A Hot Deck Imputation Procedure for Multiply Imputing Nonignorable Missing Data: The Proxy Pattern-Mixture Hot Deck

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Danielle M. Sullivan, M.S.

Graduate Program in Biostatistics

The Ohio State University

2014

Dissertation Committee:

Dr. Rebecca R. Andridge, Advisor

Dr. Bo Lu

Dr. Eloise Kaizar

Dr. Elizabeth Stasny

Abstract

Hot deck imputation is a common method for handling item nonresponse in surveys, but most implementations assume data are missing at random (MAR). We propose a new hot deck method for imputation of a partially missing outcome variable that does not assume data are MAR. We use a parametric model to create predicted means for both donors and donees under varying assumptions on the missing data mechanism, ranging from MAR to missing not at random (MNAR). When imputing a continuous outcome variable, for a given assumption on the missingness mechanism, the predicted means are used to define distances between donors and donees and probabilities of selection proportional to those distances. Multiple imputation using the hot deck is performed to create a set of completed data sets, using an approximate Bayesian bootstrap to ensure "proper" imputations. This new hot deck method creates an intuitive sensitivity analysis where imputations may be performed under MAR and under varying MNAR mechanisms, and the resulting impact on inference can be evaluated. In addition, we propose two donor quality metrics to identify situations where close matches of donor to donee are not available, which can occur under strong MNAR assumptions. We investigate bias and coverage of estimates from our proposed method through simulation and apply the method to estimation of income in the Ohio Medicaid Assessment Survey.

We extend the proposed hot deck method for multiple imputation of a binary outcome variable by assuming there exists a continuous latent variable that determines the value of the binary outcome. This allows us to use the framework developed under a continuous outcome to create predicted means assuming different missingness mechanisms. However, because the latent variable is by definition unobserved, additional steps are required to obtain the parameter estimates used in creating the predicted means and we compare two approaches of estimation. Furthermore, we modify donor selection by implementing an adjustment cell procedure. We investigate bias and coverage of estimates from our proposed method through simulation and study the sensitivity to normality. We apply the method to estimation of mean ER+ status in the Surveillance, Epidemiology, and End Results Program. In addition, we illustrate how the method can be applied to estimate regression coefficients.

For my parents, with love.

Vita

September 25, 1985	Born - Dunkirk, New York, USA

Education

2007	B.S. Mathematics, SUNY Fredonia Fredonia, NY
2009	M.S. Statistics, The Ohio State University Columbus, OH

Professional Experience

2007-2010	Graduate Teaching Associate, Department of Statistics The Ohio State University Columbus, OH
2010-Present	Graduate Research Associate, Department of Biostatistics The Ohio State University Columbus, OH

Publications

Research Publications

Pinsonneault JK, **Sullivan DM**, Sadee W, Soares CN, Hampson E, Steiner M (2013). Association study of the estrogen receptor gene *ESR1* with postpartum depression – a pilot study. *Archives of Women's Mental Health*, DOI: 10.1007/s00737-013-0373-8.

Kitzmiller JP, **Sullivan DM**, Phelps MA, Wang D, Sadee W (2013). CYP3A4/5 combined genotype analysis for predicting statin dose requirement for optimal lipid control. *Drug Metabolism and Drug Interactions* 28(1); 59-63.

Sullivan DM, Pinsonneault JK, Papp AC, Zhu H, Lemeshow S, Mash DC, Sadee W (2013). Dopamine transporter DAT and receptor DRD2 variants affect risk of lethal cocaine abuse: a gene-gene-environment interaction. *Translational Psychiatry* 3.

Fields of Study

Major Field: Biostatistics

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

This dissertation develops a new hot deck imputation procedure for item nonresponse. We do not make the assumption that missingness is at random, as the majority of hot deck imputations do. The method allows the examination of missingness mechanisms through the variation of a sensitivity parameter. The rest of this chapter presents the nomenclature of missing data, and discusses common methods of handling missing data. Chapter 2 reviews the literature that is relevant towards the proposed method. The proposed method is described and studied in Chapter 3 for imputation of a partially observed continuous variable. We formulate the steps of the method, perform a simulation study to examine its behavior, and apply the method to data from The Ohio Medicaid Assessment Survey. Chapter 4 extends the method to handle imputation of a binary variable. The theory and steps are formulated, and a simulation study is performed. The proposed method is applied to two data sets. The first imputes missing estrogen receptor (ER) status in the Surveillance, Epidemiology, and End Results (SEER) dataset, and the second illustrates how the method can be applied to a logistic regression model when the outcome is partially observed. Chapter 5 is summary, discussion and future work.

1.1 Missing Data Terminology

Incomplete data arise frequently in observational studies, surveys, and even controlled experiments. The reason the data are missing, the missingness mechanism, has an impact on the methods used for inference. The missing data mechanism can be classified as one of three types: missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR) [1]. To introduce these terms, consider the simple situation of having only one variable, Y, subject to missingness, and no covariates. MCAR implies that missingness is not dependent upon observed or unobserved values of Y and is a special case of MAR. MAR means that missingness may depend upon the observed values, but not on the unobserved Y. Specifically, we let Y be the vector of potential responses for a subject, and let M denote the missingness indicator, which takes the value 1 if Y is missing (unobserved), and 0 if Y is observed. We can write Y as $y = (y_{obs}, y_{miss})$. Under MCAR, $P(M = 1|y) = P(M = 1)$, whereas under MAR, $P(M = 1|y) = P(M = 1|y_{obs})$ and the shared assumption of these two mechanisms is the independence of missingness with the unobserved values of y [1, 2].

A more complicated situation arises when the missing data mechanism is missing not at random (MNAR). This is the case when the probability Y is missing depends upon the unobserved values (and optionally, the observed values also), i.e., $P(M = 1|y) = P(M = 1|y_{miss}, (y_{obs}))$. This is a problem because there is no formal test to determine if the data are MNAR, since the probability a value is missing is dependent upon the missing value itself. Tests can be performed to determine if the mechanism is not missing completely at random, but if the MCAR assumption is questionable, distinguishing between MAR and MNAR is not possible [3].

The missing data mechanism can either be ignorable (meaning one does not need to model it and it can be 'ignored'), or nonignorable (meaning that the mechanism should be modeled). An ignorable mechanism results when the missingness is either (1) MCAR or (2) MAR with distinctness between parameters involved in the data model and the response model. Nonignorability occurs when the data mechanism is either (1) MNAR or (2) MAR and the parameters are not distinct [1]. The assumption of ignorability can be evaluated by subject matter experts, data from outside sources, or more formally through a sensitivity analysis. Though there are two situations that lead to "nonignorability", it has become common to use MNAR and nonignorability interchangeably, and this is the nomenclature we use in this paper.

1.2 Methods for Handling Missing Data

There have been many methods developed to handle missing data, all which range in degree of implementation difficulty, ability to properly reduce bias, and ability to correctly estimate the standard error. Typically the degree of implementation difficulty depends on whether the method has already been incorporated into software methods and is easy enough to be used by non-statisticians. These may or may not be the 'best' methods to analyze any specific data set. As will be illustrated throughout this text, collecting data that are thought to be related to both missingness and the variable that is missing can help with the "worst case scenarios" of missing data.

1.2.1 Methods assuming MCAR

Suppose a dataset contains a number of subjects with a set of covariates, all of which are subject to missingness. There are two "easy" approaches to handle the missingness: a complete-case analysis (list-wise deletion) or an available-case analysis (pair-wise deletion). A complete-case analysis would remove every case that has at least one missing value for any variable prior to performing any type of analysis. Therefore, every analysis would be conducted on the same set of cases, regardless of the covariates involved. However, if missingness is scattered amongst variables, this approach could remove a large number of cases and any analysis would suffer from lack of power. An alternative is an available-case analysis. This also removes cases with missing values, but only for the variables being used in the current analysis. If sets of analyses are done, such as a series of pairwise correlations, each correlation could potentially consist of a different subset of original cases, depending on the pair of variables used. This makes use of all available data for each single analysis, but can be a problem when comparing results, since the comparisons would be done on different sets of cases with varying sample sizes.

A major assumption of both the complete-case and available-case analyses is that missingness is completely at random (MCAR). To test this assumption, one can compare the distributions of fully observed variables for respondents and nonrespondents. However, a formal test to distinguish between MNAR and MAR is not possible, because the values that matter are the values that are missing. The assumption that the data are MAR may be more believable if certain variables that are thought to be related to missingness are included in the analysis. If it is reasonable to believe the data are indeed MCAR, and the rate of missingness is low (such that the complete case analysis would only be removing a few cases) then the parameter estimates will be unbiased and only a small loss of efficiency will incur. However, if the missing data rate is high, and the data truly are not MCAR, then both the complete-case and available-case analyses could result in biased estimates and increased variance [4].

1.2.2 Methods assuming MAR

Imputation methods were developed in an attempt to recover the lost sample size and information from high rates of missingness by creating 'completed' data sets. The purpose is not necessarily to impute the correct value for the missing subject, but to regain lost information to perform inference on a population parameter. Missing values can either be imputed once (single imputation) or multiple times (multiple imputation). With many imputation methods, the resulting parameter estimates may be unbiased, but treating imputed values as if they were actually observed will result in biased variance estimates.

Single Imputation

Single imputation imputes a single value for each missing value, and is the simplest form of imputation. Mean substitution, also called unconditional mean imputation, replaces the missing values with the mean of the observed values. This drastically reduces the variance, since the same value (the mean), is imputed for every missing value. Parameter estimates may be biased if the data are not MCAR. If large values are missing at a higher rate than small values, for example, then the mean of the respondents will be lower than the population average, and imputation of the respondent mean for nonrespondents could be just as biased as the complete case analysis. Mean substitution can also be done by subgroups, if the means differ across groups (e.g., female weight versus male weight). While helping to reduce bias compared to imputing the overall mean, this still underestimates the variability in the data.

Regression imputation (or conditional mean imputation) imputes the predicted score from a regression model, i.e., $\hat{Y}_i = \mathbf{X}\hat{\boldsymbol{\beta}}$. The variable Y represents the variable with missing values and X is a fully observed variable (or set of variables). The $\hat{\beta}$'s are found using the observed cases only, but the fitted values are available for everyone in the data set (since they would all have fully observed X's).

With *un*conditional mean imputation, correlations between variables in the data are underestimated, but with conditional mean imputation, the correlations are overestimated and variances still underestimated. Consider the case where the true underlying population has a strong linear relationship between two variables, but Y values are missing at random among the X's. For the true data, a low value of X corresponds to a low value of Y. If unconditional mean imputation is used, so that the mean of the observed Y's is imputed for all those missing, then low values of X's will have missing Y imputed higher than the true value, and high values of X's will have missing Y imputed lower than the true values. This disrupts the true linear relationship between these variables, resulting in lower correlation. On the other hand, with conditional mean imputation, the missing values will be imputed to lie perfectly on the regression line, 'forcing' a stronger correlation than the true value.

Stochastic regression imputation takes conditional mean imputation one step further. Stochastic regression has the exact same $\hat{\beta}$'s, but random error is added to the imputed estimate: $\hat{Y}_i = [\hat{\beta}_0 + \hat{\beta}_1 X_i] + e_i$, with $e_i \sim N(0, \hat{\sigma}^2_{Y|X})$. Other variations involve adding a randomly chosen residual. Instead of imputing the missing values so that they fall directly on the regression line, there would be some added variability, which would decrease the overestimated correlation of the conditional mean imputation.

Stochastic regression imputation is typically done via a parametric model. A method that is considered non-parametric is hot deck imputation. Hot deck imputation finds subjects with observed values of the variable with missingness, called "donors", who are "close" matches to subjects who are missing that variable. A review of hot deck imputation can be found in [5]. There are many ways to create the set of possible donors for each unobserved subject (the donor pool), and many metrics used to define a "close" match. Once a donor has been selected, the donor's observed value is then imputed for the missing value. An advantage with hot deck imputation is that actual observed values are being imputed, as opposed to imputation of a mean, or predicted values from a regression. This method can preserve associations between variables but treating imputed values as if they were known still underestimates variances.

Multiple Imputation

Multiple imputation is the process of imputing more than one value for each missing observation, resulting in more than one complete data set. Multiple imputation (MI), in the traditional sense as formulated by Rubin [1], consists of three stages: imputation, estimation and pooling. Multiple imputation (so long as it is 'proper') accounts for the uncertainty associated with missing data by creating a set of completed data sets with different imputed values for the missing subjects in each. Proper multiple imputation must be performed to fully recover the missing variability of the imputations. Rubin states, "With ignorable nonresponse, the respondents and nonrespondents share the same parameters, but the sample mean and sample variance for respondents are not perfect estimates of these parameters, and our imputations must reflect this uncertainty to be proper" (page 123 of [1]). Rubin suggests that to make imputations proper, prior distributions should be placed on all unknown model parameters, e.g., placing priors on the β parameters in stochastic regression imputation. Schafer (1999) [6] discusses this further and gives examples. This needs to be kept in mind as we proceed later with hot deck multiple imputation, since it is an improper MI if no adjustment is made.

To implement multiple imputation, suppose the goal is to estimate a single parameter θ. As described in [7], based on D multiply-imputed datasets, the estimate is simply the mean, $\bar{\hat{\theta}}$ of the $\hat{\theta}_i's$ from each of the D datasets:

$$\bar{\hat{\theta}} = \frac{1}{D}\sum_{i=1}^{D}\hat{\theta}_i. \tag{1.1}$$

The variance associated with $\bar{\hat{\theta}}$ is given by:

$$V_T = V_W + \left(1 + \frac{1}{D}\right) V_B. \tag{1.2}$$

Here, V_W is the within-imputation variance, calculated as the average of the variance estimates $\hat{W}_i$ from the individual data sets:

$$V_W = \frac{1}{D}\sum_{i=1}^{D}\hat{W}_i, \tag{1.3}$$

and V_B is the between-imputation variance estimated as the variance of the $\hat{\theta}_i$ values across the data sets:

$$V_B = \frac{1}{D-1}\sum_{i=1}^{D}(\hat{\theta}_i - \bar{\hat{\theta}})^2. \tag{1.4}$$

Inference about θ can be performed by comparing $\bar{\hat{\theta}}$ to a t-distribution with degrees of freedom ν given by:

$$\nu = (D-1)\left(1 + \frac{D}{D+1}\frac{V_W}{V_B}\right)^2. \tag{1.5}$$

The nature of multiple imputation allows for the development of measures of the impact of missing data, as it gives both an estimate of the sampling variability as if there were no missing data (V_W), and an estimate of the uncertainty of the imputations (V_B). A simple estimate of the nonresponse bias can be found by comparing the MI estimate to the estimate using a complete case analysis [8]. One measure of the impact of nonresponse on the sampling variance of the estimate that uses V_W and V_B is the fraction of missing information (FMI) [1]. The fraction of missing information is defined as

$$\text{FMI} = \frac{V_B + V_B/D}{V_T} = \frac{V_B(1 + 1/D)}{V_T} \tag{1.6}$$

for large D. We can compare equations (1.2) and (1.6) and observe that for large D, the FMI is the fraction of the total variance which is due to the between imputation variance. If the between imputation variance is much larger than the within imputation variance, it implies we have high uncertainty in our imputations and the FMI will be close to one. If the imputations are very similar across the MI data sets, the between-imputation variance will be low, resulting in a low FMI close to zero, implying that the uncertainty pertaining to the missing values is low.

Chapter 2: Literature Review

When the missingness mechanism is nonignorable, standard imputation methods will lead to biased estimates. This leads to a challenging, but interesting, problem because the data to test nonignorability are unavailable. There has been extensive work in imputing nonignorable missing data for longitudinal settings (see [9] for a review), but not as much in cross-sectional studies and surveys. For surveys, there are imputation methods which aim to account for nonignorable missingness. Other methods use follow-up data in an attempt to reduce bias due to MNAR. Another technique is to use sensitivity analyses which study model estimates under a variety of missingness scenarios.

2.1 Nonignorable missingness

In this dissertation we consider the type of nonignorability arising when missingness depends on unobserved values. In this case, the response probability cannot be ignored and should be modeled to decrease or eliminate bias. There are two main types of models used, which differ in the factorization of the joint distribution of the data and missingness indicator. These are selection models [10] and pattern-mixture models [11]. The models have different assumptions and can lead to different estimates.

Let M denote the indicator of missingness ($M = 1$ if Y is missing, $M = 0$ if Y is observed). Factorizing according to the selection model yields:

$$p(Y, M|\theta, \psi) = p(Y|\theta)p(M|Y, \psi), \tag{2.1}$$

whereas the pattern-mixture model is written as,

$$p(Y, M|\phi, \pi) = p(Y|M, \phi)p(M|\pi), \tag{2.2}$$

where $(\theta, \psi, \phi, \pi)$ are all unknown parameters. The selection model factors the joint model into the complete data model for Y and the model for the missingness, conditional on Y. Common choices for the missingness model in the selection model framework are probit or logit models. The pattern-mixture model results in a 'mixture' of the respondents and nonrespondents by modeling the marginal distribution of M and the conditional distribution of Y given M. This assumes different parameters for the respondents and the nonrespondents, i.e., $\phi = (\boldsymbol{\mu}^{(m=0)}, \boldsymbol{\mu}^{(m=1)}, \boldsymbol{\Sigma}^{(m=0)}, \boldsymbol{\Sigma}^{(m=1)}, \pi)$, where $P(M = 1|\pi) = 1 - \pi$. When the data are missing completely at random, Y is independent of M, the selection model becomes $p(Y|\theta)p(M|\psi)$, and the pattern-mixture model becomes $p(Y|\phi)p(M|\pi)$. Hence, they are equivalent and $\theta = \phi$, $\psi = \pi$ [12]. Otherwise, they produce different models. Under nonignorable missingness, if either model holds, the other does not hold [13].

2.2 Non-imputation Approaches

We now discuss literature that attempts to handle nonignorable missingness by implementing methods other than imputation. These methods model the missingness mechanism and include several Bayesian approaches. We also discuss several methods that provide means of performing a sensitivity analysis, by either comparing a series of models, or by computing a specified value to capture the amount of nonignorability.

Baker and Laird (1988) [14] present how to model nonignorable nonresponse using log-linear models. The joint distribution of the categorical outcome, Y, the covariates, X, and the response indicator, R is factored according to the selection model. By including a $Y - R$ interaction term in the model for nonresponse, the resulting model is nonignorable. The EM algorithm is used to maximize the likelihood to find estimates for the parameters of the regression (model for X, Y), and the nonresponse parameters.

Smith et al. (1999) [15] applied the log-linear nonignorable nonresponse models of Baker and Laird to the 1992 British General Election Panel Survey. This survey was taken prior to the election and collected information on voting intention. There were also post-election results for most of the non-responders that the authors used for comparison purposes. The interest was in estimating the proportions who would vote for the various political parties: Conservative, Labour, Liberal Democratic. Following Baker and Laird, the joint distribution of $P(Y, X)$ is represented by a saturated log-linear model, while the nonresponse mechanism, $P(R|Y, X)$, is modeled through alternative specifications of a separate log-linear model. By comparing these alternative specifications, Smith et al. deem it a type of sensitivity analysis. Because the post-election data are available, they compare the estimates from fitting ignorable models and the various nonignorable models to the observed voting proportions.

One of the main findings was that some nonignorable nonresponse models gave point estimates of voting proportions and parameters on the boundary of their parameter spaces resulting in implausible results. These implausible results included no assignments of nonrespondents to the Labour party, while a large proportion of the nonrespondents was assigned to the category 'other', even though the category only represented a small proportion of the respondents. They suggest placing constraints on model parameters; specifically placing bounds on the range of the log-odds ratio of the $Y-R$ interaction term (which compares the voting intention for a nonrespondent versus a respondent).

Stasny (1988) [16] considered nonignorable nonresponse in panel data, where the interest may be in estimating the gross change from one period to another. The survey responses were categorical, and it was assumed that subjects were observed in at least one of the two periods. Five models were described, three of which were ignorable models and two were nonignorable models. All models were applied to the Current Population Survey (CPS) and the Labour Force Survey (LFS) for estimating the change in employment status in two consecutive time periods. There were two assumptions of the nonignorable models. The first was that nonresponse depended on the period and on the unobserved survey classification in the period when the individual did not respond. The second was that nonresponse depended only on the unobserved survey classification in the period when the individual did not respond. Maximum likelihood estimation was used to fit all models, but the nonignorable models were more challenging to fit because they required the parameters to be solved for simultaneously. Iterative procedures to fit all models are discussed.

In another paper by Stasny (1991) [17], an empirical Bayes procedure for estimating parameters of hierarchical models allowing for non-random nonresponse is discussed. The models were motivated by the National Crime Survey (NCS). This survey had two concerns: small-domain estimation and non-random nonresponse.

The binary response of interest was whether or not a subject had a crime committed against him/her (or their property) in the previous 6-month period. Nonresponse may not occur at random for these data for many reasons. An obvious reason is because there are crimes which people may be uncomfortable reporting, but also because people who have crimes committed against them are more likely to move, to be in jail, or not to be as trusting of interviewers. Models that account for nonrandom nonresponse should therefore be applied. Subjects within the same domain are assumed to have an equal probability of being victimized, but these probabilities are allowed to vary across domains, and also the probability that a subject responds is allowed to vary across domains and vary according to victimization status. Three parameters are modeled as coming from Beta distributions: the probability that an individual in a given domain is a victim, the probability a victim in a given domain responds, and the probability that a non-victim in a given domain responds. The entire data set is used to estimate the parameters of these models, which are then used as prior distributions for the Bayesian analysis, allowing the analysis to "borrow strength" from the full data to estimate the parameters for the small domains.

Nandram and Choi (2002) also use a Bayesian modeling approach for nonrandom nonresponse for binary data [18, 19]. Using a selection model approach they describe a model for nonignorable nonresponse and apply the methods to the NCS [18] and the National Health Interview Survey (NHIS) [19]. Both papers start with the models in Stasny (1991) [17] and Nandram and Choi (2000) [20]. These models are extended to create a nonignorable model that is centered around an ignorable model through a centering parameter with the null value of 1, implying the response does not depend on the value of the outcome (ignorable model). This is called a continuous expansion model.

For the NCS data, the expansion model for nonignorable nonresponse is

$$\begin{aligned}
&y_{ij}|p_i \overset{\text{iid}}{\sim} \text{Bernoulli}(p_i),\\
&r_{ij}|\{\pi_i, y_{ij}=0\} \overset{\text{iid}}{\sim} \text{Bernoulli}(\pi_i), \quad j=1,\ldots,n_i, i=1,\ldots,l,\\
&r_{ij}|\{\pi_i, \gamma_i, y_{ij}=1\} \overset{\text{iid}}{\sim} \text{Bernoulli}(\gamma_i\pi_i), \quad 0<\gamma_i\pi_i<1,
\end{aligned} \tag{2.3}$$

where $y_{ij} = 1$ if household j in domain i reports at least one crime and $y_{ij} = 0$ for households with no crime. The proportion of households with at least one crime, p_i, is allowed to vary across domains (areas), $i = 1,\ldots,l$. The probability that a household responds in area i is defined as $\delta_i = \pi_i\{\gamma_i p_i + (1-p_i)\}$. Response for household j in domain i, r_{ij}, is allowed to differ for those with $y_{ij} = 1$ and households with no crime ($y_{ij} = 0$). For $y_{ij} = 1$, the γ_i parameter is called a centering parameter, because it centers the nonignorable model around the ignorable model. Taking $\gamma_i = 1$ implies response does not depend on the value of y_{ij}, and leads to an ignorable nonresponsemodel. Nandram and Choi assume parameters $(p_i, \delta_i, \gamma_i)$ have a common distribution across all areas.

With the NCS data, Nandram and Choi examine the posterior mean, posterior standard deviation, and the 95% credible interval of the γ_i parameter. They found that 4 of the 10 intervals included the value 1, and thus do not provide evidence of nonignorability (in these 4 domains). They also provided the probability that $\gamma_i \leq 1$, and this probabilitiy is approximately 1 in the remaining 6 domains. This gives evidence for nonignorability in these 6 domains; the success rate (crime occurring) is lower for the respondents than for the nonrespondents.

Nandram and Choi (2002) [19] also apply their methods to The National Health Interview Survey (NHIS). In this study they consider inference on the proportion of households with at least one doctor visit in the previous 2 weeks (p_i). This can be considered an indicator of the health status in the United States. To fit the expansion model to the NHIS, the authors describe and use the Metropolis-Hastings algorithm with weighted importance sampling.

Through simulations, the authors compare two nonignorable expansion models, Stasny's nonignorable model, and an ignorable model for inference concerning p_i. One difference between the models described in this paper and the models Stasny described are that the hyper-parameters are estimated in the expansion model. All the nonignorable models produced similar results when the missingness was nonignorable, with the expansion models producing smaller intervals than the nonignorable model. When missingness was ignorable, all nonignorable models produced similar results to the ignorable model.

Baker et al. (2003) [21] performed a sensitivity analysis to study the impact of nonresponse on estimating the association between balance disorders and frequent depression. The data was from the Disability Supplement to the National Health Interview Survey. This had a complex survey design because only subjects in Phase 2 were asked about experiencing depression. Subjects entered Phase 2 if they answered “yes" to any one of a series of questions related to disability, and it is the view that these subjects are more likely to suffer from depression. There were 145,007 total adults in Phase 1, and 29,019 in Phase 2. 25,614 of these answered the depression question. Every subject was asked about balance difficulties. This resulted in three missingness scenarios for consideration: (i) missing only in depression (Y), (ii) missing in both depression (Y) and balance (X), and (iii) missing in depression (Y) with an auxiliary variable. This auxiliary variable is completely observed and strongly associatied with the partially observed Y. All models controlled for age, gender, race, health status and employment status. We only consider situation (i) because it is the most relevant for this dissertation. The outcome equaled 1 if the subject was treated frequently for depression and 0 otherwise. Let $M = 1$ if Y is missing, 0 otherwise. The model considered for Y is,

$$\begin{aligned} \text{logit}(\Pr(Y=1)) = \quad & \beta_0 \quad + \beta_B X_{\text{balance}} + \beta_A X_{\text{age}} + \beta_G X_{\text{gender}} \\ + \quad & \beta_R X_{\text{race}} + \beta_H X_{\text{health}} + \beta_W X_{\text{work}}. \end{aligned}$$

The only coefficient of interest to the authors was β_B, and they compared the corresponding odds ratios across different ignorable and nonignorable missing-data models. The model for missingness, using a main-effects ignorable missing-data model, was

$$\begin{aligned} \text{logit}(\Pr(M=0)) = \quad & \eta_0 \quad + \eta_B X_{\text{balance}} + \eta_A X_{\text{age}} + \eta_G X_{\text{gender}} \\ + \quad & \eta_R X_{\text{race}} + \eta_H X_{\text{health}} + \eta_W X_{\text{work}}. \end{aligned}$$

The main-effects nonignorable missing-data model considered adds either $(\eta_D Y)$ or $(\eta_D Y + \eta_{B*D} X_{\text{balance}} Y)$ in the above model, allowing missingness to depend on Y.

Other ignorable missing-data models considered were two- and three-way interaction hierarchical models. Similarly, to make these nonignorable, the same terms as above ($\eta_D Y$ or $\eta_D Y + \eta_{B*D} X_{\text{balance}} Y$) were added to the hierarchical models. The sensitivity analysis is a result of comparing the odds ratios ($\exp\{\beta_B\}$) from all of these different models. Baker et al. were particularly interested in making comparisons within a group of models, such as, comparing the ignorable model to the nonignorable models within the main effects model, and again within the two-way interaction models.

An Index of Local Sensitivity to Nonignorability (ISNI) is proposed in Troxel, et al. (2004) [22]. This builds upon the methods of Copas and Li (1997) [23] and provides the ability to conduct a sensitivity analysis without formally fitting any nonignorable nonresponse models. The ISNI can be used for generalized linear models with missing data on the outcome variable.

We assume there is an outcome Y subject to missingness, a set of fully observed predictors Z, and another set of fully observed predictors X that may overlap with Z. The distribution of Y, conditional on Z is represented by $f_\theta^{(Y_i|Z_i)}(y_i|z_i)$, and the interest is in estimating θ. Missingness in Y can be represented by a selection model,

$$Pr_\gamma\left(M_i = 0|Y_i = y_i, X_i = x_i\right) = h(\gamma_0' x_i + \gamma_1 y_i), \tag{2.4}$$

where we allow missingness to possibly depend on X and $h(.)$ is a function such as a logit or probit. It should be clear that if $\gamma_1 = 0$, the probability of missingness does not depend on y_i, resulting in the MAR mechanism. If $\gamma_0' = 0$ as well, then missingness is completely at random (MCAR).

The contribution to the likelihood for a subject who is observed is the likelihood of his observed values conditional on his covariates, multipled by the probability the subject is observed. The contribution to the likelihood for a subject who is unobserved can be represented as integrating over all possible values of y, multiplied by the probability the subject is missing. The idea behind the ISNI is to express the maximum likelihood estimate (MLE) of θ as a function of γ_1. The log-likelihood is expanded around values of the parameters which would lead to ignorability, i.e., $\theta =$

$\theta_0, \gamma_0 = \gamma_{00}, \gamma_1 = 0$, using a first order Taylor series approximation. This expansion is used to write the log likelihood as a function of (θ, γ_0) for fixed γ_1, to find the MLE of (θ, γ_0). Note that we will define $\nabla^2 L$ as the matrix of 2nd partial derivatives of L evaluated at $\theta = \hat{\theta}_0, \gamma_0 = \hat{\gamma}_{00}, \gamma_1 = 0$. The ISNI is,

$$\text{ISNI} = \frac{\partial \hat{\theta}(\gamma_1)}{\partial \gamma_1}\Big|_{\gamma_1=0} = -\left(\nabla^2 L_{11}\right)^{-1} \nabla^2 L_{13}. \tag{2.5}$$

The ISNI can be viewed as a rate of change and is interpreted as "the amount by which a unit change in the nonignorability parameter displaces the MLE of θ from its value $\hat{\theta}_0$ under the MAR model" ([22] page 1225).

The interpretation of the ISNI varies depending on the type of outcome (categorical or continuous) and the type of selection model used. For categorical outcomes, the interpretation is rather simple, but for continuous outcomes, the ability to transform Y to different units makes this interpretation more complex. In Chapter 4 we compare our proposed method to the methods provided here, in the context of a binary outcome, and thus we do not cover the continuous case further. For ease of interpretation the authors consider logistic selection models for all examples. This way, γ_1 can be interpreted as a "log odds ratio in the observation probability associated with a one-unit change in y." One use of the ISNI is that it can be used to adjust the MAR parameter estimate, $\hat{\theta}_0$, by noting, $\hat{\theta}(\gamma_1) \approx \hat{\theta}_0 + \text{ISNI}\gamma_1$. This gives an estimate assuming nonignorability, without ever fitting a nonignorable model. Another use of the ISNI is in assessing the degree of susceptibility to nonignorability. If the ratio of the ISNI to the standard error of the coefficient exceeds one, then the model estimates are highly susceptible to nonignorability.

