ESTIMATION OF THE POPULATION VARIANCE USING A SMOOTHING OPERATOR UNDER SIMPLE RANDOM SAMPLING

Lavender Akoth Odhiambo

(BSC Actuarial Science)

I56/PT/CTY/27344/2013

Department of Mathematics and Actuarial Sciences

A Project Submitted in Partial Fulfillment of the Requirements for the Award of the Degree of Master of Science (Statistics) in the School of Pure and Applied Sciences of Kenyatta University

June 2019

Declaration

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

Signature..... Date.....

Lavender Akoth Odhiambo

This project has been submitted for examination with my approval as the university supervisor.

Signature..... Date.....

Dr. Christopher Ouma Onyango

Kenyatta University

This project has been submitted for examination with my approval as the university supervisor.

Signature..... Date.....

Prof.Romanus Odhiambo Otieno

Meru University of Science and Technology

Dedication

I wish to dedicate this work to my loving parents, Chris Rusana and Grace Atieno for their moral and financial support.

Acknowledgment

I wish to acknowledge the project supervisors Dr.Christopher Ouma Onyango and Prof.Romanus Odhiambo Otieno who have greatly contributed to the completion of this project. I greatly thank them for sparing their time to consistently offer unlimited support and encouragement in the field of sample survey.

Table of Contents

List of Tables

List of Figures

List of Abbreviations and Acronyms

s	Sample proportion
r	Non sample proportion
$p-s$	Non sample space
i	Sample element
j	Non sample element
n	Sample size
N	Population size
x	Auxiliary variable
y	Study variable
f	Sampling fraction
e_i	Error term
p	population size
$k(t)$	Kernel function
Var	Variance
iid	Independent and identically distributed
BLUE	Best Linear Unbiased Estimator
SRSWR	Simple random sampling without replacement
MSE	Mean Squared Error
RMSE	Root Mean Squared Error
MBC	Multiplicative Bias Correction
UN	United nations
HDI	Human Development Index
GNI	Gross National Income

Abstract

Variance estimation has been a major concern in sample survey theory. The problem in estimation theory is to determine estimators that have smaller variance under a given model specification. However, existing variance estimators suffer from boundary problems and outlier sensitivity. To address this, a robust variance estimate of the ratio estimator of the population mean using a multiplicative bias correction technique under model based approach is considered. Asymptotic properties of the robust variance estimator are investigated. Also a comparative study of the existing variance estimators and the derived robust variance estimator of the population mean is studied. The results of the study show that under mild assumption, the derived variance estimator of the population mean is asymptotically more consistent and has a better coverage probability as compared to rival variance estimators of the population mean.

Chapter 1

Introduction

This chapter presents the background of the study, problem statement, objectives of the study and the significance of the study.

1.1 Background of the Study

The main aim in estimation theory is using sample data to estimate parameters of the population. Experimenters are always interested in using methods which improves precision of population parameters estimates. These parameters can be variance of the population, proportions and means of some characters of study.

Auxiliary information can be applied at both the estimation and selection stages to improve the design and achieve more efficient estimators. For instance, increased precision can be obtained when study variable Y is highly correlated with the auxiliary variable X.

Ratio estimation considers the correlation between the auxiliary variable X and the study variable Y. Whenever positively correlated information is available on the auxiliary variable and the study variable, the ratio estimator is the most suitable for estimating the population variance. For ratio estimators in sampling theory, population information of the auxiliary variable such as the coefficient of variation or the coefficient of kurtosis is often used to increase the efficiency of the estimation for the population variance. It therefore follows that a model based approach can be used to increase 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 and study variable.

Variance estimation methods that have been used in the past include the Taylor expansion, the Jacknife approach, balance repeated replication and the bootstrap technique. Taylor linearization is a method of variance estimation for statistics such as ratio, regression estimators and logistic regression coefficient estimators. This

method is applicable to any sampling design and is computationally simple. However, this method can lead to multiple variance estimators that are asymptotically design unbiased under repeated sampling.

Quenouille (1949) proposed the Jacknife procedure for correcting bias and was later refined and given its current name by Tukey (1956) for constructing confidence limits for a large class of estimators. It is similar to the bootstrap in that it involves resampling, but instead of sampling with replacement, the method samples without replacement. The Jackknife estimator of the standard error is roughly equivalent to the delta method for large samples. Wolter (1985) deduced that both the the Taylor linearization and the repeated replication methods produce a consistent but a biased estimator of the population variance. An advantage of the balanced repeated replication over Taylor linearisation as shown by Krewski and Rao (1981) is that it asymptotically holds as the number of strata increases.

Linton and Nielsen (1994) developed the multiplicative bias correction technique that assures a positive estimate and reduces the bias of the estimate with negligible increase in variance. Burr et al (2010) showed that the multiplicative bias correction approach reduces bias with insignificant increase in variance. Stephane et al (2017) in their study found out that the multiplicative bias correction technique was statistically consistent and asymptotically unbiased.

Applying a model based design to estimation is motivated by the fact that it provides a flexible way of studying the relationship between variables and also results in good estimators thus increasing their efficiency as compared to estimators obtained using design based approaches.

The main concern in this project is to apply a multiplicative bias correction technique to find a robust variance estimator of the population mean under simple random sampling. A robust variance estimator of the population mean is therefore derived with the aid of a superpopulation model. A methodology to elaborate the properties of the proposed estimator is also studied.

1.2 Problem Statement

Ratio estimation has been extensively used in sample survey because of its computational simplicity and intuitive appeal in examining the relationship between a study variable and an auxiliary variable. Ratio estimation is traditionally preferred to regression estimation because of its ease in handling large data sets unlike design based, randomization and non parametric methods of estimation. Authors have in the past

applied ratio estimation to derive efficient ratio-type estimators of the population variance by modifying the structure of existing estimators such as modified estimators of population variance using values of coefficient of variation, coefficient of kurtosis, coefficient of skewness of an auxiliary variable together with their biases and mean squared errors. Moreover, the value of quartiles and their functions are unaffected by the extreme values or the presence of outliers in the population values. For this reason, some considered the problem of estimating the population variance of the study variable using information on variance, quartiles, inter-quartile range, semi-quartile range and semi-quartile average of an auxiliary variable. The existing methods of estimating population parameters have shortcomings such as bias-variance trade off along the boundary points. This study seeks to develop a robust variance estimator of the population mean using a multiplicative bias correction approach that solves boundary problems.

1.3 Objectives of the Study

1.3.1 General Objective

To construct a robust variance estimator of the ratio estimator of the population mean using a multiplicative bias correction procedure under simple random sampling.

1.3.2 Specific Objectives

1.To derive a robust variance estimator for the ratio estimator of the population mean.
2.To compare the relative efficiency of the derived estimator to that of Royall and Cumberland (1981) and Otieno and Mwalili (2000).

1.4 Significance of Study

Survey sampling is very important in statistics. It is through sample survey that researchers can estimate population parameters using samples extracted from the population. This project focuses on estimating a robust variance of the population mean using the multiplicative bias correction approach. Unlike the Taylor linearization, Jacknife, balance repeated replication and the bootstrap technique that produce a biased estimator of the population variance, the multiplicative bias correction technique performs better by producing a consistent and unbiased estimator of the population variance. The results obtained from application of the multiplicative bias

correction technique in analysis of health, education and manufacturing data are consistent and efficient. Thus the multiplicative bias correction technique can be highly recommended for use of policy implementation and planning in education, planning, health and manufacturing sectors of the Kenyan economy.

Chapter 2

Literature Review

2.1 Introduction

Previous work related to this study are reviewed in this chapter. The estimation of population variance is very important in sectors such as agriculture, industry, medical sciences and biology which have been facing the problem of evaluating a finite population variance. A reasonable comprehension of variability is essential for better results in different fields of study.

This project is concerned with estimation of a robust variance estimator of the population mean using a multiplicative bias correction technique. The issue of choosing an efficient variance estimator has still not been solved even in the simplest setting. Many estimators of the variance have been derived as given in Rao (1969), Royall and Cumberland (1981) and Otieno and Mwalili (2000). Most of the derived estimators of variance are design based and few ones such as Royall and Cumberland and Otieno and Mwalili (2000) are model-based. Under the design based approach, inference is made based on observed sample and an assumed super population model. The sampling design becomes irrelevant under the design based approach.

The comparison of the performance of estimators of variance is theoretically made by the assumption that the auxiliary variable X and study variable Y satisfy a superpopulation model. Sometimes the results are exact. Model robustness properties of variance estimators was studied by Royall and Eberhardt (1975). The study was restricted to the robust bias properties of variance estimators when the real parameters differ from the parameters assumed in the study. Royall and Cumberland (1981) studied the conditional properties of variance estimators as a function of the sample mean. Wu (1982) and Royall and Cumberland (1981) demonstrated how some variance estimators differ significantly over a range of the sampled values in a population.

Wu (1982) and Royall and Cumberland (1981) argued that the conditional mean and the conditional MSE of a variance estimator should be close. Measurements are used to estimate the population parameter of interest. The problem with this approach is that it assumes that all samples in the population are selected. This is not possible especially because of the problems associated with the selection of the samples.

2.1.1 Review of Estimation Methods

The main approaches used in the estimation of a robust variance are the design based approach, model-based or super-population approach, model assisted approach and design assisted approach

In the design based approach, the observed values of the survey variable Y given by $y_1, y_2, ..., y_n$ are assumed to be unknown but fixed constants. In this concept a sample is drawn from the finite population and the sample measurements are employed in the estimation of the population parameter of interest.

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

$$\hat{T} = \Sigma_{ies} Y_i + \Sigma_{ier} Y_i$$

where $\Sigma_{ies} Y_i$ denotes the sample proportion and $\Sigma_{ier} Y_i$ denotes the non sample proportion.

The model assisted approach incorporates auxiliary information into the design based estimation of the population variance. It assumes the existence of a superpopulation model between the auxiliary variables and variable of interest for the sampled population. 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 the design assisted approach, the model is used to increase the efficiency of the estimators. Estimators remain typically design consistent even if the model is not correct. Since this approach has a great potential to improve the precision of the required survey estimators when the appropriate auxiliary information is available, it

often requires that this models are linear. Of the survey approaches, the model based approach has been considered to be the most consistent method of estimation.

