The calibration for stratified randomized response model

Chang-Kyoon Son[1], Ki-Hak Hong[2], and Gi-Sung Lee[3]

Abstract

This paper proposes the calibration procedure for stratified Warner's randomized response model, which suggested by Kim and Warde (2004). It is shown that the proposed calibration estimator is more efficient than the Kim and Warde's model.

Keywords: Stratified RR model, calibration, auxiliary variable.

1. Introduction

The randomized response (RR) technique suggested by Warner (1965) that minimizes underreporting of a data related to a socially undesirable or incriminating behavior questions. In RR technique, each individual respondent is provided with a randomization device by which he/she chooses one of the two questions "Do you belong to sensitive group A?" or "Do you belong to sensitive group A^c ?" with respective probabilities P and $(1-P)$ and replies "Yes" or "No" to the question chosen.

Mangat and Singh (1990) proposed a two-stage RR model that is extended the Warner model. Mangat (1994) proposed another RR model which has benefit of simplicity over that of Mangat and Singh (1990). Hong, et al. (1994) suggested a stratified RR model that applied the same randomization device to every stratum. In general, the stratified random sampling is obtained by dividing the population into non-overlapping groups called strata and selecting a simple random sample from each stratum. An RR technique using a stratified random sampling gives the group characteristics related to each stratum estimator. Also, stratified samples protect a researcher from the possibility of obtaining a poor sample.

Hong, et al. (1994) assumed the proportional sampling for a stratified sampling, whereas Kim and Warde (2004) extended the Hong, et al. model to the optimal sampling and each stratum sample provides different randomization devices. Kim and Warde (2004) showed that a stratified RR technique using an optimal allocation which is more efficient than that of using a proportional allocation. In relation to the precision of estimators of population mean or total, the statisticians are used to the generalized linear regression (GREG) estimator. Using the GREG estimator studied by Fuller (1975), Cassel, Sarndal and Wretman (1976), Isaki and Fuller(1982), and Wright (1983), it is possible to improve a posteriori, the estimate of a total of a variable of interest on the basis of auxiliary variables for which additional information is available. Deville and Sarndal(1992) and Deville, Sarndal and Sautory(1993) proposed a class of estimators derived from a re-weighting approach that addresses the same issue of variance

[1] Lecture Professor, Department of Liberal Art and Culture, Hyupsung University, Kyunggi-do.
[2] Professor, Department of Computer Science, Dongshin University, Jeonnam.
[3] Professor, Department of e-information Science, Woosuk University, Jeonbuk.

reduction called the calibration estimators.

In this paper, the problem of variance reduction of a stratified RR estimator has been considered the calibration for stratum weight of using auxiliary information.

2. Stratified Randomized Response Techniques

Let the population is divided by non-overlapping strata with a priori, a sample is a selected by simple random sampling with replacement in each stratum. Also, we assume that the number of units in each stratum is known. An each respondent in the sample stratum $h(=1,2,\cdots,L)$ is provided the randomization device R that consists of a sensitive question (A) card with probability P and its negative question (A^c) card with probability $1-P$. The respondent should answer the question by "Yes" or "No" without reporting which question card she or he has. A respondent belonging to the sample in different strata will perform same randomization devices. Let n_h be the number of units in the sample from stratum h and $n=\sum_{h=1}^{L} n_h$ be the total umber of units in the sample from all strata. Under assumption that these "Yes" or "No" reports are made truthfully and P (0<P<1, P$\neq 0.5$) is set by the researcher, the proportion of a "Yes" answer in stratum h for this procedure is

$$\lambda_h = P\pi_h + (1-P)(1-\pi_h), \text{ for } h=1,2,\cdots,L, \tag{2.1}$$

where λ_h be the proportion of "Yes" answer in stratum h, π_h is the proportion of respondents with sensitive characteristic in stratum h and P be the probability that a respondent has a sensitive question (A) card.

The maximum likelihood estimate (MLE) of π_h is

$$\hat{\pi}_h = \frac{\hat{\lambda}_h - (1-P)}{2P-1}, \tag{2.2}$$

where $\hat{\lambda}_h$ is the proportion of "Yes" answer in a sample in the stratum h.

