On Local Linear Regression Estimation of Finite Population Totals in Model Based Surveys

Conlet Biketi Kikechi*, Richard Onyino Simwa, Ganesh Prasad Pokhariyal

School of Mathematics, College of Biological and Physical Sciences, University of Nairobi, Nairobi, Kenya

Email address:
kikechiconlet@gmail.com (C. B. Kikechi), rsimwa@uonbi.ac.ke (R. O. Simwa), pokhariyal@uonbi.ac.ke (G. P. Pokhariyal)

*Corresponding author

To cite this article:
Conlet Biketi Kikechi, Richard Onyino Simwa, Ganesh Prasad Pokhariyal. On Local Linear Regression Estimation of Finite Population Totals in Model Based Surveys. American Journal of Theoretical and Applied Statistics. Vol. 7, No. 3, 2018, pp. 92-101. doi: 10.11648/j.ajtas.20180703.11

Received: February 10, 2018; Accepted: March 6, 2018; Published: March 24, 2018

Abstract: In this paper, nonparametric regression is employed which provides an estimation of unknown finite population totals. A robust estimator of finite population totals in model based inference is constructed using the procedure of local linear regression. In particular, robustness properties of the proposed estimator are derived and a brief comparison between the performances of the derived estimator and some existing estimators is made in terms of bias, MSE and relative efficiency. Results indicate that the local linear regression estimator is more efficient and performing better than the Horvitz-Thompson and Dorfman estimators, regardless of whether the model is specified or misspecified. The local linear regression estimator also outperforms the linear regression estimator in all the populations except when the population is linear. The confidence intervals generated by the model based local linear regression method are much tighter than those generated by the design based Horvitz-Thompson method. Generally the model based approach outperforms the design based approach regardless of whether the underlying model is correctly specified or not but that effect decreases as the model variance increases.

Keywords: Nonparametric Regression, Finite Population Totals, Local Linear Regression, Robustness Properties, Confidence Intervals, Model Based Surveys

1. Introduction

Integrated systems for survey designs and estimation methods to finite population inference have been considered by researchers in the past and categorised as design based approach, model assisted approach, combined inference approach and model based approach. Comparing and contrasting them in terms of their concepts of efficiency and robustness to assumptions about the characteristics of the population, it has been concluded that although none of these approaches delivers both efficiency and robustness, the model based approach seems to achieve the best compromise among the other approaches. In Chambers [1], a brief discussion on these survey strategies is accomplished. Kuo [2], Dorfman and Hall [3] and Kuk [4] apply nonparametric regression for estimating totals in finite populations.

There are two incompatible approaches for making inference from sample to population. In the traditional design-based approach, Horvitz and Thompson [5] use the probability structure of the procedure by which the sample s is selected to serve as the basis for inference in finite populations. In the model-based or predictive approach, Dorfman [6], use a regression model of the response Y on the predictor X to predict the non-sample Y's and by consequence, their total. Kikechi et al [7] employ a model based survey to estimate the unknown values of the survey variable using the local linear regression approach. In particular, the authors derive the properties of a local linear regression estimator and make variance comparisons between the derived estimator and the Nadaraya-Watson regression estimator which show that the two estimators are asymptotically equivalently efficient.

Researches done by Dorfman and Hall [3] and Chambers et al [8] have dwelt on estimating m(x), a smooth function. The expression for the asymptotic bias of this version of a non-parametric regression estimator of total does not include division by the sampling density, and so we expect the bias of a local linear regression based estimator be less sensitive to sparse x regions in the sample data. We make use of the local linear
regression technique to study the properties of the derived estimator and compare its performance with the existing estimators. Chambers and Dorfman [9] observe that the calibration estimator based on the columnar model does slightly better than the best linear unbiased estimator at high bandwidth. The estimator generally appears robust to changes in bandwidth, and gives exact unbiasedness and minimal variance for a particular weighted balanced sample.

They further noted that the estimators based on nonparametric model give approximate unbiasedness with no condition on balance and give approximate minimal variance, under approximate weighted balance. However, Fan and Gijbels [10] explore a more sophisticated method than the kernel regression, for example, the variable bandwidth local polynomial regression estimator.

Kijbels [10] explore a more sophisticated method than the kernel regression, for example, the variable bandwidth local polynomial regression estimator for the finite population total, based on local unknown regression function.

In Kikechi et al. [7], \( \widetilde{m}_{LL}(x_j) \) is derived and defined as under,

\[
\widetilde{m}_{LL}(x_j) = \sum_{i \in S} \left( \frac{S_{n,2} - S_{n,1}(x_i - x_j)}{S_{n,0}} \right) K \left( \frac{x_i - x_j}{h} \right) y_j + \sum_{i \in S} w_i(x_j) y_i,
\]

where,

\[
w_i(x_j) = \frac{S_{n,0} - S_{n,2}(x_i - x_j)}{S_{n,0}} k \left( \frac{x_i - x_j}{h} \right).
\]

3. Properties of the Local Linear Regression Estimator, \( \widetilde{T}_{LL} \)

In this section, consider the fixed equally spaced design model. The following assumptions made in Ruppert and Wand [16] are used to derive the properties of the local linear regression estimator:

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

(ii) The function \( m''(\cdot) \) 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:

\[
\int_{-\infty}^{\infty} K(x) dx = 1, \quad \int_{-\infty}^{\infty} xK(x) dx = 0, \quad \int_{-\infty}^{\infty} x^2K(x) dx > 0, \quad \int_{-\infty}^{\infty} K^2(x) dx < \infty, \quad d_k = \int_{-\infty}^{\infty} K^2(t) dt.
\]

(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 [17] imposed conditions on \( K(\cdot) \) and are only used for convenience in terms of technical arguments and thus can be relaxed.

