Model-Assisted Estimation of Population Mean in Two-Stage Cluster Sampling

Nelson Kiprono Bii
Strathmore University, Nairobi, Kenya
chrisouma2004@yahoo.co.uk

Ouma Christopher Onyango
Strathmore University, Kenyatta University, Kenya
chrisouma2004@yahoo.co.uk

Abstract

Estimation of finite population parameters has been an area of concern to statisticians for decades. This paper presents an estimation of the population mean under a model-assisted approach. Dorfman (1992), Breidt and Opsomer (2000) and Ouma *et al* (2010) carried out the estimation of finite population total on the assumption that the sample size is large and the sampling distribution is approximately normal. Unlike their researches, this paper considered a case when the sample size is small under a model-assisted approach. A model-assisted regression model was considered in a case where the cluster sizes are known only for the sampled clusters in order to predict the unobserved part of the population mean. Under mild assumptions, the proposed estimator is asymptotically unbiased and its conditional error variance tends to zero. Simulation studies show that model assisted estimation performs better than model based estimation of a finite population mean in a case where the sample size is small.

Keywords: Model-assisted surveys, Non-parametric inference, Two-stage cluster sampling.

1.1 Introduction

Sample surveys are concerned with obtaining desired information from a population. In sample survey methods, some portion of the population called a sample is used to make inferences about the entire population.

Every nation in the world uses surveys to estimate their rates of unemployment, basic prevalence of immunization against diseases, opinions about the central government, intentions to vote in an upcoming election, and people's satisfaction with goods and services that they purchase among other application areas. Surveys aid in tracking global economic trends, the rate of inflation in prices, and investments in new economic enterprises (Shewhart, 2004).

The theory of sample survey is concerned with the development of sampling strategies that yield the selection of a sample that best represents the entire population. It provides the procedures that are employed to make statistical inferences about the survey variable. It is also used in choosing the criteria for comparing various strategies while attempting to obtain optimal results from a sample survey. Generally there are four methods used in sample surveys namely model based, model assisted, design based and design assisted surveys (Ouma *et al* 2010). In this paper we used model-assisted approach to estimate the population mean in two-stage cluster sampling.

1.2 Background of the Problem

Model based methods in estimation of population parameters assume that there is an underlying model that generate survey units. However, if the assumed model is incorrect, the estimators of population parameters will certainly be incorrect. To address this problem, we wish to implement non-parametric approach to estimation of population mean then obtain a model that will be used to generate survey values given in finite populations in two stage cluster sampling. Model- assisted estimation of population total has been considered by Breidt and Opsomer (2000) under the assumption that the sample size is large and the sampling distribution is approximately normal. However, this is not always the case. In this study we consider a situation when the sample size is small and develop an estimator for estimating population mean.

In particular, the problem is to estimate the population mean defined as

$$\bar{Y} = \frac{1}{M} \sum_{i \in s \in c} \frac{y_i}{N_i} \tag{1.1}$$

1.3 Summary of the paper

The rest of the paper is organized as follows: section 2 gives a summary of the two stage cluster sampling that we proposed to use. In section 3 we introduce the estimator for the finite population mean in two stage cluster sampling. In section 4 we look at the asymptotic properties of the proposed estimator. In section 5 we give the conclusion of our paper.

2. Review of Two Stage Cluster Sampling

Consider a finite population U of M primary sample units (PSU's) or clusters labelled $1,2,\dots,M$ i.e $C = (1,2,\dots,M)$ where M is a known number. Let $N_i, i = 1,2,\dots,M$, be the number of secondary sampling units(SSU's) in the j^{th} PSU.

Let $y_{ij,} i = 1,2,\dots,M, j = 1,2,\dots,N_i$ be the value of the survey variable y for the SSU i belonging to the j^{th} PSU. We further assumed that the auxilliary data are known for all the selected clusters and the population elements. The non-sampled units are not known and we therefore use the auxiliary information together with the sampled units of the survey variable to estimate the non-sampled part of the population. We proposed to estimate the population mean defined by equation (1.1). In this paper, we generate the survey values y_i in the cluster j by the model given by

$$y_i = \mu(x_i) + e_i \tag{2.1}$$

Where e_i are independent random variables with mean zero and variance, $v(x)$. Further $v(x)$ is a smooth function of x and $v(x)$ is smooth and strictly positive, $\mu(x_i)$ is a function of the auxiliary variables.

