HHS Public Access

Author manuscript

J Biopharm Stat. Author manuscript; available in PMC 2015 March 09.

Published in final edited form as:

J Biopharm Stat. 2014 ; 24(3): 634–648. doi:10.1080/10543406.2014.888444.

A NONPARAMETRIC MULTIPLE IMPUTATION APPROACH FOR DATA WITH MISSING COVARIATE VALUES WITH APPLICATION TO COLORECTAL ADENOMA DATA

Chiu-Hsieh Hsu[1,2], **Qi Long**[3], **Yisheng Li**[4], and **Elizabeth Jacobs**[1,2]

[1]Division of Epidemiology and Biostatistics, College of Public Health, University of Arizona, Tucson, Arizona, USA

[2]Arizona Cancer Center, College of Medicine, University of Arizona, Tucson, Arizona, USA

[3]Department of Biostatistics and Bioinformatics, School of Public Health, Emory University, Atlanta, Georgia, USA

[4]Department of Biostatistics, University of Texas MD Anderson Cancer Center, Houston, Texas, USA

Abstract

A nearest neighbor-based multiple imputation approach is proposed to recover missing covariate information using the predictive covariates while estimating the association between the outcome and the covariates. To conduct the imputation, two working models are fitted to define an imputing set. This approach is expected to be robust to the underlying distribution of the data. We show in simulation and demonstrate on a colorectal data set that the proposed approach can improve efficiency and reduce bias in a situation with missing at random compared to the complete case analysis and the modified inverse probability weighted method.

Keywords

Missing at random; Multiple imputation; Nearest neighbor; Nonparametric imputation

1. INTRODUCTION

In regression analysis, sometimes some covariates are subject to missing data due to technical or financial issues, especially for nutritional studies. For example, while investigating whether vitamin D is associated with risk of cancers in order to develop prevention strategies, 25(OH)D, a metabolite of vitamin D commonly studied in epidemiological research, often is not available for all of the participants who have an observed clinical outcome due to, for example, limited financial resources for collecting the blood/tissue samples. In regression analysis, not only can missing covariate values result in a

Address correspondence to Chiu-Hsieh Hsu, Division of Epidemiology and Biostatistics, Mel and Enid Zuckerman College of Public Health and Arizona Cancer Center, University of Arizona, 1295 N. Martin A232, Campus PO Box 245211, Tucson, AZ 85724, USA; pablo1639@gmail.com.

loss of efficiency in estimation of regression coefficients, but there is also potential for bias if the missing data mechanism is nonignorable.

In addition to the covariate with missing data and the outcome, additional covariates are often collected for each study participant, which may be predictive of the missing covariate values or the probabilities of missingness. Hence, these covariates may be useful for recovering missing covariate information for the participants. There is an extensive body of literature on statistical methods that use covariates to predict either missing observations or the probabilities of missingness (Robins et al., 1994; Little and Wang, 1996; Scharstein et al., 1999, Little and Hyonggin, 2004). Most of these methods predict either the missing observations (Little and Wang, 1996) or the probabilities of missingness (Robins et al., 1994; Scharfstein et al., 1999). Only a few predict the two simultaneously (Little and Hyonggin, 2004). Furthermore, these methods directly use the covariates to predict the missing observations or the probabilities of missingness. While such an approach is usually efficient when the prediction models are correctly specified, its performance can be sensitive to the misspecification of the prediction models. To overcome this limitation, we propose a nearest neighbor-based multiple imputation approach to handling missing observations that uses covariates to predict both the missing observations and the probabilities of missingness in an indirect way. For each missing covariate observation, our nearest neighbor-based multiple imputation does not directly incorporate the covariates into estimation but only uses the covariates to select a subset of observations that have a similar covariate profile as the observation with missing covariate information. As a result, our proposed approach is expected to be more robust to the misspecification of the assumptions underlying the working parametric models. Another important feature of the proposed approach is that it allows complex covariate structures.

Multiple imputation (Rubin, 1987) is a common tool used for handling missing data. It replaces each missing value with a set of plausible values that incorporates the uncertainty about the underlying value to be imputed. We previously proposed a multiple imputation approach to impute event times for censored observations in survival analysis (Hsu et al., 2006) and to impute outcomes for subjects with missing outcomes in estimation of population mean (Long et al., 2012). We proposed using two predictive scores to define a neighborhood to impute event times for each censored case and to impute outcomes for each missing outcome case. This idea is similar to predictive mean matching (Rubin, 1986) and propensity score matching (Rosenbaum and Rubin, 1985) in the missing data literature. We derived the two predictive scores from two working regression models. We showed through simulations that the use of two working predictive scores induces a double robustness property (Robins et al., 2000). Specifically, if one of the two working models is correctly specified, the estimator based on the imputed data sets is consistent under some commonly imposed conditions. We also showed that incorporating the predictive variables into the multiple imputation method can both increase efficiency and reduce bias.

Building on our previous work in dealing with censored data in estimating survival function and missing outcomes in estimating population mean, we propose using predictive covariates to define a nearest neighborhood of similar observations for each missing covariate value and then generate imputes from this set of neighbors to estimate regression

coefficients when some covariate values are missing. Specifically, for each missing covariate observation, we will use two working models to define a set of similar observations called the imputing set. One model is a regression model for predicting the missing values. The other is a regression model for predicting the probabilities of missingness. For each missing observation, an observation is randomly drawn from the imputing set. Upon the completion of imputation, a regression model for the outcome can be developed based on the data set with imputed observations. We expect that this approach will induce a double robustness property under a missing at random (MAR) mechanism, that is, where missingness is only dependent upon the predictive covariates. The inverse probability weighting approach (Robins et al., 1994) is one of the popular existing approaches for dealing with regression with missing covariates and also has a double robustness property. We compare our multiple imputation approach with the inverse probability weighting approach.

This article is organized as follows. In the Methods section, we introduce notation used throughout the article, briefly review the inverse probability weighting approach, and describe the imputation procedures. In the Results section, we first study properties of the multiple imputation method for finite sample sizes through simulation and then demonstrate the imputation approach using baseline data from an ursodeoxycholic acid (UDCA) colorectal adenoma prevention study in which the serum 25(OH)D level was only available for some of the participants whose clinical outcomes were observed. We conclude with a discussion about the performance and potential generalizations and limitations of the proposed imputation approach.

2. METHODS

2.1. Notation

For simplicity, we consider a situation with a simple pattern of univariate nonresponse where only one covariate has missing values. Let Y denote the outcome, X_1 denote the covariate with missing observations, M denote the missingness indicator, that is, $M = 1$ if X_1 is observed and $M = 0$ otherwise, X_2 denotes the fully observed covariates that are predictive of X_1, M, or both, and $X = (1, X_1, X_2)$. Suppose there are n independent subjects in the study. We describe our proposed multiple imputation procedures for estimating the regression coefficients in the regression of Y on X in the following.

