층의 추정을 고려하였다. 또한 층화추출법에서 모평균 추정량 또는 모총합 추정량은 흔히 사용하는 공식을 사용하였다.

3.1. 최적 세부 층 경계

본 연구에서는 세 가지 방법으로 세부 층 경계를 구하였다. 이를 위해 표본 틀에 포함된 보조변수 x_i를 기준으로 (1) 등간격, (2) 분위수, (3) LH 알고리즘을 사용하였다.

(1) 등간격

표본 틀의 보조변수 x_i의 최대값 $x_{(n)}$과 최소값 $x_{(1)}$을 이용하여 $m = (x_{(n)} - x_{(1)})/L$을 구한 후 $x_{(1)} + km$, $k = 1, \ldots, L-1$로 세부 층 경계를 구한다.

(2) 분위수

표본 틀의 보조변수 x_i의 분포를 이용하여 분위수를 구하며 이 분위수를 세부 층 경계로 사용한다. 분위수는 SAS 또는 R에서 출력해 주기 때문에 쉽게 세부 층 경계를 구할 수 있다. 이제 i-분위수를 $z_{(i)}$라 하면 세부 층 경계는 다음과 같이 구한다.

$$(z_{(km)}, z_{((k+1)m)}), \quad m = \frac{100}{L},\ k = 0, \ldots, L-1.$$

(3) LH 알고리즘

LH 알고리즘은 층화추출에서 층 경계를 구하는 방법으로 흔히 절사표본설계의 층 경계 및 층 내 표본 개수를 구하기 위해 사용된다. 이 절에서는 Sim과 Shin (2014)의 내용을 수록하였다.

Lavallee와 Hidiroglou (1988)가 제안한 층화 알고리즘은 다음 식 (3.1)과 같이 전수층을 포함한 전체 층의 개수 H와 원하는 목표 변이계수(coefficient of variation, CV) c가 주어질 때 반복적 계산을 이용하여 주어진 기준을 만족시키면서 필요한 표본 규모가 최소가 되도록 층간 경계점을 찾는 방법이다. 여기서 총 표본 수 n은 W_h, S_h, 그리고 a_h의 함수이다.

$$n = N_H + \frac{\sum_{h=1}^{H} \frac{W_h^2 S_h^2}{a_h}}{c^2 \bar{X}^2 + \sum_{h=1}^{H} \frac{W_h S_h^2}{N}} \tag{3.1}$$

$n =$ 총 표본 수(total sample size)

$N =$ 모집단 수(total population size), $N = \sum_{h=1}^{H} N_h$

$N_h = h$층의 모집단 수

$S_h^2 = h$층의 모분산, $S_h^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (x_{hi} - \bar{X}_h)^2$, $\bar{X}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} x_{hi}$

$W_h = h$층의 가중치(stratum weight), $W_h = \frac{N_h}{N}$

$a_h = h$층의 표본배정 비율(sample allocation rate), $a_h = \frac{N_h S_h}{\sum_{h=1}^{H-1} N_h S_h}$

$\bar{X} =$ 전체평균, $\bar{X} = \sum_{h=1}^{H} W_h \bar{X}_h$

$c =$ 변이계수(coefficient of variation)

이 방법은 전체 표본 크기가 사후적으로 계산되며 목표 변이계수인 c가 작을수록 표본 수는 증가하고, 층의 개수가 증가할수록 전체 표본 수는 줄어든다. 각 층을 식 (3.2)와 같이 정의하고 보조변수 x_i에 대하여 n을 $H-1$개의 층간 경계점 $k = (k_1, \ldots, k_h, \ldots, k_{H-1})'$의 함수로 나타낼 때 최적값은 다음 식 (3.3)의 해로 구할 수 있다. 즉

$$U_h = \{i : k_{h-1} < x_i \leq k_h\} \tag{3.2}$$

이고 $k_1 < \cdots < k_h < \cdots < k_H - 1$, $k_0 = -\infty$, $k_H = \infty$이라 하면

$$\frac{\partial n(k)}{\partial_1} = \cdots = \frac{\partial n(k)}{\partial k_h} = \cdots = \frac{\partial n(k)}{\partial k_{H-1}} = 0 \tag{3.3}$$

을 만족하는 해를 구할 수 있다. 또한 식 (3.3)은 다시 식 (3.4)와 같은 k_h에 관한 2차식으로 표현할 수 있다.

$$\alpha_h k_h^2 + \beta_h k_h + \gamma_h = 0. \tag{3.4}$$

이제 초기값 $k^{(0)} = (k_1^{(0)}, \ldots, k_h^{(0)}, \ldots, k_{H-1}^{(0)})'$이 주어지면 식 (3.4)의 해를 반복적으로 구하게 되며 $k^{(r)}$, $r = 1, 2, \ldots$이 수렴할 때 얻어진 값을 사용하게 된다. 자세한 내용은 Lavallee와 Hidiroglou

(1988)와 Rivest (2002)를 참조하기 바란다. 또한 이 방법들에 관한 알고리즘과 R 코드는 Baillargeon과 Rivest (2011)을 참조하기 바란다. 특히 R에서는 주어진 총 표본 수에 따라 세부 층 개수와 최적의 세부 층 경계점을 제공한다. 그러나 본 연구에서처럼 다수의 세부 층을 생성할 경우에는 R 코드를 직접 사용하기가 어려운 경우도 있다. 예를 들면 25개 이상의 세부 층을 생성할 경우 R이 작동되지 않으며 매우 오랜 시간 돌아간다. 이에 본 연구에서는 모의실험 시간을 줄이기 위해 먼저 분위수로 큰 범위의 세부 층 경계를 나눈 후 나누어진 큰 범위의 세부 층을 LH 알고리즘을 이용하여 다시 세부 층으로 나눈 후 최종적인 세부 층을 구성하였다.

