UNIT III

RATIO AND REGRESSION ESTIMATION

RATIO ESTIMATION: Incorporation of more information regarding the variable under study gives better estimates provided that the information is valid and proper. Use of such auxiliary information is made in the **ratio method of estimation** to obtain an improved estimator of population mean or population total. In ratio method of estimation, auxiliary information on a variable is available which is linearly related to the variable under study and is utilized to estimate the population mean.

The auxiliary information about the population may include a known variable to which the variable of interest is approximately related. The auxiliary information typically is easy to measure, whereas the variable of interest may be expensive to measure.

For e.g.:

1) Expenditure on clothing (Y) per young female (X)
2) Y is population in year 2010 and X is the population in 2000 for given city (population in 1000's). Ratio estimation explains how much the population changes over the period of 10 years.

Notations: Let Y be the variable under study and X be any auxiliary variable which is correlated with Y. The population mean $\bar{X}$ of X variable (or equivalently the population total for X variable = X) must be known.

For e.g.:

1) x- income in year 2012, y = income in year 2017

2) x- expenditure in the year 2018, y = income in the year 2018

Let $(x_1, y_1), (x_2, y_2)\ldots\ldots\ldots(x_n, y_n)$ be the random sample of size n on paired variable (X, Y) drawn, preferably by simple random sampling without replacement (SRSWOR), from a population of size N. Then let

$Y = \Sigma Y_i$ = population total for Y values

$X = \Sigma X_i$ = population total for X values

$\bar{Y} = (1/N)\Sigma Y_i$ = population mean for Y values

$\bar{X} = (1/N) \Sigma X_i$ = population mean for X values

$R = \bar{Y}/ \bar{X} = (1/N)\Sigma Y_i / (1/N) \Sigma X_i = Y/X$ = Population Ratio

$S_y^2 = (1/N – 1)\Sigma (Y_i - \bar{Y})^2$ = Population mean square error of Y values

$S_x^2 =(1/N – 1) \Sigma (X_i - \bar{X})^2$ = Population mean square error of X values.

Let (x_i, y_i) denotes ith ordered pair observation for the sample.

$y = \Sigma y_i$ = sample total for y values, i=1, 2, ,n.

$x =\Sigma x_i$ = sample total for x values , i=1,2,…..,n.

$\bar{y} = (1/n)\Sigma y_i$ = sample mean of y values

$\bar{x} = (1/n) \Sigma x_i$ = sample mean of x values

$s_y^2= (1/n-1) \Sigma (y_i - \bar{y})^2$ = sample mean square error for y values

$s_x^2 =(1/n-1) \Sigma (x_i - \bar{x})^2$ = sample mean square error for x values

$r = \bar{y}/\bar{x} = \Sigma y_i / \Sigma x_i$ = sample ratio.

Estimator for population Ratio, its variance and estimator of variance:

We know,

$$R = \bar{Y}/\bar{X}$$

This can be estimated by the ratio estimator

$$\hat{R} = r = \bar{y}/\bar{x}$$

Theorem 1:

For large sample

(i) $E(r) - R \approx 0$

(ii) $V(r) \approx 1/\bar{X}^2 (1 - n/N)(S_r^2/n)$ where $S_r^2 = (1/ N-1)\Sigma (Y_i - RX_i)^2$

Proof:

Since sample is taken by method of SRSWOR, we have

$$E (\bar{y}) = \bar{Y} \text{ and } E (\bar{x}) = \bar{X}$$

(i) Consider $E(r) = E (\bar{y}/\bar{x}) \approx E (\bar{y})/\bar{X} = R$

$\Rightarrow E (r) \approx R$

$\Rightarrow E (r) - R \approx 0$ Hence proved

(ii) Consider $r - R = (\bar{y}/\bar{x}) - R = (\bar{y} - R\bar{x})/\bar{X}$

$$V (r) = E (r - R)^2 \approx E (\bar{y} - R\bar{x} /\bar{X})^2 = E (\bar{y} - R\bar{x})^2/ \bar{X}^2$$

$$= V (\bar{d})/ \bar{X}^2$$

Where $\bar{d} = \bar{y} - R\bar{x}$

$$E (\bar{d}) = E (\bar{y} - R\bar{x}) = E (\bar{y}) - RE (\bar{x}) = \bar{Y} - R\bar{X} = 0$$