Using equation (2) as proposed by Kikechi et al. [7], the local linear estimator of finite population total \( T \) can be estimated using,

\[
\widetilde{T}_{LL} = \sum_{i \in S} Y_i + \sum_{j \in R} \widetilde{m}_{LL}(x_j)
\]

2. The Proposed Estimator

The regression model for estimating the population total is given by,

\[
Y_i = m(X_i) + \sigma^2(X_i) \epsilon_i.
\]

Letting \( X_j \) be any point in the non-sample, and like in Dorfman [6], the estimator proposed by Kikechi et al. [7] is adopted and is defined by,

\[
\widetilde{T}_{LL} = \sum_{i \in S} Y_i + \sum_{j \in R} \widetilde{m}_{LL}(x_j)
\]

\( \widetilde{T}_{LL} \) is an estimator of the finite population total, where \( \widetilde{m}_{LL}(x_j) \) is a local linear regression estimator of \( m(x_j) \) at point \( x_j \).
\[ \bar{T}_{LL} = \sum_{i \in S} Y_i + \sum_{j \in R} \{ \sum_{i \in S} \left( \frac{S_{n,2} - S_{n,1}(x_i - x_j)}{(S_{n,0})(S_{n,2}) - (S_{n,1})} k \left( \frac{x_i - x_j}{h} \right) y_i \right) \} + \sum_{j \in R} \left( \frac{x_i - x_j}{S_{n,0}S_{n,2} - (S_{n,1})^2} \right) \sum_{i \in S} \left( S_{n,0}(x_i - x_j) - S_{n,1} \right) k \left( \frac{x_i - x_j}{h} \right) y_i \} \]

(3)

3.1. The Expectation of the Local Linear Regression Estimator, \( \bar{T}_{LL} \)

The expectation of \( \bar{T}_{LL} \) is derived as,

\[
E(\bar{T}_{LL}) = \sum_{i \in S} E(Y_i) + \sum_{j \in R} \left\{ \sum_{i \in S} \left( \frac{S_{n,2} - S_{n,1}(x_i - x_j)}{(S_{n,0})(S_{n,2}) - (S_{n,1})} k \left( \frac{x_i - x_j}{h} \right) E(Y_i) \right) \right\} + \sum_{j \in R} \left( \frac{x_i - x_j}{S_{n,0}S_{n,2} - (S_{n,1})^2} \right) \sum_{i \in S} \left( S_{n,0}(x_i - x_j) - S_{n,1} \right) k \left( \frac{x_i - x_j}{h} \right) E(Y_i) \}
\]

(4)

Using Taylor series expansion of the form,

\[ m(x_i) = m(x_j) + ht m'(x_j) + \frac{h^2 t^2}{2!} m''(x_j) + \ldots \]

(5)

Theorem 3 in Fan and Gijbels [10] is such that, under the conditions given in (i)-(v), we have,

\[
E(\bar{T}_{LL}) = \sum_{i \in S} m(x_i) + \sum_{j \in R} \left\{ \sum_{i \in S} \left( \frac{S_{n,2} - S_{n,1}(x_i - x_j)}{(S_{n,0})(S_{n,2}) - (S_{n,1})} k \left( \frac{x_i - x_j}{h} \right) m(x_i) \right) \right\} + \sum_{j \in R} \left( \frac{x_i - x_j}{S_{n,0}S_{n,2} - (S_{n,1})^2} \right) \sum_{i \in S} \left( S_{n,0}(x_i - x_j) - S_{n,1} \right) k \left( \frac{x_i - x_j}{h} \right) m(x_i) \}
\]

\[
= \sum_{i \in S} m(x_i) + \sum_{j \in R} \left\{ \sum_{i \in S} \left( S_{n,1} k \left( \frac{x_i - x_j}{h} \right) m(x_j) + ht m'(x_j) + \frac{h^2 t^2}{2!} m''(x_j) + \ldots \right) \right\} - \sum_{j \in R} \left\{ \sum_{i \in S} \left( S_{n,1} k \left( \frac{x_i - x_j}{h} \right) m(x_j) + ht m'(x_j) + \frac{h^2 t^2}{2!} m''(x_j) + \ldots \right) \right\} + \sum_{j \in R} \left( \frac{x_i - x_j}{S_{n,0}S_{n,2} - (S_{n,1})^2} \right) \sum_{i \in S} \left( S_{n,0}(x_i - x_j) - S_{n,1} \right) k \left( \frac{x_i - x_j}{h} \right) \left( m(x_j) + ht m'(x_j) + \frac{h^2 t^2}{2!} m''(x_j) + \ldots \right) \}
\]

