Journal of Statistics Applications & Probability
An International Journal

http://dx.doi.org/10.12785/jsap/030203

Estimation of Population Total using Local Polynomial Regression with Two Auxiliary Variables

*EL-Housseiny A. Rady and Dalia Ziedan**

Institute of Statistical Studies & Research, Cairo University, Cairo, Egypt

Received: 29 Sep. 2013, Revised: 12 Mar. 2014, Accepted: 16 Mar. 2014
Published online: 1 Jul. 2014

Abstract: In this paper, the estimation for finite population total of a study variable will be considered, and the local linear regression will be used. The study variable is available for the sample and is supplemented by two auxiliary variables, which are available for every element in the finite population. Also, the resampling methods will be combined with the local linear regression method to estimate the total. The comparisons between different methods will be performed based on the mean squared error (MSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). A simulation study is carried out to assess the effects.

Keywords: survey sampling, auxiliary variables, local linear regression, bootstrap and jackknife

1 Introduction

Survey sampling often supplies information about a study variable only for sampled elements. However, auxiliary information is often available for the entire population. The relationship of the auxiliary information with the study variable across the sample allows inferences about the non-sampled portion of the population. Thus, the use of auxiliary information at the estimation stage of a survey improves the precision of the estimates parameters studied. One approach to using this auxiliary information in estimation is to assume a working model describing the relationship between the study variable of interest and the auxiliary variables. Estimators are then derived on the basis of this model.

Usually a parametric approach is used to represent the relationship between the auxiliary variables and the study variable. But in some situations, the parametric model is not appropriate, and the resulting estimators do not achieve any efficiency gain over pure estimators. A natural alternative was first suggested by Kuo (1988) for the distribution function, that adopts a nonparametric approach, which does not place any restrictions on the relationship between the auxiliary data and the study variable. Other important works in this topic are Chambers et al. (1993), Drofman (1993), Drofman and Hall (1993) and Rueda and Arcos (1998).

Breidt and Opsomer (2000) used the traditional local polynomial regression estimator for the unknown regression function $m(x)$. They assume that $m(x)$ is a smooth function of x and obtained an asymptotically design-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 superpopulation model applicable to a much larger class of functions. Breidt, Claeskens, and Opsomer (1995) considered a related nonparametric model-assisted regression estimator, replacing local polynomial smoothing with penalized splines. Kim, Breidt, and Opsomer (2009) extended the local polynomial nonparametric regression estimation to two-stage sampling, in which a probability sample of clusters is selected, and then subsamples of elements within each selected cluster are obtained. In this paper, we concerned with the estimation the finite population total in the presence of the two auxiliary variables using the local polynomial regression.

2 Multiple Regression

Suppose now that the covariate is d-dimensional, where

$$X_i = (x_{i1}, x_{i2}, \ldots, x_{id})'$$

* Corresponding author e-mail: dalia_dalia444@yahoo.com

In this case,

$$Y = m(x_1, x_2, ..., x_d) + \varepsilon$$

For local linear regression, the kernel function K is defined as a function of d variables. Given a nonsingular positive definite $d \times d$ bandwidth matrix H, we define

$$K_H(x) = \frac{1}{|H|^{1/2}} K\left(H^{-1/2} x\right). \tag{1}$$

Often, one scales each covariate to have the same mean and variance and then we use the kernel

$$h^{-d} K\left(\|x\| / h\right) \tag{2}$$

where K is any one-dimensional kernel. Then there is a single bandwidth parameter h. At a target value $x = (x_1, x_2, ..., x_d)'$, the local sum of squares is given by

$$\sum_{i=1}^{n} w_i(x) \left(Y_i - a_0 - \sum_{j=1}^{d} a_j (x_{ij} - x_j)\right)^2 \tag{3}$$

where,

$$w_i(x) = K\left(\|x_i - x\| / h\right)$$

In this case, the estimator is

$$\hat{m}(x) = \hat{a}_0 \tag{4}$$

where $\hat{a} = (\hat{a}_0, \hat{a}_1, ..., \hat{a}_d)'$ is the value of $a = (a_0, a_1, ..., a_d)'$ that minimizes the weighted sums of squares. The solution $\hat{a}$ is

$$\hat{a} = \left(X'WX\right)^{-1} X'WY \tag{5}$$

where X in this case is

$$X = \begin{pmatrix} 1 & x_{11} - x_1 & \dots & x_{1d} - x_d \\ 1 & x_{21} - x_1 & \dots & x_{2d} - x_d \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} - x_1 & \dots & x_{nd} - x_d \end{pmatrix}$$

and W is the diagonal matrix whose (i, i) element. For more details [see Casella, G. *et al* (2006)].