2.2. Inverse Probability Weighting (IPW) Approach

The idea behind the inverse probability weighting (IPW) approach is intuitive and attractive. For estimating the regression coefficients in the regression of Y on X, IPW requires solving weighted estimating equations, $\sum_i \frac{M_i}{\pi_i} X_i [Y_i - E(Y_i|X_i)] = 0$, where $\pi_i = \Pr(M_i = 1)$ (i.e., the estimated probability of X_{1i} being observed). The IPW approach only includes individuals who were fully observed and its estimation performance highly relies on how well π_i is estimated. The IPW approach has been modified to include partially observed individuals into estimation as well (Robins et al., 1994). Specifically, there are two terms in the weighted estimating equation

$\sum_i \frac{M_i}{\pi_i} X_i \left[Y_i - E\left(Y_i | X_i \right) \right] + \left(1 - \frac{M_i}{\pi_i} \right) E \left\{ X_i \left[Y_i - E\left(Y_i | X_i \right) \right] | Y_i, X_{2i} \right\} = 0$. The first term (i.e., complete case analysis) is solely based on the fully observed individuals, and the second term (i.e., calibration term) is based on both fully and partially observed individuals conditional on the observed data, where a working model is fitted to predict the missing covariates. This modified IPW approach (denoted as IPW_{DR}) has been shown to have a double robustness property (Robins et al., 2000). Specifically, if at least one of π_i and $E\{X_i[Y_i - E(Y_i|X_i)]|Y_i, X_{2i}\}$ is correctly specified, the regression coefficient estimates derived from the modified IPW will be consistent under defined conditions. In this article, we compare the proposed nonparametric multiple imputation approach with IPW_{DR} in terms of robustness to misspecification of models on π_i and/or $E\{X_i[Y_i - E(Y_i|X_i)]|Y_i, X_{2i}\}$.

2.3. Imputation Procedures

For each missing covariate observation, we seek an imputing set consisting of observations from participants without missing data who are similar to the participant with a missing covariate observation as defined in the following. Five steps are used for defining the imputing set and analyzing the imputed data sets.

Step 1: Identifying the covariates predictive of the missing covariate or missingness—Standard regression analysis of the observed X_1, for example, simple linear regression when X_1 is a continuous variable, can be performed to identify all of the potential covariates that are predictive of X_1. Logistic regression of the missingness status, M, can be performed to identify all of the potential covariates that are predictive of the missingness of X_1. A higher significance level, for example, 0.10, can be used to ensure a high likelihood of inclusion of all of the potential predictive covariates, that is, X_2.

In the preceding procedures, we make an implicit assumption that all potential covariates that are predictive of X_1 and/or the missingness of X_1 are measured. When this assumption is not true, however, that is, when both working models might be misspecified, we also evaluate the robustness of the proposed procedures, in comparison to that of the existing approaches via simulations. In addition, when all relevant covariates are measured, the proposed variable selection procedure is expected to identify the correct working model(s) in large samples, provided that the proportion of the observed X_1 is bounded away from 0, under an MAR mechanism for X_1.

Step 2: Calculating predictive scores—Based on the idea behind the predictive mean matching (Rubin, 1986), we first create a scalar summary predictive score based on the fully observed variables including the predictive covariates, X_2, and the outcome, Y, which provides a profile of an individual's X_1. To achieve that, we propose to exploit the associations between (Y, X_2) and X_1 by fitting a working regression model using cases with no missing values for X_1. The working regression model can be a linear or generalized linear regression model depending on whether the variable X_1 is continuous or categorical. We then derive the predictive scores for both the nonmissing and missing cases using the working regression model. When the regression model is correctly specified, an imputing set for each missing case can be defined based on the predictive scores; the resulting multiple imputation method for assessing the association between Y and X can lead to an

improvement in efficiency of the association estimator in the case of missing completely at random (MCAR) and a consistent estimator in the case of MAR. In the latter case, if the regression model is misspecified, bias may remain because conditional on the score derived from the working regression model alone, MCAR cannot be induced within an imputing set that is defined using the score. Hence, we also investigate a working regression model that calculates a missingness score to summarize the association between (Y, X_2) and the missing status (M). One obvious choice of the working regression model is a logistic regression model, given that the missing status is a binary outcome. This idea is analogous to the propensity score matching (Rosenbaum and Rubin, 1985). Since both working models use the clinical endpoint (Y) and the predictive covariates (X_2) as covariates, each score is a linear combination of Y and X_2. Let $Z^* = (Y, X_2)$ denote the covariates included in the working regression models. The two predictive scores can then be defined as $S_x = a'Z^*$ and $S_m = b'Z^*$, where a denotes the vector of the estimates of the regression coefficients of Z^* in the working regression model for X_1 and b denotes the vector of the estimates of the regression coefficients of Z^* in the working regression model for M. To fit these two working models, variable selection will be conducted to choose a subset of the fully observed variables that are associated with X_1 and M, respectively, described earlier at Step 1. This indicates that the two working regression models can include a different set of covariates in the models. The two scores are then centered and scaled (denoted as S_{cx} and S_{cm}, respectively). This strategy summarizes the multidimensional structure of the fully observed variables into a two-dimensional summary score. The hope is that this two-dimensional summary score contains most, if not all, information about X_1 and M.

Those two working models allow complex covariate structures in the sense that they could include interactions between Z^*'s, transformation of each Z^* or a different set of the covariates in each of the two models. Note that if Y is not included in these working models, the association between Y and X_1 may be attenuated and a biased estimate of the association will result. This is because the noise added to the conditional means does not account for partial correlation of X_1 and Y given X_2 (Little, 1992).

Step 3: Defining the imputing set—We propose to calculate a distance to define similarity between subjects based on the two predictive scores, S_{cx} and S_{cm}. Specifically, the distance between subjects j and k is defined as

$d(j,k) = \sqrt{w_1[S_{cx}(j) - S_{cx}(k)]^2 + w_2[S_{cm}(j) - S_{cm}(k)]^2}$ where w_1 and w_2 are nonnegative weights that sum to 1. For each subject j with a missing X_1, this distance is then employed to define a set of, specifically, NN nearest neighbors. This neighborhood of j, denoted as $R(j, NN, w_1, w_2)$, consists of NN subjects who have the smallest NN distances from subject j based on weights w_1 and w_2. For example, $R(j, NN = 5, w_1 = 0.8, w_2 = 0.2)$ consists of five subjects with the five nearest distances from subject j based on weights $w_1 = 0.8$ and $w_2 = 0.2$ among those who have an observed X_1.

We have previously studied the combination of these two scores in survival analysis (Hsu et al., 2006) and estimation of population mean with missing outcomes (Long et al., 2011), and have shown that the use of the two working scores induces a double robustness property. We have also found that nonzero weights for w_2 are useful in reducing the bias resulting from

misspecification of the working regression model for predicting X_1, as long as the working regression model for missingness probability is not seriously misspecified. Specifically, a small weight w_2 (e.g., 0.2) will result in incorporating the score from the missing probability model into the task of defining a set of nearest neighbors. Following similar arguments in these previous studies of ours, if one of these two working regression models is correctly specified, conditional on these two scores, the covariate with missing values is independent of the missing status. Hence, within an imputing set that is defined using these two scores, the missing data mechanism becomes missing completely at random (MCAR), and we expect the combination of these two scores will have the same properties in a regression setting with a missing covariate under an MAR mechanism. We study these properties and the effects of the size of the nearest neighborhood and weights through simulations to see to what extent a double robustness property for model misspecification can be established.

