Journal of Rajasthan Academy of Physical Sciences
ISSN: 0972-6306; URL: http://raops.org.in
Vol.20, No.1&2, January-June, 2021, 49-56

LINEAR REGRESSION ESTIMATOR IN CASE OF INCOMPLETE SAMPLING FRAME

P.C. Gupta[1], Vohita Joshi[1], Pankaj Nagar[2] and Ajeet Kumar Singh[3]
[1]Professor (Rtd.), Veer Narmad South Gujarat University, Gujrat
[2]Research Scholar, Department of Statistics, University of Rajasthan, Jaipur
[3]Associate Professor, Department of Statistics, University of Rajasthan, Jaipur (Correspondence)
[4]Assstt.Professor, Department of Statistics, University of Rajasthan, Jaipur

Abstract: The paper suggests the improvement in estimator(Y) of population mean given by Agarwal & Gupta (2008) in case of incomplete sampling frame. Authors have used the auxiliary information(X) in terms of linear regression estimator. Bias and mean square error are obtained. It is shown theoretically and numerically that the proposed estimator is more efficient than the above mentioned estimator.

Keywords: Linear regression, bias, mean square error.

1. Introduction

Sampling plays an important role in daily life. This is due to its importance many sampling schemes were developed. The main purpose of developing various sampling schemes is basically to make the estimators of population parameter(s) to be more efficient. Besides developing sampling scheme the other way to improve the suchestimator is an application of auxiliary information. Ratio, Product, Ratio-cum Product and Regression methods of estimation have a great application in improving the efficiency of an estimation while using this auxiliary information. It has also been observed that Ratio Estimator works better when the two characteristics (variables) are positively correlated. But when situations are not like that then the regression estimator of the characteristic (Y, say) may be a batter choice.

The use of regression method of estimation had been briefly mentioned by Mahalonobis [10] and the theoretical basis of the regression method had been discussed in detail by Cochran [5], Choudhary and Arnav [4], Tikkiwal [14], Zarkovic [15], Mickey [11], Kuldorf [9] and Chaubey *et al*. [3]. Sometimes sampling frame was not available completely. Hansen and Hurwitz [7] has attempted first time to apply regression method of estimation in case of incomplete sample. Hansen *et al*. [8] gave predecessor-successor method of obtaining information on the missing units in the sampling frame. Singh [12] gave mathematical formulation to Hansen *et al*. method. Recently Agarwal and Gupta

[1,2], Gupta [6] contribution, to theory of estimation with incomplete sampling frame, is worthy for further development in this field of sampling.

The proposed estimator is a linear combination of regression estimator based on samples drawn from the part of population with complete frame and sample mean based on sampled values of sub-sample obtained from the part of population with incomplete frame.

2. Sampling Procedure

Let the units in the target population is "N". Among these N_1 units are in the given frame and N_2 units are left out of the frame, such that $N = N_1 + N_2$. From N_1 units n_1 units are selected through SRSWOR, then a frame of n_2 units is prepared which are occurring in between the selected units and next to it from N_2 units. From these selected n_2 units, information are gathered from $n_{2'}$ units selected with simple random sampling without replacement.

3. Estimation

Let Y_1, Y_2 and Y are respectively population total of value of characteristics of units, in the frame, units not in the frame and units of target population. y_{n_1}, $y_{n_2}$and $y'_{n_2}$are total for characteristics(Y) for n_1, n_2 and n'_2 units which are sampled from N_1, N_2, and n_2 units respectively.

The proposed estimator $\bar{y}_{wlr}$ is a weighted linear regressionestimator based on units sampled from existing frame and units sampled from the units which were between the sampled units and the next unit. We define our estimator based on (a) known value of regression coefficient, say $\beta = \beta_o$ and (b) and regression coefficient obtained from sample, $\beta_o = b$. So the estimator can be represented as follows:

$$\bar{y}_{wlr} = w_1\bar{y}_l + w_2\bar{y}_{n'_2} \tag{1}$$

where

$\bar{y}_l = \bar{y}_{n_1} + \beta_0(\bar{X}_{N1} - \bar{x}_{n1})$

$W_1 = N_1/N$; $W_2 = N_2/N$;

$W_1 = N_1/N$; $W_2 = N_2/N$;

