International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Application of Auxiliary Variables in Two-Step Semi-Parametric Multiple Imputation Procedure in Estimation of Population Mean

Onyango O Ronald[1], Christopher Ouma Onyango[2]

[1, 2]Department of Statistics and Actuarial Science, Kenyatta University, P.o.Box 43844-0100, Nairobi City

Abstract: ***Multiple imputation procedure is used in handling of item non-response. The imputation procedure is affected by model misspecification and leads to loss in efficiency and biased results. The inclusion of auxiliary variables in the sampling design helps to avoid sensitivity of inference to model misspecification and improves the precision of estimate of population mean. The main aim of this study is to incorporate auxiliary variables in the multiple imputations to improve the accuracy of the values imputed and the efficiency of point estimators. The two-step semi-parametric multiple imputation procedure is considered and modified to incorporate the auxiliary variables. The two-step semi-parametric multiple imputation procedure accounts for unequal probabilities of selection and reduces misspecification in the imputation model. In the first step a non-parametric model is used to generate a posterior predictive model that includes both item level missingness and auxiliary information. The size variables in a sample are replicated using a constrained Bayesian Bootstrap. A constrained Weighted Finite Population Bayesian Bootstrap is then used to create a population of size variables which is considered to be the value of an auxiliary variable that is closely associated with the survey variable. The imputed size variables are then used in a linear regression model to predict the survey variables for the synthetic population. A parametric model is used to impute the missing data on the survey variables in the second step. A simulation study was conducted using single stage probability-proportionate-to-size without replacement sampling design to compare the asymptotic properties of the estimator of the population mean to those obtained using the existing two-step semi-parametric multiple imputation procedure. The proposed procedure reduced bias and resulted in gain in efficiency. The 95% confidence interval coverage rates of the proposed estimator are close to nominal level when the sample size is small.***

Keywords: Multiple imputations, Bayesian bootstrap, weighted finite population Bayesian bootstrap

1. Introduction

Item non-response results when some of the individuals in the sample refuse to give responses to particular questions in a study. According to Rubin (1987), item non-response is handled by use of Multiple Imputation. The Bayesian bootstrap was suggested by Rubin in 1981.It is a development of Efron's (1979) Bootstrap. Bootstrap involves estimation of parameters by simulating the sampling distribution. In Bayesian Bootstrap, the posterior distribution of a parameter is simulated. Lo (1988) assumed a simple random sampling design and developed the Finite Population Bayesian Bootstrap (FPBB) which is based on a sampling scheme that is equivalent to polya urn. The polya posterior is closely related to the Bayesian bootstrap and the finite population Bayesian bootstrap. According to Lazar et al (2008), the polya posterior is a non-informative Bayesian approach to finite population sampling. It involves assigning a prior distribution to the population statistics using the Bayes theorem.

Auxiliary variables exist within the original data set. They are not included in the analysis but they are related to the variable of interest. They help in maintenance of the missing at random (MAR) condition. Little et al (2013) argues that, auxiliary variables are related to the probability of missingness in a variable or to the incomplete variable itself. Incorporation of auxiliary variables in incomplete data analysis takes into account the condition of missingness.

According to Strief et al (2014), the auxiliary variables can be used to create a sampling design and also in making inferences. Lazar et al (2008) proposed a constrained polya posterior to be used when there are prior information about population quantile and means of the auxiliary variables. Strief and Meeden (2014) incorporated weights which depended on auxiliary variables in the constrained polya posterior. According to Meeden (2008), inference on finite population can be obtained by incorporating auxiliary variables in the non-informative Bayesian model.

Zangeneh et al (2011) proposed a Bayesian non-parametric imputation procedure to be used in estimation of the population quantities in absence of design information on non-sampled units. A Dirichlet Process Mixture Normal (DPMN) was used in the imputation of the non-sampled units and a Bayesian Penalized spline model was used in prediction of the survey outcome variables. It was observed that using the imputed size variables in prediction of the survey outcomes results to significant gain in efficiency. The weighted FPBB was developed by Cohen (1997) and used by Zhou et al (2016) to account for item non-response. According to Zhou et al (2016) the new procedure accounted for unequal selection probabilities and reduced bias although with little loss in efficiency.

This research incorporates the auxiliary variables in the two-step semi- parametric multiple imputation procedure so as to improve efficiency and reduce the biasedness in estimation of mean. The inclusion of the auxiliary variables in the weighted FPBB is done under the assumption of MAR. Since under MAR, the condition of missingness is independent of unobserved data. The weighted FPBB model is modified so as to adjust for PPS selection by incorporating size variables and applying it in prediction of the non-sampled sizes. A model is then used to predict the survey

outcome variables for the obtained non-sampled sizes. The Constrained Bayesian Bootstrap procedure of Zangeneh et al (2011) is then used to modify the weighted FPBB of Zhou et al (2016). This results to a constrained Weighted FPBB.

This article is organized as follows. In section 2 the proposed estimator of population mean is discussed in detail. In Section 2.1 a posterior predictive distribution that incorporates auxiliary variables is developed. A constrained weighted Finite Population Bayesian Bootstrap is described in section 2.2.The asymptotic properties of the proposed estimator of population mean is discussed in section 2.3.The empirical study is described in section 3 and finally the conclusion discussed in section 4.