3 Estimation of Total in the Case of Two Auxiliary Variables

In this case

$$X = \begin{pmatrix} 1 & x_{11} - x_{1j} & x_{21} - x_{2j} \\ 1 & x_{12} - x_{1j} & x_{22} - x_{2j} \\ \vdots & \vdots & \vdots \\ 1 & x_{1n} - x_{1j} & x_{2n} - x_{2j} \end{pmatrix}, \quad j = 1, 2, ..., N$$

and

$$X' = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_{11} - x_{1j} & x_{12} - x_{1j} & \dots & x_{1n} - x_{1j} \\ x_{21} - x_{2j} & x_{22} - x_{2j} & \dots & x_{2n} - x_{2j} \end{pmatrix}, \quad j = 1, 2, ..., N$$

Let $\Delta_{1ij} = x_{1i} - x_{1j}, \qquad \Delta_{2ij} = x_{2i} - x_{2j}$

in this case, $w_{iij} = k\left(\frac{1}{h}\sqrt{\left(\Delta_{1ij}^2 + \Delta_{2ij}^2\right)}\right)$ and

$$W = \begin{pmatrix} w_{1j} & 0 & \cdots & 0 \\ 0 & w_{2j} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_{nj} \end{pmatrix}$$

where $w_{ij} = w_{ij}$. So, we will substitute in the equation$\hat{m}(x) = (X'WX)^{-1} X'WY$ by *X, W and Y* to get the estimation of the total. Hence

$$X'W = \begin{pmatrix} w_{1j} & w_{2j} & \cdots & w_{nj} \\ \Delta_{11j}w_{1j} & \Delta_{12j}w_{2j} & \cdots & \Delta_{1nj}w_{nj} \\ \Delta_{21j}w_{1j} & \Delta_{22j}w_{2j} & \cdots & \Delta_{2nj}w_{nj} \end{pmatrix}, \text{ and}$$

$$X'WX = \begin{pmatrix} \ell_{11} & \ell_{12} & \ell_{13} \\ \ell_{21} & \ell_{22} & \ell_{23} \\ \ell_{31} & \ell_{32} & \ell_{33} \end{pmatrix}, \text{ where}$$

$$\ell_{11} = \sum_i w_{ij} \qquad \ell_{12} = \sum_i \Delta_{1ij} w_{ij}$$

$$\ell_{13} = \sum_i \Delta_{2ij} w_{ij} \qquad \ell_{22} = \sum_i \Delta_{1ij}^2 w_{ij}$$

$$\ell_{23} = \sum_i \Delta_{1ij}\Delta_{2ij} w_{ij} \qquad \ell_{33} = \sum_i \Delta_{2ij}^2 w_{ij}$$

Note that: $(X'WX)$ is a symmetric matrix.
Hence, the inverse of the matrix $(X'WX)$ is

$$(X'WX)^{-1} = \frac{1}{|X'WX|} (Adj\,(X'WX))$$

The second term in the estimation of $\hat{a}$ is

$$X'WY = \begin{pmatrix} \sum_i w_{ij} y_i \\ \sum_i \Delta_{1ij} w_{ij} y_i \\ \sum_i \Delta_{2ij} w_{ij} y_i \end{pmatrix}$$

where

$$\hat{a} = (X'WX)^{-1} X'WY.$$

Since our primary interest is to compute an estimate of *Y*, the necessary computations are limited to the ones that estimate the parametera_0. Therefore, the estimator is simplified to

$$\hat{y}_j = \hat{a}_0 = e_1' (X'WX)^{-1} X'WY$$

where e_1 is a column vector with the first element equal to one, and the rest equal to zero. Then

