2019

Estimating a finite population mean under random non-response in two-stage cluster sampling with replacement.

Bii, Nelson Kiprono
Strathmore Graduate School
Strathmore University

ESTIMATING a FINITE POPULATION MEAN UNDER RANDOM NON-RESPONSE IN TWO-STAGE CLUSTER SAMPLING WITH REPLACEMENT

NELSON KIPRONO BII

Submitted in total fulfillment of the requirements for the Degree of Doctor of Philosophy in Mathematical Statistics of Strathmore University

Strathmore Institute of Mathematical Sciences

Strathmore University

June, 2019

DECLARATION

I declare this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the proposal contains no material previously published or written by another person except where due reference is made in the proposal itself.

Nelson Kiprono Bii

.....

.....

18/06/2019

Approval

The thesis of **Nelson Kiprono Bii** was received and approved by the following:

Dr. Christopher Ouma Onyango,
Senior Lecturer, Department of Statistics and Actuarial Sciences,
Kenyatta University

Prof. John W. Odhiambo,
Strathmore Institute of Mathematical Sciences,
Strathmore University

Ferdinand Othieno,
Dean, Strathmore Institute of Mathematical Sciences,
Strathmore University

Prof. Ruth Kiraka,
Dean, School of Graduate Studies,
Strathmore University

Abstract

Non-response is a regular occurrence in sample surveys. Developing estimators when non-response exists may result in large biases when estimating population parameters. In this study, a finite population mean estimator has been developed under two stage cluster sampling with replacement in the presence of random non-response. It is assumed that non-response arises in the survey variable in the second stage of cluster sampling assuming full auxiliary information is known. Weighting method of compensating for non-response has been applied. Kernel density estimation was used and the modified transformation of data method was incorporated in order to address the boundary effects due to Nadaraya-Watson estimator used in the estimation process. Asymptotic properties of the proposed estimator of the finite population mean have been derived. The performance of the proposed estimator has been compared with other estimators based on bias, mean squared error and confidence interval lengths using simulated data. The results revealed that the estimator proposed has smaller mean squared error values and shorter confidence interval lengths when compared to other estimators of the finite population mean. The bias results also indicated that the proposed estimator of finite population mean performed better than the Nadaraya-Watson and the improved Nadaraya-Watson estimators. The transformed estimator proposed to address boundary bias due to Nadaraya-Watson has also been shown to have smaller values of the bias, smaller mean squared error values and shorter confidence interval lengths compared to those of Nadaraya-Watson estimator. The results obtained can be useful in choosing efficient estimators of finite population mean for instance in demographic health sample surveys.

Contents

List of Figures

List of Tables

Abbreviations

SRS	simple random sampling
$SRSWR$	simple random sampling with replacement
var	variance
MSE	mean squared error
$p.d.f$	probability density function
KDE	kernel density estimate
$\hat{M}_{NW}(.)$	mean function of Nadaraya-Watson estimator
$PSUs$	Primary sampling units
$SSUs$	Secondary sampling units
CI	Confidence Interval
$N-W$	Nadaraya-Watson
TDM	Transformation of Data Method

Acknowledgment

First and foremost I thank Almighty God for the gift of life and good health throughout this journey. This work would not have been possible without the scholarship of Strathmore University in the Strathmore Institute of Mathematical Sciences. The Strathmore University Doctoral Academy gave me a very elaborate personal and professional guidance and taught me a great deal about both scientific research and life in general. I am especially indebted to Dr. Christopher Ouma Onyango and Professor John Odhiambo who have been very supportive of my career goals. As my lecturer and mentor, Dr. Ouma Onyango has taught me more than I could ever give him credit for here. He has shown me, by his example, what a good scientist (and person) should be.

I am grateful to all those with whom I have had the pleasure to work during my doctoral studies. I would especially like to thank Professor Vitalis Onyango Otieno and Professor Rachel Waema Mbogo who were a source of inspiration to me in many ways. The collegiality I had with my peers such as Raymond Calvin Ochieng, Paul Anthony Otieno, Titus Orwa and Evans Otieno was very inspirational during my doctoral studies. I am greatly indebted to them.

I would like to thank my family members who have been a source of inspiration in this course. Fredrick and Nicholas, my dear elder brothers have been my pillars of hope in this journey. I salute you. They are my ultimate role models. Most importantly, I wish to thank my loving and supportive wife, Mercy and my wonderful children Brian Kiptoo and Olivia Cheptoo for providing unending inspiration.

Dedication

This thesis is dedicated to my lovely wife Mercy, my dear son Brian Kiptoo and my beautiful daughter Olivia Cheptoo. I love you all. May God bless you abundantly.

CHAPTER ONE

INTRODUCTION

This chapter presents the background of the study, statement of the problem and the objectives of the study. The outline of the thesis is also highlighted in this chapter.

1.1 Background of the Study

There are many situations in practise that necessitate estimation of population parameters. Some of these parameters include but not limited to population variance, population mean or total and population proportions. Estimation of finite population mean is the subject of this study.

Statisticians need to assess the properties of different estimators so that better ones can be adopted. More often, these properties include minimal variance, unbiasedness and asymptotic mean squared error. A good sampling strategy can allow one to obtain an estimator that possess these properties. Auxiliary information has been used in this research in an attempt to produce better results in estimation of a finite population mean in presence of random non-response.

Research by statisticians have shown that failure to compensate for non-response of data can affect inference, (Khare, 2013). Hansen and Hurwitz (1946) considered the problem of estimating population mean under non-response. The proposed estimator was given by

$$\bar{y}^* = \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n}$$

where n is the sample size selected from a population of size N whereas n_1 are units that respond and n_2 are units that fail to respond in the survey ; $\bar{y}_1$ is the sample mean of the response group while $\bar{y}_2$ is the sample mean of the non-response group. $\bar{y}^*$ is an unbiased estimator for the population mean assuming that $\bar{y}_1$ and $\bar{y}_2$ are unbiased for the population mean of the response and non-response groups respectively, and has variance given by

$$var(\bar{y}^*) = \Big(\frac{1}{n} - \frac{1}{N}\Big)S_{y_1}^2 + (f-1)\frac{N_2}{N}.\frac{S_{y_2}^2}{n}$$

where $S_{y_1}^2$ is the population variance of the response group, $S_{y_2}^2$ is the population variance of the non-responding units of the population, N_2 is the population size of the non-response group and f is the finite population correction given by $\frac{n}{N}$.

El-Badry (1956) extended (Hansen and Hurwitz, 1946) methodology and further indicated how sub-sampling can be used to reduce the problem of non-response. Several authors such as (Anderson, 1979), (Holt and Elliot, 1991), (Singh and Kumar, 2009) and (Singh and Kumar, 2010) have also studied the problem of non-response under simple random sampling using various assumptions.

Onyango et al. (2010), in improving estimation of population parameters, developed estimators of population total in two stage cluster sampling in large sample approximations using model-based method. Odhiambo and Onyango (2014) considered the estimation of population total under non-response by modifying and symmetrizing Murthy's estimator given by

$$\hat{Y}_m = \frac{\sum_{i=1}^{n} P(s/i)}{P(s)}$$

where $\hat{Y}_m$ is an estimator of population total based on ordered samples $s(i)$, y_i refers to the i^{th} unit in the population, $P(s/i)$ is the conditional probability of getting the set of units that was drawn given that the i^{th} unit was drawn first, $P(s)$ is the unconditional probability of getting the set of units that was drawn such that $0 < P(s) \leq 1$. The study obtained the variance and bias of the calibrated estimator as

$$var\left(t_w^*/R\right) = \sum_{i=1}^{k} \left(N\frac{n_c}{n}\right)^2 \frac{N_c - m_c}{N_c m_c} S_c^2$$

$$Bias(t_w^*) = \sum_{i=1}^{k} \left(\frac{N}{n} - \frac{N_c}{n_c}\right)\hat{t}_{sc}$$

where N_c is the population size in class c, m_c is the number of units with complete data, N is the population size given by $N = \sum_{c=1}^{k} N_c$ partitioned into k classes such that $m = \sum_{c=1}^{k} m_c$, $n = \sum_{c=1}^{k} n_c$ where n_c is the number of units with complete data in the sample selected, S_c^2 is the population variance given by $S_c^2 = \frac{1}{N_c - 1}\left[\sum_{i=1}^{N_c} Y_i^2 - N\bar{Y}_c^{\,2}\right]$ and R is a vector of samples with complete data; t_w^* is a sequence of point estimators

given by $t_w^* = \left(y_{c1}, \cdots, y_{cm_c}\right)$.

The estimator of (Hansen and Hurwitz, 1946) has been extended by (Bahl and Tuteja, 1991) by introducing an exponential ratio-type estimator for population mean given by

$$\bar{y}_{er} = \bar{y}_{exp}\left(\frac{\bar{X} - \bar{x}}{\bar{X} + \bar{x}}\right)$$

where $\bar{X}$ and $\bar{x}$ are the population mean and sample mean of the auxiliary variables respectively, $\bar{y}_{er}$ is the exponential ratio-type estimator of the population mean $\bar{y}$. The bias of the estimator is given by

$$Bias\left(\bar{y}_{er}\right) = \left(\frac{1}{n} - \frac{1}{N}\right)\bar{Y}\left(\frac{3c_x^2}{8} - \frac{\rho C_y C_x}{2}\right)$$

whereas the MSE of the estimator is given by

$$MSE\left(\bar{y}_{er}\right) = \left(\frac{1}{n} - \frac{1}{N}\right)\bar{Y}^2\left(C_y^2 + \frac{C_x^2}{4} - \rho C_y C_x\right)S_y^2 + \frac{(f-1)N_2}{nN}S_{y2}^2$$

where S_y^2 is the population variance of the response group and S_{y2}^2 is the population variance of the non-response group of the population; N_2 is the population stratum of the non-response group and f is the finite population correction given by $\frac{n}{N}$, n is the sample size selected from a population of size N, $\bar{Y}$ is the population mean to be estimated, ρ is the correlation coefficient of all stratified data, C_x and C_y are the coefficients of variation for auxiliary and survey variables respectively.

However, this study adopted cluster sampling in two stages when non-response occurs in the second stage of sampling of study variables assuming full auxiliary random values are known in the population.

1.2 Statement of the Problem

Many authors have developed estimators for finite population mean where non-response exists in the study and auxiliary variables, for details see (Singh and Kumar, 2008), (Singh et al., 2010) and (Ismail et al., 2011). Besides, (Singh and Kumar, 2010) assumed that non-response occurs in the second stage of sampling in both the study and auxiliary variables in the estimation of finite population mean. However, there exist

cases that do not exhibit non-response at all in the auxiliary variables. In a manufacturing survey, for instance, the number of employees can be used as an auxiliary variable for the estimation of the mean of items produced in bundles. Information can be obtained completely in the number of employees while there may be non-response on the average amount of items produced in bundles. Therefore the assumption that non-response exists in both the auxiliary and survey variable may not always hold. Singh et al. (2011) noted that ratio, product and regression are examples of estimators that utilize auxiliary information in the estimation of finite population mean under non-response. The study proposed a combination of regression and ratio estimators of finite population mean. However, these estimators still need improvements in terms of efficiency and bias reduction. In the sequence of improving the estimators of finite population mean under random non-response, (Khan et al., 2014) proposed a class of estimators that uses fractional raw moments of auxiliary variables and noted that the efficiency of the estimators is increased only if the raw moments of the auxiliary variables are available. This study therefore proposed the use of auxiliary information at the estimation stage via a regression model in two stage cluster sampling with replacement. Kernel weights are used to compensate for non-response. A finite population mean is estimated assuming non-response is observed at random in the survey variable in the second stage of cluster sampling assuming full auxiliary information is available at both stage one and two.

1.3 Objectives of the Study

1.3.1 Main Objective

To estimate the mean of a finite population under random non-response in two stage cluster sampling with replacement.

1.3.2 Specific Objectives

1. To develop a mean estimator of finite population in the presence of random non-response using two stage cluster sampling method.

2. To derive the boundary bias correction of mean estimator of a finite population using the weighting method under random non-response in two stage cluster sampling with replacement.

3. To perform a comparative study of the proposed estimator, transformed estimator of the finite population mean and Nadaraya-Watson estimator using the values of the bias, mean squared error values and the confidence interval lengths.

1.4 Significance of the Study

Suppose a researcher wishes to estimate the average health insurance coverage in Nairobi County. One could take a random sample of say, 200 households. Therefore a sampling list of Nairobi households is required. If the list is not available, a census needs to be conducted. The complete coverage of Nairobi County is required so that all the households are listed; this could be very expensive. Besides, since the sample size selected is smaller in comparison to the total number of households, only a few individuals say three or four in each block need to be sampled.

Alternatively, a researcher could select ten blocks assuming the county is made up of three hundred blocks and in each block thirty households are interviewed. A household listing frame for only ten blocks need to be constructed; this is economical both temporally and financially. A list of non-respondents is then drafted so that a suitable mechanism for compensating for the non-response can be put in place. To compensate for random non-response, kernel weights are used in this study. The weights help to improve on the precision of the estimator of the finite population mean.

1.5 Scope of the Study

Estimation of finite population parameters involves mostly, either parametric or non-parametric approaches. Model-assisted non-parametric kernel regression technique is considered in this study. Two stage cluster sampling with replacement is used in presence of random non-response. Kernel weights are used to compensate for the random non-response. In particular, the Nadaraya-Watson kernel regression technique is used to develop the estimator of the finite population mean. To reduce the edge effects associated with the Nadaraya-Watson regression technique, modified transformation of data method estimator due to (Marron and Ruppert, 1994) has been proposed in this study and its asymptotic properties derived. The efficiency of the proposed estimator, transformation of data method estimator, Nadaraya-Watson and improved Nadaraya-Watson due to (Demir and Toktamiş, 2010) have been compared in terms of bias, MSE

and confidence interval lengths using simulated data.

1.6 Outline of the Thesis

This thesis has been ordered in terms of chapters as follows: The statement of the problem and objectives of the study are presented in chapter one. Literature review is discussed in chapter two. The proposed estimator and its properties have been presented in chapter three. Chapter four presents the method of reducing boundary bias arising from Nadaraya-Watson estimator. Description of the simulation experiment, presentation and discussion of the results have been done in chapter five. Areas for further research have also been suggested in chapter five.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

This chapter presents a review of non-response and methods of compensating for non-response. Estimation of Kernel Density Functions and Regression Functions is also reviewed. Methods of reducing boundary bias and methods of selecting a smoothing parameter are also discussed in this chapter. A review of how to handle the error terms of asymptotic expansions in *order* algebra is also presented in this chapter.The chapter ends with a brief summary of the literature reviewed.

2.2 Review of Non-Response

In survey sampling, non-response is one source of errors in data analysis. Non-response introduces bias in the estimation of population characteristics. It also causes samples to fail to follow the distributions determined by the original sampling design. This study seeks to reduce the non-response bias in estimating a finite mean of population in two stage cluster sampling. The use of regression models is recognized as one of the procedures for reducing bias due to non-response using auxiliary information, for details see (Andridge and Little, 2010). In practise, information on the variables of interest is not available for non-respondents but information on auxiliary variables may be available for non-respondents. It is therefore desirable to model the response behaviour and incorporate the auxiliary data into the estimation so that the bias arising from non-response can be reduced. If the auxiliary variables are correlated with the response behaviour, then the regression estimators would be more precise in estimation of population parameters, given the auxiliary information is known.

Imputation techniques have been used to account for non-response in the study variable. For instance, (Singh and Horn, 2000) applied compromised method of imputation to estimate a finite population mean under two stage cluster sampling, this method

however produced a large bias. In this study, the Nadaraya-Watson regression technique due to (Nadaraya, 1964) and (Watson, 1964) is applied in deriving the estimator for the finite population mean. Kernel weights are used to compensate for non-response.

A lot of significance is attached to efficient and cost-effective survey sampling designs in sample surveys, see for instance (Fife-Schaw, 2000). Therefore, careful design of samples based on random selection with known probabilities of population elements should be considered. This gives a target sample of intended respondents where each may provide responses to a set of survey questions that results in an array of responses. Van Buuren et al. (1999) observed that non-response occurs if some of the expected responses are missing, for instance where a whole vector of responses is missing for some sampled units or where responses are obtained for some questions and not to others in the sample selected.

The basis for statistical inference is therefore formed by a sampling design that provides a link between a sample and the population. As observed by (Holt and Elliot, 1991), a good sample survey practise is a very important approach to non-response. Besides, (Holt and Elliot, 1991) noted that collection of the intended data, if possible, is more important than to rely on subsequent methods of adjustment and compensating for non-response in the analysis stage.

Non-response is a probable source of error in sample survey estimation as observed by (Bound et al., 2001). It shows a variation from the probability sampling design. Non-response is common in most complex surveys and if the attributes of the non-responding units differ from those that respond, direct estimates obtained on the basis of the respondents will be biased. The primary goal in the analysis of survey data, therefore, is to reduce the bias due to non-response.

Various terms have been used to describe non-response by different researchers. The term 'non-response' has been adopted by (Anderson, 1979; Fuller et al., 1994; Madden et al., 1996) ; 'Missing data' is used by (Ford, 1976; Kalsbeek, 1980) ; 'Incomplete samples' is used by (Singh, 1985) among others. In this research, 'non-response' is used to explain the failure to get data from an element in a chosen sample. Under

weighting strategy, the original sampling weights for responding units are increased by use of the response probability. Imputation techniques compensate for non-response values by imputed values, (Kalton and Kasprzyk, 1982).

As observed by (Kalsbeek et al., 1980) and (Kim and Kim, 2007), two general models for non-response are available, that is, stochastic and deterministic. A deterministic model partitions a population of N elements into two mutually exclusive and exhaustive strata that is, N_1 elements responding, if selected in a sample and N_0 elements failing to respond, if selected in a sample, that is, $N = N_0 + N_1$, (Anderson, 1979).

In a stochastic model, the response probabilities, $p_i (i = 1, \cdots, N)$ may be any value between 0 and 1, whereas in deterministic model there is either $p_i = 0$ or $p_i = 1$. A stochastic model may be a more realistic model for a survey since the response probabilities are random variables. The response probabilities are mostly taken to be correlated with certain characteristics such as age, race and income for a human population survey, (for details see (Liang and Zeger, 1993)). Bethlehem (1988) and (Kim and Kim, 2007) discussed a modified Horvitz-Thompson estimator to correct for non-response problem using the weighting strategy. The estimator used was defined by

$$\hat{\bar{y}}_{HT} = N^{-1} \sum_{k=1}^{N} \left(\phi_k \hat{\bar{p}}_{HT}\right)^{-1} y_k \tau_k$$

where $\hat{\bar{p}}_{HT}$ is given by

$$\hat{\bar{p}}_{HT} = N^{-1} \sum_{k=1}^{N} \phi_k^{-1} \gamma_k$$

where $y_k, k \in U$ is the value of the k^{th} survey variable taken from a sample s selected from a finite population, $U = \left(1, 2, \cdots, N\right)$, ϕ_k is the inclusion probability given by $\phi_k = p_r(k \in s)$, γ_k is the value of the k^{th} respondent in the sample selected, s. The estimator $\hat{\bar{y}}_{HT}$ adjusts the weights by an unbiased estimator $\hat{\bar{p}}_{HT}$, of the response probabilities of the population mean which is given by

$$\bar{P}_N = N^{-1} \sum_{i=1}^{N} p_i$$

Thus an approximate bias of the estimator $\hat{\bar{y}}_{HT}$ is given by

$$Bias\left(\hat{\bar{y}}_{HT}\right) = E(\hat{\bar{y}}_{HT}) - \bar{Y} = \bar{Y}^* - \bar{Y} = \hat{\bar{P}}_N^{-1} C_p Y$$

where

$$\bar{Y}^* = N^{-1} P_N^{-1} \sum_{i=1}^{N} p_i y_i$$

and

$$C_p Y = N^{-1} \sum_{i=1}^{N} \left(p_i - \bar{P}_N\right)\left(y_i - \bar{Y}\right)$$

If $C_p Y$ is close to zero, the bias will be small, for more details see (Särndal, 1980). The adjusted Horvitz-Thompson estimator, $\hat{\bar{y}}_{HT}$, is an illustration of re-weighting measurements of respondents without using auxiliary information. This study uses auxiliary information in the estimation procedure. In what follows, a review of imputation methods for non-response are discussed.

2.2.1 Imputation Methods for Non-Response

Imputation entails compensating for non-response values by proxy values as observed by (Särndal and Lundström, 2005). Following the procedure by (Durrant et al., 2005), let $\overline{Y} = \frac{1}{N}\sum_1^N Y_i$ represents the finite population mean to be estimated. Besides, let a simple random sample, say, S with replacement be drawn from the population, $\theta = 1, 2, \ldots, N$ to estimate Y. The sample S of units has r responding units $(r < n)$ making a set R and $(n - r)$ non-responding units with the subspace $(n - r)$ having the symbol R^C in the population. For every unit $i \in R$, the value of the survey variable y_i is obtained.

