In: International Journal of Social Research Methods, 2009, 12, 4, 293-304.

Imputation Methods for Handling Item-Nonresponse in Practice: Methodological Issues and Recent Debates

Gabriele B. Durrant

Contact details:

Gabriele B. Durrant
Southampton Statistical Sciences Research Institute (S3RI)
University of Southampton
S017 1BJ, Southampton, UK
Email: gbd@soton.ac.uk
Tel: 023 8059 5481
Fax: 023 8059 3846

Word count: 5990

(without Figure 1 and Table 1 it is 5778 words)

Biographical Note:

Gabriele B. Durrant is Lecturer in Social Statistics, Southampton Statistical Sciences Research Institute (S3RI) at the University of Southampton, UK. She previously worked as Senior Research Fellow at the ESRC National Centre for Research Methods. She has recently been awarded a major 3-year research grant by the Economic and Social Research Council (ESRC) to analyse hierarchical unit-nonresponse in sample surveys.

Imputation Methods for Handling Item-Nonresponse in Practice: Methodological Issues and Recent Debates

Abstract:

Nonresponse is a major problem often faced by social scientists when analysing survey data. A range of methods exists to impute the missing responses but the choice between these methods may be difficult. This article reviews advantages and disadvantages of a range of imputation methods and provides guidance on how to use such methods in practice. The paper introduces the reader new to the imputation literature to key ideas and methods. For those already familiar with imputation the paper highlights some new developments and recent debates. The paper discusses an example from the social sciences, applying several imputation methods to a missing earnings variable. The objective is to illustrate in a real data example basic considerations when choosing between methods and to advise practitioners in the use of such methods.

Keywords: item-nonresponse, fractional imputation, multiple imputation, hot deck imputation.

1. Introduction

Social scientists often face the problem of nonresponse when analysing survey data. Not only is it common for individuals to fail to participate in a survey at all, either because of refusal or non-contact, leading to 'unit-nonresponse', but it is also common for values of specific variables to be missing, i.e. 'item-non-response', for example because respondents may regard questions about income or beliefs as unduly intrusive. A wide range of statistical methods have been developed to deal with nonresponse, but the choice between these methods can be a difficult matter.

This article reviews a range of imputation methods used within the social sciences to compensate for item-nonresponse, and provides guidance on how to use such methods in practice. A case study is presented, illustrating the choice of imputation methods for a particular application. Simple imputation methods are commonly used within the social sciences but these may not be adequate in many circumstances (Ibrahim, Chen, Lipsitz and Herring, 2005). Methods such as fractional hot deck and multiple imputation may be more appropriate. In this paper, the application of hot deck methods is presented as a means of relaxing distributional assumptions made by standard parametric model-based imputation methods. The paper introduces the novice to the principles, methods and recent debates in the imputation literature. It also provides the reader familiar with imputation methods with newer developments in this field. The choice of an appropriate imputation method may strongly depend on the data available, the

application and the purpose of the analysis, as illustrated in an example from the social sciences.

The article is structured as follows. Definitions and basic assumptions about missing data patterns are introduced in section 2. In section 3, various imputation methods of relevance to social science applications are reviewed. A case study is discussed in section 4. Some concluding remarks are made in section 5.

2. Notation and Typology of Item-Nonresponse

Item-nonresponse in surveys occurs primarily if the sample member refuses to answer a question or does not know the answer. Different types of item-nonresponse can be distinguished. The missing-data pattern can be *univariate*, meaning that the missing values only occur in a single response variable, or *multivariate*, when the missing values occur in more than one variable, as illustrated in Figure 1. Since the most appropriate choice of imputation method may depend on the underlying missing data pattern it is important to investigate the missing data structure and possible reasons for nonresponse.

[Figure 1 about here]

To describe these different types of missingness more formally and to facilitate the discussion the following notation is introduced. Let s be a sample of n units and X the complete data matrix with element x_{ik} for the ith unit and kth variable, where $i = 1,...,n$, and $k = 1,...,K$. In the presence of missing data X_{obs} refers to the observed part of the matrix and X_{mis} to the missing part. Let R be

an indicator if an item is observed or missing, i.e. a matrix with elements $r_{ik} = 1$ if x_{ik} is observed and $r_{ik} = 0$ if x_{ik} is missing. This general multivariate missing data pattern is illustrated in Figure 1(i). For the univariate case, where only one variable, say x_{iK}, is subject to nonresponse, let X^* be the matrix of all the remaining fully observed auxiliary variables, $x_i^* = (x_{i1}, x_{i2}, ..., x_{iK-1})$, as illustrated in Figure 1(ii).

