An Improved Estimation Procedure of the Mean of a Sensitive Variable Using Auxiliary Information

Tanveer AT¹ and Housila PS²

¹Department of Computer Science and Engineering, Islamic University of Science and Technology, India
²School of Studies in Statistics, Vikram University Ujjain, India

Submission: August 22, 2017; Published: October 02, 2017

*Corresponding author: Tanveer A Tarray, Department of Computer Science and Engineering, Islamic University of Science and Technology, India, Email: tanveerstat@gmail.com

Abstract
This paper proposes new ratio and regression estimators for the mean of sensitive variable utilizing information on a non-sensitive auxiliary variable. Expressions for the Biases and mean square errors of the suggested estimators correct up to first order of approximation are derived. It has been shown that the suggested new ratio and regression estimators are better than conventional unbiased estimators which do not utilize the auxiliary information, Sousa et al. [1] ratio estimator and Gupta et al. [2] regression estimator under a very realistic condition. In support of the present study we have also given the numerical illustrations.

Keywords: Ratio estimator; Regression estimator; Randomized response technique; Mean Square error; Bias; Auxiliary variable

Introduction
Let Y be the variable under study, a sensitive variable which can’t be observed directly. Let X is a non-sensitive auxiliary variable which is strongly correlated with Y. Let S be a scrambling variable independent of the study variable Y and the auxiliary variable X. The usual additive model used for gathering information on quantitative sensitive variable is due to Himmelfarb & Edgell [3]. Their model allows the interviewee to hide personal information using a scrambling variable to their response. The respondent is asked to report a scrambled response for the study variable Y (based on additive model) given by Za = Y + S, but is asked to provided a true response for the auxiliary variable X [1].

Hussain [4] have discussed the use of subtracting scrambling. Thus following Hussain [4], the respondent is asked to report a scrambled response for the study variable Y (based on subtractive model) given by Zs = Y - S, but is asked to provided a true response for the auxiliary variable X. It is interesting to mention that the proposed model generalizes both usual additive and subtractive models. Gjestvang & Singh [5] have pointed out that “the practical application of an additive model is much easier than the multiplicative model, that is, respondents may like to add two numbers rather than doing painstaking work of multiplying two numbers or dividing two numbers: thus the improvement of the additive model has its own importance in the literature”. Looking at the form the additive model, subtractive model and above arguments due to Gjestvang & Singh [5] we have introduced a new model (which is additive in nature)

\[ Z_\phi = Y + \phi S \]

where \( \phi \) is a known scalar such that \(-1 \leq \phi \leq 1\).

Thus keeping the proposed model \( Z_\phi = Y + \phi S \) in view, the respondent is asked to report a scrambled response for Y given but is asked to provide a true response for X. Let a simple random sample of size n be drawn without replacement from a finite population U=(U1, U2, ..., UN), let and respectively be the values of the study variable Y and the auxiliary variable X. Further, let \( \bar{Y} = \left(1/n\right) \sum_{i=1}^{n} Y_i \), \( \bar{X} = \left(1/n\right) \sum_{i=1}^{n} X_i \) be the sample means and \( \bar{Z}_\phi = \left(1/n\right) \sum_{i=1}^{n} Z_{\phi i} \) be the population mean for Y, X and \( Z_\phi \) respectively. We assume that the population mean \( \bar{X} \) of the auxiliary variable X is known and \( \bar{S} = E(S) = 0 \).

Thus, \( E(Z_\phi) = E(Y) \). We also define

\[ \bar{X} = \bar{X}(1 + e_x) \quad \text{and} \quad \bar{Z}_\phi = \bar{Z}_\phi(1 + e_{\phi}) \]

such that \( E(e_x) = E(e_{\phi}) = 0 \), and

\[ E(e_x^2) = \frac{(1-f)}{n} c_x^2, \quad E(e_{\phi}^2) = \frac{(1-f)}{n} c_{\phi}^2. \]
The suggested ratio estimator

We consider the following ratio estimator for the population mean \( \bar{Y} \) of the study variable Y using the known population mean of the auxiliary variable X:

\[ t_{(\phi)} = \bar{Y} + \phi \bar{S} \]  

(2.1)

We note that for \( \phi = 1 \), the proposed estimator reduces to the estimator

\[ t_{(1)} = \bar{Y} \]  

(2.2)

which is due to Sousa et al. [1], where \( \bar{S} = (1/n) \sum z_i \) . For \( \phi = 0 \), (2.1) reduces to the classical ratio estimator

\[ t_{(0)} = \bar{Y} \]  

(2.3)

based on true responses of variables Y and X.

Expressing (2.1) in terms of \( \bar{z} = \bar{y} - \phi \bar{x} \) and ex we have

\[ t_{R(\phi)} = \bar{z} + \phi \bar{S} \]  

(2.4)

We assume that |ex|<1 so that (1+ex)-1 is expandable. Expanding the right hand side of (2.4), multiplying out and neglecting terms of \( \phi \)'s having power greater than one we have

\[ t_{R(\phi)} \rightarrow (1 + e_x) \bar{z} + \phi \bar{S} \]  

(2.5)

Taking expectation of both sides of (2.5) we get the bias of to the first order of approximation as

\[ \text{Bias}(t_{R(\phi)}) = \frac{(1-f)}{n} \bar{z} (C_{x} + C_{y}) \]  

(2.6)

where \( k \) is the bias of the proposed estimator \( t_{R(\phi)} \) and the bias of the estimator \( t_{R(1)} \) due to Sousa et al. [1] are same. This fact can also be seen from (2.6) and (2.9).

