Scientific Research Publishing

Open Journal of Statistics, 2017, 7, 1053-1066
http://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

# A Multiplicative Bias Correction for Nonparametric Approach and the Two Sample Problem in Sample Survey

**Kemtim Tamboun Stephane, Romanus Odhiambo Otieno, Thomas Mageto**

Pan African University Institute of Science, Technology and Innovation, Nairobi, Kenya
Email: kemtimstephane1@gmail.com

**How to cite this paper:** Stephane, K.T., Otieno, R.O. and Mageto, T. (2017) A Multiplicative Bias Correction for Nonparametric Approach and the Two Sample Problem in Sample Survey. *Open Journal of Statistics*, **7**, 1053-1066.
https://doi.org/10.4236/ojs.2017.76073

**Received:** October 17, 2017
**Accepted:** December 26, 2017
**Published:** December 29, 2017



## Abstract

Let two separate surveys collect related information on a single population U. Consider situation where we want to best combine data from the two surveys to yield a single set of estimates of a population quantity (population parameter) of interest. This Article presents a multiplicative bias reduction estimator for nonparametric regression to two sample problem in sample survey. The approach consists to apply a multiplicative bias correction to an estimator. The multiplicative bias correction method which was proposed, by Linton & Nielsen, 1994, assures a positive estimate and reduces the bias of the estimate with negligible increase in variance. Even as we apply this method to the two sample problem in sample survey, we found out through the study of it asymptotic properties that it was asymptotically unbiased, and statistically consistent. Furthermore an empirical study was carried out to compare the performance of the developed estimator with the existing ones.

## Keywords

Multiplicative Bias Correction, Two Sample Problem, Bias

## 1. Introduction

Sometimes, it happens that two separate surveys gather related information on a variable of interest of a population, U, having perhaps distinct designs and mode of sampling. It becomes very important on how to combine the data from the two surveys.

Take as example, the students of the sub-regional institute of statistics and apply economics (ISSEA), and those of the polytechnic institute, both in different ways with different importances to collect data on unemployment in

Cameroon. Researchers at the national institute of statistics (Cameroon) are faced with the following problem: how can the data from these two distinct surveys joined together to produce a single data and have a better representation of the population?

Some great scientists have been looking into these problems for several years. The approach to this problem have been in different ways; one of which involve getting estimates of the two surveys separately and using the inverse of the estimated variances as weights to weigh them together as seen in [1]. [2] went further by using empirical likelihood method to combine information from multiple survey. Another option to this consist of putting the two data sets in a single data set, taking into account the weight on individual sample units. Developed in [3] are some of these methods which include; the pseudo-likelihood, missing information principle and iterated post-stratified estimator. After simulations on two different populations, it was concluded that, in neither population the design based ways of combining data yield best results. The iterated post-stratified estimator looks to be a very promising non-parametric way to combined data from two sources.

Just recently [4] used the Nonparametric regression, which is the model-based sampler's method of choice when there is a serious doubt about the suitability of a linear or other simple parametric models for the survey data at hand. The nonparametric regression supersedes the need for use of design weights and standard design-based weights. Recognition of this is especially helpful in confronting problems in sampling situations where design weights are missing or questionable.

This study made use of kernel smoothers, especially the Nadaraya Watson smoother. However, estimators based on Nadaraya Watson smoothing weights are normally biased in small samples and at boundary points.

There exist alternative techniques of reducing the bias. For a detailed review see [5]-[11]. These methods improve the performance of nonparametric regression at points of large curvature. But in this framework, we consider a multiplicative bias correction approach to nonparametric regression to have an estimate with a smaller bias than existing ones.

### Outline of the Paper

The remaining part of this paper is organized as follows: In Section 2, a multiplicative bias corrected estimator $\hat{T}_{MBC}$ for the finite population totals is proposed. In Section 3, the asymptotic properties of the proposed estimator are derived. In Section 4, an empirical study of the derived properties is presented. In Section 5 we give a conclusion to the paper.

## 2. Proposed Estimator

Consider a finite population, $U = 1, 2, \cdots, N$ and let $y_1, y_2, \cdots, y_n$ represent the combined random sample drawn from the population using different sampling

techniques. Suppose that to each of these $y_i's$, there is an auxiliary information $x_1, x_2, \cdots, x_n$.

Let consider the following model;

$$E\left(Y_i / X_i = x_i\right) = h\left(x_i\right) \tag{1}$$

$$\operatorname{cov}\left(Y_i, Y_j / X_i = x_i, X_j = x_j\right) = \begin{cases} \sigma^2\left(x_i\right), i = j \\ 0, i \neq j \end{cases} \tag{2}$$

where $h\left(x_i\right)$ and $\sigma^2\left(x_i\right)$ are twice continuously differentiable functions (that is lipschitz continuous). With these assumptions on $h\left(x_i\right)$ and $\sigma^2\left(x_i\right)$, one can estimate $h\left(x_i\right)$ and $\sigma^2\left(x_i\right)$ non-parametrically.

Let $\epsilon_i = Y_i - h\left(X_i\right)$ be i.i.d. with zero mean, and variance $\sigma^2$. We can refer to this set-up as the weak model. In this scheme, we can ignore which of the original samples, the $Y_i's$ are available from.

Usually in the computation of finite population total,we have the formula given by

$$T = \sum_{i \in U} y_i = \sum_{i \in s} y_i + \sum_{j \in r} y_j \tag{3}$$

where, $s$ refers to the sample and $r$ refers to the nonsampled part of the population. Since the values of the sample part is known, the process of estimating the finite population total is equivalent to predicting the nonsample part of the population.

To do this, the multiplicative bias corrected technique is employed in which case the proposed estimator of the population total is now defined as

$$\hat{T}_{MBC} = \sum_{i \in s} \frac{y_i - \hat{h}\left(x_i\right)}{\pi_i} + \sum_{j \in r} \hat{h}\left(x_j\right) \tag{4}$$

where

$\pi_i$ is the inclusion probability

$\hat{h}\left(x_i\right)$ is the multiplicative bias corrected estimator.

