Artificial Neural Network Model Based Estimation of Finite Population Total

Robert Kasisi[1], Romanus O. Odhiambo[2], Anthony G. Waititu[3]

[1, 2, 3]Jomo Kenyatta University, School of Mathematical Sciences, Department of Statistics and Actuarial Sciences
P.O Box 62000, Nairobi-Kenya

Abstract: ***In a finite population denoted U=1, 2, 3,....., N of N identifiable units, let X be the survey variable. The survey variables are observations from a super population. It is possible to get total information about these survey variables such as their total population, mean or their variance. In most cases auxiliary information about X is provided. A simple approach to using this auxiliary information is to assume a working model that describes the connection between the study variable of interest and the auxiliary variables. Estimators are then derived on the basis of this model. The best estimators are the ones that have good efficiency if the model is true, and are consistent if the model is inappropriate. In this study, we derive a nonparametric artificial neural network estimator of finite population total. The estimator is design unbiased, design consistent and asymptotic normal.***

Keywords: super population, auxiliary information, artificial neural networks, sample, survey sampling.

1. Introduction

A finite population is a list or a frame denoted U=1, 2, 3,....., N of N identifiable units. In this case N is usually known. E.g. we can have a finite population of size N=1000 .The need to get certain information about the finite population which cannot be cheaply obtained by involving every individual calls for a sample (a selection of the population) to be taken. Finite populations are of interest to government for policy making.

Total information about a population can be obtained from census data where every individual is involved in giving information. A simple way to incorporate known population totals of auxiliary variables is through ratio and regression estimation. More general situations are handled by means of generalized regression estimation (Sarndal, 1980) and calibration estimation (Deville and Sarndal, 1992). Estimation procedures have been employed in getting information from the census data, administrative registers and other surveys. However, in most cases these are challenging due to cost, time, literacy levels and other geographical factors. In these methods, part of the population referred to as the sample is used and the information about the population is inferred into the sample. For estimating the finite population total we suggest an alternative estimation procedure using artificial neural networks.

2. A Neural Network Model Based Estimator

The goal is to estimate the population mean of the survey variable so that we can get the population total, that is

$$\bar{Y} = \frac{1}{N}\sum_{I=1}^{N} y_i$$

From Deville and S¨arndal (1992) using the notion of a calibration estimators, we can define our artificial neural network estimator to be a linear combination of the observations

$$\hat{\bar{Y}} = \sum_{i=1}^{N} w_i y_i$$

With weights chosen to minimize an average distance measure from the basic design weights

$$d_i = \frac{1}{\pi_i}$$

Minimization is constrained to satisfy $\frac{1}{N}\sum_{i=1}^{N} w_i x_i = \bar{x}$, where $\bar{x}$ is the known vector of population means for the auxiliary variables. Although alternative distance measures are available in Deville and S¨arndal (1992), all resulting estimators are asymptotically equivalent to the one obtained from minimizing the chi-squared distance

$$\emptyset_s = \frac{\sum_{i\in s}(w_i - d_i)^2}{d_i q_i}$$

Where the $q_i{}'s$ are known positive weights unrelated to d_i , i.e.

$$\hat{\bar{Y}}_{nn} = \hat{\bar{Y}} + (\bar{x} - \hat{\bar{x}})^T \hat{\beta}$$

Where $\hat{\bar{Y}}_{nn}$ and $\hat{\bar{x}}$ are the Horvitz-Thompson estimators of $\bar{Y}$ and $\bar{x}$, respectively, and

$$\hat{\beta} = (\sum_{i=1}^{n} d_i q_i x_i x_i^T)\sum_{i=1}^{n} d_i q_i y_i$$

Consider the following superpopulation model

$$\begin{cases} E_\varepsilon(y_i) = f(x_i)\ for\ i = 1,2,\cdots,N \\ V \in (y_i) = \sigma^2 v(x_i)^2\ for\ i = 1,2,\cdots N \\ C_\varepsilon(y_i y_j) = 0\ for\ i \neq j \end{cases}$$

