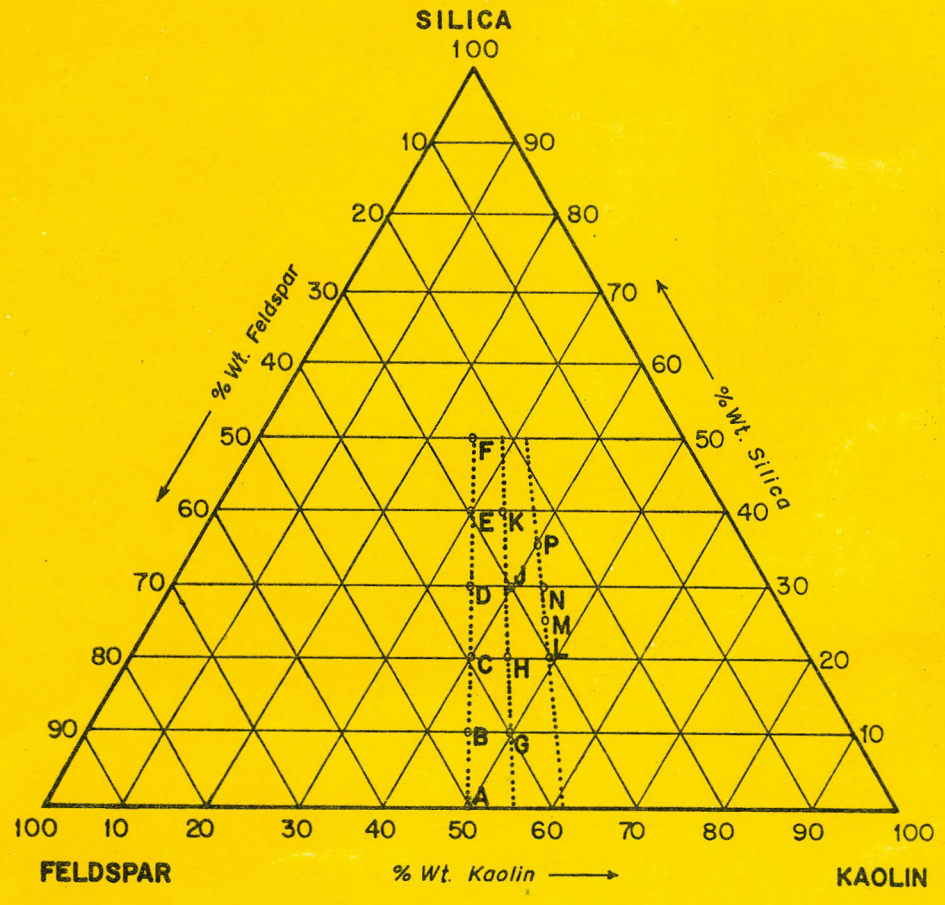
ISSN 1029–3221

EAST AFRICAN JOURNAL OF SCIENCE

Volume 2 — Part 2 December 2000



An International Journal of Pure and Applied Science

East African Journal of Science 2(2): 107–112 (2000)

Nonparametric regression method for estimating the error variance in unistage sampling

Romanus Odhiambo Otieno[1]* & Tobias Mbithi Mwalili[2]
[1]Department of Mathematics & Statistics, Jomo Kenyatta University of Agriculture and Technology, P.O. Box 62000, Nairobi, Kenya. *E-mail*: romanus@jkuat.ac.ke
[2]Department of Mathematics, Kenyatta University, P.O. Box 43844, Nairobi, Kenya.
*Corresponding author.

Nonparametric regression provides computationally intensive estimation of unknown finite population quantities. Such estimation can be more robust than inference tied to model based inference. A nonparametric procedure for estimating error variance is suggested. An empirical example is given to illustrate the performance of the derived estimator vis-a-vis the currently popular variance estimator in model based surveys.
Key words: Auxiliary variable; bandwidth; kernel estimator; ratio estimator; non-parametric regression; variance estimation.

INTRODUCTION

In this paper, we consider the application of nonparametric regression to the estimation of finite population variance based on a sample from the population. Given a population p of N identifiable units for each of which there is a variable Y of interest, with values available on a sample s of p, we wish to estimate the population total $T = \sum_P Y_i$. Parallel work on distribution may be found in Chambers *et al.* (1993) and Kuk(1993). Smith and Njenga (1992) used the nonparametric regression in the estimation of superpopulation parameters.