However, imputed values are to be derived for the non-response set of units, that is, for every $i \in R^C$, since the y_i values are missing. It is assumed that auxiliary data x_i are known for every $i \in S$. The value of the auxiliary, x_i, imputes the non-response values when $i \in R^C$, that is, for a sample S assume that the data $x_s = x_i : i \in s$ are given and $S = R \cup R^C$. Using this set up, ratio method of imputation, an example of different imputation techniques, can be defined as in equation (2.1). Using the notations of (Rancourt et al., 1994), if the i^{th} unit is to be imputed, the value $\hat{b}x_i$ is obtained, where

$\hat{b}x_i = \frac{\sum_{i\in R} y_i}{\sum_{i\in R} x_i}$. The data after imputation becomes

$$y_{\cdot i} = \begin{cases} y_i, \text{ if } i \in R \\ \hat{b}x_i, \, i \in R^C \end{cases}$$

For details, see (Singh and Horn, 2000). The imputation ratio estimator is given by

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

where $\overline{x}_n = \frac{1}{n}\sum_{i\in s} x_i$, $\overline{x}_r = \frac{1}{r}\sum_{i\in R} x_i$ and $\overline{y}_r = \frac{1}{r}\sum_{i\in R} y_i$. The bias and the mean squared error of the imputation ratio estimator due to (Singh and Horn, 2000) are given below.

$$B(\overline{y}_{RAT}) = (\frac{1}{r} - \frac{1}{n})\overline{Y}\Big(C_x^2 - \rho C_x C_y\Big)$$

where $C_x = \frac{s_x}{\overline{X}}$, $C_y = \frac{s_y}{\overline{Y}}$ and $\rho = \frac{S_{xy}}{S_x S_y}$. S_x and S_y are the standard deviations of X and Y values respectively while S_{xy} is the co-variance between X and Y. The MSE is given by

$$MSE(\overline{y}_{RAT}) = (\frac{1}{n} - \frac{1}{N})S_y^2 + \Big(\frac{1}{r} - \frac{1}{n}\Big)\Big[S_y^2 + R_1^2 S_x^2 - 2R_1 S_{xy}\Big]$$

where $R_1 = \frac{\overline{Y}}{\overline{X}}$. Ahmed et al. (2006) observed that though this method of imputation is better than other existing techniques like mean and compromised method of imputation, its bias and MSE are still large compared to rival approaches of compensating for non-response such as weighting techniques.

2.2.2 Re-weighting Method

It has been observed by authors such as (Pike, 2008) that non-response causes loss of observations and therefore re-weighting means that the weights are increased for all or almost all of the elements that fail to respond in a survey. The population mean, $\bar{Y}$, is estimated by selecting a sample of size n at random with replacement. If the responding units of item y are independent so that the probability of unit j responding in cluster i is $p_{ij} (i = 1, 2, \cdots, n; j = 1, 2, \cdots, m)$, then following the work of (Rao,

1996), an imputed estimator, $\hat{\bar{y}}_I$, for $\bar{Y}$ is given by

$$\hat{\bar{y}}_I = \frac{1}{\sum_{i,j\in s} w_{ij}} \left[\sum_{i,j\in s_r} w_{ij} y_{ij} + \sum_{i,j\in s_m} w_{ij} y^*_{ij} \right]$$

where $w_{ij} = \frac{1}{\pi_{ij}}$ gives the survey weight tied to unit j in cluster i and $\pi_{ij} = p[i, j \in s]$ is its probability of inclusion, s_r, is the set of r responding units to item y, s_m is the set of m units that failed to respond to item y so that $r + m = n$ while y^*_{ij} is the value imputed so that the missing value y_{ij} is compensated for, (Andridge and Little, 2010).

2.2.3 Estimation of Kernel Density Functions and Regression Functions

The Rosenblatt-Parzens conventional kernel density estimator (KDE) is one of the earliest in statistical literature. It is based on a random sample $x_1, \cdots, x_n$ drawn from a population with unknown density function, $f(x)$, for details see (Silverman, 1986), (Wand and Jones, 1994a) among others. The estimator is given by

$$\hat{f}(x) = \frac{1}{nb} \sum_{i=1}^{n} K\Big(\frac{x - x_i}{b}\Big)$$

where b is the bandwidth such that as $n \to \infty$, $b \to 0$ while $nb \to \infty$ and $K(.)$ is a kernel function. The choice of the bandwidth has been demonstrated widely for example in the works of (Baszczyńska, 2015).

In the regression models of the form $Y_i = m(x_i) + \epsilon_i$, for $i = 1, 2, \cdots, n, n \in N, E(\epsilon_i) = 0, var(\epsilon_i) = \sigma^2 > 0$, the approximation of an unknown function $m(.)$ can be done using kernel estimation with a bandwidth b and a kernel function K. Most commonly used kernel regression estimator is the Nadaraya-Watson estimator given by

$$m(\hat{x_i}) = \sum_{i\in s} w(x_i) Y_i = \frac{\sum_{i\in s} K\left(\frac{x - X_i}{b}\right) Y_i}{\sum_{i\in s} K\left(\frac{x - X_i}{b}\right)},$$

$$i = 1, 2, \cdots, n$$

An improvement of the Nadaraya-Watson estimator has been proposed by (Demir and Toktamiş, 2010) using local bandwidth factor λ_i determined using (Abramson, 1982) and (Läuter, 1988) algorithm. The improved Nadaraya-Watson (INW) estimator of

$m(x_i)$ is given by

$$\hat{m}_{INW}(x) = \frac{\sum_{i=1}^{n} \frac{Y_i}{\lambda_i} K\left(\frac{x-X_i}{\lambda_i b}\right)}{\sum_{i=1}^{n} \frac{1}{\lambda_i} K\left(\frac{x-X_i}{\lambda_i b}\right)} \tag{2.1}$$

where b is a smoothing parameter whereas

$$\lambda_i = \left\{\hat{f}(X_i)/a\right\}^{-\alpha}$$

where a is given by $a = \sum_{i=1}^{n} \hat{f}(X_i)/n$ while α is a parameter satisfying $0 \leq \alpha \leq 1$. Abramson (1982) and (Läuter, 1988) suggested that taking $\alpha = \frac{1}{2}$ produce good results. For details on the asymptotic properties of $m_{INW}(\hat{x}_i)$, see (Demir and Toktamiş, 2010). The improved Nadaraya-Watson estimator in equation (2.1) can be used to estimate $m(x)$, an unknown function of auxiliary random variables.

Kernel estimates give a smooth curve and also have advantage over histogram in terms of bias. The bias of a histogram estimator with bin width b is of order b, $o(b)$ while centering the kernel at each data point and using a symmetric kernel reduces this term to zero and therefore produces a bias term for the kernel estimate of order b^2, $o(b^2)$.

2.3 Methods of Reducing Boundary Bias in Estimation of Population Parameters

Estimation of population parameters, for example the population mean using kernel density estimators in presence of non-response often lead to bias due to boundary effects. This affects the optimality of the estimators for the population parameters. To address this problem, (Sheather and Jones, 1991) suggested the use of optimal bandwidth. However, this does not eliminate the boundary bias. Weighting method of compensating for non-response in two stage cluster sampling is proposed in this study using modified transformation of data method. This is because the boundary correction estimates obtained from transformation of data method are non-negative and have low variance, for details see (Wand and Jones, 1994a) and (Marron and Ruppert, 1994). The values of the auxiliary variables X_{ij} are assumed to be known for all the clusters while the values of the survey variable Y_{ij} are only known for response units in the sample selected.

Let y_{ij} represents the values of the survey variable Y_{ij} for unit i in cluster j, for $j = 1, 2, ..., N; i = 1, 2, ..., M$. The problem is to estimate the survey values of the non-response component in the second stage of sampling in the selected sample. This is done by first generating data using a regression model applied by (Ouma and Wafula, 2005) and (Onyango et al., 2010). The model is given by

$$\hat{Y}_{ij} = m(\hat{x}_{ij}) + \hat{e}_{ij} \tag{2.2}$$

where $m(.)$ is a smooth function of the auxiliary variables and $\hat{e}_{ij}$ is the residual term with mean zero and variance which is strictly positive. Auxiliary data is assumed to be known throughout the study and is therefore used to predict the non-response values.

In the following sections, different methods of reducing boundary effects when estimating the non-response values of the survey variable Y_{ij} using a function of the auxiliary data, $m(.)$, are discussed.

Let g be a probability density function with the support $[0, \infty)$ and consider non-parametric estimator of g from a random sample $X_{ij}, j = 1, 2, ..., n; i = 1, 2, ..., m$ from g. The kernel estimator of g due to (Marron and Ruppert, 1994) is given by

$$\hat{g}_{mn}(x) = \frac{1}{mnb} \sum_{i=1}^{n} \sum_{j=1}^{m} K\left(\frac{x_{ij} - X_{ij}}{b}\right) \tag{2.3}$$

such that K is a specified kernel density function, symmetric about zero over the interval $[-1, 1]$ with b as the bandwidth such that $b \to 0$ as $mn \to \infty$. The properties of $\hat{g}_{mn}$ under some smoothness assumptions for $x_{ij} \geq 0$ are:

$$E\big(\hat{g}_{mn}(x)\big) = g(x) \int_{-1}^{c} k(w)dw - bg\prime(x) \int_{-1}^{c} wk(w)dw + \frac{b^2}{2} g\prime\prime(x) \int_{-1}^{c} w^2 k(w)dw + o\big(b^2\big)$$

$$Var\big(\hat{g}_{mn}(x)\big) = g(x)(mnb)^{-1} \int_{-1}^{c} k^2(w)dw + o\left(\frac{1}{mnb}\right)$$

where $c = min\big(\frac{x}{b}, 1\big)$. For $x \geq b$, that is, the interior points, the bias of $\hat{g}_{mn}(x)$ is of order $o\big(b^2\big)$. However, at the boundary points, that is for $x \in [0, b)$, $\hat{g}_{mn}(x)$ is not consistent. In non-parametric curve estimation problems, this phenomenon is called

"boundary effects" of the estimator in equation (2.3).

2.4 Methods of Reducing Boundary Bias due to Nadaraya-Watson Estimator

Generally, kernel density estimators are known to be inconsistent when estimating a density near the finite end points of the support of the density to be estimated, (Karunamuni and Alberts, 2005). This is due to the boundary effects that occur in non-parametric curve estimation. The estimator proposed by (Bii et al., 2018) suffers from boundary problem induced by Nadaraya-Watson estimator used. Notably, it is often desirable to have an optimal bandwidth to balance the bias-variance trade-off. If K is a symmetric density function constant across estimation support, then the inference is generally simplified for unbounded support, i.e $(-\infty, \infty)$, (Härdle, 1990). But the function $m(\hat{x}_{ij})$ is not consistent at the boundary $[0, b)$ where b is the bandwidth for such a choice of K, for details see (Silverman, 1986) and (Malec and Schienle, 2014). So for $x \in [0, b)$, the bias is of order $o(b)$ instead of order $o(b^2)$ at the boundary points. To reduce the boundary effects in kernel density estimation, techniques such as data reflection, pseudo data method, boundary kernel method and transformation of data method have been proposed in literature. Discussion of each one of these methods is given below.

2.4.1 Reflection of Data Method

This is one of the frequently used methods in practise for bias reduction in kernel density estimation, (Baszczyńska, 2015). The modification of the conventional KDE consists in isolating that part of the kernel function that is outside the interval of the support of the random variable and then on its symmetrical reflection. This reflection is done in relation to the boundary support, for instance $[a; \infty)$ for left boundary of the support and $(-\infty; b]$ in case of right boundary of the support. This method has been explored by (Silverman, 1986), (Jones, 1993), (Simonoff, 2012) and (Ivanka et al., 2012). It is also known as "data-reflected technique". To apply this method, one has to add $-X_i, i = 1, \cdots, n$ to the data set. Since the kernel penalizes for lack of data on the negative axis, the estimator therefore gradually applies reduced amount of data in its

window as it approaches the boundary thus resulting in a boundary bias; the addition of $-X_i, i = 1, 2, ..., n$ compensates for the lack of data. The method uses reflections of the points in the boundary resulting in a new data set $X_1, -X_1, \cdots, X_n, -X_n$. Under the assumption that the kernel is differentiable and symmetric, the resulting estimator has zero derivative at the boundary. The estimator of $m(x)$ is defined by

$$m(\hat{x}) = \frac{1}{nb}\sum_{i=1}^{n}\left\{K\left(\frac{x - X_i}{b}\right) + K\left(\frac{x + X_i}{b}\right)\right\}$$

for $x \geq 0, m'(\hat{x}) = 0$, $m(x)$ is a function of auxiliary random variables assumed to be known throughout the study. For $x < 0$ it can be shown that $m\prime(\hat{x}) = 0$, therefore it becomes better than other methods if the underlying density has the property $m\prime(0) = 0$; if this property does not hold, the method may become cumbersome to apply.

2.4.2 Pseudo-Data Method

This method was suggested by (Cowling and Hall, 1996). In this method, data is generated outside the interval of estimation. Those data are assumed to be linear functions of order statistics in the original sample X. This transforms the data into a new set, then puts it on the negative axis. The estimator of $m(x)$ is given by

$$m(\hat{x}) = \frac{1}{nb}\left\{\sum_{i=1}^{n} K\left(\frac{x - X_i}{b}\right) + \sum_{i=1}^{m} K\left(\frac{x + X_{(-i)}}{b}\right)\right\}$$

where $m \leq n$ and

$$X_{(-i)} = -5X(\tfrac{i}{3}) - 4X(\tfrac{2}{3}i) + \frac{10}{3}X(i)$$

where $X_{(i)}$ linearly interpolates among $0, X_{(1)}, X_{(2)}, ..., X_{(n)}$ in that order, for details see (Cowling and Hall, 1996). Though this method is simple to implement and allows a minimal variance of the usual kernel estimator, the drawback is that straight data reflection corrects only for a jump in the value of the density at the ends of its support, not for discontinuities in the derivatives of the density. Therefore, the method does not adequately correct bias problems caused by the edge effects in kernel estimators of order 2 or higher .

2.4.3 Boundary Kernel Method

The boundary kernel estimate at a particular point of estimation in the boundary region is obtained by first constructing the appropriate kernel for that point. Many researchers including (Gasser and Müller, 1979), (Müller, 1991), Jones (1993) and (Zhang and Karunamuni, 2000) have explored this approach. The method applies a different kernel for estimating function at each point in the boundary region. Due to this, some kernels may not hold the symmetry property and can therefore put more weight on the positive axis. The estimator for $m(x)$ using this method is

$$m(\hat{x}) = \frac{1}{nb}\sum_{i=1}^{n} K_{\frac{c}{h(c)}}\left(\frac{x - X_i}{b_c}\right)$$

where $x = cb$, $0 \leq c \leq 1$, $h(c) = 2 - c$ Besides, K_c is such that for $0 \leq c < 1$

$$K_{(c)}(z) = \frac{12}{(1+c)^4}(1+z)\left\{(1-2c)z + \frac{3c^2 - 2c + 1}{2}\right\}\{-1 \leq z \leq 1\}$$

The boundary kernel methods normally have smaller bias though with an increase in variance. It has been noted, see for instance, (Müller, 1991) that methods dealing only with kernel adjustments without regards to data, such as boundary kernel method, mostly suffer increased variance. Besides, the corresponding estimates have negative values around the boundary points. This is due to the fact that some kernels may not be symmetric and can therefore put more weight on the positive axis. These drawbacks limit the use of this method.

2.4.4 Transformation of Data Method

This technique has been discussed by (Wand and Jones, 1994a) and (Marron and Ruppert, 1994). Original data $X_1, ..., X_n$ is transformed to $g(X_1), ..., g(X_n)$, while retaining the original data, where g is a non-negative, continuous and monotonically increasing function for $[0, \infty) \rightarrow [0, \infty)$. To use this method, one can take a one-to-one continuous function: $[0, \infty) \rightarrow [0, \infty)$. A regular kernel estimator is thus used with the transformed data set $\{g(X_1), g(X_2), ..., g(X_n)\}$. The estimator is

$$m(\hat{x}) = \frac{1}{nb}\sum_{i=1}^{n}\left(\frac{x - g(X_i)}{b}\right)$$

This method gives the estimator of the probability density function of $g(X)$ not that of X. The strength of this method is that estimates obtained using transformation of data are positive with low variance. The non-negativity property is very vital in practical applications and it is therefore worth-exploring to consider methods that result in non-negative variance of estimators. A modified version of this method is therefore proposed in this study since it is not computationally-intensive and is easier to implement compared to the other methods. In chapter four, the estimator of the finite population mean is modified using transformation of data technique, and further, its asymptotic properties are derived.

2.5 Methods of Selecting a Smoothing Parameter

Asymptotic properties of estimators of population parameters are usually a function of a smoothing parameter mostly known as a bandwidth. In non-parametric curve estimation, (Wand and Jones, 1994a) noted that the choice of the bandwidth is very critical since it determines the stability and rates of convergence of asymptotic properties of estimators of population characteristics. Moreover, reliable finite sample performance and faster convergence rates are of great importance in data analysis as noted by (Wand and Jones, 1994b).

It has been demonstrated that estimates of population parameters obtained using kernel densities should be drawn using more than one value of the bandwidth, for example (Terrell and Scott, 1992) recommended using a sequence of estimates based on the sequence of smoothing parameters starting with the sample over-smoothed bandwidth and stopping when the estimate displays some instability and error variation near the peaks. Practical importance and performance of bandwidth selection have been discussed by (Jones et al., 1992) and (Kim et al., 1994). The results showed that there is no selection criteria of a bandwidth that gives better results for all the densities. In this research, kernel weights are used in estimating the non-response component of the finite population mean. It is noted that a kernel is a function of the bandwidth hence the reason for reviewing the various methods of bandwidth selection in the next subsections.

2.5.1 Cross-Validation

This technique is one of the data-driven methods for selecting bandwidth. It comprises of least squares cross-validation as discussed by (Hardle and Marron, 1985), biased cross-validation, (Sain et al., 1994) and over-smoothing discussed by (Sheather, 2004). It has been shown by (Hall and Marron, 1987) that cross-validation is limited by a high sample variability while over-smoothing hides important characteristics of the data which may affect optimal convergence. Though these limitations may reduce the usefulness of cross-validation, it has been argued by (Silverman, 1986) that cross-validation works well for continuous samples. This study adopted a modified optimal bandwidth suggested by (Loader et al., 1999) obtained using cross-validation technique given by

$$CV(b) = (mn)^{-1} \sum_{i=1}^{n} \sum_{i=1}^{m} \left\{ Y_{ij} - \hat{m}(X_{ij}) \right\}^2 w(X_{ij})$$

where $\hat{m}(X_{ij})$ is a given data function of the auxiliary random variables, X_{ij}, while $w(X_{ij})$ is a non-negative weight function.

2.5.2 Plug-in Method

Plug-in methods of bandwidth selection that include direct plug-in, solve-the-equation and simple rule of thumb have been fronted by (Ruppert et al., 1995) and (Wand and Jones, 1994a). Ruppert et al. (1995) for instance concluded that these techniques perform fairly well especially during simulations. Goh et al. (2004) recommends consideration of a family of density estimates for a given data-set based on different values of the bandwidth. The conclusion was that such a family of estimates ought to be based around a "center point" that gives an effective choice of the bandwidth. Goh et al. (2004) suggested the adoption of the plug-in method due to (Jones et al., 1996). However, in a comprehensive research on bandwidth selectors done by (Loader et al., 1999), it was observed that plug-in methods perform poorly since they rely a lot on trial and error method which is often biased. The conclusion is that classical methods like cross-validation give more informative estimates than plug-in methods. More recently, (Lang'at, 2017) observed that plug-in methods are associated with various weaknesses which lead to less informative results and for this reason cross-validation technique should be adopted for bandwidth selection. Based on this literature, cross-validation technique is adopted in this study.

2.6 The Big "O" notation and the Little "o" notation as used in order algebra

Big O-notation is used to describe the asymptotic behaviour of functions. They are used to explain how fast a function grows or declines. For instance, when analyzing algorithms, one might find that the time (or number of steps) it takes to complete a problem of size n is given by $T(n) = 6n^2 - 3n + 9$. If the constants and the slower growing terms are ignored, one could say $T(n)$ grows at *order* of n^2 and therefore express $T(n) = O(n^2)$, (Bach et al., 1996).

In Mathematical Statistics, it is very vital to handle the error term of an approximation, for instance,

$$F(x) = x^7 + O(x^2)$$

for $x \geq 0$. This equation can be written to express the fact that the error is smaller in absolute value than some constant times x^2 if x is close enough to zero, that means the main term of $F(x)$ grows like x^7 while the correction terms are at most some constant times x^2. Furthermore, suppose $f(x)$ and $g(x)$ are two functions defined on some subset of real numbers. Then $f(x) = O(g(x))$ if and only if there exist constants a and a_0 such that

$$|f(x)| \leq a|g(x)|, \quad \forall x > a_0$$

This means $f(x)$ does not grow faster than $g(x)$ and if C is some real number, then

$$f(x) = O(g(x))$$

for $x \geq \epsilon$ if and only if $\exists$ constants $\delta > 0$ and a such that

$$|f(x)| \leq a|g(x)|, \quad \forall x > a$$

with $|x - \epsilon| < \delta$, (Jeffrey and Dai, 2008) .

If an expression is given as

$$f(x) = o(g(x))$$

as $x \to \infty$, then the o-notation means that $g(x) \neq 0$ for sufficiently large x and

$$\lim_{x \to \infty} \frac{f(x)}{g(x)} = 0$$

If this relation holds then $f(x)$ is said to be of smaller order than $g(x)$.
Big-O notations have to be true for at least one constant say a_0 while little-o notations have to be true for every positive constant, say ϵ, for more details see (Serfling, 2009). These notations have been reviewed here since they are used in chapter three and four in the derivation of asymptotic properties of the estimators.

