Generalised Model Based Confidence Intervals in Two Stage Cluster Sampling

Christopher Ouma Onyango
Center for Mathematics
Strathmore University
Nairobi, Kenya
Chrisouma2004@yahoo.co.uk

Romanus Odhiambo Otieno
Department of Statistics
Jomo Kenyatta University
Nairobi, Kenya
romanusemod@yahoo.com

George Otieno Orwa
Department of Statistics
Jomo Kenyatta University
Nairobi, Kenya
orwagoti@yahoo.com

Abstract

Chambers and Dorfman (2002) constructed bootstrap confidence intervals in model based estimation for finite population totals assuming that auxiliary values are available throughout a target population and that the auxiliary values are independent. They also assumed that the cluster sizes are known throughout the target population. We now extend to two stage sampling in which the cluster sizes are known only for the sampled clusters, and we therefore predict the unobserved part of the population total. Jan and Elinor (2008) have done similar work, but unlike them, we use a general model, in which the auxiliary values X_i are not necessarily independent. We demonstrate that the asymptotic properties of our proposed estimator and its coverage rates are better than those constructed under the model assisted local polynomial regression model.

Keywords and Phrases*:* Model Based Surveys, Bootstrapping, Two Stage Sampling

AMS 2000 subject classifications. Primary 60K35; Secondary 60K35.

1. Introduction

1.1 Background

In specifying a sampling strategy in survey sampling, there exist different approaches: the design based approach, the model assisted approach, the model-based approach and randomization-assisted model based approach. For a detailed review of these approaches, see Smith (1976), Smith (1994). Our concern is the model based approach. Ouma and Wafula (2005) reviewed the work of Chambers and Dorfman (2002) and modified the conditions. However, they limited their work to simple random sampling. Suppose that P is a finite

population of N identifiable units, Y denotes a survey variable having population values Y_i, $(i = 1,2,3,...,N)$ and X to denote an auxiliary variable with corresponding population values X_i, $(i = 1,2,3,...,N)$. If the values X_i, $(i = 1,2,3,...,N)$ are all known but the characteristic values Y_i, $(i = 1,2,3,...,N)$ are known only for a Sample, say s, of $n \leq N$ of the population elements, one way of characterizing the sample selection of the survey variable is to assume that for every unit on the sampling frame, a new variable, say S_i takes a value equivalent to the number of times which that particular population unit's value is observed. The distribution of these values defines the design of the sample survey.

Once the sample has been chosen, the values $(Y_i, i \in s)$ are known. Now, let the distribution of S_i depend on the known population values of X and suppose that one wishes to use the sample values together with the known values of X to make an inference about the unknown but finite population total $T = \sum_{i=1}^{N} Y_i$ of Y. A major concern in model based approach to statistical survey inference has been finding robust estimators for the population parameters under model misspecifications.

1.2 Outline of the paper

This paper is organised as follows. In Subsections 1.3, 1.4, 1.5, 1.6, we give a brief highlight on model based estimation, the local polynomial estimation, confidence intervals, and two stage cluster sampling respectively, in each case, pointing out some gaps that our proposed estimator attempts to fill. In Section 2, we propose an estimator for the finite population total and suggest a bootstrap confidence interval for it in Section 3. In Section 4, we derive the properties of our proposed estimator. We conclude this paper in Section 5 with a simulation experiment and some discussions.

1.3 Review of model based estimation

The model based approach to statistical survey sampling has been developed to detailed extents. In particular, we build up on the work of Dorfman (1992) who proposed a non-parametric regression estimator for the population total under a model based approach. He illustrated that the developed estimator of the population total performs better when compared to the corresponding design based estimators and linear regression estimators. The model due to Dorfman (1992) relies on the assumption that the regression line passes through the origin and that the auxiliary values X_i are independent. Suppose one or all these assumptions are incorrect. Will the prediction intervals still occupy the same nominal properties? and will the estimator of the population total still be design unbiased?

