MODEL BASED ROBUST ESTIMATION OF FINITE POPULATION TOTAL USING THE PROCEDURE OF LOCAL LINEAR REGRESSION

KIKECHI CONLET BIKETI

SCHOOL OF MATHEMATICS

COLLEGE OF BIOLOGICAL AND PHYSICAL SCIENCES

UNIVERSITY OF NAIROBI

A thesis submitted in fulfillment for the award of the Degree of Doctor of Philosophy in Mathematical Statistics

NOVEMBER 2019

DECLARATION

Candidate:

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

SIGNATURE……DATE……

KIKECHI CONLET BIKETI

REGISTRATION NUMBER: I80/51212/2016

Supervisors:

We confirm that the work reported in this thesis was carried out by the candidate under our supervision as the University Supervisors.

SIGNATURE……DATE……

PROFESSOR RICHARD ONYINO SIMWA, PhD.

SCHOOL OF MATHEMATICS,

UNIVERSITY OF NAIROBI.

SIGNATURE……DATE……

PROFESSOR GANESH PRASAD POKHARIYAL, PhD., DSc.

SCHOOL OF MATHEMATICS,

UNIVERSITY OF NAIROBI.

ACKNOWLEDGEMENT

Firstly, I convey my deep and sincere appreciation to my supervisors Professor Richard Onyino Simwa and Professor Ganesh Prasad Pokhariyal for your keen interest in this thesis. I'm grateful for your excellent and diligent supervision you demonstrated by guiding me in a professional way during the entire research period. Your continued support, tireless efforts and availability guided me to the very terminal point. Sincere thanks also go to Prof. Jairus Khalagai, Prof. Jamen Were, Prof. Jam Otieno, Prof. Moses Manene, Prof. Bernard Nzimbi, Prof. Stephen Moindi, Prof. Stephen Luketero, Prof. James Okwoyo and Prof. Isaac Kipchirchir, all from the School of Mathematics of the University of Nairobi for your criticism and great support especially at the difficult stages of proposal writing and thesis development.

To all my family members, you have been a great pillar in my life. I humbly thank my wife Evelyn Nelima Kikechi, the love of my life and the mother of my children; Wayne Tolometi Kikechi, Virginia Nakimweyi Kikechi and Ryan Biketi Kikechi. Your relentless prayers and tireless support enabled me to come this far. Thank you. I love you all.

To my dear parents, Papa Charles Kikechi and Mama Christine Kikechi, I sincerely thank you for your great inspiration and encouragement since my early intellectual pursuit. To my brothers and sisters, thank you all for your great support during the entire research period.

I wish to make a special acknowledgement to the technical team led by Prof. Antony Waititu for providing the technical support during data simulation. To the persons whose names do not appear here but contributed in one way or the other to the study, thank you all.

Lastly, I thank the Almighty God who gives life, wisdom, knowledge and ensures all things work together for the good of those who love Him and are called according to His purpose.

DEDICATION

To my beloved wife Evelyn Nelima Kikechi, my sons Wayne Tolometi Kikechi, Ryan Biketi Kikechi and my daughter Virginia Nakimweyi Kikechi. This work is also dedicated to my beloved late grandmother Virginia Nasipwondi Kikechi who washed, cooked and ate my *likhoni* for my social, economic and academic blessings that I'm enjoying now.

ABSTRACT

Nonparametric regression provides an intensive estimation of unknown finite population parameters and is frequently used to explore the association between covariates and responses. This estimation procedure is more flexible and robust than inference based on design probabilities in design based inference or on parametric regression models in model based inference. In this study, model based robust estimators of finite population total are constructed using the procedure of local linear regression. In particular, robustness properties of the derived estimators are investigated and a brief comparison between the performances of the derived estimators and some existing estimators is made in terms of the biases, variances, mean square errors, relative efficiencies, confidence intervals and average lengths of confidence intervals. The study explores the use of adaptive bandwidth to handle sparse data. The local linear procedure is extended to stratified random sampling and to two stage cluster sampling. The local linear procedure is important in the sense that it adapts well to bias problems at boundaries and in regions of high curvature and it does not require smoothness and regularity conditions required by other methods such as the boundary kernels. It has been observed that the local linear regression estimators are generally asymptotically unbiased, efficient and consistent. The results for the biases show that the local linear regression estimators are superior and dominate the Horvitz-Thompson estimator and the Linear regression estimator in all the relationships. The local linear regression estimators also dominate the Dorfman estimator in all the relationships except when the relationship is quadratic. The results for the mean square errors indicate that the local linear regression estimators are more efficient and perform better than the Horvitz-Thompson and Dorfman estimators, regardless of whether the underlying model is correctly specified or misspecified. The local linear regression estimators also outperform the linear regression estimator in all the relationships except when the relationship is linear. With respect to the relative efficiencies, results indicate that the local linear regression estimators are robust and are the most efficient estimators. The confidence intervals generated by the model based local linear method are much shorter than those generated by the design based Horvitz-Thompson method. The results also indicate that the model based approach performs better than the design based approach at 95% coverage rate. Generally, the model based approach outperforms the design based approach regardless of whether the underlying model is correctly specified or not but that effect decreases as the model variance increases.

LIST OF ABBREVIATIONS AND NOTATIONS

LLR	Local Linear Regression
LLRE	Local Linear Regression Estimators
LPR	Local Polynomial Regression
LPRE	Local Polynomial Regression Estimators
GREG	Generalized Regression Estimators
BLUE	Best Linear and Unbiased Estimators
AB	Absolute Bias
MSE	Mean Square Error
MISE	Mean Integrated Square Error
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
RE	Relative Efficiency
CI	Confidence Intervals
ALCI	Average Length of Confidence Intervals
$\bar{T}$	Estimator of the Population Total, T
$\hat{\sigma}^2(x_i)$	Estimate of the Variance of the Variable
$Cov(Y_i, Y_j)$	Covariance between Elements of the Variable

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

1.1 Background information

In many survey problems, auxiliary information is available for all elements of the population of interest. Indeed, use of auxiliary information in estimating parameters of a finite population of study variables is a central problem in sample surveys. One approach to this problem is the super population approach, in which a working model ξ describing the relationship between the auxiliary variable x and the study variable y is assumed. Estimators are sought which have good efficiency if the model is true, but maintain desirable properties like asymptotic design unbiasedness (unbiasedness over repeated sampling from the finite population) and design consistency if the model is false. Typically, the assumed models are linear models, leading to the familiar ratio and regression estimators (Cochran 1977), the Best linear and Unbiased Estimators (BLUE) (Brewer, 1963; Royall, 1970), the Generalized Regression Estimators (GREG) (Cassel et al., 1977). Royall and Herson (1973) suggested that efficiency and robustness could be combined by choosing the estimator of the population total to be optimal under a fairly realistic linear super population model, and choosing the selection procedure to ensure that this estimator was the best linear unbiased estimator under a more general family of polynomial models.

Nonparametric regression for estimating totals in finite populations has been applied by Kuo (1998), Dorfman and Hall (1993) and Kuk (1993). Such estimation is frequently more flexible and robust than inference tied to design probabilities in design based inference or to parametric regression models in model based inference.

Sample survey theory is concerned with methods of sampling from a finite population of N units and then making inferences about finite population quantities on the basis of the sample data. There are two incompatible approaches for making inference from sample to population; the more traditional design based approach, in which the probability structure of the procedure by which the sample s is selected serves as the basis for inference, and the model based or predictive approach, in which a regression model of the response Y on the predictor X is used to predict the non sample $Y's$ and by consequence, their total (Dorfman 1992). The design based estimator of total is the stratified expansion estimator given by

$$\bar{T}_{exp} = \sum_{h} \pi_h^{-1} \sum_{s_h} Y_{hi}$$

where for $h = 1,2, \dots, \pi_h$ are inclusion probabilities of Y_{hi} units in the sample S_h of the h^{th} stratum. If Y is linear in x, that is, $Y_i = \alpha + \beta x_i + \sigma_i e_i \; i = 1,2, \dots, N$, with e_i independent and identically distributed with mean 0, then an appropriate model based estimator is

$$\bar{T}_{lin} = \sum_{S} y_i + \sum_{P-S} (\bar{\alpha} + \bar{\beta} x_i)$$

where $\bar{\alpha}$ and $\bar{\beta}$ are the appropriate weighted least squares estimators of α and β respectively. The presence of inclusion probabilities is the attribute of design based estimators while the model based estimator ignores the selection probabilities. The use of nonparametric regression for inference on finite populations is firmly within the model based approach.

However, it has a much greater degree of automaticity than is generally associated with model based inference based on standard parametric models (Dorfman 1992). Consider the regression model, $Y = m(X_i) + \sigma(X_i)e$ where $E(e) = 0$, $Var(e) = 1$, X and e are independent. Researches done by Dorfman and Hall (1993) and Chambers and Dorfman (1993)

dealt on estimating $m(x)$, a smooth function. The amount of smoothing depends on the size of bandwidth. Therefore, proper choice of bandwidth is our major concern here. The problem is how close is $\bar{m}(x_j)$ to $m(x_j)$. The formula given by the above researchers does not allow us to determine the best bandwidth, because it depends on unknown quantities. The nonparametric calibration estimator that seems fairly immune to variations in bandwidth was applied but further work is required for some automatic way of selecting the bandwidth. The expression for the asymptotic bias of this version of a nonparametric regression estimator of total does not include division by the sampling density, and so the bias of a local linear regression based estimator should be less sensitive to sparse x regions in the sample data. We make use of the local linear regression procedure to study the properties of the derived estimators and compare their performances with some estimators that exist in the literature.

1.2 Basic concepts, definitions and terminologies

In local linear or local polynomial regression, a low order weighted least squares regression is fitted at each point of interest x, using data from some neighborhood around x. Letting (X_i, Y_i) be ordered pairs, consider the regression model of the form

$Y_i = m(X_i) + \sigma(X_i)\varepsilon_i$ where, $E(Y_i|X_i = x_i) = m(x_i)$

$$Cov(Y_i, Y_j|X_i = x_i, X_j = x_j) = \begin{cases} \sigma^2(x_i), & i = j \\ 0 \quad , & i \neq j \end{cases} \qquad i, j = 1,2,3, \dots, n$$

The properties of the error are given by $E(\varepsilon_i|X_i = x_i) = m(x_i)$

$$Cov(\varepsilon_i, \varepsilon_j|X_i = x_i, X_j = x_j) = \begin{cases} \sigma^2(x_i), & i = j \\ 0 \quad , & i \neq j \end{cases} \qquad i, j = 1,2,3, \dots, n$$

where $m(.)$ and $\sigma^2(.)$ are assumed to be smooth functions of $x_i's$. e and x are independent. Fan and Gijbels (1996) estimate $m(x)$, using Taylor's Expansion of the form:

$$m(x) \approx m(x_0) + (x - x_0)m'(x_0) + \frac{(x - x_0)^2}{2!}m''(x_0) + \cdots + \frac{(x - x_0)^p}{p!}m^p(x_0) \qquad (*)$$

1.2.1 Local linear regression

Local linear regression is a nonparametric technique used for smoothing scatter plots and modeling functions. Local linear regression eliminates design bias and alleviates boundary bias. More precisely, a function is locally linear at a point if and only if a tangent line exists at the said point. Thus, local linearity is the graphical manifestation of differentiability. Functions that are locally linear are called smooth. Functions are locally linear everywhere except where they have a discontinuity, that is, jumps, breaks, vertical asymptotes and sharp corners.

1.2.2 Local polynomial regression

This is a nonparametric technique which is a generalization of kernel regression. In the approximation $(*)$ of section 1.2, when $p = 0$, we refer to this as local constant regression, when $p = 1$, this is local linear regression and when $p \geq 2$, this is local polynomial regression where p is the order of the local polynomial being fitted.

1.2.3 Sample

In statistics, a sample is a finite part of a population whose properties are studied in order to gain information about the whole population. We assume that there exists a population frame or list denoted by $\underline{U} = (U_1, U_2, \dots, U_N)$ of N identified units. Typically, the population is very large, making a census or a complete enumeration of all the values in the population is impractical or impossible. The sample represents a subset of manageable size$s$ of U. Samples are collected and statistics are calculated from the samples so that one can make inferences or extrapolations from the sample to the population. This process of collecting information from a sample is referred to as sampling. The best way to avoid a biased or unrepresentative sample is to select a random sample, also known as a probability sample. A random sample is defined as a sample where the probability that any individual member from the population being

selected as part of the sample is exactly the same as any other individual member of the population. Several types of random samples are simple random sample, systematic sample, stratified random sample, and cluster random sample. A sample that is not random is called a non random sample or a nonprobability sample. Some examples of non random samples are convenience samples, judgement samples, purposive samples, quota samples, snowball samples, and quadrature no design quasi Monte Carlo methods.

1.2.4 Population

The word population or statistical population is used for all the individuals or objects on which we have to make some study about the characteristic of interest. We may be interested to know the quality of bulbs produced in a factory. The entire product of the factory in a certain period is called a population. The entire lot of anything under study is called population. All the fruit trees in a garden, all the patients in a hospital and all the cattle in a cattle yard are examples of population in different studies.

1.2.5 Finite population

A population is called finite if it is possible to account for its individuals. It may also be called an accountable population. The number of vehicles crossing a bridge every day, the number of births per year and the number of words in a book are finite populations. The number of units in a finite population is denoted by N. Thus N is the size of the population and $N < \infty$.

1.2.6 Infinite population

This is a population with a set of numbers which continue on and on forever. In this case, sampling is done with replacement from a finite population. This implies that the size of the population N tends to infinity, that is $N \to \infty$. Sampling from an infinite population is handled by regarding the population as represented by a distribution. A random sample from an infinite

population is therefore considered as a random sample from a distribution. Let us suppose that we want to examine whether a coin is fair or not. We shall toss it a very large number of times to observe the number of heads. All the tosses will make an infinite population. The number of germs in the body of a malaria patient is perhaps something which is infinite.

1.2.7 Target and sampled population

Suppose we have to make a study about the problems of the families living in rented houses in a certain big city. All the families living in rented houses are our target population. The entire target population may not be considered for the purpose of selecting a sample from the population. Some families may not be interested to be included in the sample. We may ignore some part of the target population to reduce the cost of study. The population out of which the sample is selected is called sampled population or studied population.

1.2.8 Super populations and super population models

A Super population is a hypothetical infinite population from which the finite population is a sample, (Deming and Stephan, 1941). A super population model provides an alternative framework for inference in sampling. Such models are summarized by Cassel et al. (1977) and discussed by Bolfarine and Zacks (1991).

1.2.9 Sampling

This is a statistical procedure concerned with the selection of individual observations intended to yield some knowledge about a population of concern for the purposes of statistical inference.

1.2.10 Statistics

Statistics is a mathematical body of science that pertains to the collection, analysis, interpretation or explanation, and presentation of data. While many scientific investigations

make use of data, statistics is concerned with the use of data in the context of uncertainty and decision making in the face of uncertainty. In applying statistics to a problem, it is common practice to start with a population or process to be studied. For instance, populations can refer to all persons living in a country or every atom composing a crystal.

Ideally, statisticians compile data about the entire population (an operation called census). This may be organized by governmental statistical institutes. Descriptive statistics can be used to summarize the population data. Numerical descriptors include mean and standard deviation for continuous data types (like income), while frequencies and percentages are more useful in terms of describing categorical data (like race).

When a census is not feasible, a chosen subset of the population called a sample is studied. Once a sample that is representative of the population is determined, data is collected for the sample members in an observational or experimental setting. Again, descriptive statistics can be used to summarize the sample data. However, the drawing of the sample has been subject to an element of randomness, hence the established numerical descriptors from the sample are also due to uncertainty. To still draw meaningful conclusions about the entire population, inferential statistics is needed. It uses patterns in the sample data to draw inferences about the population represented, accounting for randomness.

These inferences may take the form of: answering yes/no questions about the data (hypothesis testing), estimating numerical characteristics of the data (estimation), describing associations within the data (correlation) and modelling relationships within the data (for example, using regression analysis). Inference can extend to forecasting, prediction and estimation of unobserved values either in or associated with the population being studied; it can include extrapolation and interpolation of time series or spatial data, and can also include data mining.

1.2.11 Simple random sampling

This is a sampling technique where every item in the population has an even chance and likelihood of being selected in the sample. Here the selection of items completely depends on chance or by probability and therefore this sampling technique is also sometimes known as a method of chances.

1.2.12 Simple random sampling with replacement

This is a method of selection of n units out of the N units one by one such that at each stage of selection each unit has equal chance of being selected, that is, $1/N$. In this sampling procedure, the selections are put back into the sampling frame such that repeat selections are possible.

1.2.13 Simple random sampling without replacement

Simple random sampling without replacement of size n is the probability sampling design for which a fixed number of n units are selected from a population of N units without replacement such that every possible sample of n units has equal probability of being selected. A resulting sample is called a simple random sample. In other words, this procedure does not allow the same random selection to be made more than once.

1.2.14 Stratified random sampling

A stratum is a mutually exclusive sub population considered to be more homogeneous with respect to the characteristic investigated than the total population. Stratified sampling is a sampling procedure carried out in such a way that portions of the sample are drawn from the different strata and each stratum is sampled with at least one sampling unit. Stratified simple random sampling is simple random sampling from each stratum.

1.2.15 Two stage cluster sampling

Two stage cluster sampling is frequently used because an adequate frame of elements is not available or would be prohibitively expensive to construct, but a listing of clusters is available. In stage one, a sample of clusters is selected and in stage two, sub samples of elements within each selected cluster are obtained.

1.2.16 Sampling frame

This is a list of units in the population from which we collect data for a particular measurable characteristic or units in the population for which we make statistical inference. In the case of a simple random sample, all units from the sampling frame have an equal chance to be drawn and to occur in the sample. In the ideal case, the sampling frame should coincide with the population of interest.

1.2.17 Probability sampling

This is a scheme in which every unit in the population has a chance (greater than zero) of being selected in the sample, and this probability can be accurately determined. The combination of these traits makes it possible to produce unbiased estimates of population totals by weighting sampled units according to their probability of selection. Probability sampling includes: Simple Random Sampling, Systematic Sampling, Stratified Sampling, Probability Proportional to Size Sampling, and Cluster or Multistage Sampling. These various ways of probability sampling have two things in common; every element has a known non zero probability of being sampled and involves random selection at some point.

1.2.18 Nonprobability sampling

This is any sampling method where some elements of the population have no chance of selection (these are sometimes referred to as out of coverage or under covered), or where the probability of selection can't be accurately determined. It involves the selection of elements based on assumptions regarding the population of interest, which forms the criteria for selection. Hence, because the selection of elements is non random, nonprobability sampling does not allow the estimation of sampling errors. These conditions place limits on how much information a sample can provide about the population. Information about the relationship between the sample and the population is limited, making it difficult to extrapolate from the sample to the population.

Nonprobability sampling includes: Accidental sampling, Quota sampling and Purposive sampling. In addition, non response effects may turn any probability design into a nonprobability design if the characteristics of non response are not well understood, since non response effectively modifies each element's probability of being sampled.

1.2.19 Bandwidth

This is the parameter that controls the amount of smoothing inherent in the estimation procedures. The bandwidth parameter denoted by b which determines how large a neighbourhood of the target point is used to calculate the local average. A large bandwidth generates a smoother curve, while a small bandwidth generates a wigglier curve. To ensure that there are enough observations within the neighbourhood of a target point for parametric estimation we take a large sample size n. Hence in studying the theoretical properties of the estimator of the total we shall impose the conditions that as $n \to \infty, b \to 0$, such that $nb \to \infty$.

1.2.20 Kernel function

This is a symmetric probability density function that uses a linear classifier to solve a nonlinear problem. It entails transforming linearly inseparable data to linearly separable ones. The kernel function is what is applied on each data instance to map the original nonlinear observations into a higher dimensional space in which they become separable. For instance, consider $K(x, y) = < f(x), f(y) >$. Here K is the kernel function, x, y are n dimensional inputs. f is a map from n dimensional space to m dimensional space. $< x, y >$ denotes the dot product. usually m is much larger than n.

1.2.21 Kernel smoother parameter

In producing a kernel curve, the smoothing parameter, which is related to bandwidth, controls the smoothness of the fit. Mean integrated square error (MISE) is the value of the smoothing parameter that minimizes the mean square error (MSE) using generalized cross validation. Smoothing refers to estimating a smooth trend, usually by means of weighted averages of observations. A smoother is a device used for summarizing the trend of a response measurement $\underline{Y}$ as a function of one or more predictor variables $\underline{X}$. The term smooth is used because such averages tend to reduce randomness by allowing positive and negative random effects to partially offset each other.

1.2.22 Robustness

The degree to which an estimator can function correctly or efficiently in the presence of invalid inputs or stressful environmental conditions. For example, in examining the robustness properties of our results, we determine the ability of the estimators to withstand or overcome adverse conditions or rigorous testing.

1.2.23 Estimate and Estimator

An estimate is an approximate calculation or judgement of the value, number or quantity obtained from a sample. An estimator refers to a statistic that is used to generate an estimate once data are collected. So the estimator is the tool that can be used to estimate the population parameter of interest. An estimate is the product of one application of that tool. Like for instance, the sample total is an estimator of the population total, the sample mean is an estimator of the population mean, the sample variance is an estimator of the population variance and the sample proportion is an estimator of the population proportion.

1.2.24 Accuracy and Precision

In measurement of a set, accuracy refers to closeness of the measurements to a specific value. Precision refers to the closeness of the measurements to each other, that is, how close are estimates from different samples to each other. For example, the standard error is a measure of precision. When the standard error is small, sample estimates are more precise and when the standard error is large, sample estimates are less precise.

1.2.25 Bias

In statistics, the bias of an estimator is the difference between this estimator's expected value and the true value of the parameter being estimated. An estimator or decision rule with zero bias is called unbiased. A statistic is said to be an unbiased estimate of a given parameter when the mean of the sampling distribution of that statistic can be shown to be equal to the parameter being estimated. For example, the mean of a sample is an unbiased estimate of the mean of the population from which the sample was drawn.

1.2.26 Consistent Estimator

Consistency of an estimator means that as the sample size gets large the estimate gets closer and closer to the true value of the parameter. In statistics, a consistent estimator or asymptotically consistent estimator is an estimator or a rule for computing estimates of a parameter θ_0 having the property that as the number of data points used increases indefinitely, the resulting sequence of estimates converges in probability to θ_0.

1.2.27 Census

This is an official count or survey especially of a population. A census is the procedure of systematically acquiring and recording information about the members of a given population. The modern census is essential to international comparisons of any kind of statistics, and censuses collect data on many attributes of a population, not just how many people there are. Census is important because this process helps compile a numerical profile of a nation. A population census is a total count of the country's population, where demographic, social and economic information, as well as information about the housing conditions of the people who live in the country is gathered.

1.2.28 Finite population parameter

The finite population is the population which can be counted. The objective of the statistics is to estimate the parameters of a finite population. The population parameter is the number that describes the population. Also, the population parameter takes up a numerical value that represents the population. Moreover, the population parameter is obtained from a statistic which is calculated from a randomly selected sample of the given population. A parameter is any summary number, like an average or percentage, that describes the entire population. The population mean μ and the population proportion P are two different population parameters.

1.2.29 Descriptive inference and analytic inference

In many estimation problems, the sample is used to describe and analyse the target population from which it was selected by estimating population parameters and other descriptive and analytic inferences such as correlations. Some common parameters of interest for the finite population $\underline{Y} = (y_1, y_2, \dots, y_N)'$ are the finite population total, the finite population mean, the finite population variance and the finite population proportion respectively defined by $T = \sum_{i=1}^{N} y_i$, $\bar{T} = \frac{1}{N}\sum_{i=1}^{N} y_i$, $V(x) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y}_i)^2$ and $P = \frac{A}{N}$

On the other hand, inferences may explore properties of the process that generate the population values (Montanari and Ranalli (2003)). We assume that the finite population has been generated by a super population model $\xi = f(x, y, \varphi)$ and we are interested in estimating the population parameters $\varphi = (\alpha, \beta)$, where $y_i = \alpha + \beta x_i$. The super population model can be applied to predict the unobserved values $y_i's$ after obtaining estimates of α and β using the known auxiliary information x_i, $i = 1,2 \dots, N$.

1.3 Statement of the problem

In many complex surveys, auxiliary information about the population of interest is available. One approach to using this auxiliary information in estimation is to assume a working model ξ describing the relationship between the study variable of interest and the auxiliary variables. Estimators are then derived on the basis of this model. Estimators are sought which have good efficiency if the model is true, but maintain desirable properties like design consistency if the model is false. Often, a linear model is selected as the working model. In some situations, the linear model is not appropriate, and the resulting estimators do not achieve any efficiency gain over purely design based estimators.

Wu and Sitter (2001) proposed a class of estimators for which the working models follow a nonlinear parametric shape. However, efficient use of any of these estimators requires a priori knowledge of the specific structure of the population. This is especially problematic if the working model is to be used for many variables of interest, a common occurrence in surveys.

Because of these concerns, some researchers have considered nonparametric models instead of parametric models. The nonparametric approach does not restrict the functional form of the distribution nor does it specify the various stochastic properties such as $E_\xi(.)$, $V_\xi(.)$ and $MSE_\xi(.)$. Rather, it leaves them to cover broad classes of models, thus allowing for more robust inference than that obtained in parametric approach (Dorfman, 1992, Chambers et al.,1993 and Dorfman, 2002). Dorfman (1992), used nonparametric regression to estimate the unknown values of the survey variable using kernel regression. However, the use of local linear regression procedure in a purely model based framework is an open area that requires further study.

