Journal of Reliability and Statistical Studies (ISSN: 0974-8024)
Vol. 2, Issue 2(2009): 85-90

A COMPARISON OF REGRESSION METHODS FOR IMPROVED ESTIMATION IN SAMPLING

G.C.Misra[1], A.K.Shukla and S.K.Yadav
Department of Statistics, D. A. V. College, Kanpur, India
E Mail: 1. drmisragovind@gmail.com, govindmisra@indiatimes.com

Abstract

A linear model with an inverse term is proposed for estimation of population mean and population total in regression analysis. A comparison has been made in precisions of estimates of parameters, considering ordinary linear regression estimate, regression estimates using second degree polynomial for relationship between dependent variable (y)and auxiliary variable (x) and also with estimates obtained in regression method of estimation incorporating inverse term model for describing the relationship between y and x. The gain in efficiency is also demonstrated with the help of a data set in which comparison of various estimates of regression method of estimation has also been made with simple random sampling.

Key Words: Inverse term model, linear regression estimator, Polynomial regression, simple random sampling.

1. Introduction

The use of auxiliary information in estimation of population parameters in complex sample surveys often improves the efficiency of the estimators. There are two methods namely ratio and regression methods of estimation utilizing auxiliary information. If X is an auxiliary variable, closely related with variable under study Y and the line of regression Y on X passes through origin, the ratio method of estimation is used to estimate the population parameters. However, in most of the sample surveys it has been observed that the line of regression does not pass through origin, in such a situation the regression method of estimation is used.

We suppose that y_i and x_i are each obtained for every unit in the sample and that the population mean $\bar{X}$ of the x_i is known, the linear regression estimate of $\bar{Y}$ (the population mean of y_i) is

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

Where $\bar{y}$ and $\bar{x}$ are sample mean of the variable Y and X . The subscript lr denotes linear regression and b is an estimate of the change in Y when X is increased by unity. For detailed discussions on methods incorporating auxiliary information, references can be made of Cochran (1999) and Okafor (2002).

Matloff (1981) compared the estimate of unconditional mean of variable Y with linear estimate of variable Y and showed that linear estimate substantially improve the estimate over ordinary estimate. Jewell and Queensberry (1986) used an iterative regression method in stratified sample and demonstrated that linear estimator gives a general superior estimate of the mean in terms of efficiency.

Ekpenyong et.al (2008) considered a second order polynomial relationship of study variable (Y) with auxiliary variable (X) as

$$Y = a + bX + cX^2 + E \tag{2}$$

They have shown that it improves the efficiency of the estimates of characteristics of variables that have second order polynomial relationship with their auxiliary variable.

2. Proposed Model

We propose a linear regression model with an inverse term as

$$Y = \beta_0 + \beta_1 X + \frac{\beta_2}{X} + U, \qquad X > 0 \tag{3}$$

Where β_0, β_1 and β_2 are parameters which appear linearly in regression model (3). They are to be estimated from sample observations on X and Y. U is independently and identically distributed random variable with mean zero and fixed variance σ^2.

This model has been used earlier by Misra (1992) in some other context. For estimating population parameters of (3), several statistical techniques have been discussed in Wu (1981) and Draper and Smith (1998).

3. Estimation of Parameters and their variances

The Proposed model (3) can be written as

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + U \text{ , Where } Z = 1/X . \tag{4}$$

A regression estimator of the population mean based on (4) is given by

$$\bar{y}_{Inv} = \bar{y} - \hat{\beta}_1(\bar{x} - \bar{X}) - \hat{\beta}_2(\bar{z} - \bar{Z}) \tag{5}$$

Where $\bar{z}$ and $\bar{Z}$ are the sample and population means of variable z respectively. $\bar{y}_{Inv}$ represents regression estimate based on inverse term model. Using discussions of Cochran (1999) and Ekpenyong et.al (2008), the variance of $\bar{y}_{Inv}$ is given by

$$V(\bar{y}_{Inv}) = E\left[\bar{y} - \hat{\beta}_1(\bar{x} - \bar{X}) - \hat{\beta}_2(\bar{z} - \bar{Z})\right]^2 - \left\{E\left[\bar{y} - \hat{\beta}_1(\bar{x} - \bar{X}) - \hat{\beta}_2(\bar{z} - \bar{Z})\right]\right\}^2 \tag{6}$$

Writing the first term on right hand side of (6) as

$$E\left[\bar{y} - \hat{\beta}_1(\bar{x} - \bar{X}) - \hat{\beta}_2(\bar{z} - \bar{Z})\right]^2 = E\left\{\left[\bar{y} - \hat{\beta}_1(\bar{x} - \bar{X}) - \hat{\beta}_2(\bar{z} - \bar{Z})\right]\left[\bar{y} - \hat{\beta}_1(\bar{x} - \bar{X}) - \hat{\beta}_2(\bar{z} - \bar{Z})\right]\right\}$$

