Chapter 8

Double Sampling (Two Phase Sampling)

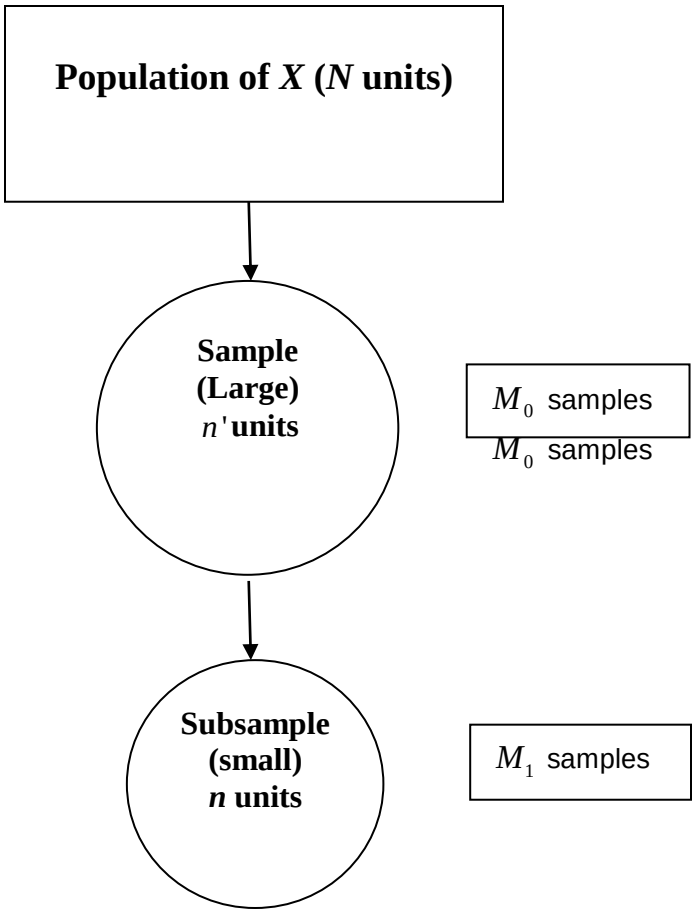
The ratio and regression methods of estimation require the knowledge of population mean of an auxiliary variable $(\bar{X})$ to estimate the population mean of study variable $(\bar{Y})$. If the information on the auxiliary variable is not available, then there are two options – one option is to collect a sample only on study variable, and use sample mean as an estimator of the population mean.

An alternative solution is to use a part of the budget for collecting information on an auxiliary variable to collect a large preliminary sample in which x_i alone is measured. The purpose of this sampling is to furnish a good estimate of $\bar{X}$. This method is appropriate when the information about x_i is on file cards that have not been tabulated. After collecting a large preliminary sample of size n' units from the population, select a smaller sample of size n from it and collect the information on y. These two estimates are then used to obtain an estimator of the population mean $\bar{Y}$. This procedure of selecting a large sample for collecting information on auxiliary variable x and then selecting a sub-sample from it for collecting the information on the study variable *y* is called double sampling or two-phase sampling. It is useful when it is considerably cheaper and quicker to collect data on x than y and there is a high correlation between x and y.

In this sampling, the randomization is done twice. First, a random sample of size n' is drawn from a population of size N and then again a random sample of size n is drawn from the first sample of size n'.

So the sample mean in this sampling is a function of the two phases of sampling. If SRSWOR is utilized to draw the samples at both the phases, then the

- number of possible samples at the first phase when a sample of size n is drawn from a population of size N is $\binom{N}{n'} = M_0$, say.
- number of possible samples at the second phase where a sample of size n is drawn from the first phase sample of size n' is $\binom{n'}{n} = M_1$, say.



Then the sample mean is a function of two variables. If τ is the statistic calculated at the second phase such that $\tau_{ij}, i = 1,2,...,M_0,\ j = 1,2,...,M_1$ with P_{ij} being the probability that i^{th} sample is chosen at the first phase and j^{th} sample is chosen at the second phase, then

$$E(\tau) = E_1\left[E_2(\tau)\right]$$

where $E_2(\tau)$ denotes the expectation over the second phase and E_1 denotes the expectation over the first phase. Thus

$$E(\tau) = \sum_{i=1}^{M_0}\sum_{j=1}^{M_1} P_{ij}\tau_{ij}$$

$$= \sum_{i=1}^{M_0}\sum_{j=1}^{M_1} P_i P_{j/i}\,\tau_{ij} \quad \text{(using } P(A\cap B) = P(A)P(B/A)\text{)}$$

$$= \underbrace{\sum_{i=1}^{M_0} P_i}_{1^{st}\text{ stage}}\ \underbrace{\sum_{j=1}^{M_1} P_{j/i}\,\tau_{ij}}_{2^{nd}\text{ stage}}$$

Variance of τ

$$
\begin{aligned}
Var(\tau) &= E\left[\tau - E(\tau)\right]^2 \\
&= E\left[(\tau - E_2(\tau)) + (E_2(\tau) - E(\tau))\right]^2 \\
&= E\left[\tau - E_2(\tau)\right]^2 + \left[E_2(\tau) - E(\tau)\right]^2 + 0 \\
&= E_1E_2\left[\tau - E_2(\tau)\right]^2 + \left[E_2(\tau) - E(\tau)\right]^2 \\
&= E_1E_2\left[\tau - E_2(\tau)\right]^2 + E_1E_2\left[E_2(\tau) - E(\tau)\right]^2 \\
&\qquad\qquad\qquad\qquad \downarrow \\
&\qquad\qquad\qquad\qquad \text{constant for } E_2 \\
&= E_1\left[V_2(\tau)\right] + E_1\left[E_2(\tau) - E_1(E_2(\tau))\right]^2 \\
&= E_1\left[V_2(\tau)\right] + V_1\left[E_2(\tau)\right]
\end{aligned}
$$

Note: The two-phase sampling can be extended to more than two phases depending upon the need and objective of the experiment. Various expectations can also be extended on similar lines.

