International Journal of Statistics and Probability; Vol. 7, No. 4; July 2018
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

On Comparison of Local Polynomial Regression Estimators for $P = 0$ and $P = 1$ in a Model Based Framework

Conlet Biketi Kikechi[1] & Richard Onyino Simwa[2]

[1] Statistics and Operations Research Section, School of Mathematics, College of Biological and Physical Sciences, University of Nairobi, Nairobi, Kenya

[2] Actuarial Science and Financial Mathematics Section, School of Mathematics, College of Biological and Physical Sciences, University of Nairobi, Nairobi, Kenya

Correspondence: Conlet Biketi Kikechi, Statistics and Operations Research Section, School of Mathematics, College of Biological and Physical Sciences, University of Nairobi, Nairobi, Kenya. Email: Kikechiconlet@gmail.com

Received: May 16, 2018 Accepted: May 31, 2018 Online Published: June 28, 2018

doi:10.5539/ijsp.v7n4p104 URL: https://doi.org/10.5539/ijsp.v7n4p104

Abstract

This article discusses the local polynomial regression estimator for $P = 0$ and the local polynomial regression estimator for $P = 1$ in a finite population. The performance criterion exploited in this study focuses on the efficiency of the finite population total estimators. Further, the discussion explores analytical comparisons between the two estimators with respect to asymptotic relative efficiency. In particular, asymptotic properties of the local polynomial regression estimator of finite population total for $P = 0$ are derived in a model based framework. The results of the local polynomial regression estimator for $P = 0$ are compared with those of the local polynomial regression estimator for $P = 1$ studied by Kikechi et al (2018). Variance comparisons are made using the local polynomial regression estimator $\bar{T}_0$ for $\mathrm{P} = 0$ and the local polynomial regression estimator $\bar{T}_1$ for $\mathrm{P} = 1$ which indicate that the estimators are asymptotically equivalently efficient. Simulation experiments carried out show that the local polynomial regression estimator $\bar{T}_1$ outperforms the local polynomial regression estimator $\bar{T}_0$ in the linear, quadratic and bump populations.

Keywords: Asymptotic Properties, Asymptotic Relative Efficiency, Finite Population, Local Polynomial Regression, Model Based Framework, Nonparametric Regression, Sample Surveys

1. Introduction

The theory of sample surveys involves principles and methods of collecting and analyzing data from a finite population of N units and then making inferences about finite population parameters on the basis of information obtained from the sample. For some early work on survey sampling theory, see Royall (1970a), Royall (1970b), Royall (1971), Smith (1976) and Pfeffermann (1993). In this study, an estimator of the finite population total is developed and its properties derived using the local polynomial regression procedure. Local polynomial regression is a nonparametric technique which is a generalization of kernel regression and is used for smoothing scatter plots and modeling functions. Under normal conditions, when $p = 0$, this is referred to as local constant regression, when $p = 1$, this is local linear regression and when $p \geq 2$, this is local polynomial regression. p is the order of the local polynomial being fit. In local polynomial regression, a low order weighted least squares regression is fit at each point of interest x, using data from some neighborhood around x (see Cleveland (1979) and Cleveland and Devlin (1988)).

Once a modeling approach is undertaken, there is a special feature in finite population estimation problems that the unknown quantities are realized values of random variables, so the basic problem has the feature of being similar to a prediction problem. In order to estimate $m(x)$ at a given point x, the association between the predictor variable and the response variable is explored. This methodology was introduced by Stone (1977). It has also been studied by Fan (1993), Fan and Gijbels (1996), Breidt and Opsomer (2000) and Kikechi et al (2017). Like in Stone (1977), the main aim of this procedure is to quantify the contribution of the covariate X to the response Y per unit value of X in order to summarize the association between the two variables, to predict the mean response for a given value X and to extrapolate the results beyond the range of the observed covariate values. A weight $k\left(\frac{x_i - x}{h}\right)$ is assigned to the point

(x_i, y_i) where h is the size of the local neighbourhood and $k(t)$ is the unimodal non-negative function. On the other hand, inferences may explore properties of the process that generate the population values (Montanari and Ranalli (2003)). An assumption is made from the fact that the finite population has been generated by a super population model $\xi = f(x, y, \varphi)$ and it is of interest to estimate the population parameters φ, where $\varphi = \alpha + \beta x_i$. The super population model can be applied to predict the unobserved values $y_i's$ after obtaining estimates of α and β using the known auxiliary information $x_i,\ i = 1,2 \ldots, N$ (see Montanari and Ranalli (2005) and Rueda and Sanchez-Borrego (2009)).

The nonparametric approach does not restrict the functional form of the distribution nor does it specify the various stochastic properties such as $E_\xi(.)$, $V_\xi(.)$ and $MSE_\xi(.)$. Rather, it leaves them to cover broad classes of models, thus allowing for more robust inference than inference obtained in parametric approach. Using the model ξ, the nonparametric estimator of total, T has been derived by Nadaraya (1964), Watson (1964), Priestly and Chao (1972), Gasser and Muller (1979), Dorfman (1992)), Chambers et al (1993) and Odhiambo and Mwalili (2000). In his study, Dorfman (1992) has been able to prove the asymptotic unbiasedness and MSE consistency of this estimator. The estimator, however suffers from sparse sample problem, and more work needs to be done to come up with another technique that can overcome this problem. This is where the local polynomial procedure comes in. See Kikechi et al (2017) and Kikechi et al (2018).