2. Proposed Estimator of Population Mean

2.1 Posterior predictive distribution that incorporates auxiliary variables

In this study a Bayesian population model that makes use of prior information available about auxiliary variables is developed. A Bayesian model in which the prior distribution for the parameter θ is specified together with a distribution for the population value conditional on θ say $P(Y/\theta)$ is considered. The posterior distribution of the non-sampled data is determined by conditioning on the sampled data. It is assumed that auxiliary variables are observed for every unit in the population. Consider a population with N units in which Y is the survey outcome variable, $\tilde{X}$ is the size variable, W is the weight, I is the sampling indicator. In this case I=1 if a unit is sampled and 0 otherwise. If a sample is selected, the units in the population can be divided into sampled and non-sampled thus Y_s, Y_{ns} , $\tilde{X}_s$, $\tilde{X}_{ns}$, W_s and W_{ns} according to the sampling indicator. Also the population is divided into the observed and missing units according to the response indicator $R = (R_s, R_{ns})$. Where R_s, the response indicator for the units observed in the sample and R_{ns} is the response indicator for the non-sampled units.

The sampled size measures are divided into those observed and those missing
$\tilde{X}_s = (\tilde{X}_{s,obs}, \tilde{X}_{s,miss})$, according to their correspondence to the survey outcome variables
, $Y_s = (Y_{s,obs}, Y_{s,miss})$. The non-sampled size measures are also divided into observed and missing, $\tilde{X}_{ns} = (\tilde{X}_{ns,obs}, \tilde{X}_{ns,miss})$.The non-sampled size measures are obtained using a constrained weighted Bayesian bootstrap. The observed and missing information on size variable is pooled together resulting to $\tilde{X}_{obs} = (\tilde{X}_{s,obs}, \tilde{X}_{ns,obs})$ and $\tilde{X}_{miss} = (\tilde{X}_{s,miss}, \tilde{X}_{ns,miss})$. The observed size variable in the entire population is then used in prediction of their associated survey variables via a linear regression model. These results to $\tilde{Y}_{obs} = (Y_{s,obs}, Y_{ns,obs})$ and $Y_{miss} = (Y_{s,miss}, Y_{ns,miss})$.The missing survey outcome variables are obtained by imputation using a parametric model.

The joint distribution of the size variable and survey variable is given by

$P(\tilde{X}, Y) = P(\tilde{X})P(Y/\tilde{X})$. The posterior predictive distribution of θ given the observed sampled size variable is $P(\theta/\tilde{X}_s)$. The posterior predictive distribution of θ is obtained by averaging the complete data on posterior of θ over the posterior predictive distribution of the missing sizes,

$$P(\theta/\tilde{X}_{s,obs}, R_s, I) = \int P(\theta/\tilde{X}_s, R, I)P(\tilde{X}_{s,miss}/\tilde{X}_{s,obs}, R_s, I)$$

$$P(\theta/\tilde{X}_{s,obs}, w_s) = P(\tilde{X}_{s,obs}, w_s/\theta)P(\theta) \quad (1)$$

The posterior predictive distribution of the non-sampled sizes given the sampled sizes is $P(\tilde{X}_{ns}/\tilde{X}_s) \alpha \int P(\tilde{X}_{ns}/\theta, \tilde{X}_s)p(\theta/\tilde{X}_s)$

$$P(\tilde{X}_{ns}/\tilde{X}_s) = \int P(\tilde{X}_{ns}/\theta, \tilde{X}_s)p(\theta/\tilde{X}_s)d\theta \quad (2)$$

The information on missing sizes and the associated survey outcome variables is incorporated into (2) and results to

$$P(\tilde{X}_{ns,obs}/\tilde{X}_s, w_s) = \int P(\tilde{X}_{ns,obs}, \tilde{X}_{miss}, Y_{miss}/\tilde{X}_{s,obs}, w_s)dY_{miss} \quad (3)$$

The non-sampled data is generated together with the missing and the observed information on both the size variable and the survey variable using the constrained weighted FPBB. Since $\tilde{X}_{miss}$ is associated with Y_{miss} the integration is done over Y_{miss}.Thus,

$$\int P(\tilde{X}_{ns,obs}, Y_{miss}/\tilde{X}_{s,obs}, Y_{s,obs}, w_s)dY_{miss} = \int P(Y_{miss}/Ys,obs,Xobs,ws)P(Xns,obs/Xs,obs,ws)dYmiss$$

(4)

Equation (4) is parameterized and then integrated with respect to the posterior distribution of θ resulting to

$$\int P(\tilde{X}_{ns,obs}, Y_{miss}/Y_{s,obs}, \tilde{X}_{s,obs}, w_s)dY_{miss} \alpha \int\int P(Y_{miss}/Ys,obs,Xobs,ws,\theta)P(Xns,obs/Xs,obs,ws,\theta)P(Ys,obs,Xs,obs,ws/\theta)P(\theta)\mathrm{d}\theta\mathrm{d}Ymiss \quad (5)$$

$$\int P(\tilde{X}_{ns,obs}, Y_{miss}/\tilde{X}_{s,obs}, Y_{s,obs}, w_s)dY_{miss} = \int\int P(Y_{miss}/Ys,obs,Xobs,ws,\theta)P(Xns,obs/Xs,obs,ws,\theta)P(\theta/Xobs,Yobs,ws,)\mathrm{d}\theta\mathrm{d}Ymiss \quad (6)$$