For a categorical data example, the authors use data on the sexual behavior of students at the University of Edinburgh in 1993, that has been analyzed by a nonignorable model elsewhere [24]. We describe their findings and compare them to those from our proposed method in Section 4.7.2.

2.3 Imputation Approaches

We now review imputation approaches that attempt to account for possibly nonignorable missingness.

Greenlees et al., (1982) [25] propose a stochastic censoring model for imputation when the response probability depends on the unobserved value of Y. Greenlees et al. assume the selection model for factorization of the joint likelihood of the response, the covariates, and the response probability is modeled via a logistic function. The procedure has two main steps. The first step maximizes the likelihood of the respondents and nonrespondents to estimate the parameters for both models. The second step uses these MLEs, along with auxiliary variables X and other covariates Z, to calculate the conditional mean of Y on nonresponse in order to impute Y for the nonrespondents. They only use single imputation, and for the purposes of this analysis are not concerned with underestimation of the variance of Y, but suggest how to account for the uncertainty in imputation.

The method is applied to impute income of nonrespondents to the 1973 Current Population Survey (CPS). The authors were able to formally test the success of their imputations because a matched data set consisting of the CPS and data from the Social Security Administration (SSA) and the Internal Revenue Service (IRS) was available, allowing the nonrespondents of the CPS to have their income value available from the SSA or the IRS. They found evidence suggesting that subjects with higher income values were less likely to respond to the survey than subjects with lower income values, implying that nonresponse is nonignorable. A key point to note is that these authors estimate the parameter associated with Y in the response model, whereas in our proposed method we allow the parameter to take on several values to conduct a sensitivity analysis.

An ideal approach to reduce the final amount of nonresponse in a survey is to take a follow-up sample. A random sample of nonrespondents is selected and further attempts are made to get a response. Glynn, Laird and Rubin (1993) discuss mixture models with multiple imputation for nonignorable nonresponse when a follow-up sample has been selected [13]. They assume that every subject in the follow-up sample has responded, forming three final sets of subjects: original respondents, respondents who were originally nonrespondents, and nonrespondents. The standard approach for estimating the mean outcome in this case is to use the double-sampling procedure. This estimate is a weighted average of the estimate obtained from the respondents and the estimate obtained from the nonrespondents in the follow-up sample, where the weights are the proportions of respondents and nonrespondents, respectively. Glynn, Laird and Rubin discuss alternatives to this approach, which should be easier to implement in more complicated settings (such as nonignorable nonresponse). These are multiple imputation of the remaining nonrespondents, using either an Approximate Bayesian Bootstrap (ABB) or imputations from a normal distribution, and model-based approaches such as large-sample Bayes estimation (under a normal selection model including follow-up) and a fully Bayesian approach (under a normal mixture model). The ABB in this setting uses the follow-up sample to impute the nonrespondents. A bootstrap of the follow-up sample is taken, and then a simple random sample of size equal to the number of remaining nonrespondents is selected and imputed for each MI dataset.

There are relationships between the fully Bayesian approach, the double-sampling inference, and standard Bayesian inference (ignoring the indicator of response). Asymptotically, the normal mixture model Bayesian inference is nearly the same as the double-sampling inference. If all the nonrespondents are successfully contacted in the follow-up sample, then the normal mixture model is the same as the standard inference obtained by ignoring the response indicators when drawing inferences.

Carpenter and Kenward (2007) [26] develop a weighting approach applied to multiple imputation estimates under MAR, allowing the study of the sensitivity of estimates to the MAR assumption. The approach is fairly simple; the MI estimates assuming MAR of an effect can be obtained by software, and then re-weighted to estimate the effect under an MNAR model. Consider the situation where a treatment effect is of interest and the model for the probability of response is assumed to be:

$$\text{logit } \Pr(R_i = 1) = \alpha + \beta I[\text{patient } i \text{ on active trt}] + \gamma X_i + \delta Y_i \tag{2.6}$$

where,

$Y_i =$ response for unit i,

$X_i =$ baseline characteristics for unit i,

$R_i =$ response indicator,

$\alpha =$ adjusted log-odds of observing Y_i,

$\beta =$ adjusted change in log-odds ratio of observing Y_i if patient was randomized to the active treatment,

$\gamma =$ adjusted change in the log-odds of observing Y_i for a 1-unit change in baseline X_i,

$\delta =$ additional change in the log-odds of observing Y_i when the response changes by one unit.

We can see that, as typical when modeling the response probability under MNAR, δ determines how much missingness depends on values of Y. Obviously, if $\delta = 0$, then missingness does not depend on Y (but still depends on other variables) and the mechanism is MAR. The MAR model can be fitted to obtain estimates of (α, β, γ). Imputations of Y can also be performed assuming MAR to produce a set of completed data sets. Imputations only have to be performed once, under the assumption that $\delta = 0$. The value of δ can be varied for a sensitivity analysis, and for their application, the authors use $\delta = 0, .3, .5$.

To implement the procedure, assume patients $i = 1, \ldots, n_1$ withdraw from the study and have missing Y, and patients $i = n_{1+i}, \ldots, n$ are observed. For the patients to be imputed, denote the dth MAR imputation by Y_i^d. For each imputation, d, weights are created as,

$$\tilde{w}_d = \exp\left(\sum_{i=1}^{n_1} -\delta Y_i^d\right) \tag{2.7}$$

$$w_d = \frac{\tilde{w}_d}{\sum_{i=1}^{d} \tilde{w}_d}. \tag{2.8}$$

Then the re-weighted versions of the MI estimates for a parameter of interest θ and associated within (V_W) and between (V_B) variances are:

$$\hat{\theta}_{MNAR} = \sum_{m=1}^{D} w_d \hat{\theta}_d \tag{2.9}$$

$$\tilde{V}_W = \sum_{d=1}^{D} w_d \hat{\sigma}_d^2 \tag{2.10}$$

$$\tilde{V}_B = \sum_{d=1}^{D} w_d (\hat{\theta}_d - \hat{\theta}_{MNAR})^2 \tag{2.11}$$

$$V_{MNAR} \approx \tilde{V}_W + (1 + 1/D)\tilde{V}_B \tag{2.12}$$

When standard MI assuming MAR is applied, rarely are more than 20 imputations used, with mean and variance estimates not improving significantly past that. However, even with $D \geq 50$, Carpenter and Kenward find that the total variance of the MNAR model is likely to be underestimated because re-weighting decreases the effective sample size. This implies the degrees of freedom for the t-distribution of the MAR imputation estimator should be decreased for the MNAR estimator.

Ressequier et al. (2011) [27] developed an R package called “SensMice" to perform a sensitivity analysis when implementing multiple imputation. First, multiple imputation is performed under the assumption of ignorable missingness. The user can then specify parameters in the function ‘sens.mice’, that dictate the relationship between the outcome and response status. This parameter (or set of parameters), is denoted by θ, and is either the odds ratio (OR) comparing the outcome among subjects with missing data to the outcome among subjects who are observed, or if the outcome is continuous, θ is the difference in the expected values for missing and observed.

Ressequier et al. applied this procedure and function to the R data set “CHAIN”, to study the relationship between poor mental health and self-reported viral load.

When considering viral load as a binary variable, the authors chose θ values of 1.2, 1.5, 2.0 (representing ORs), all indicating the belief that nonresponders have higher viral load than responders. Imputed data sets are then created for each of the values of θ. In their example, the MAR estimated OR between viral load and mental health is 2.01, whereas for MNAR with both $\theta = 1.2$ and 1.5, the OR is 1.73; for MNAR with $\theta = 2.0$, the OR is 1.75. The 95% CI is still greater than 1 for all instances. The OR for the outcome poor mental health decreases as the percent viral load is allowed to increase among nonrespondents.

The extension of hot deck imputation to nonignorable nonresponse has not been well examined. Rubin and Schenker (1991) [28] briefly describe how an ABB could be applied to handle nonignorable missing data. An ABB is one method that can be used to make hot deck MI a proper MI method. To introduce the hot deck ABB, we let n_{obs} be the number of subjects with observed Y values, Y_{obs}. An ABB draws n_{obs} subjects randomly with replacement from observed subjects. For each imputation, the set of n_{obs} values is used to create the donor pool, which is then used for hot deck imputation. Estimates of interest (for example, the mean of Y), are found for each imputation, then combined using the standard rules for MI as formulated by Rubin [1]. To extend this MI idea to nonignorability, Rubin and Schenker suggest changing the way the ABB is drawn, however, they do not study this procedure further.

Siddique and Belin [29] take the suggestion from Rubin and Schenker [28] to develop a nonignorable hot deck procedure. It builds upon the ignorable ABB hot deck Siddique and Belin develop in [30] by adjusting how the bootstrap sample of respondents is drawn. Instead of drawing from the respondents with equal probability (which corresponds to an ignorable hot deck procedure), the respondents are drawn with probability proportional to a function of Y. Hence, the probability of selection to the bootstrap sample for $y_i \in Y_{obs}$ is

$$\frac{y_i^c}{\sum_{j=1}^{n_{obs}} y_j^c}. \tag{2.13}$$

Siddique and Belin consider probabilities proportional to Y^c for values of $c = \{-1, 0, 1, 2, 3\}$. If large values of Y are assumed missing, c is chosen to be positive, whereas smaller values of Y assumed missing correspond to $c < 0$, and $c = 0$ corresponds to the ignorable hot deck. Additionally, Siddique and Belin consider probabilities proportional to how far away an observation is from a certain quantile, and coin names such as "U-shaped ABB" (for values further from the median with higher probabilities), and "Fishhook ABB" (using the distance from the first quantile). It is important to note that if values of Y are nonpositive, they need to be transformed to avoid numerical problems in the selection probabilities.

Once the bootstrap sample of nonrespondents is drawn, predictive mean matching is used on this skewed bootstrap sample. The observed values of the variable to be imputed are regressed on a chosen set of predictor variables, and using the estimated regression parameters, predicted values ($\hat{Y}$'s) are found for every subject (nonrespondents and all n_{obs} bootstrapped respondents). These predicted values are used to calculate the distance between a single nonrespondent and each respondent. The probability a respondent is chosen for imputation for a nonrespondent depends on this distance. Siddique and Belin's distance measure allows every respondent (in the bootstrap sample) to be in the donor pool and is defined as:

$$D_{0i}^k = (|\hat{y}_0 - \hat{y}_i| + \delta)^k. \tag{2.14}$$

D_{0i}^{k} is the distance between nonrespondent 0 and bootstrap sample respondent i, for "closeness" parameter k. This "closeness" parameter is used to determine the distribution of selection probabilities, with $k = 0$ equivalent to a simple random hot deck, and as k approaches ∞, a nearest neighbor hot deck. Siddique and Belin suggest a value of $k = 3$ to form a good spread of the selection probabilities [30]. Here, δ is the minimum nonzero absolute distance, and is included to insure there are no problems with division by zero in the probability of selection. Because it might be difficult before hand to select a single value of c, the authors suggest mixing the values among the MI data sets, which they call a "Mixture ABB". Their implementation uses $c = \{-1, 0, 1, 2, 3\}$ to create 5 separate MI data sets. This "averages" over a set of possible missingness mechanisms, also including the ignorable nonresponse mechanism, $c = 0$. We feel that a sensitivity analysis is more insightful and prefer an approach that compares inferences from different assumptions on the missingness mechanism (e.g., compares inference using different c values).

As stated above, Siddique and Belin suggest creating five imputation datasets each with a different value of c and use MI combining rules to obtain final estimates. This 'averages' over the different assumptions. However, as discussed in the 2012 paper by Siddique, Harel and Crespi [31], the standard MI combining rules do not apply in this situation. This is because each MI set is assuming a different model for the missingness, and the standard rules do not account for this added uncertainty.

Siddique, Harel and Crespi discuss combining rules for nested multiple imputations, with the goal of allowing researchers to have a way to use complete-data methods as opposed to model-based methods. They believe that "all imputation model uncertainty should be incorporated into one inference." Some reasons for doing this might be the need to have different model assumptions based on different time points (in a longitudinal study), or perhaps based on different groups of subjects (dropouts versus non-dropouts). They suggest choosing M missingness models, and generating N imputed datasets within each M, for a total of $M \times N$ imputation sets. The overall estimate of the quantity of interest is simply the average of the $M \times N$ point estimates. The variance of the quantity of interest has three sources of variability (as opposed to the two sources with standard MI estimates). These are the overall average of the associated variance estimates (similar to the standard within imputation variance), the within-model variance (the variance of the estimates for a given M, averaged over M), and the between-model variance (variability between the mean for M and the overall M - similar to the traditional between imputation estimate, but for a specified M). If only one model is used, the variance simplifies to the standard MI rules.

An alternative imputation method for nonignorable nonresponse is that of Kim and Kim (2012) [32], who extend Parametric Fractional Imputation (PFI) to handle nonignorable data. The data consist of fully observed $\mathbf{x}$ and a single y that is possibly MNAR. The interest is in finding the MLE of θ (a population parameter) which can be found by solving the score equations when complete data are available. Because the mechanism is MNAR, the population parameter, θ, and the parameter for the response mechanism, ϕ, must be solved for simultaneously. Under a selection model framework for the factorization of the likelihood, PFI is proposed to jointly estimate (θ, ϕ).

The observed likelihood of (θ, ϕ) is defined as,

$$L_{obs}(\theta, \phi) = \prod_{i=1}^{n} f_{obs}(m_i y_i, m_i | \mathbf{x}_i; \theta, \phi) \tag{2.15}$$

where,

$$f_{obs}(m_i y_i, m_i|\mathbf{x}_i;\theta,\phi) = \begin{cases} f_1(y_i|\mathbf{x}_i;\theta)f_2(m_i|\mathbf{x}_i,y_i;\phi), & \text{if } m=0 \\ g(\mathbf{x}_i,m_i;\theta,\phi) = \int f_1(y_i|\mathbf{x}_i;\theta)f_2(m_i|\mathbf{x}_i,y_i;\phi)dy_i, & \text{if } m=1 \end{cases} \tag{2.16}$$

Thus for the respondents the likelihood is factorized according to a selection model, and for nonrespondents, their contribution to the likelihood is integrated over the possible values of y_i.

Imputed values are generated D times for each subject with missingness, and we denote them by $y^{*(1)}_{i.mis},\ldots,y^{*(D)}_{i.mis}$. Let $y^*_{id} = (y_{i.obs}, y^{*(d)}_{i.mis})$, where $i = 1,\ldots,n$, and $d = 1,\ldots,D$. Each of the y^*_{id} values is assigned a fractional weight, $w^*_{i1},\ldots,w^*_{iD}$. These weights are importance sampling weights. The weights have a simple form when y_i is catogorical with J categories. In this case, $D = J$ imputations are used, and the corresponding fractional weights are:

$$w^*_{id} = \frac{\Pr(y_i = y_i^{*(d)}|\mathbf{x}_i,\hat{\theta})\Pr(m_i = 1|\mathbf{x}_i, y_i^{*(d)},\hat{\phi})}{\sum_{j=1}^{J}\Pr(y_i = y_i^{*(j)}|\mathbf{x}_i,\hat{\theta})\Pr(m_i = 1|\mathbf{x}_i, y_i^{*(j)},\hat{\phi})}.$$

When y_i is continuous, the fractional weights can be written as,

$$w^*_{id} = w^*_{id0} \times \frac{\{1-\pi(\mathbf{x}_i, y_i^{*(d)},\hat{\phi})\}}{\sum_{k=1}^{D} w^*_{id0}\{1-\pi(\mathbf{x}_i, y_i^{*(j)},\hat{\phi})\}}$$

where w^*_{id0} are the importance weights for $f(y_i|\mathbf{x}_i,\hat{\theta})$ assigned to $y_i^{*(j)}$, satisfying $\sum_{d=1}^{D} w^*_{id0} = 1$.

To avoid large fractional weights, Kim and Kim suggest generating imputed values from

$$\hat{f}(y_i|\mathbf{x}_i, r=0) = \frac{\hat{f}(y_i|\mathbf{x}_i, m=0)\{1/\pi(\mathbf{x}_i,y_i;\hat{\phi}^{(0)})-1\}}{\int \hat{f}(y_i|\mathbf{x}_i, m=0)\{1/\pi(\mathbf{x}_i,y_i;\hat{\phi}^{(0)})-1\}dy_i} \tag{2.17}$$

If one takes the logit$(\pi(\mathbf{x}_i, y_i;\phi)) = \phi_0 + \phi_1\mathbf{x}_i + \phi_2 y_i$, then (2.17) reduces to the case where the density of the nonrespondents is an exponential titling of the density of the respondents. This is discussed in Kim and Yu (2009) [33]. After performing imputation, the final estimates of the parameters are found using the weights. If we have complete data with corresponding sampling weights, w_i, and a consistent estimate of θ is found by solving

$$\sum_i w_i U(\mathbf{x}_i, y_i; \theta) = \mathbf{0} \tag{2.18}$$

then using the fractionally imputed data, a consistent estimate can be found by solving:

$$\sum_i \sum_{j=1}^{D} w_i w_{ij}^* U(\mathbf{x}_i, y_i^{*(j)}; \theta) = \mathbf{0} \tag{2.19}$$

The authors discuss choices for the proposal distribution, and changes which can be made to avoid extremely large fractional weights. They also discuss calibration weights to improve approximations for moderate D. For variance estimation they consider jackknife and bootstrap, and give formulas for the replicated fractional weights for both PFI and the calibrated fractional imputation (CFI).

In a subsequent article by Kim (2012) [34], the fractional imputation method is applied to nonignorable missing categorical data, when follow-up data may be available. Follow-up data may arise in the case when a sample of nonrespondents are randomly selected for follow-up (or to be re-contacted) and it is often assumed that the follow-up sample all respond. This results in three types of subjects: first-time respondents, respondents in the follow-up sample, and the remaining nonrespondents. The follow-up sample is used to fix the problem of parameter non-identifiability that arises when modeling nonignorability. When applying the PFI to partially missing categorical data, the D imputations are the D possible values of the categorical variable. The fractional weights (importance sampling weights) can be calculated using a simple application of Bayes' rule.

Kim conducted simulations to compare the complete sample estimators, the fractional imputation estimator, and the MI estimator with 10 and 100 imputations for each missing value. All results show that all estimators are mostly unbiased, but the fractional imputation estimator is more efficient than the multiple imputation estimator. This can be explained because fractional imputation is a deterministic method, whereas multiple imputation is a stochastic imputation method.

Another approach to imputing nonignorable missing data is the proxy pattern-mixture (PPM) model, developed by Andridge and Little (2011) [35]. They assume that a fully observed set of covariates Z is available (for both respondents and nonrespondents). The method first creates a proxy X for the partially missing variable Y by regressing Y on Z for respondents, and then taking X to be the predicted values using Z from all subjects. This step is similar to the predicted means calculated for the purposes of calculating distances in the Siddique and Belin (2008) hot deck [30], though their use of these means is completely different. They assume that the joint distribution of Y, X, and the missingness indicator M follows a bivariate normal pattern-mixture model [11, 12]:

$$\begin{aligned} (Y, X|M=m) &\sim N_2\left(\left(\mu_y^{(m)}, \mu_x^{(m)}\right), \Sigma^{(m)}\right), m = 0, 1 \qquad (2.20) \\ M &\sim \text{Bernoulli}(1-\pi) \\ \Sigma^{(m)} &= \begin{bmatrix} \sigma_{yy}^{(m)} & \rho^{(m)}\sqrt{\sigma_{yy}^{(m)}\sigma_{xx}^{(m)}} \\ \rho^{(m)}\sqrt{\sigma_{yy}^{(m)}\sigma_{xx}^{(m)}} & \sigma_{xx}^{(m)} \end{bmatrix} \end{aligned}$$

Here and throughout, we use a superscript (0) to denote $m = 0$ (respondent), and superscript (1) to denote $m = 1$ (nonrespondent). Likewise, a subscript R refers to respondents, and NR refers to nonrespondents. For example, $\mu_y^{(0)}$ is the population mean of Y for the respondents, whereas $\bar{y}_R$ is the sample mean of Y for the respondents.

The pattern-mixture model allows means and variances to be different for respondents ($m = 0$) and nonrespondents ($m = 1$). However, the model is underidentified. All the parameters of the respondent distribution for X and Y, $\{\mu_y^{(0)}, \mu_x^{(0)}, \sigma_{xx}^{(0)}, \sigma_{yy}^{(0)}, \rho^{(0)}\}$, and those dealing with the marginal distribution of the nonrespondent proxy X, $\{\mu_x^{(1)}, \sigma_{xx}^{(1)}\}$, are identified and easily estimable. However, the remaining parameters, $\mu_y^{(1)}, \sigma_{yy}^{(1)}$, and $\rho^{(1)}$, are not identifiable without making further assumptions.

Andridge and Little (2011) use assumptions on the missingness mechanism to make these parameters identifiable. They assume that the probability of nonresponse is a function of a linear combination of the proxy X and the outcome Y, with the proxy scaled to have the same variance as Y for the respondents,

$$\Pr\left(M = 1 | Y, X\sqrt{\frac{\sigma_{yy}^{(0)}}{\sigma_{xx}^{(0)}}}\right) = f\left(X\sqrt{\frac{\sigma_{yy}^{(0)}}{\sigma_{xx}^{(0)}}} + \lambda Y\right). \tag{2.21}$$

The sensitivity parameter, λ, determines the missingness mechanism, with $\lambda = 0$ implying an MAR mechanism and $\lambda = \infty$ implying a type of "extreme" MNAR where missingness depends entirely on Y. The assumption in (2.21), with f an unspecified function, just identifies the parameters of the model. Using these identifying restrictions, Andridge and Little estimate the marginal mean of Y and assess its sensitivity to various deviations away from MAR by varying the value of λ. They discuss and compare three methods for estimating this mean: maximum likelihood, a fully Bayesian approach, and multiple imputation.

Chapter 3: Continuous Outcome

3.1 Introduction

In this chapter we propose a new hot deck method for imputing nonignorable missing data for continuous outcomes which we call the proxy pattern-mixture (PPM) hot deck. Our method combines the distance measure-based selection probabilities of Siddique and Belin (2008) [29] with the PPM model of Andridge and Little (2011) [35]. Unlike Siddique and Belin, who incorporate nonignorability in the ABB step of the hot deck, we use an ignorable (equal probability) ABB and instead allow for nonignorability in the creation of the predicted means used to calculate donor selection probabilities. We use the PPM model conditional on a chosen value of λ to define predicted means and calculate distances from donors to nonrespondents. The entire process of creating the bootstrap sample of donors (the ABB), calculating distances from donor to nonrespondent, calculating selection probabilities, and randomly selecting donors is repeated for a set of values of λ. Doing this provides a sensitivity analysis that yields imputed "complete" data sets under different assumptions on the missingness mechanism.

3.2 Predicted Values From the Proxy Pattern-Mixture Model

To integrate the proxy pattern-mixture model into the hot deck, we need to calculate predicted means for the nonrespondents and the respondents for use in the distance function. Under the PPM model given by (2.20) and (2.21), the conditional distribution of the outcome Y given the proxy X for respondent status $M = m$ is:

$$[y_i|x_i, m_i = m] \sim N\left(\mu_y^{(m)} + \rho^{(m)}\sqrt{\frac{\sigma_{yy}^{(m)}}{\sigma_{xx}^{(m)}}}\left(x_i - \mu_x^{(m)}\right), \sigma_{yy}^{(m)} - \frac{{\sigma_{yx}^{(m)}}^2}{\sigma_{xx}^{(m)}}\right) \tag{3.1}$$

To obtain predicted means, $\hat{y}$, we need only consider the expectation of Y for the nonrespondents and respondents, $E[Y_i|x_i, m_i = m]$.

For respondents, every parameter of the mean shown in (3.1) is identified, and maximum likelihood estimates can be substituted into the equation to obtain the estimated predicted mean for the i^{th} respondent as:

$$\hat{y}_i^{(0)} = \bar{y}_R + \hat{\rho}^{(0)}\sqrt{\frac{s_{yy}^{(0)}}{s_{xx}^{(0)}}}\left(x_i - \bar{x}_R\right). \tag{3.2}$$

For nonrespondents, the identifying restrictions in (2.21) yield expressions for $\mu_y^{(1)}$ and $\rho^{(1)}$ in terms of identified parameters, and substitution of these quantities into (3.1) yields:

$$\begin{aligned} E[Y_i|X_i = x_i, m_i = 1] = \mu_y^{(0)} &+ \sqrt{\frac{\sigma_{yy}^{(0)}}{\sigma_{xx}^{(0)}}}\left(\frac{\lambda + \rho^{(0)}}{1 + \lambda\rho^{(0)}}\right)\left(\mu_x^{(1)} - \mu_x^{(0)}\right) \\ &+ \left[\frac{\sigma_{xy}^{(0)}}{\sigma_{xx}^{(1)}} + \sqrt{\frac{\sigma_{yy}^{(0)}}{\sigma_{xx}^{(0)}}}\left(\frac{\lambda + \rho^{(0)}}{1 + \lambda\rho^{(0)}}\right)\frac{\left(\sigma_{xx}^{(1)} - \sigma_{xx}^{(0)}\right)}{\sigma_{xx}^{(1)}}\right]\left(x_i - \mu_x^{(1)}\right). \end{aligned} \tag{3.3}$$

Substitution of maximum likelihood estimates for all parameters yields the predicted mean for nonrespondent i for a given λ:

$$\begin{aligned}\hat{y}_i^{(1,\lambda=\lambda)} &= \bar{y}_R + \left(\frac{\lambda + \hat{\rho}^{(0)}}{1+\lambda\hat{\rho}^{(0)}}\right)\sqrt{\frac{s_{yy}^{(0)}}{s_{xx}^{(0)}}}\left(\bar{x}_{NR} - \bar{x}_R\right) \\ &\quad + \left[\frac{s_{xy}^{(0)}}{s_{xx}^{(1)}} + \sqrt{\frac{s_{yy}^{(0)}}{s_{xx}^{(0)}}}\left(\frac{\lambda + \hat{\rho}^{(0)}}{1+\lambda\hat{\rho}^{(0)}}\right)\left(1 - \frac{s_{xx}^{(0)}}{s_{xx}^{(1)}}\right)\right]\left(x_i - \bar{x}_{NR}\right).\end{aligned} \tag{3.4}$$

The equation for the predicted mean of the nonrespondents depends on the value of λ. Following Andridge and Little (2011) and Little (1994), we perform a sensitive analysis by comparing estimates obtained using varying λ values and suggest $\lambda = \{0, 1, \infty\}$. We select nonnegative values of λ because X and Y should be positively correlated since X is a proxy for Y [12]. Estimates of the chosen parameter of interest (e.g., survey outcome mean) from $\lambda = 0$ assume MAR, while estimates from $\lambda = \infty$ assume "extreme" MNAR where missingness depends entirely on Y. The case of $\lambda = 1$ provides an in-between case where missingness depends equally on X and Y. Specifically, the values of the predicted means for the ith nonrespondent are written below for the three values of λ just discussed.

$$\begin{aligned}\hat{y}_i^{(1,\lambda=0)} &= \hat{y}_i^{(0)} \\ \hat{y}_i^{(1,\lambda=1)} &= \hat{y}_i^{(1,\lambda=0)} + \frac{s_{xy}^{(0)}}{s_{xx}^{(0)}}\left(\frac{1-\hat{\rho}^{(0)}}{\hat{\rho}^{(0)}}\right)\left(x_i - \bar{x}_R\right) + \frac{s_{xy}^{(0)}}{s_{xx}^{(1)}}\left(\frac{\hat{\rho}^{(0)}-1}{\hat{\rho}^{(0)}}\right)\left(x_i - \bar{x}_{NR}\right) \\ \hat{y}_i^{(1,\lambda=\infty)} &= \hat{y}_i^{(1,\lambda=0)} + \frac{\hat{\rho}^{(0)}+1}{\hat{\rho}^{(0)}}\left\{\frac{s_{xy}^{(0)}}{s_{xx}^{(0)}}\left(\frac{1-\hat{\rho}^{(0)}}{\hat{\rho}^{(0)}}\right)\left(x_i - \bar{x}_R\right) + \frac{s_{xy}^{(0)}}{s_{xx}^{(1)}}\left(\frac{\hat{\rho}^{(0)}-1}{\hat{\rho}^{(0)}}\right)\left(x_i - \bar{x}_{NR}\right)\right\}\end{aligned} \tag{3.5}$$

For the nonrespondents, we have written the equations for $\lambda = 1$ and $\lambda = \infty$ as functions of the predicted mean assuming MAR, $\hat{y}_i^{(0)}$. For a given nonrespondent, we can see the deviation from the MAR model for the different λ's. The value of the predicted mean for a nonrespondent will either increase or decrease for increasing λ, depending on the location of x_i in relation to $\bar{x}_{NR}$ and $\bar{x}_R$ (and the values of ρ, s_{xy}

and s_{xx}). Another observation is that if X is a weak proxy for Y meaning that $\hat{\rho}^{(0)}$ is close to zero, the predicted means under a strong nonignorable assumption can become very large. This will cause problems when one is trying to find matches and is discussed in Sections 3.4 and 3.5.

3.3 Steps of the PPM Hot Deck

We now outline the steps of the PPM hot deck. For imputation 1 (of D):

1. **Bootstrap**: Generate a bootstrap sample of the respondents by selecting n_{obs} respondents with replacement. Denote these respondent outcomes and covariate values as $\{Y_j^b, \mathbf{Z}_j^b\}$, where b distinguishes the bootstrap sample from the original sample.

2. **Proxy**: Regress Y^b on $\mathbf{Z}^b$ to create the proxy X for all nonrespondents and the bootstrap sample of respondents. Note that only respondents who are selected into the bootstrap sample have a proxy created, while all nonrespondents have a proxy created.

