Adaptive Design to Adjust for Unit Nonresponse Using an External Micro-level Benchmark

by

Julia Shin-Jung Lee

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Survey Methodology)
in the University of Michigan
2017

Doctoral Committee:

Senior Research Scientist Steven G. Heeringa, Co-Chair
Professor Trivellore E. Raghunathan, Co-Chair
Professor Roderick J. Little
Professor Richard L. Valliant

To my family

ACKNOWLEDGMENTS

I would like to express my gratitude to my committee chairs Dr. Steve Heeringa and Dr. Trivellore Raghunathan. I am grateful to Dr. Trivellore Raghunathan for suggesting the topic for my dissertation. I appreciate Dr. Steve Heeringa for his careful attention to much of my writing. Our frequent meetings over the years provided me with valuable and helpful interventions to my dissertation research.

I am most grateful to my committee members, Dr. Richard Valliant and Dr. Rod Little for their guidance and assistance on my research. I am thankful to them for continuously providing in depth insightful comments and constructive criticism. I appreciate their patience, encouragement, and immense knowledge. As great scholars as they are, they are inspirational in many ways, but especially because of their benevolent and determined mentorship. Yet mostly, I am thankful for their flexible ability to appreciate my language and laugh at my jokes.

My sincere thanks also goes to fellow students, staff, and faculty members in the program for their support and help over the years. I enjoyed the friendship, the interesting research projects, and the unbelievable culture and language at the department. Special thanks to librarians Lee Ridely and Yan Fu for their companionship at the small library of Population Study Center. I am also grateful for the tremendous assistance on my writing from Chris Feak from English Language Institute and faculty at the Sweetland Writing Center. Most importantly, I am thankful for Lara Badke; her compassion and kindness at a crucial moment helped me to reach an end to my PhD journey.

I would not have been able to finish this dissertation without the love and sacrifice of my family: to my parents, whose unconditional love has always been my encouragement and support; to my husband, whose trust and sacrifice allows for the flexibility of my schedule; to my daughter, whose confidence and independence makes me so proud; and to myself, whose unbelievable grit allows me to persevere through the direst of situations.

Finally, the submission of this dissertation concludes my journey for the PhD degree. I will always be thankful to be part of the unbelievable, world-renowned Program in Survey methodology at Michigan, where all the students are strong, all the staff are good looking, and all the faculties' wages are above average.

TABLE OF CONTENTS

Chapter

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Traditional survey design draws a representative sample and implements post-survey weighting adjustments to compensate for nonresponse. When survey participation decline renders respondents nonrepresentative, the effectiveness of post-survey weighting adjustment becomes uncertain. Recent developments to improve respondent representativeness via adaptive data collection design have delivered promising results on bias reduction.

This dissertation develops a new adaptive design to improve survey data quality, by capitalizing on a benchmark data which captures the target population. The basic idea is to adaptively draw samples that lead to representative respondents; and to compensate for nonrespondents by benchmarked imputation procedures. Respondent representativeness is enhanced by the sampling procedure as opposed to data collection, eliminating costs of nonresponse follow-up and inferential complexity due to varying data collection protocols.

The new adaptive design consists of benchmarked sequential sampling (BSS) and benchmarked multiple imputation (B-MI) procedures. The new design first improves respondent representativeness by BSS, which conforms either the frame variables alone (BSS-Z) or both frame and survey covariate information (BSS-X) to those of the benchmark. With improved respondent representativeness, the benchmarked multiple imputation recovers the population information, leading to better quality survey estimates that are less susceptible to the unknown nonresponse pattern. This design applies to surveys with rich micro-level auxiliary data and surveys that use respondents of other surveys as sampling frame.

The BSS-Z method is demonstrated using the National Health Interview Survey and Behavior Risk Factor Surveillance System; the BSS-X and the benchmarked MI methods are demonstrated using the American Community Survey, the Current Population Survey, and the Census Planning Database.

An evaluation is done between the new design of adaptive sampling and imputation and the traditional design of fixed sampling and weighting (generalized regression estimator). To assess respondent representativeness, data from the new design is compared to those of the benchmark in marginal, conditional, and descriptive statistics. To assess the quality of the survey inference, a sample mean is calculated along with its root mean square error (RMSE), bias and coverage rate. To assess whether a design is of better value, a cost-effectiveness measure is derived from RMSE and a new cost model.

CHAPTER 1

Introduction

Current survey practices are unsustainable, in part attributable to participation problems and cost inflation. Declining survey participation exacerbates the concern over nonresponse bias and the effectiveness of the traditional inferential paradigm. Strategies that aim at reducing nonresponse bias and increasing respondent representativeness often involve extensive nonresponse follow-up and prudent weighting adjustments, where the former demands higher costs and latter requires the missing at random assumption.

To improve respondent representativeness and to minimize nonresponse bias, multiple actions are often taken. Three common strategies are: 1) increasing the overall survey response rate, 2) improving respondent representativeness by adaptive data collection methods, and 3) applying post-survey weighting adjustments.

Each of the current strategies in nonresponse bias reduction has limitations that may leave this bias unchecked and/or increase the cost with uncertain benefits to survey inference. For example, increasing the overall survey response rate does not necessarily reduce nonresponse bias due to differential responses among subgroups of interest (Groves, 2006). Improving respondent representativeness by adaptive data collection methods increases cost and complicates the inferential process (Brick, 2013). Post survey weighting adjustments reduce nonresponse bias if nonresponse is missing at random (MAR) and good weighting variables exist (Lundström and Särndal, 1999).

To improve survey inference this dissertation proposes an alternative inferential paradigm that shares the principle of research on combining survey data (Schenker, et. al., 2002; Schenker and Parker, 2003; Parker, et. al., 2004; Schenker and Raghunathan, 2007; Raghunathan, et. al., 2007; Schenker, et. al., 2010). The implementation of the paradigm necessitates a survey with multi-phase setting (see section 1.1 for details), permitting adaptive and sequential improvement of survey inferences at each phase (Groves and Heeringa, 2006). The approach utilizes high quality micro-level auxiliary data (a survey or census), in a multi-phase survey setting, to sequentially guide the sampling design and post survey adjustments of a focal survey. We termed the micro-level auxiliary data the "benchmark", which serves as a surrogate for the unobserved target population. At each sampling phase, the nonrespondents render the responding sample nonrepresentative and the sample is drawn to rebalance the distribution of the respondent data and the benchmark. Appending the respondent data to the benchmark, unit nonresponse is then mimicking the monotone item nonresponse and is replaced with benchmarked multiply imputed data. Inferences are obtained by applying standard combining rules to the multiply imputed data (Rubin, 1983). Survey estimates are greatly enhanced because both sampling and imputation are guided by the benchmark, resulting in survey estimates that are less susceptible to bias due to unknown nonresponse patterns.

The existing methods in combining survey data focus on improving survey inference after survey completion, for example, Schenker and Raghunathan (2007) mentioned (1) combining estimates from a survey of households and a survey of nursing homes to extend coverage (Schenker, et. al., 2002); (2) using information from an interview survey to bridge the transition in race reporting in the United States census (Schenker and Parker, 2003); (3) combining information from an examination survey and an interview survey to improve on analyses of self-reported data (Schenker, Raghunathan, and Bondarenko, 2010); and (4) combining information from multiple interview surveys to enhance small-area estimation (Raghunathan, et. al., 2007, Dong, Elliott, and Raghunathan, 2014).

Instead of repairing after survey completion, improving inferences during survey data collection has gained growing interest, evidenced by ample articles in responsive design and adaptive data collection design in recent years. The term responsive design, first introduced by Groves and Heeringa (2006), refers to a survey design strategy that is implemented in phases. At each sequential phase of a responsive design, sample design features and survey procedures are modified with the aim of minimizing cost and errors for the final survey product. Adaptive data collection design improves survey inferences by implementing a tailored data collection strategy to nonresponse follow-up. However, varying strategies on data collection influence response propensity, complicate the inferential process, and incur higher cost. Adaptive adjustment at the sampling stage may be ideally suited for improving survey inferences during survey implementation while overcoming the limitations of adaptive data collection on inferential complexity and cost issues.

The proposed paradigm capitalizes on the benchmark data, including a sampling and an imputation step, where each step implements a widely used methodological technique that originally was applied to other statistical problems. The sampling step builds on the propensity score method that, first proposed by Rosenbaum and Rubin (1983), is widely used for observational studies to adjust for selection bias. The imputation step recovers nonrespondent and population information by multiple imputation (MI) using the Multivariate Imputation by Chained Equation (MICE) (van Buuren and Groothuis-Oudshoorn, 2011). The MICE algorithm, first proposed by Kennickell (1991), have enjoyed popularity in recent years in dealing with complex missing data problems. MICE algorithm have appeared in the literature time and again with a variety of names, including fully conditional specification (FCS) (van Buuren, 2007) and sequential regression multivariate imputation (SRMI) (Raghunathan et. al., 2001).

Finally, adaptive sampling improves respondent representativeness by over-sampling subjects with lower response propensity which induces higher costs. Some speculate that not only the reduced nonresponse bias is at the expense of higher cost but also that the bias could have been adjusted by traditional weighting methods. We evaluate the cost and error of the proposed strategy as compared to the conventional practices.

In this dissertation, I develop an alternative inferential paradigm to adjust for unit nonresponse. Through effective use of the auxiliary data, the benchmark, the objectives of this dissertation are:

- Chapter 2: to develop a new adaptive sampling method to improve respondent representativeness.
- Chapter 3: to further extend the adaptive sampling and to develop a new estimation strategy to improve survey inference.
- Chapter 4: to evaluate the new method based on cost and error criteria.

1.1 Background and Significance

The diminishing survey participation has reached the juncture that threatens the validity of weighting adjustments. The increasing difficulties in measuring a diverse society coupled with the increasing cost of conducting the same surveys motivate a major paradigm change in the survey field. One promising path forward is to combine different data sources and capitalize on the dividends of auxiliary data. Recent development on utilization of auxiliary data in adaptive tailoring of data collection protocols to nonresponse follow-up shows promising results in nonresponse bias reduction.

The premise of nonresponse bias reduction through adaptive data collection design is built upon enhancing respondent representativeness. One opportunity in enhancing respondent representativeness resides at the sampling stages for a survey with sequential replicate phases. The data collection for many contemporary surveys such as the National Health Interview Survey (NHIS), is organized into sequential phases. New replicates (e.g. monthly or quarterly samples) of the probability sample are introduced at the beginning of each sequential phase. These surveys therefore subdivide the total sample into separate replicates so that each replicate is a representative sample of the target population of interest. These sequential, ”multi-phase” sample surveys have a number of advantages, including flexibility for controlling total sample size and cost.

Traditionally, in a multi-phase survey the sampling design – its stratification, sample allocation and sample size – does not change substantially from one phase to the next (we term this conventional approach a fixed sampling design). We propose a sampling design that aims to sequentially align respondent data distributions with the population during survey data collection, which provides an opportunity to minimize the differences between respondent and population distributions. Such a design strategy is particularly useful in the context of a multi-phase survey where the data collection is often organized into several discrete phases. As successive phases of data collection are rolled out, characteristics of the respondent pool sequentially conform to known population characteristics. There are practical limits to the number of such phases that are feasible to employ in actual survey data collection. This dissertation presents examples in which the full data collection is implemented in four successive phases, corresponding to four quarters of a year.

Adaptive design strategies targeting nonresponse bias reduction during survey design and data collection have been investigated in a number of prior studies. These strategies include, but are not limited to, 1) using auxiliary and contextual data available from both respondents and nonrespondents to estimate response propensity for sample cases to guide nonresponse follow-up strategies; 2) assigning a tailored protocol to cases with low predicted response propensity in an effort to increase respondent representativeness; and 3) altering design parameters to oversample subgroups of interest. Among these strategies, the overall response rate for the focal survey is increased by case prioritization where cases with low predicted response propensity or low predicted contact probability receive different protocols.

Although the objective of these studies is nonresponse bias reduction, strategies designed to achieve this goal are often ad hoc in nature. For example, the adaptive design for the Community Advantage Panel Survey (Peytchev, et. al., 2010) implemented a revised sample allocation, after a high imbalance in the male/female distribution for several phases of data collection was observed.

The decision to adjust the sampling rate for males was based on expert judgment that further data collection would not correct the observed trend to gender imbalance in the final observed sample.

Several Norwegian surveys also implemented adaptive design where the primary objective of reducing potential nonresponse bias is achieved by improving respondent representativeness (Lagerstrom, et. al. 2010). In these surveys, respondent representativeness is evaluated by Representativeness indicators, namely, R-indexes and partial R-indexes (Schouten, et. al., 2009; Schouten, et. al., 2011; Schouten, et. al. 2007). R-indicators use propensity methods to evaluate the representativeness of the achieved respondent sample with respect to auxiliary variables that are available from external sources for both respondents and nonrespondents. R-indicators are used as a monitoring device to identify characteristics of the underrepresented subjects and to optimize the field data collection strategies.

For example, R-indicators were computed once a week in the 2006 pilot of the Level of Living Survey (LLS) test group. While monitoring LLS field work, the conditional partial-R indicator indicated an under-representation of younger age group (under 35 years), triggering an intervention at day 18 to prioritize young adults in the computer-assisted telephone interview (CATI) call schedule and to focus on mobile phone numbers (due to the hypothesis of higher response propensity on mobile phone by young adults). At day 24, a second intervention was triggered by the partial-R indicator to prioritize previous panel refusers. These interventions had a positive effect on the representativeness of the respondents when compared to the LLS control group, which implemented a traditional fixed design.

The major differences between the proposed adaptive sampling and other adaptive/responsive design described in recent literature are focus on two aspects: First, our adaptation is at the sampling design stage instead of at the data collection and nonresponse follow-up stage. Second, we balance respondent distributions with those of a benchmark instead of the selected sample.

For the first aspect, adaptive design or responsive design requires continually monitoring the data collection process and deploying alternative data collection protocols tailored to specific nonrespondents (Groves and Heeringa, 2006; Schouten, et. al., 2009; Schouten, et. al., 2011; Peytchev, et. al., 2010; Schouten, et. al., 2013). Respondent representativeness was evaluated by Representativeness indicators (Schouten, et. al., 2007; Schouten, et. al., 2009; Schouten, et. al., 2011). With different data collection protocols applied to different sample units, the estimation of response propensity and the nonresponse weighting adjustments require more understanding of the effects of data collection efforts on biases. Brick (2013) provides discussion and examples of the connection between data collection and nonresponse adjustments. The proposed adaptive sampling improves respondent representativeness at the sampling stage while keeping data collection protocols unified, resulting in more robust survey estimates.

For the second aspect, we incorporated a benchmark study that resembles the characteristics of the target population. The idea of incorporating one survey to improve analysis of another survey has been reported in other contexts (Schenker, et. al., 2007; Raghunathan, et. al., 2007; Schenker, et. al., 2009). Yet in previous research on combining surveys, the goal was to improve analysis and estimation after the fact, as opposed to guiding sampling design to obtain a representative respondent set. Our method improves respondent set representativeness through introducing of an external benchmark data source, which allows our design strategy to focus directly on respondent composition with respect to the target population. In addition, this strategy provides a statistics-based method for modifying sampling inclusion probabilities for sample replicates introduced at the successive stages of data collection. The strategy leads to the joint distribution for variables in the focal survey respondent sample converging in expectation to that of the chosen benchmark.

1.2 Benchmarked Sequential Sampling and Benchmarked Multiple Imputation Approach

The proposal is to use high quality micro-level auxiliary data to guide the sampling and imputation of a focal survey. For the micro-level auxiliary data, we consider those that 1) are available before survey implementation and 2) capture the information on the target population of interest. We term such auxiliary data as a “Benchmark” where it shares some frame variables and survey covariates with the focal survey. In a multi-phase survey setting, the benchmark-calibrated samplings sequentially improve the respondent representativeness. With improved respondent representativeness, the benchmark-calibrated imputation recovers the population information. The survey inference is derived from the joint dataset, which encompasses respondent data, imputed nonrespondent data and the benchmark data.

Chapter 2 presents a new adaptive sampling method that sequentially conforms the focal survey frame variables to those of the benchmark. At any given phase, auxiliary variables shared between the focal survey sample frame and a benchmark data source are used to estimate the propensity of being in the benchmark. The adaptive sampling rate is derived from the estimated propensity scores for the subsequent phases to restore balance between the eventual respondents and the benchmark. As this procedure is repeated, the distribution of respondent propensity scores converges to that of the benchmark, improving sample balance for not only the auxiliary variables included in the propensity score model but also survey variables of interest that are correlated with the auxiliary variables. The strategy is evaluated by applying various nonresponse mechanisms to data with different strengths of association between auxiliary variables and survey variables. This sampling strategy is illustrated via both a simulated multivariate normal dataset and data from two large government surveys (National Health Interview Survey, NHIS, and Behavior Risk Factor Surveillance System, BRFSS).

Chapter 3 extends the adaptive sampling method from the previous chapter to conform both sample frame variables and survey covariates of the focal survey to those of the benchmark. We describe a benchmark-driven mitigation and imputation (M&I) strategy, in the context of a multi-phase survey, that sequentially guides the sampling and estimation to improve survey inferences regardless of nonresponse mechanism. The M&I strategy employs a high quality benchmark to 1) (mitigate) rectify undesirable nonresponse patterns through a calibrated sequential sampling design; and 2) (impute) recover population information through calibrated Multivariate Imputation by Chained Equations (MICE), consequently achieving less biased survey estimates. The performance of the M&I strategy is evaluated by simulation experiments to mimic adaptive design under various nonresponse mechanisms, including missing not at random (MNAR). We report on the preservation of marginal and joint distributions for population estimates of three sampling designs from respondent data, completed data, and joint data. An illustration using data from the American Community Survey (ACS) and the Current Population Survey (CPS) is also presented.

Chapter 4 assesses the cost-effectiveness (*efficacy*), a function of cost and root mean square error (RMSE), for weighted and imputed estimates of a sample mean. A subject-level cost model is developed that is generalizable across survey designs. We present two post-survey adjustment methods: 1) benchmarked multiple imputation (B-MI) and 2) a conventional weighting strategy using generalized regression estimator (GREG). Benchmarked multiple imputation is the proposed alternative strategy in post-survey adjustment that is analogous to the current practice of post-survey weighting adjustments. A simulation study is conducted to evaluate the cost and error implications when the missing at random (MAR) assumption does and does not hold (i.e. missing not at random, MNAR).

Chapter 5 concludes the dissertation with discussion and suggests some future research.

1.3 Benefits and Potential Impacts

This research has several potential impacts. The most important impact is the potential of generating better inferences that are less susceptible to an unknown nonresponse mechanism. The quality of the survey inferences through the implementation of traditional survey design and post survey adjustment relies on the missing at random assumption. With high levels of nonresponse, the plausibility of the MAR assumption diminishes. Although adaptive data collection design may reduce nonresponse bias, it also increases survey cost with uncertain benefit for reducing survey errors.

Adaptive sampling design and benchmarked imputation offer an alternative approach to enhance survey inferences that minimizes nonresponse bias independent of nonresponse mechanism.

This research also provides a viable path forward in survey practice. By taking advantage of external data sources, the proposed approach improves respondent representativeness and achieves better cost-effectiveness without complicating field administration and inferential process. Prior studies in combining surveys have shown the benefit of enhanced survey estimates in reducing both sampling and non-sampling errors at the data analysis stage. This dissertation demonstrates a proactive approach to enhance survey estimates during survey implementation. With advances in technology, computation, and availability of electronic survey data and other data sources, improving survey inferences without increasing cost by effective use of auxiliary data is a feasible and sensible strategy.

The dual difficulties in survey participation and fiscal distress challenge the conventional survey inferential paradigam. When survey participation reaches the point that threatens validity of the MAR assumption, applying the fixed sampling design and drawing a design-based representative probability sample is no longer practical. The sampling methods proposed in this dissertation are flexible enough to accommodate various nonresponse mechanisms. The sampling phases could be expanded or halted, and therefore the multi-phase setting is flexible enough to accommodate unforeseen challenges in survey budget and administration resources. These procedures may be adopted by any multi-phase survey when appropriate benchmark data are available.

The benchmarked multiple imputation can be easily applied by the current computation software. Although the proposed imputation approach aims at compensating for unit nonresponse, item nonresponse can also be predicted simultaneously. This study may also stimulate a shift in how external data sources are utilized and how nonresponse is addressed in survey practice.

CHAPTER 2

Improving Respondent Representativeness Using External Micro-level Benchmark Data

Increasing unit nonresponse in surveys may render the observed data unrepresentative of the population, leading to biased inference. A standard practice is to use weighting as post survey adjustments for nonresponse bias which involve implicit or explicit assumptions concerning the missing data mechanism. Other common strategies include improvements in data collection methods aimed at increasing overall response rates or respondent representativeness. This chapter describes a new adaptive sampling strategy that uses external micro-level "benchmark" data for the survey population in a multi-phase survey to achieve improved representation in the respondent sample. At any given phase, auxiliary variables shared between the focal survey and a benchmark data source are used to estimate the propensity of being in the benchmark. The adaptive sampling rate for the subsequent phases is derived from the estimated propensity scores to restore balance between the eventual respondents and the benchmark. As this procedure is repeated, the distribution of respondent propensity scores converges to that of the benchmark, improving sample balance for not only the auxiliary variables included in the propensity score model but also survey variables of interest that are correlated with the auxiliary variables. This sampling strategy is illustrated via both a simulated multivariate normal dataset and data from two large government surveys (National Health Interview Survey, NHIS, and Behavior Risk Factor Surveillance System, BRFSS). The strategy is evaluated by various nonresponse mechanisms applied to data with different strength of association between auxiliary variables and survey variables.

2.1 Introduction

Increasing unit nonresponse in surveys may render the observed data unrepresentative of the population, calling into question the validity of survey inferences due to the potential for nonresponse bias, especially if response propensity is related to the variables of interest (Little and Vartivarian, 2005; Tourangeau, et. al., 2013).The seriousness of the nonresponse was illustrated by the Pew Research Center in 2012 that reported an average 9% response rate (or 91% nonresponse rate) for a typical telephone random digit dial (RDD) survey (Kohut, et. al., 2012). Three common strategies, applied simultaneously at times, that aim to reduce nonresponse bias are 1) increasing the overall survey response rate 2) improving respondent representativeness by tailoring data collection methods, and 3) applying post-survey nonresponse weighting adjustments. However, research shown that for strategy 1 – there is a lack of consistent association between the overall response rate and nonresponse bias (Groves, 2008; Bootsma, 2002; Barclay, 2002; Groves, 2000; Keeter, 2000; Curtin, 2000; Merkle, 2002; Groves, 2006b); for strategy 2 – tailored data collection alters response propensity, complicating nonresponse adjustments (Olsen and Groves, 2012; Schouten, et. al., 2011; Brick, 2013); for strategy 3 – unverifiable missing at random (MAR) assumption underminds the effectiveness of weighting adjustments (Brick and Kalton, 1996).

This study describes an alternative bias reduction strategy – adaptive sampling design, that incorporates an external micro-level "benchmark" data set to improve respondent representativeness in a new multi-phase survey (the focal survey). For the micro-level benchmark data, we consider those that 1) are available before survey implementation and 2) capture the information of the target population of interest. The benchmark could be a high quality survey or administrative data (see Section 2.2 for more details). Hence, surveys with linked administrative/census data and surveys using other survey respondents as the sampling frame are ideal candidates for this sampling design. The success of the proposed sampling design on bias reduction depends on the orrelations among auxiliary variables, survey covariates, and survey outcome variables.

Adaptive sampling here refers to a sampling strategy that aims to align respondent data distributions with the population using a dynamic sampling strategy. The representativeness of the respondents is enhanced across the phases by a targeted probability sampling strategy that is guided by sampling rates derived from propensity scores. Propensity scores provide a scalar summary which could potentially be based on extensive covariate information. These propensity scores express the propensity of the benchmark membership, instead of the propensity of respondent membership used in the standard nonresponse weighting adjustments. The ratio of the empirical propensity score densities of the focal and benchmark surveys guides sampling decisions for the upcoming phase. As the phases progress, the distribution of respondent propensity scores converges to that of the benchmark data, establishing sample balance for the benchmark characteristics included in the propensity score model.

The adaptive sampling strategy is particularly useful in the context of a multi-phase survey where the sampling decision is organized into several discrete phases so that characteristics of the respondent pool can be matched to known population characteristics as successive phases of the sampling design are rolled out. The proposed adaptive sampling shares the same goal of improving respondent representativeness as with the “adaptive design” and “responsive design”, yet differs fundamentally from their focus in data collection approaches (Groves and Heeringa, 2006; Schouten, et. al., 2009). Three key components that distinguish our study from that of prior work are 1) the incorporation of a benchmark survey, 2) the adaptive nature of the sampling design (as opposed to adaptive nature of data collection), and 3) the propensities of the benchmark membership (as opposed to the propensities of the response).

For the first component, we incorporate a benchmark data set that resembles the characteristics of the target population. The idea of incorporating one survey to improve analysis of another survey has been reported in other contexts (Schenker and Raghunathan, 2007; Raghunathan, et. al., 2007; Schenker, et. al., 2009). Yet in previous research on combining surveys, the goal is to improve analysis and estimation after the fact, as opposed to guiding sampling design to obtain a representative respondent set.

For the second component, the proposed adaptive sampling improves respondent representativeness at the sampling stage while keeping data collection protocols unified, resulting in more robust survey estimates. The major issue with the adaptive data collection design is that, with different data collection protocols applied to different sample units, the estimation of response propensity and the nonresponse weighting adjustments require more understanding about the effects of data collection efforts on biases. Brick (2013) gives detailed discussions and examples of the connection between data collection and nonresponse adjustments. The advantages of proposed adaptive sampling over the adaptive data collection are further discussed in Section 1.1.

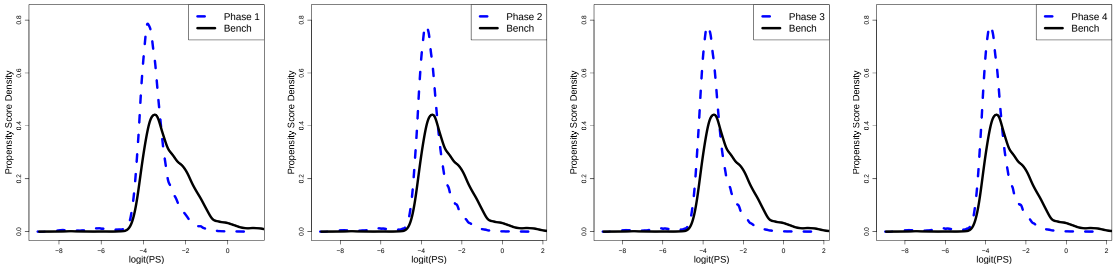
In our proposed method, the key element in improving respondent representativeness is therefore to balance multivariate distributions between respondents and the benchmark population. A natural way to balance the multivariate distributions for the focal sample and the benchmark is to balance propensity score distributions (Rosenbaum and Rubin, 1983; DAgostino, 1998). When using propensity scores to design an observational study, the goal is to match 'treated subjects' with 'control subjects' having similar measured covariates, minimizing noncomparability in the treated and control subjects and increasing the precision for estimating the 'treatment effect' (Rubin, 2002; Rubin, 2007; Austin, 2009; Hahn, et. al., 2011). This approach, however, is not a good fit with survey design for two reasons: 1) a priori matching of nonrespondents with comparable replacement subjects is not feasible since final response status for sample elements is not known prior to the survey; and 2) advance information required to assign matched substitutions to each eventual nonrespondent case is limited. Hence, the unfolding nature of the nonresponse problem as the survey data collection progresses undermines the utility of a standard propensity score matching approach that could be implemented in advance of actually conducting the survey.

Instead, in our application we model the "benchmark membership". Sampling units in the focal survey and benchmark with equal (or nearly equal) propensity scores will tend to have the same (or nearly the same) distributions on covariates included in the propensity score model (Rosenbaum and Rubin, 1985). Therefore, a respondent pool that has a propensity score distribution that is equal to that of the benchmark will also tend to have the same distributions on the covariates used to derive the propensity scores. Furthermore, benchmarking to the target population using propensity scores preserves the multivariate structure in terms of balance between the benchmark and focal surveys. Such a multivariate balance provides a more representative respondent pool for the focal survey with respect to both the marginal and joint distributions. More importantly, researchers have shown that the remaining bias after standard post-survey weighting adjustments could be further reduced had the original respondent sample been more representative (Särndal and Lundquist, 2014).

Traditional sampling strategies, labelled the "fixed sampling design" herein, in which a probability sampling design is kept stationary across data collection phases is a standard practice. With fixed sampling, the departure of respondent composition from the target population often remains as the data collection phases progress. For example, Figure 2.1 illustrates the estimates of the propensity score densities using selected variables from the 2009 NHIS and 2009 BRFSS publicly available micro-data. One can clearly see that the distributional differences of the NHIS (as Benchmark) and the BRFSS (as the respondents of the focal survey) persisted as the data collection progressed from Phase 1 to Phase 4. Without mid-survey intervention on sampling design, the resulting deviation of focal survey respondents from the target population continues at each data collection phase. The proposed strategy is illustrated via simulation studies which compare the representativeness of the respondents from adaptive sampling and fixed sampling designs under various nonresponse mechanisms.

In Section 2.2, we describe our proposed method, first presenting a simple univariate example to demonstrate the principle of our method and then formally presenting the algorithm required to

Figure 2.1: Observed BRFSS 2009 under fixed sampling design.



Estimated propensities of being in the benchmark (NHIS) of observed BRFSS 2009 under fixed sampling design where the solid line shows 2009 NHIS data and the dash-line shows the 2009 BRFSS. The similar patterns across phases indicates that, without mid-survey intervention, the nonresponse mechanism stays similar at each phase.

adjust focal survey sampling rates at each successive sampling phase to match the benchmark data set. Section 2.3 illustrates by means of a simulation study how the proposed adaptive sampling can improve the respondent representativeness under various degrees of explanatory power of the frame/auxiliary variables and various nonresponse scenarios. Section 2.4 illustrates an application example using data from 2009 BRFSS and 2009 NHIS where the BRFSS and NHIS data are combined to form a finite population. In Section 2.5 we conclude with some discussion of strengths and limitations of the proposed strategy along with areas for future research.

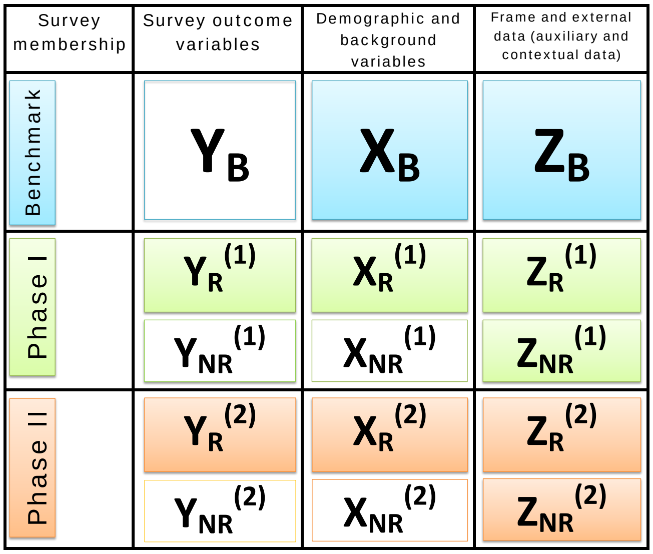
2.2 Methods

2.2.1 Data structure

Figure 2.2 illustrates the concept of our method. In the figure, shaded cells indicate available data; superscripts denote the data collection phases and subscripts indicate respondent (R), nonrespondent (NR), and benchmark (B) data. In our strategy for balancing respondent distributions, candidate data consists of focal survey variables of interest (Y), demographic and background variables (X) measured in the survey and benchmark data, and auxiliary and contextual data available from frame and external sources (Z). Currently, adaptive data collection strategies exclusively focus on using data that are available for both respondents and nonrespondents, i.e. Z_R and Z_{NR}. Alternatively, our method extends the representativeness of the respondent set through introduction of an external benchmark data source, denoted Z_B, X_B and Y_B which allows our design strategy to focus directly on respondent composition with respect to the target population.

In addition, our strategy provides a statistically-based method for modifying sampling inclusion probabilities for sample replicates introduced at the successive sampling phases such that the joint distribution for Y, X and Z in the focal survey respondent sample converges in expectation to that of the chosen benchmark; that is, $f(Y_R, X_R, Z_R) \approx f(Y_B, X_B, Z_B)$, where $f(.)$ denotes a density function. A simulation study in Section 2.3 illustrates that this goal can be achieved when $P(X_R, Z_R) \approx P(X_B, Z_B)$ and $P(Z_R) \approx P(Z_B)$, where $P(.)$ denotes the propensity density function, which can be obtained by benchmarking Z_R to Z_B through adaptive sampling.

Figure 2.2: Data available for benchmarked sequential sampling design



Survey membership	Survey outcome variables	Demographic and background variables	Frame and external data (auxiliary and contextual data)
Benchmark	Y_B	X_B	Z_B
Phase I	$Y_R^{(1)}$ $Y_{NR}^{(1)}$	$X_R^{(1)}$ $X_{NR}^{(1)}$	$Z_R^{(1)}$ $Z_{NR}^{(1)}$
Phase II	$Y_R^{(2)}$ $Y_{NR}^{(2)}$	$X_R^{(2)}$ $X_{NR}^{(2)}$	$Z_R^{(2)}$ $Z_{NR}^{(2)}$

Y denotes survey outcome variables, X denotes survey covariates, Z denotes frame and auxiliary variables that are available for both respondents and nonrespondents before data collection. B denotes benchmark, R denotes respondent, NR denotes nonrespondent. Superscript specifies the sampling phases. Shaded cells indicate available data, and nonshaded-area represents no data.

2.2.2 The Benchmarked Sequential Sampling Design

The proposed adaptive sampling strategy modifies the sampling probabilities at sequential stages in a focal survey with respect to a chosen benchmark to improve respondent representativeness. The respondent representativeness is defined as the similarity between the respondent and the benchmark propensity distributions ($P(\mathbf{Y_R}, \mathbf{X_R}, \mathbf{Z_R}) \approx P(\mathbf{Y_B}, \mathbf{X_B}, \mathbf{Z_B})$), where the propensity scores are estimated by modeling the benchmark membership (benchmark vs. focal survey), as opposed to the response membership (respondent vs. nonrespondent). Note that the proposed sampling design applies to the survey before data collection, hence response indicator is not relevant. The implementation of this strategy necessitates a focal survey with these elements: 1) the focal survey has a sequential sampling design; 2) the target population for the focal survey is captured in a population dataset (e.g. census) or a benchmark survey where micro-level information is available; and 3) the focal survey and the benchmark survey share some covariates ($\mathbf{X}$) and auxiliary variables ($\mathbf{Z}$). Note, Y is most likely not in the benchmark, otherwise there is no need to conduct the focal survey.

An example pair of benchmark and focal survey is the American Community Survey (ACS) and the National Health Interview Survey (NHIS). NHIS has a sequential sampling design where the annual sample is divided into four calendar quarters, and the sample of each calendar quarter is a probability sample of the target population. ACS captures the target population of the NHIS which is the U.S. resident civilian noninstitutionalized population. ACS and NHIS share common covariates ($\mathbf{X}$) such as the demographic and socio-economic variables. ACS and NHIS, both are address-based household surveys sponsored by U.S. government, sharing the same frame – U.S. Census Master Address File ($\mathbf{Z}$). Another example pair of benchmark and focal survey is NHIS and BRFSS. BRFSS sample could be assigned to sequential sampling phases that is similar to the NHIS – where sampling is implemented quarterly. NHIS and BRFSS share demographic and socio-economic variables ($\mathbf{X}$) as well as contextual information which includes variables listed in the Census Planning Database (PDB) such as census block level race distribution, age distribution, and gender ratio, etc. More details on the NHIS and BRFSS are described in Section 2.4.

Suppose no prior information on nonresponse patterns exist, the initial sample (phase I) of focal survey is drawn with a conventional fixed sampling design and the adaptive sampling starts at phase II. The phase I respondent data are compared to the data from the benchmark to evaluate the respondent representativeness. Note that the adaptive sampling also applies at phase I to surveys with prior nonresponse information. While nonrepresentativeness exists, subjects that are underrepresented in the focal survey respondent distribution with respect to Z will be sampled at a higher rate at the next phase of sampling design, whereas subjects that are overrepresented will be sampled in a lower rate. The derivation of the adaptive sampling rate builds upon benchmarking the respondent data at each sampling phase, using propensity scores and their empirical density function. To keep things simple, we assume that sample units are selected by simple random sampling (s.r.s). However, the points made here also apply in general to surveys with more complex designs, with more complicated technical details. As with any finite population survey, the proposed sampling strategy applies to the sampling frame without replacement.

Using simulated data we show how the objective of arriving at $P(\mathbf{Y_R}, \mathbf{X_R}, \mathbf{Z_R}) \approx P(\mathbf{Y_B}, \mathbf{X_B}, \mathbf{Z_B})$ can be achieved by the proposed benchmarked sequential sampling strategy. The strategy is derived using only the auxiliary variables (**Z**) since X and Y are not available at the sampling stage. We illustrate via simulation that $P(\mathbf{Y_R}, \mathbf{X_R}, \mathbf{Z_R}) \approx P(\mathbf{Y_B}, \mathbf{X_B}, \mathbf{Z_B})$ when $P(\mathbf{X_R}, \mathbf{Z_R}) \approx P(\mathbf{X_B}, \mathbf{Z_B})$, and which in turn can be obtained when $P(\mathbf{Z_R}) \approx P(\mathbf{Z_B})$. $P(\mathbf{Z_R}) \approx P(\mathbf{Z_B})$ is the direct result of the proposed sampling design. The simulation studies evaluate various correlation structures among Y, X and Z and how these structures affect the convergence of $P(\mathbf{Y_R}, \mathbf{X_R}, \mathbf{Z_R})$ to $P(\mathbf{Y_B}, \mathbf{X_B}, \mathbf{Z_B})$ under different missing data mechanisms.

More specifically, Figure 2.3 depicts the conceptual framework of the sampling population at each phase for a four-phase survey. The goal for the cumulative focal survey respondents propensity distribution approximating that of the population is achieved by sequential nature of the benchmarked sequential sampling, which can be described as (assuming benchmarked sampling starts at phase II),

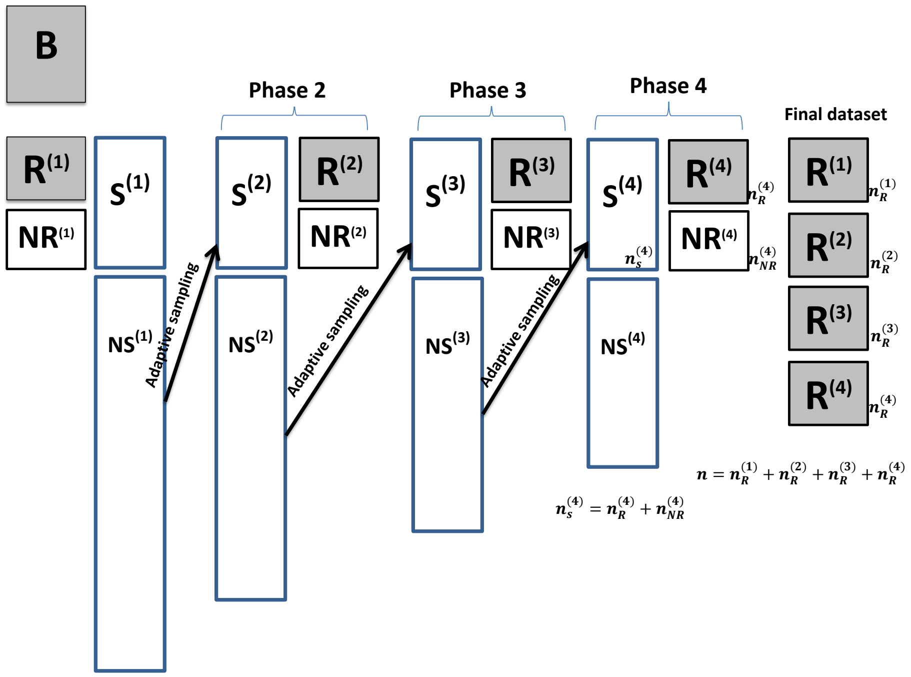
- at phase I: sampling goal is $P(\mathbf{Z}^{(1)}) = P(\mathbf{Z}_B)$, while obtaining $P(\mathbf{Z}_R^{(1)}) \neq P(\mathbf{Z}_B)$.
- at phase II: sampling goal is $Q(\mathbf{Z}^{(2)}) = P(\mathbf{Z}_B)$, with objective of $P(\mathbf{Z}_R^{(2)}|\mathbf{Z}_R^{(1)})$ approaching $P(\mathbf{Z}_B)$.
- at phase III: sampling goal is $Q(\mathbf{Z}^{(3)}) = P(\mathbf{Z}_B)$ with objective of $P(\mathbf{Z}_R^{(3)}|\mathbf{Z}_R^{(1)}, \mathbf{Z}_R^{(2)})$ further approaching $P(\mathbf{Z_B})$.
- at phase IV: sampling goal is $Q(\mathbf{Z}^{(4)}) = P(\mathbf{Z}_B)$ with objective of $P(\mathbf{Z}_R^{(4)}|\mathbf{Z}_R^{(1)}, \mathbf{Z}_R^{(2)}, \mathbf{Z}_R^{(3)}) \approx P(\mathbf{Z}_B)$.

where $Q(.)^{(k)}$ denotes density function for cumulative data from phase 1 to phase k. Suppose $Q(\mathbf{Z}_R^{(4)}) \approx P(\mathbf{Z}_B)$, then $Q(\mathbf{X}_R^{(4)}|\mathbf{Z}_R^{(4)})Q(\mathbf{Z}_R^{(4)}) \approx P(\mathbf{X}_B|\mathbf{Z}_B)P(\mathbf{Z}_B)$. That is, $Q(\mathbf{X}_R^{(4)}, \mathbf{Z}_R^{(4)}) \approx P(\mathbf{X}_B, \mathbf{Z}_B)$. When X and Z correlate with Y, $Q(\mathbf{Y}_R^{(4)}|\mathbf{X}_R^{(4)}, \mathbf{Z}_R^{(4)}) \approx P(\mathbf{Y}_B|\mathbf{X}_B, \mathbf{Z}_B)$, although $\mathbf{Y_B}$ is likely not available in the benchmark.

With regard to the selection of $\mathbf{Z}$ variables, the rule of thumb is similar to the variable selection for regression analysis. We recommend choosing $\mathbf{Z}$ variables that are correlated with $\mathbf{Y}$ and/or $\mathbf{X}$. Any types of $\mathbf{Z}$ variable and any number of $\mathbf{Z}$ variables work well. Although the inclusion of nuisance $\mathbf{Z}$ variables may not help with nonresponse bias reduction, it does not hurt either. It would be comparable to what a weighting adjustment could achieve using these $\mathbf{Z}$ variables.

The proposed strategy calls for a sequential sampling design where each sampling phase consists of a probability sample representing the target population. Aside from its flexibility of implementing benchmarked sampling to improve respondent representativeness, such a sampling design has many administrative advantages. For example, the design of sequential sampling phases where each phase represents the target population allows both the reduction or expansion of the survey phases should availability of funding and other resources change.

Figure 2.3: Proposed adaptive sampling design.



Benchmarked sequential sampling is applied to the next phase sample using previously non-sampled units. Note: S = sampled, NS = not-sampled, B = Benchmark, R= respondents, NR = nonrespondents, n = sample size. Superscript specifies the phases.

2.2.3 Illustrative Example with One Auxiliary Variable

Before the formal derivation of the method, we present a simple hypothetical example of a survey on school teachers and educators in which the auxiliary variable on the frame is gender (Z). Variables collected in the survey are education (X) and personal income (Y) where personal income is the survey outcome of interest.

