Epidemiology Biostatistics and Public Health - 2013, Volume 10, Number 1

Multiple imputation based on conditional quantile estimation

Matteo Bottai[1], Huiling Zhen[2]

Abstract

Multiple imputation is a simulation-based approach for the analysis of data with missing observations. It is widely utilized in many settings and preeminent among general approaches when the analytical method does not involve a likelihood function or this is too complex. We consider a multiple imputation method based on the estimation of conditional quantiles of missing observations given observed data. The method does not require the specification of a likelihood and has desirable features that may be useful in some practical settings. It can also be applied to impute dependent, bounded, censored and count data. In a simulation study it shows some advantage over the alternative methods considered in terms of mean squared error across all scenarios except when the data arise from a normal distribution where all methods considered perform equally well. We present an application to the estimation of percentiles of body mass index conditional on physical activity assessed by accelerometers.

Keywords: Conditional quantiles; Missing data; Multiple imputation; Quantile regression; Smoothing splines

DOI: 10.2427/8758

1 Introduction

Missing data are frequent in social and health sciences and often pose issues in data analysis. Analyses of the subset of the data with complete observations may lead to biased and inefficient inference. Numerous approaches have been suggested, and those based on multiple imputation or the likelihood function are foremost among them. Both are generally applicable, though the likelihood-based approaches are clearly not viable when the analytic method does not involve a likelihood function or the likelihood is intractable.

Multiple imputation has been applied to a wide variety of missing data problems described in several books [1, 2, 3, 4, 5] and papers [6, 7, 8, 9, 10, 11, 12, among others]. Methods for imputing multivariate data are often based on distributional assumptions about either the joint multivariate distribution of the data or a fully conditional specification of it.

Methods that relax these assumptions have also been proposed. [13] described a Bayesian approach based on a Pólya tree prior, a generalization of the Dirichlet process, that allows for imputation of continuous, discrete, and ordinal data with ignorable non-response. [14] used non-parametric Markov chain bootstrap to impute scalar and multivariate outcomes when the data are

[1] ***Correspondig Author***, Unit of Biostatistics, Institute of Environmental Medicine, Karolinska Institutet, Nobels väg 13, Karolinska Institutet, 17177 Stockholm, Sweden. *e-mail*: matteo.bottai@ki.se

[2] Division of Biostatistics, University of South Carolina, Columbia, SC. *e-mail*: hlzhen2000@yahoo.com

missing completely at random. In non-Bayesian settings [15] proposed a non-parametric method for the imputation of conditional mean, whose resulting estimator is shown to be consistent and asymptotically normal. [16] presented a non-parametric and a semi-parametric smoothing method to obtain multiple imputation estimators based on local resampling techniques. Their methods require setting smoothing parameters, and the authors give guidelines as to how to select them. [17] described a semi-parametric imputation approach based on mixtures of normal distributions, which utilizes the algorithm proposed by [18]. Recently, [19] suggested a non-parametric kernel estimator of the conditional quantiles within the context of empirical likelihood inference with missing observations for parameters defined by general estimating equations.

In this paper we present an imputation method that is based on the estimation of conditional quantiles of the missing observations given the observed data and does not require the full specification of a probability model. As [20] noted, "[t]he process of specifying the imputation model is a scientific modeling activity on its own, that comes with its own model building principles." The methods proposed may facilitate this modeling activity in some real-life settings.

The following section introduces the method for the simplest case of one only incomplete variable. Some practical suggestions for the estimation of the conditional quantiles are given in section 3. Section 4 extends the method to the general case of multiple incomplete variables. Section 5 describes the setup and the results of a simulation study. Section 6 illustrates an application to the study of the association between obesity and physical activity with epidemiological data. In section 7 we summarize the main features of the proposed method and offer few final remarks.

2 A Single Incomplete Variable

We first consider the simple, though often unrealistic, case in which only one variable has missing observations while all the others are completely observed. We present the general case of multiple incomplete variables in section 4.

Let y_i, $i = 1, \ldots, n$, be observations from n independent random variables Y_i, and $x_i = (x_{i,1}, \ldots, x_{i,p})'$ be a p-dimensional vector of covariates. We assume that the conditional cumulative distribution function of Y_i given x_i, $F(y_i|x_i)$, is continuous.

Suppose the vector x_i is completely observed for all $i \in \{1, \ldots, n\}$, while y_i has missing observations. Let $I \subset \{1, \ldots, n\}$ be the non-empty set of indexes corresponding to the units with missing values for y_i and $C \subset \{1, \ldots, n\}$ the complement of the set I, i.e. the non-empty set of indexes of the observed values for y_i. Here and throughout we further assume that the missing-data mechanism is ignorable, i.e., the probability that Y_i is observed given x_i is conditionally independent of Y_i [21].

In multiple imputation each missing value y_i, with $i \in I$, is replaced by M independent imputed values to generate M completed data sets [1, 2]. The constant integer M is usually set between 5 and 10. Each imputed value, $\hat{y}_i^{(m)}$, $m \in \{1, \ldots, M\}$, is generated by drawing one replicate from $F(y_i|x_i)$, which is generally assumed to have a known form (e.g. normal, lognormal). Instead, in the proposed method the form of $F(y_i|x_i)$ is left unspecified. Let

$$Q_{Y_i|x_i}(u) = F^{-1}(u) = \inf\{y : F(y_i|x_i) \geq u\} \tag{1}$$

be the u-th quantile of the conditional distribution of Y_i given x_i. The proposed multiple imputation method can be summarized in the following steps:

1. For each missing value y_i, with $i \in I$, generate u from a uniform distribution over $(0, 1)$.

2. Impute $\hat{y}_i^{(m)} = \hat{Q}_{Y_i|x_i}(u)$, where $\hat{Q}_{Y_i|x_i}(u)$ is a consistent estimate of $Q_{Y_i|x_i}(u)$ obtained by using all units with $i \in C$. (Details are given in section 3.)

3. Repeat steps 1 and 2 for $m = 1, \ldots, M$, to generate M completed datasets.

Step 2 above requires a point estimate of $Q_{Y|x}(u)$. Any other inference about $Q_{Y|x}(u)$ (e.g. standard errors, confidence intervals) is unnecessary.

