ROBUST ESTIMATION OF FINITE POPULATION TOTAL INCORPORATING DATA-REFLECTION TECHNIQUE IN NONPARAMETRIC REGRESSION

REUBEN CHERUIYOT LANG'AT

DOCTOR OF PHILOSOPHY

(Applied Statistics)

JOMO KENYATTA UNIVERSITY OF

AGRICULTURE AND TECHNOLOGY

2016

Robust Estimation of Finite Population Total Incorporating Data-Reflection Technique in Nonparametric Regression

Reuben Cheruiyot Lang'at

A thesis submitted in fulfillment of the degree of Doctor of Philosophy in Applied Statistics in the Jomo Kenyatta University of Agriculture and Technology

2016

DECLARATION

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

Signed:_____Date:_____

Reuben Cheruiyot Lang'at.

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

Signed: _____ Date: _____

Prof. Romanus Odhiambo Otieno.

JKUAT, Kenya

Signed: _____ Date: _____

Dr. George Otieno Orwa.

JKUAT, Kenya

DEDICATION

To my mother, Borness Kabwos, and my sons: - Kipngetich Evans and Kibet Justice.

ACKNOWLEDGEMENT

This research has been rigorous and challenging yet fulfilling and enriching academically. Doing it alone would not have been possible. There are a number of persons whose contributions enabled me to accomplish this task.

I firstly wish to thank God the Almighty for gift of life, care, knowledge and the strength throughout this period of research.

I am also very grateful to my lead supervisor Prof. Romanus Odhiambo Otieno whose support, guidance and provision of the reference materials, that I needed, is immeasurable. I thank him for his quality academic and professional guidance as well as being available for consultation.

My sincere gratitude goes to my second supervisor, Dr. George Orwa for going through my work and giving me advice and encouragement that eventually led to the improvement of this thesis.

Many thanks also go to my mother Borness Kabwos as well as my brothers and sisters, for their encouragement and persistent moral support during the entire period of research.

I appreciate my late dear wife, Lorna, for her moral support, prayers, care and understanding during the early period of this research. My daughter Mercy, whom I also lost during this period of research, was a rising star- a jewel who was a source of inspiration and one of the children who strengthened and gave a meaning to my academic struggles. May God rest their souls in eternal peace!

It is not possible to thank everyone singly, but to all whose names I have not mentioned, I am very thankful.

MAY GOD BLESS YOU ALL!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

AIDS	Acquired immunodeficiency syndrome
AMISE	Asymptotic Mean Integrated Square Error
a.s.	Converging almost surely
CD4	Cluster of differentiation 4 (immune body cells in humans)
Eff(k)	Efficiency of the kernel
HIV	Human immunodeficiency virus
i.i.d	identically and independently distributed
KDE	kernel density estimate
LSE	Least Squares estimator
MISE	Mean Integrated Square Error
MSE	Mean Square Error
NW	Nadaraya-Watson estimator
p.d.f	Probability Density function
R(K)	The roughness of the kernel
SRS	simple random sampling
Sup	supremum
var	variance
w.r.t.	with respect to

SYMBOLS

$\xrightarrow[d]{}$	approaches in distribution
$\lim_{h\to 0}$	limit as h approaches zero
$\hat{m}_{NW}(x)$	Mean function of Nadaraya-Watson estimator
$\xrightarrow[p]{}$	approaches in probability
$\hat{T}_{np}$	Nonparametric finite population total estimator that uses *NW* estimator
$\hat{T}_{npr}$	Nonparametric finite population total estimator that uses reflection technique
$\hat{T}_{HT}$	Horvitz- Thompson estimator for finite population total
$\hat{T}_{R}$	Ratio estimator for finite population total

ABSTRACT

For planning purposes, accurate information regarding population parameters of interest is essential. This information can be obtained through census or survey sampling. In sample surveys, the sampling estimation employed in a research is important since it determines the degree of accuracy. Estimation can be parametric where pre-determined parameters have been utilized or otherwise nonparametric. In nonparametric estimation, the standard kernel smoothing function has been known to suffer from the boundary bias. To overcome this, a modified kernel smoother that does not suffer significantly from this boundary effect has been proposed. This approach was found appropriate since it allows both robustness and optimality to be achieved. These properties of the proposed estimator have been investigated and the characteristics of robustness and optimality confirmed. The estimator has also been compared with the ratio estimator, the standard Nadaraya-Watson estimator as well as the design-based Horvitz-Thompson estimator using relative bias. Further to this, the Mean Square Error (MSE) as well as conditional biases were also computed to gauge the performance of the estimators. The properties of the estimators were investigated empirically and comparative analysis was made using simulated data. It is shown that the finite population total estimator whose kernel smoother was modified using reflection technique significantly addresses the bias at the boundary. This was evident from the smaller values of MSE and narrower confidence intervals, noted in the study. The relative biases also supported these findings. The study showed that the proposed estimator generally performs better than the other estimators considered.

CHAPTER ONE

INTRODUCTION

In real life situation there are many problems that involve estimation of population parameters. These parameters include population totals, means or proportions among others. This study in particular concentrated on estimation of finite population totals. Quite a number of different approaches to estimation are available in literature. To obtain an appropriate estimator one needs to evaluate the properties of such an estimator from where the desirable one can be chosen. Usually the most common desirable properties include unbiasedness, small variance and MSEs. It should be noted that a survey strategy taken may enable one to construct an estimator that posses these desirable properties. A careful implementation of such strategies often yields better results. A suitable estimation procedure that enables one to achieve this in the presence of an auxiliary variable was sought and used accordingly as is evident in the various sections of this study.

1.1 Background of the study

Suppose there is a finite population of *N* distinct and identifiable units; $U = \{1, 2, ..., N\}$. Let each population unit have the characteristic or variable of interest *Y*. It is assumed that there exist an auxiliary variable, *X*, closely associated with *Y*, which is known for the entire population.

Often researchers are faced with the problem of estimating a function of the population *Y*'s, for instance, the population total, $T = \sum_{i=1}^{N} Y_i$, the population mean $\bar{Y}$ or the population distribution function $F(y) = \frac{1}{N}\sum_{i=1}^{N} I_i(Y_i \le y)$. Some studies involving the

distribution functions may be found in (Chambers *et al,* 1992) and (Dorfman & Hall, 1993).

The focus is to estimate the finite population total, *T*. A sample *S* will be taken, so that the pair (x_i, y_i), $i= 1, 2, .., n$ is obtained for the variable *X* and its corresponding variable *Y*.

It has been assumed that *X* is a variable that is closely associated with the variable of interest *Y* and its observations are known for all elements of the population of interest. It may then be used in the design stage, estimation stage or both stages, (Hedayat & Sinha, 1991). In the presence of such an auxiliary variable a researcher can use model called *super population* model (at the estimation stage) for inference.

There are many statisticians who have used auxiliary information in their studies in the estimation stage of parametric super-population models. They include (Chambers & Danstun, 1986; Wang & Dorfman, 1996; Rao *et al*, 1990) among others.

It is known that under the parametric super-population, misspecification of the model can lead to serious errors in an inference as demonstrated in the empirical study done by (Hansen *et al*, 1983).

In the recent years efforts have been made to explore alternative ways to alleviate such a problem. It includes the use of nonparametric regression in evolving robust estimators in finite population sampling. See for example (Odhiambo & Mwalili, 2000; Dorfman, 1992; Breidt & Opsomer, 2000). Most of these researchers have used kernel smoothers. Nonparametric estimators have been found to be robust and more precise than their parametric counterparts. It is known, for instance, that a linear regression estimate will produce a large error for every sample size if the true underlying regression function is not linear and cannot be well approximated by linear functions (László *et al*, 2002).

The motivation behind the nonparametric approach in this study is that a regression curve obtained this way has four main purposes as detailed by (Härdle, 1994): - It provides a versatile method of exploring the general relationship between two variables;

secondly it enables one to make prediction of observations without any reference to a fixed parametric model; thirdly it is a tool for finding spurious observations by studying influence of isolated points and lastly it is a flexible method for interpolating between adjacent values of the auxiliary variable. The exploration of this approach is thus compelling to any researcher.

1.2 Statement of the Problem

Often researchers are faced with the problem of estimating population parameters such as population total, population mean or the population distribution function (Dorfman, 1992). In the estimation process, many researchers take advantage of existing variables (auxiliary information) that have close association with the targeted parameter in enhancing the performance of their estimators. This auxiliary information on finite population parameters is often used to increase precision of estimators of the parameters (Cochran, 1977). A popular approach to developing robust estimators is the use of kernel smoothers in nonparametric regression estimation (Racine, 2008; Irizarry & Bravo, 2010; and Breidt & Opsomer, 2009). It should however, be noted that kernel smoothers suffer from the boundary problem (Kyung-Joon & Schucany, 1998; (Karunamuni & Alberts, 2004; and Loader, 2004). Therefore, as a way of improving the nonparametric regression estimation, a robust estimator of finite population total, T, that does not suffer significantly from the boundary problem is necessary.

1.3 Objectives

1.3.1 General objective of the study

To estimate a robust finite population total incorporating data-reflection technique in nonparametric regression**.**

1.3.2 Specific objectives of the study

1. To identify an appropriate kernel for a nonparametric regression in the context of model-based estimation of finite population total.

2. To propose a nonparametric regression estimator for the finite population totals that does not suffer significantly from the boundary bias within the model-based framework.

3. To derive the asymptotic properties of bias and variance of the proposed estimator.

4. To compare the performance of the proposed estimator with the ratio, Horvitz-Thompson and the estimator due to (Dorfman, 1992) using the average relative bias, conditional bias and Mean Square Error (MSE) analyses.

1.4 Significance of the study

Usually kernel smoothers tend to perform poorly at the boundaries. This poor performance induces boundary bias in the estimator (Hastie & Loader, 1993). Researchers always endeavor to have estimators that are not biased. This study proposes a population total estimator that does not significantly suffer the drawback of this boundary bias. This is important to all research scientists and in particular the statisticians who are always keen in estimators that are of high accuracy and precision.

Nonparametric regression estimation is applicable in many other areas; this study is for instance quite useful in pathological cases such as the analysis of tissue or cells, and body fluid samples. This in turn may help to determine the area affected in the body or the amount of such fluids as blood as well as the count of relevant cells e.g. CD4 cells in HIV/AIDS patients over time. Hydrologists as well can use this approach to investigate the relationship between groundwater level fluctuations and stream-flow time series observations. The study can also benefit personnel in Agricultural sectors who may wish to estimate amount of sugar to be consumed or weight of tea exported in a year in a country as a function of element of climate among others.

1.5 Scope of the study

Kernel smoothers are commonly used in construction of nonparametric regression estimators. In this research the Nadaraya-Watson kernel estimator as used in nonparametric regression estimation was critically studied. Studies have shown that, Nadaraya-Watson estimator suffers from the boundary problem and hence requires modification. This thesis presents a nonparametric regression estimator with a simple modified kernel smoother for the population total under model-based framework. Specifically data-reflection technique has been used in modification. The asymptotic properties of the Nadaraya-Watson estimator and the proposed estimator have been investigated. Moreover the performance of this proposed estimator was assessed against the population total estimator due to (Dorfman, 1992), the ratio estimator and the design-based Horvitz-Thompson estimator, using average relative bias, MSE and conditional bias on simulated data obtained from some selected commonly used distributions as given in Table 4.2.

1.6 Limitation of the study

The framework under which the estimator proposed is used is model-based. This has been analysed and a comparison made with the conventional technique of the design-based approach. Although in literature there are many kernel functions that one can use, this study used the Gaussian kernel because of its convenience in determining the optimal bandwidth. Reflection of data technique was chosen in the modification of the kernel as a way of addressing the boundary problem induced by the standard kernel. Empirical analysis was also done using data simulated from commonly encountered distributions using Ms-Excel and R softwares as was necessary. This study limits itself to a finite population with univariate survey response variable (Y) and a univariate auxiliary variable (X), which is assumed to be available for all the population units.

1.7 Outline of the Thesis

This thesis is organized into five chapters as follows:-

The background and identification of the problem, significance and the scope of study have been done in this chapter. The notation, model used and a review of the related literature are presented in chapter two. The proposed estimator and the properties of kernel regression estimation have been derived in chapter three. The fourth chapter gives the empirical study of the estimator in simulated data. The results and discussion are given in chapter five as well as highlights of areas suggested for further research.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

In this chapter literature review has been done in areas relevant to survey sampling. Various survey strategies as well as the nonparametric regression estimation and kernel density estimation have been discussed. This chapter reviews the literature of the nonparametric regression estimation and the common survey strategies used in various statistical researches.

2.2 Survey strategies and nonparametric regression estimation

It is now well known to majority of researchers that sample surveys play a very important role in giving the much needed information in studies. Needless to say is the fact that in some occasions, its alternative (the census) technique may not be realistically practicable. To illustrate one advantage, for instance, if a doctor wanted to diagnose some disease that involves testing of blood, obviously the option to go for would be the blood sample and not all the blood from an individual as would be the case in census. A researcher using this method has four estimation approaches that can be used in statistical investigations involving survey sampling. These are the design-based approach, model-based approach, model-assisted approach and randomization-assisted approach.

For the first three approaches, their merits and demerits have been discussed in (Chambers, 2011). (Brakel & Bethlehem, 2008) have also made some literature review on these approaches. The fourth approach has been presented by (Kott, 2005). In what follows a brief account of each of them has been given.

2.2.1 Design-based approach

In this approach the values of a variable of interest of the target population are viewed as fixed quantities (constants).This implies that selection probabilities introduced with the design are used in determining the properties of estimators used to obtain expected values, variances, biases and so on. This approach is also known as classical or randomization theory. The randomization theory in a way provides a kind of nonparametric approach to inference in that no distributional assumptions are made about a random variable (Lohr, 2010).

For instance, as done by (Cornfield, 1944), if

$$I_i = \begin{cases} 1, \text{if unit } i \text{ is in the sample} \\ 0, \text{elsewhere} \end{cases}$$

Then

$$\pi_i = P(I_i = 1) = P(\text{select unit } i \text{ in sample}) = \frac{n}{N} \tag{2.1}$$

where π_i is the inclusion probabilities and I_i is the indicator variable

Thus

$$\hat{T}_{HT} = \sum_{i \in s} \pi_i^{-1} y_i \tag{2.2}$$

is the design-based expansion estimator (Horvitz & Thompson, 1952).

Statisticians, who have relied on design-based methods, like it for the capability of not only eliminating personal biases in selecting the sample but also, its use in situations where little may be known about the population.

It should, however, be noted that besides the above advantage obtaining an optimal strategy under this approach might be an impossible task where no restriction on sample size is made, a result first noted by (Godambe, 1955).

2.2.2 Model-based Approach

In this approach, the distribution is a structure innate to the population itself and is unknown but capable of being modeled. Model-based predictive inference has been discussed in detail by (Valliant *et al*, 2000). In this prediction approach as is also called, the expectations are over all possible realizations of a stochastic model (usually a linear regression model) which links a variable of interest, *Y*, with a set of auxiliary variables, *X*, (Cox, 1995). Values of interest are thought of as random variables $Y_1, Y_2, \ldots, Y_N$ generated by some model. The actual values for finite population $y_1, y_2, \ldots, y_N$ are one realization of the random variables. In the presence of the auxiliary in formation this approach supplies the link between units in the sample and those not in the sample.

The information obtained from the sample ($y_i,\ i \in s$) is thus used to predict the information of the non-sample ($y_i,\ i \notin s$).

Therefore under an SRS method one may assume that *y* is linear in *x*, hence a simple linear regression model of the form:

$$y_i = \alpha + \beta x_i + e_i \qquad \text{for } i = 1, 2, \ldots, N \tag{2.3}$$

With the assumptions that e_i is *i.i.d* with mean, 0 and variance, σ^2, then an appropriate model-based estimator is of the form:

$$\hat{T}_{lin} = \sum_{i=1}^{N} Y_i = \sum_{i \in s} Y_i + \sum_{i \notin s} \left(\hat{\alpha} + \hat{\beta} x_i \right) \tag{2.4}$$

where $\hat{\alpha}$ and $\hat{\beta}$ are the Ordinary Least Square Estimates for the parameters α and β.

Another alternative estimator that can be used under this approach is the ratio estimator. The estimator of finite population total under SRS may be given by:

$$\hat{T}_R = \hat{B}\sum_{i=1}^{N} X_i \qquad (2.5)$$

Where $\hat{B} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}$, is the ratio estimator for its equivalent population parameter, $\sum_{i=1}^{n} y_i$ is the sample total of the study variable while $\sum_{i=1}^{n} x_i$ is the equivalent for the auxiliary variable assumed to be known for the entire population. The entire population total of this auxiliary variable is given by $\sum_{i=1}^{N} X_i$. It is known that the ratio estimator is the Best Linear Unbiased Predictor (BLUP) (Cochran, 1977; Cox, 1995; and Brewer, 2002).

It should be noted that in the parametric model such as that given in (2.3) and (2.4), the choice of a model and its robustness to misspecification, is a major concern. This problem that may arise due misspecification has been tackled using varying techniques but the most predominant of all is the nonparametric method. Some of the recent studies that employ this technique include (Dorfman, 1992; Chambers *et al*, 1993; Chandran & Prajneshu, 2004; and Breidt & Opsomer, 2009) among others.

2.2.3 Model -Assisted Approach

This is an approach that still depends exclusively on randomization-based inference and estimation but optimizes them under the explicit assumption that the finite population under study is itself a sample drawn from a super population generated by a specific stochastic model.

Basically inferences are design-based while the model serves as a means to help choose between the randomization-based methods (Langat *et al*, 2007). It has been noted that

inadequacies in the super population models adopted, results in unacceptable biases (Brewer, 2002). Though researchers have suggested that selection of balanced samples could meet this difficulty, random sampling designs could make up for most of the difference by regressing on whatever variable the balanced scheme were balanced on. (Brewer, 1995) remarks that using the model assisted approach is like wearing both belt and braces to hold one's trousers. If the belt (model) should break, then one is not going to be totally embarrassed, since the braces (design-unbiasedness) should still keep things in place. For a comprehensive coverage of this approach one can see (Särndal *et al*, 1992) and (Chaudhuri & Stenger, 2005). A model assisted estimator for the population total is thus given by:

$$\hat{T} = \sum_{i \in U} \hat{y}_i + \sum_{i \in s} (y_i - \hat{y}_i)\pi_i^{-1} \tag{2.6}$$

2.2.4 Randomization – assisted model –based approach

The proponents of this approach argue that it makes little sense to average over all samples; that were not drawn as is suppose to be the case in the conventional design based approach. In this approach, therefore, inferences are made on the basis of the actual sample drawn and observed i.e the goal of the survey sampling here is that of model based. They only employ the design based method which focuses on the set of hypothetical samples that could have been drawn, to simply protect against model failure (Kott, 2005). Thus unbiasedness is the key issue.

2.3 Model-based estimation and regression techniques

The four approaches highlighted above basically stem from two broad strategies - the traditional design-based approach on the one hand which has its conceptual origin in the paper by (Neyman, 1934) and the sampling theory texts such as (Kish, 1965) and (Cochran, 1977), where inferences are based on the probability distribution induced by the sampling design with the population values being held constant. On the other hand is the other sampling strategy, model-based approach which is strongly linked to Royall

and his students, where inferences are model dependent. (Royall, 1970) gives a summary of the philosophy behind this approach. It should be noted that the nonparametric nature of the design-based Approach can easily make it an obvious methodology to robust inferences; however, there are no relevant optimality criteria that can be checked under this approach, (Chambers, 2011). Therefore, if one wants both optimality and robustness, the option is Model-based approach.

Further to this and in the presence of an auxiliary variable, it is well known that both model-based and model-assisted approaches perform better than the purely design-based approach provided that the assumed model that links study variable and auxiliary variables is appropriate, (Prasad & Subhash, 2011). This gives the basis of choosing the model-based approach in this study. Within the model-based framework, given the auxiliary variable that can inform one about the variable of interest, one can estimate the parameter required using a regression model. This can either be parametric or nonparametric.

Parametric models involve making assumptions about the underlying distribution. For instance, one may assume that the underlying distribution is normally distributed with the mean, μ and variance, σ^2. These two parameters are all that one needs to say everything about the distribution. If this assumption of normality is correct, the two numbers (mean and the variance) will be adequate as far as estimation of the parameter of interest is concerned. However, if such an assumption of the model is wrong then the resulting estimates can be seriously misleading, (DiNardo & Tobias, 2001). Fortunately on this, a promising alternative technique- the nonparametric approach that does not require specification of the underlying function is available. In this approach the data determines the functional form of the distribution required. This flexibility has made the approach popular among the statistical researchers. Nonparametric regression technique within the model-based approach often involves the use of a kernel regression estimator. Within this framework the techniques that have impressed many researchers include; the local polynomial regression estimator, spline regression and orthogonal series.

2.4 Kernel functions in nonparametric regression estimation

There are many kernel functions in literature that a researcher can use in nonparametric regression estimation. A kernel is simply a smoothing function or weight-assigning function. Different researchers make different assumptions about the functions but most are common. The common assumptions include those that require it to be symmetric and unimodal. The existence of some moments is also another common assumption though (Fan & Gijbels, 1992) require the existence of all moments. Though not universal, some researchers assume a bounded support for the kernel and that kernel is smooth, (Avery, 2010). The functions commonly used are the Gaussian function and those kernels normally derived from the Beta function with the parameter changing from 0, 1, 2, and 3 which respectively yields the uniform, Epanechnikov, biweight, and triweight kernels. Another kernel, is the triangular kernel rarely used because it lacks smoothness property (Wand & Jones, 1995).

Within the kernel window observations may receive the same weights as in histograms or weights that reduce gradually as one moves away from the target observation where the kernel is centred. The kernel used in histogram is typical of the uniform or the rectangular kernel- so called because it treats the points in a bin the same, in fact such a particular choice of a kernel is termed "naive" since weight of $\frac{1}{2}$ is assigned to all points regardless of how far or close the point is from the central point of the bin or window of the kernel (DiNardo & Tobias, 2001).

A variety of kernel functions are possible in general, but both practical and theoretical considerations limit the choice (Härdle, 1994). It is because of the thinking that points that are closer to the centre of the window of a chosen kernel, have closer association (can give more details or contribution) about the target observation. These points undoubtedly deserve to be given more weight than the ones further away (Irizarry & Bravo, 2010). This obviously cannot be done using rectangular or uniform kernels, but those that assign reducing weights further away from the centre, see Table 3.1 for the

common kernel functions. It would also be desirable to use a kernel that optimizes the error criterion measurement such as AMSE. If this is the goal of the researcher then Epanechnikov would be the right choice.

