Accounting for Complex Sample Designs in Multiple Imputation Using the Finite Population Bayesian Bootstrap

By

Hanzhi Zhou

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Survey Methodology)
in The University of Michigan
2014

Doctoral Committee:

Professor Michael R. Elliott, Co-Chair
Professor Trivellore E. Raghunathan, Co-Chair
Professor Roderick J. Little
Assistant Professor Brady T. West
Professor Richard L. Valliant, University of Maryland

DEDICATION

To my mom and dad

ACKNOWLEDGEMENTS

My utmost gratitude goes to my advisors, Mike Elliott and Trivellore Raghunathan, who have been incredibly supportive during my doctoral study. Without Mike and Raghu, I would have had a hard time developing from a student to a researcher. They should also be credited for inspiring this dissertation topic and providing regular guidance throughout this research. I would especially like to thank Mike for the late hours he has spent reviewing my dissertation as we approached the date of my defense.

Special thanks to my dissertation committee members: Rod Little, Rick Valliant and Brady West. I am grateful to Rod, whose constructive comments on my research and whose own tremendous work helped me understand everything better. I am also indebted to Rick and Brady, who carefully read many drafts of my writing and provided useful feedback throughout the development of this research.

Further thanks to faculty members in the Program in Survey Methodology (PSM): Jim Lepkowski, Steve Heeringa, Fred Conrad and Mick Couper, for providing me with interesting teaching and research assistantships over the years.

Lastly but most importantly, I would like to thank my wonderful family: my parents and my brother, their wisdoms, support and love accompanied me along the journey. I would also like to thank my terrific husband and my very best friend, Guihua,

who understood and shared every bit of my disappointments and successes through the past 10 years.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Multiple imputation (MI) is a well-established method to handle item-nonresponse in sample surveys. Survey data obtained from complex sampling designs often involve features that include unequal probability of selection, clustering and stratification. Because sample design features are frequently related to survey outcomes of interest, the theory of MI requires including them in the imputation model to reduce the risks of model misspecification and hence to avoid biased inference. However, in practice multiply-imputed datasets from complex sample designs are typically imputed under simple random sampling assumptions and then analyzed using methods that account for the design features. Less commonly-used alternatives such as including case weights and/or dummy variables for strata and clusters as predictors typically require interaction terms for more complex estimators such as regression coefficients, and can be vulnerable to model misspecification and difficult to implement.

We develop a simple two-step MI framework that accounts for complex sample designs using a weighted finite population Bayesian bootstrap (FPBB) method to generate draws from the posterior predictive distribution of the population. Imputations may then be performed assuming IID data. We propose different variations of the weighted FPBB for different sampling designs, and evaluate these methods using three studies. Simulation results show that the proposed methods have good frequentist properties and are robust to model misspecification compared to alternative approaches. We apply the proposed method to accommodate missing data in the Behavioral Risk Factor Surveillance System, the National Automotive Sampling System and the National

Health and Nutrition Examination Survey III when estimating means, quantiles and a variety of model parameters.

CHAPTER 1
INTRODUCTION

1.1 Objectives

Probability sample surveys where each member of the population has a known, non-zero probability of being selected form the backbone of empirical research in social science and public health. To assure broad representativeness, many sample survey designs use unequal probabilities of selection, selection of subjects in stages (introducing clustering) and stratification (For theoretical accounts of sampling methods, see Cochran, 1977 and Särndal et al., 1992). Accordingly, analysis methods for survey data need to take into account these complex sample design features.

It is notable that even the most well-designed sample surveys are imperfect in various ways. Missing data presents a particular challenge. Unit nonresponse occurs when sampled individuals fail to participate in the survey at all. Item nonresponse occurs when sampled individuals do not respond to certain questions. This is common in large scale surveys that include an extensive collection of questions. One principled approach to handle item nonresponse is multiple imputation (MI) (Rubin 1987). The key to success with MI lies in specifying an imputation model that reasonably describes the conditional distribution of the missing data given the observed data. Since complex sample design features frequently are related to survey variables, it is important to include them in the imputation model to reduce the risks of model misspecification and hence to avoid biased inference (Reiter, Raghunathan & Kinney, 2006). This thesis concerns methods that use

MI to deal with item nonresponse while accounting for complex sample designs. It has two *objectives*:

(1) To illustrate the bias that can arise using existing MI methods to account for complex sample designs when the imputation model is misspecified;

(2) To propose and evaluate a modified MI framework that accounts for complex sample designs with simpler modeling and with no resort to design-based estimators.

We explore the impact of the interrelationship among the data, the sampling mechanism and the missingness mechanism on the performances of the existing and the proposed MI methods. We use both simulated data and real survey data. The following *assumptions* are made throughout the thesis:

(1) There are no unit nonresponse problems;

(2) The data are missing at random (MAR).

Section 1.2 reviews approaches to survey sampling inference in the complete data context, with a focus on Bayesian approach. Section 1.3 reviews the standard multiple imputation (MI) as a well-established method to handle item nonresponse that has a Bayesian justification. The importance of accounting for complex sample designs in MI is discussed from a theoretical perspective. Section 1.4 and 1.5 point out the limitations of current techniques, and introduce the proposed method that builds upon the method reviewed in section 1.2. Finally, section 1.6 outlines the research question each chapter addresses.

1.2 Bayesian approach to survey sampling inference in the complete data context

Consider a finite population of size *N* with two types of measurement on the population units: $\{Y_i, Z_i\}$ for $i = 1,...,N$, where Y represents a single survey outcome of interest, and Z represents the design variable used for sample selection. We consider three types of *Z* in this thesis: (1) size measure for probability-proportional-to-size (PPS) sampling, (2) cluster indicators for two-stage cluster sampling, and (3) stratum indicators for stratified sample design. Typically Z is known for all units of the population to the sampler, but not necessarily to the data analyst. Let $I_i = \{I_1,...,I_N\}$ denote the vector of sample indicator variables, such that $I_i = 1$ if unit *i* is sampled and 0 otherwise. We use subscripts *s* and *ns* to denote the selected sample of size *n* and the nonsampled part of the population of size *N-n*, respectively. Thus both *Y* and *Z* divide into two parts: $\{(Y_s, Y_{ns}), (Z_s, Z_{ns})\}$. Let $w_s = \{w_i, i = 1,...,n.\}$ denote the sampling weights, e.g. $w_i = 1/\pi_i = \sum_{i=1}^{N} Z_i / nZ_i$ for a single-stage PPS sampling design. Figure 1.1 illustrates the data from a sample survey in the absence of missing data.

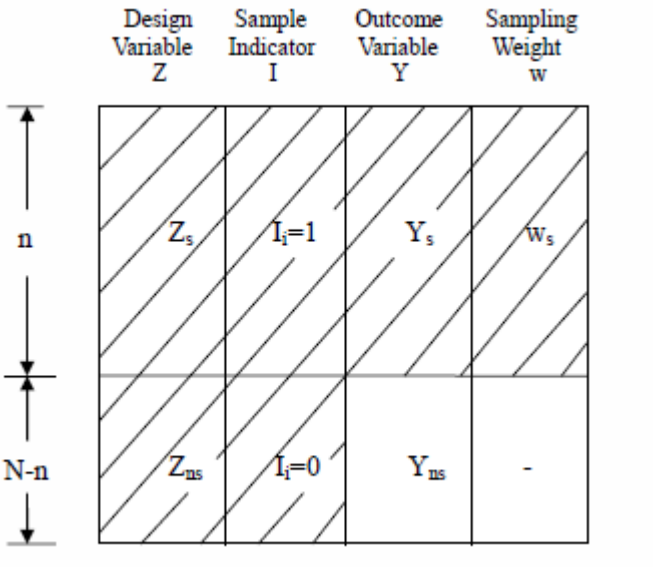

Figure 1.1 The data from a sample survey in the absence of missing data

In survey sampling, the objective is usually to learn about some population quantity denoted by $Q(Y)$, e.g. population mean/quantile/total, etc., through relating Y_s to Y_{ns} in some fashion. There are two general inferential frameworks to accomplish this: 1) the *design-based* approach (Hansen, Hurwitz & Madow, 1953; Cochran, 1977) treats the survey outcome *Y* as a fixed quantity, and imposes a random distribution on the sample inclusion indicator *I*. The statistical distribution of an estimator for $Q(Y)$ is thus induced by the sampling design. While the design-based framework brings in an objectivity element by minimizing the use of modeling assumptions, this objectivity is lost in the presence of nonsampling errors like nonresponse; 2) the *model-based* approaches, which include the frequentist modeling (Royall, 1970; Valliant et al., 2000) and the Bayesian modeling (Ericson, 1969; Basu, 1971). These regard the survey outcome *Y* as a random variable as well as *I*, and assume a model to predict Y_{ns} from Y_s. Both variants assume that *Y* comes from some parametric family of distributions indexed by the parameter θ. While the frequentist modeling treats θ as fixed, and bases inferences on repeated sampling from the model, the Bayesian modeling specifies a prior distribution on θ in addition to *Y*. The posterior distribution of θ given the observed sample $p(\theta \mid Y_s)$, and hence the posterior predictive distribution of the nonsampled population values given the sampled data $p(Y_{ns} \mid Y_s) = \int p(Y_{ns} \mid \theta, Y_s) p(\theta \mid Y_s) d\theta$, serves as the basis for inference about $Q(Y)$.

The Bayesian paradigm provides the most satisfying inferential approach to survey inference when it is done right. That is by incorporating complex sample design features to avoid sensitivity to model misspecification, using noninformative priors to

avoid subjectivity, and being frequentist calibrated, i.e. having good repeated sampling properties (Little, 2004; Little & Zheng, 2007). Bayesian finite population inference is thus proposed as a means to harmonize design- and model-based approaches for sample survey inference (Little, 2006, 2011; Gelman, 2007), and is exemplified by a variety of work in the complete data context. When design variables or the selection probabilities are known for all units in the population, Zheng and Little (2003, 2005) and Chen, Elliott and Little (2010) propose robust Bayesian predictive inference that improves efficiency over design-based estimators, using penalized splines under fairly weak model assumptions. In situations where the sizes of the non-sampled units are unavailable, Little and Zheng (2007) and Sangeneh, Keener and Little (2011) consider a two-step procedure to assure *ignorable sampling* (Sugden and Smith, 1984) with a PPS sampling design. In the first step, the nonsampled sizes (Z_{ns}) are predicted by a modified Bayesian Bootstrap (BB) procedure that adjusts for unequal probability sample selection (w_s), i.e. $p(Z_{ns} \mid Z_s, w_s)$; the nonsampled survey outcomes (Y_s) are then predicted using a penalized spline model relating Y with Z, i.e. $p(Y_{ns} \mid Y_s, Z)$.

The modified BB considered by Little and Zheng (2007) is a noninformative Bayesian method closely related to an offshoot of the Bayesian approach to surveys, known as the "Pseudo-Bayesian approach" (Ghosh & Meeden, 1997; Cohen, 1997; Dong, Elliott & Raghunathan, 2014). The modified BB is of particular interest in this thesis. Details of how our proposed methodology in this thesis builds upon and compares with these methods will be discussed in section 1.5 and in later chapters.

1.3 Standard multiple imputation as a calibrated Bayesian method to deal with item nonresponse

Consider the same finite population as in section 1.2, and assume item nonresponse occurs on the outcome variable *Y* and a covariate *X* is completely observed. Let $R_i = \{R_1, ..., R_N\}$ denote the response indicator variable for *Y*, such that $R_i = 1$ if unit *i* provides a value for *Y* and 0 otherwise. Write $R = (R_s, R_{ns})$, where R_s is observable from the sample and R_{ns} is unobservable for the nonsampled population. We use subscripts *obs* and *mis* to denote the responding and the non-responding units, respectively. Thus Y_s further breaks down to $Y_{s,obs}$ and $Y_{s,mis}$, and $Y = (Y_{s,obs}, Y_{nobs})$ if we recombine $Y_{s,mis}$ with Y_{ns} as unobserved data Y_{nobs}. The observed data may be thought as the outcome of two random processes: sampling and responding. We illustrate in Figure 1.2 the data from a sample survey with item nonresponse occurring on the survey outcome. Three types of population quantities are of interest in the thesis: the mean/proportion of *Y*, the regression coefficients of *Y* on *X*, and the quantiles of a continuous outcome *Y*.

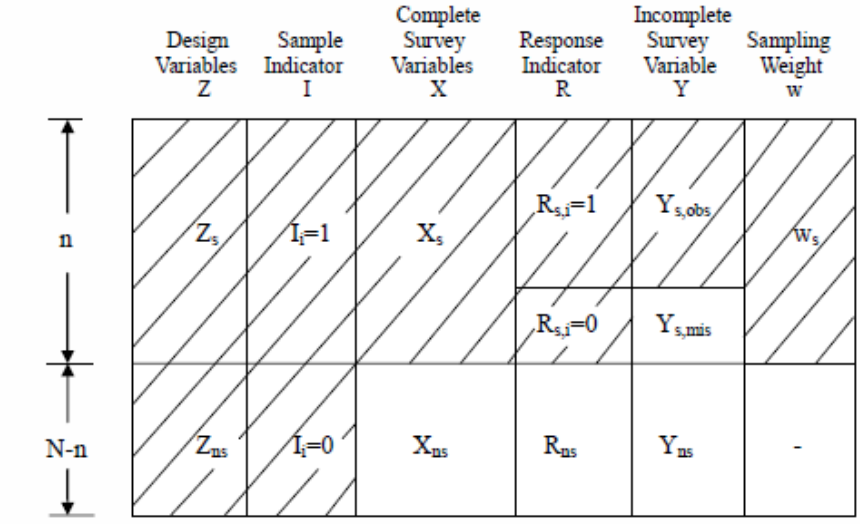

Figure 1.2 The data from a sample survey with item nonresponse on the outcome variable

The Bayesian modeling paradigm deals with nonresponse naturally, since unknowns about the finite population given the observed data can be generated from a

predictive distribution. We thus consider multiple imputation (MI) to deal with item nonresponse which has a Bayesian justification. The basic idea of MI is to replace each missing value with a set of plausible values which can then be combined in a simple way for inference using compete-data analysis techniques. The fundamental conceptualization of MI is Bayesian, where the *posterior* distribution of the model parameters θ of interest is obtained by averaging the *completed data posterior* of θ over the *posterior predictive* distribution of the missing data (Rubin, 1987, Result 3.1):

$$p(\theta \mid Y_{s,obs}, X_s, Z, R_s, I) = \int p(\theta \mid Y_s, X_s, Z, R_s, I)\, p(Y_{s,mis} \mid Y_{s,obs}, X_s, Z, R_s, I) dY_{s,mis} \qquad [1.1]$$

Assuming ignorable sampling $p(I \mid R,Y,Z) = p(I \mid R_s, Y_s, Z)$ and $p(R \mid Y, Z) = p(R \mid Y_{obs}, Z)$, [1.1] becomes $p(\theta \mid Y_{s,obs}, X_s, Z) = \int p(\theta \mid Y_s, X_s, Z)\, p(Y_{s,mis} \mid Y_{s,obs}, X_s, Z) dY_{s,mis}$, allowing the sampling and response mechanism to be ignored in the modeling.

This integration is typically accomplished using Markov Chain Monte Carlo data augmentation (Tanner & Wong, 1987). Let *t* index the iteration. Draws from the posterior predictive distribution of the missing data $p(Y_{s,mis}^{(t)} \mid Y_{s,obs}, X_s, Z)$ are obtained by iterating between draws of the model parameters conditional on the "filled in data" $p(\theta^{(t)} \mid Y_{s,obs}, Y_{s,mis}^{(t-1)},, X_s, Z)$ and imputations of the missing data conditional on the observed data and draw of the model parameter $p(Y_{s,mis}^{(t)} \mid Y_{s,obs}, \theta^{(t)}, X_s, Z)$. Rubin (1987) develops simple ***combining rules*** to obtain estimates of posterior means and variances using only a finite number (*M*) of independent draws $\left\{Y_{s,mis}^{(1)}, ..., Y_{s,mis}^{(M)}\right\}$ of the imputed data together with $Y_{s,obs}$, i.e. multiple "completed" datasets $y_{comp}^{(l)} = \left\{(Y_{s,obs}, Y_{s,mis}^{(l)}), l = 1, ..., M\right\}$. He shows that inferences obtained using these combining rules have good frequentist properties for relatively small *M* (5-20):

$$\begin{aligned} &\textit{Posterior Mean: } \hat{Q} = M^{-1}\sum_{l=1}^{M} Q(y_{comp}^{(l)}) \\ &\textit{Posterior Variance: } V = U + (1+M^{-1})B, \end{aligned} \qquad [1.2]$$

where $Q(y_{comp}^{(l)})$ is the point estimate obtained from the l^{th} completed dataset $y_{comp}^{(l)}$, $U = M^{-1}\sum_{l=1}^{M} \hat{\text{var}}\left(Q(y_{comp}^{(l)})\right)$ is the within imputation variance calculated as the average of variance estimates based on the *M* completed datasets, $B = (M-1)^{-1}\sum_{l=1}^{M}\left(\hat{Q} - Q(y_{comp}^{(l)})\right)^2$ is the between imputation variance.

Typically this is a combined design and model set up, where an imputation model is used to predict missing data in the sample. In other words, prediction of $p(Y_{s,mis}|Y_{s,obs})$ is model-based, and design-weighted estimators (i.e. $Q(y_{comp}^{(l)})$ and $\hat{\text{var}}\left(Q(y_{comp}^{(l)})\right)$) are used to estimate the finite population quantities of interest once $Y_{s,mis}$ are filled in by model predictions, i.e. predictions of $p(Y_{ns}|Y_s)$ is design-based (Reiter et al., 2006; Yuan & Little 2007a).

Rubin (1987) combines nonsampled and missing data into a single framework:

$$p(Y_{nobs}\,|\,Y_{s,obs}, X, Z, R_s, I; \theta, \varphi, \phi) \propto \int p(Y, X, Z; \theta)\, p(R\,|\,Y, X, Z; \phi)\, p(I\,|\,R, Y, X, Z; \varphi) dR_{ns} \qquad [1.3]$$

where θ, φ and ϕ denote the parameter that indexes the distribution of *Y*, *I* and *R*, respectively, and they are assumed *a priori* independent. He also gives conditions for *proper imputation* that ensure randomization validity for a *calibrated Bayes* in the sense of Little (2006, 2011). Rubin's conditions include: (i) point estimation is approximately unbiased for the scientific estimand of interest *Q(Y)*, and (ii) actual interval coverage equals the nominal interval coverage, over repeated imputation and sampling processes.

These in turn require *ignorability* assumptions of two random mechanisms--the specified sampling mechanism (*I*) and the posited missing data mechanism (*R*). The ***ignorability conditions*** (Rubin, 1976; Little, 1982) are:

- *Ignorable sampling*:

$$\begin{aligned} & p(I \mid R_s, Y, X, Z; \varphi) = p(I \mid R_s, Y_{s,obs}, X, Z; \varphi) \\ \Leftrightarrow\ & p(Y_{nobs} \mid Y_{s,obs}, X, Z, R_s, I; \theta, \varphi, \phi) = p(Y_{nobs} \mid Y_{s,obs}, X, Z, R_s; \theta, \phi) \end{aligned} \quad [1.4]$$

- *Ignorable missing data*:

$$\begin{aligned} & p(R_s \mid Y, X, Z; \phi) = p(R_s \mid Y_{s,obs}, X, Z; \phi) \\ \Leftrightarrow\ & p(Y_{nobs} \mid Y_{s,obs}, X, Z, R_s; \theta, \phi) = p(Y_{nobs} \mid Y_{s,obs}, X, Z; \theta) \end{aligned} \quad [1.5]$$

Thus explicit modeling for the sample indicator *I* and the response indicator R_s is not necessary in equation [1.3].

Rubin (1976) and Little and Rubin (2002) formalized the concept of ***missing data mechanism***. By their definition, ignorable missing data always implies missing at random (MAR). That is, given the observed data ($Y_{s,obs}, X, Z$), the missingness mechanism does not depend on the unobserved data (Y_{nobs}). MAR is a common assumption in practice and is typically assumed for implementing MI. Other missing data mechanisms include missing completely at random (MCAR), i.e. $p(R_s \mid Y, X, Z; \phi) = p(R_s \mid \phi)$, and not missing at random (NMAR), i.e. the missingness also depends on the unobserved data (Y_{nobs}). ***We focus on MAR in this thesis***. We do not consider MCAR which is often too ideal for real world sample surveys, or NMAR, which requires special statistical techniques (e.g. selection and pattern mixture models) beyond the scope of this thesis.

To make the MAR assumption plausible, we need a sufficiently rich imputation model that ideally includes in the covariate space $\{X, Z\}$ all variables that are related to

the sample selection and the missingness, and are potentially predictive of the outcome variable of interest (*Y*). For a complex sample survey, this implies that features such as stratification and clustering as well as unequal inclusion probabilities need to be built into the imputation model via design variables (Rubin, 1996).

1.4 Fully parametric techniques to account for complex sample designs in MI

Despite expert recommendations, imputers seldom account for sample designs when using available software packages to construct imputation models. They rely instead, on use of design-based estimators at the analysis stage to account for design effects. This can lead to biased point estimation and below-nominal confidence interval coverage (Reiter et al., 2006). More complex methods utilize design variables (*Z*) as covariates in models. In the setting where the design issue of interest lies in the probability of selection, Elliott and Little (2000) and Elliott (2007, 2008, 2009) account for unequal probabilities of inclusion by considering weight strata and treat the stratum means as random effects (which they term "weight smoothing models"). Their ideas to shrink the mean across weight strata are recently considered for accommodating survey weights in MI using random-effects imputation models (Carpenter et al., 2012). With stratified multistage sampling, Reiter et al. (2006) and Schenker et al. (2006) consider dummies for fixed stratum effects and fixed or random cluster effects in the imputation model. We call these MI techniques "***fully parametric MI***".

In practice, however, as more covariates and design variables are included in the model, the inferential conclusions become more valid conditionally but possibly more sensitive to model misspecification (Gelman et al., 2004). This leads to "uncongenial"

imputation (Meng, 1994)--since oftentimes, correct parametric or nonparametric approximations of the functional forms of design variables as well as their interactions with other covariates are needed. Pfeffermann (2011) points out that, modeling the relationship between the design variables and the covariates in order to integrate out the effect of the former can be complicated. In particular, uncongeniality concerns regarding the incorporation of survey weights in the imputation model for domain estimation (Kim et al., 2006), overestimation of MI variance under the fixed cluster effects model (Andridge, 2011), and convergence issues about random cluster effects models with high dimensional data and/or hierarchical data structure (Yucel & Raghunathan 2006; Zhao & Yucel, 2009) all limit the development of adequate software packages to apply these MI techniques.

The importance of accounting for sample designs in multiple imputation, and the limitation of fully parametric MI techniques to do this, creates an awkward dissonance between the theory and practice of multiple imputation. This also provides a good motivation for an alternative methodology that ideally satisfies the following properties:

- Consistent with the Bayesian derivation of MI
- Easier implementation and less expensive computation
- Less prone to model misspecification
- Do not require design-based estimators for complex population quantities of interest under complex sampling designs (e.g. domain estimation, quantile estimation, and the ratio of two medians for which there seems no readily available frequentist interval)

1.5 Proposed methodology to account for complex sample designs in MI

This thesis develops a modified MI framework (termed "*two-step MI*") to accommodate complex sample designs, which avoids the complication of modeling design variables (*Z*) and obviates the need for design-based analysis.

In the first step, we regard the nonsampled part of the target population as missing by design, and create synthetic populations that contain item-level missing data. Thus we obtain $p\left(Y_{ns}, X_{ns}, R_{ns}, w_{ns} \mid Y_s, X_s, R_s, w_s\right)$, where $Y_{ns} = (Y_{ns,obs}, Y_{ns,mis})$, w_s is the design weight for sampled units. In the second step, we multiply impute the missing values in the synthetic populations using standard parametric MI techniques assuming that the data are independent and identically distributed. Thus we obtain $p(Y_{mis} \mid Y_{obs}, X)$, where $Y_{mis} = (Y_{s,mis}, Y_{ns,mis})$, $Y_{obs} = (Y_{s,obs}, Y_{ns,obs})$. This is equivalent to a situation where a census is conducted but not all respondents answer all the survey questions, requiring multiple imputation to be performed on the entire target population.

Note that the proposed method follows a "synthesize/reverse design-then-impute" procedure, different from the procedure of Reiter (2004) who used the parametric MI to simultaneously impute missing data and generate synthetic data in a "two-step" fashion for disclosure risk limitation purpose. Reiter follows an "impute-then-synthesize" procedure in which modeling design variables in the imputation model remains an issue. *Note also that although the proposed MI framework requires one step to deal with nonsampled data, it is an intermediate step toward our ultimate goal of treating item-level missing data.*

To realize the first step in the proposed MI procedure, we consider a 'Pseudo-

Bayesian approach', i.e. the Polya posterior of Ghosh & Meeden (1997), which is equivalent to the finite population Bayesian bootstrap of Lo (1988). The approach ensures objectivity relative to a standard Bayesian approach by assuming a noninformative prior that is dominated by a nonparametric likelihood function. Because the sampling designs considered in this thesis all involve unequal selection probabilities, we consider the weighted version of the finite population Bayesian bootstrap ("weighted FPBB") (Cohen, 1997; Little & Zheng, 2007; Dong et al., 2014**).** We also develop several adapted versions of the weighted FPBB pertinent to accommodating different sample design features.

Though built upon the modified BB in Little and Zheng (2007), we have a different implementation of it because we have a different objective in the missing data context. We create the posterior joint distribution of the nonsampled population $\left(Y_{ns}, X_{ns}, R_{ns}\right)$ based on the weighted FPBB, whereas they 'impute' the unknown design variable Z_{ns} only. Our purpose is to reverse design effects while retaining population-level multivariate relationships among variables in the process. As stated previously, Z_{ns} is not imputed in our case, since we focus on model predictions of item-level missing data, and do not consider further capitalizing on *Z* to model the relationships of the design variable (*Z*) and the outcome (*Y*), as they do in their second step. Further, our implementation assumes no auxiliary information is available for the distribution of *Z* (e.g. the population mean of *Z*), and therefore do not adjust the weighted FPBB as they do to create synthetic populations that satisfy such a restriction.

1.6 Outline of chapters

Chapter 2 proposes a two-step MI framework to account for sampling weights.

We combine the method of Ghosh and Meeden (1997), who propose a Polya posterior to generate a noninformative Bayesian approach to finite population sampling, with that of Cohen (1997), who proposes a method to generate draws from Polya posteriors using data obtained from weighted sample designs in a non-clustered setting. We then modify the standard MI combining rules for inference that follow immediately from the rules developed in Raghunathan et al. (2003) for combining synthetic datasets. While the literature recognizes the importance of using weights in the analysis of complex sample survey data, methods incorporating weights in conjunction with multiple imputation to adjust for item nonresponse are underdeveloped. Imputation models that simply include weights or weight-related design variables as covariates can quickly become very complicated. They can be susceptible to misspecification if the inclusion probabilities are related to survey variables (Y) and/or the missing data mechanism in a nonlinear fashion.

The new procedure allows the sampling mechanism (I) and missing data mechanism (R) to be simultaneously disentangled, so that imputation can be performed assuming IID. As a result, data users need only to apply simple unweighted estimation methods to the imputed population datasets.

Our simulation assuming a PPS sampling design shows that the proposed method has good frequentist properties under MAR, which contrasts with standard approaches that are prone to model misspecification. We also apply the proposed method to estimation of means and linear regression, log-linear regression, and general location models using data from the BRFSS.

Chapter 3 extends the proposed MI framework to accommodate clustering effects as well as sample weight effects in two-stage unbalanced cluster samples. We extend the

approach of Chapter 2, this time combining the method of Meeden (1999), who proposes 'a two-stage Polya posterior' approach to a simple balanced two-stage cluster sample with equal selection probability at both cluster and element levels, with that of Cohen (1997). Because this approach requires that both first- and second-stage weights be known, we also propose an alternative that uses a standard Bayesian bootstrap at the resampling of the clusters. We show that this method also produces draws from the posterior predictive distribution of the population that incorporate both clustering and weighting components of the sample design while requiring only the final weights that are usually supplied in public databases. Their performances are evaluated under different population models and different degrees of clustering effects. Small and large sample behaviors of alternative methods are also investigated. While the framework of fully parametric MI does not seem to provide a direct and robust technique to deal with sample weights in hierarchical models, the proposed method turns out to be an easy alternative that recovers most of the information in the data generating mechanisms. We apply the different MI techniques to the analysis of passenger vehicle injury data from the National Automotive Sampling System – Crashworthiness Data System (NASS-CDS) survey, in particular, estimates of mean "delta-v" (instantaneous deceleration velocity) and its associated injury.

Chapter 4 develops a general purpose MI approach to account for various sample design features in a highly stratified multistage sample. Our specific focus is on evaluating the performance of the methods in comparison with existing MI techniques, with respect to several frequently encountered yet not previously addressed issues in the statistical analysis of missing data. These include: (i) accommodating stratification and

multi-stage sampling in the imputation process; (ii) the employment of logistic imputation models for estimating probabilities of rare events; and (iii) the estimation of population quantiles with multiply imputed data. In the simulation study, the proposed procedures demonstrate fairly good coverage properties for nonsmooth statistics over the entire support of the distribution for a continuous variable of interest, and yield quite stable parameter estimates in the case of sparse data. We argue that the proposed methods offer a computationally feasible solution to problems that are not well handled by current MI techniques. This method is applied to accommodate missing body mass index (BMI) data in the analysis of BMI percentiles using NHANES III data.

Chapter 5 summarizes findings in the thesis and points out both the advantages and limitations of the proposed methodology. We also point to several directions of future research, including the applications of the proposed methodology to domain/small area estimation, adaptations of it to unit nonresponse and to propagation of uncertainty in unit nonresponse adjustments, and possible extensions of it to incorporate auxiliary information.

CHAPTER 2
A TWO-STEP SEMIPARAMETRIC METHOD TO ACCOMMODATE SAMPLING WEIGHTS IN MULTIPLE IMPUTATION

2.1 Introduction

Both item nonresponse and sampling weights are typical features of survey data obtained from complex sample designs. Item nonresponse occurs when some respondents do not answer all the items in a survey questionnaire. Both "don't know" and refusal answers are considered as item nonresponse. Sampling weights arise as a correction factor to compensate for over- or under-representation of units in the target population due to unequal selection probabilities. Examples of unequal probability sampling designs include: i) disproportionate stratified sampling, where sampling units in each stratum have differential selection probabilities, and ii) probability proportional to size (PPS) sampling, in which the probability of selection for a sampling unit is proportional to a positive size measure (Z) known for all population units. For example, the Behavior Risk Factor Surveillance System (BRFSS) has both a substantial amount of missing data on income measures as well as survey weights that adjust for oversampling of adults in smaller sized households and for selection bias by poststratifying and raking to known control totals for basic demographics.

When the amount of item-level missing values is nontrivial and the data are not missing completely at random (MCAR), typical solutions for missing data like the

complete case analysis often lead to reduction of statistical power, and problematic inferences of the target population. Multiple imputation (MI) (Rubin, 1987, 1996) is a more principled method for addressing item-level missing data, which has a Bayesian conceptualization. The basic idea is to fill in missing data with *M* sets of plausible values. These values are obtained as repeated draws from the posterior predictive distribution of the missing components of the sample $Y_{s,mis}$ given its observed components $Y_{s,obs}$, i.e. $p(Y_{s,mis} \mid Y_{s,obs})$, where $p(\cdot)$ denotes the probability density function. (Typically *M* is small, e.g. *M*=3~5, but larger *M* (10~20) may be needed to obtain stable estimates of variance when the fraction of missing information is large.) The production of multiple "completed" datasets $\left\{(Y_{s,obs}, Y_{s,mis}^{(1)}),...,(Y_{s,obs}, Y_{s,mis}^{(M)})\right\}$ is typically done by an "imputer" who has access to the data to develop reasonable models for generating the predictive distribution of $Y_{s,mis}$, allowing the "analyst" to then analyze each of the *M* imputed datasets, and combine the point and variance estimates using the combining rules developed by Rubin (1987). Examples of this approach include imputation for blood alcohol concentration in the Fatal Accident Reporting System (FARS) (Heitjan & Little, 1991) and income imputation in the National Health Interview Survey (NHIS) (Schenker et al., 2006).

The implementation of multiple imputation typically assumes missing at random (MAR), that is, given the observed data, the reason for the missing data does not depend on the unobserved data. To make MAR plausible, it is important to let the imputation model condition on all variables (including sample design variables) that are either predictive of the outcome (*Y*) or the missing data mechanism (*R*). In settings where the observed data are obtained using unequal probability sampling design, however, data are

typically imputed using models where variables are assumed to be independent and identically distributed (IID). They are then analyzed using a design-weighted approach that accounts for the unequal selection probability. This can lead to biased point estimation and below-nominal confidence interval coverage.

A simple and seemingly straightforward way to incorporate sampling weights in MI is using fully parametric techniques. One option is to require the imputer's model to be conditioned on a few key design variables that determine the individuals' probabilities of inclusion, such as measure of size and stratification variables (e.g. demographics, socioeconomic status, as well as geographical characteristics, etc.). Another option is to summarize the design information by using weights as a covariate in the imputation, perhaps after log transformation or categorization in "weight strata" and modeling them as dummy indicators. However, the modeling task may be complicated by attempts to include all interactions of weights (or weight-related design variables) with other covariates in the model, particularly the interaction of the weights with domain indicators (Meng 1994; Kim et al. 2006; Seaman et al. 2006). Moreover, this approach typically requires the functional form of the interaction to be modeled correctly, perhaps using a spline or other non-parametric form to be robust against model misspecification (Elliott & Little 2000; Zheng & Little 2005; Breidt, Claeskens, & Opsomer, 2005). It can be a challenging task to come up with a robust imputation model that is attentive to sampling weights and sufficiently captures all relevant aspects of the distribution of Y of interest.

This chapter develops a modified MI framework to account for sampling weights from single-stage sampling designs. The primary goals are:

1) to propose a two-step MI procedure. In the first step, nonparametric models

are used to generate the posterior predictive distribution of the population that includes the item-level missing data. In the second step, parametric models assuming IID are used to impute the missing values in the created populations. We consider utilizing the weighted finite population Bayesian bootstrap (weighted FPBB) to account for sampling weights in the first step;

2) to illustrate the impact, during the imputation process, of ignoring sampling weights or unequal probabilities of selection, on the bias properties of both the MI point estimator and the MI variance estimator; and

3) to compare the performances of the proposed two-step MI and the fully parametric MI in terms of robustness to different degrees of model misspecification. The comparison will be made under different scenarios, defined by the associations of the design variable (Z) used for determining the selection probabilities with both the outcome variable (Y) where missing data occur, and the latent variable (T) which defines the response mechanism.

The rest of this chapter is organized as follows. Section 2.2 provides a detailed overview of the proposed two-step semiparametric multiple imputation procedure. (We term it "semiparametric" because the design feature, in particular the weights, are accommodated non-parametrically, whereas the actual imputation is conducted under a standard parametric model.) Section 2.3 discusses point estimation and inference using the MI datasets from the proposed procedure. Section 2.4 provides a simulation study in the context of a single-stage probability-proportional-to-size (PPS) sample design. We estimate population means and regression coefficients under settings where sampling weights are associated to differing degrees with both the outcome and the probability of

nonresponse, and where failure to account for design in the imputation procedure has differing degrees of impact. Section 2.5 applies the proposed procedure to estimate means, linear and loglinear regression models, and general location models describing marginal and joint distributions of income and health insurance accessibility. Section 2.6 discusses possible extensions to incorporate other design features beyond sampling weights, as well as extensions to deal with unit non-response weights.

2.2 A Two-Step Semiparametric MI Procedure

2.2.1 Overview and Notation

Bayesian finite population inference (Ericson 1969) has been proposed as a means to harmonize design- and model-based approaches for sample survey inference (Little 2004, 2011; Gelman 2007). Under this approach, we focus on the posterior predictive distribution of the finite population quantity of interest (e.g., population mean, population regression parameter) obtained from the posterior predictive distribution for the non-sampled elements of the population. To make matters more concrete, consider the setting where we have a scalar outcome *Y*, sampling weight *w* based on a single-stage PPS sampling design, and no missing data. Our complete data consist of the vector of sampling indicators *I* for the population, sampled Y_s for which $I=1$, the non-sampled Y_{ns} for which $I=0$, and similarly w_s and w_{ns}. Given the sampling weights, the sampling mechanism generating *I* is assumed to be ignorable ($p(I \mid Y, w) = p(I \mid w)$), and can be ignored in the modeling. Assuming a model for the outcome given the sampling weights $p(Y \mid \theta, w)$ parameterized by θ with prior $p(\theta)$, the posterior predictive distribution for the non-sampled elements of the population Y_{ns} is given by

$$p(Y_{ns} \mid Y_s, w_s) \propto \int p(Y_{ns} \mid Y_s, \theta, w)\, p(\theta \mid Y_s, w)\, p(w_{ns} \mid w_s)\, d\theta\, dw_{ns} \qquad [2.1]$$

Previous work has tackled estimation of this predictive distribution in a variety of manners. Zheng and Little (2003, 2005) and Chen, Elliott and Little (2010) assumed that the sampling weights were known for all subjects, so that $w_s = w$, reducing [2.1] to $p(Y_{ns} \mid Y_s, w) \propto \int p(Y_{ns} \mid Y_s, \theta, w)\, p(\theta \mid Y_s, w)\, d\theta$; these authors then obtained draws from the posterior predictive distribution under fairly weak modeling assumptions (parametric regression model for $p(Y \mid \theta, w)$ based on penalized splines). Little and Zheng (2007) and Zangeneh, Keener, and Little (2011) considered the situation in which weights (or equivalently the size measure *Z*) are observed only for the sample (as in a public use data setting), and obtained predictive draws for $p(w_{ns} \mid w_s)$ under a Dirichlet model with a non-informative (Haldane) prior; the resulting predictive draw of the population of weights w was then used as covariates in Little and Zheng to obtain posterior predictive draws of Y_{ns}. Dong, Elliott, and Raghunathan (2014) consider a different factorization of [2.1]:

$$p(Y_{ns} \mid Y_s, w_s) \propto \int p(Y_{ns}, w_{ns} \mid Y_s, w_s)\, p(Y_s, w_s)\, dw_{ns} \qquad [2.2]$$

The parameter θ is dropped because the draws of $p(Y_s, w_s)$ are made directly from the posterior of the empirical joint CDF of Y_s, w_s using a Bayesian bootstrap (BB) procedure (Rubin 1981). Draws from $p(Y_{ns}, w_{ns} \mid Y_s, w_s)$ are then made using a weighted finite population Bayesian bootstrap (FPBB) procedure described in Cohen (1997).

Here we extend the approach of Dong, Elliott, and Raghunathan to accommodate missing data. We assume that, had we taken a census of the entire population, we could have observed a vector of response indicators $R = (R_s, R_{ns})$, where R_s corresponds to the response indicators observed in the sample, and R_{ns} to the response indicators associated

with the non-sampled elements. We then divide the sampled $Y_s = (Y_{s,obs}, Y_{s,mis})$ into the fully-observed and missing elements, corresponding to the sampled *Y* values associated with $R_s = 1$ and $R_s = 0$ respectively, and similarly the non-sampled $Y_{ns} = (Y_{ns,obs}, Y_{ns,mis})$ into those that would have been observed had they been sampled ($R_{ns} = 1$), and those that would have had missing values ($R_{ns} = 0$). We also assume a fully-observable covariate $X = (X_s, X_{ns})$ consisting of the sampled and nonsampled elements respectively. Note that we can combine the observed from the sampled and nonsampled parts of the population to obtain the potentially "observable" $Y_{obs} = (Y_{s,obs}, Y_{ns,obs})$, and similarly $Y_{mis} = (Y_{s,mis}, Y_{ns,mis})$. We assume ignorable missingness, so that $p(R|Y,X,w) = p(R|Y_{obs},X,w)$, allowing *R* to be ignored in the model along with *I*. Extending [2.2] to incorporate item-level missingness (and thus the covariate *X* used for imputation) then yields

$$p(Y_{ns}, X_{ns} | Y_s, X_s, w_s) = p(Y_{ns,obs}, Y_{ns,mis}, X_{ns} | Y_{s,obs}, Y_{s,mis}, X_s, w_s) \propto \int p(Y_{ns,obs}, Y_{ns,mis}, X_{ns}, w_{ns} | Y_{s,obs}, Y_{s,mis}, X_s, w_s) p(Y_{s,obs}, Y_{s,mis}, X_s, w_s) dw_{ns} \quad [2.3]$$

To obtain the posterior predictive distribution based on fully-observed data, we need to integrate the left side of [2.3] with respect to $Y_{mis} = (Y_{s,mis}, Y_{ns,mis})$.

$$p(Y_{ns,obs}, X_{ns} | Y_{s,obs}, X_s, w_s) \propto \int p(Y_{ns,obs}, X_{ns} | Y_{mis}, Y_{s,obs}, X_s, w_s) p(Y_{mis} | Y_{s,obs}, X_s, w_s) dY_{mis}$$

To accomplish this, we reintroduce a parametric model for $p(Y|\theta, X, w)$, and integrate out with respect to the posterior distribution of θ:

$p(Y_{mis} | Y_{s,obs}, X_s, w_s) = \int p(Y_{mis} | Y_{s,obs}, X_s, \theta, w_s) p(\theta | Y_{s,obs}, X_s, w_s) d\theta$. Thus [2.3] becomes

$$p(Y_{ns,obs}, X_{ns} | Y_{s,obs}, X_s, w_s) = \iint p(Y_{ns,obs}, X_{ns} | Y_{mis}, Y_{s,obs}, X_s, w_s) p(Y_{mis} | Y_{s,obs}, X_s, \theta, w_s) p(\theta | Y_{s,obs}, X_s, w_s) d\theta dY_{mis} = p(Y_{ns,obs}, X_{ns} | Y_{s,obs}, X_s, w_s) \iint p(Y_{mis} | Y_{s,obs}, X_s, \theta, w_s) p(\theta | Y_{s,obs}, X_s, w_s) d\theta dY_{mis} \quad [2.4]$$

where the last equality in [2.4] follows from the fact that the unobserved (but potentially

observable) elements of the population are generated using the finite population Bayesian bootstrap.