Double sampling in ratio method of estimation

If the population mean $\bar{X}$ is not known, then the double sampling technique is applied. Take a large initial sample of size n' by SRSWOR to estimate the population mean $\bar{X}$ as

$$\hat{\bar{X}} = \bar{x}' = \frac{1}{n'}\sum_{i=1}^{n'} x_i \, .$$

Then a second sample is a subsample of size n selected from the initial sample by SRSWOR. Let $\bar{y}$ and $\bar{x}$ be the means of y and x based on the subsample. Then $E(\bar{x}') = \bar{X}$, $E(\bar{x}) = \bar{X}$, $E(\bar{y}) = \bar{Y}$.

The ratio estimator under double sampling now becomes

$$\hat{\bar{Y}}_{Rd} = \frac{\bar{y}}{\bar{x}}\bar{x}' .$$

The exact expressions for the bias and mean squared error of $\hat{\bar{Y}}_{Rd}$ are difficult to derive. So we find their approximate expressions using the same approach mentioned while describing the ratio method of estimation.

Let

$$\varepsilon_0 = \frac{\bar{y} - \bar{Y}}{\bar{Y}}, \quad \varepsilon_1 = \frac{\bar{x} - \bar{X}}{\bar{X}}, \quad \varepsilon_2 = \frac{\bar{x}' - \bar{X}}{\bar{X}}$$

$$E(\varepsilon_0) = E(\varepsilon_1) = E(\varepsilon_2) = 0$$

$$E(\varepsilon_1^2) = \left(\frac{1}{n} - \frac{1}{N}\right) C_x^2$$

$$E(\varepsilon_1 \varepsilon_2) = \frac{1}{\bar{X}^2} E(\bar{x} - \bar{X})(\bar{x}' - \bar{X})$$

$$= \frac{1}{\bar{X}^2} E_1\left[E_2(\bar{x} - \bar{X})(\bar{x}' - \bar{X}) \mid n'\right]$$

$$= \frac{1}{\bar{X}^2} E_1\left[(\bar{x}' - \bar{X})^2\right]$$

$$= \left(\frac{1}{n'} - \frac{1}{N}\right) \frac{S_x^2}{\bar{X}^2}$$

$$= \left(\frac{1}{n'} - \frac{1}{N}\right) C_x^2$$

$$= E(\varepsilon_2^2).$$

$$E(\varepsilon_0 \varepsilon_2) = \frac{1}{\bar{X}\bar{Y}} Cov(\bar{y}, \bar{x}')$$

$$= \frac{1}{\bar{X}\bar{Y}} Cov\left[E(\bar{y} \mid n'), E(\bar{x}' \mid n')\right] + \frac{1}{\bar{X}\bar{Y}} E\left[Cov(\bar{y}, \bar{x}') \mid n'\right]$$

$$= \frac{1}{\bar{X}\bar{Y}} Cov\left[\bar{Y}, \bar{X}\right] + \frac{1}{\bar{X}\bar{Y}} E\left[Cov(\bar{y}', \bar{x}')\right]$$

$$= \frac{1}{\bar{X}\bar{Y}} Cov\left[(\bar{y}', \bar{x}'\right]$$

$$= \left(\frac{1}{n'} - \frac{1}{N}\right) \frac{S_{xy}}{\bar{X}\bar{Y}}$$

$$= \left(\frac{1}{n'} - \frac{1}{N}\right) \rho \frac{S_x}{\bar{X}} \frac{S_y}{\bar{Y}}$$

$$= \left(\frac{1}{n'} - \frac{1}{N}\right) \rho C_x C_y$$

where $\bar{y}'$ is the sample mean of $y's$ based on the sample size n'.

$$E(\varepsilon_0\varepsilon_1) = \frac{1}{\bar{X}\bar{Y}} Cov(\bar{y}, \bar{x})$$
$$= \left(\frac{1}{n} - \frac{1}{N}\right)\frac{S_{xy}}{\bar{X}\bar{Y}}$$
$$= \left(\frac{1}{n} - \frac{1}{N}\right)\rho\frac{S_x}{\bar{X}}\frac{S_y}{\bar{Y}}$$
$$= \left(\frac{1}{n} - \frac{1}{N}\right)\rho C_x C_y$$

$$E(\varepsilon_0^2) = \frac{1}{\bar{Y}^2} Var(\bar{y})$$
$$= \frac{1}{\bar{Y}^2}\left[V_1\left\{E_2(\bar{y} \mid n')\right\} + E_1\left\{V_2(\bar{y}_n \mid n')\right\}\right]$$
$$= \frac{1}{\bar{Y}^2}\left[V_1(\bar{y}_n') + E_1\left\{\left(\frac{1}{n} - \frac{1}{n'}\right)s_y'^2\right\}\right]$$
$$= \frac{1}{\bar{Y}^2}\left[\left(\frac{1}{n'} - \frac{1}{N}\right)S_y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right)S_y^2\right]$$
$$= \left(\frac{1}{n} - \frac{1}{N}\right)\frac{S_y^2}{\bar{Y}^2}$$
$$= \left(\frac{1}{n} - \frac{1}{N}\right)C_y^2$$

where $s_y'^2$ is the mean sum of squares of y based on an initial sample of size n'.

$$E(\varepsilon_1\varepsilon_2) = \frac{1}{\bar{X}^2} Cov(\bar{x}, \bar{x}')$$
$$= \frac{1}{\bar{X}^2}\left[Cov\left\{E(\bar{x} \mid n'), E(\bar{x}' \mid n')\right\} + 0\right]$$
$$= \frac{1}{\bar{X}^2} Var(\bar{X}')$$

where $Var(\bar{X}')$ is the variance of mean of x based on an initial sample of size n'.

