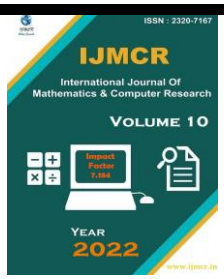
International Journal of Mathematics and Computer Research

ISSN: 2320-7167

Volume 10 Issue 10 October 2022, Page no. – 2925-2933

Index Copernicus ICV: 57.55, Impact Factor: 7.362

DOI: 10.47191/ijmcr/v10i10.02



Nonparametric Estimation of Error Variance under Simple Random Sampling without Replacement

Odhiambo Vincent[1], Hellen Waititu[2], Nyakundi Omwando Cornelious[3]

[1,2,3] Department of Mathematics and Acturial Science, The Catholic University of Eastern Africa, Nairobi- Kenya.

ARTICLE INFO

Published Online:
20 October 2022

Corresponding Author:
Odhiambo Vincent

ABSTRACT

This study adopts a nonparametric approach in the estimation 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). A mean square analysis of the $\hat{\sigma}_V^2$ estimator has been conducted, including the asymptotic behaviour of the $\hat{\sigma}_V^2$ estimator and the results show that the asymptotic distribution in a homoscedastic setting is asymptotically unbiased and consistent. The performance of the developed estimator is compared to that of other existing estimators using real data. R statistical software was utilized to analyze data and numerical results presented graphically for selected models.

KEYWORDS: Nonparametric estimation, kernel function, error variance.

1 INTRODUCTION

In sample survey, researchers are always interested in coming up with methods which improves on the asymptotic properties of population parameter estimates. This has resulted in the development of nonparametric regression models. One 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 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

$$Y_i = \psi(x_i) + e_i, i = 1,2,3,\ldots\ldots\ldots\ldots n$$

In which Y_i is the i[th] response, x_i is a univariate variable with $0 \leq x_i \leq 1$, ψ is an unknown mean function and e_i are independent and identically distributed random errors with zero mean and variance , σ^2.

In this context, assume that $x_i = i/n$ for $1 \leq i \leq n$ without loss of generality.

Estimation methods for error variance that have been proposed in the past include: kernel-based estimators, spline estimators, difference-based estimators, non-negative estimators that are unbiased in the case of a linear function, design adaptive regression and orthogonal series methods.

For difference-based estimators $\sigma^2 = E\frac{(Y_i - Y_{i-1})^2}{2}$ where Y_i and Y_{i-1} are independent with same means and variances. This was later developed by (Rice, 1986), to a first order difference based estimator. (Rice, 1986) later developed it further to the lag-Rice estimator $\hat{\sigma}_R^2(k)$ and expectation of lag-k estimator.

The results obtained by use of Rice estimator showed that, as opposed to other estimators, it obtained high uniform consistency. However, $\hat{\sigma}_R^2$ estimator underestimates bias as the sample size increases and at an Exponential model. (Gasser et. al, 1986) also developed the $\hat{\sigma}_{GSJ}^2$estimator for design points which are equidistant and is the sum of squares of second order differences. The simulation study concluded that the $\hat{\sigma}_{GSJ}^2$ estimator was sufficient in terms of Relative Efficiency. However, in terms of conditional and unconditional bias, Relative Root Mean Error (RRME), and Standard Error (SE), the $\hat{\sigma}_{GSJ}^2$ estimator has a low performance.

(Hall et. al, 1990) proved that $\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{r}(x_i)\right)^2$ with $\hat{r}$ being a kernel estimator which has asymptotic variance as$T^2 = \int x^4 dF(x) - \sigma^4$. They estimated the $\psi(x_i)$ using a weighted average $\sum_{j=1}^{n} w_{ij} Y_j$ where $w_{ij}s$ are such that $\sum_{j=1}^{n} w_{ij} = 1$ for each$i$.

They defined their i^{th} residual as

$$\hat{e}_i = Y_i - \sum_{j=1}^{n} w_{ij}\, Y_j, 1 \leq i \leq n \qquad (1)$$

(Hall et. al, 1990) then used the above i^{th}residual to propose the residual-based error variance estimator for the

homoscedastic setting .Using the r^{th}order kernel in estimating the mean function, (Hall et. al, 1990) obtained the mean squared error (MSE) of their proposed estimator. Though having a large bias, RRME, and SE; Simulation experiments showed that the $\hat{\sigma}^2_{HM}$estimator is robust and is of an acceptable accuracy. (Muller et. al, 2003) asserted that there exist an estimator for nonparametric regression models with random covariates for error variance that has a relationship with difference-based estimators.

