**Science Journal of Applied Mathematics and Statistics**
2015; 3(5): 214-224
Published online August 31, 2015 (http://www.sciencepublishinggroup.com/j/sjams)
doi: 10.11648/j.sjams.20150305.11
ISSN: 2376-9491 (Print); ISSN: 2376-9513 (Online)

SciencePG
Science Publishing Group

# Spline Regression in the Estimation of the Finite Population Total

**Joseph Kipyegon Cheruiyot**

Department of Computer and Statistics, Moi University, Eldoret, Kenya

**Email address:**
jcheruiyot77@yahoo.com

**To cite this article:**
Joseph Kipyegon Cheruiyot. Spline Regression in the Estimation of the Finite Population Total. *Science Journal of Applied Mathematics and Statistics*. Vol. 3, No. 5, 2015, pp. 214-224. doi: 10.11648/j.sjams.20150305.11

**Abstract:** This study sought to estimate finite population total using Spline regression function. It compared the Spline regression with Sample Mean estimator, design-based and model - based estimators. To measure the performance of each estimator, the study considered average bias, the efficiency by use of the mean square error and the robustness using the rate change of efficiency. In this research, five populations were used. Three of them were simulated according to the following models: linear homoscedastic, quadratic homoscedastic and linear heteroscedastic and two natural populations. The performances of the five estimators were studied under the five populations. The sudy found that Sample Mean(SM), Horvitz-Thompson (HT) and Ratio (R) estimators are not robust while Nadaraya-Watson(NW) and Periodic Spline(PS) are robust when linearity and homoscedasticity of the population structure are violated.

**Keywords:** Homoscedasticity, Population, Sample, Spline Regression, Robustness, Smoothing, Estimator

## 1. Introduction

### *1.1. Introduction*

There are two generally accepted options in studying the characteristics of finite population. The first option is a study in which every unit of the population is examined called a census. Use of a census to study a population is time consuming, expensive, often impossible and strangely enough, often inaccurate. The other option is to study the characteristic of a population by examining a part of it. The theory of survey sampling as developed during the past several decades provides us with various kinds of reasonable scientific tools for drawing samples and making valid inference about the population parameters of interest.

#### *1.1.1. Census Versus Sampling Method*

Although there are advantages with the census method, the cost, effort and the time required to conduct census may be enormous unless the population is very small. In such a case we resort to sampling that involves examination of a part of the population. Although a census operation gives a more reliable data, sampling is more appropriated when:

i. The cost of conducting census would be prohibitive.
ii. The population is large, such that it would be impossible to conduct a census.
iii. The study involves destruction of elementary units under study, such that it would be appropriate to conduct sample testing.
iv. Quick results are required, such that it would be appropriate to conduct sample survey rather than carrying out a complete count.

#### *1.1.2. Basic Ideas of Sampling and Estimation*

In the basic sampling setup, the population consists of a known finite number N of units – such as people or plots. With each unit is associated a value of a variable of interest, sometime referred to as the y-value of that unit. The y-value of each unit in the population is unknown quantity. However, the units in the population are identifiable and may be labeled with numbers 1, 2,. N. A sample of the units in the population is selected and observed. The data collected consist of the y-value for each unit in the sample together with the unit's label. The procedure by which the sample units is selected from the population is called the sampling design. With most of the well- known sampling designs, the design is determined by assigning to each possible sample the probability p(s) of selecting that sample. For example, using the simple random sampling design, the units are selected with equal and independent probability p(s).

### *1.2. Estimation Approaches*

To estimate finite population total ($Y_T$) in survey, where $Y_T = \sum_{i=1} Y_i$,

We need to have Yi ( i =1, 2, 3,., N) the survey variables and xi ( i = 1, 2, 3,., N) design variables ( Auxiliary variables). The following therefore is a list of approaches that are considered in this study in the estimation of finite population total.

#### *1.2.1. Design - Based Approach*

This is also known as classical approach. In this approach, the variables of interest of the target population are viewed as fixed quantities. Also the design introduces selection probabilities that determine the properties of estimators that are used to obtain expected values, variances, biases etc. The samples are generated by sampling design p(s) with the values $y_1, y_2, \ldots, y_n$, $x_1, x_2, \ldots, x_n$ held fixed. The repetition of sample drawing procedure forms the basis of randomization framework. The approach assumes that models have no relevance to the inferential framework. In experimental design, randomization is employed to protect the experimenter against subjective biases. Scott and Smith (1975) extended results of Blackwell. According to Fisher, randomization was relevant before the data were collected but not in the analysis of data which is in agreement with most statisticians in the experimental sciences. Randomization is therefore an insurance against selection bias.

#### *1.2.2. Model – Based (Prediction) Approach*

From Royall (1976) the concept of the super population is introduced thus: “The finite population should itself be regarded as a random sample from some infinite population”. Hence finite population is assumed to be generated as a random sample from a super population. Also noted that variable of interest are viewed as random variables and properties of estimators depend on the joint distribution of these random variables. A sample is selected from the finite population using a known sampling scheme. Then observations are made on the sample values and are then used to make predictions about the non sample values. In this case the model connects a variable of interest Y with a set of auxiliary variables X, Cox (1995). However, noted that the choice of a model and it’s robustness to misspecification is the major issue. Small deviation from a chosen model may lead to serious errors in an inference. Sometimes the models become mathematically complex while still not being suitably realistic (Thompson, 1992). For example, where model assumption of the variable being studied is that of independence it ignores the tendency in many population for nearby or related units to be correlated.

#### *1.2.3. Non-Parametric Approach*

The parametric method of estimation is used when it is assumed that the data is drawn or generated from one of the known parametric family of distributions. In many cases however, the experimenter does not know the form of the basic distribution and needs statistical techniques which are applicable regardless of the form of the distribution. These techniques are referred to as non parametric or distribution free methods. They apply to very wide families of distributions rather than only to families specified by a particular functional form. They do not require the various assumptions about the distribution of population from which the sample was obtained. The main idea behind this class of models is that the effect of an explanatory (design) variable and dependent variable of interest is not modeled as parametric, usually linear function but is kept flexible. The only assumption needed is that the effects of the explanatory variables are modeled as smooth i.e. differentiable functions. The functional shape is then to be estimated from the data by either using: Kernel based methods or Spline based methods.

