Pak. J. Statist.
2023 Vol. 39(4), 469-484

A MODEL BASED APPROACH TO THE ESTIMATION OF A FINITE POPULATION ERROR VARIANCE IN A HOMOSCEDASTIC SETTING

Winnie Mokeira Onsongo[1§], Vincent Odhiambo[2], Shaibu Osman[3] and **Kaku Sagary Nokoe[4]**

[1] Department of Statistics and Actuarial Science
University of Ghana, Legon, Ghana.
[2] Department of Mathematics and Actuarial Science,
The Catholic University of Eastern Africa, Nairobi, Kenya.
[3] Department of Basic Sciences, University of Health and
Allied Sciences, Ho, Ghana.
[4] Department of Economics and Business Administration,
Catholic University College of Ghana, Fiapre-Sunyani, Ghana.
[§] Corresponding Author Email: onsongowinnie@gmail.com

ABSTRACT

Difference-based nonparametric regression models are based on assumptions about the unknown nonparametric function and are appropriate for large sample problems. However, most of the difference-based estimators and residual-based estimators previously used do not balance between bias and variance, σ^2, which depends on the bandwidth, b, a phenomenon commonly referred to as bias-variance trade-off. As such, it is necessary to perform modification at boundary point as a measure to counter this drawback. Another drawback to these estimators is that they are generally biased due to the problem of boundary and therefore require modification at the boundary points. This study adopts a simple and explicit bias corrected estimator $\hat{\sigma}_{vo}^2$ of a finite population error variance in the setting where the variance is a constant (homoscedastic) using a model-based technique under simple random sampling without replacement (*SRSWOR*).

KEYWORDS

Difference-Based Estimators, Residual-Based Estimators, Bias Correction, Kernel Smoothing.

1. INTRODUCTION

Nonparametric regression models are widely used to describe nonlinear relationships between survey and auxiliary variables in order to avoid model misspecification that is common in design-based models. In particular, statisticians of sample survey seek to develop methods which improve on the asymptotic properties of error variance estimates. Cheng et al. (2018) developed a simple bias reduction approach for the estimation of the nonparametric regression model based on local linear regression without inflating the variance. In their simulation study, they investigated two bias-reduced estimation approaches and extended the methods to the error variance estimation problem. From the numerical results obtained, their proposed estimators improve in terms of efficiency, and

reduction of the estimation bias. Further, Opsomer et al. (2009) and Opsomer et al. (2012) described a new estimator of the design variance, under a nonparametric model for the population. The model is sufficiently flexible in surveys that considered continuous auxiliary variables observed at the population level. Their model proved to be consistent in estimating both the anticipated variance and the design variance under a non-parametric model with a univariate covariate. Another form of regression model is the ratio and regression estimators where sample units are chosen on the basis of an auxiliary variate with probability proportionate to some measure of scale. A model-based approach for the estimation of error variance can be studied in two settings namely homoscedastic and heteroscedastic settings.

A homoscedastic setting is where all data have the same error variance i.e., the variance is a constant. The homoscedastic nonparametric model is defined as:

$$Y_i = \boldsymbol{\varphi}(x_i) + \boldsymbol{\varepsilon}_i \quad (1)$$

for $i = 1,2,\ldots,n$

In which Y_i is the i^{th} response, $x_i = \frac{i}{n}$ is a univariate variable with $0 \leq x_i \leq 1$, ϕ is an unknown mean function and ε_i are independent and identically distributed random errors with zero mean and variance, σ^2. The heteroscedastic setting is where the variance is a function of x and it is not a constant instead it varies with design points. Define the heteroscedastic nonparametric model as:

$$Y_i = \boldsymbol{\varphi}(x_i) + \sqrt{V(x_i)}\boldsymbol{\varepsilon}_i \quad (2)$$

for $i = 1,2,\ldots,n$

where $\boldsymbol{\varepsilon}_i'$s are independent and identically distributed random variable with zero mean and a unit variance, and $V(x_i)$ is the variance function.

For difference-based estimators, $\sigma^2 = \frac{\boldsymbol{E}(Y_i - Y_{i-1})^2}{2}$, where Y_i and Y_{i-1} are independent with same means and variances. Rice (1984) developed a first order difference-based estimator.

$$\hat{\sigma}_R^2 = \frac{1}{2(n-1)}\sum_{i=2}^{n}(Y_i - Y_{i-1})^2 \quad (3)$$

which was later modified to a *lag-k* estimator $\hat{\sigma}_R^2(k)$ which is given by:

$$\hat{\sigma}_R^2(k) = \frac{1}{2(n-k)}\sum_{i=k+1}^{n}(Y_i - Y_{i-k})^2 \quad (4)$$

for $1 \leq k \leq n-1$.

