International Journal of Statistics and Probability; Vol. 8, No. 3; May 2019
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

A Boundary Corrected Non-Parametric Regression Estimator for Finite Population Total

Langat Reuben Cheruiyot[1], Odhiambo Romanus Otieno[2] & George O. Orwa[2]

[1]Department of Mathematics and computer Science, University of Kabianga, P. O Box 2030-20200, Kericho, Kenya

[2]Department of Statistics and Actuarial Science, Jomo Kenyatta University of Agriculture and Technology, P. O Box 62000-0100, Nairobi, Kenya

Correspondence: Langat Reuben Cheruiyot, Department of Mathematics and computer Science, University of Kabianga, P. O Box 2030-20200, Kericho, Kenya

Received: March 17, 2019 Accepted: April 25, 2019 Online Published: April 27, 2019

doi:10.5539/ijsp.v8n3p83 URL: https://doi.org/10.5539/ijsp.v8n3p83

Abstract

This study explores the estimation of finite population total. For many years design-based approach dominated the scene in statistical inference in sample surveys. The scenario has since changed with emergence of the other approaches (Model-Based, Model-Assisted and the Randomization-Assisted), which have proved to rival the conventional approach. This paper focuses on a model based approach. Within this framework a nonparametric regression estimator for finite population total is developed. The nonparametric technique has been found from previous studies to be advantageous than its parametric counterpart in terms of robustness and flexibility. Kernel smoother has been used in construction of the estimator. The challenge of the boundary problem encountered with the Nadaraya-Watson estimator has been addressed by modifying it using reflection technique. The performance of the proposed estimator has been compared to the design-based Horvitz Thompson estimator and the model –based nonparametric regression estimator proposed by (Dorfman, 1992) and the ratio estimator using simulated data.

Keywords: auxiliary information, reflection technique, kernel smoothers, nonparametric regression estimator, boundary correction

1. Background and Motivation

The goal of a researcher in survey sampling is to make estimation of the population parameters with precision and accuracy. The precision of an estimate depends on the survey strategy employed. Because of this realization, a strategy that utilizes auxiliary information has been known in literature to have an upper hand. In fact auxiliary information on finite population is often used to increase precision of estimators of parameters, (Cochran, 1977). Previous studies show that in the presence of auxiliary variable, both model based and model assisted approaches perform better than the purely design-based approach provided the assumed model that links study variable and auxiliary variables is appropriate (Prasad & Subhash, 2011). This is the reason that motivated us to carry out our study within a model- based framework. Further on this, one has the option of estimating the finite population total using a parametric or a nonparametric regression technique. Regression models give a general relationship between the response variable and the auxiliary variable. A linear regression estimate may however, produce a large error for every sample size if the true underlying function is not linear and cannot be well approximated by a linear function (László, A, Kohler, & Walk, 2002). To address this problem the non parametric regression estimation is the option to go for. The advantages as stipulated by (Härdle, 1994) include the fact that it provides a versatile method of exploring the general relationship between two variables; secondly it enables one to make prediction of observations without any reference to a fixed parametric model; thirdly it is a tool for finding spurious observations by studying influence of isolated points and lastly it is a flexible method for interpolating between adjacent values of the auxiliary variable. Notably, nonparametric estimation that use kernel densities, suffer from the boundary bias.

1.1 Statement of the Problem

Nonparametric regression estimation normally uses kernel smoothing technique. The Nadaraya-Watson estimator is the commonest of such smoothers. It has however, been known in literature that this technique induces substantial amount of bias in the estimate at the boundary. The focus of this paper therefore is to estimate the finite population total; T under the model-based framework using a technique does not suffer significantly from the boundary problem. This is developed in the next sections.

1.2 Introduction

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

Often researchers are faced with the problem of estimating a population parameters, for instance, the population total, $T = \sum_{i=1}^{N} Y_i$, or the population mean $\bar{Y}$ among others. Studies in the distribution may be found in (Chambers, Dorfman, & Wehrly, 1993)and (Dorfman, 1992).