$$\hat{y}_j = \hat{a}_0 = \sum_{a=1}^{3} s_{1a} \sum_{i=1}^{n} \Delta_{(a-1)ij} w_{ij} y_i \Big/ |X'WX| \tag{6}$$

where $\Delta_{0ij} = 1$, and

$$s_{11} = \left(\sum_i \Delta_{1ij}^2 w_{ij}\right)\left(\sum_i \Delta_{2ij}^2 w_{ij}\right) - \left(\sum_i \Delta_{1ij}\Delta_{2ij} w_{ij}\right)^2$$

$$s_{12} = \left(\sum_i \Delta_{1ij}\Delta_{2ij} w_{ij}\right)\left(\sum_i \Delta_{2ij} w_{ij}\right) - \left(\sum_i \Delta_{1ij} w_{ij}\right)\left(\sum_i \Delta_{1ij}\Delta_{2ij} w_{ij}\right)$$

$$s_{13} = \left(\sum_i \Delta_{1ij} w_{ij}\right)\left(\sum_i \Delta_{1ij}\Delta_{2ij} w_{ij}\right) - \left(\sum_i \Delta_{1ij}^2 w_{ij}\right)\left(\sum_i \Delta_{2ij} w_{ij}\right)$$

Now, our main purpose is to estimate the total (*T*). Therefore, according to Drofman (1992) the estimate of the total is

$$\hat{T} = \sum_{i=1}^{n} y_i + \sum_{j=n+1}^{N} \hat{y}_j \tag{7}$$

Substitute from equation (6) in (7), the estimated total is

$$\hat{T} = \sum_{i=1}^{n} y_i + \sum_{j\neq i}^{N} \frac{1}{D_j} \sum_{a=1}^{3} s_{1a} \sum_{i=1}^{n} \Delta_{(a-1)ij} w_{ij} y_i$$

$$= \sum_{i=1}^{n} \left(1 + \sum_{j\neq i}^{N} \frac{1}{D_j} \sum_{a=1}^{3} s_{1a} \Delta_{(a-1)ij} w_{ij}\right) y_i \tag{8}$$

where $D = |X'WX|$

4 Bootstrapping Local Linear Regression for Estimating the Total

Efron(1979) has developed a new resampling procedure named as "Bootstrap" . Bootstrap resample consists of n elements that are drawn randomly from the n original data observations with replacement (Friedl & Stampfer, 2002). The all bootstrap samples are n^n , but we choose B bootstrap samples. Bootstrapping can be done by either resampling the residuals, in which the regressors $(x_1, x_2, ..., x_d)$ are assumed to be fixed, or resampling the y_i values and their associated x_i values, in which the regressors are assumed to be random. In our study, we deal with the residuals resampling, where the bootstrap technique with nonparametric regression to estimate the total of the population will be used and the local linear regression will also be considered. Suppose we have a univariate response variable Y and two auxiliary variables X_1 and X_2 , then the nonparametric regression model is

$$Y_i = m(X_{1i}, X_{2i}) + \varepsilon_i, \quad i = 1, ..., n$$

and the bootstrap procedure based on the resampling errors can be summarized as follows:
(1) Let $Y = (Y_1, Y_2, ..., Y_n)$ denote the sample of observations was selected from the generated population. Then based on the sample Y the local linear regression estimator $\hat{m}(x)$ is given by

$$\hat{y}_i = \hat{m}(x_{1i}, x_{2i}) = e_1' \left(X'WX\right)^{-1} X'WY,$$

(2) Calculate the residuals as following
$\hat{\varepsilon}_i = Y_i - \hat{m}(x_{1i}, x_{2i})$, i = 1, 2, . . . , n.
(3) Define the centered residuals by $\tilde{\varepsilon}_i = \hat{\varepsilon}_i - \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i$.
(4) Draw with replacement a random sample of size from the residuals, $\tilde{\varepsilon}_1, \tilde{\varepsilon}_2, ..., \tilde{\varepsilon}_n$, were calculated in step (3) giving $1/n$ probability for each $\tilde{\varepsilon}_i$ values. This gives n-bootstrap sample of the residuals ε_i^*, *i = 1,2, . . . , n.* [See Stine (1985, 1990) and Wu (1986)].
(5) The bootstrap sample of observations is constructed by adding a randomly sampled residual to the original predicted value for each observation. After resampling, new observations is given by

