Hindawi
Journal of Probability and Statistics
Volume 2019, Article ID 6812795, 8 pages
https://doi.org/10.1155/2019/6812795

Research Article

Boundary Bias Correction Using Weighting Method in Presence of Nonresponse in Two-Stage Cluster Sampling

Nelson Kiprono Bii,[1] Christopher Ouma Onyango,[2] and John Odhiambo [1]

[1] *Institute of Mathematical Sciences, Strathmore University, P.O. Box 59857-00200, Nairobi, Kenya*
[2] *Department of Statistics, Kenyatta University, P.O. Box 43844-00200, Nairobi, Kenya*

Correspondence should be addressed to John Odhiambo; jodhiambo@strathmore.edu

Received 10 September 2018; Accepted 5 December 2018; Published 2 June 2019

Academic Editor: Luis A. Gil-Alana

Kernel density estimators due to boundary effects are often not consistent when estimating a density near a finite endpoint of the support of the density to be estimated. To address this, researchers have proposed the application of an optimal bandwidth to balance the bias-variance trade-off in estimation of a finite population mean. This, however, does not eliminate the boundary bias. In this paper weighting method of compensating for nonresponse is proposed. Asymptotic properties of the proposed estimator of the population mean are derived. Under mild assumptions, the estimator is shown to be asymptotically consistent.

1. Introduction

Estimation of population parameters, for example, the population mean using kernel density estimators in presence of nonresponse, often leads to bias due to boundary effects; see, for instance, [1]. This affects the optimality of the estimators for the population parameters. To address this problem, use of optimal bandwidth has been suggested in literature. However, this does not eliminate the boundary bias. Weighting method of compensating for nonresponse in two-stage cluster sampling is proposed in this study. The values of the auxiliary variables X_{ij} are assumed to be known for all the clusters while the values of the survey variable Y_{ij} are only known for response units in the sample selected.

Let y_{ij} be the values of the survey variable Y_{ij} for unit j in cluster i, for $i = 1, 2, \ldots, N$; $j = 1, 2, \ldots, M$. The problem is to estimate the survey values of the nonresponse component in the second stage of sampling in the selected sample. This is done by first generating data using a linear regression model applied by [2, 3]. The model is given by

$$\widehat{Y}_{ij} = m\left(\widehat{x_{ij}}\right) + \widehat{e_{ij}}, \tag{1}$$

where $m(\cdot)$ is a smooth function of the auxiliary variables and $\widehat{e_{ij}}$ is the residual term with mean zero and variance which is strictly positive. Auxiliary data is assumed to be known throughout the study and is therefore used to predict the nonresponse values. In the following sections, different methods of estimating the nonresponse values of the survey variable Y_{ij} using a function of the auxiliary data, $m(\cdot)$, are discussed.

Let g denote a probability density function with the support $[0, \infty)$ and consider nonparametric estimator of g based on a random sample X_{ij}, $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, m$ from g. The kernel estimator of g due to [4] is given by

$$g_{mn}(x) = \frac{1}{mnb} \sum_{i=1}^{n} \sum_{j=1}^{m} K\left(\frac{x_{ij} - X_{ij}}{b}\right), \tag{2}$$

where K is some specified density function, symmetric about zero over the interval $[-1, 1]$ with b as the bandwidth such that $b \longrightarrow 0$ as $mn \longrightarrow \infty$. The properties of g_{mn} under some smoothness assumptions for $x_{ij} \geqslant 0$ are

$$E\left(g_{mn}(x)\right) = g(x) \int_{-1}^{c} k(w)\, dw - bg'(x) \int_{-1}^{c} wk(w)\, dw$$

$$+\frac{b^2}{2}g''(x)\int_{-1}^{c}w^2k(w)\,dw+O\left(b^2\right)$$

$$\operatorname{var} g_{mn}(x)=g(x)(mnb)^{-1}\int_{-1}^{c}k^2(w)\,dw+O\left(\frac{1}{mnb}\right), \tag{3}$$

where $c=\min(x/b,1)$.

For $x\geq b$, that is, the interior points, the bias of $g_{mn}(x)$ is of order $O(b^2)$. However, at the boundary points, that is, for $x\in[0,b)$, $g_{mn}(x)$ is not consistent. In nonparametric curve estimation problems, this phenomenon is called "boundary effects" of the estimator in (2).

2. Methods of Reducing Boundary Bias due to Nadaraya-Watson Estimator

It is generally known that kernel density estimators are not consistent when estimating a density near the finite end points of the support of the density to be estimated. This is due to the boundary effects that occur in nonparametric curve estimation. The estimator proposed by [1] suffers from boundary problem induced by Nadaraya-Watson estimator used. Notably, it is often desirable to have an optimal bandwidth to balance the bias-variance trade-off. If K is a symmetric density function and fixed across a support estimation, then the inference is generally simplified for unbounded support i.e., $(-\infty,\infty)$. But the function $m(\widehat{x_{ij}})$ is not consistent at the boundary $[0,b]$ where b is the bandwidth for such a choice of K; for details see [5, 6]. So for $x\in[0,b)$, the bias is of order $o(b)$ instead of order $o(b^2)$ at the boundary points. To eliminate the boundary effects in kernel density estimation, the methods described below have been proposed in literature. A brief description of these methods is hereby provided.