One problem with missing data is that it is usually not known how the nonresponse for each variable is generated, i.e. the distribution $f(R \mid X)$, referred to as the nonresponse mechanism, where f denotes the probability density function, is unknown. It is usually necessary to make assumptions about this distribution (Rubin, 1987), which often cannot be verified.

One simple assumption is that the data are *missing completely at random* (MCAR), defined as $f(R \mid X_{obs}, X_{mis}) = f(R)$, that is the missingness depends neither on X_{obs} nor on X_{mis}. For the univariate case this means that the probability of response depends neither on the variable subject to nonresponse, x_{iK}, nor on any other variable in X. MCAR is a strong assumption, which is likely to be violated in many social science applications. A weaker assumption is that the data are *missing at random* (MAR), i.e. $f(R \mid X_{obs}, X_{mis}) = f(R \mid X_{obs})$, which implies that missingness does depend on the observed but not on the missing values. For the univariate case, under MAR the missingness may depend on x_{iK}, however, when conditioning on other variables in X this dependency is wiped out. If the probability that an item is missing is likely to depend on the variable itself (even

when conditioning on observed values) the missing-data mechanism is *not missing at random* (NMAR) (Little and Rubin, 2002). Often researchers regard the MAR assumption as a useful approximation although MAR may not strictly hold in reality. However, the impact on estimators based on imputation under such a departure may be small and MAR-based procedures may still be usable. Several imputation methods under the MAR assumption are now discussed.

3. Imputation Methods

Imputation is a method to fill in missing data with plausible values to produce a complete data set, such that procedures used for analysing complete data may be applied. The main reason for carrying out imputation is to reduce nonresponse bias, which occurs because the distribution of the missing values, assuming it was known, generally differs from the distribution of the observed items. Rather than deleting cases that are subject to item-nonresponse the sample size is maintained resulting in a potentially higher efficiency than for case deletion. *Deterministic* imputation methods produce the same imputed value for units with the same characteristics, whereas *random* imputation methods may produce a different imputed value for each case. Usually, imputation makes use of a number of auxiliary variables that are statistically related to the variable in which item-nonresponse occurs by means of an *imputation model* (Lessler and Kalsbeek, 1992). It can have, however, serious negative effects if imputed values are treated as real values. To estimate the variance of an estimator based on imputed data

adequately, special adjustments may be necessary to correct for the increase in variability due to imputation.

Under imputation let $X_{\cdot k}$ denote the vector of imputed and observed values of X_k, such that $x_{\cdot ik} = x_{ik}$ if the value is observed, if $r_{ik} = 1$, and $x_{\cdot ik} = x_{ik}^I$ if the value is missing, if $r_{ik} = 0$, where x_{ik}^I denotes the imputed value for nonrespondent i in variable k. Let θ denote the parameter of interest in the population, e.g. a mean or a regression coefficient, and $\hat{\theta}$ an estimator of θ based on the sample data X in the case of full response. Applying imputation in the presence of missing data, the so-called *imputed estimator* $\hat{\theta}_\cdot$, based on observed and imputed values, is obtained.

The aim is to choose an appropriate imputation method to obtain an approximately unbiased and efficient imputed estimator. An important aspect is robustness under misspecification of assumptions, such as assumptions about the imputation or the nonresponse model. It is recommended to carefully consider the type of analysis to be conducted, for example if the goal is to produce efficient estimates of means, proportions or other aggregated statistics, or a complete micro-data file that can be used for a variety of analyses. Other issues are the availability of variance estimation techniques and practical questions concerning implementation and computing time.

Over the last few decades a wide range of imputation methods have been developed. *Deductive methods* impute a missing value by using logical relations between variables and choosing the most plausible value. *Mean imputation* imputes

the mean of a numeric variable for each missing item within that variable. Such methods may be commonly used in the social sciences, however, they are often not adequate to handle the missing data problem, for example distributions of survey variables may be compressed and relationships between variables may be distorted (GSS, 1996; Kalton, 1983; Lessler and Kalsbeek, 1992).