3. **Predicted Values**: Calculate predicted values using maximum likelihood estimates from the bootstrapped sample of respondents and the entire sample of nonrespondents based on a chosen value of λ. Note that for respondents the predicted means do not depend on λ. However, they vary for each cycle through the PPM hot deck because of the bootstrapping step. The predicted values are

given by:

$$
\begin{aligned}
\hat{y}_i^{b(0)} &= \bar{y}_R^b + \frac{s_{xy}^{b(0)}}{s_{xx}^{b(0)}} \left(x_i^b - \bar{x}_R^b\right) \\
\hat{y}_i^{(1,\lambda=\lambda)} &= \bar{y}_R^b + \sqrt{\frac{s_{yy}^{b(0)}}{s_{xx}^{b(0)}}} \left(\frac{\lambda + \hat{\rho}^{b(0)}}{1 + \lambda\hat{\rho}^{b(0)}}\right) \left(\bar{x}_{NR} - \bar{x}_R^b\right) \\
&\quad + \left[\frac{s_{xy}^{b(0)}}{s_{xx}^{(1)}} + \left(\sqrt{\frac{s_{yy}^{b(0)}}{s_{xx}^{b(0)}}} \cdot \frac{\lambda + \hat{\rho}^{b(0)}}{\lambda\hat{\rho}^{b(0)} + 1}\right) \left(1 - \frac{s_{xx}^{b(0)}}{s_{xx}^{(1)}}\right)\right] \left(x_i - \bar{x}_{NR}\right).
\end{aligned}
\tag{3.6}
$$

We use the superscript b to denote quantities that are calculated on a bootstrapped sample, to distinguish them from quantities calculated on the whole sample. For example, $\bar{x}_R^b$ is the mean of X for the bootstrapped sample of respondents, while $\bar{x}_{NR}$ is the mean of X for the entire sample of nonrespondents.

4. **Distances**: Calculate the distance between bootstrap donor i and donee 0, using the distance measure of Siddique and Belin, given in (2.14). The distance measure requires a value for the “closeness” parameter k; we use $k = 3$ as recommended by Siddique and Belin (2008) [29].

5. **Donor Selection Probabilities**: Calculate the bootstrap donor i selection probability $l_i^{b,k}(\hat{y}_0^{(1,\lambda)})$ proportional to the distances:

$$
l_i^{b,k}(\hat{y}_0^{(1,\lambda)}) = \frac{\frac{1}{D_{0i}^k}}{\sum_{i=1}^{n_{obs}^b} \frac{1}{D_{0i}^k}}
\tag{3.7}
$$

6. **Select and Impute**: Randomly select a bootstrap donor i for donee (nonrespondent) 0 using the selection probabilities $l_i^{b,k}(\hat{y}_0^{(1,\lambda)})$. Impute the donor’s value of Y for the missing value of the donee.

7. **Repeat** steps (1-6) D times to create D complete datasets, composed of the original (pre-bootstrap) respondents and the imputed nonrespondent data. This

entire process should be completed separately for each value of λ in the sensitivity analysis.

Once the PPM hot deck has produced a set of completed data sets, standard complete data estimates can be produced for each data set and combined with standard combining rules. This will yield an estimate for a single value of λ, and the entire process is repeated for a different value of λ that assumes a different degree of nonignorability, from MAR ($\lambda = 0$) to extreme MNAR ($\lambda = \infty$).

The key difference between the method of Siddique and Belin and our proposed method is where the nonignorability adjustment occurs. In Siddique and Belin's method, nonignorability is handled in the approximate Bayesian bootstrap step. The distribution of the bootstrap sample takes different "shapes" depending on the assumption of how missingness depends on Y. If it is believed that subjects with larger values of Y are missing more often, then the bootstrap sample is set up to sample larger values of Y with higher probability, skewing the distribution. Once the nonignorable ABB sample is formed, the steps which follow are the same as they would be if the ABB were performed assuming ignorable missingness. In contrast, the PPM hot deck uses a simple ignorable ABB, and adjusts for nonignorability in the creation of selection probabilities for donors in the donor pool.

3.4 Graphical Display of Proposed Method

Figure 3.1 illustrates the steps of the proposed hot deck method for a single imputation data set by showing the distributions of various proxy values for a simulated data set (details on data generation are in Section 3.6.1). We focus first on Figure 3.1a which is for $\rho = 0.8$ and an MAR missingness mechanism, where missingness depends on Z, denoted [Z]. Shown in the top panel are the values of Y prior to deletion (True Y), the distributions of both the observed Y values (Observed Y) and those that were deleted (Missing Y), followed by the distribution of the bootstrap sample of respondent Y values (Bootstrap Y). The distribution of the bootstrap sample shows how many times a specific subject was chosen. The 2nd through 4th panels correspond to λ values of 0, 1 and ∞, respectively, and show the distributions of the predicted means ($\hat{Y}$) for the respondents (observed) and nonrespondents (missing). The last distribution (in red) shows the values of the observed Y that were imputed for the nonrespondents. With the MAR mechanism, [Z], we still observe a few respondents with large and small values of Y. In fact, the range of the observed Y is larger than that of the missing Y. For this particular imputation set, only one of the largest subjects and one of the smallest subjects were chosen for the bootstrap sample. This is an interesting observation because we can see how sparseness of donors for some covariate values might increase if specific donors are not selected for the bootstrap sample. This implies that the quality of donors could change between bootstrap samples, but the bootstrap is necessary for the extra variability required for proper MI. Examining the distribution of predicted means $\hat{Y}$, we note that the values do not change for the respondents as λ changes (as expected since (3.2) does not depend on λ). The predicted means do change for the nonrespondents; for this data set the values only slightly increase as λ increases. This also results in the largest observed subjects being imputed when $\lambda = 1$ and $\lambda = \infty$, but not in the case when $\lambda = 0$. Since the true missingness mechanism is [Z], there are still large values that can be imputed.

Figure 3.1 also demonstrates a potential problem in applying hot deck to nonignorable missing data. If we compare Figures 3.1a and 3.1b, which have the same correlation but different missingness mechanisms, we see a common problem of hot deck. When missingness depends on Y^2, denoted $[Y^2]$, such that large values of Y are missing, we see that even though the predicted means for the nonrespondents get larger as λ increases, we can only impute the largest observed value of Y. In the PPM hot deck setting, this is also the largest observed value of Y that is in the bootstrap sample for that imputation. This is the case when a parametric version of the PPM will outperform the hot deck, because it does not rely on the observed values for imputation.

We can also compare Figures 3.1a and 3.1c. These have the same missing data mechanism, $[Z]$, but 3.1a has $\rho = 0.8$, and 3.1c has $\rho = 0.2$. The main observation here is how large the values of the predicted means are for the nonrespondents when $\lambda = \infty$. This is due to the formula for the predicted means when the correlation is small and λ is large, given by (3.5). There is complete separation of predicted values between the respondents and the nonrespondents. When the procedure searches for respondents who are "close" to nonrespondents, there are none. In this case, all the distances are so large that the selection probabilities become nearly equal. Thus, the PPM hot deck selects donors with approximately equal probability, and yields results similar to the complete case analysis (MCAR).

We now propose several distance metrics to be able to tell when the predicted means are too far away, indicating that there are not enough large (or small) values of Y observed to estimate the mean in the sensitivity analysis even when using $\lambda = \infty$.

3.5 Diagnostics for Donor Quality

We want our imputed values to represent a reasonable value for each nonrespondent under the assumed missingness mechanism. This means that we not only need a large enough sample of respondents to ensure enough potential donors, but they must also exist across the values of covariates related to missingness [5]. This is an especially important consideration in the presence of nonignorable missingness. If, for example, larger values of Y are missing more often than smaller values, there will be fewer potential donors (or possibly none) with large values of Y available for imputation. Unlike a parametric imputation procedure, hot deck imputation is unable to extrapolate, since only observed values are available for imputation.

After a search of the existing literature we were unable to find commonly used metrics for measuring the quality of donors that we could adapt for the PPM hot deck. This may be because many common hot deck procedures create cells of donors, instead of allowing all respondents to serve as potential donors, and thus the number of potential donors in a cell is itself a measure of donor quality (i.e., donor availability). We therefore propose two donor quality metrics for use with the PPM hot deck to allow us to identify situations in which high quality donors are unavailable. The measures are dependent on the choice of λ because, as we will illustrate in our simulation study, quality donors may be available under an MAR assumption ($\lambda = 0$) but not under an MNAR assumption ($\lambda = \infty$).

The first donor quality metric is based on the average distance from donor to donee across all respondents and nonrespondents. We call this metric the Mean Minimum Distance (MMD). For each nonrespondent j, we calculate the minimum absolute distance to a respondent, δ_j, where $\delta_j = \min_i |\hat{y}_j^{(1,\lambda=\lambda)} - \hat{y}_i^{(0)}|; i = 1, \ldots, n_{obs}$.

To obtain the MMD these minimum differences are averaged over the nonrespondents and standardized by the standard deviation of the outcome among respondents:

$$MMD = \frac{1}{n_{mis}} \sum_{j=1}^{n_{mis}} \frac{\delta_j}{s_{yy}^{(0)}} \tag{3.8}$$

A large value of MMD indicates that the average distance from donor to donee is high. The main purpose of the metric is to identify data sets and specific values of λ where the bulk of the imputations would come from poor donors. However, individual values of δ_j could also be examined to find specific nonrespondents for whom there is no quality match.

The second donor quality metric is based on the donor selection probabilities themselves rather than on the distances. This metric is motivated by the fact that if there are no close donors and all distances are very large there will be little variation in the selection probabilities, resulting effectively in a simple random sample of the respondents used for imputation. We call this metric the Mean Variance Selection Probability (MVSP). The MVSP is obtained by first calculating the variance of the donor selection probabilities for each nonrespondent and then averaging these variances over nonrespondents,

$$MVSP = \frac{1}{n_{mis}} \sum_{j=1}^{n_{mis}} Var\left(l_i^k(\hat{y}_j^{(1,\lambda=\lambda)})\right). \tag{3.9}$$

If the MVSP is close to zero, this indicates that on average (across nonrespondents) the donor selection probabilities are close to equal, which effectively results in the PPM hot deck imputing via a simple random hot deck. This can happen when the predicted means for nonrespondents are very far away from the predicted means for the respondents, even when the closeness parameter k used in the distance measure is non-zero. As with the MMD, closer inspection of the individual variances could be used to identify specific nonrespondents with poor donor quality.

We propose calculating both donor quality metrics using the complete sample of respondents (i.e., not a bootstrap sample), though of course in one particular bootstrap sample the quality of available donors could be better than in another. These metrics could alternatively be computed for a particular bootstrap sample, though we do not consider this further as on average across the bootstrap samples of the PPM hot deck the metrics should be the same as in the pre-bootstrap sample. We suggest using both the MMD and the MVSP to determine when close matches are unobtainable.

3.6 Simulation Studies

We conducted a simulation study to assess the performance of the proposed PPM hot deck as well as to illustrate the use of the donor quality metrics. In addition to assessing bias and coverage of the new method under a range of true missingness mechanisms, we also sought to compare its performance to the Siddique and Belin nonignorable hot deck (SB hot deck).

3.6.1 Data Generation

Complete data for an outcome y_i and single covariate z_i were generated from a bivariate normal distribution, such that

$$(Z_i, Y_i) \sim N_2\left(\left(\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right), \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)\right), \quad i = 1, \ldots, n. \tag{3.10}$$

We considered two sample sizes, $n = \{400, 800\}$, and three different correlations, $\rho = \{0.8, 0.5, 0.2\}$. We note that the correlation ρ is different from the correlations $\rho^{(0)}$

and $\rho^{(1)}$ in the parametric PPM model in (2.20), since here the outcome and proxy will be unconditionally jointly normal, but in the PPM model they are bivariate normal conditional on the missing data indicator. As discussed in Andridge (2011), when missingness is linear in Z and/or Y, correlation between the proxy X and Y (before deletion) will be equal to ρ for both respondents and nonrespondents. However, if missingness is quadratic in Z and/or Y, the correlation between Y and X will be different for respondents and nonrespondents.

The missing data indicator M was generated according to a logistic regression model,

$$\text{logit}(\Pr(m_i = 1)|y_i, z_i)) = \gamma_0 + \gamma_Z z_i + \gamma_Y y_i + \gamma_{Y2} y_i^2, \tag{3.11}$$

and values of y_i were deleted when $m_i = 1$. Values of $\{\gamma_Z, \gamma_Y, \gamma_{Y2}\}$ were chosen to induce either MAR or MNAR mechanisms as shown in Table 3.1. The value of γ_0 was chosen to induce either 25% missingness or 50% missingness. Table 3.1 also lists which value of λ "matches" the missingness mechanism (i.e., MAR corresponds to $\lambda = 0$). For each replication, complete data $\{(y_i, z_i), i = 1, \ldots, n\}$ were first generated with a selected n and ρ and then missing values for y_i were induced according to one of the four mechanisms.

3.6.2 Imputation Methods

Our goal was to estimate the mean of Y. We applied both the PPM hot deck and the SB hot deck to estimate the mean, and also compared these estimates to the complete case mean. For the PPM hot deck we performed separate analyses using $\lambda = \{0, 1, \infty\}$ and for SB hot deck we used the suggested range of c values ($c = \{0, 1, 2, 3\}$) in separate analyses. We excluded the suggested value of $c = -1$

since missingness was induced such that larger values of Y were more likely to be missing; we note that the need to specify positive or negative c values is a disadvantage of the SB hot deck and discuss this further in the application in Section 3.7. For each imputation method we created 10 imputed data sets. A total of 500 replications were used for each combination of n, ρ, and missingness mechanism. To evaluate the performance of each method, we calculated the average empirical bias in the mean estimates for each method. We also calculated the actual coverage of a nominal 95% confidence interval.

In addition, the donor quality metrics described in Section 3.5 were calculated for each pre-bootstrap sample and are reported as the average MMD and average MVSP across replicates for each combination of n, ρ, missingness mechanism, and value of λ. This allows us to illustrate how these metrics could help identify scenarios in which the PPM hot deck is unable to find quality donors.

3.6.3 Results: Bias

Figure 3.2 shows the estimates of the mean of Y for the complete case analysis and each imputation method averaged across 500 replications for $n = 400$ and 25% missingness. For each population design there are three estimates for the PPM hot deck, corresponding to the three choices of λ, and four estimates for the SB hot deck, corresponding to the four choices of c.

In general the mean estimates increase for increasing λ values and for increasing c values. The spread of these estimates also tends to increase as the correlation between Y and Z decreases. The exception is for the PPM hot deck with the weakest correlation ($\rho = 0.2$). Here the estimate for $\lambda = \infty$ shifts down to be close to the value for $\lambda = 0$, instead of being much larger as it is for the stronger correlations. This is a consequence of using the distance measure in (2.14) when no close matches are available and is further discussed in Section 3.6.4. The spread of estimates for the SB hot deck is much larger than for the PPM hot deck for most scenarios. Mean estimates for the PPM hot deck with $\lambda = 0$ and the SB hot deck with $c = 0$ are virtually identical, which is to be expected since they both correspond to an MAR assumption. However, there do not appear to be other equivalences for other λ and c values.

When missingness is dependent on Z, the data are MAR. In the first panel of Figure 3.2 we see that the PPM hot deck with $\lambda = 0$ and the SB hot deck with $c = 0$ produce nearly unbiased estimates for all values of ρ. This is expected, since these correspond to an MAR mechanism. Estimates using other values of λ and c are biased, as would be expected.

When missingness is dependent on Y or Y^2, the PPM hot deck with $\lambda = \infty$ is expected to perform the best, since it assumes MNAR. For these cases it is not clear which of the SB hot deck estimates should be the most unbiased, since all c values aside from $c = 0$ model MNAR mechanisms. In the second panel of Figure 3.2 we see that when missingness depends on Y and the correlation is strong ($\rho = 0.8$), the mean estimate for the PPM hot deck with $\lambda = \infty$ is close to unbiased, but as the correlation weakens the bias becomes larger, with substantial bias for $\rho = 0.2$.

A similar pattern emerges in the third panel, where missingness depends on Y^2, with all-around worse performance. There is substantial bias for the PPM hot deck with $\lambda = \infty$ for all correlations, with bias increasing with decreasing correlation. In these simulations, for the missingness dependent on Y, $c = 2$ returns the closest to unbiased estimates. However, similar to the PPM hot deck, when missingness depends on Y^2 the SB hot deck does not produce unbiased estimates for any c value.

When missingness is dependent on $Z + Y$, we expect the PPM hot deck with $\lambda = 1$ to produce the best estimates. However, the method does not perform exactly as expected as shown in the last panel of Figure 3.2. For the weaker correlations ($\rho = 0.5, 0.2$) the estimate using $\lambda = 1$ is the least biased, but it is biased. For the strongest correlation of $\rho = 0.8$ the least biased method is $\lambda = \infty$, though $\lambda = 1$ is close. The SB hot deck gets closer with values of $c = 3$ and $c = 2$, but there is no reason why this should be the case and this result is likely specific to this particular simulation design.

For weaker proxy values and strong MNAR mechanisms, bias is high for the PPM hot deck, even when the λ value corresponds to the missingness mechanism. This is partially due to a limitation of hot deck imputation and not unique to the proposed method. Hot deck methods can only impute observed values. In our simulation, when the missing data mechanism is strongly dependent on Y, large values of Y are almost all missing. Thus the only Y values available for imputation are the values observed among respondents, which are smaller than the missing values. Fortunately, the donor quality metrics we propose can help identify these situations, as well as explain the odd results with the smallest correlation ($\rho = 0.2$), where the PPM hot deck with $\lambda = \infty$ produces mean estimates that are much smaller than would be expected, even smaller than the MAR estimates when missingness depends on $[Z + Y]$.

3.6.4 Results: Donor Quality Metrics

Table 3.2 displays the donor quality metrics calculated for the PPM hot deck on the pre-bootstrap sample and averaged across replications. These metrics help flag scenarios in which the PPM hot deck performed poorly because it could not find adequately close donors. They also explain the strange performance of the PPM hot deck with low correlations, where the estimates cluster together for all values of λ.

The MMD measures the average distance from donor to donee, across all respondents and nonrespondents. Since the distance is divided by the standard deviation of the outcome among respondents, the MMD can be interpreted as the average number of standard deviations away a donor is from a donee. Table 3.2 shows that for the MAR mechanism $[Z]$, imputations with $\lambda = 0$ will result in close matches – approximately 0.01 standard deviations away or closer on average. However, for MNAR mechanisms the MMD is larger. Looking at the "matching" λ values for each mechanism ($\lambda = \infty$ for $[Y]$ and $[Y^2]$; $\lambda = 1$ for $[Z + Y]$), as the correlation goes down, the MMD goes up. For the $[Y]$ mechanism, the MMD is reasonably small for the higher ρ values, but quite large for $\rho = 0.2$. This corresponds to the poor performance of the PPM hot deck with the low correlation in this case; an MMD of 0.510 means that on average donors were a half a standard deviation away from donees – clearly not good matches. A similar pattern is observed with the $[Z + Y]$ mechanism; larger ρ values have relatively small MMD values, but the MMD for $\rho = 0.2$ is much larger. For the $[Y^2]$ mechanism the MMD results are even more concerning – not even for the high correlation is the distance acceptable. This explains why the PPM hot deck is biased even for the “matching” λ value; respondent predicted means and nonrespondent predicted means are far apart on average, and thus there are no adequately close donors.

Similar conclusions can be drawn with the MVSP, which measures the variance in the donor selection probabilities. A large MVSP indicates good variability of the selection probabilities, meaning that some donors are closer than others. On the other hand, a small MVSP indicates that all donors are equally far away, likely because there is little overlap in the respondent and nonrespondent predicted means. Looking at Table 3.2 we can see that for the MAR mechanism $[Z]$, the MVSP is high for $\lambda = 0$ for all correlations, but for the MNAR mechanisms the MVSP decreases for smaller values of ρ even for the “matching” λ value.

The MVSP can also help explain why the PPM hot deck with $\lambda = \infty$ produces unexpected results with weak correlations. In Table 3.2, if we look at the **set** of MVSP values for each scenario, we can get an idea of how well matches are being found for each λ value. For example, for the case with missingness dependent on Z and a correlation of $\rho = 0.8$, the MVSP values are all approximately 1.7 for the three λ values, showing reasonable spread of the selection probabilities. The corresponding estimates shown in Figure 3.2 reflect this; the estimates are well spread out. However, when the correlation is $\rho = 0.2$, the MVSP is drastically lower for $\lambda = \infty$ than for the other two values: 0.07 compared to 1.74 and 1.09. This very small MVSP means that all the selection probabilities for $\lambda = \infty$ are approximately the same for all potential donors. Thus the imputations are essentially a simple random sample from the respondents, and the mean estimate gets pulled towards the complete case estimate as seen in Figure 3.2. Looking at the formula (3.5) for the predicted means, we see that the correlation estimate is in the denominator when $\lambda = \infty$, and thus small correlations will lead to larger shifts in the predicted means for nonrespondents away from the predicted means for respondents. However, the hot deck procedure simply cannot find good donors for these really large predicted means – all donors are really far away. No respondent is close, so the donor selection effectively turns into a simple random sample from the respondents.

3.6.5 Results: Coverage

Figure 3.3 shows coverage for the scenarios in the simulation experiment. For the highest correlation ($\rho = 0.8$), the PPM hot deck performs well and as expected for missingness mechanisms $[Z]$, $[Y]$, and $[Z+Y]$, with $\lambda = 0$, $\lambda = \infty$, and $\lambda = 1$ having approximately 95% coverage, respectively. For $\rho = 0.5$, the same values of λ are the closest to 95% coverage for those three mechanisms, with $[Z+Y]$ falling to about 90%. When $\rho = 0.2$, the PPM hot deck only achieves nominal coverage when the data are MAR, with $\lambda = 0$.

The PPM hot deck has a difficult time achieving nominal coverage for $[Y^2]$ for any value of λ, for all correlations, with $\lambda = \infty$ ranging from about 62% coverage when $\rho = 0.8$, to less than 10% coverage when $\rho = 0.2$. This is not unexpected, since we saw high bias for this method in these situations. The SB hot deck does have higher coverage for these scenarios using a value of $c = 3$, but still is as low as 40% coverage for $\rho = 0.2$, with the other values of c obtaining similar results as the PPM hot deck.

Overall, coverage was best for the PPM hot deck when correlation was high. For the most part, the method followed the pattern we expected, with highest coverage for the λ value corresponding to the missing data mechanism. The SB hot deck had no clear pattern linking the highest coverage to specific combinations of c-values and missingness mechanisms, similar to what we saw with the bias results.

3.6.6 Effect of Sample Size and Percent Missingness

We also examined the effect that sample size and percent missingness had on the estimates and coverage of both the PPM hot deck and the SB hot deck (Figures 3.4 and 3.5). We used combinations of $n = \{400, 800\}$ and 25% and 50% missingness. As expected, the PPM hot deck estimates are less biased with a larger sample size and lower percent missingness for a given mechanism. The difference in percent missing seems to play more of a role in bias than the sample size. In terms of coverage, for each mechanism, estimates from 25% missingness had closer to nominal coverage than estimates from 50% missingness, though it was not always the case that $n = 800$ was the best. The scenarios with $n = 400$ and 25% missingness had closer to nominal coverage for certain mechanism and correlation combinations than $n = 800$ with 25% missingness.

Siddique and Belin's hot deck with $c = 0$ performs the same as PPM hot deck with $\lambda = 0$ in all settings (the two MAR models). The values of c that were unbiased (or the least biased) for the mechanisms discussed above with $n = 400$ and 25% missingness, are still unbiased (or the least biased) for the other sample sizes and 50% missingness, with those values not varying much. For the remaining c values in each setting, the estimates do change as sample size and missingness change, with the pattern of higher estimates with 25% missingness, lower estimates with 50% missingness. The most notable difference between estimates of the PPM hot deck and the SB hot deck is that the range of the estimates for the SB hot deck becomes very wide over the values of c for 50% missingness. For example, with $\rho = 0.5$, mechanism $[Y]$, $n = 400$ with 50% missingness, the estimates range from approximately 0.81 when $c = 0$ to approximately 1.07 when $c = 3$. This compares to the range of the estimates for the PPM hot deck from approximately 0.81 when $\lambda = 0$ to approximately 0.94 when $\lambda = \infty$. Examining coverage for the SB hot deck, one observation is that for the settings in which a specific c value was unbiased, the same c suffers from undercoverage. Take $\rho = 0.5$ and mechanism $[Y]$; for all n with either 25% or 50% missingness, $c = 2$ was unbiased; and specifically was less biased than PPM hot deck with $\lambda = \infty$. However, the coverage obtained with $c = 2$ is no better than the coverage obtained with $\lambda = \infty$. This pattern holds for all cases when the SB hot deck was less biased for a value of c than the "best" λ for the PPM hot deck. This may be due to not having enough variability between MI sets to make the SB hot deck a fully proper MI procedure.

3.7 Application to OMAS 2012

The 2012 Ohio Medicaid Assessment Survey (OMAS) was a stratified simple random sample of telephone numbers of non-institutionalized adult and child populations living in residential households in Ohio [36]. There was over-sampling in the counties having the highest density of African-Americans. While there were many variables with missing data in the OMAS data, for illustrating the proxy pattern-mixture hot deck we focus on imputation of annual family income, which had the highest rate of missingness.

The imputation of income as performed by OMAS assumed that the data are MAR, using a combination of regression imputation and draws from a lognormal distribution. The goal of our application of the proxy pattern-mixture hot deck is to examine the effect that nonignorable missingness may have on estimates of the mean income. In OMAS, a subset of respondents who did not provide an exact income value did provide a range for their income; in our application we treat these subjects as nonrespondents since they did not provide an exact income value. Of the 22,929 participants in the 2012 OMAS, 7,347 did not provide their exact income, a 32% nonresponse rate. The distribution of observed income values was highly skewed with some extremely large values of income; the largest three were \$2,160,000, \$2,460,000 and \$11,999,964.

To apply the PPM hot deck, we log-transformed the income values and used linear regression to create the proxy. Fully observed (or already imputed by OMAS) covariates included region, number of adults in the household, number of children in the household, age, gender, race, education, whether or not the respondent owned his/her home, and type of insurance (none, Medicaid, other insurance). We also included the final adjusted survey weight as a covariate as recommended for survey data [37]. Starting with a model that included all pairwise interactions, backwards selection with a p-value threshold of 0.05 was used to select a final model. The final model had an R-squared of 28%, and the resulting correlation between the outcome and proxy among respondents was $\hat{\rho}^{(0)} = 0.53$, which was a moderately strong proxy. We applied the PPM hot deck with $\lambda = 0, 1, \infty$ and also calculated both donor quality metrics (MMD, MVSP) for each λ value using the pre-bootstrap sample. For comparison, we also applied the Siddique and Belin hot deck. Unlike the simulation study, we did not know a priori whether larger or smaller values of income were more likely to be missing, and thus applied their recommended range of c values, $c = \{-1, 0, 1, 2, 3\}$.

The reason for log-transforming income was to obtain the "best fitting" model for income and thus the strongest proxy. However, once donors were found, the un-transformed income values were imputed via the hot deck. We also applied the PPM hot deck using untransformed income to create the proxy and find donors; the strength of the proxy was lower ($\hat{\rho}^{(0)} = 0.30$) and as a consequence confidence intervals were wider, but mean estimates were approximately the same for all λ values (results not shown). The SB hot deck could not be applied to the untransformed income, because the highly skewed nature of income caused problems when creating the nonignorable ABB. When values of income were selected into the ABB with probability proportional to income (or positive powers of income), one very large income had a very large probability of inclusion into the bootstrap sample (> 0.5) and thus the ABB consisted of this one respondent being the only one selected into the ABB.

Figure 3.6 shows estimated means and 95% confidence intervals for income using complete cases only and after application of the PPM hot deck with $\lambda = \{0, 1, \infty\}$ and the SB hot deck with $c = \{-1, 0, 1, 2, 3\}$, using $D = 20$ multiply imputed data sets. These estimates use the final OMAS survey weights to compute weighted means; this is an advantage of the multiple imputation approach. We can see that the estimate of mean income is highest for the complete case analysis, and decreases as the value of λ increases for the PPM hot deck. As we change the missingness assumption from MCAR (the complete case estimate), to MAR ($\lambda = 0$), to MNAR ($\lambda = 1, \infty$), the estimated means decrease, and the width of the confidence intervals increases. The decreasing trend suggests that lower values of income are more likely to be missing than larger values. This is reflected in the mean proxy values; the average proxy value was 10.32 for respondents and 10.17 for nonrespondents. Neither donor quality metric indicated any problem with the PPM hot deck finding quality donors. The MMD values were approximately 0.0004 for all λ values, indicating that close matches were available on average. The MVSP values (x1000) were consistent across λ values (≈ 0.033), indicating that $\lambda = \infty$ did not result in a simple random sample as was seen in some scenarios in the simulation study.

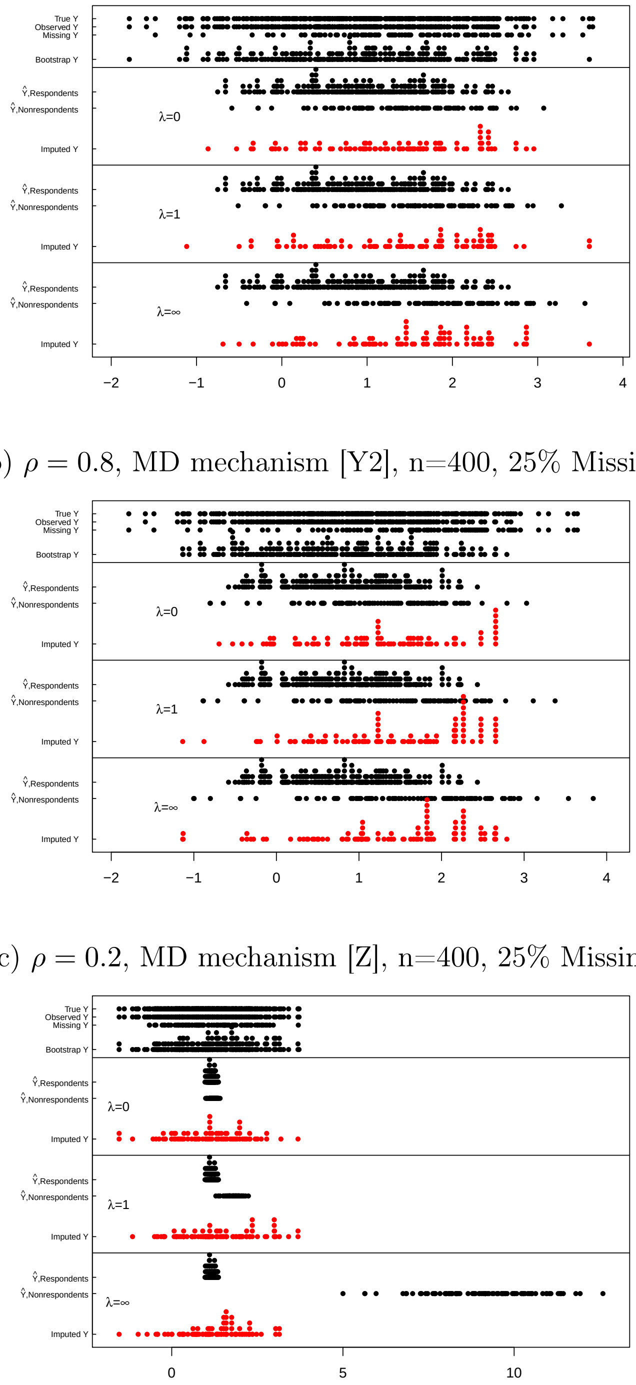
Estimates for the PPM hot deck with $\lambda = 0$ and the SB hot deck with $c = 0$ are similar, as expected, since both assume an MAR mechanism. However, for the SB hot deck, as c increases, the mean estimates increase. This is expected, since positive c values mean that larger values of the outcome will get selected into the donor pool with higher probability. However, as evidenced by the proxy means, for OMAS it appears that smaller values of income were more likely to be missing. This illustrates a disadvantage of the SB hot deck compared to the PPM hot deck. The SB hot deck requires the user to specify the direction of the nonignorable missingness, i.e., whether large or small values of the outcome are more likely to be missing. With the PPM hot deck, the method adapts to the data and will impute larger or smaller values of the outcome depending on the direction of the difference between respondents and nonrespondents as seen in the proxy.