*Kernel Based Method.*

The Kernel estimator is expressed in terms of a Kernel function which satisfies the condition;

$$\left.\begin{array}{r} \int_{-\infty}^{\infty} K(x)dx = 1 \\ \int_{-\infty}^{\infty} xK(x)dx = 0 \\ \int_{-\infty}^{\infty} x^2K(x)dx = \delta^2 < \infty \end{array}\right\} \quad (1)$$

Usually, but not always, K will be a symmetric probability density function, the normal density for instance. Therefore, according to Silverman (1986) the Kernel estimator of the density function with Kernel K is defined by,

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right) \quad (2)$$

where h is the bandwidth. It is clearly observed that the Kernel estimator is a sum of ‘bumps’ placed at the observations. Each individual bumps is created by $(nh)^{-1}K\{(x_i - x)/h\}$ and the estimate $\hat{f}$ is a resultant hump obtained by adding them up.

*Spline Based Method.*

The name, “Spline function” was given by I.J Schoenberg (1946) to the piecewise polynomial function known as univariate polynomial Splines. 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 “duck” used to position the rods at points through which it was derived to draw smooth interpolation curves passing through prescribed points. The basic idea dates back at least to Whittaker (1923). More resent papers on the subject include Wahba (1975), Smith (1979), and Silverman (1985) among others. For Kernel regression estimation a weighting scheme due to Nadaraya (1964) –Watson (1964) has been associated with random design, and a convolution type weighting scheme with fixed design based on mean square error; none of the estimators is uniformly optimal in either design. The multitude of non parametric regression estimators is an issue of considerable practical and theoretical importance. A wide class of estimators studied by Jennen Steinmetz and Gasser (1988) included fixed width Kernel estimators, smoothing spline and nearest – neighbor estimators as particular cases. No estimator is uniformly best in terms of integrated mean squared error, but the kernel estimator turns out to be the minimax optimal. Since non parametric methods are usually intended to be applicable to a broad variety of situations the minimax property is an important safeguard. Two definitions of Kernel weights enjoy particular popularity, the Nadaraya –Watson type (Nadaraya 1964, Watson 1964) and the convolution type estimator (Priestly and Chao1972, Gasser and Muller 1979). The Nadaraya-Watson method is intuitively motivated as an estimator of a conditional expectation which suggests a context where the independent variable is random. Hence this method seems suited for a situation of randomly selected design points, whose distribution is determined by the design density.

A spline function is a piecewise defined function with certain smoothness conditions. The most commonly used form is the cubic splines. There are two sorts of splines; ordinary splines and B-spline. The two spline function have the same general structure regarding the piecewise defined function such as

$$f_i(x_i)=a_{i,0}+a_{i,1}x+a_{i,2}x^2+a_{i,3}x^3 \qquad (3)$$

and the smoothing conditions. The difference is that the ordinary splines go through all the data points exactly where as B -spline do not necessarily fit the data exactly. For ordinary splines, the curve has to go through all the points hence the equation $f_i(x)=y_i$ has to be satisfied for all the points$(x_i, y_i)$. The spline function has to yield the value $y_i$ for $x_i$. The smoothing conditions too have to be fulfilled. B-spline are piecewise defined functions usually polynomial with the same smoothness conditions as ordinary spline. They are however not forced through the data points exactly, the function has simply to come close to the data points.

In estimation of finite population total, the challenge is to identify an estimator that is efficient when the population structure is not known. In this study, try to compare the spline regression with the known estimators of nonparametric (Nadaraya-Watson), Sample Mean estimator, Design-based Horvitz-Thompson estimator and Model-based Ratio estimator. The challenge is to obtain an estimator which is robust to the violation of both linearity and homoscedasticity of the population structure.

## 2. Methodology

### *2.1. Non-Parametric Estimation of the Population Total Using Kernels*

In this section, the Nadaraya-Watson Kernel estimator is considered. 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.

Consider the model

$$\text{yi} = \text{m }(x_i) + \varepsilon \text{ I} \qquad (4)$$

where $\text{m}(x_i)$ is the mean function and $\varepsilon$ i a random error term. It is assumed that the functional form of m (xi) is unknown but assumed to be smooth and continuous.

Let wi(x), i = 1, 2, …, n be the weight function known as Kernel function. The Kernel is a continuous, bounded and symmetric function which integrates to one. That is

$$\int \text{k(u)du} = 1$$

By taking kh(u) = h-1k$\left(\frac{u}{h}\right)$ to be the Kernel with band width h. The weight sequences for the Kernel smoothers as given by Nadaraya (1964) - Watson (1964) is

$$\text{wi(x)} = \frac{k_h(x_i - x)/h}{\sum_{i=1} k_h(x_i - x)/h} \qquad (5)$$

The Nadaraya -Watson estimator of m(x) in (3.1) is

$$\hat{m}(x) = \sum_i w_i(x) y_i \qquad (6)$$

Substituting 3.2 in 3.3 we have

$$\hat{m}(x) = \sum_{i=1}^{n} \frac{k_h(x_i - x)/h}{\sum_{i=1}^{n} k_h(x_i - x)/h} y_i \qquad (7)$$

The shape of the Kernel weights is determined by K, where K is a symmetric probability density function that satisfies conditions in equations 1. One unique feature of the size of the bandwidth is that the smaller it is the more concentrated are the weights around x. However, the non-parametric regression based estimator T$_{np}$ for the population total T is given by

$$\hat{T}_{np} = \sum_{i=1}^{n} y_i + \sum_{i \notin s} \hat{m}(x_i) \qquad (8)$$

where $\hat{m}(x)$ is the Nadaraya-Watson estimator give in (7). Hence by substituting (7) in (8) Nadaraya – Watson estimator of the population total becomes:

$$\hat{T}_{nw} = \sum_{i \in s} y_i + \sum_{i \notin s} \left\{ \sum_{i=1}^{n} \frac{k_h(x_i - x)/h}{\sum_{i=1}^{n} k_h(x_i - x)/h} y_i \right\} \qquad (9)$$

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

### *2.2. Properties of Nadaraya-Watson Kernel Estimator of the Population Total*

The Nadaraya-Watson Kernel regression estimator is given as in (8) and (7)

In order to find a standard measure of estimation error, the Mean Square error (MSE), the study looked at the conditional mean and variance of $\hat{T}_{np} - T$ under the model $y_i = m(x_i) + \varepsilon$.

$$\hat{T}_{np} = \sum_{i\in s} y_i + \sum_{j\notin s}\hat{m}(x_j)$$

and $T = \sum_{i\in s} y_i + \sum_{j\notin s} y_j$

So that $\hat{T}_{np} - T = \sum_{j\notin s}\left[\hat{m}(x_j) - y_j\right]$

Thus $E\left[\hat{T}_{np} - T/X_p\right] = E\left[\sum_{j\notin s}\left\{\hat{m}(x_j) - y_j\right\}\right]$

where $X_p$ is the population vector of X-values.