Where E_ε and V_ε denote expectation and variance, respectively, with respect to ε; $f(x_i)$ takes the form of a feed forward neural network with skip-layer connections and $V(.)$ is a known function of x_i. Hence,

$$f(x_i) = \sum_{q=1}^{Q}\beta q\ x_{qi} + \sum_m^M a_m\ \emptyset(\sum_{q=1}^{Q}\gamma_{qm}\, x_{qi} + \gamma_{0m}) + a_0 \quad (1)$$

Where M is the number of neurons at the hidden layer (Ripley, 1996, Chapter 5). Since we consider M as fixed, we can denote by the set of all parameters of the network, and write $f(xi)$ in (1) as $f(xi, \theta)$,

$$\theta = \{\beta_1, \cdots, \beta_q, a_0, a_1, \cdots, a_M, \gamma_{0M}, \gamma_1, \cdots, \gamma_M\}$$

Following the approach of Wu and Sitter (2001) to estimate $\tilde{\theta}$,the first step is to obtain a design-based method for estimating the model parameters and therefore obtain estimates of the regression function at x_i, for i=1,...,N , through the resulting fitted values. In other words, we first seek for an estimate $\tilde{\theta}$ of the model parameters θ based on the data from the entire finite population. We then obtain $\hat{\theta}$ a design-based estimate of $\tilde{\theta}$ based on the sampled data only. The population parameter $\tilde{\theta}$ is defined by weighted least squares with a weight decay penalty term, i.e.

$$\tilde{\theta} = argmin_{\theta \in \emptyset}\left\{\frac{1}{v(x_i)^2}(y_i - f(x_i,\theta))^2 + \frac{\lambda}{N}\sum_{i=1}^{p}\theta_i^{\,2}\right\} \quad (2)$$

where λ is a tuning parameter and p is the dimension of the parameters vector θ. The estimate $\hat{\theta}$ is defined as the solution of the design-based sample version of (2), that is

$$\hat{\theta} = argmin_{\theta \in \emptyset}\left\{\frac{1}{\pi_i}\left[\frac{1}{v(x_i)^2}(y_i - f(x_i,\theta))2 + \frac{\lambda}{N}\sum_{i=1}^{p}\theta_i^{\,2}\right]\right\} \quad (3)$$

Once the estimates $\hat{\theta}$ are obtained, the available auxiliary information is included in the estimator through the fitted values $\hat{f} = f(x_i,\hat{\theta}), for\ i = 1,2,\cdots N$.Then, we can define the neural network estimator as $\hat{\hat{Y}}_{nn} = \frac{1}{N}\sum_{i=1}^{N} w_i x_i$ where the calibrated weights w_i are sought to minimize the distance measure $\emptyset_s$ subject to $\frac{1}{N}\sum_{i=1}^{N} w_i = 1$ and $\frac{1}{N}\sum_{i=1}^{n} w_i\hat{f}_i = f(x_i,\hat{\theta})$

Using the technique of Deville and S¨arndal (1992) to derive the optimal weights, We can propose that

$$\hat{\hat{Y}}_{nn} = \hat{\hat{Y}}_{nn} + \frac{1}{N}\left\{\sum_{i=1}^{N}\hat{f}_i - \sum_{i=1}^{n} d_i\hat{f}_i\right\} \quad (4)$$

Where

$$\check{y} = \frac{\sum_{i=1}^{n} d_i q_i y_i}{\sum_{i=1}^{n} d_i q_i} \text{ And } \check{f} = \frac{\sum_{i=1}^{n} d_i q_i \hat{f}_i}{\sum_{i=1}^{n} d_i q_i}$$

We wish to combine the kernel technique to our neural network estimation. Therefore we briefly describe kernel smoothing.

A continuous kernel is denoted as k (.) and the bandwidth as h. The conditional regression estimator $\mu(x)$is the solution to a natural weighted least squares problem being the minimizer $\hat{\beta}_0$ of

$$S(\beta 0) = \sum_{i=1}^{n}(y_i - \beta_0)^2 k\left(\frac{x - xi}{h}\right) \quad (5)$$
$$= \sum_{i=1}^{n}(y_i - \beta_0)^2 w_i \quad (6)$$

