A nonparametric approach to weighted estimating equations for regression analysis with missing covariates

Peer-reviewed author version

CREEMERS, An; AERTS, Marc; HENS, Niel & MOLENBERGHS, Geert (2012) A nonparametric approach to weighted estimating equations for regression analysis with missing covariates. In: COMPUTATIONAL STATISTICS & DATA ANALYSIS, 56 (1), p. 100-113.

DOI: 10.1016/j.csda.2011.06.013
Handle: http://hdl.handle.net/1942/14374

Reference (Published version):
Creemers, An; Aerts, Marc; Hens, Niel & Molenberghs, Geert(2012) A nonparametric approach to weighted estimating equations for regression analysis with missing covariates. In: COMPUTATIONAL STATISTICS & DATA ANALYSIS, 56 (1), p. 100-113

DOI: 10.1016/j.csda.2011.06.013
Handle: http://hdl.handle.net/1942/14374

A Nonparametric Approach to Weighted Estimating Equations for Regression Analysis with Missing Covariates

An Creemers[1a], Marc Aerts[a], Niel Hens[a,b], Geert Molenberghs[a,c]

[a] *I-Biostat, Universiteit Hasselt, B-3590 Diepenbeek, Belgium*

[b] *Centre for Health Economics Research and Modeling Infectious Diseases (CHERMID), Centre for the Evaluation of Vaccination (WHO Collaborating Centre), Vaccine & Infectious Disease Institute, University of Antwerp, Antwerp, Belgium*

[c] *I-Biostat, Katholieke Universiteit Leuven, B-3000 Leuven, Belgium*

Abstract

Missing data often occur in regression analysis. Imputation, weighting, direct likelihood, and Bayesian inference are typical approaches for missing data analysis. The focus is on missing covariate data, a common complication in the analysis of sample surveys and clinical trials. A key quantity when applying weighted estimators is the mean score contribution of observations with missing covariate(s), conditional on the observed covariates. This mean score can be estimated parametrically or nonparametrically by its empirical average using the complete-case data in case of repeated values of the observed covariates, typically assuming categorical or categorized covariates. A nonparametric kernel based estimator is proposed for this mean score, allowing the full exploitation of the continuous nature of the covariates. The performance of the kernel based method is compared to that of a complete case analysis, inverse probability weighting, doubly robust estimators and multiple imputation, through simulations.

Keywords: Missing Covariates, Weighted Estimating Equations, Doubly Robustness, Mean Score Estimation, Kernel Weights

1. Introduction

Missing covariate data is a frequently encountered complication in the application of regression models. Following the taxonomy of [16], the missing data mechanism is classified as missing completely at random (MCAR), missing at random (MAR) or missing not at random (MNAR). In what follows, we will assume MAR. Naively excluding subjects

[1] Correspondence author: An Creemers, Universiteit Hasselt, Faculteit Wetenschappen, Agoralaan Gebouw D, 3590 Diepenbeek, e-mail: an.creemers@uhasselt.be, Phone: +32-11-26-8294, Fax:+32-11-26-8299

with missing covariates (so-called complete case analysis, CC) is known to possibly lead to highly inefficient estimates and, moreover, if the missing data mechanism is not completely at random, such CC-estimates can be biased. It is nowadays well recognized that the incorporation of the partially incomplete data into the analysis is a necessary and worthwhile effort to increase efficiency and reduce bias. Multiple imputation (MI, [9]) results in improved and consistent estimates, provided the imputation model is correct. Weighting by the inverse probability (IPW, [14]) of observing complete data on a unit is a conceptually simple alternative approach, involving fewer modelling assumptions but still inefficient and sensitive to the choice of the weighting model. The so-called doubly robust estimator (DR, [13]) has been developed to improve the performance of the IPW-estimators. [2] presented a nice intuitive review of these developments, contrasting these estimators from both a theoretical and a practical viewpoint. They concluded that the DR-estimator is an attractive alternative to multiple imputation. Results of MI are in general not robust to misspecification of the imputation model. DR-estimators recover a substantial proportion of the efficiency and are robust to either a wrongly specified conditional mean or a wrongly specified drop-out model but may be sensitive to correct specification of the weights and ordinarily result in efficiency loss. Although the DR-estimators exist in general, calculating them is not always straightforward. [23] discussed the numerical equivalence of the different estimators for categorical data.

There has also been some work in incorporating nonparametric components in the different approaches to deal with missing covariates. [20] investigated the properties of the IPW-estimator when the selection probabilities are estimated by kernel smoothers. [21] elaborated on this idea by also turning the (weighted) estimation equations into local linear weighted estimating equations, including additional kernel-based weights to estimate the mean parameter of interest as a smooth function of continuous covariates. Finally, [22] proposed kernel estimates for the selection probabilities and the other key quantity: the conditional mean score contribution of partially observed covariate(s).

In this paper we propose an alternative nonparametric kernel-based estimator for the conditional mean score of a partially observed case. This approach allows to estimate this quantity nonparametrically, without the need to categorize the continuous covariates, thus allowing to fully exploit the continuous nature of those covariates. It differs from the kernel estimate of [22] in using all observed variables of the partially observed cases. We compare the performance of the kernel-based method to that of a complete case analysis, inverse probability weighting, doubly robust estimators and multiple imputation, through simulations. We revisit the simulations in the review paper of [2].

The paper is organized as follows. In the next section we introduce notation while briefly defining and reviewing the CC, MI, IPW and DR methods. In Section 3, we propose our kernel based method to estimate the conditional mean score. In Section 4 we discuss the performance of the proposed method in comparison with other methods, based on simulations. Section 5 summarizes the main findings and concludes the paper.

2. Methods for Regression With Missing Covariate Data

Let y be the outcome variable of interest, x be the partially missing covariate and z the always observed covariate. The aim is to fit a parametric regression model $E_\Theta(y_i|x_i, z_i)$, where Θ is a vector of parameters. In case all observations $\{(x_i, z_i, y_i)\}_{i=1}^n$ are fully observed, the parameter vector Θ can be consistently estimated by solving the (unweighted) estimating equations

$$\sum_{i=1}^n \psi_\Theta(y_i|x_i, z_i) = 0. \tag{1}$$

For generalized linear models with canonical link function, $\psi_\Theta(y_i|x_i, z_i) = D_i\{y_i - G(\Theta' D_i)\}$ where $\Theta = (\theta_0, \theta_1, \theta_2)$, $D_i = (1, x_i, z_i)'$ and G is the link function of interest. Although all methods can be formulated for vector-valued x_i and z_i, we simplify presentation to a scalar-valued x_i and z_i.

Let δ_i indicate whether x_i is observed or not:

$$\delta_i = \begin{cases} 1, & \text{if } x_i \text{ is observed,} \\ 0, & \text{otherwise.} \end{cases}$$

We assume the probability π_i of x_i being observed to depend on (z_i, y_i) but not on x_i, i.e.,

$$\pi_i = P(\delta_i = 1|x_i, z_i, y_i) = P(\delta_i = 1|z_i, y_i). \tag{2}$$

So we assume x_i to be missing at random (MAR), according to the taxonomy introduced by [9].

In case x_i is missing for some units ($\delta_i = 0$), a complete case (CC) analyis is based on the observed estimating equations

$$\sum_{i=1}^n \delta_i \psi_\Theta(y_i|x_i, z_i) = 0. \tag{3}$$

These CC-estimating equations are no longer guaranteed to have expectation 0 at $\Theta = \Theta^*$, Θ^* being the true value, and parameter estimates are no longer guaranteed to be consistent. So, next to being inefficient because of deleting observations, a CC-analysis can lead to biased estimates [9].

