

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452

Maths 2024; 9(6): 01-07

© 2024 Stats & Maths

<https://www.mathsjournal.com>

Received: 02-08-2024

Accepted: 03-09-2024

Naveen GPDepartment of Mathematics and
Statistics CCSHAU, Hisar,
Haryana, India**Manoj Kumar**Department of Mathematics and
Statistics CCSHAU, Hisar,
Haryana, India**Alisha Mittal**Department of Anaesthesia and
Intensive Care PGIMER,
Chandigarh, India**Pankaj Das**ICAR-Indian Agricultural
Statistics Research Institute,
New Delhi, Delhi, India

Improved estimator of population mean under compromise method of imputation

Naveen GP, Manoj Kumar, Alisha Mittal and Pankaj DasDOI: <https://dx.doi.org/10.22271/math.2024.v9.i6a.1878>

Abstract

Analysing data becomes complicated when multiple values are missing, as missing data introduces bias, hampers handling and analysis, and decreases efficiency. Imputation is a technique used to address this issue by filling in missing values based on existing data and auxiliary information. Various methods are available for imputation. Once the missing observations are imputed, the complete data set can be analysed using traditional methods. Historically, methods such as Mean Imputation, Ratio Imputation, and Compromise Imputation have been employed to address missing data. In this study an alternate estimator of population mean using estimator given by Khoshnevisan (2007) under compromised imputation method is given. The expression for bias and mean squared error (MSE) up to the first order approximation are derived. An empirical study is also carried out in order to compare the efficiency of the estimator with the previously existing estimators.

Keywords: Population mean, auxiliary information, mean method of imputation, ratio method of imputation, and compromise imputation

Introduction

Nonresponses or missing values are common occurrences in sample surveys. Item nonresponse and unit nonresponse are the two types of nonresponse. Item nonresponse is deal with the imputation method, while unit nonresponse is deal with the weight method. Similarly, the most common method for solving missing values is by imputation, which replaces missing values using the existing data as a source. Additionally, many researchers have studied the auxiliary information available in order to increase the accuracy of population mean estimation simple random sampling without replacement (SRSWOR). For example, Cochran applied the auxiliary information at the estimation stage and proposed an estimator to estimate the population mean. Bahl and Tuteja (1991) [2] first proposed new ratio-type exponential method for estimating the mean of population using information on auxiliary variable and they are more efficient than the existing estimators.

Let the mean of the population $\pi = \{1, 2, \dots, N\}$ is $\bar{Y} = \frac{1}{N} \sum_{i=1}^N y_i$. A sample s of size n is drawn from the population by simple random sampling without replacement (SRSWOR) to estimate population mean ($\bar{Y}$). Let the number of responding units be r out of sampled units. The set of responding units be R and the remaining non responding units be R^c . For every units $i \in R$, the value of y_i is observed and the remaining units $i \in R^c$ are missed and have to be imputed using the imputation techniques. Here assume that the imputation has carried using the available auxiliary information x such that x_i , the values x for unit i . is known and it is positive for all $i \in s$. The following notations used are given by Lee *et al.* (1994) for the case of single value imputation. If the i^{th} unit requires the imputation, the value of $\hat{b}x_i$ is imputed where $b = \frac{\sum_{i \in R} y_i}{\sum_{i \in R} x_i}$. The data after the imputation becomes

$$y_i = \begin{cases} y_i & \text{if } i \in R \\ \hat{b}x_i & \text{if } i \in R^c \end{cases}$$

Corresponding Author:**Pankaj Das**ICAR-Indian Agricultural
Statistics Research Institute,
New Delhi, Delhi, India

The above imputation method used is known as Ratio method of imputation. The point estimator of population mean under this method of imputation is given by

$$\bar{y}_{RAT} = \bar{y}_r \frac{\bar{x}_n}{\bar{x}_r}$$

Where

$$\bar{x}_n = \frac{1}{n} \sum_{i \in S} x_i$$

$$\bar{x}_r = \frac{1}{r} \sum_{i \in R} x_i$$

$$\bar{y}_r = \frac{1}{r} \sum_{i \in R} y_i$$

Another method of imputation is mean method of imputation. In this method the data after imputation becomes

$$y_i = \begin{cases} y_i & \text{if } i \in R \\ \bar{y}_r & \text{if } i \in R^c \end{cases}$$