Where

$$w_i = k\left(\frac{x - x_i}{h}\right)$$

By differentiating equation (6) with respect to β_0 and equating to zero we get

$$\frac{\partial s(\beta_0)}{\partial(\beta_0)} = 0$$
$$-2\sum(y_i - \beta_0)\, w_i = 0$$
$$\sum(y_i - \beta_0)\, w_i = 0$$
$$\sum(y_i)\, w_i = -\beta_0\sum w_i$$
$$\hat{\mu}(xj) = \hat{\beta}_0$$
$$= \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}$$
$$= \frac{\sum_{i=1}^{n} k\left(\frac{x - x_i}{h}\right) y_i}{\sum_{i=1}^{n} k\left(\frac{x - x_i}{h}\right)}$$

For a target $x_j, j = 1,2,\cdots,N$ we have

$$\hat{\mu}(xj) = \hat{\beta}_0$$
$$= \frac{\sum_{i=1}^{n} w_{ij} y_i}{\sum_{i=1}^{n} w_{ij}}$$
$$= \frac{\sum_{i=1}^{n} k\left(\frac{x_i - x_{ij}}{h}\right) y_i}{\sum_{i=1}^{n} k\left(\frac{x_i - x_{ij}}{h}\right)}$$

Similarly

$$w_{ij} = \frac{k\left(\frac{x_i - x_j}{h}\right)}{\sum_{i=1}^{n} k\left(\frac{x_i - x_j}{h}\right)}$$

So that

$$\hat{\mu}(x_j) = \frac{\sum_{i=1}^{n}\frac{k\left(\frac{x_i - x_j}{h}\right)}{\sum_{i=1}^{n} k\left(\frac{x_i - x_j}{h}\right)} y_i}{\sum_{i=1}^{n}\frac{k\left(\frac{x_i - x_j}{h}\right)}{\sum_{i=1}^{n} k\left(\frac{x_i - x_j}{h}\right)}}$$
$$= \sum_{i=1}^{n} w_{ij}\, y_i$$

i.e $\hat{\mu}(x_j)$ is an approximation of $\mu(x_j)$with a constant weighting value of Y corresponding to x_i's closest to x_j more heavily. Alternatively, let y_s=$[y_i] i \in s$ be the n vector of y_i's obtained in the sample. Define the $n \times 1$ matrix $X_{sj} = [1]_{n\times 1}$and define the $n \times n$ matrix

$$w_{sj} = \frac{1}{h} diag\left\{k\left(\frac{x - x_i}{h}\right)\right\}_{i\in s}$$

Then a sample based estimator of $\mu(x_j)$ is given by

$$\hat{\mu}(x_j) = \left(X'_{sj}W_{sj}X_{sj}\right)^{-1}X'_{sj}W_{sj}X_{sj} = \widehat{W}_{sj}y_s$$

as long as $X'_{sj}W_{sj}X_{sj}$ is invertible.

It follows that $\left(X'_{sj}W_{sj}X_{sj}\right)^{-1}X'_{sj}W_{sj}X_{sj} = \widehat{W}_{sj}y_s$

$$= \frac{\sum_{i=1}^{n}\frac{k\left(\frac{x_i - x_j}{h}\right)}{\sum_{i=1}^{n} k\left(\frac{x_i - x_j}{h}\right)} y_i}{\sum_{i=1}^{n}\frac{k\left(\frac{x_i - x_j}{h}\right)}{\sum_{i=1}^{n} k\left(\frac{x_i - x_j}{h}\right)}}$$
$$= \sum_{i=1}^{n} w_{ij}\, y_i$$

We note that we can use the neural network package (nnet) method to obtain the mean function of the fitted values. From the kernel technique,

$$\hat{\mu}(x_j) = \left(X'_{sj}W_{sj}X_{sj}\right)^{-1}X'_{sj}W_{sj}X_{sj}$$

The weights W_{sj} are subjected to the network and learnt. Then the network adjusts the weights until they are optimal. Now the mean function of the fitted values will be $y.hat$ =nnmodel$fitted.values*maximum($y_i$).

