COMPARISON OF THE PERFORMANCE OF ESTIMATORS IN ESTIMATION OF FINITE POPULATION TOTAL

BY

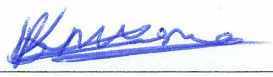
BENJAMIN K. MUEMA

Dissertation submitted for the partial fulfillment of the degree of master of science in mathematics (statistics) of Kenyatta University

JULY 2002

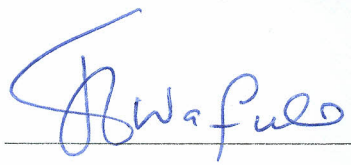
DECLARATION

This research project is my original work and has not been presented for a degree award in any other University.



Benjamin K. Muema

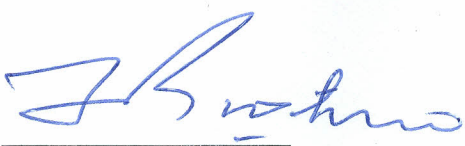
This work has been presented with our approval as the University Supervisors



DR. CHARLES WAFULA

Department of Mathematics

Kenyatta University



DR. ROMANUS ODHIAMBO

Department of Mathematics and Statistics

Jomo Kenyatta University of Agriculture and Technology

DEDICATION

This work is dedicated to my beloved Parents, **Mr. Daniel Muema Kalii** and **Mrs. Agnes Muema** who have made me to realize the virtues of hard work.

ABSTRACT

In this project we compare the performance of two different estimators of the population total. One estimator is model-based and the other one is model assisted. We look at model-based properties of the two estimators. We observe that under the general model, the biases of the two estimators are different.

ACKNOWLEDGEMENT

First and foremost my sincere thanks goes to the Almighty God, the protector, the source of all kind of knowledge and the provider of everything.

It is my pleasure to acknowledge debts to whoever has assisted in the preparation of this research project.

I would like to express my profound gratitude to my supervisor, Dr. Charles Wafula for his professional guidance, availability, patience and encouragement throughout the course of the study.

It will be difficult to forget academic guidance, encouragement and material support I received from Dr. Romanus Odhiambo, Department of Mathematics and Statistics, J.K.U.A.T. It will be impossible to clear the debt of gratitude I owe to him.

My sincere appreciation goes to my Parents for their material support, devoted patience, constant prayers and encouragement. I am also grateful to my brothers and sisters for their moral and material support.

My appreciation also goes to my former lecturers, especially Dr. Njenga, Dr. Odongo, Dr. Kahiri, Mr. Ruto and the Late Mr. Githu. To my coursemates in

the struggle, Kasungo and Karuku, your criticisms and suggestions during our course in postgraduate studies are highly appreciated.

I am also grateful to Robert Nderitu of Central Bureau of Statistics and Tobias Mwalili of software engineering for their technical assistance.

Special thanks goes to Zippy for her forbearance, encouragement and constant prayers throughout the struggle.

TABLE OF CONTENTS

CHAPTER ONE

INTRODUCTION AND LITERATURE REVIEW

1.1 INTRODUCTION

In sample surveys, the aim is to obtain the desired information from a population, which might be finite or infinite. We have two major ways of gathering the information namely census and sampling. Census involves carrying out a complete enumeration where by the whole population is observed. The second method employs a scientific mathematical process of carrying out the survey. In this process, a part of the population, referred to as a sample is used to make inference about the whole population.

The two methods are employed depending on the complexity of the survey to be carried out. Generaly, census is assumed to be more accurate than the sampling method but it is not applicable when the target population is large. This is due to certain constraints involved like time, cost, literacy and other geographical factors. These factors are prone to error, which may result to low precision. So the statistician prefers sampling method rather than census because it is deemed to be appropriate especially when dealing with complex surveys. In this chapter we briefly discuss the sample survey problem and major approaches to

inference in sample surveys. We also give the objectives and the outline of the project.

1.2 SAMPLE SURVEY PROBLEM

The survey problem is two fold:

i) How to choose the sample.

ii) How to use the observed sample values to estimate the finite population parameters.

The former is the design problem while the latter is the estimation problem.

A finite population is a collection of identifiable N units, $U = (u_1, u_2, \ldots\ldots u_N)$ where $N < \infty$, is the size of the population. Corresponding to each unit $i \in U$, is a vector of survey variables with values Y_i, $i = 1, 2, \ldots, N$.

In design stage, the sample surveyor employs a mechanism of getting a sample from the finite population. Any auxiliary information will help in choosing the sampling scheme.

In estimation stage, we use a statistic $\hat{g}(\underline{Y}_s)$ which is a function of the sample to estimate some function $G(\underline{Y})$ of the vector $Y = (Y_1, Y_2, \ldots\ldots Y_N)'$. The function $G(\underline{Y})$ can be:

a) Finite population total, $T = \sum_{i=1}^{N} Y_i$

b) Finite population mean, $\bar{T} = \frac{\sum_{i=1}^{N} Y_i}{N}$

c) Finite population variance, $\text{var}(T) = \frac{\sum_{i=1}^{N} (Y_i - \bar{Y})^2}{N}$ etc.

In this project, the focus is on the estimation problem and in particular, the estimation of finite population totals. We assume the sample has already been chosen and sample values of the characteristic of interest already obtained.

1.3 MAJOR APPROACHES TO INFERENCE IN SAMPLE SURVEYS

Despite the complexity and intensiveness of the literature on inference in sample surveys, generally we have two distinguishable approaches, namely:

i) Randomization approach (design-based approach);

ii) Model-based approach (super population approach).

From these two approaches, a third approach called model-assisted approach which combines the two has been suggested in the literature.

1.3.1 The Randomization Approach

In this approach, a randomizer comes up with a design. We note that there is no rule in choosing the design hence the choice is individualistic. Once a sample design has been chosen, we get a procedure of drawing samples of sizes n repeatedly. The framework of randomization inference is formed from the repetition of this procedure.

The values of the study variable $\underline{Y}$ are assumed to be unknown constants and the only probabilities are those from the sampling design, which is exactly known. To make inference, an estimator must be defined and the distribution of this estimator over the repeated samples must be evaluated. Otherwise, our interest in this project is not in randomization approach but in model-based approach and model-assisted approach.

Despite the proof by Godambe (1955) on non-existence of a uniformly minimum variance unbiased estimator for all possible class of populations Y, the concept of unbiasness is used to judge the goodness of an estimator. We can make an improvement on this by considering other factors like consistency and efficiency of the estimator.

1.3.2 The Model-Based Approach

Here, we start by assuming that the unknown vector $\underline{Y}$ is a random variable, hence the concept of randomization is introduced into the Y values directly. Further, we assume that the vector $\underline{Y}$ is generated as a sample from a super population.

The major problem in this approach is how to come up with a parametric probability model that relates the auxiliary variable X and the survey variable. After specification of the model, a sample selection scheme is employed to draw a sample. Hence we make predictive inference about the unobserved random variables $\underline{Y}_i$, where $i \notin s$ (s being the sample), by making use of the sample data, the selected model and the information in the sample scheme used. We estimate functions of the finite population values such as the population total T, which is indeed the core issue in this project. Instead of choosing a specific model, we can also choose a general model, for example

$$Y = m(x) + \sigma(x)\ell \qquad (1.1)$$

Where $m(\cdot)$ and $\sigma(\cdot)$ are smooth functions and the ℓ_i independently distributed with mean 0 and constant variance. We discuss this in more details in chapter two.

1.3.3 The Model-Assisted Approach

According to the argument in the literature, the above two approaches to inference are statistically valid for many finite populations. Now a very challenging question is, which of the two positions is better? Intuitively, there is no strategy between the two that is better than the other simply because each method has its own merits and demerits depending on the complexity of the survey to be considered, see Hansen et al (1983), Smith (1983), Little (1982), Royall and Pferffermann (1982).