2.1.2 Nonparametric regression estimator of the population variance

Non parametric regression has been studied by Nadaraya and Watson (1964), Hardle (1991) and Otieno and Mwalili (2000).

Dorfman (1992) introduced a non parametric regression estimator for the finite population variance based on a sample drawn from the population. Taking into consideration a simple Nadaraya Watson, the estimator of the survey variable is estimated as;

$$y = m(x_i) + \sigma(x)e \tag{2.1}$$

where m(.) is a smoother function, x_i are the auxilliary random variables that are assumed to be known for the whole population and e_i is independently distributed with mean 0 and constant variance. When estimating m(x), one possibility is to average the nearby values of Y measured by the distance $|X_i - X|$. Let k(u) be a symmetric density function for example the standard normal function. For a chosen scaling factor also known as bandwidth b, define

$$k(u) = \frac{1}{b}k(\frac{u}{b}) \tag{2.2}$$

and weights

$$W_i = \frac{k_b(xi - x)}{\Sigma k_b(xi - x)} \tag{2.3}$$

The larger b is the more equal the weights. The Nadaraya-Watson estimator of $m(x)$ is

$$\hat{m_x} = \Sigma_{j=1}^{n} W_j(x) y_i \tag{2.4}$$

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

$K(t) \geq 0$, $\int K(t)dt = 1$, $\int tK(t)dt = 0$, $\int t^2K(t)dt = k_2$, $\int_{-\infty}^{\infty}[K(t)]^2dt < \infty$, $K(t) = K(-t)$.

Under reasonable conditions on m(.) and the design points x_i, $m(x_i)$ will be consistent for $m(x_i)$,$i = 1, 2, ..., n$ as $b \to 0$, $nb \to \infty$, where $n \to \infty$.

Dorfman (1992) suggested an estimator of T as

$$\hat{T_{np}} = \Sigma_{jes} y_j + \Sigma_{iep-s} \hat{m}(x_i) \tag{2.5}$$

where $\sum_{jes} y_j$ is the sum of y values on the sample space and $\sum_{iep-s} \hat{m}(x_i)$ is an estimate of x_i in the non sample space.
For model based estimators, the estimator in equation (2.5) ignores the sampling probabilities. Dorfman (1994) derived the conditional mean and variance under equation (2.5).The derived results were:

$$E(\hat{T}_{np} - T) = \Sigma_{ier}(d_s x_j)^{-1}[\Sigma_{jes}\frac{1}{bn}k(\frac{x_i - x_j}{n})]m(x_j) \tag{2.6}$$

and

$$Var(\hat{T}_{np} - T) = \Sigma_{ies}W_i^2\sigma^2(x_i) + \Sigma_{jer}\sigma^2(x_j) \tag{2.7}$$

Where, d_s is the difference between $x_i's$, b is the bandwidth and n is the sample size. The results imply that $\hat{T}_{np}$ is a consistent estimator of T. $\hat{T}_{np}$ is bias to misspecification of $E[Y_i|X_i = x_i]$ and also to the misspecification of $Var[Y_i|X_i = x_i]$. Otieno and Mwalili (2000) derived an improved estimator of $\sigma^2(x_j)$ by smoothing e_j^2.
Let h be a smoothing parameter. Using this parameter the weight $w_i(x)$ was denoted by

$$\hat{\sigma}^2_{np}(x_i) = \Sigma_{jes}w_j^2(X_i)\hat{e_j}^2 \tag{2.8}$$

So that the derived error variance was estimated by,

$$V_n = \Sigma_{jes}w_j^2(X_i)\hat{\sigma_{np}}^2(X_i) + \Sigma_{jer}\hat{\sigma_{np}}^2(x_i) \tag{2.9}$$

A comparison of variance estimators by Otieno and Mwalili (2000) and variance estimators suggested by Royall and Cumberland (1981) showed that V_n was more efficient than V_L,V_C, V_D.

$$V_L = N^2\frac{(1-f)}{n\bar{x}_s^2}\bar{X}\bar{X}_r\Sigma\frac{e_i}{\sqrt{x_i}} \tag{2.10}$$

$$V_D = N^2\frac{(1-f)}{n^2}\bar{X}\bar{X}_r\Sigma e_i^2(1-k_i)^{-1} \tag{2.11}$$

$$V_C = \frac{N^{2(1-f)}}{n}\Sigma\frac{e_i^2}{n-1} \tag{2.12}$$

where $f = \frac{n}{N}$, $e_i = y_i - (\bar{y}_s/\bar{x}_s)x_i$

2.1.3 Selection of the Kernel Function

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

i) The kernel smoother should be easy and simple to implement.

ii) The kernel smoother should not take very small values that may result in numerical underflow in the computer.

iii) The kernel smoother should be user friendly i.e it should practically fit in both simulated and raw data.

iv) The range of the smoother should be well defined and not open as in the case of 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	**k(u)**	**Efficiency**
Uniform	$k(u) = \frac{1}{2}$	0.929
Triangular	$k(u) = (1 - \|u\|)$	0.986
Epanechnikov	$k(u) = \frac{3}{4}(1 - u^2)$	1
Quartic	$k(u) = \frac{15}{16}(1 - u^2)^2$	0.994
Triweight	$k(u) = \frac{35}{32}(1 - u^2)^3$	0.987
Tricube	$k(u) = \frac{70}{81}(1 - \|u\|^3)^3$	0.998
Gaussian	$k(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$	0.951
Cosine	$k(u) = \frac{\pi}{4} \cos(\frac{\pi}{2}u)$	0.999
Logistic	$k(u) = \frac{1}{e^u + 2 + e^{-u}}$	0.887
Sigmoid function	$k(u) = \frac{2}{\pi} \frac{1}{e^u + e^{-u}}$	0.843

2.2 Existing estimators of the population variance and their asymptotic properties

Assuming a population consisting of N distinct units of values (x_i, y_i), and $x_i > 0$ for$1 < i < N$. Take from the population a SRSWR of size n. The sample and population means of y_i and x_i are denoted by $\bar{y}$ and $\bar{x}$ respectively. The ratio estimator

$$\hat{\bar{Y}}_R = \bar{Y}\bar{X}/\bar{x} \tag{2.13}$$

is a common estimator of $\bar{Y}$. The ratio estimator is easy to use in practice and efficiently combines the covariate information in x_i when y_i when they are positively

correlated. Royall (1970) proved that the ratio estimator is the BLUE predictor under the following super population model.

$$Y_i = \beta x_i + e_i \tag{2.14}$$

where e_i are iid with mean zero and variance $\sigma^2(x_i)$.

There is no closed form for MSE $(\hat{\bar{y}_R})$ or Var $(\hat{\bar{y}_R})$. Cochran (1977) gave an approximation of V_o and V_2 as,

$$V_o = \frac{1-f}{n}\frac{1}{N-1}\Sigma_1^N((y_i - \frac{\bar{Y}}{\bar{X}})x_i)^2 \tag{2.15}$$

$$V_2 = \frac{1-f}{n}(\frac{\bar{X}}{\bar{x}})^2\frac{1}{n-1}\Sigma_1^N((y_i - \frac{\bar{Y}}{\bar{X}})x_i)^2 \tag{2.16}$$

Priestly-Chao estimator is given as

$$\hat{m}_c(x_j) = \frac{1}{nh}\Sigma_{ies}W_i(X_j)y_i$$

where $W_i(x_j) = (\frac{x_i - x_{i-1}}{h})K(\frac{x_i - x_j}{h})$

The smoothing function however 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, this particular case the sum of the weights is not equal to one but is rather 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.

Royall and Eberhardt (1975) suggested the variance estimator given by

$$V_H = V_0\frac{\bar{x_c}\bar{X}}{\bar{x}^2(1 - \frac{cx^2}{n})^{-1}} \tag{2.17}$$

where $\bar{x_c}$ is the mean of non-sampled units, c_x is the x sample coefficient of variation. V_H is approximately unbiased for more general variance patterns and is asymptotically equivalent to V_j.

Another variance estimator, which parameter follows from standard least squares theory, is

$$V_L = \frac{\frac{(1-f)}{n}\bar{x_c}\bar{X}}{\frac{1}{n-1}\Sigma e_i^2/x_i} \tag{2.18}$$

It is unbiased under the model but can be biased if $var(y_i) = \sigma^2(x_i)$ is violated in the model, Royall and Eberhardt (1975). Their empirical behavior were shown to be less efficient, biased and inconsistent in Royall and Cumberland (1981).

Cochran (1977) showed that Var $\bar{y}$ can be approximated by the approximate variance

$$Vappr = \frac{(1-f)}{n(N-1)}\Sigma(y_i - r\bar{x})^2 \tag{2.19}$$

where $f = n/N$ and $r = \bar{y}/\bar{x}$. For large samples the approximation is adequate. Cochran (1977) showed that for a sample of size ($n < 12$) Vappr can greatly underestimate MSE.

Later Royall and Cumberland (1981) suggested a closely related estimator

$$V_D = \frac{(1-f)}{n}\frac{\bar{x_c}\bar{x}}{\bar{x^2}}\frac{1}{n}\Sigma\frac{\hat{e_i^2}}{1-\frac{x_i}{n\bar{x}}} \tag{2.20}$$

Here both V_H and V_D were shown to be unbiased under the model and approximately unbiased for more general variance patterns and asymptotically equivalent to V_j.

Gasser and Muller (1979) proposed an estimator that involved sorting of X variable. The estimator is given as

$$\hat{m}_x = \Sigma_{j=1}^{n}\int_{s_j-1}^{s_j} k(u-1)dus_j$$

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

The resulting nonparametric estimator of the population variance is given as

$$\hat{T}_G = \Sigma_{ies}y_i + \Sigma_{jer}\hat{m}(x_j)$$

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 produces a good fit and the curve does not have too much variation.

Dickey and Fuller (1981) suggested a regression adjustment to V_0. The estimator due to Fuller (1981) is given by

$$V_{reg} = V_0 + \frac{(1-f)}{n}\hat{b}e_x^2(\bar{X} - \bar{x}) \tag{2.21}$$

where $\hat{b}e_x^2$ is the sample regression coefficient of $\frac{\hat{e_i}}{x_i}$.