Gasser et al. (1986) also developed the estimator

$$\hat{\sigma}_{GSJ}^2 = \frac{1}{6(n-2)}\sum_{i=3}^{n}(Y_i + Y_{i-2} - 2Y_{i-1})^2 \quad (5)$$

For design points which are equidistant, the estimator $\hat{\sigma}^2_{GSJ}$ reduces to

$$\hat{\sigma}^2_{GSJ} = \frac{2}{3(n-2)}\sum_{i=2}^{n-1}\left(\frac{1}{2}Y_{i-1} - Y_i + \frac{1}{2}Y_{i+1}\right)^2 \quad (6)$$

Hall and Marron (1990) estimated the $\phi(x_i)$ using a weighted average $\sum_{j=1}^{n} w_{ij}y_j$ where $w'_{ij}s$ are such that $\sum_{j=1}^{n} w_{ij}y_j = 1$ for each i. Their i^{th} residual is given by $\hat{e}_i = Y_i - \sum_{j=1}^{n} w_{ij}y_j$, $1 \le i \le n$. Their proposed residual-based error variance estimator for the homoscedastic setting is:

$$\hat{\sigma}^2_{HM} = \frac{\sum_{i=1}^{n}(Yi - \sum_{j=1}^{n} w_{ij}y_j)^2}{(n - 2\sum_{i=1}^{n} w_{ii} + \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}^2)} \quad (7)$$

where $w_{ij} = \frac{K(\frac{x_i - x_j}{b})}{\sum_{k=1}^{n} K(\frac{x_i - x_k}{b})}$ for $i \ge 1$ and $j \le n$, $K_b(.)$ is a kernel.

From the estimators discussed above, there is none that is efficient and appropriate for both small and large samples. They do not achieve the asymptotic optimal rate for the mean squared error. Moreover, they do not balance between bias and variance, σ^2, in that the bias decreases at the cost of increasing variance hence require modification at the boundary points as demonstrated by Buckley et al. (1988).

Although the bias-variance problem is always present in finite samples, it can be solved by use of smoothers whose asymptotic bias converges to zero while maintaining the same asymptotic variance.

2. METHODOLOGY

2.1 Proposed Estimator

In this work, we outline the procedure of estimating a robust estimator of error variance for a finite population of size $N < \infty$.

Suppose there is a data (x_i, y_i) where $1 \le i \le n$ which represents n values on the response, Y, corresponding to the n values of the independent variable X. Define the model:

$$y_i = \boldsymbol{\varphi}(x_i) + \boldsymbol{\varepsilon}_i \quad (8)$$

where $\phi(x_i)$ is a smooth function that is Lipschitz continuous and ε_i is the residual.

Suppose that equation (8) has a mean function

$$E(y_i) = \boldsymbol{\varphi}(x_i)$$

$$\boldsymbol{Cov}(\boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_j) = \begin{cases} \sigma^2(x_i), i = j \\ 0, otherwise \end{cases} \quad (9)$$

Consider n independent and identically distributed random variables $X_1, X_2, X_3, \ldots, X_n$ drawn from a population with mean μ and a finite variance, σ^2.

The finite population variance for this characteristic can be estimated as

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X}) \tag{10}$$

which is an unbiased estimator for σ^2.

Moreover

$$\frac{(n-1)\hat{\sigma}^2}{\sigma^2} \sim X^2_{(n-1)}$$

Implying that

$$Var\left(\frac{(n-1)\hat{\sigma}^2}{\sigma^2}\right) = 2(\mathrm{n}-1)$$

$$Var(\hat{\sigma}^2) = \frac{2}{(n-1)}\sigma^4 \tag{11}$$

Linton and Nielsen (1994) developed a multiplicative bias reduction method with the objective of correcting the bias-variance trade-off that mostly occurs in nonparametric regression. The technique has since been used by Burr et al. (2010) in smoothing low-resolution gamma spectra whilst Malenje et al. (2016) and Onsongo et al. (2018a, 2018b) also adopted the same methodology to smooth estimators of finite population parameters. In all cases, the technique has yielded robust estimators that are more efficient and sufficient estimators.

The bias reduction methodology is adopted in this work to estimate the error variance under the simple random sampling without replacement.

Assumptions

Assumptions used in this work are:

Assumption 1. The mean and variance are considered under a finite Fourth moment.

Assumption 2. The kernel function is smooth, bounded and twice differentiable.

Assumption 3. $x_i's$ are equispaced design points that are chosen at random from the interval [0,1]: $x_i = \frac{1}{n}$ for $1 \leq i \leq n$.

Assumption 4. The bandwidth $b \to 0$: $nb \to \infty$ as $n \to \infty$.

Define a rough estimator for the mean function in equation 8 as:

$$\boldsymbol{\varphi}\,(x_i) = \sum_{i=1}^{n} w_i(x) y_i \tag{12}$$

where $w_i(x)$ are kernel weights such that $\sum_{i\epsilon s} w_i(x) = 1$

Smoothing equation (12) yields a smooth function

$$\hat{\boldsymbol{\varphi}}(x) = \widetilde{\boldsymbol{\varphi}}(x)\bar{\beta}(x) \tag{13}$$

where $\bar{\beta}(x) = \sum_{i=1}^{n} w_{xi}\frac{y_i}{\boldsymbol{\varphi}\,(\boldsymbol{X_i})}$ is the correction factor.

First introduced by Linton and Nielsen (1994), the correction factor works by pulling the residuals in the numerator and denominator of $\frac{y_i}{\boldsymbol{\varphi}(X_i)}$ which results to a smoother function.

