APPLICATION OF AUXILIARY VARIABLES IN TWO-STEP SEMI-PARAMETRIC MULTIPLE IMPUTATION PROCEDURE IN THE ESTIMATION OF POPULATION MEAN

ONYANGO O.RONALD
I56/CE/26065/2014
DEPARTMENT OF STATISTICS AND ACTUARIAL SCIENCE

"A Project submitted in Partial Fulfilment of the Requirements for the Award of the Degree of Master of Science (Statistics) in the School of Pure and Applied Sciences of Kenyatta University”

June 2018
TABLE OF CONTENTS

DECLARATION v

CERTIFICATE OF APPROVAL vi

DEDICATION vii

ACKNOWLEDGEMENT viii

LIST OF FIGURES ix

LIST OF TABLES x

LIST OF ABBREVIATION AND ACRONYMS xi

ABSTRACT xii

1 INTRODUCTION 1

1.1 Background .......................... 1

1.2 Statement of the Problem .................. 4

1.3 Significance of Study ...................... 5

1.4 Research Objectives ........................ 6

1.4.1 The main objective ...................... 6

1.4.2 Specific objectives ........................ 6
2 LITERATURE REVIEW

2.1 Introduction ........................................ 7
2.2 The Posterior Predictive Distribution .................. 9
2.3 Weighted FPBB ......................................... 13
2.4 Estimation of the Population Statistics ................. 19

3 THE PROPOSED ESTIMATOR OF POPULATION MEAN

3.1 Introduction ........................................... 23
3.2 Posterior Predictive distribution that incorporates auxiliary variables . 23
3.3 The Constrained Weighted FPBB ......................... 27
3.4 Properties of the proposed estimator of population mean .............. 33

4 EMPIRICAL STUDY

4.1 Introduction ........................................... 36
4.2 Simulation study ........................................ 36
4.3 Simulation results ........................................ 38
4.4 Conclusion ............................................... 42
4.5 Recommendations ....................................... 43

REFERENCES ........................................... 44

APPENDICES ........................................... 45
Appendix 1: R code for using the proposed methodology
DECLARATION

I declare that this project does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any university. To the best of my knowledge it does not contain any material previously published or written by another person except where due reference has been made in the text.

The editorial assistance provided has in no way added to the substance of my project which is the product of my own research endeavors.

Onyango O.Ronald

Signature...........................................

Date.............................................
CERTIFICATE OF APPROVAL

This is to certify that this project has been submitted in partial fulfillment for the award of the degree of Master of Science (Statistics) of Kenyatta University.

Onyango O. Ronald

I56/CE/26065/2014

Signature..........................................

Date............................................

This project is submitted with the approval of my supervisor

Dr. Christopher Ouma Onyango,

School of Pure and Applied Sciences,

Department of Statistics and Actuarial Sciences,

Kenyatta University.

Signature..................................

Date............................................
DEDICATION

To my loving family, thank you for the continuous support throughout this process.
ACKNOWLEDGEMENT

Special thanks to my supervisor Dr. Christopher Ouma Onyango of Kenyatta University, School of Pure and Applied Sciences. This is for the continuous guidance and support throughout the study period. I will forever be grateful for your ever unending support and beneficial knowledge.

To my family, thank you for your support, you have kept me going along with my academic journey over the years and encouraged me to move forward in my profession. To my friends, thank you for your prayers and encouragement through this journey.

I thank all those who in one way or another played a role in this study.

Finally, I thank the above all God for making everything possible.
LIST OF FIGURES

Figure 1 : Plots of mean squared errors of imputation models that includes
Z only and X only ......................................................... 39

Figure 2 : Plots of mean squared errors of imputation models that includes
weights ................................................................. 40

Figure 3 : Plots of mean squared errors of imputation models that include
interaction between weight, auxiliary variables and other covariates . 41
LIST OF TABLES

1  Simulation results of the synthetic method proposed by [1] . . . . . . . . 19
2  Summary results of imputation model that includes auxiliary variables
    only and other covariates . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 38
3  Summary results for imputation model that includes weights. . . . . . 40
4  Summary results for imputation model that includes interaction between
    weights, auxiliary variables and other covariates. . . . . . . . . . . . . . 41
LIST OF ABBREVIATION AND ACRONYMS

MI- Multiple Imputation
BB- Bayesian Bootstrap
Weighted FPBB - Weighted Finite Population Bayesian Bootstrap
PPSWOR- Probability-Proportionate-to-Size Without replacement
PSU- Primary Sampling Units
MAR- Missing at Random
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 was to incorporate auxiliary variables in the Multiple Imputation to improve the accuracy of the values imputed and the efficiency of point estimators. The two-step semi-parametric multiple imputation procedure was considered and modified to incorporate the auxiliary variables. In the first step a non-parametric model was used to generate a posterior predictive model that includes both item level missingness and auxiliary information. The size variables in a sample were replicated using a Constrained Bayesian Bootstrap. A Constrained Weighted Finite Population Bayesian Bootstrap was then used to create a population of size variables which was considered to be the value of an auxiliary variables that is closely associated with the survey outcome variable. The imputed size variables were then used in a linear regression model to predict the survey outcome variables for the synthetic population. A parametric model was used to impute the missing data on the survey outcome variables in the second step. A simulation study was conducted using single stage probability-proportionate-to-size without replacement sampling design. The asymptotic properties of the estimator of the population mean were compared 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 were close to nominal level when the sample size was small.
1 INTRODUCTION

1.1 Background

Item non-response results when some of the individuals in the sample refuse to give responses to particular questions in a study. According to [2], item non-response is handled by use of Multiple Imputation which has Bayesian conceptualisation. This is because the posterior information is obtained from the observed samples. A finite population is considered and posterior predictive distribution is used to generate the unknown values from the observed data. Multiple imputations creates a multiple dataset which are copies of the initial complete data set. The missing data are then imputed based on the observed values. The imputation model is specified and used to impute the missing values. This imputed data is analyzed using an analysis model. Both imputation and analysis process are repeated over a large number of times. The estimates obtained from the analysis are then combined using results in [2].

Bayesian inference is one of the approaches used on inferences of finite population. In the Bayesian approach the information about population is usually included in the prior distribution. After observing the sample, the inference is usually made based on the posterior distribution of the units observed in the population.

The Bayesian bootstrap was suggested by [3]. It is a development of [4] Bootstrap. Bootstrap involves estimation of parameters by simulating the sampling distribution. In Bayesian Bootstrap, the posterior distribution of parameter is simulated. The FPBB was
developed by [5]. It is based on sampling scheme which is equivalent to polya urn. The polya posterior assumes that little prior information is available about the population. It gives a predictive distribution for the unobserved units given the observed units in a population. The polya posterior is closely related to the Bayesian Bootstrap which was suggested by [3] and also it is related to the weighted FPBB of [5]. According to [6], 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. The prior should be specified properly to avoid misspecification.

A polya posterior, $P(p_1, p_2, \ldots, p_n)$ is approximately a Dirichlet distribution when the ratio of the sample size to population size is small. The parameter vector has all ones hence it is uniform.

According to [6], inference on finite population can be obtained by incorporating auxiliary variables in the non-informative Bayesian model. [6] suggested a constrained polya posterior which is used in the presence of auxiliary variables about population mean and median. The polya posterior was extended to incorporate prior information about auxiliary variables. According to [7], the polya posterior can be adjusted to incorporate prior information such as auxiliary variables. This involves incorporating constants of median of the auxiliary variables or linear inequality constraint such as the interval in which the population mean lies.

[8] proposed a Bayesian non-parametric imputation procedure to be used in the esti-
mation of population quantities in the 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 outcome variables results to significant gain in efficiency.

[1] developed the weighted FPBB and used it to account for item non response. The new procedure accounted for the unequal probabilities. It was further noted that the new procedure reduced bias although with little loss in efficiency. The weighted FPBB also reduced model misspecification in the imputation model. This is because it prevents erroneous inclusion of interaction between the other covariates and the design variables.

In this study the weighted FPBB model was modified so as to adjust for PPS selection by incorporating size variables and applying it in prediction of the non-sampled sizes. A model was used to predict the survey outcome variables for the obtained non-sampled sizes. The Constrained Bayesian Bootstrap procedure of [8] was then used to modify the weighted FPBB of [1]. This resulted to a constrained weighted FPBB.
1.2 Statement of the Problem

In the Bayesian approach the population posterior distribution incorporates prior information in the estimation of different population parameters. Most Bayesian models are based on polya posterior. The posterior distribution has Bayes stepwise justification and it is obtained after observing the samples. The posterior do not depend on the sampling design. It uses the observed sample units and available auxiliary variables to estimate the population parameters.