3.2. 사용된 추정량

만약 응답이 관심변수 y_i와 무관하다고 판단되면 층 내에 속한 모든 자료의 가중치는 동일하다고 판단할 수 있으며 층 평균은 자료의 단순평균으로 구할 수 있다. 그러나 이 방법은 관심변수가 보조변수의 함수지만 응답률이 관심변수와 관련이 없는 경우에는 세부 층을 이용한 방법에 비해 나쁜 결과를 주는 것으로 알려져 있다 (Chung과 Shin, 2017). 다음으로 응답률이 관심변수에 따라 다르다면 층 내의 각 자료에 다른 가중치를 사용할 수 있다. 이때 현실적으로 사용할 수 있는 방법이 층을 여러 개의 세부 층으로 나눈 후 세부 층 정보를 이용하여 평균을 추정하는 것이다. 즉 층화추출법을 이용할 때 사용하는 가중평균공식을 사용할 수 있다. 다음으로 편향을 제거한 편향 보정 추정량을 사용할 수 있다. 본 연구에서는 Chung과 Shin (2017)에서 사용한 다음의 세 추정량을 비교하였다. 또한 본 연구에서는 하나의 층만을 고려하기 때문에 N을 주어진 하나의 층의 모집단 수라 표시하였으며 세부 층의 수를 L, h세부 층의 모집단 수와 표본 수를 각각 N_h, n_h, h세부 층의 가중치를 w_h 그리고 h세부 층의 i번째 자료를 y_{hi}라 표시하였다.

(1) 단순 평균 추정량

주어진 층의 가중치가 세부 층에 무관하게 일정하기 때문에 $w_h = w = N/n$가 되어 다음의 수식이 얻어진다.

$$\hat{\bar{Y}}_s = \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{n_h} w y_{hi} = \frac{1}{n}\sum_{h=1}^{L}\sum_{i=1}^{n_h} y_{hi} = \bar{y}. \tag{3.5}$$

(2) 층화 추출 추정량

세부 층의 가중치가 다르기 때문에 다음의 평균 추정량을 사용한다.

$$\hat{\bar{Y}}_{st} = \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{n_h} w_h y_{hi}. \tag{3.6}$$

(3) 편향 보정 추정량

표본 자료에서 얻어진 회귀추정량 $\hat{\beta}_0$, $\hat{\beta}_1$과 $\hat{\sigma}^2$ 그리고 응답률 모형에서 얻어진 $\hat{a}_1$을 이용하여 다음 식 (3.7)의 추정량을 사용한다.

$$\hat{\bar{Y}}_{inf} = \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{n_h} w_h \left(\hat{\beta}_0 + \hat{\beta}_1 x_i - \hat{a}_1\hat{\sigma}^2\right). \tag{3.7}$$

따라서 식 (3.5)–(3.7)의 추정량과 3.1절에서 설명한 세부 층 경계를 구하는 방법인 등간격, 분위수 그리고 LH 알고리즘을 이용하여 모의실험을 실시하였다.

4. 모의실험

4.1. 모의실험 설계

본 연구에서는 층화추출법에서 주어진 하나의 층만을 고려하였다. 이는 각 층별로 모수 추정이 이뤄어지기 때문에 하나의 층을 고려하여도 일반성을 잃지 않기 때문이다. 또한 Chung과 Shin (2017)의 결과와 비교하기 위해 동일한 모의실험 방법을 사용하였다.

- Step 1: 모집단 생성과정

  초모집단모형이 정규분포인 경우의 정보적 표본설계를 위한 모집단 자료생성 과정은 이와 같다. 특히 보조변수는 종사자 수가 흔히 사용되므로 이를 고려하기 위해 균일 분포(uniform distribution)와 절단감마분포(truncated gamma distribution)를 고려한다.

  (1) 보조변수 x_i 생성: $x_i = 100 + \gamma_i$, $i = 1, \ldots, N$

  γ_i의 분포는 $\gamma_i \overset{\text{iid}}{\sim} \text{Unif}(100, 200)$과 $\text{TGamma}(1, 100)$을 사용하고 $\text{TGamma}(1, 100)$은 절단감마분포로 0과 100 사이의 값을 갖기 위해 100 이상인 값은 버린다. 따라서 보조변수 x_i는 100에서 200 사이의 값을 갖는다. 같은 방법으로 $\text{TGamma}(0.5, 100)$인 경우도 고려한다.

  (2) 초모집단모형 : $y_i = \beta_0 + \beta_1 x_i + \epsilon$, $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$

  여기서 $\beta_0 = 10, \beta_1 = 5, \sigma^2 = 400$과 모집단 자료 수 $N = 10{,}000$을 사용한다.

- Step 2 : 표본추출과정

  생성된 모집단에서 n개의 표본을 추출한다. 추출된 자료에서 랜덤으로 무응답을 만들었으며 응답률은 지수형을 따른다.

  (3) N개의 모집단 자료에서 단순임의추출(simple random sample)로 n개의 표본을 추출한다. 이때 $n = 50, 100, 200, 300, 500, 1000$으로 다양한 표본 수를 사용한다. 이는 표본 수에 따라 최적 세부층의 개수가 달라지기 때문이다.

  (4) 추출된 n개의 표본에서 $\pi_i = \exp(a_0 + a_1 y_i)$, $\pi_i \in [0, 1]$를 계산한다. y_i의 최소값에서의 응답률을 $\pi_y^{\min}$, y_i의 최대값에서의 응답률을 $\pi_y^{\max}$라 할 때, $(\pi_y^{\min}, \pi_y^{\max}) = (0.9, 0.7)$ 그리고 $(0.7, 0.9)$를 사용하여 a_0, a_1을 구하고 y_i에 따라 응답률을 계산한다. 또한 y_i의 응답률이 모두 같은 $\pi_i = 1$인 경우도 고려하다. 이와 관련된 내용은 Chung과 Shin (2017)을 살펴보기 바란다.