The imputation performed by OMAS assumed that income values were missing at random. This analysis shows us that if this assumption is incorrect, if the missingness in income depended on the missing income values themselves, then OMAS estimates of mean income would tend to be too large. The resulting confidence intervals for

MAR ($\lambda = 0$) and extreme MNAR ($\lambda = \infty$) do overlap, but the difference in mean is around \$3,000, a not insubstantial difference.

3.8 Tables and Figures

Figure 3.1: Distributions for the PPM hot deck for a single imputed data set.



(a) $\rho = 0.8$, MD mechanism [Z], n=400, 25% Missing

(b) $\rho = 0.8$, MD mechanism [Y2], n=400, 25% Missing

(c) $\rho = 0.2$, MD mechanism [Z], n=400, 25% Missing

Table 3.1: Parameters for the missing data mechanisms in the simulation study. The variable(s) in brackets denotes the variable(s) on which missingness depends.

Missingness Mechanism	Missingness Model	Expected closest λ	Percent Missingness 25%	50%
MAR	$[Z]$	$\lambda = 0$	$-1.6 + 0.5Z$	$-0.5 + 0.5Z$
MNAR	$[Y]$	$\lambda = \infty$	$-1.6 + 0.5Y$	$-0.5 + 0.5Y$
MNAR	$[Y^2]$	$\lambda = \infty$	$-2.1 + 0.5Y^2$	$-1.0 + 0.5Y^2$
MNAR	$[Z + Y]$	$\lambda = 1$	$-2.1 + 0.5Z + 0.5Y$	$-1.0 + 0.5Z + 0.5Y$

Figure 3.2: Estimates of the mean of Y over 500 replications, for complete case analysis (CC), the proxy pattern-mixture hot deck (PHD) and the Siddique and Belin nonignorable hot deck (SB). The estimates for the three λ values are in solid symbols, and the four c-values are open symbols. The three correlations ($\rho = 0.8, 0.5, 0.2$) are grouped by the four missingness mechanisms along the x-axis. The true mean, $\mu_y = 1$, is marked by the solid horizontal line.

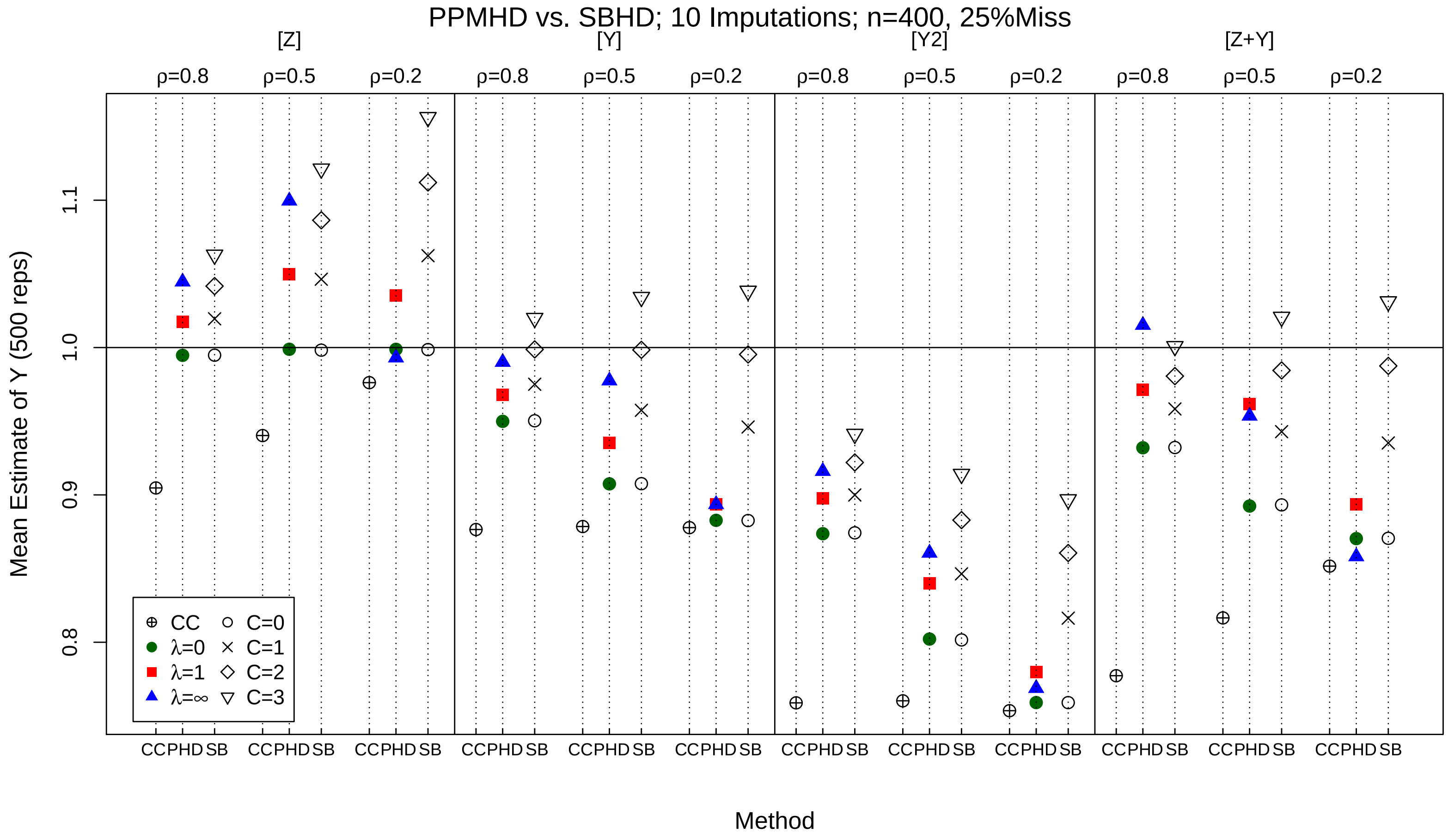


Table 3.2: Mean Minimum Distance (MMD) and Mean Variance Selection Probability (MVSP)x1000 for the simulation study, calculated on the pre-bootstrap sample of respondents. Cells shaded gray indicate where the λ value "matches" the missing data mechanism.

		Missing Data Mechanism							
		$[Z]$		$[Y]$		$[Y^2]$		$[Z+Y]$	
ρ	λ	MMD	MVSP	MMD	MVSP	MMD	MVSP	MMD	MVSP
0.8	0	0.012	1.74	0.011	1.74	0.027	1.72	0.018	1.78
	1	0.016	1.73	0.013	1.74	0.054	1.65	0.027	1.74
	∞	0.021	1.71	0.016	1.74	0.119	1.51	0.047	1.69
0.5	0	0.008	1.74	0.006	1.74	0.008	1.76	0.009	1.78
	1	0.021	1.68	0.010	1.73	0.034	1.65	0.038	1.64
	∞	0.130	1.37	0.037	1.65	0.277	1.10	0.386	0.77
0.2	0	0.003	1.74	0.002	1.74	0.002	1.78	0.003	1.78
	1	0.097	1.09	0.013	1.65	0.031	1.52	0.184	0.64
	∞	2.323	0.07	0.510	0.82	1.325	0.43	4.026	0.02

Figure 3.3: Nominal coverage for complete case analysis (CC), the proxy pattern-mixture hot deck (PHD) and the Siddique and Belin nonignorable hot deck (SB) over 500 replicates. The estimates for the three λ values are in solid symbols, and the four c-values are open symbols. The three correlations ($\rho = 0.8, 0.5, 0.2$) are grouped by the four missingness mechanisms along the x-axis. Nominal 95% coverage is marked by the solid horizontal line.

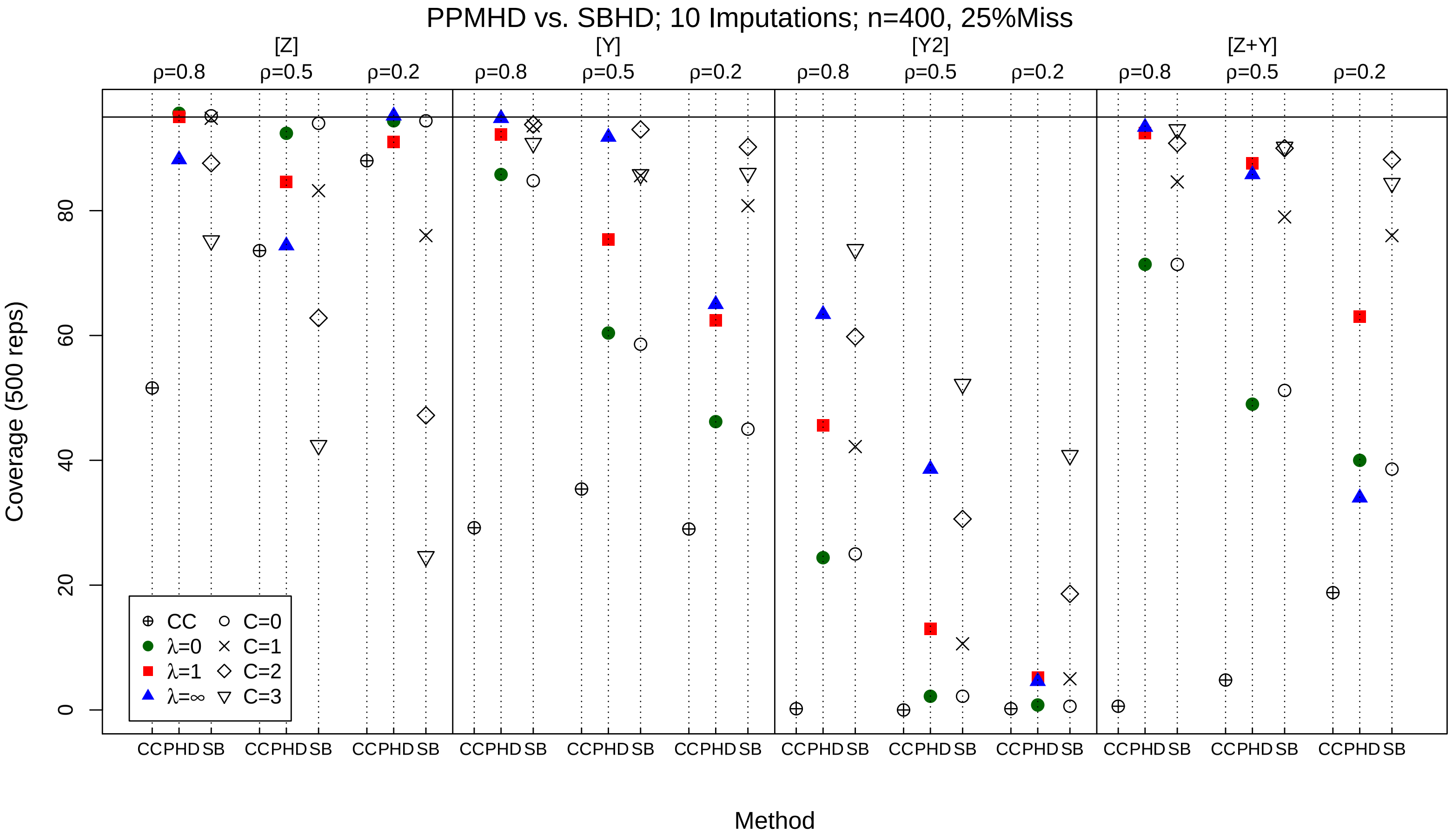
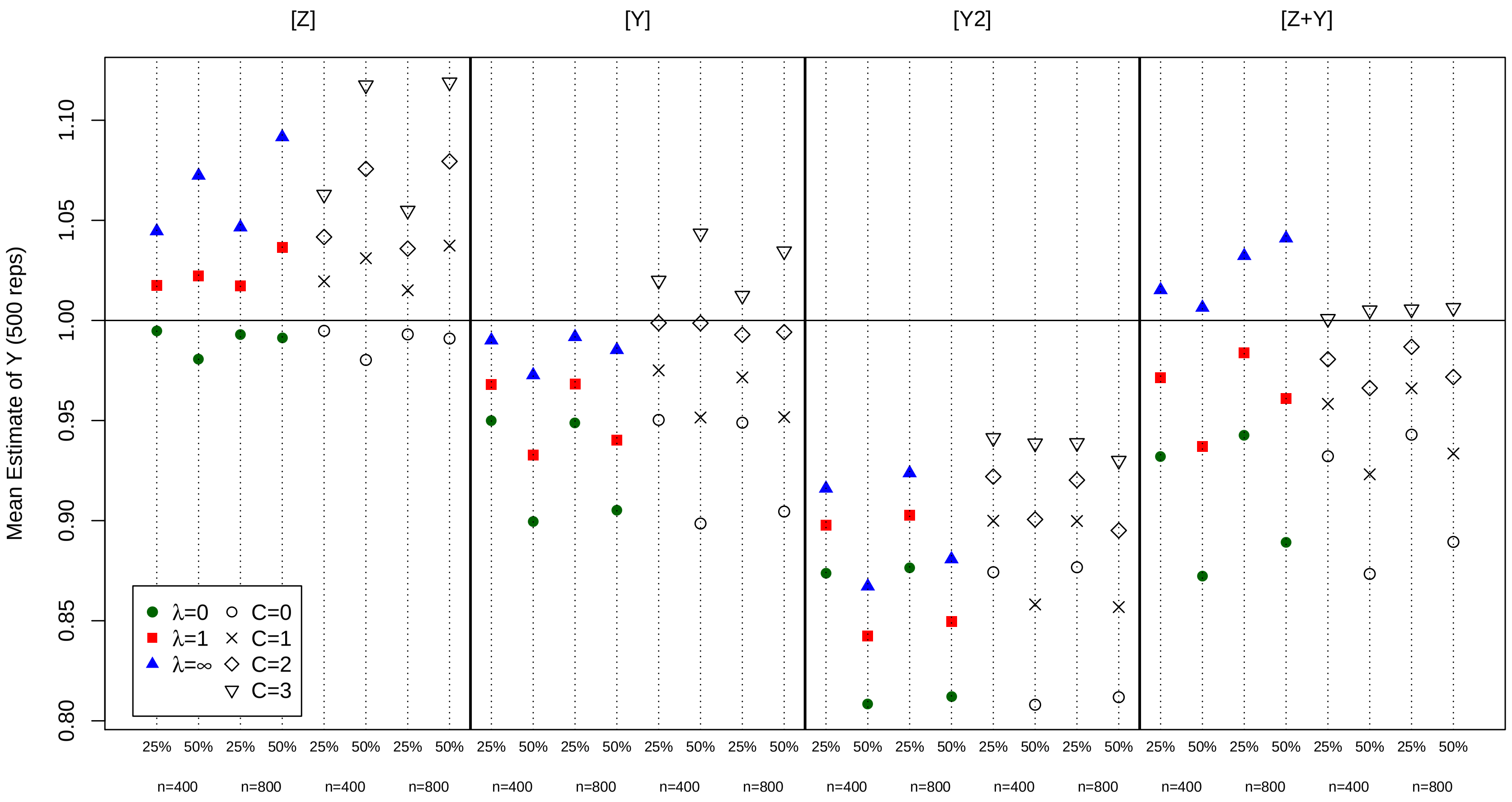


Figure 3.4: Estimates of the mean of Y over 500 replications, for the proxy pattern-mixture hot deck and the Siddique and Belin nonignorable hot deck for combinations of $n = \{400, 800\}$ and 25% or 50% missingness. The estimates for the three λ values are in solid symbols and the four c-values are open symbols. The true mean, $\mu_y = 1$, is marked by the solid horizontal line.

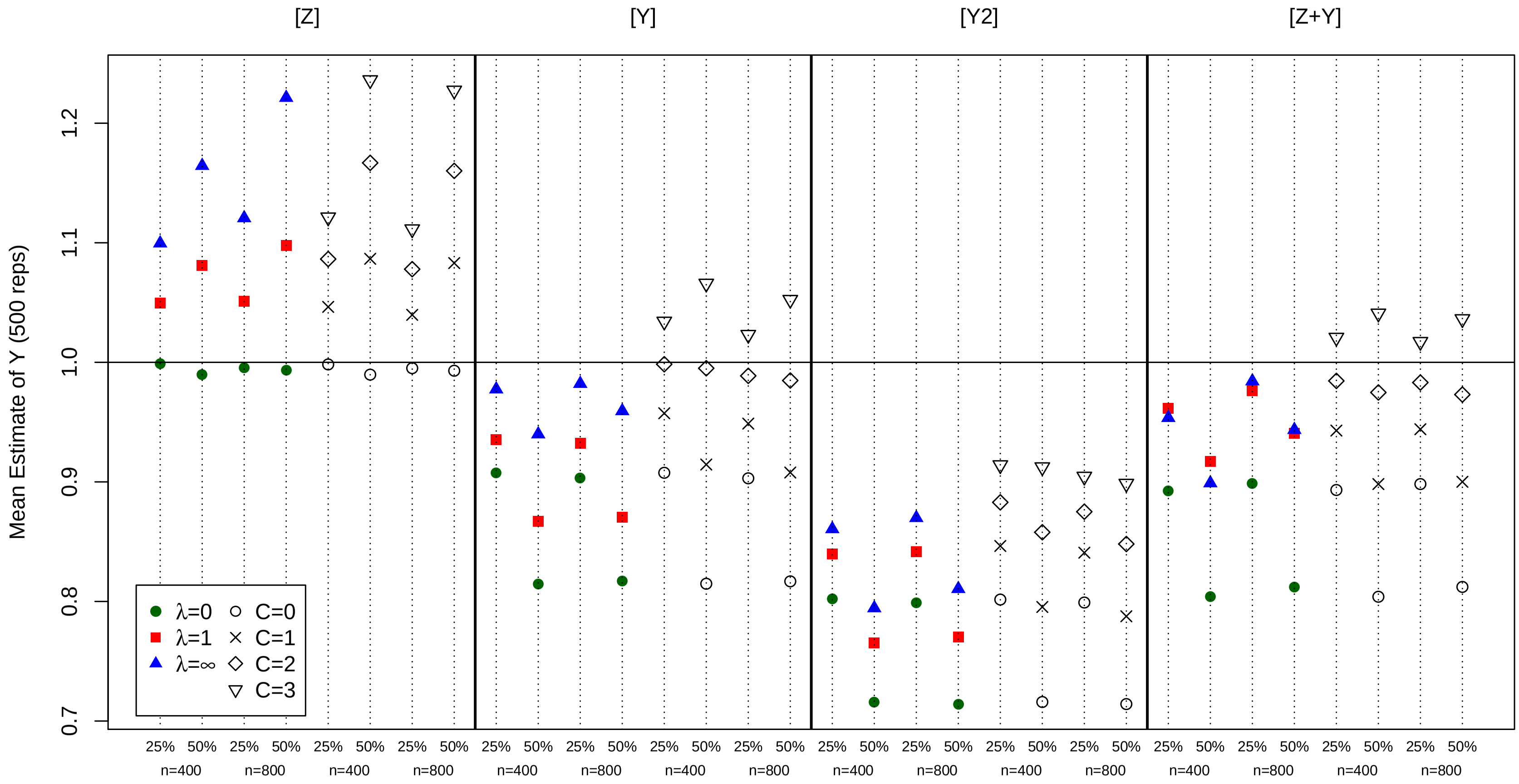


(a) $\rho = 0.8$

(Continued)

Figure 3.4: Continued.

(b) $\rho = 0.5$



(Continued)

Figure 3.4: Continued.

(c) $\rho = 0.2$

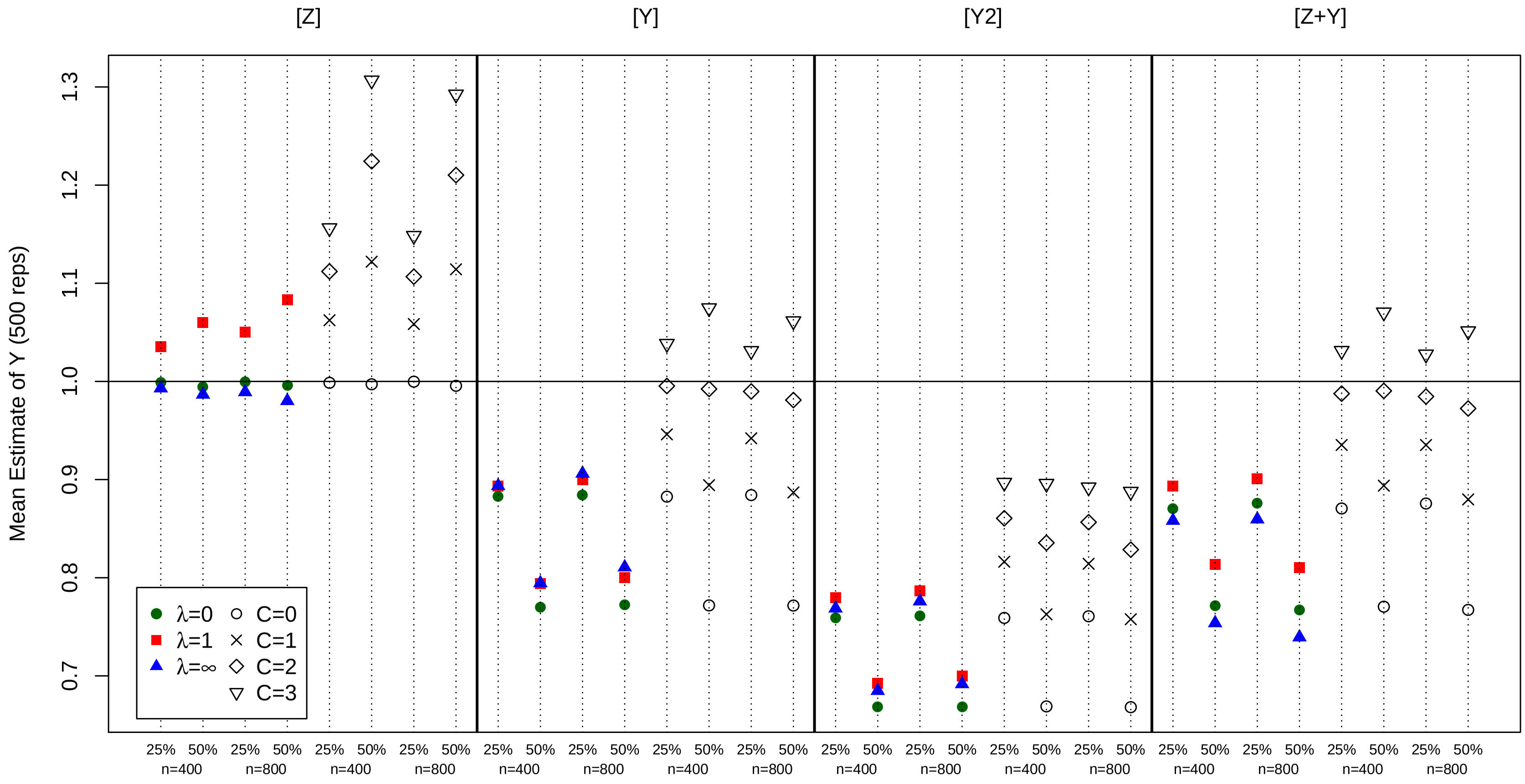
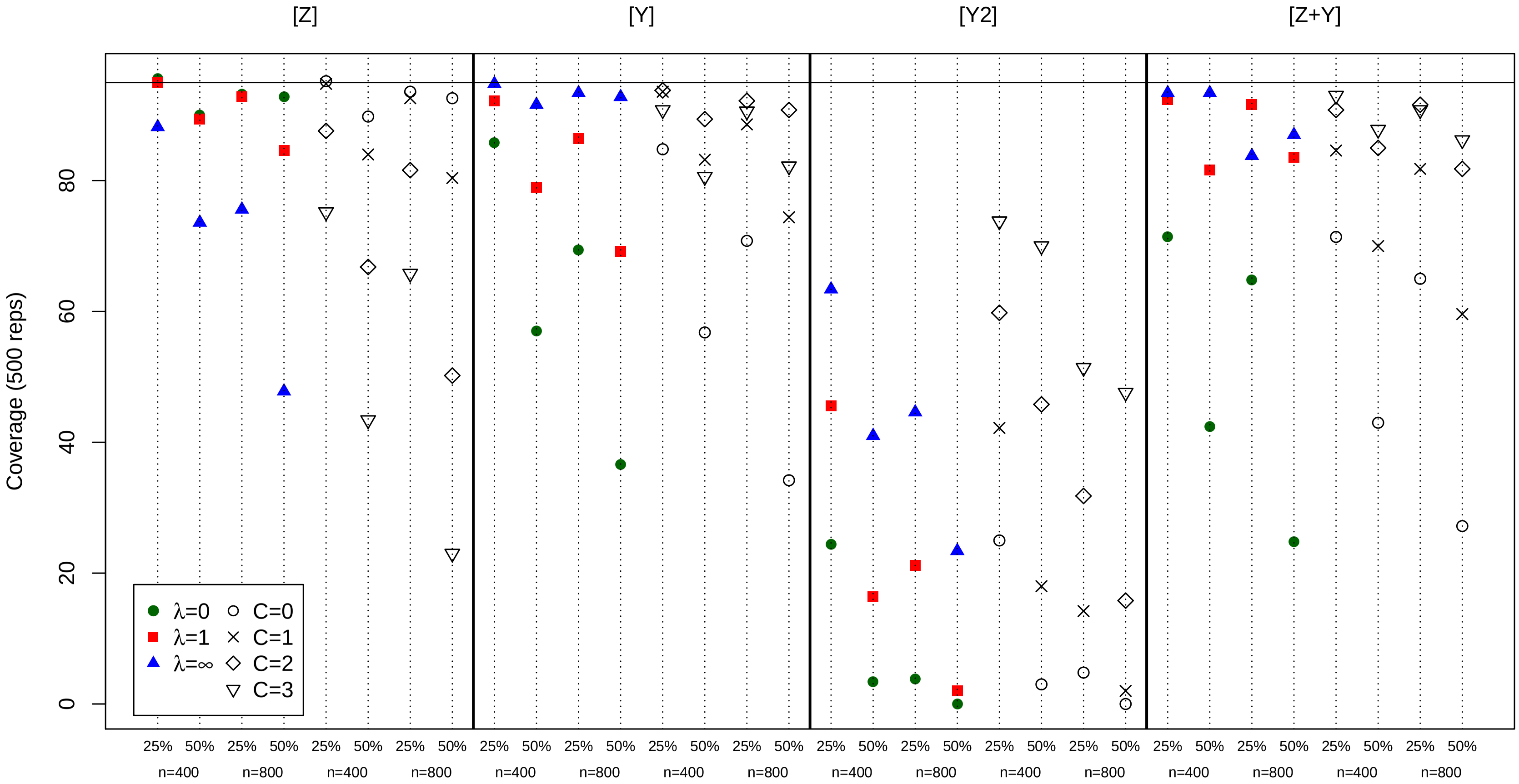


Figure 3.5: Nominal coverage over 500 replications, for the proxy pattern-mixture hot deck and the Siddique and Belin nonignorable hot deck for combinations of $n = \{400, 800\}$ and 25% or 50% missingness. The estimates for the three λ values are in solid symbols and the four c-values are open symbols. Nominal 95% coverage is marked by the solid horizontal line.

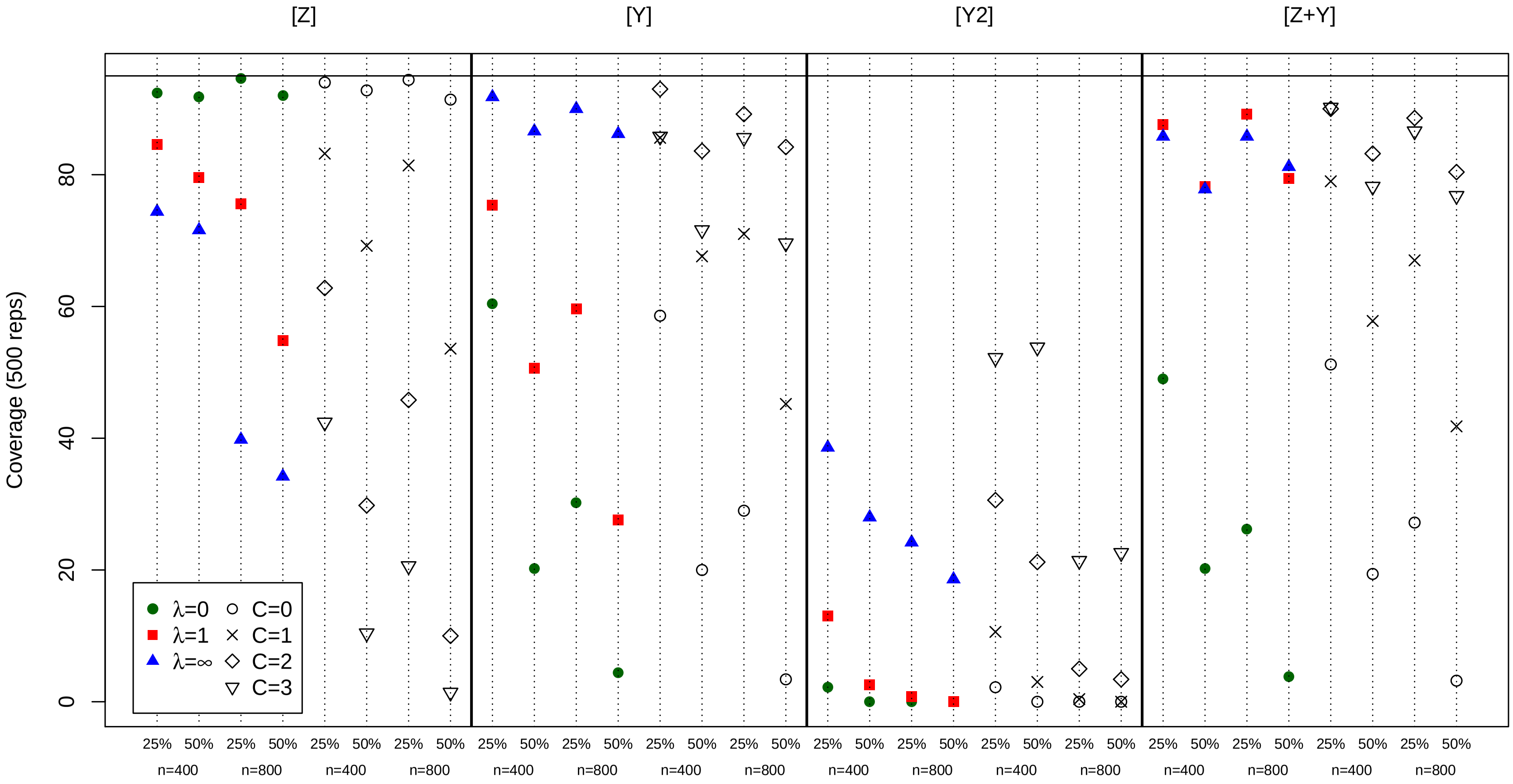
(a) $\rho = 0.8$



(Continued)

Figure 3.5: Continued.