In most research auxiliary variables are used to forecast missingness. These variables are not of the main interest but can be useful in missing data estimation. Inclusion of auxiliary variables in multiple imputations procedure improves the model by improving the precision of the estimates.

In this research the main interest was to incorporate the auxiliary variables in the two-step semi-parametric multiple imputation procedure so as to improve efficiency and reduce the biasedness in estimation of the mean. In particular, prior information on auxiliary variables was considered in derivation of the two-step semi-parametric multiple imputation procedure. The inclusion of the auxiliary variables in the weighted FPBB was done under the assumption of MAR, since under MAR the condition of missingness is independent of unobserved data.
1.3 Significance of Study

Multiple imputation procedures assume auxiliary information is absent and a simple random sampling design in generation of missing data. The existing statistical packages do not incorporate complex designs in imputation. Information on respondent and non-respondent may be available in a survey in form of auxiliary variables. In this study auxiliary variables were incorporated in multiple imputations in a complex sampling design. The developed imputation procedure accounts for the complex sampling design by undoing the weights. This study shows how to properly account the design features in imputation model so as to protect the process against model misspecification.

The existing two-step semi-parametric multiple imputation procedure assumes auxiliary information is unavailable. The developed imputation procedure shows how such information can be incorporated in the weighted FPBB and used to account for the sampling weights.

The developed imputation procedure can be applied in cases of complex sampling design. The developed imputation procedure can also be applied in small area estimation to produce subpopulation estimates known as domain. These domains are important in policy making. This is because the constrained weighted FPBB provides a Bayesian solution in small area estimation in presence of both missing and complete data.
1.4 Research Objectives

1.4.1 The main objective

To derive a modified two-step semi-parametric multiple imputation procedure that incorporates auxiliary variables in estimation of population mean.

1.4.2 Specific objectives

1. To derive a modified two-step semi-parametric multiple imputation procedure using auxiliary variables.

2. To apply the modified two-step semi-parametric multiple imputation procedure in the estimation of population mean.

3. To determine asymptotic properties of the derived estimator of population mean.

4. To perform simulation study and compare efficiency of the derived estimator with those proposed by [1].
2 LITERATURE REVIEW

2.1 Introduction

In this chapter section 2.1 reviews previous work on inclusion of auxiliary variables in multiple imputations. The existing two-step semi-parametric multiple imputation procedure is reviewed in section 2.2 and 2.3. A posterior predictive distribution is discussed in section 2.2. Section 2.3 describes a Weighted Finite Population Bayesian Bootstrap. In section 2.4, estimation of population mean using the existing two-step semi-parametric multiple imputation procedure is discussed.

Under the single stage cluster sampling design, the population is organized into clusters. These clusters are known as primary sampling units (PSU). A sample of clusters is chosen and all the units in it sampled. The selection of the cluster to be sampled is done using probability proportional to size sampling design, which entails the calculation of probability of selecting each sample. Under probability proportional to size sampling design, the probability of selecting a sampling unit is proportional to the size measure \((Z)\) which is known for all units in the population. A unit has a higher probability of being sampled if it is large. The population consist of the survey outcome variable \(Y\), a covariate \(X\) and a size measure \(Z\) based on which PPSWOR sampling is conducted.

According to [6], in finite population sampling prior information is always present in form of auxiliary variables. For instance the value of the population mean may be known or even the interval in which it lies. According to [6], auxiliary variables are available
for every unit in the sample. In a set of data both the variables of interest $Y$ and the auxiliary variables denoted by $Z$ are present. The prior information about the population can be expressed using a set of linear inequality constraints on the population value of auxiliary variables.

[9] 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.

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 maintainance of the Missing at Random (MAR) condition. According to [7], the auxiliary variables can be used to create a sampling design and also in making inferences. When creating a sampling design using the auxiliary variables, one needs to know the value of the auxiliary variables for every unit in the population.

[6] proposed a constrained polya posterior to be used when there are prior information about population quantile and means of the auxiliary variables. This is one of the ways of incorporating prior information. The constrained polya posterior is a generalised polya posterior resulting in cases where prior information about auxiliary variables is available during sampling. [7] showed how the constrained polya posterior can be used to incorporate weights for sampled units where the weights depend on auxiliary variables.
2.2 The Posterior Predictive Distribution

Consider a single stage sampling design with complete data set. Let $Y$, $W$ and $I$ be the sampling outcome of interest, weight and sampling indicator respectively for the population. Where $I = 1$ if $i^{th}$ unit is sampled and 0 otherwise. This results to $Y_s, Y_{ns}, W_s$ and $W_{ns}$ that is sampled survey outcome variable, non-sampled survey outcome variable, sampled weights and non-sampled weights respectively.

In Bayesian inference a prior distribution of the parameter $\theta$ is specified in addition to the variable of interest $Y$. The posterior predictive distribution of $\theta$ given the observed sample is $P(\theta/Y_s)$. The posterior predictive distribution of the non-sampled units given the sampled units is

$$P(Y_{ns}/Y_s) = \int P(Y_{ns}/\theta, Y_s)P(\theta/Y_s)d\theta$$

(1)

According to [3] the posterior distribution of $\theta$ is obtained by averaging the complete data posterior of $\theta$ over the posterior predictive distribution of the missing units. $P(\theta/Y_{s,obs}, X_s, Z, R_s, I) = \int P(\theta/Y_s, X_s, Z, R_s, I)P(Y_{s,mis}/Y_{s,obs}, X_s, Z, R_s, I)dY_{s,mis}$, where $R$ and $I$ are ignorable since $P(I, R/X, Y) = P(I/X, Y)P(R/X, Y)$. This implies, $P(\theta/Y_{s,obs}, X_s, Z) = \int P(\theta/Y_s, X_s, Z)P(Y_{s,mis}/Y_{s,obs}, X_s, Z)dY_{s,mis}$

This is integrated using monte carlo simulation which iterates between the draws of the parameter $\theta$ conditional to the filled in data $P(\theta^t/Y_{s,obs}, Y_s^{(t-1)}_{s,mis}, X_s, Z)$ and imputation of the missing data conditional to the observed data and the draws of $\theta$. 
\[ P(Y_{s,t}^{mis}/Y_{s,obs}, \theta^t, X_s, Z), \] where \( t \) indices iterations.

The model of the survey outcome variable given the weight is \( P(Y/\theta, W) \). This is because the presence of the sampling weights allows the mechanism which generate the vector of indicators to be ignored. This implies that the non-sampled units are obtained using a posterior distribution model given by

\[
P(Y_{ns}/Y_s, W_s) \propto \int P(Y_{ns}/Y_s, \theta, W) P(\theta/Y_s, W) P(W_{ns}/W_s) d\theta dW_{ns} \tag{2}
\]

where \( \theta \) is used to parameterize the model of the outcome given the sampling weight. [10] simplifies (2) to

\[
P(Y_{ns}/Y_s, W) \propto \int P(Y_{ns}/Y_s, \theta, W) P(\theta/Y_s, W) d\theta \tag{3}
\]

, where the sampling weight were assumed to be approximately equal to the population weight, \( W_s = W \). This was further modified by [11]. The modification involved ignoring \( \theta \) in the model to come up with,

\[
P(Y_{ns}/Y_s, W_s) \propto \int P(Y_{ns}, W_{ns}/Y_s, W_s) P(Y_s, W_s) dW_{ns} \tag{4}
\]

The draws from \( P(Y_{ns}, W_{ns}/Y_s, W_s) \) are made using a weighted FPBB procedure as suggested by [12]. The parameter \( \theta \) is ignored because the draws of \( P(Y_s, W_s) \) are made directly from the posterior of the joint cumulative distribution function (cdf) of
$Y_s, W_s$ using the Bayesian Bootstrap of [3].

