STATISTICS IN TRANSITION new series, June 2020
Vol. 21, No. 2, pp. 189–200, DOI 10.21307/stattrans-2020-020
Submitted – 26.09.2019; accepted for publishing – 25.03.2020

New linear model for optimal cluster size in cluster sampling

Alok Kumar Shukla[1], **Subhash Kumar Yadav**[2]

ABSTRACT

In this paper, a nonlinear model is proposed for improving the relationship between the size of a cluster and the variance within the cluster. This model describes the most appropriate functional relation between the within-cluster variance and the cluster size. Through this model, we can obtain the optimum size of a cluster and an estimate of the variance between clusters. The proposed model leads to further improvement in the estimation of the optimum size of a cluster, and the formula for the determination of optimum cluster size leads to explicit solution of models.

Key words: Non-linear models, optimum cluster size, four-parameter model, variance function.

1. Introduction

Regression analysis is widely used for better explanation and future prediction of any phenomenon which is assumed to develop in some patterns whether in economics or any other field. In cluster sampling, it is of interest to find the most suitable functional relationship between variance within the cluster (S_w^2) and the cluster size (M) for prediction [Singh and Chaudhary (2009)]. Smith (1938), Jessen (1942), Hansen and Hurwitz (1942), Mahalanobis (1940, 1942), Misra et al. (2010), Tiwari and Misra (2011), Shukla et al. (2013), Shukla and Yadav (2016), Lawson and Skinner (2017) etc., have discussed the problem of determining the optimum cluster size in two important contexts of variance function and cost function respectively. Scarneciu *et al.* (2017) compared various nonlinear models in determining pulmonary pressure in hyperthyroidism. Kaplan and Gurcan (2018) compared different growth curves using non-linear regression function in Japanese quail. Riazoshams *et al.* (2019) described in detail the robust nonlinear regression models with applications using R

[1] Department of Statistics, D. A-V. College, Kanpur-208001, U.P., India. E-mail: alokshukladav@gmail.com. ORCID: https://orcid.org/0000-0001-9797-7894.
[2] Corresponding Author, Department of Statistics, Babasaheb Bhimrao Ambedkar University, Lucknow-226025, U.P., India. E-mail: drskystats@gmail.com. ORCID: https://orcid.org/0000-0002-7181-8075.

software. All the functional relations given by the above authors are of similar functional form describing the relationship between the size of the cluster and variance within the cluster. It is well established in cluster sampling that sampling variance increases as the cluster size increases and it decreases with the number of clusters. The cost also decreases as cluster size decreases and increases as the number of clusters increases. Thus, it becomes important to seek a balancing point through the optimum size of the cluster and the number of clusters in the sample by minimizing the variance for a given/fixed cost or vice-versa.

Let $\bar{y}$ denote the sample mean for the characteristic y under study for the sample size n. We know that in cluster sampling, the variance of the sample mean $\bar{y}$ is given by

$$V(\bar{y}) = \frac{1-f}{n} S_b^2 \tag{1}$$

where $f = \frac{n}{N}$ is finite population correction and $S_b^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\bar{Y}_i - \bar{\bar{Y}})^2$ is the variance between cluster means,

where $\bar{Y}_i = \frac{Y_i}{M}$ is the mean of the i^{th} cluster for the characteristic under study and $\bar{\bar{Y}} = \frac{1}{N}\sum_{i=1}^{N}\bar{Y}_i = \frac{Y}{NM}$ is the population mean of NM units for N clusters each of size M with $Y = \sum_{i=1}^{N} Y_i$ as population total for the study variable.

