NONPARAMETRIC ESTIMATION OF FINITE POPULATION TOTAL

WINNIE MOKEIRA ONSONGO

(MS300-0008/12)

A Thesis submitted to Pan African University Institute for Basic Sciences Technology and Innovation in partial fulfillment of the requirement for the Degree of Master of Science in Mathematics (Statistics Option)

2014

DECLARATION

This thesis is my original work and has not been submitted to any other university for examination.

Signature:.................................. **Date**:...........................

Onsongo Winnie Mokeira

This thesis report has been submitted for examination with my approval as the University supervisor.

Signature:.. **Date**:.............................

Professor Romanus Odhiambo

Jomo Kenyatta University of Agriculture and Technology

This thesis report has been submitted for examination with my approval as the University supervisor.

Signature:.. **Date**:.................................

Dr. George Orwa

Jomo Kenyatta University of Agriculture and Technology

DEDICATION

I wish to dedicate this work to my loving parents, Josiah Nyabuto and Teresa Kiage, and siblings Welma, Welline, Weldaline, Welaine, Weavine and Wilson for their love and moral support. This work is also dedicated to my late grandfather Stanio Onsongo. You all mean the world to me.

ACKNOWLEDGEMENT

I wish to acknowledge the thesis supervisors Prof. Romanus Odhiambo and Dr. George Orwa who have greatly contributed to the completion of this thesis. I wish to thank them for sparing their precious time to consistently offer unlimited encouragement in the field of sampling and regression modeling. Their experience in the field of sample surveys has greatly contributed in the solving of this particular research problem.

I wish to acknowledge Pan African University, Institute for Basic Sciences, Technology and Innovation and Jomo Kenyatta University of Agriculture and Technology, the host university, for granting me the opportunity to conduct the study.

The whole research was fully funded through a scholarship awarded by the African Union Commission. I wish to acknowledge the financial assistance offered from the union and facilitations through the host university.

TABLE OF CONTENTS

Contents **Page**

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

$\boldsymbol{p}$	Entire population
$\boldsymbol{s}$	Sample size
$\boldsymbol{p-s}$	Non-sample size
$\boldsymbol{i}$	Sample element
$\boldsymbol{j}$	Non-sample element
MBC	Multiplicative Bias Corrected
MSE	Mean Squared Error
RMSE	Root Mean Squared Error

ABSTRACT

In this study, the problem of nonparametric estimation of finite population total using multiplicative bias correction technique is considered. A review of the model-based, design-based, model-assisted, randomization-assisted and nonparametric approaches to finite population total estimation is explored. A robust estimator of the finite population total based on multiplicative bias correction is derived. The properties of the estimator are developed and comparative study with the existing model based and design based estimators is carried to assess the performance of the estimator developed using the simulated sets of data. It is observed that the estimator is asymptotically unbiased and statistically consistent when certain conditions are satisfied. It has been shown that when the model based estimators are used in estimating the finite population total, there exists bias-variance trade-off along the boundary. The multiplicative bias corrected estimator has recorded better results in estimating the finite population total by correcting the boundary problems associated with existing model based estimators. The theoretical and empirical results led to the suggestion that the multiplicative bias corrected estimator can be highly recommended in survey sampling estimation of the finite population total.

CHAPTER ONE

INTRODUCTION

1.0 Background

In most scenarios, auxiliary information is available for all elements in the population under consideration. Auxiliary information aids in the prediction of finite population parameters and as such it forms a central part of sample surveys. This type of information is very important and is always required frequently as it acts as a basis for good planning in various sectors of the economy.

It therefore follows that a model–based approach is used to increase the precision of the estimators by incorporating auxiliary variables. As an approach to such a problem, a superpopulation model is used to describe the relationship between the auxiliary variable and the study variable.

Previous work was mostly concerned with the behaviour of these estimators under model misspecification. With this concern of robustness, it is proper to consider a nonparametric class of models since they allow the models to be correctly specified for a large class of functions.

Nonparametric regression is motivated by the fact that it provides a flexible way of studying the relationships between variables and also results in good estimators thus increasing their efficiency compared to estimators obtained using design-based approaches.

In this framework, the main concern is the prediction of population totals using a multiplicative bias correction approach to nonparametric regression. A nonparametric estimator of population totals is therefore proposed with the aid of a superpopulation model. A methodology to study the properties of the proposed estimator is also offered.

1.1 Statement of the Problem

Given a finite population P of N identifiable units, let Y denote the survey variable with population values Y_i where $i = 1,2, \dots, N$. Also let X denote the auxiliary variable with corresponding population values X_i for $i = 1,2, \dots, N$. Assume that the auxiliary values X_i are all known but the survey values Y_i are only known for a sample S of size n where $n \leq N$. Survey sampling involves the estimation of population parameters with highest precision. Existing methods of estimating the population parameters exhibit shortcomings such as bias-variance trade-off along the boundary. This study therefore focuses on the estimation of finite population total using multiplicative bias correction approach to nonparametric regression as a technique of correcting the boundary problems..

1.2 Objectives

i. General Objective

Let $Y_1, Y_2, \dots, Y_N$ be population measurements representing some survey characteristics. The goal is to estimate the population total defined as

$$T = \sum_{i=1}^{N} Y_i \qquad (1.11)$$

via the simple random sampling without replacement scheme. The survey values of interest Y_i are realized values of an assumed working model $y_i = m(x_i) + \varepsilon_i$ that incorporates auxiliary information where $m(x_i)$ is a smooth function, ε_i is the error term with $E(\varepsilon_i) = 0$ and $E({\varepsilon_i}^2) < \infty$

ii. Specific Objectives

1) To determine a nonparametric estimator of the finite population total using a multiplicative bias correction procedure.
2) To study the asymptotic properties of the proposed estimator (that is asymptotic unbiasedness, variance and consistency).
3) To compare the performance of the proposed estimator to that of Nadaraya Watson and ratio estimators using simulated data.

1.3 Rationale

Sample survey is an important field of study in statistics. It is through this field that researchers are able to estimate population parameters using samples drawn from populations of interest.

This study focuses on the estimation of finite population total using Multiplicative Bias Correction (MBC) approach. Unlike previous design-based and nonparametric methods of estimation, the proposed MBC estimator performed better by solving the boundary problems. The MBC estimator is therefore recommended for use in the precise estimation of various population parameters such as population total, mean value of population total and variation of the population total. The results obtained can be of great importance and can be used for purposes of policy implementation and planning in various sectors of the economy such as education, planning, health and manufacturing.

CHAPTER TWO

LITERATURE REVIEW

2.1 Robust Estimation of Finite Population Total

In this approach, a regression model is used to quantify the contribution of the auxiliary variable X to the survey variable Y per unit value x to summarize the relationship between the two variables to predict the survey variable for a given value x and to extrapolate the results beyond a given range of auxiliary values.

The nonparametric regression approach currently offers results to researchers where other approaches such as design based approaches have failed. Some advantages of nonparametric regression are:

1) It gives predictions of observations yet to be made without reference to a fixed parametric model.
2) It provides a tool for finding spurious observations by studying the influence of isolated points.
3) It provides a versatile method of exploring the relationship between the auxiliary variable and the survey variable.

Given a population of N identifiable units $U_1, U_2, \ldots, U_N$ the estimation of finite population total $T = \sum_{i=1}^{N} y_i$ needs to be done assuming that there exists the required auxiliary information for all the population units.

The nonparametric estimation problem is therefore concerned with the estimation of population total, T, using a model based approach where the model is of the form

$$y_i = m(x_i) + \delta^2(x_i)e_i \qquad (2.1)$$

Where $m(x_i)$ and $\delta^2(x_i)$ are smooth functions of x_i and

$$E(y_i \,/\, X_i = x_i) = m(x_i) \qquad (2.2)$$

$$Cov(y_i, y_j \,/ X_i = x_i, X_j = x_j) = \begin{cases} \delta^2(x_i) \;\; for\ i = j \\ 0\ ,\ elsewhere \end{cases} \qquad (2.3)$$

This nonparametric approach doesn't restrict the form of distribution nor does it specify the stochastic properties such as expectation, variance and Mean Squared Error.

In the next subsection, existing nonparametric estimators are reviewed.

2.1.1 The Nadaraya-Watson Estimator

The estimation of finite population totals has received considerable attention in previous research works. In his work, Dorfman (1992) introduces a nonparametric regression estimator for finite population totals based on a sample drawn from the population. Taking into consideration a population consisting of N units, the author seeks to estimate the finite population total defined by

$$T = \sum_N y_i \quad where\ i = 1, \dots, N \qquad (2.4)$$

The estimation of finite population totals was carried out by first expressing equation (2.4) as the sum of sample component and nonsample component

$$T = \sum_{i \in s} y_i + \sum_{j \in p-s} y_j \qquad (2.5)$$

Where s is the sample size and p is the population size. The task was to estimate the non-sampled values of the second part of equation (2.5). To do this and assuming availability of auxiliary variables, the author used the model

$$y_i = m(x_i) + \sigma(x_i)\varepsilon_i \qquad (2.6)$$

where ε is independent and identically distributed with zero mean and constant variance and $m(x_i)$ is a Lipchitz function, to estimate the function $m(x_i)$. Using a symmetric density function, Dorfman (1992) used the Nadaraya-Watson weights defined by $w_i(x) = \frac{k_b(x_i - x)}{\sum_{i=1}^{n} k_b(x_i - x)}$, where b is the bandwidth, to estimate $m(x)$ thus yielding the Nadaraya-Watson estimator

$$\widehat{m}(x) = \sum w_i(x)\, y_i \qquad (2.7)$$

The kernel function is always under the user's control and is defined by $K(.) = \frac{1}{nb} K\left(\frac{x_i - x_j}{b}\right)$. The assumption made is that the kernel is a symmetrical function satisfying the following properties (Silverman, 1986)

1) $K(t) \geq 0$ for all t
2) $\int K(t)dt = 1$
3) $\int tK(t)dt = 0$
4) $\int t^2 K(t)dt = k_2$
5) $\int_{-\infty}^{\infty} [K(t)]^2 dt < \infty$

6) $K(t) = K(-t)$

For $x = x_j$ in the sample, the estimator for the population total is then defined by

$$\hat{T} = \sum_{i \in s} y_i + \sum_{j \in p-s} \hat{m}(x_j) \tag{2.8}$$

The relative mean squared error of the estimator diminishes to zero as long as the standard conditions, $n \to 0\ and\ nb \to \infty,$ are met. The ratio of the bias to that of the standard error was found to be asymptotically zero, suggesting that a wide choice of bandwidth could be satisfactory in practice.