1.4 Review of the local polynomial estimation

Dorfman (1992) considered a non-parametric regression model for estimating population totals in finite populations. He proposed a non-parametric regression based estimator for the population total. To develop the estimator, he assumed that the population values were generated by a model defined as

$$Y_i = m(x_i) + e_i \tag{1.1}$$

where $i = 1, 2, 3, ..., N$, $m(.)$ is a smooth function and e_i is an independent random variable with mean zero and constant variance. The non-parametric population total estimator due to Dorfman (1992) is defined as

$$\hat{T}_D = \sum_{i \in s} Y_i + \sum_{i \in s} \hat{m}(x_i) \tag{1.2}$$

where $\hat{m}(x_i) = \sum_{i \in s} w_i x_i y_i$ and $w_i x_i = k_b(x_i - x) \Big/ \sum_s k_b(x_i - x)$ is the weight associated with the i[th] unit of the sample. Further, $k(u)$ is a symmetric density function, b a scaling factor and $k_b(u) = b^{-1} k(u/b)$. In his empirical study, Dorfman (1992) illustrated that the estimator $\hat{T}_D$ performs better when compared to the corresponding design based and linear regression estimators. These results were also confirmed by Cheng (1994) who applied non-parametric regression in estimating population parameters under conditions of missing data. Breidt and Opsomer (2000), also assumed model 1.1 and developed a new class of model-assisted non parametric regression estimators for the population total, based on local polynomial smoothing, a kernel method. Their estimator is defined as

$$\hat{T}_{OB} = \sum_{i \in s} \frac{y_i - \hat{m}_i}{\pi_i} + \sum_{i \in s} \hat{m}_i \tag{1.3}$$

where

$$i = 1, 2, 3, ..., N \text{ , } \pi_i = pr(i \in s) \text{ , } \hat{m}_i = w_{si} y_s$$

and $w_{si} = diag\left[\frac{k}{h}\left(\frac{x_i - x_j}{h} \right) \frac{1}{b\pi_j} \right]$, $j \in s$ with h denoting the bandwidth. In their simulation study, $\hat{T}_{OB}$ performs better than the Horvitz-Thompson estimator defined as

$$\hat{T}_{HT} = \sum_{i \in s} \frac{Y_i}{\pi_i} \tag{1.4}$$

However, the theory developed in Breidt and Opsomer (2000) for the local polynomial regression estimator applies only to direct element sampling designs with auxiliary information available for all elements of the population. Consequently, we offer more insight on the consistency of the coverage rates using a general super population model in two stage sampling. Ji-Yeon et al. (2009) recently extended the work of Breidt and Opsomer (2000) to two stage cluster sampling where the estimators are linear combinations of estimators of

cluster totals with weights that are calibrated to known control totals. They indicated that the local polynomial regression estimators are constructed by modeling the M points (x_i, t_i) as a realization from an infinite super population model ζ in which

$$t_i = \mu(x_i) + e_i \tag{1.5}$$

where $e_i \sim N(0, \mathrm{var}(x))$, $\mu(x)$ is a smooth function of x and $\mathrm{var}(x)$ is also smooth and strictly positive. Their estimator is defined as

$$\hat{t}_y = \sum_{i \in s} \hat{\mu}_i + \sum_{i \in s} \frac{\hat{t}_i - \hat{\mu}_i}{\pi_i} \tag{1.6}$$

where

$$\hat{\mu}_i = e_i' \left(X_{si}' W_{si} X_{si} \right)^{-1} \left(X_{si}' W_{si} \hat{t}_s \right) \tag{1.7}$$

In the equation 1.7,

$$w_{si} = diag\left[\frac{k}{h_m} \left(\frac{x_i - x_j}{h_m} \right) \frac{1}{b\pi_j} \right], \; X_{si} = \left[1, (x_j - x_i), ..., (x_j - x_i)^q \right]_{j \in s}$$

and e_i represents the first column of the identity matrix X_{si}. In their simulation results they concluded that the estimator 1.6 is more efficient than the Horvitz-Thompson and the linear regression estimators when the mean function of the super population model is non linear while being nearly as efficient when the model is linear.

Recently, Jan and Elinor (2008) considered the problem of estimating the population total in two-stage cluster sampling when cluster sizes are known only for the sampled clusters, making use of a population model arising from a variance component model. They considered the application of predictive likelihood technique in estimation of the unknown part of the population total

$$T = \sum_{i=1}^{N} \sum_{j=1}^{m_i} y_{ij} \tag{1.8}$$

where N is the number of primary sampling units or clusters and each cluster consists of m_i units which are only known for the sampled clusters, y_{ij} is the value of the variable of interest for unit j of the ith cluster. They assumed the population model defined by the equations

$$E(M_i) = \beta x_i, \; \mathrm{var}(M_i) = \sigma^2 \, \mathrm{var}(x_i), \; \mathrm{cov}(M_i M_j) = 0 \tag{1.9}$$