3.1 Regression Imputation

Regression imputation involves the use of auxiliary variables, for which the values are known for both complete units and units with missing values in the variable of interest. A regression model is fitted that relates a variable X_k to auxiliary variables X^*, i.e. the *imputation model*. In *deterministic* regression imputation the predicted values are used for imputation, which may however distort the shape of the distribution of X_k and the correlations between X_k and variables which are not used in the regression model. Under *random* regression imputation the imputed value is a randomly drawn value derived from the regression model that relates X_k to the auxiliary variables X^*. For example, if a linear model between X_k and X^* is considered a residual term is added to the predicted value from the regression, introducing randomisation and allowing for uncertainty in the predicted value. This residual may be obtained by drawing from a normal distribution with appropriate standard deviation or by computing the regression residuals from the complete cases and selecting an observed residual at random for each nonrespondent. A random regression model maintains the distribution

of the variable of interest and allows for the estimation of distributional quantities (Kalton, 1983; Nordholt, 1998). The method can make use of many categorical and numeric variables and performs well if the variable of interest is strongly related to auxiliary variables. The imputed value, however, is a predicted value either with or without an added on residual and not an actually observed or necessarily observable value. This can be a problem for imputing certain types of variables such as earnings and income variables, as illustrated in section 4. Another potential disadvantage of such a parametric approach is its sensitivity to model misspecification. If the regression model is not a good fit the predictive power of the model might be poor (Schenker and Taylor, 1996), which is of concern since the method strongly relies on the estimation of suitable predicted values.

3.2 Hot Deck Imputation Methods

Many approaches have been developed that assign the value from a unit with an observed item, the donor, to a unit with a missing value on that item, the recipient. Such imputation methods are referred to as *hot deck* methods, replacing the missing value x_{jk} with the imputed value $x_{jk}^I = x_{i'k}$ for some donor respondent i' for whom $r_{i'k} = 1$. One possibility for selecting donor values is to define so-called *imputation classes*, by allocating sample members into groups (classes), which are constructed by crossclassifying fully observed auxiliary variables. The imputed value is the response of a donor selected at random

within the relevant class. Hot deck imputation is common in practice. An advantage is that actually occurring values are used for imputation. Such methods are usually non-parametric (or semi-parametric) and avoid distributional assumptions. This is important if the data are skewed or show certain features, such as truncation and rounding effects, which are often present in social science data, or if the estimation of distributional quantities is of interest. Under hot deck imputation the imputed values will have the same distributional shape as the observed data (Rubin, 1987). For a hot deck method to work well a reasonably large sample size may be required.

3.3 Nearest-Neighbour and Predictive Mean Matching Imputation

Nearest-neighbour imputation is a hot deck method where the donor is selected by minimising a distance defined as a function of auxiliary variables (Lessler and Kalsbeek, 1992). The observed unit with the smallest distance to the nonrespondent unit is identified and its value is substituted for the missing item. *Predictive mean matching imputation* extends the idea of nearest neighbour imputation by incorporating the imputation model outlined in section 3.1 (Heitjan and Little, 1991; Little, 1988). In its simplest form it is nearest neighbour imputation where the distance is defined based on the predicted values of x_{ik} from an imputation model, denoted $\hat{x}_{ik}$. Randomisation can be introduced by defining a set of values that are closest to the predicted value and choosing one value out of that set at random for imputation (Little, 1988; Schenker and Taylor, 1996). Another form of predictive mean matching is *hot deck imputation within classes*, where the classes

are defined based on the range of the predicted values $\hat{x}_{ik}$. This achieves a more even spread of donor values within classes, reducing the variance of the imputed estimator. Donor values within classes may be drawn with or without replacement, where drawing without replacement is expected to lead to a further reduction in the variance (Durrant and Skinner, 2006a; Kim and Fuller, 2004). The method of predictive mean matching is an example of a composite method, combining elements of regression, nearest-neighbour and hot deck imputation. Since it is a semi-parametric method, which makes use of the imputation model but does not fully rely on it, it is assumed or suggested by some authors to be less sensitive to misspecifications of the underlying model than, for example, regression imputation (Schenker and Taylor, 1996).

3.4 Repeated Imputation: Multiple and Fractional Imputation

Instead of imputing one value for each missing item *repeated* imputation may be used, in the sense that M, for example $M = 3$, values are assigned for each missing item. This is done for two reasons. One is to improve the efficiency of the imputed estimator, which is the aim when using *fractional* repeated imputation (Durrant and Skinner, 2006a; Fay, 1996; Kim and Fuller, 2004), based on repeating a single (random) imputation method several times. Another reason is the simplification of variance estimation of a point estimator which is achieved when using *multiple* imputation (MI), as proposed by Rubin (1987). Here, the repeated imputed values themselves reflect uncertainty about the true but non-observed values.