Squaring both sides of (2.5) and neglecting terms of \( \phi \)'s having power greater than two we have

\[ (t_{R(\phi)} - \bar{z})^2 = \bar{z}^2 (1 + 2\phi e_x + e_x^2) \]  

(2.7)

Taking expectation of both sides of (2.7) we get the mean square error (MSE) of \( t_{R(\phi)} \) to the first degree of approximation as

\[ \text{MSE}(t_{R(\phi)}) = \frac{(1-f)}{n} \bar{z}^2 (C_{x}^2 + C_{y}^2 - 2\rho_{yx} C_{x} C_{y}) \]  

(2.8)

Expression (2.8) indicates that the MSE of the proposed estimator \( t_{R(\phi)} \) depends on the scalar \( \phi \). So there will be effect of selecting the value of \( \phi \) towards increasing or decreasing the MSE of \( t_{R(\phi)} \). So one should be very cautious about the
selection of value of $\phi$. Setting $\varphi = 1$ in (2.6) and (2.8) we get the bias and MSE of the Sousa et al. [1] estimator $t_R(1)$ to the first degree of approximation respectively as

$$\text{Bias}(t_R(1)) = -\frac{1}{n} \varphi \varphi \frac{1}{1-k} \left(1-\varphi^2\right) (2.9)$$

$$\text{MSE}(t_R(1)) = -\frac{1}{n} \varphi \varphi \frac{1}{1-k} \left[(2k-1)C_1^2 + r^2 C_1^2 (1-\varphi^2)\right] (2.10)$$

**Efficiency Comparison**

From (1.1) and (2.8) we have

$$\text{MSE}(\tau) - \text{MSE}(t_R(1)) = \frac{1}{n} \varphi \varphi \frac{1}{1-k} \left[(2k-1)C_1^2 + r^2 C_1^2 (1-\varphi^2)\right]$$

which is positive if

$$2k-1 > 0 \text{ and } (1-\varphi^2) > 0$$

i.e. if

$$k > \frac{1}{2} \text{ or } (2.12)$$

This is a condition of the classical ratio estimator $t_R$ in (2.3) to be better than the usual unbiased estimator $\overline{Y}$. It follows from (2.11) and (2.12) the proposed estimator $t_R(\varphi)$ is more efficient than the unbiased estimator $\overline{Z}$, $\overline{Z}_s$, and $\overline{Z}_\varphi$ if the conditions (2.11) holds true.

Further from (2.8) and (2.10) we have

$$\text{MSE}(t_R(1)) - \text{MSE}(t_R(\varphi)) = \frac{1}{n} \varphi \varphi \frac{1}{1-k} \left[(1+2k)\left[\varphi^2 C_1^2 (1+\varphi^2 r^2) + C_1^2 (1-2k)\right]\right]$$

which is positive if

$$| \varphi | < 1 \text{ or } (2.13)$$

Thus it follows from (2.11), (2.12) and (2.13) that the suggested estimator $t_R(\varphi)$ is more efficient than the unbiased estimator $\overline{Z}$, $\overline{Z}_s$, and $\overline{Z}_\varphi$ and the ratio type estimator $t_R(1)$ due to Sousa et al. [1].

**Remark 2.1:** If the correlation between the two variables $Z$ and $X$ is negative then one can consider the following product-type estimator for the population mean $\overline{Y}$ as

$$t_R(\varphi) = \overline{Z} \cdot \left(\frac{\varphi}{\varphi}\right) (2.14)$$

To exact bias of the proposed product-type estimator $t_P(\varphi)$ is given by

$$\text{Bias}(t_P(\varphi)) = \frac{(1-\varphi)}{n} \varphi \varphi \frac{1}{1-k} \left[\varphi^2 C_1^2 (1+\varphi^2 r^2) + C_1^2 (1-2k)\right] (2.15)$$

which is same as the bias of the classical product estimator $t_P$.

It is observed from (2.15) that the bias expression of $t_P(\varphi)$ is free from the scalar $\phi$. So whatever be the value of $\phi$, the bias of $t_P(\varphi)$ will remains same as given in (2.15).

The mean square error of the estimator $t_P(\varphi)$ to the first degree of approximation is given by

$$\text{MSE}(t_P(\varphi)) = \frac{(1-\varphi)}{n} \varphi \varphi \frac{1}{1-k} \left[C_1^2 (1+\varphi^2 r^2) + C_1^2 (1-2k)\right] (2.17)$$

which depends on the value of the scalar $\phi$. So one should be careful in selecting the value of $\phi$.