A common variance estimator is the Jackknife variance estimator V_j,

$$V_j = (1-f)\bar{x}^2\frac{n}{(n-1)}\Sigma D_j^2 \tag{2.22}$$

where D_j is the difference between the ratio $\frac{(n\bar{y}-y_j)}{(n\bar{x}-x_j)}$ and the average of these n ratios.

The jackknife estimator is independent of a superpopulation model.

Royall and Cumberland (1981) studied the model-based and sampling properties of Vj and deduced that it is approximately unbiased.

Wu (1982) proposed a general class of estimators $Vg = gV_0$, where V_0 is the sample mean of e_i^2. It was shown that the first terms of MSE (Vg) is minimized by $g_{opt} = S_{xz}\bar{X}/S_x^2\bar{Z}$ which is the population regression coefficient of $\frac{Z_i/\bar{z}}{X_i/\bar{x}}$.
Where g_{opt} is the sample regression coefficient of $\frac{z_i}{\bar{z}}$ over $\frac{x_i}{\bar{x}}$, $Z_i = e_i^2 - 2e_i\Sigma x_i e_i/\Sigma X_i$. S_x^2 and S_{xz} are the population variance and covariance respectively. The second term of Z_i accounts for the possible non zero intercept in the population when fitted by a straight line.

Isaki (1983) proposed ratio type estimator of the population variance S_y^2 when the population variance S_x^2 of the auxiliary variable X is known together with its bias and mean squared error as:

$$\hat{S}_R^2 = S_y^2 \frac{S_x^2}{s_x^2} \tag{2.23}$$

$$B(\hat{S}_R^2) = \gamma S_y^2[(\beta_{2x} - 1) - (\lambda 22 - 1)] \tag{2.24}$$

$$MSE(\hat{S}_R^2) = \gamma S_y^4[(\beta_{2y} - 1) + (\beta_{2x} - 1) - 2] - 2(\lambda 22 - 1) \tag{2.25}$$

where $\beta_{2y} = \frac{\mu_{40}}{\mu_{20}}$,$\beta_{2x} = \frac{\mu_{04}}{\mu_{02}}$,$\lambda 22 = \frac{\mu_{22}}{\mu_{20}\mu_{02}}$

When deriving the asymptotic properties of the Nadraya-Watson estimator, it becomes tedious to find the derivatives of the estimator due to the nature of the denominator of the estimator. Otieno and Mwalili(2000) therefore gave the estimator of the finite population variance as

$$\hat{T}_c = \Sigma_{ies} y_i + \Sigma_{jer} \hat{m}_c(x_j)$$

The residual sum of squares can be used to compute the regression function. It is given as,

$$\hat{m}_s(x_j) = \Sigma_{i=1}^n (y_i - g(x_i))^2$$

Breidt and Opsomer (2000) studied the Horvitz Thompson estimator of the population given by $\hat{t}_y = \Sigma_{ies} \frac{y_i}{\pi_i}$ where π is the inclusion probability. Breidt and Opsomer (2000) used the local polynomial approach and the survey values y_i. Assuming a continous kernel function k and a bandwidth h_N they defined a local polynomial estimator of degree q based on the entire population. Letting $y_u = [y_i]_{ieu_N}$ be a N vector

of survey values in the finite population and a matrix of dimension $N*(q+1)$ defined by $X_{U_i} = \begin{pmatrix} 1 & (x_1 - x_i) & \dots & (x_1 - x_i)^q \\ 1 & (x_2 - x_i) & \dots & (x_2 - x_i)^q \\ \vdots & \vdots & \ddots & \vdots \\ 1 & (x_N - x_i) & \dots & (x_N - x_i)^q \end{pmatrix}$

Define an $N*N$ matrix by $W_{U_i} = diag\{\frac{1}{h_N}K(\frac{x_j - x_i}{h_N})\}.e_i$ being a vector of 1 line in the first position and 0 otherwise. The estimator of the regression function at $m(x_i)$ is then given by;

$$m_i = e_i^{'}(X_{U_i}^{'}W_{U_i}X_{U_i})^{-1}(X_{U_i}^{'}W_{U_i}X_{U_i}) \tag{2.26}$$

as long as $(X_{U_i}^{'}W_{U_i}X_{U_i})$ is invertible. The design unbiased estimator of the population variance is then given by

$$t_y^* = \Sigma_{ies}\frac{y_i - m_i}{\pi_i} + \Sigma_{ieU_N}m_i \tag{2.27}$$

which is a generalised difference estimator with variance given by

$$V_p(t_y^*) = \Sigma_{i,jeU_N}(\pi_{i,j} - \pi_i\pi_j)\frac{y_i - m_i}{\pi_i}\frac{y_j - m_j}{\pi_j} \tag{2.28}$$

However, the estimator is based on the entire population. The sample based consistent estimator of the regression function $m(x_i)$ is given by;

$$\hat{m}_i = e_i^{'}(X_{U_i}^{'}W_{U_i}X_{U_i}^{-1})X_{U_i}^{'}W_{U_i}X_{U_i} \tag{2.29}$$

For observations less than $(q+1)$, the matrix $X_{U_i}^{'}W_{U_i}X_{U_i}$ is singular. Breidt and Opasomer (2000) therefore considered an adjusted sample based estimator that is guaranteed to exist for any sample drawn from the population. The sample based estimator of the population total is a linear combination of survey values with weights being the inverse inclusion probabilities. The estimator that uses the adjusted sample smoother 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 achieved the Godambe-Joshi lower bound.

Upadhyaya and Singh (2001) suggested a modified ratio type variance estimator using the population mean of the auxiliary variable together with its bias and mean squared error as,

$$\hat{S}_{52}^2 = s_y^2[\frac{\bar{X}}{\bar{x}}] \tag{2.30}$$

$$B(\hat{S}_{52}^2) = \gamma S_y^2[C_x^2 - \lambda_{21}C_x] \tag{2.31}$$

$$MSE(\hat{S}_{52}^2) = \gamma S_y^4[(\beta_{2y} - 1) + C_x^2 - 2\lambda_{21}C_x] \tag{2.32}$$

The proposed variance estimators above are biased but have smaller mean squared errors compared to the traditional ratio type variance estimator suggested by Isaki (1983) under certain conditions.

Zheng and Little (2003) gave the spline estimator of the population variance as

$$\hat{T}_C = \Sigma_{ies} y_i + \Sigma_{jer} \hat{m}_s(x_j)$$

Subramani and Kumarapandiyan (2015) proposed a class of modified ratio type variance estimators $\hat{S}_{p_i}^2$ for estimating the population variance S_y^2 given as

$$\hat{S}_{p_i}^2 = s_y^2[\frac{\bar{X} + w_i}{\bar{x} + w_i}], i = 1, 2, 3, ..., 51 \tag{2.33}$$

Subramani and Kumarapandiyan (2015) derived the bias and MSE of $\hat{S}_{p_i}^2$ as

$$Bias(\hat{S}_{p_i}^2) = \frac{(1-f)}{n} S_y^2(O_{pi}^2 C_x^2 - O_{pi}\lambda_{21}C_x) \tag{2.34}$$

$$MSE(\hat{S}_{p_i}^2) = \frac{(1-f)}{n} S_y^4((\beta_{2(y)} - 1) + O_{pi}^2 C_x^2 - 2O_{pi}\lambda_{21}C_x); i = 1, 2, 3, ..., 51 \tag{2.35}$$

2.3 Comparison of the relative efficiency of existing variance estimators

Based on the empirical study by Wu and Deng (1982), the estimator V_o is the least efficient among the estimators that they studied. It has unreliable t-intervals that neither estimates the MSE nor the conditional MSE of Y_R well. However it is the most commonly recommended estimator on sampling. Wu and Deng (1982) showed that the performance of the variance estimators V_o and V_2 for estimating MSE depends on the underlying populations and has no direct bearing on the performance of interval estimates.

The jackknife variance V_J gives very reliable t-intervals than V_H and V_D. All the three estimators give t-intervals that are close to V_2 for large samples, but not to V_g. The reason that V_J does so poorly for estimating MSE is because it estimates the conditional MSE well, and typically the conditional MSE varies greatly with mean of x. The estimators V_g and V_{reg} are asymptotically equivalent. The estimators are good for estimating the unconditional MSE but give unreliable t-intervals.

Wu and Deng (1982) emphasized that the reliable t-intervals seems to be related to the good performance of V_2 for estimating the conditional MSE. The problem of choosing a proper ancillary statistic and making inference conditional on it is an important one in the theory of survey sampling. Encouraged by the relative efficiency of V_2 over V_0, Wu and Deng (1982) considered the variance estimation problem in other settings.

Otieno and Mwalili (2000) studied the empirical properties of V_L, V_D,V_C and V_n in a natural population. Otieno and Mwalili (2000) found out that one can use V_n to estimate the MSE. Otieno and Mwalili (2000) further investigated how efficient the four variance estimators were on tracking the conditional MSE. In the study V_D and V_n both follow the MSE very closely but, V_C and V_L does not efficiently approximate the MSE.This indicates that V_n is a strong competitor to V_D.

The performance of the estimator proposed by Breidt and Opsomer (2000) was compared to that of other parametric and nonparametric estimators. Both the parametric and nonparametic 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 considered.

Subramani and Kumarapandiyan (2015) derived the conditions for which the proposed estimators are more efficient than the traditional and existing modified ratio type variance estimators. Subramani and Kumarapandiyan (2015) assessed the performance of the traditional estimators and the proposed estimator and observed that from a numerical representation, the bias and mean squared error of the proposed estimator was less than the mean squared error and bias of existing estimators. Subramani and Kumarapandiyan (2015) strongly recommended that the proposed modified ratio type variance estimators may be preferred over the traditional ratio type variance estimator and modified ratio type variance estimators for use in practical applications.