We assume that an auxillary variable X related to Y is available for the entire population. If this is the case, then there are at least two main rival approaches for estimating T.

If we assume that a particular model relating the variables holds, then an appropriate estimator can be based on this model. For example, under the simple linear regression model

$$Y_i = \alpha + \beta x_i + \sigma(x_i)e_i \qquad i = 1,.....N$$

with α and β unknown, $\sigma(x)$ known, and e_i identically and independently distributed with mean 0 and unknown variance, the best linear unbiased estimator of T is

$$\hat{T}_{lin} = \sum_s Y_i + \sum_{p-s} \left(\hat{\alpha} + \hat{\beta} x_i\right)$$

where $\hat{\alpha}, \hat{\beta}$ are appropriate weighted least squares estimators of α, β. Clearly the parameters α, β are essentially nuisance parameters since T is of interest. Many sample survey practitioners are uncomfortable with this approach because of uncertainties in the choice of the model i.e. the robustness problem.

The alternative is to insist that the sample be selected according to a probability design, and to assure robustness within the sampling framework by incorporating inclusion probabilities into the estimator. For example, we can stratify on the auxiliary, employ stratified equal probability random sampling without replacement and use the expansion estimator.

Naturally there are many other possibilities for design and model-based estimators, including design-based estimators that incorporate the model e.g. the combined regression estimator (Cochran, 1977). Hansen *et al.* (1983) and Royall and Cumberland (1981) give further discussion on design based versus model based approaches.

An estimator motivated by nonparametric regression, stated below, is model based in which, however, the assumptions concerning the relation of Y and X are considerably weakened. The

following section gives a review of nonparametric regression, states a nonparametric regression based estimator of the total, gives expressions for its bias and variance, and suggests how its error variance is estimated. The third section gives empirical performances of the estimators of the error variance.

AN ESTIMATOR OF TOTAL

Nonparametric regression

The idea of nonparametric regression goes back to Nadaraya (1964) and Watson (1964). A recent reference is Hardle (1991). There exist many types of nonparametric regression; here only the simple Nadaraya–Watson Kernel estimator is considered. Consider the model

$$Y = m(x) + \sigma(x)e \quad \ldots\ldots\ldots\ldots\ldots \quad (1)$$

where m (.) is a smooth function and the e_i independently distributed with mean 0 and constant variance.

Suppose we wish to estimate m(x). One possibility is to average the nearby values of Y, where "nearby" is measured by the distance $|x_i - x|$. Let k(u) be a symmetric density function, for example the standard normal function. For a chosen scaling factor ("bandwidth") b, define $k_b(u) = b^{-1}k(u/b)$, and weights $w_i(x) = k_b(X_i - x) \div \sum_{i=1}^{n} k_b(X_i - X)$. The larger b is, the more equal the weights. The Nadaraya–Watson estimator of m(x) is

$$\hat{m}(x) = \sum_{j=1}^{n} w_j(x) y_i \quad \ldots\ldots\ldots\ldots\ldots \quad (2)$$

Under reasonable conditions on m(.) and the design points x, $\hat{m}(x)$ will be consistent for m(x) as $b \to 0$, $nb \to \infty$ when $n \to \infty$

Nonparametric regression-based estimator of the total

We can let $X = x_j$, any point in the non-sample and estimate $m(x_j)$. Then analogous to $\hat{T}_{lin}$, Dorfman (1994) has suggested $\hat{T}_{np} = \sum_{j \in s} y_j + \sum_{i \in p-s} \hat{m}(X_i)$ as an estimator of T.

As with model-based estimators generally, this estimator ignores the sampling probabilities.

The conditional mean and variance of the prediction error under (1) have been derived by Dofman (1994) and can be expressed as

$$E\left(\hat{T}_{np} - T\right) = \sum_{i \in r} \{d_s(x_j)\}^{-1} \left[\sum_{j \in s} \frac{1}{bn} k\left(\frac{x_i - x_j}{n}\right) m(x_j)\right] - m(x_i)$$

and

$$Var\left(\hat{T}_{np} - T\right) = \sum_{i \in s} w_i^2(x)\sigma^2(x_i) + \sum_{j \in r} \sigma^2(x_j) \quad \ldots\ldots\ldots\ldots\ldots \quad (3)$$

