American Journal of Applied Mathematics and Statistics, 2019, Vol. 7, No. 4, 146-151
Available online at http://pubs.sciepub.com/ajams/7/4/4
Published by Science and Education Publishing
DOI:10.12691/ajams-7-4-4

A Modified Nadaraya-Watson Estimator for the Variance of the Finite Population Mean

Charlotte K Mokaya*, DR. Edward Gachangi Njenga

Department of Mathematics, Kenyatta University, Nairobi, Kenya
*Corresponding author: mokayacharlotte@yahoo.com

Received May 06, 2019; Revised June 15, 2019; Accepted July 01, 2019

Abstract The main objective of this study was to derive a nonparametric estimator for the variance of the population mean when the population structure is nonlinear and heteroscedastic. Therefore, this paper sought to investigate the performance of Nadaraya-Watson estimator with a variable bandwidth. The methodology was derived by modifying the Nadaraya-Watson estimator where the bandwidth was a function of the range of observations. The performance of the proposed estimator was compared with other estimators i.e. Ratio estimator and Nadaraya-Watson with a fixed bandwidth. To measure performance of each of the estimators, average mean squared error was considered. It was found out that the Ratio estimator performs well for linear and homoscedastic populations while the Nadaraya-Watson with fixed bandwidth performs well for nonlinear and heteroscedastic populations. However, in the light of these findings, Nadaraya-Watson estimator (with variable bandwidth) was found to perform better and most efficient than the Ratio estimator and Nadaraya-Watson estimator (with fixed bandwidth) in nonlinear and heteroscedastic populations. It was also found to be the most robust compared to the estimators considered in this study.

Keywords: *bandwidth, Nadaraya-Watson, robust, efficiency*

Cite This Article: Charlotte K Mokaya, and DR. Edward Gachangi Njenga, "A Modified Nadaraya-Watson Estimator for the Variance of the Finite Population Mean." *American Journal of Applied Mathematics and Statistics*, vol. 7, no. 4 (2019): 146-151. doi: 10.12691/ajams-7-4-4.

1. Introduction

The concept of sample survey involves obtaining information regarding the population under study and subsequently making inferences about it.

Variance estimation of a population mean in sample survey is important as it gives further information about the accuracy of the estimators. According to the works of [1], the variance estimator of the population mean is also useful in the construction of confidence intervals and hypothesis testing.

However, variance estimation parameter can be unreliable when probability sampling is used for small sample sizes. Therefore, the use of auxiliary variables is considered as it gives more information about the population. According to [2], the incorporation of auxiliary information increases precision of the estimators. Presence of auxiliary variables provides good results compared to design-based techniques. On the other hand, use of auxiliary variables requires a model and assumptions to be specified.

In this paper, linearity and homoscedasticity assumptions were considered. Estimation of the parameter using an auxiliary variable yielded accurate results when the underlying assumptions are satisfied. On the contrary, when these assumptions are violated, estimators like the Ratio estimator becomes inefficient. This leads to inaccurate computation of the confidence intervals and wrong interpretation of the results is likely to occur. Furthermore, inaccurate results of the variance lead to wrong hypothesis testing and incorrect inference of population parameters.

Therefore, statisticians resorted to the use of non-parametric estimators which are robust when linearity and homoscedasticity assumptions are violated. Examples of some nonparametric estimators include; spline functions, local polynomial regression estimator and Nadaraya-Watson estimator with fixed bandwidth. Nonparametric variance estimation has been studied by [3] and [4], among others.

In this paper, a nonparametric estimator of the population variance is proposed which uses Nadaraya-Watson estimator with variable bandwidth.

2. Review of Nadaraya-Watson Estimator with Fixed Bandwidth

Nadaraya-Watson estimator is a nonparametric estimator proposed independently by [5] and [6] to estimate the mean function of a model using a sample of size n. Watson independently proposed a simple computer method for obtaining a graph from a large number of observations while Nadaraya proposed an estimator for approximating the regression curve as per [5] and [6]. This estimator is based on locally weighted averaging.

Consider a random sample s of size n with variables $(x_i, y_i)\ldots$, (x_n, y_n) and a joint probability distribution $f(x,y)$. Let $f(x)$ be the probability density function of x. Consider a nonparametric model of the form;

$$y_i = m(x_i) + \varepsilon_i \tag{1}$$

where m(x) is an unknown regression function and ε_i are independent random errors with a mean of zero and variance of σ^2.

