Adaptive Nonparametric Variance Estimation For A Ratio Estimator

Romanus Odhiambo Otieno

Department of Mathematics and Computer Science, Jomo Kenyatta University of Agriculture and Technology, P.O. Box 62000, Nairobi, KENYA

ABSTRACT

Kernel estimators for smooth curves require modifications when estimating near end points of the support, both for practical and asymptotic reasons. The construction of such boundary kernels as solutions of variational problem is a difficult exercise.

For estimating the error variance of a ratio estimator, we suggest an alternative estimation procedure using the theory of local linear regression. The proposed estimator adapts robustly to both interior and boundary points. We also derive the asymptotic mean square error of the new estimator and conditions under which it is efficient.

1.0 INTRODUCTION

Consider a finite population of N identifiable units: $U = (U_1, \ldots\ldots, U_N)$. Suppose that to each of these units there exists two numbers (x_i, y_i) which are positively correlated and are such that $(x_i, y_i) > (0,0) \; \forall \; i \in U$, where x_i's is known $\forall \; (i \in U)$, but y_i is known only if $i \in s$, s is a subset of U chosen using a probability selection plan, P, which assigns a probability P(s), to a given s such that $P(s) \geq 0$, $\sum_{s=S} P(s) = 1$, $S = \cup\{s\}$. Given s, we can compute a statistic $\hat{T}(y)$ based on the observed y_i's $(i \in s)$ and all the prior values x_i's, where $y = (y_1, \ldots\ldots, y_N)$. Let T(y) be the finite population function (i.e. census value) of interest. The problem considered here is that of estimating the variance of $\hat{T}(y) - T(y)$.

2.0 A MODEL BASED APPROACH

A standard approach to estimating T(y) assumes that the values of y can be looked upon as realisation of some unknown random variables $Y = (Y_i, \ldots\ldots, Y_N)$ whose conditional distributions can be specified, with $X = (X_i, \ldots\ldots, X_N)$ being a conditioning parameter. This distribution is generally described by a probability model, ξ .

A commonly used model in survey sampling is (Cochran 1977):

$$E(Y_i | X_i = x_i) = \beta x_i$$

$$\text{var}(Y) = \sigma^2 x_i$$

$$\text{cov}(Y_i, Y_j) = 0 \quad \text{if } i \neq j \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

where β and σ^2 are unknown positive constants.

A Best Linear Unbiased Estimator (BLUE) for the population total $T(y) = T = \sum_{i=1}^{N} y_i$, under model (1) is the ratio estimator:

$$\hat{T}_R = \frac{\bar{y}_s}{\bar{x}_s} \sum_{i=1}^{N} x_i = \hat{R} \sum_{i=1}^{N} x_i \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

where $\hat{R} \frac{\bar{y}_s}{\bar{x}_s}$, $\bar{y}_s$, $\bar{x}_s$ are sample means of x_i 's and y_i 's respectively.

Given s, and all x_i's $\hat{T}_R$, can be computed. Once $\hat{T}_R$ has been computed the next and more formidable step is the assessment of the accuracy of $\hat{T}_R$ (i.e. the precision of $\hat{T}_R$) as an estimator of T. Under a model based approach, a popular measure of the accuracy of $\hat{T}_R$ is the variance of the prediction error.

$E = \hat{T}_R - T$. Under (1), this error variance is given as

$$\text{Var}_\xi (\hat{T}_R - T) = \left(\frac{(N - n)\bar{x}_r}{n\bar{x}_s} \right)^2 \sum_{i \in S} \sigma^2 x_i + \sum_{i \in r} \sigma^2 x_i \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

where r is the complement of s, $\bar{x}_s$, $\bar{x}_r$ are the population and non-sample means of x_i's respectively. An optimal estimator of (3) from standard weighted Least squares theory is

$$V_L\left(\frac{N(N-n)\bar{x}}{n(n-1)\bar{x}_s}\right)^2\sum_{i\in S}\frac{\hat{e}_i^2}{x_i}$$

where $\hat{e}_i = y_i - \hat{R}x_i$.

