# STATISTICAL METHODS FOR MISSING DATA IN COMPLEX SAMPLE SURVEYS

by

Rebecca Roberts Andridge

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2009

Doctoral Committee:

Professor Roderick J. Little, Chair
Professor Susan A. Murphy
Professor Trivellore E. Raghunathan
Associate Professor Michael R. Elliott

In memory of Dr. Douglas K. Richardson

# ACKNOWLEDGEMENTS

This dissertation would not be possible without the help of a large cast of characters; what follows is an attempt to offer a few words of thanks.

To my advisor, Rod, thank you for your guidance and support throughout the years of work that produced this dissertation; you are quite simply the best advisor a student could have. To the members of my dissertation committee, Raghu, Mike, and Susan, thank you for your input and help in making this dissertation all that it could be, with special thanks to Susan for filling in at the last minute. To my lost committee member, Dr. Bob Groves, thank you for your guidance and career advice, and good luck at the Census Bureau; UM will miss you.

To all my fellow UM Biostat students, thank you for making this department a friendly place to learn and work, a place that I didn't dread coming in to each day. To Tracy Green Franklin, thank you for bringing laughter and fun to the MS program, and know that Team Shafty will never be forgotten. To Annie Noone, thank you for suffering through the ups and downs of graduate school (and life) with me, including the boot, crushed skulls, and of course, "please see me."

To my amazing collaborators in the UM Department of Health Behavior, Health Education, thank you for supporting me (quite literally) and creating a workplace that has been so wonderful for the last six (six!) years that I am sad to leave. To Dr. Noreen Clark, thank you for being a supportive and encouraging mentor, and bringing together a research team that not only produced great work but also actually genuinely liked each other and could practically read each others' thoughts. To Dan Awad, thanks for so many years of making it so easy to work with you that I sometimes forgot it was work at all, and for those frequent days when we laughed so hard that it hurt. To Julie Dodge and Lara Thomas, thank you for not only being a pleasure to work with but also being joys to talk to, whether about travel, children, or the future.

To my family, thank you for your undying love and support. To my parents, David and Mary Roberts, thank you for knowing all along that this was where I'd end up, but not ever saying, "I told you so." To my brother Matthew, thank you for random photos (snowing ash in Santa Barbara?) that had the power to brighten dreary PhD days (P.S. I beat you!). To my baby sister Kathryn, thank you for being you, thus reminding me that there are more important things in life than deriving an MLE, and for not killing me for referring to you as my baby sister even after you turned 21.

And finally, to my loving husband Nathan, words are not sufficient to convey the thanks you deserve. From baking me cookies while I studied for the quals, to patiently waiting out my PhD-induced crazy spells and offering rational, clear-headed advice, to learning how to explain what I do to non-academics, to taking me on much-needed vacations, to your "equations" on Post-It notes, you have been there for me from the first to last page of this dissertation, and I cannot thank you enough. I love you $\infty$.

# TABLE OF CONTENTS

**CHAPTER**

# LIST OF FIGURES

Figure

# LIST OF TABLES

**Table**

# CHAPTER I

# Introduction

Missing data are a pervasive problem in large-scale surveys, arising when a sampled unit does not respond to a particular question (item nonresponse) or to the entire survey (unit nonresponse). This dissertation addresses two major topics under the umbrella of survey nonresponse: hot deck imputation and evaluation of nonresponse bias. Chapter II contains an extensive review of hot deck imputation, which despite being used extensively in practice has theory that is not as well developed as that of other imputation methods. One of the understudied areas discovered in this review is the topic for the subsequent chapter: Chapter III addresses the use of sample weights in the hot deck. These first chapters concern methods for imputing missing data in the case where (at worst) missingness is at random (MAR) (Rubin, 1976); the final two chapters (IV, V) focus instead on a method for estimating and correcting nonresponse bias when missingness may not be at random (NMAR). Since sample surveys are ubiquitous, not just in the health sciences but far beyond, the work in this dissertation has broad and immediate application to a wide range of survey practitioners.

A common technique for handling item nonresponse is imputation, whereby the missing values are filled in to create a complete data set that can then be analyzed with traditional analysis methods. Chapter II provides an extensive review of a specific type of imputation known as hot deck imputation, which involves replacing missing values with values from a “similar” responding unit. This method is used extensively in practice, but the theory behind the hot deck is not as well developed as that of other imputation methods, leaving researchers and analysts with no clear “correct” way to apply the hot deck and obtain inference from the completed data set. This paper describes various forms of the hot deck, including both *random* and *deterministic* versions, reviews existing research on statistical properties, and suggests some areas for future work.

One of the highlighted areas in the review is the appropriate incorporation of sample weights in the hot deck, and this is the topic of Chapter III. A key feature of complex sample surveys is that the way in which units are selected to participate leads to individual units carrying different weights in subsequent analyses. There is extensive literature on how to use these sample weights when analyzing data; despite this, there is no consensus on the incorporation of these weights when using the hot deck for imputation. The two main approaches that have been recommended, the weighted sequential hot deck (Cox, 1980) and selecting donors with probability proportional to their sample weight (Rao and Shao, 1992), require alteration of the typical hot deck implementation. This makes them either unfeasible or unattractive to users, and for this reason they are uncommon in practice, and users tend to ignore the sample weights in the imputation step. In this part of the dissertation we propose an approach to using the sample weights that does not require any alterations to software and can be easily implemented. We show through simulation that our method performs at least as well as the previously suggested methods, and in fact is superior in certain scenarios. We also demonstrate the method on data from the third National Health and Nutrition Examination Survey (NHANES III).

The final two chapters of the dissertation focus not on a particular type of imputation but on a particular type of missing data mechanism, nonignorable missingness. Chapter IV describes a novel method for assessment of nonresponse bias for the mean of a continuous survey variable $Y$ subject to nonresponse that we call proxy pattern-mixture analysis. We assume that there are a set of covariates $Z$ observed for nonrespondents and respondents, but instead of using these auxiliary data for imputation as Chapters II and III, here we use the data to estimate the potential for nonresponse bias in $Y$. To reduce dimensionality and for simplicity we reduce the covariates $Z$ to a proxy variable $X$ that has the highest correlation with $Y$, estimated from a regression analysis of respondent data. We consider adjusted estimators of the mean of $Y$ that are maximum likelihood for a pattern-mixture model with different mean and covariance matrix of $Y$ and $X$ for respondents and nonrespondents, assuming missingness is an arbitrary function of a known linear combination of $X$ and $Y$. This allows insight into whether missingness may be not at random (NMAR). We propose a taxonomy for the evidence concerning bias based on the strength of the proxy and the deviation of the mean of $X$ for respondents from its overall mean, propose a sensitivity analysis, and describe Bayesian versions of this approach. We propose using the fraction of missing information from multiple imputation under the pattern-mixture model as a measure of nonresponse bias. Methods are demonstrated through simulation and data from the NHANES III.

The proxy pattern-mixture analysis developed in Chapter IV strictly only applies to continuous survey variables, where normality is reasonable. However, categorical outcomes are ubiquitous in sample surveys. In Chapter V we propose an extension of the PPM to binary survey outcomes using probit models. The method is also extended to ordinal outcomes. In addition, the important issue of model misspecification is discussed. The methods are illustrated first through simulation and then by application to NHANES III data.

# CHAPTER II

# A Review of Hot Deck Imputation

## 2.1 Introduction

Missing data are often a problem in large-scale surveys, arising when a sampled unit does not respond to the entire survey (unit nonresponse) or to a particular question (item nonresponse). A common technique for handling item nonresponse is imputation, whereby the missing values are filled in to create a complete data set that can then be analyzed with traditional analysis methods. It is important to note at the outset that the objective of imputation is not to get the best possible predictions of the missing values, but to replace them by plausible values in order to exploit the information in the recorded variables for the incomplete cases (Little and Rubin, 2002). We consider here hot deck imputation, which involves replacing missing values with values from a "similar" responding unit. This method is used extensively in practice, but the theory behind the hot deck is not as well developed as that of other imputation methods, leaving researchers and analysts with no clear "correct" way to apply the hot deck and obtain inference from the completed data set. This paper describes various forms of the hot deck, reviews existing research on statistical properties, and highlights some areas for future work.

Hot deck imputation involves replacing missing values of one or more variables for a nonrespondent (called the recipient) with observed values from a respondent (the donor) that is similar to the nonrespondent with respect to characteristics observed by both cases. In some versions, the donor is selected randomly from a set of potential donors, which we call the donor pool; we call these methods *random hot deck methods*. In other versions a single donor is identified and values are imputed from that case, usually the "nearest neighbor" based on some metric; we call these methods *deterministic hot deck methods*, since there is no randomness involved in the selection of the donor. Other methods impute summaries of values for a set of donors, such as the mean, rather than individual values; we do not consider these as hot deck methods, although they share some common features. We note that our use of "deterministic" describes the way in which a donor is selected in the hot deck, and differs from the use of "deterministic" to describe imputation methods that impute the mean or other non-random value.

There are several reasons for the popularity of the hot deck method among survey practitioners. As with all imputation methods, the result is a rectangular data set that can be used by secondary data analysts employing simple complete-data methods. It avoids the issue of cross-user inconsistency that can occur when analysts use their own missing-data adjustments. The hot deck method does not rely on model fitting for the variable to be imputed, and thus is potentially less sensitive to model misspecification than an imputation method based on a parametric model, such as regression imputation, though implicit assumptions do exist. Additionally, only plausible values can be imputed, since values come from observed responses in the donor pool. There may be a gain in efficiency relative to complete-case analysis, since information in the incomplete cases is being retained. There is also a reduction in nonresponse bias, to the extent that there is an association between the variables defining imputation classes and both the propensity to respond and the variable to be imputed.

Section 2.2 describes some applications of the hot deck in real surveys, including the original application to the Current Population Survey (CPS). Section 2.3 discusses methods for finding "similar" units and creating donor pools. Section 2.4 consider methods for incorporating sampling weights, including weighted hot decks. Section 2.5 discusses hot decks for imputing multivariate incomplete data with monotone and more complex "swiss cheese" patterns of missingness. Theoretical properties of hot deck estimates, such as unbiasedness and consistency, are the focus of Section 2.6. Section 2.7 discusses variance estimation, including resampling methods and multiple imputation. Section 2.8 illustrates different forms of the hot deck on data from the third National Health and Nutrition Examination Survey (NHANES III), drawing comparisons between the methods by simulation. Some concluding remarks and suggestions for future research are provided in Section 2.9.

## 2.2 Examples of the Hot Deck

Historically, the term "hot deck" comes from the use of computer punch cards for data storage, and refers to the deck of cards for donors available for a nonrespondent. The deck was "hot" since it was currently being processed, as opposed to the "cold deck" which refers to using pre-processed data as the donors, i.e. data from a previous data collection or a different data set. At the U.S. Census Bureau, the classic hot deck procedure was developed for item nonresponse in the Income Supplement of the Current Population Survey (CPS), which was initiated in 1947 and has evolved since then (Ono and Miller, 1969; U.S. Bureau of the Census, 2002).

The CPS uses a sequential adjustment cell method to fill in missing items (U.S.

Bureau of the Census, 2002). The main item requiring imputation is the earnings question, but a small fraction (1-4%) missing values of other items relating to demographics, employment status, and occupation are also imputed. Each variable has its own hot deck, and imputation proceeds in a pre-specified order so that items imputed previously may be used to define adjustment cells for later variables. For example, cells to impute labor force items (e.g. employed/not employed) are defined by age, sex, and race. Then industry and occupation is imputed with cells based on age, sex, race, and employment status. Earnings can then be imputed based on age, sex, race, employment status, and industry/occupation. The number of adjustment cells ranges from approximately 100 for employment status to many thousands for earnings estimates. The records within adjustment cell sorted based on geographic location and primary sampling unit, and then values from respondents are used sequentially to impute missing values.

The hot deck is commonly used by other government statistics agencies and survey organizations to provide rectangular data sets for users. For example, the National Center for Education Statistics (NCES) uses different forms of the hot deck and alternative imputation methods even within a survey. Out of twenty recent surveys, eleven used a form of adjustment cell hot deck (sequential or random) while the remaining nine used a form of deterministic imputation (e.g. mean imputation), cold deck imputation, or a Bayesian method for MI. Within the surveys that used the hot deck, many used both random within class imputation and sequential imputation (National Center for Education Statistics, 2002).

The hot deck has been applied in epidemiologic and medical settings, although here parametric imputation methods are more common. Applications of the hot deck and comparisons with other imputation methods include Barzi and Woodward (2004) and Perez, Dennis, Gil, and Rondon (2002) in cross-sectional studies, and Twisk and de Vente (2002) and Tang, Song, Belin, and Unutzer (2005) in longitudinal studies. The lack of software in commonly used statistical packages such as SAS may deter applications of the hot deck in these settings.

Sequential hot deck methods are the most prevalent in applications, but some recent implementations have used more complex matching metrics and better methods for handling multivariate missingness; these methods are described in the following sections.

## 2.3 Methods for Creating the Donor Pool

Hot deck imputation methods share one basic property: each missing value is replaced with an observed response from a "similar" unit (Kalton and Kasprzyk, 1986). Donor pools, also referred to as imputation classes or adjustment cells, are formed based on auxiliary variables that are observed for donors and recipients. We now review the various ways in which donors can be identified. For clarity we initially focus on the use of covariate information $x$ for imputing a single variable $Y$; the case of multivariate $Y$ is discussed in Section 2.5.

### 2.3.1 Adjustment Cell Methods

The simplest method is to classify responding and nonresponding units into imputation classes, also known as adjustment cells, based on $x$ (Brick and Kalton, 1996). To create cells, any continuous covariates are categorized before proceeding. Imputation is then carried out by randomly picking a donor for each nonrespondent within each cell. Cross-classification by a number of covariates can lead to many adjustment cells. An example is imputation of income in the Current Population Survey Outgoing Rotation Group (CPS-ORG), which uses seven variables leading to 11,520 adjustment cells (Bollinger and Hirsch, 2006), some of which contain nonrespondents but no matching respondents. The usual remedy is to drop or coarsen variables until a suitable donor is found. The choice of variables for creating adjustment cells often relies on subjective knowledge of which variables are associated with the item being imputed, and predictive of nonresponse. Groups of "similar" donors could also be created empirically using branching algorithms such as CHAID or CART (Kass, 1980; Breiman and Friedman, 1993), though these methods do not seem to be widely used.

Sparseness of donors can lead to the over-usage of a single donor, so some hot decks limit the number of times $d$ any donor is used to impute a recipient. The optimal choice of $d$ is an interesting topic for research – presumably it depends on the size of the sample, and the interplay between gain in precision from limiting $d$ and increased bias from reduced quality of the matches.

The two key properties of a variable used to create adjustment cells are (a) whether it is associated with the missing variable $Y$, and (b) whether it is associated with the binary variable indicating whether or not $Y$ is missing. Table 2.1, from Little and Vartivarian (2005), summarizes the effect of high or low levels of these associations on bias and variance of the estimated mean of $Y$. Table 2.1 was presented for the case of nonresponse weighting, but it also applies for hot deck imputation. In order to see a reduction in bias for the mean of $Y$, the variables $x$ that define the donor pools must be associated both with $Y$ and with the propensity to respond, as in the bottom right cell of the table. If $x$ is associated with the propensity to respond but not with the outcome $Y$, there is an increase in variance with no compensating reduction in bias, as in the bottom left cell of the table. Using adjustment cells associated with $Y$ leads to an increase in precision, and also reduces bias if the adjustment cell variable is related to nonresponse. Attempts should thus be made to create cells that are homogeneous with respect to the item or items being imputed, and, if propensity is associated with the outcome, also with the propensity to respond.

Creating adjustment cells is not the only way of defining groups of “similar” units. A more general principle is to choose donor units that are close to the nonrespondent with respect to some distance metric. We now review these methods.

### 2.3.2 Metrics for Matching Donors to Recipients

Let $x_i = (x_{i1}, \ldots, x_{iq})$ be the values for subject $i$ of $q$ covariates that are used to create adjustment cells, and let $C(x_i)$ denote the cell in the cross-classification in which subject $i$ falls. Then matching the recipients $i$ to donors $j$ in the same adjustment cell is the same as matching based on the metric

$$d(i,j) = \begin{cases} 0 & j \in C(x_i) \\ 1 & j \notin C(x_i) \end{cases} .$$

Other measures of the “closeness” of potential donors to recipients can be defined that avoid the need to categorize continuous variables, such as the maximum deviation,

$$d(i,j) = \max_k |x_{ik} - x_{jk}|,$$

the Mahalanobis distance,

$$d(i,j) = (x_i - x_j)^T \widehat{Var}(x_i)^{-1}(x_i - x_j),$$

where $\widehat{Var}(x_i)$ is an estimate of the covariance matrix of $x_i$, or the predictive mean,

$$d(i,j) = \left(\hat{Y}(x_i) - \hat{Y}(x_j)\right)^2, \tag{2.1}$$

where $\hat{Y}(x_i)$ is the predicted value of $Y$ for nonrespondent $i$ from the regression of $Y$ on $x$ using only the respondents’ data. Inclusion of nominal variables using these metrics requires conversion to a set of dummy variables.

If all adjustment variables are categorical and main effects plus all interactions between adjustment variables are included in the regression model, predictive mean matching reduces to the adjustment cell method. Subjects with the same $x$ vector will have the same $\hat{Y}$, creating identical donor pools as for the cross-tabulation method. One advantage to defining neighborhoods via the predictive mean is that the variables $x$ that are predictive of $Y$ will dominate the metric, while the Mahalanobis metric may be unduly influenced by variables with little predictive power (Little, 1988). Using generalized linear models such as logistic regression to model the predictive means allow this metric to be used for discrete outcomes as well as continuous ones. The predictive mean neighborhood method has also been proposed in the context of statistical matching (Rubin, 1986).

One a metric is chosen there are several ways to define the set of donors for each recipient. One method defines the donor set for nonrespondent $j$ as the set of respondents $i$ with $d(i, j) < \delta$, for a pre-specified maximum distance $\delta$. A donor is then selected by a random draw from the respondents in the donor set. Alternatively, if the closest respondent to $j$ is selected, the method is called a deterministic or nearest neighbor hot deck. The widely used Generalized Edit and Imputation System uses the nearest neighbor approach, with the maximum deviation metric applied to standardized ranks to find donors (Cotton, 1991; Fay, 1999; Rancourt, 1999). A third method for selecting a donor is developed in Siddique and Belin (2008), where all respondents are eligible as donors but random selection of a donor is with probability inversely proportional to their distance from the recipient, which is defined as a monotonic function of the difference in predictive means.

As previously noted, information about the propensity to respond may help in creating the best adjustment cells. One method is to perform response propensity stratification, whereby the probability of response for a subject $p(x)$ is estimated by the regression of the response indicator on the covariates $x$, using both respondent and nonrespondent data (Little, 1986). As with the predictive mean metric, the predicted probability of response (propensity score, $\hat{p}(x)$) can be calculated for all subjects, and is itself a type of distance metric. Stratification via predictive means and response propensities are compared in the context of the hot deck in Haziza and Beaumont (2007). They show that either metric can be used to reduce nonresponse bias; however only the predictive mean metric has the potential to also reduce variance. Similar results were previously described for cell mean imputation in Little (1986). Thus, for a single variable $Y$, creating cells that are homogeneous with respect to the predictive mean is likely close to optimal; additional stratification by the propensity to respond simply adds to the variance without reducing bias. For a set of $Y$'s with the same pattern and differing predictive means, a single stratifier compromises over the set of predictive means for each variable in the set, as discussed in Section 2.5. Additional stratification by the propensity to respond may reduce bias in this setting.

### 2.3.3 Redefining the Variables to be Imputed

The natural implementation of the hot deck imputes a missing value $y_i$ of a variable $Y$ with the value $y_j$ of $Y$ from a case $j$ in the donor set. This imputation has the attractive property of being invariant to transformations of the marginal distribution of $Y$; for example imputing $Y$ yields the same imputations as imputing $\log Y$ and exponentiating those values. Improvements may result from imputing a function of $Y$ and $x$, rather than $Y$ itself. For example, if $Y$ is strongly correlated with an auxiliary variable $S$ measuring size of the unit, then it may be advantageous to treat the missing variable as the ratio $R = Y/S$. Imputing $\widehat{r_i} = r_j$ from a donor $j$, with the implied imputation $\widehat{y}_i = s_i r_j$ for $Y$, might be preferable to imputing $\widehat{y}_i = y_j$ directly from a donor, particularly if donors are chosen within adjustment cells that do not involve $S$ or are based on a crude categorization of $S$.

## 2.4 Role of Sampling Weights

We now discuss proposals for explicitly incorporating the survey design weights into donor selection.

### 2.4.1 Weighted Sequential Hot Deck

The weighted sequential hot deck procedure (Cox, 1980; Cox and Folsom, 1981) was motivated by two issues: the unweighted sequential hot deck is potentially biased if the weights are related to the imputed variable, and respondent values can be used several times as donors if the sorting of the file results in multiple nonrespondents occurring in a row. This tends to lead to estimates with excessive variance. The weighted sequential hot deck preserves the sorting methodology of the unweighted procedure, but allows all respondents the chance to be a donor and uses sampling weights to restrict the number of times a respondent value can be used for imputation. Respondents and nonrespondents are first separated into two files and sorted (randomly, or by auxiliary variables). Sample weights of the nonrespondents are rescaled to sum to the total of the respondent weights. The algorithm can be thought of as aligning both these rescaled weights and the donors' weights along a line segment, and determining which donors overlap each nonrespondent along the line (Williams and Folsom, 1981). Thus the set of potential donors for a given nonrespondent is determined by the sort order, the nonrespondent's sample weight, and the sample weights of all the donors. The algorithm is designed so that, over repeated imputations, the weighted mean obtained from the imputed values is equal in expectation to the weighted mean of the respondents alone within imputation strata. “Similarity” of donor to recipient is still controlled by the choice of sorting variables.

The weighted sequential hot deck does not appear to have been widely implemented. For example, the National Survey on Drug Use and Health (NSDUH) used it sparingly in the 2002 survey but has since switched to exclusive use of imputation via predictive mean neighborhoods (Grau, Frechtel, and Odom, 2004; Bowman, Chromy, Hunter, Martin, and Odom, 2005).

### 2.4.2 Weighted Random Hot Decks

If donors are selected by simple random sampling from the donor pool, estimators are subject to bias if their sampling weight is ignored. One approach, which removes the bias if the probability of response is constant within an adjustment cell, is to inflate the donated value by the ratio of the sample weight of the donor to that of the recipient (Platek and Gray, 1983). However, this adjustment has drawbacks, particularly in the case of integer-valued imputed value $Y$, since the imputations may no longer be plausible values. An alternative method is to select donors via random draw with probability of selection proportional to the potential donor’s sample weight (Rao and Shao, 1992; Rao, 1996). Assuming the response probability is constant within an adjustment cell, this method yields an asymptotically unbiased estimator for $Y$. Note that in contrast to the weighted sequential hot deck, the sample weights of nonrespondents are not used in determining the selection probabilities of donors.

If the values of $Y$ for donors and recipients within an adjustment cell have the same expected value, then the weighted draw is unnecessary, since unweighted draws will yield unbiased estimates. A similar situation arises in weighting adjustments for unit nonresponse, where a common approach is to compute nonresponse weights as the inverse of response rates computed with units weighted by their sampling weights. Little and Vartivarian (2003) argues that this can lead to inefficient and even biased estimates, and suggests instead computing nonresponse weights within adjustment cells that condition on the design weights and other covariates. The analogous approach to incorporating design weights in the hot deck is to use the design weight variable alongside auxiliary variables to define donor pools. Simulations suggest that that unweighted draws from these donor pools yield better imputations than weighted draws based on donor pools that are defined without including the design weights as a covariate (Andridge and Little, 2009). The caveat in utilizing weights in this manner is that if weights are not related to the outcome, an increase in variance may occur without a corresponding decrease in bias; simulations in Andridge and Little (2009) show only a modest increase, though more investigation is warranted.

## 2.5 Hot Decks for Multivariate Missing Data

Often more than one variable has missing values. Let $X = (X_1, \ldots, X_q)$ denote the fully observed items, including design variables, and let $Y = (Y_1, \ldots, Y_p)$ denote the items with missing values. If the components of $Y$ are missing for the same set of cases, the data have just two missing-data patterns, complete and incomplete cases; we call this the "two-pattern case". A more general case is "monotone missing data", where the variables can be arranged in a sequence $(Y_1, \ldots, Y_p)$ so that $Y_1, \ldots, Y_{j-1}$ are observed whenever $Y_j$ is observed, for $j = 2, \ldots, p$. This pattern results in longitudinal survey data where missing data arise from attrition from the sample. Alternatively, the missing values may occur in a general pattern – Judkins (1997) calls this a "swiss cheese pattern". We discuss hot deck methods for all three situations, moving from the simplest to the most complex.

### 2.5.1 The Two-Pattern Case

Suppose there are just two patterns of data, complete and incomplete cases. The same set of covariate information $X$ is available to create donor sets for all the missing items. One possibility is to develop distinct univariate hot decks for each variable, with different donor pools and donors for each item. This approach has the advantage that the donor pools can tailored for each missing item, for example by estimating a different predictive mean for each item and creating the donor pools for each incomplete variable using the predictive mean matching metric. However, a consequence of this method is that associations between the imputed variables are not preserved. For example, imputation may result in a former smoker with a current 2-pack per day habit, or an unemployed person with a substantial earned income. This may be acceptable if analyses of interest are univariate and do not involve these associations, but otherwise the approach is flawed.

An alternative method, which Marker, Judkins, and Winglee (2002) calls the *single-partition, common-donor* hot deck is to create a single donor pool for each nonrespondent, using for example the multivariate analog of the predictive mean metric (2.1):

$$d(i,j) = (\widehat{Y}(x_i) - \widehat{Y}(x_j))^T \widehat{Var}(y \cdot x_i)^{-1} (\widehat{Y}(x_i) - \widehat{Y}(x_j)), \tag{2.2}$$

where $\widehat{Var}(y \cdot x_i)$ is the estimated residual covariance matrix of $Y_i$ given $x_i$. A donor from this pool is used to impute all the missing items for a recipient, thereby preserving associations within the set. This approach clearly preserves associations between imputed variables, but since the same metric is used for all the variables, the metric is not tailored to each variable.

Another approach that preserves associations between $p$ variables, which we refer to as the *p-partition* hot deck, is to create the donor pool for $Y_j$ using adjustment cells (or more generally, a metric) that conditions on $X$ and $(Y_1, \ldots, Y_{j-1})$, for $j = 2, \ldots, p$, using the recipient's previously imputed values of $(Y_1, \ldots, Y_{j-1})$, when matching donors to recipients. Marker et al. (2002) calls this method the *n-partition* hot deck, here we replace $n$ by $p$ for consistency of notation. This approach allows the metric to be tailored for each item, and the conditioning on previously-imputed variables in the metric provides some preservation of associations, although the degree of success depends on whether the distance metrics for each variable $Y_j$ capture associations with $X$ and $(Y_1, \ldots, Y_{j-1})$, and the extent to which "close" matches can be found.

The single-partition and $p$-partition hot deck can be combined by dividing the variables $Y$ into sets, and applying a single partition and shared donors for variables within each set, but different partitions and donors across sets. Intuitively, the variables within each set should be chosen to be homogeneous with respect to potential predictors, but specifics of implementation are a topic for future research.

### 2.5.2 Monotone Patterns

Now suppose we have a monotone pattern of missing data, such that $Y_1, \ldots, Y_{j-1}$ are observed whenever $Y_j$ is observed, for $j = 2, \ldots, n$, and let $S_j$ denote the set of cases with $X, Y_1, \ldots, Y_j$ observed. More generally, we allow each $Y_j$ to represent a vector of variables with the same pattern. The $p$-partition hot deck can be applied to fill in $Y_1, ...Y_n$ sequentially, with the added feature that the set $S_j$ can be used as the pool of donors when imputing $Y_j$. The single-partition hot deck based on a metric that conditions on $X$ has the problem that it fails to preserve associations between observed and imputed components of $Y$ for each pattern. Such associations are preserved if the $p$-partition method is applied across the sets of variables $Y_j$, but variables within each each set are imputed using a single partition. Again, various elaborations of these two schemes could be envisaged.

### 2.5.3 General Patterns

For a general pattern of missing data, it is more challenging to develop a hot deck that preserves associations and conditions on the available information. The cyclic $p$-partition hot deck attempts to do this by iterative cycling through $p$-partition hot decks, in the manner of a Gibbs' sampler (Judkins, Hubbell, and England, 1993; England, Hubbell, Judkins, and Ryaboy, 1994). This approach is a semi-parametric analog of the parametric conditional imputation methods in the software packages IVEWare (Raghunathan, Lepkowski, Van Hoewyk, and Solenberger, 2001) and MICE (Van Buuren and Oudshoorn, 1999). In the first pass, a simple method is used to fill in starting values for all missing items. Second and later passes define partitions based on the best set of adjustment variables for each item to be re-imputed. Each variable is then imputed sequentially, and the procedure continues until convergence. Convergence in this setting is uncertain, and deciding when to stop is difficult; England et al. (1994) suggest stopping the algorithm when estimates stabilize rather than individual imputations, based on the philosophy that the goal of imputation is good inferences, rather than optimal imputations. The properties of this method remain largely unexplored.

Other approaches to general patterns have been proposed. The *full-information common-donor* hot deck uses a different single-partition common-donor hot deck for each distinct pattern of missingness in the target vector (Marker et al., 2002). Another method is that of Grau et al. (2004), who extend the idea of neighborhoods defined by predictive means to multivariate missingness. First, variables to be imputed are placed in a hierarchy, such that items higher in the hierarchy can be used for imputation of items lower in the list. Second, predictive means are determined for each item, using models built using complete cases only. For subjects with multiple missing values, the nearest neighbors are determined using the Mahalanobis distance based on the vector of predictive means for the missing items, and all values are copied from the selected donor to the recipient. All donors within a preset distance $\Delta$ are considered to be in the donor pool. Many multivariate methods seem relatively *ad hoc*, and more theoretical and empirical comparisons with alternative approaches would be of interest.

A slightly different approach is the joint regression imputation method of Srivastava and Carter (1986), which was extended to complex survey data by Shao and Wang (2002). Joint regression aims to preserve correlations by drawing correlated residuals. Srivastava and Carter (1986) suggest drawing residuals from fully observed respondents, and so with the appropriate regression model this becomes a hot deck procedure. Shao and Wang (2002) extend the method to allow flexible choice of distribution for the residuals and to incorporate survey weights. In the case of two items being imputed, if both items are to be imputed the residuals are drawn so they have correlation consistent with what is estimated from cases with all items observed. If only one item is imputed the residual is drawn conditional on the residual for the observed item. This differs from a marginal regression approach where all residuals are drawn independently, and produces unbiased estimates of correlation coefficients as well as marginal totals.

## 2.6 Properties of Hot Deck Estimates

We now review the (somewhat limited) literature on theoretical and empirical properties of the hot deck. The simplest hot deck procedure – using the entire sample of respondents as a single donor pool – produces consistent estimates only when data are missing completely at random (MCAR) (Rubin, 1976; Little and Rubin, 2002). The hot deck estimate of the mean equals the respondent mean in expectation, and the respondent mean is an unbiased estimate of the overall mean when data are MCAR. When data are not MCAR, two general frameworks for determining properties of estimates from imputed data have been developed: the imputation model approach (IM) and the nonresponse model approach (NM) (Shao and Steel, 1999; Haziza and Rao, 2006). Conditions for consistency of hot deck estimates depend on which of these two approaches is adopted.

The IM approach explicitly assumes a superpopulation model for the item to be imputed, termed the "imputation model"; inference is with respect to repeated sampling and this assumed data-generating model. The response mechanism is not specified except to assume that data are missing at random (MAR). In the case of the random hot deck this implies that the response probability is allowed to depend on auxiliary variables that create the donor pools but not on the value of the missing item itself. Brick, Kalton, and Kim (2004) show using this framework that the (weighted or unweighted) hot deck applied within adjustment cells leads to an unbiased estimator under a cell mean model; within each cell elements are realizations of independently and identically distributed random variables. For nearest neighbor imputation, Rancourt, Särndal, and Lee (1994) claim that estimates of sample means are asymptotically unbiased assuming a linear relationship between the item to be imputed and the auxiliary information, but no theoretical support is offered. Chen and Shao (2000) extend the approach of Rancourt et al. (1994) to show that the relationship between the imputed variable and the auxiliary information need not be linear for asymptotic unbiasedness to hold, with suitable regularity conditions.

Perhaps the most crucial requirement for the hot deck to yield consistent estimates is the existence of at least some donors for a nonrespondent at every value of the set of covariates that are related to missingness. To see why, consider the extreme case where missingness of $Y$ depends on a continuous covariate $x$, such that $Y$ is observed when $x < x_0$ and $Y$ is missing when $x \geq x_0$. A hot deck method that matches donors to recipients using $x$ clearly cannot be consistent when $Y$ has a non-null linear regression on $x$, since donors close to recipients are not available, even asymptotically as the sample size increases. In contrast, parametric regression imputation would work in this setting, but depends strongly on the assumption that the parametric form of the mean function is correctly specified.

In lieu of making an explicit assumption about the distribution of item values, the NM approach makes explicit assumptions about the response mechanism. Also called the quasirandomization approach (Oh and Scheuren, 1983), the NM approach assumes that the response probability is constant within an imputation cell. Inference is with respect to repeated sampling and the assumed uniform response mechanism within cells. Thus for the random hot deck to lead to unbiased estimates, the within-adjustment-cell response probability must be constant. If sample selection is with equal probability, selection of donors may be by simple random sampling to achieve unbiasedness. For unequal probabilities of selection, selection of donors with probability of selection proportional to the potential donor's sample weight leads to asymptotically unbiased and consistent mean estimates (Rao and Shao, 1992; Rao, 1996; Chen and Shao, 1999). Applications of both of these approaches to variance estimation can be found in Section 2.7.

Suppose now that the interest is in estimating either *domain* means, where a domain is a collection of adjustment cells, or *cross-class* means, defined as a subset of the population that cuts across adjustment cells (Little, 1986). The hot deck produces consistent estimates of domain and cross-class means if stratification on $x$ produces cells in which $Y$ is independent of response. Since one cannot observe the distribution of $Y$ for the nonrespondents, using all auxiliary variables to define the cells would be the best strategy. Often the dimension of $x$ is too large for full stratification, and alternative distance metrics such as the predictive mean, $\hat{Y}(x)$, or the response propensity, $\hat{p}(x)$, can be useful. Using these metrics to define adjustment cells was discussed by Little (1986). For domain means, predictive mean stratification and response propensity stratification both yield consistent estimates. For estimating cross-class means, predictive mean stratification produces estimates with zero large-sample bias, but response propensity stratification gives nonzero bias. In this case adjustment cells must be formed based on the joint distribution of response propensity and the cross-class variable in order to produce consistent estimates.

An alternative approach to the hot deck is to generate imputations as draws from the distribution of the missing values based on a parametric model. Examples of this approach include the popular regression imputation, Bayesian MI methods in SAS PROC MI (SAS Institute, Cary, NC) or the sequential MI algorithms implemented in IVEware and MICE (Raghunathan et al., 2001; Van Buuren and Oudshoorn, 1999). Little (1988) points out that the adjustment cell method is in effect the same as imputing based on a regression model that includes all high-order interactions between the covariates, and then adding an empirical residual to the predictions; imputation based on a more parsimonious regression model potentially allows more main effects and low-order interactions to be included (Lillard, Smith, and Welch, 1982; Little, 1988).

Several studies have compared parametric methods to the non-parametric hot deck David, Little, Samuhel, and Triest (1986) compared the hot deck used by the U.S. Census Bureau to impute income in the CPS to imputation using parametric models for income (both on the log scale and as a ratio) and found that the methods performed similarly. Several authors have compared hot deck imputation using predictive mean matching to parametric methods that impute predicted means plus random residuals (Lazzeroni, Schenker, and Taylor, 1990; Heitjan and Little, 1991; Schenker and Taylor, 1996). The relative performance of the methods depends on the validity of the parametric model and the sample size. When the population model matches the parametric imputation model, hot deck methods generally have larger bias and are less precise. However, the hot deck is less vlunerable to model misspecification. If a model is used to define matches, as in hot deck with predictive mean matching, it is less sensitive to misspecification than models used to impute values directly. The hot deck tends to break down when the sample size is small, since when the pool of potential donors is limited, good matches for nonrespondents are hard to find. Also, in small samples the bias from misspecification of parametric models is a smaller component of the mean squared error. Thus, parametric imputation methods become increasingly attractive as the sample size diminishes.

## 2.7 Variance Estimation

Data sets imputed using a hot deck method are often analyzed as if they had no missing values (Marker et al., 2002). In particular, variance estimates in the Current Population Survey continue to be based on replication methods appropriate for completely observed data (U.S. Bureau of the Census, 2002). Such approaches clearly understate uncertainty, as they ignore the added variability due to nonresponse. There are three main approaches to obtaining valid variance estimates from data imputed by a hot deck: (1) Explicit variance formulae that incorporate nonresponse; (2) Resampling methods such as the jackknife and the bootstrap, tailored to account for the imputed data; and (3) hot deck multiple imputation (HDMI), where multiple sets of imputations are created, and imputation uncertainty is propagated via MI combining rules (Rubin, 1987; Little, 1988). We now review these three approaches.

### 2.7.1 Explicit Variance Formulae

Explicit variance formulae for hot deck estimates can be derived in simple cases; see for example Ford (1983) for simple random sampling from the donor pool and Bailar and Bailar (1978) for the sequential hot deck. These methods make the strong and often unrealistic assumption that the data are missing completely at random (MCAR). Creating adjustment cells, applying one method separately within cells, and pooling the results eliminates bias attributable to differences in response across the cells. Alternatively, if one is willing to make some assumptions about the distribution of $Y$ in the population, several methods have been developed that lead to explicit variance formulae.

The model-assisted estimation approach of Särndal (1992) allows variance estimation under the more realistic assumption that data are missing at random (MAR). By assuming a model for the distribution of $Y$ in the population, the variance of an estimator in the presence of missingness is decomposed into a sampling variance and an imputation variance. Estimators are obtained using information in the sampling design, observed naïve values, and imputation scheme. Brick et al. (2004) extend Särndal's method to the hot deck, using the assumption that within adjustment cells ($g = 1, \ldots, G$) values of $Y$ are independent and identically distributed with mean $\mu_g$ and variance $\sigma_g^2$. They derive a variance estimator that is conditionally unbiased, given the sampling, response, and imputation indicators, and argue that conditioning on the actual number of times responding units are used as donors is more relevant than the unconditional variance which averages over all possible imputation outcomes. Cell means and variances are the only unknown quantities that need estimation to use the variance formula. The authors note that their method covers many forms of hot deck imputation, including both weighted and unweighted imputation and selection with and without replacement from the donor pool.

Chen and Shao (2000) consider variance estimation for nearest neighbor hot deck imputation and derive the asymptotic variance of the mean in the case of a single continuous outcome ($Y$) subject to missingness and a single continuous auxiliary variable ($x$). Their formula requires specification of the conditional expectation of $Y$ given $x$. In practice, one has to assume a model for the mean, such as $E(Y|x) = \alpha + \beta x$, fit the model to the observed data, and use the estimates $\hat{\alpha}$ and $\hat{\beta}$ in their variance formulas. This method produces a consistent estimate of the variance, assuming the model is correct. Of note, they show that the empirical distribution function obtained from nearest neighbor imputation is asymptotically unbiased, and so quantile estimators are also unbiased.

### 2.7.2 Resampling Methods for Single Imputation

Model-assisted methods for variance estimation are vulnerable to violations of model assumptions. A popular alternative is resampling methods. One such method is the jackknife, where estimates are based on dropping a single observation at a time from the data set. Performing a naïve jackknife estimation procedure to the imputed data underestimates the variance of the mean estimate, particularly if the proportion of nonrespondents is high. To correct this, Burns (1990) proposed imputing the full sample and then imputing again for each delete-one data set. However, this leads to overestimation when $n$ is large and requires repeating the imputation procedure

$n + 1$ times. To combat this, Rao and Shao (1992) proposed an adjusted jackknife procedure that produces a consistent variance estimate.