Variability due to imputation is introduced by generating an independent uniform random draw for imputing each missing value in each dataset in step 1 and by then creating M completed datasets in step 3.

After multiple completed datasets have been generated, estimation of any quantity of interest, θ, can be carried out as usual [1]. The M completed datasets can be analyzed separately with any statistical method of interest, and the resulting M sets of point estimates $\hat{\theta}^{(m)}$ and estimates for the sampling variance $\hat{\text{Var}}(\hat{\theta}^{(m)})$, $m = 1, \ldots, M$, can then be combined according to Rubin's rules. The combined point estimate is

$$\hat{\theta} = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}^{(m)} \tag{2}$$

with estimated sampling variance

$$\hat{\text{Var}}(\hat{\theta}) = \frac{1}{M} \sum_{m=1}^{M} \hat{\text{Var}}(\hat{\theta}^{(m)}) + \frac{M+1}{M(M-1)} \sum_{m=1}^{M} (\hat{\theta}^{(m)} - \hat{\theta})^2 \tag{3}$$

The intuition behind the proposed imputation method is based on a well known result, stated here as Lemma 1 [22, p. 52–54].

Lemma 1 (Probability integral transformation) *Let W be a variable with continuous cumulative distribution function $F_W(w)$. The probability integral transformation $U = F_W(W)$ is uniformly distributed over $(0, 1)$.*

Note that Lemma 1 does not assume that $F_W(w)$ is absolutely continuous nor strictly increasing. Step 2 of the imputation method is a direct application of the result of Lemma 1. Proposition 1 gives a formal justification for this step, and its proof is given in the Appendix.

Proposition 1 (Consistency) *If $\hat{Q}_{Y_i|x_i}(u) \xrightarrow{P} Q_{Y_i|x_i}(u)$ for every $u \in (0, 1)$, then $\hat{Y}_i^{(m)} \xrightarrow{D} Y_i$ for every $i \in \{1, \ldots, n\}$ and every $m \in \{1, \ldots, M\}$.*

Proposition 1 states that if $\hat{Q}_{Y_i|x_i}(u)$ in step 2 of the imputation method is a consistent estimator of $Q_{Y_i|x_i}(u)$, then the asymptotic conditional distribution of each imputed value $\hat{Y}_i^{(m)}$ given the observed values x_i is equal to the conditional distribution of the unobserved values Y_i given x_i.

3 Conditional Quantile Estimation

In this section we give some practical suggestions about the estimation of the conditional quantile $Q_{Y_i|x_i}(u)$ in step 2 in section 2. These may facilitate the use of the proposed imputation method in various real-life settings.

If there exist a parameter vector β_u, dependent on u, such that

$$Q_{Y|x}(u) = x'\beta_u \tag{4}$$

then the conditional quantile, $Q_{Y|x}(u)$, can be easily estimated by optimizing a mathematical linear programming problem [23]. The estimation requires no assumptions on the regression residuals and can be carried out with widely available statistical software (e.g. R, SAS, Stata).

3.1 Non-parametric models

Though the linear quantile regression model (4) makes no assumption about the distribution of the regression residual, it assumes a linear parametric form for the relationship between the quantile of the incomplete variable being imputed, $Q_{Y|x}(u)$, and the covariates, x. When this assumption is deemed untenable, non-parametric methods could be used instead. Let us consider the non-parametric model

$$Q_{Y|x}(u) = h_u(x)$$

where h_u is an unknown and unspecified non-random function dependent on u. The conditional quantile $Q_{Y|x}(u)$ can be estimated by using a consistent estimator of h_u. Among numerous non-parametric methods, one may consider quantile smoothing splines [24], which are defined as the solution to

$$\min_{h\in\mathcal{H}} \sum_{i\in B} \rho_u\{y_i - h(x_i)\} + \lambda \left\{ \int_0^1 \left| \frac{d^2}{dx^2} h(x) \right| \, dx \right\}$$

where $\rho_u(t) = |t| - (2u-1)t$, the class $\mathcal{H}$ is appropriately chosen, and the constant λ controls the degree of smoothing. When λ is sufficiently large, the solution $\hat{h}_u$ is the linear u-quantile fit to the data. The set of solutions for any given u and λ can be efficiently computed via parametric mathematical linear programming.

3.2 Variable Transformation

A possibly simpler alternative to quantile smoothing splines is to transform the outcome variable y through some convenient function, $g(y)$. In many problems it may be simpler to model the relationship between outcome and covariates after the outcome has been transformed. The logarithm and square root, for instance, are popular transforms of right-skewed, non-negative outcomes. Suppose there exists a known non-decreasing function g and a vector β_u, dependent on u, such that

$$Q_{g(Y)|x}(u) = x'\beta_u \tag{5}$$

where $Q_{g(Y)|x}(u)$ indicates the conditional quantile of the transformed outcome. This can be easily estimated by regressing $g(y)$ on x with linear quantile regression. Estimates for the conditional quantile of the untransformed outcome, $Q_{Y|x}(u)$, can then be obtained by applying the equivariance property of quantiles [23]

$$Q_{Y|x}(u) = g^{-1}\{Q_{g(Y)|x}(u)\} \tag{6}$$

which holds for any non-decreasing function g and constant u simply because $P(Y \leq y|X = x) = P\{g(Y) \leq g(y)|X = x\}$ for every random vector (Y, X).

The selection of the function g should aim to linearize the relationship between the conditional quantile the covariates and constrain its estimates within the support of the variable to be imputed. For example, [25] propose applying a logistic transform to model outcome variables which take on values that are bounded within a known interval, such as percentages bounded between 0 and 100, math achievement scores, and depression scales. [26] propose a power transformation, $g(y) = (y^{\lambda} - 1)/\lambda$ if $\lambda \neq 0$ and $g(y) = \log(y)$ if $\lambda = 0$, which is easy to implement and requires no distributional assumptions. Other ad hoc transformations can also be applied.

3.3 Dependent Data

Sampling designs may sometimes induce dependence among the data. For example, cluster, multilevel, and repeated measures (or longitudinal or panel) designs are frequently adopted. In these, observations within each cluster, level or unit repeatedly measured may be dependent on one another.