The simulation of equation (6) is done using the Gibbs sampler by iterating between the draws of $P(\theta/\tilde{X}_{obs}, Y_{obs}, w_s, Y_{miss}) = P(\theta/\tilde{X}, Y, w_s) = P(\theta/\tilde{X}, Y)$, and those of $P(Y_{miss}/\tilde{X}_{obs}, Y_{s,obs}w_s, \theta)$. The draws of the non-sampled size variables are obtained using the constrained weighted FPBB which undoes the sampling deign (weights). This makes it possible to draw units from the posterior predictive distribution that is free from the effect of unequal probability selection.

2.2 The constrained weighted FPBB

Consider a sample of size n which consists of both the survey outcome variable and its associated auxiliary variable, $(y_i, \tilde{x}_i)$ where i=1, 2… n. It is assumed that the auxiliary variables are available for all the sampled units and their corresponding total is known, $X = \sum_{i=1}^{n} \tilde{x}_i$. Also the population mean of the auxiliary variable is known since it is assumed that there exist a linear relationship between the variable of interest and the known auxiliary characteristic.

In PPS sampling the units are selected with probability proportional to the value of $\tilde{X}$ which is the size measure. The size measure is only reported for the sampled units. The number of non-sampled units (N-n) is known but their sizes are unknown hence the PPS sampling scheme is informative.

This implies that the sizes need to be adjusted due to the effect of selection. This is achieved by incorporation of the size measure in the weighted FPBB. The constrained Weighted FPBB is obtained by assuming a multinomial distribution for the observed sampled size variables and Dirichlet posterior distribution for the parameter ϕ.Thus the posterior predictive distribution of the non-sampled counts in the population is obtained by considering both the multinomial and Dirichlet distribution.

Denote the sample by $(Y_s, \tilde{X}_s, w_s, R_s) = \{(y_i, \tilde{x}_i, w_i, R_i), i = 1, \dots\dots, n\}$, where w_i is the weight attached to i^{th} unit in the sample, Y_s is the survey variable, $\tilde{X}_s$ is the size measure and R_s is the response indicator. Let $(\tilde{x}_1, \dots\dots \tilde{x}_k)$ be a set of distinct sizes for the sampled units and $\phi = \{\phi_1, \phi_2, \dots\dots \phi_k\}$ be a vector of probabilities. The sampled units are assumed to contain different vectors of values such that n=k. This implies that $P\left((y_i, \tilde{x}_i, w_i, R_i) = \tilde{x_j}/\phi\right) = \phi_j$. Let n_j be the number of sampled units with distinct sizes $\tilde{x}_j$ where $\sum_{j=1}^{k} n_j = n \quad and\ j = 1,2, \dots\dots, k$. In this case $\sum_{j=1}^{k} \tilde{x}_j = X$, where X is the total sum of all the size measures in the sample. The distribution of the sampled counts in the population is obtained by assuming multinomial distributions for the weighted count. Thus $\tilde{X}_i/\phi \sim multi(n: \phi_1, \dots\dots \phi_k)$ hence

$$P(w_1, \dots\dots\dots, w_k/\phi) \alpha \prod_{j=1}^{k} \phi_j^{w_j} \qquad (7)$$

.Where, $w_j = \frac{1}{\pi_j}$, and, $\pi_j = \frac{n\tilde{x}_j}{N\bar{X}}$.

According to Zhou et al (2016), to obtain a posterior distribution for the parameter ϕ both a multinomial distribution and a Halden prior of ϕ: $\phi \sim Dir(0, \dots\dots, 0)$ are considered. This results to a Dirichlet posterior distribution given by $\phi/w_1, \dots\dots w_k \sim Dir(w_1, \dots\dots w_k)$. Thus

$$P(\phi/w_1, \dots\dots, w_k) \alpha \prod_{j=1}^{k} \phi_j^{w_j - 1} \qquad (8)$$

To create draws for the non-sampled sizes using the constrained weighted FPBB consider the following. Let n_j' be the number of non-sampled counts with size measure $\tilde{x_j}$.This implies that $\sum_{i=1}^{k} n_j' = N - n$ .The counts of the non-sampled sizes have a posterior predictive distribution which follows a multinomial distribution with a sample size of N-n and probabilities $(\phi_1^* \dots\dots\dots \phi_K^*)$, where $\phi_j^* = \frac{c.\phi_j\,(1-\pi_j)}{\pi_j}$ .c is a constant that normalizes the expression. Thus the posterior predictive distribution of the non-sampled values in the population is given by