Estimation error of $\hat{\bar{Y}}_{Rd}$

Write $\hat{\bar{Y}}_{Rd}$ *as*

$$\hat{\bar{Y}}_{Rd} = \frac{(1+\varepsilon_0)}{(1+\varepsilon_1)}(1+\varepsilon_2)\frac{\bar{Y}}{\bar{X}}\bar{X}$$
$$= \bar{Y}(1+\varepsilon_0)(1+\varepsilon_2)(1+\varepsilon_1)^{-1}$$
$$= \bar{Y}(1+\varepsilon_0)(1+\varepsilon_2)(1-\varepsilon_1+\varepsilon_1^2-\ldots)$$
$$\simeq \bar{Y}(1+\varepsilon_0+\varepsilon_2+\varepsilon_0\varepsilon_2-\varepsilon_1-\varepsilon_o\varepsilon_1-\varepsilon_1\varepsilon_2+\varepsilon_1^2)$$

up to the terms of order two. Other terms of degree higher than two are assumed to be negligible.

Bias of $\bar{Y}_{Rd}$

$$E(\hat{\bar{Y}}_{Rd}) = \bar{Y}\left[1+0+0+E(\varepsilon_0\varepsilon_2)-0-E(\varepsilon_0\varepsilon_1)-E(\varepsilon_1\varepsilon_2)+E(\varepsilon_1^2)\right]$$

$$Bias(\hat{\bar{Y}}_{Rd}) = E(\hat{\bar{Y}}_{Rd}) - \bar{Y}$$

$$= \bar{Y}\left[E(\varepsilon_0\varepsilon_2)-E(\varepsilon_0\varepsilon_1)-E(\varepsilon_1\varepsilon_2)+E(\varepsilon_1^2)\right]$$

$$= \bar{Y}\left[\left(\frac{1}{n'}-\frac{1}{N}\right)\rho C_x C_y - \left(\frac{1}{n}-\frac{1}{N}\right)\rho C_x C_y - \left(\frac{1}{n'}-\frac{1}{N}\right)C_x^2 + \left(\frac{1}{n}-\frac{1}{N}\right)C_x^2\right]$$

$$= \bar{Y}\left(\frac{1}{n}-\frac{1}{n'}\right)\left(C_x^2 - \rho C_x C_y\right)$$

$$= \bar{Y}\left(\frac{1}{n}-\frac{1}{n'}\right)C_x(C_x - \rho C_y).$$

The bias is negligible if n is large and relative bias vanishes if $C_x^2 = C_{xy}$, i.e., the regression line passes through the origin.

MSE of $\hat{\bar{Y}}_{Rd}$:

$$MSE(\hat{\bar{Y}}_{Rd}) = E(\hat{\bar{Y}}_{Rd} - \bar{Y})^{2`}$$

$$\simeq \bar{Y}^2 E(\varepsilon_0 + \varepsilon_2 - \varepsilon_1)^2 \quad \text{(retaining the terms upto order two)}$$

$$= \bar{Y}^2 E\left[\varepsilon_0^2 + \varepsilon_1^2 + \varepsilon_2^2 + 2\varepsilon_0\varepsilon_2 - 2\varepsilon_0\varepsilon_1 - 2\varepsilon_1\varepsilon_2\right]$$

$$= \bar{Y}^2 E\left[\varepsilon_0^2 + \varepsilon_1^2 + \varepsilon_2^2 + 2\varepsilon_0\varepsilon_2 - 2\varepsilon_0\varepsilon_1 - 2\varepsilon_2^2\right]$$

$$= \bar{Y}^2\left[\left(\frac{1}{n}-\frac{1}{N}\right)C_y^2 + \left(\frac{1}{n}-\frac{1}{N}\right)C_x^2 - \left(\frac{1}{n'}-\frac{1}{N}\right)C_x^2 + 2\left(\frac{1}{n'}-\frac{1}{N}\right)\rho C_x C_y - 2\left(\frac{1}{n}-\frac{1}{N}\right)\rho C_x C_y\right]$$

$$= \bar{Y}^2\left(\frac{1}{n}-\frac{1}{N}\right)\left(C_x^2 + C_y^2 - 2\rho C_x C_y\right) + \bar{Y}^2\left(\frac{1}{n'}-\frac{1}{N}\right)C_x(2\rho C_y - C_x)$$

$$= MSE\text{(ratio estimator)} \quad + \bar{Y}^2\left(\frac{1}{n'}-\frac{1}{N}\right)\left(2\rho C_x C_y - C_x^2\right).$$

The second term is the contribution of the second phase of sampling. This method is preferred over the ratio method if

$$2\rho C_x C_y - C_x^2 < 0$$

or $\rho < \frac{1}{2}\frac{C_x}{C_y}$

Choice of n and n'

Write

$$MSE(\hat{\bar{Y}}_{Rd}) = \frac{V}{n} + \frac{V'}{n'}$$

where V and V' contain all the terms containing n and n' respectively.