Since each $\hat{\lambda}_h$ is distributed with $B(n_h, \lambda_h)$ and the selection in different strata are made independently, the MLE of π_{st} is

$$\hat{\pi}_{st} = \sum_{h=1}^{L} W_h \hat{\pi}_h = \sum_{h=1}^{L} W_h \left[\frac{\hat{\lambda}_h - (1-P)}{2P-1} \right] = \frac{P-1}{2P-1} + \frac{1}{2P-1} \sum_{h=1}^{L} W_h \hat{\lambda}_h, \tag{2.3}$$

The variance of $\hat{\pi}_{st}$ is given by

$$V(\hat{\pi}_{st}) = V\left(\sum_{h=1}^{L} W_h \hat{\pi}_h \right) = \sum_{h=1}^{L} {W_h}^2 V(\hat{\pi}_h) = \sum_{h=1}^{L} \frac{{W_h}^2}{n_h} \left[\pi_h(1-\pi_h) + \frac{P(1-P)}{(2P-1)^2} \right] \tag{2.4}$$

If the sample units are selected by simple random sampling without replacement (SRSWOR), then the variance of $\hat{\pi}_{st}$ is given by

$$V(\hat{\pi}_{st}) = V\left(\sum_{h=1}^{L} W_h \hat{\pi}_h \right) = \sum_{h=1}^{L} {W_h}^2 V(\hat{\pi}_h) = \sum_{h=1}^{L} {W_h}^2 \left[\frac{\pi_h(1-\pi_h)}{n_h}(1-f_h) + \frac{P(1-P)}{n_h(2P-1)^2} \right], \tag{2.5}$$

where $W_h = N_h / N$ is a stratum weight and $f_h = n_h / N_h$ is a sampling fraction for stratum h.

Different from Hong et, al., Kim and Warde(2004) consider that an each respondent in the sample stratum $h(=1,2,\cdots,L)$ is provided the randomization device R_h that consists of a sensitive question (A) card with

probability P_h and its negative question (A^c) card with probability $1-P_h$. The respondent should answer the question by "Yes" or "No" without reporting which question card she or he has. A respondent belonging to the sample in different strata will perform different randomization devices, each having different preassigned probabilities. Under assumption that these "Yes" or "No" reports are made truthfully and $P_h (\neq 0.5)$ is set by the researcher, the proportion of a "Yes" answer in stratum h for this procedure is

$$\lambda_h = P_h\pi_h + (1-P_h)(1-\pi_h), \text{ for } h=1,2,\cdots,L, \tag{2.6}$$

where λ_h be the proportion of "Yes" answer in stratum h, π_h is the proportion of respondents with sensitive characteristic in stratum h and P_h be the probability that a respondent in the sample stratum h has a sensitive question (A) card. The maximum likelihood estimate (MLE) of π_h is

$$\hat{\pi}_h = \frac{\hat{\lambda}_h - (1-P_h)}{2P_h - 1}, \tag{2.7}$$

where $\hat{\lambda}_h$ is the proportion of "Yes" answer in a sample in the stratum h.

Since each $\hat{\lambda}_h$ is distributed with $B(n_h, \lambda_h)$ and the selection in different strata are made independently, the MLE of π_{st} is

$$\hat{\pi}_{st} = \sum_{h=1}^{L} W_h\hat{\pi}_h = \sum_{h=1}^{L} W_h\left[\frac{\hat{\lambda}_h - (1-P_h)}{2P_h - 1}\right], \tag{2.8}$$

The variance of $\hat{\pi}_{st}$ is given by

$$V(\hat{\pi}_{st}) = V\left(\sum_{h=1}^{L} W_h\hat{\pi}_h\right) = \sum_{h=1}^{L} {W_h}^2 V(\hat{\pi}_h) = \sum_{h=1}^{L} \frac{{W_h}^2}{n_h}\left[\pi_h(1-\pi_h) + \frac{P_h(1-P_h)}{(2P_h-1)^2}\right] \tag{2.9}$$

If the sample units are selected by simple random sampling without replacement (SRSWOR), then the estimator (2.8) and its own variance (2.9) are unbiased for π_h and π_{st}, respectively. So that the variance of $\hat{\pi}_{st}$ is given by

$$V(\hat{\pi}_{st}) = V\left(\sum_{h=1}^{L} W_h\hat{\pi}_h\right) = \sum_{h=1}^{L} {W_h}^2 V(\hat{\pi}_h) = \sum_{h=1}^{L} {W_h}^2\left[\frac{\pi_h(1-\pi_h)}{n_h}(1-f_h) + \frac{P_h(1-P_h)}{n_h(2P_h-1)^2}\right], \tag{2.10}$$

where $W_h = N_h / N$ is a stratum weight and $f_h = n_h / N_h$ is a sampling fraction for stratum h.