2.1. Reflection of Data Method. This method has been explored by [5, 7, 8]. It is also known as "data-reflected technique". To apply this method, one has to add $-X_i$, $i=1,\ldots,n$ to the data set. Since the kernel is penalizing for lack of data on the negative axis, the estimator therefore gradually applies reduced amount of data in its window as it approaches the boundary, thus resulting in a boundary bias; the addition of $-X_i$, $i=1,2,\ldots,n$, compensates for the lack of data. The estimator of $m(x)$ is defined by

$$m(\widehat{x})=\frac{1}{nb}\sum_{i=1}^{n}\left\{K\left(\frac{x-X_i}{b}\right)+K\left(\frac{x+X_i}{b}\right)\right\}. \tag{4}$$

For $x\geq 0$, $m'(\widehat{x})=0$, where $m(x)$ is a function of auxiliary random variables which are assumed to be known throughout the study. For $x<0$ it can be shown that $m'(\widehat{x})=0$; therefore it becomes better than other methods if the underlying density has the property $m'(0)=0$; if this property does not hold, the method may become cumbersome to apply.

2.2. Pseudodata Method. This method was suggested by [9]. In this method, data is generated outside the interval of estimation; that is, the method generates data beyond the left endpoint of the support of the density. Those data are assumed to be linear functions of order statistics in the original sample X. This transforms the data into a new set and then puts it on the negative axis. The estimator of $m(x)$ is given by

$$m(\widehat{x})=\frac{1}{nb}\left\{\sum_{i=1}^{n}K\left(\frac{x-X_i}{b}\right)+\sum_{i=1}^{m}K\left(\frac{x+X_{(-i)}}{b}\right)\right\}, \tag{5}$$

where $m\leq n$ and

$$X_{(-i)}=-5X_{(i/3)}-4X_{(2i/3)}+\frac{10}{3}X_{(i)}, \tag{6}$$

where $X_{(i)}$ linearly interpolates among $0,X_{(1)},X_{(2)},\ldots,X_{(n)}$ in that order; for details see [9]. Though this method is simple to implement and allows a minimal variance of the usual kernel estimator, the drawback of this method is that straight data reflection corrects only for a jump in the value of the density at the ends of its support, not for discontinuities in the derivatives of the density. Therefore, the method does not adequately correct bias problems caused by the edge effects in kernel estimators of order 2 or higher order.

2.3. Boundary Kernel Method. The boundary kernel estimate at a particular point of estimation in the boundary region is obtained by first constructing the appropriate kernel for that point. Many researchers including [7, 10–12] have explored this approach. The method applies a different kernel for estimating function at each point in the boundary region. Due to this, some kernels may not hold the symmetry property and can therefore put more weight on the positive axis. The estimator for $m(x)$ using this method is given by

$$m(\widehat{x})=\frac{1}{nb}\sum_{i=1}^{n}K_{c/h(c)}\left(\frac{x-X_i}{b_c}\right), \tag{7}$$

where $x=cb$, $0\leqslant c\leqslant 1$, and $h(c)=2-c$. Besides, K_c is such that for $0\leqslant c<1$

$$K_{(c)}(z)=\frac{12}{(1+c)^4}(1+z)\left\{(1-2c)z+\frac{3c^2-2c+1}{2}\right\}\quad\{-1\leqslant z\leqslant 1\}. \tag{8}$$

The boundary kernel and related methods usually have low bias but the price for that is an increase in variance. It has been observed, see, for instance, [11], that approaches involving only kernel modifications without regard to data, such as boundary kernel method, are always associated with larger variance. Besides, the corresponding estimates tend to take negative values near the boundary points. This is due to the fact that some kernels may not be symmetric and can therefore put more weight on the positive axis. These drawbacks limit the use of this method.

2.4. Transformation of Data Method. This technique has been discussed by [4, 13]. Original data is transformed, that is, $X_1, \ldots, X_n$ is transformed to $g(X_1), \ldots, g(X_1)$, while keeping the original data, where g is a nonnegative, continuous, and monotonically increasing function for $[0, \infty) \longrightarrow [0, \infty)$. To use this method one can take a one-to-one continuous function: $[0, \infty) \longrightarrow [0, \infty)$. A regular kernel estimator is then used with the transformed data set $\{g(X_1), g(X_2), \ldots, g(X_n)\}$. The estimator is given by

$$m(\hat{x}) = \frac{1}{nb} \sum_{i=1}^{n} \left(\frac{x - g(X_i)}{b} \right). \tag{9}$$

This method gives the estimator of the probability density function of $g(X)$ not that of X. The strength of this method is that transformation-based boundary correction estimates are nonnegative and have low variance. The nonnegativity property is very vital in practical applications and it is therefore worth-exploring to consider methods that result in nonnegative estimators.

A modified version of this method is therefore proposed in this study since it is not computationally intensive and is easier to implement compared to the rest of the methods.

In the next section, the estimator of the finite population mean is modified using transformation of data technique, and further, its asymptotic properties are derived.

