AN INVESTIGATION INTO OPTIMUM INCOME IMPUTATION METHODOLOGY FOR THE SCOTTISH HOUSE CONDITION SURVEY DATA

VISHNU G GANGLANI

Dissertation submission for the award of MSc in Applied Statistics.

Date: **SEPTEMBER 2007**

in co-operation with: **THE SCOTTISH GOVERNMENT**

ABSTRACT

Missing income data in the Scottish House Condition Survey (SHCS) is problematic because it hinders full representation of fuel poverty in Scotland, and thus potentially impedes the Scottish Government's targets. Imputation methods fill missing data with plausible values so that inferences can be made on the complete dataset generated. This study examines optimum imputation methodology to adopt for the SHCS missing income data. A descriptive qualitative analysis is carried out of the previous imputation methodology used, since it did not completely impute missing income data. Missing data patterns and mechanisms are then investigated by producing relevant descriptive statistics. General exploratory analysis of the data is carried out, ranging from individual variable summary statistics to checking for multivariate normality of the dataset. Data reduction techniques are considered to reduce the large number of variables in the multivariate dataset to a relevant few. Imputation methods are then applied to the dataset. A total of nine different imputation methods (single and multiple methods) are employed on the dataset. The results are then compared to complete case (non-missing) statistics. The multiple imputation methods appear to outperform the single imputation methods on the whole. However, based on criteria of consistency, ease of implementation and efficacy, semi-parametric (logistic regression based) hot deck imputation is recommended as the optimum income imputation methodology for the SHCS in its present context.

ACKNOWLEDGEMENTS

I am grateful to Ian Mate, my line manager at The Scottish Government, and his team (Dave Cormack, Debbie Amabile, Pat Cairns and Stephen Hinchcliffe) for the opportunity to undertake such interesting research, and the accompanying autonomy and support that was provided during the course of my dissertation placement at The Scottish Government. I am also thankful to Dr. Kay Penny, my supervisor at Napier University, for her immense support in helping me progress and accomplish this study in the allotted period. Finally, many thanks to my family for their support with my decision to further my studies.

CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Preamble

The Scottish House Condition Survey (SHCS) is a large, nationally representative study in Scotland in which comprehensive information about both households and their dwellings is collected (Scottish Fuel Poverty Statement, 2002). The Scottish Government is the main organisation responsible for supporting the ministers (and ministries) with their decisions in Scotland. The SHCS is the primary source of information about the prevalence of fuel poverty in Scotland. The SHCS is carried out every year in continuous format (from 2003 onwards). Prior to 2003 surveys were conducted in 1991, 1996 and 2002. The survey is administered through interviews held at the households of respondents. The calculation of fuel poverty incorporates income values for the households. Hence, the repercussions of having missing income data are considerable. The resulting missing values for fuel poverty could inhibit achievement of The Scottish Government's target for eradicating fuel poverty by 2016 (discussed below), since the missing data could potentially include households that are fuel poor and hence undermine initiatives such as the 'Warm Deal', described below. The concepts of household income, fuel poverty and its importance are explained in this section, followed by an assertion of the aims of the study.

1.2 Household Income

Household income used in the SHCS is defined as the income of the Highest Income Householder (HIH) and his/her spouse/partner. Household income comprises all income from the following sources :

- employment, self-employment, part-time and casual work
- state benefits including Council Tax Benefit and Housing Benefit
- student grants and loans

- any other regular non-work income including non-state pensions, investments income and so forth.

The respondents are asked numerous questions pertaining to their income situation when the SHCS is carried out. These include if they are in paid work or leave, what their pay is after deduction of taxes ('take home' pay), how many hours they work per week, if their partners are in paid work, if their partners are employed and numerous other questions relating to employment income, benefit income and the above-mentioned categories. It is easy to envisage how a respondent may be reticent towards revealing such information or may genuinely not know the answer. This leads to the emergence of missing income data at the item level (item non-response).

1.3 Fuel Poverty

According to the ministerial foreword in the Scottish Fuel Poverty Statement (2002), a considerable number of people in Scotland have to choose "between staying warm or spending their money on basic necessities such as food and clothing". In other words, they are fuel poor. "A person is living in fuel poverty if, in order to maintain a satisfactory heating regime, they would be required to spend more than ten per cent of their household income (including Housing Benefit or Income Support for Mortgage Interest) on all household fuel use" (UK Fuel Poverty Strategy, 2001; s.95 Housing (Scotland) Act 2001). Fuel poverty is calculated by income (the costs of heating a property form a greater proportion of total income for those on low incomes), fuel costs (higher prices reduce the affordability of fuel) and energy efficiency (the thermal quality of the building and the efficiency of the heating source determine the amount of energy that must be purchased to heat the home adequately). The Scottish Government has asserted that it will eradicate fuel poverty in Scotland by November 2016. Steps towards this target have already been initiated via the 'Central Heating Programme' (which makes provisions of central heating for people aged over sixty) and the 'Warm Deal' (which provides grants to individuals dependent on criteria based on benefits).

1.4 Importance of Study

The code for calculating fuel poverty explicitly incorporates income (annual household income). If income data is missing, one cannot calculate fuel poverty values for those particular cases. Thus one cannot ascertain if those households are eligible for the various programmes described above to alleviate fuel poverty. More importantly, if fuel poor households form part of this missing group, The Scottish Government will not be able to meet its targets of fuel poverty eradication. After attempted imputation of missing income values by the Office of National Statistics (which is in charge of handling imputation of missing data for the SHCS), there still remained a substantial proportion (6%) of income values that were missing for the 2003/4 and 2004/5 datasets. That is, the imputation strategy was not completely successful. Based on these social, economic and technical reasons, this study is extremely important to propose more efficient imputation methods and to provide the necessary figures for missing fuel poverty values.

1.5 Current Imputation Situation

Imputation involves 'filling in' missing values so that the new values are as close as possible to the true unknown values of the data. Various imputation methods have been carried out on income data of the SHCS over the past few years. Since the recent methodology resulted in missing income data even after imputation was implemented, one of the aims of the study is to review the imputation methods previously used.

1.6 Aims of study

The aim of this research is to review the previous method(s) employed to impute missing income data in the SHCS, and to eventually develop an optimum method for imputation of these data as a recommendation to the SHCS department in The Scottish Government (through assessing the impact of various imputation methods on a relevant dataset).

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Imputation methodologies, not only provide an effective means of dealing with missing data, but also improve efficiency of any statistical inferences drawn from the datasets. The choice of technique of dealing with missing data could have serious repercussions (such as substantial bias) on a study. This is of course contingent on the percentage of missing data and the respective *missing-data pattern* (described below) of variables in a dataset. Before delving into the realm of imputation methods, it is important to have a basic understanding of the problem or science of missing data.

2.2 Missing Data

A data matrix normally comprises real numbers representing the values of continuous variables, such as age or income, or representing categories of response, which may be unordered (for example, a variable describing ethnic origins) or ordered (for example, a variable measuring level of education) (Little and Rubin, 2002). Missing data describe the unobserved entries in such a data matrix. Rules for classification of missing data are not as lucid as may be apparent. For example, whereas refusal to fill in one's income data in a survey would equate to missing data (as would unobserved entries due to external constraints in an experiment), the lack of available answers for an optimum response from a respondent in a questionnaire may not be a missing value, since the latter would have to possibly tick a 'don't know' option which truly reflects lack of expression rather than missing information as in the former situation.

Even though many statistical packages facilitate non-response identification by the use of coding unobserved data (and allotting them to different categories like 'don't know' or 'refused' in some cases), numerous statistical procedures or packages simply exclude missing data in analysis. This phenomenon is often referred to as complete-case analysis, since only

observed data are incorporated into any data analysis. When the incomplete cases form a small proportion of all cases (for example, less than 5%), complete-case analysis may be a reasonable solution to dealing with any missing data. Proponents of this methodology extend this (depending on statistical method used) to even large fractions of missing data (Allison, 2001). However, since the aim of the statistical analyst is usually to draw an inference that enables conclusions on the whole data set/sample, more appropriate methods are required to reveal the essence of the values concealed as missing entries. In multivariate datasets, where missing values occur on more than one variable, incomplete cases are usually a substantial part of the data (for example in national household survey datasets). Deleting these missing values may be inefficient, causing large amounts of potentially useful information to be discarded. This omission may introduce bias, if the incomplete observations systematically differ from those that are completely observed, rendering the latter unrepresentative of the population of all cases (Schafer, 1997).

2.3 Missing-Data Patterns

Missing entries usually follow defined structures, also referred to as *missing-data patterns* (Little and Rubin, 2002). These describe how many observations are present in the data matrix and how many are missing. Missing-data patterns are important because in some instances they determine which statistical method or technique to employ to deal with the missing data. To describe the most important missing data patterns, the following notations are used. Let Y = (Y_{ij}) denote an (n x p) rectangular data set of complete data, with *i*th row Y_i = (Y_{i1}, …., Y_{ip}) where Y_{ij} is the value of variable Y_j for subject *i*. Let M = (m_{ij}) be the missing-data indicator matrix for the missing data, such that m_{ij} = 1 if Y_{ij} is missing and m_{ij} = 0 if Y_{ij} is present. M thus defines the missing-data pattern. Illustrations of these matrices are shown in Figure 2.1 below.

Figure 2.1 Data Matrices

Matrix Y

variables

J

I	1	2	3	4	.	p
1		**?**	**?**			
2						**?**
3	**?**			**?**		
4						**?**
.	**?**	**?**				
N			**?**			

Matrix M

variables

J

I

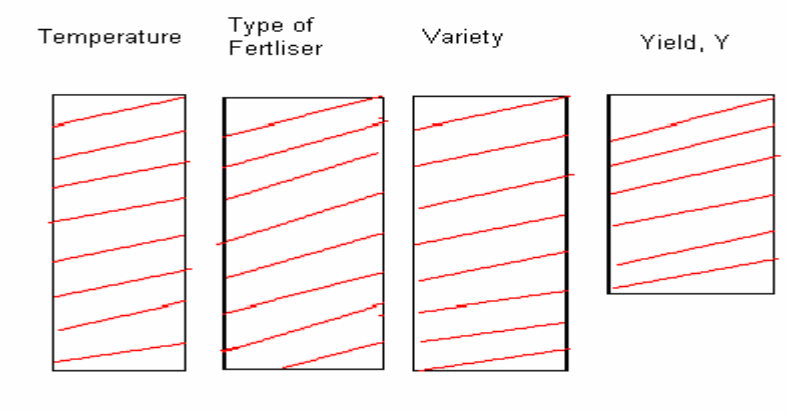
0	**1**	**1**	0	0	0
0	0	0	0	0	**1**
1	0	0	**1**	0	0
0	0	0	0	0	**1**
1	**1**	0	0	0	0
0	0	**1**	0	0	0

Some important missing-data patterns are described below.

Univariate Nonresponse

Figure 2.2 below describes a univariate missing-data pattern which emerges when missing data is confined to only one variable. It has applications in agricultural experiments, where parameters of temperature, fertilizer-type and variety are all needed to determine the yield of a crop (Hao et al., 2005). Missing values for yield could surface due to unsuccessful germination or even erroneous data entry. Techniques to deal with such missing-data patterns would aim to restore the balance of the original experimental design.

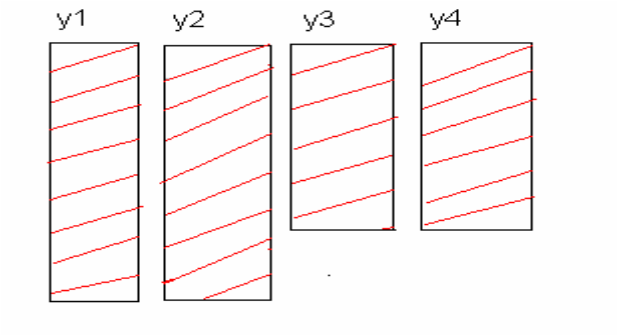
Figure 2.2: Univariate Missing-Data Pattern



Multivariate Two Patterns

When the single incomplete variable Y in Figure 2.2 is replaced by a set of variables Y_{J+1}, …., Y_K, all observed or missing on the same set of cases, another common pattern emerges. Such a pattern is usually found in unit nonresponse in sample surveys, where a questionnaire is administered and some respondents do not complete it because of noncontact, refusal or other reason . For example, if people in a subset of the SHCS refuse to answer questions relating to income (such as net pay, numbers of hours worked, etc.). The variables that are complete could be variables that deal with physical characteristics of the house or location. Figure 2.3 below depicts such a pattern. So the first two variables, y1 and y2, could be location and household size, and the latter two, y3 and y4, could be usual net pay of respondent and number of hours worked per week.

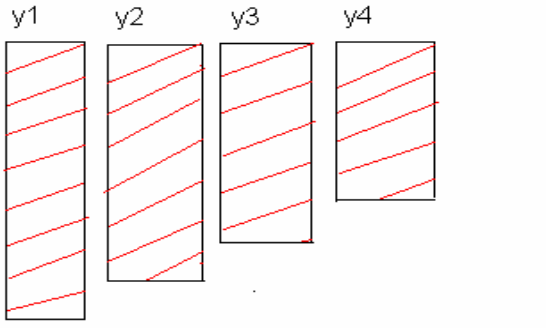
Figure 2.3: Multivariate Two Patterns



Monotone

Where variables can be arranged so that all Y_{J+1}, …., Y_K, are missing , for all J = 1, … , K – 1, the resulting pattern is called a monotone pattern. For example, in longitudinal studies, attrition (when subjects drop out prior to the end of the study and do not return) results in monotone patterns (Clayton et al., 2002). Figure 2.4 illustrates a monotone pattern with K=4. It is suggested that methods to handle missing data with monotone patterns are easier to implement than those with general patterns (Little and Rubin, 2002) which are also illustrated in Figure 2.5 below.

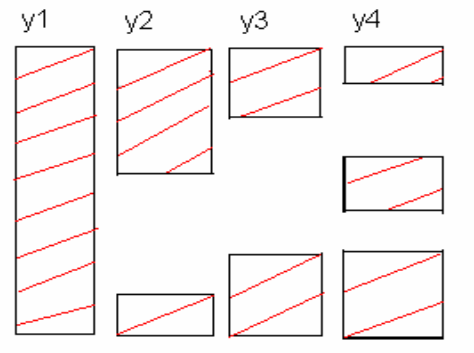
Figure 2.4: Monotone



General Pattern

A general missing-data pattern is one that has no set structure. The pattern is arbitrary. Figure 2.5 describes an example of such a dataset. Many real multivariate datasets have such missing-data patterns.

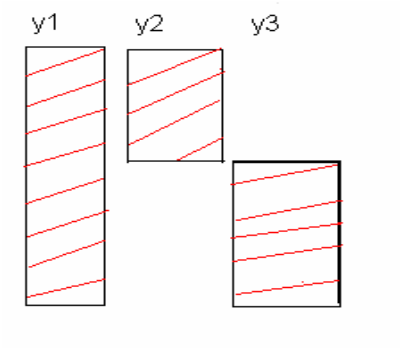
Figure 2.5: General



File Matching

In some situations, when a dataset contains numerous missing values, it possible for variables not to be observed together. This implies that some parameters may not be estimable from the data (Ziegler, 2006). An example of such a pattern is illustrated in Figure 2.6. In this pattern, y1 refers to a set of variables that is common to both data sources and fully-observed, y2 is a set of variables observed for the first data source but not the second and y3 is a set of variables observed for the second data source but not the first. This may arise in a household survey where y1 refers to location of the home, y2 refers to age of respondents only if they have partners, and y3 age of respondents if they do not have partners.

Figure 2.6: File Matching

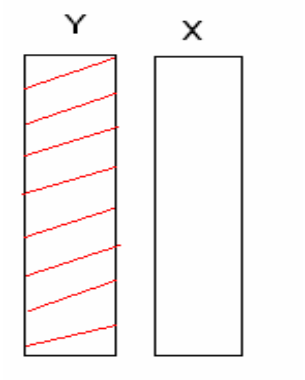


Factor Analysis

Factor analysis involves the creation of factors that describe variables in a data set. These latent factors or variables are completely missing, since nothing is known about them.

It is possible to analyse missing data of latent variables that are completely missing when performing a factor analysis (Raiko and Valpola, 2001). Figure 2.7 depicts such a missing-data pattern, where Y represents a set of variables that are fully observed and X represents a set of latent variables that are completely missing. In this case, factor analysis is considered as a multivariate regression of Y on X. That is, X can be treated as having missing values and these can be imputed using appropriate techniques. The necessary assumptions under factor analysis are maintained (For example, the components of Y are independent given X).

Figure 2.7: Factor Analysis



2.4 Missing-data Mechanisms

It is essential to understand what mechanisms give rise to missing-data patterns described above. A knowledge of the mechanism behind the missingness helps allot an optimum method of dealing with the missing data. Missing data mechanisms were formalized into theory when Rubin (1987) treated missing data indicators as random variables and assigned them a distribution. Using the same notations as before, the missing data mechanism is the conditional distribution of M (missing data indicator matrix) given Y , where φ represents unknown parameters : f (M|Y, φ).

Missing Completely at Random (MCAR) refers to the mechanism when missingness does not depend on the values of the data Y, missing or observed, that is if:

$$f(\mathrm{M}|\mathrm{Y}, \varphi) = f(\mathrm{M}|\varphi) \text{ for all } \mathrm{Y}, \varphi \qquad \text{.....(2.1)}$$

This assumption does not mean that the missing-data pattern is random but that the missingness does not depend on data values. That is, the probability of missing data on Y is unrelated to the value of Y itself or to the values of any other variables in the data set. When this assumption is satisfied for all variables, the set of individuals with complete data can be regarded as a simple random sample from the original set of observations. The MCAR assumption is easily violated when those who did not report incomes were part of a particular age group (for example over forty years old). MCAR is a strong assumption but has practical applications in many situations. For example, due to cost constraints, the assumption could be employed in research design experiments to measure variables that are expensive to measure by using a random subset and assuming MCAR for the remainder of the sample (Allison, 2001).

If Y_{obs} and Y_{mis} represent the observed and missing components of Y respectively, a less restrictive assumption than MCAR can be postulated (Rubin, 1976). That is, an assumption that affords relationships between missingness and the observed values of variables other than that in which the missingness occurs. This happens when the missingness depends only on the components Y_{obs} of Y that are observed, and not on Y_{mis} (the missing components):

$$f(\mathrm{M}|\mathrm{Y}, \varphi) = f(\mathrm{M}|\mathrm{Y}_{obs}, \varphi) \text{ for all } \mathrm{Y}_{mis}, \varphi \qquad \text{.....(2.2)}$$

This mechanism is called missing at random (MAR). There is also a heuristic definition of MAR which is described below (Schafer,1997). Let U and V be variables or non-overlapping groups of variables. If one restricts attention to units for which U is observed and equal to a constant value, *u*, MAR implies that among these units, the distribution of V is, apart from ordinary sampling variability, the same among cases for which V is observed as it is among cases for which V is missing. MAR is a weaker assumption than MCAR because it requires only that the missing values behave like a random sample of all values within subclasses defined by the observed data. That is, it allows the probability that a datum is missing to depend on the datum itself, but only indirectly through quantities that are observed. It is impossible to test whether the MAR condition is satisfied, since the values of missing data are unknown and so one cannot compare the values of those with and without missing data to see if they differ systematically on the variables (Allison, 2001).

Another mechanism, called not missing at random (NMAR), occurs when the distribution of M depends on the missing values, as well as the observed values, in the data matrix Y:

$$f(\mathrm{M}|\mathrm{Y}, \varphi) = f(\mathrm{M}|\mathrm{Y}_{obs}, \mathrm{Y}_{mis}, \varphi) \text{ for all } \mathrm{Y}_{mis}, \varphi \qquad \text{.....(2.3)}$$

To explain these mechanisms better it is useful to consider a simple data structure, such as a univariate random sample for which some units are missing (Little and Rubin, 2002). Let $\mathrm{Y} = (y_1, \ldots, y_n)^T$ where y_i represents the value of a random variable for unit I, and let $\mathrm{M} = (M_1, \ldots, M_n)$ where $M_i = 0$ for units that are observed and $M_i = 1$ for units that are missing. If the joint distribution of (y_i, M_i) is independent across units (so the probability that a unit is observed does not depend on the values of Y or M for other units), then :

$$f(Y, M|\theta, \varphi) = f(Y|\theta) f(M|Y, \varphi) = \prod_{i=1}^{n} f(y_i|\theta) \prod_{i=1}^{n} f(M_i|Y_i, \varphi) \qquad \text{.....(2.4)}$$

where $f(y_i|\theta)$ represents the density of y_i indexed by unknown parameters θ, and $f(M_i|Y_i, \varphi)$ is the probability density function of a Bernoulli distribution for the binary indicator M_i, with probability $\Pr(M_i = 1\ y_i, \varphi)$ that y_i is missing. If missingness is independent of Y, that is $\Pr(M_i = 1\ y_i, \varphi) = \varphi$, a constant that does not depend on y_i, then the missing data mechanism is MCAR (or in this case it is also MAR). If the mechanism depends on y_i the mechanism is NMAR since it depends on y_i that are missing, assuming there are some (Little and Rubin, 2002).

Missing data mechanisms can be classified as either ignorable or nonignorable. Missing data mechanisms are described as ignorable if the data are MAR and the parameters that govern the missing data process are unrelated to the parameters to be estimated (Allison, 2001). This implies there is no need to model the missing data mechanism as part of the estimation process. If the data are not MAR, one could describe the missing data mechanism as nonignorable, in which case good parameter estimates can be procured by modelling the missing data mechanism. However, the results obtained are extremely sensitive to choice of model and so good prior information is required for efficient estimates.

2.5 Methods of Dealing with Missing Data

It is possible to group methods of dealing with missing data into groups based on review papers of these methods (Gourier and Monfort, 1981 ; Little and Rubin, 1983; Schenker and Taylor, 1999; Little, 1997; Durrant, 2002; Schafer and Graham, 2002; Carter, 2006). These categories, described below, are not exhaustive and do overlap with each other through certain methodologies.

Complete case methods

When some values of variables are missing from a data set, one of the simplest methods to deal with missing data would be to completely discard incompletely recorded units, so that analysis is implemented on just the complete data (Brown and Kross, 2003). Although the procedure is simple to carry out, it could lead to serious biases when drawing inferences for subpopulations (Little and Rubin, 2002). However, there are strong proponents of these procedures, provided certain conditions are met (Allison, 2001). Allison refers to complete case analyses procedures as 'listwise deletion', implying deletion of any observations that have missing data from the sample and then applying conventional methods of analysis for complete data sets. The inherent advantages with this method is that it can be used for any kind of statistical analysis (from structural equation modelling to log-linear analysis) and no special computational methods are required. If the missing data mechanism is appropriate, for example MCAR, then complete case analysis boasts attractive statistical properties, since the reduced sample will be a random subsample of the original sample. This implies that, for any

parameter of interest, if the estimates would be unbiased for the full data set with no missing data , they will also be unbiased for the listwise deleted dataset.

The standard errors and test statistics obtained with the listwise deleted dataset will be just as appropriate as they would have been in the full data set. Following on from the notation previously introduced, this concept can be illustrated with respect to the univariate random sample data structure introduced above. If *r* is the number of complete (observed) units ($M_i = 0$), it follows that the sample size reduces from n to *r*. If we assume that the values are normally distributed and wish to make inferences about the mean, we could estimate the mean by the sample mean of the responding units with standard error $s/\sqrt{r}$, where *s* is the sample standard deviation of the responding units. This strategy is only valid if the mechanism is MCAR, since the observed cases are a random sub-sample of all the cases (Little and Rubin, 2002). The standard errors obtained would be higher because less information is used to derive them, and hence there would be loss of power. In general, it appears that listwise deletion is not robust to violations of the MCAR assumption.

It is interesting to note that listwise deletion is robust to violations of MCAR among independent variables in a regression analysis. If the probability of missing data on any of the independent variables does not depend on the values of the dependent variables, listwise deletion produces unbiased regression estimates. Since disproportionate stratified sampling on the independent variables in a regression model does not bias coefficient estimates, and a missing data mechanism that depends only on the values of the independent variables is equivalent to stratified sampling, the resulting coefficient estimates are unbiased too. This concept applies to Cox regression, logistic regression and Poisson regression in addition to linear regression (Allison, 2001).

Weighting methods

Randomization inferences from sample survey data without nonresponse commonly weigh sampled units by their design weights (Little and Rubin, 2002). These weights are inversely proportional to their probabilities of selection. For example, if y_i is the value of variable Y for unit *i* in the population, the population mean is estimated by the Horvitz-Thompson (1952) estimator:

$$(\sum_{i=1}^{n} \pi_i^{-1} y_i)(\sum_{i=1}^{n} \pi_i^{-1})^{-1}, \qquad \text{.....(2.5)}$$

where the sums are over sampled units, and π_i is the known probability of inclusion in the sample weight for unit i. Weighting procedures for nonresponse modify the weights in an attempt to adjust for nonresponse as if it were part of the sample design (Little and Rubin, 2002). The estimator then becomes:

$$\frac{\sum_{i=1}^{n}(\pi_i \hat{p}_i)^{-1} y_i}{\sum_{i=1}^{n}(\pi_i \hat{p}_i)^{-1}}, \qquad \text{.....(2.6)}$$

where the sums are now over sampled units that respond and p_i is an estimate for probability of response for unit *i* (the proportion of responding units in subclass of the sample).

Imputation

Imputation is a generic term for filling in missing data with plausible values (Schafer, 1997) and subsequently enable standard analysis of data. There are many imputation methods that be can applied to missing data. These methodologies are described in more detail in the next section (2.6).

Procedures based on models

There exist a broad class of Bayesian procedures generated by defining models for the observed data and basing inferences on the likelihood or posterior distribution under the model, with parameters estimated using methods such a maximum likelihood. These methods afford flexibility, structure due to underlying assumptions of models that are evaluated, and the availability of estimates of variance that take into account the incomplete nature of the datasets (Little and Rubin, 2002).

2.6 Focus on Imputation

General Overview of Imputation

Imputation is a method to fill in missing data with plausible values to produce a complete data set (Durrant, 2002). It involves creating a predictive distribution for the imputation based on the observed data. This distribution could be generated using explicit and implicit modelling (Little and Rubin, 2002). Explicit modelling entails a predictive distribution that is based on a formal statistical model (for example, multivariate normal), thus the assumptions are explicit. Examples of imputation by this method include mean imputation, regression imputation and stochastic regression imputation. On the other hand, implicit modelling involves an algorithm to generate the predictive distribution. This implies an underlying model and even though the assumptions are implicit, they still need to be carefully assessed to ensure that they are reasonable. Examples of imputation by this method include hot-deck imputation, substitution and cold-deck imputation. Imputation methods may also be classified according to whether they are deterministic or stochastic (Durrant, 2002). Deterministic methods produce the same imputed value for units with the same characteristics, whereas stochastic methods produce different values. Imputation makes use of a number of auxiliary variables that are statistically related to the variable in which item non-response occurs by means of an imputation model (Lessler and Kalsbeck, 1992). A more prominent classification of imputation methods is into single imputation methods and multiple imputation methods. These are described below.