Since sampling is SRSWOR and $\bar{d}$ is sample mean of d_i,

Where $d_i = y_i - Rx_i$. Therefore,

$$V (\bar{d}) = (1 - n/N)S_d^2/n$$

Where $S_d^2 = (1/N – 1)\Sigma (D_i - \bar{D})^2$ where $D_i = Y_i – RX_i$

$= (1/N – 1) \Sigma [Y_i – RX_i – (\bar{Y} - R\bar{X})]^2$

$= (1/N – 1) \Sigma (Y_i – RX_i)^2$ as $\bar{Y} - R\bar{X} = 0$

$= S_r^2$

Now, $S_r^2 = (1/N – 1) \Sigma (Y_i – RX_i)^2$

$= (1/N – 1) \Sigma [(Y_i - \bar{Y}) – (RX_i - R\bar{X})]^2$

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

$=S_y^2 – 2RS_{xy} + R^2S_x^2$……(i)

$= \bar{Y}^2(S_y^2/ \bar{Y}^2 – 2 S_{xy}/ \bar{Y} \bar{X} + S_x^2/ \bar{X}^2)$ as $R = \bar{y})/\bar{X}$

$= \bar{Y}^2 (C_{yy} – 2C_{xy} + C_{xx})$

Hence, Variance of estimator of R is

$V (r) =V(\bar{d})/\bar{X}^2$

$= (1/ \bar{X}^2) (1 – n/N)(S_r^2/n)$

$= (1/ \bar{X}^2) (1 – n/N)(S_y^2 – 2RS_{xy} + R^2S_x^2)/n$

$= (R^2/n)(1 – n/N) \bar{Y}^2 (C_{yy} – 2C_{xy} + C_{xx})$

$= (R^2/n)(1 –f)(C_{yy} – 2C_{xy} + C_{xx})$

We know that in SRSWOR, sample mean square error is an unbiased estimator of population mean square error. So estimator of S_r^2 is s_r^2.

The estimate of variance is

$\hat{v}(r) = v(\bar{d})/\bar{X}^2 = 1/ \bar{X}^2 (1 – n/N)s_r^2/n$

where $s_r^2 = (1/ n – 1)\Sigma (y_i – rx_i)^2$

$= (1/ n – 1) \Sigma [(y_i - \bar{y}) – r(x_i - \bar{x})]^2$

$$= (1/n - 1\ ) [\ \Sigma \{(y_i - \bar{y})^2 - 2r\ \Sigma(y_i - \bar{y})(x_i - \bar{x}) + r^2\ \Sigma\ (x_i - \bar{x})^2]$$

$$= s_y^2 - 2rs_{xy} + r^2 s_x^2$$

So the estimate of variance is

$$v(r) = 1/\bar{X}^2(1 - n/N)s_r^2/n$$

$$= 1/\ \bar{X}^2(1 - n/N)(s_y^2 - 2rs_{xy} + r^2 s_x^2)/n$$

Estimator for population Mean, its variance and estimator of variance:

The ratio estimate of population mean $\bar{Y}$ is

$$\hat{\bar{Y}} = (\bar{y}/\bar{x})\bar{X} = r\bar{X}$$

$$\Rightarrow \quad \hat{\bar{Y}}_R = r\bar{X}$$

The estimator r for R is biased , so $\hat{\bar{Y}}_R$ is also biased for $\bar{Y}$.

