Nonparametric Mixed Ratio Estimator for a Finite Population Total in Stratified Sampling

George Otieno Orwa
Department of Statistics and Actuarial Sciences
Jomo Kenyatta University
Nairobi

Romanus Odhiambo Otieno
Department of Statistics and Actuarial Sciences
Jomo Kenyatta University
Nairobi

Peter Nyamuhanga Mwita
Department of Statistics and Actuarial Sciences
Jomo Kenyatta University
Nairobi

Abstract

We propose a nonparametric regression approach to the estimation of a finite population total in model based frameworks in the case of stratified sampling. Similar work has been done, by Nadaraya and Watson (1964), Hansen et al (1983), and Breidt and Opsomer (2000). Our point of departure from these works is at selection of the sampling weights within every stratum, where we treat the individual strata as compact Abelian groups and demonstrate that the resulting proposed estimator is easier to compute. We also make use of mixed ratios but this time not in the contexts of simple random sampling or two stage cluster sampling, but in stratified sampling schemes, where a void still exists.

AMS 2000 subject classifications: 60K35.
Keywords: Sampling weights, Two-stage sampling.

1. Introduction

Model based surveys work against the background that an unknown value of a survey measurement is a realised value of a random variable, say X. A model for the random variable X is then sort and together with sampled data, inference is made regarding the population parameter of interest. For a detailed review of this strategy, see Hall and Patil (1996) and Rupert (2003).

The choice of an optimal model remains a concern being that mis-specifying a model leads to huge amounts of error. One way of solving this problem of model misspecifications is the use of nonparametric regression. For a review of the problems associated with model misspecifications, and how nonparametric regression has in the past been used in an attempt to solve the problems, see Hansen et al (1983) and Dorfman (1992).

Suppose now that the optimal design for a given survey is stratified random sampling. Breidt and Opsomer (2000) considered a general unequal probability

sampling design which covered stratification as a special case, and noted in their remarks that their technical assumptions hold under stratified simple random sampling. So making use of these results, we may suppose that a population of size N is divided into L disjoint strata, each of size $N_h,\ h = 1,2,3,...,L$.

For the j^{th} unit in the h^{th} stratum, let $y_{hj}, j = 1,2,3,...,N_h$ be the survey variable of interest. Further, let the auxiliary measurement x_{hj} be weakly but positively correlated with y_{hj} such that their independence is not rendered unrealistic. For each stratum, a sample of size n_h is taken without replacement, where n_h is sufficiently large with respect to N_h and $f_h = \frac{n_h}{N_h} \to 0$.

Let $W_h = \frac{N_h}{N}$ be the h^{th} stratum weight. The population total is accordingly given by

$$Y = \sum_{h=1}^{L} W_h y_h \tag{1}$$

which using separate ratios gives the estimator denoted and defined by

$$\hat{Y}_{SR} = \sum_{h=1}^{L} \frac{y_h}{x_h} X_h = \sum_{h=1}^{L} \frac{\overline{y}_h}{\overline{x}_h} X_h \tag{2}$$

where $y_h = \sum_{j=1}^{n_h} y_{hj},\ x_h = \sum_{j=1}^{n_h} x_{hj},\ \overline{y}_h = \frac{1}{n_h}\sum_{j=1}^{n_h} y_{hj},\ \overline{x}_h = \frac{1}{n_h}\sum_{j=1}^{n_h} x_{hj}$ and $X_h = \sum_{j=1}^{N_h} x_{hj}$.

Now consider the model

$$y_{hj} = \beta_h x_{hj} + \varepsilon_{hj} \tag{3}$$

where ε is the error, consider the following function.

It is logical that

$$E_\xi\left(\varepsilon_{hj}\right) \equiv E_\xi\left(\varepsilon_{hj} / X_{hj} = x_{hj}\right) \tag{4}$$

$$E_\xi\left(Y_{hj}\right) \equiv E_\xi\left(Y_{hj} / X_{hj} = x_{hj}\right) \tag{5}$$

$$cov_\xi\left(Y_{hj}, Y_{h'j'}\right) \equiv cov_\xi\left(Y_{hj}, Y_{h'j'} / X_{hj} = x_{hj}\right) \tag{6}$$

and $$cov_\xi\left(\varepsilon_{hj}, \varepsilon_{h'j'}\right) = cov_\xi\left(\varepsilon_{hj}, \varepsilon_{h'j'} / X_{hj} = x_{hj}\right) \tag{7}$$

where

$$E_\xi(y_{hj}) = \beta_h x_{hj} \tag{8}$$

$$\mathrm{cov}(y_{hj}, y_{h'j'}) = \begin{cases} \sigma^2 x_{hj}, & if\ \ h = h',\ and\ \ j = j' \\ 0, & otherwise \end{cases} \tag{9}$$

