Far East Journal of Theoretical Statistics

http://www.pphmj.com
http://dx.doi.org/10.17654/TS053050291
Volume 53, Number 5, 2017, Pages 291-311 **ISSN: 0972-0863**

ON LOCAL LINEAR REGRESSION ESTIMATION IN SAMPLING SURVEYS

Conlet B. Kikechi, Richard O. Simwa and Ganesh P. Pokhariyal

School of Mathematics
University of Nairobi
P.O. Box 30197-00100
Nairobi, Kenya
e-mail: kikechiconlet@yahoo.com
rsimwa@uonbi.ac.ke
pokhariyal@uonbi.ac.ke

Abstract

A model based survey is employed to estimate the unknown values of the survey variable, using the local linear regression approach. In particular, a local linear regression estimator in model based surveys is studied. Variance comparisons are made between the derived estimator and the Nadaraya-Watson regression estimator which show that the two estimators are asymptotically equivalently efficient.

1. Introduction

In survey sampling, auxiliary information on a finite population is regularly used to increase the precision of estimators of the finite population total. In general modeling process, complete auxiliary information is incorporated in the construction of estimators through fitted values.

Received: June 16, 2017; Revised: August 6, 2017; Accepted: August 28, 2017
2010 Mathematics Subject Classification: 62D05.
Keywords and phrases: survey sampling, auxiliary information, local linear regression estimator, model based surveys, nonparametric regression, asymptotic relative efficiency.

One approach to using this auxiliary information in estimation is to assume a working model ξ describing the relationship between the study variable of interest and the auxiliary variables. Estimators are then derived on the basis of this model. Estimators are sought which are efficient if the model is true, and which maintain desirable properties like design consistency if the model is false. Often, a linear model is selected as the working model, leading to the ratio and regression estimators [10], the best linear and unbiased estimators (BLUE) [3, 21], and the generalized regression estimators (GREG) [4].

Wu and Sitter [25] proposed a class of estimators for which the working models follow a nonlinear parametric shape. However, efficient use of any of these estimators requires a priori knowledge of the specific structure of the population. This becomes difficult if the working model is to be used for many variables of interest, a common occurrence in surveys.

Nonparametric models can also be applied. These nonparametric models do not restrict the functional form of the distribution nor does it specify the various stochastic properties such as $E_\xi(\cdot)$, $V_\xi(\cdot)$ and $MSE_\xi(\cdot)$. They allow for more robust inference than that obtained in parametric approach [11, 7, 6].

Nonparametric regression [19, 24] is such that estimation is frequently more flexible and robust than inference tied to probability distributions in design based inference or to parametric regression models in model based inference [11].

In [1], the traditional local polynomial regression estimator is used to estimate the unknown regression function $m(x)$. They assumed that $m(x)$ is a smooth function and obtained an asymptotically unbiased and consistent estimator of the finite population total. The local polynomial regression estimator has the form of the generalized regression estimator, but is based on a nonparametric super population model applicable to a much larger class of functions.

In [2], a related nonparametric model-assisted regression estimator is considered, replacing local polynomial smoothing with penalized splines. In [17], the local polynomial nonparametric regression estimation is extended to the two stage sampling, in which a probability sample of clusters is selected, and then subsamples of elements within each selected cluster are obtained. In their simulation study, the estimators are linear combinations of estimators of cluster totals with weights that are calibrated to known control totals.

In [16], the asymptotic properties of the model-assisted local linear estimator are studied under the combined inference approach. It is shown that the bias of the estimator, $\hat{m}(\cdot)$ is the same as in the identically independent distribution (iid) case but the variance equaled that from the iid case multiplied by a correction factor derived from the sampling scheme. In [23], the local polynomial fitting to a linear heteroscedastic regression mode is extended. Estimation of the finite population total in the presence of two auxiliary variables using the bootstrap method and jackknife method is considered in [20]. A comparison between the different methods is performed on the basis of mean squared error (MSE), mean absolute error (MAE) and mean absolute percentage error (MAPE).

The local linear regression procedure has potential advantages over other popular kernel based methods since it adapts well to bias problems at boundaries and in regions of high curvature. It can be tailored to work for many different distributional assumptions due to its simplicity. It does not require smoothness and regularity conditions. It is also asymptotically efficient among all linear smoothers including those produced by the kernel, orthogonal series and penalized spline methods [15, 14, 22]. In this paper, the local linear regression procedure is studied further leading to derivation of an asymptotically efficient estimator.

