LOCAL POLYNOMIAL REGRESSION ESTIMATOR OF THE FINITE POPULATION TOTAL UNDER STRATIFIED RANDOM SAMPLING: A MODEL-BASED APPROACH

SYENGO CHARLES KILUNDA

MS300-0006/15

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

2017

DECLARATION

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

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

Syengo Charles Kilunda

MS300-0006/2015

This thesis has been submitted for examination with our approval as university supervisors.

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

Prof. Romanus Odhiambo Otieno

Jomo Kenyatta University of Agriculture and Technology

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

Dr. George Otieno Orwa

Jomo Kenyatta University of Agriculture and Technology

DEDICATION

This thesis is dedicated to my fiancee Rosemary for her unequal love and devoted prayers towards me, to all my mentors Rev. Zipporah and Danson Muthiani, and Prophet T.B. Joshua for their continued encouraging sermons and prayers, and finally to all my friends, relatives and colleagues who have contributed positively through spiritual, moral and academic support during the course of the study. Above all, I dedicate this thesis to my Lord and Saviour Jesus Christ, who filled me with His power, grace and strength to sail through the challenges that came my way. Indeed, when we run out of rope amidst challenges, it is time to grab onto Faith– Faith in the finished works of our Lord Jesus Christ. Amen.

To those whose lives are Christ-centered, the Best is yet to come.

Better is NOT good enough, the Best is yet to come. Amen.

ACKNOWLEDGEMENT

My deepest gratitude to my supervisors, Professor Romanus Odhiambo Otieno and Dr. George Otieno Orwa both of Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya. I have been amazingly fortunate to have supervisors of that kind who gave me the freedom to explore on my own and at the same time the guidance to recover when my steps faltered. Many thanks to Dr. Lema of Pan African University Institute for Basic Sciences, Technology and Innovation as well as Mr. Festus Were of Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya who immensely gave me technical advice on this work.

I am also indebted to the entire family of Pan African University, Institute for Basic Sciences, Technology and Innovation with whom I have interacted during the course of my graduate studies and some of whom gave valuable advice that contributed to the improvement of this thesis.

Most importantly, this study would not have been possible without the financial support from the African Union Commission that facilitated the entire graduate study. I am grateful to the Lord God Almighty who made it possible for me to be counted worthy of the financial support by the African Union.

All views, errors and omissions of any kind are my own and should not be directed to any of the persons or organizations mentioned above.

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

RAB	Relative Absolute Bias
RE	Relative Efficiency
MSE	Mean Square Error
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MISE	Mean Integrated Square Error
AMISE	Asymptotic Mean Integrated Square Error

DEFINITION OF TERMS

U	Entire population
s	Sample
r	Nonsampled set
n	Sample size
N	Population size
b	Bandwidth
y_i	Survey measurement
x_i	Auxiliary variate
Y	Finite population total
$\bar{Y}$	Finite population mean
ξ	Superpopulation model

ABSTRACT

In this thesis, auxiliary information is used to determine an estimator of finite population total using nonparametric regression under stratified random sampling. To achieve this, a model-based approach is adopted by making use of the local polynomial regression estimation to predict the nonsampled values of the survey variable. The performance of the proposed estimator is investigated against some design-based and model-based regression estimators. From the simulation experiments, the resulting estimator records better results in the estimation of the finite population total. Generally, use of the proposed estimator leads to relatively smaller values of relative efficiency compared to other estimators.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Background Information

Sample surveys' main objective is to obtain information about the population, and then use such information to make inference about some population quantities. The information that is mostly sought about the population is usually aggregate values of various population characteristics such as total number of units or proportion of units having certain attributes. The information can be collected by either sampling methods or census. While census is the complete enumeration of units in a population, sampling methods, which consist of sample selection from a specified population, make it possible to estimate various population quantities such as population totals, means or proportions. This is done while reducing the size of survey operations. Nonparametric regression may be used in the estimation of unknown finite population quantities such as population totals, means, proportions or averages. The idea of nonparametric regression traces its origin in works by Nadaraya (1964) and Watson (1964). Nonparametric-based estimation is often more robust and flexible than inference based on parametric regression models or design probabilities (as in design-based inference) (Dorfman, 1992). In sample surveys, auxiliary information is used at the estimation stage of finite population quantities- population total or mean, say - to increase the precision of estimators of such population quantities (Montanari and Ranalli, 2003, 2005; Sánchez-Borrego and Rueda, 2009). A variety of approaches exist for construction of more efficient estimators for population total or mean, and they include model-based and design-based methods. Model-based approach in sam-

ple surveys assumes that the population under study is a realization of a random variable having a superpopulation model ξ. This model ξ is used to predict the nonsampled values of the population, and hence the finite population quantities, total Y or mean $\bar{Y}$ (Sánchez-Borrego and Rueda, 2009).

In this thesis, auxiliary information is used to determine an estimator of finite population total using nonparametric regression under stratified random sampling. To achieve this, a model-based approach is adopted by making use of the local polynomial regression estimation to predict the values of the nonsampled set.

1.2 Statement of the Problem

Model-based approach in sample surveys assumes that the population under study is a realization of random variables having a superpopulation model ξ. Such model is used to predict the nonsampled values of the population, and hence the finite population quantities Y or mean $\bar{Y}$. In this approach, auxiliary information is often used at the estimation stage to increase the precision of estimators of population total or mean. Scientists have used supplementary population information on a character x to estimate finite population total Y or mean $\bar{Y}$ of a character y under study (Montanari and Ranalli, 2003, 2005; Sánchez-Borrego and Rueda, 2009; Orwa et al., 2010; Rady and Ziedan, 2014a,b). Most of the scientists have used estimators based on simple random sampling to estimate the finite population quantities Y or mean $\bar{Y}$ in local polynomial regression. Previous works involve the construction of estimators based on simple random sampling. Elsewhere as in Orwa et al. (2010) and Ngesa et al. (2012) ratio estimators based on stratified random sampling are proposed. Various approaches have been proposed in the efficient estimation of finite population totals or means, both model-based and design-based methods. This study seeks to propose a model-based estimator of the finite population total using local polynomial regression under

stratified random sampling. Stratified estimators for finite population total Y or mean $\bar{Y}$ have proved to yield better estimators than those resulting from simple random sampling (Orwa et al., 2010; Ngesa et al., 2012). Furthermore, it has been shown in the literature that local polynomial approximation method has several attractive features including satisfactory boundary behaviour, easy interpretability, applicability for a variety of design-circumstances and nice minimax properties (see for example Fan and Gijbels, 1992; Fan, 1993, and Ruppert and Wand, 1994). Hence the need for this study.

1.3 Objectives of the Study

1.3.1 Main Objective

To investigate the theoretical properties of a local polynomial regression estimator of the finite population total under stratified random sampling in a model-based approach.

1.3.2 Specific Objectives

(i) To determine a model-based estimator of the finite population total using local polynomial regression under stratified random sampling.

(ii) To investigate the asymptotic properties (asymptotic model unbiasedness and consistency) of the proposed estimator.

(iii) To compare the performance of the proposed estimator to that of existing ones namely Horvitz-Thompson estimator, the Linear regression estimator, and the Mixed ratio estimator using simulated data.

1.4 Significance of the Study

1.4.1 Contribution to Current Knowledge

Various efficient estimators of finite population totals or means, both model-based and design-based methods have been proposed in the literature. Estimators for finite population total Y or mean $\bar{Y}$ under stratified random sampling have been proved to yield better estimators (more efficient estimators) than those resulting from simple random sampling. However, there is no work in the local polynomial regression estimation of finite population totals under stratified random sampling. This piece of work will close this gap.

1.4.2 Application to the Real World

Stratified random sampling is the most common method in many surveys and its most important to develop estimators in this context that can be used by researchers in other fields such as social sciences, economics and geology. The local polynomial regression estimator being proposed can be used to make inference about some finite population quantities such as the finite population total. Furthermore, when the variance of the proposed estimator is derived, it can be used to construct confidence intervals for the finite population total or mean.

1.5 Scope of the Study

A local polynomial regression estimator of the finite population total in a model-based approach under stratified random sampling shall be determined. Then asymptotic properties of the proposed estimator such as asymptotic model unbiasedness and consistency will be determined. This is to mean that the performance of the model will be related to how close the estimated values are to the observed values. Different criteria will be used to compare the proposed estimator

with the Horvitz-Thompson estimator, the Linear regression estimator (Cochran, 1977, p. 200) and the Mixed ratio estimator of Orwa et al. (2010). Criteria for comparison include relative absolute bias (RAB) and relative efficiency (RE).

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter reviews the various studies that have so far been done in relation to local polynomial regression estimation in various designs. It highlights some key contributions in the study of local polynomial regression and the general methodologies that have been conducted in the study of this kind.

2.2 The Paradigm of Model-based Approach

Nonparametric methods have recently been employed in the estimation procedure of finite population parameters in a model-based framework. The history of the use of auxiliary information is as old as the history of survey sampling. Neyman (1938) work may be thought as the actual work where auxiliary information was used to improve the precision of an estimator. In his work, Neyman introduced the concept of double sampling in sampling human populations when the mean of the auxiliary variate is unknown. In the development of the ratio estimator by Cochran (1940), the survey variable was highly correlated with the auxiliary variable and the regression line passed through the origin. Similarly, Cochran (1977) used auxiliary information in proposing linear regression estimators in stratified random sampling.

One of the approaches to using auxiliary information in construction of estimators is by assuming a working model that describes the relationship between the survey variable and the auxiliary variable. Estimators are then derived based on this model. At this stage, estimators are sought to have good efficiency given

that the model is true. In most cases, a linear model is assumed. Generalized regression estimators by Cassel et al. (1976) and Robinson and Sarndal (1983) including linear regression estimators and ratio estimators by Cochran (1977), and best linear unbiased estimators by Royall (1970) and Brewer (1963) as well as post-stratification estimators by Holt and Smith (1979) are all derived from the assumption of linear models.

Sometimes the linear model is an inappropriate model, and therefore, the resulting estimators do not beat the purely design-based estimators. As a result, Wu and Sitter (2001) proposed a class of estimators in which the working model assumes a nonlinear parametric model. However, the improvement of the efficiency of such estimators requires prior information about the exact parametric population structure. As a result of these concerns, several researchers have so far considered nonparametric models for ξ. For instance, Nadaraya (1964) and Watson (1964) introduced the idea of nonparametric regression to estimate unknown regression functions. Later on, Dorfman (1992) and Chambers et al. (1993) introduced model-based nonparametric kernel-regression estimators for some finite population quantities and their distribution functions.