This defect of biased estimating equations can be fixed by using weighted estimating equations, with inverse probability weights $1/\pi_i$ (IPW-estimating equations; [5, 24]):

$$\sum_{i=1}^n \frac{\delta_i}{\pi_i} \psi_\Theta(y_i|x_i, z_i) = 0. \tag{4}$$

In this way, in the spirit of the Horvitz-Thompson estimator used in the analysis of survey data [7], fully observed cases with a low probability to be observed get more weight in

order to reconstruct the subpopulation of partially observed cases, not included in the analysis. Most often, especially in a missing data setting, the probabilities $\pi(z_i, y_i)$ are unknown. Equation (3) can then be solved using an estimator $\hat{\pi}(z_i, y_i)$ from, for example a logistic regression analysis. The drawback of inefficiency remains relative to likelihood-based estimates using all partially observed cases, but paradoxically it is more efficient to use estimated rather than true weights [15].

Alternatively, [12] suggested to replace the unobserved quantity $\psi_\Theta(y_i|x_i, z_i)$ in

$$\sum_{i=1}^{n}\{\delta_i\psi_\Theta(y_i|x_i, z_i) + (1-\delta_i)\psi_\Theta(y_i|x_i, z_i)\} = 0,$$

by the expected score

$$\phi(z_i, y_i) = E\{\psi_\Theta(y_i|x_i, z_i)|z_i, y_i\}.$$

The so-called mean-score estimator is based on solving equation

$$\sum_{i=1}^{n}\{\delta_i\psi_\Theta(y_i|x_i, z_i) + (1-\delta_i)\phi(z_i, y_i)\} = 0, \tag{5}$$

with $\phi(z_i, y_i)$ replaced by an estimator $\hat{\phi}(z_i, y_i)$.

The inverse probability weights, operating in equation (4) on the fully observed cases, can also be applied to the partially observed cases in the second term in (5). This leads to the estimating equations for the so-called doubly robust (DR)-estimator [13]:

$$\sum_{i=1}^{n}\left\{\frac{\delta_i}{\pi_i}\psi_\Theta(y_i|x_i, z_i) + (1-\frac{\delta_i}{\pi_i})\phi(z_i, y_i)\right\} = 0. \tag{6}$$

The DR estimating equations are asymptotically equivalent to the full data estimating score if at least one of the two components, $\pi(z_i, y_i)$ or $\phi(z_i, y_i)$, is correctly specified. In case z and y are categorical, these components can be estimated nonparametrically by the empirical averages

$$\hat{\pi}(z_i, y_i) = \frac{\sum_{k=1}^{n}\delta_k I\{z_k = z_i, y_k = y_i\}}{\sum_{k=1}^{n} I\{z_k = z_i, y_k = y_i\}}, \quad i = 1, ..., n, \tag{7}$$

and

$$\hat{\phi}(z_i, y_i) = \frac{\sum_{k=1}^{n}\delta_k\psi_\Theta(y_k|x_k, z_k)I\{z_k = z_i, y_k = y_i\}}{\sum_{k=1}^{n}\delta_k I\{z_k = z_i, y_k = y_i\}}, \quad i = 1, ..., n. \tag{8}$$

In this categorical setting, [23] showed that some of the estimators listed above are, next to being asymptotically equivalent, also numerically the same.

In this paper we focus on the nonparametric estimation of $\phi(z_i, y_i)$. [22] proposed the following estimator

$$\hat{\phi}_{WW}(z_i, y_i) = \frac{\sum_{k=1}^{n}\delta_k\psi_\Theta(y_k|x_k, z_k)K_h\{z_k - z_i, y_k - y_i\}}{\sum_{k=1}^{n}\delta_k K_h\{z_k - z_i, y_k - y_i\}}, \quad \text{if } \delta_i = 0 \text{ for } i = 1, ..., n, \tag{9}$$

where K is a (multivariate) kernel function and $K_h(\cdot) = K(\cdot/h)$ where h is a (multivariate) smoothing parameter. The estimator $\hat{\phi}_{WW}(z_i, y_i)$ is used to replace the unobservable quantities $\psi(y_i|x_i, z_i)$ for those cases for which (z_i, y_i) is observed, but x_i is not (i.e. $\delta_i = 0$). A disadvantage of estimator (9) is that the observed part (z_i, y_i) is only used indirectly. In the next section, we propose an alternative estimator, which allows to keep that part of the data intact.

Another way to deal with missing covariate data is the use of multiple imputation [17]. Each missing value is then replaced by a vector of M imputed values. These values are obtained by randomly drawing from the predictive distribution of the missing values given the observed values. Then next, each data set completed by imputation is analyzed; afterwards, the obtained estimates for each of the datasets are combined into a single estimate. Let θ be a parameter of interest in our regression model, and $\hat{\theta}_m$ and W_m the estimated parameter and its associated variance for imputed dataset m, $m = 1, 2, \ldots, M$. Then, the combined estimate is

$$\hat{\theta}_{MI} = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_m. \tag{10}$$

The total variability associated with this estimate contains two components: the average within-imputation variance W and the between-imputation component B:

$$W = \frac{1}{M} \sum_{m=1}^{M} W_m, \qquad B = \frac{1}{M-1} \sum_{m=1}^{M} (\hat{\theta}_m - \hat{\theta}_{MI})^2. \tag{11}$$

The total variability associated with $\hat{\theta}_{MI}$ is

$$T = W + \frac{M+1}{M} B. \tag{12}$$

Under MAR and given that the imputation model is correct, multiple imputation provides consistent parameter estimates. However, when the imputation model is wrongly specified, parameter estimates might be biased.
In many applications, 3-5 imputations are sufficient to obtain excellent results. [11] state that, after a few imputations, gains rapidly diminish and thus in most situations, there simply is little advantage to producing and analyzing more than a few imputed datasets. Therefore, in our simulation studies, we will use 5 imputations (thus, M=5).

3. A Nonparametric Mean-Score Estimator

In this section, we propose the use of a new nonparametric mean score estimator. Let us start from the second term in the estimating equation (5). Estimator (9) for the mean score for the partially observed data ($\delta = 0, x = \cdot, z = z_i, y = y_i$) does not use the available

information on the (z, y) variables, which potentially constitutes a substantial part of the data. We propose the following estimator:

$$\hat{\phi}(z_i, y_i) = \frac{\sum_{k=1}^n \delta_k \psi_\Theta(y_i|x_k, z_i) K_h\{z_k - z_i, y_k - y_i\}}{\sum_{k=1}^n \delta_k K_h\{z_k - z_i, y_k - y_i\}}, \quad \text{if } \delta_i = 0 \text{ for } i = 1, ..., n. \tag{13}$$

The theoretical motivation for estimator (13) is as follows. Consider a linear model with normal homoscedastic error structure, for which $\psi_\Theta(y_i|x_i, z_i) = D_i\{y_i - \Theta' D_i\} = (1, x_i, z_i)'\{y_i - (\theta_0 + \theta_1 x_i + \theta_2 z_i)\}$, such as, for instance, the intercept score function

$$E\{\psi_\Theta^{(\text{int})}(y_i|x_i, z_i)|z_i, y_i\} = y_i - \theta_0 - E\{\theta_1 x_i|z_i, y_i\} - \theta_2 z_i, \tag{14}$$

can be estimated by the kernel estimate ($E\{\theta_1 x_i|z_i, y_i\} = E\{\theta_1 x_i|z_i, y_i, \delta_i = 1\}$, given that MAR implies that x_i and δ_i are independent, given z_i and y_i),

$$\begin{aligned}
\widehat{E}\{\psi_\Theta^{(\text{int})}(y_i|x_i, z_i)|z_i, y_i\} &= y_i - \theta_0 - \frac{\sum_{k=1}^n \delta_k \theta_1 x_k K_h\{z_k - z_i, y_k - y_i\}}{\sum_{k=1}^n \delta_k K_h\{z_k - z_i, y_k - y_i\}} - \theta_2 z_i \\
&= \frac{\sum_{k=1}^n \delta_k \{y_i - \theta_0 - \theta_1 x_k - \theta_2 z_i\} K_h\{z_k - z_i, y_k - y_i\}}{\sum_{k=1}^n \delta_k K_h\{z_k - z_i, y_k - y_i\}} \\
&= \frac{\sum_{k=1}^n \delta_k \psi_\Theta^{(\text{int})}(y_i|x_k, z_i) K_h\{z_k - z_i, y_k - y_i\}}{\sum_{k=1}^n \delta_k K_h\{z_k - z_i, y_k - y_i\}}.
\end{aligned}$$