$$E(Y_{ij}) = \mu, \; \mathrm{var}(Y_{ij}) = \tau^2, \; \mathrm{cov}(Y_{ij} Y_{ik}) = \rho\tau^2 \tag{1.10}$$

in cases where $j \neq k$ and $\rho \geq 0$. To predict the unobserved value of Z in the estimate of the population total T given by

$$\hat{T} = \sum_{i=1}^{N} \sum_{j=1}^{M_i} Y_{ij} + \hat{Z} \tag{1.11}$$

they developed a partial likelihood for Z, $L(z,y)$ from the generalized joint likelihood for the unknown quantities z and θ given by

$$L(z,y) = f_{\theta}(z,y) \tag{1.12}$$

They applied the design based Horvitz-Thompson estimator of population total,

$$\hat{T}_{HT} = \frac{x}{n_o}\sum_{i\in s}\frac{m_i\bar{y}_i}{x_i} \tag{1.13}$$

where n_0, represents the number of the primary sampling units selected in first stages and the model. In their simulation, they considered three coverage measures of Z; the model based over the joint distribution of Y and Z, the design based over the sampling design, and regarding the total sample as a stochastic variable. They concluded that for a small number and the unconditional coverage no of sampled clusters, the three intervals differ significantly, but for large n0, the three intervals are practically identical.

Further, a comprehensive simulation study of the model based and the design coverage properties of the prediction intervals indicate that for large sample sizes, the coverage measures achieve approximately the nominal level $1-\alpha$ and are slightly less than $1-\alpha$ for moderately large samples and for small sample sizes, the coverage measures are about $1-2\alpha$, being raised to $1-\alpha$ for a modified interval based on t_{n_0-2} distribution. We note that the models 1.9 and 1.10 assume that the regression line passes through the origin and that the auxiliary values X_i are considered independent. The questions raised in subsection 1.3 therefore remain unanswered.

1.5 Review of confidence intervals in survey sampling

Confidence intervals are usually constructed around point estimators in order to provide a properly scaled measure of uncertainty associated with the estimator. The conventional method is based on the assumption that the sample size is large enough for the Central Limit Theorem to hold. This is however not always true in practice.

As a consequence Do and Kokic (2001), Chambers and Dorfman (2002) applied the bootstrap method to develop model based confidence intervals to address situations where the sample sizes are not large. They also proposed modifications of the procedure to account for misspecifications in a working model. They further noted that there is greater efficiency in using of successive model refinements and estimators obtained using the bootstrap approach as opposed to their competing estimators. However, the evidence of the extended simulation study on the beef population showed that the achievement of the research did not precisely attain its goal. They therefore recommended the construction of sounder confidence intervals using the bootstrap approach. Ouma and Wafula (2005) suggested the use of a general super population model

$$Y_i = \hat{m}(x_i) + \hat{e}_i \tag{1.14}$$

where $i=1,2,3,...,N$, $m(.)$ is a smooth function, e_i is an independent random variable with mean zero and constant variance. They used a bandwidth of 1.5 and simple random sampling with replacement to generate the values of survey variable Y. In their empirical study, they established that their coverage rates were higher compared to that of Chambers and Dorfman (2002). We now extend this to two stage cluster sampling.

1.6 Review of two stage cluster sampling

Let U be a finite population of N primary sampling units psu_s or clusters labeled $1,2,..,N$, $U=1,2,...,N$ where N is a known number, M_i, $i=1,2,..,N$ be the number of secondary sampling units ssu_s in the i^{th} psu. Let y_{ij} $i=1,2,..,N$, $j=1,2,..,M_i$ be the value of the response variable Y for the ssu j belonging to the psu i. In the previous works, an assumption has been made that the element specific auxiliary data x_{ij},' $i=1,2,..,N$, $j=1,2,..,M_i$ are known for all clusters and population elements, respectively.

For our case, we assume that the cluster sizes are known only for the sampled clusters and therefore the survey values y_{ij}, $i=1,2,..,N$, $j=1,2,..,M_i$ are generated using the model

$$\hat{Y}_{ij}=m(\hat{x}_{ij})+\hat{e}_{ij} \tag{1.15}$$

with $i=1,2,..,N$ $j=1,2,..,M_i$.