From (1.3) and (2.17) we have

$$\text{MSE}(\tau) - \text{MSE}(t_P(\varphi)) = \frac{1}{n} \varphi \varphi \frac{1}{1-k} \left[C_1^2 (1+\varphi^2 r^2) + C_1^2 (1-2k)\right] (2.18)$$

Which is positive if $-(1+2k) > 0$ i.e. if $(1+2k) < 0$ as $k < -1/2 (2.19)$

Which equals to the same condition in which the classical product estimator $t_P$ is better than usual unbiased estimator $\overline{Y}$

**Empirical Study**

To judge the superiority of the proposed estimator $t_R(\varphi)$ over $\overline{Z}$, $\overline{Z}_s$, and $\overline{Z}_\varphi$ and the ratio type estimator $t_R(1)$ due to Sousa et al. [1] we have computed the percent relative efficiencies of $t_R(\varphi)$ with respect to $\overline{Z}$, $\overline{Z}_s$, and $\overline{Z}_\varphi$ by using the formulae:

$$\text{PRE}(t_R(\varphi), \tau) = \frac{\text{MSE}(\tau)}{\text{MSE}(t_R(\varphi))} \times 100$$

$$\text{PRE}(t_R(\varphi), \overline{Z}) = \frac{\text{MSE}(\tau)}{\text{MSE}(t_R(\varphi))} \times 100$$

$$\text{PRE}(t_R(\varphi), \overline{Z}_s) = \frac{\text{MSE}(\tau)}{\text{MSE}(t_R(\varphi))} \times 100$$

$$\text{PRE}(t_R(\varphi), \overline{Z}_\varphi) = \frac{\text{MSE}(\tau)}{\text{MSE}(t_R(\varphi))} \times 100$$

For the percent relative efficiency (PRE’s) computation purpose we assume for the sake of simplicity that $C_r = C_s R = \overline{Y}/\overline{X} \equiv 1 \Rightarrow \overline{Y} = \overline{X} S_s = \alpha S_r$ where $\alpha$ is a scalar in percent, (i.e. $\alpha \%$) as mentioned in Sousa et al. [1], Gupta et al. [2]. Under the above assumptions the PRE’s formulae given by (2.14), (2.15) and (2.16) respectively reduce to:

$$\text{PRE}(t_R(\varphi), \tau) = \frac{(1+\alpha^2)}{\varphi \alpha^2 + 2(1-\rho_{xy})} \times 100$$  (2.17)
We have computed the PRE’s for $\alpha=10\%$ [1,2], 20\%, 30\%, $\phi = -1 (0.25) 1$ and $\rho_{yx} = 0.55, 0.6, (0.1) 0.9$.

It is observed from Table 1-3 that:

### Table 1: The values of $PRE(t_{R(\phi)}, \overline{z}_a), PRE(t_{R(\phi)}, \overline{z}_\phi)$ and $PRE(t_{R(\phi)}, t_{R(l)})$ for different values of ($\alpha$, $\phi$, $\rho_{yx}$).

| Estimator | $\alpha$ | $\phi$ | $\rho_{yx}$ | -1.00 | -0.75 | -0.50 | -0.25 | 0.00 | 0.25 | 0.50 | 0.75 | 1.00 |
|-----------|----------|--------|-------------|-------|-------|-------|-------|-----|-----|-----|-----|-----|
| $\overline{z}_a$ | 0.55 | -1.00 | 110.99 | 111.53 | 111.91 | 112.22 | 112.14 | 111.91 | 111.53 | 110.99 |
| $\overline{z}_\phi$ | 0.55 | -0.75 | 110.99 | 111.04 | 111.08 | 111.10 | 111.11 | 111.10 | 111.08 | 111.04 | 110.99 |
| $t_{R(l)}$ | 1.00 | -1.00 | 100.00 | 100.48 | 100.83 | 101.04 | 101.11 | 101.04 | 100.83 | 100.48 | 100.00 |
| $\overline{z}_a$ | 0.60 | -1.00 | 124.69 | 125.37 | 125.86 | 126.15 | 126.25 | 126.15 | 125.86 | 125.37 | 124.69 |
| $\overline{z}_\phi$ | 0.60 | -0.75 | 124.69 | 124.83 | 124.92 | 124.98 | 125.00 | 124.98 | 124.92 | 124.83 | 124.69 |
| $t_{R(l)}$ | 1.00 | -1.00 | 100.00 | 100.54 | 101.00 | 101.24 | 101.45 | 101.24 | 101.00 | 100.54 | 100.00 |
| $\overline{z}_a$ | 0.70 | -1.00 | 165.57 | 166.77 | 167.63 | 168.33 | 168.16 | 167.63 | 166.77 | 165.57 | 165.57 |
| $\overline{z}_\phi$ | 0.70 | -0.75 | 165.57 | 166.05 | 166.39 | 166.60 | 166.60 | 166.60 | 166.05 | 165.57 | 165.57 |
| $t_{R(l)}$ | 1.00 | -1.00 | 100.00 | 100.72 | 101.24 | 101.56 | 101.56 | 101.56 | 101.24 | 100.72 | 100.00 |
| $\overline{z}_a$ | 0.80 | -1.00 | 246.34 | 249.00 | 250.93 | 252.11 | 252.50 | 252.09 | 249.00 | 246.34 | 246.34 |
| $\overline{z}_\phi$ | 0.80 | -0.75 | 246.34 | 247.92 | 249.07 | 249.77 | 250.00 | 249.77 | 249.92 | 248.92 | 246.34 |
| $t_{R(l)}$ | 1.00 | -1.00 | 100.00 | 101.08 | 101.86 | 102.34 | 102.50 | 102.34 | 101.86 | 101.08 | 100.00 |
| $\overline{z}_a$ | 0.90 | -1.00 | 480.95 | 491.19 | 498.77 | 503.43 | 505.00 | 503.43 | 498.77 | 491.19 | 480.95 |
| $\overline{z}_\phi$ | 0.90 | -0.75 | 480.95 | 490.06 | 495.06 | 498.75 | 500.00 | 498.75 | 495.06 | 489.06 | 480.95 |
| $t_{R(l)}$ | 1.00 | -1.00 | 100.00 | 102.13 | 103.70 | 104.67 | 105.00 | 104.67 | 103.70 | 102.13 | 100.00 |