3. Proposed Estimator of the Population Mean

The population mean to be estimated is given by equation (1.1). The non-sampled survey values of the population are unknown. We therefore use the sampled units of the population together with the model defined in equation (2.1) to estimate the non-sampled part of the population mean. The proposed estimator is given by

$$\hat{\bar{Y}} = \frac{1}{M}\left\{\sum_{i\in s\in c}\frac{\hat{y}_i}{N_i}\right\} \tag{3.1}$$

4. Properties of the Proposed Estimator

4.1 Asymptotic Unbiasedness of the Estimator

The estimator of population mean is given by

$$\hat{\bar{Y}} = \frac{1}{M}\left\{\sum_{i\in s\in c}\frac{\hat{y}_i}{N_i}\right\} \tag{4.1}$$

$$= \frac{1}{M}\left\{\sum_{i\in s\in c}\frac{(\mu(x_i)+e_i)}{N_i}\right\} \tag{4.2}$$

Denoting the initial sampling weights by π_i which is equal to the inverse of their selection probabilities i.e. probability of including unit i from cluster j in the sample, we have:

$$\hat{\bar{Y}} = \frac{1}{M}\left\{\sum_{i\in s\in c}\frac{\mu(x_i)}{N_i\pi_i}+\sum_{i\in s}\frac{e_i}{\pi_i N_i}\right\} \tag{4.3}$$

$$= \frac{1}{M}\left\{\sum_{i\in s\in c}\frac{\mu(x_i)}{N_i\pi_i}+\sum_{i\in s}\frac{\hat{y}_i-\mu(x_i)}{\pi_i N_i}\right\} \tag{4.4}$$

But as shown by Ouma *et al* (2010), $E(e_i)=0$, therefore the proposed estimator for the population mean becomes

$$E\left(\hat{\bar{Y}}\right) = \frac{1}{M}E\left[\sum_{i\in s\in c}\frac{\mu(x_i)}{N_i\pi_i}+\sum_{i\in s}\frac{\hat{y}_i-\mu(x_i)}{\pi_i N_i}\right] \tag{4.5}$$

$$= \frac{1}{M}E\left[\sum_{i\in s}\frac{\mu(x_i)}{N_i\pi_i}\right]+\frac{1}{M}E\sum_{i\in s}\left[\frac{\hat{y}_i-\mu(x_i)}{\pi_i N_i}\right] \tag{4.6}$$

$$= \frac{1}{M}E\left[\sum_{i\in s}\frac{\mu(x_i)}{N_i\pi_i}\right]+\frac{1}{M}E\sum_{i\in s}\left[\frac{\hat{e}_i}{\pi_i N_i}\right] \tag{4.7}$$

But

$$\frac{1}{M}\left[\frac{\sum E(\hat{e}_i)}{N_i}\right]\to 0$$

Since

$$E(e_i) = 0$$

Then

$$E\left(\hat{\bar{Y}}\right) = \frac{1}{M} E\left[\sum_{i \in s} \frac{\mu(x_i)}{N_i \pi_i}\right] \tag{4.8}$$

Let the sampling weights of the primary sampling units be given by

$$w\left(x_{ij}\right) = \frac{1}{\pi_{ij}}\ , i \in s, i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, m \tag{4.9}$$

Rupert (2003) applied (4.9) as a basis and suggested that the sampling weights in this kind of survey can be computed using

$$w * \left(x_{ij}\right) = \frac{r_i}{\pi_{ij}} \left(\frac{n}{n-1}\right)^{1/2} \tag{4.10}$$

Where r_i is the number of times the i^{th} primary sampling unit is chosen. Thus using this procedure, it follows that

$$E\left(\hat{\bar{Y}}\right) = \left(\frac{n}{n-1}\right)^{1/2} \frac{1}{M} E\left[\sum_{i \in s} \frac{\mu(x_i)}{N_i \pi_i} r_i\right] \tag{4.11}$$