The local linear regression procedure has advantages over popular kernel based methods in the sense that it adapts well to bias problems at boundaries and in regions of high curvature, it can be tailored to work for many different distributional assumptions due to its simplicity, it does not require smoothness and regularity conditions required by other methods such as the boundary kernels and having a local model (rather than just a point estimate) enables derivation of response adaptive methods for bandwidth and polynomial order selection in a straightforward manner. It is also asymptotically efficient among all linear smoothers including those produced by the kernel, orthogonal series and penalized spline methods.

Because of the unexploited good properties of this estimation procedure, there is need to explore its other properties. This procedure can be adapted to suit the case when two stage cluster sampling is applied, depending on situations whether the auxiliary information is

available at the cluster level, element level for all elements, or element level for elements in the selected clusters only. Moreover, the use of this procedure in stratified random sampling and in two stage cluster sampling in a purely model based framework remains an open field that needs further exploration.

1.4 Objectives of the study

The general objective is to construct model based robust estimators of finite population total using the procedure of local linear regression.

The specific objectives are:

i. To derive robust estimators of finite population total based on local linear regression.

ii. To investigate properties of the derived local linear regression estimators.

iii. To determine optimal bandwidth to be used in the derived estimators.

iv. To extend local linear regression estimation procedure to stratified random sampling and to two stage cluster sampling.

v. To compare the performances of the derived local linear regression estimators with some estimators that exist in the literature.

1.5 Significance of the study

The study contributes towards development of mathematical and statistical knowledge in survey sampling. The developed estimation procedure is useful to policy makers since national development is dependent on the sampling strategy employed. In addition, business and industrial sectors stand to benefit from this study by using the developed estimation procedure for prediction and thereby improving the efficiency of their internal operations.

1.6 Outline of the thesis

The rest of this thesis is organized as follows: In chapter two, a critical review of the work done by other researchers in the nonparametric estimation of the finite population parameters is accomplished. In chapter three, some robust estimators of the finite population total using the procedure of local linear regression are derived in a model based framework and their properties investigated. The local linear regression estimation procedure is adapted and extended to stratified random sampling and to two stage cluster sampling. In chapter four, a study is carried out to compare the performances of the estimators derived in chapter three with some other estimators that exist in the literature. Finally, in chapter five, a summary of the study is outlined in terms of the conclusions and recommendations for further research.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

In sample surveys, auxiliary information on a finite population is regularly used to increase the precision of estimators of the finite population parameters. We argue that under a general modeling process, complete auxiliary information is incorporated in the construction of estimators through fitted values. Once a modeling approach is undertaken, we then have a special feature in finite population estimation problems that the unknown quantities are realized values of random variables, so that the basic problem has the feature of being similar to a prediction problem. In this chapter, we review the work done on nonparametric estimation of finite population parameters in survey sampling.

2.2 Nonparametric regression

The idea of nonparametric regression is introduced by Nadaraya (1964) and Watson (1964). Several types of nonparametric regression methods such as the kernels, penalized splines and orthogonal series are in existence. Let us consider the Nadaraya-Watson estimator. Let the regression model be given by $Y = m(X) + \sigma(X)e$ where $m(x)$ is a smooth function and e is a random variable with mean zero and constant variance. Here the population generated by this model is finite and the objective is to estimate $m(x)$, a smooth function. The following table summarizes the asymptotic behavior of the Nadaraya-Watson smoother, the Gasser-Muller smoother and the Local Linear smoother, where k is a kernel and b_N is a smoothing parameter.

Table 2.1: The point wise Bias and Variance of kernel regression smoothers

Technique	Conditional Mean	Error Variance
Nadaraya-Watson smoother	$\frac{1}{2}\mathrm{m}''(\mathrm{x}) + \frac{\mathrm{m}'(\mathrm{x})\mathrm{f}_{\mathrm{x}}'(\mathrm{x})}{\mathrm{f}_{\mathrm{x}}(\mathrm{x})} \int_{-\infty}^{\infty} \mathrm{u}^2\, \mathrm{k}(\mathrm{u})\mathrm{dub}_{\mathrm{n}}^2$	$\frac{\sigma^2(\mathrm{x})}{\mathrm{f}_{\mathrm{x}}(\mathrm{x})\mathrm{nb}_{\mathrm{N}}} \int_{-\infty}^{\infty} \mathrm{k}^2\,(\mathrm{u})\mathrm{du}$
Gasser-Muller smoother	$\frac{1}{2}\mathrm{m}''(\mathrm{x}) \int_{-\infty}^{\infty} \mathrm{u}^2\, \mathrm{k}(\mathrm{u})\mathrm{dub}_{\mathrm{n}}^2$	$\frac{3\sigma^2(\mathrm{x})}{2\mathrm{f}_{\mathrm{x}}(\mathrm{x})\mathrm{nb}_{\mathrm{N}}} \int_{-\infty}^{\infty} \mathrm{k}^2\,(\mathrm{u})\mathrm{du}$
Local Linear smoother	$\frac{1}{2}\mathrm{m}''(\mathrm{x}) \int_{-\infty}^{\infty} \mathrm{u}^2\, \mathrm{k}(\mathrm{u})\mathrm{dub}_{\mathrm{n}}^2$	$\frac{\sigma^2(\mathrm{x})}{\mathrm{f}_{\mathrm{x}}(\mathrm{x})\mathrm{nb}_{\mathrm{N}}} \int_{-\infty}^{\infty} \mathrm{k}^2\,(\mathrm{u})\mathrm{du}$

The bias of the Nadaraya-Watson smoother depends on the intrinsic part $\mathrm{m}''(\mathrm{x})$ interplaying with the artifact $\mathrm{m}'(\mathrm{x})\left\{\frac{\mathrm{f}_{\mathrm{x}}'(\mathrm{x})}{\mathrm{f}_{\mathrm{x}}(\mathrm{x})}\right\}$ due to the local constant fit. Keeping $\mathrm{m}''(\mathrm{x})$ fixed, we first remark that in the highly clustered design where $|\frac{f_x'(x)}{f_x(x)}|$ is large; the bias of the Nadaraya-Watson smoother is large. This implies that the estimator cannot adapt to highly clustered designs. We also remark that when $|\mathrm{m}'(x)|$ is large, then the bias of that estimator is also large. This means that even in the case of linear regression $\mathrm{m}(x) = \alpha + \beta(x)$ with a large coefficient β, the bias of the estimator is also large. In other words, the Nadaraya-Watson smoother is not good at testing linearity. Supposing that we wish to estimate $m(x)$.

One possibility suggested by Nadaraya (1964) and Watson (1964) is that of averaging the nearby values of Y_i measured in terms of the distances $|x_i - x|$. Let $k(u)$ be a symmetric density function. For a chosen bandwidth b define, $k_b(u) = b^{-1}k\left(\frac{u}{b}\right)$ and let the Kernel based weights be $w_i(x) = \frac{k_b(x_i - x)}{\sum_{i=1}^{n} k_b(x_i - x)}$. The larger b is, the flatter and broader the density function, and the more equal the weights and vice-versa. Thus, the Nadaraya-Watson estimator of $m(x)$

is, $\bar{m}(x) = \sum_{i=1}^{n} w_i(x) y_i$ where $\bar{m}(x)$ will be consistent for $m(x)$, if $b \to 0$ as $n \to \infty$ in such a way that $nb \to \infty$ under some reasonable conditions on $m(x)$ and the design points x_i. Letting x_j be any point in the non-sample and then estimating $m(x_j)$, Dorfman (1992) adopted this procedure in estimating the finite population total defined by $\bar{T}_{np} = \sum_S Y_i + \sum_{P-S} \bar{m}(x_j) = \sum_S Y_i + \sum_{P-S} \sum_S (w_{ij} Y_i) = \sum_S Y_i + \sum_S W_i Y_i = \sum_S (1 + W_i) Y_i \ \ where \ W_i = \sum_{j \in P-S} (w_{ij})$. If we let x_j take values on a grid over the domain of interest, we generate a curve estimating $m(x_j)$ over the domain. Dorfman (1992) proved the asymptotic design unbiasedness and asymptotic MSE of the estimator. The estimator, however suffers from sparse sample problem, that is, lacks design adaptability and more work needs to be done to come up with a technique that can overcome this problem. This is where the local linear procedure comes in. Subsequently, we make use of this procedure to obtain estimators of finite population total.

In the parametric regression approach, a regression function is used to quantify the contribution of the covariate X to the response Y per unit value x to summarize the association between the two variables, to predict the mean response for a given value x and to extrapolate the results beyond the range of the observed covariate values. When using the extrapolation technique for the scatter diagram, a plot of the points $x_i's$ against $y_i's$ is constructed. If the pattern is not linear then the parametric regression approach is not suitable for adequately estimating the unknown population parameters for this type of data set and all other sets that arise in practice of this nature.

An alternative approach for estimating the nonlinear regression ought to be investigated and determined. The nonparametric approach is therefore considered a better approach than the parametric approach as it provides a versatile method of exploring a general relationship between two variables and it gives predictions of observations yet to be made without reference to a fixed parametric model. Some work on the distribution function, has been done (Chambers,

Dorfman and Wehrly, 1992; Dorfman and Hall, 1992; Smith and Njenga, 1992 and Cheng, 1993).

Consider the regression model of the form, $Y_i = m(X_i) + \sigma(X_i)\varepsilon_i$ where, $E(y_i|X_i = x_i) = m(x_i)$ and $Cov(y_i, y_j|X_i = x_i, X_j = x_j) = \begin{cases} \sigma^2(x_i), & i = j \\ 0 \quad , & i \neq j \end{cases}$ $i, j = 1,2,3 \ldots, N$. The properties of the error are given by $E(\varepsilon_i|X_i = x_i) = m(x_i)$ and $Cov(\varepsilon_i, \varepsilon_j|X_i = x_i, X_j = x_j) = \begin{cases} \sigma^2(x_i), & i = j \\ 0 \quad , & i \neq j \end{cases}$ $i, j = 1,2,3, \ldots, N$, where $m(.)$ and $\sigma^2(.)$ are assumed to be smooth functions of $x_i's$.

The nonparametric approach does not restrict the functional form of the distribution nor does it specify the various stochastic properties such as $E_\xi(.)$, $V_\xi(.)$ and $MSE_\xi(.)$. Rather, it leaves them to cover broad classes of models, thus allowing for more robust inference than obtained in parametric approach. Using the model ξ, the nonparametric estimator of total, T has been derived by Nadaraya (1964); Watson (1964); Priestly and Chao (1972); Gasser and Muller (1979); Dorfman (1992) and Otieno and Mwalili (2000).

2.3 Comparative studies

Nadaraya (1964) and Watson (1964) introduced the idea of nonparametric estimation of a regression curve using the model $Y = m(X_i) + \sigma(X_i)e$ where $m(x)$ is a smooth function and e is a random variable with mean zero and constant variance. Their objective was to estimate $m(x)$, a smooth function. The Nadaraya-Watson estimator of $m(x)$ is, $\overline{m}(x) = \sum_{i=1}^{n} W_i(x)Y_i$ and $\overline{m}(x)$ will be consistent for $m(x)$, if $b \to 0$ as $n \to \infty$ in such a way that $nb \to \infty$ under some reasonable conditions on $m(x)$ and the design points x_i.

Royall (1976) employed predictive inference based on linear models. Generalized regression estimators (Cassel et al., 1977; Sandal, 1980; Robinson and Sandal, 1983), including ratio estimators and linear regression estimators (Cochran, 1977), best linear unbiased estimators (Brewer, 1963; Royall, 1970), and post stratification estimators (Holt and Smith, 1979), are all derived from assumed linear models. Royall and Cumberland (1978) applied this predictive inference approach in a study of the properties of the ratio estimator and its variance. They concentrated on making model based methods robust to departures from assumptions.

Following Godambe (1982), a sampling strategy is defined as robust if, and only if, it attains the minimum value of the Godambe-Joshi lower bound to the expected variance (Godambe and Joshi (1965)). In the literature on the design based approach to finite population sampling, the term robustness usually has the more restrictive meaning of asymptotic design unbiasedness; Brewer (1979), Sarndal (1980), and Wright (1983). Although asymptotic design unbiasedness is necessary, it is by no means sufficient for a sampling strategy to attain the minimum value of the Godambe-Joshi lower bound if the assumed model is incorrect. It is well known, for example, that under the model, $E_\xi(Y_i) = x_i\beta \;\; and \;\; E_\xi\{(Y_i - x_i\beta)^2\} = x_i^2\sigma^2$ where β, σ are unknown and $E_\xi(.)$ denotes expectation under the model ξ, the best strategy for estimating the population total is the Horvitz-Thompson (1952) estimator together with a probability sampling design with first order inclusion probabilities π_i proportional to x_i. When the true model contains an intercept term, however, Sarndal (1980) showed that, although design unbiased, the Horvitz-Thompson estimator does not attain the minimum value of the Godambe-Joshi lower bound to the expected variance.

Given this concern with robustness, it is natural to consider a nonparametric class of models for ξ, because they allow the models to be correctly specified for much larger classes of functions. Kuo (1988), Dorfman (1992), Dorfman and Hall (1993), Chambers et al., (1993) and

Chambers (1996) adopted this approach in constructing model based nonparametric estimators. Kuo (1988) applied nonparametric regression to sample data to estimate the finite population distribution function.

In his study, Dorfman (1992) proved the asymptotic unbiasedness and asymptotic MSE of an estimator. The estimator, however suffers from sparse sample problem, and more work needs to be done to come up with another technique that can overcome this problem. In order to overcome the sparse sample problem suffered by the estimator of finite population distribution function, we introduce the local linear procedure for estimating the finite population total.

Smith and Njenga (1992) applied nonparametric regression for purposes of data exploration in analytical surveys. Dorfman and Hall (1993) developed methods and theory for nonparametric regression for finite population distribution function. Chambers (1996) described using nonparametric regression calibration successfully on multivariate data in combination with ridge regression methods. Simwa (1997) applied nonparametric models on observed reported data for HIV/AIDS incidence to determine trend patterns of the expected HIV/AIDS epidemic in both Kenya and Uganda.

Chambers and Dorfman (2002) observed that the calibration estimator based on the columnar model does slightly better than the best linear unbiased estimator at high band width. The estimator generally appears robust to changes in bandwidth, and gives exact unbiasedness and minimal variance for a particular weighted balanced sample. However, application to finite populations of methods more sophisticated than kernel regression should be explored, for example the variable bandwidth local linear regression approach of Fan and Gijbels (1996). Zheng and Little (2003) proposed a model based estimator that uses penalized spline regression, and Zheng and Little (2004) extended this estimator to two stage sampling designs.

A new type of model assisted nonparametric regression estimator for the finite population total, based on local polynomial smoothing which is a generalization of kernel regression has also been proposed. Stone (1977) introduced the theory of local linear regression using weighted least squares to fit a linear regression function to the data and evaluate this function at x. Consistent sequences of probability weight functions defined in terms of nearest neighbors are constructed and the results applied to verify the consistency of the estimators of the various quantities discussed. Cleveland (1979) and Cleveland and Devlin (1988) showed that these techniques are applicable to a wide range of problems. Theoretical work by Fan (1992, 1993), Ruppert and Wand (1994) and Ruppert et al. (1995) showed that they have many desirable theoretical properties, including adaptation to the design of the covariate(s), consistency and asymptotic unbiasedness. Wand and Jones (1995) provided a clear explanation of the asymptotic theory for kernel regression and local polynomial regression. The monograph by Fan and Gijbels (1992) and Fan and Gijbels (1996) explored a wide range of application areas of local polynomial regression techniques.

However, the application of these techniques to model-assisted survey sampling is new. Breidt and Opsomer (2000) used the traditional local polynomial regression estimator for the unknown regression function $m(x)$. They assumed that $m(x)$ is a smooth function of x and obtained an asymptotically design unbiased and consistent estimator of the finite population total. The local polynomial regression estimator has the form of the generalized regression estimator, but is based on a nonparametric super population model applicable to a much larger class of functions. Kasungo (2002) employed a model based approach in estimating the finite population total based on local polynomial regression. Simulation experiments indicated that the local polynomial regression estimator was more efficient than regression estimators when the model regression function was incorrectly specified, while being approximately as efficient when the parametric specification was correct.

Breidt et al. (2005) considered a related nonparametric model assisted regression estimator, replacing local polynomial smoothing with penalized splines. Chen et al. (2008) studied a weighted local linear regression smoother for longitudinal or clustered data. As a hybrid of the methods of Chen and Jin (2005) and Wang (2003), the proposed local linear smoother maintains the advantages of both methods in computational and theoretical simplicity, variance minimization and bias reduction. Ombui (2008) applied local polynomial regression in estimating parameters of the finite population. In his study, it was observed that the developed estimators were asymptotically unbiased, consistent and normally distributed when certain conditions were satisfied.

Kim et al. (2009) extended the local polynomial nonparametric regression estimation to two stage sampling, in which a probability sample of clusters is selected, and then subsamples of elements within each selected cluster are obtained. Two stage cluster sampling is frequently used because an adequate frame of elements is not available or would be prohibitively expensive to construct, but a listing of clusters is available. Sarndal et al. (1992) identified three cases of auxiliary information available for two stage sampling, depending on whether the information is available at the cluster level, element level for all elements, or element level for elements in selected clusters only.

Harms and Duchesne (2010) derived asymptotic properties of the model assisted local linear estimator under the combined inference approach. They showed that the bias of $\bar{m}(.)$ is the same as in the identically independent distribution (iid) case but the variance equaled that from the iid case multiplied by a correction factor derived from the sampling scheme. Su et al. (2012) outlined the idea of the extension of local polynomial fitting to a linear heteroscedastic regression model. They verified the asymptotic normality of the parameters based on numerical simulations applicable to a case of economics which indicated their method to be surely

effective in finite sample situations. Sanchez et al. (2014) estimated $m(.)$ using a modified local constant estimator for the mixed variable case. Rady and Ziedan (2014) estimated the finite population total in the presence of two auxiliary variables using the bootstrap method and jackknife method. A comparison between different methods was performed on the basis of mean squared error (MSE), mean absolute error (MAE) and mean absolute percentage error (MAPE).

Luc (2016) derived asymptotic properties of probability weighted nonparametric regression estimator under a combined inference framework for complex surveys. However, the nonparametric regression estimator considered here is the local constant estimator. Simulation studies showed that the bias of the modified nonparametric regression estimator had the same leading terms and order of probability as under the model based framework. He examined asymptotic properties under the combined inference approach and tested the performance of the estimator against the traditional model based local constant estimators. Syengo (2018) studied local polynomial regression under stratified random sampling where simulation experiments showed that the resulting estimator exhibited good properties.

Mostafa and Shan (2019) estimated finite population total from complex sample surveys in the presence of auxiliary information in a model assisted framework. Results suggested that their proposed estimators performed well relative to the other model based and model assisted estimators as well as the customary Horvitz–Thompson estimator under different levels of misspecification in the working model. Parichha et.al. (2019) described the problem of estimation of finite population mean in two phase stratified random sampling. The efficacy of the proposed methodology had been justified through empirical investigations carried out using the data set of natural population as well as the data set of artificially generated population.

CHAPTER THREE

THEORY AND METHODS

3.1 Introduction

In this chapter, we derive model based robust estimators of finite population total using the procedure of local linear regression. In particular, properties of local linear regression estimators are investigated. Local linear regression is a design adaptive nonparametric regression approach that is based on the theory of weighted least squares regression. The estimators based on this approach solve the drawbacks of the two popular kernel estimators described in chapter two due to their nature of design adaptability.

In examining the properties of local linear regression estimators, the following assumptions considered by Fan (1993) and Ruppert and Wand (1994) are used:

i. The x_j variables lie in the interval $(0,1)$.

ii. The function $m''(.)$ is continuous on $(0,1)$.

iii. The kernel $K(.)$ is symmetric and supported on $(-1,1)$. Also $K(t)$ is bounded and continuous and satisfying the following: $\int_{-\infty}^{\infty} K(x)\,dx = 1$, $\int_{-\infty}^{\infty} xK(x)\,dx = 0$, $\int_{-\infty}^{\infty} x^2K(x)\,dx \neq 0$, $\int_{-\infty}^{\infty} K^{2r}(x)\,dx < \infty$ for $r = 1,2,\ldots$.

iv. The bandwidth h is a sequence of values which depend on the sample size n and satisfying $h \to 0$ and $nh \to \infty$ as $n \to \infty$.

v. The point x_j at which the estimation is taking place satisfies $h < x_j < 1 - h$.

Consider the local polynomial regression. Then the estimate of $m(x)$ at any value of x is obtained by the minimization problem

$$\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1(x_i - x) - \beta_2(x_i - x)^2 - \cdots - \beta_p(x_i - x)^p\right)^2 K_b(x - x_i) \qquad (3.1)$$

with respect to $\beta_0, \beta_1, \ldots, \beta_p$, where $\underline{\beta} = \left(\beta_0, \beta_1, \ldots, \beta_p\right)'$. The result is therefore a weighted least squares estimator with weights $K_b(x - x_i)$. Using the notations

$$X = \begin{bmatrix} 1 & x - x_1 & \dots & (x - x_1)^p \\ 1 & x - x_2 & \cdots & (x - x_2)^p \\ & \vdots & \ddots & \vdots \\ 1 & x - x_n & \cdots & (x - x_n)^p \end{bmatrix}, \qquad \underline{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

and

$$W = \begin{bmatrix} K_b(x - x_1) & 0 & \dots & 0 \\ 0 & K_b(x - x_2) & \cdots & 0 \\ & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & K_b(x - x_n) \end{bmatrix},$$

we can compute $\bar{\beta}$ which minimizes (3.1) by usual formula for a weighted least squares estimator

$$\bar{\beta}(x) = (X'WX)^{-1}X'WY \tag{3.2}$$

Then, the local polynomial estimator of the regression function $m(x)$ is

$$\bar{m}(x) = \bar{\beta}_0(x) = \underline{e}_1'(X'WX)^{-1}X'WY \tag{3.3}$$

where $\underline{e}_1$ is the $n \times 1$ vector having 1 in the first entry and 0 elsewhere.

3.2 Sample survey strategies

The theory of sample surveys involves principles and methods of collecting and analysing data from a finite population of N units and then making inferences about finite population parameters on the basis of information obtained from the sample. Unified frameworks for survey designs and estimation methods to finite population inference have been considered by researchers in the past and classified as design based approach, model assisted approach, combined inference approach and model based approach. Comparing and contrasting them in terms of their concepts of efficiency and robustness to assumptions about the characteristics of the population, it has been concluded that although none of these approaches delivers both efficiency and robustness, the model based approach seems to achieve the best compromise among the other approaches. This study considers a model based approach to finite population

total estimation using local linear regression procedure. A brief discussion on these survey strategies is given below as described in Chambers (2003).

3.2.1 Design based approach

This is also called classical approach or randomisation approach. This approach is based on the assumption that the population values of the survey measurements are fixed constants so that there is no model generating the population values. It follows that the only probabilistic information resides in the sample selection probabilities under the prevailing random design. In this approach, each population unit $U_1, U_2, \dots, U_N$ is associated with a fixed but unknown real number which is the value of the variable under study. Inference is based on the values of the survey variable $\underline{Y} = (y_1, y_2, \dots, y_N)'$ picked through the design $p(s)$. The approach assumes that the population units are labelled and the researcher has access to the fixed value y_i of U_i population units. This becomes involving and cannot be justified.

An estimator $\bar{T}(y)$ based on the design $p(s)$ is said to be unbiased for $T(y)$ if

$$E_p(\bar{T}(y)|s) = \sum_{s \in S} \bar{T}(y)p(s) = T(y) \tag{3.4}$$

where $E_p(\bar{T}(y)|s)$ is the conditional expectation of $\bar{T}(y)$, given that the sample s is chosen through some design $p(s)$. Consider the theorem;

An estimator $\bar{T}(y) = \sum_{s \in S} l_{si} Y_i$ for the population total is unbiased for $T(y)$ where $\sum_{s \in S} l_{si} p(s) = 1$, i is the i^{th} unit and $1 \leq i \leq N$.

Proof;

$$E_p(\bar{T}(y)|s) = \sum_{s \in S} \bar{T}(y)p(s)$$

$$= \sum_{s \in S} p(s) \sum_{s \in S} l_{si} y_i$$

$$= \sum_{s \in S} Y_i \sum_{s \in S} l_{si} p(s)$$

$$= \sum_{i=1}^{N} Y_i, \; since \; \sum_{i \in S} l_{si} p(s) = 1$$

$$= T(y) \tag{3.5}$$

On the other hand, the variance of $\bar{T}(y)$ is

$$Var\left(\bar{T}(y|s)\right) = E_p\left(\bar{T}(y|s) - E_p\left(\bar{T}(y|s)\right)\right)^2 \tag{3.6}$$

However, if the estimator is biased, then its MSE is

$$MSE_p\{\bar{T}(y)|s\} = E_p\left(\bar{T}(y)|s - T(y)\right)^2$$

$$= E_p\left((\bar{T}(y)|s) - E_p\left(\hat{T}(y)|s\right) + E_p(\bar{T}(y)|s) - T(y)\right)^2$$

$$= E_p\left(\bar{T}(y)|s - E_p\bar{T}(y)|s\right)^2 + E_p\left(E_p\bar{T}(y)|s - T(y)\right)^2$$

$$= Var_p(\bar{T}(y)|s) + \left(B_p T(y)\right)^2 \tag{3.7}$$

If we minimize (3.7), then we can deduce the performance of any sampling strategy. However, in minimization criterion, problems can arise when selecting a sampling strategy between an unbiased estimator with a small variance and the biased estimator with a small MSE.