where $r = s^c$. Under some mild assumptions, Mwalili (1997) has derived the following asymptotic bias and variance expressions:

$$\text{Bias}\left(\hat{T}\right) \approx \frac{b^2}{2} \sum_{i \in r} m''(x_i) \int_0^1 u^2 k(u) du$$

and

$$Var\left(\hat{T}_{np}-T\right)\approx\frac{1}{bn}\left[\sum_{i\in s}\sum_{j\in r}\sigma^2(x_i)\int_{\frac{x_i}{h}}^{\frac{x_i-1}{h}}k(u)k\left(\frac{x_k-x_i}{h}+u\right)du\right]+\sum_{i\in r}\sigma^2(x_i)\ .$$

The above results imply that $\hat{T}_{np}\xrightarrow{P}T$ i.e. $\hat{T}_{np}$ is a consistent estimator of T. This result implies that $\hat{T}_{np}$ is bias robust to misspecification of $\mathrm{E}\left[Y_i|X_i=x_i\right]$ and also to the misspecification of $Var\left(Y_i|X_i=x_i\right)$. However, Mwalili (1997) has shown that

$$Var\left(\frac{\hat{T}_{np}}{N}\right)\approx 0\left(n^{-\frac{4}{5}}\right)$$

indicating that the bias robustness of $\hat{T}_{np}$ is attained at the cost of reduced efficiency.

The estimation of the error-variance

The error variance, (3), is clearly a function of $\sigma^2(x_i)'s$ that are unknown. Consider the squared residual

$$\hat{e}_j^{\,2}=\left(y_j-\hat{m}(x_j)\right)^2\ .$$

Under some mild assumptions on $m(x_i), x_i's$, and $W_i(x)'s$ and if n >1 it can be shown that

$$E\left(\hat{e}_j^2|X_j=x_j\right)=\sigma^2(x_j)+0(n^{-1})$$

showing that $\hat{e}_j^2$ is an asymptotically unbiased estimator of $\sigma^2(x_j)$. Hence, on taking $\hat{e}_j^{\,2}$ as a naive estimator of $\sigma^2(x_j)$, an improved estimator of $\sigma^2(x_j)$, can be obtained by smoothing $\hat{e}_j^2$ for $j\in s$ and $(x_j,y_j)s$ are sample points close to (x_i,y_i). Let h be a smoothing parameter (not necessarily equal to b). Using this parameter one can define the weight $w_i(x)$ and then write $\hat{\sigma}^2_{np}(x_i)=\sum_{j\in s}w_j(x_i)\hat{e}_j^2$ so that the error variance, (3), may be estimated by $v_n=\sum_{j\in s}w_j^2(X_i)\hat{\sigma}^2_{np}(x_i)+\sum_{j\in r}\hat{\sigma}^2_{np}(x_i)$. This estimator can never take negative values.

Assuming that $w_i(x),\sigma^2(x_i)$ are Lipschitz continuous and that the design points $X_i's$ are everywhere dense on some interval, Mwalili (1997), has verified that

$$E(V_n)\approx\sum_{j\in s}\left[w_j^2(x_i)\sigma^2(x_i)+\frac{h^2}{2}\sigma^{2'}(x_i)d_k\right]+\sum_{i\in r}\left(\sigma^2(x_i)+\frac{h^2}{2}\sigma^{2'}(x_i)d_k\right)$$

where $d_k=\int_{\frac{x_i-x_{j-1}}{h}}^{\frac{x_i-x_j}{h}}u^2k(u)du$.

If $h\to 0$ as $n\to\infty$, then $E(V_n)\approx Var\left(\hat{T}_{np}-T\right)$ asymptotically, under some mild assumptions on $\sigma^2(x_i)$. A consequence of this is that V_n is robust against model misspecification. This is the gain that one hopes to achieve when he uses this estimator against the following popular variance estimators (Royall & Cumberlard, 1981):

$$V_D = \frac{N^2(1-f)}{n^2}\frac{\bar{X}\,\bar{X}_r}{\bar{x}_s^2}\sum_{i\in s}\hat{e}_i^2(1-k_i)^{-1}$$

$$V_L = \frac{N^2(1-f)}{n\bar{x}_s^2}\bar{X}\,\bar{X}_r\sum_{i\in s}\frac{\hat{e}_i}{\sqrt{x_i}}$$

and

$$V_c = \frac{N^2(1-f)}{n}\left[\sum_{i\in s}\hat{e}_i^2/n-1\right]$$

where $f = n/N$; $\hat{e}_i = y_i - \frac{\bar{y}_s}{\bar{x}_s}x_i$, $\bar{y}_s$ *is the sample mean of* y_i's, $\bar{X}_r$, $\bar{x}_s$ represent non-sample and sample means of x_i's respectively. The estimator V_D is commonly used to estimate the variance of the ratio estimator: $\hat{T}_R = \frac{\bar{y}_s}{\bar{x}_s}\sum_P x_i$.