$$E(\hat{\bar{Y}}_R) = E(r\bar{X}) = E(r)\bar{X} \approx R\ \bar{X} = \bar{Y}$$

Variance of estimator:

$$V(\hat{\bar{y}}_R) = V(\ \hat{\bar{Y}}_R) = V(r\bar{X}) = \bar{X}^2\ V\ (r)$$

$$= (1 - f)/n[\bar{Y}^2(\ C_{yy} - 2C_{xy} + C_{xx}\ )]$$

$$= (1 - n/N)(s_y^2 - 2rs_{xy} + r^2 s_x^2)/n$$

Estimator for population total, its variance and estimator of variance:

We know , the ratio estimator of the population total Y is

$$\hat{\bar{Y}}_R = r\bar{X}$$

Variance of the estimator:

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

$= N^2(1 – n/N)(S_r^2/n)$

$= N^2(1 – n/N)(\bar{Y}^2/n)(C_{yy} – 2C_{xy} + C_{xx})$

Estimator of Variance:

$V(\hat{\bar{Y}}_R) = v(\hat{y}_R) = N^2(1 – n/N) s_r^2/n$

$= N^2(1 – n/N) (s_y^2 - 2rs_{xy} + r^2 s_x^2)/n$

Expression for bias for r:

We know,

$Cov (r, \bar{x}) = E(r \bar{x}) – E(r)E(\bar{x})$

$= E[(\bar{y}/\bar{x})\bar{x}] – E(r)E(\bar{x})$

$=E(\bar{y}) – E(r)E(\bar{x})$

$= \bar{Y} - E(r)\bar{X}$

$=>Cov(r, \bar{x}) = \bar{Y} - E(r)\bar{X}$

$=>E(r) = (\bar{Y}/ \bar{X})- Cov(r,\bar{x})/ \bar{X}$

$= R - Cov(r,\bar{x})/ \bar{X}$

$Bias = E(r) – R = - Cov(r,\bar{x})/ \bar{X}$

$= - \rho_{r\bar{x}}\sigma_r\sigma_{\bar{x}}/ \bar{X}$

$|Bias| = |E(r) – R|$

$= |- \rho_{r\bar{x}}|\sigma_r\sigma_{\bar{x}}/ \bar{X} \leq \sigma_r\sigma_{\bar{x}}$ since $|- \rho_{r\bar{x}}| \leq 1$

$=>| Bias/\sigma_r| = |[E(r) – R]/ \sigma_r| \leq \sigma_{\bar{x}}/ \bar{X}$

$=>| Bias/\sigma_r| \leq C_x$

Where C_x is the coefficient of variation of X.

Thus, if the CV($\bar{x}$) is small, the bias of $\hat{R} = r$ is small relative to SE(r). But if n is small, the bias can be large.

Bias for $\bar{y}_R$:

$$\text{Bias} = E(\bar{y}_R) - \bar{Y} = E(r\bar{X}) - R(\bar{X})$$

$$= \bar{X}[E(r) - R]$$

$$= \bar{X}[-\text{Cov}(r,\bar{x})/\bar{X}]$$

$$= -\text{Cov}(r,\bar{x})$$

$$\Rightarrow E(\bar{y}_R) - \bar{Y} = -\text{Cov}(r,\bar{x}) = -\rho_{r\bar{x}}\sigma_r\sigma_{\bar{x}}$$

So, $|E(\bar{y}_R) - \bar{Y}| = |-\rho_{r\bar{x}}\sigma_r\sigma_{\bar{x}}| \leq \sigma_r\sigma_{\bar{x}}$

Comparing efficiency of ratio estimator with ordinary estimator of SRSW)OR:

We have

$$V(\hat{\bar{Y}}_R) = (1 - n/N)(S_r^2/n)$$

And $V(\hat{\bar{Y}})_{SRSWOR} = (1 - n/N)(S_y^2/n)$

The ratio estimator is more efficient estimate of population mean than sample mean based on SRSWOR if

$$V(\hat{\bar{Y}})_{SRSWOR} - V(\hat{\bar{Y}}_R) > 0$$

$$\Rightarrow (1 - n/N)\{(S_y^2 - S_r^2)/n\} > 0$$

$$\Rightarrow (S_y^2 - S_r^2) > 0$$

$$\Rightarrow \{S_y^2 - (S_y^2 - 2R\rho S_y S_x + R^2 S_x^2)\} > 0$$

$$\Rightarrow 2R\rho S_y S_x - R^2 S_x^2 > 0$$

$$\Rightarrow \rho > RS_x/2S_y = CV_x/2CV_y \quad \text{since } R = \bar{Y}/\bar{X} \text{ and } CV_x = S_x/\bar{X},\ CV_y = S_y/\bar{Y}$$

$$\Rightarrow \rho > CV_x/2\,CV_y$$

Ratio estimator is more efficient estimate of population mean than sample mean based on SRSWOR if $\rho > CV_x/2\ CV_y$

Both are equally efficient if $\rho = CV_x/2\ CV_y$

And ratio estimator is less efficient estimate of population mean than sample mean based on SRSWOR if $\rho < CV_x/2\ CV_y$

Regression Estimation:

Regression estimation is also another method of estimation like ratio estimation. This method also uses a finite population total and the knowledge of an auxiliary variable is used which is closely related to the study variable y. If the relation between X_i and Y_i is examined and it is found to be approximately linear, but the line does not pass through the origin, linear estimates are to be used instead of ratio estimates.

Since $\hat{\bar{Y}}_r = \hat{R}X = (\bar{y}/\bar{x})X = (y/x)X$, which is of the form y = mx which is equation of straight line with slope m and the line passes through origin (0, 0). However, the linear relationship between X and Y may not pass through the origin. So a more general estimator is the regression estimator.

For the linear regression estimates, values of X_i and Y_i are measured for each unit of the sample. Let us suppose that the βpopulation mean for the auxiliary variable X_i is known, then the linear regression estimate of the population mean is defined as

$\bar{Y}_{lr} = \bar{y} + \beta$where β is the regression coefficient.

(A) Regression coefficient β is known:

(i) Estimation of Population mean, its variance and estimator of $(\bar{X} - \bar{x})$ **variance:**

Let us assume that the value of regression coefficient β is known. Let it be β_0. So the regression estimator of population mean is

$$\bar{Y}_{lr} = \bar{y} + \beta_0(\bar{X} - \bar{x})$$

Let $E(\bar{Y}_{lr}) = E\{ \bar{y} + \beta_0(\bar{X} - \bar{x})\}$

$= E(\bar{y}) + \beta_0 E(\bar{X} - \bar{x})$

$= \bar{Y} + \beta_0 E(\bar{X} - \bar{x})$

$= \bar{Y}$

Thus $\bar{Y}_{lr}$ is an unbiased estimator of $\bar{Y}$ when β is known.

Variance of $\bar{Y}_{lr}$:

$V(\bar{Y}_{lr}) = E (\bar{Y}_{lr} - \bar{Y})^2$

$= E(\bar{y} + \beta_0(\bar{X} - \bar{x}) - \bar{Y})^2$

$= E[(\bar{y} - \bar{Y}) - \beta_0(\bar{x} - \bar{X})]^2$

$= E(\bar{y} - \bar{Y})^2 - 2 \beta_0 E(\bar{y} - \bar{Y}) (\bar{x} - \bar{X}) + \beta_0^2 E (\bar{x} - \bar{X})^2$

$= V(\bar{y}) - 2\beta_0 Cov(\bar{y},\bar{x}) + \beta_0^2 V(\bar{x})$.....(a)

Since sampling is done by SRSWOR,

$V(\bar{y}) = (1 – f/n)S_y^2$, $Cov(\bar{y},\bar{x}) = (1 – f/n)S_{xy}$, $V(\bar{x})=(1 – f/n)S_x^2$