According to [5] and [6], the estimator of the mean function $m(x_i)$ is given as;

$$\hat{m}_{nw}x_i = \frac{\sum_{i=1}^{n} y_i K\left(\frac{x-x_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)} \tag{2}$$

where $h > 0$ is a fixed bandwidth parameter that controls the degree of smoothness of in equation (2) and K is the kernel function. According to [7], a kernel is a piecewise continuous function that is symmetrical at zero and integrates to 1;

$$\int K(u)\,du = 1.$$

Population mean and variance using the Nadaraya-Watson mean function with fixed bandwidth in equation (2) above was proposed by [8] and is as shown below;

$$\hat{\mu}_y = \frac{1}{N}\sum_{i\varepsilon U}\sum_{j\varepsilon s} W_h\left(x, x_j\right) y_j. \tag{3}$$

And

$$\hat{\sigma}_y{}^2 = \frac{1}{N}\sum_{i\varepsilon U}\sum_{j\varepsilon s} W_h\left(x, x_j\right)\left(y_j - \hat{\mu}\left(x_j\right)\right)^2 + \frac{1}{N}\sum_{i\varepsilon U}\left(\hat{\mu}\left(x_i\right) - \hat{\mu}_y\right)^2 \tag{4}$$

where, $W = \dfrac{w\left(\frac{x-x_j}{h}\right)}{\sum_{i=1}^{n} w\left(\frac{x-x_j}{h}\right)}$.

In the next section, we derive Nadaraya-Watson estimator with a variable bandwidth. Let us call it a modified Nadaraya-Watson estimator.

3. Modified Nadaraya-Watson Estimator of the mean and variance functions.

In this paper, we derived an estimator of the variance of the population mean using the modified Nadaraya-Watson mean function proposed by [9]. [9], modified Nadaraya-Watson mean function using a bandwidth which is a function of the range of observations. This was aimed at improving the performance of the estimator as well as making it more stable.

Let (x_i, y_i) be a pair of variables of a sample of size n where x_i is an auxiliary variable and y_i is the study variable. These variables are positively related with a joint probability distribution function (pdf) $f(x,y)$. $f(x,y)$ can defined as $f(x,y)=f(y|x).f(x)$, where $f(x)$ is the marginal density of x and $f(y|x)$ is the marginal density of y.

The density functions as follows;

$$f\left(x,y\right) = \frac{1}{nh^2}\sum_{i=1}^{n} w^*\left(\frac{x-x_i}{h}\right)\left(\frac{y-y_i}{h}\right) \tag{5}$$

and

$$f\left(x\right) = \frac{1}{nh}\sum_{i=1}^{n} w^*\left(\frac{x-x_i}{h}\right). \tag{6}$$

Using nonparametric equation (1), the residual term ε_i is estimated as below;

$$\hat{\varepsilon}_i = y_i - \hat{m}\left(x_i\right) \tag{7}$$

Taking expectation on both sides of equation **(7)**, we get,

$$\hat{m}\left(x_i\right) = E(y \mid x) \tag{8}$$

From the expression in equation **(8)**, the modified Nadaraya-Watson mean function is derived as below;

$$\hat{m}^*_{nw}\left(x\right) = \frac{\sum_{i=1}^{n} \frac{y_i}{h\lambda_i} w\left(\frac{x-x_i}{h\lambda_i}\right)}{\sum_{i=1}^{n} \frac{1}{h\lambda_i} w\left(\frac{x-x_i}{h\lambda_i}\right)}. \tag{9}$$

where h is the bandwidth parameter and λ_i is defined as $\lambda_i = \left[\dfrac{\hat{f}\left(x\right)}{IQR}\right]^{-0.5}$, whereby IQR is the interquartile range of $\hat{f}(x_i)$. The Interquartile range is used because it is not affected by the extreme values in the data. $\boldsymbol{w}$ denotes a smoothing parameter with the following properties;

i) $w(t) \geq 0$

ii)

$$\int_{-\infty}^{\infty} w(t)dt = 1 \tag{10}$$

iii) $\int_{-\infty}^{\infty} (w)^2\,(t)dt < \infty.$

Taking the square of the residual term in equation **(7)**, we getTaking expectation of this expression, we get

$$\hat{\varepsilon}_i^2 = \left(y_i - \hat{m}\left(x_i\right)\right)^2. \tag{11}$$

