Scientific Research Publishing

Open Journal of Statistics, 2016, 6, 1085-1097
http://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Local Polynomial Regression Estimator of the Finite Population Total under Stratified Random Sampling: A Model-Based Approach

Charles K. Syengo[1], Sarah Pyeye[1], George O. Orwa[2], Romanus O. Odhiambo[2]

[1]Pan African University Institute for Basic Sciences, Technology and Innovation, Nairobi, Kenya
[2]Department of Statistics and Actuarial Science, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya
Email: kilundac@gmail.com, srhpyeye@gmail.com, gorwa@jkuat.ac.ke, romanusemod@yahoo.com

How to cite this paper: Syengo, C.K., Pyeye, S., Orwa, G.O. and Odhiambo, R.O. (2016) Local Polynomial Regression Estimator of the Finite Population Total under Stratified Random Sampling: A Model-Based Approach. *Open Journal of Statistics*, **6**, 1085-1097.
http://dx.doi.org/10.4236/ojs.2016.66088

Received: September 5, 2016
Accepted: December 3, 2016
Published: December 8, 2016

Abstract

In this paper, auxiliary information is used to determine an estimator of finite population total using nonparametric regression under stratified random sampling. To achieve this, a model-based approach is adopted by making use of the local polynomial regression estimation to predict the nonsampled values of the survey variable *y*. The performance of the proposed estimator is investigated against some design-based and model-based regression estimators. The simulation experiments show that the resulting estimator exhibits good properties. Generally, good confidence intervals are seen for the nonparametric regression estimators, and use of the proposed estimator leads to relatively smaller values of RE compared to other estimators.

Keywords

Sample Surveys, Stratified Random Sampling, Auxiliary Information, Local Polynomial Regression, Model-Based Approach, Nonparametric Regression

1. Introduction

Sample surveys' main objective is to obtain information about the population, and then use such information to make inference about some population quantities. The information that is mostly sought about the population is usually aggregate values of various population characteristics, total number of units, proportion of units having certain attributes. The information can be collected by either sampling methods or census. One of the approaches to using auxiliary information in construction of estimators is by assuming a working model that describes the relationship between the survey variable and the auxiliary variable. Estimators are then derived based on this

model. At this stage, estimators are sought to have good efficiency given that the model is true. In most cases, a linear model is assumed. Generalized regression estimators by [1] and [2] including linear regression estimators and ratio estimators by [3], and best linear unbiased estimators by [4] and [5] and post-stratification estimators by [6] as well are all derived from the assumption of linear models. Sometimes the linear model fails, and therefore, the resulting estimators do not beat the purely design-based estimators. As a result, [7] proposed a class of estimators in which the working model assumes a nonlinear parametric model. The improvement of the efficiency of such estimators, however, requires prior information about the exact parametric population structure. As a result of these concerns, several researchers have so far considered nonparametric models for ξ. Nonparametric regression may be used in the estimation of unknown finite population quantities such as population totals, means, proportions or averages. The idea of nonparametric regression traces its origin in works by [8] and [9]. Nonparametric-based estimation is often more robust and flexible than inference based on parametric regression models or design probabilities (as in designed-based inference) [10]. In sample surveys, auxiliary information is used at the estimation stage of finite population quantities-population total or mean, say-to increase the precision of estimators of such population quantities [11] [12] [13].

A variety of approaches exist for construction of more efficient estimators for population total or mean, and they include model-based and design-based methods. Model-based approach in sample surveys is based on superpopulation models, which assumes that the population under study is a realization of a random variable having a superpopulation model ξ. This model ξ is used to predict the nonsampled values of the population, and hence the finite population quantities, total Y or mean $\bar{Y}$ [13]. [14] first considered nonparametric models for ξ within a model-assisted approach and obtained a local polynomial regression estimator as a generalization of the ordinary generalized regression estimator. Their simulation study shows that the proposed estimator performs relatively better than other parametric estimators. [13] improved on [14] estimator and developed a model-based local polynomial regression estimator applicable to direct sampling designs such as simple random sampling and systematic sampling. Their estimator demonstrates better performance than [14] model-assisted estimator. Their estimator also beats other parametric estimators.

In this paper, auxiliary information is used to determine an estimator of finite population total using nonparametric regression under stratified random sampling. To achieve this, a model-based approach is adopted by making use of the local polynomial regression estimation to predict the nonsampled values of the survey variable *y*. Stratified estimators for finite population total Y or mean $\bar{Y}$ have proved to yield better estimators than those resulting from simple random sampling [15] [16]. Additionally, it has been shown in the literature that local polynomial approximation method has several nice features including satisfactory boundary behaviour, easy interpretability, applicability for a variety of design-circumstances and nice minimax properties (see [17] [18] and [19]).

2. Proposed Estimator

Consider a population consisting of N units. Suppose this population is divided into H disjoint strata, each of size $N_h, h=1,2,\cdots,H$.

Let $y_{hj}, j=1,2,\cdots,N_h$ be the survey measurement for the j^{th} unit in the h^{th} stratum. Further, let $x_{hj}, j=1,2,\cdots,N_h$ be the auxiliary measurement positively correlated with y_{hi}.

From each stratum, a simple random sample of size n_h is selected without replacement, where n_h is sufficiently large with respect to N_h and $f_h = n_h/N_h \to 0$.