As shown by [27], the conditional quantile estimator is consistent when data are dependent. Proposition 1 therefore holds with dependent data and the proposed imputation method can be applied unchanged. In any given real application other related methods could also be considered [28, 29, 30].

3.4 Censored Data

The quantile regression approach can also be utilized when missing values occur because of fixed [31] or random censoring [32, 33, 34].

3.5 Discrete and Categorical Data

Predictions from quantile regression are generally appropriate for continuous, not discrete, outcome variables. When the incomplete variable to be imputed is discrete and takes on a finite number of unique values with positive probability, its quantiles are themselves discrete and cannot be modeled directly as a continuous function of a set of covariates. When the outcome is a count, the quantile regression method proposed by [35] could prove useful. In other settings binomial, ordinal, or multinomial regression may be appropriate.

4 Multiple Incomplete Variables

In this section we extend the proposed method to the more realistic scenario where multiple variables have missing observations.

We first extend the notation. Let $y_i = (y_{i,1}, \ldots, y_{i,k})'$, with $i = 1, \ldots, n$, denote replicates of a k-dimensional random vector $Y_i = (Y_{i,1}, \ldots, Y_{i,k})'$. Let $Y_{i,(-j)}$ indicate the $(k-1)$-dimensional vector defined as Y_i without its j-th element. Suppose the vector y_i has missing observations. For each element j, let $I_j \subset \{1, \ldots, n\}$ be the non-empty set of indexes corresponding to the units with missing values for $y_{i,j}$. Let $C_j = \{1, \ldots, n\} \backslash I_j$ be the set of indexes of the observed values. As in section 2, suppose the vector $x_i = (x_{i,1}, \ldots, x_{i,p})'$ is completely observed.

The proposed method for multiple incomplete variables follows the one proposed by [36] and is summarized in the following steps:

1. For each missing value $y_{i,1}$, with $i \in I_1$, generate u from a uniform distribution over $(0, 1)$ and impute $\hat{y}_{i,1}^{(m)} = \hat{Q}_{Y_{i,1}|x_i}(u)$ obtained by using all units with $i \in C_1$.

2. For each $j = 2, \ldots, k$, and for each missing value $y_{i,j}$, with $i \in I_j$, generate u from a uniform distribution over $(0, 1)$ and impute $\hat{y}_{i,j}^{(m)} = \hat{Q}_{Y_{i,j}|y_{i,1},\ldots,y_{i,j-1},x_i}(u)$ obtained by using all units with $i \in C_j$.
3. For each missing value $y_{i,j}$, with $i \in I_j$, generate u from a uniform distribution over $(0, 1)$ and impute $\hat{y}_{i,j}^{(m)} = \hat{Q}_{Y_{i,j}|y_{i,(-j)},x_i}(u)$ obtained by using all units with $i \in C_j$.
4. Repeat step 3 for R complete cycles of $j = 1, \ldots, k$. At each cycle, replace previous imputations with updated ones. This creates a single imputation sample.
5. Repeat steps 1 to 4 for $m = 1, \ldots, M$, to generate M completed datasets.

Steps 1 and 2 initialize the algorithm by imputing all missing values for each of the incomplete variables. Then steps 3 and 4 impute updated values for each variable in turn conditional on all other variables by using both observed and latest-imputed values. The cycle over all incomplete variables is repeated R times to create one completed dataset. Step 5 generates the M completed datasets.

As described by [36], the proposed approach is similar to HOMALS-like algorithms, which usually convergence fast during the first few cycles [37]. Based on our simulation study, we expect $R = 10$ cycles be generally sufficient for the proposed method.

As for the single incomplete variable case described in section 2, variability is introduced by creating M completed datasets and by generating an independent uniform random draw for imputing each missing value in each dataset. Additional variability enters at step 4, when missing values for each of incomplete variable are imputed repeatedly over R cycles within each of the M datasets. This reflects the fact that information is missing from the covariates [36].

As noted by [38, pp. 59–60], the complete stochastic mechanism for generating the random response given a set of covariates defined by $Y = x'\beta_u$ suggests that the elements of the vector β_u are dependent, for they are all generated by one replicate u of a random variable uniformly distributed over $(0, 1)$. Yet, unlike other imputation methods, which typically assume multivariate normality of the regression coefficients, their marginal distributions can take arbitrary forms.

Since the imputation method proposed is not based on modeling a fully conditional likelihood, it does not incur the risk of not converging to a sensible stationary distribution, which on the contrary may occur, if rarely, when the separate conditional-likelihood models are not compatible with any joint distribution [39].

5 Simulation Study

We examined the performance of the proposed and other imputation methods in a simulation study where we pseudo-randomly generated a large number of datasets under different, known scenarios. In the following subsections we describe the scenarios, the mechanisms to assign missing data, the methods to analyze the data, and the criteria to evaluate the adequacy and effectiveness of the alternative imputation methods. The study was performed with the statistical computer software R [40].

5.1 Generating Complete Datasets

First we generated complete datasets. Each dataset was comprised of n independent observations, with $n \in \{100, 500\}$, on an outcome variable Y and four covariates $X = (X_0, X_1, X_2, X_3)'$, with

$$Y = X'\beta + \varepsilon \,, \tag{7}$$

where $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)' = (0, 1, 2, 3)'$, X_0 was the constant intercept, X_1 and X_2 were generated from uniform distribution as $X_1 \sim U(0, 1)$ and $X_2 \sim U(-1, 2)$, and X_3 and the error term, ε, were generated from each of three different distributions: a standard normal, $N(0, 1)$, a t-Student with 3 degrees of freedom, t_3, and the chi-square with 1 degree of freedom, χ^2_1.

We focused on the performance of the various imputation methods in the simplest case in which all quantiles of the variable Y, or any non-decreasing transformation of it, were linear combinations of the covariates. A discussion on non-linear relationships is beyond the scope of the present paper.

5.2 Generating Missing Observations

In each complete dataset we assigned missing values to the variable y and the variable x_3 under each of three non-response generating mechanisms. Each observation for y and each observation for x_3 were replaced with missing data with probability p, where $p = 0.3$, $p = 0.5$, and $\text{logit}(p) = -1 + x_1$. In the first two cases the missing indicator was independent of x and y, missing completely at random [21], while in the third it depended on x_1 but not on y or x_3, missing at random. The non-ignorable missing mechanism was not considered in the present study, since all the imputation methods compared herein assumed that the mechanism was ignorable.

