Original Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–29
© The Author(s) 2013
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0962280213484401
smm.sagepub.com
SAGE

Estimation of regression quantiles in complex surveys with data missing at random: An application to birthweight determinants

Marco Geraci

Abstract

The estimation of population parameters using complex survey data requires careful statistical modelling to account for the design features. This is further complicated by unit and item nonresponse for which a number of methods have been developed in order to reduce estimation bias. In this paper, we address some issues that arise when the target of the inference (i.e. the analysis model or model of interest) is the conditional quantile of a continuous outcome. Survey design variables are duly included in the analysis and a bootstrap variance estimation approach is proposed. Missing data are multiply imputed by means of chained equations. In particular, imputation of continuous variables is based on their empirical distribution, conditional on all other variables in the analysis. This method preserves the distributional relationships in the data, including conditional skewness and kurtosis, and successfully handles bounded outcomes. Our motivating study concerns the analysis of birthweight determinants in a large UK-based cohort of children. A novel finding on the parental conflict theory is reported. **R** code implementing these procedures is provided.

Keywords

chained equations, Khmaladze tests, multiple imputation, paediatrics, weights

1 Introduction

This paper offers general guidance for conducting quantile regression (QR) analysis of complex survey data. We start considering the case in which regression quantiles are estimated from a sample of observations taken from a finite population using a complex design. We then consider estimation issues that arise when several variables of interest are partially observed. Our motivating example is a study of determinants of birthweight in children of the UK Millennium Cohort Study (MCS), a longitudinal survey of British children born at the beginning of the 21st century.[1]

Centre for Paediatric Epidemiology and Biostatistics, Institute of Child Health, University College London, UK

Corresponding author:
Marco Geraci, Centre for Paediatric Epidemiology and Biostatistics, Institute of Child Health, University College London, 30 Guilford Street, London WC1N 1EH, UK.
Email: m.geraci@ucl.ac.uk

Downloaded from smm.sagepub.com at University College London on April 29, 2015

Conditional quantiles for continuous response variables[2] have a long history in econometrics and they have been applied in many other research fields.[3] Their ability to model the location, scale and shape of the observed distribution of the response variable conditional on a set of predictors offers the opportunity to investigate a wide range of effects that location-shift models are unable to handle. In addition, QR is inherently robust to outliers in the outcome and mathematically flexible in handling transformations of the response variable that can be exploited for a variety of purposes.[4–7] In contrast, ordinary least squares methods can overshadow important associations between the predictors and the outcome distribution.

There are a number of models for conditional quantiles which, in a sense, mirror regression models that have been developed for the estimation of the location parameter, including parametric, nonparametric[8–10] and semiparametric[11] models, linear and nonlinear models and models for discrete[4,12] and survival data.[13] An overview of QR inferential and computational issues is given in Koenker's monograph.[14]

The number of applications of QR in medicine has increased in recent years. In particular, QR has been applied in paediatric research where both methodological and applied studies have led to improvements on several fronts, including the understanding of birthweight determinants,[15–19] child growth and obesity,[20–24] malnutrition,[25] cancer aetiology[26] and hypertension,[27] where the common denominator is the analysis of the tails of distributions of variables strongly associated with adverse health risks (e.g. birthweight and blood pressure).

The data used in these studies generally come from hospital registrations and national healthcare systems data collections. However, data from population-based multi-purpose surveys are another important source of information since they provide details on health and socioeconomic conditions of children, their families and the environment in which they live. Moreover, longitudinal surveys allow exploring the impact of early life exposures on long-term health outcomes from a life course perspective.

The analysis of survey data involves the estimation of population parameters and related measures of uncertainty by using models that handle design features such as probability sampling weights, stratification and cluster sampling. Accounting for unequal inclusion probabilities is necessary in order to obtain consistent estimates. A number of methodological advances have been made in the development of multivariable models for correlational studies including linear, logistic and probit regression models, survival analysis models and structural equation models.[28] The estimation of finite population quantiles has attracted much attention in the literature of survey analysis[29–32] and, recently, a general asymptotic theory for nondifferentiable survey estimators, including sample quantile estimators, has been proposed.[33] In general, there are few applications of QR in studies based on complex survey data.[34,35]

Inference for survey estimators is further complicated by nonresponse, a common problem in survey studies. This can affect all items of the survey for units who do not respond or refuse to participate in an interview (unit nonresponse), although some information for those units is typically available prior to interview and can be used for nonresponse adjustments; or it can involve only certain items due to early drop-out from the study, or questionnaire items filled with 'don't know' or 'refused' (item nonresponse).[36] The reduction of available information that follows from nonresponse has effects of varying gravity, depending on the type and degree of missingness. The much-celebrated Little and Rubin's classification of missing data[37] clarifies the conditions under which inference can lead to estimation bias if the missing process is not taken into account. In particular, missing at random (MAR) assumptions are often introduced and their tenability may be sustained by sensible modelling choices. MAR mechanisms can be thought as a middle-way between missing completely at random (MCAR) and missing not at random (MNAR) mechanisms.

Downloaded from smm.sagepub.com at University College London on April 29, 2015

Here, we consider a multiple imputation (MI) strategy based on sequential conditional regressions as opposed to joint modelling.[38] The former approach avoids specifying a joint distribution for the imputation model by using a sequence of conditional specifications for each variable whose missing values are to be imputed. This strategy is advantageous when several variables, continuous and discrete, are included in the analysis model. In particular, missing values of continuous variables are imputed using conditional quantile models.[39] A distribution-free imputation procedure based on nonparametric kernel regression to estimate the distribution function and quantiles of an incomplete response variable under MAR assumptions has been developed.[40] Other approaches to apply QR in the presence of missing data have been proposed.[41–43]

The rest of the paper is organized as follows. In Section 2, we describe the MCS data and the sampling design. In Section 3, we introduce the model of interest and related methods of analysis, including a complete case analysis of the MCS birthweight determinants. We then present an imputation procedure for when nonmonotone missingness affects both continuous and discrete variables, followed by a simulation study to assess the imputation models (Section 4). We conclude with an MI analysis of the MCS birthweight determinants (Section 5) and final remarks (Section 6).

The code to run the simulation study and the MCS analyses was written in `R`[44] and the following packages were used: `survey`,[45,46] `quantreg`[47] and `mice`.[48] Additional `R` code for custom-defined functions to implement the methods proposed in this paper is provided in Appendix B.

2 The data

The MCS is a longitudinal study of a cohort of UK children born between September 2000 and January 2002.[1] Here, we provide a background on its sampling design and briefly describe the data selected for the analysis.

The survey population was defined as all children alive and living in the UK at age 9 months and eligible to receive Child Benefit at that age[49] (all UK residents qualify for Child Benefit if they have children younger than 16 years). Details on the sampling population resulted from exclusion of children who died before 9 months of age is given by Cullis.[50] The population was stratified by UK country and, in order to adequately represent disadvantaged and ethnic minority children, stratification by these variables within country was carried out using data available at the electoral ward level.[49] In particular, population in England was stratified by: 'ethnic', children living in wards that, in the 1991 Census of Population, had an ethnic minority population of at least 30% of the total; 'disadvantaged', children living in wards other than 'ethnic' which fell into the upper quartile of the ward-based Child Poverty Index (i.e. poorest 25%) for England and Wales; and 'advantaged', children not living in wards classified as 'ethnic' or 'disadvantaged'. Wales, Scotland and Northern Ireland populations were stratified by 'disadvantaged' and 'advantaged' wards.

Families were taken from a random sample of electoral wards, the primary sampling unit (PSU), disproportionately stratified to ensure an adequate representation of disadvantaged areas and ethnic minority groups. Details on the calculation of sampling weights and adjustment for unit nonresponse are given by Plewis.[49,51]

Parents/guardians of the children were interviewed when the children were aged nine months, three, five and seven years. We abstracted data from the first sweep of the survey on birthweight, gestational age and sex of singletons for whom the main respondent at the interview and the respondent's partner were the natural parents. This gave information for 15,070 children, with the main respondent being the mother in 15,060 cases. Additional information for the mothers

Downloaded from smm.sagepub.com at University College London on April 29, 2015

comprised reported weight before pregnancy, reported height, age at delivery, parity, highest educational degree attained (including GCSE, A-level/Diploma and academic), tobacco consumption habits before, during and after pregnancy, marital status, ethnicity, antenatal care received and diabetes status. Reported weight and height of children's natural fathers were also included. Some of the questionnaire items that were missing at the first sweep were retrieved from either the second or the third sweep if available.

A summary of the dataset, including the number of missing values, is given in Table 1. The number of incomplete cases (i.e. with at least one missing item) was 3066, 20% of the sample. Paternal weight had the highest proportion of missing data (15%). For this variable and for maternal weight, we replaced 30 outliers with missing values (more details in Section 5.4).

Table 1. Details of variables included in the analysis for 15,070 MCS children.

No.	Variable	Description	Missing	Reference value/baseline	Range/No. categories
1	bw	Birthweight (g) (response, positive continuous)	37		390 – 7230
2	gestAge	Gestational age (weeks) (positive continuous)	151	37	23 – 42
3	$weight_m$	Weight of the mother (kg) (positive continuous)	815[a]	64	34 – 130
4	$height_m$	Height of the mother (cm) (positive continuous)	122	164	102 – 206
5	$weight_f$	Weight of the father (kg) (positive continuous)	2314[a]	82	41 – 154
6	$height_f$	Height of the father (cm) (positive continuous)	1495	178	120 – 213
7	age	Mother's age at delivery (years) (positive continuous)	3	30	13 – 49
8	parity	Parity (counts)	0	0	0 – 9
9	cigDecrease	Reduction in number of cigarettes smoked daily by the mother after knowing to be pregnant (semi-continuous)	11	0	0 – 60
10	IsSmoker	Indicator variable for maternal smoking (binary)	27	Non-smoker	2
11	IsNotEdu	Indicator variable for maternal educational level (binary)	54	GCSE or higher	2
12	IsGirl	Indicator variable for female (binary)	0	Male	2
13	IsNotMarried	Indicator variable for parents who are not married (binary)	0	Married	2
14	IsNotWhite	Indicator variable for maternal ethnicity other than white (binary)	48	White	2
15	IsNotCare	Indicator variable for mothers who did not receive prenatal care (binary)	33	Care received	2
16	IsDiabetic	Indicator variable for maternal diabetes status (binary)	8	Non-diabetic	2

[a]Including missing values replacing outliers.

Downloaded from smm.sagepub.com at University College London on April 29, 2015

Reference values or baselines and variable range or number of categories are detailed as they will be used in subsequent analyses.

2.1 Survey weights and finite population correction

We now provide a brief note on additional sampling details. MCS sampling weights were calculated proportionally to the fraction of electoral wards selected by systematic sampling of the population wards ordered by size within strata. Although an implicit stratification by region and ward size was introduced, this had a marginal design effect.[49] It follows that children in the same stratum received the same weight. Also, a second set of survey weights was calculated to adjust for unit nonresponse in the issued sample. Unit nonresponse is a source of concern in this specific case. In fact, we might expect that children with very low birthweights are not well represented in the MCS sample (e.g. parents refused to participate as their child was receiving postnatal care) and, thus, that a MNAR mechanism is at play. Also, infant mortality before 9 months of age[50] may have introduced additional upward bias in the left tail of the distribution. An approximate calculation based on population birthweight data (further details available upon request) suggests that only $5-6\%$ of the low birthweights not represented in the MCS sample may be accounted for by infants who died before becoming eligible for the survey.

Unless explicitly stated, we will make use of sampling weights adjusted for unit nonresponse in all subsequent analyses of the MCS data.

3 QR in complex surveys

3.1 Population model

Initially, we focus on the population model, assuming that all units have been completely observed. Let $(\mathbf{y}, \mathbf{X}, \mathcal{F})$ be the data for a sample of size n taken from a population of size N, where $\mathbf{y}$ denotes an $n \times 1$ continuous response variable, $\mathbf{X}$ denotes a $n \times q$ set of predictors with row vectors $\mathbf{x}_i'$, $i = 1, \ldots, n$ and $\mathcal{F}$ collects sampling design variables. Also, let $F_{y|X,\mathcal{F}}$ denote the unknown cumulative distribution function of y given $(\mathbf{X}, \mathcal{F})$. Our inferential target is the population conditional quantile function $Q_{y|X,\mathcal{F}} \equiv F^{-1}_{y|X,\mathcal{F}}$. That is, we want to make a statement about specific quantiles of the distribution of y conditional on predictors $\mathbf{X}$ and design $\mathcal{F}$. Matrices will be denoted with upper case bold letters (e.g. $\mathbf{U}$), while column and row vectors with lower case bold letters (e.g. $\mathbf{u}$ and $\mathbf{u}'$). Depending on the context, the symbol $\sim$ will be used interchangeably to indicate 'distributed as', 'approximately equal to' or as response-covariate separator in a linear model formula (e.g. $y \sim x$).

