Chapter 6

Regression Method of Estimation

The ratio method of estimation uses the auxiliary information which is correlated with the study variable to improve the precision which results in the improved estimators when the regression of Y on X is linear and passes through the origin. When the regression of Y on X is linear, it is not necessary that the line should always pass through the origin. Under such conditions, it is more appropriate to use the regression type estimator to estimate the population means.

In ratio method, the conventional estimator sample mean $\bar{y}$ was improved by multiplying it by a factor $\frac{\bar{X}}{\bar{x}}$ where $\bar{x}$ is an unbiased estimator of the population mean $\bar{X}$ which is chosen as the population mean of auxiliary variable. Now we consider another idea based on difference.

Consider an estimator $\bar{x}$ of $\bar{X}$ for which $E(\bar{x}-\bar{X})=0.$

Consider an improved estimator of $\bar{Y}$ as

$$\hat{\bar{Y}}^{*}=\bar{y}+\mu(\bar{x}-\bar{X})$$

which is an unbiased estimator of $\bar{Y}$ and μ is any constant. Now find μ such that the $Var(\hat{\bar{Y}}^{*})$ is minimum

$$Var(\hat{\bar{Y}}^{*})=Var(\bar{y})+\mu^{2}Var(\bar{x})+2\mu Cov(\bar{x},\bar{y})$$

$$\frac{\partial Var(\bar{Y}^{*})}{\partial\mu}=0$$

$$\Rightarrow \mu=-\frac{Cov(\bar{x},\bar{y})}{Var(\bar{x})}$$

$$=-\frac{\frac{N-n}{Nn}S_{XY}}{\frac{N-n}{Nn}S_{X}^{2}}$$

$$=-\frac{S_{XY}}{S_{X}^{2}}$$

where $S_{XY}=\frac{1}{N-1}\sum_{i=1}^{N}(X_i-\bar{X})(Y_i-\bar{Y}),\quad S_{X}^{2}=\frac{1}{N-1}\sum_{i=1}^{N}(X_i-\bar{X}).$

Consider a linear regression model $y=x\beta+e$ where y is the dependent variable, x is the independent variable and e is the random error component which takes care of the difference arising due to lack of exact relationship between x and y.

Note that the value of regression coefficient β in a linear regression model $y = x\beta + e$ of y on x obtained by minimizing $\sum_{i=1}^{n} e_i^2$ based on *n* data sets $(x_i, y_i),\ i = 1,2,..,n$ is $\beta = \frac{Cov(x,y)}{Var(x)} = \frac{S_{xy}}{S_x^2}$. Thus the optimum value of μ is same as the regression coefficient of y on x with a negative sign, *i.e.*,

$$\mu = -\beta .$$

So the estimator $\hat{\bar{Y}}^*$ with optimum value of μ is

$$\hat{\bar{Y}}_{reg} = \bar{y} + \beta(\bar{X} - \bar{x})$$

which is the regression estimator of $\bar{Y}$ and the procedure of estimation is called as the regression method of estimation.

The variance of $\hat{\bar{Y}}_{reg}$ *is*

$$Var(\hat{\bar{Y}}_{reg}) = V(\bar{y})[1 - \rho^2(\bar{x}, \bar{y})]$$

where $\rho(\bar{x}, \bar{y})$ is the correlation coefficient between $\bar{x}$ and $\bar{y}$. So $\hat{\bar{Y}}_{reg}$ would be efficient if $\bar{x}$ and $\bar{y}$ are highly correlated. The estimator $\hat{\bar{Y}}_{reg}$ is more efficient than $\bar{Y}$ if $\rho(\bar{x}, \bar{y}) \neq 0$ which generally holds.

Regression estimates with preassigned β:

If value of β is known as β_0 (say), then the regression estimator is

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

Bias of $\hat{\bar{Y}}_{reg}$:

Now, assuming that the random sample $(x_i, y_i),\ i = 1,2,..,n$ is drawn by SRSWOR,

$$\begin{aligned} E(\hat{\bar{Y}}_{reg}) &= E(\bar{y}) + \beta_0\left[\bar{X} - E(\bar{x})\right] \\ &= \bar{Y} + \beta_0\left[\bar{X} - \bar{X}\right] \\ &= \bar{Y} \end{aligned}$$