5.3 The Simulated Scenarios

Overall, we considered 18 different scenarios, which arise from combining two sample sizes, three distributions of error terms, and three non-response generating mechanisms. For each scenario we generated 1000 datasets.

5.4 Imputing the Data

We imputed each incomplete dataset by applying the proposed imputation method and three methods implemented in MICE library for the statistical package R [41]: Bayesian linear regression, predictive mean matching, and unconditional mean imputation. We set the number of imputations $M = 5$.

5.5 Analyzing the Data

In each of the five completed datasets we estimated the regression model (7). We applied four regression methods: 0.25-quantile regression, 0.50-quantile regression, 0.75-quantile regression, and least-squares linear regression. Then we combined the estimates for the regression parameter, β, and those for the marginal mean and variance of Y, μ and σ^2.

5.6 Evaluation Criteria

In each of the 18 scenarios, each comprised of 1000 simulated datasets, we evaluated the adequacy and effectiveness of the alternative multiple imputation methods. Consider the parameter vector $\theta = (\beta_0, \beta_1, \beta_2, \beta_3, \mu, \sigma^2)'$. Let θ_i, $i = 1, \ldots, 6$, denote the ith element of θ. For each parameter θ_i, let $\hat{\theta}_{i,j}$ be the combined estimate from the jth completed datasets and $\tilde{\theta}_{i,j}$ the estimate from the jth complete dataset before generating the missing observations. We evaluated the performance of the imputation methods with respect to bias, variance, and mean squared error (MSE) defined

as follows:

$$\begin{aligned}
\text{Bias}(\hat{\theta}_i) &= \frac{1}{1000}\sum_{j=1}^{1000}\left(\hat{\theta}_{i,j} - \tilde{\theta}_{i,j}\right) \\
\text{Var}(\hat{\theta}_i) &= \frac{1}{999}\sum_{j=1}^{1000}\left(\hat{\theta}_{i,j} - \frac{1}{1000}\sum_{k=1}^{1000}\hat{\theta}_{i,k}\right)^2 \\
\text{MSE}(\hat{\theta}_i) &= \frac{1}{1000}\sum_{j=1}^{1000}\left(\hat{\theta}_{i,j} - \tilde{\theta}_{i,j}\right)^2
\end{aligned}$$

These are usually defined by replacing $\tilde{\theta}_{i,j}$ with θ_i in the expressions above. With some abuse of terminology, we preferred to adopt the evaluation criteria above, which we believed were more relevant measures of the performance of the methods considered.

5.7 Simulation Results

For brevity, only a subset of all the simulation results is summarized in tabular form. The tables show the comparison of the proposed method with the best competing alternative only, which in our simulation was predictive mean matching. The tables report the observed bias, variance, and mean squared error of the estimators of the regression coefficients and of the marginal mean and variance when the sample size was 500. In Tables 1 to 3 the data were generated under missing completely at random mechanism with probability of non-response $p = 0.3$ from the χ^2_1, the t_3, and the $N(0, 1)$ distribution, respectively. In Table 4 the missing data were missing at random from the χ^2_1 distribution.

In this simulation the proposed method consistently showed slight bias and small variance. When data were generated from a $N(0, 1)$ (Table 3), all methods performed equally well. In all other scenarios (Tables 1, 2, 4) the mean squared error of the best competing alternative was greater than that of the proposed method.

As stated at the beginning of the present section, the results about the other two imputation methods, namely Bayesian linear regression and unconditional mean, are not reported in the tables. The former was comparable to predictive mean matching, though it showed erratic behavior in some of the simulation scenarios. The latter showed dramatically larger bias and variance than any other method considered. The results for sample size 100 and for a probability of non-response $p = 0.5$ were analogous and not shown.

6 An Example: Obesity and Physical Activity

The National Health and Nutrition Examination Survey (NHANES) is conducted by the Centers for Disease Control and Prevention since the early 1960's and uses a stratified, multistage probability design. During the 2003-2004 study cycle, it included an interview, physical examination, and laboratory tests.

Our research interest lay in the association between obesity and physical activity. We consider body mass index (BMI, kg/m^2) as a measure of obesity. Physical activity was assessed through accelerometers, an increasingly popular instrument [42]. Data from the accelerometers were used to calculate time spent in moderate-to-vigorous physical activity (MVPA, metabolic equivalent of task, or MET, in minutes per day).

We excluded males, participants less than 18 or greater than 49 years old, participants currently taking anti-hypertension medications, pregnant women, participants who are prevented

by impairment from walking, and participants with a history of stroke, congenital heart failure, angina, emphysema, or chronic bronchitis because these factors will affect the physical activity assessments. The final sample consisted of 1,227 individuals. Age was completely observed, BMI had 1,146 valid observations, and MVPA 362.

The sample distributions of BMI and MVPA were markedly right-skewed (third standardized moment respectively equal to 1.12 and 1.66) and leptokurtic (fourth standardized moment equal to 4.47 and 5.83). Given the sizable skewness and leptokurtosis of the sample marginal distributions, the use of imputation methods based on the multivariate normality assumption seemed inappropriate. Moreover, any potential issues resulting from the likely violation of this assumption could be exacerbated at the extreme percentiles where our inferential interest lay.

We utilized the proposed imputation method. The pair-wise relationships between age, BMI and MVPA were non-linear. After taking the logarithm of BMI and MVPA, however, they seemed approximately linear at all quantiles. We therefore utilized log(BMI) and log(MVPA) and applied the imputation model (5).

As described in section 3.2, we selected the logarithmic transform because it linearized the relationships and ensured that the imputed values for BMI and MVPA were all plausible (i.e. positive). The transform did not aim at normalizing the shape of the conditional distributions of the variables to be imputed. This would have been unnecessary, for the proposed imputation method is valid under any distribution. Indeed, the sample distributions of log(BMI) and log(MVPA) were still far from normal.

Even if the missing data pattern was not monotone, BMI had substantially fewer missing values than MVPA. Therefore, in steps 1 and 2 in section 4 we first imputed log(BMI) from age and then log(MVPA) from age and the imputed log(BMI). In steps 3 and 4 we performed $R = 10$ complete cycles, and in step 5 obtained $M = 5$ final completed datasets.