The goal of the hypothetical survey is to estimate average personal income overall, and income by education and gender. Ideally, one would like to obtain a respondent pool where the joint distribution of income, education and gender agrees or nearly agrees with that of the population. At the sampling design stage, however, only gender information is available. Income and education are available only on respondents and only after survey data is collected. Since there is information in gender about income and education, this example shows that a representative respondent pool with respect to gender will improve the representativeness of the joint distribution on gender, education and income. With respect to the proposed sampling design, the goal at each phase is to rebalance the gender distribtion of the respondent pool with respect to that of the benchmark.

Suppose there are equal numbers of females and males in the target population of interest (which is captured in the benchmark), and we plan two phases of data collection with 100 subjects for each phase. An initial simple random sample (s.r.s.) of size 100 is selected and fielded. We obtain 60 respondents, of whom 38 (63%) are female and 22 (37%) are male. In the second phase of sampling to rebalance the respondent sample we must undersample females and oversample males, so that the gender distribution in the accumulated respondents and in the target population agree.

To write out the above problem algebraically, consider the aim of obtaining 50% females in the final sample. The target female percentage of 50% is achieved by combining the female percentage from the first phase respondents, 0.6×0.63, and the second phase female sample, $0.4 \times F$, where F represents the desired female proportion for the second phase sample. We can write

$$0.4 \times F + 0.6 \times 0.63 = 0.5 \tag{2.1}$$

where $F = (0.5 - 0.6 \times 0.63)/0.4 = 0.305$.

Using mathematical notation to rewrite formula (2.1), we have

$$(1 - \pi) \times P(Z|sample) + \pi \times P_R(Z) = P_B(Z) \tag{2.2}$$

where Z denotes gender, $P(\cdot)$ denotes probability distribution, and π denotes the proportion of sample subjects who responded at the first phase of data collection. $P_R(Z)$ denotes the gender distribution of the first phase respondents, and $P_B(Z)$ denotes the benchmark gender distribution. $P(Z|sample)$, the desired female distribution F in (2.1), is now the desired gender distribution for subjects being selected in the second phase sample. Since $P_R(Z)$, $P_B(Z)$, and π are known, $P(Z|sample)$ can be computed from (2.2). We obtain $P(Z|sample) = (P_B(Z) - \pi \times P_R(Z))/(1 - \pi)$.

However, what we really want to know is how to select additional sample cases that would most likely result in the desired $P(Z|sample)$. In other words, we actually want to know $P(sample|Z)$, the sampling rate conditioning on the gender distribution. The association between $P(sample|Z)$ and $P(Z|sample)$ can be expressed as

$$P(sample|Z) = \frac{P(Z|sample)P(sample)}{P(Z)} \tag{2.3}$$

where $P(Z)$ is the population Z distribution. Putting formulas (2.2) and (2.3) together, we obtain

$$\begin{aligned} P(sample|Z) &= \frac{P_B(Z) - \pi \times P_R(Z)}{(1 - \pi)} \frac{P(sample)}{P(Z)} \\ &= \{\frac{P_B(Z)}{P(Z)} - \pi \times \frac{P_R(Z)}{P(Z)}\} \frac{P(sample)}{(1 - \pi)} \end{aligned} \tag{2.4}$$

In this one auxiliary variable example, $\pi = 0.6$, $P_R(Z) = 0.63$, and $P_B(Z) = 0.5$. Hence, assuming $P_B(Z) = P(Z)$, $P(sample|Z) = [1 - 0.6 \times (0.63/0.5)] \times P(sample)/0.4$, where $P(sample)$ can be computed using the fact that $P(sample|Z)$ is a probability and bounded by 0 and 1.

A caveat is that samples selected by $P(sample|Z)$ in equation (2.4) will suffer nonresponse. Therefore $P(Z|sample)$ in formula (2.2) can be derived to take into account anticipated nonresponse. The anticipated nonresponse rate is estimated by the nonresponse rate observed in the previous phases. Sampling with nonresponse adjusted $P(Z|sample)$ will further improve the respondent representativeness when nonresponse rate remains similar across data collection phases, such as the example in Section 2.1 for NHIS and BRFSS.

In practice, imposing differential sampling rates in different domains (e.g. gender by education) is feasible with a frame that is a rich administrative data source. In situations where frame information and external data source are limited, the goal of targeted sampling rates can be accomplished by strategies such as screening (Botman and Moriarity, 2000) or prediction of sample characteristics.

2.2.4 Computing the Sampling Rate in Successive Sampling Phases

More formally, suppose there are k sampling phases, where $k = 1, \ldots, K$. Rewrite equation (2.2) for **Z**, a vector of covariates, and generalize to k phases, we have

$$(1 - \pi^{(k)}) \times P^{(k+1)}(\mathbf{Z}|sample) + \pi^{(k)} \times Q^{(k)}(\mathbf{Z}) = P_B(\mathbf{Z}) \tag{2.5}$$

where $Q^{(k)}(\mathbf{Z})$ is the cumulative respondent **Z** distribution obtained up to the k^{th} phase, $P^{(k+1)}(\mathbf{Z}|sample)$ is the **Z** distribution desired for the $(k+1)^{th}$ phase, and $\pi^{(k)}$ denotes the cumulative sample proportion at the k^{th} phase. For example, suppose the cumulative respondent size at phase k is n_k and the cumulative sample size at phase k is N_k. Then for phase k we have $\pi^{(k)} = n_k/N_k$ and $1 - \pi^{(k)} = 1 - n_k/N_k$

Again, the goal is to derive $P^{(k+1)}(sample|\mathbf{Z})$, the sampling probability of $(k+1)^{th}$ phase conditioning on $\mathbf{Z}$. The derivation is similar to the univariate case in equation (2.4). Using Bayes formula (2.3) and $P^{(k+1)}(\mathbf{Z}|sample)$ from equation (2.5), equation (2.4) can be rewritten as

$$\begin{aligned} P^{(k+1)}(sample|\mathbf{Z}) &= \{\frac{P_B(\mathbf{Z})}{P^{(k)}(\mathbf{Z})} - \pi^{(k)}\frac{Q^{(k)}(\mathbf{Z})}{P^{(k)}(\mathbf{Z})}\}\frac{P^{(k)}(sample)}{1-\pi^{(k)}} \\ &\propto \frac{P_B(\mathbf{Z})}{P^{(k)}(\mathbf{Z})} - \pi^{(k)} \times \frac{Q^{(k)}(\mathbf{Z})}{P^{(k)}(\mathbf{Z})} \end{aligned} \tag{2.6}$$

Using the sampling rate in (2.6), the $(k+1)^{th}$ phase attempts to correct for the imbalance between the benchmark and the focal survey with respect to the $\mathbf{Z}$ distribution. This procedure is repeated for the remaining $K-k$ phases until the distribution of $\mathbf{Z}$ from the cumulative respondents matches that of the benchmark, and a pool of more representative respondents is obtained. That is, $Q^{(k)}(\mathbf{Z}) \approx P_B(\mathbf{Z})$.

However, sampling with $P(sample|\mathbf{Z})$ in equation (2.6) may not achieve balance at one round of data collection, even with a 100% response rate. Consider denoting $\mathfrak{A} = \frac{P_B(\mathbf{Z})}{P(\mathbf{Z})} - \pi \times \frac{Q(\mathbf{Z})}{P(\mathbf{Z})}$, $\mathfrak{A}$ may take negative values and values larger than one. Note that phase indicator k is omitted for simplicity. We know that $P(sample|\mathbf{Z})$ is bounded by $(0,1)$. That is, when $\mathfrak{A}$ is negative, we set $P(sample|\mathbf{Z}) = 0$; when $\mathfrak{A} > 1$, we set $P(sample|\mathbf{Z}) = 1$. Either condition causes the truncation of $P(sample|\mathbf{Z})$. With the characteristic of truncation at 0 and 1, sampling with $P(sample|\mathbf{Z})$ requires more than one round of data collection to achieve balance.

There are two special cases for $P(sample|\mathbf{Z})$. The first one is that when the focal sample distribution approximates the benchmark population, $P_B(\mathbf{Z}) = P(\mathbf{Z})$, and equation (2.6) can be simplified as

$$P^{(k+1)}(sample|\mathbf{Z}) \propto 1 - \pi^{(k)} \times \frac{Q^{(k)}(\mathbf{Z})}{P_B(\mathbf{Z})} \tag{2.7}$$

That is, the sampling rate for the $(k+1)^{th}$ phase conditioning on $\mathbf{Z}$ depends on the ratio of the k^{th} phase cumulative respondent $\mathbf{Z}$ distribution and the benchmark $\mathbf{Z}$ distribution. The second special case is that if $P_B(\mathbf{Z}) \approx Q^{(k)}(\mathbf{Z})$ and $P_B(\mathbf{Z}) \neq P^{(k)}(\mathbf{Z})$, equation (2.6) can be simplified as

$$P^{(k+1)}(sample|\mathbf{Z}) \propto \frac{P_B(\mathbf{Z})}{P^{(k)}(\mathbf{Z})} \tag{2.8}$$

That is, if respondents reach representativeness but one wanted to continue growing the sample size with additional replicates, the (relative) sampling rate becomes $\frac{P_B(\mathbf{Z})}{P^{(k)}(\mathbf{Z})}$.

As we have mentioned in Section 2.2.3, a sample selected based on $P(sample|\mathbf{Z})$ will suffer nonresponse, hence nonresponse adjusted equation (2.6) can further improve the respondent representativeness when the nonresponse pattern remains similar across data collection phases. Let ϕ denote the response probability, nonresponse adjusted (2.6) can be written as

$$P^{(k+1)}sample|\mathbf{Z}) \propto \frac{1}{\phi}\{\frac{P_B(\mathbf{Z})}{P^{(k)}(\mathbf{Z})} - \pi^{(k)} \times \frac{Q^{(k)}(\mathbf{Z})}{P^{(k)}(\mathbf{Z})}\} \tag{2.9}$$

Similar nonresponse adjustment applies to equations (2.7) and (2.8).

2.2.5 Estimating the Sampling Rate with the Propensity Score

When $\mathbf{Z}$ is univariate, estimating $P_B(\mathbf{Z})/P^{(k)}(\mathbf{Z})$ and $Q^{(k)}(\mathbf{Z})/P^{(k)}(\mathbf{Z})$ is straightforward as illustrated in Section 2.2.3. However, in most situations $\mathbf{Z}$ is a vector consisting of many covariates (multivariate). For example, covariates listed in Table 2.6 are common $\mathbf{Z}$ in typical household surveys, with both continuous and categorical variables. The joint distribution of continuous and categorical variables is complex to specify and difficult to estimate. $\mathbf{Z}$ may also include both main effects (i.e. marginal distributions) and interactions (i.e. conditional distributions) which further complicate the estimation. One strategy to reduce the multivariate distribution of $\mathbf{Z}$ to a scalar is the propensity score method (Rosenbaum and Rubin, 1983; DAgostino, 1998).

The propensity score methods, proposed by Rosenbaum and Rubin (1983), model “treatment group” membership with baseline covariates. The treated and untreated subjects with the same propensity score have similar joint distributions of observed baseline covariates. In our application, we model the "benchmark membership" using covariates common to the benchmark and the focal survey. The propensity score of interest is the probability of benchmark survey membership conditional on covariates in the model.

The propensity score is defined as $e(\boldsymbol{z}) = P(T = 1|\boldsymbol{z})$, where T is an indicator variable, representing survey membership. $T = 1$ for benchmark survey and $T = 0$ otherwise. Auxiliary variables ($\boldsymbol{Z}$) common to both surveys are included in the propensity model. The propensity scores are estimated by the simple logistic model, $logit(e(\boldsymbol{z})) = \beta_0 + \boldsymbol{\beta}'\boldsymbol{z}$ for the merged data set of size $N_B{+}N_R$, where N_B denotes the benchmark sample size and N_R denote the cumulative focal survey respondent size up until phase k. The empirical density functions of the predicted propensity scores are then derived for both the benchmark survey cases and the focal survey respondents. The estimated propensity score density functions then serve as the basis for modifying the sampling rates. Therefore, if we consider $Q^{(k)}(\mathbf{Z})$ to be the propensity score density of the focal survey cumulative respondents up until phase k, and $P_B(\mathbf{Z})$ to be the propensity score density of the target population, then $P_B(\mathbf{Z})/P^{(k)}(\mathbf{Z})$ and $Q^{(k)}(\mathbf{Z})/P^{(k)}(\mathbf{Z})$ become the ratios of propensity score density which greatly simplified the computation of these density ratios.

2.2.6 Assessing the Representativeness of the Respondents

A representative sample can be regarded as a random sample from the underlying target distribution $f(\mathbf{Z}, \mathbf{X}, \mathbf{Y})$. A representative respondent pool should reproduce and preserve the variable associations and distributions in the target population as captured by the benchmark survey. Therefore, after each data collection, there is an interest to measure respondent representativeness compared to the benchmark.

The concept of "representativeness" is similar to the concept of a *balanced sample*. In the survey context, a *balanced sample* is defined as a sample satisfying the property of $\bar{\mathbf{U}} = \bar{\mathbf{u}}$ where $\mathbf{U}$ denotes a population quantity and $\mathbf{u}$ denotes the sample quantity (Royall and Herson, et. al., 1973; Hansen, et. al., 1983; Särndal, 2011). In this study, we follow the terminology used in the propensity score literature for observational studies, the "balance" implies the similarities between two density distributions. In other words, the balance here is defined as the equivalence of the benchmark survey and the focal survey with respect to the multivariate distribution of covariates, which is expressed as $P_B(\mathbf{Y}, \mathbf{X}, \mathbf{Z}) = P(\mathbf{Y}, \mathbf{X}, \mathbf{Z})$. A high degree of balance suggests that the sample composition of the two surveys is similar, conditioning on the covariates used in the propensity score models.

Several balance diagnostics methods have been proposed for the assessment of the adequacy of propensity score matching (Austin, 2009). These diagnostic methods were developed for use in observational studies to detect the distributional differences between treatment groups with regard to baseline covariates after the application of propensity score matching. Conveniently, these methods can also be adopted to assess the balance between a benchmark survey and the focal survey after each iteration of the adaptive sampling procedure. Specifically, we use the non-parametric density estimates as the balance diagnostic measure.

The non-parametric density estimates computed using the empirical propensity scores for the benchmark and the focal survey can be compared by the Hellinger's distance function. This distance quantifies the similarity between the two probability distributions. Therefore, a large Hellinger's distance in propensity score densities between the benchmark and the focal survey suggests the dissimilarity in the covariate distributions of the two surveys. Let $P_B(\mathbf{Z})$ and $P(\mathbf{Z})$ denote the empirical propensity score probability density functions for the benchmark and the focal survey, respectively, where $\mathbf{Z}$ denotes variables in the propensity score model. The Hellinger's distance function is written as

$$H^2 = \frac{1}{2} \int \left(\sqrt{P_B(\mathbf{Z})} - \sqrt{P(\mathbf{Z})} \right)^2 d\mathbf{Z} \tag{2.10}$$

2.3 Simulation Study

This study proposes adaptive sampling rates over multiple phases of replicate sample selection that are computed with $P(\mathbf{Z})$, $P_B(\mathbf{Z})$ and $Q(\mathbf{Z})$ to improve the representativeness of the $P_R(\mathbf{Z}, \mathbf{X})$, where $\mathbf{Z}$ represents variables available on the sampling frame, and $\mathbf{Z}$ is correlated with $\mathbf{X}$ and $\mathbf{Y}$. Specifically, we demonstrate the utility of the proposed sampling rate in a sequential approach to achieve a respondent pool that better resembles the benchmark with respect to the propensity score distribution.

Data from the representative respondents obtained via sampling that utilizes propensity scores preserves the relation between variables. We show that, when compared to the respondent data from a fixed sampling plan, the respondents from adaptive sampling achieve greater similarity to the benchmark data on 1) the marginal distributions of $P(X)$, $P(Y)$, $P(Z, X)$ and $P(Z, Y)$, 2) the correlation structure of (Z, X, Y), and 3) the joint distributions $P(Z, X, Y)$. The simulation experiment validates the marginal representativeness by reporting on the means and standard deviations of the individual variables, validates the preservation of correlation structure by reporting on the variance and covariance estimates, and validates the joint distribution $P(Z, X, Y)$ by graphing the corresponding empirical propensity score densities and reporting on the Hellinger's distance between the densities.

2.3.1 Simulation Setup

Let (Z, X, Y) be multivariate normally distributed variables with the joint distribution

$$(Z, X, Y) \sim N_3(\mathbf{0}, \Sigma), \qquad \text{and} \tag{2.11}$$

$$\Sigma = \begin{pmatrix} 1 & cov(Z,X) & cov(Z,Y) \\ cov(Z,X) & 1 & 0.707 \\ cov(Z,Y) & 0.707 & 1 \end{pmatrix}$$

where $cov(Z,X)$ and $cov(Z,Y)$ are varied in this simulation study to examine the effects of Z with different explanatory power on X and on Y. From the data model of (2.11) we simulate a finite population of size $N_{POP} = 500,000$. From this finite population, a simple random sample without replacement (srswor) of size $N_B = 25,000$ is drawn to serve as the benchmark survey and the remaining $475,000$ units serve as the sampling frame for the focal surveys. From this sampling frame, two concurrent focal surveys are simulated, one using the adaptive sampling design and the other a fixed sampling design. For both focal surveys, the first phase is a srswor of size $1,000$ ($n^{(1)} = 1,000$). The two focal surveys share the same first phase sample and respondents. The two surveys start their corresponding sampling design from the second phase.

The first phase sampling strategy, a srswor of size $1,000$, is repeated for the second through fourth phases for the fixed sampling design. For the adaptive sampling design, sample allocation for each phase uses equation (2.9) and the phase samples are of size $n^{(k)} = 1,200$ for the k^{th} data collection phase, where $k = 2, \ldots, 4$.

The simulation study can be described as a 2 × 2 × 4 factorial design. The factors are:

Factor A: $cov(Z,X)$: high vs low

Factor B: $cov(Z,Y)$: high vs low

Factor C: Four nonresponse mechanisms based on the response models described below.

The $cov(Z,X)$ equals the correlation of (Z,X), denoted as $\rho_{Z,X}$, since $var(Z)$ and $var(X)$ are set to be one. Among factors A and B, a high correlation of two variables is set to be $\rho = 0.894$ (i.e. $R^2 = 0.8$) and a low correlation is $\rho = 0.447$ (i.e. $R^2 = 0.2$, where R^2 is the coefficient of determination in linear regression). The correlation between X and Y is held to be constant across simulations with $\rho_{X,Y} = 0.707$ (i.e. $R^2(X,Y) = 0.5$). The combination scenario of a low correlation in Factor A and a high correlation in Factor B is labeled “LH” herein. Similarly the other three scenarios of Factors A and B are labeled “LL”, “HL” and “HH”.

For Factor C, the nonresponse mechanisms, we simulate response status $\Re$ by drawing from Bernoulli random variables with $e(z,x,y) = Pr(\Re = 1|z,x,y)$ computed from four different response models:

1. $logit(e(z,x,y)) = Bernoulli(0.5)$. **(MCAR)**

2. $logit(e(z,x,y)) = 0.0002 + 0.8z$. **(MAR)**

3. $logit(e(z,x,y)) = 0.00001 + 0.31z + 0.61x$. **(MNAR$_X$)**

4. $logit(e(z,x,y)) = 0.00004 + 0.19z + 0.38x + 0.38y$. **(MNAR$_Y$)**

where $\Re = 1$ if subject responded, 0 otherwise. Each of these models generate an average of 50% response rate. For model 1, the response probability is independent of (Y,Z,X). For model 2, the response probability depends on Z alone. For model 3 the response probability depends on both Z and X. For model 4, the response probability depends on (Y,Z,X).

With the finite population, the procedure of a) drawing a benchmark, b) forming a sampling frame, and c) simulating two concurrent focal surveys, is carried out 100 times (trials) for each of the 16 simulation scenarios. For each trial, the following quantities are computed: (a) the means and standard deviations of Y, X, and Z, (b) estimated variance-covariance matrix among cumulative respondents, (c) propensity score density obtained by three models: model of Z alone, model with (X,Z), and model with (Y,X,Z), and (d) Hellinger’s distance on propensity score densities estimated for Benchmark and the focal survey. For each simulation scenario, we report the average of quantities (a), (b), and (d) from the 100 trials. For (c), the propensity score density plots from one of the 100 simulations are presented.

2.3.2 Simulation Results

The average response rates for each simulation scenario is shown in Table 2.1. The underlying population has a response rate of 50%. When a representative sample is drawn, we expect to see the average response rate of 50%, which is the case for the fixed design (F in Table 2.1). For adaptive sampling, we over sample the under-represented subjects (who have lower response propensity) and under sample the over-represented subjects (who have higher response propensity), the response rate is expected to be lower than 50% (A in Table 2.1). As we see in Table 2.1, the adaptive design has response rates around 40%.

For the representativeness on marginal distributions, Figure 2.4(c) shows the boxplots of the means for individual variables under various scenarios for the MNAR$_Y$ nonresponse model, where columns titled 'LL', 'LH', 'HL' and 'HH' represent the strength of correlation between (Z, X) and (Z, Y), respectively. The x-axis of each boxplot shows the four phases and the y-axis shows the mean differences from the benchmark. A horizontal zero-line aids the visual comparison to the desired representativeness, a zero difference from the benchmark. The boxplots summarize means of 100 trials from the adaptive sampling design and the red-dots show the average of the 100 means from fixed sampling design. For each simulation scenario summarized in the figure, the adaptive design brings the respondent pool closer to the benchmark than that of the fixed design, indicating that the respondent representativeness improves at each successive sampling phase. Similar results are found for MAR and MNAR$_X$ nonresponse models (see Figures 2.4(a) and 2.4(b)).

To show the preservation of correlation structure among representative respondents, the results for estimated parameters in the variance-covariance matrix for adaptive sampling and fixed sampling are shown in Figure 2.5. Panels of the figure show nonresponse models, and columns show simulation scenarios in Factor A and Factor B. For each figure, the x-axis indicates the parameters where 1 = variance of Z, 2 = variance of X, 3 = variance of Y, 4 = covariance of (Z,X), 5 = covariance of (Z,Y), and 6 = covariance of (X,Y). The y-axis represent the departure of estimated parameter values from the benchmark values. A horizontal zero line aids the visual comparison.

Table 2.1: Multivariate normal data. Average response rates from 100 simulations.

Factor A B	NR model	Design[1]	Phase 1	Phase 2	Phase 3	Phase 4
LL	MAR	A	50.0	40.9	41.8	42.5
		F	50.0	49.8	49.9	50.1
	MNAR$_X$	A	50.1	41.5	42.5	43.1
		F	50.1	49.8	50.1	50.0
	MNAR$_Y$	A	50.0	42.0	42.7	43.0
		F	50.0	50.0	50.0	50.0
LH	MAR	A	49.9	41.2	42.0	42.6
		F	49.9	50.1	50.4	50.1
	MNAR$_X$	A	50.0	41.5	42.6	43.0
		F	50.0	50.2	50.0	50.2
	MNAR$_Y$	A	49.9	41.4	42.0	42.6
		F	49.9	50.0	49.9	50.0
HL	MAR	A	49.9	40.6	41.9	42.5
		F	49.9	50.1	50.4	49.9
	MNAR$_X$	A	50.2	41.3	41.8	42.4
		F	50.2	50.0	50.0	50.1
	MNAR$_Y$	A	50.2	41.6	42.1	42.7
		F	50.2	50.0	49.9	50.0
HH	MAR	A	49.8	41.2	42.0	42.5
		F	49.8	50.1	49.9	50.0
	MNAR$_X$	A	50.3	41.2	41.9	42.3
		F	50.3	50.1	50.1	49.9
	MNAR$_Y$	A	49.7	41.0	42.1	42.3
		F	49.7	50.0	50.0	50.2

[1] F: Fixed design; A: Adaptive design.

The boxplots summarize values from the adaptive sampling approach and red-dots are average values from fixed sampling.

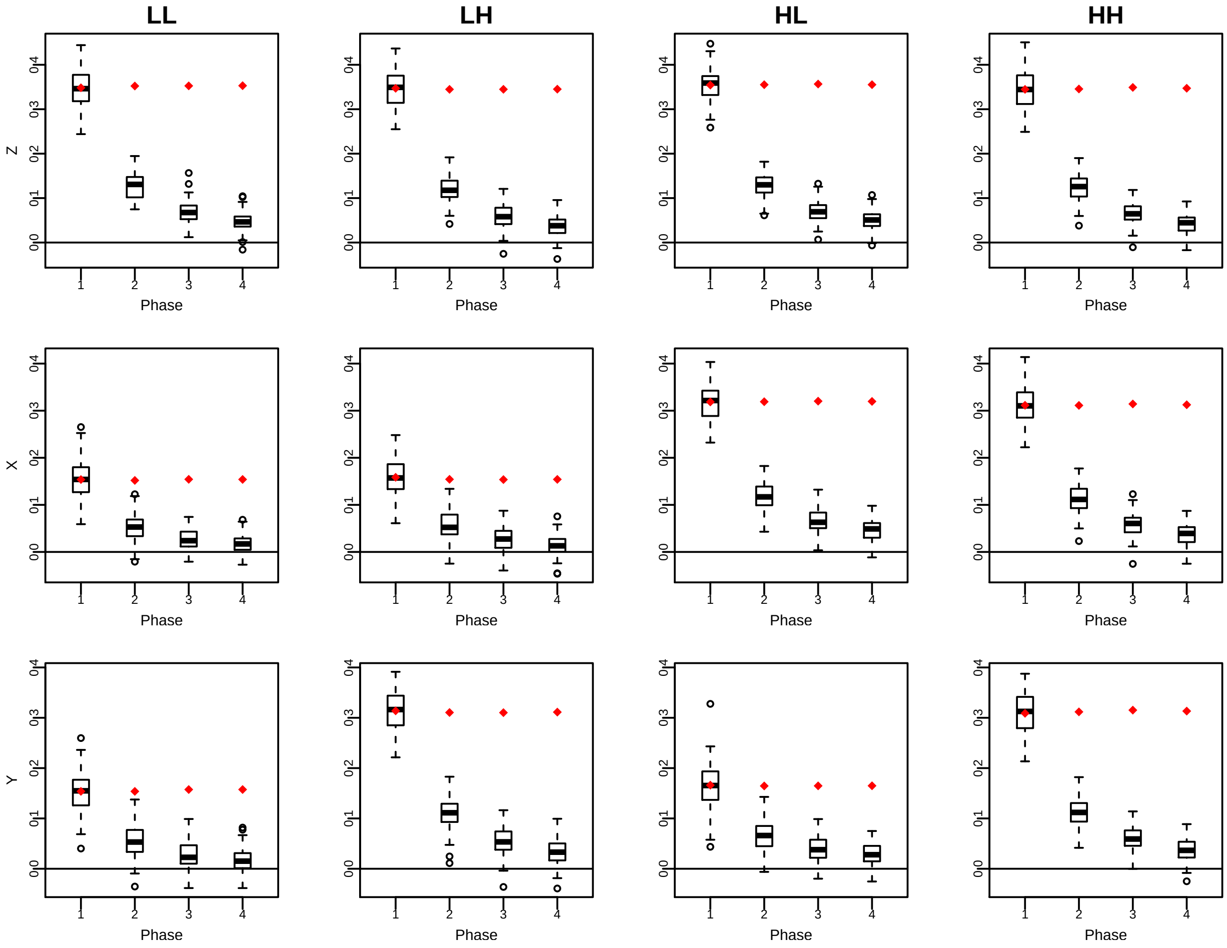
From Figure 2.5 we see that, for all simulation scenarios, the red dots (the mean of parameter estimates from the fixed sampling) fall either on the margin or outside of the boxplots, away from the zero line. Adaptive sampling estimates (the boxplots) are more similar to those of the benchmark, reflecting a respondent pool that is more representative compared to that of the fixed sampling design. In summary, the respondent pool from the adaptive sampling maintains a correlation structure among (Z, X, Y) that is closer to the correlation structure of the benchmark than the respondent pool from the fixed sampling design.

To show the preservation on joint distribution, Figure 2.6 illustrates the empirical density of propensity score distribution from benchmark membership model that includes Z, X, and Y as covariates. Figures 2.6 (a), (b), (c), and (d) show the four simulation scenarios of LL, LH, HL, and HH, respectively. Within each figure, e.g., Figure 2.6(a), the row panels show various nonresponse models while the column panels show the sampling phases. For example, the first row panel in Figure 2.6(a) illustrates that the propensity score distribution of the benchmark and the adaptive sampling focal survey converges with each sampling phase, indicating that the adaptive sampling adjustments successfully improve the joint distribution of respondent representativeness in the focal survey after each sampling phase. Similar results are seen in Figures 2.6(b), (c), and (d).

To quantify the respondent representativeness seen in Figure 2.6, we compute the *Hellinger*'s distance. As expected, the improvement in global representativeness can be seen in the decreasing values of the *Hellinger*'s distance (H) (0.162, 0.062, 0.033, and 0.020) following each of the four phases of adaptive sampling. Table 2.2 listed average Hellinger's distance from 100 simulaton for the benchmark membership propensity score model of (Z, X, Y). The numbers shown are the Hellinger's distance between $P_B(Y, X, Z)$ and $P_A(Y, X, Z)$ and Hellinger's distance between $P_B(Y, X, Z)$ and $P_F(Y, X, Z)$. In all simulaton scenarios, the propensity score density (modeling benchmark membership) from adaptive sampling design converges to that of the Benchmark, achieving the goal of $P_B(Y, X, Z) \sim P_A(Y, X, Z)$, as the sampling phases progress. Other (benchmark membership) propensity score models, that is, $P_B(X, Z)$ and $P_A(X, Z)$ and $P_B(Z)$ and $P_A(Z)$, show similar results (data not shown).

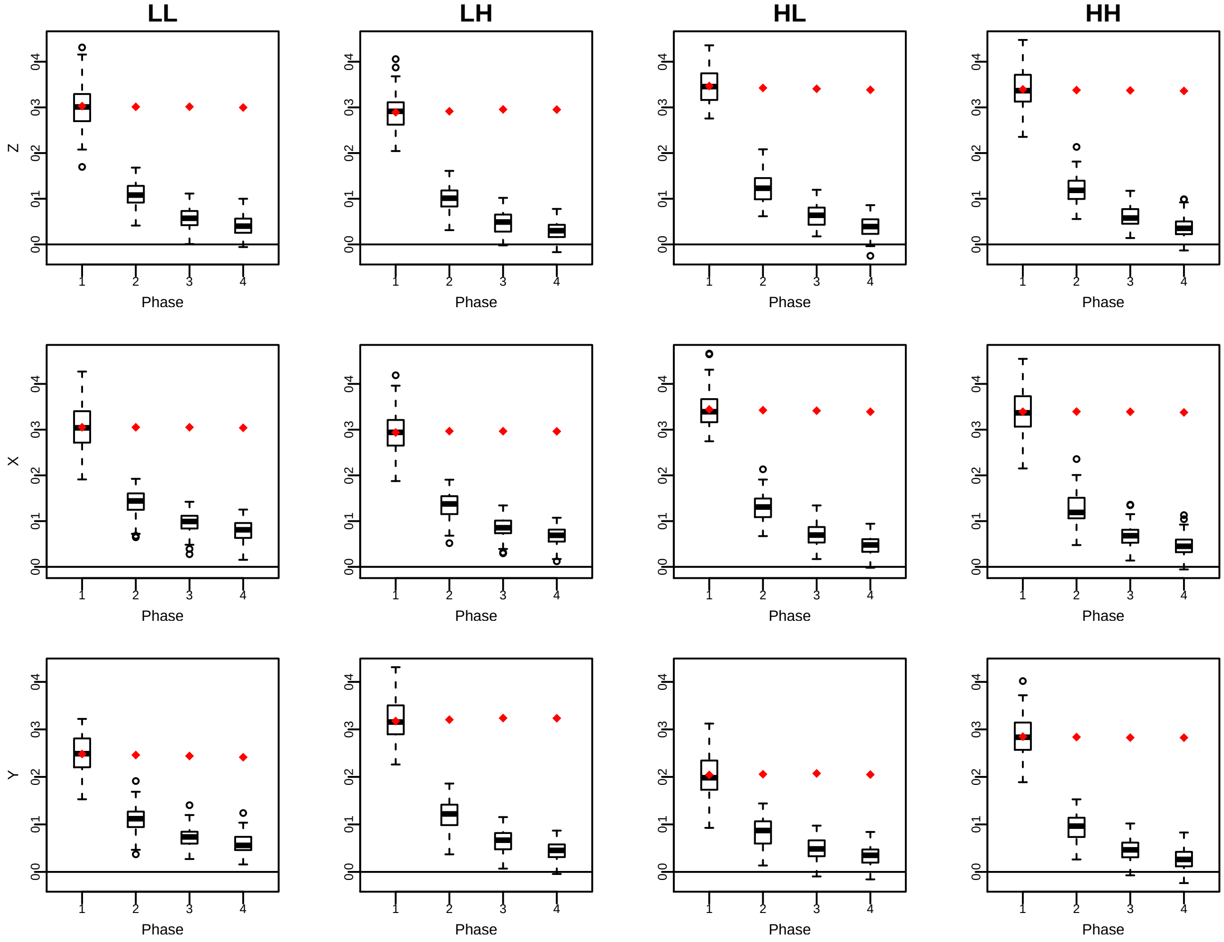
Figure 2.4: Multivariate normal data. Boxplots on means of Z, X, and Y from adaptive sampling.

(a) MAR nonresponse model



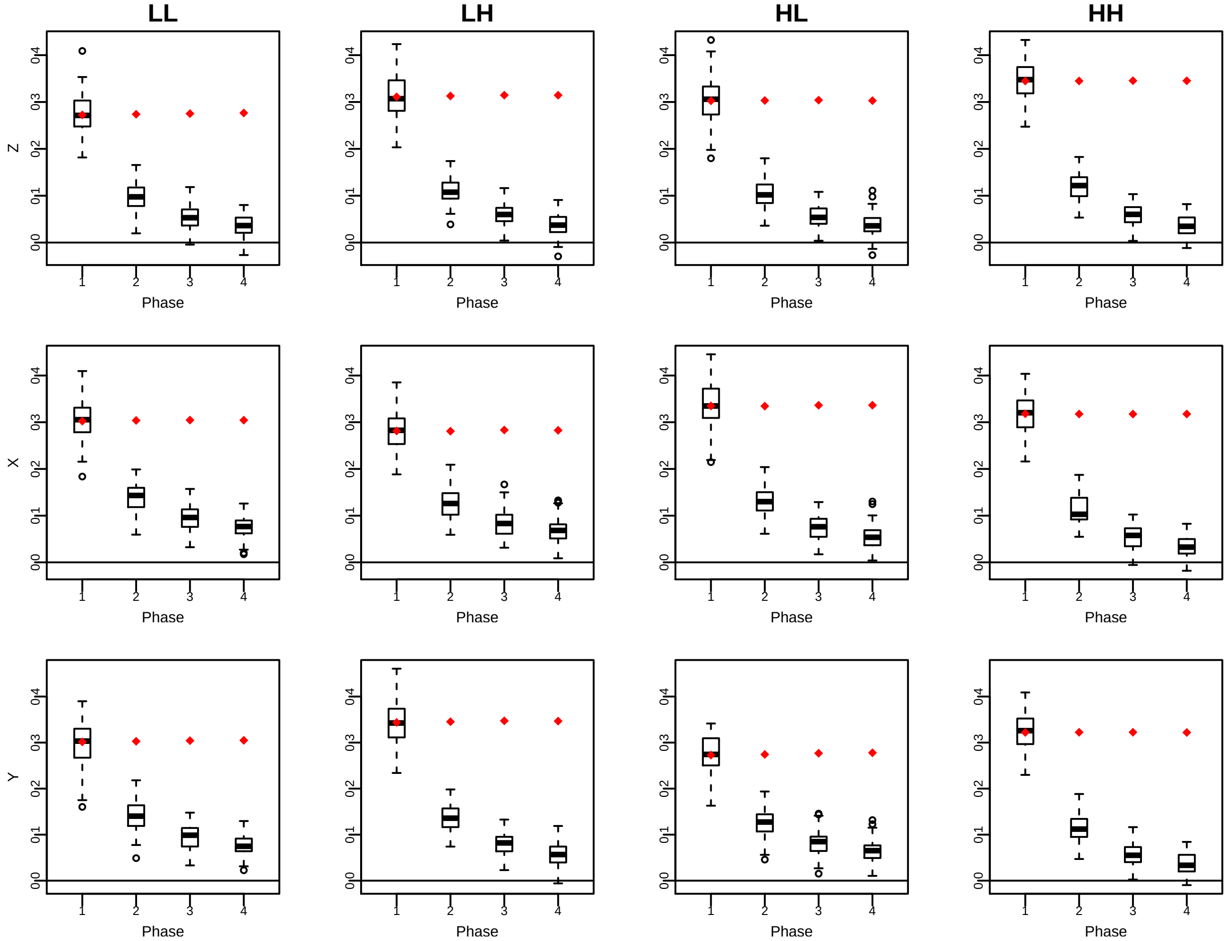
The y-axis represents bias of the estimated mean and the x-axis represents sampling phases. A horizontal zero line aids the visual reference for the location of an unbiased estimate. Boxplots are summary of 100 simulation runs for adaptive sampling design and the red dots show the average of 100 mean values from the fixed sampling design.

(b) MNAR$_X$ nonresponse model



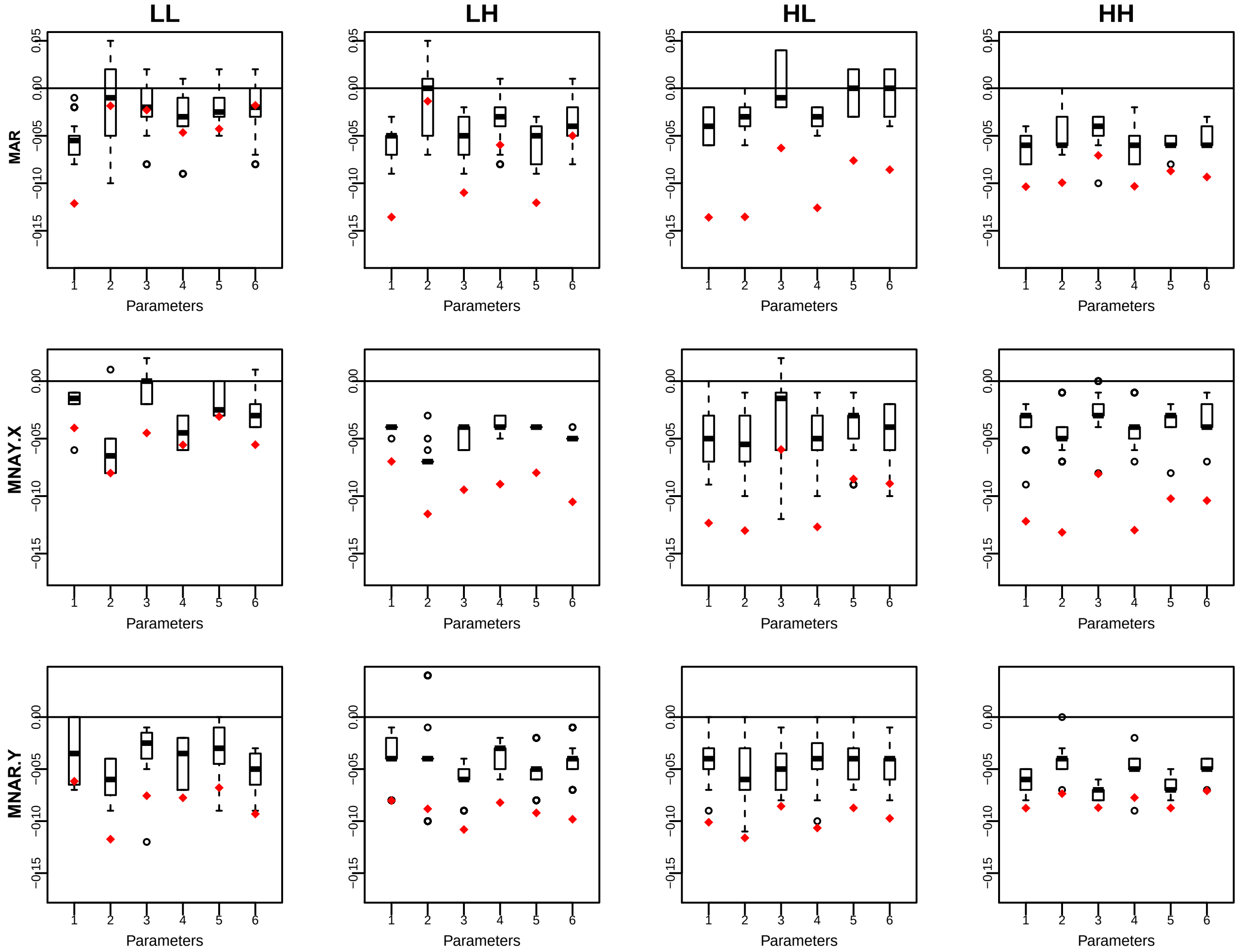
The y-axis represents bias of the estimated mean and the x-axis represents sampling phases. A horizontal zero line aids the visual reference for the location of an unbiased estimate. Boxplots are summary of 100 simulation runs for adaptive sampling design and the red dots show the average of 100 mean values from the fixed sampling design.

(c) MNAR$_Y$ nonresponse model.



The y-axis represents bias of the estimated mean and the x-axis represents sampling phases. A horizontal zero line aids the visual reference for the location of an unbiased estimate. Boxplots are summary of 100 simulation runs for adaptive sampling design and the red dots show the average of 100 mean values from the fixed sampling design.

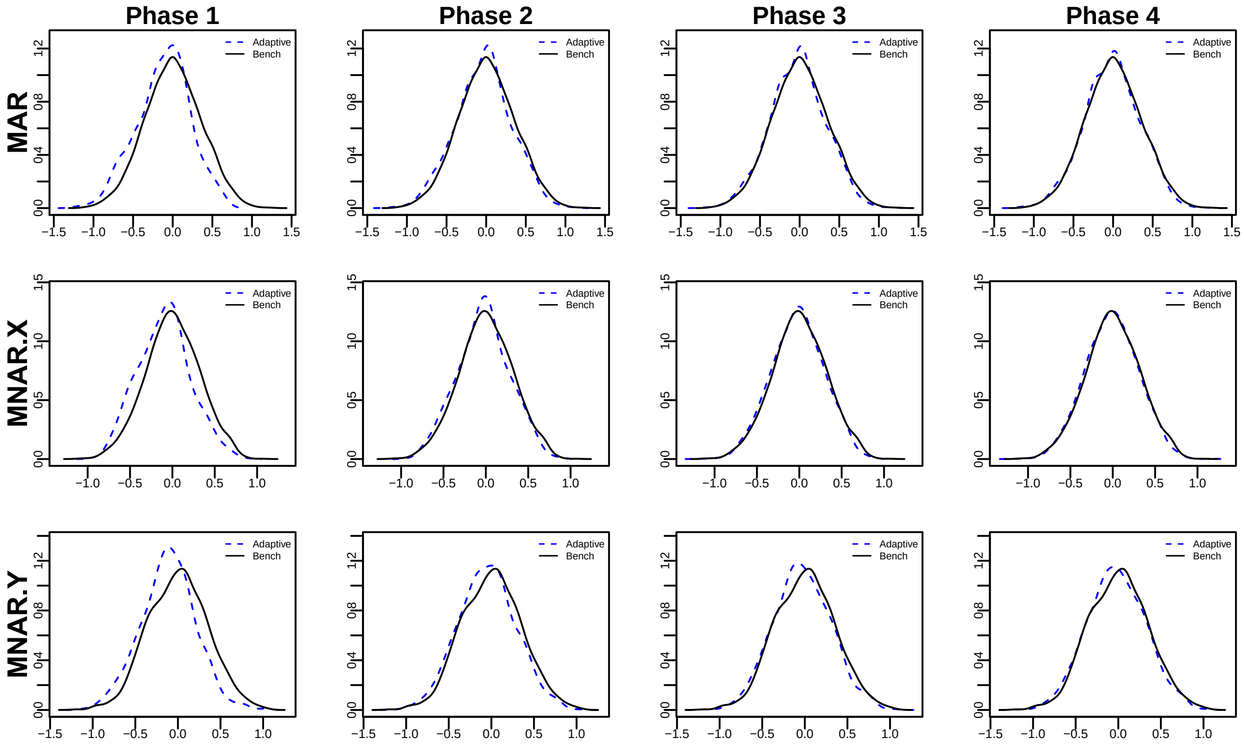
Figure 2.5: Multivariate normal data. Boxplots on components of variance-covariance matrix for (Z,X,Y) for adaptive sampling for cumulative respondents.