Step 4: Imputation schemes—For subject j who has a missing X_1, after the imputing set $R(j, NN, w_1, w_2)$ is defined, a multiple imputation scheme, denoted as $NNMI(NN, w_1, w_2)$, can be described as follows: For each subject j who has a missing covariate observation of X_1, an observation is drawn equally likely from the imputing set $R(j, NN, w_1, w_2)$. After all missing observations of X_1 are imputed, one fully imputed data set results. This procedure will be independently repeated K times to obtain K imputed data sets for use in estimation. In a linear regression setting, a small number of imputes, for example, three to five, is usually sufficient. In this article, we use $K = 5$.

Step 5: Analyzing imputed data sets—Suppose a standard regression model will be the final analysis model to study the association between Y and X for each fully imputed data set. For example, if the outcome Y is binary, a logistic regression model will be fitted to the imputed data sets. If the outcome Y is a continuous outcome, a linear regression model will be fitted to the imputed data sets. The methods for analyzing multiply imputed data sets have been well established (Rubin, 1987). Specifically, the final estimate of a regression coefficient is the average of the K regression coefficient estimates and the final variance is the sum of the sample variance (denoted as B) of the K regression coefficient estimates and the average (denoted as U) of the K variance estimates of the regression coefficient estimator. The final estimate follows a t distribution with a degree of freedom $v = (K - 1)*[1 + \{U*K/(K + 1)\}/B]^2$, and can be used for testing the null hypothesis of no association between Y and X (Rubin, 1987).

The multiple imputation procedure by itself does not incorporate the full uncertainty in the imputed values, because it does not include a first stage of an initial parameter draw; in other words, it does not incorporate the uncertainty involved in estimating the regression coefficients a and b in the working models. Multiple imputation methods can be enhanced by including a bootstrap stage, which has been shown to improve their performance (Rubin and Schenker, 1991; Heitjan and Little, 1991). Specifically, a bootstrap sample is selected with replacement from the original data set. The preceding imputation procedures are then conducted on this bootstrap sample. The imputing set for subject j is the nearest neighborhood $RB(j, NN, w_1, w_2)$ consisting of NN subjects with observed X_1 with the NN nearest distances from subject j based on weights w_1 and w_2 among those in the Bootstrap

sample. The MI method incorporating the Bootstrapping, denoted as $NNMIB(NN, w_1, w_2)$, randomly draws a value from $RB(j, NN, w_1, w_2)$ to impute the missing value. Multiple imputations are done by repeating the bootstrap stage K times. Due to the general underestimation of the uncertainty for the multiple imputation method (see, e.g., Long et al., 2011), in this article we only focus on exploring the performance of the NNMIB method.

3. RESULTS

3.1. Simulation Study

We performed a simulation study to investigate the finite sample properties of the NNMIB method in a regression setting. For each of 500 independent simulated data sets, X_1 subject to missing was generated from $N(2, 1)$ or Poisson(1), X_2 fully observed was generated from $N(2, 1)$, $N(3 - 0.5X_1, 1)$, or $N(3 - 0.5X_1, 0.5)$, Y fully observed was generated from $N(b_0 + b_1X_1 + b_2X_2, 4)$ or $N(b_0 + b_1X_1 + b_2X_2, 8)$, where $b_0 = 10$, $b_1 = 1.333$, $b_2 = -1.333$, and missing indicator for X_1, that is, M, was generated from $Pr(M = 1) = exp(r_0 + r_1X_2 + r_2Y)/[1 + exp(r_0 + r_1X_2 + r_2Y)]$, where $r_0 = -0.5$, $r_1 = -1.5$, $r_2 = 0.5$ when $X_1 \sim N(2,1)$ and $r_0 = -0.3$, $r_1 = -1.0$, $r_2 = 0.5$ when $X_1 \sim Poisson(1)$. Those parameters were chosen to control the missing rate at approximately 35%. A sample size of 100 and 200 was considered in this article. We mainly focused on comparing the estimates of the regression coefficients, b_0, b_1, b_2, for Y with X_1 and X_2 as the covariates, across the fully observed (FO), which was treated as the gold standard since all X_1 were fully observed, complete case (CC), which only included the observations without missing covariates in the analysis, double robust inverse probability weighting (IPW^{DR}), and NNMIB methods. In addition, we were also interested in exploring the effects of NN, w_1, w_2, and misspecification of the underlying distribution of X_1 conditional on Y and X_2 for the NNMIB method.

For the FO method, a linear regression model with X_1 and X_2 as the covariates was fitted to the data (Y) before the missing indicator was applied to the data. For the CC method, a linear regression model was fitted using the complete cases only. Two working regression models need to be fitted to construct the weighted estimating equations and select imputing sets for IPW_{DR} and NNMIB methods, respectively. One is a working linear regression model (M_1) for predicting X_1. The other is a working logistic regression model (M_2) for predicting missingness probabilities. Three scenarios of the two working models were considered, that is, at least one of the two working models with both Y and X_2 as the covariates in the model, including: (1) M_1 with X_2 as the covariate and M_2 with both Y and X_2 as the covariates (denoted as IPW_{DR12} and $NNMIB_{12}$), (2) M_1 with Y and X_2 as the covariates and M_2 with X_2 as the covariate (denoted as IPW_{DR21} and $NNMIB_{21}$), and (3) both models with both Y and X_2 as the covariates (denoted as IPW_{DR22} and $NNMIB_{22}$). M_1 was considered as correctly specified if both Y and X_2 were included in the model and X_1 was normally distributed; otherwise, M_1 was misspecified. This indicates that when $X_1 \sim Poissson(1)$, M_1 was misspecified even in a situation with both Y and X_2 as the covariates in the model because X_1 conditional on Y and X_2 did not follow a normal distribution. M_2 was considered as correctly specified if both Y and X_2 were included in the model; otherwise, M_2 was misspecified.

The results are provided in Tables 1-4. When X_1 was generated from a normal distribution (Tables 1 and 2), that is, the distributional assumption for the working regression model for predicting missing values was correct, the CC method had the largest bias in estimating the regression coefficients b_1 and b_2 compared to the IPW_{DR} and NNMIB methods. The bias emerged because the CC method did not take into account the MAR mechanism when estimating the regression coefficients. The bias also resulted in lower coverage rates for CC. For IPW_{DR}, the bias tended to be smaller when the working regression model for predicting missing values was correctly specified (i.e., IPW_{DR21} and IPW_{DR22}). IPW_{DR} estimates had much greater variation in terms of both SD and SE, especially for IPW_{DR22}, compared to the other methods. Each of the NNMIB methods produced estimates comparable to FO and its counterpart of the IPW methods in terms of both bias and coverage rate when NN = 3. As expected, for NNMIB the bias increased and SD and SE decreased when NN increased. In addition, the bias increased with the weight on the predictive score for missingness when the working regression model for predicting missing values was correctly specified. As the sample size increased to 200 (Table 2), the bias decreased for all NNMIB estimators and sometimes was even smaller than its counterpart of IPW_{DR}. For example, $NNMIB_{12}(3, 0.5, 0.5)$ and $NNMIB_{12}(3, 0.2, 0.8)$ had smaller bias for all three regression coefficients compared to IPW_{DR12}. $NNMIB_{21}(3,0.8,0.2)$ had smaller bias for b_2 compared to IPW_{DR21}.