But since we sample with replacement we have $1 \leq r \leq n$; Rao and Wu (1998) observed that there is considerable benefit and little loss in choosing $r = n - 1$; therefore we put $r = n - 1$, so that

$$E\left(\hat{\bar{Y}}\right) = \left(\frac{n}{n-1}\right)^{1/2} (n-1) \frac{1}{M} E\left[\sum_{i \in s} \frac{\mu(x_i)}{N_i \pi_i}\right] \tag{4.12}$$

$$= [n(n-1)]^{1/2} \frac{1}{M} E\left[\sum_{i \in s} \frac{\mu(x_i)}{N_i \pi_i}\right] \tag{4.13}$$

Let the initial sampling weights defined in (4.9) be given by the kernel weights:

$$w\left(x_{ij}\right) = \frac{(n-1) b^{-1} k\left(\frac{x_{ij} - x_{ik}}{b}\right)}{\sum_{ij \in s} (n-1) b^{-1} k\left(\frac{x_{ij} - x_{ik}}{b}\right)} \tag{4.14}$$

Where

$$\sum_{ij \in s} w\left(x_{ij}\right) = 1$$

Using (4.14) we have

$$E\left(\hat{\bar{Y}}\right) = [n(n-1)]^{1/2}\frac{1}{M}E\sum_{i\in s}\frac{\mu(x_{ij})}{N_i}\left(\frac{(n-1)b^{-1}k\left(\frac{x_{ij}-x_{ik}}{b}\right)}{\sum_{ij\in s}(n-1)b^{-1}k\left(\frac{x_{ij}-x_{ik}}{b}\right)}\right) \tag{4.15}$$

Let

$$d_s(\hat{x}_{ij}) = (n-1)b^{-1}k\left(\frac{x_{ij}-x_{ik}}{b}\right) \tag{4.16}$$

be the kernel estimator of $d_s(x_{ij})$ then

$$E\left(\hat{\bar{Y}}\right) = \frac{1}{M}[n(n-1)]^{1/2}\frac{1}{n-1}\sum_{i\in s}\frac{1}{N_i}E\left[\frac{\mu(x_{ij})(n-1)b^{-1}k\left(\frac{x_{ij}-x_{ik}}{b}\right)}{d_s(\hat{x}_{ij})}\right] \tag{4.17}$$

This consequently gives us

$$E\left(\hat{\bar{Y}}\right) = \frac{1}{M}\left(\frac{n}{n-1}\right)^{1/2}\sum_{i\in s}\left[\frac{1}{N_i}E\left(\mu(x_{ij})\right)\right]k\frac{(x_{ij}-x_{ik})}{b}\left[d_s(\hat{x}_{ij})\right]^{-1} \tag{4.18}$$

We make the following substitutions:

$$\left.\begin{array}{c} w = x_{ij} \\ u = \dfrac{w-x_{ik}}{b} \\ w = ub + x_{ik} \\ dw = bdu \end{array}\right\} \tag{4.19}$$

$$E\left(\hat{\bar{Y}}\right) = \frac{1}{M}\left(\frac{n}{n-1}\right)^{1/2}\sum_{i\in s}\left[\frac{1}{N_i}E(\mu(x_{ij}))\right]k(u)\left[d_s(\hat{x}_{ij})\right]^{-1} \tag{4.20}$$

Using substitutions in (4.19), from equation (4.20) we have

$$E\left[d_s(\hat{x}_{ij})/x_{ij}\right] = E\left[(n-1)b^{-1}\sum_{ij\in s}k\left(\frac{x_{ij}-x_{ik}}{b}\right)\right] \tag{4.21}$$

$$= \left(\frac{n-1}{b}\right)\int\sum_{ij\in s}k\left(\frac{x_{ij}-x_{ik}}{b}\right)d_s(x_{ij})\delta(x_{ij}) \tag{4.22}$$

$$= \left(\frac{n-1}{b}\right)\int\sum_{ij\in s}k(u)d_s(x_{ik}+bu)b\delta(u) \tag{4.23}$$

Where $E\left[d_s(\hat{x}_{ij})/x_{ij}\right]$ is the conditional mean of $d_s(\hat{x}_{ij})$ given the auxiliary values of $x_{ij}, ij \in s$.

