American Journal of Theoretical and Applied Statistics
2020; 9(4): 101-105
http://www.sciencepublishinggroup.com/j/ajtas
doi: 10.11648/j.ajtas.20200904.13
ISSN: 2326-8999 (Print); ISSN: 2326-9006 (Online)

Exploring Data-Reflection Technique in Nonparametric Regression Estimation of Finite Population Total: An Empirical Study

Langat Reuben Cheruiyot

Department of Mathematics and Computer Sciences, School of Science & Technology, University of Kabianga, Kericho, Kenya

Email address:
rlangat@kabianga.ac.ke

To cite this article:
Langat Reuben Cheruiyot. Exploring Data-Reflection Technique in Nonparametric Regression Estimation of Finite Population Total: An Empirical Study. *American Journal of Theoretical and Applied Statistics*. Vol. 9, No. 4, 2020, pp. 101-105. doi: 10.11648/j.ajtas.20200904.13

Received: May 6, 2020; **Accepted**: May 25, 2020; **Published**: June 3, 2020

Abstract: In survey sampling statisticians often make estimation of population parameters. This can be done using a number of the available approaches which include design-based, model-based, model-assisted or randomization-assisted model based approach. In this paper regression estimation under model based approach has been studied. In regression estimation, researchers can opt to use parametric or nonparametric estimation technique. Because of the challenges that one can encounter as a result of model misspecification in the parametric type of regression, the nonparametric regression has become popular especially in the recent past. This paper explores this type of regression estimation. Kernel estimation usually forms an integral part in this type of regression. There are a number of functions available for such a use. The goal of this study is to compare the performance of the different nonparametric regression estimators (the finite population total estimator due Dorfman (1992), the proposed finite population total estimator that incorporates reflection technique in modifying the kernel smoother), the ratio estimator and the design-based Horvitz-Thompson estimator. To achieve this, data was simulated using a number of commonly used models. From this data the assessment of the estimators mentioned above has been done using the conditional biases. Confidence intervals have also been constructed with a view to determining the better estimator of those studied. The findings indicate that proposed estimator of finite population total that is nonparametric and uses data reflection technique is better in the context of the analysis done.

Keywords: Conditional Biases, Reflection Technique, Confidence Lengths, Nonparametric Regression Estimation

1. Introduction

Many non-parametric techniques have in the recent past been used in regression estimation. They include techniques such as the *k*-nearest neighbors, local polynomial regession, spline regression, and orthogonal series [9, 19]. Besides this and in an attempt to correct the unpleasant boundary bias induced by the conventional Nadaraya-Watson estimator, many statisticians have endeavoured to modify it. Some of these include Gasser-Müller [13] and Priestley-Chao (1972). The drawback of these techniques is that their bias components were managed but at the expense of higher variability. In the framework of the model-based approach, regression estimation is paramount in obtaining estimates of the non-sample population. The flexible nature of the non-parametric technique has made it an attractive option in statistical researches [6]. The technique entails use of kernel smoothers that assign weights to observations used in estimation. In this paper we explore yet another new technique of reflection as a way of modifying the kernel smoothers with a view to minimizing the boundary bias the shortcoming of the Nadaraya-Watson estimator.

This paper has been organized as follows: in section 2, we give a brief review of the literature regarding non-parametric regression, in section 3; a new nonparametric regression estimator for finite population total is proposed. The estimator whose properties have been stated makes use of a modified kernel smoother obtained through reflection of data technique. Empirical analysis has been done in section 4 using some artificially simulated datasets. Discussion of results and conclusion is given in section 5.

2. Literature Review

A model- based non-parametric model (ξ) is conventionally of the form:

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

where Y_i- is the variable of interest

X_i-is the auxiliary variable

m-is an unknown function to be determined using sample data e_i-is error term-assumed to be $N(0, \sigma^2)$ under the model (ξ)

In nonparametric regression estimation $m(X_i)$ is an unknown function and can therefore be determined by the data sampled. Since this is a sample statistic, there are many estimators in place that have been developed by statisticians. They include the famous Nadaraya-Watson estimator which many have attempted to modify because of its weakness at the boundary. These can be found in Eubank [11] and Gasser and Müller [13].