When X_1 was generated from a Poisson distribution (Tables 3 and 4), that is, the distributional assumption for the working regression model for predicting missing values was incorrect, we mainly focused on comparing $NNMIB_{21}$ and $NNMIB_{22}$ with IPW_{DR21} and IPW_{DR22}, respectively, to examine whether NNMIB is more robust to the distributional assumption compared to IPW_{DR}. Based on Tables 3 and 4, $NNMIB_{21}$ and $NNMIB_{22}$ had a smaller bias and a coverage rate closer to FO than IPW_{DR21} and IPW_{DR22}, respectively, when more weight was put on the predictive score for missing values and NN = 3. The coverage rate was slightly off from the nominal level (i.e., 95%) for IPW_{DR21} and IPW_{DR22} due to the bias. The bias became larger when the correlation between X_1 and X_2 was stronger (Table 4).

In summary, the CC method tended to produce biased estimates, as expected. The IPW_{DR} and NNMIB methods both could produce a reasonable estimate in a situation with MAR if one of the two working regression models was correctly specified. The NNMIB method, which used the predictive covariate to recover information for missing observations, may potentially gain efficiency compared to the IPW_{DR} method and reduce bias due to MAR compared to the CC method and the IPW_{DR} method through the selection of the weights on the two predictive scores and size of the nearest neighborhood. A potential reason underlying the performance of the inverse probability weighting method in our simulations is the unstable inverse weighting in finite samples. In addition, whether the NNMIB method is asymptotically more efficient compared to the inverse probability weighting method requires additional investigation that is beyond the scope of this article and will be studied in the future research. Finally, the NNMIB method was shown to be more robust to the distributional assumption compared to the IPW_{DR} method.

3.2. Application to UDCA Data

The UDCA data consist of 1,192 patients, who underwent removal of colorectal adenomas between January 1996 and January 2000, from a colorectal adenoma prevention trial conducted at the Arizona Cancer Center (Alberts et al., 2005). Demographic information, including age, gender, and body mass index (BMI), and dietary vitamin D intake information based on the Arizona Food Frequency Questionnaire (AFFQ) (Martinez et al., 1999) were collected on all of the 1,192 participants. The vitamin D dietary intake based on the AFFQ was subject to measurement error, as vitamin D can be synthesized endogenously in the skin upon ultraviolet (UV) irradiation (Holick, 1999); therefore, a serum vitamin D metabolite was measured to obtain a more accurate measurement. However, due to a limited budget, of the 1,192 participants, only 598 (50.2%) participants were selected to perform an assay to measure the serum vitamin D level. The vitamin D metabolite employed in this study was 25(OH)D, which is the best overall marker of vitamin D status (Jacobs et al., 2007; Jacobs et al., 2008). For those participants who were not selected for the assay, their serum 25(OH)D levels were regarded as missing data. We applied the proposed nonparametric multiple imputation method to estimate the association between the size of each participant's largest baseline colorectal adenomas and serum 25(OH)D adjusting for age and gender.

Based on simple linear regression using the 598 complete cases, gender, BMI, and vitamin D intake derived from the AFFQ were significantly associated with the serum 25(OH)D level at a significance level of 0.10. On average, males tended to have a higher level of 25(OH)D compared to females with a p-value of 0.03, participants with higher vitamin D intake derived from the AFFQ tended to have a higher level of 25(OH)D with a p-value of 0.01, and participants with higher BMI tended to have a lower level of 25(OH)D with a p-value < 0.01. Based on logistic regression, gender was associated with the probability of missingness at a significance level of 0.10. Females were more likely to have missing serum 25(OH)D compared to males with a p-value of 0.05. Gender was associated with both the serum 25(OH)D level and the probability of missingness. These results implied a potential MAR mechanism for the outcome of the serum 25(OH)D levels. These variables, as well as age, were therefore used to define the predictive scores. The reason that age was also included was to assure congeniality (Meng, 1994). The proposed nearest neighbor-based multiple imputation procedure was then used to recover the information for missing serum 25(OH)D observations.

We fitted a working linear regression model to predict the serum 25(OH)D level using data from the 598 participants with gender, BMI, the vitamin D intake from the AFFQ, and the size of the largest baseline colorectal adenoma as the predictive covariates. We also fitted a logistic regression model to predict the probability of missingness using data from all of the 1,192 participants with gender and the size of the largest baseline colorectal adenoma as the predictive covariates. Two scores, as the linear combinations of the predictive covariates, were derived from the two working models. The Pearson's correlation coefficient between the two scores was −0.34, which suggested some degree of the MAR mechanism for the outcome of the serum 25(OH)D level. Hence, we expected to see improvement in both bias and efficiency of estimation by using the two scores to define a nearest neighbor for

imputation for each missing observation with the number of imputes (K) set at 5. Upon completion of the imputation, a multiple linear regression model was fitted to the imputed data sets where size of the largest baseline colorectal adenoma was the outcome variable and the imputed serum 25(OH)D level, male indicator, and age were the covariates in the model. Several combinations of the size of nearest neighborhood (NN) and weights (w_1, w_2) were used to study the performance of the nonparametric imputation method (NNMIB) and to compare with the complete case analysis (CC) and the modified inverse probability weighting method (IPW_{DR}).

The analysis results are provided in Table 5. The CC analysis showed no statistically significant association between size of the largest baseline colorectal adenoma and the serum 25(OH)D level and age with a p-value of 0.096 and 0.089, respectively, similar to what was reported for this population previously (Jacobs et al., 2008). The CC analysis also showed that male tended to have a smaller size of the largest baseline adenoma compared to female with a p-value of 0.032. Based on the findings from our simulation study and a suggested degree of MAR mechanism for the data, the CC analysis simply ignoring missing observations is expected to be biased and less efficient than the NNMIB approach. Based on Table 5, both IPW_{DR} and NNMIB methods produced different estimates of the regression coefficients than the CC analysis, especially for age and male indicator. In addition, NNMIB had much smaller estimates of standard errors for male indicator and age compared to the CC analysis. NNMIB gained about 26% and 30% efficiency for male indicator and age, respectively, by incorporating the predictive covariates into imputation. IPW_{DR} had much larger estimates of standard errors (SE) compared to the CC analysis (similar to the findings in our simulations). The changes in estimates of both regression coefficients and SE for both IPW_{DR} and NNMIB resulted in different significance findings. For IPW_{DR}, none of 25(OH)D, male indicator, and age was significantly associated with the size of the largest baseline colorectal adenoma due to larger estimates of SE. For NNMIB, male had a significantly smaller size of the largest baseline colorectal adenoma than female had, and age was not significantly associated with the size of the largest baseline colorectal adenoma. When a weight of at least 0.5 was put on the predictive score for missingness, NNMIB indicated that the participants with higher 25(OH)D had a significantly smaller size of the largest baseline colorectal adenoma than the participants with lower 25(OH)D had. Overall, the NNMIB method using the predictive covariates in the estimation had potential to improve efficiency and reduce bias in estimating the association between the size of the largest baseline colorectal adenomas and the serum 25(OH)D concentration.