The local polynomial regression is one of the most successfully applied design adaptive non parametric regression. This estimation procedure is an attractive choice due to its flexibility and asymptotic performance. Having a local model (rather than just a point estimate) enables derivation of response adaptive methods for bandwidth and polynomial order selection in a straightforward manner. The procedure has also the advantage of eliminating design bias and alleviating boundary bias. Furthermore, the method adapts well to random, fixed, highly clustered and nearly uniform designs. The weighted least squares principle to be employed in the local polynomial approximation approach, opens the way to a wealth of statistical knowledge and thus providing easy computations and generalizations. See Fan (1992), Fan (1993), Ruppert and Wand (1994) and Fan and Gijbels (1996) among others.

Kikechi et al (2018) employ a superpopulation approach to estimate the finite population total using the procedure of local linear regression. Explicitly, the authors derive robustness properties of the local linear regression estimator and carry out simulation experiments on the performances of this estimator in comparison with other estimators that exist in the literature. Results indicate that the local linear regression estimator is more efficient and performing better than the Horvitz-Thompson (1952) and Dorfman (1992) estimators, regardless of whether the model is specified or mispecified. In this paper, the local polynomial regression estimator of finite population total for $P = 0$ is studied and asymptotic properties derived. Analytical comparisons are carried out between this estimator and the local polynomial regression estimator for $P = 1$ studied by Kikechi et al (2018) which indicate that the estimators are asymptotically equivalently efficient. Simulation experiments however indicate that the local polynomial regression estimator $\bar{T}_1$ is superior and dominates the local polynomial regression estimator $\bar{T}_0$ in the linear, quadratic and bump populations.

2. Method of Constructing the Local Polynomial Regression Estimator $\bar{T}$ for $P = 0$

The superpopulation model considered for estimating the finite population total is given by,

$$Y_i = m(X_i) + \sigma^2(X_i)\varepsilon_i \tag{1}$$

Specifically, the following assumptions hold for the model considered in the nonparametric regression estimation of $m(x_i)$:

$$E(Y_i/X_i = x_i) = m(x_i)$$

$$Cov\left(Y_i, Y_j/X_i = x_i, X_j = x_j\right) = \begin{cases} \sigma^2(x_i), & i = j \\ 0 \quad , & i \neq j \end{cases} \quad i = 1,2,3,\ldots.,N \quad j = 1,2,3,\ldots,N\,. \tag{2}$$

The properties of the error are given by,

$$E(\varepsilon_i/X_i = x_i) = m(x_i)$$

$$Cov\left(\varepsilon_i, \varepsilon_j/X_i = x_i, X_j = x_j\right) = \begin{cases} \sigma^2(x_i), & i = j \\ 0 \quad , & i \neq j \end{cases} \quad i = 1,2,3,\ldots.,N \quad j = 1,2,3,\ldots,N\,. \tag{3}$$

The functions $m(x_i)$ and $\sigma^2(x_i)$ are assumed to be smooth and strictly positive. Consider the Taylor series

expansion of $m(x_i)$ expressed as,

$$m(x_i) = m(x_j + ht) = m(x_j) + htm'(x_j) + \frac{h^2t^2}{2!}m''(x_j) + \frac{h^3t^3}{3!}m'''(x_j) + \cdots$$
$$= m(x_j) + (x_i - x_j)m'(x_j) + \frac{(x_i - x_j)^2}{2!}m''(x_j) + \frac{(x_i - x_j)^3}{3!}m'''(x_j) + \cdots \quad (4)$$

The Taylor series expansion is written in a general form expressed as,

$$y_i = \alpha + (x_i - x_j)\beta + \varepsilon_i \quad (5)$$

where x_i lies in the interval $[x_j - h,\ x_j + h]$ and

$$\varepsilon_i = \frac{(x_i - x_j)^2}{2!}m''(x_j) + \frac{(x_i - x_j)^3}{3!}m'''(x_j) + \cdots$$

The constants α and β are solved using the least squares procedure by making ε_i the subject of the formulae, squaring both sides, summing over all possible sample values and applying the weights to obtain a solution to the weighted least squares problem of the form;

$$\sum_{i\in S}\varepsilon_i^2 = \sum_{i\in S}\left(y_i - \alpha - \beta(x_i - x_j)\right)^2 K\left(\frac{x_i - x_j}{h}\right) \quad (6)$$

Letting,

$$\varphi = \sum_{i\in S}\left(y_i - \alpha - \beta(x_i - x_j)\right)^2 K\left(\frac{x_i - x_j}{h}\right) \quad (7)$$

Differentiating φ with respect to α and equating to zero, gives

$$\frac{\partial\varphi}{\partial\alpha} = \sum_{i\in S} -2\left(y_i - \alpha - \beta(x_i - x_j)\right)K\left(\frac{x_i - x_j}{h}\right)\left\{\left\{\sum_{i\in S}K\left(\frac{x_i - x_j}{h}\right)\right\}^{-1}\right\} = 0 \quad (8)$$

Implying that

$$\sum_{i\in S}K\left(\frac{x_i - x_j}{h}\right)y_i = \alpha\sum_{i\in S}K\left(\frac{x_i - x_j}{h}\right) + \beta\sum_{i\in S}(x_i - x_j)K\left(\frac{x_i - x_j}{h}\right). \quad (9)$$

