The Seybold Report

ISSN 1533-9211

DOI 10.17605/OSF.IO/CUYX7

ROBUST ESTIMATOR FOR FINITE POPULATION TOTAL

AJWANG' STELLAMARIS ADHIAMBO[1], ROMANUS ODHAMBO OTIENO[2], THOMAS MAGETO[3] and DAVID ALILA[4]

[1]Department of Mathematics(Statistics option) Programme, Pan African University, Institute for Basic Sciences, Technology and Innovation(PAUSTI), Nairobi, Kenya.
[2]Department of Statistics and Actuarial Sciences, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya.
[3]Department of Statistics and Actuarial Sciences, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya.
[4]Department of Mathematics, Masinde Muliro University of Science and Technology, Kenya.

Abstract

The kernel regression estimator is a flexible and widely used nonparametric estimator that estimates a regression function. Statistical learning appears to be a promising field in which some algorithms resulting from machine learning are interpreted as statistical methods. Boosting is among the most studied machine learning techniques in this paper, a new improvement of the kernel density regression estimator is proposed, with a target of producing smaller estimates of the finite population total. The study was aimed at estimating the finite population total by incorporating the adaptive boosting technique to the nonparametric regression estimator. A numerical study using a simulated population was conducted in order to evaluate the performance of the proposed estimator and compare it with the existing one. The outcome of the proposed estimator is evaluated and presented. The simulation experiment were very promising; it shows that our modified kernel density estimator performs well in all cases, then the normal kernel density estimator.

Keywords: Nonparametric Estimation, Kernel Density Estimator, Bias Reduction, Adaboost, Finite, Population Total

1. INTRODUCTION

In many intricate surveys, the available data regarding the study population can be utilized at both the design and estimation stages to construct accurate methods for finite population quantities, such as total or mean population, to increase the precision of estimators for those population quantities. In a number of statistical problems, nonparametric regression are commonly used to describe the relationship between the response variable and certain covariates (Parzen, 1962).

Nonparametric model allows great flexibility since they do not make any assumptions hence are more powerful in prediction, nonparametric models have higher accuracy than parametric models hence they perform better. Furthermore, they can fit many kinds of functional form hence are flexible in nature. The most used approach is kernel smoothing, which dates back to (Rosenblatt et al., 1956) and (Parzen, 1962).

Many efforts have been devoted to investigating the optimal performance of kernel density estimator for the finite population totals since it has been the most widely used nonparametric method in the last several decades. Boosting has received a lot of consideration from researchers recently. It was first suggested by (Schapire, 1990) and consequently developed by

The Seybold Report
ISSN 1533-9211
DOI 10.17605/OSF.IO/CUYX7

(Freund, 1995), (Freund, Schapire, et al., 1996) and (Schapire & Singer, 1999). Boosting was explored as a way of improving the efficiency of a 'weak learner'. There are several boosting techniques.

In this paper we have considered the adaptive boosting procedure, which was first proposed by Freund et al. (1996), as a way of producing robust estimator for finite population total with an aim of producing a robust finite population total, thereby minimizing the bias of the proposed estimator.

2. KERNEL DENSITY ESTIMATOR (KDE)

The kernel density estimation a non-parametric technique for estimating probability density (pdf). It is non-parametric because it does not assume any underlying distribution for the variable. The idea of nonparametric regression goes back to (Nadaraya, 1964) and (Watson, 1964).

Consider $x_1, x_2, \ldots, x_n$ as an independent and identically distributed (iid) sample of n observations taken from a population P whose probability distribution function f(x) is not known.

The kernel density estimate $\hat{f}(x)$ of f(x) is given by;

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \quad (1)$$

where h > 0 is a smoothing parameter or bandwidth, which controls the degree of smoothness. K(x) is the kernel function, usually symmetric probability density function (pdf) satisfying $\int x\, K(x)dx = 0$ (Wand & Jones (1995). If K(x) is Gaussian distribution, then the $\hat{f}(x)$ estimated will be smooth and have derivatives of all orders.