3.2.2 Model assisted approach

In this approach, randomization based theory is treated as the only true approach to inference and models are only used to help choose between valid randomization based alternatives. This means that one chooses among randomization consistent estimation strategies by hypothesizing a reasonable and practical model, restricting attention to model unbiased estimators and selecting that strategy with minimum model expected randomization mean square error. The inference for model assisted approach is a form of randomization inference that employs

models to help determine the point estimators. The purely design based approach does not take into account the auxiliary information in its estimation of the finite population total while the purely model based approach ignores the inclusion probabilities which are based on the design used to select the sample. The model assisted approach assumes a working model ξ describing the relationship between the study variable and the auxiliary variable. Estimators are then derived on the basis of this model. These estimators have good efficiency if the model is true, but maintain desirable properties like asymptotic design unbiasedness and design consistency if the model is false.

3.2.3 Combined inference approach

It is assumed that a finite population is generated based on a selected model, where the predictor variables and outcome variable are assumed to follow a joint probability distribution. Then a sample is drawn from this population based on a probability sampling design (Pfefferman 1993). This type of estimation can be thought of having two stages; a model stage and a design stage. In the model stage, a model is selected based on the belief that it has generated the population. Nonparametric regression models are attractive in this case as they are consistent under minimal restrictions on the underlying function. The relationship between the outcome and predictor variables is estimated in the design stage, where a sample is drawn according to a specified sampling plan and the corresponding weights are included in the model.

3.2.4 Model based approach

This is a predictive approach or a super population approach. In this approach, a super population model provides an alternative framework for inference in finite population. Such models do not require the units in the super population to be identifiable. However, if the super population arises as a random permutation of identifiable units in a population, the units in the

super population are potentially identifiable. We assume that the units in the population are identifiable, and that the super population depends on this population definition. The model employed characterise the actual values, both the observed and the unobserved which are considered as a realization of a random variable Y. We employ a sampling scheme consistent with the selected model, taking into account practical considerations such as costs to draw a sample s. We let the probability be of the form $p(s|y)$. We then use the model, the sample and the information in the sampling scheme to make an inference about the unobserved random variables $Y_i's$. The super population model is

$$E(Y_i) = \beta x_i, \quad i = 1,2, \dots, N \tag{3.8}$$

$$Var(Y_i) = \sigma^2(x_i), \qquad i = 1,2, \dots, N \tag{3.9}$$

$$Cov(Y_i, Y_j) = 0, \qquad i \neq j; \;\; i,j = n+1, \dots, N \tag{3.10}$$

We let $\bar{T}(y)$ be an unbiased estimator of $T(y)$. Therefore, the estimator $\bar{T}(y)$ is said to be unbiased for $T(y)$ if $E_m\left(\bar{T}(y)|(S,\underline{Y})\right) = E_m(\bar{T}(y))$ where $E_m\left(\bar{T}(y)|(S,\underline{Y})\right)$ denotes the conditional expectation of $\bar{T}(y)$ given the sample $(S,\underline{Y})$ with respect to a given model ξ.

3.3 The proposed estimator

Consider a finite population of size N labeled $U_1, U_2, \dots \dots, U_N$. We have (x_i, Y_i), $i = 1,2, \dots \dots, N$ associated with each unit. The values $x_1, x_2, \dots \dots, x_N$ are known and can be used in the sample design, or in the estimator, or in both. The selection variable set S denotes sample of size n from T, for which y values are unknown. S is an ignorable set, meaning, given information on x, knowledge of how the sample was taken provides no additional information about y (Dorfman 1992). Let T be the finite population total defined by

$$T = \sum_{i=1}^{N} Y_i = \sum_{i \in S} y_i + \sum_{i \in R} y_i \tag{3.11}$$

where $\sum_{i\in S} y_i$ is known while $\sum_{i\in R} y_i$ is unknown such that R is an indexing set for the y values which are unknown to the investigator. The estimator of the finite population total, the bias and the error variance can be determined using the super population model of the form

$$Y_i = m(X_i) + \sigma(X_i)\varepsilon_i \tag{3.12}$$

In particular, the following assumptions hold for the model considered in the nonparametric regression estimation of $m(x_i)$, thus

$$E(Y_i|X_i) = m(x_i) \tag{3.13}$$

$$Var(Y_i|X_i) = \sigma^2(x_i) \tag{3.14}$$

$$Cov\big(Y_i, Y_j|X_i = x_i, X_j = x_j\big) = \begin{cases} \sigma^2(x_i), & i = j \\ 0 \quad , & i \neq j \end{cases} \quad i, j = 1, 2, 3, \dots, N \tag{3.15}$$

The properties of the error are given by

$$E(\varepsilon_i|X_i = x_i) = m(x_i) \tag{3.16}$$

$$Cov\big(\varepsilon_i, \varepsilon_j|X_i = x_i, X_j = x_j\big) = \begin{cases} \sigma^2(x_i), & i = j \\ 0 \quad , & i \neq j \end{cases} \quad i, j = 1, 2, 3, \dots . N \tag{3.17}$$

The functions $m(x_i)$ and $\sigma^2(x_i)$ are assumed to be smooth and strictly positive.

The proposed estimators are derived by modeling the finite population of $y_i's$, conditioned on the auxiliary variable x_i, as a realization from an infinite super population model ξ in which $Y_i = m(X_i) + \sigma(X_i)\varepsilon_i$ where ε_i are independent random variables, with mean zero and variance $v(x_i)$, $m(x_i)$ is a smooth function of x_i, and $v(x_i)$ is smooth and strictly positive. $E_\xi(Y_i|X_i) = m(x_i)$ is called the regression function, while $v(x_i) = var_\xi(Y_i)$ is called the variance function. The estimators of T are derived by noting that

$$\bar{T} = \sum_{i\in S} y_i + \left\{E\left(\sum_{i\in R} \bar{y}_i\right)\right\}$$

$$= \sum_{i \in S} y_i + \sum_{i \in R} \bar{m}(x_i) \tag{3.18}$$

Here it is observed that, $\sum_{i \in S} y_i$ is known while $\sum_{i \in R} y_i$ is unknown. The optimal predictor of this unknown quantity is

$$E\left(\sum_{i \in R} y_i\right) = \sum_{i \in R} m(x_i) \tag{3.19}$$

However $m(x_i)$ is unknown. An estimate of $m(x_i)$ is computed using the local linear procedure and then substituted in T in order to get local linear regression estimators of finite population totals defined as

$$\bar{T}_{LL} = \sum_{i \in S} y_i + \sum_{i \in R} \bar{m}_{LL}(x_i) \tag{3.20}$$

where $\bar{m}_{LL}(x_i)$ is a local linear estimator of $m(x_i)$ at point x_i

Letting x_j be any point in the non sample, and as in Dorfman (1992), we propose

$$\bar{T}_{LL} = \sum_{i \in S} Y_i + \sum_{j \in R} \bar{m}_{LL}\left(x_j\right) \tag{3.21}$$

as estimators of finite population total, where $\bar{m}_{LL}\left(x_j\right)$ is a local linear regression estimator of $m\left(x_j\right)$ at point x_j.

3.4 Construction of the local constant regression estimator of T

Local constant regression is a nonparametric conditional quantile estimation method where the order of the local polynomial being fit is equal to zero. The super population model considered for estimating the finite population total is given by (3.12). The assumptions stated in equations (3.13) (3.14) and (3.15) hold for the super population model considered in the nonparametric regression estimation of $m(x_i)$. The properties of the error are defined by equations (3.16) and (3.17). The functions $m(x_i)$ and $\sigma^2(x_i)$ are assumed to be smooth and strictly positive.

Consider the Taylor series expansion of $m(x_i)$ about x_j, which is

$$m(x_i) = m(x_j) + (x_i - x_j)m'(x_j) + \frac{(x_i - x_j)^2}{2!}m''(x_j) + \frac{(x_i - x_j)^3}{3!}m'''(x_j) \ldots.$$

with $x_i = x_j + ht$, so that

$$m(x_i) = m(x_j) + htm'(x_j) + \frac{h^2t^2}{2!}m''(x_j) + \frac{h^3t^3}{3!}m'''(x_j) + \frac{h^4t^4}{4!}m''''(x_j) + \cdots \quad (3.22)$$

The general form of the Taylor series expansion is expressed as

$$y_i = \alpha + (x_i - x_j)\beta + \varepsilon_i \quad (3.23)$$

where x_i lies in the interval $[x_j - h,\ x_j + h]$ and

$$\varepsilon_i = \frac{(x_i - x_j)^2}{2!}m''(x_j) + \frac{(x_i - x_j)^3}{3!}m'''(x_j) + \frac{(x_i - x_j)^4}{4!}m''''(x_j) + \cdots$$

The constants α and β are computed using the least squares procedure by making ε_i the subject of the formulae, squaring both sides, summing and applying the weights to obtain a solution to the weighted least squares problem of the form

$$\sum_{i \in S} \varepsilon_i^2 = \sum_{i \in S} \left(y_i - \alpha - \beta(x_i - x_j)\right)^2 K\left(\frac{x_i - x_j}{h}\right) \quad (3.24)$$

Noting that $x_i = x_j + ht$, implies that $t = \frac{x_i - x_j}{h}$ and therefore $K(t) = K\left(\frac{x_i - x_j}{h}\right)$

Letting,

$$\varphi = \sum_{i \in S} \left(y_i - \alpha - \beta(x_i - x_j)\right)^2 K\left(\frac{x_i - x_j}{h}\right) \quad (3.25)$$

and differentiating φ with respect to α and equating to zero, gives

$$\frac{\partial \varphi}{\partial \alpha} = \sum_{i \in S} -2\left(y_i - \alpha - \beta(x_i - x_j)\right) K\left(\frac{x_i - x_j}{h}\right) \left\{\left\{\sum_{i \in S} K\left(\frac{x_i - x_j}{h}\right)\right\}^{-1}\right\} = 0\,, \quad (3.26)$$

implying that

$$\sum_{i \in S} K\left(\frac{x_i - x_j}{h}\right) y_i = \alpha \sum_{i \in S} K\left(\frac{x_i - x_j}{h}\right) + \beta \sum_{i \in S} (x_i - x_j) K\left(\frac{x_i - x_j}{h}\right). \quad (3.27)$$

Letting

$$S_r(x_j;h) = \sum_{i \in S} K\left(\frac{x_i - x_j}{h}\right)(x_i - x_j)^r \qquad r = 0,1,2 \tag{3.28}$$

so that

$$S_0(x_j;h) = \sum_{i \in S} K\left(\frac{x_i - x_j}{h}\right)$$

$$S_1(x_j;h) = \sum_{i \in S} K\left(\frac{x_i - x_j}{h}\right)(x_i - x_j)$$

$$S_0(x_j;h) = \sum_{i \in S} K\left(\frac{x_i - x_j}{h}\right)(x_i - x_j)^2$$

Then it follows from equation (3.27) that

$$\sum_{i \in S} K\left(\frac{x_i - x_j}{h}\right) y_i = \alpha S_0(x_j;h) + \beta S_1(x_j;h). \tag{3.29}$$

In a similar way, differentiating φ with respect to β and equating to zero, gives

$$\frac{\partial \varphi}{\partial \beta} = \sum_{i \in S} -2\left(y_i - \alpha - \beta(x_i - x_j)\right)(x_i - x_j) K\left(\frac{x_i - x_j}{h}\right)\left\{\left\{\sum_{i \in S} K\left(\frac{x_i - x_j}{h}\right)\right\}^{-1}\right\} = 0, \tag{3.30}$$

Implying that

$$\sum_{i \in S}(x_i - x_j) K\left(\frac{x_i - x_j}{h}\right) y_i = \alpha \sum_{i \in S}(x_i - x_j) K\left(\frac{x_i - x_j}{h}\right) + \beta \sum_{i \in S}(x_i - x_j)^2 K\left(\frac{x_i - x_j}{h}\right). \tag{3.31}$$

and thus

$$\sum_{i \in S}(x_i - x_j) K\left(\frac{x_i - x_j}{h}\right) y_i = \alpha S_1(x_j;h) + \beta S_2(x_j;h). \tag{3.32}$$

Equation (3.29) is multiplied by $S_2(x_j;h)$ and equation (3.32) by $S_1(x_j;h)$ to obtain

$$S_2(x_j;h) \sum_{i \in S} K\left(\frac{x_i - x_j}{h}\right) y_i = \alpha S_0(x_j;h) S_2(x_j;h) + \beta S_1(x_j;h) S_2(x_j;h) \tag{3.33}$$

$$S_1(x_j;h) \sum_{i \in S}(x_i - x_j) K\left(\frac{x_i - x_j}{h}\right) y_i = \alpha \left(S_1(x_j;h)\right)^2 + \beta S_1(x_j;h) S_2(x_j;h) \tag{3.34}$$

Equation (3.34) is subtracted from equation (3.33) to obtain

$$S_2(x_j;h)\sum_{i\in S}K\left(\frac{x_i-x_j}{h}\right)y_i - S_1(x_j;h)\sum_{i\in S}(x_i-x_j)K\left(\frac{x_i-x_j}{h}\right)y_i$$
$$= \alpha S_0(x_j;h)S_2(x_j;h) - \alpha\left(S_1(x_j;h)\right)^2 \tag{3.35}$$

Making α the subject of the formulae, gives

$$\bar{\alpha} = \sum_{i\in S}\left\{\frac{\left(S_2(x_j;h) - S_1(x_j;h)(x_i-x_j)\right)}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2}K\left(\frac{x_i-x_j}{h}\right)y_i\right\} \tag{3.36}$$

In a similar manner, equation (3.29) is multiplied by $S_1(x_j;h)$ and equation (3.32) by $S_0(x_j;h)$ to obtain

$$S_1(x_j;h)\sum_{i\in S}K\left(\frac{x_i-x_j}{h}\right)y_i = \alpha S_0(x_j;h)S_1(x_j;h) + \beta\left(S_1(x_j;h)\right)^2 \tag{3.37}$$

$$S_0(x_j;h)\sum_{i\in S}(x_i-x_j)K\left(\frac{x_i-x_j}{h}\right)y_i = \alpha S_0(x_j;h)S_1(x_j;h) + \beta S_0(x_j;h)S_2(x_j;h) \tag{3.38}$$

Subtracting equation (3.38) from equation (3.37), gives

$$S_1(x_j;h)\sum_{i\in S}K\left(\frac{x_i-x_j}{h}\right)y_i - S_0(x_j;h)\sum_{i\in S}(x_i-x_j)K\left(\frac{x_i-x_j}{h}\right)y_i$$
$$= \beta\left(S_1(x_j;h)\right)^2 - \beta S_0(x_j;h)S_2(x_j;h) \tag{3.39}$$

Making β the subject of the formulae, gives

$$\bar{\beta} = \sum_{i\in S}\left\{\frac{\left(S_0(x_j;h)(x_i-x_j) - S_1(x_j;h)\right)}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2}K\left(\frac{x_i-x_j}{h}\right)y_i\right\} \tag{3.40}$$

Now it follows from equation (3.23) that

$$\bar{y}_i = \bar{\alpha} + (x_i - x_j)\bar{\beta} \tag{3.41}$$

If the value assigned is zero, assuming that $\bar{\beta}$ is a pre-assigned constant, then

$$\bar{y}_j = \bar{\alpha} \tag{3.42}$$

Therefore

$$\bar{m}(x_j) = \sum_{i\in S}\left\{\frac{\left(S_2(x_j;h) - S_1(x_j;h)(x_i-x_j)\right)}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2}K\left(\frac{x_i-x_j}{h}\right)y_i\right\}$$

$$= \sum_{i \in S} w_i(x_j) y_i \tag{3.43}$$

where

$$w_i(x_j) = \frac{\left(S_2(x_j;h) - S_1(x_j;h)(x_i - x_j)\right)}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2} K\left(\frac{x_i - x_j}{h}\right) y_i$$

implying that the local constant regression estimator of finite population total can be estimated using

$$\bar{T} = \sum_{i \in S} y_i + \sum_{j \in R} \bar{m}(x_j)$$

$$= \sum_{i \in S} y_i + \sum_{j \in R} \left\{ \sum_{i \in S} \left\{ \frac{\left(S_2(x_j;h) - S_1(x_j;h)(x_i - x_j)\right)}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2} K\left(\frac{x_i - x_j}{h}\right) y_i \right\} \right\} \tag{3.44}$$

3.5 Properties of the local constant regression estimator of *T*

In examining the properties of local constant regression estimators, the assumptions outlined respectively in section (3.1)and section (3.3) are considered. Fan (1993) imposed conditions on $K(.)$ and are only used for convenience in terms of technical arguments and thus can be relaxed.

3.5.1 The expectation of the local constant regression estimator of *T*

The expectation of $\bar{T}$ is

$$E(\bar{T}) = \sum_{i \in S} E(y_i) + \sum_{j \in R} \left\{ \sum_{i \in S} \left\{ \frac{\left(S_2(x_j;h) - S_1(x_j;h)(x_i - x_j)\right)}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2} k\left(\frac{x_i - x_j}{h}\right) E(y_i) \right\} \right\}$$

$$= \sum_{i \in S} m(x_i) + \sum_{j \in R} \left\{ \sum_{i \in S} \left\{ \frac{\left(S_2(x_j;h) - S_1(x_j;h)(x_i - x_j)\right)}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2} k\left(\frac{x_i - x_j}{h}\right) m(x_i) \right\} \right\} \tag{3.45}$$

Using the Taylor series expansion of the form (3.22), theorem 3 in Fan and Gijbels (1996) is such that under the conditions (i) to (v) given in section (3.1), allows

$$E(\bar{T}) = \sum_{i \in S} m(x_i) + \sum_{j \in R} \left\{ \sum_{i \in S} \left\{ \frac{S_2(x_j;h) k\left(\frac{x_i - x_j}{h}\right)}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2} \left(m(x_j) + htm'(x_j) + \frac{h^2t^2}{2!} m''(x_j) + \cdots \right) \right\} \right\}$$

$$- \sum_{j \in R} \left\{ \sum_{i \in S} \left\{ \frac{S_1(x_j;h)(x_i - x_j)}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2} k\left(\frac{x_i - x_j}{h}\right) \left(m(x_j) + htm'(x_j) + \frac{h^2t^2}{2!} m''(x_j) + \cdots \right) \right\} \right\}$$

$$= \sum_{i \in S} m(x_i) + \sum_{j \in R} \left\{ \left(\frac{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2} \right) m(x_j) \right\} + \sum_{j \in R} \left\{ \left(\frac{S_1(x_j;h)S_2(x_j;h) - S_1(x_j;h)S_2(x_j;h)}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2} \right) m'(x_j) \right\}$$

$$+ \sum_{j \in R} \left\{ \left(\frac{\left(S_2(x_j;h)\right)^2 - S_1(x_j;h)S_3(x_j;h)}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2} \right) \frac{m''(x_j)}{2!} \right\}$$

$$= \sum_{i \in S} m(x_i) + \sum_{j \in R} m(x_j) + \sum_{j \in R} \left\{ \left(\frac{\left(S_2(x_j;h)\right)^2 - S_1(x_j;h)S_3(x_j;h)}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2} \right) \frac{m''(x_j)}{2!} \right\}. \quad (3.46)$$

3.5.2 The bias of the local constant regression estimator of *T*

The bias of the local constant regression estimator, $\bar{T}$ is

$$Bias(\bar{T}) = \sum_{j \in R} \left\{ \left(\frac{\left(S_2(x_j;h)\right)^2 - S_1(x_j;h)S_3(x_j;h)}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2} \right) \frac{m''(x_j)}{2!} \right\}. \quad (3.47)$$

Therefore the asymptotic expression of the bias of the local constant regression estimator $\bar{T}$ is

$$Bias_{asy}(\bar{T}) = \sum_{j\in R}\left\{\frac{\left(n^2h^6{k_2}^2 + o(n^2h^8)\right)m''(x_j)}{2\left(n^2h^4k_2 + o(n^2h^6)\right)}\right\}$$

$$= \sum_{J\in R}\left\{\frac{1}{2}h^2k_2m''(x_j)\right\} \tag{3.48}$$

3.5.3 The variance of the local constant regression estimator of *T*

The variance of the local constant regression estimator $\bar{T}$ is estimated using the variance of the error, thus $Var(\bar{T} - T)$ is

$$Var(\bar{T}) = Var\left\{\sum_{i\in S} y_i + \sum_{j\in R}\bar{m}(x_j) - \sum_{i\in S} y_i - \sum_{j\in R} y_j\right\}$$

$$= Var\left\{\sum_{i\in S}\sum_{j\in R} w_i(x_j)y_i - \sum_{j\in R} y_j\right\}$$

$$= \sum_{j\in R}\sum_{i\in S} w_i^2(x_j)\,\sigma^2(x_i) + \sum_{j\in R}\sigma^2(x_j) \tag{3.49}$$

where,

$$w_i(x_j) = \frac{\left(S_2(x_j;h) - S_1(x_j;h)(x_i - x_j)\right)}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2}K\left(\frac{x_i - x_j}{h}\right).$$

Incorporating the results of $\bar{m}(x_j)$ so far derived, the asymptotic expression of the variance of $\bar{T}$ is expressed as

$$Var_{asy}(\bar{T}) = \frac{1}{nh}\sum_{j\in R}\sum_{i\in S}\left\{K^2\left(\frac{x_i - x_j}{h}\right)\sigma^2(x_i)\left(\frac{x_i - x_{i-1}}{h}\right)\right\}$$

$$= \sum_{j\in R}\frac{d_k}{nh}\sigma^2(x_j). \tag{3.50}$$

3.5.4 The MSE of the local constant regression estimator of *T*

Theorem I in Fan (1993) allows that under condition (ii) gives

$$MSE(\bar{T}) = \left(Bias(\bar{T})\right)^2 + Var(\bar{T})$$

$$= \left\{\sum_{j\in R}\left\{\left(\frac{\left(S_2(x_j;h)\right)^2 - S_1(x_j;h)S_3(x_j;h)}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2}\right)\frac{m''(x_j)}{2}\right\}\right\}^2 + \sum_{j\in R}\sum_{i\in S} w_i^2(x_j)\,\sigma^2(x_i)$$

$$+ \sum_{j\in R}\sigma^2(x_j) \qquad (3.51)$$

The asymptotic expression of the MSE of the local constant regression estimator $\bar{T}$ is

$$MSE_{asy}(\bar{T}) = \left\{\sum_{j\in R}\left\{\frac{1}{2}\,h^2k_2m''(x_j)\right\}\right\}^2 \qquad (3.52)$$

3.6 Construction of the local linear regression estimator of *T*

In this section, consider again the super population model for estimating the population total of the form (3.12). The assumptions stated in equations (3.13) (3.14) and (3.15) hold for the super population model considered in the nonparametric regression estimation of $m(x_i)$. The properties of the error are defined by equations (3.16) and (3.17). The functions $m(x_i)$ and $\sigma^2(x_i)$ are assumed to be smooth and strictly positive.

Using the Taylor series expansion, the expression of $m(x_i)$ is defined by (3.22) which is

$$m(x_i) = m(x_j) + (x_i - x_j)m'(x_j) + \frac{(x_i - x_j)^2}{2!}m''(x_j) + \frac{(x_i - x_j)^3}{3!}m'''(x_j) \ldots.$$

with $x_i = x_j + ht$, so that

$$m(x_i) = m(x_j) + htm'(x_j) + \frac{h^2t^2}{2!}m''(x_j) + \frac{h^3t^3}{3!}m'''(x_j) + \frac{h^4t^4}{4!}m''''(x_j) + \cdots$$

and the general form of the Taylor series expansion is defined by (3.23) which is

$$y_i = \alpha + (x_i - x_j)\beta + \varepsilon_i$$

where x_i lies in the interval $[x_j - h,\ x_j + h]$ and

$$\varepsilon_i = \frac{(x_i - x_j)^2}{2!} m''(x_j) + \frac{(x_i - x_j)^3}{3!} m'''(x_j) + \frac{(x_i - x_j)^4}{4!} m''''(x_j) + \cdots$$

Therefore, the task of estimating $m(x)$ is equivalent to the local linear regression task of estimating the intercept α. Now α and β are established in order to minimize

$$\sum_{i \in S} \left(y_j - \alpha - \beta(x_i - x_j)\right)^2 K\left(\frac{x_i - x_j}{h}\right) \tag{3.53}$$

to obtain least squares estimators of α and β

Let $\bar{\alpha}$ and $\bar{\beta}$ be the solution to the weighted least square problem (3.53). Deriving yields

$$\bar{\alpha} = \frac{\sum_{j=1}^{n} w_j y_j}{\sum_{j=1}^{n} w_j} \tag{3.54}$$

where w_j is defined in equation (3.56). Therefore, the local linear regression estimator is

$$\bar{m}_{LL}(x) = \bar{\alpha} = \frac{\sum_{j=1}^{n} w_j y_j}{\sum_{j=1}^{n} w_j} \tag{3.55}$$

where

$$w_j = K\left(\frac{x_i - x_j}{h}\right)\left(S_2(x_j; h) - (x_i - x_j)S_1(x_j; h)\right) \tag{3.56}$$

and

$$S_r(x_j; h) = \sum_{j=1}^{n} K\left(\frac{x_i - x_j}{h}\right)(x_i - x_j)^r, \quad r = 1,2 \tag{3.57}$$

We compute (3.55) and (3.56) as follows; by letting

$$Q = \sum_{j=1}^{n} \left(y_j - \alpha - \beta(x_i - x_j)\right)^2 K\left(\frac{x_i - x_j}{h}\right) \tag{3.58}$$

Differentiating (3.58) with respect to α, we get

$$\frac{\partial Q}{\partial \alpha} = \sum_{j=1}^{n} -2\left(y_j - \alpha - \beta\left(x_i - x_j\right)\right) K\left(\frac{x_i - x_j}{h}\right) \tag{3.59}$$

For the least value of Q, we have

$$\sum_{j=1}^{n} \left(y_j - \alpha - \beta\left(x_i - x_j\right)\right) K\left(\frac{x_i - x_j}{h}\right) = 0 \tag{3.60}$$

Implying that

$$\sum_{j=1}^{n} K\left(\frac{x_i - x_j}{h}\right) y_j = \alpha \sum_{j=1}^{n} K\left(\frac{x_i - x_j}{h}\right) + \beta \sum_{j=1}^{n} K\left(\frac{x_i - x_j}{h}\right)\left(x_i - x_j\right)$$

$$= \alpha S_0\left(x_j; h\right) + \beta S_1\left(x_j; h\right)$$

which is the same as (3.29).