The principal objective of the multiplicative bias corrected technique is to correct the insufficiences of the kernel smoother that is the bias problem at the boundaries. Given a pilot smoother of the regression function

$$\tilde{h}\left(x\right) = \sum_{j=1}^{n} w_{xj} Y_j \tag{5}$$

The inverse relative estimation error of the smoother at each of the observations is given by $\frac{h\left(x\right)}{\tilde{h}\left(x\right)}$.

A noisy estimate of the ratio, $\frac{h\left(x\right)}{\tilde{h}\left(x\right)}$, is given by

$$\beta\left(x\right) = \frac{Y_j}{\tilde{h}\left(X_j\right)} \tag{6}$$

Smoothing the noisy estimate $\beta\left(x\right)$ leads to

$$\tilde{\beta}(x)=\sum_{j=1}^{n} w_{xj}\beta(x) \tag{7}$$

Above gives a better estimate for the inverse of the relative estimation error at each particular observation and can therefore be used as a multiplicative correction of the pilot smoother.

$$\hat{h}(x)=\tilde{\beta}(x)\tilde{h}(x) \tag{8}$$

For both $\tilde{h}(x)$ and $\tilde{\beta}(x)$, we use the same weighting scheme;

$$w_{xj}=\frac{1}{nh}K\left(\frac{x-X_j}{h}\right) \tag{9}$$

where

$h$ is the bandwidth

$K$ is a probability density function, symmetric about zero.

$n$ is the sample size

### Bandwidth Selection Techniques

- Implement biased cross-validation (bcv).
- Implement unbiased cross-validation (ucv).
- Implements a rule-of-thumb for choosing the bandwidth of a Gaussian kernel density estimator (ndr0)
- Can use a more common variation given by Scott (1992) (ndr)

## 3. Properties of Proposed Estimator

### 3.1. Assumptions

The following assumptions are made in the estimation of $\hat{h}(x_i)$.

- The regression function is bounded and strictly positive, that is, $b \geq h(x) \geq a > 0$ for all $x$
- The regression function is twice continuously differentiable everywhere.
- $\epsilon$ has finite fourth moments and has a symmetric distribution around zero.
- The bandwidth $h$ is such that, $h \to 0$, $nh \to \infty$ and $(nh)^2 \to \infty$ as $n \to \infty$

### 3.2. Asymptotic Unbiasedness of the Proposed Estimator

We want to show that $E\left(\hat{T}_{MBC}-T\right)\to 0$ as $n \to \infty$. Under the model based, the bias of the estimator $\hat{T}_{MBC}$ is defined as follows;

$$E\left[\hat{T}_{MBC}-T\right]=E\left[\hat{T}_{MBC}\right]-E[T] \tag{10}$$

Now, we have the expected value of the proposed estimator for the finite population total given by;

$$E\left[\hat{T}_{MBC}\right]=E\left[\sum_{i\in s}\frac{y_i-\hat{h}(x_i)}{\pi_i}+\sum_{j\in r}\hat{h}(x_j)\right] \tag{11}$$

$$=E\left[\sum_{i\in s}\frac{y_i-\hat{h}(x_i)}{\pi_i}\right]+E\left[\sum_{j\in r}\hat{h}(x_j)\right] \tag{12}$$

$$= \sum_{i \in s} \frac{1}{\pi_i} E\left(y_i - \hat{h}\left(x_i\right)\right) + \sum_{U|s} E\left(\hat{h}\left(x_j\right)\right) \tag{13}$$

$E\left(\hat{h}\left(x_j\right)\right)$ is obtained by analysing the individual terms of the stochastic approximation of $\hat{h}(x)$. Let us then establish the stochastic approximatiom of $\hat{h}(x)$ as shown by (Hengartner 2009).

From (8),

$$\hat{h}(x) = \tilde{\beta}(x)\tilde{h}(x) \tag{14}$$

$$= \sum_{j=1}^{n} w_{xj} \frac{Y_j}{\tilde{h}\left(X_j\right)} \tilde{h}(x) = \sum_{j=1}^{n} w_{xj} \frac{\tilde{h}(x)}{\tilde{h}\left(X_j\right)} Y_j \tag{15}$$

$$= \sum_{j=1}^{n} w_{xj} R_j(x) Y_j \quad \text{where } R_j(x) = \frac{\tilde{h}(x)}{\tilde{h}\left(X_j\right)} \tag{16}$$

Let define, $\bar{h} = E\left(\tilde{h}(x) \middle| X_1, X_2, \cdots, X_n\right)$ then we can express $R_j(x)$ as.

$$R_j(x) = \frac{\tilde{h}(x)}{\tilde{h}\left(X_j\right)} = \left(\frac{\bar{h}(x)}{\bar{h}\left(X_j\right)}\right) * \left(\frac{\tilde{h}(x)}{\bar{h}(x)}\right) * \left(\frac{\tilde{h}\left(X_j\right)}{\bar{h}\left(X_j\right)}\right)^{-1}$$

$$= \left(\frac{\bar{h}(x)}{\bar{h}\left(X_j\right)}\right) * \left(\frac{\tilde{h}(x) - \bar{h}(x) + \bar{h}(x)}{\bar{h}(x)}\right) * \left(\frac{\tilde{h}\left(X_j\right) - \bar{h}\left(X_j\right) + \bar{h}\left(X_j\right)}{\bar{h}\left(X_j\right)}\right)^{-1}$$

$$= \left(\frac{\bar{h}(x)}{\bar{h}\left(X_j\right)}\right) * \left(\frac{\tilde{h}(x) - \bar{h}(x)}{\bar{h}(x)} + 1\right) * \left(\frac{\tilde{h}\left(X_j\right) - \bar{h}\left(X_j\right)}{\bar{h}\left(X_j\right)} + 1\right)^{-1}$$

$$= \left(\frac{\bar{h}(x)}{\bar{h}\left(X_j\right)}\right) * \left(R(x) + 1\right) * \left(R\left(X_j\right) + 1\right)^{-1}$$

Through the series expansion,

$$\left(R\left(X_j\right) + 1\right)^{-1} = \frac{1}{R\left(X_j\right) + 1} = \frac{1}{1 - \left(-R\left(X_j\right)\right)} = \sum_{n=0}^{\infty} \left[-R\left(X_j\right)\right]^n$$

$$= 1 - R\left(X_j\right) + R\left(X_j\right)^2 + \cdots$$

$$R_j(x) = \frac{\bar{h}(x)}{\bar{h}\left(X_j\right)} * \left[1 + R(x) - R\left(X_j\right) + r_j\left(x, X_j\right)\right]$$

is an approximation of the quantity R.