A key result from [2.4] is that, since the unobserved elements of the population have been generated non-parametrically, we can implement the integration in [2.4] by use of a Gibbs sampler that iterates between obtaining draws of θ conditional on the entire population, and draws of the missing data conditional on θ and the observable elements of the population

$$p(Y_{mis} \mid Y_{obs}, X, \theta, w_s) = p(Y_{mis} \mid Y_{obs}, X, \theta)$$
$$p(\theta \mid Y, X, w_s) = p(\theta \mid Y, X)$$

Since the nonparametric procedure generates draws from the joint posterior distribution of all variables for the nonsampled population, the relationships of the weights and other variables are maintained in the draws. It is thus sufficient to develop a parametric model for *Y* that does not involve the weights: $p(Y \mid \theta, X, w) = p(Y \mid \theta, X)$. This model does need to condition on weights, however, when the weights are further associated with the missing data mechanism.

Figure 2.1 shows the creation of a single imputed synthetic population dataset under the proposed two-step MI procedure: a) shows the original sample data, b) the result from the BB-weighted FPBB procedure, and c) the result from the (model-based) imputation procedure. The shaded area represents observed data and '?' represents missing data. We discuss in detail the derivation and implementation of the proposed two-step MI procedure below.

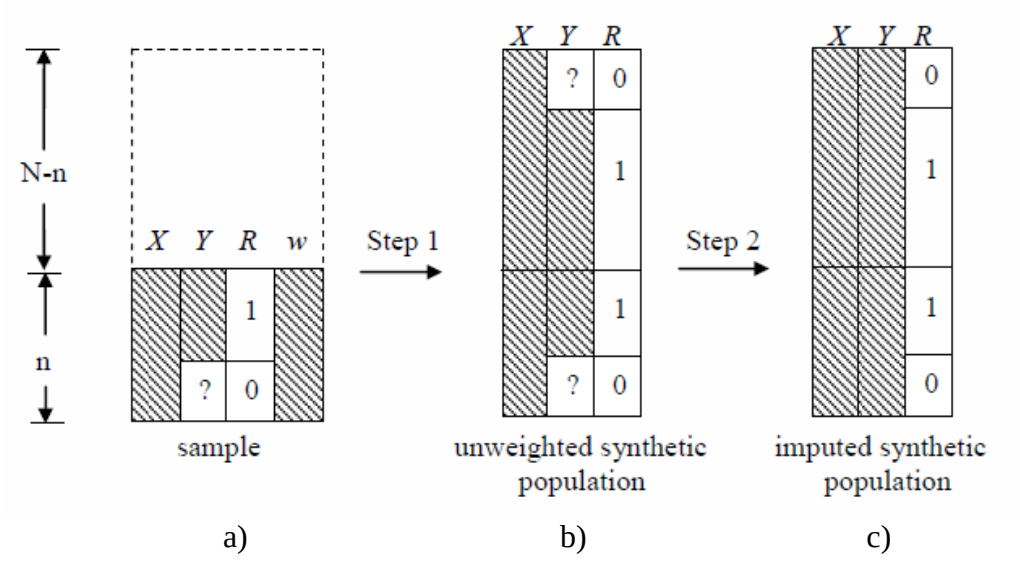

Figure 2.1 The procedure to create a single imputed synthetic dataset (Y: outcomes with item missing data; X: complete covariate; R: response indicator; w: sampling weight)

2.2.2 Step 1: Undo Sampling Weights through Synthetic Data Generation

The Pólya's Urn Scheme

The Polya urn distribution (Feller, 1968) is defined by construction as follows: suppose we have an urn containing a finite number *n* of balls of different colors. A ball is randomly drawn from the urn and another ball with the same color from outside of the urn is added back to the urn along with the originally picked one. Repeat this selection process until *m* balls have been selected; the resulting sample is termed a 'Pólya sample of size *m'*.

The Pólya Posterior

The Polya posterior (Ghosh & Meeden, 1997) derives its name from the Polya urn distribution. It is a noninformative Bayesian procedure which can be used when little or no prior information is available. One advantage of the Polya posterior is that it has a stepwise Bayes justification (Hsuan, 1979) and leads to admissible procedures.

Assume that a simple random sample of size *n* is drawn from a finite population

of size *N*. Let $y_s = \{y_1, ..., y_n\}$ denote the sample, where *y* represents the realized value of a response variable *Y*. Let $\{d_1, d_2, ..., d_K\}$ denote the set of *K* distinct values in the sample and $\lambda = \{\lambda_1, \lambda_2, ..., \lambda_K\}$ the vector of probabilities, such that $\Pr(y_i = d_j \mid \lambda) = \lambda_j$, for $i = 1, 2, ..., n,\ j = 1, ..., K,$ and $\sum_{j=1}^{K} \lambda_j = 1$. Let n_j and $n_j^{'}$ be the number of units taking value d_j in the sample and in the nonsampled part of the population, respectively, for $j = 1, 2, ..., K,$ and $\sum_{j=1}^{K} n_j = n,\ \sum_{j=1}^{K} n_j^{'} = N - n.$ Assuming a noninformative Haldane prior of λ: $\lambda \sim Dir(0, ..., 0)$, $p(\lambda) \propto \prod_{j=1}^{K} \lambda_j^{-1}$, together with a multinomial distribution for the counts of sample data: $n_1, .., n_K \mid \lambda \sim Mult(n; \lambda)$,

$p(n_1, ..., n_K \mid \lambda) = \dfrac{n!}{\prod_{j=1}^{K} n_j!} \prod_{i=1}^{K} \lambda_j^{n_j} \propto \prod_{j=1}^{K} \lambda_j^{n_j}$, yields a Dirichlet posterior distribution of λ:

$\lambda \mid n_1, ..., n_K \sim Dir(n_1, ..., n_K)$, $p(\lambda \mid n_1, ..., n_K) \propto \prod_{j=1}^{K} \lambda_j^{n_j - 1}$. The Posterior predictive distribution of counts in the nonsampled data thus follows a compound multinomial distribution $n_1^{'}, ..., n_K^{'} \mid n_1, ..., n_K \sim Mult(N - n; \lambda)$:

$$
\begin{aligned}
p(n_1^{'}, ..., n_K^{'} \mid n_1, ..., n_K) &= \frac{\int_0^1 ... \int_0^1 \prod_{j=1}^{K-1} p(n_1^{'}, ..., n_K^{'} \mid n_1, ..., n_K, \lambda)\, p(n_1, ..., n_K \mid \lambda)\, p(\lambda) d\lambda_1 ... d\lambda_{K-1}}{\int_0^1 ... \int_0^1 \prod_{j=1}^{K-1} p(n_1, ..., n_K \mid \lambda)\, p(\lambda) d\lambda_1 ... d\lambda_{K-1}} \\
&= \frac{\int_0^1 ... \int_0^1 \prod_{j=1}^{K-1} \lambda_j^{n_j + n_j^{'} - 1} (1 - \sum_{j=1}^{K-1} \lambda_j)^{n_K + n_K^{'} - 1} d\lambda_1 ... d\lambda_{K-1}}{\int_0^1 ... \int_0^1 \prod_{j=1}^{K-1} \lambda_j^{n_j - 1} (1 - \sum_{j=1}^{K-1} \lambda_j)^{n_K - 1} d\lambda_1 ... d\lambda_{K-1}} \\
&= \frac{\prod_{j=1}^{K} \Gamma(n_j + n_j^{'}) / \Gamma(n_j)}{\Gamma(N) / \Gamma(n)},
\end{aligned}
\qquad [2.5]
$$

where $\Gamma(\cdot)$ denotes the gamma function.

Ghosh and Meeden (1997) show that this posterior predictive distribution is in fact a Pólya urn distribution for the probability of seeing $n_j^{'}$ balls with color d_j, for $j = 1, 2, ..., K$ in some specified order. They also discuss the close relationship between the Polya posterior and the Dirichlet process priors (Ferguson, 1973) and the Bayesian bootstrap (BB) (Rubin, 1981). The Polya posterior is operationally and inferentially equivalent to the finite population Bayesian Bootstrap (FPBB) of Lo (1988), which is just the BB adapted to finite population sampling problems.

The weighted Polya Posterior/weighted FPBB

Formula [2.5] can be generalized to the case when the sampled units bear different weights, i.e. when the realized sample is selected with unequal probabilities (Cohen, 1997). To be consistent with the specific PPS sample we are considering in this chapter, we adapt the notation as follows: denote the PPS sample as $(Y_s, X_s, w_s, R_s) = \{(Y_i, X_i, w_i, R_i), i = 1, ..., n.\}$, where $w_i = \sum_i^N Z_i \big/ nZ_i$ for size variable *Z* and sample and population sizes *n* and *N*, and $Y_i = Y_{i,obs}$ if $R_i = 1$ and $Y_i = Y_{i,mis}$ if $R_i = 0$. Let $\{\tilde{d}_1, \tilde{d}_2, ..., \tilde{d}_K\}$ denote the set of *K* distinct vectors of (Y_i, X_i, w_i, R_i) in the sample and $\varsigma = \{\varsigma_1, \varsigma_2, ..., \varsigma_K\}$ the vector of probabilities that $\Pr\left((Y_i, X_i, w_i, R_i) = \tilde{d}_j \mid \varsigma\right) = \varsigma_j$, for $i = 1, 2, ..., n,\ j = 1, ..., K,$ and $\sum_{j=1}^K \varsigma_j = 1$. Let n_j and $n_j^{'}$ be the number of units taking vector $\tilde{d}_j$ in the sample and in the nonsampled part of the population, respectively, for $j = 1, 2, ..., K,$ and $\sum_{j=1}^K n_j = n,\ \sum_{j=1}^K n_j^{'} = N - n.$ For convenience, assume that all sampled units have distinct vector of values, i.e. $K = n$. Let w_i denote the sampling

weight for the i^{th} unit in the sample, which is normalized to sum up to N, i.e. $\sum_{i=1}^{n} w_i = N$.

Then w_i can be thought as the posterior expectation of the count of units in the population that have the same value as the i^{th} unit in the sample. Again assuming a noninformative Haldane prior of ς: $\varsigma \sim Dir(0,...,0)$ together with multinomially distributed weighted counts in the data $p(w_1,...,w_K \mid \varsigma) \propto \prod_{j=1}^{K} \varsigma_j^{w_j}$ yields a Dirichlet posterior distribution of ς: $\varsigma \mid w_1,...,w_K \sim Dir(w_1,...,w_K)$,

$$p(\varsigma \mid w_1,...,w_K) \propto \prod_{j=1}^{K} \varsigma_j^{w_j - 1} \qquad [2.6]$$

The posterior predictive distribution of counts in the nonsampled data then follows a compound multinomial distribution with an adjusted parameter $\varsigma^* = \{\varsigma_1^*,...,\varsigma_K^*\}$, i.e. $n_1^{'},...,n_K^{'} \mid w_1,...,w_K \sim Mult(N-n;\varsigma^*)$:

$$p(n_1^{'},...,n_K^{'} \mid w_1,...,w_K) = \frac{\int_0^1 ... \int_0^1 \prod_{j=1}^{K-1} (\varsigma_j^*)^{w_j + n_j^{'} - 1} (1 - \sum_{j=1}^{K-1} \varsigma_j^*)^{w_K + n_K^{'} - 1} d\varsigma_1^* ... d\varsigma_{K-1}^*}{\int_0^1 ... \int_0^1 \prod_{j=1}^{K-1} (\varsigma_j^*)^{w_j - 1} (1 - \sum_{j=1}^{K-1} \varsigma_j^*)^{w_K - 1} d\varsigma_1^* ... d\varsigma_{K-1}^*}$$
$$= \frac{\prod_{j=1}^{K} \Gamma(w_j + n_j^{'}) / \Gamma(w_j)}{\Gamma(2N-n) / \Gamma(N)}, \qquad [2.7]$$

In Little and Zheng (2007), $\varsigma_j^* = C \cdot \varsigma_j \cdot (w_j - 1)$, for $j = 1,...,K.$, where C is a constant that satisfies $\sum_j \varsigma_j^* = 1$; In this thesis, we follow the weighted Polya urn sampling suggested by Cohen (1997) and use a formula shown in equation [2.10].

The Adapted-weighted FPBB method

The adapted-weighted FPBB consists of two stages. The first stage resamples the original sample using the standard Bayesian bootstrap assuming IID, and the second

stage reverses/undoes the sampling weights using the weighted FPBB. This two-stage algorithm is similar in spirit to the standard parametric Bayesian method, where the first stage is equivalent to drawing values of the parameter (ς) from its posterior distribution given the counts in sampled data ($n_1,...,n_K$) and the second stage draws the predicted counts in the nonsampled data ($n_1^{'},...,n_K^{'}$) given the drawn parameter. Note that in Little and Zheng (2007), the first stage is replaced by drawing the parameter directly from a Dirichlet posterior distribution given by [2.6]. The method is described as follows:

- *Resampling using the standard Bayesian Bootstrap (BB)*

The standard Bayesian Bootstrap of Rubin (1981) assuming IID is used to generate *L* replicate BB samples each of size *n*, i.e. $\left\{(Y_s^{(l)}, X_s^{(l)}, w_s^{(l)}, R_s^{(l)}), l = 1,..., L.\right\}$. This essentially generates the posterior of the empirical joint CDF (denoted by f) of all the variables in the population given their realized values in the sample data set. Or equivalently, the posterior distribution of the parameter vector ς is drawn given the sample, i.e.

$$\begin{aligned} & f^{(l)}\left(Y, X, w, R\right) | \left(Y_s, X_s, w_s, R_s\right) \Leftrightarrow \\ & \left(\varsigma^{(l)} | Y_s, X_s, w_s, R_s\right) \sim Dir(n_1,...,n_K) \\ & \text{for } l = 1,..., L., \text{where } \varsigma^{(l)} = \left(\varsigma_1^{(l)},...,\varsigma_K^{(l)}\right). \end{aligned} \qquad [2.8]$$

This stage captures the sampling variability. The uncertainty in the posterior draws of the parameter $\varsigma^{(l)}$ is reflected in the varying counts of distinct units in the original sample being selected in different replicate BB samples. Let $\alpha_l(i)$ denote the number of times unit *i* is selected in the l^{th} replicate BB sample, for $l = 1,..., L.$ We incorporate this source of uncertainty in computing "the l^{th} bootstrap weight for unit *i*",

i.e. $w_i^{(l)} = w_i \cdot \alpha_l(i)$, where w_i denotes the original sampling weight for unit *i*. Note that unequal inclusion probabilities still exist in these created BB samples. The bootstrap weights are carried forward as input weights to the next stage.

- *Undo Sampling Weight using the weighted Polya posterior/weighted FPBB*

The weighted Polya posterior in equation [2.7] is used to create *B* synthetic populations for each of the *L* BB sample obtained from the previous stage, i.e. $\left\{(Y_s^{(l)}, X_s^{(l)}, R_s^{(l)}), (Y_{ns}^{(lb)}, X_{ns}^{(lb)}, R_{ns}^{(lb)})\right\}$, for $b = 1,...,B, l = 1,...,L.$ Specifically, it draws predicted counts of the distinct nonsampled units given that of the BB sample, which simultaneously adjusts for unequal inclusion probabilities, i.e.

$$\left(n_1^{(lb)'},...,n_K^{(lb)'} \mid w_1^{(l)},...,w_K^{(l)}\right) \sim Mult(N-n;\varsigma^{(l)*}), \quad \text{for } b = 1,...,B, l = 1,...,L. \qquad [2.9]$$

This stage captures the variability due to "imputing" the nonsampled cases within the l^{th} BB sample (under the same posterior draw of the parameter $\varsigma^{(l)*}$). The distribution in Equation [2.7] does not lend itself to direct calculation and thus needs to be approximated using Monte Carlo simulation. Specifically, we apply a procedure suggested by Cohen (1997) to simulate the posterior predictive distribution of the counts in nonsampled part of the population through generating *B* synthetic populations for each of the *L* Bayesian bootstrap samples:

i) Take a Pólya sample of size $N-n,$ denoted by $(Y_{ns}^{(lb)}, X_{ns}^{(lb)}, R_{ns}^{(lb)})$ from the urn $(Y_s^{(l)}, X_s^{(l)}, R_s^{(l)})$. In this process, each $(Y_i^{(l)}, X_i^{(l)}, w_i^{(l)}, R_i^{(l)})$, for $i = 1,...,n.$ in the urn is selected with probability

$$\varsigma_i^{(l)*} = \frac{w_i^{(l)} - 1 + l_{i,k-1} \times (\frac{N-n}{n})}{N-n+(k-1)\times(\frac{N-n}{n})},\ k = 1, 2, ..., N-n+1. \qquad [2.10]$$

where $w_i^{(l)}$ is the bootstrap weight for the i^{th} unit in the l^{th} replicate BB sample, and $l_{i,k-1}$ is the number of selections of unit *i* up to *(k-1)*th selection, setting $l_{i,0} = 0.$

ii) Form the weighted FPBB synthetic population

$\mathrm{P}_{(b)}^{(l)} = \left\{(Y_s^{(l)}, X_s^{(l)}, R_s^{(l)}), (Y_{ns}^{(lb)}, X_{ns}^{(lb)}, R_{ns}^{(lb)})\right\}$ so that it has exact size *N*.

The heuristic interpretation of [2.10] for adjusting the selection probability based on the initial bootstrap weight is as follows. Let $k = 1, 2, ..., N-n+1,$ and $i = 1, 2, ..., n,$ before making any FPBB selection of nonsampled units in the population from the complex BB sample (which could also be seen as n balls with distinct colors in the original urn), i.e. when $k = 1$ and $l_{i,k-1} = l_{i,0} = 0,$ the probability of selecting unit i with weight $w_i^{(l)}$ is $\frac{w_i^{(l)} - 1}{N-n}$, where $w_i^{(l)}$ represents $w_i^{(l)}$ balls with value $\{Y_i^{(l)}, X_i^{(l)}, R_i^{(l)}\}$ in the whole population (hence $w_i^{(l)} - 1$ balls outside of the urn). As we proceed with the FPBB selection, we adjust this selection probability according to the number of times each unit among the *n* sampled units was selected during the FPBB procedure, with each unit now representing $(N-n)/n$ among the $(N-n)$ units to be selected during one FPBB whenever it is selected once. After each selection, the denominator of the probability function needs to be inflated to reflect the total units being represented during the whole FPBB selection so far, while the numerator also needs to be inflated to reflect the total units represented by unit i in the process. Therefore we obtain the probability form as in

formula [2.10].

The first step (i.e. the two-stage adapted-weighted FPBB algorithm) results in the following "unweighted" synthetic populations, where *L* and *B* are the numbers of datasets generated from first- and second-stage, respectively:

$P_{(b)}^{(l)} = (Y^{(lb)}, X^{(lb)}, R^{(lb)}) = (P_{(b)obs}^{(l)}, Y_{mis}^{(lb)}), b = 1,...,B, l = 1,2,...,L.$, where

$P_{(b)obs}^{(l)} = ((Y_{s,obs}^{(l)}, X_s^{(l)}, R_s^{(l)}), (Y_{ns,obs}^{(lb)}, X_{ns}^{(lb)}, R_{ns}^{(lb)}))$ and $Y_{mis}^{(lb)} = (Y_{s,mis}^{(l)}, Y_{ns,mis}^{(lb)})$ consist of the observed and unobserved data in the lb^{th} FPBB synthetic population dataset respectively.

2.2.3 Step 2: Multiply Impute Missing Data through Parametric Models

Now that we have undone the sampling design, we are ready to perform conventional MI under an IID assumption. Following the standard MI procedure or approximations such as SRMI (Raghunathan, Lepkowski, Van Hoewyk, & Solenberger, 2001), we obtain draws from the posterior predictive distribution $p(Y_{mis}^{(lb)} \mid \mathrm{P}_{(b)obs}^{(l)})$. Without the need to include weights in the imputation model due to an IID FPBB population generated from the previous step, our task can now be concentrated on correctly modeling the covariates as well as interactions among them whenever necessary. Note that the elimination of the weights from the FPBB population does not obviate the need to account for the weights in the imputation process, if the probability of selection (I) and non-response (R) are associated with each other (i.e., $p(R \mid Y_{obs}, X, w) \neq p(R \mid Y_{obs}, X)$). This step results in *M* imputed synthetic datasets for each of the $L \times B$ FPBB synthetic populations generated from the first step,

$P_{bM}^{l} = (P_{(b)1}^{(l)}, P_{(b)2}^{(l)}, ..., P_{(b)M}^{(l)}),\ for\ b = 1,2,...,B, l = 1,2,...,L.$

2.3 Point and Variance Estimates for the Two-Step MI Procedure

In the absence of missing data, Dong, Elliott and Raghuanthan (2014) showed that, conditional on $P^{syn} = \{P_{(1)}^{(1)}, ..., P_{(B)}^{(1)}, ..., P_{(1)}^{(L)}, ..., P_{(B)}^{(L)}\}$, i.e. the L synthetic populations obtained after B FPBB samples, the posterior predictive distribution of a scalar population statistic $Q(Y) \equiv Q$ is given by

$$Q \mid P^{syn} \dot{\sim} t_{L-1}(\bar{Q}_L, (1 + L^{-1})V_L), \qquad [2.11]$$

where $t_{L-1}()$ denotes t distribution with $L-1$ degrees of freedom, $\bar{Q}_L = \frac{1}{L}\sum_l \tilde{Q}^{(l)}$ and $V_L = \frac{1}{L-1}\sum_l (\tilde{Q}^{(l)} - \bar{Q}_L)^2$ for $\tilde{Q}^{(l)} = \lim_{B\to\infty} \frac{1}{B}\sum_b q^{(lb)}$. Here $q^{(lb)}$ is the estimate of Q obtained from the b^{th} FPBB synthetic population within the l^{th} Bayesian Bootstrap sample; in practice we estimate $\tilde{Q}^{(l)}$ by $\hat{Q}^{(l)} = \frac{1}{B}\sum_b q^{(lb)}$. The result follows immediately from Section 4.1 of Raghunathan, Reiter, and Rubin (2003), and is based on the standard Rubin (1987) multiple imputation combining rules, treating Y_{ns} as missing data and Y_s as observed data. The average "within" imputation variance is zero, since the entire population is being synthesized; hence the posterior variance of Q is entirely a function of the between-imputation variance, and the degrees of freedom is simply given by the number of BB samples. The result assumes that $E(q^{(lb)}) = Q$ – a result guaranteed by the adapted-weighted FPBB estimator – as well as a sufficiently large sample size for Bayesian asymptotics to apply.

Here we have the additional need to impute the missing data within each of the synthetic population datasets, yielding $P^{imp} = \{P_{(11)}^{(1)}, ..., P_{(1M)}^{(1)}, ..., P_{(B1)}^{(1)}, ..., P_{(BM)}^{(1)}, ..., P_{(BM)}^{(L)}\}$. However, the similar result holds as in [2.11]:

$$Q \mid P^{imp} \dot{\sim} t_{L-1}(\bar{Q}_L, (1+L^{-1})V_L) \qquad [2.12]$$

where as in the fully-observed sample data case $\bar{Q}_L = \frac{1}{L}\sum_l \tilde{Q}^{(l)}$ and

$V_L = \frac{1}{L-1}\sum_l (\tilde{Q}^{(l)} - \bar{Q}_L)^2$, but now $\tilde{Q}^{(l)} = \lim_{\substack{B\to\infty \\ M\to\infty}} \frac{1}{BM}\sum_b \sum_m q^{(lbm)}$, where $q^{(lbm)}$ is an estimate of Q obtained from the m^{th} imputation of the b^{th} synthetic population within the l^{th} Bayesian Bootstrap sample; in practice we estimate $\tilde{Q}^{(l)}$ by $\hat{Q}^{(l)} = \frac{1}{BM}\sum_b \sum_m q^{(lbm)}$. The result again is based on the standard multiple imputation combining rules, where now (Y_{ns}, X_{ns}, R_{ns}) and $Y_{s,mis}$ are missing data and $(Y_{s,obs}, X_s, R_s)$ is observed; the generation of the synthetic population again sets the within imputation variance to 0. We now require $E(q^{(lbm)}) = Q$, which implies that our imputation model for Y_{mis} is correctly specified (including any associations with probabilities of selection), as well as the standard sufficiently large sample size for the *t* approximation to be reasonable.

Note that our point estimator is the same as that derived in (Reiter, 2004), but our variance estimator differs. This is a result of the fact that Reiter must condition on a released sample from the synthetic data, whereas here we are conditioning on the entire synthetic population, substantially reducing the complexity of the analytical approximation to the posterior distribution of *Q*.

2.4 Simulation Study

A simulation study is designed to investigate the inferential properties of the proposed method. The two-step MI procedure is compared with the existing fully

parametric methods under a total of 4 simulation designs defined by crossing the following two factors:

(1) Associations of the probabilities of selection with the mechanism generating the data. We call the design 'outcome relevant' if the probabilities of selection are correlated with the outcome variable *Y*, otherwise we term it an 'outcome irrelevant' design.

(2) Associations of the probabilities of selection with the mechanism generating the missing values. We use 'MAR_X' (weight independent missingness) and 'MAR_X,W (weight dependent missingness) respectively to denote respective situations where the missing data mechanism is dependent on the fully-observed covariate *X* only, and where it depends on probabilities of selection as well as the covariate.

For each of the four simulation designs, we analyze the data using three imputation models with differing degrees of model misspecification corresponding to different amounts of design information incorporated into the imputation model. *Model 1* ignores weights altogether in the imputation process, which is a procedure typically adopted by survey practitioners. Under Model 1, we assume that the data are resulted from simple random sampling for the purposes of imputation. *Model 2* includes log(weight) as a scalar summary of design information in the imputation model. This is a simple way to compensate for the naïve SRS assumption in Model 1 by adding the weight variable as a separate predictor in the imputation model. Model 2 is relatively easy to implement in practice, and has been adopted by the MI project on NHIS missing income data (Schenker et al., 2006). *Model 3* includes both log(weights) and its interactions with the covariate in the imputation model. All three imputation models will

be tested with both the fully parametric MI method and the proposed two-step synthetic MI procedure. Table 2.1 illustrates the setup of the simulation design where the differing imputation models are nested within the four cells defined by the two association factors.

Table 2.1 Strength of association of the sampling weight with both the missingness and the outcome

Association with missingness (M)	**Association with outcome variable (Y)**	
	Low	High
Low	Irrelevant design, MAR_X	Relevant design, MAR _X
High	Irrelevant design, MAR _X,W	Relevant design, MAR _X,W

Imputation Model	
1	$Y \sim X$
2	$Y \sim X + \log(weight)$
3	$Y \sim X + \log(weight) + X * \log(weight)$

2.4.1 Description of the Study Design

The simulation design is described below:

Step 1. Population data generation:

The population involves three variables: the outcome variable *Y*, a covariate *X*, and a size variable *Z* based on which probability-proportionate-to-size without replacement (PPSWOR) sampling is conducted. We consider two versions of the outcome variable *Y*, one in which the outcome is associated with the covariate, the probability of selection, and their interaction (Y_1), and the other in which the outcome is associated only with the covariate (Y_2). The joint distribution of *Z*, *X*, and *Y* is given by:

$$\log Z \sim N(2,1)$$
$$X \mid Z \sim N(0.1 * \log Z, \sigma_x^2)$$
$$Y_1 \mid X, Z \sim N(0.2 * X + 0.6 * \log Z + 0.5 * X * \log Z, \sigma_{y_1}^2)$$
$$Y_2 \mid X, Z \sim N(0.2 * X, \sigma_{y_2}^2)$$

Thus (Y_1, X, Z) constitutes the "relevant design" population and (Y_2, X, Z) constitutes the "irrelevant design" population. Figure 2.2 shows the scatter plots of Y_1 and Y_2 versus the size measure *Z*. Both populations have size *N*=4,000. For each population, we drew 100 independent samples of size *n*=200 without replacement, with inclusion probability for the i^{th} unit $\pi_i = nZ_i / \sum_{j=1}^{N} Z_j$. We call the 100 PPSWOR samples "before deletion (BD) samples".

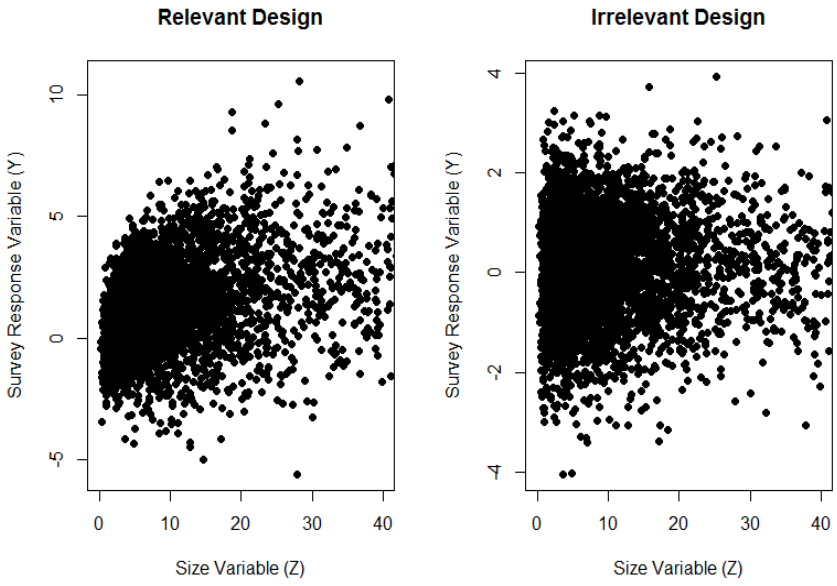

Figure 2.2 Scatter plots of survey variable *Y* versus size variable *Z*, under relevant and irrelevant designs, respectively.

Step 2. Impose missingness on the complete data under MAR mechanism:

Probit models are used as deletion functions to create missing data in the outcome

variable *Y* for each of the 100 simulations. Both *X* and *Z* are assumed to be completely observed. For both populations, we define two types of missing data mechanisms by generating two different latent variables T_1 and T_2 for the deletion function, which correspond to the MAR _X condition and MAR _X,W condition respectively. Specifically, $T_1 = -0.635 + 0.4X + e$ and $T_2 = -0.55 + 0.4X - 0.5\log Z + 0.4X * \log Z + e$, where $e \overset{iid}{\sim} N(0,1)$. The outcome is then missing if $T_j > 0$ (i.e., $P(M = 1 | T_j) = \Phi(T_j), j = 1, 2.$, where $\Phi(\cdot)$ corresponds to the standard normal CDF. This yields a moderate missingness fraction of 30% in all four designs corresponding to the four cells defined in Table 2.1.

Step 3. Impute missing data under different imputation models:

MI was performed separately for each of the 100 replication samples under each of the 12 simulation scenarios. All simulations were programmed using the R software (Version 2.14.2). In particular, the *mice* package (van Buuren & Groothuis-Oudshoorn, 2011) was used to implement the imputation, and the survey package was used for the design-based analyses under the fully parametric method (Lumley, 2004). For the proposed method, the *wtpolyap* function in package *polyapost* (Meeden & Lazar, 2012) may be used to create the weighted FPBB synthetic populations.

Step 4. Evaluation of performance:

We consider estimating the population mean of *Y* (i.e. $\bar{Y}$) and the population regression coefficients of *Y* on *X* : $Y = X^T\beta$, where $\beta = (\beta_0, \beta_1)^T$ and β_0 and β_1 are the intercept and the slope, respectively. We focus on five quantities to evaluate the performance of the two methods under comparison: bias, empirical root mean square error (RMSE), empirical interval coverage, empirical variance, and empirical estimated variance.

· *Bias.* Let $Q = (\bar{Y}, \beta_0, \beta_1)$ denote the set of the three population parameters we are interested in, and $q = (\hat{\bar{y}}, \hat{\beta}_0, \hat{\beta}_1)$ denote their corresponding point estimates (Pt.est). The raw bias of each of the parameter estimates *q* is calculated as the difference between its averaged value over the 100 simulations and the true parameter value *Q*.

$$Bias = E(q_r) - Q = \frac{1}{100}\sum_{r=1}^{100} q_r - Q$$

where q_r is the point estimate associated with the r^{th} simulation.

· *Empirical Root Mean Square Error (RMSE).* As a measure of accuracy which accounts for bias and efficiency simultaneously, RMSE is a key statistic we will consider in comparing the overall performance of the methods.

$$RMSE = \sqrt{MSE} = \sqrt{E(q_r - Q)^2} = \sqrt{\frac{1}{100}\sum_{r=1}^{100}(q_r - Q)^2}$$

· *Empirical interval coverage* (95% CI cov.). We report the actual coverage rates of a nominal 95% interval based on the 100 simulation repetitions under each simulation scenario. Normal approximation was used for inference under the proposed method.

· *Empirical Estimated Variance & Empirical Variance.* We calculated these two quantities and compared them with each other. They should be approximately equal if the variance estimator is unbiased.

$$Est.Var = E[v(q_r)] = \frac{1}{100}\sum_{r=1}^{100} V_{MI}(q_r),\ SE = \frac{1}{100}\sum_{r=1}^{100}\sqrt{V_{MI}(q_r)}$$

$$Emp.Var = Var(q_r) = \frac{1}{100-1}\sum_{r=1}^{100}[q_r - E(q_r)]^2, Emp.SE = \sqrt{Emp.Var}$$

2.4.2 Simulation Results

In deciding how many synthetic populations *B* are needed, we conducted a preliminary study based on the before deletion (BD) data. Equivalence of the BD

variance estimates to the actual sample variance estimates is desired. Simulation results are shown in Table 2.2. Note that the BD variance estimation under the synthetic method was calculated as the between synthesis variance, using the variance formula $(1+L^{-1})V_L$ in [2.11]. We observe that as we increase *B*, the variance estimate decreases, and stabilizes close to the actual sample variance for $B \geq 20$. Therefore, we use *B*=20 in the after deletion (AD) simulation.

Table 2.3 and Table 2.4 present the results from our simulation study. Each table is divided into two parts, containing the results from MAR _X condition and MAR _X,W condition respectively. Within each condition, we compare our new method with the fully parametric method. The three columns indicated by 'X','X,W' and 'XW' each correspond to the estimates under the three imputation models described in section 2.4.1.

We first examine the results in Table 2.3. When the design is relevant to the outcome variable *Y* yet uncorrelated with missingness (MAR_X), obvious advantages can be observed of the synthetic methods over the fully parametric method. For the fully parametric method to work properly under this design, the imputation model has to be correctly specified, otherwise all inferences based on this method are invalid---not only there is substantial bias associated with all three parameter estimates, but a corresponding disruption in coverage rates as well. This is particularly poor when the design is completely ignored in the model. For example, under the *X* only imputation model, the relative bias of the estimated mean and the slope is as high as $\frac{0.260}{1.343} = 19.4\%$ and $\frac{0.210}{1.435} = 14.6\%$, respectively, with 95% CI coverage rates as low as 79% and 67%, respectively. Including *W* as a main effect in the imputation model improves the

performance for estimating the mean but not the slope--the relative bias of the slope is $\frac{0.137}{1.435} = 9.5\%$ with only 79% coverage. These confirm the argument by Kim et al. (2006) and Seaman et al. (2011) that including survey weights alone in the imputation model does not guarantee valid inference if important interactions between the weight and other covariates are ignored. In contrast, our proposed method results in nearly unbiased estimates and actual coverage that is closer to the nominal level under all three models, regardless of the misspecification. Substantial gains in terms of RMSE over the model-based method are also consistently observed in all scenarios considered (e.g. 0.303 vs. 0.220 for the slope under *X* only model, and 0.201 vs. 0.185 under the interaction model). This indicates that the 'unweighting' procedure has actually played *dual roles* in the process: its effect is not limited to untying the unequal probability selection and saving the effort of design-weighted analyses afterwards, but it also captures the interactions between the design and the survey variable of interest. Thus ignoring the design in the imputation model does no harm at all.

With a relevant design that is also a correlate of missingness (MAR_X,W), the imputation model will still require the use of the design variable (here the weight) to maintain an ignorable missing data mechanism. The fully parametric method behaves similarly to the case where the design is associated only with *Y*: failure to include the weight in the imputation model substantially biases all of the estimators considered, while including the weight as a covariate corrects for bias in the mean and intercept estimator but not in the slope. The synthetic model partially corrects for these biases by providing a correct estimate of the population distribution in the presence of missing data; however, unless the imputation model is correctly specified, some biases remain.

However, the synthetic model still substantially reduces RMSE relative to the fully parametric approach for the mean and intercept estimator when the weight is ignored in the imputation model. It also has reduced RMSE when estimating the slope when the weight is included as a covariate but the interaction between the slope and the probability of selection is ignored. Otherwise, the synthetic and fully parametric approaches have similar RMSE properties. The synthetic model also has slightly conservative coverage properties, in contrast to the anti-conservative coverage of the fully parametric estimator when the model is misspecified for the estimator of interest.

With an outcome irrelevant design (Table 2.4), there are very slight effects on the estimates when compared across methods and models. Including the irrelevant design variable in the imputation model results in negligible biases and introduces some modest inefficiencies, consistent with the findings in Reiter et al. (2006).

Figures 2.3 and 2.4 examine the properties of MI variance estimators. Figure 2.3 shows scatter plots of 100 estimated standard errors (SEs) of the mean from alternative imputation methods versus the empirical SEs from actual samples before deletion under three different imputation models, with outcome relevant design. The MI variance estimates under the proposed synthetic method are consistently lower than the fully parametric method. The triangles (representing synthetic MI) are closer than the circles (representing parametric MI) to the 45 degree straight line, and the contrast is most obvious under model 1. Similar results are observed with outcome irrelevant design (not being plotted here). Figure 2.4 plots the standard error (SE) versus empirical standard error (Emp.SE) from alternative MI methods, under all 12 combinations of simulation design and imputation model, for both the mean and the slope. While the variance

estimator for the parametric MI method (represented by red dots) either over- or under-estimates when the imputation model is misspecified (i.e. under model 1 and model 2), the blue triangles are always close to the 45 degree line. This indicates the approximate unbiasedness of the variance estimator for the synthetic MI method as well as the robustness of it to model misspecification.