This also holds for the other normal score functions as well as for general score functions, because the conditional distribution $F_{x|z_i,y_i}(x)$ in the mean score

$$E\{\psi_\Theta(y_i|x_i, z_i)|z_i, y_i\} = \int_{-\infty}^{\infty} \psi_\Theta(y_i|x, z_i) dF_{x|z_i,y_i}(x), \tag{15}$$

equals $F_{x|z_i,y_i,\delta_i=1}(x)$ and can be estimated consistently by (see, e.g., [1])

$$\widehat{F}_{x|z_i,y_i,\delta_i=1}(x) = \frac{\sum_{k=1}^n \delta_k I\{x_k \leq x\} K_h\{z_k - z_i, y_k - y_i\}}{\sum_{k=1}^n \delta_k K_h\{z_k - z_i, y_k - y_i\}}. \tag{16}$$

Plugging estimator (16) in (15) leads to our estimator (13), which consequently consistently estimates the mean score $E\{\psi_\Theta(y_i|x_i, z_i)|z_i, y_i\}$. Defining the classical Nadaraya-Watson weights as $w_k(z, y) = K_h\{z_k - z, y_k - y\}/\sum_{j=1}^n K_h\{z_j - z, y_j - y\}$ and $\hat{\pi}(z, y) = \sum_{k=1}^n w_k(z, y)\delta_k$, our estimator (13) can be reformulated as

$$\hat{\phi}(z_i, y_i) = \sum_{k=1}^n \delta_k w_k(z_i, y_i) \psi_\Theta(y_i|x_k, z_i)/\hat{\pi}(z_i, y_i). \tag{17}$$

The estimator can be inserted in estimating equation (5) as well as in (6), leading to two alternative estimators for the vector of parameters Θ of interest. Plugging (17) into (5) (or analogously into (6)), leads to the following weighted estimating equations

$$\sum_{i,j=1}^n w_{ij} \psi_\Theta(y_i|x_j, z_i) = 0, \tag{18}$$

with weights (δ_{ij} denoting Kronecker delta)

$$w_{ij} = \delta_j \left\{ \delta_{ij} + (1-\delta_i) \frac{w_j(z_i, y_i)}{\hat{\pi}(z_i, y_i)} \right\},$$

and based on the $n_c = \sum_{i=1}^n \delta_i$ fully observed data $\{(x_i, y_i, z_i, \delta_i = 1)\}_{i=1}^n$ and $n_c \times (n - n_c)$ partly nonparametrically imputed data $\{(x_j, \delta_j = 1, y_i, z_i, \delta_i = 0)\}_{i,j=1}^n$.

Each of the $n - n_c$ observations $(\cdot, y_i, z_i)$ with missing x_i-value is replaced by n_c imputed observations (x_j, y_i, z_i), with $\delta_j = 1$ for $j = 1, ..., n$. These latter imputed data get weights $w_j(z_i, y_i)/\hat{\pi}(z_i, y_i)$, being larger for imputations x_j with corresponding values (z_j, y_j) closer to the (z_i, y_i)-value associated with the missing x_i, and being larger in an area with larger chance of having missing observations. Indeed, $\hat{\pi}(z_i, y_i)$ is a nonparametric kernel estimator for the probability to observe x_i at (y_i, z_i). The effect of $\hat{\pi}(z_i, y_i)$ on the weights stresses the importance of the few available but highly informative x observations in a 'sparse' area with a lot of missingness.

Now, 14 can be rewritten as

$$E\{\psi_\Theta^{(\text{int})}(y_i|x_i, z_i)|z_i, y_i\} = y_i - \theta_0 - \theta_1 E\{x_i|z_i, y_i\} - \theta_2 z_i,$$

and thus in the linear case, the nonparametric mean score estimator is imputing the unobserved x with a weighted mean of the observed x variables, with weights according to the distance between the observed part of that observation and the completers. For a nonlinear model however, this property does not hold any more.

Note that the n_c imputed values (x_j, y_i, z_i) for the single original observation $(\cdot, y_i, z_i)$, all keep the observed part (y_i, z_i) intact and get a total weight $\sum_{j=1}^n \delta_j w_j(z_i, y_i)/\hat{\pi}(z_i, y_i) = 1$, preserving the total contribution of a single observation. This also implies that the total sum of the weights w_{ij} equals the total sample size n, as $\sum_{i,j=1}^n w_{ij} = \sum_i^n \delta_i + \sum_i^n (1 - \delta_i)(\sum_{j=1}^n \delta_j w_j(z_i, y_i))/\hat{\pi}(z_i, y_i) = n$.

In general, the choice of a particular kernel K for the weights $K_h\{z_k - z_i, y_k - y_i\}$ is less crucial for kernel-based methods, as compared to the choice of the bandwidth(s) h [6]. It can be any multivariate density K with mean centered at the origin and with h referring to the covariance matrix, such as a bivariate normal density centered at $(0, 0)$ in case z and y are univariate variables. The choice of the covariance structure and the various variances, which act as smoothing parameters, is not straightforward. One reasonable simplifying option is to take an independence covariance structure such that $K_h\{z_k - z_i, y_k - y_i\} = K_{h_z}\{z_k - z_i\} \times K_{h_y}\{y_k - y_i\}$. In addition, both variance parameters can be taken identical $h_y = h_z$. Another simplifying option is to first take the Euclidean distance $||z_k - z_i, y_k - y_i||$ and next a one-dimensional density K_h with mean 0 and variance h, such as a (range truncated) normal density, Epanechnikov density, or uniform density around 0. The choice of the bandwidth(s) (h_z, h_y) or h is not straightforward. Various options to automatically select the best bandwidth were proposed in literature: cross-validation, penalizing functions and the plug-in method (see e.g. Hardle 1990). As discussed in the next paragraph we prefer to use the nearest neighbor method, by taking the k "nearest" neighbors (see e.g. [18, 10]).

Using estimator (13) in (5) and (6) leads to estimating equations based on a multiply completed, imputed dataset with weights reflecting the importance of each single imputed dataset. In practice, this multiply augmented dataset can be quite massive, given that the original number of different observations n increases up to a maximum of $n_c + (n - n_c)n_c = n_c(1 + n - n_c)$ different observations. Using kernel weights $w_j(z_i, y_i)$ with bounded support can however considerably reduce the imputed dataset. In this case, non-zero weights are given to datapoints which are 'close enough' to the partly observed datapoint. Another option to reduce the number of different observations is to use only a fraction $[\alpha n_c]$ $(0 < \alpha < 1)$ nearest neighbors in the kernel weights. This leads to a total size of $n_c + (n - n_c)[\alpha n_c] = n + (n - n_c)([\alpha n_c] - 1)$. Taking only one nearest neighbor ($[\alpha n_c] = 1$) leads to a single imputed dataset with exactly n observations, and any additional neighbor would generate $n - n_c$ extra observations (albeit appropriately downweighted).
Various simulation studies were done with different sample sizes and missingness percentages. Some of these simulation studies are shown in the next section. They show that using 3 nearest neighbors with a uniform kernel is an adequate choice, independent of the sample size and the missingness percentage. Employing this combination of 3 nearest neighbors and a uniform kernel, the difficulty of selecting a bandwidth is eliminated, since whatever bandwidth one is using, after normalizing the weights all three neighbors will have weight $\frac{1}{3}$.