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

Variance of $\hat{\bar{Y}}_{reg}$

$$
\begin{aligned}
Var(\hat{\bar{Y}}_{reg}) &= E\left[\hat{\bar{Y}}_{reg} - E(\hat{\bar{Y}}_{reg})\right]^2 \\
&= E\left[\bar{y} + \beta_0(\bar{X} - \bar{x}) - \bar{Y}\right]^2 \\
&= E\left[(\bar{y} - \bar{Y}) - \beta_0(\bar{x} - \bar{X})\right]^2 \\
&= E\left[(\bar{y} - \bar{Y})^2 + \beta_0^2(\bar{x} - \bar{X})^2 - 2\beta_0 E(\bar{x} - \bar{X})(\bar{y} - \bar{Y})\right] \\
&= Var(\bar{y}) + \beta_0^2 Var(\bar{x}) - 2\beta_0 Cov(\bar{x}, \bar{y}) \\
&= \frac{f}{n}\left[S_Y^2 + \beta_0^2 S_X^2 - 2\beta_0 S_{XY}\right] \\
&= \frac{f}{n}\left[S_Y^2 + \beta_0^2 S_X^2 - 2\beta_0 \rho S_X S_Y\right]
\end{aligned}
$$

where

$$f = \frac{N-n}{N}$$

$$S_X^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

$$S_Y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$$

ρ : Correlation coefficient between X and Y.

Comparing $Var(\hat{\bar{Y}}_{reg})$ with $Var(\bar{y})$, we note that

$$Var(\hat{\bar{Y}}_{reg}) < Var(\bar{y})$$

if $\beta_0^2 S_X^2 - 2\beta_0 S_{XY} < 0$

or $\beta_0 S_X^2\left(\beta_0 - \frac{2S_{XY}}{S_X^2}\right) < 0$

which is possible when

either $\beta_0 < 0$ and $\left(\beta_0 - \frac{2S_{XY}}{S_X^2}\right) > 0 \Rightarrow \frac{2S_{XY}}{S_X^2} < \beta_0 < 0$.

or $\beta_0 > 0$ and $\left(\beta_0 - \frac{2S_{XY}}{S_X^2}\right) < 0 \Rightarrow 0 < \beta_0 < \frac{2S_{XY}}{S_X^2}$.

Optimal value of β

Choose β such that $Var(\hat{\bar{Y}}_{reg})$ is minimum.

So

$$\frac{\partial Var(\hat{\bar{Y}}_{reg})}{\partial \beta} = \frac{\partial}{\partial \beta}\left[S_Y^2 + \beta^2 S_X^2 - 2\beta\rho S_X S_Y \right] = 0$$

$$\Rightarrow \beta = \rho \frac{S_Y}{S_X} = \frac{S_{XY}}{S_X^2}.$$

The minimum value of the variance of $\hat{\bar{Y}}_{reg}$ with optimum value of $\beta_{opt} = \frac{\rho S_Y}{S_X}$ is

$$Var_{min}(\hat{\bar{Y}}_{reg}) = \frac{f}{n}\left[S_Y^2 + \rho^2 \frac{S_Y^2}{S_X^2} S_X^2 - 2\rho \frac{S_Y}{S_X} \rho S_X S_Y \right]$$

$$= \frac{f}{n} S_Y^2 (1-\rho^2).$$

Since $-1 \le \rho \le 1,$ so

$$Var(\hat{\bar{Y}}_{reg}) \le Var_{SRS}(\bar{y})$$

which always holds true. So the regression estimator is always better than the sample mean under SRSWOR.

Departure from β:

If β_0 is the preassigned value of regression coefficient, then

$$Var_{\min}(\hat{\bar{Y}}_{reg}) = \frac{f}{n}\left[S_Y^2 + \beta_0^2 S_X^2 - 2\beta_0 \rho S_X S_Y \right]$$

$$= \frac{f}{n}\left[S_Y^2 + \beta_0^2 S_X^2 - 2\rho\beta_0 S_X S_Y - \rho^2 S_Y^2 + \rho^2 S_Y^2 \right]$$

$$= \frac{f}{n}\left[(1-\rho^2) S_Y^2 + \beta_0^2 S_X^2 - 2\beta_0 S_X^2 \beta_{opt} + \beta_{opt}^2 S_X^2 \right]$$

$$= \frac{f}{n}\left[(1-\rho^2) S_Y^2 + (\beta_0 - \beta_{opt})^2 S_X^2 \right]$$

where $\beta_{opt} = \frac{\rho S_Y}{S_X}$.