4. DISCUSSION

This article describes a nonparametric multiple imputation procedure for regression analysis with missing covariates, which uses predictive variables to recover information for missing covariate observations and is easy to implement. An attractive feature of the proposed nonparametric multiple imputation procedure is that its reliance on a correct specification of the working parametric models is weak, because the two working models are only used to identify a neighborhood of similar observations from which imputes are drawn for each missing covariate observation. After the imputation, the analysis is conducted on the original data, augmented by the imputed data. This indicates that this multiple imputation method

indirectly incorporates the information from the predictive covariates into estimation of the association. Therefore, the proposed approach is expected to be robust to misspecification of the underlying distribution of the covariate with missing observations. In contrast, most of the methods in the literature directly incorporate the information from the predictive covariates into estimation of the association, and therefore their performance will highly depend on the correctness of the model specification. Our simulation study shows that the use of this multiple imputation method has potential to lead to improved performance in estimation, in terms of both bias and efficiency. In general, the multiple imputation estimators were less variable than the estimates produced by analyzing the complete cases without using predictive covariates and the estimates derived from the double robust inverse probability weighting method. In addition, the multiple imputation estimators were more robust to the distributional assumptions on the covariate that has missing values than the double robust inverse probability weighting method.

In this article, we propose the imputation method in a linear regression setting where a covariate has missing values, and demonstrate the imputation method by analyzing a colorectal adenoma data set. The proposed imputation method can be applied to handle any data with a missing covariate and observed predictive variables of the missing covariate. The proposed imputation method can also be generalized to handle linear or generalized linear regression in which more than one covariate have missing values. In pharmaceutical studies, there are often missing data involved, especially for biomarker data. The proposed multiple imputation method can be used to recover biomarker information for the subjects with missing biomarker data.

The performance of the proposed imputation method in improving efficiency and reducing bias depends on how predictive the variables are for both the missing values and missing probabilities. In our simulations, we noticed that when the correct covariates were included in the working regression model for predicting missing values, the imputation method produced estimates with smaller bias even under a situation where the distribution of missing covariate was misspecified. This suggests that it may be more important to seek good models for predicting missing values than to find reasonable working models for both missing values and the probabilities of missingness. It is a similar case with survival analysis in that a correct specification of the working model for the failure time is more important (Hsu et al., 2006).

The adequacy of the imputation procedures will depend on the “nearness” of the imputing set. When the nearest neighborhood contains some observations that are not close enough to the missing observation, some remnant of the missing at random mechanism remains within the neighborhood, which could contribute to the bias in estimation. The “nearness” of the imputing set will depend on the correction of the specification of the working models, the quality of the parameter estimates from the two working models, especially the parameters from the working regression model for predicting missing values, the size of the nearest neighborhood, and the weights on the two predictive scores. In this article, we simply use linear regression to predict the covariate with missing observations. Potentially, when the covariate is not normal, a transformation of the covariate may be performed to better approximate a normal distribution, or a more general regression model such as the

generalized linear model may be fitted to predict the values of the missing covariate. The chosen size of the nearest neighborhood depends on both the sample size and missing rate. As for the weights on the two predictive scores, a small weight (e.g., 0.2) for the predictive score derived from the missing probability model is usually sufficient even under a MAR mechanism based on our previous study in survival analysis (Hsu et al., 2006). Sensitivity analysis can be performed to select the optimal size of the nearest neighborhood and the optimal weights (Long et al., 2011). In addition, future work of investigating the theoretical properties (i.e., double robustness and asymptotic efficiency) of the proposed nonparametric multiple imputation is required to decide whether the NNMIB method is asymptotically more efficient compared to the inverse probability weighting method.

REFERENCES

Alberts DS, Martinez ME, Hess LM, Einsphar JG, Green SB, Bhattacharyya AK, Guillen J, Krutzsch M, Batta AK, Salen G, Fales L, Koonce K, Parish D, Clouser M, Roe D, Lance P. Phase III trial of ursodeoxycholic acid to prevent colorectal adenoma recurrence. Journal of the National Cancer Institute. 2005; 97:846–853. [PubMed: 15928305]

Heitjan DF, Little RJA. Multiple imputation for the fatal accident reporting system. Applied Statistics. 1991; 40:13–29.

Holick, M. Vitamin D. In: Shils, ME.; Shike, M.; Ross, AC., editors. Modern Nutrition in Health and Disease. Williams & Wilkins; Baltimore: 1999. p. 329-345.

Hsu C-H, Taylor JMG, Murray S, Commenges D. Survival analysis using auxiliary variables via nonparametric multiple imputation. Statistics in Medicine. 2006; 25:3503–3517. [PubMed: 16345047]

Jacobs ET, Alberts DS, Benuzillo J, Hollis BW, Thompson PA, Martinez ME. Serum 25(OH)D levels, dietary intake of vitamin D, and colorectal adenoma recurrence. Journal of Steroid Biochemistry & Molecular Biology. 2007; 103:752–756. [PubMed: 17223551]

Jacbos E, Alberts DS, Foote JA, Green SB, Hollis BW, Yu Z, Martinez ME. Vitamin D insufficiency in southern Arizona. American Journal of Clinical Nutrition. 2008; 87:608–613. [PubMed: 18326598]

Little RJA. Regression with missing X's: A review. Journal of the American Statistical Association. 1992; 87:1227–1237.

Little RJA, Wang Y-X. Pattern-mixture models for multivariate incomplete data with covariates. Biometrics. 1996; 52:98–111. [PubMed: 8934587]

Little RJA, Hyonggin A. Robust likelihood-based analysis of multivariate data with missing values. Statistica Sinica. 2004; 14:949–968.

Long Q, Hsu C-H, Li Y. Doubly robust nonparametric multiple imputation for ignorable missing data. Statistica Sinica. 2012; 22:149–172. [PubMed: 22347786]

Martinez ME, Marshall JR, Graver E, Whitacre RC, Woolf K, Ritenbaugh C, Alberts DS. Reliability and validity of a self-administered food frequency questionnaire in a chemoprevention trial of adenoma recurrence. Cancer Epidemiology Biomarkers & Prevention. 1999; 8:941–946.

Meng XL. Multiple-imputation inference with uncongenial sources of input (with discussion). Statistical Science. 1994; 9:538–573.

Robins JM, Rotnitsky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association. 1994; 89:846–866.

Robins JM, Rotnitzky A, van der Laan M. Comment on 'On profile likelihood'. Journal of the American Statistical Association. 2000; 95:477–482.

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

Rubin DB. Statistical matching using file concatenation with adjusted weights and multiple mputation. Journal of Business & Economic Statistics. 1986; 4:87–94.

Rubin, DB. Multiple Imputation for Nonresponse in Surveys. Wiley; New York, NY: 1987.

Rubin DB, Schenker N. Multiple imputation in health-care databases: An overview and some applications. Statistics in Medicine. 1991; 10:585–598. [PubMed: 2057657]

Scharfstein D, Rotnitsky A, Robins JM. Adjusting for nonignorable dropout using semiparametric models. Journal of the American Statistical Association. 1999; 94:1096–1146.