Now, it is of crucial importance to know the behaviour of $V(\bar{y})$ with the cluster size M. This involves knowing the relationship between S_b^2 and M. Through analysis of variance (ANOVA) technique, S_b^2 can be found if we know

(i) The Total variance S^2 between all the NM elements in the population,

(ii) The variance S_w^2 within all M elements of the same cluster for all N clusters,

where S^2 and S_w^2 are respectively given by

$$S^2 = \frac{1}{NM-1}\sum_{i=1}^{N}\sum_{j=1}^{M}(Y_{ij} - \bar{\bar{Y}})^2 \text{ and } S_w^2 = \frac{1}{N(M-1)}\sum_{i=1}^{N}\sum_{j=1}^{M}(Y_{ij} - \bar{Y}_i)^2 .$$

Thus, the total variance S^2 of all population units can be written in the form of S_b^2 and S_w^2 as

$$S^2 = [M(N-1)S_b^2 + N(M-1)S_w^2]/(NM-1) \tag{2}$$

If N is large, we express the above equation (2) as

$$S^2 = S_b^2 + (M-1)S_w^2/M \tag{3}$$

Hence,

$$S_b^2 = S^2 - (M-1)S_w^2/M\ . \tag{4}$$

Thus, S_b^2 also depends on S_w^2.

We are considering the problem of determining the best relationship between variance function and the size of the cluster.

Jessen (1942) suggested the following relationship for S_w^2 and M by a non-linear form of model as

$$S_w^2 = \alpha M^{\beta},\ M > 1 \tag{5}$$

where α and β are the parameters of the above non linear model.

Misra *et al.* (2010) established the relationship between S_w^2 and M through an asymptotic regression model given as

$$S_w^2 = \alpha + \beta \rho^M,\ M > 1 \tag{6}$$

where α, β and ρ are the parameters of the asymptotic regression model.

Tiwari and Misra (2011) suggested a three-parameter linear regression model for the relationship between S_w^2 and M as

$$S_w^2 = \alpha + \beta M + \rho/M,\ M > 1 \tag{7}$$

where α, β and ρ are the parameters to be estimated for the above linear regression model.

2. Suggested model

In the present paper, we have proposed the following four-parameter model for the most appropriate relationship between S_w^2 and M as

$$S_w^2 = \alpha + \beta M + \delta M^2 + \rho/M,\ M > 1 \tag{8}$$

Expression (8) is a linear model in which its parameters are appearing linearly and there is no problem in assuming an additive error term in model (8).

The above model can be postulated as a statistical model as

$$S_w^2 = \alpha + \beta M + \delta M^2 + \rho / M + e_i,\ M > 1,\ i = 1,2,3,.....n \tag{9}$$

where the random variable $e_i's$ are assumed to be independently and identically normally distributed with mean zero and fixed variance σ^2 and α, β, δ and ρ are the parameters of the model (9).

2.1. Fitting of models

Draper and Smith (1998) have classified the above models in two groups; one as intrinsically linear and another as intrinsically non-linear models. The model (5) is an intrinsically linear model as it can be transformed into a linear model. Model (6) is purely nonlinear model as it cannot be transformed by means of any transformation into a linear model. The OLS method is not directly applied for estimating the parameters of model (6). The parameters of model (6) are estimated through the iterative procedure as Levenberg-Marquardt's method. Model (7) and the proposed model (8) are linear in parameters so their parameters are estimated by the method of least squares.

2.2. Goodness of fit of different models

Coefficient of Determination - R^2

The assessment of the regression model is to observe how much of the total sum of squares (TSS) has fallen into the sum of squares due to the regression (SSR).

$$R^2 = \frac{SSR}{TSS}$$

Adjusted Coefficient of Determination - R^2_{Adj}

Montgomery *et al.* (2012) have described R^2_{Adj} considering good for model comparison when the number of parameters is not equal in two models.

$$R^2_{Adj} = 1 - \left(\frac{n-1}{n-p}\right)\left(1 - R^2\right)$$

Residual Mean Square- s^2

The residual mean square is defined as

$$s^2 = \frac{SSE}{n-p}$$

where n is the number of observations and p is the number of the model parameters used and SSE is the sum of squares due to errors. A small value of s^2 reflects the appropriateness of the fitted model.

***Mean Absolute Error* (M.A.E)**

The Mean Absolute Error is defined as

$$M.A.E. = \frac{\sum_{i=1}^{n} |errors|}{n}$$

where n is the number of observations. A smaller $M.A.E.$ is preferred in fitting of various models.