The cost function is $C_0 = nC + n'C'$ where C and C' are the costs per unit for selecting the samples n and n' respectively.

Now we find the optimum sample sizes n and n' for fixed cost C_0. The Lagrangian function is

$$\varphi = \frac{V}{n} + \frac{V'}{n'} + \lambda(nC + n'C' - C_0)$$

$$\frac{\partial \varphi}{\partial n} = 0 \Rightarrow \lambda C = \frac{V}{n^2}$$

$$\frac{\partial \varphi}{\partial n'} = 0 \Rightarrow \lambda C' = \frac{V'}{n'^2}.$$

Thus $\lambda C n^2 = V$

or $n = \sqrt{\frac{V}{\lambda C}}$

or $\sqrt{\lambda}\ nC = \sqrt{VC}.$

Similarly $\sqrt{\lambda}\, n'C' = \sqrt{V'C'}.$

Thus

$$\sqrt{\lambda} = \frac{\sqrt{VC} + \sqrt{V'C'}}{C_0}$$

and so

$$\text{Optimum } n = \frac{C_0}{\sqrt{VC} + \sqrt{V'C'}} \sqrt{\frac{V}{C}} = n_{opt}, \text{ say}$$

$$\text{Optimum } n' = \frac{C_0}{\sqrt{VC} + \sqrt{V'C'}} \sqrt{\frac{V'}{C'}} = n'_{opt}, \text{ say}$$

$$Var_{opt}(\hat{\bar{Y}}_{Rd}) = \frac{V}{n_{opt}} + \frac{V'}{n'_{opt}}$$

$$= \frac{(\sqrt{VC} + \sqrt{V'C'})^2}{C_0}$$

Comparison with SRS

If X is ignored and all resources are used to estimate $\bar{Y}$ by $\bar{y}$, then required sample size $= \frac{C_0}{C}$.

$$Var(\bar{y}) = \frac{S_y^2}{C_0 / C} = \frac{CS_y^2}{C_0}$$

$$\text{Relative effiiency} = \frac{Var(\bar{y})}{Var_{opt}(\hat{\bar{Y}}_{Rd})} = \frac{CS_y^2}{(\sqrt{VC} + \sqrt{V'C'})^2}$$

Double sampling in regression method of estimation

When the population mean of the auxiliary variable $\bar{X}$ is not known, then double sampling is used as follows:

- A large sample of size n' is taken from of the population by SRSWOR from which the population mean $\bar{X}$ is estimated as $\bar{x}'$, i.e. $\hat{\bar{X}} = \bar{x}'$.
- Then a subsample of size n is chosen from the larger sample and both the variables x and y are measured from it by taking $\bar{x}'$ in place of $\bar{X}$ and treat it as if it is known.

Then $E(\bar{x}') = \bar{X}$, $E(\bar{x}) = \bar{X}$, $E(\bar{y}) = \bar{Y}$. The regression estimate of $\bar{Y}$ in this case is given by

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

where $\hat{\beta} = \frac{s_{xy}}{s_x^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ is an estimator of $\beta = \frac{S_{xy}}{S_x^2}$ based on the sample of size n.

It is difficult to find the exact properties like bias and mean squared error of $\hat{\bar{Y}}_{regd}$, so we derive the approximate expressions.

Let

$$\varepsilon_0 = \frac{\bar{y} - \bar{Y}}{\bar{Y}} \Rightarrow \bar{y} = (1+\varepsilon_0)\bar{Y}$$

$$\varepsilon_1 = \frac{\bar{x} - \bar{X}}{\bar{X}} \Rightarrow \bar{x} = (1+\varepsilon_1)\bar{X}$$

$$\varepsilon_2 = \frac{\bar{x}' - \bar{X}}{\bar{X}} \Rightarrow \bar{x}' = (1+\varepsilon_2)\bar{X}$$

$$\varepsilon_3 = \frac{s_{xy} - S_{xy}}{S_{xy}} \Rightarrow s_{xy} = (1+\varepsilon_3)S_{xy}$$

$$\varepsilon_4 = \frac{s_x^2 - S_x^2}{S_x^2} \Rightarrow s_x^2 = (1+\varepsilon_4)S_x^2$$

$$E(\varepsilon_1) = 0, E(\varepsilon_2) = 0, E(\varepsilon_3) = 0, E(\varepsilon_4) = 0$$

Define

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

$$\mu_{30} = E\left[\bar{x} - \bar{X}\right]^3$$

Estimation error:

Then

$$\begin{aligned}
\hat{\bar{Y}}_{regd} &= \bar{y} + \hat{\beta}(\bar{x}' - \bar{x}) \\
&= \bar{y} + \frac{S_{xy}(1+\varepsilon_3)}{S_x^2(1+\varepsilon_4)}(\varepsilon_2 - \varepsilon_1)\bar{X} \\
&= \bar{y} + \bar{X}\frac{S_{xy}}{S_x^2}(1+\varepsilon_3)(\varepsilon_2 - \varepsilon_1)(1+\varepsilon_4)^{-1} \\
&= \bar{y} + \bar{X}\beta(1+\varepsilon_3)(\varepsilon_2 - \varepsilon_1)(1 - \varepsilon_4 + \varepsilon_4^2 - ...)
\end{aligned}$$