The kernel has the following properties according to (Silverman, 1986).

$$\int_{-\infty}^{\infty} K(u)du = 1$$

$$\int_{-\infty}^{\infty} uK(u)du = 0$$

$$\int_{-\infty}^{\infty} u^2K(u)du > 0$$

$$K(u) = K(-u)$$

The bias of KDE is given as $u_2(K)h^2\, f''(x)/2 + O(h^2)$ which is of order $O(h^2)$ and the variance Var $(\hat{f}(x) = \frac{R(K)f(x)}{nh} + O\left(\frac{1}{nh}\right)$ which is of order O (1/nh).

2.1 Nonparametric regression estimator for finite population total

The nonparametric regression estimation provides a variety of techniques for estimating$m(x_i) = E[Y_i|X_i = x_i]$. For $x_i = x_j$ for any point in non-sample estimate of$m(x_j)$. The idea is to obtain nonparametric estimates $\hat{m}(x_j)$ for$j \in r$.

ISSN 1533-9211

DOI 10.17605/OSF.IO/CUYX7

For all the (x_j) the estimator of the population total $\hat{T}$ is

$$\hat{T} = \sum_{j\in N} \hat{m}(x_j) \quad (2)$$

This is true since it is expected that $\sum_{j\in N} \hat{m}(x_j) \approx \sum_{j\in N} Y_j$. Therefore, the estimator of the finite population total under nonparametric regression becomes;

$$\begin{aligned} \widehat{T}_{np} &= \sum_{i\in s} Y_i + \sum_r \hat{m}(x_j) \\ &= \sum_{i\in s} Y_i + \sum_r w_{ij} Y_i \\ &= \sum_{i\in s} Y_i + \sum_r w_i Y_i \end{aligned} \quad (3)$$

Where $w_i = \sum_r w_{ij}$

As with model-based estimators (see (Chambers, Dorfman, & Hall, 1992)) generally, this estimator ignores sampling probabilities (It also ignores stratum boundaries). Except for the selection of bandwidth, and possible transformation of the auxiliary, it is an automatic estimator. $\hat{T}_{np}$ Accumulates for non-sample values in the lieu of Y_j .

2.2 Adaptive Boosting Technique

Boosting is the process of combining relatively imprecise prediction models to create a more accurate one. Adaptive boosting procedure is an algorithm proposing the use of machine learning, which was initially suggested by (Freund et al., 1996). Adaptive Boosting method employs an iterative method intended to strengthen weak estimators.

3. BOOSTING THE KERNEL DENSITY ESTIMATOR FOR FINITE POPULATION TOTAL

Consider a finite population of size N of a study variable Y_i and the auxiliary variable, X_i with associated values (x_i, y_i) respectively.

In formulating the model, the dependent variable Y_i and the auxiliary variable X_i will be considered as a nonparametric super-population model (linear), which is of the form

$$Y_i = m(x_i) + e_i, \quad i = 1, 2, \ldots, n \quad (4)$$

where $e_i \sim N(0, \sigma^2)$, and $m(x_i)$ is the unknown mean function. The smoothing function of x_i will be estimated non-parametrically. The expectation and the covariance of equation 4 are presented as follows;

$$\begin{aligned} &E[Y_i | X_i = x_i] = m(x_i) \\ &= \int y\, f(x, y) \partial y \end{aligned} \quad (5)$$

$$cov[Y_i, Y_j | X_i = x_i, X_j = x_j] = \begin{cases} \sigma^2(x_i), & i = j, \quad i, j = 1, 2, \ldots, n \\ 0, & otherwise \end{cases} \quad (6)$$

The Seybold Report
ISSN 1533-9211
DOI 10.17605/OSF.IO/CUYX7

Consider a finite population total T of size N (Dorfman, 1992), given by

$$T = \sum_{i=1}^{N} Y_i = \sum_{i=1}^{n} Y_i + \sum_{i \in r} Y_i \quad (7)$$

where s is the sample and r is the non-sample.

Therefore to estimate the finite population total T we have considered taking the expectation of the non-sampled part.

$$\hat{T} = \sum_{i \in s} y_i + E(\sum_{i \in r} y_i) \quad (8)$$

Resulting to

$$= \sum_{i \in s} y_i + \sum_{i \in r} \widehat{m}(x_j) \quad (9)$$

Since $E[Y_i | X_i] = E(\sum_{i \in r} y_i) = m(x_i)$.

