VIVEKANANDA COLLEGE
THAKURPUKUR
KOLKATA-700063

NAAC ACCREDITED 'A' GRADE

Topic: Theory of Ratio and Regression Estimator

Course Title: Sample Survey

Paper: CC-8

Unit: 2

Semester: 4

Name of the Teacher: Riddhi Das Majumder

Name of the Department: Statistics

Ratio and Regression Estimators: Use of auxillary information

(1) Ratio Estimators

Frequently we come across situations in which the ratio of y to another character x is believed to be less variable than the y's themselves. In that case it would be better to estimate R, the ratio of y to x in the population, from the sample and then multiply it by the known total of x to estimate the total for y. This procedure is called ratio estimation.

Frequently we wish to estimate a ratio rather than a total or mean, for example, it is desired to estimate the total agricultural area in a region containing N communes. There are very big communes and very small communes and this makes the character y vary tremendously over the region. But the ratio of agricultural area and the popln size of the commune, which is the per capita agricultural area, would be less variable.

Let Y, and X be the total agricultural area and the total popln in the region. Then the per-capita agricultural area in the region is $R = \frac{Y}{X}$. If a simple random sample of n communes gives $\sum_{i=1}^{n} y_i$ and $\sum_{i=1}^{n} x_i$ as the total for y and x, respectively, it is natural to estimate R by $\hat{R} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \frac{\bar{y}}{\bar{x}}$ and the total of y (i.e. Y) is estimated by $\hat{Y}_R = \hat{R} \cdot X = \frac{\bar{y}}{\bar{x}} X$, where X is known total of x. It should be noted that the two problems are different, although they are connected. For estimating Y we could have used information on any character x; this information need not to be recent, but must be known for the entire population. On the otherhand, information on a sample basis is required for y as well as for x (the denominator of the ratio) if the purpose is to estimate the ratio $R = \frac{Y}{X}$ in population.

Since the theory is the same in either case, most of the subsequent results will be relate to the problem of estimating a ratio.

Bias of the ratio estimator: The following theorem gives the exact bias associated with $\hat{R}$.

Theorem: In simple random sampling, bias of the ratio estimator $\hat{R} = \frac{\bar{y}}{\bar{x}}$ is given by $B(\hat{R}) = -\frac{Cov(\hat{R}, \bar{x})}{E(\bar{x})}$

Proof: As $Cov\left(\frac{\bar{y}}{\bar{x}}, \bar{x}\right) = E(\bar{y}) - E\left(\frac{\bar{y}}{\bar{x}}\right)E(\bar{x})$, we have

$$\bar{x}\, E\left(\frac{\bar{y}}{\bar{x}}\right) = \bar{y} - Cov\left(\frac{\bar{y}}{\bar{x}}, \bar{x}\right), \Rightarrow E(\hat{R}) = R - \frac{1}{\bar{x}} Cov(\hat{R}, \bar{x})$$

Therefore $B(\hat{R}) = E(\hat{R}) - R = -\frac{1}{\bar{x}} Cov(\hat{R}, \bar{x})$

Corollary: Denoting the standard deviation of $\hat{R}$ by $\sigma(\hat{R})$, we have $B(\hat{R}) = -\frac{1}{\bar{x}} \cdot \sigma(\hat{R})\sigma(\bar{x})\rho(\hat{R}, \bar{x})$.

or $\frac{B(\hat{R})}{\sigma(\hat{R})} = -\rho(\hat{R}, \bar{x}) \cdot \frac{\sigma(\bar{x})}{\bar{x}} = -\rho(\hat{R}, \bar{x}) \cdot C.V(\bar{x})$

Hence, $\frac{|B(\hat{R})|}{\sigma(\hat{R})} \le C.V(\bar{x})$, since $|\rho(\hat{R}, \bar{x})| \le 1$.

where C.V. stands for the coefficient of variation. The same bound applies, of course, to the bias in $\hat{Y}_R$ and $\bar{Y}_R$.