We will only consider the case in which $\mathcal{F}$ defines a stratified clustered design since this represents the design that motivated our study. Let $h = 1, \ldots, H$ index population fixed strata and let m_h be the size of a sample of clusters (e.g. counties or electoral wards) randomly selected within stratum h (these also represent PSUs in our modelling). Survey weights are introduced to account for selection probability and, possibly, unit nonresponse. The following developments can be easily extended to account for other designs.

Suppose that the sample $(\mathbf{y}, \mathbf{X})$ was obtained under simple random sampling (SRS) without replacement from a large N and $n \ll N$ (i.e. sampling fraction very small) and that we wanted to fit the p-th conditional quantile function

$$Q_{y|X,\mathcal{F}}(p) = \mathbf{X}\boldsymbol{\beta}(p), \quad p \in (0, 1), \tag{3.1}$$

Downloaded from smm.sagepub.com at University College London on April 29, 2015

where $\boldsymbol{\beta}(p)$ is a $q \times 1$ vector of regression coefficients indexed by p. If we ignore the unequal sampling probability, the p-th regression quantile $\boldsymbol{\beta}(p)$ could be the estimated by minimizing the loss function[2]

$$\sum_{i=1}^{n} \rho_p\left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right) \tag{3.2}$$

with respect to $\boldsymbol{\beta}$, where $\rho_p(s) = s\left\{p - I(s < 0)\right\}$ and $s \in \mathbb{R}$, and the standard inferential theory for regression quantiles could be applied with some approximation. See Chatterjee[32] for a recent overview of inferential issues related to marginal quantiles estimated from samples in finite populations.

With data obtained from designs more complex than SRS, the estimation of population statistics and their uncertainty measures needs to take into account the design. Stratification and clustering can in principle be adjusted for by including the relevant sampling variables in the linear predictor. Generalized linear mixed models, for example, allow modelling the intra-class correlation due to survey clustering by means of random effects. Alternatively, cluster-specific parameters can be entered as fixed effects though care must be taken in controlling for standard error inflation through some form of parameter shrinkage. QR methods for repeated and clustered data have been proposed,[52–54] although models for complex hierarchical structures (e.g. more than two levels of nesting or cross-classified multilevel models) are yet to be developed. In general, dealing with survey weights in hierarchical models can be challenging.[28]

Weighting in survey estimators is used to account for unequal inclusion probabilities assigned to sample observations. In a QR context, we consider the following weighted loss function:

$$\sum_{i=1}^{n} w_i \rho_p\left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right), \tag{3.3}$$

where w_i are survey weights possibly adjusted for unit nonresponse.

The unknown regression coefficient $\boldsymbol{\beta}$ in (3.2) and (3.3) can be estimated, for example, using well-developed linear programming algorithms[47] or gradient search methods,[55] which are computationally fast. The former include classical simplex methods, suitable for problems of small to moderate size, and interior-point methods, such as Frisch–Newton algorithms,[56] recommended for large problems. As opposed to least squares estimators, quantile estimators involve nondifferentiable functions of the quantities to be estimated. The usual Taylor linearization used to obtain an approximate estimate of the variance of survey estimators cannot therefore be applied. A practical way to overcome this computational issue is to use bootstrap estimation. In Appendix A, we briefly describe the method proposed by Canty and Davison.[57–59] We also discuss preliminary results using Wang and Opsomer's[33] approach.

Details on how to fit QR models by minimizing (3.3) and estimate bootstrap standard errors in `R` are given in Appendix B. The application of these methods using the MCS data is described in the following section.

3.2 Birthweight determinants

Birthweight has long been recognized as an important surrogate for infant well-being. In particular, birthweight is a strong predictor of infant mortality and morbidity and thus provides a well-established base for clinical indicators. Least squares methods are commonly used in birthweight

Downloaded from smm.sagepub.com at University College London on April 29, 2015

analyses to assess the effects of determinants such as maternal health-related lifestyle habits (e.g. smoking and diet) and environmental factors (e.g. pollution). Mean regression, however, is generally unable to reveal effects of these determinants on the tails of the distribution which are of particular interest since small and macrosomic babies are at increased risk of morbidity and mortality. At the population level, studies in different countries show that birthweight follows approximately a normal distribution with an elongated left tail. The latter is mostly explained by preterm babies who tend to weigh less than babies born after 37 weeks of gestation. However, even at the population level, *conditional* birthweight distributions may be far from normal and distributional effects may be more complex than those explained by location-shift models.

Early QR analyses of population birthweight can be found in Abrevaya[15] and Koenker and Hallock[60] who analysed birthweights for about 200,000 US babies from the Detailed Natality Data (National Center for Health Statistics). In their analyses, the QR estimates associated with factors such as maternal smoking, age and weight gain were not constant across the conditional birthweight distribution, suggesting that these predictors may exert complex distributional effects. Cases with missing data were excluded. See also Chernozhukov and Fernández-Val[61] for an interesting application of extremal QR (i.e. very low and very high quantiles) and Abrevaya and Dahl[16] for the analysis of panel data on maternally linked births which takes into account unobserved characteristics of the mothers.

We are interested in estimating the quantiles of MCS birthweights conditional on determinants listed in Table 1. Here, we follow a conditional approach for gestational age mindful that the latter is a potential mediator of other effects. Joint modelling approaches to birthweight and gestational age have been proposed.[62] Figure 1 provides a breakdown of the birthweight histogram by gestational age. As expected, the distribution is shifted to the left for preterm infants (< 37 weeks) and departs from normality. Figure 1 also shows the histogram of gestational age, whose distribution is highly asymmetric and leptokurtic.

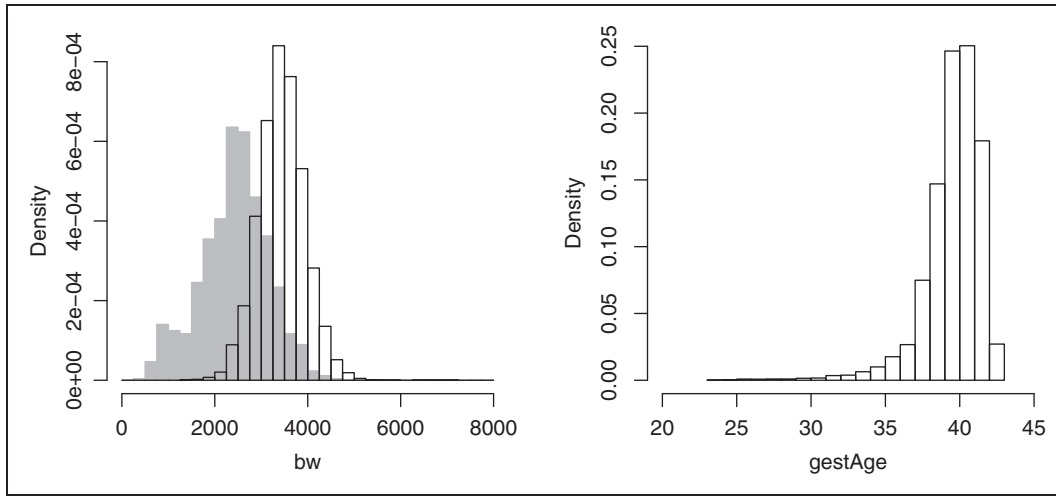

Figure 1. Left: histogram of birthweight. Children born at 37 weeks or later (white) and children born before term (grey) are contrasted. Right: histogram of gestational age.

Downloaded from smm.sagepub.com at University College London on April 29, 2015

We now proceed with a complete case analysis using the methods described in Section 3.1, under MCAR assumptions. After excluding partially observed units, there were 12,004 observations available for the analysis. The quantiles $p = 0.05$ and $p = 0.01$ of the observed birthweights were equal to, respectively, 2410 and 1626 g, approximately corresponding to standard thresholds for low (< 2500 g) and very low (< 1500 g) birthweight.

All continuous predictors were centred at their mean value, with the exception of gestational age which was centred at 37 weeks. Moreover, parental weight and height were divided by their standard deviation (internal standardization) to obtain z-scores. These scores were further reparametrized as in Griffiths et al.[63] to study the effects of differential parental weight and height contributions on birthweight, namely the effect of the half difference $(\text{w/h})\text{eight}_{hd} = \big((\text{w/h})\text{eight}_m - (\text{w/h})\text{eight}_f\big)/2$ and the independent effect of the mean (i.e. sum) $(\text{w/h})\text{eight}_{sum} = (\text{w/h})\text{eight}_m + (\text{w/h})\text{eight}_f$.

We considered the following QR model:

$$\begin{aligned} Q(p) = {} & \beta_0(p) + \beta_1(p)\text{gestAge} + \beta_2(p)\text{weight}_{hd} + \beta_3(p)\text{weight}_{sum} + \beta_4(p)\text{height}_{hd} \\ & + \beta_5(p)\text{height}_{sum} + \beta_6(p)\text{age} + \beta_7(p)\text{parity} + \beta_8(p)(-\text{cigDecrease}) + \beta_9(p)\text{IsSmoker} \\ & + \beta_{10}(p)\text{IsNotEdu} + \beta_{11}(p)\text{IsGirl} + \beta_{12}(p)\text{IsNotMarried} + \beta_{13}(p)\text{IsNotWhite} \\ & + \beta_{14}(p)\text{IsNotCare} + \beta_{15}(p)\text{IsDiabetic}. \end{aligned} \tag{3.4}$$

The intercept $\beta_0(p)$ can be interpreted as the birthweight p-th quantile for male children born at 37 weeks gestation from: married parents whose difference/mean weight/height z-scores are zero; 30-year-old, white, non-diabetic, zero-parity mothers with GCSE educational level or higher, who are non-smokers and who received prenatal care.

A sequence of 21 regression quantiles, $p \in \{0.01, 0.05, 0.1, \ldots, 0.9, 0.95, 0.99\}$, was then obtained using the methods illustrated in Section 3.1. In particular, the loss function (3.3) was weighted using MCS sampling weights. Since sampling fractions were as high as 18% due to oversampling, variance estimates were adjusted with a finite population correction. A bootstrap sample size $n_B = 100$ was used to estimate the variance of the regression quantiles.

Table 4 shows point estimates and standard errors for selected quantiles (results for all quantiles are available upon request). A discussion of the results is deferred to Section 5.

4 QR and missing data

4.1 Missing data modelling

Our estimand of interest is, again, a model of the type (3.1). However, we assume that $\mathbf{y}$ and $\mathbf{X}$ are not fully observed. Now let $(\mathbf{y}, \mathbf{X}, \mathcal{F}, \mathbf{R})$ be the data and define the augmented $n \times (q+1)$ matrix $\mathbf{Z} = (\mathbf{y}, \mathbf{X})$, with column vectors $\mathbf{z}_1, \ldots, \mathbf{z}_{q+1}$. The missing indicator matrix $\mathbf{R}$ contains nonresponse information for $\mathbf{Z}$ and has columns $\mathbf{r}_1, \ldots, \mathbf{r}_{q+1}$ whose elements $r_{i,j}$, $i = 1, \ldots, n$, $j = 1, \ldots, q+1$, take the value 1 when the corresponding item $z_{i,j}$ is observed and 0 otherwise. Let n_j and $\bar{n}_j = n - n_j$ denote the number of observed and missing values in $\mathbf{z}_j$, respectively, and let A_j be the set indexing the units i for which $r_{i,j} = 0$. In addition, let *obs* and *mis* denote restriction of variables to their observed and missing parts, respectively. Under MAR assumptions, the distribution of $\mathbf{R}$ conditional on $\mathbf{Z}_{obs}$ is independent of $\mathbf{Z}_{mis}$

$$\Pr(\mathbf{R}|\mathbf{Z}_{obs}, \mathbf{Z}_{mis}, \xi) = \Pr(\mathbf{R}|\mathbf{Z}_{obs}, \xi), \tag{4.1}$$

where ξ is a vector of unknown fixed parameters.[64]

Downloaded from smm.sagepub.com at University College London on April 29, 2015

Let $\mathbf{Z}_{-j}$ denote $\mathbf{Z}$ without the j-th column and $\theta = \left(\theta_1', \ldots, \theta_{q+1}'\right)'$ be a partitioned parameter vector, distinct from $\boldsymbol{\xi}$. Without loss of generality, suppose for the moment that $\mathbf{Z}_{-j}$ is completely observed. Equation (4.1) implies that, conditionally on $\mathbf{Z}_{-j}$, the distribution of the variable $\mathbf{z}_j$ is the same among the cases for which $\mathbf{z}_j$ is observed as it is among the cases for which $\mathbf{z}_j$ is missing,[64] that is

$$F_{obs}\left(z|\mathbf{Z}_{-j}, \mathcal{F}, \theta_j\right) = F_{mis}\left(z|\mathbf{Z}_{-j}, \mathcal{F}, \theta_j\right), \tag{4.2}$$

which corresponds to the assumption of ignorability.[65] This identity allows us to use the conditional distribution function of respondents to draw independent samples to be imputed to $\mathbf{Z}_{mis}$. The MAR assumption provides the basis for a consistent estimation of the conditional distribution. Conditioning the probability model on the sampling design is considered in order to avoid potential imputation bias.[66]

Here, we adopt a fully conditional specification (FCS) in contrast to joint modelling. For an overview and discussion of these two approaches to MI, see Van Buuren[38] and references therein. The following three stages for QR estimation are considered:

(a) Draw m samples independently using, in turn, sequential conditional models $F_{obs}\left(z|\mathbf{Z}_{-j}, \mathcal{F}, \theta_j\right)$ for all variables $\mathbf{z}_j$ with missing values. Since in general the predictors $\mathbf{Z}_{-j}$ may be incomplete, a preliminary imputation of all missing values is carried out by SRS with replacement from the observed values.[67] The preliminary imputed values are then updated with the most recent draws from the sequential conditional models. The process may be repeated several iterations in order to stabilize the results.[67]
(b) Analyse each imputed dataset $\left(\mathbf{y}_k^*, \mathbf{X}_k^*, \mathcal{F}\right)$, $k = 1, \ldots, m$, using an appropriate QR model $Q_{y|X,\mathcal{F}}(p)$ to obtain $\widehat{\boldsymbol{\beta}}_k(p)$ and its standard error $\text{se}(\widehat{\boldsymbol{\beta}}_k(p))$. The latter can be calculated, for example, using a bootstrap approach (see Appendix A for more details).
(c) Pool the m estimates $\widehat{\boldsymbol{\beta}}_k(p)$ and $\text{se}(\widehat{\beta}_k(p))$ using Rubin's rules.