Let s_h be the sample in the h^{th} stratum and r_h be the nonsampled set in the h^{th} stratum.

The population total is defined as

$$Y = \sum_{h=1}^{H}\sum_{j=1}^{N_h} y_{hj} = \sum_{h=1}^{H}\sum_{j=1}^{n_h} y_{hj} + \sum_{h=1}^{H}\sum_{j=n_h+1}^{N_h} y_{hj} \tag{1}$$

which can rewritten as

$$Y = \sum_{h=1}^{H} y_{h_s} + \sum_{h=1}^{H} y_{h_r} \tag{2}$$

where $y_{h_s} = \sum_{j=1}^{n_h} y_{hj}$ and $y_{h_r} = \sum_{j=n_h+1}^{N_h} y_{hj}$.

Once the sample has been observed, the problem of estimating Y becomes the problem of predicting the sum of the nonsampled $y_{hj}'s$. Usually, inference is made using the known sample and the model ξ.

The first component in Equation (1) is known while the second requires prediction which is the focus in this paper. In this paper, local polynomial regression method will be used to predict the unknown $y_{hj}'s$, $\forall j \in r_h$.

Suppose the distribution generating $y_{hj}'s$ is given by the superpopulation model, ξ in which

$$y_{hj} = m\left(x_{hj}\right) + e_{hj} \tag{3}$$

where $e_{hj}'s$ are independently distributed random variables with mean 0 and variance $\sigma^2\left(x_{hj}\right)$.

Then it follows that

$$E\left(y_{hj}\right) = m\left(x_{hj}\right) \tag{4}$$

$$Cov\left(y_{hj}, y_{h'j'}\right) = \begin{cases} \sigma^2\left(x_{hj}\right), & \text{for } h=h' \text{ and } j=j' \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where $\sigma^2(x)$ and $m(x)$ are assumed to be continuous and twice differentiable functions of x, and $\sigma^2(x) > 0$.

In practice, the values of $m(x)$ are unknown and so requires prediction. Adopting [13] [14] and [20] ideas, we make use of local polynomial regression of degree p, which is a generalization of the kernel smoothing, to predict the unobserved $y_{hj}'s$ in Equation (1). Let $K_b(u) = b^{-1}K(u/b)$, where K denotes a continuous kernel function and b is the bandwidth.

Then a model-based local polynomial regression estimator of the nonsampled $y'_{hj}s$ in the h^{th} stratum is given by:

$$\hat{m}_{hj} = e_1^{\mathrm{T}} \left(X_{hj}^{\mathrm{T}} W_{hj} X_{hj} \right)^{-1} X_{hj}^{\mathrm{T}} W_{hj}^{\mathrm{T}} y = w_{hj}^{\mathrm{T}} y \tag{6}$$

where $e_1 = (1, 0, 0, \cdots, 0)^{\mathrm{T}}$ is a column vector of length $p+1$; $y = \left[y_{hj} \right]_{j \in s_h}$; $W_{hj} = diag\left\{ K_b \left(x_{hj} - x_{hi} \right) \right\}_{j \in s_h}$ and $X_{hj} = \left[1, \left(x_{hj} - x_{hi} \right), \cdots, \left(x_{hj} - x_{hi} \right)^p \right]_{j \in s_h}$. Equation (6) holds as long as $X_{hj}^{\mathrm{T}} W_{hj} X_{hj}$ is a nonsingular matrix.

Now denoting the estimator for the finite population total by $\hat{Y}_{LP}$ and the estimator within the h^{th} stratum by $\hat{Y}_{LP_h}$. Therefore, in stratum *h*, the estimator of the population total based on local polynomial regression is

$$\hat{Y}_{LP_h} = y_{h_s} + \sum_{j=n_h+1}^{N_h} \hat{m}_{hj} \tag{7}$$

and the estimator for the finite population total is

$$\hat{Y}_{LP} = \sum_{h=1}^{H} \hat{Y}_{LP_h} = \sum_{h=1}^{H} \left(y_{h_s} + \sum_{j=n_h+1}^{N_h} \hat{m}_{hj} \right) \tag{8}$$

with $y_{h_s} = \sum_{j=1}^{n_h} y_{hj}$.

3. Properties of Proposed Estimator

In this section, a study is carried out on various properties of estimator (8), which may be important in practice. In doing so, the following assumptions are made:

1) The regression function $m(x)$ has a bounded second derivative.

2) The marginal density, $f_X(x)$ is continuous and $f_X(x) > 0$.

3) The conditional variance $\sigma^2(x) = \mathrm{var}(Y/X = x)$ is bounded and continuous.

4) The kernel density function $K(x)$ is bounded and continuous satisfying the following: $\int_{-\infty}^{\infty} K(x)\mathrm{d}x = 1$, $\int_{-\infty}^{\infty} xK(x)\mathrm{d}x = 0$, $\int_{-\infty}^{\infty} x^2 K(x)\mathrm{d}x > 0$ and $\int_{-\infty}^{\infty} x^{2t} K(x)\mathrm{d}x < \infty$ for $t = 1, 2, \cdots$.

These conditions on $K(\cdot)$ were imposed and used in [18] work and are purposely for the convenience of technical arguments and therefore can be relaxed.