$$P(\tilde{X}_{ns}/\tilde{X}, \phi) = \frac{\int_0^1 \dots\dots \int_0^1 P(\tilde{X}_{ns}/\tilde{X}, \phi) P(\tilde{X}/\phi) P(\phi) d\phi_1 \dots\dots d\phi_k}{\int_0^1 \dots\dots \int_0^1 P(\tilde{X}/\phi) P(\phi) d\phi_1 \dots\dots d\phi_k}$$

$$P(\tilde{X}_{ns}/\tilde{X}, \phi) = \frac{\int_0^1 \dots\dots \int_0^1 \sum_{j=1}^{k} P(n_1', \dots\dots n_k'/w_1, \dots\dots w_k, \phi) P(w_1, \dots\dots w_k/\phi) P(\phi) d\phi_1 \dots\dots d\phi_k}{\int_0^1 \dots\dots \int_0^1 \sum_{j=1}^{k} P(w_1, \dots\dots w_k/\phi) P(\phi) d\phi_1 \dots\dots d\phi_k}$$

$$P(\tilde{X}_{ns}/\tilde{X}, \phi) = \frac{\int_0^1 \dots\dots \int_0^1 \prod_{j=1}^{k} (\phi_j^*)^{n_j'} \prod_{j=1}^{k} (\phi_j^*)^{w_j - 1} d\phi_1 \dots\dots d\phi_k}{\int_0^1 \dots\dots \int_0^1 \prod_{j=1}^{k} (\phi_j^*)^{w_j - 1} d\phi_1 \dots\dots d\phi_k}$$

$$P(n_1', \dots\dots, n_k'/w_1, \dots\dots, w_k) = \frac{\int_0^1 \dots \int_0^1 \prod_{j=1}^{k-1} (\phi_j^*)^{w_j + n_j' - 1} (1 - \sum_{j=1}^{k-1} \phi_j^*)^{w_k + n_k' - 1} d\phi_1^*, \dots, d\phi_{k-1}^*}{\int_0^1 \dots \int_0^1 \prod_{j=1}^{k-1} (\phi_j^*)^{w_j - 1} (1 - \sum_{j=1}^{k-1} \phi_j^*)^{w_k - 1} d\phi_1^*, \dots, d\phi_{k-1}^*} \qquad (9)$$

In this study the adapted version of the constrained FPBB is in line with that of Zhou et al (2016). It is carried out in two stages. The first stage involves replicating the initial sample using a constrained Bayesian bootstrap to generate L replicate for the size measure. This involves drawing the posterior distribution of the parameter vector ϕ conditional on the counts in the sampled data, $(n_1, \dots\dots, n_k)$.Thus $(\phi^{(l)}/\tilde{X}_s w_s R_s) \sim Dir(n_1, \dots\dots, n_k)$, where $\phi^{(l)} = (\phi_1^{(l)}, \dots\dots, \phi_k^{(l)})$.This generates the posterior joint distribution of all the size variables in the population given the observed values in the sample. These results to $\{(\tilde{X}_s^l, w_s^l, R_s^l),\ l = 1,2, \dots\dots, L\}$. The number of times the size measure is picked from the l^{th} replicate is denoted by $h_l(i)$, this is considered in calculation of the i^{th} bootstrap weight for the size measure. Thus $w^{(l)} = w.h_l(i)$.This weights are used in the second stage.

The second stage involves undoing the sampling weights using a constrained weighted FPBB. This is achieved by creating a B unweighted population of size measures for each of the L replicates. This yields the predicted counts of the non-sampled sizes for the replicated samples. The constrained weighted FPBB (9) is approximated using the procedure proposed by Cohen (1997).The procedure involves selecting a polya sample of size measures $(\tilde{X}_{ns}^{(lb)}, R_{ns}^{(lb)})$ of size N-n from the urn $(\tilde{X}_s^{(l)}, R_s^l)$ of size n. The selection of i^{th} size measure $(\tilde{X}_i^{(l)}, w_i^{(l)}, R_i^{(l)})$ is done using the probability $\phi^{(l)*} = \frac{w_i^l - 1 + l_{i,k-1} \times \left(\frac{N-n}{n}\right)}{N - n + (k-1) \times \left(\frac{N-n}{n}\right)}$ (10)

, where $k = 1,2, \dots, N - n + 1$ and $i = 1,2, \dots, n$.In this case $w_i^{(l)}$ is the bootstrap weight for i^{th} size measure in the l^{th} replicate of the constrained Bayesian Bootstrap sample and $l_{i,k-1}$ is the number of selection of unit i such that when $k = 0, l_{i,0} = 0$. This results into a constrained weighted FPBB population of size N denoted by, $P_{(b),obs}^{(l)} = \left\{(\tilde{X}_s^{(l)}, R_s^{(l)})(\tilde{X}_{ns}^{(lb)}, R_{ns}^{(lb)})\right\}, where\ b = 1, \dots\dots, B\ and\ l = 1,,2, \dots, L$.Thus the unweighted population of size variables is given by, $P_b^{(l)} = (P_{(b),obs}^{(l)}, \tilde{X}_{miss}^{(lb)})$. Population of size variables is then used in a linear regression model to predict the survey outcome variables. This results to $P_{(b)}^{(l)} = (P_{(b),obs}^{(l)}, Y_{miss}^{(lb)})$, where $P_{(b),obs}^{(l)} = \left\{(Y_{s,obs}^{(l)}, \tilde{X}_s^{(l)}, R_s^{(l)})(Y_{ns,obs}^{(lb)}, X_{ns}^{(lb)}, R_{ns}^{(lb)})\right\}$ is the observed data and, $Y_{miss}^{(lb)} = (Y_{s,miss}^{(l)}, Y_{ns,miss}^{(lb)})$, is the missing data. The missing data on survey variables are obtained by imputation using a parametric model. The undoing of the sampling design makes it possible to conduct multiple imputations under the assumption of iid. The draws of the