***Akaike Information Criterion* (*A.I.C.*)**

Gujarati and Sangeetha (2007) have given a lot of importance to Akaike Information Criterion (A.I.C.), defined as,

$$A.I.C. = Exp\left(\frac{2p}{n}\right)\frac{RSS}{n}$$

where n is the number of observations and p is the number of parameters. RSS is residual or error sum of square.

2.3. Examination of Residuals

Analysis of the residuals (errors) is strongly recommended to decide about the suitability of a model by Draper and Smith (1998). Three important assumptions of the model are:

(i) Errors are not auto correlated.

(ii) Errors are independent.

(iii)Errors are normally distributed.

The assumptions can be verified by examining the residuals.

Test for auto correlation of errors (Durbin-Watson Test)

We test H_0 : Errors are not auto correlated (if DW test values > d_U)

Against H_1 : Errors are auto correlated (if DW test values < d_L)

where d_L and d_U are given in Draper and Smith (1998). DW Test values greater than 1.72 times d_U confirm that there is no problem of auto-correlation.

Test for independence of errors (Run Test)

We test H_0 : Errors are independent.

Against H_1 : Errors are not independent.

Test for normality (Shapiro-Wilk Test,* $n < 50$*)

We test H_0 : Errors are normally distributed.

Against H_1 : Errors are not normally distributed.

2.4. Determination of Variance Function

If the total population is considered as a single cluster containing *NM* elements, the S_w^2 will be equal to the total variance S^2. Thus, for the proposed model, we have

$$S^2 = \alpha + \beta NM + \delta (NM)^2 + \rho / NM \tag{10}$$

The between-cluster variance for the proposed model is obtained by putting (8) and (10) in (4) as,

$$\hat{S}_b^2 = [\hat{\alpha} + \hat{\beta} NM + \hat{\delta}(NM)^2 + \hat{\rho} / NM] - \frac{(M-1)}{M}[\hat{\alpha} + \hat{\beta} M + \hat{\delta} M^2 + \hat{\rho} / NM] \tag{11}$$

where $\hat{\alpha}$, $\hat{\beta}$, $\hat{\delta}$ and $\hat{\rho}$ are the estimated values of the parameters of the suggested model.

The variance of the sample mean of the characteristic under study through the suggested model in cluster sampling can be obtained as

$$V(\bar{y}) = \frac{1-f}{n} \hat{S}_b^2 \tag{12}$$

3. Empirical study

The appropriateness and model adequacy of various models have been examined by using two natural data sets from Sukhatme *et al.* (1984) and Govindarajulu (1999) respectively. The S_w^2 have been calculated for different sizes of the clusters in (acre)², with the study variable as the area under wheat crop. We have computed the estimated values of parameters, goodness of fit and residuals analysis for the models (5)-(8) given in Table-1(a) and Table-1(b). The Estimated values of S_w^2 and S_b^2 are given in Table-1(c) and Table-1(d). These values are given in Table-2(c) and Table-2(d) for the models (5)-(8) respectively. The above values have been obtained using SPSS 17.0 Statistical software.

Table-1(a). Parameter estimates for various models

Model	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\delta}$	$\hat{\rho}$
Model (5)	78.886	0.0473	-	-
Model (6)	108.171	31.530	-	0.948
Model (7)	93.813	0.012	-	-32.888
Model (8)	88.1502	0.4272	-0.0003	-21.8678

Table-1(b). Goodness of fit of models and Residuals Analysis

	Model (5)	Model (6)	Model (7)	Model (8)
R^2	0.941	0.983	0.984	0.999
R^2_{Adj}	0.921	0.966	0.978	0.998
s^2	10.479	4.410	2.756	0.0248
M.A.E.	2.078	1.208	0.9	0.055
A.I.C.	13.6643	5.8573	3.6594	0.0245
DW#	1.3457	2.2104	2.3238	3.4161
R*	0.001 (1.000)	0.109 (0.913)	0.001 (1.000)	1.200 (0.230)
W^	0.9917 (0.9784)	0.8994 (0.4037)	0.9713 (0.8749)	0.9952 (0.9926)

is Durbin and Watson Test values, * is Run test values, ^ is Shapiro-Wilk test values, the p-values are given in parentheses.