3.1. $\hat{Y}_{LP}$ Is Asymptotically Model-Unbiased

Now consider the difference:

$$\hat{Y}_{LP} - Y = \left(\sum_{h=1}^{H} y_{h_s} + \sum_{h=1}^{H} \sum_{j \in r_h} \hat{m}_{hj} \right) - \left(\sum_{h=1}^{H} y_{h_s} + \sum_{h=1}^{H} \sum_{j \in r_h} y_{hj} \right) \tag{9}$$

$$= \sum_{h=1}^{H} \sum_{j \in r_h} \left(\hat{m}_{hj} - y_{hj} \right) \tag{10}$$

$$= \sum_{h=1}^{H} \sum_{j \in r_h} \left(\left(\hat{m}_{hj} - m_{hj} \right) + \left(m_{hj} - y_{hj} \right) \right) \tag{11}$$

and taking expectation yields

$$E_\xi\left(\hat{Y}_{LP}-Y\right)=\sum_{h=1}^{H}\sum_{j\in r_h}E_\xi\left(\hat{m}_{hj}-m_{hj}\right)+\sum_{h=1}^{H}\sum_{j\in r_h}E_\xi\left(m_{hj}-y_{hj}\right) \tag{12}$$

$$=\sum_{h=1}^{H}\sum_{j\in r_h}E_\xi\left(\hat{m}_{hj}-m_{hj}\right) \tag{13}$$

since $E_\xi\left(y_{hj}\right)=m_{hj}$

i.e.

$$E_\xi\left(\hat{Y}_{LP}-Y\right)=\sum_{h=1}^{H}\sum_{j\in r_h}E_\xi\left(\hat{m}_{hj}-m_{hj}\right) \tag{14}$$

which is the bias associated with $\hat{Y}_{LP}$.

Approximating m_{hj} by Taylor series expansion about a point x_{hj} and assuming further that $n_h\to\infty$ and $b\to 0$, then observe that

$$\hat{m}_{hj}\approx m_{hj}+m'_{hj}\left(x_{hj}-x_{hi}\right)+\left(1/2\right)m''_{hj}\left(x_{hj}-x_{hi}\right)^2+\cdots. \tag{15}$$

Letting $u=\left(x_{hj}-x_{hi}\right)/b\Rightarrow ub=x_{hj}-x_{hi}$, then

$$\hat{m}_{hj}\approx m_{hj}+m'_{hj}\left(ub\right)+\left(1/2\right)m''_{hj}\left(ub\right)^2+O\left(b^2\right) \tag{16}$$

$$\Rightarrow \hat{m}_{hj}-m_{hj}\approx m'_{hj}\left(ub\right)+\left(1/2\right)m''_{hj}\left(ub\right)^2+O\left(b^2\right) \tag{17}$$

and applying expectations then

$$E_\xi\left(\hat{m}_{hj}-m_{hj}\right)=E_\xi\left(m'_{hj}\left(ub\right)+\left(1/2\right)m''_{hj}\left(ub\right)^2\right)+O\left(b^2\right) \tag{18}$$

Theorem 3 of [21] allows that under conditions (1)-(4) if $b\to 0$ and $n_h b\to\infty$,

$$\begin{aligned}&E_\xi\left(m'_{hj}\left(ub\right)+\left(1/2\right)m''_{hj}\left(ub\right)^2\right)+O\left(b^2\right)\\&\to m'_{hj}b\int uK_b\left(u\right)\mathrm{d}u+\left(1/2\right)m''_{hj}b^2\int u^2K_b\left(u\right)\mathrm{d}u+O\left(b^2\right)\end{aligned} \tag{19}$$

$$=\left(1/2\right)m''_{hj}b^2\int u^2K_b\left(u\right)\mathrm{d}u+O\left(b^2\right) \tag{20}$$

So that

$$E_\xi\left(\hat{m}_{hj}-m_{hj}\right)=\left(1/2\right)m''_{hj}b^2\int u^2K_b\left(u\right)\mathrm{d}u+O\left(b^2\right) \tag{21}$$

It implies that $E_\xi\left(\hat{m}_{hj}-m_{hj}\right)\to 0$ provided that $b\to 0$ and $n_h\to\infty$, and thus $\hat{Y}_{LP}$ is asymptotically model-unbiased.

3.2. Mean Square Error (*MSE*) of $\hat{Y}_{LP}$

The estimator (8) has the *MSE*

$$MSE\left(\hat{Y}_{LP}\right)=E_\xi\left(\hat{Y}_{LP}-Y\right)^2 \tag{22}$$

which can be decomposed as

$$MSE\left(\hat{Y}_{LP}\right)=\left[Bias\left(\hat{Y}_{LP}\right)\right]^2+Var\left(\hat{Y}_{LP}\right) \tag{23}$$

Theorem 1 of [18] allows that under Condition (1), if $b=dn_h^{-\gamma}, 0<\gamma<1$ then

$$MSE\left(\hat{Y}_{LP}\right)=\left(b^4/4\right)\sum_{h=1}^{H}\sum_{j\in r_h}\left(m''_{hj}\int_{-\infty}^{\infty}u^2K_b\left(u\right)\mathrm{d}u\right)^2 + b^{-1}\sum_{h=1}^{H}\sum_{j\in r_h}n_h^{-1}f^{-1}\left(x_{hj}\right)\sigma^2\left(x_{hj}\right)\int_{-\infty}^{\infty}K_b^2\left(u\right)+O\left(b^4+\left(n_hb\right)^{-1}\right) \tag{24}$$

Observe that Equation (24) tends to zero if $b\to 0$ and $n_hb\to\infty$ and thus $MSE\left(\hat{Y}_{LP}\right)\to 0$.