Where, $S_y^2 = (1/N – 1)\Sigma (Y_i - \bar{Y})^2$, $S_x^2= (1/N – 1)\Sigma (X_i - \bar{X})^2$

$S_{xy} = (1/N – 1) \Sigma (Y_i - \bar{Y})(X_i - \bar{X})$

Hence (a) reduces to

$$V(\bar{Y}_{lr})= (1-f/n)(S_y^2 - 2\beta_0 \rho S_x S_y + \beta_0^2 S_x^2)$$

Estimator of variance:

We have

$$V(\bar{Y}_{lr})= (1-f/n)(S_y^2 - 2\beta_0 \rho S_{xy} + \beta_0^2 S_x^2)$$

Since the sample is selected by SRSWOR,

We know that $\hat{S}_y^2=s_y^2$, $\hat{S}_x^2=s_x^2$, $\hat{S}_{xy}=s_{xy}$

where $s_y^2= (1/n-1)\Sigma(y_i - \bar{y})^2$, $s_x^2= (1/n-1)\Sigma(x_i - \bar{x})^2$, $s_{xy}=(1/n-1)\ \Sigma(x_i - \bar{x})(y_i - \bar{y})$

Estimator of Population total, its variance and estimator of variance:

We know population total $Y = N\bar{Y}$ and $\bar{Y}_{lr}$ is an unbiased estimator of $\bar{Y}$ when β is known. So estimator of population total is

$$N\ \bar{Y}_{lr} = N[\bar{y} + \beta_0(\bar{X} - \bar{x})]$$

We have $V(\bar{Y}_{lr})= (1-f/n)(S_y^2 - 2\beta_0\rho S_{xy} + \beta_0^2\ S_x^2)$

$$V(N\ \bar{Y}_{lr}) = N^2\ V(\bar{Y}_{lr}) = N^2[(1-f/n)(S_y^2 - 2\beta_0\rho S_{xy} + \beta_0^2\ S_x^2)]$$

Estimator of variance:

$$v\ (N\bar{y}_{lr})= N^2v(\bar{y}_{lr})=N^2[(1-f/n)(s_y^2- 2\beta_0 s_{xy} + \beta_0^2 s_x^2)$$

Comparison between Simple estimate and Regression estimate:

$$V(\bar{y}) = (1 – f/n)S_y^2$$

$$V(\bar{Y}_{lr}) = (1-f\ /n)(S_y^2 - 2\beta_0\rho S_{xy} + \beta_0^2\ S_x^2)$$

Comparing these two variances

$$V(\bar{Y}_{lr})< V(\bar{y})$$

$$=>(1-f\ /n)(S_y^2 - 2\beta_0\rho S_{xy} + \beta_0^2\ S_x^2)< (1 – f/n)S_y^2$$

$$= >(S_y^2 - 2\beta_0\rho S_{xy} + \beta_0^2\ S_x^2)< S_y^2$$

$$=>\beta_0\ S_x^2(\beta_0 - 2\rho S_y/S_x)< 0$$

This is possible if

Either $\beta_0 < 0$ and $(\beta_0 - 2\rho S_y/S_x)>0$

$$=>2\rho S_y/S_x < \beta_0 < 0$$

Or, $\beta_0 >0$ and $(\beta_0 - 2\rho S_y/S_x)< 0$

$=>0 < \beta_0 < 2\rho S_y/S_x$ i.e., $0 < \beta_0 < 2S_y/S_x$

(B)Regression coefficient β is unknown:

(i) Estimation of Regression coefficient β:

$\bar{y}_{lr} = \bar{y} + \beta_0(\bar{X} - \bar{x})$

$V(\bar{y}_{lr}) = (1\text{-}f/n)(S_y^2 - 2\beta S_{xy} + \beta^2 S_x^2)$

We can obtain the value of β which minimises $V(\bar{y}_{lr})$ by

$d[V(\bar{y}_{lr})]/d\beta = 0$

$=>-2S_{xy} +2 \beta S_x^2 = 0$

$=> \beta = S_{xy}/ S_x^2$

The minimum value of variance is obtained by substituting

$\beta = S_{xy}/ S_x^2$ in $V(\bar{y}_{lr})$

So the minimum value of $V(\bar{y}_{lr})$is

$V(\bar{y}_{lr}) =(1 – f/n) S_y^2(1 – \rho^2)$ where $\rho = S_{xy}/S_xS_y$

Estimator of β:

Estimate of β is $b = s_{xy}/s_x^2$

So the regression estimator of population mean is

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

Estimate of variance is

$V(\bar{y}_{lr}) = (1 – f/n)s_y^2[1 –(S_{xy}/S_yS_x)^2]$

$= (1 – f/n)s_{reg}^2$ where $s_{reg}^2 = s_y^2[1 –(S_{xy}/S_yS_x)^2]$

Bias in regression estimate of population mean:

$$\text{Bias} = E(\bar{Y}_{lr}) - \bar{Y} = E[\bar{y} + b(\bar{X} - \bar{x})] - \bar{Y}$$

$$= \bar{Y} + E[b(\bar{X} - \bar{x})] - \bar{Y}$$

$$= E[b(\bar{X} - \bar{x})]$$

$$= -\text{Cov}(b, \bar{x})$$

The regression estimator is a biased estimator if β is unknown.

Estimation of population total:

The estimate of population mean is

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

So an estimate of population total is

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

Variance of an estimate of population total is

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

$$= N^2[(1 - f/n)\, S_y^2(1 - \rho^2)]$$

Estimate of variance is

$$v(\bar{y}_{lr}) = v(\bar{y}_{lr}N) = N^2v(\bar{y}_{lr})$$

$$= N^2[(1 - f/n)s_y^2[1 - (S_{xy}/S_yS_x)^2]$$

Comparison between regression estimator and sample mean under SRSWOR:

We know,

$$V(\bar{y})_{SRSWOR} = [(1 - f/n)s_y^2]$$

$$V(\bar{y}_{lr}) = [(1 - f/n)\, S_y^2(1 - \rho^2)]$$

Since $-1 < \rho < 1$

$$V(\bar{y}_{lr}) = [(1 - f/n) S_y^2 (1 - \rho^2)] < [(1 - f/n) s_y^2] = V(\bar{y})_{SRSWOR}$$

i.e., $V(\bar{y}_{lr}) < V(\bar{y})_{SRSWOR}$ which is always true. So the regression estimator is always **better** than the sample mean under SRSWOR.

Equality holds if $\rho = 0$, i.e., there is no association between Y and X.

Comparison between regression estimator and ratio estimator:

We know,

$$V(\bar{y}_{lr}) = [(1 - f/n) S_y^2 (1 - \rho^2)]$$

$$V(\bar{y}_R) = (1-n/N)S_r^2/n = (1-f/n)[S_y^2 - 2RS_{yx} + R^2 S_x^2]$$

So, $V(\bar{y}_R) - V(\bar{y}_{lr}) = (1 - f/n)[S_y^2 - 2RS_{yx} + R^2 S_x^2 - S_y^2 (1 - \rho^2)]$

$$= (1 - f/n)[R^2 S_x^2 + S_y^2 \rho^2 - 2R\rho S_{yx}]$$

$$= (1 - f/n)[RS_x - \rho S_y]^2$$

$$= (1 - f/n) S_x^2 [R - \beta]^2 \geq 0$$

Where $\beta = S_{yx}/S_x^2 = \rho(S_y/S_x)$

$\Rightarrow V(\bar{y}_R) - V(\bar{y}_{lr}) \geq 0$

i.e., $V(\bar{y}_{lr}) \leq V(\bar{y}_R)$

So, regression estimate is always **better** than the ratio estimate.

Both are equally efficient if equality holds i.e., when $R = \beta$.