STATISTICS IN TRANSITION-new series, December 2011
Vol. 12, No. 3, pp. 579—586

A CLASS OF REGRESSION TYPE ESTIMATORS IN SURVEY SAMPLING

Govind Charan Misra[1], Subhash Kumar Yadav[2], Alok Kumar Shukla[1]

ABSTRACT

A class of linear regression models has been proposed for the estimation of population mean and total when information regarding auxiliary variate is available in survey sampling using regression method of estimation by introducing a new auxiliary variable z, which may also be a function of the auxiliary variable x. The proposed model leads to reduction in mean squared error as compared to ordinary regression method of estimation. The improvement has been demonstrated over ordinary regression estimator and also on ratio estimator with the help of an empirical example.

Key words: Auxiliary variable, Mean Squared Error, Ratio estimator, Regression type estimators.

1. Introduction

The intelligent use of auxiliary information for improving precision of the estimates has been done in sampling theory for different purposes. The auxiliary information has been used for the purposes of stratification in stratified sampling. In PPS (Probability Proportional to Size) sampling, the probabilities of selection of units are based on auxiliary information on measures of sizes. In ratio and regression methods of estimation, one uses auxiliary information so as to improve precision of the estimates of population parameters like population mean and population total etc. For detailed discussions on use of auxiliary information in sample surveys, reference can also be made of Sampath (2005) and Cochran (1999).

[1] Dept of Statistics, D.A-V College, Kanpur, India, e-mail: drmisragovind@gmail.com.
[2] Dept of Mathematics & Statistics, Dr R M L Avadh University, Faizabad, India, e-mail: drskystats@gmail.com.

In the regression method of estimation, the auxiliary variable X is correlated with the variable of interest Y and line of regression of Y on X does not pass through origin. The linear regression estimate of population mean of variable Y is defined as:

$$\bar{y}_{lr} = \bar{y} + b(\bar{X} - \bar{x}) \qquad (1)$$

Where $\bar{y}$ and $\bar{x}$ are sample mean of variable Y and X respectively. $\bar{X}$ is population mean of auxiliary variable X and is supposed to be known. $\bar{y}_{lr}$ is linear regression estimate of population mean and b is a constant Sukhatme *et al.* (1984) have described in detail , the procedures for deriving estimates of population parameters along with their biases, mean square error etc.

2. Proposed model

Motivated by Ekpenyong *et al.* (2008), we are proposing a class of linear regression type estimator by including a linear term in the ordinary linear regression estimator, which includes the estimators proposed by Ekpenyong *et al.* (2008) and Misra *et al.* (2009) as special cases. The proposed model seeks to consider following relationship among variable Y, auxiliary variables X and Z.

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + U \qquad (2)$$

Where β_0, β_1 and β_2 are parameters which appears linearly in the model (2). Z is another auxiliary variable which may also be taken as the function of X. When $Z = X^2$, it assumes the relationship considered by Ekpenyong *et al.* (2008). If $Z = \frac{1}{X}$, it takes the form of Misra *et al.* (2009). It has been shown by Misra *et al.* (2009) that their estimators of population mean and total are more efficient as compared to estimators of Ekpenyong *et al.* (2008) and ordinary linear regression estimator. U is independently and identically distributed random variable with mean zero and fixed variance σ^2.

The estimator of population mean based on (2) is given by:

$$\bar{y}_{gl} = \bar{y} - \hat{\beta}_1(\bar{x} - \bar{X}) - \hat{\beta}_2(\bar{z} - \bar{Z}) + U \qquad (3)$$

Where $\bar{y}_{gl}$ is a general regression type estimator of population mean based on proposed model defined in relation (2), $\bar{z}$ and $\bar{Z}$ are sample and population means of the variable Z.

3. Estimation of bias and variance of $\bar{y}_{gl}$

The bias and variance of $\bar{y}_{gl}$, the general regression type estimator of population mean can be calculated in following manner.