In each of the five completed dataset, we estimated the quantile regression model

$$Q_{\log(\text{BMI})}(p) = \beta_{p,0} + \beta_{p,1}\ \text{age} + \beta_{p,2}\ \ \log(\text{MVPA})$$

where $Q_{\log(\text{BMI})}(p)$ denotes the p-quantile of the conditional distribution of log(BMI) given age and log(MVPA), and $\beta_p = (\beta_{p,0}, \beta_{p,1}, \beta_{p,2})'$ is the regression coefficient vector to be estimated for the p-quantile. We considered five quantiles, $p = 0.10$, 0.25, 0.50, 0.75, and 0.90. The stratified, multistage probability design, was taken into account in the estimation. Estimation of the regression coefficients included the sampling probability weights available in the dataset. Estimation of the standard errors was performed by generating 100 stratified, cluster bootstrap samples.

Figure 1 shows the five estimated percentiles of BMI at 29 years, sample median age, on the double-log and natural scale. The latter were obtained by simply applying the inverse transform (i.e. the exponential function) to the estimates for $Q_{\log(\text{BMI})}(p)$, thus exploiting the equivariance property of quantile regression shown in equation (6).

We were particularly interested in the higher quantiles, which corresponded to the obese portion of the population whose health could be compromised. For the 90th percentile, the estimates for the coefficients from the complete-case data (356 valid observations) were $\hat{\beta}_{0.9} = (3.53, 0.00147, -0.0429)'$ with corresponding standard errors $\{\hat{\text{Var}}(\hat{\beta}_{0.9})\}^{1/2} = (0.0559, 0.00133, 0.02$ The estimates from the five completed datasets combined according to the expressions (2) and (3) were $\hat{\beta}_{0.9} = (3.67, 0.00157, -0.0665)'$ and $\{\hat{\text{Var}}(\hat{\beta}_{0.9})\}^{1/2} = (0.0921, 0.00233, 0.0338)'$. The direction of the effect of age and MVPA on BMI with the completed data was the same as those with complete-case data. The magnitude of the effects was larger with completed data by about 7% for age and 55% for log(MVPA). The standard errors were larger from completed data at the 90th percentile but smaller at the lower percentiles.

The results indicated that the 90th percentile of BMI increased with age and decreased with MVPA (Figure 1). The decrease in BMI with increasing values of MVPA was linear on the double-log scale. On the natural scale, however, the decrease was rapid for values of MVPA near zero but less and less pronounced as MVPA increased. The 90th percentile of BMI was still above 30 kg/m^2, the well accepted cut-off value for obesity in adults, even in highly active women. The analysis was extended to the other quantiles and larger sets of covariates for the imputation models. The results were congruent with those reported and not shown for brevity.

7 Final Remarks

The imputation method proposed may prove useful in some missing data problems and may be particularly appropriate when the research interest lies in the shape of the entire conditional distribution of some incomplete response, not just its mean. Under the scenarios of our simulation study, the method showed a mean squared error that was at least as small as, and sometimes considerably smaller than, that of the other methods considered.

In addition, the proposed method enjoys all the desirable features characteristic of inference about quantiles, which include that it (1) is robust to outliers, (2) makes no distributional assumption about the regression coefficients or residuals, (3) is invariant to transformations, (4) can be applied to dependent, bounded, censored and count data, and (5) its algorithm is computationally simple.

These features may prove useful when handling some known issues [10, p. 214]. For example, feature (1) may improve efficiency with respect to methods that are based on the estimation of the mean (e.g. least-squares regression), feature (2) may circumvent the need for diagnostics of model fitting (e.g., detection of influential points) and avoid convergence issues resulting from specifying conditional densities that may be incompatible with any multivariate distribution, feature (3) may alleviate modeling problems (e.g. imputations outside the support of the imputed variable, non-linear relationships), and feature (4) may extend applicability to a range of different types of data.

Appendix

In this section we present a proof of Proposition 1 that uses the result in Lemma 1 along with the fact that convergence in probability implies convergence in distribution.

Proof of Proposition 1 We use the following inequality

$$P(W_1 \leq c) \leq P(W_2 \leq c+\epsilon) + P(\ |W_1 - W_2| > \epsilon) \tag{8}$$

which holds for every random variable W_1 and W_2, every constant c, and every $\epsilon > 0$. Then

$$P(\hat{Y}_i^{(m)} \leq c) \leq P(\tilde{Y}_i \leq c+\epsilon) + P(\ |\hat{Y}_i^{(m)} - \tilde{Y}_i| > \epsilon) \tag{9}$$

$$P(\hat{Y}_i^{(m)} \leq c) \geq P(\tilde{Y}_i \leq c-\epsilon) - P(\ |\hat{Y}_i^{(m)} - \tilde{Y}_i| > \epsilon) \tag{10}$$

where the probabilities here and throughout are intended to be conditional on $X_i = x_i$, $\hat{Y}_i^{(m)} = \hat{Q}_{Y_i|x_i}(u)$ and $\tilde{Y}_i = Q_{Y_i|x_i}(u)$ with $u \sim U(0,1)$. By the consistency of $\hat{Q}_{Y_i|x_i}(u)$ for every $u \in (0,1)$,

$$\lim_{n_c \to \infty} P\{|\hat{Q}_{Y_i|x_i}(u) - Q_{Y_i|x_i}(u)| > \epsilon\} = P(\ |\hat{Y}_i^{(m)} - \tilde{Y}_i| > \epsilon) = 0 \tag{11}$$

where $n_c = \#C$, the size of the subsample with complete observations. By concatenating the inequalities (9) and (10) and using the limit (11),

$$P(\tilde{Y}_i \leq c - \epsilon) \leq \lim_{n_c \to \infty} P(\hat{Y}_i^{(m)} \leq c) \leq P(\tilde{Y}_i \leq c + \epsilon) \ . \tag{12}$$

Inverting the result in Lemma 1 gives that the random variable $\tilde{Y}_i$ is equal in distribution to Y_i, which implies that $P(\tilde{Y}_i \leq c) = P(Y_i \leq c)$. Since (12) holds for every ϵ, taking the limit $\epsilon \downarrow 0$ yields

$$\lim_{n_c \to \infty} P(\hat{Y}_i^{(m)} \leq c) = P(\tilde{Y}_i \leq c) = P(Y_i \leq c) \ .$$

◇

References

[1] Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: J. Wiley & Sons.