missing data are obtained from the posterior predictive distribution $(Y_{miss}^{(lb)}/P_{(b),obs}^{l})$.

This step yields M imputed data sets for each of the unweighted population generated above $P_{Bm}^{(l)} = (P_{(B1)}^{(l)}, P_{(B2)}^{(l)}, \dots\dots\dots, P_{(BM)}^{(l)})$. The resulting population of survey outcome variables is denoted $P^{comp} = (P_{(11)}^{(l)}, \dots\dots\dots, P_{(1M)}^{(l)}, \dots\dots\dots., P_{(B1)}^{(l)}, \dots\dots, P_{(BM)}^{(l)}, \dots\dots\dots, P_{(BM)}^{(L)})$

2.3 Properties of the proposed estimator of population mean

The missing data was first imputed using the weighted FPBB and resulted into B unweighted population of survey outcome variables. The resulting population was then imputed using a parametric model. This resulted into M imputed data sets for each of the B unweighted population, $P^{comp} = (P_{(11)}^{(l)}, \dots\dots\dots, P_{(1M)}^{(l)}, \dots\dots\dots., P_{(B1)}^{(l)}, \dots\dots, P_{(BM)}^{(l)}, \dots\dots\dots, P_{(BM)}^{(L)})$, where l=1,2,……,L , b=1,2,…..,B and m=1,2,……,M. The population statistic is estimated by $q^{(lbm)}$ which is obtained from the m^{th} imputation of b^{th} unweighted population within the l^{th} Bayesian Bootstrap sample. For complete data set the mean is given by $\bar{Q}_L = \frac{1}{L}\sum_{l=1}^{L}\tilde{Q}^{(l)}$ where $\tilde{Q}^{(l)} = \lim_{B\to\infty,M\to\infty}\frac{1}{BM}\sum_{b=1}^{B}\sum_{m=1}^{M}q^{(lbm)}$. In this case $\tilde{Q}^{(l)}$ is estimated by $\hat{Q}^{(l)} = \frac{1}{BM}\sum_{b=1}^{B}\sum_{m=1}^{M}q^{(lbm)}$, which is the point estimate of the proposed estimator. The imputation model is correctly specified if $E(q^{lbm}) = Q$.

$$\hat{Q}^{(l)} = \frac{1}{BM}\sum_{b=1}^{B}\sum_{m=1}^{M}q^{(lbm)}$$

$$\hat{Q}^{(l)} = \frac{1}{BM}\sum_{b=1}^{B}\sum_{m=1}^{M}E\,(q^{(lbm)})$$

But, $E(q^{lbm}) = Q$

$$E\big(\hat{Q}^{(l)}\big) = \frac{1}{BM}(BMQ)$$

$$E\big(\hat{Q}^{(l)}\big) = Q$$

$$\bar{Q}_L = \frac{1}{L}\sum_{l=1}^{L}\hat{Q}^{(l)}$$

$$\mathrm{E}\,(\bar{Q}_L) = \frac{1}{L}(LQ)$$

$$\mathrm{E}\,(\bar{Q}_L) = Q$$

This shows that the proposed estimator of population mean is unbiased. The variance of the proposed estimator of population mean is $V_L = \frac{1}{L-1}\sum_{l=1}^{L}(\hat{Q}^{(l)} - \bar{Q}_L)^2$

The mean squared error (MSE) is given by, $\mathrm{MSE} = \mathrm{E}(\mathrm{q}^{(\mathrm{lbm})} - \mathrm{Q})^2$

$$MSE = \frac{1}{L-1}\sum_{l=1}^{L}(\mathrm{q}^{(\mathrm{lbm})} - Q)^2$$

The coverage rate of the confidence intervals for the population mean is obtained using normal approximation. Given the confidence level (1-α)100%, the confidence interval for l^{th} bootstrap sample is obtained by $prob\left\{-Z_{\frac{\alpha}{2}} < Z < Z_{\frac{\alpha}{2}}\right\} = 1-\alpha$. Where $Z = \frac{\hat{Q}^{(l)} - \bar{Q}_L}{\frac{\sigma}{\sqrt{BM}}}$ and

$$\sigma = \sqrt{\frac{1}{\mathrm{BM}-1}\sum_{\mathrm{b=1}}^{\mathrm{B}}\sum_{\mathrm{m=1}}^{\mathrm{M}}(\mathrm{q}^{(\mathrm{lbm})} - \hat{Q}^{(l)})^2}$$