\[
= \sum_{i \in S} m(x_i) + \sum_{j \in R} \left\{ \sum_{i \in S} \left[ \frac{S_{n,2} - S_{n,1}(x_i - x_j)}{(S_{n,0})(S_{n,2}) - (S_{n,1})} k \left( \frac{x_i - x_j}{h} \right) m(x_j) \right] + \sum_{j \in R} \left\{ \sum_{i \in S} \left( S_{n,1} k \left( \frac{x_i - x_j}{h} \right) m(x_j) + ht m'(x_j) + \frac{h^2 t^2}{2!} m''(x_j) + \ldots \right) \right\} \right\} + \sum_{j \in R} \left( \frac{x_i - x_j}{S_{n,0}S_{n,2} - (S_{n,1})^2} \right) \sum_{i \in S} \left( S_{n,0}(x_i - x_j) - S_{n,1} \right) k \left( \frac{x_i - x_j}{h} \right) m(x_j) \}
\]
= \sum_{i \in S} m(x_i) + \sum_{j \in R} m(x_j) + \sum_{j \in R} \left( (x_i - x_j) m'(x_j) \right) + \sum_{j \in R} \left( \left( \frac{\left( S_{n,2} - S_{n,1} \right)}{S_{n,0}} \right)^2 + (x_i - x_j) \left( \frac{S_{n,0} - S_{n,1} \right)}{S_{n,0}^2} \right) \frac{m''(x_j)}{2} \right)
(6)

3.2. The Bias of the Local Linear Regression Estimator, $\tilde{T}_{LL}$

The bias of $\tilde{T}_{LL}$ is derived as,

$$
Bias(\tilde{T}_{LL}) = \sum_{j \in R} \left\{ (x_i - x_j) m'(x_j) \right\}
+ \sum_{j \in R} \left\{ \left( \frac{\left( S_{n,2} - S_{n,1} \right)}{S_{n,0}} \right)^2 + (x_i - x_j) \left( \frac{S_{n,0} - S_{n,1} \right)}{S_{n,0}^2} \right) \frac{m''(x_j)}{2} \right\}
(7)

3.3. The Variance of the Local Linear Regression Estimator, $\bar{T}_{LL}$

The variance of the local linear regression estimator $\bar{T}_{LL}$ is estimated using the variance of the error. Then, $Var(\bar{T}_{LL} - T)$ is taken as an estimator of $Var(\bar{T}_{LL})$.

$$
Var(\bar{T}_{LL}) = \sum_{i \in S} y_i + \sum_{j \in R} \bar{m}_{LL}(x_j) - \sum_{i \in S} y_i - \sum_{j \in R} y_j
= \sum_{i \in S} \sum_{j \in R} \omega_i(x_j) y_i + \sum_{j \in R} (x_i - x_j) \sum_{i \in S} \omega_i(x_j) y_j - \sum_{j \in R} y_j
= \sum_{j \in R} \sum_{i \in S} \omega_i(x_j) \sigma^2(x_i) + \sum_{j \in R} (x_i - x_j)^2 \sum_{i \in S} \omega_i(x_j) \sigma^2(x_i) + \sum_{j \in R} \sigma^2(x_j)
(9)
$$

where,

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

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

The asymptotic expression for the variance of $\tilde{T}_{LL}$ is given by the expression using the results of $\bar{m}_{LL}(x_j)$ in Kikechi et al [7] that have been derived, thus,

$$
Var_{asy}(\tilde{T}_{LL}) = \frac{1}{nh} \sum_{j \in R} \sum_{i \in S} k^2 \frac{(x_i - x_j)}{h} \sigma^2(x_j) \frac{(x_i - x_i(-1))}{h} + \sum_{j \in R} (x_i - x_j)^2 \sum_{i \in S} 0. \sigma^2(x_i)
= \sum_{j \in R} \frac{d_k}{nh} \sigma^2(x_j)
(10)
$$

Note that in Kikechi et al [7], $Var_{asy}(\bar{m}_{LL}(x_j)) = \frac{d_k}{nh} \sigma^2(x_j)$ and $Var_{asy}(\bar{m}_{NW}(x_j)) = \frac{d_k}{nh} \sigma^2(x_j)$.