3. Proposed Estimator of Finite Population Mean Using Modified Transformation of Data Method

Consider a finite population of size N consisting of M clusters with N_j elements in the jth cluster. Let y_{ij} denote the value of the survey variable y for unit i in cluster j, for $i = 1, 2, \ldots, N$; $j = 1, 2, \ldots, M$. To estimate the nonresponse values in the second stage of sampling a linear regression model given in (1) is used. Auxiliary data is assumed to be known throughout the study and is therefore used to predict the nonresponse values. The estimator proposed by [1] suffers from boundary bias. To obtain a nonparametric regression estimator for the finite population mean that resolves boundary bias, the function of auxiliary variables given in (10) below is used to predict the nonresponse values of the survey variable Y_{ij}; the estimator is defined by

$$m_{TDM}\left(\widehat{x_{ij}}\right) = \frac{1}{mnb} \cdot \sum_{i=1}^{n} \sum_{j=1}^{m} \left\{ K\left(\frac{x_{ij} - X_{ij}}{b} \right) + K\left(\frac{x_{ij} + g\left(X_{ij}\right)}{b} \right) \right\}, \tag{10}$$

where $m_{TDM}(\hat{x})$ is the function of auxiliary variables due to modified transformation of data method proposed. Following the work of [9], data should be generated beyond the left endpoint of the support of the density function g such that the data provides a natural adjustment of the density g outside its support. The method of data generation procedure combines the transformation and the reflection of data methods. To do this, first transform the original data X_{ij}, $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, m$ to $g(X_{ij})$, $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, m$ while retaining the original data where g is a nonnegative continuous and monotonically increasing function from $[0, \infty)$ to $[0, \infty)$. Secondly, reflect $g(X_{11}), \ldots, g(X_{mn})$ around the origin so that we have $-g(X_{11}), \ldots, -g(X_{mn})$. Consequently, using the enlarged data sample $-g(X_{ij})$, $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, m$ the new estimator of the population mean is defined by

$$E\left(\hat{\bar{\bar{Y}}}\right) = \frac{1}{M} \left\{ \sum_{i=1}^{n} \sum_{j=1}^{m} Y_{ij} + \sum_{i=n+1}^{N} \sum_{j=m+1}^{M} \widehat{y_{ij}} \right\}, \tag{11}$$

where $\widehat{y_{ij}}$ represents the estimator of the nonresponse units and can be rewritten as

$$\widehat{y_{ij}} = m_{MTD}\left(\widehat{x_{ij}}\right) = \sum_{i} \sum_{j} W_{ij}^{*}\left(x_{ij}\right) Y_{ij}, \tag{12}$$

where W_{ij}^{*} are the modified weights arising from the proposed procedure. From (12), the following equation is obtained:

$$m_{TDM}\left(\widehat{x_{ij}}\right) = \frac{\sum_{i=n+1}^{N} \sum_{j=m+1}^{M} \left\{ K\left(\left(x_{ij} - X_{ij}\right)/b\right) + K\left(\left(x_{ij} + g\left(X_{ij}\right)\right)/b\right) \right\} Y_{ij}}{\sum_{i=n+1}^{N} \sum_{j=m+1}^{M} \left\{ K\left(\left(x_{ij} - X_{ij}\right)/b\right) + K\left(\left(x_{ij} + g\left(X_{ij}\right)\right)/b\right) \right\}}. \tag{13}$$

Using (13) the expected value of the estimator of the population mean is therefore given by

$$E\left(\hat{\bar{\bar{Y}}}\right) = \frac{1}{M} \left\{ \sum_{i=1}^{n} \sum_{j=1}^{m} Y_{ij} + \frac{\sum_{i=n+1}^{N} \sum_{j=m+1}^{M} \left\{ K\left(\left(x_{ij} - X_{ij}\right)/b\right) + K\left(\left(x_{ij} + g\left(X_{ij}\right)\right)/b\right) \right\} Y_{ij}}{\sum_{i=n+1}^{N} \sum_{j=m+1}^{M} \left\{ K\left(\left(x_{ij} - X_{ij}\right)/b\right) + K\left(\left(x_{ij} + g\left(X_{ij}\right)\right)/b\right) \right\}} \right\}. \tag{14}$$

In what follows, the bias and variance of the proposed estimator are derived.

3.1. Asymptotic Bias of the Proposed Estimator

Introduction. Boundary bias occurs in the interval $[0, b)$ due to lack of data following the reduction of such data at this interval. This implies that the density function has continuity on $[0, \infty)$ and is 0 for $x_{ij} < 0$. Due to reduced amount of data, the resulting estimators are biased. This is possible if the selected bandwidth is greater than the value of x_{ij}, i.e., if $b > x_{ij}$. Consider the Nadaraya-Watson estimator given by (4.4) in [1]. In addition, consider $m(\widehat{a, b})$ for $a \in [0, 1)$ where $x_{ij} = ab$ so that for $w = (x_{ij} - X_{ij})/b$ one can obtain

$$0 \leqslant X_{ij} \leqslant \infty \Longrightarrow 0 \leqslant b(a - w) \leqslant \infty, \tag{15}$$

so that $-\infty \leqslant w \leqslant a$. Next, consider a kernel estimator given in (1) which has the support $[-1, 1]$; this means the variable w must be contained in the interval $[-1, 1]$, so that for $a \in [0, 1)$ we have $-1 \leqslant w \leqslant a$. Since the main problem is to estimate the nonresponse component of the proposed estimator, the following theorem due to [9], under certain conditions on $g(\cdot)$ and $m(\cdot)$ outlined below, is applied.