Taking expectation of this expression, we get

$$\sigma^2\left(x\right) = E(\hat{\varepsilon}_i^2 \mid X = x) = \frac{\int \hat{\varepsilon}_i^2 * f\left(x,y\right)dy}{\int f\left(x\right)} = \frac{\sum_{i=1}^{n} w\left(\frac{x-x_i}{h}\right)\left(y_i - \hat{m}\left(x_i\right)\right)^2}{\sum_{i=1}^{n} w\left(\frac{x-x_i}{h}\right)}. \tag{12}$$

Consider equation **(12)** and the proposed variance estimator for modified Nadaraya-Watson mean function in equation **(9)**, $\hat{\sigma}^{*2}_{nw}\left(x\right)$, is given as;

$$\hat{\sigma}^{*2}_{nw}(x)=\frac{\sum_{i=1}^{n}\frac{1}{h\lambda_i}w^*\left(\frac{x-x_i}{h\lambda_i}\right)\left(y_i-\hat{m}^*_{nw}(x_i)\right)^2}{\sum_{i=1}^{n}\frac{1}{h\lambda_i}w^*\left(\frac{x-x_i}{h\lambda_i}\right)} \quad (13)$$

where h denotes the bandwidth parameter and w^* is the smoothing parameter that has the same properties as w specified in equation **(10)** and $\lambda_i=\left[\frac{\hat{f}(x)}{IQR}\right]^{-0.5}$.

3.2. Modified Nadaraya-Watson Estimates of the Finite Population Mean and Variance

In this section, we derived population mean and variance using the modified Nadaraya-Watson mean and variance functions obtained in section (**3.1**).

Population mean function is defined as

$$\begin{aligned} m(y)&=E_x(E(y\mid x)=\int_{-\infty}^{+\infty}E(y|x)f(x)dx \\ &=\int_{-\infty}^{+\infty}m(x)f(x)dx. \end{aligned} \quad (14)$$

The estimate of this population mean function given in **(14)** is

$$\hat{m}(y)=\int_{-\infty}^{+\infty}\hat{m}(x)\hat{f}(x)dx. \quad (15)$$

The population variance function is defined as below;

$$\begin{aligned} \sigma^2(y)&=E_x\left[V(y|x)\right]+E_x\left(E(y|x)-m(y)\right)^2 \\ &=\int_{-\infty}^{+\infty}V(y|x)f(x)dx+\int_{-\infty}^{+\infty}\left(m(x)-m(y)\right)^2f(x)dx \quad (16)\\ &=\int_{-\infty}^{+\infty}\sigma^2(x)f(x)dx+\int_{-\infty}^{+\infty}\left(m(x)-m(y)\right)^2f(x)dx. \end{aligned}$$

and the estimate of the variance function in equation **(16)** is given as

$$\hat{\sigma}^2(y)=\int_{-\infty}^{+\infty}\hat{\sigma}^2(x)\hat{f}(x)dx+\int_{-\infty}^{+\infty}\left(\hat{m}(x)-\hat{m}_y\right)^2\hat{f}(x)dx. \quad (17)$$

Combining equation **(9)** and **(13)** we get the proposed modified Nadaraya-Watson estimators of the population mean and variance as below;

$$\hat{\mu}_y=\frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{n}w^*\left(x_j,x_i\right)y_i \quad (18)$$

$$\begin{aligned} \hat{\sigma}_y{}^2&=\frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{n}w^*\left(x_j,x_i\right)\left(y_i-\hat{m}^*(x)\right)^2 \\ &+\frac{1}{N}\sum_{j=1}^{N}\left(\sum_{i=1}^{n}w^*\left(x_j,x_i\right)y_i-\hat{m}_y\right)^2 \end{aligned} \quad (19)$$

where, $w^*\left(x_j,x_i\right)=\dfrac{\frac{1}{h\lambda_i}w^*\left(\frac{x-x_i}{h\lambda_i}\right)}{\sum_{i=1}^{n}\frac{1}{h\lambda_i}w^*\left(\frac{x-x_i}{h\lambda_i}\right)}$.