2. Proposed Estimator for population total.

Jan and Elinor [6] used the model 1.15 to define the population total as

$$T=\sum_{i=1}^{N}\sum_{j=1}^{m_i}y_{ij} \tag{2.1}$$

where N is the number of primary sampling units or clusters and each cluster consists of m_i units which are only known for the sampled clusters, y_{ij} is the value of the variable of interest for unit j of the i^{th} cluster. Referring to the same model 1.15 we may write that

$$\hat{T}=\sum_{i=1}^{N}\sum_{j=1}^{m_i}y_{ij}+\sum_{i=1}^{N}\sum_{j=m+1}^{M_i}\hat{y}_{ij}=\sum_{i\in s}\sum_{j\in s_i}Y_{ij}+\hat{Z} \tag{2.2}$$

and it follows that the problem is now reduced to the that of predicting the unobserved values z of the random variable Z. To do this, we apply the general model 1.15 to predict the values of the unobserved survey variables

$y_{ij}, i=1,2,..,N$, $j=1,2,..,M_i$. Therefore the estimate of the population total is given by

$$\hat{T}=\sum_{i=1}^{N}\sum_{j=1}^{m_i} y_{ij}+\sum_{i=1}^{N}\sum_{j=m_i+1}^{M_i}\left[\hat{m}(x_{ij})+\hat{e}_{ij}\right] \tag{2.3}$$

3. Proposed Bootstrap Confidence Interval

Under the model based approach, the sampling distribution of the estimator corresponds to the distribution of possible alternative point estimates that could arise given the selection of the same sample S from populations similar to the actual underlying population of the observed data. To construct a confidence interval for T that reflects the actual finite sample and finite population characteristics of the distribution of $\hat{T}$ we estimate such a distribution from the sample data. For our case, we make use the sample data and the working model 1.15 to generate a sequence of alternative realizations of Y using non parametric estimates of $m(x_{ij})$. Let Y_{ij}^{*} be an estimator of the values of Y, where

$$Y^{*}_{ij}=\hat{m}(x_{ij})+\hat{e}_{ij} \tag{3.1}$$

In equation 3.1, e_{ij} is selected via two stage cluster sampling with replacement from $e_{ij}: i=1,2...n,\ j=1,2,..m_i$.

Having obtained the bootstrap population, the bootstrap version $\hat{T}_1{}^{*}$ of $\hat{T}_1$, using the same sample as the parent sample, is calculated. The process is then repeated a large number, B, of times to obtain $\hat{T}_{i1}{}^{*}, \hat{T}_{i2}{}^{*}, ...\hat{T}_{iB}{}^{*}$. Then the bootstrap confidence interval is obtained using

$$\left(Q^{*}\left(\alpha/2\right), Q^{*}\left(\frac{1-\alpha}{2}\right)\right)$$

where $Q^{*}(p)$ is the p^{th} – quantile of bootstrap distribution.

4. Properties of the proposed estimator and resulting confidence interval

4.1 Unbiasedness of the model

Considering the model 1.15 we may write that

$$e_{ij}=Y_{ij}-\hat{m}(x_{ij}) \tag{4.1}$$

and

$$W^{*}(x_{ij})=W_{ij}\left[\left(1-\left(\frac{m}{n-1}\right)^{\frac{1}{2}}\right)+\left(\frac{m}{n-1}\right)^{\frac{1}{2}}\left(\frac{n}{m}\right)^{\frac{1}{2}} r_i\right] \tag{4.2}$$

where m is the bootstrap sample size, r_i is the number of times the i^{th} primary sampling unit is selected, x_{ij} is the j^{th} observation made from the i^{th} cluster, and $W(x_{ij})$ is the initial sampling weight of secondary sampling unit equal to the inverse of its selection probability, that is;

$$W(x_{ij}) = \frac{1}{\pi_{ij}} \tag{4.3}$$

with $i = 1,2,..,n$; $j = 1,2,..,n_i$.