In Figure 1, we illustrate the method of Wang & Wang and our new proposal with uniform kernel and 3 nearest neighbors. Consider the dataset shown in the upper left panel, with response variable y and a single covariate x. When some of the observations have missing x, the dataset as observed might look as in the upper right panel. The middle left and right panel illustrate the method of Wang & Wang and our new method respectively, focussing on 1 incompleter, shown with a transparent square. The method of Wang & Wang gives an extra weight to the 3 nearest neighbors, according to the distance between the y values of the incompleter and the neighbors. Higher weights are shown with bigger dots. Our new proposal is using these same three neighbors, but instead of giving these observations a higher weight, their x value is imputed in the incompleter. So, the transparent square is replaced with three 'virtual' points, with weights that sum to 1 (shown by crosses). Finally, in the lower panel, both methods are illustrated for the full dataset. In the case of Wang & Wang, a weighted complete dataset is obtained, while in the case of the new proposal a virtual imputed dataset is obtained. As mentioned earlier, in the linear case, our new method coincides with imputing the unobserved x with a weighted mean of the observed x variables. In Figure 2, the so-obtained dataset is shown. The crosses are the imputed datapoints based on these weighted averages.

In the next section we will illustrate the performance of our estimator and compare the method with the other methods, CC, MI, IPW and DR.

4. Class Size Study Data

The simulation study was motivated by the analysis of the class size study, which was carried out by Peter Blatchford and colleagues at the Institute of Education. The study was set up to look at the effects of class size on educational achievement. Further details, and the initial findings, were described in [3]. Broadly speaking, after adjusting for variables reflecting social class, improvement in basic mathematics and English over the course of pupils' reception year declines as class sizes rise above 25. The study continued to follow up students as they progressed through primary school. Attrition (due to failure to renew contact with pupils) and missing covariates (due to incomplete information from pupils and schools) are issues in the analysis of these data.
Analogously as in [2], we look into the properties of the various methods by simulating data with a similar structure to the data in the class size project. Specifically, we generate data from a four-dimensional normal distribution with mean and covariance equal to the mean and covariance of the actual data for the variables post-reception literacy score (which increases with literacy), pre-reception literacy score (which increases with literacy), eligibility for free school meals (which is coded 1 for eligible) and gender (which is coded 1 for girls). The mean μ and variance Σ of this distribution, with variables in the above order, were estimated from the data.

$$(\text{Post,Pre,Meals,Gender}) \sim N_4\left(\mu, \mathbf{\Sigma}\right), \tag{19}$$

with

$$\mu = \begin{pmatrix} 0.0201 \\ 0.0230 \\ 0.1668 \\ 0.4816 \end{pmatrix} \qquad \text{and} \qquad \mathbf{\Sigma} = \begin{pmatrix} 1.0094 & 0.6418 & -0.0710 & 0.0543 \\ 0.6418 & 0.9506 & -0.0877 & 0.04950 \\ -0.0710 & -0.0877 & 0.1390 & 0.0032 \\ 0.0543 & 0.04950 & 0.0032 & 0.2497 \end{pmatrix}. \tag{20}$$

Note that, as we simulated data from a normal distribution, all simulated data are continuous, also the variables Meals and Gender, which were discrete in the original dataset, but taken to be continuous in the simulation setting of [2]. The variable Post is used as the response variable, the other three as covariates. The distribution of Post, given Pre (P), Meals (M) and Gender (G) becomes

$$\text{Post}|\text{P,M,G} \sim N\left[\mu(\text{P,M,G}), 1.0094\right], \tag{21}$$

with

$$\mu(\text{P,M,G}) = -0.0214 + 0.6618\text{P} - 0.0952\text{M} + 0.0875\text{G}. \tag{22}$$

4.1. Missing observations in one covariate

To compare the different methods, some simulation studies were set up. In Scenario 1, data were generated according to the 4-variate distribution in (19) with sample size 4873, the sample size in the original dataset. As was also done in [2], in every run, some of the pre-reception literacy scores were set missing according to a Bernoulli trial with probability:

$$1 - \pi = [1 + \exp(0.5 + 0.5 \times \text{Post} + G - M)]^{-1} . \tag{23}$$

This yielded an average missingness percentage of around 30% over 200 runs. In each of the runs, a full case analysis (i.e. the analysis on the original data, before introducing the missingness), a complete case analysis (CC), multiple imputation (MI), inverse probability weighting (IPW), doubly robust estimation (DR), the method of Wang and Wang (WW) and our new proposal (based on equation 5) were fitted to the simulated data. For the latter two methods, a uniform kernel with 3 nearest neighbors is used. The methods are compared by the parameter estimates and by the overall mean squared error (MSE), defined as

$$MSE = \frac{1}{m} \sum_{j=1}^{m} \frac{1}{n} \sum_{i=1}^{n} \left(\hat{\mu}(P_i^{(j)}, M_i^{(j)}, G_i^{(j)}) - \mu(P_i^{(j)}, M_i^{(j)}, G_i^{(j)}) \right)^2 , \tag{24}$$

where $m = 200$, the number of simulated datasets, and n=4873, the number of observations per run. Since we simulate from a multivariate normal model, and the multiple imputation assumes this model, we expect the MI to have the best results.
Parameter estimates, sample variances of the parameter estimates and overall mean squared errors for the different methods are presented in Table 1. A nonparametric bootstrap is used to estimate the variance of the parameter estimates for all methods. For the IPW and the DR, three different results are shown: in the first case, the correctly specified parametric function is used for the missingness model, in the second case, a misspecified function with only effects of gender and meals are included, while in the third case a nonparametric generalized additive model (GAM) with spline effects for all three variables is assumed. For the MI, the correct imputation model is used. Almost all methods behave well with parameter estimates close to the true ones and an overall MSE around 0.7. However, the CC and the IPW with incorrectly specified weights perform poorly. They both have biased parameter estimates and very high mean squared errors. We also remark the bootstrap estimation of the variance behaves nice for most methods, including our new method. The empirically observed variance and the bootstrap variance are close to each other, although a bit conservative, for all parameters. In Scenario 1, results from our new proposal are thus comparable with the other methods. We will next describe a second scenario, based on the same dataset, but with a higher missingness percentage in combination with a lower sample size.

