A Note on Estimationg Variance of Finite Population Distribution Function

Sumanta Adhya
Tathagata Banerjee
Gaurangadeb Chattopadhyay

W.P. No. 2015-08-02
August 2015

INDIAN INSTITUTE OF MANAGEMENT
AHMEDABAD-380 015
INDIA

A note on estimationg variance of finite population distribution function

SUMANTA ADHYA

Department of Statistics, West Bengal State University, Barasat, Kolkata 700126, India

sumanta.adhya@gmail.com

TATHAGATA BANERJEE

Production and Quantitative methods, Indian Institute of Management, Vastrapur, Ahemedabad 380015, India

tathagata@iimahd.ernet.in

GAURANGADEB CHATTOPADHYAY

Department of Statistics, University of Calcutta, Kolkata 700035, India

gaurch@yahoo.co.in

SUMMARY

Estimating finite population distribution function is an important problem to the survey samplers since it summarizes almost all the relevant information of interest about the finite population. Moreover due to its nonlinearity estimation of variance of estimators of distribution function remains an active area of research since Chambers et al., 1992. Both analytic and resampling-based variance estimators are developed earlier. Here we poropse a bootstrap hybrid variance estimator of model-based semi-patametric estimator of finite population distribution function estimator. We prove its consistency and also show that its numerical performances are superior to analytical estimator.

Some key words: Model-based, Semi-parametric regression, P-splines, Analytical variance estimator, Bootstrap.

1. INTRODUCTION

Consider a finite population U of N distinct elements. Suppose a probability sample S of size n is taken from the population using an ignorable sampling design (Opsomer, 2009). Values of a survey variable y are observed for sample units. Then the finite population distribution function (hereafter FPDF) of survey variable y is defined by

$$F_N(t) = N^{-1}\sum_U I(y_i \le t), \tag{1}$$

where $I(y_i \le t) = 1$, if $y_i \le t$ and 0, otherwise. Now, if apart from values of y-variable the values $x_1,...,x_N$ of a vector-valued auxiliary variables x be known then following groundbreaking paper of Chambers & Dunstan (1986) a model-based predictive estimator (Valliant, 2009) of FPDF using a non-/semi-parametric regression model in the superpopulation (Dorfman & Hall, 1993) is more efficient; since it incorporates all the unit level information on the auxiliary variables through a regression robust to parametric model misspecification.

The regression model in the superpopulation is given by

$$y_i = m(x_i) + e_i,\ i \in U,$$

where $m(.)$ is an unknown but smooth function and e_i's, the error terms, are independently and identically distributed random variables following an absolutely continuous distribution function $G(.)$ with mean zero and finite variance σ_ε^2. Here we consider P-splines (Eilers & Marx, 1996; Ruppert et al., 2003) to model $m(.)$. To this end, we approximate $m(.)$ by truncated polynomial basis function

$$\mu(x;\boldsymbol{\beta}) = \beta_0 + \beta_1 x + ... + \beta_r x^r + \sum_{k=1}^{K}\beta_{r+k}(x-\kappa_k)_+^r,$$

where $\boldsymbol{\beta} = (\beta_1,...,\beta_{r+K})^T$, $(t)_+^r = t^r I(t>0)$ and κ_k's $(\kappa_1 < ... < \kappa_K)$ are fixed knots. In practice degree of spline r is low (≤ 3), $\kappa_k = \{(k+1)/(k+2)\}$-th sample quantile of x and $K = \min(35, n/4)$ (Wand, 2003). The regression coefficient $\boldsymbol{\beta}$ is obtained by minimizing penalized least square criterion

$$\sum_S \{y_i - \mu(x_i;\beta)\}^2 + \delta\sum_{k=1}^{K}\beta_{r+k}^2$$

for fixed $\delta(\geq 0)$. The quantity δ controls the amount of smoothing in $\mu(.;.)$ and its data-adaptive estimate $\hat{\delta}$ is obtained by using linear mixed model representaion of P-splines (Speed, 1991; Brumback et al., 1999; Wand 2003; Breidt et al., 2005). To define $\hat{\delta}$, let $\boldsymbol{d}_i$ and $\boldsymbol{z}_i$ be the row vectors $(1, x_i, ..., x_i^r)$ and $((x_i-\kappa_1)_+^r, ..., (x_i-\kappa_K)_+^r)$ respectively. The above semiparametric regression model can be written as a linear mixed model