However there is considerable benefit and little loss in choosing $m = n-1$. Rao and Wu (1998)

Therefore,

$$W^*(x_{ij}) = \frac{1}{\pi_{ij}}\left[\frac{n}{n-1}\right]^{\frac{1}{2}} r_i \tag{4.4}$$

$i = 1,2,..,n$; $j = 1,2,..,m_i$, and

$$E\left(\hat{e}_{ij}\right) = E\left(Y_{ij} - \hat{m}(x_{ij})\right) \tag{4.5}$$

So

$$\hat{m}(x_{ij}) = \sum_{i,j \in s} W^*(x_{ij}) Y_{ij} \tag{4.6}$$

($i = 1,2,..,n$, $j = 1,2,..,n_i$) yielding

$$E\left(\hat{m}(x_{ij})\right) = \left[\frac{n}{n-1}\right]^{\frac{1}{2}} \sum_{i \neq j} E\left[\frac{Y_{ij}}{\pi_{ij}} r_i\right] \tag{4.7}$$

Now, let the initial sampling weight of secondary sampling units $W(x_{ij}) = \frac{1}{\pi_{ij}}$ be the kernel based weights. Then we have

$$W(x_{ij}) = \frac{K_b\left(x_{ij} - x_{ik}\right)}{\sum K_b\left(x_{ij} - x_{ik}\right)} \tag{4.8}$$

with $\sum_{ij \in s} w(x_{ij}) = 1$, further, b being a scaling factor, $K_b(u) = b^{-1}K\left(u/b\right)$ and $k(u)$ is a symmetric density function which is such that $\forall u \in \Re$ with the symbols bearing their usual meanings, then

(a) $\int_{-\infty}^{\infty} k(u)\partial u = 1$, (b) $\int_{-\infty}^{\infty} k^2(u)\partial u < \infty$, (c) $\int_{-\infty}^{\infty} |u|^3 k^2(u)\partial u < \infty$ and (d) $k(u) = k(-u)$

Therefore,

$$E(\hat{e}_{ij}) = m(x_{ij}) - E[\hat{m}(x_{ij})] \tag{4.9}$$

But as $b \to 0$ and $nb \to \infty$, $\hat{m}(x_{ij}) \to m(x_{ij})$ meaning that $E\left(\hat{e}_{ij}\right) = 0$ which is the mean of $\hat{e}_{ij}$ in model 1.15, completing the proof that the proposed model is unbiased.

4.2 Asymptotic variance of the error term

From subsection 4.1, it follows that

$$Var\left[\hat{e}_{ij}\right] = E\left[\hat{e}_{ij}\right]^2 \tag{4.10}$$

Therefore

$$Var\left[\hat{e}_{ij}\right] = E\left[\hat{e}_{ij}\right]^2 = E\left[Y_{ij} - \hat{m}(x_{ij})\right]^2 = EY_{ij}^{\ 2} - 2EY_{ij}\,\hat{m}(x_{ij}) + E\left[\hat{m}(x_{ij})\right]^2 \tag{4.11}$$

which leads to

$$E(e_{ij})^2 = \sigma^2(x_{ij}) + Var\,\hat{m}(x_{ij}) \tag{4.12}$$

But

$$Var\left[\hat{m}(x_{ij})\right] = Var\left[\left(\frac{n}{n-1}\right)^{\frac{1}{2}} \sum_{i \in s} (n-1)^{-1} b^{-1} k\left(\frac{x_{ij} - x_{ik}}{b}\right) Y_{ij}\right] \left(d_s(\hat{x}_{ij})\right)^{-1} \tag{4.13}$$

where $\left[d_s\left(\hat{x}_{ij}\right)\right]^{-1} = \dfrac{1}{\sum k_b\left(x_{ij} - x_{ik}\right)}$

$$Var\left[\hat{m}(x_{ij})\right] = \left(\frac{n}{n-1}\right) \sum_{i \in s} (n-1)^{-2} b^{-2} k\left(\frac{x_{ij} - x_{ik}}{b}\right)^2 \sigma(x_{ij}) \left(d_s(\hat{x}_{ij})\right)^{-2} \tag{4.14}$$

But

$$\left[d_s(\hat{x}_{ij})\right]^{-2} = \left[d_s(x_{ij})\right]^{-2} \left[1 - \frac{b^2}{d_s(x_{ij})} k_2 d_s^{\,//}(x_{ij}) + O\left(b^3 + (n-1)^{\frac{-1}{2}} b^{\frac{-1}{2}}\right)\right] \tag{4.15}$$

and

$$k\left(\frac{x_{ij} - x_{ik}}{b}\right)\sigma(x_{ij}) = b\sigma(x_{ij}) d_s(x_{ij}) + O(b^3 + b^{\frac{1}{2}}) \tag{4.16}$$

So using equations 4.15 and 4.16 in equation 4.14 we have that

$$var\,\hat{m}(x_{ij}) = \left(\frac{n}{n-1}\right)^2 \sum_{\substack{i \in s \\ i \neq k}} (n-1)^{-2} b^{-2} k\left(\frac{x_{ij} - x_{ik}}{b}\right)^2 \sigma(x_{ij}) \left(d_s(\hat{x}_{ij})\right)^{-2} \tag{4.17}$$

Next we obtain the asymptotic expansion of $var\hat{m}(x_{ij})$ using the following theorem as a basis.