Letting

$$S_{n,l} = \sum_{i\in S}K\left(\frac{x_i - x_j}{h}\right)(x_i - x_j)^l \quad (10)$$

Then it follows from equation (9) that

$$\sum_{i\in S}K\left(\frac{x_i - x_j}{h}\right)y_i = \alpha(S_{n,0}) + \beta(S_{n,1}). \quad (11)$$

Similarly, differentiating φ with respect to β and equating to zero, gives

$$\frac{\partial\varphi}{\partial\beta} = \sum_{i\in S} -2\left(y_i - \alpha - \beta(x_i - x_j)\right)(x_i - x_j)K\left(\frac{x_i - x_j}{h}\right)\left\{\left\{\sum_{i\in S}K\left(\frac{x_i - x_j}{h}\right)\right\}^{-1}\right\} = 0 \quad (12)$$

Implying that

$$\sum_{i\in S}(x_i - x_j)K\left(\frac{x_i - x_j}{h}\right)y_i = \alpha\sum_{i\in S}(x_i - x_j)K\left(\frac{x_i - x_j}{h}\right) + \beta\sum_{i\in S}(x_i - x_j)^2K\left(\frac{x_i - x_j}{h}\right). \quad (13)$$

and thus

$$\sum_{i\in S}(x_i - x_j)K\left(\frac{x_i - x_j}{h}\right)y_i = \alpha(S_{n,1}) + \beta(S_{n,2}). \quad (14)$$

Multiplying equation (11) and equation (14) by $(S_{n,2})$ and $(S_{n,1})$ respectively, gives

$$(S_{n,2})\sum_{i\in S} K\left(\frac{x_i - x_j}{h}\right) y_i = \alpha(S_{n,0})(S_{n,2}) + \beta(S_{n,1})(S_{n,2}) \tag{15}$$

$$(S_{n,1})\sum_{i\in S} (x_i - x_j) K\left(\frac{x_i - x_j}{h}\right) y_i = \alpha(S_{n,1})^2 + \beta(S_{n,1})(S_{n,2}) \tag{16}$$

Subtracting equation (16) from equation (15), gives

$$(S_{n,2})\sum_{i\in S} K\left(\frac{x_i - x_j}{h}\right) y_i - (S_{n,1})\sum_{i\in S} (x_i - x_j) K\left(\frac{x_i - x_j}{h}\right) y_i = \alpha(S_{n,0})(S_{n,2}) - \alpha(S_{n,1})^2 \tag{17}$$

Making α the subject of the formulae, gives

$$\bar{\alpha} = \sum_{i\in S} \left\{ \frac{\left(S_{n,2} - S_{n,1}(x_i - x_j)\right)}{(S_{n,0})(S_{n,2}) - (S_{n,1})^2} K\left(\frac{x_i - x_j}{h}\right) y_i \right\} \tag{18}$$

Similarly, multiplying equation (11) and equation (14) by $(S_{n,1})$ and $(S_{n,0})$ respectively, gives

$$(S_{n,1})\sum_{i\in S} K\left(\frac{x_i - x_j}{h}\right) y_i = \alpha(S_{n,0})(S_{n,1}) + \beta(S_{n,1})^2 \tag{19}$$

$$(S_{n,0})\sum_{i\in S} (x_i - x_j) K\left(\frac{x_i - x_j}{h}\right) y_i = \alpha(S_{n,0})(S_{n,1}) + \beta(S_{n,0})(S_{n,2}) \tag{20}$$

Subtracting equation (20) from equation (19), gives

$$(S_{n,1})\sum_{i\in S} K\left(\frac{x_i - x_j}{h}\right) y_i - (S_{n,0})\sum_{i\in S} (x_i - x_j) K\left(\frac{x_i - x_j}{h}\right) y_i = \beta(S_{n,1})^2 - \beta(S_{n,0})(S_{n,2}) \tag{21}$$

Making β the subject of the formulae, gives

$$\bar{\beta} = \sum_{i\in S} \left\{ \frac{\left(S_{n,0}(x_i - x_j) - S_{n,1}\right)}{(S_{n,0})(S_{n,2}) - (S_{n,1})^2} K\left(\frac{x_i - x_j}{h}\right) y_i \right\} \tag{22}$$

Now it follows from equation (5) that

$$\bar{y}_i = \bar{\alpha} + (x_i - x_j)\bar{\beta} \tag{23}$$

If the value assigned is zero, assuming that $\bar{\beta}$ is a pre-assigned constant, then

$$\bar{y}_j = \bar{\alpha} \tag{24}$$

Therefore

$$\bar{m}(x_j) = \sum_{i\in S} \left\{ \frac{\left(S_{n,2} - S_{n,1}(x_i - x_j)\right)}{(S_{n,0})(S_{n,2}) - (S_{n,1})^2} K\left(\frac{x_i - x_j}{h}\right) y_i \right\}$$

$$= \sum_{i\in S} w_i(x_j) y_i \tag{25}$$

where

$$w_i(x_j) = \frac{\left(S_{n,2} - S_{n,1}(x_i - x_j)\right)}{(S_{n,0})(S_{n,2}) - (S_{n,1})^2} K\left(\frac{x_i - x_j}{h}\right) y_i$$