The preliminary imputation of Z_{-j} in step (a) is based on the assumption that units are exchangeable and such assumption is commonly violated in complex surveys. In the case of a stratified clustered sample, the preliminary imputation step could ideally be made more efficient by conditioning the sampling on stratification as well as clustering variables, provided that the number of observations available is sufficient for this purpose.

In the next two sections, we illustrate an imputation modelling approach to mixed (i.e. continuous and categorical) data which assumes a generalized missing pattern and takes the sampling design into account. Initially, we focus on continuous variables only; then, we briefly consider models for categorical variables.

4.2 Imputation models for continuous variables

Much of the literature on MI is dedicated to location models. A linear regression model (i.e. normal) is usually assumed for continuous data, whereas logistic, multinomial or log-linear models are used to impute categorical variables. The assumption that all the information is contained in the location parameter might, however, be too restrictive and lead to estimation bias associated with an inadequately chosen imputation model. In specific situations, distributional features like

Downloaded from smm.sagepub.com at University College London on April 29, 2015

heteroscedasticity and skewness can be accommodated by simple monotonic transformations. A general approach to handle conditional distributions that exhibit complex relationships with the model predictors is necessary when simpler strategies do not apply.

Some methods that deal with skewed continuous variables are described by White et al.,[67] including shifted-log and (marginal) Box–Cox transformations (BCT). Because of their inability to remove non-normal features other than skewness, such transformations cannot be generally qualified as methods for non-normal distributions. Moreover, they require estimation of unknown parameters from the observed data, in addition to the imputation model's parameters. In particular, White et al. suggested that if BCTs are applied to marginal rather than conditional distributions, this may not be relevant for the analysis. However, no formal discussion on this topic was provided.

Predictive mean matching[68] (PMM) is an *ad hoc* method that can be used effectively for imputation. Observed values whose predicted mean are closest to the predicted mean for the missing value are elected as 'donors' and a value taken randomly from the set of donors is imputed. In Schenker et al.,[69] partially parametric approaches to MI were compared to normal imputation. In particular, PMM and a modified PMM (local residual draw) were shown to be robust to model miss-specification and different (symmetric) error distributions. However, PMM requires a sufficient number of 'donors' from where to sample the values to be imputed. An interesting approach based on transformations of the target variable is described by He et al.,[70] where the Tukey's *gh* distribution is applied to accommodate for skewness and tail elongation in MI. This transformation was shown to be preferable to a log-transformation to reduce skewness. As in the case of BCT, the *gh* transformation requires estimating unknown parameters and a bootstrap approach was suggested. These methods[69,70] were explored in an MCAR framework. BCTs are also considered by Raghunathan et al.[71] for sequential regression multivariate imputation (SRMI) based on Metropolis–Hastings sampling. These authors warned about the computational burden of SRMI in analyses of large data sets with many variables, which are typical in survey studies.

To overcome some of the limitations of the methods mentioned above, we build an MI procedure starting from a distribution-free approach based on conditional quantile estimation proposed by Bottai and Zhen[39] in the context of SRS. Their approach exploits the well-known probability integral transformation theorem which is used in pseudo-random numbers generation: if $v \sim F$ and $u \sim \text{Unif}(0,1)$, then $F^{-1}(u) \sim F$, that is v and $F^{-1}(u)$ have the same distribution. This theorem and equation (4.2) provide our sampling framework for imputation. We apply it to continuous variables only although, in principle, extensions to discrete variables can be considered (see next section).

The aim is to impute $\bar{n}_j$ missing values of a partially observed continuous variable $\mathbf{z}_j$ within the FCS algorithm described in the previous section. A common imputation model for continuous responses is the *iid* linear model $z_{i,j} = \mathbf{Z}_{-j}\gamma + \eta_i$, with $\eta_i \sim \text{N}(0, \sigma^2)$ and $i \in A_j$. However, this location-shift model may be too restrictive in some situations. We therefore consider a distribution-free approach to sample from $F_{z_j|Z_{-j},\mathcal{F}}$, the distribution function of $\mathbf{z}_j$ conditional on $\mathbf{Z}_{-j}$ and $\mathcal{F}$:

(i) Obtain a sample $\mathbf{u}$ of size $\bar{n}_j$ independently from a standard uniform distribution. To avoid sampling in the vicinity of the boundaries which could cause computational inconveniences in step (ii), we can restrict the sampling domain to $\text{Unif}(\omega, 1-\omega)$ with ω sufficiently small, say $\omega = 0.001$. The parameter ω, as we shall see in the MCS birthweight analysis, can be used to trim or truncate the distribution of $z_{i,j}$.

Downloaded from smm.sagepub.com at University College London on April 29, 2015

(ii) Estimate the QR model

$$Q_{z_j|Z_{-j},\mathcal{F}}(u_i) = \mathbf{Z}_{-j}\gamma(u_i) \tag{4.3}$$

for each quantile u_i, $i \in A_j$.

(iii) Obtain and impute the value $z^*_{i,j} = \mathbf{Z}_{-j}\widehat{\gamma}(u_i)$, $i \in A_j$.

Suppose we fix $\omega = 0.001$ in step (i). The uniform values u can be grouped into small intervals between 0.001 and 0.999 (for a maximum number of 207 breakpoints, i.e. $u \in \{0.001, 0.002, \ldots, 0.005, 0.01, \ldots, 0.995, 0.996, \ldots, 0.999\}$) to reduce the number of quantiles that are too close one to each other and for which estimates would not differ substantially. Alternatively, rather than with pre-defined quantile breakpoints, a possibly more efficient computation could be obtained by estimating the conditional quantile function with a number of breakpoints determined by changes of solution. The latter, however, has been shown to grow with n.[72]

There are several advantages with a quantile-based imputation. Provided that the conditional quantile model is correctly specified, inferential results are valid without making assumptions about the regression error term η_i. The observed relationship between the covariates and the entire distribution of the imputed variable, not just its location parameter, is preserved. It is also worth noting that if we apply a monotone transformation h to the variable z, the equivariance property of the quantiles ensures that

$$Q_{h(z)}(p) = h\{Q_z(p)\}, \tag{4.4}$$

therefore $Q_z(p) = h^{-1}\{Q_{h(z)}(p)\}$.

This property is very useful if a transformation is applied to achieve linearity of the conditional model or to ensure that imputations lie within some interval (a, b), e.g. $(0, \infty)$ for strictly positive variables. Some authors refer to such pre-imputation transformations as pre-processing, followed by post-processing to transform the data back.[73] In our case, transformation and back-transformation take place within the estimation step (ii) (see function `mice.impute.rq` in Appendix B). As we shall see with the MCS data, transformation in MI is strictly related to diagnostics.

Suppose that z is gestational age. We could define the above interval using biologically plausible values (external bounds) or the observed range (internal bounds) as given in Table 1. We use the latter and then apply a logit function

$$h(z) \equiv \text{logit}\left(\frac{z-a}{b-a}\right) = \log\left(\frac{z-a}{b-z}\right),$$

where $a = 23 - 0.5$ and $b = 42 + 0.5$, to avoid taking logs of zero or infinity. A sample from the distribution of gestational age is simply obtained with

$$z^* = \frac{a + b \cdot \exp\left(\hat{Q}_{h(z)}\right)}{1 + \exp\left(\hat{Q}_{h(z)}\right)},$$

where z^* is bound to lie in the interval $(22.5, 42.5)$ weeks and $\hat{Q}_{h(z)}$ is the estimated quantile on the logit scale. The equivariance property has also been exploited to model censored[5,6] and bounded

Downloaded from smm.sagepub.com at University College London on April 29, 2015

outcomes.[7] Note that the equivariance property (4.4) does not apply to the expected value operator. As a consequence, transformations commonly used in mean regression[73] may introduce bias in the estimates. This topic is investigated in a simulation study (Section 4.5).

Recently, Wei et al.[43] proposed an MI estimator for regression quantiles with covariates MAR which was shown to have an advantage over complete-data methods. Their procedure shares some similarities with the QR-based procedure considered here in that their imputation method makes use of conditional quantile estimation of distribution functions as we do in equation (4.3). However, Wei et al.'s[43] estimation of the conditional density of missing covariates is partly based on the sparsity function and partly based on parametric modelling, which adds two layers of computation in the overall procedure. In addition, their MI approach is applied to independent missing covariates. In contrast, equation (4.3) can be applied to possibly dependent covariates as a result of the sequential conditional modelling approach.

4.3 Imputation models for discrete variables

Questionnaire-based surveys typically produce many variables classified as categorical. Logistic regression (binary), polytomous logistic regression (unordered categorical) and proportional odds models (ordered categorical) are typically the default choices. The former will be considered for the imputation of the MCS indicator variables (Table 1). However, it would be possible to extend the imputation model (4.3) to discrete data. Recent work in QR for count[4] and binary outcomes[74] offers interesting opportunities for developing QR-based logistic and log-linear imputation models, though computation time may increase appreciably.

4.4 Imputation model selection

There are, of course, precautions to bear in mind when applying QR-based imputation, part of which are common to MI methods in general and part are specific to quantile estimation.

More in general, the interplay between the imputation model and design features is engaging and interest lies in the 'extent [to which] the complexities of the sample design need to be incorporated into the imputation model'.[75] The consistency of the conditional quantile estimator guarantees that the asymptotic conditional distribution of each imputed value is equal to the conditional distribution of the unobserved values.[39] Here, we consider conditioning the imputation models indicated in Sections 4.2 and 4.3 on all relevant sampling design variables $\mathcal{F}$. As shown by Reiter et al.[66] in a simulation study, if the design features are related to the variable of interest but not taken into account by the imputation model, then imputations based on SRS lead to severe bias of the estimates and poor coverage of the confidence intervals. On the other hand, inclusion of irrelevant design variables may result in a loss of efficiency, though inferences will tend to be conservative. Reduced efficiency, therefore, seems to be a possibly reasonable insurance premium to pay against biased results. Yet, fitting imputation models with several strata and cluster effects, let alone their interactions with other variables, can be a formidable task. Reiter et al.,[66] for example, used a stepwise variable selection procedure. We do not pursue model selection issues here as this goes beyond the scope of our paper. It is worth stressing, however, that the application of standard techniques to discriminate among QR imputation models may not be appropriate as they focus on conditional means. A penalized approach to QR model selection has been proposed by Burgette et al.[18]

Downloaded from smm.sagepub.com at University College London on April 29, 2015

Related to this issue is the assumption that the same model applies to each quantile u. Since the application of u-specific model selection strategies for continuous values of u is unreasonable in practice, preliminary checks can be done, for example, by testing location and location-scale shift hypotheses and/or by testing subsets of variables over a specified range of quantiles u.[14, p. 95] An application of Khmaladze tests[76] is considered further in Section 5.