In model-based approach, the model ξ is used to predict the non-sampled values of the population, and hence the population quantities mean $\bar{Y}$ or total Y. In their work, Orwa et al. (2010) proposed a mixed ratio estimator for the finite population total under stratified random sampling in model based frameworks. This estimator was based on the Nadaraya-Watson kernel regression, and it generally led to a relatively small error as compared to the usual ratio estimator. Further this estimator was shown to be statistically consistent as well as asymptotically unbiased.

2.3 Local Polynomial Regression Estimation

The introduction of more general models and flexible techniques to obtain prediction of the value taken by the survey variable in non-sampled units seems of great interest more also when auxiliary information is available for each unit of the population. Breidt and Opsomer (2000) first considered nonparametric models for ξ within a model-assisted approach and obtained a local polynomial regression estimator as a generalization of the ordinary generalized regression estimator. On the other hand, Zheng and Little (2003) developed a model-based estimator based on penalized spline regression.

Sánchez-Borrego and Rueda (2009) considered a general working model through a nonparametric class of models ξ which is within the model-based approach to inference. In their work, Sánchez-Borrego and Rueda (2009) employ local polynomial regression in the estimation of the finite population mean except for that their estimator is only applicable to direct sampling designs such as simple random sampling. On the other hand, Breidt and Opsomer (2000) proposed a model-assisted nonparametric estimator for finite population total and this has attracted researchers. This estimator was based on local polynomial smoothing, which is a kernel-based technique. Although this estimator has the form of generalized regression estimator, it is based on a nonparametric superpopulation model ξ applicable to a large class of functions. Breidt and Opsomer (2000) local polynomial regression estimator only applies to direct sampling designs when auxiliary information is available for each unit of the population. When complete auxiliary information is available, the employment of more flexible methods to predict the value taken by the survey variable in non-sampled units produce more efficient estimators (Montanari and Ranalli, 2003). This is often done at the estimation stage in order to increase the precision of estimators of population parameters - population total or mean. Such auxiliary information can include

census data, administrative registers or previous surveys.

Sánchez-Borrego and Rueda (2009) improved on Breidt and Opsomer (2000) estimator and developed a model-based local polynomial regression estimator applicable to direct sampling designs such as simple random sampling and systematic sampling. This estimator is more reliable than the classical design-based approach, originally developed by Neyman (1934). The design-based approach has some weaknesses in that it is prescriptive for the choice of estimator. Godambe (1955) mentions that it lacks a theory for optimal estimation, and hence yields potentially inefficient estimates. The aforementioned limitations of classical design-based approach calls for the need to employ the model-based approach.

2.3.1 Some Local Polynomial Regression Estimators and their Theoretical Properties

Breidt and Opsomer (2000) used the traditional local polynomial regression estimator of the unknown function $m(.)$. In their work they assume that $m(.)$ is a smooth function in x and obtain consistent and design-unbiased estimators of the finite population total. Breidt and Opsomer (2000) local polynomial regression estimator has the form of the generalized regression estimator, but it is based on a nonparametric superpopulation model ξ which is applicable to a wide class of functions. Based on the empirical studies carried out, this estimator yields better results than the classical regression estimator and the post-stratification estimator of Cochran (1977), as well as the model-based nonparametric estimator of Dorfman (1992).

Sánchez-Borrego and Rueda (2009) extended Breidt and Opsomer (2000) idea to model-based approach and considered a general working model through a nonparametric class of models ξ which is within the model-based approach to inference. In their work, Sánchez-Borrego and Rueda (2009) employ local polynomial regression in the estimation of the finite population mean except for

that their estimator is applicable to direct sampling designs such as simple random sampling. Their estimator is asymptotically model-unbiased and also consistent. Based on the empirical studies, their estimator is better than the previous model-assisted local polynomial regression estimator of Breidt and Opsomer (2000) as it maintains the lowest relative efficiency in the Sine, Bump, Counties70 and Jump populations. Their estimator is asymptotically design-unbiased and a consistent estimator of the finite population total.

Most recently, Rady and Ziedan (2014b) has developed a local linear polynomial regression estimator of the finite population total. The difference between them and the aforementioned researchers is that they incorporate two auxiliary variables. They combine resampling methods together with local linear regression method in the estimation of a finite population total. Their empirical studies are based on the mean absolute error (MAE), mean squared error (MSE) and mean absolute percentage error (MAPE). More specifically, they consider two auxiliary variables and do an empirical study to compare the estimator of finite population total based on classical linear regression and local linear regression, and the effects of bootstrap and jackknife methods on these estimators. The local linear regression estimator beat the classical regression estimator when the model is misspecified, a proof of robustness.

By considering Breidt and Opsomer (2000); Sánchez-Borrego and Rueda (2009); Rady and Ziedan (2014b) ideas, local polynomial regression estimator of the finite population total in the case of stratified random sampling is suggested. In this study, the local polynomial regression estimator to stratified sampling when samples from each stratum are drawn using simple random sampling without replacement will be extended. The use of a model-based nonparametric approach (in the case of one auxiliary variable which is available for each unit of the population) will be considered.

2.4 Choice of degree, p of the polynomial

It is important to make a good choice of the appropriate degree of polynomial to fit. When choosing the degree, there is a trade-off between variance and bias. Higher order polynomials usually allow for precise fitting. This means that the bias will be small, but an increase in the degree results to an increase in the variance. However, such increase is not constant. Avery (2012) notes that the asymptotic variance of the function $\hat{m}(.)$ only increases whenever the degree, p changes from odd to even. For instance, there will be an increase in the asymptotic variance when moving from $p = 1$ to $p = 2$. However, there is no loss when moving from $p = 0$ to $p = 1$. This strongly supports the idea of choosing odd-degree polynomial (say $p = 1$) in the previous section since there is no associated cost in variance (Fan and Gijbels, 1992; Ruppert and Wand, 1994). Fan and Gijbels (1992) propose an adaptive method for selecting the correct degree of polynomial based on local factors and they allow p to vary for different points in the support of the data. The estimator that results has the property of robustness to bandwidth. This implies that if the selected bandwidth is too large, a polynomial of higher degree will be selected to better model the contours of the data. Similarly, if the selected bandwidth is too small, then a polynomial of lower degree will be fit in order to help reduce the variance and make the estimates numerically stable.

2.5 Bandwidth Selection

The choice of bandwidth, b, is critically important in local polynomial regression. Bandwidths control the complexity or the "jaggedness" of the fit. Smaller values of b will lead to less smoothing whereas larger values result to a regression curve with fewer sharp changes. Furthermore, there exists a trade-off

between bias and variance. Larger values for b will reduce the variance. This is because more local points will be included in the estimate. However, an increase in b leads to an increase in the average distance between x_0 and these local points, and consequently to a larger bias. Fan and Gijbels (1996) notes that a natural way to select bandwidth, and balance such trade-off is by minimizing the MSE.

2.6 Choice of the Kernel Function

Many possible kernel smoothers exist. However, the selected kernel should be easy to implement both practically and theoretically. Silverman (1986) listed the requirements that a kernel smoother has to meet, and they include:

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

(ii) The smoother should be user-friendly. That is, it should be theoretically and practically fit in both natural and simulated data.

(iii) The smoother should not take very small values since this may result to numerical underflow in the computer.

(iv) The range of values that the kernel smoother takes should be well-defined and not open as in the Gaussian kernel case.

Table 2.1 gives the efficiencies of several kernels with respect to the Epanechnikov kernel.

Kernel	$K(u)$	***Efficiency***
Epanechnikov	$\frac{3}{4}(1-u^2)$, $\|u\|<1$	1.000
Biweight	$\frac{15}{16}(1-u^2)^2$, $\|u\|<1$	0.9939
Triangular	$1-\|u\|$, $\|u\|<1$	0.9859
Gaussian	$\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}u^2\right)$, $-\infty<u<\infty$	0.9512
Rectangular	$\frac{1}{2}$, $\|u\|<1$	0.9295

Table 2.1: *Efficiency Relative to the Epanechnikov kernel.*

The performance of the kernel function is usually measured by the mean integrated square error (MISE) or asymptotic mean integrated square error (AMISE). Epanechnikov kernel, in this case, minimizes the AMISE and is therefore optimal. Hence the best choice for a study of this kind.

CHAPTER 3

METHODOLOGY

3.1 Introduction

The model-based approach, which is based on superpopulation models, assumes that the population under investigation is a realization of a superpopulation random variable having a superpopulation model ξ. With this model ξ, one is able to predict the nonsampled values of the population, and hence the population total Y. Superpopulation models visualize $y_i's$ as realized values of a random variable y but the sample is fixed.

Suppose there is a population of N units: $U = \{1, 2, \ldots\ldots, N\}$. Suppose a random sample s of size n is selected according to some sampling design, say d, with the first order inclusion probability π_i.

Let y_i be the value of the study variable y, for the i^{th} population element. Let also x_i be the auxiliary variate which is available for all population units in U and associated to y_i. The values of two variables $(y_i, x_i), i \in s$ are observed for the estimation of finite population total Y. Suppose the distribution generating $y_i's$ is given by ξ.

Here, the problem of estimating Y is basically the problem of predicting the sum of unobserved random variable. From the sample, inference is made about ξ and then used to predict $\sum_{i \in r} y_i$, where r is the nonsampled set.

Usually in the computation of finite population total, we have the formula given in Equation (3.1).

$$Y = \sum_{i=1}^{N} y_i = \sum_{i \in s} y_i + \sum_{i \in r} y_i \tag{3.1}$$

The first component in Equation (3.1) is known while the second requires prediction which is the focus in this particular work. Various methods exist for this purpose. In this context, focus is on the local polynomial regression, which is a kernel-based technique.