$$Y_i^* = \hat{m}(x_{1i}, x_{2i}) + \varepsilon_i^*.$$

(6) Obtain the local linear estimate from the first bootstrap sample as follows:

$$\hat{Y}_i^{*(1)} = e_1' \left(X'WX\right)^{-1} X'WY^*$$

(7) Repeat the steps 4, 5 and 6, B times.
Then, the bootstrap estimate is

$$\hat{Y}_i^* = \frac{1}{B}\sum_{r=1}^{B} \hat{Y}_i^{*(r)} \tag{9}$$

Now, we will estimate the total using local linear regression estimation with bootstrap method, since we have

$$T = \sum_{j=1}^{N} Y_j = \sum_{i=1}^{n} Y_i + \sum_{j \neq i}^{N} Y_j$$

but $\sum_{j\neq i}^{N} Y_j$ is unknown , so we will estimate it as:

$$\hat{T}^* = \sum_{i=1}^{n} Y_i + \sum_{j \neq i}^{N} \hat{Y}_j^*$$

$$= \sum_{i=1}^{n} Y_i + \sum_{j \neq i}^{N} \frac{1}{B}\sum_{r=1}^{B} \frac{1}{D_j} \sum_{a=1}^{3} s_{1a} \sum_{i=1}^{n} \Delta_{(a-1)ij} w_{ij} \hat{Y}_i^{*(r)} \tag{10}$$

5 Jackknifing Local Linear Regression for Estimating the Total

In this Section, the algorithm of estimating the total using local linear regression method with jackknife technique will be given. The technique of deleting single case from the original sample (delete one jackknife) sequentially will be used. Suppose the dataset consists of n vectors (Y_i, X_{1i}, X_{2i}), where Y_i is the study variable and X_{1i}, X_{2i} are considered auxiliary variables. For simplicity, let $x_i = (x_{1i}, x_{2i})$and $d_k = (y_k, x_k)$, $k = 1, 2, ..., n$ denote the values associated with ith observation. In this case, the set of observations is the vector $(d_1, d_2, ..., d_n)$. Then, the jackknife procedure based on delete-one is as follows.

(1) Draw n sized sample from population randomly and label the elements of the vector $d_k = (y_k, x_k)$, $k = 1, 2, ..., n$.

(2) Omit first observation of the vector $d_k = (y_k, x_k)$ and label the remaining *n-1* sized observation set $Y_{(1)}^{(J)} = (y_2, ..., y_n)$, *and* $X_{(1)}^{(J)} = (x_2, ..., x_n)$ as delete-one jackknife sample $d_{(1)}^{(J)}$.

(3) Obtain the local linear regression estimate $\hat{m}^{(J1)}(x_j)$ from $d_{(1)}^{(J)}$.

(4) Omit the second element of the vector $d_i = (y_i, x_i)$ and label remaining *n-1* sized observation set $Y_{(2)}^{(J)} = (y_1, y_3, ..., y_n)$, *and* $X_{(2)}^{(J)} = (x_1, x_3, ..., x_n)$ as $d_{(2)}^{(J)}$.

(5) Obtain the local linear regression estimate $\hat{m}^{(J2)}(x_j)$ from $d_{(2)}^{(J)}$.

(6) Similarly, omit each one of the n observations (there is n samples jackknife each of them has $n-1$ observations) and estimate the local linear regression$\hat{m}^{(Jk)}(x_j)$, where $\hat{m}^{(Jk)}(x_j)$ is the jackknife local linear regression estimate after deleting of k^{th} observation from $d_k = (y_k, x_k)$.

(7) Then, the jackknife estimate of $\hat{m}(x_j)$ is

$\hat{m}^{(J)}(x_j) = \frac{1}{n}\sum_{k=1}^{n}\hat{m}^{(Jk)}(x_j) = \frac{1}{n}\sum_{k=1}^{n}\frac{1}{D_j}\sum_{a=1}^{3}s_{1a}\sum_{i=1}^{n-1}\Delta_{(a-1)ij}w_{ij}y_{ik}$.