[2] Schafer, J. L. (1997). Analysis of Incomplete Multivariate Data. New York: Chapman & Hall.

[3] Little, R. J. and Rubin, D. B. (2002). Statistical Analysis With Missing Data. New York: J. Wiley & Sons.

[4] Allison, P. D. (2001). Missing Data. Sage University Papers Series on Quantitative Applications in the Social Sciences,07-136.Thousand Oaks, CA:Sage.

[5] McKnight, P. E., McKnight, K. M., Sidani, S., and Figueredo, A. J. (2007). Missing Data: A Gentle Introduction. New York, NY: Guilford.

[6] Shafer, J. L. (1999). Multiple imputation: a primer. Statistical Methods in Medical Research 8, 3-15.

[7] Schafer, J. L. and Graham, J. W. (2002). Multiple imputation: our view of the state of the art. Psychological Methods 7, 147-177.

[8] Horton, N. J. and Kleinman, K. P. (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. The American Statistician 61, 79-90.

[9] Harel, O. and Zhao, X.-H. (2007). Multiple imputation: Review of theory, implementation and software. Statistics in Medicine 26, 3057-3077.

[10] Kenward, M. G. and Carpenter, J. (2007). Multiple imputation: current perspectives. Statistical Methods in Medical Research 16, 199-218.

[11] Ruan, P. K. and Gray, R. J. (2008). Analyses of cumulative incidence functions via non-parametric multiple imputation. Statistics in Medicine 27, 5709-5724.

[12] Moodie, E. E. M., Delaney, J. A. C., Lefebvre, G., and Platt, R. W. (2009). Missing confounding data in marginal structural models: A comparison of inverse probability weighting and multiple imputation. International Journal of Biostatistics 4.1.

[13] Paddock, S. M. (2002). Bayesian nonparametric multiple imputation of partially observed data with ignorable nonresponse. Biometrika 89, 529-38.

[14] Zhang, L.-C. (2004). Nonparametric markov chain bootstrap for multiple imputation. Computational Statistics & Data Analysis 45, 343-353.

[15] Nielsen, S. F. (2001). Nonparametric conditional mean imputation. Journal of Statistical Planning and Inference 99, 129-150.

[16] Aerts, M., Claeskens, G., Hens, N., and Molenberghs, G. (2002). Local multiple imputation. Biometrika 89, 375-388.

[17] Di Zio, M., Guarnera, U., and Luzi, O. (2007). Imputation through finite gaussian mixture models. Computational Statistics & Data Analysis 51, 5305-5316.

[18] Hunt, L. and Jorgensen (2003). Mixture model clustering for mixed data with missing information. Computational Statistics & Data Analysis 41, 429-440.

[19] Wang, D. and Chen, S. X. (2009). Empirical likelihood for estimating equations with missing values. Annals of Statistics 37, 490-517.

[20] van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. Statistical Methods in Medical Research 16, 219-242.

[21] Rubin, D. B. (1976). Inference and missing data. Biometrika 63, 581-592.

[22] Casella, G. and Berger, R. L. (2002). Statistical Inference. Duxbury Press.

[23] Koenker, R. and Bassett, G. (1978). Regression quantiles. Econometrica 46, 33-50.

[24] Koenker, R., Ng, P., and Portnoy, S. (1994). Quantile smoothing splines.Biometrika 81, 673-680.

[25] Bottai, M., Cai, B., and McKeown, R. (2010). Logistic quantile regression. Statistics in Medicine 29, 309-317.

[26] Mu, Y. and He, X. (2007). Power transformation toward a linear regression quantile. JASA 102, 269-279.

[27] Jung, S. (1996). Quasi-likelihood for median regression models. Journal of the American Statistical Association 91, 251-257.

[28] Lipsitz, S. R., Fitzmaurice, G. M., Molenberghs, G., and Zhao, L. P. (1997). Quantile regression methods for longitudinal data with drop-outs: application to cd4 cell counts of patients infected with the human immunodeficiency virus. Journal of the Royal Statistical Society, Series C: Applied Statistics 46, 463-476.

[29] Geraci, M. and Bottai, M. (2007). Quantile regression for longitudinal data using the asymmetric laplace distribution. Biostatistics 8, 140-154.

[30] Reich, B., Bondell, H., and Wang, H. (2010). Flexible bayesian quantile regression for independent and clustered data. Biostatistics page 11 (2): 337-352.

[31] Powell, J. (1986). Censored quantile regression. Journal of Econometrics 32, 143-155.

[32] Portnoy, S. (2003). Censored regression quantiles. Journal of the American Statistical Association 98, 1001-1012.

[33] Peng, L. and Huang, Y. (2008). Survival analysis with quantile regressionmodels. Journal of the American Statistical Association 103, 637-649.

[34] Bottai, M. and Zhang, J. (2010). Laplace regression with censored data. Biometrical Journal 52, 487-503.

[35] Machado, J. A. F. and Santos-Silva, J. M. C. (2005). Quantiles for counts. Journal of the American Statistical Association 100, 1226-1237.

[36] van Buuren, S., Boshuizen, H., and Knook, D. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. Statistics in medicine 18, 681-694.

[37] Gifi, A. (1990). Nonlinear multivariate analysis. Wiley, New York.

[38] Koenker, R. (2005). Quantile Rregression. Cambridge University Press.

[39] Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J., and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey Methodology 27, 85-95.

[40] R Development Core Team (2008). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

[41] van Buuren, S. and Groothuis-Oudshoorn, K. (2010). Mice: Multivariate imputation by chained equations in R. Journal of Statistical Software, forthcoming .

[42] Caspersen, C. J., Powell, K. E., and Christenson, G. M. (1985). Physical activity, exercise, and physical fitness: definitions and distinctions for health-related research. Public Health Rep 100, 126-131.

Table 1. Bias, variance, and mean squared error of the estimators of the regression coefficients and the marginal mean and variance for the proposed quantile multiple imputation (QMI) and predictive mean matching (PMM), under missing completely at random mechanism with marginal probability of non-response 0.3, when the error term is generated from a χ^2_1 distribution with sample size 500.

