The Effect of Sample Size on Parameter Accuracy Using Ratio and Regression

Adil Altag Zaidan

Department of Mathematics, Faculty of Science, Shagra University, kingdom of Saudi Arabia
adil_zidan@hotmail.com

Abstract: The objectives of this study are to highlight the role of auxiliary variable and their importance to methods of parameter estimation and to stress to the part played by the three estimation methods i.e (ratio estimation between two variables, linear regression and mean) as well as to identify the sample size appropriate to each of the three estimation methods. The present study has generated artificial data using computer simulation, 21 sample of varying size were selected. In this study we have under taken a practical inquiry to verify whether ratio or regression estimation have more accuracy. The results revealed that the regression and mean per unit estimations are better than ratio estimation. In addition, regression estimation is better than mean estimation per unit in small samples.
[Adil Alatg zaidan. **The Effect of Sample Size on Parameter Accuracy Using Ratio and Regression.** *Nat Sci* 2014;12(7):73-80]. (ISSN: 1545-0740). http://www.sciencepub.net/nature. 12

Keywords: Estimation, bias, ratio, regression, auxiliary.

1. Introduction

Both ratio and regression estimations are intended to enhance accuracy using. auxiliary variables that are closely related to the study variables in order to reduce mean square error and the total population. Ratio estimation is regarded as the best when there is a linear relationship between variables (Y_i, X_i) passing the point of origin. However, if the linear relationship does not pass the point of origin, regression estimation will score greater accuracy. Descriptively, when precision is estimated using a *standard error*, it is thought of as the amount of fluctuation from the population parameter that we can expect by chance alone in sample estimates.(1) This study will specifically deal with ratio estimation between two variables as well regression method through study auxiliary variables are their related theorems aiding us to appreciate the impact of sample size on accuracy of estimation. Following this, a comparison will be made between the two methods to determine the appropriate sample size to make accurate estimations. In many sample surveys, the information on single (or more) auxiliary variable(s) correlated with the study variable is used for increasing the precision of estimators. A number of sampling strategies utilize the advance information about an auxiliary variable. When such information is lacking, it is sometimes relatively cheap to take a large preliminary sample in which auxiliary variable alone is measured. (2). The use of auxiliary information can increase the precision of an estimator when study variable Y is highly correlated with auxiliary variable X.(5) The problem of study stems from the impact of sample size and auxiliary variable on some sampling methods. These variables are seen as the best predictors of parameters to be evaluated through ratio and linear regression estimations once certain conditions are met. Since we to assess the value parameters, the best sampling method should be employed to obtain the most accurate parameters. Ratio and linear regression are viewed as among the most important methods to increase accuracy when using auxiliary variables.

The objectives of this study are to highlight the role of auxiliary variable and their importance to methods of parameter estimation and to stress to the part played by the three estimation methods i.e (ratio estimation between two variables, linear regression and mean) as well as to identify the sample size appropriate to each of the three estimation methods. It also aims to compare the methods of variable and regression estimation according to sample size and to uncover the cases when ratio and regression estimation have an equal degree of accuracy.

This study hypothesizes that the use of auxiliary increases the degree of accuracy in cases of small sample in ratio estimations. However, in large samples. However, in cases of large samples regression estimation will prove more accurate than former method. Also, there are differences in terms of accuracy between the two method, and the decrease in sample size, but these variable tend to dis appear if sample size increases. In additional, there are differences in the mean accuracy in cases of large samples between ration and regression estimations. Equally, there are differences between total number estimation in small sample in terms of accuracy when using ration and regression estimation techniques. The use of auxiliary information can increase the precision of an estimator when study variable y is highly correlated with auxiliary variable x. There exist situations when information is available in the form of attribute φ, which is highly correlated with y.

For example:

a) Sex and height of the persons,
b) Amount of milk produced and a particular breed of the cow,
c) Amount of yield of wheat crop and a particular variety of wheat etc.(4)

This study answer question as How do samples of different sizes perform in terms of giving an accurate representation of the underlying population? Does it depend on how large the sample is relative to the size of the population or does it depend on the absolute size of the sample? (3)

2. Ratio estimation

It is a method to evaluate population parameters through well-known sampling techniques (sample result) with the objective of enhancing accuracy of evaluation. Many studies provide measurement on the population items under study and these might be directly and strongly linked to the phenomenon whose population parameters evaluated. There for, use can be made of the existing measurements when assessing the population parameters termed auxiliary variables. To explain how the ratio technique works, we will suppose we have a population composed of N (unknown) items on the phenomenon of Y and that its items are:

$$Y_1, Y_2, Y_3, \dots Y_N$$

We will suppose that we possess previous information another phenomenon on X in this population that is related to the phenomenon under study containing the following items:

$$X_1, X_2, X_3, \dots X_N$$

We wish to test a simple random sample composed of n items in order to assess the population parameters, for the phenomenon (Y) based on the existing knowledge on phenomenon (X).