Theorem 1. *Assume that $m''(\widehat{x_{ij}})$ and $g''(\widehat{x_{ij}})$ exist and are continuous, where $g(x_{ij}) = x_{ij} + dx_{ij}^2 + Ad^2x_{ij}^3$ such that $A > 0$ and $d = m'(0)/m(0)$. Assume that $g^{-1}(0) = 0$, $g'(0) = 1$, $g''(0) = 2m'(0)/m(0)$, and $m^{(0)} = m$, $g^{(0)} = g$, where g^{-1} is the inverse function of g while $m^{(i)}$ and $g^{(i)}$ are the ith derivatives of m and g, respectively, for $i \geqslant 0$. Furthermore, let $x = ab$, where $0 \leqslant a \leqslant 1$. Assume the kernel function K is nonnegative, symmetric function with support $[-1, 1]$ such that $\int K(w)dw = 1$, $\int wK(w)dw = 0$, and $0 < \int w^2k(w)dw < \infty$.*

Using Theorem 1, the expected value of the nonresponse component is given by

$$E\left(m_{TDM}\left(\widehat{x_{ij}}\right)\right) = \frac{1}{b}E\left[K\left(\frac{x_{ij} - X_{ij}}{b}\right) + K\left(\frac{x_{ij} + g\left(X_{ij}\right)}{b}\right)\right]. \tag{16}$$

$$\begin{aligned} &= m\left(x_{ij}\right)\int_{-1}^{a} K(w)\,dw - bm'\left(x_{ij}\right)\int_{-1}^{a} wK(w)\,dw \\ &\quad + \frac{m''\left(x_{ij}\right)b^2}{2} \times \int_{-1}^{a} w^2K(w)\,dw \\ &\quad + \frac{1}{b}EK\left(\frac{x_{ij} + g\left(X_{ij}\right)}{b}\right) + o\left(b^2\right). \end{aligned} \tag{17}$$

And

$$\begin{aligned} &\frac{1}{b}EK\left(\frac{x_{ij} + g\left(X_{ij}\right)}{b}\right) \\ &\quad = \frac{1}{b}\int_{0}^{\infty} K\left(\frac{x_{ij} + g\left(y_{ij}\right)}{b}\right) f(y)\,dy. \end{aligned} \tag{18}$$

Using change of variables technique and on simplification the result becomes

$$\begin{aligned} &\frac{1}{b}EK\left(\frac{x_{ij} + g\left(X_{ij}\right)}{b}\right) \\ &\quad = \int_{a}^{1} K(w)\frac{f\left(g^{-1}(w-a)b\right)}{g'\left(g^{-1}(w-a)b\right)}dw, \end{aligned} \tag{19}$$

which simplifies to

$$\begin{aligned} \frac{1}{b}EK\left(\frac{x_{ij} + g\left(X_{ij}\right)}{b}\right) &= \int_{a}^{1} K(w)\left\{\frac{m\left(g^{-1}(0)\right)}{g'\left(g^{-1}(0)\right)} + (w-a)b \times \frac{g'\left(g^{-1}(0)\right)m'\left(g^{-1}(0) - g''\left(g^{-1}(0)m\left(g^{-1}(0)\right)\right.\right.}{\left[g'\left(g^{-1}(0)\right)\right]^3}\right. \\ &\quad + \frac{b^2}{2}(w-a)^2 \times \left[\frac{g'\left(g^{-1}(0)m''\left(g^{-1}(0)\right) - g''\left(g^{-1}(0)m\left(g^{-1}(0)\right)\right.\right.}{\left[g'\left(g^{-1}(0)\right)\right]^4}\right. \\ &\quad \left.\left. - \frac{3g''\left(g^{-1}(0)\right)\left[g'\left(g^{-1}(0)m'\left(g^{-1}(0)\right) - g''\left(g^{-1}(0)m\left(g^{-1}(0)\right)\right)\right]}{\left[g'\left(g^{-1}(0)\right)\right]^5}\right]\right\}dw + o\left(b^2\right), \end{aligned} \tag{20}$$

which on simplification yields

$$\begin{aligned} \frac{1}{b}EK\left(\frac{x_{ij} + g\left(X_{ij}\right)}{b}\right) &= m(0)\int_{a}^{1} K(w)\,dw + b\int_{a}^{1}(w-a)K(w)\,dw\left[m'(0) - g''(0)m(o)\right] + \frac{b^2}{2} \\ &\quad \cdot \int_{a}^{1}(w-a)^2K(w)\,dw\left\{m''(0) - g'''(0)m(0) - 3g''(0) \times \left[m'(0) - g''(0)m(0)\right]\right\} + o\left(b^2\right). \end{aligned} \tag{21}$$