Table 2.2 Before deletion study of the effects of the number of generated FPBB populations (*B*) on variance estimate

Parameters Of Interest	Performance Criteria	Weighted FPBB Method with *B* Synthetic Populations Created								Actual Sample
		B=1	B=5	B=10	B=15	**B=20**	B=25	B=30	B=40	
Mean	Pt. est.	1.350	1.351	1.353	1.353	1.352	1.352	1.351	1.354	1.343
	Emp.Est.Var	0.051	0.037	0.036	0.035	0.035	0.034	0.034	0.035	0.035
	Emp.Var	0.034	0.034	0.034	0.034	0.035	0.034	0.033	0.033	0.035
	RMSE	0.185	0.184	0.185	0.184	0.186	0.182	0.182	0.181	0.186
	95% CI cov.	98%	94%	95%	95%	94%	95%	95%	95%	94%
Intercept	Pt. est.	1.064	1.063	1.064	1.064	1.064	1.063	1.063	1.063	1.058
	Emp.Est.Var	0.027	0.022	0.021	0.021	0.021	0.021	0.020	0.021	0.021
	Emp.Var	0.019	0.019	0.019	0.019	0.019	0.019	0.019	0.019	0.020
	RMSE	0.137	0.136	0.137	0.137	0.138	0.138	0.137	0.136	0.142
	95% CI cov.	98%	93%	94%	94%	93%	94%	92%	93%	92%
Slope	Pt. est.	1.451	1.454	1.453	1.453	1.455	1.451	1.453	1.453	1.435
	Emp.Est.Var	0.038	0.030	0.029	0.029	0.028	0.028	0.028	0.028	0.028
	Emp.Var	0.031	0.030	0.030	0.030	0.030	0.032	0.031	0.031	0.034
	RMSE	0.182	0.181	0.180	0.181	0.179	0.183	0.182	0.181	0.187
	95% CI cov.	93%	88%	91%	89%	89%	90%	90%	88%	89%

Table 2.3 Performance of the proposed method in contrast to the fully parametric method under the relevant design condition

Actual Parameters	Performance Criteria	MAR_X						MAR_X,W					
		Standard Parametric MI			Semi-Parametric MI			Standard Parametric MI			Semi-Parametric MI		
		X	X,W	XW	X	X,W	XW	X	X,W	XW	X	X,W	XW
Mean=1.343	Bias	0.260	0.068	0.002	0.003	0.009	0.012	0.211	0.032	0.000	-0.085	-0.082	0.008
	Emp.Est.Var	0.060	0.053	0.041	0.047	0.043	0.039	0.062	0.065	0.049	0.056	0.055	0.047
	Emp.Var	0.055	0.050	0.035	0.048	0.039	0.035	0.062	0.066	0.045	0.060	0.058	0.046
	RMSE	0.351	0.231	0.187	0.217	0.196	0.185	0.327	0.258	0.211	0.258	0.253	0.214
	95%Cov	79%	93%	97%	96%	95%	95%	84%	92%	96%	95%	92%	96%
Intercept=1.058	Bias	0.218	0.038	-0.001	0.004	0.003	0.007	0.181	0.003	-0.001	-0.050	-0.072	0.007
	Emp.Est.Var	0.031	0.030	0.027	0.027	0.026	0.025	0.032	0.033	0.032	0.031	0.030	0.030
	Emp.Var	0.028	0.031	0.020	0.027	0.024	0.021	0.027	0.031	0.026	0.030	0.029	0.027
	RMSE	0.274	0.180	0.142	0.165	0.154	0.143	0.243	0.174	0.160	0.180	0.183	0.163
	95% Cov	76%	92%	96%	93%	93%	96%	76%	92%	96%	94%	94%	97%
Slope=1.435	Bias	0.210	0.137	-0.028	-0.024	-0.001	0.010	0.155	0.165	-0.003	-0.145	-0.028	0.019
	Emp.Est.Var	0.040	0.032	0.036	0.039	0.035	0.034	0.038	0.036	0.039	0.049	0.043	0.040
	Emp.Var	0.048	0.041	0.040	0.048	0.040	0.034	0.054	0.046	0.037	0.054	0.044	0.036
	RMSE	0.303	0.244	0.201	0.220	0.199	0.185	0.278	0.269	0.191	0.274	0.210	0.188
	95% Cov	67%	79%	93%	92%	91%	92%	78%	82%	97%	91%	92%	96%

Table 2.4 Performance of the proposed method in contrast to the fully parametric method under the irrelevant design condition

Actual Parameters	Performance Criteria	MAR_X						MAR_X,W					
		Standard Parametric MI			Semi-Parametric MI			Standard Parametric MI			Semi-Parametric MI		
		X	X,W	XW	X	X,W	XW	X	X,W	XW	X	X,W	XW
Mean=0.106	Bias	0.011	-0.003	-0.004	0.011	0.008	0.006	-0.006	-0.025	-0.026	-0.002	-0.005	-0.013
	Emp.Est.Var	0.019	0.022	0.022	0.018	0.019	0.020	0.024	0.029	0.029	0.017	0.023	0.026
	Emp.Var	0.017	0.022	0.025	0.020	0.021	0.022	0.013	0.021	0.022	0.017	0.021	0.021
	RMSE	0.130	0.147	0.157	0.143	0.146	0.148	0.114	0.147	0.151	0.129	0.145	0.146
	95%Cov	97%	94%	90%	93%	94%	94%	97%	94%	95%	97%	95%	96%
Intercept=0.020	Bias	0.012	0.001	0.000	0.014	0.011	0.009	-0.001	-0.021	-0.023	0.002	-0.002	-0.009
	Emp.Est.Var	0.017	0.020	0.020	0.016	0.017	0.018	0.022	0.027	0.025	0.014	0.020	0.022
	Emp.Var	0.015	0.021	0.024	0.018	0.020	0.021	0.012	0.023	0.024	0.015	0.022	0.022
	RMSE	0.124	0.144	0.155	0.136	0.141	0.143	0.110	0.152	0.157	0.123	0.146	0.149
	95% Cov	99%	96%	93%	93%	94%	94%	98%	95%	97%	93%	94%	96%
Slope=0.416	Bias	-0.008	-0.017	-0.016	-0.010	-0.012	-0.014	-0.023	-0.022	-0.008	-0.020	-0.015	-0.014
	Emp.Est.Var	0.015	0.016	0.016	0.016	0.016	0.017	0.023	0.020	0.025	0.017	0.018	0.024
	Emp.Var	0.012	0.011	0.016	0.014	0.014	0.013	0.011	0.014	0.025	0.016	0.017	0.024
	RMSE	0.108	0.108	0.126	0.119	0.118	0.116	0.108	0.120	0.158	0.128	0.129	0.155
	95% Cov	96%	96%	96%	98%	98%	99%	98%	98%	92%	95%	95%	96%

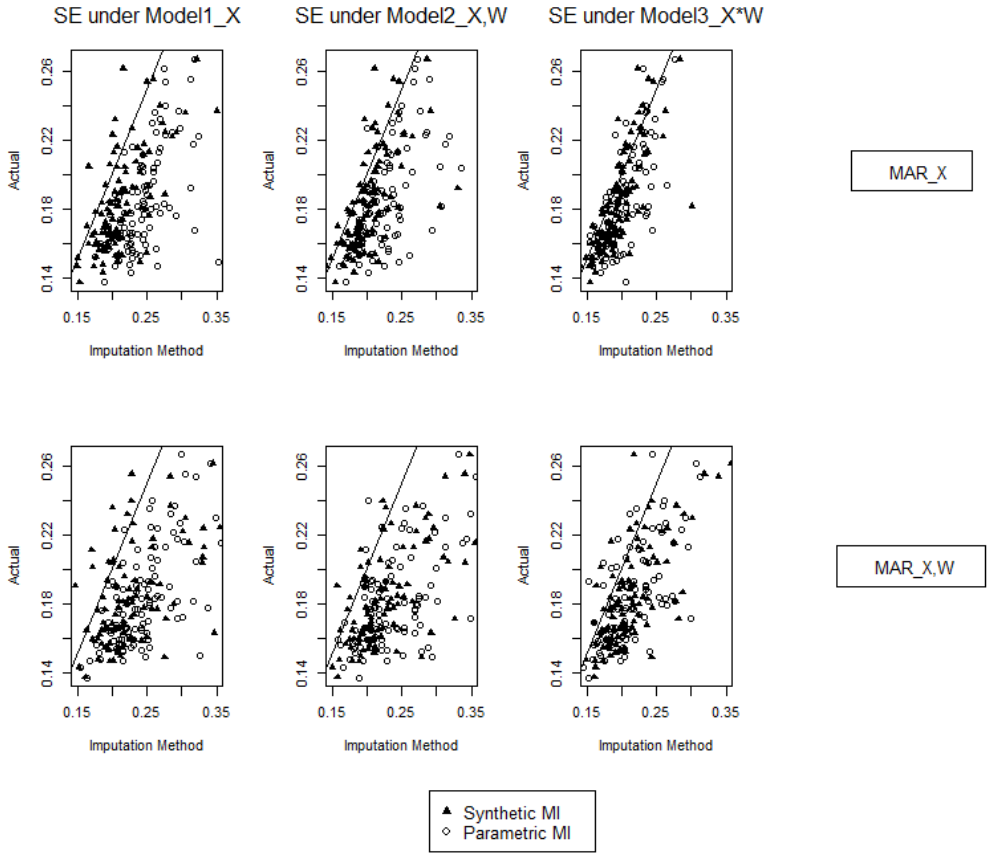

Figure 2.3 Scatter plots of 100 estimated standard errors (SEs) of the mean from alternative imputation methods (x axis) versus the empirical SEs from actual samples before deletion (y axis), under three imputation models, with outcome relevant design

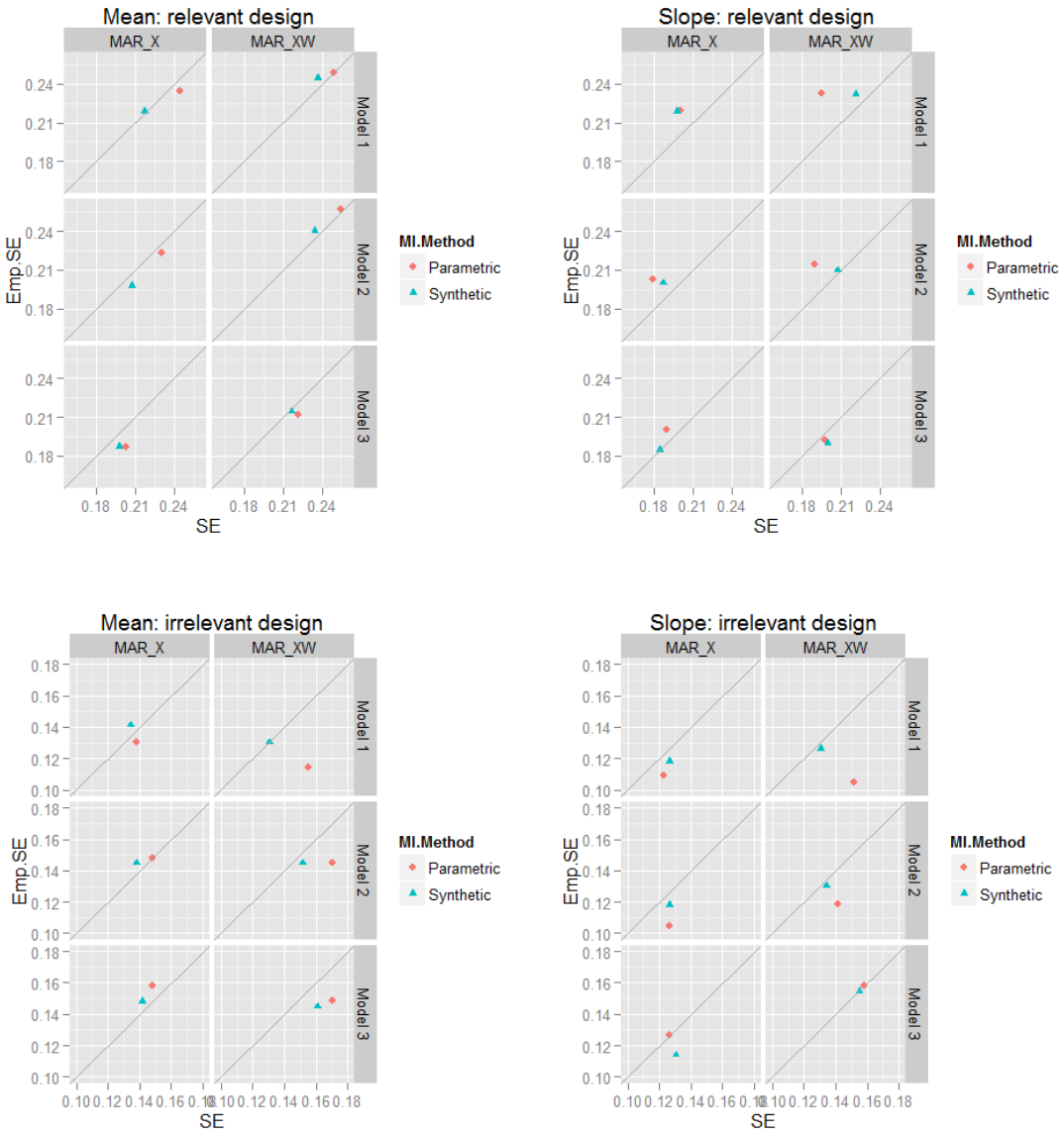

Figure 2.4 Plots of standard error (SE) versus empirical standard error (Emp.SE) from alternative MI methods, under different combinations of simulation condition and imputation model

2.5 Application to the Behavioral Risk Factor Surveillance System (BRFSS)

2.5.1 BRFSS Data

We next examine the effect of incorporating the survey weight in MI using data from one design stratum (n=388) of the 2009 Michigan BRFSS. This design stratum contains sampled households that belong to the medium-density (unlisted) telephone

numbers group. The BRFSS is a telephone survey conducted with a random sample of adults living in telephone-equipped households in the US. An independent sample of telephone numbers is used as the sampling frame; thus case weights are constructed to account for the fact that the probability of selection is proportional to the number of telephone lines and inversely proportional to the number of adults in a household. In addition, poststratification weights are used to adjust age-sex-race/ethnic distributions to Census totals. A mix of categorical and continuous variables is selected for analysis. These include health insurance coverage (yes/no), body-mass index (BMI) in kg/m^2, high blood pressure (yes/no), and five socio-demographic variables (age (in years), race (White vs. Nonwhite), annual household income (low: $\leq$ \$25,000, medium: \$25,000 - \$75,000, high: >\$75,000), gender (male/female), and employment status (yes/no/other) as well as a single design variable, the final weight. All survey variables except gender have certain degrees of missing data: income has the highest missing rate (16.5%), while others are missing 0%~6%.

2.5.2 BRFSS Imputation Method

We compare results from the fully parametric MI method with the proposed two-step semi-parametric MI method, using two imputation modeling strategies with each method: 1) assuming SRS, and 2) including the log of weights as a predictor in the model. We also include the weighted complete case analysis (CC). For the new method, we generated L=100 Bayesian bootstrap (BB) samples and created B=30 FPBB populations within each BB sample, with M=5 multiple imputations performed for each FPBB population. Since we do not know the population size in advance, and the individual final weights sum up to nearly 200,000 cases which is unrealistic to generate, we assume that N=4500 is large enough to be treated as a synthetic population. Since the degrees of

freedom is *L-1*=99, a normal distribution was used for inference.

2.5.3 BRFSS Analyses

We consider four different analyses: 1) the marginal distribution of income and health insurance accessibility (Table 2.5); 2) a linear regression model of BMI on key demographic variables as well as income (Table 2.5); 3) a loglinear model of a four-way contingency table defined by four categorical variables with no second-or-higher-order interactions (Table 2.6); and 4) a general location model for the joint distribution of the survey variables (Table 2.7).

Multivariate imputation by chained equations (*mice*) in R was used to impute the missing data under both MI methods. This technique is different from a joint modeling approach (Schafer, 1997a), which specifies a joint multivariate distribution for the missing data, this technique specifies the multivariate imputation model on a variable-by-variable basis by a set of conditional densities (Raghunathan et al., 2001; van Buuren & Groothuis-Oudshoorn, 2011). Therefore as a way to check the validity of the proposed method, we examine the joint distribution of all variables with missing data after imputation using a general location model (Olkin & Tate, 1961). However, due to sparse cells resulting from the complete cross-tabulation of all categorical variables in the original dataset, we are limited to looking at the joint distribution of BMI and age in a two-way contingency table defined by income (L,M,H) and gender (M,F). The parameters of interest are $\theta = (P, \Gamma_c, \Omega_c), c = 1,...,6.$: where *P* is a vector of length 6, i.e. the contingency table cell probability, and the bivariate normal distribution of BMI (X_1) and age (X_2) within each cell is: $X=\begin{pmatrix} X_{c1} \\ X_{c2} \end{pmatrix} \sim N_c(\Gamma_c, \Omega_c)$, where Γ_c and Ω_c denote the vector of means of BMI and age and the covariance matrix for the c^{th} cell, respectively.

We further collapse the medium and high income as one category for the loglinear analysis to avoid sparse cells.

2.5.4 BRFSS Results

Since the poststratification adjustment factor constitutes an important component of the final weight in BRFSS dataset, we presume that including the variables used to construct poststratification cells (age, race and gender in this case) in the imputation model should help in predicting the missing Y variable. A linear regression of final weights on age, sex, and race shows that these covariates explain 40% of the variance of the weights, suggesting that some unknown design variables exist that contributes to the construction of survey weights. Thus we conclude that imputation approaches that condition only on the available design variables will be insufficient to account for the sampling weights.

Table 2.5 shows that under the fully parametric MI method, including survey weights in the imputation model has a large impact on the regression coefficients of BMI on income and gender (high income: -0.47 when weights are excluded vs. -0.33 with log(weights) included; female: 2.72 vs. 2.56). In fact, these differences are particularly significant for domain estimation for whites (medium income: -2.8 vs. -2.1; female: -0.13 vs. -0.68). Under the new method, however, all estimates are similar to those from the fully parametric method with weights accounted for. Moreover, there is essentially no difference whether or not we incorporate weights into the imputation model after the sample data are synthesized. This means, that the new method can adjust for the weight effects at the synthesizing step without the need to model survey weights at the imputation step. Similar conclusions can be made from the Table 2.6 results for the log-linear model: modeling the weight in the imputation model has major impacts on the

coefficients associated with income. For example, Low income x Has insurance: -0.37 under ‘excludes weights parametric MI’ vs. -0.33 under ‘includes log(weights) parametric MI’ vs. -0.31 under both modeling strategies of ‘synthetic MI’.

As both BMI and age only have a very small fraction of missing data, we see little difference between the SRS imputation model and the model that includes log(weight) as a predictor under either MI method in the estimation of the general location model (Table 2.7), either for overall estimation or within income by gender categories. Thus we only display the results for model-based MI with weights accounted for (which is considered as the correct imputation model to use), together with the synthetic MI with the SRS model and we contrast both of them with the complete case analysis (CC). In general, two-stage semi-parametric MI provides results that are similar to the fully parametric MI method. Both seem to improve efficiency for overall estimates of BMI and age relative to those obtained from CC.

Table 2.5 Estimation of marginal distributions for income and health insurance, and linear regression coefficients for the regression of BMI on income, age and gender

			Methods									
Sample	Estimation	Variable	Complete Case		Parametric MI (M=5)				Synthetic MI (L=100, S=30, M=5)			
					Exclude weights		Include log(weights)		Exclude weights		Include log(weights)	
			Pt.est.	SE	Pt.est.	SE	Pt.est.	SE	Pt.est.	SE	Pt.est.	SE
Full Sample	Marginal	Low Income	0.50	0.04	0.50	0.04	0.52	0.05	0.52	0.04	0.51	0.04
		Medium Income	0.38	0.04	0.36	0.04	0.36	0.04	0.36	0.04	0.36	0.04
		High Income	0.12	0.03	0.14	0.03	0.12	0.03	0.13	0.03	0.13	0.03
		No insurance	0.22	0.04	0.24	0.04	0.24	0.04	0.24	0.04	0.24	0.04
	Regression	Intercept	27.0	2.75	26.1	2.02	25.8	2.05	26.3	2.29	26.2	2.30
		Medium income	0.47	1.40	0.35	1.21	0.39	1.19	0.37	0.94	0.37	0.95
		High income	0.27	1.43	-0.47	1.32	-0.33	1.37	-0.36	1.40	-0.31	1.40
		Age	0.02	0.04	0.03	0.03	0.04	0.03	0.03	0.03	0.03	0.03
		Female	2.29	1.30	2.72	1.06	2.56	1.07	2.57	1.06	2.55	1.05
Whites Domain	Marginal	Low Income	0.30	0.07	0.36	0.07	0.35	0.06	0.34	0.06	0.34	0.06
		Medium Income	0.53	0.08	0.48	0.07	0.50	0.07	0.49	0.06	0.49	0.06
		High Income	0.17	0.06	0.16	0.06	0.15	0.05	0.17	0.06	0.17	0.06
		No insurance	0.24	0.07	0.21	0.06	0.21	0.06	0.19	0.06	0.19	0.06
	Regression	Intercept	31.1	3.9	32.4	4.7	31.0	4.1	31.0	4.2	31.0	4.1
		Medium income	-1.6	3.25	-2.8	2.83	-2.1	2.72	-1.8	2.96	-1.7	2.97
		High income	-3.1	3.60	-3.5	3.42	-3.2	3.18	-3.1	3.65	-3.0	3.62
		Age	0.02	0.06	-0.01	0.06	0.02	0.06	0.02	0.06	0.02	0.05
		Female	-1.7	2.39	-0.13	2.13	-0.68	2.11	-0.80	2.17	-0.75	2.17

Table 2.6 Estimation of log-linear model for four categorical variables (collapse categories for medium and high income)

Estimation	Variable Level	Methods									
		Complete Case		Parametric MI Exclude weights		Parametric MI Include log(weights)		Synthetic MI Exclude weights		Synthetic MI Include log(weights)	
		Coef.	SE	Coef.	SE	Coef.	SE	Coef.	SE	Coef.	SE
Main effects	Low income	-0.01	0.12	0.08	0.13	0.04	0.12	0.04	0.13	0.02	0.13
	Has insurance	0.61	0.12	0.64	0.12	0.62	0.11	0.69	0.12	0.68	0.12
	White	-0.94	0.11	-1.0	0.10	-1.0	0.11	-1.1	0.12	-1.1	0.12
	Male	-0.11	0.12	-0.09	0.10	-0.07	0.10	-0.07	0.11	-0.07	0.11
Two-way Interactions	Low income x Has insurance	-0.36	0.12	-0.37	0.12	-0.33	0.13	-0.31	0.11	-0.30	0.11
	Low income x White	-0.28	0.10	-0.20	0.09	-0.20	0.09	-0.22	0.09	-0.22	0.09
	Low income x Male	-0.03	0.09	-0.03	0.09	-0.07	0.09	-0.05	0.08	-0.05	0.08
	Has insurance x White	-0.13	0.12	-0.02	0.12	-0.02	0.12	0.02	0.13	0.03	0.13
	Has insurance x Male	-0.01	0.13	-0.12	0.10	-0.14	0.10	-0.15	0.10	-0.15	0.10
	White x Male	-0.08	0.09	-0.11	0.08	-0.11	0.09	-0.11	0.08	-0.10	0.08

Table 2.7 Estimation of general location model for joint distribution of BMI, age, income and gender after MI under alternative methods (L_M = low-income male, M_M=medium-income male, H_M=high-income male; L_F = low-income female, M_F=medium-income female, H_F=high-income female)

Complete Case Analysis (design-based analysis)									
	Proportion	SE(p)	bmi	SE(bmi)	age	SE(age)	Var(bmi)	Var(age)	Cov(bmi, age)
L_M	0.230	0.045	27.2	1.86	38.7	3.81	57.4	301	-13.5
M_M	0.175	0.037	28.5	1.62	40.8	4.36	54.0	351	51.8
H_M	0.068	0.023	29.7	1.78	49.2	2.76	32.0	122	1.3
L_F	0.269	0.034	30.7	1.26	43.2	2.48	86.9	313	4.5
M_F	0.201	0.028	30.5	1.08	46.2	1.97	57.9	197	-7.2
H_F	0.058	0.014	28.9	1.04	48.1	3.62	23.6	241	-13.7
Overall	-	-	29.3	0.70	43.0	1.56	61.0	278	8.0
Model-Based MI log(weight) (design-based analysis)									
L_M	0.237	0.042	26.8	1.46	38.3	3.49	48.2	301	4.8
M_M	0.155	0.036	28.3	1.43	41.5	4.48	47.7	361	45.6
H_M	0.077	0.029	27.7	1.78	41.4	6.07	33.7	299	45.0
L_F	0.268	0.035	30.5	1.05	44.0	2.50	76.9	356	5.97
M_F	0.204	0.032	30.1	0.96	44.5	2.42	54.6	254	-2.4
H_F	0.060	0.017	29.0	1.31	45.2	4.73	24.8	262	-10.4
Overall	-	-	28.9	0.58	42.1	1.42	55.5	312	14.9
Synthetic MI SRS (simple unweighted analysis)									
L_M	0.241	0.037	27.1	1.47	38.7	2.92	45.0	275	-0.41
M_M	0.161	0.033	28.2	1.40	41.4	4.09	43.0	321	40.1
H_M	0.070	0.022	28.2	1.67	43.8	5.30	26.6	211	26.0
L_F	0.276	0.032	30.5	0.85	43.4	2.24	73.6	351	2.9
M_F	0.196	0.023	30.5	0.79	45.0	1.57	53.8	233	-4.2
H_F	0.056	0.013	29.0	1.00	46.8	3.22	25.5	231	-10.4
Overall	-	-	29.0	0.55	42.1	1.46	55.9	308	13.1

2.6 Discussion

Our primary goal was to propose a new method using the weighted finite population Bayesian Bootstrap to account for sampling weights in MI to deal with item-level missing data, and to evaluate the new method in a PPS sampling design setting. Our findings in the simulation study suggest that the new method does significantly reduce bias relative to the fully parametric methods, and with little loss in efficiency. Meanwhile, the weighted FPBB method maintains population-level multivariate relationships and potentially protects against model misspecification, for example, erroneous inclusion or exclusion of interactions between design variables and other covariates in the imputation model. A further advantage lies in that, unlike the fully parametric methods which include designs in the imputation model but still require complex survey packages to analyze the imputed datasets, the new method fully accounts for the unequal selection probabilities by unweighting them and restoring a population in a separate step. Therefore, only simple, unweighted complete-data analysis techniques are needed for inferences with the newly developed combining rules. This potentially allows a much wider variety of models to be considered with existing software, which, despite recent improvements, often cannot account straightforwardly for complex sample designs.

A limitation of the proposed method is the need to account for the design elements in the imputation procedure if the missingness mechanism requires their inclusion (e.g., if the probability of an item response is a function of the selection probability). As a practical matter, however, “undoing” the sample design to

generate the synthetic population may reduce the impact of misspecified missingness mechanisms by avoiding enhancement from misspecified data generation mechanisms.

The proposed two-stage semi-parametric multiple imputation approach has a number of possible extensions. The method developed here is designed for one-stage sample designs with independent selections and unequal probabilities of selection, as in the BRFSS sample design. Hence, extensions are required to account for multi-stage designs with clustering and stratification as part of the finite population Bayesian bootstrap. Another limitation is the assumption that no unit non-response occurs in the sample. Extensions that incorporate unit non-response provide another promising research opportunity. However, in public use samples that provide only final weights incorporating non-response adjustments, treating the final weight as a sampling weight (as we did in the BRFSS application) may be the only practical alternative.

CHAPTER 3
MULTIPLE IMPUTATION IN TWO-STAGE CLUSTER SAMPLES USING THE WEIGHTED FINITE POPULATION BAYESIAN BOOTSTRAP

3.1 Introduction

Survey data collected for social science and public health research is often clustered. Such clustered data often results from multi-stage sampling for cost and convenience reasons. For example, cluster sampling of students from schools is common in education surveys, where students are clustered by schools, or by classrooms within schools. In general population surveys, area probability sampling is considered a cost-effective way to select households because households are naturally clustered by geographic areas (e.g. PSUs/counties, blocks, etc.). In these and many other examples, it is realistic to assume that units in the same cluster tend to be more alike than units in different clusters. This leads to *clustering effects*, typically in the form of intraclass correlation (ICC) (Kish, 1965). Since a positive ICC implies an effective sample size less than the total sample, and thus increases variance estimates of statistics of interest, accounting for clustering effects is quite important.

Methods for doing this have been well-developed for complete-data approaches using either robust sandwich estimators based on Taylor Series approximations (Binder, 1993) or replication methods such as the balanced

repeated replication, jackknife, or bootstrap (Rust & Rao, 1996). These are now commonly available in statistical packages (e.g. the survey package in R, the svy suite of commands in Stata). A more model-based approach to the analysis of clustered data is multilevel modeling (Rabe-Hesketh & Skrondal, 2006), which allows the decomposition of overall variance estimates into components of variance due to different levels of sample selection. A good overview of software procedures for fitting linear mixed models is West & Galecki (2011); packages that fit nonlinear mixed models include R nlme, SAS nlmixed, and Stata gllamm (Li et al., 2011).

Concurrently with the development of methods and software for clustered sample designs, multiple imputation (MI) (Rubin, 1987) has become a standard method for dealing with item-level missing data in complex sample surveys. The need to incorporate sample design information into the imputation process has been recognized -- for example, Rubin (1996) in his MI review paper stated that any imputation model should *minimally* include major stratification and clustering indicators as well as design weights to make it proper. More generally, Schafer (1997a) also asserted that whether the estimation techniques ultimately applied to survey data are model-based or design-based, imputation models for missing values should incorporate important features of sample design. However, as this can be difficult to do in practice, imputation is typically performed under an assumption of simple random sampling.

This issue is considered in detail by Reiter, Raghunathan, and Kinney (2006). They attempt to incorporate sample design through simple modification

of existing imputation procedures such as SRMI in IVEware (Raghunathan et al., 2001), by modeling design variables (i.e. cluster membership indicators) as fixed effects in the imputation. Despite the simplicity of implementation, the fixed cluster effects model they propose is usually uncongenial to the analyst's model in the sense of Meng (1994). It can become very inefficient when the number of clusters is large. Andridge (2011) also considered fitting a fixed cluster effects imputation model to data from cluster randomized trials. However, when she used such a modeling strategy, she found an upward bias in the MI variance estimator through both analytical proof and simulation studies.

An alternative approach consistent with standard model-based approaches is to impute data using a mixed/random effects model, where each cluster has its own intercept (usually assumed to be sampled from a normal distribution with zero mean and unknown variance). Software packages allowing the use of mixed effects models for multiple imputation include: 1) the R *pan* package (Schafer & Yucel, 2002; Zhao & Schafer, 2013), for imputing continuous variables in clustered or panel designs under a multivariate normal model. Details on the MCMC imputation method used in *pan* can be found in Schafer (1997b); and 2) the *REALCOM-IMPUTE* module of the multilevel model fitting software MLwiN (Carpenter, Goldstein & Kenward, 2011). This also adopts a joint modeling approach and MCMC-based solutions but assumes a multivariate latent normal model with random effects for mixed type response variables subject to missingness. While these imputation models achieve congeniality with respect to the cluster effects and result in more efficient estimates than a fixed effect

approach, their use in practice is limited due to several shortcomings: *pan*, for example, assumes normality on binary variables, and the linear mixed effects imputation model typically leads to implausible values that needs rounding (to the nearest 0 or 1) after imputation. This can cause biased estimates of parameters (Horton et al., 2003). With *REALCOM-IMPUTE*, uncongeniality remains an issue in the sense that conditional relationships among variables in the imputation model are typically assumed linear while the models of interest usually include non-linear relationships and/or interactions. Besides, the consequences are not clear of using a probit regression for imputation for binary data that are modeled by the analyst using a logistic regression. Moreover, both procedures are confined to relatively simpler data structure with no more than two levels.

On another front, attempts to extend the variable-by-variable imputation methodology to clustered designs, e.g. the sequential hierarchical regression imputation models (*SHRIMP*) for multivariate clustered data (Yucel & Raghunathan, 2006; Zhao & Yucel, 2009; Yucel, 2011) demonstrate problems with convergence. There is evidence of poor finite-sample repeated sampling properties in logistic mixed effect models with high ICC. Other problems with the fully parametric imputation methods above include: 1) imputations drawn from these models may perform poorly if the random effects follow a non-normal distribution instead; 2) the accommodation of both sample weights and cluster effects simultaneously is particularly problematic, especially in high-dimensional settings with many missing covariates. Such an accommodation requires the estimation of interactions between clusters and weights, as well as between

weights and other quantities of interest (e.g., regression parameters). This can be difficult to achieve using a standard parametric approach.

Thus the focus of this chapter is to develop an adaptation of the two-step semi-parametric MI procedure proposed in Zhou, Elliott, and Raghunathan (2013a) for disentangling clustering effects and sampling weight effects using a weighted finite population Bayesian bootstrap (FPBB) procedure developed in Dong, Elliott, and Raghunathan (2014). The new MI procedure is designed to produce draws from the posterior predictive distribution of the population that incorporate both clustering and weighting elements of the complex sample design. Item-level missingness is incorporated in these "uncomplexed" synthetic populations; these missing data can then be imputed under an IID assumption without explicit modeling of clustering and weight effects. We consider two-stage unbalanced cluster samples obtained from unequal probability of selection, and we assume a missing at random (MAR) missing data mechanism on a single survey outcome variable (Y). The parameters of interest are the population mean of Y and population regression parameters of Y on a covariate X based on the multiply imputed data.

Section 3.2 discusses MI under simple random sampling, together with standard fixed and random effects models to incorporate cluster effects. Section 3.3 develops the newly proposed weighted FPBB procedure. Section 3.4 conducts and discusses a series of simulation studies designed to assess the repeated sampling properties of MI under the various approaches discussed in Section 3.2 and 3.3. These simulation studies use different population models, differing

numbers of clusters, differing degrees of intraclass correlation, and differing MAR mechanisms (dependent only on the fully observed covariate independent of sample weights, or dependent on both the covariate and sample weights). Section 3.5 applies the different MI procedures to the analysis of passenger vehicle injury data from the National Automotive Safety System – Crash Detection System (NASS-CDS) survey. Section 3.6 concludes with discussion and suggestions for next research steps.

3.2 Fully Parametric Imputation Methods for Clustered Sample Designs

3.2.1 Simple Random Sampling (SRS) Model

Let Y_i be the survey outcome with missing values (assumed scalar for ease of exposition) and X_i be a p-variate vector of other survey variables whose values are known for all cases in the sample, $i = 1,...,n$. We assume that Y is a member of the exponential family, and that the known transformation of $E(Y_i \mid X_i) = \mu_i$ can be modeled as a linear function of X: $g(\mu_i) = X_i^T \beta$, with $\text{Var}(Y_i \mid X_i) = \sigma_e V(\mu_i)$ for a known variance function V and a possibly estimated scale parameter σ_e. (Examples of this include Gaussian regression with $g(\mu) = \mu$, $V(\mu) = 1$, and σ_e as the [typically unknown] variance parameter; Poisson regression/log-linear model with $g(\mu) = \log(\mu)$, $V(\mu) = \mu$; and logistic regression with $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$, $V(\mu) = \mu(1-\mu)$, and $\sigma_e = 1$.) For now, we assume independence of the Y_i and equal probability of selection. A Gibbs Sampler can then be used to simulate the joint posterior of the unknown

parameters $\theta = (\beta, \sigma_e)$ and the missing data y_{mis}. Given starting values $\theta^{(0)}$ and $y_{mis}^{(0)}$, and let $f(\cdot)$ denote the probability density function, consider an iterative simulation algorithm in which the current version of the unknown parameters $\theta^{(t)} = (\beta^{(t)}, \sigma_e^{(t)})$ and the missing data $y_{mis}^{(t)}$ are updated in two steps:

(1) Draw model parameters from their posterior distributions conditional on the "filled-in data" (imputed plus observed) from the previous draw: $\theta^{(t+1)} \sim f(\theta \mid y_{obs}, y_{mis}^{(t)})$.

(2) Draw missing values in *Y* from its posterior given all other parameters drawn: $y_{mis}^{(t+1)} \sim f(y_{mis} \mid y_{obs}, \theta^{(t+1)})$.

After a sufficient number of draws have been obtained to attain convergence of the Gibbs Sampler (Gelman & Rubin, 1992), *M* widely separated draws of the missing data are combined with the observed data to form the completed data $y_{comp}^{(m)} = (y_{obs}, y_{imp}^{(m)}), m = 1, ..., M.$. Inference about the population quantity of interest *Q* is estimated using the combining rules of Rubin (1987):

$$\hat{Q} = M^{-1} \sum_{m=1}^{M} Q(y_{comp}^{(m)})$$

$$T^{1/2}(\hat{Q} - Q) \sim t_{v_m}$$

$$T = U + (1 + M^{-1}) V_B$$

$$U = M^{-1} \sum_{m=1}^{M} \hat{\text{var}}\left(Q(y_{comp}^{(m)})\right)$$

$$V_B = (M-1)^{-1} \sum_{m=1}^{M} \left(\hat{Q} - Q(y_{comp}^{(m)})\right)^2$$

$$v_m = (M-1)\left[1 + \frac{U}{(1 + M^{-1}) V_B}\right]^2$$

where $Q(y_{comp}^{(m)})$ is the point estimate obtained from the m^{th} completed dataset

$y_{comp}^{(m)}$, U is the within imputation variance calculated as the average of variance estimates based on the *M* completed datasets, V_B is the between imputation variance, v_m is the degrees of freedom associated with the *t* reference distribution for inference about *Q*.

3.2.2 Fixed Cluster Effects Model

We expand the simple random sampling model to $g(\mu_i) = X_i^T \beta + Z_i^T \gamma$, for $i = 1,...,n.$, where Z_i is a $(c-1)$ - dimension vector of dummy variables for the *c* clusters selected in the sample. Alternatively we can further expand to include interactions between the fully-observed covariates and the cluster dummy variables:

$g(\mu_i) = X_i^T \beta + Z_i^T \gamma + X_i^T \otimes Z_i^T \eta$. Imputation proceeds as in the SRS setting, with expanded parameter space $\theta = (\beta, \gamma, \eta, \sigma_e)$.

3.2.3 Random Cluster Effects Model

Here we group the variables by cluster, for the $i = 1,...,c$ clusters with $j = 1,...,m_i$ observations. Our imputation model then becomes $g(\mu_{ij}) = x_{ij}^T \beta + \omega_i$ where we assume that the random cluster effects are distributed as $\omega_i \sim N(0, \sigma_\omega)$ independently for $i = 1,...,c$. The unknown parameters in the model are now $\theta = (\beta, \sigma_\omega, \sigma_e)$ and ω_i. A Gibbs Sampler (for Gaussian regression imputations considered in this chapter) is then used to simulate the joint posterior of the unknowns. At current iteration *t*, each parameter is drawn from its respective conditional posterior distribution as follows:

$$\omega_i^{(t+1)} \sim f(\omega_i \mid y_{obs}, y_{mis}^{(t)}, \theta^{(t)}), i = 1, ..., c.$$
$$\theta^{(t+1)} \sim f(\theta \mid y_{obs}, y_{mis}^{(t)}, \tilde{\omega}^{(t+1)}), \tilde{\omega} = (\omega_1, ..., \omega_c)^T$$
$$y_{mis}^{(t+1)} \sim f(y_{mis} \mid y_{obs}, \tilde{\omega}^{(t+1)}, \theta^{(t+1)})$$

where ω_i is drawn from $N\left(\left(1+\sigma_e\sigma_\omega^{-1}\right)^{-1}\left(y_i - x_i\beta\right), \left(\sigma_e^{-1}+\sigma_\omega^{-1}\right)^{-1}\right)$; σ_ω is drawn from $\chi^{-2}_{\upsilon_2+c}$, σ_e is drawn from $\chi^{-2}_{\upsilon_1+n-p}$, where υ_1 and υ_2 are the prior specification of the hyper-parameters for the distribution of σ_e and σ_ω, respectively; β is drawn from $N\left(\left(\sum_{i=1}^{c} x_i^T x_i\right)^{-1}\left(\sum_{i=1}^{c} x_i^T\left(y_i - \omega_i\right)\right), \sigma_e\left(\sum_{i=1}^{c} x_i^T x_i\right)^{-1}\right)$; and finally the missing data $y_{j\in mis}$ are drawn from $N\left(x_{j\in mis}^T\beta + \omega_i, \sigma_\omega + \sigma_e\right)$.

3.3 Multiple Imputation using the Weighted Finite Population Bayesian Bootstrap in Clustered and Weighted Sample Designs

3.3.1 Overview

In this section, we develop a two-step multiple imputation methodology to account for complex sampling designs with two-stage cluster samples. The first step utilizes a weighted finite population Bayesian bootstrap to generate predictive draws of a population that capture the clustering and unequal probability of selection design features; the second step conducts the imputation within each of the predictive draws.

Define *I* as a (fully-observed) vector of sampling indicators for a target finite population of size *N*. To develop the population posterior predictive distribution generation for the first step, let $Y = [Y_s, Y_{ns}]$ be the survey outcome of interest where missing data occur, which divides into the sampled part Y_s (for

which $I=1$) and the nonsampled part Y_{ns} (for which $I=0$). Define separate response mechanisms for the sampled and the nonsampled part of Y: $R=(R_s, R_{ns})$, such that $Y_s=[Y_{s,obs}, Y_{s,mis}]$ further divides into the observed component $Y_{s,obs}$ and the missing component $Y_{s,mis}$, corresponding to the sampled *Y* values associated with $R_s=1$ and $R_s=0$ respectively, and similarly $Y_{ns}=[Y_{ns,obs}, Y_{ns,mis}]$ divides into those that would have been observed had they been sampled ($R_{ns}=1$), and those that would have had missing values ($R_{ns}=0$). Note that we can recombine as $Y=[Y_{obs}, Y_{mis}]$ where $Y_{obs}=[Y_{s,obs}, Y_{ns,obs}]$ and $Y_{mis}=[Y_{s,mis}, Y_{ns,mis}]$. In a similar fashion, let $X=[X_s, X_{ns}]$ be the complete covariate, and $Z=(w,C)=[(w_s,C_s),(w_{ns},C_{ns})]=[Z_s, Z_{ns}]$ be the complete design matrix which contains all the essential sample design features (here we restrict to sampling weights *w* and cluster membership indicators *C*). Note that *w* is a matrix $\{(w_{j(i)}, w_{j|i}, w_{ij}), i=1,...,n,\ j=1,...,m_i.\}$, where $w_{j(i)}$ is the cluster-level weight for element *j* in cluster *i*, $w_{j|i}$ is the element-level conditional weight for element *j* given cluster *i* is selected, and $w_{ij}=w_i w_{j|i}$ is the overall sampling weight for element *j*.