To evaluate ratio R in the population, we will choose a simple random sample with a size of n items and that the values of phenomenon (Y) for the values samples are $y_1, y_2, y_3, \dots, y_n$ moreover, the values of phenomenon (X) for the same sample items are $x_1, x_2, x_3, \dots, x_n$ while $\hat{R}$ is the ratio estimation R calculated out of the sample as follows:

$$\hat{R} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \frac{\bar{y}}{\bar{x}} \qquad \rightarrow (1)$$

The mean population estimated μ_y will be:

$$\hat{\mu}_y = \hat{R}\mu_x \qquad \rightarrow (2)$$

While the total population values Y shall be:

$$\hat{Y} = \hat{R}X \qquad \rightarrow (3)$$

As evident from the simple random sample, the $\bar{x}$ mean of the sample is the best estimate for the population mean μ_y. However, when we have x_i versions accompanying and related y_i version, there will be a more accurate estimation for the parameter μ_y as the evaluation of ratio $\hat{\mu}_y$. the purpose is to obtain an increase in the estimation accuracy taking advantage of the correlation between x_i, y_i.

If the ratio $\frac{y_i}{x_i}$ approximately equal for all units of the population, the value $\frac{\sum y_i}{\sum x_i}$ will slightly differ from one sample to another and the ratio estimation will be of great accuracy.

In random sample of the size n (large n):

$$V(\hat{Y}_R) = \frac{N^2(1-f)}{n}\frac{\sum_{i=1}^{N}(y_i - Rx_i)^2}{N-1} \qquad \rightarrow (4)$$

$$V(\hat{\bar{Y}}_R) = \frac{(1-f)}{n}\frac{\sum_{i=1}^{N}(y_i - Rx_i)^2}{N-1} \qquad \rightarrow (5)$$

$$V(\hat{R}) = \frac{(1-f)}{n\mu_x^2}\frac{\sum_{i=1}^{N}(y_i - Rx_i)^2}{N-1} \qquad \rightarrow (6)$$

Where $f = n/N$ is the sampling fraction.

3. Bias in ratio estimate:

The degree of ratio bias $\widehat{\square}$ is measured through simple random sampling $E(\hat{R} - R)$ Out of the population ratio R i,e the degree of is $E(\hat{R} - R) = E(\hat{R}) - R = \delta$ we will identify the degree of the bias δ when the sample size is large and the difference between $(\bar{x} - \mu_x)$ is small. We note from the definition:

$$\hat{R} = \frac{y}{x} = \frac{\bar{y}}{\bar{x}} = \frac{\bar{y} + \mu_y - \mu_y}{\bar{x} + \mu_x - \mu_x} = \frac{\mu_y\left(1 + \frac{\bar{y} - \mu_y}{\mu_y}\right)}{\mu_x\left(1 + \frac{\bar{x} - \mu_x}{\mu_x}\right)}$$

i.e:

$$\hat{R} = \frac{\mu_y}{\mu_x}\left(1 + \frac{\bar{y} - \mu_y}{\mu_y}\right)\left(1 + \frac{\bar{x} - \mu_x}{\mu_x}\right)^{-1} \qquad \rightarrow (7)$$

Using Taylor series $\left(1 + \frac{\bar{x} - \mu_x}{\mu_x}\right)^{-1}$, we obtain:

$$\hat{R} = \frac{\mu_y}{\mu_x}\left(1 + \frac{\bar{y} - \mu_y}{\mu_y}\right)\left(1 - \frac{(\bar{x} - \mu_x)}{\mu_x} + \frac{(\bar{x} - \mu_x)^2}{{\mu_x}^2} - \cdots\right) \qquad \rightarrow (8)$$

The Above relation can be expressed as follows:

$$\hat{R} = \frac{\mu_y}{\mu_x}\left(1 + \frac{(\bar{y} - \mu_y)}{\bar{Y}} - \frac{(\bar{x} - \mu_x)}{\mu_x} + \frac{(\bar{x} - \mu_x)^2}{{\mu_x}^2} - \frac{(\bar{y} - \mu_y)(\bar{x} - \mu_x)}{\mu_y\mu_x}\cdots\right) \rightarrow (9)$$

Using approximation according to smallness of value $(\bar{x} - \mu_x)$, we can distinguish three condition as below:

1. When $\bar{x} \cong \bar{X}$ i.e $\bar{x}$ is very close to $\bar{X}$, the equation (9) will be:

$$\hat{R} = R\left(1 + \frac{(\bar{y} - \mu_y)}{\mu_y} - 0 + 0 - \cdots\right) \qquad \rightarrow (10)$$

Considering the expectation for the tow side, we obtain:

$$E(\hat{R}) = R + RE\frac{(\bar{y}-\mu_y)}{\mu_y} = R$$

Where $E(\bar{y}-\mu_y) = 0$

$$\therefore E(\hat{R}) = R \qquad \rightarrow (11)$$

i.e $\hat{R}$ an unbiased estimation of R ratio.

2. When the value $\frac{(\bar{x}-\mu_x)^2}{{\mu_x}^2}$ is so minimal that the boundaries following it can be neglected when it approximation zero and the equation (9) will be:

$$\hat{R} = \frac{\mu_y}{\mu_x}\left(1 + \frac{(\bar{y}-\mu_y)}{\bar{Y}} - \frac{(\bar{x}-\mu_x)}{\mu_x} + 0 - 0\cdots\right) \qquad \rightarrow (12)$$

Considering the expectation of the tow side, we obtain:

$$E(\hat{R}) = R + RE\frac{(\bar{y}-\mu_y)}{\mu_y} - RE\frac{(\bar{x}-\mu_x)}{\mu_x} = R \qquad \rightarrow (13)$$

i.e $\hat{R}$ an unbiased estimation of R ratio.