The smoothing function $m(x_i)$ will be estimated non-parametrically using adaptive boosting technique, AdaBoost.

3.1 KDE - AdaBoost for finite population total

Consider the auxiliary variable X_i for $i = 1, 2, \ldots, n$. At the first step initialize the weights, that is, equal weight is assigned.

$$w_1(i) = \frac{1}{n}$$

The smoothing parameter $h_1, h_2, \ldots, h_m$ is then selected. For this study, no procedure was employed while choosing the bandwidth.

The boosting technique involves the re-weighting of data based on the loss function and so in the case of the kernel density estimation, such a measure is computed by comparing the first boosting step with the leave-one-out estimate (Silverman, 1986).

At each step m, the weak estimator is computed as follows;

$$\hat{f}_m(x) = \sum_{i=1}^{n} \frac{w_m(i)}{h} K\left(\frac{x - X_i}{h}\right) \quad (10)$$

where $K()$ is the kernel density function, h is the smoothing parameter and $w_m(i)$ is the weight of observation i at mth step and $\sum w_m(i) = 1$.

The weight of each observation is then updated in each step as;

$$w_{m+1}(i) = w_m(i) + log\left\{\frac{\hat{f}_m(x_i)}{\hat{f}_m^{(i)}(x_i)}\right\} \quad (11)$$

where $\hat{f}_m^{(i)}(x_i)$ is the leave-one-out kernel density estimator ((Silverman, 1986)) given by;

$$\hat{f}_m^{(i)}(x_i) = \frac{1}{(n-1)h} \sum_{i=1}^{n} w_m(i) K\left(\frac{x - X_i}{h}\right) \quad (12)$$

The final output is the product of all the density estimates is given by;

The Seybold Report
ISSN 1533-9211
DOI 10.17605/OSF.IO/CUYX7

$\prod_{m=1}^{M} \hat{f}_m(x_i)$ (13)

The proposed boosted estimator therefore becomes

$\hat{m}_{AB}(x) = \sum Y_i \hat{f}_m(x_i)$ (14)

Incorporating the boosted mean function to the nonparametric finite population total, the proposed regression estimator, under simple random sample without replacement, T_{npAB} is thus given by

$$\hat{T}_{npAB} = \sum_{i\in s} Y_i + \sum_{i\in r} \hat{m}_{AB}(x_i)$$

$$= \sum_{i\in s} Y_i + \sum_{i\in r} Y_i \frac{1}{n} w_m(i) \quad K\left(\frac{x-X_i}{h}\right) \quad (15)$$

As can be seen from the following, it is expected that $\hat{T}_{npAB}$ is less biased than the conventional estimator.

3.2 Asymptotic bias of the proposed estimator, $\hat{T}_{npAB}$

By definition, the bias of the proposed estimator is derived as follows; $Bias\left(\hat{T}_{npAB}\right) = E\left[\hat{T}_{npAB} - T_{np}\right]$

$$= E\left[\left(\sum_{i\in s} Y_i + \sum_{i\in r} \hat{m}_{AB}(x_i)\right) - \left(\sum_{i\in s} Y_i + \sum_{i\in r} Y_j\right)\right]$$

$$= E\left[\sum_{i=n+1}^{N} \hat{m}_{AB}(x_i) - \sum_{i=n+1}^{N} Y_i\right]$$

The bias therefore becomes

$$= E\left[\sum_{i=n+1}^{N} \hat{m}_{AB}(x_i) - \sum_{i=n+1}^{N} m(x)\right] \quad (16)$$

Therefore, the model equation (2) can be re-written as

$Y_i = \hat{m}_{AB}(x_i) + [m(x_i) - m(x) + e_i$ (17)