3.4. The MSE of the Local Linear Regression Estimator, $\bar{T}_{LL}$

Theorem I in Fan [17] allows that under condition (ii) we have,
\[ \text{MSE}(\hat{t}_{LL}) = (\text{Bias}(\hat{t}_{LL}))^2 + \text{Var}(\hat{t}_{LL}) \]

\[ = \left( \sum_{i \in R} (x_i - \hat{t}) m'(x_i) + \sum_{i \in R} \left( \frac{\left( S_{n,2} - S_{n,3} \frac{S_{n,2}}{2} \right)}{2} + (x_i - \hat{t}) \left( \frac{S_{n,0} S_{n,3} - S_{n,3} S_{n,2}}{S_{n,0} S_{n,2} - (S_{n,1})^2} \right) \right) m''(x_i) \right)^2 \]

\[ = \sum_{i \in R} \sum_{i \in S} w^2_i \sigma^2(x_i) + \sum_{i \in R} (x_i - \hat{t})^2 \sum_{i \in S} w^2_i \sigma^2(x_i) + \sum_{i \in R} \sigma^2(x_i) \]

The asymptotic expression for the MSE of the local linear regression estimator \( \hat{t}_{LL} \) is given by,

\[ \text{MSE}_{asy}(\hat{t}_{LL}) = \left( \sum_{i \in R} (x_i - \hat{t}) m'(x_i) + \sum_{i \in R} \left( h(k_2^2 + (x_i - \hat{t})k_3) m''(x_i) \right) \right)^2 \]

\[ + \sum_{i \in R} \frac{d_{xR}}{n} \sigma^2(x_i) \]

4. Simulation Study

In this section, a study is conducted on the performances of various estimators, including the proposed local linear regression estimator (2). In particular, we consider the design-based estimator, the parametric model-based estimator and the nonparametric model-based estimators.

4.1. Population Description

In this study, four populations are considered, which are generated from the regression model of the form,

\[ Y_i = m(X_i) + \sigma^2(X_i) e_i \]

where,

\[ E(Y_i / X_i = x_i) = m(x_i) \]

\[ \text{Cov}(Y_i, Y_j / X_i = x_i, X_j = x_j) = \begin{cases} \sigma^2(x_i), & i = j \\ 0, & i \neq j \end{cases} \]

The populations \( x_i \)'s are generated as independent and identically distributed (iid) uniform (0, 1) random variables. Four mean functions are considered with \( 1 \leq i \leq 200 \), namely;

- Linear: \( m_1(x) = 1 + 2(x - 0.5) \)
- Quadratic: \( m_2(x) = 1 + 2(x - 0.5)^2 \)
- Bump: \( m_3(x) = 1 + 2(x - 0.5) + \exp(-200(x - 0.5)^2) \)
- Jump: \( m_4(x) = 1 + 2(x - 0.5)I_{(x \leq 0.65)} + 0.65I_{(x > 0.65)} \)

The above mean functions represent the model specifications for the parametric and nonparametric estimators in consideration for cases where the model is correctly specified or incorrectly specified. The REG estimator is expected to be the best for \( m_1(x) \). The remaining mean functions; \( m_2(x) , m_3(x) \) and \( m_4(x) \) represent different deviations from the linear model. The errors are assumed to be independent and identically distributed (iid) random variables with mean 0 and constant variance.
The Epanechnikov kernel is used in this study for kernel smoothing on each of the populations because of its simplicity and easy computations using well designed computer programs. This is given by,

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

In Silverman. [18], the search for optimal bandwidth is done within the interval, $\frac{\sigma}{4n^{1/5}} \leq \frac{3\sigma}{2n^{1/5}}$ where $\sigma$ is the standard deviation of the $x_i$'s. In this study, the bandwidths are data driven and are determined by the least squares cross validation method.

Data simulations and computations are performed using the R computer software. A smaller population of size 200 is picked because the nonparametric local linear regression method is slower and takes more computer time to compute the estimates. The simulation has however been made exhaustive by performing 500 replications and thus the confidence in our conclusions. For each of the four 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. The population total is computed for each of the four populations and for each sample, and is defined by,

$$T(y) = \sum_{k=1}^{n} y_k$$  \hspace{1cm} (16)

The prediction errors for each of the estimators of finite population totals are computed as,

$$E_{HT} = (\bar{t}_{HT} - T)$$  \hspace{1cm} (17)

$$E_{REG} = (\bar{t}_{REG} - T)$$  \hspace{1cm} (18)

$$E_{DORF} = (\bar{t}_{DORF} - T).$$  \hspace{1cm} (19)

$$E_{LL} = (\bar{t}_{LL} - T)$$  \hspace{1cm} (20)

The biases for each of the estimators of finite population totals are computed as under,

$$B(\bar{t}_{HT}) = \sum_{i=1}^{500} \left(\frac{\bar{t}_{HT} - T}{500}\right).$$  \hspace{1cm} (21)

$$B(\bar{t}_{REG}) = \sum_{i=1}^{500} \left(\frac{\bar{t}_{REG} - T}{500}\right)$$  \hspace{1cm} (22)

$$B(\bar{t}_{DORF}) = \sum_{i=1}^{500} \left(\frac{\bar{t}_{DORF} - T}{500}\right)$$  \hspace{1cm} (23)

$$B(\bar{t}_{LL}) = \sum_{i=1}^{500} \left(\frac{\bar{t}_{LL} - T}{500}\right)$$  \hspace{1cm} (24)