Retaining the powers of $\varepsilon's$ up to order two assuming $|\varepsilon_3|<1,$ (using the same concept as detailed in the case of ratio method of estimation)

$$\hat{\bar{Y}}_{regd} \simeq \bar{y} + \bar{X}\beta(\varepsilon_2 + \varepsilon_2\varepsilon_3 - \varepsilon_2\varepsilon_4 - \varepsilon_1 - \varepsilon_1\varepsilon_3 + \varepsilon_1\varepsilon_4).$$

Bias:

The bias of $\hat{\bar{Y}}_{regd}$ upto the second order of approximation is

$$E(\hat{\bar{Y}}_{regd}) = \bar{Y} + \bar{X}\beta\left[E(\varepsilon_2\varepsilon_3) - E(\varepsilon_2\varepsilon_4) - E(\varepsilon_1\varepsilon_3) + E(\varepsilon_1\varepsilon_4)\right]$$

$$Bias(\hat{\bar{Y}}_{regd}) = E(\hat{\bar{Y}}_{regd}) - \bar{Y}$$

$$= \bar{X}\beta\left[\left(\frac{1}{n'} - \frac{1}{N}\right)\frac{1}{N}\sum\left(\frac{(\bar{x}' - \bar{X})(s_{xy} - S_{xy})}{\bar{X}S_{xy}}\right)\right.$$

$$-\left(\frac{1}{n'} - \frac{1}{N}\right)\frac{1}{N}\sum\left(\frac{(\bar{x}' - \bar{X})(s_x^2 - S_X^2)}{\bar{X}S_X^2}\right)$$

$$-\left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{N}\sum\left(\frac{(\bar{x} - \bar{X})(s_{xy} - S_{xy})}{\bar{X}S_{xy}}\right)$$

$$\left.+\left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{N}\sum\left(\frac{(\bar{x} - \bar{X})(s_x^2 - S_x^2)}{\bar{X}S_x^2}\right)\right]$$

$$= \bar{X}\beta\left[\left(\frac{1}{n'} - \frac{1}{N}\right)\frac{\mu_{21}}{\bar{X}S_{xy}} - \left(\frac{1}{n'} - \frac{1}{N}\right)\frac{\mu_{30}}{\bar{X}S_x^2} - \left(\frac{1}{n} - \frac{1}{N}\right)\frac{\mu_{21}}{\bar{X}S_{xy}} + \left(\frac{1}{n} - \frac{1}{N}\right)\frac{\mu_{30}}{\bar{X}S_x^2}\right]$$

$$= -\beta\left(\frac{1}{n} - \frac{1}{n'}\right)\left(\frac{\mu_{21}}{S_{xy}} - \frac{\mu_{30}}{S_x^2}\right).$$

Mean squared error:

$$MSE(\hat{\bar{Y}}_{regd}) = E(\bar{Y}_{regd} - \bar{Y})^2$$

$$= \left[\bar{y} + \hat{\beta}(\bar{x}' - \bar{x}) - \bar{Y}\right]^2$$

$$= E\left[(\bar{y} - \bar{Y}) + \bar{X}\beta(1 + \varepsilon_3)(\varepsilon_2 - \varepsilon_1)(1 - \varepsilon_4 + \varepsilon_4^2 - ...)\right]^2$$

Retaining the powers of $\varepsilon's$ up to order two, the mean squared error up to the second order of approximation is

$$
\begin{aligned}
MSE(\hat{\bar{Y}}_{regd}) &\simeq E\left[(\bar{y}-\bar{Y})+\bar{X}\beta(\varepsilon_2+\varepsilon_2\varepsilon_3-\varepsilon_2\varepsilon_4-\varepsilon_1-\varepsilon_1\varepsilon_3+\varepsilon_1\varepsilon_4)\right]^2 \\
&\simeq E(\bar{y}-\bar{Y})^2+\bar{X}^2\beta^2E(\varepsilon_1^2+\varepsilon_2^2-2\varepsilon_1\varepsilon_2)+2\bar{X}\beta E[(\bar{y}-\bar{Y})(\varepsilon_1-\varepsilon_2)] \\
&= E(\bar{y}-\bar{Y})^2+\bar{X}^2\beta^2E(\varepsilon_1^2+\varepsilon_2^2-2\varepsilon_1\varepsilon_2)+2\bar{X}\bar{Y}\beta E[\varepsilon_0(\varepsilon_1-\varepsilon_2)] \\
&= Var(\bar{y})+\bar{X}^2\beta^2\left[\left(\frac{1}{n}-\frac{1}{N}\right)\frac{S_x^2}{\bar{X}^2}+\left(\frac{1}{n'}-\frac{1}{N}\right)\frac{S_x^2}{\bar{X}^2}-2\left(\frac{1}{n}-\frac{1}{N}\right)\frac{S_x^2}{\bar{X}^2}\right] \\
&\quad -2\beta\bar{X}\bar{Y}\left[\left(\frac{1}{n'}-\frac{1}{N}\right)\frac{S_{xy}}{\bar{X}\bar{Y}}-\left(\frac{1}{n}-\frac{1}{N}\right)\frac{S_{xy}}{\bar{X}\bar{Y}}\right] \\
&= Var(\bar{y})+\beta^2\left(\frac{1}{n}-\frac{1}{n'}\right)S_x^2-2\beta\left(\frac{1}{n}-\frac{1}{n'}\right)S_{xy} \\
&= Var(\bar{y})+\left(\frac{1}{n}-\frac{1}{n'}\right)\left(\beta^2S_x^2-2\beta S_{xy}\right) \\
&= Var(\bar{y})+\left(\frac{1}{n}-\frac{1}{n'}\right)\left(\frac{S_{xy}^2}{S_x^4}S_x^2-2\frac{S_{xy}}{S_x^2}S_{xy}\right) \\
&= \left(\frac{1}{n}-\frac{1}{N}\right)S_y^2-\left(\frac{1}{n}-\frac{1}{n'}\right)\left(\frac{S_{xy}}{S_x}\right)^2 \\
&= \left(\frac{1}{n}-\frac{1}{N}\right)S_y^2-\left(\frac{1}{n}-\frac{1}{n'}\right)\rho^2S_y^2 \quad (\text{using } S_{xy}=\rho S_xS_y) \\
&\approx \frac{(1-\rho^2)S_y^2}{n}+\frac{\rho^2S_y^2}{n'}. \quad (\text{Ignoring the finite population correction})
\end{aligned}
$$