A suggestion has been made in the literature, where the two approaches are blended to come up with a mode-assisted approach. The resulting estimator (model-assisted estimator) has both characteristics of the design and the model. It could be design-unbiased and model unbiased. The distribution of this estimator forms the framework for inference in this integrated approach. This idea of model-assisted approach will be discussed in chapter three where we develop an estimator based on it for the population total and investigate its properties.

1.4 OBJECTIVES AND OUTLINE OF THE PROJECT

1.4.1 Objectives of the Project

In this project, we investigate the model-based properties of a model-based estimator and a model-assisted estimator for estimating a finite population total. In particular, we study their biases and variances. After investigating their properties, we compare their performance in an empirical study.

1.4.2 Outline of the Project

In chapter two, we consider the model-based approach in estimation of a finite population total. In particular we consider the application of non-parametric regression in estimating a finite population total.

In chapter three, we look at the model-assisted approach, which incorporates the Horvitz-Thompson estimator in the local polynomial set up. Here we introduce a new estimator, which does not have inclusion probability like the one suggested by Opsomer and Breidt (2000).

In chapter four we carry out an empirical study to compare the performance of these estimators and lastly in chapter five, we suggest areas for further study and give conclusions of the project

CHAPTER TWO

MODEL - BASED APPROACH

2.1 INTRODUCTION

In this chapter we consider the work done by Dorfman (1992) on application of non parametric regression to the estimation of a finite population total based on a sample from the population. From a population P of N identifiable units with study variable Y and sample values available, our aim is to estimate the population total $T = \sum_{p} Y_i$. Another work parallel to this may be found in chambers et al (1992), Dorfman and Hall (1992). Smith and Njenga (1992) applied non-parametric regression to the estimation of superpopulation parameters.

Generally in this approach we use a model, which is either specific or general. A specific model usually leads to the problem of robustness. Hansen, Madow and Tepping (1983) discusses the problem of robustness in more details.

2.2 NONPARAMETRIC REGRESSION

The concept of non-parametric regression goes back to Nadaraya (1964) and Watson (1964). Recently, we have a reference on this by Hardle (1991). In this chapter we consider the simple Nadaraya Watson kernel estimation although there are many types on this. We assume that the auxiliary information is available

for the entire population and the auxiliary variable X and study variable Y are related in a more general way, i.e. the relationship between them is not strong. We consider the general model (1.1) throughout.

The studies of the properties of the proposed estimator are conditional on the available sample and non sample values of the auxiliary variable X. We assume that sample and non sample X's are independent random variables with densities $d_s(x)$ and $d_{p-s}(x)$ respectively.

To estimate $m(x)$ in model (1.1), one method is to average the nearby values of Y_i where " nearby" is measured in terms of the distance $|x_i - x|$. Let $K_b(u) = b^{-1}K(u/b)$ where k(u) is kernel ,b is bandwidth and weight sequences for Kernel smoothers (one dimensional x)

$$w_i = \frac{K_b(x_i - x)}{\sum_{i=1}^{n} K_b(x_i - x)}$$

The Nadaraya – Watson estimator of $m(x)$ is

$$\hat{m}(x) = \sum_i w_i(x) Y_i \qquad \text{---------- (2.1)}$$

2.3 NONPARAMETRIC REGRESSION - BASED ESTIMATOR OF THE FINITE POPULATION TOTAL.

We let $x = x_j$ be any point in the non sample and estimate $m(x_j)$. Then the nonparametric regression –based estimator, $\hat{T}_{np}$, for the population total T is given by

$$\hat{T}_{np} = \sum_{s} Y_i + \sum_{p-s} \hat{m}(x_j)$$

and the prediction error is given by

$$\hat{T}_{np} - T = \sum_{p-s} \left(\hat{m}(x_J) - Y_j\right) \quad \text{.....(2.2).}$$

In the following subsection we verify the conditional mean and variance of $\hat{T}_{np} - T$ under model (1.1) as obtained by Dorfman (1992).

2.4 CONDITIONAL MEAN OF $\hat{T}_{np} - T$

From (2.2),

$$E\left[\hat{T}_{np} - T/X_p\right] = \sum_{p-s}\left[\sum_s w_i(x_j)m(x_i) - m(x_j)\right]$$

$$= \sum_{j\in p-s}\left[\frac{\sum_{i\in s}\frac{1}{nb}k\left(\frac{x_i - x_j}{b}\right)(m(x_i))}{\sum_{i\in s}\frac{1}{nb}k\left(\frac{x_i - x_j}{b}\right)} - m(x_j)\right]$$

$$= \sum_{j\in p-s}\hat{d}_s(x_j)^{-1}(nb)^{-1}\sum_{i\in s}\left[k\left(\frac{x_i - x_j}{b}\right)(m(x_i) - m(x_j))\right] \quad \ldots\ldots\ldots\ldots\ldots \quad (2.3)$$

where

$$\hat{d}_s(x_j) = (nb)^{-1}\sum_{i\in s}k\frac{(x_i - x_j)}{b}$$

is the kernel estimator of $d_s(x_j)$.

From (2.3) the relationship between the conditional mean and the selected bandwidth cannot be established. Consequently we rewrite (2.3) in a different form. To do this, we utilize the following assumptions and theorem.

Assumptions (2.0)

Let $k(u)$ be a symmetric density function with $\int uk(u)du = 0$ *and* $k_2 = \int u^2 k(u)du > 0$, assume n and N increase together such that $n/N \to \pi$, with $0 < \pi < 1$; assume sample and non sample values of x are in the interval $[c,d]$ and are generated by densities d_s and d_{p-s} respectively, both bounded away from zero on $[c,d]$, and assumed to have continuous second derivatives.

Theorem: (2.1)

If for any expression Z, $E(Z/u) = A(u) + O(B)$ and var $(Z/u) = O(c)$, then $Z = A(u) + O_p\left(B + C^{1/2}\right)$. This result follows from the Chebychev inequality.

From expression (2.3),

Let $Z = \hat{d}_s(x_j) = \frac{1}{nb}\sum_{i\in s} k\left(\frac{x_i - x_j}{b}\right)$,

then

$$E\left[\hat{d}_s\left(x_i\right)/x_j\right]=\frac{1}{nb}\sum_{i\in s}E\left[k\left(\frac{x_i-x_j}{b}\right)\right]$$

$$=\frac{1}{nb}\sum_{i\in s}\int k\left(\frac{w-x_j}{b}\right)d_s(w)dw\,.$$

By letting

$\frac{w-x_j}{b}=u$ and using the Taylor's series expansion, we have

$$E\left[\hat{d}_s\left(x_i\right)/x_j\right]=d_s\left(x_j\right)+b^2\frac{k_2}{2}d_s''\left(x_j\right)+O\left(b^3\right). \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2.4).$$

Next

$$\operatorname{var}\left[\hat{d}_s x_j / x_i\right] = \frac{1}{n^2 b^2} \sum_{i \in s} \operatorname{var}\left[k\left(\frac{x_i - x_j}{b}\right)\right]$$

$$= \frac{1}{n^2 b^2} \sum_{i \in s} \operatorname{var}\left[k\left(\frac{w - x_j}{b}\right)\right]$$

$$= \frac{1}{n^2 b^2} \sum_{i \in s} \left[E\left[k\left(\frac{w - x_j}{b}\right)\right]^2 - \left[Ek\left(\frac{w - x_j}{b}\right)\right]^2\right]$$

$$= \frac{1}{n^2 b^2} \sum_{i \in s} \left[\int k\left(\frac{w - x_j}{b}\right)^2 d_s(w) dw - \left[\int k\left(\frac{w - x_j}{b}\right) d_s(w) dw\right]^2\right]$$

$$= \frac{1}{nb}\left[d(x_j)\right] + \frac{b}{n} d_s''(x_j) + \frac{1}{nb^2} + \frac{1}{n}.$$