(b) $\rho = 0.5$



(Continued)

Figure 3.5: Continued.

(c) $\rho = 0.2$

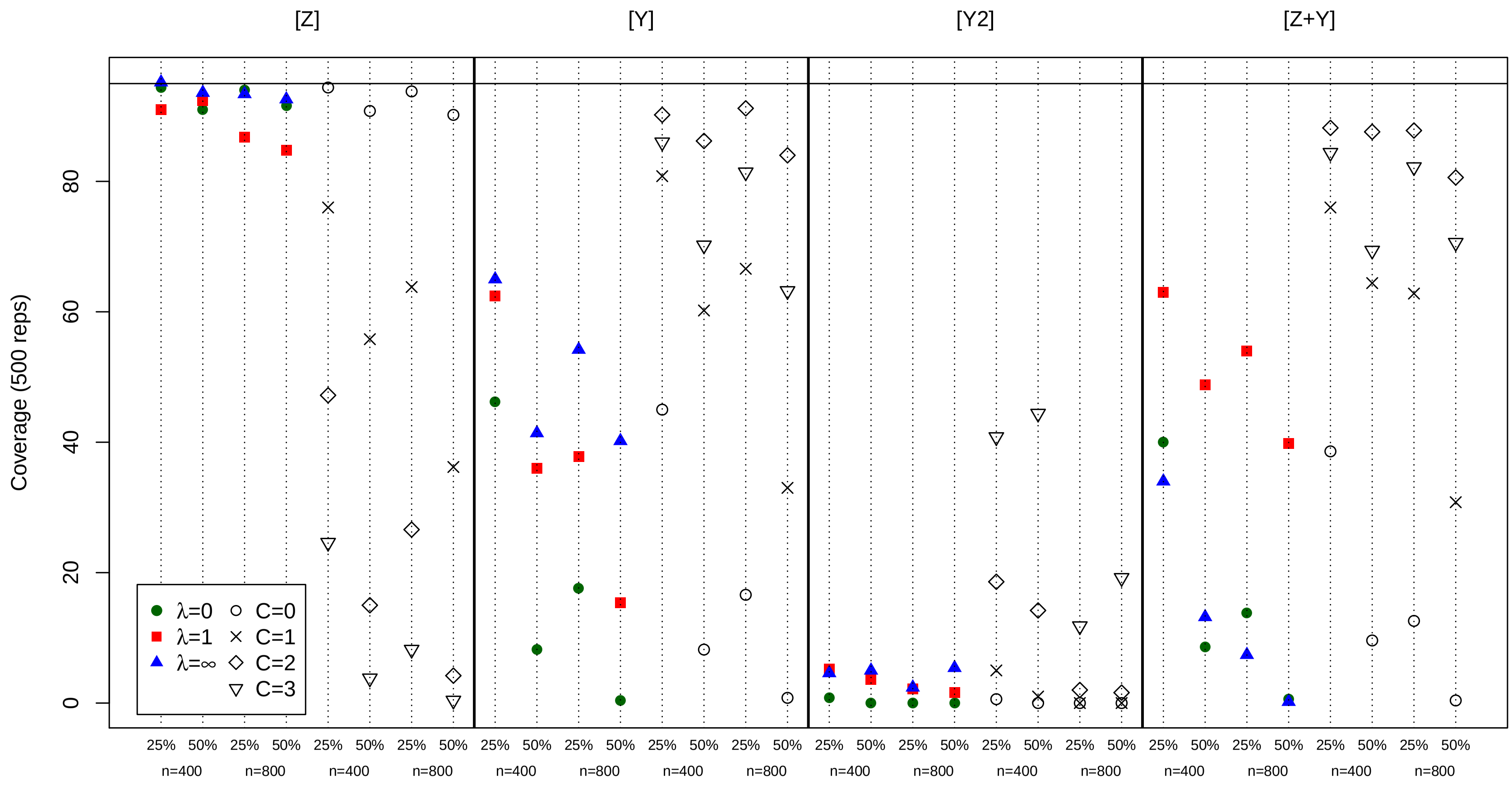
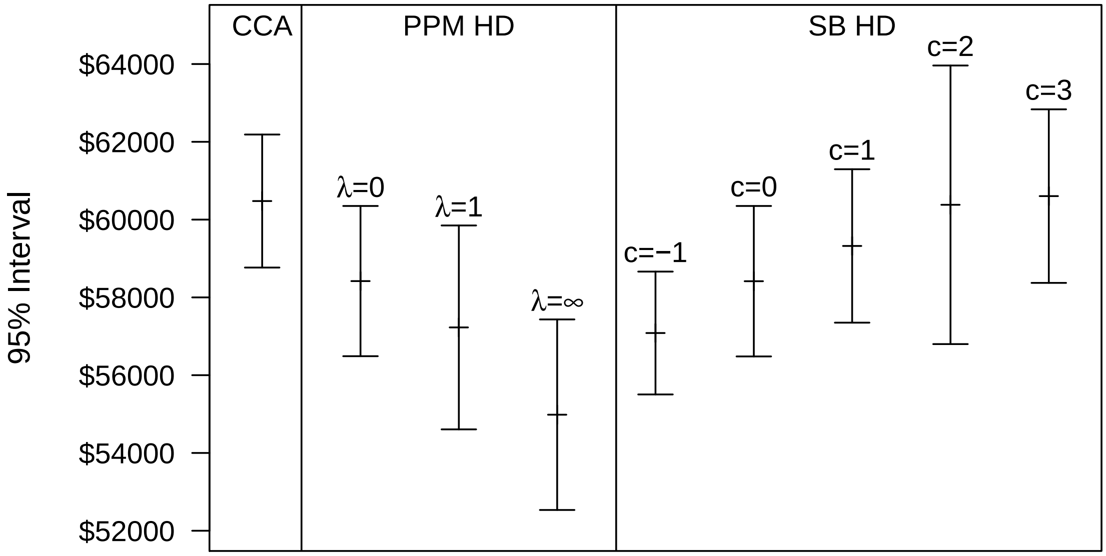


Figure 3.6: Application of the proxy pattern-mixture hot deck (PPM HD) and the Siddique and Belin hot deck (SB HD) to the 2012 OMAS to estimate mean income. The complete case mean (CCA) is also shown.



Chapter 4: Binary Outcome

We now extend the proposed proxy pattern-mixture hot deck to the imputation of a binary outcome. This should be attractive since the proposed method should not be as sensitive to departures from normality as a parametric imputation technique. As discussed in Andridge (2009) [38], when imputing a continuous variable, the pattern-mixture model is fairly robust to departures from normality. However, when imputing a binary variable, violating normality may have more severe consequences, making the hot deck method more appealing. To implement the proxy pattern-mixture hot deck for binary imputation, we follow similar steps as with a continuous outcome, but instead use the methods from [38] and modify donor selection. We note that the nonignorable hot deck of Siddique and Belin does not have a natural extension to categorical outcomes. Siddique and Harel (2009) [39] imply that they use the same predictive mean matching, using the predictive values from linear regression on a binary outcome.

4.1 The Model

We assume we have fully observed covariates Z and a partially observed binary outcome Y with missingness indicator M. Following convention, we assume there exists a latent variable U (unobserved) such that $\Pr(Y_i = 1|m_i) = \Pr(U_i > 0|m_i)$.

The covariates Z are used to form a proxy for U (instead of the outcome Y). Specifically, we take the proxy to be $X = \hat{\alpha}_0 + \hat{\boldsymbol{\alpha}}\mathbf{Z}$, where $(\hat{\alpha}_0, \hat{\boldsymbol{\alpha}})$ are estimated from $\Pr(Y = 1|\mathbf{Z}, M = 0) = \Phi(\alpha_0 + \alpha\mathbf{Z})$, the probit regression of Y on $\mathbf{Z}$ for respondents. Since $\mathbf{Z}$ is fully observed, we have X for both respondents and nonrespondents, and the pattern-mixture model is assumed for X and U. We can then apply methods from the continuous outcome case to estimate the parameters of the model under different assumptions on the missing data mechanism. However, since the latent variable U is unobserved for all subjects, we have additional requirements for estimating the parameters for respondents. We discuss two methods of estimation, Tate's full MLE method [40, 41] and a two-step method, proposed by Olsson (1982) [42]. Both have in common using a cutpoint of the distribution of U to reduce the problem to estimation of four parameters from five. We discuss the general cutpoint first, then detail the two estimation methods. We then discuss estimation for the non-respondents.

After creation of the proxy, X, and using the assumption for the latent U, we can write the conditional distribution of Y given X as

$$\begin{aligned}(Y|X = x, M = 0) &\sim \text{Binomial}(1, P_x) \qquad (4.1)\\ P_x &= P(U > 0|x, M = 0).\end{aligned}$$

The pattern-mixture model is assumed for the proxy X and the latent variable, U:

$$(X, U|M = m) \;\sim\; N\left(\begin{pmatrix}\mu_x^{(m)}\\ \mu_u^{(m)}\end{pmatrix}, \begin{pmatrix}\sigma_{xx}^{(m)} & \rho^{(m)}\sqrt{\sigma_{xx}^{(m)}\sigma_{uu}^{(m)}}\\ \rho^{(m)}\sqrt{\sigma_{xx}^{(m)}\sigma_{uu}^{(m)}} & \sigma_{uu}^{(m)}\end{pmatrix}\right) \qquad (4.2)$$

where $\rho^{(m)} = \sigma_{xu}^{(m)}/\sqrt{\sigma_{xx}^{(m)}\sigma_{uu}^{(m)}}$. This model has 11 parameters: $\{\mu_x^{(m)}, \sigma_{xx}^{(m)}, \mu_u^{(m)}, \sigma_{uu}^{(m)}, \rho^{(m)}; m = 0, 1\}$ and $\pi_1 = Pr(M = 1)$. For both respondents and nonrespondents, the joint likelihood of the sample can be factored into two pieces: the marginal distribution of X and the conditional distribution of Y given X. We begin with the respondents. Equation (4.2) implies that the marginal distribution of X given $M = 0$ is $N(\mu_x^{(0)}, \sigma_{xx}^{(0)})$. These estimates are simply $\hat{\mu}_x^{(0)} = \bar{x}_R$ and $\hat{\sigma}_{xx}^{(0)} = s_{xx}^{(0)}$. However, since U is unobserved, $\mu_u^{(0)}, \sigma_{uu}^{(0)}$ and $\rho^{(0)}$ are unidentified. Without loss of generality, we let $\sigma_{uu}^{(0)} = \frac{1}{1-\rho^{(0)2}}$ which implies $\text{Var}(U|X, M = 0) = 1$.

To formulate the cutpoint of the distribution of U, we use properties of bivariate normality to write the marginal distribution of U for the respondents as:

$$(U|m = 0) \sim N(\mu_u^{(0)}, \sigma_{uu}^{(0)}) \tag{4.3}$$

which gives,

$$\begin{aligned} \Pr(Y = 1|M = 0) = \text{P}(U > 0|M = 0) &= \Phi\left(\frac{\mu_u^{(0)}}{\sqrt{\sigma_{uu}^{(0)}}}\right) \\ &= \Phi(\omega^{(0)}). \end{aligned} \tag{4.4}$$

Here, we are defining the cutpoint of U, $\omega^{(0)} = \mu_u^{(0)}/\sqrt{\sigma_{uu}^{(0)}} = \mu_u^{(0)}/\sqrt{\frac{1}{1-\rho^{(0)2}}}$ in standard units. This has now reduced the problem of estimating three parameters, $\{\mu_u^{(0)}, \sigma_{uu}^{(0)}, \rho^{(0)}\}$, to the problem of estimating two: $\{\omega^{(0)}, \rho^{(0)}\}$.

Once the parameters for respondents ($\{\mu_x^{(0)}, \sigma_{xx}^{(0)}, \omega^{(0)}, \rho^{(0)}\}$) have been estimated, the estimates for the nonrespondents' parameters follow from the same assumptions as in the continuous outcome case. For the nonrespondents we only observe $\{(X_i^{(1)}), i = n_{obs} + 1, \ldots, n\}$ with no information for the latent U since Y is unobserved. Equation (4.2) yields the parameters: $\{\mu_x^{(1)}, \sigma_{xx}^{(1)}, \mu_u^{(1)}, \sigma_{uu}^{(1)}, \rho^{(1)}\}$. Since we observe X for the nonrespondents, the first two can be estimated as standard maximum likelihood estimates: $\{\hat{\mu}_x^{(1)}, \hat{\sigma}_{xx}^{(1)}\} = \{\bar{x}_{NR}, s_{xx}^{(1)}\}$. Estimation of those parameters dealing with U requires the following assumption on the missing data:

$$\Pr(M=1|X,U,Y) = f\left(X\frac{1}{\sqrt{\sigma_{xx}^{(0)}(1-\hat{\rho}^{(0)2})}}+\lambda U\right). \tag{4.5}$$

As in the continuous case, the proxy X is scaled to have the same standard deviation as U and λ is a sensitivity parameter that determines the extent of nonignorability.

This assumption just identifies the remaining parameters and we can write them as:

$$\begin{aligned}
\mu_u^{(1)} &= \frac{\omega^{(0)}}{\sqrt{1-\rho^{(0)^2}}}+\frac{1}{\sqrt{\sigma_{xx}^{(0)}(1-\rho^{(0)2})}}\left(\frac{\lambda+\rho^{(0)}}{\lambda\rho^{(0)}+1}\right)(\mu_x^{(1)}-\mu_x^{(0)})\\
\sigma_{uu}^{(1)} &= \frac{1}{(1-\rho^{(0)^2})}+\frac{1}{\sigma_{xx}^{(0)}(1-\hat{\rho}^{(0)2})}\left(\frac{\lambda+\rho^{(0)}}{\lambda\rho^{(0)}+1}\right)^2(\sigma_{xx}^{(1)}-\sigma_{xx}^{(0)})\\
\rho^{(1)} &= \frac{1}{\sqrt{\sigma_{xx}^{(1)}\sigma_{uu}^{(1)}}}\left(\rho^{(0)}\sqrt{\frac{\sigma_{xx}^{(0)}}{(1-\rho^{(0)^2})}}+\frac{1}{\sqrt{\sigma_{xx}^{(0)}(1-\rho^{(0)2})}}\left(\frac{\lambda+\rho^{(0)}}{\lambda\rho^{(0)}+1}\right)(\sigma_{xx}^{(1)}-\sigma_{xx}^{(0)})\right)
\end{aligned} \tag{4.6}$$

The ML estimates of $\{\mu_u^{(1)}, \sigma_{uu}^{(1)}, \rho^{(1)}\}$ are found by substitution of the previously described MLEs for $\{\mu_x^{(0)}, \sigma_{xx}^{(0)}, \rho^{(0)}, \omega^{(0)}, \mu_x^{(1)}, \sigma_{xx}^{(1)}\}$.

4.2 Estimation

We have six parameters that require explicit estimation: $\{\mu_x^{(0)}, \sigma_{xx}^{(0)}, \rho^{(0)}, \omega^{(0)}, \mu_x^{(1)}, \sigma_{xx}^{(1)}\}$. We have already discussed using the marginal distribution of X conditional on respondent status, to obtain estimates $\{\hat{\mu}_x^{(0)}, \hat{\sigma}_{xx}^{(0)}, \hat{\mu}_x^{(1)}, \hat{\sigma}_{xx}^{(1)}\} = \{\bar{x}_R, s_{xx}^{(0)}, \bar{x}_{NR}, s_{xx}^{(1)}\}$. We are now interested in the second piece of the likelihood equation that will yield estimates of $\omega^{(0)}$ and $\rho^{(0)}$, but will require more work to maximize:

$$\log L = \sum_{j=1}^{n_{obs}}\{y_j\log(P_x)+(1-y_j)\log(1-P_x)\} \tag{4.7}$$

To define P_x, we use properties of bivariate normality, to write the conditional distribution of U given X and M as:

$$(U|X = x, m = 0) \sim N\left(\mu_u^{(0)} + \frac{\rho^{(0)}}{\sqrt{\sigma_{xx}^{(0)}(1-\rho^{(0)2})}}(x - \mu_x^{(0)}), 1\right) \tag{4.8}$$

Therefore,

$$\begin{aligned} P_x &= P(U > 0|X = x, m = 0) \qquad (4.9) \\ &= \Phi\left(\frac{\omega^{(0)}}{\sqrt{1-\rho^{(0)2}}} - \frac{\rho^{(0)}}{\sqrt{\sigma_{xx}^{(0)}(1-\rho^{(0)2})}}\mu_x^{(0)} + \frac{\rho^{(0)}}{\sqrt{\sigma_{xx}^{(0)}(1-\rho^{(0)2})}}x\right) \\ &= \Phi(a^{(0)} + b^{(0)}x) \end{aligned}$$

where $b^{(0)} = \sqrt{\frac{\rho^{(0)2}}{\sigma_{xx}^{(0)}(1-\rho^{(0)2})}}$ and $a^{(0)} = \frac{\omega^{(0)}}{\sqrt{1-\rho^{(0)2}}} - b\mu_x^{(0)}$.

4.2.1 Full MLE

To find the maximum likelihood estimates of $\{\omega^{(0)}, \rho^{(0)}\}$, we follow the work of Hannan and Tate (1965) [41]. We note that this is intended for the polytomous outcome, and therefore the cutpoint is opposite our formulation. All following notation will be in the case we consider; with a dichotomous outcome such that $\Pr(Y = 1|M = 0) = \Phi(\omega^{(0)})$, and with a univariate X.

To maximize the likelihood function of the sample $(X_j^{(0)}, Y_j^{(0)})$ the parameters $(\mu_x^{(0)}, \sigma_{xx}^{(0)}, \omega^{(0)}, \rho^{(0)})$ are transformed to a new set $(\mu_x^{(0)}, \sigma_{xx}^{(0)}, a^{(0)}, b^{(0)})$ to simplify the information matrix. Using these parameters and the formula for P_x, the log-likelihood equation in (4.7) can be written as

$$l(a^{(0)}, b^{(0)}) = \sum_{j=1}^{n_R} y_j \log(\Phi(a^{(0)} + b^{(0)}x_j)) + (1 - y_j)\log(1 - \Phi(a^{(0)} + b^{(0)}x_j)).$$

Maximizing the likelihood now requires us to maximize with respect to $a^{(0)}$ and $b^{(0)}$, which are functions of $\omega^{(0)}$ and $\rho^{(0)}$. We will maximize the likelihood using R's ‘optim’ function, which maximizes a function using Nelder-Mead and requires initial values for $a^{(0)}$ and $b^{(0)}$. Hannan and Tate suggest the following equations for the initial values:

$$\begin{aligned} b^* &= \sqrt{\frac{r^{*2}}{s_{xx}(1-r^{*2})}} \\ a^* &= \frac{\omega^*}{\sqrt{1-r^{*2}}} - b^*\bar{x} \end{aligned} \tag{4.10}$$

where r^* is the biserial estimator of ρ, and ω^* is $\Phi^{-1}(\bar{y}_R)$.

Once the maximum likelihood estimates for $a^{(0)}$ and $b^{(0)}$ are found they are transformed to obtain the MLEs for $\omega^{(0)}$ and $\rho^{(0)}$:

$$\begin{aligned} \hat{\rho}^{(0)} &= \sqrt{\frac{\hat{b}^{(0)2} s_{xx}^{(0)}}{1+\hat{b}^{(0)2} s_{xx}^{(0)}}} \\ \hat{\omega}^{(0)} &= (\hat{a}^{(0)} + \hat{b}^{(0)}\bar{x})\sqrt{1-\hat{\rho}^{(0)2}}. \end{aligned} \tag{4.11}$$

4.2.2 Two-Step Estimation

An alternative estimation method for $\omega^{(0)}$ and $\rho^{(0)}$ is the two-step method of Olsson (1982) [42]. The first step is to estimate $\omega^{(0)}$ with $\hat{\omega}^{(0)} = \Phi^{(-1)}(\bar{y}_R)$. The second step uses $\hat{\omega}^{(0)}$ in the log-likelihood, and maximizes with respect to $\rho^{(0)}$ only. This formulation of the estimate of the cutpoint returns a natural estimate of the mean of Y for the respondents: $\hat{\mu}_Y^{(0)} = \Phi(\hat{\omega}^{(0)}) = \bar{y}_R$. This procedure is computationally simpler than the simultaneous estimation of the full MLE method, and will produce similar estimates when the model properties are satisfied (e.g., normality).

4.2.3 Estimates

In summary, we have the following MLEs for the respondents:

$$
\begin{align}
\hat{\mu}_x^{(0)} &= \bar{x}_R \nonumber \\
\hat{\sigma}_{xx}^{(0)} &= s_{xx}^{(0)} \nonumber \\
\hat{\omega}^{(0)}, \hat{\rho}^{(0)} &: \text{ from maximization (either full MLE or two-step)} \nonumber \\
\hat{\mu}_u^{(0)} &= \hat{\omega}^{(0)}/\sqrt{1-\hat{\rho}^{(0)^2}} = (\hat{a}^{(0)} + \hat{b}^{(0)}\bar{x}_R) \tag{4.12} \\
\hat{\sigma}_{uu}^{(0)} &= 1/(1-\hat{\rho}^{(0)^2}) \nonumber \\
\hat{\sigma}_{xu}^{(0)} &= \hat{\rho}^{(0)}\sqrt{\hat{\sigma}_{xx}^{(0)}/(1-\hat{\rho}^{(0)^2})} \nonumber
\end{align}
$$

and for the nonrespondents:

$$
\begin{align}
\hat{\mu}_x^{(1)} &= \bar{x}_{NR} \nonumber \\
\hat{\sigma}_{xx}^{(1)} &= s_{xx}^{(1)} \nonumber \\
\hat{\mu}_u^{(1)} &= \frac{\hat{\omega}^{(0)}}{\sqrt{1-\hat{\rho}^{(0)^2}}} + \frac{1}{\sqrt{s_{xx}^{(0)}(1-\hat{\rho}^{(0)2})}} \left(\frac{\lambda+\hat{\rho}^{(0)}}{\lambda\hat{\rho}^{(0)}+1}\right)(\bar{x}_{NR}-\bar{x}_R) \tag{4.13} \\
\hat{\sigma}_{uu}^{(1)} &= \frac{1}{(1-\hat{\rho}^{(0)^2})} + \frac{1}{s_{xx}^{(0)}(1-\hat{\rho}^{(0)2})} \left(\frac{\lambda+\hat{\rho}^{(0)}}{\lambda\hat{\rho}^{(0)}+1}\right)^2 (s_{xx}^{(1)} - s_{xx}^{(0)}) \nonumber \\
\hat{\rho}^{(1)} &= \frac{1}{\sqrt{s_{xx}^{(1)}\hat{\sigma}_{uu}^{(1)}}} \left(\hat{\rho}^{(0)}\sqrt{\frac{s_{xx}^{(0)}}{(1-\hat{\rho}^{(0)^2})}} + \frac{1}{\sqrt{s_{xx}^{(0)}(1-\hat{\rho}^{(0)2})}} \left(\frac{\lambda+\hat{\rho}^{(0)}}{\lambda\hat{\rho}^{(0)}+1}\right)(s_{xx}^{(1)} - s_{xx}^{(0)})\right) \nonumber
\end{align}
$$

4.3 Fitted Values

As in the case with a continuous outcome Y, to apply the bootstrap proxy hot deck, we need to match on 'something'. In the continuous case, we found $\hat{Y}$'s for all subjects, and matched using the Siddique and Belin distance function. We will follow the same steps, except we will find $\hat{U}$, where U is the latent variable for Y, and is itself unobserved. From Andridge (2009) [38], we have the following conditional distribution for nonrespondents ($m = 1$) and respondents ($m = 0$) of Y:

$$[u_i|x_i, m_i = m, \phi] \sim N\left(\mu_u^{(m)} + \rho^{(m)}\sqrt{\frac{\sigma_{uu}^{(m)}}{\sigma_{xx}^{(m)}}}\left(x_i - \mu_x^{(m)}\right), \sigma_{uu}^{(m)}\left(1 - \rho^{(m)2}\right)\right) \tag{4.14}$$

Substitution of the MLEs from (4.11) and (4.13) yields, for the respondents,

$$\begin{aligned} \hat{u}_i^{(0)} &= \hat{\mu}_u^{(0)} + \hat{\rho}^{(0)}\sqrt{\frac{\hat{\sigma}_{uu}^{(0)}}{\hat{\sigma}_{xx}^{(0)}}}\left(x_i - \hat{\mu}_x^{(0)}\right) \qquad (4.15) \\ &= \frac{\hat{\omega}^{(0)}}{\sqrt{1 - \hat{\rho}^{(0)^2}}} + \sqrt{\frac{\hat{\rho}^{(0)^2}}{s_{xx}^{(0)}(1 - \hat{\rho}^{(0)^2})}}(x_i - \bar{x}_R), \end{aligned}$$

and for the nonrespondents,

$$\begin{aligned} \hat{u}_i^{(1)} &= \hat{\mu}_u^{(1)} + \hat{\rho}^{(1)}\sqrt{\frac{\hat{\sigma}_{uu}^{(1)}}{\hat{\sigma}_{xx}^{(1)}}}\left(x_i - \hat{\mu}_x^{(1)}\right) \qquad (4.16) \\ &= \frac{\hat{\omega}^{(0)}}{\sqrt{1 - \hat{\rho}^{(0)^2}}} + \hat{g}^{\lambda}(x_i - \bar{x}_R) + \frac{\sqrt{s_{xx}^{(0)}}}{s_{xx}^{(1)}}\left(\frac{\hat{\rho}^{(0)}}{\sqrt{1 - \hat{\rho}^{(0)2}}} - \hat{g}^{\lambda}\sqrt{s_{xx}^{(0)}}\right)(x_i - \bar{x}_{NR}) \end{aligned}$$

where

$$\hat{g}^{\lambda} = \frac{1}{\sqrt{s_{xx}^{(0)}(1 - \hat{\rho}^{(0)^2})}}\left(\frac{\lambda + \hat{\rho}^{(0)}}{\lambda\hat{\rho}^{(0)} + 1}\right).$$

We can examine the value of $\hat{u}_i$ for different values of λ. We use $\lambda = 0, 1, \infty$ to represent different missingness assumptions, according to Equation (2.21).

Under the assumption of MAR, $\lambda = 0$, and $\hat{g}^{\lambda=0} = \frac{\hat{\rho}^{(0)}}{\sqrt{s_{xx}^{(0)}(1-\hat{\rho}^{(0)^2})}}$. This yields,

$$\hat{u}_i^{(1,\lambda=0)} = \frac{\hat{\omega}^{(0)}}{\sqrt{1 - \hat{\rho}^{(0)^2}}} + \frac{\hat{\rho}^{(0)}}{\sqrt{s_{xx}^{(0)}(1 - \hat{\rho}^{(0)^2})}}(x_i - \bar{x}_R) \tag{4.17}$$

which is equivalent to $\hat{u}$ for the respondents. Here, the $(x_i - x_{NR})$ term drops out because the coefficient equals 0:

$$\begin{aligned}
&\left(\frac{\hat{\rho}^{(0)}}{\sqrt{1-\hat{\rho}^{(0)2}}}\sqrt{\frac{s_{xx}^{(0)}}{s_{xx}^{(1)^2}}} - \frac{\hat{\rho}^{(0)}}{\sqrt{s_{xx}^{(0)}(1-\hat{\rho}^{(0)2})}}\frac{s_{xx}^{(0)}}{s_{xx}^{(1)}}\right) \\
&= \left(\frac{\hat{\rho}^{(0)}}{\sqrt{1-\hat{\rho}^{(0)2}}}\sqrt{\frac{s_{xx}^{(0)}}{s_{xx}^{(1)^2}}} - \frac{\hat{\rho}^{(0)}}{\sqrt{(1-\hat{\rho}^{(0)2})}}\sqrt{\frac{s_{xx}^{(0)}}{s_{xx}^{(1)^2}}}\right) \\
&= 0.
\end{aligned}$$

When $\lambda = \infty$, which is the extreme MNAR situation, $\hat{g}^{\lambda=\infty} = \frac{1}{\sqrt{s_{xx}^{(0)}\hat{\rho}^{(0)2}(1-\hat{\rho}^{(0)2})}}$. Thus,

$$\begin{aligned}
\hat{u}_i^{(1,\lambda=\infty)} &= \frac{\hat{\omega}^{(0)}}{\sqrt{1-\hat{\rho}^{(0)2}}} + \frac{1}{\sqrt{s_{xx}^{(0)}\hat{\rho}^{(0)2}(1-\hat{\rho}^{(0)2})}}(x_i - \bar{x}_R) \\
&\quad + \sqrt{\frac{s_{xx}^{(0)}}{s_{xx}^{(1)^2}}}\left(\frac{\hat{\rho}^{(0)2}-1}{\hat{\rho}^{(0)}\sqrt{1-\hat{\rho}^{(0)2}}}\right)(x_i - \bar{x}_{NR}).
\end{aligned} \tag{4.18}$$

When $\lambda = 1$, $\hat{g}^{\lambda=1} = \frac{1}{\sqrt{s_{xx}^{(0)}(1-\hat{\rho}^{(0)2})}}$. Thus,

$$\begin{aligned}
\hat{u}_i^{(1,\lambda=0)} &= \frac{\hat{\omega}^{(0)}}{\sqrt{1-\hat{\rho}^{(0)2}}} + \frac{1}{\sqrt{s_{xx}^{(0)}(1-\hat{\rho}^{(0)2})}}(x_i - \bar{x}_R) \\
&\quad + \sqrt{\frac{s_{xx}^{(0)}}{s_{xx}^{(1)^2}}}\left(\frac{\hat{\rho}^{(0)}-1}{\sqrt{1-\hat{\rho}^{(0)2}}}\right)(x_i - \bar{x}_{NR}).
\end{aligned} \tag{4.19}$$

To attempt to simplify the formulas, we can also write the $\hat{u}_i$ equations in terms of $a^{(0)}$ and $b^{(0)}$ for the different values of λ. Note that the formulation by Hannan and Tate (1965) yields $\frac{\omega^{(0)}}{\sqrt{1-\rho^{(0)2}}} = a^{(0)} + b^{(0)}\mu_x$.