$w_1 = n_1/n$; $w_2 = n_2/n$;

Est.$(W_1) = w_1$ Est.$(W_2) = w_2$;

$h = n_2/n'_2$ (1)

3.1 Bias of Proposed Estimator

3.1.1 When β_o is pre-assigned known constant value

In this case $\bar{y}_l$ is explained as follows:

$$\bar{y}_l = \bar{y}_{n_1} + \beta_0(\bar{X}_{N1} - \bar{x}_{n1}) \tag{2}$$

Where $\bar{y}_{n_1}$ is a mean of total of characteristic(Y) in the sample of size n_1 from N_1. $\bar{X}_{N_1}$ and $\bar{x}_{n_1}$ are population mean and sample mean of auxiliary characteristic(X) corresponding to Y_1 . β is a regression coefficient of Y_1 on X_1.

Using concept, given by **Sukhatme et al. [13], the** conditional expectation of proposed estimator is,

$$E(\bar{y}_{wlr}) = E_1E_2(\bar{y}_{wlr}|n_1, n_2)$$

$$= E_1E_2(w_1\bar{y}_{wlr}|n_1, n_2) + E_1E_2\left(w_2\bar{y}_{n_2'}|n_{2'}\right) \quad (3)$$

Where $E_1 and E_2$ are respectively the expectations over the first step of randomization for n_1 observations and second step of randomization over n_2 observations.

$$\mathrm{E_1E_2(w_1\bar{y}_l) = W_1\bar{Y}_{N_1}} \quad (4)$$

$$\mathrm{E_1E_2\left(w_2\bar{y}_{n_2}'\right) = W_2\bar{Y}_{N_2}} \quad (5)$$

Using (1), (4) and (5) in (4), we have

$$\mathrm{E(\bar{y}_{wlr}) = \bar{Y}_N} \quad (6)$$

Thus $(\bar{y}_{wlr})$ is an unbiased estimator.

3.1.2 When β_o is unknown constant.

In this situation β_o is replaced by sample regression coefficient (b). Therefore, in this case $\bar{y}_l$ is explained as follows:

$$\bar{y}_l = \bar{y}_{n1} + b(\bar{X}_{N1} - \bar{x}_{n1}) \text{ using estimated value of } \beta_o = b \quad (7)$$

To derive an approximation for estimate of proposed estimator, let us assume

$$\left\{\begin{array}{c} \varepsilon_0 = \frac{\bar{y}_{n1} - \bar{Y}_{N1}}{\bar{Y}_{N1}} \\ \varepsilon_1 = \frac{\bar{x}_{n1} - \bar{X}_{N1}}{\bar{X}_{N1}} \\ \varepsilon_2 = \frac{s_{(x1y1)} - S_{(X1Y1)}}{S_{(X1Y1)}} \\ \varepsilon_3 = \frac{s_{x1}^2 - S_{X1}^2}{S_{X1}^2} \end{array}\right\} \quad (8)$$

Such that $\mathrm{E(\varepsilon_i) = 0;}$ for $\mathrm{i = 0,1,2,3.}$

Also

$$E(\varepsilon_0^2) = \left(\frac{1-f_1}{n_1}\right)\frac{S_{Y1}^2}{\bar{Y}_{N1}^2};$$

$$E(\varepsilon_1^2) = \left(\frac{1-f_1}{n_1}\right)\frac{S_{X1}^2}{\bar{X}_{N1}^2}$$

$$E(\varepsilon_0\varepsilon_1) = \left(\frac{1-f_1}{n_1}\right)\frac{S_{X1Y1}}{\overline{Y}_{N1}\overline{X}_{N1}}$$

Now let,

$\bar{y}_{n_1} = \overline{Y}_{N_1}(1+\epsilon_0)$

$\bar{x}_{n_1} = \overline{X}_{N_1}(1+\epsilon_1)$

$s_{x_1y_1} = S_{X_1Y_1}(1+\epsilon_2)$

$s_{x_1}^2 = S_{X_1}^2(1+\epsilon_3)$

Then $E_1E_2(w_1\bar{y}_l) = W_1\overline{Y}_{N_1} - W_1\frac{\beta(1-f_1)}{n}\left(\frac{\mu_{21}}{S_{(XY)1}} - \frac{\mu_{30}}{S_{X1}^2}\right);\ f_1 = \frac{n_1}{N_1}$ (9)

and, $E_1E_2\left(w_2\bar{y}_{n_2'}\right) = W_2\bar{Y}_{N_2}$ (10)

