American Journal of Theoretical and Applied Statistics
2015; 4(5): 396-403
Published online September 8, 2015 (http://www.sciencepublishinggroup.com/j/ajtas)
doi: 10.11648/j.ajtas.20150405.20
ISSN: 2326-8999 (Print); ISSN: 2326-9006 (Online)

Estimation of Population Total Using Spline Functions

Gladys Gakenia Njoroge

Department of Physical Sciences, Chuka University, Chuka, Kenya

Email address:
gg.njoroge@gmail.com

To cite this article:
Gladys Gakenia Njoroge. Estimation of Population Total Using Spline Functions. *American Journal of Theoretical and Applied Statistics*. Vol. 4, No. 5, 2015, pp. 396-403.doi: 10.11648/j.ajtas.20150405.20

Abstract: This study sought to estimate finite population total using spline functions. The emerging patterns from spline smoother were compared with those that were obtained from the model-based, the model-assisted and the non-parametric estimators. To measure the performance of each estimator, three aspects were considered: the average bias, the efficiency by use of the average mean square error and the robustness using the rate of change of efficiency. We used six populations: four natural and two simulated. The findings showed that the model-based estimator works very well in terms of efficiency while the model-assisted is almost unbiased when the model is linear and homoscedastic. However, the estimators break down when the underlying model assumptions are violated. The Kernel Estimator (Nadaraya-Watson) is found to be the most robust of the five estimators considered. Between the two spline functions that we considered, the periodic spline was found to perform better. The spline functions were found to provide good results whether or not the design points were uniformly spaced. We also found out that, under certain conditions, a smoothing spline estimator and a Kernel estimator are equivalent. The study recommends that both the ratio estimator and the local polynomial estimator should be used within the confines of a linear homoscedastic model. The Nadaraya-Watson and the periodic spline estimators, both of which are non-parametric, are highly robust. The Nadaraya-Watson however is even more robust than the periodic spline.

Keywords: Population Total, Estimator, Efficiency, Homoscedasticity, Robustness

1. Introduction

The name "spline function" was given by [11] to the piecewise polynomial functions known as univariate polynomial spines. This was because of their resemblance to the curves obtained by their draftsmen using a mechanical spline- a thin flexible rod with a groove and a set of weights called "ducks" used to position the rods at points through which it was derived to draw smooth interpolating curves passing through prescribed points. The basic idea dates back at least to [16]. More recent papers on the subject include [6, 12, and 14] among others.

The available literature in statistics indicates that the approaches mostly used in estimation of population total include the model-based, the design-based and the model-assisted approaches. The non-parametric approach has also picked up especially with such works as of [5, 10] on the Kernel estimation. The spline smoothing is another non-parametric approach to estimation of finite population total. However, not much literature is available on this approach and neither has there been a lot of its application on estimation of population, as compared to the previous approaches. This study therefore sought to estimate finite population total using spline functions while using ratio estimator, local polynomial estimator and Kernel functions for a numerical comparison to determine whether the patterns of estimation would be as accurate as those derived from the use of previous approaches. To measure the performance of each estimator, we considered three aspects namely: bias, the efficiency by use of the average mean square error and the robustness using the rate of change of efficiency.

2. The Estimators

2.1. Ratio Estimator (Model-Based)

The prediction approach is based on a model. Royall [9] summarizes the philosophy behind this approach. Suppose the number of the units N in the finite population is known and that in each unit is associated a number y_i. The general problem is to choose some of the units as a sample, observe the y's for the sample units and then use those observations to estimate the value of some function $f\left(y_1, y_2, ..., y_N\right)$ of all the

y 's in the population. The prediction approach treats the numbers $y_1, y_2, ..., y_N$ as realized values of random variables $Y_1, Y_2, ..., Y_N$. After the samples have been observed, estimating $f\left(y_1, y_2, ..., y_N\right)$ entails predicting a function of the unobserved y 's. The relationships among the random variables both the auxiliary variable x and the survey variable y are expressed in a model. The general model being

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

Where $m\left(x_i\right)$ is the mean function and e_i a random error term. After selecting and observing a sample, the y 's for the sample units get to be known but the values for the non-sample units remain unknown. The ignorance of the non-sample y values implies that some functions of those values must be mathematically predicted in order to have an estimator or predictor for the full population. Suppose the study of the scatter diagram reveals that the n sample points are clustered around straight line passing through the origin. Then, the ratio $y_i / x_i, i \in s$ are more or less the same. We may then postulate the approximate relation.