It is worth noting that the difference between this kernel and the others discussed and tabulated is not significant. In fact a slight increase in the sample size brings the corresponding efficiency at par with the optimal kernel. A disadvantage noted with Epanechnikov is that it has a discontinuous first derivative which may be undesirable (Wand & Jones, 1995). This results in the choice of Gaussian function instead and thus the reasons for its preference in this study. In addition, it can also be noted that this function has an optimal bandwidth choice in the event that the underlying distribution is normal.

Researchers have found out that even with such an advantage the choice of the kernel function is not as important as that of the bandwidth itself (Faraway, 2006). It has been noted in literature that if one misses the optimal bandwidth that minimizes MSE (or other measure of accuracy) by ten percent, there is more drastic effect on the smoothing than if one selected one of the "suboptimal" kernels (Härdle, 1994).

2.5 Nonparametric regression estimation of finite population total

Nonparametric regression estimation has been carried out by many researchers in many studies. (Dorfman, 1992) did a comparison between the population total estimators constructed from the famous design-based Horvitz-Thompson estimator and the Nadaraya-Watson estimator- the nonparametric regression estimator where his findings show that the nonparametric regression estimator better reflects the structure of the data and hence yields greater efficiency. This regression estimator, however, suffered the so called boundary bias besides facing bandwidth selection challenges. (Breidt & Opsomer, 2000) did a similar study on nonparametric regression estimation of finite population total under two-stage sampling. Their study also reveals that the nonparametric regression with the application of local polynomial regression technique dominated the

Horvitz-Thompson estimator and improved greatly the Nadaraya-Watson estimator. (Breidt & Opsomer, 2009) carried out estimation of population of finite population total under two-stage sampling procedure and their results also show that the nonparametric regression estimation performs better compared to the standard parametric estimators when the model regression function is incorrectly specified, while being nearly as efficient when the parametric specification is correct. In particular the local polynomial regression estimator was applied in their research.

2.6 Asymptotic properties of estimators based on kernels

The key properties that a statistician would be interested to check given an estimator, are the variance and the bias. These two can enable one to measure the amount of accuracy and precision that an estimator has. In fact at an arbitrary fixed point, a basic measure of accuracy that takes into account both the bias and variance is the Mean Square Error (MSE) (Tsybakov, 2009). Other texts that have such literature include (Härdle, 1994; Takezawa, 2006; and Härdle *et al*, 2005). This is one of the criteria of error measurement that can be used in such statistical researches. In nonparametric regression estimation one may be interested in the cumulative amount of bias and the variance over the entire regression line. This global measure called MISE is obtained by finding the integral value of the variance and the square of the bias of the estimator (Zucchini, 2003). One can use Taylors' expansion to obtain the Asymptotic Mean Integrated Square Error (AMISE). It is from this that an optimal bandwidth can be obtained. (Manzoor *et al*, 2013) carried out similar study using these measures. Given the asymptotic properties one can discuss the speed of convergence of the estimators and determine the price to pay in a given option. It is from this vast literature that this study uses these measures in the analysis stage to compare the proposed estimator against the standard ones reported in the next section.

2.7 Performance of estimators under model-based approach

In assessing the performance of the estimators, measures that allow for comparison of the estimators are normally subjected to simulated data. (Dorfman, 1992) used the root average relative biases to make comparison between the design-based Horvitz-Thompson estimator with the model-based nonparametric regression estimator derived using the smoother first proposed by (Nadaraya, 1964) and (Watson, 1964). This study used the average relative bias to make comparison between the same design-based estimator of Horvitz-Thompson, the model-based nonparametric estimator for finite population total due to (Dorfman, 1992) and the proposed population total estimator developed in the study using the data-reflected technique. Other measures used also include the *MSE*s, the conditional biases and the construction of the confidence lengths.

2.8 Bandwidth selection

2.8.1 Introduction

Bandwidth selection is another area of concern in kernel estimation. These bandwidths vary with the kernel function chosen. An optimal bandwidth of one kernel function cannot be regarded in the same way for another function. Because of this, many researchers have been carrying out studies aimed at determining techniques of obtaining bandwidths that minimize MSE or AMSE functions that can be used with the different kernel functions preferred.

Two common ways in which this problem can be tackled have been highlighted- the "plug-in" method and the cross validation method. See section 2.8.3 for the cross-validation technique. The plug-in method simply involves the replacement of the unknown functions in the expression of interest. This is discussed in the next section.

2.8.2 Direct plug-in methods

Bandwidth selection that uses this technique has been recommended by (Wand & Jones, 1995) as well as (Fan & Gijbels, 1992). To implement the plug-in method one requires

the minimizer of *AMSE* or more usually the *AMISE*. The challenge is in the estimation of the unknown terms -the variance, σ^2, $m''(x)^2$ and $m(x)$ itself. Research has shown that the best estimate of $m(x)$ depends on higher order derivative which implies infinite regression. In other words obtaining the best estimate is impossible. As an alternative in practice, a quick simple bandwidth is used to obtain a higher order derivative which is then used in the procedure of determining the estimates. The details have been discussed in (Wand & Jones, 1995). There are a number of plug-in bandwidths estimators such as the simple rule of thumb; direct plug-in and solve-the-equation, one can see the details in (Ruppert *et al*, 1995). From the literature review done by (Avery, 2010), all the three perform well in simulation, though the rule-of-thumb tend to "under-smooth."

2.8.3 Cross validation technique

This technique of bandwidth selection is occasionally referred to as classical or conventional because it has been in use much earlier than the relatively more recent "plug-in" methods discussed in the section 2.8.2. Although the arguments fronted by the proponents of the above criteria is the tendency of the cross validation technique to "under-smooth," (Loader, 1999) still recommends this technique and argues that the noise is due to the actual trend of data and how difficult it is to have data-based bandwidth selection. In the comprehensive bandwidth analysis done, the research reveal various weaknesses in the "plug-in" methods leaving the classical methods standing better chances of being more informative in their results, if used properly. (Li & Racine, 2004) also proposed this data-driven bandwidth selection. It is for this reason that this study adopts this bandwidth selection technique. Since bandwidth selection was not really the goal of this study, a brief approach of getting this selector is given next.

Given the model in (3.3), to obtain the estimate $\hat{m}(x)$ requires Y_i. This scenario presents a problem in that Y_i is required to predict itself. By employing the leave-one-out estimator, $\hat{m}_{NW-i}(x)$, cross-validation solves this problem by replacing $\hat{m}(x)$ with $\hat{m}_{-i}(x)$.

Thus the leave-one-out estimator for the Nadaraya-Watson estimator is given by:

$$\hat{m}_{NW-i}(x) = \sum_i \frac{K\left(\frac{X_i - x}{h}\right) Y_i}{\sum_i K\left(\frac{X_i - x}{h}\right)} \tag{2.7}$$

This implies that at point x, the i^{th} observation is left out when estimating $\hat{m}(\bullet)$. This leads to the following minimization problem of the cross-validation function.

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \hat{m}_{NW-i}(X_i)\right]^2 w(X_i) \tag{2.8}$$

where $w(X_i)$ is a weighting function that may be used to assign less weight in regions of sparse data and at tail to counter large variance and boundary effect. Similarly for the reflection technique we replace $\hat{m}_{NW-i}(x)$ with $\hat{m}_{ref-i}(x)$. This automatic bandwidth selection was adopted and used in the study. In the next chapter the kernel regression estimator used to construct the finite population total estimator due to (Dorfman, 1992) has been derived. The kernel estimator was also modified using reflection, a technique which has been discussed in the chapter. It was then used in constructing the proposed estimator.

CHAPTER THREE

METHODOLOGY

3.1 Introduction

In this study, the general objective was to robustly estimate finite population total incorporating data-reflection technique in nonparametric regression models**.** This chapter focuses on the development of the nonparametric regression estimator of finite population total that is robust within the model-based framework. In proposing such an estimator, cognizance is given to the desired qualities of being more reliable as well as being flexible in terms of capturing well the structure of the data under consideration. As a result, data-driven estimators and concepts from nonparametric estimation have been adopted.

Accordingly, a nonparametric estimator was obtained with the use of the kernel function. As is always in other applications, usually the kernel function should be continuous, bounded symmetric real function which integrates to one. This kind of a function has been considered in nonparametric regression estimation constructed within the framework of the model-based inference. The standard Nadaraya-Watson regression estimator is known to suffer from the boundary bias. In the proposed estimator, modification has been made in an attempt to develop a population estimator that reduces this boundary bias significantly.

Asymptotic properties of the standard and the proposed estimator have been studied. In particular, the variance and bias of the estimator of the finite population total constructed using the Nadaraya-Watson estimator and the proposed estimator which uses data reflected technique were derived. It is worth noting that order algebra have been employed in studying the said properties. This is given in the next subsequent sections.

3.2 Notations and the framework used

Let ***U*** be a population of *N* units. That is, $U = \{1, 2, ..., N\}$ where $N < \infty$. Let each population unit have the characteristic of interest *Y*. Suppose that X is an auxiliary variable associated with the study variable Y and which is accordingly assumed to be known for all the population units. Suppose that the intention is to obtain the finite population total:

$$T = Y_1 + Y_2 + ... + Y_N = \sum_{i=1}^{N} Y_i \tag{3.1}$$

Since all the units sampled shall be observed, it is assumed that such units will be known without error, and the task therefore is to estimate the non-sampled units. In this paradigm, the population total can therefore be obtained using an equation of the form:

$$T = \sum_{i \in s} y_i + \sum_{i \notin s} y_i \tag{3.2}$$

where $\sum_{i \in s} y_i$ refers to the total of the sampled units whereas $\sum_{i \notin s} y_i$ refers to the total non-sampled units of the population. The non-sample part can be estimated using the regression model in (2.3), which is a parametric equation since the parameters α (the constant term) and β (the regression coefficient) have to be estimated using the LSE technique.

With the assumptions on the regression model (2.3) correctly stated about the underlying distribution, the parametric technique would be good and easy to compute. This requirement is however not easy to achieve in real life data. Notably, the structural nature of the parametric models greatly reduces their flexibility and therefore under such models, estimators obtained are without loss of generality, not robust. Even in cases where regression is done the context of parametric super-population models, a problem of model failure may arise due to model misspecification.

To overcome this, many researchers have employed nonparametric regression methods which have over time, become very popular. This is because they adapt to the structure innate to the data sampled for the study hence becoming flexible and robust. This guards against misassumptions in the model which could lead to incorrect conclusion regarding the estimate. The nonparametric models also provide information, independent of the researcher's thoughts or belief about the underlying distribution that relates the study variable and the auxiliary variable. Moreover, in these models, the only assumption made about the observations is that they are independent and identically distributed from an arbitrary continuous distribution.

But even with these advantages, the computational intensity that the nonparametric regression demands could have slowed down the progress of many researchers on its use in the earlier years when the technology of computers was still at the infancy stage. Fortunately it is no longer the case currently.

A model- based nonparametric model is conventionally of the form

$$Y_i = m(X_i) + e_i \qquad i=1, 2, \ldots, n \tag{3.3}$$

where Y_i- is the variable of interest

X_i-is the auxiliary variable

m-is an unknown function to be determined using sample data

e_i-is error term-assumed to be $N(0, \sigma^2)$ under the model in (3.3).

Nonparametric regression involves the use of kernel functions discussed in the next section.

3.3 Kernel functions commonly used in nonparametric regression estimation

Choosing an appropriate kernel and a suitable bandwidth is quite important in nonparametric regression. It is known that the two (i.e. kernel function and the

bandwidth), do not have the same effects–in terms of their contribution in the estimate. Previous studies reveal that compared to the kernel function which has the least impact, bandwidth selection plays a more crucial part in obtaining good estimates (Wand & Jones, 1995).

Some of the kernel functions and their efficiencies are given in the Table 3.1.

Table 3.1: Common Kernel Functions

Kernel	**Equation**	**R(K)**	**K_2(K)**	**Eff(K)**
Uniform	$K(z) = \frac{1}{2} I[\mid z \mid \leq 1]$	$\frac{1}{2}$	$\frac{1}{3}$	0.9295
Epanechnikov	$K(z) = \frac{3}{4}(1-z^2) I[\mid z \mid \leq 1]$	$\frac{3}{5}$	$\frac{1}{5}$	1.0000
Biweight	$K(z) = \frac{15}{16}(1-z^2)^2 I[\mid z \mid \leq 1]$	$\frac{5}{7}$	$\frac{1}{7}$	0.9939
Triweight	$K(z) = \frac{35}{32}(1-z^2)^3 I[\mid z \mid \leq 1]$	$\frac{350}{429}$	$\frac{1}{9}$	0.9867
Gaussian	$K(z) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2} z^2)$	$\frac{1}{2\sqrt{\pi}}$	1	0.9512
Triangular	$K(z) = (1-\mid z \mid) I[\mid z \mid \leq 1]$	$\frac{2}{3}$	$\frac{1}{6}$	0.9859

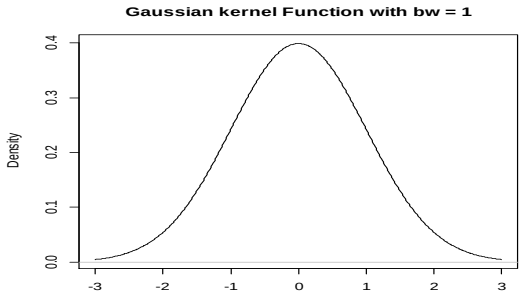
Source: (Jann, 2007).

It now well known to many statisticians in this field that the kernel that optimizes the MSE (or AMSE) - the measure of accuracy is the Epanechnikov kernel (Wand & Jones, 1995). This fact has been shown by (Epanechnikov, 1969). In Table 3.1 it is important to note that Eff(K) which represents the efficiency of the kernel has been given relative to Epanechnikov kernel -the minimizer of AMSE and $R(K) = \int_{-\infty}^{\infty} K(z)^2 dz$ is the roughness of the kernel. $K_2(K)$ is the second moment of the kernel- actually the spread ("variance") of the kernel density. From this information one can see that there are no much differences in these efficiencies, an indication that kernel selection has rather limited impact on them.

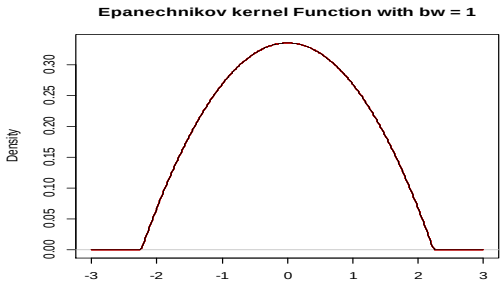
It is clear in the table that the uniform (also called the rectangular kernel) has an approximate efficiency value of 93%. The interpretation for this is that for n = 93, one can obtain an optimal AMSE with Epanechnikov kernel, while n = 100 would be required to obtain approximately the same value using the uniform kernel. Since both eventually lead to roughly the same estimate, it implies that emphasis should not be so much on choice of a kernel function. This is the reason why in many researches the choice of kernel is based on other considerations like the desired smoothness (Alberts & Karunamuni, 2007; and Zucchini, 2003).

Frequently researchers may not opt for the uniform kernel function because it assigns constant weights across the observations in its window. In particular weights of ½ are assigned to points within the distance of h (the bandwidth) away from x- the point at the centre of the window. Points farther away are all assigned zero weights because the indicator function, $I[|z| \leq 1]$, is by definition equal to 0 for all values of the scaled distance, $z = (x - X_i)/h$, that are bigger than 1.

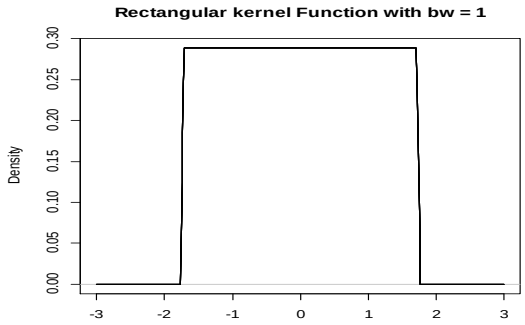
The other functions, that is, the Epanechnikov, bi-weight, tri-weight, Gaussian and triangular kernels assign relatively smaller weights progressively away from the point at the centre. They assign more weight to the points closer to the centre. It should be noted that for all the kernels shown in the Table 3.1 their indicator functions, $I[|z| \leq 1]$, are by definition also equal to 0 for all values of the scaled distance, $z = (x - X_i)/h$, that are bigger than 1, except for Gaussian which is unbounded. The individual graphs of the functions were obtained from R-software and presented in Fig. 3.1 parts (a)-(f).



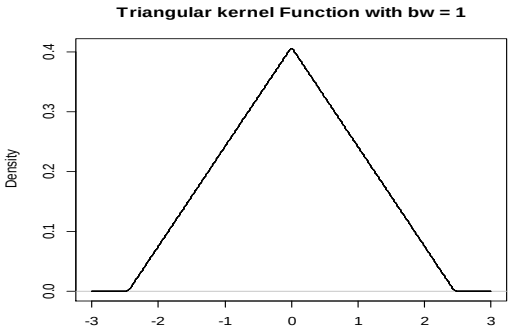
(a)



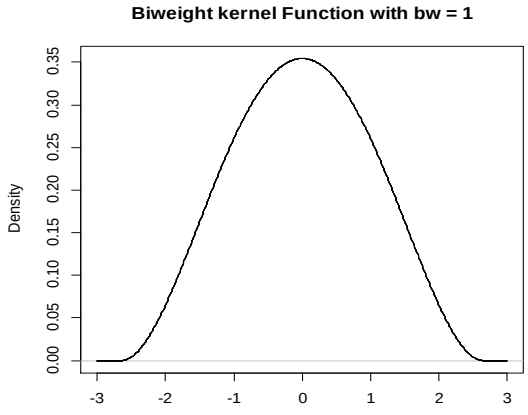
(b)



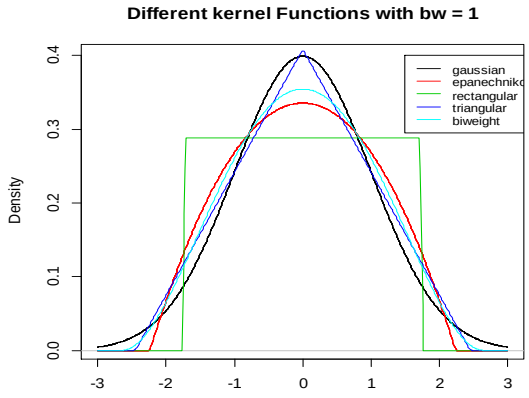
(c)



(d)



(e)



(f)

Figure 3.1: Graphs of some selected kernel functions

In Fig. 3.1 Graph (f), shows all the five selected graphs combined. Notice the shapes of each of the kernel function and how it will have an effect on the assignment of the respective weights.

3.4 Review of the Nadaraya-Watson Estimator

The idea of nonparametric regression has gained prominence in a couple of decades now. It was mentioned in the section preceding this that previous applications of the approach by the researchers were hampered by the tedious computational challenges. The recent advancement in technology and computers has enabled researchers to handle the massive computation experienced with this approach. This section gives a brief derivation of Nadaraya- Watson estimator.

Let $K(.)$ denote a kernel function which is also twice continuously differentiable, such that:

$$\text{(a) } \int K(z)dz = 1 \quad \text{(b) } \int zK(z)dz = 0 \text{ (c) } \int z^2K(z)dz := K_2(K) \tag{3.4}$$

Further, let the smoothing weight be:

$$w_i(x) = \frac{K\left(\frac{x - X_i}{h}\right)}{\sum_s K\left(\frac{x - X_i}{h}\right)}, \quad i = 1,2,...,n \tag{3.5}$$

A form of the kernel weight defined as in (3.5) was proposed by (Nadaraya, 1964) and (Watson, 1964). Since then many researchers have explored the nonparametric regression technique in estimation. Some other references include (Härdle, 1990; Takezawa, 2006; Gámiz, *et al*, 2011; and Tsybakov, 2009) among others. Herein, a simple Nadaraya-Watson Kernel estimator of $m(x)$ has been considered.

Assume a model of the form specified in (3.3), where $\sum_s (.)$ is the summation over all the sampled units and h is the bandwidth also referred to as the smoothing or tuning parameter, with $\sum_{i=1}^{n} w_i(x) = 1$. The Nadaraya-Watson estimator of $m(x)$ is therefore given by:

$$\hat{m}_{NW}(x) = \sum_{i=1}^{n} w_i(x) Y_i = \sum_{i=1}^{n} \frac{K\left(\frac{x - X_i}{h}\right) Y_i}{\sum_s K\left(\frac{x - X_i}{h}\right)} \tag{3.6}$$

Given the model in (3.3) and the conditions of the error term, the expression for the variable Y relative to variable X can be obtained as a joint *p.d.f* of $f(x, y)$ as follows:

$$\begin{aligned} m(x) &= E[Y \mid X = x] \\ &= \int y f(y \mid x) \partial y \\ &= \frac{\int y f(x, y) \partial y}{\int f(x, y) \partial y} \end{aligned} \tag{3.7}$$

The estimator for $f(x, y)$ is given by:

$$\hat{f}(x, y) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{h} K\left(\frac{x - X_i}{h}\right) \frac{1}{h} K\left(\frac{y - Y_i}{h}\right) \right) \tag{3.8}$$

Thus the numerator is obtained as:

$$\int y \hat{f}(x, y) \partial y = \frac{1}{n} \sum_{i=1}^{n} \left(\int \frac{1}{h} K\left(\frac{x - X_i}{h}\right) \frac{1}{h} y K\left(\frac{y - Y_i}{h}\right) \partial y \right) \tag{3.9}$$

On letting $z = \frac{y - Y_i}{h} \Rightarrow \partial y = h \partial z$ and $y = hz + Y_i$

Therefore

$$\int y\hat{f}(x,y)\partial y = \frac{1}{n}\sum_{i=1}^{n}\left(\int \frac{1}{h}K\left(\frac{x-X_i}{h}\right)\frac{1}{h}(hz+Y_i)K(z)h\partial z\right)$$

$$= \frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)\left[\int \frac{1}{h}hzK(z)h\partial z + \int \frac{1}{h}Y_iK(z)h\partial z\right] \tag{3.10}$$

and from (3.4), the following can be obtained;

$$\int y\hat{f}(x,y)\partial y = \frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)[0+Y_i]$$

$$= \frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)Y_i \tag{3.11}$$

The denominator can similarly be derived as follows:

$$\int \hat{f}(x,y)\partial y = \frac{1}{n}\sum_{i=1}^{n}\int\left(\frac{1}{h}K\left(\frac{x-X_i}{h}\right)\frac{1}{h}K\left(\frac{y-Y_i}{h}\right)\right)\partial y$$

$$= \frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)\int \frac{1}{h}K\left(\frac{y-Y_i}{h}\right)\partial y$$

$$= \frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)\int \frac{1}{h}K(z)h\partial z$$

$$= \frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right) \tag{3.12}$$

This is so since the integral part *w.r.t. z* equals one.