It easy to see that

$$\text{var}\left(\hat{Y}_{SR}\right)=\sum_{h=1}^{L}\left\{\frac{N_h(N_h-n_h)\overline{X}_h\overline{x}_{hr}}{n_h\overline{x}_h}\right\}\sigma^2 \tag{10}$$

which contains an unknown parameter σ^2 which must be estimated. On estimating the value of the unknown σ^2 using equations 4, 5, 6 and 8, we get that

$$\text{var}\left(\hat{Y}_{SR}\right)=\sum_{h=1}^{L}\left\{\frac{N_h(N_h-n_h)\overline{X}_h\overline{x}_{hr}}{n_h\overline{x}_h}\right\}\sum_{s}\left(\frac{\varepsilon^2{}_{hj}}{x_{hj}\left(1-k_{hj}\right)}\right) \tag{11}$$

where $k_{hj}=\dfrac{x_{hj}}{n_h\overline{x}_h}$.

Suppose now that the equation 1.6 changes for instance linearly to

$$E(Y_{hj}/X_{hj}=x_{hj})=\alpha+\beta_h X_{hj} \tag{12}$$

then it follows that

$$E_{\xi}\left[\hat{Y}_{SR}-Y\right]=\alpha\left[\sum_{h=1}\frac{X_h}{x_h}-L\right]\ \neq\ 0 \tag{13}$$

The implication of the equation 13 is that $\hat{Y}_{SR}$ is not robust to model misspecifications that can occur in the equation 6. Inevitably, an estimation approach that does not rely on parametric assumptions in the working model is therefore needed. Ratio estimation has been looked at by various researchers, but in different contexts. For example, see Srivastava (1990), Sukhatme and Sukhatme (1970) and even as early as the work of Olkin (1958). The concept of nonparametric regression has also been explored to yield desirable results in various other applications which an interested reader may review. For instance, see Cai and Brown (1998), Cheng (1994), Deng and Chikura (1990), Eilers and Marx (1996) and Grama and Nussbaum (1998).

1.1. Outline of the Paper

The rest of this paper is organised as follows. In Section 2, a nonparametric estimator $\hat{Y}_{PE}$ for finite population total Y is proposed. The asymptotic properties of the proposed estimator are derived in Section 3. In Section 4, we present an empirical study and give a conclusion to the paper in Section 5.

2. Proposed Estimator

We propose an estimator based on the model:

$$E\left(Y_{hj}\right)=\mu(x_{hj}) \tag{14}$$

$$cov\left(Y_{hj},Y_{h'j'}\right)=\begin{cases}\sigma^2(x_{hj}), & if\ \ h=h'\ and\ \ j=j'\\ 0, & otherwise\end{cases}$$

where $\mu(.)$ and $\sigma^2(.)$ are assumed to be twice continuously differentiable functions of x_{hj}. In this proposal, let $k_b(.)$ denote a kernel function, which is also twice continuously differentiable, and as expected, is such that $\int k_b(u)du = 1$. Further, let the smoothing weight in the h^{th} stratum be defined by

$$w_{hj}(x) = \frac{k_b\left(\frac{x_{hj} - x_{hi}}{b}\right)}{\sum_s k_b\left(\frac{x_{hj} - x_{hi}}{b}\right)} \tag{15}$$

with $j = 1,2,3,...,N_h$ and $i = 1,2,3,...,N_h$.

This weight defined by the equation 15 was first suggested by Nadaraya and Watson (1964) and later used in several other works. Their estimator of $\mu(.)$ in the equation 14 based on this weight was appropriately found to be

$$\hat{\mu}(x_{hj}) = \sum_s w_{hj}(x) y_{hj} \tag{16}$$

To estimate the population of the non-sampled units in the h^{th} stratum, it is assumed that $x = x_{hi}$ is any of the non-sampled units and therefore,

$$\hat{\mu}(x_{hj}) = \sum_s \frac{k_b\left(\frac{x_{hj} - x_{hi}}{b}\right) y_{hj}}{\sum_s k_b\left(\frac{x_{hj} - x_{hi}}{b}\right)} \tag{17}$$

Now, denote the nonparametric regression estimator for the population total by $\hat{Y}_{PE}$ and the estimator within stratum h by $\hat{Y}_{PEh}$. In stratum h, the population total is therefore

$$\hat{Y}_{Nph} = y_{hs} + \sum_{j=n_h+1}^{N_h} E_\xi\left(y_{hj}\right) \tag{18}$$

and the estimator of the population total is given by

$$\hat{Y}_{PE} = \sum_{h=1}^{L} y_{hs} + \sum_{h=1}^{L} \sum_s w_{hj}(x_i) y_{hj} \tag{19}$$

The equations 18 and 19 are similar to those used in previous works. We now suggest a different approach to the method of constructing the sampling weights defined in equation 15. Recall that our sampling scheme takes the various strata to be mutually disjointed and that within every stratum, simple random sampling is used.