Estimate of variance

An unbiased sample estimate of $Var(\hat{\bar{Y}}_{reg})$ is

$$Var(\hat{\bar{Y}}_{reg}) = \frac{f}{n(n-1)}\sum_{i=1}^{n}\left[(y_i - \bar{y}) - \beta_0(x_i - \bar{x})\right]^2$$
$$= \frac{f}{n}(s_y^2 + \beta_0^2 s_x^2 - 2\beta_0 s_{xy}).$$

Note that the variance of $\hat{\bar{Y}}_{reg}$ increases as the difference between β_0 and β_{opt} increases.

Regression estimates when β is computed from the sample

Suppose a random sample of size *n* on paired observations on (x_i, y_i), $i = 1,2,..,n$ is drawn by SRSWOR.

When β is unknown, it is estimated as

$$\hat{\beta} = \frac{s_{xy}}{s_x^2}$$

and then the regression estimator of $\bar{Y}$ is given by

$$\hat{\bar{Y}}_{reg} = \bar{y} + \hat{\beta}(\bar{X} - \bar{x}).$$

It is difficult to find the exact expressions of $E(\bar{Y}_{reg})$ and $Var(\hat{\bar{Y}}_{reg})$. So we approximate them using the same methodology as in the case of the ratio method of estimation.

Let

$$\varepsilon_0 = \frac{\bar{y} - \bar{Y}}{\bar{Y}} \Rightarrow \bar{y} = \bar{Y}(1 + \varepsilon_0)$$
$$\varepsilon_1 = \frac{\bar{x} - \bar{X}}{\bar{x}} \Rightarrow \bar{x} = \bar{X}(1 + \varepsilon_1)$$
$$\varepsilon_2 = \frac{s_{xy} - S_{XY}}{S_{XY}} \Rightarrow s_{xy} = S_{XY}(1 + \varepsilon_2)$$
$$\varepsilon_3 = \frac{s_x^2 - S_X^2}{S_X^2} \Rightarrow s_x^2 = S_X^2(1 + \varepsilon_3)$$

Then

$$E(\varepsilon_0) = 0, \qquad E(\varepsilon_1) = 0,$$
$$E(\varepsilon_2) = 0, \qquad E(\varepsilon_3) = 0,$$
$$E(\varepsilon_0^2) = \frac{f}{n}C_Y^2,$$
$$E(\varepsilon_1^2) = \frac{f}{n}C_X^2,$$
$$E(\varepsilon_0\varepsilon_1) = \frac{f}{n}\rho C_X C_Y$$

and

$$\bar{Y}_{reg} = \bar{y} + \frac{s_{xy}}{s_x^2}(\bar{X} - \bar{x})$$
$$= \bar{Y}(1+\varepsilon_0) + \frac{S_{XY}(1+\varepsilon_2)}{S_x^2(1+\varepsilon_3)}(-\varepsilon_1 \bar{X}).$$

The estimation error of $\hat{\bar{Y}}_{reg}$ is

$$(\hat{\bar{Y}}_{reg} - \bar{Y}) = \bar{Y}\varepsilon_0 - \beta\bar{X}\varepsilon_1(1+\varepsilon_2)(1+\varepsilon_3)^{-1}$$

where $\beta = \frac{S_{XY}}{S_X^2}$ is the population regression coefficient.

Assuming $|\varepsilon_3| < 1$,

$$(\hat{\bar{Y}}_{reg} - \bar{Y}) = \bar{Y}\varepsilon_0 - \beta\bar{X}(\varepsilon_1 + \varepsilon_1\varepsilon_2)(1 - \varepsilon_3 + \varepsilon_3^2 - ....)$$

Retaining the terms up to second power of $\varepsilon's$ and ignoring other terms, we have

$$(\hat{\bar{Y}}_{reg} - \bar{Y}) \simeq \bar{Y}\varepsilon_0 - \beta\bar{X}(\varepsilon_1 + \varepsilon_1\varepsilon_2)(1 - \varepsilon_3 + \varepsilon_3^2)$$
$$\simeq \bar{Y}\varepsilon_0 - \beta\bar{X}(\varepsilon_1 - \varepsilon_1\varepsilon_3 + \varepsilon_1\varepsilon_2)$$