Equating $\hat{m}_{AB}(x_i)$ into the model in equation (14) we have

$$= \sum_{i=n+1}^{N} \frac{w_m(i)}{n} K\left(\frac{x-X_i}{h}\right) Y_i \quad = \sum_{i=n+1}^{N} \frac{w_m(i)}{n} K\left(\widehat{\frac{x-X_i}{h}}\right) = \hat{m}_{AB}(X_i)$$

$$+ \frac{1}{nh} \sum_{i=n+1}^{N} \frac{w_m(i)}{n} K\left(\frac{x-X_i}{h}\right) [m(x_i) - m(x)]$$

$$+ \frac{1}{nh} \sum_{i=n+1}^{N} \frac{w_m(i)}{n} K\left(\frac{x-X_i}{h}\right) e_i \quad (18)$$

The equation (18) may be rewritten as

$$= \frac{1}{nh} \sum_{i=n+1}^{N} \frac{w_m(i)}{n} K\left(\frac{x-X_i}{h}\right) Y_i \quad = \frac{1}{nh} \sum_{i=n+1}^{N} \frac{w_m(i)}{n} [m(x_i + m_1(x) + m_2(x)] \quad (19)$$

Where

$$m(x) = \sum_{i=n+1}^{N} \frac{w_m(i)}{n} K\left(\widehat{\frac{x-X_i}{h}}\right) = \hat{m}_{AB}(X_i)$$

The Seybold Report
ISSN 1533-9211
DOI 10.17605/OSF.IO/CUYX7

$$m_1(x) = \frac{1}{nh}\sum_{i=n+1}^{N} \frac{w_m(i)}{n} K\left(\frac{x-X_i}{h}\right)[m(x_i) - m(x)]$$

And

$$m_2(x) = \frac{1}{nh}\sum_{i=n+1}^{N} \frac{w_m(i)}{n} K\left(\frac{x-X_i}{h}\right) e_i$$

Hence now taking the expectation of equation (19) we have

$$E[\sum_{i=n+1}^{N} \widehat{m}_{AB}(x_i) = \frac{1}{nh}\sum_{i=n+1}^{N} [m(x_i + m_1(x) + m_2(x)] \quad (20)$$

But $E[e_i|x_i] = 0$. Hence, it then follows that $E[\widehat{m}_2(x)] = 0$

Therefore, getting the expectation of $E[\widehat{m_1}(x)]$, letting $Z = \frac{u-X_i}{h}$ and by Taylor series expansion we obtain

$$E[\widehat{m_1}(x)] = \frac{N-n}{nh}(h^2 f'(x)m'(x)\int Z^2K(Z)dZ + \frac{N-n}{2nh}(h^2 f(x)m''(x)\int Z^2K(Z)dZ + O(h^4)$$

$$= \frac{N-n}{nh} h^2 K_2(K)\left[f'(x)m'(x) + \frac{1}{2} f(x)m''(x)\right] \quad (21)$$

Letting $Q(x) = f'(x)m'(x) + \frac{1}{2} f(x)m''(x)$. Equation (21) becomes

$$= \frac{N-n}{nh} h^2 K_2(K)Q(x) + O(h^4) \quad (22)$$

Clearly, from the equation (22), the bias of the proposed estimator is of order $O(h^4)$ which is lower than the bias of the normal kernel density estimator which is of order $O(h^2)$, an indication of reduced bias.

4. DESCRIPTION OF THE POPULATION

The estimation of the finite population total and the corresponding bias was be carried out using five super-population totals; linear, quadratic, exponential, sine and jump models.

The description of the set of data for the populations is summarized in equations 23, 24, 25, 26 and 27 below. The auxiliary variable for each data set has been collected and incorporated in the estimators so as to improve on the precision of the estimation since the auxiliary variable is assumed to contain important information that is necessary for the estimation of the population total.

We will perform simulation studies to illustrate the performance of the boundary bias robust estimator for the finite population total. Further, we will investigate the bias of the derived estimator. 100 values will be generated. For the sake of efficiency, we will assume that the errors terms are independent and identically distributed, with homogeneous variances, and that there is only one auxiliary variable x.