The residual-based estimator of the error variance is obtained from a regression fit for $\boldsymbol{\varphi}(x_i)$.

Define a new class of residuals such that:

$$\boldsymbol{\varepsilon}_i = y_i - \boldsymbol{\varphi}(x_i) \tag{14}$$

where $x_i's$ are equidistant design points. The error variance in the homoscedastic setting can be estimated as

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \boldsymbol{\varphi}_{-1}(x_i)\right) y_i \tag{15}$$

where $\boldsymbol{\varphi}_{-1}(x_i)$ is the estimate of $\boldsymbol{\varphi}(x_i)$ without considering the i^{th} observation. Also define the weighted average, w_{ij} as

$$w_{ij} = \frac{K\left(\frac{x_i - x_j}{b}\right)}{(n-1)bf(x_i)}$$

where $i \geq 1, j \leq n$, b refers to the bandwidth parameter and $k(.)$ is a kernel function that is bounded, symmetric around zero and compactly supported on (-1,1).

Then the new $\hat{\sigma}_{vo}^2$ estimator of error variance in a homoscedastic setting is:

$$\hat{\sigma}_{vo}^2 = \frac{1}{n}\sum y_i^2 - \frac{1}{fb}\sum_{i=1}^{n}\sum_{j\neq i} k\left(\frac{x_i - x_j}{b}\right) y_i y_j \tag{16}$$

where $f = n(n-1)$ and b is the bandwidth.

2.1.1 Asymptotic Bias of Error Variance Estimator

Define the model-based bias of $\hat{\sigma}_{vo}^2$ as;

$$\begin{aligned} Bias(\hat{\sigma}_{vo}^2) &= E[\hat{\sigma}_{vo}^2 - \sigma^2] \\ &= E[\hat{\sigma}_{vo}^2] - \sigma^2 \end{aligned} \tag{17}$$

Theorem 2.1:

$$E(\hat{\sigma}^2) = \sigma^2 + b^r A_1 + o(b^r) + O\left(\frac{1}{n}\right)$$

where

$$A_1 = \frac{(-1)^r}{r!}\int_0^1 k(y) y^r\, dy \int_0^1 \varphi(s)\varphi^{(r)}(s)\, ds$$

***Lemma* 1.**

Given the assumptions 1 to 3 above, this implies that

$$\frac{1}{n^2}\sum_{i\neq j}\sum k\left(\frac{s_i - s_j}{b}\right)\varphi(s_i)\varphi(s_j) = b\int_0^1\int_0^1 k(y)\varphi(s)\varphi(s) - b(y)dsdy + O\left(\frac{b}{n}\right)$$

From equation (17)

$$E(\hat{\sigma}_{vo}^2) = \frac{1}{n}\sum_{i=1}^{n} E(y_i^2) - \frac{1}{fb}\sum_{i=1}^{n}\sum_{j\neq i} k\left(\frac{x_i - x_j}{b}\right)E(y_i)E(y_j)$$

$$E(\hat{\sigma}_{vo}^2) = \frac{1}{n}\sum_{i=1}^{n}\{(\varphi(x_i))^2 + \sigma^2\} - \frac{1}{fb}\sum_{i=1}^{n}\sum_{j\neq i} k\left(\frac{x_i - x_j}{b}\right)\varphi(y_i)\varphi(y_j) \qquad (18)$$

Equation (18) is computed numerically as follows: Using the Riemann integral,

$$\frac{1}{n}\sum_{i=1}^{n}(\varphi(x_i))^2 = \int_0^1 \varphi^2(s)ds + O\left(\frac{1}{n}\right) \qquad (19)$$

Using lemma 1 and assumption 2

$$\frac{1}{fb}\sum_{i=1}^{n}\sum_{j\neq i} k\left(\frac{x_i - x_j}{b}\right)\varphi(x_i)\varphi(x_j) = \int_0^1\int_0^1 k(y)\varphi(s)\varphi(s - by)dsdy + O\left(\frac{1}{n}\right) \qquad (20)$$

Substituting equations (19) and (20) into equation (18) yields

$$E(\hat{\sigma}_{vo}^2) = \sigma^2 + \frac{b^r(-1)^r}{r!}\int_0^1 k(y)y^r\,dy\int_0^1 \varphi(s)\varphi^{(r)}(s)\,ds + o(b^r) + O\left(\frac{1}{n}\right) \qquad (21)$$

Since $A_1 = \frac{(-1)^r}{r!}\int_0^1 k(y)y^r\,dy\ \int_0^1 \varphi(s)\varphi^{(r)}(s)\,ds$, equation (21) reduces to

$$E(\hat{\sigma}_{vo}^2) = \sigma^2 + b^r A_1 + o(b^r) + O\left(\frac{1}{n}\right) \qquad (22)$$

Substituting equation (22) into (9)

$$Bias(\hat{\sigma}_{vo}^2) = b^r(A_1) + o(b^r) + O\left(\frac{1}{n}\right) \qquad (23)$$