3.2 Local Polynomial Regression

3.2.1 Introduction

In this subsection, local polynomial regression is discussed in a broader context. Local polynomial regression is typically used to model the relationship between a survey variable and the auxiliary variable. In the presence of an auxiliary variable, x, a natural way to predict the unknown component in Equation (3.1) is by adopting the regression model in Equation (3.2) that treats the proxy values $y_i^0 = m(x_i)$ as the predicted values of the unobserved values $y_i, i \in r$.

$$y_i = m(x_i) + e_i \tag{3.2}$$

Here, local polynomial regression estimation, a kernel-based method, is adopted to estimate the unknown function $m(x_i)$ for $i \in r$. The unknown function $m(x_i)$ is assumed to be a smooth function. The error terms are assumed to be independently distributed random variables with mean 0 and variance $\sigma^2(x)$.

The approaches employed by Breidt and Opsomer (2000),Sánchez-Borrego and Rueda (2009) and Rady and Ziedan (2014b) are used. The local polynomial kernel estimator is used to predict y_i, $i \in r$ in the context of stratified random sampling.

3.2.2 Estimation of the unknown function $m(.)$

Assuming that the unknown function $m(.)$ in Equation (3.2) is a smooth function in x. No global assumptions about $m(.)$ are made except the assumption that the function can locally be approximated with a member of a simple group of parametric functions, for instance, a straight line or a constant.

Let $K_b(u) = b^{-1}K(u/b)$, where K denotes a continuous kernel function and b is the bandwidth.

In nonparametric regression, focus is on estimation of the unknown function $m(.)$. Now consider a Taylor's expansion of the unknown function $m(.)$ for x_i in the neighborhood of a point of interest, x_0:

$$m(x_i) \approx m(x_0) + m^{'}(x_0)(x_i - x_0) + ... + m^{(p)}(x_0)(x_i - x_0)^p \frac{1}{p!} \tag{3.3}$$

Note that Taylor's theorem says that any $k-$times differentiable or continuous function can be approximated with a polynomial.

Define the local kernel weights as $w_i = K_b(x_i - x_0)$. To estimate the terms on the right hand side of Equation (3.3), we adopt the weighted least-squares regression in the following way:

Equation (3.4) is fitted

$$y_i = \beta_0 + \beta_1(x_i - x_0) + \beta_2(x_i - x_0)^2 + ... + \beta_p(x_i - x_0)^p + e_i \tag{3.4}$$

to minimize the local weighted residual sum of squares given by

$$\sum_{i=1}^{n} w_i e_i^2 = \sum_{i=1}^{n} \left[y_i - \sum_{j=0}^{p} \beta_j \left(x_i - x_0\right)^j \right]^2 K_b \left(x_i - x_0\right) \tag{3.5}$$

That is,

$$\min_{\beta} \sum_{i=1}^{n} \underbrace{\left\{y_i - \beta_0 - \beta_1(x_i - x_0) - \beta_2(x_i - x_0)^2 - ..... - \beta_p(x_i - x_0)^p\right\}^2}_{polynomial} \underbrace{K_b(x_i - x_0)}_{local} \tag{3.6}$$

where $\beta = (\beta_0, \beta_1, \beta_2, ..., \beta_p)^T$. Denoting the solution to Equation (3.6) as $\hat{\beta}(x)$. Then $\hat{m}^{(r)}(x_0) = r!\hat{\beta}_r(x)$, $(r = 0, 1, 2, ..., p)$.

Accordingly,

$$\hat{\beta}(x) = (X_{sj}^T W_{sj}^T X_{sj})^{-1} X_{sj}^T W_{sj}^T Y_s \tag{3.7}$$

as long as $X_{sj}^{'} W_{sj}^{\prime} X_{sj}$ is invertible.

Thus from Equation (3.7), a model-based local polynomial regression estimator based on the whole finite population (as in Breidt and Opsomer, 2000; Sánchez-Borrego and Rueda, 2009; Rady and Ziedan, 2014b) would be given by:

$$\hat{m}(x_0) = e_1^T (X_{sj}^T W_{sj}^T X_{sj})^{-1} X_{sj}^T W_{sj}^T Y_s = w_{sj}^T Y_s \tag{3.8}$$

where $e_1 = (1, 0, 0, ..., 0)^T$ is a column vector of length $p + 1$; $Y_s = [y_i]_{i \in s}$;

$W_{sj} = diag\{K_b(x_i - x_0)\}_{i \in s}$ and $X_{sj} = [1, (x_i - x_0), ..., (x_i - x_0)^p]_{i \in s}$.

Equation (3.8) holds as long as $X_{sj}^T W_{sj}^T X_{sj}$ is a nonsingular matrix.

Thus in estimating each $y(x)$ the following 3 steps need to be followed:

(i) Construct X_{sj} matrix for each $x \in \{x_1, ..., x_n\}$

(ii) Construct W_{sj} matrix for each $x \in \{x_1, ..., x_n\}$

(iii) Estimate $m(x)$ using equation (3.8).

Then the model-based local polynomial regression estimator for Y is given by

$$\hat{Y}_{lp} = \sum_{i \in s} y_i + \sum_{i \in r} \hat{m}_i \tag{3.9}$$

where r is the nonsampled set.

In this study, the local polynomial regression estimation in simple random sampling is extended to stratified random sampling using one auxiliary variable. The Epanechnikov kernel with bandwidth values $b = n^{-1/5}$ (see Rady and Ziedan, 2014b) (with $n = 200$), $b = 0.4$, $b = 1$ and $b = 2$ (see Orwa et al., 2010) is used for the nonparametric estimators.

3.3 Proposed Estimator

Consider a population consisting of N units. Suppose this population is divided into H disjoint strata, where h^{th} is of size N_h, $h = 1, 2, ..., H$.

Let y_{hj}, $j = 1, 2, ..., N_h$ be the survey measurement for the j^{th} unit in the h^{th} stratum. Further, let x_{hj}, $j = 1, 2, ..., N_h$ be the auxiliary measurement positively correlated with y_{hi}.

From the h^{th} stratum, a simple random sample of size n_h is selected without replacement, where n_h is sufficiently large with respect to N_h and $f_h = n_h/N_h \longrightarrow 0$.

Let s_h be the sample in the h^{th} stratum and r_h be the nonsampled set in the h^{th} stratum.

The population total is defined as

$$Y = \sum_{h=1}^{H}\sum_{j=1}^{N_h} y_{hj} = \sum_{h=1}^{H}\sum_{j=1}^{n_h} y_{hj} + \sum_{h=1}^{H}\sum_{j=n_h+1}^{N_h} y_{hj} \tag{3.10}$$

which can rewritten as

$$Y = \sum_{h=1}^{H} y_{h_s} + \sum_{h=1}^{H} y_{h_r} \tag{3.11}$$

where $y_{h_s} = \sum_{j=1}^{n_h} y_{hj}$ and $y_{h_r} = \sum_{j=n_h+1}^{N_h} y_{hj}$.

Once the sample has been observed, the problem of estimating Y becomes the problem of predicting the sum of the nonsampled $y'_{hj}s$. Usually, inference is made using the known sample and the model ξ.

The first component in Equation (3.10) is known while the second requires prediction which is the focus in this study. In this study, local polynomial regression method will be used to predict the unknown $y'_{hj}s$, $\forall j \in r_h$.

Suppose the distribution generating $y'_{hj}s$ is given by the superpopulation model, ξ in which

$$y_{hj} = m(x_{hj}) + e_{hj} \tag{3.12}$$

where $e'_{hj}s$ are independently distributed random variables with mean 0 and variance $\sigma^2(x_{hj})$.

Then it follows that

$$E(y_{hj}) = m(x_{hj}) \tag{3.13}$$

$$Cov(y_{hj}, y_{h'j'}) = \begin{cases} \sigma^2(x_{hj}), & for \quad h = h' \quad and \quad j = j' \\ 0, & otherwise \end{cases} \tag{3.14}$$

where $\sigma^2(x)$ and $m(x)$ are assumed to be continuous and twice differentiable functions of x, and $\sigma^2(x) > 0$.

From equations (3.3), (3.4), (3.5), (3.6), (3.8), and (3.9), a model-based local polynomial regression estimator of the nonsampled $y'_{hj}s$ in the h^{th} stratum is therefore given by

$$\hat{m}_{hj} = e_1^T \left(X_{hj}^T W_{hj} X_{hj}\right)^{-1} X_{hj}^T W_{hj}^T y = w_{hj}^T y \tag{3.15}$$

where $e_1 = (1, 0, 0, ..., 0)^T$ is a column vector of length $p + 1$; $y = [y_{hj}]_{j \in s_h}$;

$W_{hj} = diag\left\{K_b\left(x_{hj} - x_{hi}\right)\right\}_{j \in s_h}$ and $X_{hj} = \left[1, \left(x_{hj} - x_{hi}\right), ..., \left(x_{hj} - x_{hi}\right)^p\right]_{j \in s_h}$. Equation (3.15) holds as long as $X_{hj}^T W_{hj} X_{hj}$ is a nonsingular matrix.

Now denoting the estimator for the finite population total by $\hat{Y}_{LP}$ and the estimator within the h^{th} stratum by $\hat{Y}_{LP_h}$. Therefore, in stratum h, the estimator of the population total based on local polynomial regression is

$$\hat{Y}_{LP_h} = y_{h_s} + \sum_{j=n_h+1}^{N_h} \hat{m}_{hj} \tag{3.16}$$

and the estimator for the finite population total is

$$\hat{Y}_{LP} = \sum_{h=1}^{H} \hat{Y}_{LP_h} = \sum_{h=1}^{H} \left(y_{h_s} + \sum_{j=n_h+1}^{N_h} \hat{m}_{hj} \right) \tag{3.17}$$

with $y_{h_s} = \sum_{j=1}^{n_h} y_{hj}$.

3.4 Properties of Proposed Estimator

In this section, a study is carried out on various properties of estimator (3.17), which may be important in practice. Assumptions made are as follows:

(i) The regression function $m(x)$ has a bounded second derivative.

(ii) The marginal density, $f_X(x)$ is continuous and $f_X(x) > 0$.

(iii) The conditional variance $\sigma^2(x) = \text{var}(Y/X = x)$ is bounded and continuous.

(iv) The kernel density function $K(x)$ is bounded and continuous satisfying: $\int_{-\infty}^{\infty} K(x)dx = 1$,

$\int_{-\infty}^{\infty} xK(x)dx = 0$, $\int_{-\infty}^{\infty} x^2K(x)dx > 0$ and $\int_{-\infty}^{\infty} x^{2t}K(x)dx < \infty$ for $t = 1, 2, ...$

These conditions on $K(\cdot)$ were imposed and used in Fan (1993) work and are purposely for the convenience of technical arguments and therefore can be relaxed.