Let $\bar{y} = \bar{Y}(1+e_0)$, $\bar{x} = \bar{X}(1+e_1)$, $\hat{\beta}_1 = \beta_1(1+e_2)$, $\hat{\beta}_2 = \beta_2(1+e_3)$ and $\bar{z} = \bar{Z}(1+e_4)$

Such that $E(e_i) = 0 \quad \forall\ i = 0,1,2,3,4$ putting these values in equation (3), we get

$$\bar{y}_{gl} = \bar{Y}(1+e_0) - \beta_1(1+e_2)[\bar{X}(1+e_1) - \bar{X}] - \beta_2(1+e_3)[\bar{Z}(1+e_4) - \bar{Z}]$$
$$\bar{y}_{gl} = \bar{Y}(1+e_0) - \beta_1(1+e_2)[\bar{X}e_1] - \beta_2(1+e_3)[\bar{Z}e_4]$$
$$\bar{y}_{gl} = \bar{Y} + \bar{Y}e_0 - \beta_1(\bar{X}e_1 + \bar{X}e_1e_2) - \beta_2(\bar{Z}e_4 + \bar{Z}e_3e_4)$$
$$\bar{y}_{gl} = \bar{Y} + \bar{Y}e_0 - \beta_1\bar{X}(e_1 + e_1e_2) - \beta_2\bar{Z}(e_4 + e_3e_4)$$

Taking expectation on both the sides, we get

$$E(\bar{y}_{gl} - \bar{Y}) = \bar{Y}E(e_0) - \beta_1\bar{X}[E(e_1) + E(e_1e_2)] - \beta_2\bar{Z}[E(e_4) + E(e_3e_4)]$$
$$= -\beta_1\bar{X}\,E(e_1e_2) - \beta_2\bar{Z}\,E(e_3e_4)$$

$$Bias(\bar{y}_{gl}) = -Cov(\bar{x}, \hat{\beta}_1) - Cov(\bar{z}, \hat{\beta}_2)$$

This is negligible for large sample size. For large samples usually $Cov(\bar{x}, \hat{\beta}_1)$ decreases and it becomes zero if the joint distribution of y and x is bivariate normal. Similarly $Cov(\bar{z}, \hat{\beta}_2)$ vanishes if the joint distribution of y and z follows bivariate normal distribution. In this case the proposed regression estimator is exactly unbiased.

To the first order of approximation, by ignoring the terms with $e_i\, e_j\,(i \neq j) = 0,1,2,3,4$, we have

$$\bar{y}_{gl} - \bar{Y} = e_0\bar{Y} - e_1\beta_1\bar{X} - e_4\beta_2\bar{Z}$$

Therefore

$$V(\bar{y}_{gl}) = V(e_0\bar{Y} - e_1\beta_1\bar{X} - e_4\beta_2\bar{Z})$$

$$\begin{aligned} &= V(e_0\bar{Y}) + \beta_1^2 V(e_1\bar{X}) + \beta_2^2 V(e_4\bar{Z}) - 2\beta_1 Cov(e_0\bar{Y}, e_1\bar{X}) \\ &\qquad - 2\beta_2 Cov(e_0\bar{Y}, e_4\bar{Z}) + 2\beta_1\beta_2 Cov(e_1\bar{X}, e_4\bar{Z}) \\ &= V(\bar{y}) + \beta_1^2 V(\bar{x}) + \beta_2^2 V(\bar{z}) - 2\beta_1 Cov(\bar{y}, \bar{x}) \\ &\qquad - 2\beta_2 Cov(\bar{y}, \bar{z}) + 2\beta_1\beta_2 Cov(\bar{x}, \bar{z}) \end{aligned} \tag{4}$$

Therefore we get,

$$V(\bar{y}_{gl}) = \lambda\left[s_y^2 - 2\beta_1 s_{xy} + \beta_1^2 s_x^2 - 2\beta_2 s_{yz} + \beta_2^2 s_z^2 + 2\beta_1\beta_2 s_{xz} \right] \tag{5}$$

where $\lambda = \left(\frac{1}{n} - \frac{1}{N}\right)$

Here s_{xy}, s_{yz} and s_{zx} are estimators of the population covariances, S_{XY}, S_{YZ} and S_{ZX} respectively, while variances s_x^2, s_y^2 and s_z^2 are unbiased estimators of population variances S_X^2, S_Y^2 and S_Z^2 respectively.