$\frac{\sum_{i\in s} y_i}{\sum_{i\in s} x_i} = \frac{\sum_{i\notin s} y_i}{\sum_{i\notin s} x_i}$. Hence we can write

$$T\left(y\right) = \sum_{i\in s} y_i + \sum_{i\notin s} y_i \tag{2}$$

From which we can suggest an estimator of ȳ as

$$\hat{\bar{y}} = \left(\frac{\sum_{i\in s} y_i}{\sum_{i\in s} x_i}\right)\bar{x}$$

$$= \left[\frac{\bar{y}\left(s\right)}{\bar{x}\left(s\right)}\right]\bar{x} \tag{3}$$

where $\bar{y}\left(s\right)$ and $\bar{x}\left(s\right)$ refer to the sample means for y and x, respectively. The $\bar{x}$ is assumed to be known before hand. This estimator in (3) is popularly known as the Ratio Estimator [7]. The estimator of the population total using the model-based approach (prediction approach) thus becomes

$$\hat{T} = \sum_{i\in s} \hat{y}_i + \sum_{i\notin s} \hat{y}_i \tag{4}$$

Where

$$\hat{y}_i = \begin{cases} \mathrm{E}\left(y_i / x_i\right) & i\notin s \\ y_i & i\in s \end{cases} \tag{5}$$

substituting equation (5) in (4) gives

$$\hat{T} = \sum_{i\in s} \hat{y}_i + \sum_{i\notin s} \mathrm{E}\left(\frac{y_i}{x_i}\right) \tag{6}$$

we take $\mathrm{E}\left(y_i / x_i\right) = m\left(x_i\right)$ for the non-sample where $m\left(x_i\right)$ is linear and $v\left(x_i\right) = 1$ i.e. homoscedastic [9]. Let $\hat{m}\left(x_i\right)$ be the predictor of $m\left(x_i\right)$ of the non sample values which is given as $\hat{m}\left(x_i\right) = \left[\frac{\sum_{i\in s} y_i}{\sum_{i\in s} x_i}\right]\left(\sum_{i\notin s} x_i\right)$

Thus, our estimate of the population total under Royall's prediction model is

$$N\hat{\bar{T}}_{re} = n\bar{y}\left(s\right) + \left[\frac{\sum_{i\in s} y_i}{\sum_{i\in s} x_i}\right]\left(\sum_{i\notin s} x_i\right)$$

Therefore,

$$\hat{\bar{T}}_{re} = \frac{\bar{y}\left(s\right)}{\bar{x}\left(s\right)}.\bar{X} \tag{7}$$

$\hat{\bar{T}}_{re}$ is the ratio estimator for the population total

2.2. The Local Polynomial Regression Estimator (Model–Assisted)

Breidt and Opsomer [2], assumed that the population is generated by the super population model: $y_i = m\left(x_i\right) + \ell_i$ where ℓ_i is an independent sequence of random variables with mean zero and the variance is a smooth function of x. They employed local polynomial smoothing techniques to obtain a model-assisted regression estimator for the finite population total. We consider a finite population of N units with label set $U = \left\{1, 2, ..., N\right\}$ an auxiliary variable x_i is observed. A probability sample s is drawn from U according to a fixed size sampling design $P\left(.\right)$ where $P\left(s\right)$ is the probability of drawing the sample s. Let n be the size of s. Assume $\pi_i = p\left\{i \in s\right\} = \sum_{s:i\in s} p(s) > 0$

And

$$\pi_{ij} = p\left\{i, j \in s\right\} = \sum_{s:i,j\in s} p(s) > 0 \ \forall i, j \in U .$$

The study variable y_i is observed for each $i \in s$. The goal is to estimate $T_y = \sum_{i\in U} y_i$

Let $I_i = 1$ if $i \in s$ and $I_i = 0$ otherwise.

$Ep[I_i] = \pi_i$, where $Ep[.]$ denotes expectation with

respect to the sampling design i.e. averaging over all possible samples from the finite population.

Using this notation, an estimator $\hat{T}_y$ of T_y is said to be design-unbiased if

$$Ep\left[\hat{T}_y\right] = T_y$$

A well known design-unbiased estimator of T_y is the Horvitz-Thompson estimator,

$$\hat{T}_y = \sum_{i\in s}\frac{y_i}{\pi_i} = \sum_{i\in U}\frac{y_i I_i}{\pi_i} \text{ [4]} \quad (8)$$

The variance of the Horvitz Thompson estimator under the sampling design is

$$Var\left(\hat{T}_y\right) = \sum_{i,j\in U}\left(\pi_{ij} - \pi_i\pi_j\right)\frac{y_i y_j}{\pi_i\pi_j} \quad (9)$$

An estimator motivated by modeling the finite population of y_i's, conditioned on the auxiliary variable x_i, as a realization from a super population ξ, in which $y_i = m(x_i) + e_i$ is proposed. Given x_i, $m\left(x_i\right)$ is called the regression function, while $V\left(x_i\right)$ is the variance function.