Considering orders in magnitude we have

$$\text{var}\left[\hat{d}_s(x_i)/x_j\right] = O\left(\frac{1}{nb}\right) + O\left(\frac{b}{n}\right) + O\left(\frac{b^2}{n}\right) + O\left(\frac{1}{n}\right)$$

$$= O\left(\frac{1}{nb}\right) \text{.....(2.5).}$$

Combining (2.4) and (2.5) then

$$\hat{d}_s(x_j) = d_s(x_j) + b^2 \frac{k_2}{2} d_s''(x_j) + O_p\left(b^3 + n^{-1/2} b^{-1/2}\right) \text{.....(2.6).}$$

Next we consider the expectation of the second part of (2.3), that is

$$E\left[(nb)^{-1} \sum_{i \in s} \left[k\left(\frac{x_i - x_j}{b}\right)\left[m(x_i) - m(x_j)\right]\right] / x_j \right]$$

$$=(nb)^{-1}\sum_{i\in s}E\left[k\left(\frac{w-x_j}{b}\right)[m(w)-m(x_j)]\right]$$

$$=(nb)^{-1}\sum_{i\in s}\int k\left(\frac{w-x_j}{b}\right)[m(w)-m(x_j)]d_s(w)dw$$

$$=(nb)^{-1}\sum_{i\varepsilon s}\left[b^2k_2\left(\frac{\beta(x_j)-d_s(x_j)m''(x_j)}{2}+\frac{d_s(x_j)m''(x_j)}{2}\right)+O(b^3)\right]$$

$$=b^2\frac{k_2}{2}\beta(x_j)+O(b^{3.}) \qquad (2.7).$$

Next

$$\operatorname{var}\left[(nb)^{-1}\sum_{i\in s}k\left(\frac{x_i-x_j}{b}\right)[m(x_i)-m(x_j)]/x_j\right]=$$

$$(nb)^{-2}\sum_{i\in s}\left[\int k\left(\frac{w-x_j}{b}\right)^2\left[m(w)-m(x_j)\right]^2 d_s(w)dw\right.$$

$$\left.-\left[\int k\left(\frac{w-x_j}{b}\right)\left[m(w)-m(x_j)\right]d_s(w)dw\right]^2\right]$$

$$=(nb)^{-2}\sum_{i\in s}\left[b\left(b^2 m'(x_j)^2 d_s(x_j)k_2+O(b^3)\right)+O(b^6)\right]$$

$$=O\left(\frac{b}{n}\right)+O\left(\frac{b^2}{n}\right)+O\left(\frac{b^4}{n}\right)$$

$$=O(n^{-1}b)\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.8).$$

Therefore combining (2.7) and (2.8), then

$$(nb)^{-1}\sum_{i\in s}k\left(\frac{x_i-x_j}{b}\right)\left[m(x_i)-m(x_j)\right]$$

$$=b^2\frac{k_2}{2}\beta(x_j)+O_p\left(b^3+n^{-\frac{1}{2}}b^{\frac{1}{2}}\right).$$

Now from (2.3), we have

$$E\left[\hat{T}_{np} - T \middle/ X_p\right] = \sum_{j \in p-s} \frac{\frac{b^2}{2}\beta(x_j) + O_p\left(b^3 + n^{\frac{-1}{2}} b^{\frac{1}{2}}\right)}{d_s(x_j) + b^2 \frac{k_2}{2} d_s''(x_j) + O_p\left(b^3 + n^{-1/2} b^{-1/2}\right)}$$

$$= \sum_{j \in p-s} \frac{b^2/2\,\beta(x_j) + O_p\left(b^3 + n^{\frac{-1}{2}} b^{\frac{1}{2}}\right)}{\hat{d}_s(x_j)}.$$

But $\hat{d}_s(x_j) = d_s(x_j)(1+x)$

where $x = \dfrac{b^2 \frac{k^2}{2} d_s''(x_j)}{d_s(x_j)} + O_p\left(b^3 + n^{-1/2} b^{-1/2}\right)$.

Hence

$$E\left[\hat{T}_{np} - T \middle/ X_p\right] = \sum_{j \in p-s}\left[\int \frac{b^2 k_2 \beta(x_j)}{2 d_s(x_j)} + O_p\left(b^3 + n^{-1/2} b^{1/2}\right)\right] \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.9).$$

From (2.9) we get the mean with respect to nonsample X's to get

$$E\left[\hat{T}_{np} - T \Big/ X_p\right] = \sum_{j \in p-s} \int \frac{b^2 k_2 \beta(x_j) d_{p-s}(x_j) dx_j}{2 d_s(x_j)}$$

$$+ (N-n) O_p\left(b^3 + n^{-1/2} b^{1/2}\right)$$

$$= b^2 (N-n) \frac{K_2}{2} \int B(x) d_s(x)^{-1} d_{p-s}(x) dx + O_p\left(nb^3 + n^{1/2} b^{1/2}\right) \ldots\ldots (2.10)$$

2.5 CONDITIONAL VARIANCE OF $\hat{T}_{np} - T$

From (2.2), we derive the conditional variance of $\hat{T}_{np} - T$ as follows

$$\hat{T}_{np} - T = \sum_{p-s} \hat{m}(x_j) - \sum_{p-s} Y_j$$

but $\hat{m}(x_j) = (nb)^{-1} \sum_{i \in s} k\left(\frac{x_i - x_j}{b}\right) Y_i \Big/ (nb)^{-1} \sum_{i \in s} k\left(\frac{x_i - x_j}{b}\right)$

then

$$\hat{T}_{np} - T = \sum_{j \in p-s} \frac{\frac{1}{nb} \sum_{i \in s} k\left(\frac{x_i - x_j}{b}\right) Y_i}{(nb)^{-1} \sum_{i \in s} k\left(\frac{x_i - x_j}{b}\right)} - \sum_{j \in p-s} Y_j$$

$$= \sum_i w_i Y_j - \sum_j Y_j$$

where

$$w_i = \sum_{j \in p-s} (nb)^{-1} k\left(\frac{x_i - x_j}{b}\right) \Big/ (nb)^{-1} \sum_{i \in -s} k\left(\frac{x_i - x_j}{b}\right).$$

Therefore,

$$\text{var}\left[\hat{T}_{np} - T/X_p\right] = \text{var}\left[\sum_i w_i Y_i - \sum_j Y_j\right]$$

$$= \sum_i w_i^2 \sigma^2(x_i) + \sum_j \sigma^2(x_j) \quad \text{.....(2.11).}$$

Using assumptions 2.0 and theorem 2.1 we rewrite 2.11 in another form which is more convenient.