### Table 2: The values of $PRE(t_{R(\phi)}, \overline{z}_a), PRE(t_{R(\phi)}, \overline{z}_\phi)$ and $PRE(t_{R(\phi)}, t_{R(l)})$ for different values of ($\alpha$, $\phi$, $\rho_{yx}$).

| Estimator | $\alpha$ | $\phi$ | $\rho_{yx}$ | -1.00 | -0.75 | -0.50 | -0.25 | 0.00 | 0.25 | 0.50 | 0.75 | 1.00 |
|-----------|----------|--------|-------------|-------|-------|-------|-------|-----|-----|-----|-----|-----|
| $\overline{z}_a$ | 0.55 | -1.00 | 110.64 | 112.74 | 114.29 | 115.24 | 115.56 | 115.24 | 114.29 | 112.74 | 110.64 |
| $\overline{z}_\phi$ | 0.55 | -0.75 | 110.64 | 110.84 | 110.99 | 111.08 | 111.11 | 111.08 | 110.99 | 110.84 | 110.64 |
| $t_{R(l)}$ | 1.00 | -1.00 | 100.00 | 101.90 | 103.30 | 104.16 | 104.24 | 104.16 | 103.30 | 101.90 | 100.00 |
| $\overline{z}_a$ | 0.60 | -1.00 | 123.81 | 126.44 | 128.40 | 129.60 | 130.00 | 129.60 | 128.40 | 126.44 | 123.81 |
| $\overline{z}_\phi$ | 0.60 | -0.75 | 123.81 | 124.32 | 124.69 | 124.92 | 125.00 | 124.92 | 124.69 | 124.32 | 123.81 |
| $t_{R(l)}$ | 1.00 | -1.00 | 100.00 | 102.13 | 103.70 | 104.67 | 105.00 | 104.67 | 103.70 | 102.13 | 100.00 |
How to cite this article: Tanveer AT, Housila PS. An Improved Estimation Procedure of the Mean of a Sensitive Variable Using Auxiliary Information. Biostat Biometrics Open Acc J. 2017; 3(2): 555607. DOI: 10.19080/BBOAJ.2017.03.555607

Table 3: The values of \( \text{PRE}(t_{R(\varphi)}, \overline{Z}_a), \text{PRE}(t_{R(\varphi)}, \overline{Z}_s) \) and \( \text{PRE}(t_{R(\varphi)}, t_{R(I)}) \) for different values of \( (\alpha, \phi, \rho_{yx}) \).

| Estimator | \( \alpha \) | \( \phi \) | -1.00 | -0.75 | -0.50 | -0.25 | 0.00 | 0.25 | 0.50 | 0.75 | 1.00 |
|-----------|-------------|------------|-------|-------|-------|-------|-----|-----|-----|-----|-----|
| \( \overline{Z}_a \) | 0.70        | 162.50     | 167.07 | 170.49 | 172.61 | 173.33 | 172.61 | 170.49 | 167.07 | 162.50 |
| \( \overline{Z}_s \) | 162.50     | 164.26     | 165.57 | 166.39 | 166.67 | 166.39 | 165.57 | 164.26 | 162.50 |
| \( t_{R(I)} \) | 100.00     | 102.81     | 104.92 | 106.22 | 106.67 | 106.22 | 104.92 | 102.81 | 100.00 |
| \( \overline{Z}_a \) | 0.80        | 236.36     | 246.15 | 253.66 | 258.39 | 260.00 | 258.39 | 253.66 | 246.15 | 236.36 |
| \( \overline{Z}_s \) | 236.36     | 242.01     | 246.34 | 249.07 | 250.00 | 249.07 | 246.34 | 242.01 | 236.36 |
| \( t_{R(I)} \) | 100.00     | 104.14     | 107.32 | 109.32 | 110.00 | 109.32 | 107.32 | 104.14 | 100.00 |
| \( \overline{Z}_a \) | 0.90        | 433.33     | 467.42 | 495.24 | 513.58 | 520.00 | 513.58 | 495.24 | 467.42 | 433.33 |
| \( \overline{Z}_s \) | 433.33     | 459.55     | 480.95 | 495.06 | 500.00 | 495.06 | 480.95 | 459.55 | 433.33 |
| \( t_{R(I)} \) | 100.00     | 107.87     | 114.29 | 118.52 | 120.00 | 118.52 | 114.29 | 107.87 | 100.00 |