Putting (3.11) and (3.12) together gives the Nadaraya-Watson estimator given in (3.6). The estimator is a linear smoother since it is a linear function of the observations, Y_i. Given the sample and the Gaussian kernel function defined in Table 3.1, then at any

point x the corresponding y-estimate is given by the estimator in (3.6). This can be written as:

$$\hat{y} = \hat{m}_{NW}(x) = \sum_{i=1}^{n} w_i(x) Y_i = \sum_{i=1}^{n} \frac{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{x - X_i}{h} \right)^2 \right) Y_i}{\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{x - X_i}{h} \right)^2 \right)} \tag{3.13}$$

This gives a way of estimating the non-sample values of y given the auxiliary value of x.

The nonparametric estimator for the finite population total is thus given by:

$$\hat{T}_{np} = \sum_{i=1}^{n} y_i + \sum_{i=n+1}^{N} \hat{m}_{NW}(x_i) \tag{3.14}$$

The estimator given in equation (3.14) was first suggested by (Dorfman, 1992). For kernel regression estimator, the estimate of m at point x is obtained using a weighted function of observations in the h-neighbourhood of x. The weight given to each observation in the neighbourhood depends on the choice of kernel function.

For example a uniform kernel function would assign the same weights to all the points within its window while the bi-weight kernel function on the other hand, assigns more weight to the points closest to the target and diminishes the weights in those points that are "farthest away" from the centre of the kernel.

3.5 The "oh" notations in order algebra

This brief section gives a summary of common notations used in asymptotic theory in probability. The motivation behind it is that it helps one to approximate distribution of large sample statistics with a limiting distribution which is often much simpler to work with. Formal definitions of the "oh" notations and their variants, as well as how to work with the notations have been shown. Simple illustrations on their use have been given. Table 3.2 gives an overview of these notations.

Table 3.2: Common notations for asymptotic expressions

Notational form	The implication
$f(x) = O(g(x))$ $(x \to x_0)$	There exists a constant $\delta > 0$ such that $f(x) = O((g(x))$ $(\|x \to x_0\| \leq \delta)$.
$f(x) = O(g(x))$	There exists a constant x_0 such that $f(x) = O((g(x))$ $(x \geq x_0)$.
$f(x) = o(g(x))$	$f(x) = o(g(x))$ $(x \to \infty)$.

3.5.1 The Big "Oh"-notation

Let $f(x)$ and $g(x)$ be two functions of a real variable x. These functions are considered as $x \to \infty$. Given such functions, defined for all sufficiently large real numbers x, the expression relating the two functions is given by:

$$f(x) = O(g(x)) \tag{3.15}$$

The implication of this is that: *There exist constants c and x_0 such that:*

$$|f(x)| \leq c|g(x)| \quad \text{for } (x \geq x_0) \tag{3.16}$$

If this expression holds, then $f(x)$ is said to be of order $O(g(x))$ i.e the big O- estimate for $f(x)$. The constant c is thus the O-constant while the range $x \geq x_0$, is referred to as the range of validity of the O-estimate. In mathematics, it is often important to handle the error term of an approximation.

An illustration on the expanded form of the function of e^x to express the fact that the error is smaller in absolute value than some constant times x^3 if x is close enough to 0 is written as

$e^x = 1 + x + x^2 / 2 + O(x^3)$, for $x \geq 0$

3.5.2 The little "oh"-notation

The o-notation in the expression

$$f(x) = o\big(g(x)\big) \text{ as } x \to \infty \qquad (3.17)$$

means that $g(x) \neq 0$ for sufficiently large x and $\lim_{x\to\infty} \frac{f(x)}{g(x)} = 0$. If the relation holds, then $f(x)$ is said to be of smaller order than $g(x)$. The o-estimate is stronger than the corresponding O-estimate. Put differently, every function that is little-o of g is also big-O of g, but not every function that is big-O of g is also little-o of g.

The distinction between the earlier definition for the big-O notation and the current definition of little-o is that while the former has to be true for at least one constant c the latter must hold for every positive constant, ε, however small (Cormen, *et al*, 2001). These notations have been reviewed because they will be useful in the sections that follow.

3.6 Some Important Results from the Theory of Kernel Estimation

Condition 1

(a) *Suppose $\{X_i, i = 1, 2, \ldots, n\}$ is a collection of random variables from a distribution that is iid with the cdf M(x) and pdf m(x).*

b) *In the neighborhood of x, m(x) is bounded and twice continuously differentiable with bounded derivatives.*

It is assumed that $m(x)$ exists at x when discussing this function. The point of interest is to estimate this function without imposing assumptions on its functional form as in the parametric case.

Since $M(x) = EI(X_i \leq x)$ (3.18)

then its estimator is given by:

$$\hat{M}(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x) \quad (3.19)$$

which implies

$$\hat{M}(x) \xrightarrow[p]{} M(x)$$

In fact stronger results can be obtained (Glivenko-Cantelli Theorem, see (Van der Vaart, 1998)).

$$\sup_{x\in\Re}\left|\hat{M}(x) - M(x)\right| \xrightarrow[a.s]{} 0$$

By Central Limit Theorem hereby not stated,

$$n^{\frac{1}{2}}\left(\hat{M}(x) - M(x)\right) \xrightarrow[d]{} N\left(0, M(x)(1 - M(x))\right)$$

Furthermore, for $x_1, x_2 \in \Re$, $n^{\frac{1}{2}}\left(\hat{M}(x_1) - M(x_1)\right)$ and $n^{\frac{1}{2}}\left(\hat{M}(x_2) - M(x_2)\right)$ are jointly asymptotically normal with mean zero and covariance $M(x_1 \wedge x_2) - M(x_1)M(x_1)$,

where $x_1 \wedge x_2$ is the minimum between x_1 and x_2.

Since

$$m(x) = \frac{dM(x)}{dx} = \lim_{h \to 0} \frac{M(x+h) - M(x-h)}{2h} \tag{3.20}$$

Then the estimator is:

$$\hat{m}(x) = \frac{d\hat{M}(x)}{dx} = \lim_{h \to 0} \frac{\hat{M}(x+h) - \hat{M}(x-h)}{2h}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \frac{I\{x - h \le X_i \le x + h\}}{2h} \tag{3.21}$$

where the h in the $\hat{m}(x)$ is a function of the sample size n such that $\lim_{n \to \infty} h = 0$. Note that (3.21) may also be rewritten as an estimate of the unknown function $m(x)$ as follows (Todd, 2014):

$$\hat{m}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h} \right) \tag{3.22}$$

This is the class of standard kernel density estimators proposed by (Rosenblatt, 1956) and (Parzen, 1962).

Now with the kernel functions and the kernel estimator as defined in the earlier sections, the following conditions must hold.

Condition 2 (a) $\int K(z)dz = 1$

(b) $K(z) = K(-z)$

(c) *K is compactly supported on [−1,1] and bounded.*

(d) $\int z^2 K(z)dz \neq 0$ *say* $K_2(K)$

These conditions are used in the derivation of the properties of the density estimator as well as the regression estimator given in the next sections.

3.7 Standard kernel density estimator

An insight on how the kernel density function works is given first. This idea has been illustrated using Fig. 3.2 constructed from an artificial data set. It should be noted that while the area under the density estimate is equal to one, each of the rescaled kernel function has an area equal to $\frac{1}{n}$. This can be obtained by integration as follows.

$$\int \frac{1}{nh} K\left(\frac{x - X_i}{h}\right) dx = \frac{1}{nh} \int K(z) h dz = \frac{1}{nh} h \int K(z) dz = \frac{1}{n} \qquad (3.23)$$

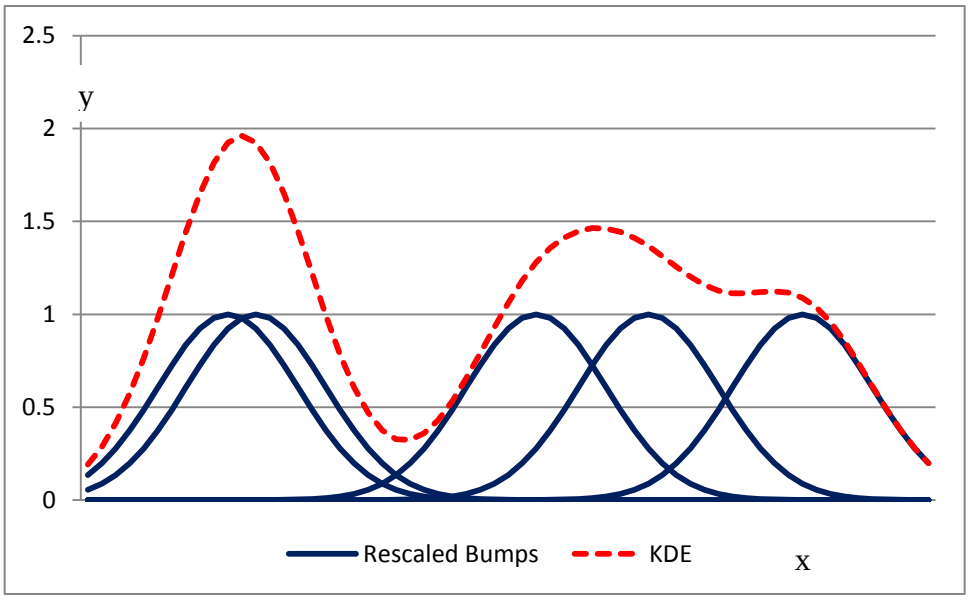


Figure 3.2: Kernel density estimate viewed as sum of bumps

This is the sum over all the rescaled kernels (the little bumps). The estimated function is a result of the vertical sum of the bumps centred on the observed values of *x*.

Unlike histograms, kernel estimates do not depend on the choice of the origin. They are smoother than histogram estimators since they inherit the smoothness of the kernel chosen and have a faster rate of convergence. As will be noticed, increasing the

bandwidth increases the amount of smoothing in the estimate (i.e. large h would give a flat estimate) while a small $h(\to 0)$ reveals the finer details of the distribution.

As stated earlier the choice of the kernel function is not very crucial as the bandwidth itself (Wand & Jones, 1995). For this reason and that of its ease in obtaining the optimal bandwidth during the bandwidth selection, the Gaussian kernel function has been used in this study. It also has an advantage in that the weights are always positive (Todd, 2014). In this study density estimators have not been pursued beyond this point since the focus of this research is that of regression estimation. In the next section the theoretical properties of Nadaraya-Watson estimator have been derived.

3.8. Properties of Nadaraya- Watson estimator

Because of the property of symmetry and clarity of the estimator let the Nadaraya-Watson kernel estimator given in (3.6) be rewritten as:

$$\hat{m}_{NW}(x) = \sum_{i=1}^{n} \frac{K\left(\frac{X_i - x}{h}\right) Y_i}{\sum_s K\left(\frac{X_i - x}{h}\right)} \tag{3.24}$$

$$= \frac{1}{\hat{f}(x)}\left[\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) Y_i\right] \tag{3.25}$$

For finite population total, the estimator is given in (3.14) that is $\hat{T}_{np}$.

The bias is then given by:

$$Bias\left(\hat{T}_{np}\right) = E\left(\hat{T}_{np} - T\right)$$

$$= E\left(\left[\sum_{i=1}^{n} y_i + \sum_{i=n+1}^{N} \hat{m}_{NW}(x_i)\right] - \left[\sum_{i=1}^{n} y_i + \sum_{i=n+1}^{N} y_i\right]\right)$$

$$= E\left(\sum_{i=n+1}^{N} \hat{m}_{NW}(x_i) - \sum_{i=n+1}^{N} y_i\right)$$

$$= E\left(\sum_{i=n+1}^{N} \hat{m}_{NW}(x_i) - \sum_{i=n+1}^{N} m(x)\right)$$

$$Bias\left(\hat{T}_{np}\right) = E\left(\sum_{i=n+1}^{N} \hat{m}_{NW}(x_i) - \sum_{i=n+1}^{N} m(x)\right) \tag{3.26}$$

But from the model in (3.3), $Y_i = m(X_i) + e_i$. This can be rewritten as;

$$Y_i = m(x) + \left[m(X_i) - m(x)\right] + e_i \tag{3.27}$$

So that from (3.25), the first term in (3.26) before taking the expectation is given by:

$$\frac{1}{\hat{f}(x)} . \sum_{i=n+1}^{N} \left(\frac{1}{nh} K\left(\frac{X_i - x}{h}\right) Y\right) = \frac{1}{\hat{f}(x)} \sum_{i=n+1}^{N} \left[\frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) m(x)\right.$$

$$+ \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) \left[m(X_i) - m(x)\right]$$

$$\left. + \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) e_i \right] \tag{3.28}$$

This may be rewritten as:

$$\frac{1}{nh\hat{f}(x)} . \sum_{i=n+1}^{N} \left(\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) Y_i\right) = \frac{1}{nh\hat{f}(x)} . \sum_{i=n+1}^{N} \left[\hat{f}(x) m(x) + m_1(x) + m_2(x)\right]$$

(3.29)

Hence taking expectation of (3.29), the estimator becomes:

$$E\left[\sum_{i=n+1}^{N} \hat{m}_{NW}(x)\right] = E\left[\frac{1}{nh} \sum_{i=n+1}^{N} \left(m(x) + \frac{\hat{m}_1(x)}{\hat{f}(x)} + \frac{\hat{m}_2(x)}{\hat{f}(x)}\right)\right] \tag{3.30}$$

Theorem 1

Under conditions 1 and 2 in section (3.6)

$$MSE[\hat{T}_{np}] = \frac{(N-n)^2 R(K)\sigma^2}{nhf(x)} + \frac{(N-n)^2}{4n^2} h^4 K_2{}^2(K)\left[m''(x) + 2\frac{f'(x)m'(x)}{f(x)} \right]^2$$

$$+ o(h^4) + o\left(\frac{(N-n)^2}{nh} + \frac{1}{nh} \right)$$

Proof

To tackle this theorem, the bias and variance terms have been derived separately as shown in the respective subsections.

3.8.1 Bias term of $\hat{T}_{np}$

From the fact that $E(e/X_i) = 0$, it will follow that $E[\hat{m}_2(x)] = 0$. Therefore for $\hat{m}_1(x)$ the following can be obtained.

$$E\sum_{i=n+1}^{N}[\hat{m}_1(x)] = \frac{1}{nh}\sum_{i=n+1}^{N} E\left\{ \sum_{i=1}^{n} K\left(\frac{X_i - x}{h} \right)[m(X_i) - m(x)] \right\}$$

$$= \left(\frac{N-n}{nh} \right) \int \left\{ K\left(\frac{u-x}{h} \right)[m(u) - m(x)] \right\} f(u)du \qquad (3.31)$$

Letting $z = \frac{u-x}{h} \Rightarrow u = x + hz$ and thus $du = hdz$, therefore:

$$E\sum_{i=n+1}^{N}[\hat{m}_1(x)] = \left(\frac{N-n}{nh} \right) \int K(z)[m(x+hz) - m(x)] f(x+hz)hdz$$

$$= \left(\frac{N-n}{n} \right) \int K(z)[m(x+hz) - m(x)] f(x+hz)dz$$

(3.32)

Applying Taylor's expansion for the v^{th} order kernels the following can be obtained:

$$m(x+hz) = m(x) + m'(x)hz + \frac{1}{2}m''(x)h^2z^2 + ... + \frac{m^{\nu}(x)h^{\nu}z^{\nu}}{\nu!} + o(h^{\nu})$$

(3.33)

And similarly

$$f(x+hz) = f(x) + f'(x)hz + \frac{1}{2}f''(x)h^2z^2 + ... + \frac{f^{\nu}(x)h^{\nu}z^{\nu}}{\nu!} + o(h^{\nu})$$

(3.34)

For the second order kernels the expansions become:

$$\left[m(x+hz) - m(x)\right] = m'(x)hz + \frac{1}{2}m''(x)h^2z^2 \tag{3.35}$$

And

$$f(x+hz) = f(x) + f'(x)hz + \frac{1}{2}f''(x)h^2z^2 \tag{3.36}$$

Restricting expansions to order $o(h^2)$ this implies that:

$$\begin{aligned} E\sum_{i=n+1}^{N}\left[\hat{m}_1(x)\right] &= \left(\frac{N-n}{n}\right)\int K(z)\left(m'(x)hz + \frac{1}{2}m''(x)h^2z^2\right)\left(f(x) + f'(x)hz\right)dz \\ &= \left(\frac{N-n}{n}\right)f(x)m'(x)h\int zK(z)dz + \left(\frac{N-n}{n}\right)f'(x)m'(x)h^2\int z^2K(z)dz \\ &\quad + \left(\frac{N-n}{n}\right)\frac{1}{2}f(x)m''(x)h^2\int z^2K(z)dz + o(h^2) \end{aligned} \tag{3.37}$$

From condition 2(c), $\int zK(z)dz = 0$ and $\int z^2K(z)dz = K_2(K)$ thus $E\left[\hat{m}_1(x)\right]$reduces to:

$$\begin{aligned} E\sum_{i=n+1}^{N}\left[\hat{m}_1(x)\right] &= \left(\frac{N-n}{n}\right)\left[\int f'(x)m'(x)h^2z^2K(z)dz + \int\frac{1}{2}f(x)m''(x)h^2z^2K(z)dz\right] + o(h^2) \\ &= \left(\frac{N-n}{n}\right)\left[f'(x)m'(x) + \frac{1}{2}f(x)m''(x)\right]h^2\int z^2K(z)dz + o(h^2) \\ &= \left(\frac{N-n}{n}\right)\left[f'(x)m'(x) + \frac{1}{2}f(x)m''(x)\right]h^2K_2(K) + o(h^2) \end{aligned}$$

(3.38)

From equation (3.30) the expected value of the second term given by $\frac{\hat{m}_1(x)}{\hat{f}(x)}$ becomes:

$$E\sum_{i=n+1}^{N}\left[\hat{m}_1(x)\right] = \left(\frac{N-n}{n}\right)\left[\frac{1}{2}m''(x)+\left[f(x)\right]^{-1}f'(x)m'(x)\right]h^2K_2(K)+o(h^2)$$

$$= \left(\frac{N-n}{n}\right)h^2K_2(K)B(x)+o(h^2) \qquad (3.39)$$

where

$$B(x) = \tfrac{1}{2}m''(x)+\left[f(x)\right]^{-1}f'(x)m'(x) \qquad (3.40)$$

From (3.26) the bias is given by $\sum_{i=n+1}^{N}\left[E\left(\hat{m}_{NW}(x)-m(x)\right)\right]$, hence from equation (3.30), it therefore follows that:

$$\text{Bias}\left[\hat{T}_{np}\right] = \left(\frac{N-n}{n}\right)h^2K_2(K)B(x)+o(h^2) \qquad (3.41)$$

That is:

$$\text{Bias}\left[\hat{T}_{np}\right] = \left(\frac{N-n}{n}\right)h^2K_2(K)\left[\frac{1}{2}m''(x)+\left[f(x)\right]^{-1}f'(x)m'(x)\right]+o(h^2) \qquad (3.42)$$

Undoubtedly the finite population total estimator using standard Nadaraya-Watson estimator is a biased estimator. A large amount of bias is induced at the boundary by the weighting function of the Nadaraya-Watson estimator. To give a pictorial view of this fact, data of size n = 100 was artificially generated using a cubic function given by the model $Y = 10 - X^3 + e$, where $X \sim U(1,2)$ and $e \sim N(0,0.5)$. The model was chosen because it showed the boundary problem clearly. The three figures, Fig.3.3-3.5, were obtained from this data. From the Fig. 3.3 the fitted curve illustrates how the Nadaraya-Watson estimator has failed to properly capture the trend on the two boundaries- there is underestimation on the left boundary and overestimation on the right. Clearly this bias term in this estimator still depends on the marginal density function of $f(x)$ and its

derivative $f'(x)$. It is also evident that the bias term is directly proportional to the square of the bandwidth *h*. This implies that the larger the bandwidth, the larger the bias and vice-versa. Cumulatively this affects the total estimator as well. This means therefore that to have a small bias one has to keep the bandwidth considerably small. The problem is how small should it be and does this small value have any effect on the other measure of variance? The insight of this has been given in Fig. 3.4

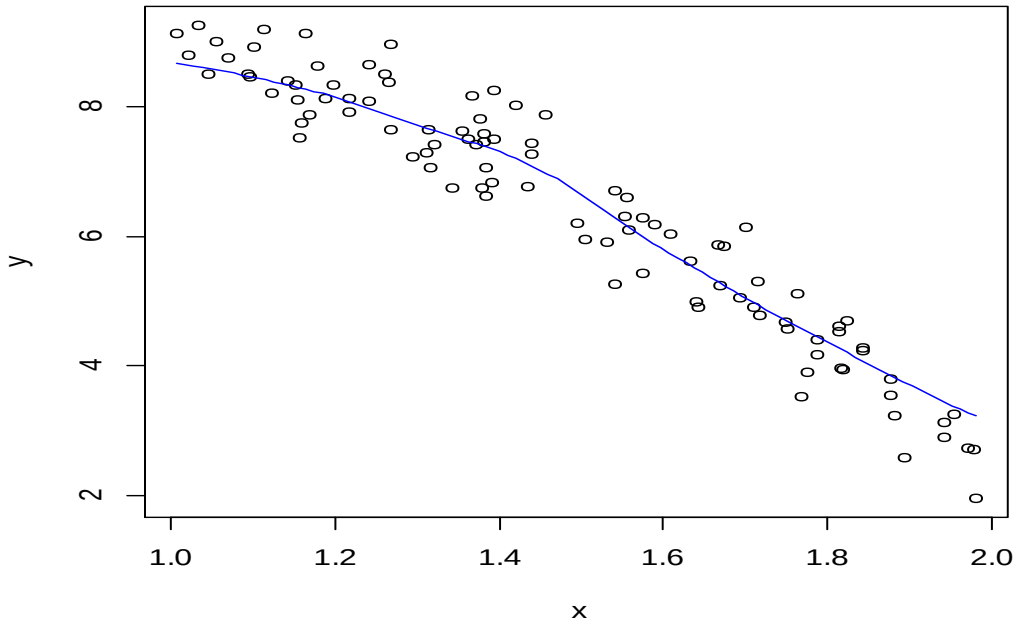


Figure 3.3: Boundary bias of Nadaraya-Watson estimator with h=0.25

From Fig. 3.3 it can be noticed the on the left boundary most of the points plotted are above the fitted line and are below it on the extreme right.