Table-1(c). Estimated S_w^2 for various models

M	Observed value of S_w^2	Estimated value of S_w^2 for model (5)	Estimated value of S_w^2 for model (6)	Estimated value of S_w^2 for model (7)	Estimated value of S_w^2 for model (8)
2	78.10	81.53	79.84	77.39	78.0695
4	84.28	84.25	82.71	85.64	84.3868
8	88.92	87.05	87.60	89.80	88.8127
16	93.50	89.95	94.75	91.96	93.5308
NM =1176	108.33	110.22	108.17	108.34	108.33

Table-1(d). Estimated S_b^2 from equation (4) for various models

M	Observed value of S_b^2 from equation (4)	Estimated value of S_b^2 for model (5)	Estimated value of S_b^2 for model (6)	Estimated value of S_b^2 for model (7)	Estimated value of S_b^2 for model (8)
2	69.28	69.45	68.25	69.64	69.2952
4	45.12	47.03	46.13	44.11	45.0399
8	30.52	34.05	31.53	29.76	30.6188
16	20.69	25.89	19.34	22.12	20.6448

Table-2(a). Parameter estimates for various models

Model	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\delta}$	$\hat{\rho}$
Model (5)	0.072	0.182	-	-
Model (6)	0.372	-0.566	-	0.966
Model (7)	0.3163	0.0000061	-	-4.311
Model (8)	-0.3039	0.0132	-0.000001	2.3414

Table-2(b). Goodness of fit of models and Residuals Analysis

	Model (5)	Model (6)	Model (7)	Model (8)
R^2	0.768	0.982	0.959	0.996
R^2_{Adj}	0.710	0.971	0.932	0.991
s^2	0.0039	0.0004	0.0009	0.0001
M.A.E.	0.0433	0.0116	0.0183	0.0045
A.I.C.	0.0051	0.0004	0.0013	0.00006
DW#	0.9083	1.8057	1.6865	3.5122
R*	-1.369	0.001	0.001	1.369
	(0.171)	(1.000)	(1.000)	(0.171)
W^	0.9648	0.9577	0.9447	0.9759
	(0.8577)	(0.8127)	(0.7175)	(0.9217)

is Durbin and Watson Test values, * is Run test values, ^ is Shapiro-Wilk test values, the p-values are given in parentheses.

Table-2(c). Estimated S_w^2 for various models

M	Observed value of S_w^2	Estimated value of S_w^2 for model (5)	Estimated value of S_w^2 for model (6)	Estimated value of S_w^2 for model (7)	Estimated value of S_w^2 for model (8)
15	0.05	0.1176	0.0361	0.0289	0.0501
20	0.08	0.1239	0.0898	0.1008	0.0770
25	0.11	0.1290	0.1349	0.1440	0.1193
30	0.18	0.1334	0.1728	0.1727	0.1694
35	0.22	0.1372	0.2046	0.1933	0.2239
NM =8820	0.37	0.3747	0.3717	0.3727	0.3699

Table-2(d). Estimated S_b^2 from equation (4) for various models

M	Observed value of S_b^2 from equation (4)	Estimated value of S_b^2 for model (5)	Estimated value of S_b^2 for model (6)	Estimated value of S_b^2 for model (7)	Estimated value of S_b^2 for model (8)
15	0.320	0.265	0.338	0.346	0.3231
20	0.294	0.257	0.286	0.277	0.2967
25	0.264	0.251	0.242	0.234	0.2553
30	0.196	0.246	0.205	0.206	0.2061
35	0.156	0.241	0.173	0.185	0.1523