4. Simulation Studies

Performance of the three estimators, that is, the Ratio estimator, Nadaraya-Watson with fixed bandwidth and the modified Nadaraya-Watson was compared using six simulated populations and one natural population. The average mean squared error criterion was used to measure efficiency of the estimators.

4.1. Description of the Study Population and Estimators

Below is the description of the populations where linear and quadratic equations were considered. The equations from the study took the below forms;

Linear Equation: $1+2*(x-0.5)+e_i$

Quadratic Equation: $1+2*(x-0.5)^2+e_i$

where the auxiliary variable X, was simulated from the uniform distribution with the interval [0,1]; $X_i \sim U[0,1]$.

The error term was simulated from the normal distribution with mean (0) and variance of (0.1); $e_i \sim N(0,0.1)$.

Table 1. Description of the study populations

Population	Description
a) $Y=m(x)+e_i$	
Population 1	Linear and Homoscedastic
Population 2	Quadratic and Homoscedastic
b) $Y=m(x)+e_i*X$	
Population 3	Linear and Heteroscedastic
Population 4	Quadratic and Heteroscedastic
c) $Y=m(x)+e_i*sqrt(X)$	
Population 5	Linear and Heteroscedastic
Population 6	Quadratic and Heteroscedastic
d) Real data was obtained from Central Bank of Kenya website on Foreign Trade Summary Imports for the period 1999-2017. In this population, the Auxiliary Variable X_i was taken to be the number of government imports in the i^{th} month in a given year. The study variable Y_i was taken to be the total imports of the i^{th} month in a particular year. And $i=1,\dots,100$.	Nonlinear and Heteroscedastic

Below is a summary of the formulae for the estimators used in this study.

Ratio Estimator: $\hat{V}(\bar{y}_R)=\frac{1-f}{n}[\sum_{i=1}^{n}\frac{(y_i-Rx_i)^2}{n-1}]$, where $f=\frac{n}{N}$ and $R=\frac{\bar{y}}{\bar{x}}$

Fixed Nadaraya-Watson: $\hat{\sigma}_y{}^2=\frac{1}{N}\sum_{j\varepsilon U}\sum_{i\varepsilon s}W^*{}_h(x,x_i)\left(y_i-\hat{\mu}(x_i)\right)^2+\frac{1}{N}\sum_{j\varepsilon U}\left(\hat{\mu}(x_i)-\hat{\mu}_y\right)^2$

Modified Nadaraya-Watson: $\hat{\sigma}_y{}^2=\frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{n}w^*\left(x_j,x_i\right)\left(y_i-\hat{m}^*_{nw}(x_i)\right)^2+\frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{n}w\left(x_j,x_i\right)\left(y_i-\hat{m}_y\right)^2$.

4.2. Description of the Computation Procedure

A population of size N=10,000 was simulated using R software. 1000 samples of size n=500 were selected using simple random sampling without replacement.

The Gaussian Kernel function defined as;

$$K(u)=\frac{1}{\sqrt{2\pi}}\exp\left(-u^2/2\right),\ -\infty<u<\infty$$

where $u=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x-x_i}{h}\right)$ was used in the study for fixed and modified Nadaraya-Watson estimators.

The performance of the kernel function depends on the choice of the bandwidth parameter. Choosing a bandwidth that balances the variance with the bias is crucial. According to [10], choice of the bandwidth can be done by data analysts either subjectively or objectively. In this research, fixed bandwidth h defined below, was obtained from unbiased (Least square) cross-validation method whose equation is as below;

$$UCVh=\int\hat{f}^2(x)-\frac{2}{n}\sum\hat{f}_{-i}(x_i),$$

where n is the number of observations and $\hat{f}_{-i}(x_i)$ is the density estimate without data point x_i. The smoothing parameter h is obtained by minimizing $UCVh$.

Average bias was obtained using the below equation;

$$Bias=\sum_{i=1}^{500}\frac{Var_{ik}-\mathrm{var}(\bar{y})}{500}.$$

where k denotes different estimators.