Substituting (21) into (17) the following is obtained:

$$\begin{aligned} E\left(m_{TDM}\left(\widehat{x_{ij}}\right)\right) &= m\left(x_{ij}\right)\int_{-1}^{a} K(w)\,dw + m(0) \\ &\cdot \int_{a}^{1} K(w)\,dw + b\left\{-m'\left(x_{ij}\right)\int_{-1}^{a} wK(w)\,dw \right. \\ &+ \left.\int_{a}^{1}(w-a)K(w)\,dw \times \left[m'(0) - g''(0)m(0)\right]\right\} \\ &+ \frac{b^2}{2}m''\left(x_{ij}\right)\int_{-1}^{a} w^2K(w)\,dw + \frac{b^2}{2} \\ &\cdot \int_{a}^{1}(w-a)^2K(w)\,dw + \left\{m''(0)\right. \\ &- \left.g'''(0)m(0) - 3g''(0)\left[m'(0) - g''(0)m(0)\right]\right\} \\ &+ o\left(b^2\right). \end{aligned} \tag{22}$$

Since $f''(0)$ exists and is continuous near 0, then for $x = ab$, we have

$$\begin{aligned} m(0) &= m\left(x_{ij}\right) - abm'\left(x_{ij}\right) + \frac{(ab)^2}{2}m''\left(x_{ij}\right) \\ &\quad + o\left(b^2\right) \\ m'\left(x_{ij}\right) &= m'(0) + abm''(0) + o(b) \\ m''\left(x_{ij}\right) &= m''(0) + o(1). \end{aligned} \tag{23}$$

Hence simplifying (22) gives

$$\begin{aligned} E\left(m_{TDM}\left(\widehat{x_{ij}}\right)\right) &= m\left(x_{ij}\right) + b\int_{a}^{1}(w-a)K(w)\,dw\left[2m'(0) - g''(0)m(0)\right] + \frac{b^2}{2}m''(0)\int_{-1}^{1} w^2K(w)\,dw - \frac{b^2}{2}\int_{a}^{1}(w \\ &- a)^2K(w)\,dw\left\{g''(0)m(0) + 3g''(0)\left[m'(0) - g''(0)m(0)\right]\right\} \\ &+ o\left(b^2\right). \end{aligned} \tag{24}$$

Thus the bias of the estimator of the nonresponse component in (14) can be expressed as

$$\begin{aligned} E\left(m_{TDM}\left(\widehat{x_{ij}}\right)\right) - m\left(x_{ij}\right) &= b\int_{a}^{1}(w-a)K(w)\,dw\left[2m'(0) - g''(0)m(0)\right] + \frac{b^2}{2}m''(0)\int_{-1}^{1} w^2K(w)\,dw - \frac{b^2}{2}\int_{a}^{1}(w \\ &- a)^2K(w)\,dw\left\{g''(0)m(0) + 3g''(0)\left[m'(0) - g''(0)m(0)\right]\right\} \\ &+ o\left(b^2\right). \end{aligned} \tag{25}$$

As $b \longrightarrow 0$, it is noted that $E(m_{TDM}(\widehat{x_{ij}})) - m(x_{ij}) \approx 0$. This shows that, for the bias to reduce, the bandwidth must tend to zero as the sample size increases; that is, as $b \longrightarrow 0$, $mn \longrightarrow \infty$.

3.2. Asymptotic Variance of the Proposed Estimator. The variance of the estimator proposed is given as

$$\begin{aligned} \operatorname{var}\left(m_{TDM}\left(\widehat{x_{ij}}\right)\right) &= \frac{1}{(mn)^2b^2} \\ &\cdot \operatorname{var}\left\{\sum_{i=n+1}^{N}\sum_{j=m+1}^{M} K\left(\frac{x_{ij} - X_{ij}}{b}\right)\right. \\ &+ \left.K\left(\frac{x_{ij} + g\left(X_{ij}\right)}{b}\right)\right\}, \end{aligned} \tag{26}$$

$$\begin{aligned} \operatorname{var}\left(m_{TDM}\left(\widehat{x_{ij}}\right)\right) &= \frac{1}{(mn)^2b^2} \\ &\cdot \sum_{i=n+1}^{N}\sum_{j=m+1}^{M} E\left\{K\left(\frac{x_{ij} - X_{ij}}{b}\right)\right. \\ &+ K\left(\frac{x_{ij} + g\left(X_{ij}\right)}{b}\right) \\ &- \left.E\left[K\left(\frac{x_{ij} - X_{ij}}{b}\right) + K\left(\frac{x_{ij} + g\left(X_{ij}\right)}{b}\right)\right]\right\}^2. \end{aligned} \tag{27}$$

which can be rewritten as follows:

$$\operatorname{var}\left(m_{TDM}\left(\widehat{x_{ij}}\right)\right) = \frac{(Mn - mn)^2}{(mn)^2b^2}E\left[K\left(\frac{x_{ij} - X_{ij}}{b}\right)\right.$$