$$y_i = \boldsymbol{d}_i^T\boldsymbol{\theta} + \boldsymbol{z}_i^T\boldsymbol{b} + e_i,$$

where $\boldsymbol{\theta}$ and $\boldsymbol{b}$ denote the vectors $(\beta_0, \beta_1, ..., \beta_r)^T$ and $(b_{r+1}, ..., b_{r+K})^T$ respectively. The errors e_i's are assumed to follow indepenpent $N(0, \sigma_e^2)$ and the random vector $\boldsymbol{b}$ is independent and follows multivariate normal with mean $\boldsymbol{0} = (0, ..., 0)^T$ and variance $\sigma_b^2\boldsymbol{I}_K$, $\boldsymbol{I}_K = diag(1, ..., 1)$. Then $\hat{\delta} = \hat{\sigma}_e^2/\hat{\sigma}_b^2$ where $\hat{\sigma}_e^2$ and $\hat{\sigma}_b^2$ be maximum likelihood (ML) or restricted maximum likelilood (REML) estimator based on sample data. The P-spline estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}_S^T\boldsymbol{X}_S + \boldsymbol{A}_{\hat{\delta}})^{-1}\boldsymbol{X}_S^T\boldsymbol{y}_S, \tag{2}$$

where $\boldsymbol{y}_S = (y_i, i \in S)^T$, $\boldsymbol{X}_S$ denotes the design matrix whose i-th row is giben by $\boldsymbol{X}_i = (1, x_i, ..., (x_i-\kappa_K)_+^r)$, $i \in S$, and penalty matrix $\boldsymbol{A}_\delta = diag(\underbrace{0, ..., 0}_{r+1}, \underbrace{\delta, ..., \delta}_{r+K})$. With the above choice of $\hat{\delta}$ the estimate $(1, x, ..., (x-\kappa_k)_+^r)\hat{\boldsymbol{\beta}}$ provides a smooth fit of regression function $m(x)$. Then the semi-parametric model-based estimator of FPDF is obtained as

$$\hat{F}_{ps}(t) = N^{-1}[\sum_S I(y_i \leq t) + n^{-1}\sum_{i\in S}\sum_{j\in\bar{S}} I\{y_i \leq t + (\boldsymbol{X}_i - \boldsymbol{X}_j)^T\hat{\boldsymbol{\beta}}\}], \tag{3}$$

where $\bar{S}(= U - S)$ is the set of non-sample units.

An immediate estimate of the variance of the prediction error may be directly obtained from the asymptotic expression of the variance of $\hat{F}_{ps}(t) - F_N(t)$ by plugging in the relevant observed quantities. For the expression of analytical variance we define following unknown quantities:

$\nu(t) = E\{e_1 I(e_1 \leq t)\}$, $G'(t) = \partial G(t)/\partial t$, $\min(u, v) = u \wedge v$, $\lim n^{-1}\sum_S \boldsymbol{X}_i = \boldsymbol{\mu}$,

$\lim n^{-1}\boldsymbol{X}_S^T\boldsymbol{X}_S = \boldsymbol{\Sigma}$, $\lim(N-n)^{-1}\sum_{i\in\bar{S}}\sum_{j\in\bar{S}} G\{(t - \boldsymbol{X}_i^T\boldsymbol{\beta}) \wedge (t - \boldsymbol{X}_j^T\boldsymbol{\beta})\} = I_1$,

$\lim(N-n)^{-1}\sum_{\bar{S}} G(t-X_i^T\beta)=I_2$, $\lim(N-n)^{-1}\sum_{\bar{S}} G'(t-X_i^T\beta)=\boldsymbol{I}_3$,

$\lim(N-n)^{-1}\sum_{\bar{S}} \boldsymbol{X}_i G'(t-\boldsymbol{X}_i^T\boldsymbol{\beta})=\boldsymbol{I}_4$, $\lim(N-n)^{-1}\sum_{\bar{S}} \nu(t-\boldsymbol{X}_i^T\boldsymbol{\beta})=I_5$, and

$\lim(N-n)^{-1}\sum_{\bar{S}} G(t-\boldsymbol{X}_i^T\boldsymbol{\beta})\{1-G(t-\boldsymbol{X}_i^T\boldsymbol{\beta})\}=I_6$,

where all above limits are finite expectations with respect to asymptotic density of x (Chambers et al., 1992; Dorfman & Hall, 1993). Also the limiting sampling fraction is $\rho\in[0,1)$. Then ignoring terms of order $o(n^{-1})$ the asymptotic variance is

$$\text{var}\{\hat{F}_{ps}(t)-F_N(t)\}\approx n^{-1}(1-f)^{-1}\{(I_1-I_2^2)+2\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}(\mu I_3-\boldsymbol{I}_4)I_5\}+n^{-1}f(1-f)I_6. \qquad (4)$$

We skip the details of the proof. The result can be obtained following similar steps as in the proof of (2.2) of Chambers et al., 1992.