So using (3), (9) and (10) we get,

$$E(\bar{y}_{wlr}) = \overline{Y}_N - W_1\frac{\beta(1-f_1)}{n}\left(\frac{\mu_{21}}{S_{(XY)_1}} - \frac{\mu_{30}}{S_{X_1}^2}\right) \quad (10)$$

So, bias in $(\bar{y}_{wlr})$ is negative with its magnitude

$$\text{Bias}(\bar{y}_{wlr}) = W_1\frac{\beta(1-f_1)}{n}\left(\frac{\mu_{21}}{S_{(XY)1}} - \frac{\mu_{30}}{S_{X1}^2}\right) \quad (11)$$

where,

$$\mu_{rs} = E[(X-\overline{X})^r(Y-\overline{Y})^s] \quad (12)$$

3.2 Variance of the proposed estimator

Then the conditional variance of proposed estimator is,

$$V(\bar{y}_{wlr}) = V(\bar{y}_{wlr}|n_1,n_2) = V_1E_2(\bar{y}_{wlr}) + E_1V_2(\bar{y}_{wlr}) \quad (13)$$

Now,

$$V_1E_2(\bar{y}_{wlr}) = V_1E_2(\bar{y}_{wlr}|n_1,n_2) = V_1E_2\left[w_1\left(\bar{y}_{n1} + b(\overline{X}_{N1} - \bar{x}_{n1})\right) + w_2\bar{y}_{n_2'}\right]$$

$$= V_1\left[w_1\left(\bar{y}_{n1} + b\left(\overline{X}_{N_1} - \bar{x}_{n_1}\right)\right)\right] + W_2^2V_1\left(\bar{y}_{n_2}\right)$$

$$= V_1\left[w_1\left(\bar{y}_{n1} + b\left(\overline{X}_{N_1} - \bar{x}_{n_1}\right)\right)\right] \quad (14)$$

Using large sample approximation we have

$$V_1E_2(\bar{y}_{wlr}) = V_1\left[w_1\{\overline{Y}_{N1}(1+\varepsilon_0) - \overline{X}_{N1}\frac{S_{(X1Y1)}}{S_{X1}^2}(\varepsilon_1)(1+\varepsilon_2)(1+\varepsilon_3)^{-1}\}\right]$$

$$= V_1\left[w_1\left\{\overline{Y}_{N_1}(1+\varepsilon_0) - \beta\overline{X}_{N_1}(\varepsilon_1 - \varepsilon_1\varepsilon_3 + \varepsilon_1\varepsilon_2)\right\}\right]$$

$$= W_1\left(\frac{1-f_1}{n}\right)\left(S_{Y_1}^2 + \beta_1^2 S_{X_1}^2 - 2\beta_1 S_{X_1Y_1}\right) \quad (16)$$

Secondly,

$$E_1V_2(\bar{y}_{wlr}) = E_1V_2(\bar{y}_{wlr}|n_1, n_2) = W_2\left(\frac{h-1}{n}\right)S_{Y2}^2 \quad (17)$$

Substituting (16) & (17) in(13) we get,

$$V(\bar{y}_{wlr}) = W_1\left(\frac{1-f_1}{n}\right)\left(S_{Y1}^2 + \beta_1^2 S_{X1}^2 - 2\beta_1 S_{X1Y1}\right) + W_2\left(\frac{h-1}{n}\right)S_{Y2}^2$$

Or,

$$V(\bar{y}_{wlr}) = W_1\left(\frac{1-f_1}{n}\right)S_{Y1}^2(1-\rho_1^2) + W_2\left(\frac{h-1}{n}\right)S_{Y2}^2 \quad (18)$$

and the estimate of the variance of proposed estimator is

$$\hat{V}(\bar{y}_{wlr}) = w_1\left(\frac{1-f_1}{n}\right)s_{y1}^2(1-{r_1}^2) + w_2\left(\frac{h-1}{n}\right)s_{y2}^2 \quad (19)$$