$$+K\left(\frac{x_{ij}+g\left(X_{ij}\right)}{b}\right)\Bigg]^2 - \frac{(Mn-mn)^2}{(mn)^2 b^2}\left\{E\left[K\left(\frac{x_{ij}-X_{ij}}{b}\right) + K\left(\frac{x_{ij}+g\left(X_{ij}\right)}{b}\right)\right]\right\}^2 = A+B, \tag{28}$$

where

$$\begin{aligned} A &= \frac{(Mn-mn)^2}{mnb^2}\int\left[K\left(\frac{x_{ij}-y_{ij}}{b}\right) + K\left(\frac{x_{ij}+g\left(y_{ij}\right)}{b}\right)\right]^2 f(y)\,dy \\ &= \frac{(Mn-mn)^2}{mnb^2}\left\{\int\left[K\left(\frac{x_{ij}-y_{ij}}{b}\right)^2 f(y)\,dy + \int K\left(\frac{x_{ij}+g\left(y_{ij}\right)}{b}\right)^2 f(y)\,dy\right] + 2\int K\left(\frac{x_{ij}-y_{ij}}{b}\right)K\left(\frac{x_{ij}+g\left(y_{ij}\right)}{b}\right) f(y)\,dy\right\} \\ &= A_1 + A_2, \end{aligned} \tag{29}$$

and

$$B = -\frac{(Mn-mn)^2}{(mn)^2 b^2}\left\{E\left[K\left(\frac{x_{ij}-X_{ij}}{b}\right) + K\left(\frac{x_{ij}+g\left(X_{ij}\right)}{b}\right)\right]\right\}^2. \tag{30}$$

$$A_1 = \frac{(Mn-mn)^2}{mnb^2}\int\left[K\left(\frac{x_{ij}-y_{ij}}{b}\right) + K\left(\frac{x_{ij}+g\left(y_{ij}\right)}{b}\right)\right]^2 f(y)\,dy. \tag{31}$$

Equation (31) can be expanded to get

$$\begin{aligned} A_1 &= \frac{(Mn-mn)^2}{mnb^2}\left\{\int\left[K\left(\frac{x_{ij}-y_{ij}}{b}\right)^2 f(y)\,dy + \int K\left(\frac{x_{ij}+g\left(y_{ij}\right)}{b}\right)^2 f(y)\,dy\right]\right\} \\ &\quad + 2\left\{\frac{(Mn-mn)^2}{mnb^2}\int\left[K\left(\frac{x_{ij}-y_{ij}}{b}\right) \cdot K\left(\frac{x_{ij}+g\left(y_{ij}\right)}{b}\right)\right] f(y)\,dy\right\}; \end{aligned} \tag{32}$$

using change of variables technique, (32) can be expressed as

$$A_1 = \frac{(Mn-mn)^2}{mnb^2}\left[b\int_{-1}^{a} K(w)^2\, m\left(x_{ij}-bw\right)dw + b\int_{a}^{1} k(w)^2 \times \frac{m\left(g^{-1}\left(x_{ij}-bw\right)\right)}{g'\left(x_{ij}-bw\right)}dw\right], \tag{33}$$

which on simplification reduces to

$$A_1 = \frac{(Mn-mn)^2}{mnb}\int_{-1}^{1} K(w)^2\,dw + O\left(\frac{1}{mnb}\right). \tag{34}$$

Next we have

$$A_2 = 2\frac{(Mn-mn)^2}{mnb^2}\int_{-1}^{a} K(w) \cdot K\left(\frac{x_{ij}+g(a-w)b}{b} \times m(a-w)b\,dw\right. . \tag{35}$$

Since $g''(\cdot)$ is continuous, Taylor's series expansion gives

$$g(a-w)b = (a-w)b + O\left(b^2\right). \tag{36}$$

Replacing $g(a-w)b$ given by (36) in (35) yields

$$A_2 = 2\frac{(Mn-mn)^2}{mnb^2}\int_{-1}^{a} K(w)\,K((2a-w)+O(b)) \cdot m((a-w)b)\,dw, \tag{37}$$

which reduces on simplification to

$$A_2 = \frac{(Mn-mn)^2}{mnb^2}\int_{-1}^{1} K(w)\,K(2a-w)\,dw + O\left(\frac{1}{mnb}\right). \tag{38}$$

Next is to evaluate B which as earlier outlined in (30) is given by

$$B = -\frac{(Mn-mn)^2}{(mn)^2 b^2}\left\{E\left[K\left(\frac{x_{ij}-X_{ij}}{b}\right) + K\left(\frac{x_{ij}+g\left(X_{ij}\right)}{b}\right)\right]\right\}^2. \tag{39}$$

Applying Taylor's series expansion and following the same procedure as for A, B would simplify to

$$B = O\left(\frac{1}{mnb}\right). \tag{40}$$

Hence putting together A, i.e., $A_1 + A_2$ and B we have

$$\begin{aligned}\operatorname{var}\left(m_{TDM}\left(\widehat{x_{ij}}\right)\right) &= A + B \\ &= \frac{(Mn - mn)^2}{mnb}\left[\int_{-1}^{1} K(w)^2\, dw \right. \\ &\left. + 2\int_{-1}^{a} K(w)\, K(2a - w)\, dw\right] + O\left(\frac{1}{mnb}\right).\end{aligned} \tag{41}$$