We will thus take a sample, *S*, so that we have (X_i, Y_i), *i*= 1, 2, .., *n*. It will be assumed that X_i's are known for all elements of the population of interest and may be used in the design stage, estimation stage or both stages (Hedayat & Sinha, 1991). Below we review the sampling strategies that can be used.

1.3 Review of Estimation Approaches in Survey Sampling

The usual inference problem in sample surveys is to estimate some summary characteristic of the population, such as the mean or the total of the *Y*-values, after observing the sample only. Various statisticians with varying points of view have proposed different approaches which one can take to make the appropriate inference. These are:-Design-based (Randomization-based) approach, Model-Based (Prediction-based/super population) approach, Model-assisted approach and Randomization-assisted model- based approach.

1.3.1 Design – Based (Randomization – Based) Approach

In this approach the values of a variable of interest of the target population are viewed as fixed quantities (constants). This implies that the selection probabilities introduced with the design are used in determining the properties of estimators used to obtain expected values, variances, biases and so on. It is also known as classical approach.

The statisticians who have relied in the design – based methods like it for the capability of elimination of personal biases in selecting the sample and its use in situations where little may be known about the population. Most researchers here look for design-unbiased methods of estimation and mind less on the nature of the population itself. This approach describes the way the sample is selected and therefore the distribution is exactly known because the designer imposes it on the population.

It should, however, be noted that besides the above advantages, obtaining an optimal strategy under this approach might be an impossible task where no restriction on the sample size is made, a result first noted by (Godambe, 1955). Both robustness and optimality cannot be achieved under this approach.

1.3.2 Model-Based Approach

In this approach, the distribution, unlike the above, is a structure innate to the population itself and is unknown but capable of being modeled. In the model – based approach or prediction theory inference, the relevant expectations are over all possible realizations of a stochastic model (usually a linear regression model), which connects a variable of interest *Y* with a set of auxiliary or benchmark variables *X*, (Cox, 1995). Statisticians using this approach view the values of interest in the population as random variables.

One area of sampling in which this super-population approach has received considerable attention is in connection with ratio and regression estimation. For example, in spatially distributed geological and ecological populations, the variable of interest of nearby units may be positively correlated, with the strength of the relationship decreasing with distance. When such tendencies are known to exist, they can be used in obtaining efficient predictors of unknown values and in devising efficient sampling procedures. This approach seems appropriate especially in sampling for resources, say, in which cost of sampling is high yet the economic incentive is strong for obtaining the most precise possible estimates for a given amount of sampling effort, as in the case of mining. When errors are modeled they are taken into account and in some way models provide for bias adjustment and assessment of the uncertainty of the estimates. Different models – for sample selection and for estimation can be developed. We, however, note that the choice of a model and its robustness to misspecification is the major issue. Small deviations from a chosen model may lead to serious errors in an inference.

1.3.3 Model-Assisted Approach

The two approaches reviewed above have their own individual strengths and weaknesses. Though for a long time they were viewed as rival approaches, some considerable researchers have made an attempt to view them as complementary and not as two competing approaches. Some references include (Brewer, 2002) and (Särndal, Swesson, & Wretman, 1992). Model-assisted approach is a method that still depends exclusively on randomization-based inference and

estimators but optimizes them under the explicit assumption that the finite population under study is itself a sample drawn from a super-population generated by a specific stochastic model. Basically, inferences are design-based while the model serves as a vehicle to help choose between the randomization based methods. Because of this, the approach may also be referred to as model-assisted design-based approach.

1.3.4 Randomization-Assisted Model-Based Approach

As opposed to the approach in 1.2.3, this approach employs design-based method to simply protect against model failure, (Kott, 2005). Here inferences remain model-based and therefore the concern is with model-unbiasedness and not design-unbiasedness, (Langat, Odhiambo, & Odongo, 2007).