(8) Using the jackknife estimate of $\hat{m}(x_j)$ in estimating the total

$$\hat{T}^{(J)} = \sum_{i=1}^{n}y_i + \sum_{j\neq i}^{N}\hat{m}^{(J)}(x_j) = \sum_{i=1}^{n}y_i + \sum_{j\neq i}^{N}\frac{1}{n}\sum_{k=1}^{n}\frac{1}{D_j}\sum_{a=1}^{3}s_{1a}\sum_{i=1}^{n-1}\Delta_{(a-1)ij}w_{ij}y_{ik} \tag{11}$$

6 Performance Criteria of the Models

The performance of the model is related with how close are the prediction values to the observed values. Three different consistency criteria are used in order to compare among different methods. These are mean square error (MSE), mean absolute error (MAE) and mean absolute percentage error (MAPE) respectively which are defined as follows:

1.$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.
2.$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$.
3.$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{|y_i|}(100\%)$.

7 Simulation studies

Sometimes in sampling, we do not usually observe all the survey information. That is, the survey variable Y is not observable for all the population units. Auxiliary variable X, is often used to estimate the unobserved survey variables. One way of overcoming the above problem is the super population approach, in which a working model relating the two auxiliary variables is assumed. In this study, we simulate data from four models, which introduced by Ye *et al* (2006), each with $Y = m(X_1, X_2) + \delta(X_1)\varepsilon$, where $\varepsilon \sim N(0,1)$.

Model (1): $m_1(x_1, x_2) = x_1 x_2$

$$\delta_1^2(x_1, x_2) = \left(x_1^2 - 0.04\right) I_{\left(x_1^2 > o.04\right)} + 0.01$$

Model (2): $m_2(x_1, x_2) = x_1 \exp\left(-2x_2^2\right)$

$$\delta_2^2(x_1, x_2) = 2.5\left(x_1^2 - 0.04\right) I_{\left(x_1^2 > o.04\right)} + 0.025$$

Model (3): $m_3(x_1, x_2) = x_1 + 2\sin(1.5x_2)$

$$\delta_3^2(x_1, x_2) = \left(x_1^2 - 0.04\right) I_{\left(x_1^2 > o.04\right)} + 0.01$$

Table 1 MSE, MAE, and MAPE of the total estimation under different methods with different sample sizes and bandwidths for model 1

$h = n^{-1/3}$									
Method	**n = 25**			**n = 50**			**n = 100**		
	MSE	**MAE**	**MAPE**	**MSE**	**MAE**	**MAPE**	**MSE**	**MAE**	**MAPE**
CLR	**404.25**	**29.63**	**85.5%**	**397.85**	**27.24**	**74.2%**	**396.56**	**25.99**	**27.0%**
LLR	**332.25**	**25.85**	**68.8%**	**325.87**	**23.49**	**57.9%**	**324.58**	**22.26**	**17.8%**
LLB	**329.22**	**21.55**	**42.1%**	**322.84**	**19.23**	**32.2%**	**321.55**	**18.01**	**6.4%**
LLJ	**336.14**	**28.14**	**72.3%**	**329.76**	**25.76**	**61.2%**	**328.47**	**24.52**	**19.6%**
$h = n^{-1/5}$									
CLR	**373.16**	**27.35**	**79.0%**	**367.25**	**25.14**	**68.5%**	**366.05**	**24.00**	**25.0%**
LLR	**306.69**	**23.86**	**63.5%**	**300.81**	**21.68**	**53.4%**	**299.61**	**20.55**	**16.4%**
LLB	**303.89**	**19.89**	**38.9%**	**298.01**	**17.75**	**29.8%**	**296.81**	**16.62**	**5.9%**
LLJ	**310.28**	**25.97**	**66.7%**	**304.40**	**23.78**	**56.5%**	**303.20**	**22.63**	**18.1%**
$h = n^{-1/7}$									
CLR	**435.35**	**31.90**	**92.1%**	**428.46**	**29.33**	**79.9%**	**427.06**	**27.99**	**29.1%**
LLR	**357.81**	**27.84**	**74.1%**	**350.94**	**25.30**	**62.3%**	**349.55**	**23.97**	**19.2%**
LLB	**354.54**	**23.21**	**45.4%**	**347.68**	**20.71**	**34.7%**	**346.28**	**19.40**	**6.9%**
LLJ	**362.00**	**30.30**	**77.8%**	**355.13**	**27.74**	**65.9%**	**353.74**	**26.41**	**21.1%**