Remarks (1) $\hat{R}$ is consistent for R in the sense that $\hat{R} \to R$ when the sample size is n.

(2) The bias associated with $\hat{Y}_R = \hat{R}X$ is $X B(\hat{R})$.

(3) $\hat{R}$ is unbiased if $\rho(\hat{R}, \bar{x}) = 0$

Theorem The approximate bias and mean square error (MSE) of the ratio estimator $\hat{R}$ are $B(\hat{R}) = \frac{\left(\frac{1}{n} - \frac{1}{N}\right)}{\bar{X}^2}(RS_x^2 - \rho S_y S_x)$

and $MSE(\hat{R}) = \frac{\left(\frac{1}{n} - \frac{1}{N}\right)}{\bar{X}^2}(S_y^2 + R^2 S_x^2 - 2R\rho S_x S_y)$

Proof: Define $e_0 = \frac{\bar{y} - \bar{Y}}{\bar{Y}}$, and $e_1 = \frac{\bar{x} - \bar{X}}{\bar{X}}$.

It may be noted that

(i) $E(e_0) = E\left(\frac{\bar{y} - \bar{Y}}{\bar{Y}}\right) = 0$ (ii) $E(e_1) = 0$

(iii) $E(e_0^2) = E\left(\frac{\bar{y} - \bar{Y}}{\bar{Y}}\right)^2 = \frac{V(\bar{y})}{\bar{Y}^2}$ (iv) $E(e_1^2) = \frac{Var(\bar{x})}{\bar{x}^2}$

(v) $E(e_0 e_1) = E\left\{\frac{(\bar{x} - \bar{X})(\bar{y} - \bar{Y})}{\bar{x}\bar{Y}}\right\} = \frac{Cov(\bar{x}, \bar{y})}{\bar{X}\bar{Y}}$

Assume that the sample size is large enough so that $|e_0| < 1$ and $|e_1| < 1$ i.e. $0 < \bar{x} < 2\bar{X}$, $0 < \bar{y} < 2\bar{Y}$.

Since $\bar{y} = \bar{Y}(1+e_0)$, $\bar{x} = \bar{X}(1+e_1)$ (SRSWOR) the estimator $\hat{R} = \frac{\bar{y}}{\bar{x}}$ can be written as $\hat{R} = \frac{\bar{Y}(1+e_0)}{\bar{X}(1+e_1)} = R(1+e_0)(1+e_1)^{-1}$

$= R\{1 + e_0 - e_1 + e_1^2 - e_0e_1 + \cdots\}$. Hence, $E(\hat{R}) - R = B(\hat{R})$

$\simeq R\{E(e_1^2) - E(e_0e_1)\} = R\left\{\frac{V(\bar{x})}{\bar{X}^2} - \frac{Cov(\bar{x},\bar{y})}{\bar{X}\bar{Y}}\right\}$

$= \frac{\left(\frac{1}{n} - \frac{1}{N}\right)}{\bar{X}^2}\left\{R S_x^2 - \rho S_x S_y\right\}$.

[In SRSWOR, $V(\bar{x}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_x^2$, $V(\bar{y}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2$ and $Cov(\bar{x},\bar{y}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_{xy} = \left(\frac{1}{n} - \frac{1}{N}\right)\rho S_x S_y$]

Again, $MSE(\hat{R}) = E(\hat{R} - R)^2 \simeq R^2 E[e_0^2 + e_1^2 - 2e_0e_1]$

, ignoring terms of degree greater than two.

Therefore $MSE(\hat{R}) \simeq R^2\left\{\frac{V(\bar{x})}{\bar{X}^2} + \frac{V(\bar{y})}{\bar{Y}^2} - 2\frac{Cov(\bar{x},\bar{y})}{\bar{X}\bar{Y}}\right\}$

$\simeq \frac{\left(\frac{1}{n} - \frac{1}{N}\right)}{\bar{X}^2}\left\{R^2 S_x^2 + S_y^2 - 2R\rho S_x S_y\right\}$.