First, for respondents,

$$\begin{aligned}
\hat{u}_i^{(0)} &= \frac{\hat{\omega}^{(0)}}{\sqrt{1-\hat{\rho}^{(0)2}}} + \sqrt{\frac{\hat{\rho}^{(0)2}}{s_{xx}^{(0)}(1-\hat{\rho}^{(0)2})}}(x_i - \bar{x}_R) \\
&= \hat{a}^{(0)} + \hat{b}^{(0)}\bar{x}_R + \hat{b}^{(0)}(x_i - \bar{x}_R) \\
&= \hat{a}^{(0)} + \hat{b}^{(0)}x_i
\end{aligned}$$

For nonrespondents,

$$\begin{aligned}
\hat{u}_i^{(m=1,\lambda=0)} &= \hat{a}^{(0)} + \hat{b}^{(0)}x_i \qquad (4.20)\\
\hat{u}_i^{(m=1,\lambda=1)} &= \hat{u}_i^{(m=1,\lambda=0)}\\
&+ \hat{b}^{(0)}\left(\frac{1-\hat{\rho}^{(0)}}{\hat{\rho}^{(0)}}\right)(x_i - \bar{x}_R) - \hat{b}^{(0)}\left(\frac{1-\hat{\rho}^{(0)}}{\hat{\rho}^{(0)}}\right)\frac{s_{xx}^{(0)}}{s_{xx}^{(1)}}(x_i - \bar{x}_{NR})\\
\hat{u}_i^{(m=1,\lambda=\infty)} &= \hat{u}_i^{(m=1,\lambda=1)}\\
&+ \frac{1}{\hat{\rho}^{(0)}}\left[\hat{b}^{(0)}\left(\frac{1-\hat{\rho}^{(0)}}{\hat{\rho}^{(0)}}\right)(x_i - \bar{x}_R) - \hat{b}^{(0)}\left(\frac{1-\hat{\rho}^{(0)}}{\hat{\rho}^{(0)}}\right)\frac{s_{xx}^{(0)}}{s_{xx}^{(1)}}(x_i - \bar{x}_{NR})\right].
\end{aligned}$$

We can see that for a given nonrespondent, increasing λ from 0 to ∞ will either increase or decrease the value of $\hat{u}$, depending on the relationship between $x_i, \bar{x}_R, \bar{x}_{NR}$, and $\rho^{(0)}$, and how different the variance of the proxy is for the respondents and nonrespondents. The same term that is added to the predicted mean when λ increases from 0 to 1 is added when λ increases from 1 to ∞ multiplied by a factor of $\frac{1}{\hat{\rho}^{(0)}}$. Therefore, if the proxy is very weak, a larger term is added, whereas a strong proxy has a smaller term. For example, the values of $\frac{1}{\rho^{(0)}}$ for values of $\rho^{(0)} = \{0.1, 0.2, 0.5, 0.8\}$ are $\{10, 5, 2, 1.25\}$.

4.4 Siddique and Belin Distance Measure

Figure (4.1) shows examples of using the Siddique and Belin distance measure on the $\hat{u}$'s for two nonrespondents in simulated data. The top panels are $\hat{u}$'s versus absolute distance between the nonrespondent (red line) and each respondent. The middle panel is $\hat{u}$ versus the distance (absolute distance to the power $k = 3$). The bottom panel is $\hat{u}$ versus the selection probability. The bottom two panels also separate the respondents into $y = 0$ and $y = 1$. The figures on the left show a pattern typical of most of the nonrespondents examined, and illustrate that even though the k is non-zero, there is only one respondent with a very large probability, and the hot deck will turn into nearest neighbor. The figures on the right illustrate the less common result, where there is not one point that stands away from the rest.

4.4.1 Use of Adjustment Cells

Due to what was observed with the selection probabilities, we propose using adjustment cells to find donors for the imputations instead of using Siddique and Belin's distance measure. Adjustment cells are a common method for hot deck imputation of a categorical variable [43]. If we use $\hat{u}$'s from the respondents and nonrespondents to form the H cells, we may have the problem that some cells only contain nonrespondents (or respondents), due to the different missingness assumptions. Because of this, we use one suggestion in [43]; we form $H = 10$ equally sized cells based on the quantiles of the $\hat{u}^{(0)}$'s of the respondents. Then each nonrespondent will be placed into the closest cell. This closeness is determined by the Euclidean distance between the $\hat{u}_i^{(1)}$ and the mean of all the $\hat{u}_i^{(0)}$'s in the cell. So, nonrespondent i will be placed in cell h, such that the following is minimized:

$$\left|\hat{u}_i^{(1)} - \frac{1}{n_h}\sum_{j=1}^{n_h}\hat{u}_j^{(0)}\right|. \tag{4.21}$$

Here, n_h is the total number of respondents in cell h. This will ensure that every cell containing a nonrespondent has respondents available to be donors, and there will also be an equal number of possible donors for each nonrespondent. This eliminates the problem described in Section (4.4) at the cost of possibly increased bias due to less than perfect matches.

Figure (4.2) illustrates how the $\hat{u}'s$ figure into the adjustment cells. The first figure shows results from simulated data that are MAR with a strong proxy ($\rho = 0.8$). The

$\hat{u}'s$ for the nonrespondents were formed using $\lambda = 0$. Each pair of boxplots represents one adjustment cell, with the unshaded boxplot the distribution of $\hat{u}_i^{(0)}$'s and the shaded boxplots the distribution of $\hat{u}_i^{(1)}$'s. The tables on the graph display, for the respondents, the number of 0's and 1's (values of Y) in each adjustment cell. We can see that there are approximately the same total number of respondents in each cell, and as the value of $\hat{u}$ increases (the adjustment cell number increases), the proportion of 1's in each cell increase. The table labeled "Imputations" is an example of a single imputation for the nonrespondents. We first notice that the number of nonrespondents in each cell varies, from 2 in cell 1 to 27 in cell 10. We also notice that the cells representing the largest $\hat{u}_i^{(1)}$ values (higher numbered cells) had 1's imputed more often. The lower figure contains information for data generated using a different distribution and missingness mechanism. The data were generated using an Exponential distribution for Z, the missingness mechanism was MNAR, and the values of $\hat{u}_i^{(1)}$ were formed using $\lambda = \infty$. Note how much larger the $\hat{u}_i^{(1)}$'s become, compared to the upper figure, with a maximum $\hat{u}_i^{(1)}$ value for the nonrespondent of approximately 14 compared to approximately 4 in the upper figure. There is also a large number of nonrespondents that are closest to the largest adjustment cell, many with very large values of $\hat{u}_i^{(1)}$. If the distance measure of Siddique and Belin were applied, this would likely turn into a simple random hot deck of all the respondents for those nonrespondents with very large predicted means. The use of adjustment cells prevents this, and forces the nonrespondents with the largest $\hat{u}_i$'s to be matched to a respondent that is among the largest $\hat{u}_i^{(0)}$'s.

4.5 Formal Steps of the PPM Hot Deck

The steps of the PPM hot deck for imputing a binary outcome are similar to those for a continuous variable, except that a proxy is created for the latent U instead of the outcome Y. We also adjust the method through use of adjustment cells to find the donors. The following are the method's steps:

For a specified λ representing a chosen missingness assumption, and chosen parameter estimation method and imputation 1 of D,

1. **Bootstrap**: Generate a bootstrap sample of the respondents by selecting n_{obs} with replacement. Denote these respondent outcomes and covariate values as $\{Y_j^b, \mathbf{Z}_j^b\}$, where b distinguishes the bootstrap sample from the original sample.

2. **Proxy**: Create the Proxy X for the latent U for the (bootstrapped) respondents and nonrespondents. Note that only respondents who are selected into the bootstrap sample have a proxy created, while all nonrespondents have a proxy created:

    (a) Fit a probit regression of Y on Z^b (the bootstrapped respondents), $\Pr(Y^b = 1|Z^b, M = 0) = \Phi(\alpha_0 + \alpha_1 Z^b)$, to obtain regression parameter estimates $(\hat{\alpha}_0, \hat{\alpha}_1)$.

    (b) Use the estimated coefficients to form the proxy, X, for all nonrespondents and all respondents in the bootstrap sample:
    $x_i = \hat{\alpha}_0 + \hat{\alpha}_1 z_i, i = 1, \ldots, n.$

3. **Predicted Values**: For the chosen estimation method, estimate the model parameters and calculate the predicted values $\hat{U}$ based on the pattern-mixture

model in (4.2) and identifying restriction in (4.5) for the selected value of λ. Under this model, predicted values for respondents (in the bootstrap sample) and nonrespondents are given by

$$
\begin{aligned}
\hat{u}_i^{b(0)} &= \hat{\mu}_u^{(0)} + \hat{\rho}^{(0)} \sqrt{\frac{\hat{\sigma}_{uu}^{(0)}}{s_{xx}^{(0)}}} \left(x_i^b - \bar{x}_R^b\right), \\
\hat{u}_i^{(1,\lambda=\lambda)} &= \hat{\mu}_u^{(0)} + \sqrt{\frac{\hat{\sigma}_{uu}^{b(0)}}{s_{xx}^{b(0)}}} \left(\frac{\lambda + \hat{\rho}^{b(0)}}{1 + \lambda \hat{\rho}^{b(0)}}\right) \left(\bar{x}_{NR} - \bar{x}_R^b\right) \qquad (4.22) \\
&\quad + \left[\hat{\rho}^{(0)} \frac{\sqrt{s_{xx}^{(0)} \hat{\sigma}_{uu}^{(0)}}}{s_{xx}^{(1)}} + \left(\sqrt{\frac{s_{uu}^{b(0)}}{s_{xx}^{b(0)}}} \cdot \frac{\lambda + \hat{\rho}^{b(0)}}{\lambda \hat{\rho}^{b(0)} + 1}\right) \left(1 - \frac{s_{xx}^{b(0)}}{s_{xx}^{(1)}}\right)\right] \left(x_i - \bar{x}_{NR}\right).
\end{aligned}
$$

We use the superscript b to denote quantities that are calculated on a bootstrapped sample, to distinguish them from quantities calculated on the whole sample. For example, $\bar{x}_R^b$ is the mean of X for the bootstrapped sample of respondents, while $\bar{x}_{NR}$ is the mean of X for the entire sample of nonrespondents. Note that for respondents the predicted means do not depend on λ. However, they vary for each cycle through the PPM hot deck because of the bootstrapping (step 1).

4. **Adjustment Cells**:

    (a) Adjustment cells are created for the bootstrapped respondents. The $\hat{U}^{(0)}$'s for the bootstrapped respondents are divided into 10 cells, with an equal number of respondents in each.

    (b) The nonrespondents are then placed into one of the 10 groups, by minimizing the Euclidean distance between the $\hat{u}_i^{(1)}$ value of the nonrespondent, and the mean of the $\hat{u}_i^{(0)}$'s in a cell.

5. **Select and Impute**: For each nonrespondent, select a donor by randomly drawing a donor from the cell with equal probability. Impute the donor's value of Y for the missing value of the donee. Repeat this process for all nonrespondents.

6. **Repeat** steps 1-5 D times to create D complete datasets, composed of the original (pre-bootstrap) respondents and the imputed nonrespondent data. This entire process should be completed separately for each value of λ in the sensitivity analysis.

4.6 Simulation

We conducted a simulation study to assess the bias and coverage of the proxy pattern-mixture hot deck when estimating the mean of a partially observed binary outcome. Of particular interest is examining the sensitivity to non-normality of the proxy X or the latent variable U. For the parametric method, the assumption of normality in the pattern-mixture model is more critical when Y is binary than when Y is continuous. For binary Y, we examine the model for the latent U and proxy X: $U = X + \epsilon$, where $X \perp e$. We consider three scenarios representing combinations of normality and various degrees of non-normality of both the proxy X and error term e. The three scenarios are:

1. Both X and ϵ are normal

2. X is nonnormal, ϵ is normal

3. ϵ is nonnormal, X is normal

Situation (1) does not violate the normality assumption and will serve as the control. Non-normality of the proxy (situation (2)) is studied in Andridge and Little

(2009) for the parametric PPM model [38]. Non-normality of X causes the marginal distribution of U for respondents to be non-normal, which causes the ML estimate of $\mu_y^{(0)}$ to be biased. Andridge and Little found this to impact the maximum likelihood method and also the fully Bayesian approach. The multiple imputation method is fairly robust in this case because the conditional distribution U given X is still normal. However, when the error term, e, is non-normal, $U|X$ is non-normal and the MI approach may be affected. The three methods which will be compared are:

1. Proposed PPM hot deck with full maximum likelihood estimation
2. Proposed PPM hot deck with Olsson's two-step estimation
3. Fully parametric PPM model using multiple imputation

Following results in [38], we expect estimates to be more robust under Olsson's two-step method. We note that we do not compare results of the three methods mentioned above to the Siddique and Belin nonignorable hot deck since, as was previously noted, the extension of their method to a binary outcome is not straightforward.

4.6.1 Data Generation

To study bias and coverage of the proxy pattern mixture hot deck, we proceed by generating a complete data set, and imposing missingness. We are particularly interested in the sensitivity to non-normality. The overview of simulation steps are below, followed by the specific details on selecting parameter values. The three distributions considered for the single covariate Z and the latent U are (1) $N(0,1)$, (2) $Gamma(\text{shape} = 4, \text{scale} = 1/2)$, and (3) $Exp(1)$. They are combined such that at least one of Z and e is Normally distributed.

We consider a sample size of $N = 400$ and correlations of $\rho = \{0.8, 0.5, 0.2\}$ to represent strong, moderate and weak proxies for U, respectively.

1. Generate covariate $Z_i, i = 1, \ldots, N$ according to specified distribution.

2. Create latent U to generate Y for a given value of ρ according to the following requirements. Specific values are discussed in Section (4.6.2) and shown in Table 4.1:

   (a) Specify α_1 such that $\text{Corr}(U, X) = \rho$

   (b) Specify α_0 such that $E[Y] = 0.3$

   (c) Generate error e for chosen distribution and set $U_i = \alpha_0 + \alpha_1 Z_i + e_i$

3. Create binary Y from the latent U:

   (a) $Y_i = I[u_i > 0], i = 1, \ldots, N$, with $I[\cdot]$ the indicator function.

4. Induce missingness of Y according to a selection model for $[M|U, X]$, inducing either MAR or MNAR.

Once U has been used to create Y and generate the missingness, it is discarded, resulting in the data $\{Z_i, Y_i, M_i\}$. We used 500 replications and 10 imputed datasets. For each replication, data were generated and missingness induced to create an incomplete data set of observed values. Multiple imputation using the PPM hot deck for all three values of λ was then performed and results combined using standard MI combining rules.

4.6.2 Choosing values for the Simulation Study

The three distributions considered are (1) $N(0,1)$, (2) $Gamma$(shape $=$ 4, scale $=$ 1/2), and (3) $Exp(1)$. These all have variance 1, and the Gamma and Exponential will be shifted to have mean 0, to make calculations simpler. Regardless of the choice of distribution, we make use of the following equations for finding the values used in the simulations.

$$X = a_0 + a_1 Z \tag{4.23}$$

$$\mu_x = E[X] = \alpha_0 + \alpha_1 E[Z] = \alpha_0 + \alpha_1 \mu_z$$

$$\sigma_{xx} = Var[X] = Var(\alpha_0 + \alpha_1 Z) = Var(\alpha_1 Z) = \alpha_1^2 Var(Z) = \alpha_1^2 \sigma_{zz}$$

$$U = X + e \tag{4.24}$$

$$\mu_u = E[U] = E[X + e] = \alpha_0 + \alpha_1 \mu_z$$

$$\sigma_{uu} = Var(U) = Var(X + e) = Var(\alpha_0 + \alpha_1 Z + e) = \alpha_1^2 \sigma_{zz} + \sigma_{ee}$$

The value of α_1 is used to maintain the proper correlation between U and X, and the value of α_0 dictates the mean of Y. For α_1, we note:

$$\begin{aligned} Cov(U,X) = Cov(X+e, X) &= Cov(\alpha_0 + \alpha_1 Z + e, \alpha_0 + \alpha_1 Z) \\ &= Cov(\alpha_1 Z, a_1 Z) + Cov(\alpha_1 Z, e) \\ &= \alpha_1^2 Var(Z) + \alpha_1 * 0 \qquad (4.25) \\ &= \alpha_1^2 \sigma_{zz} \end{aligned}$$

It follows that the correlation between U and the proxy X is given by

$$\begin{aligned} \rho(U,X) &= \frac{Cov(U,X)}{\sqrt{\sigma_{uu}\sigma_{xx}}} = \frac{\alpha_1^2 \sigma_{zz}}{\sqrt{\sigma_{uu}\alpha_1^2 \sigma_{zz}}} \\ &= \frac{\alpha_1 \sqrt{\sigma_{zz}}}{\sqrt{\sigma_{uu}}} \qquad (4.26) \end{aligned}$$

$$= \quad \frac{\alpha_1\sqrt{\sigma_{zz}}}{\sqrt{\alpha_1^2\sigma_{zz}+\sigma_{ee}}},$$

which implies,

$$\rho = \frac{a_1\sqrt{\sigma_{zz}}}{\sqrt{\alpha_1^2\sigma_{zz}+\sigma_{ee}}} \iff \alpha_1 = \frac{\sigma_{ee}}{\sqrt{\sigma_{zz}}}\frac{\rho}{\sqrt{1-\rho^2}}. \tag{4.27}$$

Therefore, since all chosen distributions have variance 1 (i.e., $\sigma_{zz} = \sigma_{ee} = 1$), we set $\alpha_1 = \rho/\sqrt{1-\rho^2}$ for all settings. The chosen distribution does matter for the value of α_0 (dictating the mean of Y), and the calculations follow.

Normal Z, Normal e

When Z_i are simulated from $N(0,1)$, and the U_i are simulated from $N(\alpha_0+\alpha_1 Z_i, 1)$ with $\alpha_1 = \rho/\sqrt{1-\rho^2}$, we have,

$$E[Y] = Pr(U > 0) = \Phi\left(\mu_u/\sqrt{\sigma_{uu}}\right).$$

In this setting, $\Phi\left(\mu_u/\sqrt{\sigma_{uu}}\right) = \Phi\left(\alpha_0/\sqrt{\alpha_1^2+1}\right)$ and α_0 is chosen to be $\alpha_0 = \Phi^{-1}(0.3)\sqrt{\alpha_1^2+1}$ such that

$$E[Y] = \Phi\left(\left(\Phi^{-1}(0.3)\sqrt{\alpha_1^2+1}\right)/\sqrt{\alpha_1^2+1}\right) = \Phi(\Phi^{-1}(0.3)) = 0.3.$$

Exponential Z, Normal e

To find α_0, we begin with the case of non-normal Z (and therefore a non-normal proxy). Recall that $U = X + e = \alpha_0 + \alpha_1 Z + e$. We let $e \sim N(0,1)$ and $Z = Z^* - 1$, where $Z^* \sim Exp(1)$ so that Z has mean 0 and variance 1. We let $e^* \sim N(\alpha_0, 1)$, and $Z^{**} \sim Exp(\alpha_1)$, to make $U = Z^{**} + e^*$ a convolution of Z^{**} and e^*. Theorem 5.2.9 in Casella and Berger [44] states that if X and Y are independent continuous random variables with pdf's $f_X(x)$ and $f_Y(y)$, then the pdf of $Z = X + Y$ is

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(w)f_Y(z-w)dw \equiv \int_{-\infty}^{\infty} f_X(z-w)f_Y(w)dw.$$

Therefore, we choose α_0 such that,

$$0.3 = E[Y = 1] = Pr(U > 0) = \int_0^\infty \int_{-\infty}^\infty f_{Z^{**}}(w) f_{e^*}(u - w) dw du \tag{4.28}$$

where $f_{Z^{**}}$ is the pdf of $Exp(\alpha_1)$ and f_{e^*} is the pdf of $N(\alpha_0 - \alpha_1, 1)$.

Gamma Z, Normal e

Similary, when $Z \sim Gamma(4, 0.5)$, we solve,

$$0.3 = E[Y = 1] = Pr(U > 0) = \int_0^\infty \int_{-\infty}^\infty f_{Z^{**}}(w) f_{e^*}(u - w) dw du \tag{4.29}$$

where $f_{Z^{**}}$ is the pdf of $Gamma(4, \alpha_1/2)$ and f_{e^*} is the pdf of $N(\alpha_0 - 2\alpha_1, 1)$ for α_0.

Normal Z, Exponential e

We next need to find α_0 in the case of non-normal e. Again, we have $U = X + e = \alpha_0 + \alpha_1 Z + e$. We let $Z \sim N(0, 1)$ and $e = e^* - 1$, where $e^* \sim Exp(1)$ so that e has mean 0 and variance 1. We then have $U = \alpha_0 + \alpha_1 Z + e^* - 1$. If we let $Z^{**} \sim N(\alpha_0 - 1, \alpha_1^2)$, and $e^* \sim Exp(1)$, then $U = Z^{**} + e^*$ is a convolution of Z^{**} and e^*.

Therefore, we choose α_0 such that,

$$0.3 = E[Y = 1] = Pr(U > 0) = \int_0^\infty \int_{-\infty}^\infty f_{Z^{**}}(w) f_{e^*}(u - w) dw du \tag{4.30}$$

where f_{Z**} is the pdf of $N(\alpha_0 - 1, \alpha_1^2)$ and f_{e^*} is the pdf of $Exp(1)$.

Normal Z, Gamma e

Similarly, when $Z \sim Gamma(4, 0.5)$, we solve,

$$0.3 = E[Y = 1] = Pr(U > 0) = \int_0^\infty \int_{-\infty}^\infty f_{Z^{**}}(w) f_{e^*}(u - w) dw du \tag{4.31}$$

where f_{Z**} is the pdf of $N(\alpha_0 - 2, \alpha_1^2)$ and f_{e^*} is the pdf of $Gamma(4, 0.5)$ for α_0.

Missingness Model Missing data on Y was induced according to the selection model:

$$\text{logit}\{\Pr(M_i = 1|Z_i, U_i)\} = \gamma_0 + \gamma_Z Z_i + \gamma_U U_i \tag{4.32}$$

We only consider MAR ([Z]) and severe MNAR ([U]). For MAR, we use $\gamma_z = 0.5$, and $\gamma_u = 0$. For extreme MNAR, we set $\gamma_z = 0$ and $\gamma_u = 0.5$ with γ_0 was chosen to create approximately 25% missingness. To obtain approximately 25% missingness, we want:

$$\Pr(M = 1|U, Z) = \frac{\exp(\gamma_0 + \gamma_Z Z + \gamma_U U)}{1 + \exp(\gamma_0 + \gamma_Z Z + \gamma_U U)} \tag{4.33}$$

If the logit equals log(0.25/0.75) we obtain, Pr(M=1 | U,Z) = (0.25/0.75)/(1 + (0.25/0.75)) = 0.25. For MAR, $E[Z] = 0$ and $\gamma_U = 0$, yielding $E[\text{logit}] = \gamma_0$ and we set $\gamma_0 = \log(0.25/0.75)$. This is true for either Z non-normal or e non-normal. Under extreme MNAR, $E[Z] = 0$, and $E[e] = 0$, implying $E[U] = \alpha_0$ and $E[\text{logit}] = \gamma_0 + \gamma_U \alpha_0$. Therefore, we simply adjust γ_0 by substracting $\gamma_U \alpha_0$ to obtain 25% missingness. The values are shown in Table 4.1.

4.6.3 Results

We now compare results from the PPM hot deck and the parametric MI method for two missingness assumptions described previously, changing the distribution of Z, and therefore the proxy, X, and also changing the distribution of e. For the PPM hot deck we also show results from using the two different MLE methods, Tate's full method (HD FMLE) and Olsson's two-step method (HD 2-Step). The two methods made a difference in [38] when changing Z – the overall mean of Y was more biased using the full MLE method. As we will see, the choice of estimation method does not significantly impact the estimates for the hot deck.

Normal Z, Normal e

Table 4.2 contains the results when both Z and the error e are normally distributed. Results obtained using the full ML estimation method and the two-step estimation method are similar, and we describe the full MLE results. For the PPM hot deck, under missing at random, $[Z]$, $\lambda = 0$ should produce the best results of all the λ's and it does (Table 4.2). It is unbiased for all correlations, (empirical bias=-0.3%,-0.3%,-0.1%, respectively), and also has nominal coverage (94.6, 95.2, 94.8). For mechanism MNAR, $[U]$, $\lambda = \infty$ has the least bias and best coverage compared to $\lambda = 0$ and 1. However, it does slightly underestimate the mean, with bias -1.1%, -1.6%, and -3.5% and has coverage less than 95%, (90.8, 88.8, 77.4), with the results being worse for the weaker proxy. The confidence interval width shows that, for each missingness mechanism and correlation, the width increases as λ increases. For each correlation and missingness mechanism, the estimate of the mean of Y does increase also as λ increases from 0 to ∞, which is indicative of the fact that larger values of Y are missing.

Gamma Z, Normal e

Tables 4.3 and 4.4 contain the results from changing the distribution of Z (and thus the distribution of X). When $Z \sim Gamma(2, 0.5)$ (Table 4.3), for both MLE methods, the appropriate λ has least bias and highest coverage for both MAR ($\lambda = 0$) and MNAR ($\lambda = \infty$), and all correlations. For correlation 0.8, $\lambda = 0$ is unbiased, with 94.4% coverage under MAR, and $\lambda = \infty$ is unbiased with $\approx$ 93% coverage under MNAR. For correlation 0.5, $\lambda = 0$ is still unbiased with nominal coverage ($\approx$ 95%) for MAR. Under MNAR, $\lambda = \infty$ is more biased than under $\rho = 0.8$, with lower coverage

($\approx 88\%$). With a weak proxy, $\lambda = 0$ is unbiased and at nominal coverage under MAR. When data are MNAR, $\lambda = \infty$ is somewhat biased, and has low coverage (77%). For all correlations, within each of the three methods, the confidence interval widths also increase with λ, under each MAR and MNAR. The amount of bias and coverage, for all situations, are comparable to when $Z \sim N(0, 1)$, in Table 4.2.

Exponential Z, Normal e

When $Z \sim Exp(1)$, which is severly skewed, for all but one case the appropriate λ has least bias and highest coverage (Table 4.4). The one case is for a strong proxy, $\rho = 0.8$, under the assumption MAR. Here, $\lambda = 1$ has slightly less bias than $\lambda = 0$, but this difference might be negligable, and the coverage for both is nominal (96.2%, $\lambda = 0$; 94.2%, $\lambda = 1$). Also, it is not surprising that $\lambda = 1$ is relatively unbiased for this case, since missingness is related to Z, and the correlation between X and Y is strong. For this distribution, there are a few cases when the full MLE and the two-step differ in results, mainly in the coverage. The biggest difference is for $\rho = 0.8$ under MNAR. The 'correct' λ, $\lambda = \infty$, the full MLE has 88.6% and two-step has 91.0% coverage. Overall, the results of the PPM hot deck with either estimation method perform similarly to when $Z \sim N(0, 1)$ and $Z \sim Gamma(4, 0.5)$, where the largest bias and the lowest coverage occur when the proxy is weak, and missingness is MNAR ($\lambda = \infty$). Confidence interval width is also comparable among the different distributions.

Normal Z, Gamma e

Tables 4.5, 4.6 display the results when changing the distribution of e. When $e \sim Gamma(2, 0.5)$, again there is no significant difference between results using the full MLE method and the two-step method (Table 4.5). When data are MAR, $\lambda = 0$ has the least bias of all λ's. The bias decreases from -0.005%, -0.004% to 0.003% as the correlation between Y and the proxy decreases. The coverage is 94% for all correlations. For $\rho = 0.8$, $\lambda = 1$ has higher coverage than $\lambda = 0$ ($\approx$ 95%), but is slightly more biased (0.006). Under MNAR, $\lambda = \infty$ has the least bias and best coverage of all λ's with all ρ's. The coverage is less than nominal for all; with the highest at 90.6% coverage for both $\rho = 0.8$ and $\rho = 0.5$. For weak proxy, $\lambda = \infty$ is biased and under coveraged, but the results compare to those when e is Normal (Table 4.2).

This was a setting that was not examined in Andridge (2009) [38]. For MAR, $\lambda = 0$ is unbiased and achieves nominal coverage. Similar for MNAR and $\lambda = \infty$. The confidence interval does get very wide when the proxy is weak, with the interval width of 0.22 under MAR and MNAR, compared to an average width of 0.1 when the proxy is strong.

Normal Z, Exponential e

For the more severe skewness, $e \sim Exp(1)$, results in Table 4.6 are similar to those when $e \sim Gamma(4, 0.5)$. A few situations have less bias and higher coverage than the other two distributions examined, specifically for the weakest proxy. For $\rho = 0.2$, under MAR, $\lambda = 0$ only has -0.001% bias and 95.4% coverage. Under MNAR (again with $\rho = 0.2$), $\lambda = \infty$ has -3.2% bias and 81% coverage. Thus both are slightly better than when e is less skewed.

PPM Hot Deck versus Parametric PPM

When the distributions of the proxy and the error term e are both normal, the parametric PPM has less bias and better coverage than the PPM hot deck for MAR with $\lambda = 0$ and MNAR with $\lambda = \infty$, for all correlations.

When the distribution of Z (and thus the proxy) is skewed, the results are comparable between methods under MAR with $\lambda = 0$. Under MNAR with $\lambda = \infty$, the parametric PPM has better coverage, due in part to a wider confidence interval, especially for the weakest proxy (0.17 versus 0.12 for the hot deck).

When the distribution of e is skewed, again the results are comparable for MAR with $\lambda = 0$ and also under MNAR with $\lambda = \infty$ when the proxy is strong. As the proxy strength decreases, the hot deck is more biased and the confidence interval width does not become large, so the coverage suffers. For $\rho = 0.2$, the confidence interval width is 0.2 for the parameteric PPM, but approximately 0.13 for the hot deck (for both gamma and exponential).