The point estimator for the above method of imputation is given by

$$\bar{y}_M = \frac{1}{r} \sum_{i \in R} y_i = \bar{y}_r$$

Singh and Horn (2000) have another imputation procedure called compromise imputation which is giving more efficient result than the other existing imputation procedures. The data after imputation becomes

$$y_i = \begin{cases} \alpha \frac{n}{r} y_i + (1 - \alpha) \hat{b} x_i & \text{if } i \in R \\ (1 - \alpha) \hat{b} x_i & \text{if } i \in R^c \end{cases}$$

The point estimator of the population mean is given by

$$\bar{y}_{COMP} = \alpha \bar{y}_r + (1 - \alpha) \bar{y}_r \frac{\bar{x}_n}{\bar{x}_r}$$

The study explores the derivation and computation procedure of an alternate estimator of population mean using estimator given by Khoshnevisan (2007) ^[5] under compromised imputation. The article is orientated so that it first discusses the estimators' theoretical foundation. In the following section, the proposed estimator and its characteristics are described. The subsequent part demonstrated how the appropriate estimator performed in comparison to the current estimators using a case study. It ends later with some general remarks.

Theoretical background

Let us define

$$\varepsilon = \frac{\bar{y}_r}{\bar{Y}} - 1$$

$$\delta = \frac{\bar{x}_r}{\bar{X}} - 1$$

$$\eta = \frac{\bar{x}_n}{\bar{X}} - 1$$

Rao and Sitter (1995) gave the concept of two-phase sampling and the mechanism of that for given r and n we have

$$E(\varepsilon) = E(\delta) = E(\eta) = 0$$

And

$$E(\varepsilon^2) = \left(\frac{1}{r} - \frac{1}{N}\right) C^2_y, E(\delta^2) = \left(\frac{1}{r} - \frac{1}{N}\right) C^2_x, E(\eta^2) = \left(\frac{1}{n} - \frac{1}{N}\right) C^2_x$$

$$E(\varepsilon\delta) = \left(\frac{1}{r} - \frac{1}{N}\right)\rho C_y C_x, E(\delta\eta) = \left(\frac{1}{n} - \frac{1}{N}\right)C^2_x, E(\varepsilon\eta) = \left(\frac{1}{n} - \frac{1}{N}\right)\rho C_y C_x$$

Where,

$$C^2_y = \frac{S^2_y}{\bar{Y}^2}$$

$$C^2_x = \frac{S^2_x}{\bar{X}^2}$$

$$\rho = \frac{S_{xy}}{(S_x S_y)}$$

The properties (Bias and MSE) of the above discussed imputation methods are as follows,
The bias and MSE of Ratio method of imputation is,

$$\diamond B(\bar{y}_{RAT}) \approx \left(\frac{1}{r} - \frac{1}{n}\right)\bar{Y}(C^2_x - \rho C_y C_x)$$

$$\diamond MSE(\bar{y}_{RAT}) \approx \left(\frac{1}{n} - \frac{1}{N}\right)S^2_y + \left(\frac{1}{r} - \frac{1}{n}\right)[S^2_y + R^2_1 S^2_x - 2R_1 S_{xy}]$$

The bias and MSE of the compromised method of imputation is,

$$\diamond B(\bar{y}_{COMP}) \approx (1 - \alpha)\left(\frac{1}{r} - \frac{1}{n}\right)\bar{Y}(C^2_x - \rho C_y C_x)$$

$$\diamond MSE(\bar{y}_{COMP}) \approx MSE(\bar{y}_{RAT}) - \left(\frac{1}{r} - \frac{1}{n}\right)\left(1 - \rho \frac{C_y}{C_x}\right)^2 \bar{Y}^2 C^2_x$$

Where,

$$\alpha = 1 - \rho \frac{C_y}{C_x} \text{ (suitable constant to reduce the variance)}$$

Proposed imputation method for missing data

By using the estimator given by Khoshnevisan (2007) [5], here we proposed a new estimator for missing data. The data after imputation takes as,

$$y_i = \begin{cases} k \frac{n}{r} y_i + (1 - k) \bar{y}_r d & \text{if } i \in R \\ (1 - k) \bar{y}_r d & \text{if } i \in R^c \end{cases}$$

where,

$$d = \left[\frac{a\bar{X} + b}{\alpha(a\bar{x}_r + b) + (1 - \alpha)(a\bar{X} + b)} \right]^g$$

where $a, b \neq 0$ are real numbers or the functions of known parameters of auxiliary variables x like standard deviation (σ_x), skewness [$\beta_1(x)$], kurtosis [$\beta_2(x)$], and correlation coefficient (ρ_x).