The problem is to estimate the variance of the prediction error $\hat{F}_{ps}(t)-F_N(t)$. An immediate estimate of the variance of the prediction error may be directly obtained from the asymptotic expression of the variance of $\hat{F}_{ps}(t)-F_N(t)$ by plugging in the relevant observed quantities. But analytical estimators generally have poor numerical performances compared to resampling-based estimators (Wu & Sitter, 2001; Lombardia et al., 2003; 2004). In our situation we notice the same (cf. Section 3). In Section 2, we propose a bootstrap hybrid estimator of prediction error. For this we adopt standard approach (Ruppert et al., 2003; Breidt et al., 2005) by assuming that the smoothing parameter δ is fixed at its observed value $\hat{\delta}$. We prove consistency of bootstrap hybrid bootstrap estimator. In Section 3 we present the results of a simulation study comparing the performances of the proposed estimator with that of the analytical variance estimator. We also apply our methodology for a real data set. Finally we give concluding remarks in Section 4.

2. HYBRID VARIANCE ESTIMATION

2.1. *Methodology*

As an alternative to analytical variance estimator, we consider a bootstrap estimator of the asymptotic variance. In model-based set-up Monte Carlo approximation to the bootstrap estimate of the asymptotic variance of a predictive estimator requires two-stage sampling (Lombardia et al., 2003). First, a finite population is generated using the estimated superpopulation model from the data; next, a bootstrap ample is selected from it. This process needs to be repeated a sufficiently large number of times for obtaining a good approximation to the bootstrap variance estimator. The procedure is highly computer intensive. To circumvent the problem we adopt a hybrid approach (Adhya et al., 2011; 2012) which avoids generating a finite population at the first stage. Wu and Sitter (2001) use similar hybrid approach for jackknife variance estimation in finite population settings. We describe it in the following.

Throughtout this Section we assme that δ is fixed at its estimated value $\hat{\delta}$. Notice that,

$$\hat{F}_{ps}(t) - F_N(t) = n^{-1}N^{-1}\sum_{i\in S}\sum_{j\in\bar{S}} I\{y_i \le t + (X_i - X_j)^T\hat{\beta}\} - N^{-1}\sum_{\bar{S}} I(y_i \le t)$$

and from the conditional independence argument given S, we have

$$\begin{aligned}\text{var}\{\hat{F}_{ps}(t) - F_N(t)\} &= \text{var}[n^{-1}N^{-1}\sum_{i\in S}\sum_{j\in\bar{S}} I\{y_i \le t + (X_i - X_j)^T\hat{\boldsymbol{\beta}}\}] \\ &\quad + \text{var}\{N^{-1}\sum_{\bar{S}} I(y_i \le t)\} \equiv V_S(\hat{\delta}) + V_{\bar{S}}(\hat{\delta}),\ \text{say.}\end{aligned}$$

A natural estimator of the second term $V_{\bar{S}}(\hat{\delta}) = N^{-2}\sum_{\bar{S}} G(t - \boldsymbol{X}_i^T\boldsymbol{\beta})\{1 - G(t - \boldsymbol{X}_i^T\boldsymbol{\beta})\}$ is

$$\hat{V}_{\bar{S}} = N^{-2}\sum_{\bar{S}} \hat{G}(t - \boldsymbol{X}_i^T\hat{\boldsymbol{\beta}})\{1 - \hat{G}(t - \boldsymbol{X}_i^T\hat{\boldsymbol{\beta}})\}, \qquad (5)$$