Finally, model misspecification may produce quantile crossing, though this will not in general represent an issue if crossing takes place outside the convex hull of the covariates.[14, p. 55] As stressed in Section 4.2, the correct specification of the imputation model is necessary for inferential results to be valid. Still, crossing may occur even when the model is correctly specified but the data are sparse in the region of interest.[77] We could expect, therefore, that for close values of u, neighbouring quantile curves will be affected by a mild form of crossing. For location and location-scale regression models, we can, however, ensure that a proper ordering of the quantiles is maintained by using restricted regression quantiles,[77,78] at the expense of additional computation time. We evaluate the performance of this adjustment by means of simulation in the next section. For other approaches to the problem of quantile crossing, see for example Chernozhukov et al.[79] and references therein.

4.5 Simulation study

In a simulation study, QR-based imputation was shown to be competitive as compared to Bayesian linear regression, PMM and unconditional mean imputation.[39] In this section, we assess the performance of this procedure in the specific case in which partially observed covariates undergo a transformation. We intentionally use a SRS design assuming an infinite population to obtain results that are easy to interpret.

We considered two data-generating models for the response y_i, $i = 1, \ldots, 1000$, namely the simple linear location-shift model

$$y_i = x_i + z_i + \epsilon_i, \tag{4.5}$$

and the heteroscedastic model

$$y_i = x_i + z_i + (1 + 2z_i)\epsilon_i. \tag{4.6}$$

We generated $x \sim \text{Unif}(0, 1)$ and $z \sim \chi^2_3/3$, independently. In both models, the error ϵ was generated from $\chi^2_3/3$. All variables are therefore strictly positive.

We then generated missing values for the variable z under a MAR mechanism by sampling $r_i \sim \text{Binom}(1000, 1 - p_i)$,

$$p_i = \frac{e^{1-\alpha y_i}}{0.1 + e^{1-\alpha y_i}},$$

where $\alpha = 2$ under model (4.5) and $\alpha = 1.5$ under model (4.6). A sample of missing values for the variable x of size $1000 - \sum r_i$ was taken without replacement under an MCAR mechanism, therefore independently from the other variables. This resulted in a nonmonotone missing data pattern: on average, x and z had 294 (range $262 - 331$) and 275 ($231 - 304$) missing values, but 501 ($447 - 559$) and 474 ($403 - 522$) overall for model (4.5) and (4.6), respectively.

Downloaded from smm.sagepub.com at University College London on April 29, 2015

Five missing data approaches were assessed: (a) complete case analysis (CC); (b) MI using linear QR as imputation model (QR); (c) as in (b), combined with pre-processing $\log x$ and $\log z$ (QR PP); (d) MI using linear regression as imputation model (LM); (e) as in (d), combined with pre-processing $\log x$ and $\log z$ (LM PP). Additionally, we assessed approach (b) using restricted QR (RQR).[77] For all MI methods, we set $m = 5$ imputed datasets and a maximum of five iterations for each imputation.

In both scenarios, we estimated the conditional quantile functions $Q_{y|x,z}(p) = \beta_0(p) + \beta_1(p)x + \beta_2(p)z$, $p \in \{0.1, 0.5\}$, where standard errors were calculated using the asymptotic variance estimator,[14] either for *iid* (4.5) or *nid* (4.6) errors. For the location-shift scenario only, we also estimated the linear conditional mean regression $\mathrm{E}(y|x,z) = \beta_0 + \beta_1 x + \beta_2 z$. In each setting, 200 replicated datasets were generated.

In Table 2, we report the Monte Carlo average of the β's point estimates obtained using the five missing data approaches (a–e) described above and, in addition, the average of the point estimates

Table 2. Average estimate of the regression coefficients and, in brackets, variance ratio for three analysis models using a complete case analysis (CC) and four different imputation models (QR, QR PP, LM, LM PP). The average model-based estimate using the full datasets (FD) is also reported.

	FD	CC	QR	QR PP	LM	LM PP
Data generated under location-shift model (4.5)						
Mean proportion of missing for x and z 29.4% (50.1% overall)						
$Q_{y\|xz}(0.1)$						
β_0	0.20 (0.94)	0.40 (3.94)	**0.23** (2.35)	0.10 (4.43)	0.32 (10.91)	0.58 (11.73)
β_1	1.00 (0.92)	0.92 (3.27)	0.92 (2.44)	**0.96** (4.52)	0.96 (10.68)	0.35 (8.59)
β_2	1.00 (1.19)	0.93 (3.65)	**1.00** (2.33)	1.02 (4.35)	0.78 (25.47)	0.73 (84.84)
$Q_{y\|xz}(0.5)$						
β_0	0.79 (1.01)	1.28 (2.99)	**0.82** (1.49)	0.86 (1.61)	1.05 (1.92)	1.14 (1.58)
β_1	1.00 (0.99)	0.79 (2.50)	0.98 (1.48)	**0.96** (1.59)	0.91 (1.94)	0.52 (1.81)
β_2	1.00 (1.00)	0.85 (2.17)	**1.00** (1.25)	0.97 (1.37)	0.86 (1.35)	0.92 (1.47)
$\mathrm{E}(y\|x,z)$						
β_0	1.00 (0.93)	1.46 (2.53)	**0.99** (1.22)	0.98 (1.32)	1.13 (1.68)	1.37 (2.40)
β_1	1.00 (1.03)	0.79 (2.37)	1.03 (1.47)	1.04 (1.55)	**0.99** (2.06)	0.45 (2.64)
β_2	1.00 (1.09)	0.86 (2.17)	**1.00** (1.15)	0.98 (1.32)	0.87 (1.42)	0.86 (4.70)
Data generated under heteroscedastic model (4.6)						
Mean proportion of missing for x and z 27.5% (47.4% overall)						
$Q_{y\|xz}(0.1)$						
β_0	0.20 (1.00)	0.80 (6.71)	0.22 (3.99)	**0.18** (4.45)	0.47 (4.35)	0.68 (5.83)
β_1	1.00 (1.05)	0.69 (5.60)	0.89 (4.47)	**0.93** (4.63)	0.88 (4.78)	0.35 (4.31)
β_2	1.40 (0.89)	**1.27** (2.18)	1.22 (2.18)	1.11 (3.11)	0.79 (3.90)	0.84 (9.41)
$Q_{y\|xz}(0.5)$						
β_0	0.81 (0.96)	1.83 (3.20)	**1.00** (1.59)	1.02 (1.86)	1.41 (1.56)	1.29 (1.64)
β_1	0.99 (0.96)	0.58 (2.72)	0.89 (1.68)	**0.92** (1.84)	0.87 (1.90)	0.47 (1.55)
β_2	2.57 (1.11)	2.27 (2.13)	**2.36** (1.28)	2.26 (1.39)	1.94 (0.76)	2.22 (1.58)

Bold denotes the lowest relative bias among the missing data approaches.

Downloaded from smm.sagepub.com at University College London on April 29, 2015

from the full dataset (FD), i.e. $\overline{\boldsymbol{\beta}}_{\mathrm{FD}} = \frac{1}{200}\sum_{j=1}^{200}\widehat{\boldsymbol{\beta}}_{j,\mathrm{FD}}$, for each analysis model. We also report the variance ratio defined as the average of the model-based estimated variances divided by the Monte Carlo variance for the FD. The mean absolute difference between point estimates and $\bar{\beta}_{\mathrm{FD}}$, divided by $\bar{\beta}_{\mathrm{FD}}$, was used as an approximation of the absolute relative bias.

For data generated under model (4.5), QR imputation outperformed all other methods. Despite the substantial fraction of missing values in each variable and overall, the absolute percentage bias of QR did not exceed 27% across the three analysis models (median 8%) (results not shown). In most occasions, the LM PP method produced heavily biased estimates, with peaks of almost 200% (median 44%), performing even worse than the CC approach. QR imputation was also very competitive in terms of variability, having the lowest variance ratio among the MI methods.

For data generated using the heteroscedastic model (4.6), QR imputation performed well as compared to the other methods. The average estimate of the regression coefficients from the two QR-based MI methods was comparable. LM and LM PP produced estimates less biased than CC in some cases, though LM PP did not perform well as compared to LM since the log-transformation worsened the relative bias of LM estimates in half of cases. The MI methods did not, however, differ greatly by variance ratio.

A sensitivity analysis with $m = 10$ imputations was carried out, and the results were almost identical to those described above.

Finally, we compared QR with RQR imputation (results not shown). As expected, estimates were insensitive to this adjustment. There was a tendency of RQR imputation to be slightly less efficient than QR imputation, though no meaningful pattern could be identified.

5 MI analysis of birthweight determinants

5.1 Missing data patterns

In this section, we give a detailed account of missing values in the MCS dataset. We also analyse the distribution of the continuous variables. The subscript $j = 1, \ldots, 16$ will indicate the variables as numbered in Table 1 (e.g. $\mathbf{z}_1$ is birthweight and $\mathbf{r}_1$ is its missing data indicator).

A visual summary of the missing pattern is given in Figure 2 which shows that, overall, the pattern is nonmonotone. The most frequent missing data patterns involved parental body size measures. In particular, paternal weight and height were both missing for 9% of all children (15,070), followed by paternal weight only (5%) and maternal weight only (3%). Each of all the other patterns occurred in less than 1% of the sample.

In the MCS, birthweights were obtained from interviews. Maternal recall of birthweight has been shown to be reliable, although with few exceptions.[80] We found a significant association between the missing indicators $\mathbf{r}_1$ and $\mathbf{r}_2$ (χ^2 test p value < 0.001), with birthweight more likely to be missing when gestational age was missing, and vice versa. In particular, the odds of gestational age being missing in women with no education and in those of non-white ethnicity were, respectively, 4 (p value < 0.001) and 3 (p value < 0.001) as compared to the baseline. There was also a positive association with parity (p value 0.03). This is consistent with reported findings[80] that MCS mothers from ethnic wards, of non-white ethnic group, unemployed and with non-zero parity were more likely than others to provide an estimate of their child's birthweight that differed from the birth registration's recorded value by 100g or more. Similarly to what suggested by Tate et al.,[80] cultural differences and language barriers could also partly explain missingness for these variables. It is

Downloaded from smm.sagepub.com at University College London on April 29, 2015

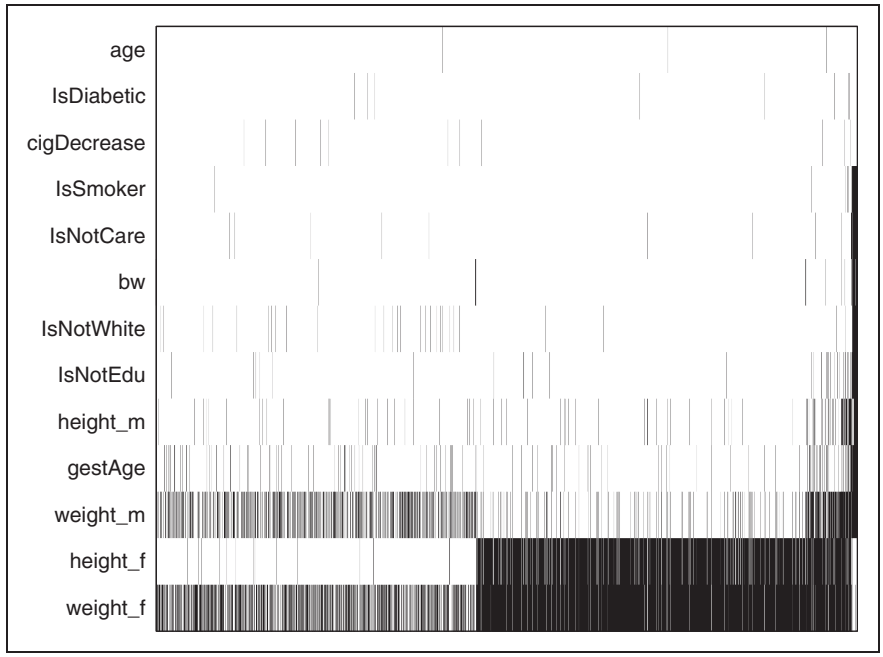

Figure 2. Missing data pattern in the MCS sample. Observations (*x*-axis) and variables (*y*-axis) are ordered by number of missing values. Black denotes missing.

reasonable to assume that conditionally on ethnicity, parity and the stratification variable, the missing data mechanism for birthweight and gestational age is ignorable.

Missing paternal weight and height accounted for 76% of the total number of incomplete cases (3066). There was a significant association between the missing indicators $\mathbf{r}_5$ and $\mathbf{r}_6$ (p value < 0.001). However, the missing data pattern showed some degree of monotonicity (Figure 2). It could be inferred that the respondents (i.e. the mothers) were able to recall their partners' height more easily that their weight, as the latter is subject to greater variation. Failing to give a measure of their partners' height implied having a poor ability to provide information on weight as well. Again, we assume a MAR mechanism for paternal weight and height as well as for the rest of the variables with missing items.

5.2 The imputation models

We now consider estimating the analysis model (3.4) after MI using the FCS algorithm described in Section 4 and we discuss some related modelling issues.