  (5) 응답한 최종 조사 자료는 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 되어 주어진 자료 수 n에 비해 약 20%가 감소한다.

- Step 3 : 층화

  얻어진 자료는 (x_i, y_i), $i = 1, \ldots, r$개이고 L개의 세부 층으로 층을 나눈다. 실제 자료 분석에서는 모집단에 보조변수 x_i의 정보만 있으므로 이를 기준으로 층을 나눈다.

  (6) 보조변수 x_i를 기준으로 등간격, 분위수 그리고 LH 알고리즘을 이용하여 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.

Table 4.1. Comparison results of $U(100, 200)$ with $r = 40$ and $(\pi_y^{\min}, \pi_y^{\max}) = (0.9, 0.7)$

L	Method	Bias			Abias			RMSE		
		$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{inf}$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{inf}$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{inf}$
3	Equal space	−7.310	−0.587	−0.487	19.144	6.872	6.863	24.085	8.580	8.576
	Percentile	−7.310	−0.493	−0.397	19.144	6.807	6.802	24.085	8.517	8.513
	LH	−7.310	−0.513	−0.409	19.144	6.778	6.795	24.085	8.525	8.530
4	Equal space	−7.310	−0.691	−0.561	19.144	5.551	5.516	24.085	7.420	7.386
	Percentile	−7.310	−0.750	−0.628	19.144	5.654	5.615	24.085	8.082	8.044
	LH	−7.310	−0.687	−0.565	19.144	5.537	5.510	24.085	7.428	7.403
5	Equal space	−7.310	−0.691	−0.589	19.144	5.038	4.966	24.085	9.359	9.314
	Percentile	−7.310	−0.694	−0.591	19.144	5.103	5.027	24.085	9.340	9.289
	LH	−7.310	−0.621	−0.526	19.144	4.992	4.927	24.085	8.853	8.805

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

(8) $\log(1/w_i) = a_0 + a_1 y_i + \eta_i$를 설정하고 단순회귀모형을 이용하여 모수 a_0, a_1을 추정한다. 여기서 오차는 등분산성을 가정한다.

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

(10) 계산된 결과를 이용하여 식 (3.5)에서 식 (3.7)의 $\hat{\bar{Y}}_s, \hat{\bar{Y}}_{st}, \hat{\bar{Y}}_{inf}$를 계산한다.

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

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

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

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

이때 사용된 반복수 $R = 3{,}000$이며 반복할 때마다 모집단을 새롭게 생성하였기 때문에 모집단 평균을 $\bar{Y}_r$로 표시하였다.

4.2. 모의실험 결과

보조변수의 분포가 균일분포 및 감마분포인 경우의 모의실험이 수행되었다. 특히 n = 50, 100, 200, 300, 500, 1000을 이용하여 모의실험을 수행하였으나 결과의 특징이 매우 유사하여 이 중에서 일부의 결과만을 수록하였다.

4.2.1. 보조변수가 균일분포일 경우 Tables 4.1–4.4는 보조변수가 균일분포인 경우의 결과이다. 다양한 표본 개수를 적용하여 모의실험이 수행되었으나 이 중에서 소표본에 해당되는 $r = 40$인 경우와 대표본인 $r = 400$인 경우의 결과만을 수록하였다. 결과를 살펴보면 등간격, 분위수 그리고 LH 알고리즘 등 어떤 방법을 사용하여도 최적 층 개수에는 큰 변화가 없으며 $r = 40$인 경우에는 $L = 4$가 $r = 400$인 경우에는 $L = 33$에서 결정되었다. 또한 응답률 $(\pi_y^{\min}, \pi_y^{\max}) = (0.9, 0.7)$과 $(\pi_y^{\min}, \pi_y^{\max}) =$

Table 4.2. Comparison results of $U(100, 200)$ with $r = 40$ and $(\pi_y^{\min}, \pi_y^{\max}) = (0.7, 0.9)$

L	Method	Bias			Abias			RMSE		
		$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{inf}$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{inf}$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{inf}$
3	Equal space	7.928	1.327	1.184	19.250	6.821	6.812	24.119	8.546	8.528
	Percentile	7.928	1.412	1.268	19.250	6.784	6.781	24.119	8.541	8.525
	LH	7.928	1.400	1.265	19.250	6.779	6.785	24.119	8.524	8.514
4	Equal space	7.928	0.532	0.395	19.250	5.544	5.522	24.119	7.386	7.361
	Percentile	7.928	0.469	0.325	19.250	5.633	5.601	24.119	8.007	7.976
	LH	7.928	0.538	0.396	19.250	5.539	5.522	24.119	7.390	7.373
5	Equal space	7.928	0.388	0.228	19.250	4.979	4.898	24.119	8.175	8.098
	Percentile	7.928	0.410	0.248	19.250	5.029	4.945	24.119	8.229	8.148
	LH	7.928	0.397	0.235	19.250	4.979	4.895	24.119	8.390	8.322

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

Table 4.3. Comparison results of $U(100, 200)$ with $r = 400$ and $(\pi_y^{\min}, \pi_y^{\max}) = (0.9, 0.7)$

L	Method	Bias			Abias			RMSE		
		$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{inf}$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{inf}$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{inf}$
30	Equal space	−7.623	−0.137	−0.005	8.708	0.842	0.812	10.444	1.052	1.011
	Percentile	−7.623	−0.134	−0.002	8.708	0.851	0.821	10.444	1.059	1.018
	LH	−7.623	−0.131	−0.001	8.708	0.848	0.814	10.444	1.057	1.015
33	Equal space	−7.623	−0.146	−0.005	8.708	0.847	0.806	10.444	1.056	1.005
	Percentile	−7.623	−0.142	−0.010	8.708	0.847	0.809	10.444	1.057	1.010
	LH	−7.623	−0.154	−0.020	8.708	0.851	0.815	10.444	1.153	1.109
35	Equal space	−7.623	−0.144	−0.008	8.708	0.848	0.809	10.444	1.060	1.009
	Percentile	−7.623	−0.149	−0.017	8.708	0.860	0.823	10.444	1.216	1.169
	LH	−7.623	−0.148	−0.016	8.708	0.854	0.815	10.444	1.169	1.119