Differentiating (3.58) with respect to β, we get

$$\frac{\partial Q}{\partial \beta} = \sum_{j=1}^{n} -2\left(y_j - \alpha - \beta\left(x_i - x_j\right)\right) K\left(\frac{x_i - x_j}{h}\right)\left(x_i - x_j\right) \tag{3.61}$$

For the least value of Q, we have

$$\sum_{j=1}^{n} \left(y_j - \alpha - \beta\left(x_i - x_j\right)\right) K\left(\frac{x_i - x_j}{h}\right)\left(x_i - x_j\right) = 0 \tag{3.62}$$

Implying that

$$\sum_{j=1}^{n} K\left(\frac{x_i - x_j}{h}\right)\left(x_i - x_j\right) y_j = \alpha \sum_{j=1}^{n} K\left(\frac{x_i - x_j}{h}\right)\left(x_i - x_j\right) + \beta \sum_{j=1}^{n} K\left(\frac{x_i - x_j}{h}\right)\left(x_i - x_j\right)^2$$

$$= \alpha S_1\left(x_j; h\right) + \beta S_2\left(x_j; h\right)$$

which is the same as (3.32).

Solving (3.29) and (3.32) simultaneously by the elimination method, yields respectively (3.33) and (3.34) thus

$$S_2(x_j;h)\sum_{j=1}^{n} K\left(\frac{x_i-x_j}{h}\right)y_j = \alpha S_0(x_j;h)S_2(x_j;h) + \beta S_1(x_j;h)S_2(x_j;h)$$

$$S_1(x_j;h)\sum_{j=1}^{n} K\left(\frac{x_i-x_j}{h}\right)(x_i-x_j)y_j = \alpha\left(S_1(x_j;h)\right)^2 + \beta S_1(x_j;h)S_2(x_j;h)$$

Now, eliminating β from (3.33) and (3.34), gives (3.36) thus

$$\bar{\alpha} = \frac{S_2(x_j;h)\sum_{j=1}^{n} K\left(\frac{x_i-x_j}{h}\right)y_j - S_{n,1}\sum_{j=1}^{n} K\left(\frac{x_i-x_j}{h}\right)(x_i-x_j)y_j}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2}$$

$$= \frac{\sum_{j=1}^{n}\left(S_2(x_j;h)K\left(\frac{x_i-x_j}{h}\right)y_j - S_1(x_j;h)K\left(\frac{x_i-x_j}{h}\right)(x_i-x_j)y_j\right)}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2}$$

$$= \frac{\sum_{j=1}^{n}\left(S_2(x_j;h) - (x_i-x_j)S_1(x_j;h)\right)K\left(\frac{x_i-x_j}{h}\right)y_j}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2}$$

$$= \sum_{j=1}^{n}\frac{\left(S_2(x_j;h) - (x_i-x_j)S_1(x_j;h)\right)}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2}K\left(\frac{x_i-x_j}{h}\right)y_j$$

which is the analogue of equation (3.55).

In a similar way, eliminating α from (3.33) and (3.34), we get

$$\bar{\beta} = \sum_{j=1}^{n}\frac{\left(S_0(x_j;h) - (x_i-x_j)S_1(x_j;h)\right)}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2}K\left(\frac{x_i-x_j}{h}\right)y_j \qquad (3.63)$$

where $S_r(x_j;h) = \sum_{i=1}^{n}(x_i-x_j)^r K\left(\frac{x_i-x_j}{h}\right) \quad r = 0,1,2$

Using the set of data provided, the estimator $\bar{\beta}$ is determined. Therefore from equation (3.23), we have

$$\bar{y}_i = \bar{\alpha} + (x_i-x_j)\bar{\beta} \qquad (3.64)$$

such that

$$\bar{M}_{LL}(x_j) = \left\{ \sum_{j=1}^{n} \frac{\left(S_2(x_j;h) - (x_i - x_j)S_1(x_j;h)\right)}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2} K\left(\frac{x_i - x_j}{h}\right) y_j \right\}$$

$$+(x_i - x_j) \sum_{j=1}^{n} \left\{ \frac{\left(S_0(x_j;h) - S_1(x_j;h)(x_i - x_j)\right)}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2} K\left(\frac{x_i - x_j}{h}\right) y_j \right\}$$

$$= \sum_{i \in S} w_i(x_j)\, y_j + (x_i - x_j) \sum_{i \in s} w_i'(x_j)\, y_j \tag{3.65}$$

where

$$w_i(x_j) = \frac{\left(S_2(x_j;h) - S_1(x_j;h)(x_i - x_j)\right)}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2} K\left(\frac{x_i - x_j}{h}\right) \tag{3.66}$$

and

$$w_i'(x_j) = \frac{\left(S_0(x_j;h) - S_1(x_j;h)(x_i - x_j)\right)}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2} K\left(\frac{x_i - x_j}{h}\right) \tag{3.67}$$

3.7 Properties of the local linear regression estimator of $\mathbf{m}(x)$

In examining the properties of the derived local linear regression estimators of $\mathrm{m}(x)$, the assumptions outlined respectively in section 3.1 and section 3.2 are considered. Fan (1993) imposed conditions on $K(.)$ and are only used for convenience in terms of the technical arguments and thus can be relaxed.

3.7.1 The expectation of the local linear regression estimator of $\mathbf{m}(x)$

The expectation of the local linear regression estimator $m(x)$ is

$$E\left(\bar{m}_{LL}(x_j)\right) = \sum_{i \in S} w_i(x_j)\, E(y_j) + (x_i - x_j) \sum_{i \in S} w_i'(x_j)\, E(y_j)$$

$$= \sum_{i \in S} \left\{ \frac{\left(S_2(x_j;h) - S_1(x_j;h)(x_i - x_j)\right)}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2} K\left(\frac{x_i - x_j}{h}\right) E(y_j) \right\}$$

$$+(x_i - x_j) \sum_{i \in S} \left\{ \frac{\left(S_0(x_j;h) - S_1(x_j;h)(x_i - x_j)\right)}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2} K\left(\frac{x_i - x_j}{h}\right) E(y_j) \right\} \quad (3.68)$$

Consider the Taylor series expansion (3.22) of the form

$$m(x_i) = m(x_j) + (x_i - x_j)m'(x_j) + \frac{(x_i - x_j)^2}{2!} m''(x_j) + \frac{(x_i - x_j)^3}{3!} m'''(x_j) + \cdots$$

so that with $x_i = x_j + ht$

$$E\left(\bar{m}_{LL}(x_j)\right) = \sum_{i \in S} \left\{ w_i(x_j) \left(m(x_j) + htm'(x_j) + \frac{h^2 t^2}{2} m''(x_j) + \cdots \right) \right\}$$

$$+(x_i - x_j) \sum_{i \in S} \left\{ w_i'(x_j) \left(m(x_j) + htm'(x_j) + \frac{h^2 t^2}{2} m''(x_j) + \cdots \right) \right\}$$

$$= \left\{ \frac{S_2(x_j;h)}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2} \right\} \left\{ S_0(x_j;h)m(x_j) + S_1(x_j;h)m'(x_j) \right.$$

$$\left. + \frac{S_2(x_j;h)}{2} m''(x_j) + \cdots \right\}$$

$$- \left\{ \frac{S_1(x_j;h)}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2} \right\} \left\{ S_1(x_j;h)m(x_j) + S_2(x_j;h)m'(x_j) \right.$$

$$\left. + \frac{S_3(x_j;h)}{2} m''(x_j) + \cdots \right\}$$

$$+ \left\{ \frac{(x_i - x_j)S_1(x_j;h)}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2} \right\} \left\{ S_1(x_j;h)m(x_j) + S_2(x_j;h)m'(x_j) \right.$$

$$\left. + \frac{S_3(x_j;h)}{2} m''(x_j) + \cdots \right\}$$

$$-\left\{\frac{(x_i-x_j)S_1(x_j;h)}{S_0(x_j;h)S_2(x_j;h)-\left(S_1(x_j;h)\right)^2}\right\}\left\{S_0(x_j;h)m(x_j)+S_1(x_j;h)m'(x_j)\right.$$

$$\left.+\frac{S_2(x_j;h)}{2}m''(x_j)+\cdots\right\}$$

$$=\left\{\frac{\left(S_0(x_j;h)S_2(x_j;h)-\left(S_1(x_j;h)\right)^2\right)+(x_i-x_j)\left(S_0(x_j;h)S_1(x_j;h)-S_0(x_j;h)S_1(x_j;h)\right)}{S_0(x_j;h)S_2(x_j;h)-\left(S_1(x_j;h)\right)^2}\right\}m(x_j)$$

$$+\left\{\frac{\left(S_1(x_j;h)S_2(x_j;h)-S_1(x_j;h)S_2(x_j;h)\right)+(x_i-x_j)\left(S_0(x_j;h)S_2(x_j;h)-\left(S_1(x_j;h)\right)^2\right)}{S_0(x_j;h)S_2(x_j;h)-\left(S_1(x_j;h)\right)^2}\right\}m'(x_j)$$

$$+\left\{\frac{\left(S_2(x_j;h)^2-S_1(x_j;h)S_3(x_j;h)\right)+(x_i-x_j)\left(S_0(x_j;h)S_3(x_j;h)-S_1(x_j;h)S_2(x_j;h)\right)}{S_0(x_j;h)S_2(x_j;h)-\left(S_1(x_j;h)\right)^2}\right\}\frac{m''(x_j)}{2}$$

$$=m(x_j)+(x_i-x_j)m'(x_j)$$

$$+\left\{\frac{\left(S_2(x_j;h)^2-S_1(x_j;h)S_3(x_j;h)\right)+(x_i-x_j)\left(S_0(x_j;h)S_3(x_j;h)-S_1(x_j;h)S_2(x_j;h)\right)}{S_0(x_j;h)S_2(x_j;h)-\left(S_1(x_j;h)\right)^2}\right\}\frac{m''(x_j)}{2}$$

(3.69)

3.7.2 The bias of the local linear regression estimator of $\mathbf{m}(\boldsymbol{x})$

The bias of the local linear regression estimator $\bar{m}_{LL}(x_j)$ is

$$Bias\left(\bar{m}_{LL}(x_j)\right)=(x_i-x_j)m'(x_j)$$

$$+\left\{\frac{\left(S_2(x_j;h)^2-S_1(x_j;h)S_3(x_j;h)\right)+(x_i-x_j)\left(S_0(x_j;h)S_3(x_j;h)-S_1(x_j;h)S_2(x_j;h)\right)}{S_0(x_j;h)S_2(x_j;h)-\left(S_1(x_j;h)\right)^2}\right\}\frac{m''(x_j)}{2}$$

(3.70)

Assuming that, $x_i's$ are fixed uniform design points in the interval $(0, 1)$, then the asymptotic expression of the bias of local linear regression estimator $\bar{m}_{LL}(x_j)$ can be obtained. According to Eubank and Speckman (1993) and Masry (1996), we have

$$\sum_{i\in S}(x_i - x_j)^l k\left(\frac{x_i - x_j}{h}\right) = nh^{l+1}k_l + o(nh^{l+3}) \tag{3.71}$$

is uniform for $x \in (0, 1)$ and $h \in H_n$, where $H_n = [C_1 n^{-E_1}, C_2 n^{-E_2}]$, $0 < E_2 < E_1 < 1$, and $C_1, C_2 > 0$.

This implies that

$$S_0(x_j; h) = nh + o(nh^3), \qquad S_1(x_j; h) = o(nh^4), \qquad S_2(x_j; h) = nh^3k_2 + o(nh^5),$$

$$S_3(x_j; h) = nh^4k_3 + o(nh^6) \quad and \quad S_4(x_j; h) = nh^5k_4 + o(nh^7)$$

such that

$$S_0(x_j; h)S_2(x_j; h) - \left(S_1(x_j; h)\right)^2 = \left(nh + o(nh^3)\right)\left(nh^3k_2 + o(nh^5)\right)\left(o(nh^4)\right)^2$$

$$= n^2h^4k_2 + o(n^2h^6) \tag{3.72}$$

$$S_2(x_j; h)^2 - S_1(x_j; h)S_3(x_j; h) = \left(nh^3k_2 + o(nh^5)\right)^2 - \left(o(nh^4)\right)\left(nh^4k_3 + o(nh^6)\right)$$

$$= n^2h^6{k_2}^2 + o(n^2h^8) \tag{3.73}$$

$$S_0(x_j; h)S_3(x_j; h) - S_1(x_j; h)S_2(x_j; h)$$

$$= \left(nh + o(nh^3)\right)\left(nh^4k_3 + o(nh^6)\right) - \left(o(nh^4)\right)\left(nh^3k_2 + o(nh^5)\right)$$

$$= n^2h^5k_3 + o(n^2h^7) \tag{3.74}$$

$$Bias_{asy}\left(\bar{m}_{LL}(x_j)\right) = (x_i - x_j)m'(x_j)$$

$$+\frac{\left(n^2h^6{k_2}^2 + o(n^2h^8) + (x_i - x_j)\left(n^2h^5k_3 + o(n^2h^7)\right)\right)m''(x_j)}{2\left(n^2h^4k_2 + o(n^2h^6)\right)}$$

$$= (x_i - x_j)m'(x_j) + \frac{h(h{k_2}^2 + (x_i - x_j)k_3)m''(x_j)}{2k_2} \tag{3.75}$$

3.7.3 The variance of the local linear regression estimator of $\mathbf{m}(\boldsymbol{x})$

The variance of the local linear regression estimator $\bar{m}_{LL}(x_j)$ is

$$Var\left(\bar{m}_{LL}(x_j)\right) = Var\left\{\sum_{i\in S} w_i(x_j)y_i + (x_i - x_j)\sum_{i\in S} w_i'(x_j)\, y_i\right\}$$

$$= \sum_{i\in S} w_i^2(x_j)\,\sigma^2(x_i) + (x_i - x_j)^2 \sum_{i\in S} w_i'^2(x_j)\sigma^2(x_i) \tag{3.76}$$

where

$$w_i^2(x_j) = \left\{\frac{\left(S_2(x_j;h) - S_1(x_j;h)(x_i - x_j)\right)}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2} k\left(\frac{x_i - x_j}{h}\right)\right\}^2, \tag{3.77}$$

and

$$w_i'^2(x_j) = \left\{\frac{\left(S_0(x_j;h)(x_i - x_j) - S_1(x_j;h)\right)}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2} k\left(\frac{x_i - x_j}{h}\right)\right\}^2 \tag{3.78}$$

The asymptotic expression of the variance of $\bar{m}_{LL}(x_j)$ is obtained as

$$w_i^2(x_j) = \left\{\left(S_2(x_j;h)K\left(\frac{x_i - x_j}{h}\right)\right.\right.$$

$$\left.\left. - S_1(x_j;h)(x_i - x_j)K\left(\frac{x_i - x_j}{h}\right)\right)\left(S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2\right)^{-1}\right\}^2$$

$$\approx \left\{\frac{1}{nh}K\left(\frac{x_i - x_j}{h}\right)\frac{\left(n^2h^4k_2 + o(n^2h^6)\right)}{\left(n^2h^4k_2 + o(n^2h^6)\right)}\right\}^2$$

$$\approx \frac{1}{n^2h^2}K^2\left(\frac{x_i - x_j}{h}\right) \tag{3.79}$$

$w_i'^2(x_j)$

$$= \left\{\frac{\left(S_0(x_j;h)S_1(x_j;h) - S_0(x_j;h)S_1(x_j;h)\right)}{\left(\left(S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2\right)\left(S_0(x_j;h)S_1(x_j;h) - S_0(x_j;h)S_1(x_j;h)\right)}\left(S_0(x_j;h)(x_i\right.\right.$$

$$\left.\left. - x_j)K\left(\frac{x_i - x_j}{h}\right) - S_1(x_j;h)K\left(\frac{x_i - x_j}{h}\right)\right)\right\}^2$$

$$\approx \left\{\frac{1}{nh}k\left(\frac{x_i - x_j}{h}\right)\frac{\left(o(n^2h^5) + o(n^2h^7) - o(n^2h^5) - o(n^2h^7)\right)}{\left(n^2h^4k_2 + o(n^2h^6)\right)}\right\}^2$$

$$\approx 0 \tag{3.80}$$

Then

$$Var_{asy}\left(\bar{M}_{LL}(x_j)\right) = \frac{1}{nh}\sum_{i\in S}K^2\left(\frac{x_i - x_j}{h}\right)\sigma^2(x_i)\left(\frac{x_i - x_{i-1}}{h}\right) + \left(x_i - x_j\right)^2\sum_{i\in S}0.\sigma^2(x_i)$$

$$= \sum_{j\in R}\frac{d_k}{nh}\sigma^2(x_j) \tag{3.81}$$

$$where \quad d_k = \int K^2(t)dt$$

3.7.4 The MSE of the local linear regression estimator of $\mathbf{m}(x)$

The MSE of the local linear regression estimator $\bar{m}_{LL}(x_j)$ is

$$MSE\left(\bar{m}_{LL}(x_j)\right) = \left\{Bias\left(\bar{m}_{LL}(x_j)\right)\right\}^2 + Var\left(\bar{m}_{LL}(x_j)\right)$$

$$= \left\{ (x_i - x_j)m'(x_j) \right.$$

$$\left. + \left(\frac{\left(S_2(x_j;h)^2 - S_1(x_j;h)S_3(x_j;h)\right) + (x_i - x_j)\left(S_0(x_j;h)S_3(x_j;h) - S_1(x_j;h)S_2(x_j;h)\right)}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2} \right) \frac{m''(x_j)}{2} \right\}^2$$

$$+ \sum_{i \in S} w_i^2(x_j)\,\sigma^2(x_i) + (x_i - x_j)^2 \sum_{i \in S} w_i'^2(x_j)\sigma^2(x_i) \qquad (3.82)$$

The asymptotic expression of the mean square error is also obtained using the asymptotic bias and asymptotic variance expressions of $\bar{m}_{LL}(x_j)$ such that

$$MSE_{asy}\left(\bar{m}_{LL}(x_j)\right) = \left\{(x_i - x_j)m'(x_j) + \frac{h\left(h{k_2}^2 + (x_i - x_j)k_3\right)m''(x_j)}{2K_2}\right\}^2 + \frac{d_k}{nh}\sigma^2(x_j) \qquad (3.83)$$

3.7.5 The unbiasedness and efficiency of the local linear regression estimator of $\mathbf{m}(x)$

The efficiency of an estimator refers to how much information it extracts about the parameter of interest from the sample. A more efficient estimator extracts more information, in some sense, from a sample of a given size. Efficiency measures information extracted by the variance of an unbiased estimator, that is, smaller variance means greater efficiency.

3.7.5.1 Introduction

An estimator is efficient if it is the minimum variance unbiased estimator. Let $X_1, .., X_n$ be a random sample from some distribution which depends on a parameter $m(x)$ and let $\bar{m}(x) =$

$\bar{m}(x)(X_1,..,X_n)$ be an estimator of $m(x)$. Then $\bar{m}(x)$ is an unbiased estimator of $m(x)$ if $E\big(\bar{m}(x)\big) = m(x)$. Thus $\bar{m}(x)$ is an asymptotically unbiased estimator of $m(x)$ if $\lim_{n\to\infty} E\big(\bar{m}(x)\big) = m(x)$. Further, $\bar{m}(x)$ is an efficient estimator of $m(x)$ if it is unbiased and its variance achieves the Cramer-Rao Lower Bound, that is if

$$Var(\bar{m}(x)\,) = \frac{1}{nI\big(m(x)\big)}\,, \tag{3.84}$$

The efficiency of an unbiased estimator $\bar{m}(x)$ of $m(x)$ is the ratio of the Cramer-Rao Lower Bound to the variance of the estimator; that is

$$Eff\big(\bar{m}(x)\big) = \frac{1/nI\big(m(x)\big)}{Var(\bar{m}(x))}\,. \tag{3.85}$$

We remark that it must be true that $Eff(\bar{m}(x)) \leq 1$. The smaller the value of the efficiency, the less efficient the estimator. Also $\bar{m}(x)$ is an asymptotically efficient estimator of $m(x)$ if it is unbiased or asymptotically unbiased such that

$$\lim_{n\to\infty} Eff\big(\bar{m}(x)\big) = 1. \tag{3.86}$$

In what follows, we make variance comparisons between the Nadaraya-Watson regression estimator and the proposed local linear regression estimator in terms of their asymptotic relative efficiency.

3.7.5.2 The asymptotic relative efficiency of the estimators of $m(x)$

The relative efficiency of two procedures is the ratio of their efficiencies, although often this concept is used where the comparison is made between a given procedure and a notional best possible procedure. The efficiencies and the relative efficiency of two procedures theoretically depend on the sample size available for the given procedure, but it is often possible to use the asymptotic relative efficiency, defined as the limit of the relative efficiencies as the sample size grows, as the principal comparison measure. If $\bar{m}_1$ and $\bar{m}_2$ are both unbiased estimators of m, then the relative efficiency of $\bar{m}_1$ to $\bar{m}_2$ is

$$RE(\bar{m}_1, \bar{m}_2) = \frac{Var(\bar{m}_2)}{Var(\bar{m}_1)}\,. \tag{3.87}$$

If $RE(\bar{m}_1, \bar{m}_2) < 1$, then $\bar{m}_2$ has a smaller variance than $\bar{m}_1$ and $\bar{m}_1$ is less efficient than $\bar{m}_2$. If $\bar{m}_1$ and $\bar{m}_2$ are both unbiased or asymptotically unbiased estimators of m, then the asymptotic relative efficiency of $\bar{m}_1$ to $\bar{m}_2$ is

$$ARE(\bar{m}_1, \bar{m}_2) = \lim_{n\to\infty} RE(\bar{m}_1, \bar{m}_2) = \lim_{n\to\infty} \frac{Var(\bar{m}_2)}{Var(\bar{m}_1)}\,. \tag{3.88}$$

Therefore, the mean regression functions, $m(x)$ for the Nadaraya-Watson regression estimator, the Dorfman regression estimator and the proposed local linear regression estimator are respectively given by

$$\bar{m}_{NW}(x_j) = \sum_{i=1}^{n} w_i(x) y_i\,. \tag{3.89}$$

$$\bar{m}_{Dorf}(x_j) = \sum_{i=1}^{n} w_i(x_j) y_i\,. \tag{3.90}$$

$$\bar{m}_{LL}(x_j) = \sum_{i\in S} w_i(x_j)\, y_j + (x_i - x_j) \sum_{i\in S} w_i'(x_j)\, y_j. \tag{3.91}$$

The variance of the Nadaraya-Watson regression estimator $\bar{m}(x_j)$ is

$$Var\left(\bar{m}_{NW}(x_j)\right) \quad = d_k \sigma^2(x_j) + \sum_{i\in s} \left\{ w_i^2(x_j) \left(ht\sigma^{2\prime}(x_j) + \frac{h^2 t^2}{2} \sigma^{2\prime\prime}(x_j) + \cdots \right) \right\} \tag{3.92}$$

The asymptotic expression for the variance of the Nadaraya-Watson regression estimator $\bar{m}_{NW}(x_j)$ is

$$Var_{asy}\left(\bar{m}_{NW}(x_j)\right) \approx \frac{d_k}{nh} \sigma^2(x_j)\,. \tag{3.93}$$

The variance of the local linear regression estimator $\bar{m}_{LL}(x_j)$ is

$$Var\left(\bar{m}_{LL}(x_j)\right) = \sum_{i\in S} w_i^2(x_j)\,\sigma^2(x_i) + \left(x_i - x_j\right)^2 \sum_{i\in S} w_i'^2(x_j)\sigma^2(x_i). \quad (3.94)$$

The asymptotic expression for the variance of the local linear regression estimator $\bar{m}_{LL}(x_j)$ is

$$Var_{asy}\left(\bar{m}_{LL}(x_j)\right) = \frac{d_k}{nh}\sigma^2(x_j)\,. \quad (3.95)$$

Thus the asymptotic relative efficiency of the Nadaraya–Watson regression estimator to the proposed local linear regression estimator is

$$ARE\left(\bar{m}_{NW}(x_j), \bar{m}_{LL}(x_j)\right) = \frac{Var_{asy}\left(\bar{m}_{LL}(x_j)\right)}{Var_{asy}\left(\bar{m}_{NW}(x_j)\right)} = \frac{\frac{d_k}{nh}\sigma^2(x_j)}{\frac{d_k}{nh}\sigma^2(x_j)} = 1\,. \quad (3.96)$$

The main objective was to obtain a consistent robust estimator using the procedure of local linear regression in model based surveys. The procedure is based on locally fitting a line rather than a constant. Unlike kernel regression, locally linear estimation would have no bias if the true model were linear. The resulting local linear estimator has minimal asymptotic variance in comparison with the Nadaraya-Watson estimator.

3.8 Properties of the local linear regression estimator of T

In investigating the properties of the derived local linear regression estimators of finite population total T, the assumptions outlined respectively in section (3.1)and section (3.3) are considered. Fan (1993) imposed conditions on $K(.)$ and can only be used for convenience in terms of the technical arguments and thus can be relaxed.