First, we establish which design variables to include in the imputation model, starting with cluster effects. We calculated complete case intraclass correlation coefficients (ICC) for all variables in Table 1, except reduction in number of cigarettes smoked (variable 9). We found that the ICC varied between 0.01 and 0.12 for variables 1–8 and between ~ 0 and 0.22 for variables 10–13, 15, and 16. As expected, the highest ICC was seen in mother's ethnicity (0.55) as this variable falls in one of the stratification domains. However, its fraction of missing values accounts for only 0.3% of the sample size. On this basis, we decided not to include clustering in the

Downloaded from smm.sagepub.com at University College London on April 29, 2015

Table 3. Analysis of shape for selected MCS continuous variables.

	bw	gestAge	weight_m	height_m	weight_f	height_f
Skewness	−0.52	−2.26	1.20	−0.11	0.58	−0.23
Kurtosis	4.95	12.30	5.40	5.02	3.83	4.36
λ	1.60	10.40	−0.80	1.60	0.00	2.40
Shapiro–Wilk	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001

imputation procedure as if would have been uninformative and the large number of clusters (~ 400) would have increased the computation time.

The MCS design stratification variable was found to be strongly associated with most of the variables. Sampling strata were therefore included in the conditional specification as categorical, so were sampling weights as continuous. The associated data matrix $\mathbf{Z}$ which was fed into `mice` had dimensions $15{,}070 \times 8$. For comparison purposes, a larger imputation model was also considered including stratum-specific effects (except for gestational age, and prenatal care and diabetes status indicators which were not found to be significantly associated with the stratification variable). As a result, the number of columns of $\mathbf{Z}$ increased from 18 to 121.

Next, we make some considerations about the continuous variables (we leave out age at delivery since the number of missing values in this variable is very small). The histogram of birthweight by gestational age and the histogram of gestational age were provided in Figure 1. Table 3 shows skewness, kurtosis, estimated (marginal) BCT parameter λ and p value of the Shapiro–Wilk test for normality after BCT. In all cases, skewness and excess kurtosis are apparent while the BCT fails to achieve normality even when these are relatively moderate, as in the case of paternal weight. Normal Q-Q plots (not shown) confirmed this result. This suggests that using mean regression as imputation model would be inappropriate.

Two pairs of variables are of particular interest in relation to MI: birthweight and gestational age, and paternal weight and height. The former pair includes the response variable and its strongest predictor (Spearman's rank correlation 0.42), while the latter includes two strongly correlated (0.47) variables with the highest proportion of missing values among all other variables. These are plotted in Figure 3 using logit scaling on the y-axis. There is an indication that a linear relationship is reasonable. Note also the outlying observations in both plots (this point will be discussed further in Section 5.4).

We tested location shift and location-scale shift hypotheses by using Khmaladze tests[76] for the linear models $\text{logit(bw)} \sim \text{gestAge}$ (sample size n=14,907) and $\text{logit}\,(\text{weight}_f) \sim \text{height}_f$ (n=12,731) on the range $p \in [0.05, 0.95]$. None of the tests rejected the null hypothesis at the 5% level except for the location shift hypothesis in the birthweight model. We repeated the tests for $\text{bw} \sim \text{gestAge}$ and $\text{weight}_f \sim \text{height}_f$. The location-scale shift hypothesis for birthweight and the location shift hypothesis for paternal weight were rejected at the 1% and 5% level, respectively. In summary, these results suggest the following: on the untransformed scale, the relationship between birthweight and gestational age is more complex than a location-scale shift model; in contrast, the relationship between paternal weight and height can be defined by a heteroscedastic model. On the logit scale, the former relationship is approximately heteroscedastic whereas the latter relationship is approximately constant over quantiles. In other words, the logit transformation has simplified the model specification.

Finally, we assigned an imputation model to each variable with missing data (Table 1). More precisely, continuous variables (1–7, 9) were imputed using the QR-based approach described in

Downloaded from smm.sagepub.com at University College London on April 29, 2015

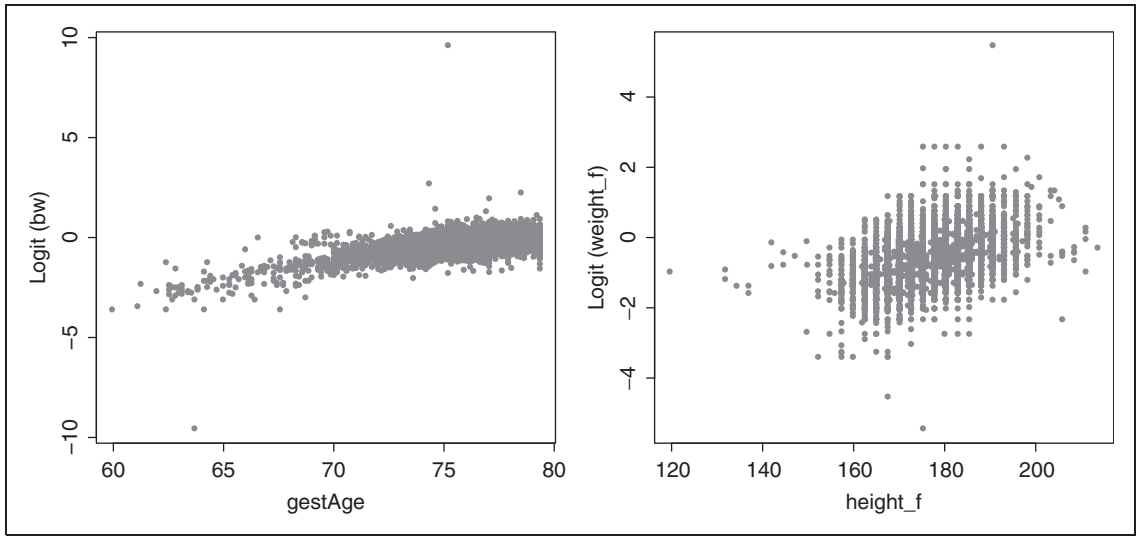

Figure 3. Left: birthweight versus gestational age. Right: paternal weight versus height. The *y* axis is on the logit scale.

Section 4.2. In particular, we set ω equal to 0.001 in step (i). Since these variables are constrained to vary within boundaries, pre-processing was carried out by applying a logit transform with internal bounds which, as seen above, gives an additional benefit for modelling. Binary variables (10, 11, 14–16) were assigned logistic regressions.

We set $m = 5$ imputed datasets and a maximum of five iterations for each imputation. These are the default values in the `R` function `mi`.[48] The QR model (3.4) was then fitted to each imputed dataset using the methods illustrated in Section 4.1. As in the complete case analysis (Section 3.2), a bootstrap sample size $n_B = 100$ was used to estimate the variance of the regression quantiles.

5.3 Results

First and foremost, the results of the analysis did not differ sensibly when using the 'reduced' (i.e. with strata effects only) or the 'full' (i.e. with stratum-specific effects) imputation models. Table 4 shows point estimates and standard errors for selected quantiles (results for all quantiles are available upon request) using the reduced model. In the following, we elaborate on parental and smoking effects which offer elements of novelty but we gloss over the other effects for the sake of brevity.

The complete sets of estimated regression quantiles for differential and mean parental weight and height are plotted in Figure 4. The coefficients of differential weight were positive and significant (p value < 0.05) for quantiles $p \geq 0.1$, with larger magnitude at higher quantiles. For $p = 0.01$ and $p = 0.05$, $\beta_2(p)$ was not significantly different from zero at the 5% level. Similarly, the quantile effects of parental mean weight were positive for all p and the magnitude of these effects was larger at higher birthweight quantiles.

The quantile effects of differential parental height had large confidence intervals. However, the least squares estimate was marginally significant (p value < 0.05). Mean parental height had a

Downloaded from smm.sagepub.com at University College London on April 29, 2015

Table 4. Point estimates of selected regression quantiles and, in brackets, standard errors for the MCS birthweight analysis.

	Complete case analysis			Multiple imputation analysis		
	Quantile			Quantile		
	0.05	0.5	0.95	0.05	0.5	0.95
Intercept	**2340** (28)	**3006** (15)	**3778** (38)	**2354** (24)	**3010** (12)	**3784** (32)
gestAge	**174** (6)	**168** (4)	**128** (9)	**170** (5)	**169** (3)	**129** (8)
$weight_{hd}$	20 (18)	**82** (11)	**119** (16)	19 (17)	**80** (9)	**114** (17)
$weight_{sum}$	**26** (7)	**45** (4)	**53** (10)	**28** (7)	**43** (4)	**53** (9)
$height_{hd}$	**37** (16)	12 (11)	22 (22)	**30** (16)	11 (10)	20 (21)
$height_{sum}$	**44** (8)	**46** (4)	**68** (11)	**45** (8)	**45** (4)	**67** (10)
age	−2 (2)	0 (1)	1 (3)	−2 (2)	0 (1)	1 (2)
parity	**58** (11)	**69** (7)	**88** (18)	**50** (10)	**62** (6)	**81** (15)
cigDecrease	**10** (2)	**5** (2)	1 (3)	**10** (2)	**5** (2)	2 (3)
IsSmoker	−56 (31)	−**58** (22)	−23 (41)	−51 (31)	−**51** (20)	−20 (39)
IsNotEdu	−**95** (34)	−22 (21)	−30 (39)	−**84** (29)	−15 (19)	−39 (32)
IsGirl	−**116** (21)	−**139** (10)	−**108** (26)	−**112** (20)	−**143** (9)	−**105** (24)
IsNotMarried	−46 (29)	−10 (16)	23 (28)	−37 (22)	−12 (14)	16 (28)
IsNotWhite	−**149** (37)	−**137** (24)	−32 (33)	−**203** (35)	−**162** (19)	−42 (31)
IsNotCare	−33 (63)	−56 (31)	147 (161)	−39 (51)	−30 (27)	61 (133)
IsDiabetic	−43 (77)	−**80** (31)	182 (142)	−83 (115)	−**88** (29)	164 (149)

Complete case analysis (Section 3) and multiple imputation analysis (Section 5). Bold denotes significant at the 5% level.

positive effect on all birthweight quantiles and such effect was approximately uniform, except at the extremes where the coefficients were larger.

These results are consistent with those reported by Griffiths et al.[63] In their study based on MCS data, a regression model for mean birthweight was fitted after incomplete parental data were discarded under MCAR assumptions. In contrast, we allowed for variables such as maternal education and ethnicity to account for a MAR data mechanism. Our analysis shows that the parental effects are not constant across the birthweight distribution. This provides important insights for the understanding of the parental conflict theory[81] according to which the father's aim is to maximize the growth of his offspring, while mothers maximize their chance of survival by constraining foetal growth. As reported by Griffiths et al., the influence of the mother wins over the father's contribution to mean birthweight. However, it seems that this influence is somehow relaxed when the survival of the child is at high risk (low birthweight quantiles) and it is strongest at the opposite end of the range (high birthweight quantiles) when her own survival is at higher risk. In other words, there may be an effect gradient in the anthropometric parental factors regulating birthweight that runs from top to bottom of the birthweight spectrum. To the best of our knowledge, these results have not been reported before and deserve further investigation.