The four approaches highlighted above basically stems from the two strategies - the traditional Design-based approach which has its conceptual origin in the paper by (Neyman, 1934) and the sampling theory texts such as (Kish, 1965) and (Cochran, 1977), where inferences are based on the probability distribution induced by the sampling design with the population values being held constant and the Model-based approach strongly linked to Royall and his students, where inferences are model dependent. (Royall, 1970) gives a summary of the philosophy behind this approach. It should be noted that the nonparametric nature of the Design-based Approach can make it an obvious methodology to robust inferences; however, there are no relevant optimality criteria that can be checked under this approach (Chambers R. , 2011). Therefore, if one wants both optimality and robustness, the option is Model-based approach.

To remove the boundary effect in kernel estimation, a number of techniques have been developed in literature especially in density estimation. For an overview of these techniques, one can see for example (Karunamuni & Alberts, 2004). This paper explores the reflection technique in addressing the boundary problem in regression estimation context.

1.4 Outline of the Paper

In section 2, the Horvitz Thompson estimator (design- based estimator) and the ratio estimator are stated. This is to be used for comparison purposes with the nonparametric regression estimator for finite population total proposed under model-based framework. Nadaraya-Watson Kernel estimation has been reviewed and in particular the bias and the variance are stated. An overview of the reflection technique as a way of fixing the bias of the estimator is given in this section and the finite population total estimator using it proposed. In section 3 simulation studies and analysis is presented. Discussion of the results and conclusion is given in Section 4.

2. Development of the Proposed Estimator

2.1 Nonparametric Regression Estimation

The interest is to obtain the finite population total:

$$T = Y_1 + Y_2 + ... + Y_N = \sum_{i=1}^{N} Y_i \tag{2.1}$$

A famous design-based estimator for the population total is the (Horvitz & Thompson, 1952) estimator given by:

$$T_{ht} = \sum_{i \in s} \pi^{-1} y_i = \sum_{i=1}^{N} I_i \pi^{-1} y_i \tag{2.2}$$

Where I_i=1 if the i^{th} observation is in the sample and zero otherwise.

From the model-based framework, since we shall observe the units sampled, to estimate the population total is equivalent to estimating the non-sampled units and summing it with the observed units using an equation of the form:

$$T = \sum_{i \in s} y_i + \sum_{i \notin s} y_i \tag{2.3}$$

The non-sample can be estimated using a regression model of the form:

$$y_i = \alpha + \beta x_i + e_i \tag{2.4}$$

This equation is parametric since the parameters α and β have to be estimated using the Least Squares Estimation technique. Parametric models are not flexible and therefore under such models, estimators obtained are not robust. It is known that under the parametric super-population, misspecification of the model can lead to serious errors in an inference as demonstrated in the empirical study by (Hansen, Madow, & Tepping, 1983). It is for this reason that many researchers use nonparametric approach. They include (Dorfman, 1992), (Chambers, Dorfman, & Wehrly, 1993), (Odhiambo & Mwalili, 2000), (Tsybakov, 2009), (Chandran & Prajneshu, 2004), and (Breidt & Opsomer, 2009) among others.

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

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

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

In non-parametric estimation, the data is allowed to determine the behaviour of the models, thus the only assumption made about the observations is that they are independent and identically distributed (*i.i.d*) from an arbitrary continuous distribution.

A model- based non-parametric model is of the form:

$$Y_i = m(X_i) + e_i \quad i=1, 2, \ldots, n \tag{2.6}$$

where Y_i- is the variable of interest

X_i-is the auxiliary variable

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

e_i-is error term-assumed to be $N(0, \sigma^2)$

The idea of non- parametric regression goes back to (Nadaraya, 1964) and (Watson, 1964). Some of the current references include (Härdle, 1990), (Takezawa, 2006), (Gámiz, Kulasekera, Limnios, & Lindqvist, 2011) and (Tsybakov, 2009) among others.