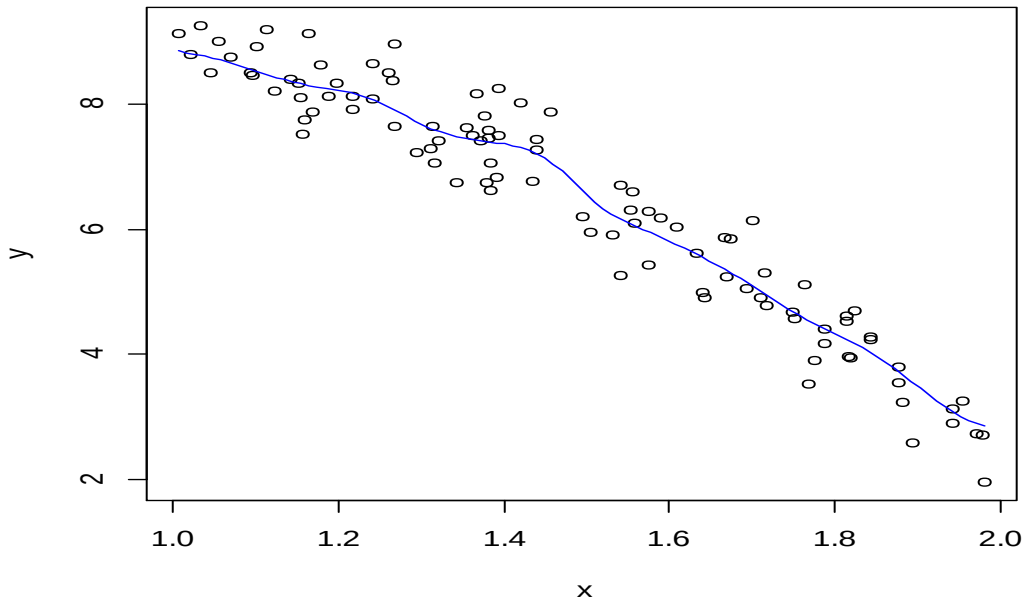


Figure 3.4: Nadaraya-Watson estimator with smaller bandwidth of h=0.15

As stated above the direct variation of the bias and the bandwidth is shown on Fig. 3.4. The smaller bandwidth of h=0.15 has slightly reduced the bias although this improvement is still insignificant at the boundary. Also one can note how the fitted line has started showing some variation.

Reducing the bandwidth to much smaller value even reduces the bias further, but the fitted line become more wiggly and unappealing. See Fig. 3.5 for this revelation. Such a curve characterized by a lot of variation is not desirable in estimation. It is clear that this reduction of the bandwidth worsens the situation and is therefore not a satisfactory solution of the problem. As to how small the bandwidth should be, depends on its impact on the variance that can be tolerated i.e. how the variance term behaves with respect to

this bandwidth. The relationship between the bandwidth and variance is shown in the derivation done in the next section.

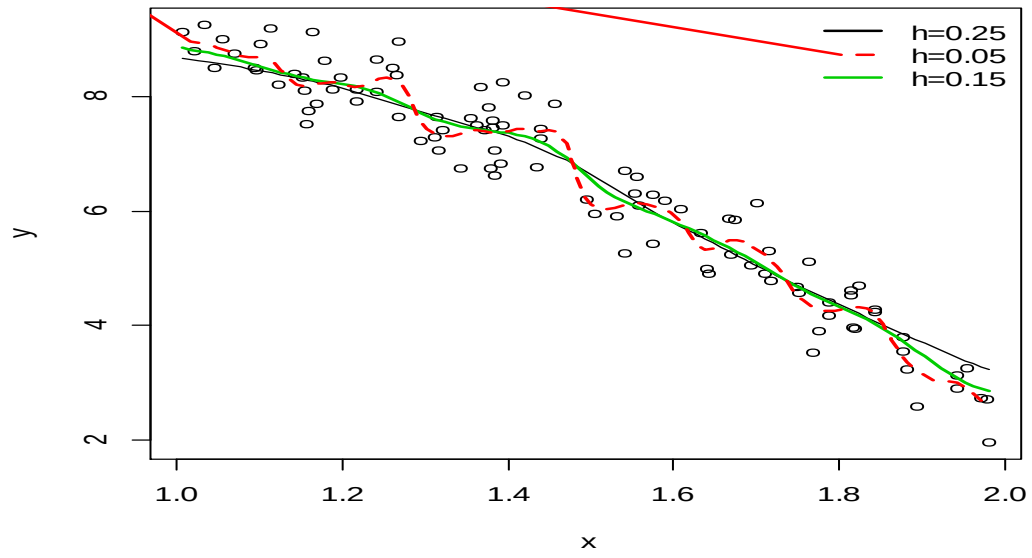


Figure 3.5: Nadaraya-Watson estimator with three different bandwidths

3.8.2 Variance term of a $\hat{T}_{np}$

From (3.28) and (3.30), it can be noted that:

$$\hat{m}_2(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)e_i \tag{3.43}$$

Thus

$$Var\sum_{i=n+1}^{N}\left[\hat{m}_2(x)\right] = Var\sum_{i=n+1}^{N}\left[\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)e_i\right]$$

$$= \frac{(N-n)^2}{n^2h^2}\sum_{i=1}^{n}\left[VarK\left(\frac{X_i - x}{h}\right)e_i\right]$$

$$Var\sum_{i=n+1}^{N}\left[\hat{m}_2(x)\right] = \frac{(N-n)^2}{nh^2}Var\left[K\left(\frac{X_i - x}{h}\right)e_i\right] \tag{3.44}$$

Writing (3.44) in terms of expectation gives the following results:

$$Var\sum_{i=n+1}^{N}\left[\hat{m}_2(x)\right]=\frac{(N-n)^2}{nh^2}\left\{E\left[K\left(\frac{X_i-x}{h}\right)e_i\right]^2-\left[E\left(K\left(\frac{X_i-x}{h}\right)e_i\right)\right]^2\right\} \tag{3.45}$$

The second term in the right hand side reduces to zero since $E(e/X_i)=0$. Note also that

$E(e/X_i)^2=\sigma^2$.

Thus

$$Var\sum_{i=n+1}^{N}\left[\hat{m}_2(x)\right]=\frac{(N-n)^2}{nh^2}E\left[K\left(\frac{X_i-x}{h}\right)^2\sigma^2\right]$$

$$=\frac{(N-n)^2}{nh^2}\int K\left(\frac{u-x}{h}\right)^2\sigma^2 f(u)du \tag{3.46}$$

Again letting $z=\frac{u-x}{h}\Rightarrow u-x=hz$ and $du=hdz$ leads to:

$$Var\sum_{i=n+1}^{N}\left[\hat{m}_2(x)\right]=\frac{(N-n)^2}{nh^2}\int K\left(\frac{u-x}{h}\right)^2\sigma^2 f(u)du$$

$$=\frac{(N-n)^2}{nh^2}\int K(z)^2\sigma^2 f(x+hz)hdz$$

$$=\frac{(N-n)^2}{nh^2}\int K(z)^2\sigma^2 f(x+hz)hdz$$

$$=\frac{(N-n)^2}{nh}\int K(z)^2dz\sigma^2 f(x)+o\left(\frac{1}{nh}\right)$$

$$Var\sum_{i=n+1}^{N}\left[\hat{m}_2(x)\right]=\frac{(N-n)^2}{nh}R(K)\sigma^2 f(x)+o\left(\frac{1}{nh}\right) \tag{3.47}$$

The variance of $\sum_{i=n+1}^{N}\left[\hat{m}_1(x)\right]$ is as follows:

$$Var\sum_{i=n+1}^{N}\left[\hat{m}_1(x)\right]=Var\sum_{i=n+1}^{N}\left\{\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{X_i-x}{h}\right)\left[m(X_i)-m(x)\right]\right\}$$

$$= \frac{(N-n)^2}{nh^2} VarK\left(\frac{X_i - x}{h}\right)\left[m(X_i) - m(x)\right]$$

$$= \frac{(N-n)^2}{nh^2} E\left[K\left(\frac{X_i - x}{h}\right)\left[m(X_i) - m(x)\right]\right]^2 - \left\{E\left[K\left(\frac{X_i - x}{h}\right)\left[m(X_i) - m(x)\right]\right]\right\}^2$$

$$= \frac{(N-n)^2}{nh^2} E\left[K\left(\frac{X_i - x}{h}\right)\left[m(X_i) - m(x)\right]\right]^2$$

$$= \frac{(N-n)^2}{nh^2}\left[\int K\left(\frac{u-x}{h}\right)^2 \left[m(u) - m(x)\right]^2\right] f(u)du \qquad (3.48)$$

And by change of variables and Taylor's expansion, this gives:

$$Var \sum_{i=n+1}^{N}\left[\hat{m}_1(x)\right] = \frac{(N-n)^2}{nh}\int K(z)^2\left[m(x+hz) - m(x)\right]^2 f(x+hz)dz$$

$$= \frac{(N-n)^2}{nh}\int K(z)^2\left[m(x) + m'(x)hz + \ldots - m(x)\right]^2\left(f(x) + f'(x)hz\right)dz$$

$$Var \sum_{i=n+1}^{N}\left[\hat{m}_1(x)\right] = O\left(\frac{(N-n)^2 h^2}{nh}\right) \qquad (3.49)$$

This is of an order smaller than $O\left(\frac{(N-n)^2}{nh}\right)$ in (3.47), hence the variance of the population total, $\hat{T}_{np}$, stated in (3.14) is therefore given by:

$$Var\left[\hat{T}_{np}\right] = Var \sum_{i=n+1}^{N}\left[m(x) + \frac{\hat{m}_1(x)}{\hat{f}(x)} + \frac{\hat{m}_2(x)}{\hat{f}(x)}\right]$$

$$= Var \sum_{i=n+1}^{N}\left[\frac{\hat{m}_2(x)}{\hat{f}(x)}\right]$$

$$Var\left[\hat{T}_{np}\right] = \frac{(N-n)^2}{\left(\hat{f}(x)\right)^2} Var \sum_{i=n+1}^{N}\left[\hat{m}_2(x)\right] \qquad (3.50)$$

From equation (3.47), this gives:

$$Var\left[\hat{T}_{np}\right] = \frac{(N-n)^2 R(K)\sigma^2}{nhf(x)} + o\left(\frac{(N-n)^2}{nh} + \frac{1}{nh}\right) \tag{3.51}$$

A critical study at this term shows that the variance still depends on the marginal density function $f(x)$. Besides this and unlike the bias, the variance is inversely proportional to the bandwidth, *h* so that an increase in the bandwidth results in a smaller variance while a larger variance is obtained for a smaller bandwidth. The implication of this is that the two components of bias and the variance cannot be looked at in isolation. It would have been quite straight forward to reduce or eliminate the bias by just reducing the bandwidth if it were not for the fact that the same action would result on an increase on the variance. This means that finding a bandwidth that allows a compromise between the two is inevitable. See Fig. 3.6 for this insight on bias-variance trade-off in relation to the size of the bandwidth. The measure that combines both the bias and the variance is the mean square error. This is explored next.

3.8.3 Mean Square Error and *AMSE* of $\hat{T}_{np}$

$$\begin{aligned} MSE\left(\hat{T}_{np}\right) &= E\left(\hat{T}_{np} - T\right)^2 \\ &= E\left(\hat{T}_{np} - E[\hat{T}_{np}] + E[\hat{T}_{np}] - T\right)^2 \\ &= E\left(\hat{T}_{np} - E[\hat{T}_{np}]\right)^2 + E\left(E[\hat{T}_{np}] - T\right)^2 \\ &\quad + 2E\left(\hat{T}_{np} - E[\hat{T}_{np}]\right)\left(E[\hat{T}_{np}] - T\right) \\ &= Var\left(\hat{T}_{np}\right) + Bias^2\left(\hat{T}_{np}\right) + 0 \end{aligned} \tag{3.52}$$

From (3.42) and (3.51), one can have:

$$MSE\left[\hat{T}_{np}\right] = \frac{(N-n)^2 R(K)\sigma^2}{nhf(x)} + o\left(\frac{(N-n)^2}{nh} + \frac{1}{nh}\right)$$

$$+\left(\left(\frac{N-n}{n}\right)h^2 K_2(K)\left[\frac{1}{2}m''(x) + \frac{f'(x)m'(x)}{f(x)}\right] + o(h^2)\right)^2$$

$$= \frac{(N-n)^2 R(K)\sigma^2}{nhf(x)} + \frac{(N-n)^2}{4n^2}h^4 {K_2}^2(K)\left[m''(x) + 2\frac{f'(x)m'(x)}{f(x)}\right]^2$$

$$+ o(h^4) + o\left(\frac{(N-n)^2}{nh} + \frac{1}{nh}\right) \qquad (3.53)$$

This completes the proof of the theorem.

For large sample, that is as n approaches N, and a sufficiently small bandwidth $MSE\left[\hat{m}(x)\right]$ in (3.53) become:

$$\approx \frac{(N-n)^2 R(K)\sigma^2}{nhf(x)} + \frac{(N-n)^2}{4n^2}h^4 {K_2}^2(K)\left[m''(x) + 2\frac{f'(x)m'(x)}{f(x)}\right]^2 \qquad (3.54)$$

$=AMSE\left(\hat{T}_{np}\right)$ since this is based on asymptotic expansions.

Equation (3.54) clearly indicates that as $h \to 0$ and $n \to N$, the *MSE* of the kernel regression estimator approaches zero. Hence the estimator is consistent. It should also be noted, though unfortunately, that in addition to the marginal density of $f(x)$ and its derivative $f'(x)$, *MSE* still depends on the functions $m(x)$ and $m''(x)$ both of which being unknown in practice. Obtaining an optimum value of h does not drop these functions either, and further depends on x making it a local bandwidth. See more details in (Härdle *et al*, 2005).

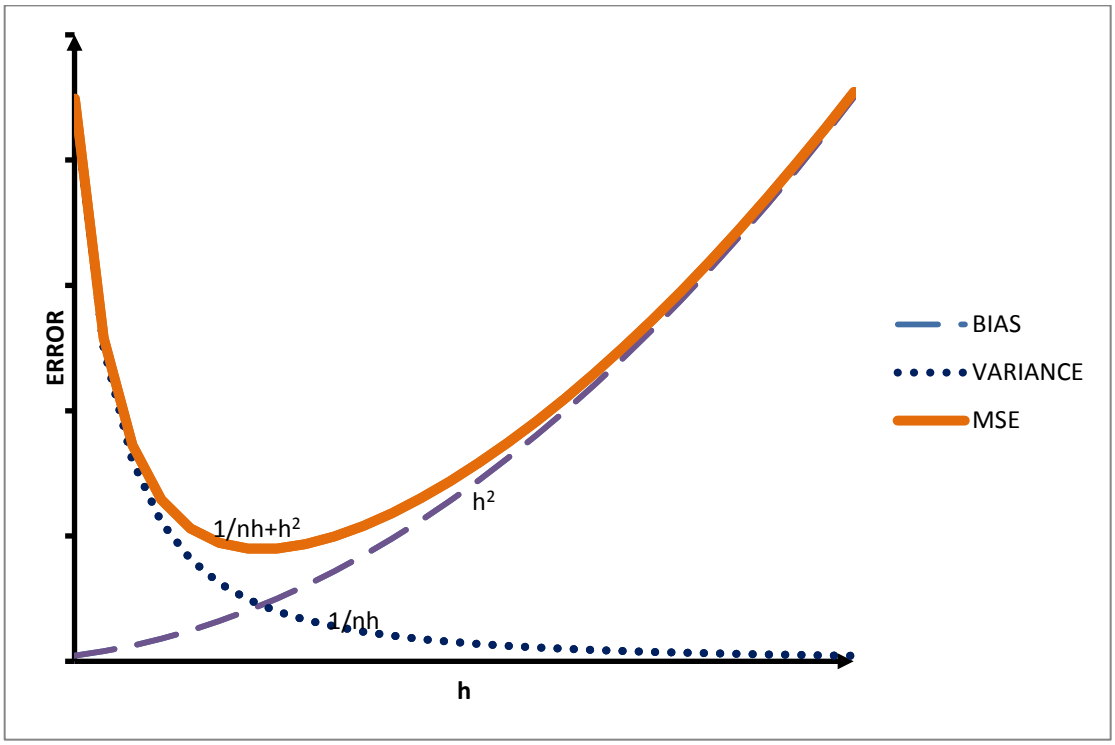


Figure 3.6: Impact of the bandwidth on the bias, variance and MSE.

3.9 Boundary Effects due to use of the Nadaraya-Watson Estimator

It is now clear that the above estimator due to (Dorfman, 1992) suffers from the boundary problem induced by the Nadaraya-Watson estimator used. It has been noted that it is necessary to have an optimal bandwidth to solve the problem of bias-variance trade-off. This bandwidth, however, does not eliminate the boundary bias. This study, therefore, proceed to propose an estimator that minimizes this effect significantly. If K is a symmetric function and fixed across the support estimation then inference are generally simplified for unbounded support i.e. $(-\infty, \infty)$. But $\hat{m}(x)$ is inconsistent at the boundary [0, h] for such a choice of K (Malec & Melanie, 2012). For $x \in [0, h)$, the bias of $m(x)$ is of order $o(h)$ rather than $o(h^2)$.

In literature many techniques of removing boundary effects in density estimation have been proposed. Four common techniques have been mentioned briefly in this study.

These are: - Reflection of Data technique, Transformation of data technique, Pseudo data methods, and Boundary kernel methods. For an overview of these techniques see (Karunamuni & Alberts, 2004). Of all these none has been taken further to regression estimation. The idea behind the Reflection of data technique is a motivation in this study. The proposed population estimator uses this technique to modify the smoothing weight. These techniques have been briefly highlighted as used in density estimation.

3.9.1 Reflection of Data Method

This technique has been proposed by researchers such as (Cline & Hart, 1991), (Schuster, 1985) and (Silverman, 1986). It is also referred to as the data-reflected technique. In this method one has to simply add $-X_1, -X_2, \ldots, -X_n$ to the data set. This addition of data is made because the kernel estimator is penalizing for lack of data on the negative axis and is therefore gradually using reduced amount of data in its window as it approaches the boundary thus resulting in a boundary bias. The estimator for this method is given by:

$$\hat{m}(x) = \frac{1}{nh}\sum_{i=1}^{n}\left\{K\left(\frac{x - X_i}{h}\right) + K\left(\frac{x + X_i}{h}\right)\right\}, \text{ for } x \geq 0 \qquad (3.55)$$

$$\hat{m}(x) = 0, \text{ for } x < 0$$

It can be shown that $\hat{m}'(x) = 0$. Hence it is a very good method if the underlying density has the property $m'(0) = 0$.

3.9.2 Transformation of data Method

This is a technique that has been studied by (Wand et al, 1991), and (Marron & Ruppert, 1994). In this method one can take a one-to-one continuous function $g : [0,\infty) \to [0,\infty)$. A regular kernel estimator is then used with the transformed data set $\{g(X_1), g(X_2), \ldots, g(X_n)\}$.

The estimator is thus given by:

$$\hat{m}(x) = \frac{1}{nh}\sum_{i=1}^{n}\left\{K\left(\frac{x - g(X_i)}{h}\right)\right\} \tag{3.56}$$

It should be noted that the estimator is not estimating the *p.d.f* of *X*, but instead that of *g*(*X*). This therefore leaves room for manipulation where one can choose *g* so as to get the data produce what one needs.

3.9.3 Pseudo Data Methods

The technique is due to (Cowling & Hall, 1996). It involves generating of data beyond the left end point of the support of the density. It is a kind of a "reflection transformation estimator," which transforms the data into a new set, then puts it on the negative axis. The estimator is given by:

$$\hat{m}(x) = \frac{1}{nh}\left[\sum_{i=1}^{n}K\left(\frac{x - X_i}{h}\right) + \sum_{i=1}^{m}K\left(\frac{x + X_{(-i)}}{h}\right)\right] \tag{3.57}$$

where $m \le n$ and

$$X_{(-i)} = -5X_{(\frac{i}{3})} - 4X_{(\frac{2i}{3})} + \frac{10}{3}X_{(i)} \tag{3.58}$$

where $X_{(t)}$ linearly interpolates among 0, $X_{(1)}$, $X_{(2)}$, …, $X_{(n)}$

3.9.4 Boundary Kernel Method

This technique has been explored by many researchers including (Gasser & Müller, 1979; Gasser *et al*, 1985; Jones, 1993; Müller, 1991; Zhang & Karunamuni, 2000). In this method a different kernel for estimating function is used at each point in the boundary region. In the investigations so far carried out, it has been noted that new kernels give up the symmetry property and put more weight on the positive axis. The estimator for this technique is given by:

$$\hat{m}(x) = \frac{1}{nh_c}\sum_{i=1}^{n}K_{(c/b(c))}\left(\frac{x - X_i}{h_c}\right) \tag{3.59}$$

where $x=ch,\ 0 \le c \le 1$, and $b(c) = 2\text{-}c$

Also $K_{(c)}\ 0 \le c \le 1, K_{(c)}(z) = \frac{12}{(1+c)^4}(1+z)\left\{(1-2c)z + \frac{3c^2-2c+1}{2}\right\}1\{-1 \le z \le 1\}$

(3.60)

The major drawback of this method is that the estimates might be negative.

In this study the use of the reflection technique has been explored in the investigation. This is because it is straight forward to calculate and implement compared to the other three techniques. Before reviewing the properties of this reflection technique, the estimator of the finite population total constructed using this technique is proposed in the next section.