Theorem 3.1

The point estimator of population mean ($\bar{Y}$) under proposed method imputation is

$$\bar{y}_p = k \bar{y}_r + (1 - k) \bar{y}_r \left[\frac{a\bar{X} + b}{\alpha(a\bar{x}_r + b) + (1 - \alpha)(a\bar{X} + b)} \right]^g$$

Proof: we have $\bar{y}_p = \frac{1}{n} \sum_{i \in S} y_i$

$$= \frac{1}{n} \left[\sum_{i \in R} y_i + \sum_{i \in R^c} y_i \right]$$

$$= \frac{1}{n} \left[\sum_{i=1}^r \left(k \frac{n}{r} y_i + (1 - k) \bar{y}_r d \right) + \sum_{i=1}^{n-r} (1 - k) \bar{y}_r d \right]$$

$$\begin{aligned}
&= \frac{1}{n} \left[n \sum_{i=1}^r y_i + (1-k)d \sum_{i=1}^r \bar{y}_r + (1-k)d \sum_{i=1}^{n-r} \bar{y}_r \right] \\
&= \frac{1}{n} \left[k \frac{n}{r} r \bar{y}_r + (1-k)dr \bar{y}_r + (1-k)d(n-r) \bar{y}_r \right] \\
&= k \bar{y}_r + (1-k)d \bar{y}_r
\end{aligned}$$

Then the point estimator is given as

$$\bar{y}_p = k \bar{y}_r + (1-k) \bar{y}_r \left[\frac{a\bar{X} + b}{\alpha(a\bar{x}_r + b) + (1-\alpha)(a\bar{X} + b)} \right]^g$$

Theorem 3.2

The bias of the proposed point estimator is given as

$$B(\bar{y}_p) = (1-k) \frac{g\alpha a \bar{X} \bar{y}}{(a\bar{X} + b)} \left(\frac{1}{r} - \frac{1}{N} \right) \left[\frac{(g+1)\alpha a \bar{X}}{2(a\bar{X} + b)} C_x^2 - \rho_{xy} C_y C_x \right]$$

Proof: see appendix 1.

Theorem 3.3

The mean squared error (MSE) of the proposed estimator $\bar{y}_p$ is given by

$$MSE(\bar{y}_p) = \bar{y}^2 \left(\frac{1}{r} - \frac{1}{N} \right) \left[C_y^2 + \frac{(1-k)^2 g^2 \alpha^2 a^2 \bar{X}^2}{(a\bar{X} + b)^2} C_x^2 - \frac{2(1-k)g\alpha a \bar{X}}{(a\bar{X} + b)} \rho_{xy} C_y C_x \right]$$

Proof: See appendix 2.

Case Study

In this section, the proposed estimator is compared with existing estimators. For this study the two population data was used which are given by Murthy (1967), the population constants are as follows.

Table 1: Data description

Parameters	Population 1	Population 2
N	80	80
n	25	25
r (assumed)	20	20
$\bar{Y}$	5182.638	5182.638
$\bar{X}$	285.125	1126.463
C_y	0.3542	0.3542
C_x	0.9485	0.7507
$\beta_1(X)$	1.2680	1.0237
$\beta_2(X)$	3.5360	2.8306
ρ_{YX}	0.9140	0.9140

Table 2: Bias of estimators.

Estimators	Population 1	Population 2
$B(\bar{y}_{Ratio})$	30.7116	16.6114
$B(\bar{y}_{comp})$	10.4823	7.1636
$B(\bar{y}_{proposed})$	1.0041	-0.6762

The above table 2 is explaining about the bias of the proposed estimator and for the previously existing estimators (Ratio estimator and compromise imputation estimator) for both the populations. According to the above population it clearly tells us that for both populations the proposed estimator is showing less bias compared to the other estimators. Hence the proposed estimator is giving better results than other existing estimators.

Table 3: MSE of estimators.

Estimators	Population 1	Population 2
$MSE(\bar{y}_{Ratio})$	203055.82	147179.33
$MSE(\bar{y}_{comp})$	98215.14	98215.14
$MSE(\bar{y}_{proposed})$	74650.61	91095.63

The table 3 showing the MSE of Proposed estimator and the existing estimators (Ratio estimator and compromise imputation estimator). From the above table for both the populations MSE of proposed estimator is less than the existing estimators by this we can conclude that the proposed estimator is preferable over the existing estimators by obtaining efficient point estimator. Figure 1 indicates that the proposed estimator has the lowest bias compared to other two estimators. Similarly, in case of MSE, the proposed estimator outperforms than the other estimators.

