IMPUTATION OF MISSING VALUES IN THE FAMILY RESOURCES SURVEY 2007-08

INTRODUCTION

Imputation is the process in which missing values in a data set are converted to non-missing values.

When a respondent answers a particular question in a survey they can state that they don't know the answer to a question, or simply refuse to give a response. Such responses are recorded and are referred to as *'missing values'.*

These values can either be left as missing, in which case you would have gaps in your data set, or replaced (*imputed*) with an estimate of the answer that the respondent would have given if they had actually answered the question.

User requirements have deemed the latter process necessary in the Family Resources Survey (FRS). The main objective of imputation is to maximise the information available to users for analysis. Furthermore, the imputation carried out simplifies the analysis for users and helps to secure the uniformity of analysis created from the FRS data sets.

It should be noted that none of the variables in the admin and care data sets are imputed. Furthermore, very few of the variables in the benefits dataset are imputed using the processes outlined below, and that benefit editing is carried out separately to the rest of imputation.

Methodology

Imputation on the FRS is carried out in three different ways. A brief overview of these methods is given here:

- **Bulk edits** – converting en masse a batch of cases with missing values that satisfy a particular characteristic to an identical value. This is a very crude method of imputation and can only be used in certain circumstances. For example, for people who don't know if they are in receipt of a particular benefit, we could:

  i) edit the answers to yes, in which case we would have to open up a record for the particular benefit and impute answers for it

  ii) edit such answers to no – which is known as *closing down routes* and is the default principle adopted in the imputation of such *routing* variables in the FRS.

- **Hotdecks** – examining the data set for non-missing cases that have similar characteristics to that missing case with the missing value, and substituting one of these non-missing values for the missing case at random. It is usual for the characteristics to bear some relationship to the variable to be imputed; the theory being that all cases matching the chosen characteristics will have similar values for the variable we are concerned with. For example we could impute rent for a household by randomly selecting a non-missing value from a case with the same number of rooms, council tax band, type of landlord and region as the case in question.

- **Algorithms** – a process in which one can predict the missing value for a particular case by looking at other relevant characteristics and applying a pre-determined set of rules (e.g. modelling council tax payments based on council tax band, local authority and entitlement to discount).

Missing Values

There are four possible types of missing values in the FRS:

.A – denotes a *'skipped'* response. Such a response occurs where a respondent has not been routed to this particular question and an answer is not therefore required and imputation is not normally necessary

.B – denotes the fact that the respondent answered *'don't know'* to the question and imputation will normally be required

.C – denotes a *'refusal'* to answer a question and, again, imputation is normally required

.D – is only output in the production of derived variables, and denotes either a mistake in the imputation process or faulty logic in the Derived Variable code. All .Ds in income and expenditure data are investigated and corrected prior to user release (there will usually be some remaining in the Care dataset as data is not imputed in this dataset).

Imputation Checking

Checks are carried out to ensure that the imputation process has not changed the distribution of the data. Examples of these checks are as follows:

- A comparison of the means, standard deviations and minimum/maximum values for each variable is undertaken both prior to and after imputation. Any large discrepancies (indicating that imputation is potentially biasing the data) are investigated.

- There can be cases in Hotdecks where we impute a large number of cases to a particular value, which is taken from one particular 'donor' case. This is a source of potential bias, and checks exist within hotdecks to monitor this. Where these checks show this to be a problem, remedial action, in the form of adjusting either the imputed value or the hotdeck, is taken.

- Finally credibility checks are run, which ensure that the data within individual cases is consistent, and feasible values have been imputed. Examples of these include:

  i) Checking that housing costs are generally less than income for cases in which components of either have been imputed.

  ii) Checking that gross income is greater than or equal to net income.

  iii) Checking that personal pension contributions are generally less than income for cases where components of either have been imputed.

Tables of Results

Table 1 provides an overall summary of imputation outlining the number of missing values initially and how many were imputed by each method. It also provides a comparison with the previous year. It should be noted that hotdecking is the most common method of imputation, followed by bulk edits.

- As with any questionnaire, a typical feature of the FRS is the gatekeeper question positioned at the top of a block of further questions, at which a particular response will open up the block. If the gatekeeper question itself is answered as 'don't know' or 'refused', the block contains skipped values for all variables within it.

- A missing gatekeeper variable could be imputed such that a further series of answers would be expected. However, these answers will not appear because a whole new route has been opened. For example, if the amount of rent is missing for a record and has since been imputed, any further questions about rent would not have been asked. From

the post-imputed database, it will appear that these questions should have been asked because a value is there for rent.

Table 2 shows the extent of imputation on the BENEFITS table. Each benefit type is listed by variable, showing the number of expected responses, the number and percentage imputed and the number left missing. Each benefit is listed on the first sheet although this is repeated for each benefit on the subsequent sheets.

Table 3 shows the extent of imputation on all tables. Each variable is listed with the number of expected responses, the number and percentage imputed and the number left missing. Apart from the BENEFITS table, where any variable has had a missing imputed all missing values for that variable will have been imputed.