[1] modified (4) by incorporating missing data. The assumption made is that a response of indicators is observed if the whole population is sampled, such that $R = (R_s, R_{ns})$. Where $R_s$ is the response indicator for the units observed in the sample and $R_{ns}$ is the response indicator for the non sampled units. The sampled units were divided into the observed units and the missing units, $Y_s = (Y_{s,obs}, Y_{s,mis})$ according to their correspondence to the response indicator. Also the non sampled units were categorised into those ones that could have been observed had they been sampled and those that could be missing, $Y_{ns} = (Y_{ns,obs}, Y_{ns,mis})$. The observed data from the sampled and non sampled parts are then pooled together giving rise to $Y_{obs} = (Y_{s,obs}, Y_{ns,obs})$ and $Y_{mis} = (Y_{s,mis}, Y_{ns,mis})$.

Further [1] ignored the non response indicator and modified (2) to incorporate item level missingness. This resulted to

$$P(Y_{ns,obs}, X_{ns}/Y_{s,obs}, X_s, W_s) = P \int P(Y_{mis}/Y_{s,obs}, X_s, W_s) dY_{mis}.$$  

In this case data is generated from

$$P(Y_{ns,obs}, X_{ns}/Y_{s,obs}, X_s, W_s) = P \int P(Y_{mis}/Y_{ns,obs}, X_{ns}, X_s, W_s) dY_{mis}, \text{ by generating the missing values together with the observed values using the weighted FPBB method.}$$

The integration is over $Y_{mis}$. This is done by assuming a parametric model for $Y/X$. Thus, $\int P(Y_{ns,obs}, X_{ns}, Y_{mis}/Y_{s,obs}, X_s, W_s) dY_{mis} = \int P(Y_{mis}/Y_{ns,obs}, X_{ns}, Y_{s,obs}, X_s, W_s) P(Y_{ns,obs}, X_{ns}/Y_{s,obs}, X_s, W_s) dY_{mis}$. 


The model was again parameterized using $\theta$ and integrated with respect to the posterior distribution of $\theta$. Thus,

$$
\int P(Y_{ns,obs}, X_{ns}/Y_{s,obs}, Y_{mis}, X_s, W_s)dY_{mis} = \int \int P(Y_{mis}/Y_{ns,obs}, X_{ns}, Y_{s,obs}, X_s, W_s, \theta)P(Y_{ns,obs}, X_{ns}/Y_{s,obs}, X_s, W_s, \theta) d\theta dY_{mis}
$$

This is implemented using Gibbs sampler in the R package. The Gibbs sampler iterates between the draws of

$$
P(\theta/Y_{ns,obs}, X_{ns}, Y_{s,obs}, X_s, W_s, Y_{mis}) = P(\theta/Y, X, W_s) = P(\theta/Y, X)$$

and

$$
P(Y_{mis}/Y_{ns,obs}, X_{ns}, Y_{s,obs}, X_s, W_s, \theta) P(\theta/Y, X) \propto P(Y/\theta, X) P(\theta/X).
$$

The missing data in the sample $Y_{mis}$ depends on observed weights which are undone using the weighted FPBB.

In the proposed study it was assumed that auxiliary variable was present for all the units in the population and this information was used in developing a posterior predictive model for the size variables.
2.3 Weighted FPBB

Weighted FPBB is used to undo the sampling design (Sampling weights). This enables one to obtain draws from the posterior predictive distribution that are free of the effect of unequal probability selection. Thus the weighted FPBB enables one to draw from the posterior predictive distribution based on unequal probability of selection sampling design without making assumptions about the data generation mechanism.

The Polya posterior is similar to the finite population Bayesian Bootstrap (FPBB) of [5]. In Polya posterior the population is assumed to be from a simple random sample where $n$ is considered to be sample size and $N$ the population size. The sample is denoted by $y_s = y_1, ..., y_n$ where $y$ is the observed value of the response variable $Y$. The set of k distinct values in the sample is denoted by $(d_1, d_2, ..., d_k)$ and a vector of probabilities by $\lambda = (\lambda_1, \lambda_2, ..., \lambda_k)$. This implies $P(y_i = d_j/\lambda) = \lambda_j$,

$$\sum_{j=1}^{k} \lambda_j = 1$$

, where $i=1,2,...,n$ and $j =1,2,...,k$. Let $n_j$ be the number of units with the value $d_j$ within a sample and $n_j'$ the number of units with the value $d_j$ in the non sampled portion of the population. This implies that

$$\sum_{j=1}^{k} n_j' = N - n$$
, where \( j = 1, 2, \ldots, k \).

To obtain a Dirichlet posterior distribution of the parameter \( \lambda \), both a multinomial distribution and a non-informative Halden prior model are assumed for the sampled values. That is

\[
\begin{align*}
n_1, \ldots, n_k / \lambda_j &\sim \text{mult}(n; \lambda) \\
\text{and} \quad \lambda; \lambda &\sim \text{Dir}(0, \ldots, 0)
\end{align*}
\]

is given by

\[
P(\lambda) \propto \prod_{j=1}^{k} \lambda_j^{-1}.
\]

Hence the Dirichlet posterior distribution is given by \( \lambda/n_1, \ldots, n_k \sim \text{Dir}(n_1, \ldots, n_k) \).

Thus

\[
P(\lambda/n_1, \ldots, n_k) \propto \prod_{j=1}^{k} \lambda_j^{n_j-1}
\]

Considering the non-sampled data the posterior predictive distribution of its count follows a multinomial distribution

\[
n_1', \ldots, n_k'/n_1, \ldots, n_k \sim \text{mult}(N - n; \lambda).
\]

That is

\[
P(n_1', \ldots, n_k'/n_1, \ldots, n_k) = \frac{\int_{f_0}^{1} \cdots \int_{f_0}^{1} \prod_{j=1}^{k-1} \lambda_j^{n_j+n_j'-1} (1 - \sum_{j=1}^{k-1} \lambda_j)^{n_k+n_k'-1} d\lambda_1, \ldots, d\lambda_{k-1}}{\int_{f_0}^{1} \cdots \int_{f_0}^{1} \prod_{j=1}^{k-1} \lambda_j^{n_j} P(n_1, \ldots, n_k/\lambda) P(\lambda) d\lambda_1, \ldots, d\lambda_{k-1}}
\]

\[
= \frac{\int_{f_0}^{1} \cdots \int_{f_0}^{1} \prod_{j=1}^{k-1} \lambda_j^{n_j+n_j'-1} (1 - \sum_{j=1}^{k-1} \lambda_j)^{n_k+n_k'-1} d\lambda_1, \ldots, d\lambda_{k-1}}{\int_{f_0}^{1} \cdots \int_{f_0}^{1} \prod_{j=1}^{k-1} \lambda_j^{n_j} (1 - \sum_{j=1}^{k-1} \lambda_j)^{n_k-1} d\lambda_1, \ldots, d\lambda_{k-1}}
\]
[1] generalised this predictive distribution to a case where the sampled data contain different weights. This was done by denoting the sample by \((Y_s, X_s, W_s, R_s) = [(y_i, x_i, w_i, R_i), i = 1, 2, ..., n]\), where

\[
w_i = \sum_{i=1}^{N} \frac{z_i}{n z_i}
\]

, \(z\) is the size variable, \(n\) is the sample size and \(N\) the population size. In this case \(Y_i = Y_{i,obs}\), is considered when \(R_i = 1\) and 0 otherwise. The set of \(k\) distinct vectors of \((y_i, x_i, w_i, R_i)\) was denoted by \((d_1, ..., d_k)\) and \(\lambda = (\lambda_1, \lambda_2, ..., \lambda_k)\) denoted the vector of probabilities that is \(P[(y_i, x_i, w_i, R_i) = d_j/\lambda] = \lambda_j\), where

\[
\sum_{j=1}^{k} n_j = n
\]

and

\[
\sum_{j=1}^{k} n'_j = N - n
\]

, \(i=1,2,...,n, j=1,2,...k\). The sampled units contain different vectors of values such that \(k=n\). The sampling weight was denoted by \(w_i\) where

\[
\sum_{i=1}^{n} w_i = N
\]

[1] came up with a dirichlet posterior distribution \(\lambda : \lambda/w_1, ..., w_k \sim Dir(w_1, ..., w_K)\)

by assuming a multinomial distribution for the weighted counts.
and a Halden prior of $\lambda : \lambda \sim Dir(0, ..., 0)$. This resulted to

$$P(\lambda / w_1, ...., w_K) \propto \prod_{j=1}^{k} \lambda_j^{w_j-1}$$

This implies that the counts in the non-sampled population follow a multinomial distribution, that is

$$n'_1, ..., n'_k / w_1, ..., w_k \sim mult(N - n; \lambda^*)$$

where $\lambda^* = (\lambda^*_1, ..., \lambda^*_k)$ is an adjusted parameter. Thus the posterior predictive distribution is

$$P(n'_1, ..., n'_k / w_1, ..., w_k) = \int_0^1 ... \int_0^1 \prod_{j=1}^{k-1} (\lambda^*_j)^{w_j + n'_j - 1} (1 - \sum_{j=1}^{k-1} \lambda^*_j)^{w_k + n'_k - 1} d\lambda^*_1, ..., d\lambda^*_{k-1}$$

where $\lambda^*_j = C.\lambda_j (w_j - 1)$ for $j=1,2,...,k$ and C is a constant.