2.4 Research Gap

The reviewed methods of estimating the population variance employed kernel smoothers in estimating regression functions. Majority of kernel smoothers suffers from boundary problems that require a refinement at the boundary points. This is such that at the boundary points, the bias of the estimators decreases at the cost of an ascending variance. There also exists a trade off between the bias and variance of the estimators. Selecting a narrow range results in a low bias and high variance while selecting a wide range results in a high bias and low variance. Also the locally weighted averages can be very biased if the regression function has a high slope. This study adopts a multiplicative bias corrected technique to estimate a robust variance of the population mean under simple random sampling.With sufficient smoothness of the density function, the multiplicative bias corrected approach reduces the sequence of the bias with no effect on the variance of the estimator.

Chapter 3

Research Methodology

3.1 Introduction

This chapter presents the estimation of variance, the variance of the ratio estimator under robust variance structure of the population mean, multiplicative bias robust variance estimator for ratio estimator using a smoother function and the asymptotic properties of the multiplicative bias corrected variance estimator.

3.2 Estimation of Variance

Estimation of variance is taken in consideration in this study. Suppose that there are units $U_1, U_2, ..., U_N$ with corresponding survey measurements $y_1, y_2, ..., y_N$ 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 variance for any set of data collected.

3.3 The variance of the ratio estimator under robust variance structure of the population mean

In this section, the exact procedure of estimating the robust variance of the population mean is presented. Let $u = (u_1, ..., u_N)$ be a finite population of size $N < \infty$. Suppose that for every unit u_i, some unknown measurement (auxiliary measurement) denoted as x_i;(i=1,2, , N) exists. It is of interest to find an estimator of the population variance i.e.

$$T = y_1 + y_2 + ...+, y_N \tag{3.1}$$

of the survey measurements y_i,s which are unknown. The estimator of equation (3.1) is deduced based on the model

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

$$E(y) = \mu(x_i)$$

$$Cov(Y_i, Y_j) = \sigma^2(x_i), fori = j, 0otherwise$$

where $\mu(x_i)$ and $\sigma^2(x_i)$ are assumed to be smooth functions of x_i mainly because the above are the simplest form of equations that describes the relationship between the auxiliary variable and the survey variable.

A simple random sample of size n is taken from the population U. The problem is how to estimate T, using known x_i, s in the entire population and the sampled values of y_i 's.

It is a common practice in survey sampling to use a ratio estimator in such context, especially when there is positive correlation between the auxiliary measurements x_i's and the survey measurements y_i' s.

Let $\bar{x}$,$\bar{y}$ be sample means of x_i's , y_i 's respectively and $\hat{X}$ and $\hat{Y}$ be the corresponding population values. Then the estimator:

$$\hat{T_R} = \frac{\bar{y}\bar{X}}{\bar{x}} = r_n\bar{X} \tag{3.2}$$

where $r_n = \frac{\bar{y}}{\bar{x}}$. $\hat{T_R}$ is called the ratio estimator of the population mean $\hat{Y}$.

A ratio estimator of the population is given by

$$\hat{T} = N\frac{\bar{y}}{\bar{x}}\bar{X} = Nr_n\bar{X}$$

The ratio estimator is generally motivated on the basis of a [superpopulation model, prediction model]. The ratio estimator is BLUE under the above model (best linear unbiased estimator). An estimator $\hat{\theta}$ is said to be model unbiased under the model if $E_{cs}(\hat{\theta}) = E_{cs}(\theta)$. that is under the above model if

$$E_{cs}(\hat{T_R}) = E_{cs}(r_n\bar{X})$$

$$E_{cs}(r_n\bar{X} = \bar{X}E_{cs}(r_n) = \frac{\bar{X}}{\bar{x}}E_{cs}(\bar{y})$$

$$= \frac{\bar{X}}{\bar{x}}\frac{1}{n}\Sigma_{i=1}^{n}E(y_i) = \beta\bar{X}$$

$$ButE_{cs}(\bar{Y}) = E_{cs}\frac{1}{n}\Sigma_{i=1}^{n}y_i = \beta\bar{X}$$

$\hat{T_R}$ is model unbiased estimator of the population mean,$\bar{Y}$. Then clearly,

$$E_{cs}(\hat{T_R} - T) = E_{cs}(\hat{T_R} - \hat{Y}) = \frac{\alpha}{\bar{x}}(\frac{\bar{X}}{\bar{x}}) = \frac{\alpha\bar{x}}{\bar{x}^2}$$

which is not equal to zero.
Clearly R is not bias robust to model misspecification: $E(Yi) = \alpha + \beta xi$.Many studies prefer using a ratio estimator R of the population mean $\hat{Y}$ in the presence of auxiliary measurements x_is, $(i = 1, ..., N)$. In such studies it is believed that the regression line passes through the origin. Assuming this to be true,this project is concerned with estimating the precision of the ratio estimator when the variance function is not linearly related to the auxiliary measurement x_i $(i = 1, 2, ..., N)$. In particular we consider the following model $Ecs(Y_i) = \beta x_i$
The robust model of the variance of y_i's is given by

$$cov(Y_i, Y_j) = \sigma^2(x_i) = j$$

$$0, i \neq 0$$

The function $\sigma^2(x_i)$ is assumed to be twice continuously differentiable. Under the above model,

$$\hat{T_R} - T = (\Sigma_{ies} y_i + r\Sigma_{ier} x_i) - (\Sigma_{ies} y_i + r\Sigma_{ier} y_i) \tag{3.3}$$

$$= r\Sigma_{ier} x_i - \Sigma_{ier} y_i$$

$$= \Sigma_s y_i (\frac{\Sigma_r x_i}{\Sigma_s x_i}) - \Sigma_r y_i$$

$$Var(\hat{T_R}) = (\frac{\Sigma_r x_i}{\Sigma_s x_i})^2 \Sigma_s \sigma^2(x_i) + \Sigma_r \sigma^2(x_i) \tag{3.4}$$

then,

$$Var(\hat{T_R}) = (\frac{\Sigma_r x_i}{\Sigma_s x_i})^2 \Sigma_{ies} \sigma^2 x_i + \Sigma_{ier} \sigma^2 x_i$$

$$= \sigma^2 \Sigma_{ier} x_i [\frac{\Sigma_{ier} x_i}{\Sigma_{ies} x_i} + 1]$$

$$= \sigma^2 \Sigma_{ier} x_i (\frac{\Sigma_r x_i + \Sigma_s x_i}{\Sigma_s x_i})$$

$$(\frac{N(N-n)\bar{X}\bar{x}_r}{n\bar{x}_s})\sigma^2$$

A number of parametric estimators for estimating $Var(\hat{T_R})$ as studied by Otieno and Mwalili (2000) include:

$$V_C = \frac{N^2(1-f)}{n}[\Sigma_{ies} \frac{\hat{e}_i^2}{(n-1)}]$$

$$V_L = \frac{N^2(1-f)}{n^2} \frac{\bar{X}\bar{X}_r}{\bar{x}^2} \Sigma_{ies} \frac{\hat{e}_i^2}{\sqrt{x_i}}$$

$$V_N = \Sigma_{jes} w_j^2(X_i)\hat{\sigma}_{np}^2(x_i) + \Sigma_{jer}\hat{\sigma}_{np}^2(x_i)$$

Where $f = \frac{n}{N}$,$\hat{e}_i = y_i - \frac{\bar{y}_s}{\bar{x}_s}x_i$,$\bar{y}_s$ is the sample mean of $y_i's$, $\bar{X}_r$, $\bar{x}_s$ represent non-sample and sample means of $x_i's$ respectively. Moreover, $\hat{\sigma}_{np}^2 = \Sigma_{jes} w_j(x_i)\hat{e}_j^2$ and $w_j(x)$ is a weight.

In the recent past a non-parametric variance estimator based on Nadaraya-Watson smoother has been proposed. This estimator is based on squared residuals

$$e_i^2 = (y_i - rx_i)^2$$

The estimator of equation (3.4) is then obtained by substituting a smoother to $\sigma^2(x_i)$ in (3.4) as

$$\hat{\sigma}^2(x_i) = \Sigma_{jes} w_j(x_i)\hat{e}_j^2$$

to give,

$$V_{Nw} = (\frac{\Sigma_r x_i}{\Sigma_s(x_i}})^2\Sigma_{ies}\hat{\sigma}_{Nw}^2(x_i) + \Sigma_{ier}\hat{\sigma}_{Nw}^2(x_i) \tag{3.5}$$

here $w_j(x_i) = \frac{1}{4}(1 - u_i^2)$. This estimator, like many other Kernel smoothers suffers from boundary problem, including outlier sensitivity.

In this project a multiplicative bias reduction procedure is used to develop a bias robust variance estimator.

3.4 Multiplicative bias robust variance estimator for ratio estimator using a smoother function

Suppose $(X_1, Y_1), (X_2, Y_2)....., (X_N, Y_N)$ are bivariate independent r.vs (X, Y). Assume all X_i's are unknown.
Define

$$\hat{e_j^2} = \sigma^2(X_i) + O(n^{-1}) = \sigma^2(X_i) + \epsilon_i \tag{3.6}$$

Consider a smoother of variance function

$$\sigma_n^2(X_i) = \Sigma_{jes} W_j(X_i)\hat{e}_j^2 \tag{3.7}$$

Then the ratio $\beta_j = \frac{e_j}{\sigma^2(X_i)}$ is a noisy estimator of $\frac{\sigma^2(X_i)}{\sigma_n^2(X_i}$.
Smoothing β_j yields

$$\hat{\alpha}(X_i) = \sigma_n^2(X_i)\beta_j \tag{3.8}$$

Equation(3.8) is used as a multiplicative correction of the pilot smoother in equation (3.7) and is defined as

$$\hat{\sigma}_n^2(X_i) = \hat{\alpha}(X)\sigma_n^2(X_i) \tag{3.9}$$

Assumptions of the study

The following assumptions are made in the estimation of $\hat{\sigma}_n^2(X_i)$

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

b)$0 < a < \sigma(x_i) < b$

c)The regression function is twice continuously differentiable everywhere.