Most of the estimators for error variance, σ^2, are in quadratic form and they are usually grouped into two classes. The first class of estimators are based on error sum of squares from some nonparametric fit to r and the estimation of $r$is done by either kernel smoothing or spline smoothing. Wahba [11] therefore proposed a residual -based estimator. (Dorfman, 1992), came up with a nonparametric regression estimator for a finite population variance which uses a sample drawn from the population and for a model based estimator, (Dorfman, 1992) developed conditional variance estimator. As observed by (Dette et.al, 1998) none of the above difference-based estimators achieve the asymptotic optimal rate for the mean squared error (MSE).

(Zheng et al, 2003), introduced the spline estimator of the population error variance. (Tong et. al, 2005) estimated error variance using a least square approach where they considered the error variance as the intercept in a simple linear regression which was obtained from the expectation of the lag-k Rice estimator. (Miya et.al, 2016) noted that the drawback with these estimators is that they fail to balance between bias and variance in that when the bandwidth,h is large the bias is also large and if h is small the variance also increases. 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. The proposed estimator in this work therefore seeks to address the shortcomings of the existing estimators. An error variance $\hat{\sigma}^2_V$ estimator which is robust in a homoscedastic setting under simple random sampling without replacement was developed. In addition to the asymptotic behaviour of the model, the $\hat{\sigma}^2_V$ estimator is unbiased by the fact that its asymptotic bias converges to zero, simple, robust, has a smaller asymptotic error variance and minimizes cost effectively.

Outline of the paper: In section 2, the proposed estimator for a finite population error variance is developed. Asymptotic properties of the developed $\hat{\sigma}^2_V$estimator are derived in section 3. In section 4, Emperical study of the results are presented and the conclusion and recommendations for future study and practitioners are given in section 5.

2 PROPOSED ESTIMATOR

Define a statistical model of the form

$$Y_i = \psi(x_i) + e_i \text{ for } i = 1,2,3\ldots\ldots\ldots,n$$

with

$$E(y) = \psi(y)(x_i)$$

$$cov(Y_i, Y_j) = \begin{cases} \sigma^2(x_i), & i = j \\ 0, & elsewhere \end{cases}$$

Where $\psi(x_i)$ is the mean function $E(Y_i/x_i)$, Y_i represents the survey variable, x_i s represents the design points and $e_i = Y_i - \psi(X_i)$ are independent and identically distributed random variable with mean zero and variance, σ^2, and the fourth moment is bounded such that $\left(E(e^4)\right) < \infty$.

In order to estimate the mean function, define

$$\bar{\psi}(x) = \sum_{i=1}^{n} w_{ij}\, Y_i, for 1 \le i \le n \quad (2)$$

but $w_{ij} = \frac{1}{h} k\left(\frac{x_i - x_j}{h}\right)$ are kernel weights and $k(.)$ is a kernel density function that is symmetric about zero, with bounded support

Putting w_{ij}into (2) we obtain

$$\bar{\psi}(x) = \sum_{i=1}^{n} \frac{1}{h} k\left(\frac{x_i - x_j}{h}\right) Y_i \quad (3)$$

which is a rough estimator of mean function in (2)

For $\bar{\psi}(x) - \psi(x)$, we obtain

$$\bar{\psi}(x) - \psi(x) = \left[\sum_{i=1}^{n} w_{xj} \,.\, \psi(X_i) - \psi(x)\right] + \sum_{i=1}^{n} w_{xj}\, e_i \quad (4)$$

Making $\bar{\psi}(x)$ the subject of the formular in equation (5), we get

$$\bar{\psi}(x) = \left[\sum W_{xj}\, \psi(x_i) - \psi(x)\right] + \sum_{i=1}^{n} W_{xj}\, e_i + \psi(x) \quad (5)$$

The estimator for the mean function is therefore

$$\hat{\psi}(x) = \bar{\psi}(x)\bar{\zeta}(x) = \left\{\left[\sum_{i=1}^{n} W_{xj}\, \psi(x_i)\right] + \sum_{i=1}^{n} W_{xj}\, e_i\right\} \sum_{i=1}^{n} W_{xj} \frac{Y_i}{\psi(X_i)} \quad (6)$$

Where $\bar{\zeta}(x) = \sum_{i=1}^{n} w_{xj} \frac{Y_i}{\psi(X_i)}$ and $\hat{e}_i = Y_i - \sum_{i=1}^{n} W_{xj} Y_i$ for $1 \le j \le n$

Motivated by Alharbi (2011), we propose a new class of error variance such that $e_i = Y_i - \psi(x_i)$ where $x_i s$ equidistant design points.