[1] considered the weighted polya urn suggested by [12] and used equation (9) to propose an adapted weighted FPBB procedure. The procedure involved two stages.

In the first stage the intial sample is resampled by use of Bayesian Bootstrap. This generates L replicate Bayesian Bootstrap sample of size n such that $(Y^t_l, X^t_l, W^t_l, R^t_l), l = 1, 2, ..., L$.

Let $t_l(i)$ denote the number of times the unit i is picked from the $l^{th}$ replicate Bayesian Bootstrap sample. This is considered in computation of $i^{th}$ bootstrap weight for unit i by $w_i^{(l)} = w_i. t_l(i)$. The bootstrap weights are then used in the second stage.
The second stage entails undoing the sampling weight by use of a weighted FPBB. A synthetic population denoted by B is created from the L Bayesian Bootstrap. This is done using the weighted FPBB shown in (8). Thus,

\[(Y^*(l), X^*(l), R^*(l)) (Y^{(lb)}_{ns}, X^{(lb)}_{ns}, R^{(lb)}_{ns}), \]

where \(b=1,2,\ldots,B\) and \(l=1,2,\ldots,L\). This draws the predicted counts of different non-sampled units given that of the Bayesian Bootstrap sample. The weighted FPBB is approximated using a monte-carlo simulation. [12] proposed a method for carrying out the simulation. It involves simulation of the posterior predictive distribution of counts in the non-sampled part by generation of B synthetic population for every L Bayesian Bootstrap sample. A polya sample, \((Y^{(lb)}_{ns}, X^{(lb)}_{ns}, R^{(lb)}_{ns})\) of size \(N-n\) is selected from the urn \((Y^*(l), X^*(l), R^*(l))\). Selection of a unit \((Y^{(l)}_i, X^{(l)}_i, W^{(l)}_i, R^{(l)}_i)\) is done using the probability given by,

\[ \lambda^{(l)*} = \frac{w^{(l)}_i - 1 + l_{i,k-1} \times (N-n) \times (k-1) \times \left(\frac{N-n}{n}\right)}{N-n + (k-1) \times \left(\frac{N-n}{n}\right)} \]  

(9)

In this case \(i=1,2,\ldots,n\) and \(k=1,2,\ldots,N-n+1\). In the equation \(w^{(l)}_i\) is the bootstrap weight for \(i^{th}\) unit in the \(l^{th}\) replicate of the Bayesian Bootstrap sample. Here \(l_{i,k-1}\) is the number of selection of unit \(i\) such that when \(k=0\), \(l_{i,0} = 0\).

Before selection of the non-sampled units from the BB sample the probability of choosing a unit \(i\) denoted by the weight \(w^{(l)}_i\) is given by \(\frac{w^{(l)}_i - 1}{N-n}\), where \(w^{(l)}_i\) represents the units with the values \((Y^{(l)}_i, X^{(l)}_i, R^{(l)}_i)\) in the entire population.

During selection the probability of selection is adjusted according to the number of
times each unit among the sampled units is selected using the weighted FPBB procedure. During the selection process the denominator is increased to indicate the total units represented by unit $i$. This yields a weighted FPBB population of size $N$

$$P^{(l)}_b = (Y^{(l)}_s, X^{(l)}_s, R^{(l)}_s)(Y^{(lb)}_{ns}, X^{(lb)}_{ns}, P^{(lb)}_{ns}).$$

According to [1] this results to unweighted synthetic population,

$$P^{(l)}_{(b)} = (Y^{(lb)}, X^{(lb)}, R^{(lb)}) = (P^{(l)}_{(b),obs}, Y^{(lb)}_{mis})$$

where $b=1,...,B$ and $l=1,2,...,L$ with $Y^{(lb)}_{mis} = (Y^{(lb)}_{s,mis}, Y^{(lb)}_{ns,mis})$, as the unobserved data in the $lb^{th}$ synthetic population and $P^{(l)}_{(b),obs} = (Y^{(l)}_s, X^{(l)}_s, R^{(l)}_s)(Y^{(lb)}_{ns,obs}, X^{(lb)}_{ns,obs}, R^{(lb)}_{ns,obs})$ which is the observed data in the $lb^{th}$ synthetic population. The missing data is then imputed using a parametric model that involves the posterior predictive distribution $(Y^{(lb)}_{mis}/P^{(l)}_{(b),obs})$.

[1] carried out a preliminary study on the complete data to determine the number of B synthetic population required. It was observed that the variance estimate stabilized at $B=20$. This was used in the simulation and resulted to the following asymptotic properties; point estimate=1.458, Estimated variance=0.33, Variance=0.032, Root Mean Squared Error =0.178 and 95% confidence interval coverage=96%. A simulation study was conducted by generating data in a situation where the outcome variable is associated with the probability of selection and where the probability of selection is associated with the mechanism of missing data generation.

A missing data generation technique where the missing data depends on observed covariates (MAR-X) was considered. Three imputation models of different degrees of
misspecification were used. Model 1 ignores weight (x), model 2 include logz (x, logz), model 3 include both logz and its interaction with the other covariates (x*logz). Model 4 and model 5 are similar to model 2 and 3 respectively but logz is is repaced with $\frac{1}{z}$ that correponds to the weights, thus model 4 (x, w) and model 5 (x*w) as shown in table 1.

[1] used input weights that were assumed to be final weights after unit non-response adjustment. It was observed that inclusion of weights as covariate corrects bias in the mean.

In this study however, a constrained weighted FPBB was developed by incorporating size variables in the weighted FPBB so as to improve efficiency and reduce bias in estimation of mean.

Table 1: Simulation results of the synthetic method proposed by [1]

<table>
<thead>
<tr>
<th>Performance criteria</th>
<th>x</th>
<th>x, logz</th>
<th>x*logz</th>
<th>x, w</th>
<th>x*w</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bias</td>
<td>0.019</td>
<td>0.012</td>
<td>0.005</td>
<td>-0.011</td>
<td>-0.056</td>
</tr>
<tr>
<td>Mean variance</td>
<td>0.047</td>
<td>0.043</td>
<td>0.040</td>
<td>0.059</td>
<td>0.089</td>
</tr>
<tr>
<td>Empirical variance</td>
<td>0.044</td>
<td>0.039</td>
<td>0.036</td>
<td>0.051</td>
<td>0.067</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.211</td>
<td>0.196</td>
<td>0.19</td>
<td>0.224</td>
<td>0.265</td>
</tr>
<tr>
<td>95 % CI coverage Rates</td>
<td>95 %</td>
<td>95 %</td>
<td>95 %</td>
<td>96 %</td>
<td>97 %</td>
</tr>
</tbody>
</table>

2.4 Estimation of the Population Statistics

The estimates of the posterior variance and mean was obtained using the combined rules developed by [2]. [2] considered a complete data set with imputed values say,
$Y_{comp} = (Y_{s,obs}, Y_{s,mis}^{(l)})$ where $l=1,2,...,M$. The mean and posterior variance is given by,

$$\hat{Q} = \frac{1}{M} \sum_{l=1}^{M} Q_{y,comp}^{(l)}$$

and $V = U + (1 + \frac{1}{M})B$ respectively. Where $Q_{y,comp}^{(l)}$ is the point estimate obtained from the $l^{th}$ complete data set and

$$U = \frac{1}{M} \sum_{l=1}^{M} Var(Q_{y,comp}^{(l)})$$

is the within imputation variance calculated as the average of variance estimates based on the $M$ complete data sets.

$$B = \frac{1}{M - 1} \sum_{l=1}^{M} [\hat{Q} - Q(Y_{comp}^{(l)})]^2$$

is the between imputation variance.