3. When the value $(\bar{x}-\mu_x)^2$ is minimal so that the boundaries following the value $\frac{(\bar{y}-\mu_y)(\bar{x}-\mu_x)}{\mu_y\mu_x}$ can be neglected due to extreme smallness and so the relationship (9) can expressed as:

$$\hat{R} = R\left(1 + \frac{(\bar{y}-\mu_y)}{\mu_y} - \frac{(\bar{x}-\mu_x)}{\mu_x} + \frac{(\bar{x}-\mu_x)^2}{{\mu_x}^2} - \frac{(\bar{y}-\mu_y)(\bar{x}-\mu_x)}{\mu_y\mu_x}\cdots\right) \rightarrow (14)$$

Considering the expectation, we obtain:

$$E(\hat{R}) = R + 0 - 0 + \frac{RE(\bar{x}-\mu_x)^2}{{\mu_x}^2} - E(\bar{y}-\mu_y)(\bar{x}-\mu_x)\frac{R}{\mu_y\mu_x} \qquad \rightarrow (15)$$

Where the variance for the variable $\bar{x}$ is:

$$E(\bar{x}-\mu_x)^2 = (1-f).\frac{S_x^2}{n}$$

And the covariance for the tow variables $\bar{x}, \bar{y}$ is:

$$E(\bar{y}-\mu_y)(\bar{x}-\mu_x) = (1-f)\rho\frac{S_xS_y}{n}$$

By substituting the variance and covariance in the equation (9), we obtain:

$$\delta = E(\hat{R}) - R = (1-f).\frac{S_x^2}{n}.\frac{1}{{\mu_x}^2}R - (1-f)\rho\frac{S_xS_y}{n}.\frac{1}{\mu_y\mu_x}.\frac{\mu_y}{\mu_x}$$

Following this, the bias δ is:

$$\delta = (1-f).\frac{1}{{\mu_x}^2}\{RS_x^2 - \rho S_yS_x\} \qquad \rightarrow (16)$$

The bias δ is zero if:

$$RS_x^2 - \rho S_yS_x = 0$$

That is:

$$E(\hat{R}) - R = 0$$

i.e $\hat{R}$ an unbiased estimation of R ratio.

Result (1):

The bias δ is zero if:

$$RS_x^2 - \rho S_yS_x = 0$$

That is:

$$RS_x^2 = \rho S_yS_x$$

Or

$$RS_x = \rho S_y \qquad \rightarrow (17)$$

The equation (17) is realized when the regression Y on X passes the original point. To prove this, we suppose the regression Y on X passing the original point is:

$$Y_i = \beta X_i \qquad \rightarrow (18)$$

It follows:

$$\beta = \frac{\mu_y}{\mu_x} = R \qquad \rightarrow (19)$$

That is the ratio R equals the regression coefficient β, i.e $\beta = R$, we substitute the R value in the relationship (18) and we obtain:

$$RS_x = \beta S_x = \frac{\sum(Y-\mu_y)(X-\mu_x)}{\sum(X-\mu_x)^2}.S_x = \frac{cov(Y,X)}{S_yS_x}.S_y = \rho S_y$$

Which equals the left side of relation (17) where:

$$\beta = \frac{\sum(Y-\mu_y)(X-\mu_x)}{\sum(X-\mu_x)^2} = \frac{cov(Y,X)}{S_x^2};\ \rho = \frac{cov(Y,X)}{S_yS_x}$$

Result (2):

The bias $\delta = E(\hat{R}-R)$ explained in the relationship (17) can be rewritten using the difference coefficient C_y, C_x in the formula:

$$E(\hat{R}-R) = (1-f).\frac{1}{{\mu_x}^2}\{RS_x^2 - \rho S_yS_x\} = (1-f).R\left\{\frac{S_x^2}{{\mu_x}^2} - \rho\frac{S_yS_x}{R{\mu_x}^2}\right\} = (1-f).R\left\{\frac{S_x^2}{{\mu_x}^2} - \rho\frac{\mu_x}{\mu_y}\frac{S_yS_x}{{\mu_x}^2}\right\} = (1-f).R\{C_x^2 - \rho C_yC_x\}$$

The bias will be zero when the regression Y on X pass the original point.

4. Variance of Ratio Estimation:

The variance of ratio estimation between two variables resulting from simple random sampling can be expressed in the relationship:

$$V(\hat{R}) = E(\hat{R}-R)^2 \qquad \rightarrow (20)$$

Theorem (1):

When $\bar{x} \cong \mu_x$, $\hat{R}$ variance can be given in the relationship:

$$V(\hat{R}) = \frac{1}{\mu_x}.\frac{(1-f)}{n}\frac{\sum_{i=1}^{N}(y_i - Rx_i)^2}{N-1} \quad \rightarrow (21)$$

Proof:

We can express the value $(\hat{R} - R)$ by the tow values $y_i - Rx_i$, μ_x using the method for calculation of the bias of ratio $\hat{R}$:

$$\hat{R} = R\left(1 + \frac{\bar{y} - \mu_y}{\mu_y}\right)$$

Here

$$\hat{R} = R + \frac{\bar{y} - \mu_y}{\mu_y}.\frac{\mu_y}{\mu_x} = R + \frac{\bar{y} - \mu_y}{\mu_x} \quad \rightarrow (22)$$