The mean squared error for each of the estimators of finite population totals are computed as under,

$$MSE(\bar{t}_{HT}) = \sum_{i=1}^{500} \left(\frac{(\bar{t}_{HT} - T)^2}{500}\right).$$  \hspace{1cm} (25)

$$MSE(\bar{t}_{REG}) = \sum_{i=1}^{500} \left(\frac{(\bar{t}_{REG} - T)^2}{500}\right)$$  \hspace{1cm} (26)
\[ \text{MSE}(\hat{T}_{\text{DORF}}) = \sum_{i=1}^{500} \left( \frac{(\hat{T}_{\text{DORF}} - \bar{T})^2}{500} \right) \]  
\[ \text{MSE}(\hat{T}_{\text{LL}}) = \sum_{i=1}^{500} \left( \frac{(\hat{T}_{\text{LL}} - \bar{T})^2}{500} \right) \]  

The absolute bias (AB) is computed in order to analyze the performances of the proposed estimator versus some specified estimators using,

\[ \text{AB}(\hat{T}_{\text{HT}}) = \sum_{i=1}^{500} \left( \frac{(\hat{T}_{\text{HT}} - \bar{T})}{500} \right) \]  
\[ \text{AB}(\hat{T}_{\text{REG}}) = \sum_{i=1}^{500} \left( \frac{(\hat{T}_{\text{REG}} - \bar{T})}{500} \right) \]  
\[ \text{AB}(\hat{T}_{\text{DORF}}) = \sum_{i=1}^{500} \left( \frac{(\hat{T}_{\text{DORF}} - \bar{T})}{500} \right) \]  
\[ \text{AB}(\hat{T}_{\text{LL}}) = \sum_{i=1}^{500} \left( \frac{(\hat{T}_{\text{LL}} - \bar{T})}{500} \right) \]

The relative efficiency (RE) which examines the robustness of various estimators, i.e. the Horvitz-Thompson estimator, the REG estimator and the Dorfman estimator versus the proposed local linear estimator is computed as under,

\[ \text{RE}(\hat{T}_{\text{HT}}, \bar{T}_{\text{LL}}) = \frac{\sum_{i=1}^{500} (\hat{\bar{T}}_{\text{LL}} - \hat{T}_{\text{HT}})^2}{\sum_{i=1}^{500} (\hat{T}_{\text{HT}} - \bar{T})^2} \]  
\[ \text{RE}(\hat{T}_{\text{REG}}, \bar{T}_{\text{LL}}) = \frac{\sum_{i=1}^{500} (\hat{\bar{T}}_{\text{LL}} - \hat{T}_{\text{REG}})^2}{\sum_{i=1}^{500} (\hat{T}_{\text{REG}} - \bar{T})^2} \]  
\[ \text{RE}(\hat{T}_{\text{DORF}}, \bar{T}_{\text{LL}}) = \frac{\sum_{i=1}^{500} (\hat{\bar{T}}_{\text{LL}} - \hat{T}_{\text{DORF}})^2}{\sum_{i=1}^{500} (\hat{T}_{\text{DORF}} - \bar{T})^2} \]  

where \( \bar{T} \) is the finite population total estimator in consideration, \( T \) is the true population total and \( R = 500 \) is the number of replications.

The confidence intervals (CI) and the average lengths (AL) of the confidence intervals of various estimators are computed as under,

\[ CI(\hat{T}_{\text{HT}}) = \sum_{i=1}^{500} \hat{T}_{\text{HT}} \pm 1.96 \sqrt{\text{Var}(\hat{T}_{\text{HT}})} \]  
\[ CI(\hat{T}_{\text{REG}}) = \sum_{i=1}^{500} \hat{T}_{\text{REG}} \pm 1.96 \sqrt{\text{Var}(\hat{T}_{\text{REG}})} \]  
\[ CI(\hat{T}_{\text{DORF}}) = \sum_{i=1}^{500} \hat{T}_{\text{DORF}} \pm 1.96 \sqrt{\text{Var}(\hat{T}_{\text{DORF}})} \]  
\[ CI(\hat{T}_{\text{LL}}) = \sum_{i=1}^{500} \hat{T}_{\text{LL}} \pm 1.96 \sqrt{\text{Var}(\hat{T}_{\text{LL}})} \]

\[ AL(\hat{T}_{\text{HT}}) = \frac{1}{500} \sum_{i=1}^{500} (C_{\text{L}}(\hat{T}_{\text{HT}}) - C_{\text{U}}(\hat{T}_{\text{HT}})) \]  
\[ AL(\hat{T}_{\text{REG}}) = \frac{1}{500} \sum_{i=1}^{500} (C_{\text{L}}(\hat{T}_{\text{REG}}) - C_{\text{U}}(\hat{T}_{\text{REG}})) \]  
\[ AL(\hat{T}_{\text{DORF}}) = \frac{1}{500} \sum_{i=1}^{500} (C_{\text{L}}(\hat{T}_{\text{DORF}}) - C_{\text{U}}(\hat{T}_{\text{DORF}})) \]  
\[ AL(\hat{T}_{\text{LL}}) = \frac{1}{500} \sum_{i=1}^{500} (C_{\text{L}}(\hat{T}_{\text{LL}}) - C_{\text{U}}(\hat{T}_{\text{LL}})) \]

where \( C_{\text{L}} \) and \( C_{\text{U}} \) are respectively the lower and upper confidence intervals within which we expect our true population total to lie with 95% confidence.