The positivity assumption on the regression function $\sigma(x_i)$ is important when performing the multiplicative bias correction.The regression function might cross the x-axis and in such a situation Glad (1998) proposed to shift all the response data by a distance a such that the new regression function is $\sigma(x_i) + a$

substituting (3.8)to (3.9) yields,

$$\hat{\sigma}_n(x_i) = \Sigma_{j=1}^n w_j(x_i) \frac{\hat{\sigma_n}(x)}{\hat{\sigma_n} x_j} y_j \tag{3.10}$$

Suppose that

$$E[\hat{\sigma}(x)|x_i, ..., x_N] = \Sigma_{j=1}^n w_j(x_i) E[Y_j] = \Sigma_{j=1}^n w_j(x_i)\sigma(X_j) = \hat{\sigma}_n(x) \tag{3.11}$$

Using $\frac{\hat{\sigma_n}(x)}{\hat{\sigma_n} x_j}$ in equation (3.10)yields

$$\frac{\hat{\sigma_n}(x)}{\hat{\sigma_n} x_j} = \frac{\bar{\sigma_n}(x)}{\bar{\sigma_n} x_j} * \frac{\hat{\sigma_n}(x)}{\hat{\sigma_n}(x_j)} * (\frac{\hat{\sigma_n}(x_j)}{\hat{\sigma_n} x_j})^{-1} \tag{3.12}$$

$$\frac{\hat{\sigma_n}(x)}{\hat{\sigma_n}(x_j)} = \frac{\bar{\sigma_n}(x)}{\bar{\sigma_n}(x_j)} * (\frac{\bar{\sigma_n}(x) + \hat{\sigma_n}(x) - \bar{\sigma_n}(x)}{\bar{\sigma_n}(x)}) * (\frac{\bar{\sigma(x_j)}}{\bar{\sigma}(x_j) + \hat{\sigma}(x_j) - \bar{\sigma}(x_j)}) \tag{3.13}$$

$$\frac{\hat{\sigma_n}(x)}{\hat{\sigma_n}(x_j)} = \frac{\bar{\sigma_n}(x)}{\bar{\sigma_n}(x_j)} * (1 + \frac{\bar{\sigma_n}(x) - \bar{\sigma_n}(x)}{\bar{\sigma_n(x)}}) * (1 + \frac{\hat{\sigma_n}(x_j) - \bar{\sigma}(x_j)}{\bar{\sigma}(x_j)})^{-1} \tag{3.14}$$

Let $\frac{\bar{\sigma_n}(x) - \bar{\sigma_n}(x)}{\bar{\sigma_n(x)}}) = b_n(x)$ and $\frac{\hat{\sigma_n}(x_j) - \bar{\sigma}(x_j)}{\bar{\sigma}(x_j)} = b_n(x_j)$.

Equation (3.14)can now be expressed as,

$$\frac{\hat{\sigma_n}(x)}{\hat{\sigma_n}(x_j)} = \frac{\bar{\sigma_n}(x)}{\bar{\sigma_n}(x_j)} * (1 + b_n(x)) * (1 + b_n(x_j))^{-1} \tag{3.15}$$

Applying the binomial expansion to equation (3.15) yields

$$(1 + b_n(x)) * (1 + b_n(x_j)^{-1}) = [1 + b_n(x)][1 - b_n(x_j) + b_n(x_j)^2]$$

This further reduces to

$$(1 + b_n(x)) * (1 + b_n(x_j))^{-1} + r_j(x, x_j) = 1 + b_n(x) - b_n(x_j) + r_j(x, X_j) \tag{3.16}$$

where $r_j(x, x_j)$ is the remainder term involving x and x_j Substituting equation (3.16) to equation (3.15) yields

$$\frac{\hat{\sigma_n}(x)}{\hat{\sigma_n}(x_j)} = \frac{\bar{\sigma_n}(x)}{\bar{\sigma_n}(x_j)} * [1 + b_n(x) - b_n(x_j) + r_j(x, x_j)] \tag{3.17}$$

Substituting equation (3.17) to equation (3.10)and using the model $Y_j = \sigma(X_j) + e_j$ yields,

$$\hat{\sigma}_n(X_i) = \Sigma_{j=1}^n w_j(\frac{\bar{\sigma_n}(x)}{\bar{\sigma_n}(x_j)} * [1 + b_n(x) - b_n(x_j) + r_j(x, x_j)][\sigma(X_j) + e_j]) \tag{3.18}$$

$$\begin{aligned}\hat{\sigma}_n(X_i) = \Sigma_{j=1}^n w_j((\frac{\bar{\sigma_n}(x)}{\bar{\sigma_n}(x_j)}\sigma(X_j) * [1 + b_n(x) - b_n(x_j) + r_j(x, x_j)]) \\ +\Sigma_{j=1}^n w_j((\frac{\bar{\sigma_n}(x)}{\bar{\sigma_n}(x_j)}e_j[1 + b_n(x) - b_n(x_j) + r_j(x, x_j)])\end{aligned} \tag{3.19}$$

$$\hat{\sigma}_n(X_i) = \Sigma_{j=1}^n w_j \frac{\bar{\sigma_n}(x)}{\bar{\sigma_n} x_j}\sigma(X_j)+\Sigma_{j=1}^n w_j \frac{\bar{\sigma_n}(x)}{\bar{\sigma_n} x_j}(e_j+\sigma(X_j)[b_n(x)-b_n(x_j)])+\Sigma_{j=1}^n w_j \frac{\bar{\sigma_n}(x)}{\bar{\sigma_n} x_j}e_j \tag{3.20}$$

$$[b_n(x) - b_n(x_j) + r_j(x, x_j)] + \Sigma_{j=1}^n w_j \frac{\bar{\sigma_n}(x)}{\bar{\sigma_n} x_j} r_j(x, x_j)[\sigma(X_j) + e_j] \tag{3.21}$$

Applying the assumption that $nh \to \infty$, in probability the remainder terms converge to 0. Therefore $r_j(x, x_j)[\sigma(X_j) + e_j] = O_p(\frac{1}{nh})$ and equation (3.20)reduces to,

$$\hat{\sigma}_n(X_i) = \Sigma_{j=1}^n w_j \frac{\bar{\sigma_n}(x)}{\bar{\sigma_n} x_j}\sigma(X_j)+\Sigma_{j=1}^n w_j \frac{\bar{\sigma_n}(x)}{\bar{\sigma_n} x_j}(e_j+\sigma(X_j)[b_n(x)-b_n(x_j)])+\Sigma_{j=1}^n w_j \frac{\bar{\sigma_n}(x)}{\bar{\sigma_n} x_j}e_j$$

$$[b_n(x) - b_n(x_j) + r_j(x, x_j)] + O_p(\frac{1}{nh}) \tag{3.22}$$

Our estimator of the population variance therefore is

$$\hat{V}_{MBC} = \Sigma_{ies} Y_i + \Sigma_{ie(p-s)}[\Sigma_{j=1}^n w_j(x; h)\frac{\hat{\sigma}_n(x)}{\hat{\sigma}_n(x_j)} + \Sigma_{j=1}^n w_j(x; h)\frac{\hat{\sigma}_n(x)}{\hat{\sigma}_n(x_j)}(e_j$$

$$\sigma(x_j)[b_n(x) - b_n(x_j)]) + \Sigma_{j=1}^n w_j(x; h)\frac{\hat{\sigma}_n(x)}{\hat{\sigma}_n(x_j)}e_j[b_n(x) - b_n(x_j)] + O_p(\frac{1}{nh}) \tag{3.23}$$

3.5 Asymptotic Unbiasedness of the MBC Estimator

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

$$E[\hat{V}_{MBC} - V] = E[\hat{V}_{MBC}] - E[V]$$

The expected value of the MBC estimator is calculated as,

$$E[\hat{V}_{MBC}] = E[\Sigma_{ies} Y_i + \Sigma_{je(p-s)}(\Sigma_{j=1}^{n} \hat{\sigma}_n(x_i))] = \Sigma_{ies} E[Y_i] + \Sigma_{ie(p-s)\Sigma_{j=1}^{n} E[\hat{\sigma}_n(x_i)]} \quad (3.24)$$

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

$$E[\hat{\sigma}_n(x_i)] = E[\Sigma_{j=1}^{n} w_j(x;h) \frac{\hat{\sigma}_n(x)}{\hat{\sigma}_n(x_j)} \sigma(x_j) + \Sigma_{j=1}^{n} w_j(x;h) \frac{\hat{\sigma}_n(x)}{\hat{\sigma}_n(x_j)} (e_j + \sigma(X_j)[b_n(x) - b_n(x_j)])$$

$$+ \Sigma_{j=1}^{n} w_j(x;h) \frac{\hat{\sigma}_n(x)}{\hat{\sigma}_n(x_j)} e_j (b_n(x) - b_n(x_j)) + O_p(\frac{1}{nh})] \quad (3.25)$$

$$= E[\Sigma_{j=1}^{n} w_j(x;h) \frac{\hat{\sigma}_n(x)}{\hat{\sigma}_n(x_j)} \sigma(x_j) + \Sigma_{j=1}^{n} w_j(x;h) A_j(x) + \Sigma_{j=1}^{n} w_j(x;h) B_j(x)] + O_p(\frac{1}{nh}) \quad (3.26)$$

Where

$$A_j(x) = \frac{\bar{\sigma_n}(x)}{\bar{\sigma_n} x_j} (e_j + \sigma(X_j)[b_n(x) - b_n(x_j)])$$

and

$$B_j(x) = \frac{\bar{\sigma_n}(x)}{\bar{\sigma_n} x_j} e_j [b_n(x) - b_n(x_j)]$$

Analyzing the first term of equation (3.26)

$$E[\Sigma_{j=1}^{n} w_j(x;h) \frac{\hat{\sigma}_n(x)}{\hat{\sigma}_n(x_j)} \sigma(x_j)]$$

yields

$$\Sigma_{j=1}^{n} w_j(x;h) \frac{\hat{\sigma}_n(x)}{\hat{\sigma}_n(x_j)} \sigma(x_j)$$

$$E[\Sigma_{j=1}^{n} w_j(x;h) \frac{\hat{\sigma}_n(x)}{\hat{\sigma}_n(x_j)} (e_j + \sigma(X_j)[b_n(x) - b_n(x_j)])]$$

$$= E[\Sigma_{j=1}^{n} w_j(x;h)] \quad (3.27)$$

This yields

$$E[\Sigma_{j=1}^{n} w_j(x;h) \frac{\bar{\sigma}_n(x)}{\bar{\sigma}_n(x_j)}] = \Sigma_{j=1}^{n} w_j(x;h) \frac{\bar{\sigma}_n(x)}{\bar{\sigma}_n(x_j)} \sigma(X_j) \quad (3.28)$$