The estimator V_L is optimal only if (1) is true, which is unlikely to hold in practice. This immediately raises the question of whether it is possible to obtain alternative estimators of (3) which are robust to miss specification of the variance model (i.e. $\text{var}(Y_i|X_i) - \sigma^2 x_i$).

3.0 Robust Variance Estimation via Local Linear Regression

If one adopts design based approach to survey sampling, this question of robustness is easily answered, just replace V_L by a design unbiased estimator of (3). For example one could use the classical variance estimator (Cochran, 1977);

$$V_e = \frac{N(N-n)}{n}\sum_{j\in S}\frac{\hat{e}_j^2}{n-1}$$

This estimator is unbiased under repeated sampling from finite population sampling, irrespective of the link between the survey variables Y and the bench mark variables.

From a model based perspective, design unbiasedness lacks appeal. This is a property that holds over repeated sampling. The survey statistician has only one set of sample data. The worry is how to protect against incorrect inference given these data.

A natural alternative is to adopt a non-parametric model based approach. That is we replace the parametric working model (1) by a non parametric model linking $\text{var}(Y_i|X_i = x_i)$ and X_i, and then use an estimator that performs reasonably with respect to this expanded model.

Let

$$E(Y_i|X_i = x_i) = \beta x_i$$

$$\text{var}(Y_i|X_i = x_i) = \sigma^2(x_i)$$

$$\text{cov}(Y_i, Y_j) = 0 \qquad \forall i \neq j \qquad \text{.....(4)}$$

where $\sigma^2(x_i)$ is differentiable up to second order. Under this model, the error variance of $\hat{T}_R$ is

$$Var_{\xi}\left(\hat{T}_R - T\right) = \left(\frac{(N-n)\bar{x}_r}{n\bar{x}_s}\right)^2 \sum_{i\in S}\sigma^2(x_i) + \sum_{i\in r}\sigma^2 x_i \qquad \text{.....(5)}$$

Noting that

$$E_{\xi}\left(e_i^2 \middle| X_i = x_i\right) = \sigma^2\left(x_i\right) + O(n^{-1}) \qquad \text{.....(6)}$$

it follows that we can take $\hat{e}_i^2$ as a naive estimator of $\sigma^2(x_i)$, and then seek an estimation procedure that achieves a more appealing estimator, based on this initial estimator.

In a small neighbourhood of x_i, it can be shown that

$$\sigma^2(x_i) \approx \sigma^2(x_i) + \sigma^{2'}(x_i)(x_j - x_i) \qquad \text{.....(7)}$$

where $\sigma^{2'}(x_i)$ is the first derivative of $\sigma^2(x_i)$. For $j \in s$, and from (6), it follows that (7) can be approximated, (if n is large) by

$$e_j^2 \approx \sigma^2(x_i) + \sigma^{2'}(x_i)(x_j - x_i) \equiv a + b(x_j - x_i).$$

Thus the problem of estimating $\sigma^2(x_i)$, and hence (by extension) (5) is equivalent to a local linear regression problem: estimating the intercept a. Now consider a weighted (local) linear regression; finding a and b to minimize:

$$\sum_{i=1}^{n}\left(\hat{e}_j^2 - a - b(x_j - x_i)\right)^2 k\left(\frac{x_i - x_j}{h}\right) \qquad \text{.....(8)}$$

where k(.) is a kernel function of the parzen type. Let $\hat{a}$ and b be the solution to the weighted Least squares problem (8). Simple calculation yields

$$\hat{a} = \frac{\sum_{j\in S} w_j \hat{e}_j^2}{\sum_{j\in S} w_j}$$

where w_j is defined by (10). Thus we define the local linear regression smoother by

$$\hat{\sigma}^2(x_i) = \frac{\sum_{j\in S} w_j \hat{e}_j^2}{\sum_{j\in S} w_j} \qquad \text{.....(9)}$$