3.4.1 $\hat{Y}_{LP}$ is Asymptotically Model-Unbiased

Now consider the difference:

$$\hat{Y}_{LP} - Y = \left(\sum_{h=1}^{H} y_{h_s} + \sum_{h=1}^{H}\sum_{j\in r_h} \hat{m}_{hj}\right) - \left(\sum_{h=1}^{H} y_{h_s} + \sum_{h=1}^{H}\sum_{j\in r_h} y_{hj}\right) \tag{3.18}$$

$$= \sum_{h=1}^{H}\sum_{j\in r_h} (\hat{m}_{hj} - y_{hj}) \tag{3.19}$$

$$= \sum_{h=1}^{H}\sum_{j\in r_h} ((\hat{m}_{hj} - m_{hj}) + (m_{hj} - y_{hj})) \tag{3.20}$$

and taking expectation yields

$$E_\xi\left(\hat{Y}_{LP} - Y\right) = \sum_{h=1}^{H}\sum_{j\in r_h} E_\xi\left(\hat{m}_{hj} - m_{hj}\right) + \sum_{h=1}^{H}\sum_{j\in r_h} E_\xi\left(m_{hj} - y_{hj}\right) \tag{3.21}$$

$$= \sum_{h=1}^{H}\sum_{j\in r_h} E_\xi\left(\hat{m}_{hj} - m_{hj}\right) \tag{3.22}$$

since $E_\xi\left(y_{hj}\right) = m_{hj}$

i.e.

$$E_\xi\left(\hat{Y}_{LP} - Y\right) = \sum_{h=1}^{H}\sum_{j\in r_h} E_\xi\left(\hat{m}_{hj} - m_{hj}\right) \tag{3.23}$$

which is the bias associated with $\hat{Y}_{LP}$.

Approximating m_{hj} by Taylor series expansion about a point x_{hj} and assuming further that $n_h \longrightarrow \infty$ and $b \longrightarrow 0$, then observe that

$$\hat{m}_{hj} \approx m_{hj} + m_{hj}^{'}(x_{hj} - x_{hi}) + (1/2)\, m_{hj}^{''}(x_{hj} - x_{hi})^2 + ... \tag{3.24}$$

Letting $u = (x_{hj} - x_{hi})/b \Longrightarrow ub = x_{hj} - x_{hi}$, then

$$\hat{m}_{hj} \approx m_{hj} + m_{hj}^{'}(ub) + (1/2)\, m_{hj}^{''}(ub)^2 + O(b^2) \tag{3.25}$$

$$\Longrightarrow \hat{m}_{hj} - m_{hj} \approx m_{hj}^{'}(ub) + (1/2)\, m_{hj}^{''}(ub)^2 + O(b^2) \tag{3.26}$$

and applying expectations then

$$E_\xi\left(\hat{m}_{hj} - m_{hj}\right) = E_\xi\left(\left(m_{hj}^{'}(ub) + (1/2)\, m_{hj}^{''}(ub)^2\right)\right) + O(b^2) \tag{3.27}$$

Theorem 3 of Fan and Gijbels (1996) allows that under conditions $(1) - (4)$ if $b \longrightarrow 0$ and $n_h b \longrightarrow \infty$,

$$E_\xi\left(m_{hj}^{'}(ub) + (1/2)\, m_{hj}^{''}(ub)^2\right) + O(b^2) \longrightarrow$$

$$m_{hj}^{'} b \int uK_b(u)du + (1/2)\, m_{hj}^{''} b^2 \int u^2 K_b(u)du + O(b^2) \tag{3.28}$$

$$= (1/2)\, m_{hj}^{''} b^2 \int u^2 K_b(u)du + O(b^2) \tag{3.29}$$

So that

$$E_\xi\left(\hat{m}_{hj} - m_{hj}\right) = (1/2)\, m_{hj}^{''} b^2 \int u^2 K_b(u)du + O(b^2) \tag{3.30}$$

It implies that $E_\xi\left(\hat{m}_{hj} - m_{hj}\right) \longrightarrow 0$ provided that $b \longrightarrow 0$ and $n_h \longrightarrow \infty$, and thus $\hat{Y}_{LP}$ is asymptotically model-unbiased.

3.4.2 Mean Square Error (MSE) of $\hat{Y}_{LP}$

The estimator (3.6) has the MSE

$$MSE(\hat{Y}_{LP}) = E_\xi\left(\hat{Y}_{LP} - Y\right)^2 \tag{3.31}$$

which can be decomposed as

$$MSE(\hat{Y}_{LP}) = \left[Bias\left(\hat{Y}_{LP}\right)\right]^2 + Var\left(\hat{Y}_{LP}\right) \tag{3.32}$$

Theorem 1 of Fan (1993) allows that under Condition (1), if the bandwidth, b is optimal i.e. $b = dn_h^{-\gamma}$ for $0 < \gamma < 1$, then

$$MSE(\hat{Y}_{LP}) = \left(\tfrac{b^4}{4}\right) \sum_{h=1}^{H} \sum_{j \in r_h} \left(m''_{hj} \int_{-\infty}^{\infty} u^2 K_b(u) du \right)^2 +$$

$$\frac{1}{b} \sum_{h=1}^{H} \sum_{j \in r_h} \frac{1}{n_h} f^{-1}(x_{hj}) \sigma^2(x_{hj}) \int_{-\infty}^{\infty} K_b^2(u) + O\left(b^4 + \frac{1}{n_h b}\right) \tag{3.33}$$

Observe that Equation (3.33) tends to zero if $b \longrightarrow 0$ and $n_h b \longrightarrow \infty$ and thus $MSE(\hat{Y}_{LP}) \longrightarrow 0$.

This shows that $\hat{Y}_{LP}$ is statistically consistent and thus useful.

3.5 Existing Estimators under Stratified Random Sampling

3.5.1 The Horvitz-Thompson Estimator

Under stratified random sampling, the Horvitz-Thompson (HT) estimator (Horvitz and Thompson, 1952) is given by,

$$\hat{Y}_{HT} = \sum_{h=1}^{H} \sum_{j=1}^{n_h} \frac{y_{hj}}{\Pi_{hj}} \tag{3.34}$$

Since the strata are independent, the variance of the HT estimator becomes

$$Var\left(\hat{Y}_{HT}\right) = \sum_{h=1}^{H} \left[\sum_{i=1}^{N_h} \sum_{j=1}^{N_h} \left(\pi_{ij} - \pi_i \pi_j\right) \frac{y_i}{\pi i} \frac{y_j}{\pi j} \right] \tag{3.35}$$

An unbiased estimator of $Var\left(\hat{Y}_{HT}\right)$ is given by,

$$\hat{Var}\left(\hat{Y}_{HT}\right) = \sum_{h=1}^{H}\left[\sum\sum_{i\in s_h j\in s_h}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\right)\frac{y_i}{\pi i}\frac{y_j}{\pi j}\right] \tag{3.36}$$

if $\pi_{ij} > 0 \ \forall \ i,j \in U_h$, for $h = 1, 2, ..., H$.

3.5.2 The Linear Regression Estimator

In Cochran (1977, p. 200), a regression estimate is given for each stratum population mean, i.e.

$$\hat{\bar{Y}}_{REG_h} = \bar{y}_{h_s} + \beta_h^o\left(\bar{X}_h - \bar{x}_h\right) \tag{3.37}$$

where $\bar{y}_{h_s}$ and $\bar{x}_h$ are stratum sample means; $\bar{X}_h$ is stratum population mean for the auxiliary variate, x; β_h^o is the regression coefficient in the h^{th} stratum.
The sample estimate for β_h^o is taken as

$$b_h^o = \frac{\sum\limits_j\left(y_{hj}-\bar{y}_{h_s}\right)\left(x_{hj}-\bar{x}_h\right)}{\sum\limits_j\left(x_{hj}-\bar{x}_h\right)^2} \tag{3.38}$$

Then given that $W_h = \frac{N_h}{N}$,

$$\hat{\bar{Y}}_{REG} = \sum_{h=1}^{H} W_h\hat{\bar{Y}}_{REG_h} \tag{3.39}$$

This estimation is appropriate when the true regression coefficients, β_h^o are thought to be varying from stratum to stratum.
From equation (3.39), a regression estimator for the total is given by,

$$\hat{Y}_{REG} = \sum_{h=1}^{H} N_h\hat{\bar{Y}}_{REG_h} = \sum_{h=1}^{H} N_h\left(\bar{y}_{h_s} + \beta_h^o\left(\bar{X}_h - \bar{x}_h\right)\right) \tag{3.40}$$

The variance $Var(\hat{Y}_{REG})$ is given by

$$Var\left(\hat{Y}_{REG}\right) = \sum_{h} \frac{N_h^2 \left(1 - f_h\right)}{n_h} S_{y_h}^2 \left(1 - \rho_h^2\right) \tag{3.41}$$

provided n_h is large enough in all strata; where ρ_h is the population correlation between y_{hj} and x_{hj}, and S_{y_h} is the population variance of y_{hj}.
And an estimator of $Var\left(\hat{Y}_{REG}\right)$ becomes

$$\hat{Var}\left(\hat{Y}_{REG}\right) = \sum_{h} \frac{N_h^2 \left(1 - f_h\right)}{n_h \left(n_h - 1\right)} \sum_{j} \left[\left(y_{hj} - \bar{y}_{h_s}\right) - b_h^o \left(x_{hj} - \bar{x}_h\right)\right]^2 \tag{3.42}$$

3.5.3 The Mixed Ratio Estimator

The mixed ratio estimator (Orwa et al., 2010) is based on the model

$$E(y_{hj}) = m(x_{hj}) \tag{3.43}$$

$$Cov(y_{hj}, y_{h'j'}) = \begin{cases} \sigma^2(x_{hj}), & for \qquad h = h' \quad and \quad j = j' \\ 0, & otherwise \end{cases} \tag{3.44}$$

where $\sigma^2(x)$ and $m(x)$ are assumed to be continuous and twice differentiable functions of x, and $\sigma^2(x) > 0$.
Letting the smoothing weight in the h^{th} stratum be given by

$$w_{hj}(x) = \frac{K_b\left(\frac{x_{hj} - x_{hi}}{b}\right)}{\sum_{s} K_b\left(\frac{x_{hj} - x_{hi}}{b}\right)} \tag{3.45}$$

Then an estimator for the nonsampled units in the h^{th} stratum becomes

$$\hat{m}(x_{hj}) = \sum_{s} w_{hj}(x) y_{hj} \tag{3.46}$$

Hence, the nonparametric regression estimator, $\hat{Y}_{PE}$, for the finite population total becomes

$$\hat{Y}_{PE} = \sum_{h=1}^{H} \sum_{j=1}^{n_h} y_{hj} + \sum_{h=1}^{H} \sum_{j=n_h+1}^{N_h} w_{hj}(x_j) y_{hj} \tag{3.47}$$

CHAPTER 4

SIMULATION STUDY

4.1 Introduction

In this section, a study is carried out on the practical performance of several estimators:

$\hat{Y}_{HT}$	Horvitz-Thompson	Equation (3.34)
$\hat{Y}_{REG}$	Linear regression	Equation (3.40)
$\hat{Y}_{PE}$	Mixed Ratio	Equation (3.47)
$\hat{Y}_{LP}$	Local polynomial with degree, $p = 1$	Equation (3.17)

Horvitz-Thompson estimator is design-based, linear regression estimator is parametric and model-based while both mixed ratio and local polynomial regression estimators are nonparametric and model-based.

