Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF
AGRICULTURAL STATISTICS 72(1) 2018 33–37

Robust Estimation in Stratified Sampling under Error-in-Variables Super Population Model

Shweta Chauhan, B.V.S. Sisodia and Dhirendra Singh
Narendra Deva University of Agriculture & Technology, Faizabad

Received 21 July 2017; Revised 09 January 2018; Accepted 23 February 2018

SUMMARY

In the present paper, robustness of the model based estimator of finite population total under error-in-variables super population model in stratified sampling has been investigated. An empirical study with real data revealed that the stratified balance sampling has minimized the percent loss in precision to the great extent due to measurement error in y along with protection against the deviation of the assumed/working model.

Keywords: Robustness, Super population model, Finite population, Stratified balance sampling, Measurement error.

1. INTRODUCTION

Measurement error is likely to occur when measuring instruments are inaccurate or biased in finite population survey sampling. If the respondents are human population, the respondents may not provide exact information because it is based on recall-basis, which may creep up response error that come under measurement error. Fuller (1987) quoted some characteristics such as age, income, unemployment etc which are generally subject to measurement error/ response error. He reported that the measurement error is about 15 per cent of the total variation for income. Bolfarine (1991) considered regression models that take such errors into account. However, he considered that samples came from bivariate normal population with constant model-error variance. Chattopadhyay and Datta (1994) extended the work of Bolfarine to stratified sampling by considering the location error-in-variables super population model. Various research workers have dealt with such situations in the past. Notably among them are Battese *et al.* (1988), Mukhopadhyay (1994), Eltinge (1994), Stefanski (2000), Ghosh & Sinha (2007), etc.

When we consider the model-based/model assisted estimation of finite population total or mean of the study variate y, it has been found generally in most of socio-economic surveys that variance of y is a function of the auxiliary variable x related to y. The structure of the variance function is generally as $V\left(y_i\right)=\sigma^2 x_i^g$, $1/2 \leq g \leq 2$, for most of the data encountered in practice, where σ^2 is variance of error term in the model (see the work of Rao and Bayless; 1969, Bayless and Rao, 1970 etc). Also for instance, Royall (1970, 1971) and Royall and Herson (1973a, 1973b) have assumed g=1. Scott *et al.* (1978) have considered g=2. Royall and Herson (1973b) have extended their work of 1973a to the stratified sampling when stratification is done on size variable (x) by assuming the following super population model

$$y_{hi} = \beta_h x_{hi} + e_{hi} x_{hi}^{1/2}, \ \text{i=1, 2, …, } N_h\text{, h=1, 2, …, H}$$

$$V\left(y_{hi}\right) = \sigma_e^2 x_{hi} \tag{1.1}$$

They referred the model (1.1) as ξ -model, denoted as $\xi\left(0,1;x_h\right)$. The strata were formed as follows: The N_1 units whose x-values are smallest form stratum 1, the next N_2 smallest units from stratum 2, and so

Corresponding author: B.V.S. Sisodia
E-mail address: bvssisodia@gmail.com

on, such that $\sum_{h=1}^{H} N_h = N$. Thus, no unit in stratum h is larger than any units in stratum h+1. They suggested a model-based separate ratio estimator of $T = \sum_{h=1}^{H}\sum_{i=1}^{N_h} y_{hi}$, the finite population total, under the model (1.1). The estimator of T they developed is

$$\hat{T} = \sum_{h=1}^{H} \hat{T}_h \tag{1.2}$$

where $\hat{T}_h = \sum_{i\in s_h} y_{hi} + \hat{\beta}_h \sum_{i\in \bar{s}_h} x_{hi}$, $\hat{\beta}_h = \sum_{i\in s_h} y_{hi} / \sum_{i\in s_h} x_{hi}$, a best linear unbiased estimator (BLUE) of β_h, s_h is a sample of n_h units and $\bar{s}_h$ is compliment of s_h in stratum h such that $s_h \cup \bar{s}_h = N_h$ and $\sum_{h=1}^{H} n_h = n$. The model variance of $\hat{T}$ is given by

$$V\left(\hat{T}\right) = \sigma_e^2 \sum_{h=1}^{H} \frac{\sum_{i\in \bar{s}_h} x_{hi}}{\sum_{i\in s_h} x_{hi}} \sum_{i=1}^{N_h} x_{hi} \tag{1.3}$$