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

Table 4.4. Comparison results of $U(100, 200)$ with $r = 400$ and $(\pi_y^{\min}, \pi_y^{\max}) = (0.7, 0.9)$

L	Method	Bias			Abias			RMSE		
		$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{inf}$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{inf}$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{inf}$
30	Equal space	7.767	0.162	0.018	8.756	0.845	0.801	10.567	1.052	1.004
	Percentile	7.767	0.162	0.020	8.756	0.855	0.809	10.567	1.060	1.010
	LH	7.767	0.169	0.024	8.756	0.848	0.803	10.567	1.059	1.008
33	Equal space	7.767	0.154	0.020	8.756	0.848	0.798	10.567	1.057	0.999
	Percentile	7.767	0.157	0.016	8.756	0.850	0.801	10.567	1.060	1.005
	LH	7.767	0.144	0.006	8.756	0.853	0.808	10.567	1.154	1.108
35	Equal space	7.767	0.147	0.011	8.756	0.854	0.804	10.567	1.088	1.036
	Percentile	7.767	0.142	0.002	8.756	0.874	0.821	10.567	1.245	1.196
	LH	7.767	0.153	0.010	8.756	0.862	0.811	10.567	1.165	1.112

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

$(0.7, 0.9)$인 두 경우에서 같은 최적 층 개수가 결정되었다. 세부 층 개수인 L에 따른 RMSE를 나타낸 Figure 4.1을 살펴보면 세 가지 방법 모두에서 RMSE가 매우 유사하다. 따라서 보조변수가 균일분포를 따르는 경우에는 어떤 방법을 사용해도 큰 문제는 없다. 그러나 세부 층 개수에 따른 RMSE 값은 매우 큰 차이를 보이고 있다. Figure 4.1을 살펴보면 $L \leq 6$에서는 RMSE가 일정 수준 유사하지

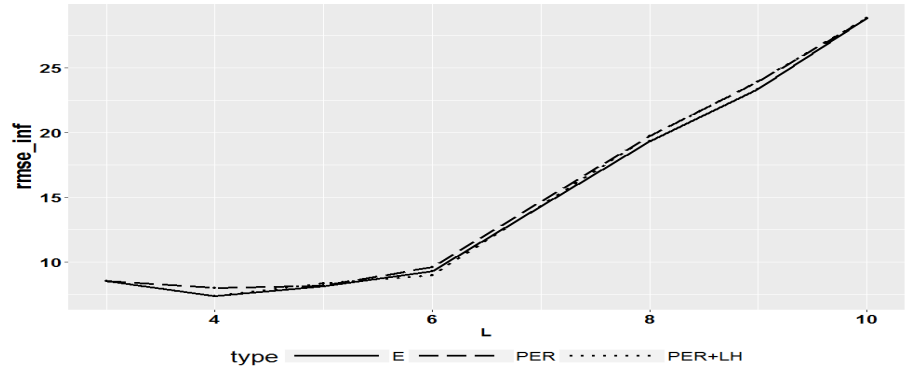

Figure 4.1. RMSE of $\hat{\bar{Y}}_{inf}$ with $U(100, 200)$, $r = 40$ and $(\pi_y^{\min}, \pi_y^{\max}) = (0.7, 0.9)$.

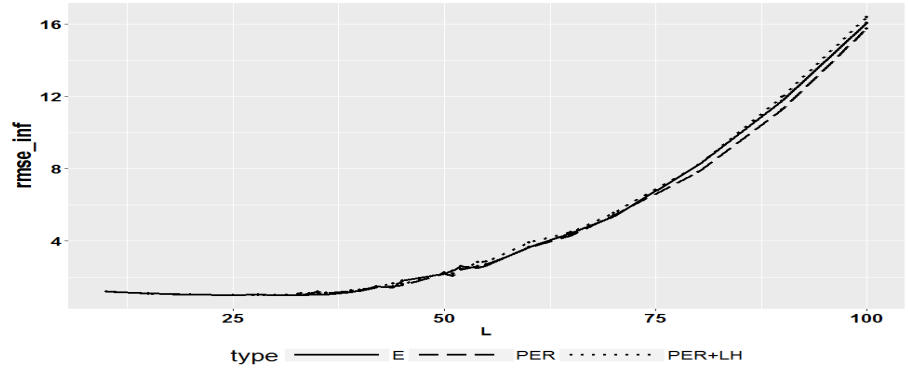

Figure 4.2. RMSE of $\hat{\bar{Y}}_{inf}$ with $U(100, 200)$, $r = 400$ and $(\pi_y^{\min}, \pi_y^{\max}) = (0.7, 0.9)$.

만 $L \geq 7$에서는 RMSE가 매우 빠르게 증가하는 것을 확인할 수 있다. 또한 $L = 10$ 또는 세부 층 내 표본 개수가 4개인 경우에서는 $\hat{\bar{Y}}_s$의 결과보다도 나빠진다. 이러한 결과는 Figure 4.2에서도 확인할 수 있다. 즉 Figure 4.2에서 $L \geq 50$인 경우에서 RMSE는 매우 크게 증가한다. 이와 같은 결과는 $n = 100, 200, 300, 1000$에서도 확인되었다. 결론적으로 최적 세부 층 개수를 사용하는 것은 매우 중요하며 세부 층의 개수가 적정 세부 층 개수에 비해 너무 많아지면 급격히 결과가 나빠지게 되므로 이 부분을 특히 주의해야 한다. 물론 그림에서는 잘 나타나지 않지만 최적 세부 층 개수에 비해 적은 세부 층 개수를 사용하게 되면 RMSE는 나빠진다.