$$= E\Big[\bar{y}^2 - \hat{\beta}_1\bar{y}(\bar{x} - \bar{X}) - \hat{\beta}_2\bar{y}(\bar{z} - \bar{Z}) - \hat{\beta}_1\bar{y}(\bar{x} - \bar{X}) + \hat{\beta}_1^2(\bar{x} - \bar{X})^2 + \hat{\beta}_1\hat{\beta}_2(\bar{x} - \bar{X})(\bar{z} - \bar{Z})$$

$$- \hat{\beta}_2\bar{y}(\bar{z} - \bar{Z}) + \hat{\beta}_1\hat{\beta}_2(\bar{x} - \bar{X})(\bar{z} - \bar{Z}) + \hat{\beta}_2^2(\bar{z} - \bar{Z})^2\Big]$$

$$= E\Big[\bar{y}^2 - 2\hat{\beta}_1\bar{y}(\bar{x} - \bar{X}) - 2\hat{\beta}_2\bar{y}(\bar{z} - \bar{Z}) + 2\hat{\beta}_1\hat{\beta}_2(\bar{x} - \bar{X})(\bar{z} - \bar{Z}) + \hat{\beta}_1^2(\bar{x} - \bar{X})^2 + \hat{\beta}_2^2(\bar{z} - \bar{Z})^2\Big] \tag{7}$$

and the second term on right hand side of (6) can be written as

$$\left\{E[\bar{y}-\hat{\beta}_1(\bar{x}-\bar{X})-\hat{\beta}_2(\bar{z}-\bar{Z})]\right\}^2$$
$$=\left[E(\bar{y})-\hat{\beta}_1E(\bar{x}-\bar{X})-\hat{\beta}_2E(\bar{z}-\bar{Z})\right]^2$$
$$=[E(\bar{y})]^2-\hat{\beta}_1E(\bar{y})(\bar{x}-\bar{X})-\hat{\beta}_2E(\bar{y})(\bar{z}-\bar{Z})-\hat{\beta}_1E(\bar{y})(\bar{x}-\bar{X})+\hat{\beta}_1\hat{\beta}_2E(\bar{x}-\bar{X})E(\bar{z}-\bar{Z})$$
$$+\hat{\beta}_1^2\left\{E(\bar{x}-\bar{X})\right\}^2-\hat{\beta}_2E(\bar{z}-\bar{Z})E(\bar{y})+\hat{\beta}_1\hat{\beta}_2E(\bar{x}-\bar{X})E(\bar{z}-\bar{Z})+\hat{\beta}_2^2\left\{E(\bar{z}-\bar{Z})\right\}^2 \quad (8)$$

Therefore, $V(\bar{y}_{Inv})=(7)-(8)$ gives

$$V(\bar{y}_{Inv})=\left[E(\bar{y}^2)-\left\{E(\bar{y})\right\}^2\right]-2\hat{\beta}_1E(\bar{x}-\bar{X})(\bar{y}-\bar{Y})-2\hat{\beta}_2E(\bar{y}-\bar{Y})(\bar{z}-\bar{Z})$$
$$+2\hat{\beta}_1\hat{\beta}_2E(\bar{x}-\bar{X})(\bar{z}-\bar{Z})+\hat{\beta}_1^2\left[E(\bar{x}^2)-\left\{E(\bar{x})\right\}^2\right]+\hat{\beta}_2^2\left[E(\bar{z}^2)-\left\{E(\bar{z})\right\}^2\right]$$
$$V(\bar{y}_{Inv})=V(\bar{y})-2\hat{\beta}_1\operatorname{cov}(\bar{y},\bar{x})-2\hat{\beta}_2\operatorname{cov}(\bar{y},\bar{z})+2\hat{\beta}_1\hat{\beta}_2\operatorname{cov}(\bar{x},\bar{z})+\hat{\beta}_1^2V(\bar{x})+\hat{\beta}_2^2V(\bar{z}) \quad (9)$$

$$So\quad V(\bar{y}_{Inv})=\frac{1-f}{n}[s_y^2-2\hat{\beta}_1s_{xy}-2\hat{\beta}_2s_{yz}+2\hat{\beta}_1\hat{\beta}_2s_{zx}+\hat{\beta}_1^2s_x^2+\hat{\beta}_2^2s_z^2] \quad (10)$$

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, s_z^2 are unbiased estimators of the population variances S_X^2, S_Y^2, S_Z^2 respectively.

We need to estimate β_1 and β_2 such that $V(\bar{y}_{Inv})$ is a minimum. By the method of ordinary least square, we differentiate partially (9) with respect to $\hat{\beta}_1$ and $\hat{\beta}_2$ to obtain the following normal equations.

$$\hat{\beta}_2\operatorname{cov}(\bar{x},\bar{z})+\hat{\beta}_1V(\bar{x})=\operatorname{cov}(\bar{y},\bar{x}) \quad (11)$$

$$\hat{\beta}_2V(\bar{x})+\hat{\beta}_1\operatorname{cov}(\bar{x},\bar{z})=\operatorname{cov}(\bar{y},\bar{z}) \quad (12)$$

Solving (11) and (12) simultaneously, we obtain

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

and

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

or

$$\hat{\beta}_1=\frac{s_{yz}s_{xz}-s_{yx}s_z^2}{s_{xz}^2-s_x^2s_z^2}\quad \&\quad \hat{\beta}_2=\frac{s_{yx}s_{xz}-s_{yz}s_x^2}{s_{xz}^2-s_x^2s_z^2} \quad (13)$$

The estimate of population total and its variance are given as

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

With the help of an example, the efficiency and the precision of the estimates of above method will be compared with the estimates obtained from linear regression, second order polynomial and simple random sampling methods. Since the model (3) has least variance hence it is more efficient and precise than other estimates.