We let

$$Z = w_i^2 = \sum_j (nb)^{-2} k\left(\frac{x_j - x_i}{b}\right) \hat{d}_s(x_j)^{-2}$$

$$+ (nb)^{-2} \sum_j \sum_{j' \neq j'} k\left(\frac{x_j - x_i}{b}\right) k\left(\frac{x_{j'} - x_i}{b}\right) \hat{d}_s(x_{j'})^{-1} \hat{d}(x_j)^{-1}.$$

Then

$$E\left[w_i^2 / x_i \right] = E\left[\sum_j (nb)^{-2} k^2 \left(\frac{x_j - x_i}{b} \right) \hat{d}_s(x_j)^{-2} / x_i \right] +$$

$$E\left[(nb)^{-2} \sum_{j_{j \neq j'}} \sum_{j'} k\left(\frac{x_j - x_i}{b} \right) k\left(\frac{x_{j'} - x_i}{b} \right) \hat{d}_s(x_{j'})^{-1} \hat{d}(x_j)^{-1} / x_i \right] \quad ...(2.12).$$

Considering the first term of (2.12) and replacing $\hat{d}(x_j)^{-2}$ with its expansion, then

$$Mn^{-2}b^{-2}E\left[k^2\left(\frac{x_j - x_i}{b} \right) d_s(x_j)^{-2} \left(1 + O_p\left(b^2 + n^{-1/2}b^{-1/2} \right) / x_i \right) \right]$$

$$= Mn^{-2}b^{-2}\left[\int k^2\left(\frac{w - x_i}{b} \right) d_s(w)^{-2} d_{p-s}(w) dw \right.$$

$$\left. + \int O_p\left(b^2 + n^{-1/2}b^{-1/2} \right) k^2\left(\frac{w - x_i}{b} \right) d_s(w)^{-2} d_{p-s}(w) dw \right]$$

$$= Mn^{-2}b^{-1} \int k^2(u) du \, d_s(x_i)^{-2} d_{p-s}(x_i) du + O\left(n^{-1} + n^{-1.5}b^{-2..5} \right) \ldots\ldots\ldots\ldots\ldots(2.13)$$

Where

$$M = N - n.$$

Considering the second term of 2.12 and assuming that j' and j are *i.i.d*, we have

$$M(M-1)(nb)^{-2}\left[E\left[k\left(\frac{x_j - x_i}{b}\right)d_s(x_j)^{-1}\left(1 - b^2\frac{k_2}{2}C(x_j)\right)\right.\right.$$

$$\left.\left. + \; O_p\left(b^3 + n^{-1/2}b^{-1/2}\right)\right] / x_i\right]^2$$

Where $C(x_i) = -d_s(x_j)^{-1} d_s''(x_j)$

$$= M(M-1)(nb)^{-2}\left[\int k\left(\frac{w-x_i}{b}\right)d_s(w)^{-1}d_{p-s}(w)dw\right.$$

$$+\int k\left(\frac{w-x_i}{b}\right)d_s(w)^{-1}b^2\frac{k_2}{2}C(w)d_{p-s}(w)dw$$

$$\left.+\int O_p\left(b^3+n^{-1/2}b^{-1/2}\right)d_s(w)^{-1}d_{p-s}(w)dw\right]^2$$

$$= M(M-1)(nb)^{-2}\left[bd_s(x_i)^{-1}d_{p-s}(x_i)+b^3\frac{k_2}{2}C(x_i)d(x_i)^{-1}d_{p-s}(x_i)\right.$$

$$\left.+\quad O\left(b^3+n^{-1/2}b^{-1/2}\right)\right]^2$$

$$= M^2n^{-2}\left[d_s(x_i)^{-2}d_{p-s}(x_i)^2+b^2\frac{k_2}{2}C'(x_i)\right]$$

$$+\quad M^2n^{-2}O\left(b^2+n^{-1/2}b^{-1.5}\right)\ldots(2.14)$$

where $C'(x_i) = C(x_i) d_s(x_i)^{-1} d_{p-s}(x_i)$.

Therefore, combining (2.13) and (2.14) then

$$E\left[w_i^2 / x_i \right] = Mn^{-2} b^{-1} \int k^2(u) du d_s(x_i)^{-2} d_{p-s}(x_i)$$

$$+ M^2 n^{-2} \left[d_s(x_i)^{-2} d_{p-s}(x_i)^2 + b^2 \frac{k_2}{2} C'(x_i) \right]$$

$$+ O\left(b^2 + n^{-1/2} b^{-1.5} \right) - - - (2.15).$$

Next we consider the conditional variance of w_i^2, i.e

$$Var\left[w_i^2 / x_i \right] = E\left[w_i^4 / x_i \right] - \left[E\left[w_i^2 / x_i \right] \right]^2 \text{--------(2.16).}$$

Now

$$E\left[w_i^4 / x_i \right] = O\left(n^{-2}b^{-3}\right) + O\left(n^{-2}b^{-2}\right)$$

$$+ O\left(n^{-1}b^{-1}\right) + d_s(x_i)^{-4} d_{p-s}(x_i)^4$$

and

$$\left[E\left[w_i^2 / x_i \right] \right]^2 = O\left(n^{-2}b^{-2}\right) + O\left(n^{-1}b^{-1}\right)$$

$$+ O\left(n^{-1}b\right) + O\left(n^{-1}b^{-1}\right).O\left(b^2 + n^{-1/2}b^{-1.5}\right)$$

$$+ d_s(x_i)^{-4} d_{p-s}(x_i)^4 .$$

From these two expressions, the dominant terms cancel and we find that

$$\text{var}\left[w_i^2 / x_i \right] = O\left(n^{-1}b^{-1}\right) \quad \text{.....(2.17)} .$$

From (2.15) and (2.17) we have

$$w_i^2 = Mn^{-2}b^{-1}\int k^2(u)du d_s(x_i)^{-2} d_{p-s}(x_i)$$

$$+ M^2 n^{-2}\left[d_s(x_i)^{-2} d_{p-s}(x_i)^2 + b^2 \frac{k_2}{2} C'(x_i) \right]$$

$$+ O_p\left(b^2 + n^{-1/2}b^{-1/2}\right) \quad \text{.....(2.18).}$$

Next we get the mean with respect to sample $X's$ and sum over i so that

$$\operatorname{var}\left[\hat{T}_{np} - T \Big/ X_p\right] = \sum_i \left[Mn^{-2}b^{-1}\int k^2(u)du \int d_s(x_i)^{-2} d_{p-s}(x_i)\sigma^2(x_i)d_s(x_i)dx_i\right]$$

$$+ \sum_i \left[M^2 M^{-2}\int \sigma^2(x_i)d_s(x_i)^{-2} d_{p-s}(x_i)^2 d_s(x_i)dx_i\right]$$

$$+ \sum_i \left[M^2 n^{-2}\int \sigma^2 x_i b^2 \frac{k_2}{2} C^*(x_i)dx_i + O_p\left(b^2 + n^{-1/2}b^{-1/2}\right)\right]$$

$$+ \sum_j \int \sigma^2(x_i)d_{p-s}(x_j)dx_j$$

$$= (N-n)n^{-1}b^{-1}\int k^2(u)du\int \sigma^2(x)d_s(x)^{-1}d_{p-s}(x)dx. +$$

$$+ (N-n)^2 n^{-1}\int \sigma^2(x)d_s(x)^{-1}d_{p-s}(x)^2 dx$$

$$+ (N-n)^2 n^{-1}b^2\frac{k_2}{2}\int \sigma^2(x)C^*(x)d_s(x)dx$$

$$+ (N-n)\int \sigma^2(x)d_{p-s}(x)dx + O_p\left(nb^2 + n^{1/2}b^{-1/2}\right) \quad \text{.....(2.19)}$$

where $C^*(x)$ is a complicated function of the derivatives $d_s(x)$ and $d_{p-s}(x)$.

CHAPTER THREE

MODEL ASSISTED APPROACH

3.1 INTRODUCTION

In this chapter a new type of model assisted nonparametric regression estimation for the finite population total based on local polynomial smoothing is considered. We investigate the model-based properties of this new estimator, i.e. its conditional mean and variance. In this approach, we introduce the concept of local polynomial regression estimation where we incorporate the Horvitz-Thompson estimator to come up with a model-assisted estimator. The concept of local polynomial regression is a general concept in kernel regression. The two can be applied to a variety of problems. See Cleveland (1979) and Cleveland Delvin (1988).

