Communications of the Korean Statistical Society Vol. 15, No. 4, 2008, pp. 597–603

The Calibration for Stratified Randomized Response Estimators

Chang-Kyoon Son[1], Ki-Hak Hong[2], Gi-Sung Lee[3], Jong-min Kim[4]

Abstract

In this paper, we propose the calibration procedure for the variance reduction of the stratified Warner's randomized response estimators, which suggested by Hong *et al.* (1994) and Kim and Warde (2004), using auxiliary information at the population level. It is shown that the proposed calibration estimators are more efficient than the ordinary Warner's estimators.

Keywords: Stratified RR model; Calibration; Auxiliary variable.

1. Introduction

The randomized response(RR) technique suggested by Warner (1965) 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 a variant of the Warner model. Mangat (1994) also proposed 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 nonoverlapping 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.* (1994) model to the optimal sampling and each stratum sample provides different randomization devices. They 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

1) Research Fellow, Korea Institute for Health and Social Affairs, Seoul 122-705, Republic of Korea.
2) Professor, Department of Computer Science, Dongshin University, Jeonnam 520-714, Republic of Korea.
3) Professor, Department of Childern Welfare, Woosuk University, Jeonbuk 565-701, Republic of Korea.
4) Professor, Statistics, Division of Science and Mathematics, University of Minnesota, Morris, MN, 56267, USA. Correspondence: jmkim@morris.umn.edu

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 *et al.* (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 Särndal (1992) and Deville *et al.* (1993) proposed a class of estimators derived from a re-weighting approach that addresses the same issue of variance reduction called the calibration estimators. In this point of view, we can consider the calibration procedure using auxiliary information to randomized response technique for a sensitive trait such as tax evasion or drug abuse in order to reduce the sampling or nonsampling error.

In this paper, we consider the calibration RR estimators in order to reduce the variance of ordinary stratified RR estimators suggested by Hong *et al.* (1994) and Kim and Warde's (2004) using auxiliary information.

The paper is organized as follows. In Section 2 we describes as the stratified RR estimator. In Section 3 we explains how calibration is carried out of the stratified RR estimator using the generalized least squares distance. In Section 4 we show that the proposed estimator is more efficient than the ordinary stratified RR estimator by numerical study. In Section 5 we make some concluding observations.

2. Stratified Randomized Response Techniques

Let the population $U = \{1, 2, \ldots, N\}$ be divided into non-overlapping strata $U = \{U_1, \ldots, U_h, \ldots, U_L\}$ each for strata size N_h, $h = 1, 2, \ldots, L$, with a priori and a sample $s = \{s_1, \ldots, s_h, \ldots, s_L\}$ with each size n_h, for $h = 1, 2, \ldots, L$, is selected by simple random sampling with replacement(SRSWR) in each stratum. We also assume that the number of units in each stratum is known. Each respondent in the sample stratum $h(= 1, 2, \ldots, L)$ 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 the 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

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

where Z_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

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

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

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

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

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

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

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

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

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*'s (1994) estimator, Kim and Warde (2004) consider that each respondent in the sample stratum $h(= 1, 2, \ldots, L)$ 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

$$Z_h = P_h \pi_h + (1-P_h)(1-\pi_h), \quad \text{for } h = 1, 2, \ldots, L, \quad (2.6)$$

where Z_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 is 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

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

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

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

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

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

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

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.

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

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

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 and a simple random sample of size n_h is selected by without replacement from each the population stratum. 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 for the stratum h. Assume that the population mean of auxiliary variable $\bar{X} = \sum_{h=1}^{L} W_h \bar{X}_h$ 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 $\pi_{st} = \sum_{h=1}^{L} W_h \pi_h$ incorporating the auxiliary variable x. We consider new weights W_h^* obtained by calibration procedure, which minimizes the chi-square distance as follows

$$G(W_h^*, W_h) = \sum_{h=1}^{L} \frac{(W_h^* - W_h)^2}{q_h W_h}, \tag{3.1}$$

subject to the benchmark constraint

$$\bar{X} = \sum_{h=1}^{L} W_h^* \bar{x}_h. \tag{3.2}$$

Using Lagrange method, we can obtain the calibration weights W_h^* are given by

$$W_h^* = W_h \left(1 + \frac{q_h \bar{x}_h}{\sum_{h=1}^{K} W_h q_h \bar{x}_h^2} \left[\bar{X} - \sum_{h=1}^{L} W_h \bar{x}_h \right] \right) = W_h g_h, \tag{3.3}$$

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

The ordinary strata weights in the stratified RR estimators (2.3) and (2.8) are adjusted by the calibrated weight (3.3), so that we can obtain some results as follows.