2. Derivation of the Local Linear Regression Estimator, $\overline{m}_{LL}(x_j)$

Consider a finite population P of size N labeled $U_1, U_2, ..., U_N$. The

pair (x_i, y_i), $i = 1, 2, ..., N$ is associated with each unit. The values $x_1, x_2, ..., x_N$ are known and can be used in the sample design, or in the estimator, or in both. The selection variable set **S** denotes sample of size n from P, for which y values are unknown. **S** is an ignorable set, that is given information on x, knowledge of how the sample was taken provides no additional information about y [11]. Let T_y be the finite population total of interest. The estimate of the population total is given by,

$$T_y = \sum_{i=1}^{N} Y_i = \sum_{i \in S} y_i + \sum_{i \in R} y_i, \tag{1}$$

where $\sum_{i \in S} y_i$ is known while $\sum_{i \in R} y_i$ is unknown such that R is an indexing set for the y-values which are unknown to the investigator.

Let

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

Such that

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

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

The estimator we propose is motivated by modeling the finite population of y_i's, conditioned on the auxiliary variable X_i as a realization from an infinite super population, ξ, in which

$$Y_i = m(X_i) + \sigma(X_i)\varepsilon_i, \tag{4}$$

where ε_i are independent random variables, with mean zero and variance $v(x_i)$. Further, $m(x_i)$ is a smooth function of x_i, and $v(x_i)$ is smooth and strictly positive.

Given x_i, then $m(x_i) = E_\xi(Y_i)$ is referred to as the regression function, while $v(x_i) = \text{var}_\xi(Y_i)$ is referred to as the variance function. The estimators of T are derived by noting that,

$$\begin{aligned} \overline{T} &= \sum_{i \in S} y_i + \left\{ E\left(\sum_{i \in R} \overline{Y}_i \right) \right\} \\ &= \sum_{i \in S} y_i + \sum_{i \in R} \overline{m}(x_i). \end{aligned} \tag{5}$$

The optimal predictor of this unknown quantity is given by

$$E\left(\sum_{i \in R} Y_i \right) = \sum_{i \in R} m(x_i). \tag{6}$$

However, $m(x_i)$ is unknown. We estimate $m(x_i)$ using the local linear procedure and then substitute it in equation (6) in order to get a local linear regression estimator of the finite population total given by

$$\overline{T}_{LL} = \sum_{i \in S} y_i + \sum_{i \in R} \overline{m}_{LL}(x_i), \tag{7}$$

where $\overline{m}_{LL}(x_i)$ is a local linear regression estimator of $m(x_i)$ at point x_i.

Letting x_j be any point in the non-sample, and as in [11], let

$$\overline{T}_{LL} = \sum_{i \in S} Y_i + \sum_{j \in R} \overline{m}_{LL}(x_j). \tag{8}$$

This defines an estimator of the finite population total, where $\overline{m}_{LL}(x_j)$ is a local linear regression estimator of $m(x_j)$ at point x_j.

Consider the regression model given in equation (2),

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

$$Cov(Y_i, Y_j/X_i = x_i, X_j = x_j) = \begin{cases} \sigma^2(x_i), & i = j, \\ 0, & i \neq j, \end{cases} \quad x_i \in x_j \pm h. \tag{9}$$

But by the Taylor series expansion, $m(x_i)$ is expressed as

$$m(x_i) = m(x_j + ht) = m(x_j) + htm'(x_j) + \frac{h^2t^2}{2}m''(x_j) + \frac{h^3t^3}{3}m'''(x_j) + \cdots, \tag{10}$$

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

$$+ \frac{(x_i - x_j)^3}{3!}m'''(x_j)\cdots. \tag{11}$$

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

$$y_i = \alpha + (x_i - x_j)\beta + \varepsilon_i, \tag{12}$$

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

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

Therefore, the task of estimating $m(x)$ is equivalent to a local linear regression task of estimating the intercept α. Now, if we consider a weighted local linear regression, we find α and β in order to minimize

$$\sum_{j=1}^{n}(y_j - \alpha - \beta(x_i - x_j))^2 K\left(\frac{x_i - x_j}{h}\right), \tag{13}$$

to obtain least squares estimators of α and β.

Equation (13) is a weighted least squares problem where the weights are given by the kernel functions $K\left(\frac{x_i - x_j}{h}\right)$. The function $K(\cdot)$ is a

symmetric probability density used in defining the estimator. The notation used here emphasizes the fact that the local linear regression is a weighted regression using data centered around x_j.

Let $\overline{\alpha}$ and $\overline{\beta}$ be the solutions to the weighted least square problem in equation (13). Then some mathematical analyses yield the following:

$$\overline{\alpha} = \left(\frac{\sum_{j=1}^{n} w_j y_j}{\sum_{j=1}^{n} w_j} \right), \tag{14}$$

where w_j is defined in equation (16). We therefore define the local linear regression estimator by

$$\overline{m}_{LL}(x) = \alpha = \left(\frac{\sum_{j=1}^{n} w_j y_j}{\sum_{j=1}^{n} w_j} \right), \tag{15}$$

where

$$w_j = K\left(\frac{x_i - x_j}{h} \right)(S_{n,2} - (x_i - x_j)S_{n,1}) \tag{16}$$

and

$$S_{n,r} = \sum_{j=1}^{n} K\left(\frac{x_i - x_j}{h} \right)(x_i - x_j)^r, \quad r = 1, 2. \tag{17}$$