Consider stratum h in isolation and let the sample size n_h from this stratum be sufficiently large. This assumption can be easily justified by our description in Section 1. Now, let the sampled elements be almost homogeneous in terms of the survey characteristics. This is usually the main goal of stratification and the elements within any stratum can usually be assumed to bear this property. Let these sampled values be such that they can be viewed as order statistics within the stratum h. This is logical being that stratification attempts to bring elements whose characteristics are very close to each other with respect to the survey problem, but this is never completely achieved. So the sampled values in the stratum can be ordered with respect to how close they are to the desired characteristics.

With this description, the sampled elements in stratum h form a compact Abelian group. Viewed as ordered values, they form a sequence whose sum is an Abelian sum, since the differences between the elements being summed has been taken to be very small.

Therefore, considering the right hand side of equation 15 let it be re-written as
The prediction error is given by

$$w_{hj}(x) = \frac{k_b\left(\frac{x_{hj} - x_{hi}}{b}\right)}{\sum_s k_b\left(\frac{x_{hj} - x_{hi}}{b}\right)} = \frac{k_b \pi_j}{\sum_{j\in s} k_b \pi_j} \tag{20}$$

with π_j being the weight of sampled unit j. Since the population size N is finite, it follows that even N_h is finite and therefore the weight in equation 20 is a countable sequence. Now as n_h has been assumed large, the Abelian sum of the denominator in equation 20 becomes

$$\sum_{j\in s} k_b \pi_j = \lim_{pr(j\in s)\to 1}\left[\sum_{j\in s}\pi_j\right]^{\sum_{j\in s}\pi_j} \tag{21}$$

which is approximated using the integral

$$\int_{j\in s}\sum_{j\in s} k_b \pi_j = \int_{j\in s}\lim_{pr(j\in s)\to 1}\left[\sum_{j\in s}\pi_j\right]^{\sum_{j\in s}\pi_j} dj \tag{22}$$

But it is known that

$$\int_x a(x)^{\sum_x a(x)}\, dx = \frac{a(x)^*}{c} \tag{23}$$

where $a(x)^*$ is a composite function of $a(x)$ containing infinitely many sub functions as may be the case when n_h is assumed large in stratum h while c is a constant which is such that c << 1. The immediate meaning of equation 23 is that

the weights described in equation 20 are purely discrete and further, they are pure functions of the auxiliary variables. It further implies that the proposed model is sufficient and also suggests unbiasedness of the model which we later prove in Section 4. The weight in the proposed estimator defined in equation 19 is therefore written as

$$w_{hj} = \frac{ck_b\pi_j}{\pi_j} \tag{24}$$

Notably from 23, the constant c << 1, meaning that this weight $w_{hj} \to k_b$ so that we now have as our proposed estimator,

$$\hat{Y}_{PE} = \sum_{h=1}^{L} y_{hs} + \sum_{h=1}^{L}\sum_{\forall s} k_b y_{hj} \tag{25}$$

But $k_b(.)$ is a kernel function assumed to be twice differentiable, and will therefore in this sampling scheme, always exist. The advantage with this setup is that we do not have to assign any weights or probabilities to the elements within the clusters, but only select a kernel and proceed to perform sampling. In fact, in the considered case where n_h is sufficiently large, $k_b \to 1$ and the estimator reduces to simply

$$\hat{Y}_{PE} = \sum_{h=1}^{L} y_{hs} + \sum_{h=1}^{L}\sum_{\forall s} y_{hj} \tag{26}$$

which is easy to compute compared to its counterpart in equation 25.

3. Properties of the Proposed Estimator

3.1 The Asymptotic Bias of the Proposed Estimator

The error in using $\hat{Y}_{PE}$ as an estimator of Y is given by

$$\hat{Y}_{PE} - Y = \sum_{h=1}^{L}\left\{\sum_{s}\frac{k_b\left(\frac{x_{hj}-x_{hi}}{b}\right)}{\sum_{s}k_b\left(\frac{x_{hj}-x_{hi}}{b}\right)}y_{hj} - y_{hr}\right\} \tag{27}$$

Since $Y_{hj} = \mu(x_{hj}) + \varepsilon_{hj}$ and given that $E_\xi\left[Y_{hj}\right] = \mu(x_{hj})$, it follows that

$$E_\xi\left[\hat{Y}_{PE} - Y\right] = \sum_{h=1}^{L}\left\{\sum_{s}\frac{k_b\left(\frac{x_{hj}-x_i}{b}\right)}{\sum_{s}k_b\left(\frac{x_{hj}-x_i}{b}\right)}\hat{\mu}(x_{hj}) - \mu(x_{hr})\right\} \tag{28}$$