This shows that $\hat{Y}_{LP}$ is statistically consistent and thus useful.

4. Simulation Study

In this section, a study is carried out on the practical performance of several estimators (see Table 1 and Table 2 for the estimators).

The first estimator is design-based, the second one is parametric and model-based while the last two are nonparametric and model-based.

4.1. Description of the Population

The working model is taken to be $E\left(y_{hj}\right)=m\left(x_{hj}\right)$, $Cov\left(y_{hj},y_{h'j'}\right)=\sigma^2$. In this study, four populations are considered, which are generated from the regression model given by

$$y_i=m\left(x_i\right)+e_i \tag{25}$$

$1\le i\le 2000$ with the following mean functions

$$Linear: m_1\left(x\right)=1+2\left(x-0.5\right) \tag{26}$$

$$Sine: m_2\left(x\right)=2+\sin\left(2\pi x\right) \tag{27}$$

$$Bump: m_3\left(x\right)=1+2\left(x-0.5\right)+\exp\left(-200\left(x-0.5\right)^2\right) \tag{28}$$

$$Jump: m_4\left(x\right)=1+2\left(x-0.5\right)I_{\{x\le 0.65\}}+0.65I_{\{x>0.65\}} \tag{29}$$

with $x\in\left[0,1\right]$. They represent a class of correct and incorrect model specifications for the estimators being considered. For m_1, $\hat{Y}_{REG}$ is expected to be the best estimator, since the model assumed is correctly specified. The rest of the mean functions: m_2, m_3 and m_4 represent various deviations from the linear model, m_1. These populations are plotted in Figure 1. For more on these populations, see [13] and [14].

The errors are assumed to be independent and identically distributed (i.i.d) normal random variables having mean 0 and standard deviation $\sigma=0.1$. They contain 2000 units and the population x_i is simulated as i.i.d uniform random variables. The

Table 1. Estimators being compared in the Simulation study.

$\hat{Y}_{HT}$	Horvitz-Thompson	[22]
$\hat{Y}_{REG}$	Linear regression	[3], p. 200
$\hat{Y}_{PE}$	Mixed Ratio	[15]
$\hat{Y}_{LP}$	Local polynomial with $p=1$	Equation (8)