Theorem

Let $k(u)$ be a symmetric density function with $\int uk(u)du = 0$ and $k_2 = \int u^2 k(u)du > 0$. Assume that n and N increase together such that $\frac{n}{N} \to \pi$ with $0 < \pi < 1$. If further the sampled and non sampled values of x are in the interval $[c,d]$ and are generated by densities d_s and d_{p-s} respectively, both bounded away from zero on $[c,d]$ and with continuous second derivatives, and if for any expression of Z, it can be shown explicitly that $E[Z/U] = A(U) + O(B)$ and $Var[Z/U] = O(C)$, then $Z = A(U) + O_p\left(B + C^{\frac{1}{2}}\right)$.

Using this theorem, we may write equation 4.17 as

$$var\hat{m}(x_{ij}) = \left(\frac{n}{n-1}\right)(n-1)^{-1}\left\{\sigma^2(x_{ij}) + O\left((n-1)^{\frac{-1}{2}} b^{\frac{-1}{2}}\right)\right\} \tag{4.18}$$

which reduces to

$$var\hat{m}(x_{ij}) = \frac{n}{(n-1)^2}\left\{\sigma^2(x_{ij}) + O\left((n-1)^{\frac{-1}{2}} b^{\frac{-1}{2}}\right)\right\} \tag{4.19}$$

and noting that as the number, $m_i = n$ of the second stage samples tends to be large, $n-1 \cong n$ so that we have

$$var\hat{m}(x_{ij}) = \frac{\sigma(x_{ij})}{(n-1)} + O\left((n-1)^{\frac{-3}{2}} b^{\frac{-1}{2}}\right) \tag{4.20}$$

again as $nb \to \infty$,

$$var\left[\hat{m}(x_i)\right] = \frac{\sigma^2(x_{ij})}{n-1} \tag{4.21}$$

hence

$$var\left[\hat{e}_{ij}\right] = \sigma^2(x_{ij}) + \frac{\sigma^2(x_{ij})}{n-1} \tag{4.22}$$

which as $n \to \infty$, reduces to

$$var\left[\hat{e}_{ij}\right] = \sigma^2(x_{ij}) \tag{4.23}$$

4.3 Conditional relative bias of the estimator for the population total

From its definition, the conditional relative bias of using $\hat{T}$ as an estimator of T is

$$E\left(\left[\hat{T}-T\right]/T\right)=\left[E\left(\hat{T}\right)-E(T)\right]/T \tag{4.24}$$

In this case,

$$\hat{T}=\sum_{i=1}^{N}\sum_{j=1}^{m_i}Y_{ij} \;=\sum_{i=1}^{N}\sum_{j=m+1}^{M_i}\left(\hat{m}(x_{ij})+\hat{e}_{ij}\right) \tag{4.25}$$

meaning that

$$\left[\left(E\left(\hat{T}\right)-E(T)\right)/T\right]=E\sum_i\sum_j\left(\hat{m}(x_{ij})+\hat{e}_{ij}\right)-E\sum_i\sum_j m\left(x_{ij}\right)+e_{ij} \tag{4.26}$$

Using equation 4.9 in equation 4.26, it can be seen that as $n\to\infty$ this bias, $\left[\left(E\left(\hat{T}\right)-E(T)\right)/T\right]$ asymptotically tends to zero.

5. Empirical Study

5.1 Description of the simulation experiment

Simulation experiments were performed in order to compare the performance of the model based regression estimator with that of the model assisted local polynomial regression estimator in two-stage element sampling due to Ji-Yeon et al. (2009).

To obtain the model based estimator for the population total, X_i are generated as independent and identically distributed on uniform (0, 1) random variables. The population consists of 100 clusters. In stage one a sample of $n_i=20$ clusters is taken which forms the primary sampling units from the total cluster size $N_i=100$ using simple random sampling with replacement.