Smoking is known to reduce birthweights[82] and this effect is approximately uniform across birthweight quantiles, as confirmed by our results (Figure 4). Although on average a decrease in smoking had little effect on birthweight (approximately 5 g per cigarette per day), at lower quantiles the coefficient was twice as much (approximately 10 g per cigarette per day). Such effect was attenuated at higher quantiles. Despite the different design and purpose of our analysis as

Downloaded from smm.sagepub.com at University College London on April 29, 2015

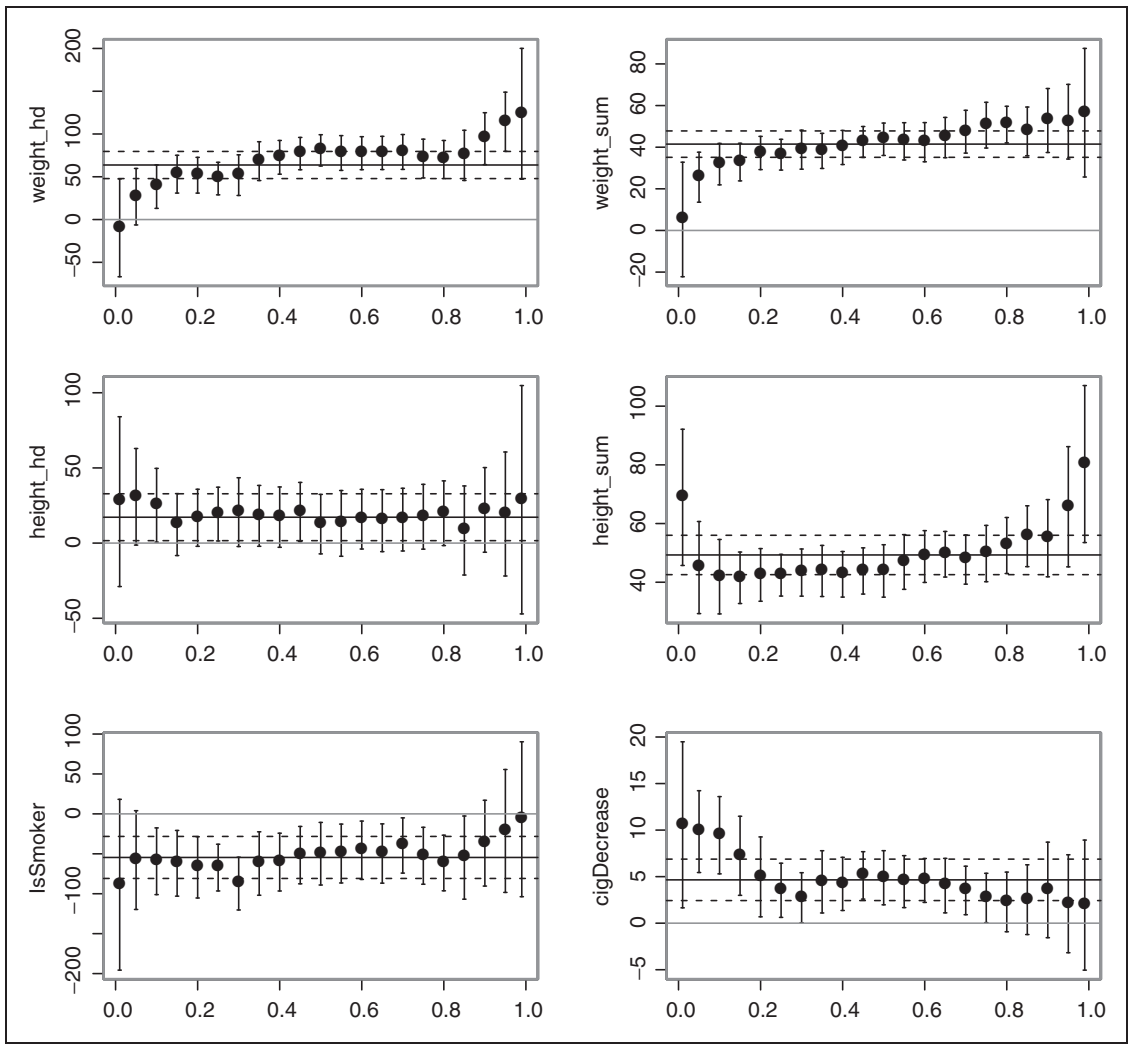

Figure 4. Estimated regression quantiles with 95% confidence intervals (error bars) for parental weight and height and smoking effects in the MCS birthweight analysis. Horizontal black lines denote mean estimates (solid) and their 95% confidence intervals (dashed).

compared to England et al.,[82] our results point out the different impact on birthweight that a reduction in smoking has at different birthweight quantiles.

Similar conclusions were reached when using the results from the complete case analysis (Table 4). In general, the latter provided estimated coefficients that were consistent with those obtained after MI, except for the effect associated with ethnicity. The magnitude of this coefficient was, on average, about 26% larger after imputation. This is not surprising given that ethnicity is a strong predictor of MCS unit nonresponse[51] and missing birthweight. It also falls in one of the stratification domains. In contrast, the standard errors of the coefficients after imputation tended to be lower than those obtained in the complete case analysis by about 10%, on average. This

Downloaded from smm.sagepub.com at University College London on April 29, 2015

can be explained by a lower within-imputation variance and by a relatively small between-imputation variance, the latter accounting on average for about 5% of the total imputation variance.

Sensitivity analyses were carried out, including: an analysis under MCAR and SRS assumptions (i.e. ignoring missingness and sampling design). QR standard errors were calculated using the asymptotic variance estimator for *nid* errors; an analysis where only sampling weights but not the strata were included in the imputation model; an analysis where strata and unadjusted weights were used. In summary, the design effect for mean and median birthweight was, respectively, 1.8 and 1.9, which indicates that design-based estimates are around 90% less efficient than an SRS for the selected outcome and this justifies the use of survey variables in the analysis model. Not including strata in the imputation model had little effect on the results, as did the use of unadjusted weights. Obviously, this has implications for the analysis of birthweight only as other MCS variables might respond differently to such modelling adjustments. Design effects for other MCS variables that are highly correlated with the sampling domain (e.g. ethnicity, income) can be as high as 20.[49]

In total, the imputation (reduced model) and analysis steps took about 43 and 24minutes, respectively, on a 64-bit operating system machine with 16 Gb of RAM and quad-core processor at 2.93 GHz. In both steps, we used a Frisch–Newton algorithm as most appropriate given the size of the problem. The Barrodale and Roberts algorithm[83] took about twice as long. Reductions of computation time could ideally be obtained by decreasing the number of breakpoints u_i in the imputation step and by employing parallel computing in the analysis step. Changing the variables visiting sequence[48] from that in Table 1 to a sequence in increasing order of missing data did not provide an appreciable computational advantage.

5.4 Diagnostics

MI methods are typically accompanied by diagnostic measures to assess the imputed values and the impact of the missing data on the parameters being estimated. In Figure 5, box-centile plots of observed and imputed values for four selected variables are reported. The distribution of each imputed sample follows the observed distribution reasonably well. The coverage of the observed range by the imputed values is nearly 100% and, as a consequence of the logit transformation, in no case the range is exceeded. A similar imputation without pre-processing (results not shown) produced values for parental weight and height outside the observed range.

Although the distributions of birthweight and gestational age stretch to very low and high points, more extreme values were implicitly excluded by trimming the distribution of z when defining upper and lower limits for u, that is $u \sim \text{Unif}(0.001, 0.999)$. In general, conditional quantiles are robust to outliers in the outcome but may be sensitive to outliers in the design matrix. As mentioned in Section 2, several outliers in parental weight were replaced by missing values. Outliers were defined as observations exceeding the quantiles 0.0005 and 0.9995 of these variables. In total, we detected 16 and 14 values (as low as 11kg and as high as 288kg) for, respectively, maternal and paternal weight. Although this method is naïve and a more sensible outlier detection approach could have been based on the bivariate distribution of weight and height, it is of little practical importance in our analysis since our aim is to assess the value of the outliers as compared to their imputed values. Figure 6 shows the bivariate plot of height and weight for both parents. In addition to the formal testing conducted earlier, the strong heteroscedasticity that characterizes the relationship between these two variables can now be seen. This distributional feature is taken into account by the QR-based imputation model as shown by the range of imputed values at different heights. For example, very low paternal weights defined as outliers had imputed values that were consistent with the

Downloaded from smm.sagepub.com at University College London on April 29, 2015

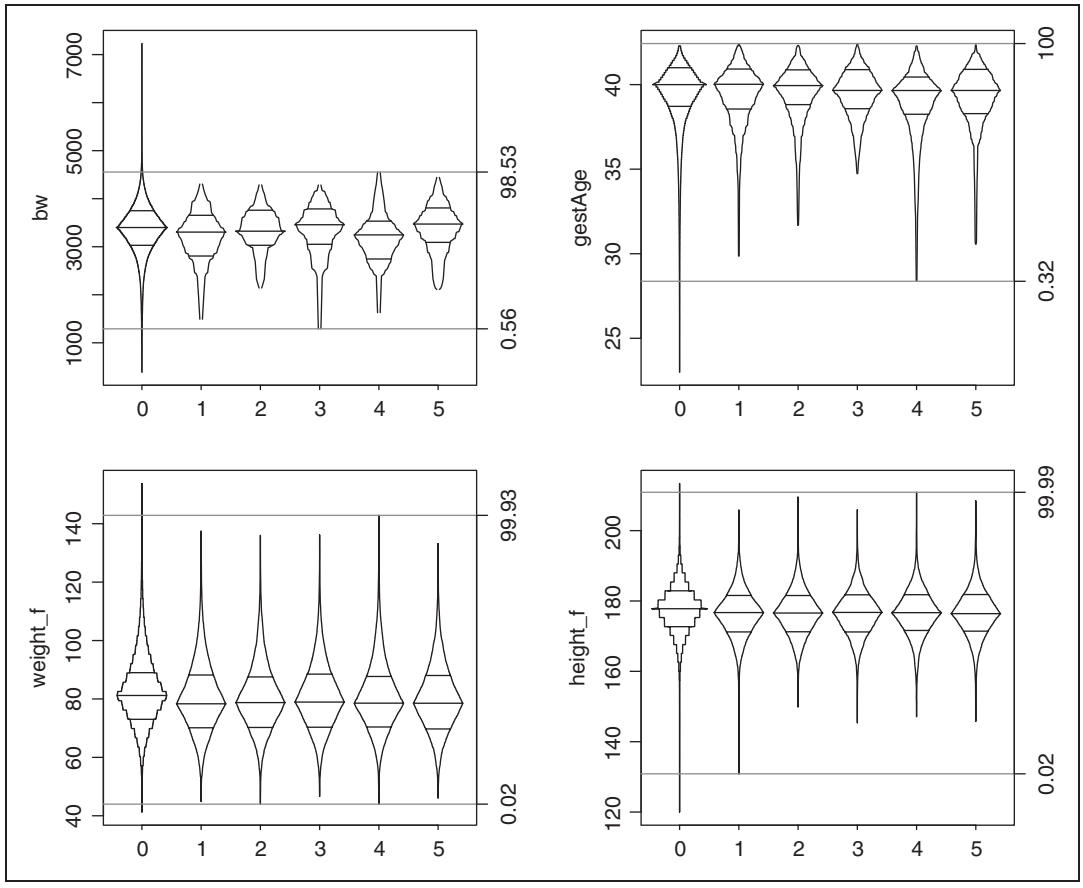

Figure 5. Box-centile plots of observed (0) and imputed (1–5) values for birthweight, gestational age, paternal weight and height. Grey lines define the centile range of coverage of the observed distribution by the minimum and maximum of all imputed values.

given height, and the spread of such imputations was consistent with the spread of the observed distribution of weight at that given height. A mean regression would not be able to preserve such distributional relationships between variables.

For each regression quantile, the rate of missing information (RMI) can be calculated as $\gamma = \frac{r+2/(df+3)}{r+1}$, where $r = \frac{\left(1+m^{-1}\right)}{\hat{\beta}(p)}$, df is the number of degrees of freedom and b is the between-imputation variance. Overall, the maximum RMI was approximately 20% and this was observed for weight_{hd}. The latter variable was obviously affected by a substantial increase in variance due to a high proportion of missing values of paternal weight. The RMI for gestational age did not exceed 8% in all quantiles but it was higher for the first regression centile (7%) as compared to the 99th (2%).

Finally, we checked the mean and variance of the imputed values at each iteration for all but the discrete variables and no systematic pattern was observed.

Downloaded from smm.sagepub.com at University College London on April 29, 2015

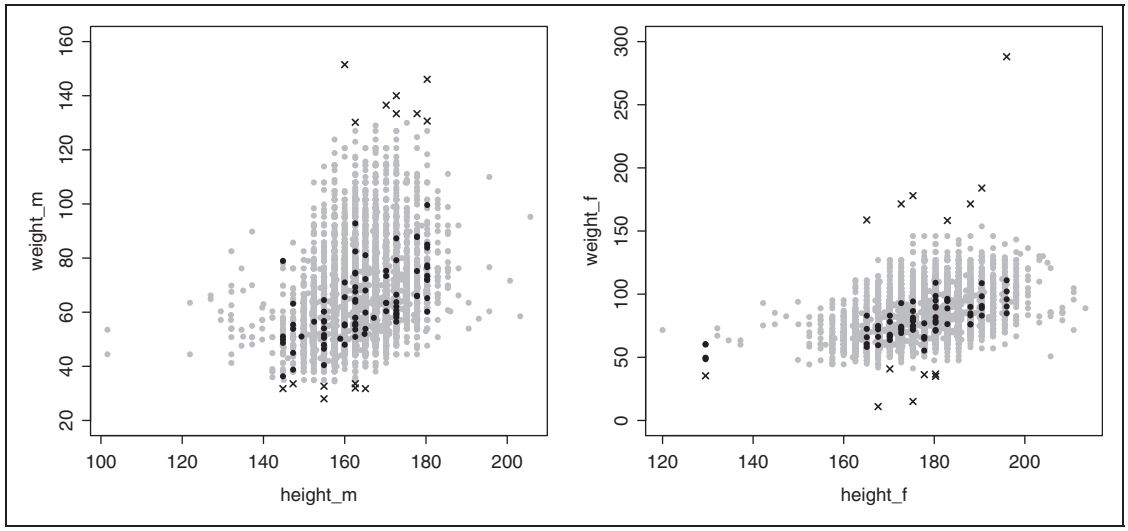

Figure 6. Plot of observed weight versus height (grey dots) for MCS mothers and fathers. Crosses denote weight outliers while black dots the corresponding imputations.