Equation 28 gives the bias associated with $\hat{Y}_{PE}$. $\mu(x_{hj})$ is approximated by applying Taylor series expansion about a point x_{hj}. Further, assume that n_h is sufficiently large and that $b \to 0$, then observe that

$$\hat{\mu}(x_{hj}) \approx \mu(x_{hi}) + \mu'(x_{hi})\left(x_{hj} - x_{hi}\right) + \frac{1}{2}\mu''(x_{hi})\left(x_{hj} - x_{hi}\right)^2 \tag{29}$$

Letting $u = \dfrac{x_{hj} - x_{hi}}{b}$, then

$$\hat{\mu}(x_{hj}) \approx \mu(x_{hi}) + \mu'(x_{hi})bu + \frac{1}{2}\mu''(x_{hi})b^2u^2 \tag{30}$$

Substituting for $\hat{\mu}(x_{hj})$, equation (1.28) becomes

$$E_\xi\left[\hat{Y}_{Np} - Y\right] \approx \sum_{h=1}^{L}\left\{\begin{array}{l} \mu(x_{hi})\dfrac{k_b(u)}{\sum_s k_b(u)} + b\mu'(x_{hi})\sum_s \dfrac{uk_b(u)}{\sum_s k_b(u)} \\ + \dfrac{b^2\mu''(x_{hj})}{2}\sum_s \dfrac{u^2k_b(u)}{\sum_s k_b(u)} - \mu(x_{hr}) \end{array}\right\} \tag{31}$$

We now make use of the following theorem.

Theorem 3.1

Assume x_h's are fixed uniform design points and regularly spaced on (0, 1), then

$$\sum_{j=1}^{n_h}\left(x_{hj} - x_{hi}\right)^l k_b\left(\frac{x_{hj} - x_{hi}}{b}\right) = n_h b^{l+1}\phi_l f(x) + O\left(n_h b^{l+3}\right)$$

where $\phi_l = \int_0^1 u^l k(u)du$. Since $f(x)$ is deterministic, $f(x) = 1$ hence

$\sum_{j=1}^{n_h}\left(x_{hj} - x_{hi}\right)^l k_b\left(\dfrac{x_{hj} - x_{hi}}{b}\right) = n_h b^{l+1}\phi_l + O\left(n_h b^{l+3}\right)$ almost surely uniformly for $x \in (0,\ 1)$ and $b \in B_n$, where $B_n = \left\{C_1 n_h^{-\varepsilon_1}, C_2 n_h^{-\varepsilon_2}\right\}$, $0 < \varepsilon_1 < \varepsilon_2$, $c_1, c_2 > 0$

Furthermore

$$\sum_{j=1}^{n_h} k_b\left(\frac{x_{hj} - x_{hi}}{b}\right)\varepsilon_{hj} \to 0 \quad \text{and} \quad \sum_{j=1}^{n_h} k_b\left(\frac{x_{hj} - x_{hi}}{b}\right)\left(\frac{x_{hj} - x_{hi}}{b}\right)\varepsilon_{hj} \to 0$$

a.s. uniformly for $x \in (0,\ 1)$ and $b \in B_n$.

Using the above Theorem, equation (1.29) can be shown to be

$$E_{\xi}\left[\hat{Y}_{Np}-Y\right] = \sum_{h=1}^{L}\left\{\mu(x_{hi})+\mu'(x_h)\left[b\frac{\varphi_1}{\varphi_0}+O(b^3)\right]+\frac{1}{2}\mu''(x_h)\left[b^2\frac{\varphi_2}{\varphi_0}+O(b^4)-\mu(x_{hr})\right]\right\} \quad (32)$$

But $\phi_0 = \int_0^1 k(u)du = 1$, $\phi_1 = \int_0^1 uk(u)du = 0$ and $\phi_2 = \int_0^1 u^2k(u)du > 0$

Hence the bias of $\hat{Y}_{PE}$ is given by

$$Bias\hat{Y}_{PE} = \sum_{h=1}^{L}\left\{\begin{matrix}\mu(x_{hi})+\mu'(x_h)O(b^3)\\ +\frac{1}{2}\mu''(x_h)\left[b^2\int_0^1 u^2k(u)du+O(b^4)\right]-\mu(x_{hr})\end{matrix}\right\} \quad (33)$$

From equation 33, it is clear that the bias of

$$Y/N \approx \frac{b^2}{2}\left[\frac{1}{N}\sum_{h=1}^{L}\mu''(x_h)\int_0^1 u^2k(u)du\right]+O\left(b^4\right) \quad (34)$$

for sufficiently large n_h and with $b \to 0$.