Proposition 3.1 For the Hong *et al*'s (1994) stratified RR estimator (2.3), the calibration estimator and its variance of π_{st} are given by

$$\hat{\pi}_{st}^{*} = \frac{P-1}{2P-1} + \frac{1}{2P-1}\sum_{h=1}^{L} W_h^{*}\hat{Z}_h = \frac{P-1}{2P-1} + \frac{1}{2P-1}\sum_{h=1}^{L} W_h g_h \hat{Z}_h. \tag{3.4}$$

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

$$V(\hat{\pi}_{st}^{*}) = \sum_{h=1}^{L} W_h^2 g_h^2 \left[\frac{\pi_h(1-\pi_h)}{n_h}(1-f_h) + \frac{P(1-P)}{n_h(2P-1)^2}\right]. \tag{3.5}$$

Proof: By simple algebraic, we can obtain the variance of $\hat{\pi}_{st}^{*}$. □

Proposition 3.2 For Kim and Warde's (2004) stratified RR estimator (2.8), the calibration estimator and its variance of the population proportion π are given by

$$\hat{\pi}_{st}^{*} = \sum_{h=1}^{L} W_h^{*}\hat{\pi}_h = \sum_{h=1}^{L} W_h g_h \left[\frac{\hat{Z}_h - (1-P_h)}{2P_h - 1}\right]. \tag{3.6}$$

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

$$V(\hat{\pi}_{st}^{*}) = \sum_{h=1}^{L} W_h^2 g_h^2 \left[\frac{\pi_h(1-\pi_h)}{n_h}(1-f_h) + \frac{P_h(1-P_h)}{n_h(2P_h-1)^2}\right]. \tag{3.7}$$

Proof: Refer to the proof of Proposition 1. □

4. Efficiency Comparisons

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 be defined by

$$\mathrm{RE}(\hat{\pi}_{st}^{*} \mid \hat{\pi}_{st}) = \frac{V(\hat{\pi}_{st})}{V(\hat{\pi}_{st}^{*})}. \tag{4.1}$$

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 (1994) RR estimator

Without loss of generality, we assume that the number of strata is two. The size of population strata is considered $N_1 = 7,000$, $N_2 = 3,000$ 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 RE is increased by the correlation from 0.1 to 0.9.

Table 4.1: Realtive Efficiencies of $\hat{\pi}_{st1}$ and $\hat{\pi}^*_{st1}$ when $n = 1,000$

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

Table 4.2: Realtive Efficiencies of $\hat{\pi}_{st2}$ and $\hat{\pi}^*_{st2}$ when $n = 1,000$

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	1.0021	1.0021
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

4.2. Kim and Warde's (2004) 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 numerical 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. It means that the variance of our proposed estimator can be reduced because our calibration estimator uses the known auxiliary information at the population level in calibration procedure. These results agree with the typical calibration estimator as Deville and Särndal (1992) and Singh *et al.* (1998).

5. Concluding Remarks

The calibration procedure is to improve the ordinary estimator by incorporating the auxiliary information. In this paper, we have 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 (2004). Especially, we have investigated the RE's by changing the values of the correlation coef-

ficient ρ, between the population proportion of a sensitive traits and the covariate. We have noticed that the REs of the proposed estimator increases as ρ increases.

References

Cassel, C. M., Särndal, C. E. and Wretman, J. H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations, *Biometrika,* **63**, 615–620.

Deville, J. C. and Särndal, C. E. (1992). Calibration estimators in survey sampling, *Journal of the American Statistical Association,* **87**, 376–382.

Deville, J. C., Särndal, C. E. and Sautory, O. (1993). Generalized raking procedures in survey sampling, *Journal of the American Statistical Association,* **88**, 1013–1020.

Fuller, W. A. (1975). Regression analysis for sample surveys, *Sankhya C,* **37**, 117–132.

Hong, K., Yum, J. and Lee, H. (1994). A stratified randomized response technique, *The Korean Journal of Applied Statistics,* **7**, 141–147.

Isaki, C. T. and Fuller, W. A. (1982). Survey design under the regression superpopulation model, *Journal of the American Statistical Association,* **77**, 89–96.

Kim, J. and Warde, W. D. (2004). A stratified Warners randomized response model, *Journal of Statistical Planning and Inference,* **120**, 155–165.

Mangat, N. S. (1994). An Improved Randomized Response Strategy, *Journal of the Royal Statistical Society, Series B,* **55**, 93–95.

Mangat, N. S. and Singh, R. (1990). An alternative randomized response proceures, *Biometrika,* **77**, 439–442.

Singh, S., Horn, S. and Yu, F. (1998). Estimation of variance of general regression estimator: Higher level calibration approach, *Survey Methodology,* **24**, 41–50.

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

Wright, R. L. (1983). Finite population sampling with multivariate auxiliary information, *Journal of the American Statistical Association,* **78**, 879–884.

[Received April 2008, Accepted June 2008]