Red dots show point estimates for fixed design for cumulative respondents. Panels show nonresponse models. Columns show simulation scenarios in Factor A and Factor B. For each figure, x-axis indicates the parameters: 1 = variance of Z, 2 = variance of X, 3 = variance of Y, 4 = covariance of (Z,X), 5 = covariance of (Z,Y), 6 = covariance of (X,Y). y-axis represent the departure of estimated parameter values from the benchmark values. A horizontal zero line aids the visual comparison. boxplots are summary of 1000 simulation values from adaptive sampling and red-dots are average value from fixed sampling.

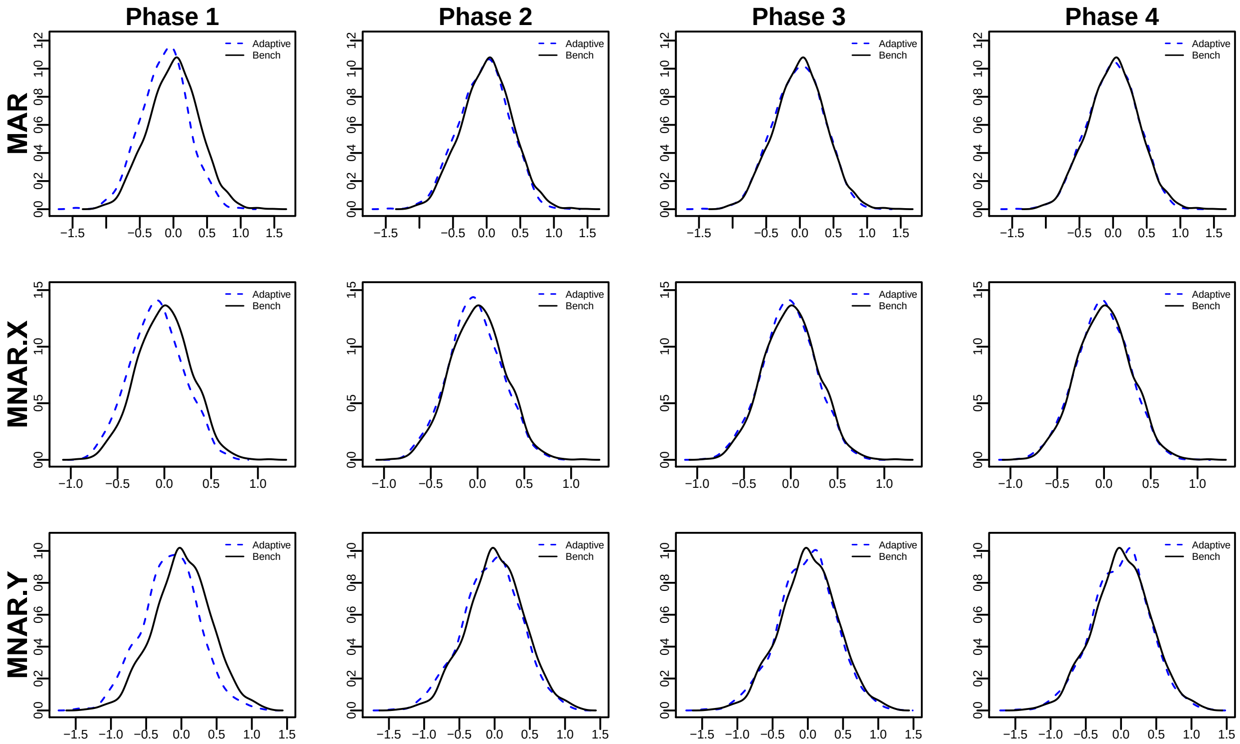
Figure 2.6: Multivariate normal data. Propensity score density plots on benchmark membership propensity score model with variables (Y, X, Z).

(a) Scenario LL



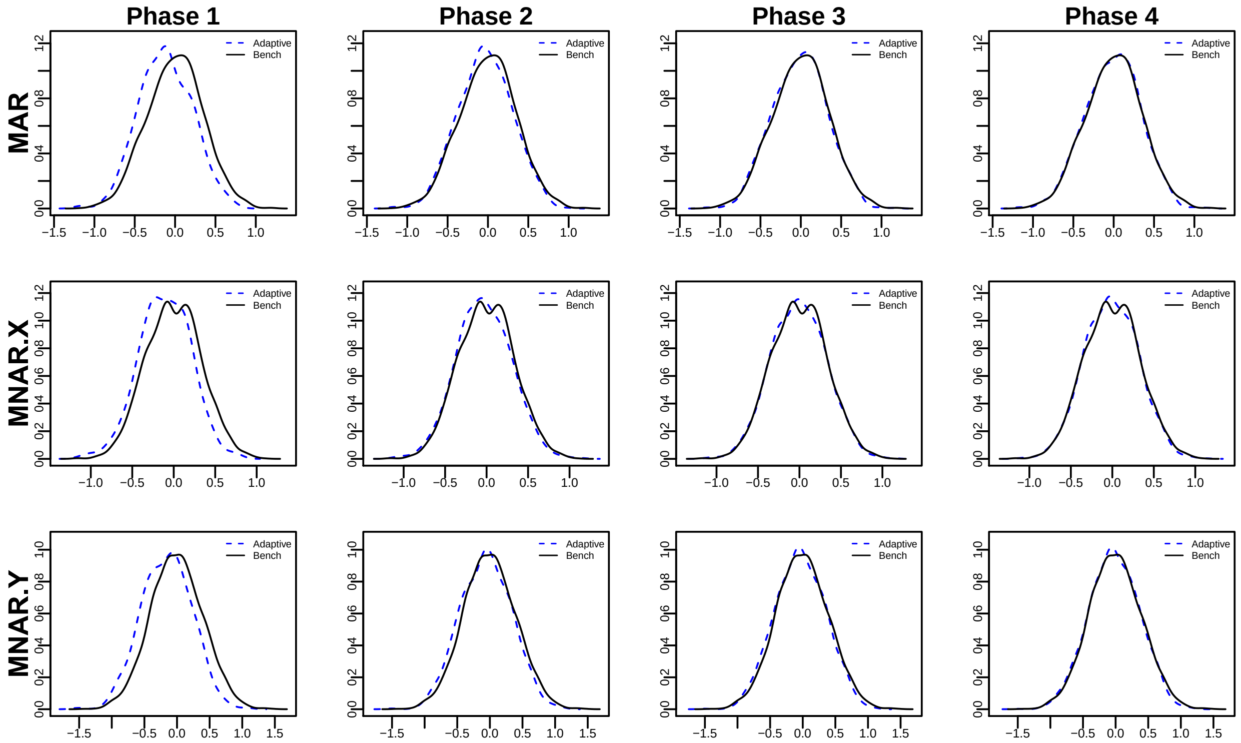
Each row panel shows different nonresponse models and each column panel shows sampling phases. x-axis represents propensity scores in logit scale.

(b) Scenario LH



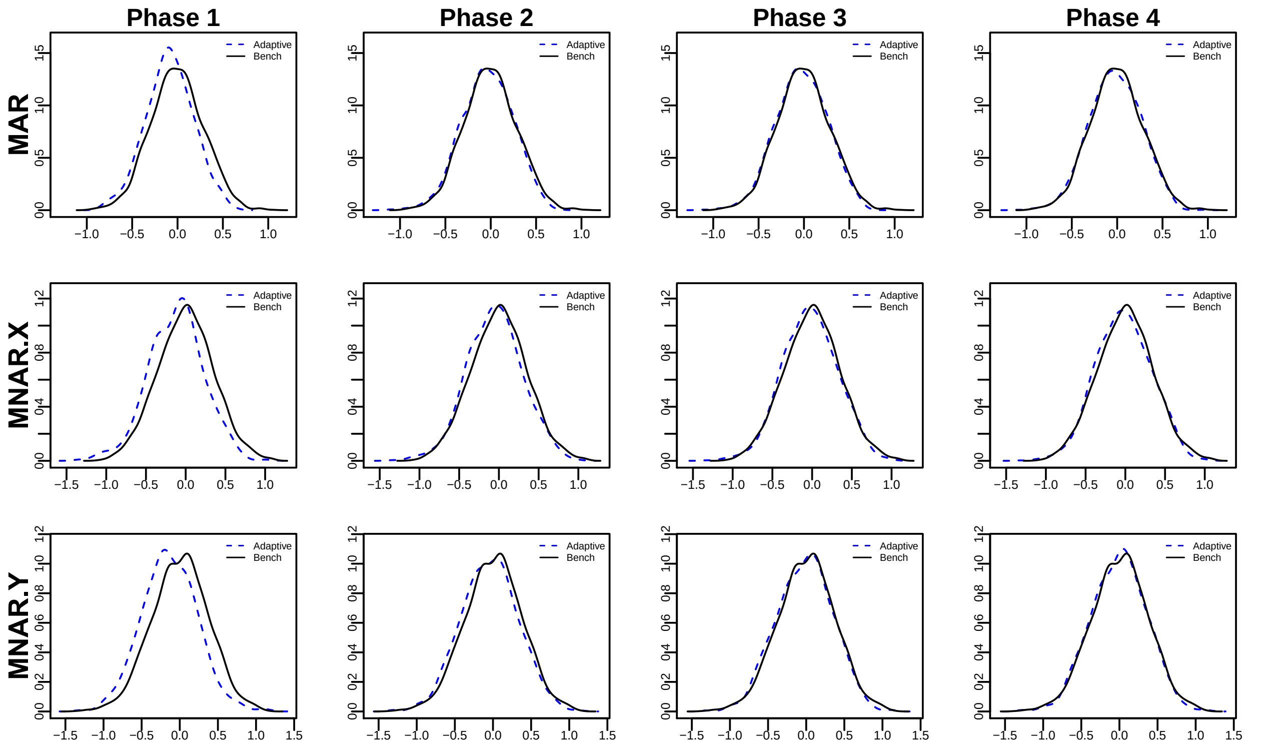
Each row panel shows different nonresponse models and each column panel shows sampling phases. x-axis represents propensity scores in logit scale.

(c) Scenario HL



Each row panel shows different nonresponse models and each column panel shows sampling phases. x-axis represents propensity scores in logit scale.

(d) Scenario HH



Each row panel shows different nonresponse models and each column panel shows sampling phases. x-axis represents propensity scores in logit scale.

2.4 Illustration with NHIS and BRFSS Data

We now demonstrate the proposed approach using data from two large scale surveys, NHIS and BRFSS. The NHIS is a face-to-face cross-sectional survey that monitors trends in illness and disability of the civilian, non-institutionalized, household population of the United States. The BRFSS is an on-going telephone health survey where data are collected monthly by each of 50 States and the District of Columbia. Both the NHIS and the BRFSS are multi-purpose health surveys that share many health related questions in common. However, researchers have reported discrepancies on survey estimates between BRFSS and NHIS, partially as a result of data collection modes (Fahimi, et. al., 2008; Nelson, et. al., 2003).

Table 2.2: Multivariate normal data. Average Hellinger's distance from 100 simulations.

Factor A B	NR model	Design[1]	Phase 1	Phase 2	Phase 3	Phase 4
LL	MAR	A	19.2	3.7	1.9	1.5
		F	19.2	18.2	17.9	17.5
	MNAR_X	A	18.9	4.5	2.5	2.0
		F	18.9	17.9	17.8	17.6
	MNAR_Y	A	19.7	4.9	2.8	2.3
		F	19.7	18.5	18.3	18.0
LH	MAR	A	19.1	3.5	1.8	1.4
		F	19.1	18.5	18.2	17.9
	MNAR_X	A	19.0	4.3	2.5	1.9
		F	19.0	17.8	17.7	17.4
	MNAR_Y	A	20.0	4.6	2.5	1.9
		F	20.0	18.9	18.4	18.1
HL	MAR	A	19.7	3.8	2.0	1.5
		F	19.7	18.5	18.1	17.9
	MNAR_X	A	19.2	3.9	2.1	1.6
		F	19.2	18.3	18.0	17.7
	MNAR_Y	A	19.3	4.4	2.6	2.0
		F	19.3	18.5	18.3	18.1
HH	MAR	A	19.3	3.6	1.8	1.4
		F	19.3	18.4	18.0	17.7
	MNAR_X	A	18.7	3.7	1.9	1.5
		F	18.7	17.8	17.6	17.4
	MNAR_Y	A	19.0	3.8	2.0	1.6
		F	19.0	18.0	17.9	17.7

Note: Values are multiplied by 1000.

[1] F: Fixed design; A: Adaptive design.

A more costly face-to-face survey mode is found to produce better data quality by some researchers(Jordan, et. al., 1980, Locander and Burton, 1976). The telephone mode BRFSS bears the problem of informative nonresponse that threaten the validity of survey estimates. Thus, the BRFSS could potentially benefit from an adaptive sampling design using NHIS as a benchmark to improve respondent representativeness. This would not only reduce the bias and variance of the survey estimates, but it can also be more cost-efficient in the sense that using a smaller face-to-face survey (e.g NHIS) to improve the respondent representativeness of a much larger telephone survey (e.g BRFSS).

The NHIS and BRFSS are one of the main motivating cases for the proposed adaptive sampling design. This is because many large government surveys like BRFSS are collected and processed in waves during the year and many government surveys use the same or similar sampling frame. With temporal phases of survey data collection, e.g., quarterly, it is natural to think about intervening in the sampling design over the annual data collection period. The simulation findings from the previous section provide evidence of promising results. We now investigate the utility of proposed method in real survey data with complex and high dimensional structure.

For illustration, NHIS 2009 and BRFSS 2009 publicly available micro-data files were combined to form a finite target population. Note, in practice one would use 2008 NHIS as the Benchmark to implement the adaptive design for the 2009 BRFSS survey. Subjects from the two surveys are assumed to be mutually exclusive. The data included subjects residing in U.S. 50 states and District of Columbia (D.C.). Subjects who are less than 18 years old and subjects with item missingness on age, sex, and race were excluded. The target population consists of $N_{POP} = 445,965$ subjects.

As in the previous simulation we create the response indicator $\Re$ for the focal surveys by drawing from Bernoulli random variables derived under four nonresponse models: 1) $\Re \sim Bernoulli(0.5)$, 2) $\Re \sim (\mathbf{Z})$, 3) $\Re \sim (\mathbf{Z}, \mathbf{X})$, and 4) $\Re \sim (\mathbf{Z}, \mathbf{X}, \mathbf{Y})$. The details of four nonresponse models of the focal surveys are listed below:

1. $logit(e(\mathbf{z}, \mathbf{x}, \mathbf{y})) = Bernoulli(0.5)$.

2. $logit(e(\mathbf{z}, \mathbf{x}, \mathbf{y})) = 0.217 + 0.7z_1 + 0.7z_2 + 0.7z_3 + 0.7z_4 + 0.7z_5$.

3. $logit(e(\mathbf{z}, \mathbf{x}, \mathbf{y})) = 0.173 + 0.68z_1 + 0.68z_2 + 0.68z_3 + 0.68z_4 + 0.68z_5 - 0.34x_1 - 0.34x_2$.

4. $logit(e(\mathbf{z}, \mathbf{x}, \mathbf{y})) = 0.147 + 0.57z_1 + 0.57z_2 + 0.57z_3 + 0.57z_4 + 0.57z_5 - 0.29x_1 - 0.29x_2 + 0.29y_1 + 0.29y_2 + 0.29y_3 + 0.29y_4 + 0.29y_5 + 0.29y_6$.

where $\mathbf{Z}$ variables are Z_1 geographic region, Z_2 age, Z_3 sex, Z_4 hispanic origin, and Z_5 race; $\mathbf{X}$ variables are X_1 marital status and X_2 education; $\mathbf{Y}$ variables are Y_1 heart attack, Y_2 stroke, Y_3 hypertension, Y_4 diabetes, Y_5 health insurance coverage, and Y_6 general health. The $\mathbf{Z}$ variables are chosen because they are typical of many household and social surveys and are often used in the weighting procedures. Each of the four nonresponse models have an average response rate of 50%. However, note that models 3 and 4 are examples of nonignorable nonresponse. The nonresponse models are assumed to be stable across sampling phases. This is a reasonable assumption since this is what was observed in BRFSS (recall Figure 2.1 in Section 2.1). In addition, the correlation structure of $(\mathbf{Z}, \mathbf{X}, \mathbf{Y})$ variables from the NHIS and the BRFSS is assumed to be similar to that encountered in surveys with a similar target population.

From the constructed finite population we draw a srswor benchmark of size $N_B = 22,299$ (i.e. 5% of the N_{POP}) and the remaining $423,666$ subjects serves as the sampling frame. We construct in parallel two simulated focal surveys, one with the adaptive sampling design and the other with the fixed sampling design. The two focal surveys share the same phase I data which is a srswor of size $1,000$. The focal survey with fixed sampling design repeats the same srswor sample of size $1,000$ for phases 2, 3, and 4 whereas the focal survey with adaptive sampling starts from phase 2 using the adaptive sampling rate as described in equation (2.9) with sample size of $n^{(k)} = 1,200$, where $k = 2, \ldots, 4$.

The procedure of drawing a benchmark and sampling four phases of two concurrent focal surveys are repeated 100 times for each nonresponse models. Our adaptive sampling procedure will be judged as effective if the resemblance of the $P_R(\mathbf{Z})$ and $P_B(\mathbf{Z})$ imply $P_R(\mathbf{Z}, \mathbf{X}) \approx P_B(\mathbf{Z}, \mathbf{X})$ and $P_R(\mathbf{Z}, \mathbf{X}, \mathbf{Y}) \approx P_B(\mathbf{Z}, \mathbf{X}, \mathbf{Y})$ in the focal survey with adaptive sampling design. We report the average of 100 runs on the marginal distributions of each variables, the estimates of covariance structure, and the joint distributions of $(\mathbf{Z}, \mathbf{X}, \mathbf{Y})$.

2.4.1 Application Results

Table 2.6 shows the convergence of the marginal distribution on individual $\mathbf{Z}$, $\mathbf{X}$, and $\mathbf{Y}$ variables for nonresponse model 4. The first column shows the population values, the second column shows the average of benchmark values from 100 trials, the third column shows the average of the mean values from phase 1 respondents, the fourth to sixth columns show the mean of adaptive design respondents from phase 2 to phase 4 (each phase includes respondents from previous phases), and the last two columns show the differences between benchmark and accumulated focal survey respondent data from adaptive design and from fixed design, respectively.

The table shows that, as expected, the average benchmark values (column 2) are similar to those of the population values (column 1). As each phase of data accumulated, we see an overall progressive improvement for the adaptive sampling design on the distributions of variables with respect to the benchmark. The last columns of Table 2.6 lists differences of focal surveys from the benchmark for the adaptive design and fixed design. As expected, the total absolute differences are larger for the fixed design than for the adaptive design for all variables. These results indicate that the covariate distributions from adaptive design resemble more closely to those of the benchmark than those of the fixed design. These results also indicate that, when $\mathbf{Z}$ is associated with $\mathbf{X}$ and $\mathbf{Y}$, improving respondent representativeness with respect to frame and auxiliary variables ($\mathbf{Z}$) also improves the representativeness of the survey covariates ($\mathbf{X}$) and survey outcome ($\mathbf{Y}$) variables. Similar findings are seen for nonresponse models 2 and 3, as shown in Tables 2.4 and 2.5, repsectively. Table 2.3 illustrates the results from nonresponse model 1. As expected, the average values of all variables from each phase remain similar across phases, reflecting a missing completely at random nonresponse scenario.

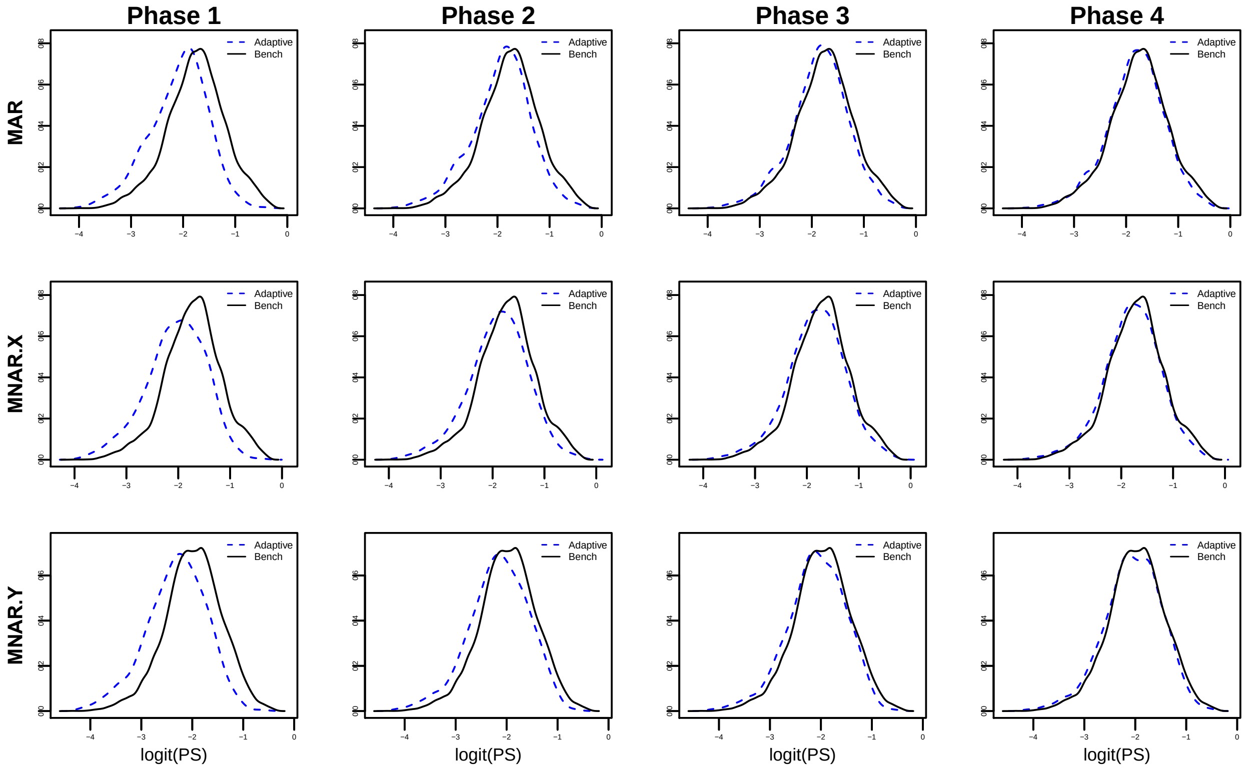
In Figure 2.7 we compare the respondent representativeness of fixed and adaptive sampling design in a multivariate fashion, via estimated propensity score densities from benchmark membership propensity score model, in short, "propensity score model" unless otherwise noted. The figure consists of four columns representing each of the four sampling replicates, and three panels representing three nonresponse models. Figure 2.7(a) shows the adaptive sampling propensity score model of the $\mathbf{Z}$ variables, Figure 2.7(b) shows the model of the $\mathbf{X}$ and $\mathbf{Z}$ variables, and Figure 2.7(c) shows the model of the $\mathbf{Y}$, $\mathbf{X}$ and $\mathbf{Z}$. The goal is to show that the $P(\mathbf{Z_R}) \sim P(\mathbf{Z_B})$ facilitates $P(\mathbf{X_R}, \mathbf{Z_R}) \sim P(\mathbf{X_B}, \mathbf{Z_B})$, which in turn facilitates $P(\mathbf{Y_R}, \mathbf{X_R}, \mathbf{Z_R}) \sim P(\mathbf{Y_B}, \mathbf{X_B}, \mathbf{Z_B})$.

The figure shows the results from nonresponse models 2,3 and 4, which clearly illustrates the convergence of propensity score density from adaptive sampling to that of the benchmark as the sampling phases progress whereas the fixed sampling distribution maintains similar shapes across phases. The patterns of convergence in adaptive sampling design are similar across different propensity score models representing the convergence of $P(\mathbf{Z})$, of $P(\mathbf{X}, \mathbf{Z})$, and of $P(\mathbf{Y}, \mathbf{X}, \mathbf{Z})$, respectively. Hellinger's distance decreases from 0.059 at phase 1 to 0.004 in phase 4 for adaptive design $(\mathbf{Y}, \mathbf{X}, \mathbf{Z})$ nonresponse model 4. Other nonresponse models have similar levels of decreasing values on Hellinger's distance. Table 2.7 shows the average Hellinger's distance from 100 simulation runs for all four nonresponse models. The Hellinger's distance for fixed design remains at the 0.059 to 0.046 level across phases reflecting the lack of improvement on respondent representativeness.

The findings from the univariate representativeness assessment in Table 2.6 echo that of the multivariate in Figure 2.7. That is, the overall representativeness assessment in Figure 2.7 shows that the propensity score density distribution from the adaptive sampling design converges to that of the benchmark at each sequence of data collection, regardless of nonresponse scenarios. The convergence in $P(\mathbf{Z_R}) \sim P(\mathbf{Z_B})$ assists in $P(\mathbf{X_R}, \mathbf{Z_R}) \sim P(\mathbf{X_B}, \mathbf{Z_B})$, which in term improves $P(\mathbf{Y_R}, \mathbf{X_R}, \mathbf{Z_R}) \sim P(\mathbf{Y_B}, \mathbf{X_B}, \mathbf{Z_B})$. The results in Table 2.6 demonstrate that the representativeness as measured by the propensity score density implies the representativeness in individual components in the propensity score model.

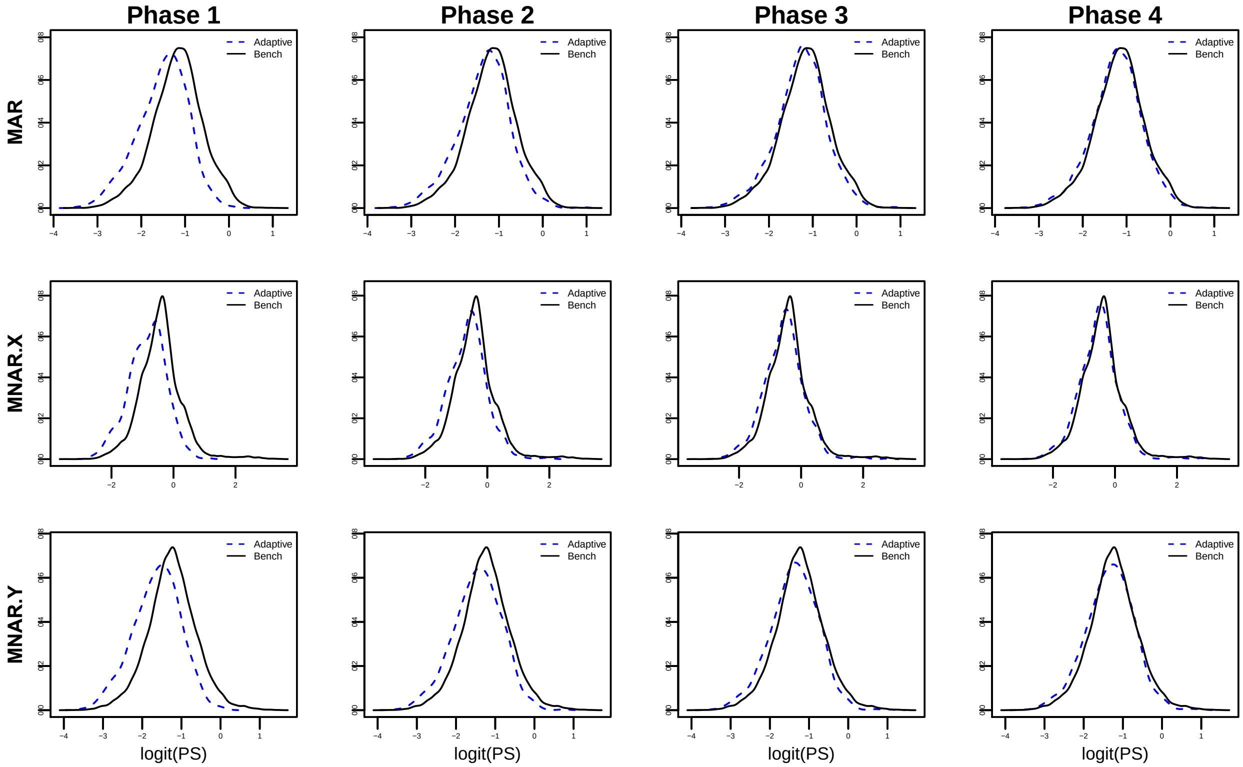
Figure 2.7: NHIS and BRFSS data. Propensity score density plots.

(a) Propensity score density $P(\mathbf{Z})$.



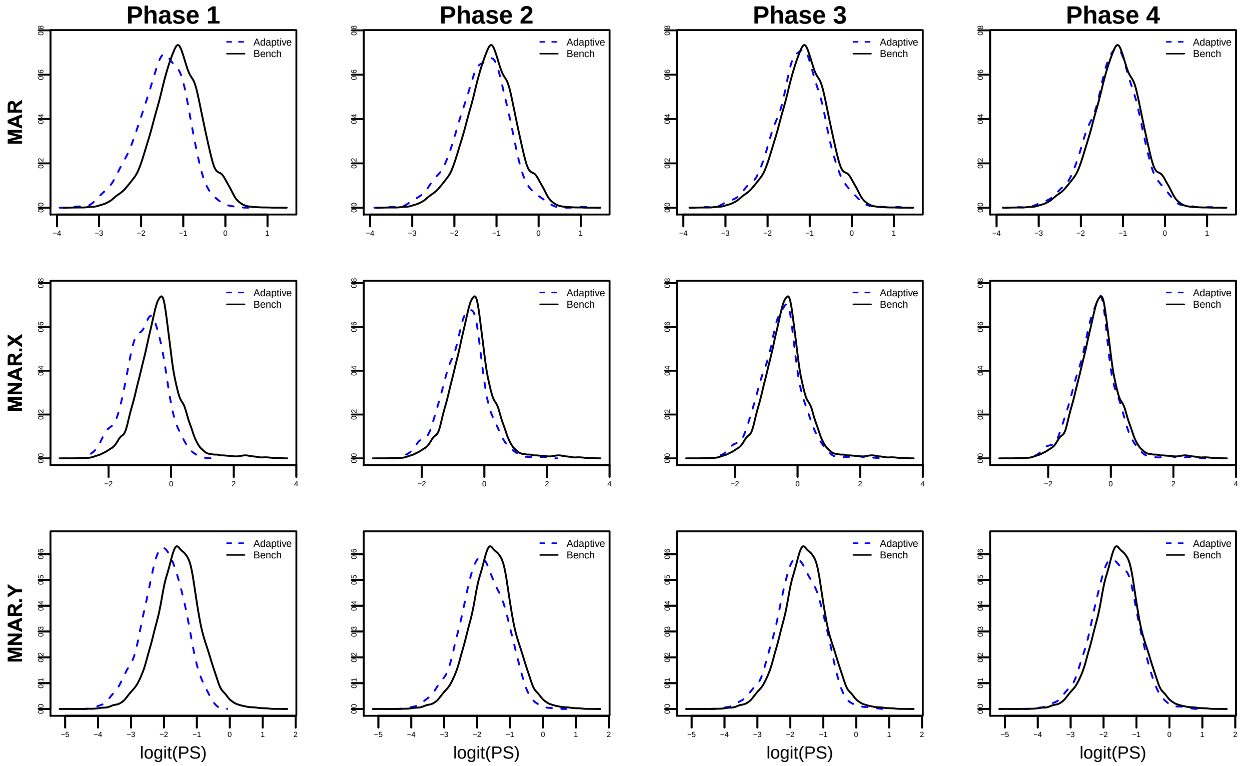
Each row panel shows different nonresponse models. Each column panel shows sampling phases.

(b) Propensity score density $P(\mathbf{Z}, \mathbf{X})$



Each row panel shows different nonresponse models. Each column panel shows sampling phases.

(c) Propensity score density $P(\mathbf{Z}, \mathbf{X}, \mathbf{Y})$.



Each row panel shows different nonresponse models. Each column panel shows sampling phases.

2.5 Discussion

This study examines the effectiveness of an adaptive sampling design using benchmarked sequential sampling method for improving respondent representativeness. Respondent representativeness at each phase is assessed by sequentially comparing the current respondent pool to an external micro-level benchmark in terms of distributions of subject characteristics. The multivariate nature of subject characteristics is summarized and simplified by propensity scores, and the sampling probabilities for the subsequent stages are defined as to maximize the balance based on the propensity score. This adaptive sampling design is studied using a simulated covariance structure (multivariate normal distribution) and observed covariance structure (NHIS and BRFSS micro-level

Table 2.3: NHIS and BRFSS data. Comparison of summary statistics between the Benchmark and focal surveys (MCAR nonresponse model).

Variable	Values	Pop Values	Bench[1] Average	Fixed Sampling Phase 1	Adaptive Sampling Phase 2	Adaptive Sampling Phase 3[2]	Adaptive Sampling Phase 4	Adaptive Design Std Diff[3] (%)	Fixed Design Std Diff[3] (%)
Z-variables									
Region	Northeast	18.27	18.27	18.40	18.30	18.10	18.20	-0.38	1.26
	Midwest	24.05	24.06	24.20	24.20	24.20	24.20	0.58	0.17
	South	31.75	31.74	31.80	31.70	31.80	31.80	0.19	-0.13
	West	25.94	25.92	25.60	25.80	25.80	25.90	-0.08	-0.46
Age	(in years)	55.34	55.33	55.21	55.27	55.36	55.41	0.16	-0.07
Gender	Male	38.41	38.44	38.30	38.30	38.30	38.30	-0.36	-0.36
Hispanic	Yes	6.60	6.60	6.70	6.60	6.50	6.40	-3.03	0.00
Race	White only	83.77	83.76	83.80	83.90	84.00	84.10	0.41	0.05
	Black only	8.64	8.65	8.70	8.70	8.70	8.60	-0.58	0.58
	AIAN[4]	1.59	1.59	1.50	1.60	1.50	1.50	-5.66	0.63
	Asian only	2.00	1.99	2.00	1.90	1.90	1.90	-4.50	0.50
	Other race	2.17	2.17	2.10	2.10	2.10	2.10	-3.23	-3.23
	Multiracial	1.83	1.83	1.80	1.70	1.70	1.70	-7.10	-1.64
X-variables									
Marital Status	Married	55.71	55.67	55.30	55.70	55.70	55.70	0.05	-0.13
	Widowed	13.84	13.83	13.80	13.80	13.80	13.90	0.51	-0.22
	Divorced	13.65	13.68	13.90	13.60	13.70	13.70	0.15	0.88
	Separated	2.08	2.08	2.10	2.10	2.10	2.10	0.96	0.96
	Never	14.08	14.10	14.30	14.20	14.10	14.00	-0.71	0.71
	Unknown	0.64	0.64	0.70	0.60	0.60	0.60	-6.25	-6.25
Education	$\leq$ Kindergarten	0.14	0.15	0.20	0.20	0.10	0.10	-35.7	35.7
	Grade 1-8	3.22	3.23	3.20	3.30	3.20	3.20	-0.93	2.17
	Grade 9-11	6.37	6.38	6.40	6.20	6.20	6.30	-1.26	0.31
	G12/GED	29.88	29.86	29.70	29.80	29.90	29.90	0.13	-0.54
	College 1-3 yrs	27.13	27.12	27.30	27.30	27.20	27.20	0.29	0.29
	College $\geq$4 yrs	33.26	33.27	33.20	33.30	33.40	33.30	0.09	-0.21
Y-variables									
HeartAttack (%)	Yes	5.74	5.74	5.60	5.60	5.60	5.60	-2.43	-0.70
Strokes (%)	Yes	3.92	3.93	3.90	4.00	3.90	3.90	-0.77	-0.77
Hypertension (%)	Yes	38.42	38.42	37.90	38.30	38.30	38.40	-0.05	-0.31
Diabetes (%)	Yes	11.90	11.91	11.70	11.80	11.80	11.90	-0.08	-0.08
Health Coverage (%)	Yes	89.28	89.28	89.30	89.30	89.30	89.30	0.02	-0.09
General Health (%)	Excellent	18.58	18.57	18.60	18.60	18.60	18.60	0.16	0.16
	Very Good	32.05	32.05	32.30	32.30	32.30	32.20	0.47	-0.16
	Good	30.45	30.47	30.30	30.20	30.30	30.30	-0.56	0.10
	Fair	13.22	13.21	13.10	13.20	13.20	13.20	-0.08	-0.08
	Poor	5.70	5.70	5.70	5.70	5.70	5.70	0.00	0.00

[1] Bench: Benchmark. [2] Phase 2 represents cumulative data, including data from both phase 1 and 2. Similarly, phase 3 and phase 4 columns reflect cumulative data up to the corresponding phases. [3] Percent absolute differences comparing phase 4 (overall sample) to benchmark. [4] AIAN : American Indian or Alaska Natives.

Table 2.4: NHIS and BRFSS data. Comparison of summary statistics between the Benchmark and focal surveys (MAR nonresponse model).

				Fixed	Adaptive	Adaptive	Adaptive	Adaptive	Fixed
Variable	Values	Pop	Bench[1]	Sampling	Sampling	Sampling	Sampling	Design	Design
		Values	Average	Phase 1	Phase 2	Phase 3[2]	Phase 4	std Diff[3]	Std Diff[3]
								(%)	(%)
Z-variables									
Region	Northeast	18.27	18.27	9.50	13.40	15.20	16.10	-11.88	-48.00
	Midwest	24.05	24.06	25.20	24.70	24.40	24.30	1.00	5.99
	South	31.75	31.74	35.40	33.60	32.90	32.60	2.71	10.27
	West	25.94	25.92	29.90	28.20	27.40	26.90	3.78	15.73
Age	(in years)	55.34	55.33	57.89	56.76	56.22	55.94	1.10	4.64
Gender	Male	38.41	38.44	49.40	43.90	41.40	40.30	4.84	29.05
Hispanic	Yes	6.60	6.60	11.80	9.40	8.30	7.70	16.67	75.76
Race	White only	83.77	83.76	70.80	77.00	79.90	81.30	-2.94	-15.23
	Black only	8.64	8.65	14.10	11.40	10.10	9.50	9.84	61.92
	AIAN[4]	1.59	1.59	3.30	2.50	2.20	2.00	25.79	101.26
	Asian only	2.00	1.99	3.80	2.90	2.60	2.40	20.50	95.5
	Other race	2.17	2.17	4.30	3.40	2.90	2.70	24.42	98.16
	Multiracial	1.83	1.83	3.60	2.80	2.40	2.20	20.22	96.72
X-variables									
Marital Status	Married	55.71	55.67	53.60	54.40	54.90	55.10	-1.02	-3.54
	Widowed	13.84	13.83	15.40	14.70	14.40	14.20	2.67	12.07
	Divorced	13.65	13.68	13.80	13.80	13.70	13.80	0.88	0.15
	Separated	2.08	2.08	2.50	2.30	2.20	2.20	5.77	15.38
	Never	14.08	14.10	14.00	14.00	14.10	14.00	-0.71	-0.71
	Unknown	0.64	0.64	0.60	0.70	0.70	0.70	9.38	9.38
Education	$\leq$ Kindergarten	0.14	0.15	0.20	0.20	0.20	0.20	35.71	35.71
	Grade 1-8	3.22	3.23	4.60	4.00	3.60	3.50	8.39	42.55
	Grade 9-11	6.37	6.38	7.60	7.00	6.70	6.60	3.45	20.72
	G12/GED	29.88	29.86	30.50	30.20	30.10	30.00	0.47	2.14
	College 1-3 yrs	27.13	27.12	26.50	26.90	27.00	27.10	-0.07	-3.39
	College $\geq$4 yrs	33.26	33.27	30.60	31.70	32.30	32.60	-2.01	-7.43
Y-variables									
HeartAttack (%)	Yes	5.74	5.74	7.20	6.50	6.20	6.10	6.27	25.44
Strokes (%)	Yes	3.92	3.93	4.70	4.40	4.20	4.10	4.34	19.64
Hypertension (%)	Yes	38.42	38.42	42.90	40.90	40.00	39.50	2.81	11.92
Diabetes (%)	Yes	11.90	11.91	14.20	13.20	12.70	12.50	4.96	19.24
Health Coverage (%)	Yes	89.28	89.28	88.30	88.70	88.90	89.00	-0.31	-0.99
General Health (%)	Excellent	18.58	18.57	16.50	17.40	18.00	18.10	-2.53	-11.68
	Very Good	32.05	32.05	29.70	30.80	31.30	31.60	-1.40	-8.58
	Good	30.45	30.47	32.10	31.30	30.90	30.70	0.76	6.67
	Fair	13.22	13.21	15.20	14.40	13.90	13.70	3.71	15.05
	Poor	5.70	5.70	6.50	6.10	5.90	5.90	3.51	15.79

[1] Bench: Benchmark. [2] Phase 2 represents cumulative data, including data from both phase 1 and 2. Similarly, phase 3 and phase 4 columns reflect cumulative data up to the corresponding phases. [3] Percent absolute differences comparing phase 4 (overall sample) to benchmark. [4] AIAN : American Indian or Alaska Natives.

Table 2.5: NHIS and BRFSS data. Comparison of summary statistics between the Benchmark and focal surveys (MNAR$_X$ nonresponse model).

