American Journal of Theoretical and Applied Statistics
2016; 5(5): 317-325
http://www.sciencepublishinggroup.com/j/ajtas
doi: 10.11648/j.ajtas.20160505.21
ISSN: 2326-8999 (Print); ISSN: 2326-9006 (Online)

SciencePG
Science Publishing Group

A Multiplicative Bias Corrected Nonparametric Estimator for a Finite Population Mean

Bonface Miya Malenje[*], **Winnie Onsongo Mokeira, Romanus Odhiambo, George Otieno Orwa**

Department of Statistics and Actuarial Sciences, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya

Email address:

bonniemiya@gmail.com (B. M. Malenje)

To cite this article:

Bonface Miya Malenje, Winnie Onsongo Mokeira, Romanus Odhiambo, George Otieno Orwa. A Multiplicative Bias Corrected Nonparametric Estimator for a Finite Population Mean. *American Journal of Theoretical and Applied Statistics*. Vol. 5, No. 5, 2015, pp. 317-325.
doi: 10.11648/j.ajtas.20160505.21

Received: September 28, 2015; **Accepted**: October 20, 2015; **Published**: September 28, 2016

Abstract: Nonparametric regression has been widely exploited in survey sampling to construct estimators for the finite population mean and total. It offers greater flexibility with regard to model specification and is therefore applicable to a wide range of problems. A major drawback of estimators constructed under this framework is that they are generally biased due to the boundary problem and therefore require modification at the boundary points. In this study, a bias robust estimator for the finite population mean based on the multiplicative bias reduction technique is proposed. A simulation study is performed to develop the properties of this estimator as well as assess its performance relative to other existing estimators. The asymptotic properties and coverage rates of our proposed estimator are better than those exhibited by the Nadaraya Watson estimator and the ratio estimator.

Keywords: Multiplicative Bias, Nonparametric Model, Finite Population Mean, Conditional Bias

1. Introduction

Sample surveys are intended to reduce the time and cost of collecting data while at the same time ensuring valid inference about population quantities. Extrapolation does not give accurate information in surveys since the sample is a subset of an entire population and therefore does not contain information on units that are not represented in the selected sample. The use of auxiliary information that is correlated to the characteristic under study has been very effective in predicting the information in the unobserved units.

Under the model based framework, a super-population model that describes the relationship between the auxiliary variable and the study variable is used to predict the non-sampled values. This has an overall effect of increasing the precision with which population quantities are estimated. Ratio and regression estimators are examples of estimators that are constructed under this framework.

One of the major challenges in using this approach lies in the selection of an optimal model. This presents a danger of model misspecification which if committed, introduces a huge amount of error in the estimates of the population parameters. A number of strategies have been proposed to solve the problems arising from model misspecification.

Nonparametric regression has been embraced as one of the ways of dealing with the problem of model misspecification. In this case, no restrictions are placed on the relationship between the auxiliary variable and the study variable of interest. This has an overall effect of improving the performance of the estimators.

A major problem that is encountered when using nonparametric kernel based regression over a finite interval such as in the estimation of finite population quantities is the bias at the boundary points. A number of techniques have been proposed in this regard and many of them have encountered various pitfalls. Our focus is to apply a multiplicative bias correction technique to the nonparametric estimation of the finite population mean and to study the asymptotic properties, coverage properties and the conditional properties of the resulting estimate.

1.1. Outline of the Paper

The rest of the paper is organized as follows. In subsections 1.2, 1.3, 1.4, we briefly highlight on model based estimation, bias-variance tradeoff and confidence intervals. A multiplicative bias corrected estimator for the finite population mean is proposed in section 2. The asymptotic properties of the proposed estimator are derived in section 3. An empirical study is given in section 4 and the conclusion of the paper is given in section 5.

1.2. Review of the Model-Based Approach to Survey Inference

The model based approach was originally proposed by Ronald. A. Fisher and comprehensively reviewed by, among others, Royall (1976, 1992), Royall and Cumberland (1981), and Chambers (1996). In this framework the survey measurements are assumed to be realized values of some random variables. It is also assumed that an auxiliary variable correlated to the variable under study is available for all units in the population. A model that describes the relationship between the study variable (survey measurements) and the auxiliary variable is then sought. The model and the sampled data are then used to predict the non-sampled values and hence finite population mean or total.

One of its main weaknesses, and a major cause of criticism, is that it is susceptible to bias arising from model misspecification. In fact, when model assumptions are seriously violated this approach can yield estimates that are even worse off than those constructed under the designed-based framework. Consequently, the focus of most research in prediction approach has been to develop strategies to counter the effects of model misspecification on inference.

More specifically, our focus is to advance the work of Dorfman (1992) who considered a similar problem of estimating the finite population total using nonparametric regression. In his work, he used the Nadaraya-Watson estimator of the mean function to predict the nonsampled values of the study variable and consequently to estimate the finite population total. In his findings he demonstrated that the developed estimator was more efficient compared to rival design based estimators.

1.3. Trade-Off Between Bias and Variance

In kernel smoothing there exists a fundamental trade-off between the bias and the variance of the estimate which is governed by the smoothing parameter. Choosing a large bandwidth reduces the variance but simultaneously increases the bias of the estimate.