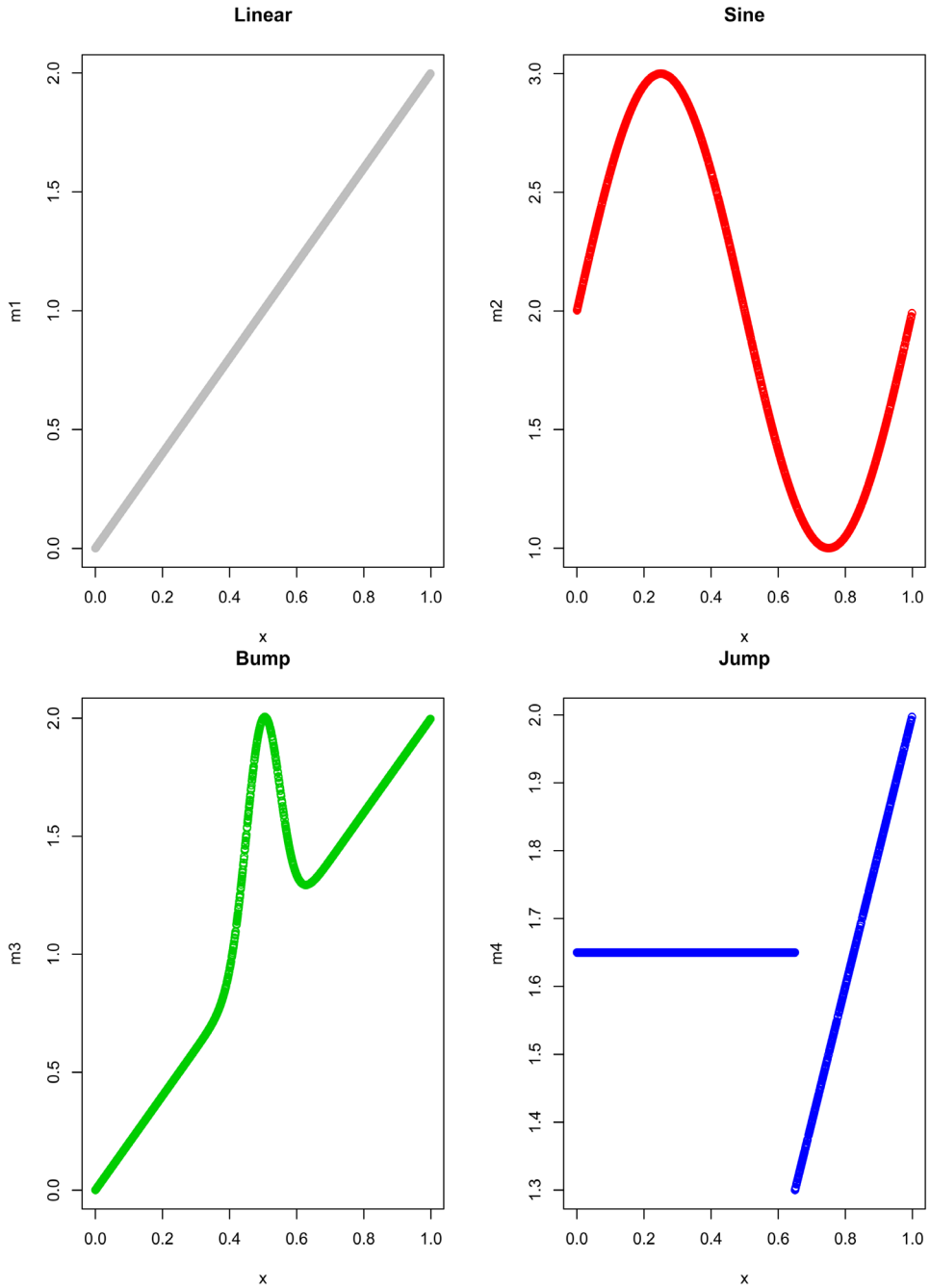

Figure 1. Plot of linear, sine, bump and jump populations.

population values $y_i's$ $(i=1,2,3,4)$ are generated from the mean functions by adding the errors $e_i's$ in each of the cases. Each of the populations is divided into 10 equal, disjoint and mutually exclusive strata which are made as homogeneous as possible to ensure that units in each stratum vary little from each other. A sample of size, $n=200$ is then taken with each stratum contributing a sample size of $n_h=20$, $(h=1,2,\cdots,10)$. 1000 samples are simulated using simple random sampling without replacement for each case.

Epanechnikov kernel,

$$K(u)=\frac{3}{4}\left(1-u^2\right)I_{\{|u|\le 1\}}, \tag{30}$$

is used for kernel smoothing on each of the populations. In each case, bandwidth values $b=n^{-1/5}$ (see [20]) (with $n=200$), $b=0.4$, $b=1$ and $b=2$ (see [15]) are considered.

Data simulations, the estimators and computations were obtained using *R* Software on a desktop.

To analyze the performance of the proposed estimator against some specified estimators, relative absolute bias (*RAB*) is computed as

$$RAB\left(\hat{\theta}\right)=\sum_{i=1}^{R}\left|\frac{\left(\hat{\theta}\left(s_i\right)-Y\right)}{Y}\right| \tag{31}$$

and the relative efficiency (*RE*) with respect to the Horvitz-Thompson (HT) estimator is computed as

$$RE\left(\hat{\theta}\right)=\frac{\sum_{i=1}^{R}\left(\hat{\theta}\left(s_i\right)-Y\right)^2}{\sum_{i=1}^{R}\left(\hat{Y}_{HT}\left(s_i\right)-Y\right)^2} \tag{32}$$

$\hat{\theta}$ is the estimator of the finite population total being considered; Y is the true population total and R is the number of replications.

The relative efficiency (*RE*) is meant to examine the robustness of the various estimators against the proposed estimator.

The confidence intervals (*CI*) and the average lengths (*AL*) of the confidence intervals of various estimators are also computed as follows:

$$CI\left(\hat{\theta}\right)=\sum_{i=1}^{R}\left(\hat{\theta}\left(s_i\right)\pm 1.96\sqrt{Var\left(\hat{\theta}\left(s_i\right)\right)}\right) \tag{33}$$

$$AL\left(\hat{\theta}\right)=\frac{1}{R}\sum_{i=1}^{R}\left(CI_U\left(\hat{\theta}\left(s_i\right)\right)-CI_L\left(\hat{\theta}\left(s_i\right)\right)\right) \tag{34}$$

where CI_U and CI_L are the upper and lower confidence limits respectively; $\hat{\theta}$ and R are as defined earlier.

4.2. Results

The results of this simulation study are summarized in Table 3 and Table 4. For each populations, $y_i's$ ($i=1,2,3,4$), the performance of each estimator is analyzed using the *RAB* and *RE*. The *RAB* indicates the measure of how close the estimator being considered is from the actual value, while the *RE* is used to check the robustness of the estimator. For instance, an estimator, $\hat{\theta}_1$, will be said to be "better" or more preferable than another one, $\hat{\theta}_2$, if its *RE* is comparably smaller. That is, if $RE\left(\hat{\theta}_1\right)<RE\left(\hat{\theta}_2\right)$, where $\hat{\theta}_1$ and $\hat{\theta}_2$ are estimators, then $\hat{\theta}_1$ is said to be "better" than $\hat{\theta}_2$.

Table 2. Summary of the formulae used in computing the respective population totals of the various estimators.

Estimator	*Formulae*
Horvitz-Thompson, $\hat{Y}_{HT}$	$\hat{Y}_{HT} = \sum_{h=1}^{H}\sum_{j=1}^{n_h}\left(y_{hj}/\Pi_{hj}\right)$
Linear regression estimator, $\hat{Y}_{REG}$	$\hat{Y}_{REG} = \sum_{h=1}^{H} N_h\left(\bar{y}_h + \beta_h^o\left(\bar{X}_h - \bar{x}_h\right)\right)$
Mixed Ratio Estimator, $\hat{Y}_{PE}$	$\hat{Y}_{PE} = \sum_{h=1}^{H} y_h + \sum_{h=1}^{H}\sum_{j\in r_h} w_{hj}\left(x_i\right) y_{hj}$, $y_h = \sum_{j=1}^{n_h} y_{hj}$
Proposed Model-based Local polynomial with $p=1$, $\hat{Y}_{LP}$	$\hat{Y}_{LP} = \sum_{h=1}^{H} y_h + \sum_{h=1}^{H}\sum_{j\in r_h} \hat{m}_{hj}$

The confidence intervals and average length of the intervals are also measured for each case. A smaller length is better because it implies that the true population total is captured within a smaller range and therefore results are more precise.