4.7 Data Applications

We apply PPM hot deck imputation to data from the Surveillance, Epidemiology, and End Results (SEER) program, from the National Cancer Institute, for imputation of estrogen receptor (ER) status for breast cancer subjects. Trends in breast cancer incidence and survival are of interest, and we estimate percent of ER positive (ER+) status each year from 1992 to 2010. In a second application, the PPM hot deck is used to impute a missing binary outcome, but instead of estimating the overall mean, we estimate the coefficients from a logistic regression analysis. The data are from Troxel [22].

4.7.1 SEER

At the time of analysis, the SEER 1992-2010 database contained information on 487,515 breast cancer subjects with 14.8% missing estrogen receptor (ER) status, as well as with missingness on other covariates. To implement the PPM hot deck, we used only cases that were complete on everything except ER status. This created a data set of $n = 348,465$ individuals with 9% missing ER status. Following Howlader (2012) [45], ER status was recategorized into 3 categories: (1) ER+ (Positive or Borderline Positive) (2) ER- (Negative) and (3) Missing ER status (Test not done, test done but results missing, or unknown).

The proxy was created using a probit regression of ER status on the following categorical variables: SEER registry (18 states), year of diagnosis (1992-2010), Hispanic (Spanish, Hispanic, Latino versus Non-Spanish-Hispanic-Latino), age (categorical), race (White, Black, American Indian/Alaskan Native, Asian or Pacific Islander), histology (ductal, lobular, mixed, other), tumor size $((0,1],(1,2],(2,3],(3,4],(5,+])$cm, grade (well differentiated, moderately differentiated, poorly differentiated, undifferentiated), nodes (positive, negative), mets (yes, no), surgery (yes, no).

The proxy pattern-mixture model hot deck imputed missing ER status for each of three missingness assumptions, represented by $\lambda = \{0, 1, \infty\}$. Estimates of model parameters were found using both the full ML estimation and the two-step method. Ten multiply imputed data sets were created for each λ and estimation method combination. They were then subset by year of diagnosis, and for each year, the mean ER+ was computed and estimates combined using standard MI combining rules.

There were only slight differences between the estimates using the full MLE method and the two-step method. While there were some differences between methods for each parameter and λ value, they did not greatly impact the hot deck method. We suspect that it is due to everyone being shifted the same amount when using possibly biased estimates.

As just stated, the differences between estimates of the mean ER+ status were negligible using hot deck, so we present only results for the two-step method. Figure 4.3 shows the trend of mean ER+ status from 1992 to 2010 for the complete cases and the three values of λ. Numerical values of percent missingness (and therefore percent imputed) are along the x-axis. Overall, the mean ER+ status increases from ≈ 0.74 to ≈ 0.83 in 2010, with occasional decreases, as observed from 2002 to 2006. For each year, the mean ER+ status decreases as λ increases from 0 to ∞ suggesting that if the missingness is extremely nonignorable, the true mean would be less than that under an MAR model. Due to the large sample size and the small percent missing, the difference between these estimates is not large; the largest difference between the estimate for $\lambda = 0$ and the estimate for $\lambda = \infty$ was 0.0074 in 1992. These estimates were 0.7467 for $\lambda = 0$, and 0.739 for $\lambda = \infty$. This compares to the nearly identical estimates in 2010: 0.832 and 0.831 for $\lambda = 0$ and $\lambda = \infty$, respectively. We can see that with the higher percentage missingness (the earlier diagnosis years), the difference between the λ estimates is larger, whereas when the percent missing decreases to 2%, as in 2010, the difference is negligible. Starting in 2004, percent missing drops to below 5.1%. As noted in [45], this drop was due to the Collaborative Staging System that was proposed in 2004 and required ER status to be recorded. There was also not a large difference between the estimates from a complete case analysis and the estimates assuming MAR ($\lambda = 0$), with the CC estimates higher than the MAR estimate in some years, yet smaller in others. The largest difference was in 1999, when the CC estimate was 0.7776 and the MAR estimate was 0.7738.

4.7.2 Comparing to ISNI

Troxel et al. (2004) apply their index of sensitivity to nonignorability (ISNI) to a sexual behavior survey data set originally analyzed by Raab and Donnelly (1999) [22, 24]. The simplified analysis consisted of using two categorical covariates, gender and faculty (medical or non-medical) to predict sexual behavior (yes/no outcome) of students at the University of Edinburgh. The response rate ranged from 59% (non-medical males) to 77.1% (medical females). We applied our PPM hot deck method to these data, using 20 MI data sets for each value of $\lambda = \{0, 1, \infty\}$.

Using the original dataset, prior to bootstrapping respondents, we can calculate the proxy X and the maximum likelihood estimates using both the full method and the two-step method. There were 37.6% subjects missing the outcome. The mean of X for the respondents was 0.624, whereas the mean of X for the nonrespondents was 0.647, suggesting that nonrespondents were more likely to respond as a 'Yes', which is what the authors believed. The estimated Pearson correlation of Y and X, was low, at 0.11. Using the full and two-step methods produced similar estimates, 0.134. This indicates the proxy is rather weak with a relatively high (for hot deck) 38% nonresponse rate.

Figure 4.4 is a visualization of the use of adjustment cells with only two dichotomous predictor variables. The $\hat{u}'s$ were calculated on the observed data (no bootstrapping). The top row shows the adjustment cells formed by the respondents, with each cell containing the number of respondents in that cell and also the percentage of respondents in the cell with $Y = 1$. The scale on the x-axis has been edited for illustration purposes. Since we only have Faculty and Gender for our predictors along with their interaction term, we have only four covariate patterns, leading to four values of $\hat{u}$. Since there are only four covariate patterns, we use those four values of $\hat{u}$ of the respondents for the adjustment cells, and everyone in the cell has the same value of $\hat{u}$, so the mean is the value of $\hat{u}$. For nonrespondents, the values of $\hat{u}$ increase on average as the value of λ increases. To place nonrespondents in an adjustment cell, we minimize the distance between the value of $\hat{u}$ and the mean $\hat{u}$ in each cell. The arrows on the figure represent which cell the nonrespondents will be placed in. For $\lambda = 0$, the $\hat{u}^{(1)}$ match those of the respondents, and they are placed in the corresponding cells. For $\lambda = 1$, all nonrespondents are closest to either the 3rd or 4th adjustment cell. Both of these cells contain 75% 1's, so taking an SRS within the cells means that on average, we will be imputing approximately 75% 1's. For $\lambda = \infty$, the $\hat{u}'s$ become very spread out, with the largest value at 21.22. If we were using the distance measure of Siddique and Belin, this might be a case when the distances are so large and this would turn into a SRS of all the respondents when imputing those large nonrespondents. Using Adjustment Cells, these nonrespondents get placed in the adjustment cell corresponding to the largest value of $\hat{u}$, which also has the largest percentage of 1's. Note that the order of the covariate patterns in relationship to the $\hat{u}$'s change between the MAR and the MNAR models.

Troxel et al. were interested in comparing the coefficients of the ignorable model with the ISNI, to the coefficients of the non-ignorable model fit in [24]. Figure 4.5 is the table of coefficient estimates from [22]. The top line is the ignorable model that

Troxel fit; the nonignorable model is one of the models in Raab (1999) [24]. They compare each coefficient from the ignorable model added to ISNI to the nonignorable model. They state that if people who failed to respond are more likely to have a positive outcome, then to go from the ignorable model to the nonignorable model, we would subtract the ISNI (e.g., for the faculty coefficient: -0.73 - 0.17 = -0.90). If they were more likely to have responded 'no', we would add the ISNI, and the nonignorable model estimate would be -0.56. They also conclude that the intercept term and the faculty term are sensitive to nonignorability, since the ratio of the ISNI to the SE is large (7.39, 1.16, respectively). On the other hand, gender and the interaction are not sensitive. It seems unintuitive for a main effect to be sensitive to nonignorability but not the interaction term.

Table 4.7 shows the MI estimates for each value of λ for the PPM hot deck with 20 imputations. When $\lambda = 0$, our results are very similar to that of the ignorable model of Troxel, with comparable standard errors. When changing from $\lambda = 0$ to $\lambda \neq 0$, we can see that the gender coefficient does not significantly change, and neither does the intercept. The Faculty coefficient is the only coefficient that really changes among the models. This is a slightly different conclusion to Troxel, which stated that both the intercept and the faculty term were sensitive to nonignorability. The coefficient of faculty does change, from -0.70 to -0.57 ($\lambda = 1$) and -0.52 ($\lambda = \infty$). This is actually comparable to the 'other' coefficient option given by Troxel. However, our interaction coefficient also decreases, whereas theirs increases. They seem to be hitting the boundary of filling everyone in as a 'yes' response, whereas our model does not hit this boundary.

4.8 Tables and Figures

Figure 4.1: For nonrespondents $i = 30$ (left figures) and $i = 40$ (right figures), plots of $\hat{u}_i^{(0)}$'s for all respondents and $\hat{u}_i^{(1)}$ for the indicated nonrespondent (red line), versus the absolute distance (upper figures), Siddique and Belin's distance measure with $k = 3$ (middle figures), and the selection probability (lower figures).

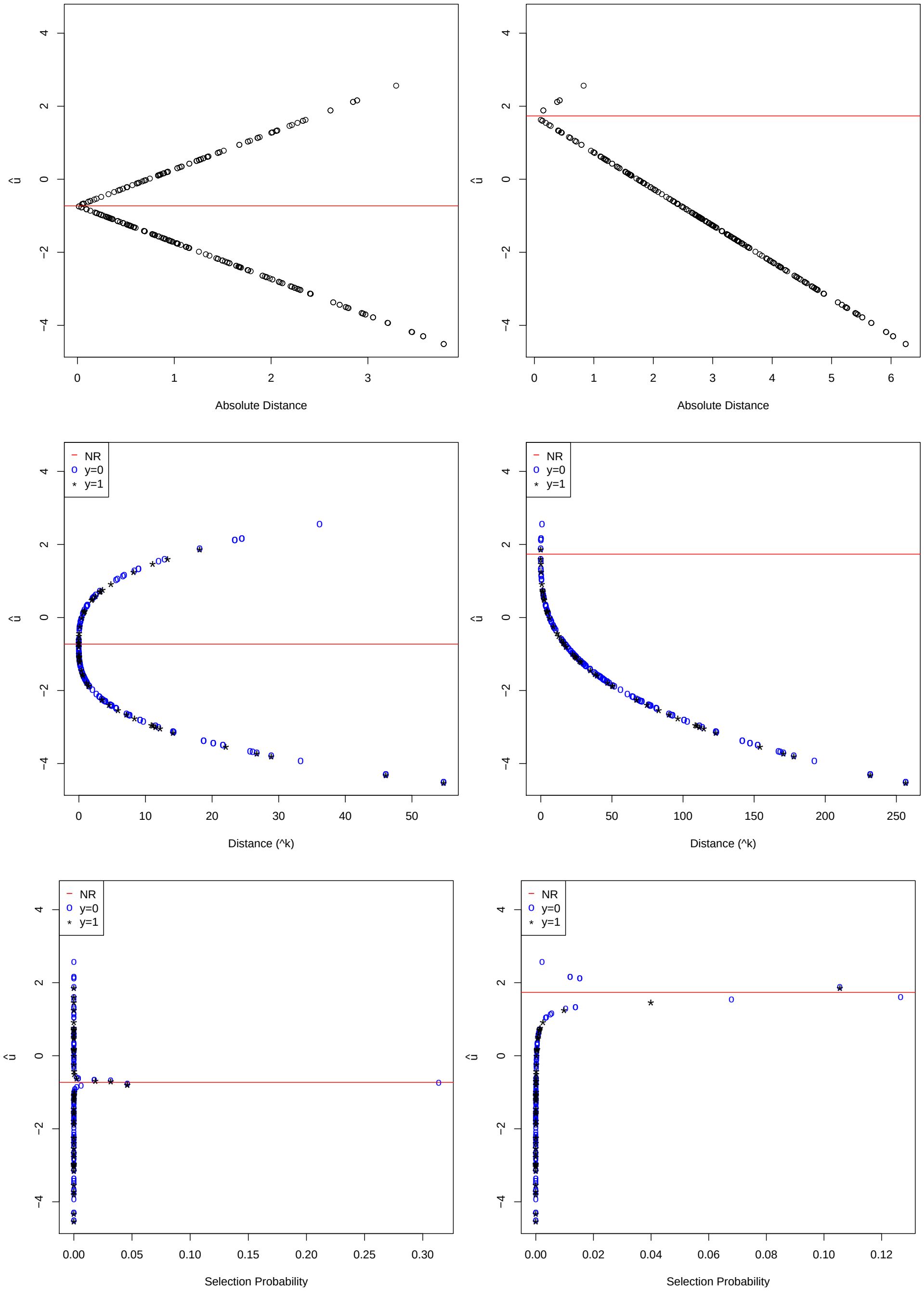
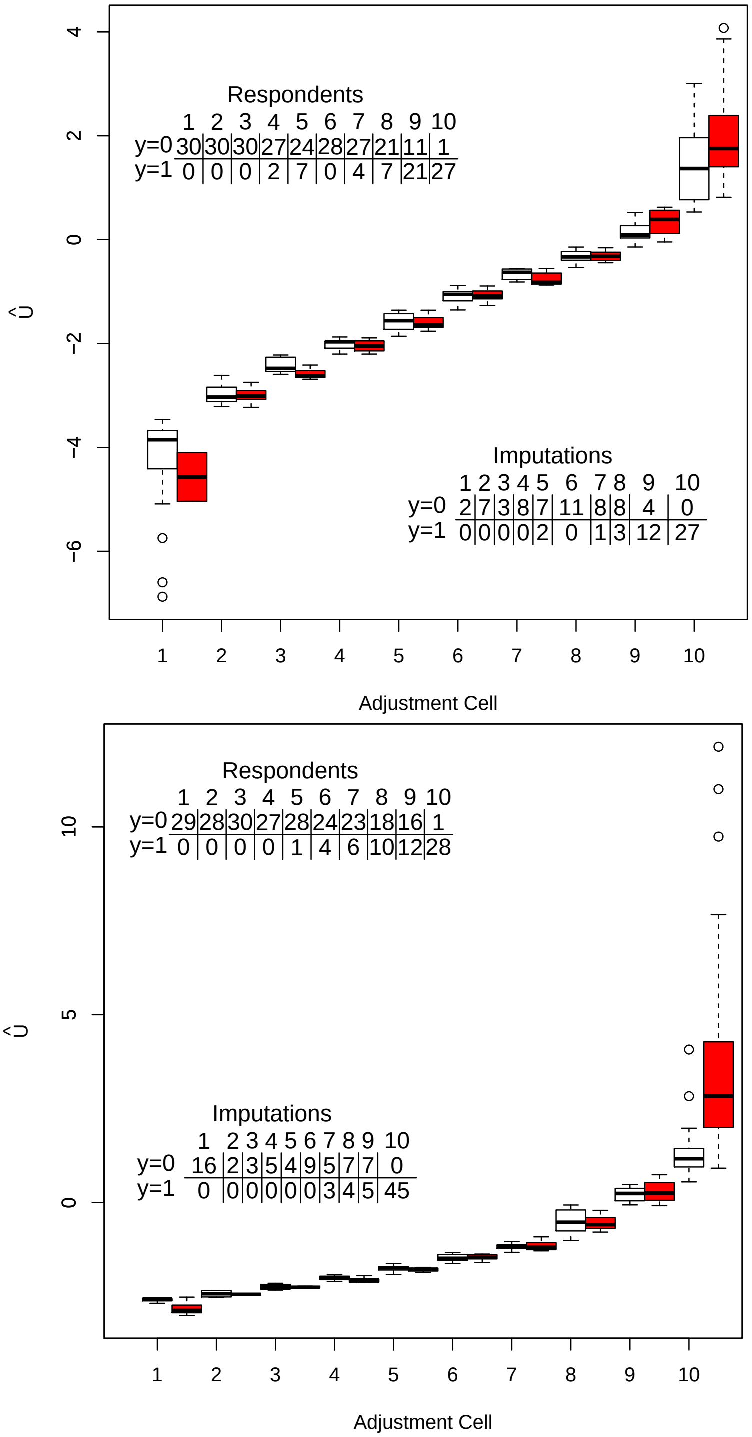


Figure 4.2: Distribution of $\hat{U}$'s versus adjustment cells. Data for top figure was generated with $\rho = 0.8$ with a normal proxy, normal error, and an MAR mechanism; $\hat{u}_i$'s created with $\lambda = 0$. Lower figure has an exponentially distributed proxy with $\rho = 0.8$, MNAR mechanism, and $\lambda = \infty$ was used to create $\hat{u}_i$'s.



Respondents

	1	2	3	4	5	6	7	8	9	10
y=0	30	30	30	27	24	28	27	21	11	1
y=1	0	0	0	2	7	0	4	7	21	27

Imputations

	1	2	3	4	5	6	7	8	9	10
y=0	2	7	3	8	7	11	8	8	4	0
y=1	0	0	0	0	2	0	1	3	12	27

Respondents

	1	2	3	4	5	6	7	8	9	10
y=0	29	28	30	27	28	24	23	18	16	1
y=1	0	0	0	0	1	4	6	10	12	28

Imputations

	1	2	3	4	5	6	7	8	9	10
y=0	16	2	3	5	4	9	5	7	7	0
y=1	0	0	0	0	0	0	3	4	5	45

Table 4.1: Parameter values used in the simulations. Each scenario assumes the other variable is Standard Normal; i.e., the rows with e ∼G(4,0.5) assume Z ∼N(0,1).

				MAR	MNAR
Z	ρ	α_0	α_1	γ_0	γ_0
$Z \sim$ N(0,1)	0.8	-0.874	1.333	-1.099	-0.662
	0.5	-6.055	0.577	-1.099	1.929
	0.2	-0.535	0.204	-1.099	-0.831
$Z \sim$ G(4,0.5)	0.8	-0.740	1.333	-1.099	-0.729
	0.5	-0.587	0.577	-1.099	-0.805
	0.2	-0.535	0.204	-1.099	-0.831
$Z \sim$ Exp(1)	0.8	-0.600	1.333	-1.099	-0.799
	0.5	-0.560	0.577	-1.099	-0.819
	0.2	-0.530	0.204	-1.099	-0.834

				MAR	MNAR
e	ρ	α_0	α_1	γ_0	γ_0
$e \sim$ N(0,1)	0.8	-0.874	1.333	-1.099	-0.662
	0.5	-6.055	0.577	-1.099	1.929
	0.2	-0.535	0.204	-1.099	-0.831
$e \sim$ G(4,0.5)	0.8	-0.820	1.333	-1.099	-0.689
	0.5	-0.490	0.577	-1.099	-0.854
	0.2	-0.380	0.204	-1.099	-0.909
$e \sim$ Exp(1)	0.8	-0.760	1.333	-1.099	-0.719
	0.5	-0.360	0.577	-1.099	-0.919
	0.2	-0.210	0.204	-1.099	-0.994

Table 4.2: Empirical bias, 95% interval coverage and average interval length for six artificial populations with $n = 400$, 25% missingness, covariate distribution $Z \sim N(0, 1)$ and error distribution, $e \sim N(0, 1)$. HD FMLE: Hot deck with full maximum likelihood; HD 2-Step: Hot deck with modified maximum likelihood; Para MI: parametric PPMA with multiple imputation. 10 imputed data sets were created for all methods. Results over 500 replicates.

			MAR			MNAR		
			Empirical	Coverage	CI	Empirical	Coverage	CI
			Bias	(%)	Width	Bias	(%)	Width
0.8	0	HD FMLE	-0.003	94.6	0.102	-0.038	67.6	0.102
		HD 2-Step	-0.003	94.6	0.102	-0.038	67.8	0.102
		Para MI	0.000	95.2	0.102	-0.033	74.0	0.101
	1	HD FMLE	0.006	94.0	0.103	-0.026	82.8	0.104
		HD 2-Step	0.006	94.0	0.103	-0.026	83.4	0.104
		Para MI	0.011	94.0	0.104	-0.019	88.2	0.103
	∞	HD FMLE	0.018	91.0	0.106	-0.011	90.8	0.107
		HD 2-Step	0.018	91.2	0.106	-0.011	91.0	0.107
		Para MI	0.024	88.8	0.108	0.000	94.6	0.112
0.5	0	HD FMLE	-0.003	95.2	0.107	-0.045	59.0	0.101
		HD 2-Step	-0.003	95.0	0.107	-0.045	59.2	0.101
		Para MI	0.000	95.4	0.107	-0.043	62.0	0.101
	1	HD FMLE	0.017	93.6	0.112	-0.034	72.8	0.105
		HD 2-Step	0.017	93.6	0.112	-0.034	73.0	0.105
		Para MI	0.024	89.8	0.113	-0.030	78.8	0.108
	∞	HD FMLE	0.043	76.2	0.130	-0.016	88.8	0.119
		HD 2-Step	0.043	76.2	0.130	-0.016	88.8	0.119
		Para MI	0.076	50.2	0.155	-0.003	94.0	0.158
0.2	0	HD FMLE	-0.001	94.8	0.107	-0.046	57.4	0.101
		HD 2-Step	-0.001	94.8	0.107	-0.046	57.4	0.101
		Para MI	0.000	95.4	0.108	-0.045	59.2	0.101
	1	HD FMLE	0.016	92.6	0.134	-0.041	68.8	0.112
		HD 2-Step	0.016	92.8	0.134	-0.041	68.8	0.112
		Para MI	0.037	79.0	0.135	-0.039	71.8	0.111
	∞	HD FMLE	0.021	92.0	0.146	-0.035	77.4	0.126
		HD 2-Step	0.021	92.2	0.146	-0.035	77.4	0.126
		Para MI	0.127	26.8	0.199	0.005	97.2	0.208

Table 4.3: Empirical bias, 95% interval coverage and average interval length for six artificial populations with $n = 400$, 25% missingness, covariate distribution $Z \sim Gamma(4, 0.5)$ and error distribution, $e \sim N(0, 1)$. HD FMLE: Hot deck with full maximum likelihood; HD 2-Step: Hot deck with modified maximum likelihood; Para MI: parametric PPMA with multiple imputation. 10 imputed data sets were created for all methods. Results over 500 replicates.

			MAR			MNAR		
			Empirical	Coverage	CI	Empirical	Coverage	CI
			Bias	(%)	Width	Bias	(%)	Width
0.8	0	HD FMLE	-0.004	94.4	0.102	-0.037	69.4	0.101
		HD 2-Step	-0.004	94.4	0.102	-0.037	69.8	0.101
		Para MI	0.001	93.6	0.100	-0.029	80.2	0.100
	1	HD FMLE	0.005	93.2	0.101	-0.024	84.8	0.102
		HD 2-Step	0.005	93.0	0.102	-0.024	85.4	0.102
		Para MI	0.012	91.6	0.101	-0.013	91.6	0.100
	∞	HD FMLE	0.015	91.0	0.103	-0.011	93.2	0.105
		HD 2-Step	0.016	90.8	0.103	-0.010	93.4	0.105
		Para MI	0.025	86.8	0.103	0.004	96.0	0.103
0.5	0	HD FMLE	-0.006	95.0	0.106	-0.045	56.2	0.101
		HD 2-Step	-0.006	94.8	0.106	-0.045	56.4	0.101
		Para MI	-0.001	95.2	0.106	-0.041	62.6	0.101
	1	HD FMLE	0.011	94.6	0.110	-0.034	72.4	0.104
		HD 2-Step	0.011	94.6	0.110	-0.034	72.8	0.104
		Para MI	0.027	85.6	0.109	-0.025	81.8	0.105
	∞	HD FMLE	0.030	84.2	0.120	-0.019	88.4	0.112
		HD 2-Step	0.030	84.8	0.120	-0.019	88.6	0.112
		Para MI	0.064	42.6	0.121	0.003	95.8	0.130
0.2	0	HD FMLE	-0.004	94.4	0.108	-0.046	56.0	0.101
		HD 2-Step	-0.004	94.0	0.108	-0.046	56.2	0.101
		Para MI	-0.002	94.4	0.108	-0.045	58.2	0.102
	1	HD FMLE	0.008	93.8	0.123	-0.042	67.4	0.109
		HD 2-Step	0.008	93.8	0.123	-0.042	67.2	0.109
		Para MI	0.038	76.8	0.130	-0.036	71.2	0.110
	∞	HD FMLE	0.016	92.2	0.139	-0.035	77.0	0.123
		HD 2-Step	0.016	92.2	0.139	-0.035	76.8	0.123
		Para MI	0.096	28.2	0.162	0.005	97.0	0.189

Table 4.4: Empirical bias, 95% interval coverage and average interval length for six artificial populations with $n = 400$, 25% missingness, covariate distribution $Z \sim Exp(1)$ and error distribution, $e \sim N(0, 1)$. HD FMLE: Hot deck with full maximum likelihood; HD 2-Step: Hot deck with modified maximum likelihood; Para MI: parametric PPMA with multiple imputation. 10 imputed data sets were created for all methods. Results over 500 replicates.

			MAR			MNAR		
			Empirical Bias	Coverage (%)	CI Width	Empirical Bias	Coverage (%)	CI Width
0.8	0	HD FMLE	-0.006	96.2	0.101	-0.038	64.2	0.100
		HD 2-Step	-0.006	96.4	0.101	-0.038	64.2	0.101
		Para MI	0.001	96.2	0.100	-0.029	78.2	0.098
	1	HD FMLE	0.003	94.2	0.101	-0.027	81.0	0.102
		HD 2-Step	0.004	94.2	0.101	-0.027	82.4	0.102
		Para MI	0.011	95.6	0.100	-0.014	89.8	0.098
	∞	HD FMLE	0.011	95.0	0.102	-0.016	88.6	0.103
		HD 2-Step	0.013	94.6	0.102	-0.015	91.0	0.103
		Para MI	0.023	90.0	0.102	0.003	95.4	0.101
0.5	0	HD FMLE	-0.006	94.2	0.106	-0.048	54.8	0.100
		HD 2-Step	-0.006	94.4	0.106	-0.048	55.0	0.100
		Para MI	0.001	96.0	0.106	-0.043	62.4	0.100
	1	HD FMLE	0.007	95.2	0.109	-0.039	69.0	0.103
		HD 2-Step	0.007	95.4	0.109	-0.039	69.8	0.103
		Para MI	0.026	86.0	0.105	-0.026	80.0	0.102
	∞	HD FMLE	0.020	92.2	0.115	-0.028	82.4	0.109
		HD 2-Step	0.020	91.8	0.115	-0.028	82.8	0.109
		Para MI	0.056	48.0	0.110	0.000	95.8	0.116
0.2	0	HD FMLE	-0.003	93.6	0.107	-0.046	55.6	0.102
		HD 2-Step	-0.003	93.6	0.107	-0.046	55.6	0.102
		Para MI	0.001	93.0	0.109	-0.045	59.2	0.102
	1	HD FMLE	0.007	93.0	0.119	-0.042	65.4	0.109
		HD 2-Step	0.007	92.8	0.119	-0.042	65.2	0.109
		Para MI	0.039	74.2	0.126	-0.035	72.8	0.110
	∞	HD FMLE	0.012	92.2	0.129	-0.035	79.6	0.121
		HD 2-Step	0.012	92.2	0.129	-0.035	79.0	0.121
		Para MI	0.080	36.8	0.143	0.002	95.0	0.171

Table 4.5: Empirical bias, 95% interval coverage and average interval length for six artificial populations with $n = 400$, 25% missingness, covariate distribution $Z \sim N(0,1)$ and error distribution, $e \sim Gamma(4, 0.5)$. HD FMLE: Hot deck with full maximum likelihood; HD 2-Step: Hot deck with modified maximum likelihood; Para MI: parametric PPMA with multiple imputation. 10 imputed data sets were created for all methods. Results over 500 replicates.

			MAR			MNAR		
			Empirical Bias	Coverage (%)	CI Width	Empirical Bias	Coverage (%)	CI Width
0.8	0	HD FMLE	-0.005	94.0	0.103	-0.041	64.0	0.101
		HD 2-Step	-0.005	94.0	0.102	-0.041	64.4	0.101
		Para MI	-0.001	95.0	0.102	-0.036	70.8	0.101
	1	HD FMLE	0.006	95.4	0.105	-0.028	78.2	0.102
		HD 2-Step	0.006	95.2	0.105	-0.028	78.2	0.102
		Para MI	0.011	94.4	0.105	-0.021	86.0	0.104
	∞	HD FMLE	0.021	88.8	0.108	-0.009	90.6	0.110
		HD 2-Step	0.021	88.6	0.108	-0.009	90.8	0.110
		Para MI	0.027	86.0	0.112	0.000	93.0	0.112
0.5	0	HD FMLE	-0.004	94.0	0.107	-0.046	55.6	0.101
		HD 2-Step	-0.004	94.0	0.107	-0.046	55.4	0.101
		Para MI	-0.002	93.6	0.107	-0.044	57.8	0.101
	1	HD FMLE	0.019	89.6	0.115	-0.035	74.4	0.106
		HD 2-Step	0.019	89.6	0.114	-0.035	74.0	0.106
		Para MI	0.025	87.2	0.115	-0.031	78.6	0.109
	∞	HD FMLE	0.047	71.2	0.137	-0.014	90.6	0.126
		HD 2-Step	0.047	71.2	0.137	-0.014	90.6	0.126
		Para MI	0.088	41.4	0.162	0.002	94.8	0.171
0.2	0	HD FMLE	0.003	93.8	0.108	-0.047	50.6	0.101
		HD 2-Step	0.003	93.8	0.108	-0.047	50.2	0.101
		Para MI	0.004	93.4	0.108	-0.047	51.8	0.101
	1	HD FMLE	0.020	91.4	0.135	-0.043	65.4	0.113
		HD 2-Step	0.020	91.0	0.135	-0.043	65.6	0.113
		Para MI	0.042	75.0	0.139	-0.041	66.0	0.112
	∞	HD FMLE	0.024	93.2	0.147	-0.038	74.8	0.128
		HD 2-Step	0.024	93.2	0.147	-0.038	75.0	0.128
		Para MI	0.130	30.0	0.220	0.004	95.4	0.216

Table 4.6: Empirical bias, 95% interval coverage and average interval length for six artificial populations with $n = 400$, 25% missingness, covariate distribution $Z \sim N(0, 1)$ and error distribution, $e \sim Exp(1)$. HD FMLE: Hot deck with full maximum likelihood; HD 2-Step: Hot deck with modified maximum likelihood; Para MI: parametric PPMA with multiple imputation. 10 imputed data sets were created for all methods. Results over 500 replicates.