Rao and Shao's jackknife method can be applied to random hot deck imputation of complex stratified multistage surveys; the more straightforward application to inference about means from a simple random sample with replacement is discussed here. Suppose that $r$ units respond out of a sample of size $n$, and the simple unweighted hot deck is applied, yielding the usual estimate $\bar{y}_{HD}$. First, the hot deck procedure is applied to create a complete data set. The estimator for each jackknife sample is calculated each time a nonrespondent value is deleted, but with a slight adjustment when respondents are deleted. Specifically, each time a respondent value is dropped the imputed nonrespondent values are each adjusted by $E(\tilde{y}_i^{(-j)}) - E(\tilde{y}_i)$, where $\tilde{y}_i$ is the imputed value for nonrespondent $i$ using the entire donor pool and $\tilde{y}_i^{(-j)}$ is the hypothetical imputed value with the $j^{th}$ respondent dropped, and expectation is with respect to the random imputation. For the random hot deck this reduces to an adjustment of $\bar{y}_R^{(-j)} - \bar{y}_R$, where $\bar{y}_R^{(-j)}$ is the mean of the remaining $(r-1)$ respondents after deleting the $j^{th}$ respondent. This adjustment introduces additional variation among the pseudoreplicates to capture the uncertainty in the imputed values that would otherwise be ignored by the naive jackknife. The adjusted jackknife variance estimate is approximately unbiased for the variance of $\bar{y}_{HD}$, assuming a uniform response mechanism and assuming the finite population correction can be ignored.

Extensions of this method to stratified multistage surveys and weighted hot deck imputation involve a similar adjustment to the jackknife estimators formed by deleting clusters; see Rao and Shao (1992) for details. Kim and Fuller (2004) describe application of the jackknife variance estimator to fractional hot deck imputation, first described by Fay (1993). A similar jackknife procedure for imputation in a without-replacement sampling scheme and for situations where sampling fractions may be non-negligable is discussed in Berger and Rao (2006). Chen and Shao (2001) show that for nearest neighbor hot deck imputation the adjusted jackknife produces overestimates of the variance since the adjustment term will be zero or near zero, similar to the difficulty in applying the jackknife to the sample median. They suggest alternative "partially adjusted" and "partially reimputed" methods that are asymptotically unbiased. Other popular resampling techniques for variance estimation include the balanced half sample method and the random repeated replication method. These methods require adjustments similar to those for the jackknife in the presence of imputed data; details are given in Shao, Chen, and Chen (1998) and Shao and Chen (1999).

Though the adjusted jackknife and its variants require only a singly-imputed data set, they are not without limitation. There must be accompanying information that indicates which values were initially nonrespondents, a feature that is not often found with public-use data sets imputed via the hot deck (or any other procedure). Additionally, the step of adjusting imputed values for each jackknife replicate requires the user to know the precise details of the hot deck method used for the imputation, including how the adjustment cells were formed and how donors were selected. In practice this means that either the end user carries out the imputation himself, or that the end user can be trusted to correctly recreate the original imputation.

The jackknife cannot be applied to estimate the variance of a non-smooth statistic, e.g. a sample quantile. A resampling method that allows for estimation of smooth or non-smooth statistics is the bootstrap (Efron, 1994), and its application to the hot deck was discussed by Shao and Sitter (1996) and Saigo, Shao, and Sitter (2001). As with the jackknife, applying a naive bootstrap procedure to a singly-imputed data set leads to underestimation. However, a simple alteration leads to a bootstrap procedure that yields consistent variance estimates. First, the hot deck is used to generate a complete data set. From this a bootstrap sample of size $n$ is drawn with replacement from the imputed sample. Instead of calculating a bootstrap estimate of $\bar{y}$ at this point, the hot deck must be reapplied and the sampled respondent values used as the donor pool for the sampled nonrespondents. Then the usual estimate $\bar{y}^{(b)}$ can be calculated for this $b^{th}$ bootstrap sample. Bootstrap samples are drawn and the imputation repeated $B$ times, and the usual bootstrap mean and variance formulae can be applied. The extra step of imputing at each bootstrap sample propagates the uncertainty, and thus yields a consistent estimate of variance. In addition, bootstrap estimates can be developed for multistage survey designs, for example by bootstrapping primary sampling units rather than individual units. As with the adjusted jackknife, the bootstrap requires knowledge of which values were imputed, which may not be available in public-use data sets. Chen and Shao (1999) consider variance estimation for singly-imputed data sets when the nonrespondents are nonidentifiable and derive design-consistent variance estimators for sample means and quantiles. The method only requires a consistent estimator of the response probability, which may be available when more detailed subject-specific response information is not, and produces an adjustment to the usual complete data variance formula (e.g. Cochran, 1977) to account for the uncertainty in imputation.

### 2.7.3 Multiple Imputation

First proposed by Rubin (1978), MI involves performing $K \geq 2$ independent imputations to create $K$ complete data sets. As before, assume that the mean of the variable $y$ subject to nonresponse is of interest. Let $\hat{\theta}_k$, $W_k$ denote the estimated mean and variance of $\bar{y}$ from the $k^{th}$ complete data set. Then the MI estimator of $\bar{y}$

is simply the average of the estimators obtained from each of the $K$ completed data sets:

$$\bar{\theta}_K = \frac{1}{K}\sum_{k=1}^{K}\hat{\theta}_k \tag{2.3}$$

The averaging over the imputed data sets improves the precision of the estimate, since the added randomness from drawing imputations from an empirical distribution (rather than imputing a conditional mean) is reduced by a factor of $1/K$. The variance of $\bar{\theta}_K$ is the sum of the average within-imputation variance and the between-imputation variance. Ignoring the finite population correction, the average within-imputation variance is

$$\bar{W_K} = \frac{1}{K}\sum_{k=1}^{K} W_k$$

and the between-imputation variance is

$$B_K = \frac{1}{K-1}\sum_{k=1}^{K}\left(\hat{\theta}_k - \bar{\theta}_K\right)^2.$$

The total variance of $\bar{\theta}_K$ is the sum of these expressions, with a bias correction for the finite number of multiply imputed data sets,

$$\text{Var}\left(\bar{\theta}_K\right) = \bar{W_K} + \frac{K+1}{K}B_K. \tag{2.4}$$

When the hot deck procedure is used to create the MI data sets, and the same donor pool is used for a respondent for all $K$ data sets, the method is not a proper MI procedure (Rubin, 1978). The method produces consistent estimates of $\bar{y}$ as $K \to \infty$ but since the predictive distribution does not properly propagate the uncertainty, its variance is an underestimate, even with an infinite number of imputed data sets. The degree of underestimation becomes important if a lot of information is being imputed.

Adjustments to the basic hot deck procedure that make it "proper" for MI have been suggested, though not widely implemented by practitioners. One such procedure is the Bayesian Bootstrap (BB) (Rubin, 1981). Suppose there are $M$ unique values, $d = (d_1, d_2, \ldots, d_M)$ of $Y$ observed among the respondents, with associated probabilities $\phi = (\phi_1, \phi_2, \ldots, \phi_M)$. Imposing an noninformative Dirichlet prior on $\phi$ yields a Dirichlet posterior distribution with mean vector $\hat{\phi} = \left(\hat{\phi_1}, \ldots, \hat{\phi_M}\right)$ with $\hat{\phi_m} = r_m/r$, where $r_m$ denotes the number of times that $d_m$ is observed among the respondents. Imputation proceeds by first drawing $\phi^*$ from the posterior distribution and then imputing values for each nonrespondent by drawing from $d$ with vector of probabilities $\phi^*$. Repeating the entire procedure $K$ times gives proper multiple imputations.

The Approximate Bayesian Bootstrap (ABB) approximates the draws of $\phi$ from the above Dirichlet posterior distribution with draws from a scaled multinomial distribution (Rubin and Schenker, 1986). First an $r$ dimensional vector $X$ is drawn with replacement from the respondents' values. Then the $n - r$ nonrespondent values are drawn with replacement from $X$. This method is easy to compute, and repeated applications will yield again yield proper multiple imputations. Variances for the ABB method are on average higher than variances for the BB method by a factor of $(r + 1)/r$, but confidence coverage for the two methods were very close and always superior to the simple hot deck in simulations in (Rubin and Schenker, 1986). Kim (2002) notes that the bias of the ABB method is not negligible when sample sizes are small and response rates are low. He suggests a modification in which the size of the vector $X$ drawn from the respondents is not $r$, but instead is a value $d$ chosen to minimize the bias in the variance for small samples. The value of $d$ depends on the total sample size and the response rate and as $n \rightarrow \infty$, $d \rightarrow r$, so that in large samples the correction is not needed. See Kim (2002) for details.

One of the biggest advantages to parametric multiple imputation is that it allows users to easily estimate variances for sample quantities besides totals and means. To achieve this with the hot deck requires modifying the imputation procedure to be "proper," via BB or ABB. However, implementation of these methods in sample settings more complex than simple random sampling (i.e. multi-stage sampling) remains largely unexplored. On the other hand, software for parametric Bayesian multiple imputation (e.g. IVEWARE) is available and can handle clustering, weighting, and other features of complex survey designs. Practitioners and agencies already using the hot deck may be unwilling to alter their imputation strategy to obtain correct variance estimates from multiple imputation, choosing instead to utilize a resampling technique. Those looking for a new multiple imputation strategy may prefer the ease of the parametric methods.

### 2.8 Detailed Example

With so many variations of the hot deck in use, we set out to compare a subset of these methods using a real data set. The third National Health and Nutrition Examination Survey (NHANES III) is a large-scale survey that has previously been used to compare imputation methods, including parametric and non-parametric and single and multiple imputation methods (Ezzati-Rice, Fahimi, Judkins, and Khare, 1993a; Ezzati-Rice, Khare, Rubin, Little, and Schafer, 1993b; Khare, Little, Rubin, and Schafer, 1993). NHANES III data were also released to the public as a multiply imputed data set (U.S. Department of Health and Human Services, 2001). Details of the survey design and data collection procedures are available in *Plan and Operation of the Third National Health and Nutrition Examination Survey* (U.S. Department of Health and Human Services, 1994).

### 2.8.1 Description of the Data

NHANES III collected data in three phases: (a) a household screening interview, (b) a personal home interview, and (c) a physical examination at a mobile examination center (MEC). The total number of persons screened was 39,695, with 86% (33,994) completing the second phase interview. Of these, only 78% were examined in the MEC. Previous imputation efforts for NHANES III focused on those individuals who had completed the second phase; weighting adjustments were used to compensate for non-response at this second stage. Since the questions asked at both the second and third stage varied considerably by age we chose to select only adults age 17 and older who had completed the second phase interview for the purposes of our example, leaving a sample of 20,050. Variables that were fully observed for the sample included age, gender, race, and household size. We focused on the imputation of diastolic blood pressure measurements (DBP) and selected additional variables from the second and third stages that we hypothesized might be related to this outcome: self-rating of health status (a five-level ordinal variable), an indicator for high blood pressure, height, and weight. In order to have a "truth" against which to measure each imputation method we selected the cases with fully observed data on all ten selected variables as our population (n=16,739, 83.5% of total sample).

### 2.8.2 Sampling and Nonresponse Mechanisms

A sample of size 800 was drawn by simple random sampling from the population for an approximate 5% sampling fraction. We utilized two separate propensity models to induce missingness in the sample. We initially fit a logistic regression model on an indicator for missingness on the DBP variable using the entire sample (n=20,050). This created predicted probabilities of non-response that mimicked the actual propensities observed in the NHANES data and ranged from 0.05 to 0.29. Variables included in the regression model were necessarily those observed for all subjects, and so were limited to age, race, sex, and household size (plus all interactions). This propensity model led to an expected 14.9% percent missing (Model 1). Our second propensity model was intended to be stronger, have a higher percent missing, and induce bias such that a complete case analysis would lead to overestimation of average DBP. The following model was used to obtain the probability of non-response for subject $i$:

$$\text{logit}(P(M_i = 1)) = -3+1.5*I(\text{age}_i < 40)+0.75*\text{female}_i+0.25*\text{Mexican-American}_i$$

where $\text{female}_i$ and $\text{Mexican-American}_i$ equal one if subject $i$ is female and Mexican-American, respectively. The individual probabilities of non-response ranged from 0.10 to 0.75, with an expected percent missing of 33.1% (Model 2).

For each of the two sets of predicted non-response probabilities, nonresponse indicators for each unit in the sample were independently drawn from a Bernoulli distribution with probabilities according to each of the two propensity models. Nonrespondent values were then deleted from the drawn sample to create the respondent data set. This process of sampling and creating nonrespondents was repeated 1,000 times.

### 2.8.3 Imputation Methods

The following imputation methods were applied to the incomplete sample: adjustment cell random hot deck, predictive mean random hot deck, propensity cell random hot deck, and parametric regression imputation. All three hot deck methods create donor pools based on a distance metric and use equal probability draws from the donors in each pool to impute for the nonrespondents. The adjustment cell hot deck used age, gender, and race to create imputation classes. In comparison, the predictive mean and the response propensity hot decks allowed incorporation of many more variables; Table 2.2 lists the imputation methods and auxiliary variables used in each method. Since a total of 18 cells were created in the cross-classification of variables in the adjustment cell method we chose to utilize a similar number of cells in the other two methods, 20 equally sized cells for both predictive mean and propensity stratification. Attempts to include more variables when defining adjustment cell strata lead to cells that were too sparse; instead of trying to collapse cells ad-hoc we opted to use the coarser cells. We required a minimum of five respondents in each imputation cell to proceed with hot deck imputation; this minimum was met in all runs. The parametric regression imputation method assumed normality for the outcome and used the same model as that which created the predictive mean strata.

For each method we applied both single and multiple imputation. Single imputation resulted in one estimator of the mean for each of the four imputation methods. A total of three methods for estimating variance for SI after random hot deck imputation were used: a naïve estimator treating the imputed values as if they were observed (SI Naïve), an exact formula (SI Formula), and the jackknife of Rao and Shao (1992) (SI RS Jackknife). For the parametric method there were two variance estimators: a naïve estimator and a bootstrap estimator (SI Bootstrap).

We applied three versions of MI to the incomplete data leading to three separate mean estimators for each imputation method: improper MI with $K = 5$ data sets (IMI 5), proper MI with $K = 5$. The simulation was carried out using the software R with the MICE package for parametric imputation (R Development Core Team, 2007; Van Buuren and Oudshoorn, 1999).

Empirical bias and root mean square error (RMSE) for each imputation method $M$ were calculated as follows,

$$\text{EBias} = \frac{1}{1000}\sum_{i=1}^{1000}(\hat{\theta}_{Mi} - \theta)$$

$$\text{RMSE} = \sqrt{\frac{1}{1000}\sum_{i=1}^{1000}(\hat{\theta}_{Mi} - \theta)^2}$$

where $\hat{\theta}_{Mi}$ is the estimate of the population mean using method $M$ for the $i$th replicate and $\theta$ is the true population parameter. Variance estimators for each method were evaluated using the empirical variance (defined as the variance of the point estimator observed in the Monte Carlo sample) and the average variance estimate. In addition to evaluating accuracy in point and variance estimators we were interested in coverage properties of the imputation methods, so the actual coverage of a nominal 95% confidence interval and average CI length were also calculated.

### 2.8.4 Results

Differences among the imputation methods were magnified under the stronger propensity mechanism (model 2); results from model 1 were similar and are not shown. Table 2.3 displays results from the simulation using propensity model 2. All methods performed well in terms of bias, with only the complete case estimate under propensity model 2 demonstrating (relatively) large bias and thus severe undercoverage (47%). The propensity strata did exhibit higher bias than either the adjustment cell or predictive mean methods but the magnitude was very small. Naïve variance estimators always underestimated the empirical variance, leading to empirical coverages for a nominal 95% interval ranging from 81-90%. Improper MI for all hot deck methods underestimated the variance, leading to coverage of only 90-91%. All other methods had near the 95% nominal coverage. Across all hot deck methods MI had lower empirical variance than SI methods, leading to shorter confidence intervals but still adequate coverage. MI with $K = 20$ showed slight gains in efficiency over $K = 5$. This simulation failed to demonstrate any major advantage of parametric imputation over the hot deck methods. Performance was very similar, with the parametric imputation having slightly lower RMSE. MI with parametric imputation had shorter confidence intervals than with either adjustment cell or predictive mean strata, however CI length was virtually identical to that of the predictive mean strata.

Figure 2.1 plots the ratio of average to empirical variance against the empirical variance for the adjustment cell (•) and predictive mean cell (▲) methods to give insight into their efficiency. Figure 2.2 similarly plots CI coverage against CI length. Again, results were similar across both propensity models; we only show the stronger mechanism. Predictive mean MI had smaller empirical variance for both propensity models with only slight underestimation, but coverage was not affected and remained at nominal levels. The jackknife following SI for both methods accurately estimated its empirical variance but was not as efficient; confidence coverage was at nominal levels but with large CI length. Overall the predictive mean method appeared to have a slight advantage over the adjustment cell method as evidenced by a gain in efficiency seen in both single and multiple imputation strategies.

This simulation used a variety of random hot deck methods to impute data in a real data set. All hot deck methods performed well and without bias, however the relationship between outcome and predictor variables was not particularly strong in this data set. Applying the predictive mean model to the complete population yielded an $R^2$ of 0.20, and this weak association may partially explain why the adjustment cell method that only used three auxiliary variables had similar results to the more flexible methods of creating the donor pools. This simulation also demonstrated the potentially severe effects of treating singly imputed data as if it were observed data, a practice that while unfortunately common in practice cannot be recommended.

## 2.9 Conclusion

The hot deck is widely used by practitioners to handle item nonresponse. Its strengths are that it imputes real (and hence realistic) values, it avoids strong parametric assumptions, it can incorporate covariate information, and it can provide good inferences for linear and nonlinear statistics if appropriate attention is paid to propagating imputation uncertainty. A weakness is that it requires good matches of donors to recipients that reflect available covariate information; finding good matches is more likely in large than in small samples. Simple hot decks based on adjustment cells have limited ability to incorporate extensive covariate information; these limitations may be ameliorated by the metric-based approaches in Section 2.3, but these methods are more complex and theory on them is largely lacking.

Our review highlights several issues with the hot deck that we feel deserve consideration. The first issue is the use of covariate information. Adjustment cell methods, while popular in their simplicity, limit the amount of auxiliary information that can be effectively used. Alternative distance metrics are more flexible and should be considered, in particular we feel the predictive mean metric shows promise. When choosing the variables for creating donor pools, the priority should be to select variables that are predictive of the item being imputed, $Y$. For example, forward selection for the regression of $Y$ on $X_1, \ldots, X_k$ might be used to choose covariates that significantly predict $Y$ and could be the basis for a predictive mean metric for defining donor pools. The response propensity is important for reducing bias, but only if it is associated with $Y$; traditional covariate selection methods could be used to determine if auxiliary information that is predictive of nonresponse is also associated with $Y$. With multiple $Y$'s, the choice of a single metric (i.e. single partition hot deck) requires compromising matches, whereas partitions for each $Y$ allow tailoring of metrics to each specific item. However, in order to preserve associations among the imputed values, each step should condition on previously imputed $Y$'s.

A second issue surrounding the hot deck is how to deal with "swiss cheese" missing data patterns. While some methods have been suggested (e.g. the cyclic p-partition hot deck), we were unable to find much theory to support these methods. More development of their theoretical properties and simulation studies of performance are needed.

The third and final issue that must be taken into consideration is how to obtain valid inference after imputation via the hot deck. As with any imputation method, it is important to propagate error, and with the hot deck this step is often overlooked. In practice, we think that the single most important improvement would be to compute standard errors that incorporate the added variance from the missing information when the fraction of missing information is substantial, by one of the sample reuse methods or MI, as discussed in Section 2.7. There has been considerable debate among methodologists about the relative merits of these two approaches, particularly under misspecified models (Meng, 1994; Fay, 1996; Rao, 1996; Rubin, 1996; Robins and Wang, 2000; Kim, Brick, Fuller, and Kalton, 2006), and more simulation comparisons of the repeated-sampling properties of these approaches would be of interest. However, either approach is superior to assuming the added variance from imputation is zero, which is implied by treating a single imputed data set as if the imputed values are real.

Despite the practical importance of the hot deck as a method for dealing with item nonresponse, the statistics literature on theory of the method and comparisons with alternative approaches is surprisingly limited, yielding opportunities for further methodological work. Other areas where more development seems possible include better ways to condition on available information in creating donor pools, ways to assess the trade-off between the size of donor pool and quality of matches, and methods for multivariate missing data with a general pattern of missingness. On the theoretical side, consistency of the hot deck has been shown under MCAR, or missing completely at random within adjustment cells, but useful conditions for consistency under MAR when conditioning on the full set of available information seem lacking. Also hot deck methods for situations where nonresponse is “nonignorable” (that is, the data are not missing at random) have not been well explored. Hopefully this review will stir some additional methodological activity in these areas.

Table 2.1: Effect of weighting adjustments on bias and variance of a mean, by strength of association of the adjustment cell variables with nonresponse and outcome (Little and Vartivarian, 2005)

| | | Association with outcome | |
|---|---|---|---|
| | | Low | High |
| Association with nonresponse | Low | Bias: –<br>Var: – | Bias: –<br>Var: ↓ |
| | High | Bias: –<br>Var: ↑ | Bias: ↓<br>Var: ↓ |

Table 2.2: Imputation methods applied to samples drawn from the NHANES III data

| Method | Imputation Cell Variables | Number of Cells |
|---|---|---|
| 1. Adjustment cells | age (categorical), gender, race | 18 |
| 2. Predictive Mean cells | age (continuous), gender, race, household size, health status, ever high BP, body mass index | 20, equally sized |
| 3. Propensity cells | age (continuous), gender, race, household size | 20, equally sized |
| 4. Parametric Model | age (continuous), gender, race, household size, health status, ever high BP, body mass index | n/a |

Table 2.3: Results from 1,000 replicates (n=800) using Propensity Model 2 (33% missing)

| Method | Variance Estimator | Empirical Bias | RMSE | Empirical Variance | Average Variance | 95% Coverage | CI Length |
|---|---|---|---|---|---|---|---|
| Before Deletion | | 0.003 | 0.468 | 0.219 | 0.211 | 95.6 | 1.80 |
| Complete case | | 1.131 | 1.262 | 0.312 | 0.307 | 47.0 | 2.17 |
| Adjustment cells | SI Naïve | -0.016 | 0.668 | 0.446 | 0.211 | 83.1 | 1.80 |
| | SI Formula | -0.016 | 0.668 | 0.446 | 0.445 | 94.5 | 2.60 |
| | SI RS Jackknife | -0.016 | 0.668 | 0.446 | 0.474 | 95.8 | 2.68 |
| | IMI 5 | -0.004 | 0.628 | 0.395 | 0.290 | 91.5 | 2.15 |
| | PMI 5 | -0.007 | 0.646 | 0.418 | 0.409 | 94.7 | 2.65 |
| | PMI 20 | -0.007 | 0.626 | 0.392 | 0.386 | 94.5 | 2.45 |
| Predictive Mean cells | SI Naïve | -0.036 | 0.672 | 0.450 | 0.213 | 82.5 | 1.81 |
| | SI Formula | -0.036 | 0.672 | 0.450 | 0.376 | 91.9 | 2.39 |
| | SI RS Jackknife | -0.036 | 0.672 | 0.450 | 0.424 | 94.2 | 2.54 |
| | IMI 5 | -0.026 | 0.627 | 0.392 | 0.288 | 90.5 | 2.14 |
| | PMI 5 | -0.031 | 0.638 | 0.406 | 0.355 | 94.3 | 2.43 |
| | PMI 20 | -0.026 | 0.620 | 0.384 | 0.337 | 94.1 | 2.28 |
| Propensity cells | SI Naïve | -0.060 | 0.665 | 0.440 | 0.216 | 82.9 | 1.81 |
| | SI Formula | -0.060 | 0.665 | 0.440 | 0.449 | 95.7 | 2.61 |
| | SI RS Jackknife | -0.060 | 0.665 | 0.440 | 0.482 | 96.6 | 2.70 |
| | IMI 5 | -0.048 | 0.637 | 0.404 | 0.295 | 91.7 | 2.17 |
| | PMI 5 | -0.049 | 0.646 | 0.415 | 0.399 | 95.0 | 2.61 |
| | PMI 20 | -0.053 | 0.625 | 0.388 | 0.377 | 94.8 | 2.42 |
| Parametric Model | SI Naïve | -0.033 | 0.692 | 0.479 | 0.208 | 81.2 | 1.78 |
| | SI Bootstrap | -0.033 | 0.692 | 0.479 | 0.490 | 94.5 | 2.73 |
| | PMI 5 | -0.031 | 0.626 | 0.391 | 0.354 | 93.8 | 2.44 |
| | PMI 20 | -0.030 | 0.610 | 0.371 | 0.333 | 93.6 | 2.27 |

Figure 2.1: Empirical variance and ratio of average to empirical variance from Propensity Model 2 (∼33% missing), for hot deck imputation within adjustment cells (●) and predictive mean cells (▲). Results from 1,000 replicates (n=800).

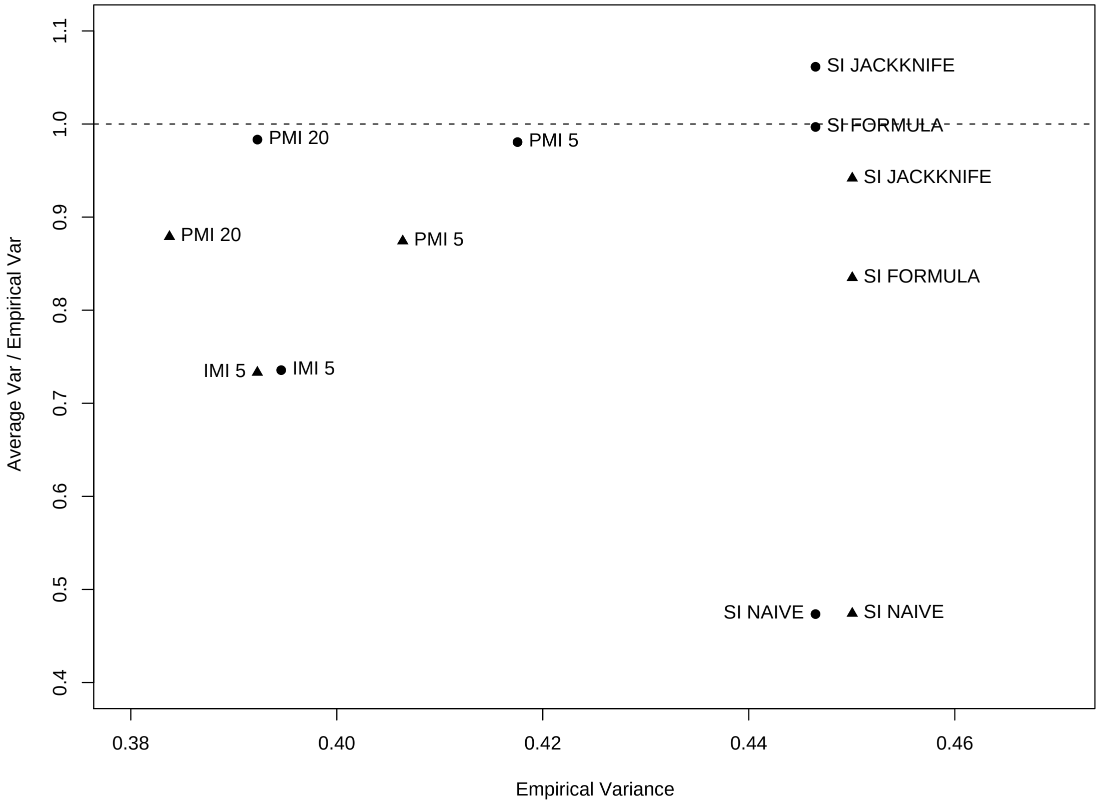
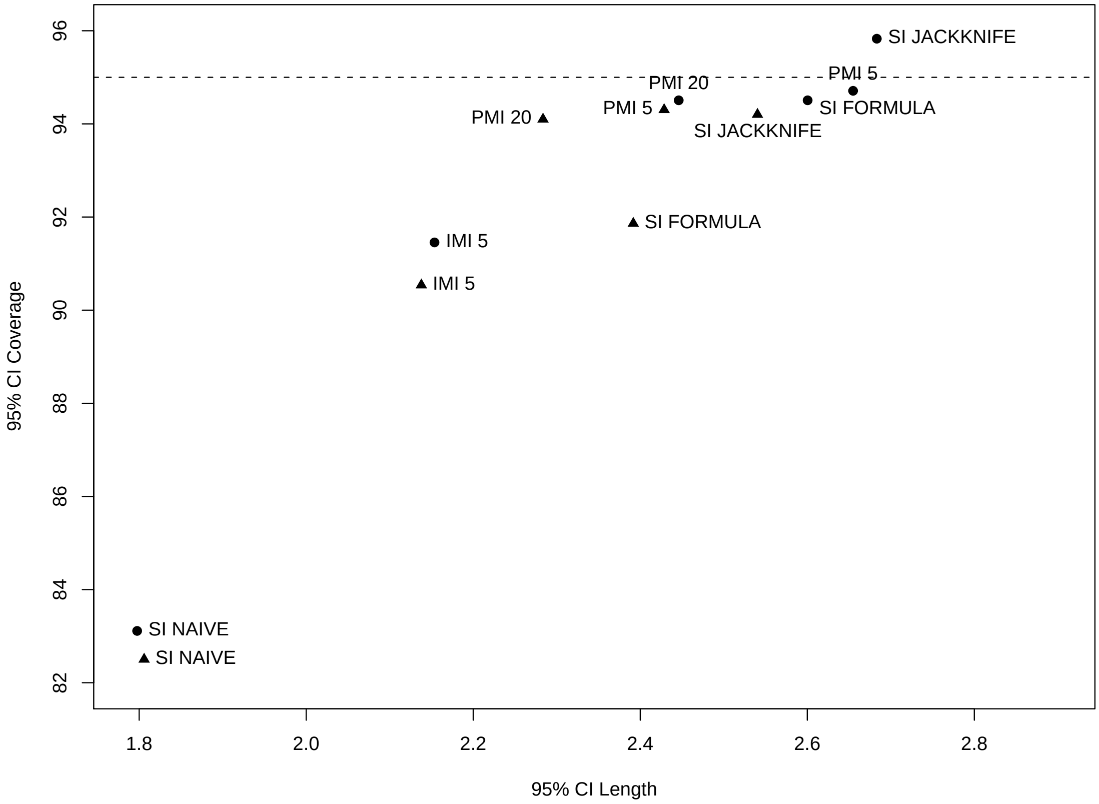


Figure 2.2: Confidence interval length and coverage from Propensity Model 2 (∼33% missing), for hot deck imputation within adjustment cells (●) and predictive mean cells (▲). Results from 1,000 replicates (n=800).



# CHAPTER III

# The Use of Sample Weights in Hot Deck Imputation

## 3.1 Introduction

Missing data are often a problem in large-scale surveys, arising when a sampled unit does not respond to the entire survey (unit nonresponse) or to a particular question (item nonresponse). We consider here imputation for item nonresponse, a common technique for creating a complete data set that can then be analyzed with traditional analysis methods. In particular we consider use of the hot deck, an imputation strategy in which each missing value is replaced with an observed response from a "similar" unit (Kalton and Kasprzyk, 1986). The hot deck method does not rely on model fitting for the variable to be imputed, and thus is potentially less sensitive to model misspecification than an imputation method based on a parametric model, such as regression imputation. It preserves the distribution of item values, unlike mean imputation which leads to a spike of values at the respondent mean. Additionally, only plausible values can be imputed, since values come from observed responses in the donor pool.

The most common method of matching donor to recipient is to divide responding and nonresponding units into imputation classes, also known as adjustment cells or donor pools, based on variables observed for all units (Brick and Kalton, 1996). To create cells, any continuous variables are categorized before proceeding. Imputation is then carried out by randomly picking a donor for each nonrespondent within each cell. These classes historically have been formed a priori based on knowledge of the subject matter and choosing variables that are associated with the missing values. In addition, variables that are predictive of nonresponse may be used to define imputation classes.

Once imputation has created a filled-in data set, analysis can proceed using the sampling weights determined by the sample design. Unlike weighting for nonresponse, where sample weights must be combined with nonresponse weights for subsequent analysis, no adjustment to the weights is necessary. However, ignoring sample weights effectively imputes using the unweighted sample distribution of respondents in an adjustment cell, which may cause bias if these respondents have differing sampling weights. In this paper we consider several ways for using the survey weights in creating donor pools and carrying out hot deck imputation. Section 3.2 reviews methods developed for incorporating sample weights into the hot deck. In Section 3.3 a simulation study compares estimators of a population mean using these methods. Section 3.4 demonstrates these methods on data from the third National Health and Nutrition Examination Survey (NHANES III).

## 3.2 Methods for Incorporating Sample Weights

Two approaches to selection from hot deck donor pools have been used: sequential and random. Sequential selection first sorts all units within a donor pool and then imputes for each missing value the closest preceding respondent value, a variant of nearest neighbor imputation. The sort order can be random, or sorting variables can be auxiliary variables presumed related to the item being imputed. In contrast, random selection imputes each missing value with a random draw from the donor pool for each nonrespondent. Neither of these methods necessarily incorporate survey design weights into donor selection.

A modification to the sequential procedure to incorporate sample weights was proposed by Cox (1980) and called the weighted sequential hot deck (WSHD). The procedure preserves the sorting methodology of the unweighted procedure, but allows all respondents the chance to be a donor and uses sampling weights to restrict the number of times a respondent value can be used for imputation. Respondents and nonrespondents are first separated into two files and sorted (randomly, or by auxiliary variables). Sample weights of the nonrespondents are rescaled to sum to the total of the respondent weights. The algorithm can be thought of as aligning both these rescaled weights and the donors' weights along a line segment, and determining which donors overlap each nonrespondent along the line (Williams and Folsom, 1981). Thus the set of donors who are eligible to donate to a given nonrespondent is a function of the sort order, the nonrespondent's sample weight, and the sample weights of all the donors. The algorithm is designed so that, over repeated imputations, the weighted mean obtained from the imputed values is equal in expectation to the weighted mean of the respondents alone within imputation strata. If response probability is constant within a cell then the WSHD leads to an unbiased estimator. "Similarity" of donor to recipient is still controlled by the choice of sorting variables.

Adjustments to the random selection method that incorporate the sample weights include inflating the donated value by the ratio of the sample weight of the donor to that of the recipient (Platek and Gray, 1983) or selecting donors via random draw with probability of selection proportional to the potential donor's sample weight (Rao and Shao, 1992; Rao, 1996). The former method has drawbacks, particularly in the case of integer-valued imputed values, since the imputations may no longer be plausible values. The latter method does not suffer from this inconsistency problem and yields an asymptotically unbiased estimator, assuming constant response probability within an adjustment cell. Note that in contrast to the weighted sequential hot deck, the sample weights of nonrespondents are not used in determining the selection probabilities of donors. We refer to this method as the weighted random hot deck (WRHD) to distinguish it from the weighted sequential hot deck (WSHD).

We suggest that neither WRHD nor WSHD are appropriate ways of incorporating design weights into the hot deck. Specifically, both the WSHD and WRHD fail to remove bias if outcome is related to the design weights and response propensity is not constant within an adjustment cell. The correct approach is to create donor pools based on stratification by auxiliary variables *and* design variables that determine the sampling weights. The goal should be to create imputation cells that are homogeneous with respect to both the outcome and the propensity to respond. Creating cells by cross-classification of both auxiliary and design variables is the best way to achieve this goal, in so far as these variables are associated with outcomes and nonresponse. With adjustment cells created in this way, draws proportional to sample weights are unnecessary and inefficient. One concern with this method is that if response is not related to the design variables, excess noise is added by over-stratifying without an accompanying bias reduction. However, simulations in Collins, Schafer, and Kam (2001) suggest that the benefits of reduction in bias outweigh the increase in variance. Little and Vartivarian (2003) demonstrated by simulation that when weighting for nonresponse adjustment, computing the unweighted response rate applied within cells defined by auxiliary and design variables was the correct approach, and that weighting the nonresponse rates using the sampling weights does not remove bias in all cases. In the next section we describe a simulation study, which shows that a similar scenario holds for the hot deck estimators.

## 3.3 Simulation Study

A simulation study was conducted to compare the performance of the various forms of the hot deck under a variety of population structures and nonresponse mechanisms. We build on the simulation in Little and Vartivarian (2003) which compared weighting estimators for the population mean. Categorical variables were simulated to avoid distributional assumptions such as normality.

### 3.3.1 Description of the Population

As in Little and Vartivarian (2003), a population of size 10000 was generated on a binary stratifier $Z$ known for all population units, a binary adjustment variable $X$ observed for the sample, and a binary survey outcome $Y$ observed only for respondents. Taking $S$ to be the sampling indicator and $R$ the response indicator, the joint distribution of these variables, say $[Z, X, Y, S, R]$ can be factorized as follows:

$$[X, Z, Y, S, R] = [X, Z][Y|X, Z][S|X, Z, Y][R|X, Z, Y, S]$$

The distributions on the right side was then defined as follows:

(a) *Distribution of $X$ and $Z$.*

The joint distribution of $[X, Z]$ was multinomial, with $\Pr(X = 0, Z = 0) = 0.3$, $\Pr(X = 1, Z = 0) = 0.4$, $\Pr(X = 0, Z = 1) = 0.2$, and $\Pr(X = 1, Z = 1) = 0.1$.

(b) *Distribution of $Y$ given $X$ and $Z$.*

Population values of the survey variable $Y$ were generated according to the logistic model

$$\operatorname{logit}(\Pr(Y=1|X,Z)) = 0.5 + \gamma_X(X-\bar{X}) + \gamma_Z(Z-\bar{Z}) + \gamma_{XZ}(X-\bar{X})(Z-\bar{Z})$$

for five choices of $\gamma = (\gamma_X, \gamma_Z, \gamma_{XZ})$ chosen to reflect different relationships between $Y$ and $X$ and $Z$. These choices are displayed in Table 3.1 using conventional linear model notation. For example, the additive logistic model $[X+Z]^Y$ sets the interaction $\gamma_{XZ}$ to zero, whereas the model $[XZ]^Y$ sets this interaction equal to 2. The models $[X]^Y$ and $[Z]^Y$ allow the outcome to depend on $X$ only and $Z$ only. The null model, where outcome is independent of $X$ and $Z$, is denoted $[\phi]^Y$.

(c) *Distribution of $S$ given $Z$, $X$, and $Y$.*

The sample cases were assumed to be selected by stratified random sampling, so $S$ is independent of $X$ and $Y$ given $Z$, that is $[S|X,Z,Y] = [S|Z]$. Two different sample sizes were evaluated. A sample of $n_0 = 125$ was drawn from the stratum with $Z = 0$ and size $n_1 = 25$ from the stratum with $Z = 1$, yielding a total sample size of 150. A larger sample of size 600 was then obtained by sampling $n_0 = 500$ and $n_1 = 100$ from the strata with $Z = 0$ and $Z = 1$ respectively.

(d) *Distribution of $R$ given $Z$, $X$, $Y$, and $S$.*

Since the response mechanism is assumed ignorable and the selection was by stratified random sampling, $R$ is independent of $Y$ and $S$ given $X$ and $Z$, i.e. $[R|Z,X,Y,S] = [R|Z,X]$. The latter was generated according to the logistic model

$$\operatorname{logit}(\Pr(R=1|X,Z)) = 0.5 + \beta_X(X-\bar{X}) + \beta_Z(Z-\bar{Z}) + \beta_{XZ}(X-\bar{X})(Z-\bar{Z})$$

where $\beta = (\beta_X, \beta_Z, \beta_{XZ})$ took the same values as $\gamma$, found in Table 3.1. As with the distribution of $Y$ given $X$ and $Z$, this yielded five models for the distribution of $R$ given $X$ and $Z$. For example, $[X+Z]^R$ refers to an additive logistic model for $R$ given $X$ and $Z$. This produced an average response rate over all simulations of 60%.