We let $k_b(u) = b^{-1}k\left(u/b\right)$ where $k(u)$ is a kernel and b is a chosen bandwidth. In addition $k(u)$ is a symmetric density function such that

$$\int k(u)\,du = 1, \int uk(u)\,du = 0,\ k_2 = \int u^2k(u)\,du > 0\,, \int k^2(u)\,du > 0 \tag{4.24}$$

Since the relationship between the conditional mean and the selected bandwidth is too complex to establish we utilize the following theorem whose proof is given in Dorfman (1992).

Theorem: Let $k(u)$ be a symmetric density function with $\int uk(u)\,du = 0$ and $\int u^2 k(u)\,du > 0$, assume that n and N increase together such that $\frac{n}{N} \to \pi$ with $0 < \pi < 1$, assume sampled and non-sampled values of x are in the interval $[c, d]$ and are generated by densities d_s and d_{p-s} respectively both bounded away from zero on $[c, d]$ and assumed to have continuous second derivatives. If for any variable Z, $E(Z/\mathrm{U} = \mathrm{u}) = A(u) + O(B)$ and $var(Z/\mathrm{U} = \mathrm{u}) = O(c)$, U being a random variable and u an observation, then

$Z = A(u) + O_p\left(B + C^{\frac{1}{2}}\right)$, where $O_p(.)$ is an order of a term with p elements.

From expression (4.16) let $Z = d_s(\hat{x}_{ij})$ i.e

$$Z = (n-1)b^{-1}k\left(\frac{x_{ij} - x_{ik}}{b}\right) \tag{4.25}$$

Then

$$E\left[d_s(\hat{x}_{ij})\Big/_{x_{ij}}\right] = (n-1)b^{-1}\sum_{ij\in s} E\left[k\left(\frac{x_{ij} - x_{ik}}{b}\right)\right] \tag{4.26}$$

$$= \left(\frac{n-1}{b}\right)\sum_{ij\in s}\int k\left(\frac{w - x_{ik}}{b}\right) d_s(w)\delta(w) \tag{4.27}$$

Using the substitution

$$\left(\frac{w - x_{ik}}{b}\right) = u \tag{4.28}$$

And applying Taylor's series expansion about a point x_{ij} we get

$$E\left[d_s(\hat{x}_{ij})\Big/_{x_{ij}}\right] = d_s(x_{ij}) + b^2\frac{k_2}{2}d_s{}''(x_{ij}) + O(b^3) \tag{4.29}$$

$$= d_s(x_{ik}) + b^2\frac{k_2}{2}d_s{}''(x_{ik}) + O(b^3) \tag{4.30}$$

$$Var\left[d_s(\hat{x}_{ij})\Big/_{x_{ij}}\right] = (n-1)^2 b^{-2}\sum_{ij\in s} Var\left[k\frac{(x_{ij} - x_{ik})}{b}\right] \tag{4.31}$$

$$= (n-1)^2 b^{-2}\sum_{ij\in s}\left[E\left[k\left(\frac{x_{ij} - x_{ik}}{b}\right)\right]^2 - \left[Ek\left(\frac{x_{ij} - x_{ik}}{b}\right)\right]^2\right] \tag{4.32}$$

$$= \left(\frac{n-1}{b}\right)^2\sum_{ij\in s}\left[E\left[k\left(\frac{w - x_{ik}}{b}\right)\right]^2 - \left[Ek\left(\frac{w - x_{ik}}{b}\right)\right]^2\right] \tag{4.33}$$

$$= \left(\frac{n-1}{b}\right)^2\sum_{ij\in s}\left\{\int k\left(\frac{w - x_{ik}}{b}\right)^2 d_s(w)\,\delta_w - \left[\int k\left(\frac{w - x_{ik}}{b}\right) d_s(w)\,\delta_w\right]^2\right\} \tag{4.34}$$