They also proved a theorem, which is stated below without proof

Theorem 1: If $n_h \propto N_h \bar{x}_h^{1/2}$, i.e. $n_h = \dfrac{n N_h \bar{x}_h^{1/2}}{\sum_{h=1}^{H} N_h \bar{x}_h^{1/2}}$, where $\bar{x}_h$ is stratum mean of x in stratum h, then under the general polynomial model $\xi\left(\delta_o, \delta_1, ..., \delta_J : x_h\right)$ of degree J, the strategy $\left[s^{+}\left(J\right), \hat{T}\right]$ is more efficient than the strategy $\left[s\left(J\right), \hat{T}\left(0,1:x\right)\right]$, where $\hat{T}\left(0,1:x\right)$ is the estimator of T under the model $\xi\left(0,1:x\right)$ when there is no stratification of the population. Note that s(J) and s^{+}(J) are simple balance and stratified balance sample (see, Royall and Herson, 1973a, 1973b).

Under optimum allocation in Theorem 1 and stratified balance sampling, the variance of $\hat{T}$ in (1.3) reduces to

$$V(\hat{T})_{opt} = \frac{\sigma_e^2}{n}\left[\left(\sum_{h=1}^{H} N_h \bar{x}_h^{1\!/\!2}\right)^2 - n N \bar{X}\right] \tag{1.4}$$

Sisodia *et al.* (2015) have made an attempt to study the effect of measurement error in y under the model (1.1). A simulation study conducted by them showed that standard error of the estimate got inflated by about 8 to 10 percent depending upon the ratio $\delta = \dfrac{\sigma_v^2}{\sigma_e^2}$, where σ_v^2 and σ_e^2 are variances of measurement error and model error, respectively.

In the present paper, an attempt has been made to examine the robustness of the predictor of finite population total if the models considered by Sisodia *et al.* (2015) deviates, i.e. these models are not perfectly correct rather some other models are correct.

2. ROBUSTNESS OF THE ESTIMATOR UNDER THE MODEL $\xi(0,1:x_h)$ WHEN STUDY VARIATE IS SUBJECT TO MEASUREMENT ERROR :

We consider the following error-in-variable super population model

$$y_{hi} = \beta_h x_{hi} + e_{hi} x_{hi}^{1/2}\ \text{, i = 1,2,…, } N_h \text{, h = 1,2…,H}$$

$$Y_{hi} = y_{hi} + v_{hi} \tag{2.1}$$

where y_{hi} and Y_{hi} are true and observed value of the study variate y, respectively. e_{hi} and v_{hi} are model and measurement error, respectively, with $E\left(e_{hi}\right) = E\left(v_{hi}\right) = 0$, $v\left(e_{hi}\right) = \sigma_e^2$, $v(y_{hi}) = \sigma_e^2 x_{hi}$ and $v(v_{hi}) = \sigma_v^2$ for all i and h. It is also assumed that e_{hi} and v_{hi} are mutually independently distributed. The model (2.1) is referred to as ξ-model and denoted as $\xi(0,1:x_h)$. The objective is to estimate $T = \sum_{h=1}^{H}\sum_{i=1}^{N_h} y_{hi}$, the finite population total of y.