where $\hat{G}(t) = n^{-1}\sum_S I(\hat{e}_i \le t)$ is empirical cumulative distribution function of e_i's based on residuals $\hat{e}_i = y_i - \boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}$, $i \in S$. The estimator $\hat{V}_{\bar{S}}$ of $V_{\bar{S}}$ is expected to exhibit negligible bias in finite samples since the model-based inference is preferred when sampling fraction n/N is small (Wu & Sitter, 2001). We estimate $V_S(\hat{\delta})$ by bootstrapping from the sampled data $\boldsymbol{S}_n = \{\boldsymbol{y}_S, \boldsymbol{x}_S\}, \boldsymbol{x}_S = (x_i, i \in S)^T$. The usual bootstrap estimator is given by

$$\hat{V}_{S,B} = \text{var}_B[n^{-1}N^{-1}\sum_{i\in S}\sum_{j\in\bar{S}} I\{y_i^* \le t + (\boldsymbol{X}_i^* - \boldsymbol{X}_j^*)^T\hat{\boldsymbol{\beta}}^*\}] \qquad (6)$$

where $\hat{\beta}^*$ is the penalized least square estimate based on bootstrap sample data $\boldsymbol{S}_n^* = \{\boldsymbol{y}_S^*, \boldsymbol{x}_S^*\}$ for a suitably chosen smoothing parameter δ^* and $\text{var}_B(.)$ denotes the variance with respect to the bootstrap distribution given the sample data $\boldsymbol{S}_n$. Using analogy with the choice of $\hat{\delta}$, we take δ^* as $\hat{\delta}^* = \sigma_e^{*2} / \sigma_b^{*2}$, where σ_e^{*2} and σ_b^{*2} are the estimates of $\boldsymbol{s}_e^2$ and $\boldsymbol{s}_b^2$ based on the bootstrap sample data $\boldsymbol{S}_n^*$. In practice the Monte Carlo approximation to $\hat{V}_{S,B}$ based on B bootstrap samples $S_b^*, b = 1,..., B$, is given by

$$B^{-1}\sum_{b=1}^{B}(u_b - \bar{u}),$$

where $u_b = n^{-1}N^{-1}\sum_{i\in S_b^*}\sum_{j\in\bar{S}} I\{y_{ib}^* \le t + (\boldsymbol{X}_{ib}^* - \boldsymbol{X}_j)^T \hat{\boldsymbol{\beta}}_b^*$, and $y_{ib}^*, \boldsymbol{X}_{ib}^*, \hat{\delta}_b^*, \hat{\boldsymbol{\beta}}_b^* (\equiv \hat{\beta}_b^*(\hat{\delta}_b^*))$ are the values of $y_i, \boldsymbol{X}_i, \delta^*$ and $\hat{\boldsymbol{\beta}}$ for the b-th bootstrap sample and $\bar{u} = B^{-1}\sum_{b=1}^{B} u_b$. Thus, from (5) and (6) the hybrid bootstrap estimate is $\hat{V}_{ps} = \hat{V}_{S,B} + \hat{V}_{\bar{S}}$.

2.2. *Consistency*

We now prove the weak consistency of the proposed hybrid estimator $\hat{V}_{ps}$ assuming $\hat{\delta}$ and δ^* as known fixed quantity. We consider a sequence of populations of sizes N_ν and sample sizes n_ν such that $n_\nu \to \infty$ and $f_\nu = n_\nu / N_\nu \to \rho \in [0,1)$, as $\nu \to \infty$. For notational convenience, we suppress the index ν below. For theoretical development, throughout we assume x to be continuous and its asymptotic density for the sample and non-sample design points are $h_S(.)$ and $h_{\bar{S}}(.)$ respectively (Chambers et al., 1992). That is,

$$n^{-1}\sum_S I(x_i \le x) \to \int_{-\infty}^{x} h_S(u)du \text{ and } (N-n)^{-1}\sum_{\bar{S}} I(x_i \le x) \to \int_{-\infty}^{x} h_{\bar{S}}(u)du \text{ as } n \to \infty.$$

For simple random sampling, since the inclusion probabilities of all the units are equal, we have $h_S(x) = h_{\bar{S}}(x)$ for all x (Dorfman and Hall, 1993) and then $N^{-1}\sum_U I(x_i \le x) \to \int_{-\infty}^{x} h(u)du$. Now we assume regularity conditions:

Condition 1. (e_i, x_i)'s are independent and identically distributed with e_k and x_k are independent, $G''(t) = \partial^2 G(t)/\partial t^2$ is continuous in t and the density of x is a continuous and bounded function.