Similarly a choice of a small bandwidth mitigates the bias but leads to an increase in the variance of the estimate. A natural way to mitigate this trade-off is to choose a bandwidth that minimizes the mean squared error of the estimate.

1.4. Review of Confidence Intervals in Survey Sampling

Sample based surveys contain a level of uncertainty to the results obtained due to the fact that they are based on a portion of the population (sample) and not the entire population. Confidence intervals are one of the statistical tests that are used to measure the likelihood of getting similar findings if the entire population is used. In other words, it measures the 'confidence' in findings from a sample survey.

Constructing confidence intervals around point estimators provide survey statisticians with a properly scaled measure of the uncertainty associated with a particular estimator of interest. A major drawback of the conventional method is that it relies on the central limit theorem which only holds for sufficiently large sample sizes. A challenge arises when modest sample sizes are encountered in practice.

As a result previous research has been concerned with the provision of alternative approaches that address the limitation of the conventional method of constructing confidence intervals. One such strategy is the bootstrap method described in Efron (1982) that has seen considerable development over the past years. (Rao & Wu, 1988) explore the application of this technique under the design based framework. Their findings are then extended to more complex survey designs by Sitter (1992a, 1992b).

R. Chambers & Dorfman (2003) describe an application of the bootstrap approach in the construction of confidence intervals under the model based approach to sample survey inference. In their work they focus on the ratio estimator as the estimator of interest. However, their empirical results obtained by using the beef population indicate that their objective of constructing sound confidence intervals for the finite population total was not attained.

Ouma & Wafula (2007) suggest the use of a general super-population model. Their methodology is simple to implement. In their study, they generated the values of the survey measurement, Y via simple random sampling with replacement. The results of their empirical study showed that their coverage rates were more satisfactory than those of R. Chambers & Dorfman (2003). Their findings are then extended to two stage cluster sampling by Onyango, Otieno, & Orwa (2010).

2. Proposed Estimator

In this section, we present the proposed procedure for estimating the finite population mean. We consider a finite population $U = \{1, 2, 3....N\}$ in which each of the sampling units is associated with a variable of interest Y. Further assume that an auxiliary variable X is available for all elements in the finite population. We describe the population units using the prediction model ξ ;

$$Y_i = \mu(x_i) + \varepsilon_i \tag{1}$$

Such that, $E(Y) = \mu(x_i)$

$$Cov(Y_i, Y_j) = \begin{cases} \sigma^2(x_i),\, for\, i = j \\ 0\, ,otherwise \end{cases} \tag{2}$$

Where $\mu(x_i)$ and $\sigma^2(x_i)$ are assumed to be smooth functions of the variable X. After obtaining sample information on the study variable Y and a census on the auxiliary variable X, the unknown population mean that is to be estimated can be written as:

$$\bar{Y} = f\,\bar{y}_s + (1-f)\,\bar{y}_{p-s} \tag{3}$$

where $\bar{y}_s = \frac{1}{n}\sum_{i\varepsilon s} y_i$, $\bar{y}_{p-s} = \frac{1}{N-n}\sum_{j\varepsilon p-s} y_j$ and consequently f is the sampling fraction. In this case i refers to the sample units and j refers to the non-sampled units. Since the sample mean is known, the process of estimating the unknown population mean $\bar{Y}$ is equivalent to predicting the unsampled part of the population. The population mean can therefore be estimated as,

$$\hat{\bar{Y}} = f\,\bar{y}_s + (1-f)\frac{1}{N-n}\sum_{j\varepsilon p-s}\hat{y}_j \tag{4}$$

To the problem of estimating the unsampled part of the population $\sum_{j\varepsilon p-s}\hat{y}_j$, we propose the estimator $\sum_{j\varepsilon p-s}\mu(x_j)$ where $\mu(x_j)$ is a smooth function. Therefore the estimator (4) becomes

$$\hat{\bar{Y}}_{MBC} = f\,\bar{y}_s + (1-f)\frac{1}{N-n}\sum_{j\varepsilon p-s}\mu(x_j) \tag{5}$$

The task is to estimate the second part of equation (5). To do this, the multiplicative bias correction technique is employed in which case the proposed estimator of the population mean is now defined as

$$\hat{\bar{Y}}_{MBC} = f\,\bar{y}_s + (1-f)\frac{1}{N-n}\sum_{j\varepsilon p-s}\hat{\mu}_n(x_j) \tag{6}$$

Where $\hat{\mu}_n(x_j)$ is as defined in equation (9). We define a pilot smoother of the regression function as

$$\tilde{\mu}_n(x) = \omega_j(x)y_j \tag{7}$$

Then the ratio $R_j = \frac{Y_j}{\tilde{\mu}(X_j)}$ is a noisy estimate of the inverse relative estimation error of the smoother $\tilde{\mu}_n$ at each of the observations given by $\frac{\mu(X_j)}{\tilde{\mu}_n(X_j)}$. Smoothing R_j yields

$$\hat{\alpha}_n(x) = \sum_{j=1}^{n}\omega_j(x;h)R_j = \sum_{j=1}^{n}\omega_j(x;h)\frac{\mu(X_j)}{\tilde{\mu}_n(X_j)} \tag{8}$$