Assuming ignorable sampling, the joint predictive distribution of the nonsampled part of the population is given as (Rubin, 1987):

$$p(Y_{ns}, X_{ns}, Z_{ns}, R_{ns} \mid Y_s, X_s, Z_s, R_s, I) \propto p(Y_{ns}, X_{ns}, Z_{ns}, R_{ns} \mid Y_s, X_s, Z_s, R_s),$$
$$\text{given } p(I \mid Y, X, Z, R) \propto p(I \mid Y_s, X_s, Z_s, R_s). \qquad [3.1]$$

Explicit modeling for I can be avoided by incorporating design variables (*Z*) into a model for $p(Y_{ns}, X_{ns}, Z_{ns}, R_{ns} \mid Y_s, X_s, Z_s, R_s)$ that conditionally eliminates

dependence on *I* (Gelman, 2007; Little, 2011). However, the practical implementation of this in a robust manner using standard parametric models can be daunting. Here we pursue a nonparametric approach using a weighted FPBB that combines a Polya sampling scheme developed by Ghosh and Meeden (1997) in the simple random sampling setting and extended by Meeden (1999) into a simple balanced cluster design setting with a second extension by Cohen (1997) to accommodate weighted data. These constitute **a sample design reversing procedure** as the first step.

The second step requires **a correct imputation model** that generates the posterior predictive distribution of missing values in the entire population, such that it is independent of the response mechanism for both sampled and nonsampled parts of the population. That is, we do not need an explicit model for R given ignorable missingness:

$$p(Y_{mis} \mid Y_{obs}, X, Z, R) \propto \begin{cases} p(Y_{mis} \mid Y_{obs}, X), \text{ if } p(R \mid Y, X, Z) \propto p(R \mid Y_{obs}, X) \\ p(Y_{mis} \mid Y_{obs}, X, Z), \text{ if } p(R \mid Y, X, Z) \propto p(R \mid Y_{obs}, X, Z) \end{cases}, \qquad [3.2]$$

Here we will proceed with a standard parametric approach, modeling the missing *Y* conditional on the observed *X* and *Z*. Note that the extent to which *Z* should be incorporated into the imputation model still depends on the specific form of the response mechanism, as shown in [3.2].

3.3.2 The Weighted-FPBB in Clustered and Weighted Sample Designs

In a simple two-stage cluster sampling (balanced case) with equal probability of selection at both cluster and element levels, Meeden (1999) proposed a "two-stage Polya posterior" approach to simulate draws that form an

entire population of clusters, and then an entire population of elements within each cluster. He showed that the posterior mean and variance are very close to standard design-based results. Cohen (1997) proposed a method to generate draws from a weighted Polya posterior using data obtained from weighted sample designs in a non-clustered setting. Here we extend the method of Meeden to a two-stage cluster sampling with unequal cluster sizes (unbalanced case) and combine it with the approach of Cohen to further handle unequal probability of selection.

Next, we propose two variations of a two-stage procedure to simulate draws from equation [3.1]. The first approach assumes that we know the sampling probabilities at both sampling stages, and uses a weighted FPBB at each stage to generate a population of clusters and then a population of elements (and hence is termed an "adapted two-stage Polya posterior" approach). The second approach requires only the final weights, as is typical in most public use surveys, and uses a Bayesian Bootstrap at the first stage to account for clustering effects, and a weighted FPBB using the final weights to generate the population of elements.

3.3.2.1 Double Weighted Finite Population Bayesian Bootstrap (SYN1)

We will call the first approach 'double weighted FPBB'. As indicated by the name, we propose to simulate the posterior predictive distribution in equation [3.1] by utilizing the weighted FPBB method at both the PSU (or cluster) level and the element level, such that a synthetic population of clusters is created at first and then a synthetic population of elements is created.

Let $D=\{(Y_{ij}, X_{ij}, Z_{ij}, R_{ij}), i=1,...,N, j=1,...,M_i\}$ denote the population of values for element j within cluster i, where N and M_i are the number of clusters and the number of elements within the i^{th} cluster in the population, respectively. Thus the population size is $\sum_{i=1}^{N} M_i = M$. Let $\{b^1,...,b^q,...,b^r, q=1,...,r\}$ be r $(1 \le r \le N)$ distinct matrices of real numbers each of dimension $|b^q_{row}| \times |b^q_{col}|$ (each column vector corresponds to a survey variable) with no row vectors in common. Each cluster in the population can take the form of one of b^q's; let t be a vector of length N, $t_i = q$ when the elements in the i^{th} cluster take on the values of b^q, for i=1,…N. Finally, let

$c_t(q) =$ the number of t_i's which equal q, for $q=1,...,r.$, and

$c^i_{t,D}(k) =$ the number of $(Y_{ij}, X_{ij}, Z_{ij}, R_{ij})$'s which equal the k^{th} row vector $b^{t_i}_k$, for $k=1,...,|b^{t_i}_{row}|, i=1,...,N.$

Suppose a two-stage cluster sample is selected from the above finite population. Thus D divides into the sampled components $D_s = (Y_s, X_s, Z_s, R_s)$ and the nonsampled components $D_{ns} = (Y_{ns}, X_{ns}, Z_{ns}, R_{ns})$. Let h and $\bar{h}$ each represents the collection of the sampled and nonsampled clusters, and thus we can break t into $t=(t_h, t_{\bar{h}})$. Let n denote the number of sampled clusters and m_i the number of sampled elements within the i^{th} sampled cluster, and thus the total sample size is $m = \sum_{i=1}^{n} m_i$. Assume $n=r$, $m_i = \| b^{t_{h_i}} \|$ the number of distinct row vectors in $b^{t_{h_i}}$, and thus $m = \sum_{i=1}^{n} \| b^{t_{h_i}} \|$ for convenience of exposition. Let

$w_{t_h}(i) =$ the first-stage sample weight of the sampled cluster $i \in h$ which equal b^q, for $i(\equiv q) = 1,...,n(\equiv r)$.

$w_{t_{h_i},D_s}(j) =$ the second-stage sample weight of the j^{th} sampled element in the i^{th} sampled cluster which equal $b_k^{t_{h_i}}$, for $j(\equiv k) = 1,...,m_i(\equiv \| b^{t_{h_i}} \|)$.

Note that $\sum_{i=1}^{n} w_{t_h}(i) = n$ and $\sum_{j=1}^{m_i} w_{t_{h_i},D_s}(j) = m_i$. Let the set of distinct clusters in the observed sample (D_s) be $\{b^1,...,b^q,...,b^r\}$, and $\varsigma = \{\varsigma_1, \varsigma_2, ..., \varsigma_r\}$ be the vector of probabilities that $\Pr(D_{s,i} = b^q \mid \varsigma) = \varsigma_q$, for $i = 1,2,...,n$, and $\sum_{q=1}^{r} \varsigma_q = 1$. Let $c_{t_h}(q)$ and $c_{t_{\bar{h}}}(q)$ be the number of clusters taking value b^q in the sampled and the nonsampled part of the population, respectively, for $q = 1,2,...,r$, and $\sum_{q=1}^{r} c_{t_h}(q) = n$ and $\sum_{q=1}^{r} c_{t_{\bar{h}}}(q) = N - n$. Let the set of distinct elements in the sampled clusters be $\{d^1,...,d^k,...,d^m\}$, and $\lambda = \{\lambda_1, \lambda_2, ..., \lambda_m\}$ be the vector of probabilities that $\Pr(D_{s,ij} = d^k \mid \lambda) = \lambda_k$, for $i = 1,2,...,n$, j=1,...,m_i, and $\sum_{k=1}^{m} \lambda_k = 1$, and $c_{t,D_s}^{i}(k)$ and $c_{t,D_{ns}}^{i}(k)$ be the number of elements taking value d^k in the sampled and the nonsampled part of the population, respectively, for $k = 1,2,...,m$, and $\sum_{k=1}^{m} c_{t,D_s}^{i}(k) = m$ and $\sum_{k=1}^{m'} c_{t,D_{ns}}^{i}(k) = M - m = m'$. Assuming a noninformative Haldane prior of ς: $\varsigma \sim Dir(0,...,0)$ together with multinomially distributed weighted counts of distinct clusters in the sample data $p(w_{t_h}(1),...,w_{t_h}(r) \mid \varsigma) \propto \prod_{q=1}^{r} \varsigma_q^{w_{t_h}(q)}$ yields a Dirichlet posterior distribution of ς:

$\varsigma \mid w_{t_h}(1),...,w_{t_h}(r) \sim Dir(w_{t_h}(1),...,w_{t_h}(r))$, and hence the posterior predictive distribution of counts of distinct clusters in the nonsampled data follows a compound multinomial distribution:

$c_{t_{\bar{h}}}(1),...,c_{t_{\bar{h}}}(r) \mid w_{t_h}(1),...,w_{t_h}(r) \sim Mult(N-n;\varsigma^*)$, where ς^* is the parameter vector adjusted for unequal selection probability for clusters as in [3.4]. In a similar fashion, the posterior predictive distribution of counts of distinct elements in the nonsampled data follows a compound multinomial distribution:

$c^i_{t,D_{ns}}(1),...,c^i_{t,D_{ns}}(m) \mid w_{t_h,D_s}(1),...,w_{t_h,D_s}(m) \sim Mult(M-m;\lambda^*)$, where λ^* is the parameter vector adjusted for unequal selection probability for elements as in [3.5]. We then have the "*adapted two-stage Polya posterior*" derived as follows (for simplicity of exposition and by an abuse of notation, we use $t_{\bar{h}}$ and t_h to represent $c_{t_{\bar{h}}}(1),...,c_{t_{\bar{h}}}(r)$ and $w_{t_h}(1),...,w_{t_h}(r)$, respectively; similarly, we use D_{ns} and D_s for $c^i_{t,D_{ns}}(1),...,c^i_{t,D_{ns}}(m)$ and $w_{t_h,D_s}(1),...,w_{t_h,D_s}(m)$, respectively):

$$
\begin{aligned}
p(D_{ns} \mid D_s) &= \frac{p(D)}{p(D_s)} = \frac{p(t) \cdot p(D \mid t)}{p(t_h) \cdot p(D_s \mid t_h)} \\
&= \frac{p(t_h, t_{\bar{h}}) \cdot p(D_{ns}, D_s \mid t_h, t_{\bar{h}})}{p(t_h) \cdot p(D_s \mid t_h)} \\
&= \frac{p(t_h, t_{\bar{h}})}{p(t_h)} \cdot \frac{p(D_{ns} \mid D_s, t_h, t_{\bar{h}}) p(D_s \mid t_h, t_{\bar{h}})}{p(D_s \mid t_h)} \\
&= p(t_{\bar{h}} \mid t_h) \cdot p(D_{ns} \mid D_s, t_h, t_{\bar{h}}) \\
&= \frac{\int_0^1 \ldots \int_0^1 \prod_{q=1}^{r-1} \varsigma_q^{w_{t'}(q)-1} (1 - \sum_{q=1}^{r-1} \varsigma_q)^{w_{t'}(r-1)-1} d\varsigma_1 \ldots d\varsigma_{r-1}}{\int_0^1 \ldots \int_0^1 \prod_{q=1}^{r-1} \varsigma_q^{w_{t_h}(q)-1} (1 - \sum_{q=1}^{r-1} \varsigma_q)^{w_{t_h}(r-1)-1} d\varsigma_1 \ldots d\varsigma_{r-1}} \\
&\times \frac{\int_0^1 \ldots \int_0^1 \prod_{k=1}^{m-1} \lambda_k^{w_{t',D_{ns}}(k)-1} (1 - \sum_{l=1}^{m-1} \lambda_k)^{w_{t',D_{ns}}(m-1)-1} d\lambda_1 \ldots d\lambda_{m-1}}{\int_0^1 \ldots \int_0^1 \prod_{k=1}^{m-1} \lambda_k^{w_{t_h,D_s}(k)-1} (1 - \sum_{l=1}^{m-1} \lambda_k)^{w_{t_h,D_s}(m-1)-1} d\lambda_1 \ldots d\lambda_{m-1}} \\
&= \left\{ \prod_{q=1}^{r} \left\{ \Gamma(w_{t'}(q)) / \Gamma(w_{t_h}(q)) \right\} \right\} / \left\{ \Gamma(N) / \Gamma(n) \right\} \\
&\times \left\{ \prod_{k=1}^{m} \left\{ \Gamma(w_{t',D_{ns}}(k)) / \Gamma(w_{t_h,D_s}(k)) \right\} \right\} / \left\{ \Gamma(M) / \Gamma(m) \right\}
\end{aligned}
\qquad [3.3]
$$

where $w_{t'}(q) = w_{t_h}(q) + c_{t_{\bar{h}}}(q)$, $w_{t_h,D_s}(k)$ is the element-level weight, and

$w_{t',D_{ns}}(k) = w_{t_h,D_s}(k) + c^i_{t,D_{ns}}(k)$.

The posterior distribution in [3.3] does not lend itself to direct calculation, and is approximated by a Monte Carlo simulation procedure described as follows:

Step 1: Apply the weighted FPBB to sampled clusters, that is, draw nonsampled clusters in the population based on the weighted Polya urn distribution. This step realizes the first factor in [3.3]. Denote the original sample of clusters by $\{c_1, c_2, ..., c_n\}$. In generating $N-n$ clusters $\{c_1^*, c_2^*, ..., c_{N-n}^*\}$ from the original *n*, we resample the sampled clusters with probability

$$\varsigma_i^* = \frac{w_{t_h}(i) - 1 + l_{i,k-1} \times \left(\frac{N-n}{n}\right)}{N - n + (k-1) \times \left(\frac{N-n}{n}\right)}, k = 1, ..., (N-n+1), i = 1, ..., n. \qquad [3.4]$$

where $l_{i,k-1}$ is the number of times that the i^{th} cluster has been sampled at the $(k-1)^{th}$ resampling, $w_{t_h}(i)$ is the weight for the i^{th} cluster which is normalized to sum up to the total number of clusters, i.e. $\sum_{i=1}^{n} w_{t_h}(i) = N$. Repeat step 1 L times to obtain L FPBB synthetic populations of clusters.

Step 2: For each repetition of step 1, form a population of clusters $\{c_1, c_2, ..., c_n, c_1^*, c_2^*, ..., c_{N-n}^*\}$. Record the number of times each of the *n* clusters from the original sample appears in the FPBB population of clusters, denoted by $\tau_i, i = 1, ..., n.$ and $\sum_{i=1}^{n} \tau_i = N$. Then update the within cluster *element-level conditional weights* to *unconditional* weights as follow: $w_{j|i}^* = w_{j|i} \times \tau_i, i = 1, ..., n,$ where $w_{j|i}$ is the inverse of the conditional probability that element j is sampled given cluster i sampled. (Note $w_{j|i}^*$ is the same as $w_{t_{h_i}, D_s}(j)$ as previously defined.) Now each observed element in these *n* clusters is associated with an updated weight that represents all nonsampled elements in both the sampled cluster it belongs to and similar nonsampled clusters in the population. Note that the resulting sample has the same sample size of elements (*m*) but different survey weights than the original sample, and we term it a 'FPBB sample'.

We then apply the weighted FPBB again to elements in the 'FPBB sample', and this realizes the second factor in equation [3.3]. In generating $M - m$ units from the *m* elements in the FPBB sample, each of the elements in the FPBB

sample is resampled with probability

$$\lambda_{j|i}^{*} = \frac{w_{j|i}^{*} - 1 + l_{j|i,k-1} \times \left(\frac{M-m}{m} \right)}{M - m + (k-1) \times \left(\frac{M-m}{m} \right)}, k = 1, ..., (M - m + 1),\ j = 1, ..., m. \qquad [3.5]$$

where $w_{j|i}^{*}$ is the updated conditional weight for the j^{th} element in the i^{th} cluster. Note $w_{j|i}^{*}$'s inherit the information from the cluster-level selection probability ς_i^{*}. Again they are normalized to sum up to the total number of elements in the entire population, i.e. $\sum_{j=1}^{m} w_{j|i}^{*} = M.$ Thus we create a single synthetic population.

Repeat step 2 *B* times to obtain *B* FPBB synthetic populations of elements.

Such a procedure captures the sampling variance while untying the sample weights at both cluster and element levels through the creation of synthetic populations. This also realizes our goal of resolving clustering effects (design effect due to cluster sampling, i.e. $deff_c = 1 + (\bar{m} - 1) * \rho$, where $\bar{m}$ is the averaged sample size within each sampled cluster and ρ is the intraclass correlation) as a sampling phenomenon. A simple illustrative example goes as follows: suppose we obtain a sample of households (HHs) from a two-stage cluster sampling design, where all US metropolitan statistical areas (MSAs) are treated as clusters to be sampled at the first stage. Suppose Detroit is among such sampled clusters. By using the proposed procedure, we first replicate the number of all 'Detroit-like MSAs' based on the sample weight for Detroit and use it to update the conditional sample weights of sampled HHs within Detroit; then we replicate the number of all nonsampled HHs in Detroit as well as that in Detroit-like MSAs in the population. The same reasoning applies to other sampled MSAs that are different

from Detroit (e.g. LA and LA-like MSAs). Once we generate the population of HHs, HH-level inference becomes straightforward without the need to consider effects due to sampling at the MSA-level.

Because only final weights w_{ij} are usually available in the sample data that are released to the public, we will often need to estimate both the cluster-level weight $w_{t_h}(i)$ in formulae [3.4] and the element-level conditional weight $w_{j|i}$ in formulae [3.5]. We propose estimating $w_{t_h}(i)$ by $\hat{w}_i = \sum_j w_{ij} / M_i$ and $w_{j|i}$ by $\hat{w}_{j|i} = w_{ij} / \hat{w}_i$, where w_{ij} is the overall sample weight as defined in section 3.3.1.

3.3.2.2 Bayesian Bootstrap — Weighted FPBB (SYN2)

A variation of the procedure proposed in subsection 3.3.1.1 is to replace the first-stage weighted FPBB for clusters with regular Bayesian Bootstrap (BB), hence we name it "BB-weighted FPBB". Operationally the difference is that, rather than obtaining a sample of clusters from a draw from a Polya posterior according to the cluster-level weights, we use a simple replication method assuming IID to capture the cluster-level sampling variance. The final sample weights (w_{ij}) instead of the adjusted element-level conditional weights (i.e. $w^*_{j|i}$) are then directly used as input in the second-stage weighted FPBB. We will show that this procedure not only serves as a handy approximation to the double weighted-FPBB due to easier implementation in practice, it can also deal with complex missing data problems as effective as the double weighted-FPBB where fully parametric MI fails. The full procedure is described as follows:

Step 1: Apply Bayesian Bootstrap (BB) (Rubin, 1981) to draw a sample of the clusters from the original sample of clusters. Unlike parametric model-based approaches, the BB can be used to obtain the distribution of a population parameter through the multinomial likelihood and an improper Haldane prior without assuming a fully parametric model for the response variable, therefore is particularly relevant and useful in the survey sampling context. See Aitkin (2008) for a comprehensive application of BB in finite population inference. In our case, we first use it to simulate the posterior for the proportions of distinct clusters in the population:

1) Draw $n-1$ i.i.d. random variates from $Unif(0,1)$ and order them as $u_{(1)}, u_{(2)}, ..., u_{(n-1)}$;
2) Calculating the gaps between $u_{(i)}$'s as $g_i = u_{(i)} - u_{(i-1)}, i = 1, ..., n-1$, where $u_{(0)} = 0$ and $u_{(n)} = 1$;
3) Sample n clusters with replacement with the vector of probabilities $g = (g_1, ..., g_n)$ to attach to the n distinct clusters in the parent sample, and record the number of times (τ_i) each cluster is selected in the 'Bayesian bootstrap (BB) sample', $\sum_{i=1}^{n} \tau_i = n.$
4) Apply the 'ultimate cluster principle' (Wolter, 2007), that is, once a given cluster is taken into the bootstrap sample, all successive stage units are taken into the sample also. Modify the initial case weight in

the parent sample as follow:

$$w_{ij}^{*} = \begin{cases} \tau_i \cdot w_{ij}, & \text{if } i^{th}\text{cluster is selected or duplicated;} \\ 0, & \text{if } i^{th}\text{cluster is not selected.} \end{cases},$$

Normalize w_{ij}^{*}'s to sum up to the population size. Note that the bootstrap sample size (m^{*}) is different from the parent sample size (m).

Step 2: Generate nonsampled elements of the population accounting for inclusion weights: Select a FPBB Pólya sample of size ($M^{*} = M - m^{*}$) from the compound multinomial distribution $mult(M^{*}; p_1, ..., p_d)$, where p_j, $j = 1, ..., d$, are the proportions of distinct values in the population and they are computed based on

$$p_j = \frac{w_{ij}^{*} - 1 + l_{ij,k-1} \times \left(\frac{M - m^{*}}{m^{*}} \right)}{M - m^{*} + (k-1) \times \left(\frac{M - m^{*}}{m^{*}} \right)},\ k = 1, ..., M - m^{*} + 1,\ j = 1, ..., d. \qquad [3.6]$$

Simulate several (*B*) copies of the nonsampled population.

3.3.3 Multiply Imputing Missing Data

Denote the $L \times B$ FPBB "unweighted" synthetic populations generated by the weighted-FPBB by $P^{syn} = \{P_{(b)}^{(l)}, b = 1, ..., B, l = 1, ..., L\}$, where $P_{(b)}^{(l)} = (Y_{(b)mis}^{(l)}, P_{(b)obs}^{(l)})$. Having untied the sampling design, we are ready to perform conventional parametric MI under an IID assumption. Following the standard MI procedure or approximations such as SRMI (Raghunathan et al., 2001), we obtain draws from the posterior predictive distribution $p(Y_{(b)mis}^{(l)} \mid P_{(b)obs}^{(l)})$, where $Y_{(b)mis}^{(l)}$ and

$P_{(b)obs}^{(l)}$ consist of the unobserved and observed data in the lb^{th} FPBB dataset respectively. Without the need to include weights or cluster-level random effects in the imputation model, our task can now be concentrated on correctly modeling the covariates as well as interactions among them whenever necessary.

Point and variance estimate then proceeds as follows. Denote the imputed synthetic datasets as $P^{imp} = \{P_{(11)}^{(1)},...,P_{(1M)}^{(1)},...,P_{(B1)}^{(1)},...,P_{(BM)}^{(1)},...,P_{(BM)}^{(L)}\}$. Note here M denotes the number of imputations created and is different from the M in the previous section which denotes population size of elements. By the standard Rubin (1987) MI combining rules, we have

$$Q \mid P^{imp} \overset{\cdot}{\sim} t_{L-1}(\bar{Q}_L, (1+L^{-1})V_L), \qquad [3.7]$$

where $\bar{Q}_L = \frac{1}{L}\sum_l \tilde{Q}^{(l)}$, $V_L = \frac{1}{L-1}\sum_l (\tilde{Q}^{(l)} - \bar{Q}_L)^2$, and $\tilde{Q}^{(l)} = \lim_{\substack{B\to\infty \\ M\to\infty}} \frac{1}{BM}\sum_b \sum_m q^{(lbm)}$,

where $q^{(lbm)}$ is an estimate of Q obtained from the m^{th} imputation of the b^{th} synthetic population within l^{th} (finite population) Bayesian Bootstrap sample, obtained as $\hat{Q}^{(l)} = \frac{1}{BM}\sum_b \sum_m q^{(lbm)}$. Note that the generation of the synthetic population sets the within imputation variance to 0 so that the posterior variance of Q can be obtained using V_L only; see Dong et al. (2014) and Zhou et al. (2013a).

Note a difference should be made to the small sample scenario when we apply these combining rules. Lo (1988) showed that the variance estimator for the FPBB mean in a simple random sample setting should be inflated by the factor $(\frac{n+1}{n-1})$ (here n is the sample size of elements). Therefore, to get the variance

estimate correct under the double-weighted FPBB, we should use $\frac{n+1}{n-1}(1+L^{-1})\, V_L$ (here n is the sample size of clusters).

3.4 Simulation Study

The simulation study considers two different clustered population data structures. For the first structure, we impose only clustering effects in the data model while leaving the weight independent of the data generating mechanism.

Structure 1: $Y_{ij} \sim \alpha_0 + \alpha_1 X_{ij} + u_i + \varepsilon_{ij}$, where $u_i \sim N(0, \sigma_u^2), \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$

For the second structure, the data model depends on both clustering effects and the measure of size (MOS) (Z) for second-stage sampling as well as its interaction with the covariate (X). Under this model, the subsampling probability is correlated with the response variable even after conditioning on the covariate:

Structure 2: $$Y_{ij} \sim \beta_0 + \beta_1 X_{ij} + \beta_2 Z_{ij} + \beta_1 X_{ij} Z_{ij} + v_i + \varepsilon_{ij},$$ where $v_i \sim N(0, \sigma_v^2), \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$

We conduct a separate factorial design for each population structure by considering the following *simulation factors*:

- *Intraclass correlation coefficients ($ICC_{y|x}$):*

We set $ICC_{y|x}$ to be 0.1 or 0.5. Note that the $ICC_{y|x}$ we examine here is a conditional *ICC* which measures residual correlation after allowing for dependence on the *X* variables, which will in general be larger than ICC_Y, the unconditional *ICC* of *Y*. We also vary the cluster size (M_i) to magnify the design effect. We purposely do not let the clustering occur on the covariates because the

impact of ignoring clustering effects in the imputation model may be abated by conditioning on X, which partially explains some of the unobserved clustering effects on Y (Andridge, 2011).

- *The number of selected clusters in the sample (n), given a fixed number of clusters (N) in the hypothetical population.*

We set two values for the first stage sample size: $n = 10$ or 40. The purpose is to observe both small and large sample behaviors for alternative methods. Meeden argues that for the "two-stage Polya posterior" to yield sensible 0.95 credible interval estimates in a simple balanced two-stage cluster sample, the number of subunits (i.e. m) selected in each sampled cluster should be large enough. We follow his suggestion by setting $m = 20$ (or $\bar{m} = 20$ under an unbalanced design), and therefore we can focus on the effects of different PSU-level sample sizes on the performance of the methods.

- *The response mechanism (R) for Y:*

Much research that investigates methods for missing data imputation only looks at one type of MAR mechanism, namely where the response indicator depends on the regular covariate matrix X. However, it is reasonable to assume that the missingness is also related to sample design information available in the sample cases. For example, in the Survey of Income and Program Participation (SIPP), low-income populations are often oversampled, and thus low-income respondents tend to have smaller weights in the sample dataset. If these families or individuals are also less likely to respond to survey questions on the specific topic, then the response mechanism can be considered as being associated with

survey weights (W) (Andridge & Little, 2009). We will look at both types of response mechanisms (denoted as MAR_X and MAR_X,W, respectively) in our simulation for both data structures.

Based on the above stated factors, we conduct a full 2×2×2 factorial design for each data structure. A total of five MI methods are compared in the simulation study. They are described as follows:

- *SRS*: the linear model where clustering effects are ignored altogether
- *FX*: the linear fixed cluster effects model described in 3.2.2
- *RE*: the linear random cluster effects model described in 3.2.3
- *SYN1*: the two-step MI method with the double-weighted FPBB as 1st step
- *SYN2*: the two-step MI method with the BB-weighted FPBB as 1st step

The imputation tasks under SRS, FX, SYN1 and SYN2 are all implemented using R *mice* package (van Buuren & Groothuis-Oudshoorn, 2011). The imputation under RE is implemented using R *pan* package (Zhao & Schafer, 2013). Specifically, we selected noninformative priors for regression parameters, and diffuse inverse-Wishart priors for variance components; we imputed the missing data 5 times taking 100 iterations between imputations, after a burn-in period of 1,000 iterations.

Finally, for imputation, we consider three model specifications:

- X-only model: $y \sim N(\beta_0 + \beta_1 * x + [b], \sigma^2)$
- Main effect model: $y \sim N(\beta_0 + \beta_1 * x + \beta_2 * w + [b], \sigma^2)$
- Interaction model: $y \sim N(\beta_0 + \beta_1 * x + \beta_2 * w + \beta_3 * x * w + [b], \sigma^2)$,

Note that the cluster effect term [*b*] is only included under FX and RE. Note also that we will employ all three models for the three fully parametric MI methods although only the interaction model is considered as "appropriate". For the two variations of synthetic MI methods, we will use X-only model for imputation throughout all scenarios.

3.4.1 Description of the Design

Step 1: Population data generation

We fix the number of clusters in the population as $N = 400$ for all data structures. Varying cluster sizes are generated from the same uniform distribution $M_i \sim Unif(50, 250), i = 1, ..., 400.$, which sum up to a population size of $N = 59501$ for population Structure 1 and $N = 59230$ for population Structure 2. For Structure 1, we let $\alpha_0 = 1, \alpha_1 = 1.5,$ and $X_{ij} \sim N(0,1).$ For Structure 2, $Z_{ij} \sim N(2,1),\ X_{ij} \sim N(0.1 * Z, 1),\ \beta_0 = 0, \beta_1 = 0.2, \beta_2 = 0.6, \beta_3 = 0.5.$ The random cluster effects are set to be independent standard normal variates, i.e. $u_i,\ \mathrm{v}_i \sim N(0,1).$ In order to generate hypothetical populations with varying ICC, we vary the random error terms as: $e_{ij}^1 \sim N(0,1)$ and $e_{ij}^2 \sim N(0,9),$ resulting in $ICC_1 = 0.5$ and $ICC_2 = 0.1$ respectively. These correspond to two hypothetical populations under each data structure.

Step 2: Drawing replications of samples

We draw clusters with probability proportional to the cluster size M_j, and $f_{1i} = n \cdot M_i / \sum_{k=1}^{N} M_k$; The subsampling methods within each sampled cluster are

different for the two data structures: we select elements using SRSWOR with an equal rate of $f_2 = 0.1$ for the first data structure, while using PPSWOR with Z as the MOS for the second data structure. As a result, the final weight is dominated by the cluster-level weight for the former and the element-level weight for the latter.

Step 3: Imposing missingness

For samples drawn from the first data structure: under MAR_X, we use a logistic function as the deletion function,

$\Pr(R=1) = \text{expit}(\lambda_0 + \lambda_1 X), (\lambda_0 = 0.8,\ \lambda_1 = -0.2)$; under MAR_X,W, we define a probit model $\Pr(R=0 \mid S) = \Phi(S)$ to generate missingness indicators, where the latent variable is $S = \eta_0 + \eta_1 X + \eta_2 \log(weight) + e,\ e \sim N(0,1^2)$.

($\eta_0 = -3.8,\ \eta_1 = -0.8,\ \eta_2 = 1$.) Similar forms of the probit model are used for data structure 2 with a different latent variable:

$S = \xi_0 + \xi_1 X + \xi_2 Z + \xi_3 X * Z + e,\ e \sim N(0,1^2)$, where we set

$\xi_0 = -0.635, \xi_1 = 0.4, \xi_2 = \xi_3 = 0$ under condition MAR_X, and

$\xi_0 = -0.55, \xi_1 = 0.4, \xi_2 = -0.5, \xi_3 = 0.4$ under condition MAR_X,W. All scenarios result in a missing rate of approximately 30% on *Y*.

Step 4: MI inference

For each scenario, 200 simulations are obtained. The number of imputations under the fully parametric MI methods is $M = 5$ and we set $L = 100, B = 20$ and $M = 5$ for the synthetic MI methods. Our focus is on estimating the population mean of *Y* and regression coefficients of *Y* on *X*. The

analyst's model is assumed $Y = b_0 + b_1 X$.

Design-based analyses are performed for original samples before deletion (termed "BD") and for fully parametric methods (SRS/FX/RE). Specifically, we use Horvitz-Thompson-type estimator for point estimation and ultimate cluster variance estimator for the within imputation variance estimation when estimating the population mean (Cochran, 1977), and we use weighted least squares estimators and sandwich variance estimators to estimate regression coefficients (Korn & Graubard, 1999).

In terms of interval estimation, since neither a sample size of *n=10* or *n=40* adequately allows for a normal approximation, we use *t* reference distribution instead to construct the 95% confidence intervals for the mean and regression coefficients. Different degrees of freedom (df) are used for different methods: 1) For the actual samples BD, complete data df is used, i.e. $v_{com} = n - p$ where *p* is the number of parameters; 2) For the fully parametric MI method, we use the small sample df derived in Barnard and Rubin (1999), i.e. $\tilde{v}_m = (\frac{1}{v_m} + \frac{1}{\hat{v}_{obs}})^{-1} = v_{com}[\{\lambda(v_{com})(1-\hat{\gamma}_m)\}^{-1} + \frac{v_{com}}{v_m}]^{-1}$, where v_m is the large-sample-finite-number-imputation df as defined in section 3.2.1, and $\hat{\gamma}_m = (1+M^{-1})\frac{V_B}{T}$ is approximately the Bayesian fraction of missing information (FMI); and 3) For the proposed synthetic MI method, because the FMI is not well-defined in the context of synthetic data in the current literature, and hence the df for constructing interval estimation for the small sample *t*-approximation synthetic data cannot be constructed using FMI, we propose using $v_{syn} = \min\{L-1, v_{com}\}$ as an

approximation. This simultaneously accounts for finite number synthetic data generation ($L-1$) and small sample size (v_{com}).

3.4.2 Results

Results pertaining to varying simulation conditions are summarized in Tables 3.1 - 3.6. The abbreviations for different methods are as previously described. We display the following five measures in the result tables: Pt.est. (point estimate) or Relbias (relative bias), SE (standard error of the estimate), Emp.SE (empirical standard error of the estimate), RMSE (root mean squared error) and Cov. (coverage rate of the nominal 95% confidence interval). We also display the interval widths (Intv.wth) in Tables 3.1 – 3.2.

- *Population structure 1*

Under population structure 1, the final sampling weight is dominated by cluster-level weights and is uncorrelated with the outcome variable, so the estimator of Y is unbiased even when we exclude weight as a predictor in the imputation model (see Table 3.1 and 3.2). Thus, our focus is on the clustering effects which only impact the variance and hence coverage properties. The average estimated fraction of missing information (FMI) using the model-based MI methods (including SRS, FX and RE) is 11.5% for the mean and 32.5% for the slope. The estimated FMI differs by ICC when estimating the mean, which is 5.4% for ICC=0.5 and 17.6% for ICC=0.1.

Despite MI's ability to propagate imputation uncertainty, the standard imputations based on SRS where clusters are ignored altogether leads to even lower standard errors (SEs) than the actual before deletion samples, thus yielding the poorest CI coverage for the estimated mean. It also leads to larger than

expected SEs when estimating the slope, since the variability due to clustering is absorbed all in the residual variance which further inflates the variance of slope, yielding overly conservative intervals. On the other hand, the fixed effects model tends to overestimate the SEs for the mean (e.g. numbers in red), and thus results in intervals that are too conservative, especially when the ICC is low (ICC=0.1). This result replicates the findings of Andridge (2011). The results hold regardless of whether the response indicator is a function of X alone (MAR_X), or both X and W (MAR_X,W).

The two newly proposed synthetic methods and the RE model have similar results when estimating the mean; their performances are generally better than FX and SRS (Table 3.1). Our synthetic methods demonstrate some advantages over RE when estimating the slope in the small sample scenario (a=10). While the use of t-corrected inference for small sample size under RE tends to overcorrect the coverage (e.g. numbers in blue), SYN1 and SYN2 yield approximately nominal coverage (Table 3.2). Again, results are similar for both the MAR_X and MAR_X,W missingness mechanisms.

Table 3.1 Performance of alternative MI methods for estimating the Mean: population 1, two-stage unbalanced sample design (tv=true value)

Simulation Parameters	Method	MAR_X						MAR_X,W					
		Pt.est.	SE.	Emp.SE	RMSE	Intv.wth	Cov.	Pt.est.	SE.	Emp.SE	RMSE	Intv.wth	Cov.
ICC=0.5, a=10 tv=1.05	BD	1.02	0.34	0.36	0.36	1.54	95.50%	1.02	0.34	0.36	0.36	1.54	95.5%
	SRS	1.03	0.28	0.36	0.36	1.33	90.50%	1.03	0.30	0.37	0.37	1.38	91.5%
	FX	1.02	0.37	0.36	0.36.	1.69	96.50%	1.03	0.37	0.37	0.37	1.68	96.5%
	RE	1.02	0.36	0.36	0.36	1.65	96.50%	1.02	0.36	0.37	0.37	1.64	95.5%
	SYN1	1.03	0.36	0.38	0.38	1.59	95.50%	1.02	0.36	0.38	0.38	1.56	95.0%
	SYN2	1.02	0.37	0.37	0.37	1.65	95.50%	1.02	0.36	0.38	0.38	1.62	95.0%
ICC =0.5, a=40 tv=1.05	BD	1.04	0.18	0.19	0.19	0.73	95.50%	1.04	0.18	0.19	0.19	0.73	95.5%
	SRS	1.04	0.15	0.19	0.19	0.59	86.50%	1.04	0.15	0.19	0.19	0.61	86.5%
	FX	1.04	0.19	0.19	0.19	0.76	95.50%	1.04	0.19	0.19	0.19	0.76	97.0%
	RE	1.04	0.18	0.19	0.19	0.74	94.50%	1.04	0.18	0.19	0.19	0.75	96.0%
	SYN1	1.04	0.18	0.20	0.20	0.72	92.50%	1.03	0.18	0.20	0.20	0.71	92.5%
	SYN2	1.05	0.19	0.20	0.20	0.77	94.50%	1.03	0.19	0.20	0.20	0.77	94.5%
ICC =0.1, a=10 tv=1.06	BD	1.06	0.42	0.43	0.43	1.89	95.00%	1.06	0.42	0.43	0.43	1.89	95.0%
	SRS	1.05	0.41	0.45	0.45	2.08	95.50%	1.07	0.43	0.46	0.46	2.27	95.5%
	FX	1.07	**0.54**	0.45	0.45	2.54	98.50%	1.07	**0.54**	0.48	0.48	2.58	98.0%
	RE	1.07	0.47	0.45	0.45	2.30	98.50%	1.08	0.49	0.47	0.47	2.40	98.0%
	SYN1	1.05	0.47	0.46	0.46	2.06	95.50%	1.07	0.48	0.47	0.47	2.10	96.0%
	SYN2	1.07	0.48	0.46	0.46	2.19	97.00%	1.07	0.48	0.48	0.48	2.19	97.0%
ICC =0.1, a=40 tv=1.06	BD	1.05	0.21	0.22	0.22	0.87	94.50%	1.05	0.21	0.22	0.22	0.87	94.5%
	SRS	1.05	0.20	0.24	0.24	0.86	92.50%	1.05	0.21	0.24	0.24	0.87	90.5%
	FX	1.05	**0.27**	0.24	0.24	1.09	99.00%	1.05	**0.27**	0.24	0.24	1.08	98.5%
	RE	1.05	0.23	0.24	0.24	0.95	93.50%	1.05	0.24	0.24	0.24	0.97	94.0%
	SYN1	1.05	0.23	0.25	0.25	0.89	94.50%	1.04	0.23	0.25	0.25	0.90	92.5%
	SYN2	1.05	0.25	0.25	0.25	1.01	96.50%	1.04	0.24	0.25	0.25	0.99	93.5%

Table 3.2 Performance of alternative MI methods for the Slope: population 1, two-stage unbalanced sample design (tv=true value)

Simulation Parameters	Method	MAR_X						MAR_X,W					
		Pt.est.	SE.	Emp.SE	RMSE	Intv.wth	Cov.	Pt.est.	SE.	Emp.SE	RMSE	Intv.wth	Cov.
ICC =0.5, a=10 tv=1.51	BD	1.51	0.11	0.12	0.12	0.48	94.5%	1.51	0.11	0.12	0.12	0.48	94.5%
	SRS	1.51	0.14	0.14	0.14	0.85	98.5%	1.53	0.17	0.16	0.16	1.04	97.5%
	FX	1.52	0.13	0.14	0.14	0.73	98.0%	1.52	0.14	0.14	0.14	0.83	98.0%
	RE	1.52	0.13	0.13	0.13	0.69	**98.0%**	1.52	0.14	0.14	0.14	0.80	**99.0%**
	SYN1	1.52	0.13	0.14	0.14	0.57	93.5%	1.52	0.14	0.16	0.16	0.62	95.0%
	SYN2	1.52	0.13	0.14	0.14	0.61	96.0%	1.52	0.15	0.15	0.15	0.71	96.5%
ICC =0.5, a=40 tv=1.51	BD	1.51	0.06	0.06	0.06	0.23	95.5%	1.51	0.06	0.06	0.06	0.23	95.5%
	SRS	1.51	0.07	0.07	0.07	0.31	93.5%	1.51	0.09	0.08	0.08	0.37	96.5%
	FX	1.51	0.07	0.07	0.07	0.28	95.5%	1.51	0.07	0.07	0.07	0.31	98.0%
	RE	1.51	0.07	0.07	0.07	0.27	95.5%	1.51	0.07	0.07	0.07	0.30	96.5%
	SYN1	1.51	0.07	0.07	0.07	0.26	93.5%	1.51	0.07	0.07	0.07	0.30	94.0%
	SYN2	1.51	0.07	0.07	0.07	0.29	95.0%	1.51	0.08	0.07	0.07	0.33	97.5%
ICC =0.1, a=10 tv=1.52	BD	1.52	0.25	0.26	0.26	1.17	96.5%	1.52	0.25	0.26	0.26	1.17	96.5%
	SRS	1.51	0.33	0.32	0.31	2.00	98.5%	1.53	0.37	0.34	0.34	2.77	100.0%
	FX	1.50	0.33	0.32	0.32	2.01	98.0%	1.52	0.37	0.34	0.34	2.60	99.0%
	RE	1.51	0.33	0.32	0.32	1.93	**99.0%**	1.54	0.35	0.35	0.35	2.50	**99.0%**
	SYN1	1.50	0.31	0.32	0.32	1.38	96.5%	1.53	0.35	0.36	0.36	1.54	96.0%
	SYN2	1.50	0.32	0.32	0.32	1.49	97.0%	1.53	0.36	0.35	0.35	1.67	98.0%
ICC =0.1, a=40 tv=1.52	BD	1.52	0.13	0.13	0.13	0.53	95.5%	1.52	0.13	0.13	0.13	0.53	95.5%
	SRS	1.52	0.16	0.16	0.16	0.69	96.0%	1.51	0.18	0.17	0.17	0.78	95.5%
	FX	1.51	0.16	0.16	0.16	0.72	96.5%	1.51	0.18	0.17	0.17	0.81	98.0%
	RE	1.51	0.16	0.16	0.16	0.68	97.5%	1.51	0.17	0.16	0.16	0.76	96.5%
	SYN1	1.51	0.15	0.16	0.16	0.60	94.0%	1.51	0.17	0.17	0.17	0.66	94.5%
	SYN2	1.51	0.16	0.16	0.16	0.65	96.0%	1.51	0.18	0.17	0.17	0.75	97.5%

- *Population structure 2:*

Under population structure 2, the final weight is dominated by the element-level weights and is associated with the outcome variable *Y* after conditioning on *X*. Therefore the clustering effects on the inference for *Y* are confounded by the weight effects.