			MAR			MNAR		
			Empirical Bias	Coverage (%)	CI Width	Empirical Bias	Coverage (%)	CI Width
0.8	0	HD FMLE	-0.004	95.2	0.102	-0.039	64.6	0.100
		HD 2-Step	-0.004	95.0	0.102	-0.039	64.6	0.100
		Para MI	0.001	96.2	0.103	-0.034	73.6	0.102
	1	HD FMLE	0.008	95.2	0.103	-0.025	80.6	0.102
		HD 2-Step	0.008	95.4	0.103	-0.025	81.0	0.102
		Para MI	0.013	94.2	0.105	-0.019	87.8	0.103
	∞	HD FMLE	0.024	85.8	0.110	-0.007	91.2	0.110
		HD 2-Step	0.024	85.2	0.110	-0.006	91.4	0.110
		Para MI	0.029	83.4	0.111	-0.001	93.2	0.110
0.5	0	HD FMLE	-0.004	93.4	0.107	-0.046	59.6	0.101
		HD 2-Step	-0.004	93.4	0.107	-0.046	59.8	0.101
		Para MI	-0.001	95.6	0.108	-0.042	64.6	0.101
	1	HD FMLE	0.020	91.0	0.115	-0.034	76.4	0.105
		HD 2-Step	0.021	90.6	0.115	-0.034	76.2	0.105
		Para MI	0.026	86.2	0.115	-0.030	80.6	0.107
	∞	HD FMLE	0.054	68.0	0.138	-0.011	93.0	0.126
		HD 2-Step	0.054	67.8	0.139	-0.011	93.2	0.126
		Para MI	0.087	40.6	0.159	-0.001	96.2	0.162
0.2	0	HD FMLE	-0.001	95.4	0.108	-0.043	61.4	0.101
		HD 2-Step	-0.001	95.4	0.108	-0.043	61.0	0.101
		Para MI	0.000	96.4	0.108	-0.043	61.0	0.101
	1	HD FMLE	0.018	92.6	0.136	-0.039	73.0	0.113
		HD 2-Step	0.018	92.8	0.136	-0.039	73.4	0.113
		Para MI	0.036	83.4	0.139	-0.037	74.4	0.111
	∞	HD FMLE	0.023	94.4	0.149	-0.032	81.0	0.127
		HD 2-Step	0.023	94.0	0.149	-0.032	81.0	0.127
		Para MI	0.125	31.4	0.220	0.009	95.6	0.211

Figure 4.3: Estimates of mean ER+ status for each year of diagnosis based on SEER data, using a complete-case analysis (CC) and the PPM hot deck with $\lambda = \{0, 1, \infty\}$ and 10 imputed data sets. Numbers along inside the x-axis are percent missing for each year.

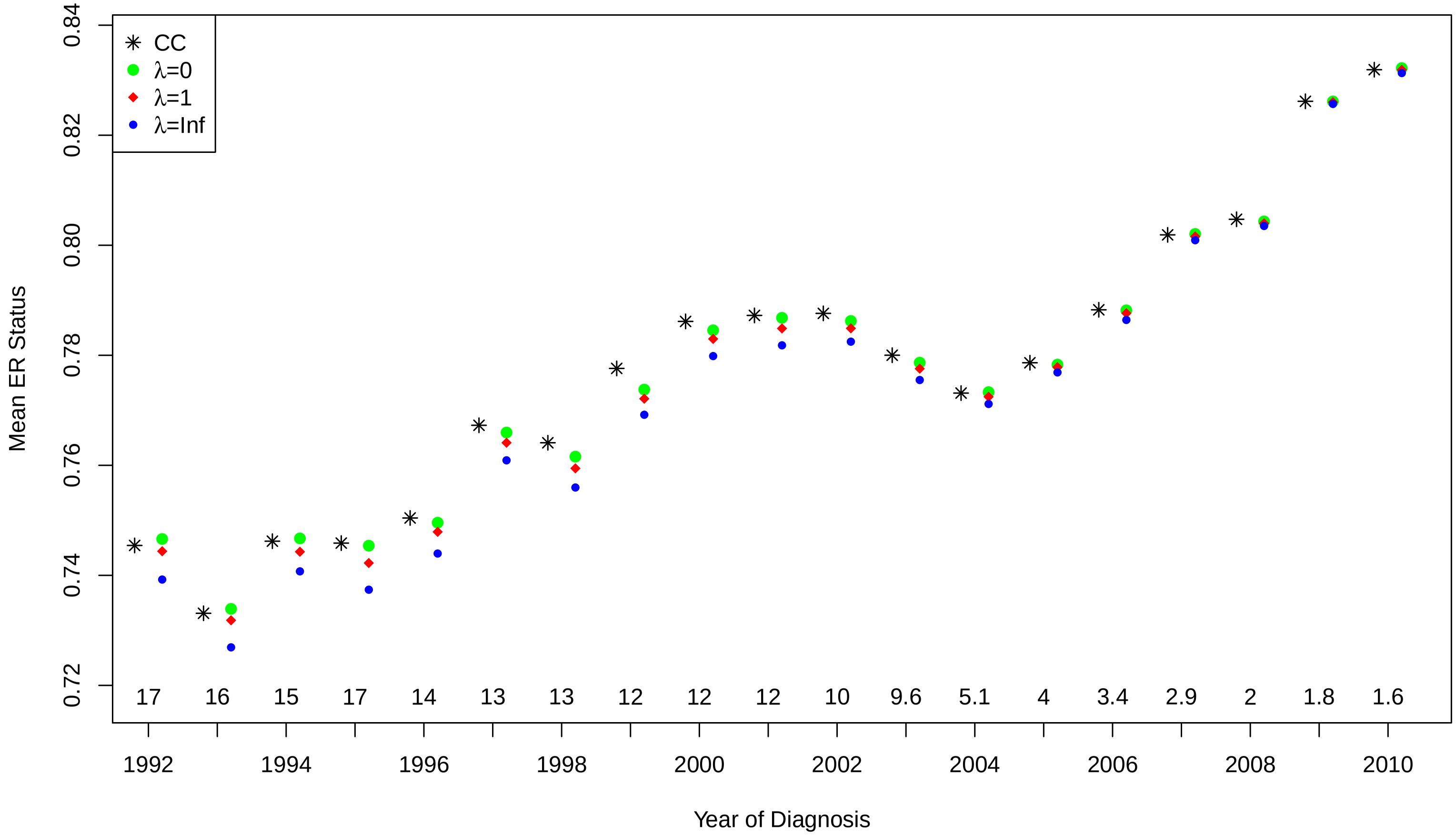


Figure 4.4: Predicted Means ($\hat{u}$'s) for the application to the Troxel data. Each box represents a covariate pattern and thus an adjustment cell. The arrows indicate the adjustment cell the nonrespondents are closest to, for each value of λ.

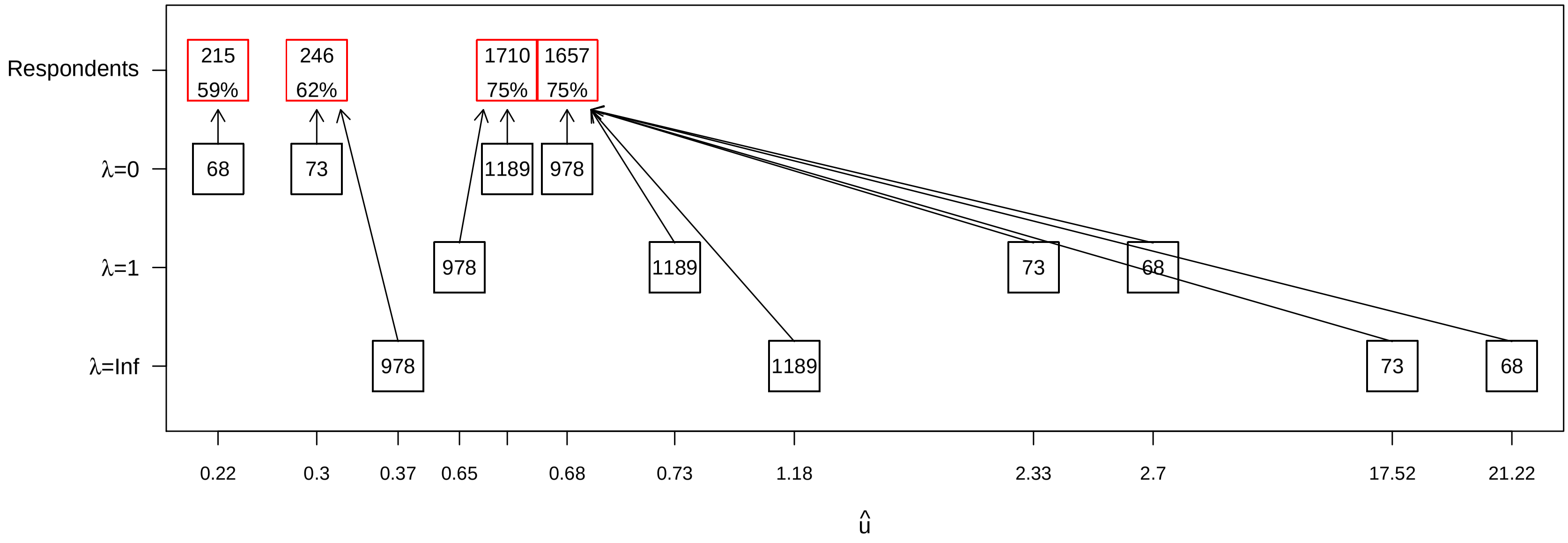


Figure 4.5: Table 3 from Troxel [22]. ML estimates of ignorable and nonignorable models and ISNI analysis.

	Intercept	Gender	Faculty	Faculty*Gender
Ignorable Model	1.08	0.03	−0.73	0.10
SE of Coefficients	0.06	0.08	0.15	0.21
Nonignorable Model	1.74	−0.05	−0.95	0.14
ISNI	−0.41	0.04	0.17	−0.03
|ISNI/SE|	7.39	0.50	1.16	0.16

Table 4.7: Results from using the PPM hot deck to estimate the model coefficients from the Troxel data, for each value of $\lambda = \{0, 1, \infty\}$.

$\lambda = 0$	Mean	SD	lb	ub	df	FMI
Intercept	1.088	0.051	0.987	1.189	204.820	0.305
Faculty	-0.719	0.143	-1.000	-0.438	488.914	0.197
Gender	0.031	0.072	-0.112	0.173	280.447	0.260
Faculty * Gender	0.087	0.199	-0.303	0.477	511.308	0.193

$\lambda = 1$	Mean	SD	lb	ub	df	FMI
Intercept	1.086	0.054	0.979	1.193	136.069	0.374
Faculty	-0.576	0.150	-0.871	-0.282	317.519	0.245
Gender	0.019	0.077	-0.134	0.171	149.834	0.356
Faculty * Gender	0.066	0.207	-0.342	0.474	325.070	0.242

$\lambda = \text{inf}$	Mean	SD	lb	ub	df	FMI
Intercept	1.028	0.112	0.797	1.258	25.820	0.858
Faculty	-0.520	0.174	-0.865	-0.174	96.756	0.443
Gender	0.021	0.155	-0.297	0.338	26.724	0.843
Faculty * Gender	0.076	0.236	-0.391	0.542	111.437	0.413

Chapter 5: Discussions and Conclusions

We have developed a new hot deck imputation procedure that does not assume ignorability of the nonresponse. We allow for the examination of the impact different types of missingness may have on estimates through a sensitivity parameter. The extension of hot deck to nonignorability has not been well-researched, with only one known method, the nonignorable hot deck of Siddique and Belin [29]. Our proposed method is first developed for imputation of a partially observed continuous outcome variable in Chapter 3, and is extended to a binary outcome in Chapter 4. In the continuous case, we combine elements of the nonignorable hot deck of Siddique and Belin with those of the parametric proxy pattern-mixture model of Andridge and Little [35]. The PPM incorporates an intuitive sensitivity analysis in the creation of the predicted means used for finding matches. For a binary outcome, donor selection is modified through the use of adjustment cells rather than the use of a distance metric.

5.1 Continuous Outcome

Chapter 3 introduced the proposed proxy pattern-mixture model (PPM) hot deck. The key distinction between the proposed PPM hot deck and the Siddique and Belin SB hot deck is where the correction for nonignorability occurs. The SB hot deck corrects for the nonignorability in the ABB step of the MI process, drawing respondents with probability to an assumed function of Y, such as proportional to Y or to Y^2, creating different "shaped" ABBs. Standard linear regression is used to estimate predicted means for calculating distances and probabilities of donor selection. Our proposed method adjusts for the nonignorability in the creation of the predicted means. An ABB is still drawn, but with equal probability, to ensure the MI procedure is proper. The proxy pattern-mixture model of Andridge and Little is incorporated into the predicted mean creation. Finally, Siddique and Belin's distance measure is used on these predicted means to find the matches between the nonrespondents and respondents for imputation.

In addition to the advantages of a hot deck procedure in general, we add the ability to study the impact of various missingness assumptions on estimates. We also show in the application to OMAS that survey design weights can easily be incorporated into the analysis. In our simulations and application we estimate means, but we could also use this method for estimation of regression coefficients, and the ability of the hot deck to preserve correlations is important for this.

We also propose two new donor quality metrics that can aid in diagnosing situations where the PPM hot deck is unable to find close matches and imputations may be suspect. The mean minimum distance (MMD) gives the average distance from donor to donee (in standard deviations) and can identify cases where there are no close matches. The mean variance of selection probabilities (MVSP) provides a summary based on the selection probabilities instead of the distances, and flags situations where the separation between donors and donees is so large that the imputation effectively turns into a simple random sampling from the respondents.

The proposed method is not without limitations. Similar to the parametric proxy pattern-mixture hot deck, λ is difficult to interpret beyond just MAR versus MNAR – by combing all the covariates into a univariate proxy variable we lose the connection between specific covariates and the probability of missingness. Another limitation is the inability of the PPM hot deck to extrapolate; only observed values can be imputed. However, this is a limitation of hot deck procedures in general, and is not unique to this proposed method. This becomes most important when we assume missingness is not at random (e.g., $\lambda = \infty$) and with a weak proxy. The predicted means for missing values might be quite different than the predicted means for respondents, such that there are no "close" donors available for imputation. However, our proposed donor quality metrics can provide some guidance, identifying when there are no close matches available for imputation, suggesting that the PPM hot deck is not the best method and perhaps the parametric proxy pattern-mixture method that can extrapolate beyond the observed values should be used.

5.2 Binary Outcome

In Chapter 4 the work done in Andridge (2009) [38] for extending the parametric PPM model to a binary outcome allows the hot deck to be naturally extended to imputation of a binary variable as well. We achieve this by assuming there is a continuous latent variable with a cutpoint that determines the value of the binary variable. The proxy X is created by a probit regression, and is considered a proxy for the latent U. The normal pattern-mixture model is then assumed for the proxy X and the latent U. Because this latent variable is unobserved for all subjects, this adds an additional problem when estimating the parameters for the respondents. We implemented and compared two methods used to estimate biserial correlation: Tate's full ML estimation and Olsson's two-step estimator.

The proposed method was further modified from the continuous case in how donors are selected. We replaced the Siddique and Belin predictive mean matching with a commonly used method of using the predicted means to create adjustment cells. Nonrespondents now each have the same number of possible donors (number of respondents in each cell), and this procedure also ensures that nonrespondents with large predicted means will be matched to respondents with the largest predicted means. This eliminates the problem of a nonrespondent being "too far away" resulting in a simple random sample of the entire set of respondents. The use of adjustment cells also allows for an examination of donor quality and availability. Plots, such as in Figure 4.2 can also be examined. Here, we can view the distribution of the proportion of $Y = 1$ for each adjustment cell, hopefully observing that the proportion increases as the values of $\hat{u}^{(0)}$'s increase. We can also gather how many nonrespondents are being placed in each cell and therefore how many respondents are actually being considered for possible donation.

We also investigated the robustness to departures from normality of both the covariates and the distribution of the latent variable. This is where we had expected to see a difference in the two ML estimation methods. When performing a full ML estimation method for estimating the overall mean of Y, the full ML procedure is biased in some cases of non-normality. However, when implemented in the hot deck, there was no major difference between the performance of the methods. We suspect that it is because the biased estimates shift everyone when creating predicted means to find matches, and therefore the subjects that are 'close' using unbiased estimates are still ‘close’ when using slightly biased estimates. Regardless of the ML method, the bias and coverage of the ‘correct’ λ with each missingness assumption was similar regardless of the distribution.

The PPM hot deck was applied to two data sets for imputation of a binary variable. For the SEER data, we imputed ER+ status and estimated the mean for each diagnosis year from 1992 to 2010. We observed a difference in the estimates for the earlier years, that suggested under a nonignorable model, the proportion of subjects ER+ would be less than the proportion under an ignorable model. However, the differences were not severe, and almost negligible in the later years, due to a very small percent of missingness.

For a different illustration of the proposed method, we applied the PPM hot deck to imputation of a binary outcome, but estimated the coefficients of a logistic regression model instead of the overall mean. Using the different values of the sensitivity parameter, we were able to obtain a regression model under different missingness assumptions. By examining how each coefficient changed (or did not change), we were able to assess which coefficients were sensitive to nonignorability.

5.3 Future Work

We now address areas of future work with the proxy pattern-mixture hot deck. The first involves donor selection in the continuous imputation case. The second is an extension for the binary method to ordinal outcomes. The third furthers the application to SEER data by performing cyclical imputation.

5.3.1 Using Adjustment Cells

Implementation of the adjustment cells for the binary case removed the donor selection problems observed in the continuous case. The next step is to implement the adjustment cell method for the continuous outcome case. The use of adjustment cells in predictive mean matching is a common way to form donor pools. We propose to implement this in the PPM method for continuous Y, in place of the distance measure of Siddique and Belin, in hopes to resolve the issue of the hot deck becoming a simple random sample in some settings. We propose to follow the suggestions of forming the adjustment cells in [43] and that were implemented in Chapter 4 for the binary outcome.

From the observations from implementing in the binary case, we expect that we can resolve the problems for finding donors when the MMD is large and the mean VSP is approximately zero. Using adjustment cells would prevent the method from turning into an SRS of the entire pool of respondents. This might also be used in conjunction with the Siddique and Belin distance measure. Consider the situation where some nonrespondents have very good matches, but others have large predicted means. In this case, the distance measure could be used for the nonrespondents with quality matches, and then the adjustment cells only used for the nonrespondents with no good matches. This would at least prevent the nonrespondent with the largest $\hat{u}$ being matched with the respondent with the smallest $\hat{u}$, while not adding bias when there are quality donors available for some.

5.3.2 Ordinal Outcomes

We extended the proxy pattern-mixture hot deck from imputation of a continuous variable to that of a categorical binary variable. The next step is naturally to extend the method to handle ordinal outcomes. Both methods for estimating biserial correlation have been extended to polyserial correlation, which is what we would need to apply.

When we had the binary outcome, there was a single cutpoint of the underlying distribution of the latent variable. When we have an outcome with J categories, we will need to estimate $J-1$ cutpoints.

The relationship between the categorical outcome Y and the latent variable U is given by:

$$Y = y_j \quad \text{if} \quad \omega_{j-1} \leq U < \omega_j, \quad j = 1, 2, \ldots, J \tag{5.1}$$

We set $\omega_0 = -\infty$ and $\omega_c = +\infty$. It is also assumed that $y_{j-1} < y_j$ for $j = 2, \ldots, J$ and $\omega_{j-1} < \omega_j$ for $j = 2, \ldots, c-1$.

As discussed in [38], the probit regression for an ordinal outcome is $\Pr(Y \leq j|Z, M = 0)$ $=\Pr(U \leq \omega_j) = \Phi(\omega_j + \alpha Z)$ and we set the proxy to $X = \hat{\alpha} Z$. The pattern-mixture model is assumed for the proxy X and the latent U, but now the latent U has $J-1$ cutpoints. Without loss of generality, we take $\mu_u^{(0)} = 0$ and $\sigma_{uu}^{(0)} = 1$. To estimate the polyserial correlation coefficient $\rho^{(0)}$ and the $J-1$ cutpoints for respondents, we describe the procedure following the two-step estimation of Olsson (1982) [42]. Now, the first step of the method solves:

$$\omega_j = \Phi^{-1}(P_j) \quad j = 1, \ldots, J-1, \tag{5.2}$$

where P_j is the cumulative marginal proportions of Y. In the binary case, we had $\omega^{(0)} = \Phi^{-1}(\bar{y}_R)$.

The 'second step' is still to maximize the log likelihood using the estimates of ω_j found in the first step, using the conditional distribution of $Y|X$:

$$\Pr(Y = j|X) = \Phi\left(\frac{\omega_j^{(0)} - \rho^{(0)} Z^*}{\sqrt{1 - \rho^{(0)2}}}\right) - \Phi\left(\frac{\omega_{j-1}^{(0)} - \rho^{(0)} Z^*}{\sqrt{1 - \rho^{(0)2}}}\right) \tag{5.3}$$

where Z^* is the standardized version of $X^{(0)}$. Then the estimate of $\rho^{(0)}$ is found by maximizing

$$l = \sum_{i=1}^{N_R} \log p(y_i|x_i) \tag{5.4}$$

with respect to $\rho^{(0)}$ only.

Once these have been estimated, the formulas for $\hat{u}$'s will follow as in Section 4.3. All steps that follow will be the same, with forming adjustment cells based on $\hat{u}^{(0)}$, placing the nonrespondent to the closest cell, and taking a simple random sample within the cell for the nonrespondent. The value of Y imputed will then be one of the J possible values Y can take.

5.3.3 SEER Imputation

Another area for future work is with the application to the SEER data set. The implementation in Chapter 4 only considered subjects that were complete cases on every covariate (used in the analysis) except ER status. This decreased the sample size from 487,515 to 348,465. We can remedy this by imputing the missingness in a cyclical manner, by starting with imputation of the variable with the least missing. Imputation of other variables would be via an ignorable (MAR) method such as standard hot deck, and imputation of ER status conditional on these previously imputed covariates would use the PPM hot deck.

Bibliography

[1] D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys.* John Wiley & Sons, Inc., 1987.

[2] M. G. Kenward and G. Molenberghs, "Likelihood based frequentist inference when data are missing at random," *Statistical Science*, vol. 13, pp. 236–247, Sept. 1998.

[3] R. J. Little, "A Test of Missing Completely at Random for Multivariate Data With Missing Values," *Journal of the American Statistical Association*, vol. 83, no. 404, pp. 1198–1202, 1988.

[4] C. K. Enders, *Applied Missing Data Analysis (Methodology in the Social Sciences).* The Guilford Press, 2010.

[5] R. R. Andridge and R. J. Little, "A Review of Hot Deck Imputation for Survey Non-response," *International Statistical Review*, vol. 78, no. 1, pp. 40–64, 2010.

[6] J. L. Schafer, "Multiple imputation: a primer," *Statistical Methods in Medical Research*, vol. 8, pp. 3–15, Mar. 1999.

[7] R. J. Little and D. B. Rubin, *Statistical Analysis with Missing Data.* Wiley, 2 ed., 2002.

[8] J. Wagner, "The Fraction of Missing Information as a Tool for Monitoring the Quality of Survey Data," *Public Opinion Quarterly*, vol. 74, pp. 223–243, Mar. 2010.

[9] J. G. Ibrahim and G. Molenberghs, "Missing data methods in longitudinal studies: a review," *Test*, vol. 18, pp. 1–43, May 2009.

[10] J. J. Heckman, "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement*, vol. 5, no. 4, pp. 475–492, 1976.

[11] R. J. Little, "Pattern-Mixture Models for Multivariate Incomplete Data," *Journal of the American Statistical Association*, vol. 88, no. 421, pp. 125–134, 1993.

[12] R. J. Little, “A class of pattern-mixture models for normal incomplete data,” *Biometrika*, vol. 81, no. 3, pp. 471–483, 1994.

[13] R. J. Glynn, N. M. Laird, and D. B. Rubin, “Multiple Imputation in Mixture Models for Nonignorable Nonresponse With Follow-ups,” *Journal of the American Statistical Association*, vol. 88, no. 423, pp. 984–993, 1993.

[14] S. G. Baker and N. M. Laird, “Regression Analysis for Categorical Variables With Analysis Regression Outcome Subject to Nonignorable Nonresponse,” *Journal of the American Statistical Association*, vol. 83, no. 401, pp. 62–69, 1988.

[15] P. W. F. Smith, C. J. Skinner, and P. S. Clarke, “Allowing for non-ignorable non-response in the analysis of voting intention data,” *Journal of the Royal Statistical Society Series C*, vol. 48, no. 4, pp. 563–577, 1999.

[16] E. A. Stasny, “Modeling Nonignorable Nonresponse in Categorical Panel Data with an Example in Estimating Gross Labor-Force Flows,” *Journal of Business and Economic Statistics*, vol. 6, no. 2, pp. 207–219, 1988.

[17] E. A. Stasny, “Hierarchical Models for the Probabilities of a Survey Classification and Nonresponse : An Example From the National Crime Survey,” *Journal of the American Statistical Association*, vol. 86, no. 414, pp. 296–303, 1991.

[18] B. Nandram and J. W. Choi, “Hierarchical Bayesian Nonresponse Models for Binary Data From Small Areas With Uncertainty About Ignorability,” *Journal of the American Statistical Association*, vol. 97, no. 458, pp. 381–388, 2002.

[19] B. Nandram and J. W. Choi, “A Bayesian analysis of a proportion under non-ignorable non-response,” *Statistics in Medicine*, vol. 21, pp. 1189–212, May 2002.

[20] B. Nandram and J. W. Choi, “Bayes Empirical Bayes Estimation of a Proportion under Nonignorable Nonresponse,” in *Proceedings of the Survey Research Methods Section, ASA*, pp. 215–220, 2000.

[21] S. G. Baker, C.-W. Ko, and B. I. Graubard, “A sensitivity analysis for nonrandomly missing categorical data arising from a national health disability survey,” *Biostatistics*, vol. 4, pp. 41–56, Jan. 2003.

[22] A. B. Troxel and D. F. Heitjan, “An Index of Local Sensitivity to Nonignorability,” *Statistica Sinica*, vol. 14, pp. 1221–1237, 2004.

[23] J. B. Copas and H. G. Li, “Inference for Non-random Samples,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 59, pp. 55–95, Feb. 1997.

[24] G. M. Raab and C. A. Donnelly, “Information on sexual behaviour when some data are missing,” *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 48, no. 1, pp. 117–133, 1999.

[25] J. S. Greenlees, W. S. Reece, and K. D. Zieschang, “Imputation of Missing Values when the Probability of Response Depends on the Variable Being Imputed,” *Journal of the American Statistical Association*, vol. 77, p. 251, June 1982.

[26] J. R. Carpenter, M. G. Kenward, and I. R. White, “Sensitivity analysis after multiple imputation under missing at random: a weighting approach,” *Statistical Methods in Medical Research*, vol. 16, pp. 259–75, June 2007.

[27] N. Ressequier, R. Giorgi, and X. Paoletti, “Sensitivity Analysis When Data Are Missing Not-at-random,” *Epidemiology*, vol. 22, pp. 282–283, Mar. 2011.

[28] D. B. Rubin and N. Schenker, “Multiple Imputation in Health-Care Databases: An Overview and Some Applications,” *Statistics in Medicine*, vol. 10, pp. 585–598, 1991.

[29] J. Siddique and T. R. Belin, “Using an Approximate Bayesian Bootstrap to Multiply Impute Nonignorable Missing Data,” *Computational Statistics and Data Analysis*, vol. 53, no. 2, pp. 405–415, 2008.

[30] J. Siddique and T. R. Belin, “Multiple imputation using an iterative hot-deck with distance-based donor selection,” *Statistics in Medicine*, vol. 27, no. October 2006, pp. 83–102, 2008.

[31] J. Siddique, O. Harel, and C. M. Crespi, “Addressing Missing Data Mechanism Uncertainty Using Multiple-Model Multiple Imputation: Application to a Longitudinal Clinical Trial,” *Annals of Applied Statistics*, vol. 6, no. 4, pp. 1814–1837, 2012.

[32] J. Y. Kim and J. K. Kim, “Parametric fractional imputation for nonignorable missing data,” *Journal of the Korean Statistical Society*, vol. 41, pp. 291–303, Sept. 2012.

[33] J. K. Kim and C. L. Yu, “A semi-parametric approach to fractional imputation for nonignorable missing data,” in *Proceedings of the Survey Research Methods Section, ASA*, no. 1994, pp. 2603–2610, 2009.

[34] J. Y. Kim, “Parametric Fractional Imputation for Non-Ignorable Categorical Missing Data With Follow-Up,” *Australian & New Zealand Journal of Statistics*, Aug. 2012.

[35] R. R. Andridge and R. J. Little, “Proxy Pattern-Mixture Analysis for Survey Nonresponse,” *Journal of Official Statistics*, vol. 27, no. 2, pp. 153–180, 2011.

[36] RTI International, “2012 Ohio Medicaid Assessment Survey: Sample Design and Methodology,” tech. rep., 2012.

[37] J. K. Kim, J. Michael Brick, W. a. Fuller, and G. Kalton, "On the bias of the multiple-imputation variance estimator in survey sampling," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, pp. 509–521, June 2006.

[38] R. R. Andridge and R. J. Little, "Extensions of Proxy Pattern-Mixture Analysis for Survey Nonresponse," in *Proceedings of the Survey Research Methods Section, ASA*, pp. 2468–2482, 2009.

[39] J. Siddique and O. Harel, "MIDAS : A SAS Macro for Multiple Imputation Using Distance-Aided Selection of Donors," *Journal of Statistical Software*, vol. 29, no. 9, 2009.

[40] R. F. Tate, "Applications of Correlation Models for Biserial Data," *Journal of the American Statistical Association*, vol. 50, no. 272, pp. 1078–1095, 1955.

[41] R. F. Tate and J. F. Hannan, "Estimation of the parameters for a multivariate normal distribution when one variable is dichotomized," *Biometrika*, pp. 664–668, 1965.

[42] U. Olsson, F. Drasgow, and N. J. Dorans, "The Polyserial Correlation Coefficient," *Psychometrika*, vol. 47, no. 3, pp. 337–347, 1982.

[43] D. Haziza and J.-F. Beaumont, "On the Construction of Imputation Classes in Surveys," *International Statistical Review*, vol. 75, pp. 25–43, Apr. 2007.

[44] G. Casella and R. L. Berger, *Statistical Inference.* Duxbury Thomson Learning, 2 ed., 2002.

[45] N. Howlader, A.-M. Noone, M. Yu, and K. A. Cronin, "Use of imputed population-based cancer registry data as a method of accounting for missing information: application to estrogen receptor status for breast cancer," *American Journal of Epidemiology*, vol. 176, pp. 347–56, Aug. 2012.