Replacing both $Y_j$ and $R_j$ in (16), we obtain

$$\hat{h}(x) = \sum_{j=1}^{n} w_{xj} \frac{\bar{h}(x)}{\bar{h}\left(X_j\right)} \left[1 + R(x) - R\left(X_j\right) + r_j\left(x, X_j\right)\right] \left(h\left(X_j\right) + \epsilon_j\right)$$

$$= \sum_{j=1}^{n} w_{xj} \frac{\bar{h}(x)}{\bar{h}\left(X_j\right)} h\left(X_j\right) + \sum_{j=1}^{n} w_{xj} \frac{\bar{h}(x)}{\bar{h}\left(X_j\right)} \left(\epsilon_j + h\left(X_j\right)\right) \left(R(x) - R\left(X_j\right)\right)$$

$$+ \sum_{j=1}^{n} w_{xj} \frac{\bar{h}(x)}{\bar{h}\left(X_j\right)} \left(R(x) - R\left(X_j\right)\right) \epsilon_j + \sum_{j=1}^{n} w_{xj} \frac{\bar{h}(x)}{\bar{h}\left(X_j\right)} r_j\left(x, X_j\right) \left(h\left(X_j\right) + \epsilon_j\right)$$

Using the assumption $nh \to \infty$ the remainder term turns to zero in

probability and the expression reduces to;

$$\hat{h}(x)=\sum_{j=1}^{n} w_{xj}\frac{\bar{h}(x)}{\bar{h}(X_j)}h(X_j)+\sum_{j=1}^{n} w_{xj}\frac{\bar{h}(x)}{\bar{h}(X_j)}\left(\epsilon_j+h(X_j)\right)\left(R(x)-R(X_j)\right)$$
$$+\sum_{j=1}^{n} w_{xj}\frac{\bar{h}(x)}{\bar{h}(X_j)}\left(R(x)-R(X_j)\right)\epsilon_j+0_p\left(\frac{1}{nh}\right)$$

To solve Equation (16), we need to find $E\left(\hat{h}(x_j)\right)$ hence,

$$E\left(\hat{h}(x_j)\right)=E\left[\sum_{j=1}^{n} w_{xj}\frac{\bar{h}(x)}{\bar{h}(X_j)}h(X_j)+\sum_{j=1}^{n} w_{xj}\frac{\bar{h}(x)}{\bar{h}(X_j)}\left(\epsilon_j+h(X_j)\right)\left(R(x)-R(X_j)\right)\right.$$
$$\left.+\sum_{j=1}^{n} w_{xj}\frac{\bar{h}(x)}{\bar{h}(X_j)}\left(R(x)-R(X_j)\right)\epsilon_j+0_p\left(\frac{1}{nh}\right)\right]$$
$$=\sum_{j=1}^{n} w_{xj}\frac{\bar{h}(x)}{\bar{h}(X_j)}E\left(h(X_j)\right)+\sum_{j=1}^{n} w_{xj}\frac{\bar{h}(x)}{\bar{h}(X_j)}E(\epsilon_j)+\sum_{j=1}^{n} w_{xj}\frac{\bar{h}(x)}{\bar{h}(X_j)}h(X_j)$$
$$\times E\left(R(x)-R(X_j)\right)+\sum_{j=1}^{n} w_{xj}\frac{\bar{h}(x)}{\bar{h}(X_j)}\left(R(x)-R(X_j)\right)E(\epsilon_j)+0_p\left(\frac{1}{nh}\right)$$
$$=\sum_{j=1}^{n} w_{xj}\frac{\bar{h}(x)}{\bar{h}(X_j)}h(X_j)+\sum_{j=1}^{n} w_{xj}\frac{\bar{h}(x)}{\bar{h}(X_j)}h(X_j)E\left(\frac{\tilde{h}(x)}{\bar{h}(x)}-\frac{\tilde{h}(X_j)}{\bar{h}(X_j)}\right)+0_p\left(\frac{1}{nh}\right)$$

since $E(\epsilon_j=0)$

$$E\left(\hat{h}(x_j)\right)=\sum_{j=1}^{n} w_{xj}\frac{\bar{h}(x)}{\bar{h}(X_j)}h(X_j)+o_p\left(\frac{1}{nh}\right)\quad\text{since }\bar{h}(x)=E\left(\tilde{h}(x)\right)\tag{17}$$

Hence,

$$E\left[\hat{T}_{MBC}\right]=\sum_{i\in s}\frac{1}{\pi_i}E(y_i)-\left(\sum_{j=1}^{n} w_{xj}\frac{\bar{h}(x)}{\bar{h}(X_i)}h(X_i)+o_p\left(\frac{1}{nh}\right)\right)$$
$$+\sum_{U|s}\left(\sum_{j=1}^{n} w_{xj}\frac{\bar{h}(x)}{\bar{h}(X_i)}h(X_j)\right)+o_p\left(\frac{1}{nh}\right)\tag{18}$$