After a series of simulations, the estimator was found to be more efficient than the design-based estimators, Dorfman (1992). The larger the bandwidth, the broader and flatter the density function, the more equal are the weights and the smoother the estimated function. However, the estimator was found to be of greatest efficiency when the variance was assumed to be proportional to the square of the auxiliary variable.

Also in their work, Dorfman and Hall (1993) considered the estimation of the distribution function of a given variable over a finite population for a given sample of units. They defined the distribution function for a variable of interest y by

$F(t) = \frac{1}{N}\left\{\sum_i I(y_i \leq t) + \sum_j I(y_j \leq t)\right\}$ where i indicates sample values, j indicates nonsample units and $I(.)$ is the standard indicator function. The task is to estimate the nonsample values.

To estimate $F(.)$ Dorfman and Hall (1993) worked under the assumption that a regression relationship exists between the survey variable and the auxiliary variables. Dorfman and Hall (1993) considered three relationships that are likely to arise in practice namely:

1) $y_i = a + bx_i + \varepsilon_i$ for $i = 1, \dots, N$ with $E(\varepsilon_i = 0)$, $Var(\varepsilon_i) = \sigma^2$ and $Cov(\varepsilon_i, \varepsilon_j) = 0$
2) $y_i = m(x_i) + \varepsilon_i$ for $i = 1, \dots, N$ with $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2$ and $m(x_i)$ is a smooth function.
3) A relationship between $h(y_i) \equiv I(y_i \leq t)$ and x_i was assumed so that $E(h(y_i)) = H(x_i)$ where $H(x_i)$ is a smooth function.

Dorfman and Hall (1993) considered design-based and model-based estimators when either model (1) or (2) was utilized, and design calibrated and model calibrated estimators when model (1) was utilized.

The nonparametric calibrated estimator performed best and its bias was of the same order as that of model-based estimators when the model was correctly specified but it did not share in their vulnerability if the model was misspecified.

2.1.2 Local Polynomial Estimator

Breidt and Opsomer (2000) re-looked at the Horvitz–Thompson estimator of the population total given by $\hat{t}_y = \sum_{i \in s} \frac{y_i}{\pi_i}$ where π_i is the inclusion probability.

The Horvitz-Thompson estimator does not depend on auxiliary information and therefore their work was to improve on the efficiency of the estimator by incorporating the auxiliary information into the model of population totals. They used the local polynomial approach and the survey values y_i are realized values of the model defined in equation (2.6).

Using a continuous kernel function K and a bandwidth h_N, they defined a local polynomial kernel estimator of degree q based on the entire population. By letting $y_U = [y_i]_{i \in U_N}$ be a N-vector of survey values in the finite population, they defined a matrix of dimension $N \times (q+1)$ by $X_{U_i} = \begin{bmatrix} 1 & x_1 - x_i & \cdots & (x_1 - x_i)^q \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_N - x_i & \cdots & (x_N - x_i)^q \end{bmatrix}$ and a $N \times N$ matrix by $w_{U_i} = diag\left\{\frac{1}{h_N} K\left(\frac{x_j - x_i}{h_N}\right)\right\}$

With e_1 being a vector with a 1 in the 1^{st}position and 0 elsewhere, the estimator of the regression function at $m(x_i)$ is then given by

$$m_i = e_1'\left(X_{U_i}' W_{U_i} X_{U_i}\right)^{-1} X_{U_i}' W_{U_i} y_U \qquad (2.9)$$

as long as $X_{U_i}' W_{U_i} X_{U_i}$ is invertible.

The designed-unbiased estimator of the population total, t_y is then given by

$$t_y^* = \sum_{i \in s} \frac{y_i - m_i}{\pi_i} + \sum_{i \in U_N} m_i \qquad (2.10)$$

which is a generalized difference estimator with variance defined by

$$V_P\left(t_y^*\right) = \sum_{i,j\in U_N}\left(\pi_{ij} - \pi_i\pi_j\right)\frac{y_i - m_i}{\pi_i}\frac{y_j - m_j}{\pi_i} \tag{2.11}$$

However, the estimator in equation (2.9) is based on the entire finite population. The sample based consistent estimator of the regression function $m(x_i)$ is then given by

$$\hat{m}_i^o = e_1'\left(X_{s_i}'W_{s_i}X_{s_i}\right)^{-1}X_{s_i}'W_{s_i}y_s \tag{2.12}$$

The regression estimator of the population total is

$$\tilde{t}_y^o = \sum_{i\in s}\frac{y_i - \hat{m}_i^o}{\pi_i} + \sum_{i\in U_N}\hat{m}_i^o \tag{2.13}$$

For observations less than $(q+1)$, the matrix $X_{U_i}'W_{U_i}X_{U_i}$ is singular. Therefore Breidt and Opsomer (2000) considered an adjusted sample based estimator that is guaranteed to exist for any sample drawn from the population. The adjusted sample smoother is given by

$$\hat{m}_i = e_1'\left(X_{s_i}'W_{s_i}X_{s_i} + diag\left\{\frac{\delta}{N^2}\right\}_{j=1}^{q+1}\right)^{-1} X_{s_i}'W_{s_i}y_s \tag{2.14}$$

The estimator of the population total is

$$\tilde{t}_y = \sum_{i\in s}\frac{y_i - \hat{m}_i}{\pi_i} + \sum_{i\in U_N}\hat{m}_i \tag{2.15}$$

The sample based estimator of population totals is a linear combination of the survey values with weights being the inverse inclusion probabilities. The estimator,$\tilde{t}_y$, that uses the adjusted sample smoother in equation (2.14) was found to be asymptotically design unbiased and design consistent. The variance of the estimator was also found to be design unbiased and design consistent for the asymptotic mean squared error. The estimator satisfied the property of asymptotic normality and was found to be robust in the sense that it attained the Godambe-Joshi lower bound.

The performance of the estimator was compared with that of other parametric and nonparametric estimators. Both parametric and nonparametric regression estimators performed better than the Horvitz-Thompson estimator. However, the local polynomial regression estimator by Breidt and Opsomer (2000) was the best estimator among the nonparametric estimators that were considered.

In their work, Odhiambo and Mwalili (2000) considered the application of nonparametric regression to the estimation of finite population error variance for a given sample drawn from the population. The error variance obtained by Dorfman (1992) was a function of $\sigma^2(x_i)'s$ that are unknown.

By considering the squared residual $\hat{e}_j^2 = \left(y_j - \widehat{m}(x_j)\right)^2$ and using some mild assumptions on $m(x_i)$ and $x_i's$, the authors showed that $E\left(\hat{e}_j^2 / X_j = x_j\right) = \sigma^2(x_j) + O(n^{-1})$ is an asymptotic unbiased estimator of $\sigma^2(x_j)$. They obtained an improved estimator of $\sigma^2(x_j)$ by smoothing $\hat{e}_j^2$ for $j \in s$ and $(x_j, y_j)'s$ being sample points close to (x_i, y_i).

Letting the smoothing parameter be h and defining the weights $w_i(x)$ then $\hat{\sigma}_{np}^2(x_i) = \sum_{j \in s} W_j(x_i)\, \hat{e}_j^2$ so that the error variance is estimated by

$$V_n = \sum_{j \in s} W_j^2(X_i)\hat{\sigma}_{np}^2(x_i) + \sum_{j \in r} \hat{\sigma}_{np}^2(x_i) \tag{2.16}$$

The expectation of equation (2.16) yields

$$E(V_n) \approx \sum_{j \in s} \left[W_j^2(X_i)\sigma^2(x_i) + \frac{h^2}{2}\sigma^2(x_i)d_k \right] + \sum_{i \in r} \left[\sigma^2(x_i) + \frac{h^2}{2}\sigma^2(x_i)d_k \right] \tag{2.17}$$

where $d_k = \int_{\frac{x_i - x_{j-1}}{h}}^{\frac{x_i - x_j}{h}} u^2 k(u)du.$

For $h \to 0$ and $n \to \infty$, $E(V_n) \approx Var\left(\hat{T}_{np} - T\right)$ asymptotically. Thus V_n was found to be robust against model misspecification.

Ombui (2008) used local polynomial regression in the estimation of finite population totals. Using the model defined in equation (2.6), he applied the technique of using a strip of data around the covariate X in order to fit a line through the set of data (x_i, y_i) $i = 1,2, \dots, n$

The estimator yielded better results in estimating the finite population total. Moreover, the estimator was found to be asymptotically unbiased, consistent and normally distributed when certain conditions were satisfied.

2.1.3 The Gasser-Muller Estimator

When deriving the asymptotic properties of the Nadaraya-Watson estimator, it becomes tedious to find the derivatives of the estimator due to the nature of the denominator of the estimator. Gasser and Muller (1979) proposed an estimator that involved the sorting of X variable. The estimator is given by

$$\widehat{m}(x) = \sum_{j=1}^{n} \int_{s_{j-1}}^{s_j} k(u-1)du\, s_j \tag{2.18}$$

Where $s_j = \frac{1}{2}(x_j + x_{j+1}), x_0 = -\infty$ and $x_{n+1} = \infty$.

The corresponding nonparametric estimator of finite population total is therefore given as

$$\widehat{T}_G = \sum_{i\in s} y_i + \sum_{j\in r} \widehat{m}(x_j) \tag{2.19}$$

2.1.4 The Priestly-Chao Estimator

According to Priestly and Chao (1972), the Priestly-Chao weight is described by the relation $w_i(x_j) = \left(\frac{x_i - x_{i-1}}{h}\right) K\left(\frac{x_i - x_j}{h}\right)$ such that the Priestly-Chao estimator is given by

$$\widehat{m}_c(x_j) = \frac{1}{nh} \sum_{i\in s} w_i(x_j) y_i \tag{2.20}$$

However this smoothing function has a shortcoming when one needs to extrapolate various values of the survey variable. Furthermore, unlike the usual weighting scheme where the weights sum to one, in this particular case the sum of the weights is not equal to one but rather the sum is an approximation.