Remark:

(1) $B(\hat{Y}_R) = B(N\bar{x}\hat{R}) = N\left(\frac{N-n}{Nn}\right)\frac{1}{\bar{X}}\left\{R S_x^2 - \rho S_x S_y\right\}$

and $MSE(\hat{Y}_R) = N^2 \frac{N-n}{nN}\left\{R^2 S_x^2 + S_y^2 - 2R\rho S_x S_y\right\}$.

(2) The quantity $\frac{Bias}{S.E}$, which is the same for $\hat{R}$, $\hat{Y}_R$, $\hat{\bar{Y}}_R$, may be expressed as $\frac{Bias}{S.E} = C.V(\bar{x}) \cdot \frac{(RS_x - \rho S_y)}{\sqrt{\{R^2 S_x^2 + S_y^2 - 2R\rho S_x S_y\}}}$

where $C.V(\bar{x}) = \frac{\sqrt{V(\bar{x})}}{\bar{X}} = \frac{\left(\frac{1}{n} - \frac{1}{N}\right)^{\frac{1}{2}} S_x}{\bar{X}}$.

■ The following theorem gives the condition under which the ratio estimator will be more efficient than the conventional expansion estimator, or estimator based on the mean per unit ($\bar{y}$)

Theorem: The ratio estimator $\hat{Y}_R = \frac{\bar{y}}{\bar{x}} X = \hat{R} X$ is more efficient than the expansion estimator $\hat{Y}$ ■, with simple random sample, if $\rho > \frac{1}{2}\frac{C.V(x)}{C.V(y)} = \frac{1}{2}\left(\frac{S_x}{\bar{X}}\right)\Big/\left(\frac{S_y}{\bar{Y}}\right)$.

Proof: $V(\bar{y}) > MSE(\hat{Y}_R)$, under SRS

$\Rightarrow N^2 \frac{N-n}{nN} S_y^2 > N^2 \frac{N-n}{Nn} \{R^2 S_x^2 + S_y^2 - 2R\rho S_x S_y\}$ approximately.

$\Rightarrow S_y^2 > \{S_y^2 + R^2 S_x^2 - 2R\rho S_x S_y\}$

$\Rightarrow \rho > \frac{1}{2} \cdot R \cdot \frac{S_x}{S_y} = \frac{1}{2} \cdot \left(\frac{S_x}{\bar{x}}\right) \Big/ \left(\frac{S_y}{\bar{y}}\right) = \frac{1}{2} \cdot \frac{C.V(x)}{C.V(y)}$, if $R > 0$

Hence the proof. [The theorem holds for large samples]

Estimated MSE under simple random sampling:

Note that $\sum_{i=1}^{N} [Y_i - RX_i]^2 = \sum_{i=1}^{N} [Y_i - \bar{Y} + \bar{Y} - RX_i]^2$

$= \sum_{i=1}^{N} [Y_i - \bar{Y} + R\bar{X} - RX_i]$, since $R = \frac{\bar{Y}}{\bar{X}}$.

$= \sum_{i=1}^{N} [Y_i - \bar{Y}]^2 + R^2 \sum_{i=1}^{N} [X_i - \bar{X}]^2 - 2R \sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y})$

$\Rightarrow \frac{1}{N-1} \sum_{i=1}^{N} [Y_i - RX_i]^2 = S_y^2 + R^2 S_x^2 - 2R\rho S_x S_y$.

Then, we have $MSE(\hat{R}) = \left(\frac{1}{n} - \frac{1}{N}\right) \frac{1}{\bar{X}^2} \cdot \frac{1}{N-1} \sum_{i=1}^{N} (Y_i - RX_i)^2$

Therefore a reasonable estimator for the MSE of the ratio estimate is

$$v(\hat{R}) = \left(\frac{1}{n} - \frac{1}{N}\right) \cdot \frac{1}{\bar{x}^2} \cdot \frac{\sum_{i=1}^{n} (y_i - \hat{R}x_i)^2}{n-1} \quad \text{where } \hat{R} = \frac{\bar{y}}{\bar{x}}.$$

The estimator is biased.