A simple kernel estimator at an arbitrary point x as presented by Priestley and Chao (1972) can be written as:

$$\hat{m}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) Y_i \tag{2}$$

where h is the bandwidth, sometimes referred to as the tuning parameter or window width. $K(.)$ denotes a kernel function which is also twice continuously differentiable, symmetrical and having support within the bounded interval [-1, 1] such that:

$$\text{(a) } \int K(z)dz = 1 \quad \text{(b) } \int zK(z)dz = 0 \quad \text{(c) } \int z^2 K(z)dz := K_2(K)(<\infty) \tag{3}$$

For the derivation of the asymptotic bias term and even the variance term, one can see Kyung-Joon and Shucany [15]. They are respectively given by:

$$\text{Bias}[\,\hat{m}(x)\,] = \frac{1}{2}h^2 m''(x)K_2(K) + o(h^2) \tag{4}$$

and

$$\text{Var}[\,\hat{m}(x)\,] = \frac{\sigma^2}{nh}R(K) + o\left(\frac{1}{nh}\right) \tag{5}$$

where $R(K) = \int K(Z^2)dz$

The direct proportionality of the bias and the bandwidth means a small bandwidth will reduce it. While this is true for the bias a similar action of decreasing the bandwidth increases the variance making the regression curve to be wiggly. The implication of this scenario is that an optimal bandwidth that minimizes the mean square error (MSE) is necessary. Although with the use of the knowledge of calculus it is possible to obtain, such a bandwidth has never provided a solution to the boundary menace. Following this, Gasser and Müller [13] proposed optimal boundary kernels to address the problem. They suggested multiplying the truncated kernel at the boundary by a linear function. A generalized jackknife approach was proposed by Rice [16]. Eubank and Speckman [12] suggested the use of "bias reduction theorem" to remove the boundary effects. Schuster [18] gave another technique of correcting the boundary bias by using reflection of data method in density estimation. The same idea has also been reviewed by Albert and Karunamuni [1] among others, but notably within density estimation. This technique has further been examined in this paper but in the context of regression estimation. The technique is applied in estimating the finite population total and its performance has been analysed against other known estimators such as:

The ratio estimator given by:

$$\hat{T}_R = \hat{B}\sum_{i=1}^{N} X_i \tag{6}$$

Where $\hat{B} = \dfrac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}$, is the estimator for its equivalent population parameter, $\sum_{i=1}^{n} y_i$ is the sample total of the study variable while and $\sum_{i=1}^{n} x_i$ is the equivalent for the auxiliary variable assumed to be known for the entire population. The entire population total of this auxiliary variable is given by $\sum_{i=1}^{N} X_i$. It is known that the ratio estimator is the Best Linear Unbiased Predictor (BLUP), Cochran [7], Cox [8] and Brewer [5].

Another approach to estimation is the design-based estimator suggested by Horvitz-Thompson [14] is given by:

$$\hat{T}_{HT} = \sum_{i \in s} \pi_i^{-1} y_i \tag{7}$$

While the nonparametric regression estimator proposed by Dorfman [10] for finite population total is:

$$\hat{T}_{np} = \sum_{i=1}^{n} y_i + \sum_{i=n+1}^{N} \hat{m}_{NW}(x_i) \tag{8}$$

where $\hat{m}_{NW}(x_i)$ is the Nadaraya-watson estimator.

As noted above, this estimator suffers from boundary effects. But even with that weakness the nonparametric techniques in regression estimation have been known to outperform its counterparts-the fully parametric and semiparametric techniques. Dorfman [10] did a comparison between the population total estimators constructed from the famous design-based Horvitz-Thompson estimator and the Nadaraya-

Watson estimator- the nonparametric regression estimator where he found out that the nonparametric regression estimator better reflects the structure of the data and hence yields greater efficiency. This regression estimator, however, suffered the so called boundary bias besides facing bandwidth selection challenges. Breidt and Opsomer [3] did a similar study on nonparametric regression estimation of finite population total under two-stage sampling. Their study also reveals that the nonparametric regression with the application of local polynomial regression technique dominated the Horvitz-Thompson estimator and improved greatly the Nadaraya-Watson estimator. Breidt et al [4] carried out estimation of population of finite population total under two-stage sampling procedure and their results also show that the nonparametric regression estimation is superior to the standard parametric estimators when the model regression function is incorrectly specified, while being nearly as efficient when the parametric specification is correct.

We also propose an estimator under this nonparametric regression in the model-based framework.