The estimators $\hat{Y}_{PE}$ and $\hat{Y}_{LP}$ are tested under the same bandwidth choice *i.e.* $b = n^{-1/5}$ (with $n = 200$), $b = 0.4$, $b = 1$ and $b = 2$. Results of this simulation are shown in Table 3 and Table 4 below.

Table 3 shows the *RAB*'s and *RE*'s of the various estimators with respect to the Horvitz-Thompson estimator ($\hat{Y}_{HT}$). Table 4 shows the confidence intervals and their average lengths.

In most scenarios, $\hat{Y}_{LP}$ is better than the parametric estimators, but the parametric estimator, $\hat{Y}_{REG}$, performs best when the model is correctly specified, as Table 3 shows. This occurs both in the linear and the bump populations, where in the former, a strong linear relationship holds between the variables while in the latter, the function is linear over most of its range despite a "bump" for a small part of the range of $x_{hi}'s$.

When the model is completely misspecified as in the Sine and Jump populations, a greater efficiency can be achieved by the nonparametric regression estimators. This can be seen in Table 3 for the Sine and Jump populations: the nonparametric estimators ($\hat{Y}_{LP}$ and $\hat{Y}_{PE}$) are more efficient than their parametric opponent, $\hat{Y}_{REG}$.

When the underlying superpopulation model is completely unknown, a reasonable choice for finite population total estimation would be the nonparametric estimators such as $\hat{Y}_{LP}$ and $\hat{Y}_{PE}$ with small bandwidth choices. This can be seen in Table 3 and Table 4.

In this study, $\hat{Y}_{LP}$ is sometimes seen to perform much bettter but not as worse as $\hat{Y}_{PE}$, and hence the proposed estimator, $\hat{Y}_{LP}$ emerges as the best performing among the nonparametric estimators being considered here (see Table 3). A good overall performance is observed with the proposed estimator, with smaller values of *RAB* and *RE* than the model-based competitor $\hat{Y}_{PE}$ for every population and fixed bandwidth under consideration.

Despite $\hat{Y}_{LP}$ being relatively the best estimator, its performance is significantly affected by the bandwidth choices. As the bandwidth size increases, some amount of efficiency is lost (see Table 3).

Table 3. Relative absolute bias (*RAB*) and relative efficiency (*RE*) based on 1000 replications of simple random sampling within strata from four fixed populations of size $N = 2000$. Sample size is $n = 200$.

Population	*b*	$\hat{Y}_{HT}$		$\hat{Y}_{REG}$		$\hat{Y}_{PE}$		$\hat{Y}_{LP}$	
		RAB	*RE*	*RAB*	*RE*	*RAB*	*RE*	*RAB*	*RE*
Linear	0.3465724	0.03212401	1	0.005778929	0.03155733	0.03321496	1.067811	0.03201888	0.9959899
	0.4	0.03212401	1	0.005778929	0.03155733	0.0335352	1.089573	0.0320533	0.9965037
	1	0.03212401	1	0.005778929	0.03155733	0.03434122	1.144951	0.03210449	0.9991698
	2	0.03212401	1	0.005778929	0.03155733	0.03272264	1.037753	0.03212023	0.9997907
Estimated Total	b = 0.3465724	1941.427		1943.161		1939.52		1941.248	
	b = 0.4	1941.427		1943.161		1938.807		1941.167	
	b = 1	1941.427		1943.161		1937.391		1941.419	
	b = 2	1941.427		1943.161		1940.336		1941.424	
Population Total		1943.052							
Sine	0.3465724	0.01855193	1	0.03836453	4.286723	0.02072086	1.243534	0.01657321	0.7990398
	0.4	0.01855193	1	0.03836453	4.286723	0.02082649	1.255919	0.01685303	0.826246
	1	0.01855193	1	0.03836453	4.286723	0.0201947	1.183826	0.01810882	0.9576443
	2	0.01855193	1	0.03836453	4.286723	0.01895357	1.043951	0.0184607	0.9908383
Estimated Total	b = 0.3465724	4071.066		4114.031		4080.316		4056.493	
	b = 0.4	4071.066		4114.031		4081.685		4054.513	
	b = 1	4071.066		4114.031		4079.156		4066.007	
	b = 2	4071.066		4114.031		4073.04		4070.166	
Population Total		4071.383							
Bump	0.3465724	0.03109618	1	0.01449569	0.2130984	0.03243536	1.085912	0.03100986	0.9935966
	0.4	0.03109618	1	0.01449569	0.2130984	0.03289121	1.116063	0.03319303	1.123072
	1	0.03109618	1	0.01449569	0.2130984	0.03357809	1.165075	0.0321397	1.061732
	2	0.03109618	1	0.01449569	0.2130984	0.03165829	1.036739	0.03106365	0.9988702
Estimated Total	b = 0.3465724	2186.49		2192.769		2188.266		2172.2	
	b = 0.4	2186.49		2192.769		2195.394		2151.329	
	b = 1	2186.49		2192.769		2200.689		2161.91	
	b = 2	2186.49		2192.769		2189.318		2182.232	
Population Total		2187.923							
Jump	0.3465724	0.004845022	1	0.02483609	26.07389	0.005616896	1.353566	0.007676967	2.274792
	0.4	0.004845022	1	0.02483609	26.07389	0.0056205	1.35023	0.007750974	2.329744
	1	0.004845022	1	0.02483609	26.07389	0.005181882	1.155266	0.005505162	1.259671
	2	0.004845022	1	0.02483609	26.07389	0.004852543	1.006773	0.004872778	1.006966
Estimated Total	b = 0.3465724	3299.185		3321.699		3288.857		3322.128	
	b = 0.4	3299.185		3321.699		3288.415		3322.202	
	b = 1	3299.185		3321.699		3291.326		3309.116	
	b = 2	3299.185		3321.699		3297.485		3300.881	
Population Total		3300.252							