3.4.1 Multiple Imputation

The basic idea of multiple imputation is as follows: impute the missing values using an appropriate imputation model that incorporates random imputation, repeat this M times, carry out the analysis of interest in each of the M resulting datasets and combine the estimates using Rubin's rules (Rubin, 1987). A combined estimate of θ is obtained as the average of the complete-data point estimates

$$\hat{\theta}. = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}.^{(m)} \qquad (1)$$

where $\hat{\theta}.^{(m)}$ is the imputed estimator for the m^{th} imputation. To obtain a variance estimate for $\hat{\theta}.$, the average of the complete-data variance estimates - the so-called within-imputation variance - and the variance estimate of the complete-data point estimates - the so-called between-imputation variance - have to be calculated and combined. The differences in the M results obtained from the M complete-data sets can be used as a measure of uncertainty caused by missing data. For this to work the multiple imputations need to fulfil certain conditions, i.e. they need to be *proper*. Using the definition in Schafer (1997), multiple imputations are said to be proper if they are independent realizations of the distribution of the missing values given the observed, $f(X_{mis} \mid X_{obs})$. An advantage of MI is that it is possible to produce complete micro-data files that can be used for a variety of analyses, beneficial when the dataset is analysed by a range of researchers. MI is currently maybe the most general approach, in particular for social scientists carrying out a number of different analyses and with missing values in several variables. Having to use multiple datasets,

however, may also be more burdensome for secondary analysts. Also, careful consideration needs to be given to the definition of the imputation model and the relationship to the analysis model for the method to work well. Reviews of MI can be found in Rubin (1996), Schafer (1997, 1999), Zhang (2003), Schafer and Olsen (1998), Allison (2000), Sinharay, Stern and Russell (2001) and Schafer and Graham (2002), of which the latter four are less technical and refer to applications in the social sciences.

In practice, different ways exist for generating (proper) multiple imputations. For example data augmentation algorithms, defined in a Bayesian framework, may be used (Lipsitz, Zhao and Molenberghs, 1998; Rubin, 1996; Schafer, 1997). The idea here is to solve iteratively tractable complete data problems. The algorithm consists of a series of so-called *imputation* and *posterior steps*. In the imputation step, the missing values are imputed based on an imputation model, e.g. using regression imputation. In the posterior step, the parameters for that imputation model are drawn from a distribution - the so-called posterior distribution - given the complete data from the imputation step (Allison, 2000; Schafer, 1997). Such an approach, however, is fully parametric and requires making assumptions about underlying distributions, such as multivariate normality, which may not be appropriate in some applications. Semi-parametric or non-parametric approaches that make less or even no distributional assumptions about the variable to be imputed may be advantageous. Such a non-parametric approach to MI is the *approximate Bayesian bootstrap* (ABB) (Rubin and Schenker, 1986). Defining

imputation classes, the donors within each class are sampled (bootstrapped) with replacement, generating a sample of the same size as the number of respondents that are available in each class. For each nonrespondent in a class one donor is selected with replacement from the set of bootstrapped respondents for that class at random. This is repeated M times.

An alternative semi-parametric MI approach is to incorporate a hot deck method in the imputation step of the MI data augmentation procedure. The novelty here is to use a form of predictive mean matching in the imputation step instead of regression imputation with the aim of relaxing residual assumptions, commonly made in standard data augmentation procedures. The approaches implemented in the imputation step are: (i) hot deck imputation within classes, and (ii) nearest neighbour imputation, where the classes and the nearest neighbours are defined based on the range of the predicted values of the imputation model respectively (for further details see Durrant and Skinner, 2006b). Such a combination of methods may have advantages, for example overcoming distributional assumptions by using a hot deck method and at the same time providing a simple variance estimation formula by using MI.

Some cases, however, have been reported where the MI variance estimation formula does not perform very well, depending on, for example, the point estimator of interest or the way the multiple imputations are generated, such as under special cases of the ABB (Allison, 2000; Fay, 1996; Heitjan and Little, 1991; Kim and Fuller, 2004; Nielsen, 2003; Rao, 1996).

3.4.2 Fractional Imputation

Under fractional imputation the estimator $\hat{\theta}.$ can be expressed in the same way as under multiple imputation in (1). The method, however, views the imputed estimator as a weighted estimator with fractional weights $1/M$ for each of the M imputed values. Examples of fractional imputation are the use of repeated random hot deck and repeated predictive mean matching imputation. The main aim of this approach is to improve the efficiency of the imputed point estimator. Kim and Fuller (2004) find that fractional imputation is more efficient than MI based on the same number of repeated imputations due to the additional variability in the MI methods required to reflect uncertainty in the parameter estimates. An advantage of fractional imputation is that it can be easily defined based on hot deck imputation, which makes less distributional assumptions in comparison to a fully parametric method, imputes actually observed values and can preserve distributional properties of the data.