				Fixed	Adaptive	Adaptive	Adaptive	Adaptive	Fixed
Variable	Values	Pop	Bench[1]	Sampling	Sampling	Sampling	Sampling	Design	Design
		Values	Average	Phase 1	Phase 2	Phase 3[2]	Phase 4	Std Diff[3]	Std Diff[3]
								(%)	(%)
Z-variables									
Region	Northeast	18.27	18.27	9.80	13.80	15.50	16.50	-9.69	-46.36
	Midwest	24.05	24.06	25.60	24.90	24.60	24.40	1.41	6.82
	South	31.75	31.74	34.30	32.80	32.30	32.10	1.13	8.06
	West	25.94	25.92	30.30	28.60	27.60	27.10	4.55	16.50
Age	(in years)	55.34	55.33	57.55	56.56	56.13	55.91	1.05	3.76
Gender	Male	38.41	38.44	50.40	44.60	42.00	40.60	5.62	30.88
Hispanic	Yes	6.60	6.60	11.10	9.00	8.10	7.50	13.64	68.18
Race	White only	83.77	83.76	72.40	77.60	80.10	81.40	-2.82	-13.56
	Black only	8.64	8.65	12.80	10.80	9.90	9.40	8.68	46.88
	AIAN[4]	1.59	1.59	3.20	2.50	2.10	1.90	1.95	94.97
	Asian only	2.00	1.99	3.90	3.00	2.60	2.40	20.50	95.50
	Other race	2.17	2.17	4.20	3.30	2.80	2.60	19.82	98.16
	Multiracial	1.83	1.83	3.50	2.80	2.40	2.20	20.22	96.72
X-variables									
Marital Status	Married	55.71	55.67	63.10	61.20	60.10	59.50	6.87	13.52
	Widowed	13.84	13.83	12.60	12.90	13.10	13.10	-5.27	-10.33
	Divorced	13.65	13.68	11.40	12.00	12.50	12.80	-6.45	-17.44
	Separated	2.08	2.08	1.20	1.30	1.40	1.50	-27.88	-37.50
	Never	14.08	14.10	11.60	12.40	12.80	13.00	-7.81	-17.05
	Unknown	0.64	0.64	0.10	0.20	0.20	0.20	-68.75	-84.38
Education	$\leq$ Kindergarten	0.14	0.15	0.30	0.20	0.20	0.20	35.71	35.71
	Grade 1-8	3.22	3.23	3.30	3.20	3.00	3.00	-7.14	5.28
	Grade 9-11	6.37	6.38	6.20	6.00	5.90	5.90	-7.54	-4.40
	G12/GED	29.88	29.86	30.90	30.40	30.40	30.40	1.81	3.15
	College 1-3 yrs	27.13	27.12	26.40	26.70	26.80	26.90	-0.81	-2.29
	College $\geq$ 4 yrs	33.26	33.27	33.00	33.50	33.60	33.70	1.29	-0.81
Y-variables									
HeartAttack (%)	Yes	5.74	5.74	7.00	6.30	6.10	6.00	4.53	20.21
Strokes (%)	Yes	3.92	3.93	4.40	4.10	4.00	4.00	1.79	9.44
Hypertension (%)	Yes	38.42	38.42	41.30	39.80	39.20	38.90	1.25	8.02
Diabetes (%)	Yes	11.90	11.91	13.60	12.80	12.50	12.20	2.44	15.04
Health Coverage (%)	Yes	89.28	89.28	89.10	89.40	89.40	89.40	0.13	-0.31
General Health (%)	Excellent	18.58	18.57	17.30	18.10	18.30	18.60	0.16	-7.37
	Very Good	32.05	32.05	30.50	31.10	31.50	31.60	-1.40	-4.21
	Good	30.45	30.47	32.10	31.50	31.20	31.00	1.74	4.70
	Fair	13.22	13.21	14.20	13.60	13.40	13.20	-0.08	7.49
	Poor	5.70	5.70	5.90	5.70	5.70	5.60	-1.75	5.26

[1] Bench: Benchmark. [2] Phase 2 represents cumulative data, including data from both phase 1 and 2. Similarly, phase 3 and phase 4 columns reflect cumulative data up to the corresponding phases. [3] Percent absolute differences comparing phase 4 (overall sample) to benchmark. [4] AIAN : American Indian or Alaska Natives.

Table 2.6: NHIS and BRFSS data. Comparison of summary statistics between the Benchmark and focal surveys (MNAR$_Y$ nonresponse model).

Variable	Values	Pop Values	Bench[1] Average	Fixed Sampling Phase 1	Adaptive Sampling Phase 2	Adaptive Sampling Phase 3[2]	Adaptive Sampling Phase 4	Adaptive Design Std Diff[3] (%)	Fixed Design Std Diff[3] (%)
Z-variables									
Region	Northeast	18.3	18.3	11.3	14.5	15.8	16.6	-9.3	-38.8
	Midwest	24.0	24.1	25.6	24.8	24.5	24.3	0.8	5.4
	South	31.7	31.7	34.1	32.9	32.5	32.2	1.6	8.5
	West	25.9	25.9	29.1	27.8	27.2	26.8	3.5	12.0
Age	(in years)	55.3	55.3	59.0	57.4	56.6	56.2	1.5	6.8
Gender	Male	38.4	38.4	47.9	43.9	41.7	40.6	5.7	25.0
Hispanic	Yes	6.6	6.6	10.3	8.8	8.1	7.6	15.2	56.1
Race	White only	83.8	83.8	73.4	78.0	80.1	81.4	-2.9	-12.3
	Black only	8.6	8.7	12.1	10.5	9.8	9.4	8.1	39.5
	AIAN[4]	1.6	1.6	3.2	2.5	2.2	2.0	25.0	87.5
	Asian only	2.0	2.0	3.8	3.0	2.6	2.4	20.0	90.0
	Other race	2.2	2.2	4.2	3.4	2.9	2.7	22.7	90.9
	Multiracial	1.8	1.8	3.3	2.7	2.3	2.1	16.7	88.9
X-variables									
Marital Status	Married	55.7	55.7	60.5	59.7	59.1	58.7	5.3	8.6
	Widowed	13.8	13.8	14.9	14.2	13.8	13.7	-0.7	8.7
	Divorced	13.6	13.7	12.1	12.4	12.7	12.8	-6.6	-11.8
	Separated	2.1	2.1	1.4	1.5	1.6	1.6	-23.8	-33.3
	Never	14.1	14.1	11.0	12.1	12.6	13.0	-7.8	-22.7
	Unknown	0.6	0.6	0.1	0.2	0.2	0.2	-66.7	-83.3
Education	$\leq$ Kindergarten	0.1	0.1	0.2	0.2	0.2	0.2	100.0	100.0
	Grade 1-8	3.2	3.2	3.8	3.4	3.3	3.2	0.0	18.8
	Grade 9-11	6.4	6.4	6.6	6.3	6.2	6.2	-3.1	4.7
	G12/GED	29.9	29.9	32.1	31.2	30.8	30.6	2.3	6.4
	College 1-3 yrs	27.1	27.1	26.7	26.9	27.1	27.1	0.0	-1.9
	College $\geq$4 yrs	33.3	33.3	30.7	31.9	32.5	32.8	-1.5	-7.2
Y-variables									
HeartAttack (%)	Yes	5.7	5.7	10.0	8.4	7.6	7.2	26.3	73.7
Strokes (%)	Yes	3.9	3.9	6.8	5.7	5.2	4.9	25.6	74.4
Hypertension (%)	Yes	38.4	38.4	49.9	45.8	43.6	42.4	10.4	30.0
Diabetes (%)	Yes	11.9	11.9	18.3	16.0	14.8	14.1	18.5	53.8
Health Coverage (%)	Yes	89.3	89.3	92.2	91.5	91.2	90.9	1.8	3.1
General Health (%)	Excellent	18.6	18.6	12.1	14.1	15.3	16.0	-14.0	-35.5
	Very Good	32.1	32.0	30.0	31.1	31.6	31.9	-0.3	-6.9
	Good	30.5	30.5	32.7	32.0	31.5	31.2	2.3	7.5
	Fair	13.2	13.2	17.0	15.5	14.8	14.4	9.1	29.6
	Poor	5.7	5.7	8.3	7.3	6.8	6.5	14.0	43.9

[1] Bench: Benchmark. [2] Phase 2 represents cumulative data, including data from both phase 1 and 2. Similarly, phase 3 and phase 4 columns reflect cumulative data up to the corresponding phases. [3] Percent absolute differences comparing phase 4 (overall sample) to benchmark. [4] AIAN : American Indian or Alaska Natives.

Table 2.7: NHIS and BRFSS data. Average Hellinger's distance from 100 simulations.

Nonresponse model	Design[1]	Phase 1	Phase 2	Phase 3	Phase 4
MCAR	A	11.1	3.8	2.4	1.8
	F	11.1	4.5	2.9	2.1
MAR	A	49.6	9.0	3.9	2.5
	F	49.6	42.3	40.3	39.6
MNAR$_X$	A	52.9	11.6	5.6	3.8
	F	52.9	45.3	42.7	41.4
MNAR$_Y$	A	59.0	13.0	6.2	4.1
	F	59.0	50.3	47.4	46.3

Note: Values shown are multiplied by 1000.

[1] F: Fixed design; A: Adaptive design.

data). In both implementations, the adaptive design shows its capacity to obtain a more representative respondent pool.

There are several reasons why achieving a representative respondent pool is appealing. First, a representative respondent pool allows one to perform simple analysis with limited correction for nonresponse adjustment. Second, survey estimates for a representative respondent pool are usually more robust to departures from the assumed form of the underlying nonresponse mechanism than survey estimates on weighting over an ordinary respondent pool, primarily because of reduced reliance on the model extrapolations. Third, even if the assumption for the nonresponse mechanism underlying a statistical adjustment is correct, the variance of the survey estimates will be lower in the representative respondent pool than in an ordinary respondent pool.

The goal of obtaining representative respondents is the same for the proposed adaptive sampling design and other adaptive designs published in recent literature (see Section 1.1). Three major aspects distinguish the proposed adaptive design and other adaptive design. First, other adaptive designs are adaptive in "data collection" strategy whereas the proposed strategy focuses on being adaptive in sampling design. Second, other adaptive designs emphasize the balance of respondents with respect to the selected sample, whereas the proposed strategy proposes to "balance" to the benchmark, a surrogate of the target population. Third, other adaptive designs use ad hoc sampling strategy at nonresponse follow-up to 'fix' the imbalance after the fact, which becomes increasingly complex with the increase in the number of covariates to balance, whereas under the proposed strategy it is relatively easy to incorporate both main effects and interactions even if numerous covariates are added to the balancing task. The simulation examples demonstrated the simplicity of deriving the adaptive sampling rate needed to achieve agreement with the target population. Our strategy was superior in two ways: 1) easy to implement and less sensitive to the number of covariates for balancing and 2) survey respondents become more representative of the target population in a multivariate fashion which preserves the variance-covariance structure of the sample composition.

One may question the need to improve respondent representativeness since nonresponse weighting adjustments correct for nonresponse bias. However, standard nonresponse adjustments exclusively focus on the use of auxiliary information which is often limited and might not be informative for nonresponse bias reduction. Suppose the missing at random assumption holds and rich auxiliary information are available for both respondents and nonrespondents, weighting adjustments still suffer two disadvantages. First, weighting adjustments designed to attenuate nonresponse bias could increase the variances of the estimates (Little and Vartivarian, 2005). Second, the magnitude of the nonresponse adjustment could be dampened by further weighting adjustments which are performed to ensure that the distribution of sample characteristics conforms to those of the target population (Little and Vartivarian, 2003). The proposed adaptive sampling strategy offers an alternative solution to solely relying on weighting to reduce nonresponse bias. The proposed strategy decreases the nonresponse bias by first improving the respondent representativeness and consequently minimizing corrective weighting adjustments, increasing the precision of the survey estimates.

Several limitations may render the proposed strategy ill-favored. In terms of weighting versus adaptive sampling, post-survey weighting is applicable to any survey design whereas the adaptive sampling requires a 1) multi-phase design to allow the incremental adjustments and 2) micro-level benchmark data that captures the target population of interest. The inherent property of the strategy suggests that the bias reduction is only as good as the benchmark data available. In addition, it may not be common to have auxiliary variables $\mathbf{Z}$ that are related to survey variables $\mathbf{Y}$, in such case the bias reduction effect relies on the correlation of (Z, X) and (X, Y). Without them, there are no bias reductions. With respect to the cost, the proposed design may increase survey cost if contacting a new subject costs more than recontacting a nonrespondent when the probabilities of getting a response are the same. The proposed design also requires covariate information on nonrespondents at each phase, which may not be available for some populations.

CHAPTER 3

Benchmarked Sequential Sampling and Benchmarked Multiple Imputation

Current solutions for unit nonresponse focus on improving nonresponse follow-up and enhancing post-survey weighting adjustment. We propose an alternative inferential paradigm to adjust for unit nonresponse using micro-level auxiliary data that captures the same features of the target population, referred to as "benchmark" hereafter. We describe a benchmark-driven mitigation and imputation (M&I) strategy, in the context of a multi-sampling survey, that sequentially guides the sampling and estimation to improve survey inferences regardless of the nonresponse mechanism. The M&I strategy employs a high quality benchmark to 1) (mitigate) rectify undesirable nonresponse patterns through a calibrated sequential sampling design; and 2) (impute) recover population information through calibrated multivariate imputation by chained equations (MICE), thus achieving less biased survey estimates. The performance of the M&I strategy will be evaluated by simulation experiments to mimic adaptive design under various nonresponse mechanisms including missing not at random (MNAR). An illustration using data from the American Community Survey (ACS) and the Current Population Survey (CPS) is also presented. We report on the preservation of marginal and joint distribution for population estimates of three sampling designs from respondent data, completed data(respondent and imputed nonrespondent data), and joint data (completed data and benchmark data).

3.1 Introduction

This study discusses methods applicable to surveys with rich micro-level auxiliary data, such as establishment and recurrent surveys and surveys that use other survey respondents as a sampling frame. For the micro-level auxiliary data, we consider those that 1) are available before survey implementation and 2) captures the information on the target population. We term such auxiliary data as a “Benchmark”. We propose a benchmarked sequential sampling and benchmarked imputation strategy to amplify the dividends of the auxiliary data and minimize the bias of unit nonresponse, including missingness not at random (MNAR) (Little and Rubin, 2002).

Unit nonresponse is traditionally managed by the post-survey adjustments to the weights for the respondents. Sample weighting adjustments conforming the respondent data to sample totals, such as weighting class adjustments and response propensity weighting, were shown to reduce the bias of the estimated mean (Little, 1986; Bethlehem, 1988; Kalton and Maligalig, 1991). Population weighting adjustments conforming the respondent data to population totals, such as post-stratification and calibration weighting, were reported to reduce nonresponse bias and variance, as well as bias due to incomplete coverage (Deville and Särndal, 1992; Lundström and Särndal, 1999; Särndal and Lundström, 2005).

In practice, variance can be managed and measured reasonably well. The reason is that inflated variance due to unit nonresponse is primarily the consequences of smaller than anticipated sample size and increased variation of the survey weights (Brick, 2013). Sample size could be expanded. Effective variance reduction can be achieved by prudent weighting and weight trimming methods. Conversely, the reduction of bias through weighting methods requires the missing at random (MAR) assumption and good weighting variables, though promising theoretically, remain illusive in practice. Statements such as “... bias is very difficult to measure” (Brick and Jones, 2008) and “...bias is the dominant component of the nonresponse-related error in the estimates” (Brick, 2013) made it clear that it was the nonresponse bias that captivated survey researchers.

Aside from the validity of the MAR assumption, the concerns over the effectiveness of weighting strategy on bias reduction exacerbated over rapidly falling response rates. For instance, survey practitioners are calibrating the current norm of 9–20% respondents in random digit dial surveys to represent the entire sample, instead of calibrating historically 80–$90+\%$ of respondents. The growing interest in mitigating nonresponse bias before weighting adjustments is evident, especially from the ample articles in recent years on adaptive design and related topics (Groves and Heeringa, 2006; Wagner, 2008; Schouten et. al., 2009, 2011, 2013). The premise of adaptive design at curtailing nonresponse bias is that representative respondents result in less biased estimates (Schouten et. al., 2016; Särndal and Lundquist, 2014). These adaptive strategies are guided by auxiliary data to assess respondent representativeness, to target nonrespondents of interest, and to optimize resources on nonresponse follow-up. The improvement in respondent representativeness is achieved by intervening in data collection adaptively, implemented during nonresponse follow-up. (Wagner, 2008; Schouten et al., 2009, 2011, 2013, Särndal and Lundquist, 2014).

While improved respondent representativeness may reduce bias, doing so by varying strategies on data collection complicates the inferential process. This is because response propensity is dynamic and depends not only on auxiliary variables but also on data collection strategies. Researchers have shown that response propensity of a sampled unit changes as the data collection protocol changes (Schouten et al. 2011; Olsen and Groves, 2012; Brick, 2013). Therefore, maintaining a standardized data collection protocol has at least two advantages: the ease of field administration and the ease of inferential process. Instead of adaptive data collection, the adaptive sampling design as proposed in Chapter 2 provides an opportunity to improve respondent representativeness while maintaining a standardized data collection protocol on all subjects, and therefore a coherent inferential process.

The adaptive sampling method described in Chapter 2 aims at benchmarking frame variables (denoted by $\mathbf{Z}$) from the sample respondents (R) to those of the benchmark (B), obtaining $f_R(\mathbf{Z}) \approx f_B(\mathbf{Z})$. Amending results from chapter 2, we propose to match not only the frame variables but also the survey covariates (denoted by $\mathbf{X}$). Similar to Chapter 2, the proposed design requires a multi-phase survey setting in which replicates of the full sample are sequentially introduced at each sampling phase. The $\mathbf{X}$ information needed at the sampling stage is predicted using models fitted to $\mathbf{Z}$ on the frame and the respondent data from previous phases. Following the sampling, we propose using benchmarked multivariate imputation by chained equations (MICE) to adjust for unit nonresponse (see Section 3.2.4 for more details).

Traditionally, imputation has been the standard procedure used to compensate for item nonresponse (Brick and Jones, 2008; Brick, 2013; Brick and Kalton, 1996). There is relatively little published research on imputation for unit nonresponse. An exception is by Rässler and Schnell (2004) who suggest that multiple imputation is a superior strategy than post-survey weighting adjustment for nonresponse bias correction. Imputation as an alternative for unit nonresponse adjustment has been primarily explored by researchers with a model-based perspective (Gelman, 2007; Little and Rubin, 2013). Their principal argument is based on the preposition that effective bias reduction can only be achieved by modeling survey outcome variables (Greenlees, et al., 1982; Beaumont, 2000). In our strategy, by combining the benchmark and the respondent set, unit nonresponse mimics a special case of item nonresponse and, as a result, warrants the imputation strategy (see Section 3.2.1 and figure 3.1 for more details), fulfilling the modeling perspective on bias reduction.

The key component, the benchmark data, is used to calibrate both the sampling and imputation in a sequence of sampling-imputation procedures. Through iterative benchmarking, the population structure approximated by benchmark is incorporated into sampling to restore balance between the respondent and the benchmark (mitigation step), attenuating the undesirable nonresponse pattern. Similarly, the population structure is embedded in the imputation models derived from iterative model fitting using combined data from benchmark and the respondent (imputation step), enhancing results from the sampling procedure. By integrating sampling and imputation to adjust for unit nonresponse, we show that this inferential paradigm better preserves population distribution and reduces nonresponse bias, especially for MNAR mechanisms. This unit nonresponse study is set apart from other survey design and weighting literature through the preservation of joint distribution and the bias reduction at both design and estimation stages.

The performance of the M&I strategy will be evaluated by simulation experiments to mimic adaptive design under various nonresponse mechanisms, including “missing not at random” (MNAR). Survey estimates from a focal survey are compared to those derived from the benchmark. The decision to stop the M&I procedure is based on marginal and conditional distributions of survey data as well as on objectives of the survey. An application to the Current Population Survey (CPS) and the American Community Survey (ACS) data will serve to illustrate how the proposed strategy is likely to perform in practice.

In Section 3.2, we define notation and describe the rationale and methods for the proposed strategy. Section 3.3 demonstrates the proposed strategy by presenting a simulation study, its design and setup, using multivariate normal data with one covariate (X), one survey outcome (Y) and one frame variable (Z). Section 3.4 gives the results of the application using CPS and ACS data. In Section 3.5 we conclude with some discussion of strengths, limitations, and the applicability of the proposed strategy.

3.2 Method

In a multi-phase survey design, our adaptive strategy uses benchmarked sequential sampling and imputation to improve survey estimates under unit nonresponse. At each phase, benchmarked sequential sampling derives the sampling probability with an objective of achieving $P(\mathbf{X}_R, \mathbf{Z}_R) \approx P(\mathbf{X}_B, \mathbf{Z}_B)$, mitigating the nonresponse pattern to improve respondent representativeness. An improved respondent pool from the sampling step prepares the way for benchmarked multiple imputation to recover population information. After each mitigation-and-imputation cycle, we evaluate marginal and conditional distributions of survey variables to assess resemblance between the benchmark (D_B), the respondent set (D_R), and the completed sample ($D_R \cup \hat{D}_{NR}$). The increased similarity of marginal and conditional distributions towards those of the benchmark produces better survey estimates, especially for MNAR unit nonresponse. Before discussing the steps necessary to implement the adaptive design, we discuss the pivotal role of a benchmark in our strategy.

3.2.1 The Benchmark Data

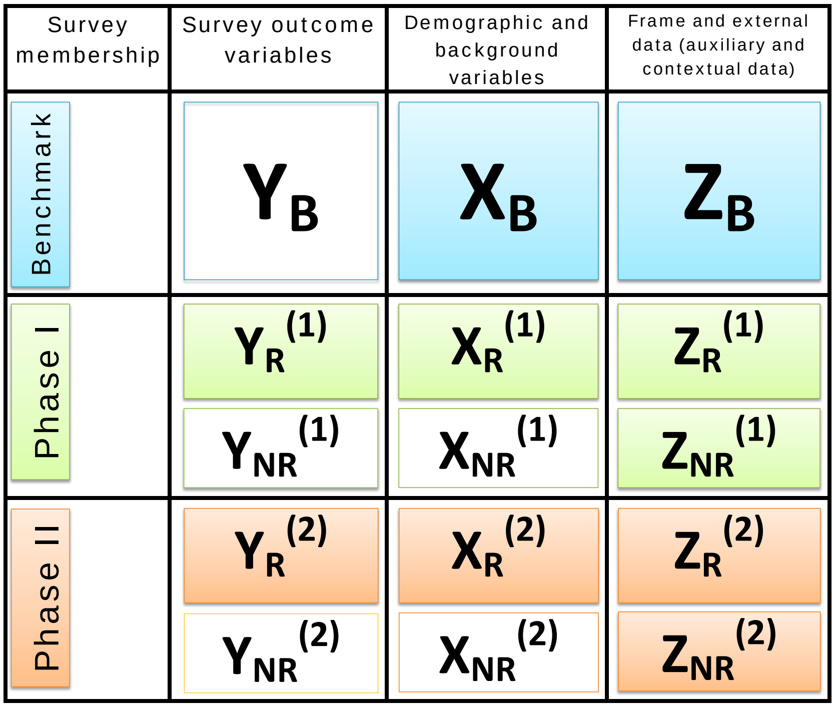
The "benchmark" here is defined as a micro-level auxiliary data that captures the characteristics of the target population, that is $f(\mathbf{Y}_B, \mathbf{X}_B, \mathbf{Z}_B) \approx f(\mathbf{Y}_{pop}, \mathbf{X}_{pop}, \mathbf{Z}_{pop})$. For example, if the U.S. non-institutionalized household civilian population is the target, the ACS publicly-available micro-data may serve as the benchmark. The current convention in this situation is to use the marginal or cross-classified counts of selected ACS variables as the control totals for calibration weighting adjustment (Valliant, et.al., 2013).

In calibration weighting adjustment, the control totals guide the survey weights to conform the weighted respondent data to the population counts. In comparison, the proposed strategy uses the benchmark to guide the sampling and imputation so as to conform the respondent data (D_R) and completed data ($D_C = D_R \bigcup \hat{D}_{NR}$) to the population. While similar in principle, the proposed method capitalizes on the micro-level data in the benchmark to recover population multivariate structure among covariates in the focal survey.

Figure 3.1 illustrates the conceptual data structure available in a hypothetical two-phase survey, where we show the benchmark data set D_B, the focal survey respondent sets $D_R^{(1)}$ and $D_R^{(2)}$, and the nonrespondent sets $D_{NR}^{(1)}$ and $D_{NR}^{(2)}$. Note, the superscript denotes the sampling phases. The shaded areas represent data available before imputation and the blank areas represent missing data to be imputed. In the proposed method, benchmark and the focal survey share some $\mathbf{X}$ and $\mathbf{Z}$, whereas $\mathbf{Y}$ is available for respondents and usually not available in the benchmark. This data structure mimics a special case of monotone item nonresponse, leading to a natural consideration to imputation strategy.

"Frame variables" ($\mathbf{Z}$) in this study is a label for any auxiliary data that are available before data collection and serves as information for sampling purposes. This could include information frequently found on the sampling frame, administrative records, paradata, and geographic contextual data, etc. Our strategy requires overlapping frame variables between the benchmark survey and the focal survey. The label of “benchmark” and ”frame data” depends on the function of the auxiliary information which can be distinguished and visualized based on their roles in the data shown in Figure 3.1.

Figure 3.1: Data available for M&I strategy



Survey membership	Survey outcome variables	Demographic and background variables	Frame and external data (auxiliary and contextual data)
Benchmark	Y_B	X_B	Z_B
Phase I	$Y_R^{(1)}$	$X_R^{(1)}$	$Z_R^{(1)}$
	$Y_{NR}^{(1)}$	$X_{NR}^{(1)}$	$Z_{NR}^{(1)}$
Phase II	$Y_R^{(2)}$	$X_R^{(2)}$	$Z_R^{(2)}$
	$Y_{NR}^{(2)}$	$X_{NR}^{(2)}$	$Z_{NR}^{(2)}$

3.2.2 Mitigating Step: Benchmarked Sequential Sampling

Chapter 2 proposed a benchmarked sequential sampling method, BSS-Z, to improve respondent representativeness by using the frame variables. In a multi-phase survey setting, the frame variables $\mathbf{Z}$ are used to derive sampling probabilities to sequentially achieve $P(\mathbf{Z}_R) \approx P(\mathbf{Z}_B)$. Extending the BSS-Z method, we now derive sampling probabilities at each phase using $(\mathbf{Z}, \hat{\mathbf{X}})$. The new method aims at obtaining $P(\mathbf{X_R}, \mathbf{Z_R}) \approx P(\mathbf{X_B}, \mathbf{Z_B})$ after several sampling phases. The sampling probabilities are adjusted at each sampling phase to restore the resemblance between benchmark and cumulative respondents with respect to $\mathbf{Z}$ and $\mathbf{X}$. Through iterative benchmarking, this dynamic sampling strategy mitigates undesirable nonresponse patterns when $\mathbf{Z}$ and $\mathbf{X}$ correlate with $\mathbf{Y}$. We termed this new approach the BSS-X method.

To illustrate the main idea of benchmarked sequential sampling, suppose one wants to conduct a four-phase focal survey ($k = 1, \ldots, 4$) and no prior information is available on nonresponse patterns. The phase I sample follows a fixed probability sampling design such that the sample is drawn to be representative of the benchmark (i.e. the surrogate of the target population of interest). While phase I nonresponse mechanism renders phase I respondent non-representative, our goal is to restore the resemblance between the benchmark and the cumulative respondents at each successive phase with respect to $P(\mathbf{Z}, \mathbf{X})$. This goal at phase k can be written as

$$\pi^{(k)} \times P^{(k)}(\mathbf{Z}, \mathbf{X}|sample) + (1 - \pi^{(k)}) \times Q_R^{(k-1)}(\mathbf{Z}, \mathbf{X}) = P_B(\mathbf{Z}, \mathbf{X}) \tag{3.1}$$

where $Q_R^{(k-1)}(\mathbf{Z}, \mathbf{X})$ denotes the distribution of $(\mathbf{Z}, \mathbf{X})$ from cumulative respondents up to phase $(k - 1)$. Let $\pi^{(k)} = n^{(k)}/N_k$ where $n^{(k)}$ is the sample size for phase k and N_k is the cumulative sample size up to phase k. Then $\pi^{(k)}$ denotes the proportion of the new subjects we expect to interview in phase k, and hence $1 - \pi^{(k)}$ is the proportion of cumulative responding subjects obtained up to phase k.

By combining (3.1) with the Bayes formula (2.3), the sampling rate for phase k is derived as

$$P^{(k)}(sample|\mathbf{Z}, \mathbf{X}) = \{\frac{P_B(\mathbf{Z}, \mathbf{X})}{P(\mathbf{Z}, \mathbf{X})} - (1 - \pi^{(k)})\frac{Q_R^{(k-1)}(\mathbf{Z}, \mathbf{X})}{P(\mathbf{Z}, \mathbf{X})}\}\frac{P^{(k)}(sample)}{\pi^{(k)}} \tag{3.2}$$

where $P(\mathbf{Z}, \mathbf{X})$ denotes the distribution of $(\mathbf{Z}, \mathbf{X})$. Equation (3.2) is a direct extension from equation (2.6) in Chapter 2.

Estimating $P_B(.)/P(.)$ and $Q_R(.)/P(.)$ are straightforward in a univariate case where the covariate of interest on the sampling frame could be age, gender or race, etc. For a multivariate situation like $P_B(\mathbf{Z}, \mathbf{X})/P(\mathbf{Z}, \mathbf{X})$, Similar to Chapter 2, we use propensity scores in place of $P_B(\mathbf{Z}, \mathbf{X})$ and $P(\mathbf{Z}, \mathbf{X})$. The propensity scores are estimated by combining $(\mathbf{Z}, \mathbf{X})$ data from the benchmark and the focal survey respondents, and modeling the probability of survey membership T ($T = 1$ for benchmark, and $T = 0$ otherwise) using the logistic regression. The estimated propensity score, that is, the predicted probability of "being in the benchmark survey", replaces $P_B(\mathbf{Z}, \mathbf{X})$ in formula (3.2), leading to a much simplified computation for $P^{(k)}(sample|\mathbf{Z}, \mathbf{X})$.

The propensity score densities of the benchmark and focal surveys are estimated with a non-parametric density function, and consequently the $P^{(k)}(sample|\mathbf{Z}, \mathbf{X})$ from (3.2) are computed, where $P^{(k)}(sample|\mathbf{Z}, \mathbf{X})$ is the sampling probability for the sampling units with characteristics of $(\mathbf{Z}, \mathbf{X})$ at phase k. For phase $k+1$, the propensity score model is refit using respondent information accumulated up to phase k, and the $P^{(k+1)}(sample|\mathbf{Z}, \mathbf{X})$ is updated accordingly. This implies that subjects with the same $(\mathbf{Z}, \mathbf{X})$ characteristics may have different sampling probability at each phase.

However, $\mathbf{X}$ is not available on the sampling frame and therefore is unknown before data collection occurs. We predict $\mathbf{X}_s^{(k)}$ through MICE using data $\mathbf{Z}_s^{(k)}$, $(\mathbf{Z}_B, \mathbf{X}_B)$ and cumulative $(\mathbf{Z}_R, \mathbf{X}_R)$ up to phase $(k-1)$, where $s = R \cup NR$. The average of imputed $\mathbf{X}$ replaces $\mathbf{X}$ in equation (3.2) for the sampling purpose. After data collection, the observed $\mathbf{X}$ replaces the predicted $\mathbf{X}$ for inferences of D_R.

3.2.3 Imputation Step: Benchmarked Multiple Imputation

Despite results at the mitigating step, the ultimate goal is to recover population information through accurate estimation. Our imputation strategy appends benchmark data to the respondent data (as shown in figure 3.1) which situates the unit nonresponse to mimic the item nonresponse. This set up serves as our imputation strategy for unit nonresponse, and the completed data ($D_C = D_R \cup \hat{D}_{NR}$)

are calibrated to the benchmark such that

$$P(\hat{\mathbf{Y}}_B|\mathbf{X}_B, \mathbf{Z}_B) \approx P(\mathbf{Y}_R, \hat{\mathbf{Y}}_{NR}|\mathbf{X}_R, \hat{\mathbf{X}}_{NR}, \mathbf{Z}_R, \hat{\mathbf{Z}}_{NR}).$$

The quality of imputed values can be evaluated by comparing the inferences from the benchmark data and the completed focal survey data.

Note that the equation above has $\hat{Z}_{NR}$ implying imputed Z_{NR}. This is because in the benchmarked sequential sampling design nonrespondents are those oversampled subjects with underrepresented characteristics from previous phases. When combining observed $\mathbf{Z}_R$ with $\mathbf{Z}_{NR}$, the marginal distribution of $\mathbf{Z}$ deviates from that of the benchmark, causing skewed results of $\mathbf{Y}$ and $\mathbf{X}$ from imputation. Therefore, instead of keeping observed values of $\mathbf{Z}_{NR}$ from the sampling frame, we impute them.

We implement the multivariate imputation by chained equations (MICE) to predict the nonrespondent values (van Buuren, 2007; Raghunathan, et. al. 2001). The MICE is an iterative regression prediction process that cycles through each variable with missing values, and models each variable conditional on others. MICE fits each variable with separate models and can handle complexities such as bounds or survey skip patterns, hence provides more accurate predicted values, resulting in better survey estimates (Raghunathan, et. al. 2001; White, et. al., 2011). The imputations themselves are predicted values from these regression models, and the predicted values are draws from the posterior predictive distributions obtained using Gibbs sampling (van Buuren, et. al. 2011). For example, variable types from ACS and CPS include continuous, binary, unordered categorical, and ordered categorical. Some of these variables have bounds and restrictions. MICE procedure imputes these variables using models appropriate to the variable type (e.g. linear regression for continuous variables) and handles special conditions accordingly.

Stuart, et al., (2009) gave the following clear and brief summary on the MICE procedures:

1. The variable with the least missingness (var 1) is imputed conditional on all variables with

no missingness.

2. The variable with the second least missingness is then imputed conditional on the variables with no missing values and var 1, and so on.

3. After all the variables have been cycled through in this way (one iteration), there are no longer any missing values in the data (one imputed data set).

4. Steps 1 to 3 is then repeated using this data set with no missing values.

The imputation literature recommends including variables in the imputation model that are predictive of the unknown values and that will be used in the subsequent analysis, especially variables that are likely to be associated with the variables that need to be predicted (White, et. al. 2011; Azur et. al. 2011). Furthermore, studies have recommended establishing an imputation model that is more general than the analysis model, which can be achieved by including additional auxiliary variables that will not be used in the analysis but that can improve the predictions (Collins et al, 2001; Schafer, 2003). When more predictors are included in the prediction model, it is more plausible that unknown depends only on observed characteristics and not on those that are missing, and the MAR assumption becomes more tenable (Raghunathan, et. al. 2001). Detail descriptions on practical implementation of MICE for issues such as number of imputation, models forms, variables to include, et. al. can be found in White, et. al., (2011).

The main difference distinguishing the imputation in this study and the imputation for a conventional missing data problem is that, in this study, the goodness of fit and diagnostics of imputation models can be evaluated by direct comparison of the marginal and conditional distributions of $\mathbf{X}$ and $\mathbf{Z}$ variables for the imputed data and the benchmark data. Well imputed $\mathbf{X}$ and $\mathbf{Z}$ enhance the imputation of $\mathbf{Y}$, when $\mathbf{X}$ and $\mathbf{Z}$ carry information on $\mathbf{Y}$. By calibrating completed data to match that of the population we reduce nonresponse bias even if nonrespondents are missing not at random.

3.2.4 Adaptive Design

Suppose there is an ongoing multi-phase focal survey implementing the M&I strategy, the goal is to decide if further data collection (i.e., more sampling phases) is needed, or whether the data collected so far is adequate to minimize unit nonresponse bias through imputation, regardless of nonresponse mechanism. This decision is based on factors such as the quality of the data that are already collected, how different the respondents are from the benchmark, and the properties of the imputed data with respect to the inferences.

We propose comparing the respondent data (D_R) and the benchmark data (D_B) to assess the M step and comparing the completed data ($D_C = D_R \bigcup \hat{D}_{NR}$) and the benchmark to assess the I step. The goals and methods corresponding to each step of the benchmarked adaptive survey design can be summarized as the following:

1. Draw phase $(k-1)$ sample by formula (3.2) and collect respondent data.
   - Evaluate similarity between D_R and D_B aiming at $P(\mathbf{X}_R, \mathbf{Z}_R) \approx P(\mathbf{X}_B, \mathbf{Z}_B)$.
2. Impute $(\mathbf{Y}_B, \mathbf{Y}_{NR}^{(K-1)}, \mathbf{X}_{NR}^{(K-1)}, \mathbf{Z}_{NR}^{(K-1)})$ and $(\mathbf{Y}_s^{(k)}, \mathbf{X}_s^{(k)})$ based on $(\mathbf{X}_B, \mathbf{Z}_B, \mathbf{Y}_R^{(K-1)}, \mathbf{X}_R^{(K-1)}, \mathbf{Z}_R^{(K-1)}, \mathbf{Z}_s^{(k)})$ using MICE.
   (Note, capital K denotes cumulative data up to k^{th} phase and lowercase k denotes single k^{th} phase. s denotes sample.)
   - Evaluate similarity between D_C and D_B aiming at $P(\mathbf{Y}_R, \hat{\mathbf{Y}}_{NR} | \mathbf{X}_R, \hat{\mathbf{X}}_{NR}, \mathbf{Z}_R, \hat{\mathbf{Z}}_{NR}) \approx P(\hat{\mathbf{Y}}_B | \mathbf{X}_B, \mathbf{Z}_B)$.
3. Consider:
   - If $P^{(K-1)}(\mathbf{Y}_R, \hat{\mathbf{Y}}_{NR} | \mathbf{X}_R, \hat{\mathbf{X}}_{NR}, \hat{\mathbf{Z}}_R, \hat{\mathbf{Z}}_{NR}) \approx P(\hat{\mathbf{Y}}_B | \mathbf{X}_B, \mathbf{Z}_B)$ then stop.
   - If $P^{(K-1)}(\mathbf{Y}_R, \hat{\mathbf{Y}}_{NR} | \mathbf{X}_R, \hat{\mathbf{X}}_{NR}, \hat{\mathbf{Z}}_R, \hat{\mathbf{Z}}_{NR}) \neq P(\hat{\mathbf{Y}}_B | \mathbf{X}_B, \mathbf{Z}_B)$ then repeat steps 1 and 2.

Since X is imputed at mitigation stage, the quality of the imputation adds a level of uncertainty on the performance in addition to the truncation properties for the sampling method (see section 2.2.4).

The advantage of M&I over weighting is that M&I strategy utilizes the joint probability of $(X, Y|Z)$ whereas weighting does not.

3.2.5 Evaluation

To evaluate the quality of the data produced by our strategy, we gauge it against benchmark parameters. We are interested in respondent data (D_R), completed data (D_C), and joint data (D_J), where $D_J = D_C \cup D_B$. Comparison of D_R and D_B tells us the performance of the sampling design (the mitigation step), whereas the assessment between D_C and D_B tells us the performance of the M&I strategy. We focus the evaluation on three categories: 1) marginal distribution 2) conditional distribution, and 3) accuracy of parameter estimates.

To evaluate benchmarked sequential sampling, we compute the difference between D_R and D_B on the following estimates from the respondent data: 1) the probability of $(X > c|Z > c)$, where c is a constant, 2) the correlation between X and Z (denoted as $\rho_{z,x}$) and 3) the difference of $\bar{Y}_R$ and $\bar{Y}_B$.

To evaluate benchmarked multiple imputation, we compute the difference between D_C and D_B on the following : 1) the probability of $(Y > c|X > c, Z > c)$, 2) the correlation between Y and X (denoted as $\rho_{y,x}$) and 3) the difference of Y_C and $\bar{Y}_B$. Note, measures derived from imputed data are combined using Rubin's rules (1987).

Finally, to evaluate the M&I strategy we assess the accuracy of parameter estimates on joint data (J) by reporting on 1) $Bias(\bar{Y}_J)$, 2) $\rho_{y,x}$, 3) $P(Y > c)$, 4) $P(Y > c|X > c)$, 5) $P(X > c|Z > c)$.

For the difference of ρ's between two samples, we report on RMSE, 95% confidence interval (CI) width, 95% CI coverage rate (CR), and bias. For example, the difference of $\rho_{y,x}$ between D_C

and D_B, denoted by d_ρ for simplification, its RMSE is computed by $RMSE = \sqrt{MSE}$, and

$$MSE(d_\rho) = B(d_\rho)^2 + V(d_\rho) \tag{3.3}$$

where $B(d_\rho)$ is the bias of a point estimate d_ρ ($d_\rho = \rho_{C(y,x)} - \rho_{B(y,x)}$), estimated by $bias(\hat{d}_\rho) = \sum_S(\hat{d}_{\rho_S} - d_{\rho_N})/r$. $\hat{d}_{\rho S}$ is the estimate from sample S, and d_{ρ_N} is the finite population parameter. $d_{\rho_N} = 0$ in our case. $V(d_\rho)$ is the empirical variance of $\hat{d}_\rho$, estimated by $Var(\hat{d}_\rho) = \sum_S(\hat{d}_{\rho S} - \bar{d}_\rho)^2/r$, where $\bar{d}_\rho = \sum_S \hat{d}_\rho/r$. r denotes simulation replicates.

The 95% confidence interval width is computed by

$$Width_{95} = 2 \times 1.96 \times SE, \tag{3.4}$$

where $SE = \sqrt{V(d_\rho)}$. The CR, the percentages of intervals that include d_{ρ_N}, are based on the nominal 95 percent confidence intervals ($\hat{d}_\rho \pm 1.96\hat{V}^{1/2}$) computed for each of the r simulations for each simulation scheme.

3.3 Simulation Study

The simulation study investigates the performance of M&I strategy on multivariate normal data in terms of bias reduction and recovery of population structure.

3.3.1 Simulation Design and Nonresponse Models

We simulate a finite population of size $N_{POP} = 500,000$ that comprises of three variables (Y, X, Z) from a multivariate normal distribution with the following joint distribution.

$$(Z, X, Y) \sim N_3(\mathbf{0}, \Sigma), \qquad \text{and} \tag{3.5}$$

$$\Sigma = \begin{pmatrix} 1 & cov(Z,X) & cov(Z,Y) \\ cov(Z,X) & 1 & cov(X,Y) \\ cov(Z,Y) & cov(X,Y) & 1 \end{pmatrix}$$

where $cov(Z, X)$ and $cov(Z, Y)$ are varied to examine the effects of Z with varying degrees of explanatory power on X and on Y. And $cov(X, Y)$ are varying to examine the effects of X with various degrees of explanatory power on Y. From this finite population, a simple random sample without replacement (srswor) of size $N_B = 25,000$ (5% of the finite population) is drawn to serve as the benchmark data and the remaining $475,000$ units serve as the sampling frame for the focal surveys. From this sampling frame, three concurrent focal surveys are simulated, one implements the proposed benchmarked sequential sampling design BSS-X, another implements the benchmarked sequential sampling design BSS-Z from chapter 2, and the third implements the conventional fixed sampling design (F).

For all three focal surveys, the first phase is a srswor of size $1,000$ ($n^{(1)} = 1,000$). This sample size is carried on for phase II to IV for the F sampling design. For BSS-X and BSS-Z designs, the k^{th} phase samples are of size $n^{(k)} = 1,200$, where $k = 2, 3, 4$. The larger sample sizes are to account for higher selection probability on subjects who are less likely to respond. These sample sizes in BSS-X and BSS-Z designs result in comparable number of respondents as with the fixed design such that we are imputing the same number of missing units at the imputation step and arriving in the same sizes of completed data.

The simulation study can be described as a $2 \times 2 \times 2 \times 4$ factorial design. The factors are:

Factor A: $cov(Z, X)$: high vs low

Factor B: $cov(Z, Y)$: high vs low

Factor C: $cov(X, Y)$: high vs low

Factor D: Four nonresponse mechanisms based on the nonresponse models described below.

For design factors A and B, a high correlation of two variables is set to be $\rho = 0.894$ (i.e. $R^2 = 0.8$) and a low correlation is $\rho = 0.447$ (i.e. $R^2 = 0.2$, where R^2 is the coefficient of determination in simple linear regression). For factor C, a high correlation between X and Y is $\rho_{X,Y} = 0.707$ (i.e. $R^2(X,Y) = 0.5$) and a low correlation is $\rho_{X,Y} = 0.3$ (i.e. $R^2(X,Y) = 0.09$). The combination scenario of a low correlation (L) in Factor A, a high correlation (H) in Factor B, and a high correlation (H) is Factor C is labeled "LHH" herein. Similarly the other seven scenarios of Factors A, B and C are labeled "LLL", "LLH", "HLL" ,"HLH", "HHL", and "HHH".