Condition 2. Sampling design does not involve the auxiliary variable x.

For simplicity, we consider simple random sampling design in (*iii*) to prove consistency. Consistency of complex sampling designs, e.g., stratified sampling and probability proportional to size sampling is not considered. For our case the density of x is the asymptotic density $h(.)$.

Condition 3. $\lim n^{-1}\sum_S x_i^{4r} = \int x^{4r} h(x)dx = O(1)$.

Next for any $\boldsymbol{\beta}$, $F(t;\boldsymbol{\beta}) = \int\int G\{t - \boldsymbol{X}_1^T\boldsymbol{\beta}_0 + (\boldsymbol{X}_1 - \boldsymbol{X}_2)^T\beta\}h(x_1)h(x_2)dx_1dx_2$, which entails $F(t;\boldsymbol{\beta}_0) = \int G(t - \boldsymbol{X}^T\boldsymbol{\beta}_0)h(x)dx$ for true parameter value $\boldsymbol{\beta}_0$, where $\boldsymbol{X}_l^T = (1, x_l, ..., (x_l - \kappa_K)_+^r)$, $l = 1,2$, and $\boldsymbol{X}^T = (1, x, ..., (x - \kappa_K)_+^r)$. Note that $F(t,\boldsymbol{\beta})$ is a distribution function. Further we assume that

Condition 4. The partial derivative $\partial F(t,\boldsymbol{\beta})/\partial\boldsymbol{\beta} = (\partial F(t,\boldsymbol{\beta})/\partial\beta_o, ..., \partial F(t,\boldsymbol{\beta})/\partial\beta_{r+K})$ is not equal to the null vector at the true parameter value $\boldsymbol{\beta}_0$.

Now we state the following theorems. Proof is outlined in the Appendix.

THEOREM 1. *Under conditions* 1-4, *for fixed and known penalty parameters* $\hat{\delta}$ *and* $\hat{\delta}^*$ $\hat{V}_{ps}$ *is weakly consistent for* $\text{var}\{\hat{F}_{ps}(t) - F_N(t)\}$.

In finding the variance we ignore the uncertainty associated with the penalty parameter estimate. However, the impact of ignoring the uncertainty in its estimate does not seem to be serious. This fact follows from Section 2.2 of Breidt et al., 2005 by bringing an analogy of the sample and the corresponding bootstrap sample in our set up with the "finite population" and the probability "sample" drawn from it in Breidt et al.'s paper. This interpretation immediately implies that bootstrap choice of penalty parameter is consistent for penalty parameter obtained from sample data.

3. NUMERICAL RESULTS

3.1. *Model-based simulation*

We report the results of a limited simulation study comparing the performance of the hybrid bootstrap estimator $\hat{F}_{ps}$ to that of analytical estimator obtained by estimating (4) at the quantiles x_q for $q = 0.25, 0.50$ and 0.75 for six superpopulation models given in Table 1. We consider six different choices of $m(x)$ for the superpopulation model $y = m(x) + e$, where x and e are generated respectively from the $U(0,1)$ and $N(0,1)$ distributions. The chosen population size (N) and the sample size (n) (1000, 100) resulting in sampling fraction (f) equal to 0.10.

For analytical estimator we require to estimate (4). The estimators is

$$\hat{V}_A = n^{-1}(1-f)^{-1}\{(\hat{I}_1 - \hat{I}_2^2) + 2\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1}(\hat{\mu}\hat{I}_3 - \hat{\boldsymbol{I}}_4)\hat{I}_5\} + n^{-1} f(1-f)\hat{I}_6,$$

where $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\Sigma}}$ and $\hat{I}_i$'s $(i \neq 4)$ and $\hat{\boldsymbol{I}}_4$ are obtained by repacing unknown quantities $G(t)$, $G'(t)$ and $\nu(t)$ with their estimates $\hat{G}(t)$, $\hat{G^{'}}(t) = (nb_n)^{-1}\sum_S \mathrm{K}\{(\hat{e}_i - t)/b_n\}$ and $\hat{\nu}(t) = n^{-1}\sum_S \hat{e}_i$ $I(\hat{e}_i \leq t)$ respectively; K(.) is a kernel function, usually a density function, and b_n is the bin width. We use two choices of the kernel function (Wu & Sitter; 2001): (*i*) standard normal kernel with $b_n = 1.059\hat{\sigma}n^{-1/5}$ and (*ii*) standard logistic kernel with $b_n = 4\hat{\sigma}/n$ where $\hat{\sigma}^2 = (n-2)^{-1}\sum_S\{\hat{e}_i - n^{-1}\sum_S \hat{e}_i\}^2$. We denote the corresponding variance estimators by $\hat{V}_{A,1}$ and $\hat{V}_{A,2}$ respectively. The steps are:

1. Generate $(x_i, i = 1,...,N)$ randomly from uniform (0,1) distribution. For each x_i generate e_i independently from $N(0,0.01)$ distribution. Then y_i's are generated using $m(x_i)$'s and ε_i's. This forms a finite population of size $N = 1000$.

2. A simple ramdom without replacement (SRSWOR) sample of size $n = 100$ is drawn from the finite population generated in step1 and then compute the P-spline estimator $\hat{F}_{ps}(x_q)$,and estimators of its variance $\hat{V}_{A,1}$ and $\hat{V}_{A,2}$. We use linear spline model given by $\mu(x;\boldsymbol{\beta}) = \beta_0 + \beta_1 x + \sum_{k=1}^{K}\beta_{1+k}(x-\kappa_k)_+^r$, where the choices of knots (κ_K's) and K are mentioned in Section 1.

3. Compute the bootstrap hybrid estimator $\hat{V}_{ps}$ based on 250 bootstrap samples drawn from the sample obtained in step 2.

4. The steps 1-3 are repeated $R = 500$ times. Let $\hat{F}_{ps}^{r}, \hat{V}_{A,1}^{r}, \hat{V}_{A,2}^{r}$ and $\hat{V}_{ps}^{r}$ denote the values of $\hat{F}_{ps}, \hat{V}_{A,1}, \hat{V}_{A,2}$ and $\hat{V}_{ps}$ respectively at r-th repetition, $r = 1,..., R$, $R = 200$.

For comparing the performances of the variance estimators' we use the standard criteria viz., relative stability (*RST*) and ratio of standard errors (*RSE*). For any generic estimator v, $RST(v)$ (Wu & Sitter; 2001) and $RSE(v)$ (Breidt et al., 2005) are defined as $RST(v) = [MSE(v)]^{1/2} / MSE^{1/2}$ and $RSE(v) = (\bar{v} / MSE)^{1/2}$, where $MSE(v) = R^{-1}\sum_{r=1}^{R}(v_r - \bar{v})^2$, $\bar{v} = R^{-1}\sum_{r=1}^{R} v_r$, $MSE = R^{-1}\sum_{r=1}^{R}(\hat{F}_{ps}^{r} - q)^2$ and v_r is the value of v at the r-th iteration. On inspection of the Table 2 the following patterns emerge:

(*i*) With regard to *RST*, hybrid estimate appears to be the most stable although the stability depends on the simulation settings viz., super population model considered, the quantile x_q at which the FPDF estimated and the sample size. For the analytical estimates the choice of kernel does not seem to have any significant impact.

(*ii*) With regard to *RSE*, it is interesting to notice that the analytical variance estimates are always giving overestimates of the true variance. On the other hand, most of the time the hybrid estimate is close to the *MSE*. Like *RST*, its value also depends on the simulation settings considered and choice of the kernel does not seem to be important.

3.2. *Empirical study*

To study the performances of eleven estimators of variance estimators in a design-based simulation study we consider a fixed finite population of 140 families living in United Kingdom. For each family, food expenses (y) and family income (*x*) (Zhang, 2004) are known. Note that the auxiliary information on x is available for all the population units. We repeatedly draw random samples of size 35 from it and estimate the variance estimators $\hat{V}_{A,1}$, $\hat{V}_{A,2}$ and $\hat{V}_{ps}$ for food expenses. We study the performances of the bootstrap hybrid variance estimator along withthe analytical estimators under simple random sampling (SRSWOR) in this design-based set-up. The plan of the simulation experiment is exactly same as above except step 1. Here repeated sampling is performed from a fixed finite

population. To compare the performances we evaluate *RST* and *RSE* of the estimators in three quantiles ($q = 0.25, 0.50$ and 0.75) based on $R = 500$ samples. Results are shown in Table 3. The patterns are very similar to that obtained in model-based simulation studies.