Therefore, using equation (3.21), the local linear regression estimator of finite population total T can be estimated as

$$\bar{T}_{LL} = \sum_{i\in S} y_i + \sum_{j\in R} \bar{m}_{LL}(x_j)$$

$$= \sum_{i\in S} y_i + \sum_{j\in R}\left\{\sum_{i\in S}\left\{\frac{\left(S_2(x_j;h) - S_1(x_j;h)(x_i - x_j)\right)}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2} K\left(\frac{x_i - x_j}{h}\right) y_i\right\}\right\}$$

$$+ \sum_{j\in R}\left\{\left(\frac{x_i - x_j}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2}\right)\sum_{i\in S}\left\{\left(S_0(x_j;h)(x_i - x_j)\right.\right.\right.$$

$$\left.\left.\left. - S_1(x_j;h)\right) k\left(\frac{x_i - x_j}{h}\right) y_i\right\}\right\} \tag{3.97}$$

3.8.1 The expectation of the local linear regression estimator of T

The expectation of the local linear regression estimator $\bar{T}_{LL}$ is

$$E(\bar{T}_{LL}) = \sum_{i\in S} E(y_i) + \sum_{j\in R}\left\{\sum_{i\in S}\left\{\frac{\left(S_2(x_j;h) - S_1(x_j;h)(x_i - x_j)\right)}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2} K\left(\frac{x_i - x_j}{h}\right) E(y_i)\right\}\right\}$$

$$+ \sum_{\boldsymbol{j\in R}}\left\{\left(\frac{x_i - x_j}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2}\right)\sum_{\boldsymbol{i\in S}}\left\{\left(S_0(x_j;h)(x_i - x_j)\right.\right.\right.$$

$$\left.\left.\left. - S_1(x_j;h)\right) K\left(\frac{x_i - x_j}{h}\right) E(y_i)\right\}\right\} \tag{3.98}$$

Using Taylor series expansion (3.22) which is

$$m(x_i) = m(x_j) + htm'(x_j) + \frac{h^2t^2}{2!}m''(x_j) + \frac{h^3t^3}{3!}m'''(x_j) + \cdots,$$

$$E(\bar{T}_{LL}) = \sum_{\boldsymbol{i\in S}} m(x_i) + \sum_{j\in R}\left\{\sum_{i\in S}\left\{\frac{\left(S_2(x_j;h) - S_1(x_j;h)(x_i - x_j)\right)}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2} K\left(\frac{x_i - x_j}{h}\right) m(x_i)\right\}\right\}$$

$$+\sum_{\boldsymbol{j\in R}}\left\{\left(\frac{x_i-x_j}{S_0(x_j;h)S_2(x_j;h)-\left(S_1(x_j;h)\right)^2}\right)\sum_{\boldsymbol{i\in S}}\left\{\left(S_0(x_j;h)(x_i-x_j)\right.\right.\right.$$

$$\left.\left.\left.-S_1(x_j;h)\right)K\left(\frac{x_i-x_j}{h}\right)m(x_i)\right\}\right\}$$

$$=\sum_{\boldsymbol{i\in S}}m(x_i)+\sum_{j\in R}\left\{\sum_{i\in S}\left\{\frac{S_2(x_j;h)\left(\frac{x_i-x_j}{h}\right)}{S_0(x_j;h)S_2(x_j;h)-\left(S_1(x_j;h)\right)^2}\left(m(x_j)+htm'(x_j)\right.\right.\right.$$

$$\left.\left.\left.+\frac{h^2t^2}{2!}m''(x_j)+\cdots\right)\right\}\right\}$$

$$-\sum_{j\in R}\left\{\sum_{i\in S}\left\{\frac{S_1(x_j;h)(x_i-x_j)}{S_0(x_j;h)S_2(x_j;h)-\left(S_1(x_j;h)\right)^2}K\left(\frac{x_i-x_j}{h}\right)\left(m(x_j)+htm'(x_j)\right.\right.\right.$$

$$\left.\left.\left.+\frac{h^2t^2}{2!}m''(x_j)+\cdots\right)\right\}\right\}$$

$$+\sum_{j\in R}\left\{\frac{(x_i-x_j)}{S_0(x_j;h)S_2(x_j;h)-\left(S_1(x_j;h)\right)^2}\sum_{i\in S}\left\{S_0(x_j;h)(x_i-x_j)K\left(\frac{x_i-x_j}{h}\right)\left(m(x_j)+htm'(x_j)\right.\right.\right.$$

$$\left.\left.\left.+\frac{h^2t^2}{2!}m''(x_j)+\cdots\right)\right\}\right\}$$

$$-\sum_{j\in R}\left\{\frac{(x_i-x_j)}{S_0(x_j;h)S_2(x_j;h)-\left(S_1(x_j;h)\right)^2}\sum_{i\in S}\left\{S_1(x_j;h)K\left(\frac{x_i-x_j}{h}\right)\left(m(x_j)+htm'(x_j)\right.\right.\right.$$

$$\left.\left.\left.+\frac{h^2t^2}{2!}m''(x_j)+\cdots\right)\right\}\right\}$$

$$= \sum_{i \in S} m(x_i) + \sum_{j \in R} \left\{ \left(\frac{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2} \right) m(x_j) \right\}$$

$$+ \sum_{j \in R} \left\{ \left(\frac{S_0(x_j;h)S_1(x_j;h) - S_0(x_j;h)S_1(x_j;h)}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2} \right) (x_i - x_j) m(x_j) \right\}$$

$$+ \sum_{j \in R} \left\{ \left(\frac{S_1(x_j;h)S_2(x_j;h) - S_1(x_j;h)S_2(x_j;h)}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2} \right) m'(x_j) \right\}$$

$$+ \sum_{j \in R} \left\{ \left(\frac{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2} \right) (x_i - x_j) m'(x_j) \right\}$$

$$+ \sum_{j \in R} \left\{ \left(\frac{\left(S_2(x_j;h)\right)^2 - S_1(x_j;h)S_3(x_j;h)}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2} \right) \frac{m''(x_j)}{2} \right\}$$

$$+ \sum_{j \in R} \left\{ \left(\frac{S_0(x_j;h)S_3(x_j;h) - S_1(x_j;h)S_2(x_j;h)}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2} \right) (x_i - x_j) \frac{m''(x_j)}{2} \right\}$$

$$= \sum_{i \in S} m(x_i) + \sum_{j \in R} m(x_j) + \sum_{j \in R} \{(x_i - x_j) m'(x_j)\}$$

$$+ \sum_{j \in R} \left\{ \left(\left(\frac{\left(S_2(x_j;h)\right)^2 - S_1(x_j;h)S_3(x_j;h)}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2} \right) \right.$$

$$\left. + (x_i - x_j) \left(\frac{S_0(x_j;h)S_3(x_j;h) - S_1(x_j;h)S_2(x_j;h)}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2} \right) \right) \frac{m''(x_j)}{2} \right\} \quad (3.99)$$

3.8.2 The bias of the local linear regression estimator of T

The bias of the local linear regression estimator $\bar{T}_{LL}$ is

$$Bias(\bar{T}_{LL}) = \sum_{j \in R} \{(x_i - x_j) m'(x_j)\}$$

$$+\sum_{j\in R}\left\{\left(\left(\frac{\left(S_2(x_j;h)\right)^2 - S_1(x_j;h)S_3(x_j;h)}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2}\right)\right.$$

$$\left.+\left(x_i - x_j\right)\left(\frac{S_0(x_j;h)S_3(x_j;h) - S_1(x_j;h)S_2(x_j;h)}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2}\right)\right)\frac{m''(x_j)}{2}\Bigg\} \qquad (3.100)$$

$$Bias_{asy}(\bar{T}_{LL}) = \sum_{j\in R}\{(x_i - x_j)m'(x_j)\}$$

$$+\sum_{j\in R}\left\{\frac{\{n^2h^5{k_2}^2 + o(n^2h^7) + (x_i - x_j)(n^2h^4k_3 + o(n^2h^6))\}m''(x_j)}{2(n^2h^3k_2 + o(n^2h^5))}\right\}$$

$$=\left\{\sum_{j\in R}(x_i - x_j)m'(x_j)\right\} + \sum_{j\in R}\left\{\frac{h(hk_2{}^2 + (x_i - x_j)k_3)m''(x_j)}{2k_2}\right\} \qquad (3.101)$$

3.8.3 The variance of the local linear regression estimator of *T*

The variance of the local linear regression estimator $\bar{T}_{LL}$ is

$$Var(\bar{T}_{LL}) = Var\left\{\sum_{i\in S} y_i + \sum_{j\in R}\bar{M}_{LL}(x_j) - \sum_{i\in S} y_i - \sum_{j\in R} y_j\right\}$$

$$= Var\left\{\sum_{i\in S}\sum_{j\in R} w_i(x_j)y_i + \sum_{j\in R}(x_i - x_j)\sum_{i\in S} w_i'(x_j)y_j - \sum_{j\in R} y_j\right\}$$

$$= \sum_{j\in R}\sum_{i\in S} w_i^2(x_j)\,\sigma^2(x_i) + \sum_{j\in R}(x_i - x_j)^2\sum_{i\in S} w_i'^2(x_j)\sigma^2(x_i) + \sum_{j\in R}\sigma^2(x_j) \qquad (3.102)$$

where,

$$w_i(x_j) = \frac{S_2(x_j;h) - S_1(x_j;h)(x_i - x_j)}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2}K\left(\frac{x_i - x_j}{h}\right)$$

$$w_i'(x_j) = \frac{S_0(x_j;h) - S_1(x_j;h)(x_i - x_j)}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2} K\left(\frac{x_i - x_j}{h}\right)$$

Applying the results of $\bar{M}_{LL}(x_j)$ so far derived, the asymptotic expression of the variance of $\bar{T}_{LL}$ is

$$Var_{asy}(\bar{T}_{LL}) = \frac{1}{nh}\sum_{j\in R}\sum_{i\in S} k^2\left(\frac{x_i - x_j}{h}\right)\sigma^2(x_i)\left(\frac{x_i - x_{i-1}}{h}\right) + \sum_{j\in R}(x_i - x_j)^2 \sum_{i\in S} 0.\sigma^2(x_i)$$

$$= \sum_{j\in R}\frac{d_k}{nh}\sigma^2(x_j) \qquad (3.103)$$

3.8.4 The MSE of the local linear regression estimator of *T*

The MSE of the local linear regression estimator $\bar{T}_{LL}$ is

$$MSE(\bar{T}_{LL}) = \{Bias(\bar{T}_{LL})\}^2 + Var(\bar{T}_{LL})$$

$$= \left\{\sum_{j\in R}(x_i - x_j)m'(x_j) + \sum_{j\in R}\left\{\left(\frac{\left(S_2(x_j;h)\right)^2 - S_1(x_j;h)S_3(x_j;h)}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2}\right) + (x_i - x_j)\left(\frac{S_0(x_j;h)S_3(x_j;h) - S_1(x_j;h)S_2(x_j;h)}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2}\right)\right\}\frac{m''(x_j)}{2}\right\}^2$$

$$= \sum_{j\in R}\sum_{i\in S} w_i^2(x_j)\,\sigma^2(x_i) + \sum_{j\in R}(x_i - x_j)^2 \sum_{i\in S} w_i'^2(x_j)\sigma^2(x_i) + \sum_{j\in R}\sigma^2(x_j) \qquad (3.104)$$

The asymptotic expression for the MSE of the local linear regression estimator $\bar{T}_{LL}$ is given by

$$MSE_{asy}(\bar{T}_{LL}) = \left\{\sum_{j\in R}(x_i - x_j)m'(x_j) + \sum_{j\in R}\left\{\frac{h\left(h{k_2}^2 + (x_i - x_j)k_3\right)m''(x_j)}{2k_2}\right\}\right\}^2$$

$$+\sum_{j\in R}\frac{d_k}{nh}\sigma^2(x_j) \tag{3.105}$$

3.8.5 The asymptotic relative efficiency of the estimators of *T*

The relative efficiency of two procedures is the ratio of their efficiencies, but it is often possible to use the asymptotic relative efficiency, defined as the limit of the relative efficiencies as the sample size grows, as the principal measure of comparison. Let $\bar{T}_0$ be the local constant regression estimator of finite population total and $\bar{T}_1$ be the local linear regression estimator of finite population total.

If $\bar{T}_0$ and $\bar{T}_1$ are both unbiased estimators of T, then the relative efficiency of $\bar{T}_0$ to $\bar{T}_1$ is

$$RE(\bar{T}_0, \bar{T}_1) = \frac{Eff(\bar{T}_1)}{Eff(\bar{T}_0)} = \frac{Var(\bar{T}_1)}{Var(\bar{T}_0)}. \tag{3.106}$$

If $\bar{T}_0$ and $\bar{T}_1$ are both asymptotically unbiased estimators of T, then the asymptotic relative efficiency of $\bar{T}_0$ to $\bar{T}_1$ is given by

$$ARE(\bar{T}_0, \bar{T}_1) = \lim_{n\to\infty} RE(\bar{T}_0, \bar{T}_1) = \lim_{n\to\infty}\frac{Var(\bar{T}_1)}{Var(\bar{T}_0)}. \tag{3.107}$$

Therefore, the finite population total functions for the local constant regression estimator $\bar{T}_0$ and for the local linear regression estimator $\bar{T}_1$ are respectively given by

$$\bar{T}_0 = \sum_{i\in S} y_i + \sum_{j\in R}\left\{\sum_{i\in S}\left\{\frac{\left(S_2(x_j;h) - S_1(x_j;h)(x_i - x_j)\right)}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2}K\left(\frac{x_i - x_j}{h}\right)y_i\right\}\right\}. \tag{3.108}$$

$$\bar{T}_1 = \sum_{i\in S} \mathrm{Y}_i + \sum_{j\in R}\left\{\sum_{i\in S}\left\{\frac{\left(S_2(x_j;h) - S_1(x_j;h)(x_i - x_j)\right)}{S_0(x_j;h)S_2(x_j;h) - \left(S_1(x_j;h)\right)^2}k\left(\frac{x_i - x_j}{h}\right)y_i\right\}\right\}$$

$$+\sum_{j\in R}\left\{\left(\frac{x_i-x_j}{S_0(x_j;h)S_2(x_j;h)-\left(S_1(x_j;h)\right)^2}\right)\sum_{i\in S}\left\{\left(S_0(x_j;h)(x_i-x_j)\right.\right.\right.$$

$$\left.\left.\left.-S_1(x_j;h)\right)k\left(\frac{x_i-x_j}{h}\right)y_i\right\}\right\}. \quad (3.109)$$

The variance of the local constant regression estimator $\bar{T}_0$ is

$$Var(\bar{T}_0) = \sum_{j\in R}\sum_{i\in S} w_i^2(x_j)\,\sigma^2(x_i) + \sum_{j\in R}\sigma^2(x_j) \quad (3.110)$$

The asymptotic expression for the variance of the local constant regression estimator $\bar{T}_0$ is

$$Var_{asy}(\bar{T}_0) = \sum_{j\in R}\frac{d_k}{nh}\sigma^2(x_j) \quad (3.111)$$

The variance of the local linear regression estimator $\bar{T}_1$ is

$$Var(\bar{T}_1) = \sum_{j\in R}\sum_{i\in S} w_i^2(x_j)\,\sigma^2(x_i) + \sum_{j\in R}(x_i-x_j)^2\sum_{i\in S} w_i'^2(x_j)\sigma^2(x_i) + \sum_{j\in R}\sigma^2(x_j) \quad (3.112)$$

The asymptotic expression for the variance of the local linear regression estimator $\bar{T}_1$ is

$$Var_{asy}(\bar{T}_1) = \sum_{j\in R}\frac{d_k}{nh}\sigma^2(x_j)\,. \quad (3.113)$$

Thus the asymptotic relative efficiency of the local constant regression estimator $\bar{T}_0$ to the local linear regression estimator $\bar{T}_1$ is

$$ARE(\bar{T}_0,\bar{T}_1) = \lim_{n\to\infty} RE(\bar{T}_0,\bar{T}_1) = \lim_{n\to\infty}\left\{\frac{Var_{asy}(\bar{T}_1)}{Var_{asy}(\bar{T}_0)}\right\} = \lim_{n\to\infty}\left\{\frac{\sum_{j\in R}\frac{d_k}{nh}\sigma^2(x_j)}{\sum_{j\in R}\frac{d_k}{nh}\sigma^2(x_j)}\right\} = 1 \quad (3.114)$$

3.9 Extension to stratified random sampling

3.9.1 Introduction

In stratified sampling, the population of N units is first divided into H sub populations of

$$N_1 + N_2 + N_3 + N_H = N \quad (3.115)$$

such that

$$\sum_{h=1}^{H} N_h = N \tag{3.116}$$

The sub populations are called strata. In order to obtain the full benefits from stratification, the values of N_h must be known. When the strata have been determined, a random sample is drawn from each stratum, the drawings being made independently in different strata. The sample sizes within the strata are denoted by $n_1, n_2, n_3, \ldots, n_H$, respectively.

If a simple random sample is taken in each stratum, the whole procedure is described as stratified random sampling. Stratification may produce a gain in precision in the estimates of characteristics of the whole population. It may be possible to divide a heterogeneous population into subpopulations, each of which is internally homogeneous. This is suggested by the name strata, with its implication of a division into layers. If each stratum is homogeneous, in the sense that the measurements vary little from one unit to another, a precise estimate of any stratum mean can be obtained from a small sample in that stratum. These estimates can then be combined into a precise estimate for the whole population. In what follows, we extend the use of the local linear regression to stratified random sampling.

3.9.2 The proposed estimator

In this section, the local linear regression estimators of finite population total under stratified random sampling are derived. Suppose this population consisting of N units is divided into H different strata of size $N_h, h = 1,2, \ldots, H$. Let $y_{hj},\ j = 1,2, \ldots, N_h$ be the survey measurement for the j^{th} unit in the h^{th} stratum. Further, let $x_{hj},\ j = 1,2, \ldots, N_h$ be the auxiliary measurement. A simple random sample of size n_h is selected without replacement, where n_h is sufficiently large with respect to N_h. Let also s_h be the sample values in the h^{th} stratum and r_h be the non sample values in the h^{th} stratum. Consider the super population model below

$$E(Y_{hj}) = m(x_{hj}) \tag{3.117}$$

$$Var(Y_{hj}) = \sigma^2(x_{hj}) \tag{3.118}$$

$$Cov(Y_{hj}, Y_{h'j'}) = \begin{cases} \sigma^2(x_{hj}), & h = h', \quad j = j' \\ 0, & otherwise \end{cases} \tag{3.119}$$

The functions $m(x_{hj})$ and $\sigma^2(x_{hj})$ are assumed to be smooth and strictly positive.

This model implies that

$$Y_{hj} = m(x_{hj}) + \varepsilon_{hj} \tag{3.120}$$

where properties of the error terms are

$$E(\varepsilon_{hj}) = 0 \tag{3.121}$$

$$Cov(\varepsilon_{hj}, \varepsilon_{h'j'}) = \begin{cases} \sigma^2(x_{hj}), & h = h', \quad j = j' \\ 0, & otherwise \end{cases} \tag{3.122}$$

We now use the local linear regression procedure to estimate $m(x_{hj})$ using (3.91). We denote our local linear regression estimator by $\bar{m}_{LL}(x_{hj})$. We assume that the auxiliary (prior) information is available for the entire finite population. We let k denote the kernel which is a continuous, bounded real function that integrates to one, that is, $\int k(u)du = 1$. Further, let the kernel weight in the h^{th} stratum be

$$w_{hj}(x_{hi}) = \frac{k\left(\frac{x_{hj} - x_{hi}}{b}\right)}{\sum_s k\left(\frac{x_{hj} - x_{hi}}{b}\right)} \quad i, j = 1,2, \dots, N_h, \;\; h = 1,2, \dots, L \tag{3.123}$$

where $\sum_s(.)$ implies summation over all the sampled units and $w_{hj}(.)$ is a symmetric density function while b is the bandwidth that determines how large a neighborhood of the target point is used to calculate the local average. This form of weight is suggested by Nadaraya (1964) and Watson (1964) and it is such that $\sum_s w_{hj}(x_h) = 1$.

Using this idea, we suggest a local linear polynomial regression estimator of the non sampled $y_{hj}'s$ in the h^{th} stratum which is

$$\bar{m}_{LL}(x_{hj}) = e_1'(X_{hj}' W_{hj} X_{hj})^{-1} X_{hj}' W_{hj} X_{hj} = \bar{y}_{hj} \tag{3.124}$$

where $e_1 = (1,0,0,\dots,0)'$ is a column vector of length $p+1$, $W_{hj} = diag\left(k(x_{hj} - x_{hi})\right)$ and $X_{hj} = \left(1, (x_{hj} - x_{hi}), \dots, (x_{hj} - x_{hi})^p\right)$

In order to estimate the population total of the non sampled units in the h^{th} stratum, we let $x = x_{hi}$ be any point in the non sampled units. Then it follows from (3.124) that

$$\bar{m}_{LL}(x_{hj}) = \sum_s \left\{ \frac{k\left(\frac{x_{hj} - x_{hi}}{b}\right)}{\sum_s k\left(\frac{x_{hj} - x_{hi}}{b}\right)} \right\} \bar{y}_{hj} \tag{3.125}$$

We denote the local linear regression estimator of finite population total, T as $\bar{T}_{LL}$ and hence the local linear regression estimators of finite population total within stratum h are denoted by $\bar{T}_{LLh}$. In stratum h, the population total can be partitioned into observed and unobserved components, assuming that the sample from the h^{th} stratum is rearranged so that $N_h - n_h$ are non-sample values, then

$$T_{LLh} = \sum_{j=1}^{n_h} y_{hj} + \sum_{j=n_h+1}^{N_h} y_{hj}$$

$$= y_{hs} + y_{hr} \tag{3.126}$$

$$where\ \ y_{hs} = \sum_{j=1}^{n_h} y_{hj} \quad and \quad y_{hr} = \sum_{j=n_h+1}^{N_h} y_{hj} \tag{3.127}$$

therefore the local linear estimator of finite population total within stratum h is

$$\bar{T}_{LLh} = y_{hs} + \sum_{j=n_h+1}^{N_h} E(y_{hj})$$

$$= y_{hs} + \sum_{j=n_h+1}^{N_h} \bar{m}_{LL}(x_{hj})$$

$$= y_{hs} + \bar{m}_{LL}(x_{hr}) \tag{3.128}$$

$$where, \quad \bar{m}_{LL}(x_{hr}) = \sum_{j=n_h+1}^{N_h} \bar{m}_{LL}(x_{hj}) \tag{3.129}$$

The local linear regression estimator of finite population total under stratified sampling is

$$\bar{T}_{LL} = \sum_{\mathrm{h}=1}^{H} \bar{T}_{LLh}$$

$$= \sum_{\mathrm{h}=1}^{H} y_{hs} + \sum_{\mathrm{h}=1}^{H} \sum_{j=n_h+1}^{N_h} \bar{m}_{LL}(x_{hj})$$

$$= \sum_{\mathrm{h}=1}^{H} \sum_{j=1}^{n_h} y_{hj} + \sum_{\mathrm{h}=1}^{H} \sum_{j=n_h+1}^{N_h} \bar{m}_{LL}(x_{hj}) \quad (3.130)$$

3.9.3 Properties of the local linear regression estimator under stratified sampling

In this section we explore properties of the estimator of finite population total derived using the local linear procedure. The necessary and sufficient conditions outlined by Fan (1993) and Ruppert and Wand (1994) to investigate the properties of the estimator are considered. Specifically, these assumptions are stated in section (3.1)and section (3.3).

The function for the finite population total is given as

$$T = \sum_{h=1}^{L} y_h \quad (3.131)$$

where y_h is the sum of all the units in stratum h, that is

$$y_h = \sum_{j=1}^{N_h} y_{hj} \quad (3.132)$$

Therefore, we define the prediction error for estimating the finite population total within stratum h by

$$\bar{T}_{LLh} - T = \sum_{h=1}^{L} (y_{hs} + \bar{y}_{hr}) - \sum_{h=1}^{L} y_h$$

$$= \sum_{h=1}^{L} \left(y_{hs} + \bar{y}_{hr} - (y_{hs} + y_{hr})\right)$$

$$= \sum_{h=1}^{L} (\bar{y}_{hr} - y_{hr})$$

$$= \sum_{h=1}^{L} \left\{ \sum_{s} w_{hj}(x_{hi}) y_{hj} - y_{hr} \right\} \tag{3.133}$$

where $y_H = y_{hs} + y_{hr}$, and y_{hs} is the sum of all the sampled units whereas y_{hr} is the sum of all the non sampled units in the h^{th} stratum.