4.2.2. 보조변수가 감마분포일 경우 보조변수의 분포가 감마분포인 경우에는 보조변수 값이 클 확률이 작기 때문에 이 경우에 얻어지는 자료의 개수는 적어진다. 따라서 등간격으로 층을 나누게 되면 보조변수 값이 큰 세부 층에 포함되는 자료 개수가 적어져 가중치의 변동성이 커질 수 있다. 따라서 보조변수의 분포를 고려하여 세부 층의 층 경계를 나누는 것이 타당할 수 있다. 감마분포인 경우에도 다양한 표본 개수를 이용하여 모의실험이 수행되었지만 Tables 4.5–4.8에는 Gamma$(1, 100)$이고 $r = 40$과 400인 결과를 수록하였다. 결과를 살펴보면 $L = 4$와 33에서 RMSE가 가장 작은 값을 갖는다. 이 결과는 등간격의 경우와 일치하는 것으로 보조변수의 분포에 크게 영향을 받지 않는 것으로 판단된다. 그러나 등간격인 경우에 비해 분위수 또는 LH 알고리즘을 사용하였을 때 RMSE 면에서 우수한 결과를 주는 것을 확인할 수 있다. 특히 Figure 4.3과 Figure 4.4를 살펴보면 등간격을 사용하지 않는 것이 타당

Table 4.5. Comparison results of Gamma$(1, 100)$ with $r = 40$ and $(\pi_y^{\min}, \pi_y^{\max}) = (0.9, 0.7)$

L	Method	Bias			Abias			RMSE		
		$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{inf}$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{inf}$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{inf}$
3	Equal space	−7.094	−0.862	−0.746	18.994	6.744	6.738	23.641	8.360	8.363
	Percentile	−7.094	−0.876	−0.741	18.994	7.005	6.996	23.641	8.762	8.739
	LH	−7.094	−0.767	−0.634	18.994	6.901	6.897	23.641	8.586	8.584
4	Equal space	−7.094	−0.959	−0.881	18.994	5.711	5.695	23.641	8.945	8.918
	Percentile	−7.094	−0.528	−0.390	18.994	5.733	5.711	23.641	7.953	7.932
	LH	−7.094	−0.501	−0.390	18.994	5.501	5.478	23.641	7.573	7.540
5	Equal space	−7.094	−1.554	−1.485	18.994	5.761	5.693	23.641	13.146	13.094
	Percentile	−7.094	−0.592	−0.457	18.994	5.017	5.000	23.641	8.011	7.980
	LH	−7.094	−0.980	−0.869	18.994	5.248	5.181	23.641	10.576	10.516

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

Table 4.6. Comparison results of Gamma$(1, 100)$ with $r = 40$ and $(\pi_y^{\min}, \pi_y^{\max}) = (0.7, 0.9)$

L	Method	Bias			Abias			RMSE		
		$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{inf}$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{inf}$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{inf}$
3	Equal space	7.490	1.075	0.940	19.460	6.890	6.886	24.300	8.542	8.541
	Percentile	7.490	1.083	0.974	19.460	7.074	7.067	24.300	9.717	9.706
	LH	7.490	1.127	1.009	19.460	6.870	6.868	24.300	8.610	8.610
4	Equal space	7.490	0.261	0.080	19.460	5.732	5.709	24.300	8.834	8.806
	Percentile	7.490	0.765	0.650	19.460	5.756	5.727	24.300	7.957	7.942
	LH	7.490	0.774	0.632	19.460	5.498	5.477	24.300	6.854	6.810
5	Equal space	7.490	−0.237	−0.438	19.460	5.387	5.307	24.300	10.956	10.911
	Percentile	7.490	0.251	0.143	19.460	5.020	4.992	24.300	8.521	8.482
	LH	7.490	−0.027	−0.176	19.460	5.160	5.110	24.300	9.840	9.791

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

Table 4.7. Comparison results of Gamma$(1, 100)$ with $r = 40$ and $(\pi_y^{\min}, \pi_y^{\max}) = (0.9, 0.7)$

L	Method	Bias			Abias			RMSE		
		$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{inf}$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{inf}$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{inf}$
32	Equal space	−7.317	−0.229	−0.141	8.279	0.927	0.884	10.065	1.547	1.503
	Percentile	−7.317	−0.192	−0.061	8.279	0.880	0.838	10.065	1.294	1.250
	LH	−7.317	−0.175	−0.049	8.279	0.872	0.829	10.065	1.178	1.133
33	Equal space	−7.317	−0.284	−0.186	8.279	0.979	0.928	10.065	1.804	1.755
	Percentile	−7.317	−0.175	−0.042	8.279	0.864	0.817	10.065	1.085	1.028
	LH	−7.317	−0.169	−0.042	8.279	0.867	0.821	10.065	1.087	1.034
35	Equal space	−7.317	−0.313	−0.222	8.279	0.999	0.954	10.065	1.902	1.863
	Percentile	−7.317	−0.184	−0.051	8.279	0.879	0.826	10.065	1.168	1.109
	LH	−7.317	−0.188	−0.058	8.279	0.881	0.832	10.065	1.199	1.143

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

한 것을 확인할 수 있다. 또한 Tables 4.9–4.12에는 Gamma$(0.5, 100)$이고 $r = 40, 400$의 결과가 수록되어 있다. 이 결과를 살펴보면 분위수를 사용한 경우가 다른 층 경계점을 사용한 결과에 비해 매우 우수한 것을 확인할 수 있다. Figure 4.5와 Figure 4.6을 보면 이를 더욱 확실하게 확인할 수 있다. 반면 최적의 세부 층 개수는 다른 두 분포 결과와 매우 유사하다. 이 결과에서도 세부 층의 개수가 너무 많아

Table 4.8. Comparison results of Gamma(1, 100) with $r = 40$ and $(\pi_y^{\min}, \pi_y^{\max}) = (0.7, 0.9)$