Table 2 MSE, MAE, and MAPE of the total estimation under different methods with different sample sizes and bandwidths for model 2

$h = n^{-1/3}$									
Method	**n = 25**			**n = 50**			**n = 100**		
	MSE	**MAE**	**MAPE**	**MSE**	**MAE**	**MAPE**	**MSE**	**MAE**	**MAPE**
CLR	**513.09**	**37.60**	**108.6%**	**504.97**	**34.57**	**94.2%**	**503.32**	**32.99**	**34.3%**
LLR	**421.70**	**32.81**	**87.3%**	**413.61**	**29.81**	**73.4%**	**411.97**	**28.25**	**22.6%**
LLB	**426.64**	**35.71**	**91.7%**	**418.55**	**32.69**	**77.7%**	**416.90**	**31.12**	**24.9%**
LLJ	**417.85**	**27.35**	**53.5%**	**409.76**	**24.40**	**40.9%**	**408.12**	**22.86**	**8.1%**
$h = n^{-1/5}$									
CLR	**502.83**	**36.85**	**106.4%**	**494.87**	**33.88**	**92.3%**	**493.26**	**32.33**	**33.6%**
LLR	**413.27**	**32.16**	**85.5%**	**405.34**	**29.22**	**72.0%**	**403.73**	**27.68**	**22.2%**
LLB	**409.49**	**26.80**	**52.4%**	**401.57**	**23.92**	**40.1%**	**399.96**	**22.40**	**7.9%**
LLJ	**418.11**	**35.00**	**89.9%**	**410.17**	**32.04**	**76.2%**	**408.56**	**30.50**	**24.4%**
$h = n^{-1/7}$									
CLR	**519.31**	**38.06**	**109.9%**	**511.09**	**34.99**	**95.4%**	**509.43**	**33.39**	**34.7%**
LLR	**426.81**	**33.21**	**88.3%**	**418.62**	**30.18**	**74.3%**	**416.96**	**28.59**	**22.9%**
LLB	**422.92**	**27.68**	**54.1%**	**414.73**	**24.70**	**41.4%**	**413.06**	**23.14**	**8.2%**
LLJ	**431.81**	**36.15**	**92.9%**	**423.62**	**33.09**	**78.7%**	**421.96**	**31.50**	**25.2%**

Model (4): $m_4(x_1,x_2) = \sin(x_1+x_2) + 2\exp(-2x_2^2)$

$$\delta_4^2(x_1,x_2) = 3\left(x_1^2 - 0.04\right) I_{\left(x_1^2 > o.04\right)} + 0.03$$

The populations of X_1 and X_2 are generated as independent and identically distributed (iid) Uniform (-2, 2) random variables.

The simulation experiments will be performed to compare the performance of the local linear regression estimator with the classic linear regression estimator. Also, the effects of the bootstrap and the jackknife techniques on those estimators will be studied. The simulation will be carried out as follows:

1. Firstly, we generate population of size $N = 1000$ as above.
2. The simple random samples will be chosen from the population and different sizes will be considered, namely $n = 25, 50, \text{and } 100$ respectively

Table 3 MSE, MAE, and MAPE of the total estimation under different methods with different sample sizes and bandwidths for model 3