Now $\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \hat{R}x_i)^2 = \frac{1}{n-1} \left\{ \sum_{i=1}^{n} y_i^2 + \hat{R}^2 \sum_{i=1}^{n} x_i^2 - 2\hat{R} \sum_{i=1}^{n} x_i y_i \right\}$

$= \frac{1}{n-1} \sum_{i=1}^{n} \{(y_i - \bar{y}) - \hat{R}(x_i - \bar{x})\}^2 = \{s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R} s_{xy}\}$

Where $s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$

Hence, $\boxed{v(\hat{R}) = \left(\frac{1}{n} - \frac{1}{N}\right) \frac{1}{\bar{x}^2} \{s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R} s_{xy}\}}$

Since $\hat{Y}_R = \hat{R}\bar{X}$, $\boxed{v(\hat{Y}_R) = N^2 \left(\frac{1}{n} - \frac{1}{N}\right) \{s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R} s_{xy}\}}$

Unbiased Ratio-type estimators:

To modify the usual ratio estimator itself so that a ratio-type estimator is obtained that is unbiased under a simple random sampling. Actually, the estimator $\bar{r} = \frac{1}{n}\sum_{i=1}^{n} r_i = \frac{1}{n}\sum_{i=1}^{n} \frac{y_i}{x_i}$ is corrected for its bias to obtain an unbiased estimator.

Theorem: In simple random sampling, an unbiased estimator of $R = \bar{Y}/\bar{X}$ is given by

$$\hat{R} = \bar{r} + \frac{(N-1)n}{N(n-1)} \cdot \frac{\bar{y} - \bar{r}\bar{x}}{\bar{x}}$$

Proof: $\frac{1}{N}\sum_{i=1}^{N} R_i (x_i - \bar{x})$, where $R_i = \frac{Y_i}{X_i}$, $i = 1(1)N$.

$$= \frac{1}{N}\sum_{i=1}^{N} (Y_i - \bar{x} R_i) = \bar{Y} - \bar{X} \cdot \frac{1}{N}\sum_{i=1}^{N} R_i = \bar{Y} - \bar{X} \cdot E(r_i)$$

But in simple random sampling $E(\bar{r}) = E(r_i)$.

Hence, bias in $\bar{r} = E(\bar{r}) - R = -\frac{1}{\bar{X}N}\sum_{i=1}^{N} R_i (x_i - \bar{x})$. ——— (*)

Again, an unbiased estimator of $\frac{1}{N-1}\sum_{i=1}^{N} R_i (x_i - \bar{x})$ is $\frac{1}{n-1}\sum_{i=1}^{n} r_i (x_i - \bar{x}) = \frac{1}{n-1}\sum_{i=1}^{n} (y_i - r_i \bar{x})$

$$= \frac{n}{n-1}(\bar{y} - \bar{r}\bar{x}).$$

From (*), bias in $\bar{r} = E(\bar{r}) - R = -\frac{(N-1)n}{N(n-1)} \cdot E\left\{\frac{\bar{y} - \bar{r}\bar{x}}{\bar{x}}\right\}$

$$\Rightarrow E\left\{\bar{r} + \frac{(N-1)n}{N(n-1)} \cdot \frac{\bar{y} - \bar{r}\bar{x}}{\bar{x}}\right\} = R$$

Hence, $\hat{R}_* = \bar{r} + \frac{(N-1)n}{N(n-1)} \cdot \left(\frac{\bar{y} - \bar{r}\bar{x}}{\bar{x}}\right)$ is an unbiased estimator of $R = \frac{\bar{Y}}{\bar{X}}$.

Remark:

(i) The corresponding unbiased estimator of the pop. total Y is $\hat{Y}_{R_*} = \hat{R}_* X = \bar{r} X + \frac{(N-1)n}{n-1}(\bar{y} - \bar{r}\bar{x})$.