But noting that

$$\sum_s w_{hj}(x_{hi}) y_{hj} = \frac{k\left(\frac{x_{h1}-x_{hi}}{b}\right)}{\sum_s k\left(\frac{x_{h1}-x_{hi}}{b}\right)} y_{h1} + \frac{k\left(\frac{x_{h2}-x_{hi}}{b}\right)}{\sum_s k\left(\frac{x_{h2}-x_{hi}}{b}\right)} y_{h2} + .. + \frac{k\left(\frac{x_{hn_h}-x_{hi}}{b}\right)}{\sum_s k\left(\frac{x_{hn_h}-x_{hi}}{b}\right)} y_{hn_h} \tag{3.134}$$

we have

$$\bar{T}_{LLh} - T = \sum_{h=1}^{L} (\bar{y}_{hr} - y_{hr})$$

$$= \sum_{h=1}^{L} \left\{ \frac{k\left(\frac{x_{h1} - x_{hi}}{b}\right)}{\sum_s k\left(\frac{x_{h1} - x_{hi}}{b}\right)} y_{h1} + \frac{k\left(\frac{x_{h2} - x_{hi}}{b}\right)}{\sum_s k\left(\frac{x_{h2} - x_{hi}}{b}\right)} y_{h2} + \cdots + \frac{k\left(\frac{x_{hn_h} - x_{hi}}{b}\right)}{\sum_s k\left(\frac{x_{hn_h} - x_{hi}}{b}\right)} y_{hn_h} - y_{hr} \right\} \tag{3.135}$$

which is simply

$$\bar{T}_{LLh} - T = \sum_{h=1}^{L} \left\{ \sum_{s} \frac{k\left(\frac{x_{hj} - x_{hi}}{b}\right)}{\sum_s k\left(\frac{x_{hj} - x_{hi}}{b}\right)} y_{hj} - \mathrm{y_{hr}} \right\} \tag{3.136}$$

Since $Y_{hj} = m(x_{hj}) + \varepsilon_{hj}$, and given that $E(Y_{hj}) = m(x_{hj})$, then it follows that

$$E(\bar{T}_{LLh} - T)$$

$$= \sum_{h=1}^{L} \left\{ \frac{k\left(\frac{x_{h1} - x_{hi}}{b}\right)}{\sum_s k\left(\frac{x_{h1} - x_{hi}}{b}\right)} m_{h1} + \frac{k\left(\frac{x_{h2} - x_{hi}}{b}\right)}{\sum_s k\left(\frac{x_{h2} - x_{hi}}{b}\right)} m_{h2} + \cdots + \frac{k\left(\frac{x_{hn_h} - x_{hi}}{b}\right)}{\sum_s k\left(\frac{x_{hn_h} - x_{hi}}{b}\right)} \right\} m(x_{\mathrm{hj}})$$

$$- m(x_{\mathrm{hr}})$$

$$= \sum_{h=1}^{L} \left\{ \sum_{s} \frac{k\left(\frac{x_{hj} - x_{hi}}{b}\right)}{\sum_s k\left(\frac{x_{hj} - x_{hi}}{b}\right)} m(x_{\mathrm{hj}}) - m(x_{\mathrm{hr}}) \right\} \tag{3.137}$$

Equation (3.137) is the prediction bias associated with $\bar{T}_{LLh}$. In order to approximate this bias, we apply a Taylor series expansion about a point x_h and assume that $m(x_{hj})$ is smooth, $n_h \to \infty$ and $b \to 0$. Then it is observed that

$$\bar{m}_{LL}(x_{hj}) \approx m(x_{\mathrm{h}}) + m'(x_h)(x_{hj} - x_h) + \frac{1}{2} m''(x_h)(x_{hj} - x_h)^2 \tag{3.138}$$

Letting $u = \frac{x_{hj} - x_h}{b}$ so that $bu = x_{hj} - x_h$, then it implies that

$$\bar{m}_{LL}(x_{hj}) \approx m(x_{\mathrm{h}}) + m'(x_h)bu + \frac{1}{2} m''(x_h)\mathrm{b}^2 u^2 \tag{3.139}$$

Therefore equation (3.137) becomes

$$E(\bar{T}_{LLh} - T) \approx \sum_{h=1}^{L} \left\{ \sum_{s} \frac{k(u)}{\sum_s k(u)} \left(m(x_{\mathrm{h}}) + m'(x_h)bu + \frac{1}{2} m''(x_h)\mathrm{b}^2 u^2 \right) - m(x_{\mathrm{hr}}) \right\}$$

$$\approx \sum_{h=1}^{L} \left\{ m(x_{\mathrm{h}}) \sum_{s} \frac{k(u)}{\sum_s k(u)} + bm'(x_h) \sum_{\mathrm{s}} \frac{uk(u)}{\sum_s k(u)} + \frac{\mathrm{b}^2 m''(x_h)}{2} \sum_{\mathrm{s}} \frac{u^2 k(u)}{\sum_s k(u)} - m(x_{\mathrm{hr}}) \right\} \tag{3.140}$$

Equation (3.140) can be expressed as

$$E(\bar{T}_{LLh} - T)$$

$$= \sum_{h=1}^{L} \left\{ m(x_{\mathrm{h}}) + m'(x_h) \left(\frac{n_h b^2 \phi_1 + O(n_h b^4)}{n_h b \phi_0 + O(n_h b^3)} \right) + \frac{1}{2} m''(x_h) \left(\frac{n_h b^3 \phi_2 + O(n_h b^5)}{n_h b \phi_0 + O(n_h b^3)} \right) - m(x_{\mathrm{hr}}) \right\}$$

$$= \sum_{h=1}^{L} \left\{ m(x_{\mathrm{h}}) + m'(x_h) \left(b \frac{\phi_1}{\phi_0} + O(b^3) \right) + \frac{1}{2} m''(x_h) \left(b^2 \frac{\phi_2}{\phi_0} + O(b^4) \right) - m(x_{\mathrm{hr}}) \right\}$$

$$= \sum_{h=1}^{L} \left\{ m(x_{\mathrm{h}}) + m'(x_h) \left(b \frac{\phi_1}{\phi_0} + O(b^3) \right) + \frac{1}{2} m''(x_h) \left(b^2 \frac{\phi_2}{\phi_0} + O(b^4) - m(x_{\mathrm{hr}}) \right) \right\} \tag{3.141}$$

Noting that, $\phi_0 = \int_0^1 k(u)du = 1$, $\phi_1 = \int_0^1 uk(u)du = 0$ and $\phi_2 = \int_0^1 u^2k(u)du > 0$, then the bias for estimating the finite population total within stratum h is given by

$$E(\bar{T}_{LLh} - T) = \sum_{h=1}^{L}\left\{m(x_h) + m'(x_h)O(b^3) + \frac{1}{2}m''(x_h)\left(b^2\phi_2 + O(b^4)\right) - m(x_{\mathrm{hr}})\right\}$$

$$= \sum_{h=1}^{L}\left\{m(x_h) + m'(x_h)O(b^3) + \frac{1}{2}m''(x_h)\left(b^2\int_0^1 u^2k(u)du + O(b^4)\right) - m(x_{\mathrm{hr}})\right\} \tag{3.142}$$

We deduce from equation (3.142), that the bias is given by

$$\frac{b^2}{2}\left\{\frac{1}{N}\sum_{\mathrm{h=1}}^{L} m''(x_h)\int_0^1 u^2k(u)du\right\} + O(b^4) \tag{3.143}$$

Thus as $b \to 0$, $E(\bar{T}_{LLh} - T) \to 0$. Therefore the bias of $\bar{T}_{LLh}$ is asymptotically design unbiased estimator of the finite population total, T.

The asymptotic mean square error of the estimator for finite population total $\bar{T}_{LLh}$ is defined by

$$MSE(\bar{T}_{LLh}) = Var(\bar{T}_{LLh}) + \left(B_m(\bar{T}_{LLh})\right)^2 \tag{3.144}$$

Considering theorem 1 of Fan (1993) and the assumptions therein stated, we have

$$MSE(\bar{T}_{LLh}) \approx \frac{1}{b}\frac{1}{N^2}\left\{\sum_{h=1}^{L}\frac{1}{n_h}\sigma^2(x_{hi})C_k\right\} + \left\{\frac{b^2}{2}\left(\frac{1}{N}\sum_{h=1}^{L} m''_{LL}(x_{hi})\int_0^1 u^2k(u)du\right)\right\}^2$$

$$\approx \frac{1}{b}\frac{1}{N^2}\left\{\sum_{h=1}^{L}\frac{1}{n_h}\sigma^2(x_{hi})C_k\right\} + \frac{b^4}{4}\left\{\frac{1}{N}\sum_{h=1}^{L} m''_{LL}(x_{hi})d_k\right\}^2 \tag{3.145}$$

where $d_k = \int_0^1 u^2k(u)du$, $C_k = \int k_b^2(u)\,du$

Clearly it is seen that equation (3.145) approaches zero in such a way that if $b \to 0$ and $nb \to \infty$, then $MSE(\bar{T}_{LLh}) \to 0$

3.10 Extension to two stage cluster sampling

3.10.1 Introduction

Let $Y_1, Y_2, \dots, Y_M$ denote the finite population elements and $X_1, X_2, \dots, X_M$ denote the associated covariates. Consider this finite population (T) of size N primary units or clusters labeled $\underline{U} = (U_1, U_2, \dots, U_N)$. Let M_i , $i = 1,2, \dots, N$ be the number of secondary units in the i^{th} primary unit. In this case, N is assumed to be known, but the $M_i's$ are unknown before sampling. Let $y_{ij}, i = 1,2, \dots, N,\ j = 1,2, \dots, M_i$ be the value of the interest variable for the j^{th} secondary unit belonging to the i^{th} primary unit. The relationship between the study variable and the auxiliary variable is described by the two stage nonparametric super population model which is

$$y_{ij} = m(x_{ij}) + \varepsilon_{ij}, \quad i = 1,2, \dots, n; \quad j = 1,2, \dots, M_i \tag{3.146}$$

The following assumptions hold for the model considered in the nonparametric regression estimation of finite population total

$$E_m(y_{ij}) = m(x_{ij}),\ Var_m(y_{ij}) = \sigma_\mu^2 + \sigma_\varepsilon^2, \qquad Cov_m(y_{ij}) = 0, \tag{3.147}$$

where E_m, var_m and Cov_m denote the expectation, the variance and the covariance under the model distribution.

3.10.2 The proposed estimator

The main purpose is to estimate the finite population total under two stage cluster sampling design which is

$$T = \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij} \tag{3.148}$$

where N is the number of primary sampling units (clusters), $M_i's$ are unknown before sampling and y_{ij} is the value of the survey variable of interest for the j^{th} secondary unit of the i^{th} cluster. Suppose that the selected primary units are n and the selected secondary units are m_i, $m =$

$\sum_{i=1}^{n} m_i$. Assuming the two stage nonparametric regression model (3.146) and using local linear regression procedure, we suggest an estimator of the finite population total of the form

$$\bar{T}_{LL} = \sum_{i=1}^{n}\sum_{j=1}^{m_i} y_{ij} + \sum_{i=1}^{n}\sum_{j=m_i+1}^{M_i} y_{ij} + \sum_{i=n+1}^{N}\sum_{j=1}^{M_i} y_{ij} \tag{3.149}$$

The first component of the finite population total $\bar{T}_{LL}$ represents the population total for the observed items, the second component represents the population total for the unobserved items in sample clusters and the third component represents the population total for the unobserved items in non sample clusters.

The second and third components of finite population total $\bar{T}_{LL}$ are estimated respectively using

$$\sum_{i=1}^{n}\sum_{j=m_i+1}^{M_i} \bar{y}_{ij} = \sum_{i=1}^{n}\sum_{j=m_i+1}^{M_i} \bar{m}_i(x_{ij}) \tag{3.150}$$

and

$$\sum_{i=n+1}^{N}\sum_{j=1}^{M_i} \bar{y}_{ij} = \sum_{i=n+1}^{N}\sum_{j=1}^{M_i} \bar{m}_i(x_{ij}) \tag{3.151}$$

Therefore estimation of the mean regression function $m(x)$ for each of the selected i^{th} cluster is obtained using the procedure of local linear regression. Now under local linear regression, and for each of the selected i^{th} cluster, we have

$$X = \begin{pmatrix} 1 & x - x_1 \\ 1 & x - x_2 \\ & \vdots \\ 1 & x - x_n \end{pmatrix} \qquad \underline{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

and

$$W = \begin{bmatrix} K_b(x - X_1) & 0 & \dots & 0 \\ 0 & K_b(x - X_2) & \cdots & 0 \\ & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & K_b(x - X_n) \end{bmatrix},$$

Thus, the estimation of $\beta(x)$ for each of the selected i^{th} cluster is

$$\bar{\beta}_i(x) = (X_i' W_i X_i)^{-1} X_i' W_i Y_i \tag{3.152}$$

and the local linear regression estimator of $m(x)$ for each of the i^{th} selected cluster is

$$\bar{m}_i(x) = \bar{\beta}_{i0} = e_1'(X_i' W_i X_i)^{-1} X_i' W_i Y_i \tag{3.153}$$

3.10.3 Properties of the local linear regression estimator of $\boldsymbol{m(x)}$

In order to examine the properties of the estimator (3.153), we need assumptions outlined in section (3.10.1) and the sample size n to be sufficiently large. It therefore follows from the definition of the estimator that the expectation of this estimator is given by

$$E\big(\bar{m}_i(x)\big) = e_1'(X_i' W_i X_i)^{-1} X_i' W_i M \tag{3.154}$$

where the vector $M = \big(m(x_1), m(x_2), \dots, m(x_n)\big)'$ contains the true regression function values at each of the $x_i's$. We denote the point at which m is being estimated simply by x_j. Therefore, for the local linear regression, we have

$$X_i = \begin{pmatrix} 1 & (x_1 - x_j) \\ 1 & (x_2 - x_j) \\ & \vdots \\ 1 & (x_n - x_j) \end{pmatrix} \tag{3.155}$$

Now, by Taylor's theorem, for any $x_j \in (0, 1)$, we can write

$$m_i(x_i) = m_i(x_j) + (x_i - x_j) m_i'(x_j) + \frac{(x_i - x_j)^2}{2!} m_i''(x_j) + \frac{(x_i - x_j)^3}{3!} m_i'''(x_j) + . \tag{3.156}$$

so that

$$M = X_i \begin{pmatrix} m_i(x_j) \\ m_i'(x_j) \end{pmatrix} + \frac{1}{2!} m_i''(x_j) \begin{pmatrix} 1 & (x_1 - x_j)^2 \\ 1 & (x_2 - x_j)^2 \\ & \vdots \\ 1 & (x_n - x_j)^2 \end{pmatrix} + \frac{1}{3!} m_i'''(x_j) \begin{pmatrix} 1 & (x_1 - x_j)^3 \\ 1 & (x_2 - x_j)^3 \\ & \vdots \\ 1 & (x_n - x_j)^3 \end{pmatrix} + \cdots$$

(3.157)

The first term in the expansion of $\bar{m}_i(x_j)$ which is the true regression function is

$$e_1'(X_i'W_iX_i)^{-1}(X_i'W_iX_i)\begin{pmatrix} m_i(x_j) \\ m_i'(x_j) \end{pmatrix} = e_1'\begin{pmatrix} m_i(x_j) \\ m_i'(x_j) \end{pmatrix} = m_i(x_j) \tag{3.158}$$

The bias of the estimator $\bar{m}_i(x_j)$ is

$$E\left[\bar{m}_i(x_j)\right] - m_i(x_j) = e_1'(X_i'W_iX_i)^{-1}(X_i'W_i)\frac{1}{2!}m_i''(x_j)\begin{pmatrix} 1 & (x_1 - x_j)^2 \\ 1 & (x_2 - x_j)^2 \\ & \vdots \\ 1 & (x_n - x_j)^2 \end{pmatrix} + \cdots$$

$$= e_1'(X_i'W_iX_i)^{-1}(n^{-1})(X_i'W_i)\frac{1}{2!}m_i''(x_j)\begin{pmatrix} 1 & (x_1 - x_j)^2 \\ 1 & (x_2 - x_j)^2 \\ & \vdots \\ 1 & (x_n - x_j)^2 \end{pmatrix} + \cdots \tag{3.159}$$

We note that if m_i is a linear function, then $m_i^{(r)}(x_j) = 0 \quad \forall\, r \geq 2$ so that the local linear estimator is exactly unbiased when m_i is a linear function. In order to find the leading bias term for the general functions m_i, we note that

$$(n^{-1})(X_i'W_iX_i) = \begin{pmatrix} \hat{S}_0(x_j; h) & \hat{S}_1(x_j; h) \\ \hat{S}_1(x_j; h) & \hat{S}_2(x_j; h) \end{pmatrix} \tag{3.160}$$

and

$$(n^{-1})(X_i'W_i) = \begin{pmatrix} (x_1 - x_j)^2 \\ (x_2 - x_j)^2 \\ . \\ . \\ (x_n - x_j)^2 \end{pmatrix} = \begin{pmatrix} \hat{S}_2(x_j; h) \\ \hat{S}_3(x_j; h) \end{pmatrix} \quad (3.161)$$

where $\hat{S}_r(x_j; h) = n^{-1} \sum_{i=1}^{n} (x_i - x_j)^r K\left((x_i - x_j)/h\right)$ for $r = 0,1,2,3$.

Since the first derivative $K^{(1)}$ of the kernel is assumed to be bounded, we can approximate the functions $\hat{S}_r(x_j; h)$ by integrals. In order to perform this, we need the following conditions and the sample size n to be sufficiently large:

i. The function $m''(.)$ is continuous on $(0, 1)$.

ii. The kernel K is symmetric and supported on $(-1, 1)$. Also K has a bounded first derivative.

iii. The bandwidth h is a sequence of values which depend on the sample size n and satisfying $h \to 0$ and $nh \to \infty$ as $n \to \infty$.

iv. The point x_j at which the estimation is taking place satisfies $h < x_j < 1 - h \quad \forall_n \geq n_0$ where n_0 is fixed.

Thus, we have

$$\hat{S}_r(x_j; h) = \int_0^1 (y - x_j)^r K\left((y - x_j)/h\right) dy + O(n^{-1})$$

$$= h^{r+1} \int_{-x_j/h}^{(1-x_j)/h} u^r K(u) du + O(n^{-1})$$

$$= h^{r+1} \int_0^1 u^r K(u) du + O(n^{-1}) \tag{3.162}$$

By the symmetry and compact support of K, the odd moments of K are all zero and so we have

$$n^{-1} X_i' W_i X_i = \begin{pmatrix} \hat{S}_0(x_j; h) & \hat{S}_1(x_j; h) \\ \hat{S}_1(x_j; h) & \hat{S}_2(x_j; h) \end{pmatrix}$$

$$= \begin{pmatrix} h + O(n^{-1}) & O(n^{-1}) \\ O(n^{-1}) & h^3 \sigma_K^2 + O(n^{-1}) \end{pmatrix} \tag{3.163}$$

where $\sigma_K^2 = \int_{-1}^1 u^2 K(u) du$ and

$$(n^{-1})(X_i' W_i) = \begin{pmatrix} (x_1 - x_j)^2 \\ (x_2 - x_j)^2 \\ \cdot \\ \cdot \\ (x_n - x_j)^2 \end{pmatrix} = \begin{pmatrix} \hat{S}_2(x_j; h) \\ \hat{S}_3(x_j; h) \end{pmatrix}$$

$$= \begin{pmatrix} h^3 \sigma_K^2 + O(n^{-1}) \\ O(n^{-1}) \end{pmatrix} \tag{3.164}$$

The following expression for the leading bias term is obtained using some straight forward matrix algebra

$$E\left[\bar{m}_i(x_j)\right] - m_i(x_j) = \frac{1}{2} h^2 \sigma_K^2 m_i^2(x_j) + o(h^2) + O(n^{-1}) \tag{3.165}$$

Examining the asymptotic variance of $\bar{m}_i(x_j)$, we have

$$Var\left(\bar{m}_i(x_j)\right) = e_1' (X_i' W_i X_i)^{-1} (X_i' W_i V W_i X_i) (X_i' W_i X_i)^{-1} e_1$$

$$= (n^{-1}) e_1' n (X_i' W_i X_i)^{-1} (n^{-1}) (X_i' W_i V W_i X_i) n (X_i' W_i X_i)^{-1} e_1 \tag{3.166}$$

where $V = diag(\sigma_e^2, \ldots, \sigma_e^2)$. And using approximations analogous to those used above, we have

$$n^{-1} (X_i' W_i V W_i X_i) = n^{-1} \sum_{i=1}^n K\left((x_i - x_j)/h\right)^2 \sigma_K^2 \begin{pmatrix} 1 & (x_i - x_j) \\ (x_i - x_j) & (x_i - x_j)^2 \end{pmatrix}$$

$$= \begin{pmatrix} h\sigma_K^2 R(K) + o(n^{-1}) & O(n^{-1}) \\ O(n^{-1}) & h^3\sigma_K^2 \int u^2 K(u)^2 du + O(n^{-1}) \end{pmatrix} \quad (3.167)$$

where $R(K) = \int K(u)^2 du$. We can combine these expressions to obtain

$$Var\left(\overline{m}_i(x_j)\right) = \frac{1}{nh}\sigma_K^2 R(K) + o\{(nh)^{-1}\} \quad (3.168)$$

3.11 Chapter Summary

So far we have studied the properties of the local linear regression estimators of finite population total in a model based framework. We have also extended the local linear regression procedure to stratified random sampling and to two stage cluster sampling. Analytical comparisons show that the local linear regression estimators are consistent and have minimal asymptotic variance in comparison with the Nadaraya-Watson regression estimator. In the next chapter, we carry out a study to compare the performances of the proposed local linear regression estimators of finite population total with some other estimators that exist in the literature.

CHAPTER FOUR

EMPIRICAL STUDY

4.1 Introduction

We have indicated that the local linear regression estimators are not only asymptotically design unbiased but are also consistent estimators. In this chapter, a study is carried out to compare the performances of the derived local linear regression estimators with some other estimators that exist in the literature. In particular, the estimator that is more efficient than the other one under different circumstances is determined. The design based estimators, the parametric model based estimators and the nonparametric model based estimators are considered in our simulation experiments.

4.2 Population description

In this section, four data sets are considered, which are again generated from the super population model (3.12) having different mean functions: Linear (L), Quadratic (Q), Bump (B) and Jump (J). The assumptions stated in equations (3.13) (3.14) and (3.15) still hold for the super population model considered in the nonparametric regression estimation of $m(x_i)$. The properties of the error are defined by equations (3.16) and (3.17). The idea is to estimate the finite population total T using the super population model (3.12) having different mean functions $m(x_i)$: L, Q, B and J.

The four sets of observations are generated as independent and identically distributed (iid) random variables from a uniform distribution over $(0, 1)$. The observations $x_i's$ are generated from the model (3.12) with respect to the mean functions $m_j(x_i)$: L, Q, B, J with $1 \leq i \leq 200$, $j = 1,2,3,4$

$m_1(x_i): L = 1 + 2(x_i - 0.5)$

$m_2(x_i): Q = 1 + 2(x_i - 0.5)^2$

$m_3(x_i): B = 1 + 2(x_i - 0.5) + \exp(-200(x_i - 0.5)^2)$

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

where in $m_4(x_i)$, the indicator functions $I_{(x \leq 0.65)}$ and $I_{(x > 0.65)}$ equal 1 if the event occurs and 0 otherwise.

The above mean functions represent the model specifications for the parametric and nonparametric estimators in consideration for cases where the model is correctly specified or incorrectly specified. The linear regression estimator is expected to be the best under the linear relationship as the model is correctly specified. The remaining mean functions; $m_2(x_i)$, $m_3(x_i)$ and $m_4(x_i)$ represent different deviations from the linear model. The errors are assumed to be independent and identically distributed (iid) random variables with mean 0 and constant variance.