with $w_j \equiv k\left(\frac{x - x_j}{n}\right)\left[S_{n,2} - \left(x - x_j\right)S_{n,2}\right]$ .....(10)

where $S_{n,L} = \sum_{j=1}^{n} k\left(\frac{x - x_j}{h}\right)\left(x - x_j\right)^L$, $L = 1,2,\ldots\ldots$ .....(11)

and h is the bandwidth parameter that controls the amount of smoothing to be done. This idea is an extension of Stone (1977), who used a kernel function $k(x) = \frac{1}{2}\left[|x| \le 1\right]$. It follows that $\hat{\sigma}^2(x_i)$ is a weighted average of the squared residuals and is called a linear smoother in curve estimation. Also by intuition it is clear that $\sigma^2(x_i)$ is estimated by $\hat{b}$, defined by

$$\hat{b} = \frac{\sum_{j=S} k\left(\frac{x_i - x_j}{h}\right)\hat{e}_j^2\left(x_i - x_j\right) - \hat{a}\sum_{j\in S} k\left(\frac{x_i - x_j}{h}\right)}{\sum_{j\in S} k\left(\frac{x_i - x_j}{h}\right)\left(\frac{x_i - x_j}{h}\right)^2}$$

substitution of (9) in (5) yields a robust estimator of (3) as

$$v_{LR} = \left[\frac{(N-n)\bar{x}_r}{n\bar{x}_s}\right]^2 \sum_{i\in S}\hat{\sigma}^2(x_i) + \sum_{i\in r}\hat{\sigma}^2(x_i)$$

4.0 Asymptotic Properties of the New Variance Estimator

Assume that the finite population under consideration comes from a sequence of populations $\{P_k\}_{k=1}^{\infty}$ each of size N_k with $N_k \ge N_{k-1}$. Let $\{S_k\}_{k=1}^{\infty}$ be a sequence of corresponding samples, each of size n_k, with $n_k \ge n_{k-1}$. Let as $k \to \infty$, $f_k = \frac{n_k}{N_k} \to 0$, n_k, $N_k \to \infty$ Suppose that as these developments take place, the sample and the population averages converge to non zero constants. Supposing further that

(i.) $\sigma^2(x_i)$ has a bounded and continuous second derivative;

(ii.) The kernel function k(.) is a bounded density function with

$$\int xk(x)dx = 0 \text{ and } \int x^2k(x)dx < \infty$$

Let, in the sequel,

$$c_k = \int_{-\infty}^{\infty} u^2k(u)du, \qquad d_k = \int_{-\infty}^{\infty} k^2(u)du$$

If the above conditions apply and if in addition, the x_i's are regularly spaced in [0,1], then the relationship of the asymptotic mean square error, bias, variance of V_{LR} and the bandwidth can be specified by the following theorem:

Theorem:

If $h \to \infty$, and $nh \to \infty$, then for $x \in (a_o, b_o)$ the estimator V_{LR} has the conditional relative mean square error given by

$$E_\xi\left\{\left(\frac{V_{LR}}{Var_\xi(\hat{T}_R - T)} - 1\right)^2 \middle| X_1 = x_1, X_2 = x_2, \ldots\ldots\ldots, X_n = x_n\right\} \approx$$

$$\frac{h^4}{4}\left(\frac{c_k \sum_{i\in S} \sigma^{2''}(x_i)}{\sum_{i\in S} \sigma^2(x_i)}\right)^2 + \frac{1}{nh}\left[\sum_{i\in S}\sum_{k\in S}\left(\mu_4(x_i) - \mu_2^2(x_i)\right)\int k(u)k\left(\frac{x - x_i}{h} - u\right)du\right]$$

where $\sigma^{2''}(.)$ is the second derivative $\sigma^2(.)$

A sketch of the proof of the theorem follows. Observe that for large n,

$$w_j = k\left(\frac{x - x_j}{h}\right)\left[S_{n,2} - (x - x_j)S_{n,1}\right] \approx k\left(\frac{x - x_j}{h}\right)$$