But $\hat{m}(x_j) = \sum w_i(x_j)y_j = \dfrac{\sum_{i\in s} h^{-1}K\left(\frac{x_i-x_j}{h}\right)y_i}{h^{-1}\sum_{i\in s}k\left(\frac{x_i-x_j}{h}\right)}$

$$= \sum_{i\in s}K\left(\frac{x_i-x_j}{h}\right)\left[\hat{d}_s(x_j)\right]^{-1}(nh)^{-1}y_i$$

where $\hat{d}_s(x_j) = (nh)^{-1}\sum_{i\in s}K\left(\frac{x_i-x_j}{h}\right)$ is the standard Nadaraya-Watson estimator of the density $d_s(x_j)$.

Hence $E\left[\hat{T}_{np} - T/X_p\right] = E\left[\sum_{j\notin s}\left\{\hat{m}(x_j) - y_j\right\}\right]$

$$= E\left[\sum_{j\notin s}\left\{\sum_{i\in s}K\left(\frac{x_i-x_j}{h}\right)\left[\hat{d}_s(x_j)\right]^{-1}(nh)^{-1}y_i - y_j\right\}\right]$$

$$= E\left[\sum_{j\notin s}\left[\hat{d}(x_j)\right]^{-1}(nh)^{-1}\left\{\sum_{i\in s}K\left(\frac{x_i-x_j}{h}\right)y_i - \hat{d}_s(x_j)(nh)y_j\right\}\right]$$

$$= E\left[\sum_{j\notin s}\left[\hat{d}_s(x_j)\right]^{-1}(nh)^{-1}\left\{\sum_{i\in s}K\left(\frac{x_i-x_j}{h}\right)y_i - \sum_{i\in s}K\left(\frac{x_i-x_j}{h}\right)y_j\right\}\right]$$

$$= \sum_{j\notin s}\left[\hat{d}_s(x_j)\right]^{-1}(nh)^{-1}\sum_{i\in s}K\left(\frac{x_i-x_j}{h}\right)E\{y_i - y_j\}$$

Since $E(y) = m(x)$ under the model, we have;

$$E\left[\hat{T}_{np} - T/X_p\right]$$

$$= \sum_{i\in s}\left[\hat{d}_s(x_j)\right]^{-1}(nh)^{-1}\sum_{i\in s}K\left(\frac{x_i-x_j}{h}\right)\{m(x_i) - m(x_j)\}.$$

Next, we look at the conditional error variance;

$$V\left(\hat{T}_{np} - T/X_p\right) = \mathrm{Var}\sum_{j\notin s}\left[\left\{\sum_{i\in s}K\left(\frac{x_i-x_j}{h}\right)\left[\hat{d}_s(x_j)\right]^{-1}(nh)^{-1}y_i\right.\right.$$

$$= \sum_{j\notin s}\mathrm{Var}\left[\sum_{i\in s}K\left(\frac{x_i-x_j}{h}\right)\left[\hat{d}_s(x_j)\right]^{-1}(nh)^{-1}y_i - y_j\right]$$

Since $\mathrm{Var}Y = \delta^2(x)$ under the model, we have

$$V\left(\hat{T}_{np} - T/X_p\right) = \sum_{i\in s} w_i^2\delta^2(x_i) + \sum_{j\notin s}\delta^2(x_j)$$

where $w_i = \sum_{j\notin s}K\left(\frac{x_i-x_j}{h}\right)\left[\hat{d}_s(x_j)\right]^{-1}(nh)^{-1}$.

### *2.3. Spline Regression Estimator of the Population Total*

Wahba (1975) has shown that Kernel smoothing estimator $\hat{m}(x)$ is closely related to smoothing

Splines estimator when it is represented approximately as a linear function of the data values yi. Hence there exists a weight function F (z,xi) such that

$$\hat{m}(x) = \frac{1}{n}\sum_{i=1}^{n} F(z,x)y_i$$

where the function F(z,x) is defined as

$$F(z,x) = \frac{1}{f(x)}\frac{1}{h}k\left(\frac{z-x}{h}\right) \qquad (10)$$

Hence we have

$$\hat{m}(x) = \frac{1}{f(x)}\frac{1}{h}k\left(\frac{z-x}{h}\right)$$

Substituting $\hat{m}(x)$ in

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

We get the smoothing spline estimator of the population Total $\hat{T}_{ss}$ as

$$\hat{T}_{ss} = \sum_{i\in s} y_i + \sum_{i\notin s}\frac{1}{n}\sum_{i=1}^{n}\frac{1}{f(x)}.\frac{1}{h}k\left(\frac{z-x}{h}\right)y_i \qquad (11)$$

where K(u) is defined as

K(u) = 0.5 exp(-|u | /1.41)sin(( | u| /1.41) + π/4)

and the function K(u) has the following properties;

$$\left.\begin{array}{l}\int K(u)du = 1\\ \int u\,K(u)du = 0\\ \int K^2(u)du < \infty\\ \int x^2\,K(u)du = \int x^3\,K(u)du = 0\\ \text{and}\int x^4\,K(u)du = -1\end{array}\right\} \qquad (12)$$

We can see that the properties of the function K(.) above are similar to those given for the Kernel function but can take negative values as well. Hence the smoothing spline estimator corresponds approximately to a Kernel type estimator of order 4. Eubank (1988) has shown that if the function m(.) is assumed to be periodic then $\hat{m}(x)$ corresponds to a spline estimator with a fixed bandwidth parameter h and weights F(z,x) = hw(u/h) where h = $\lambda^{1/4}$ and w(u) = $\lambda w(u/\lambda)$

the estimator corresponding to the periodic spline is

$$\hat{m}(x)=\frac{1}{n}\sum_{i=1}^{n}hF(z,x_i)y_i$$

where F(.) is as defined in (10),hence giving

$$\hat{m}(x)=\frac{1}{n}\sum_{i=1}^{n}h.\frac{1}{f(x_i)}.\frac{1}{h}k\left(\frac{z-x_i}{h}\right)y_i \qquad (13)$$

Since $n^{-1}F(z,x)$ does not sum to one, we divide the weights by their sum and we denote the modified weights by $F_R(z,x_i)$, then

$$FR(z,xi) = F((z-x_i)/h)\Big/\sum_i F((z-x_i)/h) \qquad (14)$$

therefore the $F_m(z,x_i)$ periodic spline estimator of the function m(x)is given by;

$$\hat{m}_R(x)=\sum_{i=1}^{n}\frac{F((z-x_i)/h)}{\sum_i F((z-x_i)/h)}y_i \qquad (15)$$