3.10 Proposed estimator of finite population total ($\hat{T}_{npr}$)

Let the variables X and Y be auxiliary and the study variables respectively as earlier stated. To obtain a robust nonparametric regression estimator for the finite population total that addresses the boundary effect, the function given in equation (3.55) has been modified and used to construct the required estimator so that the non-sample part in equation (3.2) can be estimated using it, to give the following proposed estimator:

$$\hat{T}_{npr} = \sum_{i=1}^{n} y_i + \sum_{i=n+1}^{N} \hat{m}_{ref}(x_i) \qquad (3.61)$$

Where the first term $\sum_{i=1}^{n} y_i$ is the sample total observed and therefore under model-based approach it will not be necessary to be estimated while the second term $\sum_{i=n+1}^{N} \hat{m}_{ref}(x_i)$ is the non-sample total term that is to be estimated nonparametrically using the reflection technique. As noted earlier the Nadaraya-Watson estimator induces a bias at the boundary. This is because at the boundary the interval where $x \in [0, h)$, the symmetric

kernel has decreased amount or lacks data on part of its window. The data-reflected technique therefore provides the data through reflection method so that this information is put on the negative axis thereby supplying the kernel with the information required on this section. The following steps give the procedure on how it works.

Let the $\{(X_1,Y_1), (X_2, Y_2),..., (X_n ,Y_n)\}$ be the set of n observations in the sample. The data is augmented by adding the reflections of all the points in the boundary, to give the set $\{(X_1, Y_1), (-X_1, Y_1), (X_2, Y_2), (- X_2, Y_2) ..., (-X_n, Y_n), (X_n , Y_n)\}$. If a kernel estimate $m^*(x)$ is constructed from this data set of size $2n$, then an estimate based on the original data can be given by putting $\hat{m}(x) = 2m^*(x)$, for $x \geq 0$, and zero otherwise. This gives the modified general weight function given by:

$$\hat{m}_{ref}(x) = \sum_{i=1}^{n} \frac{\left\{ K\left(\frac{x - X_i}{h}\right) + K\left(\frac{x + X_i}{h}\right)\right\} Y_i}{\sum_{i=1}^{n}\left\{ K\left(\frac{x - X_i}{h}\right) + K\left(\frac{x + X_i}{h}\right)\right\}} \qquad (3.62)$$

It can be shown that the estimate will always have zero derivative at the boundary, provided the kernel is symmetric and differentiable. In practice it will not usually be necessary to reflect the whole data set, since if X_i/h is sufficiently large, the reflected point $- X_i/h$ will not be felt in the calculation of $m^*(x)$ for $x \geq 0$, and hence reflection of points near 0 is all that is needed. (Silverman, 1986) in his example, states that if K is the Gaussian kernel there is no practical need to reflect points beyond $X_i > 4h$.

The next section reviews some properties that are unique to this modified kernel density estimator.

3.11 Properties of data-reflected estimator for population total ($\hat{T}_{npr}$)

3.11.1 The kernel estimator at the boundary

The interest in this study is in the boundary problem which occurs in the interval $[0, h)$. This is as a result of lack of information which follows due to truncation of such

information at this interval, i.e. the density function is continuous on [0, ∞) and is 0, for $x< 0$. This reduced amount of information leads to serious bias during the estimation and as such the estimate becomes inaccurate. The boundary problem arises when the value of x is smaller than the chosen value of the bandwidth. In the case of the standard kernel estimator of $m(x)$ given by (3.23) consider $\hat{m}(c.h)$ for $c \in [0,1)$, where $x = c.h$, then for $z = \frac{x-u}{h}$,

one can have

$$0 \le u \le \infty \Rightarrow 0 \le h(c-z) \le \infty \Rightarrow c \ge z \ge -\infty$$

For a kernel function which has the support [-1, 1], the variable z must lie within [-1, 1]. But $c \in [0,1)$, hence $c \ge z \ge -1$.

This implies that for the density estimation the expectation of the estimator is:

$$E[\hat{m}(x)] = \int_{-1}^{c} K(z)m(x+hz)dz \tag{3.63}$$

Taylor's expansion yields:

$$E[\hat{m}(x)] = m(x)\int_{-1}^{c} K(z)dz + hm'(x)\int_{-1}^{c} zK(z)dz + \frac{h^2}{2}m''(x)\int_{-1}^{c} z^2K(z)dz + o(h^2) \tag{3.64}$$

For the case of regression estimation considered in this study, it can be deduced from (3.30) that (3.64) results in:

$$\sum_{i=n+1}^{N} E[\hat{m}(x)] = \frac{N-n}{n\hat{g}(x)}\left[g(x)m'(x)h\int_{-1}^{c} zK(z)dz + \frac{1}{2}g(x)m''(x)h^2\int_{-1}^{c} z^2K(z)dz \right.$$

$$\left. - g'(x)m'(x)h^2\int_{-1}^{c} z^2K(z)dz \right] \tag{3.65}$$

This estimator will only be unbiased and consistent asymptotically if $x \ge h$ i.e. $c \ge 1$. The implication of this is that the expected value can only reach half the original value.

That is:

$$E[\hat{m}(0)] = \frac{1}{2} m(0) + O(h) \tag{3.66}$$

It should be noted that:

$$\int_{-\infty}^{\infty} \hat{m}(x)dx = 1, \text{ and also that } \int_{0}^{\infty} m(x)dx = 1,$$

But $$\int_{0}^{\infty} \hat{m}(x)dx = \frac{1}{nh} \sum_{i=1}^{n} \int_{0}^{\infty} K\left(\frac{x - X_i}{h} \right) dx$$

On letting $z = \left(\frac{x - X_i}{h} \right) \Rightarrow x = X_i + hz \Rightarrow dx = hdz$, the following is obtained:

$$\int_{0}^{\infty} \hat{m}(x)dx = \frac{1}{nh} \sum_{i=1}^{n} h \int_{-\frac{X_i}{h}}^{\infty} K(z)dz < 1, \text{ if } \exists i \in \{1,...,n\} : X_i < h \tag{3.67}$$

an indication that the density does not live up to the condition of being a *p.d.f* about its support at the boundary. One way of correcting this boundary problem is by use of data-reflected technique. Due to symmetry of the kernel function one can look at the reflection estimator as:

$$E\left(\hat{m}_{ref}(x)\right) = E\left[\hat{m}(x)\right] + E\left[\hat{m}(-x)\right] \tag{3.68}$$

3.11.2 The Bias of Data-reflected Estimation Technique in Regression

It can be shown that this reflection estimator being symmetric around the origin further has the condition:

$$K'(-z) = -K'(z) \tag{3.69}$$

So that

$$m'_{ref}(0) = \sum_{i=1}^{n} \frac{\left\{\frac{1}{h}K'\left(\frac{-X_i}{h}\right) + \frac{1}{h}K'\left(\frac{X_i}{h}\right)\right\}Y_i}{\sum_{i=1}^{n}\left\{\frac{1}{h}K'\left(\frac{-X_i}{h}\right) + \frac{1}{h}K'\left(\frac{X_i}{h}\right)\right\}} \cong 0 \qquad (3.70)$$

The implication of this is that the reflection estimator satisfies the so-called shoulder condition always. At the boundary the decreased amount of the data suggest concavity of the density in the vicinity of the origin. As a consequence, these kernels tend to misinterpret the local concavity as an indication of a mode over the strictly positive region, (Hirukawa & Sakudo, 2015). This is a condition where a given density, say *m*, has a shoulder at 0, i.e. $m'(0) = 0$. See for instance (Mack *et al*, 1999). The graphs in Fig. 3.7 show this impression. It should be noted that except for the graph on exponential data, the other two on cauchy and normal data satisfy the shoulder condition.

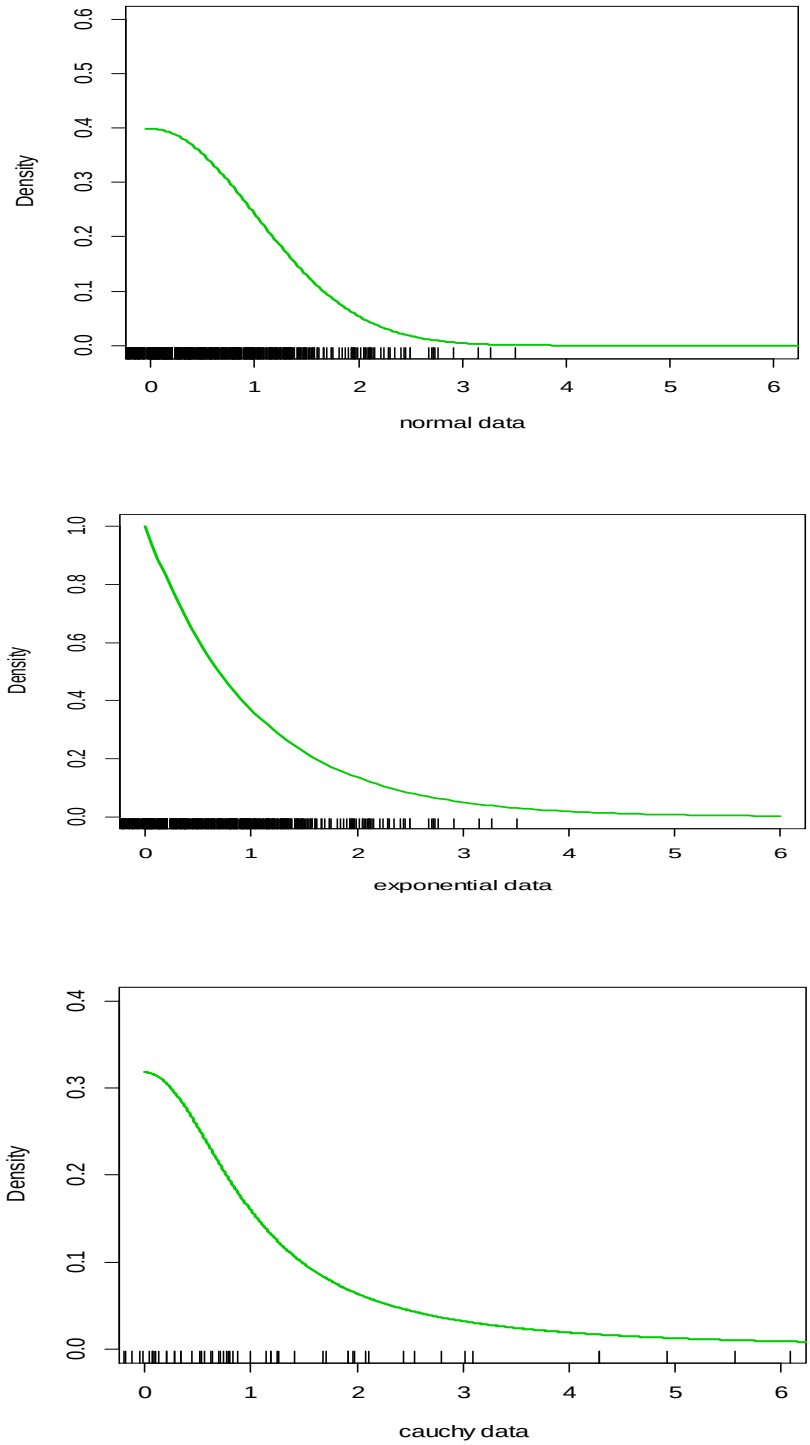


Figure 3.7: Shoulder condition

The first term in the right hand side of equation (3.68) is already given in (3.65); therefore, proceeding to look at the second term and noting that $Y_i = m(x) + [m(X_i) - m(x)] + e_i$ gives.

$$\sum_{i=n+1}^{N}\left(E[\hat{m}(-x)]\right) = \sum_{i=n+1}^{N}\frac{1}{nh\hat{g}(x)}\int_{0}^{\infty}K\left(\frac{-x-u}{h}\right)[m(u)-m(x)]g(u)du \qquad (3.71)$$

But $x = ch,\ c \in [0,1)$.

Thus

$$\sum_{i=n+1}^{N}\left(E[\hat{m}(-x)]\right) = \frac{(N-n)}{nh\hat{g}(x)}\int_{0}^{(1-c)}K\left(\frac{-x-u}{h}\right)[m(u)-m(x)]g(u)du$$

(3.72)

Since $\frac{-x-u}{h} \geq -1 \Rightarrow u = -x + h,$ then

$$\sum_{i=n+1}^{N}\left(E[\hat{m}(-x)]\right) = \frac{(N-n)}{nh\hat{g}(x)}\int_{-c}^{-1}K(z)[m(-x-hz)-m(x)]g(-x-hz)hdz$$

$$= \frac{(N-n)}{nh\hat{g}(x)}\int_{-1}^{-c}K(z)[m(x-(2x+hz))-m(x)]g(x-(2x+hz))dz$$

(3.73)

Taylor's expansion yields:

$$[m(x-(2x+hz))-m(x)] = m(x) - m'(x)(2x+hz) - \frac{1}{2}m''(x)(2x+hz)^2 + \ldots - m(x)$$

$$= -m'(x)(2x+hz) - \frac{1}{2}m''(x)(2x+hz)^2 + \ldots \quad (3.74)$$

and

$$g(x-(2x+hz)) = g(x) - g'(x)(2x+hz) + \ldots \qquad (3.75)$$

Therefore the product of this expansion is:

$$= -m'(x)g(x)(2x+hz) - \frac{1}{2}g(x)m''(x)(2x+hz)^2 + g'(x)m'(x)(2x+hz)^2$$

$$+ \frac{1}{2}g'(x)m''(x)(2x+hz)^3 \qquad (3.76)$$

Thus

$$\sum_{i=n+1}^{N}\left(E\left[\hat{m}(-x)\right]\right) = \frac{(N-n)}{nh\hat{g}(x)}\left[-2xm'(x)g(x)\int_{-1}^{-c}K(z)dz - hm'(x)g(x)\int_{-1}^{-c}zK(z)dz \right.$$

$$+ 2x^2 g(x)m''(x)\int_{-1}^{-c}K(z)dz + 2xhm''(x)g(x)\int_{-1}^{-c}zK(z)dz$$

$$\left. + \frac{h^2}{2}m''(x)g(x)\int_{-1}^{-c}z^2K(z)dz \right] + o(h^2) \qquad (3.77)$$

Because of the property of symmetry the following equality holds:

$$\left.\begin{aligned} &\int_{-1}^{c}K(z)dz = 1 - \int_{-1}^{-c}K(z)dz \\ &\int_{-1}^{c}zK(z)dz = \int_{-1}^{-c}zK(z)dz \\ &\int_{-1}^{c}z^2K(z)dz = K_2(K) - \int_{-1}^{-c}z^2K(z)dz \end{aligned}\right\} \qquad (3.78)$$

where $K_2(K) := \int_{-1}^{1}z^2K(z)dz \neq 0$, see equation (3.4)

Thus putting together the results in (3.77) with that of (3.65) yields in the following:

$$E\left[\sum_{i=n+1}^{N}\hat{m}_{ref}(x)\right]=\left(\frac{N-n}{n}\right)\left[\frac{h^2}{2}m''(x)\int_{-1}^{1}z^2K(z)dz\right.$$

$$+2h[g(x)]^{-1}m'(x)\int_{c}^{1}(z-c)K(z)dz$$

$$\left.+2h^2m''(x)\left(c^2\int_{-1}^{-c}K(z)dz+c\int_{-1}^{-c}zK(z)dz\right)\right]+o\left(h^2\right)$$

$$=\left(\frac{N-n}{n}\right)\left[\frac{h^2}{2}m''(x)\int_{-1}^{1}z^2K(z)dz\right.$$

$$+2h\left(m'(0)g(x)]^{-1}+chm''(0)+o(h)\right)\int_{c}^{1}(z-c)K(z)dz$$

$$\left.+2h^2m''(x)\left(c^2\int_{-1}^{-c}K(z)dz+c\int_{-1}^{-c}zK(z)dz\right)\right]+o\left(h^2\right) \quad (3.79)$$

The bias for the estimator of the finite population total, T_{npr}, given in equation (3.61) would therefore be given by:

$$Bias\left[T_{npr}\right]=\left(\frac{N-n}{n}\right)\left[\frac{h^2}{2}m''(x)\int_{-1}^{1}z^2K(z)dz\right.$$

$$+2h\left(m'(0)[g(x)]^{-1}+chm''(0)+o(h)\right)\int_{c}^{1}(z-c)K(z)dz$$

$$\left.+2h^2m''(x)\left(c^2\int_{-1}^{-c}K(z)dz+c\int_{-1}^{-c}zK(z)dz\right)\right]+o\left(h^2\right) \quad (3.80)$$

This clearly shows that within the boundary interval, the estimator still has a bias of order *h* while at the interior interval the expectation coincides with that of the standard kernel estimator. Notably, however, if the underlying density, *m*, has a shoulder at 0, i.e. $m'(0)=0$, the term of order *h* drops out thereby making the bias to be order h^2.

3.11.3 The variance of data-reflected kernel regression estimation technique

Similarly the variance can be computed as follows:

$$Var\left[T_{npr}\right] = E\left[T_{npr}\right]^2 - \left[E[T_{npr}]\right]^2$$

$$\operatorname{var}\left[\sum_{i=n+1}^{N}\left(\hat{m}_{ref}(x)\right)\right] = \frac{(N-n)^2}{nh^2[g(x)]^2}\operatorname{var}\left[K\left(\frac{X_i+x}{h}\right)+K\left(\frac{X_i-x}{h}\right)e\right]$$

$$= \frac{(N-n)^2}{nh^2[g(x)]^2}\left\{E\left[\left(K\left(\frac{X_i+x}{h}\right)+K\left(\frac{X_i-x}{h}\right)\right)^2 e^2\right]\right.$$

$$\left. -\frac{1}{n}\left[\frac{1}{h}E\left(K\left(\frac{X_i+x}{h}\right)+K\left(\frac{X_i-x}{h}\right)\right)e\right]^2\right\} \qquad (3.81)$$

The second term of (3.81) is zero, thus procedure of computing the first term is as follows:

$$E\left[\left(K\left(\frac{X_i+x}{h}\right)+K\left(\frac{X_i-x}{h}\right)\right)^2\right] = E\left(K\left(\frac{X_i+x}{h}\right)\right)^2 + E\left(K\left(\frac{X_i-x}{h}\right)\right)^2$$

$$+ 2E\left[K\left(\frac{X_i+x}{h}\right)K\left(\frac{X_i-x}{h}\right)\right]$$

$$= \int_0^\infty K\left(\frac{u+x}{h}\right)^2 g(u)du + \int_0^\infty K\left(\frac{u-x}{h}\right)^2 g(u)du$$

$$+ 2\int_0^\infty K\left(\frac{u+x}{h}\right)K\left(\frac{u-x}{h}\right)g(u)du$$

$$= h\int_{-1}^{c} K(z)^2 g(x+hz)dz + h\int_{-1}^{c} K(z)^2 g(x-(2x-hz))dz$$

$$+ 2h\int_{-1}^{c} K(z)K(z-2c)g(x+hz)dz$$

$$= h\int_{-1}^{c} K(z)^2\left(g(x) - hzg'(x) + \frac{h^2}{2}z^2 g''(x) + o(h^2)\right)dz$$

$$+ h\int_{-1}^{c} K(z)^2\left(g(x) - (2x+hz)g'(x) + \frac{(2x+hz)^2}{2}g''(x) + o(h^2)\right)dz$$

$$+ 2h\int_{-1}^{\infty} K(z)K(z-2c)\left(g(x) - hzg'(x) + \frac{h^2}{2}z^2 g''(x) + o(h^2)\right)dz$$

(3.82)

And from the property of symmetry of the kernel function, the following equality is obtained:

$$\int_{-1}^{c} K(z)^2 dz + \int_{-1}^{c} K(z)^2 dz = 2\int_{-1}^{-c} K(z)^2 dz + \int_{-c}^{c} K(z)^2 dz$$

$$= 2\int_{c}^{1} K(z)^2 dz + 2\int_{0}^{c} K(z)^2 dz$$

$$= 2\int_{0}^{1} K(z)^2 dz$$

$$= \int_{-1}^{1} K(z)^2 dz \qquad (3.83)$$

With this, therefore, the variance is given by:

$$\operatorname{var}\left(T_{npr}\right)=\frac{(N-n)^2\sigma^2}{nh^2\left[\hat{g}(x)\right]^2}\left(hg(x)\int_{-1}^{1}K(z)^2dz+O\left(h^2\right)\right)+O\left(n^{-1}\right)$$

$$=\frac{(N-n)^2\sigma^2}{nhg(x)}\int_{-1}^{1}K(z)^2dz+O\left(n^{-1}\right)\approx\frac{(N-n)^2\sigma^2}{nhg(x)}R(K) \quad (3.84)$$

where $R(K)=\int_{-1}^{1}K(z)^2dz$.

From the derivation of the bias and the variance, it was noted that the estimated function always fulfills the shoulder condition. This condition unfortunately is imposed even for functions whose true density does not satisfy the shoulder condition. Further to this, is that while the reflection estimator has a low variance its bias is fairly high, but still better than the Nadaraya-Watson estimator, where the shoulder condition is not satisfied. Though so, it should be noted that an impressive thing with this technique is that it is easy to calculate and at the same time very good for densities that fulfill the shoulder condition.

Moreover, the other advantage of the technique is that the estimate is a density i.e. it integrates to 1 over the whole real axis without assigning any mass on the negative axis.

3.11.4 Illustration

We illustrate the standard kernel and the effect of modifying it in the following example whose data has been simulated. The resulting curve is shown in Fig. 3.8 and Fig. 3.9.

Table 3.3: Simulated data $X_i \sim U(0,1)$, $Y_i=m(X_i)+e_i$, $m(X_i)= 10+X_i^3$, $e_i \sim N(0, 1)$

X	0.08	0.12	0.15	0.29	0.38	0.49	0.50	0.56	0.61	0.74
Y	9.31	10.34	11.08	10.98	9.59	10.95	11.79	9.52	10.47	10.28

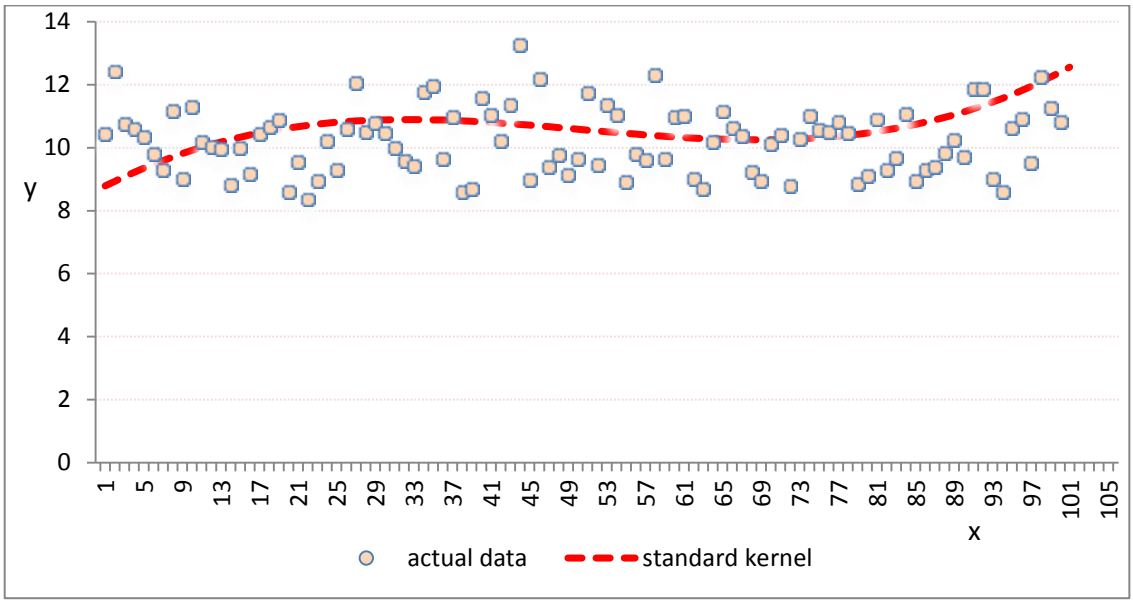


Figure 3.8: Showing Nadaraya-Watson kernel smoother with the bandwidth of h=0.5.