I. For fixed values of \( \alpha = 10 \%, 20 \%, 30 \% \), larger gain in efficiency is observed by using the proposed estimator \( t_{R(\varphi)} \) over the conventional unbiased estimators \( \overline{Z}_a, \overline{Z}_s \) which do not utilize the auxiliary information.

II. For \( \alpha = 10 \% \) the gain in efficiency by using the proposed estimator \( t_{R(\varphi)} \) over the ratio type estimator \( t_{R(I)} \) of Sousa et al.’s (2010) is marginal while for \( \alpha = 20 \% \) and 30 \% are substantial.

III. For fixed values of \( (\alpha, \rho_{yx}) \), the values of, \( \text{PRE}(t_{R(\varphi)}, \overline{Z}_a), \text{PRE}(t_{R(\varphi)}, \overline{Z}_s) \) and \( \text{PRE}(t_{R(\varphi)}, t_{R(I)}) \) increase as the value of \( \phi \) increases up to ‘zero’ and starts decreasing when it goes beyond ‘zero’.

IV. The maximum gain in efficiency is observed when \( \phi = 0 \), which is obvious because proposed additive model \( \hat{Z} = \beta x + \phi x^2 \) becomes free from the scrambling.

V. For fixed value of \( (\rho_{xy}, \alpha) \), the values of \( PRE(t_{R(\phi)}), PRE(t_{R(\phi)}), PRE(t_{R(\phi)}t_{R(\phi)}) \) increase as the values of the correlation coefficient \( \rho_{xy} \) increases.

Overall we conclude that the proposed estimator \( t_{R(\phi)} \) is to be preferred in practice when:

i. The standard deviation of the scrambling variable \( S \) is closer to the standard deviation of the auxiliary variable \( X \).

ii. The value of \( \phi \) is closer to ‘zero’ and the value of correlation coefficient \( \rho_{xy} \) is larger.

Proposed Regression Estimator

To obtain the regression estimator of the population mean \( \bar{Y} \) we first define the difference estimator for \( \bar{Y} \) as

\[ t_d = \bar{Z} + d(\bar{X} - \bar{X}) \]  

(3.1)

where \( d \) is a suitably chosen constant. It is easy to verify that the difference estimator \( t_d \) is unbiased estimator of the population mean \( \bar{Y} \).

The variance of the estimator \( t_d \) is given by

\[ V(t_d) = V(\bar{Z}) + d^2V(\bar{X}) - 2dCov(\bar{Z}, \bar{X}) \]  

(3.2)

which is minimized for

\[ d = \frac{S_x}{S_z} = \beta_{x,z} = d_x \text{ (say)} \]  

(3.3)

where

\[ S_y = \frac{1}{N} \sum_{i=1}^{N} (z_i - \bar{z})(y_i - \bar{y}) \]  

\[ \beta_{x,z} = \text{population regression coefficient of } Z \text{ on } X. \]

Substitution of (3.3) in (3.1) yields the resulting optimum difference estimator for the population mean \( \bar{Y} \) as

\[ t_{d0} = \bar{Z} + \beta_{x,z}(\bar{X} - \bar{X}) \]  

(3.4)

We note that the value of \( \beta_{x,z} \) is unknown in practice. In such a situation we replace \( \beta_{x,z} \) by its consistent estimate

\[ \hat{\beta}_{x,z} = \frac{s_{z,x}}{s_z} \]  

(3.5)

where \( \hat{\beta}_{x,z} \) is the sample regression coefficient of \( Z \) and \( X \) and \( S_z = Y + \phi \hat{S} \) is the scrambled response on \( Y \);

\[ s_{z,x} = \frac{1}{(N-1)} \sum_{i=1}^{N} (z_i - \bar{z})(x_i - \bar{x}) \]

and \( s_z^2 = \frac{1}{(N-1)} \sum_{i=1}^{N} (x_i - \bar{x})^2 \) are unbiased estimators of \( S_{z,x} \) and \( S_z^2 \) respectively. Thus the resulting regression estimator for the population mean \( \bar{Y} \) is given by

\[ t_r = \bar{Z} + \hat{\beta}_{x,z}(\bar{X} - \bar{X}) \]  

(3.6)

To obtain the bias of the regression estimator \( t_r \) we further write

\[ s_{z,x} = S_{z,x}^2(1 + e) \]  

and \( s_z^2 = S_z^2(1 + e) \)

such that

\[ E(e) = 0 \]

and from Sukhatme & Sukhatme [6] we have

\[ E(e) \approx \frac{1}{n} \frac{\mu(x, z)}{S_{x,y}}, \quad E(e) \approx \frac{1}{n} \frac{\mu(x, z)}{S_{x,y}} \]

where

\[ \mu(x, z) = E(x(z - \bar{z})) \]

and \( \mu(x, z) = E(x(z - \bar{x})). \]