2.1 *Assumptions of the study*

1. The mean and variance is considered under a finite Fourth moment.
2. The kernel function is smooth, bounded and twice differentiable.
3. For a bandwidth $h \to 0$: $nh \to \infty$ as $n \to \infty$

Using assumptions 1-3 above, the error variance can be estimated as

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \hat{\psi}_{-i}(x_i)\right) Y_i \quad (7)$$

Where $\hat{\psi}_{-i}(x_i)$ is the estimate of $\psi(x_i)$ without considering the $i - th$ observations

In order to estimate$\psi(x_i)$, we define the weighted average, w_{ij} as $w_{ij} = \frac{K\left(\frac{x_i - x_j}{h}\right)}{(n-1)gh(x_i)}$ where$i \ge 1, j \le n$.

Since the weight function satisfies the constraint $\sum_{j \ne i}^{n} w_{ij} = 1$ for each i, then the $\psi(x)$ is estimated by $\hat{\psi}_{-1}(x_i) = \sum_{j \ne i} w_{ij} Y_i$ which does not consider the i^{th} observation of $\psi(x_i)$ at point x_i

From (9), (11) and equation for w_{ij}, the new $\hat{\sigma}_V^2$ estimator of error variance in a homoscedastic setting is:

$$\hat{\sigma}_V^2 = \frac{1}{n}\sum Y_i^2 - \frac{1}{gh}\sum_{i=1}^{n}\sum_{j\neq i} k\left(\frac{x_i - x_j}{h}\right) Y_i Y_j$$

(8)

Where $g = n(n-1)$ and h is the bandwidth

3 *Properties of the estimator*

3.1 *Bias of the estimator*

Define the bias of the developed estimator as

$E[\hat{\sigma}^2{}_V - \sigma] = E[\sigma_V^2] - E[\sigma]$

with

$$E(\sigma_V\ \ ) = \frac{1}{n}\sum_{i=1}^{n}\sum_{i=1}^{n}\left\{\left(\psi(x_i)\right)^2 + \sigma^2\right\} - \frac{1}{gh}\sum_{i=1}\sum_{j\neq i} k\left(\frac{x_i - x_j}{h}\right)\psi(x_i)\psi(x_j)$$

This yields

$$E(\hat{\sigma}_V^2) = \hat{\sigma}_V^2 + \frac{1}{n}\sum_{i=1}^{n}\left(\psi(x_i)\right)^2 - \frac{1}{gh}\sum_{i=1}^{n}\sum_{j\neq i} k\left(\frac{x_i - x_j}{h}\right)\psi(x_i)\psi(x_j) \qquad (9)$$

Mathematical analysis of equation (9) yields

$$E(\hat{\sigma}_V^2) = \hat{\sigma}_V^2 + \int_0^1 \psi^2(s)ds - \int_0^1 \psi^2(s)ds + \frac{t^r(-1)^r}{r!}\int_0^1 k(y)y^r dy \int_0^1 \psi(s)\psi^r(s)ds + o(t^r) + O\left(\frac{1}{n}\right)$$

(10)

Since the kernel function is bounded, then equation (10) converges such that $E(\hat{\sigma}_V^2) = \hat{\sigma}_V^2$

This confirms that the $\hat{\sigma}_V^2$ estimator is a true parameter and is unbiased.

3.2 *Consistency of the estimator*

Define the mean square error of the estimator as $MSE\left(\hat{\sigma}_V^2\right) = E\left(\hat{\sigma}_V^2 - \sigma^2\right)^2$ then

$$MSE(\hat{\sigma}_V^2) = Bias(\hat{\sigma}_V^2) + Variance(\hat{\sigma}_V^2) \qquad (11)$$

3.0.1 Theorems

1. $E\left(\hat{\sigma}_V^2\right) = \sigma^2 + h^r\frac{(-1)^r}{r!}\int_0^1 k(y)y^r dy \int_0^1 \psi(s)\psi^{(r)}(s)ds + o(h^r) + O\left(\frac{1}{n}\right)$.
2. $Var\left(\hat{\sigma}_V^2\right) = \frac{1}{n}\frac{(-1)^r}{r!}\int_0^1 k(y)y^r dy \int_0^1 \psi(s)\psi^{(r)}(s)ds + \frac{1}{n^2h}2\sigma^4\int_0^1 k^2(y)dy + 4\psi^2\int_0^1 k^2(y)dy\int_0^1 \psi(x)dx + o\left(\frac{1}{n^2h}\right)$