2.2 Review of the Nadaraya-Watson Estimator

The idea of non- parametric regression has gained prominence in a couple of decades now. This section gives a brief derivation of Nadaraya- Watson estimator.

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

$$\text{(a)} \int K(z)dz = 1 \quad \text{(b)} \int zK(z)dz = 0 \ \text{(c)} \int z^2K(z)dz := K_2(K)(<\infty) \tag{2.7}$$

Further, let the smoothing weight be:

$$w_i(x) = \frac{K\left(\frac{x-X_i}{h}\right)}{\sum_s K\left(\frac{x-X_i}{h}\right)}, \ i=1,2,\ldots,n \tag{2.8}$$

A form of the kernel weight defined as in (2.8) was proposed by (Nadaraya, 1964) and (Watson, 1964). Since then many researchers have explored the nonparametric regression technique in estimation. Some of the current references include (Härdle, 1990), (Takezawa, 2006), (Gámiz, et al, 2011) and (Tsybakov, 2009) among others. Herein, a simple Nadaraya-Watson Kernel estimator of m(x) has been considered.

Assume a model of the form specified in (2.6), where $\sum_s(.)$ is the summation over all the sampled units and *h* is the bandwidth also referred to as the smoothing or tuning parameter, with $\sum_{i=1}^{n} w_i(x) = 1$. The Nadaraya-Watson estimator of *m*(*x*) is therefore given by:

$$\hat{m}_{NW}(x) = \sum_{i=1}^{n} w_i(x)Y_i = \frac{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)Y_i}{\sum_s K\left(\frac{x-X_i}{h}\right)} \tag{2.9}$$

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

$$\hat{T}_{np} = \sum_{i=1}^{n} y_i + \sum_{i=n+1}^{N} \hat{m}_{NW}(x_i) \tag{2.10}$$

The estimator given in equation (2.10) was first suggested by (Dorfman, 1992). For kernel regression estimator, the estimate of *m* at point *x* is obtained using a weighted function of observations in the *h*-neighbourhood of *x*. The weight given to each observation in the neighbourhood depends on the choice of kernel function.

The bias is then given by:

$$\begin{aligned} Bias\left(\hat{T}_{np}\right) &= E\left(\hat{T}_{np} - T\right) \\ &= E\left(\left[\sum_{i=1}^{n} y_i + \sum_{i=n+1}^{N} \hat{m}_{NW}(x_i)\right] - \left[\sum_{i=1}^{n} y_i + \sum_{i=n+1}^{N} y_i\right]\right) \\ &= E\left(\sum_{i=n+1}^{N} \hat{m}_{NW}(x_i) - \sum_{i=n+1}^{N} y_i\right) \\ &= E\left(\sum_{i=n+1}^{N} \hat{m}_{NW}(x_i) - \sum_{i=n+1}^{N} m(x)\right) \\ Bias\left(\hat{T}_{np}\right) &= E\left(\sum_{i=n+1}^{N} \hat{m}_{NW}(x_i) - \sum_{i=n+1}^{N} m(x)\right) \end{aligned} \tag{2.11}$$

It can be shown that this is given by:

$$\text{Bias}\left[\hat{T}_{np}\right] = \left(\frac{N-n}{n}\right) h^2 K_2(K)\left[\frac{1}{2} m''(x) + \left[f(x)\right]^{-1} f'(x) m'(x)\right] + o(h^2) \tag{2.12}$$

And the variance is given by:

$$Var\left[\hat{T}_{np}\right] = \frac{(N-n)^2 R(K)\sigma^2}{nhf(x)} + o\left(\frac{(N-n)^2}{nh} + \frac{1}{nh}\right) \tag{2.13}$$

The derivation of this can be found in (Hansen, 2009).