Therefore $\hat{f}_i = \hat{\mu}(x_j) = \left(X'_{sj}y.hatX_{sj}\right)^{-1}X'_{sj}y.hat_{sj}y_s$

In other words $y.hat = k\left(\frac{x_i - x_j}{h}\right)$

3. Asymptotic P roperties of Ar tificial Neu ral Network Estimator

3.1 Unbiasness

For a design expectation $E_p$and model expectation ε_p then, $\lim_{N\to\infty}\{E_p(Y_{nn})\} = Y$

Next,

$$\lim_{n\to\infty}\{E_p(Y_{nn})\} = \lim_{N\to\infty} E_p\left\{\frac{y_i}{\pi_i} + \left(\sum_{i=1}^{N}\hat{\mu}_i - \sum_{i=1}^{n}\frac{\hat{\mu}_i}{\pi_i}\right)\right\}$$

$$= \lim_{n\to\infty}\{E_p(Y_{nn})\} = \lim_{N\to\infty} E_p\left\{\frac{y_i I_i}{\pi_i} + \left(\sum_{i=1}^{N}\hat{\mu}_i - \sum_{i=1}^{N}\frac{E_p(I_i)\hat{\mu}_i}{\pi_i}\right)\right\}$$

But, $E(I_i)\pi_i$
Hence

$$\lim_{N\to\infty} E_p\left\{\sum_{i=1}^{n} y_i + \left(\sum_{i=1}^{N}\hat{\mu}_i - \sum_{i=1}^{N}\hat{\mu}_i\right)\right\}$$

$$= \lim_{N\to\infty} E_p \sum_{i=1}^{n} y_i$$

$$= N\bar{Y}$$

$$= Y$$

3.2 Consistency

By Chebchev technique

$$p[|X_n - \theta| > \varepsilon] \leq \frac{E_p|Y_{nn} - Y|^2}{\varepsilon^2}$$

Then

$$p[|Y_{nn} - Y| > \varepsilon] \leq \frac{E_p|Y_{nn} - Y|^2}{\varepsilon^2}$$

We know Y_{nn} is unbiased, therefore has a bias =0.
From, $\mathrm{MSE}(Y_{nn}) = \mathrm{Var}(Y_{nn}) + (bias(Y_{nn}))^2 = \mathrm{Var}(Y_{nn})$
Next

$$p[|Y_{nn} - Y| > \varepsilon] \leq \frac{\mathrm{Var}(Y_{nn})}{\varepsilon^2}$$

$$= \lim_{N\to\infty} P[|Y_{nn} - Y|] > \varepsilon] \leq \lim_{N\to\infty} \frac{\mathrm{Var}(Y_{nn})}{\varepsilon^2}$$

but

$$\lim_{N\to\infty} \frac{\mathrm{Var}(Y_{nn})}{\varepsilon^2} = 0$$

Due to convergence in probability then,
$\lim_{N\to\infty} P[|Y_{nn} - Y|] > \varepsilon \to 0$
Hence Y_{nn} is consistent.

3.3 Asymptotic Normality

$$\hat{Y}_{nn} = \frac{y_i}{\pi_i} + \left(\left(\sum_{i\to\infty}^{N}\hat{\mu}_i + \sum_{i=1}^{n}\frac{\hat{\mu}_i}{\pi_i}\right)\right)$$

Then $\frac{\frac{1}{N}(\hat{Y}_{nn} - Y)}{\sqrt{var(\frac{1}{N}\hat{Y}_{nn})}} \to N(0,1)$ As $N \to \infty$ implies $\frac{\frac{1}{N}(Y_{nn} - Y)}{\sqrt{var(\frac{1}{N}Y_{nn})}} \to$

$$N(0,1)$$

4. Some Results Displayed

Using R statistical package we simulate two populations of x as independent and identically distributed uniform (0,1) and gamma (1,1) random variables. The populations are of size N=300.Samples of size n=30 are generated by simple random sampling. The population size is considered large enough for several samples and the sample size is 10 percent of population size. For each population of x ,mean function , and bandwidth ,100 replicate samples are generated and the estimates calculated .The population is kept fixed during these 100 replicates in order to be able to evaluate the design averaged performance of the estimators. We consider four mean functions:

1. Linear $2 + 5x$
2. Quadratic $(2 + 5x)^2$
3. Exponential $exp(-8x)$
4. $Cycle1$ $2 + sin(2\pi x)$

We report on some performance of several estimators.