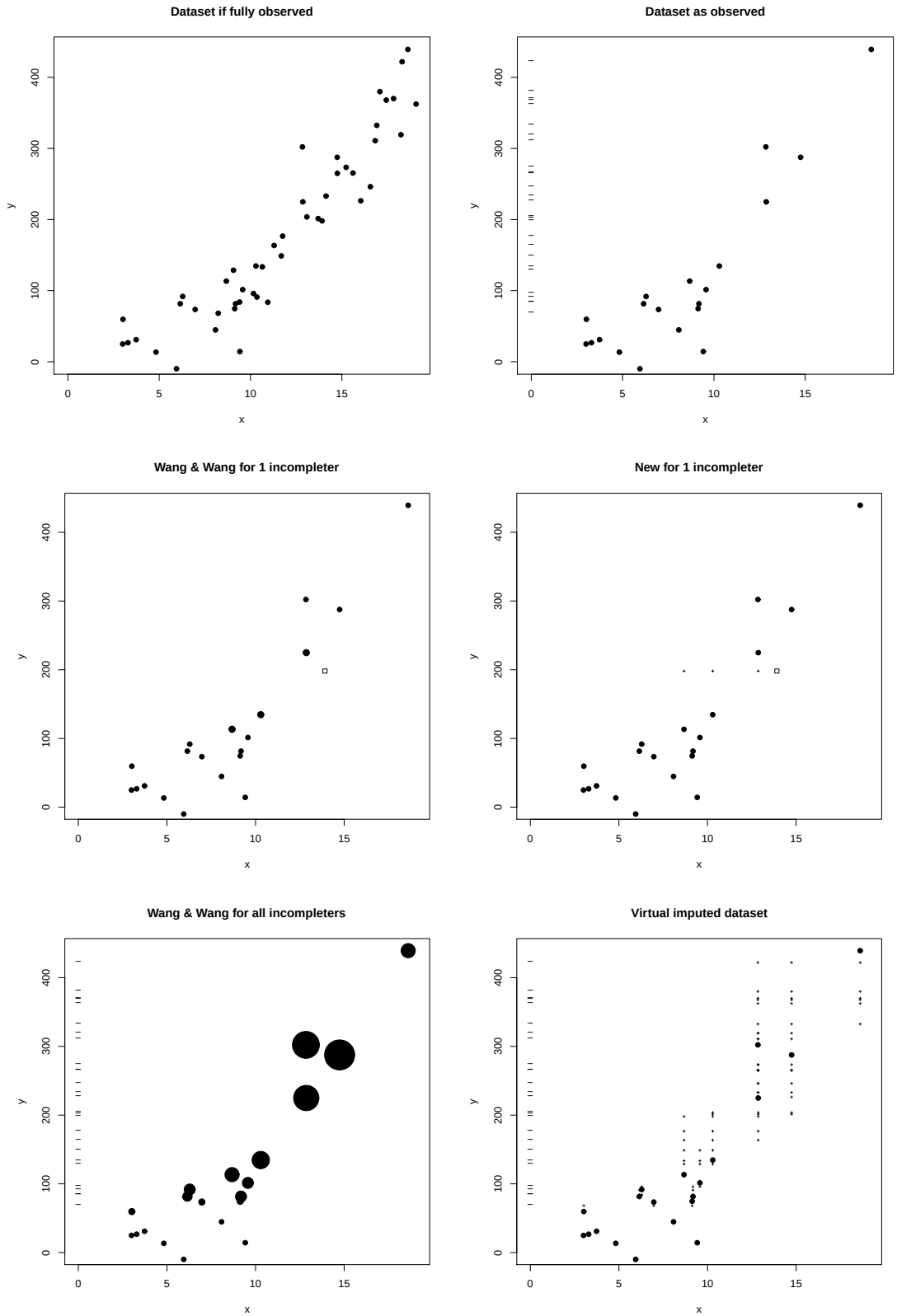

Figure 1: Illustration of our new proposal versus the method of Wang & Wang: dataset if fully observed (upper left), dataset as observed (upper right), method of Wang & Wang focussing on one incompleter (middle left), new proposal focussing on one incompleter (middle right), method of Wang & Wang (lower left) and new proposal (lower right). Both methods are based on the 3 nearest neighbors with uniform kernel function.

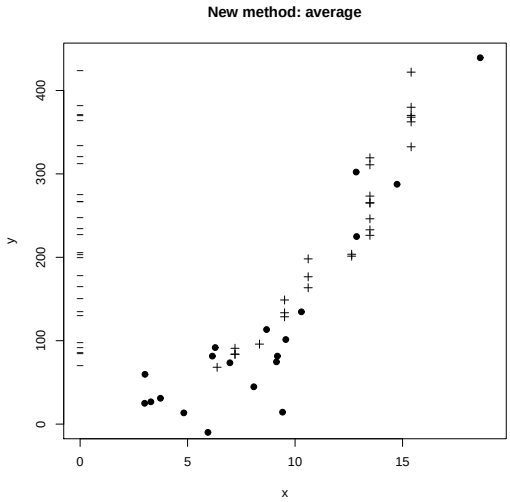

Figure 2: New proposal in the linear case: the dataset that is obtained by imputing the unobserved x with a weighted mean of the observed x variables. The crosses are the imputed datapoints based on these weighted averages.

Table 1: Results Scenario 1: Mean Parameter Estimates (AvEst), Sample Variances (VarEst) and Mean Estimated Variances using nonparametric bootstrap (VarBoot) for the different methods in the Class-Size Simulation Study, analogue to [2]. The last column is the overall Mean Squared Error.

Method	Intercept			Pre			Meals			Gender			MSE ($\times 10^{-4}$)
	AvEst	VarEst (10^{-2})	VarBoot (10^{-2})	AvEst	VarEst (10^{-2})	VarBoot (10^{-2})	AvEst	VarEst (10^{-2})	VarBoot (10^{-2})	AvEst	VarEst (10^{-2})	VarBoot (10^{-2})	
True	-0.0214			0.6618			-0.0952			0.0875			
Full	-0.0237	0.2372	0.2245	0.6606	0.1380	0.1299	-0.0982	0.8447	0.8453	0.0905	0.5727	0.5718	5.198
CC	0.0862	0.4133	0.4218	0.6441	0.1992	0.2001	-0.0445	1.2649	1.2851	0.0355	0.7330	0.7319	409.949
MI	-0.0227	0.3143	0.3033	0.6603	0.1742	0.1774	-0.0982	1.0534	1.0617	0.0893	0.6245	0.6300	7.102
IPW correct	-0.0226	0.4244	0.4529	0.6615	0.2538	0.2543	-0.0977	1.4904	1.4731	0.0884	0.9111	1.2133	7.796
IPW wrong	0.0872	0.4359	0.4455	0.6436	0.2265	0.2381	-0.0422	1.4076	1.3814	0.0337	0.8129	0.8046	413.908
IPW nonpara	-0.0221	0.3435	0.3889	0.6605	0.1992	0.2371	-0.0973	1.1920	1.1929	0.0881	0.7204	0.7311	7.092
DR correct	-0.0228	0.3212	0.3241	0.6609	0.1975	0.1924	-0.0979	1.1612	1.1573	0.0891	0.6630	0.6783	7.077
DR wrong	-0.0230	0.3194	0.3265	0.6609	0.1988	0.2047	-0.0976	1.1482	1.1516	0.0894	0.6482	0.6551	7.038
DR nonpara	-0.0228	0.3214	0.3209	0.6609	0.1957	0.2164	-0.0977	1.1492	1.1592	0.0891	0.6624	0.6843	7.086
WangWang	-0.0209	0.3407	0.3604	0.6570	0.2010	0.1901	-0.0999	1.2170	1.2007	0.0861	0.6814	0.6506	7.520
New	-0.0249	0.3249	0.3333	0.6588	0.2057	0.2213	-0.1074	1.1691	1.1789	0.0926	0.6548	0.6677	7.243

Again, in each run, data were generated according to the 4-variate distribution (19). A sample size of 400 was used instead of 4873 to reduce computation time. In every run, some of the pre-reception literacy scores were set missing according to the following rule:

$$1 - \pi = [1 + \exp(0.2 + \text{Post} - 1.5G + 0.1M)]^{-1} \,. \tag{25}$$

This yielded an average missingness percentage of around 60% over the 200 runs.
The results of Scenario 2 are summarized in Table 2 and show some difference. First, as expected and as in Scenario 1, the CC analysis and the IPW with incorrectly specified weights show again biased parameter estimates and a high MSE compared to the other methods. IPW with the correctly specified models for the missingness function has parameter estimates close to the true ones, and has highly reduced mean squared error. For the DR estimates, it is clearly seen that parameter estimates are robust against misspecification of the missingness model. Since only the missingness model was misspecified, and the mean function was the correct one, The DR estimating equations are asymptotically equivalent to the full data estimating score, and thus provide good results.The parameter estimates in WW seem to be biased, but the overall MSE is comparable to the IPW from a correctly specified missingness model. Further, the imputation methods (MI and our new method) behave very good in terms of overall MSE and are of the same order as the doubly robust MSEs.