Let k denote a continuous kernel function and let h denote the bandwidth. We begin by defining the Local polynomial Kernel estimator of degree q based on the entire finite population. Let $Y_U = \left[Y_i\right] i \in U$ be the N-vector of y_i's in the finite population.

Define the $N\times\left(q+1\right)$ matrix as

$$X_{Ui} = \begin{bmatrix} 1, x_1 - x_i \ldots \left(x_1 - x_i\right)^q \\ 1, x_n - x_i \ldots \left(x_n - x_i\right)^q \end{bmatrix} = \left[1, x_j - x_i \ldots \left(x_j - x_i\right)^q\right]_{j\in U}$$

and define the $n\times n$ matrix,

$w_{Ui} = diag\left\{\frac{1}{h}k\left(\frac{x_j - x_i}{h}\right)\right\}_{j\in U}$ the Kernel weights where $1 \le i \le n$

$h > 0$ is the smoothing parameter (bandwidth). Let ℓ_r represent a vector with a 1 in the r^{th} position and 0 elsewhere. The local polynomial kernel estimator of the regression function at x_i, based on the entire finite population is then given by

$$m_i = \ell_i\left(x'_{Ui} w_{Ui} x_{Ui}\right)^{-1} x'_{Ui} w_{Ui} x_U = w'_{Ui} y_U \quad (10)$$

which is well defined as long as $x'_{Ui} w_{Ui} x_{Ui}$ is invertible.

Since only y_i in $s \subset U$ are known, m_i is replaced by a sample-based consistent estimator to make its calculation possible.

Let $y_s = \left[y_i\right]_{i\in s}$ be the n-vector y_i's obtained in the sample.

Define the $n\times\left(q+1\right)$ matrix, $x_{si} = \left[1, x_j - x_i \ldots \left(x_j - x_i\right)^q\right]_{j\in s}$

And define the $n\times n$ matrix, $w_{si} = diag\left\{\frac{1}{\pi_j h}k\left(\frac{x_j - x_i}{h}\right)\right\}_{j\in s}$ a sample design-based estimator of m_i is then given by $\hat{m}_i = \ell_1^1\left(x'_{si} w_{si} x_{si}\right)^{-1} x'_{si} w_{si} y_s = w'_{si} y_s$ as long as $x'_{si} w_{si} x_{si}$ is invertible. $\ell_1^1 = \left(1,0,0,\ldots,0\right)$ which is a $a\left(q+1\right)$ vector.

The above shows that the local polynomial estimators linear smoothers are of the form $\sum_{i=1}^{n} w_i\left(x\right) y_i$

The coefficient of the linear combination depends on the degree q of the polynomial approximation. We note that for $q = 0$, the estimator reduces to the Nadaraya-Watson estimator [1]. Now, based on the proposed estimator in equation (6) $\hat{T} = \sum_{i\in s}\frac{y_i}{\pi_i}$, and assuming that $q = 0$ throughout, due to mathematical complexity, then the local polynomial regression estimator for the finite population total is given by

$$\hat{T}_{lp} = \sum_{i\in s}\frac{y_i - \hat{m}\left(x_i\right)}{\pi_i} + \sum_{i=1}^{N}\hat{m}\left(x_i\right) \quad (11)$$

where $\hat{m}\left(x_i\right)$ is the sample estimator for $m\left(x_i\right)$. Substituting equation (9) in (11) above gives

$$\hat{T}_{lp} = \sum_{i\in s}\frac{y_i - w'_{si} y_s}{\pi_i} + \sum_{i=1}^{N} w'_{si} y_s \quad (12)$$

2.3. Kernel Estimation

We consider the Nadaraya-Watson Kernel estimator. It is assumed that the auxiliary information is available for the entire population and the auxiliary variable x and the study variable y are related in a more general way. The studies of the properties of the proposed estimator are conditional on the available sample and non-sample values of the auxiliary variable x. A conceptually simple approach to a representation of the weight sequences $\left\{w_i\left(x\right)\right\}, i = 1,2,\ldots,n$ is to describe the shape of the weight function $w_i\left(x\right)$ by a density function with a scale parameter

that adjusts the size and the form of the weights near x. This function is commonly referred to as Kernel K. The Kernel is continuous, bounded and symmetric function which integrates to one,

$$\int K\left(u\right)du = 1 \tag{13}$$

To estimate $m\left(x\right)$ in model (1) one method is to average the nearby values of y_i where "nearby" is measured in terms of the distance $|x_i - x|$.