3. Calibration for the stratified RR estimators

Let the population consists of L strata with N_h units in the stratum h from which a simple random sample of size n_h is selected without replacement. Then the total number of population size $N = \sum_{h=1}^{L} N_h$ and sample size $n = \sum_{h=1}^{L} n_h$ as defined in Section 2. Now, in order to calibrate the stratum weight $W_h = N_h / N$, we should define the covariate x, which associated with $\hat{\pi}_h$. Let $\bar{x}_h$ and $\bar{X}_h$ are the sample and population means of covariate x in the stratum h. Assume the population mean $\bar{X} = \sum_{h=1}^{L} W_h\bar{X}_h$ is accurately known. Let $\hat{\pi}_h$ and π_h are the sample and population proportions of a sensitive characteristic. The purpose is to estimate $\pi = \sum_{h=1}^{L} W_h\pi_h$ by incorporating the auxiliary variable x. We consider a new weight W_h^* obtained by calibration procedure, which minimizes the chi-square distance as follows

$$G(W_h^*, W_h) = \sum_{h=1}^{L} \frac{(W_h^* - W_h)^2}{q_h W_h}, \tag{3.2}$$

subject to the benchmark constraint

$$\overline{X} = \Sigma_{h=1}^{L} W_h^* \overline{x}_h \tag{3.3}$$

Using Lagrange method, we can obtain the calibration weight W_h^* is given by

$$W_h^* = W_h + \frac{W_h q_h \overline{x}_h}{\sum_{h=1}^{L} W_h q_h \overline{x}_h^2}\left[\overline{X} - \sum_{h=1}^{L} W_h \overline{x}_h\right] = W_h\left(1 + \frac{q_h \overline{x}_h}{\sum_{h=1}^{L} W_h q_h \overline{x}_h^2}\left[\overline{X} - \sum_{h=1}^{L} W_h \overline{x}_h\right]\right) = W_h g_h . \tag{3.4}$$

where $g_h = 1 + q_h \overline{x}_h \left(\Sigma_{h=1}^{L} W_h q_h \overline{x}_h^2\right)^{-1}\left[\overline{X} - \Sigma_{h=1}^{L} W_h \overline{x}_h\right]$ be the g-weight for stratum h, and q_h be a constant weight for determining the type of estimator.

Thus the calibration estimator of π is

$$\hat{\pi}_{st}^* = \frac{P-1}{2P-1} + \frac{1}{2P-1}\sum_{h=1}^{L} W_h^* \hat{\lambda}_h = \frac{P-1}{2P-1} + \frac{1}{2P-1}\sum_{h=1}^{L} W_h g_h \hat{\lambda}_h \tag{3.5}$$

The variance of $\hat{\pi}_{st}^*$ is

$$\begin{aligned} V(\hat{\pi}_{st}^*) &= \frac{1}{(2P-1)^2} V\left(\sum_{h=1}^{L} W_h g_h \hat{\lambda}_h\right) \\ &= \sum_{h=1}^{L} W_h^2 g_h^2 \left[\frac{\pi_h(1-\pi_h)}{n_h}(1-f_h) + \frac{P(1-P)}{n_h(2P-1)^2}\right] \end{aligned} \tag{3.6}$$

Whereas, Kim and Warde's calibrated estimator of the population proportion π is given by

$$\hat{\pi}_{st}^* = \sum_{h=1}^{L} W_h^* \hat{\pi}_h = \sum_{h=1}^{L} W_h g_h \left[\frac{\hat{\lambda}_h - (1-P_h)}{2P_h - 1}\right] \tag{3.7}$$

The variance of $\hat{\pi}_{st}^*$ is

$$V(\hat{\pi}_{st}^*) = V\left(\sum_{h=1}^{L} W_h^* \hat{\pi}_h\right) = \sum_{h=1}^{L} W_h^2 g_h^2 \left[\frac{\pi_h(1-\pi_h)}{n_h}(1-f_h) + \frac{P_h(1-P_h)}{n_h(2P_h-1)^2}\right] \tag{3.8}$$

4. Efficiency comparison

We perform the efficiency comparison of the ordinary and the calibrated estimators by the way of variance comparison. Let the relative efficiency (RE) of two variances is defined by

$$RE_i = \frac{V(\hat{\pi}_{sti})}{V(\hat{\pi}_{sti}^*)} \tag{4.1}$$

where the index i means that $i=1$ for the Hong et al. estimator and $i=2$ for the Kim and Warde's estimator.

To get the full benefit from stratification, the population proportions for the sensitive trait in strata are assumed to be possibly different. For the calibration estimator, the covariate x is proportional to the population proportion of a

sensitive trait.

4.1 Hong, et al 's RR estimator

Without loss of generality, we assume that the number of strata L=2. The size of population strata is considered $N_1 = 7000$, $N_2 = 3000$ and the counterpart $n_1 = 700$, $n_2 = 300$. Let the selection probabilities of sensitive question P=0.6 to 0.9 by 0.1 increments. Table 4.1 shows that the calibrated RR estimator is more efficient than the Hong, et al's RR estimator. Also, the RE1 is increased by the correlation from 0.1 to 0.9.