Single and Multiple Imputations

Single imputation involves replacing missing values with single appropriate values from an imputation method, whereas multiple imputation involves the provision of multiple random replacements for each missing value to enable quantification of variance (Durrant, 2002). That is, m completed data sets are created, each of which is analysed in turn and then combined to produce overall summary statistics (Rubin,1987). If D is the variance associated with the parameter θ, the estimates from the mth dataset are denoted $\hat{\theta}^m = \theta (Y_{obs,} Y_{mis}^m)$ and $\hat{D}^m = D (Y_{obs,} Y_{mis}^m)$, m= 1,…, M. Applying Rubin's (1987) rules results in the

following combined multiple imputation point estimate of θ. By averaging those from all the imputed data sets :

$$\hat{\theta} = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta} \qquad \text{…..(2.7)}$$

Applying similar logic (Rubin, 1987) , the mean of the complete-data point variance estimates, called the within-imputation variance could be calculated as:

$$\hat{D} = \frac{1}{M}\sum_{m=1}^{M}\hat{D} \qquad \text{…..(2.8)}$$

Between-imputation variance is the variance estimate of the complete-data point estimates and is calculated as:

$$\hat{B} = \frac{1}{M-1}\sum_{m=1}^{M}\left(\theta^{m} \hat{} - \theta\right)^{2} \qquad \text{…..(2.9)}$$

Combining both variance estimates above by including an adjustment term (1 + 1/M) for finite M, gives the overall variance estimate associated with θ. as:

$$\hat{T} = \hat{D} + \left(1 + \frac{1}{M}\right)\hat{B} \qquad \text{….(2.10)}$$

The above formula provides an approximately unbiased estimator of the variance (Durrant, 2002). Multiple imputation enables analysis of micro-data files. Imputation models need to be chosen with respect to the analysis that would be carried out on the resulting imputed data sets. They need to incorporate any associations among variables, for example interactions (Schafer and Olsen, 1998) and need to preserve the structure of the dataset (for example,

maintenance of a hierarchical multilevel structure (Carpenter and Goldstein, 2005)). A minimum of two datasets are required to perform multiple imputation (Schafer, 1997).

Case for Imputation

Imputation is mainly carried out to reduce nonresponse bias, which occurs because the distribution of missing values generally differs from the distribution of observed values (Durrant, 2002). Imputation enables the recreation of a balanced design so that the methods used for analysing the complete data can be applied in many situations. It affords higher efficiency than case deletion since sample sizes are maintained. It also results in high precision since use of observed auxiliary information for cases with nonresponse is made (Schafer and Graham, 2002). However, imputation does have drawbacks if the imputed values are treated as real values. Imputation could increase bias in a dataset (for example, if the relationship between known and unknown variables is poor) (Kalton,1983). Thus special adjustment methods are required to correct for the increase in variability due to both nonresponse and imputation.

Choice of Imputation Method

Imputation methods must be chosen on the basis of the target of the study. Single imputation methods function by finding point estimators for missing values based on the imputation method used. For example, if unconditional means were imputed, the point estimator would be the mean of the complete cases. The standard variance estimation for a point estimator valid for complete data may lead to severe underestimation of true variance if applied to observed and imputed data (Rao and Shao, 1992). Although standard variance estimation would not suffice in the presence of imputation, other methods such as multiple imputation and jackknife variance estimators attempt to deal with this problem of low variability (Shao and Sitter, 1996). Some important imputation methods are considered below.

2.7 Imputation Methods

2.7.1 Single Imputation Methods

Deductive and Mean Imputation Methods

The simplest methods of performing imputation include deductive methods and mean imputations. Deductive methods impute missing values by using logical relations between variables and derive values for the missing item with high probability (GSS, 1996), whereas an unconditional mean imputation method involves imputing the overall mean of a numeric variable for each missing item within that variable. The latter could involve imputing class means, if classes are defined on some explanatory variables (Durrant, 2002). Despite the obvious virtues of simplicity in implementation, these methods lead to distortion of the relationships between variables since the distribution of the variables are compressed (Lessler and Kalsbeck,1992). One needs to consider these demerits before applying such methods to solve a missing data problem.

Regression Methods

Predictive regression imputation (deterministic regression imputation) involves use of one or more auxiliary variables, where there are known values for auxiliary variables in observations where the variable analysed has missing values . The regression model involves fitting y to auxiliary variables x_i. This is referred to as the imputation model. Predicted values from this model are then used to impute the missing values in y. Linear regression is usually used for numerical variables, and logistic regression is used for binary variables. Regression methods also distort the shape of the distribution of the variable being imputed and the correlation between variables (Kalton, 1983). However, random regression methods could circumvent this problem by maintaining the variable distribution (Nordholt, 1998). This involves adding an error term to the regression process. Random regression imputation involves imputing a value for the variable y of interest as a random draw from the conditional distribution of y given x. A residual term is added to the predicted value from the regression, which reflects the uncertainty in the predicted value. The residual term is obtained by drawing from a normal distribution, or by computing the regression residuals from the complete cases and selecting an observed residual at random for each nonrespondent (Durrant, 2002). Regression methods use a lot of information (such as numerous categorical and numeric variables) to produce predictions which could be very accurate. However, this imputed value is still a

predicted value (unlike hotdecking described below) and is sensitive to model misspecification of the regression model (Schenker and Taylor, 1996). The predictive power of a regression model could potentially be poor (Little and Rubin, 2002).

Hot deck Methods

Hot deck methods refer to imputation methods that assign the value from a record within an observed item, referred to as the donor, to a record with a missing value on that item (Little, 1986). There are many ways of imputing under this method. The simplest approach is to impute for each missing item the response of a randomly selected case for the variable of interest (Durrant, 2002). The imputed values emanate from observations with common variables or characteristics as the missing observation. Since these variables enable donation of real values into the missing variables, they are also known as donor or matching variables. Donor categories could thus be created, and donor values could then be selected randomly from these categories. The merits of hot deck methods include that the values imputed are real values, rather than predicted values. In addition, the methods are usually semi or non-parametric, thus distributional assumptions (and their respective potential violations) are not a concern. The imputed values maintain a similar distributional shape to the observed data (Rubin, 1987). Hot deck methods are effective with large sample sizes (Durrant, 2002).

Nearest-Neighbour Imputation

This is a deterministic donor method (described above) in which the donor is selected on the basis of a minimal distance, which is a function of the auxiliary variables (Kalton, 1983). The observed unit with the smallest distance to the nonrespondent unit (from a selected auxiliary variable) is identified and its value is substituted for the missing item according to the variable of interest (Durrant, 2002). This method shares the merits of donor methods in general. That is, actual observed values are used for imputation. In addition, it could introduce geographic effects if the cases were ordered geographically (Durrant, 2002). The nearest-neighbour approach is known to estimate missing distributions correctly (Chen and Shao, 2000). In the event of repeated donor values being used for imputation, penalties could be imposed on using a donor value several times (Kalton, 1983). This curbs any resulting inflation in variance.

Predictive Mean Matching Imputation

This is basically a deterministic hot deck donor method that incorporates regression models to facilitate donor class selection, thus rendering it semi-parametric. That is, donor variables are selected as the covariates from regressing the variable with missing values on the complete auxiliary variables. The classes could be randomly selected, thus incorporating a random component (error term) to the method. Because of the semi-parametric nature of this method, it is less sensitive (and thus more robust) to underlying model misspecifications than other model-oriented approaches (for example, regression imputations) (Schenker and Taylor, 1996). Predictive mean matching imputations are usually applied in large surveys that have missing data for relevant variables. For example, this method was used to impute SHCS income data in 2002 by regressing the relevant income variables on complete auxiliary variables to generate matching variables to carry out the imputation.

EM Algorithm for Data with Missing Values

The EM algorithm can aid single (or multiple imputations) by finding the necessary parameters based on the incomplete data set. Essentially, the EM algorithm (Dempster et al, 1977) is a technique that finds maximum likelihood estimates in parametric models for incomplete data. The EM algorithm is an iterative procedure that finds the maximum likelihood estimator (MLE) of the parameter vector by repeating the expectation and maximisation steps:

1. The expectation E-step involves calculation of the conditional expectation of the complete-data loglikelihood, given the observed data and parameter estimates.

2. The maximization M-step involves deduction of parameter estimates to maximise the complete-data loglikelihood from the expectation step.

Both steps are iterated until the iterations converge.

2.7.2 Multiple Imputation Methods

It is possible for all the single imputation methods described above to be repeated multiple times and combined for inference (Rubin, 1987), thus making them multiple imputation

methods. Some of these methods are described again below for elaboration (for example, regression methods). Other methods include specialised algorithms and models that are strictly multiple imputation methods. Some of these methods are also described below.

Regression Method

This method is best suited to missing data with a monotone pattern (described above) and assumes that the data is from a multivariate normal distribution. In the regression method, a regression model is fitted for each variable with missing values, with the previous variables as covariates (Durrant,2002). Based on the fitted regression coefficients, a new regression model is simulated from the posterior predictive distribution of the parameters and is used to impute the missing values for each variable (Rubin, 1987). The process is repeated sequentially for variables with missing values.

Propensity Score Method

Like regression methods, this also is best suited to monotone missing data patterns. The propensity score method was originally designed for a randomized experiment with repeated measures on the response variables. The goal was to impute the missing values on the response variables. The method uses only the covariate information that is associated with whether the imputed variable values are missing. It does not use correlations among variables. A propensity score is the conditional probability of assignment to a particular treatment given a vector of observed covariates (Rosenbaum and Rubin, 1983). In the propensity score method, for each variable with missing values, a propensity score is generated for each observation to estimate the probability that the observation is missing. The observations are then grouped based on these propensity scores, and an approximate Bayesian bootstrap imputation (Rubin 1987, p. 124) is applied to each group (Lavori, Dawson, and Shera 1995).

MCMC Method

This applies to arbitrary missing data patterns and assumes that the data is from a multivariate normal distribution. The Markov Chain Monte Carlo (MCMC) method originated in physics as a tool for exploring equilibrium distributions of interacting molecules. In statistics, it is

used to generate pseudo-random draws from multidimensional and otherwise intractable probability distributions via Markov chains. Markov chains are processes describing trajectories where successive quantities are described probabilistically according to the values of their immediate predecessors. In many cases, these processes tend to equilibrium and limiting quantities follow an invariant distribution (Gamerman, 1997). In MCMC simulation, one constructs a Markov chain long enough for the distribution of the elements to stabilize to a stationary distribution, which is the distribution of interest. By repeatedly simulating steps of the chain, the method simulates draws from the distribution of interest (Schafer, 1997). MCMC has been applied as a method for exploring posterior distributions in Bayesian inference. That is, through MCMC, one can simulate the entire joint posterior distribution of the unknown quantities and obtain simulation-based estimates of posterior parameters that are of interest. In many incomplete data problems, the observed-data posterior $p(Y_{obs})$ is intractable and cannot be easily simulated. However, when Y_{obs} is augmented by an estimated/simulated value of the missing data Y_{mis}, the complete-data posterior $p(Y_{obs}, Y_{mis})$ is much easier to simulate. Critics may regard the use of a prior distribution as subjective and artificial. It can alternatively be viewed as a "necessary evil" (Schafer and Graham, 2002). Moreover, in some problems prior distributions can be formulated to reflect a state of relative ignorance about the parameters, thus mitigating the effect of subjective inputs. The influence of the prior diminishes as the sample size increases. Since multiple imputation relies on large sample approximations for the complete-data distribution, the prior rarely exerts a major influence on the results. This fact challenges many of the drawbacks postulated with parametric stochastic models (Allison, 2001).

2.8 Software for Imputation

It is essential to briefly describe the types of software available, since they ascertain whether many of the procedures described above can be implemented in practice, if at all. SPSS[1] has a missing value analysis procedure which performs basic imputations using EM methods and regression. It is a limited procedure and excludes multiple imputation techniques and other basic automated imputation methods like hot deck imputation. STATA[2] incorporates hot

[1] http://www.spss.com/ accessed : 3 September 2007
[2] http://www.stata.com/ accessed: 3 September 2007

deck imputation methods based on approximate Bayesian bootstrap, regression imputations and multiple imputation methods. It enables combination of results from the latter using Rubin's rules (Rubin, 1987). It does not include data augmentation methods. R[3] incorporates hot deck, regression methods and parametric model-based procedures. If assumptions of parametric models are not met, these methods are note executed. SAS[4] possesses the procedures PROC MI and PROCMIANALYZE to perform multiple imputations. SAS allows multiple imputation via three methods: regression, propensity score and MCMC (all described above). SOLAS[5] performs mean imputation, hot deck imputation and regression imputation. Multiple imputation can be carried out using regression or propensity score methods. SOLAS performs imputations on monotone data only. This is a limitation of SOLAS. There also exist many free software to implement multiple imputations such as NORM, CAT, MIX[6] and IVEware[7]. There are many specialised software which impute for specific purposes. An example of such a software is CANCEIS (Bankier et al.,2001). This was developed by Statistics Canada in ASCII C language to impute missing data for the Census. It employs the nearest neighbour imputation methodology described earlier to impute values so that the complete data set retains as many attributes as possible as the original dataset and closely resembles it. It differs from the Fellegi/Holt methodology (which provides theory for nearest neighbour imputation methods) by searching for nearest-neighbour donors first and then determines the minimum change imputation action, rather than identifying the minimum number of variables to impute first (Fellegi and Holt, 1976). This has numerous computational advantages (Bankier et al., 2001).

2.9 Applied Methods

There exist a variety of imputation methods that have been applied on large important surveys worldwide. CANCEIS (described above), a software that applies the nearest neighbour imputation methodology, is used in national household surveys and censuses. For

[3] http://www.r-project.org/ accessed: 3 September 2007
[4] http://www.sas.com/ accessed: 3 September 2007
[5] SOLAS is a specialised software produced for dealing with missing data, produced by Statistical Solutions, http://www.statssol.ie/html/dolas/solas_home.html accesed: 3 September 2007
[6] These three software are based on Schafer's (1997) algorithms and are available freely on the internet, http:/www.stat.rtu.edu~jls/misoftwa.html accessed: 3 September 2007
[7] This software was produced by the University of Michigan and is available on the internet, http://www.irv.umich.edu/src/swp/ive/ accessed: 3 September 2007

example, it was used in the 2006 Canadian census, and was distributed for national census analysis to UK, Italy, Brazil, USA, New Zealand and Australia (Bankier, 2006). The Office of National Statistics (ONS), which currently imputes SHCS missing income data, also used CANCEIS for data editing and imputation. Semi-parametric hot deck regression methods using SAS have been implemented in other household surveys (Beissel and Skinner, 2002). The National Centre for Social Research used this method to impute the SHCS missing income data prior to the ONS methodology. Other methods applied have included using EUROMOD, an algorithm that exploits the tax and contribution rules to convert net income information into gross amounts for European countries (Immervoll et al., 2003). Multiple imputation methods have also been applied to impute data in large surveys with SAS as the software medium. For example, the Household Survey of Healthcare for Communities 1997/8 (Tang et al., 2001) uses SAS to impute missing data.

2.10 Conclusion

The merits of imputation methods to deal with missing data range from enabling complete-data analysis to enhancing precision of estimates. The need to deal with missing data in a calculated manner is immense, especially since many data sets include missing values that are not (completely) missing at random. However, contingent on the imputation method employed, imputed values need to be treated with caution. In the words of Rubin, the "seductive" yet "dangerous" process may not be the panacea we consider it to be (Rubin, 1983).

CHAPTER 3

METHODOLOGY

3.1 Data Preparation

The first SHCS dataset (2003/4) that was imputed by ONS was used for analysis for this study. The dataset contained 3870 observations and 3894 variables. These comprised mainly social and demographic variables. Another dataset containing derived variables was present for the same year. Income is a derived variable, made up of components of the social variables. Both datasets were merged (using merging or 'set' commands in SAS and 'match-filing' in SPSS (programming)). This enabled a holistic view of both the derived and social variables. The dataset included original component income variables (for example: employment status of respondents, hours worked , etc.) and imputed versions of these. However the derived income variable contained both original and imputed values in the data column. The missing values on this (post-imputed) income variable were analysed. The dataset with original (pre-imputed) missing derived income values was not available. In order to carry out any analysis on the data, it was imperative to have a dataset that reflected the pre-imputation situation. This would enable implementation of various imputation methodologies for comparison and would also summarise the original missing data situation. To revert to the original missing income values, it was necessary to 'work backwards' from the post-imputation to the pre-imputation situation. The relevant SPSS codes for the ONS were analysed. These were split into five subsets of codes which were all brought together to provide income values. The arduous 'working backwards' process involved replacing any redundant variables, deleting variables that did not exist (for example, din3), replacing all the imputed variables (distinguished by 'i's attached to their original names) with original variable names and specifying correct locations of files using 'Get File' statements in SPSS programming. After several iterations of recoding, the original income dataset was finally procured and a frequency table was created to check the proportion of original income missing values in the dataset. This would now enable imputation at the income level. A copy of one section of the original SPSS programming code and the modified code is enclosed in the Appendix 1.

3.2 Exploratory Analysis of Data

The high-dimensional multivariate dataset had almost 4000 variables. For the purposes of complete analysis, frequency tables (using 'proc freq' statements in SAS and 'fre' statements in SPSS programming) were produced for each variable. This gave an indication of the percentage of missing values inherent in each variable and described their distributions. Summary statistics were produced for variables where appropriate. Some of the variables were recoded for sake of better interpretation and manipulation in analysis. For example, council tax paid in pounds sterling was recoded into a grouped variable based on bands, called council tax band using 'if/then/else' statements in SAS. Some assumptions were made when recoding certain variables. For example, when recoding the council tax (payable in pounds sterling) variable into band intervals, it was assumed that these values included any amendments for benefits and exemptions the household was entitled to. This assumption was made by checking the (frequency) distribution of the council tax paid variable. The fact that the frequency split into more divisions than the eight council tax band values expected confirmed that the variable represented the amount the respondent actually paid rather than fixed band values. The missing data structure was analysed using SOLAS. It was found that the missing data pattern for the data (including income variables) was arbitrary. SOLAS was used to convert this to monotone data.

To better manage the data in terms of the income, it was sought to reduce the number of accompanying variables. Some variables were eliminated on the logical basis of their (lack of) relationships with income. For example, variables describing inter-relationships of people in house (such as relationships between the seventh person and the sixth person in the house) were eliminated because they would be unlikely to have any impact on the income variable. Correlations (using Pearson's correlation coefficient) between numeric variables and associations between categorical variables (using Chi-square tests) were carried out where it was thought that a group (or pair) of variables in the dataset were measuring the same concept. The significance of the tests determined respective inclusion or elimination. Fortunately, most of the variables that were eliminated in this way had high percentages of missing values (which would have rendered their analysis futile anyway). Variables with high missing value percentages were also eliminated if they were considered not relevant to the income variables based on logic and theory. There still remained a large chunk of variables

(with no missing values). These were checked for outliers for data validation purposes, and were respectively modelled to produce better reduction in numbers.

Two types of regression modelling, one algorithm method (EM) and a principal component analysis were implemented. These are described below. The variables in the dataset were also checked to see if they followed a multivariate normal distribution, which was a core assumption for some of the imputation methods that were implemented. This was tested by initially checking the distributions of individual variables and any interrelationships (using correlation or association measures), and then eventually creating Chi-Square graphs of squared distances. Some of the variables were transformed to check if this would improve the multivariate distribution of the dataset.

It was also of interest to investigate if the post- or pre-imputed missing data were biased in any way. This investigation of missing data mechanism in operation was carried out by creating cross-tabulations of relevant variables with whether income was missing or not.

3.3 Intermediary Modelling

The following intermediate analyses were carried out to help render the dataset more tractable and to provide the necessary parameters or covariates required for certain imputation methods applied on the dataset. These analyses are classified into the 'intermediary modelling' section because the results from such methods are not the end product of the study, but rather a bridge to enable efficient implementation of the subsequent imputation methods applied.

Principal Component Analysis

Principal component analysis (Cody and Smith, 2005) is best known for variable reduction. However, applying this using SAS in a normal fashion would be erroneous because a correlation or covariance matrix is usually required which is generated from continuous variables. Many of the variables in the dataset are categorical. Because of this, a polychoric correlation matrix was generated by calculating the polychoric correlations between these non-numeric variables (Raskin and Novacek, 1988). This was then input into the normal 'proc factor' procedure in SAS with method specified as principal component analysis

('method = prin') to generate a reduced data set for use in imputation. A scree plot was produced to substantiate selection of principal components (Cody and Smith, 2005).

EM Algorithm Imputation

Using the SAS 'proc mi' statement, the EM algorithm imputation method (Schafer,1997) was carried out to produce parameters for subsequent use in multiple imputation using the maximum likelihood estimator theory. The resulting parameter estimates for the dataset were noted (mean and covariance matrices). In writing the SAS code for the EM algorithm, dummy variables were created for all categorical variables. The creation of dummy variables enables comparison with a base dummy variable, which was chosen, and also helps distinguish between categorical and continuous variables.

Multiple Linear Regression Modelling

Multiple linear regression attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. The dependent variable (income) is represented by y_1 below and the independent variables (comprising sociodemographic variables) are represented by x_{in} below. β_i describes the coefficients of regression for each of the independent x variables. The error term associated with the model is ε_i.Formally, the model for multiple linear regression, given *n* observations, is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots \beta_p x_{ip} + \varepsilon_i \text{ for } i = 1,2, \ldots n \qquad \ldots\ldots(3.1)$$

Multiple linear regression modelling was used as an intermediate step for imputation and to further reduce variable numbers based on subsequent refinements of the model. The derived income variable (weekly household income) was modelled using 'proc reg' in SAS. Dummy variables were created for the categorical variables in the model, so that they were not misconstrued as being continuous. Multicollinearity and Variance Inflation were checked for as part of the refinement process. Those with high values were eliminated and the model

rerun. Parsimonious models were sought based on the fit statistic of the model (R-square values). Despite being employed initially, step-wise automatic selection procedures were avoided as the model got more parsimonious to preclude the modelling of noise. Residual error terms of the models were checked for normality to test compatibility with assumptions of multiple regression.

Logistic Regression Modelling

Logistic regression is used when the dependent variable is binary. In logistic regression the logarithm of the odds of the proportion is used as the dependent variable. This is known as the logit of p.

$$\log it(p) = \ln\left(\frac{p}{1-p}\right) \qquad \text{.....(3.2)}$$

This is modelled as a linear combination of the explanatory variables so that the regression equation being fitted is

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \varepsilon, \qquad \text{.....(3.3)}$$

and the error structure is assumed to be binomial.

As with multiple regression modelling, logistic regression modelling was used as an intermediate step for imputation and to further reduce variable numbers based on subsequent refinements of the model. The dependent variable in this case was whether the derived income variable (weekly household income) was missing or not. A code was written using 'if/then/else' statements in SAS to create this new dependent variable. After eliminating any highly significant correlated predictor variables, logistic regressions were carried out for the derived income variable (weekly household) using 'proc logistic' in SAS. This enabled insertion of categorical variables in the 'class' statement. Appropriate reference parameters were chosen to make the resulting odd ratios interpretable (by comparison with base values). Exponents of the coefficients were coded to calculate these odd ratios. Receiver Operator Characteristic (ROC) curves (which plot sensitivity against '1-specificity') were plotted and

low deviance values were chosen as criteria to check fit of the model in the refinement process. Sensitivity is a ratio consisting of the number of correctly classified events over the total number of events (Ott and Longnecker, 2000). Specificity is a ratio consisting of the number of correctly classified non-events over the total number of non-events (Ott and Longnecker,2000).

3.4 Derived Versus Component Income Variables

Since the purpose of the study was to impute missing income values in as simple a manner as possible, preference was given to impute at the overall derived income level rather than component income variable level. The saving in computation time afforded time to explore more methods of imputing, and maintained, if not enhanced, accuracy of the procedure. This is because if each of the component variables are imputed, their imputed (not real) values are used to finally arrive at the derived income variable. This may increase the variance associated with the values of the overall derived income variable.

3.5 Imputation Methods

With a more parsimonious and tractable data set created from the above –mentioned procedures, various imputation methods were implemented.

3.5.1 Single Imputation Methods

Method 1: Single Unconditional Mean Imputation

Despite the prima facie demerits of distortion posed by this method, it was still implemented for comparative purposes. The means of complete-case income variables were calculated using SAS (‘proc means’ procedure) , and these were imputed in the dataset to replace the respective missing values (Allison, 2001). No residuals were added when this method was applied.

Method 2: Single Simple Hot Deck Imputation

A simple non-parametric hot deck imputation code was written in SAS using the 'proc surveyselect' method (see Appendix 7). This SAS statement obviates the need of macro statements. The code was written so that derived income values were imputed through a random selection (with replacement) of donor values, based on random probability allotting from the uniform distribution. This method represented the simplest form of hot deck imputation (Carter, 2006), since there was only one large donor pool of variables. The donors were not split into subgroups.

Method 3: Single Hot Deck (Regression) Imputation

Using the final multiple regression linear regression model (intermediary model) described above, the predictor variables were entered as the hot deck donors in SOLAS for the derived income variable. That is, the covariates from the multiple regression intermediary model (described above) were used to divide the variables into subgroups according to values of the donors. Imputed values were randomly chosen from these subgroups to replace missing values on the income variable. For example if the tenure type was the only covariate and was owner-occupied, a value was selected randomly from the corresponding subgroup. SOLAS automatically checks for the presence of sufficient members in a donor class, and collapses the necessary class if it does not meet the criterion of sufficient membership, before imputing. Imputed data sets were produced based on these methods (SHCS Technical Report, 2002).

Method 4: Single Hot Deck (Logistic Regression) Imputation

Using the intermediary logistic regression model , the predictor variables were entered as the hot deck donors in SOLAS for the income variable. That is, the covariates from the logistic regression intermediary model (described above) were used to subdivide the variables into donor subgroups. Random retrieval (from donor subgroups) was selected as an option in SOLAS to replace the missing values on the income variable. SOLAS automatically checks for the presence of sufficient members in a donor class, and collapses the necessary class if it does not meet the criterion of sufficient membership, before imputing. Imputed data sets were produced based on these methods (SHCS Technical Report, 2005/6).

Method 5: Single Predicted Mean Imputation

Regression imputation was implemented in SOLAS to impute values. The covariates were selected from the multiple regression intermediary model previously created. This time regression was used as the method of imputation. That is, missing values were imputed by inserting the corresponding covariate values in the regression model for each missing income observation. This method was included as an enhancement of the unconditional mean method carried out earlier (Little and Rubin, 2002).

3.5.2 Multiple Imputation Methods

Method 6: Multiple Simple Hot Deck Imputation

To check the performance of a fractional imputation method, the simple hot deck imputation method described above (Method 2) was repeated five times by adjusting the seed values randomly in the SAS code to produce five different imputed datasets for further analysis (Durrant, 2002). Like the single simple hot deck imputation method, no donor subgroups were created for each dataset.

Method 7: Multiple Predictive Model Based (Regression) Imputation

The set of covariates in the intermediary multiple regression model was used to impute the missing values in the income variable using SOLAS. First, the predicted model was estimated from the observed data. Using this estimated model (similar to format of equation 3.1), new linear regression parameters were randomly drawn from their Bayesian posterior distribution. The randomly drawn values were used to generate the imputations, which include random deviations from the model's predictions (Carter, 2006). In the system, multiple regression estimates of parameters were obtained using the method of least squares. SOLAS automatically designs any dummy variables pertaining to nominal variables in the regression equation.

Method 8: Multiple Predictive Model Based (Propensity Score) Imputation

Propensity Score methods and Approximate Bayesian Bootstrap (Durrant, 2002) were incorporated into a model to carry out this procedure in SOLAS. Like the previous method,

the missing data pattern was made monotone, but in this method the covariates used were that from the intermediary logistic regression model. SOLAS automatically creates a missingness variable so that conditional probability of missingness, given the vector of observed covariates, can be calculated. The conditional probability of missingness is also called the propensity score. The propensity scores, generated by using the covariates of the logistic regression model, were sorted and divided into five equal sized groups. For each missing data entry of the income variable, a subset of observed values was found such that their assigned propensity scores were close to the assigned propensity score of the missing values to be imputed. The set of observed values used to generate the imputations were the observed values of the subset of cases where this missing data entry belongs. Thus, hot decking was used to impute missing values based on the donor pool.