Applying Taylor's series expansion about a point x_{ik} on $d_s(x_{ik}+bu)$, we have

$$E\left[d_s(\hat{x}_{ij})\Big/_{x_{ij}}\right] = (n-1)^{-1}\sum_{ij\in s}\left[d_s(x_{ij})\int k(u)du + bd_s{}'(x_{ij})\int uk(u)\,du + \frac{b^2}{2}d_s{}''(x_{ij})\int u^2k(u)du + \frac{b^3}{3!}d_s{}''(x_{ij})\int u^3\,k(u)du + \cdots\right] \quad (4.35)$$

By the theorem stated above, this reduces to

$$E\left[d_s(\hat{x}_{ij})\Big/_{x_{ij}}\right] = d_s(x_{ik}) + b^2\frac{k_2}{2}d_s{}''(x_{ik}) + \cdots \quad (4.36)$$

$$Var\left[d_s(\hat{x}_{ij})\Big/_{x_{ij}}\right] = E\left[d_s(\hat{x}_{ij})\Big/_{x_{ij}}\right]^2 - \left[E\left[d_s(\hat{x}_{ij})\Big/_{x_{ij}}\right]\right]^2 \quad (4.37)$$

The second term in equation (4.37) is given in equation (4.36), we therefore need to evaluate $E\left[d_s(\hat{x}_{ij})\Big/_{x_{ij}}\right]^2$ as follows:

$$E\left[\hat{d}_s(x_{ij})\Big/_{x_{ij}}\right]^2 = E\left[(n-1)^2\left(\sum_{j\in s}k\left(\frac{x_{ij}-x_{ik}}{b}\right)\right)^2\right] \quad ,i\in s, \quad (4.38)$$

$$= E\left[\left(\frac{n-1}{b}\right)^2\sum_{i\in s}k\left(\frac{x_{ij}-x_{ik}}{b}\right)\sum_{i\in s}k\left(\frac{x_{ij}{}'-x_{ik}}{b}\right)\right] \quad (4.39)$$

But

$$E\left[\sum_{i\in s}k\left(\frac{x_{ij}-x_{ik}}{b}\right)\Big/_{x_{ij}}\right]^2 = \sum_{i\in s}\int\left[k\left(\frac{x_{ij}-x_{ik}}{b}\right)\Big/_{x_{ij}}\right]^2 d_s(x_{ij})\delta_{x_{ij}}, \quad (4.40)$$

Applying the substitutions in (4.19) equation (4.40) becomes

$$E\left[\sum_{i\in s}k\left(\frac{x_{ij}-x_{ik}}{b}\right)\Big/_{x_{ij}}\right]^2 = \sum_{i\in s}\int k^2(u)\,d_s(x_{ik}+bu)b\delta u \quad (4.41)$$

Using Taylor's series expansion about point x_{ij} in equation (4.41) we get

$$E\left[\sum_{i\in s} k\left(\frac{x_{ij}-x_{ik}}{b}\right)\Big/ x_{ij}\right]^2 = \sum_{i\in s}\int k^2(u)\, b\left\{d_s(x_{ij}) + bud_s{}'(x_{ij}) + \frac{(bu)^2}{2} d_s{}''(x_{ij}) + O(b^3)\right\} du \quad (4.42)$$

and

$$E\sum_{ij}\sum k\left(\frac{x_{ij}-x_{ik}}{b}\right) k\left(\frac{x_{ij}{}'-x_{ik}}{b}\right) = b^2 \sum\sum \int k(u)\, d_s(x_{ik} + bu)du \int k(u')\, d_s(x_{ik}+bu')du' \quad (4.43)$$

Expanding the right hand side of equation (4.43) we get

$$b^2\sum\sum\left[d_s(x_{ij})\int k(u)du + bd_s{}'(x_{ij})\int uk(u)\delta u + \frac{b^2}{2} d_s{}''(x_{ij})\int u^2 k(u)du + O(b^3)\right] \quad (4.44)$$

Therefore

$$E\sum_{ij}\sum k\left(\frac{x_{ij}-x_{ik}}{b}\right) k\left(\frac{x_{ij}{}'-x_{ik}}{b}\right) = b^2\sum\sum\left[d_s(x_{ij}) + \frac{b^2}{2} d_s{}''(x_{ij})k_2\right]\left[d_s(x_{ij}) + \frac{b^2}{2} d_s{}''(x_{ij})k_2\right] \quad (4.45)$$

by the above theorem.