4. Comparison with Other Estimator

For same layout estimator given by Agarwal & Gupta [2] is:

$$\bar{y}_{AG} = w_1\bar{y}_{n_1} + w_2\bar{y}_{n_2'} \quad (20)$$

with variance

$$V(\bar{y}_{AG}) = W_1\left(\frac{1-f_1}{n}\right)S_{Y_1}^2 + W_2\left(\frac{h-1}{n}\right)S_{Y_2}^2 \quad (21)$$

From equations (18) and (21) we have

$$V(\bar{y}_{AG}) - V(\bar{y}_{wlr}) = W_1\left(\frac{1-f_1}{n}\right)S_{Y_1}^2\rho_1^2 \quad (22)$$

which is always non negative. Hence estimator $\bar{y}_{wlr}$ is more efficient than the $\bar{y}_{AG}$ Agarwal & Gupta (2008).

5. Numerical Illustration

In support of our findings, we here explain it with a hypothetical example of the data of 50 factories related with number of workers(X) and number of absentees(Y) as the data for incomplete frame is not available in any form.

X: NUMBER OF WORKERS

Y: NUMBER OF ABSENTEES

Table 1

S.N	1	2*[1]	3*[2]	4	5*[3]	6	7	8*[4]	9	10*[5]
X	95	79	30	45	28	142	125	81	43	53
Y	9	7	3	2	3	8	9	10	6	2

S.N	11	12	13*[6]	14	15*[7]	16	17	18*[8]	19	20*[9]
X	148	89	57	132	47	43	116	65	103	52
Y	16	4	5	13	4	9	12	8	9	8
S.N	21*[10]	22*[11]	23*[12]	24	25*[13]	26*[14]	27	28	29	30
X	57	64	75	69	63	83	124	31	96	42
Y	14	6	6	8	5	7	13	2	23	13
S.N	31*[15]	32*[16]	33	34	35	36	37	38	39*[17]	40*[18]
X	85	91	73	159	54	69	61	164	132	82
Y	18	14	7	18	13	14	1	35	21	5
S.N	41	42	43	44*[19]	45	46*[20]	47	48	49	50
X	33	86	41	50	80	50	62	72	105	90
Y	4	11	10	9	20	15	9	18	12	7

*represent the Y_i units which were not available in the frame.

N= 50; N_1=30; N_2=20;

$\bar{X} = 78.32;\ \bar{Y} = 10.10;$

$\bar{X}_1 = 86.4;\ \bar{X}_2 = 66.20;$

$\bar{Y}_1 = 11.1667;\ \bar{Y}_2 = 8.5.$

An estimate of total number of workers in factories under study with the help of sample values x is given below by predecessor-successor method:

Table-2

R.N.	11	19	20	23	27
M_i	4	0	0	0	0
X	103	54	69	33	62

$\bar{m} = \frac{1}{n_1}\sum_{i=1}^{n_1} M_i = \frac{4}{5}$=0.8 $\Rightarrow N_1\bar{m} = 24$, $T_X = 30(1 + 0.8)44.2 = 2396.8 \approx 2397$

Estimated number of workers, in the factories, are 2397.

In population, the values of M_i are: 2,1,1,1,1,1,1,,4,2,2,2,1,1.

Then, $S_M^2 = \frac{1}{N_1 - 1}\sum_{i=1}^{N_1}(M_i - \bar{M})^2$ = 0.7692

$V(\bar{m}) = \left(\frac{1}{n_1} - \frac{1}{N_1}\right)S_M^2 = 0.0256$

and, $S.E.(\bar{m}) = 0.1601$

6. Estimation of absentees in the factories under study

The estimate of absentees using our estimate can be found according to layout as:

N= 50; N_1= 30; N_2=20;n=15; n_1 = 9; n_2 = 6; $n_{2'}$ = 4.