L	Method	Bias			Abias			RMSE		
		$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{inf}$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{inf}$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{inf}$
32	Equal space	7.312	0.121	−0.061	8.531	0.881	0.843	10.262	1.177	1.140
	Percentile	7.312	0.120	−0.019	8.531	0.865	0.831	10.262	1.082	1.040
	LH	7.312	0.126	−0.018	8.531	0.868	0.830	10.262	1.085	1.042
33	Equal space	7.312	0.108	−0.066	8.531	0.889	0.850	10.262	1.218	1.174
	Percentile	7.312	0.108	−0.027	8.531	0.873	0.834	10.262	1.133	1.089
	LH	7.312	0.121	−0.020	8.531	0.871	0.829	10.262	1.084	1.037
35	Equal space	7.312	0.084	−0.088	8.531	0.894	0.857	10.262	1.296	1.257
	Percentile	7.312	0.100	−0.038	8.531	0.893	0.848	10.262	1.244	1.194
	LH	7.312	0.096	−0.045	8.531	0.891	0.852	10.262	1.242	1.197

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

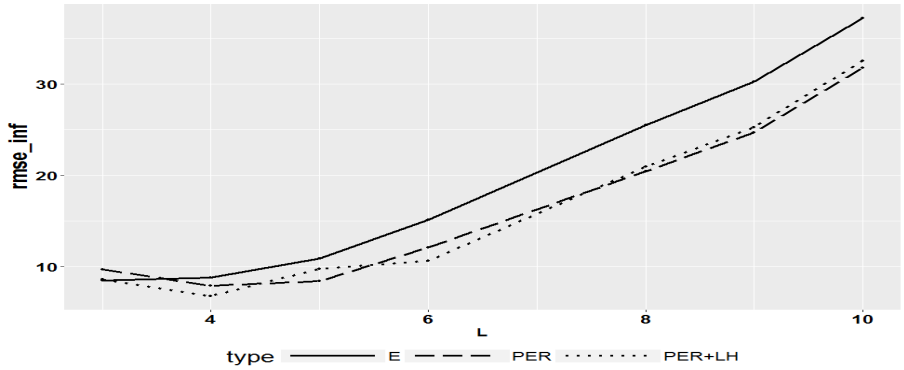

Figure 4.3. RMSE of $\hat{\bar{Y}}_{inf}$ with Gamma(1, 100), $r = 40$ and $(\pi_y^{\min}, \pi_y^{\max}) = (0.7, 0.9)$

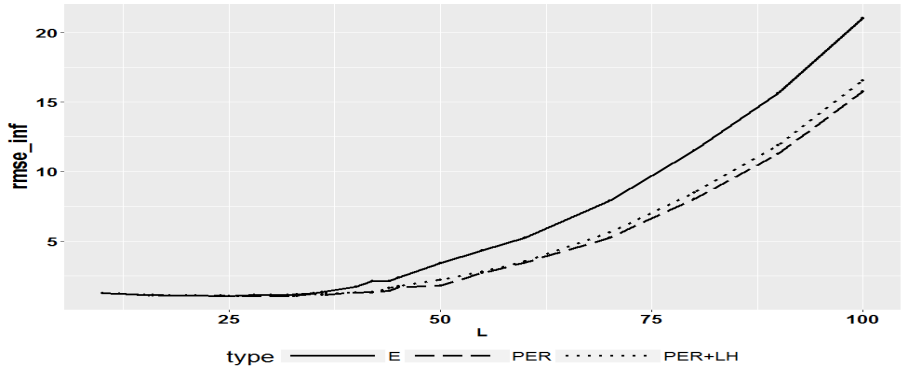

Figure 4.4. RMSE of $\hat{\bar{Y}}_{inf}$ with Gamma(1, 100), $r = 400$ and $(\pi_y^{\min}, \pi_y^{\max}) = (0.7, 0.9)$

지면 결과가 급격히 나빠지는 것을 확인할 수 있다.

4.3. 최적 표본 수

4.2절에는 다양한 표본 개수 결과 중 일부만을 수록하였다. 다음의 Table 4.13은 RMSE를 최소로 하는 최적 세부 층 개수와 평균 최적 세부 층 개수 그리고 평균 최적 세부 층 표본 수를 정리한 표이다. 전체

Table 4.9. Comparison results of Gamma$(0.5, 100)$ with $r = 40$ and $(\pi_y^{\min}, \pi_y^{\max}) = (0.9, 0.7)$

L	Method	Bias			Abias			RMSE		
		$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{inf}$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{inf}$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{inf}$
4	Equal space	−5.832	−5.213	−5.176	16.847	9.817	9.786	21.042	20.260	20.212
	Percentile	−5.832	−0.821	−0.646	16.847	6.668	6.630	21.042	8.846	8.804
	LH	−5.832	−1.303	−1.177	16.847	5.870	5.804	21.042	11.088	11.022
5	Equal space	−5.832	−9.802	−9.850	16.847	13.333	13.310	21.042	25.678	25.678
	Percentile	−5.832	−0.473	−0.282	16.847	5.804	5.749	21.042	7.395	7.311
	LH	−5.832	−2.339	−2.237	16.847	6.590	6.499	21.042	15.369	15.307
6	Equal space	−5.832	−15.588	−15.661	16.847	18.326	18.277	21.042	31.161	31.139
	Percentile	−5.832	−0.858	−0.681	16.847	5.538	5.454	21.042	9.230	9.159
	LH	−5.832	−1.830	−1.663	16.847	6.238	6.123	21.042	13.495	13.403

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

Table 4.10. Comparison results of Gamma$(0.5, 100)$ with $r = 40$ and $(\pi_y^{\min}, \pi_y^{\max}) = (0.7, 0.9)$