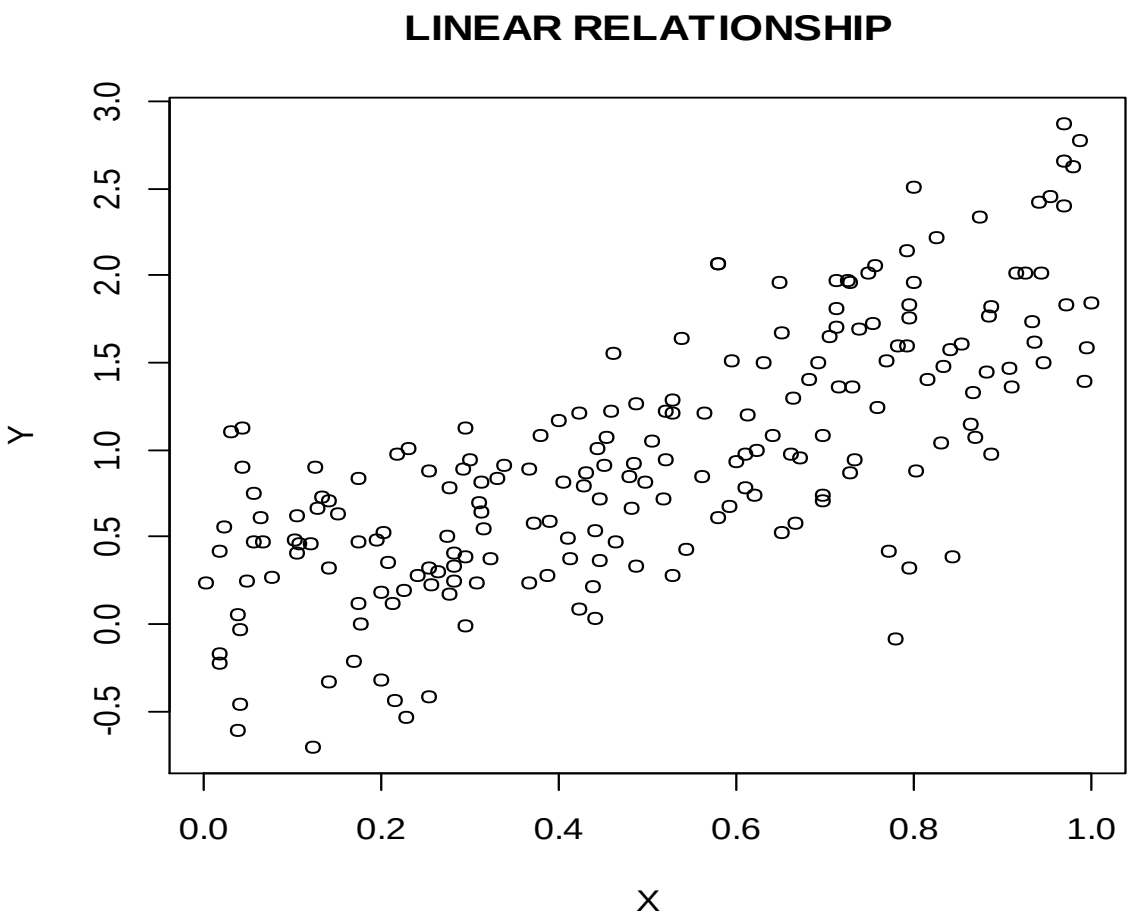

Figure 4.1: A scatter diagram for the linear relationship

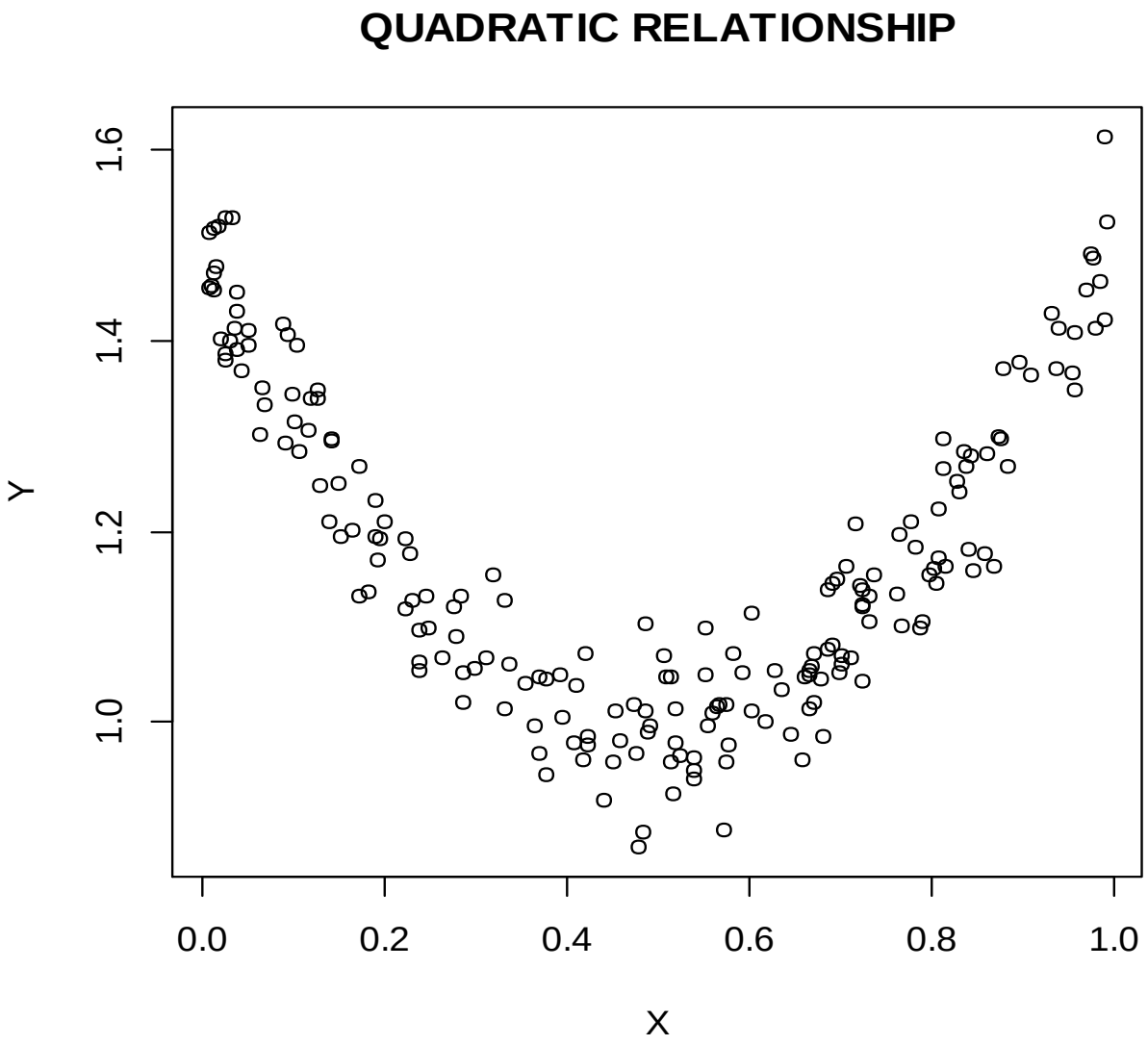

Figure 4.2: A scatter diagram for the quadratic relationship

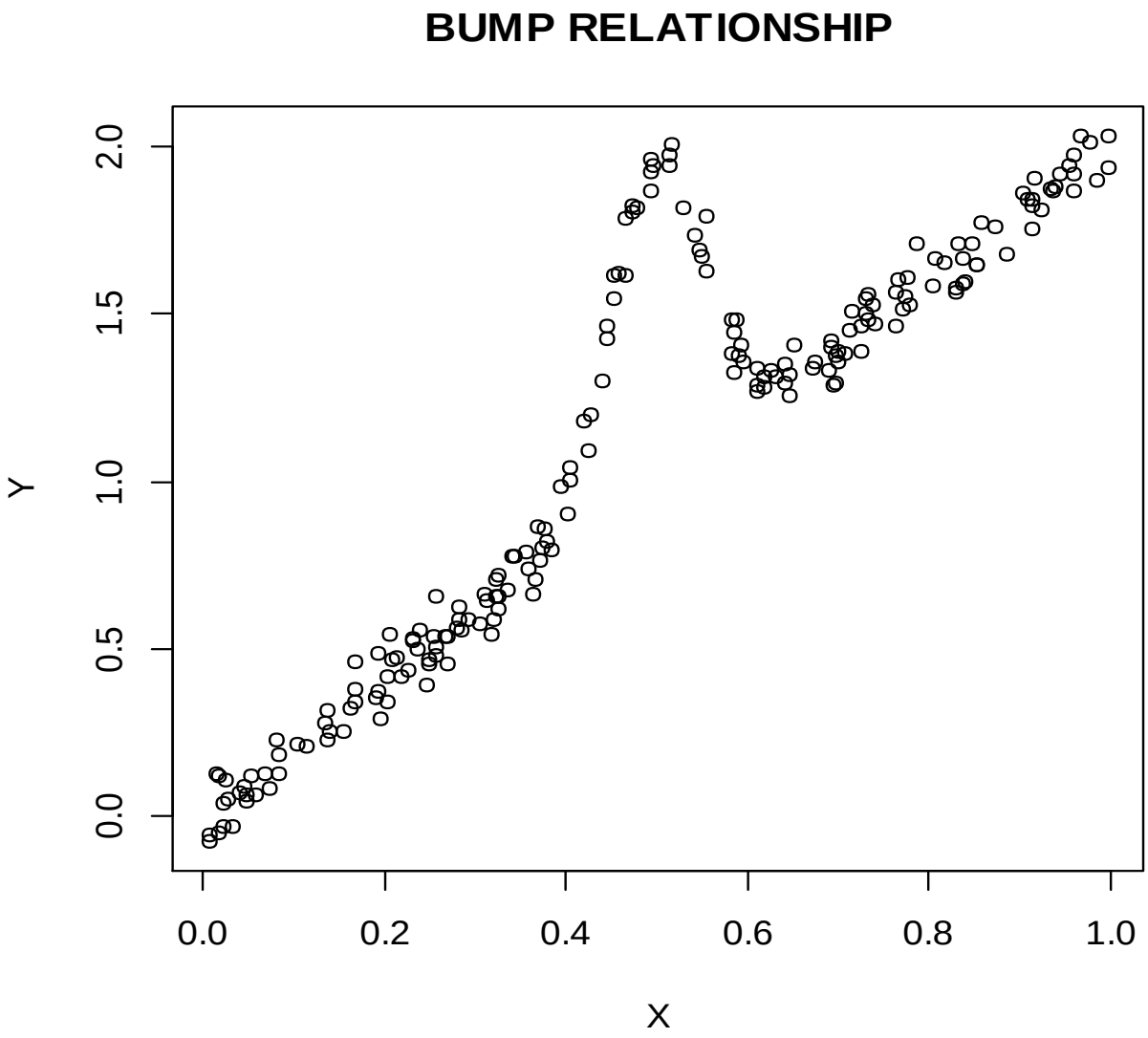

Figure 4.3: A scatter diagram for the bump relationship

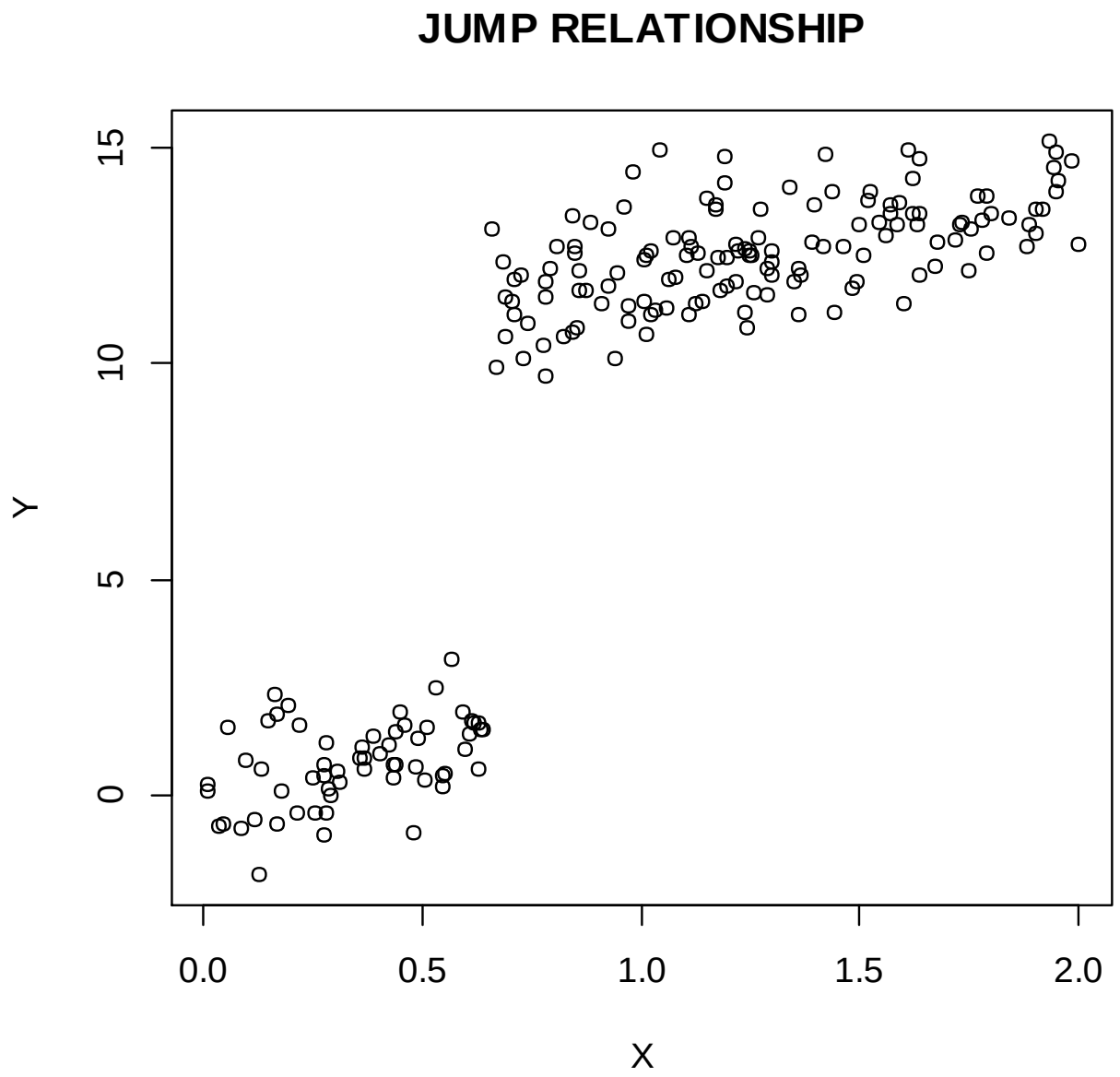

Figure 4.4: A scatter diagram for the jump relationship

Table 4.1: Notation for estimators used for comparison in the simulation study

$\bar{T}_{HT}$	$Horvitz-Thompson$	$Horvitz\ and\ Thompson$ (1952)
$\bar{T}_{REG}$	$Linear\ Regression$	$Cochran$ (1977)
$\bar{T}_{DORF}$	$Dorfman$	$Dorfman$ (1992)
$\bar{T}_{LL}$	$Local\ linear$	$Proposed\ Estimator$

Table 4.2: Formulae for computing estimator of finite population total

Estimator	**Formula**
$\bar{T}_{HT}$	$\bar{T}_{HT} = \sum_{i=1}^{n} \frac{y_i}{\pi_i}$
$\bar{T}_{REG}$	$\bar{T}_{REG} = \sum_{i \in S} y_i + \sum_{i \in R} (\bar{\alpha} + \bar{\beta} x_i)$
$\bar{T}_{DORF}$	$\bar{T}_{DORF} = \sum_{S} y_i + \sum_{P-s} \bar{m}(x_j)$
$\bar{T}_{LL}$	$\bar{T}_{LL} = \sum_{i \in S} y_i + \sum_{j \in R} \bar{m}_{LL}(x_j)$

4.3 The choice of the kernel function

We have several available kernel functions but the selected kernel should be theoretically good and be practical. According to Silverman (1986), we choose a function that satisfies the following conditions:

i. Small values should be minimized as they may cause numerical underflow in the computer.
ii. The kernel function should be user friendly, practical and theoretically fit in both simulated and raw data.
iii. The kernel function should be easy and simple to construct.
iv. The function should have its range well defined and not open as the Gaussian.

A comparison of the various kernels has been done by determining the efficiency of every kernel and the results are as shown in table (4.3) below:

Table 4.3: The kernel functions with respective efficiency

Kernel	k(t)		Efficiency
Epanechnikov	$\frac{3}{4\sqrt{5}}\left(1-\frac{1}{5}t^2\right)$	$\|t\| < \sqrt{5}$	1.0000
	0	otherwise	
Biweight	$\frac{15}{16}(1-t^2)^2$	$\|t\|<1$	0.9939
	0	otherwise	
Triangular	$1-\|t\|$	$\|t\|<1$	0.9859
Gaussian	$\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}t^2}$	$-\infty < t < \infty$	0.9512
Rectangular	$\frac{1}{2}$	$\|t\| < 1$	0.9295
	0	otherwise	

4.4 The choice of the bandwidth

Application of the selected kernel function, requires the specification and establishment of the bandwidth, h. The bandwidth is the standard deviation of the kernel and several bandwidths are available in practice. The proper selection of the bandwidth is always affected by the role for which the total estimate is to be applied. If the objective of population total estimation is to explore the data in order to suggest possible models and hypotheses, then it will probably be sufficient to select the bandwidth subjectively. When using population total estimation for presenting inferences, there is a case for under smoothing. Somehow, the user can do further smoothing by visual method but cannot easily unsmooth. However, many applications require an automatic choice of bandwidth. The automatic choice can nonetheless be used as a starting point for subsequent subjective adjustment. Scientists comparing their results will want to make reference to a standardized procedure. If the population total estimation is to be used on large data sets, then a user friendly and automatic method is necessary.

4.4.1 The subjective method

This is a natural procedure for selecting the bandwidth by plotting out several curves and choosing the estimate that fits the values of the data well in conjunction with the nature of the data. Accordingly, the process of investigating several plots of data, all smoothed by different amount of bandwidths, may give more insight into the data than merely considering a single automatically produced curve. In this procedure, several bandwidths are suggested and the optimal bandwidth is selected by visual inspection to pick that bandwidth that constructs the best curve.

4.4.2 The least squares cross validation method

This is an automatic procedure that totally depends on the data for selecting the bandwidth. We let the quantity $\Phi(h)$ be given by the relation

$$\Phi(h) = \sum_{k=1}^{L} \sum_{j \in r} \left\{ \left(\hat{m}_{hj}(x_j) - y_j \right)^2 w_i(x_j) \right\} \qquad L = 1{,}000 \tag{4.1}$$

The method is based on minimizing the quantity $\Phi(h)$ over the quantity h in L randomly selected samples from the target population. This method was initially suggested by (Rademo, 1982; Bowman, 1984; Hall, 1983 and Stone, 1984) for the density estimation.

The basic principle of least squares cross validation procedure is to construct the data themselves and then minimize the estimator over, h to give the choice of the bandwidth for the estimator of the finite population total, $\bar{T}$ given by

$$\bar{T} = \sum_{i \in S} y_i + \sum_{j \in R} \bar{m}(x_j) \tag{4.2}$$

The concern of picking the bandwidth from the interval arises. The Epanechnicov kernel is used for kernel smoothing because of its efficiency using well designed computer programs which is

$$K(t) = \frac{3}{4\sqrt{5}} \left(1 - \frac{1}{5} t^2 \right), \qquad |t| < \sqrt{5} \tag{4.3}$$

In Silverman (1986), the search for optimal bandwidth is done within the interval, $\frac{\sigma}{4n^{1/5}} \leq h \leq \frac{3\sigma}{2n^{1/5}}$ where σ is the standard deviation of the $x_i's$. In this study, the bandwidths are data driven and are determined by the least squares cross validation method.

We perform data simulations and computations using R computer software. We picked on a smaller population of size $N = 200$ because the nonparametric local linear regression method is slower and takes more computer time to compute the estimates. The simulation has however been made exhaustive by performing $r = 500$ replications and thus the confidence in our

conclusions. For each of the four data sets of size $N = 200$, samples are generated by simple random sampling without replacement using sample size $n = 60$. For each combination of mean function, standard deviation and bandwidth, 500 replicate samples are selected and the estimators calculated. Now for each of the four data sets and for each sample, we compute the finite population total given by

$$T = \sum_{i=1}^{N} y_i \tag{4.4}$$

The prediction errors for each of the estimators of finite population total are computed as

$$E_{HT} = (\bar{T}_{HT} - T) \tag{4.5}$$

$$E_{REG} = (\bar{T}_{REG} - T) \tag{4.6}$$

$$E_{DORF} = (\bar{T}_{DORF} - T) \tag{4.7}$$

$$E_{LL} = (\bar{T}_{LL} - T) \tag{4.8}$$

The biases for each of the estimators of finite population total are computed as

$$B(\bar{T}_{HT}) = \sum_{i=1}^{500} \left\{ \frac{\bar{T}_{HT} - T}{500} \right\} \tag{4.9}$$

$$B(\bar{T}_{REG}) = \sum_{i=1}^{500} \left\{ \frac{\bar{T}_{REG} - T}{500} \right\} \tag{4.10}$$

$$B(\bar{T}_{DORF}) = \sum_{i=1}^{500} \left\{ \frac{\bar{T}_{DORF} - T}{500} \right\} \tag{4.11}$$

$$B(\bar{T}_{LL}) = \sum_{i=1}^{500} \left\{ \frac{\bar{T}_{LL} - T}{500} \right\} \tag{4.12}$$

The mean square errors for each of the estimators of finite population total are computed as

$$MSE(\bar{T}_{HT}) = \sum_{I=1}^{500}\left\{\frac{(\bar{T}_{HT} - T)^2}{500}\right\} \tag{4.13}$$

$$MSE(\bar{T}_{REG}) = \sum_{I=1}^{500}\left\{\frac{(\bar{T}_{REG} - T)^2}{500}\right\} \tag{4.14}$$

$$MSE(\bar{T}_{DORF}) = \sum_{I=1}^{500}\left\{\frac{(\bar{T}_{DORF} - T)^2}{500}\right\} \tag{4.15}$$

$$MSE(\bar{T}_{LL}) = \sum_{I=1}^{500}\left\{\frac{(\bar{T}_{LL} - T)^2}{500}\right\} \tag{4.16}$$

For an unbiased estimator, the MSE is the variance of the estimator. Like the variance, MSE has the same units of measurement as the square of the quantity being estimated. Note also that

$$MSE(\bar{\theta}) = Var_{\bar{\theta}}(\bar{\theta}) + Bias(\bar{\theta}, \theta)^2 \tag{4.17}$$

The variances for the estimators of finite population total are computed as

$$VAR(\bar{T}_{HT}) = \sum_{I=1}^{500}\left\{\frac{\left(\bar{T}_{HT} - E(\bar{T})\right)^2}{500}\right\} \tag{4.18}$$

$$VAR(\bar{T}_{REG}) = \sum_{I=1}^{500}\left\{\frac{\left(\bar{T}_{REG} - E(\bar{T})\right)^2}{500}\right\} \tag{4.19}$$

$$VAR(\bar{T}_{DORF}) = \sum_{I=1}^{500}\left\{\frac{\left(\bar{T}_{DORF} - E(\bar{T})\right)^2}{500}\right\} \tag{4.20}$$

$$VAR(\bar{T}_{LL}) = \sum_{I=1}^{500}\left\{\frac{\left(\bar{T}_{LL} - E(\bar{T})\right)^2}{500}\right\} \tag{4.21}$$

We compute the absolute bias (AB) in order to analyse the performance of the proposed estimator versus some specified estimators using

$$AB(\bar{T}_{HT}) = \sum_{i=1}^{500} \left| \frac{(\bar{T}_{HT} - T)}{500} \right| \tag{4.22}$$

$$AB(\bar{T}_{REG}) = \sum_{i=1}^{500} \left| \frac{(\bar{T}_{REG} - T)}{500} \right| \tag{4.23}$$

$$AB(\bar{T}_{DORF}) = \sum_{i=1}^{500} \left| \frac{(\bar{T}_{DORF} - T)}{500} \right| \tag{4.24}$$

$$AB(\bar{T}_{LL}) = \sum_{i=1}^{500} \left| \frac{(\bar{T}_{LL} - T)}{500} \right| \tag{4.25}$$

The relative efficiency (RE) which examines the robustness of various estimators, that is, the Horvitz-Thompson estimator, the REG estimator and the Dorfman estimator versus the proposed local linear estimator is computed as

$$RE(\bar{T}_{HT}, \bar{T}_{LL}) = \frac{\sum_{i=1}^{500}(\bar{T}_{LL} - T)^2}{\sum_{i=1}^{500}(\bar{T}_{HT} - T)^2} \tag{4.26}$$

$$RE(\bar{T}_{REG}, \bar{T}_{LL}) = \frac{\sum_{i=1}^{500}(\bar{T}_{LL} - T)^2}{\sum_{i=1}^{500}(\bar{T}_{REG} - T)^2} \tag{4.27}$$

$$RE(\bar{T}_{DORF}, \bar{T}_{LL}) = \frac{\sum_{i=1}^{500}(\bar{T}_{LL} - T)^2}{\sum_{i=1}^{500}(\bar{T}_{DORF} - T)^2} \tag{4.28}$$

where $\bar{T}$ is the finite population total estimator in consideration, T is the true population total and $r = 500$ is the number of replications.

The confidence intervals (CI) and the average lengths (AL) of the confidence intervals of various estimators are computed as

$$CI(\bar{T}_{HT}) = \sum_{i=1}^{500}\left(\bar{T}_{HT} \pm 1.96\sqrt{Var(\bar{T}_{HT})}\right) \tag{4.29}$$

$$CI(\bar{T}_{REG}) = \sum_{i=1}^{500}\left(\bar{T}_{REG} \pm 1.96\sqrt{Var(\bar{T}_{REG})}\right) \tag{4.30}$$

$$CI(\bar{T}_{DORF}) = \sum_{i=1}^{500}\left(\bar{T}_{DORF} \pm 1.96\sqrt{Var(\bar{T}_{DORF})}\right) \tag{4.31}$$

$$CI(\bar{T}_{LL}) = \sum_{i=1}^{500}\left(\bar{T}_{LL} \pm 1.96\sqrt{Var(\bar{T}_{LL})}\right) \tag{4.32}$$

$$AL(\bar{T}_{HT}) = \frac{1}{500}\sum_{i=1}^{500}\left(CI_U(\bar{T}_{HT}) - CI_L(\bar{T}_{HT})\right) \tag{4.33}$$

$$AL(\bar{T}_{REG}) = \frac{1}{500}\sum_{i=1}^{500}\left(CI_U(\bar{T}_{REG}) - CI_L(\bar{T}_{REG})\right) \tag{4.34}$$

$$AL(\bar{T}_{DORF}) = \frac{1}{500}\sum_{i=1}^{500}\left(CI_U(\bar{T}_{DORF}) - CI_L(\bar{T}_{DORF})\right) \tag{4.35}$$

$$AL(\bar{T}_{LL}) = \frac{1}{500}\sum_{i=1}^{500}\left(CI_U(\bar{T}_{LL}) - CI_L(\bar{T}_{LL})\right) \tag{4.36}$$

where CI_L and CI_U are respectively the lower and upper confidence intervals within which we expect our true population total to lie with 95% confidence.

4.5 Results

The results for the absolute biases, mean square errors, relative efficiencies, confidence intervals and average lengths of confidence intervals for the various estimators are provided in tables 4.4, 4.5, 4.6, 4.7, 4.8 and 4.9 respectively.