Tables 3.3 - 3.4 display the results of SYN1 and SYN2 in the complete data context (SYN1_BD, SYN2_BD) for estimation of the mean and slope, respectively. Both versions of the proposed method result in point and variance estimates very close to that obtained from the actual replication samples (BD). This clearly indicates the ability of the synthetic procedure to untie both the weight and the clustering effects. There are even slight gains in efficiency (i.e. smaller RMSE) of the proposed methods over the design-based BD estimates.

Tables 3.5 - 3.6 display the results of different MI methods in the analysis of missing data for estimating the mean and slope, respectively. When estimating the mean under MAR_X condition, all three fully parametric methods yield sizable biases and large RMSE in the point estimates under the X-only model, and at least the main effects model is required in this case for these biases to disappear. But if the missingness also relates to the weight (i.e. under MAR_X,W mechanism), then only the interaction model gives approximately unbiased estimates. (For example, the relative biases are about 6.5% for SRS_x,w, FX_x,w and RE_x,w.) For the slope, however, the correct (interaction) model is always required for obtaining unbiased estimates, regardless of the missingness

mechanisms. (For example, the relative biases are about 9.0% under MAR_X and 14.0% under MAR_X,W, for SRS_x,w, FX_x,w and RE_x,w.) As before, the SRS method underestimates the variance for the mean and overestimates the variance for the slope because the clustering effects are ignored. The FX method with small ICC overestimates the variance and hence the confidence interval coverage for MAR_X; and this problem becomes more pronounced for MAR_X,W. (For example, the SE./Emp.SE./Cov. are 0.584/0.538/99.0% for MAR_X and 0.638/0.567/99.5% for MAR_X,W, even under the correct model FX_x,w,x*w). With misspecified imputation models under all three fully parametric methods, the magnitude of biases and undercoverage is amplified by the extra association between the weight and the response mechanism in addition to its association with the outcome variable.

Our synthetic MI performs consistently well in all scenarios, and is superior over the fully parametric methods in this combination of population structure and sampling design. By using the simplest X-only imputation model, it not only yields comparable RMSE as the FX and RE using the appropriate interaction model, it also results in the smallest relative bias for point estimates. This is because, despite the operational and inferential similarity to the frequentist bootstrap (Efron, 1979), the (finite population) Bayesian bootstrap simulates the posterior distribution of the true parameter (φ) instead of the sampling distribution of a statistic estimating that parameter ($\hat{\varphi}$). As a result, the point estimates based on our method average closer to the true value than those obtained from fully parametric methods which more resemble the statistic based

on actual before deletion samples. There are slight underestimation for the variance (i.e. SE. < Emp.SE.) with SYN1 relative to SYN2 as well as the appropriate fully parametric models (i.e. FX_x,w,x*w and RE_x,w,x*w), but this does not affect the overall conclusion.

Furthermore, while the correct imputation model in the form of [3.2] is required for our semi-parametric method to break associations between the sample design (*W*) and response mechanism (*R*) under a MAR_X,W condition, the correct structure in the observed synthetic population data greatly reduce the effect of this association in the ultimate inference, even if the design is not included in the second-step imputation.

Table 3.3 Performance of alternative methods for estimating the *mean*: population 2, two-stage unbalanced sample design, before deletion results (tv=true value).

Simulation Factor	Method	Summary Statistics					
		Pt.est.	SE.	Emp.SE	RMSE	Intv.wth	Cov.
ICC =0.5, a=40; tv=1.542	BD	1.498	0.179	0.176	0.181	0.723	96.5%
	SYN1_BD	1.507	0.176	0.179	0.182	0.712	95.0%
	SYN2_BD	1.514	0.178	0.175	0.177	0.719	95.5%
ICC =0.1, a=40; tv=1.541	BD	1.493	0.233	0.232	0.237	0.944	94.5%
	SYN1_BD	1.502	0.230	0.235	0.237	0.931	93.5%
	SYN2_BD	1.509	0.231	0.232	0.234	0.936	94.5%
ICC =0.5, a=10; tv=1.542	BD	1.511	0.354	0.369	0.370	1.603	97.5%
	SYN1_BD	1.512	0.362	0.377	0.377	1.640	98.0%
	SYN2_BD	1.521	0.358	0.371	0.370	1.620	98.0%
ICC =0.1, a=10; tv=1.541	BD	1.507	0.474	0.517	0.517	2.145	95.5%
	SYN1_BD	1.505	0.485	0.520	0.520	2.196	95.0%
	SYN2_BD	1.527	0.470	0.511	0.510	2.126	95.0%

Table 3.4 Performance of alternative methods for estimating the *slope*: population 2, two-stage unbalanced sample design, before deletion results (tv=true value).

Simulation Factor	Method	Summary Statistics					
		Pt.est.	SE.	Emp.SE	RMSE	Intv.wth	Cov.
ICC =0.5, a=40; tv=1.265	BD	1.236	0.095	0.102	0.106	0.384	96.5%
	SYN1_BD	1.246	0.089	0.097	0.099	0.361	94.0%
	SYN2_BD	1.248	0.093	0.099	0.100	0.378	95.5%
ICC =0.1, a=40; tv=1.255	BD	1.226	0.174	0.170	0.172	0.706	94.0%
	SYN1_BD	1.237	0.166	0.166	0.166	0.671	92.0%
	SYN2_BD	1.241	0.173	0.168	0.168	0.700	94.5%
ICC =0.5, a=10; tv=1.265	BD	1.251	0.171	0.206	0.206	0.787	92.5%
	SYN1_BD	1.267	0.169	0.193	0.193	0.780	91.0%
	SYN2_BD	1.268	0.173	0.190	0.190	0.800	93.5%
ICC =0.1, a=10; tv=1.255	BD	1.253	0.330	0.357	0.356	1.520	95.5%
	SYN1_BD	1.270	0.338	0.351	0.350	1.557	95.5%
	SYN2_BD	1.274	0.335	0.343	0.342	1.545	95.0%

Table 3.5 Performance of alternative MI methods for estimating the *mean*: population structure 2, two-stage unbalanced sampling design.

Simulation Factor	Method	MAR_X						MAR_X,W					
		Relbias	SE.	Emp.SE	RMSE	Intv.wth	Cov.	Relbias	SE.	Emp.SE	RMSE	Intv.wth	Cov.
ICC=0.5, a=40 true value =1.542	BD	-2.9%	0.179	0.176	0.181	0.723	96.5%	-2.9%	0.179	0.176	0.181	0.723	96.5%
	SRS_x	13.0%	0.165	0.186	0.274	0.671	74.5%	20.0%	0.169	0.185	0.359	0.691	57.0%
	SRS_x,w	1.8%	0.161	0.187	0.188	0.655	89.5%	6.5%	0.167	0.191	0.215	0.683	86.0%
	SRS_x,w,x*w	-2.9%	0.156	0.181	0.186	0.635	90.0%	-2.7%	0.162	0.187	0.191	0.662	91.0%
	FX_x	13.2%	0.197	0.183	0.273	0.802	83.5%	20.1%	0.201	0.188	0.362	0.819	65.5%
	FX_x,w	1.8%	0.191	0.181	0.182	0.775	96.5%	6.3%	0.197	0.189	0.212	0.801	93.5%
	FX_x,w,x*w	-2.9%	0.187	0.179	0.184	0.759	96.0%	-2.8%	0.191	0.182	0.187	0.778	98.0%
	RE_x	13.2%	0.192	0.184	0.275	0.780	83.0%	20.0%	0.196	0.188	0.362	0.798	66.0%
	RE_x,w	1.8%	0.187	0.183	0.185	0.758	94.5%	6.4%	0.192	0.187	0.211	0.782	92.5%
	RE_x,w,x*w	-3.0%	0.184	0.178	0.184	0.745	95.5%	-2.7%	0.188	0.180	0.185	0.763	97.5%
	SYN1_x	-2.1%	0.185	0.190	0.192	0.750	94.5%	0.6%	0.189	0.198	0.198	0.767	94.0%
	SYN2_x	-1.9%	0.188	0.183	0.185	0.761	95.5%	0.8%	0.193	0.192	0.192	0.779	95.5%
ICC=0.1, a=40 true value =1.541	BD	-3.1%	0.233	0.232	0.237	0.944	94.5%	-3.1%	0.233	0.232	0.237	0.944	94.5%
	SRS_x	12.7%	0.244	0.242	0.311	1.003	86.5%	20.3%	0.260	0.237	0.392	1.077	79.0%
	SRS_x,w	1.3%	0.246	0.258	0.258	1.009	93.5%	6.7%	0.265	0.265	0.284	1.103	94.0%
	SRS_x,w,x*w	-3.2%	0.244	0.252	0.257	1.005	94.0%	-2.7%	0.269	0.266	0.268	1.126	95.5%
	FX_x	12.8%	0.287	0.237	0.309	1.170	92.0%	20.5%	0.304	0.243	0.398	1.246	87.0%
	FX_x,w	1.3%	0.285	0.250	0.250	1.164	97.0%	6.3%	0.306	0.270	0.286	1.261	97.0%
	FX_x,w,x*w	-3.3%	0.283	0.254	0.259	1.155	97.0%	-2.8%	0.307	0.267	0.270	1.267	96.5%
	RE_x	13.0%	0.263	0.240	0.313	1.076	90.0%	20.3%	0.279	0.241	0.395	1.148	85.5%
	RE_x,w	1.3%	0.261	0.259	0.259	1.067	95.5%	6.6%	0.280	0.261	0.280	1.152	95.0%
	RE_x,w,x*w	-3.4%	0.261	0.254	0.259	1.068	96.5%	-2.6%	0.283	0.258	0.260	1.170	97.5%
	SYN1_x	-2.7%	0.254	0.266	0.268	1.029	92.0%	0.6%	0.262	0.270	0.270	1.058	94.5%
	SYN2_x	-2.4%	0.262	0.261	0.263	1.059	94.0%	0.6%	0.267	0.265	0.264	1.080	93.5%

Simulation Factor	Method	MAR_X						MAR_X,W					
		Relbias	SE.	Emp.SE	RMSE	Intv.wth	Cov.	Relbias	SE.	Emp.SE	RMSE	Intv.wth	Cov.
ICC=0.5, a=10 true value =1.542	BD	-2.0%	0.354	0.369	0.370	1.603	97.5%	-2.0%	0.354	0.369	0.370	1.603	97.5%
	SRS_x	12.8%	0.326	0.364	0.413	1.564	91.5%	20.4%	0.333	0.390	0.500	1.622	87.0%
	SRS_x,w	1.8%	0.316	0.370	0.370	1.519	95.5%	6.4%	0.326	0.381	0.392	1.596	93.0%
	SRS_x,w,x*w	-2.9%	0.308	0.372	0.374	1.479	94.5%	-2.6%	0.320	0.373	0.374	1.591	98.0%
	FX_x	13.8%	0.391	0.366	0.423	1.847	95.5%	20.4%	0.397	0.397	0.506	1.887	93.0%
	FX_x,w	2.5%	0.377	0.367	0.368	1.779	98.0%	6.4%	0.388	0.397	0.408	1.845	97.0%
	FX_x,w,x*w	-2.0%	0.369	0.372	0.373	1.741	98.5%	-2.7%	0.381	0.379	0.380	1.812	98.5%
	RE_x	13.7%	0.381	0.364	0.420	1.801	95.0%	20.2%	0.387	0.394	0.502	1.841	93.0%
	RE_x,w	2.3%	0.371	0.366	0.366	1.752	97.5%	6.4%	0.382	0.392	0.403	1.821	96.5%
	RE_x,w,x*w	-2.3%	0.363	0.369	0.370	1.713	98.0%	-2.8%	0.374	0.374	0.376	1.787	98.0%
	SYN1_x	-1.6%	0.377	0.381	0.381	1.705	98.0%	1.6%	0.383	0.399	0.398	1.734	95.5%
	SYN2_x	1.9%	0.377	0.375	0.374	1.707	96.0%	1.6%	0.379	0.405	0.405	1.716	93.0%
ICC=0.1, a=10 true value =1.541	BD	-2.2%	0.474	0.517	0.517	2.145	95.5%	-2.2%	0.474	0.517	0.517	2.145	95.5%
	SRS_x	12.6%	0.490	0.492	0.528	2.428	96.5%	19.3%	0.522	0.514	0.593	2.698	97.0%
	SRS_x,w	1.1%	0.491	0.538	0.537	2.466	96.0%	4.3%	0.530	0.543	0.546	2.790	97.5%
	SRS_x,w,x*w	-3.6%	0.491	0.540	0.541	2.476	97.0%	-4.2%	0.544	0.552	0.555	2.983	98.0%
	FX_x	14.1%	0.575	0.488	0.533	2.788	98.0%	18.9%	0.606	0.519	0.595	3.011	98.5%
	FX_x,w	2.1%	0.570	0.532	0.532	2.791	98.5%	4.2%	0.613	0.578	0.580	3.084	99.0%
	FX_x,w,x*w	-2.3%	0.569	0.538	0.538	2.785	99.0%	-4.4%	0.622	0.567	0.570	3.187	99.5%
	RE_x	13.5%	0.532	0.485	0.527	2.591	97.0%	18.6%	0.561	0.514	0.587	2.818	97.5%
	RE_x,w	1.2%	0.536	0.529	0.528	2.652	98.0%	4.3%	0.581	0.558	0.561	3.013	98.5%
	RE_x,w,x*w	-3.4%	0.533	0.532	0.533	2.644	99.5%	-4.3%	0.588	0.555	0.558	3.096	99.0%
	SYN1_x	-2.3%	0.524	0.563	0.562	2.373	96.0%	-1.0%	0.537	0.570	0.569	2.438	92.5%
	SYN2_x	-1.9%	0.527	0.547	0.547	2.382	95.5%	-0.4%	0.536	0.571	0.570	2.425	92.0%

Table 3.6 Performance of alternative MI methods for estimating the *slope*: population structure 2, two-stage unbalanced sampling design.

Simulation Factor	Method	MAR_X						MAR_X,W					
		Relbias	SE.	Emp.SE	RMSE	Intv.wth	Cov.	Relbias	SE.	Emp.SE	RMSE	Intv.wth	Cov.
ICC=0.5, a=40 true value =1.265	BD	-2.3%	0.095	0.102	0.106	0.384	96.5%	-2.3%	0.095	0.102	0.106	0.384	96.5%
	SRS_x	14.5%	0.112	0.104	0.211	0.462	65.5%	22.3%	0.120	0.101	0.299	0.501	35.0%
	SRS_x,w	8.8%	0.108	0.100	0.149	0.443	83.5%	13.8%	0.117	0.100	0.201	0.488	74.0%
	SRS_x,w,x*w	-2.1%	0.110	0.112	0.115	0.454	96.0%	-2.5%	0.129	0.124	0.128	0.539	98.5%
	FX_x	14.4%	0.109	0.103	0.209	0.445	62.5%	22.3%	0.107	0.102	0.300	0.444	29.5%
	FX_x,w	8.5%	0.103	0.101	0.148	0.420	81.5%	13.5%	0.104	0.101	0.198	0.432	65.0%
	FX_x,w,x*w	-2.4%	0.105	0.110	0.114	0.430	96.0%	-2.5%	0.113	0.119	0.123	0.467	97.5%
	RE_x	14.6%	0.107	0.105	0.213	0.440	59.5%	22.1%	0.106	0.101	0.297	0.441	28.0%
	RE_x,w	8.6%	0.101	0.100	0.148	0.415	81.5%	13.6%	0.103	0.098	0.198	0.424	64.0%
	RE_x,w,x*w	-2.5%	0.105	0.110	0.114	0.427	96.0%	-2.5%	0.113	0.119	0.123	0.465	97.0%
	SYN1_x	-1.8%	0.107	0.116	0.120	0.434	93.0%	-3.3%	0.118	0.133	0.142	0.479	90.0%
	SYN2_x	-2.4%	0.113	0.114	0.118	0.459	93.5%	-3.7%	0.123	0.131	0.139	0.498	93.5%
ICC=0.1, a=40 true value =1.255	BD	-2.3%	0.174	0.170	0.172	0.706	94.0%	-2.3%	0.174	0.170	0.172	0.706	94.0%
	SRS_x	14.6%	0.211	0.190	0.264	0.871	91.5%	22.6%	0.234	0.187	0.339	0.989	84.0%
	SRS_x,w	8.8%	0.210	0.186	0.216	0.869	98.0%	14.1%	0.233	0.190	0.259	0.984	92.5%
	SRS_x,w,x*w	-2.2%	0.214	0.210	0.212	0.888	95.0%	-2.7%	0.254	0.237	0.239	1.083	98.0%
	FX_x	14.5%	0.214	0.184	0.258	0.883	92.0%	22.8%	0.233	0.197	0.347	0.981	80.5%
	FX_x,w	8.5%	0.212	0.186	0.214	0.977	97.5%	13.5%	0.235	0.196	0.259	0.988	93.5%
	FX_x,w,x*w	-2.2%	0.217	0.209	0.210	0.899	96.0%	-2.5%	0.250	0.238	0.239	1.060	98.5%
	RE_x	15.1%	0.209	0.190	0.268	0.863	87.0%	22.2%	0.229	0.193	0.339	0.963	82.5%
	RE_x,w	8.7%	0.207	0.186	0.215	0.854	97.5%	13.7%	0.227	0.187	0.254	0.955	92.5%
	RE_x,w,x*w	-2.5%	0.213	0.207	0.209	0.881	96.0%	-2.5%	0.246	0.235	0.237	1.046	99.0%
	SYN1_x	-3.5%	0.207	0.219	0.221	0.837	93.0%	-5.5%	0.232	0.248	0.254	0.938	90.5%

	SYN2_x	-2.5%	0.214	0.218	0.222	0.866	93.0%	-4.1%	0.241	0.244	0.251	0.975	94.5%
Simulation Factor	Method	MAR_X						MAR_X,W					
		Relbias	SE.	Emp.SE	RMSE	Intv.wth	Cov.	Relbias	SE.	Emp.SE	RMSE	Intv.wth	Cov.
ICC=0.5, a=10 true value =1.265	BD	-1.1%	0.171	0.206	0.206	0.787	92.5%	-1.1%	0.171	0.206	0.206	0.787	92.5%
	SRS_x	14.9%	0.207	0.225	0.294	1.112	92.5%	22.8%	0.224	0.213	0.359	1.308	92.0%
	SRS_x,w	9.4%	0.199	0.219	0.249	1.066	94.5%	13.6%	0.217	0.213	0.274	1.261	96.5%
	SRS_x,w,x*w	-0.9%	0.208	0.237	0.236	1.111	97.5%	-1.2%	0.239	0.253	0.253	1.389	98.5%
	FX_x	15.3%	0.198	0.224	0.296	1.028	88.5%	23.1%	0.196	0.203	0.355	1.089	85.5%
	FX_x,w	9.7%	0.186	0.218	0.249	0.953	93.0%	13.9%	0.188	0.206	0.271	1.027	92.0%
	FX_x,w,x*w	-0.6%	0.192	0.230	0.229	0.988	97.0%	-1.3%	0.206	0.243	0.243	1.142	99.5%
	RE_x	15.3%	0.196	0.222	0.294	1.018	87.5%	23.1%	0.193	0.201	0.354	1.082	85.5%
	RE_x,w	9.7%	0.188	0.218	0.250	0.967	92.5%	13.8%	0.189	0.205	0.270	1.054	94.0%
	RE_x,w,x*w	-0.9%	0.192	0.225	0.225	0.985	96.5%	-1.3%	0.207	0.242	0.242	1.150	98.5%
	SYN1_x	0.1%	0.196	0.231	0.231	0.902	92.0%	-1.1%	0.206	0.254	0.254	0.950	91.5%
	SYN2_x	-0.2%	0.205	0.228	0.227	0.945	94.0%	-1.5%	0.216	0.251	0.251	0.998	92.5%
ICC=0.1, a=10 true value =1.255	BD	-0.2%	0.330	0.357	0.356	1.520	95.5%	-0.2%	0.330	0.357	0.356	1.520	95.5%
	SRS_x	16.6%	0.404	0.401	0.450	2.219	97.0%	23.6%	0.447	0.412	0.506	2.769	98.5%
	SRS_x,w	10.8%	0.402	0.395	0.416	2.218	98.5%	13.5%	0.447	0.408	0.441	2.770	100.0%
	SRS_x,w,x*w	0.7%	0.421	0.436	0.435	2.376	97.0%	-0.7%	0.484	0.476	0.475	3.205	99.5%
	FX_x	17.1%	0.407	0.404	0.456	2.215	98.0%	23.7%	0.442	0.403	0.500	2.711	99.5%
	FX_x,w	11.2%	0.403	0.398	0.421	2.190	98.5%	14.0%	0.445	0.404	0.440	2.698	99.5%
	FX_x,w,x*w	1.8%	0.422	0.441	0.441	2.393	98.5%	-0.6%	0.485	0.473	0.472	3.232	99.5%
	RE_x	17.0%	0.400	0.395	0.448	2.172	98.5%	23.5%	0.431	0.400	0.496	2.661	98.5%
	RE_x,w	11.3%	0.404	0.399	0.422	2.235	98.5%	13.9%	0.444	0.402	0.437	2.786	100.0%
	RE_x,w,x*w	0.7%	0.417	0.427	0.426	2.375	98.0%	-0.6%	0.483	0.472	0.471	3.286	100.0%
	SYN1_x	-0.2%	0.400	0.444	0.443	1.843	92.5%	-3.7%	0.441	0.500	0.500	2.033	94.0%
	SYN2_x	1.2%	0.411	0.440	0.439	1.894	94.0%	-1.8%	0.454	0.481	0.481	2.096	96.0%

3.5 Application to NASS-CDS data

The National Automotive Sampling System Crashworthiness Data System (NASS-CDS) publishes micro-level crash record datasets that are obtained from a representative three-stage probability sample design: geographic region (county or groups of counties), police departments, and police-reported crashes. The design also has unequal probabilities of selection, with vehicles in more severe crashes much more likely to be sampled than vehicles in less severe crashes. The sample is selected annually from all police-reported crashes that resulted in at least one vehicle having to be towed from the scene for damage (http://www.nhtsa.gov/NASS).

A key feature of this car-crash data is the Delta-V measure (roughly defined as the "instantaneous" change in velocity). This measure of crash intensity has two qualities that make it quite useful as a tool for testing our imputation method. First, a very high proportion of car-crash cases are missing this variable; second, there exists a rich set of covariates which are potentially good predictors in the imputation for Delta-V. These include basic demographics, road condition, type of vehicle, injury condition of the driver, number of vehicles in the crash, etc. The survey variables selected for imputation and analysis purpose are summarized in Table 3.7. We estimate: 1) the mean of Delta-V as a continuous variable ($\bar{Y}_{Delta-V}$), and 2) the odds ratio of having severe injury or certain types of injury given varying levels of Delta-V.

Table 3.7 Summary of selected survey variables for imputation

Selected Variables	**Values and labels**
Delta V	Continuous
driver's age	Continuous
maximum ais injury	1-not injured/minor injury; 2-severe injury;
direction of force	1-frontal; 2-right side; 3-left side; 4-rear; 5-other
driver's gender	1-male; 2-female
light condition	1-daylight; 2-dark; 3-dark, lighted; 4-dawn/dusk
road surface condition	1-dry; 2-wet; 3-other
vehicle type	1-passenger car; 2-truck
model year	1->2006; 2-<=2006
vehicle make	1-american; 2-japanese; 3-korean; 4-european&other
body region	1-head; 2-thorax; 3-lower extremity; 4-other
primary sampling unit	1,2,…,24.
final survey weight	Continuous

The analysis sample is obtained by merging four different data files at both occupant level and vehicle level. We retain only the driver data, and end up with $n = 7525$ cases in the final sample, where nearly half the sample cases (45%) are missing the Delta-V measure. There are 24 PSUs selected in the first sampling stage (out of a total of 1195 PSU's in the population which were grouped into 12 strata based on geographic region and type, with at least two PSU's selected from each stratum). We regard the stratum effects (at both first- and third-stage) as being reflected in the unequal selection probability or sample weights, and analyze the data using an ultimate-cluster approach with focus on the PSU-level clustering effects. This results in a moderate ICC of 0.1 on Delta-V among the complete cases.

As a preliminary analysis, we examined features of the design variables and the missing data pattern before imputation. The logistic regression coefficients of the response indicator on covariates are mostly significant, suggesting at least a MAR missing data mechanism. Therefore all these covariates are included as predictors for

imputation. A linear regression of final weights on all the survey variables indicates that several covariates as well as the Delta-V measure are strong predictors of final weights. Yet all together they explain only 10% of the variance in weights; therefore incorporating weights in the imputation is essential. Because the estimated population size based on the sum of the weights is in excess of 1 million, we assume for operational simplicity that $N = 80,000$, which is large enough for the proposed method to work properly.

Table 3.8 displays the results of applying different MI methods to the NASS-CDS Delta-V measure. Different modeling strategies are used under each method according to the extent the design information is modeled. Fully parametric imputation that does not include the weight appears to overestimate the mean Delta-V, as would be expected, since more severe crashes are overrepresented in the unweighted sample. When weights are included in the imputation, the fixed effect model FX_X,W tends to result in wider confidence intervals. The FPBB method (SYN2) mimics the performance of the random effects model with weights in the imputation (RE_X,W). Results are broadly similar for estimates of the odds ratios of injury. Here the complete case model appears to underestimate the effect of Delta-V on injury risk, suggesting that injury cases are more likely to have high Delta-V measures missing. The SYN2 approach even enjoys some gains in precision when estimating the odds ratio of having head injury, where all imputation methods yield similar point estimates.

Table 3.8 Estimating mean Delta-V, odds ratio of severe injury given Delta-V, and odds ratio of head injury given Delta-V (high Delta-V=in excess of 35 kph; medium Delta-v=15-35 kph; low delta-V=less than 15 kph). CC=complete case; SRS=imputation under simple random sampling assumption; FX=imputation using fixed cluster effects; RE=imputation using random cluster effects; SYN2=Results from 2-stage finite population Bayesian bootstrap model using Bayesian bootstrap-finite population Bayesian bootstrap for synthetic population generation.

	Methods	CC	SRS		FX		RE		SYN2			
	Models	CC	X	X,W	X	X,W	X	X,W	X	X,W	C	X,W,C
Estimates	$\bar{Y}_{Delta-V}$	26.7	29.8	27.9	29.6	27.3	29.6	27.9	28	27.6	28.1	27.4
	SE	0.88	0.88	0.96	1.12	1.05	1.09	0.88	0.85	0.89	0.81	0.89
	$OR^{\text{Severe Injury}}_{\text{high vs. low DV}}$	5.03	5.78	6.23	5.43	6.38	5.53	6.71	6.55	7.49	6.23	7.29
	95% CI	[2.53, 9.99]	[3.16, 10.6]	[3.37, 11.5]	[2.99, 9.86]	[3.45, 11.8]	[3.23, 9.48]	[3.74, 12.0]	[3.41, 12.6]	[3.85, 14.6]	[3.26, 11.9]	[3.70, 14.4]
	$OR^{\text{Severe Injury}}_{\text{medium vs. low DV}}$	1.28	1.58	1.52	1.61	1.77	1.56	1.77	1.81	1.86	1.76	1.80
	95% CI	[0.612, 2.68]	[0.838, 2.98]	[0.862, 2.68]	[0.917, 2.83]	[0.994, 3.15]	[0.911, 2.67]	[1.01, 3.11]	[1.07, 3.06]	[1.05, 3.29]	[1.03, 3.00]	[0.992, 3.27]
	$OR^{\text{Head Injury}}_{\text{high vs. low DV}}$	1.05	1.55	1.45	1.69	1.52	1.58	1.36	1.62	1.60	1.72	1.56
	95% CI	[0.581, 1.90]	[0.76, 3.16]	[0.751, 2.80]	[0.904, 3.16]	[0.765, 3.02]	[0.706, 3.54]	[0.919, 2.65]	[1.02, 2.56]	[1.09, 2.35]	[1.09, 2.71]	[1.09, 2.23]

3.6 Discussion

Two-stage cluster sampling is a popular sampling scheme in survey research because it is cost effective and easy to administer. However, fully parametric MI does not work well with two-stage cluster sampling when sampling units at both stages are selected with unequal probabilities unless the sampling probabilities are properly accounted for. As shown in the simulation study, when model misspecification with respect to the sampling weights is present, both fixed effects and random effects imputations lead to biases in estimates, and hence invalid inferences**.** As Rabe-Hesketh and Skrondal (2006) put it, dealing with sampling weights in hierarchical/multilevel models can be challenging from either a computational or a modeling standpoint. We propose an alternative two-step MI approach, where the first step generates synthetic populations with missing data that account for weights and clusters, and the second step imputes under an IID assumption.

We propose two different approaches for the two-stage Bayesian nonparametric synthetic data generation from two-stage cluster sampling designs. The first (SYN1) uses a "FPBB-FPBB" approach, generating clusters from a finite-population Bayesian bootstrap and population elements from a finite-population Bayesian bootstrap from those FPBB-generated clusters. The "adapted two-stage Polya posterior" under this approach is different from the "two-stage Polya posterior" of Meeden (1999) in two ways. First, we extend Meeden's simple two-stage cluster sampling (balanced case) with equal selection probability to an unbalanced case with unequal selection probabilities, for applications in the missing data context. Second, while Meeden considered a single variable (Y), we consider the joint distribution of several survey variables (Y, X, Z, R).

Our procedure correctly restores the population configuration of an outcome variable (*Y*) that is built on complex relationships of the regular covariate (*X*) and the sample design variables (*Z*). Such population-level multivariate relationships are not guaranteed by Meeden's procedure (according to results from a side simulation study not reported here). Our second approach (SYN2) uses a "BB-FPBB" procedure, resampling the clusters using a Bayesian bootstrap, and then generating the population elements from a finite-population Bayesian bootstrap. Simulation results show that both approaches are insensitive to either sample size or population structure. However, because SYN1 requires knowing both cluster-level weights and within cluster element-level conditional weights while SYN2 only needs the final weights, we recommend SYN2 for general purposes.

In this chapter, we focus on the mean structure of the imputation model for *Y* and use design-based analysis to account for weight and clustering effects after fully parametric MI. In our future research, we will investigate the robustness of the proposed new method in comparison with other MI methods against potentially incompatible normal assumptions on random cluster effects and random errors (Yucel & Demirtas, 2010). We will also consider extensions to stratified, clustered, and unequal probability of selection sample designs.

CHAPTER 4
A SYNTHETIC MULTIPLE IMPUTATION PROCEDURE FOR MULTI-STAGE COMPLEX SAMPLES

4.1 Introduction

Stratified multistage sampling is the most common type of sample design for large-scale surveys conducted by the U.S. federal statistical agencies. Examples of surveys using stratified multistage sample selection include the Current Population Survey (CPS) by the Census Bureau, the National Health Interview Survey (NHIS) and the National Health and Nutrition Examination Survey (NHANES) conducted by the National Center for Health Statistics, and the National Assessment of Educational Progress (NAEP) conducted by the National Center for Educational Statistics. This type of sample design combines the advantages of both stratification (for statistical efficiency) and cluster sampling (for cost and logistical efficiency). It thus may be as precise as a simple random sample or a single-stage stratified sample design, but costs significantly less (Murphy, 2008). Under this design, the primary sampling units (PSUs) are stratified in such a way that they are homogeneous with respect to a stratum-level aggregate of the variable(s) of interest. To permit a maximum degree of stratification and thus variance reduction, it is common practice to define a large number of strata where only a small number of PSUs are selected in each stratum.

On the one hand, such highly stratified multistage sample designs facilitate the data collection process while assuring broad representativeness of the target population.

On the other hand, because of the complexities involved, i.e. complex sample design features including stratification, clustering and unequal selection probability, these sample designs require sophisticated statistical methods at the analysis stage of survey data. When missing data are present, the analysis of complex survey data becomes particularly challenging. Taking the NAEP as an example, missing data may occur at two different levels: 1) survey nonresponse at the PSU level, if some sampled schools fail to participate in the entire survey or some school-level measures are missing; 2) survey nonresponse at the ultimate sampling unit level, if within participating schools, some students fail to provide responses to items in the survey questionnaire. We consider using multiple imputation (MI) (Rubin, 1976, 1987) to deal with item-level missing data at the ultimate sampling unit level. In particular, we consider the role of complex sample designs in the MI procedure.

Reiter, Raghunathan and Kinney (2006) demonstrated the importance of simultaneously accounting for stratum effects and clustering effects in multiple imputation. They showed that when design features were ignored in the imputation model, biases would occur on the estimated parameter, even if a design-based analysis method was applied on the imputed data. Current MI methods typically include dummy variables to represent strata as well as PSUs nested within each stratum in the imputation model. When necessary, they also identify statistically significant interactions between these dummies with other covariates through routine variable selection procedures such as stepwise regression (Reiter et al., 2006; Schenker et al., 2006). Such a modeling strategy is not only operationally burdensome but also inferentially inefficient when there are hundreds of strata in the sample design, and the sample in each stratum consequently

becomes sparse. For example, the Census Bureau's CPS design groups 1768 nonself-representing PSUs into 220 strata.

A possibly better strategy is to consider clusters as random effects while treating strata as either fixed (using dummies) or random effects. However, many of the popular software packages that implement multiple imputation (e.g. SAS MI procedure, R packages *mice* or *mi*, and *IVEware*) cannot simply be adapted to such a mixed effects approach. While a few recent software modules (such as R package *pan* and MLwiN module *REALCOM-IMPUTE*) have started to consider mixed effects or multilevel modeling for imputation purposes, they typically assume normal or latent normal distribution for variables with missing data. Their performances for missing categorical variables (binary in particular) are unclear. Moreover, little research has formally investigated their use to incorporate strata as well as clusters.

To circumvent these problems with fully parametric model-based imputation techniques, we develop a modification of the two-step semi-parametric MI method proposed by Zhou, Elliott and Raghunathan (2013a, 2013b). The idea was to separate the need to account for complex sample designs from the treatment of missing data. In the first step, they reversed the sample designs through synthetic population data generation. They developed different variations of a weighted finite population Bayesian bootstrap (FPBB) (Cohen, 1997; Little & Zheng, 2007; Dong et al., 2014) for untying the sampling weights and clustering effects. In the second step, they imputed missing values in the created synthetic population based on a much simpler imputation model that assumes IID (identically independently distributed). To account for stratum effects, we propose an adapted version of their procedure in this chapter. The new procedure combines a

replication variance estimation method (Efron, 1979; Kovar, Rao, & Wu, 1988; Rao & Wu, 1988; Rao, Wu, & Yue, 1992; Rust & Rao, 1996) with the weighted FPBB. Under a standard missing at random (MAR) assumption (Little & Rubin 2002), our method requires neither complicated modeling of strata and clusters nor design-based analyses of the imputed data.

Although our method is applicable to multiple imputation in general settings, we focus in this chapter on the estimation of two quantities: quantile estimates for a continuous variable, and estimates of rare proportions and their associated logistic regression estimates. We consider a stratified two-stage sample design and investigate a full range of quantiles including tail behaviors. While design-based methods for quantile estimation from complex survey data have been developed (Francisco & Fuller, 1991; Woodruff, 1952), quantile estimation after imputation is rarely addressed in the literature, to our knowledge. This is despite the rapid development and increasing popularity of MI. We also consider MI for incomplete binary variables, with a focus on rare outcomes. It is well known that maximum likelihood estimation of logistic regression models typically suffers from small sample bias, the degree of which is strongly dependent on the number of sample cases in the less frequent of the two categories (King & Zeng, 2001). Thus when the dependent binary variable represents the occurrence of rare events, the logit coefficients can be substantially biased even with a simple IID data structure. Random effects logistic models are commonly used for fitting clustered binary data, however, these models rely heavily on asymptotic theory assumptions, which may not be met in sparse samples. All these issues might extend naturally to the missing data context. As shown by Zhao and Yucel (2009), sequential MI for binary data missing completely at

random in a multilevel setting suffers from severe bias and poor coverage in estimating probabilities that are close to 0 or 1, particularly when the intraclass correlation is high.

The objectives of this paper are: i) to develop an adapted version of the two-step synthetic MI method of Zhou et al. (2013a, 2013b) as a way to account for stratification, in addition to clustering and unequal inclusion probability; and ii) to demonstrate the effectiveness of the new method, with respect to quantile estimation and logistic regression for binary rare events data, as compared with existing fully parametric imputation strategies. Section 4.2 discusses the imputation strategies under three different models: simple random sample, fixed effects for clusters/strata, and random effects for cluster/strata. Section 4.3 introduces the newly proposed procedure and the MI inference rules for quantile estimation under this method. Section 4.4 presents a Monte Carlo simulation study as the validation tool to assess the repeated sampling properties of MI under the various approaches. Section 4.5 applies different MI procedures to the analysis of body mass index on youth data from the third National Health and Nutrition Examination Survey (NHANES III). Some concluding remarks follow in Section 4.6. We focus on the two-PSU-per-stratum design in this chapter, although the methods we develop can accommodate any number of PSUs per stratum.

4.2 Fully parametric imputation methods for the two-PSU per stratum design

Here we briefly describe fully parametric multiple imputation techniques with complex sample design features incorporated to different degrees. We assume the missing data Y_i is a member of the exponential family, and that there are fully observed covariates X_i (a $(p+1)$- dimension vector) such that $g\left(E(Y_i \mid X_i)\right) = X_i\beta$ for a known link function

$g(\bullet)$ (e.g. $g(u) = \log(\frac{u}{1-u})$ for binary outcomes (logistic regression), $g(u) = \log(u)$ for count outcomes (Poisson regression), or $g(u) = u$ for continuous outcomes (Gaussian regression)).