By assumption 4, $Bias(\hat{\sigma}_{vo}^2) \to 0$

2.1.2 Asymptotic Variance of Error Variance Estimator

From theorem 2.1, the squared bias is given by

$$(E(\hat{\sigma}_{vo}^2) - \sigma^2)^2 = b^{2r}{A_1}^2 + o(b^{2r}) + o\left(\frac{1}{n^2 b}\right)$$

and the variance is given by

$$Var(\hat{\sigma}_{vo}^2) = E(\hat{\sigma}_{vo}^2)^2 - \left(E(\hat{\sigma}_{vo}^2)\right)^2 \tag{24}$$

Theorem 2.2:

$$Var(\hat{\sigma}^2) = \frac{1}{n}A_2 + \frac{1}{n^2 b}A_3 + o\left(\frac{1}{n^2 b}\right)$$

where

$$A_2 = \mu_4 - \sigma^4$$

$$A_2 = 2\sigma^4 \int_0^1 k^2(y)dy + 4\varphi^2 \int_0^1 k^2(y)dy \int_0^1 \varphi(x)dx$$

For proof of theorems 2.1 and 2.2 see Alharbi (2011).

$$\begin{aligned} Var(\hat{\sigma}^2) = {} & \frac{1}{n^2}\sum_{i=1}^{n}[\mu_4 - \sigma^4 + 4\mu_3\varphi(x_i) + 4\sigma^2\varphi^2(x_i)] \\ & - \frac{2}{fb}\sum_{i\neq j}\sum k\left(\frac{x_i - x_j}{b}\right)\left[\mu_3\varphi(x_j) + \mu_3\varphi(x_i) + 4\sigma^2\varphi(x_i)\varphi(x_j)\right] \\ & + \frac{2}{(fb)^2}\sum_{i\neq j}\sum k^2\left(\frac{x_i - x_j}{b}\right)\left[\sigma^4 + \sigma^2\varphi^2(x_i) + \sigma^2\varphi^2(x_j)\right] \\ & + \left(\frac{\sigma}{fb}\right)^2\sum\sum_{i\neq j\neq k}\sum k\left(\frac{x_i - x_j}{b}\right)k\left(\frac{x_i - x_k}{b}\right)\varphi(x_j)\varphi(x_k) \\ & + \left(\frac{\sigma}{fb}\right)^2\sum\sum_{i\neq j\neq k} k\left(\frac{x_i - x_j}{b}\right)k\left(\frac{x_k - x_i}{b}\right)\varphi(x_j)\varphi(x_k) \\ & + \left(\frac{\sigma}{fb}\right)^2\sum\sum_{i\neq k\neq d}\sum k\left(\frac{x_i - x_k}{b}\right)k\left(\frac{x_k - x_d}{b}\right)\varphi(x_i)\varphi(x_d) \\ & + \left(\frac{\sigma}{fb}\right)^2\sum\sum_{i\neq k\neq d}\sum k\left(\frac{x_i - x_j}{b}\right)k\left(\frac{x_k - x_j}{b}\right)\varphi(x_i)\varphi(x_k) \end{aligned} \tag{25}$$

where $\mu_4 = E(y_i - \varphi(x_i))^4$, $\mu_3 = E(y_i - \varphi(x_i))^3$ and $\sigma^4 = (E(y_i - \varphi(x_i))^2)^2$.
A mathematical computation and stochastic approximation of equation (25) yields

$$\begin{aligned} Var(\hat{\sigma}_{vo}^2) = {} & \frac{1}{n}[\mu_4 - \sigma^4] \\ & + \frac{1}{n^2 b}\left[2\sigma^4 \int_0^1 k^2(y)dy + 4\varphi^2 \int_0^1 k^2(y)dy \int_0^1 \varphi^2(x)dx\right] \\ & + o\left(\frac{1}{n^2 b}\right) \end{aligned} \tag{26}$$

Applying assumptions 1 and 4 in equation 26 implies $\lim_{n\to\infty} Var(\hat{\sigma}_{vo}^2) = 0$.
Consequently, the estimator $\hat{\sigma}_{vo}^2$ is consistent.

2.1.3 Simulated Data Analysis

The theory developed in the previous section is tested using simulated data. Simulation experiments were done to study the performance of the developed estimator. A population of 1,000 auxiliary values x_i are identified from natural data obtained from the uniform distribution on the interval [0,1].

The corresponding survey values y_i are generated using the super-population model of the form

$$y_i = \boldsymbol{\mu}(x_i) + \sigma(x_i)\boldsymbol{\varepsilon}_i$$

Two statistical models are used to simulate measurements of the survey variable Y

1. $y_i = 1 + 2(x_i - 0.5)^2 + \boldsymbol{\varepsilon}_i$
2. $y_i = exp(-4x_i) + \boldsymbol{\varepsilon}_i$

where $\boldsymbol{\varepsilon}_i \sim N(0,1)$ in the simulation exercise.

Statistical properties of the estimator were simulated based on samples of sizes $n = 250$, $n = 500$, $n = 700$ and $n = 900$. These simple random samples of sizes n are drawn without replacement from a population of size $N = 1{,}000$. Consequently, there is a total of $\binom{N}{n}$ possible samples that are to be realized from this population. All the samples, irrespective of the sample size n, are non-overlapping i.e. they are unique.