$h = n^{-1/3}$									
Method	**n = 25**			**n = 50**			**n = 100**		
	MSE	**MAE**	**MAPE**	**MSE**	**MAE**	**MAPE**	**MSE**	**MAE**	**MAPE**
CLR	**466.44**	**34.18**	**98.7%**	**459.06**	**31.43**	**85.7%**	**457.57**	**29.99**	**31.2%**
LLR	**383.36**	**29.83**	**79.4%**	**376.01**	**27.10**	**66.8%**	**374.51**	**25.68**	**20.6%**
LLB	**379.86**	**24.86**	**48.6%**	**372.51**	**22.19**	**37.2%**	**371.02**	**20.78**	**7.4%**
LLJ	**387.85**	**32.47**	**83.4%**	**380.50**	**29.72**	**70.7%**	**379.00**	**28.29**	**22.7%**
$h = n^{-1/5}$									
CLR	**444.68**	**32.59**	**94.1%**	**437.64**	**29.96**	**81.7%**	**436.21**	**28.59**	**29.7%**
LLR	**365.47**	**28.44**	**75.6%**	**358.46**	**25.84**	**63.6%**	**357.04**	**24.48**	**19.6%**
LLB	**362.14**	**23.70**	**46.3%**	**355.13**	**21.15**	**35.5%**	**353.70**	**19.81**	**7.0%**
LLJ	**369.75**	**30.95**	**79.5%**	**362.74**	**28.33**	**67.4%**	**361.32**	**26.97**	**21.6%**
$h = n^{-1/7}$									
CLR	**478.88**	**35.10**	**101.3%**	**471.30**	**32.26**	**87.9%**	**469.77**	**30.79**	**32.0%**
LLR	**393.59**	**30.62**	**81.5%**	**386.03**	**27.83**	**68.5%**	**384.50**	**26.37**	**21.1%**
LLB	**389.99**	**25.53**	**49.9%**	**382.44**	**22.78**	**38.2%**	**380.91**	**21.34**	**7.5%**
LLJ	**398.20**	**33.33**	**85.6%**	**390.64**	**30.51**	**72.5%**	**389.11**	**29.05**	**23.3%**

Table 4 MSE, MAE, and MAPE of the total estimation under different methods with different sample sizes and bandwidths for model 4

$h = n^{-1/3}$									
Method	**n = 25**			**n = 50**			**n = 100**		
	MSE	**MAE**	**MAPE**	**MSE**	**MAE**	**MAPE**	**MSE**	**MAE**	**MAPE**
CLR	**539.73**	**41.02**	**118.4%**	**550.88**	**37.71**	**102.8%**	**549.08**	**35.99**	**37.4%**
LLR	**440.04**	**35.79**	**95.2%**	**451.21**	**32.52**	**80.1%**	**449.42**	**30.82**	**24.7%**
LLB	**435.84**	**29.84**	**58.3%**	**447.01**	**26.62**	**44.6%**	**445.22**	**24.94**	**8.8%**
LLJ	**445.42**	**38.96**	**100.1%**	**456.60**	**35.67**	**84.8%**	**454.80**	**33.95**	**27.2%**
$h = n^{-1/5}$									
CLR	**522.42**	**38.29**	**110.5%**	**514.15**	**35.20**	**95.9%**	**512.48**	**33.59**	**34.9%**
LLR	**429.37**	**33.41**	**88.9%**	**421.13**	**30.36**	**74.8%**	**419.46**	**28.76**	**23.0%**
LLB	**425.45**	**27.85**	**54.4%**	**417.21**	**24.85**	**41.7%**	**415.54**	**23.27**	**8.2%**
LLJ	**434.40**	**36.36**	**93.4%**	**426.16**	**33.29**	**79.1%**	**424.48**	**31.69**	**25.4%**
$h = n^{-1/7}$									
CLR	**565.95**	**41.48**	**119.8%**	**557.00**	**38.13**	**103.9%**	**555.18**	**36.39**	**37.9%**
LLR	**465.15**	**36.19**	**96.3%**	**456.22**	**32.89**	**81.0%**	**454.41**	**31.16**	**24.9%**
LLB	**470.60**	**39.39**	**101.2%**	**461.67**	**36.06**	**85.7%**	**459.86**	**34.33**	**27.5%**
LLJ	**460.90**	**30.17**	**59.0%**	**451.98**	**26.92**	**45.1%**	**450.17**	**25.21**	**8.9%**

Secondly, for each sample, we estimate the total $T = \sum_{i=1}^{n} Y_i + \sum_{j \neq i}^{N} m(x_j)$. The linear regression and the local linear regression will be used to estimate $m(x)$. Also, the bootstrap and the jackknife techniques will be combined with those regression methods to estimate $m(x)$. We consider the normal kernel function with different bandwidth values $h = n^{-1/3}, n^{-1/5}$ and $n^{-1/7}$ for the local linear regression, each simulation setting is applied to all four models and repeated $M = 1000$ times.