Table 4.4: The AB of the estimators with respect to the four mean functions

THE ABSOLUTE BIAS (AB)				
	HORVITZ-THOMPSON (HT)	**LINEAR REGRESSION (REG)**	**DORFMAN (DORF)**	**LOCAL LINEAR (LL)**
Linear	139.1395	3.650095	3.628214	3.626798
Quadratic	163.4725	1.226636	0.403125	0.4323062
Bump	157.7427	2.018801	0.4777851	0.4087753
Jump	1219.668	21.785	9.760465	9.485367

Table 4.5: The MSE of the estimators with respect to the four mean functions

THE MEAN SQUARE ERROR (MSE)				
	HORVITZ-THOMPSON (HT)	**LINEAR REGRESSION (REG)**	**DORFMAN (DORF)**	**LOCAL LINEAR (LL)**
Linear	514.9775	15.36639	15.74559	15.47903
Quadratic	453.5207	1.521063	0.1713249	0.160443
Bump	548.131	4.551133	0.2942485	0.1894413
Jump	35691.94	512.8734	110.7915	97.02299

Table 4.6: The RE of the estimators to the proposed estimator

THE RELATIVE EFFICIENCY (RE)			
	HORVITZ-THOMPSON (HT)	**LINEAR REGRESSION (REG)**	**DORFMAN (DORF)**
	Relative Efficiency	**Relative Efficiency**	**Relative Efficiency**
Linear	0.09467563	0.8093	0.95664
Quadratic	0.000464731	0.9954403	0.962707
Bump	0.0002038478	0.02743355	0.9433107
Jump	0.003577862	0.1901854	0.9706123

Table 4.7: The CI of the estimators with respect to the four mean functions

THE 95% CONFIDENCE INTERVALS (CI)								
	HORVITZ-THOMPSON (HT)		**LINEAR REGRESSION (REG)**		**DORFMAN (DORF)**		**LOCAL LINEAR (LL)**	
	Lower Limit	**Upper Limit**	**Lower Limit**	**Upper Limit**	**Lower Limit**	**Upper Limit**	**Lower Limit**	**Upper Limit**
Linear	65.4358	78.3565	62.9204	63.2486	62.7598	63.0129	62.6295	63.0638
Quadratic	61.7471	62.4128	60.2974	60.3065	60.2583	60.2785	60.4442	60.4762
Bump	88.4308	92.8534	93.0109	93.1452	92.0642	93.3488	91.9164	93.1867
Jump	503.684	565.581	479.946	495.731	460.767	479.153	465.117	483.178

Table 4.8: The ALCI of the estimators with respect to the four mean functions

THE AVERAGE LENGTH OF CONFIDENCE INTERVALS				
	HORVITZ-THOMPSON (HT)	**LINEAR REGRESSION (REG)**	**DORFMAN (DORF)**	**LOCAL LINEAR (LL)**
Linear	12.92073	0.3282467	0.2532001	0.4342478
Quadratic	0.6656047	0.009090092	0.02025908	0.03197243
Bump	4.422574	0.1342954	1.284649	1.270295
Jump	61.8971	15.78477	18.38621	18.06073

Table 4.9: The Bias and MSE for $\overline{T}_0$ and $\overline{T}_1$ in the three selected mean functions

	Linear		**Quadratic**		**Bump**	
	$\overline{T}_0$	$\overline{T}_1$	$\overline{T}_0$	$\overline{T}_1$	$\overline{T}_0$	$\overline{T}_1$
BIAS	5.507608	3.777348	4.7372	0.45116	5.293896	0.4187236
MSE	100.8874	15.40735	18.40769	0.1601695	43.9272	0.1896261

4.6 Discussions

In this section, results for the bias (B), the mean square error (MSE), the relative efficiency (RE), the confidence intervals (CI) and the average length of confidence intervals (ALCI) are discussed. The bias of an estimator $\bar{\theta}$ of a parameter θ is the difference between the expected value of $\bar{\theta}$ and θ; that is, $Bias(\bar{\theta}) = E(\bar{\theta}) - \theta$. An estimator whose bias is identically equal to

0 is called an unbiased estimator and satisfies $E(\bar{\theta}) = \theta$ for all θ. The larger the bias, the poorer the estimators. The mean square error (MSE) measures the average squared difference between the estimator $\bar{\theta}$ and the parameter θ, which is a somewhat reasonable measure of performance for estimators. The MSE of an estimator $\bar{\theta}$ of a parameter θ is the function of θ defined by $E(\bar{\theta} - \theta)^2$ and this is denoted as $MSE_{\bar{\theta}}$. Thus, MSE has two components, one that measures the variability of the estimator (precision) and the other one that measures its bias (accuracy). An estimator that has good MSE properties has small combined variance and bias.

The relative efficiency of two estimators is the ratio of their efficiencies. If $\bar{\theta}_1$ and $\bar{\theta}_2$ are both unbiased estimators of θ, then the efficiency of $\bar{\theta}_1$ relative to $\bar{\theta}_2$ is $Eff(\bar{\theta}_1, \bar{\theta}_2) = Var(\bar{\theta}_2)/Var(\bar{\theta}_1)$. If this is less than 1, then it implies that $Var(\bar{\theta}_2) < Var(\bar{\theta}_1)$ and therefore $\bar{\theta}_2$ has a smaller variance than $\bar{\theta}_1$ and so $\bar{\theta}_2$ is preferred. Finally, confidence intervals consist of a range of values (interval) that act as good estimates of the unknown population parameter. The best performing confidence interval is the one whose coverage rate is close to the true population and with the shortest length.

4.6.1 The absolute bias

The biases for different estimators are summarised in table 4.4. In all the relationships considered, the Horvitz-Thompson estimator was the poorest resulting in large biases as compared to the other three finite population total estimators. The bias for the local linear regression estimators are much lower than those of the other three estimators. For all the biases computed, the local linear regression estimators are superior and dominate the Horvitz-Thompson estimator and the Linear regression estimator in all the relationships. The local linear regression estimators also dominate the Dorfman estimator in all the relationships except when the relationship is quadratic. The biases under the model based approach are also much lower than those under the design based approach under different relationships.

4.6.2 The mean square error

The MSE for different estimators are summarised in table 4.5. Generally, the estimator with a smaller MSE is regarded as the most efficient one. The local linear regression estimators are more efficient and performing better than the Horvitz-Thompson and Dorfman estimators, regardless of whether the model is specified or misspecified. The local linear regression estimators also outperform the linear regression estimator in all the relationships except when the relationship is linear. In general, local linear regression estimation removes a bias term from the kernel estimator, that makes it have better behavior near the boundary of the $x's$ and smaller MSE everywhere.

4.6.3 The relative efficiency

Table 4.6 examines the robustness of various estimators, that is, the Horvitz-Thompson estimator, the REG estimator and the Dorfman estimator versus the proposed local linear regression estimators. The results in the table show that relative efficiency of the Horvitz-Thompson estimator, the REG estimator and the Dorfman estimator to the proposed local linear regression estimators is less than 1. This implies that the proposed local linear regression estimators have a smaller variance than the three estimators and thus the three estimators are less efficient than the local linear regression estimators. Generally, the local linear regression estimators outperform the HT estimator, the REG estimator and the DORF estimator in all the relationships. The local linear regression estimators are therefore robust and are the most efficient estimators.

4.6.4 The confidence intervals and their average lengths

In table 4.7 and table 4.8, the confidence intervals and the average lengths of the confidence intervals are also measured for each case. A smaller length is better because it implies that the true population total is captured within a smaller range and therefore results are more precise.

The confidence intervals generated by the model based local linear method are much shorter than those generated by the design based Horvitz-Thompson method, regardless of whether the underlying model is specified or misspecified. The confidence intervals also indicate that the local linear regression method dominates the REG and Dorfman methods when the model is incorrectly specified. Generally, the model based estimators are much far better than the traditional design based estimators. The results show that the model based approach outperforms the design based approach at 95% coverage rate.

4.6.5 Comparison of estimators with respect to the bias and MSE in selected functions

In estimating $\bar{m}(x_j)$ for the local constant regression estimator $\bar{T}_0$, $\bar{\beta}$ has been assumed to be a pre assigned constant and in particular the value assigned is zero. It has therefore been shown in chapter three that the estimator $\bar{m}(x_j)$ is biased leading to a biased estimation of the finite population total. On the other hand, when estimating $\bar{m}(x_j)$ for the local linear regression estimator $\bar{T}_1$, the value of β is not pre-assigned but rather determined by the set of data provided and thus minimizing the bias. Analytically, variance comparisons are explored using the local constant regression estimator $\bar{T}_0$ and the local linear regression estimator $\bar{T}_1$ in which results indicate that the estimators are asymptotically equivalently efficient. It is observed that the biases and MSEs computed in table (4.9) for the local linear regression estimator $\bar{T}_1$ are small in all the three selected mean functions. The results therefore indicate that the local linear regression estimator $\bar{T}_1$ is superior and dominates the local constant regression estimator $\bar{T}_0$ for the linear, quadratic and bump relationships. Simulation experiments show that $\bar{T}_1$ is the most efficient estimator.

CHAPTER FIVE

CONCLUSIONS AND RECOMMENDATIONS

5.1 Summary

In this study, a model based approach for estimating the finite population total using the procedure of local linear regression has been considered. The local linear regression procedure has much strength and in particular, it is important in the following sense:

i. Adapts well to bias problems at boundaries and in regions of high curvature.
ii. Easy to understand and interpret.
iii. Methods have been developed that provide fast computation for one or more independent variables.
iv. Because of its simplicity, can be tailored to work for many different distributional assumptions.
v. Having a local model (rather than just a point estimate) enables derivation of response adaptive methods for bandwidth and polynomial order selection in a straightforward manner.
vi. Does not require smoothness and regularity conditions required by other methods such as boundary kernels.
vii. The estimate is linear in the response provided the fitting criterion is least squares and model selection does not depend on the response.

In chapter one, the background information and review of the basic concepts, definitions and terminologies applicable in the theory of sample surveys have been accomplished. The problem statement has been outlined and in addition, the objectives of the study and significance of the study have been stated. In chapter two, a critical review of the literature has been presented. Various approaches for estimating the finite population total have been discussed and in

particular, a detailed review of the nonparametric regression procedure for estimating the finite population parameters in different frameworks has been accomplished. In chapter three, we considered a design adaptive nonparametric approach based on weighted local linear regression estimators for estimating the finite population total. In particular, the local linear regression estimators have been derived in a model based framework. Likewise, asymptotic properties of the derived local linear regression estimators have been examined. The local linear procedure has been extended to stratified random sampling and to two stage cluster sampling.

Chapter four is on empirical study where simulation experiments have been conducted to compare the performances of the derived local linear regression estimators with some estimators described in chapter one and chapter two.

5.2 Conclusions

The results for the biases, the mean square errors, the relative efficiencies, the confidence intervals and the average length of confidence intervals with respect to the various estimators have been presented. The bias results show that the local linear regression estimators dominate the Horvitz-Thompson estimator for the linear, quadratic, bump and jump relationships. The MSE results show that the local linear regression estimators perform better than the Horvitz-Thompson estimator and Dorfman estimator, irrespective of the model specification or misspecification. Results also show that the local linear regression estimators are robust and are most efficient.

The results further indicate that the confidence intervals generated by the model based local linear regression procedure are much shorter than those generated by the design based Horvitz-Thompson method, regardless of whether the model is specified or misspecified. It has been observed that the model based approach outperforms the design based approach at 95% coverage rate.

Generally, the local linear regression estimators are not only superior to the popular kernel regression estimators, but are also the best among all linear smoothers including those produced by orthogonal series and penalized spline methods. The estimators adapt well to bias problems at boundaries and in regions of high curvature and do not require smoothness and regularity conditions required by other methods such as the boundary kernels.

Furthermore, the local constant and local linear regression estimators $\bar{T}_0$ and $\bar{T}_1$ of finite population total have been studied and compared. Analytically, variance comparisons have been explored using the local constant regression estimator $\bar{T}_0$ and the local linear regression estimator $\bar{T}_1$ in which results indicate that the estimators are asymptotically equivalently efficient. Simulation experiments carried out in terms of the biases and MSEs show that the local linear regression estimator $\bar{T}_1$ outperforms the local constant regression estimator $\bar{T}_0$ in all the three selected mean functions and therefore, $\bar{T}_1$ is the most efficient estimator.

5.3 Recommendations for further research

i. In this study, the local linear regression procedure has been extended to stratified random sampling and to two stage cluster sampling. But its contribution in the context of performance of estimators is lacking. Further research is needed as far as comparison in performances of the estimators are concerned with respect to stratified random sampling and two stage cluster sampling. Again, apart from a population being stratified into a fixed number of homogeneous strata, most surveys are multi level in nature and therefore more work is required. For example, individuals within households, households within locations, locations within divisions, divisions within counties and so on. In particular, introducing extra error terms so that one has mixed effects regression models within the local linear regression model will be of interest.

ii. After the simulations, real practical application is lacking. However, there are many data sets available to demonstrate the use of the local linear procedure to real data. For example, the Kenya demographic and health survey data which is readily available upon request. Again, there are many dependent variables to use in this data sets with many independent variables. Therefore, more work is required and recommended as far as real practical applications are concerned with respect to the simple random sampling, the stratified random sampling and the two stage cluster sampling techniques.

REFERENCES

Bolfarine, H. and Zacks, S. (1991). "Bayes and Minimax Prediction in Finite Populations." *Journal of Statistical Planning and Inference*, 28 (2), 139-151.

Bowman, A.W. (1984). "An Alternative Method of Cross Validation for the Smoothing of Density Estimation." *Biometrika*, 71, 353-360

Breidt, F.J. and Opsomer, J.D. (2000). "Local Polynomial Regression Estimation in Survey Sampling." *Annals of statistics*, 28, 1026-1053.

Breidt, F.J., Claeskens, G. and Opsomer, J.D. (2005). "Model Assisted Estimation for Complex Surveys Using Penalized Splines." *Biometrika* 92 *(4)*. 831-846.

Brewer, K.R.W. (1963). "Ratio Estimation in Finite Populations: Some Results Deductible from the Assumption of an Underlying Stochastic Process." *Austral. J. Statist*. 5 93–105.

Brewer, K.R.W. (1979). "A Class of Robust Sampling Designs for Large Scale Surveys." *J. Amer Statist. Assoc,* 74, 911-915.

Cassel, C.M., Sarndal, C.E. and Wretman, J.H. (1977). "Foundations of Inference in Survey Sampling." *Wiley, New York.*

Chambers, R.L. (1996). "Robust Case Weighting for Multipurpose Establishment Surveys." *Journal of official statistics* Vol. 12, 3-32.

Chambers, "R. L. (2003). "Which Sample Survey Strategy? A Review of Three Different Approaches." *Southampton statistical Sciences Research Institute, University of Southampton*

Chambers, R.L. and Dorfman, A.H. (2002). "Nonparametric Regression with Complex Survey Data." *Survey Methods Research Bureau of Labor Statistics.*

Chambers, R.L., Dorfman, A.H. and Wehrly, T.E. (1993). "Bias Robust Estimation in Finite Populations using Nonparametric Calibration." *J. Amer Statist. Assoc.* 88, 268-277

Chen, K. and Jin, Z. (2005). "Local Polynomial Regression Analysis for Clustered Data." *Biometrika*, 92, 59-74.

Chen, K. Fan, J. and Jin, Z. (2008). "Design Adaptive Minimax Local Linear Regression for Longitudinal/Clustered Data." *Statistica Sinica*, 18, pg. 515-534

Cleveland, W.S. (1979). "Robust Locally Weighted Regression and Smoothing Scatter Plots." J. *Amer. Statist. Assoc.* 74 829–836.

Cleveland, W.S. and Devlin, S. (1988). "Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting." *J. Amer. Statist. Assoc.* 83 596–610.

Cochran, W.G. (1977). "Sampling Techniques." New York: *John Wiley and sons Ltd*.

Deming, W.E. and Stephan, F. (1941). "On the Interpretation of Censuses as Samples." Proceedings of the Section on Survey Research Methods, *J. Amer. Statist. Assoc.* 36, 45-49.

Dorfman, A.H. (1992). "Nonparametric Regression for Estimating Totals in Finite Populations." Proceedings of the Section on Survey Research Methods, *J. Amer. Statist. Assoc.* 622-625.

Dorfman, A.H. and Hall P. (1993). "Estimators of the Finite Population Distribution Function using Nonparametric Regression*." Annals of Statistics*. Vol. 21, 1452-1475.

Eubank, R. (1988). Spline Smoothing and Nonparametric Regression. *New York: Dekker*

Eubank, R. and Speckman, L. (1993). "Local Polynomial Fitting Adopted to the Autoregressive Context for Modeling Nonlinear Time Series under some Missing Conditions." *Journal of the American Statistical Association,* Vol. 88, 424 1287-1301.

Fan, J. (1992). "Design Adaptive Nonparametric Regression." *J. Amer Statist. Assoc* Vol. 87, Pg. 998–1004.

Fan, J. (1993). "Local Linear Regression Smoothers and their Minimax Efficiencies." *Annals of Statistics.* 21 196–216.

Fan, J. and Gijbels, I. (1992). "Variable Bandwidth and Local Linear Regression Smoothers *Annals of Statistics.* 21(4), 2008–2036.

Fan, J. and Gijbels, I. (1996). "Local Polynomial Modeling and its Applications." *Chapman and Hall, London*.

Gasser, T. and Muller, H.G. (1979). "Kernel Estimation of Regression Function." Smoothing techniques for curve estimation, (ed. T. Gasser and Rosenblatt) New York; *Springer Verlag*, 23-68

Godambe, V.P. (1955). "A Unified Theory of Sampling from Finite Populations." *Journal of the Royal Statistical Society,* Ser. B, 17, 269-278.

Godambe, V.P. (1982). "Estimation in Survey Sampling: Robustness and Optimality." *J. Amer Statist. Assoc*, 77, 393-406.

Godambe, V.P., and Joshi, V.M. (1965). "Admissibility and Bayes Estimation in Sampling Finite Populations." *Annals of Mathematical Statistics,* 36, 1707-1 742.

Hall, P. (1983). "Large Sample Optimality of Least Squares Cross Validation in Density Estimation." *Annals of Statistics,* 11, 1156-1174

Hall, P. and Turlach, B.A. (1997). "Interpolation Methods for adapting to Sparse Design in Nonparametric Regression." *J. Amer Statist. Assoc,* Vol. 92 Pg. 466-472

Hardle, W. (1989). "Applied Nonparametric Regression Analysis." *Cambridge: Cambridge University Press.*

Harms T. and Duchesne P. (2010). "On Kernel Nonparametric Regression Designed for Complex Survey Data." *metrika,* 72, 111-138

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

Horvitz, D.G. and Thompson, D.J. (1952). "A Generalization of Sampling without Replacement from a Finite Universe." *Journal of American Statistical Association*, 47, 663–685.

Kasungo, K. (2002). "Model Based Approach to Finite Population Total Estimation by the use of Local Polynomial Regression." *Project, Kenyatta University, Kenya*.

Kim, J. Breidt, F. and Opsomer, J. (2009). "Nonparametric Regression Estimation of Finite Population Totals under Two Stage Cluster Sampling." *Technical report, Department of Statistics*, Colorado State University.

Kuk, A. (1993). "A Kernel Method for Estimating Finite Population Distribution Functions using Auxiliary Information." *Biometrika.* Vol. 80, 385-392.

Kuo, L. (1988). "Classical and Prediction Approaches to Establishing Distribution Functions from Survey Data." Proceedings of the Section on Survey Research Methods. *J. Amer Statist. Assoc* 280-285

Loftsgaarden, D.O. and Quesenberry, C.P. (1965). "A Nonparametric Estimate of a Multivariate Density Function." *The Annals of Mathematical Statistics* 36, 3, 1049-1051.

Luc C. (2016). "Nonparametric Kernel Regression using Complex Survey Data." Job market paper.

Masry, E. (1996). "Multivariate Local Polynomial Regression for Time Series, Uniform Strong Consistence and Rates." *Journal of Time Series Analysis*, 17, 571-599

Montanari, G.E. & Ranalli, M.G. (2003). "Nonparametric Methods in Survey Sampling." In: Vinci, M., Monari, P., Mignani, S. and Montanari, A., Eds., *New Developments in Classification and Data Analysis, Springer*, Berlin, 203-210.

Mostafa, S.A. & Shan, Q. J (2019). "Finite Population Model Assisted Estimation Using Combined Parametric and Nonparametric Regression Smoothers." 13: 58. https://doi.org/10.1007/s42519-019-0060-9

Nadaraya, E.A. (1964). "Estimating Regression." *Theory of probability and application,* 10, 186-190.

Ombui T.M. (2008). "Robust Estimation of Population Total Using Local Polynomial Regression." *Thesis, Jomo Kenyatta University of Agriculture and Technology, Kenya.*

Otieno, R. (1995). "Robust Variance Estimation for Finite Population Sampling." *Thesis, Kenyatta University, Kenya.*

Otieno, R. and Mwalili, T. (2000). "Nonparametric Regression for Finite Population Estimation." *East African Journal of Science*, 11, part 2, 107-112

Parichha P., Basu K. and Bandyopadhyay, A. (2019). "Development of Estimation Procedure of Population Mean in Two Phase Stratified Sampling." *Open Access Peer-Reviewed Chapter-ONLINE FIRST, http://dx.doi.org/10.5772/intechopen.8285027*

Parzen, E. (1962). "On Estimating of a Probability Density Function and Mode." *Annals of Mathematical Statistics*, 33, 1065-1076

Pfeffermann D. (1993). "The Role of Sampling Weights when Modeling Survey Data." *international statistics review*, 61, 317-337.

Priestley, M.B and Chao, M.T. (1972). "Nonparametric Function Fitting." *Journal of Royal Statistical Society* B 34, 384-392

Rady E. A. and Ziedan D. (2014). "Estimation of Population Total using Local Polynomial Regression with Two Auxiliary Variables." *J. Stat. Appl. Pro*.3, No. 2, 129-136.

Rady E. A. and Ziedan D. (2014). "A New Technique for Estimation of Total using Nonparametric Regression Under Two Stage Cluster Sampling." *Applied Mathematical Sciences. Vol*.8, No. 74, 3647-3659.

Robinson, P.M. and Sarndal, C.E. (1983). “Asymptotic Properties of the Generalized Regression Estimation in Probability Sampling.” Sankhya: *The Indian Journal of Statistics*, Series B 45, 240–248.

Rosenblatt, M. (1956). “Remarks of some Nonparametric Estimates of Density Function.” *Annals of Mathematical Statistics*, 27, 832-837

Royall, R.M. (1970). “On Finite Population Sampling Under Certain Linear Regression Models.” *Biometrika* 57, 377–387.

Royall, R.M. (1976), “The Linear Least Squares Prediction Approach to Two-Stage Sampling.” *J. Amer Statist. Assoc*, 71, 651-664.

Royall, R.M., and Cumberland, W.G. (1978). “An Empirical Study of the Ratio Estimator and Estimators of its Variance.” *J. Amer Statist. Assoc.* 76, 66-77

Royall, R.M., and Herson, J. (1973a). “Robust Estimation in Finite Populations I.” *J. Amer Statist. Assoc*, 68, 880-889.

Ruppert, D. and Wand, M.P. (1994). “Multivariate Locally Weighted Least Squares Regression.” *Ann. Statist.* 22 1346–1370.

Ruppert, D., Sheather, S.J. (1995) and Wand, M.P. (1994). “Multivariate Locally Weighted Least Squares Regression.” *Ann. Statist.* 22 1346–1370.

Sanchez B.I, Opsomer J., Rueda, M. and Arcos, A. (2014). “Nonparametric Estimation with Mixed Data Types in Survey Sampling.” *Rev mat complut*, 27, 685-700

Sarathi P.D. (2007) “Adaptive Design for Locating the Maxima of a Regression Function; A Nonparametric Approach.” *Indian Statistical Institute*.

Sarndal, C.E. (1980). “On Inverse Weighting Versus Best Linear Unbiased Weighting in Probability Sampling.” *Biometrika* 67, 639–650.

Sarndal, C.E., Swenson, B. and Wretman, J. (1992). “Model Assisted Survey Sampling,” *Springer*, New York

Silverman, B. (1986). "Densty Estimation for Statistics and Data Analysis," *Monographs on Statistics and Applied Probability*, Chapman and Hall, London.

Simwa, R.O. (1987). "Mathematical and Statistical Analysis of HIV/AIDS Epidemic with Reference to Kenya and Uganda." *Thesis, Makerere University, Uganda*.

Smith T.M. (1976). "The Foundations of Survey Sampling" *Journal Royal Statistical Society Association* 139, Part 2, pp. 183-204

Smith T.M. and Njenga, E. (1992). "Robust Model Based Methods for Analytic Surveys." *Survey methodology*; 18, 187-208.

Stone, C.J. (1977). "Consistent Nonparametric Regression (with discussion)." *Annals of statistics*, Vol. 5, No. 4, 595-645

Stone, C.J. (1984). "An Asymptotic Optimal Window Selection Rule for Kernel Density Estimates." *Annals of statistics*, 12, 1285-1297

Su, L., Zhao Y and Yan T (2012). "Two Stage Method Based on Local Polynomial Fitting for a Linear Heteroscedastic Regression Model and its Applications in Economics." *Hindawi Publishing Corporation*, 2012, pages 1-7

Syengo C.K. (2018). "Local Polynomial Regression Estimator under Stratified Random Sampling." *Project, Pan African University of Science and Technology, Kenya*.

Wafula, C. (1997). "Model Based Analysis of the Variance Estimators for the Combined Ratio Estimator." *Journal of Agriculture, Science and Technology*, 1, 23-33

Wahba, G. (1973). "Smoothing Noisy Data by Spline Functions." *Numerical mathematics* 24, 383-394.

Wahba, G. (1975). "Optimal Convergence Properties of Variable Knot, Kernel and Orthogonal Series Methods for Density Estimation." *Annals of statistics* 3, 15-29.

Wand, M.P. and Jones, M.C. (1995). "Kernel Smoothing," Chapman and Hall, London

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

Wright, R.L. (1983). “Finite Population Sampling with Multivariate Auxiliary Information.” *J. Amer Statist. Assoc,* 78, 879-884.

Wu, C. and Sitter, R.R. (2001). “*A* Model Calibration Approach to Using Complete Auxiliary information from survey data.” *J. Amer Statist. Assoc* 96, 185–193.

Zheng, H. and Little, R.J. (2003). “Penalized Spline Model Based Estimation of the Finite Population Total from Probability Proportional to Size Samples.” *Journal of Official Statistics* 19, 99–117.

Zheng, H. and Little, R.J. (2004). “Penalized Spline Nonparametric Mixed Models for Inference about a Finite Population Mean from Two Stage Samples.*” Survey Methodology* 30, 209–218.