Table 2: Results Scenario 2: Mean Parameter Estimates (AvEst), Sample Variances (VarEst) and Mean Estimated Variances using nonparametric bootstrap (VarBoot) for the different methods in the Class-Size Simulation Study with missing Pre. The last column is the overall Mean Squared Error.

Method	Intercept			Pre			Meals			Gender			MSE ($\times 10^{-2}$)
	AvEst	VarEst	VarBoot	AvEst	VarEst	VarBoot	AvEst	VarEst	VarBoot	AvEst	VarEst	VarBoot	
True	-0.0214			0.6618			-0.0952			0.0875			
Full	-0.0214	0.0532	0.0552	0.6631	0.0363	0.0403	-0.0846	0.1055	0.1050	0.0816	0.0716	0.0765	0.4872
CC	0.2161	0.0753	0.0742	0.5900	0.0572	0.0629	-0.0702	0.1570	0.1579	0.2253	0.1175	0.1212	11.5500
MI	-0.0230	0.0705	0.0673	0.6612	0.0533	0.0571	-0.0810	0.1434	0.1392	0.0841	0.0987	0.1022	1.2050
IPW correct	-0.0121	0.0840	0.0859	0.6414	0.0858	0.0812	-0.0747	0.2241	0.2089	0.0899	0.1820	0.1561	2.6000
IPW wrong	0.2188	0.0766	0.0769	0.5875	0.0604	0.0655	-0.0649	0.1657	0.1659	0.2178	0.1229	0.1285	11.6500
IPW nonpara	-0.0051	0.0724	0.1531	0.6333	0.0751	0.1207	-0.0932	0.1848	0.2757	0.1128	0.1359	0.2144	1.8760
DR correct	-0.0186	0.0734	0.0748	0.6679	0.0614	0.0586	-0.0824	0.1595	0.1471	0.0741	0.1126	0.1041	1.2710
DR wrong	-0.0222	0.0690	0.0721	0.6639	0.0537	0.0538	-0.0789	0.1448	0.1390	0.0779	0.0990	0.0975	1.0210
DR nonpara	-0.0191	0.0717	0.0772	0.6677	0.0623	0.0605	-0.0791	0.1591	0.1519	0.0767	0.1087	0.1067	1.2580
WangWang	-0.0217	0.0735	0.0815	0.6025	0.0608	0.0725	-0.1343	0.1850	0.1948	0.1916	0.1232	0.1377	2.0560
New	-0.0275	0.0693	0.0764	0.6396	0.0592	0.0930	-0.1444	0.1357	0.1439	0.0832	0.0965	0.1070	1.1394

4.2. Missing observations in two covariates

Methods like the complete case analysis and inverse probability weighting are easy to implement, but also known to be biased and/or inefficient. The previous section illustrated that multiple imputation, doubly robust estimation and our new proposal behave well when dealing with one missing covariate. Although the doubly robust estimators exist more generally [19], it is not completely clear how to calculate them in settings where several covariates are missing [2]. Our new proposal can easily be generalized to the case of several missing covariates. In this section, we will study the behavior of the different methods for the case of two incomplete covariates.

In Scenario 3, again the 4 variables were generated according to the distribution in (20) and with sample size 400. In every run, missingness was introduced in both the variables Pre-reception literacy score and Meals, according to the following rules:

$$\begin{cases} 1-\pi_1 = [1+\exp(0.2+\text{Post}+1.5G+0.4M)]^{-1}\,, \\ 1-\pi_2 = [1+\exp(0.2+1.7\text{Post}+0.6P+0.4G)]^{-1}\,, \end{cases}$$

where $1-\pi_1$ is the probability that variable Pre is missing, while $1-\pi_2$ is the probability that Meals is not observed, given that Pre is observed. In this way, the variables Meals and Pre will not be missing at the same time, and thus the probability that Meals is missing does only depend on the observed values of Pre. Therefore, the missing data mechanism is MAR. This yielded an average missingness percentage of 56%. In variables Pre and Meals, the average missingness percentage was 32% and 24%, respectively. While the complete case analysis will throw away 56% of the information in the dataset and most other methods will use this 56% only implicitly, our new proposal and the multiple imputation will use all available information explicitly in the analysis.

To each of the 200 simulated datasets, a full case analysis, a CC analysis, MI, IPW, WW, and our new proposal were fitted. As one of the reviewers suggested, it might be interesting to look at multiple imputation when the imputed values are obtained from a nonparametric estimator of the predictive distribution of the missing values. Therefore, we included such an estimator in the analysis. Using generalized additive models, the predictive distribution of the missing values is estimated and 5 imputed values are obtained. Parameter estimates, sample variances, mean estimated variances using nonparametric bootstrap and overall MSEs of all considered methods are summarized in Table 3. As before, the CC analysis and the IPW with incorrectly specified weights exhibit biased estimates and high MSE. The MSEs from the other methods are substantially lower. Multiple imputation behaves very well and has the smallest MSE, almost as small as the full case analysis. This was to be expected, given that the data are simulated from a multivariate normal distribution, assumed by the multiple imputation method. The nonparametric imputation still behaves well, but has larger MSE compared to the parametric imputation. The larger MSE is due to the more biased parameter estimates, variances are still small for the nonparametric version. The MSE of the nonparametric imputation is comparable to that of the IPW with nonparametric weights. Our new proposal does not make any assumption about the distribution, but it also behaves very good in terms of MSE. Compared to the method of Wang and Wang, MSE reduces with 1.3373×10^{-2}.

As expected, the advantage of using all available information seems to be more clear in the missing-more-covariate case. Whereas only the completely observed records contribute to the CC and IPW analyses, also the partly observed records with one of the two covariates observed contribute to the estimation procedure. Thus, compared to the missing-one-covariate case, more information can be gained in the missing-two-covariates case and it therefore might result in a stronger reduction in MSE compared to methods that only use the fully observed observations explicitly.
The behavior of DR estimates could not be investigated in Scenario 3, since no extension to the missing-more-covariates case is available for this method.

Table 3: Results Scenario 3: Mean Parameter Estimates (AvEst), Sample Variances (VarEst) and Mean Estimated Variances using nonparametric bootstrap (VarBoot) for the different methods in the Class-Size Simulation Study with missing Pre and Meals. The last column is the overall Mean Squared Error.