A comparative study was done based on estimators proposed by Rice (1984), Gasser et al. (1986) and Hall and Marron (1990) herein denoted as $\hat{\sigma}_R^2$, $\hat{\sigma}_{GSJ}^2$ and $\hat{\sigma}_{HM}^2$ respectively.

2.1.4 Conditional and Unconditional Properties

Unconditional properties such as unconditional biases, Relative Efficiency (RE), Relative Root Mean Error (RRME) and Standard Error (SE) are determined for all proposed estimators.

1,000 samples were randomly selected without replacement from a meta-population of $\binom{N}{n}$ samples and the sample realizations were used to compute the sample estimates. The parameter estimates were then obtained by averaging over the sample estimates.

The estimators of the error variance were computed as:

$$\hat{\sigma}_{vo}^2 = \frac{1}{1000}\sum_{i=1}^{1000}\hat{\sigma}_{vo_i}^2$$
$$\hat{\sigma}_{R}^2 = \frac{1}{1000}\sum_{i=1}^{1000}\hat{\sigma}_{R_i}^2$$
$$\hat{\sigma}_{GSJ}^2 = \frac{1}{1000}\sum_{i=1}^{1000}\hat{\sigma}_{GSJ_i}^2 \quad (27)$$
$$\hat{\sigma}_{HM}^2 = \frac{1}{1000}\sum_{i=1}^{1000}\hat{\sigma}_{HM_i}^2$$

The estimated bias was computed as

$$Bias(\hat{\sigma}_{vo}^2) = \frac{1}{1000}\sum_{i=1}^{1000}\left(\hat{\sigma}_{vo_i}^2 - \sigma_e^2\right)$$
$$Bias(\hat{\sigma}_{R}^2) = \frac{1}{1000}\sum_{i=1}^{1000}\left(\hat{\sigma}_{R_i}^2 - \sigma_e^2\right)$$
$$Bias\left(\hat{\sigma}_{GSJ}^2\right) = \frac{1}{1000}\sum_{i=1}^{1000}\left(\hat{\sigma}_{GSJ_i}^2 - \sigma_e^2\right) \quad (28)$$
$$Bias(\hat{\sigma}_{HM}^2) = \frac{1}{1000}\sum_{i=1}^{1000}\left(\hat{\sigma}_{HM_i}^2 - \sigma_e^2\right)$$

where σ_e^2 is the true error variance. The relative root mean error was computed using the algorithms.

$$RRME(\hat{\sigma}_{vo}^2) = \sqrt{\frac{1}{1000}\sum_{i=1}^{1000}\left(\hat{\sigma}_{vo_i}^2 - \sigma_e^2\right)}$$
$$RRME(\hat{\sigma}_{R}^2) = \sqrt{\frac{1}{1000}\sum_{i=1}^{1000}\left(\hat{\sigma}_{R_i}^2 - \sigma_e^2\right)}$$
$$RME\left(\hat{\sigma}_{GSJ}^2\right) = \sqrt{\frac{1}{1000}\sum_{i=1}^{1000}\left(\hat{\sigma}_{GSJ_i}^2 - \sigma_e^2\right)} \quad (29)$$
$$RRME(\hat{\sigma}_{HM}^2) = \sqrt{\frac{1}{1000}\sum_{i=1}^{1000}\left(\hat{\sigma}_{HM_i}^2 - \sigma_e^2\right)}$$

The standard errors of the estimators were computed as:

$$SE(\hat{\sigma}_{vo}^2) = \sqrt{\frac{1}{1000}\sum_{i=1}^{1000}\left(\hat{\sigma}_{vo_i}^2 - \sigma_e^2\right)^2}$$
$$SE(\hat{\sigma}_R^2) = \sqrt{\frac{1}{1000}\sum_{i=1}^{1000}\left(\hat{\sigma}_{R_i}^2 - \sigma_e^2\right)^2}$$
$$SE\left(\hat{\sigma}_{GSJ}^2\right) = \sqrt{\frac{1}{1000}\sum_{i=1}^{1000}\left(\hat{\sigma}_{GSJ_i}^2 - \sigma_e^2\right)^2} \quad (30)$$
$$SE(\hat{\sigma}_{HM}^2) = \sqrt{\frac{1}{1000}\sum_{i=1}^{1000}\left(\hat{\sigma}_{HM_i}^2 - \sigma_e^2\right)^2}$$

The efficiency of the estimator $\hat{\sigma}_{vo}^2$ relative to the other estimators $\hat{\sigma}_{HM}^2$, $\hat{\sigma}_{GSJ}^2$ and $\hat{\sigma}_R^2$ is computed as:

$$RE(\hat{\sigma}_{vo}^2) = \frac{\sum_{i=1}^{1000}\left(\hat{\sigma}_{vo_i}^2 - \sigma_e^2\right)^2}{\sum_{i=1}^{1000}\left(\hat{\sigma}_{R_i}^2 - \sigma_e^2\right)^2}$$
$$RE(\hat{\sigma}_{vo}^2) = \frac{\sum_{i=1}^{1000}\left(\hat{\sigma}_{vo_i}^2 - \sigma_e^2\right)^2}{\sum_{i=1}^{1000}\left(\hat{\sigma}_{GSJ_i}^2 - \sigma_e^2\right)^2} \quad (31)$$
$$RE(\hat{\sigma}_{vo}^2) = \frac{\sum_{i=1}^{1000}\left(\hat{\sigma}_{vo_i}^2 - \sigma_e^2\right)^2}{\sum_{i=1}^{1000}\left(\hat{\sigma}_{HM_i}^2 - \sigma_e^2\right)^2}$$

Results obtained from the simulation are presented in Tables 1 and 2.