Clearly, $\hat{\bar{Y}}_{regd}$ is more efficient than the sample mean SRS, i.e. when no auxiliary variable is used.

Now we address the issue of whether the reduction in variability is worth the extra expenditure required to observe the auxiliary variable.

Let the total cost of the survey is

$$C_0 = C_1n + C_2n'$$

where C_1 and C_2 are the costs per unit observing the study variable y and auxiliary variable x, respectively.

Now minimize the $MSE\,(\hat{\bar{Y}}_{regd})$ for fixed cost C_0 using the Lagrangian function with Lagrangian multiplier λ as

$$\varphi = \frac{S_y^2(1-\rho^2)}{n} + \frac{\rho^2 S_y^2}{n'} + \lambda(C_1 n + C_2 n' - C_0)$$

$$\frac{\partial \varphi}{\partial n} = 0 \Rightarrow -\frac{1}{n^2} S_y^2(1-\rho^2) + \lambda C_1 = 0$$

$$\frac{\partial \varphi}{\partial n'} = 0 \Rightarrow -\frac{1}{n'^2} S_y^2 \rho^2 + \lambda C_2 = 0$$

Thus $n = \sqrt{\frac{S_y^2(1-\rho^2)}{\lambda C_1}}$

and $n' = \frac{\rho S_y}{\sqrt{\lambda C_2}}$.

Substituting these values in the cost function, we have

$$C_0 = C_1 n + C_2 n'$$

$$= C_1 \sqrt{\frac{S_y^2(1-\rho^2)}{C_1 \lambda}} + C_2 \sqrt{\frac{\rho^2 S_y^2}{\lambda C_2}}$$

$$\text{or } C_0 \sqrt{\lambda} = \sqrt{C_1 S_y^2(1-\rho^2)} + \sqrt{C_2 \rho^2 S_y^2}$$

$$\text{or } \lambda = \frac{1}{C_0^2}\left[S_y \sqrt{C_1(1-\rho^2)} + \rho S_y \sqrt{C_2} \right]^2.$$

Thus the optimum values of n and n' are

$$n'_{opt} = \frac{\rho S_y C_0}{\sqrt{C_2}\left[S_y \sqrt{C_1(1-\rho^2)} + \rho S_y \sqrt{C_2} \right]}$$

$$n_{opt} = \frac{C_0 S_y \sqrt{1-\rho^2}}{\sqrt{C_1}\left[S_y \sqrt{C_1(1-\rho^2)} + \rho S_y \sqrt{C_2} \right]}.$$

The optimum mean squared error of $\hat{\bar{Y}}_{regd}$ is obtained by substituting $n = n_{opt}$ and $n' = n'_{opt}$ as

$$MSE(\hat{\bar{Y}}_{regd})_{opt} = \frac{S_y^2(1-\rho^2)\left[\sqrt{C_1}\left(\sqrt{C_1 S_y^2(1-\rho^2)}+\rho S_y\sqrt{C_2}\right)\right]}{C_0\sqrt{S_y^2(1-\rho^2)}}$$

$$+\frac{S_y^2\rho^2\sqrt{C_2}\left[S_y\left(\sqrt{C_1(1-\rho^2)}+\rho S_y\sqrt{C_2}\right)\right]}{\rho S_y C_0}$$

$$=\frac{1}{C_0}\left[S_y\sqrt{C_1(1-\rho^2)}+\rho S_y\sqrt{C_2}\right]^2$$

$$=\frac{S_y^2}{C_0}\left[\sqrt{C_1(1-\rho^2)}+\rho\sqrt{C_2}\right]^2$$

The optimum variance of $\bar{y}$ under SRS for SRS where no auxiliary information is used is

$$Var(\bar{y}_{SRS})_{opt} = \frac{C_1 S_y^2}{C_0}$$

which is obtained by substituting $\rho = 0, C_2 = 0$ in $MSE(\hat{\bar{Y}}_{SRS})_{opt}$. The relative efficiency is

$$RE = \frac{Var(\bar{y}_{SRS})_{opt}}{MSE(\hat{\bar{Y}}_{regd})_{opt}} = \frac{C_1 S_y^2}{S_y^2\left[\sqrt{C_1(1-\rho^2)}+\rho\sqrt{C_2}\right]^2}$$

$$=\frac{1}{\left[\sqrt{1-\rho^2}+\rho\sqrt{\frac{C_2}{C_1}}\right]^2}$$

$$\leq 1.$$

Thus the double sampling in regression estimator will lead to gain in precision if

$$\frac{C_1}{C_2} > \frac{\rho^2}{\left[1-\sqrt{1-\rho^2}\right]^2}.$$

Double sampling for probability proportional to size estimation:

Suppose it is desired to select the sample with probability proportional to an auxiliary variable x but information on x is not available. Then, in this situation, the double sampling can be used. An initial sample of size n' is selected with SRSWOR from a population of size N, and information on x is collected for this sample. Then a second sample of size n is selected with replacement and with probability proportional to x from the initial sample of size n'. Let $\bar{x}'$ denote the mean of x for the initial sample of size n', Let $\bar{x}$ and $\bar{y}$ denote means respectively of x and y for the second sample of size n. Then we have the following theorem.