Equation (8) 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 in equation (7). This yields the smoother;

$$\hat{\mu}_n(x_i) = \hat{\alpha}_n(x)\tilde{\mu}_n(x) \tag{9}$$

Using equation (8) and (9) easily yields

$$\hat{\mu}_n(x) = \sum_{j=1}^{n}\omega_j(x;h)\frac{\tilde{\mu}(x)}{\tilde{\mu}_n(X_j)}Y_j \tag{10}$$

The ratio $\frac{\tilde{\mu}(x)}{\tilde{\mu}_n(X_j)}$ in equation (10) can be expressed as

$$\frac{\tilde{\mu}_n(x)}{\tilde{\mu}_n(X_j)} = \frac{\bar{\mu}_n(x)}{\bar{\mu}_n(X_j)} \times \frac{\tilde{\mu}_n(x)}{\bar{\mu}_n(x)} \times \left(\frac{\tilde{\mu}_n(X_j)}{\bar{\mu}_n(X_j)}\right)^{-1} \tag{11}$$

For simplicity, we let $V_n(x) = \frac{\tilde{\mu}_n(x)-\bar{\mu}_n(x)}{\bar{\mu}_n(x)}$ and $V_n(X_j) = \frac{\tilde{\mu}_n(X_j)-\bar{\mu}_n(X_j)}{\bar{\mu}_n(X_j)}$. Equation (11) becomes,

$$\frac{\tilde{\mu}_n(x)}{\tilde{\mu}_n(X_j)} = \frac{\bar{\mu}_n(x)}{\bar{\mu}_n(X_j)} \times \left(1+V_n(x)-V_n(X_j)+r_j(x,X_j)\right) \tag{12}$$

Where $r_j(x,X_j)$ is the remainder term that involves the terms x and X_j. Using equation (12) and utilizing the model $Y_j = \mu(x_j)+\varepsilon_j$ in equation (10) we obtain

$$\hat{\mu}_n(x) = \sum_{j=1}^{n}\omega_j(x;h)\left\{\frac{\bar{\mu}_n(x)}{\bar{\mu}_n(X_j)} \times \left[1+V_n(x)-V_n(X_j)+r_j(x,X_j)\right]\left[\mu(X_j)+\varepsilon_j\right]\right\} \tag{13}$$

Using the assumption $nh \to \infty$ and $n \to \infty$ the remainder terms converge to zero in probability and equation (13)reduces to

$$\begin{aligned} &= \sum_{j=1}^{n}\omega_j(x;h)\frac{\bar{\mu}_n(x)}{\bar{\mu}_n(X_j)}\mu(X_j) + \sum_{j=1}^{n}\omega_j(x;h)\frac{\bar{\mu}_n(x)}{\bar{\mu}_n(X_j)}\left[\varepsilon_j + \mu(X_j)(V_n(x)-V_n(X_j))\right] \\ &+ \sum_{j=1}^{n}\omega_j(x;h)\frac{\bar{\mu}_n(x)}{\bar{\mu}_n(X_j)}\varepsilon_j\left[V_n(x)-V_n(X_j)\right] + O_p\left(\frac{1}{nh}\right) \end{aligned} \tag{14}$$

The proposed estimator for the finite population mean can then be expressed as

$$\bar{Y}_{MBC} = f\bar{y}_s + (1-f)\frac{1}{N-n}\sum_{j\in p-s}\{\ \sum_{j=1}^{n}\omega_j(x;h)\frac{\bar{\mu}_n(x)}{\bar{\mu}_n(X_j)}\mu(X_j)$$

$$+\sum_{j=1}^{n}\omega_j(x;h)\frac{\bar{\mu}_n(x)}{\bar{\mu}_n(X_j)}\left[\varepsilon_j+\mu(X_j)(V_n(x)-V_n(X_j))\right]+\sum_{j=1}^{n}\omega_j(x;h)\frac{\bar{\mu}_n(x)}{\bar{\mu}_n(X_j)}\varepsilon_j\left[V_n(x)-V_n(X_j)\right]+O_p\left(\frac{1}{nh}\right)\ \} \tag{15}$$

3. Properties of the Proposed Estimator

3.1. The Asymptotic Bias of the Proposed Estimator

Under the model based framework, the bias of the estimator $\bar{Y}_{MBC}$ is defined as;

$$E_\xi\left[\bar{Y}_{MBC}-\bar{Y}\right]=E_\xi\left[\bar{Y}_{MBC}\right]-E_\xi\left[\bar{Y}\right] \tag{16}$$

Next, the expected value of the proposed estimator for the finite population mean is given by

$$E_\xi\left[\bar{Y}_{MBC}\right]=E_\xi\left[f\ \bar{y}_s+(1-f)\frac{1}{N-n}\sum_{p-s}\hat{\mu}(x_j)\right]=f\ E_\xi(\bar{y}_s)+(1-f)\frac{1}{N-n}\sum_{p-s}E_\xi\ [\hat{\mu}(x_j)] \tag{17}$$

The calculation of $E_\xi\,[\hat{\mu}(x_j)]$ is obtained by analyzing the individual terms of the stochastic approximation of the estimator in equation (9) which are given by equation (14).