Bias of $\hat{\bar{Y}}_{reg}$

Now the bias of $\hat{\bar{Y}}_{reg}$ up to the second order of approximation is

$$E(\hat{\bar{Y}}_{reg} - \bar{Y}) \simeq E\left[\bar{Y}\varepsilon_0 - \beta\bar{X}(\varepsilon_1 + \varepsilon_1\varepsilon_2)(1 - \varepsilon_3 + \varepsilon_3^2)\right]$$
$$= -\frac{\beta\bar{X}f}{n}\left[\frac{\mu_{21}}{\bar{X}S_{XY}} - \frac{\mu_{30}}{\bar{X}S_X^2}\right]$$

where $f = \frac{N-n}{N}$ and $(r, s)^{\text{th}}$ cross-product moment is given by

$$\mu_{rs} = E\left[(x - \bar{X})^r (y - \bar{Y})^s\right]$$

So that

$$\mu_{21} = E\left[(x - \bar{X})^2 (y - \bar{Y})\right]$$
$$\mu_{30} = E\left[(x - \bar{X})^3\right].$$

Thus

$$E(\hat{\bar{Y}}_{reg}) = -\frac{\beta f}{n}\left[\frac{\mu_{21}}{S_{XY}} - \frac{\mu_{30}}{S_X^2}\right].$$

Also,

$$
\begin{aligned}
E(\hat{\bar{Y}}_{reg}) &= E(\bar{y}) + E[\hat{\beta}(\bar{X} - \bar{x})] \\
&= \bar{Y} + \bar{X}E(\hat{\beta}) - E(\hat{\beta}\bar{x}) \\
&= \bar{Y} + E(\bar{x})E(\hat{\beta}) - E(\hat{\beta}\bar{x}) \\
&= \bar{Y} - Cov(\hat{\beta}, \bar{x}) \\
Bias(\hat{\bar{Y}}_{reg}) &= E(\hat{\bar{Y}}_{reg}) - \bar{Y} = -Cov(\hat{\beta}, \bar{x})
\end{aligned}
$$

MSE of $\hat{\bar{Y}}_{reg}$

To obtain the MSE of $\hat{\bar{Y}}_{reg}$, consider

$$E(\hat{\bar{Y}}_{reg} - \bar{Y})^2 \approx E\left[\varepsilon_0\bar{Y} - \beta\bar{X}(\varepsilon_1 - \varepsilon_1\varepsilon_3 + \varepsilon_1\varepsilon_2)\right]^2$$

Retaining the terms of $\varepsilon's$ up to the second power second and ignoring others, we have

$$
\begin{aligned}
E(\hat{\bar{Y}}_{reg} - \bar{Y})^2 &\approx E\left[\varepsilon_0^2\bar{Y}^2 + \beta^2\bar{X}^2\varepsilon_1^2 - 2\beta\bar{X}\bar{Y}\varepsilon_0\varepsilon_1\right] \\
&= \bar{Y}^2E(\varepsilon_0^2) + \beta^2\bar{X}^2E(\varepsilon_1^2) - 2\beta\bar{X}\bar{Y}E(\varepsilon_0\varepsilon_1) \\
&= \frac{f}{n}\left[\bar{Y}^2\frac{S_Y^2}{\bar{Y}^2} + \beta^2\bar{X}^2\frac{S_X^2}{\bar{X}^2} - 2\beta\bar{X}\bar{Y}\rho\frac{S_XS_Y}{\bar{X}\bar{Y}}\right] \\
MSE(\hat{\bar{Y}}_{reg}) &= E(\hat{\bar{Y}}_{reg} - \bar{Y})^2 \\
&= \frac{f}{n}(S_Y^2 + \beta^2S_X^2 - 2\beta\rho S_XS_Y)
\end{aligned}
$$

Since $\beta = \frac{S_{XY}}{S_X^2} = \rho\frac{S_Y}{S_X}$,

so substituting it in MSE $(\hat{\bar{Y}}_{reg})$, we get

$$MSE(\hat{\bar{Y}}_{reg}) = \frac{f}{n}S_Y^2(1-\rho^2).$$

So up to the second order of approximation, the regression estimator is better than the conventional sample mean estimator under SRSWOR. This is because the regression estimator uses some extra information also. Moreover, such extra information requires some extra cost also. This shows a false superiority in some sense. So the regression estimators and SRS estimates can be combined if the cost aspect is also taken into consideration.