For nonresponse mechanisms (factor D), we simulate response status $\Re$ by drawing from Bernoulli random variables with $e(z,x,y) = Pr(\Re = 1|z,x,y)$ computed from four different nonresponse models:

1. $logit(e(z,x,y)) = Bernoulli(0.5)$. **(MCAR)**

2. $logit(e(z,x,y)) = 0.00020 + 0.8z$. **(MAR)**

3. $logit(e(z,x,y)) = 0.00001 + 0.31z + 0.61x$. **(MNAR$_X$)**

4. $logit(e(z,x,y)) = 0.00004 + 0.19z + 0.38x + 0.38y$. **(MNAR$_Y$)**

where $\Re = 1$ if subject responded, 0 otherwise. Each of these models generate an average of 50% response rate. For model 1, the response probability is independent of (Y,Z,X), i.e. MCAR. For model 1, the response probability depends on Z alone, i.e. MAR. For model 2 the response probability depends on both Z and X, labeled as MNAR$_X$. For model 3, the response probability depends on (Y,Z,X), labeled as MNAR$_Y$. Note, model 2 is MNAR since response depends on X, which is not observed for nonrespondents.

For the imputation procedure, five imputations are implemented where each imputation runs for 10 iterations. The imputation models include main effects of Y,X and Z. Let $n^{(K)}$ denote the intended sample size up to phase k. Of these $n_R^{(K)}$ responded and $n_{NR}^{(K)}$ did not. There are two options to obtain a completed data of size $n^{(K)}$. The first option applies to fixed design (F).

At each phase, the Y_{NR}, X_{NR} and Y_B are imputed by MICE based on the cumulative respondents Y_R, X_R, Z_R, nonrespondent Z_{NR}, and the benchmark X_B, Z_B.

The second option is to impute the nonrespondents, not only for $(\mathbf{Y}_{NR}^{(K)}, \mathbf{X}_{NR}^{(K)})$ but also for $(\mathbf{Z}_{NR}^{(K)})$. This option is preferred for the BSS-X and BSS-Z designs. Note that Z_{NR} is imputed as described in section 3.2.3. In addition, X values corresponding to Z of the next phase sampling frame is also imputed for BSS-X design, since $\hat{X}$ on the sampling frame is necessary for the next phase sampling.

For all three designs, imputation is carried out by using all available data collected so far. Therefore imputed values at phase k are always replaced by the imputed values at phase $k+1$ for both sampling and inference.

With the finite population generated by the data model 3.5, the procedure of a) drawing a benchmark, b) forming a sampling frame, c) simulating three concurrent focal surveys, and d) imputing nonrespondents, is carried out 1000 times (trials) for each of the 32 simulation scenarios. For each trial, the evaluation measures mentioned in section 3.2.5 are computed. For each simulation scenarios, we report the average quantities from the 1000 trials.

3.3.2 Simulation Results

Recall that the goals are to achieve the convergence of the following distributions:

- Mitigating step: $f(\mathbf{X}_R|\mathbf{Z}_R) \approx f(\mathbf{X}_B|\mathbf{Z}_B)$.
- Imputation step: $f(\mathbf{Y}_R, \hat{\mathbf{Y}}_{NR}|\mathbf{X}_R, \hat{\mathbf{X}}_{NR}, \mathbf{Z}_R, \hat{\mathbf{Z}}_{NR}) \approx f(\hat{\mathbf{Y}}_B|\mathbf{X}_B, \mathbf{Z}_B)$ given that $f(\mathbf{X}_R, \hat{\mathbf{X}}_{NR}, \mathbf{Z}_R, \hat{\mathbf{Z}}_{NR}) \approx f(\mathbf{X}_B, \mathbf{Z}_B)$.

Since the simulations show little difference between the two levels of factor C, i.e. scenarios $cor(X, Y) = Low$ and $cor(X, Y) = High$, we focus on reporting the results of $cor(X, Y) = Low$. We abbreviate the simulation scenario labels to factors A and B, unless otherwise noted. For example, a "LH" indicates factor A $cor(Z, X) = Low$ and factor B $cor(Z, Y) = High$, while factor C is low.

3.3.2.1 Benchmarked sequential sampling

To evaluate M-step objective of $f(\mathbf{X}_R|\mathbf{Z}_R) \approx f(\mathbf{X}_B|\mathbf{Z}_B)$, we compare D_R and D_B on three measures: $P(X > 1|Z > 1)$, $\rho_{z,x}$, and $\bar{Y}$.

We first show the results of $P_R(X > 1|Z > 1) \approx P_B(X > 1|Z > 1)$ from the three sampling designs. We compute $d_{X|Z}$, the difference of $P_R(X > 1|Z > 1)$ and $P_B(X > 1|Z > 1)$. The smaller the $d_{X|Z}$ implies the closer the $P_R(X > 1|Z > 1)$ and $P_B(X > 1|Z > 1)$. Figure 3.2 illustrates the phase IV point estimates and their corresponding confidence intervals for $d_{X|Z}$ for the various simulation scenarios.

In the figure, the point estimates of $d_{X|Z}$ and their corresponding confidence intervals are represented by a horizontal error bar, where the dots are point estimates. For each plot, the x-axis represents the value of $d_{X|Z}$ and the Y-axis shows the combination scenarios of factors A and B, respectively. A zero vertical line is added to aid the visual comparison. In the grand view, the horizontal panels show the nonresponse mechanisms and factor C, and the vertical panels show the sampling designs. For example, the first two horizontal panels are labeled "MAR H" and "MAR L", indicating MAR nonresponse with $cor(X, Y)$ being high and low, respectively.

Under the MAR missingness mechanism, the point estimates of $d_{X|Z}$ are zero among BSS-Z and BSS-X designs while they all deviate from zero for the F design. For MNAR$_X$ and MNAR$_Y$, the confidence intervals cover zero in all but one scenario. The exception is the fixed design with nonresponse pattern MNAR$_Y$ under the LLH scenario. In general, zero tends to fall on the edges of the confidence interval for the F design, whereas zero tends to be the point estimates or close to the point estimates of $d_{X|Z}$ in BSS-X and BSS-Z designs. This pattern is evident by the larger variation of point estimates and several nearly missed CI band noncoverage of the zero line from the F design when comparing to those of the BSS-X and BSS-Z designs. Under MNAR$_X$ and

MNAR$_Y$ $d_{X|Z}$ point estimates range from 0.025 to 0.055 for the F design and range from 0.01 to 0.025 for the BSS-X and BSS-Z designs. Among all simulation scenarios, the performances of BSS-Z is slightly better than those of the BSS-X on $d_{X|Z}$ with respect to the point estimates. However, all their CIs cover zero.

We then assess the preservation of the correlation between Z and X (denoted as $\rho_{z,x}$). Table 3.1 shows the average of 1000 trials on $d_{\rho_{z,x}}$ at phase IV for various simulation scenarios, where $d_{\rho_{z,x}} = \rho_{R(z,x)} - \rho_{B(z,x)}$. The rows show the combination scenarios of factors A by B by nonresponse mechanisms. For example, four rows corresponding to MAR represent LL, HL, LH and HH scenarios of cor(Z,X) and cor(Z,Y) combinations. Columns show the four statistical properties of the estimates, including root mean squared error (RMSE), 95% confidence interval width (CI width), coverage rate (CR), and bias.

Among all simulation scenarios, BSS-X performs better than BSS-Z, and BSS-Z is better than F, with respect to RMSE, CR, and bias. For bias, BSS-X and BSS-Z are in general about half or less of those from the F design. Most strikingly, when $cor(Z, X) = Low$, biases from BSS-X are 0 and 1 under MNAR$_X$ and 2 and 20 under MNAR$_Y$, whereas the corresponding biases for BSS-Z design are 39, 39, 34, 35 and for F design are 88, 90, 75, and 80.

Similarly, for coverage rate, BSS-X and BSS-Z in general are at least two-fold that of the F design. Under MAR missingness, the CRs for BSS-X are 82, 41, 81, 50 for LL, HL, LH, HH scenarios, respectively, and the corresponding CRs for the F design are 32, 3, 32, 3. The greatest benefits of BSS-X are seen under MNAR$_X$ and MNAR$_Y$. When $cor(Z, X) = Low$, BSS-X coverage rates are 95 and 96 for MNAR$_X$ and 94 and 88 for MNAR$_Y$. The corresponding coverage rates for BSS-Z are 55, 55, 66, 62, and for F are 4, 1, 9, 5.

The results of RMSE are similar to those of the bias. With regard to CI width, the results are mixed among three designs, although BSS-Z in general has narrower CI width than F and BSS-X.

Figure 3.3 illustrates four sampling phases of $d_{(\bar{Y}_R,\bar{Y}_B)}$ in horizontal boxplots where $d_{(\bar{Y}_R,\bar{Y}_B)} = \bar{Y}_R - \bar{Y}_B$. The vertical panels show the three sampling designs and horizontal panels show the nonresponse mechanisms. For each plot, the x-axis represents the value of $d_{(\bar{Y}_R,\bar{Y}_B)}$ and the y-axis indicates the combination scenarios of $cor(Z,X)$ and $cor(Z,Y)$ under $cor(Y,X) = Low$.

Each box shows the distribution of $d_{(\bar{Y}_R,\bar{Y}_B)}$ from 1000 trials. Each cluster of four boxes (in different colors) represents the four phases of sampling where phase I is the top box. The most obvious pattern in figure 3.3 is the contrast of stationary (F) versus the shifting (BSS-X and BSS-Z) of the $d_{(\bar{Y}_R,\bar{Y}_B)}$ across sampling phases over all missingness patterns. In all cases, the $d_{(\bar{Y}_R,\bar{Y}_B)}$ of the BSS-X and BSS-Z designs decrease as the sampling phases progress (judging by the boxes shifting toward the zero line) while the $d_{(\bar{Y}_R,\bar{Y}_B)}$ of the F design remains the same across phases within a given scenario. The shifting of the boxes (towards zero) over sampling phases suggests $P(\mathbf{X}_R|\mathbf{Z}_R) \approx P(\mathbf{X}_B|\mathbf{Z}_B)$ for these two designs.

These findings suggest that the BSS-X and BSS-Z designs improve respondent representativeness by better preserving $f(X|Z)$ and $\rho_{z,x}$, even under MNAR missingness. Furthermore, the BSS-X has some advantages over the BSS-Z with respect to achieving $P(\mathbf{X}_R|\mathbf{Z}_R) \approx P(\mathbf{X}_B|\mathbf{Z}_B)$ and preserving $\rho_{z,x}$.

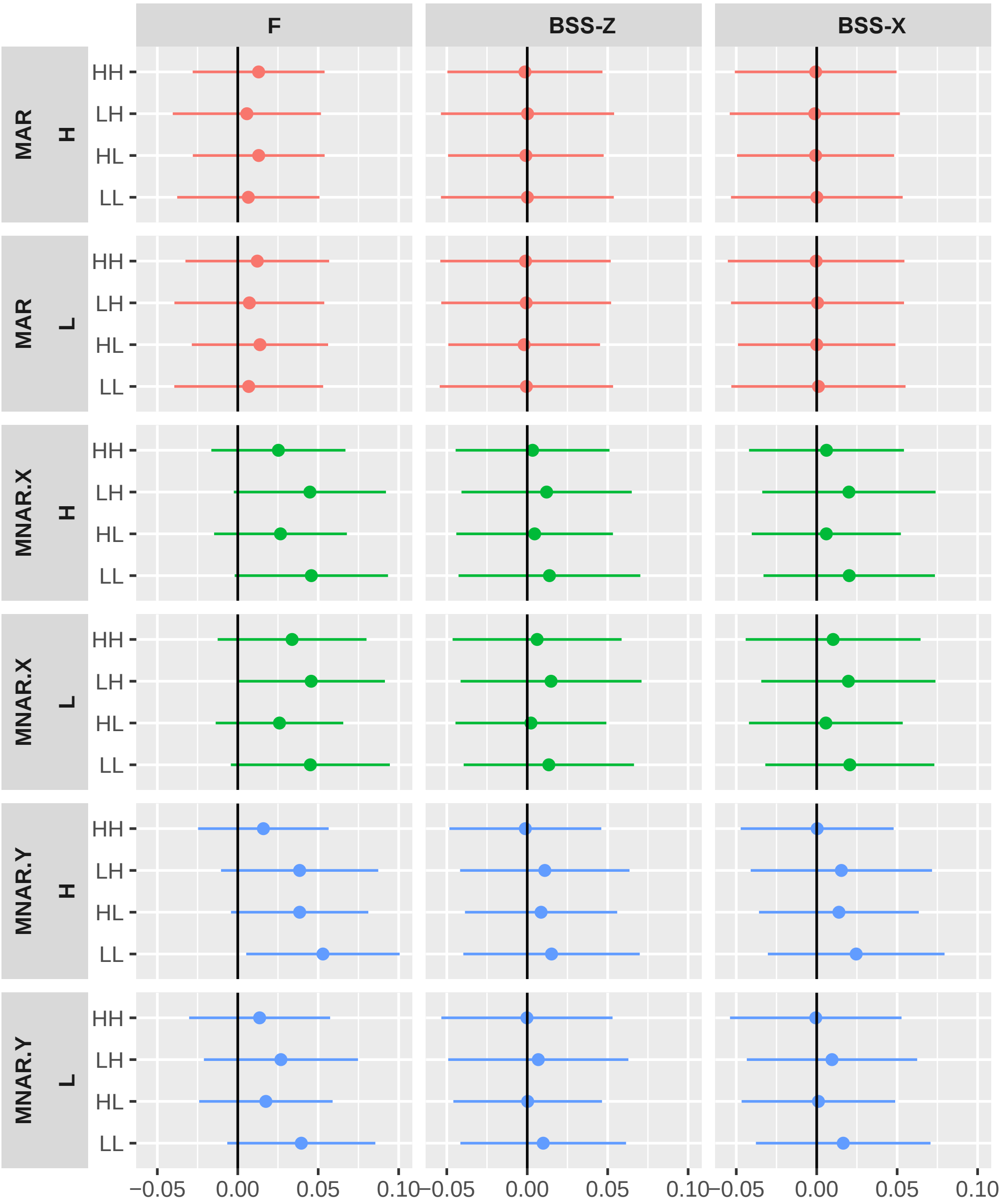
3.3.2.2 Benchmarked multiple imputation

To evaluate I-step objective of $f(\mathbf{Y}_R, \hat{\mathbf{Y}}_{NR}|\mathbf{X}_R, \hat{\mathbf{X}}_{NR}, \mathbf{Z}_R, \hat{\mathbf{Z}}_{NR}) \approx f(\hat{\mathbf{Y}}_B|\mathbf{X}_B, \mathbf{Z}_B)$, we compare D_C and D_B on three measures: $P(Y > 1|X > 1, Z > 1)$, $\rho_{y,x}$, and $\bar{Y}$.

We first assess whether $P_C(Y > 1|X > 1, Z > 1) \approx P_B(Y > 1|X > 1, Z > 1)$ by computing their differences $d_{\hat{Y}|X,Z}$. The smaller the $d_{\hat{Y}|X,Z}$ implies the closer the $P_C(Y > 1|X > 1, Z > 1)$ and $P_B(Y > 1|X > 1, Z > 1)$. Figure 3.4 illustrates phase IV $d_{\hat{Y}|X,Z}$ point estimates and their corresponding 95% confidence intervals for each simulation scenarios. The layout of figure 3.4 is identical to that of figure 3.2.

In all scenarios, the point estimates of $d_{\hat{Y}|X,Z}$ are zero or near zero for BSS-X and BSS-Z designs while they depart from zero for the F design. In several cases, the 95% confidence intervals of $d_{\hat{Y}|X,Z}$ for F design do not cover zero, e.g. LHH and LHL scenarios under MAR and MNAR$_Y$

Figure 3.2: Difference of $P_R(X > 1|Z > 1)$ and $P_B(X > 1|Z > 1)$ by sampling design at phase IV.



Values are point estimates and corresponding 95% confidence intervals. Horizontal panels show nonresponse mechanism (MAR, MNAR$_X$, MNAR$_Y$) and $cor(X, Y)$ (High,Low) and vertical panels show sampling designs. Within each plot, x-axis shows the values of $d_{x|z}$, and y-axis represents predictive power of Z on X and Y (i.e. $cor(Z, X)$ and $cor(Z, Y)$). For example, "LH" indicates a low correlation between Z and X ($\rho_{Z,X} = 0.2$) and a high correlation between Z and Y ($\rho_{Z,Y} = 0.8$). F = fixed sampling design, BSS-Z = benchmarked sequential sampling using Z, and BSS-X = benchmarked sequential sampling using $(Z, \hat{X})$. A vertical zero line is added to aid visual comparison.

Table 3.1: Difference of $\rho_{z,x}$ between D_R and D_B.

		RMSE[1]			**CI width**[2]			**Coverage**[3]			**Bias**[4]		
		F	**AZ**	**AX**	**F**	**AZ**	**AX**	**F**	**AZ**	**AX**	**F**	**AZ**	**AX**
MAR	**LL**	60	35	33	89	85	88	32	74	82	56	27	23
	HL	114	57	54	109	87	88	3	34	41	110	53	49
	LH	59	34	33	88	85	87	32	74	81	54	27	24
	HH	100	53	49	99	84	88	3	39	50	97	48	44
MNAR$_X$	**LL**	91	44	25	88	82	99	4	55	95	88	39	0
	HL	119	59	48	103	82	102	1	25	64	116	55	40
	LH	92	44	27	83	84	107	1	55	96	90	39	1
	HH	115	56	39	97	84	115	1	33	88	112	52	25
MNAR$_Y$	**LL**	78	41	28	86	86	108	9	66	94	75	34	2
	HL	105	54	48	107	83	92	6	33	58	102	49	41
	LH	83	41	31	86	85	94	5	62	88	80	35	20
	HH	105	51	49	105	84	88	4	42	50	101	46	44

Note, for RMSE, CI width, and bias, values shown are average of 1000 trials multiplied by 1000.
AZ = BSS-Z; AX = BSS-X.

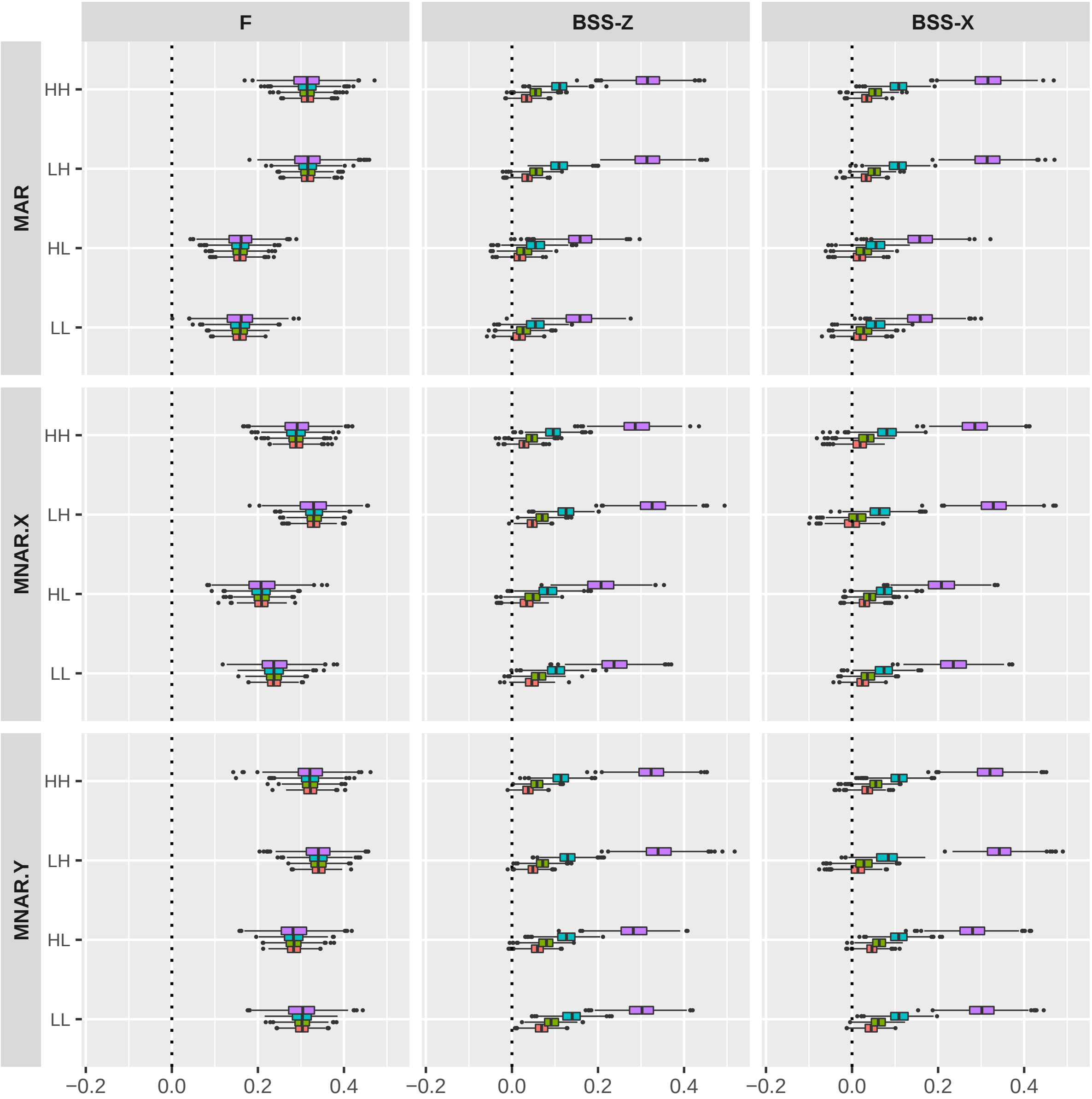
[1] RMSE=$\sqrt{MSE}$, where MSE is computed from equation (3.3).

[2] CI width is computed from equation (3.4).

[3] Coverage is the percentage of 95% CI among r imputations that includes zero. Note: values are multiplied by 100.

[4] Bias = $\sum_r(\rho_{R(z,x)} - \rho_{B(z,x)})/r$.

Figure 3.3: Boxplots for difference of $\bar{Y}_R$ and $\bar{Y}_B$ by sampling design.



Horizontal panels show nonresponse mechanisms and vertical panels show sampling designs. x-axis indicates the values of $d_{\bar{Y}_R,\bar{Y}_B}$, where $d_{\bar{Y}_R,\bar{Y}_B} = \bar{Y}_R - \bar{Y}_B$. y-axis indicates the predictive power of Z on X and Z on Y. For example, "LH" indicates a low correlation between Z and X ($\rho_{Z,X} = 0.2$), and a high correlation between Z and Y ($\rho_{Z,Y} = 0.8$). $\rho_{X,Y} = 0.2$ is held constant across all plots shown. F = fixed sampling design, BSS-Z = benchmarked sequential sampling using Z, and BSS-X = benchmarked sequential sampling using $(Z, \hat{X})$. A vertical zero line is added to aid visual comparison.

missingness. In many cases for F design, zero falls on the boundary of the lower bound of the 95% confidence intervals. Overall, the most striking pattern is the larger variation of $d_{\hat{Y}|X,Z}$ point estimates in F design in contrast to those in BSS-X and BSS-Z designs. This pattern is consistent for all nonresponse mechanisms. One interesting observation is that, not only the point estimates of $d_{\hat{Y}|X,Z}$ for F design depart from zero, the 95% CI does not cover zero for two out of the eight scenarios under MAR missingness. When comparing BSS-X to BSS-Z design, they are comparable with respect to $d_{\hat{Y}|X,Z}$.

The assessment on the preservation of $\rho_{y,x}$ from completed data is shown on table 3.2. Values on the table are the average of $d_{\rho_{y,x}}$ at phase IV over 1000 trials, where $d_{\rho_{y,x}} = \rho_{R(y,x)} - \rho_{B(y,x)}$. The layout of table 3.2 is identical to that of table 3.1. We found that under MAR missingness the three designs are comparable with respect to RMSE, yet the bias and CR from F design trump those of the BSS-X and BSS-Z designs.

For MNAR$_X$, BSS-X and BSS-Z designs have better CI width and CR. BSS-X and BSS-Z designs also have better bias values when $cor(Z, X)$ is low, whereas F design has better bias values when $cor(Z, X)$ is high. Perhaps the greatest advantage of BSS-X and BSS-Z over F design is under MNAR$_Y$ missingness, especially for LL, HL, and LH scenarios, where better values are seen for RMSE, CR and bias. A somewhat surprising finding is that for HH scenario under MNAR$_Y$, F design outperforms BSS-X and BSS-Z designs on RMSE, CR, and bias. When comparing BSS-X to BSS-Z designs, the BSS-X design in general performs better on RMSE, bias and CR. BSS-X and BSS-Z are comparable over CI width.

Figure 3.5 illustrates four sampling phases of $d_{(\bar{Y}_C,\bar{Y}_B)}$ in horizontal boxplots where $d_{(\bar{Y}_C,\bar{Y}_B)} = \bar{Y}_C - \bar{Y}_B$. For each plot, the x-axis represents the value of $d_{(\bar{Y}_C,\bar{Y}_B)}$ and the y-axis indicates the combination scenarios of $cor(Z, X)$ and $cor(Z, Y)$ given that $cor(Y, X) = Low$. The layout of figure 3.5 is identical to that of figure 3.3.

Each box shows the distribution of $d_{(\bar{Y}_C,\bar{Y}_B)}$ from 1000 trials. Each cluster of four boxes (in different colors) represents the four phases of sampling where phase I is the top box. For the fixed design, differences between $\bar{Y}_C$ and $\bar{Y}_B$ are eliminated under MAR. Although these differences remain under MNAR$_X$ and MNAR$_Y$ as the data collection phases progress, the differences are largely minimized when Z has high predictive power on X and Y (HH scenario).

Similar to figure 3.3, the most obvious pattern in figure 3.5 is the contrast patterns of stationary (F) versus the shifting (BSS-X and BSS-Z) on $d_{(\bar{Y}_C,\bar{Y}_B)}$ over sampling phases under MNAR$_X$ and MNAR$_Y$. In these cases, the $d_{(\bar{Y}_C,\bar{Y}_B)}$ of the BSS-X and BSS-Z designs decreases as the sampling phases progress (judging by the boxes shifting toward the zero line) while the $d_{(\bar{Y}_R,\bar{Y}_B)}$ of the F design remains the same over sampling phases for any given scenario.

The most interesting finding is that, under MAR, $d_{(\bar{Y}_C,\bar{Y}_B)}$ for all three designs centered around zero. This is different from the findings in figure 3.3 where $d_{(\bar{Y}_R,\bar{Y}_B)}$ of the F design remains apart from zero over sampling phases.

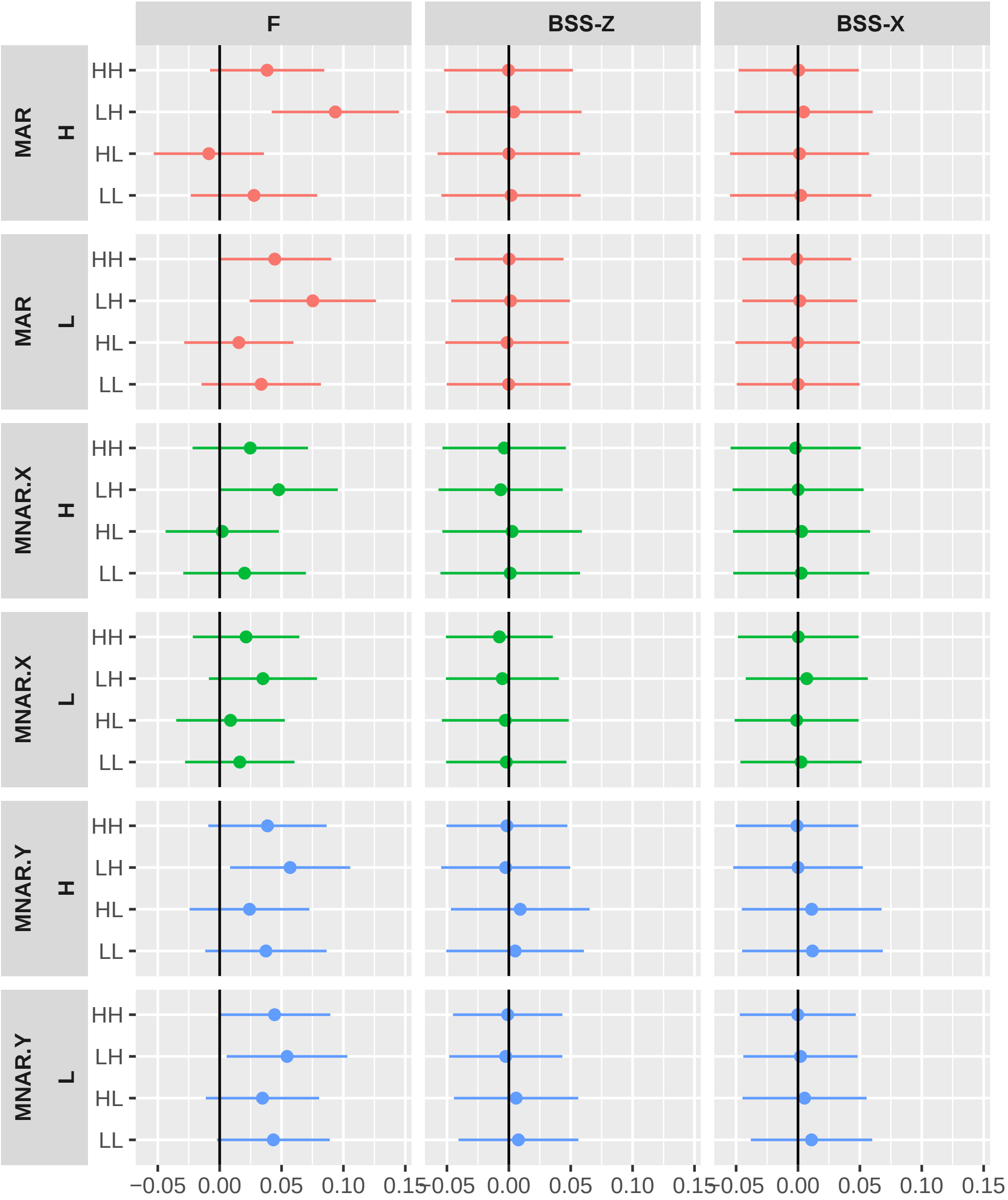
By comparing figures 3.5 to 3.3, we find that imputation step completely eliminates the differences between $\bar{Y}_C$ and $\bar{Y}_B$ for F design under MAR. This is not surprising since imputation under MAR is comparable to weighting adjustments. In addition, under MNAR$_X$ and MNAR$_Y$, the imputation also eliminates almost all of the differences between $\bar{Y}_C$ and $\bar{Y}_B$ for F design in HH scenario. Furthermore, for LH and HL scenarios imputation reduces the differences between $\bar{Y}_C$ and $\bar{Y}_B$ by at least 50% for F design. For LL scenario, the differences between $\bar{Y}_C$ and $\bar{Y}_B$ is reduced by 30%.

3.3.2.3 Accuracy of the estimation

Figure 3.6 illustrates $Bias(\bar{Y}_J)$ in boxplots for scenario $cor(X,Y) = Low$ from the three sampling designs. The layout of this figure is the same as figures 3.3 and 3.5 where each horizontal panel represents a nonresponse scenario and each vertical panel shows a sampling design.

Under MAR, all boxplots are centered around zero indicating that the benchmarked multiple imputation strategy successfully recovers the population information for $\bar{Y}_J$ at all phases regardless of the sampling design. Similar results are found for MNAR$_X$ missingness, although there are

Figure 3.4: Difference of $P_C(Y > 1|X > 1, Z > 1)$ and $P_B(Y > 1|X > 1, Z > 1)$ by sampling design at phase IV.



Values are point estimates of $d_{y|x,z}$ and corresponding 95% confidence intervals. $d_{y|x,z} = P_C(Y > 1|X > 1, Z > 1) - P_B(Y > 1|X > 1, Z > 1)$. Horizontal panels show nonresponse mechanisms (MAR, MNAR$_X$, MNAR$_Y$) and factor C ($cor(X, Y) = High/Low$); vertical panels show sampling designs. Within each plot, x-axis shows the values of $d_{y|x,z}$, and y-axis represents predictive power of Z on X and Y (i.e. $cor(Z, X)$ and $cor(Z, Y)$). For example, “LH” indicates a low correlation between Z and X and a high correlation between Z and Y. F = fixed sampling design, BSS-Z = benchmarked sequential sampling using Z, and BSS-X = benchmarked sequential sampling using $(Z, \hat{X})$. A vertical zero line is added to aid visual comparison.

Table 3.2: Difference of $\rho_{y,x}$ between D_C and D_B.

		RMSE			CI width			Coverage			Bias		
		F	AZ	AX	F	AZ	AX	F	AZ	AX	F	AZ	AX
MAR	LL	20	20	21	79	78	80	95	93	94	1	4	3
	HL	21	22	21	83	78	77	94	91	92	2	8	7
	LH	14	15	14	55	52	52	93	91	92	1	6	5
	HH	14	16	16	54	44	45	92	81	84	1	12	12
MNAR$_X$	LL	24	22	21	83	82	82	91	93	94	11	6	0
	HL	21	22	21	83	83	80	93	93	94	1	7	4
	LH	22	17	15	57	55	56	78	88	94	16	10	3
	HH	14	16	13	52	45	48	93	85	94	3	11	4
MNAR$_Y$	LL	39	25	24	79	80	81	62	87	91	33	15	12
	HL	35	27	26	85	77	77	74	85	86	28	18	17
	LH	22	18	16	58	53	54	81	86	90	16	11	8
	HH	14	18	17	55	46	44	93	79	79	1	13	13

Note, for RMSE, CI width, and bias, values shown are average of 1000 trials multiplied by 1000.
AZ = BSS-Z; AX = BSS-X.

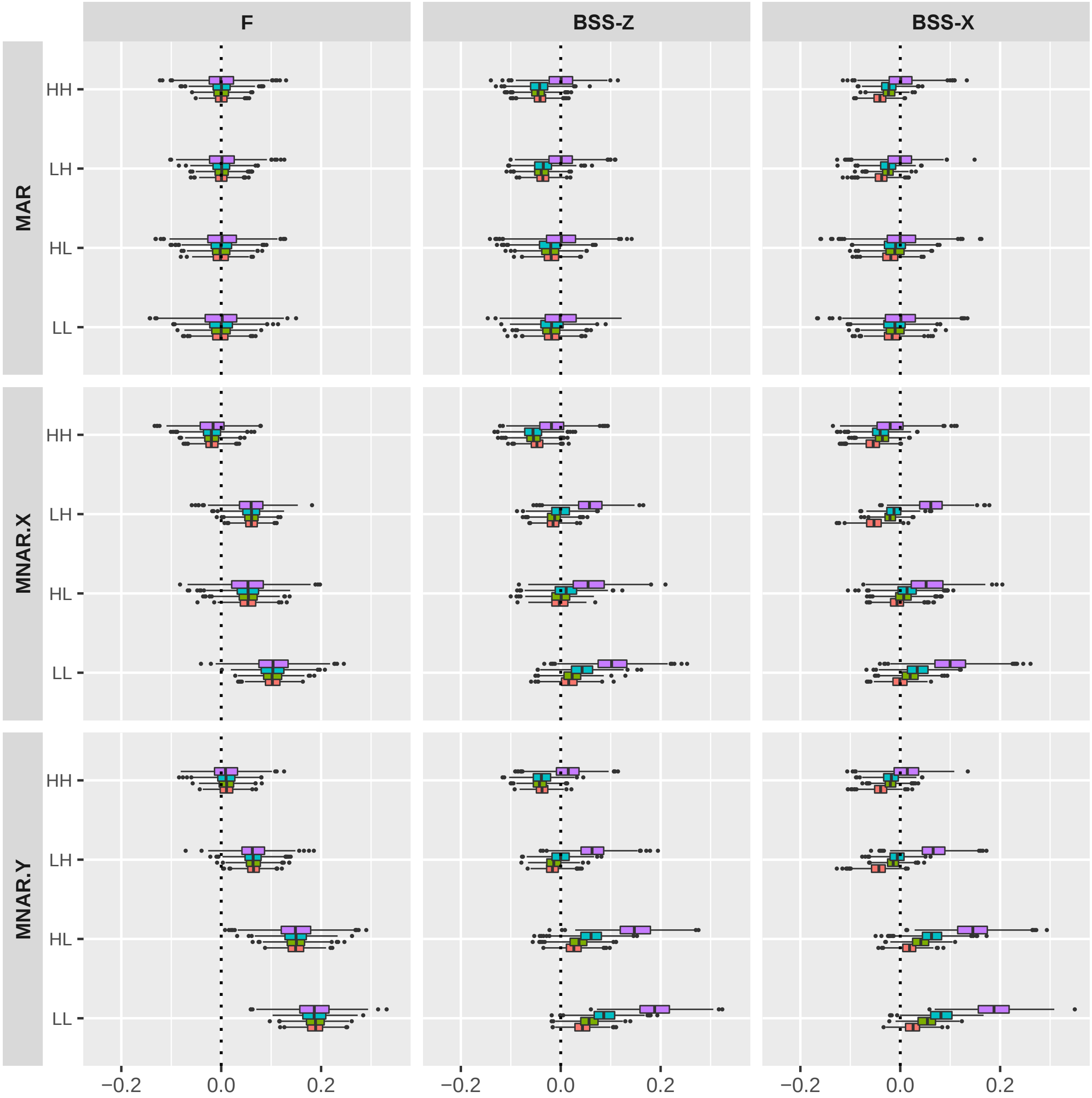
[1] RMSE=$\sqrt{MSE}$, where MSE is computed from equation (3.3).

[2] CI width is computed from equation (3.4).

[3] Coverage is the percentage of 95% CI among r imputations that includes zero. Note: values are multiplied by 100.

[4] Bias = $\sum_r (\rho_{C(z,x)} - \rho_{B(z,x)})/r$.

Figure 3.5: Boxplots for difference of $\bar{Y}_C$ and $\bar{Y}_B$ by sampling design.



Horizontal panels show nonresponse mechanism and vertical panels show predictive power of Z on X and Y. x-axis indicates the values of $d_{\bar{Y}_C,\bar{Y}_B}$. $d_{\bar{Y}_C,\bar{Y}_B} = \bar{Y}_C - \bar{Y}_B$. y-axis indicates the predictive power of Z on X and Y. For example, "LH" indicates a low correlation between Z and X ($\rho_{Z,X} = 0.2$, and a high correlation between Z and Y ($\rho_{Z,Y} = 0.8$). $\rho_{X,Y} = 0.2$ is held constant across all plots shown. F = fixed sampling design, BSS-Z = benchmarked sequential sampling using Z, and BSS-X = benchmarked sequential sampling using $(Z, \hat{X})$. A vertical zero line is added to aid visual comparison.

small biases remain for F design when $cor(Z, Y) = L$.

When nonresponse is related to Y (MNAR$_Y$), the M&I strategy of BSS-X design recovers the population information for $\bar{Y}_J$ when $cor(Z, Y) = H$. When $cor(Z, Y) = L$, minimal bias remains after four phases of data collection. Similar results are seen for BSS-Z design. Under MNAR$_Y$, $Bias(\bar{Y}_J)$ of F design remains, and $Bias(\bar{Y}_J)$ is relatively larger when $cor(Z, Y) = L$ as comparing to when $cor(Z, Y) = H$. The scale of the bias increases with the increase of the sampling phases when $cor(Z, Y) = L$.

Table 3.3 shows the assessment on the preservation of $\rho_{y,x}$ from joint data. This table has the identical layout as Tables 3.1 and 3.2. Values on the table are the average of $d_{\rho_{y,x}}$ at phase IV over 1000 trials, where $d_{\rho_{y,x}} = \rho_{R(y,x)} - \rho_{B(y,x)}$. The results are similar to those of the $Bias(\bar{Y}_J)$ where benchmarked multiple imputation strategy successfully recovers the population $\rho_{y,x}$ under MAR and MNAR$_X$ regardless of sampling design. These findings are evident based on the comparable numbers among three designs in RMSE, CI width, CR, and bias, although bias of $\rho_{y,x}$ for F design is slightly larger than that of BSS-X and BSS-Z when $cor(Z, X) = L$. When comparing BSS-X to BSS-Z designs under MNAR$_X$, the BSS-X design performs better on bias for all scenarios.

For MNAR$_Y$, three designs are comparable for HH scenario. For scenarios LL, LH, HL, BSS-X and BSS-Z designs have noticeable advantage over F design in RMSE, CR, and bias. This finding is consistent with the results of $\rho_{y,x}$ from completed data that the greatest advantage of BSS-X and BSS-Z over F design is under MNAR$_Y$ missingness, especially for LL, HL, and LH scenarios. When comparing BSS-X to BSS-Z designs, BSS-X design performs slightly better on CI width when $cor(Z, X) = H$.

Table 3.4 shows the percent bias of three probability measures: $\mathfrak{M}_1 = P(Y > 1)$, $\mathfrak{M}_2 = P(Y > 1|X > 1)$, and $\mathfrak{M}_3 = P(X > 1|Z > 1)$ from joint data. Under MAR, three designs are comparable although F design are mostly unbiased (for $\mathfrak{M}_1$ and $\mathfrak{M}_3$) whereas BSS-X and BSS-Z often have small residual bias (especially for $\mathfrak{M}_1$ and $\mathfrak{M}_2$).

Under MNAR$_X$, three designs are comparable for $\mathfrak{M}_1$; BSS-X and BSS-Z outperform F for

$\mathfrak{M}_3$; and results are mixed for $\mathfrak{M}_2$. Under MNAR$_Y$, three designs are comparable for $\mathfrak{M}_1$, $\mathfrak{M}_2$ and $\mathfrak{M}_3$ for HH scenario. For LL, LH and HL scenarios, BSS-X and BSS-Z designs have better results on minimizing the bias.

These findings suggest that the benchmarked sequential sampling effectively enhances the respondent representativeness, and the benchmarked multiple imputation when applied to unit nonresponse sufficiently minimizes the nonresponse bias. Overall, the major benefit of the M&I strategy focuses on the MNAR$_Y$ missingness, especially when both or one of the $cor(Z, X)$ and $cor(Z, Y)$ are low.

Table 3.3: Difference of $\rho_{y,x}$ between D_J and population true value.

		RMSE			**CI width**			**Coverage**			**Bias**		
		F	**AZ**	**AX**	**F**	**AZ**	**AX**	**F**	**AZ**	**AX**	**F**	**AZ**	**AX**
MAR	**LL**	20	20	21	79	79	80	94	94	95	1	1	1
	HL	21	20	20	82	78	79	94	94	94	2	2	1
	LH	11	11	11	42	42	42	96	96	96	0	1	0
	HH	7	7	7	26	27	28	98	98	98	0	1	1
MNAR$_X$	**LL**	22	21	21	86	83	82	95	94	94	3	1	0
	HL	21	21	21	82	83	81	94	94	94	1	2	1
	LH	12	11	11	44	44	44	95	96	96	4	2	1
	HH	7	7	7	29	28	27	98	98	99	0	2	0
MNAR$_Y$	**LL**	35	24	24	81	82	83	72	91	91	28	11	11
	HL	36	24	24	82	80	78	71	89	89	29	13	12
	LH	13	11	12	44	42	44	93	95	94	6	3	3
	HH	7	8	7	28	28	27	98	97	98	2	3	2

Note, for RMSE, CI width, and bias, values shown are average of 1000 trials multiplied by 1000.
AZ = BSS-Z; AX = BSS-X.