4. Results and discussion

From Table-1(b) and Table-2(b), it is easily evident that the value of R^2 for the competing models ranges from [0.941 0.984] and [0.768 0.982] respectively for Data Set-1 and Data Set-2, while that for the suggested model is 0.999 and 0.996 respectively. The value of R_{Adj}^2 for the competing models lies between [0.921 0.978] and [0.710 0.971] respectively while that for the suggested model between 0.998 and 0.991 respectively. The values of the s^2 for the competing models range from [2.756 10.479] and [0.0039 0.0004] while for the proposed model are 0.0248 and 0.0001 for Data Set-1 and Data Set-2 respectively. The values of M.A.E. are between [0.9 2.208] and [0.0183 0.0433] for the models in comparison while for the suggested models they are 0.055 and 0.0045 for Data Set-1 and Data Set-2 respectively. The values of A.I.C. lie between [3.6594 13.6643] and [0.0004 0.0051] for the competing models while these of proposed models are 0.0245 and 0.00006 for Data Set-1 and Data Set-2 respectively. Other measures are also better for the suggested model as compared to competing models.

Figure-1 and Figure-2 show the graph of R^2 and R_{Adj}^2 and s^2, M.A.E. and A.I.C. for the suggested and the competing models respectively for Data Set-1 while Figure-3 and Figure-4 for Data Set-2.

Figure-1. R^2 and R^2_{Adj} for data set-1 for various models

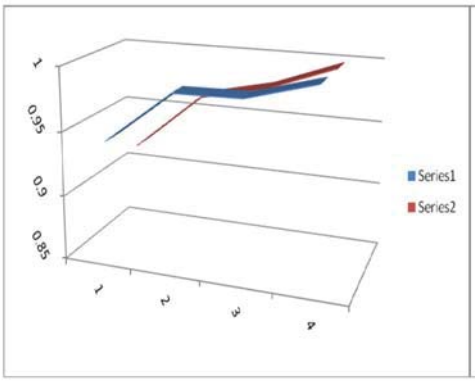


Figure-2. s^2, MAE and AIC for data set-1 for various models

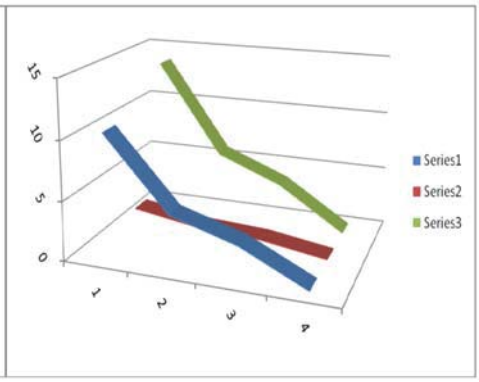


Figure-3. R^2 and R^2_{Adj} for data set-2 for various models

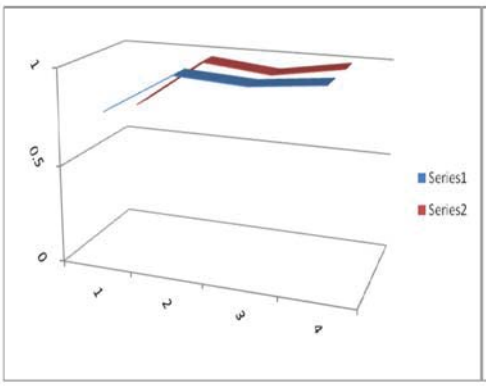
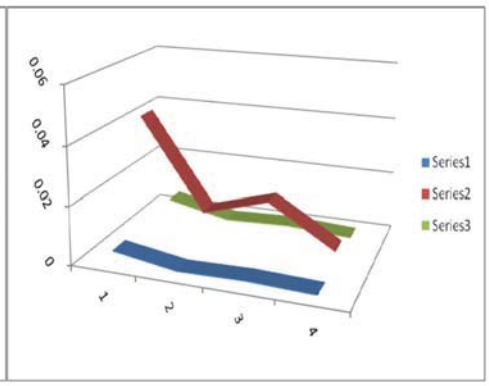


Figure-4. s^2, MAE and AIC for data set-2 for various models



5. Conclusion

In the present manuscript, we have proposed a four-parameter linear regression model for enhanced estimation of the variance function for clustered data. The parameters of the proposed model have been estimated through a well-known method of least squares. The proposed model and many other linear and nonlinear models have been fitted for the real data sets. The suggested model is compared with the competing linear and non-linear models. It has been shown that the proposed model fits well in comparison with other models for variance function in cluster sampling as it has

lesser mean residual error and other good measures of adequacy. Thus, it is recommended to use the proposed model for improved estimation of variance function, the relation between within-cluster variance and cluster size, between cluster variance and the cluster size in cluster sampling.