The above expression can be reduced by considering a limited Taylor series of $\frac{h(X_j)}{\bar{h}(X_j)}$ about a point *x*. Hence

$$\frac{h(X_j)}{\bar{h}(X_j)}=\frac{h(x)}{\bar{h}(x)}+(X_j-x)\left(\frac{h(x)}{\bar{h}(x)}\right)'+(X_j-x)^2\left(\frac{h(x)}{\bar{h}(x)}\right)''+o_p(1)\tag{19}$$

Now, substituting the first two terms in (18) gives

$$E\left[\hat{T}_{MBC}\right]=\sum_{i\in s}\frac{1}{\pi_i}E(y_i)-E\left(\hat{h}(x_i)\right)+\sum_{U|s}\left(\sum_{j=1}^{n} w_{xj}\bar{h}(x)\left(\frac{h(x)}{\bar{h}(x)}+(X_j-x)\left(\frac{h(x)}{\bar{h}(x)}\right)'\right)\right.\tag{20}$$
$$\left.+o_p\left(\frac{1}{nh}\right)\right.$$

But $\sum_{j=1}^{n} w_{xj}=1$ and $\sum_{j=1}^{n}(X_j-x)w_{xj}=0$, therefore

$$E\left[\hat{T}_{MBC}\right]=\sum_{U|s}h(x)+o_p\left(\frac{1}{nh}\right) \tag{21}$$

Furthermore,

$$E(T)=\sum_{i\in s}E(y_i)+\sum_{j\in r}E(y_j)=\sum_{i\in s}\bar{y}+\sum_{j\in r}h(x)$$

Hence the asymptotic bis of the estimator is given by

$$BIAS\left(\hat{T}_{MBC}\right)=E\left(\frac{\hat{T}_{MBC}-T}{N}\right)=\frac{1}{N}\sum_{i\in s}\bar{y}+o_p\left(\frac{1}{nh}\right)$$

The bias of $\hat{T}_{MBC}$ will be of order $o_p\left(\frac{1}{nh}\right)$. Thus it converges to zero at a faster rate compared to the existing non-parametric estimators which generally converge at the rate $o_p\left(h^2\right)$.

### 3.3. Asymptotic Variance of the Proposed Estimator

The variance of the finite population total is given by;

$$\begin{aligned}Var\left[\hat{T}_{MBC}\right]&=Var\left[\sum_{i\in s}\frac{y_i-\hat{h}(x_i)}{\pi_i}+\sum_{j\in r}\hat{h}(x_j)\right]\\&=Var\left[\sum_{i\in s}\frac{y_i-\hat{h}(x_i)}{\pi_i}\right]+Var\left[\sum_{j\in r}\hat{h}(x_j)\right]\\&=\sum_{i\in s}\left(\frac{1}{\pi_i}\right)^2Var\left(y_i-\hat{h}(x_i)\right)+\sum_{U|s}Var\left(\hat{h}(x_j)\right)\end{aligned}$$

Firstly,

$$Var\left(\hat{h}(x_j)\right)=Var\left(\sum_{j=1}^{n}w_{xj}\frac{\bar{h}(x)}{\bar{h}(X_j)}\left[1+R(x)-R(X_j)+r_j(x,X_j)\right]\left(h(X_j)+\epsilon_j\right)\right) \tag{22}$$

Using the assumption $nh\to\infty$, the remainder terms converge to zero in probability. Therefore $r_j(x,X_j)\left(h(X_j)+\epsilon_j\right)=0_p\left(\frac{1}{nh}\right)$ and Equation (22) reduces to

$$Var\left(\hat{h}(x_j)\right)=Var\left(\sum_{j=1}^{n}w_{xj}\frac{\bar{h}(x)}{\bar{h}(X_j)}\left[1+R(x)-R(X_j)\right]\left(h(X_j)+\epsilon_j\right)+0_p\left(\frac{1}{nh}\right)\right) \tag{23}$$

Truncating the binomial expansion at the first term yields

$$\begin{aligned}Var\left(\hat{h}(x_j)\right)&=Var\left(\sum_{j=1}^{n}w_{xj}\frac{\bar{h}(x)}{\bar{h}(X_j)}y_j\right)+0_p\left(\frac{1}{(nh)^2}\right)\\&=\sum_{j=1}^{n}\left(w_{xj}\frac{\bar{h}(x)}{\bar{h}(X_j)}\right)^2\sigma^2(x_j)+0_p\left(\frac{1}{(nh)^2}\right)\end{aligned}$$

Simplify the above expression by considering the first and second part of the Taylor series of $\frac{\sigma^2(x_j)}{\overline{h}^2(X_j)}$. So we obtain

$$Var\left(\hat{h}(x_j)\right) = \sum_{j=1}^{n}\left(w_{xj}\right)^2 \sigma^2(x_j) + 0_p\left(\frac{1}{(nh)^2}\right) \tag{24}$$

Therefore,

$$Var\left[\hat{T}_{MBC}\right] = \sum_{i\in s}\left(\frac{1}{\pi_i}\right)^2 \sigma^2(x_i) + \sum_{U}\sum_{j=1}^{n}\left(w_{xj}\right)^2 \sigma^2(x_j) + 0_p\left(\frac{1}{(nh)^2}\right) \tag{25}$$

Thus the asymptotic variance is given by

$$Var\left(\frac{\hat{T}_{MBC}}{N}\right) = \frac{1}{N^2}\sum_{i\in s}\left(\frac{1}{\pi_i}\right)^2 \sigma^2(x_i) + \frac{1}{N^2}\sum_{U}\sum_{j=1}^{n}\left(w_{xj}\right)^2 \sigma^2(x_j) + 0_p\left(\frac{1}{(nh)^2}\right) \tag{26}$$

This implies that $\hat{T}_{MBC}$ is more efficient than the usual non-parametric regression estimator proposed by Dorfman (1992).