ISSN 1533-9211

DOI 10.17605/OSF.IO/CUYX7

The superpopulation models below were be considered in the study;

Linear model

$$y_i = 1 + 2(x_i - 0.75) + e_i \quad (23)$$

Quadratic model

$$y_i = 1 + 2(x_i - 0.75)^2 + e_i \quad (24)$$

Exponential model

$$y_i = exp\{-4x_i\} + e_i \quad (25)$$

Jump model

$$y_i = 1 + 2(x_i - 0.75) + 0.25 + e_i \quad (26)$$

Sine model

$$y_i = 1 + sin(2\pi x_i) + e_i \quad (27)$$

The auxiliary variable is assumed to be distributed uniformly in the interval N[0, 1] and The error term is defined as a standard normal variable on (0, 1), defined as $e_i \sim N(0,1)$. Then, using the models above (in equations 23 : 27) we will compute the values for the response variable y.

A random sample of 100 was randomly selected from the generated data through simple random sampling without replacement.

5. SIMULATIONS

In this work we investigated how well the Adaptive boosting technique performs for two iterations, I.e m=2. The Tables 1, 2, 3, 4 and 5 and figure 1 below illustrate the behavior of the proposed finite population total estimator for various super-population models.

Table 1: Finite population total Estimate for linear model

	$\widehat{T}_{np}$	$\widehat{T}_{npAB_1}$	$\widehat{T}_{npAB_2}$
n=20	51.60925	43.17438	43.15549
n=35	50.27206	44.79664	44.17372
n=50	48.88752	45.47158	44.7088
n=85	46.35591	44.93168	44.7398

Table 2: Finite population total Estimates for Quadratic model

	$\widehat{T}_{np}$	$\widehat{T}_{npAB_1}$	$\widehat{T}_{npAB_2}$
n=20	144.1784	120.2492	120.2031
n=35	140.3838	124.8412	123.1222
n=50	136.458	126.8414	124.7286
n=85	129.0207	125.4413	124.9201

The Seybold Report
ISSN 1533-9211
DOI 10.17605/OSF.IO/CUYX7

Table 3: Finite population total Estimates for Exponential model

	$\hat{T}_{np}$	$\hat{T}_{npAB_1}$	$\hat{T}_{npAB_2}$
n=20	127.3841	126.2486	125.1751
n=35	105.0021	104.6434	104.3088
n=50	80.28447	80.47077	80.63151
n=85	47.62623	47.71608	47.78484

Table 4: Finite population total Estimates for Jump model

	$\hat{T}_{np}$	$\hat{T}_{npAB_1}$	$\hat{T}_{npAB_2}$
n=20	81.20815	67.54266	67.51416
n=35	79.03095	70.09063	69.11925
n=50	76.78285	71.16759	69.97648
n=85	72.59102	70.33901	70.04221

Table 5: Finite population total Estimates for Sine model

	$\hat{T}_{np}$	$\hat{T}_{npAB_1}$	$\hat{T}_{npAB_2}$
n=20	114.0265	94.15194	94.11835
n=35	110.8455	97.78002	96.43586
n=50	107.5239	99.38051	97.73005
n=85	101.2687	98.34453	97.93547

Clearly, from the above tables, the proposed estimator resulted to better finite population totals, $\hat{T}_{npAB_1}$ and $\hat{T}_{npAB_2}$ for the five super-population models under study as compared to the $\hat{T}_{np}$.

5.1 Performance of proposed estimator at different population sizes

This paper considered a sample of size 100 which was partitioned, in order to study the behaviour of the proposed finite population total at different sample sizes. The T_{npAB} after partitioning were as follows;

$$\hat{T}_{npAB} = \sum_{i=1}^{20} Y_i + \sum_{i=1}^{80} \widehat{m}_{AB}(x_i)$$

$$\hat{T}_{npAB} = \sum_{i=1}^{35} Y_i + \sum_{i=1}^{65} \widehat{m}_{AB}(x_i)$$

$$\hat{T}_{npAB} = \sum_{i=1}^{50} Y_i + \sum_{i=1}^{50} \widehat{m}_{AB}(x_i)$$

$$\hat{T}_{npAB} = \sum_{i=1}^{20} Y_i + \sum_{i=1}^{80} \widehat{m}_{AB}(x_i)$$

ISSN 1533-9211

DOI 10.17605/OSF.IO/CUYX7

$$\hat{T}_{npAB} = \sum_{i=1}^{85} Y_i + \sum_{i=1}^{15} \widehat{m}_{AB}(x_i)$$

The figures below illustrate the comparison of the proposed estimator, T_{npAB}, after the application of the AdaBoost. The results, clearly shows that the boosted finite population, T_{npAB}, outperformes the $\hat{T}_{np}$. It can be seen that the overall finite population totals of the suggested estimator remain at their lowest throughout, both on the first and second boosting for different sample sizes.