In the case of complete data [11] showed how to obtain the mean and variance conditional on the population. A synthetic population obtained from the $B$ FPBB samples say, $P_{syn} = (P_{(1)}^{(1)},..., P_{(B)}^{(1)},..., P_{(1)}^{(L)},..., P_{(B)}^{(L)})$ was considered. The posterior predictive distribution of a population statistic $Q(Y) = Q$ is given by, $Q/P_{syn} \sim t_{L-1}(\bar{Q}_L, (1 + L^{-1})V_L)$. Where $t_{L-1}$ denotes the t distribution with $L-1$ degrees of freedom. In this
case the mean is estimated by,

$$\tilde{Q}^{(l)} = \lim_{B \to \infty} \frac{1}{B} \sum_{b=1}^{B} q^{(lb)}$$

The population mean is given by

$$\bar{Q}_{L} = \frac{1}{L} \sum_{l=1}^{L} \tilde{Q}^{(l)}$$

and the variance by

$$V_{L} = \frac{1}{L-1} \sum_{l=1}^{M} (\tilde{Q}^{(l)} - \bar{Q}_{L})^{2}$$

The estimator of $Q$ is given by $q^{(lb)}$ which is obtained from the $b^{th}$ FPBB synthetic population of $l^{th}$ BB sample. This implies that

$$\tilde{Q}^{(l)} = \hat{Q}^{(l)} = \frac{1}{B} \sum_{b=1}^{B} q^{(lb)}$$

[1] imputed the missing data within each synthetic population data set and obtained

$$P^{imp} = P^{(l)}_{(11)}, \ldots, P^{(l)}_{(1M)}, \ldots, P^{(l)}_{(BM)}, \ldots, P^{(L)}_{(BM)}.$$ This implies that $Q/P^{imp} \sim t_{L-1}(\bar{Q}_{L}, (1+ L^{-1})V_{L})$. This means that

$$\hat{Q}^{(l)} = \lim_{B \to \infty, M \to \infty} \frac{1}{BM} \sum_{b} \sum_{m} q^{lbm}$$
with $q^{(lbm)}$ being an estimate of $Q$ which is obtained from the $m^{th}$ imputation of $b^{th}$ synthetic population within $l^{th}$ BB sample. The statistic $\bar{Q}^{(l)}$ is estimated by

$$\bar{Q}^{(l)} = \frac{1}{BM} \sum_{b=1}^{B} \sum_{m=1}^{M} q^{(lbm)}$$
3 THE PROPOSED ESTIMATOR OF POPULATION MEAN

3.1 Introduction

In this chapter the estimator of population mean is discussed in detail. Section 3.2 describes a posterior predictive distribution that incorporates auxiliary variables. Thereafter, a constrained weighted FPBB version is introduced in section 3.3. The properties of the applied estimator of population mean are discussed in section 3.4.

3.2 Posterior Predictive distribution that incorporates auxiliary variables

In this study the assumptions made were:

- The data is missing at random (MAR)
- Auxiliary variables are all known in the population.
- The auxiliary variables are assumed to be independent and identically distributed.

In this study a Bayesian population model that makes use of prior information available about auxiliary characteristics was developed. A Bayesian model in which the prior distribution for the parameter \( \theta \) is specified together with a distribution for the population value conditional on \( \theta \) say \( P(Y/\theta) \) was considered. The posterior distribution of the
non-sampled data is determined by conditioning on the sampled data. It was assumed that auxiliary variables were observed for every unit in the population.

Consider a population with N units in which Y is the survey outcome variable, \( \hat{X} \) is the size variable, W is the weight, I is the sampling indicator and R is the response indicator. If a sample is selected, the units in the population can be divided into sampled and non-sampled. Also the population can be divided into the observed and missing units according to the response indicator \( R = (R_s, R_{ns}) \), where \( R_s \) is the response indicator for the observed units in the sample and \( R_{ns} \) is the response indicator for the non-sampled units.

The sampled size measures were divided into those observed and those missing \( \hat{X}_s = (\hat{X}_{s,obs}, \hat{X}_{s,ns}) \) according to their correspondence to the survey outcome variables \( Y_s = (Y_{s,obs}, Y_{s,ns}) \). The non-sampled size measures were also divided into observed and missing, \( \hat{X}_{ns} = (\hat{X}_{ns,obs}, \hat{X}_{ns,ns}) \). The non sampled size variables were obtained using a constrained weighted Finite population Bayesian Bootstrap. The observed and missing information on size variables were pooled together resulting to \( \hat{X}_{obs} = (\hat{X}_{s,obs}, \hat{X}_{ns,obs}) \) and \( \hat{X}_{mis} = (\hat{X}_{s,ns}, \hat{X}_{ns,ns}) \). The observed size variables in the entire population was then used in prediction of their associated survey variables via a linear regression model. These resulted to \( Y_{obs} = (Y_{s,obs}, Y_{ns,obs}) \) and \( Y_{mis} = (Y_{s,ns}, Y_{ns,ns}) \). The missing survey outcome variables were obtained by multiple imputation using a parametric model.
The joint distribution of the size variable and survey outcome variable was given by
\( P(\tilde{X}, Y) = P(\tilde{X})P(Y/\tilde{X}) \). The posterior predictive distribution of \( \theta \) given the observed sampled size variable is \( P(\theta/\tilde{X}_s) \). The distribution of \( i^{th} \) size variable in the sample was given by \( P(\tilde{X}_i/\theta_i) \). The marginalised density of the size variable was given by \( P(\tilde{X}/\theta) = P(\tilde{X}_s/\theta) + P(\tilde{X}_{ns}/\theta) \). The posterior predictive distribution of \( \theta \) was obtained by averaging the complete data on posterior of \( \theta \) over the posterior predictive distribution of the missing sizes. That is

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

\[
P(\theta/\tilde{X}_{s,obs}, W_s) = P(\tilde{X}_{s,obs}, W_s/\theta)P(\theta) \quad (10)
\]

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

\[
P(\tilde{X}_{ns}/\tilde{X}_s) = \int P(\tilde{X}_{ns}/\theta, \tilde{X}_s)P(\theta/\tilde{X}_s)d\theta \quad (11)
\]

The information on missing size variables and associated survey outcome variables was incorporated into (11) and resulted to

\[
P(\tilde{X}_{ns,obs}/\tilde{X}_s, W_s) = \int P(\tilde{X}_{ns,obs}, \tilde{X}_{mis}, Y_{mis}/\tilde{X}_{s,obs}, W_s)dY_{mis} \quad (12)
\]
The non-sampled data was generated together with the missing and observed information on the size variable and survey outcome variable using the constrained weighted FPBB. Since $\tilde{X}_{mis}$ is associated with $Y_{mis}$ the integration was done over $Y_{mis}$. Thus,

$$\int P(\tilde{X}_{ns,obs}, Y_{mis}/Y_{s,obs}, \tilde{X}_{s,obs}, W_{s})dY_{mis} = \int P(Y_{mis}/Y_{s,obs}, \tilde{X}_{obs}, W_{s})P(\tilde{X}_{ns,obs}/\tilde{X}_{s,obs}$$

$$W_{s})dY_{mis}$$

Equation (13) was parameterized and then integrated with respect to the posterior distribution of $\theta$ resulting to

$$\int P(\tilde{X}_{ns,obs}, Y_{mis}/Y_{s,obs}, \tilde{X}_{s,obs}, W_{s})dY_{mis} \propto \int \int P(Y_{mis}/Y_{s,obs}, \tilde{X}_{obs}, W_{s}, \theta)P(\tilde{X}_{ns,obs}/\tilde{X}_{s,obs}$$

$$W_{s}, \theta)P(Y_{s,obs}, \tilde{X}_{s,obs}, W_{s}/\theta)P(\theta)d\theta dY_{mis}$$

$$\int P(\tilde{X}_{ns,obs}, Y_{mis}/Y_{s,obs}, \tilde{X}_{s,obs}, W_{s})dY_{mis} = \int \int P(Y_{mis}/Y_{s,obs}, \tilde{X}_{obs}, W_{s}, \theta)P(\tilde{X}_{ns,obs}/\tilde{X}_{s,obs}$$

$$W_{s}, \theta)P(\theta/\tilde{X}_{obs}, Y_{obs}, W_{s})d\theta dY_{mis}$$

The simulation of equation (14) was done using the Gibbs sampler by iterating between the draws of $P(\theta/\tilde{X}_{s,obs}, Y_{obs}, W_{s}, Y_{mis}) = P(\theta/\tilde{X}, Y, W_{s}) = P(\theta/\tilde{X}, Y)$ and those of $P(Y_{mis}/\tilde{X}_{s,obs}, Y_{s,obs}, W_{s}, \theta)$. The draws of the non-sampled size variables were obtained using the constrained weighted
FPBB which undoes the sampling design (weights). This made it possible to draw units from the posterior predictive distributions that were free from the effect of unequal probability selection.