Theorem:

(1) An unbiased estimator of the population mean $\bar{Y}$ is given as

$$\hat{\bar{Y}} = \frac{x'_{tot}}{n'n}\sum_{i=1}^{n}\left(\frac{y_i}{x_i}\right),$$

where x'_{tot} denotes the total for x in the first sample.

(2) $Var(\hat{\bar{Y}}) = \left(\frac{1}{n'}-\frac{1}{N}\right)S_y^2 + \frac{(n'-1)}{N(N-1)nn'}\sum_{i=1}^{N}\frac{x_i}{X_{tot}}\left(\frac{\frac{y_i}{x_i}}{\frac{X_{tot}}{}} - Y_{tot}\right)^2$, where X_{tot} and Y_{tot} denote the totals of x and y respectively in the population.

(3) An unbiased estimator of the variance of $\hat{\bar{Y}}$ is given by

$$Var(\hat{\bar{Y}}) = \left(\frac{1}{n'}-\frac{1}{N}\right)\frac{1}{n(n'-1)}+\left[x'_{tot}\sum_{i=1}^{n}\frac{y_i^2}{x_i}-\frac{x'^2_{tot}(A-B)}{n'(n-1)}\right]+\frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{x'_{tot}y_i}{n'x_i}-\hat{\bar{Y}}\right)$$

where $A = \left(\sum_{i=1}^{n}\frac{y_i}{x_i}\right)^2$ and $B = \sum_{i=1}^{n}\frac{y_i^2}{x_i^2}$

Proof. Before deriving the results, we first mention the following result proved in varying probability scheme sampling.

Result: In sampling with varying probability scheme for drawing a sample of size n from a population of size N and with replacement.

(i) $\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$ is an unbiased estimator of the population mean $\bar{y}$ where $z_i = \frac{y_i}{Np_i}$, p_i being the probability of selection of i^{th} unit. Note that y_i and p_i can take any one of the N values $Y_1, Y_2, ..., Y_N$ with initial probabilities $P_1, P_2, ..., P_N$, respectively.

(ii) $Var(\bar{z}) = \frac{1}{nN^2}\left[\sum_{i=1}^{N}\frac{Y_i^2}{P_i} - N^2\bar{Y}^2\right] = \frac{1}{nN^2}\sum_{i=1}^{N}P_i\left(\frac{Y_i}{P_i}-\bar{Y}\right)^2..$

(iii) An unbiased estimator of the variance of $\bar{z}$ is

$$Var(\bar{z}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{Np_i}-\bar{z}\right)^2..$$

Let E_2 denote the expectation of $\hat{\bar{Y}}$, when the first sample is fixed. The second is selected with probability proportional to x, hence using the result (i) with $P_i = \frac{x_i}{x'_{tot}}$, we find that

$$E_2\left(\frac{\hat{\bar{Y}}}{n'}\right) = E_2\left[\frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{n'\frac{x_i}{x'_{tot}}}\right]$$
$$= E_2\left[\frac{x'_{tot}}{nn'}\sum_{i=1}^{n}\left(\frac{y_i}{x_i}\right)\right]$$
$$= \bar{y}'$$

where $\bar{y}'$ is the mean of y for the first sample. Hence

$$E(\hat{\bar{Y}}) = E_1\left[E_2\left(\hat{\bar{Y}} \mid n'\right)\right]$$
$$= E_1(\bar{y}_{n'})$$
$$= \hat{\bar{Y}},$$

which proves the part (1) of the theorem. Further,

$$Var(\hat{\bar{Y}}) = V_1E_2\left(\hat{\bar{Y}} \mid n'\right) + E_1V_2\left(\hat{\bar{Y}} \mid n'\right)$$
$$= V_1(\bar{y}') + E_1V_2\left(\hat{\bar{Y}} \mid n'\right)$$
$$= \left(\frac{1}{n'} - \frac{1}{N}\right)S_y^2 + E_1V_2\left(\hat{\bar{Y}} \mid n'\right).$$

Now, using the result (ii), we get

$$V_2\left(\hat{\bar{Y}} \mid n'\right) = \frac{1}{nn'^2}\sum_{i=1}^{n'}\frac{x_i}{x'_{tot}}\left(\frac{y_i}{\frac{x_i}{x'_{tot}}} - y'_{tot}\right)^2$$
$$= \frac{1}{nn'^2}\sum_{i=1}^{n'}\sum_{i<j}^{n'}x_ix_j\left(\frac{y_i}{x_i} - \frac{y_j}{x_j}\right)^2,$$

and hence

$$E_1V_2\left(\hat{\bar{Y}} \mid n'\right) = \frac{1}{nn'^2}\frac{n'(n'-1)}{N(N-1)}\sum_{i=1}^{N}\sum_{i<j}^{n'}x_ix_j\left(\frac{y_i}{x_i} - \frac{y_j}{x_j}\right)^2,$$

using the probability of a specified pair of units being selected in the sample is $\frac{n'(n'-1)}{N(N-1)}$. So we can express

$$E_1 V_2\left(\hat{\bar{Y}} \mid n'\right) = \frac{1}{nn'^2}\frac{n'(n'-1)}{N(N-1)}\sum_{i=1}^{N}\frac{x_i}{X_{tot}}\left(\frac{\frac{y_i}{x_{tot}}}{\frac{X_{tot}}{}} - Y_{tot}\right)^2.$$

Substituting this in $V_2\left(\hat{\bar{Y}} \mid n'\right)$, we get

$$Var(\hat{\bar{Y}}) = \left(\frac{1}{n'} - \frac{1}{N}\right)S_y^2 + \frac{(n'-1)}{nn'N(N-1)}\sum_{i=1}^{N}\frac{x_i}{X_{tot}}\left(\frac{\frac{y_i}{x_i}}{\frac{}{X_{tot}}} - Y_{tot}\right)^2.$$

This proves the second part (2) of the theorem.