We determine results in equations (16) and (17) as follows; by letting,

$$Q = \sum_{j=1}^{n} (y_j - \alpha - \beta(x_i - x_j))^2 K\left(\frac{x_i - x_j}{h} \right). \tag{18}$$

Differentiating equation (18) with respect to α, we get,

$$\frac{\partial Q}{\partial \alpha} = \sum_{j=1}^{n} -2(y_j - \alpha - \beta(x_i - x_j))K\left(\frac{x_i - x_j}{h} \right). \tag{19}$$

For the least value of α, we have

$$\sum_{j=1}^{n}(y_j - \alpha - \beta(x_i - x_j))K\left(\frac{x_i - x_j}{h}\right) = 0. \tag{20}$$

This implies that,

$$\sum_{j=1}^{n}K\left(\frac{x_i - x_j}{h}\right)y_j = \alpha\sum_{j=1}^{n}K\left(\frac{x_i - x_j}{h}\right) + \beta\sum_{j=1}^{n}K\left(\frac{x_i - x_j}{h}\right)(x_i - x_j)$$

$$= \alpha(S_{n,0}) + \beta(S_{n,1}). \tag{21}$$

Differentiating equation (18) with respect to β, we get

$$\frac{\partial Q}{\partial \beta} = \sum_{j=1}^{n} - 2(y_j - \alpha - \beta(x_i - x_j))K\left(\frac{x_i - x_j}{h}\right)(x_i - x_j). \tag{22}$$

For the least value of β, we have

$$\sum_{j=1}^{n}(y_j - \alpha - \beta(x_i - x_j))K\left(\frac{x_i - x_j}{h}\right)(x_i - x_j) = 0. \tag{23}$$

This implies that

$$\sum_{j=1}^{n}K\left(\frac{x_i - x_j}{h}\right)(x_i - x_j)y_j = \alpha\sum_{j=1}^{n}K\left(\frac{x_i - x_j}{h}\right)(x_i - x_j)$$

$$+ \beta\sum_{j=1}^{n}K\left(\frac{x_i - x_j}{h}\right)(x_i - x_j)^2$$

$$= \alpha(S_{n,1}) + \beta(S_{n,2}). \tag{24}$$

Solving equations (21) and (24) simultaneously by the elimination method, we have

$$S_{n,2}\sum_{j=1}^{n}K\left(\frac{x_i - x_j}{h}\right)y_j = \alpha(S_{n,0})(S_{n,2}) + \beta(S_{n,1})(S_{n,2}), \tag{25}$$

$$S_{n,1}\sum_{j=1}^{n} K\left(\frac{x_i - x_j}{h}\right)(x_i - x_j)y_j = \alpha(S_{n,1})^2 + \beta(S_{n,1})(S_{n,2}). \tag{26}$$

Now, eliminating β from equations (25) and (26), we get

$$\overline{\alpha} = \frac{S_{n,2}\sum_{j=1}^{n} k\left(\frac{x_i - x_j}{h}\right)y_j - S_{n,1}\sum_{j=1}^{n} K\left(\frac{x_i - x_j}{h}\right)(x_i - x_j)y_j}{(S_{n,0})(S_{n,2}) - (S_{n,1})^2}$$

$$= \frac{\sum_{j=1}^{n}\left(S_{n,2}k\left(\frac{x_i - x_j}{h}\right)y_j - S_{n,1}K\left(\frac{x_i - x_j}{h}\right)(x_i - x_j)y_j\right)}{(S_{n,0})(S_{n,2}) - (S_{n,1})^2}$$

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

$$= \sum_{j=1}^{n}\frac{(S_{n,2} - (x_i - x_j)S_{n,1})}{(S_{n,0})(S_{n,2}) - (S_{n,1})^2} K\left(\frac{x_i - x_j}{h}\right)y_j \tag{27}$$

which is analogous to equation (14).

In a similar way, eliminating α from equations (24) and (25), we get

$$\overline{\beta} = \sum_{j=1}^{n}\frac{(S_{n,0} - (x_i - x_j)S_{n,1})}{(S_{n,0})(S_{n,2}) - (S_{n,1})^2} K\left(\frac{x_i - x_j}{h}\right)y_j, \tag{28}$$

where $S_r(x_j;\, h) = \sum_{i=1}^{n}(x_i - x_j)^r K\left(\frac{x_i - x_j}{h}\right)$, $r = 0, 1, 2$.