Let $K_h\left(u\right) = h^{-1}K\left(\frac{u}{h}\right)$ be the Kernel with bandwidth h.

The weight sequences for the Kernel smoothers (for one dimensional *x*) is given by

$$w_i\left(x\right) = \frac{K_h\left(x_i - x\right)}{\sum_{i=1} K_h\left(x_i - x\right)} \tag{14}$$

This form of Kernel weights (13) was proposed by [8, 15].

The Nadaraya-Watson estimator of $m\left(x\right)$ in (1) is

$$\hat{m}\left(x\right) = \sum_i w_i\left(x\right)y_i \tag{15}$$

On substituting (13) in (14) we get

$$\hat{m}\left(x\right) = \frac{\sum_{i=1}^{n} K_h\left(x_i - x\right)y_i}{\sum_{i=1}^{n} K_h\left(x_i - x\right)} \tag{16}$$

The shape of the Kernel weights is determined by k. One unique feature of the size of the bandwidth is that the smaller it is the more concentrated are the weights around *x*.

Selection of the bandwidth is the important part of the Kernel estimation method. When selecting the bandwidth we need to consider the error in our selection. This is the deeper reason why precision has to be measured in terms of point wise Mean Squared Error (MSE), the sum of variance and squared bias. The MSE is given by

$E\left[\hat{m}\left(x\right) - m\left(x\right)\right]^2$ which tends to zero for the Kernel estimator.

$\hat{m}_k$, if $h \to \infty$ and $\frac{h}{n} \to 0$.

The non-parametric regression-based estimator, $\hat{T}_{np}$, for the population total T is given by

$$\hat{T}_{np} = \sum_{i=1}^{n} y_i + \sum_{i \notin s} \hat{m}\left(x_i\right) \tag{17}$$

where $\hat{m}\left(x\right)$ is the Nadaraya-Watson estimator in (15).

Therefore the Nadaraya-Watson estimator of the population total is given by substituting (15) in (16) which gives

$$\hat{T}_{nw} = \sum_{i \in s} y_i + \sum_{i \notin s} \left\{ \frac{\sum_{i=1}^{n} k_n\left(x_i - x\right)y_i}{\sum_{i=1}^{n} k_h\left(x_i - x\right)} \right\} \tag{18}$$

where $\hat{T}_{nw}$ represents the Nadaraya-Watson estimator of the population total.

2.4. The Spline Smoothing

A measure of the rapid local variation of a curve can be given by a roughness penalty such as the integrated square second derivative. Various penalties have been suggested and used. For example, [3], but $\int \left(m''\right)^2$ is most convenient for our purpose. Using this measure, we define the modified sum of squares as

$$s\left(m\right) = \sum \left\{y_i - m\left(x_i\right)\right\}^2 + \lambda \int m''\left(x\right)^2 dx \tag{19}$$

The idea behind spline estimation then, is to find the function $m\left(x\right)$ such that the following minimization problem is solved

$$\min_{m(.)} \left\{ \sum_{i=1}^{n} \left(y_i - m\left(x_i\right)\right)^2 + \lambda \int \left(m''\left(x\right)\right)^2 dx \right\} \text{[3]} \tag{20}$$

The parameter $\lambda > 0$ is a smoothing parameter which controls the trade-off between smoothness and goodness of fit to the data. If $\lambda \to \infty$ the minimization of (21) gives a linear fit whereas letting $\lambda \to 0$ gives a wiggly function. The larger the value of λ, the more the data will be smoothed to produce the curve estimate. However, the basic underlying idea of penalising a measure of goodness of fit by one of roughness was described by [16].Equation (21) shows that the function to be minimized consists of two components: first, the deviation of the fitted function from the observed values should be minimized which gives the goodness of the fit. Second, complex functions are penalised by the second term in (21), as measured by the second order derivative. From [3] and from the quadratic nature of equation (21), the spline smoother $\hat{m}\left(x\right)$ is linear in the observations y_i in the sense that there exists a weight function $G\left(s,x\right)$ such that

$$\hat{m}\left(x\right) = \frac{1}{n}\sum\nolimits_{i=1}^{n} G\left(s,x_i\right)y_i \tag{21}$$