There were a total of $5 \times 5 = 25$ combinations of population structures and nonresponse mechanisms in the simulation study and two different sample sizes. A total of 1000 replicate populations of $(X, Z, Y, S, R)$ were generated for each of the $25 \times 2$ combinations.

### 3.3.2 Estimators

A total of seven methods for estimating the population mean were employed. Four versions of the hot deck were used to impute missing values, followed by computing the usual sample-weighted Horvitz-Thompson estimator for the population mean. The four hot deck methods are summarized in Table 3.2. All hot deck methods stratify on X, that is, perform imputation separately for units with $X = 0$ and $X = 1$. The weighted hot deck methods, wrhd(x) and wshd(x), use information in $Z$ in determining donor probabilities, in contrast to uhd(xz), which imputes within cells additionally defined by $Z$, and uhd(x), which ignores the information in $Z$. We implemented the wshd(x) in both a sorted (by $Z$, within adjustment cells) and unsorted form, the results were similar and we report only the unsorted results. In addition, three weighting estimators were used to estimate the population average without imputation, shown in Table 3.3. The weighting estimators wrr(x) and urr(xz) are analogous to the hot deck methods wrhd(x) and uhd(xz), respectively. We expected to see higher variance with the hot deck methods, but parallel results in terms of bias. For each replicate we also calculated the complete-case estimate using the Horvitz-Thompson estimator, with weights unadjusted for nonresponse.

Finally, for comparison purposes we calculated the before-deletion estimate using the Horvitz-Thompson estimator, that is, before sampled units with $R = 0$ had their $Y$ values deleted. This captures simulation variance in measures of bias and acts as a benchmark for evaluating increases in root mean square error due to nonresponse.

Empirical bias and root mean square error (RMSE) for each method $M$ were calculated as follows:

$$\text{EBias} = \frac{1}{1000}\sum_{i=1}^{1000}(\hat{\theta}_{Mi} - \theta_i) \tag{3.1}$$

$$\text{RMSE} = \sqrt{\frac{1}{1000}\sum_{i=1}^{1000}(\hat{\theta}_{Mi} - \theta_i)^2} \tag{3.2}$$

where $\hat{\theta}_{Mi}$ is the estimate of the population mean using method $M$ for the $i$th replicate and $\theta_i$ is the full population mean for the $i$th replicate. Selected pairs of hot deck estimators were compared to determine if differences in performance were statistically significant. The average difference between a pair of estimators was calculated as

$$\bar{d} = \frac{1}{1000}\sum_{i=1}^{1000}|\hat{\theta}_{BDi} - \hat{\theta}_{1i}| - |\hat{\theta}_{BDi} - \hat{\theta}_{2i}| \tag{3.3}$$

where for the $i$th replicate $\hat{\theta}_{BDi}$ is the estimated sample mean before-deletion of cases due to non-response and $\hat{\theta}_{1i}$ and $\hat{\theta}_{2i}$ are estimates found after imputation with the two different hot deck methods being compared.

### 3.3.3 Results

Tables 3.4 and 3.5 display the empirical bias for all seven methods as well as the complete case and before-deletion estimates for the smaller and larger sample sizes.

Tables 3.6 and 3.7 show the percent increase in RMSE for each method over the before-deletion method for sample sizes $n = 150$ and $n = 600$ respectively. Table 3.8 displays $\bar{d}$ ($\times 10,000$) for the comparison of uhd(xz) with each of the other three hot deck methods for the smaller sample size; results were similar for the larger sample size and are not shown. Differences that are statistically significant from zero based on a t-test are asterisked ($* = p < 0.05$, $** = p < 0.01$).

As shown in Table 3.4, the unweighted hot deck using cells based on $X$ and $Z$, uhd(xz), has small empirical bias in all population structures. With this method, the expected outcome and response propensity are constant within a cell, regardless of the model for $Y$ and $R$, so imputation leads to an unbiased estimate of the population mean. This is similar to the weighting estimator that uses unweighted response rates but stratifies on both $X$ and $Z$, urr(xz), which also has low empirical bias over all populations. Not surprisingly, the hot deck estimator that ignores $Z$, uhd(x), is biased for situations where $Y$ depends on $Z$, since the dependence on $Z$ cannot be ignored. However, the weighted hot decks (wrhd(x) and wshd(x)) do not correct the bias for all these cases. When the response propensity does not depend on $Z$, both wrhd(x) and wshd(x) have low bias, since the response propensity is constant within their adjustment cells (based on $X$ only). If the response propensity is not constant within adjustment cells, as in populations where $R$ depends on $Z$, then wrhd(x) and wshd(x) are biased and in fact have larger bias than the method that ignores $Z$, though we believe this to be an artifact of the simulation design and cannot conclude that uhd(x) would always outperform wrhd(x) and wshd(x) in these situations. This parallels the performance of the weighting methods that stratify on $X$ only (wrr(x), urr(x)), which have similar performance with two exceptions. As noted in Little and Vartivarian (2003), wrr(x) outperforms urr(x) where $R$ depends on both $X$ and $Z$

and $Y$ depends on $X$ but not $Z$ (specifically rows 11 and 12 of Table 3.4). This is not seen with the hot deck methods; all hot deck methods have low bias for populations where the outcome $Y$ does not depend on $Z$, regardless of the model for $R$. When $Y$ depends only on $X$, both the weighted and unweighted respondent means are unbiased within cells defined by $X$. Thus the hot deck methods are all unbiased, as over repeated imputations they impute the (weighted) respondent mean to the nonrespondents. For the weighting methods, using unweighted response rates as in urr(x) yields biased estimates of the response rate, and thus biased estimates of the overall mean, and weighting the response rates as in wrr(x) corrects this bias.

All hot deck and weighting methods perform well in terms of bias when the outcome is independent of $X$ and $Z$, regardless of the response model. Of note, in comparing the average absolute errors, wshd(x) has statistically significantly lower empirical bias than uhd(xz) when $Y$ does not depend on $Z$, though the size of the difference is small compared to the differences seen when uhd(xz) outperforms the weighting methods.

When missingness is independent of $X$ and $Z$, that is, missingness is completely at random (Rubin, 1976), the complete case estimator is unbiased. Nonresponse adjustment via any of these methods is unnecessary but not harmful in almost all cases. All hot deck and weighting methods produce unbiased estimates with one exception: the unweighted hot deck that ignores $Z$, uhd(x), induces bias when the outcome is dependent on $Z$ (populations 5, 10, and 20). In this case the nonresponse compensation has an adverse effect and is dangerous, demonstrating the need to condition on as much auxiliary data as is available.

A crude summary of the overall performance of the methods is the average of the percent increase in RMSE over all populations, shown at the bottom of Tables 3.6

and 3.7. The best overall hot deck method under both sample sizes is uhd(xz), which as expected has higher RMSE than the best overall weighting method, urr(xz). Differences between uhd(xz) and other hot deck methods follow similar patterns for both sample sizes but are exaggerated with the larger sample size ($n = 600$). The worst hot deck method is the weighted random hot deck, with a higher overall RMSE than the sequential version. Somewhat surprisingly, the unweighted hot deck showed lower overall RMSE than both the weighted hot decks and two of the weighting methods (wrr(x), urr(x)). Though uhd(x) is biased in more scenarios, the magnitude of the bias is much lower than wrhd(x), wshd(x), wrr(x), and urr(x), and this difference drives the difference in RMSE. We reiterate that this finding is likely an artifact of the simulation design, and in fact though the bias is smaller, uhd(x) is biased for a larger number of populations than the weighted hot deck methods. The sequential version of the weighted hot deck (wshd(x)) has lower RMSE than wrhd(x) in all populations for both sample sizes, and in fact has the lowest (or in one case just slightly larger than the lowest) RMSE among hot deck methods when $Y$ does not depend on $X$ or $Z$.

Overall, the unweighted hot deck that stratifies on both design and covariate information is robust under all scenarios, and the expected increase in RMSE when response does not depend on the design variable was not severe. In fact uhd(xz) had very similar RMSE to the unweighted method that stratified on $X$ only, uhd(x), in the ten populations where $Y$ did not depend on $Z$, demonstrating that over-stratifying at least in this case did not lead to a notable increase in variance. Of the weighted hot deck methods, the sequential version performed slightly better than the method using weighted draws from the donor pools.

## 3.4 Application

The third National Health and Nutrition Examination Survey (NHANES III) was a large-scale stratified multistage probability sample of the noninstitutionalized U.S. population conducted during the period from 1988 to 1994 (U.S. Department of Health and Human Services, 1994). NHANES III collected data in three phases: (a) a household screening interview, (b) a personal home interview, and (c) a physical examination at a mobile examination center (MEC). The total number of persons screened was 39,695, with 86% (33,994) completing the second phase interview. Of these, only 78% were examined in the MEC. Previous imputation efforts for NHANES III focused on those individuals who had completed the second phase; weighting adjustments are used to compensate for non-response at this second stage. Since the questions asked at both the second and third stage varied considerably by age we chose to select only adults age 20 and older who had completed the second phase interview for the purposes of our example, leaving a sample size of 18,825. Design variables that were fully observed for the sample included age, gender, race, and household size.

In order to demonstrate the hot deck methods on a continuous outcome we used systolic blood pressure measured at the MEC examination (SBP, defined as the average of three recorded measurements). The nonresponse rate was 16%. As our stratification variable ($X$) we chose a self-reported health status variable (Excellent/Very Good/Good/Fair/Poor) from the household interview. Since only 6% of subjects reported the lowest level of health status, the lowest two categories (Fair/Poor) were combined, leaving 4 strata. The $Z$ variables were the design variables: gender (2 levels), race (3 levels), age (3 levels), and household size (3 levels). The goal was to estimate the population mean of SBP.

In order to demonstrate the effect of larger nonresponse rates we increased the missingness as follows. First, we fit a logistic regression model on an indicator for missingness of SBP using the entire sample (n=18,825), using main effects for health status and all design variables as predictors, leaving the variables age and log(household size) as continuous. This created predicted probabilities of non-response mimicking the actual propensities observed in the NHANES data and ranging from 0.05 to 0.39. The mean probability for respondents was 0.15; in order to double the missingness to 32% we required an additional 19% of the respondents to have missing values, so each predicted probability was increased by 0.04. Nonresponse indicators for each respondent were then independently drawn from a Bernoulli distribution with these predicted probabilities and values were subsequently deleted from the sample to create a second data set.

The four different imputation strategies implemented in the simulation study were applied to each of the two data sets. The weighted hot deck methods, wrhd(x) and wshd(x), stratified by health status and used the sample weights to determine donor probabilities within the donor pools. The most naive hot deck method, uhd(x) stratified by health status and ignored the sample weights, and the fully stratified method, uhd(xz) stratified by both health status and the design variables for a total of 215 donor cells (one cell was empty). Complete case estimates were also calculated. In order to obtain measures of variability and better compare estimates, imputation was via the Approximate Bayesian Bootstrap (Rubin and Schenker, 1986). Within each adjustment cell the respondent values were resampled with replacement to form a new pool of potential donors and the imputation method (wrhd(x), wshd(x), uhd(x), uhd(xz)) was then applied to this bootstrapped donor pool. This method is easy to compute, and repeated applications yield proper multiple imputations. A total of 10 multiply imputed data sets were created for each method, the Horvitz-Thompson estimator of the mean SBP calculated for each data set, and resulting inference obtained using the combining rules of Rubin (1987).

Resulting mean estimates and 95% confidence intervals are displayed in Figure 3.1 for both the original 16% missingness and the induced 32% missingness. The larger level of nonresponse showed more exaggerated differences in performance between the methods. For both scenarios the weighted hot deck methods (wrhd(x) and wshd(x)) lead to intervals that are close to the complete case estimates. The uhd(xz) method generates estimates that are higher than those of the weighted methods, with the difference becoming more exaggerated with the larger amount of nonresponse. The mean estimate for uhd(xz) is the same across moth missingness scenarios, which is comforting since the overall mean should be the same in both cases, while both wrhd(x) and wshd(x) parallel the complete case estimate and show a downward shift under 32% missingness. The unweighted hot deck that ignores the weights (uhd(x)) also shows a downward shift as missingness increases. One feature that is evident with these data that did not appear in the simulations is the increase in variance with uhd(xz) – for the larger amount of missingness the confidence interval for uhd(xz) is larger than that of the weighted methods, though the difference is minimal. Though the “truth”’ is not available for this real data set, the performance of uhd(xz) appears to be the most robust as it produces similar estimates under both missingness mechanisms.

## 3.5 Conclusion

The simulation study suggests strongly that the two forms of sample-weighted hot deck (WSHD and WRHD) do not correct for bias when the outcome is related to the sampling weight and the response propensity, and are inferior to the method that uses the sampling weight as a stratifying variable when forming adjustment cells. The simulation study focused on estimating a mean and was deliberately kept simple, but it varied systematically the key elements of the problem, namely the relationship between the outcome and the response propensity and the sampling stratum and adjustment cell variable. It seems to us unlikely that more complex simulations will lead to different conclusions, although admittedly this possibility cannot be ruled out. The conclusions parallel similar results for weighting nonresponse adjustments in Little and Vartivarian (2003). Weighting adjustments are a bit more efficient than the hot deck, since the latter is effectively adding noise to the estimates to preserve distributions. However, the hot deck is a more flexible approach to item nonresponse than weighting, and the added noise from imputing real values from donors can be reduced by applying the hot deck repeatedly to generate multiply-imputed data sets (Rubin, 1987). Since a benefit of the hot deck is the preservation of associations among variables, future evaluation of these methods when estimating a second order relation such as a correlation or regression coefficient would be of interest. However we conjecture that methods that condition on the design information would outperform sample-weighted hot deck methods for these kinds of estimands, as they do for the mean.

The main drawback to creating adjustment cells that stratify on sampling strata as well as other covariate information is that it may lead to a large number of cells, and hence some cells where there are no donors for a case with missing values. With an extensive set of covariates $X$ and $Z$, imputation based on the multiple regression of $Y$ on $X$ and $Z$ maintains the logic of the suggested approach while accommodating extensive sets of covariates. Specifically, a hot deck approach is to create adjustment cells based on the predicted means from the regression of $Y$ on $X$ and $Z$, or to generate donors for incomplete cases based on predictive mean matching (Little, 1986). For a review of recent extensions of hot deck adjustment cell methods, including predictive mean matching, see Andridge and Little (2008).

Table 3.1: Models for $Y$ given $X$, $Z$

| | $\gamma_X$ | $\gamma_Z$ | $\gamma_{XZ}$ |
|---|---|---|---|
| $[XZ]^Y$ | 2 | 2 | 2 |
| $[X+Z]^Y$ | 2 | 2 | 0 |
| $[X]^Y$ | 2 | 0 | 0 |
| $[Z]^Y$ | 0 | 2 | 0 |
| $[\phi]^Y$ | 0 | 0 | 0 |

Table 3.2: Hot Deck Methods

| Method | | Adjustment Cells | Draws |
|---|---|---|---|
| wrhd(x) | Weighted Random Hot Deck | $X$ | Proportional to sample weight |
| wshd(x) | Weighted Sequential Hot Deck | $X$ | n/a |
| uhd(x) | Unweighted Hot Deck | $X$ | Equal probability |
| uhd(xz) | Unweighted Hot Deck | $X$ and $Z$ | Equal probability |

Table 3.3: Weighting Methods

| Method | | Adjustment Cells | Response Rate |
|---|---|---|---|
| wrr(x) | Weighted Response Rate | $X$ | Weighted |
| urr(x) | Unweighted Response Rate | $X$ | Unweighted |
| urr(xz) | Unweighted Response Rate | $X$ and $Z$ | Unweighted |

Table 3.4: 1,000 x (Average Empirical Bias) of 1000 replicate samples (n=150)

| Generated Model for $Y$ and $R$ | | | Hot deck estimators | | | | Weighting estimators | | | Complete | Before |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | $[\,]^Y$ | $[\,]^R$ | wrhd(x) | wshd(x) | uhd(x) | uhd(xz) | wrr(x) | urr(x) | urr(xz) | case | deletion |
| 1 | $XZ$ | $XZ$ | 22 | 22 | *4* | *-4* | 21 | 17 | -4 | 66 | 0 |
| 2 | $XZ$ | $X+Z$ | 37 | 37 | 21 | *1* | 37 | 27 | 2 | 71 | 2 |
| 3 | $XZ$ | $X$ | *-2* | *-2* | -13 | *-2* | -2 | -2 | -1 | 57 | 0 |
| 4 | $XZ$ | $Z$ | 30 | 28 | 14 | *-1* | 29 | 27 | -1 | 21 | -1 |
| 5 | $XZ$ | $\phi$ | *0* | *0* | -13 | 1 | 0 | 0 | 0 | 0 | 1 |
| 6 | $X+Z$ | $XZ$ | 37 | 37 | 10 | *0* | 37 | 33 | 1 | 78 | 2 |
| 7 | $X+Z$ | $X+Z$ | 59 | 59 | 34 | *2* | 59 | 51 | 1 | 87 | 0 |
| 8 | $X+Z$ | $X$ | -3 | -3 | -27 | *-1* | -3 | -2 | -1 | 59 | -1 |
| 9 | $X+Z$ | $Z$ | 39 | 41 | 21 | *1* | 41 | 39 | 2 | 33 | 0 |
| 10 | $X+Z$ | $\phi$ | *0* | -1 | -18 | -1 | -1 | 0 | 0 | 0 | 0 |
| 11 | $X$ | $XZ$ | *0* | 1 | *0* | 1 | 0 | -6 | 0 | 65 | -1 |
| 12 | $X$ | $X+Z$ | *0* | -1 | *0* | -1 | 0 | -16 | 0 | 54 | -1 |
| 13 | $X$ | X | 1 | *0* | -1 | 1 | 0 | 1 | 0 | 84 | 0 |
| 14 | $X$ | $Z$ | *-1* | *-1* | *-1* | -2 | -1 | -4 | -1 | -13 | 1 |
| 15 | $X$ | $\phi$ | -1 | *0* | *0* | -1 | -1 | -1 | -1 | -1 | 1 |
| 16 | $Z$ | $XZ$ | 36 | 37 | 11 | *-1* | 36 | 38 | 0 | 20 | 0 |
| 17 | $Z$ | $X+Z$ | 52 | 52 | 29 | *-2* | 52 | 58 | -3 | 33 | -2 |
| 18 | $Z$ | $X$ | -2 | *-1* | -25 | *-1* | -2 | -1 | 0 | -17 | 0 |
| 19 | $Z$ | $Z$ | 43 | 41 | 20 | *-2* | 41 | 42 | -2 | 44 | 0 |
| 20 | $Z$ | $\phi$ | *-3* | -4 | -23 | *-3* | -4 | -3 | -3 | -3 | -2 |
| 21 | $\phi$ | $XZ$ | -2 | -1 | 1 | *0* | -1 | -1 | 0 | -1 | -1 |
| 22 | $\phi$ | $X+Z$ | *-2* | -3 | *-2* | -3 | -3 | -3 | -3 | -2 | -1 |
| 23 | $\phi$ | $X$ | *0* | -1 | -2 | -2 | -2 | -2 | -1 | -1 | -1 |
| 24 | $\phi$ | $Z$ | *1* | *1* | 2 | 2 | 1 | 1 | 1 | 1 | 1 |
| 25 | $\phi$ | $\phi$ | *4* | *4* | *4* | *4* | 4 | 4 | 4 | 4 | 2 |
| Mean | | | 14 | 14 | 2 | -1 | 14 | 12 | 0 | 30 | 0 |
| Mean absolute average empirical bias | | | 15 | 15 | 12 | 2 | 15 | 15 | 1 | 33 | 1 |

Smallest absolute empirical average bias among hot deck methods shown in italics.

Table 3.5: 1,000 x (Average Empirical Bias) of 1000 replicate samples (n=600)

| Generated Model for $Y$ and $R$ | | | Hot deck estimators | | | | Weighting estimators | | | Complete | Before |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | $[\,]^Y$ | $[\,]^R$ | wrhd(x) | wshd(x) | uhd(x) | uhd(xz) | wrr(x) | urr(x) | urr(xz) | case | deletion |
| 1 | $XZ$ | $XZ$ | 26 | 25 | 8 | *0* | 25 | 21 | 1 | 68 | 1 |
| 2 | $XZ$ | $X+Z$ | 35 | 35 | 20 | *0* | 35 | 25 | 0 | 68 | 0 |
| 3 | $XZ$ | $X$ | 2 | *1* | -14 | 2 | 1 | 2 | 2 | 59 | -1 |
| 4 | $XZ$ | $Z$ | 32 | 31 | 16 | *1* | 31 | 29 | 1 | 23 | 0 |
| 5 | $XZ$ | $\phi$ | -1 | -1 | -13 | *0* | 0 | 0 | 0 | 0 | 0 |
| 6 | $X+Z$ | $XZ$ | 36 | 37 | 9 | *1* | 37 | 33 | 0 | 78 | 0 |
| 7 | $X+Z$ | $X+Z$ | 57 | 58 | 32 | *0* | 58 | 49 | 0 | 86 | 0 |
| 8 | $X+Z$ | $X$ | -1 | -1 | -26 | *0* | -1 | 0 | 0 | 62 | 0 |
| 9 | $X+Z$ | $Z$ | 40 | 40 | 21 | *0* | 40 | 38 | 0 | 32 | 0 |
| 10 | $X+Z$ | $\phi$ | *0* | *0* | -18 | *0* | 0 | 0 | 0 | 0 | -1 |
| 11 | $X$ | $XZ$ | 2 | 1 | *0* | *0* | 1 | -5 | 1 | 67 | 1 |
| 12 | $X$ | $X+Z$ | *0* | *0* | *0* | -1 | 0 | -17 | -1 | 55 | 0 |
| 13 | $X$ | $X$ | *0* | *0* | -1 | -1 | 0 | 0 | 0 | 84 | 1 |
| 14 | $X$ | $Z$ | *2* | *2* | *2* | *2* | 2 | -2 | 2 | -10 | 1 |
| 15 | $X$ | $\phi$ | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 16 | $Z$ | $XZ$ | 37 | 37 | 11 | *1* | 37 | 39 | 1 | 21 | 0 |
| 17 | $Z$ | $X+Z$ | 56 | 56 | 31 | *1* | 56 | 61 | 2 | 36 | 1 |
| 18 | $Z$ | $X$ | -1 | -1 | -25 | *0* | -1 | -1 | 0 | -16 | 0 |
| 19 | $Z$ | $Z$ | 43 | 43 | 22 | *1* | 43 | 44 | 0 | 46 | 1 |
| 20 | $Z$ | $\phi$ | *-1* | *-1* | -20 | *-1* | -1 | -1 | -1 | -1 | 0 |
| 21 | $\phi$ | $XZ$ | *0* | *0* | *0* | *0* | 0 | 0 | 0 | 0 | 0 |
| 22 | $\phi$ | $X+Z$ | -1 | -1 | -1 | *0* | 0 | 0 | 0 | -1 | 0 |
| 23 | $\phi$ | $X$ | *1* | *1* | *1* | *1* | 1 | 1 | 1 | 1 | 0 |
| 24 | $\phi$ | $Z$ | *0* | *0* | *0* | *0* | 0 | 0 | 0 | 0 | 0 |
| 25 | $\phi$ | $\phi$ | 1 | *0* | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Mean | | | 14 | 14 | 2 | -1 | 14 | 12 | 0 | 30 | 0 |
| Mean absolute average empirical bias | | | 15 | 15 | 12 | 2 | 15 | 15 | 1 | 33 | 1 |

Smallest absolute empirical average bias among hot deck methods shown in italics.

Table 3.6: Percent increase in RMSE compared to before deletion estimate, 1000 replicate samples (n=150)

| Generated Model for $Y$ and $R$ | | | Hot deck estimators | | | | Weighting estimators | | | Complete |
|---|---|---|---|---|---|---|---|---|---|---|
| | $[]^Y$ | $[]^R$ | wrhd(x) | wshd(x) | uhd(x) | uhd(xz) | wrr(x) | urr(x) | urr(xz) | Case |
| 1 | $XZ$ | $XZ$ | 66 | 56 | *49* | 53 | 55 | 53 | 39 | 108 |
| 2 | $XZ$ | $X + Z$ | 77 | 71 | 57 | *45* | 68 | 62 | 35 | 115 |
| 3 | $XZ$ | $X$ | 60 | *52* | 59 | 63 | 48 | 48 | 48 | 89 |
| 4 | $XZ$ | $Z$ | 55 | 45 | 37 | *29* | 45 | 43 | 22 | 42 |
| 5 | $XZ$ | $\phi$ | 40 | *31* | 42 | 37 | 26 | 26 | 25 | 29 |
| 6 | $X + Z$ | $XZ$ | 83 | 78 | 58 | *49* | 75 | 71 | 38 | 139 |
| 7 | $X + Z$ | $X + Z$ | 116 | 109 | 74 | *40* | 108 | 97 | 31 | 159 |
| 8 | $X + Z$ | $X$ | 57 | *49* | 67 | 52 | 47 | 43 | 43 | 91 |
| 9 | $X + Z$ | $Z$ | 61 | 60 | 36 | *26* | 56 | 53 | 18 | 49 |
| 10 | $X + Z$ | $\phi$ | 43 | *33* | 51 | 39 | 30 | 29 | 26 | 31 |
| 11 | $X$ | $XZ$ | 57 | 48 | 49 | *47* | 45 | 50 | 37 | 107 |
| 12 | $X$ | $X + Z$ | 50 | *42* | 44 | 48 | 39 | 51 | 35 | 82 |
| 13 | $X$ | X | 53 | *41* | 47 | 55 | 39 | 40 | 44 | 132 |
| 14 | $X$ | $Z$ | 33 | 28 | 28 | *26* | 24 | 26 | 17 | 34 |
| 15 | $X$ | $\phi$ | 34 | *27* | 29 | 38 | 21 | 22 | 23 | 26 |
| 16 | $Z$ | $XZ$ | 90 | 82 | 70 | *62* | 78 | 84 | 50 | 50 |
| 17 | $Z$ | $X + Z$ | 99 | 93 | 65 | *51* | 90 | 105 | 38 | 53 |
| 18 | $Z$ | $X$ | 74 | *59* | 89 | 65 | 55 | 56 | 50 | 47 |
| 19 | $Z$ | $Z$ | 80 | 73 | 52 | *39* | 68 | 70 | 29 | 74 |
| 20 | $Z$ | $\phi$ | 50 | *40* | 68 | 47 | 35 | 35 | 32 | 35 |
| 21 | $\phi$ | $XZ$ | 59 | *48* | 53 | 53 | 46 | 48 | 40 | 32 |
| 22 | $\phi$ | $X + Z$ | 47 | *41* | 46 | 46 | 39 | 43 | 34 | 28 |
| 23 | $\phi$ | $X$ | 63 | *49* | 53 | 61 | 46 | 46 | 50 | 30 |
| 24 | $\phi$ | $Z$ | 37 | 33 | *32* | 34 | 28 | 29 | 22 | 29 |
| 25 | $\phi$ | $\phi$ | 43 | *32* | 38 | 46 | 29 | 29 | 30 | 29 |
| Mean Percent | | | 61 | 53 | 52 | 46 | 50 | 50 | 34 | 65 |

Lowest percent increase in RMSE among hot deck methods shown in italics.

Table 3.7: Percent increase in RMSE compared to before deletion estimate, 1000 replicate samples (n=600)

| Generated Model for $Y$ and $R$ | | | Hot deck estimators | | | | Weighting estimators | | | Complete |
|---|---|---|---|---|---|---|---|---|---|---|
| | $[]^Y$ | $[]^R$ | wrhd(x) | wshd(x) | uhd(x) | uhd(xz) | wrr(x) | urr(x) | urr(xz) | case |
| 1 | $XZ$ | $XZ$ | 102 | 91 | 58 | *53* | 90 | 78 | 37 | 259 |
| 2 | $XZ$ | $X+Z$ | 130 | 124 | 77 | *40* | 123 | 93 | 33 | 257 |
| 3 | $XZ$ | $X$ | 65 | *53* | 73 | 63 | 52 | 51 | 48 | 215 |
| 4 | $XZ$ | $Z$ | 105 | 98 | 55 | *34* | 96 | 88 | 23 | 71 |
| 5 | $XZ$ | $\phi$ | 37 | *31* | 51 | 34 | 24 | 24 | 21 | 26 |
| 6 | $X+Z$ | $XZ$ | 143 | 138 | 59 | *50* | 136 | 122 | 39 | 312 |
| 7 | $X+Z$ | $X+Z$ | 241 | 237 | 129 | *42* | 237 | 200 | 32 | 375 |
| 8 | $X+Z$ | $X$ | 71 | *57* | 112 | 59 | 53 | 49 | 44 | 243 |
| 9 | $X+Z$ | $Z$ | 153 | 149 | 74 | *30* | 147 | 137 | 20 | 113 |
| 10 | $X+Z$ | $\phi$ | 43 | *30* | 65 | 35 | 26 | 25 | 22 | 28 |
| 11 | $X$ | $XZ$ | 57 | *44* | 52 | 46 | 43 | 49 | 35 | 256 |
| 12 | $X$ | $X+Z$ | 51 | *41* | 46 | 42 | 39 | 64 | 32 | 191 |
| 13 | $X$ | X | 60 | *49* | 54 | 57 | 47 | 47 | 47 | 320 |
| 14 | $X$ | $Z$ | 29 | 26 | 28 | *24* | 22 | 22 | 16 | 37 |
| 15 | $X$ | $\phi$ | 40 | *34* | 37 | 39 | 26 | 27 | 27 | 31 |
| 16 | $Z$ | $XZ$ | 169 | 163 | 79 | *60* | 160 | 170 | 44 | 86 |
| 17 | $Z$ | $X+Z$ | 238 | 232 | 124 | *52* | 230 | 261 | 41 | 135 |
| 18 | $Z$ | $X$ | 76 | *59* | 129 | 64 | 56 | 57 | 47 | 70 |
| 19 | $Z$ | $Z$ | 179 | 178 | 89 | *45* | 174 | 178 | 28 | 186 |
| 20 | $Z$ | $\phi$ | 54 | *38* | 93 | 44 | 34 | 34 | 29 | 34 |
| 21 | $\phi$ | $XZ$ | 57 | *48* | 47 | 52 | 43 | 45 | 34 | 33 |
| 22 | $\phi$ | $X+Z$ | 57 | *49* | *49* | 50 | 45 | 49 | 37 | 35 |
| 23 | $\phi$ | $X$ | 66 | *51* | 56 | 63 | 47 | 47 | 48 | 32 |
| 24 | $\phi$ | $Z$ | 45 | *37* | *37* | *37* | 32 | 32 | 24 | 32 |
| 25 | $\phi$ | $\phi$ | 40 | *32* | 38 | 43 | 26 | 26 | 26 | 26 |
| Mean Percent | | | 92 | 84 | 68 | 46 | 80 | 79 | 33 | 136 |

Lowest percent increase in RMSE among hot deck methods shown in italics.

Table 3.8: Pairwise comparisons of average absolute error ($\bar{d}$ x 1,000) of hot deck methods (n=150)

| Generated model for $Y$ and $R$ | | | uhd(xz) and | uhd(xz) and | uhd(xz) and |
|---|---|---|---|---|---|
| | $[\,]^Y$ | $[\,]^R$ | wrhd(x) | wshd(x) | uhd(x) |
| 1 | $XZ$ | $XZ$ | -5.2** | -1.7 | 0.7 |
| 2 | $XZ$ | $X+Z$ | -11.2** | -8.1** | -2.9** |
| 3 | $XZ$ | $X$ | -0.4 | 3.4** | 0.2 |
| 4 | $XZ$ | $Z$ | -10.0** | -6.5** | -2.6** |
| 5 | $XZ$ | $\phi$ | 0.3 | 4.6** | -2.4** |
| 6 | $X+Z$ | $XZ$ | -12.9** | -10.9** | -3.2** |
| 7 | $X+Z$ | $X+Z$ | -31.2** | -29.3** | -11.9** |
| 8 | $X+Z$ | $X$ | -2.1 | 0.8 | -7.2** |
| 9 | $X+Z$ | $Z$ | -17.1** | -16.9** | -4.8** |
| 10 | $X+Z$ | $\phi$ | -2.0* | 3.0** | -5.6** |
| 11 | $X$ | $XZ$ | -0.4 | 3.2** | 0.6 |
| 12 | $X$ | $X+Z$ | 2.2* | 5.1** | 3.5** |
| 13 | $X$ | $X$ | 1.1 | 6.3** | 2.2* |
| 14 | $X$ | $Z$ | -0.5 | 3.0** | 0.7 |
| 15 | $X$ | $\phi$ | 1.5 | 4.6** | 1.2 |
| 16 | $Z$ | $XZ$ | -11.6** | -8.3** | -3.0** |
| 17 | $Z$ | $X+Z$ | -23.6** | -21.9** | -8.5** |
| 18 | $Z$ | $X$ | -5.7** | -0.1 | -10.3** |
| 19 | $Z$ | $Z$ | -18.0** | -14.8** | -6.4** |
| 20 | $Z$ | $\phi$ | -2.7** | 2.6** | -7.7** |
| 21 | $\phi$ | $XZ$ | -1.1 | 4.1** | 0.2 |
| 22 | $\phi$ | $X+Z$ | 4.3** | 6.8** | 3.4** |
| 23 | $\phi$ | $X$ | -1.2 | 4.3** | 1.6 |
| 24 | $\phi$ | $Z$ | 0.2 | 3.2** | 1.7* |
| 25 | $\phi$ | $\phi$ | 0.0 | 4.8** | 1.0 |

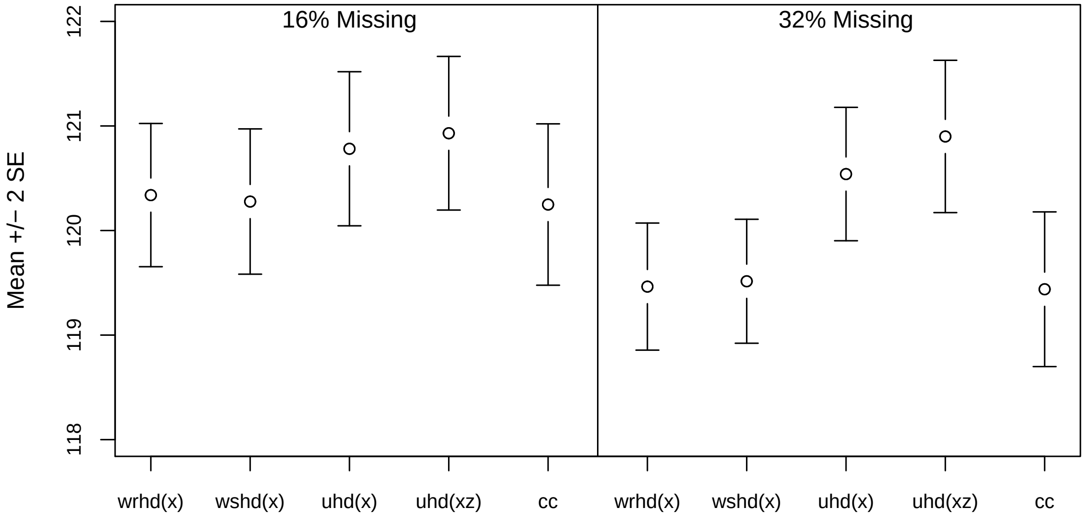
Negative value: First estimator does better

Positive value: Second estimator does better

* Significance at the 5 percent level

** Significance at the 1 percent level

Figure 3.1: Estimates of mean SBP for NHANES III data, after imputation with different hot deck methods. Original missingness was 16%; artificially increased missingness was 32%. Results from 10 multiply-imputed data sets. cc=Complete Case.



# CHAPTER IV

# Proxy Pattern-Mixture Analysis for Survey Nonresponse

## 4.1 Introduction

Missing data are often a problem in large-scale surveys, arising when a sampled unit does not respond to the entire survey (unit nonresponse) or to a particular question (item nonresponse). In this paper we focus on the adjustment for and measurement of nonresponse bias in a single variable $Y$ subject to missing values, when a set of variables $X$ are measured for both respondents and nonrespondents. With unit nonresponse this set of variables is generally restricted to survey design variables, except in longitudinal surveys where variables are measured prior to dropout. With item nonresponse, the set of observed variables can include survey items not subject to nonresponse, and hence is potentially more extensive. With a set of variables $Y$ subject to nonresponse, our methods could be applied separately for each variable, but we do not consider here methods for multivariate missing data where variables are missing for different sets of cases.

Limiting the impact of nonresponse is an important design goal in survey research, and how to measure and adjust for nonresponse is an important issue for statistical agencies and other data collectors, particularly since response rates are on the decline. Current U.S. federal standards for statistical surveys state, "Nonresponse bias analyses must be conducted when unit or item response rates or other factors suggest the potential for bias to occur," (Office of Management and Budget, 2006, p. 8) and go on to suggest that unit nonresponse rates of less than 80% require such an analysis. However, specific analysis recommendations are lacking, focusing on methods for accurately calculating response rates. While the response rate is clearly an important feature of the problem, there is a tension between increasing response rates and increasing response error by including respondents with no inclination to respond accurately. Indeed, some studies have shown that response rates are a poor measure of nonresponse bias (e.g. Curtain, Presser, and Singer, 2000; Keeter, Miller, Kohut, Groves, and Presser, 2000).

There are three major components to consider in evaluating nonresponse: the amount of missingness, differences between respondents and nonrespondents on characteristics that are observed for the entire sample, and the relationship between these fully observed covariates and the survey outcome of interest. Each facet provides some information about the impact of nonresponse, but no single component completely tells the story. Historically the amount of missingness, as measured by the response rate, has been the most oft-used metric for evaluating survey quality. However, response rates ignore the information contained in auxiliary covariates. Federal reports have recommended the second component, evaluating nonresponse based on differences between respondents and nonrespondents (Federal Committee on Statistical Methodology, 2001). A related approach is to focus on measures based on the response propensity, the estimated probability of response given the covariates, which is the auxiliary variable that is most different between respondents and nonrespondents. Measures such as the variability of nonresponse weights indicate the potential of weighting for nonresponse bias reduction, and lack of variability can suggest missingness is completely at random. Though response propensity analyses are appealing, nonresponse bias depends on the strength of the correlation between the survey variable of interest and the probability of response, and bias will vary across items in a single survey (Bethlehem, 2002; Groves, 2006).

The final component is the value of the auxiliary information in predicting survey outcomes. Suppose $Y$ is a survey outcome subject to nonresponse, $X$ is an auxiliary variable observed for respondents and nonrespondents, and missing values of $Y$ are imputed by predictions of the regression of $Y$ on $X$ estimated using the respondent sample. If data are missing completely at random, the variance of the mean of $Y$ based on the imputed data under simple random sampling is asymptotically

$$Var(\widehat{\mu}_y) = \frac{\sigma_{yy}}{r}\left(1 - \frac{n-r}{n}\rho^2\right),$$

where $n$ is the sample size, $r$ is the number of respondents, $\sigma_{yy}$ is the variance of $Y$, and $\rho$ is the correlation between $X$ and $Y$ (see Little and Rubin, 2002, equation 7.14). The corresponding fraction of missing information, the loss of precision from the missing data, is

$$FMI = \frac{n/\sigma_{yy} - Var^{-1}(\hat{\mu}_y)}{n/\sigma_{yy}}.$$

This fraction varies from the nonresponse rate $(n-r)/n$ when $\rho^2 = 0$ to 0 when $\rho^2 = 1$. With a set of covariates $Z$, imputation based on the multiple regression of $Y$ on $Z$ yields similar measures, with $\rho^2$ replaced by the squared coefficient of determination of the regression of $Y$ on $Z$. This approach is attractive since it gives appropriate credit to the availability of good predictors of $Y$ in the auxiliary data as well as a high response rate, and arguably good prediction of the survey outcomes is a key feature of good covariates; in particular, conditioning on a covariate $Z$ that is a good predictor of nonresponse but is unrelated to survey outcomes simply results in increased variance without any reduction in bias (Little and Vartivarian, 2005). A serious limitation with this approach is that it is more focused on precision than bias, and it assumes the data are missing at random (MAR); that is, missingness of $Y$ is independent of $Y$ after conditioning on the covariates $Z$ (Rubin, 1976). Also, this approach cannot provide a single measure of the impact of nonresponse, since by definition measures are outcome-specific.