Expressing (3.6) in terms of \( e \)’s we have

\[ t_r = \bar{Z} + \hat{\beta}_{x,z}(\bar{X} - \bar{X}) + \sum_{i=1}^{N} \hat{\beta}_{x,z} e_i \]

(3.7)

We assume that \( |e| < 1 \) so that \((1 + e)^2 \) is expandable. Now expanding the right hand side of (3.7), multiplying out and neglecting terms of \( e \)’s having power greater than two we have

\[ t_r \approx \bar{Z} + \hat{\beta}_{x,z}(\bar{X} - \bar{X}) + \sum_{i=1}^{N} \hat{\beta}_{x,z} e_i \]

(3.8)

Noting that \( \bar{Z} \approx \bar{Y} \) in (3.8) we have

\[ t_r \approx \bar{Y} + \hat{\beta}_{x,z}(\bar{X} - \bar{X}) + \sum_{i=1}^{N} \hat{\beta}_{x,z} e_i \]

(3.9)

Taking expectation of both sides of (3.9) we get the bias of \( t_r \) to the first degree of approximation as

\[ E(t_r) = \bar{Y} + \hat{\beta}_{x,z}(\bar{X} - \bar{X}) + \sum_{i=1}^{N} \hat{\beta}_{x,z} E(e_i) \]

(3.10)

Showing that the proposed regression estimator \( t_r \) is a biased estimate. The bias will be negligible if the sample size \( n \) is sufficiently large.

Squaring both sides of (3.10) and neglecting terms of \( e \)’s having power greater than two we have

\[ (t_r - \bar{Y})^2 \approx (\bar{Y}^2 + \bar{X}^2 \hat{\beta}_{x,z}^2 e_i^2 - 2\bar{Y} \bar{X} \hat{\beta}_{x,z} e_i e_i) \]

(3.11)

Taking expectation of both sides of (3.11) we get the mean square error of \( t_r \) to the first degree of approximation as

\[ MSE(t_r) = \frac{1}{n} \frac{\sum_{i=1}^{N} (\hat{\beta}_{x,z}^2 e_i^2)}{S_z^2} \]  

(3.12)

We note that

\[ \beta_{x,z} = \frac{S_{z,x}}{S_z} = \frac{\mu_{x,z}}{\mu_{x,y}} \]

and

\[ s_z^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2 \]

\[ \rho_{xy} = \frac{s_{x,y}}{\sqrt{s_x s_y}} \]

(3.13)

In the light of (3.13), the expression (3.12) reduces to:

\[ MSE(t_r) = \frac{1}{n} \frac{\sum_{i=1}^{N} (\hat{\beta}_{x,z}^2 e_i^2 + \rho_{x,z}^2 s_z^2 - 2\rho_{x,z}^2 s_z^2)}{S_z^2} \]

\[ = \frac{1}{n} \frac{S_z^2 [1 + \rho^2 (S_z^2 / S_y^2) - \rho_{x,z}^2]}{S_z^2} \]
It is observed from (3.14) that the MSE of \( t \) depends on the scalar \( \phi \). So the value \( \phi \) will effect the MSE of \( t \). Thus one should be very cautious about the selection of the value of scalar \( \phi \). Assuming linear relationship between \( Y \) and \( X \) Gupta et al. (2012) suggested the following regression estimator for the population mean \( \bar{Y} \) as

\[
\hat{\mu}_{\text{Reg}} = \bar{z} + \hat{\beta}_{\text{z,x}}(\overline{X} - \overline{X}) \tag{3.15}
\]

where \( \hat{\beta}_{\text{z,x}} \) is the sample regression coefficient between \( Z \) and \( X \) and \( \overline{X} = Y + X \) is the scrambled response on \( Y \). Setting \( \phi = 1 \) in (3.10), one can easily get the bias of Gupta et al.'s (2012) regression estimator \( \hat{\mu}_{\text{Reg}} \) as

\[
\text{Bias}(\hat{\mu}_{\text{Reg}}) = \frac{1}{n}(1 - \frac{1}{n}) \beta_{\text{z,x}} \left[ \frac{\mu_z(x, z_x)}{s_z} - \frac{\mu_z(x, z_x)}{s_z} \right] \tag{3.16}
\]

where \( \mu_z(x, z_x) = E[(x - \overline{X})^2(\overline{Z} - \overline{Z})] \) and \( \mu_z(x, z_x) = E(x - \overline{X})^2 \).

The mean square of the Gupta et al.'s (2012) regression estimator \( \hat{\mu}_{\text{Reg}} \) to the first degree of approximation is given by

\[
\text{MSE}(\hat{\mu}_{\text{Reg}}) = \frac{1}{n}(1 - \frac{1}{n}) \beta_{\text{z,x}} \left[ \frac{\mu_z(x, z_x)}{s_z} - \frac{\mu_z(x, z_x)}{s_z} \right] \tag{3.17}
\]

which can be also obtained from (3.14) just by setting \( \phi = 1 \).