Sisodia *et al.* (2015) have shown that the estimator $\hat{T}_1 = \sum_{h=1}^{H} \hat{T}_h$ under the model 2.1 is model unbiased estimator of T, where $\hat{T}_h$ is model based unbiased estimator of $T_h = \sum_{i=1}^{N_h} y_{hi}$, given by

$$\hat{T}_h = \sum_{i\in s_h} Y_{hi} + \hat{\beta}_h \sum_{i\in \bar{s}_h} x_{hi} \tag{2.2}$$

where $\hat{\beta}_h = \sum_{i\in s_h} Y_{hi} \Big/ \sum_{i\in s_h} x_{hi}$, which is least square estimate of β_h under model (2.1). Variance of $\hat{T}_1$ is given by

$$V(\hat{T}_1) = \sum_{h=1}^{H}\left[\sigma_e^2 \frac{\sum_{i=1}^{N_h} x_{hi}}{\sum_{i\in s_h} x_{hi}} \sum_{i\in \bar{s}_n} x_{hi} + \sigma_v^2 n_h \left(\frac{\sum_{i=1}^{N_h} x_{hi}}{\sum_{i\in s_h} x_{hi}}\right)^2\right] \tag{2.3}$$

Now, suppose that the working model $\xi(0,1:x_h)$ is not true but the true model is

$$y_{hi}=\alpha_h+\beta_h x_{hi}+e_{hi}x_{hi}^{1/2}\text{, i=1,2…N}_\text{h}\text{, h =1,2…,H}$$

$$Y_{hi}=y_{hi}+v_{hi} \qquad \text{…(2.4)}$$

where α_h is y-intercept and notations assumptions and definitions are same as in model (2.1). The model 2.4 is referred to as ξ -model and denoted as $\xi(1,1:x_h)$.

We wish to examine the property of $\hat{T}_1$ under the model $\xi(1,1:x_h)$. If the estimator $\hat{T}_1$ is used under the model $\xi(1,1:x_h)$, it is pertinent to examine whether $\hat{T}_1$ holds same properties as in model $\xi(0,1:x_h)$ even under the model $\xi(1,1:x_h)$. If not, we need to find out some criteria under which $\hat{T}_1$ holds same properties even under model $\xi(1,1:x_h)$. For this we derive the expectation of $\hat{T}_1$ and its model variance under model $\xi(1,1:x_h)$.

The model expectation of estimator $\hat{T}_1$ is given by

$$E\left[\hat{T}_1-T\right]=E\left[\sum_{h=1}^{H}\hat{T}_h-\sum_{h=1}^{H}T_h\right]$$

$$=E\left[\sum_{h=1}^{H}\left(\sum_{i\in s_h}Y_{hi}+\frac{\sum_{i\in s_h}Y_{hi}}{\sum_{i\in s_h}x_{hi}}\sum_{i\in\bar{s}_h}x_{hi}-\sum_{i=1}^{N_h}y_{hi}\right)\right]$$

$$=\sum_{h=1}^{H}E\left[\sum_{i\in s_h}\left(\alpha_h+\beta_h x_{hi}+e_{hi}x_{hi}^{1/2}+v_{hi}\right)+\frac{\sum_{i\in s_h}\left(\alpha_h+\beta_h x_{hi}+e_{hi}x_{hi}^{1/2}+v_{hi}\right)}{\sum_{i\in s_h}x_{hi}}\sum_{i\in\bar{s}_h}x_{hi}-\sum_{i=1}^{N_h}\left(\alpha_h+\beta_h x_{hi}+e_{hi}x_{hi}^{1/2}\right)\right]$$

$$E\left[\hat{T}_1-T\right]=\sum_{h=1}^{H}\alpha_h\left[n_h\frac{\sum_{i\in\bar{s}_h}x_{hi}}{\sum_{i\in s_h}x_{hi}}-\left(N_h-n_h\right)\right] \qquad (2.5)$$