Implying that the finite population total estimator $\bar{T}$ for $P = 0$ can be estimated using

$$\begin{aligned} \bar{T} &= \sum_{i\in S} y_i + \sum_{j\in R} \bar{m}(x_j) \\ &= \sum_{i\in S} y_i + \sum_{j\in R} \left\{ \sum_{i\in S} \left\{ \frac{\left(S_{n,2} - S_{n,1}(x_i - x_j)\right)}{(S_{n,0})(S_{n,2}) - (S_{n,1})^2} K\left(\frac{x_i - x_j}{h}\right) y_i \right\} \right\} \end{aligned} \tag{26}$$

3. Properties of the Local Polynomial Regression Estimator $\bar{T}$ for $P = 0$

In deriving the properties of the local polynomial regression estimator, the following assumptions are made according to Ruppert and Wand (1994):

(i) The x_j variables lie in the interval $(0,1)$.

(ii) The function $m''(.)$ is bounded and continuous on $(0,1)$.

(iii) The kernel $K(t)$ is symmetric and supported on $(-1,1)$. Also $K(t)$ is bounded and continuous satisfying the following: $\int_{-\infty}^{\infty} K(x)\,dx = 1$, $\int_{-\infty}^{\infty} xK(x)\,dx = 0$, $\int_{-\infty}^{\infty} x^2K(x)\,dx > 0$, $\int_{-\infty}^{\infty} K^2x\,dx < \infty$, $d_k = \int_{-\infty}^{\infty} K^2(t)\,dt$

(iv) The bandwidth h is a sequence of values which depend on the sample size n and satisfying $h \to 0$ and $nh \to \infty$, as $n \to \infty$.

(v) The point x_j at which the estimation is taking place satisfies $h < x_j < 1 - h$.

Fan (1993) imposed conditions on $K(.)$ and are only used for convenience in terms of technical arguments and thus can be relaxed.

3.1 The Expectation of the Local Polynomial Regression Estimator $\bar{T}$ for $P = 0$

The expectation of $\bar{T}$ *for* $P = 0$ is derived as,

$$\begin{aligned} E(\bar{T}) &= \sum_{i\in S} E(y_i) + \sum_{j\in R}\left\{\sum_{i\in S}\left\{\frac{\left(S_{n,2} - S_{n,1}(x_i - x_j)\right)}{(S_{n,0})(S_{n,2}) - (S_{n,1})^2} k\left(\frac{x_i - x_j}{h}\right) E(y_i)\right\}\right\} \\ &= \sum_{i\in S} m(x_i) + \sum_{j\in R}\left\{\sum_{i\in S}\left\{\frac{\left(S_{n,2} - S_{n,1}(x_i - x_j)\right)}{S_{n,0}S_{n,2} - (S_{n,1})^2} k\left(\frac{x_i - x_j}{h}\right) m(x_i)\right\}\right\} \end{aligned} \tag{27}$$

Using the Taylor series expansion of the form,

$$m(x_i) = m(x_j) + htm'(x_j) + \frac{h^2t^2}{2!}m''(x_j) + \cdots, \tag{28}$$

Theorem 3 in Fan and Gijbels (1996) is such that under the conditions given in (i)-(v), allows

$$\begin{aligned} E(\bar{T}) &= \sum_{i\in S} m(x_i) + \sum_{j\in R}\left\{\sum_{i\in S}\left\{\frac{S_{n,2}k\left(\frac{x_i - x_j}{h}\right)}{S_{n,0}S_{n,2} - (S_{n,1})^2}\left(m(x_j) + htm'(x_j) + \frac{h^2t^2}{2!}m''(x_j) + \cdots\right)\right\}\right\} \\ &\quad - \sum_{j\in R}\left\{\sum_{i\in S}\left\{\frac{S_{n,1}(x_i - x_j)}{S_{n,0}S_{n,2} - (S_{n,1})^2} k\left(\frac{x_i - x_j}{h}\right)\left(m(x_j) + htm'(x_j) + \frac{h^2t^2}{2!}m''(x_j) + \cdots\right)\right\}\right\} \\ &= \sum_{i\in S} m(x_i) + \sum_{j\in R}\left\{\left(\frac{S_{n,0}S_{n,2} - (S_{n,1})^2}{S_{n,0}S_{n,2} - (S_{n,1})^2}\right)m(x_j)\right\} + \sum_{j\in R}\left\{\left(\frac{S_{n,1}S_{n,2} - S_{n,1}S_{n,2}}{S_{n,0}S_{n,2} - (S_{n,1})^2}\right)m'(x_j)\right\} \\ &\quad + \sum_{j\in R}\left\{\left(\frac{(S_{n,2})^2 - S_{n,1}S_{n,3}}{S_{n,0}S_{n,2} - (S_{n,1})^2}\right)\frac{m''(x_j)}{2}\right\} \\ &= \sum_{i\in S} m(x_i) + \sum_{j\in R} m(x_j) + \sum_{j\in R}\left\{\left(\frac{(S_{n,2})^2 - S_{n,1}S_{n,3}}{S_{n,0}S_{n,2} - (S_{n,1})^2}\right)\frac{m''(x_j)}{2}\right\}. \end{aligned} \tag{29}$$