Substituting (15) in (6)

We have the Periodic Spline Estimator of the population total $\hat{T}_{ps}$ as

$$\hat{T}_{ps}=\sum_{i\in s}y_i+\sum_{i\notin s}\left\{\sum_{i=1}^{n}\frac{F((z-x_i)/h)}{\sum_i F((z-x_i)/h)}y_i\right\} \qquad (16)$$

hence $E\left[\hat{T}_{ps}-T/X_p\right]=E\left[\sum_{j\notin s}\left\{\hat{m}_R(x_j)-y_j\right\}/X_p\right]=E\left[\sum_{j\notin s}\left\{\sum_{i=1}^{n}\frac{F((z-x_j)/h)}{\sum_{i\in s}F((z-x_i)/h)}y_i-y_j\right\}/X_P\right]$

Let $\hat{d}_F(x_i)=\sum_{i\in s}F((z-x_i)/h)$,

Then $E\left[\hat{T}_{ps}-T/X_p\right]=E\left[\sum_{j\notin s}\left\{\sum_{i=1}^{n}F((z-x_i)/h)\left[\hat{d}_F(x_i)\right]^{-1}y_i-y_j\right\}/X_P\right]=E\left[\sum_{j\notin s}\left[\hat{d}_F(x_i)\right]^{-1}\left\{\sum_{i=1}^{n}F\left(\frac{z-x_i}{h}\right)y_i-\hat{d}_F(x_i)y_j\right\}/X_P\right]$

$$=E\left[\left[\sum_{j\notin s}\left[\hat{d}_F(x_i)\right]^{-1}\right]\left\{\sum_{i=1}^{n}F\left(\frac{z-x_i}{h}\right)y_i-\sum_{i=1}^{n}F\left(\frac{z-x_i}{h}\right)y_j\right\}/X_P\right]=\left[\sum_{j\notin s}\left[\hat{d}_F\right]^{-1}\sum_{i=1}^{n}F\left(\frac{z-x_i}{h}\right)E\{y_i-y_j\}/X_P\right]$$

$$=\sum_{j\notin s}\left[\hat{d}_F\right]^{-1}\sum_{i=1}^{n}F\left(\frac{z-x_i}{h}\right)\{m_R(x_i)-m_R(x_j)\}$$

Next, we consider the conditional error variance;

$$V\left(\hat{T}_{ps}-T/X_p\right)=\mathrm{Var}\sum_{j\notin s}\left[\left\{\sum_{i\in s}F\left(\frac{z-x_j}{h}\right)\left[\hat{d}_F(x_i)\right]^{-1}y_i-y_j\right\}/X_P\right]=\sum_{j\notin s}\mathrm{Var}\left[\left[\sum_{i\in s}F\left(\frac{z-x_i}{h}\right)\left[\hat{d}_F[x_i]\right]^{-1}y_i-y_j\right]/X_P\right]$$

let $W_i=\sum_{i\in s}F\left(\frac{z-x_i}{h}\right)\left[\hat{d}_F(x_i)\right]^{-1}$ hence $V\left(\hat{T}_{ps}-T/X_p\right)=\sum_{j\notin s}W_i^2\delta^2(x_i)+\sum_{j\notin s}\delta^2(x_j)$.

# 3. Empirical Results and Discussion

To compare the performance of the five estimators, that is, Horvitz Thompson, the Ratio estimator, Sample Mean estimator, the Nadaraya - Watson Kernel estimator and periodic spline estimator as spline regression estimator so as to identify a robust estimators, the study simulated three populations based on the following models; linear Homoscedastic model, Quadratic Homoscedastic model and Linear Heteroscedastic model. Also the study used two real populations. The criteria for comparing these estimators are average bias, mean square error and the rate of change of efficiency as a measure of robustness.

## *3.1. The Choice of the Kernel and Bandwidth*

This study used the Gaussian Kernel in Nadaraya - Watson estimator of the population total which is defined as

$$K(u)=\frac{1}{\sqrt{2\pi}}e^{\frac{-1}{2}(u)^2},-\infty<u<\infty$$

where $u=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x-x_i}{h}\right)$. Assume that the Kernel function K satisfies the conditions given in equation 1. An optimal bandwidth for Nadaraya-Watson smoother was chosen within the interval $\left[\frac{\delta}{4n^{1/5}}\le h\le\frac{3\delta}{2n^{1/5}}\right]$ where $\delta$ is the standard deviation of $x_i$'s, ( Silverman, 1986). Therefore, the bandwidth h used was chosen to be the centre point h= 7/8 $\delta n^{-1/5}$. The Kernel function used in the periodic spline is K (u) = 0.5 exp(-|u | /1.41)sin(( | u| /1.41) + π/4) (Wahba, 1975)

### *3.2. Description of the Study Population and Estimators*

The artificial population was simulated in the following manner.

a) In artificial population I, 76 data points were generated according to the model;

$$Y_i = 1 + \alpha\left(x_i - \frac{1}{4}\right) + (\varepsilon_i) \text{where } \varepsilon_i \approx N(0, \delta^2),\ x_i \approx U[0,1] \text{ and}$$

$$\alpha = 0.5.$$

b) In artificial population II, we again generated 76 data points according to the model

$$Y_i = 1 + \alpha\left(x_i - \frac{1}{4}\right)^2 + \varepsilon_i .$$

c) In artificial population III, once more 76 data points were generated according to the model

$$Y_i = 1 + \alpha\left(x_i - \frac{1}{4}\right) + \varepsilon_i x_i . \text{ Where } x_i \ \ , \alpha \text{ and } \varepsilon_i \text{ in b and c}$$

are the same as in population I

d) The Real population IV, was obtain from the Kenya National Bureau of Statistics (KNBS) for the population census done in Kenya in 2009. In this population, i considered the Auxiliary variable Xi to be the number of households in the ith District and study variable Yi the total population by District except for Nairobi province where Divisions are used instated of Districts, where i = 1,2,., 76. Our variable of interest Y is the population total.

e) Population V, this population has variable X describing shares a customer already possessed (Acquired) versus shares applied for (Booked) in a stock exchange brokerage farm, variable Y, both expressed in Kshs. Again i selected 76 data points in this population. The average bias and Mean Square Error of the population total were computed for each of the following five estimators: Sample Mean $\hat{Y}_{SM}$, Horvitz-Thompson $\hat{Y}_{HT}$ , Ratio estimator $\hat{Y}_{R}$ , Nadaraya-Watson $\hat{T}_{nw}$ and periodic spline $\hat{T}_{ps}$.