### 3.4. Asymptotic Mean Square Error

The asymtotic mean square error of the estimator $\hat{T}_{MBC}$ is given by

$$MSE\left[\hat{T}_{MBC}\right] = Var\left[\hat{T}_{MBC}\right] + \left[Bias\left(\hat{T}_{MBC}\right)\right]^2 \tag{27}$$

$$\begin{aligned} MSE\left[\hat{T}_{MBC}\right] = {} & \frac{1}{N^2}\sum_{i\in s}\left(\frac{1}{\pi_i}\right)^2 \sigma^2(x_i) + \frac{1}{N^2}\sum_{U}\sum_{j=1}^{n}\left(w_{xj}\right)^2 \sigma^2(x_j) \\ & + 0_p\left(\frac{1}{(nh)^2}\right) + \left[\frac{1}{N}\sum_{i\in s}\overline{y} + o_p\left(\frac{1}{nh}\right)\right]^2 \end{aligned} \tag{28}$$

As $n\to\infty$ and $h\to\infty$, the $MSE\left[\hat{T}_{MBC}\right]$ turns to 0 indicating that, the proposed estimator is statistically consistent.

## 4. Empirical Study

### 4.1. Population

In this section, the theory developed in the previous section was tested using a set of simulation studies, with a mix of survey designs, and employing various approaches to selecting the best bandwidths. We employ a population U of countries in the world of size, $N = 188$, with auxiliary variable $x$ = gross national product (GNI) and variable of interest $y$ = human development index(HDI), of interest is the population total of the HDI, $y = \sum_{l\in U} y_l$.

Figure 1 below shows the scatter diagram of the population. Where HDI is on the vertical axis and GNI on the horizontal axis, where there exist a quadratic relationship between the two variables.

We suppose, for each run of the experiment that two samples are taken:

Sample 1 ( $s_1$ ): srswor $\left(n_1 = 32\right)$

Sample 2 ( $s_2$ ): stratsrs-four strata equal in each, and 8 units taken at random

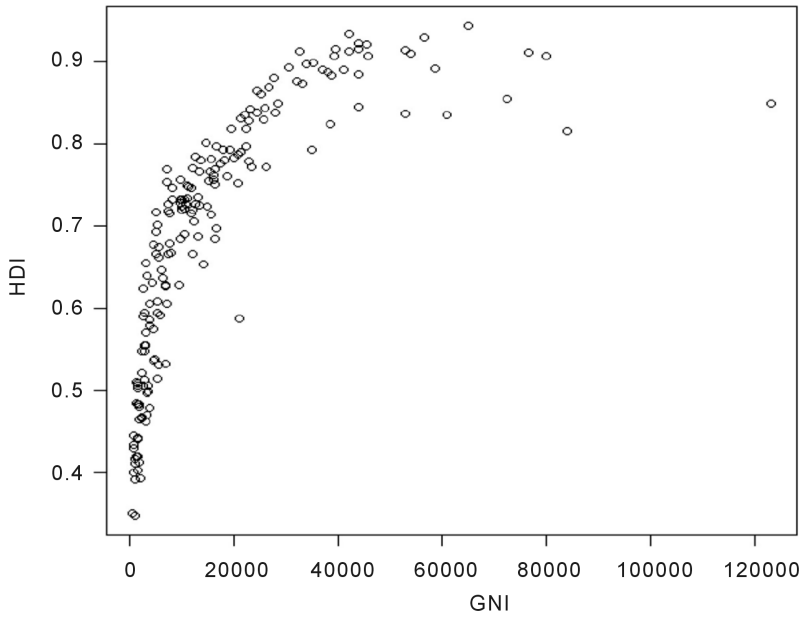



**Figure 1.** Scatter diagram.

in each, so that $n_2 = 32$. The total experiment consists of 500 runs of pairs of samples. Table 1 gives the estimators considered.

For an estimator $\hat{T}$ we considered three measures of relative success across the 500 runs:

i) Unconditional relative bias measured as ratio of mean value (across runs) to target

$$\text{Bias} = \sum_{runs}\left(\hat{T} - T\right)\Big/T$$

ii) Unconditional relative root mean square error divided by target

$$\text{rmse} = \sqrt{\left(\sum_{runs}\left(\hat{T} - T\right)\right)^2}\Big/T$$

## 4.2. Results

Results obtained are tabulated in Table 2.

From the results obtained, we observe that the unbiased cross validation approach is a viable means of selecting bandwidth as it gives the lowest bias and root mean square error across all the estimators. The proposed estimator to the two sample problem gives better estimates of the population total compared to those realized using the estimator proposed by [12], and [4] respectively.

Furthermore, we study the conditional performances of the selected estimators. 500 samples obtained were sorted by the values of the mean of the auxiliary variable and put in 25 groups each containing 20 values. We then compute the bias and root mean square error of each group. The plots of conditional performances against the average of the sorted mean auxiliary variable. We then report the behaviour of the conditional bias for the different bandwidth.

**Table 1.** Estimators.

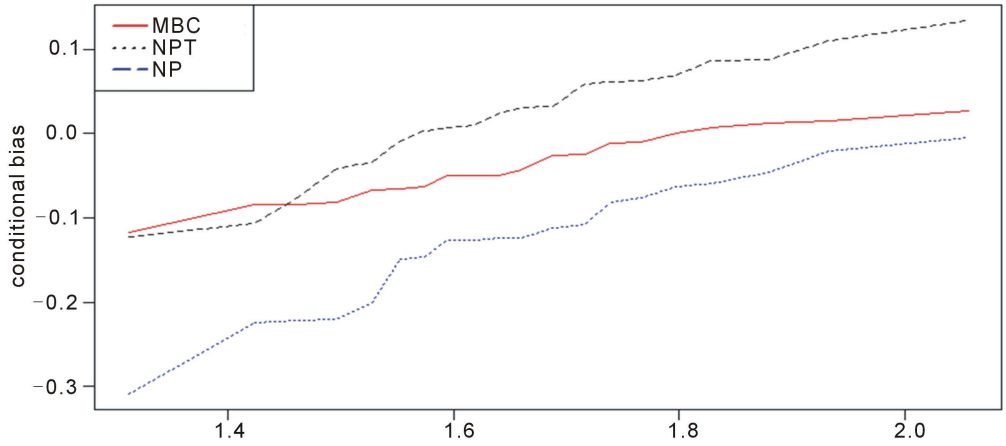
| Estimator | Formula | Comment |
|---|---|---|
| Non parametric (NP) Regression | $\hat{T}_{NP} = \sum_{i\in s} y_i + \sum_{j\in r} \hat{h}(x_j)$ | |
| Nonparametric (NPT) regression, twiced | $\hat{T}_{NPT} = \sum_{i\in s} \frac{y_i - \hat{h}(x_i)}{\pi_i} + \sum_{j\in U} \hat{h}(x_j)$ | $\pi_i =$ Inclusion probabilities |
| Multiplicative (MBC) Bias Corrected | $\hat{T}_{MBC} = \sum_{i\in s} \frac{y_i - \hat{h}(x_i)}{\pi_i} + \sum_{j\in U} \hat{h}(x_j)$ | $\pi_i =$ Inclusion probabilities |