2.3 The Proposed Estimator of Finite Population Total

An estimator of finite population total of the form given below is hereby proposed:

$$\hat{T}_{npr} = \sum_{i=1}^{n} y_i + \sum_{i=n+1}^{N} \hat{m}_{ref}(x_i) \tag{2.14}$$

where the first term $\sum_{i=1}^{n} y_i$ is the sample total observed and therefore under model-based approach it will not be necessary to be estimated while the second term $\sum_{i=n+1}^{N} \hat{m}_{ref}(x_i)$ is the non-sample total term that is to be estimated non-parametrically using the reflection technique. As noted earlier the Nadaraya-Watson estimator induces a bias at the boundary. This is because at the boundary the interval where $x \in [0, h)$, the symmetric kernel has decreased amount or lacks data on part of its window. The data-reflected technique therefore provides the data through reflection method so that this information is put on the negative axis thereby supplying the kernel with the information required on this section. The following simple steps give the procedure on how it works. Let the $\{(X_1,Y_1), (X_2, Y_2), \ldots, (X_n,Y_n)\}$ be the set of *n* observations in the sample. If the data is augmented by adding the reflections of all the points in the boundary, to give the set $\{(X_1, Y_1), (-X_1, Y_1), (X_2, Y_2), (-X_2, Y_2) \ldots, (-X_n, Y_n), (X_n, Y_n)\}$. If a kernel estimate $m^*(x)$ is constructed from this data set of size 2*n*, then an estimate based on the original data can be given by putting $\hat{m}(x) = 2m^*(x)$, for $x \geq 0$, and zero otherwise. This gives the modified general weight function given by:

$$\hat{m}_{ref}(x) = \sum_{i=1}^{n} \frac{\left\{K\left(\frac{x - X_i}{h}\right) + K\left(\frac{x + X_i}{h}\right)\right\} Y_i}{\sum_{i=1}^{n} \left\{K\left(\frac{x - X_i}{h}\right) + K\left(\frac{x + X_i}{h}\right)\right\}} \tag{2.15}$$

It can be shown that the estimate will always have zero derivative at the boundary, provided the kernel is symmetric and differentiable. The estimate has also been shown under the section on properties of the data-reflected technique that it is

a p.d.f for the symmetric kernel. In practice it will not usually be necessary to reflect the whole data set, since if Xi/h is sufficiently large, the reflected point - Xi/h will not be felt in the calculation of m*(x) for x> 0, and hence reflection of points near 0 is all that is needed. (Silverman, 1986) in his example, states that if K is the Gaussian kernel there is no practical need to reflect points beyond Xi > 4h.

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

2.3.1 The Kernel Estimator at the Boundary

The interest in this study is in the boundary problem which occurs in the interval [0, *h*). This is as a result of lack of information which follows due to truncation of such information at this interval, i.e. the density function is continuous on [0, ∞) and is 0, for *x*< 0. This reduced amount of information leads to serious bias during the estimation and as such the estimate becomes inaccurate. The boundary problem arises when the value of *x* is smaller than the chosen value of the bandwidth. In the case of the standard kernel estimator of $m(x)$, consider

$$\hat{m}(c.h) \quad \text{for} \quad c \in [0,1), \text{ where } x = c.h, \text{ then for } \quad z = \frac{u-x}{h},$$

one can have

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

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

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

$$E[\hat{m}(x)] = \int_{-1}^{c} K(z)m(x+hz)dz \tag{2.16}$$

Taylor's expansion yields:

$$E[\hat{m}(x)] = m(x)\int_{-1}^{c} K(z)dz + hm'(x)\int_{-1}^{c} zK(z)dz + \frac{h^2}{2}m''(x)\int_{-1}^{c} z^2K(z)dz + o(h^2) \tag{2.17}$$

For the case of regression estimation considered in this study, it can be deduced that (2.17) results in:

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

$$- g'(x)m'(x)h^2\int_{-1}^{c} z^2K(z)dz\Bigg] \tag{2.18}$$

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

That is:

$$E[\hat{m}(0)] = \frac{1}{2}m(0) + O(h) \tag{2.19}$$

It should be noted that:

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

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

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

$$\int_{0}^{\infty} \hat{m}(x)dx = \frac{1}{nh}\sum_{i=1}^{n} h\int_{-\frac{X_i}{h}}^{\infty} K(z)dz < 1, \text{ if } \quad \exists i \in \{1,\ldots,n\} : X_i < h \tag{2.20}$$

an indication that the density does not live up to the condition of being a *p.d.f* about its support at the boundary. One

way of correcting this boundary problem is by use of data-reflected technique. Due to symmetry of the kernel function one can look at the reflection estimator as:

$$E\left(\hat{m}_{ref}(x)\right) = E\left[\hat{m}(x)\right] + E\left[\hat{m}(-x)\right] \tag{2.21}$$

2.3.2 The Bias of Data-Reflected Estimation Technique in Regression

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

$$K'(-z) = -K'(z) \tag{2.22}$$

So that

$$m'_{ref}(0) = \sum_{i=1}^{n} \frac{\left\{\frac{1}{h}K'\left(\frac{-X_i}{h}\right) + \frac{1}{h}K'\left(\frac{X_i}{h}\right)\right\}Y_i}{\sum_{i=1}^{n}\left\{\frac{1}{h}K'\left(\frac{-X_i}{h}\right) + \frac{1}{h}K'\left(\frac{X_i}{h}\right)\right\}} \equiv 0 \tag{2.23}$$

The implication of this is that the reflection estimator satisfies the so-called shoulder condition always. At the boundary the decreased amount of the data suggest concavity of the density in the vicinity of the origin. As a consequence, these kernels tend to misinterpret the local concavity as an indication of a mode over the strictly positive region, (Hirukawa & Sakudo, 2015). This is a condition where a given density, say m , has a shoulder at 0, i.e. $m'(0) = 0$. See for instance, (Mack, Quang, & Zhang, 1999). The graphs in Fig. 3.7 show this impression.

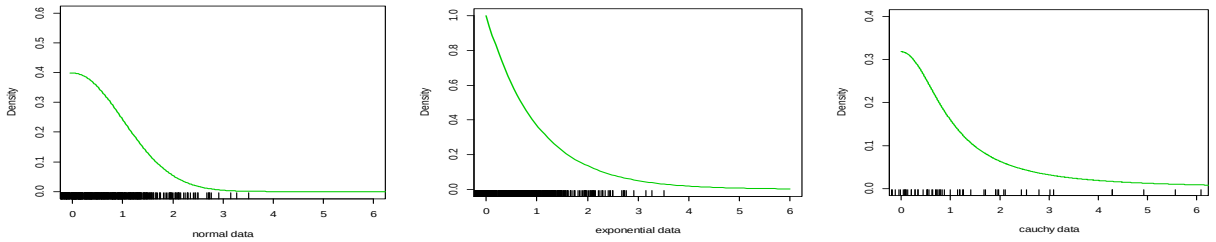

Figure 3.7. Shoulder condition: Except the 2nd, these densities satisfy the condition

The first term in the right hand side of expression (2.21) is already given above in (2.18); therefore, proceeding to look at the second term and noting that $Y_i = m(x) + \left[m(X_i) - m(x)\right] + e_i$ gives.

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

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

Thus

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

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

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

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

Taylor's expansion yields:

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

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

and

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

Therefore the product of this expansion is:

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

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

Thus

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

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

$$+ \frac{h^2}{2}m''(x)g(x)\int_{-1}^{-c}z^2K(z)dz\Bigg] + o(h^2) \tag{2.24}$$

Because of the property of symmetry the following equality holds:

$$\left.\begin{aligned}
&\int_{-1}^{c}K(z)dz = 1 - \int_{-1}^{-c}K(z)dz \\
&\int_{-1}^{c}zK(z)dz = \int_{-1}^{-c}zK(z)dz \\
&\int_{-1}^{c}z^2K(z)dz = K_2(K) - \int_{-1}^{-c}z^2K(z)dz
\end{aligned}\right\} \tag{2.25}$$

where $K_2(K) := \int_{-1}^{1}z^2K(z)dz \neq 0$, see equation (2.7)