Where $\mu_y = R\mu_x$

From the relation (22), we obtain:

$$\hat{R} - R = \frac{\bar{y} - R\mu_x}{\mu_x} \quad \rightarrow (23)$$

Also, where $\bar{x} \cong \mu_x$, the relation (23) can be rewritten in the form:

$$\hat{R} - R = \frac{\bar{y} - R\bar{x}}{\mu_x} \quad \rightarrow (24)$$

Where y_i, x_i are the means of the two samples from the two groups and from the relation (24) we find:

$$\hat{R} - R = \frac{1}{\mu_x}\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - Rx_i)\right]$$

$$= \frac{1}{\mu_x}\left(\frac{1}{n}\sum_{i=1}^{n}(z_i)\right) \quad \rightarrow (25)$$

Where $z_i = y_i - Rx_i$

Here:

$$\bar{z} = \frac{1}{n}\sum_{i=1}^{n}(y_i - Rx_i) = \frac{1}{n}\sum_{i=1}^{n} z_i \quad \rightarrow (26)$$

By substituting the relation (26) in (25), we obtain:

$$\hat{R} - R = \frac{1}{\mu_x}.\bar{z} \quad \rightarrow (27)$$

Where $\bar{x}$ can be considered as the mean deviation of the value of the sample size n. Consequently, the mean deviation of the population size N can be calculated as follows:

$$\bar{Z} = \frac{1}{N}\sum_{i=1}^{N} z_i = \frac{1}{N}\sum_{i=1}^{N}(y_i - Rx_i)$$

Where the variance $V(\hat{R})$ is as follows:

$$V(\hat{R}) = E(\hat{R} - R)^2 = \frac{1}{{\mu_x}^2}E(\bar{z} - \bar{Z})^2 \quad \rightarrow (28)$$

However, the value $E(\bar{z} - \bar{Z})^2$ is a variance of the mean deviation $\bar{z}$ for a simple random sample n from population N and which can be expressed as follows:

$$V(\hat{R}) = \frac{1}{{\mu_x}^2}(1-f)\frac{S_z^2}{n}$$

$$= \frac{1}{{\mu_x}^2}\left[(1-f)\frac{\sum_{i=1}^{N}(z_i - \bar{Z})^2}{n(N-1)}\right]$$

Since $\bar{Z} = 0$:

$$V(\hat{R}) = \frac{1}{{\mu_x}^2}(1-f)\frac{S_z^2}{n} = \frac{1}{{\mu_x}^2}\frac{(1-f)}{n}\frac{\sum_{i=1}^{N} z_i^2}{(N-1)}$$

And since $z_i = y_i - Rx_i$:

$$V(\hat{R}) = \frac{1}{{\mu_x}^2}\frac{(1-f)}{n}\frac{\sum_{i=1}^{N}(y_i - Rx_i)^2}{(N-1)} \quad \rightarrow (29)$$

Result (3):

The relationship (28) can be rewritten using y_i, x_i variance and correlation coefficient as follows:

$$V(\hat{R}) = \frac{(1-f)}{n}\frac{1}{{\mu_x}^2}\left[S_y^2 + R^2S_x^2 - 2\rho RS_yS_x\right] \quad \rightarrow (30)$$

From (29) where $\bar{y} = R\mu_x$ we find:

$$\frac{1}{{\mu_x}^2}\frac{(1-f)}{n}\frac{\sum_{i=1}^{N}(y_i - Rx_i)^2}{(N-1)}$$

$$= \frac{1}{{\mu_x}^2}\frac{(1-f)}{n}\frac{1}{(N-1)}\sum_{i=1}^{N}[(y_i - \bar{Y}) - R(x_i - \mu_x)]^2$$

$$= \frac{1}{{\mu_x}^2}\frac{(1-f)}{n}\left[S_y^2 - 2Rcov(y_i, x_i) + S_x^2\right]$$

$$= \frac{1}{{\mu_x}^2}\frac{(1-f)}{n}\left[S_y^2 + R^2S_x^2 - 2\rho RS_yS_x\right]$$

Where $\rho = \frac{cov(y,x)}{S_yS_x}$ and the right side (30) is at minimum. When ρ is at maximum, that is $\rho = 1$. this means the pair value (y_i, x_i) are on straight line and hence $V(\hat{R})$ is at minimum value.

5. Regression Estimation:

Regression estimation resembles ratio estimation using the auxiliary variable x_i correlated to y_i. While the relationship is virtually linear the line does not pass the original point. The estimation is, there for, based on linear regression y_i on x_i instead of the ration between two variables.(6)

Estimation of linear regression has many uses. For instance, if we can easily obtain a value for an attribute for each unit, and if we can use other costly method to obtain the correct value X_i for the same attribute for a simple random sample, we can employ either of the estimation to reach the accurate mean estimation or the total value. Regression estimation is consistent but biased and can be overlooked in the large samples.

We suppose that we obtain y_i, x_i for each unit of the sample and that the mean population x_i is known μ_x , the linear regression μ_y for the mean population y_i is:

$$\bar{y}_{lr} = \bar{y} + b(\mu_x - \bar{x}) \quad \rightarrow (31)$$

Where:

$\bar{y}$ is the mean of the measurements of y_i in the random sample size (n).