3.2 The Bias of the Local Polynomial Regression Estimator $\bar{T}$ for $P = 0$

The bias of $\bar{T}$ is given by

$$Bias(\bar{T}) = \sum_{j\in R}\left\{\left(\frac{(S_{n,2})^2 - S_{n,1}S_{n,3}}{S_{n,0}S_{n,2} - (S_{n,1})^2}\right)\frac{m''(x_j)}{2}\right\}. \tag{30}$$

Therefore the asymptotic expression of the bias of the local polynomial regression estimator $\bar{T}$ is

$$\begin{aligned} Bias_{asy}(\bar{T}) &= \sum_{j\in R}\left\{\frac{\left(n^2h^6{k_2}^2 + o(n^2h^8)\right)m''(x_j)}{2\left(n^2h^4k_2 + o(n^2h^6)\right)}\right\} \\ &= \sum_{J\in R}\left\{\frac{1}{2}h^2k_2m''(x_j)\right\} \end{aligned} \tag{31}$$

3.3 The Variance of the Local Polynomial Regression Estimator $\bar{T}$ for $P = 0$

The variance of the local polynomial regression estimator $\bar{T}$ is estimated using the variance of the error, thus $Var(\bar{T} - T)$ is derived as

$$\begin{aligned} Var(\bar{T}) &= Var\left\{\sum_{i\in S} y_i + \sum_{j\in R} \bar{m}(x_j) - \sum_{i\in S} y_i - \sum_{j\in R} y_j\right\} \\ &= Var\left\{\sum_{i\in S}\sum_{j\in R} w_i(x_j) y_i - \sum_{j\in R} y_j\right\} \\ &= \sum_{j\in R}\sum_{i\in S} w_i^2(x_j)\,\sigma^2(x_i) + \sum_{j\in R}\sigma^2(x_j) \end{aligned} \tag{32}$$

where,

$$w_i(x_j) = \frac{\left(S_{n,2} - S_{n,1}(x_i - x_j)\right)}{(S_{n,0})(S_{n,2}) - (S_{n,1})^2} K\left(\frac{x_i - x_j}{h}\right).$$

The asymptotic expression for the variance of $\bar{T}$ is given by the expression using the results of $\bar{m}(x_j)$ that have been derived, thus

$$\begin{aligned} Var_{asy}(\bar{T}) &= \frac{1}{nh}\sum_{j\in R}\sum_{i\in S}\left\{K^2\left(\frac{x_i - x_j}{h}\right)\sigma^2(x_i)\left(\frac{x_i - x_{i-1}}{h}\right)\right\} \\ &= \sum_{j\in R}\frac{d_k}{nh}\sigma^2(x_j). \end{aligned} \tag{33}$$

3.4 The MSE of the Local Polynomial Regression Estimator $\bar{T}$ for $P = 0$

Theorem I in Fan (1993) allows that under condition (ii) gives,

$$\begin{aligned} MSE(\bar{T}) &= \{Bias(\bar{T})\}^2 + Var(\bar{T}) \\ &= \left\{\sum_{j\in R}\left\{\left(\frac{(S_{n,2})^2 - S_{n,1}S_{n,3}}{S_{n,0}S_{n,2} - (S_{n,1})^2}\right)\frac{m''(x_j)}{2}\right\}\right\}^2 + \sum_{j\in R}\sum_{i\in S} w_i^2(x_j)\,\sigma^2(x_i) + \sum_{j\in R}\sigma^2(x_j) \end{aligned} \tag{34}$$

The asymptotic expression for the MSE of the local polynomial regression estimator $\bar{T}$ is given by

$$MSE_{asy}(\bar{T}) = \left\{\sum_{j\in R}\left\{\frac{1}{2}\,h^2 k_2 m''(x_j)\right\}\right\}^2 \tag{35}$$

Note that results for the local polynomial regression estimator of finite population total $\bar{T}$ for $P = 1$ have been derived by Kikechi et al (2018).

3.5 The Asymptotic Relative Efficiency

The relative efficiency of two procedures is the ratio of their efficiencies, but it is often possible to use the asymptotic relative efficiency, defined as the limit of the relative efficiencies as the sample size grows, as the principal measure of comparison. Let $\bar{T}_0$ be the local polynomial regression estimator of finite population total for $P = 0$ and $\bar{T}_1$ be the local polynomial regression estimator of finite population total for $P = 1$ as studied by Kikechi et al (2018).

If $\bar{T}_0$ and $\bar{T}_1$ are both unbiased estimators of T, then the relative efficiency of $\bar{T}_0$ to $\bar{T}_1$ is given by,

$$Eff(\bar{T}_0, \bar{T}_1) = \frac{Var(\bar{T}_1)}{Var(\bar{T}_0)}. \tag{36}$$

If $\bar{T}_0$ and $\bar{T}_1$ are both asymptotically unbiased estimators of T, then the asymptotic relative efficiency of $\bar{T}_0$ to $\bar{T}_1$ is given by,

$$ARE(\bar{T}_0, \bar{T}_1) = \lim_{n\to\infty} Eff(\bar{T}_0, \bar{T}_1) = \lim_{n\to\infty}\frac{Var(\bar{T}_1)}{Var(\bar{T}_0)}. \tag{37}$$