Analyzing the first term of the expression in equation (14) yields

$$E_\xi\left[\sum_{j=1}^{n}\omega_j(x;h)\frac{\bar{\mu}_n(x)}{\bar{\mu}_n(X_j)}\mu(X_j)\right]=\sum_{j=1}^{n}\omega_j(x;h)\frac{\bar{\mu}_n(x)}{\bar{\mu}_n(X_j)}E_\xi(\mu(X_j))=\sum_{j=1}^{n}\omega_j(x;h)\frac{\bar{\mu}_n(x)}{\bar{\mu}_n(X_j)}\mu(X_j) \tag{18}$$

An analysis of the second term of equation (14) gives

$$E_\xi\left[\sum_{j=1}^{n}\omega_j(x;h)\frac{\bar{\mu}_n(x)}{\bar{\mu}_n(X_j)}\left[\varepsilon_j+\mu(X_j)(V_n(x)-V_n(X_j))\right]\right] \tag{19}$$

$$=E_\xi\left[\sum_{j=1}^{n}\omega_j(x;h)\left\{\frac{\bar{\mu}_n(x)}{\bar{\mu}_n(X_j)}\varepsilon_j+\frac{\bar{\mu}_n(x)}{\bar{\mu}_n(X_j)}\mu(X_j)\left(\frac{\tilde{\mu}_n(x)-\bar{\mu}_n(x)}{\bar{\mu}_n(x)}\right)-\frac{\bar{\mu}_n(x)}{\bar{\mu}_n(X_j)}\mu(X_j)\left(\frac{\tilde{\mu}_n(X_j)-\bar{\mu}_n(X_j)}{\bar{\mu}_n(X_j)}\right)\right\}\right] \tag{20}$$

Which further reduces to

$$0+0-\sum_{j=1}^{n}\omega_j(x;h)\frac{\bar{\mu}_n(x)\mu_n(X_j)}{\bar{\mu}_n(X_j)}+\sum_{j=1}^{n}\omega_j(x;h)\frac{\bar{\mu}_n(x)\mu_n(X_j)}{\bar{\mu}_n(X_j)}=0 \tag{21}$$

Lastly analyzing the third term of (14) gives the following results

$$=E_\xi\left[\sum_{j=1}^{n}\omega_j(x;h)\frac{\bar{\mu}_n(x)}{\bar{\mu}_n(X_j)}\left(V_n(x)-V_n(X_j)\right)\right] \tag{22}$$

$$=E_\xi\left[\sum_{j=1}^{n}\omega_j(x;h)\left\{\frac{\bar{\mu}_n(x)}{\bar{\mu}_n(X_j)}\varepsilon_j\left(\frac{\tilde{\mu}_n(x)-\bar{\mu}_n(x)}{\bar{\mu}_n(x)}\right)-\frac{\bar{\mu}_n(x)}{\bar{\mu}_n(X_j)}\varepsilon_j\left(\frac{\tilde{\mu}_n(X_j)-\bar{\mu}_n(X_j)}{\bar{\mu}_n(X_j)}\right)\right\}\right] \tag{23}$$

$$=\sum_{j=1}^{n}\omega_j(x;h)\left\{\frac{\bar{\mu}_n(x)}{\bar{\mu}_n(X_j)}\left(\frac{\tilde{\mu}_n(x)-\bar{\mu}_n(x)}{\bar{\mu}_n(x)}\right)E_\xi\{\varepsilon_j\}-\frac{\bar{\mu}_n(x)}{\bar{\mu}_n(X_j)}\left(\frac{\tilde{\mu}_n(X_j)-\bar{\mu}_n(X_j)}{\bar{\mu}_n(X_j)}\right)E_\xi\{\varepsilon_j\}\right\}=0 \tag{24}$$

Consequently, putting the results obtained in the analysis of the above terms equation (14) reduces to,

$$E_{\xi}\left[\hat{\mu}_n(x)\right]=\sum_{j=1}^{n}\omega_j(x;h)\frac{\bar{\mu}_n(x)}{\bar{\mu}_n(X_j)}\mu(X_j)+O_P\left(\frac{1}{nh}\right) \tag{25}$$

And the expected value of the proposed estimator takes the form

$$E_{\xi}\left[\bar{Y}_{MBC}\right]=f\,E_{\xi}(\bar{y}_s)+(1-f)\frac{1}{N-n}\sum_{p-s}\left(\sum_{j=1}^{n}\omega_j(x;h)\frac{\bar{\mu}_n(x)}{\bar{\mu}_n(X_j)}\mu(X_j)\right)+O_P\left(\frac{1}{nh}\right) \tag{26}$$

A simplified version of the above expression is obtained by considering a limited version of the Taylor series expansion of the ratio $\frac{\bar{\mu}_n(x)}{\bar{\mu}_n(X_j)}$ about a point x which gives the following expression.

$$\frac{\mu(X_j)}{\mu_n(X_j)}=\frac{\mu_n(x)}{\bar{\mu}_n(x)}+(X_j-x)\left(\frac{\mu_n(x)}{\bar{\mu}_n(x)}\right)'+\frac{1}{2}(X_j-x)^2\left(\frac{\mu_n(x)}{\bar{\mu}_n(x)}\right)''(1+O_p(1)) \tag{27}$$