Efficiency Comparisons

From (1.1) and (3.14) we have

\[
\text{MSE}(\hat{\mu}_{\text{Reg}}) = \frac{1}{n}(1 - \frac{1}{n}) \beta_{\text{z,x}} \left[ \frac{\mu_z(x, z_x)}{s_z} - \frac{\mu_z(x, z_x)}{s_z} \right] \tag{3.18}
\]

which is always positive if

\[
|\phi| > 0 \tag{3.19}
\]

i.e. if

\[
|\phi| < 1 \tag{3.19}
\]

From (1.3), (2.8), (3.14) and (3.17) we have

\[
\text{MSE}(\hat{\mu}_{\text{Reg}}) = \frac{1}{n}(1 - \frac{1}{n}) \beta_{\text{z,x}} \left[ \frac{\mu_z(x, z_x)}{s_z} - \frac{\mu_z(x, z_x)}{s_z} \right] > 0 \tag{3.20}
\]

\[
\text{MSE}(\hat{\mu}_{\text{Reg}}) = \frac{1}{n}(1 - \frac{1}{n}) \beta_{\text{z,x}} \left[ \frac{\mu_z(x, z_x)}{s_z} - \frac{\mu_z(x, z_x)}{s_z} \right] > 0 \tag{3.21}
\]

> 0 provided \( R \neq \beta_{yx} \)

\[
\text{MSE}(\hat{\mu}_{\text{Reg}}) = \frac{1}{n}(1 - \frac{1}{n}) \beta_{\text{z,x}} \left[ \frac{\mu_z(x, z_x)}{s_z} - \frac{\mu_z(x, z_x)}{s_z} \right] > 0 \tag{3.22}
\]

> 0 provided \(|\phi| < 1\)

It follows from (3.19), (3.20), (3.21) and (3.22) that the proposed estimator \( t_{\alpha} \) is more efficient than:

(i) The conventional unbiased estimator \( \bar{Z} \)

(ii) The usual unbiased estimator \( \bar{Z}_\phi \).

(iii) The ratio estimator \( t_{\alpha} \) considered by Sousa et al. (2010) unless \( R = \beta_{yx} \), the case where both the estimators \( t_{\alpha} \) and \( t_{\alpha} \) are equally efficient.

Empirical Study

To judge the merits of the suggested regression estimator \( t_{\alpha} \) over Gupta et al. [2] regression estimator \( \hat{\mu}_{\text{Reg}} \) we have computed the percent relative efficiency of the suggested estimator \( t_{\alpha} \) with respect to Gupta et al.'s (2012) estimator \( \hat{\mu}_{\text{Reg}} \) by using the formula:

\[
\text{PRE}(t_{\alpha}, \hat{\mu}_{\text{Reg}}) = \frac{(1 + \phi^2)(1 - \rho_{yx})}{(1 + \phi^2)(1 - \rho_{yx})} \times 100 \tag{3.23}
\]

Under the assumption \( S_y \cong S_x \) and \( S_x = \alpha S_x \), where \( \alpha \) is a scalar in percent (i.e. \( \alpha \% \)), the \( \text{PRE}(t_{\alpha}, \hat{\mu}_{\text{Reg}}) \) reduces to

\[
\text{PRE}(t_{\alpha}, \hat{\mu}_{\text{Reg}}) = \frac{(1 + \alpha u - \rho_{yx})}{(1 + \phi^2)(1 - \rho_{yx})} \times 100 \tag{3.24}
\]

We have computed the values of \( \text{PRE}(t_{\alpha}, \hat{\mu}_{\text{Reg}}) \) in (3.24) for \( \alpha = 10 \% \), 20 \% , 30 \% and \( \rho_{yx} = 0.55, (0.6) 0.9 \) and the finding are depicted in Tables 4-6.

Tables 4-6 clearly indicate that the values of \( \text{PRE}(t_{\alpha}, \hat{\mu}_{\text{Reg}}) \) are larger than 100. So the proposed regression estimator \( t_{\alpha} \) more efficient than that of Gupta et al. [2] regression estimator \( \hat{\mu}_{\text{Reg}} \) when \( |\phi| < 1 \). There is considerable gain in efficiency by using the proposed regression estimator \( t_{\alpha} \) over Gupta et al.'s (2012) regression estimator \( \hat{\mu}_{\text{Reg}} \) when the value of \( \phi \) is in the neighborhood of ‘origin’, the value of \( \rho_{yx} \) is closer to ‘unity’ and the value of \( \alpha \) is moderately large. Thus in such situations our recommendation is to use the proposed regression estimator \( t_{\alpha} \) as long as \( |\phi| < 1 \).