3. Proposed Estimator

$$\hat{T}_{npr} = \sum_{i=1}^{n} y_i + \sum_{i=n+1}^{N} \hat{m}_{ref}(x_i) \tag{9}$$

where the first term $\sum_{i=1}^{n} y_i$ is the sample total observed and therefore under model-based approach it will not be necessitate estimation while the second term $\sum_{i=n+1}^{N} \hat{m}_{ref}(x_i)$ is the non-sample total term that is to be estimated non-parametrically using the reflection technique. The data-reflected technique therefore provides the data through reflection method so that this information is put on the negative axis thereby supplying the kernel with the information required on this section.

3.1. Data Reflection Procedure

The following simple steps give the procedure on how reflection of data is done. Let the $\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$ be the set of n observations in the sample. If the data is augmented by adding the reflections of all the points in the boundary, to give the set $\{(X_1, Y_1), (-X_1, Y_1), (X_2, Y_2), (-X_2, Y_2)\ldots, (-X_n, Y_n), (X_n, Y_n)\}$. If a kernel estimate $m^*(x)$ is constructed from this data set of size $2n$, then an estimate based on the original data can be given by putting $\hat{m}(x) = 2m^*(x)$, for $x \geq 0$, and zero otherwise. This gives the modified general weight function given by:

$$\hat{m}_{ref}(x) = \sum_{i=1}^{n} \frac{\left\{K\left(\frac{x-X_i}{h}\right) + K\left(\frac{x+X_i}{h}\right)\right\} Y_i}{\sum_{i=1}^{n}\left\{K\left(\frac{x-X_i}{h}\right) + K\left(\frac{x+X_i}{h}\right)\right\}} \tag{10}$$

It can be shown that the estimate will always have zero derivative at the boundary, provided the kernel is symmetric and differentiable. The estimate has also been shown under the section on properties of the data-reflected technique that it is a *p.d.f* for the symmetric kernel. In practice it will not usually be necessary to reflect the whole data set, since if X_i/h is sufficiently large, the reflected point - X_i/h will not be felt in the calculation of $m^*(x)$ for $x \geq 0$, and hence reflection of points near 0 is all that is needed. Silverman [17] in his example, states that if K is the Gaussian kernel there is no practical need to reflect points beyond $X_i > 4h$.

3.2. Asymptotic Properties of the Proposed Estimator

It can be shown (one can see Albers [2] for similar derivation under the density estimation) that the asymptotic bias and the variance of the proposed estimator are respectively given by:

$$Bias\left[T_{npr}\right] = \left(\frac{N-n}{n}\right)\left[\frac{h^2}{2} m''(x)\int_{-1}^{1} z^2 K(z)dz + 2h\left(m'(0)[g(x)]^{-1} + chm''(0) + o(h)\right)\int_{c}^{1}(z-c)K(z)dz + 2h^2 m''(x)\left(c^2\int_{-1}^{-c} K(z)dz + c\int_{-1}^{-c} zK(z)dz\right)\right] + o\left(h^2\right) \tag{11}$$

and

$$\begin{aligned}\operatorname{var}\left(T_{npr}\right) &= \frac{(N-n)^2\sigma^2}{nh^2\left[\hat{g}(x)\right]^2}\left(hg(x)\int_{-1}^{1} K(z)^2\,dz + O\left(h^2\right)\right) + O\left(n^{-1}\right) \\ &= \frac{(N-n)^2\sigma^2}{nhg(x)}\int_{-1}^{1} K(z)^2\,dz + O\left(n^{-1}\right) \approx \frac{(N-n)^2\sigma^2}{nhg(x)} R(K)\end{aligned} \tag{12}$$

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

4. Empirical Study

To examine the performance of the proposed estimator, simulation was done from various common distributions and analysis was done to compare them based on their confidence lengths and conditional biases. Table 1 gives the models used in simulation.

Table 1. *Equations of models simulated.*

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

4.1. Unconditional 95% C.I for the Respective Population Total Estimators

The 95% confidence interval of each of the estimators was

also computed using the formula given by; $T = \hat{T} \pm Z_{\alpha/2}\sqrt{Var(\hat{T})}$ and the interval length is therefore the difference between the upper limit and the lower limit. The results are presented in table 2.