Substituting the first two terms of the expansion given by equation (26) we obtain

$$E_{\xi}\left[\bar{Y}_{MBC}\right]=f\bar{y}_s+(1-f)\frac{1}{N-n}\sum_{p-s}\left\{\sum_{j=1}^{n}\omega_j(x;h)\bar{\mu}_n(x)\left(\frac{\mu_n(x)}{\bar{\mu}_n(x)}+(X_j-x)\left(\frac{\mu_n(x)}{\bar{\mu}_n(x)}\right)'\right)\right\}+O_P\left(\frac{1}{nh}\right) \tag{28}$$

It is easy to verify that $\sum_{j=1}^{n}\omega_j(x;h)=1$ and $\sum_{j=1}^{n}(X_j-x)\omega_j(x;h)=0$

We therefore obtain the results

$$E_{\xi}\left[\bar{Y}_{MBC}\right]=fE(\bar{y}_s)+(1-f)\frac{1}{N-n}\sum_{p-s}\omega_j(x;h)\mu(x)+O_P\left(\frac{1}{nh}\right) \tag{29}$$

Using model (1) it can also be shown that

$$E_{\xi}\left[\bar{Y}\right]=f\,E_{\xi}(\bar{y}_s)+(1-f)\frac{1}{N-n}\sum_{p-s}\mu(x) \tag{30}$$

Hence the asymptotic bias of the proposed estimator is given by

$$BIAS\left[\bar{Y}_{MBC}\right]=\frac{1}{N}\left(\sum_{j\in p-s}\left\{\sum_{p-s}\omega_j(x;h)\mu(x)\right\}-\sum_{j\in p-s}\mu(x)\right)+O_P\left(\frac{1}{nh}\right) \tag{31}$$

The bias of $\bar{Y}_{MBC}$ 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 $O_P\left(h^2\right)$.

3.2. The Asymptotic Variance of the Proposed Estimator

We can express the estimator of the finite population mean as

$$\bar{Y}_{MBC}=f\,\bar{y}_s+(1-f)\frac{1}{N-n}\sum_{j\in p-s}\left\{\sum_{j=1}^{n}\omega_j(x;h)\left(\frac{\bar{\mu}_n(x)}{\bar{\mu}_n(X_j)}[\mu(X)+\varepsilon_j]\left[1+V_n(x)-V_n(X_j)+r_j(x,X_j)\right]\right)\right\} \tag{32}$$

Using the assumption $nh\to\infty$ the remainder terms $r_j(x,X_j)$ converge to zero in Probability and the above expression reduces to,

$$\bar{Y}_{MBC} = f\bar{y}_s + (1-f)\frac{1}{N-n}\sum_{j\in p-s}\left\{\sum_{j=1}^{n}\omega_j(x;h)\left(\frac{\bar{\mu}_n(x)}{\bar{\mu}_n(X_j)}[\mu(X)+\varepsilon_j]\left[1+V_n(x)-V_n(X_j)\right]\right)\right\} + O_P\left(\frac{1}{nh}\right) \tag{33}$$

Truncating the binomial expansion at the first term yields

$$\bar{Y}_{MBC} = f\,\bar{y}_s + (1-f)\frac{1}{N-n}\sum_{j\in p-s}\left\{\sum_{j=1}^{n}\omega_j(x;h)\left(\frac{\bar{\mu}_n(x)}{\bar{\mu}_n(X_j)}[\mu(X)+\varepsilon_j]\right) + O_P\left(\frac{1}{nh}\right)\right\} \tag{34}$$

The variance of the estimator is then defined by

$$Var\left[\bar{Y}_{MBC}\right] = \frac{1}{N^2}\left[\sum_{i\in s}Var(Y_i) + \sum_{j\in p-s}\left\{\sum_{j=1}^{n}\omega_j(x;h)\frac{\bar{\mu}_n(x)}{\bar{\mu}_n(X_j)}Var(Y_j)\right\} + O_P\left(\frac{1}{n^2h^2}\right)\right] \tag{35}$$

$$= \frac{1}{N^2}\left[\sum_{i\in s}\sigma^2(x_i) + \sum_{j\in p-s}\omega_j(x;h)^2\left\{\frac{\bar{\mu}_n(x)}{\bar{\mu}_n(X_j)}\right\}^2\sigma^2(x_i) + O_P\left(\frac{1}{n^2h^2}\right)\right] \tag{36}$$

Using the Taylor series expansion of the term $\frac{\sigma^2(x_i)}{\{\bar{\mu}_n(X_j)\}^2}$ the variance expression above can be written as

$$Var\left[\bar{Y}_{MBC}\right] = \frac{1}{N^2}\sum_{i\in s}\sigma^2(x_i) + \frac{1}{N^2}\left[\sum_{j\in p-s}\sum_{j=1}^{n}\omega_j(x;h)^2\sigma^2(x_i)\right] + O_P\left(\frac{1}{n^2h^2}\right) \tag{37}$$

This implies that is more efficient than the usual non-parametric regression estimator
Proposed by Dorfman (1992)