4.2.1 Standard regression model assuming SRS

Based on the maximum likelihood (ML) estimates $\hat{\beta}$ and the associated asymptotic covariance matrix $\hat{V}(\hat{\beta})$ for the generalized linear model $g\left(E(Y_i \mid X_i)\right) = X_i\beta$, the posterior predictive distribution of the parameters can be constructed, which is then used to impute the missing values (Rubin, 1987, pp. 169-170). The steps used to generate imputed values are summarized as follows:

1) At current iteration *t*, draw new regression parameters $\beta^{(t+1)}$ (a row vector with length $(p+1)$) from their normal approximation posterior predictive distribution $N(\hat{\beta}, \hat{V}(\hat{\beta}))$: $\beta^{(t+1)} = \hat{\beta}^{(t)} + \hat{V}_c^{'} z$, where $\hat{\beta}^{(t)}$ is the ML estimator of β based on the observed data X and Y^{obs} together with the filled-in $Y^{imp(t)}$, $\hat{V}_c^{'}$ is the upper triangular matrix in the Cholesky decomposition of covariance matrix $\hat{V}(\hat{\beta}^{(t)})$, and z is a vector of $p+1$ independent random normal deviates.

2) If the distribution of Y_i has an unknown scale parameter σ^2, draw $\delta \sim \chi^2_{n-p-1}$ and compute $\sigma^{2(t+1)} = \sum_i \left(Y_i - g^{-1}(X_i^T \beta^{(t+1)})\right)^2 / \delta$.

3) For an observation with missing Y_i, draw $Y_{i \in mis}$ from the assumed distribution with mean $g^{-1}(X_{i \in mis}^T \beta^{(t+1)})$ and scale $\sigma^{2(t+1)}$.

Point and variance estimates of the regression parameters can then be obtained using the usual MI combining rules (Rubin, 1987, p. 76). For the p^{th} component of the regression parameter:

$$\hat{\beta}_p = \frac{1}{M}\sum_{m=1}^{M}\hat{\beta}_p^{(m)} \quad , \qquad [4.1]$$

$$\hat{V}(\hat{\beta}_p) = \frac{1}{M}\sum_{m=1}^{M}\hat{V}(\hat{\beta}_p^{(m)}) + \frac{M+1}{M(M-1)}\sum_{m=1}^{M}(\hat{\beta}_p^{(m)} - \hat{\beta}_p)^2 \qquad [4.2]$$

and

$$\frac{\left(\hat{\beta}_p - \beta_p\right)}{\sqrt{\hat{V}(\hat{\beta}_p)}} \sim t_\nu, \nu = (M-1)\left(1 + \frac{\sum_{m=1}^{M}\hat{\beta}_p^{(m)}}{\frac{(M+1)}{(M-1)}\sum_{m=1}^{M}(\hat{\beta}_p^{(m)} - \hat{\beta}_p)^2}\right)^2 , \qquad [4.3]$$

where $m = 1,...,M$ imputations are taken from draws widely separated to have practically eliminated autocorrelation. Multivariate combining rules for the joint distribution of $\hat{\beta}$ are available as well (Schafer, 1997a, pp. 112-118).

4.2.2 Appropriate fixed effects model (FX_APR)

Compared to the predictive model using standard generalized linear regression, we can add dummy variables indicating stratum and cluster memberships to account for stratification and clustering effects. Note we also need to include the log transformation of sampling weight as a predictor if the missing data mechanism depends on weights, to make the imputation model truly appropriate. The model takes the following form:

$$g\left(E(Y_i \mid X_i)\right) = X_i\beta + D_i\gamma + E_i\eta + [\zeta \log(w_i)] , \qquad [4.4]$$

where D_i is a $1\times(H-1)$ row vector of dummies representing the H strata, and E_i is a

$1\times Q$ row vector of dummies representing the clusters nested within each stratum. Note that $Q=\sum_h Q_h - H$, where Q_h is the number of clusters in each stratum; in the case of the two-PSU per stratum case, $Q=H.$ The dummy part of the design matrix thus takes a block diagonal matrix form as follows:

$$\left[\begin{array}{c|ccc|ccc} X_1 & A_1 & 0 & \cdots & B_1 & 0 & \cdots & 0 \\ X_2 & 0 & A_2 & \cdots & 0 & B_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ X_H & 0 & 0 & \cdots & 0 & 0 & \cdots & B_H \end{array}\right]_{n\times(\mathrm{p}+H+Q)},$$

$$A_h = \overset{D_h}{\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}}_{n_h\times 1}, \quad B_h = \overset{E_1 \quad \cdots \quad E_{Q_h-1}}{\begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix}}_{n_h\times(Q_h-1)} \qquad h=1,...,H-1.$$

where X_h is a $n_h\times(p+1)$ matrix containing the values of covariates for observations in stratum $h,$ and $\sum_h n_h = n,\ h=1,...,H.$, A_h is a $n_h\times 1$ matrix containing a vector of ones for observations in stratum *h*, and B_h is an $n_h\times(Q_h-1)$ matrix containing vectors of dummy variables for the Q_h PSUs in the h^{th} stratum. To ensure identifiability, the dummy for the H^{th} stratum and the Q_h^{th} PSU are dropped. The parameter space under this model is expanded as $\theta=(\beta,\gamma,\eta,\zeta)$, and the steps for imputation are similar as in the SRS setting.

4.2.3 Appropriate mixed effects model (RE_APR)

As there are only two PSUs selected from each stratum, it is not feasible to model clusters as random effects separately within each stratum. Here we pool all *Q*+*H* clusters in the sample, and model them using a single random effect term. The imputation model

is specified as follow:

$$g\left(E(Y_j \mid X_j)\right) = X_j\beta + D_j\gamma + u_i + [\zeta \log(w_j)], \qquad [4.5]$$

where $u_i \sim N(0, \sigma_u^2)$ is a random intercept term representing cluster effects, for $i = 1, ..., (Q+H),$ and σ_u^2 denotes the between cluster variance. Other terms are as previously defined. (In the two-PSU-per-stratum case, $Q+H = 2H.$) This MI strategy obtains the imputed values in the following steps:

1) At current iteration *t*, fit the above generalized linear mixed effects regression model using the observed data *X* and Y^{obs} together with the filled-in $Y^{imp(t-1)}$. The inference about $\theta^{(t)} = (\vartheta^{(t)}, \sigma_u^{2(t)})$, where $\vartheta^{(t)} = (\beta^{(t)}, \gamma^{(t)}, \zeta^{(t)})$, is based on the marginal likelihood function $L = \prod_i \prod_j \int_u p(y_{ij}^{(t-1)} \mid \theta^{(t)}) \varphi(u_i; 0, \sigma_u^{2(t)}) du_i$, where $y_{ij}^{(t-1)}$ is either the observed value or the filled-in value from the $(t-1)^{th}$ draw and $\varphi(\bullet; \mu, \sigma^2)$ is the normal density with mean μ and variance σ^2. When Y_i is non-normal, we employ adaptive Gaussian quadrature (Pinheiro & Bates, 1995) (with the number of quadrature points set as 10) to evaluate the integral and obtain the maximum likelihood estimates for the parameters ($\hat{\theta}^{(t)} = (\hat{\vartheta}^{(t)}, \hat{\sigma}_u^{2(t)})^T$) and the associated variance-covariance matrix of $(\hat{\vartheta}^{(t)}, \hat{\sigma}_u^{2(t)})^T$ given by $\hat{\Sigma}^{(t)}$ through numerical methods.
2) Obtain a draw from the normal approximation to the posterior predictive distribution of these parameters: $\theta^{*(t)} = \hat{\theta}^{(t)} + \begin{bmatrix} V_{\vartheta}^{'} \\ V_{\sigma_u}^{'} \end{bmatrix} \times z$, where *z* is a vector containing $p+Q+H+1 = p+2H+1$ independent random normal variates,

$\begin{bmatrix} V_{\vartheta}^{'} \\ V_{\sigma_u}^{'} \end{bmatrix}$ are components of the cholesky root of $\hat{\Sigma}^{(t)}$ corresponding to the fixed effects and random cluster effects, respectively.

3) Any residual scale parameter is drawn as in the SRS setting.

4) Finally, the missing values are drawn from the following distribution:

$$Y_{i\in mis}^{(t)} \mid X_{i\in mis}, \vartheta^{*(t)}, u_j^{*(t)} \sim g^{-1}\left(X_{i\in mis}\beta^{*(t)} + D_i\gamma^{*(t)} + u_j^{*(t)} + \zeta^{*(t)}\log(w_{i\in mis})\right)$$

Point and variance estimates and 95% confidence intervals are then obtained using the standard Rubin MI combining rules as previously described.

4.3 Synthetic MI using the weighted FPBB for stratified samples

In this section, we extend the two-step multiple imputation methodology proposed by Zhou et al. (2013a, 2013b) to a stratified two-stage sample design where a combination of complex sampling techniques are considered, namely, stratification, clustering and unequal inclusion probability. We develop methods for an unrestricted number of clusters per stratum, but for our simulations and application we focus on the special case of two primary sampling units (PSUs) selected per stratum, which mimics the form of a public use dataset that is commonly released for analyses.

4.3.1 Synthetic data generation to account for complex sample designs

Consider a finite population *P*, which is stratified into *H* strata with N_h PSUs in the h^{th} stratum, and hence the population size of PSUs is $\sum_{h=1}^{H} N_h = N.$ For the h^{th} stratum, select n_h PSUs with/without replacement from some probability sampling plan, independently across strata, and hence the total sample size of PSUs is $\sum_{h=1}^{H} n_h = n.$

Subsampling of m_{hi} elements (treated as the ultimate sampling units in this example) from a total of M_{hi} is then conducted within the i^{th} sampled PSU of the h^{th} stratum, for $i = 1,...,n_h,\ h = 1,2,...,H.$ Hence the overall sample size and population size of elements are $\sum_{h=1}^{H}\sum_{i=1}^{n_h} m_{hi} = \sum_{h=1}^{H} m_h = m$ and $\sum_{h=1}^{H}\sum_{i=1}^{N_h} M_{hi} = \sum_{h=1}^{H} M_h = M$, respectively, where m_h and M_h are sample size and population size of elements for the h^{th} stratum, respectively. The population consists of four types of survey variables: a single outcome *Y*, a single covariate *X*, a design matrix $Z = [S, C, w]$ including the stratum indicators (*S*), the cluster indicators (*C*) and the sample weight (*w*), and the response indicator *R*. Let $D = (D_s, D_{ns}) = \{\left(Y_{hij}, X_{hij}, Z_{hij}, R_{hij}\right), h = 1,...,H, i = 1,...,N_h, j = 1,...,M_{hi}\}$ denote the population of values measured on the survey variables, which is divided into the sampled component (D_s) and the nonsampled (D_{ns}) component.

We generate synthetic populations using a two-stage procedure. The first stage accommodates stratification and clustering, and the second weighting. We have two broad approaches. The first, which we term SYN1, assumes that first (cluster-level) and second (PSU-level) stage sample weights are available for the analysis, and implements a weighted finite population Bayesian bootstrap (FPBB) at each level to generate the synthetic population. The second, which we term SYN2, assumes that only final weights are available for the analysis, and uses a Bayesian bootstrap to account for stratification and clustering at the first stage, and the weighted FPBB to account for the final weight at the second stage.

4.3.1.1 Double Weighted Finite Population Bayesian Bootstrap (SYN1)

For the h^{th} stratum, let $t_{s,h}$ and $t_{ns,h}$ index the sampled and nonsampled clusters,

respectively, and $\{b^1,...,b^q,...,b^{r_h}, q=1,...,r_h\}$ be the r_h $(1 \le r_h \le N_h)$ distinct matrices of real numbers each of dimension $|b^q_{row}| \times |b^q_{col}|$ with no row vectors in common. Each cluster in the stratum can take the form of one of b^q's. Let $t_{hi} = q$ when the i^{th} cluster takes on the values of b^q, for $i = 1,..., N_h$. Assume $n_h = r_h$ and $m_{hi} = \| b^{t_{s,hi}} \|$ (the number of distinct row vectors in $b^{t_{s,hi}}$) for convenience of exposition. Let $w_{t_{s,h}}(i)$ be the sample weight of the i^{th} sampled cluster in the h^{th} stratum which equals b^q, for $i = 1,..., n_h$. and $w_{t_{s,h_i},D_{s,h}}(j)$ be the sample weight of the j^{th} sampled element in the i^{th} sampled cluster which equals $b_k^{t_{s,hi}}$, for $j = 1,..., m_{hi}$. Finally, let $c_{t_{s,h}}(q)$ and $c_{t_{ns,h}}(q)$ be the number of sampled and nonsampled clusters that equal b^q, and $c^{hi}_{t_h,D_{s,h}}(k)$ and $c^{hi}_{t_h,D_{ns,h}}(k)$ be the number of sampled and nonsampled elements that equal $b_k^{t_{s,hi}}$.

Zhou et al. (2013b) showed that, <u>*within a stratum h*</u>, the Polya posterior for the counts of distinct unobserved elements $D_{ns,h}$ is given by

$$p(D_{ns,h} \mid D_{s,h}) = \left\{ \prod_{q=1}^{r_h} \left\{ \Gamma(w_{t_h'}(q)) / \Gamma(w_{t_{s,h}}(q)) \right\} \right\} / \left\{ \Gamma(N_h) / \Gamma(n_h) \right\} \times \left\{ \prod_{k=1}^{m_h} \left\{ \Gamma(w_{t_h',D_{ns,h}}(k)) / \Gamma(w_{t_{s,h},D_{s,h}}(k)) \right\} \right\} / \left\{ \Gamma(M_h) / \Gamma(m_h) \right\}, \quad [4.6]$$

where $w_{t_h'}(q) = w_{t_{s,h}}(q) + c_{t_{ns,h}}(q)$ and $w_{t_h',D_{ns,h}}(k) = w_{t_{s,h},D_{s,h}}(k) + c^{hi}_{t_h,D_{ns,h}}(k)$, for $m_h = \sum_{k=1}^{m_h} c^{hi}_{t_h,D_{s,h}}(k)$ and $m_h' = M_h - m_h = \sum_{k=1}^{m_h'} c^{hi}_{t_h,D_{ns,h}}(k)$. The ***full posterior*** is then given by the product of the posteriors within each stratum, since these strata are independent and all strata in the population are in the sample:

$$p(D_{ns} \mid D_s) = \prod_{h=1}^{H} p(D_{ns,h} \mid D_{s,h}) \,. \qquad [4.7]$$

A Monte Carlo procedure to simulate from this posterior distribution is then given as follows:

(i) Draw the $N_h - n_h$ nonsampled clusters in the population based on the Polya posterior distribution independently for each stratum. Each of the sampled clusters is resampled with probability

$$\varsigma_{hi} = \frac{w_{t_{s,h}}(i) - 1 + l_{hi,k-1} \times \left(\frac{N_h - n_h}{n_h} \right)}{N_h - n_h + (k-1) \times \left(\frac{N_h - n_h}{n_h} \right)}, k = 1, ..., N_h - n_h + 1, \qquad [4.8]$$

where $l_{hi,k-1}$ is the number of times that the i^{th} cluster in the h^{th} stratum has been resampled at the *(k-1)*th resampling, and $w_{t_{s,h}}(i)$ is the weight for the i^{th} sampled cluster in the h^{th} stratum which is normalized to sum up to the total number of clusters, i.e. $\sum_{i=1}^{n_h} w_{t_{s,h}}(i) = N_h$.

(ii) From Step 1, form a population of clusters $\{c_{11}, c_{12}, ..., c_{1n_1}, c_{11}^*, c_{12}^*, ..., c_{1N_1-n_1}^*, ..., c_{H1}, c_{H2}, ..., c_{Hn_H}, c_{H1}^*, c_{H2}^*, ..., c_{HN_H-n_H}^*\}$. Record the number of times each of the clusters from the original sample appears in the FPBB population of clusters, denoted by $\tau_{hi}, i = 1, ..., n_h, h = 1, ..., H.$, and $\sum_{h=1}^{H} \sum_{i=1}^{n_h} \tau_{hi} = N$. Then update the within cluster *element-level conditional weights* as follows: $w_{j|hi}^* = w_{j|hi} \times \tau_{hi}, i = 1, ..., n_h, h = 1, ..., H.$, where $w_{j|hi}$ is the inverse of the conditional probability that element *j* is selected given cluster *i* in stratum *h* is selected. Now pool all elements from these clusters together and treat

them as a single *FPBB sample* (i.e., as if they have no stratum or cluster boundaries). Note that this FPBB sample has the same sample size $m = \sum_{h=1}^{H} \sum_{i=1}^{n_h} m_{hi}$ but different sampling weights than the original sample. We then apply the weighted FPBB again to these pooled elements to generate $M - m$ units from the *m* units in the FPBB sample. We resample from each of the resampled clusters $M - m$ elements, cycling through $M - m$ times and resampling with probability

$$\lambda_{j|hi} = \frac{w^{*}_{j|hi} - 1 + l_{hij,k-1} \times \left(\frac{M-m}{m} \right)}{M - m + (k-1) \times \left(\frac{M-m}{m} \right)}, \quad k = 1, ..., (M - m + 1), \qquad [4.9]$$

where $l_{hij,k}$ is the number of times that the j^{th} element in the i^{th} cluster in the h^{th} stratum has been resampled at the k^{th} resampling, and $w_{j|hi}$ is the updated conditional weight for the j^{th} element in the i^{th} cluster in the h^{th} stratum. Again, they are normalized to sum up to the total number of units in the entire population, i.e. $\sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{j|hi} = M$. Thus we create a single synthetic population.

Repeat step 2 *B* times to obtain *B* FPBB synthetic populations.

(iii) Repeat steps 1-2 *L* times to obtain *L* bootstrap samples, yielding $L \times B$ FPBB populations $P^{syn}_{(lb)} = (P^{syn}_{(lb)obs}, P^{syn}_{(lb)mis}), l = 1, ..., L, b = 1, ...B.$, each of which consists of both responding elements and nonresponding elements on a vector of variables $\{Y, X, Z, R\}$.

4.3.1.2 Bootstrap — Weighted Finite Population Bayesian Bootstrap (SYN2)

Because we often do not know the first- and second-stage weights in public-use dataset, we consider an alternative to the procedure proposed in subsection 4.3.1.1. Rather than obtaining a sample of clusters from a draw from a Polya posterior, we use replication methods (Rust & Rao, 1996) to capture the cluster-level sampling variance. The final sampling weights instead of the adjusted element-level conditional weights are then directly used as input in the second-stage weighted FPBB. We use Rao and Wu (1988)'s rescaling bootstrap, which is a generalized extension of McCarthy and Snowden (1985)'s "with replacement bootstrap". The reason for choosing replication methods, particularly this special type of bootstrap, is three-fold: 1) replication methods in general are simple and direct to implement, and the formation of replicate samples further creates replicate weights that inherently represent the effects of a complex sample design in addition to reflecting the effects of a wide range of reweighting techniques such as calibration weighting; 2) among various replication methods, while the standard delete-1 jackknife is known to give an inconsistent variance estimator for a quantile as a classic example of nonsmooth statistics, the bootstrap yields sensible estimates for a variety of estimators. In fact, for stratified multistage samples, both *consistency* of the bootstrap variance estimators and confidence intervals for both smooth statistics and nonsmooth statistics have been established by researchers. Examples include functions of sample means by Rao and Wu (1988), and sample quantiles and sample low income proportions by Shao and Chen (1998). *Asymptotic consistency* of the Balanced Repeated Replication (BRR) and the bootstrap has also been established by Shao and Wu (1992) and Rust and Rao (1996). The bootstrap is adopted here because its application extends readily to more PSU sample allocations other than two-PSU-per-stratum design; 3) The Rao-Wu variant

of the conventional bootstrap yields adequate and stable variance estimates when the sample sizes are small, which occurs most often with stratified multistage sampling where only a small number of PSUs are selected within each stratum. Once the PSUs have been sampled, we continue with the weighted FPBB approach to complete the synthetic population data generation. The proposed procedure is as follows:

(i) Select a sample of $n_h^* = n_h - 1$ PSUs from the parent sample in each stratum via SRSWR sampling;

(ii) Apply the "ultimate cluster principle" (Wolter, 2007), that is, once a PSU is taken into the bootstrap replicate, all elements in that PSU are taken into the replicate also. Thus, we obtain our first bootstrap sample;

(iii) Repeat the previous steps L times to obtain L bootstrap samples $\{Boot\_l, l = 1, ..., L.\}$;

(iv) Within each bootstrap sample, update the element-level sampling weights as:

$$w_{hij}^* = w_{hij} \times (\tau_{hi} \frac{n_h}{n_h^*}) = \begin{cases} = \dfrac{n_h}{n_h - 1} w_{hij}, \text{ if the } i^{th} \text{ PSU selected in the bootstrap sample} \\ = 0, \text{ if the } i^{th} \text{ PSU not selected in the bootstrap sample} \end{cases}$$

As w_{hij}^* itself implicitly carries over the strata and PSU information in addition to unequal inclusion probability, we can drop the subscripts hi henceforth by pooling all elements in the bootstrap sample regardless of which stratum and PSU they originally came from. Normalize w_j^*'s to sum up to m^*: $\sum_{j=1}^{m^*} w_j^* = m^*$, where m^* is the bootstrap sample size.

(v) For the l^{th} bootstrap sample, $l = 1, ..., L$, apply the weighted FPBB algorithm to create an entire population $D = (D_{ns}, D_s^*)$ based on the posterior predictive

distribution of elements in the nonsampled population

$D_{ns} = \{(Y_j, X_j, Z_j, R_j), j = m^* + 1, ..., M\}$ given the elements in the bootstrap sample $D_s^* = \{(Y_j, X_j, Z_j, R_j), j = 1, ..., m^*\}$:

$$p(D_{ns} \mid D_s^*) = \frac{\int_0^1 ... \int_0^1 \prod_{j=1}^{K-1} \lambda_j^{w_j^* + r_j - 1} (1 - \sum_{j=1}^{K-1} \lambda_j)^{w_K^* + r_K - 1} d\lambda_1 ... d\lambda_{K-1}}{\int_0^1 ... \int_0^1 \prod_{j=1}^{K-1} \lambda_j^{w_j^* - 1} (1 - \sum_{j=1}^{K-1} \lambda_j)^{w_K^* - 1} d\lambda_1 ... d\lambda_{K-1}}$$
$$= \frac{\prod_{j=1}^{K} \Gamma(w_j^* + r_j) / \Gamma(w_j^*)}{\Gamma(M) / \Gamma(m^*)}, \qquad [4.10]$$

where r_j is the number of elements in the nonsampled FPBB population that take on the same value as the j^{th} element in the bootstrap sample, for $j = 1, ..., K(\equiv m^*)$. Operationally, we draw a Polya sample of size $M^* = M - m^*$ from $mult(M^*; \lambda_1, ..., \lambda_K)$ where the selection probability $\lambda_k, k = 1, ..., K.$ is a function of w_j^*:

$$\lambda_k = \frac{w_j^* - 1 + 1_{j,k-1} \times \left(\frac{M^*}{m^*} \right)}{M^* + (k-1) \times \left(\frac{M^*}{m^*} \right)}, \; k = 1, ..., M^* + 1., \qquad [4.11]$$

Repeat Step **(v)** for *B* times to obtain $L \times B$ FPBB populations.

4.3.2 Imputation of the synthesized populations

Once the set of FPBB synthetic populations $P^{syn} = \{P_{(b)}^{(l)}, l = 1, ..., L, b = 1, ..., B\}$, where $P_{(b)}^{(l)} = (Y_{(b)mis}^{(l)}, P_{(b)obs}^{(l)})$ are created using either SYN1 method or SYN2 method, we generate imputations $P^{imp} = \{P_{(ba)}^{(l)}, l = 1, ..., L, b = 1, ..., B, a = 1, ..., A.\}$ from the posterior predictive distribution $p(Y_{(b)mis}^{(l)} \mid P_{(b)obs}^{(l)})$ based on a parametric model that does not

condition on sample design features, i.e. a model taking similar form as the SRS model stated in section 4.2.1. We consider imputations based on the covariate (*X*) only (SYN1_srs or SYN2_srs) or imputations that include the log of the sample weights in the linear predictors (SYN1_lwt or SYN2_lwt).

We obtain the MI inference by applying the combining rules developed in Zhou et al. (2013a) to the $L \times B \times A$ estimates, based on both the observed set $P_R = \{P_{(b)obs}^{(l)}, b = 1,...,B, l = 1,...,L\}$ and the imputed set $P_{\bar{R}} = \{Y_{(ba)mis}^{(l)}, l = 1,...,L, b = 1,...,B, a = 1,...,A\}$ of the synthetic populations, where R and $\bar{R}$ represent responding and nonresponding, respectively. We estimate the *population mean of* Y by calculating the mean of the synthetic population

$$\hat{\bar{Y}}_{lba}^{P} = [\sum_{i \in P_R} y_i + \sum_{j \in P_{\bar{R}}} y_j] / M , \qquad [4.12]$$

To estimate *a generalized linear regression parameter*, we solve the regression score function

$$U(\beta)_{lba}^{P} = [\sum_{i \in P_R} x_i^T (y_i - g^{-1}(x_i^T \beta)) + \sum_{j \in P_{\bar{R}}} x_j^T (y_j - g^{-1}(x_j^T \beta))], \qquad [4.13]$$

where $g(\bullet)$ is the link function for the transformation of the mean, so that $\hat{\beta}_{lba}^{P} = \{\beta : U(\beta)_{lba}^{P} = 0\}$. For *quantile estimation*, we proceed by first obtaining the empirical distribution function based on the lba^{th} imputed synthetic population:

$$\hat{F}_{lba}^{P}(y) = [\sum_{i \in P_R} I(y_i < y) + \sum_{j \in P_{\bar{R}}} I(y_j < y)] / M , \qquad [4.14]$$

Then we estimate the γ^{th} quantile (q_γ) as:

$$\hat{q}_{\gamma,lba} = (\hat{F}_{lba}^{P})^{-1}(\gamma) = \inf\{y : \hat{F}_{lba}^{P}(y) \geq \gamma\} , \qquad [4.15]$$

The MI point estimator for the population statistic of interest *Q* (mean, regression

estimator, quantile) is then given by the mean of the lba^{th} point estimators:

$$\hat{Q}_{MI} = \frac{1}{LBA}\sum_l \sum_b \sum_b \hat{Q}_{lba}\,, \qquad [4.16]$$

The MI variance estimator is:

$$\hat{V}_{MI} = (1+L^{-1})V_L = (1+L^{-1})\frac{1}{L-1}\sum_l (\hat{Q}_l - \hat{Q}_{MI})^2, \text{ where } \hat{Q}_l = \frac{1}{BA}\sum_b \sum_a \hat{Q}_{lba}\,, \qquad [4.17]$$

We then construct the 95% interval estimate for quantiles based on *t* reference distribution with degrees of freedom equal to $\min\{v_{com} = \sum_h n_h - H, v_{syn} = L-1\}$. These results arise from the fact that, by the standard Rubin (1987) MI combining rules, we have

$$Q \mid P^{imp} \dot{\sim} t_{L-1}(\bar{Q}_L, (1+L^{-1})V_L)\,, \qquad [4.18]$$

where $\bar{Q}_L = \frac{1}{L}\sum_l \tilde{Q}^{(l)}$, $V_L = \frac{1}{L-1}\sum_l (\tilde{Q}^{(l)} - \bar{Q}_L)^2$, and $\tilde{Q}^{(l)} = \lim_{\substack{B\to\infty \\ A\to\infty}} \frac{1}{BA}\sum_b \sum_a \hat{Q}_{lba}$. Replacing $\tilde{Q}^{(l)}$ with its finite simulation estimator $\hat{Q}_l$ replaces $\bar{Q}_L$ with $\hat{Q}_{MI}$ and gives the results above. Note that the generation of the synthetic population sets the within imputation variance to 0 so that the posterior variance of *Q* can be obtained using the between-bootstrap variance only; see Dong et al. (2014) and Zhou et al. (2013a). The result assumes that $E(\hat{q}_{ba}) = Q$ – a result guaranteed by our Bayesian bootstrap estimator if the imputation model is also correct – as well as a sufficiently large sample size for the *t* approximation to be reasonable.

Lo (1988) showed that the variance estimator for the FPBB mean in a simple random sample setting should be inflated by the factor $(\frac{n+1}{n-1})$. Thus in the double-weighted FPBB (SYN1) setting, a small sample correction to the variance estimate needs

to be used when the number of clusters per stratum is small. When $n_h = a$ is a constant across all strata, we use $\frac{n_h+1}{n_h-1}(1+L^{-1})\,V_L$; otherwise we suggest $\frac{\bar{n}_h+1}{\bar{n}_h-1}(1+L^{-1})\,V_L$, where $\bar{n}_h = H^{-1}\sum_h n_h$.

4.4 Simulation Study

We conducted a simulation study to investigate the performance of the proposed method for incorporating stratified cluster sampling effects in multiple imputation. We targeted three population statistics: 1) population quantiles, 2) proportions of binary event data, and 3) logistic regression parameters relating the covariate to the binary data. The simulation is a 2×2 factorial design based on the following factors: 1) keeping the first stage sampling plan constant, we let the subsampling rate f_2 of elements within sampled clusters be a) independent of or b) dependent on the stratum effects, and 2) assume a) the missingness on the *Y*-variable (continuous or binary) depends only on the covariate (X) (MAR_X), or b) depends on both X and the final sampling weight W (MAR_X,W).

We focus on a two-PSU-per-stratum sample design, both because it is a common design, especially in public use settings, and because it is a "limiting case" in terms of the number of PSUs per stratum. In addition to the two variants of our synthetic MI estimators, we consider standard parametric MI under the SRS, appropriate fixed effect (FX_APR), and appropriate random effect (RE_APR) models.

4.4.1 Description of the Design

- **Data generation**

Let i be the index for strata, j be the index for clusters, and k be the index for elements. Suppose there are 50 strata in the population. First, the number of PSUs in each stratum was randomly determined according to a uniform distribution, i.e. $C_i \sim Unif(2,54), i = 1,...,50$; second, the number of population elements within PSUs was randomly generated as $N_{ij} \sim Unif(20,80), i = 1,...,50,\ j = 1,...,C_i$. Thus we obtained a population of size $N = 67385$. The complete data for four survey variables $Y = (Y_1, Y_2, Y_3, Y_4)^T$ were generated from a super-population model according to a two-step process: In the first step, Y_1 and Y_2 were randomly selected from a bivariate linear mixed effects model; let $N_2(\cdot)$ denote a bivariate normal distribution function:

$$\begin{pmatrix} Y_{1ijk} \\ Y_{2ijk} \end{pmatrix} \sim N_2(\mu, \Sigma), \text{ where } \mu = \begin{bmatrix} \beta_1 + S_i + u_{1ij} + \varepsilon_{1ijk} \\ \beta_2 + u_{2ij} + \varepsilon_{2ijk} \end{bmatrix}, \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}, \qquad [4.19]$$

Let $\beta_1 = \beta_2 = 15$ be the fixed covariate effects, $S_i = \frac{i}{5}$ be the fixed stratum effects, and let $\begin{bmatrix} u_{1ij} & u_{2ij} \end{bmatrix}^T$ and $\begin{bmatrix} \varepsilon_{1ijk} & \varepsilon_{2ijk} \end{bmatrix}^T$ be the random cluster effects and random error terms drawn from two independent bivariate normal distributions: $N_2(0, \Sigma_u)$ and $N_2(0, \Sigma_\varepsilon)$. Elements of Σ_u were set as: $\sigma^2_{u_1} = 4, \sigma^2_{u_2} = 1, \sigma_{u_1 u_2} = 0.2,$ and elements of Σ_ε were set as: $\sigma^2_{\varepsilon_1} = 4, \sigma^2_{\varepsilon_2} = 3, \sigma_{\varepsilon_1 \varepsilon_2} = 1.732.$ This results in conditional intraclass correlations (ICC) of Y_1 and Y_2 as $\rho_{Y_1} = 0.5$ and $\rho_{Y_2} = 0.25$ (note that the unconditional ICC for the two variables may be smaller than these values). In the second step, a random effects logistic regression model (Anderson & Aitkin, 1985; Stiratelli, Laird, & Ware, 1984) was used to simulate two binary outcome variables Y_3 and Y_4 as a function of Y_2. Under this model, a random effect is added to the linear part of the logistic regression model for each element

in the cluster. The conditional mean of Y_{3ijk} and Y_{4ijk} is

$$\pi_{ijk} = E(Y_{\cdot ijk} \mid Y_{2ijk}, u_{\cdot ij}) = \Pr(Y_{\cdot ijk} = 1 \mid Y_{2ijk}, u_{\cdot ij}) = \frac{e^{\alpha_0 + \alpha_1 S_i + \alpha_2 Y_{2ijk} + u_{\cdot ij}}}{1 + e^{\alpha_0 + \alpha_1 S_i + \alpha_2 Y_{2ijk} + u_{\cdot ij}}}, \qquad [4.20]$$

where $u_{3ij} \sim N\left(0, 6^2\right)$, $u_{4ij} \sim N\left(0, 10^2\right)$ and $\alpha = (\alpha_0, \alpha_1, \alpha_2)^T$ is the vector of fixed covariate effects. We fixed $\alpha_2 = 1.5$ and vary α_0 and α_1 to obtain two different binary variables Y_{3ijk} and Y_{4ijk}, with either moderate ($\alpha_0 = -5, \alpha_1 = -1.5$) or rare probabilities ($\alpha_0 = -8, \alpha_1 = -6$). Given $u_{\cdot ij}$, the $Y_{\cdot ijk}$'s in the cluster are independent Bernoulli variables, that is, $Y_{\cdot ijk} \mid u_{\cdot ij} \sim Bern(\pi_{ijk})$.

Figure 4.1 shows the correlations among variables in the simulated population, where the different shades of grey represent different degrees of association between any of the two variables. The darker shades indicate higher correlation. All survey outcome variables (Y_1, Y_3, Y_4) have moderate to strong (0.2~0.8) stratum effect (H or *strID*) and clustering effect (U_1, U_3, U_4), indicating that accounting for these effects in the analysis of missing data is essential.

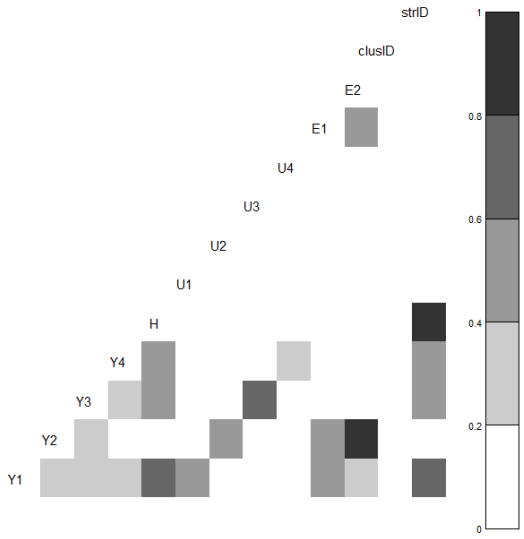

Figure 4.1 Correlation among variables in the simulated population (darker shades = higher correlation)

- **The Sample Design**

Within each stratum, we drew a two-stage cluster sample according to the following procedure: first, we drew a sample of 2 PSUs without replacement with probability proportional to the cluster size $f_{1ij} = \frac{2 * N_{ij}}{\sum_j N_{ij}}$; second, we sampled elements from each sampled cluster by two different subsampling schemes: 1) sampling probability independent of S_i which was defined in [4.19]: SRS with an equal sampling fraction of $f_{2k|ij} = 1/5$; 2) sampling probability related to S_i: SRS with varying sampling fractions across strata, i.e. $f_{2k|ij} = \text{expit}(-0.8 - 0.12 * S_i)$, where $\text{expit}(x) = 1/(1+e^{-1}(x))$. An average of 1122 elements are selected in each of the 200 simulation replications. The distributions of sampling weights are shown in Figure 4.2. The distributions of sampling weights under the two subsampling schemes are generally very similar with somewhat more skewness under subsampling scheme 2.

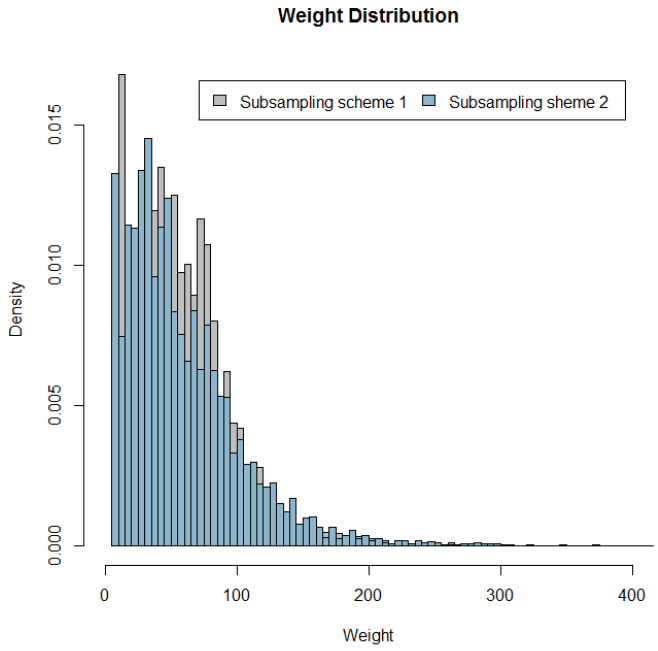

Figure 4.2 Distribution of weights under the two subsampling schemes

- **Imposing missingness**

Throughout the simulation study, we assume that Y_2 is always completely observed and we impose missing values on Y_1, Y_3 and Y_4 independently according to the following deletion function conditional on Y_2 and/or log transformation of the weight:

$$\Pr(R=0\,|\,Y_2,W)=\frac{\exp(\lambda_0+\lambda_1*Y_2+\lambda_2*\log(W))}{1+\exp(\lambda_0+\lambda_1*Y_2+\lambda_2*\log(W))}, \qquad [4.21]$$

where *R* is the response indicator and *W* is the overall sample weight. Setting $\lambda_2=0$, we obtain the first MAR mechanism (i.e. MAR_X, note we treat Y_2 as *X* here), under which we further set $\lambda_0=3.42,\lambda_1=-0.2$ and $\lambda_0=-2.58,\lambda_1=0.2$ for deleting values on Y_1 and Y_3,Y_4, respectively. Setting $\lambda_2=-0.6$, we obtain the second MAR mechanism (i.e. MAR_X,W), under which we fix $\lambda_1=0.2$ and set two values on λ_0 ($=-0.274$ or -0.33)

for deleting values on all three outcome variables under subsampling scheme 1 and subsampling scheme 2, respectively. All deletion functions result in approximately 40% missingness.

- **Parametric Multiple Imputation**

Both simple random sample SRS (including SRS, SYN1_srs and SYN2_srs) and fixed-effected model FX_APR can be implemented in R (R Core Team, 2013) using *mice* routines; for the logistic model associated with the binary outcome, the method '*logreg*' must be specified. We use the *pan* package (Schafer, 1997b) in R for the mixed effects imputation (RE_APR) for the missing continuous outcome; logistic mixed effects imputation is programmed in SAS for the missing binary outcome, as there is no package readily available for use.

- **Parameters of interest and inference**

We focus on inference for the following population parameters: the mean of the continuous variable Y_1, the mean of the binary variables Y_3 and Y_4 (i.e. Bernoulli proportions), linear regression coefficients of Y_1 on Y_2, logistic regression coefficients of Y_3 (or Y_4) on Y_2, and the population percentiles of the continuous variable Y_1.

Weighted analyses and sandwich variance estimators accounting for strata and clusters are used to estimate smooth statistics (including proportions and regression parameters) under the three fully parametric MI methods. For estimating quantiles of the distribution of a continuous survey variable, we construct the sample-weighted point estimator with confidence intervals based on the test inversion method (Francisco & Fuller, 1991). We chose the test-inversion method instead of Woodruff's method (Woodruff, 1952) despite the computational intensity, because the literature suggests that

it may outperform Woodruff in heavily stratified samples or in small-to-moderate-sized samples (Kovar et al., 1988), and these are the sample designs we are particularly interested in in this chapter. Based on the a^{th} imputed dataset, the empirical distribution function can be written as

$$\hat{F}^{(a)}(y) = [\sum_{S_R} w_{hij} I(y_{hij}^{obs} < y) + \sum_{S_{\bar{R}}} w_{hij} I(y_{hij}^{(a)} < y)] / \sum_{S} w_{hij} , \qquad [4.22]$$

where S_R and $S_{\bar{R}}$ are subsets of the sample data S, consisting of respondents and nonrespondents, respectively. The estimator $\hat{F}(y)$ and its associated estimated variance $v(\hat{F}(y))$ can then be obtained using the variance estimator proposed by Francisco and Fuller (1991) together with standard Rubin combining rules as previously described. The sample γ^{th} quantile estimator thus is $\hat{q}_\gamma = (\hat{F})^{-1}(\gamma)$, with 95% asymptotic confidence interval given by

$$[L,U] = \left[[\hat{F}]^{-1}\left(\gamma - t_{0.025}\sqrt{\text{var}(\hat{F}(q_\gamma))}\right),\ [\hat{F}]^{-1}\left(\gamma + t_{0.025}\sqrt{\text{var}(\hat{F}(q_\gamma))}\right)\right], \qquad [4.23]$$

4.4.2 Results

Figures 4.3 through 4.6 plot the point estimates as well as their upper and lower confidence interval lines for 19 population quantiles (from 0.05 to 0.95 with an increment of 0.05). These were obtained from the two proposed finite population Bayesian bootstrap procedures (SYN1 and SYN2). These graphs demonstrate visually how the proposed methods work in incorporating complex sample design features for estimating population quantiles. From left to right in each figure, we compare the proposed methods in the absence of missing data (synthetic BD), the proposed methods in the presence of missing data under mechanism MAR_X (synthetic ADX) and under MAR_X,W

(synthetic ADXW), with the design-based quantile estimation method based on the actual replication samples (Complete Data). The perfect overlapping of the red plot and black plot indicates that all point and interval estimates obtained from the synthetic BD (both SYN1 and SYN2) are identical to those using the design-based method. This provides good evidence that both proposed procedures were able to accommodate all sample design features to produce synthetic populations that behaved as simple random samples from the underlying true population. In the missing data setting, the point estimates are generally very close. The variance estimates increased as expected — note the green dashed lines always encompass the black dashed lines — due to the added noise from multiply imputing missing data. Hence, we further investigate the performance of the synthetic MI in comparison with fully parametric MI methods, by looking at several key summary measures under the four simulation conditions in Table 4.1.