For proof of theorem, see (Alharbi, 2011)

Using theorem 1 and theorem 2, this yields

$$MSE(\hat{\sigma}_V^2) = h^{2r}\frac{(-1)^r}{r!}\int_0^1 k(y)y^r dy \int_0^1 \psi(s)\psi^{(r)}(s)ds + \frac{1}{n^2h}2\sigma^4\int_0^1 k^2(y)dy + 4\psi^2\int_0^1 k^2(y)dy\int_0^1 \psi(x)d + o\left(\frac{1}{n^2h}\right) + o(h^{2r})$$

(12)

Further analysis results in

$$MSE(\hat{\sigma}_V^2) = \frac{1}{N^2}\left\{\frac{1}{n}(\mu_4 - \sigma^4) + \frac{1}{n^2h}\left[2\sigma^4\int_0^1 k^2(y)dy + 2\sigma^2\int_0^1 k^2(y)dy\int_0^1 \psi^2(x)dx\right] + o\left(\frac{1}{n^2h}\right)^2\right\}$$

(13)

where $\mu_4 = E\left[\left(Y_i - \psi(x_i)\right)^4\right]$

As $n \to \infty$ implying also that $n^2 \to \infty$ and h $\to$ 0, the mean squared error in equation (13) reduces to zero.

$$MSE[\hat{\sigma}_V^2] = \frac{1}{nN^2}(\mu_4 - \sigma^4) \qquad (14)$$

This implies that $n \to \infty$ and considering also that $N \to \infty$, the RHS in equation (14) converges and reduces to zero.

Therefore, the Mean Square Error of the $\hat{\sigma}_V^2$ estimator reduce to zero asymptotically implying that the estimator is consistent in the mean square error.

4 *Data analysis*

A simulation study was conducted based on daily data obtained from the Kenyan Capital markets Authority for the years 2020 and 2021. Data on daily shares on banking and investments for the years 2020 and 2021 to assess the performance of our proposed $\hat{\sigma}_V^2$ estimator. The survey variable Y was simulated using a cosine mean function $\varphi_5(x) = \frac{3}{4}cos(10\pi x)$.

Samples of different sizes were drawn via simple random sampling without replacement and various properties of the estimator were evaluated. These properties include bias, Relative Efficiency (RE), Relative Root Mean Error (RRME), Standard Error (SE) and Confidence intervals. The performance of the developed estimator was then compared with that of other existing estimators such as

1. The Rice (1984) estimator $\hat{\sigma}_R^2 = \frac{1}{2(n-1)}\sum_{i=2}^{n}(Y_i - Y_{i-1})^2$
2. The GSJ estimator $\hat{\sigma}_{GSJ}^2 = \frac{2}{3(n-2)}\sum_{i=2}^{n-1}\left(\frac{1}{2}Y_{i-1} - Y_i + \frac{1}{2}Y_{i+1}\right)^2$
3. The H&M (1990) estimator $\hat{\sigma}_{HM}^2 = \frac{\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{n} w_{ij}Y_j\right)^2}{\left(n - 2\sum_{i=1}^{n} w_{ii} + \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}^2\right)}$