Below is a summary of the formulae used in computing their respective population total.

The following are scatter diagrams showing the distributions of the five populations mentioned above.

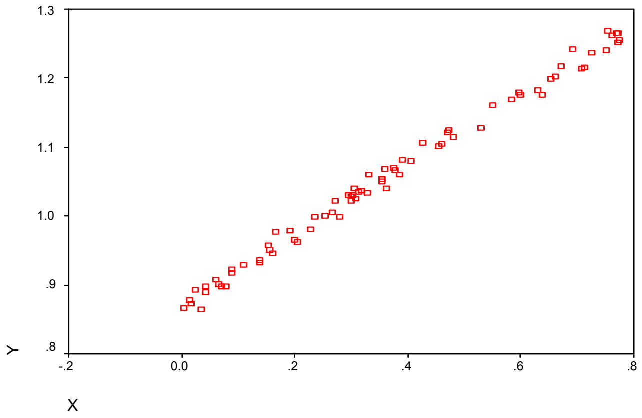

***Figure 1.*** *The population is linear with homoscedastic variance structure.*

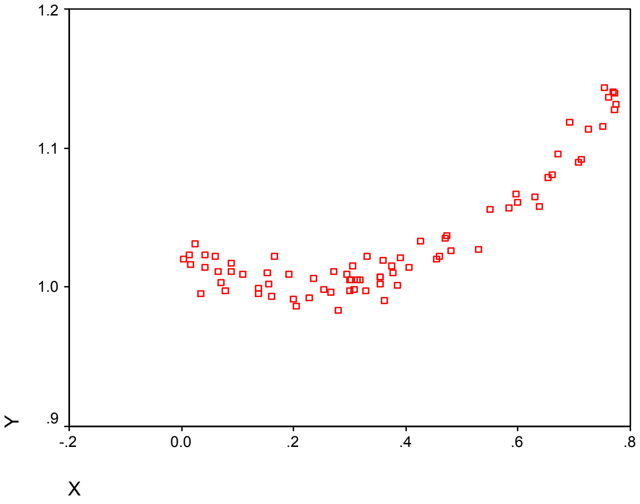

***Figure 2.*** *The population is quadratic with homoscedastic variance structure.*

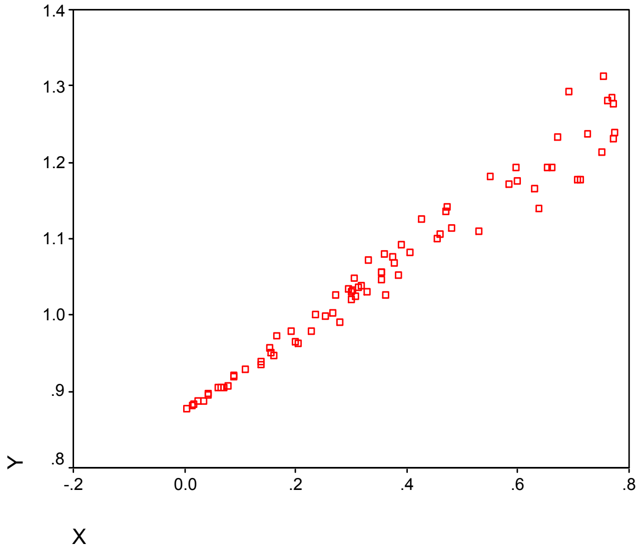

***Figure 3.*** *Linear population with heteroscedastic variance structure.*

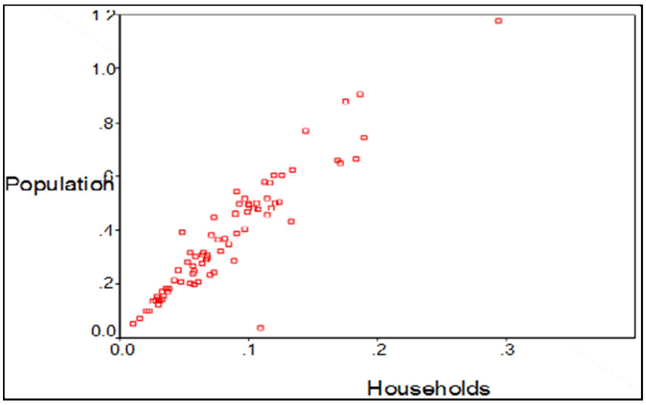

***Figure 4.*** *Kenya population census- 2009 (in millions).*

This population appears to be linear with heteroscedastic variance structure.

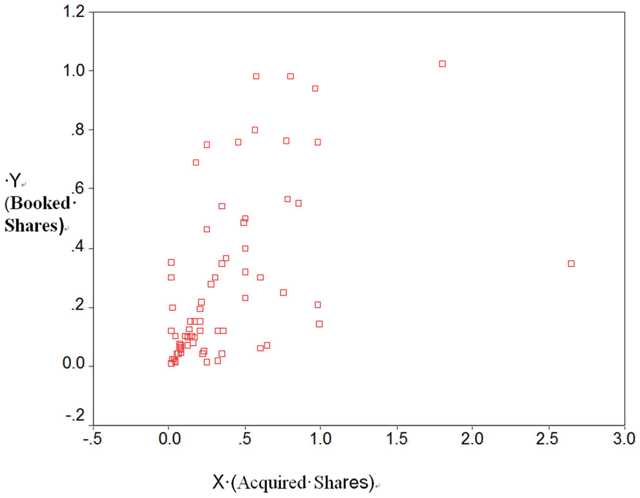

***Figure 5.*** *Stock Exchange Shares (in millions Ksh).*

The acquired shares and booked shares in this population structure appear to be uncorrelated.

### *3.3. Description of the Computation Procedure*

For each artificial population of size 76, samples of size n = 40 were generated by simple random sampling without replacement and 30 replicate samples were selected and estimates computed. Similarly, for the real population of size 76, samples of each size 40 were replicated 30 by SRSWOR and the estimators of the population total computed. For the case of Horvitz-Thompson, the sample units xi's are selected with unequal probabilities. To select a sample with unequal probabilities with Horvitz-Thompson weights, we have $\pi_i$ , the probability of the unit i being included in the sample such that $\pi_i = p_i\left[S+1-\frac{p_i}{1-p_i}\right]$

where $p_i = \frac{x_i}{\sum_{i=1}^{n} x_i}$ *and* $S = \sum_{i=1}^{n}\frac{p_i}{1-p_i}$ . Hence the estimate of the population total is obtained as $\hat{Y}_{\mathrm{HT}} = \sum_{i=1}^{n}\frac{y_i}{\pi_i}$ . For each of the population, we compute the true population total $\mathrm{Y} = \sum_{\mathrm{i=1}}^{\mathrm{N}} \mathrm{y_i}$ .