3.3. The Asymptotic Mean Squared Error of the Proposed Estimator

The Mean squared error of $\bar{Y}_{MBC}$ is given by

$$MSE\left[\bar{Y}_{MBC}\right] = Var\left[\bar{Y}_{MBC}\right] + \left[Bias\left(\bar{Y}_{MBC}\right)\right]^2 \tag{38}$$

Substituting the expressions for the variance and the bias in the above equation yields,

$$= \frac{1}{N^2}\sum_{i\in s}\sigma^2(x_i) + \frac{1}{N^2}\left[\sum_{j\in p-s}\sum_{j=1}^{n}\omega_j(x;h)^2\sigma^2(x_i)\right] + O_P\left(\frac{1}{n^2h^2}\right) + \left[O_P\left(\frac{1}{nh}\right)\right]^2 \tag{39}$$

$$= \frac{1}{N^2}\sum_{i\in s}\sigma^2(x_i) + \frac{1}{N^2}\left[\sum_{j\in p-s}\sum_{i=1}^{n}\omega_j(x;h)^2\sigma^2(x_i)\right] + O_P\left(\frac{1}{n^2h^2}\right) \tag{40}$$

$$\leq \frac{1}{N}\left[\frac{1}{N}\sum_{i\in s}\sigma^2(x_i)\right] + \frac{1}{N}\left[\frac{1}{N-n}\sum_{j\in p-s}\sigma^2(x_i)\right] \leq 0 \tag{41}$$

As $n \to \infty$ and $h \to \infty$ the mean squared error tends to zero indicating that the proposed estimator is statistically consistent.

4. Empirical Study

We perform a simulation experiment in order to investigate the statistical properties of the proposed estimator as well as compare its performance to that of the Nadaraya-Watson and the ratio estimators. We consider a case where only one auxiliary variable is available and generate them as independent and identically distributed on uniform (0, 1) random variables. We examine five simulated populations generated from the following regression model,

$$Y_i = \mu(x_i) + \varepsilon_i \quad 1 \leq i \leq 2000$$

With mean functions

$$\mu_1(x) = 1 + 2(x-0.5) \qquad \textit{(Linear)}$$

$$\mu_2(x) = 1 + 2(x-0.5)^2 \qquad \textit{(Quadratic)}$$

$$\mu_3(x) = 1 + 2(x-0.5)I_{[x\leq 0.65]} + 0.65I_{[x\geq 0.65]} \qquad \textit{(Jump)}$$

$$\mu_4(x) = 2 + 2\sin(2\pi x) \qquad \textit{(Sine)}$$

$$\mu_5 = exp(-8x) \qquad \textit{(Exponential)}$$

$$\boldsymbol{\mu}_6 = 1 + 2(\boldsymbol{x} - 0.5) + \exp(-200(\boldsymbol{x} - 0.5)^2) \qquad \textit{(Bump)}$$

The linear function $\mu_1(x)$ is correct specification for the ratio estimator and therefore it is expected that the ratio estimator will perform better than the other estimators under this model because it is rightly specified.

The errors are independent and identically distributed with zero means and standard deviation $\sigma = 1$. Five hundred samples of size 500 were generated using simple random sampling without replacement. The sampling is done with indices due to the assumed relationship between the study variable and the auxiliary variable that has to be reflected in the simulation. We compare the performance of the proposed estimator, $\overline{Y}_{MBC}$ with the Nadaraya Watson estimator $\overline{Y}_{NW}$, ratio estimator, $\overline{Y}_{RATIO}$.

The following diagrams represent the plots of the linear, quadratic, jump, exponential and the sine populations.