Previous work has focused on distinct measures based on these considerations, but has not integrated them in a satisfactory way. We propose a new method for nonresponse bias measurement and adjustment that takes account all three aspects, in a way which we find intuitive and satisfying. In particular, it gives appropriate credit for predictive auxiliary data, without making the MAR assumption, which is implicit in existing propensity and prediction methods; our methods are based on a pattern-mixture model (Little, 1993) for the survey outcome that allows missingness to be not at random (NMAR) and assesses the sensitivity of estimates to deviation from MAR. We prefer a sensitivity analysis approach over approaches that require strong distributional and other assumptions on the missingness mechanism for estimation such as the selection models arising from the work of Heckman (1976). For more discussion of this point see for example Little and Rubin (2002, chap. 15) and citations therein. As a measure of the impact of nonresponse, we propose using the estimated fraction of missing information, obtained through multiple imputation under the pattern-mixture model with a range of assumptions about the nonresponse mechanism.

Section 4.2 introduces our approach to the nonresponse problem and describes the general framework, and Section 4.3 details the corresponding pattern-mixture model analysis. Section 4.4 describes three different estimation approaches: maximum likelihood, a Bayesian approach, and multiple imputation. Section 4.5 discusses the use of the fraction of missing information from multiple imputation under the pattern-mixture model as a measure of nonresponse bias. Section 4.6 describes a set of simulation studies to demonstrate the assessment of nonresponse bias using these methods. Section 4.7 applies these methods to data from NHANES III. Section 4.8 presents discussion, including extensions of the proposed method.

## 4.2 General Framework

We consider the problem of assessing nonresponse bias for estimating the mean of a survey variable $Y$ subject to nonresponse. For simplicity, we initially consider an infinite population with a sample of size $n$ drawn by simple random sampling. Let $Y_i$ denote the value of a continuous survey outcome and $Z_i = (Z_{i1}, Z_{i2}, \ldots, Z_{ip})$ denote the values of $p$ covariates for unit $i$ in the sample. Only $r$ of the $n$ sampled units respond, so observed data consist of $(Y_i, Z_i)$ for $i = 1, \ldots, r$ and $Z_i$ for $i = r+1, \ldots, n$. In particular this can occur with unit nonresponse, where the covariates $Z$ are design variables known for the entire sample or with item nonresponse. Of primary interest is assessing and correcting nonresponse bias for the mean of $Y$.

For simplicity and to reduce dimensionality, we replace $Z$ by a single proxy variable $X$ that has the highest correlation with $Y$. This proxy variable can be estimated by regressing $Y$ on $Z$ using the respondent data, including important predictors of $Y$ , as well as interactions and nonlinear terms where appropriate. The regression coefficients are subject to sampling error, so in practice $X$ is estimated rather than known, but we address this complication later. Let $\rho$ be the correlation of $Y$ and $X$, which we assume is positive. If $\rho$ is high (say, 0.8) we call $X$ a strong proxy for $Y$ and if $X$ is low (say, 0.2) we call $X$ a weak proxy for $Y$. The distribution of

$X$ for respondents and nonrespondents provides the main source of information for assessing nonresponse bias for $Y$.

Let $\bar{y}_R$ denote the respondent mean of $Y$, which is subject to nonresponse bias. We consider adjusted estimators of the mean $\mu_y$ of $Y$ of the form

$$\hat{\mu}_y = \bar{y}_R + g(\hat{\rho})\sqrt{\frac{s_{yy}}{s_{xx}}}(\bar{x} - \bar{x}_R), \tag{4.1}$$

where $\bar{x}_R$ is the respondent mean of $X$, $\bar{x}$ is the sample mean of $X$, and $s_{xx}$ and $s_{yy}$ are the respondent sample variances of $X$ and $Y$. Note that since the proxy $X$ is the conditional mean of $Y$ given $X$ it will have lower variance than $Y$. Rearranging terms yields the standardized bias in $\bar{y}_R$ as a function of the standardized bias in $\bar{x}_R$,

$$\frac{\hat{\mu}_y - \bar{y}_R}{\sqrt{s_{yy}}} = g(\hat{\rho})\frac{\bar{x} - \bar{x}_R}{\sqrt{s_{xx}}}. \tag{4.2}$$

Some comments on the estimator (1) follow. The classical regression estimator is obtained when $g(\hat{\rho}) = \hat{\rho}$, and this is an appropriate choice when missingness depends on the proxy $X$. It is also appropriate more generally when the data are missing at random (MAR), that is, missingness depends on $Z$, if $Y|Z$ is normal, and models are well specified. This is true because under MAR, the partial association between the residual $Y - X$ and the missing data indicator (say $M$) is zero.

In general, we may want the weight $g(\hat{\rho})$ given to the standardized proxy data to increase with the strength of the proxy, and $g(\hat{\rho}) \to 1$ as $\hat{\rho} \to 1$, that is, as the proxy variable converges towards the true variable $Y$. The size of the deviation, $d = \bar{x} - \bar{x}_R$, and its standardized version, $d^* = d/\sqrt{s_{xx}}$, is a measure of the deviation from missing completely at random (MCAR), and as such is the “observable” component of nonresponse bias for $Y$. Specific choices of $g(\hat{\rho})$ based on a pattern-mixture model are presented in the next section.

The information about nonresponse bias for $Y$ depends on the strength of the proxy, as measured by $\hat{\rho}$, and the deviation from MCAR, as measured by the size of $d$. We consider four situations, ordered from what we consider most favorable to least favorable from the point of view of the quality of this information for nonresponse bias assessment and adjustment.

1. If $X$ is a strong proxy (large $\hat{\rho}$), and $d$ is small, then the adjustment via (4.1) is small and the evidence of a lack of nonresponse bias in $Y$ is relatively strong, since it is not evident in a variable highly correlated with $Y$. This is the most favorable case.

2. If $X$ is a strong proxy, and $d$ is large, then there is strong evidence of response bias in respondent mean $\bar{y}_R$ but good information for correcting the bias using the proxy variable via (4.1). Since an adjustment is needed, model misspecification is a potential issue.

3. If $X$ is a weak proxy (small $\hat{\rho}$), and $d$ is small, then the adjustment via (4.1) is small. There is some evidence against nonresponse bias in the fact that $d$ is small, but this evidence is relatively weak since it does not address the possibility of bias from unobserved variables related to $Y$.

4. If $X$ is a weak proxy, and $d$ is large, then the adjustment via (4.1) depends on the choice of $g(\hat{\rho})$, although it is small under the MAR assumption when $g(\hat{\rho}) = \hat{\rho}$. There is some evidence that there is nonresponse bias in $Z$ in the fact that $d$ is large, but no evidence that there is bias in $Y$ since $Z$ is only weakly related to $Y$. The evidence against bias in $Y$ is however relatively weak since there may be bias from other unobserved variables related to $Y$. This is the least favorable situation.

In the next section we consider specific choices of $g(\hat{\rho})$ based on a pattern-mixture model analysis that reflects this hierarchy.

## 4.3 The Pattern-Mixture Model

Let $M$ denote the missingness indicator, such that $M = 0$ if $Y$ is observed and $M = 1$ if $Y$ is missing. We assume $E(Y|Z, M = 0) = \alpha_0 + \alpha Z$, and let $X = \alpha Z$. For simplicity we assume in this section that $\alpha$ is known, that is, we ignore estimation error in $\alpha$. We focus on the joint distribution of $[Y, X, M]$ which we assume follows the bivariate pattern-mixture model discussed in Little (1994). This model can be written as follows:

$$(4.3) \qquad \begin{gathered} (Y, X|M = m) \sim N_2\left((\mu_y^{(m)}, \mu_x^{(m)}), \Sigma^{(m)}\right) \\ M \sim Bernoulli(1-\pi) \\ \Sigma^{(m)} = \begin{bmatrix} \sigma_{yy}^{(m)} & \rho^{(m)}\sqrt{\sigma_{yy}^{(m)}\sigma_{xx}^{(m)}} \\ \rho^{(m)}\sqrt{\sigma_{yy}^{(m)}\sigma_{xx}^{(m)}} & \sigma_{xx}^{(m)} \end{bmatrix} \end{gathered}$$

where $N_2$ denotes the bivariate normal distribution. Of primary interest is the marginal mean of $Y$, which can be expressed as $\mu_y = \pi\mu_y^{(0)} + (1-\pi)\mu_y^{(1)}$. This model is underidentified, since there is no information on the conditional normal distribution for $Y$ given $X$ for nonrespondents ($M = 1$). However, Little (1994) shows that the model can be identified by making assumptions about how missingness of $Y$ depends on $Y$ and $X$. Specifically if we assume that

$$(4.4) \qquad \Pr(M = 1|Y, X) = f(X + \lambda^* Y),$$

for some unspecified function $f$ and known constant $\lambda^*$, the parameters are just identified by the condition that

$$(4.5) \qquad ((Y, X) \perp M | f(X + \lambda^* Y))$$

where $\perp$ denotes independence. The resulting ML estimate of the mean of $Y$ averaging over patterns is

$$\hat{\mu}_y = \bar{y}_R + \frac{s_{xy} + \lambda^* s_{yy}}{s_{xx} + \lambda^* s_{xy}} (\bar{x} - \bar{x}_R), \tag{4.6}$$

where $s_{xx}, s_{xy}$ and $s_{yy}$ are the sample variance of $X$, the sample covariance of $X$ and $Y$, and the sample variance of $Y$ for respondents (Little (1994)).

We apply a slight modification of this model in our setting, rescaling the proxy variable $X$ to have the same variance as $Y$, since we feel this enhances the interpretability of the model (4.4) for the mechanism. Specifically we replace (4.4) by

$$\Pr(M = 1|Y, X) = f(X\sqrt{\frac{\sigma_{yy}^{(0)}}{\sigma_{xx}^{(0)}}} + \lambda Y) = f(X^* + \lambda Y), \tag{4.7}$$

where $X^*$ is the proxy variable $X$ scaled to have the same variance as $Y$ in the respondent population, and $\lambda = \lambda^* \sqrt{\sigma_{xx}^{(0)}/\sigma_{yy}^{(0)}}$. The parameters are just identified by the condition that

$$((Y, X) \perp M | f(X^* + \lambda Y)) \tag{4.8}$$

where $\perp$ denotes independence. We call the model defined by (4.3) and (4.7) a proxy pattern-mixture (PPM) model. By a slight modification of the arguments in (Little, 1994), the resulting maximum likelihood estimate of the overall mean of $Y$ has the form of (4.1) where

$$g(\hat{\rho}) = \frac{\lambda + \hat{\rho}}{\lambda \hat{\rho} + 1}, \tag{4.9}$$

and $\hat{\rho}$ is the respondent sample correlation. Note that regardless of $\lambda$, $g(\hat{\rho}) \rightarrow 1$ as $\hat{\rho} \rightarrow 1$, so this choice of $g$ satisfies the desirable property previously described.

### 4.3.1 Properties of the Missing Data Mechanism

There are limitless ways to model deviations from MAR, and any method needs to make assumptions. Thus, the assumption about the missing data mechanism, given by (4.7), is a key to the proposed method, and deserves careful consideration. The assumption (4.7) is quite flexible and covers a wide range of models relating $X$ ($X^*$) and $Y$ to $M$. In particular, it is more flexible than the well-known Heckman selection model (Heckman, 1976), which assumes that missingness is linear in $X$ and $Y$. For example, the PPM model encompasses not only mechanisms that are linear in $X$ or linear in $Y$, but also ones that are quadratic in $X$ or quadratic in $Y$. A broad class of mechanisms are those that depend on both $X$ and $Y$, potentially including quadratic terms and the interaction of $X$ and $Y$, that is

$$\text{logit}(\Pr(M=1|Y,X)) = \gamma_0 + \gamma_1 X + \gamma_2 X^2 + \gamma_3 Y + \gamma_4 Y^2 + \gamma_5 XY \tag{4.10}$$

If we take $f(\cdot)$ in (4.7) to be quadratic, we obtain a missingness mechanism for the PPM that is a specific subset of this general model. The PPM missing data mechanism is

$$\begin{aligned}\text{logit}(\Pr(M=1|Y,X)) &= \alpha_0 + \alpha_1(X+\lambda Y) + \alpha_2(X+\lambda Y)^2 \\ &= \alpha_0 + \alpha_1 X + \alpha_1\lambda Y + \alpha_2 X^2 + \alpha_2\lambda^2 Y^2 + 2\alpha_2 XY \qquad (4.11)\end{aligned}$$

Assuming this more general model, the ability of the PPM to produce an unbiased estimate depends on whether there is a value of $\lambda$ that makes (4.11) close to the true mechanism (4.10). In particular, we note that the sign of the $X$ and $Y$ terms (and similarly the $X^2$ and $Y^2$ terms) must be the same; however since $X$ is a proxy for $Y$ we feel that this assumption is not unreasonable.

An alternative method when data may be not missing at random is to specify a selection model, which factors the joint distribution of $Y$ and $M$ given $X$ into the conditional distribution of $M$ given $Y$ and $X$ as in (4.10) and the marginal distribution of $Y$ given $X$ (e.g. Heckman, 1976; Diggle and Kenward, 1994; Little and Rubin, 2002). The approach requires full specification of the distribution of $M$

given $Y$ and $X$. In contrast, our pattern-mixture model avoids the need to specify the function $f$ that relates missingness of $Y$ to $X^* + \lambda Y$, although it shares with the corresponding selection model the assumption that missingness depends on $Y$ and $Z$ only through the value of $X^* + \lambda Y$. One reason for restricting the dependence on the set of variables $Z$ to the combination $X^*$ is that, under the normality assumption, dependence on missingness of $Y$ on other combinations (say $U = \delta Z$) does not result in bias in the mean of $Y$, since $Y$ is conditionally independent of $U$ given $X^*$. Reduction to $X^*$ limits the analysis to just one sensitivity parameter ($\lambda$) and so is much simpler than an analysis that models departures from MAR for each of the individual $Z$'s. Another advantage of our model is that likelihood-based analysis is much simpler than selection models, which require iterative algorithms.

### 4.3.2 Other Properties of the Model

Suppose $\lambda$ is assumed to be positive, which seems reasonable given that $X$ is a proxy for $Y$. Then as $\lambda$ varies between 0 (missingness depends only on $X$) and infinity (missingness depends only on $Y$), $g(\hat{\rho})$ varies between $\hat{\rho}$ and $1/\hat{\rho}$. This result is intuitively very appealing. When $\lambda = 0$ the data are MAR, since in this case missingness depends only on the observed variable $X$. In this case $g(\hat{\rho}) = \hat{\rho}$, and (4.1) reduces to the standard regression estimator described above. In this case the bias adjustment for $Y$ increases with $\hat{\rho}$, as the association between $Y$ and the variable determining the missing data mechanism increases. On the other hand when $\lambda = \infty$ and missingness depends only on the true value of $Y$, $g(\hat{\rho}) = 1/\hat{\rho}$ and (4.1) yields the inverse regression estimator proposed by Brown (1990). The bias adjustment thus decreases with $\hat{\rho}$, reflecting the fact that in this case the bias in $Y$ is attenuated in the proxy, with the degree of attenuation increasing with $\hat{\rho}$.

### 4.3.3 Sensitivity Analysis

There is no information in the data to inform the choice of $\lambda$ . Little (1994) proposes a sensitivity analysis, where the estimate defined by (4.1) and (4.9) are considered for a range of values of $\lambda$ between 0 and infinity; the latter is the most extreme deviation from MAR, and estimates for this case have the highest variance. Indeed for small $\hat{\rho}$, the estimate with $\lambda$ set to infinity is very unstable, and it is undefined when $\hat{\rho} = 0$. We suggest a sensitivity analysis using $\lambda = (0, 1, \infty)$ to capture a range of missingness mechanisms. In addition to the extremes, we use the intermediate case of $\lambda = 1$ that weights the proxy and true value of $Y$ equally because the resulting estimator has a particularly convenient and simple interpretation. In this case $g(\hat{\rho}) = 1$ regardless of the value of $\hat{\rho}$, implying that the standardized bias in $\bar{y}_R$ is the same as the standardized bias in $\bar{x}_R$. In general, the stronger the proxy, the closer the value of $\hat{\rho}$ to one, and the smaller the differences between the three estimates.

## 4.4 Estimation Methods

### 4.4.1 Maximum Likelihood

The estimator described by (4.1) and (4.9) is maximum likelihood (ML) for the pattern-mixture model. Large-sample variances are given by Taylor series calculations as in Little (1994) (details given in Appendix, Section 4.9.1), though this approximation may not be appropriate for small samples. Additionally, the ML estimate and corresponding inference does not take into account the fact that the regression coefficients that determine $X$ are subject to sampling error. Better methods incorporate this uncertainty, such as the Bayesian methods described below.

### 4.4.2 Bayesian Inference

An alternative to ML is Bayesian inference, which allows us to incorporate the uncertainty in $X$ and which may perform better in small samples. Let $M$ denote the missingness indicator, and let $\alpha$ be a the vector of regression parameters from the regression of $Y$ given $Z$ that creates the proxy (i.e. $X = \alpha Z$). Let $Z \longrightarrow (X, V)$ be a (1-1) tranformation of the covariates. Letting [] denote distributions, we factor the joint distribution of $Y$, $X, V$, $M$, and $\alpha$ as follows:

$$[Y, X, V, M, \alpha] = [Y, X|M, \alpha][M][\alpha][V|Y, X, M, \alpha] \tag{4.12}$$

We leave the last distribution for $V$ unspecified, and assume in (4.12) that $M$ is independent of $\alpha$. We assume the standard linear regression model creates the proxy $X$; the $Y_i$ are independent normal random variables with mean $X = Z\alpha$ and variance $\phi^2$. We place non-informative priors on all parameters and draw from their posterior distributions. For each draw of the parameters we recalculate the proxy using the draws of $\alpha$ and then scale using the draw of $\sigma_{xx}^{(0)}$ and $\sigma_{yy}^{(0)}$. Throughout the remainder of this and the following section we take $X$ to denote this scaled version of the proxy.

Draws from the posterior distribution are obtained using different algorithms for the cases with $\lambda = 0$ and $\lambda = \infty$, as detailed below. In the case of intermediate values of $\lambda$ the algorithm for $\lambda = \infty$ is applied to obtain draws from the joint distribution of $(X, X + \lambda Y)$ and then these draws are transformed to obtain the parameters of the joint distribution of $(X, Y)$ (details in Appendix, Section 4.9.2). In the equations that follow, let $s_{jj}$ be the sample variance of $j$, $b_{jk.k}$ and $s_{jj.k}$ be the regression coefficient of $k$ and the residual variance from the regression of $j$ on $k$, and (0) and (1) denote quantities obtained from respondents and nonrespondents, respectively. The sample size is $n$ with $r$ respondents, and $p$ is the number of covariates $Z$ that create the proxy.

First we consider the model with $\lambda = 0$. This implies that missingness depends only on $X$, so the distribution of $Y$ given $X$ is the same for respondents and nonrespondents. Thus the intercept and regression coefficient, $\beta_{y0.x}^{(m)}$ and $\beta_{yx.x}^{(m)}$, are the same for $M = 0$ and $M = 1$. Draws of the identifiable parameters are computed in the following sequence:

1. $1/\phi^2 \sim \chi^2_{(r-p-1)}/((r-p-1)s^{(0)}_{yy.z})$
2. $\alpha \sim N(\hat{\alpha}, \phi^2(Z^TZ)^{-1})$
3. $\pi \sim Beta(r+0.5, n-r+0.5)$
4. $1/\sigma^{(0)}_{xx} \sim \chi^2_{(r-1)}/(rs^{(0)}_{xx})$
5. $\mu^{(0)}_x \sim N(\bar{x}_R, \sigma^{(0)}_{xx}/r)$
6. $1/\sigma^{(0)}_{yy.x} \sim \chi^2_{(r-2)}/(rs^{(0)}_{yy.x})$
7. $\beta^{(0)}_{yx.x} \sim N\left(b_{yx.x}, \dfrac{\sigma^{(0)}_{yy.x}}{rs^{(0)}_{xx}}\right)$
8. $\beta^{(0)}_{y0.x} \sim N(\bar{y}_R - \beta^{(0)}_{yx.x}\bar{x}_R, \sigma^{(0)}_{yy.x}/r)$
9. $1/\sigma^{(1)}_{xx} \sim \chi^2_{(n-r-1)}/((n-r)s^{(1)}_{xx})$
10. $\mu^{(1)}_x \sim N(\bar{x}_{NR}, \sigma^{(1)}_{xx}/(n-r))$

Draws from the posterior distribution of $\mu_y$ are obtained by substituting these draws into $\mu_y = \beta^{(0)}_{y0.x} + \beta^{(0)}_{yx.x}\mu_x$ where $\mu_x = \pi\mu^{(0)}_x + (1-\pi)\mu^{(1)}_x$.

When $\lambda = \infty$, the resulting assumption is that missingness depends only on $Y$, so the distribution of $X$ given $Y$ is the same for respondents and nonrespondents, i.e. $\beta^{(m)}_{x0.y}$ and $\beta^{(m)}_{xy.y}$ are the same for $M = 0$ and $M = 1$. Draws are obtained in a similar fashion as before. Steps 1 through 3 remain the same, but steps 4 through 10 are replaced by the following:

4. $1/\sigma_{yy}^{(0)} \sim \chi^2_{(r-1)}/(rs_{yy}^{(0)})$
5. $\mu_y^{(0)} \sim N(\bar{y}_R, \sigma_{yy}^{(0)}/r)$
6. $1/\sigma_{xx.y}^{(0)} \sim \chi^2_{(r-2)}/(rs_{xx.y}^{(0)})$
7. $1/\sigma_{xx}^{(1)} \sim \chi^2_{(n-r-1)}/((n-r)s_{xx}^{(1)})$
8. $\beta_{xy.y}^{(0)} \sim N\left(b_{xy.y}, \dfrac{\sigma_{xx.y}^{(0)}}{rs_{yy}^{(0)}}\right)$
9. $\beta_{x0.y}^{(0)} \sim N(\bar{x}_R - \beta_{xy.y}^{(0)}\bar{y}_R, \sigma_{xx.y}^{(0)}/r)$
10. $\mu_x^{(1)} \sim N(\bar{x}_{NR}, \sigma_{xx}^{(1)}/(n-r))$

To satisfy parameter constraints, the drawn value of $\sigma_{xx}^{(1)}$ from step 7 must be larger than the drawn value of $\sigma_{xx.y}^{(0)}$ from step 6; if this is not the case then these draws are discarded and these steps repeated. Draws from the posterior distribution of $\mu_y$ are obtained by substituting these draws into

$$\mu_y = \pi\mu_y^{(0)} + (1-\pi)\frac{\mu_x^{(1)} - \beta_{x0.y}^{(0)}}{\beta_{xy.y}^{(0)}}.$$

### 4.4.3 Multiple Imputation

An alternative method of inference for the mean of $Y$ is multiple imputation (Rubin, 1978). We create $K$ complete data sets by filling in missing $Y$ values with draws from the posterior distribution, based on the pattern-mixture model. Draws from the posterior distribution of of $Y$ are obtained by first drawing the parameters from their posterior distributions as outlined in Section 4.4.2, dependent on the assumption about $\lambda$, and then drawing the missing values of $Y$ based on the conditional distribution of $Y$ given $X$ for nonrespondents ($M = 1$),

$$(4.13)\qquad [y_i|x_i, m_i = 1, \phi_{(k)}] \sim N\left(\mu^{(1)}_{y(k)} + \frac{\sigma^{(1)}_{yx(k)}}{\sigma^{(1)}_{xx(k)}}\left(x_i - \mu^{(1)}_{x(k)}\right), \sigma^{(1)}_{yy(k)} - \frac{{\sigma^{(1)}_{yx(k)}}^{2}}{\sigma^{(1)}_{xx(k)}}\right)$$

where the subscript $(k)$ denotes the $k$th draws of the parameters. For the $k$th completed data set, the estimate of $\mu_y$ is the sample mean $\bar{Y}_k$ with estimated variance $W_k$. A consistent estimate of $\mu_y$ is then given by $\hat{\mu}_y = \frac{1}{K}\sum_{k=1}^{K}\bar{Y}_k$ with $\text{Var}(\hat{\mu}_y) = \bar{W}_K + \frac{K+1}{K}B_K$, where $\bar{W}_K = \frac{1}{K}\sum_{k=1}^{K}W_k$ is the within-imputation variance and $B = \frac{1}{K-1}\sum_{k=1}^{K}(\bar{Y}_k - \hat{\mu}_y)^2$ is the between-imputation variance.

An advantage of the multiple imputation approach is the ease with which complex design features like clustering, stratification and unequal sampling probabilities can be incorporated. Once the imputation process has created complete data sets, design-based methods can be used to estimate $\mu_y$ and its variance; for example the Horvitz-Thompson estimator can be used to calculate $\bar{Y}_k$. Incorporating complex design features into the model and applying maximum likelihood or Bayesian methods is less straightforward, though arguably more principled. See for example Little (2004) for more discussion.

## 4.5 Quantifying Nonresponse Bias

We propose using the estimated fraction of missing information (FMI), obtained through multiple imputation under the PPM model with different nonresponse mechanism assumptions, as a measure of nonresponse bias. The FMI due to nonresponse is estimated by the ratio of between-imputation to total variance under multiple imputation (Little and Rubin, 2002). Traditionally one applies this under the assumption that data are missing at random, but we propose its use under the pattern-mixture model where missingness is not at random. FMI is influenced by both the strength of the proxy ($\rho$) and the size of the deviation from MCAR ($d$). For the purposes of illustration we use the standardized deviation $d^*$ so it is the same regardless of whether $X$ has been scaled.

Figure 4.1 is a plot of simulated data showing FMI as a function of $\rho$ for different values of $d^*$ and the response rate. Separate estimates of FMI are obtained for different nonresponse assumptions ($\lambda = 0, 1, \infty$). For all nonresponse mechanisms, as the strength of the proxy ($\rho$) increases, the FMI decreases, eventually reaching zero for a perfect proxy ($\rho = 1$). Across all values of $\rho$ and $d^*$ FMI is smallest when assuming $\lambda = 0$, largest when assuming $\lambda = \infty$, and falls in between when $\lambda = 1$.

When missingness is at random ($\lambda = 0$) and $d^*$ is small, the FMI is approximately equal to the nonresponse rate for $\rho = 0$ and decreases as the strength of the proxy increases. For all values of $\lambda$, larger $d$ leads to elevated FMI, though these differences are relatively small compared to the effect of $\rho$. The FMI is larger for lower response rates across all values of $\rho$, though differences are more severe with a strong proxy than with a weak one.

With NMAR mechanisms, the FMI is greatly inflated above the response rate for weak proxies, but rapidly declines to levels similar to those of the MAR assumption. The relative gains from a moderately correlated proxy are larger for NMAR mechanisms than for the MAR mechanism. For example, for small $d$ and 50% missingness the gain from moving from $\rho = 0$ to $\rho = 0.5$ is a decrease in FMI from 50% to 43% when $\lambda = 0$ but from nearly 100% to 75% when $\lambda = \infty$. Clearly the presence of strong predictors is of the utmost importance in identifying and removing nonresponse bias; the sensitivity of FMI to $\rho$ illustrates this.

## 4.6 Simulation Studies

We now describe a set of simulation studies designed to (1) illustrate the effects of $\rho$, $d^*$, and sample size on PPM estimates of the mean of $Y$, (2) assess confidence coverage of ML, Bayes and MI inferences, and (3) demonstrate the performance of the PPM model when data arise from a selection model with a range of nonresponse mechanisms. All simulations and data analysis were performed using the software package R (R Development Core Team, 2007).

### 4.6.1 Numerical Illustration of PPMA

Our first objective with the simulation studies was to numerically illustrate the taxonomy of evidence concerning bias based on the strength of the proxy and the deviation of its mean. We created a total of eighteen artificial data sets in a 3x3x2 factorial design. A single data set was generated for each combination of $\rho = \{0.8, 0.5, 0.2\}$, $d^* = \{0.1, 0.3, 0.5\}$ and $n = \{100, 400\}$ as follows. A single covariate $Z$ was generated for both respondents and nonrespondents with the outcome $Y$ generated only for respondents. Respondent data were created as pairs $(z_i, y_i), i = 1 \ldots r$ with $z_i \sim N(0, \rho^2)$ and $y_i = 1 + z_i + e_i$, where $e_i \sim N(0, 1 - \rho^2)$. Nonrespondent data were $Z's$ only, generated from $z_i \sim N(2\rho d^*, \rho^2)$ for $i = r+1 \ldots n$. The nonresponse rate was fixed at 50%. This data structure was chosen so that the variance of the complete case mean would be constant (and equal to one) across different choices of $\rho$ and $d^*$, and so that varying $\rho$ would not affect $d^*$ and vice-versa. $R^2$ values that corresponded to the selected $\rho$ were 64%, 25%, and 4%, covering a range likely to be encountered in practice.

For each of the eighteen data sets, estimates of the mean of $Y$ and its precision were obtained for $\lambda = (0, 1, \infty)$. For each value of $\lambda$, three 95% intervals were calculated:

(a) ML: the maximum likelihood estimate ± 2 standard errors (large-sample approximation)

(b) PD: the posterior median and 2.5th to 97.5th posterior interval based on 5000 draws from the posterior distribution of $\mu_Y$ as outlined in Section 4.4.2

(c) MI: mean ± 2 standard errors from 20 multiply imputed data sets.

Posterior median and quantiles were used because initial evaluations showed that the posterior distribution of $\mu_Y$ was skewed and had extreme outliers for small $\rho$ and large $\lambda$. The complete case estimate (± 2 standard errors) was also computed for each data set; note that the expected value of the respondent mean and corresponding confidence interval is constant across all values of $\rho$ and $d$ for each $n$.

**Results**

Results from applying the three estimation methods to each of the nine data sets with $n = 100$ are displayed in Figure 4.2. The complete case estimate is shown alongside 95% intervals estimated by maximum likelihood, multiple imputation, and the posterior distribution, for $\lambda = (0, 1, \infty)$. For each population the PD intervals are longer than the ML and MI intervals for all choices of $\lambda$, especially for weak proxies and $\lambda = \infty$. Results for $n = 400$ were similar and are not shown.

Populations with a strong proxy ($\rho = 0.8$) do not show much variation across values of $\lambda$; there is evidence that nonresponse bias is small for small $d$ and there is good information to correct the potential bias for larger values of $d$. For moderately strong proxies ($\rho = 0.5$) the intervals increase in length, with differences between PD and ML becoming more exaggerated as $d$ increases. As expected, when the proxy is weak ($\rho = 0.2$) we see large intervals for models that assume missingness is not at random ($\lambda \neq 0$); this reflects the fact that we are in the worst-case scenario where there is not much information in the proxy to estimate the nonresponse bias. Notice that in this simulation the true mean of $Y$ is not known; we simply illustrate the effect of various values of $\rho$ and $\lambda$ on the sensitivity analysis.

### 4.6.2 Confidence Coverage

The second objective of the simulation was to assess coverage properties for each of the three estimation methods. We generated 500 replicate data sets as before for each of the eighteen population designs and computed the actual coverage of a nominal 95% interval and median interval length. The Bayesian intervals were based on 1000 draws from the posterior distribution. Coverage is based on the unreasonable assumption that the assumed value of $\lambda$ equals the actual value of $\lambda$. This is unrealistic, but coverages are clearly not valid when the value of $\lambda$ is misspecified, and uncertainty in the the choice of $\lambda$ is captured by the sensitivity analysis.

#### Results

Table 4.1 displays the nominal coverage and median CI width for each of the eighteen populations. For populations with a strong or moderately strong proxy ($\rho = 0.8, 0.5$) coverage is at or above nominal levels for all three methods, for both the smaller and larger sample sizes and for all levels of $d$. For these populations, PD inference is slightly more conservative; intervals are larger than ML for most populations. However, when the proxy is weak, ML coverage is below nominal levels for larger values of $\lambda$, while both PD and MI have coverage close to nominal levels. Wiwth small sample size and weak proxies, taking $\lambda = \infty$ leads to large confidence intervals, since draws of $\beta_{xy.y}$ approach zero. The $\lambda = \infty$ model requires a strong proxy or large sample size to provide reliable estimates of $\mu_y$.

### 4.6.3 Missing Data Mechanisms under a Selection Model

In our final simulation we generated complete data under a selection model framework, induced missingness according to a range of missing data mechanisms, and applied the PPM sensitivity analysis to evaluate its coverage. The selection model factorization implies marginal normality, while the PPM assumes conditional normality, so in this simulation the distributional assumptions of the PPM are violated. Simulated data were pairs $(z_i, y_i)$ for $i = 1 \ldots n$ from a bivariate normal distribution such that $EZ = EY = 1$, $Var(Z) = Var(Y) = 1$, and $Cov(Z, Y) = \rho$. The missing data indicator $M$ was generated according to a logistic model,

$$\text{logit}(\Pr(M = 1|Y, Z)) = \gamma_0 + \gamma_Z Z + \gamma_{Z2} Z^2 + \gamma_Y Y + \gamma_{Y2} Y^2,$$

for eight choices of $\gamma = \{\gamma_0, \gamma_Z, \gamma_{Z2}, \gamma_Y, \gamma_{Y2}\}$ chosen to reflect different nonresponse mechanisms, including both MAR and NMAR scenarios. The choices of $\gamma$ are displayed in Table 4.2, and are labeled using conventional linear model notation. These models for $M$ led to approximately 50% missingness in populations where the missing data mechanism was linear in $Z$ and $Y$, and a slightly lower proportion of missing values in the populations that were quadratic in $Z$ and/or $Y$. We note that unlike the previous simulations, $\rho$ is specified as the correlation between $Y$ and the covariate $Z$ in the entire sample, not the respondents only. For populations where nonresponse is linear in $Z$ and/or $Y$, the induced correlation between $Y$ and the proxy $X$ is the same for both respondents and nonrespondents and is equal to $\rho$. However, when missingness is quadratic in $Z$ and/or $Y$, the correlation between $Y$ and the proxy is attenuated in the respondents and stronger in the nonrespondents.

There were two different sample sizes, $n = \{100, 400\}$, and three different correlations, $\rho = \{0.8, 0.5, 0.2\}$. We generated 500 replicate data sets for each of the two sample sizes, three correlation levels, and eight nonresponse mechanisms and applied our PPM sensitivity analysis with $\lambda = 0, 1, \infty$ to estimate the mean of $Y$. As before we calculated three 95% intervals for the mean of $Y$ (ML, PD, and MI) and computed the actual coverage and length of a nominal 95% interval, noting that $\mu_Y = 1$ for all populations. Bayesian intervals were based on 1000 draws from the posterior distribution. We also calculated the coverage of the sensitivity analysis as a whole, that is, the percent of the replicates where at least one of the three intervals ($\lambda = 0, 1, \infty$) covered the population mean.

**Results**

Results from the 24 populations with $n = 400$ are shown in Figures 4.3a–c; coverage was higher for the smaller sample size since confidence intervals were wider for all values of $\lambda$ and is not shown. There were four nonresponse mechanisms where, aside from distributional assumptions, there was a value of $\lambda$ that corresponded to the true missingness mechanism: $\lambda = 0$ for mechanisms that depended only on $Z$ ($[Z]$ and $[Z^2]$) and $\lambda = \infty$ for mechanisms that depended only on $Y$ ($[Y]$ and $[Y^2]$). For these populations, coverage was approximately at nominal levels for the corresponding value of $\lambda$ for all estimation methods and for all three levels of correlation $\rho$.

The remaining four nonresponse mechanisms had missingness dependent on some combination of $Z$ and $Y$, the toughest situation for the PPM. For missingness mechanisms $[Z+Y]$ and $[Z^2+Y^2]$ there is in theory a value of $\lambda$ that yields the corresponding PPM, however it might not be one of the three in the sensitivity analysis. In these situations the PPM performed well, with at or near nominal coverage for one of the $\lambda$ values for all three levels of $\rho$, and at least one interval covering the truth in almost 100% of the replicates. The final two missingness mechanisms, $[Z^2 + Y]$ and $[Z + Y^2]$, do not correspond to any value of $\lambda$; these are situations where the PPM is likely to show poor performance. In fact, no method reached nominal coverage levels, except for the weak proxy ($\rho = 0.2$) where confidence interval lengths were extremely large for $\lambda = \infty$ (note the increase in range of the plot). However, as a whole the sensitivity analysis performed better for these populations in that at least one interval covered the truth at closer to nominal levels (at or above nominal levels for ML and MI, at worst 83% for PD, results not shown).

As expected, confidence interval lengths were larger for PD and MI than for ML, particularly for the weaker proxies. However, this did not always lead to improved coverage. By construction the confidence intervals for PD were not symmetrical, and for $\lambda = 1$ and $\lambda = 0$ they were heavily skewed due to draws of $\beta_{xy.y}$ that approached zero. When the point estimates were biased (for example, for $[Z + Y^2]$ and $\lambda = \infty$), the skewness tended to lead to undercoverage for PD, while the symmetric intervals of ML and MI had higher coverage. These differences were exaggerated in the weaker proxies, where better coverage was driven by large confidence interval widths, not by unbiased point estimates.

Overall, the PPM sensitivity analysis performed well in a setting where it was not the "correct" model. This final simulation demonstrated the flexibility of the method, as it had good coverage for a wide range of nonresponse mechanisms, including both linear and quadratic functions of the covariate and the outcome.

## 4.7 Application

The third National Health and Nutrition Examination Survey (NHANES III) was a large-scale stratified multistage probability sample of the noninstitutionalized U.S.

population conducted during the period from 1988 to 1994 (U.S. Department of Health and Human Services, 1994). NHANES III collected data in three phases: (a) a household screening interview, (b) a personal home interview, and (c) a physical examination at a mobile examination center (MEC). The total number of persons screened was 39,695, with 86% (33,994) completing the second phase interview. Of these, only 78% were examined in the MEC. Since the questions asked at both the second and third stage varied considerably by age we chose to select only adults age 17 and older who had completed the second phase interview for the purposes of our example, leaving a sample size of 20,050. We chose to focus on estimating nonresponse bias for three body measurements at the MEC exam: systolic blood pressure (SBP), diastolic blood pressure (DBP), and body mass index (BMI). The nonresponse rates for these three items was 15%, 15%, and 10% respectively. It has been suggested that nonresponse in health surveys may be related to health (Cohen and Duffy, 2002), hence these measures may potentially be missing not at random.

In order to reflect nonresponse due to unit nonresponse at the level of the MEC exam we chose to only include fully observed covariates to create the proxies; variables that were fully observed for the sample included age, gender, race, and household size. The design weight was also used as a covariate in creating the proxies. Linear regression was used to create the proxies, with the final models chosen with backwards selection starting from a model that contained second-order interactions. To better approximate a normal distribution, BMI values were log-transformed. Systolic blood pressure displayed both the largest correlation between outcome and the proxy and the largest deviation in the proxy, with $\hat{\rho} = 0.6$, $d = 1.0$ and $d^* = 0.08$. Diastolic blood pressure had $\hat{\rho} = 0.33$, $d = 0.05$, and $d^* = 0.01$, while log(BMI) had the weakest proxy at $\hat{\rho} = 0.24$ and essentially no deviation with $d = -0.0004$ and $d^* = -0.008$.