3.2 LOCAL POLYNOMIAL REGRESSION ESTIMATION

In local polynomial regression estimation, we consider a model which is more general, i.e. a model which is parametrically unspecified. This is different from traditional regression estimation (Ratio and linear regression estimation)

which makes use of parametrically specified models. Model (1.1) will be considered in this case.

To obtain a local polynomial regression estimator, we assume a polynomial regression model locally around x. Given that the regression function $m(\cdot)$ has derivatives up to order p, then by a Taylor approximation,

$$m(z) \cong \sum_{j=0}^{p} \frac{m^j(x)}{j!}(z-x)^j \equiv \sum_{j=0}^{p} \beta_j(x)(z-x)^j$$

Where $\beta_j(x)$ is a smooth and z is in a neighborhood of x. We use the weighted least squares method, locally to obtain $\hat{m}(x)$ the estimator for $m(x)$, i.e. with

$$\beta_x \equiv (\beta_0(x), \beta_1(x) \ldots \beta_p(x))^t .$$

We minimize the following expression with respect to β_x, i.e.

$$\text{Min}\ \beta_x \quad \sum_{i=1}^{n}\left(Y_i - \sum_{j=0}^{p}\beta_j(x)(x_i - x)^j\right)^2 w_i(x) \quad \ldots\ldots\ldots\ldots\ldots\ldots (3.1).$$

[Augustyns (1997)]

Considering a local polynomial regression model with kernel weights

$$w_i(x)=\frac{1}{b}k\frac{(x_i-x)}{b},$$

where $k(\cdot)$, is the kemel function and $b>0$ the smoothing parameter (bandwidth), equation (3.1) becomes

$$\text{Min}\,\beta_x \quad \sum_{i=1}^{n}\left(Y_i-\sum_{j=0}^{p}\beta_j(x)(x_i-x)^j\right)^2\frac{1}{b}k\left(\frac{xi-x}{b}\right).....(3.2).$$

To get the solution of equation (3.2),

let

$Y=(Y_1\,Y_2...Y_n)'$ be the vector of $Y_i's$ in the sample.

Define

$$X_x=\begin{pmatrix}1,(x_1-x.),.....,(x_1-x)^p\\ .\\ .\\ .\\ 1,(x_n-x),.....,(x_n-x)^p\end{pmatrix}$$

a $n\,x\,p+1$ design matrix and

$w_x = diag \underset{1 \le i \le n}{\left(\frac{1}{b} k \frac{(x_i - x)}{b}\right)}$, a $n\,x\,n$ matrix.

Then (3.2) can be written as

$$\text{Min}\ \beta_x (Y - X_x\beta_x)' W_x (Y - X_x\beta_x)$$

$$\text{Min}\,\beta_x \left[Y' W_x Y - Y' W_x X_x \beta_x - (X_x\beta_x)' W_x Y + (X_x\beta_x)' W_x X_x \beta_x \right] \ldots\ldots\ldots (3.3).$$

To minimize the above with respect to β_x, we have

$$\frac{\partial F(\cdot)}{\partial \beta_x} = -2X_x' W_x Y + 2X_x' W_x X_x \beta_x = 0$$

where $F(\cdot)$ represents (3.3).

Then

$$X_x' W_x X_x \beta_x = X_x' W_x Y$$

which implies that

$$\hat{\beta}_x = \left(X_x' W_x X_x \right)^{-1} X_x' W_x Y .$$

The estimator for $m(x)$ is

$$\hat{m}(x) = \hat{\beta}_0(x) = \ell_1' \left(X_x W_x X_x \right)^{-1} \left(X_x' W_x Y \right)$$

where $\ell_1' = (1,0,0...,0)$ which is a $(p+1)$ vector.

We note that $\hat{\beta}_j(x), j = 1,2,...p$ are the estimators of the derivatives of the regression function $m(\cdot)$ at x up to order p. The above shows that local polynomial estimators are linear smoothers of the form

$$\sum_{i=1}^{n} w_i(x) Y_i$$

The coefficient of the linear combination depends on the degree p of the polynomial approximation. We note that for $p=0$, the estimator reduces to the Nadaraya-Watson estimator [Augustyns (1997)].

3.3 LOCAL POLYNOMIAL REGRESSION ESTIMATOR

Now let the proposed estimator of the population total be the Horvitz-Thompson estimator

$$\hat{T} = \sum_{i \in s} \frac{Y_i}{\pi_i}$$

This estimator is design-unbiased for any design $p(\cdot)$. The sample weights do not depend on the variables of interest $Yi's$, meaning that the sample weights can be applied directly to any other variables of interest in the same sample. Due to mathematical complexity we assume that $p=0$ throughout. Then, under this assumption, the local polynomial regression estimator for the finite population total based on the Horvitz-Thompson estimator is given by

$$\hat{T}_{lp} = \sum_{i \in s} \frac{Y_i - \hat{m}(x_i)}{\pi_i} + \sum_{i=1}^{N} \hat{m}(x_i) \quad \text{.....(3.4)}$$

where $\hat{m}(x_i)$ is the sample estimator for $m(x_i)$.

Opsomer and Breidt (2000) suggests another estimator different from this new one in the sense that it has inclusion probability in the smoothing weight. Our proposed estimator has a major advantage over the one suggested above, since it is easy to compute as compared to the other which complicates the algebra in calculation of its bias and variance, although the presence of the inclusion probabilities in the smoothing weights makes the sample based estimator $\hat{m}(x)$ a design-consistent estimator of the finite population smooth $m(x)$.

Now the error $T_{lp} - T$ is given by

$$\hat{T}_{lp} - T = \sum_{\tilde{s}} (\hat{m}(x_i) - Y_i) - \sum_{s} \left(\frac{1}{\pi_i} - 1 \right) (\hat{m}(x_i) - Y_i) \quad \text{.....(3.5)}$$

Let i and k refer to sample values and j to nonsample values. Then

$$\hat{T}_{lp} - T = \sum_{j} \left(\sum_{i} w_i(x_i) Y_i - Y_j \right) - \sum_{i} \left(\frac{1}{\pi_i} - 1 \right) \left[\sum_{k \in s} w_k(x_i) Y_k - Y_i \right] \quad \text{.....(3.6)}$$

where $\sum_i w_i(x_i)Y_i = \hat{m}(x_i)$ and $\sum_{k\in s} w_k(x_i)Y_k = \hat{m}(x_i)$.

3.4 THE CONDITIONAL MEAN OF $\hat{T}_{lp} - T$

The conditional mean of $\hat{T}_{lp} - T$ under the model (1.1) is given by

$$E\left[\hat{T}_{lp} - T \Big/ X_p\right] = E\left[\sum_j \left[\sum_i w_i(x_i)Y_i - Y_j\right]\right]$$

$$- E\left[\sum_i \left(\frac{1}{\pi_i} - 1\right)\left[\sum_{k\in s} w_k(x_i)Y_k - Y_i\right]\right]$$

$$= \sum_j \left[\sum_i w_i(x_i)E(Y_i) - E(Y_j)\right] - \sum_i \left(\frac{1}{\pi_i} - 1\right)\left[\sum_{k\in s} w_i(x_i)E(Y_k) - E(Y_i)\right]$$

$$=\sum_j\left[\frac{\sum_i\frac{1}{nb}k\left(\frac{x_j-x_i}{b}\right)m(x_i)}{(nb)^{-1}\sum_i k\left(\frac{x_i-x_j}{b}\right)}-m(x_j)\right]$$

$$-\sum_j\left(\frac{1}{\pi_i}-1\right)\left[\frac{\sum_{k\in s}\frac{1}{nb}k\left(\frac{x_j-x_i}{b}\right)m(x_k)}{(nb)^{-1}\sum_{k\in s}k\left(\frac{x_k-x_i}{b}\right)}-m(x_i)\right]$$

$$=\sum_j\hat{d}_s(x_j)^{-1}\frac{1}{nb}\sum_i k\left(\frac{x_j-x_i}{b}\right)\left[m(x_i)-m(x_j)\right]$$

$$+\sum_i\left(\frac{1}{\pi_i}-1\right)\hat{d}_s(x_i)^{-1}\frac{1}{nb}\sum_{k\in s}k\left(\frac{x_j-x_i}{b}\right)\left[m(x_i)-m(x_k)\right].....(3.7)$$