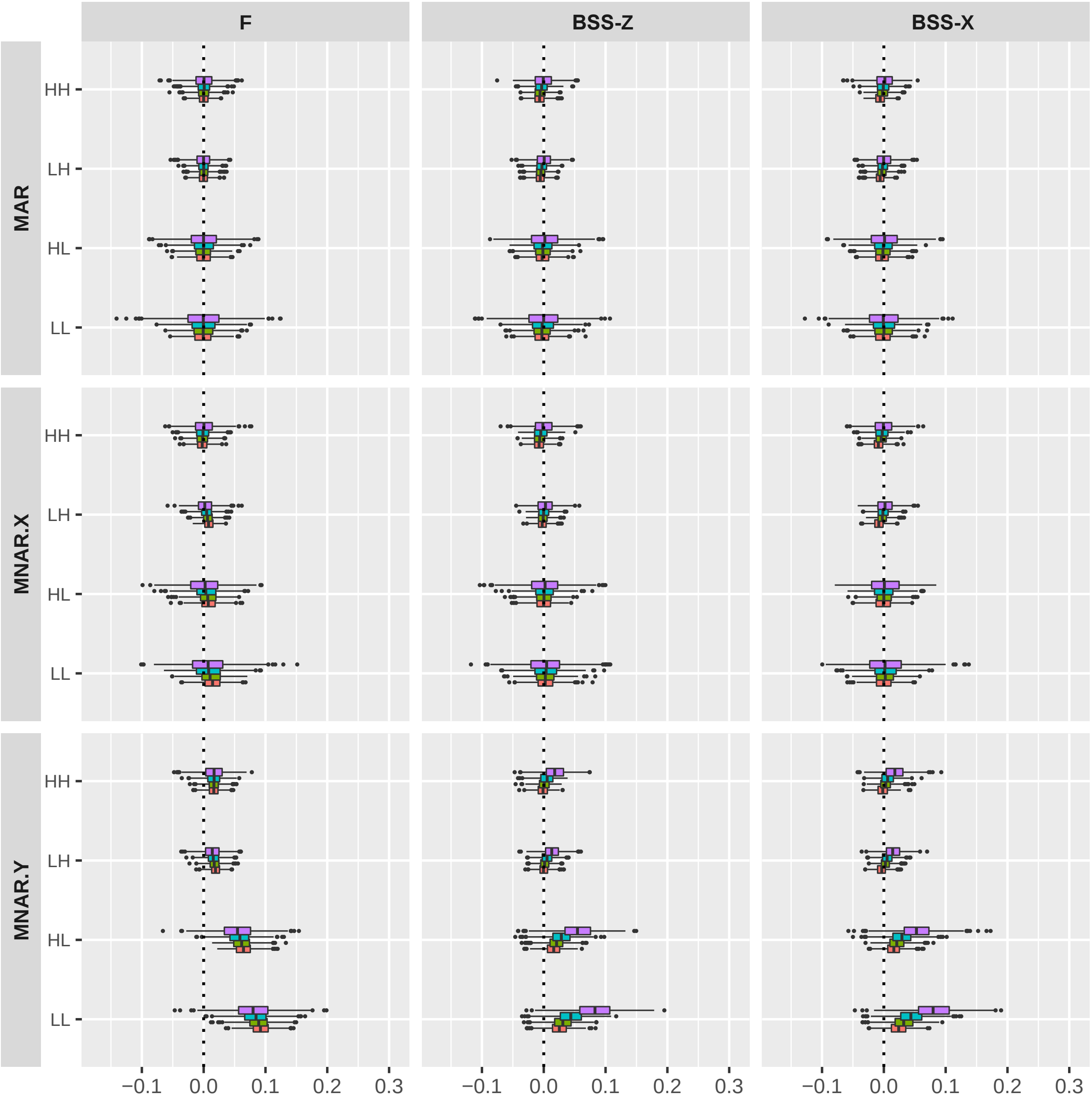
[1] RMSE=$\sqrt{MSE}$, where MSE is computed from equation (3.3).

[2] CI width is computed from equation (3.4).

[3] Coverage is the percentage of 95% CI among r imputations that includes zero. Note: values are multiplied by 100.

[4] Bias = $\sum_r (\rho_{J(z,x)} - \rho_{B(z,x)})/r$.

Figure 3.6: Boxplots for bias of $\bar{Y}$ from joint data ($D_R \cup D_{NR} \cup D_B$).



Horizontal panels show nonresponse mechanism and vertical panels show sampling designs. x-axis indicates the value of $bias(\bar{Y})$ and y-axis indicates the predictive power of Z on X and on Y. For example, "LH" indicates a low correlation between Z and X ($\rho_{Z,X} = 0.2$, and a high correlation between Z and Y ($\rho_{Z,Y} = 0.8$). $\rho_{X,Y} = 0.2$ is held constant across all plots shown. F = fixed sampling design, BSS-Z = benchmarked sequential sampling using Z, and BSS-X = benchmarked sequential sampling using $(Z, \hat{X})$. A vertical zero line is added to aid visual comparison.

Table 3.4: Percent bias on $\mathfrak{M}_1$, $\mathfrak{M}_2$, and $\mathfrak{M}_3$ for joint data.

		$\mathfrak{M}_1$			$\mathfrak{M}_2$			$\mathfrak{M}_3$		
			BSS	**BSS**		**BSS**	**BSS**		**BSS**	**BSS**
		F	**Z**	**X**	**F**	**Z**	**X**	**F**	**Z**	**X**
MAR	**LL**	0	1	1	1	1	2	0	0	0
	HL	0	1	1	1	1	3	0	1	1
	LH	0	2	1	1	2	1	0	0	0
	HH	0	2	2	0	2	3	0	1	1
MNAR$_X$	**LL**	1	0	1	3	1	0	6	3	3
	HL	0	1	1	1	0	2	2	0	0
	LH	1	2	2	5	3	2	7	2	2
	HH	2	3	2	4	4	1	3	1	1
MNAR$_Y$	**LL**	28	7	7	23	7	5	6	1	2
	HL	25	5	6	20	1	5	1	0	0
	LH	4	1	1	1	1	2	3	1	1
	HH	1	1	1	1	1	1	0	0	0

Note: $\mathfrak{M}_1 = P(Y > 1)$, $\mathfrak{M}_2 = P(Y > 1|X > 1)$, and $\mathfrak{M}_3 = P(X > 1|Z > 1)$.

3.4 Application to CPS Data

Our goal is to examine the relative performances of the three sampling designs in situations that are likely to occur in practice. We conducted a repeated sampling experiment using an extract of the public-use microdata sample from the 2010 Current Population Survey (CPS), 2010 American Community Survey (ACS), and the Census Planning Database (PDB).

3.4.1 The Sample Population

CPS and ACS share many common variables and have similar target populations. Both use the Master Address File (MAF) as the sample frame and together they represent a sensible example to implement our proposed framework. In this practical application, CPS is considered the focal survey, ACS is the benchmark survey, and PDB contains the frame information that is shared for the two surveys. Our goal is to recover the population information (e.g.Census estimates) by imputing the CPS nonresponse through MICE using CPS respondent data, benchmark data (ACS) and the frame data (PDB).

We first line up the sample domains between ACS and CPS and excluded subjects who reside in only one of the ACS or CPS domains. For example, in ACS, subjects from Puerto Rico and/or living in group quarters are excluded from our sample because CPS does not collect data from these special populations. We further excluded subjects who are younger than 18 years old and/or have missing value data on the PDB variable "percent mail-out mail-back". This leaves a combined ACS and CPS data with $N = 666,188$ individuals. A random sample of 50% (N=333,094) is selected for the analysis, serving as the sample population (i.e., sampling frame).

The selected five X variables shared by ACS and CPS are age, sex, race (white, black, and other), education (less than high school, high school and more than high school), and Hispanic origin (yes vs no). The age variable is continuous with an average of 49.5 years old. The other four variables (Sex, race, education, Hispanic origin) are categorical. These variables are chosen because they are typical of weighting models applied to this type of survey. There are 19 selected frame variables (Z), all are continuous quantities (percentages) except for geographic region (4 levels). (See Table 3.5).

The original designs for both ACS and CPS are multi-stage stratified samples. For our illustration, we treat the data as a simple random sample. We combine the ACS and CPS data as the sampling population. PDB is based on the Census geographical boundaries, data are subject to non-sampling error, whereas ACS data are subject to sampling error. For our purpose, we consider PDB data as true values.

3.4.2 Study Designs

Of the $333,094$ subjects in the sample population we randomly selected 5% ($n_B = 16,655$) as the benchmark survey. The remaining 316,439 subjects comprise the sample frame for the focal surveys. Three concurrent focal surveys (F, BSS-X, BSS-Z) draw samples independently from this sample frame. Each focal survey has four sampling phases and the sample size for k^{th} phase is $n_k = 1000$ for the fixed design, where $k = 1, \ldots, 4$; $n_1 = 1000$ and $n_k = 1200$ for the adaptive sampling designs (BSS-X, BSS-Z), where $k = 2, 3, 4$.

Four outcome variables are created such that X and Z variables have different predictive strength for Y. Table 3.6 lists regression coefficients for the four models that created the four outcome variables $Y_1, \ldots, Y_4$. The last row on table 3.5 indicates the R-squared of all the variables on Y. Table 3.6 lists the predictive power of X-variables and Z-variables for each Y. These Y variables can be considered as index of well-being.

Three schemes of unit nonresponse mechanisms are evaluated: MAR, MNAR$_X$ and MNAR$_Y$. To create unit nonresponse, response probabilities are generated using the coefficients listed below.

1. $logit(e(\mathbf{Z})) = 0.01 + \mathbf{Z}\beta$ **(MAR)**

2. $logit(e(\mathbf{Z}, \mathbf{X})) = 0.01 + \mathbf{Z}\beta + \mathbf{X}\gamma$ **(MNAR$_X$)**

3. $logit(e(\mathbf{Z}, \mathbf{X}, y)) = 0.01 + \mathbf{Z}\beta + \mathbf{X}\gamma + 0.15y$ **(MNAR$_Y$)**

In the three models above, **Z** is an $n \times 18$ matrix, **X** is an $n \times 7$ matrix, and y is an $n \times 1$ vector. The columns of **Z** and **X** correspond to the variables listed in Table 3.5. For the all three models, β is an 18×1 vector and γ is a 7×1 vector. For the MAR model, all elements of β are equal to 0.17. For the MNAR$_X$ model, all elements of β and γ are equal to 0.16. For the MNAR$_Y$ model, all elements of β and γ are equal to 0.15. The response status for sampled cases is simulated by drawing Bernoulli random variables based on these response probabilities, creating an average of 50% unit nonresponse for each nonresponse mechanism.

At the imputation step, all imputation models include Y and all X and Z variables. The continuous variables, other than age and Y$_s$, are standardized by their corresponding mean and sd and modeled as a normal distribution. Age and Ys are modeled using predictive mean matching (pmm) because normal transformed data resulted in poor predictions. The binary variables, sex and Hispanic origin, are predicted using logistic regression. Categorical variables, including geographic region, education and race, are predicted using polytomous regression. Imputation diagnostics include the scatter plot of imputed Y versus respondent Y values, residuals plotted against propensity scores, and density plots of imputed vs benchmark data. Five imputations are implemented and each imputation runs for 10 iterations.

The experiment is conducted by first drawing simple random sample (s.r.s) as the benchmark using the population described in section $3.4.1$. Once the benchmark is selected, the phase samples of focal surveys are drawn from the sample frame, excluding the benchmark subjects. Response status (respondent/nonrespondent) is assigned to each sampled unit according to the given response scheme. The MICE procedure imputes the unit nonresponse of focal survey phase sample for each of the M = 5 imputed datasets. The new imputation is carried out after each phase of data collection, using the cumulative respondent data, discarding previous imputed data. The estimated mean values of Y are computed for the respondents, completed data, and joint data and variance are estimated using the MI variance estimation method suggested by Rubin (1987). The entire process is repeated 100 times (trials) for each nonresponse scheme.

For the evaluation of M&I strategy we report on the bias of the following measures.

1. $\bar{Y}_j$.

2. $\mathfrak{C}_1$: $\%Y_j > \mu_j$, where μ_j denotes population mean for Y_j.

3. $\mathfrak{C}_2$: $\%$ age> 50 and $Y_j > \mu_j$, and

4. $\mathfrak{C}_3$: $\%$ Black, college degree, and $Y_j > \mu_j$.

where $j = 1, \ldots, 4$. These estimates are averaged over 100 replicates for each of the following data sets: respondents, completed data, and joint data.

Table 3.5: CPS example. Population coefficients that created Y.

	Y_1	Y_2	Y_3	Y_4
Intercept	121.023	-484.399	-13.861	-575.622
X variables				
Age	1.831	1.834	4.078	4.077
Sex	2.356	1.615	3.803	3.755
Hispanic origin	1.315	2.346	3.801	3.588
Race White	1.33	2.091	4.082	4.367
Race Black	1.42	1.729	4.185	5.675
Education HS	0.391	1.936	3.647	5.086
Education College	0.982	3.127	3.467	4.541
Z variables				
Region MW	-0.996	-0.323	0.175	0.797
Region S	-0.709	-0.108	-0.054	0.31
Region W	0.568	0.961	-0.901	0.577
Female(%)	0.053	0.139	0.05	0.129
under 5(%)	0.029	0.027	-0.02	0.163
65plus(%)	0.093	0.18	0.08	0.169
Hisp(%)	0.073	0.157	0.069	0.163
Blk alone(%)	0.081	0.154	0.07	0.166
AIAN alone(%)	0.08	0.116	0.083	0.161
Asian alone(%)	0.067	0.154	0.084	0.161
NH NHOPI alone(%)	0.083	0.169	0.03	0.123
NH SOR(%)	-0.249	0.097	-0.03	0.184
Female No HB(%)	0.047	0.28	0.112	0.149
Prns in HHD(%)	0.074	0.162	0.073	0.165
Rel Child Under 5(%)	0.087	0.234	0.118	0.162
Vacant Units CEN 2010(%)	0.06	0.157	0.077	0.166
MailBack Area Count(%)	0.083	0.168	0.068	0.167
TEA Mail Out Mail Back(%)	0.072	0.163	0.072	0.159
R^2	0.098	0.296	0.296	0.492
Population mean	658.994	592.882	627.626	612.805

Table 3.6: CPS example. Population R^2 between Y and X and Z variables.

Scenarios	$R^2(Y\|Z)$	$R^2(Y\|X)$
Y_1 LL	0.05	0.05
Y_2 LH	0.05	0.25
Y_3 HL	0.25	0.08
Y_4 HH	0.24	0.27

3.4.3 Results

Figure 3.7 illustrates results of $Bias(\bar{Y})$ in boxplots from joint data of the CPS application. The layout of this figure is identical to that of figures 3.3, 3.5, and 3.6. The findings are consistent with those of the simulation study on multivariate normal data.

The benchmarked sequential sampling, BSS-X and BSS-Z, successfully improved respondent representativeness, judging by the declining $bias(\bar{Y})$ from respondent data as the sampling phases progress. The benchmarked multiple imputation successfully reduces the $bias(\bar{Y})$ for all three sampling designs on completed data, yet the degree of bias reduction is greater for BSS-X and BSS-Z designs, especially for MNAR$_Y$ missingness under LL, LH, and HL scenarios when comparing to the F design.

For joint data, the population estimates of $\bar{\mathbf{Y}}$ are recovered through benchmarked multiple imputation for MAR and MNAR$_X$ missingness, regardless of sampling design and the degree of respondent representativeness. The greatest benefit of the M&I strategy, however, occurs under the MNAR$_Y$ missingness where the improvement in respondent representativeness through M-step demonstrated its return by minimizing the bias of $\bar{Y}$ through I-step, regardless of the predictive power of $\mathbf{X}$ and $\mathbf{Z}$ on Y. On the other hand, the I-step followed by the F design reduces $Bias(\bar{Y})$ for MNAR$_Y$ if $cor(Z, Y) = H$. BSS-X and BSS-Z designs are comparable with respect to $Bias(\bar{Y})$ from joint data.

Table 3.7 shows the bias of three percentage measures, $\%$ $Y_j > \mu_j$, $\%$ (age > 50 and $Y_j > \mu_j$), $\%$ (Black College Graduates and $Y_j > \mu_j$), computed from respondent, completed, and joint data, where $j = 1, \ldots, 4$. As expected, we found that when looking at respondent data, BSS-X and BSS-Z designs perform better than F design, while the results are mixed when comparing BSS-X and BSS-Z. The results depend on the measures, the missingness mechanism, and predictive power of Z and X on Y. For example, the results for $\%Y > \mu$ shows that BSS-X produces smaller bias than BSS-Z under MAR, yet BSS-X produces larger bias than BSS-Z under MNAR$_X$. For MNAR$_Y$, BSS-X produces smaller bias when X has lower predictive power on Y (Y_1, Y_3) whereas

BSS-Z produces smaller bias when X has higher predictive power on Y (Y_2, Y_4).

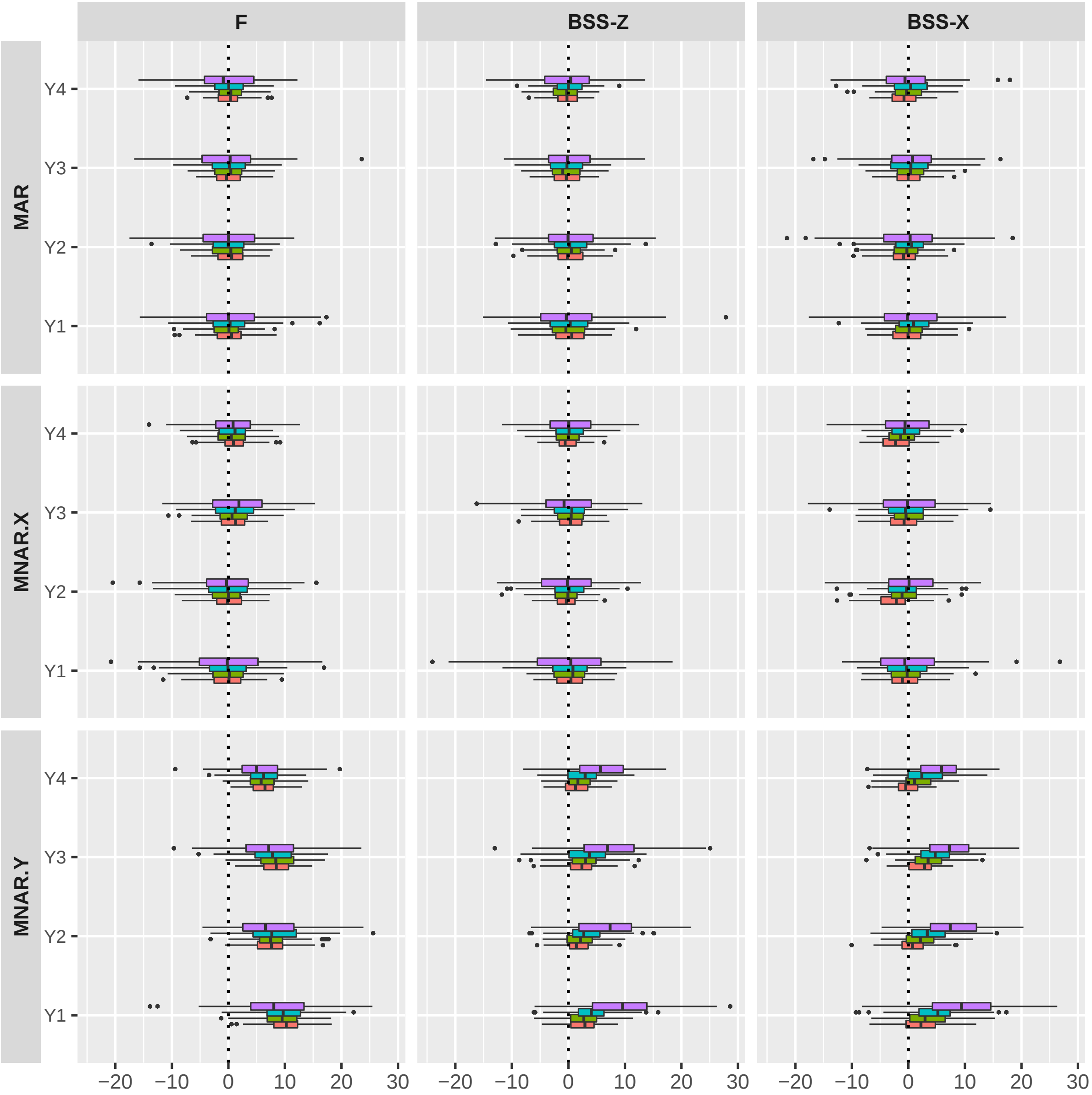
When looking at the results from the completed data, F design in general has an advantage over BSS-X and BSS-Z under MAR. BSS-Z has advantage over BSS-X under MNAR$_X$. For MNAR$_Y$, BSS-X in general reduces the bias more when $cor(Y, X) = L$ whereas BSS-Z reduces more bias when $cor(Y, X) = H$. Finally, estimates from joint data show comparable results under MAR and MNAR$_X$ over three sampling designs. The greatest benefit of the M&I strategy appears for joint data under MNAR$_Y$ missingness, although the results also depend on the particular estimates.

3.5 Discussion

In this chapter, a "design with the estimation in mind" strategy is presented. Our goal is to reproduce the population estimates under unit nonresponse by using a benchmark survey. Our two main objectives are 1) improving respondent representativeness and 2) recovering population information under unknown nonresponse mechanism. The first objective is accomplished by the reduction in the level of differences between respondents and the benchmark through benchmarked sequential sampling. The second objective is achieved by minimization of nonresponse bias in estimates derived from benchmarked multiple imputation. By capitalizing on the benchmark (the auxiliary variables), we provide a new paradigm for survey inference that leads to more robust statistics, eliminating bias not only under ignorable nonresponse but also under one type of non-ignorable nonresponse (MNAR$_X$).

Guided by the benchmark, a propensity score based sampling probability is used to monitor and tailor the sampling decision sequentially, attenuating the impact of undesirable nonresponse patterns. We show that adaptive sampling mediates the undesirable nonresponse and achieves $f(Z_R, X_R) \approx f(Z_B, X_B)$. Later, we implement MICE to perform the benchmarked multiple imputation. Imputation models iteratively fit to the benchmark data (X_B, Z_B) and the respondent data (Y_R, X_R, Z_R), predicting missing information from unit nonresponse to construct a completed

Figure 3.7: CPS example. Bias of $\bar{Y}$ from joint data.



Horizontal panels show nonresponse mechanism and vertical panels show sampling designs. F = fixed sampling design, BSS-Z = benchmarked sequential sampling using Z, and BSS-X = benchmarked sequential sampling using $(Z, \hat{X})$. A vertical zero line is added to aid visual comparison.

Table 3.7: CPS example. Bias of $\mathfrak{C}_1$, $\mathfrak{C}_1$, $\mathfrak{C}_1$ for respondent, completed and joint data.

		Respondent			**Completed**			**Joint**		
		F	**AZ**	**AX**	**F**	**AZ**	**AX**	**F**	**AZ**	**AX**
Bias($\mathfrak{C}_1$)										
MAR	**LL**	22	6	0	0	2	5	1	0	2
	HL	22	5	5	1	1	2	0	0	2
	LH	55	13	1	2	2	10	1	0	1
	HH	54	11	1	2	2	15	1	1	3
MNAR$_X$	**LL**	26	6	18	5	1	16	0	3	3
	HL	33	9	16	14	2	13	2	0	2
	LH	53	13	46	4	3	41	0	0	8
	HH	63	14	44	14	2	36	3	2	7
MNAR$_Y$	**LL**	53	13	9	34	8	8	29	7	6
	HL	59	17	3	41	11	1	24	7	6
	LH	79	18	39	28	5	32	22	6	2
	HH	86	20	37	37	6	31	19	5	0
Bias($\mathfrak{C}_2$)										
MAR	**LL**	11	4	1	1	0	1	1	1	2
	HL	9	2	2	1	0	2	2	0	0
	LH	25	7	0	1	2	5	0	0	1
	HH	25	6	0	0	0	8	1	0	2
MNAR$_X$	**LL**	24	7	2	14	4	3	2	1	2
	HL	28	7	1	18	5	1	2	0	1
	LH	35	10	17	12	4	15	1	1	3
	HH	42	10	15	18	2	13	3	0	3
MNAR$_Y$	**LL**	41	10	3	33	9	2	16	4	2
	HL	47	14	7	38	10	6	15	4	3
	LH	52	13	14	28	8	10	13	3	1
	HH	61	14	10	37	8	9	12	3	1
Bias($\mathfrak{C}_3$)										
MAR	**LL**	7	0	0	0	2	1	0	1	0
	HL	6	1	1	0	1	1	0	0	0
	LH	7	1	1	0	0	1	0	0	0
	HH	8	1	0	0	0	2	0	0	0
MNAR$_X$	**LL**	7	1	4	1	0	5	0	0	1
	HL	6	1	4	1	0	4	0	0	1
	LH	8	0	6	0	1	6	0	0	1
	HH	9	1	6	1	1	5	0	1	1
MNAR$_Y$	**LL**	8	1	4	2	0	4	1	0	0
	HL	8	2	4	3	1	4	2	1	0
	LH	9	1	6	1	0	5	1	0	1
	HH	9	2	5	2	0	5	1	1	1

data $(\hat{Y}_C, \hat{X}_C, \hat{Z}_C)$ of size n_s. Under MCAR (not shown) and MAR (Response model 2: $R \sim Z$),$f(Y_B, X_B, Z_B) = f(\hat{Y}, \hat{X}, \hat{Z})$, is the direct result of imputation framework. Yet this result also holds for response model 3 where $R \sim (X, Z)$. Lastly, we show that under response model 4: $R \sim (Y, X, Z)$, the bias was completely eliminated when Z has strong correlation with Y (scenarios LH and HH); when Z has weak correlation with Y, $f(\hat{Y}, \hat{X}, \hat{Z})$ still preserves a high level of population structure with respect to the marginal distribution, joint distribution, and the correlation structure (see Table 3.3), although biases of the estimates are not completely eliminated (see Table 3.4).

The benefit of this strategy is twofold: 1) through each sampling phase, survey respondents become more representative of the target population, attenuating undesirable nonresponse, leading to less nonresponse bias, 2) benchmarked multiple imputation models build on representative respondent data and are better able to recover the population structure. Note that for the first benefit, it is true even if nonresponse occurs in a rare population (e.g. Black and Hispanics) as long as there are enough of them on the sampling frame to be selected from. It is also true for non-proportionally allocated or non-self-weighting samples as possibilities for the focal survey. In such cases, focal survey respondents need to be weighted appropriately before the assessment of $P(\mathbf{X}_R|\mathbf{Z}_R) \approx P(\mathbf{X}_B|\mathbf{Z}_B)$ and the sampling probability in (3.2) can be extended with higher technical complexity to include design weights.

Benchmarked sequential sampling using both Z and X does not out-perform the benchmarked sequential sampling using Z alone. This is likely due to the fact that imputations do not reflect the real values but a distributional information. Our sampling strategy requires accurately selecting the targeted subjects in the micro-data level which call for knowing the real value of $\mathbf{X}$. Our results show that sampling by matching the observed Z values appear to be comparable to sampling by Z and $\hat{X}$.

Previous research on unit nonresponse imputation has not produced definitive results when compared to the conventional nonresponse weighting adjustments. The main problem is that the imputation models are built upon respondent data alone, whereas weighting adjustments take advantage of data $\mathbf{Z}$ that are also known for the unit nonrespondents. Imputation may recover the nonrespondent information when nonresponse is missing at random. However, when missingness is non-ignorable, the differential distributions for respondents and nonrespondents undermine the ability of imputation to effectively restore nonrespondent information. The proposed research overcomes this pitfall by employing an external micro-level benchmark that captures the data structure of the target population, hence delivers desired results.

The proposed framework has the following limitations. (1) As with any simulation, the findings are restricted to the situations posited in this study. The simulation experiment used only a few variables from ACS and the Census PDB. A common survey has a lot more variables and the intercorrelation among variables would be more complex. (2) Although simulation scenarios included MAR and MNAR nonresponse mechanisms, the nonresponse mechanisms are relatively simple, with known structures. The findings in this study are important but more simulation conditions are needed to generalize the results. (3) The MNAR nonresponse mechanism was studied in a single continuous outcome variable. Analysis on multivariate outcomes with a mixture of distributions may lead to different conclusions. (4) In this study, the sample size of the focal survey is merely a small fraction (5%) of the benchmark survey. Such sample size differences implicate a small fraction of predictions are computed based on a set of stable chained equation models, which warrants easy iteration convergence. Clearly, in practice these sample size conditions may not be attainable leading to imprecise prediction models and convergence difficulties. (5) Finally, the key ingredient in the proposed framework is a high quality benchmark that resembles target population of interest and shares many common variables with the focal survey. Such a benchmark is only a proxy of the population and may need to be constructed from several sources.

Given these limitations, mediating nonresponse mechanisms by benchmarked sequential sampling can be recommended for surveys with suitable benchmark information, especially when the proportion of sample size in focal survey is small relative to that of the benchmark, and the number of variables are moderate (around 10-20) for constructing the model fitting. To improve prediction on focal survey data, strategies recommended for the implementation of MICE algorithm for missing data imputation are readily applicable for the proposed framework and should be diligently followed (Abayomi, et. al. 2008; He et. al., 2009; Schafer, 1999, 2003; Stuart, et. al. 2009; White, et. al., 2011).

Our research has important implication on the era of increasing organic data, data not collected from probability samples. Combining data from various sources to produce information not available from either data sources is not only inevitable but also sensible with respect to time and cost. The proposed new inferential paradigm is simple, straightforward and readily applicable with current statistical software.

In conclusion, a micro-level benchmark provides an opportunity to improve population estimates through benchmarked sequential sampling and imputation. And the benchmarked multiple imputation by chained equations is a feasible and effective way to recover population regardless of the nonresponse mechanism. This paper has outlined the use of a micro-level benchmark for benchmarked sequential sampling and benchmarked imputation by chained equations for prediction on focal survey nonrespondents, focusing on the utility of improving population estimates for nonresponse mechanism investigated here. Further research is warranted to investigate the flexibility of the proposed method with respect to more complex nonresponse mechanisms.

CHAPTER 4

Cost and Error Evaluation

Improving respondent representativeness by adaptive sampling designs is associated with reducing nonresponse bias. Some speculate that reduced bias is at the expense of increased cost. This chapter uses simulations to investigate cost and error of two adaptive sampling designs and compare their performances with the conventional fixed design.

The simulations study the cost-effectiveness (*efficacy*) for weighted and imputed estimates of a sample mean under different unit nonresponse mechanisms, including missing at random (MAR) and missing not at random (MNAR). The cost-effectiveness measure "*Efficacy*" is derived from root mean square error and cost, where cost is estimated from a new subject-level cost model. RMSE of a sample mean is computed from two estimation methods: one uses a conventional calibration weighting estimator and the other uses benchmarked multiple imputation estimator. Other evaluation measures include RMSE per unit cost, bias and coverage rate. The comparison is done between the new design of adaptive sampling and imputation and the traditional design of fixed sampling and weighting (generalized regression estimator).

4.1 Introduction

This chapter concerns cost and errors from unit nonresponse when implementing adaptive survey design by benchmarked sequential sampling methods. We use simulation studies to compare error per unit cost (PUC) and cost-effectiveness (*efficacy*) from fixed and adaptive designs when the MAR assumption required by weighting methods does and does not hold. The error here is nonresponse error, but implicitly also encompasses coverage and sampling errors. This is fundamentally different from the classical optimizing problem on cost and error that minimizes sampling error for a fixed total survey cost (Groves, 2004). Our goal is to examine the relative performances of different sample designs under the same set of auxiliary variables and experimental conditions. Nonresponse error here is defined as root mean square error (RMSE) of the sample mean, a measure combining bias and variance. A subject-level cost model is developed. A simple random sample (s.r.s) of a four-phase single stage survey from a simulated population is used to evaluate these designs and estimation methods.

Two types of adaptive survey designs appear in the recent literature, one targets sampling probability and the other focuses on data collection. Adaptive data collection design is primarily a nonresponse follow-up strategy which applies differential data collection protocols to different subgroups of the nonrespondents (Groves and Heeringa, 2006; Wagner, 2008; Schouten, et al., 2009). Adaptive sampling design obtains representative respondents through differential sampling probabilities among over- and under-represented subjects while maintaining a coherent data collection protocol (see Chapters 2 and 3). Both types of adaptive designs aim at improving respondent representativeness, reducing nonresponse bias. This chapter focuses on adaptive sampling designs.

Adaptive sampling designs by benchmarked sequential sampling methods as described in Chapters 2 and 3 improve respondent representativeness by capitalizing on a micro-level benchmark data which serves as a surrogate for the unobserved target population of interest. The benchmark data shares sample frame and survey covariate information with the focal survey. One design (BSS-Z) derives its sampling probability by matching the benchmark frame information, and the other design (BSS-X) matches both frame and survey covariate information. In the multi-phase setting where new sample replicates are introduced at each phase, the benchmarked sequential sampling adjusts the sampling probability such that cumulative distributions of respondent characteristics for the focal survey converge to those for the benchmark data. Chapter 3 shows that both adaptive sampling designs outperform conventional fixed design with respect to bias of descriptive survey measures under ignorable and nonignorable nonresponse.

A general concern for adaptive sampling designs is their cost. For example, Section 2.3 demonstrated a simulated population with an average true response rate of 50%. Samples drawn by the adpative sampling designs consisted of more subjects with lower response propensities, hence yielded lower response rates. Furthermore, to reach a comparable respondent size the total sample size of adaptive sampling design is 15% larger than that of the fixed sampling design. While subjects with lower response propensity are costlier to obtain a response, larger samples cost more. Yet these incurred higher costs are prompted by the intent of bias reduction. To evaluate whether the reduced bias offsets the increased cost, we compute the error per unit cost. Ultimately one prefers a design that not only produces smaller error per unit cost but also costs less. A cost-effectiveness scale, *efficacy*, is derived to consider both cost and error where a larger *efficacy* represents a design of better value. We suggest that, when designing a survey, *efficacy* provides a more sensible assessment.

However, the use of cost models to design the sample is uncommon. Judkins, Waksberg, and Northrup (1990) reported difficulties in using existing cost data from an ongoing survey to construct a model for National Health Interview Survey (NHIS), a three-stage probability sample of households. They discussed the differences of planning purpose cost data and administrative cost data, in addition to the scarcity of detailed information on survey costs. Shimizu, et. al., (2001) illustrated an example of approximating variable costs using NHIS administrative data, where many assumptions were made to assign the aggregated administrative cost data to the case level. Groves (2004) points out that the survey costs in reality are often nonlinear functions of sample size, discontinuous over large sample ranges, vary stochastically and have limited domain of applicability; making it not generalizable from one survey to another. Given these challenges, our goal is to establish a cost model for the purpose of comparing the sampling design and estimation methods, instead of developing a cost model for sample design purpose. We limit our discussion to survey costs that occur before nonresponse follow-up.

While the cost model has its challenges, variance estimation is relatively straightforward. Variance estimation is tied to the inference methods, most commonly weighting and imputation, for survey estimates with unit nonresponse (Kalton and Maligalig, 1991). We study variance estimators for estimates of a sample mean from both weighting and imputation strategies. Both strategies share the same objective of compensating for missing data and of making inferences from the sample to the target population. A conventional weighting strategy is implemented by a sequence of weights (design weight, nonresponse weight and a calibration method) that applies to the respondent data to reflect the variation of sample design, nonresponse and population structure (Kish, 1965; Deville and Särndal, 1992; Brick and Kalton, 1996; Valliant, et. al., 2013). The benchmarked multiple imputation (MI) is implemented by appending benchmark data to respondent data and imputing the nonrespondent data using multivariate imputation by chained equation (MICE) (see Sect. 3.2.3). Both methods have been applied to unit nonresponse, although weighting strategy is the current norm (Brick and Kalton, 1996; Rässler and Schnell, 2004; Brick and Jones, 2008) Weighting methods to adjust for unit nonresponse produce unbiased survey estimates under MAR. In the case of informative auxiliary data, Sect. 3.3.2.2 show that benchmarked multiple imputation for unit nonresponse is effective in removing bias especially under MNAR.

In Section 4.2, we describe methods of adaptive sampling design, variance estimation with weighted and imputed data, the cost model under consideration, and evaluation measures. In Section 4.3, we outline the study population, the sample designs used in the simulations, and the models used to generate the missing data and to implement the weighting and imputations. Section 4.4 gives the results of simulation. The last section concludes with an analysis of error PUC and *efficacy* of the adaptive sampling design compared with those under traditional fixed sampling design and gives some discussion about the methods and their applicability.

4.2 Method

We assume that suitable benchmark data are available for the planning and developing of the adaptive sampling designs. We also assume that the focal survey is organized into sequential phases and the total sample is divided into separate replicates that are introduced at the beginning of each sequential phase. See chapter 2 Figure 1 for data structure that facilitates the implementation of benchmarked sequential sampling designs. Two post-survey adjustment strategies are investigated: the conventional post survey weighting strategy and the benchmarked multiple imputation (B-MI) strategy. The frame variables used for weighting are also used for B-MI. While B-MI uses micro-level data, the final step of the weighting adjustment (i.e. the calibration) uses aggregated data as control totals. For our purpose, we ignore the uncertainty of the control total.

4.2.1 The Sample Designs

The fixed sampling design refers to the conventional sampling design where a sample is drawn to be representative of the underlying target population. The sample design remains unchanged across sampling phases. Subjects with the same characteristics have the same sampling probability across sampling phases.

In contrast to the fixed sampling design the sampling probabilities of the adaptive sampling design for a given characteristic vary across phases, depending on the distribution of the cumulative respondent pool. We study two types of benchmarked sequential sampling designs, BSS-Z and BSS-X, that are given in Chapter 3.

Suppose $\mathbf{Y}$ denotes survey outcome variables, $\mathbf{X}$ denotes survey covariates, and $\mathbf{Z}$ denotes frame variables. The benchmarked sequential sampling designs call for an external benchmark survey that shares a subset of $\mathbf{X}$ and $\mathbf{Z}$ variables with the focal survey.

The core idea of the benchmarked sequential sampling design is that a k-phase focal survey is conducted. Benchmark data are available where $P_B(\mathfrak{A})$ denotes the distribution of some variables $\mathfrak{A}$. After obtaining the phase $k-1$ respondent data, the goal is to restore the resemblance between the benchmark and the cumulative respondents, denoted as $Q^{(k-1)}(\mathfrak{A})$, with respect to the distribution of $\mathfrak{A}$ variables. This goal can be written as

$$\pi^{(k)} \times P^{(k)}(\mathfrak{A}|sample) + (1-\pi^{(k)}) \times Q_R^{(k-1)}(\mathfrak{A}) = P_B(\mathfrak{A}) \tag{4.1}$$

where π denotes the proportion of the new sample to be added. The adaptive sampling probability at phase k, $P^{(k)}(sample|\mathfrak{A})$, can be solved by combining formula (4.1) with a Bayes formula.

The benchmarked sequential sampling design, BSS-Z, derives sampling probability by matching the $\mathbf{Z}$ distribution of the focal survey and the benchmark survey. Its sampling probability for subject i at phase $k+1$ is

$$\begin{aligned} P^{(k+1)}(sample|\mathbf{Z}) &= \{\frac{P_B(\mathbf{Z})}{P^{(k)}(\mathbf{Z})} - \pi^{(k)} \times \frac{Q^{(k)}(\mathbf{Z})}{P^{(k)}(\mathbf{Z})}\}\frac{P^{(k)}(sample)}{1-\pi^{(k)}} \\ &\propto \frac{P_B(\mathbf{Z})}{P^{(k)}(\mathbf{Z})} - \pi^{(k)} \times \frac{Q^{(k)}(\mathbf{Z})}{P^{(k)}(\mathbf{Z})} \end{aligned} \tag{4.2}$$

where $P_B(\mathbf{Z})$ is the $\mathbf{Z}$ distribution of benchmark, $P^{(k)}(\mathbf{Z})$ is population Z distribution at k^{th} phase, $Q^{(k)}(\mathbf{Z})$ is the cumulative respondent $\mathbf{Z}$ distribution obtained up to the k^{th} phase, and $1-\pi^{(k)}$ denotes the cumulative sample proportion at the k^{th} phase. In practice, $P^{(k)}(\mathbf{Z})$ is approximated by the $\mathbf{Z}$ distribution in sampling frame. In most cases, $P^{(k)}(\mathbf{Z}) = P(\mathbf{Z})$ for all k.

The BSS-X design derives sampling probability through benchmarking $\mathbf{Z}$ and $\mathbf{X}$, where $\mathbf{X}$ for the subsequent phases are imputed. The sampling probability for subject i at phase $k+1$ is

$$P^{(k+1)}(sample|\mathbf{Z},\mathbf{X}) = \{\frac{P_B(\mathbf{Z},\mathbf{X})}{P^{(k)}(\mathbf{Z},\mathbf{X})} - \pi^{(k)} \times \frac{Q^{(k)}(\mathbf{Z},\mathbf{X})}{P^{(k)}(\mathbf{Z},\mathbf{X})}\}\frac{P^{(k)}(sample)}{1-\pi^{(k)}} \tag{4.3}$$

where $P^{(k)}(\mathbf{Z},\mathbf{X})$ denotes the distribution of $(\mathbf{Z},\mathbf{X})$ from sampling frame. In most cases, $P^{(k)}(\mathbf{Z},\mathbf{X}) = P(\mathbf{Z},\mathbf{X})$ for all k.

4.2.2 The Estimation Methods

We study inference for a sample mean by two estimation methods: the weighting method (WT) and the benchmarked multiple imputation method (MI).

We denote the full sample by D. The subset of $n_R^{(k)}$ respondents at phase k sampling is denoted as $D_R^{(k)}$, and the subset of nonrespondents is denoted as $D_{NR}^{(k)}$. Four phases are implemented for a total sample size of $\sum_{i=1}^{K} n^{(i)} = n$, where $K = 4$. We then denote $D_R = \bigcup\limits_{k=1}^{K} D_R^{(k)}$ and $D_{NR} = \bigcup\limits_{k=1}^{K} D_{NR}^{(k)}$.

The weighted estimator of a sample mean is $\bar{y}_w = \sum_{i \in D_R} w_i y_i / \sum_{i \in D_R} w_i$, where y_i is the i^{th} respondent reported value and w_i is the corresponding survey weight. The imputed estimator of a sample mean is $\bar{y}_m = \sum_{i \in D_R} w_i y_i / \sum_{i \in D_R} w_i + \sum_{j \in D_{NR}} w_j \bar{\tilde{y}}_j / \sum_{j \in D_{NR}} w_j$, where $\bar{\tilde{y}}_j$ is the average of imputed values for unit j in the nonrespondent set.

4.2.2.1 Weighting

Weighting methods are the current convention to adjust for unit nonresponse. Respondents are weighted by a sequence of weights to account for the uncertainty by sample design, nonresponse, and the deviation from the target population. A weighted mean estimate can be written as

$$\bar{y}_w = \sum_{i \in D_R} d_i u_i g_i y_i / \sum_{i \in D_R} d_i u_i g_i$$

where d_i denotes design weight, u_i denotes nonresponse weight, and g_i denotes weight from a calibration strategy.

For the nonresponse weight, we predict the probability of response by fitting a logistic regression model using frame variables that are available for both respondents and nonrespondent. The u_i is the inverse of the estimated response probability. The nonresponse adjusted weights are further calibrated to match the control totals, which are the total of benchmark data frame variables.

Among the model-assisted calibration strategies, including poststratification, raking, and the generalized regression estimator (GREG), we choose GREG which is known for improved efficiency over traditional weighting methods by borrowing strength from auxiliary information (Deville and Särndal, 1992). GREG weighting adjusts the weights so that the weighted sample totals of covariates conform to population totals (i.e., benchmark data totals in our case).

4.2.2.2 Benchmarked multiple imputation

Literature on imputation for unit nonresponse is scarce and has mixed findings (Rässler and Schnell, 2004). We study benchmarked multiple imputation (B-MI) which is applicable to the particular situation when a micro-level benchmark data capturing a target population of interest is available (see section 3.2.1). Benchmarked multiple imputation differs from other imputation adjustments for unit nonresponse in that nonrespondents are imputed using both the benchmark data and the respondent data, instead of respondent data alone. The inclusion of the benchmark data conforms the unit nonresponse to a special case of item nonresponse, facilitating the imputation rationale. We adopted Multivariate Imputation by Chained Equation (MICE) as the imputation strategy (see section 3.2.3).