**Table 2.** Empirical results.

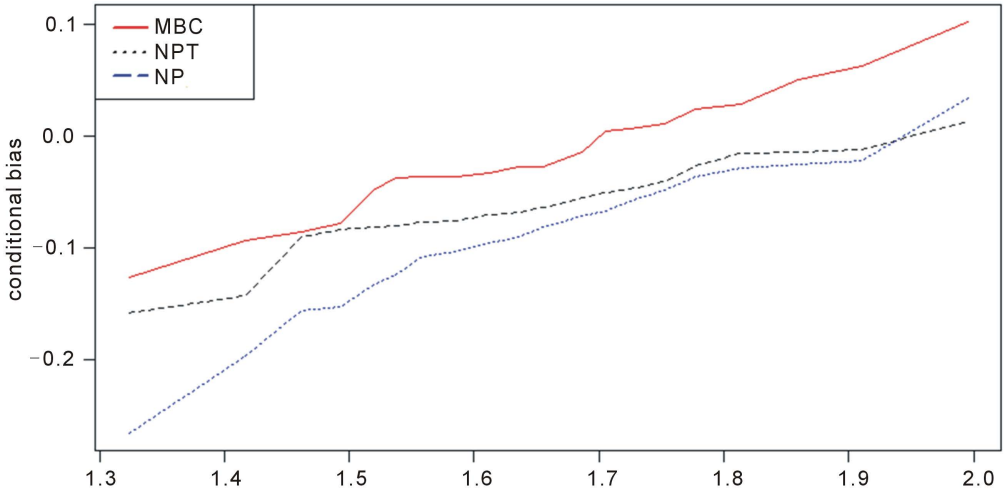
| Estimators | Bandwidth | Bias/T | 10 rmse/T |
|---|---|---|---|
| **NP(one sample)** | **ndr** | 0.25 | 19.63 |
| | **ndr0** | 0.26 | 20.14 |
| | **bcv** | 0.11 | 20.71 |
| | **ucv** | 0.37 | 17.10 |
| **NP(s1Us2)** | **ndr** | 0.01 | 10.5 |
| | **ndr0** | 0.01 | 10.49 |
| | **bcv** | 0.45 | 11.19 |
| | **ucv** | 0.05 | 8.22 |
| **NPT** | **ndr** | 0.05 | 9.93 |
| | **ndr0** | 0.24 | 10.32 |
| | **bcv** | 0.39 | 10.83 |
| | **ucv** | 0.09 | 8.54 |
| **MBC** | **ndr** | 0.20 | 10.23 |
| | **ndr0** | 0.02 | 9.97 |
| | **bcv** | 0.23 | 10.17 |
| | **ucv** | 0.01 | 8.20 |

Figure 2 and Figure 3 indicate the conditional bias and conditional root mean square respectively, with each of the plot drawn at different bandwidth. The population mean of auxiliary variable $x$ was found to be 1.701. Under the conditional bias plots, it is observed that, the proposed estimator outperforms the two currently used estimatorsin terms of conditional biases especially with the unbiased cross-validation and the biased cross-validation method of

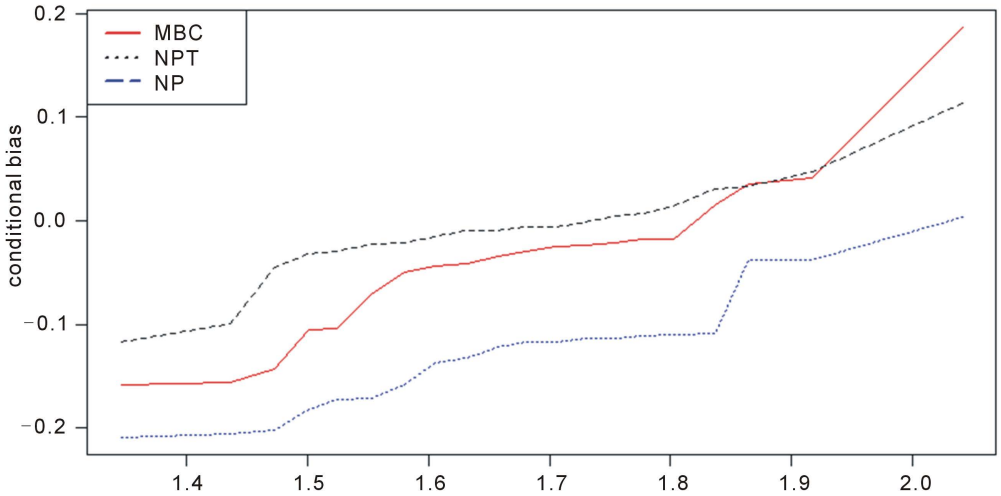
selecting bandwidth. This trend persist in the case of conditional root mean square error.