Comparison of $\hat{\bar{Y}}_{reg}$ with ratio estimate and SRS sample mean estimate

$$MSE(\hat{\bar{Y}}_{reg}) = \frac{f}{n} S_Y^2 (1-\rho^2)$$

$$MSE(\hat{\bar{Y}}_R) = \frac{f}{n} (S_Y^2 + R^2 S_X^2 - 2\rho R S_X S_Y)$$

$$Var_{SRS}(\bar{y}) = \frac{f}{n} S_Y^2.$$

(i) As $MSE(\hat{\bar{Y}}_{reg}) = Var_{SRS}(\bar{y})(1-\rho^2)$ and because $\rho^2 < 1$, so $\hat{\bar{Y}}_{reg}$ is always superior to $\bar{y}$.

(ii) $\hat{\bar{Y}}_{reg}$ is better than $\hat{\bar{Y}}_R$ if $MSE(\hat{\bar{Y}}_{reg}) \le MSE(\hat{\bar{Y}}_R)$

or if $\frac{f}{n} S_Y^2 (1-\rho^2) \le \frac{f}{n} (S_Y^2 + R^2 S_X^2 - 2\rho R S_X S_Y)$

or if $(RS_X - \rho S_Y)^2 \ge 0$

which always holds true.

So regression estimate is always superior to the ratio estimate upto the second order of approximation.

Regression estimates in stratified sampling

Under the set up of stratified sampling, let the population of N sampling units be divided into k strata. The strata sizes are $N_1, N_2, .., N_k$ such that $\sum_{i=1}^{k} N_i = N$. A sample of size n_i on (x_{ij}, y_{ij}), $j = 1, 2, .., n_i$, is drawn from i^{th} strata (i = 1,2,..,k) by SRSWOR where x_{ij} and y_{ij} denote the j^{th} unit from i^{th} strata on auxiliary and study variables, respectively.

In order to estimate the population mean, there are two approaches.

1. Separate regression estimator

- Estimate regression estimator

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

from each stratum separately, i.e., the regression estimate in the i^{th} stratum is

$$\hat{\bar{Y}}_{reg(i)} = \bar{y}_i + \beta_i (\bar{X}_i - \bar{x}_i).$$

- Find the stratified mean as the weighted mean of $\hat{\bar{Y}}_{reg(i)}$ $i = 1, 2, .., k$ as

$$\hat{\bar{Y}}_{sreg} = \sum_{i=1}^{k} \frac{N_i \hat{\bar{Y}}_{reg(i)}}{N}$$

$$= \sum_{i=1}^{k} [w_i \{\bar{y}_i + \beta_i (\bar{X}_i - \bar{x}_i)\}]$$

where $\beta_i = \frac{S_{ixy}}{S_{ix}^2}$, $w_i = \frac{N_i}{N}$.

In this approach, the regression estimator is separately obtained in each of the strata and then combined using the philosophy of the stratified sample. So $\hat{\bar{Y}}_{sreg}$ is termed as separate regression estimator,

2. Combined regression estimator

Another strategy is to estimate $\bar{x}$ and $\bar{y}$ in the $\hat{\bar{Y}}_{reg}$ as respective stratified mean. Replacing $\bar{x}$ by $\bar{x}_{st} = \sum_{i=1}^{k} w_i \bar{x}_i$ and $\bar{y}$ by $\bar{y}_{st} = \sum_{i=1}^{k} w_i \bar{y}_i$, we have

$$\hat{\bar{Y}}_{creg} = \bar{y}_{st} + \beta(\bar{X} - \bar{x}_{st}).$$

In this case, all the sample information is combined first and then implemented in regression estimator, so $\hat{\bar{Y}}_{reg}$ is termed as combined regression estimator.

Properties of separate and combined regression

In order to derive the mean and variance of $\hat{\bar{Y}}_{sreg}$ and $\hat{\bar{Y}}_{creg}$, there are two cases

- when β is pre-assigned as β_0
- when β is estimated from the sample.

We consider here the case that β is pre-assigned as β_0. Other case when β is estimated as $\hat{\beta} = \frac{S_{xy}}{S_x^2}$ can be dealt with the same approach based on defining various $\varepsilon's$ and using the approximation theory as in the case of $\hat{\bar{Y}}_{reg}$.