L	Method	Bias			Abias			RMSE		
		$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{inf}$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{inf}$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{inf}$
4	Equal space	6.985	−2.533	−2.742	18.428	8.634	8.613	23.386	17.206	17.198
	Percentile	6.985	0.994	0.928	18.428	6.681	6.633	23.386	8.864	8.820
	LH	6.985	0.087	−0.050	18.428	5.596	5.533	23.386	8.916	8.839
5	Equal space	6.985	−6.333	−6.633	18.428	11.076	11.044	23.386	21.786	21.842
	Percentile	6.985	0.790	0.712	18.428	5.839	5.779	23.386	7.653	7.560
	LH	6.985	−0.828	−0.993	18.428	5.978	5.892	23.386	12.992	12.939
6	Equal space	6.985	−11.740	−12.058	18.428	15.416	15.406	23.386	27.403	27.473
	Percentile	6.985	−0.116	−0.211	18.428	5.782	5.661	23.386	10.207	10.128
	LH	6.985	−0.850	−0.935	18.428	6.131	6.001	23.386	12.391	12.303

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

Table 4.11. Comparison results of Gamma$(0.5, 100)$ with $r = 400$ and $(\pi_y^{\min}, \pi_y^{\max}) = (0.9, 0.7)$

L	Method	Bias			Abias			RMSE		
		$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{inf}$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{inf}$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{inf}$
4	Equal space	−6.382	−3.253	−3.236	7.460	3.732	3.698	8.962	6.073	6.045
	Percentile	−6.382	−0.161	−0.011	7.460	0.860	0.822	8.962	1.077	1.033
	LH	−6.382	−0.166	−0.019	7.460	0.874	0.833	8.962	1.171	1.124
5	Equal space	−6.382	−3.933	−3.925	7.460	4.378	4.344	8.962	6.780	6.756
	Percentile	−6.382	−0.167	−0.018	7.460	0.865	0.821	8.962	1.081	1.029
	LH	−6.382	−0.179	−0.033	7.460	0.869	0.825	8.962	1.191	1.140
6	Equal space	−6.382	−4.434	−4.429	7.460	4.846	4.820	8.962	7.254	7.232
	Percentile	−6.382	−0.159	−0.024	7.460	0.872	0.826	8.962	1.121	1.064
	LH	−6.382	−0.183	−0.041	7.460	0.876	0.829	8.962	1.215	1.162

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

적으로 보조변수의 분포는 크게 영향을 주지 않는다. 이는 층 경계점을 구하는 방법으로 보조변수의 분위수를 사용하였기 때문인 것으로 판단된다. 최소 세부 층 개수는 4개 이상이고 전체적으로 세부 층 내에 9개에서 13개 정도의 세부 층 표본 수를 갖도록 한다면 실제 자료 분석에서 우수한 결과를 얻을 수 있을 것으로 판단된다. 결론적으로 r이 100개 이내인 경우에는 세부 층에 9–10개 정도의 자료가 포함

Table 4.12. Comparison results of Gamma(0.5, 100) with $r = 400$ and $(\pi_y^{\min}, \pi_y^{\max}) = (0.7, 0.9)$

L	Method	Bias			Abias			RMSE		
		$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{inf}$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{inf}$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{inf}$
4	Equal space	6.574	−1.819	−2.074	7.784	2.638	2.660	9.511	4.616	4.707
	Percentile	6.574	0.151	0.024	7.784	0.900	0.855	9.511	1.129	1.077
	LH	6.574	0.148	0.017	7.784	0.909	0.864	9.511	1.213	1.162
5	Equal space	6.574	−2.281	−2.543	7.784	3.054	3.086	9.511	5.129	5.229
	Percentile	6.574	0.140	0.016	7.784	0.900	0.852	9.511	1.129	1.074
	LH	6.574	0.126	−0.003	7.784	0.904	0.865	9.511	1.243	1.198
6	Equal space	6.574	−2.617	−2.882	7.784	3.346	3.399	9.511	5.492	5.608
	Percentile	6.574	0.144	0.006	7.784	0.916	0.860	9.511	1.203	1.137
	LH	6.574	0.117	−0.016	7.784	0.923	0.870	9.511	1.302	1.247

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

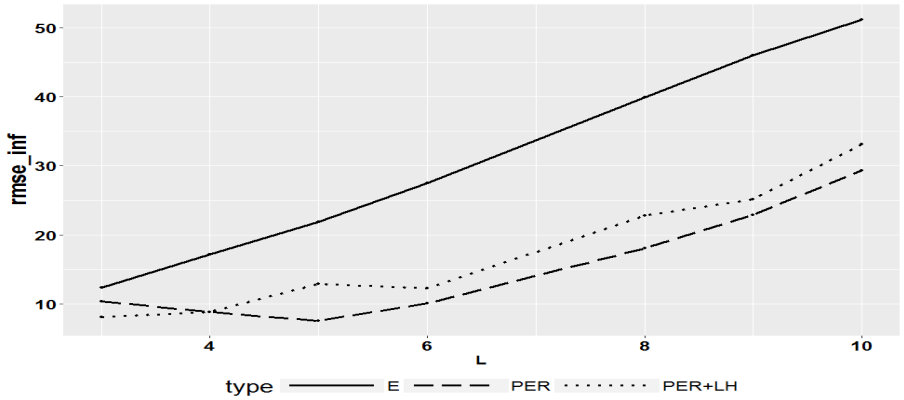

Figure 4.5. RMSE of $\hat{\bar{Y}}_{inf}$ with Gamma(0.5, 100), $r = 40$ and $(\pi_y^{\min}, \pi_y^{\max}) = (0.7, 0.9)$.