Moreover, this estimator assumes that the data set is ordered such that $x_{i-1} < x_i$ and the weights are only applicable to instances where the auxiliary variable is restricted to some interval. Odhiambo (1995) therefore gave the estimator of finite population total as

$$\widehat{T}_C = \sum_{i\in s} y_i + \sum_{j\in r} \widehat{m}_c(x_j) \tag{2.21}$$

2.1.5 The Spline Estimator

In this technique, the residual sum of squares is used to compute the regression function and is given by

$$\widehat{m}_s(x_j) = \sum_{i=1}^{n}(y_i - g(x_i))^2 \tag{2.22}$$

Where $g(x)$ is a curve restricted to the functional form. The distance can be reduced by using any $g(x)$ that is used to interpolate the data.

This technique yields good results because it will produce a good fit and the curve does not have too much variation. Zheng and Little (2003) therefore gave the spline estimator for finite population total to be

$$\hat{T}_C = \sum_{i\in s} y_i + \sum_{j\in r} \widehat{m}_s(x_j) \tag{2.23}$$

2.2 Selection of the Kernel Function

There exists many possible kernel smoothers but the selected kernel should be easy to implement both theoretically and practically. Silverman (1986) gave requirements that ought to be met by the smoother. The requirements are:

1) The kernel smoother should be easier and simple to construct.
2) The kernel smoother should not take very small values because they may result in numerical underflow in the computer.
3) The kernel smoother should be user friendly i.e it should theoretically and practically fit in both simulated and raw data.

4) The range of the smoother should be well defined and not open as in the case of the Gaussian kernel.

Table 2.1 gives the efficiency of various kernels with respect to the Epanechnkov kernel.

Table 2.1: Efficiency Relative to Epanechnkov Kernel

Kernel	$\boldsymbol{K(t)}$	**Efficiency**
Epanechnkov	$\begin{cases} \frac{3}{4\sqrt{5}}\left(1-\frac{t^2}{5}\right) & \lvert t\rvert < \sqrt{5} \\ 0 & otherwise \end{cases}$	1.0000
Biweight	$\begin{cases} \frac{15}{16}(1-t^2)^2 & \lvert t\rvert < 1 \\ 0 & otherwise \end{cases}$	0.9939
Triangular	$\begin{cases} \quad 1-\lvert t\rvert & \\ 0 & elsewhere \end{cases} \lvert t\rvert < 1$	0.9859
Gaussian	$\frac{1}{\sqrt{2\pi}}e^{\frac{-1t^2}{2}} \quad -\infty < t < \infty$	0.9512
Rectangular	$\begin{cases} \frac{1}{2} & \lvert t\rvert < 1 \\ 0 & elsewhere \end{cases}$	0.9295

2.6 Research Gap

Having reviewed the various methods on nonparametric estimation of finite population total, all methods employed kernel smoothers in the estimation of regression functions. Most kernel smoothers have boundary problems and require modifications at the boundary points. That is, towards the boundary points the bias of the estimators decreases at the cost of an increasing variance. Moreover, there exists trade-off between the bias and variance of the estimators. Selecting a narrow window results in low bias and high variance while selecting a wide window yields a high bias and low variance. Moreover, locally weighted averages can be highly biased if the regression function has a significant slope.

Further, there exists no framework for the selection of optimal bandwidth for the kernel smoothers. For small bandwidth, the tails of the density function are more wiggly and smoother when the bandwidth is wider.

This study uses multiplicative bias corrected approach to the nonparametric estimation of finite population total. In the multiplicative bias corrected approach, the estimator has low bias with no cost to the variance. Under sufficient smoothness of the density function, the multiplicative bias corrected technique reduces the order of the bias with no effect on the variance of the estimator.

CHAPTER THREEE

METHODOLOGY

3.1 Types of Populations

There are two types of population mainly the finite population and the infinite population. In the finite population, units are known, distinct and the size is known while in the infinite population, the units may be distinct but the size is not known with certainty.

3.2 Estimation of Finite Population

Estimation of finite population totals is taken into consideration in this study. Suppose that there are sampling units $U_1, U_2, \dots, U_N$ with corresponding survey measurements $y_1, y_2, \dots, y_N$ for the survey variable $Y.$ If all the units are labeled and supposing that in each unit it is possible to collect survey measurements, then it is possible to determine the finite population total for any set of data collected.

3.3 Approaches Used to Estimate Finite Population Total

The main approaches used in the estimation of finite population total are:

a) The classical approach (design based)

b) Model-based or super-population approach

c) Model assisted approach

d) Design assisted approach

3.3.1 Classical Approach

In the classical approach, the observed values of the survey variable Y given by $y_1, y_2, \dots, y_N$ are assumed to be unknown but fixed constants. In this concept, a sample is drawn from the finite population and the sample measurements are then used to estimate the population parameter of interest. Standard sampling designs are well discussed in Cochran (1977). The problem with this approach is that it is assumed that all samples in the population are selected. This is not possible mainly because of the problems associated with selection of samples.

3.3.2 Model-Based or Super-Population Approach

In the model-based approach, an assumption that the actual survey measurements $y_1, y_2, \dots, y_N$ are realized values of the random vector $Y_1, Y_2, \dots, Y_N$ is made. In this approach, knowledge is summarized using a model defined by $Y_i = m(X_i) + e_i$ for $i = 1,2, \dots, N$ where $m(X_i)$ is a smooth function and e_i is a sequence of independent and identically distributed random variables with zero mean and finite variance. The estimator of the population total is then defined as

$\hat{T} = \sum_{i=1}^{N} Y_i = \sum_{i \in s} Y_i + \sum_{i \in r} Y_i$ where $\sum_{i \in s} Y_i$ denote the sample proportion and $\sum_{i \in r} Y_i$ denote the non sample proportion. The problem of estimating the population total T therefore reduces to the problem of estimating the non-sample values $\sum_{i \in r} Y_i$.

3.3.3 Model Assisted Approach

Model assisted survey estimation of population total is a well-known approach that incorporates auxiliary information into the design-based estimation of finite population total. This approach assumes the existence of a superpopulation model between the auxiliary variables and the variable of interest for the population to be sampled. The population quantities of interest are estimated in such a way that the design-based properties of the estimators can be established. This contradicts the model-based approach for which the design-based inference is not possible.

In this approach, the model is used to increase the efficiency of the estimators, but even when the model is not correct, estimators typically remain design-consistent. Since the model assisted estimation has a great potential to improve the precision of the required survey estimators when the appropriate auxiliary information is available, it often requires that these models are linear or should at least have a known parametric shape.

Of these survey approaches, the model based approach has been considered to be the most consistent method of estimation.

3.4 Multiplicative Bias Corrected Approach

In this section, the exact procedure of estimating the population total for a finite population is now presented. In this it is assumed that there are sampling units $U_1, U_2, \dots, U_N$ with corresponding survey measurements $Y_1, Y_2, \dots, Y_N$ so that the population total is denoted by T and defined as

$$T = \sum_{i=1}^{N} Y_i \tag{3.1}$$

The estimator of equation (3.1) is proposed based on the model

$$Y = \mu(x_i) + \varepsilon_i$$

$$E(Y) = \mu(x_i) \tag{3.2}$$

$$Cov(Y_i, Y_j) = \begin{cases} \sigma^2(x_i), \ for\ i = j \\ 0, \quad otherwise \end{cases}$$

Where $\mu(x_i)$ and $\sigma^2(x_i)$ are assumed to be smooth functions of x_i. This is mainly because equation (3.2) is the simplest form of equation that describes the relationship between the auxiliary variable and the survey variable.

Suppose that to each of these $Y_i's$, some auxiliary information $X_1, X_2, \dots, X_N$ is available and that these auxiliary variables are to be considered in the estimation process. Then the construction may be re-written to take the predictive form

$$T = \sum_{i\in s} Y_i + \sum_{i\epsilon p-s} Y_i \tag{3.3}$$

Where $\sum_{i\in s} Y_i$ provides the proportion that is truly observed with $\sum_{i\epsilon p-s} Y_i$ providing the proportion that is not observed but estimated using the corresponding auxiliary information.

To estimate equation (3.3), several methods may be employed. In this study, an estimator is proposed as

$$\hat{T} = \sum_{i\in s} Y_i + \sum_{i\epsilon p-s} \hat{Y}_i \tag{3.4}$$

In equation (3.4), $\sum_{i\epsilon p-s} Y_i$ may be difficult to state with exactness and thus the problem reduces to that of predicting $\sum_{i\epsilon p-s} Y_i$.

To this problem, the estimator $\sum_{i\in p-s} \mu(x_i)$ is proposed where $\mu(x_i)$ is a smooth function. Therefore the estimator in equation (3.4) becomes

$$\hat{T}_{MBC} = \sum_{i\in s} Y_i + \sum_{i\epsilon p-s} \mu(x_i) \tag{3.5}$$

The task is to estimate the second part of equation (3.5). To do this, the multiplicative bias corrected technique is employed in which case the proposed estimator of the population total is now defined as

$$\hat{T}_{MBC} = \sum_{i\in s} Y_i + \sum_{i\epsilon p-s} \hat{\mu}_n(x_i) \tag{3.6}$$

Where $\hat{\mu}_n(x_i)$ is as defined in equation (3.9)

Suppose that $(X_1, Y_1), (X_2, Y_2), \dots, (X_N, Y_N)$ are N independent pairs of random variables (X, Y) with real values. Assuming that the explanatory variable X has a probability density f and thus model the dependence of the univariate survey variable Y to the explanatory variable X through a non-parametric model $Y = \mu(X) + \varepsilon$. The function $\mu(X)$ is smooth and the error term has zero mean and finite variance that is independent of the covariate X.

Define a pilot smoother of the regression function as

$$\tilde{\mu}_n(x) = \sum_{j=1}^{n} w_i\,(x) y_j \tag{3.7}$$

Then the ratio $\beta_j = \frac{y_i}{\tilde{\mu}_n(x_j)}$ is a noisy estimate of the inverse relative estimation error of the smoother $\tilde{\mu}_n$ given by $\frac{\mu(X_j)}{\tilde{\mu}_n(X_j)}$.