For each outcome, estimates of the mean and confidence intervals for $\lambda = (0, 1, \infty)$ were obtained using maximum likelihood (ML), 5000 draws from the posterior distribution (PD), and multiple imputation with $K = 20$ data sets (MI). Additionally, since NHANES III has a complex survey design we obtained estimates using multiple imputation with design-based estimators of the mean using the survey weights (MI wt). Design-based estimators were computed using the "survey" routines in R, which estimate variances using Taylor series linearizations (Lumley, 2004).

Mean estimates and confidence intervals are displayed in Figures 4.4, 4.5, and 4.6. The three methods, ML, PD, and MI, produce similar estimates and confidence intervals across all three outcomes and all values of $\lambda$. The intervals for weighted MI are larger than those for either of the non-design-adjusted methods, and for both SBP and BMI there is also a shift in the mean estimates for the weighted estimators, consistent for all values of $\lambda$, reflecting the impact on these outcomes of the oversampling in NHANES of certain age and ethnic groups. The choice of $\lambda$ has a larger impact on the mean estimate for the SBP and DBP measurements; assuming MAR would result in significantly different mean estimates than assuming NMAR. BMI has a weak proxy and a small deviation so there is some evidence against nonresponse bias (small $d$) but this evidence is weak (small $\rho$).

Table 4.3 shows the estimates of FMI for each outcome under each missingness mechanism for multiple imputation analyses that ignore design weights (FMI) and incorporate them (FMIwt). The weighted estimates of FMI are considerably smaller than the unweighted estimates; the same between-imputation variance is coupled with increased within-imputation variability due to incorporation of the sample design. Larger values of $\lambda$ result in larger estimates of FMI. When the proxy is strong, as with SBP, FMI remains relatively low even when assuming NMAR. For BMI which has a weak proxy but also essentially zero deviation, if one is willing to assume $\lambda = 0$ then the FMI is low and close to the nonresponse rate, as there appears to be little deviation from MCAR. However, since the proxy is weak, as soon as one assumes NMAR the estimates of FMI become drastically larger, as high as 83%.

### 4.8 Discussion

The PPM analysis of nonresponse bias we propose has the following attractive features: it integrates all the various components of nonresponse noted in the introduction into a single sensitivity analysis. It is the only analysis we know of that formally reflects the hierarchy of evidence about bias in the mean suggested in the introduction, which we believe is realistic. It is easy to implement, since the ML form is simple to compute, and the Bayesian simulation is noniterative, not requiring iterative Markov Chain Monte Carlo methods that pervade more complex Bayesian methods and might deter survey practitioners; the MI method is also non-iterative, and allows complex design features to be incorporated in the within-imputation component of variance. PPM analysis includes but does not assume MAR, and it provides a picture of the potential nonresponse bias under a reasonable range of MAR and non-MAR mechanisms. It gives appropriate credit to the existence of good predictors of the observed outcomes. When data are MAR, it is the squared correlation between the covariates and the outcome that drives the reduction in variance, which means that covariates with rather high correlations are needed to have much impact. An interesting implication of our PPM analysis is that if the data are not MAR, covariates with moderate values of correlation, such as 0.5, can be useful in reducing the sensitivity to assumptions about the missing data mechanism. We suggest that emphasis at the design stage should be on collection of strong auxiliary data to help evaluate and adjust for potential nonresponse, not solely on obtaining the highest possible response rate.

The PPM analysis employs a sensitivity analysis to assess deviations from MAR, in contrast with some selection model approaches that attempt to use the data to estimate parameters that capture deviations from MAR (e.g. Heckman, 1976; Diggle and Kenward, 1994). These models are technically identified in situations where pattern-mixture models are not, but estimation of the NMAR parameters is still based on strong and unverifiable structural and distributional assumptions, and these assumptions are more transparent in the pattern-mixture factorization, since differences between respondents and nonrespondents are directly modeled (Little and Rubin, 2002). The sensitivity analysis for PPM analysis only varies one sensitivity parameter, $\lambda$, but still manages to capture a range of assumptions on the missing data mechanism. Both the standard and reverse regression estimators are contained in the PPM framework, which are familiar to survey practitioners.

A drawback of the PPM analysis is that by reducing the auxiliary data to the single proxy $X^*$, the coefficient $\lambda$ is not associated with any particular covariate and hence is difficult to interpret, since the effects on missingness on individual covariates $Z_j$ are lost. The pattern-mixture model proposed by Daniels and Hogan (2000) in the context of longitudinal data, uses a location-scale parameterization to model differences in the marginal distribution of $(Y, Z)$ for respondents and nonrespondents. This model is more readily interpretable than our approach, but it is very underidentified, even with a single $Z$ it has three unidentified parameters, and additional specification is needed to limit the number of parameters to be varied in a sensitivity analysis. Modeling the conditional distribution of $Y$ given $Z$ for respondents and nonrespondents, as in PPM analysis, focuses more directly on the distribution that is not identified, namely the distribution of $Y$ given $Z$ for nonrespondents. A reasonable alternative to the PPM model allows the intercept of this regression to differ for respondents and nonrespondents but the regression coefficients and residual variance to be the same. This results in a simple nonignorable model with just one sensitivity parameter, the difference in intercepts. However, it is hard to assess how much of a difference in intercepts is plausible, and this model does not readily distinguish between strong and weak proxies of $Y$. Allowing the regression coefficients of individual $Z_j$'s in this model to differ for respondents and nonrespondents provides more flexibility, at the expense of adding more unidentified parameters, particularly when there is more than one covariate. Our approach trades off interpretability for parsimony, allowing a single parameter to model deviations from MAR.

Another limitation of our analysis is that it focuses only on the mean of a particular outcome $Y$, so it is outcome-specific. Thus, in a typical survey with many outcomes, the analysis needs to be repeated on each of the key outcomes of interest and then integrated in some way that reflects the relative importance of these outcomes. This makes life complicated, but that seems to us inevitable. An unavoidable feature of the problem is that nonresponse bias is small for variables unrelated to nonresponse, and potentially larger for variables related to nonresponse. Measures that do not incorporate relationships with outcomes, like the variance of the nonresponse weights, cannot capture this dimension of the problem. Presenting the fraction of missing information over a range of key survey variables and a range of values of $\lambda$ seems valuable for capturing the full scope of the potential nonresponse bias.

The pattern-mixture model that justifies the proposed analysis strictly only applies to continuous survey variables, where normality is reasonable, although we feel it is still informative when applied to non-normal outcomes. Extensions to categorical variables appear possible via probit models, and many other extensions can be envisaged, including extensions to other generalized linear models. PPM analysis can be applied to handle item nonresponse by treating each item subject to missing data separately, and restricting the covariates to variables that are fully observed. However, this approach does not condition fully on the observed information, and extensions for general patterns of missing data would be preferable. Our future work on PPM analysis will focus on developing these extensions.

## 4.9 Appendix

### 4.9.1 Large-Sample Variance of the MLE

We want to find the large-sample variance of the maximum likelihood estimate of $\mu_y$, given by

$$\hat{\mu}_y = \bar{y}_R + \sqrt{\frac{s_{yy}}{s_{xx}}}\left(\frac{\lambda+\hat{\rho}}{\lambda\hat{\rho}+1}\right)(\bar{x}-\bar{x}_R),$$

where $\bar{x}_R$ and $\bar{y}_R$ are respondent means of $X$ and $Y$, $\bar{x}$ is the sample mean of $X$, and $s_{xx}$, $s_{yy}$, and $s_{xy}$ are the respondent sample variances and covariance of $X$ and $Y$. Let $h = \sqrt{\frac{s_{yy}}{s_{xx}}}\left(\frac{\lambda+\hat{\rho}}{\lambda\hat{\rho}+1}\right)$ so that the MLE can be expressed as $\hat{\mu}_y = \bar{y}_R + h \times (\bar{x}-\bar{x}_R)$. Using Taylor expansion,

$$\begin{aligned}
\widehat{\mathrm{Var}}(\hat{\mu}_y) &= \widehat{\mathrm{Var}}(\bar{y}_R + h \times (\bar{x}-\bar{x}_R)) \\
&= \widehat{\mathrm{Var}}(\bar{y}_R) + (\bar{x}-\bar{x}_R)^2\widehat{\mathrm{Var}}(h) + h^2\widehat{\mathrm{Var}}(\bar{x}-\bar{x}_R) + 2h\widehat{\mathrm{Cov}}(\bar{y}_R, \bar{x}-\bar{x}_R) \\
&= \frac{s_{yy}}{r} + (\bar{x}-\bar{x}_R)^2\widehat{\mathrm{Var}}(h) + h^2\left(\frac{\hat{\sigma}_{xx}}{n} + \frac{s_{xx}}{r} - 2\frac{s_{xx}}{n}\right) - 2h\left(\frac{n-r}{n}\right)\frac{s_{xy}}{r} \\
&= \frac{\hat{\sigma}_{yy}}{n} + (\bar{x}-\bar{x}_R)^2\widehat{\mathrm{Var}}(h) + \left(\frac{n-r}{nr}\right)\left(h^2 s_{xx} - 2h s_{xy} + s_{yy}\right) \qquad (4.14)
\end{aligned}$$

since $\mathrm{Cov}(\bar{y}_R, h) = 0$ and $\mathrm{Cov}(\bar{x}-\bar{x}_R, h) = 0$. To find the variance of $h$, we rewrite

$\hat{\rho}$ in terms of variances and covariances and express $h$ as

$$h = \frac{s_{yy}}{s_{xx}} \left( \frac{\lambda\sqrt{s_{xx}s_{xy}} + s_{xy}}{\lambda s_{xy} + \sqrt{s_{xx}s_{yy}}} \right).$$

For bivariate normal $(X_i, Y_i)$, applying the central limit theorem and delta method yields

$$\sqrt{r}[(s_{xx}, s_{yy}, s_{xy})^T - (\sigma_{xx}, \sigma_{yy}, \sigma_{xy})^T] \xrightarrow{d} N(\underline{0}, \Sigma)$$

where

$$\Sigma = \begin{bmatrix} 2\sigma_{xx}^2 & 2\sigma_{xy}^2 & 2\sigma_{xx}\sigma_{xy} \\ 2\sigma_{xy}^2 & 2\sigma_{yy}^2 & 2\sigma_{yy}\sigma_{xy} \\ 2\sigma_{xx}\sigma_{xy} & 2\sigma_{yy}\sigma_{xy} & \sigma_{xy}^2 + \sigma_{xx}\sigma_{yy} \end{bmatrix}$$

Applying the delta method, the asymptotic variance of $h$ is given by $\nabla h^T \Sigma \nabla h / r$, which after some calculations yields

(4.15)
$$\begin{aligned} \widehat{\text{Var}}(h) = \frac{\left(s_{xx}s_{yy} - s_{xy}^2\right)}{r s_{xx}^2 \left(\sqrt{s_{xx}s_{yy}} + \lambda s_{xy}\right)^4} \times & \left\{ s_{xx}^2 s_{yy}^2 (1 - \lambda^2 + \lambda^4) \right. \\ & + 2 s_{xx} s_{yy} s_{xy} \lambda (3\lambda s_{xy} + \sqrt{s_{xx}s_{yy}}(1 + \lambda^2)) \\ & \left. + \lambda s_{xy}^3 (\lambda s_{xy} + 2\sqrt{s_{xx}s_{yy}}(1 + \lambda^2)) \right\}. \end{aligned}$$

Plugging (4.15) into (4.14) completes the calculation of $\widehat{\text{Var}}(\hat{\mu}_y)$.

### 4.9.2 Posterior Draws for Intermediate Values of $\lambda$

Let $W = X + \lambda Y$ be the linear combination of $X$ and $Y$. When $\lambda \in (0, \infty)$, missingness depends on $W$, which is not observed for the nonrespondents. Thus we apply the algorithm for $\lambda = \infty$ to the pair $(X, W)$ and obtain draws from this joint distribution. This results in posterior draws of the following set of parameters:

$(\phi^2, \alpha, \pi, \sigma_{ww}^{(0)}, \mu_w^{(0)}, \sigma_{xx.w}^{(0)}, \sigma_{xx}^{(0)}, \beta_{xw.w}^{(0)}, \beta_{x0.w}^{(0)}, \mu_x^{(0)})$. Draws from the marginal mean of $W$ are obtained by transforming these draws with,

$$\mu_w = \pi\mu_w^{(0)} + (1-\pi)\frac{\mu_x^{(1)} - \beta_{x0.w}^{(0)}}{\beta_{xw.w}^{(0)}}.$$

Since $W = X + \lambda Y$, we have $Y = (W - X)/\lambda$ and thus draws from the marginal mean of $Y$ are obtained by the transformation $\mu_y = (\mu_w - \mu_x)/\lambda$.

Table 4.1: Coverage and median confidence interval length for eighteen artificial populations. ML: Maximum likelihood; PD: Posterior distribution; MI: 20 multiply imputed data sets. Results over 500 replicates.

| | | | n=100 | | | | | | n=400 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Population | | | Coverage | | | CI Width | | | Coverage | | | CI Width | | |
| $\rho$ | $d$ | $\lambda$ | ML | PD | MI | ML | PD | MI | ML | PD | MI | ML | PD | MI |
| 0.8 | 0.1 | 0 | 93 | 94 | 93 | 0.46 | 0.47 | 0.47 | 95 | 94 | 94 | 0.23 | 0.23 | 0.23 |
| | | 1 | 95 | 95 | 95 | 0.47 | 0.48 | 0.48 | 95 | 95 | 95 | 0.24 | 0.24 | 0.24 |
| | | $\infty$ | 95 | 95 | 96 | 0.51 | 0.52 | 0.52 | 95 | 95 | 94 | 0.25 | 0.25 | 0.25 |
| 0.8 | 0.3 | 0 | 94 | 94 | 94 | 0.48 | 0.49 | 0.49 | 96 | 95 | 95 | 0.24 | 0.24 | 0.24 |
| | | 1 | 96 | 96 | 96 | 0.50 | 0.51 | 0.51 | 96 | 95 | 96 | 0.25 | 0.25 | 0.25 |
| | | $\infty$ | 96 | 95 | 96 | 0.55 | 0.56 | 0.56 | 96 | 95 | 96 | 0.27 | 0.27 | 0.27 |
| 0.8 | 0.5 | 0 | 95 | 96 | 95 | 0.52 | 0.53 | 0.53 | 96 | 95 | 95 | 0.26 | 0.26 | 0.26 |
| | | 1 | 96 | 97 | 96 | 0.54 | 0.56 | 0.55 | 96 | 95 | 97 | 0.27 | 0.27 | 0.27 |
| | | $\infty$ | 97 | 96 | 97 | 0.62 | 0.64 | 0.64 | 97 | 96 | 96 | 0.31 | 0.31 | 0.31 |
| 0.5 | 0.1 | 0 | 93 | 93 | 93 | 0.52 | 0.53 | 0.54 | 94 | 93 | 94 | 0.26 | 0.26 | 0.27 |
| | | 1 | 95 | 96 | 95 | 0.56 | 0.59 | 0.59 | 95 | 95 | 96 | 0.29 | 0.28 | 0.29 |
| | | $\infty$ | 97 | 95 | 97 | 0.84 | 0.98 | 0.96 | 96 | 95 | 95 | 0.41 | 0.42 | 0.43 |
| 0.5 | 0.3 | 0 | 93 | 94 | 94 | 0.54 | 0.56 | 0.56 | 94 | 94 | 94 | 0.27 | 0.27 | 0.28 |
| | | 1 | 96 | 97 | 96 | 0.59 | 0.64 | 0.64 | 95 | 95 | 96 | 0.3 | 0.31 | 0.31 |
| | | $\infty$ | 96 | 96 | 97 | 1.0 | 1.2 | 1.2 | 95 | 95 | 96 | 0.51 | 0.52 | 0.53 |
| 0.5 | 0.5 | 0 | 95 | 95 | 95 | 0.58 | 0.6 | 0.61 | 94 | 94 | 95 | 0.29 | 0.29 | 0.3 |
| | | 1 | 97 | 97 | 98 | 0.64 | 0.73 | 0.72 | 95 | 96 | 97 | 0.33 | 0.35 | 0.35 |
| | | $\infty$ | 96 | 97 | 96 | 1.3 | 1.6 | 1.6 | 97 | 96 | 96 | 0.66 | 0.68 | 0.69 |
| 0.2 | 0.1 | 0 | 93 | 94 | 94 | 0.55 | 0.56 | 0.57 | 94 | 93 | 93 | 0.28 | 0.27 | 0.28 |
| | | 1 | 94 | 96 | 96 | 0.64 | 0.72 | 0.72 | 95 | 95 | 95 | 0.33 | 0.33 | 0.34 |
| | | $\infty$ | 94 | 97 | 97 | 2.5 | 9.9 | 9.0 | 94 | 97 | 96 | 1.2 | 1.7 | 1.6 |
| 0.2 | 0.3 | 0 | 94 | 95 | 94 | 0.57 | 0.59 | 0.6 | 95 | 94 | 94 | 0.29 | 0.29 | 0.29 |
| | | 1 | 87 | 96 | 94 | 0.66 | 0.98 | 0.97 | 95 | 96 | 97 | 0.34 | 0.38 | 0.38 |
| | | $\infty$ | 87 | 96 | 93 | 4.7 | 23 | 19 | 90 | 97 | 94 | 2.3 | 3.4 | 3.3 |
| 0.2 | 0.5 | 0 | 95 | 95 | 95 | 0.62 | 0.63 | 0.65 | 96 | 95 | 94 | 0.31 | 0.31 | 0.32 |
| | | 1 | 86 | 98 | 97 | 0.73 | 1.7 | 1.4 | 95 | 97 | 98 | 0.36 | 0.45 | 0.45 |
| | | $\infty$ | 85 | 96 | 94 | 7.4 | 39 | 32 | 90 | 97 | 96 | 3.5 | 5.6 | 5.3 |

Table 4.2: Parameters in the model for $M$ given $Z$ and $Y$ for the third simulation.

| Model | $\gamma_0$ | $\gamma_Z$ | $\gamma_{Z2}$ | $\gamma_Y$ | $\gamma_{Y2}$ |
|---|---|---|---|---|---|
| $[Z]$ | -0.5 | 0.5 | 0 | 0 | 0 |
| $[Z^2]$ | -1 | 0 | 0.5 | 0 | 0 |
| $[Y]$ | -0.5 | 0 | 0 | 0.5 | 0 |
| $[Y^2]$ | -1 | 0 | 0 | 0 | 0.5 |
| $[Z+Y]$ | -1 | 0.5 | 0 | 0.5 | 0 |
| $[Z^2+Y^2]$ | -2 | 0 | 0.5 | 0 | 0.5 |
| $[Z^2+Y]$ | -1.5 | 0 | 0.5 | 0.5 | 0 |
| $[Z+Y^2]$ | -1.5 | 0.5 | 0 | 0 | 0.5 |

Table 4.3: Fraction of missing information (FMI) estimates from NHANES III data for three outcomes. SBP = systolic blood pressure; DBP = diastolic blood pressure; BMI = body-mass index, log-transformed to approximate normality. FMIwt denotes estimation incorporating the survey design.

| Outcome | Missing (%) | $\hat{\rho}$ | $d$ | $d^*$ | $\lambda$ | FMI (%) | FMIwt (%) |
|---|---|---|---|---|---|---|---|
| SBP | 15 | 0.60 | 1.0 | 0.079 | 0 | 9.6 | 10 |
| | | | | | 1 | 13 | 8.3 |
| | | | | | $\infty$ | 24 | 11 |
| DBP | 15 | 0.33 | 0.050 | 0.011 | 0 | 16 | 5.8 |
| | | | | | 1 | 24 | 8.3 |
| | | | | | $\infty$ | 75 | 30 |
| BMI | 9.7 | 0.24 | -0.00042 | -0.0084 | 0 | 12 | 3.6 |
| | | | | | 1 | 18 | 5.8 |
| | | | | | $\infty$ | 83 | 53 |

Figure 4.1: Fraction of missing information as a function of $\rho$, $d$, and nonresponse rate.

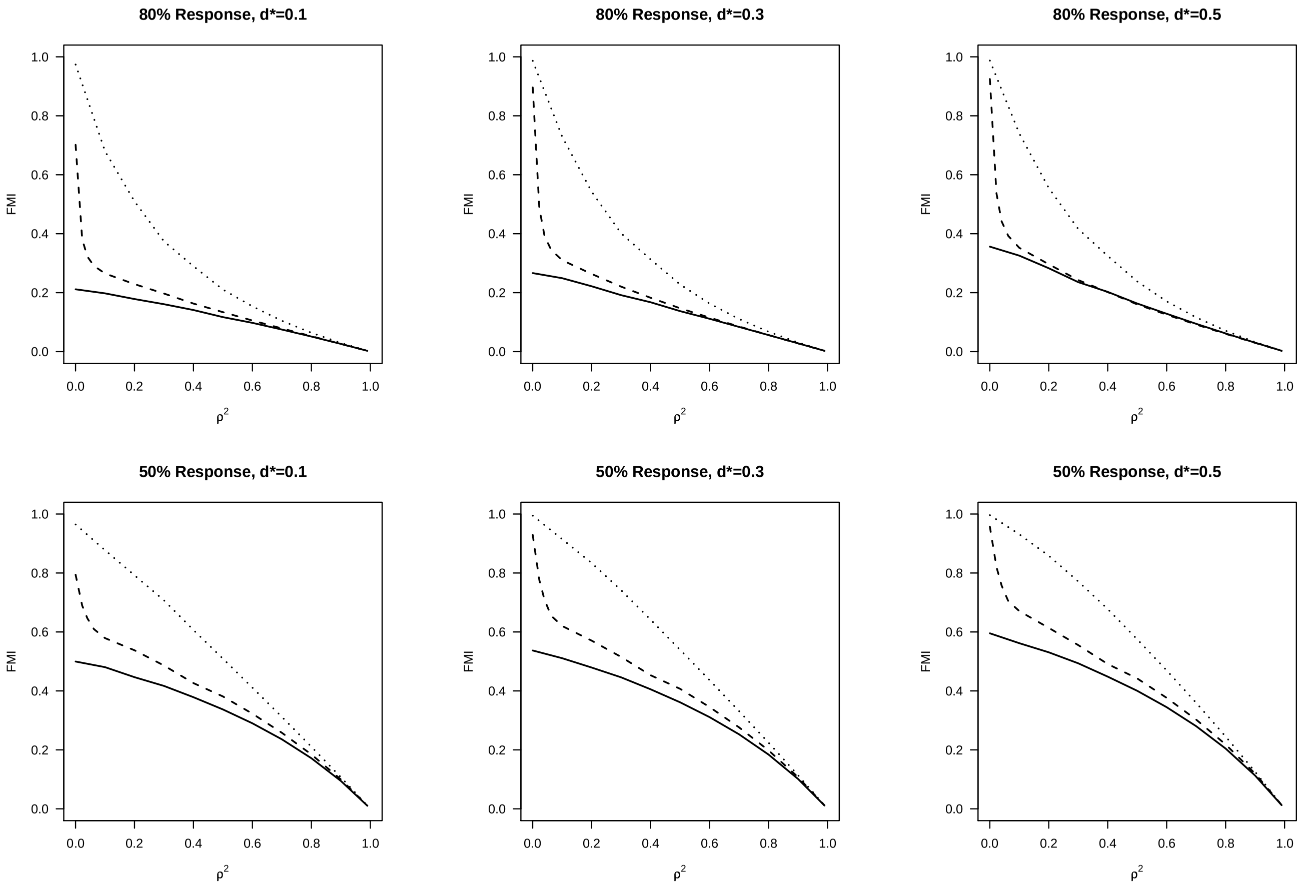


Figure 4.2: 95% confidence intervals for nine generated data sets ($n = 100$) for $\lambda = (0, 1, \infty)$. Numbers below intervals are the interval length. CC: Complete case; ML: Maximum likelihood; PD: Posterior distribution; MI: 20 multiply imputed data sets.

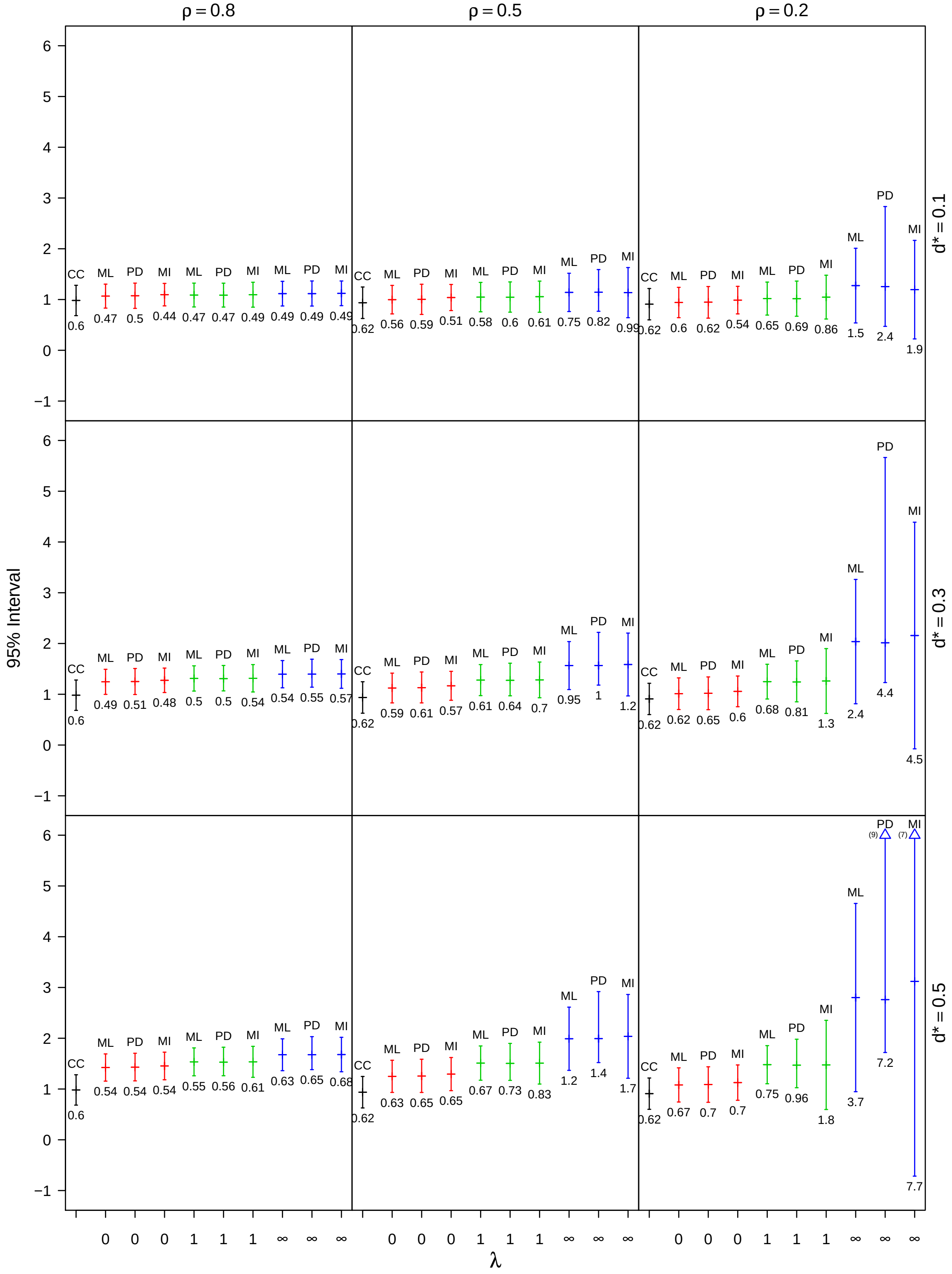
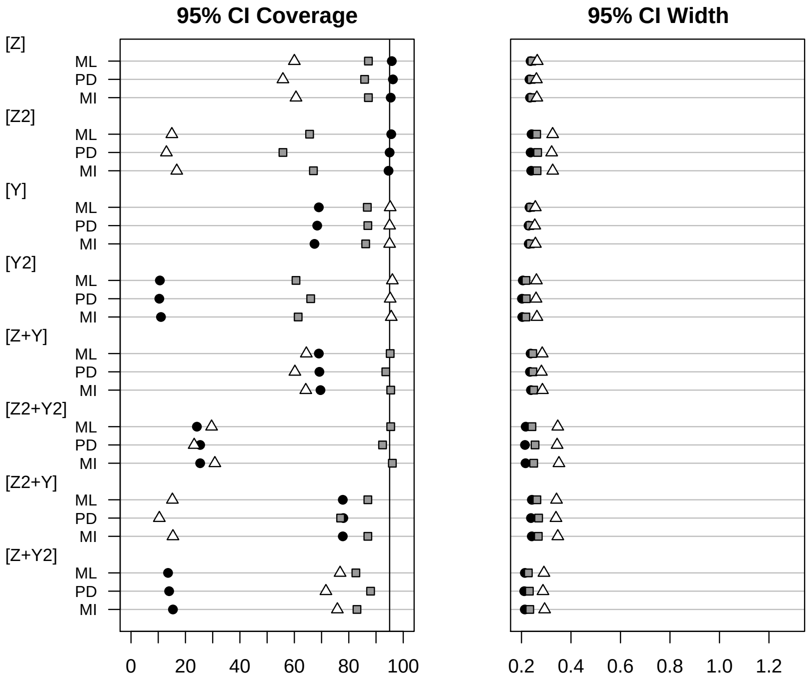
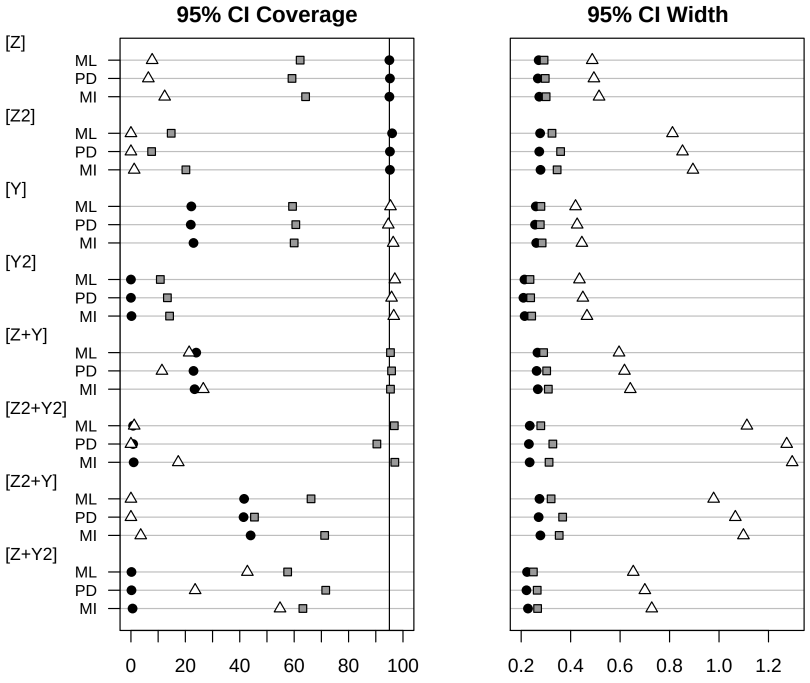


Figure 4.3: Coverage and median CI length for twenty-four artificial populations for $\lambda = 0$ (●), $\lambda = 1$ (■), and $\lambda = \infty$ ($\Delta$), with (a) $\rho = 0.8$; (b) $\rho = 0.5$; (c) $\rho = 0.2$. ML: Maximum likelihood; PD: Posterior distribution; MI: 20 multiply imputed data sets. Results over 500 replicates with $n = 400$.

(a) $\rho = 0.8$



(b) $\rho = 0.5$



(c) $\rho = 0.2$

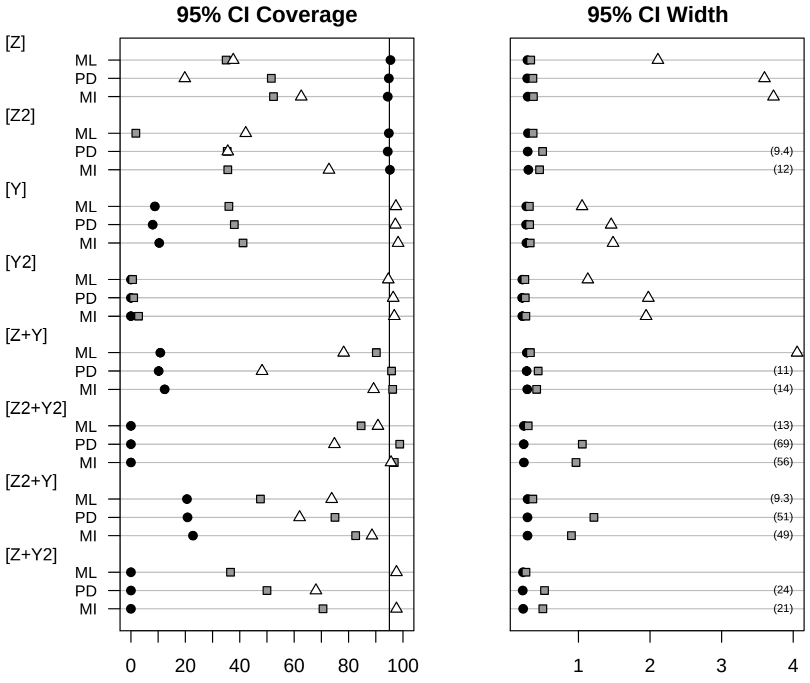


Figure 4.4: Estimates of mean SBP for $\lambda = (0, 1, \infty)$ based on NHANES III adult data. Numbers below intervals are the interval length. CC: Complete case; CC wt: Complete case with estimation incorporating the survey design; ML: Maximum likelihood; PD: Posterior distribution; MI: 20 multiply imputed data sets; MIwt: 20 multiply imputed data sets with estimation incorporating the survey design.

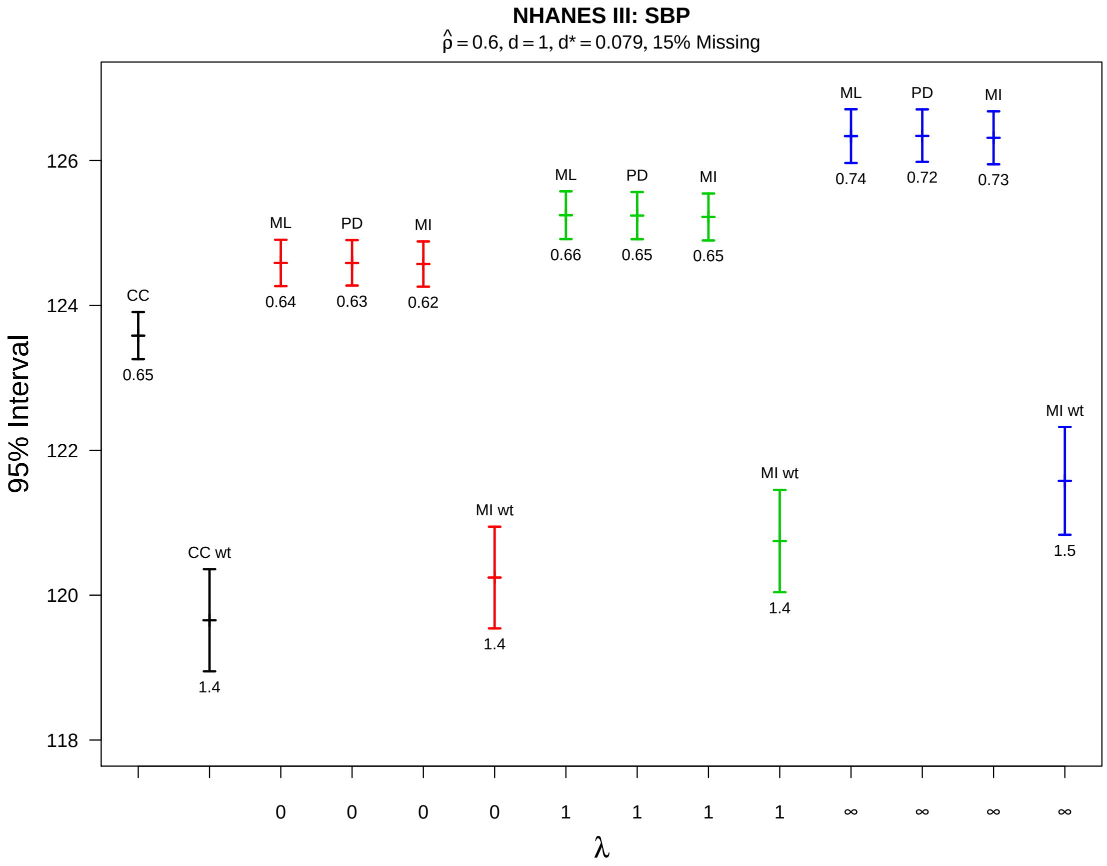


Figure 4.5: Estimates of mean DBP for $\lambda = (0, 1, \infty)$ based on NHANES III adult data. Numbers below intervals are the interval length. CC: Complete case; CC wt: Complete case with estimation incorporating the survey design; ML: Maximum likelihood; PD: Posterior distribution; MI: 20 multiply imputed data sets; MIwt: 20 multiply imputed data sets with estimation incorporating the survey design.