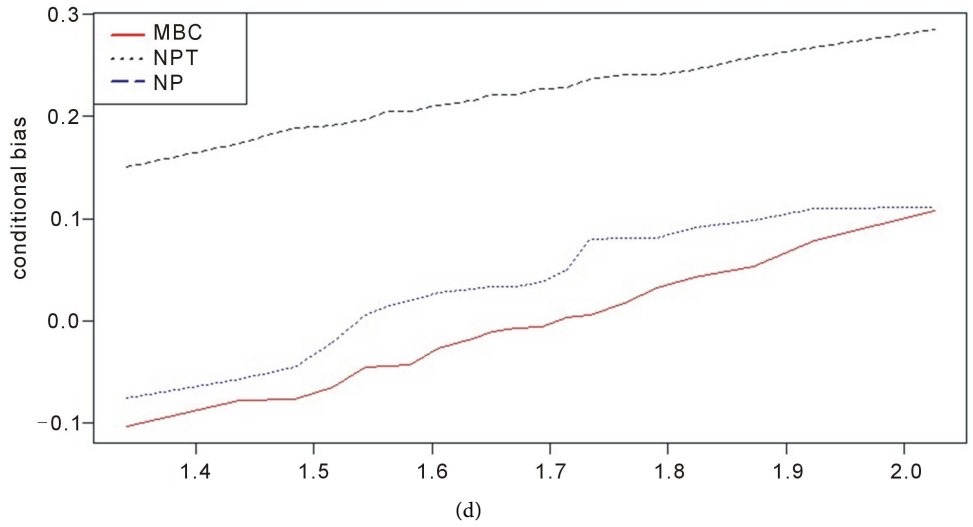

(a)




(b)

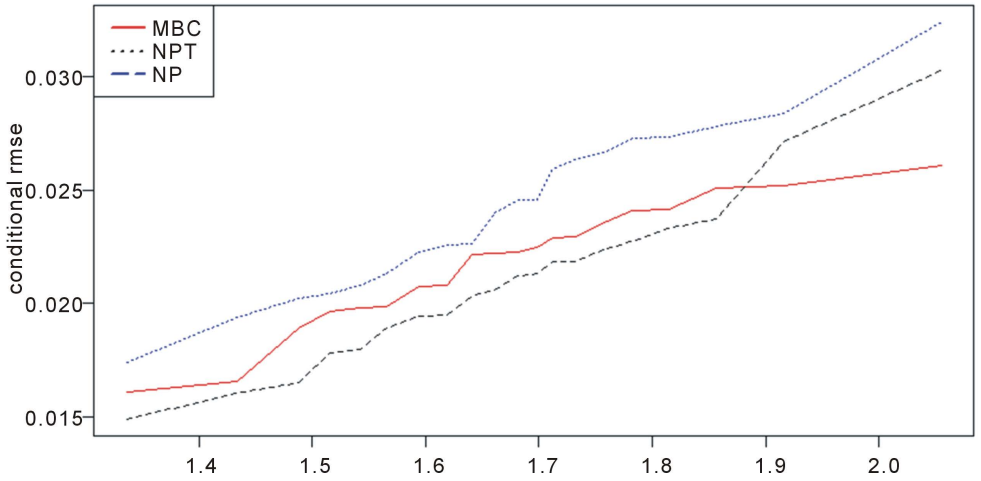



(c)

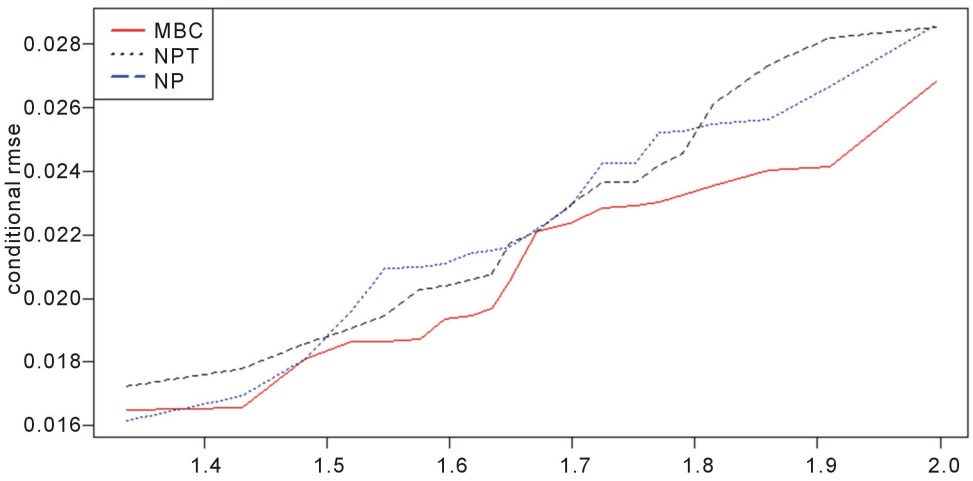



(d)

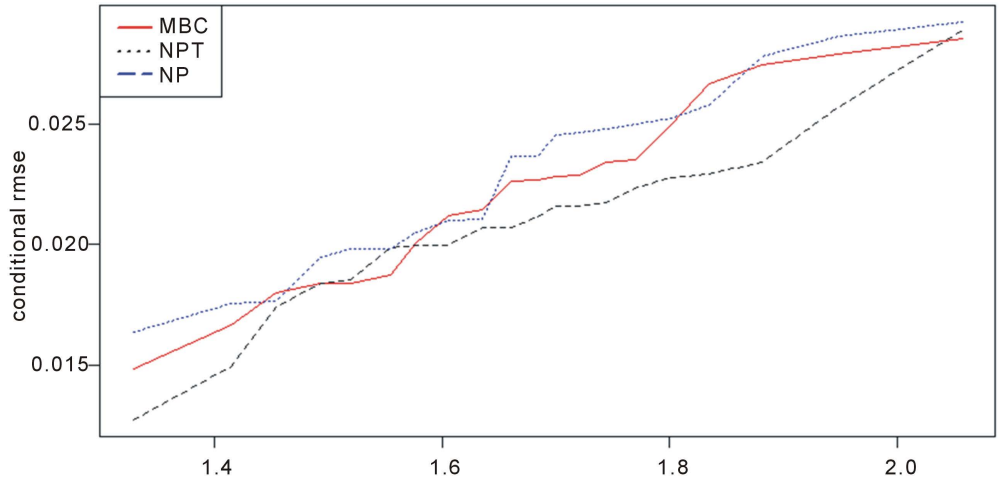
**Figure 2.** Plots indicating the conditional biases of three estimators. (a) Biased cross-validation (bcv); (b) Rule-of-thumb for choosing the bandwidth of a Gaussian kernel density estimator (ndr0); (c) Common common variation given by Scott (1992); (d) Unbiased cross-validation (ucv).