In this section, the estimator $\overline{\beta}$ is determined using the set of data provided. Therefore from equation (12), we have

$$\bar{y}_i = \overline{\alpha} + (x_i - x_j)\overline{\beta} \tag{29}$$

such that

$$\overline{m}_{LL}(x_j) = \sum_{i\in S}\left\{\frac{(S_2(x_j;\,h) - S_1(x_j;\,h)(x_i - x_j))}{S_0(x_j;\,h)S_2(x_j;\,h) - S_1(x_j;\,h)^2} K\left(\frac{x_i - x_j}{h}\right) y_j\right\}$$

$$+ (x_i - x_j)\sum_{i\in S}\left\{\frac{(S_0(x_j;\,h) - S_1(x_j;\,h)(x_i - x_j))}{S_0(x_j;\,h)S_2(x_j;\,h) - S_1(x_j;\,h)^2} K\left(\frac{x_i - x_j}{h}\right) y_j\right\}$$

$$= \sum_{i\in S} w_i(x_j) y_j + (x_i - x_j)\sum_{i\in S} w_i'(x_j) y_j, \tag{30}$$

where

$$w_i(x_j) = \frac{(S_2(x_j;\,h) - S_1(x_j;\,h)(x_i - x_j))}{S_0(x_j;\,h)S_2(x_j;\,h) - S_1(x_j;\,h)^2} K\left(\frac{x_i - x_j}{h}\right) \tag{31}$$

and

$$w_i'(x_j) = \frac{(S_0(x_j;\,h) - S_1(x_j;\,h)(x_i - x_j))}{S_0(x_j;\,h)S_2(x_j;\,h) - S_1(x_j;\,h)^2} K\left(\frac{x_i - x_j}{h}\right). \tag{32}$$

3. Properties of the Local Linear Regression Estimator, $\overline{m}_{LL}(x_j)$

In deriving the properties of the local linear regression estimator, we need to make the following assumptions as in [22], namely,

(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:

$$\int_{-\infty}^{\infty} K(x)dx = 1,\ \int_{-\infty}^{\infty} xK(x)dx = 0,\ \int_{-\infty}^{\infty} x^2K(x)dx > 0,$$

$$\int_{-\infty}^{\infty} K^2 x dx < \infty,\ d_k = \int_{-\infty}^{\infty} K^2(t)dt,$$

(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 $h < x_j < 1 - h$.

In [15] some conditions are imposed on $K(\cdot)$ which are used only for convenience in terms of technical arguments and thus can be relaxed.

3.1. Expectation of the local linear regression estimator, $\overline{m}_{LL}(x_j)$

Therefore, it follows from the definition of the estimator in equation (30) that the expectation of $\overline{m}_{LL}(x_j)$ is

$$E(\overline{m}_{LL}(x_j)) = \sum_{i \in S} w_i(x_j)E(y_j) + (x_i - x_j)\sum_{i \in S} w_i'(x_j)E(y_j)$$

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

$$+ (x_i - x_j)\sum_{i \in S}\left\{\frac{(S_{n,0} - S_{n,1}(x_i - x_j))}{(S_{n,0})(S_{n,2}) - (S_{n,1})^2} K\left(\frac{x_i - x_j}{h}\right)E(y_j)\right\}. \quad (33)$$

Consider the Taylor series,

$$m(x_i) = m(x_j) + (x_i - x_j)m'(x_j) + \frac{(x_i - x_j)^2}{2!} m''(x_j) + \cdots, \quad (34)$$

for the local linear regression procedure in a small neighbourhood of a point x_j.

Theorem 3 in [15] is such that, under the conditions given in (i)-(v),

$$E(\overline{m}_{LL}(x_j))$$

$$= \sum_{i \in S}\left\{w_i(x_j)\left(m(x_j) + htm'(x_j) + \frac{h^2t^2}{2} m''(x_j) + \cdots\right)\right\}$$

$$+(x_i - x_j)\sum_{i\in S}\left\{w_i'(x_j)\left(m(x_j) + htm'(x_j) + \frac{h^2t^2}{2}m''(x_j) + \cdots\right)\right\}$$

$$= \left\{\frac{S_{n,2}}{(S_{n,0})(S_{n,2}) - (S_{n,1})^2}\right\}\left\{(S_{n,0})m(x_j) + (S_{n,1})m'(x_j) + \frac{(S_{n,2})}{2}m''(x_j) + \cdots\right\}$$

$$-\left\{\frac{S_{n,1}}{(S_{n,0})(S_{n,2}) - (S_{n,1})^2}\right\}\left\{(S_{n,1})m(x_j) + (S_{n,2})m'(x_j) + \frac{(S_{n,3})}{2}m''(x_j) + \cdots\right\}$$

$$+\left\{\frac{(x_i - x_j)S_{n,1}}{(S_{n,0})(S_{n,2}) - (S_{n,1})^2}\right\}\left\{(S_{n,1})m(x_j) + (S_{n,2})m'(x_j) + \frac{(S_{n,3})}{2}m''(x_j) + \cdots\right\}$$