The Horvitz Thompson estimator is a design based estimator while the others are nonparametric estimators which are model based. The Epanechnikov kernel

$$k(u) = \frac{3}{4}(1 - u^2), u \leq 1$$

is used for all four nonparametric estimators. Several bandwidths are considered (h=0.1, h=0.25, h=0.5, h=0.75, h=1 and h=2) to help see how efficiency of the estimators vary with bandwidth. The second bandwidth is based on the ad hoc rule of $\frac{1}{4}th$ the data range. The bandwidths h=1 and h=2 are large bandwidths relative to the data range,[0,1]. For the linear mean function, Y_{nn} and Y_{lp} the results show equal performance evident from equal mean squared errors for both uniform and gamma distributions. We therefore examine how much efficiency is lost if we used the other estimators. The other mean functions represent departures from the linear model. For quadratic function Y_{nn} performs better followed by Y_{lp} (linear), except for a small portion for the range of x i.e for (h=0.1, h=0.5, and h=0.75 Y_{lp} (linear) performs better under the gamma distribution. The biases at these turning points for Y_{lp} (linear) are seen to be less compared to those of Y_{nn} .For the exponential mean function under uniform distribution, Y_{nn} performs better followed by Y_{lp}(linear).It is interesting to see the cycle and exponential mean functions yield similar MSE values under gamma distribution. The performance of any estimator,$\hat{Y}$ in $\{Y_{HT}, Y_{nw}, Y_{nn}, Y_{lp}\}$ is evaluated using its relative bias R_B and MSEs.The relative bias is defined as

$$R_B = \frac{\sum_{r=1}^{B}(\hat{Y} - Y)}{R \times Y}$$

where R is the replicate number of samples. We evaluate the actual design variance and estimated the mean squared error as $MSE(\hat{Y}) = var(\hat{Y}) + (R_B)^2$

We also consider an estimate of the mean square error

$$MSE(\hat{Y}) = \frac{\sum_{r=1}^{R}(\hat{Y}_r - Y)^2}{R}$$

Where $\hat{Y}_r$ is calculated from the R^{th} simulated sample.

5. Table of Results

Table 1: Comparative MSEs for the nonparametric estimators for a quadratic mean function under uniform distribution

uniform	*MSE of nn*	*MSE of local polynomial*	*MSE of Local linear*	*MSE of ht*
h= 0.1	44.0644	50.36182	168.5472	36196.61
h= 0.25	44.0644	50.36182	51.19504	36196.61
h= 0.5	44.0644	79.97721	145.6657	36196.61
h= 0.75	44.0644	127.6176	162.6657	36196.61
h= 1.0	44.0644	146.0463	168.5472	36196.61
h= 2.0	44.0644	164.5679	174.183	36196.61

Table 2: Comparative biases for the nonparametric estimators for a quadratic mean function under uniform distribution

uniform	*Bias. nn*	*Bias. local polynomial*	*Bias. Local linear*	*Bias.ht*
h= 0.1	0.004388394	0.1610076	0.08051538	0.1257754
h=0.25	0.004388394	0.004691507	0.004730158	0.1257754
h= 0.5	0.004388394	0.005912141	0.00797885	0.1257754
h=0.75	0.004388394	0.007468217	0.00858267	0.1257754
h= 1.0	0.004388394	0.007989266	0.00858267	0.1257754
h= 2.0	0.004388394	0.008480749	0.008724983	0.1257754

Table 3: Comparative MSEs for the nonparametric estimators for a quadratic mean function under gamma distribution

gamma	*MSE of nn*	*MSE of local polynomial*	*MSE of local linear*	*MSE of ht*
h=0.1	22431.97	1171973	7735.465	231856.3
h= 0.25	22431.97	2318298	628123.9	231856.3
h= 0.5	22431.97	9334244	6453.919	231856.3
h= 0.75	22431.97	230634	5626.875	231856.3
h= 1.0	22431.97	43149.25	756015.2	231856.3
h= 2.0	22431.97	596.9408	836.358	231856.3

Table 4: Comparative Absolute biases for the nonparametric estimators for a quadratic mean function under gamma distribution

gamma	*Bias nn*	*Bias. Local polynomial*	*Bias. Local linear*	*Bias.ht*
h= 0.1	0.05004112	0.361703	0.02938574	0.1501117
h= 0.25	0.05004112	0.5087191	0.2647988	0.1501117
h= 0.5	0.05004112	1.020782	0.0268414	0.1501117
h= 0.75	0.05004112	0.1604557	0.02506265	0.1501117
h= 1.0	0.05004112	0.06940328	0.2905084	0.1501117
h= 2.0	0.05004112	0.008163174	0.0096625	0.1501117

6. About Efficiency

Considering the MSEs of the various estimators, we make several observations.Y_{nn} Performs exceptionally well under linear and quadratic functions. Y_{lp} also performs well since its itself linear, and hence is almost a true model for the linear function. For most of the other mean functions, Y_{nn} retained consistent efficiencies. Therefore from our results we are able to meet our objective that the artificial neural network outperforms the kernel and local polynomial estimators.