1. Separate regression estimator

Assume β is known, say β_0. Then

$$\hat{\bar{Y}}_{sreg} = \sum_{i=1}^{k} w_i[\bar{y}_i + \beta_{0i}(\bar{X}_i - \bar{x}_i)]$$

$$E(\hat{\bar{Y}}_{sreg}) = \sum_{i=1}^{k} w_i\left[E(\bar{y}_i) + \beta_{0i}\left(\bar{X}_i - E(\bar{x}_i)\right)\right]$$

$$= \sum_{i=1}^{k} w_i[\bar{Y}_i + (\bar{X}_i - \bar{X}_i)]$$

$$= \bar{Y}.$$

$$Var(\hat{\bar{Y}}_{sreg}) = E\left[\hat{\bar{Y}}_{sreg} - E(\hat{\bar{Y}}_{sreg})\right]^2$$

$$= E\left[\sum_{i=1}^{k} w_i\bar{y}_i +_i + \sum_{i=1}^{k} w_i\beta_{0i}(\bar{X}_i - \bar{x}_i) - \bar{Y}\right]^2$$

$$= E\left[\sum_{i=1}^{k} w_i(\bar{y}_i - \bar{Y}) - \sum_{i=1}^{k} w_i\beta_{0i}(\bar{x}_i - \bar{X}_i)\right]^2$$

$$= \sum_{i=1}^{k} w_i^2 E(\bar{y}_i - \bar{Y}_i)^2 + \sum_{i=1}^{k} w_i^2\ \beta_{0i}^2 E(\bar{x}_i - \bar{X}_i)]^2 - 2\sum_{i=1}^{k} w_i^2\beta_{0i}E(\bar{x}_i - \bar{X}_i)(\bar{y}_i - \bar{Y}_i)$$

$$= \sum_{i=1}^{k} w_i^2 Var(\bar{y}_i) + \sum_{i=1}^{k} w_i^2\beta_{0i}^2 Var(\bar{x}_i) - 2\sum_{i=1}^{k} w_i^2\beta_{0i}Cov(\bar{x}_i, \bar{y}_i)$$

$$= \sum_{i=1}^{k} \frac{w_i^2 f_i}{n_i}(S_{iY}^2 + \beta_{0i}^2 S_{iX}^2 - 2\beta_{0i}S_{iXY})]$$

$Var(\hat{\bar{Y}}_{sreg})$ is minimum when $\beta_{0i} = \dfrac{S_{iXY}}{S_{iX}^2}$ and so substituting β_{0i}, we have

$$V_{\min}(\hat{\bar{Y}}_{sreg}) = \sum_{i=1}^{k}\left[\frac{w_i^2 f_i}{n_i}(S_{iY}^2 - \beta_{0i}^2 S_{iX}^2)\right]$$

where $f_i = \dfrac{N_i - n_i}{N_i}$.

Since SRSWOR is followed in drawing the samples from each stratum, so

$$E(s_{ix}^2) = S_{iX}^2$$

$$E(s_{iy}^2) = S_{iY}^2$$

$$E(s_{ixy}) = S_{iXY}$$

Thus an unbiased estimator of variance can be obtained by replacing S_{iX}^2 and S_{iY}^2 by their respective unbiased estimators s_{ix}^2 and s_{iy}^2, respectively as

$$Var(\hat{\bar{Y}}_{sreg}) = \sum_{i=1}^{k}\left[\frac{w_i^2 f_i}{n_i}(s_{iy}^2 + \beta_{oi}^2 s_{ix}^2 - 2\beta_{0i}s_{ixy})\right]$$

and

$$Var_{\min}(\hat{\bar{Y}}_{sreg}) = \sum_{i=1}^{k}\left[\frac{w_i^2 f_i}{n_i}(s_{iy}^2 - \beta_{oi}^2 s_{ix}^2)\right].$$

2. Combined regression estimator:

Assume β is known as β_0. Then

$$\hat{\bar{Y}}_{creg} = \sum_{i=1}^{k} w_i \bar{y}_i + \beta_0 (\bar{X} - \sum_{i=1}^{k} w_i \bar{x}_i)$$

$$E\left(\hat{\bar{Y}}_{creg}\right) = \sum_{i=1}^{k} w_i E(\bar{y}_i) + \beta_0 [\bar{X} - \sum_{i=1}^{k} w_i E(\bar{x}_i)]$$

$$= \sum_{i=1}^{k} w_i \bar{Y}_i + \beta_0 [\bar{X} - \sum_{i=1}^{k} w_i \bar{X}_i]$$