$$-\left\{\frac{(x_i - x_j)S_{n,1}}{(S_{n,0})(S_{n,2}) - (S_{n,1})^2}\right\}\left\{(S_{n,0})m(x_j) + (S_{n,1})m'(x_j) + \frac{(S_{n,2})}{2}m''(x_j) + \cdots\right\}$$

$$= \left\{\frac{((S_{n,0})(S_{n,2}) - (S_{n,1})^2) + (x_i - x_j)((S_{n,0})(S_{n,1}) - (S_{n,0})(S_{n,1}))}{(S_{n,0})(S_{n,2}) - (S_{n,1})^2}\right\}m(x_j)$$

$$+\left\{\frac{((S_{n,1})(S_{n,2}) - (S_{n,1})(S_{n,2})) + (x_i - x_j)((S_{n,0})(S_{n,2}) - (S_{n,1})^2)}{(S_{n,0})(S_{n,2}) - (S_{n,1})^2}\right\}m'(x_j)$$

$$+\left\{\frac{((S_{n,2})^2 - (S_{n,1})(S_{n,3})) + (x_i - x_j)((S_{n,0})(S_{n,3}) - (S_{n,1})(S_{n,2}))}{(S_{n,0})(S_{n,2}) - (S_{n,1})^2}\right\}\frac{m''(x_j)}{2}$$

$$= m(x_j) + (x_i - x_j)m'(x_j)$$

$$+\left\{\frac{((S_{n,2})^2 - (S_{n,1})(S_{n,3})) + (x_i - x_j)((S_{n,0})(S_{n,3}) - (S_{n,1})(S_{n,2}))}{(S_{n,0})(S_{n,2}) - (S_{n,1})^2}\right\}$$

$$\times \frac{m''(x_j)}{2}. \tag{35}$$

3.2. The biasness of the local linear regression estimator, $\overline{m}_{LL}(x_j)$

The bias of the estimator $\overline{m}_{LL}(x_j)$ is expressed as

$$Bias(\overline{m}_{LL}(x_j)) = (x_i - x_j)m'(x_j) + \left\{\frac{((S_{n,2})^2 - (S_{n,1})(S_{n,3})) + (x_i - x_j)((S_{n,0})(S_{n,3}) - (S_{n,1})(S_{n,2}))}{(S_{n,0})(S_{n,2}) - (S_{n,1})^2}\right\} \times \frac{m''(x_j)}{2}. \quad (36)$$

The asymptotic expression of the bias of $\overline{m}_{LL}(x_j)$ may be obtained by making the assumption that, x_i's are fixed uniform design points in (0, 1) [18, 13]. Therefore,

$$\sum_{i \in S} (x_i - x_j)^l k\left(\frac{x_i - x_j}{h}\right) = nh^{l+1}k_l + o(nh^{l+3}), \quad (37)$$

is almost surely uniform for $x \in (0, 1)$ and $h \in H_n$, where $H_n = [C_1 n^{-E_1}, C_2 n^{-E_2}]$, $0 < E_2 < E_1 < 1$, and $C_1, C_2 > 0$.

This implies that,

$$S_{n,0} = nh + o(nh^3), \quad S_{n,1} = o(nh^4), \quad S_{n,2} = nh^3k_2 + o(nh^5),$$

$$S_{n,3} = nh^4k_3 + o(nh^6) \text{ and } S_{n,4} = nh^5k_4 + o(nh^7)$$

such that

$$(S_{n,0})(S_{n,2}) - (S_{n,1})^2 = \{nh + o(nh^3)\}\{nh^3k_2 + o(nh^5)\}\{o(nh^4)\}^2 = n^2h^4k_2 + o(n^2h^6), \quad (38)$$

$$(S_{n,2})^2 - (S_{n,1})(S_{n,3}) = \{nh^3k_2 + o(nh^5)\}^2 - \{o(nh^4)\}\{nh^4k_3 + o(nh^6)\} = n^2h^6k_2^2 + o(n^2h^8), \quad (39)$$

$$S_{n,0}S_{n,3} - S_{n,1}S_{n,2} = \{nh + o(nh^3)\}\{nh^4k_3 + o(nh^6)\} - \{o(nh^4)\}\{nh^3k_2 + o(nh^5)\}$$

$$= n^2h^5k_3 + o(n^2h^7), \tag{40}$$

$$Bias_{asy}(\overline{m}_{LL}(x_j))$$

$$= (x_i - x_j)m'(x_j)$$

$$+ \frac{\{n^2h^6k_2^2 + o(n^2h^8) + (x_i - x_j)(n^2h^5k_3 + o(n^2h^7))\}m''(x_j)}{2(n^2h^4k_2 + o(n^2h^6))}$$

$$= (x_i - x_j)m'(x_j) + \frac{h(hk_2^2 + (x_i - x_j)k_3)m''^{(x_j)}}{2k_2}. \tag{41}$$