Where,

$$G\left(s,x\right) = \frac{1}{f\left(x\right)}.\frac{1}{n\left(x\right)}K\left(\frac{s-x}{n\left(x\right)}\right) \tag{22}$$

with the Kernel function K given by

$$K\left(u\right)=\frac{1}{2}\exp\left(-\mid u\mid/\sqrt{2}\right)\sin\left(\mid u\mid/\sqrt{2}+\pi/4\right)[14] \quad (23)$$

and the local bandwidth $h\left(x\right)$ satisfies

$$h\left(x\right)=\lambda^{\frac{1}{4}}n^{-\frac{1}{4}}f\left(x\right)^{-\frac{1}{4}} \quad (24)$$

It has been assured that n is large and that the design points have local density $f\left(x\right)$, in that the proportion of x_i in an interval of length dx near x is approximately $f\left(x\right)dx$. Equation (23) above applies for large n provided s is not too near the edge of the interval on which the data lie, and λ is not too big or too small.

After obtaining the spline smoother $\hat{m}\left(x\right)$ in equation (22), we then can substitute this value in the equation (16) to obtain the population total as from $\hat{m}\left(x\right)=\frac{1}{n}\sum_{i=1}^{n}G\left(s,x_i\right)y_i$ and

$$G\left(s,x\right)=\frac{1}{f\left(x\right)}.\frac{1}{h\left(x\right)}k\left(\frac{s-x}{h\left(x\right)}\right)$$

$$\hat{m}\left(x\right)=\frac{1}{n}\sum_{i=1}^{n}y_i\frac{1}{f\left(x\right)}.\frac{1}{h\left(x\right)}k\left(\frac{s-x}{h\left(x\right)}\right)$$

substituting in $\hat{T}_{np}=\sum_{i\in s}y_i+\sum_{i\notin s}\hat{m}\left(x_i\right)$

we get the smoothing spline estimator of the population total, $\hat{T}_{ss}$ as

$$\hat{T}_{ss}=\sum_{i\in s}y_i+\sum_{i\notin s}\frac{1}{n}\sum_{i=1}^{n}\frac{1}{f\left(x\right)}.\frac{1}{h\left(x\right)}k\left(\frac{s-x}{h\left(x\right)}\right)y_i \quad (25)$$

While the periodic Spline Estimator of the Population Total $\hat{T}_{ps}$ is obtained as

$$\hat{T}_{ps}=\sum_{i\in s}y_i+\sum_{i\notin s}\left\{\frac{\sum_{i=1}^{n}G\left(s-\frac{x_i}{n}\right)y_i}{\sum_{i=1}^{n}G\left(s-\frac{x_i}{n}\right)}\right\} \quad (26)$$

3. Empirical Study

We present the analysis and results of the five estimators i.e. the ratio, the local polynomial, the Nadaraya-Watson Kernel, the spline smoother and the periodic spline. We used four natural and two artificial populations in the study.

3.1. Description of the Study Populations

In artificial population I, we generated 100 data points according to the linear homoscedastic model:

$y_i=0.25x_i+\ell_i$ with $\ell_i\sim N\left(0,1\right)$ and $x_i\sim U[0,1]$

In artificial population II, we again generated 100 data points according to the quadratic homoscedastic model:

$y_i=0.5+0.25x_i+1.5x_i^{\,2}+\ell_i$ with $\ell_i\sim N\left(0,1\right)$

$x_i\sim U[0,1]$

We obtained the natural populations from the Kenya Central Bureau of Statistics ofbetween 2006and 2014. The description of each of the populations is given in the table 3.1 below.

Table 3.1. *Description of the four natural populations.*

Population	Data Points	Description	
		y_i	x_i
I	100	Value (in millions) of Road Transport equipment Imported.	Quantity (number) of Road Transport Equipment Imported.
II	126	Value in thousands of principle articles traded.	Quantity (units) of principle Articles Traded.
III	130	Total number of employees engaged per industry.	Total number of firms and Establishments per industry.
IV	130	Total outputs per industry in a manufacturing sector.	Total inputs per Industry in the manufacturing sector.

Scatter plots drawn for each of the four natural populations (Population I-IV) were used to deduce the form of the population structures as below:

Population I: the structure of the population could be non-linear and heteroscedastic

Population II: the structure of the population could be linear and heteroscedastic.

Population III: the structure of the population could be linear and heteroscedastic

Population IV: the structure of the population could be linear and homoscedastic.

Population V and IV were the artificial populations with known population structures:

Population V: is of a linear homoscedastic model and passing through the origin**.**

Population VI: is of a quadratic homoscedatic model.