We need to estimate β_1 and β_2 such that $V(\bar{y}_{gl})$ is a minimum. Using the method of ordinary least square, we differentiate partially (4) with respect to $\hat{\beta}_1$ and $\hat{\beta}_2$ and obtain following normal equations.

$$\hat{\beta}_2 \text{Cov}(\bar{x}, \bar{z}) + \hat{\beta}_1 V(\bar{x}) = \text{Cov}(\bar{y}, \bar{x}) \tag{6}$$

$$\hat{\beta}_2 V(\bar{z}) + \hat{\beta}_1 \text{Cov}(\bar{x}, \bar{z}) = \text{Cov}(\bar{y}, \bar{z}) \tag{7}$$

Solving (6) and (7) simultaneously, we obtain

$$\hat{\beta}_1 = \frac{\text{Cov}(\bar{y}, \bar{z})\text{Cov}(\bar{x}, \bar{z}) - \text{Cov}(\bar{y}, \bar{x})V(\bar{z})}{[\text{Cov}(\bar{x}, \bar{z})]^2 - V(\bar{x})V(\bar{z})}$$

$$\hat{\beta}_2 = \frac{\text{Cov}(\bar{y}, \bar{x})\text{Cov}(\bar{x}, \bar{z}) - \text{Cov}(\bar{y}, \bar{z})V(\bar{x})}{[\text{Cov}(\bar{x}, \bar{z})]^2 - V(\bar{x})V(\bar{z})}$$

And the expressions for variance of ordinary linear regression estimator and ratio estimator are

$$V(\bar{y}_{lr}) = \lambda\left[s_y^2 - 2\beta_1 s_{xy} + \beta_1^2 s_x^2\right] \qquad (8)$$

and $V(\bar{y}_R) = \lambda\left[s_y^2 + R^2 s_x^2 - 2R s_{xy}\right]$ (9)

where $R = \dfrac{\bar{y}}{\bar{x}}$

The estimate of population total (y_{gl}) and its variance using proposed estimator $\bar{y}_{gl}$, are as follows

$$y_{gl} = N\,\bar{y}_{gl}$$

$$V(y_{gl}) = N^2 V(\bar{y}_{gl})$$

4. Comparison with ordinary regression estimator

Now we shall compare the variances of ordinary regression estimator and proposed regression estimator and it is found that proposed regression estimator is more efficient as compare to ordinary regression estimator. Using relations (5) and (8), we have

$$V(\bar{y}_{lr}) - V(\bar{y}_{gl}) = \lambda\left[2\hat{\beta}_2 s_{yz} - \hat{\beta}_2^2 s_z^2 - 2\hat{\beta}_1\hat{\beta}_2 s_{zx}\right]$$

Now

$$2\hat{\beta}_2 s_{yz} - \hat{\beta}_2^2 s_z^2 - 2\hat{\beta}_1\hat{\beta}_2 s_{zx}$$

$$= \hat{\beta}_2(2s_{yz} - \hat{\beta}_2 s_z^2 - 2\hat{\beta}_1 s_{zx})$$

$$= \frac{(s_{yx}s_{zx} - s_{yz}s_x^2)}{(s_{xz}^2 - s_x^2 s_z^2)}\left[2s_{yz} - \frac{(s_{yx}s_{zx} - s_{yz}s_x^2)}{(s_{xz}^2 - s_x^2 s_z^2)}s_z^2 - 2\frac{(s_{yz}s_{zx} - s_{yx}s_z^2)}{(s_{xz}^2 - s_x^2 s_z^2)}s_{zx}\right]$$

$$= \frac{(s_{yx}s_{zx} - s_{yz}s_x^2)}{(s_{xz}^2 - s_x^2 s_z^2)^2}\left[2s_{yz}(s_{xz}^2 - s_x^2 s_z^2) - (s_{yx}s_{zx} - s_{yz}s_x^2)s_z^2 - 2(s_{yz}s_{zx} - s_{yx}s_z^2)s_{zx}\right]$$

$$= \frac{(s_{yx}s_{zx} - s_{yz}s_x^2)}{(s_{xz}^2 - s_x^2 s_z^2)^2}\left[2s_{yz}s_{xz}^2 - 2s_{yz}s_x^2 s_z^2 - s_{yx}s_{zx}s_z^2 + s_{yz}s_x^2 s_z^2 - 2s_{yz}s_{zx}s_{zx} + 2s_{yx}s_z^2 s_{zx}\right]$$