Hence

$$E_\xi\left[\hat{\sigma}^2(x_i)\right] \approx \frac{\sum_{j\in S} k\left(\frac{x_i - x_j}{h}\right) E\left(\hat{e}_j^2 \middle| X_j = x_j\right)}{\sum_{j\in S} k\left(\frac{x - x_j}{h}\right)}$$

$$\approx \frac{\frac{1}{nh}\sum_{j\in S} k\left(\frac{x_i - x_j}{h}\right) E\left(\hat{e}_j^2 \middle| X_j = x_j\right)}{\frac{1}{nh}\sum_{j\in S} k\left(\frac{x - x_j}{h}\right)}$$

$$\approx \frac{\frac{1}{nh}\sum_{j\in S} k\left(\frac{x - x_j}{h}\right) \sigma^2(x_j)}{\frac{1}{nh}\sum_{j\in S} k\left(\frac{x - x_j}{h}\right)}$$

$$\approx \frac{\frac{1}{nh}\sum_{j\in S} k\left(\frac{x-x_j}{h}\right)\left[\sigma^2(x_j)+hu\sigma^{2'}(x_i)+\frac{h^2u^2}{2!}\sigma^{2''}(x_i)\right]}{\frac{1}{nh}\sum_{j\in S} k\left(\frac{x-x_j}{h}\right)}$$

$$\approx \sigma^2(x_i)+\frac{h'\sigma^{2''}(x_i)}{2}\int u^2 k(u)du$$

$$\approx \sigma^2(x_i)+\frac{h^2\sigma^{2''}(x_i)c_k}{2}$$

Hence the asymptotic bias of V_{LR} under the expanded model is

$$\approx \left[\frac{(N-n)\bar{x}_r}{n\bar{x}_s}\right]^2 \sum_{i\in S}\frac{h^2\sigma^{2''}(x_i)c_k}{2}$$

From this it follows that the asymptotic relative bias of V_{LR} is

$$\approx \frac{\left[\frac{(N-n)\bar{x}_r}{n\bar{x}_s}\right]^4\left[\sum_{i\in S}\sigma^{2''}(x_i)c_k\right]^2\frac{h^4}{4}}{\left[\frac{(N-n)\bar{x}_r}{n\bar{x}_s}\right]^4\left[\sum_{i\in S}\sigma^2(x_i)\right]^2}$$

$$\approx \frac{h^4}{4}\left[\frac{c_k\sum_{i\in S}\sigma^{2''}(x_i)}{\sum_{i\in S}\sigma^2(x_i)}\right]^2$$

Next we derive the asymptotic variance of V_{LR}

Now

$$\text{var}\left[\left(\frac{(N-n)\bar{x}_r}{n\bar{x}_s}\right)^2\sum_{i\in S}\hat{\sigma}^2(x_i)+\sum_{i\in r}\hat{\sigma}^2(x_i)\right]=\left(\frac{(N-n)\bar{x}_r}{n\bar{x}_s}\right)^4\sum_{i\in S}\sum_{k\in r}\text{cov}\left(\hat{\sigma}^2(x_i),\hat{\sigma}^2(x_k)\right)$$

$$+\sum_{i\in S}\sum_{k\in r}\text{cov}\left(\hat{\sigma}^2(x_i),\hat{\sigma}^2(x_k)\right)$$

$$+2\left(\frac{(N-n)\bar{x}_r}{n\bar{x}_s}\right)^2\sum_{i\in S}\sum_{k\in r}\text{cov}\left(\hat{\sigma}^2(x_i),\hat{\sigma}^2(x_k)\right)$$

where

$$\text{cov}\left(\hat{\sigma}^2(x_i),\hat{\sigma}^2(x_k)\right)=\sum_{j\in S}\sum_{L\in r}k\left(\frac{x_i-x_j}{h}\right)\left(\frac{x_k-x_L}{h}\right)\text{cov}\left(r_j^2,r_L^2\right)$$