6 Discussion

Complex survey estimation and MI are two important statistical areas in the analysis of large population-based epidemiological and socioeconomic studies. In this paper, we considered the application of quantile modelling to the analysis of birthweight determinants in a cohort of UK children, providing an extensive and detailed investigation of methods that can be easily implemented using available software and little additional programming effort. The problem of missing data was addressed by developing an MI approach under MAR assumptions, based on an FCS algorithm in which imputation of continuous variables takes into account the distributional features of the observed data. We discussed model selection issues associated with MI in general and QR-based imputation in particular and explored some practical aspects in the MCS analysis.

We proposed some strategies on how to use transformations for handling outliers and bounded variables, and, in the MCS birthweight study, we showed that transformations may also be beneficial to the specification of the form of quantile functions. However, transformations that are typically considered as a remedy for data that do not conform to some well-known distribution are often unsatisfactory, may require estimating additional parameters, and pose delicate problems for interpreting the results. Features such as skewness or kurtosis, which is traditionally associated with peakedness and tail weight of a distribution, are not to be seen as a nuisance but rather as informative on the unknown data-generating process. Moreover, their conditional relationship with the covariates under study is often too complex to be *normalized* by a simple marginal transformation. In contrast, quantiles are able to discriminate among distributional shapes in a natural way, as also demonstrated by an increasing number of proposals of quantile-based measures of asymmetry and kurtosis.[84–86] QR offers a flexible and powerful means to preserve the information in the data. In addition, it handles desired data transformations in a simple yet mathematically rigorous fashion.

But the flexibility associated with QR may come at a cost: the difficulty in choosing among a number of models for different quantiles. As stressed by Van Buuren[38] and Schafer,[75] avoiding

Downloaded from smm.sagepub.com at University College London on April 29, 2015

specifying a joint distribution for the data and the missing data mechanism does not clearly remove other model selection issues. Moreover, accounting for the sampling design variables in the MI procedure can be laborious from a computational and a modelling standpoint. In addition, analytical demonstration that the imputation is 'proper'[65] may be difficult except in trivial cases.[64, p. 145]

It is worth stressing that analysis and imputation models need not be the same. It is, of course, a logical consequence that a location-shift hypothesis for the imputation model is incompatible with the motivation that has led, in the first place, to a conditional quantile approach for the model of interest. Separating imputation and analysis models can be advantageous.[75] In non-normal conditions, QR-based imputation is an excellent alternative to mean imputation when the target of the analysis is the conditional mean,[39] even when the data undergo some form of pre-processing as shown in our simulation study. It needs to be investigated whether other inferential targets (e.g. scale and shape indices) can benefit from this approach.

Funding

The Centre for Paediatric Epidemiology and Biostatistics benefits from funding support from the Medical Research Council in its capacity as the MRC Centre of Epidemiology for Child Health (G0400546). The UCL Institute of Child Health receives a proportion of funding from the Department of Health's NIHR Biomedical Research Centres funding scheme.

Acknowledgements

We thank two anonymous reviewers, whose constructive comments led to an improved manuscript; Jianqiang Wang and Jean Opsomer who kindly provided the R code to replicate their simulation study; and Matteo Bottai for providing a copy of the paper entitled 'Multiple imputation based on conditional quantile estimation'.

References

1. Smith K and Joshi H. The Millennium Cohort Study. *Popul Trends* 2002; **107**: 30–34.
2. Koenker R and Bassett G. Regression quantiles. *Econometrica* 1978; **46**: 33–50.
3. Yu KM, Lu ZD and Stander J. Quantile regression: applications and current research areas. *J R Stat Soc Ser D* 2003; **52**: 331–350.
4. Machado JAF and Santos Silva JMC. Quantiles for counts. *J Am Stat Assoc* 2005; **100**: 1226–1237.
5. Portnoy S. Censored regression quantiles. *J Am Stat Assoc* 2003; **98**: 1001–1012.
6. Bottai M and Zhang J. Laplace regression with censored data. *Biom J* 2010; **52**: 487–503.
7. Bottai M, Cai B and McKeown RE. Logistic quantile regression for bounded outcomes. *Stat Med* 2009; **9**: 309–317.
8. Yu KM and Jones MC. Local linear quantile regression. *J Am Stat Assoc* 1998; **93**: 228–237.
9. Koenker R, Ng P and Portnoy S. Quantile smoothing splines. *Biometrika* 1994; **81**: 673–680.
10. Thompson P, Cai YZ, Moyeed R, et al. Bayesian nonparametric quantile regression using splines. *Comput Stat Data Anal* 2010; **54**: 1138–1150.
11. Heagerty PJ and Pepe MS. Semiparametric estimation of regression quantiles with application to standardizing weight for height and age in US children. *J R Stat Soc Ser C* 1999; **48**: 533–551.
12. Lee D and Neocleous T. Bayesian quantile regression for count data with application to environmental epidemiology. *J R Stat Soc Ser C* 2010; **59**: 905–920.
13. Koenker R and Geling O. Reappraising medfly longevity. *J Am Stat Assoc* 2001; **96**: 458–468.
14. Koenker R. *Quantile regression*. New York: Cambridge University Press, 2005.
15. Abrevaya J. The effects of demographics and maternal behavior on the distribution of birth outcomes. *Empir Econ* 2001; **26**: 247–257.
16. Abrevaya J and Dahl CM. The effects of birth inputs on birthweight: evidence from quantile estimation on panel data. *J Bus Econ Stat* 2008; **26**: 379–397.
17. Burgette LF and Reiter JP. Modeling adverse birth outcomes via confirmatory factor quantile regression. *Biometrics* 2012; **68**: 92–100.
18. Burgette LF, Reiter JP and Miranda ML. Exploratory quantile regression with many covariates: an application to adverse birth outcomes. *Epidemiology* 2011; **22**: 859–866.
19. Mudd LM, Pivarnik J, Holzman CB, et al. Leisure-time physical activity in pregnancy and the birth weight distribution: where is the effect? *J Phys Act Health* 2012; **9**: 1168–1177.
20. Börnhorst C, Hense S, Ahrens W, et al. From sleep duration to childhood obesity – what are the pathways? *Eur J Pediatr* 2012; **171**: 1029–1038.

Downloaded from smm.sagepub.com at University College London on April 29, 2015

21. Wei Y and He XM. Conditional growth charts (with Discussion). *Ann Stat* 2006; **34**: 2069–2097.
22. Wei Y, Pere A, Koenker R, et al. Quantile regression methods for reference growth charts. *Stat Med* 2006; **25**: 1369–1382.
23. Chen K and Müller HG. Conditional quantile analysis when covariates are functions, with application to growth data. *J R Stat Soc Ser B* 2012; **74**: 67–89.
24. Beyerlein A, Toschke AM and von Kries R. Breastfeeding and childhood obesity: shift of the entire BMI distribution or only the upper parts? *Obesity* 2008; **16**: 2730–2733.
25. Fenske N, Kneib T and Hothorn T. Identifying risk factors for severe childhood malnutrition by boosting additive quantile regression. *J Am Stat Assoc* 2011; **106**: 494–510.
26. Birch JM, Geraci M, Alston RD, et al. Low birthweight and aetiology of childhood liver tumours in North West England. *Pediatr Blood Cancer* 2010; **55**: 932.
27. Bayer O, Neuhauser H and Von Kries R. Sleep duration and blood pressure in children: a cross-sectional study. *J Hypertension* 2009; **27**: 1789–1793.
28. Heeringa S, West BT and Berglund PA. *Applied survey data analysis*. Boca Raton, FL: Chapman & Hall/CRC, 2010.
29. Woodruff RS. Confidence intervals for medians and other position measures. *J Am Stat Assoc* 1952; **47**: 635–646.
30. Sedransk J and Meyer J. Confidence intervals for the quantiles of a finite population: simple random and stratified simple random sampling. *J R Stat Soc Ser B* 1978; **40**: 239–252.
31. Shao J. L-statistics in complex survey problems. *Ann Stat* 1994; **22**: 946–967.
32. Chatterjee A. Asymptotic properties of sample quantiles from a finite population. *Ann Inst Stat Math* 2011; **63**: 157–179.
33. Wang JCQ and Opsomer JD. On asymptotic normality and variance estimation for nondifferentiable survey estimators. *Biometrika* 2011; **98**: 91–106.
34. Geraci M and Salvati N. The geographical distribution of the consumption expenditure in Ecuador: estimation and mapping of the regression quantiles. *Stat Appl – Ital J Appl Stat* 2007; **19**: 167–183.
35. Li Y, Graubard BI and Korn EL. Application of nonparametric quantile regression to body mass index percentile curves from survey data. *Stat Med* 2010; **29**: 558–572.
36. He Y, Zaslavsky A, Landrum M, et al. Multiple imputation in a large-scale complex survey: a practical guide. *Stat Meth Med Res* 2010; **19**: 653–670.
37. Little RJA and Rubin DB. *Statistical analysis with missing data*, 2nd ed. Hoboken, NJ: Wiley, 2002.
38. Van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. *Stat Meth Med Res* 2007; **16**: 219–242.
39. Bottai M and Zhen H. Multiple imputation based on conditional quantile estimation. *Epidemiol Biostat Pub Health* 2013; **19**(1). DOI: 10.2427/8758.
40. Cheng PE and Chu CK. Kernel estimation of distribution functions and quantiles with missing data. *Stat Sin* 1996; **6**: 63–78.
41. Lipsitz SR, Fitzmaurice GM, Molenberghs G, et al. Quantile regression methods for longitudinal data with drop-outs: application to CD4 cell counts of patients infected with the human immunodeficiency virus. *J R Stat Soc Ser C* 1997; **46**: 463–476.
42. Yuan Y and Yin G. Bayesian quantile regression for longitudinal studies with nonignorable missing data. *Biometrics* 2010; **66**: 105–114.
43. Wei Y, Ma Y and Carroll RJ. Multiple imputation in quantile regression. *Biometrika* 2012; **99**: 423–438.
44. R Development Core Team. *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, 2011.
45. Lumley T. Analysis of complex survey samples. *J Stat Software* 2004; **9**: 1–19.
46. Lumley T. *Survey: analysis of complex survey samples*, R package version 3.24-1. 2011.
47. Koenker R. *quantreg: quantile regression*, R package version 4.65. 2011.
48. Van Buuren S and Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations in R. *J Stat Software* 2011; **45**: 1–67.
49. Plewis I. *The Millennium Cohort Study: technical report on sampling*, 4th ed. London: Centre for Longitudinal Studies, 2007.
50. Cullis A. *Infant mortality in the Millennium Cohort Study (MCS) sample areas*. London: Centre for Longitudinal Studies, 2004.
51. Plewis I. Non-response in a birth cohort study: the case of the Millennium Cohort Study. *Int J Soc Res Methodol* 2007; **10**: 325–334.
52. Koenker R. Quantile regression for longitudinal data. *J Multivar Data Anal* 2004; **91**: 74–89.
53. Geraci M and Bottai M. Quantile regression for longitudinal data using the asymmetric Laplace distribution. *Biostatistics* 2007; **8**: 140–154.
54. Geraci M and Bottai M. Linear quantile mixed models. *Stat Comput* 2013. DOI: 10.1007/s11222-013-9381-9.
55. Geraci M. *lqmm: linear quantile mixed models*, R package version 1.02. 2012.
56. Portnoy S and Koenker R. The Gaussian hare and the Laplacian tortoise: computability of squared-error versus absolute-error estimators. *Stat Sci* 1997; **12**: 279–300.
57. Canty AJ and Davison AC. Resampling-based variance estimation for labour force surveys. *J R Stat Soc Ser D* 1999; **48**: 379–391.
58. Presnell B and Booth JG. *Resampling methods for sample survey*. Florida: Department of Statistics, University of Florida, 1994.
59. Davison AC and Hinkley DV. *Bootstrap methods and their application*. Cambridge, UK: Cambridge University Press, 1997.
60. Koenker R and Hallock KF. Quantile regression. *J Econ Perspect* 2001; **15**: 143–156.
61. Chernozhukov V and Fernández-Val I. Inference for extremal conditional quantile models, with an application to market and birthweight risks. *Rev Econ Stud* 2011; **78**: 559–589.
62. Schwartz SL, Gelfand AE and Miranda ML. Joint Bayesian analysis of birthweight and censored gestational age using finite mixture models. *Stat Med* 2010; **29**: 1710–1723.
63. Griffiths LJ, Dezateux C, Cole TJ, et al. Differential parental weight and height contributions to offspring birthweight and weight gain in infancy. *Int J Epidemiol* 2007; **36**: 104–107.
64. Schafer JL. *Analysis of incomplete multivariate data*. London: Chapman & Hall, 1997.
65. Rubin DB. *Multiple imputation for nonresponse in surveys*. New York: Wiley, 1987.
66. Reiter JP, Raghunathan TE and Kinney SK. The importance of modeling the sampling design in multiple imputation for missing data. *Surv Methodol* 2006; **32**: 143–150.
67. White IR, Royston P and Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Stat Med* 2011; **30**: 377–399.