2.7 Summary of Literature Review

In summary, estimation of a finite population mean in presence of non-response has been done by many researchers using various methods of compensating for non-response. For example (Singh and Horn, 2000) and (Durrant et al., 2005) used imputation techniques to compensate for non-response in the estimation of a finite population mean. However, the estimators developed using imputation technique produced a large bias. Kernel weights derived using Nadaraya-Watson technique are proposed in this study to compensate for non-response. However, this is limited by the boundary bias. Transformation of data method discussed by (Wand and Jones, 1994a) and (Marron and Ruppert, 1994) and further extended by (Baszczyńska, 2015) has been proposed in this study to correct the boundary effects due to Nadaraya-Watson kernel weights. Moreover, (Marron and Ruppert, 1994) observed that the selection of the bandwidth used in kernel density estimations is a critical problem in determining the asymptotic properties of estimators of finite population parameters. For that reason, the cross-validation technique due to (Loader et al., 1999) was adopted in this study using two stage cluster sampling since it gives more informative estimates than other methods such as plug-in method.

CHAPTER THREE

THE PROPOSED ESTIMATOR

3.1 Introduction

In this chapter, an estimator of a finite population mean is proposed in the presence of non-response. Also, asymptotic properties of the proposed estimator such as the bias and the mean squared error are derived.

3.2 The Proposed Estimator of a Finite Population Mean

Consider a finite population of size N consisting of M clusters with N_j elements in the j^{th} cluster. A sample of m clusters is selected so that n_{1i} units respond and n_{2i} units fail to respond. Let y_{ij} denote the value of the survey variable y for unit j in cluster i, for $i = 1, 2, \cdots, N, j = 1, 2, \cdots, N_i$ and let population mean be given by

$$\overline{\overline{Y}} = \frac{1}{N} \sum_{i=1}^{M} \sum_{j=1}^{N_i} Y_{ij} \tag{3.1}$$

Let an estimator of the finite population mean be defined by $\hat{\overline{\overline{Y}}}$ as follows

$$\hat{\overline{\overline{Y}}} = \frac{1}{M} \left\{ \frac{1}{n_{1i}} \sum_{i \in s} \sum_{j \in s} \frac{y_{ij}}{\pi_{ij}} \delta_{ij} + \frac{1}{n_{2i}} \sum_{i \in s} \sum_{j \notin s} \left(1 - \frac{1}{\pi_{ij}}\right) \hat{Y}_{ij} \delta_{ij} \right\} \tag{3.2}$$

where δ_{ij} is an indicator variable defined by

$$\delta_{ij} = \begin{cases} 1, \text{ if } j^{th} \text{ unit in the } i^{th} \text{ cluster responds} \\ 0, \text{ elsewhere} \end{cases} \tag{3.3}$$

and n_{1i} and n_{2i} are the number of units that respond and those that fail to respond respectively. π_{ij} is the probability of selecting the i^{th} unit from the j^{th} cluster into the sample while $\hat{Y}_{ij}$ estimates the survey values, Y_{ij}. Let $w(x_{ij}) = \frac{1}{\pi_{ij}}$, where $0 < \pi_{ij} < 1$, be the inverse of the second order inclusion probabilities and x_{ij} be the i^{th} auxiliary

random variable from the j^{th} cluster. It follows that (3.2) becomes

$$\hat{\bar{\bar{Y}}} = \frac{1}{M}\left\{\frac{1}{n_{1i}}\sum_{i\in s}\sum_{j\in s} w(x_{ij})\delta_{ij}y_{ij} + \frac{1}{n_{2i}}\sum_{i\in s}\sum_{j\notin s}\Big(1 - w(x_{ij})\Big)\hat{Y}_{ij}\delta_{ij}\right\} \tag{3.4}$$

Suppose δ_{ij} is known to be an indicator variable with probability of success δ^*_{ij}, then, $E(\delta_{ij}) = p_r(\delta_{ij} = 1) = \delta^*_{ij}$ and $Var(\delta_{ij}) = \delta^*_{ij}(1 - \delta^*_{ij})$. Thus, the expected value of the estimator is given by

$$E(\hat{\bar{\bar{Y}}}) = \frac{1}{M}\left\{\frac{1}{n_{1i}}\sum_{i\in s}\sum_{j\in s} w(x_{ij})\delta^*_{ij}y_{ij} + \frac{1}{n_{2i}}\sum_{i\in s}\sum_{j\notin s} E\Big(\Big(1 - w(x_{ij})\Big)\hat{Y}_{ij}\Big)\delta^*_{ij}\right\} \tag{3.5}$$

Assuming non-response in the second stage of sampling, the problem is therefore to estimate the values of $\hat{Y}_{ij}$. To do this, a regression model applied by (Ouma and Wafula, 2005) and (Onyango et al., 2010) given below is used;

$$\hat{Y}_{ij} = m(\hat{x}_{ij}) + \hat{e}_{ij} \tag{3.6}$$

where $m(.)$ is a smooth function of the auxiliary variables and $\hat{e}_{ij}$ is the residual term with mean zero and variance which is strictly positive. Substituting equation (3.6) in equation (3.5) the following result is obtained:

$$\begin{aligned} E(\hat{\bar{\bar{Y}}}) = \frac{1}{M}\Bigg\{&\frac{1}{n_{1i}}\sum_{i\in s}\sum_{j\in s} w(x_{ij})\delta^*_{ij}y_{ij} \\ &+ \frac{1}{n_{2i}}\sum_{i\in s}\sum_{j\notin s} E\Big(1 - w(x_{ij})\Big)\Big(m(\hat{x}_{ij}) + \hat{e}_{ij}\Big)\delta^*_{ij}\Bigg\} \end{aligned} \tag{3.7}$$

Assume $n_{1i} = n_{2i} = n$, this assumption was taken as a special case since there are cases where $n_{1i} \neq n_{2i}$. Using this assumption, equation (3.7) was then simplified to obtain

$$\begin{aligned} E(\hat{\bar{\bar{Y}}}) = \frac{1}{Mn}\Bigg\{&\sum_{i\in s}\sum_{j\in s} w(x_{ij})\delta^*_{ij}y_{ij} \\ &+ \sum_{i\in s}\sum_{j\notin s} E\Big(1 - w(x_{ij})\Big)\Big(m(\hat{x}_{ij}) + \hat{e}_{ij}\Big)\delta^*_{ij}\Bigg\} \end{aligned} \tag{3.8}$$

The second term in equation (3.8) is simplified as follows:

$$\frac{1}{Mn}\left\{\sum_{i\in s}\sum_{j\notin s}E\Big(1-w(x_{ij})\Big)\Big(m(\hat{x}_{ij})+\hat{e}_{ij}\Big)\delta^*_{ij}\right\}$$
$$=\frac{1}{Mn}\left\{\sum_{i\in s}\sum_{j\notin s}E\Big(1-w(x_{ij})\Big)m(\hat{x}_{ij})\delta^*_{ij}\right\} \quad (3.9)$$
$$+\frac{1}{Mn}\left\{\sum_{i\in s}\sum_{j\notin s}E\Big(1-w(x_{ij})\Big)\delta^*_{ij}e_{ij}\right\}$$

But $E\left(m(x_{ij})\right)=m(\hat{x}_{ij})=m(x_{ij})$, (Dorfman, 1992). Thus the following equation is obtained:

$$\frac{1}{Mn}\left\{\sum_{i\in s}\sum_{j\notin s}E\Big(1-w(x_{ij})\Big)\Big(m(\hat{x}_{ij})+\hat{e}_{ij}\Big)\delta^*_{ij}\right\}$$
$$=\frac{1}{Mn}\left\{\sum_{i=\,n+1}^{N}\sum_{j=\,m+1}^{M}\delta^*_{ij}m(x_{ij})-w(x_{ij})\delta^*_{ij}m(x_{ij})\right\} \quad (3.10)$$
$$+\frac{1}{Mn}\left\{\sum_{i=\,n+1}^{N}\sum_{j=\,m+1}^{M}E(e_{ij}\delta^*_{ij})-E\left(w(x_{ij})\left(e_{ij}\delta^*_{ij}\right)\right)\right\}$$

Equation (3.10) is simplified as follows:

$$\frac{1}{Mn}\left\{\sum_{i\in s}\sum_{j\notin s}E\Big(1-w(x_{ij})\Big)\Big(m(\hat{x}_{ij})+\hat{e}_{ij}\Big)\delta^*_{ij}\right\}$$
$$=\frac{1}{Mn}\Bigg\{\Big(M-(m+1)\Big)\Big(N-(n+1)\Big)\Big[\delta^*_{ij}m(x_{ij})$$
$$-w(x_{ij})\delta^*_{ij}m(x_{ij})\Big]+(M-(m+1))(N-(n+1)) \quad (3.11)$$
$$\times\Big[\delta^*_{ij}E(e_{ij})-E(e_{ij})\delta^*_{ij}w(x_{ij})\Big]\Bigg\}$$

But as shown by (Onyango et al., 2010), $E(e_{ij})=0$, thus on simplification, equation (3.11) reduces to

$$\frac{1}{Mn}\left\{\sum_{i\in s}\sum_{j\notin s}E\Big(1-w(x_{ij})\Big)\Big(m(\hat{x}_{ij})+\hat{e}_{ij}\Big)\delta^*_{ij}\right\}$$
$$=\frac{\Big\{\Big(M-(m+1)\Big)\Big(N-(n+1)\Big)\Big\}}{Mn}\times\Delta_1 \quad (3.12)$$

where $\Delta_1 = \left\{ \delta_{ij}^* m(x_{ij}) \left(1 - w(x_{ij})\right) \right\}$

But $w(x_{ij}) = \frac{1}{\pi_{ij}}$ so that equation (3.12) is thus equivalent to:

$$\begin{aligned} \frac{1}{Mn}\Bigg\{ \sum_{i\in s}\sum_{j\notin s} E\Big(1 - w(x_{ij})\Big)\Big(m(\hat{x}_{ij}) + \hat{e}_{ij}\Big)\delta_{ij}^* \Bigg\} \\ = \frac{\Big\{\Big(M - (m+1)\Big)\Big(N - (n+1)\Big)\Big\}}{Mn} \times \Delta_2 \end{aligned} \tag{3.13}$$

where $\Delta_2 = \left\{ \delta_{ij}^* m(x_{ij}) \left(\frac{\pi_{ij}-1}{\pi_{ij}}\right) \right\}$. Assume the sample sizes are large i.e. as $n \to N$ and $m \to M$, equation (3.13) simplifies to

$$\begin{aligned} \frac{1}{Mn}\Bigg\{ \sum_{i\in s}\sum_{j\notin s} E\Big(1 - w(x_{ij})\Big)\Big(m(\hat{x}_{ij}) + \hat{e}_{ij}\Big)\delta_{ij}^* \Bigg\} \\ = \frac{1}{Mn}\left\{ \delta_{ij}^* m(x_{ij}) \left(\frac{\pi_{ij} - 1}{\pi_{ij}}\right) \right\} \end{aligned} \tag{3.14}$$

Combining equation (3.14) and the first term in equation (3.8) the following result is obtained:

$$E(\hat{\bar{\bar{Y}}}) = \frac{1}{Mn}\left\{ \sum_{i\in s}\sum_{j\in s} E\left(m(x_{ij})\right) E\left(\frac{\delta_{ij}^*}{\pi_{ij}}\right) y_{ij} + \sum_{i\in s}\sum_{j\notin s} \delta_{ij}^* \left(m(\hat{x}_{ij})\right) \left(\frac{\pi_{ij} - 1}{\pi_{ij}}\right) \right\} \tag{3.15}$$

Since the first term represents the response units, their values are all known and hence $\delta_{ij}^* = 1$. The problem is to estimate the non-response units in the second term. The problem now reduces to that of estimating the function $m(\hat{x_{ij}})$, which is a function of the auxiliary variables. Hence the expected value of the estimator of finite population mean under non-response is given as;

$$E(\hat{\bar{\bar{Y}}}) = \frac{1}{Mn}\left\{ \sum_{i\in s}\sum_{j\in s} y_{ij} + \sum_{i\in s}\sum_{j\notin s} m(\hat{x}_{ij}) \left(\frac{\pi_{ij} - 1}{\pi_{ij}}\right) \right\} \tag{3.16}$$

where $0 < \pi_{ij} < 1$. In order to derive the asymptotic properties of the expected value of the proposed estimator in equation (3.16), first a review of Nadaraya-Watson estimator is given below:

3.3 Review of Nadaraya-Watson Estimator

Given a random sample of bi-variate data $(x_i, y_i), \cdots, (x_n, y_n)$ having a joint p.d.f $g(x, y)$ with the regression model given by $\hat{Y}_{ij} = m(\hat{x_{ij}}) + \hat{e_{ij}}$ as in equation (3.6), where $m(.)$ is unknown. It is assumed that the error term satisfies the following conditions:

$$E(e_{ij}) = 0, Var(e_{ij}) = \sigma_{ij}^2, Cov(e_i, e_j) = 0, for \quad i \neq j \tag{3.17}$$

Furthermore, let $K(.)$ denote a kernel density function which is twice continuously differentiable with the following conditions due to (Nadaraya, 1964; Watson, 1964):

$$\left.\begin{array}{c} \int_{-\infty}^{\infty} k(w)dw = 1 \\ \int_{-\infty}^{\infty} wk(w)dw = 0 \\ \int_{-\infty}^{\infty} k^2(w)dw < \infty \\ \int_{-\infty}^{\infty} w^2k(w)dw = d_k \\ k(w) = k(-w) \end{array}\right\} \tag{3.18}$$

In addition, let the smoothing weights be given by

$$w(x_{ij}) = \frac{K\left(\frac{x-X_{ij}}{b}\right)}{\sum_{i\in s}\sum_{j\in s} K\left(\frac{x-X_{ij}}{b}\right)}, i = 1, 2, \cdots, n, j = 1, 2, \cdots, m \tag{3.19}$$

where b is a smoothing parameter, normally referred to as the bandwidth such that $\sum_{i\in s}\sum_{j\in s} w(x_{ij}) = 1$.

Using equation (3.19), the Nadaraya-Watson estimator of $m(x_{ij})$ is

$$m(\hat{x}_{ij}) = \sum_{i\in s}\sum_{j\in s} w(x_{ij})Y_{ij} = \frac{\sum_{i\in s}\sum_{j\in s} K\left(\frac{x-X_{ij}}{b}\right)Y_{ij}}{\sum_{i\in s}\sum_{j\in s} K\left(\frac{x-X_{ij}}{b}\right)}, \tag{3.20}$$
$$i = 1, 2, \cdots, n, j = 1, 2, \cdots, m$$

Given the model $\hat{Y}_{ij} = m(\hat{x}_{ij}) + \hat{e}_{ij}$ and the conditions of the error term as stated in (3.17) above, the expression of the expected value of the survey variable Y_{ij} relative to

the auxiliary variable X_{ij} can be given as a joint p.d.f of $g(x_{ij}, y_{ij})$ as follows:

$$\begin{aligned} m(x_{ij}) &= E\left(Y_{ij}/X_{ij} = x_{ij}\right) \\ &= \int y\, g\left[y/x\right] dy \end{aligned} \tag{3.21}$$

Equation (3.21) can therefore be written as

$$m(x_{ij}) = \frac{\int y\, g\left(x, y\right) dy}{\int g\left(x, y\right) dy} \tag{3.22}$$

where $\int g\left(x, y\right) dy$ is the marginal density of X_{ij}. The numerator and the denominator of equation (3.22) can be estimated separately using kernel functions as follows: $g(x, y)$ is estimated by:

$$\hat{g(x, y)} = \frac{1}{mn} \sum_i \sum_j \left(\frac{1}{b} K \left(\frac{x - X_{ij}}{b} \right) \frac{1}{b} K \left(\frac{y - Y_{ij}}{b} \right) \right) \tag{3.23}$$

and

$$\int y\hat{g(x, y)} dy = \frac{1}{mn} \sum_i \sum_j \int \left(\frac{1}{b} K \left(\frac{x - X_{ij}}{b} \right) \frac{1}{b} K \left(\frac{y - Y_{ij}}{b} \right) \right) y dy \tag{3.24}$$

Using change of variables technique, let

$$\left.\begin{aligned} w &= \tfrac{y - Y_{ij}}{b} \\ y &= wb + Y_{ij} \\ dy &= b dw \end{aligned}\right\} \tag{3.25}$$

so that

$$\int y\hat{g(x, y)} dy = \frac{1}{mn} \sum_i \sum_j \int \frac{1}{b} K \left(\frac{x - X_{ij}}{b} \right) \frac{1}{b} \left(bw + Y_{ij} \right) K(w) b dw \tag{3.26}$$

this reduces to

$$\int yg(\hat{x},y)dy = \frac{1}{mnb}\sum_i\sum_j K\left(\frac{x-X_{ij}}{b}\right)\left[\int wK(w)bdw + \frac{1}{b}Y_{ij}\int K(w)bdw\right] \tag{3.27}$$

From the conditions specified in equation (3.18) , equation (3.27) can be simplified to

$$\int yg(\hat{x},y)dy = \frac{1}{mnb}\sum_i\sum_j K\left(\frac{x-X_{ij}}{b}\right)[0+Y_{ij}] \tag{3.28}$$

which reduces to

$$\int yg(\hat{x},y)dy = \frac{1}{mnb}\sum_i\sum_j K\left(\frac{x-X_{ij}}{b}\right)Y_{ij} \tag{3.29}$$

Following the same procedure, the denominator of equation (3.22) reduces to

$$\int g(\hat{x},y)dy = \frac{1}{mn}\sum_i\sum_j \int\left(\frac{1}{b}K\left(\frac{x-X_{ij}}{b}\right)\frac{1}{b}K\left(\frac{y-Y_{ij}}{b}\right)\right)dy \tag{3.30}$$

Re-writing equation (3.30) gives

$$\int g(\hat{x},y)dy = \frac{1}{mnb}\sum_i\sum_j \int\left(K\left(\frac{x-X_{ij}}{b}\right)\frac{1}{b}K\left(\frac{y-Y_{ij}}{b}\right)\right)dy \tag{3.31}$$

Using change of variable technique as in equation (3.25), equation (3.31) can be re-written as follows

$$\int g(\hat{x},y)dy = \frac{1}{mnb}\sum_i\sum_j K\left(\frac{x-X_{ij}}{b}\right)\int\frac{1}{b}K(w)bdw \tag{3.32}$$

which yields

$$\int g(\hat{x},y)dy = \frac{1}{mnb}\sum_i\sum_j K\left(\frac{x-X_{ij}}{b}\right) \tag{3.33}$$

since $\int \frac{1}{b}K(w)bdw$ is a p.d.f and therefore integrates to 1.

It follows from equation (3.29) and (3.33) that the estimator $m(\hat{x_{ij}})$ is as given in equation (3.20). Given a random sample and a specified kernel function, then for a given auxiliary value x_{ij}, the corresponding $y-$estimate is obtained by the estimator

outlined in equation (3.20), which can be written as:

$$\hat{y}_{ij} = m_{NW}(\hat{x_{ij}}) = \sum_i \sum_j W_{ij}(x_{ij}) Y_{ij} \tag{3.34}$$

where $m_{NW}(\hat{x_{ij}})$ is the Nadaraya-Watson estimator for estimating the unknown function $m(.)$, for details see (Nadaraya, 1964; Watson, 1964). Since the choice of the kernel function is not critical for the performance of the kernel regression estimator, a simplified normal kernel having mean 0 and variance 1 is used in this study. This is given by

$$K(w) = \frac{1}{\sqrt{2\pi}} e^{-\left(\frac{w^2}{2}\right)} = \frac{1}{\sqrt{2\pi}} e^{-\left(\frac{\left(\frac{x - X_{ij}}{b}\right)^2}{2}\right)} \tag{3.35}$$

In this case, the Nadaraya-Watson kernel estimation at any point x_{ij} is given by

$$\hat{y}_{ij} = m_{NW}(\hat{x}_{ij}) = \frac{\sum_i \sum_j \frac{1}{\sqrt{2\pi}} e^{-\left(\frac{\left(\frac{x - X_{ij}}{b}\right)^2}{2}\right)} Y_{ij}}{\sum_i \sum_j \frac{1}{\sqrt{2\pi}} e^{-\left(\frac{\left(\frac{x - X_{ij}}{b}\right)^2}{2}\right)}} \tag{3.36}$$

where b is the bandwidth.

This provides a way of estimating for instance the non-response values of the survey variable Y_{ij}, given the auxiliary values x_{ij}, for a specified kernel function.