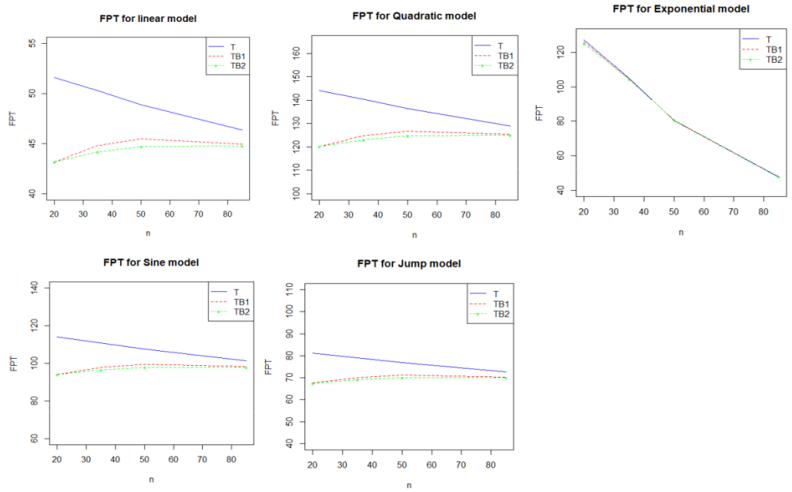

Figure 1: Plot of the Finite population Total for the five population models at different sample sizes n

It is clearly visible from Figure 1 that the $\hat{T}_{np}$ was reducing with increasing n for all the superpopulation models. However, both T_{npAB1} and T_{npAB2} were increasing with increasing n upto a point, followed by a decrease after n = 50 for linear, Quardratic, jump and Sine models. The exponential model resulted into a reducing finite population total throughout.

6. CONCLUSION

From the summary tables and the figures above, it can be seen clearly that the proposed estimator $\hat{T}_{npAB}$ has a bias of order $O(h^4)$ which converges faster. Moreover, it results to the finite population total $\hat{T}_{npAB}$ with a very smaller estimator both on first and second boosting as compared to the normal finite population total, $\hat{T}_{np}$.

Acknowledgements

We are grateful to anonymous referees for detailed and helpful comments that led to significant improvements in this paper.

The Seybold Report
ISSN 1533-9211
DOI 10.17605/OSF.IO/CUYX7

References

1. Chambers, R., Dorfman, A. H., & Hall, P. (1992). Properties of estimators of the finite popula-tion distribution function. Biometrika, 79(3), 577–582.
2. Dorfman, A. H. (1992). Nonparametric regression for estimating totals in finite populations. In Proceedings of the section on survey research methods (pp. 622–625).
3. Freund, Y. (1995). Boosting a weak learning algorithm by majority. Information and computation, 121(2), 256–285.
4. Freund, Y., Schapire, R. E., et al. (1996). Experiments with a new boosting algorithm. In icml (Vol. 96, pp. 148–156).
5. Nadaraya, E. A. (1964). On estimating regression. , 9, 141–142.
6. Parzen, E. (1962). On estimation of a probability density function and mode. The annals of mathematical statistics, 33(3), 1065–1076.
7. Rosenblatt, M., et al. (1956). Remarks on some nonparametric estimates of a density function. The Annals of Mathematical Statistics, 27(3), 832–837.
8. Schapire, R. E. (1990). The strength of weak learnability. Machine learning, 5(2), 197–227.
9. Schapire, R. E., & Singer, Y. (1999). Improved boosting algorithms using confidence-rated predictions. Machine learning, 37(3), 297–336.
10. Silverman, B. W. (1986). Density estimation for statistics and data analysis (Vol. 26). CRC press.
11. Watson, G. S. (1964). Smooth regression analysis. , 359–372