After some algebra, it can be shown that

$$\operatorname{cov}\left(r_j^2, r_L^2\right) = \begin{cases} \operatorname{Var}(y^2)(1+O(1)) + O(n^{-1}), & j = L \\ O(n^{-1}) \end{cases}$$

$$\approx \begin{cases} \left[\mu_4(x_i) - \mu_2^2(x_i)\right](1+O(1), & i = L \\ 0, \text{Otherwise} \end{cases}$$

where

$$\mu_4(x_i) = E_\xi\left(Y_i^4 \middle| X_i = x_i\right),$$
$$\mu_2(x_i) = E_\xi\left(Y_i^2 \middle| X_i = x_i\right)$$

Thus

$$\operatorname{var}\left[\frac{V_{LR}}{\operatorname{var}_\xi\left(\hat{T}_R - T\right)}\right] \approx \frac{\left[\sum_{i\in S}\sum_{k\in S}\left[\mu_4(x_i) - \mu_2^2(x_i)\right](1+O(1))\int k(u)k\left(\frac{x_i - x_k}{h} - u\right)du\right]}{\left[\frac{(N-n)\bar{x}_r}{n\bar{x}_s}\right]^4\left[\sum_{i\in S}\sigma^2(x_i)\right]^2}$$

$$\approx \frac{\sum_{i\in S}\sum_{k\in S}\left[\mu_4(x_i) - \mu_2^2(x_i)\right](1+O(1))\int k(u)k\left(\frac{x_i - x_j}{h} - u\right)du}{\left[\sum_{i\in S}\sigma^2(x_i)\right]^2}$$

Hence the asymptotic relative mean square error of V_{LR} is

$$\approx \frac{h^4}{4}\left[\frac{c_k\sum_{i\in S}\sigma^{2''}(x_i)}{\sum_{i\in S}\sigma^2(x_i)}\right]^2 + \frac{1}{nh}\left[\frac{\sum_{i\in S}\sum_{k\in S}\left[\mu_4(x_i) - \mu_2^2(x_i)\right](1+O(1))\int k(u)k\left(\frac{x_i - x_k}{h} - u\right)du}{\left[\sum_{i\in S}\sigma^2(x_i)\right]^2}\right]$$

Remark 1: Unlike the usual kernel regression (Nadaraya (1964), Watson (1964)) estimators, the local linear variance estimation procedure proposed here is not susceptible to boundary effects. It thus has a wider scope and is theoretically more appealing than the kernel procedures.

Remark 2: The above theorem illustrates the following points:

V_{LR} is bias robust if $h \to 0$ as $n \to \infty$, V_{LR} is efficient if $nh \to \infty$ as $n \to \infty$.

5.0 CONCLUSION

The conclusion is that local linear estimation can give estimators of the variance that adapts robustly to the position of the point in the support. The resulting variance estimator has good bias robustness properties that are generally lacking in V_L and V_C for a given sample. Unlike the usual kernel techniques, the estimation procedure suggested here does not require modifications when estimating near end points of the support.

ACKNOWLEDGEMENTS

I am grateful to Ringa Kaingu in the department of Mathematics and Computer Science, JKUAT for ably typing this work. My thanks also go to the JAST Editorial Board, and the anonymous referees whose comments greatly improved the quality of the first version of this paper.

REFERENCES

Cochran W.G. (1977) *Sampling techniques*, Wiley and sons.

Nadarasa E.A. (1964) *On estimating regression*. Theory of probability applications, **9** 141 - 142.

Stone, C. J. (1977) *Consistent Nonparametric Regression*, The Annals of Statistics, **5**, 595 - 620.

Watson G. S. (1964) *Smooth Regression Analysis*, Sakhya ser.A, **26**, 359 - 372.