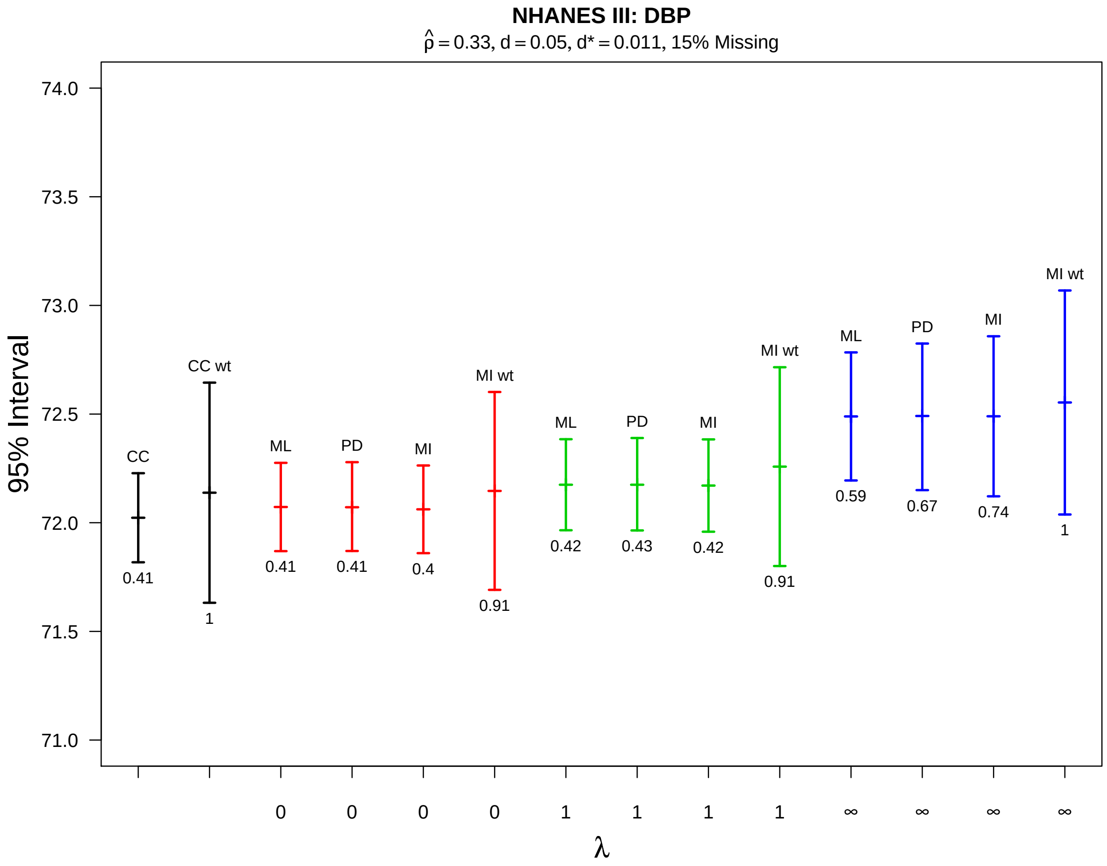
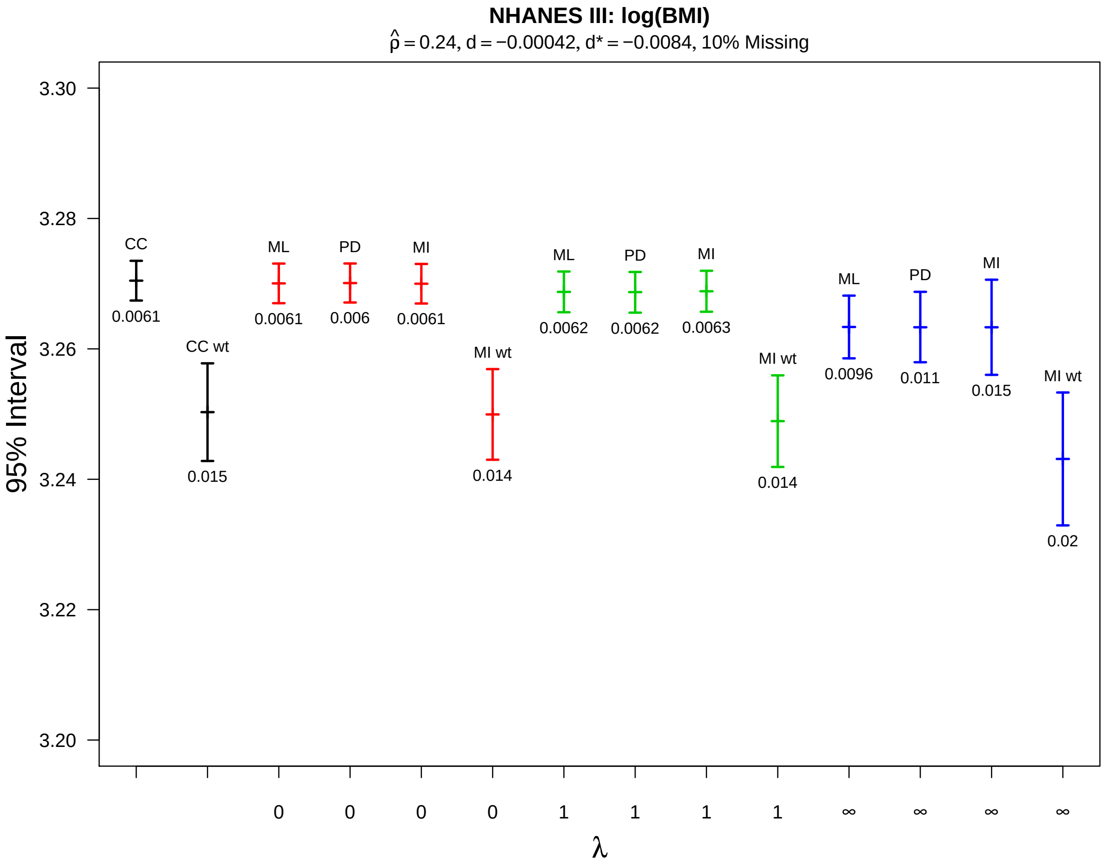


Figure 4.6: Estimates of mean BMI (log-transformed) for $\lambda = (0, 1, \infty)$ based on NHANES III adult data. Numbers below intervals are the interval length. CC: Complete case; CC wt: Complete case with estimation incorporating the survey design; ML: Maximum likelihood; PD: Posterior distribution; MI: 20 multiply imputed data sets; MIwt: 20 multiply imputed data sets with estimation incorporating the survey design.



## CHAPTER V

# Extensions of Proxy Pattern-Mixture Analysis

### 5.1 Introduction

Response rates for large-scale surveys have been steadily declining in recent years (Curtain, Presser, and Singer, 2005), increasing the need for methods to analyze the impact of nonresponse on survey estimates. There are three major components to consider in evaluating nonresponse: the amount of missingness, differences between respondents and nonrespondents on characteristics that are observed for the entire sample, and the relationship between these fully observed covariates and the survey outcome of interest. Current methods to handle nonresponse in surveys have tended to focus on a subset of these components, however, the impact of nonresponse cannot be fully understood without all three pieces. In addition, historically the focus has been on situations were data are assumed to be missing at random (Rubin, 1976), with less attention paid to the case when missingness may be not at random (NMAR), that is, depend on the unobserved outcome itself. In this paper we propose a method for estimating population proportions in survey samples with nonresponse that includes but does not assume ignorable missinginess.

A limited amount of work has been done in the area of nonignorable nonresponse for categorical outcomes in survey data. Some examples include Stasny (1991), who used a hierarchical Bayes nonignorable selection model to study victimization in the National Crime Survey. Extensions of this approach by Nandram and Choi (2002a) and Nandram and Choi (2002b) use continuous model expansion to center the nonignorable model on an ignorable model, in the manner of Rubin (1977). Similar methods are developed for multinomial outcomes in Nandram, Han, and Choi (2002) and Nandram, Liu, Choi, and Cox (2005) and used to study health outcomes in the third National Health and Nutrition Examination Survey (NHANES III). The main difference between our proposed approach and these previous methods is the method of modeling the missing data. There are two general classes of models for incomplete data, selection models and pattern-mixture models (Little and Rubin, 2002). Previous work on nonresponse models in surveys has tended to favor the selection model; we use a pattern-mixture approach. The pattern-mixture approach requires explicit assumptions on the missing data mechanism and naturally leads to a sensitivity analysis, whereas the selection model approach requires strong distributional assumptions to (often weakly) identify parameters. In addition, methods for categorical nonresponse have tended to be limited to the case when auxiliary data are also categorical. However, auxiliary variables may be continuous; our proposed method does not require that continuous variables be categorized before inclusion in the model.

The work in this paper is an extension of our previously described proxy pattern-mixture analysis (PPMA) for a continuous outcome; In Section 5.2 we briefly review the continuous outcome PPMA before describing its extension to binary outcomes in Section 5.3. Section 5.4 discusses three different estimation approaches, maximum likelihood, a Bayesian approach, and multiple imputation, and the sensitivity of each method to model misspecification. In Section 5.5 we extend the binary case to ordinal outcomes. These methods are illustrated first through simulation in Section 5.6 and then by application to NHANES III data in Section 5.7. Section 5.8 presents some concluding remarks.

## 5.2 Review of the Proxy Pattern-Mixture Model

Proxy pattern-mixture analysis was developed for the purpose of assessing nonresponse bias for estimating the mean of a continuous survey variable $Y$ subject to nonresponse. For simplicity, we initially consider an infinite population with a sample of size $n$ drawn by simple random sampling. Let $Y_i$ denote the value of a continuous survey outcome and $Z_i = (Z_{i1}, Z_{i2}, \ldots, Z_{ip})$ denote the values of $p$ covariates for unit $i$ in the sample. Only $r$ of the $n$ sampled units respond, so observed data consist of $(Y_i, Z_i)$ for $i = 1, \ldots, r$ and $Z_i$ for $i = r + 1, \ldots, n$. In particular this can occur with unit nonresponse, where the covariates $Z$ are design variables known for the entire sample or with item nonresponse. Of primary interest is assessing and correcting nonresponse bias for the mean of $Y$.

To reduce dimensionality and for simplicity we reduce the covariates $Z$ to a single proxy variable $X$ that has the highest correlation with $Y$, estimated from a regression analysis of $Y$ on $Z$ using respondent data. Let $\rho$ be the correlation of $Y$ and $X$, which we assume is positive. If $\rho$ is high (say, 0.8) we call $X$ a strong proxy for $Y$ and if $X$ is low (say, 0.2) we call $X$ a weak proxy for $Y$. In addition to the strength of the proxy as measured by $\rho$, an important factor is the deviation from missing completely at random (MCAR) as measured by the difference between the overall mean of the proxy and the respondent mean of the proxy, $d = \bar{x} - \bar{x}_R$. The distribution of $X$ for respondents and nonrespondents provides the main source of information for assessing nonresponse bias for $Y$. We consider adjusted estimators of the mean of

$Y$ that are maximum likelihood for a pattern-mixture model with different mean and covariance matrix of $Y$ and $X$ for respondents and nonrespondents, assuming missingness is an arbitrary function of a known linear combination of $X$ and $Y$. This allows insight into whether missingness may be not at random (NMAR).

Specifically, we let $M$ denote the missingness indicator, such that $M = 0$ if $Y$ is observed and $M = 1$ if $Y$ is missing. We assume that the joint distribution of $[Y, X, M]$ follows the bivariate pattern-mixture model discussed in Little (1994). This model is underidentified, since there is no information on the conditional normal distribution for $Y$ given $X$ for nonrespondents ($M = 1$). However, Little (1994) shows that the model can be identified by making assumptions about how missingness of $Y$ depends on $Y$ and $X$. For the proxy pattern-mixture we assume that,

$$\Pr(M = 1|Y, X) = f(X\sqrt{\frac{\sigma_{yy}^{(0)}}{\sigma_{xx}^{(0)}}} + \lambda Y) = f(X^* + \lambda Y), \tag{5.1}$$

where $X^*$ is the proxy variable $X$ scaled to have the same variance as $Y$ in the respondent population. By a slight modification of the arguments in (Little, 1994), the resulting maximum likelihood estimate of the overall mean of $Y$ is,

$$\hat{\mu}_y = \bar{y}_R + \frac{\lambda + \hat{\rho}}{\lambda\hat{\rho} + 1}\sqrt{\frac{s_{yy}}{s_{xx}}}(\bar{x} - \bar{x}_R), \tag{5.2}$$

where $\bar{x}_R$ and $\bar{y}_R$ are the respondent means of $X$ and $Y$, $s_{xx}$ and $s_{yy}$ are the respondent sample variances of $X$ and $Y$, and $\bar{x}$ is the overall sample mean of $X$.

The parameter $\lambda$ is a sensitivity parameter; there is no information in the data with which to estimate it. Different choices of $\lambda$ correspond to different assumptions on the missing data mechanism. We assume that $\lambda$ is positive, which seems reasonable given that $X$ is a proxy for $Y$. Then as $\lambda$ varies between 0 (missingness depends only on $X$) and infinity (missingness depends only on $Y$), $g(\hat{\rho}) = (\lambda + \hat{\rho})/(\lambda\hat{\rho} + 1)$ varies between $\hat{\rho}$ and $1/\hat{\rho}$. When $\lambda = 0$ the data are MAR, since in this case missingness depends only on the observed variable $X$. In this case $g(\hat{\rho}) = \hat{\rho}$, and (5.2) reduces to the standard regression estimator. In this case the bias adjustment for $Y$ increases with $\hat{\rho}$, as the association between $Y$ and the variable determining the missing data mechanism increases. On the other hand when $\lambda = \infty$ and missingness depends only on the true value of $Y$, $g(\hat{\rho}) = 1/\hat{\rho}$ and (5.2) yields the inverse regression estimator proposed by Brown (1990). The bias adjustment thus decreases with $\hat{\rho}$, reflecting the fact that in this case the bias in $Y$ is attenuated in the proxy, with the degree of attenuation increasing with $\hat{\rho}$.

For assessing potential nonresponse bias in the mean of $Y$, we suggest a sensitivity analysis using $\lambda = (0, 1, \infty)$ to capture a range of missingness mechanisms. In addition to the extremes, we use the intermediate case of $\lambda = 1$ that weights the proxy and true value of $Y$ equally because the resulting estimator has a particularly convenient and simple interpretation. In this case $g(\hat{\rho}) = 1$ regardless of the value of $\hat{\rho}$, implying that the standardized bias in $\bar{y}_R$ is the same as the standardized bias in $\bar{x}_R$. In general, the stronger the proxy, the closer the value of $\hat{\rho}$ to one, and the smaller the differences between the three estimates.

## 5.3 Extension of PPMA to a Binary Outcome

The proxy pattern-mixture analysis described above strictly only applies to continuous survey variables, where normality is reasonable. However, categorical outcomes are ubiquitous in sample surveys. In this section we extend PPMA to binary outcomes using a latent variable approach. Let $Y_i$ now denote the value of a partially missing binary survey outcome, and $Z_i = (Z_{i1}, Z_{i2}, \ldots, Z_{ip})$ denote the values of $p$ fully observed covariates for unit $i$ in the sample. As before, only $r$ of the $n$ sampled units respond, so observed data consist of $(Y_i, Z_i)$ for $i = 1, \ldots, r$ and $Z_i$ for

$i = r+1, \ldots, n$. Of interest is the proportion of units in the population with $Y = 1$.

For simplicity and to reduce dimensionality, we replace $Z$ by a single continuous proxy variable $X$, estimated by a probit regression of $Y$ on $Z$ using the respondent data,

$$\Pr(Y = 1|Z, M = 0) = \Phi(\alpha_0 + \alpha Z). \tag{5.3}$$

We take $X = \hat{\alpha}_0 + \hat{\alpha} Z$ to be the linear predictor from the probit regression, rather than the predicted probability, so that its support is the real line. The regression coefficients $\alpha$ are subject to sampling error, so in practice $X$ is estimated rather than known. The choice of the probit link, rather than alternatives such as the logit link, is due to the latent variable motivation of probit regression. We assume that $Y$ is related to a continuous normally distributed latent variable $U$ through the rule that $Y = 1$ when the latent variable $U > 0$. The latent (respondent) data are then related to the covariates through the linear regression equation, $U = \alpha_0 + \alpha Z + \epsilon$, where $\epsilon \sim N(0, 1)$.

This latent variable approach motivates application of the normal proxy pattern-mixture (PPM) model to the latent variable $U$ and proxy $X$. If we could observe $U$ for the respondents, application of the PPM model would be straightforward. Taking $M$ to be the missing data indicator, we assume that the joint distribution of $[U, X, M]$ follows the bivariate pattern-mixture model:

$$\begin{gathered}
(U, X|M = m) \sim N_2\left((\mu_u^{(m)}, \mu_x^{(m)}), \Sigma^{(m)}\right) \\
M \sim Bernoulli(1 - \pi) \\
\Sigma^{(m)} = \begin{bmatrix} \sigma_{uu}^{(m)} & \rho^{(m)}\sqrt{\sigma_{uu}^{(m)}\sigma_{xx}^{(m)}} \\ \rho^{(m)}\sqrt{\sigma_{uu}^{(m)}\sigma_{xx}^{(m)}} & \sigma_{xx}^{(m)} \end{bmatrix},
\end{gathered} \tag{5.4}$$

where $N_2$ denotes the bivariate normal distribution. Note that the parameter $\rho^{(m)}$

is the correlation between the latent variable $U$ and the constructed proxy $X$. As with the continuous outcome PPM model the parameters $\mu_u^{(1)}$, $\sigma_{uu}^{(1)}$, and $\rho^{(1)}$ are unidentifiable without further model restrictions. Since $U$ is completely unobserved, $\sigma_{uu}^{(0)}$ is also not identifiable and without loss of generality can be fixed at an arbitrary value. Following convention we set $\sigma_{uu}^{(0)} = 1/(1-{\rho^{(0)}}^2)$ so $\mathrm{Var}(U|X, M=0) = 1$.

We identify the model by making assumptions about how missingness of $Y$ depends on $U$ and $X$. As with the continuous outcome PPM model, we modify the arguments in Little (1994) and assume that

$$\text{(5.5)} \qquad \Pr(M=1|U,X) = f(X\sqrt{\frac{\sigma_{uu}^{(0)}}{\sigma_{xx}^{(0)}}} + \lambda U) = f(X^* + \lambda U),$$

where $X^*$ is the proxy variable $X$ scaled to have the same variance as $U$ in the respondent population. An important feature of this mechanism is that when $\lambda > 0$, i.e. under NMAR, the missingness in the binary outcome $Y$ is being driven by $X$ and by the completely unobserved latent $U$. This allows for a “smooth” missingness function in the sense that conditional on $X$ the probability of missingness may lie on a continuum instead of only taking two values (as would be the case if missingness depended on $Y$ itself). Of primary interest is the marginal mean of $Y$, which is given by,

$$\text{(5.6)} \quad \mu_y = \Pr(Y=1) = \Pr(U>0) = \pi\Phi\left(\mu_u^{(0)}/\sqrt{\sigma_{uu}^{(0)}}\right) + (1-\pi)\Phi\left(\mu_u^{(1)}/\sqrt{\sigma_{uu}^{(1)}}\right),$$

where $\Phi(\cdot)$ denotes the standard normal CDF.

### 5.3.1 Summary of Evidence about Nonresponse Bias

The information about nonresponse bias in the mean of $Y$ is contained in the strength of the proxy as measured by $\rho^{(0)}$ and the deviation in the proxy mean, $d = \bar{x} - \bar{x}_R$. Strong proxies (large $\rho^{(0)}$) and small deviations (small $d$) lead to decreased uncertainty and higher precision in estimates, even under NMAR, while weak proxies (low $\rho^{(0)}$) and large deviations (large $d$) lead to increased uncertainty, especially when missingness depends on $Y$. In the case of the continuous outcome, both $\rho^{(0)}$ and $d$ were directly interpretable, since $\hat{\rho}^{(0)}$ was the square root of the $R^2$ from the regression model that built the proxy $X$ and the deviation $d$ was on the same scale as the (linear) outcome $Y$. With a binary outcome, we lose these neat interpretations of $\rho^{(0)}$ and $d$, though their usefulness as markers of the severity of the nonresponse problem ($d$) and our ability to make adjustments to combat the problem ($\rho^{(0)}$) remains. The information about nonresponse bias in $Y$ is contained in $X$, with $X$ now a proxy for the latent $U$ instead of the partially observed outcome $Y$ itself.

Another issue unique to the binary case (and subsequent extension to ordinal $Y$) is that the size of the nonresponse bias in $Y$, i.e. the difference in mean between respondent and overall means, depends not only on the size of the deviation on the latent scale ($d$) but also on the respondent mean itself. In the continuous case, the bias in $\bar{y}_R$ is a linear function of $d$ (see (5.2)); a deviation $d$ has the same (standardized) effect on the overall mean regardless of the value of $\bar{y}_R$. However, in the binary case the deviation is on the latent scale, and only the bias in $U$ is location-invariant. When transformed to the binary outcome, different $d$ values will lead to different size biases, depending on the respondent mean of $Y$. The use of the standard normal CDF to transform $U$ to $Y$ drives this; the difference $\Phi(a+d)-\Phi(a)$ is not merely a function of $d$ but also depends on the value of $a$.

## 5.4 Estimation Methods

### 5.4.1 Maximum Likelihood

Maximum likelihood (ML) estimators for the distribution of $U$ given $X$ for nonrespondents follow directly from the continuous outcome PPM model,

$$\begin{aligned}
\hat{\mu}_u^{(1)} &= \hat{\mu}_u^{(0)} + g \times \sqrt{\frac{\hat{\sigma}_{uu}^{(0)}}{\hat{\sigma}_{xx}^{(0)}}}(\bar{x}_{NR} - \bar{x}_R) \\
\hat{\sigma}_{uu}^{(1)} &= \hat{\sigma}_{uu}^{(0)} + g^2 \times \frac{\hat{\sigma}_{uu}^{(0)}}{\hat{\sigma}_{xx}^{(0)}}(\hat{\sigma}_{xx}^{(1)} - \hat{\sigma}_{xx}^{(0)}) \qquad (5.7) \\
g &= \frac{\lambda + \hat{\rho}^{(0)}}{\lambda \hat{\rho}^{(0)} + 1}.
\end{aligned}$$

Plugging these estimates into (5.6) yields the ML estimate of the mean of $Y$. The ML estimates of the parameters of the distribution of $X$ are the usual estimators, however, estimators for $\mu_u^{(0)}$ and $\rho^{(0)}$, and therefore $\sigma_{uu}^{(0)}$, are not immediately obvious since the latent $U$ is unobserved even for respondents. To obtain these estimates, we note that the correlation $\rho^{(0)}$ is the biserial correlation between the binary $Y$ and continuous $X$ for the respondents. Maximum likelihood estimation of the biserial correlation coefficient was first studied by Tate (1955a,b), who showed that a closed form solution does not exist. The parameters $\rho^{(0)}$ and $\omega^{(0)} = \mu_u^{(0)}/\sqrt{\sigma_{uu}^{(0)}}$ (referred to as the cutpoint) must be jointly estimated through an iterative procedure such as a Newton-Raphson type algorithm. It is important to note that the ML estimate of $\omega^{(0)}$ is not the inverse probit of the respondent mean of $Y$, i.e. the ML estimate of the mean of $Y$ for respondents is not $\bar{y}_R$.

An alternative method of estimating the biserial correlation coefficient is the *two-step method*, proposed by Olsson, Drasgow, and Dorans (1982) in the context of the polyserial correlation coefficient. In the first step, the cutpoint $\omega^{(0)}$ is estimated by $\hat{\omega}^{(0)} = \Phi^{-1}(\bar{y}_R)$, so that the ML estimate of the respondent mean of $Y$ is $\bar{y}_R$. Then a conditional maximum likelihood estimate of $\rho^{(0)}$ is then computed, given the other parameter estimates. This method is computationally simpler than the full ML estimate, and also has the attractive property of returning the logical estimate $\hat{\mu}_y^{(0)} = \bar{y}_R$.

The large sample variance of the full ML estimate of $\mu_y$ is obtained through Taylor series expansion and inversion of the information matrix. The properties of the two-step estimator are not well studied, so variance estimates are obtained with the bootstrap.

### 5.4.2 Bayesian Inference

The ML estimate ignores the uncertainty inherent in the creation of the proxy $X$. An alternative approach is to use a Bayesian framework that allows incorporation of this uncertainty. Since $U$ is unobserved, we propose using a data augmentation approach. We place noninformative priors on the regression parameters $\alpha$ and use a Gibbs sampler to draw the latent $U$ for respondents (Albert and Chib, 1993). Conditional on $\alpha$ (and therefore on the created proxy $X$), $U$ follows a truncated normal distribution,

$$(U|Y, \alpha, M = 0) = (U|Y, X, M = 0) \sim N(X, 1) = N(\alpha Z, 1) \quad \text{truncated at the left by 0 if } Y = 1 \text{ and at the right by 0 if } Y = 0. \tag{5.8}$$

Then given the augmented continuous $U$ we draw $\alpha$ from its posterior distribution, which also follows a normal distribution,

$$(\alpha|Y, U, M = 0) \sim N((Z^T Z)^{-1} Z^T U, (Z^T Z)^{-1}), \tag{5.9}$$

and recreate the proxy $X = \alpha Z$.

This data augmentation allows for straightforward application of the Bayesian estimation methods for continuous PPMA. For a chosen value of $\lambda$, we apply the

PPM algorithm as described in Chapter IV to the pair $(X, U)$ to obtain draws of the parameters of the joint distribution of $X$ and $U$. Since $U$ is unobserved even for the respondents, after each draw of the parameters from the PPM model, $X$ is recreated for the entire sample and $U$ is redrawn for the respondents given the current set of parameter values as described in the data augmentation approach above. Note that this does not require a draw of the latent data for nonrespondents. Draws from the posterior distribution of $\mu_y$ are obtained by substituting the draws from the Gibbs sampler into (5.6).

### 5.4.3 Multiple Imputation

An alternative method of inference is multiple imputation (Rubin, 1978). For a selected $\lambda$ we create $K$ complete data sets by filling in missing $Y$ values with draws from the posterior distribution, based on the pattern-mixture model. For a given draw of the parameters $\phi = (\mu_u^{(1)}, \mu_x^{(1)}, \sigma_{uu}^{(1)}, \sigma_{xx}^{(1)}, \rho^{(1)})$ from their posterior distribution as Section 5.4.2, we draw the latent $U$ for nonrespondents based on the conditional distribution,

$$(5.10) \qquad [u_i | x_i, m_i = 1, \phi_{(k)}] \sim N\left(\mu_{u(k)}^{(1)} + \frac{\sigma_{ux(k)}^{(1)}}{\sigma_{xx(k)}^{(1)}}\left(x_i - \mu_{x(k)}^{(1)}\right), \sigma_{uu(k)}^{(1)} - \frac{{\sigma_{ux(k)}^{(1)}}^{2}}{\sigma_{xx(k)}^{(1)}}\right)$$

where the subscript $(k)$ denotes the $k$th draws of the parameters. In order to reduce auto-correlation between the imputations due to the Gibbs sampling algorithm for drawing the parameters, we thin the chain for the purposes of creating the imputations. The missing $y_i$ are then imputed as $y_i = I(u_i > 0)$, where $I()$ is an indicator function taking the value 1 if the expression is true. For the $k$th completed data set, the estimate of $\mu_y$ is the sample mean $\bar{Y}_k$ with estimated variance $W_k$. A consistent estimate of $\mu_y$ is then given by $\hat{\mu}_y = \frac{1}{K}\sum_{k=1}^{K} \bar{Y}_k$ with $\text{Var}(\hat{\mu}_y) = \bar{W}_K + \frac{K+1}{K} B_K$, where $\bar{W}_K = \frac{1}{K}\sum_{k=1}^{K} W_k$ is the within-imputation variance and $B = \frac{1}{K-1}\sum_{k=1}^{K}(\bar{Y}_k - \hat{\mu}_y)^2$

is the between-imputation variance.

As with the continuous PPMA, an advantage of the multiple imputation approach is the ease with which complex design features like clustering, stratification and unequal sampling probabilities can be incorporated. Once the imputation process has created complete data sets, design-based methods can be used to estimate $\mu_y$ and its variance; for example the Horvitz-Thompson estimator can be used to calculate $\bar{Y}_k$.

### 5.4.4 Sensitivity to a Non-normally Distributed Proxy

A crucial assumption of the PPM model for both continuous and binary outcomes is that of bivariate normality of $X$ and $Y$ or $U$. The continuous outcome PPM model is relatively robust to departures from this assumption and only relies on linear combinations of first and second moments in estimating the mean of $Y$. However, for binary outcomes the normality assumption plays a more crucial role, made clear with a simple example. Suppose the proxy $X$ is normally distributed in the respondent population, with $[X|M=0] \sim N(\mu_x^{(0)}, \sigma_{xx}^{(0)})$. We assume that, for respondents, the latent variable $U = X + e$ where $e \sim N(0,1)$, such that $\Pr(Y=1|M=0) = \Pr(U > 0|M=0)$. Then the conditional and marginal respondent distributions of $U$ along with the mean of $Y$ are given by,

$$\begin{aligned}
[U|X, M=0] &\sim N(X, 1) \\
[U|M=0] &\sim N(\mu_x^{(0)}, 1 + \sigma_{xx}^{(0)}) \\
\mu_y^{(0)} = \Pr(U > 0|M=0) = \Phi\left(\mu_u^{(0)}/\sqrt{\sigma_{uu}^{(0)}}\right) &= \Phi\left(\mu_x^{(0)}/\sqrt{1+\sigma_{xx}^{(0)}}\right)
\end{aligned}$$

However, if the distribution of $X$, $f_X(x)$, is not normal, then the conditional distribution $[U|X, M=0]$ is the same but the marginal distribution is no longer normal. Now $\Pr(U > 0) = \int_0^\infty f_U(u)\, \mathrm{d}u$ where $f_U(u)$ is the convolution of the error distribution $N(0,1)$ and $f_X(x)$. Thus even the estimate of the respondent mean of $Y$ will be biased, despite the fact that $Y$ is fully observed for the respondents.

Even though PPMA can provide unbiased estimates of the mean and variance of $U$ in the case when $Z$ is not normally distributed (like the continuous PPMA), the transformation to the mean of $Y$ is only accurate when $Z$ is normally distributed. Both the full ML estimation and Bayesian methods will produce biased estimates of $\mu_y$ if $X$ deviates away from normality. The two-step ML method is less sensitive to non-normality, since it estimates $\mu_y^{(0)}$ by $\bar{y}_R$. Multiple imputation also is less sensitive to departures from normality since imputation is based on the conditional distribution $[U|X, M]$ which is normal by definition of the latent variable and is not affected by non-normal $X$.

We propose modifying the Bayesian method to attempt to reduce sensitivity to deviations from normality in the proxy $X$. The modification is an extension of the multiple imputation approach: at each iteration of the Gibbs sampler, the latent $U$ for nonrespondents is drawn conditional on the current parameter values, and the subsequent draw of $\mu_y^{(1)}$ is taken to be $\mu_y^{(1)} = \frac{1}{n-r}\sum_{i=r+1}^{n} I(U_i > 0)$. A similar method of obtaining an estimator for the respondent mean does not work, as draws of $U$ for the respondents in the Gibbs sampler are conditional on the observed $Y$ and thus the resulting draw will always be $\bar{y}_R$. To avoid this, we can take one of two approaches. An obvious extension is to redraw the latent $U$ conditional only on the current draws of the proxy and the parameters, with the subsequent draw of $\mu_y^{(0)}$ is taken to be $\mu_y^{(0)} = \frac{1}{n-r}\sum_{i=r+1}^{n} I(U_i > 0)$. The drawback of this method (Modification 1) is that variances may actually be overestimated since we are essentially imputing the observed binary outcome $Y$ for the respondents. Alternatively, we can use the average of the predicted probabilities for the respondents as a draw of $\mu_y^{(0)}$, i.e.

$\frac{1}{r}\sum_{i=1}^{r}\Phi^{-1}(X_i)$. This is actually a draw of the conditional mean of $Y$ (conditional on $X$) and so its posterior distribution will underestimate the variance of $\mu_y^{(0)}$. To combat this we take a bootstrap sample of the $X_i$ before calculating the mean of the predicted probabilities (Modification 2).

## 5.5 Extension to Ordinal Outcome

Suppose instead of a binary outcome we observe a partially missing ordinal outcome $Y$, where $Y_i$ takes one of $J$ ordered values, $1, \dots, J$. As with the binary case we assume there is an underlying latent continuous variable $U$, related to the observed $Y$ through the rule that $Y = j$ if $\gamma_{j-1} < U < \gamma_j$ for $j = 1, \dots, J$, with $\gamma_0 = -\infty$ and $\gamma_J = \infty$. This latent structure motivates an extension of probit regression to ordinal outcomes (e.g. Agresti, 2002, chap. 7), which we apply to the respondent data:

$$\Pr(Y \leq j|Z, M=0) = \Pr(U \leq \gamma_j) = \Phi(\gamma_j + \alpha Z). \tag{5.11}$$

We take $X = \hat{\alpha}Z$ to be the proxy, noting that the intercepts $\{\gamma_j\}$ are the cutpoints of the latent variable $U$ and are not used in the construction of the proxy. As with the binary $Y$ we apply the proxy pattern-mixutre model (5.4) to the joint distribution of the proxy $X$ and latent $U$, with assumption (5.5) on the missing data mechanism to make the model identifiable. Of interest are the marginal probabilities that $Y = j$ for $j = 1, \dots, J$, averaged across missing data patterns, given by:

(5.12)

$$\begin{aligned}\Pr(Y=j) &= \pi\Pr(\gamma_{j-1} < U \leq \gamma_j|M=0) + (1-\pi)\Pr(\gamma_{j-1} < U \leq \gamma_j|M=1)\\ &= \pi\left[\Phi\left(\frac{\gamma_j - \mu_u^{(0)}}{\sqrt{\sigma_{uu}^{(0)}}}\right) - \Phi\left(\frac{\gamma_{j-1} - \mu_u^{(0)}}{\sqrt{\sigma_{uu}^{(0)}}}\right)\right]\\ &\quad + (1-\pi)\left[\Phi\left(\frac{\gamma_j - \mu_u^{(1)}}{\sqrt{\sigma_{uu}^{(1)}}}\right) - \Phi\left(\frac{\gamma_{j-1} - \mu_u^{(1)}}{\sqrt{\sigma_{uu}^{(1)}}}\right)\right]\end{aligned}$$

### 5.5.1 Estimation Methods

Resulting maximum likelihood estimates of the parameters $\mu_u^{(1)}$ and $\sigma_{uu}^{(1)}$ have the same form (5.7) as in the binary case. The ML estimates of the parameters of the distribution of $X$ for respondents and nonrespondents are the usual estimators. This leaves $\mu_u^{(0)}$, $\sigma_{uu}^{(0)}$, $\rho^{(0)}$, and $\gamma = \{\gamma_j\}$ to be estimated. Without loss of generality we take $\mu_u^{(0)} = 0$ and $\sigma_{uu}^{(0)} = 1$ and obtain MLEs for the correlation $\rho^{(0)}$ and cutpoints $\gamma$. This reduces to the problem of estimating the polyserial correlation between the ordinal $Y$ and continuous $X$, first considered by Cox (1974). As with the binary case, there is no closed-form solution and an iterative solution is required. The MLE of the marginal probabilities of $Y$ are obtained by substituting these estimates into (5.12). As with the binary case, the ML estimate of $\Pr(Y = j|M = 0)$ is not $\frac{1}{r}\sum_{i=1}^{r} I(y_i = j)$. As an alternative, the *two-step method* of Olsson et al. (1982) estimates $\gamma$ with the inverse normal distribution function evaluated at the the cumulative proportions of $Y$ for the respondents. The ML estimate of $\rho^{(0)}$ is then obtained by maximizing the likelihood conditional on these estimates of $\hat{\gamma}$. As with the binary case, the two-step method is appealing because the estimates of $\Pr(Y = j|M = 0)$ will be the sample proportions. Large sample variances are obtained for full ML through Taylor series expansion, and via the bootstrap for the two-step estimator.

Bayesian estimation for ordinal $Y$ is similar to the binary case. With noninformative priors on the regression parameters $\alpha$ we again use the data augmentation approach of Albert and Chib (1993) to draw the latent $U$ for respondents and apply the continuous PPM model to the latent $U$ and proxy $X$. The posterior distribution of $U$ given $\alpha$ (or equivalently the proxy $X = \alpha Z$) and $\gamma$ is given by a truncated normal distribution,

$$(U|Y = j, \alpha, \gamma, M = 0) \sim N(X, 1) \quad \text{truncated at the left (right) by } \gamma_{j-1}(\gamma_j). \tag{5.13}$$

The conditional posterior distribution of $\alpha$ is multivariate normal as before, given by (5.9). The posterior distribution of $\gamma_j$ given $U$ and $\alpha$ is uniform on the interval $[\max\{\max\{U_i : Y_i = j\}, \gamma_{j-1}\}, \min\{\min\{U_i : Y_i = j+1\}, \gamma_{j+1}\}]$. Application of the continuous PPM model to the latent data proceeds in an iterative fashion as in Section 5.4.2, with draws from the posterior distribution of the marginal probabilities of $Y$ obtained by substituting draws from the PPMA algorithm into (5.12).

Multiple imputation for the ordinal outcome is a straightforward extension of the Bayesian method. For a chosen $\lambda$, draws of the latent $U$ for nonrespondents are obtained in the same manner as the binary case, using (5.10). The missing $Y$ are then imputed as $Y_i = I(\gamma_{j+1} < U_i \leq \gamma_j)$.

As with the binary case, the model is sensitive to an incorrect specification of the distribution of the proxy $X$. The two-step ML estimation and multiple imputation are less sensitivity to deviation away from normality, but both the full ML estimation and Bayesian method will produce biased results, even for estimates of the respondent probabilities. We propose similar modifications to the Bayesian method as in the binary case. We draw the latent $U$ for nonrespondents conditional on the current parameter values and a draw of $\Pr(Y = j|M = 1)$ is taken to be $\frac{1}{n-r}\sum_{i=r+1}^{n} I(\gamma_{j+1} < U_i \leq \gamma_j)$ for $j = 1, \ldots, J$. Again there are two methods of obtaining draws of the respondent probabilities. In the first method we redraw $U$ for the respondents, not conditioning on $Y$, and estimate probabilities as for the nonrespondents. The second method draws a bootstrap sample of respondents and uses the average over the $r$ respondents of the predicted probabilities given the current $\alpha$ (for each of the $J$

categories) as draws of the posterior probabilities.

## 5.6 Simulation Studies

We now describe a set of simulation studies designed to (1) illustrate the effects of $\rho$, $d^*$, and sample size on PPMA estimates of the mean of a binary outcome $Y$, (2) assess confidence coverage of ML, Bayes and MI inferences when model assumptions are met, and (3) assess confidence coverage of the various estimation methods when the normality assumption is incorrect. All simulations and data analysis were performed using the software package R (R Development Core Team, 2007).

### 5.6.1 Numerical Illustration of Binary PPMA

Our first objective with the simulation studies was to numerically illustrate the taxonomy of evidence concerning bias based on the strength of the proxy and the deviation of its mean. We created a total of eighteen artificial data sets in a 3x3x2 factorial design with a fixed nonresponse rate of 50%. A single data set was generated for each combination of $\rho = \{0.8, 0.5, 0.2\}$, $d^* = \{0.1, 0.3, 0.5\}$ and $n = \{100, 400\}$ as follows. A single covariate $Z$ was generated for both respondents and nonrespondents, with $z_i \sim N(0, 1), i = 1, \ldots, r$ for respondents and $z_i \sim N(d^*/(1 - r/n), 1), i = r + 1, \ldots, n$ for nonrespondents. For respondents only, a latent variable $u_i$ was generated as $[u_i|z_i] \sim N(a_0 + a_1 z_i, 1)$, with an observed binary $Y$ then created as $y_i = 1$ if $u_i > 0$. We set $a_1 = \rho/\sqrt{1-\rho^2}$ so that $\text{Corr}(Y, X|M = 0) = \rho$ and choose $a_0 = \Phi^{-1}(0.3)\sqrt{1 + a_1^2}$ so that the expected value of $Y$ for respondents was 0.3. In this and all subsequent simulations the latent variable $U$ was used for data generation and then discarded; only $Y$ and $Z$ were used for the proxy pattern-mixture analysis.

For each of the eighteen data sets, estimates of the mean of $Y$ and its variance were obtained for $\lambda = (0, 1, \infty)$. For each value of $\lambda$, three 95% intervals were calculated:

(a) ML: the (full) maximum likelihood estimate ± 2 standard errors (large-sample approximation),

(b) PD: the posterior median and 2.5th to 97.5th posterior interval based on 2000 cycles of the Gibbs sampler as outlined in Section 5.4.2, with a burn-in of 20 iterations,

(c) MI: mean ± 2 standard errors from 20 multiply imputed data sets, with a burn-in of 20 iterations and imputing on every hundredth iteration of the Gibbs sampler.

The two-step ML estimator and two modifications to the Bayes estimator to handle non-normal proxies were also calculated. Since the simulated covariate data were normally distributed, the modified estimators yield similar results and are not shown. The complete case estimate (± 2 standard errors) was also computed for each data set.

#### Results

Figure 5.1 shows the resulting 95% intervals using each of the three estimation methods for the nine data sets with $n = 400$, plotted alongside the complete case estimate. The relative performances of each method for the data sets with $n = 100$ are similar to the results with $n = 400$ (with larger interval lengths); results are not shown. We note that in this simulation the true mean of $Y$ is not known; we simply illustrate the effect of various values of $\rho$ and $d^*$ on the sensitivity analysis and compare the different estimation methods.

For populations with strong proxies ($\rho = 0.8$), ML, PD, and MI give nearly identical results. For these populations there is not a noticable increase in the length of the intervals as we move from $\lambda = 0$ to $\lambda = \infty$, suggesting that even in the case of a large deviation ($d^* = 0.5$) there is good information to correct the potential bias.

For weaker proxies we begin to see differences among the three methods. When $\lambda = 0$ (MAR) the three methods yield similar inference, but for nonignorable mechanisms the intervals for PD and MI tend to be wider than those for ML. For both Bayesian methods (PD, MI) the interval width increases as we move from $\lambda = 0$ to $\lambda = \infty$, with a marked increase in length when $\rho = 0.2$. The ML estimate displays different behaviour; its intervals actually get very small for the weak proxies and large $d$. This is due to the unstable behaviour of the MLE near the boundary of the parameter space. For weak proxies (small $\rho$), the MLE of $\sigma_{uu}^{(1)}$ as given in (5.7) can be zero or negative if the nonrespondent proxy variance is smaller than the respondent variance. If it is negative, we set $\hat{\sigma}_{uu}^{(1)} = 0$. Since the MLE of the mean of $Y$ is given by $\mu_y^{(1)} = \Phi\left(\mu_u^{(1)}/\sqrt{\sigma_{uu}^{(1)}}\right)$, a zero value for $\sigma_{uu}^{(1)}$ causes $\hat{\mu}_y^{(1)}$ to be exactly 0 or 1 depending on the sign of $\mu_u^{(1)}$. The large sample variance will then be small since the estimate of $\sigma_{uu}^{(1)}$ is zero, and interval widths will be small relative to the PD or MI intervals.