Smoothing β_j yields

$$\hat{\alpha}(x) = \sum_{j=1}^{n} w_j(x;h)\,\beta_j = \sum_{j=1}^{n} w_j(x;h)\,\frac{y_i}{\tilde{\mu}_n(x_j)} \qquad (3.8)$$

Equation (3.8) can then be used as a multiplicative correction of the pilot smoother in equation (3.7) which can now be defined by

$$\hat{\mu}_n(x_i) = \hat{\alpha}(x)\tilde{\mu}_n(x) \qquad (3.9)$$

Assumptions

The following assumptions are made in the estimation of $\hat{\mu}(x_i)$

a) The regression function is bounded and strictly positive i.e

b) $0 < a \leq \mu(x_i) \leq b.$

c) The regression function is twice continuously differentiable everywhere.

d) ε_i has finite fourth moments and has a symmetric distribution around zero.

e) The bandwidth h is such that $h \to 0, nh \to \infty, nh^3 \to \infty\ as\ n \to \infty.$

The positivity assumption on the regression function,$\mu(x_i)$, is classical when performing multiplicative bias correction. It is important to note that the regression function might cross the $x - axis$. In such a situation, Glad (1998) proposes to shift all the response data by a distance a such that the new regression function is

$\mu(x_i) + a.$

Using equation (3.8) in equation (3.9) easily yields

$$\hat{\mu}_n(x_i) = \sum_{j=1}^{n} w_j(x;h)\,\frac{\tilde{\mu}_n(x)}{\tilde{\mu}_n(x_j)}\,y_j \qquad (3.10)$$

Now suppose that

$$E[\tilde{\mu}_n(x)/x_1,\dots,x_N] = \sum_{j=1}^{n} w_j(x;h)\,E[Y_j] = \sum_{j=1}^{n} w_j(x;h)\mu(X_j) = \bar{\mu}_n(x) \qquad (3.11)$$

Then using $\frac{\tilde{\mu}_n(x)}{\tilde{\mu}_n(x_j)}$ found in equation (3.10) yields

$$\frac{\tilde{\mu}_n(x)}{\tilde{\mu}_n(x_j)} = \frac{\bar{\mu}_n(x)}{\bar{\mu}(X_j)} \times \frac{\tilde{\mu}_n(x)}{\bar{\mu}_n(x)} \times \left(\frac{\tilde{\mu}_n(x_j)}{\bar{\mu}(X_j)}\right)^{-1} \qquad (3.12)$$

$$\frac{\tilde{\mu}_n(x)}{\tilde{\mu}_n(x_j)} = \frac{\bar{\mu}_n(x)}{\bar{\mu}(X_j)} \times \left(\frac{\bar{\mu}_n(x)+\tilde{\mu}_n(x)-\bar{\mu}_n(x)}{\bar{\mu}_n(x)}\right) \times \left(\frac{\bar{\mu}(X_j)+\tilde{\mu}_n(x_j)-\bar{\mu}(X_j)}{\bar{\mu}(X_j)}\right)^{-1} \qquad (3.13)$$

$$\frac{\tilde{\mu}_n(x)}{\tilde{\mu}_n(x_j)} = \frac{\bar{\mu}_n(x)}{\bar{\mu}(X_j)} \times \left(1 + \frac{\tilde{\mu}_n(x)-\bar{\mu}_n(x)}{\bar{\mu}_n(x)}\right) \times \left(1 + \frac{\tilde{\mu}_n(x_j)-\bar{\mu}(X_j)}{\bar{\mu}(X_j)}\right)^{-1} \qquad (3.14)$$

For ease of derivation let $\frac{\tilde{\mu}_n(x)-\bar{\mu}_n(x)}{\bar{\mu}_n(x)} = b_n(x)$ and $\frac{\tilde{\mu}_n(x_j)-\bar{\mu}(X_j)}{\bar{\mu}(X_j)} = b_n(X_j)$. Equation (3.14) therefore becomes

$$\frac{\tilde{\mu}_n(x)}{\tilde{\mu}_n(x_j)} = \frac{\bar{\mu}_n(x)}{\bar{\mu}(X_j)} \times \left(1 + b_n(x)\right) \times \left(1 + b_n(X_j)\right)^{-1} \qquad (3.15)$$

Applying the binomial expansion to $\left(1 + b_n(x)\right) \times \left(1 + b_n(X_j)\right)^{-1}$gives

$$\left(1 + b_n(x)\right) \times \left(1 + b_n(X_j)\right)^{-1} = [1 + b_n(x)]\left[1 - b_n(X_j) + b_n(X_j)^2\right]$$

which further reduces to

$$\left(1 + b_n(x)\right) \times \left(1 + b_n\left(X_j\right)\right)^{-1} = 1 + b_n(x) - b_n\left(X_j\right) + r_j\left(x, X_j\right) \tag{3.16}$$

where $r_j\left(x, X_j\right)$ is the remainder term that involves the terms x and X_j.

Using equation (3.16) in equation (3.15) yields

$$\frac{\widetilde{\mu}_n(x)}{\widetilde{\mu}_n(x_j)} = \frac{\overline{\mu}_n(x)}{\overline{\mu}(X_j)} \times \left[1 + b_n(x) - b_n\left(X_j\right) + r_j\left(x, X_j\right)\right] \tag{3.17}$$

Substituting equation (3.17) into equation (3.10) and using the model $Y_j = \mu\left(X_j\right) + \varepsilon_j$ one obtains

$$\hat{\mu}_n(x_i) = \sum_{j=1}^{n} w_j(x; h) \left\{\frac{\overline{\mu}_n(x)}{\overline{\mu}(X_j)} \times \left[1 + b_n(x) - b_n\left(X_j\right) + r_j\left(x, X_j\right)\right]\left[\mu\left(X_j\right) + \varepsilon_j\right]\right\}$$

(3.18)

$$\begin{gathered}\hat{\mu}_n(x_i) = \\ \sum_{j=1}^{n} w_j(x; h) \left\{\frac{\overline{\mu}_n(x)}{\overline{\mu}(X_j)} \mu\left(X_j\right)\left[1 + b_n(x) - b_n\left(X_j\right) + r_j\left(x, X_j\right)\right]\right\} + \\ \sum_{j=1}^{n} w_j(x; h) \left\{\frac{\overline{\mu}_n(x)}{\overline{\mu}(X_j)} \varepsilon_j\left[1 + b_n(x) - b_n\left(X_j\right) + r_j\left(x, X_j\right)\right]\right\}\end{gathered} \tag{3.19}$$

$$\begin{gathered}\hat{\mu}_n(x_i) = \sum_{j=1}^{n} w_j(x; h) \frac{\overline{\mu}_n(x)}{\overline{\mu}(X_j)} \mu\left(X_j\right) + \sum_{j=1}^{n} w_j(x; h) \frac{\overline{\mu}_n(x)}{\overline{\mu}(X_j)} \left\{\varepsilon_j + \mu\left(X_j\right)\left[b_n(x) - \right.\right. \\ \left.\left. b_n\left(X_j\right)\right]\right\} + \sum_{j=1}^{n} w_j(x; h) \frac{\overline{\mu}_n(x)}{\overline{\mu}(X_j)} \varepsilon_j\left[b_n(x) - b_n\left(X_j\right)\right] + \\ \sum_{j=1}^{n} w_j(x; h) \frac{\overline{\mu}_n(x)}{\overline{\mu}(X_j)} r_j\left(x, X_j\right)\left[\mu\left(X_j\right) + \varepsilon_j\right]\end{gathered} \tag{3.20}$$

Using the assumption $nh \to \infty$, the remainder terms converge to zero in probability. Therefore $r_j(x, X_j)[\mu(X_j) + \varepsilon_j] = O_P\left(\frac{1}{nh}\right)$ and equation (3.20) reduces to

$$\hat{\mu}_n(x_i) = \sum_{j=1}^{n} w_j(x;h)\frac{\bar{\mu}_n(x)}{\bar{\mu}(X_j)}\mu(X_j) + \sum_{j=1}^{n} w_j(x;h)\frac{\bar{\mu}_n(x)}{\bar{\mu}(X_j)}\{\varepsilon_j + \mu(X_j)[b_n(x) - b_n(X_j)]\} + \sum_{j=1}^{n} w_j(x;h)\frac{\bar{\mu}_n(x)}{\bar{\mu}(X_j)}\varepsilon_j[b_n(x) - b_n(X_j)] + O_P\left(\frac{1}{nh}\right) \tag{3.21}$$

Our estimator for finite population total in equation (3.6) therefore becomes

$$\hat{T}_{MBC} = \sum_{i\in s} Y_i + \sum_{i\epsilon p-s}\left\{\sum_{j=1}^{n} w_j(x;h)\frac{\bar{\mu}_n(x)}{\bar{\mu}(X_j)}\mu(X_j) + \sum_{j=1}^{n} w_j(x;h)\frac{\bar{\mu}_n(x)}{\bar{\mu}(X_j)}\{\varepsilon_j + \mu(X_j)[b_n(x) - b_n(X_j)]\} + \sum_{j=1}^{n} w_j(x;h)\frac{\bar{\mu}_n(x)}{\bar{\mu}(X_j)}\varepsilon_j[b_n(x) - b_n(X_j)] + O_P\left(\frac{1}{nh}\right)\right\} \tag{3.22}$$

3.1 Asymptotic Unbiasedness of the Proposed Estimator

Under the model based approach, the bias of the estimator $\hat{T}_{MBC}$ is defined by

$$E[\hat{T}_{MBC} - T] = E[\hat{T}_{MBC}] - E[T] \tag{3.23}$$

Next, the expected value of the proposed estimator for population total is calculated. Now

$$E[\hat{T}_{MBC}] = E\left[\sum_{i\in s} Y_i + \sum_{j\epsilon p-s}\left\{\sum_{j=1}^{n}\hat{\mu}_n(x_i)\right\}\right] = \sum_{i\in s} E[Y_i] + \sum_{i\epsilon p-s}\sum_{j=1}^{n} E[\hat{\mu}_n(x_i)] \tag{3.24}$$

The calculation of $E[\hat{\mu}_n(x_i)]$ is based on establishing a stochastic approximation of the estimator $\hat{\mu}_n(x_i)$ in which each term can be directly analyzed.