Example: [Data Source: Desraj (1972), page 89]

The population of size is 30 and the sample size is 8. The observations corresponding to sample numbers 12, 02, 22, 21, 03, 08, 10, 07 gave following results.

$s_{xy} = 3.82321,\ s_{yz} = -0.40706,\ s_{zx} = -0.40901,\ s_x^2 = 3.55357,$
$s_y^2 = 4.29696,\ s_z^2 = 0.070861\ \hat{\beta}_1 = 1.23553,\ \hat{\beta}_2 = 1.38714.$

With the help of these results the following table is prepared

Estimates / Methods	Inverse term model	Second degree polynomial	Linear regression	simple random sampling without replacement
Mean($\bar{y}$)	4.30367	4.34509	4.21338	3.13750
$V(\bar{y})$	0.137886	0.166376	0.183672	4.29696
Total(y)	129.1102	130.3527	126.4014	94.1250
$V(y)$	124.0976	149.7388	165.3044	3867.2640

Table: Summary of estimates of means, totals and their variances

From above table the efficiencies of estimate of population mean using relation (3) over population estimate using model (1), (2) and simple random sampling without replacement is given below:

E_1 [The efficiency of regression estimate using relation (3) over regression estimate using relation (2)]

$$E_1 = 120.66\%$$

E_2 [The efficiency of regression estimate using relation (3) over regression estimate using relation (1)]

$$E_2 = 133.21\%$$

E_3 [The efficiency of regression estimate using relation (3) over regression estimate using simple random sampling]

$$E_3 = 3116.31\% .$$

4. Result and Discussion

From above results, it is observed that use of the inverse term model in regression method of estimation gives estimates of population mean and population total which have least variances as compared to other estimates (ordinary regression estimate, second degree polynomial regression estimate and simple random sampling estimate). Hence, regression estimate using inverse term model provides the most efficient estimates of the population parameters.

It has also been observed that the precision of estimates of parameters in regression method of estimation depends on nature or pattern of relationship between auxiliary variable and variable under study (dependent variable). The relationship using inverse term model improve the precision of estimation of estimates as compared to ordinary linear regression estimates, regression estimates using polynomial relationship.

Ekpenyong et.al (2008) have suggested use of a polynomial of degree two for describing relationship between x and y and called it non-linear relation. In fact the relationship between x and y considered by Ekpenyong et.al (2008) is linear in its parameters and is not non-linear. The method of classical least squares is directly applicable to estimates the parameters of the model (2). It is only in nonlinear models that least squares technique can not be directly applied to estimate its parameters and estimation procedure is carried out by iterative methods and estimates of parameters possess asymptotic properties. For detailed discussion on estimation procedures in linear and nonlinear models, reference can be made of Draper & Smith (1998), Saber & Wild (1989) and Bates & Watts (1988).

5. Conclusion

The use of inverse term model in regression estimation method provides more efficient and precise estimates of characteristic under study as compared to polynomial regression method of Ekpenyong et al (2008) and traditional linear regression method. It also reflects the effect of nature of relationship between characteristic under study and auxiliary variable in regression method of estimation in sampling theory.

Acknowledgement

This work has been funded by U.G.C. under Major research project F.No. 33-56/2007(SR)

References

1. Bates, D.M. and Watts, D.G. (1988). Non-linear Regression Analysis and Its Applications. John Wiley, New York.
2. Cochran, W.G. (1999). Sampling Techniques. John Wiley & Sons.
3. Des Raj. (1972). The Design of Sample Surveys. McGraw-Hill, New York.
4. Draper, N.R. and Smith, H. (1998). Applied Regression Analysis. John Wiley, New York.

5. Ekpenyong, E.J., Okonnah, M.I., John, E.D. (2008). Polynomial (Non Linear) Regression Method for Improved Estimation Based on Sampling. Journal of Applied Sciences, 8(8), p. 1597-1599.
6. Jewell, N.P. and Queensberry, C.P. (1986). Regression Analysis Based on Stratified Sample. Biometrika, 73 (3), p. 605-661.
7. Matloff, N.S.(1981). Use of Regression Functions for Improved Estimation of Means. Biometrika, 68 (3), p. 685-689.
8. Misra, G.C.(1992). "Some Estimation Procedures in Non linear Statistical Models" Unpublished Thesis for Ph.D. , C.S.J.M.University , Kanpur.
9. Okafor, F.C.(2002). Sample Survey theory with Applications. Afro-Orbis Publisher, Nsukka, Nigeria.
10. Saber, G.A.F and Wild, C.J.(1989). Non-Linear Regression. Wiley, New York.
11. Wu, C.F.(1981). Asymptotic Theory of Non-Linear Least Squares Estimation. Ann. Stat., 9, p. 501-513.