### 3.3 The Constrained Weighted FPBB

Consider a finite population of size N. Let Y be the survey outcome variable and \( \tilde{X} \) be the associated size variable. Let I be the sampling indicator and R the response indicator. Thus the observed data consist of \((Y_{\text{obs}}, \tilde{X}_{\text{obs}}, I, R)\)

Also consider a sample of size n. The sample consists of both the variable of interest and its associated auxiliary characteristic, \((y_i, \tilde{x}_i)\) where \(i=1,2,...,n\). In this study it was assumed that the size variables were available for all the sampled units and their corresponding total was known,

\[
X = \sum_{i=1}^{n} \tilde{x}_i
\]

Also it was assumed that the population mean of the auxiliary variable was known since it was assumed that there exist a linear relationship between the characteristic of interest and the known auxiliary characteristic.

In Probability proportional to size sampling design units were selected with probability proportional to \(\tilde{X}\) which is the size measure. The size measure is only reported for the sampled units. The number of the non-sampled units (N-n) is known but their sizes are unknown hence the probability proportional to size sampling scheme is said to be in-
formative. 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.

In this study the weighted FPBB model was modified so as to adjust for probability proportional to size sampling design and was used to predict the non-sampled sizes. Thereafter, a model was used to predict the survey outcome variable for the obtained non-sampled sizes. The constrained Bayesian Bootstrap procedure of [8] was used to modify the weighted FPBB of [1]. This resulted to a constrained Weighted FPBB.

Denote the distribution of the $i^{th}$ size variable in the sample by $f(\tilde{x}_i/\phi_i)$ where $i=1,2,...,n$. The constrained Weighted FPBB was obtained by assuming a multinomial distribution for the observed sampled size variables and Dirichlet posterior distribution for the parameter $\phi$. Thus the posterior predictive distribution for the non-sampled counts in the population was obtained by considering both the multinomial and Dirichlet distribution.

Denote the sample by $(Y_s, \tilde{X}_s, W_s, R_s) = (y_i, \tilde{x}_i, w_i, R_i), i = 1, 2, ..., n$, 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 $(\tilde{x}_1, ..., \tilde{x}_k)$ be a set of distinct sizes for the sampled units and $\phi = (\phi_1, ..., \phi_k)$ be a vector of probabilities. The sampled units were assumed to contain different vectors of values such that $n=k$. This implies that $P((y_i, x_i, w_i, R_i) = \tilde{x}_j/\phi) = \phi_j$. Let $n_j$ be the
number of sampled units with distinct sizes $\tilde{x}_j$, where

$$\sum_{j=1}^{k} n_j = n$$

and $i=1,2,...k$. 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 was obtained by assuming a multinomial distribution for the weighted counts. Thus $\tilde{x}_i/\phi \sim \text{multi}(n : \phi_1, ..., \phi_k)$ hence

$$P(w_1, ..., w_k/\phi) \propto \prod_{j=1}^{k} \phi_j^{w_j} \quad (15)$$

where $w_j = \frac{1}{\pi_j}$ and $\pi_j = \frac{n \tilde{x}_j}{X}$. To obtain a posterior distribution for the parameter $\phi$ both a multinomial distribution and a Halden prior of $\phi; \phi \sim \text{Dir}(0, ..., 0)$ were considered. This resulted to Dirichlet posterior distribution that was given by $\phi/w_1, ..., w_k \sim \text{Dir}(w_1, ..., w_k)$. Thus

$$P(\phi/w_1, ..., w_k) \propto \prod_{j=1}^{k} \phi_j^{w_j-1} \quad (16)$$

To create draws for the non-sampled counts using the constrained weighted FPBB the following was considered. Let $n'_j$ be the number of non-sampled counts with size mea-
The counts of non-sampled sizes have a posterior predictive distribution which follows a multinomial distribution with a sample size of \( N - n \) and probabilities \( (\phi_1^*, \ldots, \phi_k^*) \) where \( \phi_j^* = \frac{c \phi_j (1 - \pi_j)}{\pi_j} \). Where \( c \) is a constant that normalizes the expression. Thus the posterior predictive distribution of the non-sampled counts in the population was given by

\[
P(\tilde{X}_{ns}/\tilde{X}, \phi) = \int_0^1 \cdots \int_0^1 P(\tilde{X}_{ns}/\tilde{X}, \phi) P(\tilde{X}/\phi) P(\phi) d\phi_1, \ldots, d\phi_k\]

\[
P(\tilde{X}_{ns}/\tilde{X}, \phi) = \int_0^1 \cdots \int_0 \sum_{j=1}^k P(n_1'/w_1, \ldots, w_k/\phi) P(w_1, \ldots, w_k/\phi) P(\phi) d\phi_1, \ldots, d\phi_k\]

\[
P(\tilde{X}_{ns}/\tilde{X}, \phi) = \int_0^1 \cdots \int_0 \sum_{j=1}^k P(n_1'/w_1, \ldots, w_k/\phi) P(w_1, \ldots, w_k/\phi) P(\phi) d\phi_1, \ldots, d\phi_k\]

\[
P(n_1', \ldots, n_k'/w_1, \ldots, w_k) = \int_0^1 \cdots \int_0 \sum_{j=1}^k P(n_1'/w_1, \ldots, w_k/\phi) P(w_1, \ldots, w_k/\phi) P(\phi) d\phi_1, \ldots, d\phi_k\]

(17)

In this study the adapted version of the constrained FPBB was in line with that of [1]. It was carried out in two stages. The first stage involved replicating the initial sample using a constrained Bayesian Bootstrap to generate \( L \) replicate of the size measure. This involved drawing the posterior distribution of the parameter vector \( \phi \) conditional on the counts in the sampled data, \( (n_1, \ldots, n_k) \). Thus \( (\phi^{(l)}/\tilde{X}_s, W_s, R_s) \sim Dir(n_1, \ldots, n_k) \), where \( \phi^{(l)} = (\phi_1^{(l)}, \ldots, \phi_K^{(l)}) \). This generated the posterior joint distribution of all the size
variables in the population given the observed values in the sample.

This resulted to \((\tilde{X}^{(l)}_s, W^{(l)}_s, R^{(l)}_s)\), \(l = 1, 2, \ldots, L\). The number of times the size measure was picked from the \(i^{th}\) replicate was denoted by \(h_l(i)\), this was considered in calculation of the \(i^{th}\) bootstrap weight for the size measure. Thus \(w^{(l)}_i = w_i.h_l(i)\). The weights were used in the second stage.

The second stage involved undoing the sampling weights using a constrained weighted FPBB. This was achieved by creating a B unweighted population of size measures for each of the L replicates. This yielded the predicted counts of the non-sampled sizes for the replicated samples. The constrained weighted FPBB (17) was approximated using the procedure proposed by [12]. The procedure involved selecting a polya sample of size measures \((\tilde{X}^{(lb)}_{ns}, R^{(lb)}_{ns})\) of size \(N-n\) from the urn \((\tilde{X}^{(l)}_s, R^{(l)}_s)\) of size \(n\). The selection of \(i^{th}\) size measure \((\tilde{X}^{(l)}_i, W^{(l)}_i, R^{(l)}_i)\) was done using the probability.

\[
\phi^{(l)*} = \frac{w^{(l)}_i - 1 + l_{i,k-1} \times \left(\frac{N-n}{n}\right)}{N - n + (k - 1) \times \left(\frac{N-n}{n}\right)}
\]

(18)

where \(k=1,2,\ldots,N-n+1\) and \(i=1,2,\ldots,n\). In this case \(w^{(l)}_i\) 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 resulted into a constrained weighted FPBB population of size \(N\) denoted by

\[P^{(l)}_{(b),obs} = \left((\tilde{X}^{(l)}_s, R^{(l)}_s) (\tilde{X}^{(lb)}_{ns}, R^{(lb)}_{ns})\right)\text{ where } b=1,2,\ldots,B\text{ and } l=1,2,\ldots,L.\]

Thus the unweighted population of size variables was given by, \(P^{(l)}_{(b)} = (P^{(l)}_{(b),obs}; \tilde{X}^{(lb)}_{mis}).\)
This population of size variables was then used in a linear regression model to predict the population of the survey outcome variables. This resulted to unweighted population of the survey outcome variables denoted by $P_{\text{pl}}(b) = (P_{\text{s,obs}}(l), \tilde{X}_s(l), R_s(l))$ where $P_{\text{s,obs}}(l) = ((Y_{s,obs}^{(l)}, X_s^{(l)}, R_s^{(l)}), (Y_{ns,obs}^{(l)}, \tilde{X}_n^{(l)}, R_n^{(l)}))$ is the observed data and $Y_{\text{mis}}^{(lb)} = (Y_{s,\text{mis}}^{(lb)}, Y_{ns,\text{mis}}^{(lb)})$ is the missing data.