3.2 Asymptotic Variance if the Proposed Estimator

Let the variance of $\hat{Y}_{PE}$ be denoted by $Var(\hat{Y}_{PE})$ so that,

$$Var\left(\hat{Y}_{PE}\right) = E\left(\hat{Y}_{Np}-Y\right)^2 \quad (35)$$

with $Y=\sum_{h=1}^{L}y_h$ where $y_h = \sum_{j=1}^{n_h}y_{hj} + \sum_{j=n_h+1}^{N_h}y_{hj} = \sum_{j\in s}y_{hj} + \sum_{j\notin s}y_{hj}$

So that we may now write that

$$Y = \sum_{h=1}^{L}\left(y_{hs}+y_{hr}\right)\left[\hat{Y}_{Np}-Y\right] = \sum_{h=1}^{L}\left\{\sum_s w_{hj}(x_h)y_{hj}-y_{hr}\right\} \quad (36)$$

Hence

$$Var\left(\hat{Y}_{PE}\right) = \sum_{h=1}^{L}\text{var}\left\{\sum_s w_{hj}(x_h)y_{hj}-y_{hr}\right\} = \sum_{h=1}^{L}\text{var}\left\{\sum_s w_{hj}(x_h)y_{hj}-y_{hr}\right\} \quad (37)$$

Which may be expanded to the form

$$Var\left(\hat{\bar{Y}}_{PE}\right) = \sum_{h=1}^{L}\left\{\sum_{s}\left(w_{hj}(x_h)\right)^2 \operatorname{var}\left(y_{hj}\right) + \operatorname{var}\left(y_{hr}\right) - 2\operatorname{cov}\left[\left(\sum_{j\in s} w_{hj}(x_h)\right) y_{hj}, y_{hr}\right]\right\} \tag{38}$$

which can be shown to reduce to

$$Var\left(\hat{\bar{Y}}_{PE}\right) = \sum_{h=1}^{L}\left\{\hat{W}_h\,\sigma^2(x_h) + \left(N_h - n_h\right)\sigma^2(x_h)\right\} \tag{39}$$

where $\hat{W}_h = \left(\sum_{j\in s} w_{hj}(x_h)\right)^2$ and the variance of the mean error to be

$$Var\left(\hat{\bar{Y}}_{PE}\right) = \sum_{h=1}^{L}\left\{\frac{\hat{W}_h}{N^2}\sigma^2(x_h) + \left[\frac{\left(N_h - n_h\right)}{N^2}\right]\sigma^2(x_h)\right\} \tag{40}$$

Which leads to

$$Var\left(\hat{\bar{Y}}_{PE}\right) = \frac{1}{N}\sum_{h=1}^{L}\frac{\hat{W}_h}{N}\sigma^2(x_h) + \frac{1}{N}\sum_{h=1}^{L}\frac{N_h}{N}\left(1 - f_h\right)\sigma^2(x_h) \tag{41}$$

The unknown $\sigma^2(.)$ in (40) can be estimated using its estimator, $\hat{\sigma}^2(.)$

where $\hat{\sigma}^2\left(x_h\right) = \sum_{j\in s} w_h\left(x_{hj}, x_{hi}\right)\hat{\varepsilon}_{hj}^2$ and $\hat{\varepsilon}_{hj}^2 = \left(y_{hj} - \hat{\mu}\left(x_{hj}\right)\right)^2$ and hence

$$Var\left(\hat{\bar{Y}}_{PE}\right) = \left(\sum_{h=1}^{L}\left\{\hat{W}_h\,\hat{\sigma}^2\left(x_h\right) + \left(N_h - n_h\right)\hat{\sigma}^2\left(x_h\right)\right\}\right) \tag{42}$$

In equation 41 in which n_h is sufficiently large and $N_h \to \infty$, $f_h = \frac{n_h}{N_h} \to O(n^{-\delta})$, as earlier stated in Section 3.1 and $\frac{N_h}{N} \to O(N^{-\delta}), \quad 0 < \delta < 1$.

Meaning that

$$V\operatorname{ar}\left[\frac{\hat{Y}_{PE} - Y}{N}\right] \approx \frac{1}{N}\sum_{h=1}^{L}\frac{\hat{W}_h}{N}\hat{\sigma}^2(x_h) \tag{43}$$

The asymptotic variance of $\left[\frac{\hat{Y}_{PE} - Y}{N}\right]$ is therefore

$$V\operatorname{ar}\left[\frac{\hat{Y}_{PE} - Y}{N}\right] = \frac{1}{N}\sum_{h=1}^{L}\frac{\left(\sum_{s} w_{hj}(x_h)\right)^2}{N}\hat{\sigma}^2(x_h) \tag{44}$$