It is often stressed that single value and fractional imputation do not reflect sampling variability under nonresponse correctly leading to underestimation of the variance (Schafer, 1999; Schafer and Graham, 2002; Sinharay et al., 2001). However, this is only true if imputed values are regarded as observed values. If a variance estimation technique is applied that takes account of the particular imputation method single and fractional imputation will lead to valid inference. A number of such estimation techniques have been developed in recent years, including two-phase, model-assisted and replication approaches (Durrant and

Skinner, 2006a ; Kim and Fuller, 2004; Rao, 1996; Skinner and Rao, 2002), to mention only a few.

4. A Case Study: Estimating Pay Distributions in the Presence of Item-Nonresponse

4.1 Example from the UK Labour Force Survey

To illustrate the properties of the methods described in section 3 and to demonstrate important considerations when applying imputation in practice, an application from the social sciences is discussed. The focus is on the choice of imputation methods to estimate a distribution function, with regards to bias, efficiency, robustness to model assumptions and ease of implementation. The illustration is motivated by the problem of estimating pay distributions of hourly pay in the United Kingdom based on Labour Force Survey (LFS) data, important in evaluating the impact of National Minimum Wage legislations (Stuttard and Jenkins, 2001). In this survey, the variable of interest, hourly pay of employees, denoted x_{iK}, is missing for some cases, whereas other variables in the dataset, denoted as a vector x_i^*, such as gender, occupation, qualification, industry section and other pay information, are fully observed. The aim is to estimate the distribution of x_{iK} by imputing the missing values using information on the fully observed variables x_i^*. For more information on the particular estimation problem and the data see Durrant and Skinner (2006a) and Stuttard and Jenkins

(2001). Although the example is based on a specific problem, it illustrates different properties of imputation methods and the choice between them.

In this application, the parameter θ, a function of the distribution of x_{iK} in the population of employees U, may be expressed as the proportion of employees earning below a certain pay threshold y, such as the National Minimum Wage. The parameter can then be expressed as $\theta = \frac{1}{N}\sum_{i \in U} I(x_{iK} \leq y)$, where $I(.)$ indicates if a sample member earns below this threshold or not. Of particular interest is the estimation of the proportion of employees earning below or around the National Minimum Wage. The variable x_{iK} is missing for a number of cases and various imputation methods are considered for estimating this parameter under the assumption of MAR. The imputed estimator may be written as

$$\hat{\theta}_{.} = \frac{1}{n}\sum_{i=1}^{n} I(x_{iK}^{I} \leq y). \qquad (2)$$

When applying imputation careful consideration needs to be given to the aim of the analysis, the estimator of interest, the type of data available, the missing data pattern and the properties of possible imputation methods in the context of the specific application. An investigation of observed cases of the hourly pay variable x_{iK} showed that truncation and rounding effects are an important feature of this variable (Durrant and Skinner, 2006a), since certain hourly pay rates are more likely to occur than others (e.g. employees may be more likely to earn £6.60/hour rather than £6.57). In addition, the variable is skewed, caused by the uneven distribution of pay in the population with only a few employees having a

very high income. To estimate its distribution correctly the imputation method should ideally reproduce such features. The point estimators of interest are the proportion of employees with pay below the National Minimum Wage (NMW), denoted $\hat{\theta}_{.1}$, and the proportion with pay between the NMW and £5/hour, denoted $\hat{\theta}_{.2}$. (In spring 2000, shortly after the introduction of the legislation, the NMW was £3.60 per hour for employees aged 22 and over.) Let us now consider the imputation methods described in section 3 for this application.

4.2 Imputation Approaches for LFS Application

Methods such as mean imputation do not seem suitable for this application, since they may distort the shape of the distribution of x_{iK} and lead to bias in the estimator of interest. Under the MAR assumption where the conditional distribution of x_{iK} for the nonrespondents is assumed to be the same as for the respondents, i.e. $f(x_{iK} \mid x_i^*, r_{iK} = 0) = f(x_{iK} \mid x_i^*, r_{iK} = 1)$, it would appear 'natural' to use the conditional distribution of x_{iK} given x_i^* fitted to respondent data, $\hat{f}(x_{iK} \mid x_i^*, r_{iK} = 1)$, and to draw the imputed values x_{iK}^I from this estimated distribution at the values x_{jK} observed for the nonrespondents, $r_{jK} = 0$. Regression imputation appears to be an obvious choice, representing the conditional distribution by a parametric regression model, such as $\ln(x_{iK}) = g(x_i^*; \beta) + e_i$, where $g(.)$ is a function of the covariates x_i^*, allowing for non-linear and interaction terms, β is a vector of regression parameters and e_i are the residuals. (The logarithmic transformation on x_{iK} is used to address the skewness of the variable and to approximate normality, which is common