$$prob\left\{-Z_{\frac{\alpha}{2}} < \frac{\hat{Q}^{(l)} - \bar{Q}_L}{\frac{\sigma}{\sqrt{BM}}} < Z_{\frac{\alpha}{2}}\right\} = 1-\alpha$$

$$prob\left\{-\hat{Q}^{(l)} - \frac{\sigma}{\sqrt{BM}}Z_{\frac{\alpha}{2}} < \bar{Q}_L < -\hat{Q}^{(l)} + \frac{\sigma}{\sqrt{BM}}Z_{\frac{\alpha}{2}}\right\} = 1-\alpha$$

$$prob\left\{\hat{Q}^{(l)} + \frac{\sigma}{\sqrt{BM}}Z_{\frac{\alpha}{2}} > \bar{Q}_L > \hat{Q}^{(l)} - \frac{\sigma}{\sqrt{BM}}Z_{\frac{\alpha}{2}}\right\} = 1-\alpha$$

$$prob\left\{\hat{Q}^{(l)} - \frac{\sigma}{\sqrt{BM}}Z_{\frac{\alpha}{2}} < \bar{Q}_L < \hat{Q}^{(l)} + \frac{\sigma}{\sqrt{BM}}Z_{\frac{\alpha}{2}}\right\} = 1-\alpha$$

Thus the confidence interval is $\left(\hat{Q}^{(l)} - \frac{\sigma}{\sqrt{BM}}Z_{\frac{\alpha}{2}},\ \hat{Q}^{(l)} + \frac{\sigma}{\sqrt{BM}}Z_{\frac{\alpha}{2}}\right)$. The coverage rates are obtained by computation of the proportions of samples for which the population mean is contained in the confidence interval.

3. Empirical Study

3.1 Description of the study population

In the study design the survey outcome variable is assumed to be associated with the size variable, the probability of selection and their interaction. A missing data generation technique that does not depend on design information is considered, MAR-X. A population which consist of the survey outcome variables, size variables and other covariate is generated with the following joint distributions, logZ~N (2, 1), the survey weights Z=exp (logZ), size variables that are closely associated with the survey outcome variable. X ~ N (0.1*logZ, d_x^2) , fully observed covariate information. Y~ N (0.1*X+0.5*logZ+0.6*x*logZ, d_y^2), the survey outcome variables of main interest

3.2 Simulation Study

A simulation study was conducted to assess the proposed estimator of population mean. The simulation design was as follows. In steps one a population of size N=4000 is considered and used to obtain independent samples of size n=200 of size variables using PPSWOR sampling design. Step two involves generation of unweighted population of size variables. Each of the replicated samples is simulated using 'wtpolyap' function in the polyapost package. The number of simulations to be done is denoted as B.

Step three involves prediction of survey outcome variables for the generated unweighted population of size variables. This is done using linear regression function in R package. This results to a complete data of survey variables which is referred to as "Before deletion population". In step four, missing data is created for each of the replicated samples. A probit model is used as a deletion function to create missing data on each of the replicated samples. Both X and Z is assumed to be fully observed. The missing data generation technique is considered in the generation of latent variables for the deletion function. Thus $T_1 = -0.635 + 0.4x + e$, where, $e \sim N(0,1)$ .The survey outcome variable is

considered to be missing if $T_j > 0$.This may be done using the function 'simsem' in R.

In step five missing data is imputed for each of the replicated samples using the 'mice' package in R. This is done using three different models of misspecification; the first model includes size variables, Z. The second model includes both size variables and weights say Z, logZ. The third model includes the interaction of size variables and weights, logZ*Z. Step six involves calculation of mean squared error and the 95% confidence interval coverage rates of the proposed estimator. Mean squared error (MSE) $=\sqrt{(E(q_r) - Q)^2}$ $=\sqrt{\frac{1}{L}\sum_{r=1}^{L}(q_r - Q)^2}$.

The 95% confidence interval coverage rate of the proposed estimator is calculated based on the L replicates. The confidence interval of the estimator of population mean is obtained using $\left(\hat{Q}^{(l)} - 1.96\frac{\sigma}{\sqrt{BM}}\ ,\ \hat{Q}^{(l)} + 1.96\frac{\sigma}{\sqrt{BM}}\right)$.

2.4 Simulation Results

Two critical statistics examined are mean squared error and 95% confidence interval coverage rate. All are calculated based on B=20, M=5 and L=20, 30…, 100. The mean of the generated data is 1.456 and variance is 0.046.The simulation results are as shown below.