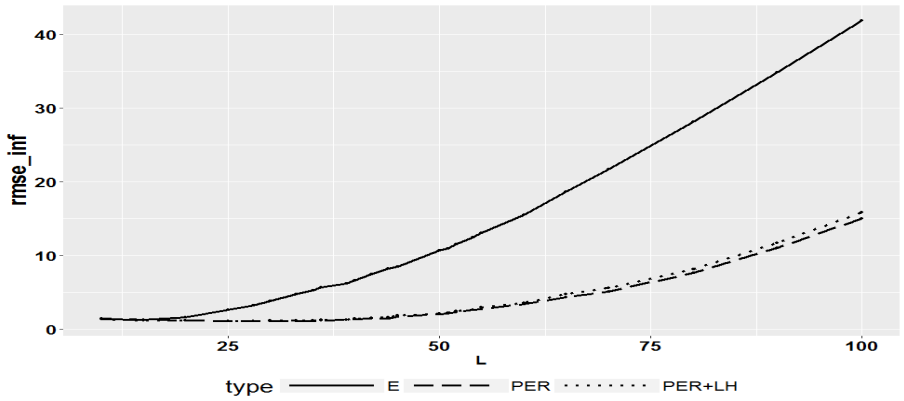

Figure 4.6. RMSE of $\hat{\bar{Y}}_{inf}$ with Gamma(0.5, 100), $r = 400$ and $(\pi_y^{\min}, \pi_y^{\max}) = (0.7, 0.9)$.

되도록 하고 100개 이상에서는 12–13개 정도의 자료가 세부 층에 포함되도록 세부 층의 개수를 정하면 큰 문제가 없을 것으로 판단된다. 물론 4.1절과 4.2절에서도 언급하였듯이 세부 층의 개수가 너무 많아 세부 층에 포함된 표본 수가 적은 경우에는 결과가 매우 나빠진다.

Table 4.13. Optimal sample size and optimal number of substrata

distribution	40	50	80	100	160	200	240	300	400	500	800
$U(100, 200)$	4	6	9	10	15	18	20	27	33	40	57
Gamma$(1, 100)$	4	5	7	10	12	12	18	25	33	36	54
Gamma$(0.5, 100)$	5	6	8	10	15	16	20	27	35	36	70
Average number of substrata	4.3	5.7	8	10	14.0	15.3	19.3	26.3	33.7	37.3	60.3
Mean of sample size	9.3	8.8	10	10	11.4	13.1	12.4	11.4	11.9	13.4	13.3

5. 결론

최근 응답률이 관심변수의 지수함수로 얻어질 때 발생하는 편향을 감소 또는 제거하는 방법이 연구되었다. 이 방법에서 편향의 크기에 관련이 있는 모수를 정확히 추정하기 위해서는 최적의 세부 층 경계와 세부 층 개수를 사용하는 것이 중요하다. 이에 본 연구에서는 실제 자료 분석에서 사용할 수 있는 세부 층 경계와 세부 층 개수를 정하는 방법을 제안하였다. 모의실험 결과 세부 층 경계를 구하는 방법으로 분위수를 사용하는 것이 타당하며 특히 보조변수의 분포가 감마분포와 같이 오른쪽으로 긴 꼬리가 있는 경우에 매우 효과적이다. 또한 최적 세부 층 개수는 주어진 층의 표본 규모에 큰 영향을 받으며 전체적으로 세부 층내 자료 수는 9–13개 정도가 적당한 것으로 나타났다. 물론 모든 경우를 고려하지 않고 또한 이론적으로 나온 결과가 아니기 때문에 이 결과를 일반화하는 것에는 신중할 필요가 있지만 실질적인 자료분석에 충분히 적용될 수 있을 것으로 판단된다.

References

Baillargeon, S. and Rivest L.-P. (2011). The construction of stratified designs in R with the package stratification, *Survey Methodology*, **37**, 53–65.

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.

Lavallee, P. and Hidiroglou, M. (1988). On the stratification of skewed populations, *Survey Methodology*, **14**, 33–43.

Lee, Y. and Shin, K. I. (2017). A study on sensitivity of representativeness indicator in survey sampling, *The Korean Journal of Applied Statistics*, **30**, 69–82.

Pfeffermann, D. Krieger, A. M., and Rinott, Y. (1998). Parametric distributions of complex survey data under informative probability sampling, —em 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.

Rivest, L. P. (2002). A generalization of Lavallee and Hidiroglou algorithm for stratifications in business survey, *Survey Methodology*, **28**, 191–198.

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

Sim, S. H. and Shin, K. I. (2014). A composite estimate for cut-off sampling using cost function, *The Korean Journal of Applied Statistics*, **27**, 43–59.

모의실험을 기반으로 지수형 응답률 보정을 위한 세부 층 결정에 관한 연구

민주원[a] · 신기일[a,1]

[a]한국외국어대학교 통계학과

(2018년 7월 23일 접수, 2018년 8월 17일 수정, 2018년 9월 9일 채택)

요 약

정보적 표본설계 기법을 적용하여 무응답의 영향을 줄이기 위한 연구가 진행되고 있다. 특히 초모집단모형(super population model)에 포함된 오차의 분포가 정규분포를 따르고 응답률이 지수함수를 따를 때 지수형 응답률 정보를 모수추정에 사용함으로써 추정의 정확성이 향상되는 것으로 알려져 있다. 최근 Chung과 Shin (2017)은 정보적 표본설계의 가중치를 구하기 위해 세부 층을 등간격으로 나누는 방법을 고려하였으며 세부 층의 개수가 추정의 정확성에 영향을 주는 것을 확인하였다. 이에 본 연구에서는 주어진 표본 규모에 따른 최적의 세부 층 개수와 최적의 층 경계를 구하기 위해 등간격, 분위수, LH 알고리즘을 이용하여 층을 나누는 방법을 살펴보았으며 모의실험을 통하여 각 방법의 결과를 비교하였다. 또한 다양한 형태의 보조변수 분포를 이용하여 실무에서 사용할 수 있는 세부 층 경계와 세부 층 개수를 정하는 기준을 제안하였다.

주요용어: LH 알고리즘, 층화추출, 회귀추정량, 표본 포함확률, 초모집단모형

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

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