Average MSE was obtained using equation;

$$MSE=\sum_{i=1}^{500}\frac{\left(Var_{ik}-\mathrm{var}(\bar{y})\right)^2}{500}+\left(Bias(\bar{y})\right)^2.$$

Relative change of efficiency (RCE) was also calculated from the equation below;

$$RCE(i)=\frac{\begin{bmatrix}MSE(\bar{y}),inpopulation(i+1)\\-MSE(\bar{y}),inpopulation(1)\end{bmatrix}}{MSE(\bar{y}),inpopulation(1)}$$

i=Ratio estimator, fixed Nadaraya-Watson and modified Nadaraya-Watson estimators.

The following are scatter plots showing distributions of seven populations analyzed.

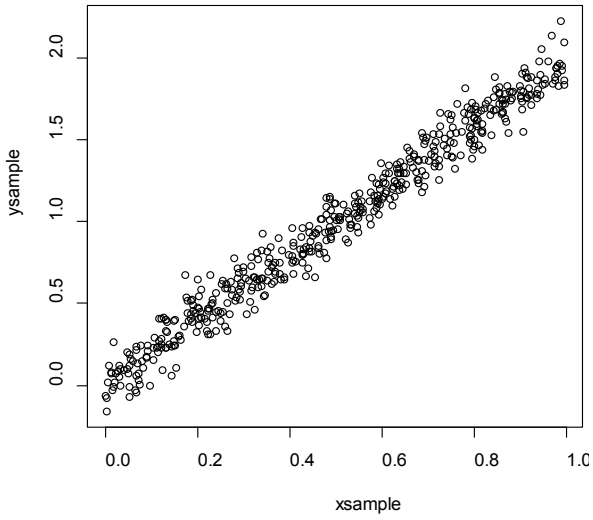

Figure 1. Linear population with homoscedastic variance structure

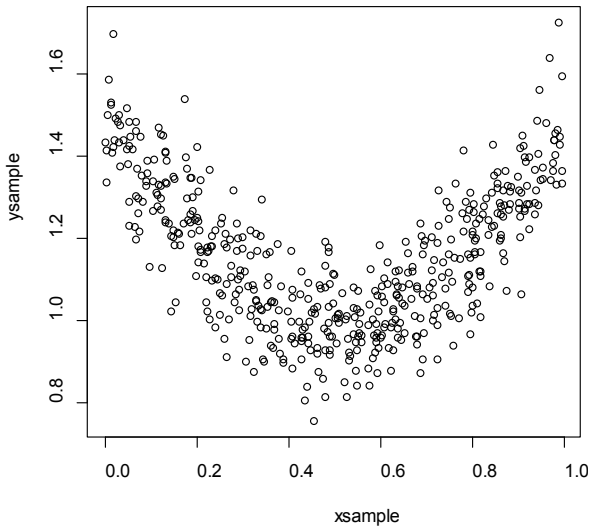

Figure 2. Quadratic population with homoscedastic variance structure

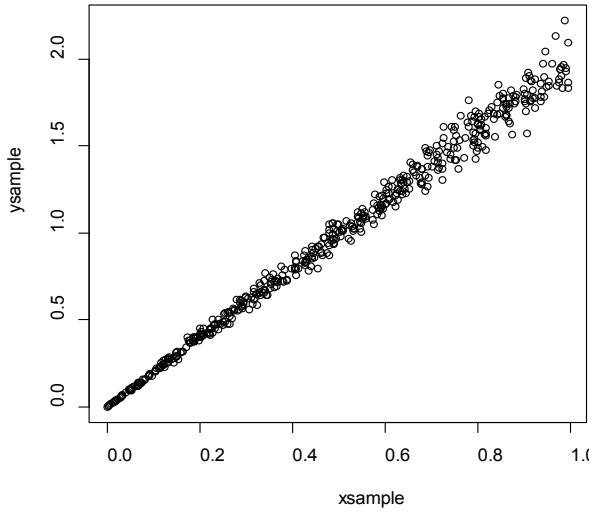

Figure 3. Linear population with heteroscedastic variance structure