7. Conclusions

We have derived an artificial neural network estimator for finite population total. The properties of this estimator outclass the existing nonparametric estimators such the kernel and local polynomial estimators. We also note that this estimator remains invariant (i.e. gives same result) under different bandwidths. The only closest competitor of this estimator is the linear local polynomial estimator. However our estimator is more applicable since we do not have to determine the degrees to use. We have also found that if the mean $\mu(x)$of a sample is known, then we can use this information to find the mean of the non-sampled elements which leads to overall population estimation. Our objectives have been achieved that the artificial neural network estimator outperforms kernel estimators and also local polynomial estimators. We have also successfully derived the asymptotic properties of the estimator. These are asymptotic unbiasness, design consistency and asymptotic normality. The overall remark is that an artificial neural network estimator is an improvement of existing nonparametric and parametric estimators.

References

[1] Breidt,F.J Kim,J.Y and Opsormer,J.D.(2005) Nonparametric regression of finite population totals under Two-stage sampling. *Annals of Statistics.25, 1026-1053.*

[2] Breidt,F.J and Opsormer,J.D.(2000).Local polynomial regression estimators in survey sampling. *Annals of Statistics.28, 1026-1053.*

[3] Cassel,et al (1976).Some results on generalized different estimation and generalized regression estimation for finite populations.*Biometrika 63,615-620.*

[4] Firth,D.and Bennet,K.E(2006).Robust models in probability sampling *.journal of Royal Statistical Society.B,17,267-278*

[5] K. Deb, S. Agrawal, A. Pratab, T. Meyarivan, "A Fast Elitist Non-dominated Sorting Genetic Algorithms for Multiobjective Optimization: NSGA II," KanGAL report 200001, Indian Institute of Technology, Kanpur, India, 2000. (technical report style)

[6] Godambe,V.P and Thompson, M.E(1973).Estimation in sampling Theory with exchangeable prior distributions. *Annals of Statistics.1,1212-21*

[7] B He ,t Oki,F Sun,D Komon,S Kanae ,Y Wang Estimating monthly total nitrogen concentration in streams by using artificial neural networks, *journal of environmental,2011*

[8] Wu,C. and Sitter,R.R(2001).Efficient estimation of quadratic finite population functions in the presence of auxiliary information. *Journal of American Statistical Association, 97,535-43.*

[9] William G.Cochran(1992),Sampling Techniques, *third edition,44-49,364-382*

[10] Godambe,V.P(1995)A Unified Theory of Sampling from finite populations.*journal of Royal Statistical Society.B,17,267-278*

[11] MA Mazurowski,PA Habas,JM Zurada JY L O Training neural network classifiers for medical decision making *Neural networks,2008*

[12] Isaki,C.T.and Fuller,W.A.(1982),Survey design under the regression super population model.

Author Profile

Robert K asisi received a BSc in Mathematics and Computer Science from Jomo Kenyatta University of Agriculture and Technology in the year 2011; He is currently pursuing a Master of Science in Applied Statistics at Jomo Kenyatta University of Agriculture and Technology.

Romanus O. Odhiambo is a professor of Statistics. He is currently working as a professor at Jomo Kenyatta University of Agriculture and Technology in the college of Pure and Applied Science, Department of Statistics and Actuarial Science. He holds PhD (Statistics) degree conferred to him at Kenyatta University, a Master's of Science (Statistics) degree from Kenyatta University

and Executive Master of Organizational Development (EMOD) from the United States International University (USIU).

Anthony G. Waititu holds a PhD in Applied Statistics. He is currently a senior lecturer in the department of Statistics and Actuarial Sciences, Jomo Kenyatta University of Agriculture and Technology.