Method 9: Multiple Markov Chain Monte Carlo Imputation

This model-based Monte Carlo Markov Chain method (Schafer,1997) was used to impute the derived income variable where it was missing by specifying a method that generated five datasets. This was done using a code with the 'proc mi' statement (see Appendix 8). Monotone conversion (even though assumed for the dataset) was not carried out here because this method is especially suited to arbitrary missing data patterns, hence being the most relevant method to apply for imputation to the multivariate dataset. The process goes through many (default of two hundred) iterations of imputation and posterior parameter-estimation steps until it converges. Time series and autocorrelation function plots were created in SAS for the derived income variable to check for convergence. In writing the SAS code for the MCMC model, dummy variables were created for all categorical variables. The creation of dummy variables enables comparison with a base dummy variable, which was chosen, and also helps distinguish between categorical and continuous variables.

3.6 Note on Disclosure/ File Conversion

Throughout this study, a standalone computer was used at The Scottish Government to carry out any SAS or SOLAS analysis. Files were converted from SPSS format into SAS files by changing them into portable format first, and then invoking the SPSS engine in SAS. Because of disclosure risks and other security precautions, the dataset used was modified into a subset by dropping all postcode/address variables (using 'drop' statement in SAS/ SPSS

programming). This dataset was then burned onto CDs to use on the standalone computer. It was ensured at all times that no data was stored on the standalone computer and all output, excluding programming code, was transferred onto CD to be printed from the network computer or to be stored on the hard drive of the network computer. This was an essential negotiated protocol agreed so as not to compromise in anyway the interests or stakes in the sensitive nature of privileged information present in the dataset.

3.7 Summary of Analysis

Wherever possible, the mean and standard deviation for each imputation method was calculated (using 'proc means' in SAS) to summarise the imputation performed. However, to put all methods in a comparative context, the new income dataset was categorised into council tax bands (which was one the variables produced from recoding the council tax variable) for each method, to check whether the income values generated made any sense. Therefore, we would expect the income values to increase as the council tax band level elevated (from A to H).[8]

[8] It should be noted that the council tax band itself contained missing values. These were not imputed so accuracy of comparison with income values could be maintained. The observations with these missing values were deleted from the data set when the income-value check, describe above, was carried out. This shrunk the imputed income value percentage from 6.8% to 5.4%, which implied that 80% of the imputed values could still be checked.

CHAPTER 4

RESULTS I : DATA DESCRIPTION AND QUALITATIVE ANALYSIS OF PREVIOUS IMPUTATION METHODS

4.1 Descriptive Overview of Variables

The original 2003/4 SHCS dataset (of social variables), described in the Methodology section, contained 3870 observations and 3894 variables. These comprised identification variables, household size (number of people in household), sex of all the members in the household, ages of each member in the household, the interrelationships of all the members in the household, their marital/partner status, tenure type, the number of children they have (if any) and their respective attributes (age, schooling, etc.), the number of dependents (including children) on the householder and their attributes, ethnicity of householders, the physical geographical features of home/facilities and surroundings (for example, number of bathrooms, security features, central heating features, if there is a post office in the area, opinions on neighbourhood, needs of householders, etc.), the working/wealth status of the householders and other working attributes (for example whether self employed or unemployed, hours worked, if they owned holiday homes, savings, etc.). Other variables the survey measures include benefits the householders receive (if any), the expenses of the household (for example fuel costs), the sanitary conditions of the household (for example, whether any of the rooms have mould), the health status of the householders (for example, disabilities, if any of the children wheeze, etc.), the property value of the household and the benefits status of the householders(if they receive benefits/grants, if they are pensioners, etc.). Based on this, it can be appreciated how these variables would help reflect the fuel poverty status of Scotland.

4.2 Missing Values

The values of the variables described above represented the post-imputed dataset. That is, it included imputed values of all variables that were imputed. Despite imputation, there was a

(circa six percent) proportion of missing values for the derived income variable. Table 4.1 reflects the situation.

Table 4.1: Post-imputation Missing Values For Derived Income Variable

Household income band

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	< £100 p.w.	246	6.4	6.8	6.8
	£100 -199.99 p.w.	778	20.1	21.4	28.1
	£200 -299.99 p.w.	736	19.0	20.2	48.3
	£300 -399.99 p.w.	526	13.6	14.4	62.8
	£400 -499.99 p.w.	458	11.8	12.6	75.3
	£500 -699.99 p.w.	529	13.7	14.5	89.9
	£700+	369	9.5	10.1	100.0
	Total	3642	94.1	100.0	
Missing	System	228	5.9		
Total		3870	100.0		

Table 4.1 above shows that 5.9% of the income values were still missing after imputation.

As described in the Methodology section, it was imperative to convert the dataset (after merging with the derived variables) to the pre-imputation situation for the derived income variable for further analysis. This enabled analysis of the missing values of the original dataset (before any imputation was carried out). The refined code (described in the Methodology section) produced the original dataset correctly. Table 4.2 describes the missing values of this dataset (produced from the SPSS programming code).

Table 4.2: Pre-imputation Missing Values For Derived Income Variable

Household income band

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	< £100 p.w.	389	10.1	10.8	10.8
	£100 -199.99 p.w.	932	24.1	25.8	36.6
	£200 -299.99 p.w.	643	16.6	17.8	54.4
	£300 -399.99 p.w.	435	11.2	12.1	66.5
	£400 -499.99 p.w.	417	10.8	11.6	78.1
	£500 -699.99 p.w.	465	12.0	12.9	91.0
	£700+	326	8.4	9.0	100.0
	Total	3607	93.2	100.0	
Missing	System	263	6.8		
Total		3870	100.0		

Table 4.2 shows that the original income missing values comprised 6.8% of the dataset. This table gives an indication of the efficiency of the imputed method that had been applied. That is, the missing value percentage (considered in terms of the whole dataset) fell by 1%.

4.3 Missing Data Pattern

Drawing from the literature review, one can now appreciate the missing pattern of the variables of this rectified dataset. Because frequency tables reveal that almost sixty percent of the total number of variables in the dataset contain missing values, this dataset has an arbitrary missing pattern. However, since variable reduction procedures were implemented (some results of which are in the next section) so that complete variables (apart from income variables) remained, and for the convenience of the monotone analyses that were carried out, the dataset was converted into a monotone pattern (using SAS and SOLAS) procedures. The fact that many variables have missing values together with the unordered pattern of missing data substantiates that the missing data pattern is arbitrary. However, by presuming that only the income variable has missing values and making the necessary modifications for it, the application of procedures (described in the Methodology section) can render the pattern monotone.

4.4 Analysis of Variables

Each variable was analysed by creating frequency tables. This showed how many values, if any, were missing. The presence of outliers was checked for. Summary statistics, when applicable, were investigated. For example, it was found that the mean age of the highest income householder was 53 years old (with a standard deviation of 17) and that of his/her partner was 50 years old with a standard deviation of 14 approximately. The mean number of adults in the household was 2 with a standard deviation of approximately 1. The mean number of children in the household was close to zero (0.03). The average council tax paid by individuals in the population was £1033.32 with a standard deviation of £558.44. Table 4.3 exemplifies how correlations between numeric variables were investigated. It shows that weekly total household income and council tax paid per year have a positive correlation coefficient that is significant at the 1% significance level. Table 4.4 depicts how associations between categorical or nominal variables were investigated using Chi-Square tests. It shows that a person having children in the household is associated with the number of babies in the

household, meaning those with no babies in the household are not likely to have children in the household. Figure 4.1 shows how the distributions of each variable were investigated. It reveals that the age of the highest income householder deviates from the normal distribution. Figure 4.2 describes the results of investigating multivariate normality (which was an assumption for some of the imputation models used) of the dataset as a whole after checking for correlations/associations between variables and individual distributions. It shows the Chi-Square plot of squared distances ,and reveals some deviation from the normal line. Tests of normality (for example, Kolmogorov-Smirnov) for many variables showed that they were not normally distributed. Figure 4.3 shows that logarithmic transformations of variables resulted in a similar Chi-Square plot. The plots in both Figures 4.2 and 4.3 indicated that the distribution of the dataset did deviate from multivariate normality. However, since many of the models subsequently applied would be robust to deviations from multivariate normality, we could assume that the distribution of the dataset was multivariate normal.

Table 4.3: Correlation Between Total Household Income and Council Tax

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations		
	WKHHINC	**CTAX**
WKHHINC Weekly total household income (pounds)	1.00 3607.00	0.42 <.0001 3261.00
CTAX Council tax paid per year (£)	0.42 <.0001 3261.00	1.00 3448.00

Table 4.4: Association Between Person Having Children and Number of Babies in Household

Frequency Percent	**Table of NBABY by HASCHD**			
	NBABY(NUMBER OF CHILDREN UNDER 1)	**HASCHD(PERSON HAS CHILDREN IN HOUSEHOLD)**		**Total**
		Yes	**No**	
	No	1426	2377	3803
		36.85	61.42	98.27
	Yes	67	0	67
		1.73	0	1.73
	Total	1493	2377	3870
		38.58	61.42	100
	Statistic	**DF**	**Value**	**Prob**
	Chi Square	1	108.55	<0.0001

Figure 4.1: Distribution of Age of Highest Income Householder

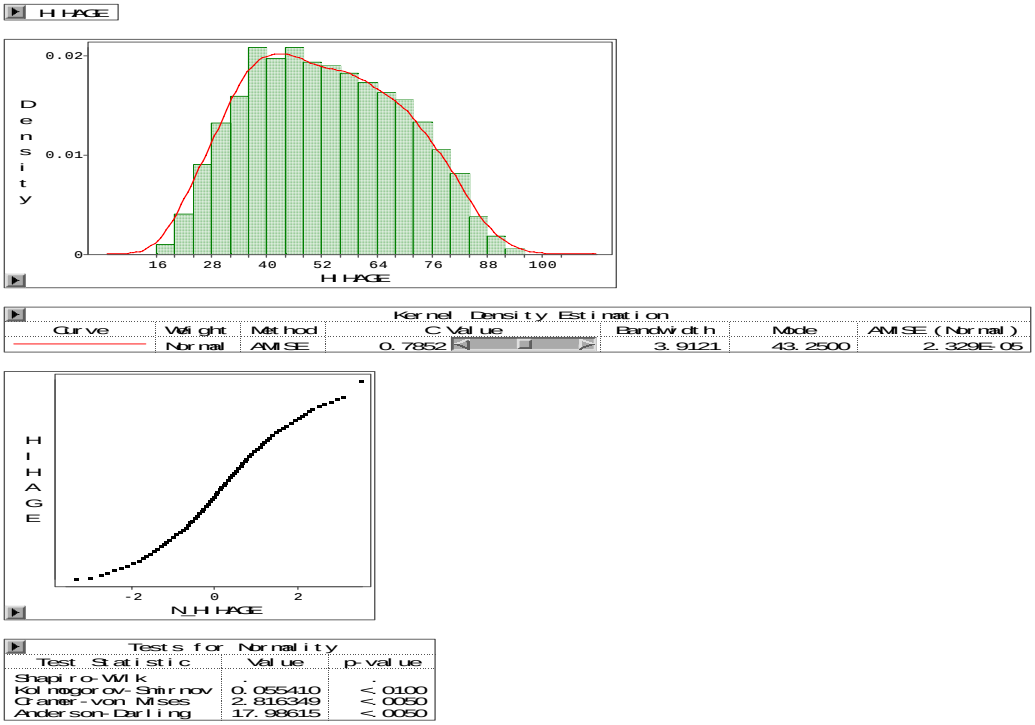


Figure 4.2: Chi-square plot of squared distances of original variables

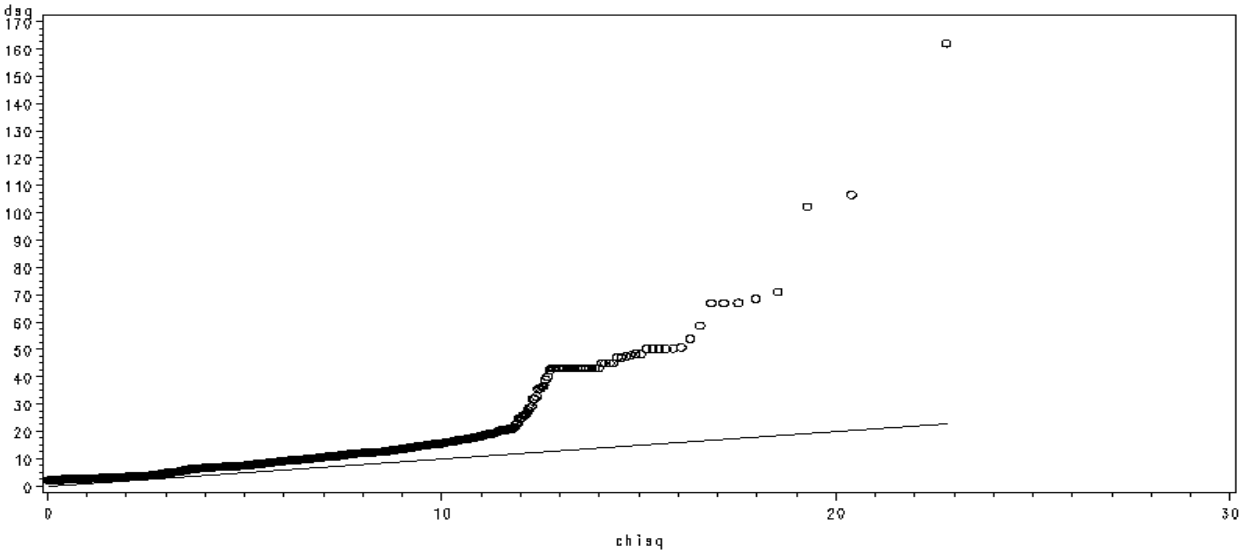
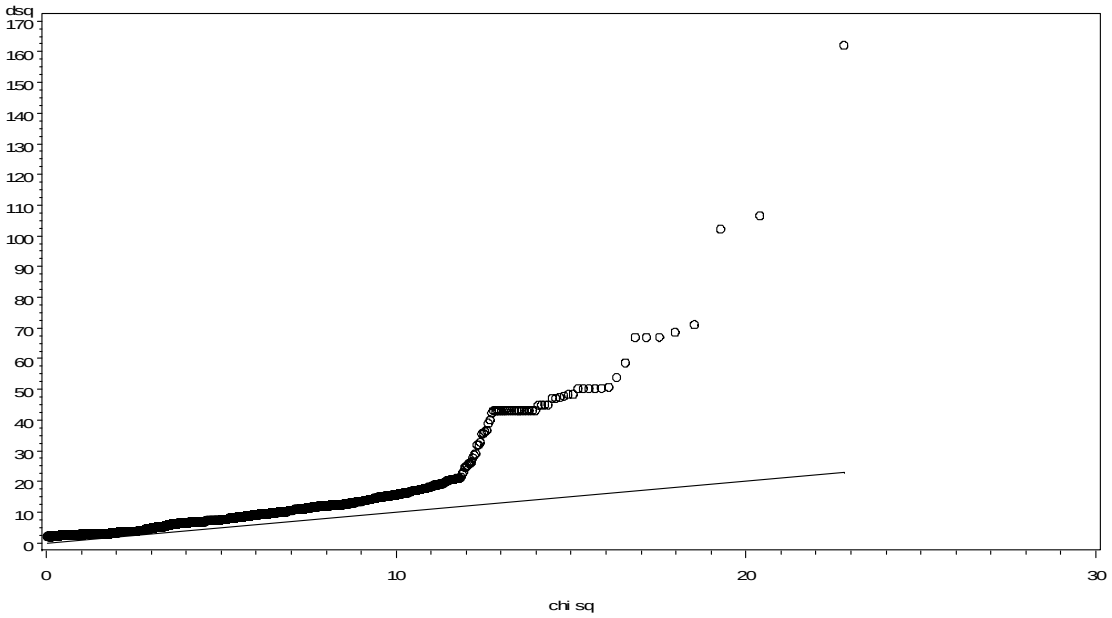


Figure 4.3: Chi-square plot of squared distances of transformed variables



4.5 Missing Data Mechanisms

It is of interest to find out if the post- or pre-imputed missing data were biased in any way or if they in fact followed a random pattern. This is an essential data-description step and helps identify the missing data mechanism operating in this data set. To investigate this, cross-tabulations of relevant variables with whether the income was missing or not were constructed and analysed for association using the Chi-Square test. For example, Table 4.5 describes the analysis for the number of adults in the household. Because the post-imputed missing data set was actually a subset of the pre-imputed data set, both gave similar results. It was found that there was a significant association between the number of adults living in the household and missingness of income, where majority of the missing income data comprised households with just one or two adults.

Similar analysis showed that the missing data mostly had no children in the household. There was also a significant association between tenure and income missingness, with majority of the households with missing income data being owner-occupied. Even

though the proportions of males and females with the missing data was approximately equal, the cross-tabulations still showed a significant association between sex of highest income householder and missingness of income. That is, according to the Chi-square test there was a higher proportion of males in the missing dataset than females. The majority of the missing data population fell in the older age brackets, with the range 65-74 for both highest income householder and his/her partner being the most populated. The missing data comprised mainly of households with no pensionable age householders. Most of the households in the missing data set were single pensioner and single adult households. Most of these households also had no children under the age of five, and contained more people aged over sixty. Most of the missing data households did have a long-term sick/disabled person in the household.

It was found that all of these associations were significant at the 1% significance level. Details of all the cross-tabulations for both pre- and post-imputed missing data are enclosed in Appendices 9 and 10 respectively. This implies that the missingness was definitely not missing completely at random, and thus was assumed as missing at random. With respect to this one can appreciate why simplistic methods like listwise deletion would not be appropriate as an imputation method.

Table 4.5: Cross-tabulation of Number of Adults in Household with Income (Missing or Not)

Frequency Col Pct	Table of M by NUMADULT							
	M	NUMADULT(NUMBER OF ADULTS IN HOUSEHOLD)						Total
		1	2	3	4	5	6	
	income	1217	1824	415	127	18	6	3607
		88.64	95.1	98.34	95.49	100	100	
	missing income	156	94	7	6	0	0	263
		11.36	4.9	1.66	4.51	0	0	
	Total	1373	1918	422	133	18	6	3870

4.6 Qualitative Analysis: Description of Previous Imputation Methods Used on SHCS Data

Prior to the current group that handles missing income data imputation for the SHCS, income imputation was carried out (until 2002) by National Centre for Social Research according to a

set of requirements specified by Communities Scotland. Hot deck imputation was used, and the relevant characteristics or donor classes were chosen using regression analysis. Very large and very small values were excluded from the imputation classes. This imputation method was quite successful (there was a negligible percentage of missing values after imputation).

From 2003 onwards, the Office of National Statistics (ONS) was in charge of imputing income values using the Canadian Census Edit and Imputation System (CANCEIS), which was developed to perform minimum change nearest neighbour imputation. CANCEIS (Bankier et al., 2001) is a generic system written in ANSI C and has the functionality to work with a variety of data types in which the user supplies his/her own data rules and requirements. Design of CANCEIS is based on Nearest Neighbour Imputation Methodology (NIM). This implies that CANCEIS still used the hot deck method to impute, but the principle was based on minimising the distance of donors to the missing values so that the distribution of the data set was maintained. The variables that comprise the calculation of income (including other income variables such as benefits) were chosen to be imputed for both Highest Income Householder (HIH) and his/her partner (see Appendix 12). Donor variables were, however, chosen by modelling the missingness using a logistic regression model to find a common group of matching variables for the variables chosen to be imputed (see Appendix 12). After carrying out the imputation, it was found that missing income data still persisted. Reasons attributed to this could range from the absence of full donor classes, to rigidity of constraints applied in CANCEIS (Bankier, 2001) or even a break down in the algorithm process. The absence of full donor classes was investigated by constructing frequency tables of the matching variables to check if they contained any missing values. Tables 4.6 and 4.7 describe the frequency tables of relevant matching variables that were created.

Table 4.6: Frequency table describing whether in receipt of household income or not matching variable

RECEIVING HOUSING BENEFIT

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	YES	671	17.3	55.1	55.1
	NO	546	14.1	44.9	100.0
	Total	1217	31.4	100.0	
Missing	9	4	.1		
	System	2649	68.4		
	Total	2653	68.6		
Total		3870	100.0		

Table 4.6 above shows that the 'receiving housing benefit' matching variable contained 68.4% of missing values, and Table 4.7 below shows 'working tax credit' proxy matching variable contained 92.6% of missing values. Unless these values were coded as system-missing , this could pose problematic. It would imply that a large percentage of the donor values were missing and so would not have been included when subdividing into donor groups to perform hot deck imputation.

In addition to the missing values problem described above, the exact methodology that CANCEIS was involved in was not known. The exact constraints were unknown , and hence the extent of their rigidity could not be ascertained. Similarly, details of logistic regression models performed to determine matching variables were brief.

Table 4.7: Frequency table describing 'period: working tax credit' variable as a proxy for 'whether receiving working tax credit or not'

Period: working tax credit

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	One week	101	2.6	35.4	35.4
	Two weeks	28	.7	9.8	45.3
	Four weeks	39	1.0	13.7	58.9
	Calendar month	112	2.9	39.3	98.2
	One year/12 months/52 weeks	4	.1	1.4	99.6
	None of these	1	.0	.4	100.0
	Total	285	7.4	100.0	
Missing	System	3585	92.6		
Total		3870	100.0		

The imputation procedure carried out on the 2003/4 could be described as a predominantly automated one (using CANCEIS). Reasons attributed to the missingness could be maintained as application of rigid constraints and insufficient donor members for some matching variables (described above).

CHAPTER 5

RESULTS II: INTERMEDIARY ANALYSES AND IMPUTATION RESULTS

5.1 Intermediary Analysis: Principal Component Analysis

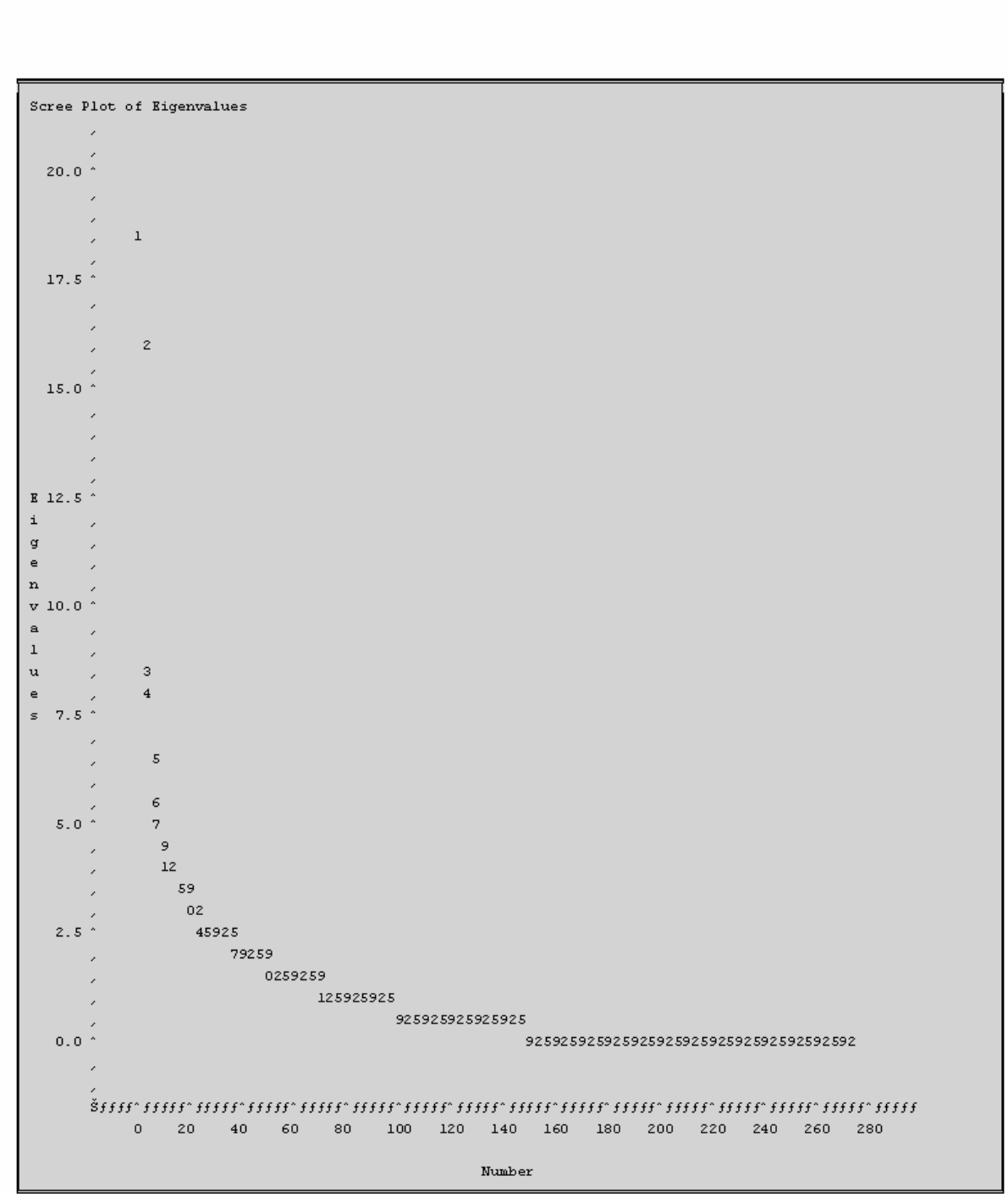
An extensive polychoric correlation matrix of the multivariate dataset was successfully created. This has not been included here because the dimensions ran into thousands. The results of using this matrix to perform a principal component analysis of the data, yielded retention of eighty two factors by the eigenvalue criterion (being greater than one). Table 5.1 below provides a snapshot of the eigenvalues of the factors (first fifteen), and the respective variance explained by each factor. As observed below, the variance explained by each factor is not very high. That is, each factor explains a very small portion of the variance.

Table 5.1: Eigenvalues of Factors

Eigenvalues of the Correlation Matrix : Total = 273 Average = 1				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
1	18.26	2.13	0.07	0.07
2	16.13	7.77	0.06	0.13
3	8.36	0.58	0.03	0.16
4	7.78	1.33	0.03	0.19
5	6.45	0.94	0.02	0.21
6	5.51	0.44	0.02	0.23
7	5.07	0.22	0.02	0.25
8	4.85	0.30	0.02	0.27
9	4.54	0.12	0.02	0.29
10	4.42	0.23	0.02	0.31
11	4.20	0.15	0.01	0.32
12	4.04	0.22	0.01	0.33
13	3.82	0.06	0.01	0.34
14	3.76	0.20	0.01	0.35
15	3.57	0.11	0.01	0.36

The scree plot associated with the analysis is described in Figure 5.1 below. The points on the scree plot overlap.

Figure 5.1: Scree Plot of Eigenvalues



There was no obvious identification of factors in terms of grouping (or explaining) similar variables. That is, for the purpose of this analysis, the factors (generated from using the principal components method in SAS) did not have applied implications.

5.2 Intermediary Analysis: EM Algorithm

A non-informative prior was used in the EM algorithm. The algorithm procedure also produced missing data patterns. Table 5.2 below lists the missing data patterns with corresponding frequencies and percents for the final variables retained in the model. In Table 5.3, "X" implies that the variable is observed in the corresponding group and "." means that the variable is missing. The table shows that 218 of the income values were missing.

Table 5.2: Missing Data Patterns

Missing Data Patterns										
Group	**Weekly Household Income**	**Tenure : LA/Other Public Sector**	**Tenure: Housing Association/ Co-ops**	**Tenure: Private Rented**	**Tenure: Joint Users**	**Household Has Children: No**	**Employer Pension: No**	**Investment Income: No**	**Frequency**	**Percent**
1	*X*	*X*	*X*	*X*	*X*	*X*	*X*	*X*	3076	93.38
2	.	*X*	*X*	*X*	*X*	*X*	*X*	*X*	218	6.62

Table 5.3 describes the initial parameter estimates for EM. That is, it shows the mean and covariance matrix for the variables based on the complete case data.