Therefore, the estimators of finite population totals for $\bar{T}_0$ and $\bar{T}_1$ are respectively given by,

$$\bar{T}_0 = \sum_{i\in S} y_i + \sum_{j\in R}\left\{\sum_{i\in S}\left\{\frac{\left(S_{n,2} - S_{n,1}(x_i - x_j)\right)}{(S_{n,0})(S_{n,2}) - (S_{n,1})^2} K\left(\frac{x_i - x_j}{h}\right) y_i\right\}\right\}. \tag{38}$$

$$\bar{T}_1 = \sum_{i\in S} \mathrm{Y}_i + \sum_{j\in R}\left\{\sum_{i\in S}\left\{\frac{\left(S_{n,2} - S_{n,1}(x_i - x_j)\right)}{(S_{n,0})(S_{n,2}) - (S_{n,1})^2} k\left(\frac{x_i - x_j}{h}\right) y_i\right\}\right\}$$
$$+ \sum_{\boldsymbol{j\in R}}\left\{\left(\frac{x_i - x_j}{S_{n,0}S_{n,2} - (S_{n,1})^2}\right)\sum_{\boldsymbol{i\in S}}\left\{(S_{n,0}(x_i - x_j) - S_{n,1})k\left(\frac{x_i - x_j}{h}\right) y_i\right\}\right\}. \tag{39}$$

The variance of the local polynomial regression estimator $\bar{T}_0$ is given by,

$$Var(\bar{T}_0) \quad = \sum_{j\in R}\sum_{i\in S} w_i^2(x_j)\,\sigma^2(x_i) + \sum_{j\in R}\sigma^2(x_j) \tag{40}$$

The asymptotic expression for the variance of the local polynomial regression estimator $\bar{T}_0$ is estimated by,

$$Var_{asy}(\bar{T}_0) = \sum_{j\in R}\frac{d_k}{nh}\sigma^2(x_j) \tag{41}$$

The variance of the local polynomial regression estimator $\bar{T}_1$ is given by,

$$Var(\bar{T}_1\ ) = \sum_{j\in R}\sum_{i\in S} w_i^2(x_j)\,\sigma^2(x_i) + \sum_{j\in R}(x_i - x_j)^2\sum_{i\in S} w_i'^2(x_j)\sigma^2(x_i) + \sum_{j\in R}\sigma^2(x_j) \tag{42}$$

The asymptotic expression for the variance of the local polynomial regression estimator $\bar{T}_1$ is estimated by,

$$Var_{asy}(\bar{T}_1) = \sum_{j\in R}\frac{d_k}{nh}\sigma^2(x_j)\,. \tag{43}$$

Note that in Kikechi e tal (2017), $Var_{asy}\left(\bar{m}_{LL}(x_j)\right) = \frac{d_k}{nh}\sigma^2(x_j)$ and $Var_{asy}\left(\bar{m}_{NW}(x_j)\right) = \frac{d_k}{nh}\sigma^2(x_j)$

Thus the asymptotic relative efficiency of the local polynomial regression estimator $\bar{T}_0$ to the local polynomial regression estimator $\bar{T}_1$ derived by Kikechi et al (2018) is given by,

$$ARE(\bar{T}_0, \bar{T}_1) = \lim_{n\to\infty} Eff(\bar{T}_0, \bar{T}_1) = \lim_{n\to\infty}\left\{\frac{Var_{asy}(\bar{T}_1)}{Var_{asy}(\bar{T}_0)}\right\} = \lim_{n\to\infty}\left\{\frac{\sum_{j\in R}\frac{d_k}{nh}\sigma^2(x_j)}{\sum_{j\in R}\frac{d_k}{nh}\sigma^2(x_j)}\right\} = 1\,. \tag{44}$$

4. Simulation Study

4.1 Description of the Data Sets

In this section, simulation experiments are carried out to evaluate the performance of the estimators. The data are generated from the regression model of the form,

$$Y_i = m(X_i) + \sigma^2(X_i)\varepsilon_i \qquad i = 1,2,\dots,n \tag{45}$$

The data sets are obtained by simulation using specific models having relations of the form,

$$y_i = 1 + 2(x - 0.5) + \varepsilon_i \tag{46}$$

$$y_i = 1 + 2(x - 0.5)^2 + \varepsilon_i \tag{47}$$

$$y_i = 1 + 2(x - 0.5) + \exp(-200(x - 0.5)^2 + \varepsilon_i \tag{48}$$

for the linear, quadratic and bump populations respectively. The $x_i's$ are generated as independent and identically distributed (iid) uniform (0, 1) random variables. The errors are assumed to be independent and identically distributed (iid) random variables with mean 0 and constant variance. The analysis and comparison in terms of performance is based on the local polynomial regression estimator $\bar{T}_0$ and the local polynomial regression estimator $\bar{T}_1$. The Epanechnicov kernel given is used for kernel smoothing on each of the populations due to its simplicity and easy computations using well designed computer programs and is defined as,

$$\frac{3}{4\sqrt{5}}\left(1-\frac{1}{5}t^2\right)|t|<\sqrt{5} \tag{49}$$

The bandwidths are data driven and are determined by the least squares cross validation method. For each of the three artificial populations of size 200, samples are generated by simple random sampling without replacement using sample size $n=60$. For each combination of mean function, standard deviation and bandwidth, 500 replicate samples are selected and the estimators calculated.