Note that a population of size N=100 was generated from where a simple random sample of size $n=10$ was taken and the corresponding pairs of x and y were as shown in Table 3.3.This data subjected to Kernel smoothing function used by (Dorfman, 1992). The figure also shows a scatter diagram of the actual pairs of points plotted for purposes of showing boundary effects of the kernel smoothing. Clearly there is a bias at the boundaries. Most points lie above the curve on the left and below the line on the right. Fig.3.9 shows both the regression estimation obtained using the *NW*-estimator and that modified using the reflection technique.

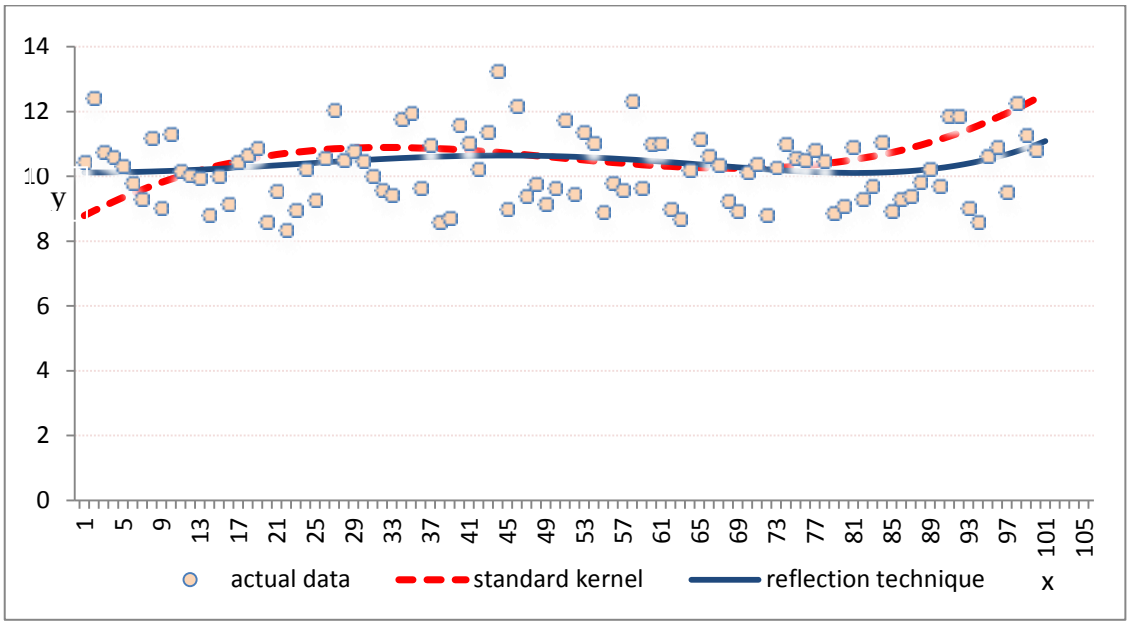


Figure 3.9: Effect of kernel modification in regression estimation

In Fig. 3.9 the graph of the standard Nadaraya-Watson and the modified kernel regression estimators which uses reflection technique given by (3.62), have been shown. It clearly addresses the bias at the boundary and some sections of the interior part of the curve especially where the observations taken are not equally spaced. The bandwidth of h=0.5 was maintained. The standard kernel regression estimation fit is given by the broken line while the modified one is given by the continuous line. It should be noted that the biggest impact is at the boundary. In the next chapter more analytical graphs have been given and discussed based the techniques and theories derived in this chapter.

CHAPTER FOUR

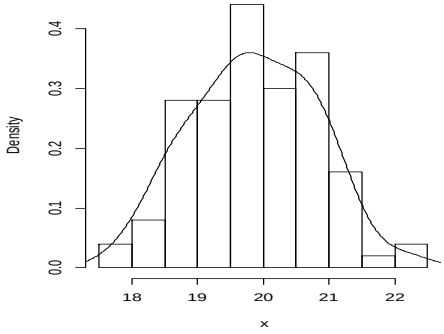
EMPIRICAL STUDY

4.1 Introduction

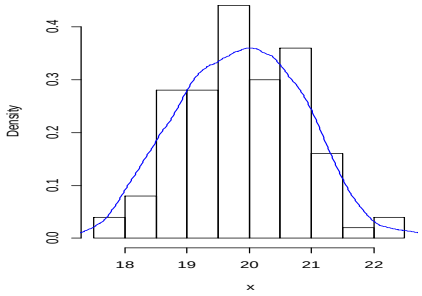
In this chapter an empirical analysis has been done using data simulated using the R software (R Core Team, 2014) and in some limited cases where it was convenient Ms-Excel was also used. Some few graphs have been done using Ms-Excel while majority of them have been done using the R software. The datasets were artificially generated, using selected distributions and whose importance and/or the use in similar researches has been explained, were used during the analysis of this study. R-codes used to obtain figures and tabulated values presented, have been given at the appendix section of this thesis.

4.2 Kernel functions in regression estimation

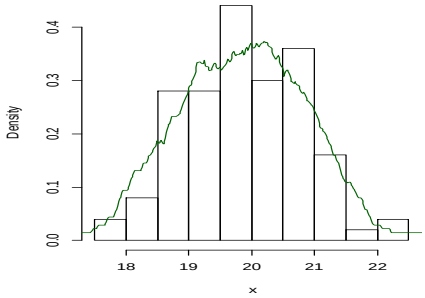
Identification of an appropriate kernel was part of the objectives dealt with on in this study. Within the theoretical framework and because of its smoothness property, Gaussian kernel function was identified. With a normal random sample of size 100, simulation using the R software was done and this kernel function was compared with the rest. The influence of the different selected kernel functions was checked, with the bandwidths and data kept constant. The resulting graphs on this study are given in Fig. 4.1 for the density estimation and regression estimation in the subsequent graphs in Fig. 4.2. Other graphs have been given in Fig. 4.3-4.4.



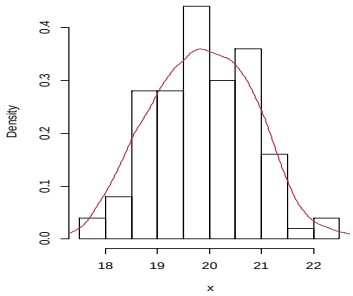
(a) Gaussian



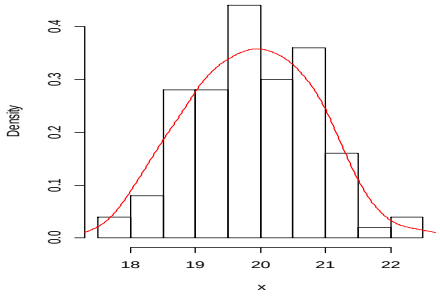
(b) Epanechnikov



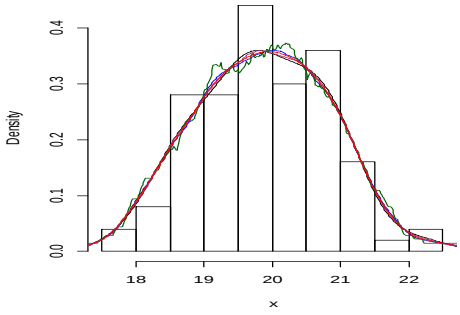
(c) Rectangular



(d) Triangular



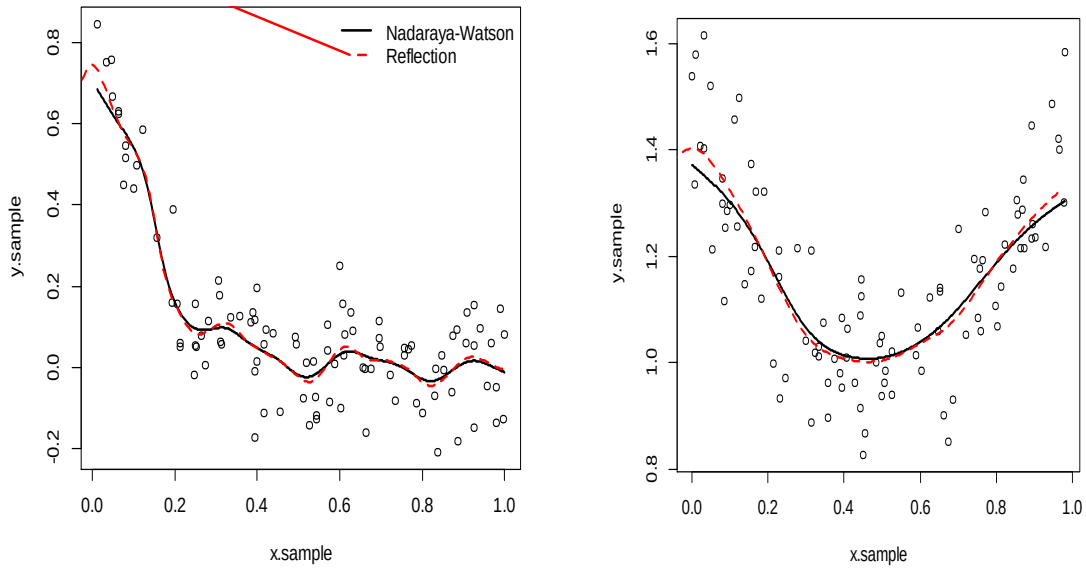
(e) Biweight



(f) Combined Graphs

Figure 4.1: Graphs obtained using various kernel functions with bw=0.39

Two models namely; quadratic and exponential were chosen for this purpose. This was done because the regression curves of the figures of these models were clearer than the others at the boundary. The graphs in Fig. 4.1 are from an artificial data set simulated using a random normal distribution with mean 20 and standard deviation of 1. Note, however, that every kernel function has its own optimal bandwidth. These bandwidths may not necessarily be the same. The histograms only serve to give a rough picture on the data used in the density plot.



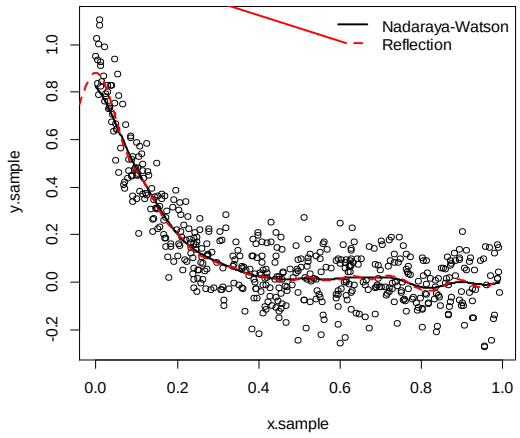
(a) (b)

Figure 4.2: Comparative regression Graphs for Nadaraya-Watson and Reflection estmators

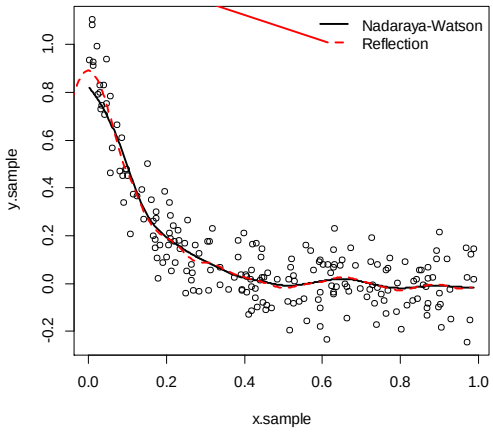
In Fig. 4.2 the sample size of n = 100, was kept constant for the exponential model in (a) and the quadratic model in (b). The two graphs in the figure show how the reflection technique addressed the boundary problem.

4.3 Nadaraya-Watson and Data-reflection Technique in regression

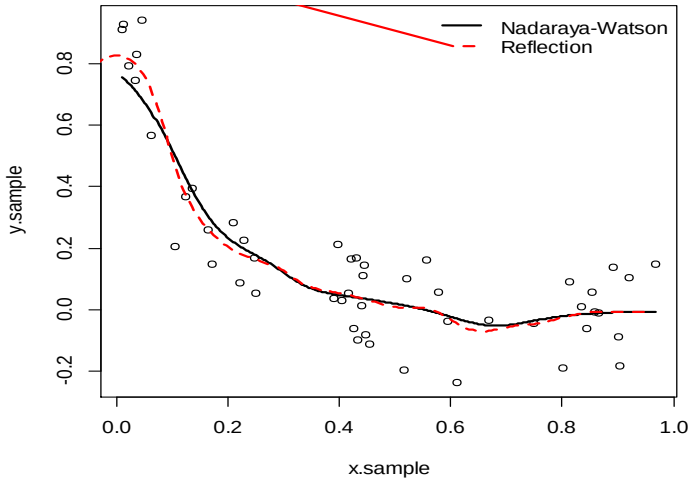
In this section the Nadaraya- Watson and data-reflection technique were compared. To achieve this data was simulated artificially and regression curves were fitted to reveal how best each of the techniques could capture the data. The bandwidths were obtained using the cross-validation technique whose merits had been discussed under the literature review. The simulation was done for varied sample sizes ranging from $n = 50$ to $n = 500$, for two of the models chosen to facilitate this comparison. The quadratic and the exponential models were chosen because they reveal the boundary biases clearly. The resulting graphs are presented in Fig. 4.3 for the exponential model and Fig. 4.4 for the quadratic model.



(a)



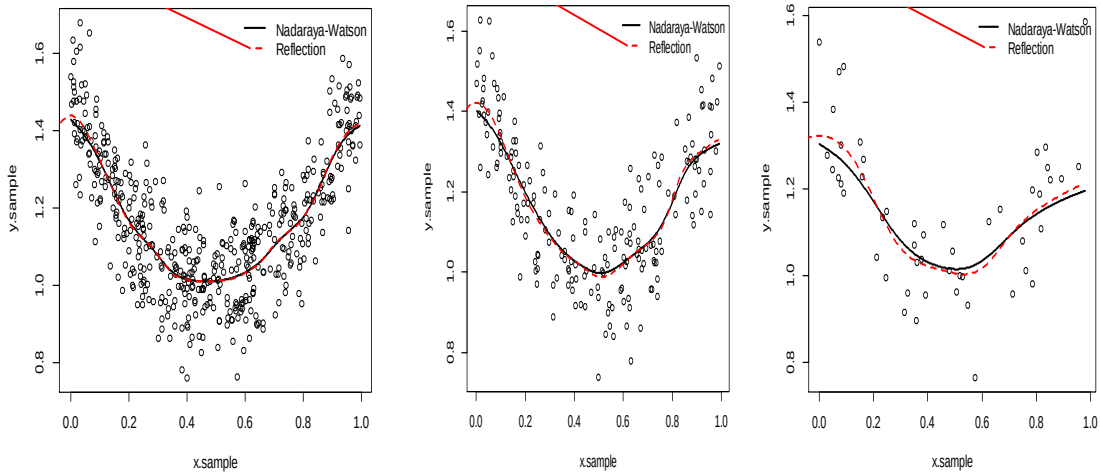
(b)



(c)

Figure 4.3: Comparing Nadaraya-Watson with Reflection estimator in regression estimation (Exponential Model with varying sample sizes)

Fig. 4.3 show the graphs of the exponential model for the same estimators of Nadaraya-Watson and data-reflection technique, but with varying sample sizes of (a) n = 500, (b) n = 200 and (c) n = 50 respectively. Fig 4.4 has similar sample sizes for the respective parts for the quadratic model.



(a) (b) (c)

Figure 4.4: Comparing Nadaraya-Watson with Reflection estimator in regression estimation (Quadratic Model with varying sample sizes)

4.4 Average relative biases of the estimators

This section gives an empirical study that facilitates comparison between the two famous approaches in survey sampling- the Design-based and the Model-based approaches. Six models given in Table 4.2 have been used for this purpose. In addition to this comparison was made between the three model-based nonparametric regression techniques. To achieve this, simulations were performed both for the response variable Y and the corresponding auxiliary variable X for a populations of size N=2000 from where

2000 simple random samples of size n=500 were drawn and used for estimation. In each case estimates of population totals were obtained using the designed-based Horvitz-Thompson estimator, $\hat{T}_{HT}$, the other three model-based estimators, that is, the ratio estimator, $\hat{T}_R$, nonparametric regression estimator due to (Dorfman, 1992), $\hat{T}_{np}$, and the nonparametric regression estimator proposed in this study, $\hat{T}_{npr}$. The average relative biases of the finite population totals got using the three estimation techniques were obtained using the relation: $\left(\left[\sum_{i=1}^{2000}\hat{T}_i \Big/ 2000\right] - T\right)/T$ where T is the actual population total and $\hat{T}_i$ is one of the estimators of the population total computed from the i^{th} sample. A cross validation data generated bandwidth was used in the simulation. Table 4.1 gives the summary of these results.

Table 4.1: Summary of respective estimators and their average relative biases for population totals

MODEL	$\hat{T}_{npr}$	$\hat{T}_{np}$	$\hat{T}_{HT}$	$\hat{T}_R$
LINEAR	0.09731943	-0.06678663	0.6766549	-0.05929314
QUADRATIC	-0.1741573	-2.458993	-0.8319519	0.4905587
SINE	-0.002352638	-1.150008	-0.9798779	-1.39459
EXPONENTIAL	-0.1586322	-2.716807	-0.269894	-0.2317445
JUMP	0.4267462	-0.862683	0.2380192	1.351407
BUMP	-0.1310614	1.265547	-0.1032644	0.2410446

It can be noticed that some of the values of the average relative biases are negative while others are positive representing underestimation or overestimation respectively. It can also be seen that for the data simulated most of the estimates obtained using the estimator due to (Dorfman, 1992) and those of the ratio estimator had slightly larger

biases in most of the models. The designed-based Horvitz-Thompson estimator had biases that were fairly smaller than ratio estimator in three of the models but leaves the proposed standing out as the best among these estimators in overall performance.

4.5 MSE (AMSE) of the different population total estimators studied

In this section the study focuses on assessing the performance of the proposed estimators. Data was simulated to compare the standard kernel estimator and the modified kernel estimator. To make this comparison possible the measures for the MSEs were computed for the six models whose details are given in Table 4.2. These distributions have been chosen because they are often applicable in reality especially in social and behavioural sciences. Bumps and the jumps are often useful during events such as bio-surveillance where there may be an unusually high rate of occurrence of certain events like disease outbreaks or rainfall within certain period in a given region of a country. Time taken between two successive breakdowns of a machine after repair is an example that can constitute an exponential distribution, while sine distributions are useful in phenomena whose occurrences are periodic. It is for such useful applications of the distributions that motivated us to simulate data using them. The other functions were useful in facilitating comparison of the estimators. The summary of the results are given in Table 4.3

Table 4.2 Equations of models simulated

Model	Equation
Linear	$1+2(x-0.5)+e \sim N(0,1)$
Quadratic	$1+2(x-0.5)^2+e \sim N(0,1)$
Jump	$1+2(x-0.5)I_{x\leq 0.65}+0.65I_{x>0.65}+e \sim N(0,1)$
Sine	$2+\sin(2\pi x)+e \sim N(0,1)$
Exponential	$\exp(-8x)+e \sim N(0,1)$
Bump	$1+2(x-0.5)+\exp\left(-200(x-0.5)^2\right)+e \sim N(0,1)$

Table 4.3 Summary of the results for the unconditional MSE (Obtained from 2000 iterations and sample sizes of n=500)

MODEL	$\hat{T}_{npr}$	$\hat{T}_{np}$	$\hat{T}_{HT}$	$\hat{T}_{R}$
LINEAR	10.09797	12.39128	332.0872	8.85685
QUADRATIC	10.74908	17.59172	29.24937	8338.517
SINE	11.40551	32.64659	505.6168	10154.5
EXPONENTIAL	10.71374	19.90012	54.49664	1035.598
JUMP	11.1304	14.55378	22.88779	13395.2
BUMP	11.28804	42.52958	396.9592	92.61349

It was noted earlier on that it is not adequate to look at the bias term without considering the variance of an estimator. MSE is the measure of accuracy that puts the two properties into perspective. From the values presented in Table 4.3, the designed-based Horvitz-Thompson estimator and the model-based ratio estimators have large values. This scenario reveals that for the data obtained from the models considered, nonparametric regression estimators are better. The two estimators $\hat{T}_{npr}$ and $\hat{T}_{np}$ generally have smaller values. Comparing the two, shows that $\hat{T}_{npr}$ has even smaller values than $\hat{T}_{np}$.

4.6 Unconditional 95% C.I for the respective population total estimators

The 95% confidence interval of each of the estimators was also computed using the formula given by; $T = \hat{T} \pm Z_{\alpha/2}\sqrt{Var(\hat{T})}$ and the interval length is therefore the difference between the upper limit and the lower limit. The results are presented in Table 4.4.