Method	Intercept			Pre P			Meals M			Gender G			MSE ($\times 10^{-2}$)
	AvEst	VarEst	VarBoot	AvEst	VarEst	VarBoot	AvEst	VarEst	VarBoot	AvEst	VarEst	VarBoot	
True	-0.0214			0.6618			-0.0952			0.0875			
Full	-0.0289	0.0766	0.0550	0.6629	0.0527	0.0402	-0.0990	0.1520	0.1045	0.0938	0.1022	0.0758	1.0838
CC	0.4553	0.1381	0.0895	0.4827	0.0857	0.0632	-0.0980	0.1771	0.1413	-0.0319	0.1627	0.1086	23.2139
MI	-0.0732	0.0751	0.0736	0.6482	0.0511	0.0540	-0.1203	0.1322	0.1443	0.1287	0.0900	0.0923	1.0920
MI nonpara	-0.1786	0.0661	0.0678	0.4392	0.0424	0.0475	-0.3563	0.1438	0.1275	0.2452	0.0864	0.0932	7.3970
IPW correct	0.0603	0.1786	0.1253	0.5471	0.1315	0.0946	0.0876	0.3170	0.2259	0.0364	0.2317	0.1580	7.2872
IPW wrong	0.4653	0.1376	0.0907	0.4789	0.0855	0.0644	-0.1042	0.1848	0.1444	-0.0441	0.1652	0.1128	23.7179
IPW nonpara	0.1145	0.1728	0.1273	0.5380	0.1170	0.0903	-0.0805	0.2874	0.2242	0.0074	0.1937	0.1424	7.2161
WangWang	0.0723	0.0947	0.1021	0.5473	0.0817	0.0819	-0.1304	0.1794	0.2041	0.0206	0.1304	0.1403	3.3906
New	-0.1073	0.0732	0.0829	0.6019	0.0727	0.0776	-0.1884	0.1378	0.1586	0.1680	0.0925	0.1025	2.0533

4.3. Missing observations in three covariates

In this section, a fourth Scenario will be investigated. We will illustrate the behavior of the different methods when three covariates have missing observations. Therefore, the original study was extended with another covariate, which will be referred to as *Extra*. Data were simulated from the five-variate normal distribution

$$(\text{Post,Pre,Meals,Gender,Extra}) \sim N_5\left(\mu, \mathbf{\Sigma}\right), \tag{26}$$

with

$$\mu = \begin{pmatrix} 0.0201 \\ 0.0230 \\ 0.1668 \\ 0.4816 \\ 1.5327 \end{pmatrix} \quad \text{and} \quad \mathbf{\Sigma} = \begin{pmatrix} 1.0094 & 0.6418 & -0.0710 & 0.0543 & -0.3247 \\ 0.6418 & 0.9506 & -0.0877 & 0.04950 & -0.1572 \\ -0.0710 & -0.0877 & 0.1390 & 0.0032 & 0.1378 \\ 0.0543 & 0.04950 & 0.0032 & 0.2497 & -0.4876 \\ -0.3247 & -0.1572 & 0.1378 & -0.4876 & 6.8516 \end{pmatrix}, \tag{27}$$

and with sample size 400. Missingness was introduced in the three original covariates: Pre-reception literacy score, Meals and Gender. These variables were set missing with probabilities respectively

$$\begin{cases} 1-\pi_1 = \left[1+\exp(0.2+\text{Post}+1.5G+0.4M+E)\right]^{-1}, \\ 1-\pi_2 = \left[1+\exp(0.2+1.7\text{Post}+0.6P+0.4G+E)\right]^{-1}, \\ 1-\pi_3 = \left[1+\exp(-5+1.1\text{Post}+P+0.2M+2E)\right]^{-1}, \end{cases}$$

where $1-\pi_1$ is the probability that variable Pre is missing, $1-\pi_2$ is the probability that Meals is not observed, given that Pre is observed, and finally, $1-\pi_3$ is the probability that Gender is not observed, given both Pre and Meals are observed. In this way, only one covariate can be missing in the same observation, and thus MAR is guaranteed. This yielded an average total missingness percentage of 65%: 21% in Pre, 15% in Meals and 29% in Gender. As in the other Scenarios, a CC analysis, MI, IPW, WW and our new proposal were fitted. Parameter estimates, sample variances, mean estimated variances using nonparametric bootstrap and overall MSEs are summarized in Table 4. All methods, except the multiple imputation and the new proposal have high Mean Squared Errors. Multiple imputation behaves best in terms of MSE (3.8944×10^{-2}), followed by our new proposal (6.2015×10^{-2}).

Table 4: Results Scenario 4: Mean Parameter Estimates (AvEst), Sample Variances (VarEst) and Mean Estimated Variances using nonparametric bootstrap (VarBoot) for the different methods in the Class-Size Simulation Study with missing Pre and Meals and Gender. The last column is the overall Mean Squared Error.

Method	Intercept			Pre			Meals			Gender			Extra			MSE ($\times 10^{-2}$)
	AvEst	VarEst	VarBoot	AvEst	VarEst	VarBoot	AvEst	VarEst	VarBoot	AvEst	VarEst	VarBoot	AvEst	VarEst	VarBoot	
True	0.0448			0.6628			-0.0649			0.0309			-0.0287			
Full	0.0484	0.0958	0.0660	0.6615	0.0531	0.0403	-0.0671	0.1582	0.1051	0.0358	0.1164	0.0829	-0.0296	0.0238	0.0158	1.4617
CC	0.5656	0.2642	0.1908	0.5637	0.1009	0.0713	-0.0570	0.2761	0.1764	0.0091	0.2010	0.1371	-0.1159	0.0580	0.0408	27.0392
MI	-0.0482	0.0638	0.0721	0.5971	0.1412	0.1324	-0.1079	0.1537	0.1347	0.0329	0.0575	0.0708	0.0060	0.0174	0.0708	3.8944
IPW correct	0.1740	0.3926	0.2617	0.5706	0.1414	0.1015	-0.0942	0.3807	0.2637	0.0004	0.2900	0.2068	-0.0573	0.0806	0.0538	16.1286
IPW wrong	0.7499	0.3548	0.2554	0.5401	0.1412	0.0956	-0.0708	0.3412	0.2299	-0.0060	0.2387	0.1820	-0.1560	0.0790	0.0552	48.6663
IPW nonpara	0.2622	0.3691	0.2621	0.5684	0.1299	0.0951	-0.0928	0.3515	0.2524	0.0060	0.2527	0.1942	-0.0708	0.0751	0.0524	16.0922
WangWang	0.4278	0.2941	0.2787	0.5602	0.1121	0.1174	-0.1052	0.3121	0.3100	0.0142	0.2491	0.2416	-0.1143	0.0641	0.0608	19.6930
New	-0.1459	0.1027	0.1219	0.5117	0.0739	0.0805	-0.2182	0.1235	0.1466	0.1070	0.1386	0.1498	0.0217	0.0230	0.0265	6.2015

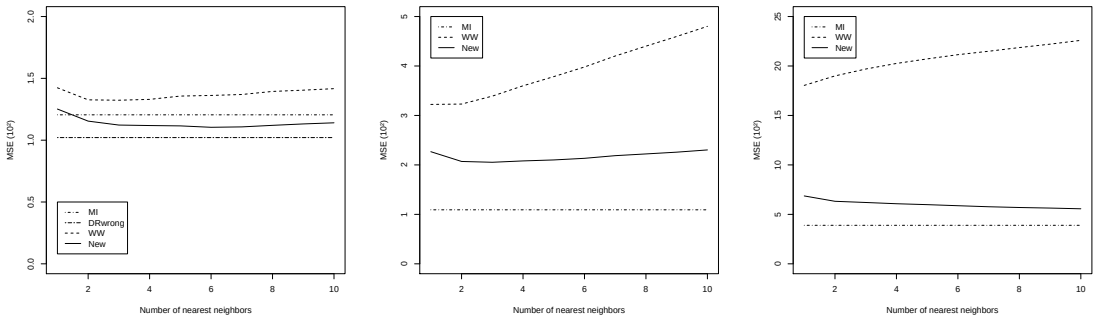

Figure 3: Sensitivity of new method and method of Wang and Wang to the choice of number of nearest neighbors. Left: 1 missing covariate (sample size 400), middle: 2 missing covariates, right: 3 missing covariates. MSE for Multiple Imputation and Doubly Robust Estimation are included for comparison.

4.4. Sensitivity of method to number of nearest neighbors

In previous simulation settings, 3 nearest neighbors were used with a uniform kernel function. However, the results might be sensitive to the number of nearest neighbors used in the analysis.