$$E[\hat{\mu}_n(x_i)] = E\left\{\sum_{j=1}^{n} w_j(x;h)\frac{\bar{\mu}_n(x)}{\bar{\mu}(X_j)}\mu(X_j) + \sum_{j=1}^{n} w_j(x;h)\frac{\bar{\mu}_n(x)}{\bar{\mu}(X_j)}\{\varepsilon_j + \mu(X_j)[b_n(x) - b_n(X_j)]\} + \sum_{j=1}^{n} w_j(x;h)\frac{\bar{\mu}_n(x)}{\bar{\mu}(X_j)}\varepsilon_j[b_n(x) - b_n(X_j)] + O_P\left(\frac{1}{nh}\right)\right\} \quad (3.25)$$

$$E[\hat{\mu}_n(x_i)] = E\left\{\sum_{j=1}^{n} w_j(x;h)\frac{\bar{\mu}_n(x)}{\bar{\mu}(X_j)}\mu(X_j) + \sum_{j=1}^{n} w_j(x;h)\,A_j(x) + \sum_{j=1}^{n} w_j(x;h)\,B_j(x)\right\} + O_P\left(\frac{1}{nh}\right)$$

(3.26)

Where

$A_j(x) = \frac{\bar{\mu}_n(x)}{\bar{\mu}(X_j)}\{\varepsilon_j + \mu(X_j)[b_n(x) - b_n(X_j)]\}$ and

$B_j(x) = \frac{\bar{\mu}_n(x)}{\bar{\mu}(X_j)}\varepsilon_j[b_n(x) - b_n(X_j)]$

Analyzing the first term of equation (3.26)

$$E\left\{\sum_{j=1}^{n} w_j(x;h)\frac{\bar{\mu}_n(x)}{\bar{\mu}(X_j)}\mu(X_j)\right\} = \sum_{j=1}^{n} w_j(x;h)\frac{\bar{\mu}_n(x)}{\bar{\mu}(X_j)}E[\mu(X_j)]$$

which yields

$$E\left\{\sum_{j=1}^{n} w_j(x;h)\frac{\bar{\mu}_n(x)}{\bar{\mu}(X_j)}\mu(X_j)\right\} = \sum_{j=1}^{n} w_j(x;h)\frac{\bar{\mu}_n(x)}{\bar{\mu}(X_j)}\mu(X_j) \quad (3.27)$$

This is mainly because $\mu(X_j)$ is the mean function given in equation (3.2)

Analyzing the second term of equation (3.26)

$$E\left\{\sum_{j=1}^{n} w_j(x;h)\frac{\bar{\mu}_n(x)}{\bar{\mu}(X_j)}\left\{\varepsilon_j+\mu(X_j)\left[b_n(x)-b_n(X_j)\right]\right\}\right\}=E\left\{\sum_{j=1}^{n} w_j(x;h)\left[\frac{\bar{\mu}_n(x)}{\bar{\mu}(X_j)}\varepsilon_j+\frac{\bar{\mu}_n(x)}{\bar{\mu}(X_j)}\mu(X_j)\left\{\frac{\tilde{\mu}_n(x)-\bar{\mu}_n(x)}{\bar{\mu}_n(x)}\right\}-\frac{\bar{\mu}_n(x)}{\bar{\mu}(X_j)}\mu(X_j)\left\{\frac{\tilde{\mu}_n(x_j)-\bar{\mu}(X_j)}{\bar{\mu}(X_j)}\right\}\right]\right\} \tag{3.28}$$

$$E\left\{\sum_{j=1}^{n} w_j(x;h)\frac{\bar{\mu}_n(x)}{\bar{\mu}(X_j)}\left\{\varepsilon_j+\mu(X_j)\left[b_n(x)-b_n(X_j)\right]\right\}\right\}=\sum_{j=1}^{n} w_j(x;h)\frac{\bar{\mu}_n(x)}{\bar{\mu}(X_j)}E\left[\varepsilon_j\right]+\sum_{j=1}^{n} w_j(x;h)\frac{\mu(X_j)}{\bar{\mu}(X_j)}E\left[\varepsilon_j\right]-\sum_{j=1}^{n} w_j(x;h)\frac{\bar{\mu}_n(x)\mu(X_j)}{\bar{\mu}(X_j)^2}E\left[\tilde{\mu}_n(x_j)\right]+\sum_{j=1}^{n} w_j(x;h)\,E\left[\frac{\bar{\mu}_n(x)\mu(X_j)}{\bar{\mu}(X_j)}\right] \tag{3.29}$$

$$E\left\{\sum_{j=1}^{n} w_j(x;h)\frac{\bar{\mu}_n(x)}{\bar{\mu}(X_j)}\left\{\varepsilon_j+\mu(X_j)\left[b_n(x)-b_n(X_j)\right]\right\}\right\}=0+0-\sum_{j=1}^{n} w_j(x;h)\frac{\bar{\mu}_n(x)\mu_n(X_j)\bar{\mu}(X_j)}{\bar{\mu}(X_j)^2}+\sum_{j=1}^{n} w_j(x;h)\frac{\bar{\mu}_n(x)\mu_n(X_j)}{\bar{\mu}(X_j)}E[1] \tag{3.30}$$

$$E\left\{\sum_{j=1}^{n} w_j(x;h)\frac{\bar{\mu}_n(x)}{\bar{\mu}(X_j)}\left\{\varepsilon_j+\mu(X_j)\left[b_n(x)-b_n(X_j)\right]\right\}\right\}=0+0-\sum_{j=1}^{n} w_j(x;h)\frac{\bar{\mu}_n(x)\mu_n(X_j)}{\bar{\mu}(X_j)}+\sum_{j=1}^{n} w_j(x;h)\frac{\bar{\mu}_n(x)\mu_n(X_j)}{\bar{\mu}(X_j)} \tag{3.31}$$

$$E\left\{\sum_{j=1}^{n} w_j(x;h)\frac{\bar{\mu}_n(x)}{\bar{\mu}(X_j)}\left\{\varepsilon_j+\mu(X_j)\left[b_n(x)-b_n(X_j)\right]\right\}\right\}=0 \tag{3.32}$$

Analyzing the third term of equation (3.26)

$$E\left\{\sum_{j=1}^{n} w_j(x;h)\frac{\bar{\mu}_n(x)}{\bar{\mu}(X_j)}\varepsilon_j\left[b_n(x)-b_n(X_j)\right]\right\}=E\left\{\sum_{j=1}^{n} w_j(x;h)\left[\frac{\bar{\mu}_n(x)}{\bar{\mu}(X_j)}\varepsilon_j\left\{\frac{\tilde{\mu}_n(x)-\bar{\mu}_n(x)}{\bar{\mu}_n(x)}\right\}-\frac{\bar{\mu}_n(x)}{\bar{\mu}(X_j)}\varepsilon_j\left\{\frac{\tilde{\mu}_n(x_j)-\bar{\mu}(X_j)}{\bar{\mu}(X_j)}\right\}\right]\right\} \tag{3.33}$$

$$E\left\{\sum_{j=1}^{n} w_j(x;h)\frac{\bar{\mu}_n(x)}{\bar{\mu}(X_j)}\varepsilon_j\left[b_n(x)-b_n\left(X_j\right)\right]\right\} = \sum_{j=1}^{n} w_j(x;h)\frac{\bar{\mu}_n(x)}{\bar{\mu}(X_j)}\left\{\frac{\tilde{\mu}_n(x)-\bar{\mu}_n(x)}{\bar{\mu}_n(x)}\right\}E\left[\varepsilon_j\right] - \sum_{j=1}^{n} w_j(x;h)\frac{\bar{\mu}_n(x)}{\bar{\mu}(X_j)}\left\{\frac{\tilde{\mu}_n(x_j)-\bar{\mu}(X_j)}{\bar{\mu}(X_j)}\right\}E\left[\varepsilon_j\right] \tag{3.34}$$

$$E\left\{\sum_{j=1}^{n} w_j(x;h)\frac{\bar{\mu}_n(x)}{\bar{\mu}(X_j)}\varepsilon_j\left[b_n(x)-b_n\left(X_j\right)\right]\right\} = 0 \tag{3.35}$$

Therefore equation (3.26) reduces to

$$E[\hat{\mu}_n(x_i)] = \sum_{j=1}^{n} w_j(x;h)\frac{\bar{\mu}_n(x)}{\bar{\mu}(X_j)}\mu\left(X_j\right) + O_P\left(\frac{1}{nh}\right) \tag{3.36}$$

This means that $E\left[\hat{T}_{MBC}\right]$ will be given by the expression

$$E\left[\hat{T}_{MBC}\right] = \sum_{i\in s}\overline{Y_{\imath}} + \sum_{i\in p-s}\left\{\sum_{j=1}^{n} w_j(x;h)\frac{\bar{\mu}_n(x)}{\bar{\mu}(X_j)}\mu\left(X_j\right)\right\} + O_P\left(\frac{1}{nh}\right) \tag{3.37}$$

Equation (3.36) can be simplified by considering a Taylor's series expansion of the ratio $\frac{\mu(X_j)}{\bar{\mu}(X_j)}$ about the point x.

This is done as follows

$$\frac{\mu(X_j)}{\bar{\mu}(X_j)} = \frac{\mu(x)}{\bar{\mu}_n(x)} + \left(X_j - x\right)\left(\frac{\mu(x)}{\bar{\mu}_n(x)}\right)' + \frac{1}{2}\left(X_j - x\right)^2\left(\frac{\mu(x)}{\bar{\mu}_n(x)}\right)'' + \left(1 + O_P(1)\right) \tag{3.38}$$

Using equation (3.38) in equation (3.37) yields

$$E\left[\hat{T}_{MBC}\right] = \sum_{i\in s}\overline{Y_{\imath}} + \sum_{i\in p-s}\left\{\sum_{j=1}^{n} w_j(x;h)\bar{\mu}_n(x)\left\{\frac{\mu(x)}{\bar{\mu}_n(x)} + \left(X_j - x\right)\left(\frac{\mu(x)}{\bar{\mu}_n(x)}\right)' + \frac{1}{2}\left(X_j - x\right)^2\left(\frac{\mu(x)}{\bar{\mu}_n(x)}\right)'' + \left(1 + O_P(1)\right)\right\}\right\} + O_P\left(\frac{1}{nh}\right) \tag{3.39}$$

Considering the first two terms of the Taylor's series expansion, equation (3.39) reduces to

$$E\left[\hat{T}_{MBC}\right] = \sum_{i\in s} \overline{Y}_\imath + \sum_{i\in p-s}\left\{\sum_{j=1}^{n} w_j(x;h)\bar{\mu}_n(x)\left\{\frac{\mu(x)}{\bar{\mu}_n(x)} + \left(X_j - x\right)\left(\frac{\mu(x)}{\bar{\mu}_n(x)}\right)'\right\}\right\} + O_P\left(\frac{1}{nh}\right) \tag{3.40}$$