(ii) An unbiased estimator of the popl. mean $\bar{Y}$ is $\hat{\bar{Y}}_{R_*} = \hat{R}_* \bar{x}$

$$= \bar{r}\bar{x} + \frac{(N-1)n}{N(n-1)}(\bar{y} - \bar{r}\bar{x}).$$

Linear Regression Estimator:

Like the ratio estimator, the linear regression estimate is designed to increase precision by the use of an auxiliary variate x_i that is correlated with y_i. The ratio estimator is at its best when the relation between y and x is a straight line through the origin, that is, $y - kx = 0 \Leftrightarrow y/x = k$. When the relation between y_i and x_i is examined, it may be found that although the relation is (approximately) linear, the line does not go through the origin. This suggests an estimator based on the linear regression of y on x rather than on the ratio of the variables.

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 estimator of $\bar{Y}$, the popln mean of y_i, is

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

where b is an estimator of the change in y when x is increased by unit. The rationale behind this estimator is that if $\bar{x}$ is below average we should expect $\bar{y}$ also to be below average by an amount $b(\bar{X} - \bar{x})$ because of the regression of y_i on x_i. For an estimator of the popln Y, we take $\hat{Y}_{lr} = N\bar{y}_{lr}$.

Suppose that we can take a rapid estimate x_i of some characteristic for every unit and can also, by some more costly method, determine the correct value y_i of the characteristic for a simple random sample of the units. For an example, an eye estimate of the volume of timber was made on each of a population of $\frac{1}{10}$-acre plots, and the actual timber volume was measured for a simple random sample of the plots. The regression estimate

$$\bar{y} + b(\bar{X} - \bar{x})$$

adjusts the sample mean of the actual measurements by the regression of the actual measurements on the rapid estimates.

By a suitable choice of b, the regression estimate includes as particular cases both the mean per unit and the ratio estimate. Obviously if b is taken as zero, $\bar{y}_{lr}$ reduces to $\bar{y}$.

If $b = \frac{\bar{y}}{\bar{x}}$, $\bar{y}_{lr} = \bar{y} + \frac{\bar{y}}{\bar{x}}(\bar{X} - \bar{x}) = \frac{\bar{y}}{\bar{x}}\bar{X} = \hat{Y}_R$

Regression Estimator when $\bar{Y}$ is computed from the sample

Let $y = \bar{y} + B(x - \bar{x})$ be the pop^n regression line of y on x, where $B = \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{N}(x_i-\bar{x})^2}$ is the pop^n regression coeff^t.

Here, 'b' must be (the sample estimate of B, that is) the least squares estimate of B, that is, $b = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$

(i) Bias of the Linear regression Estimator:

Introduce the variate $e_i = y_i - \bar{y} - B(x_i - \bar{x})$.

The properties of e_i are : $\sum_{i=1}^{N} e_i = 0$

and $\sum_{i=1}^{N} e_i(x_i-\bar{x}) = \sum_{i=1}^{N}(y_i-\bar{y})(x_i-\bar{x}) - B\sum_{i=1}^{N}(x_i-\bar{x})^2 = 0$

, by defn. of B.

Now, $b = \sum_{i=1}^{n} y_i(x_i-\bar{x}) \Big/ \sum_{i=1}^{n}(x_i-\bar{x})^2$

$= \left\{\sum_{i=1}^{n}[e_i+\bar{y}+B(x_i-\bar{x})](x_i-\bar{x})\right\} \Big/ \sum_{i=1}^{n}(x_i-\bar{x})^2$

$= B + \left\{\sum_{i=1}^{n} e_i(x_i-\bar{x}) \Big/ \sum_{i=1}^{n}(x_i-\bar{x})^2\right\}$

We have $E(\bar{y}_{lr}) = \bar{y} - E\{b(\bar{x}-\bar{x})\}$. Thus one expression for bias is $-E\{b(\bar{x}-\bar{x})\} = -\mathrm{Cov}(b, \bar{x})$.