practice for earnings variables.) Using the predicted values from this model for imputation, however, may lead to serious underestimation of θ. Random regression imputation, denoted Reg Imp, can address this problem, setting $\ln(x_{iK}^I) = g(x_i^*;\hat{\beta}) + \hat{e}_i$, where $\hat{\beta}$ is an estimator of β based on respondent data and $\hat{e}_i$ is a randomly selected residual, either drawn from a normal distribution or as an empirical residual from respondent data.

Instead of single imputation multiple imputation is used next, to take advantage of the simple variance estimation formula. A data augmentation procedure is implemented using regression imputation in the imputation step (as in Schafer, 1997), denoted DA-Reg Imp(*M*). The number of repeated imputations is given in parentheses, here *M*=10. However, for this application it was found that the imputed values do not reproduce truncation and step effects of the hourly pay distribution, because the imputed values would take on any value on the scale and not the value of the nearest decimal place. This may lead to bias around such effects (Durrant and Skinner, 2006b). In addition, the residual assumptions made under such regression imputation, when drawing residuals from a normal distribution with a constant variance with respect to the auxiliary variables, may not hold in this application. In particular, the assumption of a constant variance seemed likely to be violated (Durrant and Skinner, 2006b), resulting in adding on inappropriate residuals to the predicted values. This illustrates a potential inadequacy of standard parametric (single or multiple) imputation approaches for this application. The effects of such methods (Reg Imp(1) and DA-Reg Imp(10))

when applied to LFS data can be seen in Table 1, leading to quite different estimates than for the following hot deck imputation methods, indicating an overestimation of θ_1 and underestimation of θ_2. The sensitivity towards misspecification of model assumptions for such parametric methods has been further illustrated in Durrant and Skinner (2006b).

[Table 1 about here]

In contrast, hot deck imputation methods relax such residual assumptions. The imputed value from a donor is a genuine value and such methods seem more suitable for this application. The basic hot deck imputation method considered is predictive mean matching, denoted PMM(1). In addition to bias, it is of interest to consider the efficiency of the point estimator under imputation. A number of approaches to reducing the variance inflation effect under nearest neighbour imputation, due to the multiple usage of donors, are considered. One possibility is to define imputation classes based on the range of the predicted values and to draw donors by simple random sampling within classes, denoted IC. The variance will be smaller if donors are drawn without replacement. Another approach for reducing the variance is to employ repeated imputed values, based on fractional imputation. This is implemented by repeating the imputation class method 10 times, denoted IC(10), ensuring that at least 10 donor values are available in each class. In addition, the predictive mean matching imputation based on nearest neighbour is extended to fractional imputation, denoted PMM(10), by taking the 5 nearest donor neighbours above and below the

predicted value of the nonrespondent. Since these forms of hot-deck imputation still make assumptions about the form of the imputation model these approaches are referred to as semi-parametric methods.

The use of such forms of repeated imputation 'naturally' leads to the implementation of MI taking into account the uncertainty of the parameters of the imputation model. This may provide a simpler way to estimate the variance of the point estimator but may risk a less efficient point estimator. For this application, non- or semi-parametric forms of MI seem more suitable, such as the approximate Bayesian bootstrap using imputation classes as suggested in Rubin and Schenker (1986), where the classes are defined based on the predicted values, denoted ABB-IC(10). Another possibility is to implement predictive mean matching imputation within a data augmentation procedure to generate MI by a) hot deck imputation within classes, denoted DA-IC(10), and b) predictive mean matching based on nearest neighbour imputation, denoted DA-PMM(10). Table 1 shows that the hot deck methods, either under single, fractional or multiple imputation, lead to very similar point estimates for θ_1 with slightly greater variation for θ_2. It was found that the hot deck methods are able to reproduce certain features of the hourly pay distributions, such as step and truncation effects, and perform better than the parametric approaches when applied to LFS data. These methods also provide a tool to compensate for departures from residual assumptions, such as non-constant variance. The two regression methods do not seem to perform as well, indicated for example by the higher estimates for θ_1 in Table 1. For comparison, estimates based on observed

cases are also provided. For the spring quarter the variable hourly pay was missing for 9182 and observed for 7176 cases. We can see that ignoring the missing data problem will in this case lead to significant overestimation, since the missing data reflect information on employees who are not paid by the hour but for example monthly or annually and who generally have a higher income than hourly paid employees.