### 4.2 Results

The results for the absolute biases, mean squared errors, relative efficiencies, confidence intervals and average length of confidence intervals for the various estimators are provided in tables 3, 4, 5, 6 and 7 respectively.

**Table 3. The Absolute Bias of various Estimators in the four Populations.**

|                      | HORVITZ-THOMPSON (HT) | LINEAR REGRESSION (REG) | DORFMAN (DORF) | LOCAL LINEAR (LL) |
|----------------------|-----------------------|-------------------------|----------------|------------------|
| Linear               | 139.1395              | 3.650095                | 3.628214       | 3.626798        |
| Quadratic            | 163.4725              | 1.226636                | 0.403125       | 0.4323062       |
| Bump                 | 157.7427              | 2.018801                | 0.4777851      | 0.4087753       |
| Jump                 | 1219.668              | 21.785                  | 9.760465       | 9.485367        |

**Table 4. The Mean Squared Error (MSE) of various Estimators in the four Populations.**

|                      | HORVITZ-THOMPSON (HT) | LINEAR REGRESSION (REG) | DORFMAN (DORF) | LOCAL LINEAR (LL) |
|----------------------|-----------------------|-------------------------|----------------|------------------|
| Linear               | 514.9775              | 15.36639                | 15.74559       | 15.47903        |
| Quadratic            | 453.5207              | 1.521063                | 0.1713249      | 0.160443        |
| Bump                 | 548.131               | 4.551133                | 0.2943485      | 0.1894413       |
| Jump                 | 35691.94              | 110.7915                | 110.7915       | 97.02299        |

**Table 5. The Relative Efficiency of various Estimators versus the proposed Local Linear Estimator.**

|                      | HORVITZ-THOMPSON (HT) | LINEAR REGRESSION (REG) | DORFMAN (DORF) |
|----------------------|-----------------------|-------------------------|----------------|
| Relative Efficiency  | Relative Efficiency   | Relative Efficiency     |                |
| Linear               | 0.99467563            | 0.8093                  | 0.936644       |
| Quadratic            | 0.001546731           | 0.9954403               | 0.962707       |
| Bump                 | 0.0002038478          | 0.02743355              | 0.9433107      |
| Jump                 | 0.003577862           | 0.1901854               | 0.9706123      |
bias of an estimator
parameter
between the expected value of
properties has small combined variance and bias.
bias (accuracy). An estimator that has good MSE
and dominates the Horvitz-Thompson estimator and th e
The bias for the Local Linear estimator is much low er than
compared to the other three finite population total  estimators.
for all the bi ases
estimator was the poorest resulting in large biases  as
(MSE), relative efficiency, confidence intervals an d
mean squared error (MSE) measures the average squar ed
¼̅
3
0
for an estimator. The MSE of an estimator
which is a somewhat reasonable measure of performan ce
estimator (precision) and the other one that measur es its
4.3. Discussion of Results
In this section, results of the bias, the mean square error
(MSE), relative efficiency, confidence intervals and average length of confidence intervals are discussed. The bias of an estimator \( \hat{\theta} \) of a parameter \( \theta \) is the difference between the expected value of \( \hat{\theta} \) and \( \theta \); that is, \( \text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta \). An estimator whose bias is identically equal to
0 is called an unbiased estimator and satisfies \( E(\hat{\theta}) = \theta \) for all \( \theta \). The larger the bias, the poorer the estimator. The mean squared error (MSE) measures the average squared difference between the estimator \( \hat{\theta} \) and the parameter \( \theta \), which is a somewhat reasonable measure of performance for an estimator. The MSE of an estimator \( \hat{\theta} \) of a parameter \( \theta \) is the function of \( \theta \) defined by \( E(\hat{\theta} - \theta)^2 \), and this is denoted as \( MSE_\hat{\theta} \). Thus, MSE has two components, one that measures the variability of the estimator (precision) and the other one that measures its bias (accuracy). An estimator that has good MSE properties has small combined variance and bias.

The relative efficiency of two estimators is the ratio of their efficiencies. If \( \hat{\theta}_1 \) and \( \hat{\theta}_2 \) are both unbiased estimators of \( \theta \), then the efficiency of \( \hat{\theta}_1 \) relative to \( \hat{\theta}_2 \) is \( \text{Eff}(\hat{\theta}_1, \hat{\theta}_2) = \frac{\text{Var}(\hat{\theta}_1)}{\text{Var}(\hat{\theta}_2)} \). If this is less than 1, then it implies that \( \text{Var}(\hat{\theta}_2) < \text{Var}(\hat{\theta}_1) \) and therefore \( \hat{\theta}_2 \) has a smaller variance than \( \hat{\theta}_1 \) and so \( \hat{\theta}_2 \) is preferred. Finally, confidence intervals consist of a range of values (interval) that act as good estimates of the unknown population parameter. The best performing confidence interval is one whose coverage rate is close to the true population and its length small.