Results of the simulation experiment are illustrated in tables 1 and 2 below

Table 1: Unconditional properties of the estimator using a cosine model

Sample Size (n)	Estimator	Estimate	Bias	RRME	SE	RE	C.I (90%)	
	C.I (95%)			Lower	Upper	Lower	Upper	
	VO	1.0347	0.0347	0.1863	0.0643	12.0965	1.0258	1.0436
	1.0241	1.0453						
n=250	**RICE**	1.0272	0.0272	0.1650	0.1621	1.9048	1.0009	1.0535
	0.9959	1.0585						
	HM	1.0865	0.0865	0.2941	0.0942	5.6415	1.0804	1.0927
	1.0792	1.0938						
	GSJ	1.1394	0.1394	0.3734	0.2129	1.1044	1.1129	1.1659
	1.1079	1.1710						
	VO	1.2526	0.2526	0.5026	0.2538	0.0508	1.2486	1.2567
	1.2478	1.2574						
n=500	**RICE**	1.1275	0.1275	0.3571	0.1648	0.1205	1.1103	1.1447
	1.1070	1.1480						
	HM	1.3428	0.3428	0.5855	0.3431	0.0278	1.3407	1.3449
	1.3403	1.3453						
	GSJ	1.2817	0.2817	0.5308	0.2957	0.0374	1.2669	1.2965
	1.2641	1.2994						
	VO	1.0699	0.0699	0.2644	0.0832	1.9896	1.0624	1.0773
	1.0610	1.0787						
n=700	**RICE**	0.9349	-0.0651	0.25517	0.1028	1.3039	0.9218	0.9480
	0.9193	0.9505						
	HM	1.0924	0.0924	0.3040	0.0996	1.3887	1.0863	1.0985
	1.0851	1.0997						
	GSJ	1.0598	0.0598	0.2445	0.0951	1.5241	1.0476	1.0719
	1.0453	1.0743						
	VO	1.2145	0.2145	0.4632	0.2148	0.0480	1.2129	1.2162
	1.2126	1.2165						
n=900	**RICE**	1.0872	0.0872	0.2953	0.1108	0.1803	1.0759	1.0985
	1.0738	1.1006						
	HM	1.3846	0.3846	0.6202	0.3848	0.015	1.3830	1.3863
	1.3827	1.3866						
	GSJ	1.2601	0.2601	0.5100	0.2662	0.0313	1.2507	1.2694
	1.2490	1.2712						

Table 2: Conditional properties of the estimator using a cosine model

Sample size (n)	Estimator	Bias	RRME	SE	RE
	VO	0.0347	0.0422	0.0474	3.7471
n=250	**RICE**	0.0272	0.1302	0.1307	2.4902
	HM	0.0865	0.0897	0.0914	1.7692
	GSJ	0.1394	0.1736	0.1742	2.0568
	VO	0.2526	0.2532	0.2537	1.1359
n=500	**RICE**	0.1275	0.1373	0.1378	4.9155
	HM	0.3428	0.3428	0.3430	0.8405

	GSJ	0.2817	0.2817	0.2819	1.1554
	VO	0.0699	0.0751	0.0773	0.9936
n=700	**RICE**	-0.0651	0.0839	0.0842	1.9644
	HM	0.0924	0.0951	0.0961	0.9407
	GSJ	0.0598	0.0779	0.0782	2.0133
	VO	0.2145	0.2145	0.2147	1.5752
n=900	**RICE**	0.0872	0.0941	0.0943	7.8572
	HM	0.3846	0.3846	0.3848	0.8793
	GSJ	0.2601	0.2601	0.2602	1.3640