Table 1. Computational Formulae for the Local Polynomial Regression Estimators $\bar{T}_0$ and $\bar{T}_1$

Estimator	Formulae
$LPRE,\bar{T}_0$	$\bar{T}_0=\sum_{i\in S}Y_i+\sum_{j\in R}\bar{m}_0(x_j)$
$LPRE,\bar{T}_1$	$\bar{T}_1=\sum_{i\in S}Y_i+\sum_{j\in R}\bar{m}_1(x_j)$

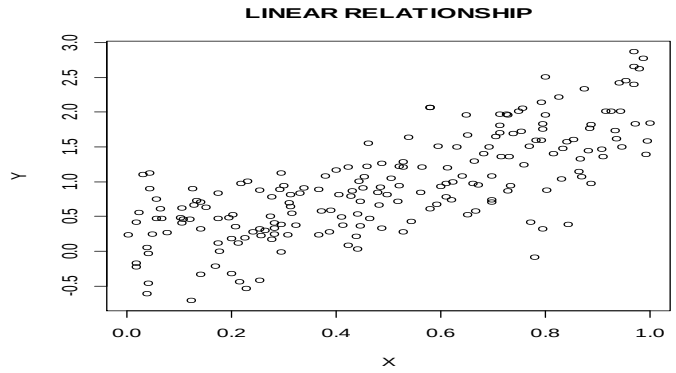

Figure 1. Scatter Diagram for the Linear Population

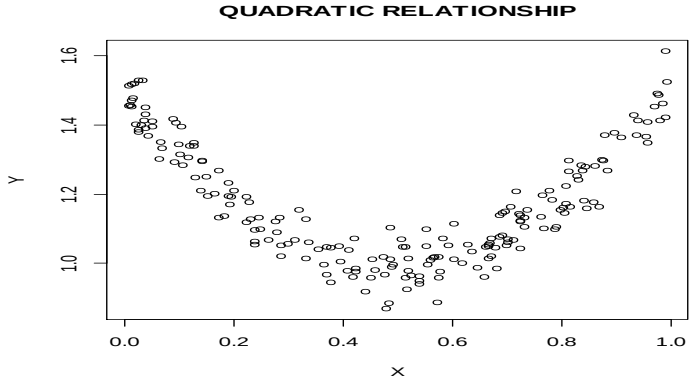

Figure 2. Scatter Diagram for the Quadratic Population

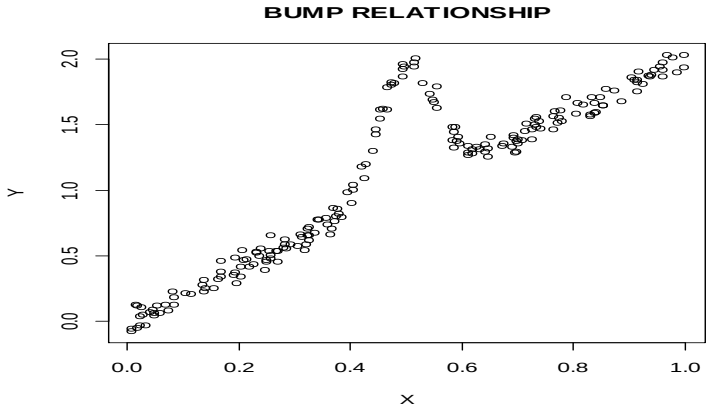

Figure 3. Scatter Diagram for the Bump Population

4.2 Results

The results of the bias and mean squared error (MSE) for the local polynomial regression estimator $\bar{T}_0$ for $P = 0$ and the local polynomial regression estimator $\bar{T}_1$ for $P = 1$ in the linear, quadratic and bump populations are provided in the table below.

Table 2. The Bias and MSE for $\bar{T}_0$ and $\bar{T}_1$ in the Three Artificial Populations

	Linear		**Quadratic**		**Bump**	
	$\bar{T}_0$	$\bar{T}_1$	$\bar{T}_0$	$\bar{T}_1$	$\bar{T}_0$	$\bar{T}_1$
BIAS	5.507608	3.777348	4.7372	0.45116	5.293896	0.4187236
MSE	100.8874	15.40735	18.40769	0.1601695	43.9272	0.1896261

5. Discussion

In estimating $\bar{m}(x_j)$ for the local polynomial regression estimator $\bar{T}_0$, $\bar{\beta}$ has been assumed to be a pre-assigned constant and in particular the value assigned is zero. It has therefore been shown in section 2 that the estimator $\bar{m}(x_j)$ is biased leading to a biased estimation of the finite population total. On the other hand, when estimating $\bar{m}(x_j)$ for the local polynomial regression estimator $\bar{T}_1$, the value of $\bar{\beta}$ is not pre-assigned but rather determined by the set of data provided and thus minimizing the bias. With regard to asymptotic relative efficiency, there is no difference in the performance of the local polynomial regression estimator $\bar{T}_0$ studied in this paper and the local polynomial regression estimator $\bar{T}_1$ studied by Kikechi et al (2018). The reason for this being that their ratio converges to 1 as n becomes large, see equation (44). This therefore implies that the estimators are asymptotically equivalently efficient. However, it is observed from simulation experiments conducted that the biases and MSEs computed in table 2 for the local polynomial regression estimator $\bar{T}_1$ are small in all the three populations. The results therefore indicate that the local polynomial regression estimator $\bar{T}_1$ is superior and dominates the local polynomial regression estimator $\bar{T}_0$ for the linear, quadratic and bump populations.