Table 2. *Summary results for the unconditional confidence interval lengths.*

MODEL	$\hat{T}_{npr}$	$\hat{T}_{np}$	$\hat{T}_{HT}$	$\hat{T}_{R}$
LINEAR	10.70248	11.51506	60.75419	10.4797
QUADRATIC	10.67005	11.09624	18.99733	72.39773
SINE	10.77354	19.22728	75.80554	185.6727
EXPONENTIAL	10.47873	12.36678	25.15132	30.85746
JUMP	11.18108	12.98228	16.63127	97.24218
BUMP	11.12733	20.55477	65.69977	31.66942

Notice that the confidence lengths given by the proposed estimator in the first column are the least of all except for the ratio estimator under the linear model.

4.2. Conditional Performance of the Respective Population Total Estimators

To study the conditional performance of the estimators, the sample means $\bar{x}_i$'s were calculated and ranked in ascending order while maintaining the corresponding estimates, $\hat{T}_i$'s, of the finite population totals. Forty groups of 50 samples each were then obtained as per the new order of the rankings. From each of these groups the sample means of the auxiliary variable were averaged to give, $\bar{\bar{x}}_j$, the mean of the sample means of the j^{th} group (j= 1, 2, …, 40). The corresponding population totals i.e $\hat{\bar{T}}_j$'s for the various population estimators studied were also computed and used to calculate the respective conditional biases for the models given. The results have been plotted in the Figures 1-4.

The figures portray that the proposed estimator is better placed than the other estimators examined in terms of posting a smaller conditional bias.

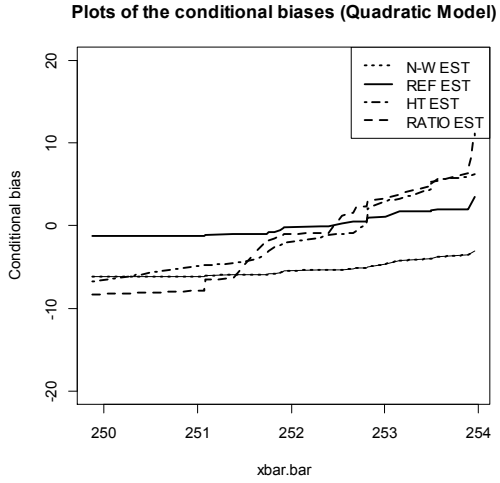

Figure 1. *Comparison of conditional bias for the respective finite population total estimators (Quadratic model).*

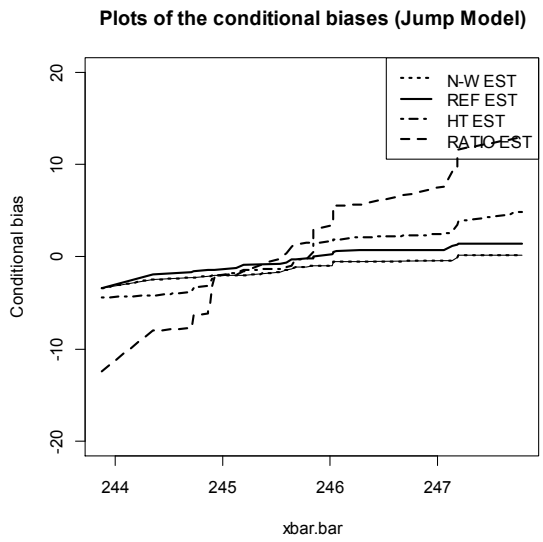

Figure 2. *Comparison of conditional bias for the respective finite population total estimators (Jump model).*

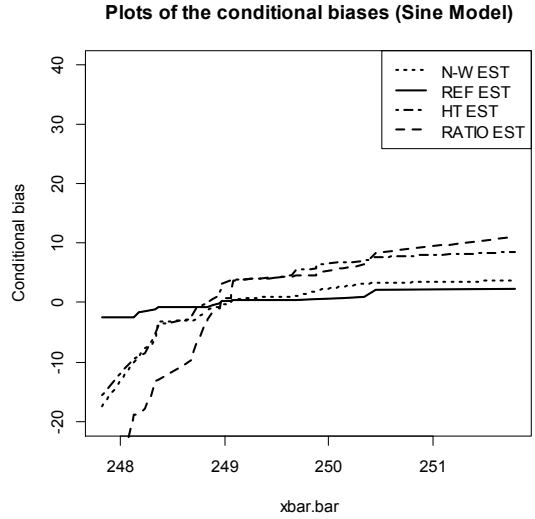

Figure 3. *Comparison of conditional bias for the respective finite population total estimators (Sine model).*