This leads to

$$b^2(n-1)(n-2)\left[d_s(x_{ij}) + b^2 d''(x_{ij})k_2 d_s(x_{ij}) + O(b^3)\right] \times \left[d_s(x_{ij}) + \frac{b^2}{2} d_s(x_{ij})k_2 + O(b^3)\right] \quad (4.46)$$

Equation (4.46) leads to

$$E\sum_{ij}\sum k\left(\frac{x_{ij}-x_{ik}}{b}\right) k\left(\frac{x_{ij}{}'-x_{ik}}{b}\right) = b^2\,(n-1)(n-2)\left[d_s{}^2(x_{ij}) + \frac{b^2}{2} d''(x_{ik})k_2 d_s(x_{ik}) + O(b^3)\right] \quad (4.47)$$

$$= b^2(n-1)(n-2)[k_2 d_s(x_{ik})d''(x_{ik}) + O(b^3)] \quad (4.48)$$

Using this result in equation (4.43) gives

$$E\left[\sum_{i\in s} k\left(\frac{x_{ij}-x_{ik}}{b}\right)\Big/ x_{ij}\right]^2 = (n-1)b^{-2}\left[(n-1)bd_s(x_{ij})\int k^2(u)du + bd_s(x_{ij})\int uk^2(u)\delta u + \frac{b^2}{2}d''(x_{ik})\int u^2k(u)\,du + O(b^3)b^2(n-1)(n-2)(b^2k_2d_s(x_{ik})d''(x_{ik}) + O(b^3))\right] \quad (4.49)$$

This simplifies to

$$E\left[\sum_{i\in s} k\left(\frac{x_{ij}-x_{ik}}{b}\right)\Big/ x_{ij}\right]^2 = (n-1)b^{-1}\left[d_s(x_{ij})\int k^2(u)du + bd_s{}'(x_{ij})\int uk^2(u)du + \frac{b^2}{2}d'(x_{ik})\int u^2k(u)\,du + O(b^3)\right] \quad (4.50)$$

Therefore

$$Var[d_s(x_{ij})] = \frac{(n-1)}{b}d_s(x_{ik})\int k^2(u)du + \frac{1}{n-1}d_s{}'(x_{ik})\int k^2(u)du + \frac{b}{2(n-1)}d_s{}''(x_{ik})\int u^2k(u)du + b^2\frac{(n-1)}{(n-2)}[b^2k_2d_s(x_{ik})d_s{}''(x_{ik}) + O(b^3)] - [b^2k_2d_s(x_{ik})d''(x_{ik}) + O(b^3)] \quad (4.51)$$

Thus we get

$$Var[d_s(x_{ij})] = O\left[\frac{(n-1)^{-1}}{b}\right] + O\left[\left(\frac{1}{n-1}\right)b\right] + O[b^2] + \cdots \quad (4.52)$$

Hence

$$Var\left[d_s(\hat{x}_{ij})\Big/ x_{ij}\right] = O[b^2] \quad (4.53)$$

From equation (4.20)

$$E\left(\hat{\bar{Y}}\right) = \bar{Y} + O_p\left(b^2 + b^{-\frac{1}{2}}\right) - \left[b^2\frac{d''(x_{ik})}{2d_s(x_{ik})}\right] \quad (4.54)$$

As $b \to 0$, and for n small,

$$E\left(\hat{\bar{Y}}\right) = \bar{Y} \tag{4.55}$$

And therefore the proposed estimator is asymptotically unbiased.