In Figure 3, the MSE is plotted as a function of the number of nearest neighbors for Scenarios 2, 3 and 4. The method of Wang and Wang seems to be more sensitive to the number of nearest neighbors, compared to the new method. For our new proposal, the dependence on the number of nearest neighbors is rather limited. In the case of 1 missing covariate, there is a minimum MSE when using 6 nearest neighbors, but the difference with 3 nearest neighbors is only 0.0181 (or a difference of 1.6%). In the case of 2 covariates with missing observations, the minimum MSE for the new proposal is reached for 3 nearest neighbors. In the case of 3 missing covariates, a bit more sensitivity can be detected: the MSE is decreasing with increasing number of nearest neighbors. The minimum is situated at 10 nearest neighbors, which gives a difference of 0.6373 (or 10.2%) with the 3 nearest neighbors.

These are only some illustrations of the sensitivity of the proposed method to the number of nearest neighbors. In general can be said that our new proposal is not too sensitive to the number of nearest neighbors. Although in some applications results could be better with another number of nearest neighbors, using 3 neighbors is an adequate choice.

4.5. Computational Remarks

As was already mentioned in Section 3, the virtual imputed dataset can be quiet massive. If the sample size is 4873 with a missingness percentage of 30%, like was the case in the first simulation study, the size of the virtual imputed dataset becomes 4,990,293 using all neighbors. When the missingness percentage increases to 60%, the size of the virtual imputed dataset would increase to 5,700,825. For this reason, it is useful to use only a fraction of the nearest neighbors to impute. In the simulation studies described before,

only 3 nearest neighbors were taken into account. This means, for the previous examples, that the virtual imputed datasets will have size 7,797 and 10,721, respectively, reducing computation time. In Table 5, the computation times (in seconds) of the various methods and for the different simulation studies are shown. The nonparametric versions of the IPW and DR take considerably more time than the others, which is mainly due to the nonparametric estimation of the missingness model. Further, doubly robustness, the method of Wang & Wang, and our new proposal take more time than the other methods, especially when the sample size is large. However, for smaller sample sizes the difference with the other methods becomes smaller, and for $n = 400$, the method of Wang & Wang and our new proposal are even faster than the doubly robust estimates. The missingness percentage seems to have less impact for our new proposal, the computation times for 40% and 60% missingness are more or less the same.

Table 5: Computation time (in seconds) for the different methods and different datasets

	One incomplete covariate			Two incomplete covariates
	$n = 4873$	$n = 400$		$n = 400$
	30% missing	30% missing	60% missing	56% missing
Full	0.05	0.04	0.02	0.03
CC	0.05	0.04	0.04	0.03
MI	0.68	0.53	0.53	1.10
IPW correct	0.26	0.14	0.15	0.13
IPW wrong	0.30	0.26	0.20	0.12
IPW nonpara	84.75	6.88	6.31	4.05
DR correct	5.56	5.21	4.05	/
DR wrong	5.56	4.06	4.11	/
DR nonpara	109.74	11.39	10.33	/
WangWang	19.53	1.48	1.45	2.98
New	19.74	1.40	1.45	2.71

5. Discussion and Final Remarks

In this paper, an overview of methods to handle missing covariate data is presented. Also, the use of a nonparametric kernel based mean score estimator is proposed. In line with current common wisdom, naive methods like the complete case analysis should be avoided in practice, since it is known that they lead in general to inefficient and biased estimates. Multiple imputation and inverse probability weighting lead to improved estimates, but only when the imputation and weighting models are correctly specified. The doubly

robust estimators do not suffer from this property, but generalization to several missing covariates is not straightforward in this case. The method of Wang & Wang is, similar to our method, kernel-based, but less efficient in the sense that not all available information is used. The new method we propose is using all available information directly in the estimating equations.
In the simulation studies, two methods could be indicated as behaving poor: the complete case analysis and the inverse probability weighting with incorrectly specified weights. The multiple imputation behaves very well in the simulation studies, since it was based on the correct imputation model. Our method also behaves very well in the missing-one-covariate, missing-two-covariates and missing-three-covariates case. In the new proposal, we used three nearest neighbors with a uniform kernel, and we showed that the method is not too sensitive to the choice of the number of nearest neighbors. There are options to find the best bandwidth and/or number of nearest neighbors (for example leave-one-out-cross-validation), but these are computational intensive procedures. We therefore suggest to use three nearest neighbors with a uniform kernel.

Acknowledgements

This work has been funded by the IAP research network nr P6/03 of the Belgian Government (Belgian Science Policy).
For the simulation studies we used the infrastructure of the VSC - Flemish Supercomputer Center, funded by the Hercules foundation and the Flemish Government - department EWI.
We thank both referees for their constructive remarks that led to an improved version of this manuscript.

References

[1] Aerts, M., Janssen, P., Veraverbeke, N. (1994) Bootstrapping regression quantiles. *Journal of Nonparametric Statistics* **4**, 1–20.

[2] Carpenter, J.R., Kenward, M.G. and Vansteelandt, S. (2006) A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *J. R. Statist. Soc. A* **169**, 571–584.

[3] Blatchford, P., Goldstein, H., Martin, C., and Browne, W. (2002) A study of class size effect in English school reception year classes. *Br. Educ. Res. J.* **28**, 169–185.

[4] Efron, B. and Tibshirani, R.J. (1993). *An introduction to the bootstrap.* Chapman & Hall.

[5] Flanders, W. and Greenland, S. (1991) Analytic methods for two-stage case-control studies and other stratified designs. *Statistics in Medicine* **10**, 739–747.

[6] Hardle, W. (1989). *Applied nonparametric regression.* Cambridge University Press.

[7] Horvitz, D.G. and Thompson, D.J.(1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**, 663-685.

[8] Kutner, M.H., Nachtsheim, C.J., Neter, J. and Li, W. (1996). *Applied Linear Statistical Models.* Boston: McGraw-Hill.

[9] Little, R. and Rubin, D. (1987). *Statistical Analysis with Missing Data.* New York: Wiley.

[10] Loader, C. (1999). *Local regression and Likelihood. Statistics and Computing.* Springer-Verlag: New York.

[11] Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data.* New York: Springer.

[12] Reilly, M., Pepe, M.S. (1995) A mean-score method for missing and auxiliary covariate data in regression models. *Biometrika*, **82**, 299–314.

[13] Robins, J., Rotnitzky, A. and Zhao, L. (1994) Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, **89**, 846–866.

[14] Robins, J. and Rotnitzky, A. (1995) Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, **90**, 122–129.

[15] Robins, J., Rotnitzky, A. and Zhao, L. (1995) Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, **90**, 106–121.

[16] Rubin, D. (1976). Inference and Missing Data. *Biometrika*, **63** 581–592.

[17] Rubin, D. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley

[18] Stone, C. (1977). Consistent nonparametric regression. *Annals of Statistics*, **5** 595–645.

[19] Van der Laan, M.J. and Robins, J. (2003). *Unified Methods for Censored Longitudinal Data and Causality.* New York: Springer

[20] Wang, C.Y., Wang, S., Zhao, L.P., Ou, S.T. (1997) Weighted semiparametric estimation in regression analysis with missing covariate data. *J. Amer. Statist. Assoc.*, **92**, 512–525.

[21] Wang, C., Wang, S., Gutierrez, R. and Carroll, R. (1998) Local linear regression for generalized linear models with missing data. *Annals of Statistics*, **26**, 1028–1050.

[22] Wang, S., Wang, C.Y. (2001) A note on kernel assisted estimators in missing covariate regression. *Statistics and Probability letters*, **55**, 439–449.

[23] Wang, C.Y., Lee, S., Chao, E.C.(2007) Numerical equivalence of imputing scores and weighted estimators in regression analysis with missing covariates. *Biostatistics*, **8(2)**, 468–473.

[24] Zhao, L.P. and Lipsitz, S. (1992) Design and analysis of two-stage studies. *Statistics in Medicine*, **11**, 769–782.