Since the outcome is binary, we obtain a natural upper and lower bound for the mean of $Y$ by filling in all missing vales with zeros or all with ones. These bounds are shown in dotted lines in Figure 5.1. For strong proxies, even with a large deviation this upper bound is not reached, suggesting that even in the worst-case NMAR scenario where missingness depends entirely on the outcome the overall mean would not be this extreme. However, for the weakest proxy ($\rho = 0.02$) we see that even for the smallest deviation the intervals for PD and MI cover these bounds. This is due to the weak information about $Y$ contained in the proxy. The PD intervals are highly skewed and the MI intervals are exaggerated in length. The posterior distribution of $\mu_y$ is bimodal, with modes at each of the two bounds obtained when all missings are zeros or all ones. Thus the posterior interval essentially covers the entire range of possible values of $\mu_y$. Similarly for MI the imputed data sets have imputed values that are either all zeros or all ones. This causes very large variance and thus large intervals, and since by construction the intervals are symmetric for MI, they are even larger than the posterior intervals from PD. As previously discussed, the ML method gives extremely small intervals for the weak proxies, with the point estimate at the upper bound.

### 5.6.2 Confidence Coverage, Normally Distributed Proxy

The second objective of the simulation was to assess coverage properties for each of the three estimation methods when model assumptions are met, i.e. when the proxy is normally distributed. We generated data using the same set-up as Section 5.6.1. We fixed $d^* = 0.3$ and varied $\rho = \{0.8, 0.5, 0.2\}$ and $n = \{100, 400\}$ for a total of six populations, and generated 500 replicate data sets for each population. For each population we applied the proxy pattern-mixture model using each of $\lambda = \{0, 1, \infty\}$, with the assumption that the assumed value of $\lambda$ equals the actual value of $\lambda$. This lead to a total of eighteen hypothetical populations, and for each we computed the actual coverage of a nominal 95% interval and median interval length. We also calculated the relative empirical bias for each estimator. Assuming that the $\lambda$ value is correct is unrealistic, but coverages are clearly not valid when the value of $\lambda$ is misspecified, and uncertainty in the the choice of $\lambda$ is captured by the sensitivity analysis.

A total of six estimators for the mean of the binary outcome $Y$ and its variance were obtained for each of the eighteen data sets. These included the usual maximum likelihood (ML Full), posterior distribution (PD A), and multiple imputation (MI) estimators as in the previous section, as well as the three modified estimators: the two-step maximum likelihood estimator (ML 2-step) and two modifications to the Bayesian estimator as described in Section 5.4.4 (PD B, PD C). Confidence intervals for the two-step ML estimator were based on 500 bootstrap samples. Posterior intervals for all three PD methods were based on 1000 draws from the Gibbs sampler as the chains were quick to converge.

#### Results

Table 5.1 displays the average empirical relative bias, nominal coverage, and median CI width for the eighteen populations. For the smaller sample size ($n = 100$), all methods suffer from slight undercoverage, even when the proxy is strong. This undercoverage is exaggerated in the populations with the weakest proxy ($\rho = 0.2$) and when $\lambda = \infty$, where all the methods are negatively biased. With 50% nonresponse, these small samples have only 50 observed data points, and estimation of the distribution of the latent variable from the binary observations is challenging. No method displays consistently better performance in the small sample size, though the larger interval lengths of PD B (redrawing the latent $U$ for nonrespondents) and MI yield slightly improved performance.

Differences between the methods emerge with the larger sample size ($n = 400$). All methods perform well when the proxy is strong ($\rho = 0.8$), though the second modification to the Bayesian method (PD C — bootstrapping the predicted probabilities) consistently shows a small amount of undercoverage. As expected, the interval widths for the alternative modification to the Bayesian method (PD B) are wider than the usual PD method (PD A), with PD B actually overcovering for several populations, most notably when $\lambda = 0$ or 1. There does not seem to be much difference between the two ML methods for any of the populations, though for the smaller sample size we see slightly wider confidence intervals for the two-step method.

As was evident in the previous simulation, when $\rho = 0.2$ and $\lambda = \infty$ the confidence interval length for ML Full is much smaller than any of the other methods, and this leads to slight undercoverage.

### 5.6.3 Confidence Coverage, Non-Normally Distributed Proxy

As a final objective for the simulation study we wanted to assess the performance of the modifications to the maximum likelihood and Bayesian estimation methods for binary $Y$ when the normality assumption of the proxy was violated. Since by definition this is a situation where the model is violated, we cannot generate data as in the previous two sets of simulations. Instead, complete data were generated in a selection model framework and missingness was induced via different missingness mechanisms. The sample size was fixed at $n = 400$ since the previous simulation showed difficulty in distinguishing performances of the methods for smaller $n$. Three different distributions for a single covariate $Z$ were selected: (a) Normal$(0, 1)$, (b) Gamma$(4, 0.5)$, (c) Exponential$(1)$. These distributions were chosen to evaluate the effect of both moderate skew (Gamma) and severe skew (Exponential). The normally distributed covariate was chosen to serve as a reference; the selection model implies marginal normality, while the PPM model assumes conditional normality, so even with a normally distributed covariate the distributional assumptions of the PPM model are violated.

Data were generated as follows. For each of the three $Z$ distributions the covariate $z_i, i = 1, \ldots, n$ was generated. Then for each of $\rho = \{0.8, 0.5, 0.2\}$ the latent $u_i$ was generated from $[u_i|z_i] \sim N(a_0 + a_1 z_i, 1)$, where $a_1 = \rho/\sqrt{1-\rho^2}$ so that $\text{Corr}(Y, X) = \rho$. Values of $a_0$ were chosen so that the expected value of $Y$ was approximately 0.3, where the binary outcome $Y$ was then created as $y_i = 1$ if $u_i > 0$. The missing data indicator $m_i$ was generated according to a logistic model,

$$\text{logit}(\Pr(m_i = 1|u_i, z_i)) = \gamma_0 + \gamma_Z z_i + \gamma_U u_i,$$

and values of $y_i$ were deleted when $m_i = 1$. The two different mechanisms selected were MAR, with $\gamma_Z = 0.5, \gamma_U = 0$, and extreme NMAR, with $\gamma_Z = 0, \gamma_U = 0.5$. Aside from the discrepancy of marginal versus conditional normality, these two mechanisms correspond to $\lambda$ values of 0 and $\infty$, respectively. For both scenarios, values of $\gamma_0$ were selected to induce approximately 50% missingness.

The process of generating $\{z_i, u_i, y_i, m_i\}$, and inducing missingness was repeated 500 times for each of the eighteen populations. The same six estimators for the mean of the binary outcome $Y$ and its variance were obtained for each of the eighteen data sets as in the previous section. For the MAR mechanism, $\lambda$ was taken to be zero, and for NMAR $\lambda = \infty$.

**Results**

When $Z$ is normally distributed, results are similar to the previous simulation, as seen in Table 5.2a. All methods are unbiased across all scenarios except when $\rho = 0.2$ under NMAR. For this population there is a small bias but all methods except ML Full still achieve nominal coverage, and in fact many show higher than nominal coverage. The consistently best performing methods are ML 2-step, PD B, and MI, which reach nominal coverage in all scenarios. PD C shows undercoverage (as in the previous simulation) when the proxy is strong, and also slight undercoverage under MAR. As was previously seen, ML Full has intervals that are too short when missingness is not at random and the proxy is weak ($\rho = 0.2$), and thus exhibits very poor coverage. The two-step ML fixes this problem, since the bootstrap is used for variance estimation instead of the large-sample approximation, though the intervals are nearly twice as long as other methods that actually show overcoverage.

Table 5.2b shows results for the slightly skewed proxy, when $Z$ has a Gamma distribution. The methods that rely the most on the underlying normality assumption of the PPMA, ML Full and PD A, show bias for the stronger proxies under both missingness mechanisms and hence tend to undercover. When missingness is at random, as before the best performers are ML 2-step, PD B, and MI, with PD C showing undercoverage. The more difficult populations are under NMAR. For both $\rho = 0.8$ and $\rho = 0.5$ all methods exhibit some bias, though ML Full and PD A are the most biased, and subsequently all methods fail to acheive nominal coverage. The exception is MI, which is at nominal coverage for all but one scenario. For the weakest proxy ($\rho - 0.2$) ML Full again shows undercoverage, while the two-step ML corrects this problem. However, it does so with very large confidence intervals relative to the Bayesian methods which reach nominal coverage.

Results for $Z$ having an Exponential distribution are displayed in Table 5.2c. The results are similar to the Gamma case, with larger biases and lower coverage rates across all populations. With a severely skewed proxy, the PPM model actually performs the worst with a strong proxy; under both MAR and NMAR it is difficult for any estimation method to reach nominal coverage. As the strength of the proxy weakens, under MAR ML 2-step and PD B reach nominal coverage, while the unmodified ML Full and PD A methods remain biased and have poor coverage.

Overall, the best performing method is MI, which achieves nominal or just under nominal coverage for all three distributions of $Z$, including the severely skewed Exponential, and under both missingness mechanisms with all strengths of proxies. This result is not surprising. Even though MI uses the fully parametric PPM model to generate posterior draws of the parameters, these draws are subsequently used to impute the missing $Y$ values via the conditional distribution of $[U|X, M = 1]$. Even if the proxy is not normally distributed, the conditional distribution of the latent variable given the proxy is normal by definition, and so MI should be the least sensitive to departures away from normality in the proxy.

The one other method that does reasonably well in most scenarios is the first modification to the Bayesian draws, PD B. As with MI, this method conditions on the proxy and draws the latent $U$ and thus outperforms the unmodified Bayesian method that relies entirely on the joint normality of $U$ and the proxy $X$. PD B achieves at or near nominal coverage for strong proxies across all levels of skewness, but exhibits overcoverage for weaker proxies. This is to be expected, since in this modification the latent $U$ for respondents are redrawn unconditional on the observed $Y$, which is effectively imputing the observed $Y$, and certainly has the potential to add unnecessary variability, as was noted in Section 5.4.4.

## 5.7 Application

The third National Health and Nutrition Examination Survey (NHANES III) was a large-scale stratified multistage probability sample of the noninstitutionalized U.S. population conducted during the period from 1988 to 1994 (U.S. Department of Health and Human Services, 1994). NHANES III collected data in three phases: (a) a household screening interview, (b) a personal home interview, and (c) a physical examination at a mobile examination center (MEC). The total number of persons screened was 39,695, with 86% (33,994) completing the second phase interview. Of these, only 78% were examined in the MEC. Since the questions asked at both the second and third stage varied considerably by age we chose to select only adults age 17 and older who had completed the second phase interview for the purposes of our example, leaving a sample size of 20,050.

We selected two binary variables and one ordinal variable: an indicator for low income, defined as being below the poverty threshold, an indicator for hypertension, defined as having a systolic blood pressure above 140 mmHg, and a three-level bone mineral density (BMD) variable (normal / osteopenia / osteoporosis). The nonresponse rates for these three items were 15%, 11%, and 22% respectively. In order to reflect nonresponse due to unit nonresponse at the level of the MEC exam we chose to only include fully observed covariates to create the proxies; variables that were fully observed for the sample included age, gender, race, and household size. The (log-transformed) design weight was also used as a covariate in creating the proxies. The final models were chosen with backwards selection starting from a model that contained all second-order interactions.

Both hypertension and BMD had strong proxies and relatively large deviations, with $\hat{\rho} = 0.67$ and $d* = 0.065$ for hypertension and $\hat{\rho} = 0.63$ and $d = -0.064$ for BMD. Income had a slightly weaker proxy, with $\hat{\rho} = 0.47$, but also a smaller deviation with $d = 0.035$.

For each outcome, estimates of the probilities and confidence intervals for $\lambda = (0, 1, \infty)$ were obtained using maximum likelihood (ML), 1000 draws from the posterior distribution with a burn-in of 20 draws (PD), and multiple imputation with $K = 20$ data sets (MI), again with a burn-in of 20 draws and imputing on every hundredth iteration. Additionally, since NHANES III has a complex survey design we obtained estimates using multiple imputation with design-based estimators of the mean using the survey weights (MI wt). Design-based estimators were computed using the "survey" routines in R, which estimate variances using Taylor series linearizations (Lumley, 2004).

Estimated proportions and confidence intervals are displayed in Figures 5.2, 5.3, and 5.4. The intervals for weighted MI are larger than those for any of the non-design-adjusted methods, and for all three outcomes there is also a shift in the mean estimates for the weighted estimators, consistent for all values of $\lambda$, reflecting the impact on these outcomes of the oversampling in NHANES of certain age and ethnic groups. The deviations are not negligble for any of the three outcomes, as evidenced by the shift in the estimates as we move from $\lambda = 0$ to $\lambda = \infty$. However, all three outcomes have moderately strong proxies, so the width of confidence intervals even in the extreme case of $\lambda = \infty$ are not inflated too much above the length of the intervals under MAR ($\lambda = 0$).

For both hypertension and BMD we see a difference in the estimates for full maximum likelihood (ML Full) and the unmodified Bayesian method (PD A) compared to all the other estimators. The distribution of the proxies for each of the three outcomes is shown in Figure 5.5, separately for respondents and nonrespondents. We can see that the proxies for both hypertension and BMD are skewed, while the proxy for income does not appear to be exactly normally distributed but is basically symmetric. The sensitivity of the full ML and Bayesian method to non-normality is an issue of skewness. These deviations from symmetry have the effect of shifting mean estimates considerably, as seen in Figures 5.2 and 5.4. Though we do not know the true proportions, since the modified methods condition on the proxy when estimating the proportions and yield the respondent proportion as the respondent means, for these skewed proxies using the modified Bayesian methods, multiple imputation, or the two-step ML estimator seems to be the wisest choice.

The two modifications to the Bayesian method, labeled PD B and PD C in the figures, do not yield identical inference. In particular the first modification, redrawing the latent $U$ for respondents, seems to be overestimating variance relative to the two-step ML estimator (ML 2step) and multiple imputation estimator (MI). Conversely, the modifcation that bootstraps the predicted probabilities seems to be slightly underestimating variability.

## 5.8 Discussion

In this paper we have extended the previously developed proxy pattern-mixture analysis to handle binary and ordinal data, which are ubiquitous in sample survey data. As with a continuous outcome, this novel method integrates the three key components that contribute to nonresponse bias: the amount of missingness, differences between respondents and nonrespondents on characteristics that are observed for the entire sample, and the relationship between these fully observed covariates and the survey outcome of interest. The analysis includes but does not assume that missingness is at random, allowing the user to investigate a range of non-MAR mechanisms and the resulting potential for nonresponse bias. For the binary case, it is common to investigate what the estimates would be if all nonresponding units were zeros (or ones), and in fact the binary PPMA produces these two extremes when the proxy is weak.

An attractive feature of the continuous outcome PPMA is its ease of implementation; a drawback of the extension to binary (and ordinal) outcomes is a loss of some of this simplicity. By introducing a latent variable framework we reduce the problem to one of applying the continuous PPMA to a latent variable, but since this underlying continuous latent variable is unobserved even for nonrespondents, application is more complex. Closed-form solutions are no longer available for the maximum likelihood approach, and Bayesian methods require iteration using Gibbs sampling. However, the ML solutions are good starting points for the Gibbs sampler and only very short burn-in periods are required.

An additional level of complexity in the binary and ordinal case is the effect of skewed proxies. Where the continuous PPMA is relatively robust to departures from bivariate normality in the proxy and outcome, the binary and ordinal cases rely heavily on the normality assumption. The assumption of normality of the proxy is crucial and even slight deviations away from normality will cause biased results. To relax the dependence on the normality assumption we introduced modified estimators that appear to not only perform better when the normality assumption is violated but also maintain good performance if the normality assumption holds.

We have described three different estimation methods for the categorical PPMA, maximum likelihood, fully Bayesian, and multiple imputation. In our investigations the consistently best performer is multiple imputation, MI does not require a modification to handle skewed proxies, while both the maximum likelihood and Bayesian methods require modified estimators. In addition, incorporation of design weights in estimating the mean is straightforward with MI, as once the model-based imputation is completed a design-based estimator of the mean can be applied in a straightforward manner.

Future work will work to extend PPMA to domain estimation, an important issue in practice. In particular, we are interested in the case where there is a continuous outcome and a binary domain indicator. When the domain indicator is fully observed (for example, gender in the NHANES data), application of the continuous PPM model is straightforward; the domain indicator can be included in the model that creates the proxy, or the entire continuous PPM method can be applied separately for the two domains. The more complex case is when the domain indicator and outcome are jointly missing. We have begun work on this aim, using methods similar to that of Little and Wang (1996), who extend the bivariate pattern-mixture model to the multivariate case when there are two patterns of missingness.

Figure 5.1: 95% intervals for nine generated data sets ($n = 400$) for $\lambda = (0, 1, \infty)$. Numbers below intervals are the interval length. Dotted lines are the point estimates obtained by filling in all ones or all zeros for missing values. CC: Complete case; ML: Maximum likelihood; PD: Posterior distribution; MI: 20 multiply imputed data sets.

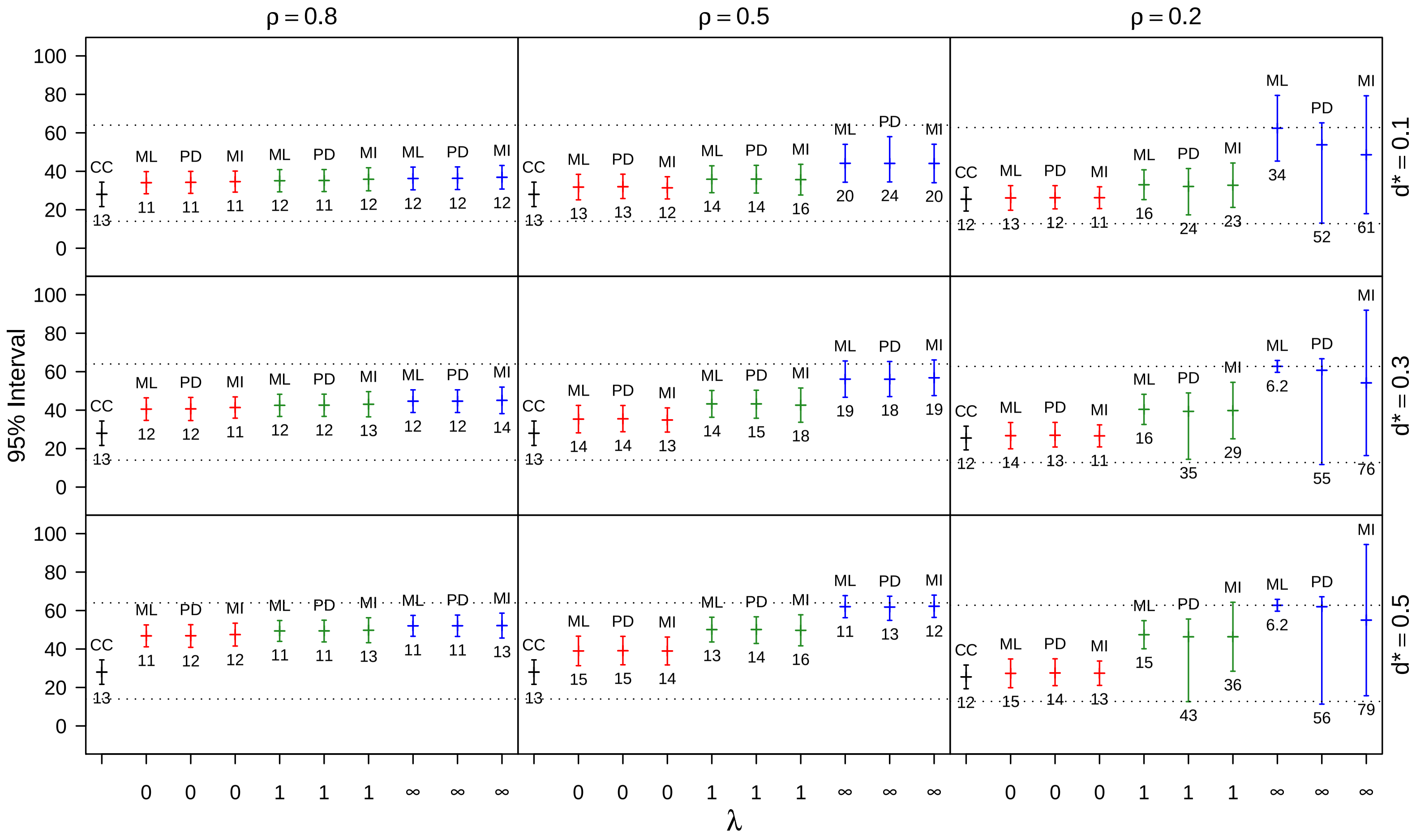


Figure 5.2: Estimates of the proportion hypertensive for $\lambda = (0, 1, \infty)$ based on NHANES III adult data. Numbers below intervals are the interval length. CC: Complete case; CC wt: Complete case with estimation incorporating the survey design; ML Full: Maximum likelihood; ML 2step: Two-step Maximum likelihood; PD A: Posterior distribution; PD B: Modification 1 to Bayesian method; PD C: Modification 2 to Bayesian method; MI: 20 multiply imputed data sets; MIwt: 20 multiply imputed data sets with estimation incorporating the survey design.

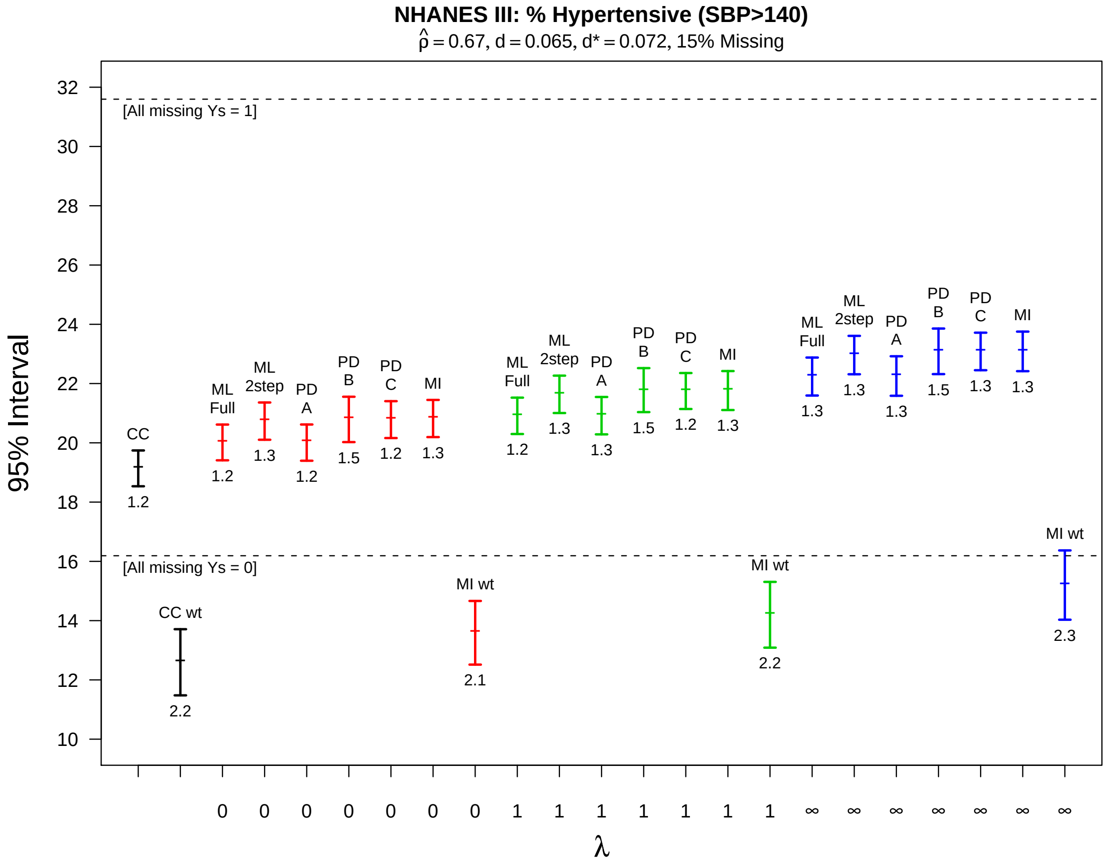


Figure 5.3: Estimates of proportion low income for $\lambda = (0, 1, \infty)$ based on NHANES III adult data. Numbers below intervals are the interval length. CC: Complete case; CC wt: Complete case with estimation incorporating the survey design; ML Full: Maximum likelihood; ML 2step: Two-step Maximum likelihood; PD A: Posterior distribution; PD B: Modification 1 to Bayesian method; PD C: Modification 2 to Bayesian method; MI: 20 multiply imputed data sets; MIwt: 20 multiply imputed data sets with estimation incorporating the survey design.

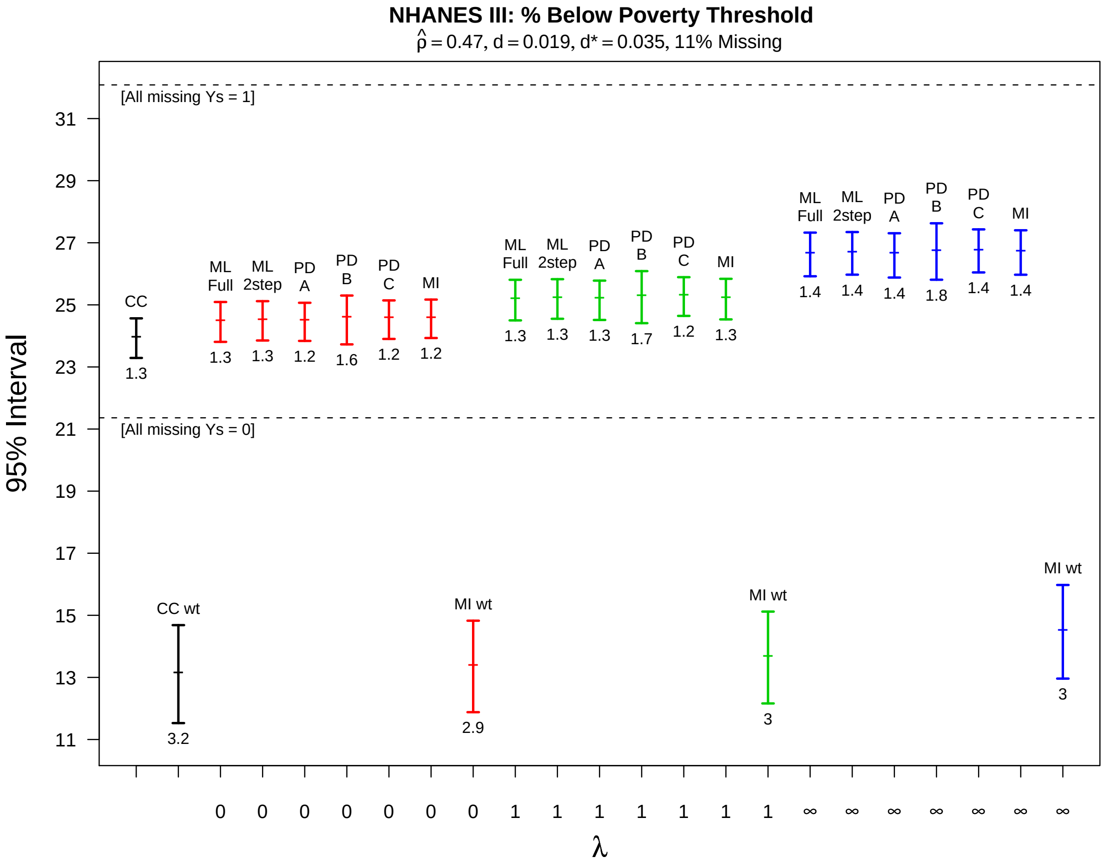


Figure 5.4: Estimates of proportions in each of three categories of BMD for $\lambda = (0, 1, \infty)$ based on NHANES III adult data. Numbers below intervals are the interval length. CC: Complete case; CC wt: Complete case with estimation incorporating the survey design; ML Full: Maximum likelihood; ML 2step: Two-step Maximum likelihood; PD A: Posterior distribution; PD B: Modification 1 to Bayesian method; PD C: Modification 2 to Bayesian method; MI: 20 multiply imputed data sets; MIwt: 20 multiply imputed data sets with estimation incorporating the survey design.

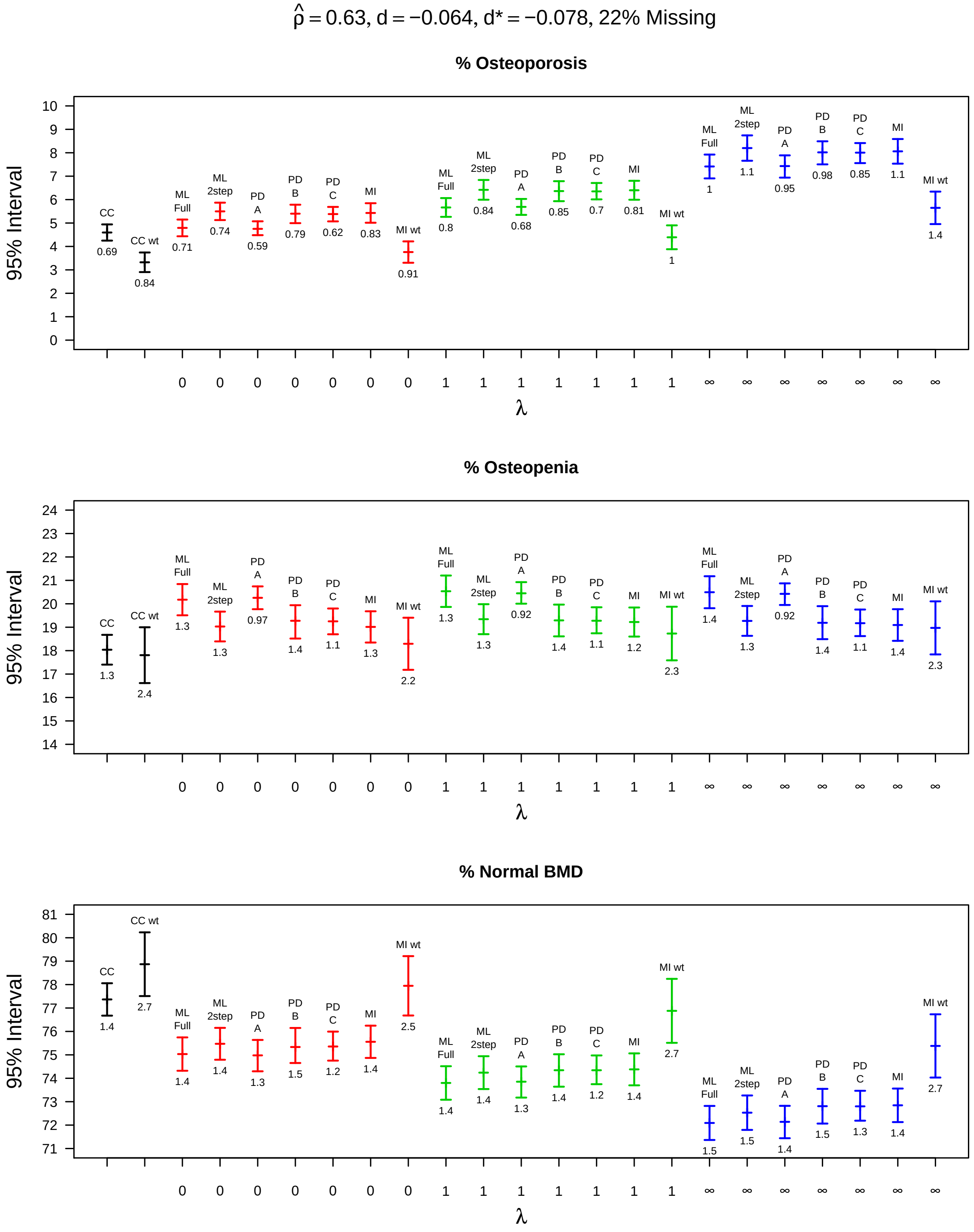


Figure 5.5: Distribution of the respondent and nonrespondent proxies for each of the three outcomes, based on NHANES III adult data. Superimposed line is the normal distribution with mean and variance equal to the sample values.

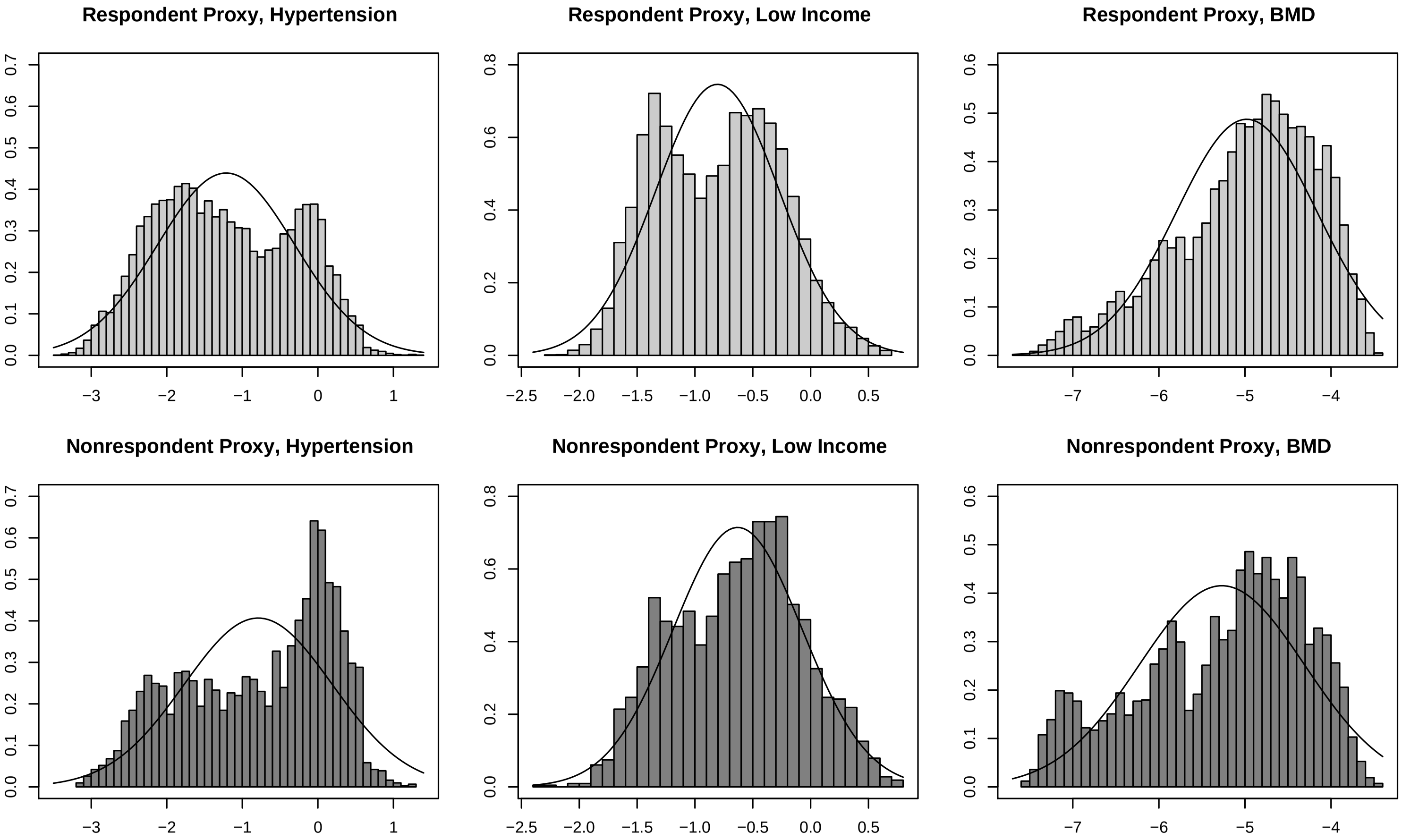


Table 5.1: Average relative empirical bias, 95% interval coverage and median interval length for eighteen artificial populations with $d^* = 0.3$ and (a) $\rho = 0.8$; (b) $\rho = 0.5$; (c) $\rho = 0.2$. ML Full: Maximum likelihood; ML 2-step: modified maximum likelihood; PD A: Posterior distribution; PD B: Modification 1 to PD; PD C: Modification 2 to PD; MI: 20 multiply imputed data sets. Results over 500 replicates.

(a) $\rho = 0.8,\ d^* = 0.3$

| | | $n = 100$ | | | $n = 400$ | | |
|---|---|---|---|---|---|---|---|
| $\lambda$ | Method | Relative Bias (%) | Coverage (%) | CI Width | Relative Bias (%) | Coverage (%) | CI Width |
| 0 | ML Full | -0.6 | **91.2** | 0.24 | -0.4 | 94.0 | 0.12 |
| | ML 2-step | -0.2 | 93.8 | 0.25 | -0.3 | 96.0 | 0.13 |
| | PD A | 0.0 | **92.4** | 0.23 | -0.2 | 94.2 | 0.12 |
| | PD B | -0.1 | **92.4** | 0.24 | -0.3 | 94.8 | 0.13 |
| | PD C | -0.4 | **89.8** | 0.22 | -0.4 | **92.2** | 0.11 |
| | MI | -0.8 | **91.0** | 0.23 | -0.2 | 94.0 | 0.12 |
| 1 | ML Full | -0.8 | **92.2** | 0.24 | -0.3 | 94.0 | 0.12 |
| | ML 2-step | -0.3 | 93.6 | 0.25 | -0.2 | 94.8 | 0.13 |
| | PD A | -0.7 | **92.2** | 0.23 | -0.3 | 93.8 | 0.12 |
| | PD B | -0.8 | **92.4** | 0.25 | -0.4 | 94.8 | 0.13 |
| | PD C | -0.9 | **90.2** | 0.22 | -0.4 | **92.4** | 0.11 |
| | MI | -2.6 | **91.0** | 0.23 | -1.2 | 93.2 | 0.12 |
| $\infty$ | ML Full | -0.7 | **92.6** | 0.25 | -0.2 | **93.0** | 0.13 |
| | ML 2-step | -0.2 | 94.6 | 0.27 | 0.0 | 94.8 | 0.13 |
| | PD A | -0.9 | 93.4 | 0.25 | -0.2 | 93.4 | 0.13 |
| | PD B | -1.1 | 93.4 | 0.26 | -0.3 | 94.8 | 0.13 |
| | PD C | -1.0 | **90.6** | 0.24 | -0.3 | **92.2** | 0.12 |
| | MI | -2.3 | **91.6** | 0.25 | -1.4 | **92.2** | 0.13 |

Bolded values are below 1.96 simulation standard errors.
Italicized values are above 1.96 simulation standard errors.