Table 1

Monte Carlo results (based on 500 replicates): Linear regression with missing covariate where N = 100, $X_1 \sim N(2,1)$, $X_2 \sim N(2,1)$, $Y \sim N(b_0 + b_1 * X_1 + b_2 * X_2, 4)$, missing rate = 0.33, Spearman correlation coefficient between X_1 and X_2: 0.00, Spearman correlation coefficient between X_1 and Y: 0.29, and Spearman correlation coefficient between X_2 and Y: −0.29

	b_0 = 10.000				b_1 = 1.333				b_2 = −1.333			
Method	**Est**[a]	**SD**[b]	**SE**[c]	**CR**[d]	**Est**	**SD**	**SE**	**CR**	**Est**	**SD**	**SE**	**CR**
FO	10.044	1.248	1.213	94.2	1.332	0.421	0.406	93.8	−1.360	0.416	0.404	94.4
CC	10.259	1.352	1.273	92.6	1.031	0.454	0.448	90.0	−0.374	0.514	0.493	49.2
IPW_{DR12}	10.519	4.616	13.685	90.2	1.157	1.441	4.146	87.2	−1.409	1.275	4.308	94.2
IPW_{DR21}	10.052	1.621	1.810	94.0	1.351	0.639	0.754	94.0	−1.376	0.514	0.593	94.4
IPW_{DR22}	9.755	3.073	2.667	93.8	1.452	0.986	1.150	94.0	−1.296	0.927	0.936	94.0
$NNMIB_{12}$	M_1: misspecified; M_2: correctly specified											
(3,0.8,0.2)[e]	10.631	1.413	1.494	92.8	1.055	0.520	0.576	94.8	−1.389	0.448	0.444	94.2
(3,0.5,0.5)	10.494	1.461	1.560	94.0	1.103	0.539	0.607	95.8	−1.353	0.454	0.449	94.4
(3,0.2,0.8)	10.409	1.507	1.600	93.6	1.124	0.564	0.625	96.4	−1.319	0.450	0.450	94.8
(5,0.8,0.2)	10.724	1.357	1.479	93.4	1.010	0.489	0.571	94.8	−1.395	0.445	0.439	95.0
(5,0.5,0.5)	10.583	1.413	1.529	94.0	1.065	0.509	0.595	95.4	−1.367	0.446	0.442	94.4
(5,0.2,0.8)	10.516	1.468	1.567	94.4	1.070	0.531	0.608	94.8	−1.332	0.447	0.442	94.4
$NNMIB_{21}$	M_1: correctly specified; M_2: misspecified											
(3,0.8,0.2)	10.219	1.511	1.473	92.6	1.266	0.581	0.575	94.0	−1.375	0.453	0.430	93.4
(3,0.5,0.5)	10.355	1.491	1.473	92.6	1.205	0.560	0.576	94.6	−1.392	0.453	0.434	93.6
(3,0.2,0.8)	10.565	1.423	1.460	93.8	1.110	0.537	0.569	94.4	−1.414	0.449	0.436	94.0
(5,0.8,0.2)	10.286	1.502	1.464	93.0	1.235	0.568	0.570	94.0	−1.381	0.454	0.428	92.8
(5,0.5,0.5)	10.432	1.457	1.456	94.2	1.167	0.554	0.565	94.2	−1.397	0.451	0.431	93.6
(5,0.2,0.8)	10.626	1.386	1.440	94.2	1.079	0.510	0.559	95.4	−1.417	0.446	0.433	93.6
$NNMIB_{22}$	M_1: correctly specified; M_2: correctly specified											
(3,0.8,0.2)	10.134	1.534	1.516	94.0	1.292	0.587	0.596	94.0	−1.347	0.457	0.430	92.8
(3,0.5,0.5)	10.184	1.536	1.535	93.8	1.263	0.591	0.603	93.0	−1.343	0.457	0.435	93.6
(3,0.2,0.8)	10.231	1.540	1.587	92.8	1.220	0.583	0.622	94.6	−1.320	0.457	0.441	94.4
(5,0.8,0.2)	10.211	1.518	1.489	92.8	1.258	0.581	0.582	93.4	−1.357	0.452	0.428	93.2

	$b_0 = 10.000$				$b_1 = 1.333$				$b_2 = -1.333$			
Method	**Est**[a]	**SD**[b]	**SE**[c]	**CR**[d]	**Est**	**SD**	**SE**	**CR**	**Est**	**SD**	**SE**	**CR**
(5,0.5,0.5)	10.241	1.517	1.512	93.0	1.235	0.571	0.592	94.4	−1.349	0.452	0.431	93.8
(5,0.2,0.8)	10.328	1.512	1.542	93.0	1.176	0.563	0.600	94.8	−1.331	0.451	0.435	94.6

[a] Average of 500 point estimates.

[b] Empirical standard deviation of 500 point estimates.

[c] Average of 500 estimated standard errors.

[d] Coverage rate of 500, 95% confidence intervals.

[e] (NN, w_1, w_2).

Table 2

Monte Carlo results (based on 500 replicates): Linear regression with missing covariate where N = 200, $X_1 \sim N(2,1)$, $X_2 \sim N(2,1)$, $Y \sim N(b_0 + b_1 * X_1 + b_2 * X_2, 4)$, missing rate = 0.33, Spearman correlation coefficient between X_1 and X_2: 0.00, Spearman correlation coefficient between X_1 and Y: 0.29 and Spearman correlation coefficient between X_2 and Y: −0.29