3.3. Variance of the local linear regression estimator, $\overline{m}_{LL}(x_j)$

$$Var(\overline{m}_{LL}(x_j)) = Var\left\{\sum_{i\in S} w_i(x_j)y_i + (x_i - x_j)\sum_{i\in S} w_i'(x_j)y_i\right\}$$

$$= \sum_{i\in S} w_i^2(x_j)\sigma^2(x_i) + (x_i - x_j)^2\sum_{i\in S} w_i'^2(x_j)\sigma^2(x_i), \tag{42}$$

where

$$w_i^2(x_j) = \left\{\frac{(S_{n,2} - S_{n,1}(x_i - x_j))}{(S_{n,0})(S_{n,2}) - (S_{n,1})^2}\,k\left(\frac{x_i - x_j}{h}\right)\right\}^2 \tag{43}$$

and

$$w_i'^2(x_j) = \left\{\frac{(S_{n,0}(x_i - x_j) - S_{n,1})}{(S_{n,0})(S_{n,2}) - (S_{n,1})^2}\,k\left(\frac{x_i - x_j}{h}\right)\right\}^2. \tag{44}$$

The asymptotic expression of the variance of $\overline{m}_{LL}(x_j)$ is obtained as

$w_i^2(x_j)$

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

$$\approx \left\{\frac{1}{nh}k\left(\frac{x_i - x_j}{h}\right)\frac{(n^2h^4k_2 + o(n^2h^6))}{(n^2h^4k_2 + o(n^2h^6))}\right\}^2$$

$$\approx \frac{1}{n^2h^2}k^2\left(\frac{x_i - x_j}{h}\right), \tag{45}$$

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

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

$$\approx \left\{\frac{1}{nh}k\left(\frac{x_i - x_j}{h}\right)\frac{(o(n^2h^5) + o(n^2h^7) - o(n^2h^5) - o(n^2h^7))}{(n^2h^4k_2 + o(n^2h^6))}\right\}^2$$

$$\approx 0. \tag{46}$$

Then

$$Var_{asy}(\overline{m}_{LL}(x_j)) = \frac{1}{nh}\sum_{i\in S}k^2\left(\frac{x_i - x_j}{h}\right)\sigma^2(x_i)\left(\frac{x_i - x_{i-1}}{h}\right)$$

$$+ (x_i - x_j)^2\sum_{i\in S}0.\sigma^2(x_i)$$

$$= \sum_{j\in R}\frac{d_k}{nh}\sigma^2(x_j), \tag{47}$$

where

$$d_k = \int k^2(t)dt.$$

3.4. The MSE of the local linear regression estimator, $\overline{m}_{LL}(x_j)$

Theorem 1 in [14] allows that under condition (ii) we have

$$MSE(\overline{m}_{LL}(x_j))$$

$$= \{Bias(\overline{m}_{LL}(x_j))\}^2 + Var(\overline{m}_{LL}(x_j))$$

$$= \left\{(x_i - x_j)m'(x_j) + \left(\frac{((S_{n,2})^2 - S_{n,1}S_{n,3}) + (x_i - x_j)(S_{n,0}S_{n,3} - S_{n,1}S_{n,2})}{(S_{n,0})(S_{n,2}) - (S_{n,1})}\right)\frac{m''(x_j)}{2}\right\}^2$$

$$+ \sum_{i \in S} w_i^2(x_j)\sigma^2(x_i) + (x_i - x_j)^2 \sum_{i \in S} w_i'^2(x_j)\sigma^2(x_i). \quad (48)$$

The asymptotic expression of the mean square error is also obtained using the asymptotic bias and asymptotic variance expressions of $\overline{m}_{LL}(x_j)$ such that,

$$MSE_{asy}(\overline{m}_{LL}(x_j)) = \left\{(x_i - x_j)m'(x_j) + \frac{h(hk_2^2 + (x_i - x_j)k_3)m''(x_j)}{2k_2}\right\}^2$$

$$+ \frac{d_k}{nh}\sigma^2(x_j). \quad (49)$$

3.5. Unbiasedness and efficiency of the local linear regression estimator, $\overline{m}_{LL}(x_j)$

The efficiency of an estimator refers to how much information it extracts about the parameter of interest from the sample. A more efficient estimator extracts more information, in some sense, from a sample of a given size. Efficiency measures information extracted by the variance of an unbiased estimator, that is, smaller variance means greater efficiency.