Table 4: The values of \( \text{PRE}(t_{\alpha}, \hat{\mu}_{\text{Reg}}) \) for different values of \( (\alpha, \phi, \rho_{yx}) \).

| \( \phi \) | \( \rho_{yx} \) | -1.00 | -0.75 | -0.50 | -0.25 | 0.00 | 0.25 | 0.50 | 0.75 | 1.00 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.55 | 100.00 | 100.62 | 101.07 | 101.34 | 101.43 | 101.34 | 101.07 | 100.62 | 100.00 |
| 0.60 | 100.00 | 100.68 | 101.17 | 101.46 | 101.56 | 101.46 | 101.17 | 100.68 | 100.00 |
| 0.70 | 100.00 | 100.85 | 101.46 | 101.84 | 101.96 | 101.84 | 101.96 | 100.85 | 100.00 |
| 0.80 | 100.00 | 101.20 | 102.07 | 102.60 | 102.78 | 102.60 | 102.07 | 101.20 | 100.00 |
| 0.90 | 100.00 | 102.24 | 103.90 | 104.92 | 105.26 | 104.92 | 103.90 | 102.24 | 100.00 |

How to cite this article: Tanveer AT, Housila PS. An Improved Estimation Procedure of the Mean of a Sensitive Variable Using Auxiliary Information. Biostat Biometrics Open Acc J. 2017; 3(2): 555607. DOI: 10.19080/BBOAJ.2017.03.555607
Table 5: The values of $PRE(t_{R\mu}, \hat{\mu}_{Reg})$ for different values of $(\alpha, \phi, \rho_{yx})$.

| Estimator | $\phi$ | $\rho_{yx}$ | -1.00 | -0.75 | -0.50 | -0.25 | 0.00 | 0.25 | 0.50 | 0.75 | 1.00 |
|-----------|-------|-------------|-------|-------|-------|-------|------|------|------|------|------|
| $\hat{\mu}_{Reg}$ | 0.55  | 100.00      | 102.43| 104.24| 105.36| 105.73| 105.36| 104.24| 102.43| 100.00|
|           | 0.60  | 100.00      | 102.64| 104.62| 105.84| 106.25| 105.84| 104.62| 102.64| 100.00|
|           | 0.70  | 100.00      | 103.29| 105.77| 107.32| 107.84| 107.32| 105.77| 103.29| 100.00|
|           | 0.80  | 100.00      | 104.58| 108.11| 110.34| 111.11| 110.34| 108.11| 104.58| 100.00|
|           | 0.90  | 100.00      | 108.24| 115.00| 119.48| 121.05| 119.48| 115.00| 108.24| 100.00|

Table 6: The values of $PRE(t_{R\mu}, \hat{\mu}_{Reg})$ for different values of $(\alpha, \phi, \rho_{yx})$.

| Estimator | $\phi$ | $\rho_{yx}$ | -1.00 | -0.75 | -0.50 | -0.25 | 0.00 | 0.25 | 0.50 | 0.75 | 1.00 |
|-----------|-------|-------------|-------|-------|-------|-------|------|------|------|------|------|
|           | 0.55  | 100.00      | 105.26| 109.38| 112.00| 112.90| 112.00| 109.38| 105.26| 100.00|
|           | 0.60  | 100.00      | 105.70| 110.19| 113.07| 114.06| 113.07| 110.19| 105.70| 100.00|
|           | 0.70  | 100.00      | 107.02| 112.68| 116.36| 117.65| 116.36| 112.68| 107.02| 100.00|
|           | 0.80  | 100.00      | 109.59| 117.65| 123.08| 125.00| 123.08| 117.65| 109.59| 100.00|
|           | 0.90  | 100.00      | 116.36| 131.76| 143.13| 147.37| 143.13| 131.76| 116.36| 100.00|

Acknowledgement

The authors are thankful to the Editor-in-Chief, and to the anonymous learned referee for his valuable suggestions regarding improvement of the paper.

References

1. Sousa R, Shabbir J, Real PC, Gupta S (2010) Ratio estimation of the mean of a sensitive variable in the presence of auxiliary information. Jour Statist Theo and Pract 4(3): 495-497.
2. Gupta S, Shabbir J, Sousa R, Real PC (2012) Estimation of mean of a sensitive variable in the presence of auxiliary information. Commun. Statist Theo and Metho 41: 2394-2404.
3. HimmeFarb S, Edgell SE (1980) Additive constant model: a randomized response technique for eliminating evasiveness to quantitative response questions. Psych Bull 87: 525-530.
4. Hussain Z (2012) Improvement of Gupta and Thornton scrambling model through double use of randomization device. Int Jour Acad Res Bus Soc Sci 2(6): 91-97.
5. Gjestvang CR, Singh S (2009) An improved randomized response model: Estimation of mean. Jour Appl Statist 36(12): 1361-1367.
6. Sukhatme PV, Sukhatme BV (1970) Sampling theory of surveys with applications. (2nd Rev Edn) Ind Soci Agri Statist, New Delhi, India.

Your next submission with Juniper Publishers

will reach you the below assets

- Quality Editorial service
- Swift Peer Review
- Reprints availability
- E-prints Service
- Manuscript Podcast for convenient understanding
- Global attainment for your research
- Manuscript accessibility in different formats (Pdf, E-pub, Full Text, Audio)
- Unceasing customer service

Track the below URL for one-step submission
https://juniperpublishers.com/online-submission.php