(b) $\rho = 0.5,\ d^* = 0.3$

| | | $n = 100$ | | | $n = 400$ | | |
|---|---|---|---|---|---|---|---|
| $\lambda$ | Method | Relative Bias (%) | Coverage (%) | CI Width | Relative Bias (%) | Coverage (%) | CI Width |
| 0 | ML Full | -0.1 | **92.6** | 0.27 | -0.3 | 94.6 | 0.14 |
| | ML 2-step | 0.1 | 93.2 | 0.28 | -0.2 | 94.2 | 0.14 |
| | PD A | 0.8 | 93.4 | 0.26 | 0.0 | 94.8 | 0.13 |
| | PD B | 0.6 | 95.8 | 0.30 | -0.1 | *97.6* | 0.15 |
| | PD C | 0.2 | **91.8** | 0.25 | -0.2 | 93.2 | 0.13 |
| | MI | -0.5 | **91.2** | 0.26 | 0.7 | **92.6** | 0.13 |
| 1 | ML Full | -0.5 | **91.0** | 0.28 | -0.1 | 94.8 | 0.14 |
| | ML 2-step | -0.4 | **92.0** | 0.28 | -0.1 | 95.2 | 0.14 |
| | PD A | -1.0 | 93.6 | 0.28 | -0.1 | 95.8 | 0.14 |
| | PD B | -1.2 | 96.2 | 0.32 | -0.2 | *97.4* | 0.16 |
| | PD C | -1.2 | **92.0** | 0.27 | -0.2 | 94.2 | 0.14 |
| | MI | -3.4 | **91.6** | 0.27 | -1.0 | 96.0 | 0.15 |
| $\infty$ | ML Full | -1.7 | **91.0** | 0.34 | 1.1 | 94.4 | 0.19 |
| | ML 2-step | 0.7 | 94.0 | 0.39 | 1.6 | **93.0** | 0.20 |
| | PD A | -4.7 | 95.2 | 0.33 | 0.3 | 96.8 | 0.19 |
| | PD B | -4.5 | 96.6 | 0.36 | 0.5 | 96.8 | 0.20 |
| | PD C | -4.5 | 94.6 | 0.33 | 0.4 | 96.2 | 0.19 |
| | MI | -5.7 | 94.2 | 0.35 | -0.3 | 94.8 | 0.20 |

Bolded values are below 1.96 simulation standard errors.
Italicized values are above 1.96 simulation standard errors.

(c) $\rho = 0.2,\ d^* = 0.3$

| | | $n = 100$ | | | $n = 400$ | | |
|---|---|---|---|---|---|---|---|
| $\lambda$ | Method | Relative Bias (%) | Coverage (%) | CI Width | Relative Bias (%) | Coverage (%) | CI Width |
| 0 | ML Full | 0.0 | **93.0** | 0.28 | -0.2 | 95.0 | 0.14 |
| | ML 2-step | 0.0 | 93.2 | 0.28 | -0.2 | 94.8 | 0.14 |
| | PD A | 1.3 | 94.2 | 0.26 | 0.1 | 94.4 | 0.13 |
| | PD B | 1.0 | *97.2* | 0.31 | 0.1 | *97.4* | 0.16 |
| | PD C | 0.7 | 93.2 | 0.25 | 0.0 | 93.4 | 0.13 |
| | MI | -0.7 | **93.0** | 0.26 | 1.6 | 94.0 | 0.13 |
| 1 | ML Full | -7.0 | **80.2** | 0.29 | -0.6 | 93.4 | 0.15 |
| | ML 2-step | -6.9 | **80.8** | 0.30 | -0.6 | 93.4 | 0.15 |
| | PD A | -10 | 95.2 | 0.37 | -1.5 | *97.0* | 0.18 |
| | PD B | -10 | 96.4 | 0.41 | -1.6 | *97.8* | 0.20 |
| | PD C | -10 | 95.2 | 0.36 | -1.5 | 96.8 | 0.17 |
| | MI | -11 | **91.8** | 0.37 | -3.3 | *97.6* | 0.18 |
| $\infty$ | ML Full | -19 | **75.8** | 0.31 | -5.6 | **92.4** | 0.20 |
| | ML 2-step | -18 | **82.0** | 0.66 | -4.7 | 95.2 | 0.30 |
| | PD A | -25 | **80.6** | 0.53 | -8.1 | **91.4** | 0.22 |
| | PD B | -25 | **84.8** | 0.55 | -8.1 | 94.4 | 0.23 |
| | PD C | -25 | **80.6** | 0.54 | -8.0 | **91.2** | 0.22 |
| | MI | -28 | **87.0** | 0.60 | -11 | 94.0 | 0.21 |

Bolded values are below 1.96 simulation standard errors.
Italicized values are above 1.96 simulation standard errors.

Table 5.2: Average relative empirical bias, 95% interval coverage and median interval length for eighteen artificial populations with $n = 400$ and covariate distributions (a) Normal; (b) Gamma; (c) Exponential. ML Full: Maximum likelihood; ML 2-step: modified maximum likelihood; PD A: Posterior distribution; PD B: Modification 1 to PD; PD C: Modification 2 to PD; MI: 20 multiply imputed data sets. Results over 500 replicates.

(a) $Z \sim \text{Normal}(0, 1)$

| | | MAR<br>$\Pr(M = 1 \mid Z, U) = f(Z)$ | | | NMAR<br>$\Pr(M = 1 \mid Z, U) = f(U)$ | | |
|---|---|---|---|---|---|---|---|
| $\rho$ | Method | Relative Bias (%) | Coverage (%) | CI Width | Relative Bias (%) | Coverage (%) | CI Width |
| 0.8 | ML Full | -0.2 | **93.0** | 0.12 | 0 | 93.4 | 0.13 |
| | ML 2-step | -0.1 | 93.6 | 0.12 | 0 | 94.4 | 0.14 |
| | PD A | 0.3 | **92.8** | 0.12 | 0 | 94.0 | 0.13 |
| | PD B | 0.1 | 94.4 | 0.12 | -0.3 | 94.6 | 0.14 |
| | PD C | 0.0 | **91.4** | 0.11 | -0.3 | **92.2** | 0.13 |
| | MI | 0.0 | 93.6 | 0.12 | 0.1 | 93.6 | 0.14 |
| 0.5 | ML Full | 0.1 | 94.4 | 0.13 | -1.6 | **91.6** | 0.20 |
| | ML 2-step | 0.2 | 95.0 | 0.13 | -1.6 | *97.2* | 0.27 |
| | PD A | 0.5 | 93.6 | 0.13 | 0.4 | 95.6 | 0.23 |
| | PD B | 0.4 | 96.6 | 0.15 | 0.2 | 96.4 | 0.24 |
| | PD C | 0.3 | **92.8** | 0.12 | 0.2 | 95.0 | 0.23 |
| | MI | 0.4 | 93.6 | 0.13 | 0.4 | 96.2 | 0.25 |
| 0.2 | ML Full | -0.1 | 93.8 | 0.13 | -1.2 | **57.0** | 0.22 |
| | ML 2-step | -0.1 | 94.4 | 0.13 | -1.2 | 96.0 | 0.57 |
| | PD A | 0.3 | 93.2 | 0.13 | 5.2 | *99.0* | 0.35 |
| | PD B | 0.2 | 96.6 | 0.16 | 5.1 | *99.2* | 0.36 |
| | PD C | 0.1 | **92.4** | 0.12 | 5.0 | *98.8* | 0.35 |
| | MI | 0.3 | 93.4 | 0.13 | 4.1 | *97.8* | 0.37 |

Bolded values are below 1.96 simulation standard errors.
Italicized values are above 1.96 simulation standard errors.

(b) $Z \sim \text{Gamma}(4, 0.5)$

| | | MAR | | | NMAR | | |
|---|---|---|---|---|---|---|---|
| | | $\Pr(M = 1 \mid Z, U) = f(Z)$ | | | $\Pr(M = 1 \mid Z, U) = f(U)$ | | |
| $\rho$ | Method | Relative Bias (%) | Coverage (%) | CI Width | Relative Bias (%) | Coverage (%) | CI Width |
| 0.8 | ML Full | 7.9 | **88.0** | 0.12 | 12 | **78.2** | 0.12 |
| | ML 2-step | 2.7 | 93.4 | 0.12 | 10 | **84.2** | 0.12 |
| | PD A | 8.2 | **85.4** | 0.12 | 12 | **77.4** | 0.12 |
| | PD B | 0.4 | 93.2 | 0.12 | 4.5 | **92.4** | 0.12 |
| | PD C | 0.3 | **90.8** | 0.11 | 4.6 | **89.4** | 0.11 |
| | MI | 0.4 | 93.4 | 0.12 | 4.8 | 93.6 | 0.13 |
| 0.5 | ML Full | 2.3 | **92.4** | 0.13 | 8.3 | **85.0** | 0.15 |
| | ML 2-step | 0.9 | 93.4 | 0.13 | 7.5 | **90.8** | 0.19 |
| | PD A | 2.7 | **91.8** | 0.13 | 8.4 | **84.2** | 0.16 |
| | PD B | 0.5 | 96.6 | 0.15 | 5.2 | **91.8** | 0.17 |
| | PD C | 0.3 | **92.2** | 0.12 | 5.2 | **89.0** | 0.16 |
| | MI | 0.4 | 93.4 | 0.13 | 5.4 | **93.0** | 0.17 |
| 0.2 | ML Full | 0.0 | 94.0 | 0.14 | 1.1 | **68.4** | 0.21 |
| | ML 2-step | -0.1 | 94.8 | 0.14 | 1.0 | 96.2 | 0.43 |
| | PD A | 0.5 | **93.0** | 0.13 | 6.5 | *97.0* | 0.30 |
| | PD B | 0.1 | *97.4* | 0.16 | 5.5 | *97.8* | 0.30 |
| | PD C | 0.0 | **92.4** | 0.13 | 5.5 | 96.8 | 0.29 |
| | MI | 0.2 | 94.4 | 0.13 | 5.4 | 96.4 | 0.31 |

Bolded values are below 1.96 simulation standard errors.
Italicized values are above 1.96 simulation standard errors.

(c) $Z \sim$ Exponential(1)

| | | MAR<br>$\Pr(M = 1 \mid Z, U) = f(Z)$ | | | NMAR<br>$\Pr(M = 1 \mid Z, U) = f(U)$ | | |
|---|---|---|---|---|---|---|---|
| $\rho$ | Method | Relative Bias (%) | Coverage (%) | CI Width | Relative Bias (%) | Coverage (%) | CI Width |
| 0.8 | ML Full | 16 | **66.4** | 0.13 | 21 | **37.4** | 0.11 |
| | ML 2-step | 5.5 | **89.6** | 0.12 | 16 | **56.8** | 0.11 |
| | PD A | 17 | **63.8** | 0.13 | 21 | **36.2** | 0.11 |
| | PD B | 0.4 | **92.6** | 0.12 | 5.2 | **90.6** | 0.11 |
| | PD C | 0.3 | **88.6** | 0.10 | 5.3 | **88.4** | 0.10 |
| | MI | 0.4 | **92.0** | 0.11 | 5.3 | 94.0 | 0.12 |
| 0.5 | ML Full | 4.0 | **92.4** | 0.14 | 14 | **71.4** | 0.13 |
| | ML 2-step | 1.7 | 93.6 | 0.13 | 12 | **81.2** | 0.17 |
| | PD A | 4.5 | **90.4** | 0.13 | 14 | **69.8** | 0.13 |
| | PD B | -0.1 | 96.8 | 0.14 | 6.0 | **91.8** | 0.15 |
| | PD C | -0.1 | **92.8** | 0.12 | 6.1 | **85.8** | 0.13 |
| | MI | 0.0 | 93.4 | 0.13 | 6.2 | 93.2 | 0.15 |
| 0.2 | ML Full | 0.1 | 93.8 | 0.14 | -0.5 | **65.4** | 0.18 |
| | ML 2-step | -0.2 | 94.2 | 0.14 | -0.6 | 94.4 | 0.41 |
| | PD A | 0.5 | 93.4 | 0.13 | 6.6 | 94.8 | 0.25 |
| | PD B | -0.2 | *97.0* | 0.15 | 4.1 | *98.0* | 0.26 |
| | PD C | -0.3 | **92.6** | 0.12 | 4.1 | *97.2* | 0.25 |
| | MI | -0.1 | **93.0** | 0.13 | 4.3 | 96.6 | 0.26 |

Bolded values are below 1.96 simulation standard errors.
Italicized values are above 1.96 simulation standard errors.

# CHAPTER VI

## Summary and Future Work

This thesis described several different problems pertaining to nonresponse in complex sample surveys and suggested novel methods to tackle these problems. The first half of the dissertation (Chapters II and III) dealt with hot deck imputation, a particular method for "filling in the holes" left by missing data in a sample survey. The second half of the dissertation (Chapters IV and V) focused not on a particular method of imputation but rather on evaluating the magnitude of potential bias brought on by the missing data, using a novel analysis method we called a proxy pattern-mixture analysis.

Chaper II contained an extensive review of hot deck imputation. Though survey practitioners tend to favor this simple form of imputation, there is a glaring lack of unifying theory and implementation tends to be relatively ad hoc. We found that no consensus exists as to the best way to apply the hot deck and obtain inferences from the completed data set. We described a range of different forms of the hot deck and different uses of the hot deck in practice, including the U.S. Census Bureau's hot deck for the Current Population Survey (CPS). We outlined existing research on its statistical properties, highlighting several areas in which the current research is lacking and would be good places for future work. We also provided an extended example of variations of the hot deck applied to the third National Health and Nutrition Examination Survey (NHANES III).

In Chapter II we considered a particular aspect of hot deck imputation, the use of sample weights. The naive approach is to ignore sample weights in creation of adjustment cells, which effectively imputes the unweighted sample distribution of respondents in an adjustment cell, potentially causing bias. Alternative approaches have been proposed that use weights in the imputation by incorporating them into the probabilities of selection for each donor (Cox, 1980; Rao and Shao, 1992). In this chapter we showed by simulation that these weighted hot decks do not correct for bias when the outcome is related to the sampling weight and the response propensity. We describe the correct approach, which is to use the sampling weight as a stratifying variable alongside additional adjustment variables when forming adjustment cells. We also demonstrated the practical use of our method through application to NHANES III data.

In Chapter IV we turned our focus to a different aspect of survey nonresponse: the potential for bias, particularly in the case when missingness might be related to the missing data themselves. We were unable to find existing methods that we felt adequately integrated the major factors contributing to the potential for bias, and thus were motivated to create a novel method. We introduced proxy pattern-mixture analysis (PPMA), a new method for assessing and adjusting for nonresponse bias for a continuous survey outcome when there is some fully observed auxiliary information available. PPMA is based on constructing a proxy for the partially missing outcome, and we proposed a taxonomy for the evidence concerning bias based on the strength of this proxy and the deviation of the mean of auxiliary information for respondents from its overall mean. We described several different estimation methods and introduced the fraction of missing information under the PPMA model as a simple measure of the magnitude of the nonresponse bias problem for a given data set. We proposed a sensitivity analysis to capture a range of potential missingness mechanisms, and illustrated the method through both simulation and application to NHANES III data.

Chapter V dealt with the natural extension of the PPMA to categorical outcomes. Through probit models and a latent variable framework we developed methods for evaluating nonresponse bias in binary and ordinal outcomes. We described multiple estimation methods and the sensitivity to model assumptions for each one. We proposed modifications to the less robust estimators to allow flexibility in handling situations where model assumptions may be violated. Simulations were used to illustrate the method and investigate the properties of the various estimation methods. Finally, NHANES III data were used to demonstrate the applicability of the method to real data.

There is much future work that will arise from this dissertation. As described in Chapter II, a major area that deserves attention is the so-called "swiss cheese pattern of missing data. Historically, hot deck methods have been studied when only one variable is subject to nonresponse. Though methods for general patterns of missingness have been suggested, there is little existing research on their empirical or theoretical properties. Future comparison of existing methods and development of hybrid approaches would be beneficial.

The other work that will grow out of this dissertation pertains to the proxy pattern-mixture analysis. In Chapter IV we describe the PPMA for continuous outcomes, and there are several potential extensions of this work. Thus far we have focused only on estimation of a mean; a useful extension would be to expand the approach to estimate potential bias in regression estimands. We developed the PPMA in the setting of unit nonresponse, and in the future will adapt the method to handle multivariate patterns of nonresponse arising from item nonresponse. Finally, extensions to handle panel surveys with more than two waves would also be of use.

In Chapter V the PPMA was extended to binary outcomes, and the next obvious step is domain estimation, an area of large interest to survey practitioners. Some discussion of this future aim follows. We will consider the case where there is a continuous outcome $Y$ and a binary domain indicator $D$. When $D$ is fully observed, application of the continuous PPM model is straightforward; the domain indicator can be included in the model that creates the proxy, or the entire continuous PPM method can be applied separately for the two domains. The more complex case is when $D$ is missing when $Y$ is missing, which we consider here. The primary interest is evaluating nonresponse bias for the domain means of $Y$.

Since both $Y$ and $D$ are partially missing, the proxy pattern-mixture analysis requires the creation of two proxies, one for $Y$ and one for $Y$. Let $X_1$ be the proxy for $Y$, created by a linear regression of $Y$ on the covariates $Z$ for the respondents. Since $D$ is binary, we create a proxy $X_2$ for $D$, created by a probit regression of $D$ on the covariates $Z$, assuming the latent variable $U$ underlies $D$. With two proxies we have two measures of the strength of the proxies, $\rho_{1y}^{(0)}$ and $\rho_{2u}^{(0)}$ for the proxies for $Y$ and $D$, respectively. There are also two measures of deviation, $d_1 = \bar{x}_1 - \bar{x}_{1R}$ and $d_2 = \bar{x}_2 - \bar{x}_{2R}$. The basic ranking of levels of evidence still holds; strong proxies and small deviations lead to the least uncertainty, with weak proxies and large deviations leading to the greatest uncertainty. However, there is added complexity when trying to estimate nonresponse bias for a domain mean. One could imagine a scenario where you have a strong proxy for $Y$, and thus a good handle on nonresponse bias for the overall mean of $Y$, but a weak proxy for $D$, and thus weak information with which to estimate the potential for bias in $Y$ within a domain.

The extension of the PPM to domain estimation extends the PPM model in a similar manner as the work of Little and Wang (1996), who extend the bivariate pattern-mixture model to the multivariate case when there are two patterns of missingness. We assume that the joint distribution of the proxies $X_1$ and $X_2$, outcome $Y$, and latent domain variable $U$ follow a multivariate pattern mixture model,

(6.1)

$$(X_1, X_2, Y, U|M=m) \sim N_4\left((\mu_1^{(m)}, \mu_2^{(m)}\mu_y^{(m)}, \mu_u^{(m)}), \Sigma^{(m)}\right)$$

$$M \sim Bernoulli(1-\pi)$$

$$\Sigma^{(m)} = \begin{bmatrix} \sigma_{11}^{(m)} & \sigma_{12}^{(m)} & \rho_{1y}^{(m)}\sqrt{\sigma_{11}^{(m)}\sigma_{yy}^{(m)}} & \sigma_{1u}^{(m)} \\ \sigma_{12}^{(m)} & \sigma_{22}^{(m)} & \sigma_{2y}^{(m)} & \rho_{2u}^{(m)}\sqrt{\sigma_{22}^{(m)}\sigma_{uu}^{(m)}} \\ \rho_{1y}^{(m)}\sqrt{\sigma_{11}^{(m)}\sigma_{yy}^{(m)}} & \sigma_{2y}^{(m)} & \sigma_{yy}^{(m)} & \sigma_{yu}^{(m)} \\ \sigma_{1u}^{(m)} & \rho_{2u}^{(m)}\sqrt{\sigma_{22}^{(m)}\sigma_{uu}^{(m)}} & \sigma_{yu}^{(m)} & \sigma_{uu}^{(m)} \end{bmatrix}.$$

Since $U$ is a latent variable we fix $\sigma_{uu}^{(0)}$ to an arbitrary value as in the binary PPM. The model is underidentified, and identifying restrictions are needed to estimate the parameters of the conditional distribution $[Y, U|X_1, X_2, M=1]$. In the language of Little and Wang (1996) the model is “just-identified” in that the number of partially observed variables is equal to the number of fully observed variables, enabling estimation without further restrictions in the extreme case where missingness depends only on $Y$ and $U$.

Once the PPM has yielded estimates of the joint distribution of $Y$ and $U$ for the nonrespondents, estimates of the domain means are obtained by considering the conditional distribution of $Y$ given $U$. The domain mean $(D=1)$ of $Y$ for $M=m$

is given by,

$$E[Y|U>0,M=m]=\mu_{y|d}^{(m)}=\mu_{y}^{(m)}+\frac{\rho_{yu}^{(m)}\sqrt{\sigma_{yy}^{(m)}}}{\sqrt{2\pi}\Phi\left(\frac{\mu_u^{(m)}}{\sigma_{uu}^{(m)}}\right)}\exp\left\{\frac{-{\mu_u^{(m)}}^2}{2\sigma_{uu}^{(m)}}\right\} \tag{6.2}$$

Thus the domain mean averaged over patterns is given by $\pi\mu_{y|d}^{(0)}+(1-\pi)\mu_{y|d}^{(1)}$.

Since the domain is a binary variable, the domain estimate will be sensitive to deviation from normality of the proxy $X_2$. Modifications to estimation methods are needed to ensure robustness, similar to the two-step maximum likelihood estimation and the modified Bayesian methods in the binary case.

# BIBLIOGRAPHY

Agresti, A. (2002), *Categorical Data Analysis*, Wiley: New York.

Albert, J. H. and Chib, S. (1993), "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, 669–679.

Andridge, R. R. and Little, R. J. A. (2008), "A Review of Hot Deck Imputation for Survey Nonresponse," Submitted to *International Statistical Review.*

— (2009), "The Use of Sample Weights in Hot Deck Imputation," *Journal of Official Statistics*, 25, 21–36.

Bailar, J. C. and Bailar, B. A. (1978), "Comparison of Two Procedures for Imputing Missing Survey Values," in *American Statistical Association Proceedings of the Survey Research Methods Section*, pp. 462–467.

Barzi, F. and Woodward, M. (2004), "Imputations of Missing Values in Practice: Results from Imputations of Serum Cholesterol in 28 Cohort Studies," *American Journal of Epidemiology*, 160, 34–45.

Berger, Y. G. and Rao, J. N. K. (2006), "Adjusted Jackknife for Imputation Under Unequal Probability Sampling Without Replacement," *Journal of the Royal Statistical Society B*, 68, 531–547.

Bethlehem, J. (2002), "Weighting Nonresponse Adjustments Based on Auxiliary Information," in *Survey Nonresponse*, eds. Groves, R., Dillman, D., Eltinge, J., and Little, R., New York: Wiley, chap. 18, pp. 275–287.

Bollinger, C. R. and Hirsch, B. T. (2006), "Match Bias from Earnings Imputation in the Current Population Survey: The Case of Imperfect Matching," *Journal of Labor Economics*, 24, 483–519.

Bowman, K., Chromy, J., Hunter, S., Martin, P., and Odom, D. (eds.) (2005), *2003 NSDUH Methodological Resource Book*, Rockville, MD: Substance Abuse and Mental Health Services Administration, Office of Applied Studies.

Breiman, L. and Friedman, J. H. (1993), *Classification and Regression Trees*, New York: Chapman & Hall.

Brick, J. M. and Kalton, G. (1996), "Handling Missing Data in Survey Research," *Statistical Methods in Medical Research*, 5, 215–238.

Brick, J. M., Kalton, G., and Kim, J. K. (2004), "Variance Estimation with Hot Deck Imputation Using a Model," *Survey Methodology*, 30, 57–66.

Brown, C. H. (1990), "Protecting Against Nonrandomly Missing Data in Longitudinal Studies," *Biometrics*, 46, 143–155.

Burns, R. M. (1990), "Multiple and Replicate Item Imputation in a Complex Sample Survey," in *U.S. Bureau of the Census Proceedings of the Sixth Annual Research Conference*, pp. 655–665.

Chen, J. and Shao, J. (1999), "Inference with Survey Data Imputed by Hot Deck when Imputed Values are Nonidentifiable," *Statistica Sinica*, 9, 361–384.

— (2000), "Nearest Neighbor Imputation for Survey Data," *Journal of Official Statistics*, 16, 113–141.

— (2001), "Jackknife Variance Estimation for Nearest-Neighbor Imputation," *Journal of the American Statistical Association*, 96, 260–269.

Cochran, W. G. (1977), *Sampling Techniques*, Wiley: New York, 3rd ed.

Cohen, G. and Duffy, J. C. (2002), "Are Nonrespondents to Health Surveys Less Healthy than Respondents," *Journal of Official Statistics*, 18, 13–23.

Collins, L., Schafer, J., and Kam, C. (2001), "A Ccomparison of Inclusive and Restrictive Missing-Data Strategies in Modern Missing-Data Procedures," *Psychological Methods*, 6, 330–351.

Cotton, C. (1991), "Functional Description of the Generalized Edit and Imputation System," Tech. rep., Statistics Canada.

Cox, B. G. (1980), "The Weighted Sequential Hot Deck Imputation Procedure," in *American Statistical Association Proceedings of the Survey Research Methods Section*, pp. 721–726.

Cox, B. G. and Folsom, R. E. (1981), "An Evaluation of Weighted Hot Deck Imputation for Unreported Health Care Visits," in *American Statistical Association Proceedings of the Survey Research Methods Section*, pp. 412–417.

Cox, N. R. (1974), “Estimation of the Correlation Between a Continuous and a Discrete Variable,” *Biometrics*, 30, 171–178.

Curtain, R., Presser, S., and Singer, E. (2000), “The Effects of Response Rate Changes on the Index of Consumer Sentiment,” *Public Opinion Quarterly*, 64, 413–428.

— (2005), “Changes in Telephone Survey Nonresponse over the Past Quarter Century,” *Public Opinion Quarterly*, 69, 87–98.

Daniels, M. J. and Hogan, J. W. (2000), “Reparameterizing the Pattern Mixture Model for Sensitivity Analyses under Informative Dropout,” *Biometrics*, 56, 1241–1248.

David, M., Little, R. J. A., Samuhel, M. E., and Triest, R. K. (1986), “Alternative Methods for CPS Income Imputation,” *Journal of the American Statistical Association*, 81, 29–41.

Diggle, P. and Kenward, M. G. (1994), “Informative Drop-Out in Longitudinal Data Analysis,” *Applied Statistics*, 43, 49–93.

Efron, B. (1994), “Missing Data, Imputation, and the Bootstrap,” *Journal of the American Statistical Association*, 89, 463–475.

England, A. M., Hubbell, K. A., Judkins, D. R., and Ryaboy, S. (1994), “Imputation of Medical Cost and Payment Data,” in *American Statistical Association Proceedings of the Survey Research Methods Section*, pp. 406–411.

Ezzati-Rice, T. M., Fahimi, M., Judkins, D., and Khare, M. (1993a), “Serial Imputation of NHANES III With Mixed Regression and Hot-Deck Imputation,” in *American Statistical Association Proceedings of the Survey Research Methods Section*, pp. 292–296.

Ezzati-Rice, T. M., Khare, M., Rubin, D. B., Little, R. J. A., and Schafer, J. L. (1993b), “A Comparison of Imputation Techniques in the Third National Health and Nutrition Examination Survey,” in *American Statistical Association Proceedings of the Survey Research Methods Section*, pp. 303–308.

Fay, R. E. (1993), “Valid Inferences from Imputed Survey Data,” in *American Statistical Association Proceedings of the Survey Research Methods Section*, pp. 41–48.

— (1996), “Alternative Paradigms for the Analysis of Imputed Survey Data,” *Journal of the American Statistical Association*, 91, 490–498.

— (1999), “Theory and Application of Nearest Neighbor Imputation in Census 2000,” in *American Statistical Association Proceedings of the Survey Research Methods Section*, pp. 112–121.

Federal Committee on Statistical Methodology (2001), “Statistical Policy Working Paper 31: Measuring and Reporting Sources of Error in Surveys,” Tech. rep., Executive Office of the President of the United States of America.

Ford, B. L. (1983), “An Overview of Hot-Deck Procedures,” in *Incomplete Data in Sample Surveys*, eds. Madow, W. G., Olkin, I., and Rubin, D. B., Academic Press: New York, vol. 2, pp. 185–207.

Grau, E. A., Frechtel, P. A., and Odom, D. M. (2004), “A Simple Evaluation of the Imputation Procedures Used in HSDUH,” in *American Statistical Association Proceedings of the Survey Research Methods Section*, pp. 3588–3595.

Groves, R. M. (2006), “Nonresponse Rates and Nonresponse Bias in Household Surveys,” *Public Opinion Quarterly*, 70, 646–675.

Haziza, D. and Beaumont, J.-F. (2007), “On the Construction of Imputation Classes in Surveys,” *International Statistics Review*, 75, 25–43.

Haziza, D. and Rao, J. N. K. (2006), “A Nonresponse Model Approach to Inference Under Imputation for Missing Survey Data,” *Survey Methodology*, 32, 53–64.

Heckman, J. J. (1976), “The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and a Simple Estimator for Such Models,” *The Annals of Economic and Social Measurement*, 5, 475–492.

Heitjan, D. F. and Little, R. J. A. (1991), “Multiple Imputation for the Fatal Accident Reporting System,” *Applied Statistics*, 40, 13–29.

Judkins, D. R. (1997), “Imputing for Swiss Cheese Patterns of Missing Data,” in *Proceedings of Statistics Canada Symposium 97*.

Judkins, D. R., Hubbell, K. A., and England, A. M. (1993), “The Imputation of Compositional Data,” in *American Statistical Association Proceedings of the Survey Research Methods Section*, pp. 458–462.

Kalton, G. and Kasprzyk, D. (1986), “The Treatment of Missing Survey Data,” *Survey Methodology*, 12, 1–16.

Kass, G. V. (1980), “An Exploratory Technique for Investigating Large Quantities of Categorical Data,” *Applied Statistics*, 29, 119–127.

Keeter, S., Miller, C., Kohut, A., Groves, R. M., and Presser, S. (2000), "Consequences of Reducing Nonresponse in a National Telephone Survey," *Public Opinion Quarterly*, 64, 125–148.

Khare, M., Little, R. J. A., Rubin, D. B., and Schafer, J. L. (1993), "Multiple Imputation of NHANES III," in *American Statistical Association Proceedings of the Survey Research Methods Section*, pp. 297–302.

Kim, J. K. (2002), "A Note on Approximate Bayesian Bootstrap," *Biometrika*, 89, 470–477.

Kim, J. K., Brick, J. M., Fuller, W. A., and Kalton, G. (2006), "On the Bias of the Multiple-Imputation Variance Estimator in Survey Sampling," *Journal of the Royal Statistical Society B*, 68, 509–521.

Kim, J. K. and Fuller, W. (2004), "Fractional Hot Deck Imputation," *Biometrika*, 91, 559–578.

Lazzeroni, L. G., Schenker, N., and Taylor, J. M. G. (1990), "Robustness of Multiple-Imputation Techniques to Model Misspecification," in *American Statistical Association Proceedings of the Survey Research Methods Section*, pp. 260–265.

Lillard, L., Smith, J. P., and Welch, F. (1982), "What Do We Really Know About Wages: The Importance of Non-reporting and Census Imputation," Tech. rep., Rand Corporation, Santa Monica, CA.

Little, R. J. A. (1986), "Survey Nonresponse Adjustments for Estimates of Means," *International Statistics Review*, 54, 139–157.

— (1988), "Missing-Data Adjustments in Large Surveys," *Journal of Business and Economic Statistics*, 6, 287–296.

— (1993), "Pattern-Mixture Models for Multivariate Incomplete Data," *Journal of the American Statistical Association*, 88, 125–134.

— (1994), "A Class of Pattern-Mixture Models for Normal Incomplete Data," *Biometrika*, 81, 471–483.

— (2004), "To Model or Not to Model? Competing Modes of Inference for Finite Population Sampling," *Journal of the American Statistical Association*, 99, 546–556.

Little, R. J. A. and Rubin, D. B. (2002), *Statistical Analysis with Missing Data*, Wiley: New York, 2nd ed.

Little, R. J. A. and Vartivarian, S. (2003), “On Weighting the Rates in Non-Response Weights,” *Statistics in Medicine*, 22, 1589–1599.

— (2005), “Does Weighting for Nonresponse Increase the Variance of Survey Means?” *Survey Methodology*, 31, 161–168.

Little, R. J. A. and Wang, Y. (1996), “Pattern-Mixture Models for Multivariate Incomplete Data with Covariates,” *Biometrics*, 52, 98–111.

Lumley, T. (2004), “Analysis of complex survey samples,” *Journal of Statistical Software*, 9, 1–19.

Marker, D. A., Judkins, D. R., and Winglee, M. (2002), “Large-Scale Imputation for Complex Surveys,” in *Survey Nonresponse*, Wiley: New York, pp. 329–341.

Meng, X. L. (1994), “Multiple Imputation Inferences with Uncongenial Sources of Input (with discussion),” *Statistical Science*, 9, 538–573.

Nandram, B. and Choi, J. W. (2002a), “A Bayesian Analysis of a Proportion Under Non-Ignorable Non-Response,” *Statistics in Medicine*, 21, 1189–1212.

— (2002b), “Hierarchical Bayesian Nonresponse Models for Binary Data from Small Areas with Uncertainty about Ignorability,” *Journal of the American Statistical Association*, 97, 381–388.

Nandram, B., Han, G., and Choi, J. W. (2002), “A Hierarchical Bayesian Nonignorable Nonresponse Model for Multinomial Data from Small Areas,” *Survey Methodology*, 28, 145–156.

Nandram, B., Liu, N., Choi, J. W., and Cox, L. (2005), “Bayesian Non-response Models for Categorical Data from Small Areas: An Application to BMD and Age,” *Statistics in Medicine*, 24, 1047–1074.

National Center for Education Statistics (2002), “NCES Statistical Standards,” Tech. rep., U.S. Department of Education.

Office of Management and Budget (2006), “Standards and Guidelines for Statistical Surveys,” Tech. rep., Executive Office of the President of the United States of America.

Oh, H. L. and Scheuren, F. S. (1983), “Weighting Adjustments for Unit Nonresponse,” in *Incomplete Data in Sample Surveys*, eds. Madow, W. G., Olkin, I., and Rubin, D. B., Academic Press: New York, vol. 2, pp. 143–184.

Olsson, U., Drasgow, F., and Dorans, N. J. (1982), “The Polyserial Correlation Coefficient,” *Psychometrika*, 47, 337–347.

Ono, M. and Miller, H. P. (1969), "Income Nonresponses in the Current Population Survey," in *American Statistical Association Proceedings of the Social Statistics Section*, pp. 277–288.

Perez, A., Dennis, R. J., Gil, J. F. A., and Rondon, M. A. (2002), "Use of the Mean, Hot Deck and Multiple Imputation Techniques to Predict Outcome in Intensive Care Unit Patients in Colombia," *Statistics in Medicine*, 21, 3885–3896.

Platek, R. and Gray, G. B. (1983), "Imputation Methodology: Total Survey Error," in *Incomplete Data in Sample Surveys*, eds. Madow, W. G., Olkin, I., and Rubin, D. B., Academic Press: New York, vol. 2, pp. 249–333.

R Development Core Team (2007), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., and Solenberger, P. (2001), "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models," *Survey Methodology*, 21, 85–95.

Rancourt, E. (1999), "Estimation with Nearest Neighbor Imputation at Statistics Canada," in *American Statistical Association Proceedings of the Survey Research Methods Section*, pp. 131–138.

Rancourt, E., Särndal, C. E., and Lee, H. (1994), "Estimation of the Variance in the Presence of Nearest Neighbor Imputation," in *American Statistical Association Proceedings of the Survey Research Methods Section*, pp. 888–893.

Rao, J. N. K. (1996), "On Variance Estimation with Imputed Survey Data," *Journal of the American Statistical Association*, 91, 499–506.

Rao, J. N. K. and Shao, J. (1992), "Jackknife Variance Estimation with Survey Data Under Hot Deck Imputation," *Biometrika*, 79, 811–822.

Robins, J. M. and Wang, N. (2000), "Inference for Imputation Estimators," *Biometrika*, 87, 113–124.

Rubin, D. B. (1976), "Inference and Missing Data (with Discussion)," *Biometrika*, 63, 581–592.

— (1977), "Formalizing Subjective Notions About the Effect of Nonrespondents in Sample Surveys," *Journal of the American Statistical Association*, 72, 538–542.

— (1978), "Multiple Imputation in Sample Surveys - a Phenomenological Bayesian Approach to Nonresponse," in *American Statistical Association Proceedings of the Survey Research Methods Section*, pp. 20–34.

— (1981), “The Bayesian Bootstrap,” *Annals of Statistics*, 9, 130–134.

— (1986), “Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations,” *Journal of Business and Economic Statistics*, 4, 87–94.

— (1987), *Multiple Imputation for Nonresponse in Surveys*, Wiley: New York.

— (1996), “Multiple Imputation After 18+ Years,” *Journal of the American Statistical Association*, 91, 473–489.

Rubin, D. B. and Schenker, N. (1986), “Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Non-Response,” *Journal of the American Statistical Association*, 81, 366–374.

Saigo, H., Shao, J., and Sitter, R. R. (2001), “A Repeated Half-Sample Bootstrap and Balanced Repeated Replications for Randomly Imputed Data,” *Survey Methodology*, 27, 189–196.

Särndal, C. E. (1992), “Methods for Estimating the Precision of Survey Estimates When Imputation Has Been Used,” *Survey Methodology*, 18, 241–252.

Schenker, N. and Taylor, J. M. G. (1996), “Partially Parametric Techniques for Multiple Imputation,” *Computational Statistics and Data Analysis*, 22, 425–446.

Shao, J. and Chen, J. (1999), “Approximate Balanced Half Sample and Repeated Replication Methods for Imputed Survey Data,” *Sankhya, Series B*, 61, 187–201.

Shao, J., Chen, Y., and Chen, Y. (1998), “Balanced Repeated Replication for Stratified Multistage Survey Data under Imputation,” *Journal of the American Statistical Association*, 93, 819–831.

Shao, J. and Sitter, R. R. (1996), “Bootstrap for Imputed Survey Data,” *Journal of the American Statistical Association*, 91, 1278–1288.

Shao, J. and Steel, P. (1999), “Variance Estimation for Survey Data With Composite Imputation and Nonnegligible Sampling Fractions,” *Journal of the American Statistical Association*, 94, 254–265.

Shao, J. and Wang, H. (2002), “Sample Correlation Coefficients Based on Survey Data Under Regression Imputation,” *Journal of the American Statistical Association*, 97, 544–552.

Siddique, J. and Belin, T. R. (2008), “Multiple imputation using an iterative hot-deck with distance-based donor selection,” *Statistics in Medicine*, 27, 83–102.

Srivastava, M. S. and Carter, E. M. (1986), "The Maximum Likelihood Method for Non-Response in Sample Surveys," *Survey Methodology*, 12, 61–72.

Stasny, E. A. (1991), "Hierarchical Models for the Probabilities of a Survey Classification and Nonresponse: An Example from the National Crime Survey," *Journal of the American Statistical Association*, 86, 296–303.

Tang, L., Song, J., Belin, T. R., and Unutzer, J. (2005), "A Comparison of Imputation Methods in a Longitudinal Randomized Clinical Trial," *Statistics in Medicine*, 24, 2111–2128.

Tate, R. F. (1955a), "Applications of Correlation Models for Biserial Data," *Journal of the American Statistical Association*, 50, 1078–1095.

— (1955b), "The Theory of Correlation Between Two Continuous Variables When One is Dichotomized," *Biometrika*, 42, 205–216.

Twisk, J. and de Vente, W. (2002), "Attrition in Longitudinal Studies: How to Deal with Missing Data," *Journal of Clinical Epidemiology*, 55, 329–337.

U.S. Bureau of the Census (2002), "Technical Paper 63," Tech. rep., U.S. Government Printing Office.

U.S. Department of Health and Human Services (1994), "Plan and Operation of the Third National Health and Nutrition Examination Survey, 1988-94," Tech. rep., National Center for Health Statistics, Centers for Disease Control and Prevention.

— (2001), "Third National Health and Nutrition Examination Survey (NHANES III, 1988-1994): Multiply Imputed Data Set. CD-ROM, Series 11, No. 7A." Tech. rep., National Center for Health Statistics, Centers for Disease Control and Prevention.

Van Buuren, S. and Oudshoorn, C. G. M. (1999), "Flexible Multivariate Imputation by MICE," Tech. rep., TNO Prevention and Health, Leiden.

Williams, R. L. and Folsom, R. E. (1981), "Weighted Hot-Deck Imputation of Medical Expenditures Based on a Record Check Subsample," in *American Statistical Association Proceedings of the Survey Research Methods Section*, pp. 406–411.