The missing data on the survey outcome variables were obtained by imputation using a parametric model. The undoing of the sampling design made it possible to conduct multiple imputation under the assumption of independent and identically distributed.

The draws of the missing data were obtained from the posterior predictive distribution $(Y_{\text{mis}}^{(lb)}/P_{\text{pl}}(b))$. This step yielded $M$ imputed data sets for each of the unweighted population generated above $P_{\text{pl}}(B) = (P_{(B1)}^{(l)}, P_{(B2)}^{(l)}, \ldots, P_{(BM)}^{(l)})$. The resulting population of survey outcome variables was denoted by,

$$P_{\text{comp}} = (P_{(11)}^{(l)}, \ldots, P_{(1M)}^{(l)}, \ldots, P_{(B1)}^{(l)}, \ldots, P_{(BM)}^{(l)}, \ldots, P_{(BM)}^{(L)}),$$

where $m=1,2,\ldots,M$ and $l=1,2,\ldots,L$.

The proposed estimator of population mean was given by

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

where

$$\bar{Q}^{(l)} = \lim_{B \to \infty,M \to \infty} \frac{1}{BM} \sum_{b=1}^{B} \sum_{m=1}^{M} q_{\text{lm}}^{(lbm)}$$

and $q_{\text{lm}}^{(lbm)}$ is the population statistic obtained from the $m^{th}$ imputation of $b^{th}$ unweighted
population within the $l^{th}$ Bayesian Bootstrap sample.

### 3.4 Properties of the proposed estimator of population mean

In this study, the missing data was first imputed using the weighted FPBB which resulted to B unweighted population of survey outcome variables. Each of the B unweighted population was then imputed using a parametric model. This resulted into M imputed data sets for each of the B unweighted population,

$$\mathcal{P}_{\text{comp}} = (P^{(l)}_{(11)}, \ldots, P^{(l)}_{(1M)}, \ldots, P^{(l)}_{(B1)}, \ldots, P^{(l)}_{(BM)}, \ldots, P^{(L)}_{(BM)}), \quad \text{where } b=1,2,\ldots,B, \ m=1,2,\ldots,M$$

and $l=1,2,\ldots,L$.

The population statistic was estimated by $q^{(l_{bm})}$ which was obtained from the $m^{th}$ imputation of $b^{th}$ unweighted population within the $l^{th}$ Bayesian Bootstrap sample. For complete data set the mean was given by

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

, where

$$\hat{Q}^{(l)} = \lim_{B \to \infty, M \to \infty} \frac{1}{BM} \sum_{b=1}^{B} \sum_{m=1}^{M} q^{(l_{bm})}$$

In this case $\hat{Q}^{(l)}$ was estimated by

$$\hat{Q}^{(l)} = \frac{1}{BM} \sum_{b=1}^{B} \sum_{m=1}^{M} q^{(l_{bm})}$$
, which is the point estimate of the proposed estimator.

The imputation model is correctly specified if $E(q_{lbm}) = Q$.

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

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

$$E(\bar{Q}_L) = \frac{1}{L} (LQ)$$
This showed that the proposed estimator of population mean was unbiased. The variance of the proposed estimator of population mean was given by

\[ V_L = \frac{1}{L-1} \sum_{l=1}^{L} (\hat{Q}^{(l)} - \bar{Q}_L)^2 \]

The mean squared error (MSE) was given by

\[ \text{MSE} = \frac{1}{L-1} \sum_{l=1}^{L} (q^{(lbm)} - Q)^2 \]

The coverage rate of the confidence intervals for the population mean was obtained using normal approximation. Given the confidence level \((1 - \alpha)100\%\), the confidence intervals for \(l^{th}\) bootstrap sample was obtained by

\[
\text{prob}(-Z_{\frac{\alpha}{2}} < Z < Z_{\frac{\alpha}{2}}) = 1 - \alpha, \text{ where } Z = \frac{\hat{Q}^{(l)} - \bar{Q}_L}{\sigma/\sqrt{BM}} \text{ and}
\]

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

\[
\text{prob}(-Z_{\frac{\alpha}{2}} < \frac{\hat{Q}^{(l)} - \bar{Q}_L}{\sigma/\sqrt{BM}} < Z_{\frac{\alpha}{2}}) = 1 - \alpha
\]

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

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

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

Thus the confidence intervals was given by \( \hat{Q}^{(l)} \pm \frac{\sigma}{\sqrt{BM}} Z_{\frac{\alpha}{2}} \). The coverage rates were obtained by computation of the proportions of samples for which the population mean was contained in the confidence intervals.
4 EMPIRICAL STUDY

4.1 Introduction

This chapter is structured as follows: section 4.1 describes the study design in details. The simulation study is described in section 4.2 and the results discussed in section 4.3. The chapter concludes with discussion in section 4.4 and recommendations in section 4.5.

In the study design the survey outcome variable was 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 was considered, MAR-X. A population which consist of the survey outcome variables, size variables and other covariate was generated with the following joint distributions:

- $\log Z \sim N(2, 1)$, the survey weights. $Z = \exp(\log Z)$, size variables that are closely associated with the survey outcome variable.
- $X \sim N(0.1 \times \log Z, \sigma^2_x)$, fully observed covariate information.
- $Y \sim N(0.1 \times X + 0.5 \times \log Z + 0.6 \times X \times \log Z, \sigma^2_y)$, the survey outcome variables of main interest.

4.2 Simulation study

A simulation study was conducted to assess the proposed estimator of population mean. The simulation design was as follows. In step one a population of size $N=4000$ was con-
sidered and used to obtain independent samples of size $n = 200$ of size variables using PPSWOR sampling design.

Step two involved generation of unweighted population of size variables. Each of the replicated samples was simulated using ‘wtpolyap’ function in the polyapost package. The number of simulation to be done was denoted by $B$.

Step three involved prediction of survey outcome variables for the generated unweighted population of size variables. This was done using linear regression function in R package. This resulted to a complete data of survey outcome variables which was referred to as “Before deletion population”.

In step four, missing data was created for each of the replicated samples. A probit model was used as a deletion function to create missing data on each of the replicated samples. Both $X$ and $Z$ were assumed to be fully observed. The missing data generation technique was 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 was considered to be missing if $T_j > 0$. This may be done using the function ‘simsem’ in R.

In step five, missing data was imputed for each of the replicated samples using ‘mice’ package in R. This was done using three different models of misspecification; the first model included size variables only, $Z$. The second model included both size variables and weights say $Z, \log Z$. The third model included the interaction of size variables and weights, $Z^* \log Z$. 
Step six involved calculation of mean squared error and the 95% confidence interval coverage rates of the proposed estimator.

\[ MSE = \sqrt{\frac{1}{L} \sum_{r=1}^{L} (q_r - Q)^2} \]

The 95% confidence Interval Coverage rate of the proposed estimator was calculated based on the L replicates. The confidence interval of the estimator of population mean was obtained using \( \hat{Q}^{(i)} \pm 1.96 \frac{\sigma}{\sqrt{BM}} \).

4.3 Simulation results.

Two critical statistics examined were mean squared error and 95% confidence interval coverage rate. All are calculated based on \( B=20, M=5 \) and \( L=20,\ldots,100 \). The mean of the generated data was 1.456 and variance was 0.046. The simulation results are as given in table 2.