Define $\hat{Y}_r$ as the population total estimator, where r = SM, R, HT, NW, and PS. Then $\hat{Y}_r = \frac{\sum_{i=1}^{R}\hat{Y}_{ir}}{R}$ where $\hat{Y}_{ir}$ is population total estimate of the ith sample and rth estimator while R is the number of sample replicates.Hence the bias of each estimator of populations total were computed as $Bias\left(\hat{Y}_r\right) = E\left[\hat{Y}_r - Y\right]$ Thus the average bias for each estimator for both the real and artificial population totals are

$$Bias\left(\hat{Y}_r\right)_{population(k)} = \frac{\sum_{i=1}^{30}\left(\hat{Y}_{ir} - Y\right)}{30}$$

where k = 1, 2, 3, 4, 5.

We define the mean square error to be $MSE\left(\hat{Y}_r\right) = Var\left(\hat{Y}_r\right) + \left[Bias\left(\hat{Y}_r\right)\right]^2$

where $Var\left(\hat{Y}_r\right)$ is the unconditional variance of the estimator over the 30 replicates for the artificial and natural populations. Therefore, the Mean Square Error in the estimation of both the artificial and natural populations is given by:

$$MSE\left(\hat{Y}_r\right)_{population(k)} = \frac{\sum_{i=1}^{30}\left[\hat{Y}_{ir} - Y\right]^2}{30} + \left[Bias\left(\hat{Y}_r\right)_{populaton(k)}\right]^2$$

The Relative Change in Efficiency (RCE) for each estimator was given by

$$RCE(\mathrm{j}) = \frac{MSE\left(\hat{Y}_r\right) in\ pop(\mathrm{j}+1) - MSE\left(\hat{Y}_r\right) in\ pop\,1}{MSE\left(\hat{Y}_r\right) in\ pop 1}$$

Where j = 1,2,3,4.

### *3.4. Results and Interpretations*

The results of this study are summarized in Tables 1to 5. On each population the performance of each estimator is analyzed using the average bias and mean square error. The average bias is an indication of the measure of how closed an estimator is from the true value, while the MSE is used to assess efficiency of an estimator. For example an estimator will be said to be more efficiency than another, if its MSE is comparably smaller i.e if MSE ($T_1$) < MSE ($T_2$), where $T_1$ and $T_2$ are estimators, then $T_1$ is said to be more efficient than $T_2$.

***Table 1.** Summary of the formulae used in computing their respective population total.*

| Estimator | Formula |
|---|---|
| Sample Mean(SM) $\overset{\wedge}{Y}_{SM}$ | $\overset{\wedge}{Y}_{SM} = N\bar{y}$ |
| Horvitz-Thompson(HT) $\overset{\wedge}{Y}_{HT}$ | $\overset{\wedge}{Y}_{HT} = \sum_{i \in s} \frac{y_i}{\pi_i}$ |
| Ratio(R) $\overset{\wedge}{Y}_{R}$ | $\overset{\wedge}{Y}_{R} = \frac{\bar{y}}{\bar{x}} X$ |
| Nadaraya-Watson(NW) $\overset{\wedge}{T}_{nw}$ | $\overset{\wedge}{T}_{nw} = \sum_{i \in s} y_i + \sum_{i \notin s} \left\{ \sum_{i=1}^{n} \frac{k_h(x_i - x)/h}{\sum_{i=1}^{n} k_h(x_i - x)/h} y_i \right\}$ |
| Periodic-Spline(PS) $\overset{\wedge}{T}_{ps}$ | $\overset{\wedge}{T}_{ps} = \sum_{i \in s} y_i + \sum_{i \notin s} \left\{ \sum_{i=1}^{n} \frac{F((z - x_i)/h)}{\sum_{i} F((z - x_i)/h)} y_i \right\}$ |

***Table 2.** Population I (Linear and homoscedastic).*

| | SM | HT | Ratio | NW | PS |
|---|---|---|---|---|---|
| Estimate | 82.38245 | 82.38245 | 83.2557 | 82.17422 | 82.61311 |
| Bias | 2.090872 | 2.09187 | 2.96412 | 1.88264 | 2.32153 |
| Var | 16.06147 | 59.24723 | 49.725193 | 17.95833 | 19.84259 |
| MSE | 20.43321 | 63.61897 | 58.511200 | 21.50266 | 25.23209 |

Population Total 80.29158

In population I, i noted that from the low values of the bias that all the five estimators perform well under these conditions. However, Nadaraya-Watson has the least bias followed by SM, Horvitz-Thompson, Periodic spline and Ratio estimator in that order. Looking at MSE of this population, SM estimator has the lowest MSE, followed by Nadaraya-Watson, periodic spline, and Ratio. H-T estimator has the highest MSE in this population. However, the values of the MSE of these estimators on this population are lowest as compared to those obtained in the other populations. This implies that these estimators have high efficiency in linear and homoscedastic population structure. Though the sample mean with the least MSE is the most efficient in this population.

***Table 3.** Population II (Quadratic and homoscedastic).*

| | SM | HT | Ratio | NW | PS |
|---|---|---|---|---|---|
| Estimate | 80.50299 | 88.61431 | 88.28095 | 82.24362 | 82.75191 |
| Bias | 1.99378 | 10.1051 | 9.77174 | 3.73441 | 4.2427 |
| Var | 308.0725 | 1871.043 | 2037.5316 | 35.25991 | 38.1297 |
| MSE | 312.0477 | 1973.156 | 2133.018 | 49.20572 | 56.13020 |
| Population Total | | 78.50921 | | | |

In population II, i noted that SM has the least absolute bias followed by Nadaraya-Watson, periodic spline, Horvitz-Thompson, and lastly the Ratio estimator. Next, looking at MSE, the Nadaraya-Watson and periodic spline both have low MSE followed by SM, H-T and lastly Ratio estimator. Here we note that the Nadaraya-Watson is the best estimator for a quadratic and homoscedastic population while Ratio estimator has the highest MSE thus making it the least efficient estimator for this population. This is true because the ratio estimator is based on the assumption of linearity which when violated the estimator as expected breaks down.