Table 5.3: Initial Parameter Estimates for EM

	Weekly Household Income	**Tenure : LA/Other Public Sector**	**Tenure: Housing Association/ Co-ops**	**Tenure: Private Rented**	**Tenure : Joint Users**	**Household Has Children: No**	**Employer Pension: Yes**	**Investment Income: Yes**
MEAN	344.631	0.180	0.090	0.070	0.002	0.618	0.211	0.112
COVARIANCE MATRIX								
Weekly Household Income	103485	0	0	0	0	0	0	0
Tenure : LA/Other Public Sector	0	0.147	0	0	0	0	0	0
Tenure: Housing Association/ Co-ops	0	0	0.082	0	0	0	0	0
Tenure: Private Rented	0	0	0	0.065	0	0	0	0
Tenure:Joint Users	0	0	0	0	0.002	0	0	0
Household Has Children:No	0	0	0	0	0	0.236	0	0
Employer Pension:Yes	0	0	0	0	0	0	0.167	0
Investment Income:Yes	0	0	0	0	0	0	0	0.099

Table 5.4 below displays the 'Iteration History' for the EM algorithm. This shows convergence of the model. Table 5.5 displays the maximum likelihood estimates (mean and covariance matrix) of a multivariate distribution from the dataset. That is, it shows the parameter estimates after the EM algorithm is applied.

Figure 5.4: EM (MLE) Iteration History

EM (MLE) Iteration History		
Iteration	**-2 Log L**	**Weekly Household Income**
0	6036.87	344.63
1	4832.92	344.63
2	4828.31	340.71
3	4828.23	340.04
4	4828.22	339.93
5	4828.22	339.91
6	4828.22	339.91
7	4828.22	339.91
8	4828.22	339.91

Table 5.5: EM (MLE) Parameter Estimates

	Weekly Household Income	**Tenure : LA/Other Public Sector**	**Tenure: Housing Association/ Co-ops**	**Tenure: Private Rented**	**Tenure: Joint Users**	**Household Has Children: No**	**Employer Pension: Yes**	**Investment Income: Yes**
MEAN	339.9096	0.1797	0.0899	0.0695	0.0015	0.6178	0.2113	0.1117
	COVARIANCE MATRIX							
Weekly Household Income	103415	-20.4085	-9.5446	-2.8935	-0.0532	-37.6794	-32.6663	11.6613
Tenure : LA/Other Public Sector	-20.4085	0.1474	-0.0162	-0.0125	-0.0003	-0.0033	-0.0079	-0.0152
Tenure: Housing Association/ Co-ops	-9.5446	-0.0162	0.0818	-0.0062	-0.0001	-0.0003	-0.0062	-0.0091
Tenure: Private Rented	-2.8935	-0.0125	-0.0062	0.0647	-0.0001	0.0065	-0.0056	-0.0029
Tenure: Joint Users	-0.0532	-0.0003	-0.0001	-0.0001	0.0015	-0.0003	-0.0003	-0.0002
Household Has Children: No	-37.6794	-0.0033	-0.0003	0.0065	-0.0003	0.2361	0.0571	0.0108
Employer Pension: Yes	-32.6663	-0.0079	-0.0062	-0.0056	-0.0003	0.0571	0.1666	0.0131
Investment Income: Yes	11.6613	-0.0152	-0.0091	-0.0029	-0.0002	0.0108	0.0131	0.0992

5.3 Intermediary Analysis : Multiple Linear Regression Modelling

Table 5.6 below shows the ANOVA table from modelling the derived income variable (weekly household income) with relevant independent variables. The table shows that the independent variables reliably predict weekly household income (at the 1% significance level).

Table 5.6: ANOVA Table

Number of Observations Read		3870			
Number of Observations Used		3607			
Number of Observations with Missing Values		263			
Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	12	109451664	9120972	133.87	<.0001
Error	3594	244865667	68132		
Corrected Total	3606	354317332			

Table 5.7 below shows that the independent variables predict 30.89% of the variance in weekly household income.

Table 5.7: Overall Model Fit

Root MSE	261.02	**R-Square**	0.31
Dependent Mean	345.10	**Adj R-Sq**	0.31
Coeff Var	75.64		

Table 5.8 below shows the parameter estimates for the independent variables in the model.

Table 5.8: Parameter Estimates for Regression Model

Variable	**DF**	**Parameter Estimate**	**Stan-dard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**	**95% Confidence Limits**	
Intercept	**1**	454.84	37.35	12.18	<.0001	0	381.61	528.06
Employer Pension: Yes	**1**	-58.24	12.35	-4.72	<.0001	1.36	-82.44	-34.04
Investment Income: Yes	**1**	118.49	13.63	8.69	<.0001	1.04	91.77	145.21
Whether Working in Reference Week : No	**1**	-155.28	10.75	-14.44	<.0001	1.49	-176.36	-134.20
Working in Reference Week: No Information	**1**	-153.86	261.18	-0.59	0.5558	1.00	-665.94	358.22
Partner Age Group: 25-34	**1**	94.29	40.36	2.34	0.0195	6.31	15.16	173.41
Partner Age Group: 35-44	**1**	114.86	39.02	2.94	0.0033	9.99	38.36	191.36
Partner Age Group: 45-54	**1**	61.58	39.19	1.57	0.1162	9.31	-15.25	138.42
Partner Age Group: 55-64	**1**	-50.73	39.69	-1.28	0.2013	8.67	-128.54	27.09
Partner Age Group: 65-74	**1**	-114.01	41.63	-2.74	0.0062	6.23	-195.64	-32.38
Partner Age Group: 75-84	**1**	-132.22	46.52	-2.84	0.0045	3.12	-223.42	-41.03
Partner Age Group: 85+	**1**	-200.44	123.14	-1.63	0.1037	1.11	-441.88	40.99
Partner Age Group: No Partner	**1**	-146.44	38.15	-3.84	0.0001	18.63	-221.24	-71.65

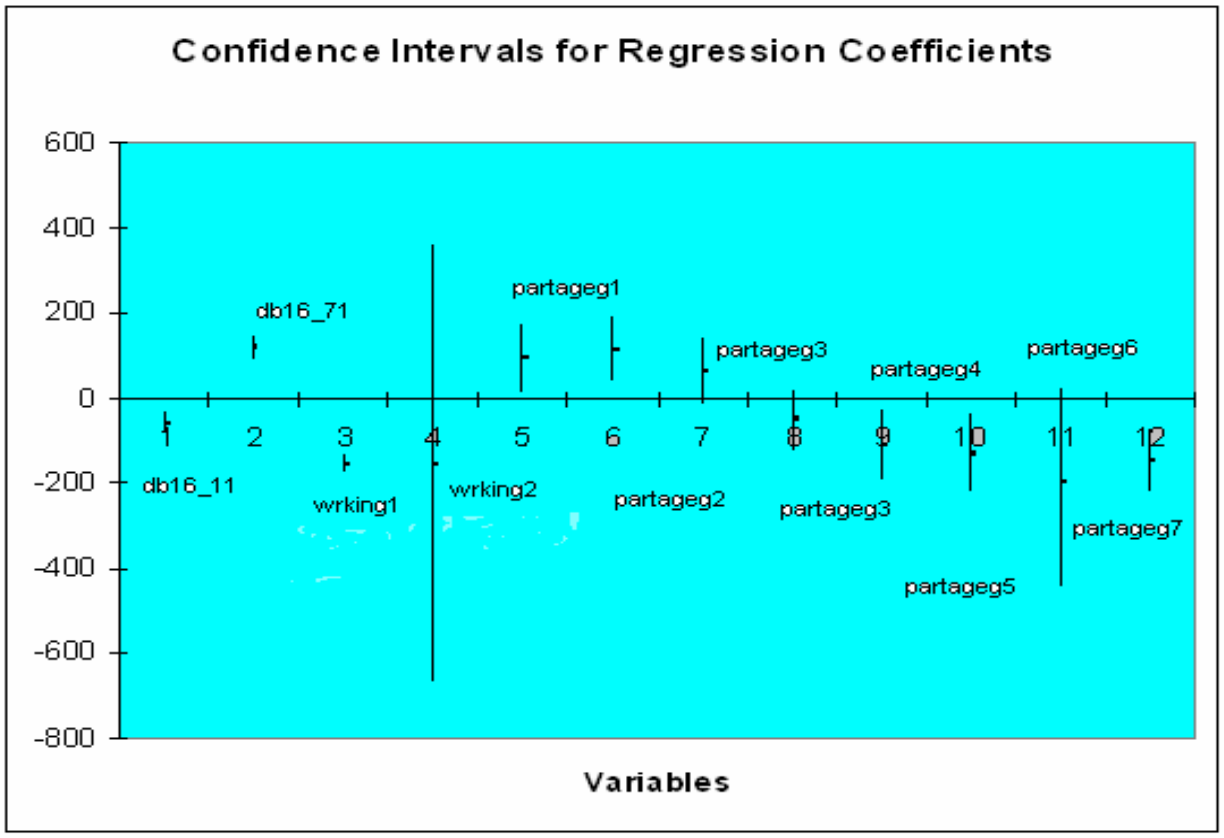
Table 5.8 shows that the variable 'Partner Age Group: No Partner' has the highest variance inflation. The parameter estimates can be used to predict the income model. Therefore, the model for this is:

Weekly Household Income = 454.84 -58.24***(Employer Pension: Yes)** + 118.49***(Investment Income: Yes)** -155.28***(Whether Working in Reference Week : No)** -153.86***(Working in Reference Week: No Information)** +94.29***(Partner Age Group: 25-34)** +114.86***(Partner Age Group: 35-44)** + 61.58***(Partner Age Group: 45-54)** -50.73***(Partner Age Group:55-64)** -114.01***(Partner Age Group: 65-74)** -132.22***(Partner Age Group: 75-84)** -200.44***(Partner Age Group:85+)** -146.44***(Partner Age Group:No Partner)**

The t-tests from the table for the parameters show that the most of the coefficients differ from zero significantly (at the 5% level, albeit some are significant at higher significance levels). For the occupier/pension variable, the income is £58 more for those with pensions than for those with no pensions. On the other hand, for the investment income variable, the investment income is £118 higher for those respondents with investment incomes than for those with none. For those working in the reference week, those who were not working had incomes lower by £155 and those who provided no information had incomes lower by £153 than those who were working. For those with partners in the age group 25-34, for every unit (year) increase in age, there was £94 increase in predicted income. Similarly for those aged 35-44, there was £114 units increase in predicted income for every year increase in age. For those aged 45-54, there was a £62 increase for every unit increase in age. For those aged in the higher age groups, for example, those in the age range 55-74, there was a £114 decrease in predicted income for every unit (year) increase in age. This falling trend continued for the higher age groups. For those with no partners, and hence no partner age information, there was a £146 decrease in predicted income for every unit increase.

A plot of the confidence intervals of the coefficients of parameters in this model is provided in figure 5.2 .

Figure 5.2: Confidence Intervals Plot



The large confidence intervals in the plots above for wrking2 (no information about working in reference week) and partageg6 (those in the age group of 85 years and above) imply that the sample size or population of these variables may have been quite small.

5.4 Intermediary Analysis :Logistic Regression Modelling

Table 5.9 below describes the convergence status. The model convergence status shows that the relative gradient convergence criterion was satisfied, thus the maximum likelihood algorithm converged.

Table 5.9 : Model Convergence Status

Model Convergence Status		
Convergence criterion (GCONV=1E-8) satisfied.		
Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	1924.05	1746.59
SC	1930.31	1821.73
-2 Log L	1922.05	1722.59

Table 5.10 below confirms the validity or significance of the logistic regression model. The likelihood ratio shows that none of the predictor values are equal to zero, because the Chi-square test is significant at the 1% significance level.

Table 5.10 : Significance of Model

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	199.45	11	<.0001
Score	212.18	11	<.0001
Wald	176.46	11	<.0001
Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
Tenure	4	89.23	<.0001
Has Child	1	50.11	<.0001
Number of Householders	3	6.28	0.0985
Respondent Person Number	3	22.90	<.0001

Table 5.11 reveals binary regression estimates for the parameters in the model. The logistic regression model models the log odds of a positive response (probability modelled is M (missingness) =1) as a linear combination the predictor variables. This can be written as **Log [p / (1-p)] = -3.96 + 1.15*(Tenure: LA/Other Public) + 1.60*(Tenure: Housing Association) + 0.33*(Tenure: Private Rented) -13.39*(Joint Owners) + 1.24*(Household Has Children: No) – 0.40 *(Number of Householders: 2) + 3.59*(Person Number of Respondent: 3) + 3.08*(Person Number of Respondent: 4)**

Table 5.11 : Model Parameters

Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-3.96	0.20	383.85	<.0001
Tenure	**LA/Public Sector**	1	1.15	0.17	47.75	<.0001
Tenure	**Housing Association**	1	1.60	0.19	74.24	<.0001
Tenure	**Private Rented**	1	0.33	0.28	1.41	0.24
Tenure	**Joint Owners**	1	-13.39	1915.10	0.00	0.99
Household Has Children	**No**	1	1.24	0.17	50.11	<.0001
Number of Householders	**2**	1	-0.40	0.17	5.56	0.02
Number of Householders	**3**	1	0.60	0.95	0.40	0.53
Number of Householders	**4**	1	-11.84	397.50	0.00	0.98
Respondent Person Number	**2**	1	0.07	0.23	0.10	0.75
Respondent Person Number	**3**	1	3.59	0.86	17.44	<.0001
Respondent Person Number	**4**	1	3.08	1.27	5.86	0.02

The positive coefficient for the tenure (LA/Public sector) variable suggests that the odds of having a missing income value for those who have houses provided by the public sector or local authority is (approximately three times) higher than for those who live in owner occupied houses. Similarly, the positive coefficient of the tenure (housing association) variable indicates that the odds of having a missing income value for those living in housing associations is (about five times) greater than for those living in owner occupied houses. The same applies to the tenure (private rented) variable, where the odds of having a missing income value is (about two times) higher for those living in private rented homes is higher than for those living in their own homes. However, the negative parameter coefficient of the tenure (joint owners) variable shows that the odds of joint owners having a missing income value is lower than for those living in single owner occupied homes.

The positive coefficient of the 'household has children: no' variable means that the odds of having a missing income value for those with no children is about four times higher than those with children.

The negative parameter coefficient of the 'number of householders: 2' variable means that the odds of having a missing income value for those with two householders is about a third less than those of households with one householder. The positive coefficient of the 'number of householders: 3' variable shows that the odds of a missing income value occurring in households with three members is about twice that of households with one member. However, it should be noted that this coefficient is not significant at the 5% significance level. The negative coefficient of householders with four members means that the odds of missing income values occurring for those households are lower than for those with one household. However, it should be noted that this coefficient is not significant at the 5% significance level.

The positive coefficient of the 'person number of respondent: 2' means that the odds of income missing when the respondent being the second person in the household is two times that of the respondent being the first person in the household. Similarly if the respondent is the third person in the household the odds of having missing income values is thirty six times higher than that of the respondent being the first person in the household. The same applies if the respondent is the fourth person in the household. This time the odds of missing income are twenty two times that of the first person in the household.

The odds ratios together with the respective confidence intervals are summarised in Table 5.12 below.

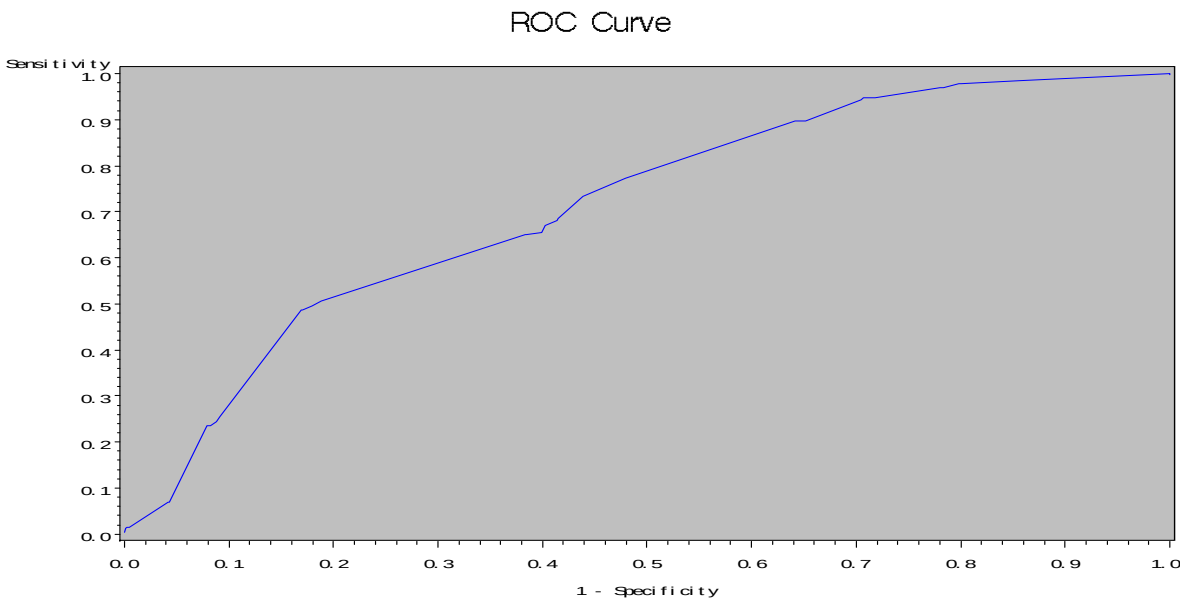
Table 5.12 : Odds Ratios and Confidence Intervals

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
'Tenure: LA/Other Public' vs 'Tenure: Owner Occupied'	3.16	2.28	4.38
'Tenure: Housing Association' vs 'Tenure: Owner Occupied'	4.97	3.45	7.16
'Tenure: Private Rented' vs 'Tenure: Owner Occupied'	1.39	0.81	2.41
'Tenure: Joint Owners' vs 'Tenure: Owner Occupied'	<0.001	<0.001	>999.999
'Household Has Children: No' vs 'Household Has Children:Yes'	3.45	2.45	4.86
'Number of Householders: 2' vs 'Number of Householders: 1'	0.67	0.48	0.94
'Number of Householders: 3' vs 'Number of Householders: 1'	1.82	0.28	11.76
'Number of Householders: 4' vs 'Number of Householders: 1'	<0.001	<0.001	>999.999
'Person Number of Respondent: 2' vs 'Person Number of Respondent: 1'	1.08	0.68	1.70
'Person Number of Respondent: 3' vs 'Person Number of Respondent: 1'	36.24	6.72	195.45
'Person Number of Respondent: 4' vs 'Person Number of Respondent: 1'	21.69	1.80	261.62

The table above shows that the variables describing joint ownership (tenure) and four householders have wide confidence intervals, implying that they may have few values or members.

The ROC curve plotted is displayed in Figure 5.3 together with the table of associated probabilities and observed responses. The area under the ROC curve was found to be 0.74 approximately, which is reasonably close to 1.

Figure 5.3: ROC Curve



5.5 IMPUTATION RESULTS

5.5.1 Single Imputation Methods: Results

Method 1: Single Unconditional Mean Imputation

Summary income data after single unconditional mean imputation are presented in Table 5.13, together with comparative summary statistics of the complete case income data.

Table 5.13: Mean Income Data By Council Tax Band After Mean Imputation

	Overall Post-Imputation Mean	Standard Deviation	Complete Case Mean	Complete Case Standard Deviation
	344.63	300.64	344.63	311.35
Council Tax Band	**Post-Imputation Mean**	**Standard Deviation**	**Complete Case Mean**	**Complete Case Standard Deviation**
A	341.00	248.78	340.72	258.14
B	328.12	236.19	327.14	243.10
C	350.15	257.45	350.63	268.60
D	337.57	251.52	337.11	259.54
E	348.65	445.26	348.90	458.78
F	353.57	250.50	354.21	259.30
G	339.70	265.57	339.26	277.28
H	354.74	355.41	355.33	365.67

The table above shows that the mean values per council tax band are quite close to those of the complete case mean values. However, the standard deviations of the mean are lower than the complete case standard deviations of the means. There appears to be a linear trend of income increasing through council tax bands A to H, as observed from the table, that is in conformity with the pattern depicted by the complete case means.

Method 2: Single Simple Hot Deck Imputation

Table 5.14 describes the summary income data with comparative complete case statistics after single simple hot deck imputation was carried out on the dataset.

Table 5.14: Mean Income Data By Council Tax Band After Hot Deck Imputation

	Post-Imputation Overall Mean	Standard Deviation	Complete Case Mean	Complete Case Standard Deviation
	342.04	308.23	344.63	311.35
Council Tax Band	**Post-Imputation Mean**	**Standard Deviation**	**Complete Case Mean**	**Complete Case Standard Deviation**
A	339.21	259.20	340.72	258.14
B	324.19	238.69	327.14	243.10
C	342.81	263.85	350.63	268.60
D	336.28	256.04	337.11	259.54
E	351.00	449.88	348.90	458.78
F	363.35	262.41	354.21	259.30
G	342.68	279.87	339.26	277.28
H	353.16	359.29	355.33	365.67

As with the first imputation method, table 5.14 above demonstrates that the mean values per council tax band are quite close to those of the complete cases mean values. However, the standard deviations of the means are lower than the complete case standard deviations of the means. In comparison with Method 1, Method 2 yields a lower overall mean and a higher overall standard deviation. Table 5.14 reveals a general linear trend of income increasing through council tax bands A to H, as is the case with the complete case means.

Method 3: Single Hot Deck (Regression) Imputation

The income data after single semi-parametric regression hot deck imputation are summarised in Table 5.15, together with comparative summary statistics of the complete case income data as with the previous methods. As with the Method 1, Table 5.15 shows that the mean values per council tax band are close to those of the complete cases mean values. However, the standard deviations of the mean are again lower than the complete case standard deviations of the mean. Table 5.15 shows that the overall mean income for this method is smaller than both Methods 1 and 2. The standard deviation of the overall mean is smaller too when compared with Methods 1 and 2. There appears to be an approximate linear trend of income increasing through council tax bands A to H. This conforms with the pattern of the complete case data.

Table 5.15: Mean Income Data By Council Tax Band After Regression Hot Deck Imputation

	Post-Imputation Overall Mean	Standard Deviation	Complete Case Mean	Complete Case Standard Deviation
	336.40	305.68	344.63	311.35
Council Tax Band	**Post-Imputation Mean**	**Standard Deviation**	**Complete Case Mean**	**Complete Case Standard Deviation**
A	333.64	254.34	340.72	258.14
B	318.74	241.95	327.14	243.10
C	341.00	262.92	350.63	268.60
D	329.41	257.55	337.11	259.54
E	340.57	447.51	348.90	458.78
F	348.13	258.25	354.21	259.30
G	325.92	272.72	339.26	277.28
H	347.24	357.62	355.33	365.67

Method 4: Single Hot Deck (Logistic Regression) Imputation

The income data after single semi-parametric logistic regression hot deck imputation are summarised in Table 5.16, together with summary statistics of the complete case income data.

Table 5.16: Mean Income Data By Council Tax Band After Logistic Regression Hot Deck Imputation

	Overall Post-Imputation Mean	Standard Deviation	Complete Case Mean	Complete Case Standard Deviation
	338.09	305.80	344.63	311.35
Council Tax Band	**Post-Imputation Mean**	**Standard Deviation**	**Complete Case Mean**	**Complete Case Standard Deviation**
A	333.68	256.60	340.72	258.14
B	322.86	239.47	327.14	243.10
C	337.23	262.61	350.63	268.60
D	332.66	255.57	337.11	259.54
E	345.20	450.12	348.90	458.78
F	346.82	255.34	354.21	259.30
G	329.64	270.22	339.26	277.28
H	350.01	357.72	355.33	365.67

Table 5.16 reveals results similar to those of Method 3. The overall mean and standard deviation values are slightly higher than that from Method 3. The means and standard deviations of the resulting values are lower than those from the complete case data, and the

(approximately linear) trend of income values across council tax bands A to H is similar to that of the complete case data.

Method 5: Single Predicted Mean Imputation

The SOLAS regression equation used for imputing missing values under this methods was : Weekly Household Income = 794.45 – 35.67*(Partner Age Group) – 160.90*(Whether Working in Reference Week) – 88.42*(Pension Income) + 119.83*(Investment Income)

Table 5.17 depicts the mean income results and comparative complete case statistics after single predicted mean imputation was performed on the data set. It reveals that the overall mean and standard deviations of income data after imputation are smaller than those for the complete case data. The overall mean is as small as Methods 3 and 4, but the standard deviation is slightly smaller than both Methods 3 and 4. The approximate increasing income value trend across council tax bands A to H is similar to that of the complete case data.

Table 5.17: Mean Income Data By Council Tax Band After Predicted Mean Imputation

	Overall Post-Imputation Mean	**Standard Deviation**	**Complete Case Mean**	**Complete Case Standard Deviation**
	336.63	304.07	344.63	311.35
Council Tax Band	**Post-Imputation Mean**	**Standard Deviation**	**Complete Case Mean**	**Complete Case Standard Deviation**
A	332.59	252.93	340.72	258.14
B	318.47	239.98	327.14	243.10
C	340.96	262.17	350.63	268.60
D	331.59	254.87	337.11	259.54
E	342.30	446.93	348.90	458.78
F	346.79	254.18	354.21	259.30
G	326.11	270.77	339.26	277.28
H	347.44	357.45	355.33	365.67

5.5.2 Multiple Imputation Methods : Results

Method 6: Multiple Simple Hot Deck Imputation

One of the five income datasets generated after repeated simple hot deck imputation is summarised in Table 5.18 below, with comparative summary statistics of the complete case income data.

Table 5.18: Mean Income Data By Council Tax Band After Multiple Hot Deck Imputation For First Dataset

	Overall Post-Imputation Mean	Standard Deviation	Complete Case Mean	Complete Case Standard Deviation
	341.67	294.59	344.63	311.35
Council Tax Band	**Mean**	**Standard Deviation**	**Complete Case Mean**	**Complete Case Standard Deviation**
A	342.65	310.72	340.72	258.14
B	338.91	256.28	327.14	243.10
C	336.88	250.39	350.63	268.60
D	346.50	268.69	337.11	259.54
E	333.43	255.21	348.90	458.78
F	349.53	458.94	354.21	259.30
G	346.59	256.00	339.26	277.28
H	344.28	277.47	355.33	365.67

Table 5.18 shows that for one of the datasets generated, the overall mean is slightly lower than that of the complete case mean income data. The standard deviation of this mean is much smaller than the complete case data, and is the smallest of all the other imputation methods. In addition, there is a linear trend of increasing mean income across council tax bands but it deviates from the pattern of the complete case data. For example, the mean income increases when moving from band C to D for the imputed data, whereas that for the complete case data increases for the same council tax bands.

Table 5.19 shows the combined summary statistics of all the five datasets generated from the imputation method (including the overall variance associated with multiple imputation).

Table 5.19: Mean Income Data By Council Tax Band After Multiple Hot Deck Imputation For Combined Datasets

Post-Imputation Overall Mean		Standard Deviation		Complete Case Mean		Complete Case Standard Deviation	
348.83		308.36		344.63		311.35	
Council Tax Band	**Post-Imputation Mean**	**Within - Imputation Variance**	**Between-Imputation Variance**	**Overall Variance**	**Standard Deviation**	**Complete Case Mean**	**Complete Case Standard Deviation**
A	339.95	72032.82	2.36	72035.77	268.39	340.72	258.14
B	330.89	63381.21	37.85	63428.52	251.85	327.14	243.10
C	344.76	68395.25	28.58	68430.97	261.59	350.63	268.60
D	338.35	66690.50	25.40	66722.26	258.31	337.11	259.54
E	371.35	175143.88	786.54	176127.1	419.68	348.90	458.78
F	352.26	95442.98	39.32	95492.14	309.02	354.21	259.30
G	340.94	73730.81	13.59	73747.8	271.57	339.26	277.28
H	354.10	118593.90	35.54	118638.3	344.44	355.33	365.67

Unlike the first dataset, Table 5.19 shows that the overall mean generated is higher than that of the complete case mean. The standard deviation is also higher than that of the first dataset, but lower than the complete case data. The linear pattern is similar to that of the complete case data. In addition, the combined results show the overall variance of the imputation process per council tax band, since this is a multiple imputation procedure.