Recalling that $w_{hj}(x_h) = \dfrac{k_b(u)}{\sum_{j\in s} k_b(u)}$,

$\sum_s k_b(u) = \sum_s k_b\left(\dfrac{x_{hj} - x_{hi}}{b}\right)$ = $n_h b + O(n_h b^3)$, and that

$$w_{hj}(x_h) = \frac{k_b(u)}{n_h b + O(n_h b^3)} \approx \frac{k_b(u)}{n_h b}$$

we have that

$$V\text{ar}\left[\frac{\hat{Y}_{PE} - Y}{N}\right] \approx \frac{1}{N^2}\sum_{h=1}^{L}\frac{1}{n_h b}\sum_s \frac{k_b^2(u)}{n_h b}\hat{\sigma}^2(x_h) \tag{45}$$

Consider n_h equidistant points within the h^{th} stratum such that the spacing between any two consecutive points is n_h^{-1} then $\dfrac{1}{n_h} = \left(x_{hj} - x_{hj-1}\right)$

$$V\text{ar}\left[\frac{\hat{Y}_{PE} - Y}{N}\right] \approx \frac{1}{N^2}\sum_{h=1}^{L}\frac{1}{n_h b}\sum_s \frac{k_b^2(u)}{b}\left(x_{hj} - x_{hj-1}\right)\sigma^2(x_h) \tag{46}$$

Which we may write as

$$V\text{ar}\left[\frac{\hat{Y}_{PE} - Y}{N}\right] \approx \frac{1}{N^2}\sum_{h=1}^{L}\frac{1}{n_h b}\sum_s \frac{k_b^2(u)}{b}dx_{hj}\sigma^2(x_h) \tag{47}$$

Let $\dfrac{x_{hj} - x_{hi}}{b} = u$, $bu = x_{hj} - x_{hi}$ so that $bdu = dx_{hj}$.

Using this transformation in equation 47, we may write that

$$V\text{ar}\left[\frac{\hat{Y}_{PE} - Y}{N}\right] \approx \frac{1}{N^2}\sum_{h=1}^{L}\frac{\sigma^2(x_h)}{n_h b}\sum_{j\in s} \frac{k_b^2(u)}{b}bdu \tag{48}$$

For a continuous set of points, this expression becomes

$$V\text{ar}\left[\frac{\hat{Y}_{PE} - Y}{N}\right] \approx \frac{1}{b}\frac{1}{N^2}\sum_{h=1}^{L}\frac{1}{n_h}\sigma^2(x_h)C_k \tag{49}$$

where $C_k = \int k_b^2(u)du$.

3.3 Asymptotic Mean Squared Error

The MSE of $\hat{Y}_{PE}$ is given by

$$MSE_{\xi}\left(\hat{Y}_{PE}\right) = V\,\mathrm{ar}\left(\hat{Y}_{PE}\right) + \left[Bias\left(\hat{Y}_{PE}\right)\right]^2 \tag{50}$$

From equations 34 and 49 the following the following results are immediate consequences

$$MSE_{\xi}\left(\hat{Y}_{PE}\right) \approx \frac{1}{b}\frac{1}{N^2}\left[\sum_{h=1}^{L}\frac{1}{n_h}\hat{\sigma}^2(x_h)C_k\right] + \left\{\frac{b^2}{2}\left[\frac{1}{N}\sum_{h=1}^{L}\mu''(x_h)\int_0^1 u^2k(u)du\right]\right\}^2 \tag{51}$$

$$MSE_{\xi}\left(\hat{Y}_{PE}\right) \approx \frac{1}{b}\frac{1}{N^2}\left[\sum_{h=1}^{L}\frac{1}{n_h}\sigma^2(x_h)C_k\right] + \frac{b^4}{4}\left[\frac{1}{N}\sum_{h=1}^{L}\mu''(x_h)d_k\right]^2 \tag{52}$$

where $d_k = \int_0^1 u^2k(u)du$

Now differentiating equation 52 with respect to *b* and leads to

$$\frac{\delta}{\delta b} = -\frac{1}{b^2}\left[\frac{1}{N^2}\sum_{h=1}^{L}\frac{1}{n_h}\sigma^2(x_h)C_k\right] + b^3\left[\frac{1}{N}\sum_{h=1}^{L}\mu''(x_h)d_k\right]^2 \tag{53}$$

And by equating the equation (53) to zero, the optimum *b* therefore becomes

$$b_{opt} = \left[\frac{C_1}{C_2}\right]^{\frac{1}{5}} \text{ where } C_1 = \sum_{h=1}^{L}\frac{1}{n_h}\sigma^2(x_h)C_k \text{ and } C_2 = \left[\sum_{h=1}^{L}\mu''(x_h)d_k\right]^2$$