This is because $\sigma(X_j)$ is the variance function. Analyzing the second term of equation (3.26)

$$E[\Sigma_{j=1}^{n} w_j(x;h)\frac{\bar{\sigma}_n(x)}{\bar{\sigma}_n(x_j)}(e_j + \sigma(X_j)((b_n)(x) - b_n(X_j))]$$

$$= E[\Sigma_{j=1}^{n} w_j(x;h)(\frac{\bar{\sigma}_n(x)}{\bar{\sigma}_n(x_j)}e_j + \frac{\bar{\sigma}_n(x)}{\bar{\sigma}_n(x_j)}\sigma(X_j)(\frac{\bar{\sigma}_n(x) - \bar{\sigma}_n(x)}{\bar{\sigma}_n(x)}) - \frac{\bar{\sigma}_n(x)}{\bar{\sigma}_n(x_j)}\sigma(X_j)(\frac{\bar{\sigma}_n(x_j) - \bar{\sigma}_n(x_j)}{\bar{\sigma}_n(x_j)})]$$

$$= \Sigma_{j=1}^{n} w_j(x;h)\frac{\bar{\sigma}_n(x)}{\bar{\sigma}_n(x_j)}E[e_j] + \Sigma_{j=1}^{n} w_j(x;h)\frac{\bar{\sigma}_n(x)}{\bar{\sigma}_n(x_j)}E[e_j] - \Sigma_{j=1}^{n} w_j(x;h)\frac{\bar{\sigma}_n(x)\sigma(x_j)}{\bar{\sigma}(X_j)^2}E[\bar{\sigma}_n(X_j)]$$

$$+\Sigma_{j=1}^{n} w_j(x;h)E[\frac{\bar{\sigma}_n(x)\sigma(x_j)}{\bar{\sigma}(X_j)}] \tag{3.29}$$

$$[\Sigma_{j=1}^{n} w_j(x;h)\frac{\bar{\sigma}_n(x)}{\bar{\sigma}_n(x_j)}(e_j + \sigma(X_j)((b_n)(x) - b_n(X_j))] = 0 + 0 - \Sigma_{j=1}^{n} w_j(x;h)\frac{\bar{\sigma}_n(x)\sigma_n(X_j)\bar{\sigma}(X_j)}{\bar{\sigma}(X_j)^2}$$

$$+\Sigma_{j=1}^{n} w_j(x;h)\frac{\bar{\sigma}_n(x)\sigma_n(X_j)}{\bar{\sigma}(X_j)}E[1] \tag{3.30}$$

$$E[\Sigma_{j=1}^{n} w_j(x;h)\frac{\bar{\sigma}_n(x)}{\bar{\sigma}_n(x_j)}(e_j + \sigma(X_j)((b_n)(x) - b_n(X_j))]$$

$$= 0 + 0 - \Sigma_{j=1}^{n} w_j(x;h)\frac{\bar{\sigma}_n(x)\sigma_n(X_j)}{\bar{\sigma}(X_j)} + \Sigma_{j=1}^{n} w_j(x;h)\frac{\bar{\sigma}_n(x)\sigma_n(X_j)}{\bar{\sigma}(X_j)} \tag{3.31}$$

Thus

$$E[\Sigma_{j=1}^{n} w_j(x;h)\frac{\bar{\sigma}_n(x)}{\bar{\sigma}_n(x_j)}(e_j + \sigma(X_j)((b_n)(x) - b_n(X_j))] = 0 \tag{3.32}$$

Analyzing the third term of equation (3.26)

$$E[\Sigma_{j=1}^{n} w_j(x;h)\frac{\bar{\sigma}_n(x)}{\bar{\sigma}_n(x_j)}e_j((b_n)(x) - b_n(X_j))] =$$

$$E(\Sigma_{j=1}^{n} w_j(x;h)[\frac{\bar{\sigma}_n(x)}{\bar{\sigma}_n(x_j)}e_j\frac{\bar{\sigma}_n(x) - \bar{\sigma}_n(x)}{\bar{\sigma}_n(x)}] - \frac{\bar{\sigma}_n(x)}{\bar{\sigma}_n(x_j)}e_j(\frac{\bar{\sigma}_n(x_j) - \bar{\sigma}(x_j)}{\bar{\sigma}_n(x_j)})] \tag{3.33}$$

$$E[\Sigma_{j=1}^{n} w_j(x;h)\frac{\bar{\sigma}_n(x)}{\bar{\sigma}_n(x_j)}e_j((b_n)(x) - b_n(X_j))] =$$

$$\Sigma_{j=1}^{n} w_j(x;h)\frac{\bar{\sigma}_n(x)}{\bar{\sigma}_n(x_j)}[\frac{\bar{\sigma}_n(x) - \bar{\sigma}_n(x)}{\bar{\sigma}_n(x)}]E[e_j] - \Sigma_{j=1}^{n} w_j(x;h)\frac{\bar{\sigma}_n(x)}{\bar{\sigma}_n(x_j)}[\frac{\bar{\sigma}_n(x_j) - \bar{\sigma}_n(x_j)}{\bar{\sigma}_n(x_j)}]E[e_j] \tag{3.34}$$

Therefore

$$E[\Sigma_{j=1}^{n} w_j(x;h)\frac{\bar{\sigma}_n(x)}{\bar{\sigma}_n(x_j)}e_j((b_n)(x) - b_n(X_j))] = 0 \tag{3.35}$$

equation (3.26)thus reduces to

$$E[\hat{\sigma}_n(x_i)] = \Sigma_{j=1}^{n} w_j(x;h)\frac{\bar{\sigma}_n(x)}{\bar{\sigma}_n(x_j)}\sigma(X_j) + O_p(\frac{1}{nh}) \tag{3.36}$$

Thus $E[\hat{V}_{MBC}]$ is given by the expression

$$E[\hat{V}_{MBC}] = \Sigma_{ies}\bar{Y}_i + \Sigma_{ie(p-s)}[\Sigma_{j=1}^{n} w_j(x;h)\frac{\bar{\sigma}_n(x)}{\bar{\sigma}_n(x_j)}\sigma(X_j)] + O_p(\frac{1}{nh}) \tag{3.37}$$

Equation (3.36) can be simplified by taking a Taylor series expansion of the ratio $\frac{\bar{\sigma}_n(x)}{\bar{\sigma}_n(x_j)}$ about the point x as follows.

$$\frac{\bar{\sigma}_n(x)}{\bar{\sigma}_n(x_j)} = \frac{\bar{\sigma}_n(x)}{\bar{\sigma}_n(x_j)} + (X_j - x)(\frac{\sigma(x)}{\bar{\sigma}_n(x)})' + \frac{1}{2}(X_j - x)^2(\frac{\sigma(x)}{\bar{\sigma}_n(x)})'' + (1 + O_p) \tag{3.38}$$

Substituting equation (3.38) to equation (3.37) yields

$$E[\hat{V}_{MBC}] = \Sigma_{ies}\bar{Y}_i + \Sigma_{ie(p-s)}[\Sigma_{j=1}^{n} w_j(x;h)\bar{\sigma}_n(x)(\frac{\sigma(x)}{\bar{\sigma}_n(x)} + (X_j - x)(\frac{\sigma(x)}{\bar{\sigma}_n(x)})'$$

$$+\frac{1}{2}(X_j - x)^2\frac{\sigma(x)}{\bar{\sigma}_n(x)})'' + (1 + O_p))] + O_p(\frac{1}{nh}) \tag{3.39}$$

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

$$E[\hat{V}_{MBC}] = \Sigma_{ies}\bar{Y}_i + \Sigma_{ie(p-s)}[\Sigma_{j=1}^{n} w_j(x;h)\hat{\sigma}_n(x)(\frac{\sigma(x)}{\hat{\sigma}_n(x)} + (X_j - x)(\frac{\sigma(x)}{\hat{\sigma}_n(x)})')] + O_p(\frac{1}{nh}) \tag{3.40}$$

since $\Sigma_{j=1}^{n} w_j(x;h) = 1$ and $\Sigma_{j=1}^{n} w_j(x;h)(X_j - x) = 0$,equation (3.40)can be expressed as

$$E[\hat{V}_{MBC}] = \Sigma_{ies}\bar{Y}_i + \Sigma_{ie(p-s)}[\Sigma_{j=1}^{n} w_j(x;h)\sigma(x)] + O_p(\frac{1}{nh}) \tag{3.41}$$

Also we have,

$$T = \Sigma_{ies}Y_i + \Sigma_{ie(p-s)}Y_i \tag{3.42}$$

$$E[T] = E[\Sigma_{ies}Y_i + \Sigma_{ie(p-s)}Y_i] \tag{3.43}$$

Substituting equation (3.41) and equation (3.43) into equation (3.23)yields,

$$E[\hat{V}_{MBC} - T] = \Sigma_{ies}Y_i + \Sigma_{ie(p-s)}[\Sigma_{j=1}^{n} w_j(x;h\sigma(x)] + O_p(\frac{1}{nh}) - [\Sigma_{ies}Y_i + \Sigma_{ie(p-s)}\sigma(x)] \tag{3.44}$$

$$E[\hat{V}_{MBC} - T] = \Sigma_{ie(p-s)}[\Sigma_{j=1}^{n} w_j(x;h)\sigma(x)] - \Sigma_{ie(p-s)}\sigma(x) + O_p(\frac{1}{nh}) \tag{3.45}$$

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

$$Bias[\frac{\hat{V}_{MBC}}{N}] = E[\frac{\hat{V}_{MBC} - T}{N}] = \frac{1}{N}[\Sigma_{ie(p-s)}\Sigma_{j=1}^{n}(w_j(x;h)\sigma(x)) - \Sigma_{ie(p-s)}\sigma(x)] + O_p(\frac{1}{nh}) \tag{3.46}$$