Thirdly, the mean square error (MSE) of the total (T) under the two types of the regression methods will be calculated. Also, the mean absolute error (MAE) and the mean absolute percentage error (MAPE) will be calculated.

Finally, the effects of the bootstrap and the jackknife techniques on the estimation of total (T) will be studied, these effects based on the bias, MSE, MAE, MAPE.

Tables 1, 2, 3 and 4 reveals the values of the mean squared error (MSE), mean absolute error (MAE) and the mean absolute percentage error (MAPE) of the estimators for the four models, when the sample size (n) has different values $n = 25, 50,$ and 100 and the bandwidth has values $h = n^{-1/3}, n^{-1/5}$ and $n^{-1/7}$.

8 Results of the Simulation Study

Tables 1, 2, 3 and 4 summarize the following conclusions about our simulation study:

1.For the four models the local linear regression estimator dominates the classical linear regression estimator when the regression model is incorrectly specified.
2.The local linear regression estimator with bootstrap is overall the best choice for all models and bandwidths under study.
3.The effect of the bootstrap on the estimator is better than the jackknife at the most.
4.The bandwidth $h = n^{-1/5}$ is the best choice at the most for all models.
5. For all estimators as the sample size increases the mean squared error (MSE), the mean absolute error (MAE) and the man absolute percentage error (MAPE) decrease, for the three bandwidths (h) considered and for the four models.

Abbreviation: CLR: classical linear regression, LLR: local linear regression, LLB: local linear regression with bootstrap and LLJ: local linear regression with jackknife.

References

[1] Breidt, F.J. and Opsomer, J.D., Local polynomial regression estimators in survey sampling. Annals of Stati, (2000).
[2] Casella, G. Fienberg, S., and Olkin, I, All of Nonparametric Statistics, Springer, New York, (2006).
[3] Chambers RL, Drofman AH, Wehrly TE. Bias robust estimation in finite populations using nonparametric calibration. J Am Stat Assoc., **88**, 268–277 (1993).
[4] Drofman AH, Hall P. Estimators of the finite population distribution function using nonparametric regression. Ann Stat., **16**, 1452–1475 (1993).
[5] Drofman AH. A comparison of design-based and model-based estimators of the finite population distribution function. Aust J Stat., **35**, 29–41 (1993).
[6] Friedl, H. and Stampfer, E., "Jackknife Resampling", Encyclopedia of Environmetrics, **2**, 1089-1098 , (2002).
[7] Kim, J Y, Breidt, FJ, and Opsomer, JD. Nonparametric Regression Estimation of Finite Population Totals under Two-Stage Sampling. Technical Report, Department of Statistics, Colorado State University, (2009).
[8] Kuo, L. Classical and prediction approaches to estimating distribution functions from survey data, In Proceedings of the section on survey research methods, , Amer. Statist. Assoc., Alexandria, VA, 280-285 (1988).
[9] Rueda, M., Arcos, A. On estimating the median from survey data using multiple auxiliary information. Metrika , **54**, 59–76 (2001).
[10] Sahinler, S. and Topuz, D.,"Bootstrap and Jackknife Resampling Algorithms for Estimation of Regression Parameters", Journal of Applied Quantitative methods, **2**, 188-199 (2007).
[11] Stine, R., Bootstrap prediction intervals for regression, J. Amer. Statist. Assoc, **80**, 1026-1031 (1985).
[12] Stine, R., Modern Methods of Data Analysis; Edit: by John Fox, Scotland, 325-373 (1990)
[13] Wu, C.F.J. Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis, Annals of Statistics, **14**, 1343-1350 (1986).
[14] Ye, A., Hyndman, R, J., and Li, Z., Local linear multivariate regression with variable bandwidth in the presence of heteroscedasticity, Working Paper 8/06, Department of Econometrics and Business Statistics, Monash University.