4.2 Description of the Population

For comparison of the proposed estimator with the Horvitz-Thompson estimator (Horvitz and Thompson, 1952), the linear regression estimator (Cochran, 1977, p. 200), and the mixed ratio estimator of Orwa et al. (2010), simulated populations are used. The working model is taken to be $E(y_{hj}) = m(x_{hj})$, $Cov(y_{hj}, y_{h'j'}) = \sigma^2$, for $h = h'$ and $j = j'$ (i.e. constant variance).

In this study, four populations are considered, which are generated from the regression model given by

$$y_i = m(x_i) + e_i \tag{4.1}$$

$1 \leq i \leq 2,000$ with the following mean functions

$$\text{Linear: } m_1(x) = 1 + 2(x - 0.5)$$

$$\text{Sine: } m_2(x) = 2 + sin(2\pi x)$$

$$\text{Bump: } m_3(x) = 1 + 2(x - 0.5) + \exp(-200(x - 0.5)^2)$$

$$\text{Jump: } m_4(x) = 1 + 2(x - 0.5)I_{\{x \leq 0.65\}} + 0.65I_{\{x > 0.65\}}$$

with $x \in [0, 1]$. They represent a class of correct and incorrect model specifications for the estimators being considered. If it is assumed that the model is linear, then it would be interesting to check how much efficiency is lost by deviating from this assumption of linearity and assuming that the underlying model is smooth. For m_1, $\hat{Y}_{REG}$ is expected to be the best estimator, since the model assumed is correctly specified. The rest of the mean functions:- m_2, m_3 and m_4 represent various deviations from the linear model, m_1. These populations are plotted in Fig. 4.1. They were used by Breidt and Opsomer (2000) and the Sine, Bump and Jump populations were used by Sánchez-Borrego and Rueda (2009).

The errors are assumed to be independent, identically distributed (i.i.d) normal variables with mean 0 and standard deviation $\sigma = 0.1$. They contain $2,000$ units and the population x_i is simulated as i.i.d uniform random variables. The populations, $y_i's$, are generated from the mean functions by adding the errors $e_i's$ in each of the cases.

Data simulations, the estimators and computations were obtained using R Software on a desktop.

In order to study the practical performance of the proposed estimator, each of the populations (i.e. $y_i's$) is divided into 10 equal, disjoint and mutually exclusive strata which are made as homogenous as possible to ensure that units in each stratum vary little from each other. A sample of size, $n = 200$ is then taken with each stratum contributing a sample size of $n_h = 20$, $(h = 1, 2, ..., 10)$. 1000 samples are simulated using simple random sampling without replacement for each case.

Epanechnikov kernel,

$$K(u) = \frac{3}{4}\left(1 - u^2\right) I_{\{|u| \leq 1\}},$$

is used for kernel smoothing on each of the populations. In each case, bandwidth values $b = n^{-1/5}$ (see Rady and Ziedan, 2014b) (with $n = 200$), $b = 0.4$, $b = 1$ and $b = 2$ (see Orwa et al., 2010) are considered.

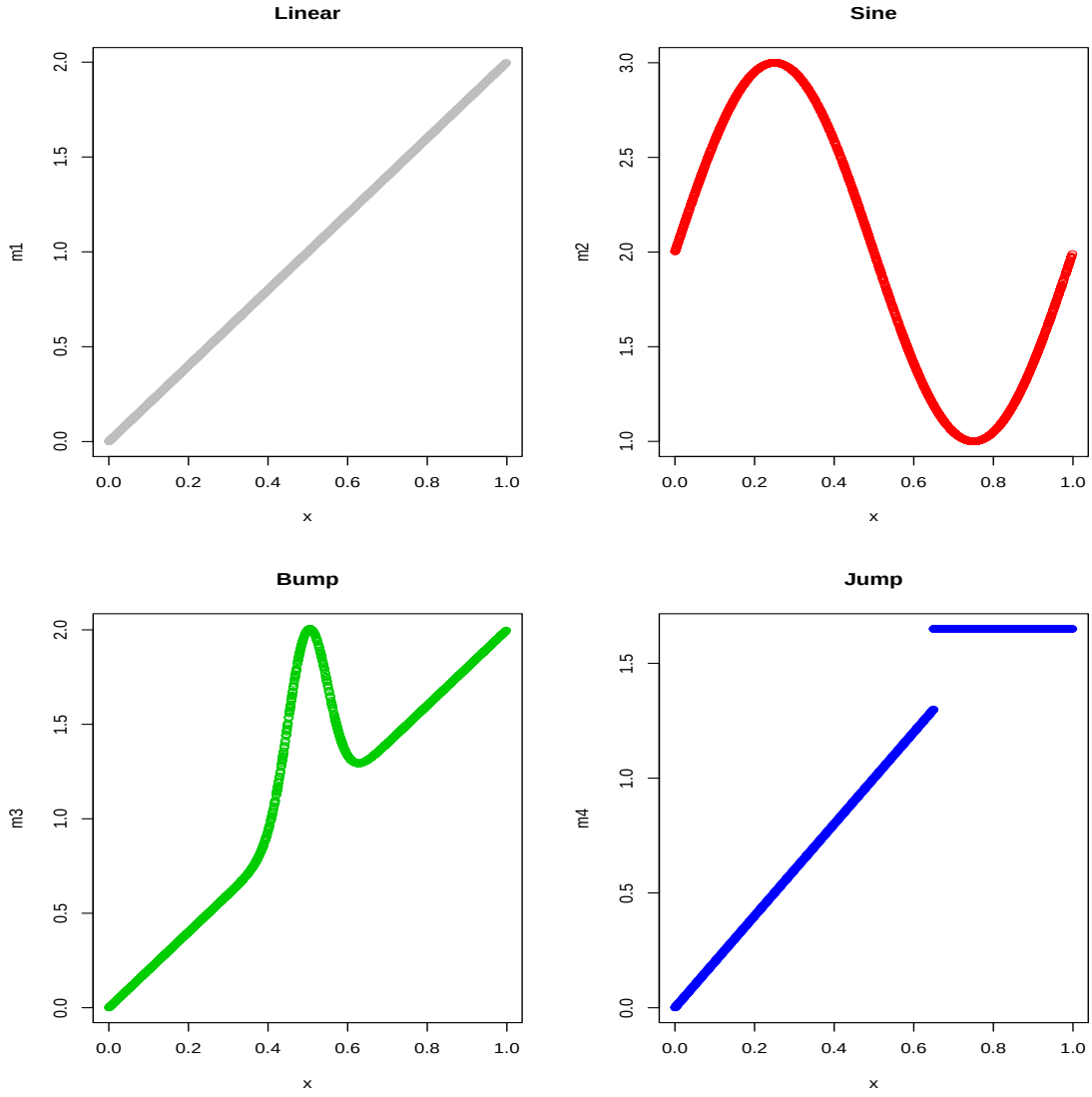

Figure 4.1: *Plot of Linear, Sine, Bump and Jump populations*

4.3 Performance Criteria of the Proposed Estimator

To analyze the performance of the proposed estimator against some specified estimators, relative absolute bias (RAB) is computed as

$$RAB(\hat{\theta}) = \sum_{i=1}^{R} \left| \frac{\left(\hat{\theta}(s_i) - Y\right)}{Y} \right| \tag{4.2}$$

and the relative efficiency (RE) with respect to the Horvitz-Thompson (HT) estimator is computed as

$$RE(\hat{\theta}) = \frac{\sum_{i=1}^{R} \left(\hat{\theta}(s_i) - Y\right)^2}{\sum_{i=1}^{R} \left(\hat{Y}_{HT}(s_i) - Y\right)^2} \tag{4.3}$$

$\hat{\theta}$ is the estimator of the finite population total being considered; Y is the true population total and R is the number of replications.

The relative efficiency is meant to examine the robustness of the various estimators against the proposed estimator.

4.4 Results

The results of this simulation study are summarized in Table 4.2. For each populations, $y_i's$ $(i = 1, 2, 3, 4)$, the performance of each estimator is analyzed using the RAB and RE. The RAB indicates the measure of how close the estimator being considered is from the actual value, while the RE is used to check the robustness of the estimator. For instance, an estimator, $\hat{\theta}_1$, will be said to be "better" or more preferrable than another one, $\hat{\theta}_2$, if its RE is comparably smaller. That is, if $RE(\hat{\theta}_1) < RE(\hat{\theta}_2)$, where $\hat{\theta}_1$ and $\hat{\theta}_2$ are estimators, then $\hat{\theta}_1$ is said to be "better" than $\hat{\theta}_2$.