Table 2: Summary results of imputation model that includes auxiliary variables only and other covariates only

Imputation model	Sample size	20	30	40	50	60	70	80	90	100
Z only	MSE	0.144	0.104	0.100	0.096	0.087	0.078	0.071	0.057	0.053
	95% CI cov.	1.000	1.000	1.000	0.980	0.917	0.943	0.900	0.900	0.870
X only	MSE	0.145	0.105	0.102	0.098	0.088	0.078	0.066	0.059	0.052
	95% CI cov.	0.950	0.966	0.975	0.980	0.917	0.914	0.913	0.922	0.920

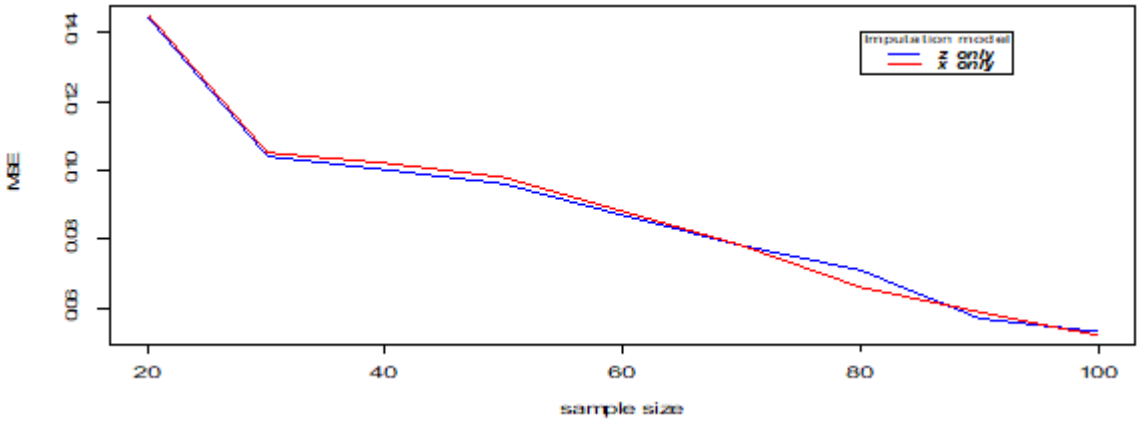

Figure 1: Plots mean squared errors of imputation models that includes Z only and X only

Table 2 displays the summary results of imputation model that includes auxiliary variables, Z only and other covariates, X only. Figure 1 compares this two imputation models in terms of mean squared error. Under Z only imputation model, most of the mean squared errors are observed to be lower than those for X only imputation. The coverage rates decreases with increase in sample size. Under X only imputation model, the coverage rates are closer to the nominal level.

Table 3: Summary results for imputation model that includes weights.

Imputation model	Sample size	20	30	40	50	60	70	80	90	100
Z, logZ	MSE	0.145	0.103	0.099	0.095	0.086	0.077	0.071	0.056	0.054
	95% CI cov.	1.000	1.000	1.000	0.980	0.917	0.943	0.900	0.900	0.880
X, logZ	MSE	0.146	0.110	0.104	0.098	0.088	0.077	0.066	0.058	0.051
	95% CI cov.	0.950	0.966	0.975	0.98	0.917	0.914	0.913	0.922	0.920

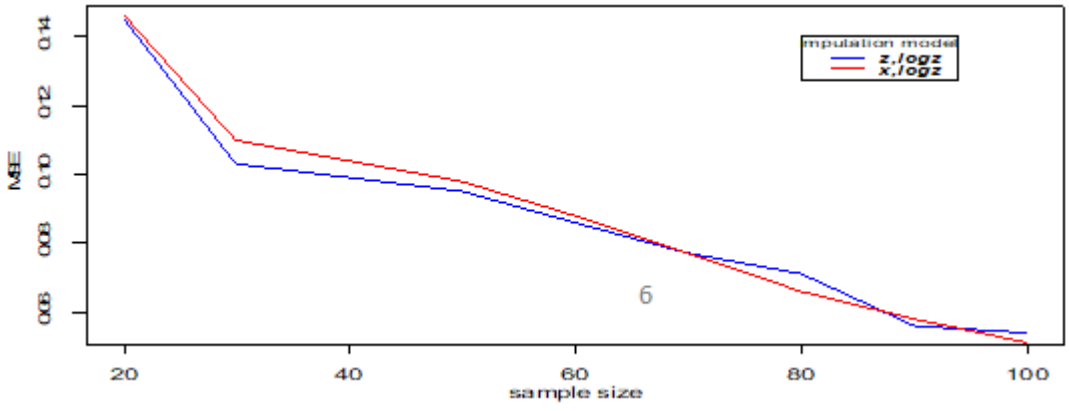

Figure 2: Plots of mean squared errors of imputation model that includes weights

Table 3 displays summary results of imputation model that incorporates survey weights. The mean squared errors are lower than those in table 1.This implies imputation models that includes weights outperforms imputation model that includes X only and Z only. The imputation model that includes both auxiliary variables and weight has coverage rates which are higher than the nominal level when the sample size is small. The imputation model that includes both auxiliary variables and weight outperforms that which includes both weights and other covariates.