Using theorem (2.1) and assumptions (2.0) the first term of (3.7) is equivalent to

$$b^2(N-n)\left(\frac{k_2}{2}\right)\int\beta(x)d_s(x)^{-1}d_{p-s}(x)dx+O_p\left(nb^3+n^{1/2}b^{1/2}\right)\ .....(3.8).$$

We now consider the expansion of the second term of the equation (3.7) i.e.

$$\sum_i \left(\frac{1}{\pi_i}-1\right)\hat{d}_s(x_i)^{-1}\frac{1}{nb}\sum_{k\in s}k\left(\frac{x_i-x_k}{b}\right)[m(x_i)-m(x_k)]$$

$$=\sum_i \left(\frac{1}{\pi_i}-1\right)\left[b^2\left(\frac{k_2}{2}\right)\int\beta(x)d_s(x)^{-1}d_{p-s}(x)dx+O_p\left(b^3+n^{-1/2}b^{1/2}\right)\right]\ldots\ldots\ldots\ldots(3.9).$$

Assuming simple random sampling design where $\pi_i=\frac{n}{N}$, (3.9) becomes

$$(N-n)b^2\left(\frac{k_2}{2}\right)\int\beta(x)d_s(x)^{-1}d_{p-s}(x)dx+O_p\left(nb^3+n^{1/2}b^{1/2}\right)\ldots\ldots\ldots\ldots\ldots\ldots(3.10)$$

combining (3.8) and (3.10), we have

$$E\left[\hat{T}_{lp}-T\Big/X_p\right]=b^2(N-n)k_2\int\beta(x)d_s(x)^{-1}d_{p-s}(x)dx+O_p\left(nb^3+n^{1/2}b^{1/2}\right)\ldots\ldots\ldots(3.11)$$

3.5 CONDITIONAL VARIANCE OF $\hat{T}_{lp}-T$

So far we have shown that

$$\hat{T}_{lp}-T=\sum_j\left(\hat{m}(x_j)-Y_j\right)-\sum_i\left(\frac{1}{\pi_i}-1\right)\left(\hat{m}(x_i)-Y_i\right).$$

But $\sum_j \left(\hat{m}(x_j) - Y_j\right) = \sum_i w_i Y_i - \sum_j Y_j$ .....(3.12)

where $$w_i = \sum_j \frac{1}{nb} k\left(\frac{x_j - x_i}{b}\right) \Bigg/ (nb)^{-1} \sum_i k\left(\frac{x_i - x_j}{b}\right)$$

and

$$\sum_i \left(\frac{1}{\pi_i} - 1\right)\left(\hat{m}(x_i) - Y_i\right) = \sum_i \left(\frac{1}{\pi_i} - 1\right)\left[(nb)^{-1} \frac{\sum_{k \in s} k\left(\frac{x_i - x_k}{b}\right) Y_k}{(nb)^{-1} \sum_{k \in s} k\left(\frac{x_i - x_k}{b}\right)} - Y_i\right]$$

$$= \sum_{k \in s} \tilde{w}_k Y_k - \sum_{i \in s} \left(\frac{1}{\pi_i} - 1\right) Y_i \text{.....(3.13)}$$

where

$$\tilde{w}_k = \sum_i \left(\frac{1}{\pi_i} - 1\right) \frac{(nb)^{-1} k\left(\frac{x_i - x_k}{b}\right)}{(nb)^{-1} \sum_k k\left(\frac{x_i - x_k}{b}\right)}.$$

REMARK:

$$\text{If } w_k = \sum_i \left(\frac{1}{nb}\right) k \frac{(x_i - x_k)}{b} \Big/ (nb)^{-1} \sum_{k \in s} k \frac{(x_i - x_k)}{b}$$

then $\tilde{w}_k$ has more weight than w_k provided the inclusion probability π_i is less than ½. Otherwise $\tilde{w}_k$ and w_k have the same weight if $\pi_i = ½$.

Now we combine equations (3.12) and (3.13) to get

$$T_{lp} - T = \sum_i w_i Y_i + \sum_j Y_j + \sum_{k \in s} Y_i - \sum_{k \in s} \tilde{w}_k Y_k + \sum_i \left(\frac{1}{\pi_i} - 1\right) Y_i .$$

Let $\tilde{w}_k = z_i$ then

$$T_{lp} - T = \sum_i \left(w_i - z_i + \left(\frac{1}{\pi_i} - 1\right)\right) Y_i - \sum_j Y_j .$$

Therefore

$$\text{var}\left[T_{lp} - T / X_p\right] = \text{var}\left[\sum_i \left(w_i - z_i + \left(\frac{1}{\pi_i} - 1\right)\right) Y_i - \sum_j Y_j\right]$$

$$= \sum_i w_i^2 \sigma^2(x_i) + \sum_i z_i^2 \sigma^2(x_i) + \sum_i \left(\frac{1}{\pi_i} - 1\right)^2 \sigma^2(x_i) - 2\sum_i w_i z_i \sigma^2(x_i) + 2\sum_i \left(\frac{1}{\pi_i} - 1\right) w_i \sigma^2(x_i)$$

$$-2\sum_i \left(\frac{1}{\pi_i} - 1\right) z_i \sigma^2(x_i) + \sum_j \sigma^2(x_j) \quad \text{.....(3.14).}$$

From the previous chapter, we have shown that

$$\sum_i w_i^2 \sigma^2(x_i) + \sum_j \sigma^2(x_j)$$

$$= Mn^{-1}b^{-1} \int k^2(u)du \int \sigma^2(x) d_s(x)^{-1} d_{p-s}(x)dx + M^2 n^{-1} \int \sigma^2(x) d_s(x)^{-1} d_{p-s}(x)dx$$

$$+ \quad M^2 n^{-1} b^2 \frac{k^2}{2} \int \sigma^2(x) \overset{*}{C}(x) d_s(x)dx + (N-n) \int \sigma^2(x) d_{p-s}(x)dx + O_p\left(nb^2 + n^{1/2}b^{-1/2}\right) \quad \text{.....(3.15)}$$

.For the remaining section of equation (3.14) we expand w_i and z_i to get

$$w_i = Mn^{-1} \int k(u)du d_s(x_i)^{-1} d_{p-s}(x_i) + O_p\left[\left(n^2 b + n^{1/2} b^{-1/2}\right)^{1/2}\right]$$

and

$$z_i = \sum_i \left(\frac{1}{\pi_i} - 1\right) n^{-1} \int k(u)du + O_p\left[\left(n^2 b + n^{1.5} b^{-1.5}\right)^{1/2}\right].$$

Next we multiply z_i with w_i and integrate with respect to samples *Xi's,* to get

$$w_i z_i = \sum_i \left(\frac{1}{\pi_i} - 1\right) Mn^{-2} \int d_s(x_i)^{-1} d_{p-s}(x_i) d_s(x_i) dx_i + O_p\left(n^2 b + n^{1.5} b^{-1.5}\right) \text{.....(3.17).}$$