In stage two, from each selected clusters, say $i,(i=1,2,...n_i)$ we select sample m_{ij}, $j=1,2,...50$, from $j=1,2,...,m_k,...1{,}000$ that is the j^{th} sample from a fixed selected i^{th} cluster using simple random sampling with replacement from total M_k=1000 elements. We consider the variable of interest Y_{ij}, $j=1,2,...,m_k,...,M_k$ which are known only for the sample and using the known auxiliary variables x_{ij}, $j=1,2,...,M_k$ we generate the non sample values using the model given in 1.14.

To simulate bootstrap values of M_k independent samples of size m_k we use simple random sampling with replacement within cluster i in order to obtain the

bootstrap population values and the model $Y_{ij}^{*} = m(\hat{x}_{ij}) + \hat{e}_{ij}$ to obtain $y^{*}_{i1}, y^{*}_{i2}, y^{*}_{i3}, ...y^{*}_{iM_k}$. Further, we let $K(u) \sim U[0,1]$ so that the model based regression estimator for the population total is given by equation 2.3 where $\hat{m}(x_{ij})$ is defined by equation 1.14 and $\hat{e}_{ij} \sim \left[0, \sigma^2\left(x_{ij}\right)\right]$.

This procedure is repeated a large number 1000 of times such that we have $\hat{T}_{i1}^{*}, \hat{T}_{i2}^{*}, \hat{T}_{i3}^{*}, \ldots, \hat{T}_{i1000}^{*}$.We then construct the 95% confidence intervals for population total $\hat{T}^{*}_{i}, i = 1,2,...N$. Similarly, we compute the local polynomial estimator for the population total suggested by Ji-Yeon et al. (2009) given in equation 1.6. For each mean function values of x_{ij}, each study variable y_{ij} $(j = 1,2.)$ for m_k values from M_k elements are generated as

$$y_{jk} = \frac{\mu_j\left(x_{ij}\right)}{M_k} + \frac{e_{jk}}{M_k^{1/2}} \tag{5.1}$$

where, y_{jk} is the k^{th} observation made from the j^{th} cluster, and

$$e_{ij} \sim N\left(0, \sigma^2\left(x_{ij}\right)\right).$$

Using above bootstrap procedure, the bootstrap estimate of the population total

$$\hat{T}^{*}_{i} = \sum_{i \in c} m\left(\hat{x}_{ij}\right) + \sum_{i \in s}\left(\frac{y_i - m_i}{\pi_i}\right) \tag{5.2}$$

is calculated, where $\pi_i = \Pr\{i \in s\}$, $\hat{m}_i = diag\left\{\frac{1}{h}k\left(\frac{x_{ij} - x_{ik}}{h}\right)\frac{1}{\pi_{ij}}\right\}, i, j \in s$ and $\hat{m}(x_{ij})$ is as defined in equation 1.4. Similarly, we construct the confidence interval for the population total T and compare performance of the developed model on estimation of T with that due to Ji-Yeon et al. (2009) on Local Polynomial regression estimation in two-stage sampling.

In computation of the model assisted local polynomial regression estimators, we again adopt a method due to Ji-Yeon et al. (2009). This we do as a means to having a realistic comparative study. We therefore apply the Epanechnikov kernel $k(u) = \frac{3}{4}\left(1 - u^2\right)\mathrm{I}_{\{|u| \le 1\}}$ and different bandwidths for computation of the Local polynomial regression estimator of population totals. This helps to compare the bias, mean squared errors and confidence interval lengths of the estimators using both the model based and model assisted local polynomial regression approaches.

5.2 Simulation Results

Table 1 gives the results of Mean Squared Error of the model based MSEmb and the Local Linear Polynomial regression estimator of the population total in two stage cluster sampling.

Table 1: Mean Squared Error (MSE)

Band width	LP	MB
0.005	0.7845631	1.2684580
0.006	0.7961103	1.7980100
0.007	0.7528211	1.4856560
0.008	0.7523094	1.4641440
0.009	0.7909287	1.3896480
0.010	0.7740691	0.1187740
0.020	0.7615400	0.0846216
0.030	0.7653660	0.0752471
0.040	0.7639690	0.0720307
0.050	0.7621990	0.0672295
0.060	0.7740690	1.1703404
0.070	0.7615400	0.7534386
0.080	0.7551610	0.7534386

It can be seen that at lower bandwidths the *MSE* for the model based estimators is higher compared to that of the Local Polynomial regression estimator. As the bandwidth increases, the *MSEmb* drastically reduces and approximately remains low. It is important to note that an increase in the bandwidth does not significantly change the *MSE* for the Local Polynomial Regression Estimators *LPRE*. Generally, the Model based estimator is more efficient than the Local polynomial estimators of the totals. Table 2 is a summary of bias for the model based estimator of the population total and the Local Polynomial regression estimator.