3.5.1. Introduction

An estimator is efficient if it is the minimum variance unbiased estimator. Let $X_1, ..., X_n$ be a random sample from some distribution which

depends on a parameter T and let $\overline{T} = \overline{T}(X_1, ..., X_n)$ be an estimator of T. Then $\overline{T}$ is an unbiased estimator of T if $E(\overline{T}) = T$, $\overline{T}$ is an asymptotically unbiased estimator of T if $\lim_{n\to\infty} E(\overline{T}) = T$, $\overline{T}$ is an efficient estimator of T if it is unbiased and its variance achieves the Cramer-Rao lower bound; that is if

$$Var(\overline{T}) = \frac{1}{nI(T)}, \tag{50}$$

and the efficiency of an unbiased estimator $\overline{T}$ of T is the ratio of the Cramer-Rao lower bound to the variance of the estimator; that is,

$$Eff(\overline{T}) = \frac{1/nI(T)}{Var(\overline{T})}. \tag{51}$$

We remark that it must be true that $Eff(\overline{T}) \leq 1$. The smaller the value of the efficiency, the less efficient the estimator. Also $\overline{T}$ is an asymptotically efficient estimator of T if it is unbiased or asymptotically unbiased such that

$$\lim_{n\to\infty} Eff(\overline{T}) = 1. \tag{52}$$

In what follows, we make variance comparisons between the Nadaraya-Watson regression estimator [19] and the proposed local linear regression estimator, in terms of their asymptotic relative efficiency.

3.5.2. Asymptotic relative efficiency

The relative efficiency of two procedures is the ratio of their efficiencies, although often this concept is used where the comparison is made between a given procedure and a notional best possible procedure. The efficiencies and the relative efficiency of two procedures theoretically depend on the sample size available for the given procedure, 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 comparison measure.

If $\overline{T}_1$ and $\overline{T}_2$ are both unbiased estimators of T, then the relative efficiency of $\overline{T}_1$ to $\overline{T}_2$ is given by

$$Eff(\overline{T}_1, \overline{T}_2) = \frac{Var(\overline{T}_2)}{Var(\overline{T}_1)}. \tag{53}$$

If $Eff(\overline{T}_1, \overline{T}_2) < 1$, then $\overline{T}_2$ has a smaller variance than $\overline{T}_1$ and $\overline{T}_1$ is less efficient than $\overline{T}_2$.

If $\overline{T}_1$ and $\overline{T}_2$ are both unbiased or asymptotically unbiased estimators of T, then the asymptotic relative efficiency of $\overline{T}_1$ to $\overline{T}_2$ is given by,

$$ARE(\overline{T}_1, \overline{T}_2) = \lim_{n\to\infty} Eff(\overline{T}_1, \overline{T}_2) = \lim_{n\to\infty} \frac{Var(\overline{T}_2)}{Var(\overline{T}_1)}. \tag{54}$$

The mean regression functions, $m(x)$ for the Nadaraya-Watson regression estimator and the proposed local linear regression estimator, respectively, can be expressed as follows:

$$\overline{m}_{NW}(x_j) = \sum_{i=1}^{n} w_i(x) y_i, \tag{55}$$

$$\overline{m}_{LL}(x_j) = \sum_{i\in S} w_i(x_j) y_j + (x_i - x_j) \sum_{i\in S} w_i'(x_j) y_j. \tag{56}$$

The variance of the Nadaraya-Watson regression estimator $\overline{m}(x_j)$ is given by [19],

$$Var(\overline{m}_{NW}(x_j)) = d_k \sigma^2(x_j) + \sum_{i\in s} \left\{ w_i^2(x_j) \left(ht\sigma^{2'}(x_j) + \frac{h^2 t^2}{2} \sigma^{2''}(x_j) + \cdots \right) \right\}. \tag{57}$$

The asymptotic expression for the variance of the Nadaraya-Watson regression estimator $\overline{m}_{NW}(x_j)$ is estimated by [19],

$$Var_{asy}(\overline{m}_{NW}(x_j)) \approx \sum_{j\in R} \frac{d_k}{nh} \sigma^2(x_j). \tag{58}$$

The variance of the local linear regression estimator $\overline{m}_{LL}(x_j)$ is given by

$$Var(\overline{m}_{LL}(x_j)) = \sum_{i\in S} w_i^2(x_j)\sigma^2(x_i) + (x_i - x_j)^2 \sum_{i\in S} w_i'^2(x_j)\sigma^2(x_i). \quad (59)$$

The asymptotic expression for the variance of the local linear regression estimator $\overline{m}_{LL}(x_j)$ is estimated by

$$Var_{asy}(\overline{m}_{LL}(x_j)) = \sum_{j\in R} \frac{d_k}{nh}\sigma^2(x_j). \quad (60)$$

Thus, the asymptotic relative efficiency of the Nadaraya-Watson regression estimator to the proposed local linear regression estimator is given by

$$ARE(\overline{m}_{NW}(x_j), \overline{m}_{LL}(x_j)) = \frac{Var(\overline{m}_{LL}(x_j))}{Var_{asy}(\overline{m}_{NW}(x_j))} = \frac{\sum_{j\in R}\frac{d_k}{nh}\sigma^2(x_j)}{\sum_{j\in R}\frac{d_k}{nh}\sigma^2(x_j)} = 1. \quad (61)$$

4. Discussion

The main objective was to obtain a consistent robust estimator using the procedure of local linear regression in model based surveys. The procedure is based on locally fitting a line rather than a constant. Unlike kernel regression, locally linear estimation would have no bias if the true model were linear. The resulting local linear estimator has minimal asymptotic variance in comparison with the Nadaraya-Watson estimator.