	Bias		Variance		Mean Squared Error		
	QMI	PMM	QMI	PMM	QMI	PMM	Ratio
0.25-Quantile Regression							
β_0	-0.0014	0.0212	0.0017	0.0045	0.0010	0.0040	3.9681
β_1	0.0009	0.0019	0.0040	0.0143	0.0026	0.0120	4.5851
β_2	-0.0001	-0.0035	0.0005	0.0013	0.0003	0.0009	3.2463
β_3	0.0025	-0.0041	0.0002	0.0006	0.0002	0.0005	2.9734
0.50-Quantile Regression							
β_0	-0.0054	0.0242	0.0130	0.0191	0.0080	0.0131	1.6378
β_1	0.0038	0.0040	0.0305	0.0496	0.0199	0.0347	1.7428
β_2	-0.0025	-0.0050	0.0035	0.0056	0.0022	0.0037	1.6554
β_3	0.0058	0.0069	0.0017	0.0024	0.0009	0.0017	1.7548
0.75-Quantile Regression							
β_0	-0.0178	-0.0100	0.0722	0.0731	0.0484	0.0502	1.0372
β_1	0.0109	0.0117	0.1585	0.1460	0.1176	0.1188	1.0107
β_2	0.0020	0.0027	0.0180	0.0192	0.0118	0.0136	1.1551
β_3	0.0074	0.0110	0.0086	0.0101	0.0052	0.0072	1.3987
Least-Squares Regression							
β_0	-0.0125	-0.0072	0.0268	0.0362	0.0144	0.0180	1.2529
β_1	0.0077	0.0121	0.0584	0.0820	0.0359	0.0433	1.2044
β_2	0.0003	-0.0039	0.0073	0.0102	0.0040	0.0054	1.3685
β_3	0.0049	-0.0031	0.0033	0.0055	0.0016	0.0034	2.1297
Marginal Mean and Variance							
μ	-0.0026	-0.0044	0.0108	0.0112	0.0024	0.0026	1.1021
σ^2	-0.0063	-0.0249	0.4992	0.5176	0.1219	0.1367	1.1208

Table 2. Bias, variance, and mean squared error of the estimators of the regression coefficients and the marginal mean and variance for the proposed quantile multiple imputation (QMI) and predictive mean matching (PMM), under missing completely at random mechanism with marginal probability of non-response 0.3, when the error term is generated from a t_3 distribution with sample size 500.

	Bias		Variance		Mean Squared Error		
	QMI	PMM	QMI	PMM	QMI	PMM	Ratio
0.25-Quantile Regression							
β_0	-0.0628	-0.0050	0.0408	0.0392	0.0317	0.0296	0.9323
β_1	-0.0133	-0.0193	0.1024	0.1005	0.0702	0.0826	1.1756
β_2	-0.0023	-0.0090	0.0126	0.0127	0.0078	0.0086	1.0962
β_3	0.0024	-0.0100	0.0051	0.0058	0.0033	0.0039	1.1864
0.50-Quantile Regression							
β_0	0.0024	0.0091	0.0293	0.0309	0.0170	0.0210	1.2346
β_1	-0.0041	-0.0082	0.0755	0.0839	0.0430	0.0570	1.3239
β_2	-0.0019	-0.0083	0.0087	0.0100	0.0051	0.0064	1.2498
β_3	0.0020	-0.0089	0.0033	0.0038	0.0021	0.0027	1.2584
0.75-Quantile Regression							
β_0	0.0698	0.0242	0.0428	0.0426	0.0287	0.0280	0.9760
β_1	-0.0031	-0.0018	0.1052	0.1023	0.0638	0.0758	1.1885
β_2	-0.0043	-0.0093	0.0126	0.0128	0.0074	0.0082	1.1197
β_3	0.0045	-0.0069	0.0051	0.0053	0.0031	0.0033	1.0715
Least-Squares Regression							
β_0	-0.0011	0.0036	0.0437	0.0488	0.0220	0.0267	1.2158
β_1	-0.0043	-0.0037	0.1100	0.1271	0.0593	0.0721	1.2151
β_2	-0.0008	-0.0081	0.0130	0.0150	0.0065	0.0081	1.2431
β_3	0.0140	-0.0135	0.0070	0.0109	0.0055	0.0087	1.5733
Marginal Mean and Variance							
μ	-0.0015	-0.0021	0.0157	0.0155	0.0037	0.0039	1.0622
σ^2	-0.0492	-0.0993	4.7169	6.2335	0.5974	1.2979	2.1726

Table 3. Bias, variance, and mean squared error of the estimators of the regression coefficients and the marginal mean and variance for the proposed quantile multiple imputation (QMI) and predictive mean matching (PMM), under missing completely at random mechanism with marginal probability of non-response 0.3, when the error term is generated from a $N(0,1)$ distribution with sample size 500.

	Bias		Variance		Mean Squared Error		
	QMI	PMM	QMI	PMM	QMI	PMM	Ratio
0.25-Quantile Regression							
β_0	0.0012	-0.0004	0.0262	0.0245	0.0145	0.0154	1.0617
β_1	0.0026	0.0041	0.0645	0.0568	0.0382	0.0395	1.0354
β_2	-0.0004	-0.0026	0.0074	0.0074	0.0039	0.0044	1.1060
β_3	0.0011	-0.0036	0.0058	0.0060	0.0035	0.0038	1.0941
0.50-Quantile Regression							
β_0	0.0009	0.0046	0.0211	0.0215	0.0110	0.0117	1.0704
β_1	0.0010	-0.0017	0.0557	0.0528	0.0279	0.0298	1.0695
β_2	-0.0018	-0.0050	0.0070	0.0069	0.0043	0.0042	0.9838
β_3	0.0005	-0.0030	0.0050	0.0052	0.0028	0.0032	1.1352
0.75-Quantile Regression							
β_0	0.0022	0.0069	0.0234	0.0237	0.0139	0.0146	1.0519
β_1	-0.0016	-0.0061	0.0603	0.0572	0.0359	0.0396	1.1038
β_2	-0.0004	-0.0027	0.0071	0.0068	0.0044	0.0048	1.0916
β_3	0.0009	-0.0020	0.0061	0.0061	0.0033	0.0038	1.1675
Least-Squares Regression							
β_0	-0.0031	-0.0026	0.0183	0.0186	0.0073	0.0080	1.0918
β_1	0.0042	0.0046	0.0458	0.0468	0.0192	0.0206	1.0729
β_2	-0.0003	-0.0018	0.0052	0.0054	0.0022	0.0025	1.1266
β_3	0.0012	-0.0035	0.0041	0.0043	0.0017	0.0020	1.1378
Marginal Mean and Variance							
μ	0.0003	0.0005	0.4990	0.4986	0.0009	0.0009	1.0247
σ^2	0.0021	-0.0025	0.0911	0.0919	0.0229	0.0255	1.1115