AN EMPIRICAL EXAMPLE

Properties of the four variance estimators V_L, V_D, V_c and V_n are studied in one natural population (NP) whose scatter plot is given in figure 1. This population of size 195 is the production by mass of dairy products by different farms in Kenya for the year 1992 and 1996 (source: Kenya Bureau of Statistics). The 1992 served as the auxiliary variable while the 1996 as the variable under study, $y_i\ (i = 1,2,.....195)$. Using simple random sampling without replacement we drew 1000 samples of size n = 40 from the population. Epanechnikov kernel $k(u) = \frac{1}{4}(1-u^2)$ if $|u|<1$ or zero otherwise; was used in the study whenever $\hat{T}_{np}$ and V_n were employed. An optimal bandwidth was selected using Silverman (1986) criterion: $\frac{\sigma}{4n^{1/5}} \le h \le \frac{3\sigma}{2n^{1/5}}$

For each sample we computed $T = \sum_{i=1}^{N} y_i$, the prediction errors: $E_R = \hat{T}_R - T$; $E_{mw} = \hat{T}_{np} - T$.

The variance estimators V_L, V_D, V_c and V_n were computed for each sample. The mean square error values: ; $MSE_{mw} = \sum_{1000} E_{mw}^2/1000$ were computed, where $\sum_{1000}$ denotes summation over 1000 samples. The results were $MSE_{mw} = 8.3717\ E-04$; $MSE_R = 9.2551\ E-04$; indicating that there is no significant difference between $\hat{T}_R$ *and* $\hat{T}_{np}$ for this particular population. Hence one can use either $\hat{T}_{np}$ or $\hat{T}_R$ to estimate T. This implies that one can use V_n to estimate MSE_R. This served as a motivation for using this population to compare the variance estimators. To see how each variance estimator is effective in tracking the conditional MSE_R, we sorted the samples in an increasing order and then grouped the samples into groups of 50 samples in that order. We then computed conditional $MSE_R = \sum_{1000} E_R^2/1000$; the conditional variance estimates

$\sum_{50}^{V}(\,)/_{50}$ for each of the four variance estimators in each of the 20 groups. The graphs of increasing $\overline{x_s}$ against MSE_R and $\sum_{50}^{V}(\,)/_{50}$ were plotted to gauge how different variance estimators are effective in tracking the conditional MSE_R. The results are given in figure 1. Clearly both V_D and V_n follow MSE_R very closely indicating that V_n is a strong competitor to V_D. Both V_C and V_L performed rather poorly.