5. Conclusions

Based on an example from the social sciences the paper illustrates basic considerations in relation to imputation and provides guidance on the choice of imputation methods and their implementation in practice. It is important to consider the type of analysis, the type of point estimator of interest and any impacts of the choice of method on model robustness, bias and efficiency. The paper aims to clarify some recent misconceptions and to highlight some new developments in this field. In particular, the paper tries to present a more balanced view of the different types of imputation methods available and to reconcile different approaches.

Multiple imputation is an important and powerful form of imputation and has the advantage of comparatively simple variance estimation. Fractional hot deck imputation is introduced with the aim of improving the efficiency and the sensitivity to model misspecifications of an imputed estimator. Parametric regression imputation, including standard multiple imputation, may not seem

appropriate for situations where parametric assumptions are likely to be violated. Hot deck methods may better preserve distributional properties, important for many applications in the social sciences. To take advantage of multiple imputation and hot-deck properties, semi-parametric forms of MI may be favourable. Research is needed to further develop semi- and non-parametric imputation methods and to make such methods more easily accessible to social scientists.

6. References

[1] Allison, P.D. (2000). Multiple Imputation for Missing Data, A Cautionary Tale, *Sociological Methods and Research*, 28, 3, 301-309.

[2] Durrant, G.B. and Skinner, C. (2006a). Using Missing Data Methods to Correct for Measurement Error in a Distribution Function, *Survey Methodology*, 32, 25-36.

[3] Durrant, G.B. and Skinner, C. (2006b). Using Data Augmentation to Correct for Nonignorable Nonresponse when Surrogate Data are Available: An Application to the Distribution of Hourly Pay, *Journal of the Royal Statistical Society, Series A,* 169, 605-623.

[4] Fay, R.E. (1996). Alternative Paradigms for the Analysis of Imputed Survey Data, *Journal of the American Statistical Association*, 91, 434, 490-498.

[5] Government Statistical Service (GSS) (1996). Report of the Task Force on Imputation, *Methodology Series*, 3, London. Retrieved from

http://www.statistics.gov.uk/downloads/theme_other/GSSMethodology_No_03_v2.pdf on 25 August 2008.

[6] Heitjan, D.F. and Little, R. (1991). Multiple Imputation for the Fatal Accident Reporting System, *Journal of the Royal Statistical Society*, *Applied Statistics*, 40, 1, 13-29.

[7] Ibrahim, J.G., Chen, M.H., Lipsitz, S.R. and Herring, A.H. (2005). Missing-Data Methods for Generalised Linear Models: A Comparative Review, *Journal of the American Statistical Association*, 100, 469, 332-346.

[8] Kalton, G. (1983). *Compensating for Missing Survey Data*, Research Report Series, Institute for Social Research, Survey Research Centre, University of Michigan, Michigan.

[9] Kim, J.K. and Fuller, W. (2004). Fractional Hot Deck Imputation, *Biometrika*, 91, 3, 559-578.

[10] Lessler, J.T. and Kalsbeek W.D. (1992). *Nonsampling Error in Surveys*, New York, Chichester.

[11] Lipsitz, S.R., Zhao, L.P. and Molenberghs, G. (1998). A Semiparametric Method of Multiple Imputation, *Journal of the Royal Statistical Society*, *Series B*, 60, 1, 127-144.

[12] Little, R.J.A. (1988). Missing-Data Adjustments in Large Surveys, *Journal of Business and Economic Statistics*, 6, 3, 287-301.

[13] Little, R.J.A. and Rubin, D.B. (1990). The Analysis of Social Science Data with Missing Values, *Sociological Methods and Research*, 18, 3, 292-326.

[14] Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, New York.

[15] Nielsen, S.F. (2003). Proper and Improper Multiple Imputation, *International Statistical Review*, 71, 3, 593-627.

[16] Nordholt, E.S. (1998). Imputation: Methods, Simulation, Experiments and Practical Examples, *International Statistical Review*, 66, 2, 157-180.

[17] Rao, J.N.K. (1996). On Variance Estimation with Imputed Survey Data, *Journal of the American Statistical Association*, 91, 434, 499-506.

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

[19] Rubin, D.B. (1996). Multiple Imputation after 18+ Years, *Journal of the American Statistical Association*, 91, 434, 473-489.