Method 7: Multiple Predictive Model Based (Regression) Imputation

One of the five income datasets (third) generated after multiple regression imputation is summarised in Table 5.20 below, as was done for Method 6. Table 5.20 shows the results from the third dataset generated as part of the multiple imputation procedure. The overall mean appears to be lower than the complete case mean. On the other hand, the standard deviation is higher than that of the complete case mean. The increasing trend across council tax bands is apparent as with the complete case data, but the trend differs to that of the complete case data. For this method, council tax band E has the highest mean income value, whereas for the complete case data council tax band H has the highest mean income value.

Table 5.20: Mean Income Data By Council Tax Band After Multiple Regression Imputation For Third Dataset

Overall Post-Imputation Mean	Standard Deviation	Complete Case Mean	Complete Case Standard Deviation	Overall Post-Imputation Mean
335.20	314.56	344.63	311.35	335.20
Council Tax Band	**Post-Imputation Mean**	**Standard Deviation**	**Complete Case Mean**	**Complete Case Standard Deviation**
A	330.43	263.19	340.72	258.14
B	315.56	254.11	327.14	243.10
C	340.61	271.95	350.63	268.60
D	328.36	273.99	337.11	259.54
E	345.60	449.01	348.90	458.78
F	349.12	266.95	354.21	259.30
G	315.41	289.04	339.26	277.28
H	342.94	364.15	355.33	365.67

As with Method 6, the combined results for all the datasets generated for Method 7 are depicted in Table 5.21.

Table 5.21: Mean Income Data By Council Tax Band After Multiple Regression Imputation For Combined Datasets

Post-Imputation Overall Mean		Standard Deviation		Complete Case Mean		Complete Case Standard Deviation	
346.43		313.93		344.63		311.35	
Council Tax Band	**Post-Imputation Mean**	**Within - Imputation Variance**	**Between-Imputation Variance**	**Overall Variance**	**Standard Deviation**	**Complete Case Mean**	**Complete Case Standard Deviation**
A	331.95	69098.86	2.48	69101.95	262.87	340.72	258.14
B	317.38	62709.15	10.91	62722.79	250.45	327.14	243.10
C	343.64	75132.15	12.98	75148.38	274.13	350.63	268.60
D	328.67	71044.06	6.08	71051.66	266.56	337.11	259.54
E	391.13	203477.66	2896.54	207098.3	455.08	348.90	458.78
F	350.29	68788.19	21.46	68815.02	262.33	354.21	259.30
G	321.53	81584.89	24.95	81616.08	285.69	339.26	277.28
H	346.14	131900.30	6.09	131907.9	363.19	355.33	365.67

Table 5.21 reveals that for the combined datasets, the mean income was actually larger than that of the complete case mean. The overall standard deviation was higher than that of the complete case mean as with the third dataset. Similarly, council tax band E has the highest

mean income value for the imputed data, whereas council tax band H has the highest value for the complete case data. This confirms that the trend of income across council tax bands differs from that of the complete case data. Table 5.21 also provides the overall variance per council tax band of the imputation procedure, since this is a multiple imputation procedure.

Method 8: Multiple Predictive Model Based (Propensity Score) Imputation

One (the second) of the five income datasets generated after multiple propensity score imputation is summarised in Table 5.22 below, with comparative summary statistics of the complete case income data.

Table 5.22: Mean Income Data By Council Tax Band After Multiple Propensity Score Imputation For Second Dataset

Overall Mean	Standard Deviation	Complete Case Mean	Complete Case Standard Deviation	Overall Mean
339.79	306.40	344.63	311.35	339.79
Council Tax Band	**Mean**	**Standard Deviation**	**Complete Case Mean**	**Complete Case Standard Deviation**
A	335.11	257.40	340.72	258.14
B	319.33	240.47	327.14	243.10
C	348.92	264.39	350.63	268.60
D	329.16	255.75	337.11	259.54
E	346.21	448.75	348.90	458.78
F	352.21	257.00	354.21	259.30
G	332.42	272.12	339.26	277.28
H	348.81	358.95	355.33	365.67

Table 5.22 shows that the overall mean for the second data set is smaller than that of the complete case data. The standard deviation is also lower than that of the complete case mean. The linear pattern of mean income approximately increasing with council tax band differs from that of the complete case mean. Council tax band F appears to have the highest mean income for imputed data, whereas council tax band H has the highest mean income for the complete case data.

Table 5.23 shows the combined results of all five datasets under propensity score multiple imputation. It shows the within and between variances (to give overall variance) associated with the multiple imputation procedure per council tax band.

Table 5.23: Mean Income Data By Council Tax Band After Multiple Propensity Score Imputation For Combined Datasets

Overall Mean		Standard Deviation		Complete Case Mean		Complete Case Standard Deviation	
348.83		307.00		344.63		311.35	
Council Tax Band	**Mean**	**Within - Imputation Variance**	**Between-Imputation Variance**	**Overall Variance**	**Standard Deviation**	**Complete Case Mean**	**Complete Case Standard Deviation**
A	339.95	65148.36	2.36	65154.42	255.25	340.72	258.14
B	330.89	58773.25	37.85	58781.41	242.45	327.14	243.10
C	344.76	70207.62	28.58	70235.15	265.02	350.63	268.60
D	338.35	66137.06	25.40	66155.93	257.21	337.11	259.54
E	371.35	200899.91	786.54	204207.7	451.89	348.90	458.78
F	352.26	65807.87	39.32	65812.28	256.54	354.21	259.30
G	340.94	74369.33	13.59	74386.73	272.74	339.26	277.28
H	354.10	129591.41	35.54	129598.3	360.00	355.33	365.67

The combined results shown in the table above conforms with the results of the second dataset describe in Table 5.22, with the exception of the overall mean income being higher than the complete case mean. The standard deviation remains lower than the complete case mean and the linear pattern differs from the complete case mean, with council tax band E containing the highest mean income band this time (instead of council tax band H as in the complete case data). In this respect the mean income pattern is similar to that of Method 7.

Method 9: Multiple Markov Chain Monte Carlo Imputation

Table 5.24 shows the model information from SAS of performing the multiple Markov Chain Monte Carlo Imputation method on the dataset. The procedure completed 200 burn-in iterations before each imputation. The procedure used a noninformative (Jeffreys prior) to derive the posterior mode from the EM algorithm as the starting values for the MCMC process.

Table 5.24: Model Information

Model Information	
Data Set	Work.ctsmall
Method	MCMC
Multiple Imputation Chain	Multiple Chains
Initial Estimates for MCMC	EM Posterior Mode
Start	Starting Value
Priors	Jeffreys
Number of Imputations	5
Number of Burn-in Iterations	200
Seed for random number generator	55417

Table 5.25 below shows the posterior mean and covariance matrix generated from the complete case data.

Table 5.25: EM Posterior Mode Estimates

	Weekly Household Income	**Tenure: LA/Other Public Sector**	**Tenure: Housing Association/ Co-ops**	**Tenure: Private Rented**	**Tenure: Joint Users**	**Household Has Children: No**	**Employer Pension: Yes**	**Investment Income: Yes**
MEAN	339.9100	0.1800	0.0900	0.0700	0.0020	0.6180	0.2110	0.1120
COVARIANCE MATRIX								
Weekly Household Income	1031160	20.3530	9.5190	2.8860	0.0530	37.5770	32.5770	11.630
Tenure : LA/Other Public Sector	20.3530	0.1470	0.0160	0.0120	0.0002	0.0032	0.0079	0.0152
Tenure: Housing Association/ Co-ops	9.5126	0.0161	0.08156	0.0062	0.0001	0.0003	0.0062	0.0091
Tenure: Private Rented	2.8856	0.0125	0.0062	0.0645	0.0001	0.0065	0.0056	0.0029
Tenure: Joint Users	0.0531	0.0003	0.0001	0.0001	0.0015	0.0003	0.0003	0.0001
Household Has Children: No	37.5767	0.0032	0.0003	0.0065	0.0003	0.2355	0.0569	0.0108
Employer Pension:Yes	32.5773	0.0079	0.0062	0.0055	0.0003	0.0569	0.1662	0.0131
Investment Income:Yes	11.6295	0.01518	0.0091	0.0029	0.0002	0.01079	0.01309	0.0990

The parameters estimated in Table 5.25 are used to simulate (imputed) missing values. Table 5.26 shows the initial parameter estimates after simulation of missing values based on the mean and covariance matrix.

Table 5.26: Initial Parameter Estimates

	Weekly Household Income	Tenure : LA/Other Public Sector	Tenure: Housing Association/ Co-ops	Tenure: Private Rented	Tenure: Joint Users	Household Has Children: No	Employer Pension: Yes	Investment Income: Yes
MEAN	339.91	0.18	0.09	0.07	0.00	0.62	0.21	0.11
COVARIANCE MATRIX								
Weekly Household Income	103116	-20.35	-9.52	-2.89	-0.05	-37.58	-32.58	11.63
Tenure : LA/Other Public Sector	-20.35	0.1470	-0.0161	-0.0125	-0.0003	-0.0032	-0.0079	-0.0152
Tenure: Housing Association/ Co-ops	-9.519	-0.016	0.082	-0.006	0.000	0.000	-0.006	-0.009
Tenure: Private Rented	-2.89	-0.0125	-0.0062	0.0645	-0.0001	0.0065	-0.0056	-0.0029
Tenure: Joint Users	-0.05	-0.0003	-0.0001	-0.0001	0.0015	-0.0003	-0.0003	-0.0002
Household Has Children: No	-37.58	-0.0032	-0.0003	0.0065	-0.0003	0.2355	0.0569	0.0108
Employer Pension: Yes	-32.58	-0.0079	-0.0062	-0.0056	-0.0003	0.057	0.166	0.013
Investment Income: Yes	11.63	-0.015	-0.009	-0.003	0.000	0.011	0.013	0.099

The process of simulating missing values and generating posterior distributions is repeated till the model converges. Table 5.27 displays the multiple imputation variance after the model (iterative imputation and posterior steps) converges.

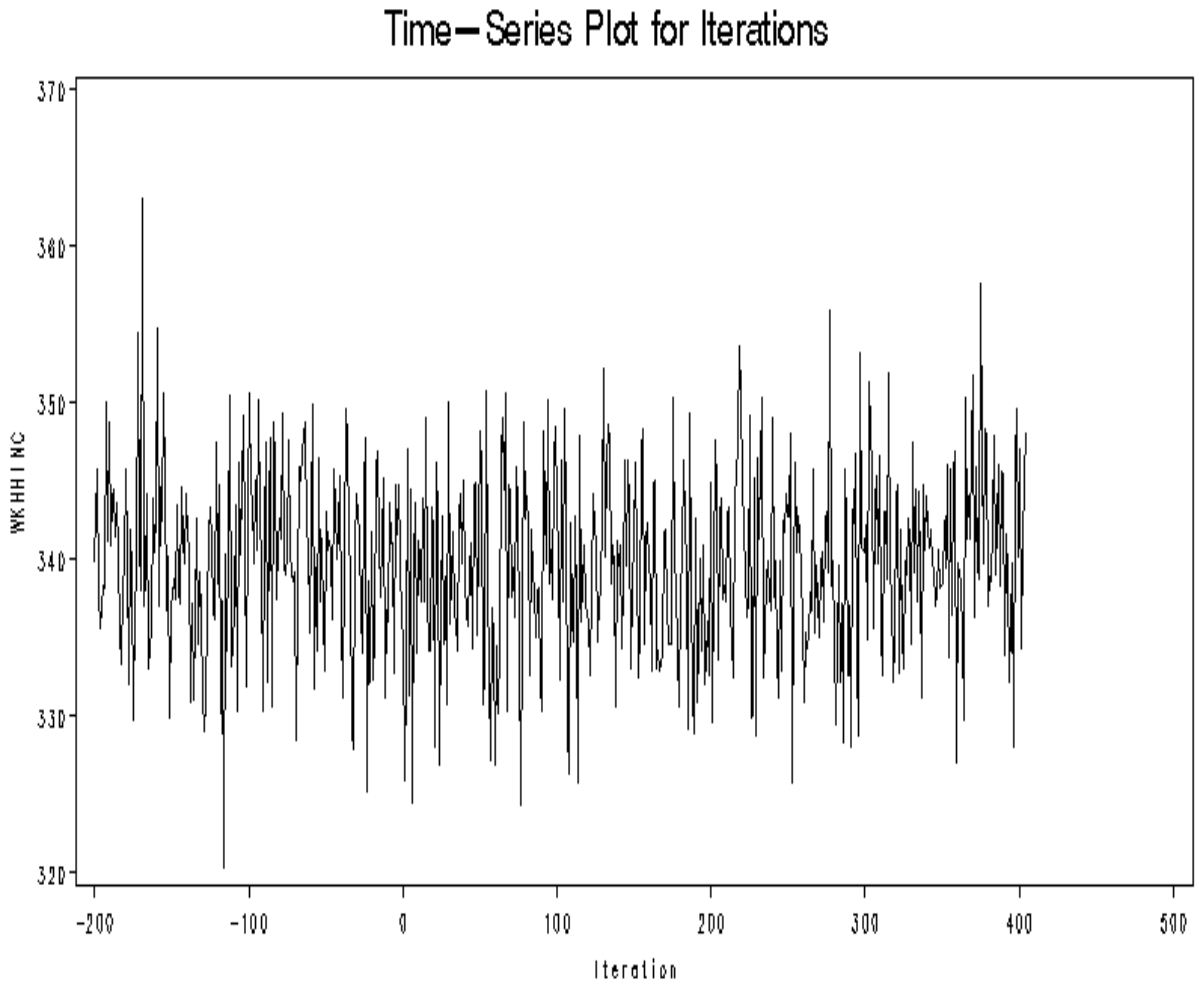
Table 5.27: Multiple Imputation Variance Information

Multiple Imputation Variance Information							
Variable	**Variance**			**DF**	**Relative Increase in Variance**	**Fraction Missing Information**	**Relative Efficiency**
	Between	**Within**	**Total**				
Weekly Household Income	4.06	31.43	36.29	206.38	0.15	0.14	0.97

Table 5.27 also reflects the relative efficiency of the imputation procedure. It was found to be 97% efficient in imputing missing values.

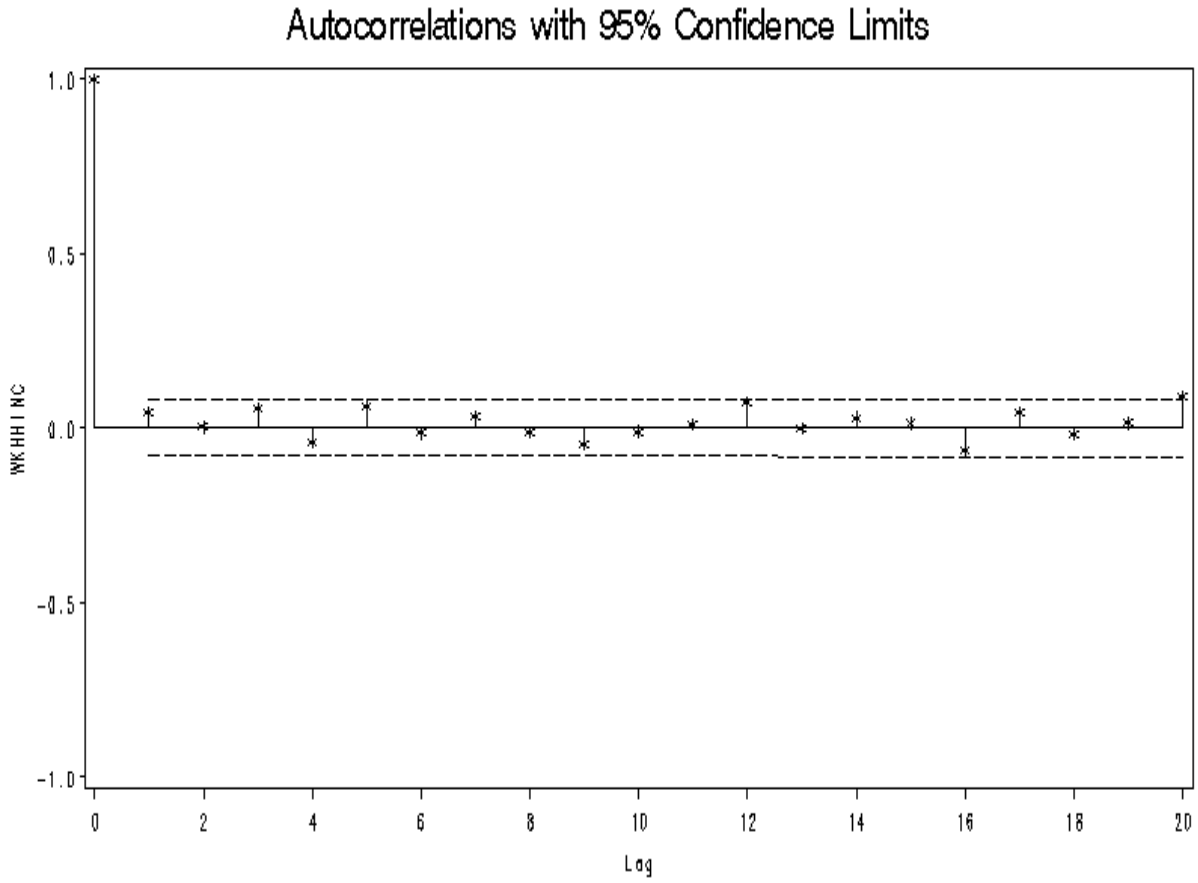
Convergence of the model was confirmed by the time series plot in Figure 5.4. This is a scatter plot of successive parameter estimates against the iteration number. The time series plot appears stationary, and no trends are apparent.

Figure 5.4: Time Series Plot of Iterations



To further substantiate convergence, an autocorrelation function plot provided in Figure 5.8 displayed below. The autocorrelations tend to zero, thus confirming convergence.

Figure 5.8: Autocorrelations with 95% Confidence Intervals



The parameter estimates after the multiple imputation process is summarised in Table 5.28 .

Table 5.28: MCMC Multiple Imputation Parameter Estimates

Variable	Mean	Std Error	DF	t for H0: Mean=Mu0	Pr > |t|
Weekly Household Income	339.69	6.02	206.38	56.38	<.0001

Table 5.28 shows that the mean income value of the whole dataset set after multiple imputation is £339.69. It also shows that the mean income value is non-zero (with respect to the significant t-test).

The income data after multiple regression imputation (combined imputation result) is summarised in Table 5.29 below, together with comparative summary statistics of the complete case income data.

Table 5.29: Mean Income Data By Council Tax Band After MCMC Imputation For Combined Datasets

	Overall Mean	Standard Deviation	Complete Case Mean	Complete Case Standard Deviation
	339.69	312.37	344.63	311.35
Council Tax Band	**Mean**	**Standard Deviation**	**Complete Case Mean**	**Complete Case Standard Deviation**
A	334.51	263.62	340.72	258.14
B	326.90	248.84	327.14	243.10
C	346.68	271.36	350.63	268.60
D	332.29	263.23	337.11	259.54
E	345.06	451.73	348.90	458.78
F	349.18	262.06	354.21	259.30
G	331.26	282.24	339.26	277.28
H	349.57	363.60	355.33	365.67

Table 5.29 shows that the overall mean from this method is lower than the complete case mean, but the standard deviation is higher than that of the complete case income data. The linear patterns are similar to that of the complete case mean, with council tax band H having the highest mean income value for both complete case and imputed data.

5.5.3 : Summary of Imputation Methods : Results

Table 5.30 table summaries each imputation method and the difference from the complete case mean and standard deviation. It reveals that Method 1(mean imputation) gives no difference in mean between the imputed mean income data and complete case mean income data, but gives the highest negative difference between the standard deviation from the imputed data and complete case standard deviation. Similarly, Method 3 (hot deck regression imputation) gives the highest negative mean income difference. Method 7 gives the lowest positive mean income difference and the highest positive standard deviation difference. Therefore, Method 1 appears to minimise the absolute difference between the imputed and complete case means, whereas Method 9 (MCMC imputation) appears to provide the

minimum absolute difference between standard deviations of means from imputed and complete case data.

Table 5.30: Differences Between Complete Case Statistics and Post-Imputed Statistics

Imputation Method	Difference in Mean	Difference in Standard Deviation
Method 1: Single Unconditional Mean Imputation	0	-10.71
Method 2: Single Simple Hot Deck Imputation	-2.59	-3.12
Method 3: Single Hot Deck (Regression) Imputation	-8.23	-5.67
Method 4: Single Hot Deck (Logistic Regression) Imputation	-6.54	-5.55
Method 5: Single Predicted Mean Imputation	-8	-7.28
Method 6: Multiple Simple Hot Deck Imputation	4.2	-2.99
Method 7: Multiple Predictive Model Based (Regression) Imputation	1.8	2.58
Method 8: Multiple Predictive Model Based (Propensity Score) Imputation	4.2	-4.35
Method 9 : Multiple Markov Chain Monte Carlo Imputation	-4.94	1.02

CHAPTER 6

CONCLUSIONS

6.1 Summary of Results

The revival of the original income data set through SPSS programming code revealed the original percentage of missing values, and confirmed that the method employed by the ONS to impute income data using CANCEIS was not fully efficient. Reasons attributed to incomplete efficiency could be attributed to rigid constraints specified in CANCEIS, and the lack of populated donors (that is, donors with many missing values). The latter was confirmed by the frequency tables of some donor variables that had many missing values. The investigation of missing data mechanism, by cross tabulations performed on relevant variables, showed that the missing data was biased. That is , there was an association at the 5% significance level between the missingness of data and the variables being compared. This enabled assumption of the MAR mechanism, since the associations between missingness and auxiliary variables rule out the MCAR mechanism (Rubin and Little, 2002). The missing data pattern revealed was arbitrary. However, for the purposes of this analysis, it was converted to a monotone dataset. The variables in the dataset were found to deviate from multivariate normality. However, multivariate normality was assumed since multiple imputation methodologies would be robust to such deviations (Meng,1994).

Principal component analysis resulted in eighty two relevant factors. There was no clear 'elbow' in the scree plot. The factors generated did not seem to explain any logical groupings/patterns of variables. This did not serve the purpose of data reduction. Relevant EM parameters were estimated to demonstrate the parameters for the MCMC model. The intermediary multiple linear regression model was significant at the 5% level. The intermediary logistic regression model was also significant at the 5% level. Approximately 30% of the variance was explained by the models. The area under the ROC curve of the logistic regression model was reasonably close to 1, hence confirming the validity of the model.

In terms of the difference between the mean income from imputed and complete case data, Table 5.30 (Chapter 5) reveals that the single unconditional mean imputation (Method 1) is the best method since the difference appears to be zero. Following this, multiple linear regression imputation (Method 7) appeared to deviate the least from the complete case mean. The largest difference of means between imputed and complete case data was from single hot

deck regression imputation (Method 3). On the whole, the multiple imputation methods appeared to deviate less from the complete case mean than the single imputation methods. All the multiple imputation methods, but the MCMC imputation (Method 9), deviated positively from the complete case mean. The deviations from the complete case mean for all methods were not substantial. Even Method 3 just deviated from the complete case mean by approximately eight units (pounds sterling). In percentages, this is just 3% of the complete case mean value.

Table 5.30 (Chapter 5) also showed the deviations of the imputed data from the standard deviations of the complete case means. Single mean imputation (Method 1), which was regarded as the best method from the mean criteria above, deviated the most (negatively) from the complete case standard deviation. Thus, the mean imputation method (Method 1) is not an appropriate method for imputation, because it underestimates the standard deviation of the mean income substantially. This would be expected because unconditional mean imputations distort the distribution of variables and their interrelationships in the dataset (Durrant, 2002). The reason why the mean imputation gave no difference between the imputed mean and complete case mean is because of the low original percentage of missing data in the dataset. All the methods, but the multiple regression imputation (Method 7) and the MCMC imputation (Method 9), deviated positively from the complete case standard deviation. This is expected of multiple imputation methods (Methods 7 and 9) because they incorporate variance due to imputation. The MCMC method, in particular, also incorporates variances due to uncertainty of parameter estimates (Schafer, 1997), thus the standard deviation of this methods is expected to be greater than the complete case standard deviation. The other multiple imputation methods (for example Methods 6 and 8) did not overestimate the standard deviation because they may have been less compatible with the missing data pattern (which was predominantly arbitrary). On the whole, multiple imputation methods deviated less than single imputation methods from the complete case standard deviation.

The individual summary results for each imputation method described in Chapter 5 also depicted the pattern of mean income values across council tax bands. It appears that all the single imputation methods (Methods 1 to 5) and the MCMC multiple imputation method (Method 9) conform to the approximate linear trend of the complete case pattern. The rest of the methods, including the multiple linear regression imputation (Method 7), deviate from the trend the complete case means follow, although the general trend remains linear (with a positive slope) from council tax band A to H. This criterion renders Method 7 less

appropriate, despite the low deviations of mean and standard deviation from the complete case statistics.

Based on the above criteria, it would appear MCMC imputation (Method 9) would be the optimum method to apply. Not only is it best suited to arbitrary missing data patterns (which the SHCS dataset demonstrates), but it also incorporates additional variance due to imputation and uncertainty of prior estimates. It deviates from the complete case mean just by 4 units (pounds). It is easy to implement using in-house SAS software. The major criticism of MCMC imputation appears to be the problem of subjective prior distributions. This is circumvented by specifying non-informative priors, as was done in this case. In addition, the numerous iterations mitigate the effect of an inaccurate prior (Schafer and Graham, 2002). The method is robust to violations of assumptions of multivariate normality when the sample size is large (Meng, 1994), and the SHCS dataset is a very large sample (almost 4000 observations). Hence, the MCMC method (Method 9) would be the best method from an academic perspective.

In choosing an optimum method, the deviations from complete case summary statistics are considered together with the practicality of implementation of the method. The overview of the methods revealed that all of the methods performed well (in terms of deviations from complete case statistics). Despite the merits of the MCMC method, it may be inconvenient to maintain five datasets. The criteria thus considered for recommendation of optimum methodology are efficacy of method (in terms of deviations), ease of implementation and consistency. This results in the single hot deck logistic regression imputation (Method 4) as the optimum method to adopt for the SHCS in its current context. It deviates less from the complete case statistics than some of the other imputation methods, only one dataset is maintained (since it is a single imputation method) and it applies the same logic as the previous imputation method applied by the ONS. The CANCEIS imputation had involved choosing matching variables by performing a logistic regression of missingness and then hotdecking based on the covariates. The hotdeck logistic regression imputation method would thus be more consistent with past methodology. In addition, it has additional merits of maintaining distribution and relationships between variables in the dataset (Durrant, 2002), and would be less reliant on modelling assumptions since it is semiparametric (Allison, 2001). This means only the matching variable selection is dependent on modelling assumptions, and not the hot deck procedure itself. Single hot deck (logistic regression) imputation would thus be the optimum method to implement.

6.2 Limitations of Study

The primary limitation of this study was that only one dataset was used for analysis. This dataset was used as representative of the SHCS data on the whole. Another limitation was that the dataset was converted to a monotone missing pattern, despite being arbitrary. Even though this is common imputation practice, it slightly robs the process of its statistical integrity.

The intermediary multiple regression model and logistic regression model were constructed using a reductionist approach, since the main purpose of these models was to select a few matching variables rather than being the ultimate basis of inference. The models could have potentially been more extensive in terms of explanatory appeal.

The optimum method recommended was performed using SOLAS. Since the Community Analytical Services department of the Scottish Government does not have this software in house, the method cannot be implemented unless a code is written for it in SPSS programming or SAS. Another limitation worth mentioning is that many of the models used in imputations deviated from the necessary assumptions. However, the robust nature of some the methods does not make this limitation a critical one.