Since $E\left[\hat{T}_1-T\right]\neq 0$, therefore, the estimator ˆ is biased if it is used in $\xi(1,1:x_h)$. Further, we have derived the mean square error of $\hat{T}_1$ as

$$MSE\left(\hat{T}_1\right)=\sum_{h=1}^{H}E\left(\hat{T}_h-T_h\right)^2$$

$$MSE\left(\hat{T}_1\right)=\sum_{h=1}^{H}\alpha_h^2\left[n_h\frac{\sum_{i\in\bar{s}_h}x_{hi}}{\sum_{i\in s_h}x_{hi}}-\left(N_h-n_h\right)\right]^2+\sigma_e^2\sum_{h=1}^{H}\left[\frac{\sum_{i\in\bar{s}_h}x_{hi}}{\sum_{i\in s_h}x_{hi}}\sum_{i=1}^{N_H}x_{hi}\right]+n_h\sigma_v^2\sum_{h=1}^{H}\left[\left(\frac{\sum_{i=1}^{N_h}x_{hi}}{\sum_{i\in s_h}x_{hi}}\right)^2\right]$$

(2.6)

The first term of (2.6) is square of bias and second term is the variance of $\hat{T}_1$ under the model $\xi(1,1:x_h)$

3. CHARACTERISTICS OF THE SAMPLES IN WHICH THE ESTIMATOR $\hat{T}_1$ IS UNBIASED UNDER THE MODEL $\xi(1,1:x_h)$:

Consider the bias expression (2.5) of $\hat{T}_1$ under the model $\xi(1,1:x_h)$ i.e

$$B\left[\hat{T}_1\right]=\sum_{h=1}^{H}\alpha_h\left[n_h\frac{\sum_{i\in\bar{s}_h}x_{hi}}{\sum_{i\in s_h}x_{hi}}-(N_h-n_h)\right] \qquad (3.1)$$

The above bias reduces to zero if $\overline{x}_{\bar{s}_h}=\overline{x}_{s_h}=\overline{x}_h$, i.e. if the sample is stratified balance sample, denoted as $s_h=s_h(1)$ as referred by Royall & Herson (1973b).

Thus, we summarize the above results in the form of following theorem.

Theorem 3.1: When the estimator $\hat{T}_1$ developed under the model $\xi(0,1:x_h)$ is used in the model $\xi(1,1:x_h)$, it remains unbiased under stratified balance sampling with variance

$$V\left[\hat{T}_1\right]=\sigma_e^2\sum_{h=1}^{H}\frac{N_h\left(N_h-n_h\right)}{n_h}\overline{x}_h+\sigma_v^2\sum_{h=1}^{H}\frac{N_h^2}{n_h} \qquad (3.2)$$

The variance given in (3.2) can further be minimized for optimum allocation of n_h. We minimize this for given cost function with respect to n_h

$$c=\sum_{h=1}^{H}n_h c_h \qquad (3.3)$$

where c_h is the cost of enumeration per unit in the hth stratum. The function to be minimized is

$$\varphi(n_h,\lambda)=\sigma_e^2\left[\sum_{h=1}^{H}N_h\left(\frac{N_h}{n_h}-1\right)\overline{x}_h+\delta\sum_{n=1}^{H}N_h^2\Big/n_h\right]+\lambda(c-\sum n_h c_h)$$

(3.4)

Differentiating (3.4) with respect to n_h and equating to zero, we get a solution for optimum value of n_h as follows

$$n_n=\frac{nN_h\sqrt{\overline{x}_h+\delta}\Big/\sqrt{c_h}}{\sum N_h\sqrt{\overline{x}_h+\delta}\Big/\sqrt{c_h}} \qquad (3.5)$$

For c_h= c for all h, it reduces to Neyman type allocation as

$$n_h = \frac{n.N_h\sqrt{\bar{x}_h+\delta}}{\sum N_h\sqrt{\bar{x}_h+\delta}} \qquad (3.6)$$