Downloaded from smm.sagepub.com at University College London on April 29, 2015

68. Little RJA. Missing-data adjustments in large surveys. *J Bus Econ Stat* 1988; **6**: 287–296.
69. Schenker N and Taylor JMG. Partially parametric techniques for multiple imputation. *Comput Stat Data Anal* 1996; **22**: 425–446.
70. He Y and Raghunathan TE. Tukey's *gh* distribution for multiple imputation. *Am Stat* 2006; **60**: 251–256.
71. Raghunathan TE, Lepkowski JM, Hoewyk JV, et al. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Surv Methodol* 2001; **27**: 85–95.
72. Portnoy S. Asymptotic behavior of the number of regression quantile breakpoints. *SIAM J Sci Stat Comput* 1991; **12**: 867–883.
73. Su YS, Gelman A, Hill J, et al. Multiple imputation with diagnostics (mi) in R: opening windows into the black box. *J Stat Software* 2011; **45**: 1–31.
74. Kordas G. Smoothed binary regression quantiles. *J Appl Econometrics* 2006; **21**: 387–407.
75. Schafer JL. Multiple imputation in multivariate problems when the imputation and analysis models differ. *Stat Neerland* 2003; **57**: 19–35.
76. Koenker R and Xiao ZJ. Inference on the quantile regression process. *Econometrica* 2002; **70**: 1583–1612.
77. He X. Quantile curves without crossing. *Am Stat* 1997; **51**: 186–192.
78. Zhao QS. Restricted regression quantiles. *J Multivar Anal* 2000; **72**: 78–99.
79. Chernozhukov V, Fernandez-Val I and Galichon A. Quantile and probability curves without crossing. *Econometrica* 2010; **78**: 1093–1125.
80. Tate AR, Dezateux C, Cole TJ, et al. Factors affecting a mother's recall of her baby's birth weight. *Int J Epidemiol* 2005; **34**: 688–695.
81. Moore T and Haig D. Genomic imprinting in mammalian development: a parental tug-of-war. *Trends Genet* 1991; **7**: 45–49.
82. England LJ, Kendrick JS, Wilson HG, et al. Effects of smoking reduction during pregnancy on the birth weight of term infants. *Am J Epidemiol* 2001; **154**: 694–701.
83. Koenker RW and d'Orey V. Algorithm AS 229: computing regression quantiles. *J R Stat Soc Ser C* 1987; **36**: 383–393.
84. Groeneveld RA and Meeden G. Measuring skewness and kurtosis. *J R Stat Soc Ser D* 1984; **33**: 391–399.
85. Groeneveld RA. A class of quantile measures for kurtosis. *Am Stat* 1998; **52**: 325–329.
86. Wang J and Serfling R. Nonparametric multivariate kurtosis and tailweight measures. *J Nonparametric Stat* 2005; **17**: 441–456.
87. Gutenbrunner C and Jurečková J. Regression rank scores and regression quantiles. *Ann Stat* 1992; **20**: 305–330.

Appendix A – Standard errors for regression quantiles

For regression quantiles obtained from minimizing loss functions of the type (3.2), standard error estimation can be performed using, for example, sandwich-type estimators or bootstrap techniques.[14] Inference can also be carried out based on regression rank scores.[87]

For complex survey samples, we consider the method described by Canty and Davison.[57–59] Briefly, this bootstrap approach is designed to mimic the effect of sampling without replacement and it applies a calibration of the sample to ensure that the post-strata marginal totals agree with known population margins. This method is also implemented in the `R` package `survey`.[45,46] See also Chatterjee[32] for a bootstrap approach based on sampling with replacement where the bootstrap estimates are rescaled to guarantee consistency of the sample variance.

Let $\widehat{\boldsymbol{\beta}}(p)$ be the regression quantile obtained from minimizing (3.3). Also let $\{\widehat{\boldsymbol{\beta}}_b(p)\colon b = 1, \ldots, n_B\}$ be a set of n_B bootstrap estimates obtained from the sample data using Canty and Davison's method.[57] Let us define the matrix $\mathbf{B}$ with row vectors $(\widehat{\boldsymbol{\beta}}_b(p) - \overline{\boldsymbol{\beta}}(p))'$, where $\overline{\boldsymbol{\beta}}(p) = \frac{1}{n_B}\sum_{b=1}^{n_B}\widehat{\boldsymbol{\beta}}_b(p)$ is the $q \times 1$ vector of element-wise averages of bootstrap estimates. An estimate of the variance-covariance matrix of the estimator of the regression coefficients is given by

$$\widehat{\mathbf{V}} = \mathbf{B}'\mathbf{B} \cdot \sigma \tag{A.1}$$

where $\sigma = \frac{\bar{m}}{(\bar{m}-1)(n_B-1)}$ is the overall scaling factor and $\bar{m} = \frac{H}{\sum_h m_h^{-1}}$.[46] The boostrap estimated standard error is given by $\widehat{\text{se}(\boldsymbol{\beta}(p))} = (\hat{v}_1^{1/2}, \ldots, \hat{v}_q^{1/2})'$, where $\hat{v}_j$ is the jth diagonal element of $\widehat{\mathbf{V}}$. Alternatively, one can use the mean squared error to take the bias $\overline{\boldsymbol{\beta}}(p) - \widehat{\boldsymbol{\beta}}(p)$ into account.

The estimation of the variance of regression quantiles is an area to be investigated further. Although bootstrap is a practical approach which is already implemented in statistical software, other promising design-consistent estimation methods have been proposed and these might offer substantial computational advantages. Wang and Opsomer[33] recently developed a theory for nondifferentiable survey estimators. They studied analytical and replication-based design-consistent variance estimators, which rely on the choice of a bandwidth parameter for the kernel

Downloaded from smm.sagepub.com at University College London on April 29, 2015

regression used to estimate the smooth limit of the nondifferentiable functions. In a simulation experiment on sample quantiles, their proposed variance estimators were shown to perform well compared to a naïve jackknife. However, the relative bias depended on the kernel bandwidth and further numerical and theoretical studies on this front are needed.[33] Here, we report the results of a comparison study of Canty and Davison's survey bootstrap estimator with Wang and Opsomer's proposed estimators. We used a simulation setting as in Wang and Opsomer and estimated nine population quantiles $p \in \{0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95, 0.99\}$, including four additional tail quantiles as they are of particular interest in, for example, birthweight and body mass index studies. We calculated the sample quantiles by inverting the population cumulative distribution function[33] and, in addition, by minimizing a weighted loss function as in (3.3) with an intercept only. A Gaussian kernel with bandwidth value set to 0.2 and a desired sample size of 400 were used. As expected, the results showed that the two weighted survey estimators produced the same quantile estimates. The analytic and jackknife variance estimators performed well as compared to bootstrap ($n_B = 100$): the percentage relative bias of the three estimators ranged, respectively, from -31.3 to 34.7, from -27.2 to 35.7 and from -26.8 to 92.4. The bias, as expected, was larger for quantiles farther away from the median.

Appendix B – R code

B.1. Regression quantiles

An example on how to fit a median regression model using survey data is provided below. The reader is referred to the documentation available in `survey`[45,46] and `quantreg`[47] packages for further details. First, the object `mydesign` specifying the design is defined. This is then replicated using the function `as.svrepdesign`:

```
mydesign <- svydesign(ids = ~mycluster, strata = ~mystrata, fpc = ~myfpc,
data = mydata, nest = TRUE, weights = ~myweights)
bootdesign<- as.svrepdesign(mydesign, type = "bootstrap", replicates = 100)
```

Quantile estimation is carried out using the following syntax:

```
fit<- withReplicates(bootdesign, quote(coef(rq(y ~x, tau=0.5, weights=.weights,
method="fn"))))
```

The resulting fitted object `fit` contains the estimated regression coefficients and their bootstrap variances. This object can be then passed to the following custom-defined function to produce a summary table, including p values:

```
format.rq.svy<- function(x, rdf){
V<- attr(x, "var")
FLAG<- length(V)== 1
se<- if(FLAG) sqrt(V) else sqrt(diag(V))
val<- cbind(as.matrix(x), se, NA, NA)
if(FLAG) val<- matrix(val, nrow=1)
val[,3]<- val[,1]/val[,2]
val[,4]<- 2 * (1 - pt(abs(val[, 3]), rdf))
colnames(val)<- c("Value","Std. Error", "t value", "Pr(>|t|)")
rownames(val)<- names(x)
return(val)
}
```

Downloaded from smm.sagepub.com at University College London on April 29, 2015

where the argument `rdf` specifies the residual degrees of freedom (i.e. $n - q$) for approximate p value calculation using t-distributions.

B.2. Multiple imputation by chained equations

The `mice`[48] routines are extremely flexible in that any method specified in the `method` argument of the main function mice(data, m=5, method="newmethod",...) will be looked for as the function `mice.impute.newmethod(y, ry, x, ...)`. The value returned by the latter has to be a vector of the same length as the number of missing values being imputed. An example of unoptimized code for QR imputation with logit transform argument `logit` is given as follows:

```
mice.impute.rq<- function (y, ry, x, logit=FALSE, omega=0.001,
method.rq="fn",...){
x<- cbind(1, as.matrix(x))
n<- sum(!ry)
p<- ncol(x)
u<- round(runif(n, omega, 1 - omega)*1e3)
u<- ifelse(u %in% c(1:4,996:999), u/1e3, (u - u %% 5)/1e3)
taus<- unique(u) # group quantiles
nt<- length(taus)
 xobs<- x[ry,]
yobs<- if(!logit) y[ry] else logit(y[ry],...)
xmis<- x[!ry,]
fit<- matrix(NA, p, nt)
for(j in 1:nt){
fit[,j]<- as.numeric(rq.fit(xobs, yobs, tau=taus[j],
method=method.rq)$coefficients)} # from package quantreg
ypred<- xmis%*%fit # n times nt matrix
ypred<- diag(ypred[,match(u, taus)]) # diagonal of n times n matrix
val<- if(!logit) ypred else invlogit(x=ypred, x.r=attr(yobs,"range"))
return(val)
}
logit<- function(x, x.r=NULL, epsilon=0.5){
if(is.null(x.r)) x.r<- range(x, na.rm=TRUE)+c(-epsilon, epsilon)
val<- log((x - x.r[1])/(x.r[2] - x))
attr(val, "range")<- x.r
return(val)
}
invlogit<- function(x, x.r=NULL){
if(is.null(x.r)) x.r<- attr(x, "range")
x<- exp(x)
val<- (x.r[1]+x.r[2]*x)/(1+x)
attr(val, "range")<- x.r
return(val)
}
```

Downloaded from smm.sagepub.com at University College London on April 29, 2015

In general, the above functions can be easily modified and generalized to include additional arguments such as, for example, estimation control parameters. Imputation based on restricted regression quantiles can be carried out using the following functions:

```
mice.impute.rrq<- function (y, ry, x, omega=0.001, method.rq="fn"){
x<- cbind(1, as.matrix(x))
n<- sum(!ry)
p<- ncol(x)
u<- round(runif(n, omega, 1 - omega)*1e3)
u<- ifelse(u %in% c(1:4,996:999), u/1e3, (u - u %% 5)/1e3)
taus<- unique(u)
nt<- length(taus)
xobs<- x[ry,]
 yobs<- y[ry]
xmis<- x[!ry,]
 fit<- rrq.fit(xobs, yobs, tau=taus, method=method.rq)$coef
 ypred<- xmis%*%as.matrix(fit)
val<- diag(ypred[,match(u, taus)])
return(val)
}
rrq.fit<- function(x, y, tau, method="fn"){
fit.lad<- rq.fit(x, y, tau=0.5, method=method)
r.lad<- fit.lad$residuals
r.abs<- abs(fit.lad$residuals)
beta<- fit.lad$coefficients
fit.lad<- rq.fit(x, r.abs, tau=0.5, method=method)
s.lad<- fit.lad$fitted.values
gamma<- fit.lad$coefficients
nt<- length(tau)
zeta<- rep(NA, nt)
for(i in 1:nt){
zeta[i]<- rq.fit(s.lad, r.lad, tau=tau[i], method=method)$coefficients}
val<- if (nt>1)
apply(outer(gamma, zeta, "*"), 2, function(x, b) x+b, b=matrix(beta))
else beta+zeta * gamma
if(nt>1) colnames(val)<- tau
return(list(coef=val, c=zeta, beta=beta, gamma=gamma))
}
```

Downloaded from smm.sagepub.com at University College London on April 29, 2015