	b_0 = 10.000				b_1 = 1.333				b_2 = −1.333			
Method	**Est**[a]	**SD**[b]	**SE**[c]	**CR**[d]	**Est**	**SD**	**SE**	**CR**	**Est**	**SD**	**SE**	**CR**
FO	10.001	0.827	0.856	95.4	1.354	0.290	0.286	96.0	−1.358	0.293	0.286	95.4
CC	10.248	0.873	0.896	95.6	1.049	0.308	0.313	85.0	−0.368	0.350	0.348	22.2
IPW_{DR12}	10.447	1.226	1.608	89.6	1.155	0.555	0.616	87.6	−1.382	0.338	0.459	94.2
IPW_{DR21}	9.969	0.996	1.142	96.6	1.378	0.422	0.450	95.4	−1.360	0.334	0.345	93.8
IPW_{DR22}	9.941	1.200	1.241	95.6	1.379	0.512	0.510	93.4	−1.336	0.340	0.357	94.0
$NNMIB_{12}$	M_1: misspecified; M_2: correctly specified											
(3,0.8,0.2)[e]	10.475	0.966	1.085	94.2	1.122	0.381	0.419	93.8	−1.373	0.316	0.321	95.4
(3,0.5,0.5)	10.335	1.044	1.134	94.8	1.173	0.421	0.439	95.2	−1.343	0.317	0.327	94.8
(3,0.2,0.8)	10.263	1.103	1.196	94.6	1.191	0.447	0.469	96.0	−1.313	0.318	0.329	95.8
(5,0.8,0.2)	10.552	0.926	1.068	93.6	1.085	0.359	0.410	94.2	−1.380	0.313	0.317	95.4
(5,0.5,0.5)	10.418	0.991	1.100	94.4	1.136	0.395	0.425	93.8	−1.352	0.315	0.322	95.6
(5,0.2,0.8)	10.320	1.053	1.153	95.0	1.165	0.417	0.450	95.2	−1.320	0.312	0.321	96.2
$NNMIB_{21}$	M_1: correctly specified; M_2: misspecified											
(3,0.8,0.2)	10.121	1.061	1.096	94.8	1.296	0.436	0.435	93.6	−1.353	0.318	0.308	94.4
(3,0.5,0.5)	10.219	1.041	1.075	94.0	1.254	0.425	0.423	94.0	−1.368	0.319	0.311	94.6
(3,0.2,0.8)	10.393	0.979	1.049	94.0	1.176	0.396	0.409	92.8	−1.388	0.318	0.311	94.4
(5,0.8,0.2)	10.162	1.016	1.071	94.8	1.277	0.415	0.424	94.0	−1.359	0.314	0.304	94.8
(5,0.5,0.5)	10.274	0.995	1.061	94.2	1.228	0.406	0.419	95.2	−1.372	0.313	0.308	95.0
(5,0.2,0.8)	10.463	0.940	1.043	94.8	1.145	0.375	0.409	94.8	−1.396	0.315	0.308	94.8
$NNMIB_{22}$	M_1: correctly specified; M_2: correctly specified											
(3,0.8,0.2)	10.078	1.094	1.136	95.2	1.310	0.453	0.456	93.8	−1.338	0.320	0.308	94.0
(3,0.5,0.5)	10.083	1.143	1.165	93.8	1.299	0.470	0.467	93.6	−1.328	0.323	0.313	94.8
(3,0.2,0.8)	10.116	1.152	1.193	93.6	1.268	0.472	0.473	93.8	−1.311	0.323	0.319	94.8
(5,0.8,0.2)	10.110	1.054	1.109	95.0	1.295	0.435	0.442	93.2	−1.344	0.313	0.305	94.6

	$b_0 = 10.000$				$b_1 = 1.333$				$b_2 = -1.333$			
Method	**Est**[a]	**SD**[b]	**SE**[c]	**CR**[d]	**Est**	**SD**	**SE**	**CR**	**Est**	**SD**	**SE**	**CR**
(5,0.5,0.5)	10.130	1.090	1.131	95.2	1.277	0.446	0.450	94.6	−1.333	0.317	0.309	95.0
(5,0.2,0.8)	10.173	1.080	1.155	95.4	1.242	0.435	0.455	96.0	−1.319	0.315	0.313	95.0

[a] Average of 500 point estimates.

[b] Empirical standard deviation of 500 point estimates.

[c] Average of 500 estimated standard errors.

[d] Coverage rate of 500, 95% confidence intervals.

[e] (NN, w_1, w_2).

Table 3

Monte Carlo results (based on 500 replicates): Linear regression with missing covariate where N = 200, X_1 ~ Poisson(1), X_2 ~ N(3 − 0.5 * X_1, 1), Y ~ N(b_0 + b_1 * X_1 + b_2 * X_2, 4), missing rate = 0.37, Spearman correlation coefficient between X_1 and X_2: −0.42, Spearman correlation coefficient between X_1 and Y: 0.40, and Spearman correlation coefficient between X_2 and Y: −0.43

	b_0 = 10.000				b_1 = 1.333				b_2 = −1.333			
Method	**Est**[a]	**SD**[b]	**SE**[c]	**CR**[d]	**Est**	**SD**	**SE**	**CR**	**Est**	**SD**	**SE**	**CR**
FO	9.921	0.977	0.944	94.6	1.348	0.334	0.320	95.2	−1.307	0.288	0.285	95.6
CC	9.911	1.125	0.995	92.6	1.095	0.348	0.327	87.0	−0.483	0.367	0.339	30.4
IPW_{DR12}	10137	1.367	1.307	93.0	1.239	0.571	0.558	91.0	−1.343	0.377	0.389	95.2
IPW_{DR21}	9.428	1.213	1.174	90.0	1.536	0.454	0.481	90.2	−1.143	0.368	0.360	90.8
IPW_{DR22}	9.543	1.403	1.589	89.6	1.494	0.557	0.758	89.4	−1.188	0.398	0.403	91.4
$NNMIB_{12}$	M_1: misspecified; M_2: correctly specified											
(3,0.8,0.2)[e]	10.349	1.152	1.137	92.6	1.155	0.428	0.443	93.4	−1.408	0.334	0.325	92.2
(3,0.5,0.5)	10.264	1.211	1.173	92.6	1.198	0.467	0.464	92.0	−1.387	0.344	0.332	92.2
(3,0.2,0.8)	10.180	1.257	1.209	93.0	1.240	0.501	0.484	93.6	−1.365	0.352	0.341	92.2
(5,0.8,0.2)	10.439	1.113	1.115	93.4	1.112	0.409	0.431	93.8	−1.431	0.325	0.319	92.0
(5,0.5,0.5)	10.324	1.172	1.153	93.6	1.170	0.447	0.456	94.2	−1.402	0.335	0.327	94.2
(5,0.2,0.8)	10.252	1.213	1.193	94.4	1.208	0.480	0.478	93.8	−1.384	0.342	0.335	93.0
$NNMIB_{21}$	M_1: misspecified; M_2: misspecified											
(3,0.8,0.2)	10.164	1.207	1.167	93.8	1.244	0.459	0.456	93.8	−1.362	0.348	0.355	93.6
(3,0.5,0.5)	10.260	1.168	1.146	92.6	1.197	0.442	0.446	93.2	−1.386	0.338	0.329	93.2
(3,0.2,0.8)	10.373	1.133	1.122	92.6	1.141	0.421	0.432	92.8	−1.414	0.330	0.323	91.0
(5,0.8,0.2)	10.208	1.156	1.144	95.4	1.224	0.437	0.444	93.8	−1.374	0.336	0.329	93.6
(5,0.5,0.5)	10.324	1.134	1.134	94.6	1.165	0.423	0.441	93.8	−1.402	0.330	0.325	93.2
(5,0.2,0.8)	10.474	1.092	1.104	92.8	1.093	0.402	0.424	94.6	−1.439	0.319	0.317	92.2
$NNMIB_{22}$	M_1: misspecified; M_2: correctly specified											
(3,0.8,0.2)	10.068	1.235	1.211	93.8	1.288	0.474	0.478	92.6	−1.336	0.355	0.345	92.4
(3,0.5,0.5)	10.085	1.243	1.241	94.8	1.283	0.486	0.498	93.4	−1.340	0.354	0.350	92.6
(3,0.2,0.8)	10.084	1.266	1.252	93.6	1.289	0.504	0.506	93.0	−1.341	0.356	0.351	92.4
(5,0.8,0.2)	10.112	1.194	1.176	93.8	1.272	0.455	0.462	93.6	−1.351	0.345	0.336	93.2

	$b_0 = 10.000$				$b_1 = 1.333$				$b_2 = -1.333$			
Method	**Est**[a]	**SD**[b]	**SE**[c]	**CR**[d]	**Est**	**SD**	**SE**	**CR**	**Est**	**SD**	**SE**	**CR**
(5,0.5,0.5)	10.142	1.212	1.201	94.4	1.261	0.471	0.479	93.0	−1.358	0.346	0.340	93.8
(5,0.2,0.8)	10.157	1.218	1.219	93.8	1.259	0.483	0.491	94.6	−1.362	0.344	0.343	93.0

[a] Average of 500 point estimates.