<Table 4.1> The relative efficiencies of $\hat{\pi}_{st1}$ and $\hat{\pi}^{*}_{st1}$ when n=1000

ρ	π_1	π_2	W_1	W_2	P			
					0.6	0.7	0.8	0.9
0.1	0.1	0.2	0.7	0.3	1.00053	1.00053	1.00053	1.00053
	0.3	0.4	0.7	0.3	1.00040	1.00040	1.00040	1.00040
0.5	0.1	0.2	0.7	0.3	1.00203	1.00203	1.00203	1.00204
	0.3	0.4	0.7	0.3	1.00166	1.00166	1.00166	1.00166
0.7	0.1	0.2	0.7	0.3	1.00258	1.00258	1.00258	1.00259
	0.3	0.4	0.7	0.3	1.00221	1.00221	1.00221	1.00222
0.9	0.1	0.2	0.7	0.3	1.00309	1.00309	1.00309	1.00310
	0.3	0.4	0.7	0.3	1.00282	1.00282	1.00283	1.00283

4.2 Kim and Warde's RR estimator

We assume that the number of strata, the size of population strata and the counterpart are the same as Section 4.1, respectively. Let the selection probabilities of sensitive question P_1=0.6 to 0.9 by 0.1 increments for stratum 1, and P_2 is different from P_1. It is difficult to derive the mathematical condition of the RE comparison between (2.5) and (3.6), so we perform to an empirical study on RE. We investigate the RE by different ρ, the correlation coefficient of π and x. We would expect that the RE is increased by ρ. From Table 4.2, we showed that the proposed calibration estimator is more efficient than Kim and Warde's estimator, because our calibration estimator uses the known auxiliary information at the population level in calibration procedure. Also. we can reduce the variance of that estimator. These results agree with the typical calibration estimator as Deville and Sarndal (1992) and Singh, Horn and Mohl (1998).

<Table 4.2> The relative efficiencies of $\hat{\pi}_{st2}$ and $\hat{\pi}^{*}_{st2}$ when n=1000

ρ	π_1	π_2	W_1	W_2	P_1							
					0.6		0.7		0.8		0.9	
					P_2		P_2		P_2		P_2	
					0.7	0.8	0.8	0.9	0.9	0.95	0.93	0.95
0.1	0.1	0.2	0.7	0.3	1.0005	1.0005	1.0005	1.0005	1.0005	1.0005	1.0005	1.0005
	0.3	0.4	0.7	0.3	1.0003	1.0003	1.0003	1.0003	1.0003	1.0003	1.0004	1.0004
0.5	0.1	0.2	0.7	0.3	1.0019	1.0019	1.0019	1.0019	1.0020	1.0019	1.0020	1.0020
	0.3	0.4	0.7	0.3	1.0016	1.0016	1.0016	1.0016	1.0016	1.0016	1.0016	1.0016
0.7	0.1	0.2	0.7	0.3	1.0025	1.0025	1.0025	1.0025	1.0025	1.0025	1.0025	1.0025
	0.3	0.4	0.7	0.3	1.0021	1.0021	1.0021	1.0021	1.0021	1.0021	1.0022	1.0022
0.9	0.1	0.2	0.7	0.3	1.0030	1.0030	1.0031	1.0031	1.0031	1.0031	1.0031	1.0031
	0.3	0.4	0.7	0.3	1.0028	1.0028	1.0028	1.0028	1.0028	1.0028	1.0028	1.0028

5. Concluding Remarks

The calibration procedure is to improve the ordinary estimator by incorporating the auxiliary information. In this paper, we have been derived the calibration estimator for the stratified randomized response model which suggested by Kim and Warde (2004), our proposed calibration estimator is more efficient than that of Kim and Warde's. Especially, we have been investigated the RE for different the correlation coefficient ρ, between the population proportion of a sensitive traits and the covariate, so that the RE of proposed estimator is increased by ρ.

References

[1] Cochran, W. G. (1977), Sampling Techniques, 3rd Edition, Wiley, New York.

[2] Deville, J.C., and Sarndal, C.E. (1992), Calibration estimators in survey sampling, *Journal of the American Statistical Association*, 87, 376-382.

[3] Kim, J. and Flueck, J. A. (1978), Modifications of the randomized response technique for sampling without replacement, *Proceedings of Survey Research Section. American Statistical Association*, 346-350.

[4] Kim, J. and Warde, W. D. (2004), A stratified Warner's randomized response model, *Journal of Statistical Planning and Inference*, 120, 155-165.

[5] Warner, S. L. (1965), Randomized response: a survey technique for eliminating evasive answer bias, *Journal of the American Statistical Association*, 60, 63-69.