The use of council tax bands to assess the logic of the income results could have been prone to error. This would have been the case if an individual paid less council tax because he or she received a discount based on living alone, rather than based on his/her household income. However, council tax band was the most appropriate variable to use for comparative purposes for this study.

6.3 Further Work

The first step after this piece of work, would be to prepare an in-house SAS or SPSS code so that the recommended imputation methodology could be used by the SHCS department.

It would also be of interest to consider missing values of any latent variables in the data sets and follow on from the principal component analysis initiated with this dataset. For example, some of the variables in the dataset could be grouped together to measure factors

that could be investigated. Assuming sufficient data was available on income values and heating equipment, fuel poverty for example could be a factor that would be investigated based on these observed available variables.

A more extensive piece of work to check the validity of imputation values would be to mimic imputed values by simulating a dataset with missing values in the same distribution as they were in the original set. These values could them be compared with results from imputation on the dataset to ascertain how effective the imputation methodology was.

It would also be useful to repeat imputation methods on as many SHCS data sets as possible to confirm the efficacy of the methods. The intermediate models could also be made more explanatory by inputting more variables, to achieve more extensive and less parsimonious models. Multiple models could be generated and the best one chosen.

It would also be of interest to maintain the arbitrary pattern of the dataset and conduct a Markov Chain Monte Carlo imputation on the dataset without imputing any of the variables to check how well it performs, since in theory (Schafer, 1997) this would be the best method suited to such a data pattern.

Other further work could involve imputing at the component-variable level to check for any discrepancies in results from component-variable level imputation and overall income- level imputation (which was used in this study).

References

Allison, P.D. (2001): Missing Data, Iowa: Sage.

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

Allison, P.D. (2001): Missing Data, *Sage University Papers Series on Quantitative Applications in the Social Sciences*, series no. 07-136, Thousand Oaks.

Bankier, M. (1999): Experience with the New Imputation Methodology used in the 1996 Canadian Census with extension for future Censuses, *Proceedings of the Workshop on Data Editing*, UN/ECE, Italy (Rome).

Bankier, M. et al. (2001): Efficient Methodology Within the Canadian Census Edit and Imputation System (CANCEIS), *Proceedings of the Annual Meeting of the American Statistical Association August 5-9.*

Beaumont, J.F. (2003): The System for Estimation of Variance due to Nonresponse and Imputation (SEVANI), *The Imputation Bulletin*, 3,1, 2003, 6-9.

Binder, D.A. and Sun, W. (1996): Frequency Valid Multiple Imputation for Surveys with a Complex Design, *Proceedings of the Section on Survey Research Methods*, *American Statistical Association*, 281-286.

Burren, v. S., Boshuizen, H.C. and Knock, D.L. (1999): Multiple Imputation of Missing BloodPressure Covariates in Survival Analysis, *Statistics in Medicine*, 18, 681-696.

Carlin, J. B., Li, N., Greenwood, P. and Coffey, C. (2003): Tools for Analyzing Multiple Imputed Datasets, *The Stata Journal*, 3, 3, 226-244.

Carter, R.(2006): Solutions for Missing Data in Structural Equation Modelling, *Research in Practice in Assessment*, 1,1,1-5

Chambers, R. (2003): Evaluation Criteria for Statistical Editing and Imputation, *National Statistical Methodological Series,* 28, Office for National Statistics, London.

Chen, J. and Shao, J. (2000): Nearest Neighbour Imputation for Survey Data, *Journal of Official Statistics*, 16, 2, 113-131.

Chen, J. and Shao, J. (2001): Jackknife Variance Estimation for Nearest Neighbour Imputation, *Journal of the American Statistical Association*, 96, 453, 260-269.

Cody, R.P. and Smith, J.K. (2005): Applied Statistics and the SAS Programming Language, New York: Prentice Hall

Continuous Scottish House Condition Survey Technical Report Year 3 (2005/06).

David, M. H., Little, R., Samuhel, M. and Triest, R. (1983): Imputation Models based on the Propensity to Respond, *Proceedings of the Business and Economic Statistics Section*, *American Statistical Association*, 168-173.

Deville, J.C. and Särndal, C.E. (1994): Variance Estimation for the Regression Imputed Horvitz-Thompson Estimator, *Journal of Official Statistics*, 10, 4, 381-394.

Durrant, G.B. and Skinner, C. (2005a): Using Missing Data Methods to Correct for Measurement Error in a Distribution Function, *Survey Methodology*, forthcoming.

Durrant, G.B. and Skinner, C. (2005b): 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,* forthcoming.

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

Freedman, V.A. and Wolf, D.A. (1995): A Case Study on the Use of Multiple Imputation, *Demography*, 32, 3, 459-470.

Fuller, W. and Kim, J.K. (2005): Replicated Nearest Neighbour Imputation, *Bulletin of the International Statistical Institute*, 55th Session, Sydney.

Gamerman, D.(1997) : Markov Chain Monte Carlo, London: Chapman & Hall

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (1998): *Bayesian Data Analysis*, London.
Government Statistical Service (GSS) (1996): Report of the Task Force on Imputation, *Government Statistical Service Methodology Series*, 3, London.
Gourieroux, C. and Monfort, A. (1981): On The Problem of Missing Data in Linear Models, *The Review of Economic Studies*, 84, 4, 571-586
Haziza, D. (2002): Genesis, Generalised System for Imputation Simulation, *The Imputation Bulletin*, 2, 2, 8-10.
Heitjan, D.F. (1994): Ignorability in General Incomplete-Data Models, *Biometrika*, 81, 4, 701-708.
Heitjan, D.F. and Landis, J.R. (1994): Assessing Secular Trends in Blood Pressure, A Multiple Imputation Approach, *Journal of the American Statistical Association*, 89, 427, 750-759.
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.
Heitjan, D.F. and Rubin, D.B. (1990): Inference from Coarse Data via Multiple Imputation with Application to Age Heaping, *Journal of the American Statistical Association*, 85, 410, 304-314.
Hirsch, B.T. and Schumacher, E.J. (2004): Match Bias in Wage Gap Estimates Due to Earnings Imputation, *Journal of Labour Economics*, 22, 3, 689-721.
Horton, N.J. and Lipsitz, S.R. (2001): Multiple Imputation in Practice: Comparison of Software
Horvitz, D.G. and Thompson, D.J. (1952): A Generalization of Sampling Without Replacement from a Finite Universe, *Journal of American Statistical Association*, 47, 663-685
Housing Scotland Act 2001 avalaible at:
http://www.opsi.gov.uk/legislation/scotland/acts2001/20010010.htm *accessed 3 Septemeber 2007)*
HOX, J.J. (1999): A Review of Current Software for Handling Missing Data, *Kwantitatieve Methoden*, 62, 123-138.
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.
IVEware Software Information available at:
http://www.irv.umich.edu/src/swp/ive/ *accessed 3 September 2007.*
Jinn, J.H. and Sedransk, J. (1989): Effect on Secondary Data Analysis of Common Imputation Methods, *Sociological Methodology*, 19, 213-241.
Kalton, G. (1983): *Compensating for Missing Survey Data*, Michigan.
Kalton, G. and Kasprzyk, D. (1982): Imputing for Missing Survey Responses, *Proceedings of the Section on Survey Research Methods*, *American Statistical Association*, 22-31.
Kalton, G. and Kish, L. (1984): Some Efficient Random Imputation Methods, *Communications in Statistics*, *Part A*, *Theory and Methods*, 13, 1919-1939.
Kim, J.K. (2002): A Note on Approximate Bayesian Bootstrap Imputation, *Biometrika*, 89, 2, 470-477
Kim, J.K. and Fuller, W. (2004): Fractional Hot Deck Imputation, *Biometrika*, 91, 3, 559-578.
Lessler, J.T. and Kalsbeek W.D. (1992): *Nonsampling Error in Surveys*, New York, Chichester.

Lillar, L., Smith, J.P. and Welch, F. (1986): What do We Really Know About Wages?, The Importance of Nonreporting and Census Imputation, *Journal of Political Economy*, 94, 3, 489-506.
Lipsitz, S.R., Zhao, L.P. and Molenberghs, G. (1998): A Semiparametric Method of Multiple Imputation, *Journal of the Royal Statistical Society*, *Series B*, *Statistical Methodology*, 60, 1, 127-144.
Little, R.J.A. (1986): Survey Nonresponse Adjustments for Estimates of Means, *InternationalStatistical Review*, 54, 2, 139-157.
Little, R.J.A. (1988, Dec.): A Test of Missing Completely at Random for Multivariate Data with Missing Values, *Journal of the American Statistical Association*, 83, 404, 1198-1202.
Little, R.J.A. (1988, July): Missing-Data Adjustments in Large Surveys, *Journal of Business and Economic Statistics*, 6, 3, 287-301.
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.
Little, R.J.A. and Rubin, D.B. (2002): *Statistical Analysis with Missing Data*, New York.
Manski, C. (1995): *Identification Problems in the Social Sciences*, Cambridge.
Manski, C. (2005): Partial Identification with Missing Data: Concepts and Findings, *International Journal of Approximate Reasoning*, 39, 2-3, 151-165.
Meng, X.L. (1994) Multiple-imputation inferences with uncongenial sources of input (with discussion). *Statistical Science*, 10, 538 - 573.
Missing Data Information available at:
http://www.missingdata.org *accessed 27 July 2007*.
Multiple Imputation Software Information available at:
http:/www.stat.rtu.edu~jls/misoftwa.html *accessed 3 September 2007.*
Nielsen, S.F. (2003): Proper and Improper Multiple Imputation, *International Statistical Review*, 71,3, 593-627.
Nordholt, E.S. (1998): Imputation: Methods, Simulation, Experiments and Practical Examples, *International Statistical Review*, 66, 2, 157-180.
Ott, R.L. and Longecker, M.(2000): An Introduction to Statistical Methods and Data Analysis, California: Duxbury
Polychoric Correlation Original Macro Code available at:
http://support.sas.com/samples_app/00/sample00512_1_polychor.sas.txt accessed 3 September 2007.
Raghunathan, T.E., Lepkowski, J.M. van Hoewyk M., Solenberger P.W. (2001): A Multivariate Technique for Multiply Imputing Missing Values using a Sequence of Regression Models, *Survey Methodology*, 27, 85-95.
R Software Information available at:
http://www.r-project.org *accessed 3 September 2007*.
Rancourt, E. (1999): Estimation with Nearest Neighbour Imputation at Statistics Canada, in:*Proceedings of the Section on Survey Research Methods*, *American Statistical Association*, 131-138.
Rao, J.N.K. (1996): On Variance Estimation with Imputed Survey Data, *Journal of the American Statistical Association*, 91, 434, 499-506
Rao, J.N.K. and Shao, J. (1992): Jackknife Variance Estimation with Survey Data under Hot Deck Imputation, *Biometrika*, 79, 4, 811-822.
Rao, J.N.K. and Sitter, R.R. (1995): Variance Estimation under Two-Phase Sampling with
Applications to Imputation for Missing Data, *Biometrika*, 82, 2, 453-460.

Rao, J.N.K., Shao J. (1999): Modified Balanced Repeated Replication for Complex Survey Data, *Biometrika*, 86, 2, 403-415.
Raskin and Novacek (1988): A Principal-Components Analysis of the Narcissistic Personality Inventory and Further Evidence of Its Construct Validity, *Journal of Personality and Scoial Psychology*, 54, 5, 890-902.
Robins, J.M. and Rotnitzky, A. (1995): Semiparametric Efficiency in Multivariate Regression Models with Missing Data, *Journal of the American Statistical Association*, 90, 429, 122-129.
Robins, J.M., Rotnitzky, A. and Zhao, L.P. (1994): Estimation of Regression Coefficients when some Regressors are not Always Observed, *Journal of the American Statistical Association*, 89, 427,846-866.
Rosenbaum, P.R. and Rubin, D.B. (1983): The Central Role of the Propensity Score in Observational Studies for Causal Effects, *Biometrika*, 70, 1, 41-55.
Rosenbaum, P.R. and Rubin, D.B. (1985): Constructing a Control-Group Using Mutlivariate Matched Sampling Methods that Incorporate the Propensity Score, *The American Statistician*, 39, 1, 33-38.
Royston, P. (2004): Multiple Imputation of Missing Values, *The Stata Journal*, 4, 3, 227-241.
Rubin, D.B. (1987): *Multiple Imputation for Nonresponse in Surveys*, New York, Chichester.
Rubin, D.B. (1996): Multiple Imputation after 18+ Years, *Journal of the American Statistical Association*, 91, 434, 473-489.
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.
Särndal, C.E., Swensson, B. and Wretman, J. (1992): *Model Assisted Survey Sampling*, New York.
SAS Software Information available at:
http://www.sas.com/ : *accessed 3 September 2007.*
Scottish Fuel Poverty Statement 2002 available at:
http://www.scotland.gov.uk/Publications/2002/08/15258/9951 *accessed 3 September 2007*.
Schafer J.L. (1999): Multiple Imputation: A Primer, *Statistical Methods in Medical Research*, 8, 3-15.
Schafer, J. L. (1997): *Analysis of Incomplete Multivariate Data*, London: Chapman & Hall
Schafer, J.L. (1999): Multiple Imputation: A Primer, Statistical Methods in Medical Research, 8, 3-15.
Schafer, J.L. (2001): Multiple Imputation with Pan, in: Collins, L.M. and Sayer, A.G. (eds), *New Methods for the Analysis of Change*, Washington, 2001, 357-377.
Schafer, J.L. and Graham, J.W. (2002): Missing Data: Our View of the State of the Art, *Psychological Methods*,7, 2, 147-177.
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
Schenker, N. and Taylor, J.M.G. (1996): Partially Parametric Techniques for Multiple Imputation,*Computational Statistics and Data Analysis*, 22, 425-446.
Shao, J and Sitter, R.R. (1996): Bootstrap for Imputed Survey Data, *Journal of the American Statistica lAssociation*, *Theory and Methods*, 91, 435, 1278-1286.

Shao, J. and Steel, P. (1999): Variance Estimation for Survey Data with Composite Imputation and Nonnegligible Sampling Fractions, *Journal of the American Statistical Association*, *Theory and Methods*, 94, 445, 254-265.
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.
Skinner, C. and Rao, J.N.K. (2002): Jackknife Variance Estimation for Multivariate Statistics.
SOLAS Software Information available at:
http://www.statssol.ie/html/dolas/solas_home.html *accessed 3 September 2007.*
SPSS Software Information available at :
http://www.spss.com *accessed 3 September 2007.*
STATA Software Information available at:
http://www.stata.com *accessed 3 September 2007*.
Technical Report Scottish Fuel Poverty 2002 available at:
http://www.scotland.gov.uk/Topics/Statistics/SHCS/technicalnotefuelpoverty *accessed 3 September 2007.*
UCLA Paper Examples available at:
http://www.ats.ucla.edu/stat/sas/paperexamples/schafer1/default.htm *accessed 3 Septemeber 2007*

APPENDICES

Appendix 1: Sample of SPSS Programming Code Before and After Modification To Generate Original Dataset

Post-imputation Code For Earned Income

```
********************************************** shcearndv_imputed.sps
**********************************************.

******************************** EARNINGS INCOME DERIVED
VARIABLES**********************************.

*****based on GHS 2003

***** look at cases.

*****Variables created in this program are   NETPAY ntsecjob seproit seprnet.

* Modified by DKD 12/10/05.
* Takes into account new imputed variable names from Methodology.
* These are Respondent:    Din1, Din4, Din5, Din6, Din7
* Partner:            Din35, Din36, Din37, Din39, Din40, Din41, Din42, Din56a, Din56, Din57, Din58i, Din59
* Imputed variables are the same as thse, but with an appended "i".



recode din4i (1 thru 999997=1) (else=copy) into din4test.
missing values din4test ().

fre vars = din4test.

**** cases with DK usual net.

temp.
select if (din4test= 999999).
list variables = caseid din4i to din7i, din8a to din16.

**** if DK usual net (Din4) asks usual gross (DIN8) = 3 cases.
```

**** if DK usual net (Din4) and no usual gross (DIN8), asks for last net (DIN12) = 2 cases + 1 DK .
**** if DK usual net (Din4) no usual gross, dK last net asks for last gross (DIN16) = DK only.

**** check age lt 16.
temp.
select if (dvage lt 16).
list variables = caseid dvage din4i to din7i.

**** investigate pay periods 1 hour and 1 day.
temp.
select if (din5i = 91 or din5i=92).
list variables = caseid din2 din3a din3b din4i din5i din7i din8.
**** only one case, paid £9 per hour but DK how many hours worked in week.

temp.
select if din5i=97.
list variables = caseid stat din1i, din2 din3, din4i to din7i,din8a to to din19.
*** 4 cases.

**********.
********** NETPAY **********.
missing values din4i ().

DO IF DVAGE LT 16.
. COMPUTE NETPAY = -9.
ELSE IF DIN1i =6.
. COMPUTE NETPAY = -7.
ELSE IF (DIN2=1 AND DIN3a = 2).
. COMPUTE NETPAY=-7.
ELSE IF (DIN2=2 AND DIN3B = 2).
. COMPUTE NETPAY=-7.
ELSE IF DIN4i =999998.
. COMPUTE NETPAY = -9.
ELSE IF (RANGE(DIN1i,1,5)).
. DO IF DIN3a = 1 or DIN3b = 1.
. DO IF (DIN4i = 999999 AND DIN8 GE 0).
. DO IF RANGE (DIN9,1,4) OR RANGE (DIN9,13,52).
. COMPUTE NETPAY = DIN8/DIN9 * 100.

```
.   ELSE IF DIN9 = 5.
.         COMPUTE NETPAY = DIN8 * 12/52 * 100.
.   ELSE IF DIN9 = 7.
.         COMPUTE NETPAY = DIN8 * 6/52 * 100.
.   ELSE IF RANGE (DIN9,8,10).
.         COMPUTE NETPAY = DIN8 *DIN9/52 * 100.
.   ELSE IF RANGE (DIN9,90,92).
.         COMPUTE NETPAY = DIN8 * 100.
.   ELSE IF RANGE (DIN9,95,99).
.         COMPUTE NETPAY = -8.
.   END IF.
. END IF.

. DO IF (DIN4i = 999999  AND DIN12 GE 0).
.   DO IF RANGE (DIN13,1,4) OR RANGE (DIN13,13,52).
.         COMPUTE NETPAY = DIN12/DIN13 * 100.
.   ELSE IF DIN13 = 5.
.         COMPUTE NETPAY = DIN12 * 12/52 * 100.
.   ELSE IF DIN13 = 7.
.         COMPUTE NETPAY = DIN12 * 6/52 * 100.
.   ELSE IF RANGE (DIN13,8,10).
.         COMPUTE NETPAY = DIN12 *DIN13/52 * 100.
.   ELSE IF RANGE (DIN13,90,92).
.         COMPUTE NETPAY = DIN12 * 100.
.   ELSE IF RANGE (DIN13,95,99).
.         COMPUTE NETPAY = -8.
.   END IF.
. END IF.

. DO IF (DIN4i = 999999  AND DIN16 GE 0).
.   DO IF RANGE (DIN17,1,4) OR RANGE (DIN17,13,52).
.         COMPUTE NETPAY = DIN16/DIN17 * 100.
.   ELSE IF DIN17 = 5.
.         COMPUTE NETPAY = DIN16 * 12/52 * 100.
.   ELSE IF DIN17 = 7.
.         COMPUTE NETPAY = DIN16 * 6/52 * 100.
.   ELSE IF RANGE (DIN17,8,10).
.         COMPUTE NETPAY = DIN16 *DIN17/52 * 100.
.   ELSE IF RANGE (DIN17,90,92).
.         COMPUTE NETPAY = DIN16 * 100.
```

```
.    ELSE IF RANGE (DIN17,95,99).
.          COMPUTE NETPAY = -8.
.    END IF.
.  END IF.

.  DO IF (DIN4i = 999999 AND DIN8a NE 1).
.    COMPUTE NETPAY = -8.
.  END IF.

.  DO IF (DIN4i GE 0).
.    DO IF RANGE (DIN5i,1,4) OR RANGE (DIN5i,13,52).
.          COMPUTE NETPAY = DIN4i/DIN5i * 100.
.    ELSE IF DIN5i = 5.
.          COMPUTE NETPAY = DIN4i * 12/52 * 100.
.    ELSE IF DIN5i = 7.
.          COMPUTE NETPAY = DIN4i * 6/52 * 100.
.    ELSE IF RANGE (DIN5i,8,10).
.          COMPUTE NETPAY = DIN4i *DIN5i/52 * 100.
.    ELSE IF RANGE (DIN5i,90,92).
.          COMPUTE NETPAY = DIN4i * 100.
.    ELSE IF RANGE (DIN5i,95,99).
.          COMPUTE NETPAY = -8.
.    ELSE IF DIN4i = -9 AND DIN5i LT 90.
.          COMPUTE NETPAY = -9.
.    ELSE IF DIN5i =-8.
.          COMPUTE NETPAY=-8.
.    ELSE.
.          COMPUTE NETPAY = -9.
.    END IF.
.  END IF.

. END IF /* DIN3a  = 1 or DIN3b = 1.

END IF /* DVAGE LT 16.

temp.
select if (DIN4i = 999999  AND DIN8 GE 0).
list variables = caseid din4i din8 din9 netpay.

compute netpayx=netpay.
```

```
fre vars =netpayx.

recode netpayx (300 thru 150000=999)(SYSMIS=11111111) (else=copy).
VAR LABEL NETPAY  'Usual net weekly pay - employees (pence per wk)'
          /NETPAYx 'Usual net weekly pay - employees (pence per wk) - recoded'.
Value label NETPAY NETPAYx      -9 'DNA/child/proxy/NO INT/refused'
                                -8 'Don t know'
                                -7 'not working'
                                 0 'No pay received'.

missing values netpayx (-9 -8 -7).
freq vars=netpayx.

temp.
select if (NETPAYx=11111111).
list variables = caseid din1i, din2 to din3b, din4i to din 7i, din8a TO din9 netpay.

temp.
select if netpayx ne  999.
list variables = caseid netpayx din1i, din2 to din3a to din3b, din4i to din 7i, din8a TO din19.

**** 5 cases earn more than £80k - fine.

missing values netpayx ().

temp.
select if (netpayx=-8 or netpayx=-9).
list variables = caseid netpayx din1i, din2 to din3a to din3b, din4i to din 7i, din8a TO din19.

*** cases with amounts but period = 95 or 97 will need imputing ***.

missing values netpayx (-9 -8 -7).

recode netpay (-7=0) (sysmis=-9) (else=copy).
Value label NETPAY      -9 'DNA/child/proxy/NO INT/refused'
                        -8 'Don t know'
                         0 'No pay received'.

*********************************************************************************.
```

```
********** GHS Asks about bonuses but SHC doesn't so section deleted**********.

fre vars = din2 din3b din27 to din32.

*******************************************************************************.
********** NTSECJOB *********************************.
missing values din27 din30 ().

DO IF DVAGE LT 16.
+ COMPUTE NTSECJOB = -9.
ELSE IF DIN1i = 6.
+ COMPUTE NTSECJOB = -7.
ELSE IF (DIN2=1).
+ COMPUTE NTSECJOB = -7.
ELSE IF DIN27 =999998.
+ COMPUTE NTSECJOB= -9.
ELSE IF DIN3b=2.
+ COMPUTE NTSECJOB = -7.
ELSE IF DIN3b = 1.
+ COMPUTE NTSECJOB = 0.

. DO IF (DIN27 = 999999  AND DIN30 = 999999).
.   COMPUTE NTSECJOB = -8.
. END IF.

. DO IF (DIN27 = 999999  AND DIN30 GE 0).

.   DO IF RANGE (DIN31,1,4) OR RANGE (DIN31,13,52).
.     COMPUTE NTSECJOB = DIN30/DIN31 * 100.
.   ELSE IF DIN31 = 5.
.     COMPUTE NTSECJOB = DIN30 * 12/52 * 100.
.   ELSE IF DIN31 = 7.
.     COMPUTE NTSECJOB = DIN30 * 6/52 * 100.
.   ELSE IF RANGE (DIN31,8,10).
.     COMPUTE NTSECJOB = DIN30 *DIN31/52 * 100.
.   ELSE IF RANGE (DIN31,90,92).
.     COMPUTE NTSECJOB = DIN30 * 100.
.   ELSE IF RANGE (DIN31,95,99).
.     COMPUTE NTSECJOB = -8.
.   END IF.
```

```
. END IF  /* (DIN27 = 999999  AND DIN30 GE 0).

. DO IF (DIN27 GE 0).

.   DO IF RANGE (DIN28,1,4) OR RANGE (DIN28,13,52).
.    COMPUTE NTSECJOB = DIN27/DIN28 * 100.
.   ELSE IF DIN28 = 5.
.    COMPUTE NTSECJOB = DIN27 * 12/52 * 100.
.   ELSE IF DIN28 = 7.
.    COMPUTE NTSECJOB = DIN27 * 6/52 * 100.
.   ELSE IF RANGE (DIN28,8,10).
.    COMPUTE NTSECJOB = DIN27 *DIN28/52 * 100.
.   ELSE IF RANGE (DIN28,90,92).
.    COMPUTE NTSECJOB= DIN27 * 100.
.   ELSE IF RANGE (DIN28,95,99).
.    COMPUTE NTSECJOB = -8.
.   ELSE IF DIN27 = -9 AND DIN28 LT 90.
.    COMPUTE NTSECJOB = -9.
.   ELSE IF DIN28 =-8.
.    COMPUTE NTSECJOB=-8.
.   ELSE.
.    COMPUTE NTSECJOB = -9.
.   END IF.

. END IF /* (DIN27 GE 0).

END IF /* DVAGE LT 16.

compute ntsecjox=ntsecjob.
recode ntsecjox (50 thru 75000=999)(else=copy).

VAR LABELS NTSECJOB 'Net weekly - other jobs (pence per wk)'.
VAR LABELS NTSECJOx 'Net weekly - other jobs (pence per wk) - recoded'.

Value label  NTSECJOB  ntsecjox -9 'DNA/CHILD/PROXY/NO INT/refused'
                                 -8 'DK'
                                 -7 'not working/only 1 job/selfemp'
                                  0 'No earnings'.
missing values  ntsecjox (-9 -8 -7).
```

```
freq vars= ntsecjox.

missing values ntsecjox ().

temp.
select if (ntsecjox = -8 or ntsecjox= -9).
list variables = caseid ntsecjob dvage din1i, din2, din3a, din3b,din4i din27 to din32.

recode ntsecjob (-7=0) (sysmis=-9) (else=copy).
Value label ntsecjob       -9 'DNA/child/proxy/NO INT/refused'
                            -8 'Don t know'
                             0 'No 2nd pay received'.

*************************************************************************************
*******.
********************************** SEPROFIT
*********************************************.

**** includes Self employed in 2nd job *****************.

fre vars = din3a din3b din20a to din24.

missing values DIN1i din20 DIN24 ().

compute selfemp = 2.
DO IF (DIN3a=2) or (DIN3b=2).
.            COMPUTE selfemp=1.
END IF.

VAR LABEL selfemp  'whether self-employed'.
Value label selfemp 1 'yes'   2 'no'.

fre vars = selfemp.

DO IF DVAGE LT 16.
. COMPUTE SEPROFIT = -9.
ELSE IF DIN1i = 8.
. COMPUTE SEPROFIT = -9.
ELSE IF DIN1i = 6.
. COMPUTE SEPROFIT = -7.
```

```
ELSE IF (selfemp=2).
. COMPUTE SEPROFIT = -7.
ELSE IF (DIN20a = 2  AND (DIN24 GE 0 and din24 LE 99997)).
. COMPUTE SEPROFIT = DIN24/52 * 100.
ELSE IF (DIN20a = 2  and DIN24=999999).
. COMPUTE SEPROFIT = -8.
ELSE IF (DIN20a=3).
. COMPUTE SEPROFIT = -8.
ELSE IF (DIN20a=4 OR DIN20=999998).
. COMPUTE SEPROFIT = -9.
ELSE IF (RANGE(DIN1i,1,5)) AND (selfemp=1).
. COMPUTE SEPROFIT = 0.

. DO IF (DIN20 = 999999).
.   COMPUTE SEPROFIT = -8.
. END IF.

. DO IF (DIN20 GE 0 AND DIN20 LE 999997).

.   DO IF RANGE (DIN21,1,4) OR RANGE (DIN21,13,52).
.     COMPUTE SEPROFIT = DIN20/DIN21 * 100.
.   ELSE IF DIN21 = 5.
.     COMPUTE SEPROFIT = DIN20 * 12/52 * 100.
.   ELSE IF DIN21 = 7.
.     COMPUTE SEPROFIT = DIN20 * 6/52 * 100.
.   ELSE IF RANGE (DIN21,8,10).
.     COMPUTE SEPROFIT = DIN20 *DIN21/52 * 100.
.   ELSE IF RANGE (DIN21,90,92).
.     COMPUTE SEPROFIT = DIN20 * 100.
.   ELSE IF RANGE (DIN21,95,99).
.     COMPUTE SEPROFIT = -8.
.   ELSE IF DIN20 = -9 AND DIN21 LT 90.
.     COMPUTE SEPROFIT = -9.
.   ELSE IF DIN21 =-8.
.     COMPUTE SEPROFIT=-8.
.   ELSE.
.     COMPUTE SEPROFIT = -9.
.   END IF.

. END IF /* (DIN20 GE 0 AND DIN20 LE 999997).
```

```
END IF /*DVAGE LT 16.

compute seprofix=SEPROFIT.
recode seprofix (100 thru 150000=999)(else=copy).

Variable Label SEPROFIT  'Self employed weekly earnings'.
Variable Label  seprofix 'Self employed weekly earnings - recoded'.

Value label  SEPROFIT seprofix -9 'DNA/CHILD/PROXY/NO INT'
                              -8 'DK'
                              -7 'not WORKING OR NOT self emp'
                               0 'No profit'.
missing values seprofix (-9 -8 -7).
freq vars=seprofix .

recode seprofix (sysmis=111111111).
temp.
select if (seprofix = 111111111) .
list variables = caseid seprofix din1i, din2 to din3b,din4i selfemp din20a to din24.

temp.
select if (seprofix lt 999) or (seprofix gt 999) .
list variables = caseid seprofix din1i din2 to din3b din4i din20a to din24.
*******2 cases with earnings  £80k and £90k pa OK *************************.



*************28 cases given gross pay -need to convert to net **************.



*******************************************************************************************
*********************************************
***                          NATIONAL INSURANCE CONTRIBUTIONS - RATES AND THRESHOLDS
                             ***
***                                                  03/04            04/05
                  ***
*** NIC Class 2 Threshold                           £4,095           £4,215
                  ***
*** NIC Class 2 Weekly flat-rate                    £2.00            £2.05
                  ***
***
                  ***
```