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

$$= \bar{Y}.$$

Thus $\hat{\bar{Y}}_{creg}$ is an unbiased estimator of $\bar{Y}$.

$$Var(\hat{\bar{Y}}_{creg}) = E[\bar{Y}_{creg} - E(\bar{Y}_{creg})]^2$$

$$= E[\sum_{i=1}^{k} w_i \bar{y}_i + \beta_0 (\bar{X} - \sum_{i=1}^{k} w_i \bar{x}_i) - \bar{Y}]^2$$

$$= E[\sum_{i=1}^{k} w_i (\bar{y}_i - \bar{Y}) - \beta_0 \sum_{i=1}^{k} w_i (\bar{x}_i - \bar{X}_i)]^2$$

$$= \sum_{i=1}^{k} w_i^2 Var(\bar{y}_i) + \beta_0^2 \sum_{i=1}^{k} w_i^2 Var(\bar{x}_i) - 2\sum_{i=1}^{k} w_i^2 \beta_0 Cov(\bar{x}_i, \bar{y}_i)$$

$$= \sum_{i=1}^{k} \frac{w_i^2 f_i}{n_i} \left[S_{iY}^2 + \beta_0^2 S_{iX}^2 - 2\beta_0 S_{iXY} \right].$$

$Var(\hat{\bar{Y}}_{creg})$ is minimum when

$$\beta_0 = \frac{Cov(\bar{x}_{st}, \bar{y}_{st})}{Var(\bar{x}_{st})}$$

$$= \frac{\sum_{i=1}^{k} \frac{w_i^2 f_i}{n_i} S_{iXY}}{\sum_{i=1}^{k} \frac{w_i^2 f_i}{n_i} S_{iX}^2}$$

and the minimum variance is given by

$$Var_{\min}(\hat{\bar{Y}}_{creg}) = \sum_{i=1}^{k} \frac{w_i^2 f_i}{n_i} (S_{iY}^2 - \beta_0^2 S_{iX}^2).$$

Since SRSWOR is followed to draw the sample from strata, so using $E\left(s_{ix}^2\right) = S_{ix}^2$, $E\left(s_{iy}^2\right) = S_{iy}^2$ and $E\left(s_{ixy}\right) = S_{iXY}$, we get the estimate of variance as

$$Var(\hat{\bar{Y}}_{creg}) = \sum_{i=1}^{k} \left[\frac{w_i^2 f_i}{n_i} (s_{iy}^2 + \beta_o^2 s_{ix}^2 - 2\beta_{0i} s_{ixy}) \right]$$

and

$$Var_{\min}(\hat{\bar{Y}}_{creg}) = \sum_{i=1}^{k} \left[\frac{w_i^2 f_i}{n_i} (s_{iy}^2 - \beta_{oi}^2 s_{ix}^2) \right] .$$

Comparison of $\hat{\bar{Y}}_{sreg}$ and $\hat{\bar{Y}}_{creg}$:

The variance of $\hat{\bar{Y}}_{sreg}$ is minimum when $\beta_{0i} = \beta_0$ for all *i*.

The variance of $\hat{\bar{Y}}_{creg}$ is minimum when $\beta_0 = \frac{Cov(\bar{x}_{st}, \bar{y}_{st})}{Var(\bar{x}_{st})} = \beta_0^*$.

The minimum variance is $Var(\hat{\bar{Y}}_{creg})_{\min} = Var(\bar{y}_{st})(1-\rho_*^2)$ where $\rho_* = \frac{Cov(\bar{x}_{st}, \bar{y}_{st})}{\sqrt{Var(\bar{x}_{st})Var(\bar{y}_{st})}}$.

$$Var(\hat{\bar{Y}}_{creg}) - Var(\hat{\bar{Y}}_{sreg}) = \sum_{i=1}^{k} (\beta_{0i}^2 - \beta_0^2) \frac{w_i^2 f_i}{n_i} S_{iX}^2$$

$$Var(\hat{\bar{Y}}_{creg})_{\min} - Var(\hat{\bar{Y}}_{sreg})_{\beta_{0i}=\beta_0} = \sum_{i=1}^{k} \frac{f_i}{n_i} (\beta_{0i} - \beta_0)^2 w_i^2 S_{iX}^2$$

$$\geq 0$$

which is always true.

So if the regression line of *y* on *x* is approximately linear and the regression coefficients do not vary much among the strata, then separate regression estimate is more efficient than combined regression estimator.