Table 4. Estimated lower and upper confidence limits and corresponding average lengths based on 1000 replications of simple random sampling within strata from four fixed populations of size $N = 2000$. Sample size is $n = 200$. (LCL is the Lower Confidence Limit, UCL is the Upper Confidence Limit and AL is the Average Length).

Population	*b*	$\hat{Y}_{HT}$			$\hat{Y}_{REG}$			$\hat{Y}_{PE}$			$\hat{Y}_{LP}$		
		LCL	UCL	AL	LCL	UCL	AL	LCL	UCL	AL	LCL	UCL	AL
Linear	0.3465724	1905.431	1977.423	71.992	1919.139	1967.183	48.044	1934.86	1944.18	9.32	1936.249	1946.247	9.998
	0.4	1905.431	1977.423	71.992	1919.139	1967.183	48.044	1934.250	1943.364	9.114	1936.169	1946.165	9.996
	1	1905.431	1977.423	71.992	1919.139	1967.183	48.044	1933.711	1941.071	7.360	1936.418	1946.420	10.002
	2	1905.431	1977.423	71.992	1919.139	1967.183	48.044	1936.733	1943.938	7.206	1936.424	1946.424	9.999
Population Total		1943.052											
Sine	0.3465724	4026.580	4115.552	88.973	4044.296	4183.766	139.470	4074.654	4085.978	11.324	4050.937	4062.049	11.113
	0.4	4026.580	4115.552	88.973	4044.296	4183.766	139.470	4076.156	4087.213	11.057	4049.014	4060.012	10.998
	1	4026.580	4115.552	88.973	4044.296	4183.766	139.470	4074.650	4083.661	9.012	4060.254	4071.760	11.506
	2	4026.580	4115.552	88.973	4044.296	4183.766	139.470	4068.589	4077.491	8.902	4064.498	4075.834	11.336
Population Total		4071.383											
Bump	0.3465724	2146.545	2226.434	79.889	2156.490	2229.048	72.558	2183.234	2193.299	10.065	2166.839	2177.560	10.721
	0.4	2146.545	2226.434	79.889	2156.490	2229.048	72.558	2190.473	2200.315	9.842	2145.980	2156.678	10.698
	1	2146.545	2226.434	79.889	2156.490	2229.048	72.558	2196.621	2204.758	8.137	2156.582	2167.238	10.656
	2	2146.545	2226.434	79.889	2156.490	2229.048	72.558	2185.320	2193.315	7.995	2176.909	2187.554	10.645
Population Total		2187.923											
Jump	0.3465724	3290.027	3308.344	18.317	3127.463	3515.934	388.471	3287.902	3289.813	1.912	3321.078	3323.179	2.101
	0.4	3290.027	3308.344	18.317	3127.463	3515.934	388.471	3287.47	3289.36	1.89	3321.172	3323.232	2.060
	1	3290.027	3308.344	18.317	3127.463	3515.934	388.471	3290.409	3292.244	1.835	3308.167	3310.065	1.898
	2	3290.027	3308.344	18.317	3127.463	3515.934	388.471	3296.569	3298.401	1.832	3299.932	3301.829	1.897
Population Total		3300.252											

Additionally, a keen look at the estimated totals in Table 3 shows that: as the bandwidth increases, the local linear regression estimator, $\hat{Y}_{LP}$ becomes equivalent to the linear regression estimator, $\hat{Y}_{REG}$. This shows that the bandwidth has an effect on the mean square error of $\hat{Y}_{LP}$. Particularly, for whichever bandwidth that is considered in this study, $\hat{Y}_{LP}$ essentially dominates $\hat{Y}_{REG}$ for all the populations except Linear and Bump populations, where $\hat{Y}_{REG}$ is competitive. Further, $\hat{Y}_{LP}$ essentially dominates $\hat{Y}_{HT}$ for all populations except in the Jump population, where $\hat{Y}_{HT}$ dominates all estimators being considered. The overall performance of $\hat{Y}_{LP}$ is consistently good as long as the bandwidth remains small in this particular study.