It is easy to show that

$\sum_{j=1}^{n} w_j(x;h) = 1$ and $\sum_{j=1}^{n}\left(X_j - x\right)w_j(x;h) = 0$. Therefore equation (3.40) can be written as

$$E\left[\hat{T}_{MBC}\right] = \sum_{i\in s} \overline{Y}_\imath + \sum_{i\in p-s}\left\{\sum_{j=1}^{n} w_j(x;h)\mu(x)\right\} + O_P\left(\frac{1}{nh}\right) \tag{3.41}$$

From equation (3.3) we have

$$T = \sum_{i\in s} Y_i + \sum_{i\epsilon p-s} Y_i$$

$$E[T] = E\left\{\sum_{i\in s} Y_i + \sum_{i\epsilon p-s} Y_i\right\} \tag{3.42}$$

$$E[T] = \sum_{i\in s} \overline{Y}_\imath + \sum_{i\epsilon p-s} \mu(x) \tag{3.43}$$

Substituting equations (3.41) and (3.43) back into equation (3.23) yields

$$E\left[\hat{T}_{MBC} - T\right] = \sum_{i\in s} \overline{Y}_\imath + \sum_{i\in p-s}\left\{\sum_{j=1}^{n} w_j(x;h)\mu(x)\right\} + O_P\left(\frac{1}{nh}\right) - \left[\sum_{i\in s} \overline{Y}_\imath + \sum_{i\epsilon p-s} \mu(x)\right] \tag{3.44}$$

$$E\left[\hat{T}_{MBC} - T\right] = \sum_{i\in p-s}\left\{\sum_{j=1}^{n} w_j(x;h)\mu(x)\right\} - \sum_{i\epsilon p-s} \mu(x) + O_P\left(\frac{1}{nh}\right) \tag{3.45}$$

Hence the bias of $\hat{T}_{MBC}$ is given by

$$Bias\left[\frac{\hat{T}_{MBC}}{N}\right] = E\left[\frac{\hat{T}_{MBC}-T}{N}\right] = \frac{1}{N}\left\{\sum_{i\in p-s}\left\{\sum_{j=1}^{n} w_j(x;h)\mu(x)\right\} - \sum_{i\epsilon p-s}\mu(x)\right\} + O_P\left(\frac{1}{nh}\right)$$

(3.46)

The bias of $\hat{T}_{MBC}$ will be of order $O_P\left(\frac{1}{nh}\right)$. Thus it converges to zero at a faster rate compared to the existing non-parametric estimators which generally converge at the rate $O_P(h^2)$.

3.2 Asymptotic Variance of the Proposed Estimator

Using equation (3.21), the estimator of finite population total is given by

$$\hat{T}_{MBC} = \sum_{i\in s} Y_i + \sum_{i\in p-s}\Bigg\{\sum_{j=1}^{n} w_j(x;h)\left\{\frac{\bar{\mu}_n(x)}{\bar{\mu}(X_j)} \times \left[\mu(X_j) + \varepsilon_j\right]\left[1 + b_n(x) - b_n(X_j) + r_j(x, X_j)\right]\right\}\Bigg\} \qquad (3.47)$$

where $r_j(x, X_j)$ is the remainder term that involves the terms x and X_j.

Using the assumption $nh \to \infty$, the remainder terms converge to zero in probability. Therefore $r_j(x, X_j)\left[\mu(X_j) + \varepsilon_j\right] = O_P\left(\frac{1}{nh}\right)$ and equation (3.47) reduces to

$$\hat{T}_{MBC} = \sum_{i\in s} Y_i + \sum_{i\in p-s}\Bigg\{\sum_{j=1}^{n} w_j(x;h)\left\{\frac{\bar{\mu}_n(x)}{\bar{\mu}(X_j)} \times \left[\mu(X_j) + \varepsilon_j\right]\left[1 + b_n(x) - b_n(X_j)\right]\right\}\Bigg\} + O_P\left(\frac{1}{nh}\right) \qquad (3.48)$$

Truncating the binomial expansion at the first term yields

$$\hat{T}_{MBC} = \sum_{i\in s} Y_i + \sum_{i\in p-s}\left\{\sum_{j=1}^{n} w_j(x;h)\left\{\frac{\bar{\mu}_n(x)}{\bar{\mu}(X_j)}\times\left[\mu(X_j)+\varepsilon_j\right]\right\}\right\} + O_P\left(\frac{1}{nh}\right) \qquad (3.49)$$

The variance of the estimator is then defined by

$$Var\left[\hat{T}_{MBC}\right] = Var\Bigg\{\sum_{i\in s} Y_i + \sum_{i\in p-s}\Bigg[\sum_{j=1}^{n} w_j(x;h)\Bigg\{\frac{\bar{\mu}_n(x)}{\bar{\mu}(X_j)}\times\left[\mu(X_j)+\varepsilon_j\right] + O_P\left(\frac{1}{nh}\right)\Bigg\}\Bigg]\Bigg]\Bigg\}$$

(3.50) $$Var\left[\hat{T}_{MBC}\right] = Var\left\{\sum_{i\in s} Y_i\right\} + Var\left[\sum_{i\in p-s}\left\{\sum_{j=1}^{n} w_j(x;h)\left\{\frac{\bar{\mu}_n(x)}{\bar{\mu}(X_j)}\times Y_j\right\}\right\}\right] + O_P\left(\frac{1}{n^2h^2}\right) \qquad (3.51)$$

$$Var\left[\hat{T}_{MBC}\right] = Var\left\{\sum_{i\in s} Y_i\right\} + \sum_{i\in p-s}\left[\sum_{j=1}^{n} w_j(x;h)\frac{\bar{\mu}_n(x)}{\bar{\mu}(X_j)}Var(Y_j)\right] + O_P\left(\frac{1}{n^2h^2}\right)$$

(3.52)

$$Var\left[\hat{T}_{MBC}\right] = \sum_{i\in s}\sigma^2(x_i) + \sum_{i\in p-s}\sum_{j=1}^{n} w_j(x;h)^2\left\{\frac{\bar{\mu}_n(x)}{\bar{\mu}(X_j)}\right\}^2\sigma^2(x_i) + O_P\left(\frac{1}{n^2h^2}\right)$$

(3.53)

Obtaining the Taylor’s series expansion of the ratio $\frac{\sigma^2(x_i)}{\bar{\mu}(X_j)^2}$ in the second part of equation (3.53) gives

$$Var\left[\hat{T}_{MBC}\right] = \sum_{i\in s}\sigma^2(x_i) + \sum_{i\in p-s}\sum_{j=1}^{n} w_j(x;h)^2\,\sigma^2(x_i) + O_P\left(\frac{1}{n^2h^2}\right) \qquad (3.54)$$

This means that the asymptotic variance of $\left[\frac{\hat{T}_{MBC}}{N}\right]$ will be given by

$$Var\left[\frac{\hat{T}_{MBC}}{N}\right] = \frac{1}{N^2}\sum_{i\in s}\sigma^2(x_i) + \frac{1}{N^2}\left[\sum_{i\in p-s}\sum_{j=1}^{n} w_j(x;h)^2\,\sigma^2(x_i)\right] + O_P\left(\frac{1}{n^2h^2}\right) \qquad (3.55)$$

This implies that $\hat{T}_{MBC}$ is more efficient than the usual non-parametric regression estimator proposed by Dorfman (1992).

3.3 Asymptotic Mean Squared Error

The mean squared error of $\hat{T}_{MBC}$ is given by

$$\text{MSE}\left[\hat{T}_{MBC}\right] = Var\left[\frac{\hat{T}_{MBC}}{N}\right] + \left[Bias\left[\frac{\hat{T}_{MBC}}{N}\right]\right]^2 \qquad (3.56)$$

$$Bias\left[\frac{\hat{T}_{MBC}}{N}\right] = E\left[\frac{\hat{T}_{MBC}-T}{N}\right] = O_P\left(\frac{1}{nh}\right) \qquad (3.57)$$

Using equations (3.55) and (3.57) in equation (3.56) yields

$$\text{MSE}\left[\hat{T}_{MBC}\right] = \frac{1}{N^2}\left[\sum_{i\in s}\sigma^2(x_i) + \sum_{i\in p-s}\sum_{j=1}^{n} w_j(x;h)^2\,\sigma^2(x_i)\right] + O_P\left(\frac{1}{n^2h^2}\right) + \left(O_P\left(\frac{1}{nh}\right)\right)^2 \qquad (3.58)$$

$$\text{MSE}\left[\hat{T}_{MBC}\right] = \frac{1}{N^2}\left[\sum_{i\in s}\sigma^2(x_i) + \sum_{i\in p-s}\sum_{j=1}^{n} w_j(x;h)^2\,\sigma^2(x_i)\right] + O_P\left(\frac{1}{n^2h^2}\right) \qquad (3.59)$$

As $n \to \infty$ and $h \to 0$, the mean squared error in equation (3.58) tends to zero, that is, $\text{MSE}\left[\hat{T}_{MBC}\right] \to 0$.

This shows that the estimator $\hat{T}_{MBC}$ is statistically consistent and therefore useful.

CHAPTER FOUR

EMPIRICAL STUDY

4.1 Description of the Population

In this chapter, the theory developed in the previous chapter is tested using simulated data. The estimation of the population total and the corresponding mean squared error will be carried out using two sets of data, namely linear and quadratic, that make use of simulated data. The analysis and comparison on the performance of the estimates will be based on the ratio, Nadaraya-Watson and the Multiplicative Bias Corrected estimators.

The description of the set of data for the populations is summarized in Table 4.1. The auxiliary variable for each data set has been collected and incorporated in the estimators so as to improve on the precision of the estimation since the auxiliary variable is assumed to contain important information that is necessary for the estimation of the population total.

Table 4.1: Characteristics of Data Set Used

Population description	**X**	**Y**
Linear	$X_i \sim u(0,1), e_i \sim N(0,1)$	$Y_i = 2x_i + e_i$
Quadratic	$X_i \sim u(0,1), e_i \sim N(0,1)$	$Y_i = 1 + 2(x_i - 0.5)^2 + e_i$

The data set are artificial data that were obtained by simulation using user-designed computer program. Each data set is described below

The first data set was obtained through simulation by use of a linear model which has the relation

$Y_i = 2x_i + e_i$ (4.1)

The random variable X is simulated using a rectangular distribution that takes the values that are equally likely from 0 to 1 inclusive. It is assumed that (x_i, y_i), $i = 1,2, \dots, N$ are independent and identically distributed random variables. The error term e_i is a standard normal variable defined as $e_i \sim N(0,1)$.