It can be noted that the variance, $\operatorname{var}(m_{TDM}(\widehat{x_{ij}}))$, is decreasing in mnb. This is because as $mn \longrightarrow \infty$, the bandwidth $b \longrightarrow 0$ and hence $mnb \longrightarrow \infty$; that is, the bandwidth decreases but not at a faster rate than the sample size. Thus for a large sample size, the variance is reduced significantly. Both the bias and the variance must become small as $mn \longrightarrow \infty$ for the estimator to be optimal. That means the bandwidth has to decrease but not at a faster rate than the sample size. This suffices to establish the consistency of the estimator. That is, for all x_{ij}, $m_{TDM}(\widehat{x_{ij}}) \longrightarrow m(x_{ij})$ in probability as $mn \longrightarrow \infty$.

4. Simulation Study and Discussion of the Results

The simulation study was carried out using R statistical Package (R code). To obtain the estimator for the finite population mean, $\overline{\overline{Y}}$, the auxiliary variables X_{ij} were generated as identically and independently distributed random variables on $U(0, 1)$. The population consisted of 30 clusters. In stage one, a sample of $m_i = 10$ clusters was chosen using simple random sampling with replacement (*SRSWR*) which constituted the primary sampling units (PSUs).

In stage two, from each selected clusters, say i, ($i = 1, \ldots, m_i$), a sample m_{ij}, $j = 1, 2, \ldots, 100$ from $j = 1, 2, \ldots, m_k, \ldots, 900$ was selected; that is, the jth sample from a fixed selected i^{th} cluster was selected using *SRSWR* from a total $M_k = 900$ elements.

Consider the survey variables Y_{ij}, $j = 1, 2, \ldots, m_k, \ldots, M_k$ that are known only for the respondents in the sample. Using known auxiliary variables, x_{ij}, $j = 1, 2, \ldots, m_k$, nonresponse values were generated using the model $\widehat{Y}_{ij} = m(\widehat{x}_{ij}) + \widehat{e}_{ij}$ using SRSWR within the ith cluster.

Moreover, let $K(u) \sim U[0, 1]$, $\widehat{e}_{ij} \sim N(0, 1)$ such that $m(\widehat{x}_{ij})$ is a function of auxiliary random variables generated using linear, sine, and quadratic data functions outlined in the following subsection.

This procedure was repeated iteratively to obtain $\widehat{\overline{\overline{Y}}}_{i1}, \ldots, \widehat{\overline{\overline{Y}}}_{in}$. 95% confidence intervals (CI) were then constructed for the estimators of population means $\widehat{\overline{\overline{Y}}}_i$, $i = 1, 2$ which corresponded to the proposed estimator and Nadaraya-Watson estimators of finite population means, respectively.

A normal kernel with mean 0 and variance 1 was used since it has smooth and continuous derivatives at every data point. To maintain stability in terms of the variation of the random values simulated, an optimal bandwidth obtained using the cross-validation technique was used.

Table 1: Equations of data functions simulated.

Data function	Equation
Linear	$1 + 2(x - 0.5)$
Quadratic	$1 + 2(x - 0.5)^2$
Sine	$2 + \sin(2\pi x)$
Bump	$1 + 2(x - 0.5) + \exp\left\{-200(x - 0.5)^2\right\}$

Table 2: Summary results of bias.

Data function	Proposed Estimator	Nadaraya-Watson Estimator
Linear	-0.16774	-0.40747
Quadratic	0.02981	0.06921
Sine	-0.29083	-1.17087
Bump	-0.18757	-0.51888

4.1. Equations of Data Functions of $m(\widehat{x}_{ij})$ Simulated. These data functions are normally used in statistics for data simulations since they are widely applicable in real life; see, for instance, [14, 15]. Sine functions are used to model periodic events such as light waves and average temperature variations throughout the year while quadratic functions are used in physics to describe trajectory followed by objects thrown upward at an angle whereas in economics quadratic functions can be used to develop profit and loss functions. Linear functions are widely applicable, for example, in establishing the relationships between a dependent variable and two or more independent variables, e.g., analyzing the linear relationship between the price, supply, and demand of a commodity. Bump functions used in such events as bio-surveillance for modeling disease-outbreaks or floods within a certain limit of time in a given place and can also be used in curve fitting, uncertainty analysis, and approximation of nonlinear relationships in scattered data. Equations of data functions simulated are presented in Table 1.

4.2. Simulation Results. The results of the data simulated are presented in Tables 2, 3 and 4. For details on Nadaraya-Watson estimator see [16]. The Nadaraya-Watson estimator was used for comparison with the proposed estimator.

It can be noted in Table 2 that the values of the bias for the proposed estimator are relatively smaller than those of the Nadaraya-Watson estimator for all the data functions simulated except for the quadratic data function where the values of the bias are close to each other for the two estimators. This may be attributed to the reflection of the transformed data at the boundaries of the support of the kernel density function used. The transformation of data method was proposed to address the boundary bias arising from Nadaraya-Watson technique. Hence the proposed estimator clearly resolved the bias due to Nadaraya-Watson estimation technique.