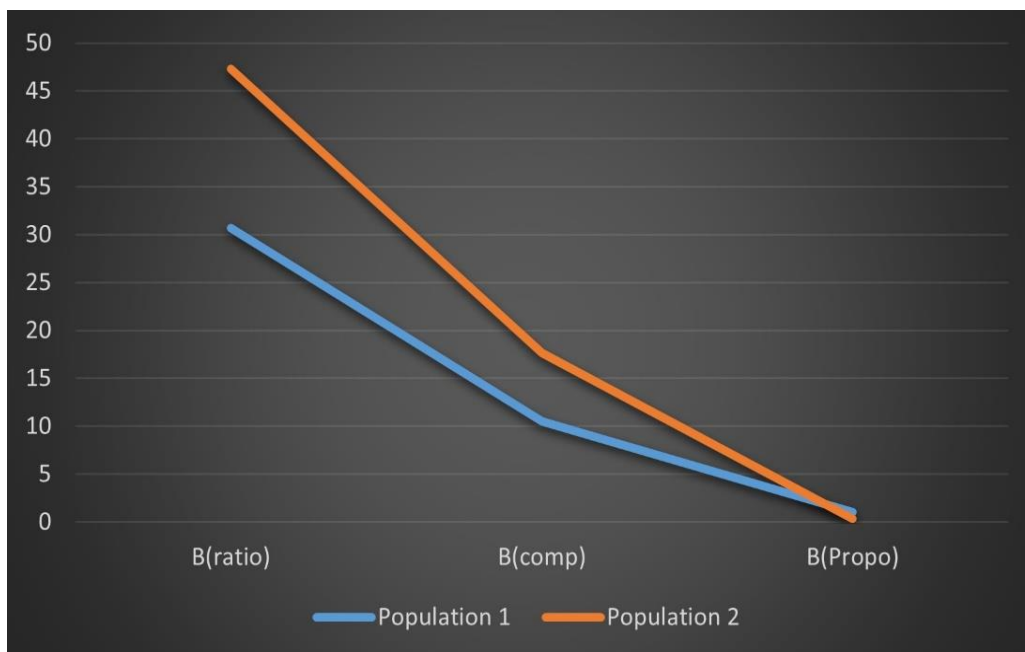


Fig 1: Bias of the estimators

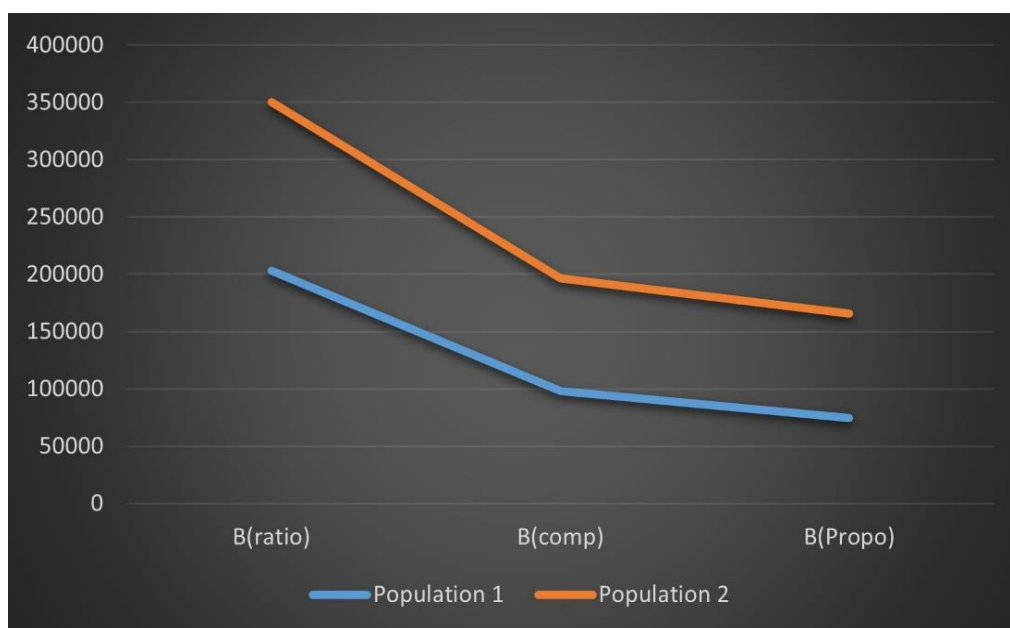


Fig 2: Mean squared of the estimator

Conclusion

A problem of non-response for a specific unit or units in the population occurs frequently in surveys involving the medical profession, social surveys, and household surveys etc... Due to missing values in the data set, this non-sampling error may creep and produce wrong inferences. In the present study, the alternative estimator of population mean under compromised imputation method has been proposed. Its bias and MSE also been proposed. For the proposed estimator the efficiency has been checked and prove to be more efficient than the existing estimators like ratio method of imputation, mean method of imputation and compromised imputation.