3.4 Asymptotic Bias of the Mean estimator, $\hat{\bar{\bar{Y}}}$

Equation (3.16) may be written as

$$E(\hat{\bar{\bar{Y}}}) = \frac{1}{Mn} \left\{ \sum_{i=1}^{n} \sum_{j=1}^{m} Y_{ij} + \sum_{i=n+1}^{N} \sum_{j=m+1}^{M} m\left(\hat{x}_{ij}\right) \right\} \tag{3.37}$$

Re-writing equation (3.34) using the property of symmetry associated with Nadaraya-Watson estimator,

$$m\left(\hat{x}_{ij}\right) = \frac{\sum_{i \in s} \sum_{j \in s} K\left(\frac{X_{ij} - x_{ij}}{b}\right) Y_{ij}}{\sum_{i \in s} \sum_{j \in s} K\left(\frac{X_{ij} - x_{ij}}{b}\right)}, i = 1, 2, \cdots, n; j = 1, 2, \cdots, m \tag{3.38}$$

Following the procedure by (Takezawa, 2005), equation (3.38) can be re-written as

$$m\left(\hat{x}_{ij}\right) = \frac{1}{g(\hat{x}_{ij})}\left[\frac{1}{mnb}\sum_{i}\sum_{j} K\left(\frac{X_{ij} - x_{ij}}{b}\right) Y_{ij}\right] \tag{3.39}$$

where $g(\hat{x}_{ij})$ is the estimated marginal density of auxiliary variables X_{ij}. But for a finite population mean, the expected value of the estimator is given in equation (3.37). The bias is given by

$$Bias\left(\hat{\overline{\overline{Y}}}\right) = E(\hat{\overline{\overline{Y}}} - \overline{\overline{Y}}) \tag{3.40}$$

$$\begin{aligned} Bias\left(\hat{\overline{\overline{Y}}}\right) = E\Bigg\{ & \frac{1}{Mn}\Bigg[\sum_{i=1}^{n}\sum_{j=1}^{m} Y_{ij} + \sum_{i=n+1}^{N}\sum_{j=m+1}^{M} m(\hat{x}_{ij})\Bigg] \\ & - \frac{1}{Mn}\Bigg[\sum_{i=1}^{n}\sum_{j=1}^{m} Y_{ij} + \sum_{i=n+1}^{N}\sum_{j=m+1}^{M} Y_{ij}\Bigg]\Bigg\} \end{aligned} \tag{3.41}$$

which reduces to

$$Bias\left(\hat{\overline{\overline{Y}}}\right) = \frac{1}{Mn}E\left\{\sum_{i=n+1}^{N}\sum_{j=m+1}^{M} m(\hat{x}_{ij}) - \sum_{i=n+1}^{N}\sum_{j=m+1}^{M} Y_{ij}\right\} \tag{3.42}$$

Re-writing the regression model given by $Y_{ij} = m(X_{ij}) + e_{ij}$ as

$$Y_{ij} = m(x_{ij}) + \left[m(X_{ij}) - m(x_{ij})\right] + e_{ij} \tag{3.43}$$

and substituting it in equation (3.39) gives

$$\begin{aligned} m\left(\hat{x}_{ij}\right) = \frac{1}{g(\hat{x}_{ij})}\Bigg[\frac{1}{mnb}\sum_{i}\sum_{j} K\Big(\frac{X_{ij} - x_{ij}}{b}\Big)\Big(m(x_{ij}) + \left[m(X_{ij}) - m(x_{ij})\right] \\ + e_{ij}\Big)\Bigg] \end{aligned} \tag{3.44}$$

Hence the first term in equation (3.42) before taking expectation is given as:

$$\begin{aligned}
&\frac{1}{Mn}\left\{\frac{\frac{1}{mnb}\sum_{i=n+1}^{N}\sum_{j=m+1}^{M}K\left(\frac{X_{ij}-x_{ij}}{b}\right)Y_{ij}}{g(\hat{x}_{ij})}\right\} \\
&\qquad=\frac{1}{Mng(\hat{x}_{ij})}\left\{\frac{1}{mnb}\sum_{i=n+1}^{N}\sum_{j=m+1}^{M}K\left(\frac{X_{ij}-x_{ij}}{b}\right)m(x_{ij})\right. \\
&\qquad+\frac{1}{mnb}\sum_{i=n+1}^{N}\sum_{j=m+1}^{M}K\left(\frac{X_{ij}-x_{ij}}{b}\right)\Big[m(X_{ij})-m(x_{ij}))\Big] \\
&\qquad\left.+\frac{1}{mnb}\sum_{i=n+1}^{N}\sum_{j=m+1}^{M}K\left(\frac{X_{ij}-x_{ij}}{b}\right)e_{ij}\right\}
\end{aligned} \tag{3.45}$$

Simplifying equation (3.45) the following is obtained:

$$\begin{aligned}
&\frac{1}{Mn}\left\{\frac{1}{mnb\,g(\hat{x}_{ij})}\sum_{i=n+1}^{N}\sum_{j=m+1}^{M}K\left(\frac{X_{ij}-x_{ij}}{b}\right)Y_{ij}\right\} \\
&\qquad=\frac{1}{Mn}\left(\frac{1}{mnb\,g(\hat{x}_{ij})}\right)\left\{\sum_{i=n+1}^{N}\sum_{j=m+1}^{M}g(\hat{x}_{ij})m(x_{ij})\right. \\
&\qquad\qquad\left.+m_1(\hat{x}_{ij})+m_2(\hat{x}_{ij})\right\}
\end{aligned} \tag{3.46}$$

where

$$m_1(\hat{x}_{ij})=\frac{1}{mnb}\sum_{i=n+1}^{N}\sum_{j=m+1}^{M}K\left(\frac{X_{ij}-x_{ij}}{b}\right)\Big[m(X_{ij})-m(x_{ij})\Big] \tag{3.47}$$

$$m_2(\hat{x}_{ij})=\frac{1}{mnb}\sum_{i=n+1}^{N}\sum_{j=m+1}^{M}K\left(\frac{X_{ij}-x_{ij}}{b}\right)e_{ij} \tag{3.48}$$

Taking conditional expectation of equation (3.46) leads to

$$\begin{aligned}
E\Big[\sum_{i=n+1}^{N}\sum_{j=m+1}^{M}m(\hat{x}_{ij})/x_{ij}\Big]&=\frac{1}{Mn}E\left[\frac{1}{mnb}\sum_{i=n+1}^{N}\sum_{j=m+1}^{M}\left[m(x_{ij})\right.\right. \\
&\qquad\left.\left.+\frac{m_1(\hat{x}_{ij})}{g(\hat{x}_{ij})}+\frac{m_2(\hat{x}_{ij})}{g(\hat{x}_{ij})}\right]\right]
\end{aligned} \tag{3.49}$$

To obtain the relationship between the conditional mean and the selected bandwidth, the following theorem suggested by (Dorfman, 1992) was applied.

Theorem 3.1 *Let $K(w)$ be a symmetric density function with $\int wk(w)dw = 0$ and $\int w^2k(w)dw = k_2$. Assume n and N increase together such that $\frac{n}{N} \to \pi$ with $0 < \pi < 1$. Besides, assume the sampled and non-sampled values of x are in the interval $[c, d]$ and are obtained by densities d_s and d_{p-s} respectively where both are bounded away from zero on $[c, d]$ with continuous second derivatives. If for any variable Z, $E(Z/U = u) = A(u) + O(B)$ and $Var(Z/U = u) = O(C)$, then $Z = A(u) + O_p(B + C^{\frac{1}{2}})$.*

Applying this theorem, the conditional mean squared error, partitioned into conditional variance and bias can be obtained as

$$MSE(\hat{\bar{\bar{Y}}}/x_{ij}) = Var(\hat{\bar{\bar{Y}}}/x_{ij}) + \Big(Bias(\hat{\bar{\bar{Y}}}/x_{ij})\Big)^2 \tag{3.50}$$

The conditional variance and bias terms can thus be separately obtained. From the conditions stated in (3.17) it follows that $E(e_{ij}/X_{ij}) = 0$. Therefore, $E\big[m_2(\hat{x}_{ij}) = 0\big]$. Thus, $E\big[m_1(\hat{x}_{ij})\big]$ can be obtained as follows:

$$\begin{aligned} E\sum_{i=n+1}^{N}\sum_{j=m+1}^{M}\Big[m_1(\hat{x}_{ij})\Big] = \frac{1}{Mn}\Big(\frac{1}{mnb}\Big)E\Big\{\sum_{i=n+1}^{N}\sum_{j=m+1}^{M}K\Big(\frac{X_{ij}-x_{ij}}{b}\Big) \\ \times\Big[m(X_{ij}) - m(x_{ij})\Big]\Big\} \end{aligned} \tag{3.51}$$

Using substitution and change of variable technique below

$$\left.\begin{aligned} w &= \tfrac{V-x_{ij}}{b} \\ V &= x_{ij} + bw \\ dV &= bdw \end{aligned}\right\} \tag{3.52}$$

Equation (3.51) can be simplified to:

$$\begin{aligned} E\sum_{i=n+1}^{N}\sum_{j=m+1}^{M}\big[m_1(\hat{x}_{ij})\big] = \frac{1}{Mn}\Big\{\frac{Mn-mn}{mnb}\int k(w)\Big[m\Big(x_{ij}+bw\Big) \\ - m(x_{ij})\Big]\int g\big(x_{ij}+bw\big)bdw\Big\} \end{aligned} \tag{3.53}$$

which simplifies to

$$E\sum_{i=n+1}^{N}\sum_{j=m+1}^{M}\Big[m_1(\hat{x}_{ij})\Big] = \frac{1}{Mn}\Bigg\{\frac{Mn-mn}{mn}\int k(w)\Big[m(x_{ij}+bw) - m(x_{ij})\Big]\int g\Big(x_{ij}+bw\Big)dw\Bigg\} \tag{3.54}$$

Using Taylor's series expansion about the point x_{ij}, the k^{th} order kernel can be derived as follows:

$$g\big(x_{ij}+bw\big) = g(x_{ij}) + g\prime(x_{ij})bw + \frac{1}{2}g\prime\prime(x_{ij})b^2w^2 + \cdots + \frac{1}{k!}g^k(x_{ij})b^kw^k + o(b^2) \tag{3.55}$$

Similarly,

$$m\big(x_{ij}+bw\big) = m(x_{ij}) + m\prime(x_{ij})bw + \frac{1}{2}m\prime\prime(x_{ij})b^2w^2 + \cdots + \frac{1}{k!}m^k(x_{ij})b^kw^k + o(b^2) \tag{3.56}$$

Expanding up to the 3^{rd} order kernels, equation (3.56) becomes

$$\big[m\big(x_{ij}+bw\big) - m(x_{ij})\big] = m\prime(x_{ij})bw + \frac{1}{2}m\prime\prime(x_{ij})b^2w^2 + \frac{1}{3!}m\prime\prime\prime(x_{ij})b^3w^3 + o(b^2) \tag{3.57}$$

In a similar manner, the expansion of equation (3.54) up to order $o(b^2)$ is given by:

$$E\sum_{i=n+1}^{N}\sum_{j=m+1}^{M}\big[m_1(\hat{x}_{ij})\big] = \frac{1}{Mn}\Bigg\{\frac{Mn-mn}{mn}\int k(w)\big(m\prime(x_{ij})bw + \frac{1}{2}m\prime\prime(x_{ij})b^2w^2\big)\big(g(x_{ij}) + g\prime(x_{ij})\big)bwdw\Bigg\} \tag{3.58}$$

Simplifying equation (3.58) gives:

$$E\sum_{i=n+1}^{N}\sum_{j=m+1}^{M}[m_1(\hat{x}_{ij})] = \frac{1}{Mn}\Big\{(\frac{Mn-mn}{mn})g(x_{ij})m\prime(x_{ij})b\int wk(w)dw + (\frac{Mn-mn}{mn})g\prime(x_{ij})m\prime(x_{ij})b^2\int w^2k(w)dw + (\frac{Mn-mn}{mn})\frac{1}{2}g(x_{ij})m\prime\prime(x_{ij})b^2 \times\int w^2k(w)dw + o(b^2)\Big\} \tag{3.59}$$

Using the conditions stated in equation (3.18), the derivation in equation (3.59) can further be simplified to obtain:

$$E\sum_{i=n+1}^{N}\sum_{j=m+1}^{M}\Big[m_1(\hat{x}_{ij})\Big] = \frac{1}{Mn}(\frac{Mn-mn}{mn})\Big[g\prime(x_{ij})m\prime(x_{ij}) + \frac{1}{2}g(x_{ij})m\prime\prime(x_{ij})\Big]b^2d_k + o(b^2) \tag{3.60}$$

Hence the expected value of the second term in equation (3.49) then becomes:

$$E\sum_{i=n+1}^{N}\sum_{j=m+1}^{M}[m_1(\hat{x}_{ij})] = \frac{1}{Mn}\Big\{(\frac{Mn-mn}{mn})\Big[\frac{1}{2g(\hat{x}_{ij})}m\prime\prime(x_{ij})g(x_{ij}) + \frac{g\prime(x_{ij})m\prime(x_{ij})}{g(\hat{x}_{ij})}\Big]b^2d_k + o(b^2)\Big\} \tag{3.61}$$

Simplifying equation (3.61) gives:

$$E\sum_{i=n+1}^{N}\sum_{j=m+1}^{M}[m_1(\hat{x}_{ij})] = \frac{1}{Mn}\Big\{(\frac{Mn-mn}{mn})b^2d_kC(x) + o(b^2)\Big\} \tag{3.62}$$

where $C(x) = [g(\hat{x}_{ij})]^{-1}\Big[\frac{1}{2}m\prime\prime(x_{ij})g(x_{ij}) + g\prime(x_{ij})m\prime(x_{ij})\Big]$ and $d_k = \int w^2k(w)dw$. Using equation of the bias given in (3.40) and the conditional expectation in equation (3.49), the following equation for the conditional bias of the estimator was obtained:

$$Bias\,(\hat{\bar{Y}}/x_{ij}) = \frac{1}{Mn}\Big\{(\frac{Mn-mn}{mn})b^2d_kC(x) + o(b^2)\Big\} \tag{3.63}$$

3.5 Asymptotic variance of the estimator of the population mean

Using equation (3.37), the variance of the estimator is given as

$$Var(\hat{\bar{\bar{Y}}}/x_{ij}) = Var\left\{ \frac{1}{Mn}\left\{ \sum_{i=1}^{n}\sum_{j=1}^{m} Y_{ij} + \sum_{i=n+1}^{N}\sum_{j=m+1}^{M} m\left(\hat{x}_{ij}\right) \right\} \right\} \tag{3.64}$$

$$= \left(\frac{1}{Mn}\right)^2 Var\left\{ \sum_{i=1}^{n}\sum_{j=1}^{m} Y_{ij} + \sum_{i=n+1}^{N}\sum_{j=m+1}^{M} m\left(\hat{x}_{ij}\right) \right\} \tag{3.65}$$

where $m\left(\hat{x}_{ij}\right)$ is given by

$$m\Big(\hat{x}_{ij}\Big) = \frac{1}{g(\hat{x}_{ij})}\left[\frac{1}{mnb}\sum_{i}\sum_{j} K\Big(\frac{X_{ij}-x_{ij}}{b}\Big)Y_{ij}\right] \tag{3.66}$$

where $g(\hat{x}_{ij})$ is the estimated marginal density of auxiliary variables X_{ij}, see equation (3.33). Re-writing the regression model $Y_{ij} = m(X_{ij}) + e_{ij}$ as $Y_{ij} = m(x_{ij}) + \Big[m(X_{ij}) - m(x_{ij})\Big] + e_{ij}$ and substituting in equation (3.66) leads to

$$\begin{aligned} Var(\hat{\bar{\bar{Y}}}/x_{ij}) = \left(\frac{1}{Mn}\right)^2 Var\Bigg\{ & \sum_{i=1}^{n}\sum_{j=1}^{m} Y_{ij} \\ & + \left(\frac{1}{mnb\, g(\hat{x}_{ij})}\right)\left\{ \sum_{i=n+1}^{N}\sum_{j=m+1}^{M} g(\hat{x}_{ij})m(x_{ij}) + m_1(\hat{x}_{ij}) + m_2(\hat{x}_{ij}) \right\}\Bigg\} \end{aligned} \tag{3.67}$$

From equation (3.48),

$$m_2(\hat{x}_{ij}) = \frac{1}{mnb}\sum_{i=n+1}^{N}\sum_{j=m+1}^{M} K\Big(\frac{X_{ij}-x_{ij}}{b}\Big)e_{ij} \tag{3.68}$$

Hence

$$Var\sum_{i=n+1}^{N}\sum_{j=m+1}^{M}\left[m_2(\hat{x}_{ij})\right] = \frac{1}{(Mn)^2}\left(\frac{Mn-mn}{mnb}\right)^2 \sum_{i=1}^{n}\sum_{j=1}^{m} Var(D_x) \tag{3.69}$$

where $D_x = K\left(\frac{X_{ij}-x_{ij}}{b}\right)e_{ij}$. Expressing equation (3.69) in terms of expectation the following equation is obtained

$$Var\sum_{i=n+1}^{N}\sum_{j=m+1}^{M}\left[m_2(\hat{x}_{ij})\right] = \frac{1}{(Mn)^2}\left[\frac{(Mn-mn)^2}{mnb^2}\right]\left\{E[D_x]^2-[E(D_x)]^2\right\} \quad (3.70)$$

Using the fact that the conditional expectation $E(e_{ij}/X_{ij}) = 0$, the second term in equation (3.70) reduces to zero. Therefore,

$$Var\sum_{i=n+1}^{N}\sum_{j=m+1}^{M}\left[m_2(\hat{x}_{ij})\right] = \frac{1}{(Mn)^2}\left[\frac{(Mn-mn)^2}{mnb^2}\right]\sigma_{ij}^2 \quad (3.71)$$

where $E\left(e_{ij}/X_{ij}\right)^2 = \sigma_{ij}^2$.

Let $X = X_{ij}$, and $x = x_{ij}$ and make the following substitutions

$$\left.\begin{aligned} w &= \tfrac{X-x}{b} \\ X-x &= bw \\ dX &= bdw. \end{aligned}\right\} \quad (3.72)$$

so that

$$Var\sum_{i=n+1}^{N}\sum_{j=m+1}^{M}\left[m_2(\hat{x}_{ij})\right] = \frac{(Mn-mn)^2}{mnb^2(Mn)^2}\int K\left(\frac{X-x}{b}\right)^2\sigma_x^2\, g(X)dX \quad (3.73)$$

Equation (3.73) can also be written as

$$Var\sum_{i=n+1}^{N}\sum_{j=m+1}^{M}\left[m_2(\hat{x}_{ij})\right] = \frac{(Mn-mn)^2}{mnb^2(Mn)^2}\int K(w)^2\sigma_x^2\, g\left(x+bw\right)bdw \quad (3.74)$$

which can be simplified to get:

$$Var\sum_{i=n+1}^{N}\sum_{j=m+1}^{M}\left[m_2(\hat{x}_{ij})\right] = \frac{(Mn-mn)^2}{mnb(Mn)^2}\int K(w)^2\sigma_x^2\, g(x)dw + o\left(\frac{1}{mnb}\right) \quad (3.75)$$

Following the same procedure for getting the variance of $m_2(\hat{x}_{ij})$,
$Var\sum_{i=n+1}^{N}\sum_{j=m+1}^{M}\left[m_1(\hat{x}_{ij})\right]$ can similarly be obtained as follows:

$$Var\sum_{i=n+1}^{N}\sum_{j=m+1}^{M}\left[m_1(\hat{x}_{ij})\right]=\frac{1}{(Mn)^2}Var\sum_{i=n+1}^{N}\sum_{j=m+1}^{M}\left[\frac{1}{mnb}\sum_{i=1}^{n}\sum_{j=1}^{m}K\left(\frac{X_{ij}-x_{ij}}{b}\right)\right]\times\left[m(X_{ij})-m(x_{ij})\right] \tag{3.76}$$

Equation (3.76) can be re-written as

$$Var\sum_{i=n+1}^{N}\sum_{j=m+1}^{M}\left[m_1(\hat{x}_{ij})\right]=\frac{(Mn-mn)^2}{mnb^2(Mn)^2}VarK\left(\frac{X_{ij}-x_{ij}}{b}\right)^2\left[m(X_{ij})-m(x_{ij})\right]^2g(X)dX \tag{3.77}$$

where $X = bw + x$ so that $dX = bdw$. Changing variables and applying Taylor's series expansion about the point x_{ij}

$$Var\sum_{i=n+1}^{N}\sum_{j=m+1}^{M}\left[m_1(\hat{x}_{ij})\right]=\frac{(Mn-mn)^2}{mnb^2(Mn)^2}\int K(w^2)\left[m(x+bw)-m(x)\right]^2\times g(x+bw)dw \tag{3.78}$$

which gives

$$Var\sum_{i=n+1}^{N}\sum_{j=m+1}^{M}\left[m_1(\hat{x}_{ij})\right]=\frac{(Mn-mn)^2}{mnb^2(Mn)^2}\int K(w^2)\left[m(x)+m\prime(x)bw+\cdots-m(x)\right]^2\left(g(x)+g\prime(x)bw\right)dw \tag{3.79}$$

Following the procedure by (Bach et al., 1996) and simplifying, equation (3.79) reduces to

$$Var\sum_{i=n+1}^{N}\sum_{j=m+1}^{M}\left[m_1(\hat{x}_{ij})\right]=O\left[\frac{(Mn-mn)^2b^2}{mnb}\right] \tag{3.80}$$

For large samples, as $n \to N, m \to M$ and $b \to 0$, then $mnb \to \infty$. Hence the variance in equation (3.79) asymptotically tends to zero, i.e, $Var\sum_{i=1}^{N}\sum_{j=1}^{M}\left[m_1(\hat{x}_{ij})\right] \to 0$

so that the variance of the estimator of the population mean reduces to

$$Var\left(\hat{\overline{\overline{Y}}}/x_{ij}\right) = \frac{(Mn-mn)^2}{mnb(Mn)^2}\sum_{i=n+1}^{N}\sum_{j=m+1}^{M} Var\left[m(x_{ij}) + \frac{m_1(\hat{x}_{ij}) + m_2(\hat{x}_{ij})}{g(\hat{x}_{ij})}\right] \tag{3.81}$$

Simplifying equation (3.81) leads to

$$Var\left(\hat{\overline{\overline{Y}}}/x_{ij}\right) = \frac{(Mn-mn)^2}{mnb(Mn)^2\left[g(\hat{x}_{ij})\right]^2} Var\left\{\sum_{i=n+1}^{N}\sum_{j=m+1}^{M}\left[m_2(\hat{x}_{ij})\right]\right\} \tag{3.82}$$

Substituting equation (3.75) into (3.82) yields the following:

$$Var\left(\hat{\overline{\overline{Y}}}/x_{ij}\right) = \frac{1}{(Mn)^2}\left\{\frac{(Mn-mn)^2\int K(w)^2\sigma^2_{x_{ij}}dw}{mnbg(\hat{x}_{ij})} + o\left[\frac{(Mn-mn)^2}{mnb} + \frac{1}{mnb}\right]\right\} \tag{3.83}$$

This can further be simplified to get

$$Var\left(\hat{\overline{\overline{Y}}}/x_{ij}\right) = \frac{1}{(Mn)^2}\left\{\frac{(Mn-mn)^2H(w)\sigma^2_{x_{ij}}}{mnbg(\hat{x}_{ij})} + o\left[\frac{(Mn-mn)^2}{mnb} + \frac{1}{mnb}\right]\right\} \tag{3.84}$$

where $H(w) = \int K(w)^2dw$.