Table 1
Unconditional Properties of the Estimators using a Quadratic Mean Function

Sample Size (n)	Estimator	Estimate	Bias	RRME	SE	RE
$n = 250$	$\hat{\sigma}^2_{vo}$	1.3509	0.3509	0.5924	0.3747	0.3636
	$\hat{\sigma}^2_{R}$	1.1150	0.1150	0.3391	0.2756	0.6718
	$\hat{\sigma}^2_{HM}$	2.1559	1.1559	1.0751	1.1654	0.0376
	$\hat{\sigma}^2_{GSJ}$	6.9217	5.9217	2.4335	6.1885	0.0013
$n = 500$	$\hat{\sigma}^2_{vo}$	1.1802	0.1802	0.4245	0.1911	0.2017
	$\hat{\sigma}^2_{R}$	1.1719	0.1719	0.4146	0.2304	0.1389
	$\hat{\sigma}^2_{HM}$	2.4757	1.4757	1.2148	1.4784	0.0034
	$\hat{\sigma}^2_{GSJ}$	6.2683	5.2683	2.2953	5.3589	0.0003
$n = 700$	$\hat{\sigma}^2_{vo}$	1.0611	0.0611	0.2471	0.0809	0.9846
	$\hat{\sigma}^2_{R}$	1.1622	0.1622	0.04027	0.2006	0.1603
	$\hat{\sigma}^2_{HM}$	2.6346	1.6346	1.2785	1.6372	0.0024
	$\hat{\sigma}^2_{GSJ}$	6.2310	5.2310	2.2871	5.2810	0.0002
$n = 900$	$\hat{\sigma}^2_{vo}$	1.1732	0.1732	0.4161	0.1733	0.0372
	$\hat{\sigma}^2_{R}$	1.1521	0.1521	0.3900	0.1878	0.0332
	$\hat{\sigma}^2_{HM}$	3.0684	2.0684	1.4382	2.0684	0.0003
	$\hat{\sigma}^2_{GSJ}$	7.3155	6.3155	2.5131	6.3659	0.00003

The negative and positive values of the bias in Tables 1 and 2 imply underestimation and overestimation respectively.

Using a quadratic mean function, the estimator $\hat{\sigma}^2_{vo}$ has a smaller bias, RRME, SE and RE compared to the estimators $\hat{\sigma}^2_{HM}$ and $\hat{\sigma}^2_{GSJ}$ for all sample sizes chosen except in an exponential model where they have a better RE.

Generally, the proposed $\hat{\sigma}^2_{vo}$ estimator performs better, followed closely by $\hat{\sigma}^2_{R}$ estimator. $\hat{\sigma}^2_{HM}$ and $\hat{\sigma}^2_{GSJ}$ proved to be inefficient estimators in all models and at all sample sizes chosen.

Table 2
Unconditional Properties of the Estimators using an Exponential Mean Function

Sample Size (n)	Estimator	Estimate	Bias	RRME	SE	RE
$n = 250$	$\hat{\sigma}^2_{vo}$	1.4859	0.4859	0.6971	0.6478	0.9918
	$\hat{\sigma}^2_{R}$	1.1099	0.1099	0.3315	0.2828	5.2027
	$\hat{\sigma}^2_{HM}$	2.2942	1.2942	1.1376	1.4391	0.2010
	$\hat{\sigma}^2_{GSJ}$	5.9378	4.9378	2.2221	5.4459	0.0140
$n = 500$	$\hat{\sigma}^2_{vo}$	1.7643	0.7643	0.8742	0.8740	22.7668
	$\hat{\sigma}^2_{R}$	0.9307	-0.0693	0.2633	0.1903	480.3945
	$\hat{\sigma}^2_{HM}$	2.4371	1.4371	1.1988	1.5334	7.3955
	$\hat{\sigma}^2_{GSJ}$	6.0612	5.0612	2.2497	5.3579	0.6058
$n = 700$	$\hat{\sigma}^2_{vo}$	1.3322	0.3322	0.5764	0.3914	8.0229
	$\hat{\sigma}^2_{R}$	0.8228	-0.1772	0.4210	0.2055	29.1018
	$\hat{\sigma}^2_{HM}$	1.6338	0.6338	0.7961	0.6813	2.6479
	$\hat{\sigma}^2_{GSJ}$	3.7873	2.7873	1.6695	2.8572	0.1505
$n = 900$	$\hat{\sigma}^2_{vo}$	1.3139	0.3139	0.5602	0.3568	19.3795
	$\hat{\sigma}^2_{R}$	0.8141	-0.1859	0.4311	0.2037	59.4349
	$\hat{\sigma}^2_{HM}$	1.5938	0.5938	0.7706	0.6205	6.4079
	$\hat{\sigma}^2_{GSJ}$	3.6936	2.6936	1.6412	2.7214	0.3331

The conditional performance of the estimator $\hat{\sigma}^2_{vo}$ was done and was compared with the performance of the other estimators. To do this, 1,000 random samples, all of size 250, were selected and the mean of the auxiliary values x_i was computed for each sample. These sample means were then sorted in ascending order and further grouped into clusters of size 25. The mean of means i.e. $\bar{\bar{X}}$ is then computed as $\bar{\bar{X}} = \frac{1}{40} = \sum_{i=1}^{40} \bar{X}_i$.