Acknowledgments

We also thank two anonymous referees and the editor for their helpful comments and suggestions that have substantially improved the presentation of this article.

Data availability statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. Adhvaryu D, Gupta PC. On some alternative sampling strategies using auxiliary information. *Metrika*. 1983;30(1):217-226.
2. Bahl S, Tuteja R. Ratio and product type exponential estimators. *J Inform Optim Sci*. 1991;12(1):159-164.
3. Chodjuntug K, Lawson N. A chain ratio exponential-type compromised imputation for mean estimation: case study on ozone pollution in Saraburi, Thailand. *J Probab Stat*; c2020.
4. Gira AA. Estimation of population mean with a new imputation method. *Appl Math Sci*. 2015;9(34):1663-1672.
5. Khoshnevisan M, Singh R, Chauhan P, Sawan N. A general family of estimators for estimating population mean using known value of some population parameter(s). *Far East J Theor Stat*.
6. Murthy MN. Sampling theory and methods. 2nd ed. Calcutta: Statistical Publishing Society; c1967.
7. Prasad S. A study on new methods of ratio exponential type imputation in sample surveys. *Hacet J Math Stat*. 2018;47(5):1281-1301.
8. Singh AK, Singh P, Singh VK. Exponential-type compromised imputation in survey sampling. *J Stat Appl Probab*. 2014;3(2):211.
9. Singh BK, Gogoi U. Estimation of population mean using ratio-cum-product imputation techniques in sample survey. *Res Rev J Stat*. 2018;7:38-49.
10. Singh GN. On the improvement of product method of estimation in sample surveys. *J Ind S Agri Stat*. 2003;56(3):267-265.
11. Singh GN, Maurya S, Khetan M, Kadilar C. Some imputation methods for missing data in sample surveys. *Hacet J Math Stat*. 2016;45(6):1865-1880.
12. Singh HP, Pal SK. A new chain ratio-ratio-type exponential estimator using auxiliary information in sample surveys. *Int J Math Appl*. 2015;55:7.
13. Singh S, Horn S. Compromised imputation in survey sampling. *Metrika*. 2000;51(3):267-276.

Appendix – 1: To prove 3.2 theorem, we have the point estimator as

$$\begin{aligned}\bar{y}_p &= k\bar{y}_r + (1-k)\bar{y}_r \left[\frac{a\bar{X} + b}{\alpha(a\bar{x}_r + b) + (1-\alpha)(a\bar{X} + b)} \right]^g \text{ expressing in terms of } \varepsilon, \delta \text{ and } \eta \\ \bar{y}_p &= k\bar{y}(1+\varepsilon) + (1-k)\bar{y}(1+\varepsilon) \left[\frac{a\bar{X} + b}{\alpha[a(1+\delta)\bar{X} + b] + (1-\alpha)(a\bar{X} + b)} \right]^g \\ &= k\bar{y}(1+\varepsilon) + (1-k)\bar{y}(1+\varepsilon) \left[\frac{a\bar{X} + b}{a\bar{X} + b + \alpha a \delta \bar{X}} \right]^g \\ &= k\bar{y}(1+\varepsilon) + (1-k)\bar{y}(1+\varepsilon) \left[1 + \frac{\alpha a \delta \bar{X}}{a\bar{X} + b} \right]^{-g} \\ &= k\bar{y}(1+\varepsilon) + (1-k)\bar{y}(1+\varepsilon) \left[1 - \frac{g\alpha a \bar{X} \delta}{(a\bar{X} + b)} + \frac{g(g+1)\alpha^2 a^2 \bar{X}^2 \delta^2}{2(a\bar{X} + b)^2} \dots \right] \\ &= k\bar{y}(1+\varepsilon) + \left[\bar{y} - k\bar{y} + \bar{y}\varepsilon - k\bar{y}\varepsilon - \frac{g\alpha a \bar{X} \bar{y} \delta}{(a\bar{X} + b)} + \frac{g\alpha k a \bar{X} \bar{y} \delta}{(a\bar{X} + b)} - \frac{g\alpha a \bar{X} \bar{y} \varepsilon \delta}{(a\bar{X} + b)} + \frac{g\alpha a \bar{X} \bar{y} k \varepsilon \delta}{(a\bar{X} + b)} + \frac{g(g+1)\alpha^2 a^2 \bar{X}^2 \bar{y} \delta^2}{2(a\bar{X} + b)^2} \right. \\ &\quad \left. - \frac{k g (g+1) \alpha^2 a^2 \bar{X}^2 \bar{y} \delta^2}{2(a\bar{X} + b)^2} \right] \\ \bar{y}_p &= \bar{y} + \varepsilon \bar{y} - \frac{(1-k)g\alpha a \bar{X} \bar{y} \delta}{(a\bar{X} + b)} - \frac{(1-k)g\alpha a \bar{X} \bar{y} \varepsilon \delta}{(a\bar{X} + b)} + \frac{(1-k)g(g+1)\alpha^2 a^2 \bar{X}^2 \bar{y} \delta^2}{2(a\bar{X} + b)^2} \\ \bar{y}_p - \bar{y} &= \bar{y}\varepsilon - \frac{(1-k)g\alpha a \bar{X} \bar{y}}{(a\bar{X} + b)} \left[\delta + \varepsilon \delta - \frac{g(g+1)\alpha a \bar{X} \delta^2}{2(a\bar{X} + b)} \right]\end{aligned}$$