It is notable that the variance term still depends on the marginal density function, $g(x_{ij})$ of the auxiliary variables X_{ij}. It can also be observed that the variance is inversely related to the smoothing parameter, b. This implies that an increase in b results in a smaller variance. However, increasing the bandwidth would give a larger bias. Hence the bias and the variance of the estimated population mean are inversely proportional based on the bandwidth. A bandwidth that provides a compromise between the two measures would therefore be desirable.

3.6 MSE of the Estimator of Finite Population Mean

The MSE combines the bias and the variance terms of the estimator, $(\hat{\overline{\overline{Y}}})$, that is,

$$MSE(\hat{\overline{\overline{Y}}}) = E\left(\hat{\overline{\overline{Y}}} - \overline{\overline{Y}}\right)^2 \tag{3.85}$$

Equation (3.85) can be re-written as:

$$MSE(\hat{\overline{\overline{Y}}}) = E\big(\hat{\overline{\overline{Y}}} - E[\hat{\overline{\overline{Y}}}] + E[\hat{\overline{\overline{Y}}}] - \overline{\overline{Y}}\big)^2 \tag{3.86}$$

Expanding equation (3.86) gives

$$MSE(\hat{\overline{\overline{Y}}}) = E\big(\hat{\overline{\overline{Y}}} - E[\overline{\overline{Y}}]\big)^2 + E\big(E[\hat{\overline{\overline{Y}}} - [\overline{\overline{Y}}]]\big)^2 + 2E\big(\hat{\overline{\overline{Y}}} - E[\overline{\overline{Y}}]\big)\big([\hat{\overline{\overline{Y}}}] - \overline{\overline{Y}}\big) \tag{3.87}$$

It can be deduced from equation (3.87) that

$$MSE(\hat{\overline{\overline{Y}}}) = Var(\hat{\overline{\overline{Y}}}) + Bias^2(\hat{\overline{\overline{Y}}}) + 0 \tag{3.88}$$

Using the theorem due to Dorfman (1992), the conditional mean squared error is thus given by

$$MSE(\hat{\overline{\overline{Y}}}/x_{ij}) = Var(\hat{\overline{\overline{Y}}}/x_{ij}) + \Big(Bias(\hat{\overline{\overline{Y}}}/x_{ij})\Big)^2 \tag{3.89}$$

Combining the conditional bias in equation (3.63) and the conditional variance in equation (3.84) and conditioning on the auxiliary values x_{ij} of the auxiliary variables X_{ij}, the following equation is obtained

$$\begin{aligned} MSE\big(\hat{\overline{\overline{Y}}}/X_{ij} = x_{ij}\big) = \frac{1}{(Mn)^2}\Bigg\{ & \frac{(Mn-mn)^2 H(w)\sigma^2_{x_{ij}}}{mnbg(\hat{x}_{ij})} \\ & + o\Bigg(\frac{1}{Mn}\Big\{\frac{(Mn-mn)^2}{mnb} + \frac{1}{mnb}\Big\}\Bigg)\Bigg\} \\ & + \Big(Bias(\hat{\overline{\overline{Y}}}/x_{ij})\Big)^2 \end{aligned} \tag{3.90}$$

which gives

$$\begin{aligned} MSE\big(\hat{\overline{\overline{Y}}}/X_{ij} = x_{ij}\big) = \frac{1}{(Mn)^2}\Bigg\{ & \frac{(Mn-mn)^2 H(w)\sigma^2_{x_{ij}}}{mnbg(\hat{x}_{ij})} \\ & + \Bigg[\frac{(Mn-mn)^2}{4(mn)^2(Mn^2)} b^2 d_k^2 \Bigg[m''(x_{ij})g(x_{ij}) + \frac{2g'(x_{ij})m'(x_{ij})}{g(\hat{x}_{ij})}\Bigg]^2 \\ & + o\Bigg(\frac{1}{Mn}\Big\{\frac{(Mn-mn)^2}{mnb} + \frac{1}{mnb}\Big\}\Bigg)\Bigg]\Bigg\} \end{aligned} \tag{3.91}$$

Where $H(w) = \int K(w)^2 dw$, $d_k = \int w^2 K(w) dw$,
$C(x) = [g(\hat{x}_{ij})]^{-1}\Big[\frac{1}{2}m''(x_{ij})g(x_{ij}) + g'(x_{ij})m'(x_{ij})\Big]$ as used earlier. Therefore from

equation (3.91) it can be noted that if the sample size is large, that is as $n \to N$ and $m \to M$, the MSE of $\hat{\bar{\bar{Y}}}$ due to the kernel tends to zero for a sufficiently small bandwidth, b. The estimator $\hat{\bar{\bar{Y}}}$ is therefore asymptotically consistent since its MSE converges to zero in probability.

However, it can be observed that the MSE is still a function of $m''(x_{ij})$, $g(x_{ij})$, $g'(x_{ij})$ and $m'(x_{ij})$ which are unknown. Optimal bandwidth, b, may not eliminate these functions either. In the next chapter, weighting method of compensating for non-response in two stage cluster sampling is proposed using modified transformation of data method.

CHAPTER FOUR

BOUNDARY BIAS CORRECTION USING THE WEIGHTING METHOD UNDER RANDOM NON RESPONSE IN TWO STAGE CLUSTER SAMPLING WITH REPLACEMENT

4.1 Introduction

In this chapter, transformation of data method is used to develop an estimator for finite population mean. It is one of the methods of reducing boundary bias in kernel density estimations as discussed in the literature review.

4.2 Proposed Estimator of Finite Population Mean using Modified Transformation of Data Method

Consider a finite population of size N consisting of M clusters with N_j elements in the i^{th} cluster. Let y_{ij} denote the value of the survey variable y for unit j in cluster i, for $i = 1, 2, ..., N; j = 1, 2, ..., M$. To estimate the non-response values in the second stage of sampling, a regression model given in (2.2) is used. Auxiliary data is assumed to be known throughout the study and is therefore used to predict the non-response values. The estimator proposed in chapter three due to Nadaraya-Watson suffers from boundary bias. Following the work of (Baszczyńska, 2015), a non-parametric regression estimator for the finite population mean that resolves boundary bias is obtained. The function of auxiliary variables given in equation (4.1) below is used to predict the non-response values of the study variable Y_{ij}; the estimator is defined by

$$m_{TDM}(\hat{x}_{ij}) = \frac{1}{mnb}\sum_{i=1}^{n}\sum_{j=1}^{m}\left\{K\left(\frac{x_{ij}-X_{ij}}{b}\right)+K\left(\frac{x_{ij}+g(X_{ij})}{b}\right)\right\} \tag{4.1}$$

where $m_{TDM}(\hat{x})$ is the function of auxiliary variables due to modified transformation of data method proposed. Following the work of (Cowling and Hall, 1996), data should be generated beyond the left endpoint of the support of the density function g such that the data provides a natural adjustment of the density g outside its support. The data

generation procedure combines the transformation and the reflection of data methods. To do this, first transform the original data $X_{ij}, j = 1, 2, \cdots, n; i = 1, 2, \cdots, m$ to $g(X_{ij}), j = 1, 2, \cdots, n; i = 1, 2, \cdots, m$ while retaining the original data where g is a non-negative, continuous and monotonically increasing function from $[0, \infty)$ to $[0, \infty)$. Secondly, reflect $g(X_{11}), \cdots, g(X_{mn})$ around the origin so that we have $-g(X_{11}), \cdots, -g(X_{mn})$. Consequently, using the enlarged data sample $-g(X_{ij}), j = 1, 2, \cdots, n; i = 1, 2, \cdots, m$ the new mean estimator of the population is defined by

$$\hat{\bar{Y}} = \frac{1}{Mn}\left\{ \sum_{i=1}^{n}\sum_{j=1}^{m} Y_{ij} + \sum_{j=n+1}^{N}\sum_{i=m+1}^{M} \hat{y}_{ij} \right\} \tag{4.2}$$

where $\hat{y_{ij}}$ represents the estimator of the non-response units and can be re-written as

$$\hat{y}_{ij} = m_{TDM}(\hat{x}_{ij}) = \sum_{i}\sum_{j} W_{ij}^{*}(x_{ij}) Y_{ij} \tag{4.3}$$

where $W_{ij}^{*}(x_{ij}) = \sum_{i=n+1}^{N}\sum_{j=m+1}^{M} \frac{\left\{K\left(\frac{x_{ij}-X_{ij}}{b}\right)+K\left(\frac{x_{ij}+g(X_{ij})}{b}\right)\right\}}{\sum_{i=n+1}^{N}\sum_{j=m+1}^{M}\left\{K\left(\frac{x_{ij}-X_{ij}}{b}\right)+K\left(\frac{x_{ij}+g(X_{ij})}{b}\right)\right\}}$ are the modified weights arising from the proposed procedure. From equation (4.3), the following equation is obtained

$$m_{TDM}(\hat{x}_{ij}) = \sum_{i=n+1}^{N}\sum_{j=m+1}^{M} \frac{\left\{K\left(\frac{x_{ij}-X_{ij}}{b}\right) + K\left(\frac{x_{ij}+g(X_{ij})}{b}\right)\right\} Y_{ij}}{\sum_{i=n+1}^{N}\sum_{j=m+1}^{M}\left\{K\left(\frac{x_{ij}-X_{ij}}{b}\right) + K\left(\frac{x_{ij}+g(X_{ij})}{b}\right)\right\}} \tag{4.4}$$

Using equation (4.4), the estimator of the population mean is therefore given by

$$\begin{aligned} \hat{\bar{Y}} = \frac{1}{Mn}\Bigg\{ & \sum_{i=1}^{n}\sum_{j=1}^{m} Y_{ij} \\ & + \sum_{i=n+1}^{N}\sum_{j=m+1}^{M} \frac{\left\{K\left(\frac{x_{ij}-X_{ij}}{b}\right) + K\left(\frac{x_{ij}+g(X_{ij})}{b}\right)\right\} Y_{ij}}{\sum_{i=n+1}^{N}\sum_{j=m+1}^{M}\left\{K\left(\frac{x_{ij}-X_{ij}}{b}\right) + K\left(\frac{x_{ij}+g(X_{ij})}{b}\right)\right\}} \Bigg\} \end{aligned} \tag{4.5}$$

In what follows, some properties of the proposed estimator are derived.

4.3 Asymptotic Bias of the Transformation of Data Method Estimator

Boundary bias occurs in the interval $[0, b)$ due to lack of data following the reduction of such data at this interval. This implies that the density function has continuity on $[0, \infty)$ and is 0 for $x_{ij} < 0$. Due to reduced amount of data, the resulting estimators are biased. This is possible if the selected bandwidth is greater than the value of x_{ij} i.e if $b > x_{ij}$. Consider the Nadaraya-Watson estimator given by equation (3.20). In addition, consider $m(\hat{ab})$ for $a \in [0, 1)$ where $x_{ij} = ab$ so that for $w = \frac{x_{ij} - X_{ij}}{b}$ one can obtain

$$0 \leq X_{ij} \leq \infty \Rightarrow 0 \leq b(a - w) \leq \infty \tag{4.6}$$

so that $-\infty \leq w \leq a$. Next, consider a kernel estimator given in equation (2.3) which has the support $[-1, 1]$; this means the variable w must be contained in the interval $[-1, 1]$, so that for $a \in [0, 1)$ we have $-1 \leq w \leq a$. Since the main problem is to estimate the non-response component of the proposed estimator, the following theorem due to (Cowling and Hall, 1996), under certain conditions on $g(.)$ and $m(.)$ outlined below is applied.

Theorem 4.2 *Assume that $m''(\hat{x_{ij}})$ and $g''(\hat{x_{ij}})$ exist and are continuous, where $g(x_{ij}) = x_{ij} + dx_{ij}^2 + Ad^2x_{ij}^3$ such that $A > 0$ and $d = \frac{m'(0)}{m(0)}$. Assume that $g^{-1}(0) = 0$, $g'(0) = 1$, $g''(0) = \frac{2m'(0)}{m(0)}$ and $m^{(0)} = m$, $g^{(0)} = g$, where g^{-1} is the inverse function of g while $m^{(i)}$ and $g^{(i)}$ are the i^{th} derivatives of m and g respectively for $i \geq 0$. Furthermore, let $x = ab$, where $0 \leq a \leq 1$. Assume the kernel function K is non-negative, symmetric function with support $[-1, 1]$ such that $\int K(w)dw = 1$, $\int wK(w)dw = 0$ and $0 < \int w^2k(w)dw < \infty$.*

The second term in equation (4.2) represents the non-response component of the estimator. Since $\hat{y}_{ij} = m_{TDM}(\hat{x}_{ij})$ as expressed in equation (4.3), the theorem stated above due to (Cowling and Hall, 1996) can be used to obtain the expected value of the non-response component as follows

$$E\big(m_{TDM}(\hat{x}_{ij})\big) = \frac{1}{b}E\left[K\left(\frac{x_{ij} - X_{ij}}{b}\right) + K\left(\frac{x_{ij} + g(X_{ij})}{b}\right)\right] \tag{4.7}$$

which can be expanded to obtain

$$E\big(m_{TDM}(\hat{x}_{ij})\big) = m(x_{ij})\int_{-1}^{a} K(w)dw - bm\prime(x_{ij})\int_{-1}^{a} wK(w)dw + \frac{m\prime\prime(x_{ij})b^2}{2}\int_{-1}^{a} w^2K(w)dw + \frac{1}{b}EK\left(\frac{x_{ij}+g(X_{ij})}{b}\right) + o\big(b^2\big) \tag{4.8}$$

and

$$\frac{1}{b}EK\left(\frac{x_{ij}+g(X_{ij})}{b}\right) = \frac{1}{b}\int_0^{\infty} K\left(\frac{x_{ij}+g(y_{ij})}{b}\right)f(y)dy \tag{4.9}$$

Using change of variables technique and simplifying, equation (4.9) becomes,

$$\frac{1}{b}EK\left(\frac{x_{ij}+g(X_{ij})}{b}\right) = \int_a^1 K(w)\frac{f\Big(g^{-1}(w-a)b\Big)}{g\prime\Big(g^{-1}(w-a)b\Big)}dw \tag{4.10}$$

which expands to

$$\begin{aligned}\frac{1}{b}EK\left(\frac{x_{ij}+g(X_{ij})}{b}\right) = \int_a^1 K(w)\Bigg\{&\frac{m\big(g^{-1}(0)\big)}{g\prime\big(g^{-1}(0)\big)} + (w-a)b \\ &\times \frac{g\prime\big(g^{-1}(0)\big)m\prime\big(g^{-1}(0) - g\prime\prime\big(g^{-1}(0)m\big(g^{-1}(0)\big)\big)}{\Big[g\prime\big(g^{-1}(0)\big)\Big]^3} \\ &+ \frac{b^2}{2}(w-a)^2\Bigg[\frac{g\prime\big(g^{-1}(0)m\prime\prime\big(g^{-1}(0)\big) - g\prime\prime\big(g^{-1}(0)m\big(g^{-1}(0)\big)\big)}{\Big[g\prime\big(g^{-1}(0)\big)\Big]^4} \\ &- \frac{3g\prime\prime\big(g^{-1}(0)\big)\Big[g\prime\big(g^{-1}(0)m\prime\big(g^{-1}(0)\big) - g\prime\prime\big(g^{-1}(0)m\big(g^{-1}(0)\big)\big)\Big]}{\Big[g\prime\big(g^{-1}(0)\big)\Big]^5}\Bigg]\Bigg\}dw \\ &+ o(b^2)\end{aligned} \tag{4.11}$$

which on simplification yields

$$\begin{aligned}\frac{1}{b}EK\left(\frac{x_{ij}+g(X_{ij})}{b}\right) &= m(0)\int_a^1 K(w)dw + b\int_a^1 (w-a)K(w)dw \\ &\quad\times\Big[m\prime(0) - g\prime\prime(0)m(0)\Big] + \frac{b^2}{2}\int_a^1 (w-a)^2K(w)dw \\ &\quad\times\Big\{m\prime\prime(0) - g\prime\prime(0)m(0) - 3g\prime\prime(0)\Big[m\prime(0) \\ &\quad - g\prime\prime(0)m(0)\Big]\Big\} + o(b^2)\end{aligned} \tag{4.12}$$

Substituting equation (4.12) in equation (4.8) the following is obtained

$$\begin{aligned}E\big(m_{TDM}(\hat{x}_{ij})\big) &= m(x_{ij})\int_{-1}^a K(w)dw + m(0)\int_a^1 K(w)dw \\ &\quad + b\Big\{-m\prime(x_{ij})\int_{-1}^a wK(w)dw + \int_a^1 (w-a)K(w)dw \\ &\quad\times\Big[m\prime(0) - g\prime\prime(0)m(0)\Big]\Big\} + \frac{b^2}{2}m\prime\prime(x_{ij})\int_{-1}^a w^2K(w)dw \\ &\quad + \frac{b^2}{2}\int_a^1 (w-a)^2K(w)dw + \Big\{m\prime\prime(0) - g\prime\prime(0)m(0) \\ &\quad - 3g\prime\prime(0)\Big[m\prime(0) - g\prime\prime(0)m(0)\Big]\Big\} + o(b^2)\end{aligned} \tag{4.13}$$

Since $f\prime\prime(0)$ exists and is continuous near 0, then for $x = ab$ we have

$$\left.\begin{aligned} m(0) &= m(x_{ij}) - abm\prime(x_{ij}) + \tfrac{(ab)^2}{2}m\prime\prime(x_{ij}) + o(b^2) \\ m\prime(x_{ij}) &= m\prime(0) + abm\prime\prime(0) + o(b) \\ m\prime\prime(x_{ij}) &= m\prime\prime(0) + o(1)\end{aligned}\right\}$$

Simplifying equation (4.13) gives

$$\begin{aligned}E\big(m_{TDM}(\hat{x}_{ij})\big) &= m(x_{ij}) + b\int_a^1 (w-a)K(w)dw\Big[2m\prime(0) - g\prime\prime(0)m(0)\Big] \\ &\quad + \frac{b^2}{2}m\prime\prime(0)\int_{-1}^1 w^2K(w)dw - \frac{b^2}{2}\int_a^1 (w-a)^2K(w)dw \\ &\quad\times\Big\{g\prime\prime(0)m(0) + 3g\prime\prime(0)\Big[m\prime(0) - g\prime\prime(0)m(0)\Big]\Big\} + o(b^2)\end{aligned} \tag{4.14}$$

It follows that the bias of the estimator of the non-response component in equation (4.5) can be expressed as

$$\begin{aligned} E\big(m_{TDM}(\hat{x}_{ij})\big) - m(x_{ij}) = b\int_a^1 (w-a)K(w)dw\Big[2m\prime(0) - g\prime\prime(0)m(0)\Big] \\ + \frac{b^2}{2}m\prime\prime(0)\int_{-1}^1 w^2K(w)dw - \frac{b^2}{2}\int_a^1 (w-a)^2K(w)dw \\ \times\Big\{g\prime\prime(0)m(0) + 3g\prime\prime(0)\Big[m\prime(0) - g\prime\prime(0)m(0)\Big]\Big\} + o(b^2) \end{aligned} \tag{4.15}$$

As $b \to 0$, it is noted that $E\big(m_{TDM}(\hat{x}_{ij})\big) - m(x_{ij})$ is approximately equal to zero. This shows that for the bias to reduce, the bandwidth must tend to zero as the sample size increases, as $b \to 0$, $mn \to \infty$.