We also square z_i and integrate the result with respect to sample *Xi's,* to get

$$z_i^2 = \sum_i \left(\frac{1}{\pi_i} - 1\right)^2 n^{-2} \int d_s(x_i) d_x + O_p\left[n^2 b + n^{1.5} b^{-1.5}\right] \text{.....(3.18).}$$

Lastly we replace equations (3.15), (3.16), (3.17) and (3.18) in equation (3.14) so that

$$\text{var}\left[T_{lp} - T / X_p\right] = Mn^{-1} b^{-1} \int k^2(u) du \int \sigma^2(x) d_s(x)^{-1} d_{p-s}(x) dx + M^2 n^{-1} \int \sigma^2(x) d_s(x)^{-1} d_{p-s}(x)^2 dx$$

$$+ M^2 n^{-1} b^2 \frac{k_2}{2} \int \sigma^2(x) \overset{*}{C}(x) d_s(x) dx + (N-n) \int \sigma^2(x) d_{p-s}(x) dx$$

$$+ \sum_i \sum_i \left(\frac{1}{\pi_i} - 1\right)^2 n^{-2} \int \sigma^2(x) d_s(x) dx + \sum_i \left(\frac{1}{\pi_i} - 1\right)^2 \int \sigma^2(x) d_s(x) dx$$

$$-2\sum_i\sum_i\left(\frac{1}{\pi_i}-1\right)Mn^{-2}\int\sigma^2(x)d_{p-s}(x)dx+$$

$$2\sum_i\left(\frac{1}{\pi_i}-1\right)\sum_i\left(\frac{1}{\pi_i}-1\right)n^{-1}\int\sigma^2(x)d_s(x)dx$$

$$+Op\left[n^3b+n^{5/2}b^{-1.5}\right].$$

Assuming simple random design, the above simplifies to

$$\text{var}\left[T_{lp}-T/X_p\right]=Mn^{-1}b^{-1}\int k^2(u)du\int\sigma^2(x)d_s(x)^{-1}d_{p-s}(x)dx+M^2n^{-1}\int\sigma^2(x)d_s(x)^{-1}d_{p-s}(x)^2dx$$

$$+\ M^2n^{-1}b^2\frac{k_2}{2}\int\sigma^2(x)\overset{*}{C}(x)d_s(x)dx+(N-n)\int\sigma^2(x)d_{p-s}(x)dx\ +$$

$$\left(\frac{N-n}{n}\right)^2\int\sigma^2(x)d_s(x)dx-.\frac{(N-n)^2}{n}\int\sigma^2(x)d_s(x)dx+O_p\left[n^3b+n^{5/2}b^{-1.5}\right].....(3.19).$$

3.6 CONCLUSION

Under the assumptions given in the project, we observe that the bias based on the local polynomial regression estimator is large compared to the bias based on nonparametric regression estimator.

Again the error variance of $\hat{T}_{lp} - T$ is small than the error variance of $\hat{T}_{np} - T$ provided n is large. Therefore, we conclude that nonparametric regression estimator $\hat{T}_{np}$ is a good estimator in terms of bias compared to the local polynomial regression estimator $\hat{T}_{lp}$.

CHAPTER FOUR

EMPIRICAL STUDY

4.1 INTRODUCTION

In this chapter, we carry out an empirical study to compare the performance of the two estimators discussed in chapters two and three. Two artificial populations and one natural population are used in the study.

4.2 DESCRIPTION OF THE STUDY POPULATIONS

In artificial population I, 300 data points were generated according to the model

$Y_i = \beta x_i + \ell_i$ with $\ell_i \sim N(0,\sigma^2)$, $x_i \sim U[0,1]$ mutually independent across i and β a constant (100).In artificial population II, we again generate 300 data points according to the model $Y_i = \beta x_i + x_i \ell_i$ where x_i, ℓ_i and β are the same as in artificial population I.

The natural population of size 46 was obtained from the Central Bureau of Statistics. In this population, the variable of interest Y_i is the total income while the auxiliary variable X_i is the total expenditure both from the same district .

4.3 DESIGN OF THE STUDY

For each of the two artificial populations, 500 samples of size 100 were drawn by simple random sampling without replacement. The same procedure is used in the natural population but with 100 samples of size 10. The *Epanechnikov Kernel*

$k(u)=\frac{3}{4}(1-u^2)$ if $|u| \leq 1$ or zero otherwise was used in the study for the two nonparametric estimators considered.

We search for an optimal bandwidth for *Nadaraya-Watson Smoother* within the interval

$$\left(\frac{\sigma}{4n^{1/5}} \leq b \leq \frac{3\sigma}{2n^{1/5}}\right)$$

Where σ is the standard deviation of $Xi's$ [Silverman (1986)].

4.4 DESCRIPTION OF THE COMPUTATION PROCEDURE

For each of the three populations, we compute

$T = \sum_{i=1}^{N} Y_i$ and for the $i\ th$ sample, $i = 1,2,...500$ (for the artificial population) and

$i = 1,2,...100$ (for the natural population),

$\hat{T}_{npi}$ and $\hat{T}_{lpi}$ the values of $\hat{T}_{np}$ and $\hat{T}_{lp}$ respectively.

The biases for the artificial populations were computed as

$$\sum_{i=1}^{500} \frac{\left(\hat{T}_{npi} - T\right)}{500}$$

and

$$\sum_{i=1}^{500} \frac{\left(\hat{T}_{lpi} - T\right)}{500}.$$

For the natural population, the biases were computed as

$$\sum_{i=1}^{100} \frac{\left(\hat{T}_{npi} - T\right)}{100}$$

and

$$\sum_{i=1}^{100} \frac{\left(\hat{T}_{lpi} - T\right)}{100}.$$

The two mean square errors for each of the two artificial populations were computed as

$$\sum_{i=1}^{500} \frac{\left(\hat{T}_{npi} - T\right)^2}{500}$$

and

$$\sum_{i=1}^{500} \frac{\left(\hat{T}_{lpi} - T\right)^2}{500}.$$

For the natural population, the two mean squares errors were computed as

$$\sum_{i=1}^{100} \frac{\left(\hat{T}_{npi} - T\right)^2}{100}$$

and

$$\sum_{i=1}^{100} \frac{\left(\hat{T}_{lpi} - T\right)^2}{100}.$$

4.5 RESULTS

The results of this study are summarized in table I and in Figure 1.

Table 1: Biases and mean square errors of the estimators

1 ARTIFICIAL POPULATION I		
ESTIMATOR	BIAS	MEAN SQUARE ERROR (MSE)
$\hat{T}_{np}$	790.1541	1307605.3929
$\hat{T}_{lp}$	100.7838	1773713.8530
2 ARTIFICIAL POPULATION II		
ESTIMATOR	BIAS	MEAN SQUARE ERROR (MSE)
$\hat{T}_{np}$	-2.3852	514380.8937
$\hat{T}_{lp}$	-22.4549	426765.6454
3 NATURAL POPULATION		
ESTIMATOR	BIAS	MEAN SQUARE ERROR (MSE)
$\hat{T}_{np}$	-20.6693	12244.5396
$\hat{T}_{lp}$	-20.6693	12244.5396