$\bar{x}$ is the mean of the measurements x_i, and that $\mu_x$is the mean of population.

b is the regression coefficient.
$\bar{y}_{lr}$ is the mean estimation through linear regression.
To estimate the population Y, we shall take:

$$\hat{Y}_{lr} = N\bar{y}_{lr} \quad \rightarrow (32)$$

6.Regression Estimation when the regression coefficient b is known:

In many statistical studies, we might need to estimate regression coefficient. This is determined through studies or previous data. We might rely phenomenon. then, the regression coefficient approaches one and we select:
First: when $b = 1$, the regression estimation is:

$$\bar{y}_{lr} = \bar{y} + (\mu_x - \bar{x}) \quad \rightarrow (33)$$

The formula for the population can also he written as:

$$\bar{y}_{lr} = \mu_x + (\bar{y} - \bar{x}) \quad \rightarrow (34)$$

This provides another explanation for the regression estimation $(\bar{y}_{lr})$, since it equals the approximate value of the real mean with the addition of the numerical bias (that is the difference between the mean in the present estimations and previous studies).
Second: when $b = 0$ the estimation is:

$$\bar{y}_{lr} = \bar{y} \quad \rightarrow (35)$$

Third: when $b = \frac{\bar{y}}{\bar{x}}$ the estimation is:

$$\bar{y}_{lr} = \bar{y} + \frac{\bar{y}}{\bar{x}}(\mu_x - \bar{x}) = \frac{\bar{y}}{\bar{x}}\mu_x = \hat{\bar{Y}}_R \rightarrow (36)$$

Theorem (2):

In the simple random method where b_0 is predetermined constant, the linear regression estimation is:

$$\bar{y}_{lr} = \bar{y} + b_0(\mu_x - \bar{x}) \quad \rightarrow (37)$$

Is unbiased towards the mean, that is $E(\bar{y}_{lr}) = \mu_y$ and the variance is:

$$V(\bar{y}_{lr}) = \frac{1-f}{n}.\frac{\sum_{i=1}^{n}\left[(y_i - \mu_y) - (x_i - \mu_x)\right]^2}{N-1}$$
$$= \frac{1-f}{n}\left(S_y^2 - 2b_0S_{yx} + b_0^2S_x^2\right) \quad \rightarrow (38)$$

The sample variance is:

$$v(\bar{y}_{lr}) = \frac{1-f}{n}.\frac{\sum_{i=1}^{n}[(y_i - \bar{y}) - (x_i - \bar{x})]^2}{N-1}$$
$$= \frac{1-f}{n}\left(s_y^2 - 2b_0s_{yx} + b_0^2s_x^2\right) \quad \rightarrow (39)$$

What is the optimal value for b_0?
From theorem(2)

$$V(\bar{y}_{lr}) = \frac{1-f}{n}\left(S_y^2 - 2b_0S_{yx} + b_0^2S_x^2\right) \rightarrow (40)$$

By taking the first derivation for b_0, we obtain:

$$\frac{\partial V}{\partial b_0} = \frac{1-f}{n}\left(0 - 2S_{yx} + 2b_0S_x^2\right) = 0$$
$$b_0S_x^2 = S_{yx}$$
$$\therefore b_0 = \frac{S_{yx}}{S_x^2}$$

The variance of the minimal limit for $\bar{y}_{lr}$ shall be:

$$V_{min}(\bar{y}_{lr}) = \frac{1-f}{n}S_y^2(1-\rho^2) \rightarrow (41)$$

From theorem (2)

$$V(\bar{y}_{lr}) = \frac{1-f}{n}\left(S_y^2 - 2b_0S_{yx} + b_0^2S_x^2\right)$$

However, $b_0 = \frac{S_{yx}}{s_x^2}$

$$\therefore V(\bar{y}_{lr}) = \frac{1-f}{n}\left[S_y^2 - 2S_{yx}\left(\frac{S_{yx}}{S_x^2}\right) + S_x^2\left(\frac{S_{yx}}{S_x^2}\right)^2\right]$$
$$= \frac{1-f}{n}\left[S_y^2 - 2\left(\frac{S_{yx}}{S_x}\right)^2 + \left(\frac{S_{yx}}{S_x}\right)^2\right]$$
$$= \frac{1-f}{n}\left[S_y^2 - \left(\frac{S_{yx}}{S_x}\right)^2\right]$$
$$= \frac{1-f}{n}.S_y^2\left[1 - \left(\frac{S_{yx}}{S_x}\right)^2\right]$$
$$= \frac{1-f}{n}.S_y^2(1-\rho^2)$$

The relationship between the regression coefficient b_0 and the correlation coefficient ρ:
Since

$$b_0 = \frac{S_{yx}}{S_x^2}$$

By multiplying the right side by $\frac{S_y}{S_y}$, we find:

$$b_0 = \frac{S_{yx}}{S_x^2}.\frac{S_y}{S_y} = \frac{S_{yx}}{S_yS_x}.\frac{S_y}{S_x} = \rho.\frac{S_y}{S_x}$$
$$\therefore b_0 = \rho.\frac{S_y}{S_x}$$