***Table 4.** Population III (Linear and heteroscedastic).*

| | SM | HT | Ratio | NW | PS |
|---|---|---|---|---|---|
| Estimate | 74.74086 | 82.06381 | 83.4728 | 80.31023 | 82.88512 |
| Bias | -2.82623 | 3.49672 | 4.905757 | 1.74314 | 4.318024 |
| Var | 20.81065 | 1467.187 | 1189.2757 | 25.4385 | 28.73993 |
| MSE | 28.79823 | 1479.414 | 1213.3417 | 28.47704 | 47.38526 |
| Population Total | | 78.56709 | | | |

In population III, noted that Nadaraya-Watson has the least absolute bias, followed by SM, Horvitz-Thompson, Periodic spline and lastly Ratio estimator. Considering the MSE, Nadaraya-Watson and SM have a low MSE followed by periodic spline, Ratio and the H-T estimator in that order. Nadaraya –Watson and SM become the best estimators of this population which is linear and heteroscedastic population.

***Table 5.** Population IV (Kenya population census-1999).*

| | SM | HT | Ratio | NW | PS |
|---|---|---|---|---|---|
| Estimate | 29.69973 | 26.436061 | 30.33481 | 29.8734 | 31.40218 |
| Bias | 1.31815 | -1.945519 | 1.953226 | 1.491818 | 3.020605 |
| Var | 29.31531 | 640.95164 | 560.92100 | 109.86438 | 186.017 |
| MSE | 31.05283 | 644.76673 | 564.7361 | 112.0899 | 195.1441 |
| Population Total 28.38158 | | | | | |

In population IV, noted that Sample Mean has the least bias followed by Nadaraya-Watson,

H-T, Ratio and lastly periodic spline. Looking at MSE, SM has the least MSE, followed by Nadaraya-Watson, Periodic spline, Ratio and H-T estimator. Thus, SM has proved to be the best estimator for this real population which appears to be linear and with heteroscedastic variance from the scatter diagram.

***Table 6.*** *Population V (stock exchange shares).*

| | OLS | HT | Ratio | NW | PS |
|---|---|---|---|---|---|
| Estimate | 22.90294 | 14.37602 | 17.61762 | 21.3429 | 20.95047 |
| Bias | 2.98857 | -5.53835 | -2.29675 | 1.42853 | 1.0361 |
| Var | 666.5136 | 1419.2356 | 1681.9498 | 197.4472 | 102.8545 |
| MSE | 675.4452 | 1449.9089 | 1687.2246 | 199.48789 | 103.9280 |
| Population | Total | 19.91437 | | | |

This population appears to be neither linear nor homoscedastic from the scatter diagram Figure

5. In this population, Periodic spline has the least absolute bias, next is Nadaraya-Watson, Ratio, SM, and lastly Horvitz-Thompson estimator. As concerns the MSE, Periodic spline has the least MSE thus proving to be the best estimator for this population whose structure is not known. It is followed by Nadaraya-Watson, SM, H-T and Ratio estimator.

***Table 7.*** *Mean Square Error.*

| | SM | HT | Ratio | NW | PS |
|---|---|---|---|---|---|
| POP I | 20.43321 | 63.61897 | 58.5112 | 21.50266 | 25.23209 |
| POP II | 312.0477 | 1973.156 | 2133.018 | 49.20572 | 56.13020 |
| POP III | 28.79823 | 1479.414 | 1213.3417 | 28.47704 | 47.38526 |
| POP IV | 31.052832 | 644.76673 | 564.7361 | 112.0899 | 195.1441 |
| POP V | 675.4452 | 1449.9089 | 1687.2246 | 199.48789 | 103.9280 |

***Table 8.*** *Relative Change in Efficiency (RCE).*

| | SM | HT | Ratio | NW | PS |
|---|---|---|---|---|---|
| RCE I | 14.271595 | 30.015214 | 35.454867 | 1.288355 | 1.224556 |
| RCE II | 0.40938355 | 22.254290 | 19.7369136 | 0.3243495 | 0.877976 |
| RCE III | 0.51972363 | 9.134818 | 8.6517607 | 4.2128388 | 5.2143488 |
| RCE IV | 32.056245 | 21.7905 | 27.835925 | 8.2773587 | 3.1188819 |

Finally, the study compared the relative Change in Efficiency (RCE) among the five estimators. First, was the case when linearity assumption of the population structure is violated. Considering the RCE I, in Table 7 that the nonparametric estimators, Nadaraya-Watson and Periodic Spline have low RCE. This imply that they are the least sensitive to the violation of the linearity structure of the population and hence the most Robust among the five estimators. They are then followed by the SM, and Ratio estimators. Nevertheless, Horvitz-Thompson estimator is the least Robust among them as far as the violation of linearity assumption of the population structure is concerned. Secondly RCE II, investigate the violation of the Homoscedastic assumption in a population structure. Considering the RCE II, Table 7 that the Nadaraya-Watson, Periodic Spline and SM have the lowest RCE. This imply that they are the least sensitive to the change of structure of the population and hence the most robust among the five when homoscedastic assumption is violated.

On the other hand the Ratio and Horvitz-Thompson are least robust to the violation of homoscedastic condition on the population structure. Next we consider RCE 111. SM is having the least value. Next on the list is Nadaraya-Watson, Periodic spline, Ratio and Horvitz-Thompson estimators. However, we have also noted that all values of RCE 111 are quite low. This implies that though SM is the most robust to the change in the population structure, the low value shows that the other estimators are also robust and we conclude that population I is almost similar in structure to population IV, though it seems that homoscedastic condition is violated.

Lastly, in RCE IV, The Periodic Spline estimator has the least value of RCE thus becoming the most robust estimator to the change of population structure from linear and homoscedastic to the structure which is non linear and non homoscedastic. Nadaraya-Watson also proved to be robust to the same change in the structure. However, Ratio and Horvitz-Thompson estimators proved to be highly sensitive to the changes in the population structure. These two estimators are therefore less robust as compared to the other two non parametric estimators. The least robust estimator on this list as fur as this population is concern is SM. Therefore, Periodic spline has proved to be robust when both linearity and homoscedastic conditions are violated.

# 4. Conclusions and Recommendations

### *4.1. Conclusions*

This study has revealed that the spline regression estimator performed impressively well in all aspects considered: bias, efficiency and robustness. We noted that it performed well in linear homoscedastic model and in quadratic homoscedastic model. However, even when the homoscedasticity assumption was violated it still performed well. We therefore conclude that Periodic Spline estimator is a robust estimator. It is therefore recommended to be used as a suitable estimator of the population total when the structure of the population is unknown. It has also been noted that the Nadaraya-Watson estimator performs well in the linear homoscedastic model and also when the linearity conditions is violated. It also suffices to mention that its performance was unquestionably impressive in the linear heteroscedastic model clearly indicating that it is robust to the violation of linearity and homoscedastic condition.