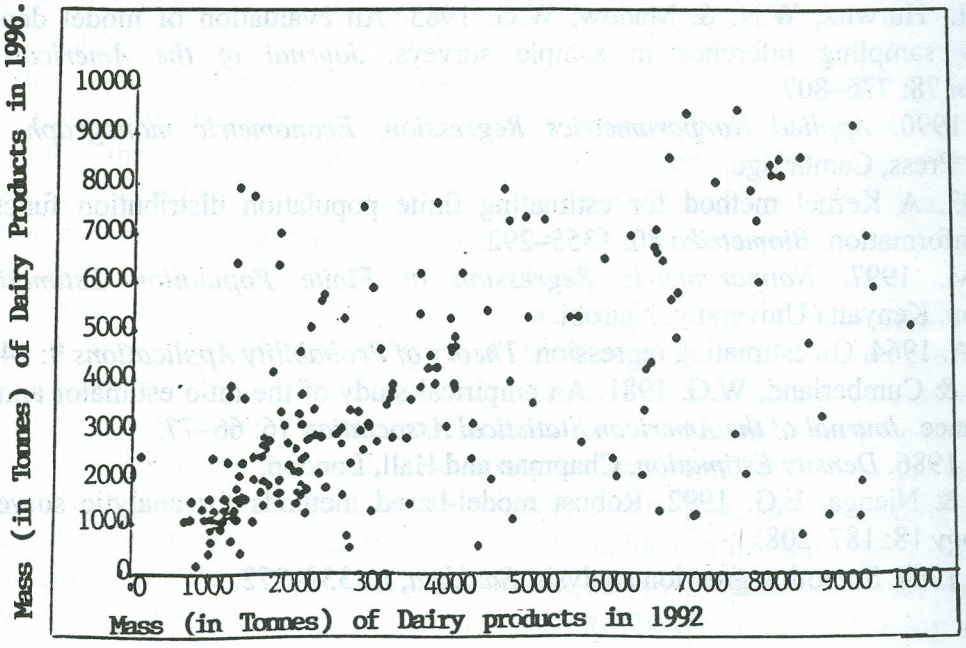


Figure 1. Scatter diagram for the population.

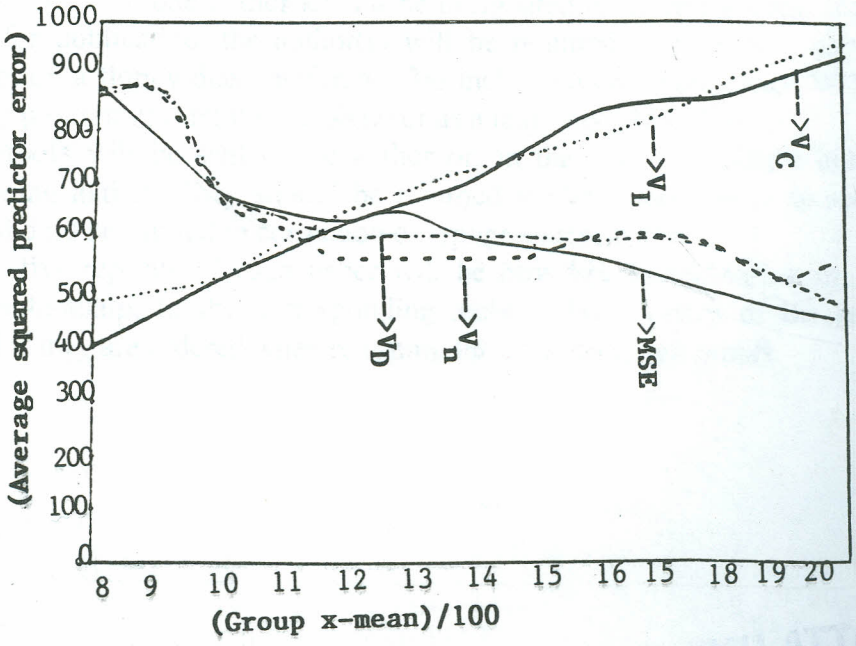


Figure 2. Average squared predictor error.

REFERENCES

Chambers, R.L., Dorfman, A.H. & Wehrly, T.E. 1993. Bias robust estimation in finite populations using nonparametric calibration. *Journal of the American Statistical Association* 88: 268–277.

Cochran, W.G. 1977 *Sampling Techniques*. Edn. 3. Wiley Eastern Publication, New York.

Dorfman, A.H. 1994. Nonparametric regression estimation of a finite population total. Unpublished manuscript.

Hansen, M.H., Hurwitz, W.N. & Madow, W.G. 1983. An evaluation of model dependent and probability sampling inference in sample surveys. *Journal of the American Statistical Association* 78: 776–807.

Hardle, W. 1990. *Applied Nonparametrics Regression, Econometric monograph.* Cambridge University Press, Cambridge.

Kuk, A. 1993. A Kernel method for estimating finite population distribution functions using auxiliary information. *Biometrika* 80: 1355–292.

Mwalili, T.M. 1997. *Nonparametric Regression to Finite Population Estimation.* M.Sc. Dissertation, Kenyatta University, Nairobi.

Nadaraya, E.A. 1964. On estimating regression. *Theory of Probability Applications* 9: 141–142.

Royall, R.M. & Cumberland, W.G. 1981. An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association* 76: 66–77.

Silverman, *. 1986. *Density Estimation.* Chapman and Hall, London.

Smith, T.M. & Njenga, E.G. 1992. Robust model-based methods for analytic surveys. *Survey Methodology* 18: 187–208.

Watson, G.S. 1964. Smooth regression analysis. *Sankhya*, A: 359–372.