Table 4.1 compares the average width $\times 10^2$ and average coverage rates of the 95% CI of $q(\alpha)$, where $\alpha = 0.05, 0.10, 0.25, 0.50, 0.75, 0.90$ and $0.95,$ corresponding to seven selected population quantiles. Among all methods considered, the SRS imputation model yields the poorest coverage. This results from the compounding effects of biases and variance underestimation, due to ignoring stratum effects and clustering effects, respectively. As we increase the dependence of both the sampling mechanism and response mechanism on stratum effects and sampling weights, the performance of SRS gets even worse, as exhibited by the markedly increased RelBias and decreased coverage rates. In addition, ignoring stratum and/or weight effects that are highly relevant to either mechanism seems to impact the median and 2^{nd} and 3^{rd} quartiles more than the tail quantiles under SRS, as evident in the relatively lower coverage rates in the right block of

Table 4.1.

The FX_APR model (Reiter et al., 2006; Rubin, 1996; Schenker et al., 2006), generally performs fairly well in our simulation study with respect to estimation of population quantiles. There is some modest underestimation of the small percentile quartiles with the second stage sampling constant. The RE_APR model also performs well, with the exception of moderate to high overcoverage when the second stage sampling probability is associated with the stratum mean and the missingness mechanism.

In contrast, our synthetic MI (SYN2 in particular) compares favorably with all of its competitors, and in most cases yield comparable results to the RE_APR, which is regarded as a "gold standard" as it is compatible with the data generating mechanism. There is some undercoverage when the stratified double-weighted FPBB estimator (SYN1) is used, perhaps due to the fact that the Lo small-sample adjustment is not as accurate when $n_h = 2$. However, use of a stratified bootstrap-weighted FPBB estimator (SYN2) generally eliminates this issue. Although an imputation model assuming SRS suffices for the synthetic MI method in most scenarios, we need to include the sampling weight as a predictor when the outcome Y and the response indicator R are strongly associated with each other through the sampling mechanism I, as is the case with the second subsampling scheme, when both the missingness indicator and the second-stage sampling rate are functions of the stratum mean.

Tables 4.2 and 4.3 compare the absolute relative bias $relbias = 100 \times \frac{|\hat{\theta} - \theta_{complete}|}{\theta_{complete}}\%$, RMSE and 95% nominal CI coverage for the estimated mean/proportions of Y_1, Y_3 and Y_4, and the slopes of the three outcome variables on Y_2,

respectively. ($\theta_{complete}$ is the estimated parameter with complete data, and $\hat{\theta}$ is the estimated parameter under one of the different MI methods.) As in the estimation of the quantiles, the SRS imputation model is biased and has poor coverage as it ignores stratum and cluster effects. Again, dependence of subsampling on stratum effects and dependence of response on sampling weights damage the performance of SRS even further.

FX_APR generally performs well in estimating the mean of a continuous variable (Y_1) and a regular binary variable (Y_3) with moderate probability as well as the slopes. However, it fails for proportion estimation for rare events data (Y_4), yielding biased point estimates and less than nominal coverage throughout all scenarios. One interpretation might be that overfitting occurs when including too many dummies to account for fixed strata and cluster effects. This damages the predictive efficacy when the fitted model is used for drawing missing values. The problem is particularly prominent when the logistic fixed effects imputation model is used along with the current sampling design where an average of only 10 elements are selected per PSU within each stratum; and this results in even more substantial biases on $\bar{Y}_4$ than the SRS model.

Compared with FX_APR, RE_APR avoids the overfitting issue through shrinkage effects: note that under RE_APR, we pooled all PSUs from all strata as if there were no strata bounds, and the stratum effects can be thought as being implicitly modeled in the random intercept term ($u_j = I_h + u_{h(j)}$). This in some sense alleviates the small sample issue though not a complete solution (e.g. the estimated $\bar{Y}_4$ under RE_APR is still moderately biased).

As in the quantile estimation setting, our synthetic MI compares favorably with

all of its competitors, and in most cases yields comparable results to the RE_APR for estimation of means and logistic regression parameters. In the case of rare events data, our proposed new method increases the analytical size through generating synthetic population data thus is even superior to RE_APR, consistently yielding negligible biases and close to nominal coverage. The impact of ignoring the weights in the imputation (under MAR_X,W mechanism) is less than in the quantile estimation setting, with the exception of the estimation of the continuous mean $\bar{Y}_1$, where including the weight is required to obtain approximately correct coverage.

A disadvantage of the method lies in the relative inefficiency for estimating nonlinear parameters (regression coefficients) (e.g. the synthetic MI results in unbiased point estimates but a larger RMSE than the two model-based MI methods). This is typical in that nonparametric methods cannot typically compete with their fully parametric counterparts under the correct model, and is a tradeoff made to improve robustness to model misspecification, and, in our setting, simplicity in implementation.

(a) (b) (c)

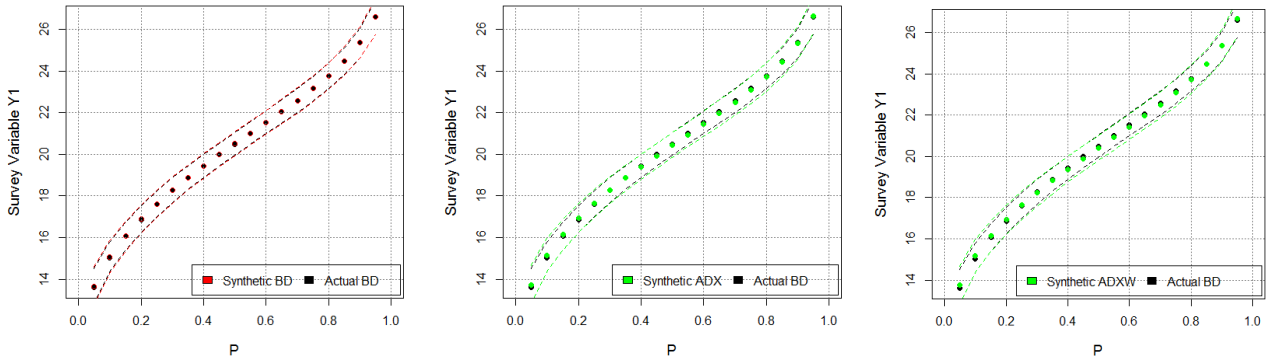

Figure 4.3 Comparison of point and interval estimation for 19 population quantiles using the Stratified Boot-FPBB and the design-based complete data analysis for subsampling scheme1 ($f_2 \propto$ constant). (a) before deletion (b) imputation based on covariates under MAR_X (c) imputation based on covariates and sampling weights under MAR_X,W.

(a) (b) (c)

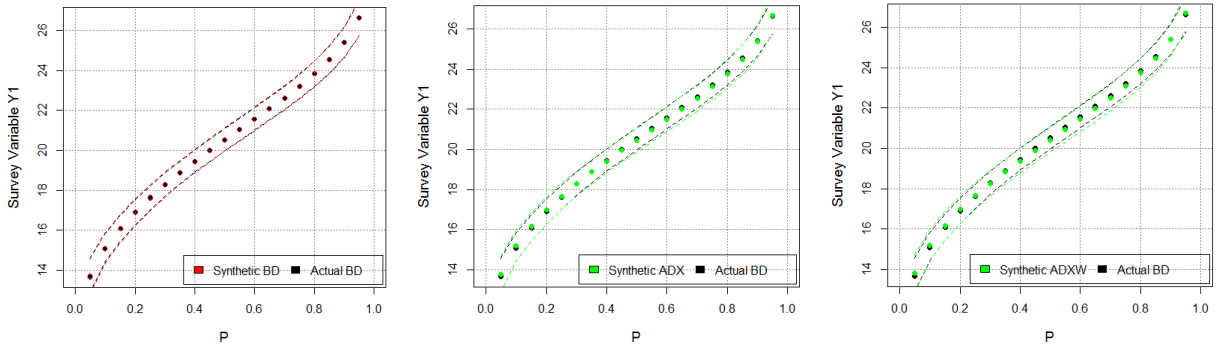

Figure 4.4 Comparison of point and interval estimation for 19 population quantiles using the Stratified Boot-FPBB and the design-based complete data analysis for subsampling scheme2 ($f_2 \propto h(S_i)$). (a) before deletion (b) imputation based on covariates under MAR_X (c) imputation based on covariates and sampling weights under MAR_X,W.

(a) (b) (c)

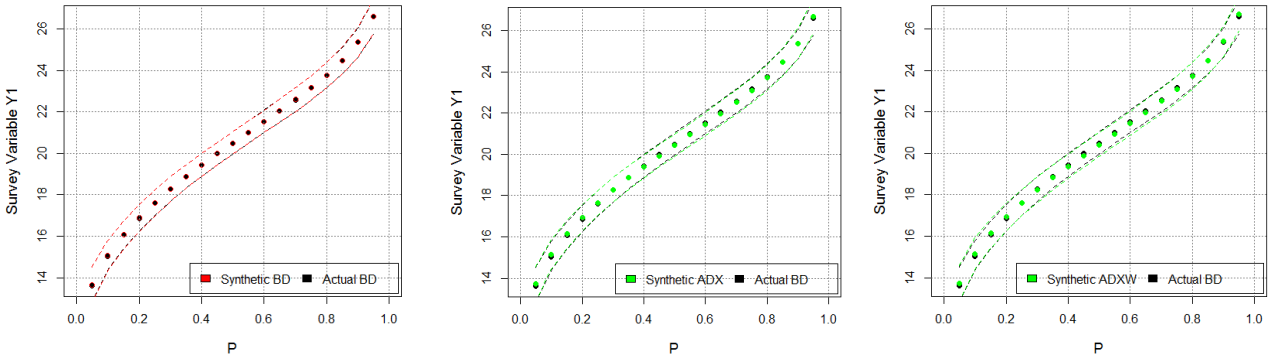

Figure 4.5 Comparison of point and interval estimation for 19 population quantiles using the stratified double weighted-FPBB and the design-based complete data analysis for subsampling scheme1 ($f_2 \propto$ constant). (a) before deletion (b) imputation based on covariates under MAR_X (c) imputation based on covariates and sampling weights under MAR_X,W.

(a) (b) (c)

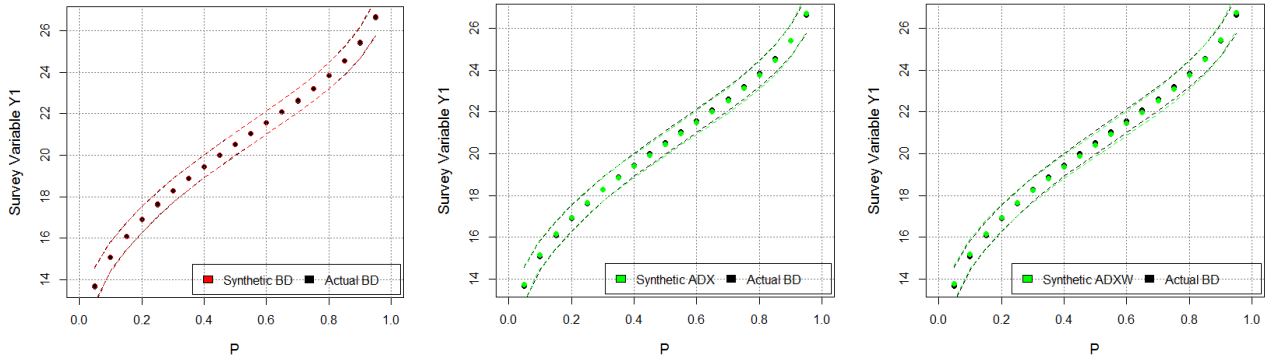

Figure 4.6 Comparison of point and interval estimation for 19 population quantiles using the stratified double weighted-FPBB and the design-based complete data analysis for subsampling scheme2 ($f_2 \propto h(S_i)$). (a) before deletion (b) imputation based on covariates under MAR_X (c) imputation based on covariates and sampling weights under MAR_X,W.

Table 4.1 Comparison of average width $\times 10^2$ and 95% CI coverage rates of $q(\alpha)$ for $\alpha = 0.05, 0.10, 0.25, 0.50, 0.75, 0.90$ and 0.95.

Sampling Scheme	Missingness Mechanism	Methods	Average width of 95% CI x 10^2							95% CI coverage						
			0.05	0.1	0.25	0.5	0.75	0.9	0.95	0.05	0.1	0.25	0.5	0.75	0.9	0.95
$f_2 \propto$ const.	Complete Data	Actual	170	144	123	106	116	142	165	90.5%	92.5%	94.5%	93.5%	95.5%	91.5%	91.0%
		Syn1BD	172	144	126	105	117	146	165	90.5%	90.0%	95.0%	94.0%	95.0%	94.0%	89.0%
		Syn2BD	182	154	132	113	122	150	171	94.0%	95.0%	96.0%	94.5%	96.5%	96.0%	92.5%
	MAR_X	SRS	165	134	112	101	108	132	158	93.0%	91.5%	86.0%	82.5%	83.0%	89.0%	93.5%
		FX_APR	171	143	120	105	116	146	172	92.5%	90.5%	90.5%	92.5%	93.5%	94.0%	95.0%
		RE_APR	184	154	131	115	125	156	186	93.5%	94.0%	93.0%	97.5%	95.5%	95.5%	97.0%
		Syn1_srs	171	145	123	109	122	148	165	91.0%	89.5%	92.5%	95.0%	90.5%	89.5%	94.0%
		Syn2_srs	182	158	134	118	129	156	175	93.5%	93.0%	94.5%	96.5%	94.5%	94.5%	95.0%
	MAR_X,W	SRS	178	146	120	109	110	139	163	89.0%	81.0%	70.5%	69.0%	80.0%	90.0%	91.0%
		FX_APR	186	153	126	115	125	155	190	89.5%	92.5%	93.5%	95.5%	92.5%	92.5%	96.0%
		RE_APR	197	166	140	127	136	168	197	95.0%	97.0%	97.0%	98.0%	96.0%	95.0%	96.0%
		Syn1_srs	173	150	124	111	119	146	163	91.5%	92.0%	93.0%	91.5%	90.0%	94.0%	92.5%
		Syn2_srs	183	160	134	119	123	153	172	93.5%	95.5%	96.5%	92.5%	92.0%	93.0%	95.5%
		Syn1_lwt	174	151	126	115	124	148	166	90.0%	89.0%	93.0%	94.5%	90.5%	96.0%	94.0%
		Syn2_lwt	184	161	136	122	132	155	174	92.0%	93.0%	95.5%	96.0%	94.5%	96.0%	95.0%
$f_2 \propto h(S_i)$	Complete Data	Actual	170	143	120	110	121	148	169	92.5%	94.5%	95.0%	96.0%	92.5%	87.5%	87.5%
		Syn1BD	177	142	120	108	121	152	175	91.0%	92.5%	92.0%	94.5%	92.5%	87.5%	87.5%
		Syn2BD	182	152	128	116	126	154	178	95.0%	97.0%	96.0%	97.0%	94.5%	90.0%	90.5%
	MAR_X	SRS	175	139	121	111	116	141	169	86.5%	73.0%	57.0%	48.5%	61.0%	72.0%	80.5%
		FX_APR	174	142	121	113	124	162	202	95.5%	95.0%	98.0%	95.5%	93.5%	92.5%	95.5%
		RE_APR	181	150	128	119	131	168	205	94.0%	96.5%	97.0%	96.5%	97.0%	94.0%	96.0%
		Syn1_srs	166	140	119	111	126	156	180	93.5%	94.0%	96.5%	92.5%	92.0%	91.0%	90.0%
		Syn2_srs	179	152	129	119	132	162	185	94.5%	95.5%	98.0%	96.5%	95.0%	93.5%	92.5%
	MAR_X,W	SRS	191	157	127	117	122	147	168	47.0%	31.5%	9.5%	8.0%	30.0%	60.0%	73.5%
		FX_APR	186	153	125	119	138	179	227	96.5%	97.0%	93.5%	96.5%	97.0%	95.0%	94.5%
		RE_APR	190	161	135	131	148	184	220	98.0%	99.5%	97.5%	98.0%	98.5%	96.5%	95.0%
		Syn1_srs	168	146	124	114	128	155	174	94.0%	92.5%	84.0%	73.0%	76.0%	87.5%	88.0%
		Syn2_srs	184	160	135	122	134	160	179	95.0%	95.5%	88.0%	77.5%	79.0%	87.0%	89.0%
		Syn1_lwt	168	143	121	113	130	158	176	92.5%	92.5%	94.5%	92.0%	89.0%	92.0%	91.5%
		Syn2_lwt	178	155	131	122	138	166	185	96.0%	95.5%	95.5%	92.5%	95.5%	95.0%	93.0%

Table 4.2 Comparison of RelBias, RMSE and 95% CI coverage rates for the mean of Y1 and proportions of Y3 and Y4,

Population True Value: $\bar{Y}_1 = 20.4$, $P_{Y_3} = 0.608$, $P_{Y_4} = 0.117$

Sampling Scheme	**Missingness Mechanism**	**Methods**	**RelBias**			**RMSE**			**95% CI coverage**		
			$\bar{Y}_1$	P_{Y_3}	P_{Y_4}	$\bar{Y}_1$	P_{Y_3}	P_{Y_4}	$\bar{Y}_1$	P_{Y_3}	P_{Y_4}
$f_2 \propto$ const. **Actual samples BD:** $\bar{Y}_1 = 20.3$ $P_{Y_3} = 0.604$ $P_{Y_4} = 0.117$	Complete Data	Actual	-	-	-	0.220	0.042	0.024	95.0%	94.0%	90.5%
		Syn1BD	0.0%	0.0%	0.0%	0.221	0.042	0.024	94.5%	94.0%	91.5%
		Syn2BD	0.0%	0.0%	0.0%	0.222	0.043	0.024	95.0%	94.5%	93.0%
	MAR_X	SRS	0.8%	1.6%	10.8%	0.309	0.041	0.028	76.9%	90.0%	85.0%
		FX_APR	0.0%	1.3%	39.2%	0.243	0.040	0.054	91.0%	96.5%	72.5%
		RE_APR	0.0%	1.3%	15.1%	0.236	0.040	0.026	93.0%	93.5%	91.0%
		Syn1_srs	0.0%	0.3%	0.4%	0.255	0.044	0.025	94.5%	93.5%	91.5%
		Syn2_srs	0.0%	0.2%	0.4%	0.254	0.044	0.025	97.0%	95.0%	94.5%
	MAR_X,W	SRS	1.4%	2.8%	19.4%	0.398	0.042	0.035	72.0%	85.5%	77.5%
		FX_APR	0.0%	2.7%	48.4%	0.260	0.042	0.065	91.5%	96.0%	60.0%
		RE_APR	0.1%	0.3%	6.8%	0.250	0.041	0.022	97.5%	95.5%	86.0%
		Syn1_srs	0.4%	1.4%	4.2%	0.285	0.043	0.026	92.0%	95.5%	91.5%
		Syn2_srs	0.5%	1.4%	4.4%	0.283	0.043	0.026	96.5%	95.0%	96.0%
		Syn1_lwt	0.0%	0.6%	0.3%	0.273	0.045	0.027	95.5%	93.5%	89.0%
		Syn2_lwt	0.0%	0.5%	0.0%	0.271	0.045	0.026	96.0%	96.0%	94.0%
$f_2 \propto h(S_i)$ **Actual samples BD:** $\bar{Y}_1 = 20.4$ $P_{Y_3} = 0.609$ $P_{Y_4} = 0.117$	Complete Data	Actual	-	-	-	0.218	0.037	0.023	96.0%	97.5%	92.0%
		Syn1BD	0.0%	0.0%	0.0%	0.220	0.037	0.023	93.5%	94.0%	92.0%
		Syn2BD	0.0%	0.0%	0.3%	0.219	0.038	0.023	96.0%	97.0%	94.0%
	MAR_X	SRS	2.4%	4.7%	29.6%	0.540	0.048	0.045	42.0%	80.5%	62.5%
		FX_APR	0.0%	1.5%	42.0%	0.237	0.036	0.058	94.0%	97.0%	70.5%
		RE_APR	0.2%	1.6%	16.1%	0.230	0.039	0.025	96.5%	93.5%	91.5%
		Syn1_srs	0.1%	0.0%	0.9%	0.266	0.042	0.025	92.5%	95.5%	91.5%
		Syn2_srs	0.1%	0.1%	0.5%	0.266	0.042	0.025	94.0%	96.0%	93.5%
	MAR_X,W	SRS	4.4%	9.2%	54.0%	0.912	0.067	0.071	6.5%	56.0%	34.5%
		FX_APR	0.1%	1.2%	55.3%	0.288	0.037	0.074	93.5%	95.5%	55.0%
		RE_APR	0.0%	0.7%	5.1%	0.239	0.038	0.022	97.5%	95.5%	87.0%
		Syn1_srs	1.5%	3.3%	15.0%	0.401	0.045	0.033	77.5%	91.5%	88.0%
		Syn2_srs	1.5%	3.2%	15.0%	0.400	0.045	0.033	82.0%	94.5%	91.5%
		Syn1_lwt	0.1%	0.2%	0.9%	0.281	0.042	0.025	89.5%	93.0%	91.0%
		Syn2_lwt	0.0%	0.1%	1.2%	0.278	0.043	0.025	93.5%	95.5%	92.5%

Table 4.3 Comparison of RelBias, RMSE and 95% CI coverage rates for the regression coefficients of Y1, Y3 and Y4 on Y2,
Population True Value: $\beta_{1,Y_1|Y_2} = 0.488$, $\beta_{1,Y_3|Y_2} = 0.227$, $\beta_{1,Y_4|Y_2} = 0.083$

Sampling Scheme	**Missingness Mechanism**	**Methods**	**RelBias**			**RMSE**			**95% CI coverage**		
			$\beta_{1,Y_1\|Y_2}$	$\beta_{1,Y_3\|Y_2}$	$\beta_{1,Y_4\|Y_2}$	$\beta_{1,Y_1\|Y_2}$	$\beta_{1,Y_3\|Y_2}$	$\beta_{1,Y_4\|Y_2}$	$\beta_{1,Y_1\|Y_2}$	$\beta_{1,Y_3\|Y_2}$	$\beta_{1,Y_4\|Y_2}$
$f_2 \propto$ const. **Actual samples BD:** $\beta_{1,Y_1\|Y_2} = 0.481$ $\beta_{1,Y_3\|Y_2} = 0.232$ $\beta_{1,Y_4\|Y_2} = 0.086$	Complete Data	Actual	-	-	-	0.103	0.065	0.098	98.0%	96.0%	90.0%
		Syn1BD	0.4%	1.1%	1.9%	0.104	0.067	0.098	96.0%	93.5%	88.0%
		Syn2BD	0.2%	2.8%	5.0%	0.103	0.067	0.100	98.0%	97.5%	91.5%
	MAR_X	SRS	4.6%	4.6%	24.7%	0.110	0.071	0.100	93.0%	90.0%	91.0%
		FX_APR	0.2%	1.0%	44.7%	0.103	0.063	0.087	97.0%	97.0%	92.5%
		RE_APR	0.3%	2.1%	22.6%	0.100	0.056	0.068	98.0%	95.5%	95.0%
		Syn1_srs	0.0%	0.5%	2.8%	0.114	0.079	0.111	95.5%	93.0%	88.0%
		Syn2_srs	0.2%	3.0%	4.4%	0.115	0.082	0.111	96.5%	96.5%	94.5%
	MAR_X,W	SRS	7.3%	7.5%	45.6%	0.121	0.070	0.100	93.0%	90.5%	87.0%
		FX_APR	0.4%	1.7%	53.5%	0.114	0.064	0.087	96.5%	96.0%	91.5%
		RE_APR	0.2%	6.5%	22.9%	0.105	0.054	0.073	97.5%	96.0%	96.0%
		Syn1_srs	3.6%	2.7%	9.7%	0.123	0.076	0.105	94.5%	91.5%	91.0%
		Syn2_srs	3.5%	0.5%	4.6%	0.121	0.076	0.107	96.5%	96.0%	93.0%
		Syn1_lwt	1.8%	1.4%	2.8%	0.121	0.075	0.104	95.5%	93.0%	90.0%
		Syn2_lwt	2.2%	1.5%	2.1%	0.120	0.075	0.106	96.5%	96.0%	96.5%
$f_2 \propto h(S_i)$ **Actual samples BD:** $\beta_{1,Y_1\|Y_2} = 0.481$ $\beta_{1,Y_3\|Y_2} = 0.229$ $\beta_{1,Y_4\|Y_2} = 0.090$	Complete Data	Actual	-	-	-	0.108	0.066	0.088	95.0%	96.0%	95.0%
		Syn1BD	0.1%	0.6%	2.2%	0.109	0.068	0.089	95.0%	95.0%	93.0%
		Syn2BD	0.4%	2.9%	6.5%	0.109	0.069	0.090	95.0%	96.5%	96.0%
	MAR_X	SRS	12.8%	9.1%	52.0%	0.136	0.074	0.096	89.5%	90.0%	88.0%
		FX_APR	0.5%	0.6%	43.5%	0.114	0.069	0.079	93.5%	95.0%	97.0%
		RE_APR	0.8%	2.5%	19.0%	0.111	0.061	0.065	95.0%	95.5%	97.0%
		Syn1_srs	0.4%	0.7%	5.6%	0.126	0.082	0.097	94.0%	92.0%	91.5%
		Syn2_srs	0.0%	3.5%	2.7%	0.124	0.082	0.098	95.0%	94.0%	96.5%
	MAR_X,W	SRS	17.6%	12.4%	69.5%	0.141	0.069	0.101	86.0%	94.0%	83.0%
		FX_APR	0.4%	5.7%	42.2%	0.118	0.066	0.082	93.5%	95.5%	55.0%
		RE_APR	1.7%	3.1%	30.1%	0.111	0.054	0.073	97.5%	98.0%	97.5%
		Syn1_srs	6.7%	3.1%	23.0%	0.136	0.073	0.093	93.0%	94.0%	94.5%
		Syn2_srs	7.4%	0.4%	19.0%	0.136	0.075	0.095	96.0%	97.5%	97.0%
		Syn1_lwt	0.9%	0.9%	6.0%	0.130	0.075	0.092	93.0%	95.5%	93.5%
		Syn2_lwt	1.7%	2.6%	3.3%	0.126	0.076	0.094	97.0%	98.0%	97.5%

4.5 Application to NHANES III

We apply our method to the National Health and Nutrition Examination Survey (NHANES) III (1988-1994), which is designed to provide national estimates of health and nutritional status of the civilian noninstitutionalized population of the United States aged 2 months and older (NCHS 1994). The data is obtained from a stratified, multistage area probability sampling design with oversampling of certain age and ethnicity groups. For confidentiality and computational reasons, the public use data provides two pseudo PSUs per stratum. Another unique feature of NHANES is that data are collected through both interview and actual physical examinations of the sampled persons. Both unit- and item-level nonresponse occurs in both components of the survey, and there is a particularly high missing rate on the body mass index (BMI) measure for youth data in the physical examination component (30%). As a popular measure of overweight status and obesity, the percentiles of BMI for children and youths are of particular interest for public health reasons. The upper percentiles and the lower percentiles are also closely monitored for overweight and underweight status, respectively. As a result, we restrict our analysis sample to children and youths 2 months to 16 years of age.

We estimate population quantiles (from 0.05 to 0.95 with an increment of 0.05 along with two extreme percentiles: 0.03 and 0.97) of BMI for children and youths by gender. We also estimate the proportion of such a population being covered by health insurance, overall and by race. To assure congenial inference, we include the following variables that are either of primary interest in the substantive analysis, or are important predictors for BMI measures in the imputation model: age, gender, race, education, mother's BMI, father's BMI and family income (Yuan & Little, 2007a). We compared

three different methods in treatment of the missing data: 1) Complete case analysis (CC) with design-based estimation; 2) fully parametric model-based MI using design-based estimation, within which we apply both an imputation model assuming SRS and the appropriate model conditional on all three sample design features (i.e. dummy variables indicating cluster and stratum memberships as well the log transformation of sampling weights); and 3) our proposed finite population Bayesian bootstrap method (using SYN2_lwt, since we do not have separate weights for the first and second stages of sampling). Estimates of the median BMI and the proportion of children with health insurance are given in Table 4.4. The CC method appears to overestimate both the median of the BMI measure and health insurance coverage for full sample and race domains relative to the MI approaches, and yields the widest confidence intervals or largest standard errors as a result of decreased sample size. On the other hand, the median of BMI obtained from synthetic MI is quite similar to that from the model-based MI, while demonstrating some advantages in efficiency by yielding shorter intervals. The generally lower health insurance coverage estimates under the synthetic MI relative to model-based MI might be attributable to the fact that the synthetic MI are able to capture certain interactions between the sample design variables and the regular covariate matrix which are not explicitly modeled in the fully model-based MI.

Table 4.4 Alternative methods in estimating the median of BMI and the health insurance coverage rate, for full sample and by gender and race, respectively

Variable	Domain	Methods		
		CC	Model-based MI	Synthetic MI
BMI	Overall	17.2 [17.1, 17.4]	17.1 [16.9, 17.3]	17.0 [16.9, 17.2]
	Male	17.2 [16.9, 17.4]	17.0 [16.7, 17.2]	17.0 [16.8, 17.2]
	Female	17.3 [17.0, 17.7]	17.1 [16.8, 17.4]	17.1 [16.8, 17.3]
Health Insurance	Overall	0.785 (0.020)	0.778 (0.019)	0.761 (0.019)
	White	0.822 (0.018)	0.815 (0.017)	0.799 (0.016)
	Non-White	0.645 (0.036)	0.643 (0.033)	0.634 (0.036)

Figure 4.7 displays a visual comparison of the percentile estimation for the three methods under consideration. We look at how those methods perform in three different percentile ranges by gender domains: the middle percentiles from 0.5 to 0.75, the upper percentiles from 0.90 to 0.97 and the lower percentiles from 0.03 to 0.1. We chose these percentile ranges because: first, the extreme lower and upper percentiles of BMI are typically used to monitor under- and over-weight for children and youths; second, there is evidence that gender difference exists in these BMI percentile ranges (particularly when age is considered, i.e. growth patterns in BMI). In general, both MI methods result in very similar BMI estimates, and they are lower than those obtained from CC analysis. This makes sense, because by comparing the distributions of age for complete cases and for missing cases on the BMI measure, we found that younger children are more susceptible to missingness, and therefore CC analysis tends to overestimate BMI by excluding those younger missing cases. The inclusion of the age variable as a predictor in the imputation model corrects such an overestimation. The magnitude of this correction for boys is bigger than that for girls in estimating the lower percentiles (0.03, 0.05). By examining a report on BMI-for-age percentiles by gender released from CDC

(http://www.cdc.gov/nchs/data/series/sr_11/sr11_246.pdf), we find that baby boys (corresponding to the lower quantiles here) have relatively higher BMI, which might be the explanation. Therefore, it is convincing to believe that our synthetic MI method is comparable to the appropriate model-based MI in adjusting for potentially incorrect estimation under CC analysis.

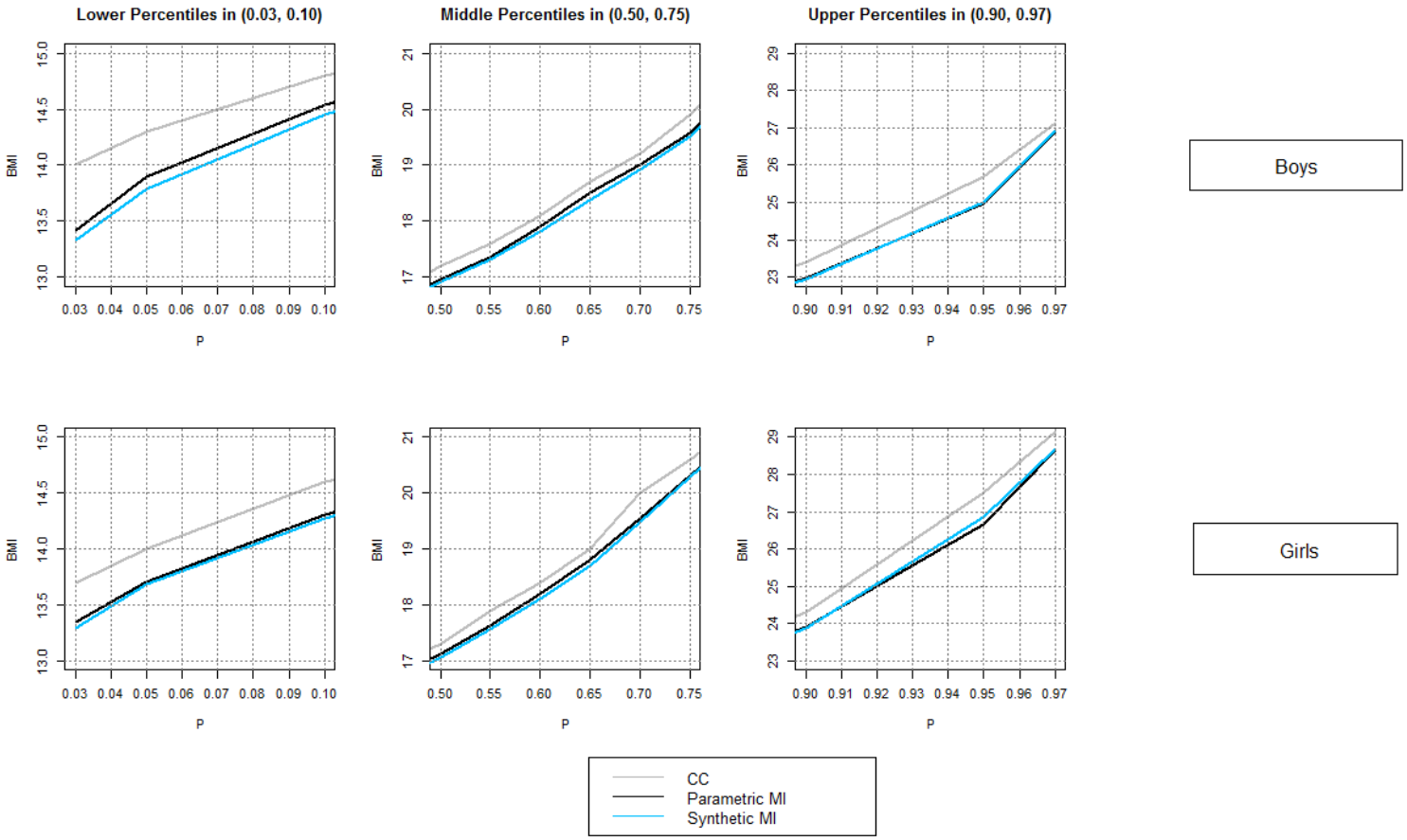

Figure 4.7 Comparison of methods for quantile estimation of BMI, by gender

4.6 Discussion

While multiple imputation has become a popular option for the analysis of missing data, open issues remain in its practical application with complex sample survey data. The complex features of sampling compounded with nonresponse in survey data often result in a rather complicated data structure, which prevents the standard MI techniques (such as a multivariate normal model assuming simple random sampling) from being applied straightforwardly. In this paper, we develop a general purpose approach to account for various design features in a highly stratified two-stage sample, using the two-step synthetic MI framework proposed by Zhou et al. (2013a). We have focused on evaluating the performance of the new method compared with existing methods, with respect to several missing data issues frequently encountered in large population-based socioeconomic and epidemiological studies. These include: i) accommodating stratification and multi-stage sampling in the imputation process; ii) the employment of nonstandard or non-normal imputation models for estimating probabilities of rare events; and iii) the estimation of population quantiles with multiply imputed data.

We demonstrate that the coverage properties of the proposed method are fairly good for nonsmooth statistics. Specifically, our stratified variations of the weighted Polya posterior exhibits robustness to the loss function for estimating upper and lower tails of the distribution function where even the appropriate model-based method (i.e. FX_APR) fails. In contrast with existing fully parametric MI methods, most of which perform poorly when applied to rare outcome binary data, the proposed method yields quite stable parameter estimates regardless of the rarity of the outcome.

It is worth stressing that our method requires only the simplest form of imputation model and combining rules for inference, because the effects of complex sample designs and the effect of estimating the nuisance parameter (e.g. regression coefficient in the regression imputation scheme is an example of a nuisance parameter, since our main parameters of interest are the population mean/quantiles of Y) in imputation are both correctly reflected in the replication variance estimation given the design-reversed and multiply imputed synthetic populations. The simplicity of the imputation model and the inference rules mean that any higher-level and nonlinear interactions in the covariate data, including those with the weights, clusters, or strata, will automatically be captured in the synthesizing step. However, when the imputation is conducted parametrically, as we do here, such design variable interactions will still need to be considered if they are associated with the missingness mechanism, although the impact of misspecification will generally be attenuated. Similarly, not-missing-at-random mechanisms that are dependent on the missing values are not accommodated in this framework.

Future research will investigate the inferential properties of the proposed method in situations where auxiliary information on all population units is available, using a constrained version of the Polya posterior. Two other possible research directions include: (i) extending the two-step synthetic MI framework to deal with unit nonresponse problems, and (ii) extending it to deal with generating synthetic data for disclosure risk limitation.

CHAPTER 5
CONCLUSION AND FUTURE RESEARCH

5.1 Contribution

Complex sample survey estimation and multiple imputation (MI) for survey nonresponse are two important areas in survey research. Coming up with a proper imputation model that is attentive to all complex sample design features can be difficult in many practical settings. In this thesis, we modified the classical MI framework of Rubin (1987) by dividing up the need to account for both sample designs and missing data into two separate steps. In the first step, we reversed the sampling mechanism by utilizing a *noninformative Bayesian approach* to finite population sampling. In the second step, we imputed the missing values in the created pseudo-population by constructing a *parametric Bayesian model* for the missing data under an IID assumption. Thus, the new framework stays within the Bayesian fundamentals underlying the standard MI theory. Yet it also generalizes naturally to increased *flexibility* for imputation with complex sample design surveys, because once a population with missing values is synthesized in a fashion that accommodates the sample design, a variety of imputation methods not limited to model-based MI can then be applied, for example, nonparametric hot deck imputation (Andridge & Little, 2010).

Chapters 2 to 4 develop different procedures but are methodologically interrelated within the general conceptual framework of the proposed two-step MI, in the sense that the weighted finite population Bayesian bootstrap (FPBB), or equivalently the weighted

Polya posterior, serves as the basic approach to the different procedures developed in each chapter.

Chapter 2 derives the theoretical formulation of the two-step MI from the traditional MI, filling a gap in the literature resulting from the fact that most current methods cannot adequately accommodate sampling weights. *Chapter 3* addresses the compound effects of clustering and unequal selection probability in the imputation process. It derives a posterior predictive distribution of the population that improves upon the "two-stage Polya posterior" of Meeden (1999). The posterior under the new procedure relaxes an inherent exchangeability assumption about the correlational structure among survey variables between the sampled clusters and nonsampled clusters not guaranteed by the population data generation mechanism, yet required by Meeden's "two-stage Polya posterior". It therefore allows for a wider range of population data structures to be considered. *Chapter 4* further extends the two-step MI approach into the area of stratified, clustered, and unequal probability-of-selection sample designs. It also explores the potential of the new method to deal with quantile estimation, and the estimation of proportions for binary rare events in the presence of item nonresponse.