Conditional means and biases were then computed for the estimators $\hat{\sigma}^2_{vo}$, $\hat{\sigma}^2_{R}$, $\hat{\sigma}^2_{HM}$ and $\hat{\sigma}^2_{GSJ}$. The results obtained from this simulation exercise are presented in tables 3 and 4.

A comparison between the unconditional and conditional properties given sample sizes $n = 250, 500$ and 900, the developed estimator $\hat{\sigma}^2_{vo}$ performs better than $\hat{\sigma}^2_{HM}$ and $\hat{\sigma}^2_{GSJ}$ since it has relatively smaller bias, RRME and SE.

The Rice estimator, $\hat{\sigma}^2_{R}$, has slightly performed better than our estimator and this can be attributed to the fact that it took into consideration the distance between the variables in its

simulation study. Nonetheless, the developed estimator $\hat{\sigma}_{vo}^2$ aimed at striking a balance between the bias-variance trade off which the $\hat{\sigma}_R^2$ estimator did not consider.

For $n = 700$, our estimator has the smallest bias and SE hence performs better than these other estimators.

Table 3
Conditional Properties of the Estimators using a Quadratic Mean Function

Sample Size (n)	Estimator	Bias	RRME	SE	RE
$n = 250$	$\hat{\sigma}_{vo}^2$	0.3509	0.3596	0.3653	1.1408
	$\hat{\sigma}_R^2$	0.1150	0.2098	0.2105	5.6836
	$\hat{\sigma}_{HM}^2$	1.1559	1.1559	1.1613	0.3585
	$\hat{\sigma}_{GSJ}^2$	5.9217	5.9217	5.9233	0.0751
$n = 500$	$\hat{\sigma}_{vo}^2$	0.1802	0.1802	0.1878	1.0224
	$\hat{\sigma}_R^2$	0.1719	0.1719	0.1891	2.4970
	$\hat{\sigma}_{HM}^2$	1.4757	1.4757	1.4778	0.1301
	$\hat{\sigma}_{GSJ}^2$	5.2683	5.2683	5.2689	0.0376
$n = 700$	$\hat{\sigma}_{vo}^2$	0.0611	0.0671	0.07171	3.6518
	$\hat{\sigma}_R^2$	0.1622	0.1722	0.1726	2.7380
	$\hat{\sigma}_{HM}^2$	1.6346	1.6346	1.6365	0.1463
	$\hat{\sigma}_{GSJ}^2$	5.2310	5.2310	5.2313	0.0463
$n = 900$	$\hat{\sigma}_{vo}^2$	0.1732	0.1732	0.1733	2.6424
	$\hat{\sigma}_R^2$	0.1521	0.1521	0.1560	5.8642
	$\hat{\sigma}_{HM}^2$	2.0684	2.0684	2.0684	0.2214
	$\hat{\sigma}_{GSJ}^2$	6.3155	6.3155	6.3157	0.0736

Table 4
Conditional Properties of the Estimators using an Exponential Mean Function

Sample Size (n)	Estimator	Bias	RRME	SE	RE
$n = 250$	$\hat{\sigma}^2_{vo}$	0.4849	0.5360	0.53926	1.1106
	$\hat{\sigma}^2_{R}$	0.1099	0.2174	0.2182	6.1669
	$\hat{\sigma}^2_{HM}$	1.2942	1.2978	1.2999	0.5269
	$\hat{\sigma}^2_{GSJ}$	4.9378	4.9378	4.9417	0.1217
$n = 500$	$\hat{\sigma}^2_{vo}$	0.7643	0.7724	0.7749	6.4346
	$\hat{\sigma}^2_{R}$	-0.0693	0.1576	0.1581	14.5912
	$\hat{\sigma}^2_{HM}$	1.4371	1.4371	1.4382	2.4715
	$\hat{\sigma}^2_{GSJ}$	5.0612	5.0612	5.0633	0.4973
$n = 700$	$\hat{\sigma}^2_{vo}$	0.3322	0.3402	0.3412	1.7073
	$\hat{\sigma}^2_{R}$	-0.1772	0.1831	0.1834	7.7139
	$\hat{\sigma}^2_{HM}$	0.6338	0.6338	0.6343	0.9154
	$\hat{\sigma}^2_{GSJ}$	2.7873	2.7873	2.7878	0.1877
$n = 900$	$\hat{\sigma}^2_{vo}$	0.3139	0.3173	0.3184	2.5588
	$\hat{\sigma}^2_{R}$	-0.1859	0.1864	0.1866	18.9647
	$\hat{\sigma}^2_{HM}$	0.5938	0.5938	0.5940	1.0812
	$\hat{\sigma}^2_{GSJ}$	2.6936	2.6936	2.6937	0.2304