Figures 1-4 show the performance of the estimator for the various sample sizes

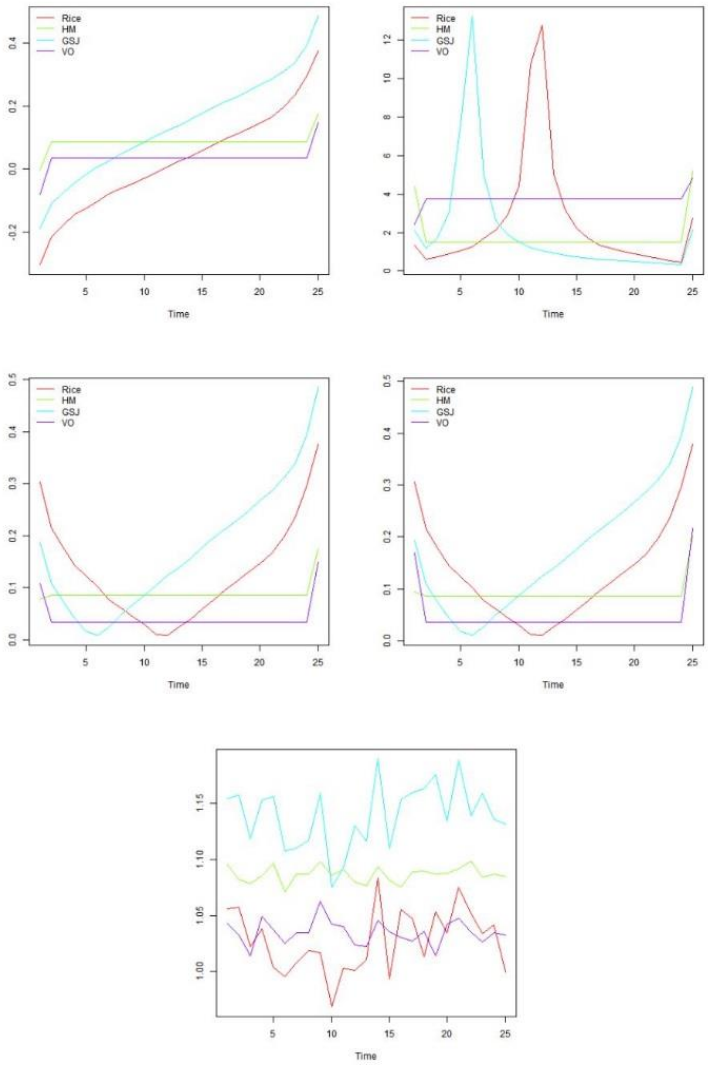


Figure 1: Conditional biases, RE, RRME, SE and Means for the estimators using a cosine model at n=250

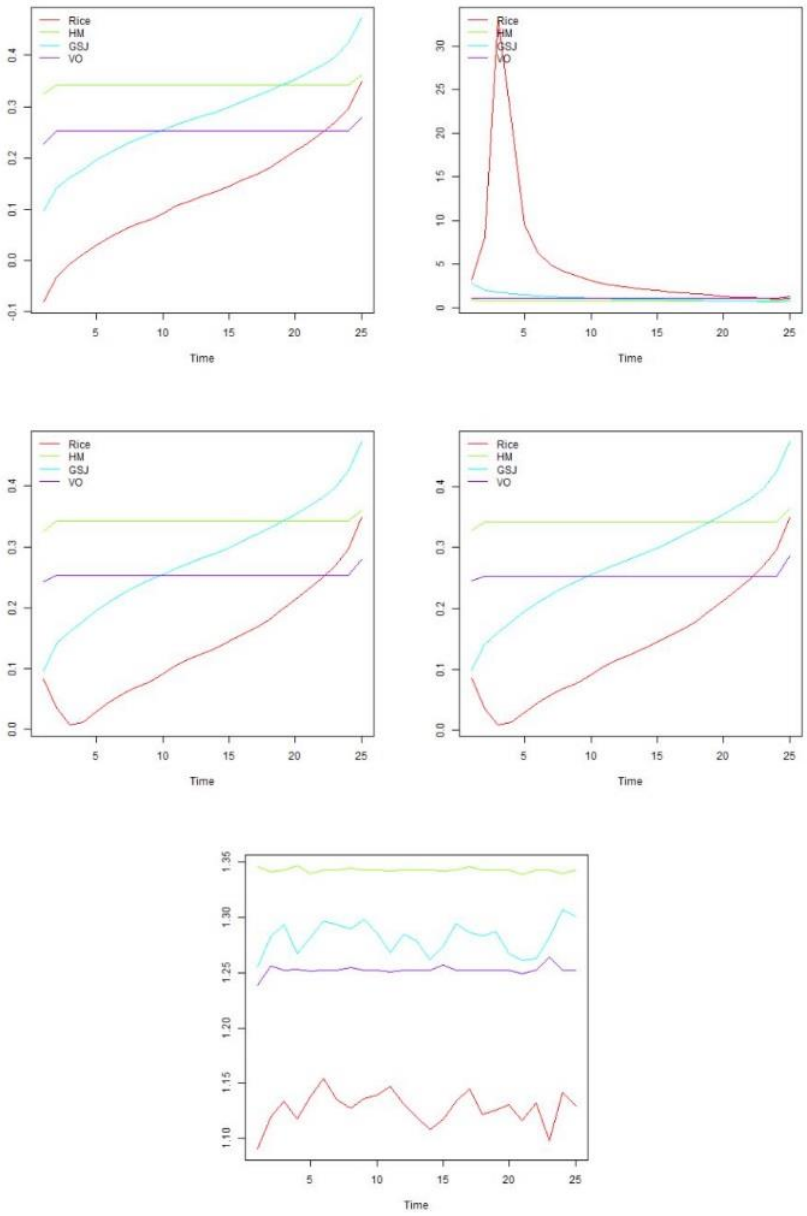


Figure 2: Conditional biases, RE, RRME, SE and Means for the estimators using a cosine model at n=500