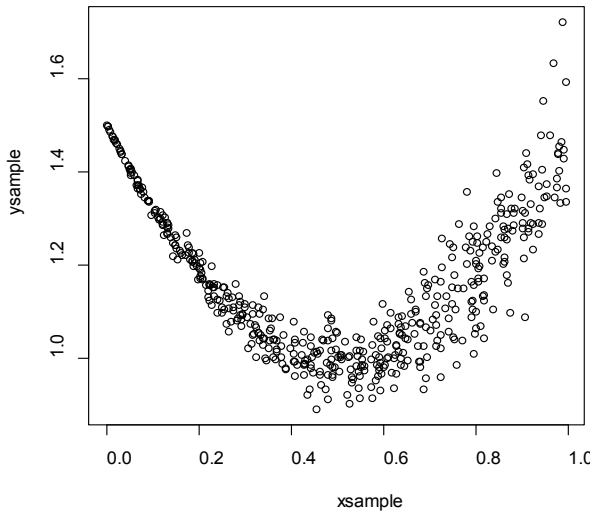

Figure 4. Population is quadratic with heteroscedastic variance structure

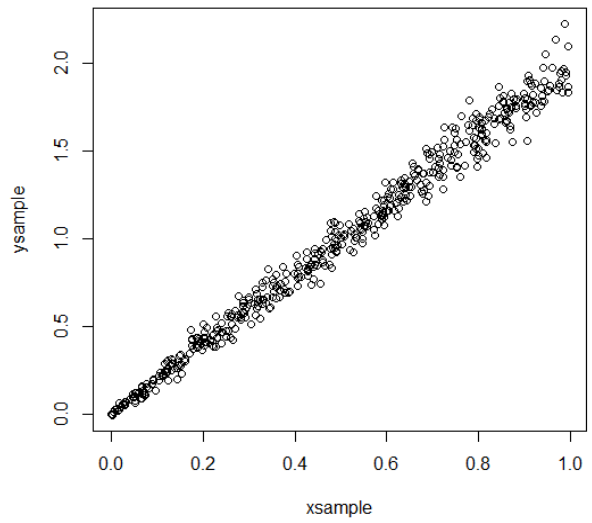

Figure 5. Linear population with heteroscedastic variance structure

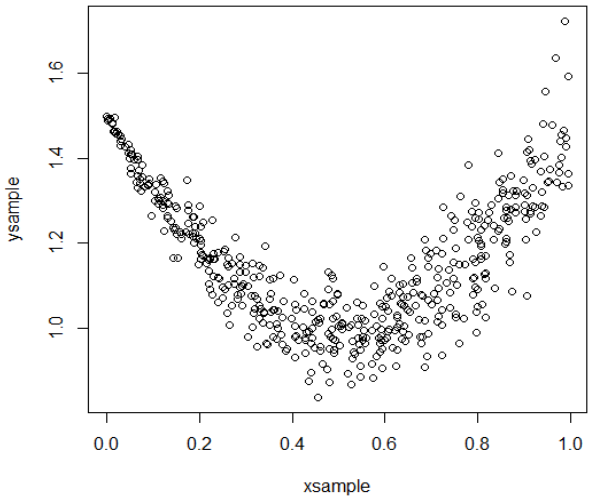

Figure 6. Population is nonlinear and heteroscedastic

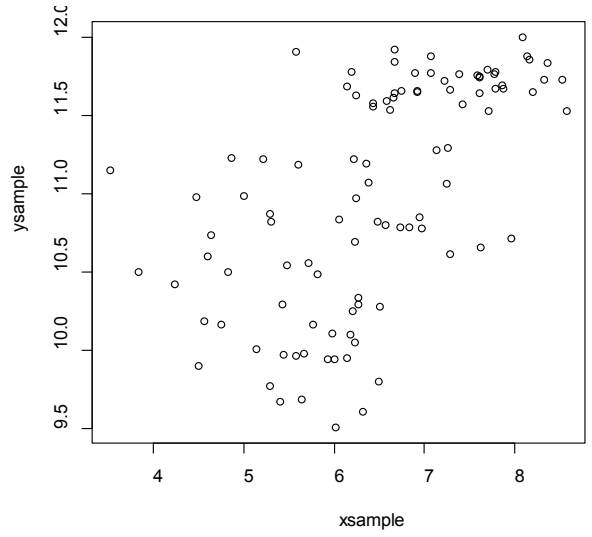

Figure 7. Real population that is neither linear nor homoscedastic

4.3. Results and Interpretations

In Table 2, Ratio estimator has the smallest variance for populations I, III and V followed by Nadaraya-Watson with fixed bandwidth estimator. Modified Nadaraya-Watson estimator comes in last. For populations II, IV, VI and VII which are nonlinear homoscedastic and nonlinear heteroscedastic, our proposed estimator has the smallest variance estimate followed by Nadaraya-Watson estimator with fixed bandwidth. Ratio estimator comes in third.