4.4 Asymptotic Variance of the Transformation of Data Method Estimator

The variance of the transformed estimator proposed is given as

$$\begin{aligned} Var\big(m_{TDM}(\hat{x}_{ij})\big) = \frac{1}{(mn)^2b^2}Var\Bigg\{\sum_{i=n+1}^{N}\sum_{j=m+1}^{M} K\Big(\frac{x_{ij}-X_{ij}}{b}\Big) \\ + K\Big(\frac{x_{ij}+g(X_{ij})}{b}\Big)\Bigg\} \end{aligned} \tag{4.16}$$

$$\begin{aligned} Var\big(m_{TDM}(\hat{x}_{ij})\big) = \frac{1}{(mn)^2b^2}\sum_{i=n+1}^{N}\sum_{j=m+1}^{M} E\Bigg\{K\Big(\frac{x_{ij}-X_{ij}}{b}\Big) \\ + K\Big(\frac{x_{ij}+g(X_{ij})}{b}\Big) - E\Bigg[K\Big(\frac{x_{ij}-X_{ij}}{b}\Big) \\ + K\Big(\frac{x_{ij}+g(X_{ij})}{b}\Big)\Bigg]\Bigg\}^2 \end{aligned} \tag{4.17}$$

which can be re-written as follows

$$\begin{aligned} Var\big(m_{TDM}(\hat{x}_{ij})\big) &= \frac{(Mn-mn)^2}{(mn)^2b^2}E\left[K\Big(\frac{x_{ij}-X_{ij}}{b}\Big)+K\Big(\frac{x_{ij}+g(X_{ij})}{b}\Big)\right]^2 \\ &\quad -\frac{(Mn-mn)^2}{(mn)^2b^2}\left\{E\left[K\Big(\frac{x_{ij}-X_{ij}}{b}\Big)+K\Big(\frac{x_{ij}+g(X_{ij})}{b}\Big)\right]\right\}^2 \\ &= A+B \end{aligned} \tag{4.18}$$

where

$$\begin{aligned} A &= \frac{(Mn-mn)^2}{mnb^2}\int\left[K\Big(\frac{x_{ij}-y_{ij}}{b}\Big)+K\Big(\frac{x_{ij}+g(y_{ij})}{b}\Big)\right]^2 f(y)dy \\ &= \frac{(Mn-mn)^2}{mnb^2}\left\{\int\left[K\Big(\frac{x_{ij}-y_{ij}}{b}\Big)^2 f(y)dy+\int K\Big(\frac{x_{ij}+g(y_{ij})}{b}\Big)^2\right.\right. \\ &\quad \left.\left.\times f(y)dy\right]+2\int K\Big(\frac{x_{ij}-y_{ij}}{b}\Big)K\Big(\frac{x_{ij}+g(y_{ij})}{b}\Big)f(y)dy\right\} \\ &= A_1+A_2 \end{aligned} \tag{4.19}$$

and

$$B = -\frac{(Mn-mn)^2}{(mn)^2b^2}\left\{E\left[K\Big(\frac{x_{ij}-X_{ij}}{b}\Big)+K\Big(\frac{x_{ij}+g(X_{ij})}{b}\Big)\right]\right\}^2 \tag{4.20}$$

$$A_1 = \frac{(Mn-mn)^2}{mnb^2}\int\left[K\Big(\frac{x_{ij}-y_{ij}}{b}\Big)+K\Big(\frac{x_{ij}+g(y_{ij})}{b}\Big)\right]^2 f(y)dy \tag{4.21}$$

Equation (4.21) can be expanded to get

$$\begin{aligned} A_1 &= \frac{(Mn-mn)^2}{mnb^2}\left\{\int\left[K\Big(\frac{x_{ij}-y_{ij}}{b}\Big)^2 f(y)dy+\int K\Big(\frac{x_{ij}+g(y_{ij})}{b}\Big)^2\right.\right. \\ &\quad \left.\left.\times f(y)dy\right]\right\}+2\left\{\frac{(Mn-mn)^2}{mnb^2}\int\left[K\Big(\frac{x_{ij}-y_{ij}}{b}\Big)K\Big(\frac{x_{ij}+g(y_{ij})}{b}\Big)\right]\right. \\ &\quad \left.\times f(y)dy\right\} \end{aligned} \tag{4.22}$$

Using change of variables technique, equation (4.22) can be expressed as

$$A_1 = \frac{(Mn-mn)^2}{mnb^2}\Big[b\int_{-1}^{a} K(w)^2 m(x_{ij}-bw)dw + b\int_{a}^{1} K(w)^2 \times \frac{m(g^{-1}(x_{ij}-bw))}{g\prime(x_{ij}-bw))}dw\Big] \tag{4.23}$$

which on simplification reduces to

$$A_1 = \frac{(Mn-mn)^2}{mnb}\int_{-1}^{1} K(w)^2 dw + o(\frac{1}{mnb}) \tag{4.24}$$

Next we have,

$$A_2 = 2\frac{(Mn-mn)^2}{mnb^2}\int_{-1}^{a} K(w)K\Big(\frac{x_{ij}+g(a-w)b}{b}\Big) \times m(a-w)bdw \tag{4.25}$$

Since $g\prime\prime(.)$ is continuous, Taylor's series expansion gives

$$g(a-w)b = (a-w)b + o(b^2) \tag{4.26}$$

Replacing $g(a-w)b$ given by equation (4.26) in equation (4.25) yields

$$A_2 = 2\frac{(Mn-mn)^2}{mnb^2}\int_{-1}^{a} K(w)K\big((2a-w)+o(b)\big)m\big((a-w)b\big)dw \tag{4.27}$$

which reduces on simplification to

$$A_2 = 2\frac{(Mn-mn)^2}{mnb^2}\int_{-1}^{1} K(w)K(2a-w)dw + o(\frac{1}{mnb}) \tag{4.28}$$

Next is to evaluate B which as earlier outlined in equation (4.20) is given by

$$B = -\frac{(Mn-mn)^2}{(mn)^2b^2}\left\{E\left[K\Big(\frac{x_{ij}-X_{ij}}{b}\Big)+K\Big(\frac{x_{ij}+g(X_{ij})}{b}\Big)\right]\right\}^2 \tag{4.29}$$

Applying Taylor's series expansion and following the same procedure as for A, B would simplify to

$$B = o(\frac{1}{mnb}) \tag{4.30}$$

Hence putting together A i.e $A_1 + A_2$ and B we have

$$Var\big(m_{TDM}(\hat{x}_{ij})\big) = A + B = \frac{\big(Mn - mn\big)^2}{mnb}\Big[\int_{-1}^{1} K(w)^2 dw + 2\int_{-1}^{a} K(w)K\big(2a - w\big)dw\Big] + o\big(\frac{1}{mnb}\big) \tag{4.31}$$

It can be noted that $Var\big(m_{TDM}(\hat{x}_{ij})\big)$ is decreasing in mnb. This is because as $mn \to \infty$, the bandwidth $b \to 0$ and hence $mnb \to \infty$; that is, the bandwidth decreases but not at a faster rate than the sample size. Thus for a large sample size, the variance is reduced significantly.

4.5 Conclusion

It is noted that the the variance is decreasing in mnb whereas the bias is increasing in b. Both the bias and the variance must become small as $mn \to \infty$ for the estimator to be optimal. Therefore, as $mn \to \infty$, the bandwidth must tend to zero, $b \to 0$ and $mnb \to \infty$. That means the bandwidth must decrease but not at a faster rate than the sample size. This suffices to establish the consistency of the transformation of data method estimator proposed. That is, for all x_{ij}, $m_{TDM}(\hat{x}_{ij}) \to m(x_{ij})$ in probability as $mn \to \infty$.

CHAPTER FIVE

SIMULATION STUDY, PRESENTATION AND DISCUSSION OF THE RESULTS

5.1 Introduction

In this chapter, simulation and discussion of the results for the various estimators is carried out.

5.2 Significance of Simulation

The comparison of the performance of the various estimators of finite population mean in terms of the bias, variance, the MSE and the confidence interval lengths may not be easy to establish due to the complex nature of their asymptotic expressions. This necessitated the simulation study to be conducted. The simulation results give researchers an insight on the performance of a given estimator of a population parameter. In this study, the aim of the simulation exercise was to conduct a comparative study on the efficiency of the proposed estimator against Nadaraya-Watson estimator and modified transformation of data estimators. The data were generated using both linear and non-linear data functions. Non-linear data functions included quadratic, sine, exponential, bump and jump functions. The results are presented using graphs and tables.

5.3 Description of the simulation experiment

To obtain the estimator for the finite population mean, $\bar{\bar{Y}}$, the auxiliary variables X_{ij} were generated as identically and independently distributed random variables on $U(0,1)$. The population consisted of 30 clusters. In stage one, a sample of $m_i = 10$ clusters was chosen by simple random sampling with replacement which constituted primary sampling units (PSUs).

In stage two, from each selected cluster, say $i, (i = 1, \cdots, m_i)$ a sample $m_{ij}, j =$

$1, 2, \cdots, 120$ from $j = 1, 2, \cdots, m_k, \cdots, 1200$ was selected, that is, the j^{th} sample from a fixed selected i^{th} cluster was selected using $SRSWR$ from a total $M_k = 1200$ elements.

Consider the survey variable $Y_{ij}, j = 1, 2, \cdots, m_k, \cdots, M_k$ that are known only for the respondents in the sample. Using known auxiliary variables, $X_{ij}, j = 1, 2, \cdots, m_k$, non-response values were generated using the model $\hat{Y}_{ij} = m(\hat{x}_{ij}) + \hat{e}_{ij}$ using SRSWR within the i^{th} cluster.

Moreover, let $K(u) \sim U[0, 1], \hat{e}_{ij} \sim N(0, 1)$ such that the estimator of a finite population mean is given by equation (3.7) where $m(\hat{x}_{ij})$ is a function of auxiliary random variables generated using different data functions outlined in the following subsection. This procedure was repeated iteratively to obtain $\hat{\bar{Y}}_{i1}, \cdots, \hat{\bar{Y}}_{in}$. 95% confidence intervals (CI) were then constructed for the estimators of population means $\hat{\bar{Y}}_i, i = 1, 2, 3, 4$ which corresponded to the proposed estimator, Nadaraya-Watson estimator, transformation of data method and improved Nadaraya-Watson estimators of finite population means respectively.

A normal kernel having a mean 0 and standard deviation 1 was used since it has smooth and continuous derivatives at every data point. To maintain stability in terms of the variation of the random values simulated, an optimal bandwidth obtained using the cross-validation technique was used. Linear and non-linear functions of $m(\hat{x}_{ij})$ were used in simulation of data to compare the performance of the estimators of the population mean.

5.4 Equations of Data Functions of $m(\hat{x}_{ij})$ Simulated

Table 5.1: Equations of Data Functions Simulated

Data function	Equation
Linear	$1 + 2(x - 0.5)$
Quadratic	$1 + 2(x - 0.5)^2$
Sine	$2 + sin(2\pi x)$
Exponential	$exp(-8x)$
Bump	$1 + 2(x - 0.5) + exp\{-200(x - 0.5)^2\}$
Jump	$1 + 2(x - 0.5)I_{x \leq 0.65} + 0.65I_{x \geq 0.65}$

These data functions have been applied by various authors for data simulations, see

for instance (Kim et al., 2003) and (Lang'at, 2017). The data functions in table 5.1 that were used for simulation are widely applicable in real life and are often used in statistics for data simulations. Exponential function for instance is used for modeling successive times between two rare events such as the number of misprints on a page of a textbook while jumps and bumps are mostly used in such events as bio-surveillance for modeling disease-outbreaks or floods within a certain limit of time in a given place. Bump functions are also used in curve fitting, uncertainty analysis and approximation of non-linear relationships in scattered data. Sine functions are used to model periodic events such as light waves and average temperature variations throughout the year while quadratic functions are used in physics to describe trajectory followed by objects thrown upward at an angle whereas in economics quadratic functions can be used to develop profit and loss functions.

5.5 Simulation Results

Results of the data simulated are illustrated using tables and graphs in the following section. R codes used in the simulation are provided in the appendix.

Table 5.2: Results of Biases Simulated from a Linear Data Function

Population sizes	300	450	600	750	900	1200
Proposed Est.	-0.17243	-0.15902	-0.1522	-0.1677	-0.1494	-0.13351
N-W	-0.3615	0.4315	-0.33124	-0.40746	-0.4228	-0.3959
Transformed Est.	-0.00159	-0.00064	-0.00213	-0.00148	-0.00150	-0.0015
Improved N-W	-0.18152	-0.21607	-0.16668	-0.20447	-0.21217	-0.19871

Table 5.3: Results of Biases Simulated from Exponential Data Function

Population sizes	300	450	600	750	900	1200
Proposed Est.	-0.02437	-0.01552	-0.01909	-0.01228	-0.02158	-0.01865
N-W	0.83211	0.59492	0.72252	0.58382	0.58567	0.52649
Transformed Est.	-0.00562	-0.00439	-0.00412	-0.00367	0.00057	-0.00589
Improved N-W	0.41325	0.29527	0.35919	0.29008	0.29312	0.26030

Table 5.4: Results of Biases Simulated from Sine Data Function

Population sizes	300	450	600	750	900	1200
Proposed	-0.34306	-0.3321	-0.31596	-0.29083	-0.28649	-0.28522
N-W	-1.1712	-1.1942	-1.231	-1.17087	-1.2288	-1.2457
Transformed	-0.01914	-0.01647	-0.05211	-0.02913	-0.03182	-0.02729
Improved N-W	-0.59519	-0.60534	-0.64161	-0.5999	-0.63029	-0.6365

Tables 5.2, 5.3 and 5.4 present bias values for the data generated from different mean functions indicated. Negative values imply underestimation while positive values of the bias indicate overestimation of the finite population mean by the different estimators. The transformation of data method estimator has relatively smaller values of the bias than the rest of the estimators with the smallest being observed in table 5.3 on exponential mean function. In fact a closer look at the values of bias for the transformation of data method estimator reveals that the bias is tending to zero as the population size increases from 300 to 900. The proposed estimator also follows closely with smaller bias values as noted in the three tables. It is noted that the transformation of data method estimator was proposed to correct boundary bias due to Nadaraya-Watson estimators and this has clearly been evident in its relatively smaller values of the bias.

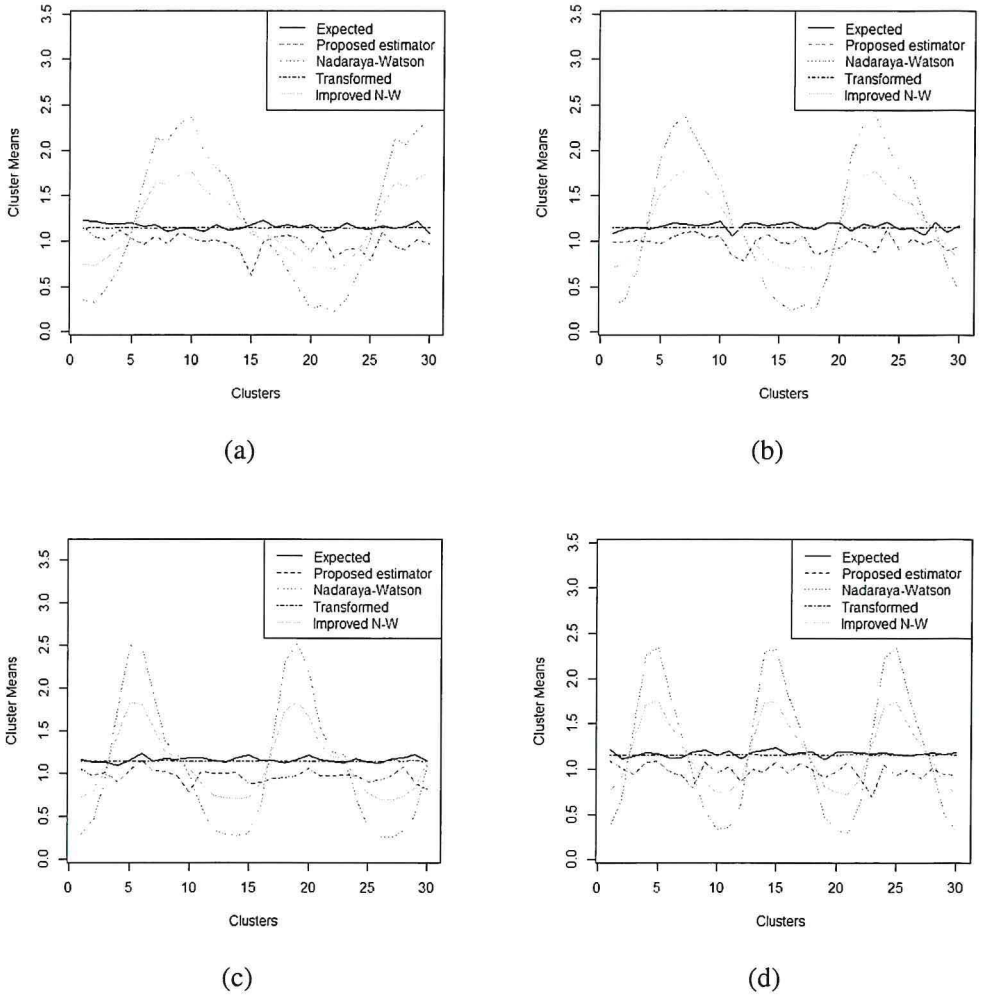

Figure 5.1: Graphs showing how the cluster means estimate the population mean for data simulated using the quadratic function for different population sizes. Figure 5.1a was obtained from a population of size 600, figure 5.1b was obtained from a population of size 750, figure 5.1c was obtained from a population of size 900 whereas figure 5.1d was obtained from a population of size 1200.

From these graphs it is clear that the transformed estimator is vary close to the expected values of the population means. The proposed estimator closely follows in the second place while the Nadaraya-Watson and improved Nadaraya-Watson estimators are relatively farther away from the expected population values of the mean, though on minimal frequencies they coincide with the expected values of the population mean as can be seen from the graphs 5.1a to 5.1d.

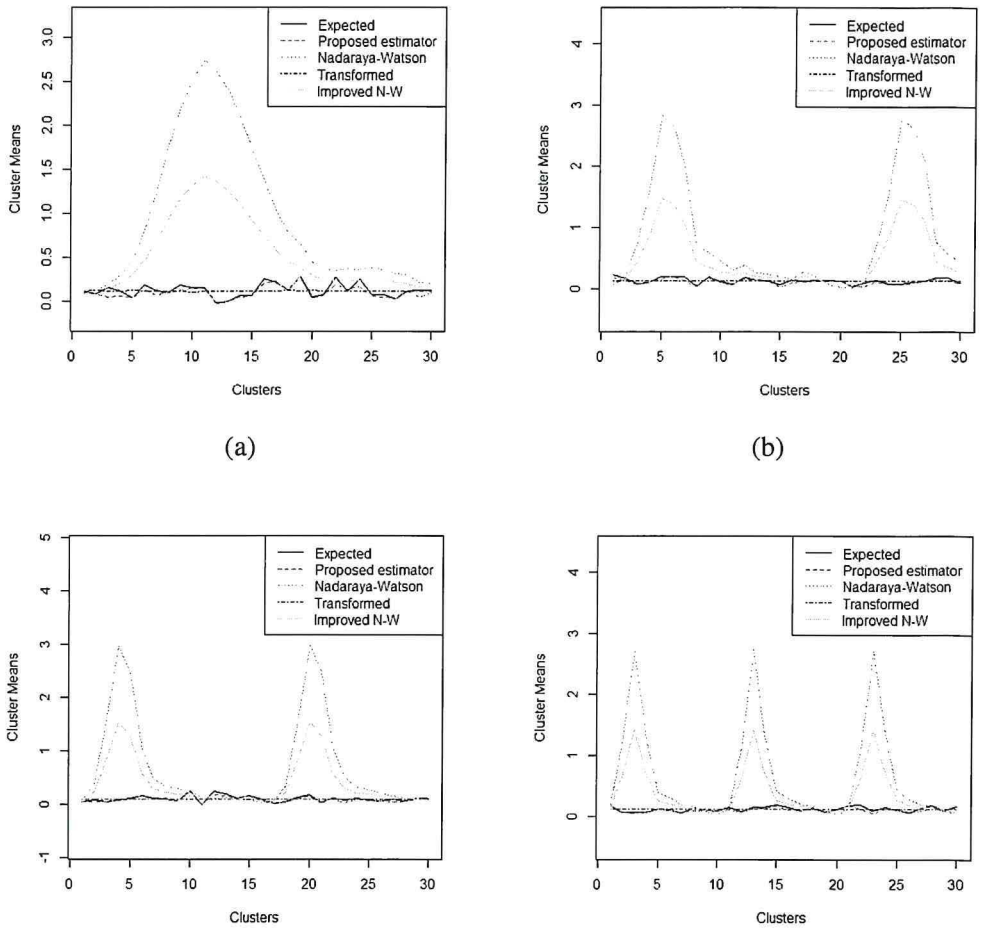

Figure 5.2: Graphs showing how the cluster means estimate the population mean for data simulated using the exponential function for different population sizes. Figure 5.2a was obtained from a population of size 450, figure 5.2b was obtained from a population of size 600, figure 5.2c was obtained from a population of size 750 whereas figure 5.2d was obtained from a population of size 900.

In all these graphs, it can be observed that the cluster means for the proposed estimator closely follows the expected means of the population values. This is more conspicuous in graph 5.2a. Secondly, it is noted that the transformation of data method estimator averages the points below and above the expected trend-line due to reflection of the transformed data at the boundaries. This estimator was proposed in chapter four to correct for boundary effects due to Nadaraya-Watson regression technique used. From graphs 5.2a to 5.2d, it can also be noted that both the Nadaraya-Watson and its improved counterpart have got their cluster means relatively farther away from the expected values of the population mean with Nadaraya-Watson having the farthest cluster means away

from the expected means.

Table 5.5: Summary Results of Bias

Data function	Proposed Estimator	N-W	Transformed Estimator	Improved N-W
Linear	-0.16774	-0.40747	-0.00147	-0.20447
Quadratic	0.02981	0.06921	-0.00958	-0.17728
Sine	-0.29083	-1.17087	-0.02913	-0.5999
Exponential	-0.01228	0.58383	-0.00368	0.29008
Bump	-0.18757	-0.51888	-0.00089	-0.25989
Jump	-0.2537	-0.30344	-0.01307	-0.15825

It can be noted in table 5.5 that the bias for the proposed estimator is smaller than those for the other respective estimators. However, the bias for the transformed estimator is even much smaller than all the other methods. This may be attributed to the reflection of the transformed data at the boundaries of the support of the kernel density function used. The transformation of data method was proposed in chapter four to address the boundary bias arising from Nadaraya-Watson technique.