3.2. Design of the Study

For each of the six populations, 500 samples of size 50 were drawn by Simple Random Sampling without replacement. The Epanechnikov Kernel defined as

$$K\left(u\right)=\begin{cases}\frac{3}{4}\left(1-u^2\right) & if \quad \mid u\mid\leq 1\\ 0 & Otherwise\end{cases}$$

was used in the study for the Local Polynomial Estimator and the Nadaraya-Watson Kernel Estimator. An optimal bandwidth for Nadaraya-Watson smoother within the interval $\left[\frac{\sigma}{4n^{\frac{1}{5}}} \le h \le \frac{3\sigma}{2n^{\frac{1}{5}}}\right]$ was sought where σ is the standard deviation of x_i's. The Kernel function used in the spline smoothing and periodic spline is $K(u) = \frac{1}{2}\exp\left[-\frac{|u|}{\sqrt{2}}\right]\sin\left[\frac{|u|}{\sqrt{2}} + \frac{\pi}{4}\right]$ [14], with the local bandwidth $h(x)$ satisfying $h(x) = \lambda^{\frac{1}{4}} n^{-\frac{1}{4}} f(x)^{-\frac{1}{4}}$

3.3. Description of the Computation Procedure

For each of the six populations, we computed the true population total $T = \sum_{i=1}^{N} y_i$, where N is the number of units in each population. The estimator of population total $\hat{T}_{ik}$, was then obtained for each population using the five different estimators as follows;

Ratio Estimator: $\hat{T}_{re} = \frac{\overline{y}(s)}{\overline{x}(s)}\overline{x}$ Local polynomial:

$$\hat{T}_{lp} = \sum_{i\in s}\frac{y_i - w_{si}^{'} y_s}{\pi_i} + \sum_{i=1}^{N} w_{si}^{'} y_s$$

Nadaraya-Watson: $\hat{T}_{nw} = \sum_{i\in s} y_i + \sum_{i\notin s}\left\{\frac{\sum_{i=1}^{n} k_n(x_i - x) y_i}{\sum_{i=1}^{n} k_h(x_i - x)}\right\}$

Smoothing Spline:

$$\hat{T}_{ss} = \sum_{i\in s} y_i + \sum_{i\notin s}\frac{1}{n}\sum_{i=1}^{n}\frac{1}{f(x)}\cdot\frac{1}{h(x)} k\left(\frac{s-x}{h(x)}\right) y_i$$

Periodic Spline: $\hat{T}_{ps} = \sum_{i\in s} y_i + \sum_{i\notin s}\left\{\frac{\sum_{i=1}^{n} G\left(s - \frac{x}{n}\right) y_i}{\sum_{i=1}^{n} G\left(s - \frac{x}{n}\right)}\right\}$

To compare the five estimators, the average biases and the average Mean Square Errors (MSE) for each population were calculated. For population five and six, the relative change in efficiency was calculated to measure the robustness of the estimators. The Average Bias for each estimator was calculated as;

Average Bias $\left(\hat{T}_k\right) = \sum_{i=1}^{500}\frac{\hat{T}_{ik} - T}{500}$ where k denotes the different estimators.

The Average Mean Square Error for each estimator was obtained from

Average MSE $\left(\hat{T}_k\right) = \sum_{i=1}^{500}\frac{\left(\hat{T}_{ik} - T\right)^2}{500}$.

The Relative change in efficiency (RCE) for each estimator was given by

$$\text{RCE} = \frac{Efficiencyinpopulation(6) - efficiencyinpopulation\,(5)}{efficiencyinpopulation\,(5)}$$

3.4. Results

The results of this study were summarized in Tables 3.2, 3.3, 3.4, 3.5 and 3. 6 below:

Table 3.2. *True Population Totals.*

	Pop 1	Pop 2	Pop 3	Pop 4	Pop 5	Pop 6
Population Sums	131.002	598.124317	510.177	178.7683	12.18925	111.4207

Table 3.3. *Estimates of Population Totals.*

	Pop 1	Pop 2	Pop 3	Pop 4	Pop 5	Pop 6
Nadaraya-Watson	135.4742	617.397269	509.3305	186.1202	11.53005	111.2965
Smoothing Spline	90.64416	1836.954517	317.3828	295.2271	22.32936	211.1834
Local Polynomial	131.8575	484.6628646	395.1619	139.7997	12.23147	113.7409
Ratio Estimator	163.0781	623.9877722	534.3458	188.1737	16.65574	152.2789
Periodic Spline	129.4973	598.0745695	449.508	173.8463	11.23823	104.4373