Thus putting together the results in (2.24) with that of (2.18) yields in the following:

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

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

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

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

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

$$+ 2h^2m''(x)\left(c^2\int_{-1}^{-c}K(z)dz + c\int_{-1}^{-c}zK(z)dz\right)\Bigg] + o\left(h^2\right) \tag{2.26}$$

The bias for estimator of the finite population total, T_{npr}, given in equation (2.16) would therefore be given by:

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

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

$$\left.+2h^2m''(x)\left(c^2\int_{-1}^{-c}K(z)dz+c\int_{-1}^{-c}zK(z)dz\right)\right]+o\left(h^2\right) \tag{2.27}$$

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

2.3.3 The Variance of Data-Reflected Kernel Regression Estimation Technique

Similarly the variance can be computed as follows:

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

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

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

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

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

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

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

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

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

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

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

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

$$+h\int_{-1}^{-c}K(z)^2\left(g(x)-(2x+hz)g'(x)+\frac{(2x+hz)^2}{2}g''(x)+o\left(h^2\right)\right)dz \tag{2.28}$$

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

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

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

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

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

$$= \int_{-1}^{1} K(z)^2\,dz \tag{2.29}$$

With this, therefore, the variance is given by:

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

$$= \frac{(N-n)^2\sigma^2}{nhg(x)}\int_{-1}^{1} K(z)^2\,dz + O(n^{-1}) \approx \frac{(N-n)^2\sigma^2}{nhg(x)} R(K) \tag{2.30}$$

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

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

When the variance and the square of the bias term are summed up the MSE/AMSE error criterion is obtained.

3. Empirical Study

This section gives an empirical study that facilitates comparison between the two famous approaches in survey sampling- the Design-based and the Model-based approaches. Further to this, simulation study also compares the proposed estimator with the model-based ratio estimator and that due to (Dorfman, 1992). Six models given in table 3 have been used for this purpose. To achieve this, simulations were performed both for the response variable Y and the corresponding auxiliary variable X for a populations of size N=2000 from where 2000 simple random samples of size n=500 were drawn and used for estimation. In each case estimates of population totals were obtained using the Designed-based Horvitz-Thompson estimator, $\hat{T}_{HT}$, the other three model-based estimators, that is, the ratio estimator, $\hat{T}_R$, nonparametric regression estimator due to (Dorfman, 1992), $\hat{T}_{np}$, and the non-parametric regression estimator proposed in this study, $\hat{T}_{npr}$. The average relative biases of the finite population totals got using the three estimation techniques were obtained using the relation: $\left(\left|\sum_{i=1}^{2000}\hat{T}_i/2000 - T\right|\right)/T$ where T is the actual population total and $\hat{T}_i$ is one of the estimators of the population total computed from the ith sample. A cross validation data generated bandwidth was used in the simulation. See table 1 for these summaries.

Table 1. Summary of respective estimators and their relative biases for population totals

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

The *MSE*s of the respective estimators and models were also computed. The results are reported in table 2. Table 3 gives the equations of the models simulated.

Table 2. Summary results for the unconditional MSE (Obtained from 2000 iterations and sample sizes of n=500)

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

Table 3. Equations of models simulated

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

4. Discussion of Results

Table 1 and 2 presents the average relative biases of the various estimators studied i.e. the nonparametric regression estimator, T_{np}, nonparametric regression estimator (with kernel modified), T_{npr}, and the Horvitz Thompson estimator, T_{HT}, as well as the ratio estimator, T_R. The bandwidth was selected using the data-driven cross-validation technique.

Except for the ratio estimator in the linear model and the Horvitz-Thompson estimator in the bump model the proposed estimator proved superior of all in the average relative biases computed. It can also be seen that the unconditional *MSE*s are the smallest for the proposed estimator virtually beating all the rest of the estimators except for only the ratio estimator under the linear model where this fact can be attributed to the fact that it is the best linear unbiased estimator.