Taking expectation on both side

$$E(\bar{y}_p) - \bar{Y} = \bar{y}E(\varepsilon) - \frac{(1-k)g\alpha a \bar{X} \bar{y}}{(a\bar{X} + b)} \left[E(\delta) + E(\varepsilon\delta) - \frac{g(g+1)\alpha a \bar{X}}{2(a\bar{X} + b)} E(\delta^2) \right]$$

$$B(\bar{y}_p) = (1-k) \frac{g\alpha a \bar{X} \bar{y}}{(a\bar{X} + b)} \left(\frac{1}{r} - \frac{1}{N} \right) \left[\frac{(g+1)\alpha a \bar{X}}{2(a\bar{X} + b)} C_x^2 - \rho_{xy} C_y C_x \right]$$

Appendix – 2: To prove 3.3 theorem, we have

$$(\bar{y}_p - \bar{y})^2 = \bar{y}^2 \varepsilon^2 + \frac{(1-k)^2 g^2 \alpha^2 a^2 \bar{X}^2 \bar{y}^2 \delta^2}{(a\bar{X} + b)^2} - \frac{2(1-k)g\alpha a \bar{X} \bar{y}^2 \varepsilon \delta}{(a\bar{X} + b)}$$

Now apply expectation on both side

$$\begin{aligned} E[(\bar{y}_p - \bar{y})^2] &= \bar{y}^2 E[\varepsilon^2] + \frac{(1-k)^2 g^2 \alpha^2 a^2 \bar{X}^2 \bar{y}^2}{(a\bar{X} + b)^2} E[\delta^2] - \frac{2(1-k)g\alpha a \bar{X} \bar{y}^2}{(a\bar{X} + b)} E[\varepsilon \delta] \\ &= \bar{y}^2 \left(\frac{1}{r} - \frac{1}{N} \right) C_y^2 + \frac{(1-k)^2 g^2 \alpha^2 a^2 \bar{X}^2 \bar{y}^2}{(a\bar{X} + b)^2} \left(\frac{1}{r} - \frac{1}{N} \right) C_x^2 - \frac{2(1-k)g\alpha a \bar{X} \bar{y}^2}{(a\bar{X} + b)} \left(\frac{1}{r} - \frac{1}{N} \right) \rho_{xy} C_y C_x \\ MSE(\bar{y}_p) &= \bar{y}^2 \left(\frac{1}{r} - \frac{1}{N} \right) \left[C_y^2 + \frac{(1-k)^2 g^2 \alpha^2 a^2 \bar{X}^2}{(a\bar{X} + b)^2} C_x^2 - \frac{2(1-k)g\alpha a \bar{X}}{(a\bar{X} + b)} \rho_{xy} C_y C_x \right] \end{aligned}$$

In order to obtain the value of k differentiate w.r.t. (1-k) and equate to zero, then

$$k = 1 - \left(\frac{a\bar{X} + b}{g\alpha a \bar{X}} \right) \rho_{xy} \frac{C_y}{C_x}$$