Table 3.4. *Average Bias.*

Nadaraya-Watson	4.472212	19.27295201	-0.84652	7.351878	-0.6592	-0.12418
Smoothing Spline	-40.3578	1238.8302	-192.794	116.4588	10.1401	99.7627
Local Polynomial	0.855476	-113.461452	-115.015	-38.9686	0.042218	2.320245
Ratio Estimator	32.07607	25.86345519	24.16878	9.405401	4.466489	40.85822
Periodic Spline	-1.5047	-0.0497475	-60.669	-4.92197	-0.95103	-6.98339

Table 3.5. *Average Mean Square Error.*

Nadaraya-Watson	372.2935	19113.57612	3152.551	508.5213	0.965757	25.60268
Smoothing Spline	2157.35	1714144.509	43854.85	16611.31	103.8781	9994.21
Local Polynomial	4168.499	109812.6846	6061.498	1345.238	20.49219	1731.601
Ratio Estimator	332.4187	24818.69519	15134.26	1791.341	0.568978	31.90509
Periodic Spline	2448.407	18474.74869	110964	675.3513	1.418111	70.75136

Table 3.6. *Relative Change in Efficiency (RCE).*

Estimator	Nadaraya-Watson	Smoothing spline	Local polynomial	Ratio Estimator	Periodic spline
RCE	25.51048	95.21094	83.50053	55.07438	48.89127

3.5. Discussion of the Results

For population I which is approximately non-linear and heteroscedastic, the bias of local polynomial estimator $\left(\hat{T}_{lp}\right)$ is the smallest compared to the rest, making it the best estimator for this population. Periodic spline has the smallest bias for population II which is approximately linear and heteroscedastic. On the other hand, Nadaraya-Watson has the lowest bias for population III which is also approximately linear and heteroscedastic. In population four (approximately linear and homoscedastic), we notice that the periodic spline has the lowest bias, hence becoming a good estimator for this population. Table 3.4 shows that generally all the estimators have low biases in population V compared to the rest of the populations. The lowest bias however is of the local polynomial estimator which makes it a good estimator for the linear homoscedastic model. We further notice that Nadaraya-Watson estimator has the smallest bias in population VI, making it the best estimator for the non-linear homoscedastic model.

We next consider the performance of each estimator across the six populations in terms of average biases as shown in table 3.4. The Nadaraya-Watson estimator performed relatively well in all the populations. It, however, did best in populations three and six which are linear and heteroscedastic and quadratic and homoscedastic respectively. The smoothing spline on the other hand, had the largest bias in all the populations. It had its best performance with a linear homoscedastic population. For the Local polynomial estimator, we notice that it had the lowest bias in population one which is linear and heteroscedastic and population five which is linear and homoscedastic. Its bias in population six, which is quadratic and homoscedastic, is also relatively low. When it comes to Ratio Estimator, we notice that generally its performance is low compared to the other estimators but better than the smoothing spline. Its best performance is in population three which is approximately linear and heteroscedastic.

Then we moved on to the Average Mean Square Error (AMSE) in table 3.5. The smaller the AMSE, the higher the efficiency of the estimator for the given population. In population I, the lowest AMSE was given by the Ratio Estimator while in population II, it was the periodic spline. Nadaraya-Watson had the lowest AMSE in population III and IV while for Population V it was the Ratio Estimator. On the other hand, the Nadaraya-Watson was the most efficient estimator for the non-linear homoscedastic population VI.

Finally, we compared the Relative Change in Efficiency (RCE) among the five estimators. We noticed from table 3.6 that the Nadaraya-Watson had the lowest RCE. The implication here was that it is the least sensitive to the change of structure of the population and hence the most Robust among the five estimators. It was then followed by the Periodic Spline, the Ratio Estimator and the Local polynomial. The Smoothing Spline was the least Robust among them.

4. Summary, Conclusions and Recommendations

4.1. Summary of the Findings

The research set out to estimate population total using spline functions. However, other estimators of the population total were also involved for comparative purposes. In all the six populations considered, the Periodic spline had a smaller average bias, had less average AMSE and was found to be more robust than the Smoothing Spline. The Nadaraya-Watson estimator performed generally well in terms of the average bias, efficiency and robustness. It had very small biases in both linear and non-linear homoscedastic models. The bias in heteroscedastic models was also relatively low. Its efficiency was equally higher in most of the populations and it also had the lowest RCE value out of the five estimators considered.

The local polynomial estimator was found to be almost unbiased for a linear homoscedastic model. Its bias however goes up when a non-linear homoscedastic population is considered. In terms of efficiency, the estimator is far more efficient in a linear homoscedastic model than a non-linear one. It has a high RCE value.