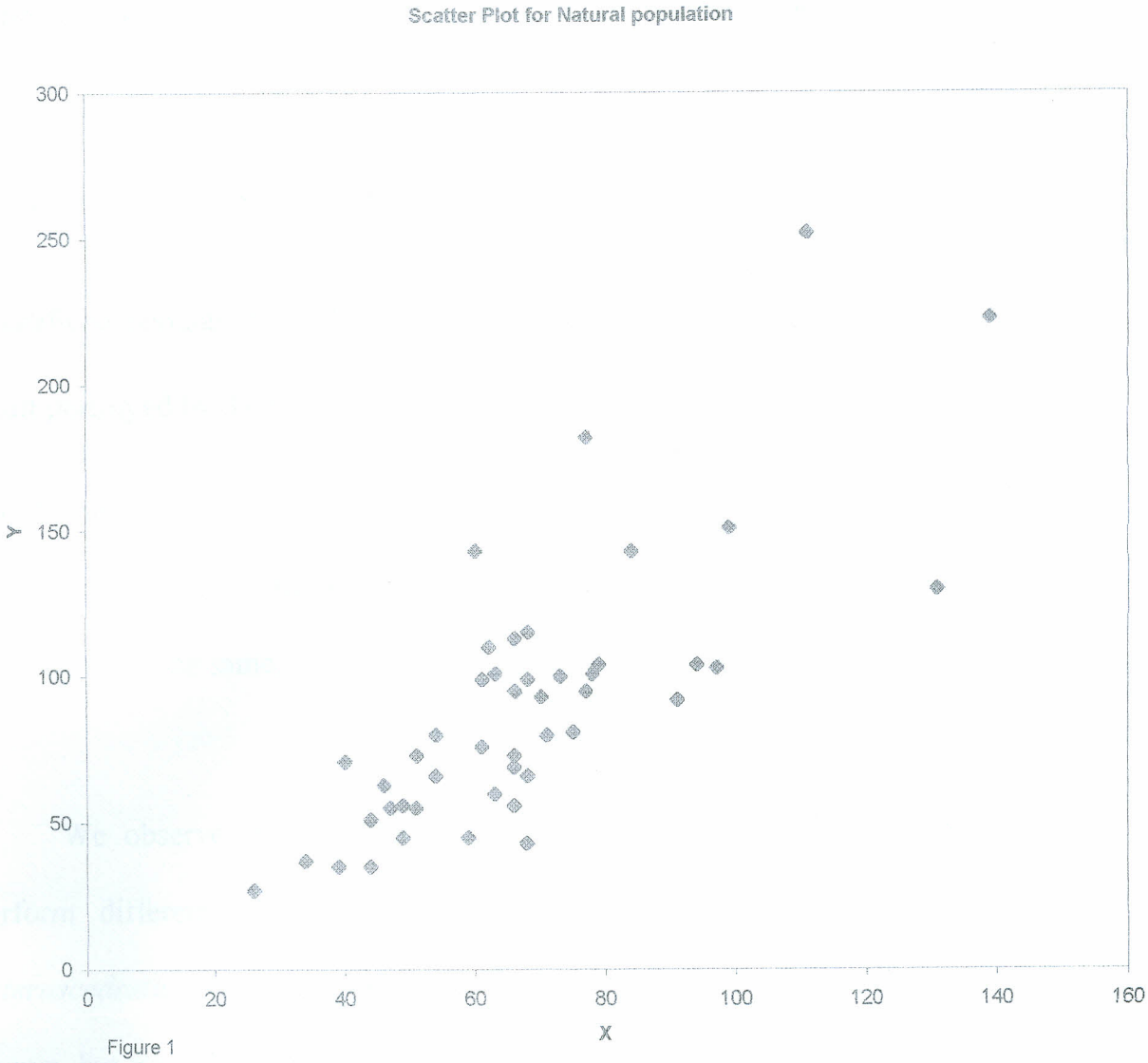

Scatter Plot for Natural population
300
250
200
150
100
50
0
Y
0
20
40
60
80
100
120
140
160
X

Figure 1

4.6 DISCUSSION OF THE RESULTS

In artificial population one, we notice that the bias of $\hat{T}_{lp}$ is small compared to the bias of $\hat{T}_{np}$. This shows that as far as this population is concerned, $\hat{T}_{lp}$ is better than $\hat{T}_{lp}$.

The mean square error of $\hat{T}_{lp}$ is greater than the mean square error of $\hat{T}_{np}$.

In artificial population II, $\hat{T}_{np}$ proves to be a good estimator compared to $\hat{T}_{lp}$, a result portrayed by theoretical work.

For the natural population, there is no significant difference between the two estimators. Theoretically if the sampling function approaches one, the two estimators are the same.

We observe that for the two artificial populations, the two estimators perform differently depending on whether the model is *homoscedastic* or *heteroscedastic*. In natural population, the structure of the population is not known, but from the scatter diagram, this structure could be linear.

CHAPTER FIVE

CONCLUSIONS AND SUGGESTIONS FOR FURTHER STUDY

5.1 INTRODUCTION

In this last chapter, we give our concluding remarks and outline suggestions for further areas of study, which have emerged during the course of our project.

5.2 CONCLUSIONS

The entire research project was about the comparison of the performance of two different estimators in estimation of two finite population totals. Generally, the two approaches to sample survey that is model-based approach and model-assisted approach are statistically valid. We cannot determine which one is better than the other, since each one of them has different conditions, which favours its validity. This is outlined by what we have observed in our project for the performance of the two estimators.

Theoretically, we have discovered that the incorporation of the *Horvitz-Thompson* estimator in local polynomial regression estimation does not improve the performance of the resulting model-assisted estimator. This was witnessed in the calculation of its bias, a result that is our major contribution to knowledge in this area of study.

5.3 AREAS FOR FURTHER RESEARCH

In the course of our research project, areas for further study have emerged. For example in our theoretical work, we only considered the case when p=0. We did not consider the case of higher order derivatives of m (x), that is the case where p=1,2,3, etc. this is an open area fro further study.

Despite the fact that we considered a general design initially, we went ahead and considered the specific case of simple random sampling design in model-assisted approach and in the empirical work. The case for other designs was not considered, a proposal still open for further research.

REFERENCES

Augustyns, I. (1997). Local Polynomial Smoothing of Sparse Multinomial Data. Unpublished Ph. D thesis, Limburgs University.

Cleveland, W. S (1979). Robust Locally Weighted Regression and Smoothing Scatter Plots, Journal of American Statistical Association. 74, 829-836.

Cleveland, W. S. And Delvin, S (1988). Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting, Journal of American Statistical Association 83, 596-610.

Chambers, R. L., Dorfman, A. H, and Hall P. (1992). Properties of Estimators of the Finite Distribution Function, Biometrika 79, 577-582.

Dorfman, A. H. (1992) Non Parametric Regression for Estimating Totals in Finite Population. Proceedings of the Section on Survey Research Methods, Journal of the American Statistical Association, 622-625.

Dorfman, A. H And Hall, P. (1993). Estimators of The Finite Population Distribution Function Using Non Parametric Regression, Annals of Statistics, 21, 1452-1475.

Godambe, W. A. (1955). A Unified Theory of Sampling from Finite Populations. Journal of Royal Statistical Society. B, 17, 269 - 278

Hansen, M.H, Maelow, W.Cy And Tepping, B.J (1983). An Evaluation of Model - Dependent and Probability Sampling Inferences in Sample Surveys, Journal of the American Statistical Association, 78, 776-793.

Hardle, W. (1991). Smoothing Techniques. London Springlet - Verlag.

Little, R.E (1982). Models for Non- Response in Sample Surveys, Journal of the American Statistical Association.77, 237-250.

Nadaraya, E.A (1964). On Estimating Regression. Theory of Probability Application 9, 141-142.

Royall, R.M And Pfeffermann, D. (1982). Balanced Samples and Robust Bayesian inference in Finite Population Sampling. Biometrika. 69, 401-410.

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

Smith, T.M.F (1983). On the validity of inferences from non- random samples, Journal of Royal Statistical Society, A, 146, 394-403.

Smith, T.M.F and Njenga, E. (1992).

Robust Model - based methods for analytic surveys, Survey methodology 18, 187-208.

Watson, G.S (1964). Smooth Regression Analysis. Sankhya. A, 359 - 372.