4. DISCUSSIONS

The proposed hybrid methodology based on bootstrapping is starightforoward and can be appliled to any finite population parameter which is either exactly or approximately takes form $N^{-1}\sum_U H(y_i)$, $H(.)$ being a nonlinear function may depend on auxiliary variable. It is interesting to extend our methodology for other nonlinear parameters, viz., Gini index, low income proportion etc. (Goga et al., 2013).

Since non-sample observations can be considered as missing by design, our methodology resembles the bootstrap methodology for missing data when missingness mechanism is missing at random (Efron, 1994)., Hence we actually adopt Efron (1994)'s "nonparametric bootstrap" and it is advantageously does not depends on the missingness mechanism. So from Efron's argument, resampling from the design sample doesn't depend on the sampling design as long as it is ignorable. This facilitates our hybrid methodology to extend to more complex sampling designs, viz., probability proportional to size sampling. We need to devbelop the underelying theory in details and study the design effects. Notice that the conventional two-stage bootstrap (Lombardia et al., 2003; 2004) is actually the "full-mechanism bootstrap" (Efron, 1994) and thus operatinally does not provide the estimate conditional variance considered in model-based apparoch with non-ignirable designs. Obviously, for designs that are not ignorable, two-stage bootstrap has to be executed.

Note: For proofs of the results please contact the authors.

REFERENCES

ADHYA, S., BANERJEE, T., and CHATTOPADHYAY, G. (2011). Inference on polychotomous responses in finite populations. *Scand. J. of Staist*. 38, 788-800.

ADHYA, S., BANERJEE, T., and CHATTOPADHYAY, G. (2012). Inference on finite population categorical response: nonparametric regression-based predictive approach. *AStA Advs. in Statist. Analysis*. 96, 69-98.

BREIDT, F. J., OPSOMER, J. D., and CLAESKENS, G. (2005). Model-assisted estimation of complex surveys using penalised splines. *Biometrika*. 92, 831-846.

BRUNMACK, B. A., RUPPERT, D., and WAND, M. P. (1999). Comment on "Variable selection and function estimation in additive nonparametric regression using a data-based prior" by Silvey, T.S., Kohn, R., and Wood, S., *J. of the Amer. Statist. Assoc*. 94, 794-997.

CHAMBERS, R. L., and DUNSTAN, R. (1986). Estimating distribution functions from survey data. *Biometrika*. 73, 597-604.

CHAMBERS, R. L., DORFMAN, A. H., and HALL, P. (1992). Properties of estimators of the finite population distribution function. *Biometrika*. 79, 577-582.

DORFMAN, A. H., and HALL, P. (1993). Estimators of finite population distribution function using nonparametric regression. *Ann. of Statist.* 21, 1452-1475.

EFRON, B. (1994). Missing data, imputation and the bootstrap. *J. of Amer. Statist. Assoc*. 89, 463-475.

EILERS, P. H. C., and MARX, B. D. (1996). Flexible smoothing with B-splines and penalties (with discussions). *Statist. Sci*. 11, 89-121.

FULLER, W. (2009). *Sampling Statistics*. Wiley, New Jersey.

GOGA, C. and RUIZ-GAZEN, A. (2014). Efficient estimation of non-linear finite population parameters by non-parametrics. *J. of Roy. Statist. Soc*. 76, 113-140.

JANSEEN, P., and VERAVERBEKE, N. (1992). Bootstrapping U-statistics with estimated parameters. *Commun. in Statist.-Theor. and Meth*. 21, 1585-1603

LOMBARDIA, M. J., GONZALEZ-MANTEGIA, W., and PRADA-SANCHEZ, J. M. (2003). Bootstrapping the Chamberts- Dunstan estimate of a finite population distribution function. *J. of. Statist. Plann. and Inf.* 106, 367-388.

LOMBARDIA, M. J., GONZALEZ-MANTEGIA, W., and PRADA-SANCHEZ, J. M. (2004). Bootstrapping Dorfman-Hall-Chambers-Dunstan estimator of finite population distribution function. *J. of Nonpar. Statist*. 16, 63-90.

OOSOMER, J. D. (2009). Introduction to part 4. *Handbook of Statistics*, Vol. 29B : *Sample surveys*: *Inference and analysis*, *Edited by Rao, C.R., and Pfeffermann, D.*, North Holland, Amsterdam, 1-9.

RANDLES, R. H. (1982). On the asymptotic normality of statistics with estimated parameters. *Ann .of Statist.* 10, 462-474.