Under the allocation (3.6), the variance expression (3.2) comes out to be

$$V\left[\hat{T}_1\right]_{opt} = \frac{\sigma_e^2}{n}\left[\sum_{h=1}^{H}\left(\frac{N_h\bar{x}_h}{\sqrt{\bar{x}_h+\delta}}\right)\sum_{h=1}^{H}N_h\sqrt{\bar{x}+\delta} - nN\bar{x} + \delta\sum_{h=1}^{H}\left(\frac{N_h}{\sqrt{\bar{x}_h+\delta}}\right)\left(\sum_{h=1}^{H}N_h\sqrt{\bar{x}_h+\delta}\right)\right]$$

$$= \frac{\sigma_e^2}{n}\left[\sum_{h=1}^{H}\frac{N_h}{\sqrt{\bar{x}_h+\delta}}(\bar{x}_h+\delta)\sum_{h=1}^{H}N_h\sqrt{\bar{x}_h+\delta} - nN\bar{X}\right]$$

$$V\left[\hat{T}_1\right]_{opt} = \frac{\sigma_e^2}{n}\left[\left(\sum_{h=1}^{H}N_h\sqrt{\bar{x}_h+\delta}\right)^2 - nN\bar{X}\right] \qquad (3.7)$$

We summarize the above results in the form of the following theorem.

Theorem 3.2: When the estimator $\hat{T}_1$ developed under the model $\xi(0,1:x)$ is used in the model $\xi(1,1:x_h)$, it remains unbiased under stratified balance sampling. Its optimum variance under allocation (3.6) is given by

$$V\left[\hat{T}_1\right]_{opt} = \frac{\sigma_e^2}{n}\left[\left(\sum_{h=1}^{H}N_h\sqrt{\bar{x}_h+\delta}\right)^2 - nN\bar{X}\right]$$

4. RELATIVE EFFICIENCY OF THE ESTIMATOR :

The relative efficiency of $\hat{T}_1$ with measurement error in y as compared to $\hat{T}_1$ without measurement in y is obtained as

$$E = \frac{V\left(\hat{T}\right)_{opt}}{V\left(T_1\right)_{opt}}$$

$$E = \frac{\left(\sum_{h=1}^{H}N_h\bar{x}_h^{1/2}\right)^2 - nN\bar{x}}{\left(\sum_{h=1}^{H}N_h(\bar{x}_h+\delta)^{1/2}\right)^2 - nN\bar{X}} \qquad (4.1)$$

Obviously, the value of E is less than one. Therefore, there is a loss in precision of the estimator if there is measurement error. The percent loss in precision due to measurement in error in y is obtained as

$$L(y) = (1-E)\times 100$$

$$L(y) = \frac{\left(\sum_{h=1}^{H}N_h(\bar{x}_h+\delta)^{1/2}\right)^2 - \left(\sum_{h=1}^{H}N_h\bar{x}_h^{1/2}\right)^2}{\left(\sum_{h=1}^{H}N_h(\bar{x}_h+\delta)^{1/2}\right)^2 - nN\bar{X}}\times 100 \qquad (4.2)$$

It is evident from the above expression that the percent loss in precision depends on stratum size and its mean, δ, population size N, overall sample size n and population mean $\bar{X}$.

5. SOME NUMERICAL FINDINGS AND CONCLUDING REMARKS

In order to assess the percent loss in precision, it has been computed with some real data.

To find out the value of L(y), we use the data in Sukhatme and Sukhatme, 1970, page: 152, Table 4.1 of the example 4.1 on pages 150-151. The data are related to a sample survey for estimation of livestock numbers conducted in Etawah district during the year 1951. The auxiliary variable x related to the study variable y (livestock number) is agricultural area. The data according to the range of agricultural area are classified into the classes (strata): 0-100, 101-200, 201-300, 301-400, 401-600, 601-1000 and greater than 1000 acres.

From the analysis of the data, Sukhatme & Sukhatme (1970) have pointed out that the relationship between Y_{hi} and x_{hi} (h = 1, 2, …, H, the strata) is approximately linear and passes through the origin, i.e. $E\left(Y_{hi}/x_{hi}\right) = \beta_h x_{hi}$. They have also indicated that $V(Y_{hi}/x_{hi})$ appears to vary as x increases up to 1000 acres but not beyond that. That means the data up to 1000 acres satisfies the model given in (2.1). Therefore, using the data in the Table 4.1 up to 1000 acres, the values of L(y) have been computed for different values of $\delta = \frac{\sigma_v^2}{\sigma_e^2}$=0.75, 1.00 and 1.25, and sample size n=64 and 80. The population size N=319. The population mean $\bar{X} = 367.5$.The results are presented in the Table 5.1