*** NIC Class 4 lower threshold £4,615 £4,745
***

*** NIC Class 4 upper threshold £30,940 £31,720
***

*** NIC Class 4 8% 8%
***

***NIC Class 4 rate above upper threshold 1% 1%
***

****
****

****
****

*** TAX RATES AND ALLOWANCES ***

***	03/04	04/05	***
*** Starting Rate	10%	10%	***
*** Starting Rate threshold	1960	2020	***
*** Basic Rate	22%	22%	***
*** Basic rate threshold	30,500	31,400	***
*** Higher Rate	40%	40%	***
***			***
*** Personal allowance	4,615	4,745	***
*** Age related upper threshold	6,610	6,830	***
*** 65-74			***

****
****

```
temp.
select if din22=2.
list variables = caseid startdat dvage seprofit din1i din2 to din3b din4i din20a to din24.

**********calculate tax liability **********.

missing values seprofit (-8,-9).
compute seprofpa=seprofit*52/100.

compute totax=0.
compute totNI=0.
do if din22=2.

. do if seprofpa lt 4615.
```

```
.   compute totax=0.
. else if (seprofpa ge 4615) and (seprofpa lt 6575).
.   compute totax=(seprofpa-4615)*0.1.
. else if (seprofpa ge 6575) and (seprofpa lt 30500).
.   compute totax=((1960)*0.1) + ((seprofpa-6575)*0.22).
. else if (seprofpa ge 30500).
.   compute totax=((1960*0.1) + (23925*0.22) + ((seprofpa-30500)*0.4)).
. end if.

end if /* din22=2.

fre vars = totax.
temp.
select if din22=2.
list variables = caseid startdat totax dvage seprofpa din1i din2 to din3b,din4i din20a to din24.

***** haven't changed tax liability for age 65-74.

**********calculate NI liability **********.

do if din22=2.

. do if seprofpa lt 4095.
.   compute totNI=0.
. else if (seprofpa ge 4095) and (seprofpa lt 4615).
.   compute totNI=(2*52).
. else if (seprofpa ge 4615) and (seprofpa lt 30940).
.   compute totNI=(2*52) + ((seprofpa-4615)*0.08).
. else if (seprofpa ge 30940).
.   compute totNI=(2*52) + (26325*0.08) + ((seprofpa-30940)*0.01).
. end if.

end if /* din22=2.

fre vars = totNI.

temp.
select if din22=2.
list variables = caseid startdat totNI dvage seprofpa din1i din2 to din3b,din4i din20a to din24.
```

```
MISSING VALUES SEPROFIT DIN22().

FRE VARS = DIN22.

compute seprnet=seprofit.
do if (seprofit=-7) .
.  compute seprnet=-7.
else if (seprofit=-8) .
.  compute seprnet=-8.
else if (seprofit=-9) .
.  compute seprnet=-9.
else if (din22=2).
.  compute seprnet=seprofit-totax-totNI.
else if (din22=1).
.  compute seprnet=seprofit.
end if.



Variable Label SEPRnet  'Self employed net weekly earnings'.

Value label  SEPRNET  -9 'DNA/CHILD/PROXY/NO INT'
                      -8 'NA'
                      -7 'not self emp'
                       0 'No profit'.
missing values seprnet (-9 -8 -7).
fre vars = seprnet.

recode seprnet (sysmis=111111111) into seprnx.
temp.
select if (seprnx = 111111111) .
list variables = caseid seprofix din1i din2 to din3b, din4i selfemp din20a to din24.



temp.
select if din22=2.
list variables = caseid startdat seprnet totNI totax dvage seprofpa din1i din2 to din3b, din4i
        din20a to din24.
```

```
recode seprnet (-7=0) (sysmis=-9) (else=copy).
Value label seprnet       -9 'DNA/child/proxy/NO INT/refused'
                              -8 'Don t know'
                               0 'No self emp pay received'.


missing values netpay ntsecjob seprnet ().

do if (din1i=6).
. compute HIHearn=-7.
else if (netpay=-9) or (ntsecjob=-9) or (seprnet=-9).
. compute HIHearn=-9.
else if (netpay=-8) or (ntsecjob=-8) or (seprnet=-8).
. compute HIHearn=-8.
else.
. compute HIHearn=netpay+ntsecjob+seprnet.
end if.


Variable Label HIHearn  HIH net weekly earnings (pence)'.

Value label  HIHearn  -9 'DNA/CHILD/PROXY/NO INT'
                      -8 'NA'
                      -7 'not working'
                       0 'No earnings'.

missing values HIHearn (-9 -8 -7).

fre vars = HIHearn.

recode HIHearn (-7=0) (sysmis=-9).
missing values HIHearn (-9 -8).
fre vars = HIHearn.


temp .
select if HIHearn lt -9.
list variables = caseid HIHearn netpay dvage din4i din5i din9 din13 ntsecjob din27 din30  seprnet selfemp.
```

list variables = HIHearn netpay ntsecjob seprnet /cases = from 1 to 50.

Pre-imputation Code For Earned Income

********** shcearndv_again.sps
**********.

********** EARNINGS INCOME DERIVED VARIABLES**********.

** getting pre-imputed data

** caseid replaced with uprn

** redundant or erroneous variables deleted

** 'i's deleted so that imputed variables are not included and original values can be recorded

**GET

**FILE='N:\SPSSSYSTEM\SOCIAL\S03data_all.sav'.

recode din4 (1 thru 999997=1) (else=copy) into din4test.

missing values din4test ().

fre vars = din4test.

**** cases with DK usual net.

temp.

select if (din4test= 999999).

list variables = uprn din4 to din7, din8a to din16.

**** if DK usual net (Din4) asks usual gross (DIN8) = 3 cases.

**** if DK usual net (Din4) and no usual gross (DIN8), asks for last net (DIN12) = 2 cases + 1 DK .

**** if DK usual net (Din4) no usual gross, dK last net asks for last gross (DIN16) = DK only.

**** check age lt 16.

temp.

select if (dvage lt 16).

list variables = uprn dvage din4 to din7.

```
**** investigate pay periods 1 hour and 1 day.
temp.
select if (din5 = 91 or din5=92).
list variables = uprn din2 din3a din3b din4 din5 din7 din8.
**** only one case, paid £9 per hour but DK how many hours worked in week.

temp.
select if din5=97.
list variables  = uprn stat din1, din2 din3a, din4 to din7,din8a to din19.
*** 4 cases.



*********************************************************************************.
********************************** NETPAY ***************************************.
missing values din4 ().

DO IF DVAGE LT 16.
. COMPUTE NETPAY = -9.
ELSE IF din1 =6.
. COMPUTE NETPAY = -7.
ELSE IF (DIN2=1 AND DIN3a  = 2).
. COMPUTE NETPAY=-7.
ELSE IF (DIN2=2 AND DIN3B  = 2).
. COMPUTE NETPAY=-7.
ELSE IF din4 =999998.
. COMPUTE NETPAY = -9.
ELSE IF (RANGE(din1,1,5)).
. DO IF DIN3a  = 1 or DIN3b = 1.
.   DO IF (din4 = 999999  AND DIN8 GE 0).
.    DO IF RANGE (DIN9,1,4) OR RANGE (DIN9,13,52).
.          COMPUTE NETPAY = DIN8/DIN9 * 100.
.    ELSE IF DIN9 = 5.
.          COMPUTE NETPAY = DIN8 * 12/52 * 100.
.    ELSE IF DIN9 = 7.
.          COMPUTE NETPAY = DIN8 * 6/52 * 100.
.    ELSE IF RANGE (DIN9,8,10).
.          COMPUTE NETPAY = DIN8 *DIN9/52 * 100.
.    ELSE IF RANGE (DIN9,90,92).
.          COMPUTE NETPAY = DIN8 * 100.
```

```
.   ELSE IF RANGE (DIN9,95,99).
.         COMPUTE NETPAY = -8.
.   END IF.
. END IF.

. DO IF (din4 = 999999  AND DIN12 GE 0).
.   DO IF RANGE (DIN13,1,4) OR RANGE (DIN13,13,52).
.         COMPUTE NETPAY = DIN12/DIN13 * 100.
.   ELSE IF DIN13 = 5.
.         COMPUTE NETPAY = DIN12 * 12/52 * 100.
.   ELSE IF DIN13 = 7.
.         COMPUTE NETPAY = DIN12 * 6/52 * 100.
.   ELSE IF RANGE (DIN13,8,10).
.         COMPUTE NETPAY = DIN12 *DIN13/52 * 100.
.   ELSE IF RANGE (DIN13,90,92).
.         COMPUTE NETPAY = DIN12 * 100.
.   ELSE IF RANGE (DIN13,95,99).
.         COMPUTE NETPAY = -8.
.   END IF.
. END IF.

. DO IF (din4 = 999999  AND DIN16 GE 0).
.   DO IF RANGE (DIN17,1,4) OR RANGE (DIN17,13,52).
.         COMPUTE NETPAY = DIN16/DIN17 * 100.
.   ELSE IF DIN17 = 5.
.         COMPUTE NETPAY = DIN16 * 12/52 * 100.
.   ELSE IF DIN17 = 7.
.         COMPUTE NETPAY = DIN16 * 6/52 * 100.
.   ELSE IF RANGE (DIN17,8,10).
.         COMPUTE NETPAY = DIN16 *DIN17/52 * 100.
.   ELSE IF RANGE (DIN17,90,92).
.         COMPUTE NETPAY = DIN16 * 100.
.   ELSE IF RANGE (DIN17,95,99).
.         COMPUTE NETPAY = -8.
.   END IF.
. END IF.

. DO IF (din4 = 999999 AND DIN8a NE 1).
.   COMPUTE NETPAY = -8.
. END IF.
```

```
.  DO IF (din4 GE 0).
.   DO IF RANGE (din5,1,4) OR RANGE (din5,13,52).
.          COMPUTE NETPAY = din4/din5 * 100.
.   ELSE IF din5 = 5.
.          COMPUTE NETPAY = din4 * 12/52 * 100.
.   ELSE IF din5 = 7.
.          COMPUTE NETPAY = din4 * 6/52 * 100.
.   ELSE IF RANGE (din5,8,10).
.          COMPUTE NETPAY = din4 *din5/52 * 100.
.   ELSE IF RANGE (din5,90,92).
.          COMPUTE NETPAY = din4 * 100.
.   ELSE IF RANGE (din5,95,99).
.          COMPUTE NETPAY = -8.
.   ELSE IF din4 = -9 AND din5 LT 90.
.          COMPUTE NETPAY = -9.
.   ELSE IF din5 =-8.
.          COMPUTE NETPAY=-8.
.   ELSE.
.          COMPUTE NETPAY = -9.
.   END IF.
.  END IF.

. END IF /* DIN3a  = 1 or DIN3b = 1.

END IF /* DVAGE LT 16.

temp.
select if (din4 = 999999  AND DIN8 GE 0).
list variables = uprn din4 din8 din9 netpay.

compute netpayx=netpay.

fre vars =netpayx.

recode netpayx (300 thru 150000=999)(SYSMIS=11111111) (else=copy).
VAR LABEL NETPAY  'Usual net weekly pay - employees (pence per wk)'
          /NETPAYx 'Usual net weekly pay - employees (pence per wk) - recoded'.
Value label NETPAY NETPAYx      -9 'DNA/child/proxy/NO INT/refused'
                                -8 'Don t know'
```

```
                        -7 'not working'
                         0 'No pay received'.

missing values netpayx (-9 -8 -7).
freq vars=netpayx.

temp.
select if (NETPAYx=11111111).
list variables = uprn din1, din2 to din3b, din4 to din7, din8a TO din9 netpay.

temp.
select if netpayx ne  999.
list variables = uprn netpayx din1, din2 to din3a din3b, din4 to din7, din8a TO din19.

**** 5 cases earn more than £80k - fine.

missing values netpayx ().

temp.
select if (netpayx=-8 or netpayx=-9).
list variables = uprn netpayx din1, din2 to din3a din3b, din4 to din7, din8a TO din19.

*** cases with amounts but period = 95 or 97 will need imputing ***.

missing values netpayx (-9 -8 -7).

recode netpay (-7=0) (sysmis=-9) (else=copy).
Value label NETPAY      -9 'DNA/child/proxy/NO INT/refused'
                        -8 'Don t know'
                         0 'No pay received'.

*********************************************************************************.
******************** GHS Asks about bonuses but SHC doesn't so section deleted***********.

fre vars = din2 din3b din27 to din32.

*********************************************************************************.
***************** NTSECJOB ********************************.
missing values din27 din30 ().
```

```
DO IF DVAGE LT 16.
+ COMPUTE NTSECJOB = -9.
ELSE IF din1 = 6.
+ COMPUTE NTSECJOB = -7.
ELSE IF (DIN2=1).
+ COMPUTE NTSECJOB = -7.
ELSE IF DIN27 =999998.
+ COMPUTE NTSECJOB= -9.
ELSE IF DIN3b=2.
+ COMPUTE NTSECJOB = -7.
ELSE IF DIN3b = 1.
+ COMPUTE NTSECJOB = 0.

. DO IF (DIN27 = 999999  AND DIN30 = 999999).
.   COMPUTE NTSECJOB = -8.
. END IF.

. DO IF (DIN27 = 999999  AND DIN30 GE 0).

.   DO IF RANGE (DIN31,1,4) OR RANGE (DIN31,13,52).
.     COMPUTE NTSECJOB = DIN30/DIN31 * 100.
.   ELSE IF DIN31 = 5.
.     COMPUTE NTSECJOB = DIN30 * 12/52 * 100.
.   ELSE IF DIN31 = 7.
.     COMPUTE NTSECJOB = DIN30 * 6/52 * 100.
.   ELSE IF RANGE (DIN31,8,10).
.     COMPUTE NTSECJOB = DIN30 *DIN31/52 * 100.
.   ELSE IF RANGE (DIN31,90,92).
.     COMPUTE NTSECJOB = DIN30 * 100.
.   ELSE IF RANGE (DIN31,95,99).
.     COMPUTE NTSECJOB = -8.
.   END IF.

. END IF  /* (DIN27 = 999999  AND DIN30 GE 0).

. DO IF (DIN27 GE 0).

.   DO IF RANGE (DIN28,1,4) OR RANGE (DIN28,13,52).
.     COMPUTE NTSECJOB = DIN27/DIN28 * 100.
.   ELSE IF DIN28 = 5.
```

```
.    COMPUTE NTSECJOB = DIN27 * 12/52 * 100.
.  ELSE IF DIN28 = 7.
.    COMPUTE NTSECJOB = DIN27 * 6/52 * 100.
.  ELSE IF RANGE (DIN28,8,10).
.    COMPUTE NTSECJOB = DIN27 *DIN28/52 * 100.
.  ELSE IF RANGE (DIN28,90,92).
.    COMPUTE NTSECJOB= DIN27 * 100.
.  ELSE IF RANGE (DIN28,95,99).
.    COMPUTE NTSECJOB = -8.
.  ELSE IF DIN27 = -9 AND DIN28 LT 90.
.    COMPUTE NTSECJOB = -9.
.  ELSE IF DIN28 =-8.
.    COMPUTE NTSECJOB=-8.
.  ELSE.
.    COMPUTE NTSECJOB = -9.
.  END IF.

. END IF /* (DIN27 GE 0).

END IF /* DVAGE LT 16.

compute ntsecjox=ntsecjob.
recode ntsecjox (50 thru 75000=999)(else=copy).

VAR LABELS NTSECJOB 'Net weekly - other jobs (pence per wk)'.
VAR LABELS NTSECJOx 'Net weekly - other jobs (pence per wk) - recoded'.

Value label  NTSECJOB  ntsecjox -9 'DNA/CHILD/PROXY/NO INT/refused'
                                 -8 'DK'
                                 -7 'not working/only 1 job/selfemp'
                                  0 'No earnings'.
missing values  ntsecjox (-9 -8 -7).
freq vars= ntsecjox.

missing values ntsecjox ().

temp.
select if (ntsecjox = -8 or ntsecjox= -9).
list variables = uprn ntsecjob dvage din1, din2, din3a, din3b,din4 din27 to din32.
```

```
recode ntsecjob (-7=0) (sysmis=-9) (else=copy).
Value label ntsecjob       -9 'DNA/child/proxy/NO INT/refused'
                            -8 'Don t know'
                             0 'No 2nd pay received'.

*************************************************************************************
*******.
********************************** SEPROFIT
*********************************************.

**** includes Self employed in 2nd job *****************.

fre vars = din3a din3b din20a to din24.

missing values din1 din20 DIN24 ().

compute selfemp = 2.
DO IF (DIN3a=2) or (DIN3b=2).
.            COMPUTE selfemp=1.
END IF.

VAR LABEL selfemp  'whether self-employed'.
Value label selfemp 1 'yes'   2 'no'.

fre vars = selfemp.

DO IF DVAGE LT 16.
. COMPUTE SEPROFIT = -9.
ELSE IF din1 = 8.
. COMPUTE SEPROFIT = -9.
ELSE IF din1 = 6.
. COMPUTE SEPROFIT = -7.
ELSE IF (selfemp=2).
. COMPUTE SEPROFIT = -7.
ELSE IF (DIN20a = 2  AND (DIN24 GE 0 and din24 LE 99997)).
. COMPUTE SEPROFIT = DIN24/52 * 100.
ELSE IF (DIN20a = 2  and DIN24=999999).
. COMPUTE SEPROFIT = -8.
ELSE IF (DIN20a=3).
. COMPUTE SEPROFIT = -8.
```

```
ELSE IF (DIN20a=4 OR DIN20=999998).
. COMPUTE SEPROFIT = -9.
ELSE IF (RANGE(din1,1,5)) AND (selfemp=1).
. COMPUTE SEPROFIT = 0.

. DO IF (DIN20 = 999999).
.   COMPUTE SEPROFIT = -8.
. END IF.

. DO IF (DIN20 GE 0 AND DIN20 LE 999997).

.   DO IF RANGE (DIN21,1,4) OR RANGE (DIN21,13,52).
.     COMPUTE SEPROFIT = DIN20/DIN21 * 100.
.   ELSE IF DIN21 = 5.
.     COMPUTE SEPROFIT = DIN20 * 12/52 * 100.
.   ELSE IF DIN21 = 7.
.     COMPUTE SEPROFIT = DIN20 * 6/52 * 100.
.   ELSE IF RANGE (DIN21,8,10).
.     COMPUTE SEPROFIT = DIN20 *DIN21/52 * 100.
.   ELSE IF RANGE (DIN21,90,92).
.     COMPUTE SEPROFIT = DIN20 * 100.
.   ELSE IF RANGE (DIN21,95,99).
.     COMPUTE SEPROFIT = -8.
.   ELSE IF DIN20 = -9 AND DIN21 LT 90.
.     COMPUTE SEPROFIT = -9.
.   ELSE IF DIN21 =-8.
.     COMPUTE SEPROFIT=-8.
.   ELSE.
.     COMPUTE SEPROFIT = -9.
.   END IF.

. END IF /* (DIN20 GE 0 AND DIN20 LE 999997).

END IF /*DVAGE LT 16.

compute seprofix=SEPROFIT.
recode seprofix (100 thru 150000=999)(else=copy).

Variable Label SEPROFIT  'Self employed weekly earnings'.
Variable Label  seprofix 'Self employed weekly earnings - recoded'.
```

```
Value label  SEPROFIT seprofix -9 'DNA/CHILD/PROXY/NO INT'
                         -8 'DK'
                         -7 'not WORKING OR NOT self emp'
                          0 'No profit'.
missing values seprofix (-9 -8 -7).
freq vars=seprofix .

recode seprofix (sysmis=111111111).
temp.
select if (seprofix = 111111111) .
list variables = uprn seprofix din1, din2 to din3b,din4 selfemp din20a to din24.

temp.
select if (seprofix lt 999) or (seprofix gt 999) .
list variables = uprn seprofix din1 din2 to din3b din4 din20a to din24.
*******2 cases with earnings  £80k and £90k pa OK ************************.


*************28 cases given gross pay -need to convert to net **************.


*******************************************************************************************
*********************************************
***                     NATIONAL INSURANCE CONTRIBUTIONS - RATES AND THRESHOLDS
                        ***
***                                                 03/04           04/05
                  ***
*** NIC Class 2 Threshold                          £4,095          £4,215
                  ***
*** NIC Class 2 Weekly flat-rate                   £2.00           £2.05
                  ***
***
                  ***
*** NIC Class 4  lower threshold                   £4,615          £4,745
                  ***
*** NIC Class 4  upper threshold                   £30,940         £31,720
                  ***
*** NIC Class 4                                    8%              8%
                  ***
***NIC  Class 4 rate above upper threshold  1%                1%
          ***
```

```
*******************************************************************************************
*********************************************

*******************************************************************************************
*********************************************
***                          TAX RATES AND ALLOWANCES                          ***
***                          03/04          04/05                              ***
*** Starting Rate            10%            10%                                ***
*** Starting Rate threshold  1960           2020                               ***
*** Basic Rate               22%            22%                                ***
*** Basic rate threshold     30,500         31,400                             ***
*** Higher Rate              40%            40%                                ***
***                                                                            ***
*** Personal allowance       4,615          4,745                              ***
*** Age related upper threshold 6,610       6,830                              ***
***       65-74                                                                ***
*******************************************************************************************
*********************************************

temp.
select if din22=2.
list variables = uprn startdat dvage seprofit din1 din2 to din3b din4 din20a to din24.

**********calculate tax liability **********.

missing values seprofit (-8,-9).
compute seprofpa=seprofit*52/100.

compute totax=0.
compute totNI=0.
do if din22=2.

. do if seprofpa lt 4615.
.   compute totax=0.
. else if (seprofpa ge 4615) and (seprofpa lt 6575).
.   compute totax=(seprofpa-4615)*0.1.
. else if (seprofpa ge 6575) and (seprofpa lt 30500).
.   compute totax=((1960)*0.1) + ((seprofpa-6575)*0.22).
. else if (seprofpa ge 30500).
.   compute totax=((1960*0.1) + (23925*0.22) + ((seprofpa-30500)*0.4)).
```

```
. end if.

end if /* din22=2.

fre vars = totax.
temp.
select if din22=2.
list variables = uprn startdat totax dvage seprofpa din1 din2 to din3b,din4 din20a to din24.

***** haven't changed tax liability for age 65-74.

**********calculate NI liability **********.

do if din22=2.

. do if seprofpa lt 4095.
.   compute totNI=0.
. else if (seprofpa ge 4095) and (seprofpa lt 4615).
.   compute totNI=(2*52).
. else if (seprofpa ge 4615) and (seprofpa lt 30940).
.   compute totNI=(2*52) + ((seprofpa-4615)*0.08).
. else if (seprofpa ge 30940).
.   compute totNI=(2*52) + (26325*0.08) + ((seprofpa-30940)*0.01).
. end if.

end if /* din22=2.

fre vars = totNI.

temp.
select if din22=2.
list variables = uprn startdat totNI dvage seprofpa din1 din2 to din3b,din4 din20a to din24.


MISSING VALUES SEPROFIT DIN22().

FRE VARS = DIN22.

compute seprnet=seprofit.
do if (seprofit=-7) .
```

```
.  compute seprnet=-7.
else if (seprofit=-8) .
.  compute seprnet=-8.
else if (seprofit=-9) .
.  compute seprnet=-9.
else if (din22=2).
.  compute seprnet=seprofit-totax-totNI.
else if (din22=1).
.  compute seprnet=seprofit.
end if.



Variable Label SEPRnet  'Self employed net weekly earnings'.

Value label  SEPRNET  -9 'DNA/CHILD/PROXY/NO INT'
                      -8 'NA'
                      -7 'not self emp'
                       0 'No profit'.
missing values seprnet (-9 -8 -7).
fre vars = seprnet.

recode seprnet (sysmis=111111111) into seprnx.
temp.
select if (seprnx = 111111111) .
list variables = uprn seprofix din1 din2 to din3b, din4 selfemp din20a to din24.



temp.
select if din22=2.
list variables = uprn startdat seprnet totNI totax dvage seprofpa din1 din2 to din3b, din4
          din20a to din24.



recode seprnet (-7=0) (sysmis=-9) (else=copy).
Value label seprnet      -9 'DNA/child/proxy/NO INT/refused'
                          -8 'Don t know'
                           0 'No self emp pay received'.
```

```
missing values netpay ntsecjob seprnet ().

do if (din1=6).
. compute HIHearn=-7.
else if (netpay=-9) or (ntsecjob=-9) or (seprnet=-9).
. compute HIHearn=-9.
else if (netpay=-8) or (ntsecjob=-8) or (seprnet=-8).
. compute HIHearn=-8.
else.
. compute HIHearn=netpay+ntsecjob+seprnet.
end if.


Variable Label HIHearn  HIH net weekly earnings (pence)'.

Value label  HIHearn  -9 'DNA/CHILD/PROXY/NO INT'
                       -8 'NA'
                       -7 'not working'
                        0 'No earnings'.

missing values HIHearn (-9 -8 -7).

fre vars = HIHearn.

recode HIHearn (-7=0) (sysmis=-9).
missing values HIHearn (-9 -8).
fre vars = HIHearn.


temp .
select if HIHearn lt -9.
list variables = uprn HIHearn netpay dvage din4 din5 din9 din13 ntsecjob din27 din30  seprnet selfemp.

list variables = HIHearn netpay ntsecjob seprnet /cases = from 1 to 50.
```