Following this, since

$$\bar{y}_{lr} = \bar{y} + b_0(\mu_x - \bar{x})$$

We can this write:

$$\bar{y}_{lr} = \bar{y} + \rho.\frac{S_y}{S_x}(\mu_x - \bar{x})$$

7. A comparison of the Three Estimations:

The three variances for the estimation of the mean of population $\bar{y}$ to be compare are as follows:(6)

$$V(\bar{Y}_{lr}) = \frac{(1-f)}{n}S_y^2(1-\rho^2)$$
$$V(\bar{Y}_R) = \frac{(1-f)}{n}(S_y^2 + R^2S_x^2 - 2R\rho S_yS_x)$$
$$V(\bar{Y}) = \frac{(1-f)}{n}S_y^2$$

It is apparent that the regression estimation variance is less than the mean variance for each unit unless $\rho = 0$. Also, the ration estimation is more optimal than the mean estimation for each unit if there is a strong relationship between (x, y), that is $\rho > 0.5$.
The ratio estimation is less variance if:

$$S_y^2 + R^2S_x^2 - 2R\rho S_yS_x < S_y^2$$

This mean that if:

$$S_y^2 + R^2S_x^2 - 2R\rho S_yS_x < S_y^2$$
$$2R\rho S_yS_x > R^2S_x^2$$
$$2\rho S_y > RS_x$$

$$\rho > R\frac{S_x}{2S_y}$$

The regression estimation variance will less than the ratio estimation if:

$$S_y^2(1-\rho^2) < (S_y^2 + R^2S_x^2 - 2R\rho S_yS_x)$$
$$-S_y^2\rho^2 < R^2S_x^2 - 2R\rho S_yS_x$$
$$S_y^2\rho^2 - 2R\rho S_yS_x + R^2S_x^2 > 0$$
$$(\rho S_y - RS_x)^2 > 0$$
$$(\rho\frac{S_y}{S_x} - R)^2 > 0 \quad or \quad (b-R)^2 > 0$$

Hence, the regression estimation is more accurate than ratio estimation unless b=R. This occurs when (x, y) are have a relationship of a straight line passing through the original point.

8. Relative Efficiency:

$$Eff\left(\hat{\bar{Y}}, \hat{\bar{Y}}_R\right) = \frac{V(\hat{\bar{Y}})}{V(\hat{\bar{Y}}_R)} \;, Eff\left(\hat{\bar{Y}}, \hat{\bar{Y}}_{lr}\right)$$
$$= \frac{V(\hat{\bar{Y}})}{V(\hat{\bar{Y}}_{lr})} \;, Eff\left(\hat{\bar{Y}}_R, \hat{\bar{Y}}_{lr}\right)$$
$$= \frac{V\left(\hat{\bar{Y}}_R\right)}{V\left(\hat{\bar{Y}}_{lr}\right)}$$

9. Practical Aspects of Regression and Ratio Estimation:

The present study has generated data using computer simulation. 21 sample of varying size were selected. The size of auxiliary variable totaled 720 with the a mean of 69. The initial sample results are displayed in the following table (Table 1):

Table(1-a)

	s_y	s_x	s_y^2	s_x^2	b
$Samp_1$	3.104	3.237	9.632	10.478	0.152
$Samp_2$	2.833	2.404	8.024	5.78	0.174
$Samp_3$	2.685	3.735	7.209	13.953	0.07
$Samp_4$	3.076	2.444	9.463	5.975	0.164
$Samp_5$	3.095	2.782	9.52	7.742	0.125
$Samp_6$	2.672	2.425	8.867	7.139	0.197
$Samp_7$	3.306	2.499	10.931	6.243	0.098
$Samp_8$	3.122	3.079	9.744	9.483	0.289
$Samp_9$	3.059	3.184	9.361	10.135	0.136
$Samp_{10}$	2.745	2.572	7.537	6.614	0.158
$Samp_{11}$	2.978	2.794	8.867	7.809	0.052
$Samp_{12}$	2.779	2.987	7.724	8.922	0.059
$Samp_{13}$	2.943	3.286	8.662	10.798	0.015
$Samp_{14}$	2.834	2.951	8.034	8.71	0.039
$Samp_{15}$	2.948	2.844	8.689	8.087	0.017
$Samp_{16}$	2.869	2.947	8.23	8.684	0.035
$Samp_{17}$	3.123	2.939	9.755	8.64	0.088
$Samp_{18}$	3.077	3.019	9.466	9.117	0.031
$Samp_{19}$	2.915	3.078	8.499	9.477	0.01
$Samp_{20}$	3.098	3.016	9.597	9.098	0.022
$Samp_{21}$	2.97	3.034	8.822	9.205	0.031

Table(1-b)

$\hat{R}$	ρ	$\bar{y}$	$\bar{x}$	μ_x	n	N
1	0.158	67.64	67.66	71	9	720
0.988	0.148	68.9	69.56	71	12	720
0.983	0.097	68.75	69.98	71	15	720
0.979	0.13	67.65	69.09	71	18	720
0.987	0.113	67.99	68.9	71	20	720
0.986	0.179	69.52	70.47	71	24	720
0.995	0.074	69.12	69.46	71	27	720
1.02	0.285	69.28	67.9	71	30	720
0.995	0.142	68.73	69.06	71	50	720
1.006	0.148	69.68	69.25	71	100	720
0.999	0.049	68.87	68.96	71	140	720
0.994	0.063	69	69.44	71	200	720
0.999	0.017	69.03	69.07	71	270	720
1.002	0.041	69.25	69.04	71	300	720
1	0.016	69.02	69.02	71	330	720
1	0.036	69.15	69.15	71	350	720
1	0.083	69.12	69.13	71	400	720
0.991	0.031	68.67	69.39	71	450	720
1	0.01	68.95	68.96	71	465	720
1	0.021	69.01	68.99	71	475	720
1.001	0.032	69.1	69	71	500	720