Table 2: Summary Results of Bias

Band width	LP	MB
0.005	-0.0443350	0.00399122
0.006	-0.0381211	0.00200561
0.007	-0.0380143	-0.0004676
0.008	-0.0384639	-0.0036316
0.009	-0.0404816	-0.0001817
0.010	-0.0985294	-0.0032099
0.020	-0.0360501	-0.0046855
0.030	-0.0396870	-0.0265196
0.040	-0.0536516	-0.0156076
0.050	-0.0552965	-0.0201331
0.060	-0.0985294	-0.0384731
0.070	-0.0360501	-0.0538332
0.080	-0.0496699	-0.0538332

The bias for the model based estimator is much lower than those of the Local Polynomial regression estimators. The large bias associated with the Local Polynomial estimators are reflected in the values of its estimators which are much lower than the true simulated population total of 99.5078. This can best be attributed to the choice of the variance. The precision of estimation can be improved by choosing a smaller value of the variance. Table 3 now presents a summary of the estimated population totals for the model based and the Local Polynomial regression estimators in two stage sampling.

Table 3: Summary Results of Estimated Population Totals

Band width	LP	MB
0.005	60.81535	99.27500
0.006	61.38900	99.01420
0.007	61.49330	99.92000
0.008	60.93400	99.20200
0.009	59.02400	99.32200
0.010	58.64900	97.47400
0.020	60.24100	93.09100
0.030	59.81800	88.08000
0.040	59.34000	83.90000
0.050	59.30700	79.47900
0.060	58.64900	59.49500
0.070	60.24000	59.82600
0.080	59.82400	59.82600

Table 4 now gives the coverage rates of 95 % confidence interval lengths for model based and Local Polynomial Regression models.

Table 4: Confidence Interval Lengths

Band width	LP	MB
0.01	38.97447	1.531668
0.02	40.74652	1.315887
0.03	39.52278	1.227330
0.04	38.87707	1.212220
0.05	38.72655	1.148093
0.06	38.93319	38.85571
0.07	38.86739	40.20200
0.08	38.54396	38.60400

The confidence intervals generated by the model based method are much tighter than those generated by the Local polynomial method at lower bandwidths but at larger bandwidths, both the LPRE and the model based estimators of population total perform poorly. We note that the best performing confidence interval is one whose coverage rate is close to the true population total and its length small. Consequently, the model based estimators are far better than their local

polynomial regression estimators. The results in general show that the model based approach outperforms the model assisted method at 95% coverage rate. The bias under model based approach is also much lower.

References

1. Breidt, F. and Opsomer, J. (2000). Local polynomial regression estimators in survey sampling. *The Annals of Statistics*, 28:1026–1053.
2. Chambers, R. and Dorfman, A. (2002). 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.
3. Cheng, P. (1994). Nonparametric estimation of mean functional with data missing at random. *Journal of the American Statistical Association*, 89: 81–87.
4. Do, K. and Kokic, P. (2001). *Bootstrap Variance and confidence interval estimation for model- based surveys*. Australia National University.
5. Dorfman, R. (1992). Nonparametric regression for estimating totals in finite population. In Section on Survey Research Methods, *Journal of the American Statistical Association*, pages 622–625.
6. Jan, F. and Elinor, Y. (2008). Two stage sampling from a predictive point of view when the cluster sizes are unknown. Biometrika, 95 (1): 187–204.
7. Ji-Yeon, K., Breidt, F., and Opsomer, J. (2009). Nonparametric regression estimation of finite population totals under two-stage cluster sampling. Technical report, Department of Statistics, Colorado State University.
8. Ouma, C. and Wafula, C. (2005). Bootstrap confidence interval for model based surveys. East African Journal of Statistics, 1: 14–18.
9. Rao, J. and Wu, C. (1998). Re sampling inference with complex survey data. *Journal of the American Statistical Association*, 83: 231–241.
10. Smith, T. (1976). The foundations of survey sampling a review. *Journal of the Royal Statistical Society, Series A*, 139: 183–198.
11. Smith, T. (1994). Sample surveys 1975, 1990. an age reconciliation. *International Review*, 62: 3–34.