Asymptotically, there is no difference in the performance of the Nadaraya-Watson regression estimator and the proposed local linear regression estimator. The reason for this being that their ratio converges to 1 as n becomes large, see equation (61). Thus, the two estimators are equivalently asymptotically efficient.

Acknowledgement

Special thanks are due to the National Research Fund (NRF), by the Government of Kenya, for funding this research.

The authors thank the anonymous referees for their valuable suggestions which led to the improvement of the manuscript.

References

[1] F. J. Breidt and J. D. Opsomer, Local polynomial regression estimation in survey sampling, Ann. Statist. 30 (2000), 952-975.

[2] F. J. Breidt, G. Claeskens and J. D. Opsomer, Model-assisted estimation for complex surveys using penalised splines, Biometrika 92(4) (2005), 831-846.

[3] K. R. W. Brewer, Ratio estimation in finite populations: some results deductible from the assumption of an underlying stochastic process, Austral. J. Statist. 5 (1963), 93-105.

[4] C. M. Cassel, C. E. Sarndal and J. H. Wretman, Foundations of Inference in Survey Sampling, Wiley, New York, 1977.

[5] R. L. Chambers and A. H. Dorfman, Robust Sample Survey Inference via Bootstrapping and Bias Correction; the Case of the Ratio Estimator, Technical Report, Southampton Statistical Sciences Research Institute, University of Southampton, 2002.

[6] R. L. Chambers and A. H. Dorfman, Nonparametric Regression with Complex Survey Data, Survey Methods Research Bureau of Labor Statistics, 2002.

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

[8] K. Chen, J. Fan and Z. Jin, Design-adaptive minimax local linear regression for longitudinal/clustered data, Statist. Sinica 18 (2008), 515-534.

[9] K. Chen and Z. Jin, Local polynomial regression analysis for clustered data, Biometrika 92 (2005), 59-74.

[10] W. G. Cochran, Sampling Techniques, New York, John Wiley, 1977.

[11] A. H. Dorfman, Non-parametric regression for estimating totals in finite populations, Proceedings of the Section on Survey Research Methods, Amer. Statist. Assoc. 1992, pp. 622-625.

[12] A. H. Dorfman and P. Hall, Estimators of the finite population distribution function using non-parametric regression, Ann. Statist. 21 (1993), 1452-1475.

[13] R. Eubank and L. Speckman, Local polynomial fitting in adopted to the autoregressive context for modeling nonlinear time series under some missing conditions, Amer. J. Stat. Assoc. 88(2) (1993), 1287-1301.

[14] J. Fan, Local linear regression smoothers and their minimax efficiencies, Ann. Stat. 21 (1993), 196-216.

[15] J. Fan and I. Gijbels, Local Polynomial Modeling and its Applications, Monographs on Statistics and Applied Probability, Chapman and Hall, London, 1996.

[16] T. Harms and P. Duchesne, On kernel nonparametric regression designed for complex survey data, Biometrika 72 (2010), 111-138.

[17] J. Kim, F. Breidt and J. Opsomer, Nonparametric Regression Estimation of Finite Population Totals Under Two Stage Cluster Sampling, Technical Report, Department of Statistics, Colorado State University, 2009.

[18] E. Masry, Multivariate local polynomial regression for time series analysis. Uniform strong consistency and rates, J. Time Series Anal. 17 (1996), 571-599.

[19] E. A. Nadaraya, On estimating regression, Theory Probab. Appl. 9(1) (1964), 141-142.

[20] E. A. Rady and D. Ziedan, Estimation of population total using local polynomial regression with two auxiliary variables, J. Stat. Appl. Pro. 3(2) (2014), 129-136.

[21] R. M. Royall, On finite population sampling under certain linear regression models, Biometrika 57 (1970), 377-387.

[22] D. Ruppert and M. P. Wand, Multivariate locally weighted least squares regression, Ann. Statist. 22 (1994), 1346-1370.

[23] L. Su, Y. Zhao and T. Yan, Two Stage Method Based on Local Polynomial Fitting for a Linear Hederoscedastic Regression Model and its Applications in Economics, Hindawi Publishing Corporation, 2012, pp. 1-7.

[24] G. S. Watson, Smooth regression analysis, Sankhya, Ser. A 17 (1964), 359-372.

[25] C. Wu and R. R. Sitter, A model-calibration approach to using complete auxiliary information from survey data, J. Amer. Statist. Assoc. 96 (2001), 185-193.