Sample values for n_1 =9

Table-3

R.N	3	5	20	8	6	10	11	26	22	Mean
x_{n_1}	142	43	69	132	148	116	103	80	164	110.778
y_{n_1}	8	6	14	13	16	12	9	20	35	14.778

Sample values for n_2 =6

Table-4

R.N.	6	8	4	9	1	7	Mean
x_{n_2}	57	65	81	52	79	47	63.5
y_{n_2}	5	8	10	8	7	4	7

Sample values for $n_{2'}$ =4

Table-5

R.N	6	3	4	1	Mean
$x_{n_2'}$	47	81	52	57	59.25
$y_{n_2'}$	4	10	8	5	6.75

7. Calculations

7.1 Variance and estimate of variance

$S_Y^2 = 41.2347$
$S_{Y_1}^2 = 48.4885$
$S_{Y_2}^2 = 27.8421$
$S_{X_1Y_1} = 148.9655$
$S_{X_1}^2 = 1503.60$
$s_y^2 = 60.5238$
$s_{y_1}^2 = 75.6944$
$s_{y_2}^2 = 4.8$
$s_{x_1}^2 = 3160.667$
$s_{x_1y_1} = 171.1298$

$\rho_1 = 0.5517$
$\rho_1^2 = 0.3044$
$r_1 = 0.4877$
$r_1^2 = 0.2379$
$W_1 = 0.6$
$W_2 = 0.4$
$w_1 = 0.6$
$w_2 = 0.4$
$f_1 = 0.3$
$f_2 = 0.3$
$f = 0.3$

$V(\bar{y}_{wlr}) = 1.3156$
$\hat{V}(\bar{y}_{wlr}) = 1.6792$
$SE(\bar{y}_{wlr}) = 1.1469$
$\widehat{SE}(\bar{y}_{wlr}) = 1.2958$
$V(\bar{y}_{AG}) = 1.7289$
$\hat{V}(\bar{y}_{AG}) = 2.1834$
$SE(\bar{y}_{AG}) = 1.3148$
$\widehat{SE}(\bar{y}_{AG}) = 1.4776$

8. Conclusions

From the above results we conclude that our proposed estimator $\bar{y}_{wlr}$is more efficient than $\bar{y}_{AG}$.

Acknowledgement: The authors are thankful to the Referee for valuable comments and suggestions.

References

[1] Agarwal, B. and Gupta, P.C. (2007). Synergism in incomplete sampling design. Management change! 183-190.

[2] Agarwal, B. and Gupta, P.C. (2008). Estimation from incomplete sampling frame in case of simple random sampling. MASA (USA), **3**, 113-117.

[3] Chaubey, Y.P., Singh, M. and Dwivedi, T.D. (1984). A note on an optimal property of regression estimator. Bio. Journal, **26**(4), 465-467.

[4] Choudhary, A. and Arnab, R. (1982). On unbiased product-type estimators. Journal of Indian Society of Agricultural Statistics; **34**, 65-69.

[5] Cochran, W.G. (1942). Sampling Theory when sample units are of unequal size. JASA, **37**, 199-212.

[6] Gupta, P.C. (2013). A review of the problem of estimation from incomplete sampling frame. JRSA, **2**(2) (1993), 8-13.

[7] Hansen, M.H. and Hurwitz, W.N. (1943). On the theory of sampling from finite population. AMS, **14**, 333-362.

[8] Hansen, M.H., Hurwitz, W.N. and Jabine, T.H. (1963). The use of imperfect lists for probability sampling at the Bureau of Census. Invited paper, 34th session, Int. Inst. Ottawa, Canada, 497-517.

[9] Kulldorf (1963). Some problems of optimum allocation for sampling on two occasions. Rev. Stat. Inst., **31**(1), 24-57.

[10] Mahalonobis, P.C. (1941). Report on the sampling Techniques for Forecasting the Bark-yield of Cinchona. Plants Experiment Series B, I.S.I.

[11] Mickey, M.R. (1959). Some finite population unbiased ratio and regression estimators. JASA, 54, 594-612. Rev. Stat. Inst., **24**, 52-63.

[12] Singh, R. (1983). On the use of incomplete frames in sample surveys. Bio. Journal **25**(6); 545-549.

[13] Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S. and Ashok, C. (1984). Sampling Theory of surveys with applications, Iowa State University Press, Ames, Iowa, U.S.A., III revised edition.

[14] Tikkiwal, B.D. (1960). On the theory of classical regression and double sampling method. JRSS, (B), **22**, 131-135.

[15] Zarkovic, S.S. (1956). An illustration of some characteristic situations in the application of the difference estimate. Rev. Stat. Inst., **24**, 52-63.