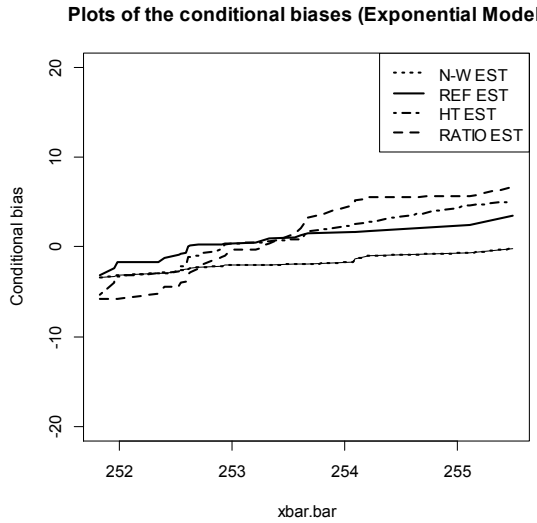

Figure 4. *Comparison of conditional bias for the respective finite population total estimators (Exponential model).*

5. Conclusion

The proposed estimator of the finite population total that uses the reflection technique shows narrower confidence lengths as opposed to the others considered in the study. The smaller 95% confidence lengths is a characteristic of a better estimator that is more precise and accurate.

Further the graphs given in the figures above shows that the proposed estimator outwits the others. The graphs show that the proposed estimator is almost conditionally unbiased.

It can therefore be concluded that based on the analysis done in this study reflection technique can be of benefit in correcting the boundary bias usually experienced with the use of kernel estimators in regression estimation.

References

[1] Alberts T and Karunamuni R. J. (2007). Boundary correction methods in Kernel density estimation. Presentation slides downloaded from internet.

[2] Albers M. G. (2011). Boundary Estimation of Densities with Bounded Support. Swiss Federal Institute of Technology Zurich. Masters' Thesis.

[3] Breidt, B. F. and Opsomer, J. D. (2000). Local Polynomial regression in Survey Sampling. *Annals of Statistics* 28 (August): 1026-1053.

[4] Breidt, F. J. and Opsomer, J. (2009). Nonparametric and Semiparametric Estimation in Complex Surveys, Handbooks of Statistics, vol. 29B, eds. D. Pfeffermann and C. R. Rao, 103-121.

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

[6] Čížek, P., & Sadıkoglu, S. (2020). *Robust nonparametric regression: A review.* Retrieved from https://doi.org/10.1002/wics.1492.

[7] Cochran, W. G. (1977), *Sampling Techniques* (3rd ed.), New York: John Wiley.

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

[9] Dhekale, B. S., Sahu, P. K., Vishwajith, K. P., Mishra, P., & Narsimhaiah, L. (2017). Application of parametric and nonparametric regression models for area, production and productivity trends of tea (Camellia sinensis) in India. *Indian Journal of Ecology, 44* (2), 192-200.

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

[11] Eubank, R. L., 1988. Spline Smoothing and Nonparametric Regression. Marcel Dekker, New York.

[12] Eubank, R. L., Speackman, P. L., (1991). A bias reduction theorem with application in nonparametric regression, *Scandinavian Journal of Statistics*. 18, 211- 222.

[13] Gasser, Th., Miiller, H. G., (1979). In: Gasser, Th., Rosenblatt, M. (Eds.) Kernel estimation of regression functions: in Smoothing Techniques for Curve Estimation. Springer, Heidelberg, pp. 23-68.

[14] Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

[15] Kyung-Joon C. and Schucany W. R. (1998), Nonparametric kernel regression estimation near endpoints *Journal of Statistical Planning and Inference* 66 289-304.

[16] Rice, J., 1984. Boundary modification for kernel regression. *Communication in Statistics*. A 13, 893-900.

[17] Silverman, B. W. (1986). Density Estimation for Statistics and Data Analysis, London: Chapman and Hall.

[18] Schuster E. F (1985). Incorporating support into nonparametric estimators of densities, *Communication in Statistics*. A 14, 1123-36.

[19] Tibshirani, R., & Wasserman, L. (2015). *Nonparametric Regression- Statistical Machine Learning.* Retrieved May 22, 2020, from http://www.stat.cmu.edu.