6. Conclusion

In this article the local polynomial regression estimators $\bar{T}_0$ and $\bar{T}_1$ of finite population totals have been studied in a model based framework. Analytically, variance comparisons are explored using the local polynomial regression estimator $\bar{T}_0$ for $\mathrm{P} = 0$ and the local polynomial regression estimator $\bar{T}_1$ for $\mathrm{P} = 1$ in which results indicate that the estimators are asymptotically equivalently efficient. Simulation experiments carried out in terms of the biases and MSEs show that the local polynomial regression estimator $\bar{T}_1$ outperforms the local polynomial regression estimator $\bar{T}_0$ in all the three artificial populations and therefore, $\bar{T}_1$ is the most efficient estimator.

References

Breidt, F. J., & Opsomer, J. D. (2000). Local Polynomial Regression Estimation in Survey Sampling. *Annals of statistics, 28*, 1026-1053.

Chambers, R. L., Dorfman, A. H., & Wehrly, T. E. (1993). Bias robust estimation in finite populations using nonparametric calibration. *J. Amer Statist Assoc., 88*, 268-277.

Cleveland, W. S. (1979). Robust Locally Weighted Regression and Smoothing Scatter Plots. *J. Amer. Statist. Assoc. 74*, 829–836.

Cleveland, W. S., & Devlin, S. (1988). Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *J. Amer. Statist. Assoc. 83*, 596–610.

Dorfman, A. (1992). Nonparametric Regression for Estimating Totals in Finite Populations, Proceedings of the Section on Survey Research Methods. *American Statistical Association*, 622-625.

Fan, J. (1992). Design Adaptive Nonparametric Regression. *Journal of American Statistical Association, 87*, 998-1004.

Fan, J. (1993). Local Linear Regression Smoothers and Their Minimax Efficiencies. *Annals of Statistics, 21*, 196–216. https://doi.org/10.1214/aos/1176349022

Fan, J., & Gijbels, I. (1996). Local Polynomial Modeling and its Applications. London: Chapman and Hall.

Gasser, T., & Muller, H. G. (1979). Kernel Estimation in Regression Functions. *Smoothing Techniques for Curve Estimation*, 23-68.

Horvitz, D. G., & Thompson, D. J. (1952). A Generalization of Sampling without Replacement from a Finite Universe. *Journal of American Statistical Association, 47*, 663-685. https://doi.org/10.1080/01621459.1952.10483446

Kikechi, C. B., Simwa, R. O., & Pokhariyal, G. P. (2017). On Local Linear Regression Estimation in Sampling Surveys. *Far East Journal of Theoretical Statistics, 53*(5), 291-311. . https://doi.org/10.17654/TS053050291

Kikechi, C. B., Simwa, R. O., & Pokhariyal, G. P. (2018). On Local Linear Regression Estimation of Finite Population Totals in Model Based Surveys. *American Journal of Theoretical and Applied Statistics, 7*(3), 92-101. . https://doi.org/10.11648/j.ajtas.20180703.11

Montanari, G. E., & Ranalli, M. G. (2003). Nonparametric Methods in Survey Sampling. In: Vinci, M., Monari, P., Mignani, S. and Montanari, A., Eds., New Developments in Classification and Data Analysis, Springer, Berlin, 203-210.

Montanari, G. E., & Ranalli, M. G. (2005). Nonparametric Model Calibration Estimation in Survey Sampling. *Journal of the American Statistical Association, 100*, 1429-1442. https://doi.org/10.1198/016214505000000141

Nadaraya, E. A. (1964). On Estimating Regression. *Theory of Probability Applications, 10*, 186-190.

Odhiambo, R. O., & Mwalili, T. (2000). Nonparametric Regression for Finite Population Estimation. *East African Journal of Science, II*(2), 107-112.

Pfeffermann, D. (1993). The Role of Sampling Weights When Modeling Survey Data. *International Statistical Review, 61*(2), 317-337. https://doi.org/10.2307/1403631

Priestley, M. B., & Chao, M. T. (1972). Nonparametric Function Fitting. *Journal of the Royal Statistical Society, B34*, 384-392.

Royall, R. M. (1970a). On Finite Population Sampling under certain Linear Regression Models. *Biometrika, 57*, 377-387

Royall, R. M. (1970b). Finite Population Sampling-On Labels in Estimation. *Journal of the Annals of Mathematical Statistics, 41*, 1774-1779.

Royall, R. M. (1971). Linear Regression Models in Finite Population Sampling Theory Holt, Rinhart and Winston, Toronto, Canada, 54, 499-513.

Rueda, M. & Sanchez-Borrego, I. (2009). A Predictive Estimator of Finite Population Mean Using Nonparametric Regression. *Computational Statistics 24*, 1-14. https://doi.org/10.1007/s00180-008-0140-x

Ruppert, D., & Wand, M. P. (1994). Multivariate Locally Weighted Least Squares Regression. *Annals of Statistics, 22*, 1346-1370. https://doi.org/10.1214/aos/1176325632

Smith, T. M. (1976). The Foundations of Survey Sampling. *Journal of Royal Statistical Society Association, 139*, Part 2 183-204.

Stone, C. (1977). Consistent Nonparametric Regression. *Annals of Statistics, 5*, 595-645.

Watson, G. (1964). Smooth Regression Analysis. *Sankhya Series A, 26*, 359-372.