The second data set was obtained through simulation by use of a quadratic model which has the relation

$Y_i = 1 + 2(x_i - 0.5)^2$ for $i = 1,2, \dots, N$

(4.2)

The random variable X is simulated using a rectangular distribution that takes the values that are equally likely from 0 to 1 inclusive. It is assumed that (x_i, y_i), $i = 1,2, \dots, N$ are independent and identically distributed random variables. The error term e_i is a standard normal variable defined as $e_i \sim N(0,1)$.

In all variables, 500 simple random samples without replacement of size n=250 were selected. In each selected sample, the estimate of the population total and the estimate of the mean squared error are computed.

4.2 Unconditional Properties for the Parametric and Nonparametric Estimators

The estimates of the bias and the mean squared error of the finite population total for the ratio, Nadaraya-Watson and Multiplicative Corrected estimators are recorded, analyzed and conclusions are made. In each variable, 500 simple random samples without replacement of size n=250 were selected and unconditional results for the estimators were computed and analyzed.

In the study, a population of size $N = 1000$ was simulated using a computer program. Five hundred samples of size 250 were generated using simple random sampling. The Epanechnkov kernel was used in the smoothing process. A comparison between the multiplicative bias corrected estimator denoted by $\hat{T}_{MBC}$, that proposed by Dorfman (1992) which is also denoted by $\hat{T}_{NW}$ and the ratio estimator which is denoted by $\hat{T}_R$ was done. The biases were computed as $(\hat{T}_{MBC} - \mathrm{Y})$, $(\hat{T}_{NW} - Y)$ and $(\hat{T}_R - Y)$ respectively. Root Mean Squared Errors (RMSE) were also computed for each estimator where $(\hat{T}_{MBC} - \mathrm{Y}) = \sqrt{\frac{1}{500}\sum_{s\in S=500}(\hat{T}_{MBC} - \mathrm{Y})^2}$, $(\hat{T}_{NW} - \mathrm{Y}) = \sqrt{\frac{1}{500}\sum_{s\in S=500}(\hat{T}_{NW} - \mathrm{Y})^2}$ and $(\hat{T}_R - \mathrm{Y}) = \sqrt{\frac{1}{500}\sum_{s\in S=500}(\hat{T}_R - \mathrm{Y})^2}$ respectively.

Table 4.2 presents the unconditional biases and Root Mean Squared Errors (RMSE) for the multiplicative bias corrected estimator denoted by $\hat{T}_{MBC}$, that proposed by Dorfman (1992) and the ratio estimator.

Table 4.2: Unconditional Biases and RMSE

Mean function	$\hat{T}_{MBC}$		$\hat{T}_{NW}$		$\hat{T}_{R}$	
	Bias	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**
Linear	2037.44	5220.837	2856.32	6908.85	1125.67	3968.14
Quadratic	850.53	1008.5954	2444.83	3008.41	4230.605	7635.81

4.3 Conditional Properties for the Parametric and Nonparametric Estimators

Further, the samples were grouped into groups of 20 so that there were 25 groups. For each group $\bar{\bar{X}} = \frac{1}{50}\sum_{i=1}^{20}\bar{x}_i$ was computed. $\hat{\bar{T}}_{MBC} = \frac{1}{50}\sum_{i=1}^{20}\hat{T}_{MBC.i}$ was also computed. The conditional bias for each group was computed as $\hat{\bar{T}}_{MBC} - \bar{Y}$ where $\bar{Y}$ is the population mean for the survey measurements and $\bar{x}_i$ is the sample mean for the auxiliary variables.

The graphs below illustrate the behaviour of the conditional bias for each estimator when various mean functions were used. The figure 2 shows the conditional bias when linear mean functions was used and figure 2 shows the conditional bias when a quadratic mean function was used.

From figure 4.1, the ratio estimator performed well when a linear mean function was used. This is mainly because the ratio estimator is the Best Linear Unbiased Estimator (BLUE). It can be observed that biases to the left of the population mean of the auxiliary variable, $\bar{X} = 0.5049$, are large but they systematically reduce towards the right. It can also be noted that at the point $\bar{X} = 0.5049$, the bias associated with the ratio estimator is negligible. This is owed to the fact that the bias of the ratio estimator is very minimal when we have a balanced sample.

In figure 4.2, the quadratic mean function was used, the proposed estimator gives better estimates of the population total compared to those realized using the estimator proposed by Dorfman (1992) and the ratio estimator. It can be observed that biases to the left of the population mean of the auxiliary variable, $\bar{X} = 0.5085$, are large but they systematically reduce towards the right.