[b] Empirical standard deviation of 500 point estimates.

[c] Average of 500 estimated standard errors.

[d] Coverage rate of 500, 95% confidence intervals.

[e] (NN, w_1, w_2).

Table 4

Monte Carlo results (based on 500 replicates): Linear regression with missing covariate where N = 200, X_1 ~ Poisson(1), X_2 ~ N(3 − 0.5 * X_1, 0.5), Y ~ N(b_0 + b_1 * X_1 + b_2 * X_2, 8), missing rate = 0.40, Spearman correlation coefficient between X_1 and X_2: −0.68, Spearman correlation coefficient between X_1 and Y: 0.23, and Spearman correlation coefficient between X_2 and Y: −0.22

	b_0 = 10.000				b_1 = 1.333				b_2 = −1.333			
Method	**Est**[a]	**SD**[b]	**SE**[c]	**CR**[d]	**Est**	**SD**	**SE**	**CR**	**Est**	**SD**	**SE**	**CR**
FO	10.128	3.451	3.505	95.8	1.304	0.786	0.806	97.0	−1.364	1.116	1.137	94.6
CC	11.604	3.282	3.311	92.4	0.786	0.748	0.739	87.8	0.151	1.065	1.098	71.0
IPW_{DR12}	11.686	4.694	4.821	94.4	0.864	1.272	1.161	90.8	−1.786	1.438	1.529	94.6
IPW_{DR21}	8.457	5.195	5.253	94.0	1.756	1.401	1.415	90.6	−0.835	1.592	1.611	93.8
IPW_{DR22}	9.248	6.337	6.451	92.8	1.529	1.720	1.846	91.0	−1.043	1.921	2.064	94.0
$NNMIB_{12}$	M_1: misspecified; M_2: correctly specified											
(3,0.8,0.2)[e]	11.162	4.671	4.972	95.0	1.040	1.254	1.300	94.0	−1.636	1.432	1.522	94.8
(3,0.5,0.5)	10.781	5.102	5.283	94.8	1.167	1.421	1.423	93.4	−1.517	1.542	1.601	94.6
(3,0.2,0.8)	10.599	5.490	5.642	94.2	1.246	1.592	1.582	92.6	−1.452	1.637	1.685	93.8
(5,0.8,0.2)	11.294	4.350	4.808	96.6	1.001	1.134	1.255	96.6	−1.681	1.350	1.474	95.6
(5,0.5,0.5)	10.922	4.773	5.046	94.6	1.131	1.300	1.356	94.8	−1.569	1.460	1.534	94.8
(5,0.2,0.8)	10.777	4.971	5.209	94.6	1.205	1.426	1.456	93.6	−1.522	1.497	1.566	94.4
$NNMIB_{21}$	M_1: misspecified; M_2: misspecified											
(3,0.8,0.2)	10.634	4.840	4.382	92.2	1.187	1.275	1.105	92.2	−1.493	1.484	1.369	94.0
(3,0.5,0.5)	11.144	4.504	4.421	93.8	1.036	1.152	1.115	92.4	−1.643	1.395	1.378	93.8
(3,0.2,0.8)	11.674	4.011	4.391	95.8	0.879	0.979	1.102	95.0	−1.799	1.264	1.366	95.2
(5,0.8,0.2)	10.725	4.738	4.345	94.0	1.165	1.235	1.096	90.6	−1.523	1.456	1.357	94.4
(5,0.5,0.5)	11.209	4.299	4.368	94.8	1.018	1.088	1.100	93.4	−1.665	1.335	1.361	95.0
(5,0.2,0.8)	11.830	3.824	4.339	95.0	0.833	0.924	1.089	95.4	−1.847	1.212	1.349	95.0
$NNMIB_{22}$	M_1: misspecified; M_2: correctly specified											
(3,0.8,0.2)	10.047	5.358	4.785	91.2	1.375	1.475	1.244	88.2	−1.311	1.625	1.481	92.2
(3,0.5,0.5)	10.015	5.546	5.165	92.2	1.401	1.564	1.388	89.8	−1.295	1.668	1.577	92.2
(3,0.2,0.8)	10.184	5.653	5.624	93.0	1.377	1.648	1.579	89.8	−1.332	1.684	1.686	93.0
(5,0.8,0.2)	10.192	5.190	4.674	93.0	1.337	1.407	1.217	91.8	−1.360	1.578	1.446	92.8

	$b_0 = 10.000$				$b_1 = 1.333$				$b_2 = -1.333$			
Method	**Est**[a]	**SD**[b]	**SE**[c]	**CR**[d]	**Est**	**SD**	**SE**	**CR**	**Est**	**SD**	**SE**	**CR**
(5,0.5,0.5)	10.174	5.272	4.909	92.6	1.361	1.471	1.316	90.8	−1.350	1.596	1.504	93.4
(5,0.2,0.8)	10.368	5.227	5.154	92.8	1.337	1.510	1.442	91.8	−1.406	1.569	1.554	92.8

[a] Average of 500 point estimates.

[b] Empirical standard deviation of 500 point estimates.

[c] Average of 500 estimated standard errors.

[d] Coverage rate of 500, 95% confidence intervals.

[e] (NN, w_1, w_2).

Table 5

UDCA study: Regression analysis for the size of the largest baseline adenoma

	25(OH)D		Male		Age	
Method	**Est[a] (SE[b])**	***p*** **[c]**	**Est (SE)**	***p***	**Est (SE)**	***p***
CC	−0.042 (0.025)	0.096	−1.038 (0.483)	0.032	−0.046 (0.027)	0.089
IPW_{DR}	−0.046 (0.052)	0.376	−0.799 (0.893)	0.371	−0.018 (0.026)	0.489
NNMIB						
(3,0.8,0.2)	−0.046 (0.021)	0.028	−0.798 (0.352)	0.023	−0.018 (0.019)	0.343
(3,0.5,0.5)	−0.045 (0.020)	0.024	−0.806 (0.352)	0.022	−0.018 (0.019)	0.343
(3,0.2,0.8)	−0.039 (0.025)	0.119	−0.821 (0.355)	0.021	−0.018 (0.019)	0.343
(5,0.8,0.2)	−0.044 (0.026)	0.091	−0.802 (0.352)	0.023	−0.017 (0.019)	0.371
(5,0.5,0.5)	−0.043 (0.021)	0.041	−0.799 (0.350)	0.022	−0.018 (0.019)	0.343
(5,0.2,0.8)	−0.037 (0.019)	0.051	−0.816 (0.351)	0.020	−0.017 (0.019)	0.371

[a] Estimate of regression coefficient.

[b] Estimate of standard error.

[c] *p*-Value.