From the Tables 3 and 4 the estimator $\hat{\sigma}^2_{vo}$ has a smaller conditional bias, a smaller conditional RRME, and a smaller conditional SE than $\hat{\sigma}^2_{HM}$ and $\hat{\sigma}^2_{GSJ}$ in all models and all sample sizes chosen. The performance of the estimator $\hat{\sigma}^2_{vo}$ under conditional properties is consistent with the performance under unconditional properties. Tables 2 and 4 show the performance of the estimators for unconditional and conditional properties respectively using an exponential mean function.

At $n = 250$, the developed estimator $\hat{\sigma}^2_{vo}$ has a smaller bias than $\hat{\sigma}^2_{HM}$ and $\hat{\sigma}^2_{GSJ}$.

For sample sizes $n = 500, 700$ and 900, the $\hat{\sigma}^2_{vo}$ has a smaller bias than all other estimators.

The Rice estimator $\hat{\sigma}^2_{R}$, underestimates since it has a negative bias at the above sample sizes.

3. CONCLUSION

In this article, we have proposed a model-based estimator for the error variance in a finite population with the objective of improving the efficiency and precision in relation to bias-variance trade-off. The objective of this work was to develop a robust estimator of error variance for a finite population in a homoscedastic setting under simple random sampling without replacement (SRSWOR) using a model-based approach.

Kernel smoothers were utilized as tools for developing the estimator $\hat{\sigma}_{vo}^2$ to tackle the problem associated with bias-variance trade off at the boundary points. Theoretical properties were derived and a Monte Carlo simulation study was done to compare the performance of the developed estimator to that of other existing estimators. The simulation results confirm that the proposed estimator $\hat{\sigma}_{vo}^2$ is statistically consistent in estimating the error variance.

ACKNOWLEDGMENT

Authors profoundly acknowledged the encouragement and support from members of department. Our appreciation goes to the numerous review comments and suggestions. Authors have expressed their gratitude for such a wonderful support and comments.

Source of Funding:

There are no sources of funding for this research. Authors are solely responsible for the entire cost of this research.

Conflict of Interest:

Authors declare that there is no conflict of interest regarding the publication of this paper.

REFERENCES

1. Alharbi, Y.F. (2011). *Error variance estimation in nonparametric regression models*. Doctoral dissertation, University of Birmingham.
2. Buckley, M.J., Eagleson, G.K. and Silverman, B.W. (1988). The estimation of residual variance in nonparametric regression. *Biometrika*, 75(2), 189-199.
3. Burr, T., Hengartner, N., Matzner-Lober, E., Myers, S. and Rouviere, L. (2010). Smoothing low resolution gamma spectra. *IEEE Transactions on Nuclear Science*, 57(5), 2831-2840.
4. Cheng, M.Y., Huang, T., Liu, P. and Peng, H. (2018). Bias reduction for nonparametric and semiparametric regression models. *Statistica Sinica*, 28(4), 2749-2770.
5. Gasser, T., Sroka, L. and Jennen-Steinmetz, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika*, 73(3), 625-633.
6. Hall, P. and Marron, J.S. (1990). On variance estimation in nonparametric regression. *Biometrika*, 77(2), 415-419.
7. Hall, P., Kay, J.W. and Titterington, D.M. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika*, 77(3), 521-528.
8. Linton, O. and Nielsen, J.P. (1994). A multiplicative bias reduction method for nonparametric regression. *Statistics & Probability Letters*, 19(3), 181-187.

9. Malenje, B.M., Mokeira, W.O., Odhiambo, R. and Orwa, G.O. (2016). A multiplicative bias corrected nonparametric estimator for a finite population mean. *American Journal of Theoretical and Applied Statistics*, 5(5), 317-325.
10. Onsongo, W.M., Otieno, R.O. and Orwa, G.O. (2018a). Bias Reduction Technique for Estimating Finite Population Distribution Function under Simple Random Sampling without Replacement. *International Journal of Statistics and Applications*, 8(5), 259-266.
11. Onsongo, W.M., Otieno, R.O. and Orwa, G.O. (2018b). Nonparametric estimation of distribution function for stratified populations. *International Journal of Probability and Statistics*, 7(5), 125-129.
12. Opsomer, J.D., Francisco-Fernández, M. and Li, X. (2009). Model-based nonparametric variance estimation for systematic sampling in a forestry survey. http://dm.udc.es/profesores/mario/ficheros/Nonpara_model_varianceFinal1.pdf
13. Opsomer, J.D., Francisco-Fernández, M. and Li, X. (2012). Model-based non-parametric variance estimation for systematic sampling. *Scandinavian Journal of Statistics*, 39(3), 528-542.
14. Rice, J. (1984). Bandwidth choice for nonparametric regression. *The Annals of Statistics*, 12(4), 1215-1230.