### *4.2. Recommendation*

i. From the findings of our research, the Horvitz-Thompson (design-based) estimator and the Ratio estimator (model-based) should be used within the confines of a linear homoscedastic model. They are not appropriate for use when the structure of the population is not known.

ii. The two estimators, Nadaraya-Watson and periodic spline estimators; are suitable for use in linear homoscedastic model and even when the assumptions of the model are violated sensitivity to the change of population structure is relatively low and hence are classified as highly robust.

*Notation*

1. N = size of the finite population generally assumed to be known.
2. n = sample size.
3. x = design variable. Its values can either be made

available before hand or in the course of data collection.

4. y = the survey variable or variable under study.
5. $\bar{y}$= n-1 $\sum_{i=1}^{n} y_i$ = sample mean.
6. s2 = sample variance = $\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$
7. σ 2(sigma) = population variance = $\frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2$
8. $\bar{Y}$ = $\frac{1}{N}\sum_{I=1}^{N} Y_i$ the finite population mean.
9. $Y_T$ = $\sum_{i=1}^{N} Y_i$ = Finite population total.
10. Srswor – Abbreviation of simple random sampling without replacement.
11. Ksh – Kenya shilling.
12. SM=Sample Mean
13. HT = Horvitz-Thompson
14. R = Ratio
15. NW = Nadaraya-Watson
16. PS = Periodic-Spline

## Acknowledgements

First and foremost, I thank Almighty God for His abundant grace, faithfulness, care, knowledge and the strength throughout this period of study. Secondly, I sincerely thank my supervisor, Dr. Njenga, for his profession guidance, patience, his availability for consultation and provision of the reference materials that I needed in the course of this research project. My gratitude also goes to the other team of my lecturers in the Department of Mathematics (statistics), Kenyatta University for their support and very educative lectures. Special thanks also go to J.P. Wanjala of KNBS who closely journey with us under institution based program for his encouragement. Lastly, I am grateful to Mr. Felix Wasike of Kenyatta University for his invaluable technical assistance.

## References

[1] Cassel, C.M., Sarndal, C.E. and Wretman, J.H. (1977). Foundation of inference in survey sampling. New York: Wiley.

[2] Chambers, R.L (2003) which sample survey strategy? A review of three different approaches. Southampton Statistical Sciences Research Institute. University of Southampton

[3] Cochran, W.G. (1940). The estimation of the yields of cereal experiments by sampling for the ratio of grain to total produce. Journal of Agricultural Science, 30. 262-275.

[4] Cochran, W.G. (1977). Sampling Techniques, 3rd edition. New York: Wiley.

[5] Cox, B. G (1995) Business survey methods. New York: John Wiley

[6] Gasser, T. and Muller, H.G. (1979). Kernel estimation of regression functions. In smoothing techniques for curves estimation. Heidelberg: Springer –verlag 23-68.

[7] Hubback, J.A (1927). Sampling for rice yield in Bihar and Orissa. Imperial Agricultural Research Institute, Pusa, Bulletin No. 166 (represented in Sankhya, 1946, 7, 281-294.

[8] Kiaer, A.N (1897). The representative method of statistical survey. (1976 English translation of the original Norwegian), Oslo. Central Bureau of Statistic Norway.

[9] Kim, J.-Y. (2004). Nonparametric Regression Estimation in Survey Sampling. Ph.D. thesis, Iowa State University

[10] Laplace, P. S. (1820). A philosophical essay on probabilities. English translation, Dover, 1951

[11] Mukhopadhyay,P.(2005). Theory and methods of Survey Sampling, Prentice-Hall of India Private Limited, New Delhi

[12] Nadaraya, E.A. (1964). On Estimating Regression. Theory of Probability Application 9, 141-142

[13] Newman, J (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. Journal of the Royal Statistic Society, 97,558-625.

[14] Odhiambo, R. O. And Etwasi, W. Non-response weighting adjustment approach sample survey unpublished MSC Dissertation (2006) Jomo Kenyatta University of Agriculture and Technology.

[15] Sukhatme, P.V. and Sukhatme, B.V. (1970) Sampling Theory of Surveys with Applications.

[16] Priestly, M.B. and Chao, M.T. (1972). Non parametric function fitting. Statistic Society. B 34,385,392.

[17] Royall, R.M. (1971).Linear regression models in finite population sampling theory. In V.P Godambe and D.a. Spott. Foundations of Statistical Inference. Holt, Rinehart and Winston, Toronto 259-279.

[18] Royall, R.M. And Herson, J. (1973a). Robust estimations in finite populations I. Journal of the American Statistical Association, 68,880-889.

[19] Royall, R.M. And Herson, J. (1973b). Robust estimations in finite populations II. Journal of the American Statistical Association, 68,890-893.

[20] Schoenberg, I. J. (1946). Spline functions and the problem of graduation. Proc. Nat. Acad. Sci. USA 52, 947-950.

[21] Silverman, B.W. (1984). Spline smoothing: the equivalent variable Kernel method. The Annals of Statistics, 12, 3, 898-916.

[22] Silverman, B.W. (1985). Some aspects of the spline smoothing approach to non- parametric regression curve fitting. Journal of the Royal Statistical Society 13, 47, 1-52.

[23] Silverman, B. W. (1986).Density Estimation for statistics and Data Analysis Chapman and Hall.

[24] Smith, P. (1979). Splines as a useful and convenient statistical tool. American Statistical Journal 33, 57-62.

[25] Smith, T.M.F., and Njenga, E. (1992). Robust model-based methods for analytical surveys. Survey Methodology 18, 2, 187-208.

[26] Wahba, G. (1975). Optimal convergence properties of a variable knots, Kernel and Orthogonal series methods for density estimation. Ann. Statistic. 3, 15-29.

[27] Whittaker, E. (1923). On a new method of graduation. Proc. Edinburgh math. Soc 41, 63-75.

[28] Zarkovic, S. (1956). Note on the history of sampling methods in Russia. Journal of the Royal Statistical Society Series A, 119, 336-338.

[29] Zheng, H. and Little, R.J. (2004). Penalized Spline Nonparametric Mixed Models for Inference about a Finite Population Mean from Two-Stage Samples. Survey Methodology 30, 209–218.