Table 4. Bias, variance, and mean squared error of the estimators of the regression coefficients and the marginal mean and variance for the proposed quantile multiple imputation (QMI) and predictive mean matching (PMM), under missing at random mechanism with missing probability of non-response $p = 1/[1 + \exp(1 - x_1)]$, when the error term is generated from a χ^2_1 distribution with sample size 500.

	Bias		Variance		Mean Squared Error		
	QMI	PMM	QMI	PMM	QMI	PMM	Ratio
0.25-Quantile Regression							
β_0	-5e-04	0.0348	0.0019	0.0073	0.0012	0.0067	5.5744
β_1	8e-04	-0.0030	0.0049	0.0317	0.0035	0.0262	7.4684
β_2	-6e-04	-0.0045	0.0006	0.0022	0.0004	0.0018	4.3195
β_3	4e-03	-0.0099	0.0004	0.0014	0.0002	0.0013	5.5015
0.50-Quantile Regression							
β_0	0.0008	0.0347	0.0140	0.0230	0.0098	0.0177	1.8024
β_1	-0.0068	0.0108	0.0314	0.0691	0.0278	0.0570	2.0502
β_2	0.0017	-0.0009	0.0037	0.0064	0.0029	0.0049	1.6651
β_3	0.0056	0.0035	0.0021	0.0034	0.0015	0.0024	1.6816
0.75-Quantile Regression							
β_0	0.0019	0.0160	0.0734	0.0809	0.0556	0.0612	1.1015
β_1	-0.0005	0.0004	0.1665	0.1714	0.1520	0.1704	1.1207
β_2	-0.0043	-0.0063	0.0194	0.0219	0.0165	0.0182	1.1032
β_3	0.0058	0.0107	0.0092	0.0119	0.0067	0.0096	1.4454
Least-Squares Regression							
β_0	-0.0027	0.0128	0.0300	0.0440	0.0173	0.0239	1.3855
β_1	0.0047	-0.0106	0.0725	0.1144	0.0522	0.0700	1.3416
β_2	-0.0017	-0.0077	0.0084	0.0129	0.0056	0.0078	1.3903
β_3	0.0026	-0.0092	0.0036	0.0077	0.0024	0.0057	2.3696
Marginal Mean and Variance							
μ	0.0020	-0.0042	0.0114	0.0114	0.0037	0.0040	1.0671
σ^2	0.0089	-0.0305	0.5332	0.5342	0.1868	0.1957	1.0474

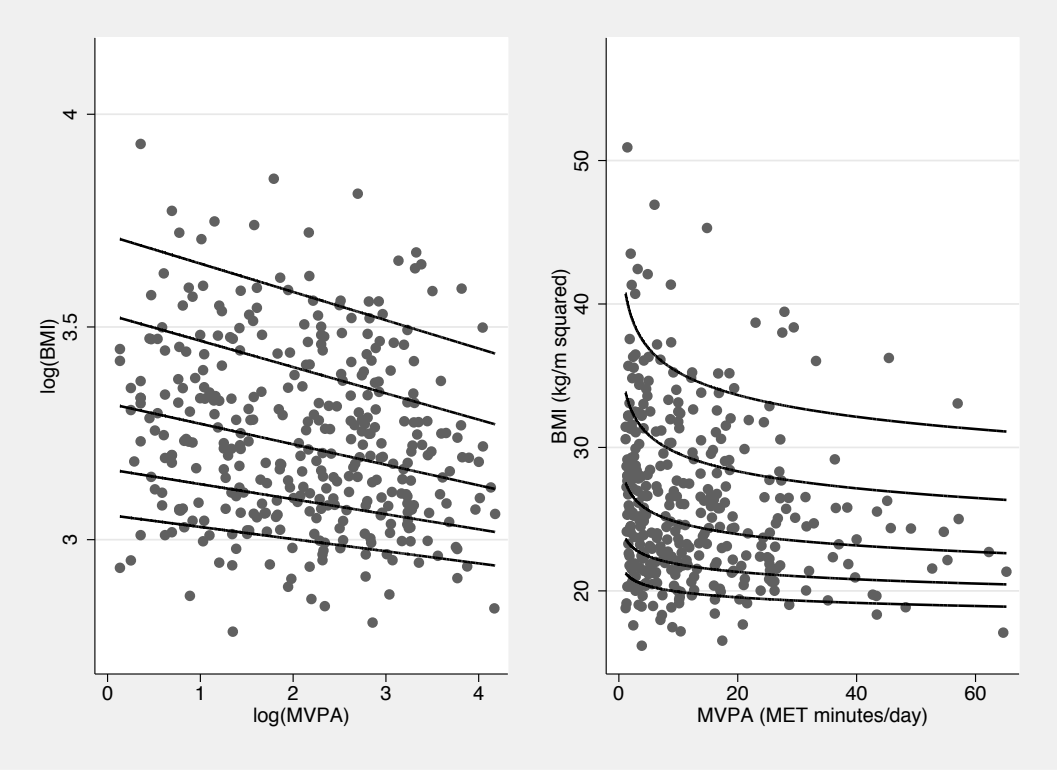

Figure 1. Scatter plot of body mass index (BMI) versus moderate-to-vigorous physical activity (MVPA) on double-log (left-hand-side panel) and natural (right-hand-side panel) scale. The solid lines from bottom to top represent the 10th, 25th, 50th, 75th, and 90th percentiles at age 29, sample median age, estimated by multiple imputation.