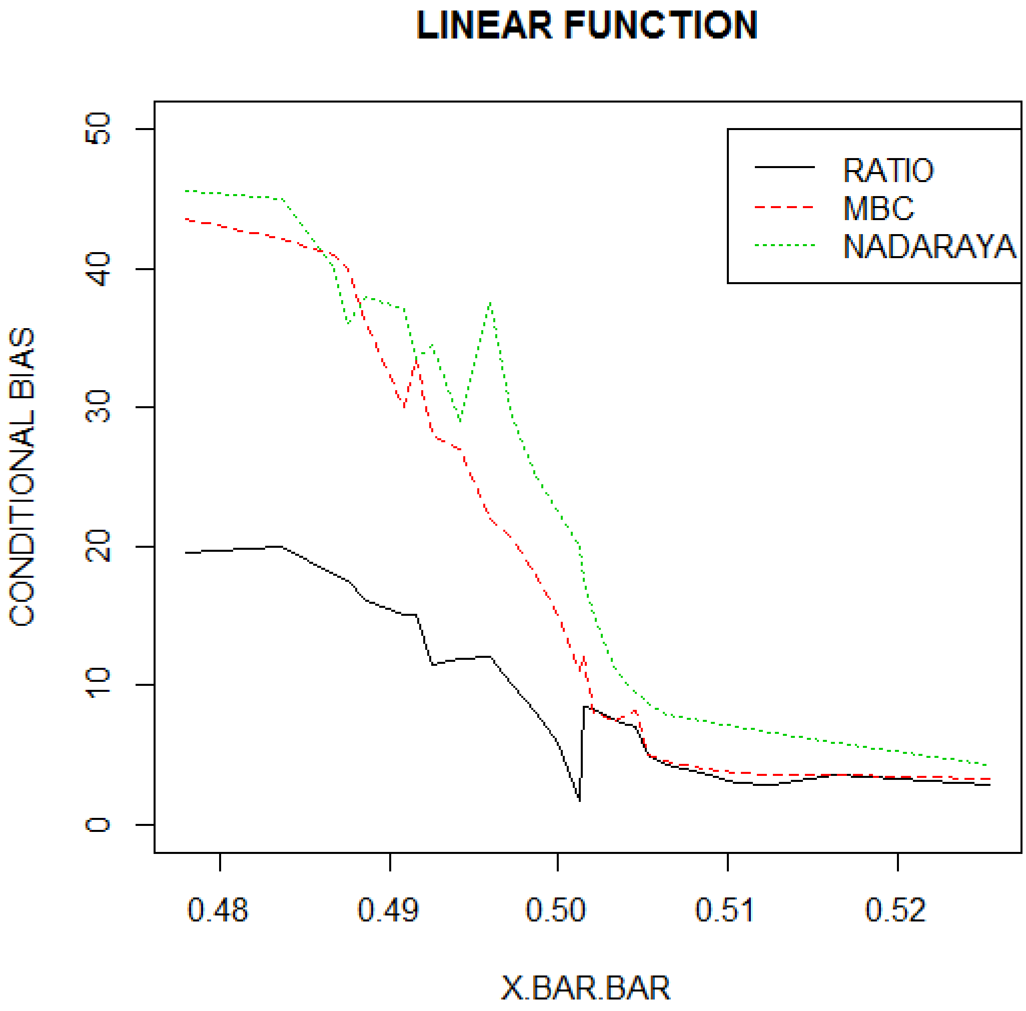

Figure 4.1: Conditional bias using a linear data set

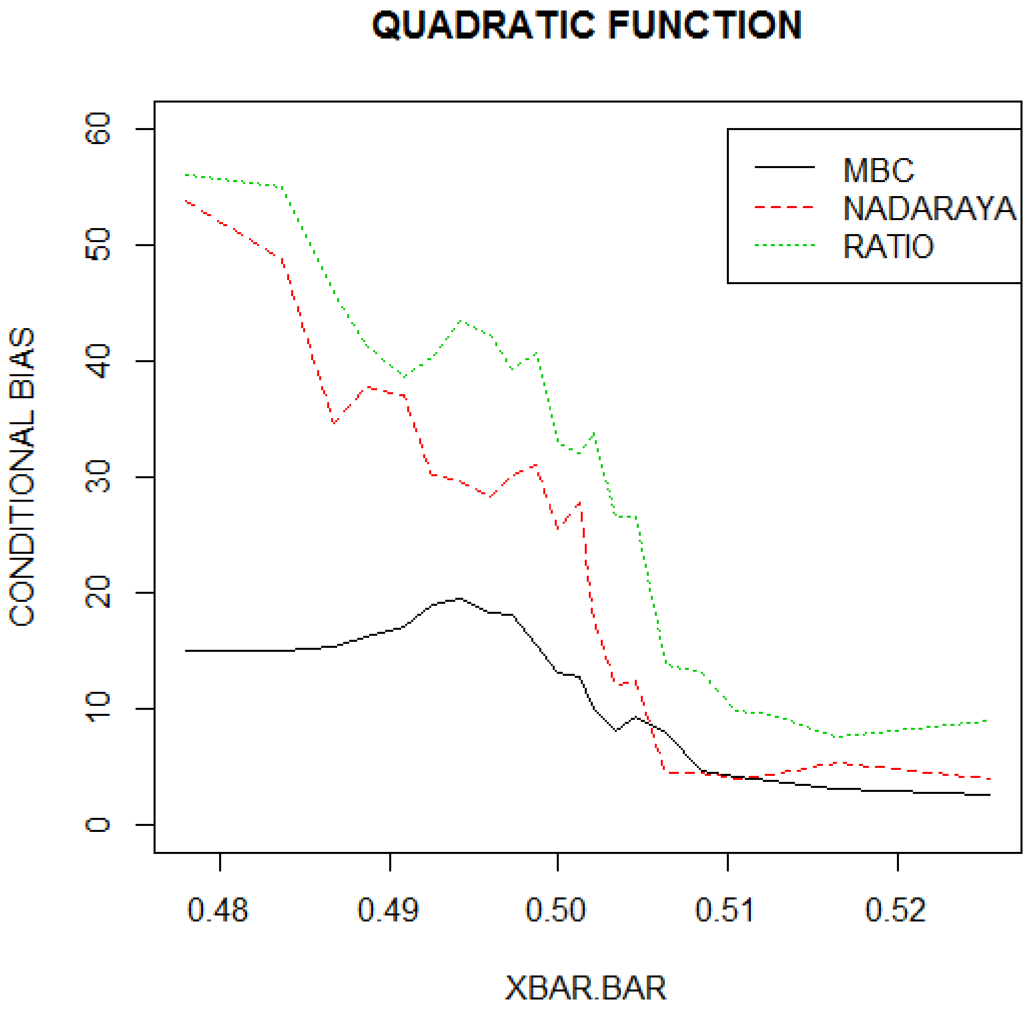

Figure 4.2: Conditional bias using a quadratic data set

CHAPTER FIVE

CONCLUSION AND RECOMMENDATION

5.1 Conclusion

The main objective of this study was to obtain a consistent estimator of finite population total using the multiplicative bias correction technique. As a way of achieving this, a pilot smoother was utilized and the resulting nonparametric estimator was found to be a useful tool in the correction of boundary bias. The methodology used possesses a kind of robustness in the sense that the multiplicative factor $\widetilde{m}_n(x)$ is bounded. The method is easy to implement and has good asymptotic properties both theoretically and practically.

5.2 Recommendation

In this study, a single auxiliary variable was considered. The use of more than one auxiliary variable ought to be investigated and the performance of the resulting estimator be compared to determine if it yields better estimation of finite population total.

Independence of survey variables y_i and y_j was assumed in the study of asymptotic properties of the estimator derived in the previous chapter. The investigation on the nature of the results if dependence of the observations is still an open area for extension of this problem of estimating the finite population total.

REFERENCES

Bierens, H. J. (1987), Kernel estimators of regression functions, in: Bewley, ed., Advances in Econometrics, (Cambridge University Press, Cambridge).

Breidt, F.J. and Opsomer, J.D. (2000), Local polynomial regression estimators in survey sampling, The Annals of Statistics, Vol.28, No.4, pp.1026-1053.

Cochran, W.G. (1977), Sampling techniques, Third edition, New York: John Wiley and Sons.

Dorfman, A.H. (1992), Nonparametric regression for estimating totals in finite populations, Proceedings of the section on Survey Research Methods, American Statistical Association, pp.622-625.

Di Marzio, M. and Taylor, C. (2008), On boosting kernel regression, Journal of Statistical Planning and Inference, Vol. 138, No. 8, pp. 2483-2498.

Dorfman, A.H. and Hall, P. (1993), Estimators of the finite population distribution function using nonparametric regression, The Annals of Statistics, Vol.21, No.3, pp.1452-1475.

Fan, J. (1992), Design-adaptive nonparametric regression, Journal of the American Statistical Association, Vol.87, No.420, pp.998-1004.

Glad, I.K (1998), Parametrically guided non-parametric regression, Scandinavian Journal of Statistics, Vol.25, pp. 649-668.

H$\ddot{a}$rdle, W. (1986), A note on jackknifing kernel regression function estimators, IEEE Trans. Inf. Theory, Vol. 32, pp.298-300.

Hirukawa, M. (2010), Nonparametric multiplicative bias correction for kernel-type density estimation on the unit interval, Computational Statistics and Data analysis, Vol.54, pp. 473-495.

Jones, M. C., Linton, O. and Nielsen, J. P. (1995), A simple bias reduction method for kernel density estimation, Biometrika, Vol.82, pp.327-338.

Jones, M. C., Signorini, D. F. and Hjort, N. L. (1999), On multiplicative bias correction in kernel density estimation, The Indian Journal of Statistics, Vol.61, Series A, Pt.3, pp. 422-430.

Linton, O. and Nielsen, J. P. (1994), A multiplicative bias reduction method for nonparametric regression, Statistics & Probability Letters, Vol.19, pp. 181-187.

Mageto, T. (2008), Robust estimation of finite population total using local polynomial regression, PhD Thesis, Jomo Kenyatta University of Agriculture and Technology.

M$\ddot{u}$ller, H. G. and U. Stadm$\ddot{u}$ller (1987a), Variable bandwidth kernel estimators of regression curves, Annals of Statistics, Vol.15, pp.610-625.

Odhiambo, R. and Mwalili, T. (2000), Nonparametric regression method for estimating the error variance in unistage sampling, East African Journal of Science, Vol.2, No.2, pp.107-112.

Odhiambo, R. (1995), Robust variance estimation for finite population sampling. Thesis, Kenyatta University.

Ombui, T.M. (2008), Robust estimation of finite population total using local polynomial regression. Thesis, Jomo Kenyatta University of Agriculture and Technology.

Opsomer, J.D., Moisen G.G. and Kim J.Y. (2001), Model-assisted estimation of forest resources with generalized additive models, Proceedings of the Annual Meeting of the American Statistical Journal, August 5-9.

Priestly, M.B. and Chao, M.T. (1972), Nonparametric function fitting, Journal of the Royal Statistical Society, B34, pp.384-392.

Zeng, H. and Little, R. (2003), Inference for the population total from probability-proportional-to-size samples based on predictions from penalized spline nonparametric model. The Berkely Electronic Press.

APPENDIX

Appendix I: Multiplicative Bias Correction Simulation

```
set.seed(123)

x=runif(1000,0,1)

x1=sample(x,size=20,replace=FALSE,prob=NULL)

x1bar=mean(x1)

x1bar

y=rnorm(1000,0,1)

sumy=sum(y)

sumy

ybar=mean(y)

ybar

y1=sample(y,size=980,replace=FALSE,prob=NULL)

y1sum=sum(y1)

j=20

u=seq(-1,1,2/979)
```

```
ku=0.75*(1-u^2)

ku

c=sum(ku)

c

w=ku/c

w

sumw=sum(w)

sumw

m=sum(w*y1)

m[1]=w[1]*y1[1]

for(i in 1:980)

{

m=sum(w*y1[i])

}

m

outsample=j*m
```

```
T=y1sum+outsample

T

Taverage=(1/1000)*T

Taverage

unconditionalbias=Taverage-ybar

unconditionalbias
```

Appendix II: Nadaraya Watson Estimator Simulation

```
u=seq(-1,1,2/(M-1))

ku=0.75*(1-u^2)

ku

c=sum(ku)

c

w=ku/c

w

sumw=sum(w)

sumw

m=sum(w*y1)
```

```
m[1]=w[1]*y1[1]
for(i in 1:M)
{
m=sum(w*y1[i])
}
m
outsample=j*m
Tnw=y1sum+outsample
Tnw
```

Appendix III: Linear Regression for Conditional Bias

```
conditional_bias1=c(19.5,20,18,17.5,16,15.09,15,11.5,11.9,12.09,10,8.01,5.88,1.69,8.5,
8.3,7.5,7,5,4.2,3.7,3,2.8,3.5,2.9)
length(conditional_bias1)
conditional_bias2=c(43.5,42.1,41,40,36,30,33.5,28,27,22,20.5,18,15,11,12,8,7.5,8.09,5,
4.5,4.0,3.75,3.6,3.5,3.2)
length(conditional_bias2)
```

```
conditional_bias3=c(45.5,45,40,36,38,37,33.5,34.5,29,37.5,29.3,25,22.5,20,17.5,15,11,
9.5,8.76,7.8,7.5,7,6.5,5.9,4.2)

length(conditional_bias3)

xbar.bar=c(0.4780,0.4836,0.4867,0.4875,0.4886,0.4908,0.4916,0.4925,0.4942,0.4959,0.
4972,0.4987,0.4999,0.5012,0.5015,0.5021,0.5034,0.5046,0.5053,0.5063,0.5085,0.5105,
0.5126,0.5164,0.5255)

length(xbar.bar)

plot(xbar.bar,conditional_bias1,type="l",col="1",lty=1,ylim=c(0,50),xlab="X.BAR.BA
R",ylab="CONDITIONAL BIAS",main="LINEAR FUNCTION")

lines(xbar.bar,conditional_bias2,type="l",col="2",lty=2)

lines(xbar.bar,conditional_bias3,type="l",col="3",lty=3)

legend(0.51,50,c("RATIO","MBC","NADARAYA"),col=c(1,2,3),lty=c(1,2,3))
```

Appendix IV: Quadratic Regression for Conditional Bias

```
conditional_bias1=c(14.9,15,15.32,16.09,16.95,18.88,19.5,18.32,18,15.5,13,12.71,9.93,
8.02,9.29,7.8,4.63,4.11,3.76,3.02,2.51)

conditional_bias2=c(53.8,48.8,34.6,37.8,37,30.2,29.7,28.2,30.1,31.1,25.5,27.8,17.8,12.
1,12.3,4.4,4.5,3.9,4.32,5.33,4)

length(conditional_bias2)
```

```
conditional_bias3=c(56,55,45.8,41.3,38.7,40.3,43.4,42.2,39.3,40.7,33,32.1,33.8,26.5,26.6,13.8,13.1,9.8,9.5,7.6,8.9)

length(conditional_bias3)

xbar.bar=c(0.4780,0.4836,0.4867,0.4886,0.4908,0.4925,0.4942,0.4959,0.4972,0.4987,0.4999,0.5012,0.5021,0.5034,0.5046,0.5063,0.5085,0.5105,0.5126,0.5164,0.5255)

plot(xbar.bar,conditional_bias1,type="l",col="1",lty=1,ylim=c(0,60),xlab="XBAR.BAR",ylab="CONDITIONAL BIAS",main="QUADRATIC FUNCTION")

lines(xbar.bar,conditional_bias2,type="l",col="2",lty=2)

lines(xbar.bar,conditional_bias3,type="l",col="3",lty=3)

legend(0.51,60,c("MBC","NADARAYA","RATIO"),col=c(1,2,3),lty=c(1,2,3))
```