A few selected graphs plotted using data generated from the different data functions stated in table 5.1 are given in the following pages. They show how the different estimators considered performed in terms of the bias.

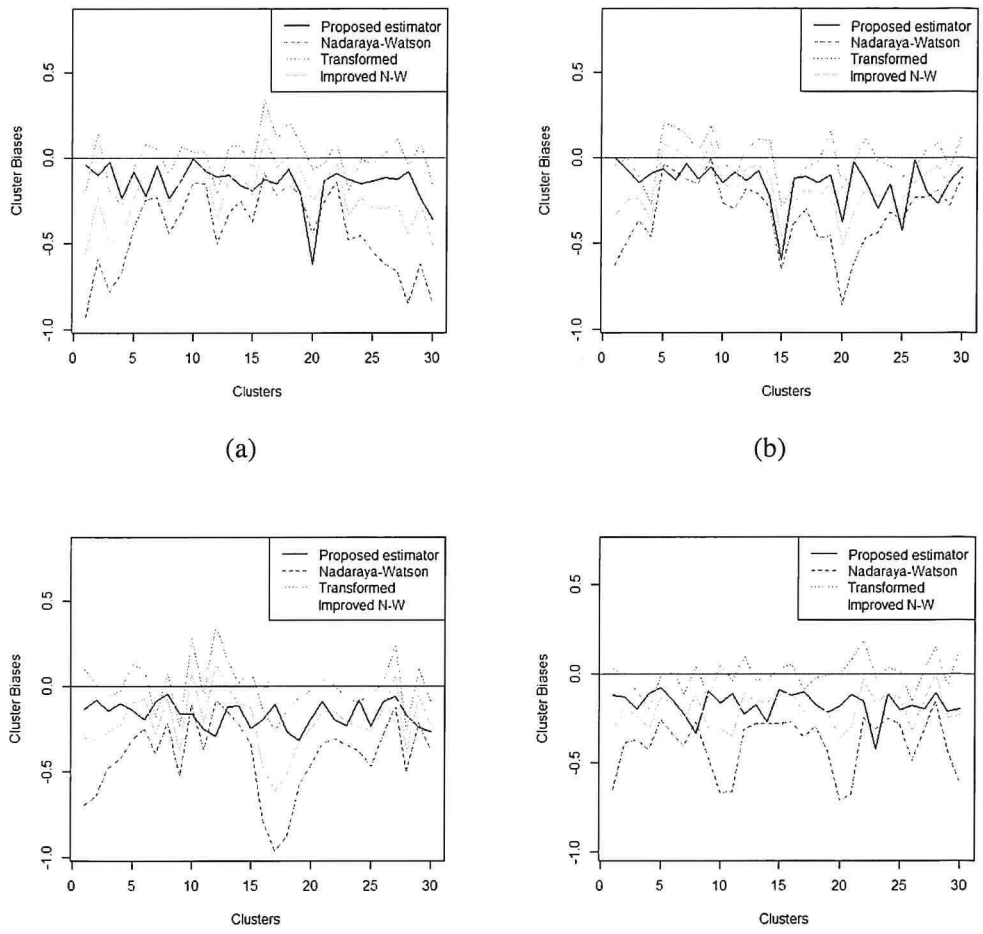

Figure 5.3: Graphs showing the biases of the estimators simulated using the linear function for different population sizes. Figure 5.3a was obtained from a population of size 450, figure 5.3b was obtained from a population of size 600, figure 5.3c was obtained from a population of size 750 whereas figure 5.3d was obtained from a population of size 1200.

Due to reflection of transformed data at the boundaries, the transformation of data method as noted from the graphs 5.3a to 5.3d crosses the line 0.0 most frequently than the rest of the estimators considered. The proposed estimator follows in the second position while the Nadaraya-Watson and imroved Nadaraya-Watson estimators are relatively more biased than the rest of the estimators.

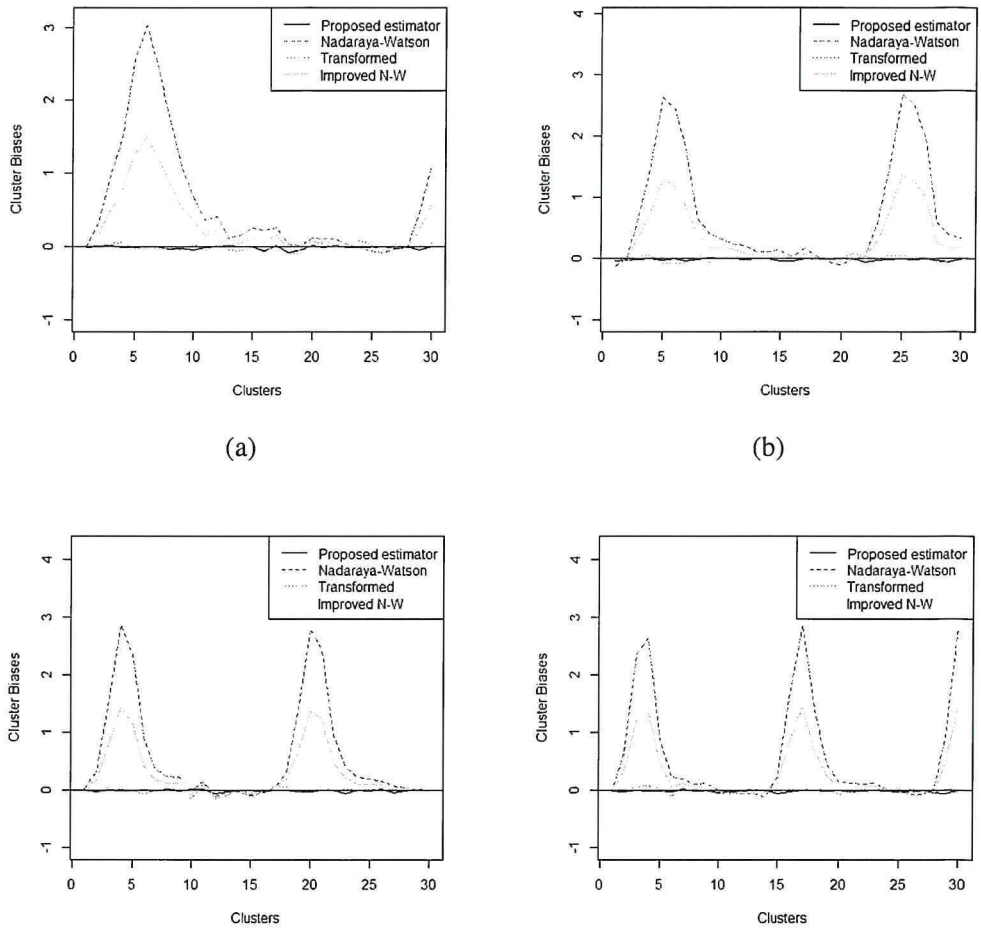

Figure 5.4: Graphs showing the biases of the estimators simulated using the exponential function for different population sizes. Figure 5.4a was obtained from a population of size 450, figure 5.4b was obtained from a population of size 600, figure 5.4c was obtained from a population of size 750 whereas figure 5.4d was obtained from a population of size 900.

Using exponential mean function, it can be seen that in all the graphs from 5.4a to 5.4d that the proposed estimator and the transformation of data method estimator are very close to zero compared to Nadaraya-Watson and improved Nadaraya-Watson estimators. In fact the transformation of data method estimator almost coincides completely with the zero-trend-line (benchmark for measuring unbiasedness) of the bias as seen in graph 5.4c.

Table 5.6: Results of MSEs Simulated from Linear Data Function

Population sizes	300	450	600	750	900	1200
Proposed Est.	0.06136	0.05592	0.04332	0.04867	0.04162	0.03843
N-W	0.15127	0.22155	0.13214	0.19015	0.2087	0.1852
Transformed Est.	0.13065	0.18618	0.10972	0.16603	0.17878	0.1567
Improved N-W	0.03802	0.05555	0.03339	0.04783	0.05246	0.04662

Table 5.7: Results of MSEs Simulated from Sine Data Function

Population sizes	300	450	600	750	900	1200
Proposed	0.17689	0.15257	0.15174	0.13688	0.13642	0.13321
N-W	1.39212	1.47663	1.55518	1.41025	1.54450	1.58529
Transformed	1.3718	1.42615	1.5157	1.37093	1.50988	1.55179
Improved N-W	0.35932	0.37915	0.42154	0.36981	0.40596	0.41353

Table 5.8: Results of MSEs Simulated from Exponential Data Function

Population sizes	300	450	600	750	900	1200
Proposed	0.00548	0.00208	0.00239	0.00314	0.00210	0.00163
N-W	1.4425	1.09392	1.3780	1.1673	1.1959	0.95237
Transformed	0.69241	0.35393	0.52204	0.34085	0.34301	0.27719
Improved N-W	0.35807	0.2722	0.3430	0.29075	0.29916	0.23655

Mean squared error combines both the variance and the squared bias terms of an estimator. The MSE values presented in tables 5.6, 5.7 and 5.8 were generated using different data functions as indicated. It can be noted that the the MSE values for the proposed estimator are relatively smaller than the rest of the estimators considered. The transformation of data method estimator follows closely in the second place with smaller MSE values compared to Nadaraya-Watson and improved Nadaraya-Watson estimators. Nadara-Watson has the largest MSE values than any other estimator considered. Comparing the MSE values for these estimators shows that the proposed estimator outperformed the rest of the estimators.

Table 5.9: Summary Results of MSEs

Data function	Proposed Estimator	N-W	Transformed Estimator	Improved N-W
Linear	0.04867	0.19015	0.16603	0.04782
Quadratic	0.04451	0.56209	0.00479	0.14017
Sine	0.13688	1.41025	1.3709	0.36981
Exponential	0.00314	1.16734	0.34085	0.29075
Bump	0.06229	0.32087	0.26924	0.08046
Jump	0.08790	0.94584	0.09208	0.23835

Efficiency of the mean estimator of the population mean was obtained using its mean squared error. This is illustrated in table 5.9. Measures for the MSE were simulated for purposes of comparison. Comparatively, the proposed estimator of the finite population mean outperforms the rest of the estimators in terms of efficiency as noted from the table. This is because the MSE of the proposed estimator is relatively smaller compared to Nadaraya-Watson and improved Nadaraya-Watson estimator except for the quadratic data function where the transformation of data method estimator stands out. Graphs of MSE for selected data functions are presented below.

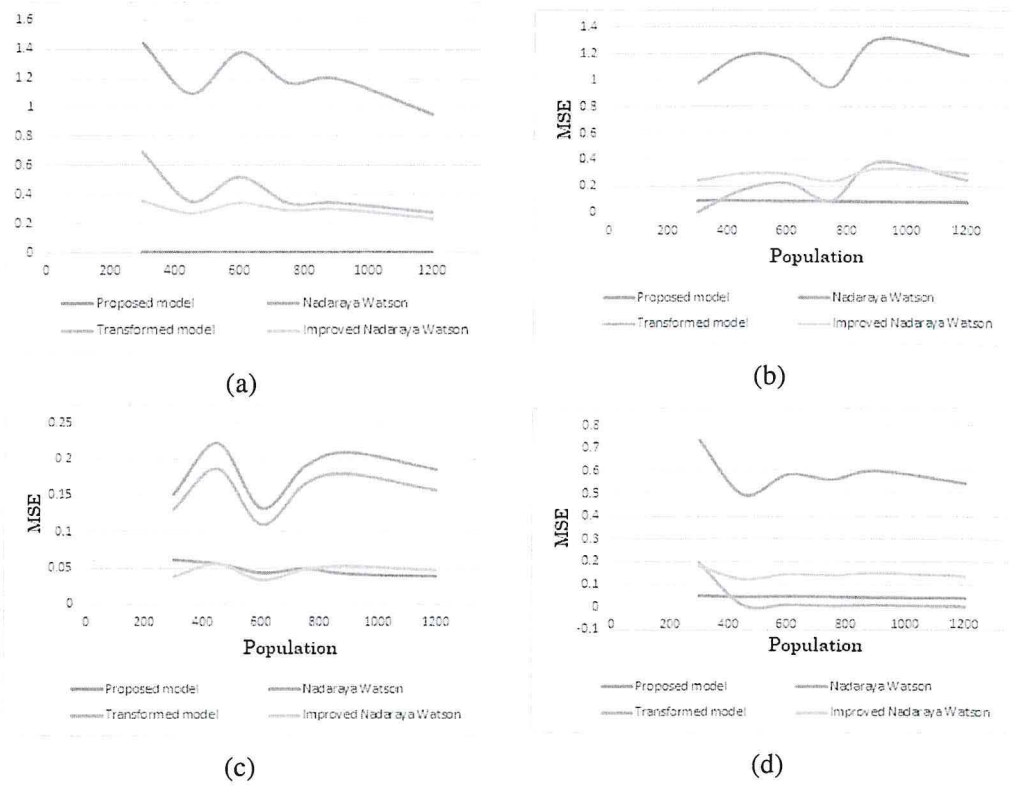

Figure 5.5: Figure 5.5a is a graph of MSE for the estimators for data simulated using an exponential data function, figure 5.5b is a graph of MSE for the estimators for data simulated using a jump data function, figure 5.5c is a graph of MSE for the estimators for data simulated using a linear data function whereas figure 5.5d is a graph of MSE for the estimators for data simulated using a quadratic data function.

The MSE for the proposed estimator tends to zero as the population size increases from 300 to 1200 as observed in graphs 5.5a to 5.5d. However, thesere seems to be lack of consistency in the transformation estimator from all these graphs of MSE for different mean functions. For instance from the exponential data function in graph 5.5a, the transformation of data method estimator displays a large MSE compared to graph 5.5d for quadratic data function where its MSE is very close to zero.

Table 5.10: Results of CI Lengths Simulated from Quadratic Data Function

Population sizes	300	450	600	750	900	1200
Proposed	0.87910	0.82502	0.82701	0.76973	0.79899	0.85104
N-W	3.36413	2.75610	2.99434	2.93895	3.03853	2.89298
Transformed	1.73943	0.34203	0.37874	0.27133	0.32993	0.12025
Improved N-W	1.67154	1.37648	1.49507	1.46763	1.52239	1.44527

Table 5.11: Results of CI Lengths Simulated from Exponential Data Function

Population sizes	300	450	600	750	900	1200
Proposed	0.29014	0.17864	0.19169	0.2195	0.17969	0.15844
N-W	4.7081	4.0999	4.6017	4.2353	4.2869	3.8255
Transformed	3.26188	2.33210	2.83228	2.28859	2.29583	2.06383
Improved N-W	2.34569	2.04514	2.29581	2.1137	2.14408	1.90655

Table 5.12: Results of CI Lengths Simulated from Bump Data Function

Population sizes	300	450	600	750	900	1200
Proposed	1.06388	0.96892	0.87839	0.9784	0.84913	0.90361
N-W	2.18623	2.31586	1.96303	2.22051	2.29845	2.13277
Transformed	1.98238	2.12953	1.83649	2.03401	2.11702	1.93430
Improved N-W	1.09189	1.15803	0.98723	1.11193	1.14911	1.06800

Confidence intervals are normally constructed around point estimators to provide a properly scaled measure of uncertainty associated with an estimator of finite population parameter of interest. Shorter confidence interval lengths means the estimator is close to the true population values being estimated. In view of this, confidence intervals were generated at 95% and from the results seen in table 5.10, 5.11 and 5.12, the proposed estimator has tighter confidence interval lengths than the other estimators considered.

Table 5.13: Summary Results of 95% Confidence Interval (CI) Lengths

Data function	Proposed Estimator	N-W	Transformed Estimator	Improved N-W
Linear	0.92695	1.70934	1.59725	0.85725
Quadratic	0.76973	2.93895	0.77713	1.46763
Sine	1.52697	4.88850	4.82599	2.54509
Exponential	0.21954	4.23531	2.28859	2.11372
Bump	0.97842	2.22051	2.03401	1.11193
Jump	1.09441	3.81237	1.18948	1.91380

The 95% upper and lower confidence intervals were generated for the estimators of finite population mean using the formula $\bar{\bar{Y}} = \hat{\bar{\bar{Y}}} \pm Z_{\frac{\alpha}{2}}\left(\sqrt{var(\hat{\bar{\bar{Y}}})}\right)$ and subsequently the confidence interval lengths were obtained. The results of these confidence interval lengths are given in table 5.13 above. A good confidence interval has a coverage rate closer to the true population mean being estimated and therefore its length has to be small. From table 5.13, it can be noted that the confidence interval lengths for the proposed estimator are shorter than the other estimators. Therefore, it can safely be concluded that the estimator developed in this research is superior to its rival estimators at 95% coverage rate.

5.6 Summary Discussion of the Results

Estimating finite population parameters is very important both in sample survey theory and practise. This study focused on the estimation of a finite population mean using two stage cluster sampling scheme in presence of non-response in the second stage of sampling.

Using inclusion probabilities, a sample of clusters were chosen from the population from which some elements in the selected clusters failed to respond. Non-response indicator variable was then used to obtain response elements and therefore estimate the non-response elements in the selected sample. Kernel weights were then used to compensate for non-response in the estimation process. Nadaraya-Watson kernel estimator was used and found to be affected by the boundary effects since the kernel density function uses reduced amount of data at the boundaries because it does not detect data outside the boundaries of its support.

In this research, one of the aims was to first develop an estimator for finite population mean under non-response in two stage cluster sampling. The estimator was therefore developed and besides, its bias and mean squared error were derived asymptotically. This was done in chapter three.

To reduce the bias due to Nadaraya-Watson estimator, various methods were reviewed in chapter two and consequently a method was proposed for this study. Modified transformation of data method was proposed and its asymptotic properties were obtained. In chapter five, a simulation study was done to check the performance of the proposed estimators with the existing ones. MSE, bias and confidence interval lengths were obtained for all the estimators considered using various data functions stated in table 5.1.

The bias results in table 5.5 revealed that the proposed estimator has smaller bias than its rival estimators except for the transformed estimator which does well due to transformation and reflection of data at the boundaries; this conforms to the explanation in chapter two on boundary bias reduction. From table 5.9, it is notable that the proposed estimator has smaller values of MSE compared to other estimators. The graphs in appendix also show that the proposed estimator performed well in terms of the MSE compared to the others.

The confidence interval lengths were generated at 95% level and the results tabulated in table 5.9 indicate shorter confidence interval lengths for the proposed estimator than those of other estimators.

From the simulation results discussed, it can be concluded that the proposed estimator performs better in terms of MSE than that of Nadaraya-Watson, (Nadaraya, 1964) and Watson (1964) and (Demir and Toktamiş, 2010). Hence the study improves on the existing non-parametric techniques in estimation of population parameters.

This research however is limited by the choice of the bandwidth. From the literature discussed in chapter two on bandwidth selection, no universal method exists for choosing the bandwidth, though this research adopted the cross-validation technique. Besides, a normal kernel with mean 0 and standard deviation 1 was used in the simulation experiment. Other kernels such as Epanechnikov could also be used in the simulation process.

5.7 Conclusion

The main aim of this study was to estimate a finite population mean in presence of random non-response using two-stage cluster sampling with replacement. Simulation experiment conducted revealed that the proposed estimator is more efficient than Nadaraya-Watson, improved Nadaraya-Watson and transformation of data method estimators. The proposed estimator has smaller bias, lower mean squared error values and tighter confidence interval lengths compared to its rival estimators.

The transformation of data method estimator proposed in chapter four has also proved to be efficient in correcting boundary bias associated with existing kernel-based regression estimators, in particular the Nadaraya-Watson estimator. This can be seen in table 5.2 from the lower values of the bias of the transformation of data method estimator compared to those of other estimators.

5.8 Suggestions for Further Research

This study considered boundary correction method for the density. It would be interesting to try complicated estimators that automatically caters for boundary bias. Besides, this research adopted an optimal bandwidth obtained from cross-validation technique of bandwidth selection. As outlined in chapter two on smoothing parameter selection criteria, no one criteria does well in all types of densities with regards to bias-variance trade-off. Though bandwidth selection was not the main focus of this study, a research to develop the most optimal way of selecting bandwidths could be done so that better estimators of population parameters may be obtained.

References

Abramson, I. S. (1982). On bandwidth variation in kernel estimates-a square root law. *The annals of Statistics*, pages 1217–1223.

Ahmed, M., Al-Titi, O., Al-Rawi, Z., and Abu-Dayyeh, W. (2006). Estimation of a population mean using different imputation methods. *Statistics in Transition 7, 6: 1247*, 1264.

Anderson, H. (1979). On nonresponse bias and response probabilities. *Scandinavian Journal of Statistics*, pages 107–112.

Andridge, R. R. and Little, R. J. (2010). A review of hot deck imputation for survey non-response. *International statistical review*, 78(1):40–64.

Bach, E., Shallit, J. O., Shallit, J., and Jeffrey, S. (1996). *Algorithmic number theory: Efficient algorithms*, volume 1. MIT press.

Bahl, S. and Tuteja, R. (1991). Ratio and product type exponential estimators. *Journal of information and optimization sciences*, 12(1):159–164.

Baszczyńska, A. (2015). Bias reduction in kernel estimator of density function in boundary region. *Metody Ilościowe w Badaniach Ekonomicznych*, 16(1):7–16.

Bethlehem, J. G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, 4(3):251–260.