Estimator	***Formulae***
Horvitz-Thompson, $\hat{Y}_{HT}$	$\hat{Y}_{HT} = \sum_{h=1}^{H}\sum_{j=1}^{n_h} \frac{y_{hj}}{\Pi_{hj}}$
Linear regression estimator, $\hat{Y}_{REG}$	$\hat{Y}_{REG} = \sum_{h=1}^{H} N_h\left(\bar{y}_{h_s} + \beta_h^o\left(\bar{X}_h - \bar{x}_h\right)\right)$
Mixed Ratio Estimator, $\hat{Y}_{PE}$	$\hat{Y}_{PE} = \sum_{h=1}^{H}\sum_{j=1}^{n_h} y_{hj} + \sum_{h=1}^{H}\sum_{j=n_h+1}^{N_h} w_{hj}(x_j)y_{hj}$
Proposed Model-based local polynomial, $\hat{Y}_{LP}$	$\hat{Y}_{LP} = \sum_{h=1}^{H} y_{h_s} + \sum_{h=1}^{H}\sum_{j=n_h+1}^{N_h} \hat{m}_{hj}$

Table 4.1: *Summary of the formulae used in computing the respective population totals of the various estimators*

The following plots demonstrate the effects of increasing the bandwidth on the proposed estimator $(\hat{Y}_{LP})$. They demostrate the effects of bandwidth on the complexity or "jaggedness" of the fit. Figures 4.2 - 4.5 represent the graphs of samples from various simulated $y-$ populations (Linear, Sine, Bump and Jump populations respectively) against the samples from the $x-$ population. In this illustration, four bandwidth values are used: $b = n^{-1/5}$ (with $n = 200$), $b = 0.4$, $b = 1$ and the data-driven bandwidth computed using the function "regCVBwSelC" in the R package "locpol". Smaller values of bandwidth, b, results to less smoothing while larger values yield curves with fewer sharp changes (see

Figures 4.2 - 4.5). In this illustration, high bandwidth values oversmooths the nonparametric regression curves as expected.

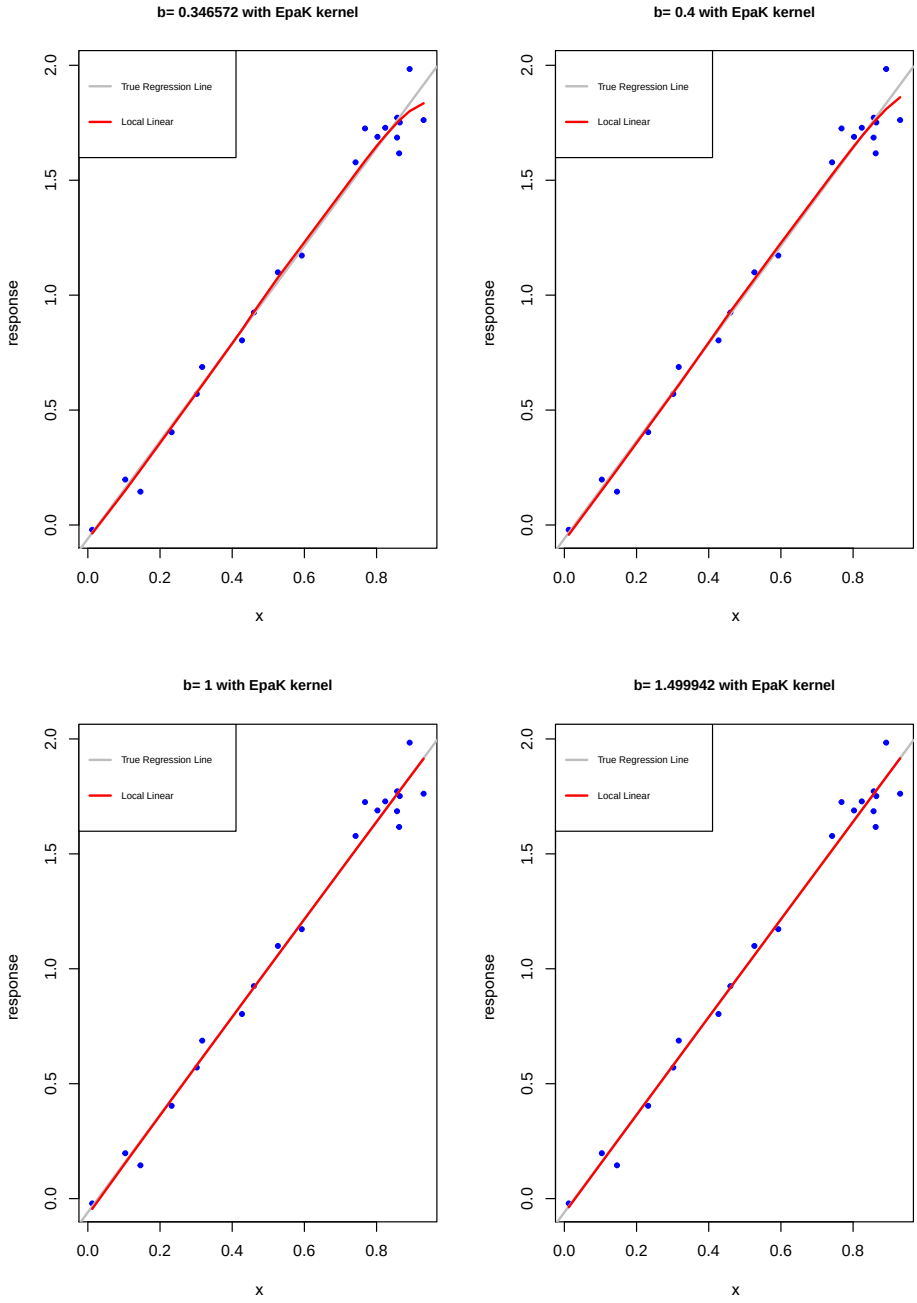

Figure 4.2: *Plots of the Simulated data (Stratum 1 of Linear population) with the true regression curve (gray line) and local linear smoother using an Epanechnikov kernel and various bandwidth values (red line)*

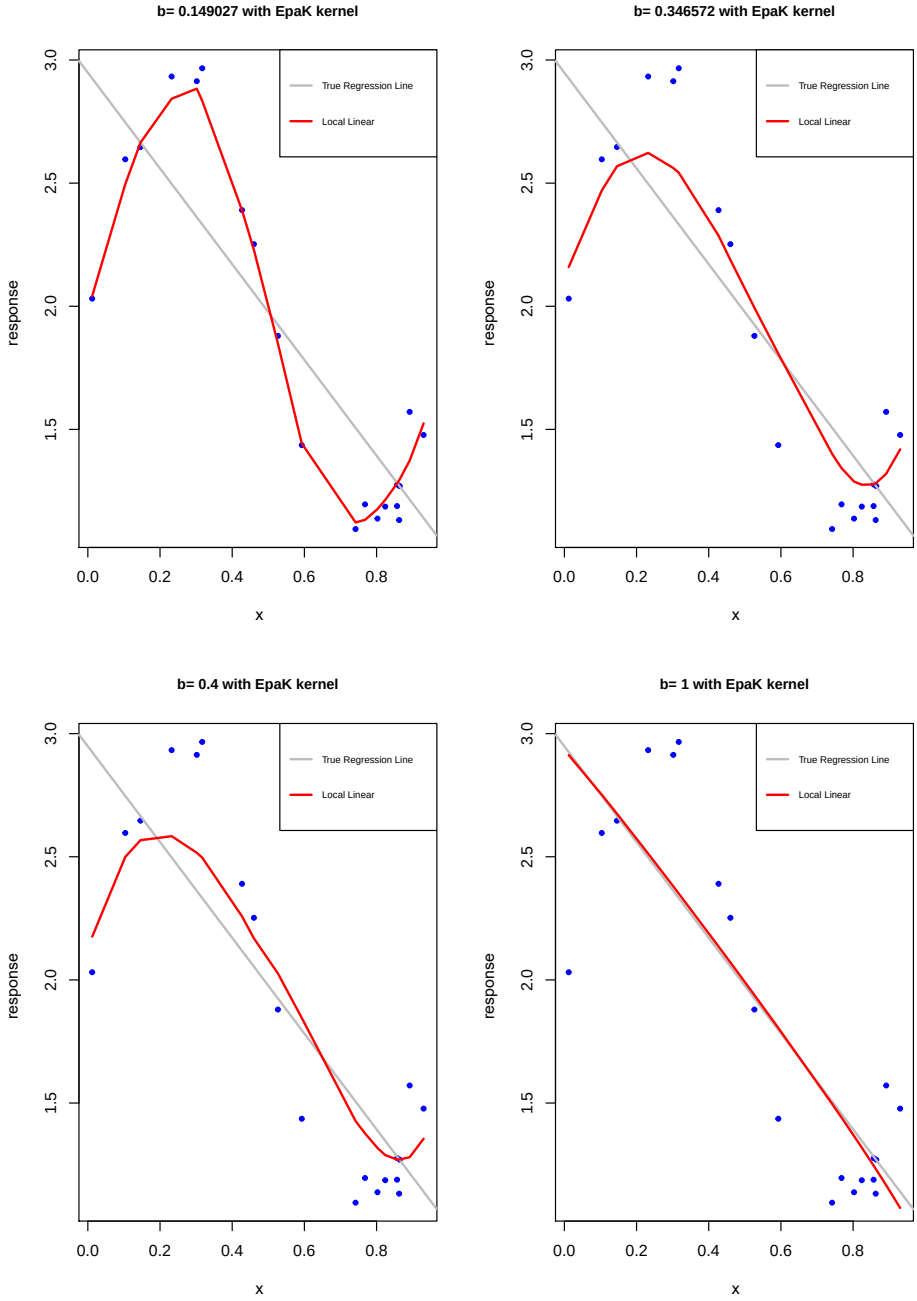

Figure 4.3: *Plots of the Simulated data (Stratum 1 of Sine population) with the true regression curve (gray line) and local linear smoother using an Epanechnikov kernel and various bandwidth values (red line)*