The squared average mean errors of the estimators are calculated to assess their efficiency. The results are as shown in Table 3. In light of the statistics and tests above, modified Nadaraya-Watson is most efficient in populations with nonlinear structure compared to the other two estimators. Therefore, modified Nadaraya-Watson estimator is most robust when the linear structure of the populations is violated.

Table 2. Variance Estimators

	Pop I	Pop II	Pop III	Pop IV
Ratio Estimator	0.0085	0.4436	0.0028	0.4322
Nadaraya-Watson (fixed bandwidth)	0.1162	0.0099	0.1404	0.0112
Modified Nadaraya-Watson(variable bandwidth)	0.1269	0.0063	0.1570	0.0103
	Pop V	Pop VI	Pop VII	
Ratio Estimator	0.0042	0.4355	1.4332	
Nadaraya-Watson (fixed bandwidth)	0.1307	0.0104	0.8472	
Modified Nadaraya-Watson(variable bandwidth)	0.1448	0.0083	0.8388	

Table 3. Average Mean Square Error

	Pop I	Pop II	Pop III	Pop IV
Ratio Estimator	0.1216	0.6133	0.1152	0.5978
Nadaraya-Watson (fixed bandwidth)	0.1687	0.0104	0.1795	0.0114
Modified Nadaraya-Watson (variable bandwidth)	0.1744	0.0069	0.1898	0.0105
	Pop V	Pop VI	Pop VII	
Ratio Estimator	0.1168	0.6026	2.1984	
Nadaraya-Watson (fixed bandwidth)	0.1744	0.0107	0.9175	
Modified Nadaraya-Watson(variable band width)	0.1828	0.0086	0.9306	

5. Conclusion

Following the results of our data analysis, we noted that the Ratio estimator performs well for linear and homoscedastic model but when there is violation of the model structure, the estimator breaks down. Therefore, it can be concluded that Ratio estimator is not efficient when linear and homoscedastic assumptions of a population are violated. Nadaraya-Watson with fixed bandwidth and Nadaraya-Watson with variable bandwidth estimators performed well in nonlinear heteroscedastic populations. However, the proposed estimator performed even better compared to the Nadaraya-Watson with fixed bandwidth as it was the most efficient amongst the estimators considered in this study.

References

[1] Yuejin, Z., Yebin, C. and Tiejun, T. (2014). A Least Squares Method for Variance Estimation in Heteroscedastic Nonparametric Regression. Hindawi Publishing Corporation, 1-14.

[2] Wu, C. and Sitter, R. (2001). Variance estimation for the finite population distribution function with complete auxiliary information.The Canadian Journal of Statistics, 29, 289-307.

[3] Hall, P. and Marron, J.S. (1990). On Variance estimation in Nonparametric Regression. Biomerika, 77, 415-419.

[4] Shen, S. and Mei, C. (2009). Estimation of the Variance Function in Heteroscedastic Linear Regression Models. Communications in Statistics – Theory and Methods, 38, 1098-1112.

[5] Nadaraya, E.A. (1964). On Estimating Regression. Theory of probability application, 9,141-142.

[6] Watson, G. (1964). Smooth Regression Analysis. Sankhya, 26, 359-372.

[7] Hardle, W. (1994). Applied Nonparametric Regression. Cambridge University Press.

[8] Njenga, E. and Smith, T.M.F. (1992). Robust Model-Based Methods for Analytic Surveys. Survey Methodology, 18,187-208.

[9] Aljuhan, H. and Alturuk, I. (2014). Modification of the Adaptive Nadaraya-Watson Kernel Regression Estimator. Academic Journals, 9, 966-971.

[10] Fan and Irene, G. (1992). Variable Bandwidth and Local Linear Regression Smoothers. The Annals of Statistics, 20, 2008-2036.