4.1 Conclusion and Recommendation

From the study, the proposed estimator gave very satisfactory results. The population totals arrived at were closer to the actual totals than the other techniques considered. This fact is evidenced by the small relative biases as well as the respective *MSE*s. Reflection technique can therefore be taken as a way of correcting the boundary bias in regression estimation.

References

Breidt, F. J., & Opsomer, J. (2009). Nonparametric and Semiparametric Estimation in Complex Surveys. (D. Pfeffermann, & C. R. Rao, Eds.) *Handbooks of Statistics, 29B*, 103-121. https://doi.org/10.1016/S0169-7161(09)00227-2

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

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

Chambers, R. L., Dorfman, A. H., & Wehrly, T. E. (1993). Bias Robust estimation in finite populations using nonparametric calibration. *Journal of American Statistics Association, 88*(421), 226–277. https://doi.org/10.1080/01621459.1993.10594319

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

Cochran, W. G. (1977). *Sampling Techniques* (3rd ed.). New York: John Wiley. https://doi.org/10.1002/9781118150504

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

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

Gámiz, M. L., Kulasekera, K. B., Limnios, N., & Lindqvist, B. H. (2011). *Applied Nonparametric Statistics in Reliability.* London: Springer verlag. https://doi.org/10.1007/978-0-85729-118-9

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

Hansen, M. H., Madow, W. G., & Tepping, B. J. (1983). An evaluation of model dependent and probability sampling inferences in sample surveys. *Journal of American Statistical Association, 78*(384), 776-793. https://doi.org/10.1080/01621459.1983.10477018

Härdle, W. (1990). *Applied Nonparametric Regression Analysis.* Cambridge: Cambridge University Press. https://doi.org/10.1017/CCOL0521382483

Härdle, W. (1994). *Applied Nonparametric Regression Analysis.* Cambridge: Cambridge University Press. https://doi.org/10.2307/2533418

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

Hirukawa, M., & Sakudo, M. (2015). Family of the Generalized Gamma Kernels: A Generator of Asymmetric Kernels for Nonnegative Data. Submitted paper. *Journal of Nonparametric Statistics, 27*(1), 41-63. https://doi.org/10.1080/10485252.2014.998669

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association, 47*(260), 663-685. https://doi.org/10.1080/01621459.1952.10483446

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

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

Kott, P. S. (2005). Randomization-assisted model-based survey sampling. *Journal of Statistical Planning and Inference, 129*(1), 263-277. https://doi.org/10.1016/j.jspi.2004.06.052

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

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

Mack, Y. P., Quang, P., & Zhang, S. (1999). Kernel estimation in transect sampling without the shoulder condition. *Communications in Statistics - Theory and Methods, 28*(10), 2277–2296. https://doi.org/10.1080/03610929908832422

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and Application, 9*(1), 141-142. https://doi.org/10.1137/1109020

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society, 97*(4), 558–622. https://doi.org/10.2307/2342192

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

Prasad, N. G., & Subhash, R. L. (2011, November 2010). Improved prediction in finite population sampling using convex combination of parametric and nonparametric models. *Sankhya B, 72*(2), 189-201. https://doi.org/10.1007/s13571-011-0009-9

Royall, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika, 57*(2), 377-387. https://doi.org/10.1093/biomet/57.2.377

Särndal, C. E., Swesson, B., & Wretman, J. N. (1992). *Model- Assisted Survey sampling* . New York: Springer Verlag. https://doi.org/10.1007/978-1-4612-4378-6

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* London: Chapman and Hall. https://doi.org/10.1007/978-1-4899-3324-9

Takezawa, K. (2006). *Introduction to Nonparametric Regression.* Hoboken, New Jersey: John Wiley. https://doi.org/10.1002/0471771457

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation.* New York: Springer Science+Business Media, LLC. https://doi.org/10.1007/b13794

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