Calibration Estimator for Randomized Response Model of Quantitative Attribute

손창균(동국대),신재동(한국보건사회연구원)

전북대학교

2021.05.27

1. Introduction

- Warner (1965) first proposed a randomized response model to estimate the proportion of sensitive attributes through randomization devices to protect the respondents' identity or privacy.
- Greenberg et al. (1971) proposed an unrelated question model and extended it to cases with quantitative attributes, after which Eriksson (1973), Poole (1974), and Pollock and Bek (1976). Loynes (1976) proposed a forced answer technique that forces answer "yes" instead of sensitive and mutually exclusive question of the Warner model.
- Carr et al. (1982) used a conditional randomization device to obtain sensitive information by using Loynes' forced answered model only to respondents who answered "yes" through a randomization device consisting of forced and less sensitive question.

1. Introduction

- On the other hand, the RRM considers an estimator under the assumption of true response and complete response, but there is a possibility that non response may occurs in the actual survey field.
- The nonrepresentative problem caused by this nonresponse is a direct cause of nonresponse bias. To overcome for this problem, we can consider the weighting adjustment or sample replacement.
- Weighting adjustment is usually applied in two step, in first step it is being non-response adjustment constructing the nonresponse homogeneous group and in secod step calibration using population information.

1. Introduction

- Diana and Perri (2009) proposed the calibraion estimator for a various quantitative attribute randomized response models.
- However, these studies did not consider the level of available auxiliary information, and if the level of auxiliary information is available at the population and the sample level, the precision of estimator can be improved by calibrated it according to the level of auxiliary information.
- In this study, we apply the calibration method to increase the level of estimator in estimating sensitive attribute using the RRM of quantitative unrelated attributes.

2. RRM for Quantitative Unrelated Attribute

- The study of quantitative attribute, such as the number of drug abuse or amount of tax evasion, begins with Greenberg et al. (1971). They proposed an unrelated question model and extended it to a case with quantitative attribute, suggesting a RRM that can obtain information about quantitative sensitive attribute.
- In estimating unknown population mean of sensitive variables, Greenberg et al., which extends the unrelated question technique to quantitative attributes, are as follows.

2. RRM for Quantitative Unrelated Attribute

- For a sensitive question, suppose that a sensitive variable X has a continuous density function $g(\cdot)$, and let Y be a unrelated attribute variable with a density function $h(\cdot)$, where the forms of the probability function of the sensitive and unrelated attribute variables of each individual are both continuous (or discrete) and the expected value should be exist in either case.
- Respondents answer to the selected variable through a randomization device that has a selection probability p of a sensitive variable X and $1 - p$, a probability of selecting an unrelated variable Y.

- If the ith respondent answers Z_i, then Z_i has the probability density function as follows

$$f(Z_i) = pg(Z_i) + qh(Z_i) \ , \ \ i = 1, 2, , N. \tag{1}$$

- The population mean of Z is given by

$$\mu_Z = p\mu_X + q\mu_Y. \tag{2}$$

- The population variance of Z is

$$\sigma_Z^2 = p\sigma_X^2 + q\sigma_Y^2 + pq(\mu_X - \mu_Y)^2. \tag{3}$$

- Let $Z_1, Z_2, \cdots, Z_n$ are the answers of the respondents with size n for the simple random sample from a population with size N, then the sample mean and variance of responds are given by

$$\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i, \tag{4}$$

$$s_Z^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(Z_i - \bar{Z})^2. \tag{5}$$

- An unbiased estimator $\hat{\mu}_X$ of μ_X and its variance are respectively

$$\hat{\mu}_X = \frac{\bar{Z} - q\mu_Y}{p}, \tag{6}$$

$$V(\hat{\mu}_X) = \frac{\sigma_Z^2}{np^2} \tag{7}$$

where μ_Y assumed to be known.