Table 4.4 Summary results for the unconditional confidence interval lengths

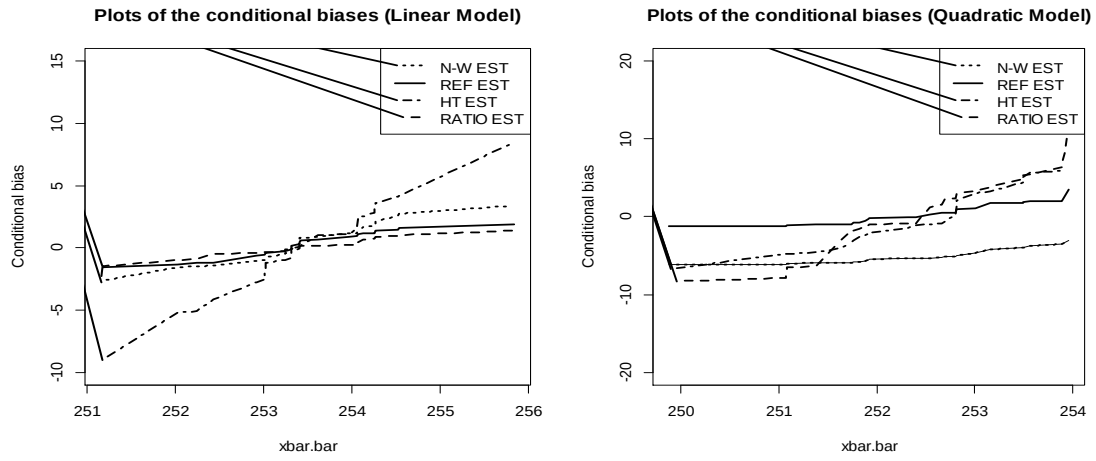
MODEL	$\hat{T}_{npr}$	$\hat{T}_{np}$	$\hat{T}_{HT}$	$\hat{T}_{R}$
LINEAR	10.70248	11.51506	60.75419	10.4797
QUADRATIC	10.67005	11.09624	18.99733	72.39773
SINE	10.77354	19.22728	75.80554	185.6727
EXPONENTIAL	10.47873	12.36678	25.15132	30.85746
JUMP	11.18108	12.98228	16.63127	97.24218
BUMP	11.12733	20.55477	65.69977	31.66942

From Table 4.4, the values of the estimators $\hat{T}_{npr}$ and $\hat{T}_{np}$ reveal shorter confidence lengths compared to $\hat{T}_{HT}$ and $\hat{T}_{R}$. This is also an indication that the nonparametric regression estimators dominate the other estimators. As well, it is clear that $\hat{T}_{npr}$ has the smallest figures overall except in the linear model where $\hat{T}_{R}$ has the least value. In other words $\hat{T}_{R}$ would only be better for the linear model while being poor in all others.

4.7 Conditional performance of the respective population total estimators

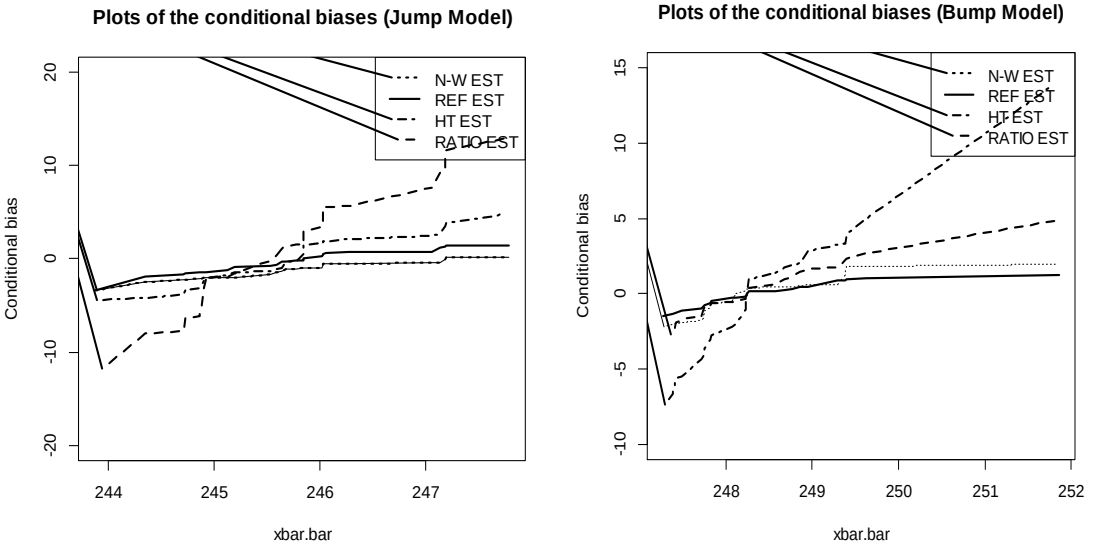
To study the conditional performance of the estimators, the sample means $\bar{x}_i's$ were calculated from 10,000 samples taken from the population already generated and ranked in ascending order while maintaining the corresponding estimates, $\hat{T}_i's$, of the finite population totals. Fifty groups of 200 samples each were then obtained as per the new order of the rankings. From each of these groups the sample means of the auxiliary variable were averaged to give, $\bar{\bar{x}}_j$, the mean of the sample means of the j^{th} group (j= 1, 2, …, 50). The corresponding population totals i.e $\hat{\bar{T}}_j's$ for the various population

estimators studied were also computed and used to calculate the respective conditional biases for the models given. The resulting graphs have been plotted in the Fig. 4.5-4.7.



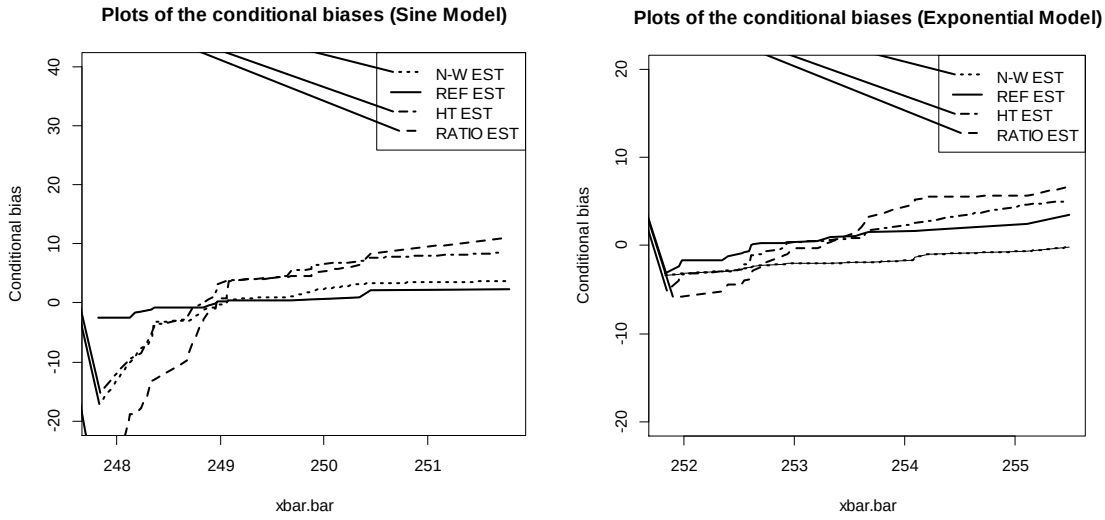
(a) (b)

Figure 4.5: Comparison of conditional bias for the respective finite population total estimators (linear & Quadratic models)



(a) (b)

Figure 4.6: Comparison of conditional bias for the respective finite population total estimators (Jump & Bump models)



(a) (b)

Figure 4.7: Comparison of conditional bias for the respective finite population total estimators (Sine & Exponential models)

It can be noted that from the results obtained from all the models simulated the reflection estimator dominates the rest of the estimators considered in that it is almost conditionally unbiased except for the linear model in Fig. 4.5 part (a) where the ratio estimator is better. In this model the proposed estimator comes close at the second position after it.

CHAPTER FIVE

RESULTS, DISCUSSION AND RECOMMENDATION

5.1 RESULTS AND DISCUSSION

From the onset of this research, it was noted that estimators that utilize the standard kernel estimators such as that due to (Dorfman, 1992) suffer from the boundary problem. It was therefore inevitable that a better estimator ought to be sought. In particular, modification of the Nadaraya-Watson kernel estimator was proposed with a view to alleviating the problem.

The first objective set out was that of identifying the appropriate kernel for a nonparametric regression in the context of model-based estimation. The study revealed that quite a number of them have the suitable features such as that of assignment of weights so that the farthest point is given the least while that closest to the central point of the window of the kernel function receives the most. These kernel functions included the Gaussian, Epanechnikov, bi-weight, and the tri-weight among others. There is the so-called "naive" kernel function also variously referred to as the uniform or rectangular kernel function which assigns weights of ½ uniformly within the window. The constant assignment of the weights is not a pleasant thing as it is believed that the points closer to a given observation have more information than those that are farther away. Also the density function, and in the case of this study the regression curve is not smooth because of the jumps over the short intervals of the window while being constant elsewhere. This can be seen in Fig. 4.1 part (c). These jumps often end up affecting the smoothness of the estimate. Also noted in previous study by (Wand & Jones, 1995) is the triangular function which (Avery, 2010) reports to be lacking the smoothness property. On such grounds these kernels were dropped.

As for the others stated they have almost similar characteristics but Epanechnikov is the optimal kernel. This function posses smooth properties but has discontinuous first derivative thus making the Gaussian function to be the best substitute.

Secondly the next objective of this study was to propose a nonparametric regression estimator that uses the modified kernel estimator and the identified kernel within the model-based approach. This study developed a robust estimator of finite population total whose kernel estimator was modified using data reflection technique and as evidenced from the analysis of the relative biases presented in Table 4.1, it was possible to significantly reduce the boundary bias. The relative biases indicate that the proposed estimator is superior to the other estimators in all the models used except for the linear model where the ratio estimator dominated. But though so, it is known in literature that the ratio estimator is the Best Linear Unbiased Predictor and therefore is expected to perform better in that case. It is worth noting that even though the proposed estimator is not the best under the linear model, the results obtained are still satisfactory. The graphs of the conditional biases in Fig. 4.5-4.7 also indicate that the proposed estimator dominates the other estimators. The graphs show that while the other estimators have larger conditional biases, the proposed estimator is almost conditionally unbiased. This good performance of the reflection estimator was also evident with the comparative regression graphs given in Fig. 4.2 and those given in Fig. 4.3-4.4 where the boundary correction reveals itself.

The proposed technique can therefore be recommended for removal of the boundary effect in regression estimation of finite population total. The tighter confidence interval lengths reported about this estimator in Table 4.4 is an indication of how it is better in terms of efficiency and reliability.

To enable the investigation on the asymptotic properties to be carried out simulation based on the theoretical results derived in the study was done. The smaller values of the *MSE*s reported for the proposed finite population total estimator, T_{npr}, in Table 4.3

proved that this estimator is far much better than the rest of the estimators compared across the models used. From the simulated data it was found out that change in the sample size does not have any impact on the bias.

From the nonparametric point of view, it can therefore, be concluded that the proposed estimator performs better than the one of (Dorfman, 1992) and can therefore be recommended for bias correction at the boundary. This adds knowledge to the already existing techniques in regression estimation.

5.2 RECOMMENDATION

From this study it would be recommended that a suitable bandwidth that strikes a balance between the bias and the variance should be found for optimal results. As noted from this study and previous ones, this can be achieved using various bandwidth selectors. Like many other researchers the data generated cross-validation approach was used, but though so, this remains an area that requires further research because currently no selector can be used universally. The graphs obtained using varied bandwidths in Fig. 5.1 can be compared. The smaller bandwidth notably reduced the bias to a certain extent but a further reduction on this bandwidth came with a cost of increased variance.

REFERENCES

Alberts, T., & Karunamuni, J. R. (2007). Boundary correction methods in Kernel density estimation. Presentation slides . Retrieved from www.math.utah.edu/~alberts/talks/KernelEstimation.pdf

Avery, M. (2010). Literature Review for Local Polynomial Regression. Unpublished manuscript. Retrieved from www4.ncsu.edu/~mravery/AveryReview2.pdf

Bowman, A. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika, 71*(2), 353-360.

Brakel, J., & Bethlehem, J. (2008). *Model- based Estimation for Official Statistics.* Voorburg/Heerlen: Statistics Netherlands. Retrieved from http://www.cbs.nl

Breidt, B. F., & Opsomer, J. D. (2000, August). Local Polynomial regression in Survey Sampling. *Annals of Statistics, 28*(4), 1026-1053.

Breidt, F. J., & Opsomer, J. (2009). Nonparametric and Semiparametric Estimation in Complex Surveys. (D. Pfeffermann, & C. R. Rao, Eds.) *Handbooks of Statistics, 29B,* 103-121.

Brewer. (1995). Combining design-based and model-based inference. Chapter 30 in Business survey methods. New York: John Wiley.

Brewer. (2002). *Combined survey sampling inference: weighing Basu's Elephants.* London: Arnold a member of the Hodder Headline Group.

Chambers, R. (2011). *Which sample survey strategy? A Review of three different approaches.* Working Paper. Centre for Statistical and Survey Methodology, University of Wollongong.

Chambers, R. L., & Danstun. (1986). Estimating Distribution Functionsfrom Survey data. *Biometrika, 73*(3), 597-604.

Chambers, R. L., Dorfman, A. H., & Wehrly, T. E. (1993). Bias Robust estimation in finite populations using nonparametric calibration. *Journal of American Statistics Association, 88*(421), 226–277.

Chandran, K. P., & Prajneshu. (2004). *Computation of Growth rates in Agriculture: Nonparametric Regression Approach.* New Delhi: Indian Agricultural Statistics Research Institute.

Chaudhuri, A., & Stenger, H. (2005). *Survey Sampling: Theory and Methods. 2nd ed.* (2nd ed.). New York: Chapman and Hall/CRC Taylor and Francis group.

Cline, D. B., & Hart, J. D. (1991). Kernel Estimation of Densities of Discontinuous Derivatives. *Statistics, 22*(1), 69-84.

Cochran, W. G. (1977). *Sampling Techniques* (3rd ed.). New York: John Wiley.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2001). *Introduction to Algorithms* (2nd ed.). Massachusetts London, England: The MIT Press Cambridge.

Cornfield. (1944). On samples from finite populations. *Journal of the American Statistics Association, 39*(226), 236-239.

Cowling, A., & Hall, P. (1996). On Pseudodata Methods for Removing Boundary Effects in Kernel Density Estimation. *Journal of the Royal Statistical Society ser. B*, 551-563.

Cox, B. G. (1995). *Business survey methods.* New York : John Wiley.

DiNardo, J., & Tobias, J. L. (2001). Nonparametric Density and Regression Estimation. *Journal of Economic Perspectives, 15*(4), 11-28.

Dorfman, A. H. (1992). Nonparametric Regression for Estimating Totals in Finite Populations. In proceedings of the section on Survey Research Methods. *American Statistics Association*, 622-625.

Dorfman, A. H., & Hall, P. (1993). Estimators of the finite population distribution function using nonparametric regression. *Annals of Statistics*, 1452-1475.

Epanechnikov, V. (1969). Nonparametric estimation of a multidimensional probability density. *Teoriya Veroyatnostej i Ee Primeneniya, 14*(1), 156–162.

Fan, J., & Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. 20,. *Annals of Statistics, 20*, 2008-2036.

Faraway, J. (2006). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models.* New York: Chapman & Hall/CRC Taylor & Francis Group, LLC.

Gámiz, M. L., Kulasekera, K. B., Limnios, N., & Lindqvist, B. H. (2011). *Applied Nonparametric Statistics in Reliability.* London: Springer verlag.

Gasser, T., & Müller, H. G. (1979). Kernel Estimation of Regression Functions. In Smoothing Techniques for Curve Estimation. 23-68. (T. Gasser, & M. Rosenblatt, Eds.) Heidelberg: Springer-Verlag.

Gasser, T., Müller, H. G., & Mammitzsch, V. (1985). Kernels for Nonparametric Curve Estimation. *Journal of the Royal Statistical Society Ser. B, 47*, 238-252.

Godambe, V. (1955). A Unified theory of sampling from finite populations. *Journal of the Royal statistical society. Series B, 28*, 310-328.

Hansen, B. E. (2009). Lecture Notes on Nonparametrics.University of Wisconsin.

Hansen, M. H., Madow, W. G., & Tepping, B. J. (1983). An evaluation of model dependent and probability sampling inferences in sample surveys. *Journal of American Statistical Association, 78*(384), 776-793.

Härdle, W. (1990). *Applied Nonparametric Regression Analysis.* Cambridge: Cambridge University Press.

Härdle, W. (1994). *Applied Nonparametric Regression Analysis.* Cambridge: Cambridge University Press.

Härdle, W., Müller, M., Sperlich, S., & Werwatz, A. (2005). *Nonparametric and Semiparametric Models-An Introduction.* . Berlin Heidelberg: Springer verlag.

Hastie, T. J., & Loader, C. R. (1993). Local regression: Automatic kernel carpentry (with discussion). *Statistical Science, 8*, 120-143.

Hedayat, A. S., & Sinha, B. K. (1991). *Design and Inference in finite sampling* . New York: John Wiley.

Hirukawa, M., & Sakudo, M. (2015). Family of the Generalized Gamma Kernels: A Generator of Asymmetric Kernels for Nonnegative Data. Submitted paper. *Journal of Nonparametric Statistics, 27*(1), 41-63.

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association, 47*(260), 663-685.

Irizarry, R. A., & Bravo, H. C. (2010). Smoothing. Lecture 7 notes. Retrieved from www.cbcb.umd.edu/~hcorrada/practicalML/pdf/lectures/smoothing.pdf

Jann, B. (2007). Univariate kernel density estimation. A paper downloaded from internet.

Jones, M. C. (1993). Simple Boundary Correction for Kernel Density Estimation. *Statistics and Computing, 3*(3), 135-146.

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

Karunamuni, R., & Alberts, T. (2004). On the boundary correction in Kernel density estimation. *A paper presented in the Fifth Biennial IISA International Conference on Statistics, Probability and Related Areas held at the University of Georgia.* Athens, Georgia.

Kish, L. (1965). *Survey Sampling.* New York: John Wiley and Sons.

Kott, P. S. (2005). Randomization-assisted model-based survey sampling. *Journal of Statistical Planning and Inference, 129*(1), 263-277.

Kyung-Joon, C., & Schucany, W. R. (1998). Nonparametric kernel regression estimation near endpoints. *Journal of Statistical Planning and Inference, 66*(2), 289-304.

Langat, R. C., Odhiambo, R. O., & Odongo, L. (2007). Model-Assisted estimation of Finite population Total in Stratified random sampling. Unpublished MSc Thesis. Nairobi: Kenyatta University.

László, G., A, K., Kohler, M., & Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression.* New York: Springer-Verlag.

Li, Q., & Racine, J. (2004). Cross-Validated Local Linear Nonparametric Regression. *Statistica Sinica, 14*, 485-512.

Loader, C. (1999). *Local Regression and Likelihood.* New York: Springer-Verlag.

Loader, C. (2004). *Smoothing: Local Regression Techniques.* Retrieved from Papers / Humboldt-Universität Berlin, Center for Applied Statistics and Economics (CASE), No. 2004,12: http://hdl.handle.net/10419/22186

Lohr, S. L. (2010). *Sampling: Design and analysis* (2nd ed.). New York: Duxbury Press.

Mack, Y. P., Quang, P., & Zhang, S. (1999). Kernel estimation in transect sampling without the shoulder condition. *Communications in Statistics - Theory and Methods, 28*(10), 2277–2296.

Malec, P., & Melanie, S. (2012). *Nonparametric Kernel Density near the Boundary.* Humboldt-Universit¨at zu Berlin: Institute for Statistics and Econometrics.

Manzoor, M. A., Akbar, A., & Ullah, M. A. (2013). Performance of Nonparametric Regression Estimation with Diverse Covariates. *Pakistan Journal of Social Sciences (PJSS), 33*(1), 77-85.

Marron, J. S., & Ruppert, D. (1994). Transformations to reduce boundary bias in kernel density estimation. *Journal of the Royal Statistical Society. Series B, 56*, 653–671.

Müller, H. G. (1991). Smooth Optimum Kernel Estimators Near Endpoints. *Biometrika, 78*(3), 521-530.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and Application, 9*(1), 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. *Journal of the Royal Statistical Society*, *97*(4), 558–622.

Odhiambo, R., & Mwalili, S. (2000). Nonparametric regression for Finite Population Estimation. *East African Journal of Statistics, II*(part 2), 107-118.

Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Statistics, 33*(3), 1065– 1076.

Prasad, N. G., & Subhash, R. L. (2011). Improved prediction in finite population sampling using convex combination of parametric and nonparametric models. *Sankhya B, 72*(2), 189-201.

R Core Team. (2014). *R: A language and environment for statistical computing.* Retrieved from R Foundation for Statistical Computing, Vienna, Austria: http://www.R-project.org/

Racine, J. S. (2008). Nonparametric Econometrics. *A Primer. Foundations and Trends in Econometrics, 3*(1), 1–88.

Rao, J. N., Kovar, J. G., & Mantel, H. J. (1990). On Estimating Distribution Functions and Survey Data using Auxiliary Information. *Biometrika, 77*(2), 365-375.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Statistics, 27*(3), 832–837.

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

Ruppert, D., Sheather, S. J., & Wand, M. P. (1995). An Effective Bandwidth Selector for Local Least Squares Regression. *Journal of the American Statistical Association, 90*(432), 1257-1270.

Särndal, C. E., Swesson, B., & Wretman, J. N. (1992). *Model- Assisted Survey sampling* . New York: Springer Verlag.

Schuster, E. (1985). Incorporating Support Constraints into Nonparametric Estimators of Densities. *Communications in Statistics- Theory and Methods, 14*(5), 1123–1136.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* London: Chapman and Hall.

Takezawa, K. (2006). *Introduction to Nonparametric Regression.* Hoboken, New Jersey: John Wiley.

Todd, P. E. (2014). Lecture Notes on Nonparametric Density and Regression Estimation. Retrieved from www.athena.sas.upenn.edu/petra/class721/panelnotes2.pdf

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation.* New York: Springer Science+Business Media, LLC.

Valliant, R., Dorfman, A. H., & Royall, R. (2000). *Finite population sampling and inference: A prediction approach.* . New York: John Wiley.

Van der Vaart, A. W. (1998). *Asymptotic statistics.* Cambridge: Cambridge university press.

Wand, & Jones. (1995). *Kernel Smoothing,* . New York: Chapman and Hall.

Wand, M. P., Marron, J. S., & Ruppert, D. (1991). Transformations in Density Estimation (with discussion). *Journal of the American Statistical Association, 86*(414), 343-361.

Wang, S., & Dorfman, A. H. (1996). A New Estimator for the Finite Population Distribution Function. *Biometrika, 83*(3), 639-652.

Watson, G. S. (1964). Smooth regression analysis. *Sankhya, Series A*, 359-372.

Zhang, S., & Karunamuni, R. J. (2000). On Nonparametric Density Estimation at the Boundary. *Nonparametric Statistics, 12*(2), 197-221.