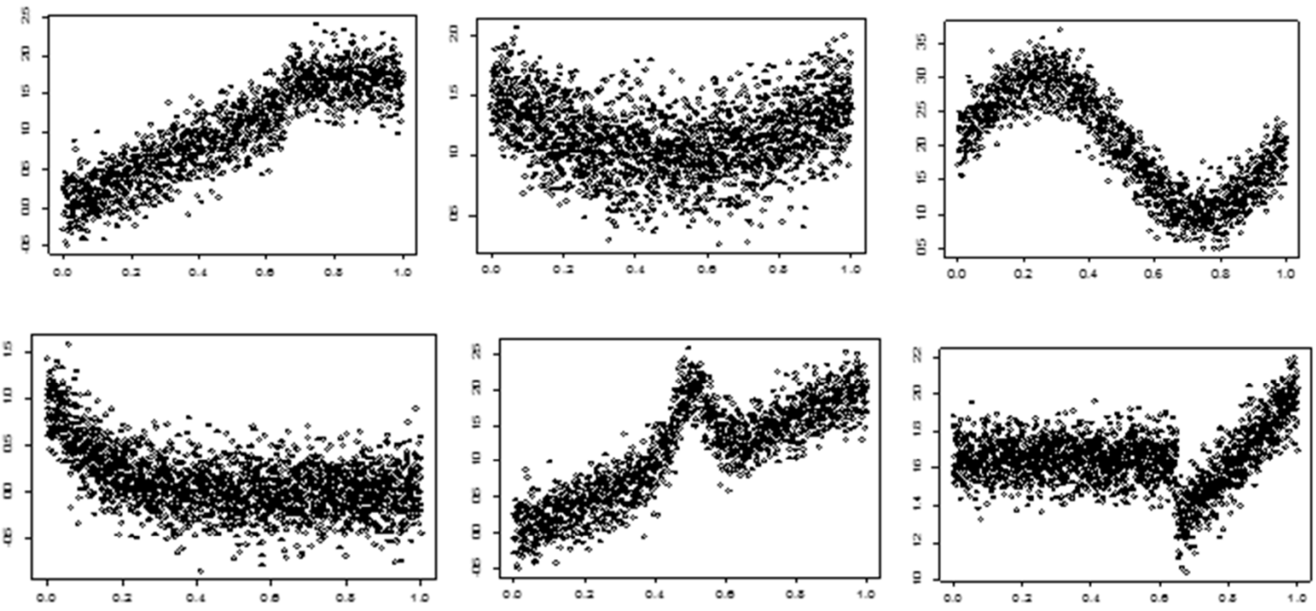

Fig. 1. Plots of the linear, quadratic,sine, exponential, bump and the jump populations.

Table 1. Summary Results of the unconditional Biases and Mean Squared Error.

Mean function	$\overline{Y}_{MBC}$		$\overline{Y}_{NW}$		$\overline{Y}_{RATIO}$	
	BIAS	MSE	BIAS	MSE	BIAS	MSE
Linear	0.000292	1.31010	-0.000845	1.311201	0.000117	0.97453
Quadratic	-0.000169	1.00253	-0.000174	1.00410	0.001431	1.49366
Sine	-0.000243	1.49014	0.001275	1.49177	0.004628	4.18035
Exponential	-0.000770	1.03117	-0.001467	1.03211	0.030857	2.06845
Jump	-0.000733	0.99325	-0.00089	0.99480	0.002222	1.90373
Bump	0.000639	1.36964	0.001424	1.37083	0.052467	3.03956

The unconditional biases are computed as $\frac{1}{500}\sum(\overline{Y}_{MBC} - \overline{Y})$, $\frac{1}{500}\sum(\overline{Y}_{NW} - \overline{Y})$, $\frac{1}{500}\sum(\overline{Y}_{RATIO} - \overline{Y})$. The unconditional Mean Squared Errors are computed for each of the estimators. We also computed the 95% confidence interval lengths for each of the estimators under the different populations.

Table 1 gives the results of the unconditional Biases and the unconditional Mean Squared Error of the multiplicative bias corrected Nadaraya Watson Estimator, $\overline{Y}_{MBC}$ the Nadaraya-Watson estimator $\overline{Y}_{NW}$, and the ratio estimator, $\overline{Y}_{RATIO}$ applied to finite mean estimation for different mean functions. It can be seen that the bias of the multiplicative bias corrected estimator is much lower than those of the Nadaraya-Watson estimator and the corresponding MSE of the multiplicative bias corrected estimator is also lower than that of the Nadaraya Watson estimator for each of the mean functions.

Table 2 gives a comparison of the coverage probabilities of the three estimators for the different mean functions. The coverage probabilities for the Multiplicative bias corrected estimator are closer to the nominal value than are the coverage probabilities for the Nadaraya-Watson estimator .The coverage ability of $\overline{Y}_{MBC}$ is better than that of $\overline{Y}_{NW}$ .The ratio estimator has the best coverage ability under the linear mean function and outperforms the other two estimators.

***Table 2.** Summary results of the unconditional coverage probabilities.*

Mean function	$\bar{Y}_{MBC}$	$\bar{Y}_{NW}$	$\bar{Y}_{RATIO}$
Linear	0.907	0.884	0.937
Quadratic	0.805	0.793	0.034
Sine	0.918	0.896	0.133
Exponential	0.933	0.876	0.234
Jump	0.952	0.915	0.016
Bump	0.948	0.929	0.643

Table 3 gives a comparison of the 95% confidence interval lengths for the Multiplicative Bias corrected estimator, Nadaraya Watson estimator and the ratio estimator for the different mean functions. The confidence intervals generated by the Multiplicative bias corrected estimator are much tighter than those generated by the Nadaraya Watson estimator and the ratio estimator. The results indicate that the multiplicative bias corrected estimator outperforms the usual non-parametric regression estimator proposed by Dorfman (1992) at 95% coverage rate.

***Table 3.** Summary results of the unconditional Bootstrap confidence interval lengths.*

Mean function	$\bar{Y}_{MBC}$	$\bar{Y}_{NW}$	$\bar{Y}_{RATIO}$
Linear	0.1449937	0.198121	0.126388
Quadratic	0.034785	0.063455	0.155434
Sine	0.047185	0.170652	0.257946
Exponential	0.043652	0.077384	0.132733
Jump	0.093663	0.126653	0.174328
Bump	0.040595	0.129329	0.148295

To study the conditional performance of the selected estimators, the 500 samples obtained were sorted by the value of $\bar{x}$ into groups of 20 samples so that we had a total of 25 groups. We then computed the empirical means and bias within each group. The plots of the conditional biases versus $(\bar{x}-\bar{X})$ obtained for the three estimators under the different functions all indicated similar results. We report the behavior of the conditional bias under the linear mean function.