Now, $-\mathrm{Cov}(b,\bar{x}) = -\mathrm{Cov}\left(B + \frac{\sum_{i=1}^{n} e_i(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}, \bar{x}\right)$

$= -\mathrm{Cov}\left(\sum_{i=1}^{n} e_i(x_i-\bar{x}), \bar{x}\right) \Big/ \sum_{i=1}^{n}(x_i-\bar{x})^2 = -\mathrm{Cov}\left\{\sum_{i=1}^{n} e_i(x_i-\bar{x}) + n\bar{e}(\bar{x}-\bar{x}), \bar{x}\right\} \Big/ \sum_{i=1}^{n}(x_i-\bar{x})^2$

$= -E\left\{\sum_{i=1}^{n} e_i(x_i-\bar{x})(\bar{x}-\bar{x})\right\} \Big/ nS_x^2 + E\left\{\bar{e}(\bar{x}-\bar{x})^2\right\} \Big/ S_x^2$ [replacing $\sum_{i=1}^{n}(x_i-\bar{x})^2$ by its leading term nS_x^2]

$= -E(\bar{u}-\bar{U})(\bar{x}-\bar{x}) \Big/ S_x^2$, where $u_i = e_i(x_i-\bar{x})$ and $\bar{U}=0$.

$= -\left(\frac{1}{n}-\frac{1}{N}\right) \cdot E(u_i-\bar{U})(x_i-\bar{x}) \Big/ S_x^2 = -\frac{1-f}{n} \cdot E\{e_i(x_i-\bar{x})^2\} \Big/ S_x^2$

, which is the bias turns out to be in the estimator $\bar{y}_{lr}$.

to The approximate MSE of the regression estimator under SRSWOR

Theorem: If b is the least square estimate of B and $\bar{y}_{lr} = \bar{y} + b(\bar{X} - \bar{x})$, then in SRSWOR of size n, with n large,

MSE $\simeq V(\bar{y}_{lr}) \simeq \frac{1-f}{n} S_y^2(1-\rho^2)$, where $\rho = \frac{S_{yx}}{S_x S_y}$ is the population correlation between y and x.

Proof: The sampling error of $\bar{y}_{lr}$ arises from the quantity

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

As an approximation, replace $\bar{y}_{lr}$ by $\bar{y}^*_{lr} = \bar{y} + B(\bar{X} - \bar{x})$ where B is the population linear regression coefficient of y on x.

The error committed in this approximation is $(B-b)(\bar{X}-\bar{x})$

Note that $(b-B) = O(\frac{1}{\sqrt{n}})$ and $(\bar{x}-\bar{X}) = O(\frac{1}{\sqrt{n}})$, hence $(b-B)(\bar{x}-\bar{X})$ is of order $\frac{1}{n}$ in SRS. Again $V(\bar{y}^*_{lr})$ is of order $\frac{1}{n}$, since it is the variance of the sample mean of the variate $(y - Bx)$.

Hence, $E(\bar{y}_{lr} - \bar{Y})^2 = E\{\bar{y}^*_{lr} - \bar{Y} + (B-b)(\bar{x}-\bar{X})\}^2$

$= V(\bar{y}^*_{lr}) + E[(b-B)^2(\bar{x}-\bar{X})^2] + 2E[(\bar{y}^*_{lr} - \bar{Y})(b-B)(\bar{x}-\bar{X})]$

Now, $E[(b-B)^2(\bar{x}-\bar{X})^2] \le \{E(b-B)^4 E(\bar{x}-\bar{X})^4\}^{1/2}$, which is of order $1/n^2$. Similarly, $E[(b-B)(\bar{y}^*_{lr} - \bar{Y})(\bar{x}-\bar{X})]$

$\le \{E(b-B)^2\}^{3/4}\{E(\bar{y}^*_{lr} - \bar{Y})^4 . E(\bar{x}-\bar{X})^4\}^{1/4}$, which is of order $1/n^{3/2}$

Thus, the large sample variance (MSE) of the regression estimator $\bar{y}_{lr}$ is $V(\bar{y}_{lr}) \simeq V(\bar{y}^*_{lr}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2(1-\rho^2) = \frac{1-f}{n} . S_y^2(1-\rho^2)$

Sample estimate of the MSE or Variance:

Note that $V(\bar{y}_{lr}) \simeq \frac{1-f}{n} S_e^2$ where $S_e^2 = S_y^2(1-\rho^2)$.