RUPPERT, D., WAND, M.P. and CAROLL, R. J. (2003). *Semiparametric regression*. Cambridge University Press, New York.

SPEED, T. (1991). Comment to Robinson's paper on "That BLUP is a good thing: the estimation of random effects" by Robinson, G.K., *Statist. Sci*. 6, 42-44.

WAND, M. P. (2003). Smoothing and mixed model. *Comp. Statist*. 18, 223-249.

WU, C., and SITTER, R. R. (2001). Variance estimation for the finite population distribution function with complete auxiliary information. *Canad. J. of Statist.* 29, 289-307.

VALLIANT, R. (2009). Model-based prediction of finite population totals. *Sample surveys: Inference and analysis*, *Handbook of Statistics*, *Vol. 29B*: *Sample surveys*: *Inference and analysis*, *Edited by Rao, C.R.*, and *Pfeffermann D.*, North-Holland, Amsterdam, Chapter 23, 11-31.

Table 1. *Choices of the regression function* $m(.)$

Linear	$m(x) = 1 + 2(x - 0.5)$
Exponential	$m(x) = \exp(-8x)$
Bump	$m(x) = 1 + 2(x - 0.5) + \exp(-200(x - 0.5)^2)$
Jump	$m(x) = (0.35 + 2(x - 0.5))I_{\{x \le 0.65\}} + 0.65 I_{\{x > 0.65\}}$
Quadratic	$m(x) = 1 + 2(x - 0.5)^2$
Cycle	$m(x) = 2 + \mathrm{Sin}(2\pi x)$

Table 2. *Relative stabilities and ratio of standard errors of the three variance estimators for six possible population models with population size N=1000 and f=0.10.*

Model	Quantiles	*Relative stability*			*Ratio of standard errors*		
		$\hat{V}_{A,1}$	$\hat{V}_{A,2}$	$\hat{V}_{ps}$	$\hat{V}_{A,1}$	$\hat{V}_{A,2}$	$\hat{V}_{ps}$
Linear	0.25	3.1545	3.1543	0.4566	2.0284	2.0283	1.1657
	0.50	2.5165	2.5164	1.0091	1.8642	1.8641	1.3984
	0.75	4.5182	4.5179	1.1873	2.3373	2.3372	0.9683
Exponential	0.25	1.6887	1.6857	0.2654	1.6323	1.6319	1.0681
	0.50	0.9085	0.9079	0.2564	1.3679	1.3675	1.0487
	0.75	0.9491	0.9803	0.3820	1.3416	1.3418	1.0286
Bump	0.25	0.8625	0.8609	0.5219	1.3329	1.3325	1.0182
	0.50	0.5978	0.6006	0.4249	1.2301	1.2303	1.0018
	0.75	0.7156	0.7168	0.6299	1.2765	1.2766	0.9821
Jump	0.25	2.4857	2.4911	0.4275	1.8346	1.8359	1.0815
	0.50	0.1989	0.2032	0.2033	1.1269	1.1258	1.0059
	0.75	0.2204	0.2194	0.5277	0.9943	0.9915	0.7471
Quadratic	0.25	0.9259	0.9231	0.1887	1.3799	1.3785	0.9735
	0.50	0.8024	0.8021	0.2702	1.3268	1.3263	0.9311
	0.75	1.0354	1.0364	0.3301	1.4131	1.4133	0.8805
Cyclical	0.25	0.9735	0.9729	0.4196	1.3870	1.3868	1.0104
	0.50	2.9604	2.9608	0.6904	1.9607	1.9606	1.0838
	0.75	1.5130	1.5126	0.6148	1.5639	1.5638	1.0206

Table 3. *Relative stabilities (RST) and relative standard errors (RSE) of three variance estimators and their attained coverage's for the confidence intervals* based on "*Food Consumption Data" for three quartiles of FPDF.*

Estimator	*Quantiles*	*Relative stability*	*Ratio of standard errors*
$\hat{V}_{A,1}$	0.25	1.8693	1.6567
	0.50	1.2921	1.4898
	0.75	0.7815	1.3017
$\hat{V}_{A,2}$	0.25	1.8719	1.6489
	0.50	1.2894	1.4875
	0.75	0.7821	1.3007
$\hat{V}_{ps}$	0.25	0.4977	1.0983
	0.50	0.7052	1.1621
	0.75	0.5913	1.1239