[20] Rubin, D.B. and Schenker, N. (1986). Multiple Imputation for Interval Estimation from Simple Random Samples With Ignorable Nonresponse, *Journal of the American Statistical Association*, 81, 394, 366-374.

[21] Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*, London: Chapman and Hall.

[22] Schafer J.L. (1999). Multiple Imputation: A Primer, *Statistical Methods in Medical Research*, 8, 3-15.

[23] Schafer, J.L. and Graham, J.W. (2002). Missing Data: Our View of the State of the Art, *Psychological Methods*, 7, 2, 147-177.

[24] Schafer, J.L. and Olsen, M.K. (1998). Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective, *Multivariate Behavioral Research*, 33, 545-571.

[25] Schenker, N. and Taylor, J.M.G. (1996). Partially Parametric Techniques for Multiple Imputation, *Computational Statistics and Data Analysis*, 22, 425-446.

[26] Sinharay, S., Stern, H.S. and Russell, D. (2001). The Use of Multiple Imputation for the Analysis of Missing Data, *Psychological Methods*, 6, 317-329.

[27] Skinner, C. and Rao, J.N.K. (2002). Jackknife Variance Estimation for Multivariate Statistics under Hot Deck Imputation From Common Donors, *Journal of Statistical Planning and Inference*, 102, 1, 421-422.

[28] Stuttard, N. and Jenkins, J. (2001). Measuring Low Pay Using the New Earnings Survey and the Labour Force Survey, *Labour Market Trends*, 55-66.

[29] Zhang, P. (2003). Multiple Imputation: Theory and Method, *International Statistical Review*, 71, 3, 581-592.

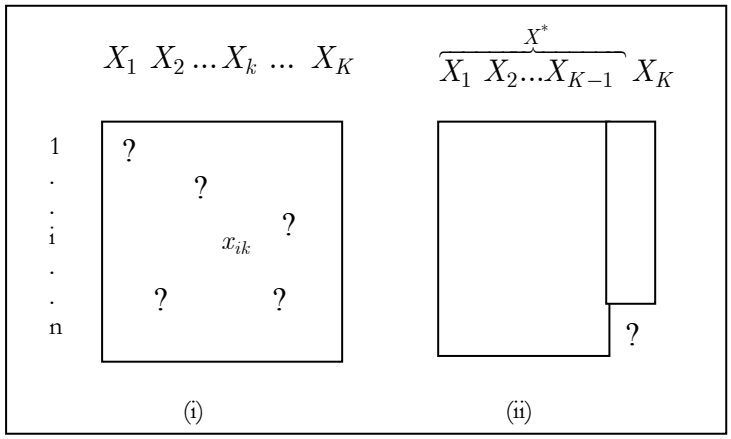

Figure 1: Missing data patterns in sample s:

(i) general multivariate pattern, where missingness (indicated by a question mark) may occur in any variable $X_1, X_2, \ldots, X_K$ and for any item i in the sample. The element x_{ik} for the ith unit and the kth variable is observed here.

(ii) univariate pattern, where only one variable X_K is subject to missing data and all remaining variables $X_1, X_2, \ldots, X_{K-1}$, denoted X^*, are fully observed.

(Source: adapted from Little and Rubin, 1990).

Imputation Method		$\hat{\theta}_{\cdot 1}$ in %	$\hat{\theta}_{\cdot 2}$ in %
Random Regression Imputation	REG IMP(1)	1.24	26.27
Predictive Mean Matching	PMM(1)	0.51	29.06
Predictive Mean Matching	PMM(10)	0.50	29.04
Hot Deck Imputation Within Classes	IC(10)	0.49	29.07
Data Augmentation With Regression	DA-REG IMP(10)	1.22	26.11
Data Augmentation With Predictive Mean Matching	DA-PMM(10)	0.50	28.78
Data Augmentation With Hot Deck Within Classes	DA-IC(10)	0.51	26.99
Approximate Bayesian Bootstrap With Hot Deck Within Classes	ABB-IC(10)	0.51	27.98
No imputation method - using observed cases only	No method	1.86	35.94

Table 1: Estimates of the proportion of employees with pay below the National Minimum Wage, denoted $\hat{\theta}_{\cdot 1}$, and the proportion with pay between the minimum wage and £5/hour, denoted $\hat{\theta}_{\cdot 2}$, given in percent, under various fractional and multiple imputation methods applied to the March-May 2000 quarter of the LFS. The number of imputations M is indicated in parentheses.