Zucchini, W. (2003). Applied Smoothing Techniques Part 1: Kernel Density Estimation. Retrieved from www.staff.ustc.edu.cn/~zwp/teach/Math-Stat/kernel.pdf

APPENDICES

Appendix 1: Graph on boundary correction

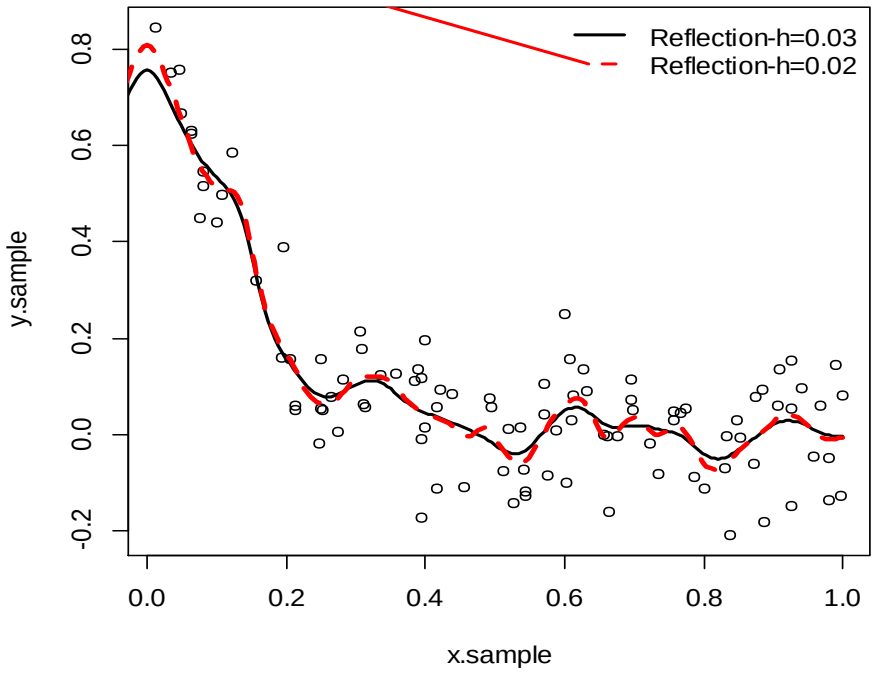
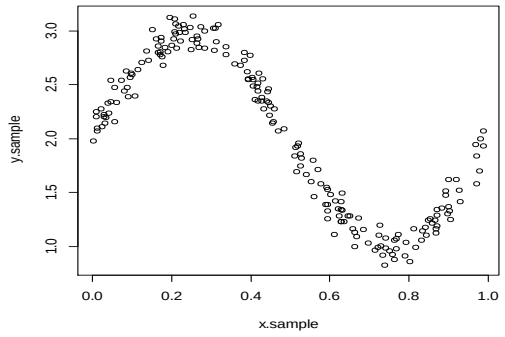
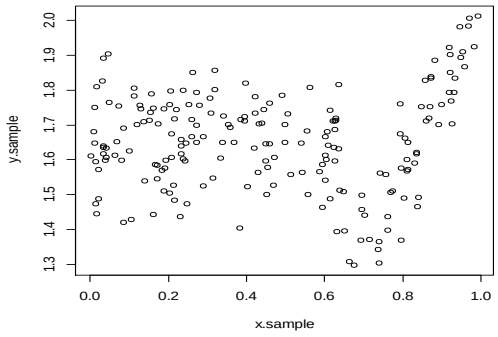


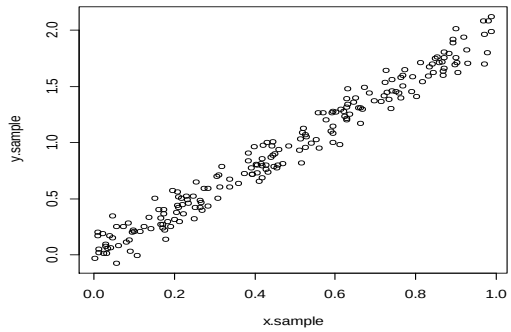
Figure 5.1: Reflection estimator in regression estimation (Exponential Model with sample size, n=100 and two different bandwidths)



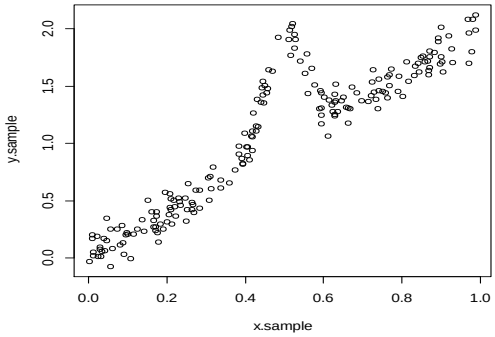
(a)



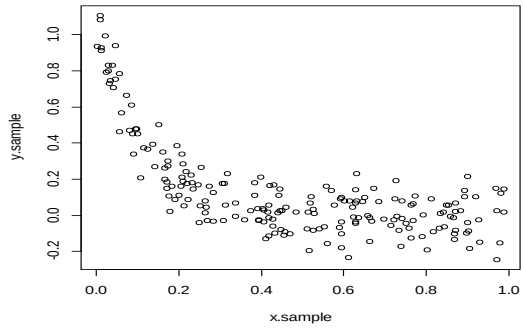
(b)



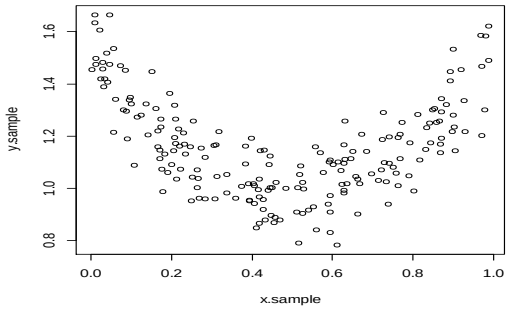
(c)



(d)



(e)



(f)

Figure 5.2: Scatter plots for the respective models used in simulation (a) Sine (b) Jump (c) Linear (d)Bump (e) exponential and(f)Quadratic

Appendix 2: R-codes for various Graphs and results tabulated

R-codes for Fig.3.3

```
library(KernSmooth)
 ##simulating data
 set.seed(12)
 x<-c(runif(100,1,2))
 e<-c(rnorm(100,0,0.5))
 y<-c(10-x^3+e)
 plot(x,y)
 nonpar.reg.nw1 <- ksmooth(x, y, kernel="normal", bandwidth=0.25)
 lines(nonpar.reg.nw1,col=4)
```

R-codes for Fig.3.4-3.5

```
set.seed(12)
 x<-c(runif(100,1,2))
 e<-c(rnorm(100,0,0.5))
 y<-c(10-x^3+e)
 plot(x,y)
 nonpar.reg.nw1 <- ksmooth(x, y, kernel="normal", bandwidth=0.25)
 nonpar.reg.nw2 <- ksmooth(x, y, kernel="normal", bandwidth=0.15)
 nonpar.reg.nw3 <- ksmooth(x, y, kernel="normal", bandwidth=0.05)
 nonpar.reg.nw4 <- ksmooth(x, y, kernel="normal", bandwidth=0.02)
 lines(nonpar.reg.nw1,col=1)
 lines(nonpar.reg.nw2,col=3,lwd=2)
 lines(nonpar.reg.nw3,col=2,lwd=2,lty=2)
 lines(nonpar.reg.nw4,col=4,lwd=2,lty=2)
 legend("topright", c("h=0.25",
"h=0.05","h=0.15","h=0.02"),
 col=1:4,lwd=2:2,lty=1:2, bty="n")
**********************************************************************
```

R-codes for comparison of various kernel functions in Fig 4.1

```
set.seed(10)
x<-rnorm(100,20)
hist(x)
hist(x, prob=T,
main="Histogram With Fitted Density Curve", bw="nrd")
density (x,bw="nrd", kernel="gaussian")
density (x,bw="nrd", kernel="epanechnikov")
density (x,bw="nrd", kernel="rectangular")
density (x,bw="nrd", kernel="triangular")
density (x,bw="nrd", kernel="biweight")

lines(density(x, bw="nrd",kernel="gaussian"),
      col="black",
      lwd=1)
lines(density(x, bw="nrd",kernel="epanechnikov"),
      col="blue",
      lwd=1)
lines(density(x, bw="nrd",kernel="rectangular"),
      col="dark green",
      lwd=1)
lines(density(x, bw="nrd",kernel="triangular"),
      col="maroon",
      lwd=1)
lines(density(x, bw="nrd",kernel="biweight"),
      col="red",
      lwd=1)
```

****

*

R-codes for comparison of graphs of NW and reflection regression estimators in Fig. 4.2-4.4

```
library(MASS)
library(sm)
library(KernSmooth)
set.seed(500)
```

```
E<-rnorm(1000,mean=0,sd=.1)    #the random error
X<-runif(1000, min = 0, max = 1)  #the explanatory variable
mx=1-X+exp(-200*(X-0.5)^2)
mx=2+sin(2*pi*X)#Sine
Ix <- function(x,h) 1*(x >= h)
mx = 1 + 2*(X-.5)*Ix(X,.65) +0.65*(1-Ix(X,.65))
mx=2+sin(2*pi*X)
mx=1+2*(X-0.5)
mx=1+2*(X-0.5)+exp(-200*(X-0.5)^2)
mx=exp(-8*X)
mx=1+2*((X-0.5)^2)

 #the regression function
Y <- mx+E

sindex=sample(1:1000,200)
x.sample=X[sindex]
xreflect=c(x.sample,-x.sample)
xreflect
xnosample=setdiff(X,x.sample)
y.sample=Y[sindex]
yreflect=c(y.sample,y.sample)
yreflect
ynosample=setdiff(Y,y.sample)
data1=data.frame(xnosample)

plot(x.sample,y.sample)
title(main="Optimum cross validation bandwidth, n=200,Jump")
#title(main="Bandwidth=0.02, n=00")
hm1 <- hcv(x.sample,y.sample)
hm1
#H1=(ucv(x.sample,nb=1000,min(x.sample),max(x.sample)))
#H1
hm2 <- hcv(xreflect,yreflect)
hm2

#NADARAYA
fit1 <- locpoly(x.sample, y.sample,degree=0, kernel = "normal",bandwidth = hm1)
lines(fit1,col=1,lwd=2)
```

```
#fit2 <- locpoly(x.sample, y.sample, degree=1, kernel = "normal",bandwidth = hm)
#REFLECTION
fit3 <- locpoly(xreflect, yreflect, degree=0, kernel = "normal",bandwidth = hm2)
lines(fit3,col="red",lty=2,lwd=2)
#HT
fit4=(length(Y)/length(y.sample))*y.sample
legend("topright", c("Nadaraya-Watson",
"Reflection"),
 col=1:2,lwd=2:2,lty=1:2, bty="n")
**********************************************************************************
*******
```

R-codes for results tabulated

```
library(MASS)
require(sm)
E<-rnorm(1000,mean=0,sd=.1)    #the random error
X<-runif(1000, min = 0, max = 1)  #the explanatory variable
mx=1-X+exp(-200*(X-0.5)^2)
mx=2+sin(2*pi*X)#Sine
Ix <- function(x,h) 1*(x >= h)
mx = 1 + 2*(X-.5)*Ix(X,.65) +0.65*(1-Ix(X,.65))
mx=2+sin(2*pi*X)
mx=1+2*(X-0.5)
mx=1+2*(X-0.5)+exp(-200*(X-0.5)^2)
mx=exp(-8*X)
mx=1+2*((X-0.5)^2)

  #the regression function
Y <- mx+E

mf=mf2=TT=0
j=0
BIASR=BIASND=BIASHT=BIASRATIO=TOTALS=0
  MSEND=0
  MSER=MSEHT=MSERATIO=0
TTND=TTR=TTHT=TTLL=TRATIO=0
```

```
 means=0
 varND=varR=varHT=varRATIO=numeric()

while(j<=10000)
{
sindex=sample(1:1000,500)#Selection of the sample
x.sample=X[sindex]
xreflect=c(x.sample,-x.sample)
xreflect
xnosample=setdiff(X,x.sample)
y.sample=Y[sindex]
yreflect=c(y.sample,y.sample)
yreflect
ynosample=setdiff(Y,y.sample)
data1=data.frame(xnosample)
#H<-hcv(xnosample,ynosample)
H3=(ucv(xreflect,nb=1000,min(x.sample),max(x.sample)))

H1=(ucv(x.sample,nb=1000,min(x.sample),max(x.sample)))
#H=(ucv(x.sample,nb=1000,min(x.sample),max(x.sample)))*100
#H2=(ucv(xnosample,nb=1000,min(xnosample),max(xnosample)))


#H2=(ucv(xreflect,nb=1000,min(xreflect),max(xreflect),tol=0.01))*100

#H2=(ucv(xreflect,nb=1000))*100

nad1=ksmooth(x.sample,y.sample,kernel="normal",bandwidth =0.2322215,x.points=xnosample)
nad2=ksmooth(xreflect,yreflect,kernel="normal",bandwidth =0.03205294,x.points=xnosample)
nad3=loess(y.sample~x.sample,span=.5)
#nad4 <- locpoly(x.sample, y.sample, bandwidth = 0.25)
#nad4<-locpoly(x.sample,y.sample,bandwidth=H)
#model1=npreg(xdat=x.sample,ydat=y.sample,bws=H,regtype="ll")
#pred=predict(model1,newdata=data.frame(x.sample),exdat=xnosample)

#computing variance of ratio
Vr=0
for(i in 1:length(y.sample))
{
```

```
f=length(y.sample)/length(Y)
x_bar=mean(x.sample)
r=mean(y.sample)/mean(x.sample)
Vr[i]=(y.sample[i]-r*x.sample[i])^2
}
VAr=((1-f)/(length(y.sample)*(length(y.sample)-1))*x_bar^2)*sum(Vr)

summary(nad3)
mf=nad1$y
mf2=nad2$y
mf1=predict(nad3,data=data1,se=TRUE)
   TTND[j]=sum(c(y.sample,as.vector(mf)))
   TTR[j]=sum(c(y.sample,as.vector(mf2)))
   TTHT[j]=sum((length(Y)/length(y.sample))*y.sample)
   #TTLL[j]=sum(c(y.sample,as.vector(pred)))
   TRATIO[j]=(sum(y.sample)/sum(x.sample))*sum(X)
   varND[j]=var(c(y.sample)) +var(as.vector(mf))
   varR[j]= var(c(y.sample)) +var(as.vector(mf2))
   varHT[j]= var((length(Y)/length(y.sample))*y.sample)
   varRATIO[j]=var((y.sample)/(x.sample))

   BIASND[j]=(TTND[j]-sum(Y))
   BIASR[j]=(TTR[j]- sum(Y))
   BIASHT[j]=(TTHT[j]- sum(Y))
   BIASRATIO[j]=(TRATIO[j]- sum(Y))
   MSEND[j]=varND[j]+(BIASND[j]^2)
   MSER[j]=varR[j]+(BIASR[j]^2)
   MSEHT[j]=varHT[j]+(BIASHT[j]^2)
   MSERATIO[j]=varRATIO[j]+(BIASRATIO[j]^2)
    TOTALS[j]=sum(x.sample)

j=j+1
}
#MSERATIO=var(TRATIO)-(mean(BIASRATIO))^2

Totals=cbind(sum(Y),mean(TTND),mean(TTR),mean(TTHT),mean(TRATIO))
Totals
```

```
uncond=c(mean(BIASND),mean(MSEND),mean(BIASR),mean(MSER),mean(BIASHT),mean(MSEHT),mean(BIAS
RATIO),mean(MSERATIO))
RESULTS<-matrix(uncond,1,8)
colnames(RESULTS)=c("BIASND","MSEND","BIASR","MSER","BIASHT","MSEHT","BIASRATIO","MSERAT
IO")
RESULTS
M=sum(Y)
V=mean(X)

ptotals=c(mean(TOTALS[1:20])-V,mean(TOTALS[21:40])-V,mean(TOTALS[41:60])-V,
mean(TOTALS[61:80])-V,mean(TOTALS[81:100])-V,mean(TOTALS[101:120])-V,
mean(TOTALS[121:140])-V,mean(TOTALS[141:160])-V,mean(TOTALS[161:180])-V,
mean(TOTALS[181:200])-V,mean(TOTALS[201:220])-V,mean(TOTALS[221:240])-V,
mean(TOTALS[241:260])-V,mean(TOTALS[261:280])-V,mean(TOTALS[281:300])-V,
mean(TOTALS[301:320])-V,mean(TOTALS[321:340])-V,mean(TOTALS[341:360])-V,
mean(TOTALS[361:380])-V,mean(TOTALS[381:400])-V,mean(TOTALS[401:420])-V,
mean(TOTALS[421:440])-V,mean(TOTALS[441:460])-V,mean(TOTALS[461:480])-V,
mean(TOTALS[481:500])-V)
ptotals



condRbias=c(mean(TTR[1:20])-M,mean(TTR[21:40])-M,mean(TTR[41:60])-M,
mean(TTR[61:80])-M,mean(TTR[81:100])-M,mean(TTR[101:120])-M,
mean(TTR[121:140])-M,mean(TTR[141:160])-M,mean(TTR[161:180])-M,
mean(TTR[181:200])-M,mean(TTR[201:220])-M,mean(TTR[221:240])-M,
mean(TTR[241:260])-M,mean(TTR[261:280])-M,mean(TTR[281:300])-M,
mean(TTR[301:320])-M,mean(TTR[321:340])-M,mean(TTR[341:360])-M,
mean(TTR[361:380])-M,mean(TTR[381:400])-M,mean(TTR[401:420])-M,
mean(TTR[421:440])-M,mean(TTR[441:460])-M,mean(TTR[461:480])-M,
mean(TTR[481:500])-M)
condRbias



condNDbias=c(mean(TTND[1:20])-M,mean(TTND[21:40])-M,mean(TTND[41:60])-M,
mean(TTND[61:80])-M,mean(TTND[81:100])-M,mean(TTND[101:120])-M,
mean(TTND[121:140])-M,mean(TTND[141:160])-M,mean(TTND[161:180])-M,
mean(TTND[181:200])-M,mean(TTND[201:220])-M,mean(TTND[221:240])-M,
mean(TTND[241:260])-M,mean(TTND[261:280])-M,mean(TTND[281:300])-M,
```

```
mean(TTND[301:320])-M,mean(TTND[321:340])-M,mean(TTND[341:360])-M,
mean(TTND[361:380])-M,mean(TTND[381:400])-M,mean(TTND[401:420])-M,
mean(TTND[421:440])-M,mean(TTND[441:460])-M,mean(TTND[461:480])-M,
mean(TTND[481:500])-M)
condNDbias

condHTbias=c(mean(TTHT[1:20])-M,mean(TTHT[21:40])-M,mean(TTHT[41:60])-M,
mean(TTHT[61:80])-M,mean(TTHT[81:100])-M,mean(TTHT[101:120])-M,
mean(TTHT[121:140])-M,mean(TTHT[141:160])-M,mean(TTHT[161:180])-M,
mean(TTHT[181:200])-M,mean(TTHT[201:220])-M,mean(TTHT[221:240])-M,
mean(TTHT[241:260])-M,mean(TTHT[261:280])-M,mean(TTHT[281:300])-M,
mean(TTHT[301:320])-M,mean(TTHT[321:340])-M,mean(TTHT[341:360])-M,
mean(TTHT[361:380])-M,mean(TTHT[381:400])-M,mean(TTHT[401:420])-M,
mean(TTHT[421:440])-M,mean(TTHT[441:460])-M,mean(TTHT[461:480])-M,
mean(TTHT[481:500])-M)
condHTbias

condTRATIObias=c(mean(TRATIO[1:20])-M,mean(TRATIO[21:40])-M,mean(TRATIO[41:60])-M,
mean(TRATIO[61:80])-M,mean(TRATIO[81:100])-M,mean(TRATIO[101:120])-M,
mean(TRATIO[121:140])-M,mean(TRATIO[141:160])-M,mean(TRATIO[161:180])-M,
mean(TRATIO[181:200])-M,mean(TRATIO[201:220])-M,mean(TRATIO[221:240])-M,
mean(TRATIO[241:260])-M,mean(TRATIO[261:280])-M,mean(TRATIO[281:300])-M,
mean(TRATIO[301:320])-M,mean(TRATIO[321:340])-M,mean(TRATIO[341:360])-M,
mean(TRATIO[361:380])-M,mean(TRATIO[381:400])-M,mean(TRATIO[401:420])-M,
mean(TRATIO[421:440])-M,mean(TRATIO[441:460])-M,mean(TRATIO[461:480])-M,
mean(TRATIO[481:500])-M)
condTRATIObias
```

****

R-codes for conditional bias plots for the various models in Fig. 4.5-4.7

```
#par(mfrow=c(2,3))
plot(sort(ptotals),sort(condNDbias),type='l',xlab='xbar.bar',ylab='Conditional bias',main="Plots of the conditional biases (Linear Model)",ylim=c(-10,15))
lines(sort(ptotals),sort(condNDbias),lty='dotted',xlab='xbar.bar',ylab='Conditional bias',lwd=c(2))
#plot(sort(ptotals),sort(condRbias),ltype='dotted',xlab='xbar.bar',ylab='Conditional bias',main="Plots of the conditional biases(Linear Model)")
CHT=jitter(condHTbias, factor = 25000, amount = NULL)
```

```
RTT=jitter(condTRATIObias, factor = 10000, amount = NULL)
lines(sort(ptotals),sort(condRbias),lty='solid',xlab='xbar.bar',
ylab='Conditional bias',lwd=c(2))

lines(sort(ptotals),sort(condHTbias),lty='dotdash',xlab='xbar.bar',
ylab='Conditional bias',lwd=c(2))
lines(sort(ptotals),sort(condTRATIObias),lty='dashed',xlab='xbar.bar',
ylab='Conditional bias',lwd=c(2))
legend("topright",c("N-W EST","REF EST","HT EST","RATIO
EST"),lty=c("dotted","solid","dotdash","dashed"),lwd=c(2,2,2,2))
NND<-sort(TTND)
REFT<-sort(TTR)
RRRT<-sort(TRATIO)
HOV<-sort(TTHT)
NNDCONFINT<-NND[9500]-NND[500]
REFTCONFINT<-REFT[9500]-REFT[500]
RATIOCONFINT<-RRRT[9500]-RRRT[500]
HTCONFINT<-HOV[9500]-HOV[500]
confs<-cbind(NNDCONFINT,REFTCONFINT,RATIOCONFINT,HTCONFINT)
confs
```