Table 5.1. Percent loss in precision L(y) for different value of δ , n=64 and 80, and N=319 (population size) :

δ	L(y) in percent	
	n=64	**n=80**
0.75	0.29	0.32
1.00	0.40	0.43
1.25	0.50	0.54

It is evident from the results of the above table that percent loss in precision is quite marginal below one percent. A simulation study conducted by Sisodia *et al.* (2015) without stratified balance sampling had shown that the percent losses in precision due to measurement error in y were between 8 to 10 percent. Therefore, it is very obvious that use of stratified balance sampling has two-fold advantages (i) it protects the optimality of the estimator against the deviation of the assumed model and (ii) it minimizes the percent loss in precision of the estimate to the great extent if there is measurement error in y.

REFERENCES

Battese, G.E., Harter, R.M. and Fuller, W.A. (1988). An error components models for predicition of county crop areas using survey and satellite data*. J. Amer. Statist. Assoc.*, **83**, 28-36.

Bayless, D.L. and Rao, J.N.K. (1970). An empirical study of stabilities of estimators and variance estimators in unequal probabilities sampling (n = 3 or 4). *J. Amer. Statist. Assoc.*, **65,** 1645-1667.

Bolfarine, H. (1991). Finite population prediction under error-in-variables super population models. *The Canadian J. Statist.,* **19,** (2), 191-207.

Chattopadhyay, M.K. and Datta, G.S. (1994). Finite population prediction for stratified sampling under error-in-variables superpopulation models. Sankhya : *Ind. J. Statist.,* **56**, Series B, pt 3, pp. 415-433.

Eltinge L. Jhon (1994). A Finite population approach to approximating small measurement error effects in regression *Sankhya: Ind. J. Statist..* **56,** Series(B), pt 2; pp 234-250.

Fuller,W.A. (1987). Measurement error models. *Wiley*, New York.

Ghosh, M. and Sinha, K. (2007). Empirical bayes estimation in finite population sampling under functional measurement error models. *J. Statist. Plg. Inf.* **137** : 2759-2773

Mukhopadhyay, P. (1994). Prediction in finite population under error-in-variables superpopulation models. *J. Statist. Plg. Inf.*, **41,** 151-161.

Rao, J. N. K. and Bayless, D. L. (1969). An empirical study of stabilities of estimators and variance estimators in unequal probability sampling of two units per stratum. *J. Amer. Statist. Assoc.,* **64**, 540-559.

Royall, R.M. (1970). On finite population sampling theory under certain linear regression models, *Biometrika*, **57**(2), 377-387.

Royall, R.M. (1971). Linear regression models in finite population sampling theory, in Godambe, V.P. and Sprott, D.A. eds. Foundations of Statistical Inference, Toronto Holt, Rinehart and Winston of Canada Ltd.

Royall, R. M. and Herson, J. (1973a). Robust estimation in finite populations. I. *J. Amer. Statist. Assoc.*, **68,** 880-889.

Royall, R. M. and Herson, J. (1973b). Robust estimation in finite populations II. Stratification on a size variable. *J. Amer. Statist. Assoc.*, **68,** 890-893.

Sisodia, B.V.S., Singh, Amar, Mourya, K.K. and Rai, V.N. (2015). Estimation of finite population total in stratified sampling under error-in-variables super population model. *J. Ind. Soc. Agril. Statist.*, **69**(2), 127-133.

Scott, A.J., Brewer, K.R.W. and Ho, E.W.H. (1978). Finite population sampling and robust estimation. *J. Amer. Statist. Assoc.*, **73**, 359-361.

Stefanski, L.A. (2000). Measurement error model. *J. Amer. Statist. Assoc.„* **95**(452), 1353-1358.

Sukhatme, P.V. and Sukhatme, B.V. (1970). Sampling theory of surveys with applications. Second revised edition, Iowa State University Press, Ames, Iowa, U.S.A.