Using this b_{opt} in the Theorem 3.1 gives

$$MSE_{\xi}\left(\hat{Y}_{PE}\right) \approx \frac{1}{N^2}\left\{\left[\sum_{h=1}^{L}\mu''(x_h)d_k\right]^{2/5}\left(\begin{array}{l}\left[\sum_{h=1}^{L}\frac{1}{n_h}\sigma^2(x_h)C_k\right]^{4/5} \\ +\frac{1}{4}\left[\sum_{h=1}^{L}\sigma^2(x_h)C_k\right]\end{array}\right)\right\} \tag{54}$$

Hence MSE for the population total under this model is

$$MSE\left(\hat{Y}_{PE}\right) \approx \left\{\left[\sum_{h=1}^{L}\mu''(x_h)d_k\right]^{2/5}\left(\begin{array}{l}\left[\sum_{h=1}^{L}\frac{1}{n_h}\sigma^2(x_h)C_k\right]^{4/5} \\ +\frac{1}{4}\left[\sum_{h=1}^{L}\sigma^2(x_h)C_k\right]\end{array}\right)\right\} \tag{55}$$

If the averages are bounded as $N \to \infty$ and $b \to 0$ then equation $49 \to 0$ This shows that our estimator is statistically consistent and therefore useful.

$$b_{opt} = \left\{ \frac{\sum_{h=1}^{L} \frac{1}{n_h} \sigma^2(x_h) C_k}{\left[\sum_{h=1}^{L} \mu''(x_h) d_k \right]^2} \right\}^{1/5} \tag{56}$$

Since b is asymptotically equivalent to $n_h^{-1/5}$, it follows that an optimal local rate of convergence for $\hat{Y}_{PE}$ is $O\left(n_h^{-1/5}\right)$ as expected. See Bashtannyk and Hyndman (2001), Wu (2003) and Wu and Luan (2003) for examples. In fact, this conclusion implies that the model is actually unbiased, since past works bearing this property, already demonstrated unbiasedness. See Chambers et al. (1992) and Chambers and Dorfman (1992) and Silverman (1986).

4. Empirical Study

4.1. Description of the Population

For our experiment, we simulated a population, $X \sim N\left(0, \sigma^2\left(x_{hj}\right)\right)$.

We stratified this population, first considering the distance of each value from the mean, and then by considering the distance of each value from the median to give two separate populations. In these populations, we used the mixed ratio estimator due to Nadaraya and Watson (1964), which we denote by $\hat{Y}_{SR}$, and the proposed estimator $\hat{Y}_{PE}$ to estimate the population total. We did this for several mean functions so as to make enough comparison. A sample of size 100 was taken with each stratum contributing a sample size proportional to the number of units in it. Simple random sampling is done 250 times for each case. Epanechnikov kernel, defined by, was used for the kernel smoothing on the various populations. We computed the biases as $\left(\hat{Y}_{SR} - Y\right)$ and $\left(\hat{Y}_{NP} - Y\right)$, respective average of the variances $\left(\hat{Y}_{SR} - Y\right)$ and $\left(\hat{Y}_{NP} - Y\right)$, and the Mean Squared Errors. We also computed the 95% confidence interval for each of the populations.

4.2. Results

From Table 1, the $\hat{Y}_{SR}$ gives a better estimation of the population total, while $\hat{Y}_{PE}$ underestimates the population total. However, the Root Mean Square Error *RMSE* of $\hat{Y}_{PE}$ is superior. The coverage ability of $\hat{Y}_{PE}$ is also better than that of $\hat{Y}_{SR}$.

Table 1: MSE from the Random Samples

	$STE < -1.96$	$STE > 1.96$	$\lvert STE \rvert > 1.96$	*Bias*	*RMSE*
$\hat{Y}_{SR}$	16.5	21.5	62.0	1,098	17,170
$\hat{Y}_{PE}$	22.0	9.0	69.0	−21,511	15,578
$\hat{Y}_{SR}$	22.5	14.5	63.5	−6,782	16,111
$\hat{Y}_{PE}$	23.0	8.5	68.5	−10,745	15,234

Table 2 presents the conditional relative biases due to the two estimators for the various chosen mean functions. For each of the mean functions, various values of the bandwidths were tried.