The bias of $\hat{V}_{MBC}$ will be of order $O_p(\frac{1}{nh})$.Thus it converges to 0 at a faster rate compared to the existing non-parametric estimators which generally converge at the rate of $O_p(h^2)$

3.6 Asymptotic Variance of the MBC estimator

Using equation (3.21),the estimator of the robust variance is given by;

$$\hat{V}_{MBC} = \Sigma_{ies}Y_i + \Sigma_{ie(p-s)}[\Sigma_{j=1}^{n} w_j(x;h)\frac{\hat{\sigma}_n(x)}{\hat{\sigma}_n(x_j)} + \Sigma_{j=1}^{n} w_j(x;h)\frac{\hat{\sigma}_n(x)}{\hat{\sigma}_n(x_j)}(e_j$$

$$\sigma(x_j)[b_n(x) - b_n(x_j)]) + \Sigma_{j=1}^{n} w_j(x;h)\frac{\hat{\sigma}_n(x)}{\hat{\sigma}_n(x_j)}e_j[b_n(x) - b_n(x_j)] + r_j(x, X_j) \quad (3.47)$$

where $r_j(x, X_j)$ is the remainder term that involves the terms x and X_j using the assumption that $nh \to \infty$,the terms converge to zero in probability. Therefore $r_j(x, X_j)[\sigma(X_j) + e_j] = O_p(\frac{1}{nh})$ and equation (3.47) reduces to

$$\hat{V}_{MBC} = \Sigma_{ies}Y_i + \Sigma_{ie(p-s)}[\Sigma_{j=1}^{n} w_j(x;h)\frac{\hat{\sigma}_n(x)}{\hat{\sigma}_n(x_j)} * [\hat{\sigma}_n(x_j) + e_j][1 + b_n(x) - b_n(x_j)] + O_p(\frac{1}{nh}) \quad (3.48)$$

Truncating the binomial expansion at the first term yields

$$\hat{V}_{MBC} = \Sigma_{ies}Y_i + \Sigma_{ie(p-s)}[\Sigma_{j=1}^{n} w_j(x;h)\frac{\hat{\sigma}_n(x)}{\hat{\sigma}_n(x_j)} * [\hat{\sigma}_n(x_j) + e_j] + O_p(\frac{1}{nh}) \quad (3.49)$$

The variance of the estimator is then defined by;

$$Var[\hat{V}_{MBC}] = Var[\Sigma_{ies}Y_i + \Sigma_{ie(p-s)}[\Sigma_{j=1}^{n} w_j(x;h)\frac{\hat{\sigma}_n(x)}{\hat{\sigma}_n(x_j)} * [\hat{\sigma}_n(x_j) + e_j] + O_p(\frac{1}{nh})] \quad (3.50)$$

$$= Var[\Sigma_{ies}Y_i] + Var[\Sigma_{ies}Y_i] + Var[\Sigma_{ie(p-s)}[\Sigma_{j=1}^{n} w_j(x;h)\frac{\hat{\sigma}_n(x)}{\hat{\sigma}_n(x_j)} * [\hat{\sigma}_n(x_j) + e_j] + (O_p(\frac{1}{nh}))^2 \quad (3.51)$$

$$Var[\hat{V}_{MBC}] = Var[\Sigma_{ies}Y_i] + [\Sigma_{ie(p-s)}\Sigma_{j=1}^{n} w_j(x;h)\frac{\hat{\sigma}_n(x)}{\hat{\sigma}_n(x_j)} * Var(Y_j)] + (O_p(\frac{1}{nh}))^2 \quad (3.52)$$

$$Var[\hat{V}_{MBC}] = \Sigma_{ies}\sigma^2(x_i) + \Sigma_{ie(p-s)}\Sigma_{j=1}^{n}(w_j(x;h))^2(\frac{\hat{\sigma}_n(x)}{\hat{\sigma}_n(x_j)})^2\sigma^2(x_i) + O_p(\frac{1}{nh}))^2 \quad (3.53)$$

Obtaining the Taylor series expansion of the ratio $\frac{\sigma^2(x_j)}{\sigma(x_j)^2}$ in the second part of the equation (3.53) gives

$$Var[\hat{V}_{MBC}] = \Sigma_{ies}\sigma^2(x_i) + \Sigma_{ie(p-s)}\Sigma_{j=1}^{n}(w_j(x;h))^2\sigma^2(x_i) + O_p(\frac{1}{nh}))^2 \quad (3.54)$$

Thus the asymptotic variance of $[\frac{\hat{V}_{MBC}}{N}]$ is given by;

$$Var[\frac{\hat{V}_{MBC}}{N}] = \frac{1}{N^2}\Sigma_{ies}\sigma^2(x_i) + \frac{1}{N^2}\Sigma_{ie(p-s)}\Sigma_{j=1}^{n}(w_j(x;h))^2\sigma^2(x_i + O_p(\frac{1}{nh}))^2 \quad (3.55)$$

3.7 Asymptotic Mean squared of the MBC estimator

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

$$MSE[\hat{V}_{MBC}] = Var[\frac{\hat{V}_{MBC}}{N}] + [Bias[\frac{\hat{V}_{MBC}}{N}]^2] \tag{3.56}$$

$$Bias[\frac{\hat{V}_{MBC}}{N}] = E[\frac{\hat{V}_{MBC} - T}{N}] = O_p(\frac{1}{nh}) \tag{3.57}$$

using equation (3.55) and (3.57) in equation (3.56) yields

$$MSE[\hat{V}_{MBC}] = \frac{1}{N^2}[\Sigma_{ies}\sigma^2(x_i) + \Sigma_{ie(p-s)}\Sigma_{j=1}^{n}(w_j(x;h))^2\sigma^2(x_i)] + O_p(\frac{1}{nh}))^2] \tag{3.58}$$

As $n \to \infty$ and $h \to 0$, the mean squared error in equation (3.58) tends to zero. This shows that the estimator $\hat{V}_{MBC}$ is asymptotically consistent and thus very useful.

Chapter 4

Empirical Study

4.1 Introduction

This chapter presents the description of the population and the results of the asympotic variance estimators.

4.2 Simulation Procedure

A simulation experiment is performed in order to investigate the statistical properties of the proposed estimator as well as compare its performance to that studied by Otieno and Mwalili (2000). The unconditional MSE and Biases are computed for each of the variance estimators. R statistical programming version R.2.12.1 is employed to simulate the coverage probabilities using randomly selected samples of size n=50, n=100, n=200, n=500 and n=1000. The samples of n=50, n=100, n=200, n=500 and n=1000 are generated using simple random sampling without replacement. A comparison of the performance of the proposed estimator and the variance estimators studied by Otieno and Mwalili (2000) over the randomly selected samples is done. The biases of the proposed variance estimator and those studied by Otieno and Mwalili (2000) are calculated. The Root Mean squared Error(RMSE) for the proposed variance estimators is calculated as given below,

$$RMSE(V_{MBC}) = \sqrt{\frac{1}{N}(\hat{V}_{MBC} - T)^2}$$

The results are displayed in a table and was compared to the following rival estimators,

$$RMSE(V_N) = \sqrt{\frac{1}{N}(\hat{V}_N - T)^2}$$

$$RMSE(V_C) = \sqrt{\frac{1}{N}(\hat{V}_C - T)^2}$$

$$RMSE(V_L) = \sqrt{\frac{1}{N}(\hat{V}_L - T)^2}$$

4.3 Simulation Results

Table 4.1 represents the unconditional biases and root Mean Squared Errors (RMSE) for the multiplicative bias corrected estimator and those studied by Otieno and Mwalili (2000). Table 4.1 show that the bias and the RMSE of the multiplicative bias corrected estimator is lower than the bias and mean squared error of the variance estiamtors studied by Otieno and Mwalili(2000).

Table 4.1: Unconditional biases and RMSE from simulated data

Estimator	Bias	RMSE
V_C	0.37235	25258.25
V_L	0.45917	19945.48
V_N	0.19287	19856.87
V_{MBC}	0.01432	18872.54

Table 4.2 gives a comparison of the coverage probabilities of the four variance estimators for different population sizes. The coverage probabilities for the Multiplicative bias corrected estimator is closer to the nominal value of 0.95 than are the coverage probabilities for the other three rival variance estimators. The variance estimator V_L has a better coverage ability under a small sample size of n=50.

Table 4.2: Summary of the unconditional coverage probabilities from the simulted data.

Sample size	V_{MBC}	V_N	V_L	V_C
n=50	0.905	0.885	0.936	0.899
n=100	0.892	0.792	0.685	0.714
n=200	0.919	0.895	0.562	0.435
n=500	0.952	0.915	0.654	0.642
n=1000	0.948	0.929	0.643	0.742

4.4 Real data analysis of the population

A population of size 188 was obtained from United Nations development Programme 2015. UN studied development in 188 countries. UN grouped development in the countries as either very high human development, high human development, medium human development or low human development. Kenya tops the list in countries under low medium development as per the UN statistics 2015 and ranks at number 145 among the 188 countries studied. The UN study used Human Development Index (HDI), Life expectancy at birth, Expected years of schooling, Mean years of schooling, Gross National Income (GNI) per capita and GNI per capita rank minus HDI to rank human development index in the 188 countries.

In this study a relationship between HDI and GNI is considered. HDI acts as the auxiliary variable while GNI acts as the variable under study $y_i(i = 1, 2, 3, ..., 188)$. Figure 4.1 gives a scatter plot of HDI against GNI and also demonstrates a line of best fit between GNI and HDI.

From the scatter plot we observe a quadratic relationship between HDI and GNI. Fitting a linear regression model to the data we obtain a linear model equation of the form $GNI = 89490HDI - 44953$. The correlation coefficient between GNI and HDI is approximately 0.74. This indicates a strong positive linear relation between GNI and HDI.