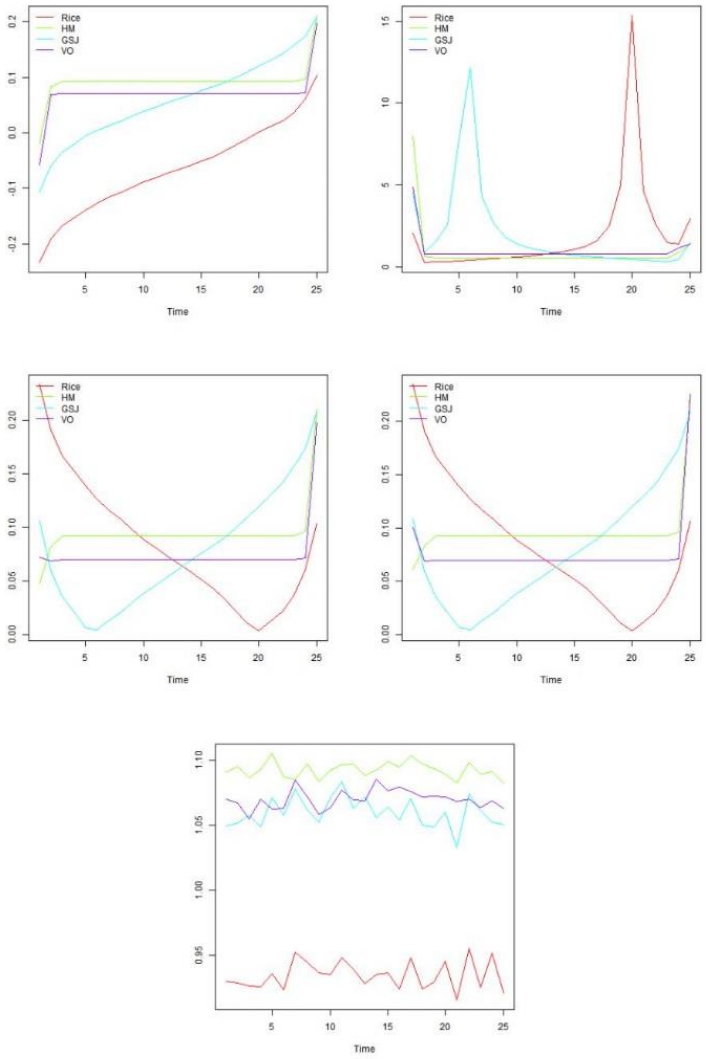


Figure 3: Conditional biases, RE, RRME, SE and Means for the estimators using a cosine model at n=700