MICE is described in detail in van Buuren (2007) and van Buuren and Groothuis-Oudshoorn (2011). As with the item nonresponse approach, the MICE assumes ignorable missingness when conditioning on all other variables in the model. For each missing unit, m units are imputed, creating m completed datasets. The point estimate is the average of m imputations, $\bar{Y}_m = \frac{1}{m}\sum_{j=1}^{m}\bar{Y}_j$. The variance estimation takes into account within and between imputation variations, as suggested by Rubin (1987) and Rubin and Schenker (1991).

$$T = \bar{U} + \big(1 + \frac{1}{m}\big)B \tag{4.4}$$

where $\bar{U}$ is the average variance within the imputations and B is the variance between the imputations.

4.2.3 The Cost Model

The most cited survey cost model is of a linear form. For example, the cost model for a two-stage design with a total cost of C can be written as

$$C = C_0 + C_1 n_1 + C_2 n_2 \tag{4.5}$$

where C_0 denotes the fixed overhead cost, C_1 denotes the sampling or operation cost corresponding to n_1 primary sampling units, PSU, (e.g., housing units), and C_2 denotes the variable cost for n_2 secondary sampling units (e.g. subjects).

Overhead cost is the fixed cost of conducting a survey regardless of numbers of PSUs and sample subjects. Sampling/operation cost refers to costs such as the implementation of sample design (e.g. mapping and listing housing units, etc., field representative training, and the deployment of the field staff). The variable cost is the costs which increases with increases in sample size at each sampling stage. Example variable costs include hours, miles and other expenses related to locating, contacting and interviewing (when possible) sampled units.

As the formula states, the sampling/operation costs and variable costs are frequently considered separately. Sampling cost sometimes is included in the overhead cost C_0, not affecting allocation. Operation cost, depending on the survey design, may or may not increase with the increase in sample sizes. For example, operation cost increases discontinuously in a household face-to-face interview survey where a fixed ratio of field supervisor to field representative is needed. When the sample size increases so that an extra field representative needs to be hired; the hiring leads to the higher ratio of supervisor and staff, which in turn necessitates the hiring of a supervisor; in this situation the operation cost increases with the increase in the sample size. On the other hand, if the increase in the sample size fits into current field representatives' workload, then the operation cost would not increase. Variable cost (C_2) is the cost that by definition links to the direct cost of sampled subjects.

Note that the setup of the cost model in (4.5) implies that all subjects, regardless of their characteristics and response propensities, have identical variable cost, C_2. This assumption is often not true. Groves (2004) pointed out that there is evidence that the efficiency of interviewers increases with the number of interviews completed. Therefore, the cost per sampled subject is a decreasing function of sample size within an interviewer. The model in (4.5) also did not take into account the discontinuous increase in operation cost, as the example described above.

In practice, survey costs are affected by the response propensity, in addition to the sample design, operational plan and the sample size. The reason is that cases with lower response propensity are likely to incur higher cost. Therefore, a cost model needs to take into sensible consideration the differential cost of sampling and operation between subjects with lower and higher response propensities, and presents a fair comparison on the variable cost for these subjects. We focus on survey cost before nonresponse follow-up. This is because the adaptive sampling designs improve respondent representativeness by oversampling the under-represented subjects, instead of conducting nonresponse follow-up. Conducting nonresponse follow-up increases the inferential complexity (Brick, 2013).

We propose a cost model that specifies the individual-level costs and reflects the differences between subjects with lower (likely nonrespondent) and higher (likely respondent) response propensities. For example, the sampling cost incurred by a likely nonrespondent may include the use of special frame, samples from areas with strong minority concentrations, sub-sampling within household, and screening (extra screener households). Operation cost incurred by likely nonrespondents may be allocating field representatives to a rural area or inner city, and special training for field representatives on communication skills. The variable cost for a likely nonrespondent is higher, perhaps due to the reasons such as harder to obtain a contact (requiring more visits) and higher probability of rejection (requiring a more experienced interviewer and costlier data collection protocols).

With respect to the differential cost between adaptive sampling and fixed sampling designs, adaptive sampling designs over-sample subjects with lower response propensity, incurring higher sampling, operation and variable costs. Suppose the overhead cost of conducting a survey is comparable between fixed and adaptive sampling designs, we can leave it out from the cost model for our purpose. Let total cost of subject i, excluding overhead cost, be written as

$$C_i = A_i + B_i \tag{4.6}$$

where A_i denotes the sample/operation cost and B_i denotes the variable cost. The setting of formula (4.6) suggests that, depending on their characteristics, each subject has their own sampling/operation and variable costs. Using this model, the total cost will be higher for a sample with higher number of likely nonrespondents than a sample with lower number of likely nonrespondents.

We assume that (a) subject i incurs the variable cost B_i, regardless of the sample design and (b) subject i incurs sampling/operation cost A_i regardless of the sample design. That is, if subjects i and j have the same response propensities, subject i is in fixed sampling design and subject j is in adaptive sampling design, they incur the same total cost, $C_i = C_j$, regardless of the sample design. We can simplify the comparison by considering the cost model as the inverse of the response propensity. That is,

$$C_i \propto 1/p_i \tag{4.7}$$

where p_i is the response propensity for subject i. Formula (4.7) takes into account of the differential cost between likely respondents and nonrespondents. This cost model is subject-specific, stochastic, and simple.

4.2.4 Evaluation Measures

Evaluation focuses on two aspects: root mean square error per unit cost (RMSE PUC) and cost-effectiveness (*efficacy*). RMSE PUC assesses the sizes of the RMSE for a fixed cost. *Efficacy* (*Eff*) takes into account both error and cost and evaluate the "best value" design. Cost unit is defined as the inverse of the estimated response propensity.

The RMSE PUC is computed by

$$RMSE_{PUC} = RMSE/C_s \tag{4.8}$$

where $RMSE = \sqrt{MSE}$ and C_s is the total cost for sample S. $C_s = \sum_{i \in S} C_i$, and $C_i \propto 1/p_i$, where p_i denotes response propensity for subject i. MSE is mean square error, estimated by

$$MSE(\theta) = B(\theta)^2 + V(\theta) \tag{4.9}$$

where $B(\theta)$ is the bias of a point estimate θ, estimated by $bias(\hat{\theta}) = \sum_S (\hat{\theta}_S - \theta_N)/r$. $\hat{\theta}_S$ is the estimate from sample S, and θ_N is the finite population parameter. $V(\theta)$ is the empirical variance of $\hat{\theta}$, estimated by

$$Var(\hat{\theta}) = \sum_S (\hat{\theta}_S - \bar{\theta})^2/r, \tag{4.10}$$

where $\bar{\theta} = \sum_S \hat{\theta}/r$ and r denotes simulation replicates. We also compute Bias PUC and SE PUC. Bias PUC is estimated by $Bias_{PUC} = bias(\hat{\theta})/C_s$. Standard error PUC is estimated by $SE_{PUC} = \sqrt{V(\theta)}/C_s$.

RMSE PUC decreases when errors decrease, a desirable scenario. However, given the same error sizes, designs with larger costs also yield smaller RMSE PUC, perplexing the comparison. The ultimate preference would be a design that not only produces smaller errors per cost unit but also costs less. We propose an *efficacy* measure that increases when errors decrease and/or costs decrease. The *efficacy* is computed by

$$\textit{Eff} = f(Q)/g(C) \tag{4.11}$$

where Q stands for $Quality$, proportional to the inverse of the error. $f(Q)$ denotes a quality function and $g(C)$ denotes a cost function. Example *efficacy* measures are *Eff* $= (RMSE)^{-1}/C_s$ and *Eff* $= (Bias)^{-1}/\sqrt{C_s}$.

The relationship between quality and cost is likely to be complex and survey-dependent. For illustration purpose, we compute $\textit{Eff}_{RMSE}$, $\textit{Eff}_{Bias}$ and $\textit{Eff}_{SE}$.

$$\textit{Eff}_{RMSE} = (RMSE)^{-1}/g(C_s), \tag{4.12}$$

Similarly, $\textit{Eff}_{Bias} = (Bias)^{-1}/g(C_s)$ and $\textit{Eff}_{SE} = (SE)^{-1}/g(C_s)$. Note that GREG estimates are unbiased under MAR and B-MI estimates are unbiased under MAR and MNAR$_X$, therefore $\textit{Eff}_{Bias}$ is only defined under MNAR$_X$ and MNAR$_Y$ for GREG estimates and MNAR$_Y$ for B-MI estimates. Three $g(C_s)$ are evaluated: $g(C_s) = C_s$, $g(C_s) = \sqrt{C_s}$, and $g(C_s) = log(C_s)$. These cost functions imply linear, square root and logarithm relationship between the quality and the cost, respectively. When comparing sample designs, larger *efficacy* is preferable.

We also consider relative efficiency, 95% confidence interval width and coverage rates. The relative efficiency is computed by Rel.eff $= SE_F/SE_{BSS.Z}$, where larger Rel.eff values indicate better efficiency for adaptive designs. The 95% confidence interval width is computed by Width$_{95} = 2 \times 1.96 \times SE$. The percentages of intervals that include θ_N are based on the nominal 95 percent confidence intervals ($\hat{\theta} \pm 1.96\hat{V}^{1/2}$) computed for each of the r simulations for each simulation scheme.

4.3 Design of the Simulation Study

The simulations study the 1) root mean square error per unit cost (RMSE PUC), 2) cost-effectiveness (*efficacy*), and 3) properties of the bias and variance for weighted and imputed estimates of the sample mean under various unit nonresponse mechanisms.

4.3.1 Study Population and Sample Design

We simulate a finite population of size $N_{POP} = 500,000$ that comprises of three variables (Y, X, Z) from a multivariate normal distribution with the following joint distribution.

$$(Z, X, Y) \sim N_3(\mathbf{0}, \Sigma), \qquad \text{and} \tag{4.13}$$

$$\Sigma = \begin{pmatrix} 1 & cov(Z,X) & cov(Z,Y) \\ cov(Z,X) & 1 & cov(X,Y) \\ cov(Z,Y) & cov(X,Y) & 1 \end{pmatrix}$$

where $cov(Z, X)$ and $cov(Z, Y)$ are varied to examine the effects of Z with various degrees of explanatory power on X and on Y. And $cov(X, Y)$ are varied to examine the effects of X with various degrees of explanatory power on Y. From this finite population, a simple random sample without replacement (srswor) of size $N_B = 25,000$ (5% of the finite population) is drawn to serve as the benchmark data and the remaining $475,000$ units serve as the sampling frame for the focal surveys. From this sampling frame, three concurrent focal surveys are simulated: the conventional fixed sampling design (F), and the adaptive sampling designs (BSS-Z and BSS-X), as described in Sections 2.2, 3.2.2 and 4.2.1.

For all three focal surveys, the first phase is a srswor of size $1,000$ ($n^{(1)} = 1,000$). This sample size is carried on for phase II to IV for the F sampling design. For BSS-X and BSS-Z designs, the k^{th} phase samples are of size $n^{(k)} = 1,200$, where $k = 2, 3, 4$. The larger sample sizes are to account for higher selection probability on subjects who are less likely to respond. These sample sizes in BSS-Z and BSS-X design result in comparable number of respondents as with the fixed design such that we are weighting and imputing the same number of missing units at the estimation step and arriving in the same sizes of completed data.

The simulation study can be described as a $2 \times 2 \times 2 \times 4$ factorial design. The factors are:

Factor A: $cov(Z, X)$: high vs low

Factor B: $cov(Z, Y)$: high vs low

Factor C: $cov(X, Y)$: high vs low

Factor D: Four nonresponse mechanisms based on the response models described below.

Among factors A and B, a high correlation of two variables is set to be $\rho = 0.894$ (i.e. $R^2 = 0.8$) and a low correlation is $\rho = 0.447$ (i.e. $R^2 = 0.2$, where R^2 is the coefficient of determination in simple linear regression). For factor C, a high correlation between X and Y is $\rho_{X,Y} = 0.707$ (i.e. $R^2(X, Y) = 0.5$) and a low correlation is $\rho_{X,Y} = 0.3$ (i.e. $R^2(X, Y) = 0.09$). The combination scenario of a low correlation (L) in Factor A, a high correlation (H) in Factor B, and a high correlation (H) is Factor C is labeled “LHH” herein. Similarly the other seven scenarios of Factors A, B and C are labeled “LLL”, “LLH”, “HLL” ,“HLH”, “HHL”, and “HHH”.

4.3.2 Missing Data Mechanisms and Imputation Methods

For nonresponse mechanisms (factor D), we simulate response status $\Re$ by drawing from Bernoulli random variables with $e(z, x, y) = Pr(\Re = 1|z, x, y)$ computed from four different nonresponse models:

1. $logit(e(z, x, y)) = Bernoulli(0.5)$. **(MCAR)**

2. $logit(e(z, x, y)) = 0.00020 + 0.8z$. **(MAR)**

3. $logit(e(z,x,y)) = 0.00001 + 0.31z + 0.61x$. **(MNAR$_X$)**

4. $logit(e(z,x,y)) = 0.00004 + 0.19z + 0.38x + 0.38y$. **(MNAR$_Y$)**

where $\Re = 1$ if subject responded, 0 otherwise. Each of these models generate an average of 50% response rate. For model 1, the response probability is independent of (Y, Z, X), i.e. MCAR. For model 2, the response probability depends on Z alone, i.e. MAR. For model 3 the response probability depends on both Z and X, labeled as MNAR$_X$. For model 4, the response probability depends on (Y, Z, X), labeled as MNAR$_Y$. Note, model 3 is MNAR since response depends on X, which is not observed for nonrespondents.

For the imputation procedure, five imputations are implemented where each imputation runs for 10 iterations. The imputation models include main effects of Y,X and Z. Let $n^{(K)}$ denote the intended sample size up to phase k, of it $n_R^{(K)}$ responded and $n_{NR}^{(K)}$ did not. There are two options to obtain a completed data of size $n^{(K)}$. The first option applies to F design. At each phase, the Y_{NR}, X_{NR} and Y_B are imputed by MICE using a linear model based on the cumulative respondents Y_R, X_R, Z_R, nonrespondent Z_{NR}, and the benchmark X_B, Z_B.

The second option is to impute the nonrespondents, not only for $(\mathbf{Y}_{NR}^{(K)}, \mathbf{X}_{NR}^{(K)})$ but also for $(\mathbf{Z}_{NR}^{(K)})$. This option is preferred for BSS-Z and BSS-X designs. Note that Z_{NR} is imputed as described in section 3.3 of Chapter 3. In addition, X values corresponding to Z of the next phase sampling frame is also imputed for BSS-X design, since $\hat{X}$ on the sampling frame is necessary for the next phase sampling.

For all three designs, imputation is carried out by using all available data collected so far. Therefore imputed values at phase k are always replaced by the imputed values at phase $k+1$ for both sampling and inference.

With the finite population generated by the data model (4.13), the procedure of a) drawing a benchmark, b) forming a sampling frame, c) simulating three concurrent focal surveys, and d) imputing nonrespondents, is carried out $1,000$ times (trials) for each of the 32 simulation scenarios.

For each trial, the evaluation measures mentioned in Sect. $4.2.4$ are computed. For each simulation scenarios, we report the average quantities from the $1,000$ trials.

4.4 Simulation Results

We first present the estimated cost from the three sample designs to show the higher costs from adaptive designs. To compare designs of higher cost with smaller bias to designs of lower cost with bigger bias, we consider a measure, root mean square error per unit cost (RMSE PUC), that takes into account of both cost and error. While a design may have smaller RMSE PUC, a preferable design should also cost less, that is, cost-effective. To assess the cost-effectiveness for the three sample designs we report on *efficacy* (Eff), a measure that increases when cost decreases and/or error decreases. Lastly we show bias PUC, SE PUC and 95 percent confidence interval coverage rates as described in Sect. $4.2.4$ for GREG and B-MI estimators, respectively.

4.4.1 Cost

For a total sample of $1,000$ subjects with an average of 50% response rate, the average cost for the fixed design is approximately $9,508$ cost units. To obtain a comparable number of respondents, adaptive designs have to sample 1200 subjects which require an average cost of $13,024$ and $13,273$ cost units for BSS-Z design and BSS-X design, respectively. The relative cost of adaptive designs are 37% and 40% higher than fixed design. The higher cost of the adaptive designs is attributed to the objective of reducing nonresponse biases by oversampling of under-represented subjects. Evaluating survey cost alone may not be as pertinent as considering survey errors per unit cost since the later contains information on survey quality generated by the cost. Below we show results from several evaluation criteria to assess a survey design on its cost and error.

4.4.2 RMSE Per Unit Cost

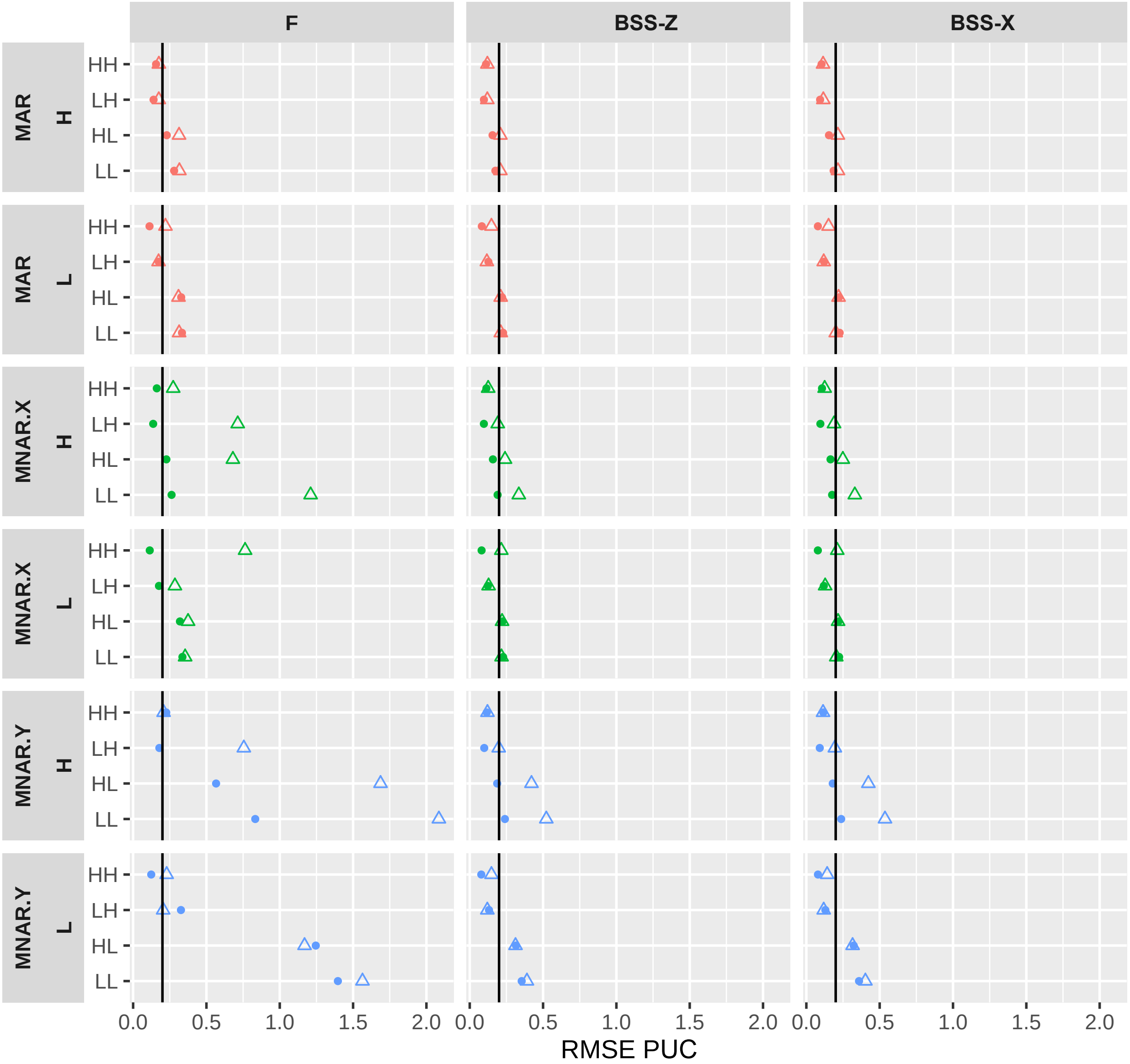
Figure 4.1 shows the results of the simulations for the GREG (△) and the B-MI (●) estimates from the $10,000$ trials for each of the 24 simulation schemes. The figure gives the RMSE PUC by three sample designs for estimating $\bar{Y}$. A vertical line at 0.2 is drawn to aid visual comparison. All three sample designs performed very well under MAR missingness with unbiased point estimates and roughly comparable RMSE per unit cost. Under MNAR$_X$, all three sample designs had comparable RMSE PUC over B-MI estimates. However, for the GREG estimates, the two adaptive sampling methods were substantial improvements over the traditional fixed sampling design.

When missingness is MNAR$_Y$, the results depend on the strength of the auxiliary (weighting) variables. When the auxiliary variables (**Z**) are strong, i.e. $\rho_{Z,Y} = H$, RMSE PUC are comparable among three sample designs within the bounds of simulation error, for both B-MI and GREG estimates. The one exception where B-MI outperforms GREG for the fixed sampling design is under LHH, suggesting that B-MI estimators capitalize on the high correlation between X and Y to reduce RMSE while GREG does not.

When the auxiliary variables are weak, the two adaptive sampling designs outperformed the traditional fixed sampling design substantially, for both B-MI and GREG estimators. This is because weak weighting variables cannot effectively reduce bias and variance (Little and Vartivarian, 2005), adaptive designs with representative respondents are less susceptible to weak weighting variables. If weak auxiliary variables compound with the high $\rho_{X,Y}$, the B-MI estimators have greater advantage over the GREG estimators. This advantage is most pronounced for the fixed sampling design.

Since the RMSE PUC from the two adaptive designs, BSS-Z and BSS-X, are comparable, the graphs presented hereinafter (Figures 4.2, 4.3, 4.4, and 4.5) only show the results for the BSS-Z design with the F design. A downside of RMSE PUC is that, for designs with the same error sizes, the costlier one will yield a smaller RMSE PUC.

Figure 4.1: Root mean square error per unit cost (RMSE PUC) by sampling design at phase IV.



△ denotes GREG estimates and ● denotes the B-MI estimates. Horizontal panels show nonresponse mechanism (MAR, MNAR$_X$, MNAR$_Y$) and $cor(X, Y)$ (High,Low) and vertical panels show sampling designs. Within each plot, x-axis shows the values of RMSE PUC, and y-axis represents predictive power of Z on X and Y (i.e. $cor(Z, X)$ and $cor(Z, Y)$). For example, "LH" indicates a low correlation between Z and X ($\rho_{Z,X} = 0.2$) and a high correlation between Z and Y ($\rho_{Z,Y} = 0.8$). F = fixed sampling design, BSS-Z = benchmarked sequential sampling using Z, and BSS-X = benchmarked sequential sampling using $(Z, \hat{X})$. A vertical line at $x = 0.2$ is added to aid visual comparison.

4.4.3 RMSE *Efficacy*

Efficacy assesses cost-effectiveness with larger numbers representing designs of "better value". Ratio of *efficacy* provides an easy measure to evaluate the relative cost-effectiveness from two designs. A ratio larger than one indicates the numerator design is more cost-effective. Figure 4.2 shows the ratio of the *efficacy* measures, BSS-Z divided by F (denoted by $ER_F^{BSS.Z} = ER =$ Eff$^{BSS.Z}$/Eff^F) for the GREG (△) and the B-MI (●) estimates from the $10,000$ trials for each of the 24 simulation schemes. Vertical lines at $ER = 1$ are drawn to aid visual comparison. There are three vertical lines for each plot, with each line at $ER = 1$ from one of the three cost functions (see Sect. 4.2.4). For example, the solid black line is drawn at $ER = 1$ for cost model 1. The location of $ER = 1$ varies by the cost function, suggesting that the relative *efficacy* of two designs hinges on the relationship between their quality and cost.

For Eff$_{RMSE}$ under MAR, F design is more cost-effective since all three designs have comparable RMSE but fixed design costs less. Similarly, under MNAR$_X$ all three designs have comparable RMSE over B-MI estimates and fixed design costs less, hence fixed design is of better value. For the GREG estimates, BSS-Z design is more cost-effective than fixed design, echoing the results of RMSE PUC.

When missingness is MNAR$_Y$, adaptive designs are more cost-effective with the exceptions of HHH and HHL scenarios. For these two scenarios, Eff$_{RMSE}$ for the F design is slightly better, comparable, and slightly worse than adaptive designs under linear, square root, and logarithm cost function, respectively, for both B-MI and GREG estimates. When the auxiliary variables are weak, the two adaptive designs are more cost-effective than the fixed sampling design, for both B-MI and GREG estimators. The advantage of adaptive designs in *efficacy* is more pronounced for GREG estimators, which is attributed to the poor performance of GREG under MNAR with weak weighting variables. This finding echoes that of the RMSE PUC described above.

The results of *efficacy* suggest that when errors can be eliminated from post-survey adjustments, fixed design is of better value. Otherwise, depending on the relationship between quality and cost, adaptive designs are often more cost-effective, especially under MNAR with weak weighting variables.

In the following sections, we report on Eff$_{bias}$ and Eff$_{SE}$, bias PUC, SE PUC, and 95% coverage rate. We show the results from the GREG estimates, followed by the results for the B-MI. GREG results reflect the current norm of practice whereas the B-MI results illustrate the improvement over GREG by taking into account of X in the post-survey adjustments.

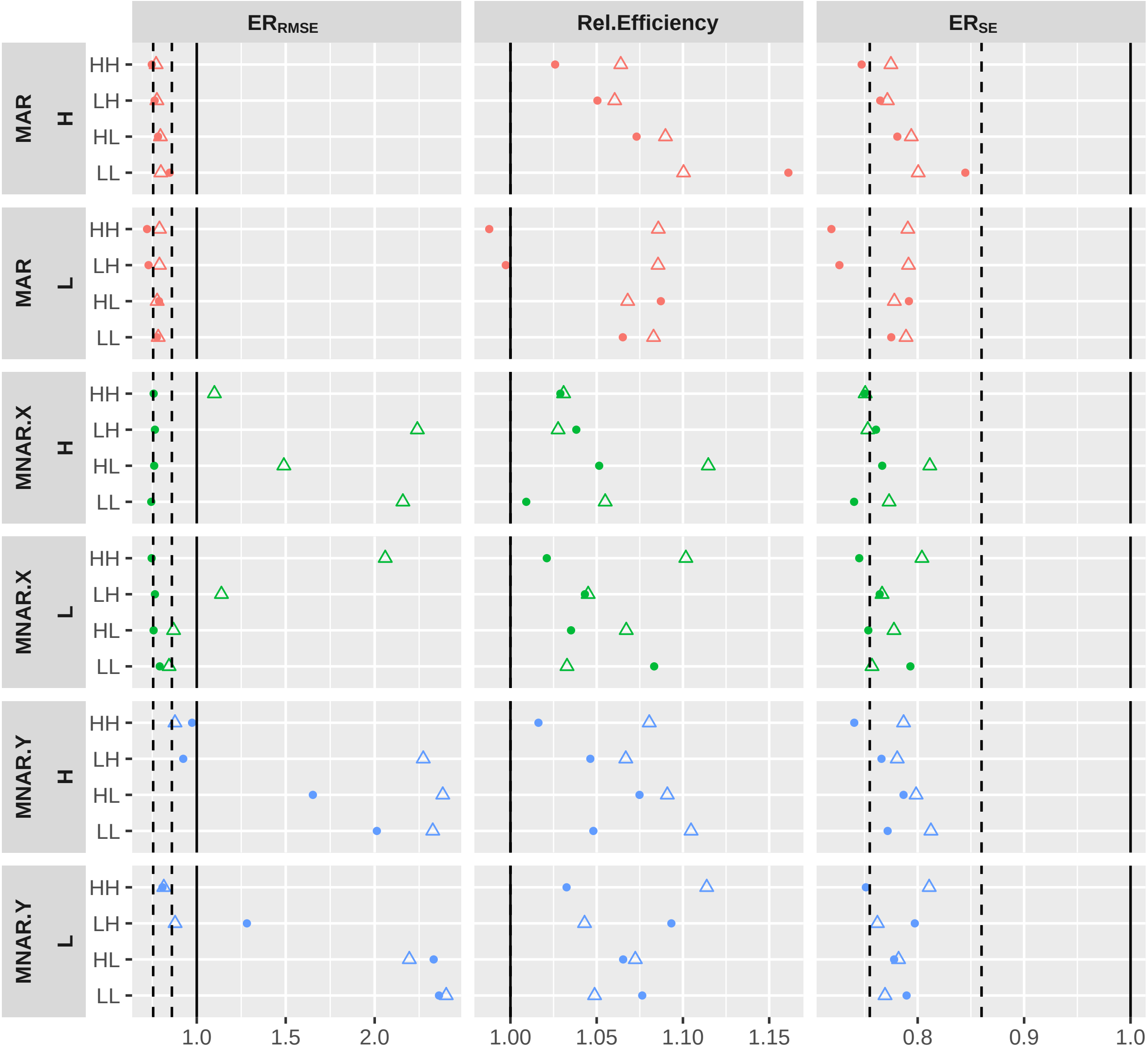
4.4.4 GREG Estimates

Theoretically the point estimates should be nearly unbiased with GREG weighted respondent data if missingness is MAR (Deville and Särndal, 1992). As illustrated in the right hand panel of Figure 4.3, the expected outcome of zero bias holds for the three sample designs. For MNAR missingness, the bias of fixed design is greater than that of the adaptive designs since adaptive designs obtain a more representative respondent set. With greater cost and smaller bias, the bias per unit cost for the adaptive design is expected to be noticeably more favorable than for the fixed design. The graph of bias per unit cost in Figure 4.3 (first column) is consistent with this expectation.

The advantage of adaptive designs on bias PUC under MNAR is mostly echoed by the Eff$_{bias}$, with the exceptions at the situations of strong auxiliary variables. When biases are minimized by strong auxiliary variables, the advantage of adaptive designs decreases(with respect to their smaller biases relative to its cost), resulting in diminished cost-effectiveness. This situation can be seen under HHH and HHL scenarios. On the other hand, the advantage of adaptive design is especially pronounced under MNAR$_Y$ when weighting variables are weak, i.e. $\rho_{zy} = L$. When ER values are greater than one, they indicate better *efficacy* on BSS-Z design.

While the bias of the point estimates under MNAR$_X$ and MNAR$_Y$ in the two adaptive sampling designs are small (always less than 0.05 per unit cost), they still may be important if the ratio of the bias to the standard error (SE) is relatively large. A large bias-to-SE ratio suggests that the coverage rate can be much lower than the nominal level (Cochran, 1977). For the adaptive

Figure 4.2: *Efficacy* and efficiency ratio of BSS-Z and F at phase IV.



△ denotes GREG estimates and ● denotes the B-MI estimates. Horizontal panels show nonresponse mechanism (MAR, MNAR$_X$, MNAR$_Y$) and $cor(X, Y)$ (High,Low) and vertical panels show: RMSE *efficacy* ratio (ER_{RMSE}), relative efficiency (R.Efficiency) and SE *efficacy* ratio (ER_{SE}). Within each plot, x-axis shows the values of corresponding *efficacy* or efficiency measures, and y-axis represents predictive power of Z on X and Y (i.e. $cor(Z, X)$ and $cor(Z, Y)$). For example, "LH" indicates a low correlation between Z and X ($\rho_{Z,X} = 0.2$) and a high correlation between Z and Y ($\rho_{Z,Y} = 0.8$). F = fixed sampling design and BSS-Z = benchmarked sequential sampling using Z. Three vertical lines (one solid and two dashed lines) for ER.RMSE and ER.SE are added to aid visual comparison where, for ER_{RMSE} and ER_{SE}, each vertical line indicates $ER = 1$ for the corresponding cost models. The solid line is based on linear cost model, the first dashed line is based on logarithm cost model, and the second dashed line is based on the square root cost model. For R.Efficiency (F/BSS-Z), a vertical line at R.Efficiency $= 1$ is added where larger values indicate better efficiency for the BSS-Z design.

designs GREG estimates, the ratios never exceed 0.07, 2 and 3.5 for MAR, MNAR$_X$ and MNAR$_Y$ missingness, respectively. Much larger ratios occur for fixed design estimates, especially under MNAR$_Y$ missingness, where five of the eight simulation schemes have ratios above 5. As expected, the corresponding coverage rates for these large ratios are much lower than the nominal level, close to zero in some cases as discussed later.

The variance is estimated empirically, as shown in equation (4.10). Theoretically good weighting variables reduce not only bias but also the variance of the estimates. That is, the variances are smaller when $\rho_{zy} = H$ as comparing to when $\rho_{zy} = L$. These observations hold for all three sample designs. The variance estimates are consistent with this theoretical result. The variances are also smaller when weights are less variable, such as, for example under a representative respondent set. We found that, for a given simulation scheme, the variances of the two adaptive designs are consistently smaller than those of the fixed design. The graph of Rel.eff in Figure 4.2 (second column) is consistent with this result.

With larger variance estimates and greater costs in fixed designs, the standard error (SE) per unit cost for the fixed design is larger than that of the adaptive designs. This result is shown in the second column in Figure 4.3. The large differences between the F and BSS-Z designs in SE PUC are observed for all three nonresponse mechanisms. These differences are greater when the weighting variables are weak.

Whether the advantage of adaptive designs on SE PUC translates to better *efficacy* depends on the cost function. For example, the third column on figure 4.2 (△) shows the *efficacy* ratio of SE for BSS-Z and F designs. When the cost function is linear or square root, the ER_{SE} values are less than one for all simulation scenarios, suggesting better *efficacy* on fixed design. For a logarithm cost function, however, the adaptive designs have better *efficacy*.

For a given simulation scheme the variance differences in three designs are small, leading to similar confidence intervals. However, the coverage rates of these confidence intervals may prove to be more insightful. The third column in Figure 4.3 shows that the coverage rates of the 95%

confidence intervals for the estimates under MAR missingness are generally close to the nominal 95 percent level for both F and BSS-Z designs.

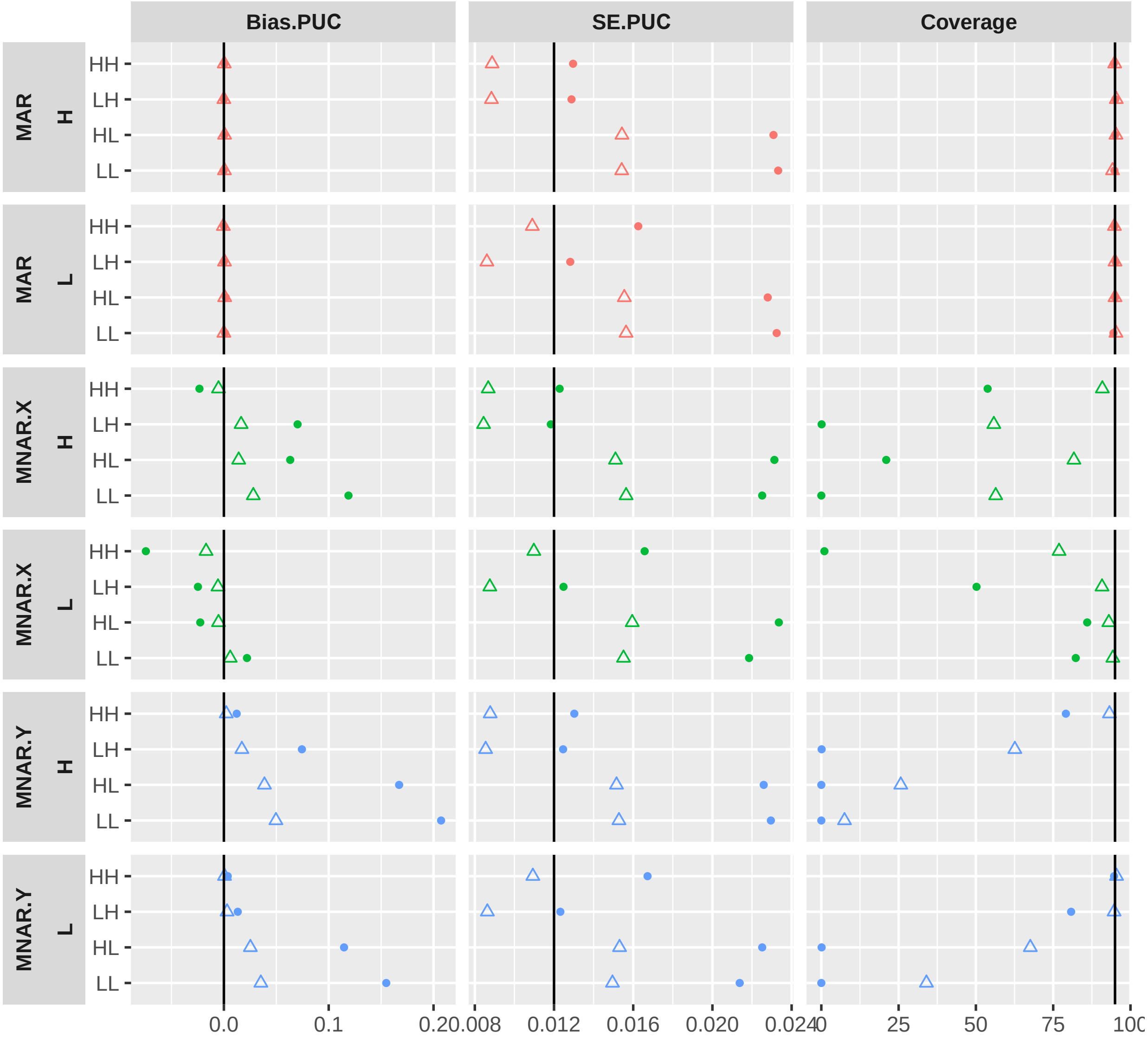
The coverage rates for BSS-Z design under MNAR$_X$ are between 80% and 95%, with the exception of LHH, LLH, and HHL schemes (coverage rates of 55%, 55%, and 76%, respectively). For these three schemes the F design has zero coverage rate, which is attributed to the relatively large bias in point estimate. Under MNAR$_Y$ the coverage rates are zero for F design with the exception of HHH, HHL, and LHL schemes. These schemes have coverage rates of 80%, 95%, 86% for the F design, and 94%, 95%, and 95% for the BSS-Z design. The BSS-Z design point estimates for these three schemes are nearly unbiased. For those schemes where F design has zero coverage rate, the coverage rates for BSS-Z design range from 25% to 68%, with the exception of LLH scheme (8%) which corresponds to the largest bias on point estimate. Overall, adaptive sampling designs produce confidence intervals with coverages that are vast improvements over those for intervals based on fixed sampling design.

4.4.5 Benchmarked Multiple Imputation Estimates

Benchmarked multiple imputation (B-MI) applied to unit nonresponse can be more effective than weighting strategies in reducing bias. Section 3.3.2.2 reported favorable results on bias reduction for point estimates and multivariate correlation preservation for MAR and MNAR$_X$ missingness. These results are attributed to the effective utilization of micro-level data as well as the information of X covariates. However, micro-level benchmark information may not be available for all surveys, it is unrealistic to favor one estimation method over the other. For this reason, we report the results of B-MI with respect to the gains over traditional weighting method in the case where micro-level data are available. Weighting strategy in most surveys is still the only feasible estimation method.

Figure 4.4 presents the results of the simulations for the B-MI estimates in the same format as used in the GREG estimates. Note that scales for Figures 4.3 and 4.4 differ from each other. The most noticeable pattern is the zero biases and their corresponding nominal level coverage rates un-

Figure 4.3: GREG estimates for 1) bias per unit cost (Bias.PUC), 2) standard error per unit cost (SE.PUC), and 3) 95% coverage rates by sampling design at phase IV.



△ denotes BSS-Z design estimates and ● denotes the F design estimates. Horizontal panels show nonresponse mechanism (MAR, MNAR$_X$, MNAR$_Y$) and $cor(X, Y)$ (High,Low) and vertical panels show evaluation measures (Bias.PUC, SE.PUC, and coverage rate). Within each plot, x-axis shows the values of the corresponding evaluation measures, and y-axis represents predictive power of Z on X and Y (i.e. $cor(Z, X)$ and $cor(Z, Y)$). For example, "LH" indicates a low correlation between Z and X ($\rho_{Z,X} = 0.2$) and a high correlation between Z and Y ($\rho_{Z,Y} = 0.8$). Vertical lines are added to aid visual comparison. For Bias.PUC, a zero line is added. For SE.PUC, vertical line is at 0.012. For 95% coverage rates, the vertical line is at 95.

der MAR and MNAR$_X$ for both BSS-Z and F designs. Recall the results of GREG estimates where both adaptive designs and fixed design show biased estimates under MNAR$_X$. This is because nonresponse depends on X, not on Y, hence GREG estimates of Y are biased for not using X in the weight adjustment. The fact that biases can be eliminated by B-MI under MNAR$_X$ suggests the importance of incorporating X variables in the post-survey adjustment when unit nonresponse depends on X. These results also imply that, under MAR and MNAR$_X$, cost-effectiveness (Eff$_{bias}$) is better for the fixed design than for adaptive designs (for B-MI estimates) since the former cost less. The graph of ER_{bias} in Figure 4.2 (second column) is consistent with this expectation.

Under MNAR$_Y$, the results of B-MI estimates are similar to those for GREG estimates where the bias is not completely eliminated by B-MI even if Z is highly predictive of Y. Highly informative Z still reduce more bias than weak Z, and adaptive designs have smaller bias PUC than the fixed design. However, The advantage of adaptive designs over the fixed design in reducing bias is most pronounced when Z is weak. These findings are also echoed in *efficacy* where adaptive designs have higher values on Eff$_{bias}$ as comparing to the fixed design.

The coverage rates reflect similar results as those of the GREG estimates. In addition, when $\rho_{xy} = H$, the B-MI estimates substantially improve the coverage rates for the adaptive designs over those of the GREG estimates, again suggesting the importance of incorporating X variables in the post-survey adjustments.

The variance (V) is estimated empirically, as shown in equation (4.10). For SE ($SE = \sqrt{V}$), SE per unit cost, and Rel.eff, the results are similar to the GREG estimators. That is, for a given simulation scheme, the variances of the adaptive designs are consistently smaller than those of the fixed design, suggesting adaptive designs to be of better efficiency, as shown by the (larger than one) values of Rel.eff (second column (Rel.eff) in Figure 4.2). With samller SE and larger cost, adaptive designs yield smaller SE PUC than fixed design. In terms of *efficacy*, Eff_{SE} for the fixed design is larger than the adaptive designs when cost function is either linear or square root. When the cost funtion is logarithm, adaptive design Eff$_{SE}$ is larger than the fixed deisgn except when weighting variables are strong (i.e., $\rho_{z,y} = H$).