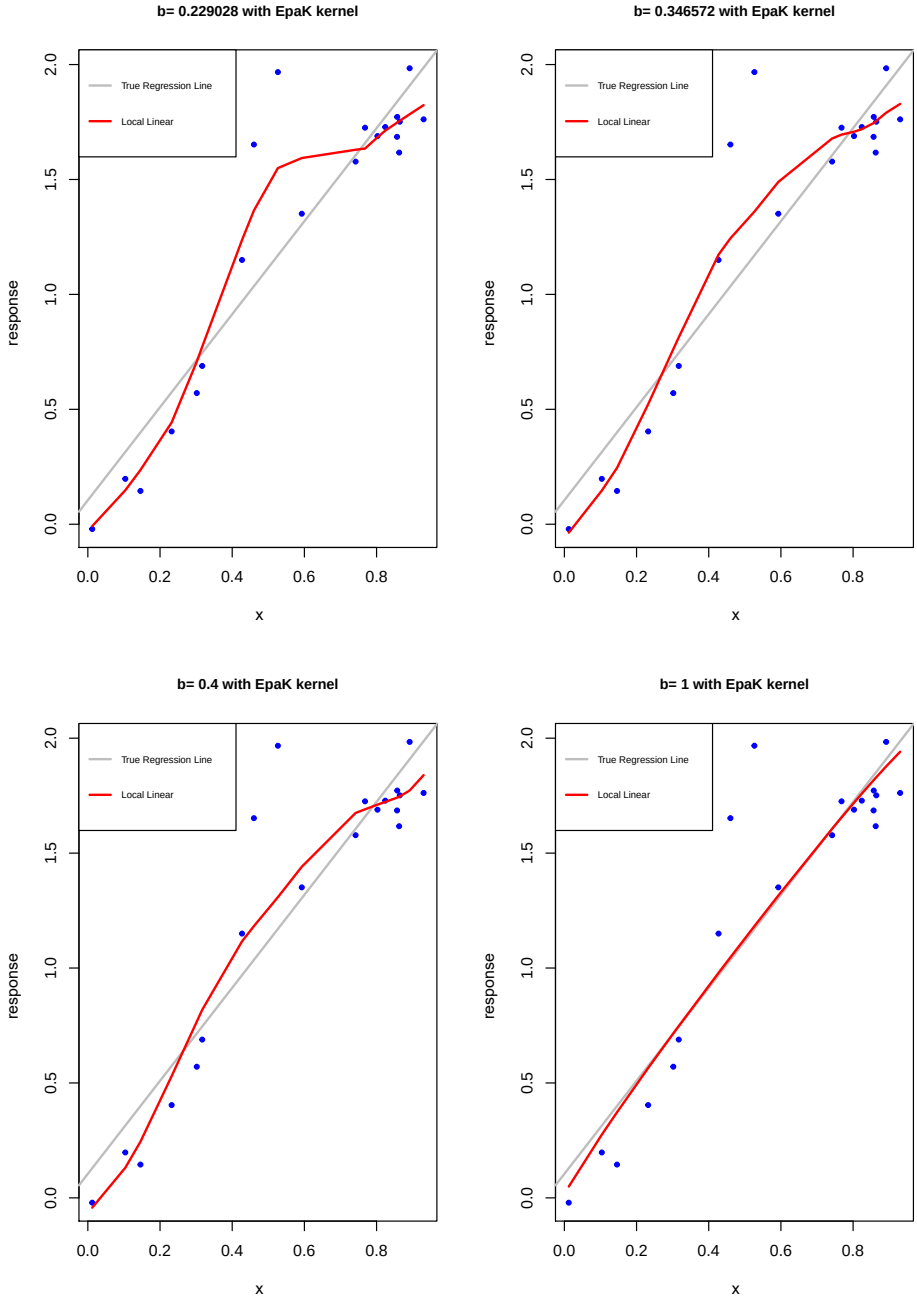

Figure 4.4: *Plots of the Simulated data (Stratum 1 of Bump population) with the true regression curve (gray line) and local linear smoother using an Epanechnikov kernel and various bandwidth values (red line)*

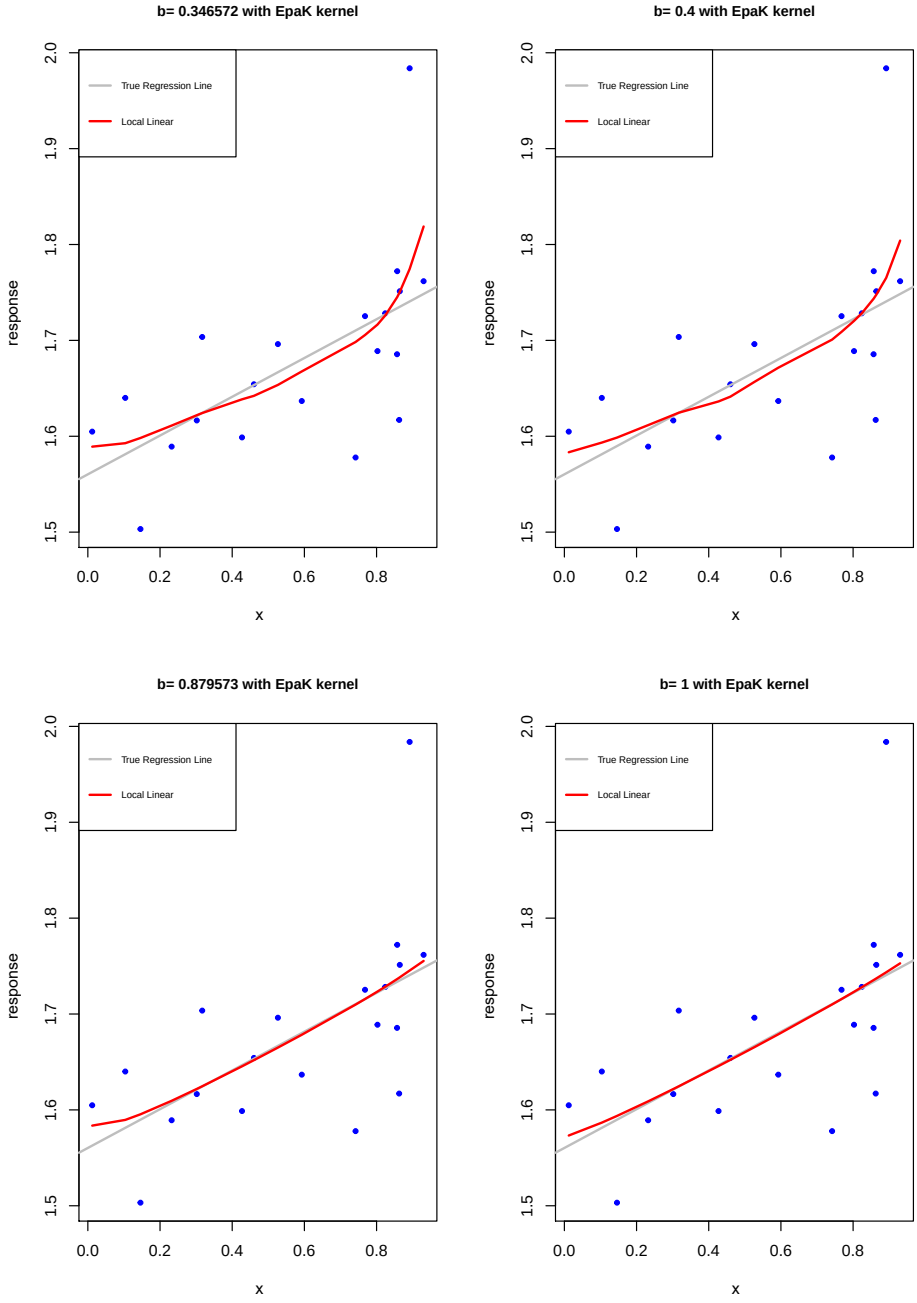

Figure 4.5: *Plots of the Simulated data (Stratum 1 of Jump population) with the true regression curve (gray line) and local linear smoother using an Epanechnikov kernel and various bandwidth values (red line)*

The estimators $\hat{Y}_{PE}$ and $\hat{Y}_{LP}$ are tested under the same bandwidth choice i.e. $b = n^{-1/5}$ (with $n = 200$), $b = 0.4$, $b = 1$ and $b = 2$. Results of this simulation are shown in Table 4.2.

Population	b	$\hat{Y}_{HT}$		$\hat{Y}_{REG}$		$\hat{Y}_{PE}$		$\hat{Y}_{LP}$	
		RAB	RE	RAB	RE	RAB	RE	RAB	RE
Linear	0.3465724	0.03212401	1	0.005778929	0.03155733	0.03321496	1.067811	0.03201888	0.9959899
	0.4	0.03212401	1	0.005778929	0.03155733	0.0335352	1.089573	0.0320533	0.9965037
	1	0.03212401	1	0.005778929	0.03155733	0.03434122	1.144951	0.03210449	0.9991698
	2	0.03212401	1	0.005778929	0.03155733	0.03272264	1.037753	0.03212023	0.9997907
Estimated Total	$b = 0.3465724$	1941.427		1943.161		1939.52		1941.248	
	$b = 0.4$	1941.427		1943.161		1938.807		1941.167	
	$b = 1$	1941.427		1943.161		1937.391		1941.419	
	$b = 2$	1941.427		1943.161		1940.336		1941.424	
Population Total		1943.052							
Sine	0.3465724	0.01855193	1	0.03836453	4.286723	0.02072086	1.243534	0.01657321	0.7990398
	0.4	0.01855193	1	0.03836453	4.286723	0.02082649	1.255919	0.01685303	0.826246
	1	0.01855193	1	0.03836453	4.286723	0.0201947	1.183826	0.01810882	0.9576443
	2	0.01855193	1	0.03836453	4.286723	0.01895357	1.043951	0.0184607	0.9908383
Estimated Total	$b = 0.3465724$	4071.066		4114.031		4080.316		4056.493	
	$b = 0.4$	4071.066		4114.031		4081.685		4054.513	
	$b = 1$	4071.066		4114.031		4079.156		4066.007	
	$b = 2$	4071.066		4114.031		4073.04		4070.166	
Population Total		4071.383							
Bump	0.3465724	0.03109618	1	0.01449569	0.2130984	0.03243536	1.085912	0.03100986	0.9935966
	0.4	0.03109618	1	0.01449569	0.2130984	0.03289121	1.116063	0.03319303	1.123072
	1	0.03109618	1	0.01449569	0.2130984	0.03357809	1.165075	0.0321397	1.061732
	2	0.03109618	1	0.01449569	0.2130984	0.03165829	1.036739	0.03106365	0.9988702
Estimated Total	$b = 0.3465724$	2186.49		2192.769		2188.266		2172.2	
	$b = 0.4$	2186.49		2192.769		2195.394		2151.329	
	$b = 1$	2186.49		2192.769		2200.689		2161.91	
	$b = 2$	2186.49		2192.769		2189.318		2182.232	
Population Total		2187.923							
Jump	0.3465724	0.004845022	1	0.02483609	26.07389	0.005616896	1.353566	0.007676967	2.274792
	0.4	0.004845022	1	0.02483609	26.07389	0.0056205	1.35023	0.007750974	2.329744
	1	0.004845022	1	0.02483609	26.07389	0.005181882	1.155266	0.005505162	1.259671
	2	0.004845022	1	0.02483609	26.07389	0.004852543	1.006773	0.004872778	1.006966
Estimated Total	$b = 0.3465724$	3299.185		3321.699		3288.857		3322.128	
	$b = 0.4$	3299.185		3321.699		3288.415		3322.202	
	$b = 1$	3299.185		3321.699		3291.326		3309.116	
	$b = 2$	3299.185		3321.699		3297.485		3300.881	
Population Total		3300.252							