We observed that this estimator is relatively highly biased across the six populations considered. However in terms of efficiency, it was the most efficient of the five estimators for a linear homoscedastic model. The efficiency went down when a non-linear homoscedastic population was considered. The RCE value is relatively high. We also observed that the periodic spline and the Nadaraya-Watson estimators gave results that were quite similar in terms of bias, efficiency and robustness.

4.2. Conclusions and Recommendations

We observed from this study that the two spline functions considered perform quite differently. The periodic spline performed better than the smoothing spline in all the aspects considered: bias, efficiency and robustness. We, therefore,

concluded that the periodic spline is a better estimator than the smoothing spline in a case of a linear homoscedastic model and even when the model assumptions have been violated. It was also shown that the Nadaraya-Watson estimator performed well in the linear homoscedastic model and also when the conditions were violated. It had the lowest RCE value. Therefore, we came to the conclusion that, Nadaraya-Watson estimator was the most robust of the five estimators. The results also showed the periodic spline and the Nadaraya-Watson estimators to be quite similar. Thus, we concluded from both the theoretical results and the empirical study that spline smoothing corresponds approximately to smoothing by a Kernel method thus concurring with the theoretical observation made by [13].

The local polynomial estimator was very sensitive to model assumption violation and we therefore concluded that it is not robust. The results also indicated that the radio estimator was the most efficient of the five estimators for a linear homoscedastic model. Nevertheless, when these conditions are violated, the estimator completely breaks down. We conclude that this estimator is not robust to the violation of the linear and homoscedastic conditions.

From the findings of the study, we gave the following recommendations:

1. Both the ratio estimator (model-based) and the local polynomial (model -assisted) estimator should be used within the confines of a linear homoscedastic model. They are not appropriate for use when the model is unspecified or when the linear and homoscedastic assumptions are violated.
2. The Nadaraya-Watson and the periodic spline estimators, both of which are non-parametric, should be used in case of a linear and homoscedastic model and even when the model assumptions are violated. Their sensitivity to the change of structure of the population is relatively low and hence are highly robust. The Nadaraya-Watson, however, is even more robust than the periodic spline.

References

[1] Aerts, M., Augustyns, I. and Janssen, P., "Smoothing Sparse Multinomial Data Using Local Polynomial Fitting," *Journal of Nonparametric Statistics*, 8, 127-147, 1997.

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

[3] Cardot, H., "Local Roughness Penalties for Regression Splines,"*Computational Statistics*, 17, 89-102, 2002.

[4] Fuller, W.A, *Sampling Statistics*, Wiley, Hoboken, 2009.

[5] Harms, T. and Duchesne, P., "On Kernel Non- Parametric Regression Designed for Complex Survey", *Metrika*, 72 (1), 111-138, July2010.

[6] Kauermann, G., Krivobokova, T. and Fahrmeir, L., "Some Asymptotic Results on Generalized Penalized Spline Smoothing," *J. R. Statistic. Soc.*Series B, 71, 487-503, 2009.

[7] Lu, J. and Yan, Z., "A Class of Ratio Estimators of a Finite Population Mean Using Two Auxiliary Variables,"PLoS ONE 9(2): e89538.doi:10.1371/journal.pone.0089538, 2014.

[8] Nadaraya, E. A., "On Estimating Regression," Jour. Theory Probab. Appl.9 (1), 141-142, 1964.

[9] Royall, R. M., "Likelihood Functions in Finite Population Sampling Theory,"*Biometrika,* 63, 605-614, 1976.

[10] Sarda, P. and Vieu, P., *Kernel Regression in Smoothing and Regression: Approaches Computation and Application*, Ed M. G. Schimek, Wiley Series in Probability and Statistics, 2000, 43-70.

[11] Schoenberg, I. J., "Spline Functions and the Problem of Graduation,"*Proc. Nat. Acad. Sci. U.S.A.*, 52, 947-950, 1946.

[12] Schumaker, L. L., *Spline Functions: Computational Methods*, SIAM, Philadelphia, 2015.

[13] Silverman, B. W., "Spline Smoothing: The Equivalent Variable Kernel Method," *The Annals of Statistics*, 12(3), 898-916, 1984.

[14] Wahba, G., "Smoothing Noisy Data with Spline Functions,"*Numerische Mathematik*, 24, 383-393, 1975.

[15] Watson, G. S., "Smooth Regression Analysis,"*Sankhya Ser. A*.,26, 359-372, 1664

[16] Whittaker, E., "On a New Method of Graduation,"*Proc. Edinburgh Math. Soc.*, 41, 63-75, 1923.