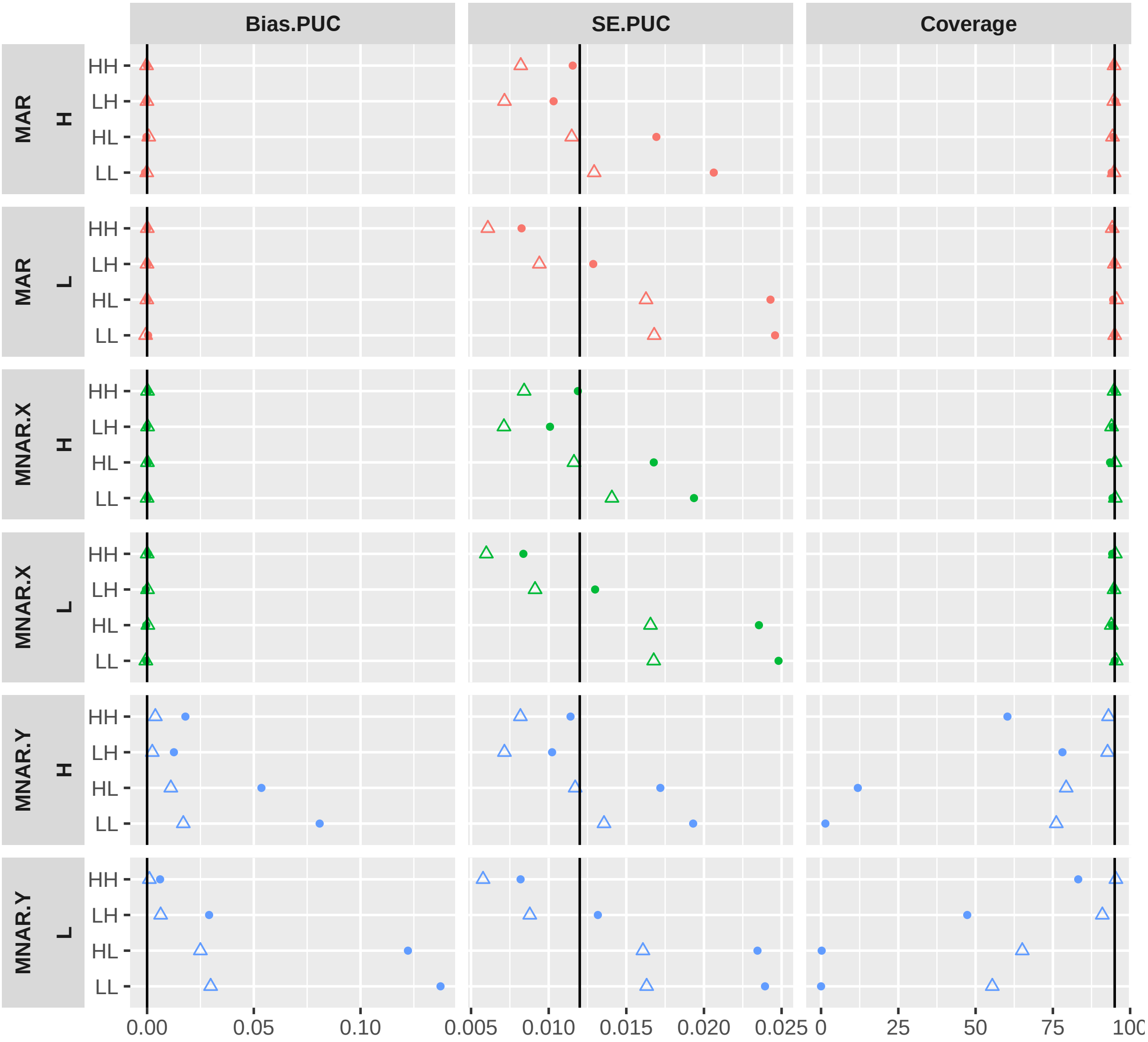
Under MAR and MNAR$_X$ the bias-to-SE ratio are zero for all three designs since the B-MI point estimates are unbiased. For MNAR$_Y$ the adaptive designs have lower bias-to-SE ratio than the fixed design, consist with the results of coverage rates. In addition, adaptive designs bias-to-SE ratio for B-MI estimates are smaller than those of the GREG estimates. This is, however, not always true for the fixed design. Under MNAR$_Y$, the ratio for B-MI estimates are in general smaller than those of the GREG estimates, with the exception of HHH, HHL and LHL simulation schemes. Not surprisingly, the GREG estimates have better 95 percent confidence interval coverage rates for these three schemes comparing to the B-MI estimates, suggesting that the inclusion of X variable in the post-survey adjustment may not necessary improve the results under MNAR$_Y$ when weighting variables have strong correlation with Y.

4.5 Conclusion and Discussion

In today's survey climate with tight budgets and high nonresponse, considering a survey design with favorable cost-effectiveness is preeminent. This study employed simulations to examine the cost and error implications when using the benchmarked sequential sampling methods to improve respondent representativeness, under a single-stage four-phase simple random sample design. We developed a subject-level cost model and presented simulation variance from two types of post-survey adjustments for error estimation. The cost model was nonlinear, stochastic, and inversely proportionate to subjects response propensity. For the variance estimators, we showed a current practice for post-survey adjustments, GREG estimate, and an alternative adjustment, benchmarked multiple imputation (B-MI) estimate. The benchmarked multiple imputation offered some insights for additional gains over GREG on bias reduction in situations where micro-level benchmark data and X covariates were available.

Under MAR, both GREG and B-MI produced unbiased point estimates regardless of sample

Figure 4.4: Benchmarked multiple imputation estimates for 1) bias per unit cost (Bias.PUC), 2) standard error per unit cost (SE.PUC), and 3) 95% coverage rates by sampling design at phase IV.



△ denotes BSS-Z design estimates and ● denotes the F design estimates. Horizontal panels show nonresponse mechanism (MAR, MNAR$_X$, MNAR$_Y$) and $cor(X, Y)$ (High,Low) and vertical panels show evaluation measures (Bias.PUC, SE.PUC, and coverage rate). Within each plot, x-axis shows the values of the corresponding evaluation measures, and y-axis represents predictive power of Z on X and Y (i.e. $cor(Z, X)$ and $cor(Z, Y)$). For example, "LH" indicates a low correlation between Z and X ($\rho_{Z,X} = 0.2$) and a high correlation between Z and Y ($\rho_{Z,Y} = 0.8$). Vertical lines are added to aid visual comparison. For Bias.PUC, a zero line is added. For SE.PUC, vertical line is at 0.012. For 95% coverage rates, the vertical line is at 95.

design. For MNAR$_X$, B-MI point estimates were unbiased while those of the GREG were biased. This is an anticipated result since GREG used only Z for weighting adjustment whereas B-MI used both Z and X in the adjustment. Although bias persisted under MNAR$_Y$ for GREG and B-MI estimates, it was largely reduced when Z had high correlation with Y and X. And B-MI estimates had smaller bias than GREG estimates when $\rho_{XY} = H$, regardless the strength of Z. When bias persisted, the adaptive designs outperformed the fixed design for bias and cost-effectiveness. The differences in the biases of the sample designs support claims that, when MAR assumption is violated, adaptive designs are superior in general.

When the MAR assumption was violated, adaptive designs consistently had more favorable bias properties, hence better coverage rates. When the biases were small, both GREG and B-MI estimates often produced confidence intervals with the nominal coverage levels. When the point estimates were seriously biased, both variance estimators produced confidence intervals that covered at far less than the nominal rate. Poor coverage rates are expected for GREG estimates under MNAR$_Y$, especially when $\rho_{Z,Y}$ is low. This is because the MAR assumption is required for the post-survey adjustment to arrive at unbiasedness of the point estimate. B-MI estimates had better coverage rates under MNAR$_Y$. The differences in the coverage rates of the two estimators are consistent to support claims that, when MAR assumption is violated, B-MI estimators are superior in general.

Adaptive designs also produced smaller variances on point estimates compared to the fixed design. For B-MI estimators, the differences in general were small especially when Z was strong predictor of Y. The differences on variance estimates for the designs were more pronounced under GREG estimators. Taking into account of both bias and variance, we computed root mean square error (RMSE). The results suggested that the RMSE by both WT and MI performed well in single stage samples estimates of survey mean; B-MI performs much better when MAR assumption was violated.

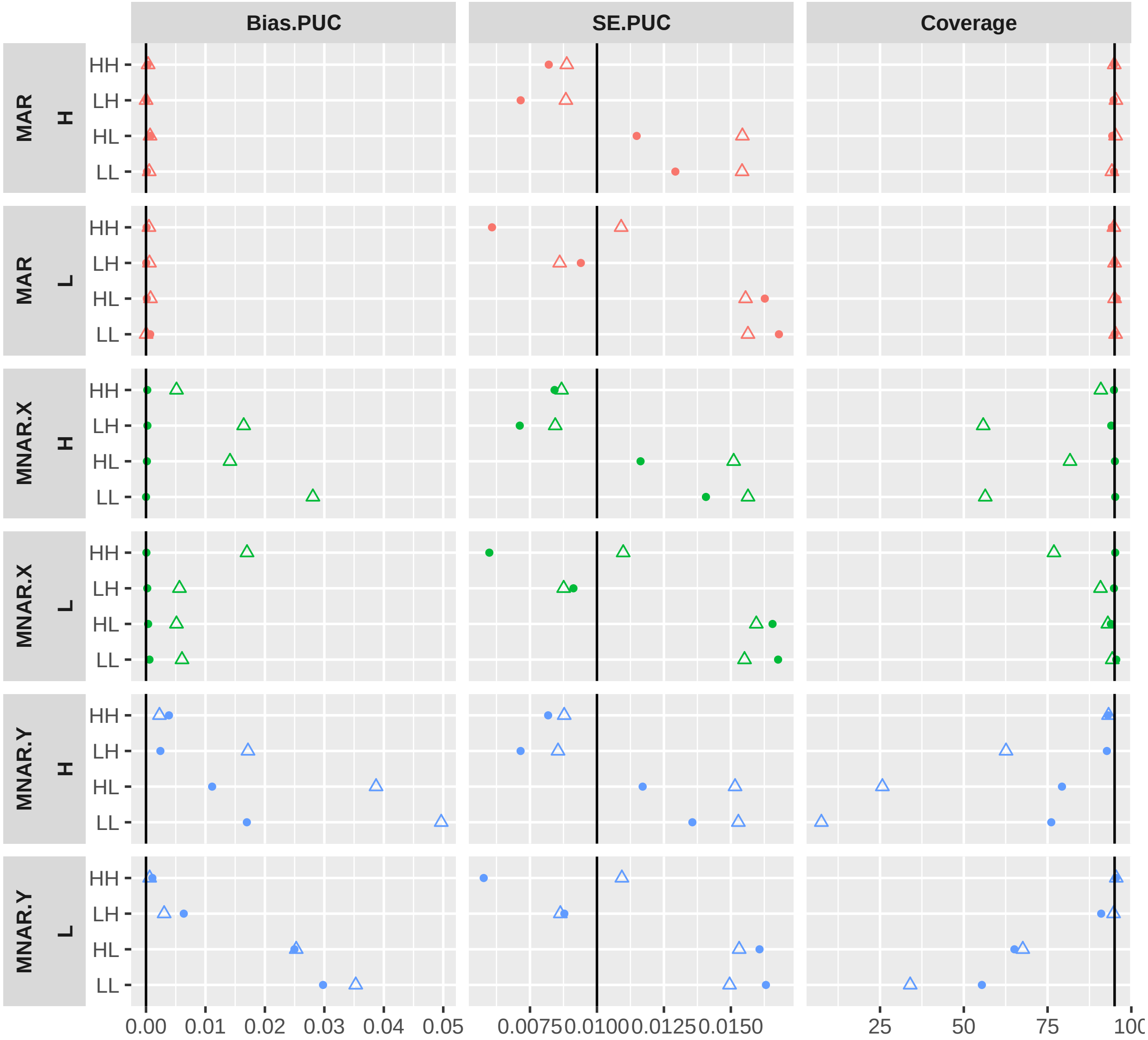
Overall, adaptive designs were more cost-effective than the fixed design when the MAR assumption was violated and weighting variables were weak. Specifically, for GREG estimates adpative designs offered better quality per unit cost under MNAR$_X$ and MNAR$_Y$. B-MI strategy minimized bias effectively for both types of sample designs, the cost-effectiveness of adaptive designs were more evident under MNAR$_Y$ when $\rho_{Z,Y} = L$. For both adaptive and fixed designs, GREG and B-MI estimates were comparable under MAR, and B-MI outperformed GREG estimates under MNAR with $\rho_{X,Y} = H$ or under HHL scenario. For adpative sampling designs, B-MI estimates outperformed GREG estimates in cost-effectiveness. For the fixed design, B-MI outperforms GREG with the exception of when $\rho_{Z,Y}$ was high.

This article does not intend to answer the question of whether we should weight or impute for unit nonresponse. Both methods have their strengths, weaknesses and applicability. The point here is that with budget shortage and increasing nonresponse, strategies that can effectively capitalize on the available auxiliary data become increasingly important in minimizing the survey error and maximizing the cost-effectiveness.

Our simulation results showed that B-MI (using both Z and X information) outperformed GREG weighting (using only Z) in reducing bias and variance of the population mean estimate, for both the fixed design and adaptive design, and under both ignorable and nonignorable nonresponse. We attribute this to two factors: First, B-MI takes advantage of the correlation structure among auxiliary variables which provides better variance reduction than GREG weighting strategy. Second, B-MI is model-based. It models survey variables by their distributions and therefore is able to better predict the values of nonrespondents. In contrast, the GREG weighting strategy uses same variables to develop weights to correct the entire dataset. However, B-MI is only applicable to situations where a micro-level benchmark data is available. Weighting, on the other hand, needs relatively little information on non-respondents. The choice of weighting versus imputing should be a data-driven decision.

While variance estimation is relatively straightforward for the nonresponse adjustment, the development of cost models has received less attention among statisticians. James, et. al. (1997)

Figure 4.5: BSS-Z design GREG and B-MI estimates for 1) bias per unit cost (Bias.PUC), 2) standard error per unit cost (SE.PUC), and 3) 95% coverage rates at phase IV.



△ denotes GREG estimates and • denotes the B-MI estimates. Horizontal panels show nonresponse mechanism (MAR, MNAR$_X$, MNAR$_Y$) and $cor(X, Y)$ (High,Low) and vertical panels show evaluation measures (Bias.PUC, SE.PUC, and coverage rate). Within each plot, x-axis shows the values of the corresponding evaluation measures, and y-axis represents predictive power of Z on X and Y (i.e. $cor(Z, X)$ and $cor(Z, Y)$). For example, "LH" indicates a low correlation between Z and X ($\rho_{Z,X} = 0.2$) and a high correlation between Z and Y ($\rho_{Z,Y} = 0.8$). Vertical lines are added to aid visual comparison. For Bias.PUC, a zero line is added. For SE.PUC, vertical line is at 0.01. For 95% coverage rates, the vertical line is at 95.

developed a spreadsheet-based survey budgeting system for the National Center for Education Statistics for computer assisted telephone interview (CATI) surveys (James et al., 1997). A set of input parameters are: 1) a project time schedule broken down by activity, e.g. survey design, staff recruitment, data collection, data processing, report preparation. 2) sample design factors, e.g. sample targets, average interview length, etc., broken down by strata or subpopulation 3) projected sample performance factors, e.g. incidence rates, response rates, etc. 4) CATI factors, e.g. number of calling centers, workload distribution across centers, paid hours and calling hours per shift, interviewer/supervisor ratios, etc. 5) unit cost for field and office personnel (including fringes), telephone time, equipment printing, postage, etc. However, we are unable to find literature on the performance evaluation and survey design examples of this system.

Judkins, et. al. (1990) discussed the difficulty in fitting a complex cost model using both administrative data and data from special cost studies, and the difficulty of optimization efforts with simple cost considerations. The main reason of these difficulties is that the purposes of tracking cost by survey organizations do not coincide with the needs of cost modelers. The cost modeler would like to associate costs with individual cases whereas costs monitored by organizations focus on paying the line items and field staff (Judkins, et. al. 1990). Tracking costs at the case level often requires more complicated approach to recording time and expense which requires the design of project specific applications. All of these issues limit the development and application of cost data on survey design and evaluation.

In this study, we use a simple model that relates the cost to the inverse of the response propensity estimated for the sampled unit. Although it does not reflect all of the small subtleties of a large scale data collection such as interviewer variability, scale economies, learning efficiencies, and extreme efforts to convert initial nonrespondents, this simple approach has the benefit of being generalizable from one survey to another. We assume the volume of work (i.e. sampling, operation, and data collection cost) for subjects with the same response probability remain the same regardless of sample design. In some situations the variable cost of a sampled subject is inversely proportional to its response propensity regardless of sample design, but sampling and operation costs drastically differs depending on the sample design. In such cases, the presented cost model would only account for partial differences on cost between the two designs, and the benefits on error per unit costs for adaptive designs may change. This awaits further investigation.

We show the findings of one sample mean. When applied to multiple outcome variables, one can aggregate the bias or RMSE over all variables. The results of this study give practitioners of conventional sample design empirical evidence that adaptive sampling design is both cost effective and less susceptible to unknown missingness mechanisms.

CHAPTER 5

Conclusion

This dissertation develops a new inferential paradigm to improve survey data quality, by capitalizing on benchmark data which captures the target population of interest. The inferential paradigm consists of sampling and imputation stages. This paradigm improves survey inference for surveys with rich micro-level auxiliary data and surveys that use respondents of other surveys as a sampling frame. The new paradigm first improves respondent representativeness by benchmarked sequential sampling that conforms the frame and covariate information of the focal survey respondents to those of the benchmark. With improved respondent representativeness, benchmarked multiple imputation recovers the population information, leading to better quality survey estimates that are less susceptible to bias under unknown nonresponse mechanisms. A new subject-level cost model is used to evaluate the cost-effectiveness of the proposed paradigm relative to the cost-effectiveness of a traditional paradigm. A traditional paradigm is referred to a fixed sampling design with post-survey weighting adjustment.

5.1 Chapter Overview

Improving respondent representativeness by adaptive design is associated with reducing nonresponse bias (Sarndal 2014, Schouten, et. al. 2016). Current methods for adaptive designs focus on data collection. Adaptive data collection designs are primarily a nonresponse follow-up strategy which applies differential data collection protocols to different subgroups of the nonrespondents. Although intricate strategies targeting nonresponse follow-up improve respondent representativeness, it is a task that not only complicates the inferential process but also inflicts additional cost (i.e. costs related to nonresponse follow-up). In contrast, this dissertation develops adaptive sampling designs that obtains representative respondents through differential sampling probabilities for over- and under-represented subjects while maintaining a coherent data collection protocol. Two adaptive sampling designs are proposed, BSS-Z and BSS-X. Guided by the benchmark, a propensity score based sampling probability is used to tailor the sampling decision sequentially, attenuating the impact of undesirable nonresponse mechanisms. In a multi-replicate survey setting, the BSS-Z method sequentially conforms the frame variables of focal survey to those of the benchmark, improving the representativeness of the respondent data. Employing a similar mechanism, the BSS-X method conforms not only the frame but also survey covariates of focal survey to those of the benchmark. Both sampling designs are evaluated by simulation experiments to mimic adaptive designs under various nonresponse mechanisms, including two types of not missing at random (NMAR) mechanisms, NMAR$_X$ and NMAR$_Y$. The results show that the respondent representativeness improves at each successive sampling phase. The greatest benefit to these sampling paradigm is that representative respondent pool maintains a similar variance-covariance structure to that of the benchmark, thereby producing less biased descriptive statistics.

Prior studies on unit nonresponse imputation obtained mixed results, perhaps in part caused by the fact that imputation models were built by the respondent data alone. Focal survey respondent data often bear unknown nonresponse patterns, thereby producing mixed results on inferences. In this dissertation, the imputation of unit nonresponse is guided by the benchmark. The benchmarked multiple imputation (B-MI), implemented after obtaining representative respondent data, are better able to recover the population structure and eliminates bias not only under ignorable nonresponse (MAR) but also under one type of nonignorable nonresponse (NMAR$_X$). We implemented Multivariate Imputation for Chained Equation (MICE) to perform the benchmarked multiple imputation.

The imputation models iteratively fit to the benchmark data and the cumulative respondent data, predicting missing information for unit nonrespondents to construct a completed dataset, i.e., respondent data and imputed nonrespondent data. The completed data preserves a high level of population structure with respect to marginal distribution and joint distribution, although biases of the estimates are not completely eliminated under MNAR.Y missingness. The point estimates of the sample mean are unbiased for both MAR and NMAR$_X$. The greatest benefit of the proposed approach is nonresponse bias reduction under NMAR$_X$ and NMAR$_Y$ missingness.

Conventional wisdom suggests that the reduced error often is a tradeoff from increased cost. We use simulation to examine the cost and error implications under a single-stage four-phase simple random sample design. We developed a cost model and evaluated the cost-effectiveness of proposed paradigm. The subject-level cost model is nonlinear, stochastic, and inversely proportionate to subjects response propensity. For error estimation we presented two variance estimators, a current practice for post-survey adjustments (i.e., GREG estimate), and an alternative adjustment (i.e., benchmarked MI estimate). The benchmarked MI offers some insights for additional gains over GREG on bias reduction in situations where micro-level data and survey covariates are available.

Overall, while we found that the traditional fixed sampling and weighting adjustments outperformed the proposed strategy on cost-effectiveness when missingness is MAR, the proposed strategy outperformed the traditional strategy when missingness is NMAR. Although bias persists under NMAR, the adaptive designs outperformed the fixed design for both bias reduction and cost-effectiveness. In practice, it is not possible to assess whether the unit nonresponse is MAR or NMAR without obtaining additional data for nonrespondents (Brick, 2013). But where real world limits exist, the advantage for the proposed strategy was even more pronounced when auxiliary variables are weak and/or survey variable (Y) is strongly correlated with survey covariates (X). The differences in bias reduction and cost-effectiveness of the sample designs support claims that, when the MAR assumption is violated, adaptive designs are superior in general.

When the MAR assumption is violated, adaptive designs consistently have more favorable bias properties, thereby better coverage rates for confidence intervals. When the point estimates are seriously biased, Benchmarked MI estimates have better coverage rates under NMAR$_Y$. The differences in the coverage rates of the two estimators are consistent to support claims that, when the MAR assumption is violated, Benchmarked MI estimators are superior in general.

In summary, the proposed adaptive designs are more cost-effective than the fixed sampling design when the MAR assumption is violated and weighting variables are weak predictors of survey outcome variables. Our research has important implications for the era of increasing survey cost and increasing availability of digital data, e.g., administrative data, medical records, and paradata. Combining data from various sources to produce information not available from a single data source is not only inevitable but also sensible with respect to time and cost. The time is ripe for a new path forward. The proposed strategy is simple, straightforward and readily applicable with current statistical software. Most importantly, the proposed inferential paradigm would serve as an alternative and cost-effective survey design strategy in improving the quality of survey inference.

5.2 Limitations

The issues of incomparability and issues of modelling are common challenges in the research on combining surveys (Schenker and Raghunathan, 2007). The sources of incomparability could come from the modes of interviewing, the survey contexts, the sample design, the survey questions, and the types of respondents and/or the sources of respondent information. The modeling issues for combining surveys range from variable selection to the forms of the models, and in some cases, small sample sizes.

For example, Schenker et. al. (2010) used clinical information from the National Health and Nutrition Examination Survey (NHANES) to improve on analyses of the National Health Interview Survey (NHIS), adapting an imputation-based strategy. NHIS respondents provided information from their memory for themselves and household members whereas NHANES respondents received physical health examination. Such incomparability of two survey designs in respondents (self vs. proxy) and sources of respondent information (self-reported vs. examination) requires careful thoughts when models fitted to the NHANES are applied to the data from the NHIS.

Incomparability and its related issues limit our methods and deserve further investigation.

5.3 Future Research

Here we list some extensions that one could pursue.

In chapters 2, 3, and 4 we illustrated our strategy for a simple random sample design. Surveys nowadays, however, are often multipurpose with complex sample designs. Extending the idea of benchmarked sequential sampling and benchmarked multiple imputation strategy to account for complex survey design could further generalize our method.

For example, the sample design of the National Health Interview Survey (NHIS) is a multistage area probability sample where the first stage consists of strata formed by primary sampling units (e.g. counties), the second stage are clusters formed by secondary sampling units (e.g. housing units) and at the third stage black, Hispanic and Asian are oversampled with a higher rate.

When using NHIS as the benchmark, a focal survey may have the same multistage sample design, or may have a different sample design. In the first case, the complex sample design is identical for both the benchmark and the focal survey. A more realistic and generalized situation is the second case where the focal survey has a different complex survey design from the benchmark survey.

In chapter 4 we evaluated the cost and error properties of our strategy. We compared the errors from the benchmarked multiple imputation and the GREG estimates using the empirical variances of the adaptive design. A more technically involved extension would be to develop the analytical variance formula of the adaptive design. The development of such formula enable reserachers to evaluate the error of their survey estimates in a direct fashion with much less computation.

Recall that, at each mitigation and imputation iteration, our method monitors the convergence between the benchmark and the focal survey on frame variables alone (BSS-Z) or frame variables and the survey covariates (BSS-X). A direct extension of our method is to monitor the convergence of survey outcome variables (Y). Survey outcome variables are predicted for the benchmark data and the nonrespondent data using imputation models iteratively fitted to all the observed data in the benchmark and the focal survey. Monitoring the changes in differences of the predicted survey outcome variables between the benchmark and the focal survey across phases may provide information on the necessity of further data collection.

Quantifying the relationship between these changes and the cost may serve as a tool for cost-and-error optimization. Also, similar estimates on survey outcome variables between the benchmark and the focal survey suggest information saturation for imputation models. Quantifying the changes of predicted survey outcome variables may be developed into a survey stopping rule. For example, one could stop the survey data collection when the difference between the benchmark and the focal survey is equal to or smaller than a predefined distance measure. And moreover, this quantification may result in the development of a new representativeness indicator or bias indicator that measures the similarity of the benchmark and the focal survey.

BIBLIOGRAPHY

[1] Abayomi K, Gelman A and Levy M. “Diagnostics for multivariate imputations.” *Journal of the Royal Statistical Society, series C* 2008; 57(3): 273–291

[2] The American Community Survey, http://www.census.gov/acs/www/

[3] Austin PC. “Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched sample.” *Statistics in Medicine* 2009; 28: 3083–3107.

[4] Azur MJ, Stuart EA, Frangakis C, Leaf PJ. “Multiple imputation by chained equations: what is it and how does it work?” *Int J Methods Psychiatr Res* 2011; 20(1):40–49.

[5] Barclay S, Todda C, Finlayb I, Grande G, Wyattc P. “Not another questionnaire! Maximizing the response rate, predicting non-response and assessing non-response bias in postal questionnaire studies of GPs.” *Family Practice* 2002; 19(1), 105–111.

[6] Beaumont JF. “An estimation method for nonignorable nonresponse.” *Survey Methodology* 2000; 26(2), 131–136.

[7] Bethlehem JG. “Reduction of nonresponse bias through regression estimation.” *Journal of Official Statistics* 1988; 4(3), 251–260.

[8] Bootsma-van der Wiel A, van Exel E, de Craen AJM, Gussekloo J, Lagaay AM, Knook DL, Westendorp RGJ. “A high response is not essential to prevent selection bias: Results from the Leiden 85-plus study,” *Journal of Clinical Epidemiology* 2002; 55(11), 1119–1125.

[9] Botman SL and Moriarity C. “Design and estimation for the National Health Interview Survey, 1995–2004”, National Center for Health Statistics. *Vital Health Stat* 2000; DHHS Publication No. (PHS) 2000-1330. 2(130).

[10] BRFSS 2009 public use microdata files. Available for download at http://www.cdc.gov/brfss/technical_infodata/surveydata/2009.htm [accessed on February 15, 2012]

[11] Brick JM, Kalton G. “Handling missing data in survey research.” *Statistical Methods in Medical Research* 1996; 5: 215–238.

[12] Brick JM, Jones ME. “Propensity to respond and nonresponse bias.” *Metron* 2008; 66(1), 51–73.

[13] Brick JM. “Unit nonresponse and weighting adjustments: A critical review.” *Journal of Official Statistics* 2013; 29(3): 329–353.

[14] Bryant KJ, Weidman L. “Developing cost models for CATI surveys.” *Proceedings of Survey Research Methods Section, American Statistical Association* 1987, pp767–771.

[15] Cochran WG. *Sampling techniques.* Wiley: New York, 1977.

[16] Collins LM, Schafer JL, Kam CM. “A comparison of inclusive and restrictive strategies in modern missing data procedures.” *Psychological Methods*. 2001; 6:330–351.

[17] Couper MP, Wagner J. “Using Paradata and adaptive Design to Manage Survey Nonresponse,” Int. Statistical Inst.: Proc. 58th World Statistical Congress, 2011, Dublin (Session IPS033), 542–548.

[18] Curtin R, Presser S, Singer E. “The Effects of Response Rate Changes on the Index of Consumer Sentiment,” *The Public Opinion Quarterly*. 2000; 64(4), 413–428.

[19] D’Agostino Jr. RB. “Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group.” *Statistics in Medicine* 1998; 17, 2265–2281.

[20] Daniels MJ, Hogan JW.“Reparameterizing the pattern mixture model for sensitivity analyses under informative dropout.” *Biometrics* 2004; 56(4): 1241–1248.

[21] de Leew ED, de Heer W. “Trends in household survey nonresponse: a longitudinal and international comparison”, in *Survey Nonresponse* 2002; eds. Groves RM, Dillman DA, Eltinge JL, and Little RJA, New York: John Wiley & Sons, 41–54.

[22] Deville JC, Särndal CE.“Calibration Estimator in Survey Sampling.” *JASA* 1992, 87(418): 376–382.

[23] Fahimi M, Link M, Schwartz DA, Levy P, Mokdad A. “Tracking chronic disease and risk behavior prevalence as survey participation declines: statistics from the Behavioral Risk Factor Surveillance System and other national surveys.” *Prev Chronic Dis* 2008; 5(3), A80.

[24] Gelman A. “Struggles with survey weighting and regression modeling.” *Statistical Science* 2007; 22(2), 153–164.

[25] Glynn RJ, Laird NM, Rubin DB. “Selection modeling versus mixture modeling with nonignorable nonresponse.” in *Drawing inferences from self-selected samples* 1986; 115–142. Springer.

[26] Glynn RJ., Laird NM, Rubin DB. “Multiple imputation in mixture models for nonignorable nonresponse with follow-ups.” *Journal of the American Statistical Association* 1993; 88(423): 984–993.

[27] Greenlees JS, Reece WS, Zieschang KD. ”Imputation of Missing Values When the Probability of Response Depends on the Variable Being Imputed.” *Journal of the American Statistical Association* 1982; 251–261.

[28] Groves RM, Couper M. *Nonresponse in Household Interview Surveys.* Wiley: New York, 1998.

[29] Groves RM, Singer E, Corning A. “Leverage-Saliency theory of survey nonresponse: Description and an illustration.” *Public Opinion Quarterly* 2000; 64(3): 299–308.

[30] Groves RM., et al. *Survey nonresponse.* Wiley: New York, 2002.

[31] Groves RM. *Survey Errors and Survey Costs.* Wiley: New York, 2004.

[32] Groves RM, Heeringa SG. “Adaptive design for household surveys: tools for actively controlling survey errors and costs.” *Journal of the Royal Statistical Society, Series A*. 2006; 169(3):439–457.

[33] Groves RM. “Nonresponse rates and nonresponse bias in household surveys.” *The public Opinion Quarterly* 2006; 70(5):646–675.

[34] Groves RM, Peytchev E. “The impact of nonresponse rates on nonresponse bias,” *The Public Opinion Quarterly*. 2008; 72(2):167–189.

[35] Hansen M, Madow W, Tepping B. “An evaluation of model-dependent and probability-sampling inferences in sample surveys.” *Journal of the American Statistical Association* 1983; 78(384):776–793.

[36] Hahn J, Hirano K, Karlan D. “Adaptive experimental design using the propensity score.” *Journal of Business and Economic Statistics*. 2011; 29(1):96–108.

[37] He Y, Zaslavsky AM, Landrum MB, Harrington DP, Catalano P. “Multiple imputation in a large-scale complex survey: A practical guide”. *Statistical Methods in Medical Research* 2009; 1-18.

[38] James, B, Krotki, K, and Rathbun, A. “A Survey Budgeting System”, *Proceedings of the Social Statistics Section (Survey Research Methods Section), American Statistical Association* 1997; 719–723

[39] Jordan LA, Marcus AC, and Reeder LG. “Response Styles in Telephone and Household Interviewing: A Field Experiment.” *Public Opinion Quarterly* 1980; 44(2):210–222.

[40] Judkins, D, Waksberg, J, Northrup, D. "Cost Functions for NHIS and implications for design", *Proceedings of Survey Research Methods Section, American Statistical Association* 1990; 34–43.

[41] Kalsbeek WD, Botman SL, Massey JT, Liu PW. "Cost-Efficiency and the Number of Allowable Call Attempts in the National Health Interview Survey." *Journal of Official Statistics* 1994; 10(2):133–152.

[42] Kalton G, Maligalig DS. "A Comparison of Methods of Weighting Adjustment for Nonresponse." *Proceedings of the 1991 Annual Research Conference* 1991; 409–428.

[43] Keeter S, Miller C, Kohut A, Groves R, Presser S. "Consequences of Reducing Nonresponse in a National Telephone Survey." *The Public Opinion Quarterly* 2000; 64(2):125–148.

[44] Kennickell, AB. "Imputation of the 1989 Survey of Consumer Finances: Stochastic Relaxation and Multiple Imputation." presented at the 1991 Joint Statistical Meeting, 1991 Atlanta, Georgia.

[45] Kish L. *Survey Sampling.* Wiley: New York, 1965.

[46] Kohut A, Keeter S, Doherty C, Dimock M, Christian L. "Assessing the Representativeness of Public Opinion Surveys." *Pew research report* 2012; Available at http://www.people-press.org/2012/05/15/assessing-the-representativeness-of-public-opinion-surveys/, [accessed 15 July 2012].

[47] Lagerström BO, Björshol E. "Fieldwork Monitoring. Indicators and Data Control", Work plan and Preliminary Findings - Pilot, Statistics Norway. 2010, Available at http://www.risq-project.eu/papers/RISQ-Deliverable-8-1.pdf, [accessed 15 July 2012].

[48] Lepkowski JM, Mosher WD, Groves RM, et al. "Adaptive design, weighting, and variance estimation in the 2006–2010 National Survey of Family Growth," National Center for Health Statistics. *Vital Health Stat* 2013; 2(158).

[49] Little RJA. "Models for nonresponse in sample surveys." *Journal of the American Statistical Association* 1982; 77(378):237–250.

[50] Little RJA. "Survey nonresponse adjustments for estimates of means." *International Statistical Review* 1986; 54(2):139–157

[51] Little, Roderick JA. "Modeling the drop-out mechanism in repeated-measures studies." *Journal of the American Statistical Association* 1995; 90(431):1112–1121.

[52] Little RJA, Rubin D.B. *Statistical Analysis with Missing Data*. Wiley: New York, 2002.

[53] Little R, Vartivarian S. “On weighting the rates in non-response weights.” *Statistics in Medicine* 2003; 22(9):1589–1599.

[54] Little R, Vartivarian S. “Does Weighting for Nonresponse Increase the Variance of Survey Means?” *Survey Methodology* 2005; 31(2):161–168.

[55] Little, RJA, Rubin DB. Discussion for “Unit Nonresponse and Weighting Adjustments: A Critical Review.” *Journal of Official Statistics* 2013; 29(3):363–366.

[56] Locander WB, Burton JP. “The Effect of Question Form on Gathering Income Data by Telephone.” *Journal of Marketing Research* 1976; 13(2):189–192.

[57] Lundström, Sixten, and Carl-Erik Särndal. “Calibration as a standard method for treatment of nonresponse.” *Journal of Official Statistics* 1999; 15(2):305–327.

[58] Merkle D, Edelman M. “Nonresponse in exit polls: A comprehensive analysis,” In *Survey nonresponse* 2002; eds. R. M. Groves, D. A. Dillman, J. L. Eltinge, and R. J. A. Little, New York: John Wiley & Sons, pp. 243–258.

[59] Nandram B, Choi JW. “Hierarchical Bayesian nonresponse models for binary data from small areas with uncertainty about ignorability.” *Journal of the American Statistical Association*, 2002; 97(458):381–388.

[60] Nandram B, Choi JW. “A Bayesian analysis of a proportion under nonignorable nonresponse.” *Statistics in Medicine* 2002; 21(9):1189–1212.

[61] Nandram B, Choi JW. “Hierarchical Bayesian nonignorable nonresponse regression models for small areas: an application to the NHANES data.” *Survey Methodology* 2005; 31(1):73–84.

[62] Nelson DE, Powell-Griner E, Town M, Kovar MG. “A comparison of national estimates from the National Health Interview Survey and the Behavioral Risk Factor Surveillance System.” *American Journal of Public Health* 2003; 93:1335–1341.

[63] NHIS 2009 public use microdata files, available for download at http://www.cdc.gov/nchs/nhis/nhis_2009_data_release.htm [accessed on February 15, 2012]

[64] Olsen K, Groves RM. “An Examination of Within-Person Variation in Response Propensity over the Data Collection Field Period.” *Journal of Official Statistics* 2012; 28:29–51.

[65] Parker JD, Schenker N, Ingram DD, Weed JA, Heck KE, Madans JH. “Bridging between Two Standards for Collecting Information on Race and Ethnicity: an Application to Census 2000 and Vital Rates.” *Public Health Reports* 2004; 119:192–205.

[66] Peytchev A, Rosen J, Riley S, Murphy J, Lindblad M. “Reduction of Nonresponse Bias in Surveys through Case Prioritization,” *Survey Research Methods* 2010; 4(1):21–29.

[67] Peytchev A. “Adaptive design in telephone survey data collection.” In *Proceedings of Household Survey Nonresponse Workshop*. 2010; available at http://www.nonresponse.org/index.php?fl=2&lact=1&bid=621&avtor=509&parent=3, [accessed on July 15, 2012]

[68] Raghunathan TW, Lepkowski JM, Van Hoewyk J, Solenbeger P. “A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology* 2001; 27:85–95.

[69] Raghunathan TE, Xie D, Schenker N, Parsons V, Davis W, Rancourt E, Dodd K. “Combining Information from Multiple Surveys for Small Area Estimation: A Bayesian Approach.” *Journal of American Statistical Association* 2007; 102(478):474–486.

[70] Rässler S, Schnell R. “Multiple imputation for unit-nonresponse versus weighting including a comparison with a nonresponse follow-up study.” No 65/2004, Discussion Papers, Statistics and Econometrics, University Erlangen-Nuremberg,

[71] Reist B. “Early Experience of Adaptive Design Work in the NSCG.” Presentation at the Federal Economic Statistics Advisory Committee Meeting, June 2014, available at ftp://ftp.census.gov/adrm/fesac/2014-06-13_reist.pdf, [accessed on Sep 30, 2014]

[72] Rosenbaum PR, Rubin DB. “The central role of the propensity score in observational studies for causal effects.” *Biometrika* 1983; 70:41–55.

[73] Rosenbaum PR, Rubin DB. “Constructing a control group using multivariate matched sampling methods that incorporate the propensity score.” *American Statistician* 1985; 39: 33–38.

[74] Royall R, Herson J. “Robust Estimation in Finite Populations I.” *Journal of the American Statistical Association* 1973; 68(344):880–889.

[75] Rubin DB. *Multiple imputation for nonresponse in surveys*. Wiley: New York, 1987.

[76] Rubin DB, Schenker N. “Multiple Imputation in Health Care Databases: An Overview and Some Applications.” *Statistics in Medicine* 1991; 10:585–598.

[77] Rubin DB. “Using propensity scores to help design observational studies: Application to the tobacco litigation.” *Health Services and Outcomes Research Methodology* 2002; 2:169–188.

[78] Rubin DB. “The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials.” *Statistics in Medicine* 2007; 26(1): 20–36.

[79] Särndal CE, Lundström S. *Estimation in surveys with nonresponse.* Wiley: New York, 2005.

[80] Särndal CE. “The 2010 Morris Hansen Lecture Dealing with Survey Nonresponse in Data Collection, in Estimation,” (with Discussions) by M. Brick and R. Tourangeau, *Journal of Official Statistics* 2011; 27(1):1–21.

[81] Särndal CE, Lundquist P. "Accuracy in Estimation with Nonresponse: A Function of Degree of Imbalance and Degree of Explanation." *J Surv Stat Methodology* 2014; 2(4):361–387.

[82] Schafer JL. *Analysis of Incomplete Multivariate Data.* Chapman & Hall: London, 1997.

[83] Schafer JL. "Multiple imputation: a primer." *Statistical Methods in Medical Research.* 1999; 8:3–15.

[84] Schafer JL. "Multiple imputation in multivariate problems when the imputation and analysis models differ". *Statistica Neerlandica* 2003; 57:19–35.

[85] Schenker N, Gentleman JF, Rose D, Hing E, Shimizu IM. "Combining estimates from complementary surveys: a case study using prevalence estimates from national health surveys of households and nursing homes." *Public Health Reports* 2002; 117:393–407.

[86] Schenker N, Parker JD. "From single-race reporting to multiple-race reporting: using imputation methods to bridge the transition." *Statistics in Medicine* 2003; 22:1571–1587. DOI: 10.1002/sim.1512

[87] Schenker N, Raghunathan TE. "Combining information from multiple surveys to improve measures of health." *Statistics in Medicine* 2007; 26:1802–1811.

[88] Schenker N, Raghunathan TE, Bondarenko I. "Improving on analyses of self-reported data in a large-scale health survey by using information from an examination-based survey." *Statistics in Medicine* 2010; 29(5):533–545.

[89] Schouten B, Cobben F. 2007. "R-indexes for the comparison of different fieldwork strategies and data collection modes," Discussion paper 07002, Voorburg/Heerlen, The Netherlands: Statistics Netherlands.

[90] Schouten B, Cobben F, Bethlehem J. "Indicators of Representativeness of Survey Nonresponse," *Survey Methodology* 2009; 35:101–113.

[91] Schouten B, Shlomo N, Skinner C. "Indicators for monitoring and improving representativeness of response." *Journal of Official Statistics* 2011; 27:231–253.

[92] Schouten B, Calinescu M, Luiten A. "Optimizing quality of response through adaptive survey designs." *Survey Methodology* 2013; 39(1):29–58.

[93] Schouten B, Cobben F, Lundquist P, Wagner J. "Does more balanced survey response imply less non-response bias?" *Journal of Royal Statistical Society, Series A* 2016; 179(3): 727–748.

[94] Shimizu I, Lan F. "Approximation of Variable Costs for the National Health Interview Survey", *Proceedings of Survey Research Methods Section, American Statistical Association* 2001; 3197–3202.

[95] Stuart EA, Azur M, Frangakis C, Leaf P. “Multiple imputation with large data sets: a case study of the Children’s Mental Health Initiative.” *Am J Epidemiol* 2009; 169(9): 1133–1139.

[96] Tang G, Little RJA, Raghunathan TE. “Analysis of multivariate missing data with nonignorable nonresponse.” *Biometrika* 2003; 90(4):747–764.

[97] The Census Planning Database, http://www.census.gov/research/data/planning_database/

[98] Tourangeau R, Plewes T. ”The growing problem of nonresponse”, in *Nonresponse in Social Science Surveys: A Research Agenda*. 2013: eds. Roger Tourangeau and Thomas Plewes, Washington, D.C.: The National Academies Press, pp. 23.

[99] Valliant R, Dever JA, Kreuter F. *Practical Tools for Designing and Weighting Survey Samples* Springer: New York, 2013

[100] van Buuren S., “Multiple Imputation of Discrete and Continuous Data by Fully Conditional Specification,” *Statistical Methods in Medical Research* 2007; 16(3):219–242. http://dx.doi.org/10.1177/0962280206074463

[101] van Buuren S., Groothuis-Oudshoorn K. “mice: Multivariate Imputation by Chained Equations in R.” *Journal of Statistical Software* 2011; 45(3):1–67

[102] Wagner J. “Adaptive Survey Designs to Reduce Nonresponse Bias.” Ph.D. Dissertation, University Of Michigan, 2008.

[103] White IR, Royston P, Wood AM. “Multiple Imputation using Chained Equations: Issues and guidance for practice.” *Statistics in Medicine* 2011; 30(4):377–399. DOI: 10.1002/sim.4067