$$=\frac{(s_{yx}s_{zx}-s_{yz}s_x^2)}{(s_{xz}^2-s_x^2s_z^2)^2}\left[-s_{yz}s_x^2s_z^2+s_{yx}s_{zx}s_z^2\right]$$

$$=\frac{(s_{yx}s_{zx}-s_{yz}s_x^2)}{(s_{xz}^2-s_x^2s_z^2)^2}\left[s_{yx}s_{zx}-s_{yz}s_x^2\right]s_z^2$$

$$=\frac{(s_{yx}s_{zx}-s_{yz}s_x^2)^2s_z^2}{(s_{xz}^2-s_x^2s_z^2)^2}>0$$

$$\Rightarrow \quad V(\overline{y}_{lr})-V(\overline{y}_{gl})>0$$

Which shows that the estimator $\overline{y}_{gl}$ is more efficient as compared to estimator $\overline{y}_{lr}$, as it has lesser mean squared error. We shall demonstrate this result with the help of following example in which Z has been considered as a function of X .

5. Numerical example

Example: The following table describes the estimates population mean and population total and their variances after ignoring f.p.c. using different models, the data given in Des Raj (1972), page 89. The size of the population has been considered as 30 and a random sample without replacement of size 8 has been drawn from it. The observations corresponding to sample numbers 12, 02,22,21,03,08,10,07 gave following results.

Estimates **Methods**	**Mean ($\overline{y}$)**	$V(\overline{y})$	**Total (y)**	$V(y)$
Model ($z=1/x^2$)	**4.0308**	**0.1302**	**120.9265**	117.1428
Model ($z=1/x$)	**4.3036**	**0.1378**	**129.1102**	124.0976
Model ($z=\sqrt{x}$)	**4.0598**	**0.1551**	**121.7963**	139.6262
Model ($z=x^2$)	**4.3451**	**0.1663**	**130.3527**	149.7388
Linear Regression	**4.2133**	**0.1836**	**126.4014**	165.3044
Ratio Estimator	3.1375	0.9636	94.1250	867. 2074

It is observed that estimates of population mean and population total obtained from $\overline{y}_{gl}$ are more efficient as compared to estimates of population mean and population total obtained from $\overline{y}_{lr}$ and also ratio estimator.

6. Conclusions

The inclusion of a linear term in ordinary linear regression estimator improves the precision of the estimates of population parameters such as population mean and population total etc. The proposed model is in general form, which includes the models of Ekpenyong *et al.* (2008) and also Misra *et al.* (2009) as particular cases. It has been shown with the help of an example that the proposed model provides more precise estimates of population mean and population total as compared to linear regression estimator as well as ratio estimator of population mean and population total. The inclusion of additional linear term has considered four cases in which z, a function of x has been considered as $\frac{1}{x^2}, \frac{1}{x}, \sqrt{x}$ and x^2. The second and fourth values of z correspond to Misra *et al.* (2009) and Ekpenyong *et al.* (2008) respectively. In all these four cases of z, the estimates of population mean and total are more efficient as compared to traditional linear regression estimator and traditional ratio estimator.

Acknowledgement

Authors are very thankful to the referees for their valuable suggestions in improving this manuscript and the financial assistance has been given by University Grant Commission, New Delhi, India under Major Research Project F.No.33-56/2007 (SR).

REFERENCES

COCHRAN, W.G. (1999). Sampling Techniques, John Wiley & Sons.

DES RAJ (1972). The design of sampling surveys, McGraw-Hill, New York.

EKPENYONG, E. J., OKONNAH, M.I. and JOHN, E.D. (2008), Polynomial (Non-linear) Regression.

Method for Improved Estimation Based on Sampling, Journal of Applied Sciences, 8(8), 1597-1599.

MISRA, G.C., SHUKLA, A.K. AND YADAV, S.K. (2009). A Comparison of Regression Methods for Improved Estimation in Sampling, Journal of reliability and statistical studies, Vol. 2, Issue2, 85-90.

SUKHATME, P.V., SUKHATME, B.V., SUKHATME, S. and ASOK, C. (1984). Sampling Theory of Surveys with Applications, Indian Society of Agricultural Statistics. Sampath, S. (2005). Sampling Theory and Methods, Narosa Publishing House, India.