4.2 The Bias of the Proposed Estimator

The bias of the proposed estimator is given by

$$Bias = E\left(\hat{\bar{Y}}\right) - \bar{Y} \tag{4.56}$$

$$E\left(\hat{\bar{Y}}\right) - \bar{Y} = \frac{1}{M}\left[E\left\{\sum_{i\in s\in c}\frac{\mu(x_i)}{N_i\pi_i} + \sum_{i\in s}\frac{e_i}{N_i\pi_i}\right\} - \sum_{i\in s\in c}\frac{y_i}{N_i}\right] \tag{4.57}$$

$$= \frac{1}{M}\sum_{i\in s\in c}\frac{E\left(\mu(x_i)\right)}{N_i\pi_i} + \frac{1}{M}\sum_{i\in s}\frac{E(e_i)}{N_i\pi_i} - \frac{1}{M}\sum_{i\in s\in c}\frac{y_i}{N_i} \tag{4.58}$$

But

$$\mathrm{E}(e_i) = 0$$

Hence

$$\frac{1}{\mathrm{M}}\sum_{\mathrm{i}\in\mathrm{s}}\frac{\mathrm{E}(\mathrm{e_i})}{\mathrm{N_i}\pi_\mathrm{i}} = 0 \tag{4.59}$$

$$E\left(\hat{\bar{Y}}\right) - \bar{Y} = \frac{1}{M}\sum_{i\in s\in c}\frac{E\left(\mu(x_i)\right)}{N_i\pi_i} - \frac{1}{M}\sum_{i\in s\in c}\frac{y_i}{N_i} \tag{4.60}$$

But the inclusion probability in two-stage cluster sampling is given by

$$\pi_{i=}\frac{1}{Mn} \tag{4.61}$$

Therefore

$$\mathrm{E}\left(\hat{\bar{\mathrm{Y}}}\right) - \bar{\mathrm{Y}} = n\sum_{i\in s\in c}\frac{E\left(\mu(x_i)\right)}{N_i\pi_i} - \frac{1}{M}\sum_{i\in s\in c}\frac{y_i}{N_i} \tag{4.62}$$

We have

$$\sum_{i\in s}\frac{1}{N_i} = \frac{1}{N_1} + \frac{1}{N_2} + \cdots + \frac{1}{N_n} = \frac{n}{N}; \tag{4.63}$$

Since

$$\textstyle\sum_i N_i = N$$

$$E\left(\hat{\bar{Y}}\right) - \bar{Y} = \frac{n^2}{N}\sum_{i\in s\in c}\frac{E\left(\mu(x_i)\right)}{\pi_i} - \frac{n}{MN}\sum_{i\in s\in c}y_i \tag{4.64}$$

As *n* becomes small the bias asymptotically tends to zero i.e

$$E\left(\hat{\bar{Y}}\right) - \bar{Y} \to 0 \tag{4.65}$$

4.3 Asymptotic Error Variance of the Proposed Estimator

$$Error\ Variance = Var\left[E\left(\hat{\bar{Y}}\right) - \bar{Y}\Big/_{x_{ij}}\right] \tag{4.66}$$

$$= VarE\left(\hat{\bar{Y}}\right) + Var(\bar{Y}) \tag{4.67}$$

$$= VarE\left(\hat{\bar{Y}}\right) + \sigma^2(x_{ij}) \tag{4.68}$$

due to i.i.d of auxiliary variables.

But

$$VarE\left(\hat{\bar{Y}}\right) = var\left[\frac{1}{M}\left(\frac{n}{n-1}\right)^{1/2}\sum_{i\in s}\left[\frac{1}{N_i}E\left(\mu(x_{ij})\right)\right]k\,\frac{(x_{ij}-x_{ik})}{b}\left[d_s(\hat{x}_{ij})\right]^{-1}\right] \tag{4.69}$$

$$= \frac{1}{M^2}\left(\frac{n}{n-1}\right)\sum_{i\in s}\frac{1}{{N_i}^2}Var\left[k(u)\left[d_s(\hat{x}_{ij})\right]^{-1}\right]Var\left(\mu(x_{ij})\right) \tag{4.70}$$

$$= \frac{1}{M^2}\left(\frac{n}{n-1}\right)\sum_{i\in s}\frac{1}{{N_i}^2}Var\left[k(u)\left[d_s(\hat{x}_{ij})\right]^{-1}\right]\sigma^2(x_{ij}) \tag{4.71}$$