4.3.1. The Absolute Bias
The biases for different estimators are summarised in table
3. In all the populations considered, the Horvitz-Thompson estimator was the poorest resulting in large biases as compared to the other three finite population total estimators. The bias for the Local Linear estimator is much lower than those of the other three estimators. For all the biases computed, the Local Linear Regression estimator is superior and dominates the Horvitz-Thompson estimator and the Linear Regression estimator for all the populations. The Local Linear estimator also dominates the Dorfman estimator for all the populations except when the population is quadratic.

4.3.2. The Mean Squared Error (MSE)
The MSE for different estimators are summarised in table 4. Generally the estimator with a smaller MSE is regarded as the most efficient one. The Local Linear Regression estimator is more efficient and performing better than the Horvitz-Thompson and Dorfman estimators, regardless of whether the model is specified or unspecified. The Local Linear estimator also outperforms the Linear Regression estimator in all the populations except when the population is linear. The Local Linear Regression estimator is not only superior to the popular Kernel Regression estimators, but it is also the best among all linear smoothers including those produced by orthogonal series and spline methods. In general, Local Linear estimation removes a bias term from the kernel estimator, that makes it have better behavior near the boundary of the \( x \)'s and smaller MSE everywhere.

4.3.3. The Relative Efficiency
Table 5 examines the robustness of various estimators i.e. the Horvitz-Thompson estimator, the REG estimator and the Dorfman estimator versus the proposed Local Linear estimator. The results in the table show that relative efficiency of the proposed Local Linear estimator to the Horvitz-Thompson estimator, the REG estimator and the Dorfman estimator is less than 1. This implies that the proposed Local Linear estimator has a smaller variance than the three estimators and thus the three estimators are less efficient than the Local Linear estimator. Generally, the Local Linear estimator outperforms the HT estimator, the REG estimator and the DORF estimator in all the populations. The Local Linear estimator is therefore robust and the most efficient estimator.

4.3.4. The Confidence Intervals and Their Average Length
The confidence intervals and average length of the

Table 6. The Confidence Intervals of various Estimators with respect to the four populations.

| THE 95% CONFIDENCE INTERVALS | HORVITZ-THOMPSON (HT) | LINEAR REGRESSION (REG) | DORFMAN (DORF) | LOCAL LINEAR (LL) |
|-----------------------------|----------------------|-------------------------|----------------|------------------|
| Lower Limit | Upper Limit | Lower Limit | Upper Limit | Lower Limit | Upper Limit | Lower Limit | Upper Limit |
| Linear | 65.43579 | 78.35652 | 62.92036 | 63.24861 | 62.75978 | 63.01298 | 62.62953 | 63.06378 |
| Quadratic | 61.74714 | 62.41275 | 60.29736 | 60.30645 | 60.25827 | 60.27853 | 60.44418 | 60.47615 |
| Bump | 88.43077 | 92.85335 | 93.01087 | 93.14516 | 92.06424 | 93.34889 | 91.91642 | 93.18671 |
| Jump | 503.6836 | 565.5807 | 479.9458 | 495.7306 | 460.7667 | 479.1529 | 465.1171 | 483.1778 |

Table 7. The Average Length of Confidence Intervals of various Estimators.

| THE AVERAGE LENGTH OF CONFIDENCE INTERVALS | HORVITZ-THOMPSON (HT) | LINEAR REGRESSION (REG) | DORFMAN (DORF) | LOCAL LINEAR (LL) |
|------------------------------------------|----------------------|-------------------------|----------------|------------------|
| Linear | 12.92073 | 0.3282467 | 0.2532001 | 0.4342478 |
| Quadratic | 0.6656047 | 0.090909092 | 0.02025908 | 0.03197243 |
| Bump | 4.422574 | 0.1342954 | 1.284649 | 1.270295 |
| Jump | 61.8971 | 15.78477 | 18.38621 | 18.06073 |
intervals are also measured for each case. A smaller length is better because it implies that the true population total is captured within a smaller range and therefore results are more precise. The confidence intervals generated by the model based Local Linear method are much tighter than those generated by the design based Horvitz-Thompson method, regardless of whether the model is specified or misspecified. The confidence intervals also indicate that the Local Linear method dominates the REG and Dorfman methods when the model is incorrectly specified. Generally, the model based estimators are much far better than the traditional design based estimators. The results show that the model based approach outperforms the design based approach at 95% coverage rate. The biases under the model based approach are also much lower than those for the design based approach in different populations.

4.4. Conclusion

In this paper, a model based estimator of finite population total has been constructed using the procedure of Local Linear regression. The Local Linear regression estimator has been derived and robustness properties studied. Results of the bias, mean squared error, relative efficiency, confidence intervals and average length of confidence intervals for the various estimators have been provided.