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

<table>
<thead>
<tr>
<th>Model</th>
<th>Sample size</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>70</th>
<th>80</th>
<th>90</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z only</td>
<td>MSE</td>
<td>0.144</td>
<td>0.104</td>
<td>0.100</td>
<td>0.096</td>
<td>0.087</td>
<td>0.078</td>
<td>0.071</td>
<td>0.057</td>
<td>0.053</td>
</tr>
<tr>
<td></td>
<td>95% CI Cov.</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>98%</td>
<td>92%</td>
<td>94%</td>
<td>90%</td>
<td>90%</td>
<td>87%</td>
</tr>
<tr>
<td>X only</td>
<td>MSE</td>
<td>0.145</td>
<td>0.105</td>
<td>0.102</td>
<td>0.098</td>
<td>0.088</td>
<td>0.078</td>
<td>0.066</td>
<td>0.059</td>
<td>0.052</td>
</tr>
<tr>
<td></td>
<td>95% CI Cov.</td>
<td>95%</td>
<td>97%</td>
<td>98%</td>
<td>98%</td>
<td>92%</td>
<td>91%</td>
<td>91%</td>
<td>92%</td>
<td>92%</td>
</tr>
</tbody>
</table>

Table 2 displays the summary results of imputation model that included auxiliary variables, Z only and other covariates, X only.
Figure 1: Plots of mean squared errors of imputation models that includes Z only and X only

Figure 1 compares these two imputation models in terms of mean squared error. Under Z only imputation model, most of the mean squared errors were observed to be lower than those for X only imputation. The coverage rates decreased with increase in sample size. Under X only imputation model, the coverage rates were closer to the nominal level.
Table 3: Summary results for imputation model that includes weights.

<table>
<thead>
<tr>
<th>Model</th>
<th>Sample size</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>70</th>
<th>80</th>
<th>90</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z,logZ</td>
<td>MSE</td>
<td>0.145</td>
<td>0.103</td>
<td>0.099</td>
<td>0.095</td>
<td>0.086</td>
<td>0.077</td>
<td>0.071</td>
<td>0.056</td>
<td>0.054</td>
</tr>
<tr>
<td></td>
<td>95% CI Cov.</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>98%</td>
<td>92%</td>
<td>94%</td>
<td>90%</td>
<td>90%</td>
<td>88%</td>
</tr>
<tr>
<td>X*logZ</td>
<td>MSE</td>
<td>0.146</td>
<td>0.110</td>
<td>0.104</td>
<td>0.098</td>
<td>0.088</td>
<td>0.077</td>
<td>0.066</td>
<td>0.058</td>
<td>0.051</td>
</tr>
<tr>
<td></td>
<td>95% CI Cov.</td>
<td>95%</td>
<td>97%</td>
<td>98%</td>
<td>98%</td>
<td>92%</td>
<td>91%</td>
<td>91%</td>
<td>92%</td>
<td>92%</td>
</tr>
</tbody>
</table>

Table 3 displays summary results of imputation model that incorporates survey weights.

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

Figure 2 compares the mean squared error of imputation model that includes both auxiliary variables and weights to those that includes other covariates and weights. The mean squared errors were lower than those in table 1. This implies imputation model that included weights outperformed imputation model that included X only and Z only.
The imputation model that included both auxiliary variables and weights had coverage rates which were higher than the nominal level when the sample size was small. The imputation model that included both auxiliary variables and weights outperformed that which included both weights and other covariates.

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

<table>
<thead>
<tr>
<th>Model</th>
<th>Sample size</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>70</th>
<th>80</th>
<th>90</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Z^\log Z$</td>
<td>MSE</td>
<td>0.146</td>
<td>0.100</td>
<td>0.096</td>
<td>0.092</td>
<td>0.085</td>
<td>0.077</td>
<td>0.070</td>
<td>0.056</td>
<td>0.053</td>
</tr>
<tr>
<td></td>
<td>95% CI Cov.</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>98%</td>
<td>92%</td>
<td>94%</td>
<td>90%</td>
<td>90%</td>
<td>87%</td>
</tr>
<tr>
<td>$X^\log Z$</td>
<td>MSE</td>
<td>0.155</td>
<td>0.110</td>
<td>0.104</td>
<td>0.098</td>
<td>0.088</td>
<td>0.077</td>
<td>0.066</td>
<td>0.059</td>
<td>0.050</td>
</tr>
<tr>
<td></td>
<td>95% CI Cov.</td>
<td>95%</td>
<td>97%</td>
<td>98%</td>
<td>98%</td>
<td>92%</td>
<td>91%</td>
<td>91%</td>
<td>92%</td>
<td>92%</td>
</tr>
</tbody>
</table>

Table 4 shows summary results of imputation model that included interaction between weights, auxiliary variables and other covariates.

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

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

4.4 Conclusion

The aim of this study was to propose a two-step semi-parametric multiple imputation procedure that incorporates auxiliary variables and apply it in the estimation of population mean. The first step involved replication of the size variables using constrained bayesian bootstrap and imputation of the non-sampled size variables using constrained weighted finite population bayesian bootstrap. The second step involved conducting parametric multiple imputation.

The proposed imputation procedure was applied in the estimation of population mean in the presence of auxiliary variables and resulted to unbiased estimates and gain in efficiency. The mean squared errors of the proposed estimator of population mean were observed to decrease with increase in sample size in all the imputation models. This is because a large sample has characteristics similar to those of the population hence the mean squared errors tend to converge to its population counterpart. The 95% confidence
interval rates of the proposed estimator were closer to nominal level when the sample size was small.

The proposed procedure outperformed that suggested by [1] since it reduced bias and resulted to gain in efficiency.

4.5 Recommendations

The proposed procedure assumed a linear relationship between the survey variables and auxiliary variables. In presence of non-linear relationship, Bayesian Penalized Spline model can be used in prediction of the survey variables before conducting multiple imputation.
References


APPENDICES

Appendix 1 : R code for using the proposed methodology

```r
set.seed (4000)
library (survey)
library (polyapost)
library (simsem)
library (mice)

Generation of parent sample data,
logz=rnorm (200, 2, 1)
z=exp (logz)
x=rnorm (200, 0.1*logz, )
y=rnorm (200, 0.1*x + 0.5*logz + 0.6*logz*x, )
Y=mean (y) , mean of the generated data
dtm1=data.frame (y, z, x, logz) , sample data set for analysis

Draw L Bayesian Bootstrap samples for the parent sample (L=1, 2, .., 100)
dsgn<-svydesign (ids= 1, strata=NULL, nest=FALSE, data=dtm1, weights=logz)
dsgn.r < -as.svrepdesign (design=dsgn, type="subbootstrap", replicates=100)
k= repwt<-as.matrix (dsgn.r repweights)
k1=k [, 1] , l^{th} bootstrap weights

Use linear regression to obtain size measures from the replicated weights
```
lm (z \sim k)

z1 = \text{fit} = 54.35 + 0.2297 \times k1

Draw weighted FPBB for the size measures for (i in 1:L)

w1 = \text{wt.polyap(z1, k1, 20)} , B=20

lm (y \sim z1)

yc = \text{fit} = -32.8310 + 0.6282 \times z1 , l^{th} bootstrap size measures

Impose missing data on the unweighted population of survey variables for (i in 1:L)

lm (x \sim y)

x1 = \text{fit} = 0.1929 + 0.3403 \times yc , l^{th} bootstraps covariates

dtm2 = \text{data.frame(x1, yc)}

script = 'yc \sim -0.635 + 0.4 \times x1'

missing = miss (logit=script)

dtm3 = impose (missing, dtm2)

ymis = dtm3 [,2]

Multiple imputation of missing data

dtm4 = \text{data.frame(ymis, w1)}

imp = mice (dtm4)

Com = complete (imp, 5) , m=5 is the number of multiple imputations

q = com [,1]

Analysis of the imputed data
\[ m = \text{sum}(q)/220 \quad \text{sample mean} \]

\[ v = \text{sum}((q-m)^2)/99 \quad \text{sample variance} \]

\[ \text{Mse} = \text{sum}((q-\bar{Y})^2)/99 \quad \text{Mean squared error} \]

\[ \text{LCL} = m - (1.96*\text{sqrt}(v)/\text{sqrt}(1100)) \quad \text{lower class limit} \]

\[ \text{Ucl} = m + (1.96*\text{sqrt}(v)/\text{sqrt}(1100)) \quad \text{upper class limit.} \]