To calculate the mean estimation for these samples, we use the following relationships:

1. Unit mean estimation

$$\hat{\bar{Y}} = \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

2. Mean estimation using the ratio of two variable:

$$\hat{\bar{Y}}_R = \hat{R}\mu_x \,, \hat{R} = \frac{\sum y_i}{\sum x_i} =$$

3. Mean estimation using simple linear regression:

$$\hat{\bar{Y}}_{lr} = \bar{y} + b(\mu_x - \bar{x}) \;, \quad b = \rho\frac{S_y}{S_x}$$

10.Calculating variances of the above mean estimation:

1. Unit mean estimation variance calculation:

$$V\left(\hat{\bar{Y}}\right) = \frac{(1-f)}{n}.s_y^2 \,, \qquad f = \frac{n}{N}$$

2. Calculation of mean estimation variance using the ratio between two variable:

$$V\left(\hat{\bar{Y}}_R\right) = \frac{(1-f)}{n}\left[S_y^2 - 2\rho RS_yS_x + R^2S_x^2\right], \quad f = \frac{n}{N}$$

3. Calculation of mean estimation variance using simple linear regression:

$$V\left(\hat{\bar{Y}}_{lr}\right) = \frac{(1-f)}{n}.S_y^2[1-\rho^2], \quad f = \frac{n}{N}$$

11.Calculation of mean sample confidence intervals:

1. Calculation unit mean confidence intervals (95%):

$$CI = \bar{y} \pm (Z)S.E(\bar{y})$$

2. Calculation the mean estimation confidence intervals (95%) using ration between two variables:

$$CI = \bar{y}_R \pm (Z)S.E(\bar{y}_R)$$

3. Calculation the mean estimation confidence intervals (95%) using simple linear regression :

$$CI = \bar{y}_{lr} \pm (Z)S.E(\bar{y}_{lr})$$

12.The results of the estimation of means and variance are presented in the tables below:

Table(2-a)

	$\hat{\bar{Y}}$	$\hat{\bar{Y}}_R$	$\hat{\bar{Y}}_{lr}$	$V(\hat{\bar{Y}})$
$Samp_1$	67.64	71	68.15	1.056
$Samp_2$	68.7	70.15	68.95	0.657
$Samp_3$	68.75	69.79	68.82	0.471
$Samp_4$	67.65	69.51	67.96	0.513
$Samp_5$	67.99	70.08	68.25	0.463
$Samp_6$	69.52	70.01	69.62	0.288
$Samp_7$	69.12	70.65	69.27	0.389
$Samp_8$	69.28	72.42	70.18	0.311
$Samp_9$	68.73	70.65	68.99	0.174
$Samp_{10}$	69.68	71.43	69.96	0.065
$Samp_{11}$	68.87	70.92	68.98	0.051
$Samp_{12}$	69	70.57	69.09	0.028
$Samp_{13}$	69.03	70.92	69.06	0.02
$Samp_{14}$	69.25	71.14	69.33	0.016
$Samp_{15}$	69.02	71	69.05	0.014
$Samp_{16}$	69.15	71	69.21	0.012
$Samp_{17}$	69.12	71	69.28	0.01
$Samp_{18}$	68.76	70.36	68.81	0.008
$Samp_{19}$	68.95	71	68.97	0.006
$Samp_{20}$	69.01	71	69.05	0.007
$Samp_{21}$	69.1	71.12	69.13	0.005

Table(2-b)

$V(\hat{\bar{Y}}_R)$	$V(\hat{\bar{Y}}_{lr})$	ρ	$\hat{R}$	b
1.857	1.03	0.158	1	0.152
0.956	0.643	0.148	0.988	0.174
0.88	0.466	0.097	0.983	0.07
0.719	0.504	0.13	0.979	0.164
0.736	0.457	0.113	0.987	0.125
0.426	0.278	0.179	0.986	0.197
0.566	0.387	0.074	0.995	0.098
0.448	0.286	0.285	1.02	0.289
0.31	0.171	0.142	0.995	0.136
0.104	0.063	0.148	1.006	0.158
0.091	0.051	0.049	0.999	0.052
0.056	0.028	0.063	0.994	0.059
0.044	0.02	0.017	0.999	0.015
0.031	0.016	0.041	1.002	0.039
0.027	0.014	0.016	1	0.017
0.023	0.013	0.36	1	0.035
0.019	0.01	0.083	1	0.088
0.014	0.008	0.031	0.991	0.031
0.014	0.006	0.01	1	0.01
0.013	0.007	0.021	1	0.022
0.011	0.005	0.032	1.001	0.031

And to reach the total sums of estimation for the sample, we shall use following relationship:

1. Estimation of the total sum using unit mean:

$$\hat{Y} = N\hat{\bar{Y}}$$

2. Estimation of the total sum using ratio estimation between two variable:

$$\hat{Y}_R = N\hat{\bar{Y}}_R$$

3. Estimation of the total sum using simple linear regression:

$$\hat{Y}_{lr} = N\hat{\bar{Y}}_{lr}$$

13. Calculating variance of total sum estimate:

1. Calculating the unit total sum variance:

$$V(\hat{Y}) = N^2 V(\hat{\bar{Y}})$$

2. Calculating total sum estimate variance using ratio of two variable:

$$V(\hat{Y}_R) = N^2 V(\hat{\bar{Y}}_R)$$

3. Calculating total sum estimate variance using simple linear regression:

$$V(\hat{Y}_{lr}) = N^2 V(\hat{\bar{Y}}_{lr})$$

14. The results of the estimations of sample total sums and variance are presented in the following tables:

Table(3-a)

	$\hat{Y}$	$\hat{Y}_R$	$\hat{Y}_{lr}$
$Samp_1$	48701	51120	49068
$Samp_2$	49464	50508	49644
$Samp_3$	49500	50249	49550
$Samp_4$	48708	50047	48931
$Samp_5$	48953	50458	49140
$Samp_6$	48953	50810	49745
$Samp_7$	49766	50868	49874
$Samp_8$	49882	52142	50530
$Samp_9$	49486	50868	49673
$Samp_{10}$	50170	51430	50371
$Samp_{11}$	49586	51062	49666
$Samp_{12}$	49680	50810	49745
$Samp_{13}$	49680	50810	49745
$Samp_{14}$	49870	51221	49918
$Samp_{15}$	49694	51120	49716
$Samp_{16}$	49788	51120	49831
$Samp_{17}$	49766	51120	49882
$Samp_{18}$	49507	50659	49543
$Samp_{19}$	49644	51120	49658
$Samp_{20}$	49687	51120	49716
$Samp_{21}$	49752	51206	49774

Table(3-b)

$V(\hat{Y})$	$V(\hat{Y}_R)$	$V(\hat{Y}_{lr})$
547430	962669	533952
340589	495590	333331
244166	456192	24574
265939	372729	261274
240019	381542	236909

14515	29030	14515
201658	293414	200620
161222	232243	147744
90202	160704	90202
33696	53913	32659
26438	47174	26438
14515	29030	14515
14515	29030	14515
8294	16070	8294
7258	13997	7258
6221	11923	6739
5184	98496	5184
4147	7258	4147
3110	7258	3110
3629	6739	3629
2592	5702	2592

15. The results of the relative efficiency of the sample variance are displayed in the following table:

Table(4)

	$Eff(\hat{\bar{Y}},\hat{\bar{Y}}_R)$	$Eff(\hat{\bar{Y}},\hat{\bar{Y}}_{lr})$	$Eff(\hat{\bar{Y}}_R,\hat{\bar{Y}}_{lr})$
$Samp_1$	0.569	1.025	1.803
$Samp_2$	0.687	1.022	1.487
$Samp_3$	0.535	1.011	1.888
$Samp_4$	0.713	1.018	1.427
$Samp_5$	0.629	1.013	1.611
$Samp_6$	0.676	1.063	1.532
$Samp_7$	0.687	1.005	1.463
$Samp_8$	0.694	1.091	1.572
$Samp_9$	0.561	1.018	1.813
$Samp_{10}$	0.925	1.032	1.651
$Samp_{11}$	0.56	1	1.784
$Samp_{12}$	0.5	1	2
$Samp_{13}$	0.045	1	2.2
$Samp_{14}$	0.516	1	1.938
$Samp_{15}$	0.519	1	1.929
$Samp_{16}$	0.522	1	1.917
$Samp_{17}$	0.526	1	1.2
$Samp_{18}$	0.571	1	1.75
$Samp_{19}$	0.429	1	2.333
$Samp_{20}$	0.538	1	1.857
$Samp_{21}$	0.455	1	2.2

Conclusion:

In this study we have under taken a practical inquiry to verify whether ratio or regression estimation have more accuracy. The results in the table that the value $\rho < 0.5$ in the all instances. Also, regression and mean per unit estimations are better than ratio estimation. In addition, regression estimation is better than mean estimation per unit in small samples. However, in the large samples, regression and mean estimation are superior to ratio estimation and they have equal efficiency.

References

1. Brown J.D, Questions and answers about language testing statistics: Sample size and Statistical Precision, (University of Hawai'i at Manoa), Shiken: JALT Testing & Evaluation SIG Newsletter, 11 (2) August 2007 (p. 21 - 24).
2. Housila.P, Singh.1, Sarjinder Singh, and Jong-Min Kim, General Families of Chain Ratio Type Estimators of the Population Mean Withknown Coefficient of Variation of the Second Auxiliray of Variable in Two Phase Sampling, Journal of the Korean Statistical Society (2006), 35: 4, pp 377–395.
3. Margaret. H. Smith, A Sample/Population Size Activity:Is it the sample size of the sample as a fraction of the population that matters?Pomona College Journal of Statistics Education Volume 12, Number 2 (2004), www.amstat.org/publications/jse/v12n2/smith.html
4. Singh.R, Chauhan.P, Sawan.N. and Smarandache. F, Ratio Estimators in Simple Random Sampling
5. Using Information on Auxiliary Attribute, Department of Statistics, Banaras Hindu University. Email: smarand@unm.edu
6. Sachin.M, and singh.R, A family of Estimators of Population Mean using Information on Point Bi-Serial and Phi Correlation Coefficient, International jornal of statistics & Economics, volume 10, (2013).
7. William.G. Cochran, Sampling Techniques, USA, New York, (1977).

6/13/2014