Bii, N. K., Onyango, C. O., and Odhiambo, J. (2018). Estimating a finite population mean under random non response in two stage cluster sampling with replacement. In *Journal of Physics: Conference Series*, volume 1132, page 012071. IOP Publishing.

Bound, J., Brown, C., and Mathiowetz, N. (2001). Measurement error in survey data. In *Handbook of econometrics*, volume 5, pages 3705–3843. Elsevier.

Cowling, A. and Hall, P. (1996). On pseudodata methods for removing boundary effects in kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 551–563.

Demir, S. and Toktamiş, Ö. (2010). On the adaptive nadaraya-watson kernel regression estimators. *Hacettepe Journal of Mathematics and Statistics*, 39(3).

Dorfman, A. H. (1992). Nonparametric regression for estimating totals in finite populations. In *Proceedings of the Section on Survey Research Methods*, pages 622–625. American Statistical Association Alexandria, VA.

Durrant, G. B. et al. (2005). Imputation methods for handling item-nonresponse in the social sciences: a methodological review. *ESRC National Centre for Research Methods and Southampton Statistical Sciences Research Institute. NCRM Methods Review Papers NCRM/002*.

El-Badry, M. (1956). A sampling procedure for mailed questionnaires. *Journal of the American Statistical Association*, 51(274):209–227.

Fife-Schaw, C. (2000). Surveys and sampling issues. *Research methods in psychology*, 2:88–104.

Ford, B. L. (1976). *Missing data procedures: a comparative study*. Sampling Studies Section, Sample Surveys Research Branch, Statistical Reporting Service, US Department of Agriculture.

Fuller, W. A., Loughin, M. M., and Baker, H. D. (1994). Regression weighting for the 1987-88 national food consumption survey. *Survey Methodology*, 20:75–85.

Gasser, T. and Müller, H.-G. (1979). Kernel estimation of regression functions. In *Smoothing techniques for curve estimation*, pages 23–68. Springer.

Goh, C. et al. (2004). Bandwidth selection for semiparametric estimators using the m-out-of-n bootstrap. Technical report, Working Paper.

Hall, P. and Marron, J. S. (1987). Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation. *Probability Theory and Related Fields*, 74(4):567–581.

Hansen, M. H. and Hurwitz, W. N. (1946). The problem of non-response in sample surveys. *Journal of the American Statistical Association*, 41(236):517–529.

Härdle, W. (1990). *Applied nonparametric regression*. Number 19. Cambridge university press.

Hardle, W. and Marron, J. S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *The Annals of Statistics*, pages 1465–1481.

Holt, D. and Elliot, D. (1991). Methods of weighting for unit non-response. *The Statistician*, pages 333–342.

Ismail, M., Shahbaz, M. Q., and Hanif, M. (2011). Pak. j. statist. 2011 vol. 27 (4), 467-476 a general class of estimator of population mean in presence of non–response. *Pak. J. Statist*, 27(4):467–476.

Ivanka, H., Jan, K., and Jiri, Z. (2012). *Kernel Smoothing in MATLAB: theory and practice of kernel smoothing*. World scientific.

Jeffrey, A. and Dai, H. H. (2008). *Handbook of mathematical formulas and integrals*. Elsevier.

Jones, M., Marron, J. S., and Sheather, S. J. (1992). *Progress in data-based bandwidth selection for kernel density estimation*. Department of Statistics [University of North Carolina at Chapel Hill].

Jones, M. C. (1993). Simple boundary correction for kernel density estimation. *Statistics and Computing*, 3(3):135–146.

Jones, M. C., Marron, J. S., and Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, 91(433):401–407.

Kalsbeek, W., McLaurin, R., Miller, J., et al. (1980). The national head and spinal cord injury survey: major findings. *Journal of Neurosurgery*, pages S19–31.

Kalsbeek, W. D. (1980). A conceptual review of survey error due to nonresponse. In *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pages 131–136.

Kalton, G. and Kasprzyk, D. (1982). Imputing for missing survey responses. In *Proceedings of the section on survey research methods, American Statistical Association*, volume 22, page 31. American Statistical Association Cincinnati.

Karunamuni, R. J. and Alberts, T. (2005). On boundary correction in kernel density estimation. *Statistical Methodology*, 2(3):191–212.

Khan, M., Shabbir, J., Hussain, Z., and Al-Zahrani, B. (2014). A class of estimators for finite population mean in double sampling under nonresponse using fractional raw moments. *Journal of Applied Mathematics*, 2014.

Khare, B. (2013). Separate generalized estimators for population mean in the presence of non-response in stratified random sampling. *Advances in Statistics and Optimization*, page 150.

Kim, J. K. and Kim, J. J. (2007). Nonresponse weighting adjustment using estimated response probability. *Canadian Journal of Statistics*, 35(4):501–514.

Kim, J.-Y., Breidt, F. J., and Opsomer, J. D. (2003). Nonparametric regression estimation of finite population totals under two-stage sampling. *preprint*.

Kim, W., Park, B., and Marron, J. S. (1994). Asymptotically best bandwidth selectors in kernel density estimation. *Statistics & Probability Letters*, 19(2):119–127.

Lang'at, R. C. (2017). *Robust Estimation of Finite Population Total Incorporating Data-Reflection Technique in Nonparametric Regression*. PhD thesis, COPAS, JKUAT.

Läuter, H. (1988). Silverman, bw: Density estimation for statistics and data analysis. chapman & hall, london–new york 1986, 175 pp.,£ 12.—. *Biometrical Journal*, 30(7):876–877.

Liang, K.-Y. and Zeger, S. L. (1993). Regression analysis for correlated data. *Annual review of public health*, 14(1):43–68.

Loader, C. R. et al. (1999). Bandwidth selection: classical or plug-in? *The Annals of Statistics*, 27(2):415–438.

Madden, L., Hughes, G., and Munkvold, G. (1996). Plant disease incidence: Inverse sampling, sequential sampling, and confidence intervals when observed mean incidence is zero. *Crop Protection*, 15(7):621–632.

Malec, P. and Schienle, M. (2014). Nonparametric kernel density estimation near the boundary. *Computational Statistics & Data Analysis*, 72:57–76.

Marron, J. S. and Ruppert, D. (1994). Transformations to reduce boundary bias in kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 653–671.

Müller, H.-G. (1991). Smooth optimum kernel estimators near endpoints. *Biometrika*, 78(3):521–530.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142.

Odhiambo, O. D. and Onyango, C. O. (2014). Estimation of population total in the presence of missing values using a modified murthy's estimator and the weight adjustment technique. *American Journal of Applied Mathematics and Statistics*, 2(3):163–167.

Onyango, C. O., Otieno, R. O., and Orwa, G. O. (2010). Generalised model based confidence intervals in two stage cluster sampling. *Pakistan Journal of Statistics and Operation Research*, 6(2).

Ouma, C. and Wafula, C. (2005). Bootstrap confidence intervals for model-based surveys. *East African Journal of Statistics*, 1(1):84–90.

Pike, G. R. (2008). Using weighting adjustments to compensate for survey nonresponse. *Research in Higher Education*, 49(2):153–171.

Rancourt, E., Lee, H., and Särndal, C. (1994). Bias corrections for survey estimates from data with ratio imputed values for confounded non-response. *Survey Methodology*, 20:137–147.

Rao, J. (1996). On variance estimation with imputed survey data. *Journal of the American Statistical Association*, 91(434):499–506.

Ruppert, D., Sheather, S. J., and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, 90(432):1257–1270.

Sain, S. R., Baggerly, K. A., and Scott, D. W. (1994). Cross-validation of multivariate densities. *Journal of the American Statistical Association*, 89(427):807–817.

Särndal, C. E. (1980). On π-inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, 67(3):639–650.

Särndal, C.-E. and Lundström, S. (2005). *Estimation in surveys with nonresponse*. John Wiley & Sons.

Serfling, R. J. (2009). *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons.

Sheather, S. J. (2004). Density estimation. *Statistical science*, pages 588–597.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(3):683–690.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, volume 26. CRC press.

Simonoff, J. S. (2012). *Smoothing methods in statistics*. Springer Science & Business Media.

Singh, H. P. and Kumar, S. (2008). A general family of estimators of finite population ratio, product and mean using two phase sampling scheme in the presence of non-response. *Journal of Statistical Theory and Practice*, 2(4):677–692.

Singh, H. P. and Kumar, S. (2009). A general procedure of estimating the population mean in the presence of non-response under double sampling using auxiliary information. *SORT-Statistics and Operations Research Transactions*, 33(1):71–84.

Singh, H. P. and Kumar, S. (2010). Estimation of mean in presence of non-response using two phase sampling scheme. *Statistical papers*, 51(3):559–582.

Singh, H. P., Kumar, S., et al. (2011). Combination of regression and ratio estimate in presence of nonresponse. *Brazilian journal of Probability and Statistics*, 25(2):205–217.

Singh, H. P., Kumar, S., and Kozak, M. (2010). Improved estimation of finite-population mean using sub-sampling to deal with non response in two-phase sampling scheme. *Communications in Statistics—Theory and Methods*, 39(5):791–802.

Singh, R. (1985). Estimation from incomplete data in longitudinal surveys. *Journal of Statistical Planning and Inference*, 11(2):163–170.

Singh, S. and Horn, S. (2000). Compromised imputation in survey sampling. *Metrika*, 51(3):267–276.

Takezawa, K. (2005). *Introduction to nonparametric regression*, volume 606. John Wiley & Sons.

Terrell, G. R. and Scott, D. W. (1992). Variable kernel density estimation. *The Annals of Statistics*, pages 1236–1265.

Van Buuren, S., Boshuizen, H. C., and Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in medicine*, 18(6):681–694.

Wand, M. P. and Jones, M. C. (1994a). *Kernel smoothing*. Crc Press.

Wand, M. P. and Jones, M. C. (1994b). Multivariate plug-in bandwidth selection. *Computational Statistics*, 9(2):97–116.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372.

Zhang, S. and Karunamuni, R. J. (2000). Boundary bias correction for nonparametric deconvolution. *Annals of the Institute of Statistical Mathematics*, 52(4):612–629.

Appendix: R Codes

```
#n=2000
N=40 #size of the cluster
k=30 #number of clusters
Pop=N*k
Pop
set.seed(500)
#X1=rnorm(Pop,12.5,5)
E<-rnorm(Pop,mean=0,sd=.1) #the random error
X<-runif(Pop, min = 0, max = 1) #the explanatory variable
#various types of data functions
mx=1-X+exp(-200*(X-0.5)^2)
mx.1=2+sin(2*pi*X)#Sine
Ix <- function(x,h) 1*(x >= h)
mx.2 = 1 + 2*(X-.5)*Ix(X,.65) +0.65*(1-Ix(X,.65))
mx.3=2+sin(2*pi*X)
mx.4=1+2*(X-0.5)
mx.5=1+2*(X-0.5)+exp(-200*(X-0.5)^2)
mx.6=exp(-8*X)
mx.7=1+2*((X-0.5)^2)
Y <- mx.5+E
DataR=data.frame(Y)
DR11=array(as.matrix.data.frame(DataR),dim=c(N,1,k))
#DR11 #Total population means
mean(DR11)
mean(Y)
mean(DR11[,,1])
mean(DR11[,,2])
mean(DR11[,,3])
m11=mean(DR11[,,1])
for  (i in 2:k)
```

```
m11[i]=mean(DR11[,,i])
#m11
mean(m11)#cluster mean of means
#plot(m11,type="l")
v11=var(DR11[,,1])
for  (i in 2:k)
v11[i]=var(DR11[,,i])
#v11
mean(m11)#cluster mean of means
#plot(m11,type="l")
#delta=sample(c(0,1), replace=TRUE, size=n)
#Simulating the cluster indicator variables (delta)
set.seed(1)
delt=runif(Pop,0.3,1.5)
#ceiling(delta)
#floor(delta)
delta=round(delt)
#delta
sum(delta)
#simulating inclusion probabilities
phi=runif(Pop,0,.5)
#round(phi)
#sample size of the response
#n1=sum(delta)
#n1
#sample size of non-response
#n2=n-n1
#n2
#DR11*delta
#Y*delta
length(Y)
```

```
A=Y*delta
B=A/phi
#C=sum(B)/n1
#C
D=1-(1/phi)
E=D*A
#F=sum(E)/n2
G=B+E
sum(G)
sum(G)/n
mean(Y)
S=sum(Y)
Ss=sum((Y*phi))
Pop=N*k
Pop
Samplesize=round((Ss/S)*Pop)
Samplesize
#Deltas within the clusters
DeltaR=data.frame(delta)
Delta11=array(as.matrix.data.frame(DeltaR),dim=c(N,1,k))
#Delta11 #Total population means
#inclusion probabilities within the clusters
phiR=data.frame(phi)
phi11=array(as.matrix.data.frame(phiR),dim=c(N,1,k))
#phi11 #Total population means
A1=DR11*Delta11
B2=A1/phi11
C1=1-(1/phi11)
C2=C1*A1
D1=B2+C2
#means of response
meanB11=mean(B2[,,1])
```

```
for  (i in 2:k)
meanB11[i]=mean(B2[,,i])
#meanB11
#means of non-response
meanC21=mean(C2[,,1])
for  (i in 2:k)
meanC21[i]=mean(C2[,,i])
#meanC21
E2=meanB11+meanC21 #estimated means for the clusters
#estimated mean
Ybar2=mean(E2)
Ybar2
mean(D1)
mean(DR11)#population mean
var(E2) #variance of the estimated means
var(DR11)
#Nadaraya watson
library(MASS)
library(sm)
library(KernSmooth)
b=hcv(Y)
b
fitpol <- locpoly(Y,degree=0, kernel = "normal",
bandwidth = b)#weight
weight=fitpol$y
#fitpol$y
#inclusion probabilities within the clusters
#substituted by the N-W weights
weightR=data.frame(weight)
weights=array(as.matrix.data.frame(weightR),dim=c(N,1,k))
#weights #Total population means
A11=DR11
```

```
B22=A11*phi11 #sample with no non-response
B22a=B22*Delta11 #sample with non-response
C11=1-(1/phi11)
C22=weights*DR11
C222=C22*Delta11
D11=B22+C222
E22=B22a+weights #clusters updated by the N-W
#cluster means of the N-W
meanE22=mean(E22[,,1])
for  (i in 2:k)
meanE22[i]=mean(E22[,,i])
meanE22
#generating the trasformed data from the reflection
mf=mf2=TT=0
ynosample means=0
mf=mf2=TT=0
j=0
MEAN=0
MTTR=0
ynosamplemeans=0
while(j<=Pop)
{
sindex=sample(1:Pop,Samplesize)
x.sample=X[sindex]
xreflect=c(x.sample,-x.sample)
#xreflect
xnosample=setdiff(X,x.sample)
y.sample=Y[sindex]
yreflect=c(y.sample,-y.sample)
yreflect
ynosample=setdiff(Y,y.sample)
data1=data.frame(xnosample)
```

```
H3=(ucv(xreflect,nb=Pop,min(x.sample),max(x.sample)))
H1=(ucv(x.sample,nb=Pop,min(x.sample),max(x.sample)))
nad1=ksmooth(x.sample,y.sample,kernel="normal",
bandwidth =b,x.points=xnosample)
nad2=ksmooth(xreflect,yreflect,kernel="normal",
bandwidth =b,x.points=xnosample)
nad3=loess(y.sample~x.sample,span=.5)
Vr=0
for(i in 1:length(y.sample))
{
f=length(y.sample)/length(Y)
x_bar=mean(x.sample)
r=mean(y.sample)/mean(x.sample)
Vr[i]=(y.sample[i]-r*x.sample[i])^2
}
VAr=((1-f)/(length(y.sample)*(length(y.sample)-1))
*x_bar^2)*sum(Vr)
summary(nad3)
mf=nad1$y
mf2=nad2$y
mf1=predict(nad3,data=data1,se=TRUE)
MTTR[j]=mean(c(y.sample,as.vector(mf2)))
#mean of the predicted Y
#MEAN[j]=mean(x.sample)
j=j+1
}
MTTR
#clusters for the transformed data
MTTRR=data.frame(MTTR)
MTTRR1=array(as.matrix.data.frame(MTTRR),
dim=c(N,1,k))
MTTRR1
```

```
#cluster means of the transformed data
meanMTTR=mean(MTTRR1[,,1])
for  (i in 2:k)
meanMTTR[i]=mean(MTTRR1[,,i])
meanMTTR
MeanTR=mean(meanMTTR)
#mean of the transformed model
VarTR=var(meanMTTR)
#variance of the transformed model
#Estimated mean
Ybar22=mean(meanE22)
Ybar22
#mean(D11)
mean(DR11)#population mean
var(meanE22)#variance of the estimated
means under N-W
var(DR11)#Population variance
#Estimation of the bias
#1. Proposed model
Bia=E2-m11
Bias1=mean(Bia)
Bias1#Bias of the proposed model
MSE=var(E2)+Bias1^2
MSE#MSE of the proposed model
H1=(sum(mean(E2)/k))-mean(DR11)
RBias1=H1/mean(DR11)
RBias1
#2. N-W
Bia1=meanE22-m11
Bias2=mean(Bia1)
Bias2#Bias of the N-W
MSE1=var(meanE22)+Bias2^2
```

```
MSE1#MSE of the proposed model
#relative bias
H=(sum(meanE22)/k)-mean(DR11)
RBias2=H/mean(DR11)
RBias2
#3. The transformed model
Bia2=meanMTTR-m11
Bias3=mean(Bia2)#Bias of the transformed
MSE2=var(meanMTTR)+Bias2^2
H3=(sum(mean(meanMTTR))/k)-mean(DR11)
RBias3=H3/mean(DR11)
RBias3
##############################
#Output
##############################
#From the population
mean(DR11)#population mean
var(DR11)#Population variance
b #used bandwidth
##################################
##################################
#proposed model
##################################
Ybar2 #mean estimate
var(E2) #variance of the estimated means
Bias1#Bias of the proposed model
MSE#MSE of the proposed model
##################################
#Estimation under the N-W
##################################
Ybar22 #mean estimate N-W
var(meanE22)#variance of the estimated means under N-W
```

```
Bias2#Bias of the N-W
MSE1#MSE of the N-W
##############################################
###################################
#Estimation under the transformed model
###################################
MeanTR #mean of the transformed model
VarTR #variance of the transformed model
Bias3#Bias of the transformed model
MSE2#MSE of the transformed model
##############################################
#confindence intervals
#proposed model
Ybar2ConL=Ybar2-1.96*sqrt(MSE)
Ybar2ConU=Ybar2+1.96*sqrt(MSE)
Ybar2ConL
Ybar2ConU
##########################################
#N-W model
Ybar22ConL=Ybar22-1.96*sqrt(MSE1)
Ybar22ConU=Ybar22+1.96*sqrt(MSE1)
Ybar22ConL
Ybar22ConU
##########################################
#transformed model
Ybar33ConL=MeanTR-1.96*sqrt(MSE2)
Ybar33ConU=MeanTR+1.96*sqrt(MSE2)
Ybar33ConL
Ybar33ConU
##########################################
#improved N-W computations
TRNW=meanE22+meanMTTR
```

```
meanTRNW=TRNW/2
mean(meanTRNW)
mean(Y)
TRNWB=Bia1+Bia2
varTRNWB=TRNWB/2
Bias4=mean(varTRNWB)
Bias4#Bias of the improved N-W
MSE4=var(meanTRNW)+Bias4^2
##################################
#Estimation under the Improved N-W
##################################
mean(meanTRNW) #mean estimate Improved N-W
var(meanTRNW)
#variance of the estimated means under N-W
Bias4#Bias of the improved N-W
MSE4#MSE of the improved N-W
#confindence intervals of improved N-W
YbarTNWConL=mean(meanTRNW)-1.96*sqrt(MSE4)
YbarTNWConU=mean(meanTRNW)+1.96*sqrt(MSE4)
YbarTNWConL
YbarTNWConU
##############################################
#plotting the cluster means
#library(ggplot2)
Clusters=c(1:k)
plot(Clusters,m11, ylim=c(0.2,2), type="l",
lty=1,col=1,ylab="Cluster Means",cex=2)
#cluster means for the population
lines(Clusters,E2, type="l", lty=2,col=2,cex=2)
#cluster means for the proposed model
lines(Clusters,meanE22,type="l",lty=3,col=3,cex=2)
lines(Clusters,meanMTTR,type="l",lty=4,col=4,cex=2)
```

```
lines(Clusters,meanTRNW,type="l",lty=5,col=5,cex=2)
#abline(h=mean(Y), col="yellow")
legend("topright", c("Expected", "Proposed estimator ",
"Nadaraya-Watson", "Transformed","Improved N-W"),
col = c(1,2,3,4,5),lty = c(1,2,3,4,5))
#plotting the cluster Biases
#library(ggplot2)
Clusters=c(1:k)
plot(Clusters,Bia, ylim=c(-0.9,1), type="l",
lty=1,col=1,ylab="Cluster Biases",cex=2)
lines(Clusters,Bia1, type="l", lty=2,col=2,cex=2)
#cluster means for the proposed model
lines(Clusters,Bia2,type="l",lty=3,col=3,cex=2)
lines(Clusters,varTRNWB,type="l",lty=5,col=5,cex=2)
abline(h=0, col="blue")
legend("topright", c("Proposed estimator",
"Nadaraya-Watson", "Transformed","Improved N-W"),
col = c(1,2,3,5),lty = c(1,2,3,5))
#lines(Clusters,m11impu,type="l",lty=5)
```