5. Conclusion

In this study, performance of the proposed estimator has been investigated against some design-based and model-based regression estimators. The *RE* values of the proposed estimator are in general close to one. It has been shown that for whichever bandwidth considered, $\hat{Y}_{LP}$ essentially dominates $\hat{Y}_{REG}$ for all the populations except

Linear and Bump populations, where $\hat{Y}_{REG}$ is competitive. Further, $\hat{Y}_{LP}$ essentially dominates $\hat{Y}_{HT}$ for all populations except in the Jump population, where it dominates all estimators being considered. Generally, good confidence intervals are seen for the nonparametric regression estimators, and use of the proposed estimator leads to relatively smaller values of *RE* compared to other estimators. We conclude that nonparametric regression approach under stratified random sampling using the proposed estimator yields good results.

Acknowledgements

Special thanks to the African Union (AU) for the funding that saw the success of this research.

References

[1] Cassel, C.M., Särndal, C.E. and Wretman, J.H. (1976) Some Results on Generalized Difference Estimation and Generalized Regression Estimation for Finite Populations. *Biometrika*, **63**, 615-620.

[2] Robinson, P.M. and Sarndal, C.E. (1983) Asymptotic Properties of the Generalized Regression Estimation in Probability Sampling. *The Indian Journal of Statistics*, *Series B*, **45**, 240-248.

[3] Cochran, W.G. (1977) Sampling Techniques. J. Wiley, New York.

[4] Royall, R.M. (1970) On Finite Population Sampling Theory under Certain Linear Regression Models. *Biometrika*, **57**, 377-387.

[5] Brewer, K.R.W. (1963) Ratio Estimation in Finite Populations: Some Results Deductible from the Assumption of an Underlying Stochastic Process. *Australian Journal of Statistics*, **5**, 93-105.

[6] Holt, D. and Smith, T.M. (1979) Post Stratification. *Journal of the Royal Statistical Society*, *Series A*, **142**, 33-46.

[7] Wu, C.B. and Sitter, R.R. (2001) A Model-Calibration Approach to Using Complete Auxiliary Information from Survey Data. *Journal of the American Statistical Association*, **96**, 185-193.

[8] Nadaraya, E.A. (1964) On Estimating Regression. *Theory of Probability and Applications*, **9**, 141-142.

[9] Watson, G.S. (1964) Smooth Regression Analysis. Sankhya, Series A, 359-372.

[10] Dorfman, A.H. (1992) Nonparametric Regression for Estimating Totals in Finite Population. In Section on Survey Research Methods. *Journal of American Statistical Association*, 622-625.

[11] Montanari, G.E. and Ranalli, M.G. (2003) Nonparametric Methods in Survey Sampling. In: Vinci, M., Monari, P., Mignani, S. and Montanari, A., Eds., *New Developments in Classification and Data Analysis*, Springer, Berlin, 203-210.

[12] Montanari, G.E. and Ranalli, M.G. (2005) Nonparametric Model Calibration Estimation in Survey Sampling. *Journal of the American Statistical Association*, **100**, 1429-1442. https://doi.org/10.1198/016214505000000141

[13] Sanchez-Borrego, I.R. and Rueda, M. (2009) A Predictive Estimator of Finite Population Mean Using Nonparametric Regression. *Computational Statistics*, **24**, 1-14.

https://doi.org/10.1007/s00180-008-0140-x

[14] Breidt, F.J. and Opsomer, J.D. (2000) Local Polynomial Regression Estimators in Survey Sampling. *The Annals of Statistics*, **28**, 1026-1053.

[15] Orwa, G.O., Otieno, R.O. and Mwita, P.N. (2010) Nonparametric Mixed Ratio Estimator for a Finite Population Total in Stratified Sampling. *Pakistan Journal of Statistics and Operation Research*, **4**, 21-35. https://doi.org/10.18187/pjsor.v6i1.149

[16] Ngesa, O.O., Orwa, G.O., Otieno, R.O. and Murray, H.M. (2012) Multivariate Ratio Estimator of the Population Total under Stratified Random Sampling. *Open Journal of Statistics*, **2**, 300-304. https://doi.org/10.4236/ojs.2012.23036

[17] Fan, J. and Gijbels, I. (1992) Variable Bandwidth and Local Linear Regression Smoothers. *The Annals of Statistics*, **20**, 2008-2036. https://doi.org/10.1214/aos/1176348900

[18] Fan, J. (1993) Local Linear Regression Smoothers and Their Minimax Efficiencies. *The Annals of Statistics*, **21**, 196-216. https://doi.org/10.1214/aos/1176349022

[19] Ruppert, D. and Wand, M.P. (1994) Multivariate Locally Weighted Least Squares Regression. *The Annals of Statistics*, **22**, 1346-1370. https://doi.org/10.1214/aos/1176325632

[20] Rady, E.-H.A. and Ziedan, D. (2014) Estimation of Population Total Using Local Polynomial Regression with Two Auxiliary Variables. *Journal of Statistics Applications & Probability*, **3**, 129-136. https://doi.org/10.12785/jsap/030203

[21] Fan, J. and Gijbels, I. (1996) Local Polynomial Modelling and Its Applications. Chapman and Hall, London.

[22] Horvitz, D.G. and Thompson, D.J. (1952) A Generalization of Sampling without Replacement from a Finite Universe. *Journal of American Statistical Association*, **47**, 663-685. https://doi.org/10.1080/01621459.1952.10483446