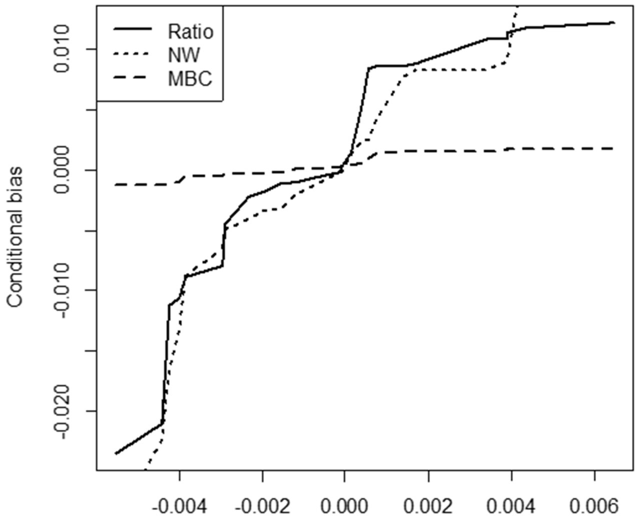

***Fig. 2.** Plots of the conditional biases of $\bar{Y}_{MBC}$, $\bar{Y}_{NW}$ and $\bar{Y}_{RATIO}$ versus $(\bar{x}-\bar{X})$.*

Figure 2 reports the behavior of the conditional bias plotted against the average values of $(\bar{x}-\bar{X})$. The Nadaraya Watson estimator and the ratio estimator are substantially conditionally biased. The Multiplicative bias corrected estimator on the other hand is approximately conditionally unbiased.

5. Conclusions and Recommendations

The aim of this study was to develop a bias robust estimator for the finite population mean using the multiplicative bias correction approach to nonparametric regression. The study reveals that the derived estimator is more efficient than the Nadaraya Watson estimator. The proposed estimator has smaller bias, lower mean squared error, better coverage ability and tighter confidence interval lengths compared with the Nadaraya-Watson estimator. It is also approximately conditionally unbiased. It has therefore proved to be efficient in correcting boundary problems that are associated with the existing nonparametric regression smoothers.

References

[1] A. H. Dorfman. (1992). Nonparametric regression for estimating totals in finite pop- ulations, proceedings of the section on survey research methods. *American Statistical Association, 1992, pp. 622-625.*

[2] Breidt, F. J., & Opsomer, J. D. (2000, 08). Local polynomial regression estimators in survey sampling. Ann. Statist., 28 (4), 1026– 1053.

[3] Chambers, R., & Dorfman, A. (2003). Robust sample survey inference via boot- strapping and bias correction: The case of the ratio estimator.

[4] Chambers, R. L., Dorfman, A. H., & Wehrly, T. E. (1993). Bias robust estimation in finite populations using nonparametric calibration. *Journal of the American Statistical Association*, *88* (421), 268-277. Cochran, W. G. (2007). *Sampling techniques*. John Wiley & Sons.

[5] Efron, B. (1982). the jackknife, the bootstrap and other resampling plans. Society for Industrial and Applied Mathematics.

[6] Fan, J., & Gijbels, I. (1992, 12).Variable bandwidth and local linear regression smoothers *Ann. Statist.*, *20* (4), 2008–2036.

[7] Gasser, T., & Müller, H.-G. (1979). *Kernel estimation of regression functions*.

[8] Godambe, V. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 269–278.

[9] Hengartner, N., Matzner-Lober, E., Rouviere, L., & Burr, T. (2009). Multiplicative bias corrected nonparametric smoothers.

[10] Kwong-Aquino, A. A. H. (2011). Nonparametric model-based predictive estimation in survey sampling. *The Philippine Statistician*.

[11] Linton, O., & Nielsen, J. P. (1994). A multiplicative bias reduction method for nonparametric regression. *Statistics & Probability Letters*, *19* (3), 181–187.

[12] Onyango, C. O., Otieno, R. O., & Orwa, G. O. (2010). Generalised model based confidence intervals in two stage cluster sampling. *Pakistan Journal of Statistics and Operation Research*, *6* (2), 101–115.

[13] Ouma, C., & Wafula, C. (2007). Bootstrap confidence intervals for model-based surveys. *East African Journal of Statistics*, *1* (1), 84–90.

[14] Pensky, M., Vidakovic, B., et al. (1999). Adaptive wavelet estimator for nonparametric density deconvolution. The Annals of Statistics, 27 (6), 2033–2053.

[15] Priestley, M., & Chao, M. (1972). Non-parametric function fitting. *Journal of the Royal Statistical Society. Series B (Methodological)*, 385–392.

[16] Rao, J. N., & Wu, C. (1988). Resampling inference with complex survey data.

[17] Rueda, M., & Sánchez-Borrego, I. (2009). A predictive estimator of finite population mean using nonparametric regression. Computational Statistics, 24 (1), 1–14.

[18] Zheng, H., & Little, R. (2004). Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline non- parametric model. Journal of Official statistics.

[19] Zheng, H., & Little, R. J. (2003). Penalized spline model-based estimation of the finite populations total from probability-proportional-to-size samples. *Journal of offi statistics-Stockholm-*, *19* (2), 99–118.