Efficiency of the mean estimator of the population mean was obtained by its MSE. This is illustrated in Table 3. Measures for the MSE were simulated for purposes of comparison. Comparatively, the proposed estimator of the finite population mean outperforms the Nadaraya-Watson estimator in terms of efficiency as noted from the table. This

Table 3: Summary results of MSEs.

Data function	Proposed Estimator	Nadaraya-Watson Estimator
Linear	0.04867	0.19015
Quadratic	0.04451	0.56209
Sine	0.13688	1.41025
Bump	0.06229	0.32087

Table 4: Summary results of 95% confidence interval lengths.

Data function	Proposed Estimator	Nadaraya-Watson Estimator
Linear	0.92695	1.70934
Quadratic	0.76973	2.93895
Sine	1.52697	4.88850
Bump	0.97842	2.22051

is because the MSE of the proposed estimator is relatively smaller compared to Nadaraya-Watson estimator in all the data functions. Hence the proposed estimator is more efficient than the Nadaraya-Watson estimator as illustrated in Table 3.

The 95% upper and lower confidence intervals were generated for the estimators of finite population mean using the formula $\overline{\overline{Y}} = \hat{\overline{\overline{Y}}} \pm Z_{\alpha/2}(\sqrt{\mathrm{var}(\hat{\overline{\overline{Y}}})})$ and subsequently the confidence interval lengths were obtained. The results of these confidence interval lengths are presented in Table 4. A good confidence interval has a coverage rate closer to the true population mean being estimated and therefore its length has to be small. From Table 4, it can be observed that the confidence interval lengths for the proposed estimator are much smaller than those of Nadaraya-Watson estimator in all the data functions simulated. Therefore, it can be concluded that the estimator developed in this paper has a tighter confidence interval length and is superior to its rival Nadaraya-Watson estimator at 95% coverage rate.

5. Conclusion

The proposed estimator of finite population mean has been shown to be better than the Nadaraya-Watson estimator using various performance criteria such as the bias, the mean squared error, and the confidence interval lengths. The results are tabulated in Tables 2, 3, and 4. Most importantly, it is shown in Table 4 that the proposed estimator has got tighter confidence interval lengths at 95% level; hence it produces estimates that are closer to the true population values being estimated.

Data Availability

The data used to support the theoretical findings were generated via simulation using R statistical package.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] N. Kiprono Bii, O. C. Onyango, and J. Odhiambo, "Estimating a finite population mean under random non-response in two stage cluster sampling with replacement," *Journal of Statistics*, vol. 7, no. 05, p. 834, 2017.

[2] C. Ouma and C. Wafula, "Bootstrap confidence intervals for model-based surveys," *East African Journal of Statistics*, vol. 1, no. 1, pp. 84–90, 2005.

[3] O. C. Onyango, R. O. Otieno, and G. O. Orwa, "Generalised model based confidence intervals in two stage cluster sampling," *Pakistan Journal of Statistics and Operation Research*, vol. 6, no. 2, pp. 127–139, 2010.

[4] J. S. Marron and D. Ruppert, "Transformations to reduce boundary bias in kernel density estimation," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 56, no. 4, pp. 653–671, 1994.

[5] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, London, UK, 1986.

[6] P. Malec and M. Schienle, "Nonparametric kernel density estimation near the boundary," *Computational Statistics & Data Analysis*, vol. 72, pp. 57–76, 2014.

[7] M. C. Jones, "Simple boundary correction for kernel density estimation," *Statistics and Computing*, vol. 3, no. 3, pp. 135–146, 1993.

[8] J. S. Simonoff, *Smoothing Methods in Statistics*, Springer, New York, NY, USA, 2012.

[9] A. Cowling and P. Hall, "On pseudodata methods for removing boundary effects in kernel density estimation," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 3, pp. 551–563, 1996.

[10] T. Gasser and H.-G. Müller, "Kernel estimation of regression functions," in *Smoothing Techniques for Curve Estimation*, Lecture Notes in Math., pp. 23–68, Springer, Berlin, 1979.

[11] H.-G. Müller, "Smooth optimum kernel estimators near endpoints," *Biometrika*, vol. 78, no. 3, pp. 521–530, 1991.

[12] S. Zhang and R. J. Karunamuni, "Boundary bias correction for nonparametric deconvolution," *Annals of the Institute of Statistical Mathematics*, vol. 52, no. 4, pp. 612–629, 2000.

[13] M. P. Wand and M. C. Jones, *Kernel Smoothing, ser. Monographs on Statistics and Applied Probability*, Chapman and Hall, New York, NY, USA, 1995.

[14] J.-Y. Kim, F. J. Breidt, and J. D. Opsomer, "Nonparametric regression estimation of finite population totals under two-stage sampling," *Preprint*, 2003.

[15] R. C. Langat, *Robust Estimation of Finite Population Total Incorporating Data-Reflection Technique in Nonparametric Regression [Ph.D. thesis]*, JKUAT, COPAS, USA, 2017.

[16] A. H. Dorfman, "Nonparametric regression for estimating totals in finite populations," in *Proceedings of the Section on Survey Research Methods*, pp. 622–625, American Statistical Association Alexandria, Alexandria, EGYPT, 1992.