3. Two-Step Calibration Method

- Auxiliary information for two-step calibration are (1) observation values in the sample and (2) the values of frame in the given popualtion. That is, auxiliary information are being two levels population and sample.
- The auxiliary vector $x_k = (x_{1k}^{'}, x_{2k}^{'})^{'}$, where x_{1k} be the vector of J_1 values known for all $k \in U$, and x_{2k} is the vector of $J_2 = J - J_1$ values recorded by observation of units k in the sample only.
- Let's define inclusion probabilities v_k and v_{kl} as following, respectively

$$v_k = \sum_{k \in s} p(s), \quad v_{kl} = \sum_{kl \in s} p(s), \tag{8}$$

- Let the variable of interest is y_k then the estimate of the population total is defined as

$$\hat{\tau}_y = \sum_{k \in s} \frac{y_k}{v_k} = \sum_{k \in s} d_k y_k. \tag{9}$$

3. Two-Step Calibration Method

- Basically, the calibration improves the original sampling weight which has noncoverage or nonresponse bias due to population change or refusal to response using auxiliary information in population or sample level.
- We calibrate the original sampling weights d_k to improve the original estimator. Lw w_k be the new weight via calibration procedure proposed by Deville and Sarndal (1992).
- In two-step calibration, a distance function $G(\dot{})$ is minimized by new weight w_{1k} in first-step calibration for the population level

$$G(w_{1k}, d_{1k}) = \sum_{s \in k} \frac{(w_{1k} - d_{1k})^2}{2d_{1k}} \tag{10}$$

subject to the calbration equation

$$\sum_{s \in k} w_{1k} x_{1k} = \sum_{U} x_{1k} = \tau_{x1} \tag{11}$$

3. Two-Step Calibration Method

- The calibrated weights are given by

$$w_{1k} d_{1k} (1 + x_{1k}^{'} \lambda_1) = d_{1k} F(x_{1k}^{'} \lambda_1) \tag{12}$$

  where $F() = (\partial G / \partial w)^{-1}$, and λ_1 is unknown vector value of Lagrange multiplier in the sample.
- We also use the initial weights $w_{1k} d_{2k}$ at second-step calibration, so that it is reasonable to use for second-step weight because the final weight depends on the first-step weight.
- The final (second-phase) weight w_k is determined by minimizing

$$G(w_k, d_k) = \sum_{s \in k} \frac{(w_k - w_{1k} d_{2k})^2}{2 w_{1k} d_{2k}} \tag{13}$$

  subject to the calbration equation

$$\sum_{s \in k} w_k x_k = \sum_{s} w_{1k} x_k \tag{14}$$

3. Two-Step Calibration Method

- The final calibraion weight is given using the two-step calibration procedure as above

$$\begin{aligned} w_k &= w_{1k} d_{2k} (1 + x_k^{'} \lambda_2) \\ &= d_{1k} d_{2k} (1 + x_{1k}^{'} \lambda_1)(1 + x_k^{'} \lambda_2) \\ &= d_k F(x_{1k}^{'} \lambda_1) F(x_k^{'} \lambda_2) \end{aligned} \tag{15}$$

- For the two-step calibration, we can define the function F in the first-step, by assumption SRSWOR and the auxilairy variabe $x_{1k} = 1$, then

$$\begin{aligned} F(x_{1k}^{'} \lambda_1) &= 1 + x_{1k}^{'} \lambda_1 \\ &= 1 + \left[\sum_s d_{1k} x_{1k} x_{1k}^{'} \right]^{-1} \left[\tau_{1x} - \hat{\tau}_{1x} \right] x_{1k} \\ &= 1 + \left[\sum_s d_{1k} \right]^{-1} \left[N - \sum_s d_{1k} \right] = N / \hat{N}_1 . \end{aligned} \tag{16}$$

where $\hat{N}_1 = \sum_s d_{1k}$.

- We can obtain the function of λ_2, $F(x_k^{'}\lambda_2)$ as follows

$$\begin{aligned} F(x_k^{'}\lambda_2) &= 1 + x_k^{'}\lambda_2 \\ &= 1 + \left[\sum_s w_{1k} d_{2k} x_{1k} x_{1k}^{'}\right]^{-1} [\hat{\tau}_{1x} - \hat{\tau}_{2x}] x_k \end{aligned} \tag{17}$$

- If the auxiliary variable for the second-step is $x_k = x_{1k} = x_{2k} = 1$ with SRSWOR, then the $F(x_k^{'}\lambda_2)$ becomes

$$F(\lambda_2) = 1 + \hat{N}_1 \left(\frac{1}{\hat{N}_2} - \frac{1}{N}\right) \tag{18}$$

where $\hat{N}_2 = \sum_s d_k$.