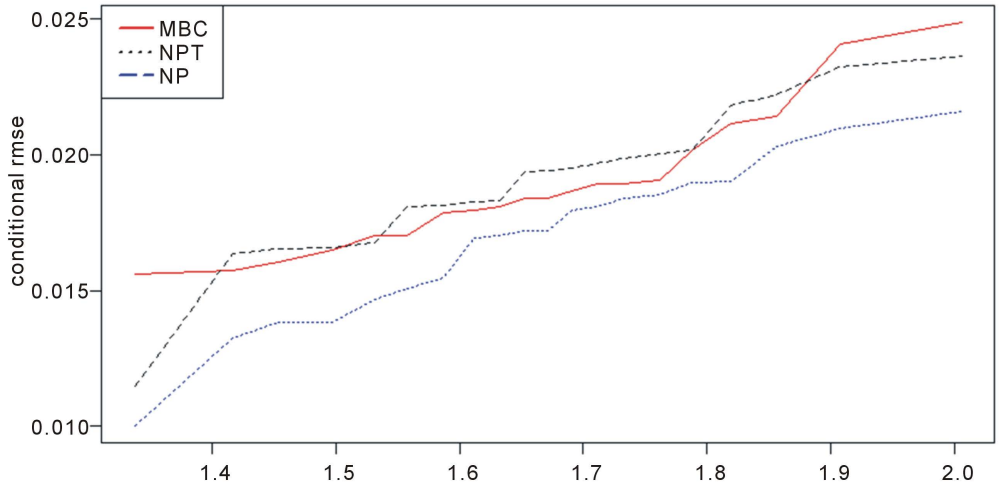

(a)




(b)




(c)




(d)

**Figure 3.** Plots indicating the conditional root mean square error of three estimators. (a) Biased cross-validation (bcv); (b) rule-of-thumb for choosing the bandwidth of a Gaussian kernel density estimator (ndr0); (c) Common variation given by Scott (1992); (d) unbiased cross-validation (ucv).

## 5. Conclusion

The aim of this study was to develop an estimator with the lowest bias for the finite population total using the multiplicative bias corrected approach to non parametric regression. This study reveals that the proposed estimator is more efficient than the modified nonparametric estimator (NPT). With a suitable bandwidth selection (ucv), the proposed estimator has the smallest bias and root mean square error values. It has therefore proven to be efficient in resolving the boundary value problem that is associated with the existing nonparametric smoothers.

## Acknowledgements

My first appreciation goes to my supervisors Professor Odhiambo and Doctor

Mageto for accompanying me through this work. Also, alot of thanks to the African Union for providing for this scientific reseach and placing such confident in its youth. Lastly but not the least, thanks to my family for their support.

## References

[1] Merkouris, T. (2004) Combining Independent Regression Estimators from Multiple Surveys. *Journal of the American Statistical Association*, **99**, 1131-1139. https://doi.org/10.1198/016214504000000601

[2] Wu, C.B. (2004) Combining Information from Multiple Surveys through the Emperical Likelihood Method. *The Canadian Journal of Statistics*, **32**, 15-26. https://doi.org/10.2307/3315996

[3] Dorfman, A.H. (2008) The Two Sample Problem. *Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods*, Denver, 3-7 August 2008.

[4] Dorfman, A.H. (2009) Nonparametric Regression and the Two Sample Problem. *Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods*, Washington DC, August 1-6 2009, 277-270.

[5] Marron, J.S. and Härdle, W. (1986) Random Approximations to Some Measures of Accuracy in Nonparametric Curve Estimation. *Journal of Multivariate Analysis*, **20**, 91-113. https://doi.org/10.1016/0047-259X(86)90021-7

[6] Bierens, H.J. (1987) Kernel Estimators of Regression Functions. *Advances in Econometrics: Fifth World Congress*, Cambridge University Press, Cambridge, 99-144. https://doi.org/10.1017/CCOL0521344301.003

[7] Muller, H.-G. and Stadtmuller, U. (1987) Variable Bandwidth Kernel Estimators of Regression Curves. *The Annals of Statistics*, **15**, 182-201. https://doi.org/10.1214/aos/1176350260

[8] Linton, O. and Nielsen, J.P. (1994) A Multiplicative Bias Reduction Method for Nonparametric Regression. *Statistics & Probability Letters*, **19**, 181-187. https://doi.org/10.1016/0167-7152(94)90102-3

[9] Fan, J.Q. (1992) Design-Adaptive Nonparametric Regression. *Journal of the American Statistical Association*, **87**, 998-1004. https://doi.org/10.1080/01621459.1992.10476255

[10] Hirukawaa, M. and Sakudo, M. (2014) Nonnegative Bias Reduction Methods for Density Estimation Using Asymmetric Kernels. *Computational Statisticsand Data Analysis*, **92**, 112-123. https://doi.org/10.1016/j.csda.2014.01.012

[11] Hengartner, N. and Matzner-Løber, E., Rouviere, L. and Burr, T, (2009) Multiplicative Bias Corrected Nonparametric Smoothers. arXiv Preprint arXiv:0908.0128.

[12] Dorfman, A.H. (1992) Nonparametric Regression for Estimating Totals in Finite Populations. *Proceedings of the Section on Survey Research Methods*, American Statistical Association Alexandria, Washington DC, 622-625.