Table 2: Conditional Relative Biases from the Two Populations

	*pop.*1	*pop.*1	*pop.*2	*pop.*2
mean functions	$\hat{Y}_{PE}$	$\hat{Y}_{SR}$	$\hat{Y}_{PE}$	$\hat{Y}_{SR}$
linear				
$h = 0.1$	0.002102	0.052537	0.026044	0.244286
$h = 0.25$	0.034306	7.081053	0.048733	0.238996
$h = 1$	0.021130	7.081123	0.035501	0.238999
$h = 2$	0.015000	7.081232	0.027226	0.225987
quadratic				
$h = 0.1$	0.034904	0.066913	0.052033	3.982488
$h = 0.25$	0.052177	1.317905	0.074353	3.989672
$h = 1$	0.022990	1.319005	0.086174	3.997947
$h = 2$	0.011171	1.364600	0.040354	3.998730
exp *onential*				
$h = 0.1$	0.350423	0.370265	0.408932	7.389913
$h = 0.25$	0.211236	28.32795	0.705700	7.713993
$h = 1$	2.135405	30.24339	2.529527	9.450606
$h = 2$	1.216383	29.43096	1.710575	8.787640
cycle				
$h = 0.1$	0.014433	0.012207	0.014464	0.432200
$h = 0.25$	0.034446	0.545414	0.100755	0.576848
$h = 1$	0.035104	0.315127	0.120530	0.607393
$h = 2$	0.037353	0.627709	0.103034	0.683317

The biases due to $\hat{Y}_{SR}$ are remarkably larger compared to those of $\hat{Y}_{PE}$ for the first population. Though the biases due to the separate ratio estimator are

positive, those of the proposed estimator are generally smaller; a manifestation that $\hat{Y}_{SR}$ tends to overestimate the population total while $\hat{Y}_{PE}$ may underestimate the population total but only in a case where n_h is not large enough. For the second population, $\hat{Y}_{PE}$ performs better than $\hat{Y}_{SR}$, but again it is worth noting that in a case where the sample size is small, there is a risk of $\hat{Y}_{PE}$ underestimating the population total.

5. Conclusion

Use of $\hat{Y}_{PE}$ has in general led a relatively smaller error compared to the usual separate ratio estimator. We can therefore conclude that nonparametric regression approach in stratified sampling using the modified kernel smoothing yields very good results.

References

1. Bashtannyk, D. and Hyndman, R. (2001). Bandwidth selection for kernel conditional density estimation. *Comput. Statist. Data Anal*., 36:279-298.
2. Breidt, F.J. and Opsomer, J.D. (2000). Local polynomial regression estimators in survey sampling. *The Annals of Statistics,* 28:1056-1053.
3. Cai, T.T. and Brown, L.D. (1998). Wavelet shrinkage for nonequispaced samples. *The Annals of Statistics,* 26:1783-1799.
4. Chambers, R.L. and Dorfman, A.H. (1992). Bias robust estimation in finite population using nonparametric calibration. *Journal of the American Statistical Association,* 88:268-277.
5. Chambers, R.L., Dorfman, A.H. and Hall, P. (1992). Properties of estimators of the finite population distribution function. *Biometrika,* 79(3):577-582.
6. Cheng, P. (1994). Nonparametric estimation of mean functional with data missing at random. *Journal of the American Statistical Association,* 89: 81-87.
7. Deng, L. and Chikura, R.S. (1990). On the ratio and regression estimation in finite population sampling. *Journal of the American Statistical Association,* 44:282-284.
8. Dorfman, R. L. (1992). Nonparametric regression for estimating totals in finite population. In *Section on Survey Research Methods, Journal of American Statistical Association*, pages 622 -625.
9. Eilers, P. and Marx, B. (1996). Flexible smoothing with b-splines and penalties (with discussion). *Statistical Science,* 11:89-121.

10. Grama, I. and Nussbaum, M. (1998). Asymptotic equivalence for nonparametric generalized linear models. *Probab. Theory Related Fields,* 111:167-214.

11. Hall, P. and Patil, P. (1996). On the choice of smoothing parameter, threshold and truncation in nonparametric regression by wavelet methods. *Journal of the Royal Statistical Society, Series B,* 58:361-377.

12. Hansen, M. H., Madow, W. G. and Tepping, B. J (1983). An evaluation of model- dependent and probability sampling inferences in sample surveys. *Journal of American Statistical Association*, 73:776-793.

13. Nadaraya, E. A. and Watson, J. (1964). On estimating regression, *Theory Probability Application,* 10:186-190.

14. Olkin, I. (1958). Multivariate ratio estimation for finite populations. *Biometrika,* 45:154-165.

15. Ruppert, D. (2003). *Semiparametric Regression*. Cambridge, UK: Cambridge University Press.

16. Silverman, B. (1986). *Density estimation for statistics and Data analysis: Monographs on Statistics and Applied Probability.* Chapman and Hall, London.

17. Srivastava, S.K. (1990). An estimation of the mean of the finite population using several auxiliary variables. *Journal of Indian Statistical Association,* 3:201-223.

18. Sukhatme, P. and Sukhatme, B. (1970). *Sampling Theories of Survey with Applications*. Lowa State University Press, Ames.

19. Wu, C. (2003). Optimal calibration estimators in survey sampling. *Biometrika,* 90:927-951.

20. Wu, C. and Luan, Y. (2003). Optimal calibration estimators under two phase sampling. *Journal of Official Statistics,* 19:119-131.