Table 4.2: *Relative absolute bias (RAB) and Relative efficiency (RE) based on* 1000 *replications of simple random sampling within strata from four fixed populations of size* $N = 2000$. *Sample size is* $n = 200$. *The nonparametric estimators are computed with bandwidths* $b = 0.3465724$, $b = 0.4$, $b = 1$ *and* $b = 2$, *and Epanechnikov kernel.*

Table 4.2 shows the RAB's and RE's of the various estimators with respect to the Horvitz-Thompson estimator ($\hat{Y}_{HT}$). In most scenarios, $\hat{Y}_{LP}$ is better than the parametric estimators, but the parametric estimator- $\hat{Y}_{REG}$ performs best when the model is correctly specified. This occurs both in the linear and the bump populations, where in the former, a strong linear relationship holds between the variables while in the latter, the function is linear over most of its range despite a "bump" for a small part of the range of $x'_{hi}s$.

When the model is completely misspecified as in the Sine and Jump populations, a greater efficiency can be achieved by the nonparametric regression estimators and Horvitz-Thompson estimator. This can be seen in Table 4.2 for the Sine and Jump populations: the Horvitz-Thompson estimator ($\hat{Y}_{HT}$) and the nonparametric estimators ($\hat{Y}_{LP}$ and $\hat{Y}_{PE}$) are more efficient than their parametric opponent, $\hat{Y}_{REG}$.

When the underlying superpopulation model is completely unknown, a reasonable choice for finite population total estimation would be the nonparametric estimators such as $\hat{Y}_{LP}$ and $\hat{Y}_{PE}$ with small bandwidth choices.

In this study, $\hat{Y}_{LP}$ is sometimes seen to perform much better but not as worse as $\hat{Y}_{PE}$, and hence the proposed estimator, $\hat{Y}_{LP}$ emerges as the best performing among the nonparametric estimators being considered here. A good overall performance is observed with the proposed estimator, with smaller values of RAB and RE than the model-based competitor $\hat{Y}_{PE}$ for every population and fixed bandwidth under consideration.

Despite $\hat{Y}_{LP}$ being relatively the best estimator, its performance is significantly affected by the bandwidth choices. As the bandwidth size increases, some amount of efficiency is lost. Additionally, a keen look at the estimated totals in Table 4.2 shows that: as the bandwidth increases, the local linear regression estimator, $\hat{Y}_{LP}$ becomes equivalent to the linear regression estimator, $\hat{Y}_{REG}$. This shows that the bandwidth has an effect on the mean square error of $\hat{Y}_{LP}$. Partic-

ularly, for whichever bandwidth that is considered in this study, $\hat{Y}_{LP}$ essentially dominates $\hat{Y}_{REG}$ for all the populations except Linear and Bump populations, where $\hat{Y}_{REG}$ is competitive. Further, $\hat{Y}_{LP}$ essentially dominates $\hat{Y}_{HT}$ for all population except in the Jump population, where $\hat{Y}_{HT}$ dominates all estimators being considered. The overall performance of $\hat{Y}_{LP}$ is consistently good as long as the bandwidth remains small in this particular study.

CHAPTER 5

CONCLUSION AND RECOMMENDATION

5.1 Conclusion

The main objective of this study was to investigate the theoretical properties of a local polynomial regression estimator of the finite population total under stratified random sampling. To achieve this, a model-based estimator of the finite population total using local polynomial regression was determined in the case of stratified random sampling. The resulting estimator was found to be asymptotically model-unbiased and consistent, and therefore a useful tool in sample surveys.

Through a simulation experiment, performance of the proposed estimator was investigated against some design-based and model-based regression estimators. The RE values of the proposed estimator are in general close to one. It has been shown that for whichever bandwidth value considered, $\hat{Y}_{LP}$ essentially dominates $\hat{Y}_{REG}$ for all the populations except Linear and Bump populations, where $\hat{Y}_{REG}$ is competitive. Further, $\hat{Y}_{LP}$ essentially dominates $\hat{Y}_{HT}$ for all populations except in the Jump population, where it dominates all estimators being considered. This shows that the proposed local linear estimator, $\hat{Y}_{LP}$ can likely be an improvement over the linear regression estimator, $\hat{Y}_{REG}$ and the Horvitz-Thompson estimator, $\hat{Y}_{HT}$ when the relationship between the survey variable of interest and the auxiliary variable is non-linear. Generally, use of the proposed estimator leads to relatively smaller values of RE compared to other estimators. We conclude that nonparametric regression approach under stratified random sampling using the proposed estimator yields good results.

5.2 Recommendations and Suggestions for further study

Firstly, the bandwidths used in this study were predetermined. There is need to investigate the performance of the proposed estimator under optimal bandwidths generated from the data.

Secondly, a single auxiliary variable was used. The use of two or more auxiliary variables need to be investigated and then performance of the resulting estimator be determined against other rival estimators under stratified random sampling.

Thirdly, in the simulation study, the error variances were considered to be constant (i.e. homoscedastic). It will be interesting to investigate the performance of the estimators when the error variances are functions of the auxiliary variable (i.e. heteroscedastic).

REFERENCES

Avery, M. (2012). Literature review for Local polynomial regression.

Breidt, F. J. and Opsomer, J. D. (2000). Local polynomial regression estimators in Survey sampling. *The Annals Of Statistics*, 28:1026–1053.

Brewer, K. R. W. (1963). Ratio estimation in finite populations: some results deductible from the assumption of an underlying stochastic process. *Australian Journal of Statistics*, 5:93–105.

Cassel, C. M., Särndal, C. E., and Wretman, J. H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63(3):615–620.

Chambers, R. L., Dorfman, A. H., and Wehrly, T. E. (1993). Bias robust estimation in finite populations using nonparametric calibration. *Journal of American Statistical Association*, 88:268–277.

Cochran, W. G. (1940). The estimation of the yields of the cereal experiments by sampling for the ratio of grain to total produce. *Journal of Agricultural science*, 30:262–275.

Cochran, W. G. (1977). *Sampling techniques*. J. Wiley, New York.

Dorfman, A. H. (1992). Nonparametric regression for estimating totals in finite population. In section on Survey Research Methods. *Journal of American Statistical Association*, pages 622–625.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *The Annals Of Statistics*, 21(1):196–216.

Fan, J. and Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. *The Annals Of Statistics*, 20(4):2008–2036.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications*. Chapman and Hall, London.

Godambe, V. P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 17(2):269–278.

Holt and Smith, T. M. (1979). Post stratification. *Journal of the Royal Statistical Society, Series A*, 142:142, 33–46.

Horvitz, D. G. and Thompson, D. J. (1952). A Generalization of sampling without replacement from a finite universe. *Journal of American Statistical Association*, 47(260):663– 685.

Montanari, G. E. and Ranalli, M. G. (2003). Nonparametric methods in survey sampling. In new developments in classification and data analysis, (Eds., M. Vinci, P. Monari, S. Mignani, A. Montanari). *Springer*, pages 1–9.

Montanari, G. E. and Ranalli, M. G. (2005). Nonparametric model calibration estimation in survey sampling. *Journal Of The American Statistical Association*, 100(472):1429–1442.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Prob. and Applic.*, 9:141–142.

Neyman, J. (1934). On the two different aspects of the representative method: The method of Stratified sampling and the method of purposive selection. *J R Stat Soc Ser A*, 97:558–606.

Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, 33(201):101–116.

Ngesa, O. O., Orwa, G. O., Otieno, R. O., and Murray, H. M. (2012). Multivariate ratio estimator of the population total under Stratified random sampling. *Open Journal of Statistics*, 2:300–304.

Orwa, G. O., Otieno, R. O., and Mwita, P. N. (2010). Nonparametric mixed ratio estimator for a finite population total in Stratified sampling. *Pakistan Journal of statistics and operation research*, 4(1):21–35.

Rady, E.-H. A. and Ziedan, D. (2014a). A new technique for estimation of total using nonparametric regression under two stage sampling. *Applied Mathematical Sciences*, 8(74):3647–3659.

Rady, E.-H. A. and Ziedan, D. (2014b). Estimation of population total using local polynomial regression with Two auxiliary variables. *Journal of Statistics Applications & Probability*, 3(2):129–136.

Robinson, P. M. and Sarndal, C. E. (1983). Asymptotic properties of the generalized regression estimation in probability sampling. *The Indian Journal of Statistics, Series B*, 45:240–248.

Royall, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57(2):377–387.

Ruppert, D. and Wand, M. (1994). Multivariate locally weighted least squares regression. *The Annals Of Statistics*, 22(3):1346–1370.

Sánchez-Borrego, I. R. and Rueda, M. (2009). A predictive estimator of finite population mean using nonparametric regression. *Computational Statistics*, 24(1):1–14.

Silverman, B. (1986). *Density estimation for statistics and data analysis*. Chapman and Hall, London.

Watson, G. S. (1964). Smooth regression analysis. *Sankhya, Ser. A*, pages 359–372.

Wu, C. and Sitter, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96(453):185–193.

Zheng, H. and Little, R. J. (2003). Penalized spline model-based estimation of the finite population total from probability-proportional-to-size samples. *Journal of Official Statistics*, 19:99–117.