Table 4: Summary results for imputation model that includes interaction between weights, auxiliary variables and other covariates

Imputation model	Sample size	20	30	40	50	60	70	80	90	100
z*logz	MSE	0.146	0.100	0.096	0.092	0.085	0.077	0.070	0.056	0.053
	95% CI cov.	1.000	1.000	1.000	0.980	0.917	0.943	0.900	0.900	0.870
x*logz	MSE	0.155	0.110	0.104	0.098	0.088	0.077	0.066	0.059	0.050
	95% CI cov.	0.950	0.966	0.975	0.980	0.917	0.914	0.913	0.922	0.920

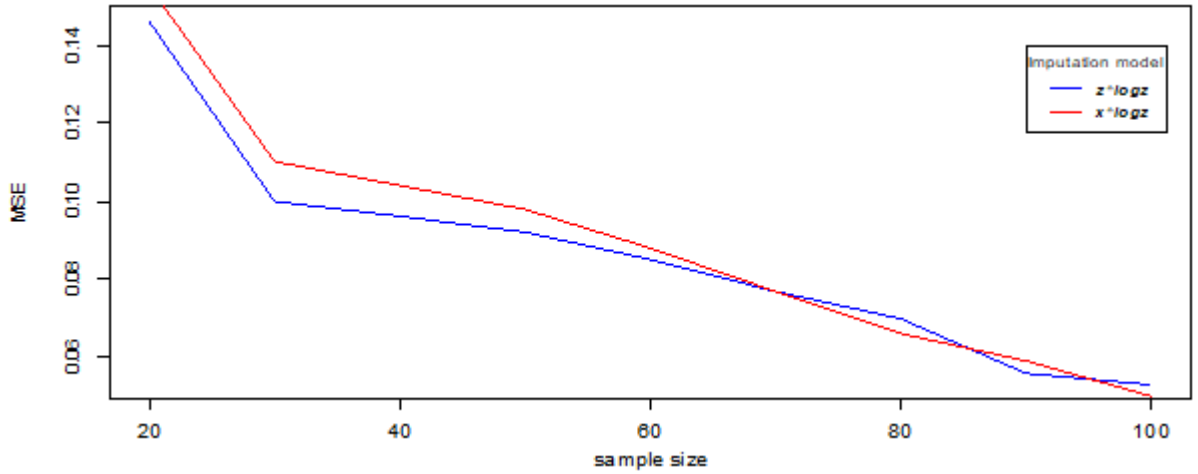

Figure 3: Plots of mean squared errors of imputation models that include interaction between weight, auxiliary variables and other covariates

Table 4 shows summary results of imputation models that include interaction between weights, auxiliary variables and other covariates. The imputation model that includes interaction between auxiliary variable and weights outperforms the one that includes interaction between the weights and other covariates. This is vice-versa when the sample size is large. The imputation model that includes interaction between weights, auxiliary variables and other covariates outperforms the other two imputation models. This implies inclusion of interaction in the imputation model results to unbiased results and coverage rates which are closer to the nominal level.

4. Conclusion

The aim of this study was to propose a two-step semi-parametric multiple imputation procedure that incorporates auxiliary variables and to apply it in estimation of population mean. The results show that inclusion of auxiliary variables in the imputation model results to unbiased estimates and gain in efficiency. The 95% confidence interval coverage rate of the proposed estimator is closer to nominal level when the sample size is small.

References

[1] Cohen, M.P. (1997). The Bayesian Bootstrap and Multiple Imputations for Unequal Probability Sample Designs. *Proceedings of the Research Methods Section, American Statistical Association*.635-638.

[2] Donald, R. (1981). The Bayesian Bootstrap. *Annals of Statistics.9* (1):130-134.

[3] Dong, Q, Elliott, M.R., and Raghunathan, T.E. (2014). A Nonparametric Method to Generate Synthetic Populations to Adjust for Complex Sample Designs. *Survey Methodology.40* (1):29-46.n

[4] Efron, B. (1979). Bootstrap methods. Another look at the jackknife, *Annals of statistics.7* 1-26.

[5] Lazar, R, Meeden, G., and Nelson, D. (2008).A non-informative Bayesian approach to finite population sampling using Auxiliary variables. *Survey methodology.34* (1):51-64.

[6] Lo, A.Y. (1988). A Bayesian Bootstrap for a Finite Population. *The Annals of Statistics.16* (4):1684-1695.Accessed via http://www.jstor.org/stable/2241787

[7] West, B., and Little, R.J. (2013). Non-response Adjustment of Survey Estimates Based on the Auxiliary Variables subject to errors. *Journal of Royal Statistical Society (Applied statistics):62(2)* 213-231.**doi**: 10.1111/j.1467-9876.2012. 01058. X

[8] Strief, J., and Meeden, G. (2014). Objective Stepwise Bayes weights in Survey sampling. *Survey methodology.39* (1):1-27.

[9] Zangeneh, S. Z., Keener, R., and Little, R. J. A. (2011). Bayesian nonparametric estimation of finite population quantities in absence of design information on nonsampled units. Proceedings of the Joint Statistical Meetings.

[10] Zhou, H, Elliott, M, R., and Raghunathan, T. E. (2016).A Two-step Semiparametic method to

accommodate sampling weights in multiple imputations. *Journal of the International Biometric Society.72* (1): 242-252**.** doi**:** 10.1111/biom.12413

[11] Zhou, H, Elliott, M.R and Raghunathan, T.E. (2016).Multiple imputations in two stage cluster samples using Finite population Bayesian Bootstrap. *Journal of survey statistics and method logy.4* (3):139-170.doi: 10.1093/jssam/smv031.