Acknowledgment

The authors expresses their heartfelt gratitude to the Editor and Editior-in-Chief along with the learned referees for critically examining the manuscript which improved the earlier draft.

REFERENCES

DRAPER, N. R., SMITH, H., (1998). Applied regression analysis. 3rd Ed., John Wiley & Sons,.

GUJARATI, D. N., SANGEETHA, (2007). Basic Econometrics, 4th Ed, Tata McGraw-Hill.

HANSEN, M. H., HURWITZ, W. N., (1942). Relative efficiencies of various sampling units in population enquiries, *Journal of American Statistical Association*, 37, pp. 89–94.

JESSEN, R. J., (1942 (). Statistical investigation of sample survey for obtaining farm facts, *Iowa Agricultural Experiment Station, Research Bulletin*, p. 304.

KAPLAN, S., GURCAN, E. K., (2018). Comparison of growth curves using non-linear regression function in Japanese quail, *Journal of Applied Animal Research*, 46, 1, pp. 112–117.

LAWSON, N., SKINNER, C., (2017). Estimation of a cluster-level regression model under nonresponse within clusters, *Metron*, 75, pp. 319–331.

MAHALANOBIS, P. C., (1940). A sample survey of acreage under jute in Bengal, *Sankhya*, 4, pp. 511–530.

MAHALANOBIS, P. C., (1942). General report on the sample census of area under jute in Bengal, *Indian Central Jute Committee*.

MISRA, G. C., YADAV, S. K., SHUKLA, A. K., RAJ, B., (2010). Use of a non-linear model for improved estimation in cluster sampling, *Journal of Reliability and Statistical Studies*, 3, 2, pp. 73–78.

MONTGOMERY D. C., PECK E. A., VINING G. C., (2012). Introduction to Linear Regression Analysis, 5th Ed., Wiley.

RIAZOSHAMS, H., MIDI, H., GHILAGABER, G., (2019). Robust Nonlinear Regression: with Applications using R, 1st Ed, Wiley.

SCARNECIU, C. C., SANGEORZAN, L., RUS, H., SCARNECIU, V. D., VARCIU, M. S., ANDREESCU, O., SCARNECIU, I., (2017). Comparison of linear and non-linear regression analysis to determine Pulmonary Pressure in Hyperthyroidism, *Pakistan Journal of Medical Sciences*, 33, 1, pp. 111–120.

SINGH, D., CHAUDHARY, F. S., (2009). Theory and Analysis of Sample Survey Designs, New Age International.

SMITH, H. F., (1938). An empirical law describing heterogeneity in the yields of agricultural crops, *Journal of Agricultural Science*, 28, pp. 1–23.

SUKHATME, P. V., SUKHATME, B. V., SUKHATME, S., ASOK, C., (1984). Sampling theory of surveys with applications, *Indian Society of Agricultural Statistics*.

SHUKLA, A.K., YADAV, S. K., (2016). Asymptotic Non-Linear Models for Uniformity Trial Experiments, *Elixir International Journal,* 94, 1, pp. 40042–40044.

SHUKLA, A. K., YADAV, S. K., MISRA, G. C., (2013). A Linear Model for Uniformity Trial Experiments, *Statistics in Transition-new series*, 14, 1, pp. 161–170.

TIWARI, R. B., MISRA, G. C., (2011). Estimation of optimum cluster size, IFRSA's *International Journal of Computing*, l, 4, pp. 717–722.

ZAKKULA G., (1999). Elements of Sampling Theory and Methods, Printice Hall Publication.