In Chapters 3 and 4, we considered two variations of a two-stage procedure to create a synthetic population with missing data. One applies the weighted FPBB at each level of sampling and hence requires known design weights for units at all stages of sampling. The other employs replication methods at the PSU level and only applies the weighted FPBB to the ultimate sampling units using final weights (products of the design weights at each stage). Since the final weights are often the only design information available in public-use databases, the latter version has important practical value. The

simulation results demonstrated good repeated sampling properties of the proposed methodology in several sampling design settings, including PPS, two-stage clustering and highly stratified two-stage cluster samples. The proposed method also performed well with respect to estimation of means, regression parameters, and quantiles.

The proposed MI framework provides advantages over the existing fully parametric model-based MI in three ways:

(i) ***modeling***: the "untying" step of the proposed MI procedure recovers most of the information about the data generating mechanisms, including the data model, the sampling process and part of the response process. Thus a simple imputation model assuming IID suffices (in most cases) for the resulting pseudo population. This strikes a balance between *congeniality* and *sparsity* simultaneously required of the model, something difficult to attain with the fixed effects modeling strategy (as shown in the simulation results in chapter 4)

(ii) ***computation***: the proposed algorithms can be easily programmed in the R language (example codes for their applications on the BRFSS data and the NHANES data are provided in the Appendix). To be quite conservative for the proposed method to work properly, we purposefully generated a very large number of synthetic populations (i.e. L and S) for both the simulation study and the real data application. This required, for example, about 1.5 hours for imputing the missing BMI in the NHANES III data, with $L=50$ at the PSU level, $S=5$ at the element level and $M=5$ multiple imputations, on a 2.50 GHz Intel Core i5 laptop. Meanwhile, the proposed method does not need
complex numerical integration methods as is the case with random effects

modeling. Hence, they offer a feasible alternative to the random effects model which is well-known to have convergence issues with high dimensional categorical data;

(iii) ***analysis***: the weighted Polya sampling draws all variables jointly, and hence preserves the population-level multivariate relationships that can then be used for a variety of analyses (as evident by the real data applications). Since the relationships of the survey variables with the sample weights are also maintained in the draws, once we simulate sensible copies of the entire population, these analyses become fairly straightforward without appeal to design-based estimation (as is the case with quantile estimation).

5.2 Limitations

A full retrospective appreciation of the work done in this thesis also reveals several limitations of the proposed methods that suggest directions for further investigation. A major limitation lies in the fact that the simulation studies are restricted to one-stage element sampling or at most a stratified two-stage cluster sampling. Methods are not fully developed for *multistage designs* that involve more than two stages of sampling. For example, for the application on NHANES III, we considered the data as a two-stage sample, and assumed that lower-than-PSU-level sampling design effects could be neglected. In seeking for an improved solution, analysts should consider mapping multistage sampling designs to an "ultimate cluster design" (Kalton, 1979) before implementing the proposed procedure developed in this thesis.

Another limitation is the lack of development for the *small sample degrees of*

freedom (df) under the newly derived combining rules. As the method is a combination of multiple imputation and synthetic data generation, we attempted to draw theoretical support from both literatures (Raghunathan et al., 2003; Reiter, 2003, 2004; Reiter & Raghunathan, 2007; Rubin, 1987). However, despite the small sample *df* derived in Barnard and Rubin (1999) that is aimed at MI inference at the sample-level, the fraction of missing information (FMI) is not well-defined in the context of synthetic data in the current literature. Hence *df* for constructing interval estimation for the small sample t-approximation synthetic data cannot be constructed using FMI. In this regard, a next step is to break down the variance estimator under the new MI combining rules into components that may be used to define a sensible FMI.

The *efficiency loss* under the proposed method in a complex sample design setting is a further, though expected, limitation, resulting from the gains in robustness to model assumptions by using a computational nonparametric technique. We should note that, the proposed method is typically not as efficient (with higher RMSE) as its fully parametric counterparts with correctly specified imputation model. Improvement might be made through the use of auxiliary information, which will be discussed in the next section.

In addition, the validity of our method relies strongly on sample size. In chapter 4, we circumvented this problem by assuming a relatively large number of strata (H=50) in the sample. Hence, the synthetic procedure was still based on a reasonably large sample scenario, even though only 2 PSUs were selected within each stratum. The method will not work when both the number of strata and the number of clusters within strata are quite small and if single PSU strata exist in the sample. In such cases, we need to consider adaptation of the method by either collapsing strata or splitting PSUs. See Rubin

(1981) for a discussion of inappropriateness in using bootstrap and Bayesian bootstrap.

5.3 Future Research

The two-step MI framework developed in this thesis has meaningful extensions in several directions.

Throughout the thesis, I consider only the statistical adjustments for item-level missing data, assuming no unit nonresponse in the selected sample. A natural extension of the method would be to adapt the two-step MI framework to ***incorporate unit nonresponse***. Unit nonresponse is typically accounted for by weighting or weight adjustments, for which the sampling weight of non-respondents is spread over respondents using either weighting class adjustments or propensity modeling. When replication variance estimators (such as BRR and Jackknife) are used, it remains a practical question whether it is necessary to repeat every weighting step (including adjustments for unknown eligibility, unit nonresponse adjustments and post-stratification using auxiliary variables) separately for each replicate sample in order to produce consistent or approximately unbiased variance estimates (Valliant, 2004). Findings from previous research are mixed. Some evidence suggests the need to recalculate weights for each replicate (Lemeshow, 1979; Valliant, 1993); while others suggest that such extra efforts make little difference (Rust, 1987) or even lead to overestimation of the variance in small samples (Valliant, 2004). On all accounts, there is a need to develop an adapted version of the two-step MI method that ***propagates the uncertainty in unit nonresponse adjustments***, given the fact that a replication method and hence a replication-type variance estimator is used under the proposed MI framework. It would also be interesting

to compare the variance estimator under such an adapted two-step MI method with that resulting from an expedient method that does not propagate this source of variability. A feasible way to achieve this in a single-stage sampling design would follow a three-step procedure: 1) generate a Bayesian bootstrap (BB) sample from the parent sample S (including the unit response indicator variable, the base weight w_{base}, and the item-level missing data); 2) for each BB sample, apply regular unit nonresponse adjustments to the base weights for the respondent set S_R, using available auxiliary variables for the entire sample, and obtain the nonresponse (NR)-adjusted weight w_{NR}; and 3) for each NR-adjusted bootstrap sample, apply the weighted Polya posterior with w_{NR} as the input weight in the algorithm to create synthetic populations (with item-level missing data). The subsequent imputation method that assumes IID as well as the combining rules for inference would follow as described in chapter 2.

A second valuable extension would be to investigate ***domain/small area estimation***. Statistical agencies and survey organizations routinely produce estimates for subpopulations, called domains, to aid public and private sectors in effective policy making. Despite the growing demand, reliable domain estimates are difficult to obtain using standard frequentist approaches, primarily because domains are typically defined after a sample was selected. When cut across planned strata, they usually have very small associated sample size, leading to less precise estimates than those for the whole population. Complications of domain estimation and methods to tackle them are accounted in the complete data context to some degree (Cochran, 1977; Ghosh & Rao, 1994). When missing data are present, multiple imputation for domain estimation remains an issue, especially when the domain size is small and response rate is low (Kim,

Brick, Fuller, & Kalton, 2006; Meng, 1994; Seaman, White, Copas, & Li, 2011). Theoretically, we should include domain indicators Z as well as their interactions with the survey weight W in the imputation scheme for the MI variance estimator to be approximately unbiased. In practice, however, many domain estimates of interest in the substantive analysis will not be identified at the time that the imputation was carried out. Even if they do, to model the interaction term $W*Z$ for all possible domains is not feasible. In this case, the proposed two-step MI method is a logical choice because the association of the weights with survey variables in the population data can be resolved by generating synthetic populations. Once the numerous complete copies of the population are created and imputed, the simulated population-level estimates can be simply combined. This differs from the standard MI method where inference is based on explicit design-based domain estimators that may be lacking for certain estimands, e.g. the difference or ratio estimator of two domain medians.

Another related extension of the proposed MI procedure would be to incorporate ***auxiliary information*** at the first step of synthetic data generation. Auxiliary information is routinely used in finite population sampling to facilitate inference. For example, ratio and regression estimators are used when the population mean of an auxiliary variable is known a priori. The procedures developed in this thesis all assumed an absence of prior auxiliary information. Our future work will focus on adapting the current procedures by capitalizing on such information to improve the efficiency of imputation and inference as a whole. Lazar, Meeden, and Nelson (2008) showed that the Polya posterior can be easily adapted to incorporate different levels of prior information. In a similar fashion, I can restrict the weighted-FPBB to create pseudo-populations that satisfy certain constraints

specified by the prior information, such as known population means or ranges about the auxiliary variable(s). This would in effect extend the work of Sangeneh et al. (2011) to the missing data context. On the other hand, since the essence of all small area estimation methods lies in the use of available auxiliary variables (Rao, 2003), I believe a constrained version of the weighted-FPBB would also provide a nonparametric Bayesian solution for small area estimation in both complete data and missing data settings. Yet another utilization of such auxiliary information lies in the aforementioned nonresponse adjustments in each BB sample.

We considered a special type of missing data mechanism which depends on the selection probability (or sample weights). On the one hand, the design-inversing step of the new method ensures a correct estimate of the population distribution in the presence of missing data and thus reduces the impact of misspecified missingness mechanisms by avoiding enhancement from misspecified data generation mechanisms. On the other hand, elimination of the weights from the self-weighting FPBB population does not obviate the need to account for the weights in the imputation process to attain valid inference. Future extensions of this work could use other imputation methods to account for such a weight-dependent MAR missingness, for example, the weighted hot deck by Andridge and Little (2009). In some cases, the missingness may also depend on the unobserved random clustering effects and thus becomes nonignorable. We do not consider such scenarios here, since our method is proposed to address item nonresponse under the MAR assumption. But the method may be adapted to deal with ***nonignorable missing data*** problem under a two-stage cluster sampling design, by combining with the work of Yuan and Little (2007a, 2007b, 2008), Andridge and Little (2011) and West and Little (2013).

APPENDIX

1. R code for using the proposed two-step MI method on BRFSS

```
require(mice)
require(survey)
set.seed(seed #)

#########################################################
#dt: sample data set for analysis, with recoded variables;
#N: synthetic population size;
#Bt1: number of BB samples created for stage 1;
#Bt2: number of weighted FPBB populations created for stage 2;
#Mt: number of multiple imputations;
#ps.n: sample size of parent sample;
#########################################################

SynMI <- function(dt, N, Bt1, Bt2, Mt, ps.n) {
##Step 1: synthesize populations;
#Stage 1: draw bootstrap samples from the parent sample;
      #Normalize the final weights to sum up to the synthetic population size;
      dt[,"X_FINALWT"] <- dt[,"X_FINALWT"]*N/sum(dt[,"X_FINALWT"])
      dsgn <- svydesign(ids = ~1, strata = NULL, nest = FALSE, data = dt, weights =
~ X_FINALWT)
      dsgn.r <- as.svrepdesign(design = dsgn, type = "subbootstrap", replicates = Bt1)
      repwt <- as.matrix(dsgn.r$repweights)
      repwt[repwt==0] <- NA

      #Set up a data frame to store results from loglinear analysis;
      logcf <- matrix(0,20,Bt1)

#Stage 2: within each bootstrap sample, draw weighted FPBB synthetic populations;
      for (j in 1:Bt1){
            st.bb <- cbind(dt,repwt[,j])
            #Delete those units with zero replicate weights for each bootstrap sample;
            st.BB <- na.omit(st.bb)
            #Recode those 9999 back to NA for imputation;
            st.BB[st.BB==9999] <- NA
            st.rp <- st.BB[rep(1:nrow(st.BB),round(st.BB$repwt)),]
            ns <- ps.n-1
            Samwts <- rep((ps.n/ns)*st.BB$X_FINALWT,round(st.BB$repwt))
```

```
        lgcf <- matrix(0,20,Bt2)
        #Write the algorithm for creating weighted FPBB populations based on Cohen;
        for(boot in 1: Bt2){
                l <- rep(0,ns)
                        for(k in 1:(N-ns)){
                        l <- l+rmultinom(1,1,((Samwts-1)+l*((N-ns)/ns))/((N-ns)+(k-
                        1)*((N-ns)/ns)))
                        }
        income <- as.factor(c(rep(st.rp[,6],l),st.rp[,6]))
        bmi <- as.numeric(c(rep(st.rp[,10],l),st.rp[,10]))
        bphigh <- as.factor(c(rep(st.rp[,9],l),st.rp[,9]))
        hlthplan <- as.factor(c(rep(st.rp[,1],l),st.rp[,1]))
        age <- as.numeric(c(rep(st.rp[,2],l),st.rp[,2]))
        racew <- as.factor(c(rep(st.rp[,5],l),st.rp[,5]))
        educa <- as.factor(c(rep(st.rp[,8],l),st.rp[,8]))
        employ <- as.factor(c(rep(st.rp[,7],l),st.rp[,7]))
        gender <- as.factor(c(rep(st.rp[,3],l),st.rp[,3]))
        lgwt <- as.numeric(log(c(rep(st.rp[,4],l),st.rp[,4])))

##Step 2: Multiple imputation of synthetized populations;
    #Imputation model ignores the design;
      temp1 <- data.frame(cbind(income, bmi, bphigh, hlthplan, age, racew, educa,
      employ, gender))
    #Imputation model includes log of final weights;
      temp2 <- data.frame(cbind(income, bmi, bphigh, hlthplan, age, racew, educa,
      employ, gender, lgwt))
    # Impute for each synthetic population;
      temp_impx <- mice(temp1)
      temp_impxw <- mice(temp2)
      mlx <- data.frame()
      mlxw <- data.frame()

      for (u in 1:Mt){
            mult <- as.vector(rep(u,nrow(temp1)))
            mlx <- rbind(mlx,cbind(complete(temp_impx, u),mult))
            mlxw <- rbind(mlxw,cbind(complete(temp_impxw, u),mult))
      }

      lcf <- matrix(0,20,Mt)

      ##Analysis of the imputed synthetic populations;

            for (v in 1:Mt){
                  for (mult in v:v){
                        stx <- mlx[mlx["mult"]==v,]
                              stxw <- mlxw[mlxw["mult"]==v,]
```

```
                    dsgnx <- svydesign(id=~1, strata= NULL, weights= NULL,
                    data=stx)
                    dsgnxw <- svydesign(id=~1, strata= NULL, weights= NULL,
                    data=stxw)

                    #loglinear analysis on the imputed data;
                    ax <- svyloglin(~income+hlthplan+racew+gender,dsgnx)
                    bx <- update(ax,~.^2)
                    axw <- svyloglin(~income+hlthplan+racew+gender,dsgnxw)
                    bxw <- update(axw,~.^2)
                    lcf[,v] <- c(coef(bx),coef(bxw))
                }
            }

            #Calculate the synthetic MI point and variance estimates;
            lgcf[ ,boot] <- apply(lcf,1,mean)
            print(boot)
        }
        logcf[ ,j] <- apply(lgcf,1,mean)
        print(j)
        }
smpm_log <- apply(logcf,1,mean)
smpv_log <- (1+1/Bt1)*apply(logcf,1,var)
stat <- cbind(smpm_log,smpv_log)
write.table(stat,file="D:\\Dissertation\\paper1\\brfss_syn_loglin.csv",row.names=FALSE,
sep=",")
}

############;
##Example run;
dt<-read.csv("D:\\Dissertation\\ paper1\\brfss09.csv")
#Need to collapse the medium and high income categories to avoid sparse cells;
dt$INCOME[dt$INCOME==3]<-2
#missing data were coded as '9999' in data set brfss09.csv;
SynMI(dt=dt, N=4500, Bt1=100, Bt2=30, Mt=5, ps.n=388)
```

2. R code for using the proposed two-step MI method on NHANES III

```
require(survey)
require(mice)
require(polyapost)
set.seed(seed #)

syn_bmi<-function(dt, N, Bt1, Bt2, Mt){

##Step 1: Generate synthetic populations with missing data;
#Stage 1: Create bootstrap samples from the parent sample;
        dsgn <- svydesign(ids = ~predcl, strata = ~pstrat, nest = TRUE, data = dat,
        weights = ~predwt)
        dsgn.RW <- as.svrepdesign(design = dsgn, type = "subbootstrap", replicates = Bt1)
        dim(dsgn.RW$repweights)
        repwt<-as.matrix(dsgn.RW$repweights)
        repwt[repwt==0]<-NA
        dim(repwt)

        #set up arrays to hold point estimates from bootstrap samples;
        btm<-matrix(0,nrow=Bt1,ncol=3)
        btqt<-matrix(0,nrow=Bt1,ncol=21)
        btqtm<-matrix(0,nrow=Bt1,ncol=21)
        btqtf<-matrix(0,nrow=Bt1,ncol=21)

        for (j in 1:Bt1){
                st.bb<-cbind(dat,repwt[,j])
                #delete those units with zero weights for each bootstrap sample;
                st.BB<-na.omit(st.bb)
                #recode those 999 back to NA so that the mice package can be used for
                imputation;
                st.BB$pybmi[st.BB$pybmi==999]<-NA

                #need to calculate the replicate weights;
                Samwt<-st.BB[,9]*st.BB[,13]
                #normalize again the adjusted weights;
                Samwts<-Samwt*N/sum(Samwt)
                np<-nrow(st.BB)
                ids<-seq(np)
                ns<-N-np

##Stage 2: Create unweighted synthetic populations within each bootstrap sample;
#Set up arrays to hold point estimates from imputed unweighted synthetic populations;
        fbm<-matrix(0,nrow=Bt2,ncol=3)
        fbqt<-matrix(0,nrow=Bt2,ncol=21)
        fbqtm<-matrix(0,nrow=Bt2,ncol=21)
```

```
fbqtf<-matrix(0,nrow=Bt2,ncol=21)

for(boott in 1:Bt2){
        l<-vector()
        smp<-wtpolyap(ids, Samwts, ns)
        #input the adjusted weights in the weighted Polya sampling algorithm;
        for (k in 1:np){
        l<-c(l,length(smp[smp==k]))
        }
    #check if the vector of l sum up to the number of synthetic population size;
    sum(l);

    predY1<-c(rep(st.BB[,1],l)) #bmi
    predY2<-c(rep(st.BB[,2],l)) #race
    predY3<-c(rep(st.BB[,3],l)) #gender
    predY4<-c(rep(st.BB[,4],l)) #income
    predY5<-c(rep(st.BB[,5],l)) #education
    predY6<-c(rep(st.BB[,6],l)) #mother's bmi
    predY7<-c(rep(st.BB[,7],l)) #father's bmi
    predY8<-c(rep(st.BB[,8],l)) #age
    predwt1<-c(rep(st.BB[,9],l))
    predlwt<-log(predwt1) #log of sample weight
    predCID<-c(rep(st.BB[,12],l)) #cluster ID
    predSTID<-c(rep(st.BB[,11],l)) #stratum ID

##Step 2: Multiple imputation of the unweighted synthetic populations;

    #use the imputation model including log of weight as a predictor (syn_lwt);
    temp1<-data.frame(cbind(predY1, predY2, predY3, predY4, predY5, predY6,
    predY7, predY8, predlwt))
    temp1_imp<-mice(temp1,method="norm", m=Mt)
    ml<-complete(temp1_imp, 'long')
    ml$bmit<-exp(ml$predY1) #back transform bmi to its normal scale
    mlmale<-subset(ml, predY3==1)
    mlfem<-subset(ml, predY3==2)
    multm<-cbind(as.vector(by(ml$bmit,ml$.imp,mean)),
    as.vector(by(mlmale$bmit,mlmale$.imp,mean)),
    as.vector(by(mlfem$bmit,mlfem$.imp,mean)))
    multqt<-sapply(with(ml,by(ml,.imp,function(x)quantile(x$bmit,
    c(0.03,seq(0.05,0.95,0.05),0.97)))),as.vector)
    multqtm<-sapply(with(mlmale,by(mlmale,.imp,function(x)quantile(x$bmit,
    c(0.03,seq(0.05,0.95,0.05),0.97)))),as.vector)
    multqtf<-sapply(with(mlfem,by(mlfem,.imp,function(x)quantile(x$bmit,
    c(0.03,seq(0.05,0.95,0.05),0.97)))),as.vector)

        fbm[boott,]<-t(apply(multm,2,mean))
```

```
            fbqt[boott,]<-t(apply(multqt,1,mean))
            fbqtm[boott,]<-t(apply(multqtm,1,mean))
            fbqtf[boott,]<-t(apply(multqtf,1,mean))
            print(boott)
            }

    btm[j,]<-t(apply(fbm,2,mean))
    btqt[j,]<-t(apply(fbqt,2,mean))
    btqtm[j,]<-t(apply(fbqtm,2,mean))
    btqtf[j,]<-t(apply(fbqtf,2,mean))
    print(j)
    }

    smpm<-apply(btm,2,mean)
    smpv<-(1+1/Bt1)*apply(btm,2,var)
    smpse<-sqrt(smpv)
    smpqt<-apply(btqt,2,mean)
    smpqtv<-(1+1/Bt1)*apply(btqt,2,var)
    smpqtse<-sqrt(smpqtv)
    smpqtm<- apply(btqtm,2,mean)
    smpqtvm<-(1+1/Bt1)*apply(btqtm,2,var)
    smpqtsem<-sqrt(smpqtvm)
    smpqtf<-apply(btqtf,2,mean)
    smpqtvf<-(1+1/Bt1)*apply(btqtf,2,var)
    smpqtsef<-sqrt(smpqtvf)

tt<-cbind(smpqt,smpqtm,smpqtf,smpqtse,smpqtsem,smpqtsef)
ss<-cbind(smpm,smpse)
write.table(tt,file="D:/Dissertation/paper3/nhanes/synbmiqt_lwt.csv",row.names=FALSE
,sep=",")
write.table(ss,file="D:/Dissertation/paper3/nhanes/synbmimn_lwt.csv",row.names=FALS
E,sep=",")
}

#############;
##Example run;
syn_bmi(dt=dt, N=100000, Bt1=50, Bt2=5, Mt=5)
dt<-read.csv("D:/Dissertation/paper3/nhanes/synbmi.csv")
#Set the synthetic population size about 10 times the sample size;
N<-100000
#Normalize the weights to sum up to the assumed synthetic population size;
dt[,"predwt"]<-dt[,"predwt"]*N/sum(dt[,"predwt"])
sum(dt$predwt)
#Recode the missing values to 999;
dat[is.na(dat)]<-999
```

BIBLIOGRAPHY

Aitkin, M. (2008). Application of the Bayesian Bootstrap in Finite Population Inference. *J Off Stat, 24*(1), 21-51.

Anderson, D., & Aitkin, M. (1985). Variance Component Models with Binary Response: Interviewer Variability. *Journal of the Royal Statistical Society. Series B, 47*(2), 203-210.

Andridge, R. R. (2011). Quantifying the Impact of Fixed Effects Modeling of Clusters in Multiple Imputation for Cluster Randomized Trials. *Biometrical Journal 53*(1), 57-74.

Andridge, R. R., & Little, R. J. (2009). The Use of Sample Weights in Hot Deck Imputation. *Journal of Official Statistics, 25*(1), 21-36.

Andridge, R. R., & Little, R. J. (2010). A Review of Hot Deck Imputation for Survey Non-response. *Int Stat Rev, 78*(1), 40-64. doi: 10.1111/j.1751-5823.2010.00103.x

Andridge, R. R., & Little, R. J. (2011). Proxy Pattern-Mixture Analysis for Survey Nonresponse. *Journal of Official Statistic, 27*(2), 153-180.

Barnard, D., & Rubin, D. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika, 86*, 948-955.

Basu, D. (1971). An Essay on the Logical Foundation of Survey Sampling, Part I. *Foundations of Statistical Inference, Toronto: Holt, Rinehart & Winston*, 203-242.

Binder, D. A. (1993). On the Variances of Asymptotically Normal Estimators from Complex Surveys. *International Statistical Review, 51*, 279-292.

Breidt, F. J., Claeskens, G., & Opsomer, J. D. (2005). Model-assisted estimation for complex surveys using penalised splines. *Biometrika, 92*, 831-846.

Carpenter, J. R. (2012). Using survey weights with multiple imputation — a multilevel approach. *http://www.lse.ac.uk/statistics/events/SpecialEventsandConferences/CarpenterJR.pdf*.

Carpenter, J. R., Goldstein, H., & Kenward, M. G. (2011). REALCOM-IMPUTE Software for Multilevel Multiple Imputation with Mixed Response Types. *J Stat Softw, 45*(5), 1-14.

Chen, C. An Introduction to Quantile Regression and the QUANTREG Procedure. *SAS®Users Group International Paper Presentation Guidelines, Cary, NC: SAS Institute Inc.*

Chen, Q., Elliott, M. R., & Little, R. J. (2010). Bayesian penalized spline model-based inference for finite population proportions in unequal probability sampling. *Survey Methodology, 36*, 22-34.

Cochran, W. G. (1977). Sampling Techniques (Third Edition), New York: J Wiley & Sons, New York.

Cohen, M. P. (1997). The Bayesian Bootstrap and Multiple Imputation for Unequal Probability Sample Designs. *ASA Proceedings of the Section on Survey Research Methods*, 635-638.

Dong, Q., Elliott, M. R., & Raghunathan, T. E. (2014). A Nonparametric Method to Generate Synthetic Populations to Adjust for Complex Sample Designs. *Survey Methodology (accepted)*.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics, 7*,

1-26.
Elliott, M. R. (2007). Bayesian Weight Trimming for Generalized Linear Regression Models. *Survey Methodology, 33*(1), 23-34.
Elliott, M. R. (2008). Model Averaging Methods for Weight Trimming. *J Off Stat, 24*(517-540).
Elliott, M. R. (2009). Model Averaging Methods for Weight Trimming in Generalized Linear Regression Models. *J Off Stat, 25*, 1-20.
Elliott, M. R., & Little, R. J. (2000). Model-based alternatives to trimming survey weights. *J Off Stat, 16*, 191-209.
Elliott, M. R., Reslera, A., Flannagan, C. A., & Rupp, J. D. (2010). Appropriate analysis of CIREN data: Using NASS-CDS to reduce bias in estimation of injury risk factors in passenger vehicle crashes. *Accident Analysis and Prevention, 42*, 530-539.
Ericson, W. A. (1969). Subjective Bayesian Models in Sampling Finite Populations. *Journal of the Royal Statistical Society. Series B,, 31*(2), 195-233.
Feller, W. (1968). An Introduction to Probability Theory and Its Applications, volume I. Wiley.
Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics, 1*, 209-230.
Francisco, C. A., & Fuller, W. A. (1991). Quantile Estimation with a Complex Survey Design. *Annals of Statististics, 19*, 454-469.
Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science, 22*(2), 153-164.
Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). Bayesian Data Analysis, Chapman & Hall/CRC.
Gelman, A., & Rubin, D. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science, 7*(4), 457-472.
Ghosh, M., & Meeden, G. (1997). Bayesian Methods for Finite Population Sampling. Chapman & Hall.
Ghosh, M., & Rao, J. N. K. (1994). Small Area Estimation: An Appraisal. *Statistical Science, 9*(1), 55-76.
Girard, C. (2009). The Rao-Wu Rescaling Bootstrap: From theory to practice *Proceedings of FCSM*.
Gross, S. (1980). Median estimation in sample surveys. *ASA Proceedings of the Section on Survey Research Methods*.
Hansen, M., Hurwitz, W., & Madow, W. (1953). Sample Survey Methods and Theory. New York: John Wiley & Sons, Inc.
Hansen, M. H., Madow, W. G., & Tepping, B. J. (1983). An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys. *Journal of the American Statistical Association, 78*(384), 776-793.
Heitjan, D. F., & Little, R. J. A. (1991). Multiple Imputation for the Fatal Accident Reporting System. *Applied Statistics, 40*, 13-29.
Hinkins, S., Oh, H. L., & Scheuren, F. (1997). Inverse Sampling Design Algorithms. *Survey Methodology, 23*(1), 11-21.
Horton, N., Lipsitz, S., & Parzen, M. (2003). A Potential for Bias When Rounding in Multiple Imputation. *Am Stat, 57*(4), 229-232.

Hsuan, F. (1979). A Stepwise Bayes Procedure. *Annals of Statistics, 7*(860-868).
Kalton, G. (1979). Ultimate cluster sampling. *Journal of the Royal Statistical Society, Series A, 142*(2), 210-222.
Kim, J. K., Brick, J. M., Fuller, W. A., & Kalton, G. (2006). On the bias of the multiple-imputation variance estimator in survey sampling. *Journal of the Royal Statistical Society, Series B, 68*, 509-521.
King, G., & Zeng, L. (2001). Logistic Regression in Rare Events Data. *Political Analysis, 9*, 137-163.
Kish, L. (1965). Survey Sampling. New York: Wiley.
Korn, E. L., & Graubard, B. I. (1999). Analysis of Health Surveys. New York: Wiley. .
Kovar, J. G., Rao, J. N. K., & Wu, C. F. J. (1988). Bootstrap and Other Methods to Measure Errors in Survey Estimates. *Canadian Journal of Statistics, 16*, 25-45.
Lazar, R., Meeden, G., & Nelson, D. (2008). A noninformative Bayesian approach to finite population sampling using auxiliary variables. *Survey Methodology, 34*(1), 51-64.
Lemeshow, S. (1979). The Use of Unique Statistical Weights for Estimating Variances with the Balanced Half-Sample Technique. *Journal of Statistical planning and Inference, 3*, 315-323.
Li, B., Lingsma, H. F., Steyerberg, E. W., & Lesaffre, E. (2011). Logistic random effects regression models: a comparison of statistical packages for binary and ordinal outcomes. *BMC Med Res Methodol, 11:77*.
Little, R. J. (1982). Models for non-response in sample surveys. *Journal of the American Statistical Association, 77*, 237-249.
Little, R. J. (2004). To Model or Not to Model? Competing Modes of Inference for Finite Population Sampling. *Journal of the American Statistical Association, 99*, 546-556.
Little, R. J. (2006). Calibrated Bayes: A Bayes/Frequentist Roadmap. *Am Stat, 60*(3), 213-223.
Little, R. J. (2011). Calibrated Bayes, for Statistics in General, and Missing Data in Particular (with Discussion and Rejoinder). *Statistical Science, 26*(2), 162-186.
Little, R. J., & Rubin, D. B. (2002). Statistical Analysis with Missing Data (Second Edition), New York: J Wiley & Sons, New York.
Little, R. J., & Vartivarian, S. (2005). Does Weighting for Nonresponse Increase the Variance of Survey Means? *Survey Methodology, 31*, 161-168.
Little, R. J., & Zheng, H. (2007). The Bayesian Approach to the Analysis of Finite Population Surveys. *Bayesian Statistics, 8*, 283-302.
Lo, A. Y. (1986). Bayesian statistical inference for sampling a finite population. *The Annals of Statistics, 14*, 1226-1233.
Lo, A. Y. (1988). A Bayesian Bootstrap for a Finite Population. *The Annals of Statistics, 16*(4), 1684-1695.
Lumley, T. (2004). Analysis of complex survey samples. *J Stat Softw, 9*(1), 1-19.
McCarthy, P. J., & Snowden, C. B. (1985). The Bootstrap and Finite Population Sampling. *Vital and Health Statistics, 2–95. Public Health Service Publication 85-1369, U. S. Government Printing Office, Washington.*
Meeden, G. (1999). A Non-Informative Bayesian Approach for Two-Stage Cluster Sampling *Sankhya. Series B, 61*, 133-144.

Meeden, G., & Lazar, R. (2012). polyapost: Simulating from the Polya posterior. R package version 1.1-2. http://CRAN.R-project.org/package=polyapost.
Meeden, G., & Vardeman, S. (1991). A Noninformative Bayesian Approach to Interval Estimation in Finite Population Sampling. *Journal of the American Statistical Association, 86*, 972-980.
Meng, X. L. (1994). Multiple Imputation Inferences with Uncongenial Sources of Imput. *Statistical Science, 9*(4), 538-558.
Murphy, P. (2008). An Overview of Primary Sampling Units (PSUs) in Multi-Stage Samples for Demographic Surveys. *ASA Proceedings of the Section on Survey Research Methods*, 2856-2863.
Nelson, D., & Meeden, G. (2006). Noninformative Nonparametric Finite Population Quantile Estimation. *Journal of Statistical planning and Inference, 136*, 53-67.
Olkin, I., & Tate, R. F. (1961). Multivariate Correlation Models with Mixed Discrete and Continuous Variables. *Annals of Mathematical Statistics, 32*, 448-465.
Pfeffermann, D. (1993). The Role of Sampling Weights When Modeling Survey Data. *International Statistical Review, 61*(2), 317-337.
Pfeffermann, D. (2011). Modelling of complex survey data: Why model? Why is it a problem? How can we approach it? *Survey Methodology, 37*(2), 115-136.
Pfeffermann, D., & Sverchkov, M. Y. (2003). Fitting Generalized Linear Models under Informative Sampling. *In Analysis of Survey Data. (R.L. Chambers and C.J. Skinner). New York: Wiley.*, 174-195.
Pinheiro, J. C., & Bates, D. M. (1995). Approximations to the Log-likelihood Function in the Nonlinear Mixed-effects Model. *Journal of Computational and Graphical Statistics, 4*, 12-35.
R Core Team. (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.
Rabe-Hesketh, S., & Skrondal, A. (2006). Multilevel Modeling of Complex Survey Data. *Journal of the Royal Statistical Society, Series A, 169*(4), 805-827.
Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology, 27*(1), 85-95.
Raghunathan, T. E., Reiter, J. P., & Rubin, D. B. (2003). Multiple Imputation for Statistical Disclosure Limitation. *J Off Stat, 19*(1), 1-16.
Rao, J. N. K. (2003). Small Area Estimation. Wiley, New York.
Rao, J. N. K., & Wu, C. F. J. (1988). Resampling Inference with Complex Survey Data. *Journal of the American Statistical Association, 83*, 231-241.
Rao, J. N. K., Wu, C. F. J., & Yue, K. (1992). Some Recent Work on Resampling Methods for Complex Surveys. *Survey Methodology, 18*, 209-217.
Reiter, J. P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology, 29*, 181-188.
Reiter, J. P. (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodology, 30*(2), 235-242.
Reiter, J. P., & Raghunathan, T. E. (2007). The Multiple Adaptations of Multiple Imputation. *Journal of the American Statistical Association, 102*, 1462-1471.
Reiter, J. P., Raghunathan, T. E., & Kinney, S. K. (2006). The Importance of Modeling

the Sampling Design in Multiple Imputation for Missing Data. *Survey Methodology, 32*(2), 143-149.
Royall, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika, 57*, 377-387.
Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*, 581-592.
Rubin, D. B. (1981). The Bayesian bootstrap. *Annals of Statistics, 9*, 130-134.
Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: Wiley.
Rubin, D. B. (1996). Multiple Imputation After 18+ Years. *Journal of the American Statistical Association, 91*(434), 473-489.
Rubin, D. B., & Schenker, N. (1986). Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse. *Journal of the American Statistical Association, 81*(394), 366-374.
Rust, K. (1987). Practical Problems in Sampling Error Estimation. Bulletin of the International Statistical Institute. Invited paper. 10.3, 1-18.
Rust, K., & Rao, J. N. K. (1996). Variance Estimation for Complex Estimators in Sample Surveys. *Statistics in Medical Research, 5*, 381-397.
Sarndal, C., Swensson, B., & Wretman, J. (1992). Model Assisted Survey Sampling. Springer-Verlag.
Schafer, J. L. (1997a). Analysis of Incomplete Multivariate Data. Chapman & Hall. London.
Schafer, J. L. (1997b). Imputation of missing covariates under a multivariate linear mixed model. *Technical report 97-04, Dept. of Statistics, The Pennsylvania State University, http://www.stat.psu.edu/reports/1997/tr9704.pdf*.
Schafer, J. L. (1999). Multiple Imputation: A Primer. *Stat Methods Med Res, 8*, 3-15.
Schafer, J. L., & Yucel, R. M. (2002). Computational Strategies for Multivariate Linear Mixed-Effects Models with Missing Values. *Journal of Computational and Graphical Statistics, 11*(2), 421-442.
Schenker, N., Raghunathan, T. E., Chiu, P., Makuc, D. M., Zhang, G., & Cohen, A. J. (2006). Multiple Imputation of Missing Income Data in the National Health Interview Survey. *Journal of the American Statistical Association, 101*(475), 924-933.
Seaman, S. R., White, I. R., Copas, A. J., & Li, L. (2011). Combining Multiple Imputation and Inverse-Probability Weighting. *Biometrics, 68*, 129-137.
Shao, J., & Chen, Y. (1998). Bootstrapping Sample Quantiles Based on Complex Survey Data under Hot Deck Imputation. *Statistica Sinica, 8*, 1071-1085.
Shao, J., & Wu, C. F. J. (1992). Asymptotic Properties of the Balanced Repeated Replication Method for Sample Quantiles. *Annals of Statistics, 20*, 1571-1593.
Stiratelli, R., Laird, N., & Ware, J. (1984). Random-effects Models for Serial Observations with Binary Response. *Biometrics, 40*, 961-971.
Sugden, R., & Smith, T. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika, 71*(3), 495-506.
Tanner, M. A., & Wong, W. H. (1987). The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association, 82*, 528-540.
Valliant, R. (1993). Poststratification and Conditional Variance Estimation. *Journal of the American Statistical Association, 88*, 89-96.
Valliant, R. (2004). The Effect of Multiple Weighting Steps on Variance Estimation.

Journal of Official Statistic, 20(1), 1-18.
Valliant, R., Dorfman, A., & Royall, R. M. (2000). Finite Population Sampling and Inference: A Prediction Approach. New York: John Wiley. .
van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *J Stat Softw, 45*(3), 1-67. URL http://www.jstatsoft.org/v45/i03/.
West, B. T., & Galecki, A. T. (2011). An Overview of Current Software Procedures for Fitting Linear Mixed Models. *Am Stat, 65*(4), 274-282.
West, B. T., & Little, R. J. (2013). Nonresponse Adjustment of Survey Estimates Based on Auxiliary Variables Subject to Error. *Journal of the Royal Statistical Society, Series C (Applied Statistics), 62* (2), 213-231.
Wolter, K. M. (2007). Introduction to Variance Estimation. New York: Springer-Verlag.
Woodruff, R. (1952). Interval for Medians and Other Position Measures. *Journal of the American Statistical Association, 47*(260), 635-646.
Yuan, Y., & Little, R. J. (2007a). Parametric and Semiparametric Model-Based Estimates of the Finite Population Mean for Two-Stage Cluster Samples with Item Nonresponse. *Biometrics, 63*, 1172-1180.
Yuan, Y., & Little, R. J. (2007b). Model-Based Estimates of the Finite Population Mean for Two-Stage Cluster Samples with Unit Non-Response. *Appl. Statist., 56*(1), 79-97.
Yucel, R. M. (2011). Random covariances and mixed-effects models for imputing multivariate multilevel continuous data. *Statistical Modeling, 11*, 351-370.
Yucel, R. M., & Demirtas, H. (2010). Impact of Non-Normal Random Effects on Inference by Multiple Imputation: A Simulation Assessment. *Comput Stat Data Anal, 54*(3), 790-801.
Yucel, R. M., & Raghunathan, T. E. (2006). Sequential Hierarchical Regression Imputation (shrimp). *ASA Proceedings of the Health Policy Statistics Section. American Statistical Association, Alexandria, VA.*
Zangeneh, S. Z., Keener, R. W., & Little, R. J. (2011). Bayesian nonparametric estimation of finite population quantities in absence of design information on nonsampled units. *ASA Proceedings of the Joint Statistical Meetings*.
Zhao, E., & Yucel, R. M. (2009). Performance of Sequential Imputation Method in Multilevel Applications. *ASA Proceedings of the Section on Survey Research Methods*, 2800-2810.
Zhao, J., & Schafer, J. L. (2013). pan: Multiple imputation for multivariate panel or clustered dataR package version 0.9.
Zheng, H., & Little, R. J. (2003). Penalized spline model-based estimation of the finite population total from probability-proportional-to-size samples. *Journal of Official Statistic, 19*, 99-117.
Zheng, H., & Little, R. J. (2004). Penalized Spline Nonparametric Mixed Models for Inference about a Finite Population Mean from Two-Stage Samples. *Survey Methodology, 30*(2), 209-218.
Zheng, H., & Little, R. J. (2005). Inference for the Population Total from Probability-Proportional-to-Size Samples Based on Predictions from a Penalized Spline Nonparametric Model. *J Off Stat, 21*, 1-20.
Zhou, H., Elliott, M. R., & Raghunathan, T. E. (2013a). A Two-Step Semiparametric

Method to Accommodate Sampling Weights in Multiple Imputation. *submitted.*

Zhou, H., Elliott, M. R., & Raghunathan, T. E. (2013b). Multiple Imputation in Two-Stage Cluster Samples Using Finite Population Bayesian Bootstrap. *In Preparation.*