The bias results show that the Local Linear estimator dominates the Horvitz-Thompson estimator for the linear, quadratic, bump and jump populations. The MSE results show that the Local Linear estimator is performing better than the Horvitz-Thompson estimator and Dorfman estimator, irrespective of the model specification or misspecification. Results further indicate that the confidence intervals generated by the model based Local Linear procedure are much tighter than those generated by the design based Horvitz-Thompson method, regardless of whether the model is specified or misspecified. It has been observed that the model based approach outperforms the design based approach at 95% coverage rate.

Generally, the Local Linear Regression estimator is not only superior to the popular kernel regression estimators, but it is also the best among all linear smoothers including those produced by orthogonal series and spline methods. The estimator adapts well to bias problems at boundaries and in regions of high curvature and it does not require smoothness and regularity conditions required by other methods such as boundary kernels. Simulation experiments carried out on the proposed Local Linear regression estimator in comparison with some estimators that exist in the literature indicate that the proposed estimator is robust and is the most efficient estimator.

Acknowledgements

Firstly, I convey my deep and sincere appreciation to my instructors Prof. Simwa and Prof. Pokhariyal for their academic and professional guidance that led to the successful completion of the manuscript. I’m grateful for their excellent and diligent comments and suggestions throughout the research period.

Sincere thanks also go to Prof. Weke, Prof. Khalagai, Prof. Manene, Dr. Were and Dr. Luketero of the School of Mathematics of the University of Nairobi for their continued support, criticism, tireless efforts and availability that guided me to the very terminal point.

My family members have been a great pillar in my life. I humbly thank my wife Evelyn Nelima and my children Wayne Tolometi, Virginia Nakikmwei and Ryan Biketi for their moral support and understanding during the entire research period. I sincerely thank Papa Charles Kikechi and Mama Christine Kikechi for their great inspiration and encouragement since my early intellectual pursuit.

I wish to make a special acknowledgement to the technical team led by Prof. Antony Waititu for providing the technical support during data simulation. Special thanks are also due to the National Research Fund (NRF), by the Government of Kenya, for funding this research. The project or funding reference number is NRF/1/PhD/137.

Lastly, the authors thank the anonymous referees for their valuable comments and suggestions which led to the improvement of the manuscript.

References

[1] R. L. Chambers, "Which sample survey strategy? A review of three different approaches," Southampton statistical Sciences Research Institute, University of Southampton, 2003.

[2] L. Kuo, "Classical and prediction approaches to establishing distribution functions from survey data. Proceedings of the section on survey research method," American statistical Association, pp. 280-285, (1998).

[3] A. H. Dorfman and P. Hall, "Estimators of the finite population distribution function using non-parametric regression," Annals of statistics, vol. Vol 21, pp. 1452-1475, (1993).

[4] A. Kuk, "A Kernel method for estimating finite population distribution functions using auxiliary information," Biometrika, vol. Vol 80, pp. 385-392, (1993).

[5] V. P. Horvitz and D. J. Thompson, "A Generalization of Sampling Without Replacement From a Finite Universe," Journal of the American Statistical Association, Vols. 68, pp. 880-889, (1952).

[6] A. H. Dorfman, "Non-Parametric Regression for Estimating Totals in Finite populations, Proceedings of the Section on Survey Research Method.,” American Statistical Association, pp. 622-625, (1992).

[7] C. B. Kikechi, R. O. Simwa and G. P. Pokhariyal, "On Local Linear Regression Estimation in Sampling Surveys," Far east Journal of Theoretical Statistics., vol. 53, no. 5, pp. 291-311, (2017).

[8] R. L. Chambers, A. H. Dorfman and T. E. Wehrly, "Bias robust estimation in finite populations using nonparametric calibration," J. Amer Statist Assoc., Vols. 88., pp. 268-277, (1993).
[9] R. L. Chambers and A. H. Dorfman, "Nonparametric regression with complex survey data," Survey Methods Research Bureau of Labor Statistics, (2002).

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

[11] H. Zheng and R. J. Little, "Penalized Spline Model-Based Estimation of the Finite Population Total from Probability-Proportional-To-Size Samples," Journal of official Statistics, vol. 19, p. 99–117, (2003).

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

[13] F. J. Breidt and J. D. Opsomer, "Local Polynomial Regression Estimation in Survey Sampling," Annals of Statistics, vol. 28, pp. 1026-1053, (2000).

[14] B. I. Sanchez, J. D. Opsomer, M. Rueda and A. Arcos, "Nonparametric Estimation with mixed data types in survey sampling," Rev Mat complut., vol. 27, pp. 685-700, (2014).

[15] C. Luc, "Nonparametric kernel regression using complex survey data," Job market paper, (2016).

[16] D. Ruppert and M. P. Wand, "Multivariate locally weighted least squares regression," Annals of Statistics, vol. 22, p. 1346–1370, (1994).

[17] J. Fan, "Local linear regression smoothers and their minimax efficiencies," Annals of Statistics, vol. 21, p. 196–216, (1993).

[18] B. Silverman, "Density Estimation for Statistics and Data Analysis," in Monographs on Statistics and Applied Probability, London, Chapman and Hall, (1986).