Substituting equation (4.71) in equation (4.68) we get

$$Var\left[E\left(\hat{\bar{Y}}\right) - \bar{Y}\Big/_{x_{ij}}\right] = \sigma^2(x_{ij}) + \frac{1}{M^2}\left(\frac{n}{n-1}\right)\sum_{i\in s}\frac{1}{{N_i}^2}Var\left[k(u)\left[d_s(\hat{x}_{ij})\right]^{-1}\right]\sigma^2(x_{ij}) \tag{4.72}$$

But as shown above, the conditional variance of $d_s(\hat{x}_{ij})$ is given by (4.53) i.e

$$Var\left[d_s(\hat{x}_{ij})\Big/_{x_{ij}}\right] = O[b^2]$$

Which tends to zero as $b \to 0$, and hence the second term on the right of equation (4.72) vanishes. Therefore the error variance becomes

$$Var\left[E\left(\hat{\bar{Y}}\right) - \bar{Y}\Big/_{x_{ij}}\right] = \sigma^2(x_{ij}) \tag{4.73}$$

Table 1: Summary of the Mean Squared Error Values (MSE)

The table below shows the results of the mean squared errors (MSE) of the estimators of the finite population mean in two stage cluster sampling at different bandwidths.

Bandwidth	Model-Assisted	Model-Based
0.001	0.006646769	4.328651
0.002	0.001167328	4.071455
0.003	0.09797738	5.799239
0.004	0.2305878	4.33418
0.005	0.1958319	6.575275
0.006	0.1321953	5.486596
0.007	1.758383	4.897115

From **Table 1** above we observe that at different bandwidths considered, the mean squared errors obtained in estimating population mean using the model-assisted estimator are significantly smaller than those incurred in using model-based estimator when the sample size is small i.e when $n < 30$ as used in the empirical study.

Table 2: Summary of the bias for the Two Estimators

Bandwidth	Model-Assisted	Model-Based
0.001	0.02105037	0.4000981
0.002	0.02789657	0.05316963
0.003	-0.08081971	-0.3859103
0.004	-0.123986	-0.1279632
0.005	-0.1142605	0.001984305
0.006	-0.1442672	-0.189994
0.007	-0.3423822	0.5324499

From **Table 2** above we observe that the bias associated with the model assisted estimator increases insignificantly while that of model based estimator fluctuates. However, the bias due to model assisted estimator is comparatively low with respect to that of model based estimation at the different bandwidths considered.

It can be noted also that there is a trade-off between these two approaches as illustrated by the graphs. So none of the approaches is overally better but they can be compared at specific bandwiths.

5. Conclusion

The main aim of the paper was to estimate a finite population mean using model assisted approach in two stage cluster sampling when the sample size is small i.e $n < 30$. The empirical study compared the performance of a model based estimator of the finite

population mean with the proposed estimator. Empirical results show that model-assisted approach performed better than model-based approach when the sample size is small. Comparison was done on the basis of the mean squared errors and biases of the two estimators.

References

1. Dorfman, A.H (1992). "Non-Parametric Regression for Estimating Totals in Finite Populations". Proceedings of the Section on Survey Research Methods, American Statistical Association 622-625.
2. Breidt, F.J., and Opsomer, J. D. (2000). Local Polynomial Regression Estimators in Survey Sampling. The Annals of Statistics, 28, 1026-1063.
3. Ouma *et al* (2010). Generalized Model Based Confidence Intervals in Two Stage Cluster Sampling. Far East Journal of Statistics, 171-184.
4. Shewhart, A.W (2004). Properties and Use of Shewhart Method. Sequential Analysis, 2007, 26, 171-193.
5. Ruppert, *et al* (2003), *Transformations and Weighting In Regression*, London: Chapman and Hall.
6. Rao, J. N. K. and Wu, C. F. J. (1998): Resampling Inference with Complex Survey Data Journal of the American Statistical Association 83, 231-241.