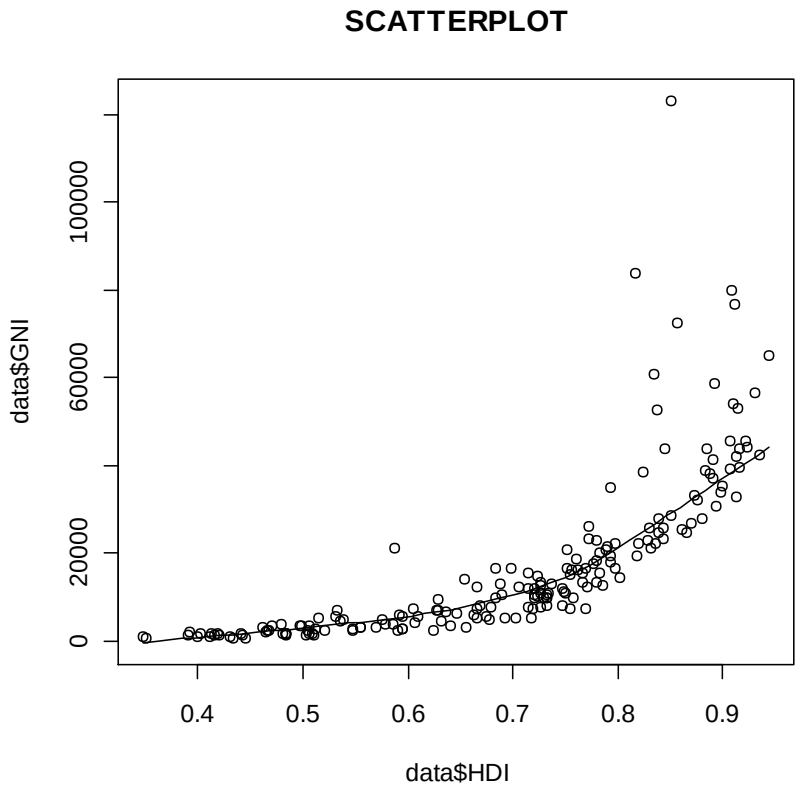

Figure 4.1: Scatter diagram for the population

Table 4.3 presents the unconditional biases and RMSE for the multiplicative bias corrected estimator and those studied by Otieno and Mwalili (2000) from the actual data set.

Table 4.3: Unconditional biases and RMSE from actual data

Estimator	Bias	RMSE
V_C	0.56432	29342.45
V_L	0.32927	20045.54
V_N	0.189287	19756.67
V_{MBC}	0.01498	18986.35

Employing R statistical programming version R.2.12.1, simulate the coverage probabilities using randomly selected samples of size n=50, n=100, n=200, n=500 and n=1000 from the UN data. Table 4.4 gives a summary of the results of coverage probabilities using a nominal value of 0.95.

Table 4.4: Summary of the unconditional coverage probabilities from the actual data set

Sample size	V_{MBC}	V_N	V_L	V_C
n=50	0.912	0.798	0.942	0.832
n=100	0.862	0.763	0.621	0.798
n=200	0.934	0.821	0.451	0.524
n=500	0.962	0.926	0.689	0.653
n=1000	0.949	0.931	0.664	0.792

4.5 Comparison of simulated data and real data

The proposed variance estimator V_{MBC} has a small bias and root mean squared error as compared to the estimators derived by Royall and Cumberland (1981) and Otieno and Mwalili (2000) as observed from the simulated and real data set analysis. The bias of the proposed estimator V_{MBC} is almost close to zero. Thus the proposed variance estimator V_{MBC} is asymptotically consistent and can be recommended in practical applications. From both the simulated and actual data,V_{MBC} has a better coverage probability closer to the nominal value of 0.95 as compared to the other rival variance estimators.

4.6 Conclusion

The main objective of this study was to construct a robust variance estimator of the ratio estimator of the population mean using a multiplicative bias correction technique. As a way of achieving this, a pilot smoother was utilized and the resulting variance estimator using the multiplicative bias correction technique was found to be a useful tool in correcting of boundary bias. The methodology used possesses a kind of robustness in the sense that the multiplicative factor is bounded. The method is easy to implement and has good asymptotic properties both theoretically and practically. We have also assessed the performances of our estimator with that of the existing estimators for simulated data and real data sets. It is observed from the numerical comparison that our proposed estimator V_{MBC} is more efficient than existing variance estimators. The derived estimator V_{MBC} has better coverage probability as compared to rival variance estimators of the population mean.

4.7 Recommendation for further study

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

Independence of survey variables y_i and y_j was assumed in the study of asymptotic properties of the estimator derived. The investigation of the nature of the results if dependence of the observations is still an open area for study.

References

Chambers, R.L., Dorfman, A.H.Wehrly, T.E.(1993). Bias robust estimation in finite populations using non parametric calibration. Journal of the American Statistical Association $88 : 268 - 277$.

Cochran, W.G.(1977) Sampling Techniques. Edn. 3. Wiley Eastern Publication, New York.

Hansen, M. H. , Hurwitz, W. N. and Madow, W.O.(1983). An evaluation of model dependent and probability sampling inference in sample surveys. Journal of the American Statistical Association 78:776.-807.

Otieno and Mwalili (2000). Nonparametric regression method for estimating the error variance in unistage sampling. East African Journal of Science 2(2):107-112 (2000).

Nadaraya, E.pt.(1964). On estimating regression. Theory of Probability Applications 9: 141-142.

Royall, R. M.and Cumberland, W.O.(1981). An empirical study of the ratio estimator and estimators of its variance. Journal of the American Statistical Association 76:66-77.

Silverman, (1986). Density Estimation.Chapman and Hall, London.

J.Subramani and G.Kumarapandiyan (2015). A class of ratio estimators for estimation of population variance jamsi, 11 (2015), No.1

Dorfman, A.H. (2009) Nonparametric Regression and the Two Sample Problem. Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods, Washington DC, August 1-6 2009, 277-270.

Wolter, K.M. (1985). Introduction to Variance Estimation. New York: Springer-Verlag.

Quenouille, M. H. (1949). Approximate tests of correlation in time-series. J.R. Statistic. Soc. B 11, 68-84.

Linton, O. and Nielsen, J.P. (1994). A Multiplicative Bias Reduction Method for Nonparametric Regression. Statistics and Probability Letters , 19, 181-187.

Dorfman, A.H. (1992) Nonparametric Regression for Estimating Totals in Finite Populations. Proceedings of the Section on Survey Research Methods , American Statistical Association Alexandria, Washington DC, 622-625.

Brewer, K. R. W., Ratio estimation and finite populations: Some results deducible from the assumption of an underlying stochastic process, Australian Journal of Statistics 5 (1963), 93-105.

Wu, C. F., Estimation of variance of the ratio estimator, Biometrika 69 (1982), 183-189.

Isaki, C.T. (1983). Variance estimation using auxiliary information. Journal of the American Statistical Association, 78:117-123.

Upadhyaya, L.N. and Sing, H.P. (2001). Estimation of population standard deviation using auxiliary information. American Journal of Mathematics and Management Sciences, 21(3-4), 345-358.

Royall, R. M. and Eberhardt, K. R. (1975). Variance estimates for the ratio estimator, Sankya, Ser. C 37, 43-52.

Royall, R. M. (1970). On finite population sampling theory under certain linear regression models. Biometrika 57, 377-387.

Krewski, D. ; Rao, J. N. K. (1981). Inference From Stratified Samples: Properties of the Linearization, Jackknife and Balanced Repeated Replication Methods. Ann. Statist no. 5, 1010–1019.

Priestly, M. B. and Chao, M. T. (1972). Non-Parametric Function Fitting. Journal of the Royal Statistical Society, Series B, 34, 385–392.

Gasser, T. and Müller, H (1981). Kernel Estimation of Regression Functions. In Smoothing Techniques for Curve Estimation Springer-Verlag, pp. 23-68, 1979.

Dickey, D. and Fuller W. (1981). Likelihood Ratio Statistics for Autoregressive Time Series with a Unit Root .Econometrica, 49: 1057-1072.

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

C.F. Jeff Wu and L.Yuan .Deng (1982).Estimation of Variance of the Ratio Estimator: An Empirical Study. Proceedings of a Conference Conducted by the Mathematics Research Center, the University of Wisconsin–Madison, November 4–6, 1981

Appendix I: Multiplicative Bias Correction Simulation

$> data < -read.delim(file.choose(), header = T)$

$> data$
$xbar = mean(HDI)$

$xbar$

$sumy = sum(GNI)$

$> linearMod < -lm(GNI\ HDI)$

$> linearMod < -lm(data\text{GNI} \quad \text{data}HDI)$

$> print(linearMod)$

$Call:$

$lm(formula = data\text{GNI} \quad \text{data}HDI)$

$Coefficients:$

$(Intercept)dataHDI$

$> scatter.smooth(x = data\text{HDI, y=data}GNI, main = "SCATTERPLOT")$

$> cor(data\text{HDI, data}GNI)$

Appendix II: Conditional Bias Regression

$> n < -50$

$> data$

$> mse < -numeric(n)$

$> bias < -numeric(n)$

$> variance < -numeric(n)$

$> for(iin1:n)$

$+mse[i] < -MSE((1-data[i])*Z, mu)$

$+bias[i] < -mu*data[i]$

$+variance[i] < -(1-data[i])^2$

$plot(x, mse, type = "o", col = "blue", pch = "o", lty = 1, ylim = c(0,1000), xlab = "(Groupmean)/100", ylab = "Averagesquaredpredictorerror", main = "AverageSquaredPredictorError")$

$> lines(x, vc, type = "o", col = "red", lty = 2)$

$> lines(x, vl, type = "o", col = "green", lty = 3)$

$> lines(x, vn, type = "o", col = "orange", lty = 4)$

$> lines(x, vmbc, type = "o", col = "black", lty = 5)$

$legend("bottom", legend = c("MSE","VC","VL","VN","VMBC"), col = c("blue","red","green","orange","black"), lty = 1:5, cex = 0.8)$

Appendix III: Coverage probabilities

$N = c(50, 100, 200, 500, 1000)$

$CP = array(0, length(N))$

$for(kin1:length(N))$

$[n = N[k]$

$X = array(rexp(n*1000), c(1000, n))$

$M = apply(X, 1, mean)$

$MX = apply(X, 1, max)$

$MN = apply(X, 1, min)$

$C = (MX - MN)/(2*sqrt(n))$

$ci = ((M - C < 1)(M + C > 1))$
CP[k] = mean(ci)