4. Calibration estimator for Randomized Response Model of Quantitative Attribute

- The RRM for quantitaive unrelated attribute described as Section 2 can be rewitten as follows

$$\begin{aligned} \hat{\mu}_X &= \frac{\bar{Z} - q\mu_Y}{p}, \\ &= \frac{1}{np}\sum_{k=1}^{n} Z_k - \frac{q\mu_y}{p} \\ &= \frac{1}{Np}\sum_{k=1}^{n} d_k Z_k - \frac{q\mu_y}{p} \end{aligned} \tag{19}$$

4. Calibration estimator for Randomized Response Model of Quantitative Attribute

- Then the two-stp calibrated RR estimator for quantitaive unrelated attribute

$$\hat{\mu}_{Xcal} = \frac{1}{Np}\sum_{s} d_k F(x_{1k}'\lambda_1)F(x_k'\lambda_2)Z_k - \frac{q\mu_y}{p} \quad (20)$$

where the final calibrated weight w_k is

$$\begin{aligned} w_k &= d_k F(x_{1k}'\lambda_1)F(x_k'\lambda_2) \\ &= d_k N\left(\frac{1}{\hat{N}_1} + \frac{1}{\hat{N}_2} - \frac{1}{N}\right) \end{aligned} \quad (21)$$

4. Calibration estimator for Randomized Response Model of Quantitative Attribute

- The variance of calibrated RR estimator is

$$V(\hat{\mu}_{Xcal}) = \sum_{k} \sum_{l} (w_k w_l - w_{kl}) (w_k Z_k) (w_l Z_l) \qquad (22)$$

5. Efficiency Comparison

- The relative efficiecy (RE) is defined as follows

$$RE(\hat{\mu}_{Xcal}|\hat{\mu}_X) = \frac{V(\hat{\mu}_X)}{V(\hat{\mu}_{Xcal})} \tag{23}$$

- Also we generate the artificial population by the correlation between Z_k and x_k with size of $N = 10,000$

$$Z_k = 30 + \sqrt{100(1-\rho^2)}x_k + \rho\sqrt{100}y_k, \tag{24}$$

  where we assume the correlation $\rho = 0.2, 0.7$ respctively.
- The selection probability of question in RRM $p = 0.1$ to 0.9 by 0.1, we compare the efficiency of the proposed estimator.

5. Efficiency Comparison

<Table 1> RE with $n = 500$, $n = 1,000$ and $n = 3,000$

p	n=500		n=1,000		n=3,000	
	$\rho = 0.2$	$\rho = 0.7$	$\rho = 0.2$	$\rho = 0.7$	$\rho = 0.2$	$\rho = 0.7$
0.1	1.0623	1.0624	1.1314	1.1319	1.507	1.5086
0.2	1.0728	1.073	1.1537	1.154	1.5928	1.5941
0.3	1.0834	1.0834	1.176	1.1762	1.6789	1.6798
0.4	1.0939	1.094	1.1983	1.1985	1.765	1.7655
0.5	1.1045	1.1047	1.2206	1.2207	1.8511	1.8513
0.6	1.115	1.1154	1.2429	1.2429	1.9373	1.937
0.7	1.1257	1.1263	1.2654	1.2651	2.0235	2.0226
0.8	1.1361	1.1354	1.2878	1.2874	2.1097	2.1083
0.9	1.1465	1.1454	1.3105	1.3097	2.1958	2.1944

6. Conclusions

- We find the two-step calibration estimator for RRM of quantiative unrealted attributed by assumption of levels of auxiliary information.
- The numerical comparison shows the proposed estimator is more efficient than the exsited RR estimator.
- We will consider the sampling scehme as two-phase sampling for RRM unrelated quntitative attribute.