We now consider the estimation of $Var(\hat{\bar{Y}})$. Given the first sample, we obtain

$$E_2\left[\frac{1}{n}\sum_{i=1}^{n}\frac{y_i^2}{p_i}\right] = \sum_{i=1}^{n'} y_i^2,$$

where $p_i = \frac{x_i}{x_{tot}'}$. Also, given the first sample,

$$E_2\left[\frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{n' p_i} - \hat{\bar{Y}}\right)^2\right] = V_2(\hat{\bar{Y}}) = E_2(\hat{\bar{Y}}^2) - \bar{y}'^2.$$

Hence

$$E_2\left[\hat{\bar{Y}}^2 - \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{n' p_i} - \hat{\bar{Y}}^2\right)^2\right] = \bar{y}'^2.$$

Substituting $\hat{\bar{Y}} = \frac{x_{tot}'}{n'n}\sum_{i=1}^{n}\left(\frac{y_i}{x_i}\right)$ and $p_i = \frac{x_i}{x_{tot}'}$ the expression becomes

$$E_2\left[\frac{x'^2}{nn'^2(n-1)}\left\{\left(\sum_{i=1}^{n}\frac{y_i}{x_i}\right)^2 - \left(\sum_{i=1}^{n}\frac{y_i^2}{x_i^2}\right)\right\}\right] = \bar{y}'^2$$

Using

$$E_2\left[\frac{1}{n}\sum_{i=1}^{n}\frac{y_i^2}{p_i}\right] = \sum_{i=1}^{n'} y_i^2,$$

we get

$$E_2\left[\frac{1}{n}\sum_{i=1}^{n} y_i^2 \frac{x_{tot}^{'}}{x_i} - \frac{x_{tot}^{'2}}{nn'(n-1)}(A-B)\right] = \sum_{i=1}^{n'} y_i^2 - n'\bar{y}'^2$$

where $A = \left(\sum_{i=1}^{n} \frac{y_i}{x_i}\right)^2$, and $B = \sum_{i=1}^{n} \frac{y_i^2}{x_i^2}$ which further simplifies to

$$E_2\left[\frac{1}{n(n'-1)}\left\{x_{tot}^{'}\sum_{i=1}^{n}\frac{y_i^2}{x_i} - \frac{x_{tot}^{'2}(A-B)}{n'(n-1)}\right\}\right] = s_y^{'2},$$

where $s_y^{'2}$ is the mean sum of squares of y for the first sample. Thus, we obtain

$$E_1E_2\left[\frac{1}{n(n'-1)}\left\{x_{tot}^{'}\sum_{i=1}^{n}\frac{y_i^2}{x_i} - \frac{x_{tot}^{'2}(A-B)}{n'(n-1)}\right\}\right] = E_1(s_y^{'2}) = S_y^2 \qquad (1)$$

which gives an unbiased estimator of S_y^2. Next, since we have

$$E_1V_2\left(\hat{\bar{Y}} \mid n'\right) = \frac{1}{nn'}\frac{(n'-1)}{N(N-1)}\sum_{i=1}^{N}\frac{x_i}{X_{tot}}\left(\frac{\frac{y_i}{x_i}}{\frac{x_i}{X_{tot}}} - Y_{tot}\right)^2,$$

and from this result we obtain

$$E_2\left[\frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i x_{tot}^{'}}{n'x_i} - \hat{\bar{Y}}\right)^2\right] = V_2\left(\hat{\bar{Y}} \mid n'\right).$$

Thus

$$E_1E_2\left[\frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{x_{tot}^{'}y_i}{n'x_i} - \hat{\bar{Y}}\right)^2\right] = \frac{(n'-1)}{nn'N(n-1)}\sum_{i=1}^{N}\frac{x_i}{X_{tot}}\left(\frac{\frac{y_i}{x_i}}{\frac{x_i}{X_{tot}}} - Y_{tot}\right)^2 \qquad (2)$$

when gives an unbiased estimator of

$$\frac{(n'-1)}{nn'N(n-1)}\sum_{i=1}^{N}\frac{x_i}{X_{tot}}\left(\frac{\frac{y_i}{x_i}}{\frac{x_i}{X_{tot}}} - Y_{tot}\right)^2.$$

Using (1) and (2) an unbiased estimator of the variance of $\hat{\bar{Y}}$ is obtained as

$$Var(\hat{\bar{Y}}) = \left(\frac{1}{n'} - \frac{1}{N}\right)\frac{1}{n(n'-1)}\left[x_{tot}^{'}\sum_{i=1}^{n}\frac{y_i^2}{x_i} - \frac{x_{tot}^{'2}(A-B)}{n'(n-1)}\right] + \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{x_{tot}^{'}y_i}{n'x_i} - \hat{\bar{Y}}\right)^2$$

Thus, the theorem is proved.