Note that, an unbiased estimator of $S_e^2 = \frac{1}{N-1}\sum_{i=1}^{N}(e_i - \bar{e})^2$ is $s_e^2 = \frac{\sum_{i=1}^{n}(e_i - \bar{e})^2}{n-1}$

Now, $e_i - \bar{e} = y_i - \bar{y} - B(x_i - \bar{x}) = \{y_i - \bar{y} - b(x_i - \bar{x})\} + (b-B)(x_i - \bar{x})$

The 2nd term on the right, of order $\frac{1}{\sqrt{n}}$, may be neglected in relation to the 1st term, which is of order unity.

Hence, in large sample $s_e^2 \simeq \frac{1}{(n-1)}\sum_{i=1}^{n}\{y_i - \bar{y} - b(x_i - \bar{x})\}^2$ & so on $= S_y^2(1-\rho^2)$

estimate of S_e^2. The estimator $\frac{1}{n-2}\sum_{i=1}^{n}\{y_i - \bar{y} - b(x_i - \bar{x})\}^2$ is

an unbiased since it is used in regression theory

... an estimate of $V(\bar{y}_{lr})$ or $MSE(\bar{y}_{lr})$, valid in large samples, we may use

$$v(\bar{y}_{lr}) = \frac{1-f}{n(n-2)} \sum_{i=1}^{n} \{(y_i - \bar{y}) - b(x_i - \bar{x})\}^2$$

$$= \frac{1-f}{n(n-2)} \left\{ \sum_{i=1}^{n} (y_i - \bar{y})^2 - \frac{\left[\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})\right]^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right\}$$

Comparison with the Ratio and the Mean per unit, in large Sample :

Theorem: Under simple random sampling, with large sample, $V(\bar{y}) > MSE(\bar{y}_{lr})$ and $MSE(\bar{y}_R) > MSE[\bar{y}_{lr}]$

proof:

$MSE(\bar{y}_{lr}) \approx V(\bar{y}_{lr}) = \frac{N-n}{Nn} S_y^2 (1-\rho^2)$, [regression]

$MSE(\bar{y}_R) \approx V(\bar{y}_R) = \frac{N-n}{Nn} (S_y^2 + R^2 S_x^2 - 2R\rho S_x S_y)$, [Ratio]

$$V(\bar{y}) = \frac{N-n}{Nn} S_y^2 \quad \text{(Mean per unit)}.$$

Since $|\rho| < 1$, $(1-\rho^2) < 1 \Rightarrow V(\bar{y}_{mu}) \geq V(\bar{y}_{lr})$ or $MSE(\bar{y}_{lr})$

Now, $MSE(\bar{y}_R) - MSE(\bar{y}_{lr}) = \frac{N-n}{Nn} \{S_y^2 + R^2 S_x^2 - 2R\rho S_x S_y - S_y^2 + S_y^2 \rho^2\}$

$$= \frac{N-n}{Nn} \{RS_x - S_y\rho\}^2 > 0.$$

The the regression estimator is more precise than the ratio estimator unless $B = R$ ⇔ $y = kx$, that is, the relation between y and x is a straight line through the origin.

$MSE(\hat{\bar{Y}}_R) - MSE(\bar{y}_{lr}) = \left(\frac{1}{n} - \frac{1}{N}\right)(RS_x - S_y\rho)^2 \geq 0$

'=' holds iff $RS_x = S_y\rho$, iff $R = \rho \cdot \frac{S_y}{S_x} = B$.

which is true if $y = kx$, i.e if the relation b/n y and x is a st. line passing through origin.