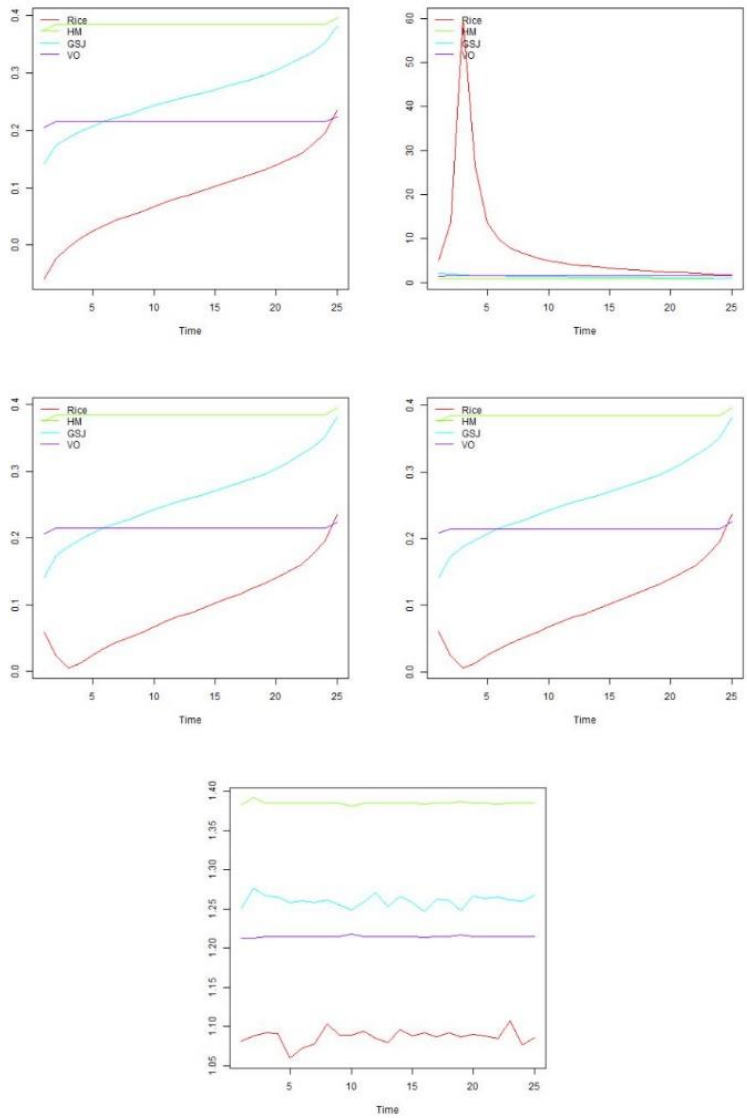


Figure 4: Conditional biases, RE, RRME, SE and Means for the estimators using a cosine model at n=900

From the tables and figures, the developed $\hat{\sigma}_V^2$ estimator performed comparatively well in all simulation sets.

From the above analysis, our developed $\hat{\sigma}_V^2$ estimator competes well with these known estimators and from the real data plots the purple curve for the $\hat{\sigma}_V^2$ estimator is closer to the x-axis indicating a smaller conditional bias, RRME, SE, RE, and Means.

5 CONCLUSION AND RECOMMENDATIONS FOR FUTURE STUDY

The main objective of this work was to develop a robust estimator of error variance for a finite population. To achieve this, a weighted average and kernel smoothers were utilized as tools for developing a $\hat{\sigma}_V^2$ estimator that solved the bias-variance trade off at the boundary points. The methodology employed possesses some kind of robustness since the asymptotic properties of the proposed $\hat{\sigma}_V^2$ estimators were derived. From the numerical and graphical comparison, though our developed $\hat{\sigma}_V^2$ estimator underestimates at some points due to the fact that we did not take care of the distance between observations, it is observed that our estimator has a relatively smaller bias, smaller variance, smaller Relative Root Mean Error, smaller standard error and smaller relative efficiency than the estimators of Rice (R) ; Gasser, Stroka and Jennen-steinmetz (GSJ) ; Hall&Marron (HM) . For future study one can take care of the distance between observations and develop an error variance estimator for finite population in a heteroscedastic setting using another real data set.

REFERENCES

1. Alharbi, Ye.F. (2011). Error Variance Estimation in nonparametric regression models. University of Birmingham, MPhil dissertation.
2. Dette, H.Munk, A. and Wagner (1998).Estimating the variance in non-parametric regression-What is a

reasonable choice? , Journal of the Royal Statistical society B 60:751-764.

3. Dorfman, A.H. (1992).Non parametric regression for Estimating Totals in finite populations. proceeding of the section on survey Research methods, American statistical Association Alexandria Washington DC,622-625.
4. Gasser, T., Sroka.L. and Jennen-Steinmetz C.(1986). Residual Variance and Residual pattern in non-linear regression. Biometrika 73, 625-633
5. Gasser.T. and Muller, H. (1981). Kernel Estimation of Regression Functions. In smoothing Techniques for curve estimation springer-Ver lag, pp.23-68,1979.
6. Hall. P., Kay, J.W and Tetterington, D.M (1990).Asymptotically optimal difference-based estimation of Variance in nonparametric regression. Biometrika 77, 521-528.
7. Hall P. and Carroll, R.J. (1986).Variance function estimation in regression. The effect of estimating the mean. J.Roy. statist.soc.ser.B. (methodological) 51, 3-14.
8. Bonface Miya Malenje, Winnie Onsongo Mokeira, Romanus Odhiambo, George Orwa. A Multiplicative Bias Corrected Nonparametric Estimator For a Finite Population Mean. American Journal of Theoretical and Applied statistics.vol.5, No.5,2016, pp.317-325.
9. Muller, U.U., Shick, A. and Welfemer, W. (2003). Estimating the error Variance in nonparametric regression by a covariate-matched U-statistics. statistics 37,179-188.
10. Rice .J. (1984).Bandwidth choice for nonparametric regression.Ann.statist.12, 1215-1230.
11. Wahba, G. (1990). Spline models for Observational Data, SIAM, Philadelphia .CBMS-NSF Regional Conference Series in Applied Mathematics, Vol.59.
12. Zheng, H., & Little, R.J. (2003). Penalized spline model-based estimation of the finite populations total from probability-proportional-to-size samples. Journal of of statistics-stockholm-, 19(2), 99-118.