Appendix 2: Principal Component Analysis

Principal Component Analysis SAS Code Using the Polyhcor Macro

```
libname impute spss 'E:\origfin.POR';
/*invokes spss portable file for conversion*/

data orig;
        set impute.origfin;
run;
/*converst file into spss format to be stored in temporary library*/


%inc "C:\Documents and Settings\G\Desktop\imputation\polychor.sas";
/*invokes stored polychor macro*/
%polyhcor
(var = ua urbanind dvhsize sex dhc8a dvmardf dv9age1 workage numadult
numadul2 numchild numchil2 numssex numcpart nummpart numhhldr numch18
numprima relsize dvhrpnum hrppart dt8 keyhh keyhhpt respdnt respsex respmar
partner partno sspart isndep numdepch haschd hasdep hasndep nchild ndepc nndepc
nbaby nc5und1 nc5to9_1 nc1015_1 ncu16_1 nc1618_1 singpar dt1 dhc12 ethnic
dt13 dt25 dt42 dt43 dt44 dt45 dt50 dt52 dt54a dt60 df1 df6 dr18a dr18b dr18d dr20
dr23 dpa1 dpa3 dpa4 dpa8 dh2 dh5a dh11 dh13 dhe1 wrking dvilo4a benintro dst3a
dst1 db15 recall2 dwelladd nounits nssec8 proptyp floors dweltyp othhome lgtres
hout dhc13_1 nation_1 nation_2 nation_3 nation_4 nation_5 nation_6 dt47_1 dt47_2
dt47_3 dt47_4 dt47_5 dt47_6 dt47_7 dt47_8 dt47_9 dt47_10 dt47_11 dt47_12 dt47_13
dt47_14 dt47_15 dt47_16 dt47_17 dt47_18 dt47_19 dt47_20 dt47_21 dt47_22 dt51_1
dt56_1 dt56_2 dt56_3 dt56_4 dt56_5 dt57_1 dt58_1 dt58_2 dt58_3 dt58_4 dt58_5 dt58_6
dt58_7 dt58_8 dt58_9 dt67_1 dt67_2 dt67_3 dt67_4 dt67_5 dt67_6 dt67_7 dt67_8 dt67_9
dt67_10 dt67_11 dt67_12 dt67_13 dt67_14 dt67_15 dt67_16 dt67_17 dt67_19 dt67_20
dt67_21 dc1_1 dc1_2 dc1_3 dc1_4 dc1_5 dc1_11 dc8a_1 dc8a_2 dc8a_3 dc8a_4 dc8a_5
dc8a_7 dc8a_8 dc8a_9 dc8a_10 dc8a_11 dc8a_12 dc8a_13 dc8a_14 dc8a_15 dc8a_16
dc10_1 dc10_2 dc10_4 dc10_6 dc10_7 dc10_8 dc10_9 dc10_10 db16_1 db16_2 db16_3
db16_4 db16_5 db16_7 db16_8 db16_10 dr2_1 dr2_2 dr2_3 dr2_4 dr2_6 dr2_7 dr2_8 dr2_9
dr2_10 dr2_11 dr2_12 dr2_13 dr2_14 dr2_15 dr2_16 dr2_17 dr2_18 dr2_19 dr2_20 dr2_21
dr2_22 dr2_23 dr2_24 dr2_25 dr2_26 dr2_27 dr2_28 dr2_29 dr2_30 dr2_31 dr2_35 dh1_3
```

```
dh1_4 dh1_5 dh1_6 dh1_7 dh1_8 dh1_9 dh1_10 dh5_2 dh5_3 dh5_4 dh5_5 dh5_6 dh5_7
dh5_8 dh14_1 dh14_2 dh14_3 dh14_4 dh14_5 dh14_6 dh14_7 dh14_8 dh14_9 dh14_10
dh14_11 dh14_12 dh14_13 dh14_14 dh14_15 dh14_16 dh14_17 ntsecjox selfemp spouse
hhx hhz hih tenure outten rtbten hihsex hihageg partageg pension hhtype under5 sixtypls
health factor);
/*specifies variables for macros*/
proc print noobs;
run;

proc factor method = prin scree;
/*principal component with scree plot requested*/
var /*variables for analysis specified*/
ua urbanind dvhsize sex dhc8a dvmardf dv9age1 workage numadult
numadul2 numchild numchil2 numssex numcpart nummpart numhhldr numch18
numprima relsize dvhrpnum hrppart dt8 keyhh keyhhpt respdnt respsex respmar
partner partno sspart isndep numdepch haschd hasdep hasndep nchild ndepc nndepc
nbaby nc5und1 nc5to9_1 nc1015_1 ncu16_1 nc1618_1 singpar dt1 dhc12 ethnic
dt13 dt25 dt42 dt43 dt44 dt45 dt50 dt52 dt54a dt60 df1 df6 dr18a dr18b dr18d dr20
dr23 dpa1 dpa3 dpa4 dpa8 dh2 dh5a dh11 dh13 dhe1 wrking dvilo4a benintro dst3a
dst1 db15 recall2 dwelladd nounits nssec8 proptyp floors dweltyp othhome lgtres
hout dhc13_1 nation_1 nation_2 nation_3 nation_4 nation_5 nation_6 dt47_1 dt47_2
dt47_3 dt47_4 dt47_5 dt47_6 dt47_7 dt47_8 dt47_9 dt47_10 dt47_11 dt47_12 dt47_13
dt47_14 dt47_15 dt47_16 dt47_17 dt47_18 dt47_19 dt47_20 dt47_21 dt47_22 dt51_1
dt56_1 dt56_2 dt56_3 dt56_4 dt56_5 dt57_1 dt58_1 dt58_2 dt58_3 dt58_4 dt58_5 dt58_6
dt58_7 dt58_8 dt58_9 dt67_1 dt67_2 dt67_3 dt67_4 dt67_5 dt67_6 dt67_7 dt67_8 dt67_9
dt67_10 dt67_11 dt67_12 dt67_13 dt67_14 dt67_15 dt67_16 dt67_17 dt67_19 dt67_20
dt67_21 dc1_1 dc1_2 dc1_3 dc1_4 dc1_5 dc1_11 dc8a_1 dc8a_2 dc8a_3 dc8a_4 dc8a_5
dc8a_7 dc8a_8 dc8a_9 dc8a_10 dc8a_11 dc8a_12 dc8a_13 dc8a_14 dc8a_15 dc8a_16
dc10_1 dc10_2 dc10_4 dc10_6 dc10_7 dc10_8 dc10_9 dc10_10 db16_1 db16_2 db16_3
db16_4 db16_5 db16_7 db16_8 db16_10 dr2_1 dr2_2 dr2_3 dr2_4 dr2_6 dr2_7 dr2_8 dr2_9
dr2_10 dr2_11 dr2_12 dr2_13 dr2_14 dr2_15 dr2_16 dr2_17 dr2_18 dr2_19 dr2_20 dr2_21
dr2_22 dr2_23 dr2_24 dr2_25 dr2_26 dr2_27 dr2_28 dr2_29 dr2_30 dr2_31 dr2_35 dh1_3
dh1_4 dh1_5 dh1_6 dh1_7 dh1_8 dh1_9 dh1_10 dh5_2 dh5_3 dh5_4 dh5_5 dh5_6 dh5_7
dh5_8 dh14_1 dh14_2 dh14_3 dh14_4 dh14_5 dh14_6 dh14_7 dh14_8 dh14_9 dh14_10
dh14_11 dh14_12 dh14_13 dh14_14 dh14_15 dh14_16 dh14_17 ntsecjox selfemp spouse
hhx hhz hih tenure outten rtbten hihsex hihageg partageg pension hhtype under5 sixtypls
health factor;
```

```
run;
```

Appendix 3: EM Algorithm SAS Code

```
libname impute spss 'E:\origfin.POR';
/*invokes portable spss file*/
data orig;
        set impute.origfin;
run;
/*conversion into sas format*/
/*em algorithm for MLE*/
proc mi data = orig seed=55417 simple nimpute =0;
/*seed value specified to enable replication of results*/
/*nimpute=0 so that only parameters are estimated*/
em itprint outem=outem;
var annhhinc wkhhinc incband benintro wrking dt25 respdnt haschd dt45;
/*variables for analysis*/
run;
```

Appendix 4: Multinormality Test For Dataset With Relevant Variables

(Chisquare Plot of Square Distances Before and After Logarithmic Transformations of Variables)

```
libname new 'C:\Documents and Settings\G\Desktop\NEW';
/*Defines library in which file is stored*/

data new.origtran (keep = annhhinc wkhhinc incband respdnt haschd dt25 dt45 wrking
logwkhhinc logannhhinc logincband
logrespdnt loghaschd logdt25 logdt45 logwrking);
/*subsets relevant variables*/
        set impute.origfin;
        if annhhinc = . then delete;
        if wkhhinc = . then delete;
        if incband = . then delete;
        if respdnt = . then delete;
        if haschd = . then delete;
        if dt25 = . then delete;
        if dt45 = . then delete;
        if wrking = . then delete;
/*deletes missing values from variables*/
        logwkhhinc = log(wkhhinc);
        logannhhinc = log(annhhinc);
        logincband = log(incband);
        logrespdnt = log(respdnt);
        loghaschd = log(haschd);
        logdt25 = log(dt25);
        logdt45 = log(dt45);
        logwrking = log(wrking);
/*defines new logtransformations of variables*/
run;

proc univariate normal plot;
var annhhinc wkhhinc respdnt incband haschd dt25 dt45 wrking logannhhinc logwkhhinc
logincband logrespdnt loghaschd logdt25 logdt45 logwrking;
```

```
run;
/*checks individual distributions of variables*/

proc princomp data = new.origtran cov std out = b noprint;
var logannhhinc logwkhhinc logincband logrespdnt loghaschd logdt25 logdt45 logwrking;
data chiq;
set b;
dsq=uss (of prin1-prin8);
proc sort;
by dsq;
proc means noprint;
var dsq;
output out=chiqn n=totn;
data chiqq;
if (_n_=1) then set chiqn;
set chiq;
chisq = cinv(((_n_-.5)/totn),4);
symbol1 v=circle c=black;
symbol2 i=join c=black;
proc gplot;
plot (dsq chisq)*chisq/overlay;
run;
/*Chi-square plot of square distances created*/

data origtra (keep = wkhhinc annhhinc incband respdnt dt25 dt45 wrking);
set new.origtran;
run;


proc princomp data = origtra cov std out = be noprint;
var logannhhinc logwkhhinc logincband logrespdnt loghaschd logdt25 logdt45 logwrking;
data chiq;
set be;
dsq=uss (of prin1-prin8);
proc sort;
```

```
by dsq;
proc means noprint;
var dsq;
output out=chiqn n=totn;
data chiqq;
if (_n_=1) then set chiqn;
set chiq;
chisq = cinv(((_n_-.5)/totn),4);
symbol1 v=circle c=black;
symbol2 i=join c=black;
proc gplot;
plot (dsq chisq)*chisq/overlay;
run;
/*Chi-square plot of squared distances created in similar manner as above but this time for transformed variables*/
```

Appendix 5: Final SAS Code For Intermediary Multiple Regression Model

```
libname new spss 'E:\ctsmaller.por';
/*invoking portable spss file into sas*/

data ctsmall;
set new.ctsm;
/*conversion into sas file*/
dt251 = dt25 =2;
dt252 = dt25 = 3;
dt253 = dt25 = 4;
dt254 = dt25 = 5;
dt255 = dt25 = 6;
haschd1 = haschd = 2;
db16_11 = db16_1 = 1;
db16_71 = db16_7 = 1;
/*dummy variables created since these are categorical variables*/
run;

proc reg data = ctsmall;
/*regression procedure*/
model wkhhinc = dt251 dt252 dt253 dt254 dt255 haschd1 db16_11 db16_71/vif collin;
/*model statement with dependent variable and independent predictor variables, together with variance inflation information and multicollinearity requested*/
run;
```

Appendix 6: Final SAS Code For Intermediary Logistic Regression Model

```
libname impute spss 'E:\origfin.POR';
/*invoking portable SPSS file into SAS*/
data orig;
        set impute.origfin;
if numhhldr = 4 then numhhldr= 3;
if respdnt = 4 then respdnt = 3;
if respdnt = 3 then respdnt = 2;
/*collapsing categories with few values to plot confidence intervals plot*/
run;
/*conversion of file into SAS*/

proc logistic data = orig descending;
/*logistic regression option with descending option for easier odds ratio interpretation*/
class dt25 haschd numhhldr respdnt/ param = ref ref=first;
/*categorical variables inserted with specification of parameters for comparison*/
model m = dt25 haschd numhhldr respdnt / expb outroc=roc1;
/*model statement with m representing the missingness; odds ratio specified by
exponentiating b; roc curve specified*/
symbol1 i=join v=none c=blue;
proc gplot data=roc1;
title 'ROC Curve';
plot _sensit_ * _1mspec_=1/ vaxis = 0 to 1 by .1 cframe =ligr;
run;
/*roc curve requested*/
```

Appendix 7: Simple Hot Deck Imputation Code

```
data simple (keep = respmar respdnt nbaby haschd wkhhinc);
 set impute.origfin;
run;
/*conversion of file into sas format*/

data hotdeck_simple;
set table1 (where = (wkhhinc ~ = .) );
run;
/*subsetting data where income values are missing*/


proc surveyselect data = hotdeck_simple method =URS sampsize=3870 rep = 1 seed = 12345 out = hotde;
id _all_;
run;
/*macro-like method specifying sampling with replacement and seed value for replication purposes*/


data hotd;
set hotde;
do i = 1 to numberhits;
r = uniform (0);
/*uniform distribution selected for randomising selection*/
output;
end;
rename wkhhinc = wkhhinc_imp;
/*renaming variable to differentiate those that were hotdeck imputed from the original data*/
keep r wkhhinc;
run;

proc sort data = hotr;
by r;
run;
/*sorting data so it can be merged */

data mergeh;
merge ctsmall hotr;
if wkhhinc ~=. then wkhhinc_imp = wkhhinc;
```

```
run;

proc sort data=mergeh;
by ctaxb;

proc means mean std;
var wkhhinc_imp ;
by ctaxb;
run;
```

/*sorting and generating means for each council tax band to produce the output for the results*/

Appendix 8: Markov Chain Monte Carlo Method SAS Code

```
libname impute spss 'E:\origfin.POR';
/*invoking spss portable file into SAS*/
data orig;
        set impute.origfin;
run;
/*conversion of file into SAS*/

/*mcmc method*/
proc mi data = orig seed = 55417 nimpute = 5 out=origout;
/*five datasets generated; seed value specification enables replication; out dataset prevents
alteration of original dataset*/
mcmc chain=multiple displayinit initial=em(itprint);
/*multiple chain method specified*/
var annhhinc incband wkhhinc respdnt haschd dt25 dt45 wrking;
/*variables specified*/
run;

/*checking convergence of model*/
proc mi data = orig seed = 55417 noprint nimpute =5;
mcmc timeplot(mean(wkhhinc)) acfplot(mean(wkhhinc));
/*time plot and autocorrelation function plots generated to check convergence*/
var annhhinc incband wkhhinc respdnt haschd dt25 dt45 wrking;
run;
```

Appendix 9: Cross Tabulations of Pre-imputed Missing Data to Determine Missing Data Mechanism

Frequency Percent Row Pct Col Pct	Table of M by NUMADULT						
M	**NUMADULT(NUMBER OF ADULTS IN HOUSEHOLD)**						**Total**
	1	**2**	**3**	**4**	**5**	**6**	
income	1217 31.45 33.74 88.64	1824 47.13 50.57 95.10	415 10.72 11.51 98.34	127 3.28 3.52 95.49	18 0.47 0.50 100.00	6 0.16 0.17 100.00	3607 93.20
missing income	156 4.03 59.32 11.36	94 2.43 35.74 4.90	7 0.18 2.66 1.66	6 0.16 2.28 4.51	0 0.00 0.00 0.00	0 0.00 0.00 0.00	263 6.80
Total	1373 35.48	1918 49.56	422 10.90	133 3.44	18 0.47	6 0.16	3870 100.00

Statistics for Table of M by NUMADULT

Statistic	DF	Value	Prob
Chi-Square	5	76.4953	<.0001
Likelihood Ratio Chi-Square	5	79.4602	<.0001
Mantel-Haenszel Chi-Square	1	58.1184	<.0001
Phi Coefficient		0.1406	
Contingency Coefficient		0.1392	
Cramer's V		0.1406	

Frequency Percent Row Pct Col Pct	Table of M by NUMCHILD							
M	**NUMCHILD(NUMBER OF CHILDREN IN HOUSEHOLD)**							**Total**
	0	**1**	**2**	**3**	**4**	**5**	**6**	
income	2601 67.21 72.11 91.71	465 12.02 12.89 96.88	401 10.36 11.12 98.28	116 3.00 3.22 96.67	20 0.52 0.55 95.24	4 0.10 0.11 100.00	0 0.00 0.00 0.00	3607 93.20
missing income	235 6.07 89.35 8.29	15 0.39 5.70 3.13	7 0.18 2.66 1.72	4 0.10 1.52 3.33	1 0.03 0.38 4.76	0 0.00 0.00 0.00	1 0.03 0.38 100.00	263 6.80
Total	2836 73.28	480 12.40	408 10.54	120 3.10	21 0.54	4 0.10	1 0.03	3870 100.00

Statistics for Table of M by NUMCHILD

Statistic	DF	Value	Prob
Chi-Square	6	53.1972	<.0001
Likelihood Ratio Chi-Square	6	54.1067	<.0001
Mantel-Haenszel Chi-Square	1	26.3035	<.0001
Phi Coefficient		0.1172	
Contingency Coefficient		0.1164	
Cramer's V		0.1172	

WARNING: 36% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

Frequency Percent Row Pct Col Pct	Table of M by TENURE				
M	**TENURE**				**Total**
	owner-occupier	**LA/ other public**	**HA/co-op**	**private-rented**	
income	2458 63.51 68.15 96.09	624 16.12 17.30 88.01	281 7.26 7.79 82.16	244 6.30 6.76 93.49	3607 93.20
missing income	100 2.58 38.02 3.91	85 2.20 32.32 11.99	61 1.58 23.19 17.84	17 0.44 6.46 6.51	263 6.80
Total	2558 66.10	709 18.32	342 8.84	261 6.74	3870 100.00

Statistics for Table of M by TENURE

Statistic	DF	Value	Prob
Chi-Square	3	129.6803	<.0001
Likelihood Ratio Chi-Square	3	111.2032	<.0001
Mantel-Haenszel Chi-Square	1	59.7679	<.0001
Phi Coefficient		0.1831	
Contingency Coefficient		0.1801	
Cramer's V		0.1831	

Frequency Percent Row Pct Col Pct	Table of M by HIHSEX		
M	**HIHSEX(sex of HIH)**		**Total**
	male	**female**	
income	2169 56.05 60.13 94.10	1438 37.16 39.87 91.88	3607 93.20
missing income	136 3.51 51.71 5.90	127 3.28 48.29 8.12	263 6.80
Total	2305 59.56	1565 40.44	3870 100.00

Statistics for Table of M by HIHSEX

Statistic	DF	Value	Prob
Chi-Square	1	7.2188	0.0072
Likelihood Ratio Chi-Square	1	7.1109	0.0077
Continuity Adj. Chi-Square	1	6.8733	0.0087
Mantel-Haenszel Chi-Square	1	7.2169	0.0072
Phi Coefficient		0.0432	
Contingency Coefficient		0.0431	
Cramer's V		0.0432	

Fisher's Exact Test	
Cell (1,1) Frequency (F)	2169
Left-sided Pr <= F	0.9969
Right-sided Pr >= F	0.0046
Table Probability (P)	0.0015
Two-sided Pr <= P	0.0076

Frequency Percent Row Pct Col Pct	Table of M by HIHAGEG								
M	**HIHAGEG(age of HIH grouped)**								**Total**
	16-24	**25-34**	**35-44**	**45-54**	**55-64**	**65-74**	**75-84**	**85+**	
income	114 2.95 3.16 94.21	458 11.83 12.70 97.24	770 19.90 21.35 96.13	706 18.24 19.57 95.02	652 16.85 18.08 90.93	529 13.67 14.67 91.68	317 8.19 8.79 87.09	61 1.58 1.69 80.26	3607 93.20
missing income	7 0.18 2.66 5.79	13 0.34 4.94 2.76	31 0.80 11.79 3.87	37 0.96 14.07 4.98	65 1.68 24.71 9.07	48 1.24 18.25 8.32	47 1.21 17.87 12.91	15 0.39 5.70 19.74	263 6.80
Total	121 3.13	471 12.17	801 20.70	743 19.20	717 18.53	577 14.91	364 9.41	76 1.96	3870 100.00

Statistics for Table of M by HIHAGEG

Statistic	DF	Value	Prob
Chi-Square	7	76.5359	<.0001
Likelihood Ratio Chi-Square	7	70.8817	<.0001
Mantel-Haenszel Chi-Square	1	60.6757	<.0001
Phi Coefficient		0.1406	
Contingency Coefficient		0.1393	
Cramer's V		0.1406	

Frequency Percent Row Pct Col Pct	Table of M by PARTAGEG									
M	**PARTAGEG(age of HIHs partner or spouse)**									**Total**
	16-24	**25-34**	**35-44**	**45-54**	**55-64**	**65-74**	**75-84**	**85+**	**no partner/spouse**	
income	49 1.27 1.36 100.00	287 7.42 7.96 98.63	523 13.51 14.50 98.31	476 12.30 13.20 96.95	425 10.98 11.78 94.65	264 6.82 7.32 90.72	101 2.61 2.80 92.66	5 0.13 0.14 83.33	1477 38.17 40.95 89.41	3607 93.20
missing income	0 0.00 0.00 0.00	4 0.10 1.52 1.37	9 0.23 3.42 1.69	15 0.39 5.70 3.05	24 0.62 9.13 5.35	27 0.70 10.27 9.28	8 0.21 3.04 7.34	1 0.03 0.38 16.67	175 4.52 66.54 10.59	263 6.80
Total	49 1.27	291 7.52	532 13.75	491 12.69	449 11.60	291 7.52	109 2.82	6 0.16	1652 42.69	3870 100.00

Statistics for Table of M by PARTAGEG

Statistic	DF	Value	Prob
Chi-Square	8	92.7107	<.0001
Likelihood Ratio Chi-Square	8	108.1576	<.0001
Mantel-Haenszel Chi-Square	1	87.7744	<.0001
Phi Coefficient		0.1548	
Contingency Coefficient		0.1530	
Cramer's V		0.1548	

Frequency Percent Row Pct Col Pct	Table of M by PENSION				
M	**PENSION(number of pensionable age householders)**				**Total**
	0.00	**1.00**	**2.00**	**3.00**	
income	2445 63.18 67.78 94.95	741 19.15 20.54 88.96	419 10.83 11.62 91.09	2 0.05 0.06 100.00	3607 93.20
missing income	130 3.36 49.43 5.05	92 2.38 34.98 11.04	41 1.06 15.59 8.91	0 0.00 0.00 0.00	263 6.80
Total	2575 66.54	833 21.52	460 11.89	2 0.05	3870 100.00

Statistics for Table of M by PENSION

Statistic	DF	Value	Prob
Chi-Square	3	39.5513	<.0001
Likelihood Ratio Chi-Square	3	37.0243	<.0001
Mantel-Haenszel Chi-Square	1	24.7497	<.0001
Phi Coefficient		0.1011	
Contingency Coefficient		0.1006	
Cramer's V		0.1011	

WARNING: 25% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

Frequency Percent Row Pct Col Pct	Table of M by HHTYPE								
M	**HHTYPE(Household type)**								**Total**
	single adult	**small adult**	**single parent**	**small family**	**large family**	**large adult**	**older smaller**	**single pensioner run**	
income	499 12.89 13.83 88.79	646 16.69 17.91 95.14	214 5.53 5.93 93.04	516 13.33 14.31 98.66	276 7.13 7.65 98.22	391 10.10 10.84 97.02	561 14.50 15.55 91.82	504 13.02 13.97 86.75	3607 93.20
missing income	63 1.63 23.95 11.21	33 0.85 12.55 4.86	16 0.41 6.08 6.96	7 0.18 2.66 1.34	5 0.13 1.90 1.78	12 0.31 4.56 2.98	50 1.29 19.01 8.18	77 1.99 29.28 13.25	263 6.80
Total	562 14.52	679 17.55	230 5.94	523 13.51	281 7.26	403 10.41	611 15.79	581 15.01	3870 100.00

Statistics for Table of M by HHTYPE

Statistic	DF	Value	Prob
Chi-Square	7	106.4484	<.0001
Likelihood Ratio Chi-Square	7	114.4271	<.0001
Mantel-Haenszel Chi-Square	1	4.4934	0.0340
Phi Coefficient		0.1658	
Contingency Coefficient		0.1636	
Cramer's V		0.1658	

Frequency Percent Row Pct Col Pct	Table of M by UNDER5		
M	**UNDER5(child under 5 in household)**		**Total**
	No	**Yes**	
income	3240 83.72 89.83 92.70	367 9.48 10.17 97.87	3607 93.20
missing income	255 6.59 96.96 7.30	8 0.21 3.04 2.13	263 6.80
Total	3495 90.31	375 9.69	3870 100.00

Statistics for Table of M by UNDER5

Statistic	DF	Value	Prob
Chi-Square	1	14.2514	0.0002
Likelihood Ratio Chi-Square	1	18.6420	<.0001
Continuity Adj. Chi-Square	1	13.4480	0.0002
Mantel-Haenszel Chi-Square	1	14.2478	0.0002
Phi Coefficient		-0.0607	
Contingency Coefficient		0.0606	
Cramer's V		-0.0607	

Fisher's Exact Test	
Cell (1,1) Frequency (F)	3240
Left-sided Pr <= F	1.745E-05
Right-sided Pr >= F	1.0000
Table Probability (P)	1.286E-05
Two-sided Pr <= P	3.245E-05

Frequency Percent Row Pct Col Pct	Table of M by SIXTYPLS		
M	**SIXTYPLS(person aged 60 or over in household)**		**Total**
	No	**Yes**	
income	2299 59.41 63.74 95.32	1308 33.80 36.26 89.71	3607 93.20
missing income	113 2.92 42.97 4.68	150 3.88 57.03 10.29	263 6.80
Total	2412 62.33	1458 37.67	3870 100.00

Statistics for Table of M by SIXTYPLS

Statistic	DF	Value	Prob
Chi-Square	1	45.0412	<.0001
Likelihood Ratio Chi-Square	1	43.4129	<.0001
Continuity Adj. Chi-Square	1	44.1609	<.0001
Mantel-Haenszel Chi-Square	1	45.0295	<.0001
Phi Coefficient		0.1079	
Contingency Coefficient		0.1073	
Cramer's V		0.1079	

Fisher's Exact Test	
Cell (1,1) Frequency (F)	2299
Left-sided Pr <= F	1.0000
Right-sided Pr >= F	3.347E-11
Table Probability (P)	1.939E-11
Two-sided Pr <= P	6.114E-11

Frequency Percent Row Pct Col Pct	Table of M by HEALTH			
M	**HEALTH(Any long-term sick/disabled in household)**			**Total**
	No	**Yes**	**Unobtainable**	
income	2148 55.50 59.55 94.96	1454 37.57 40.31 90.70	5 0.13 0.14 100.00	3607 93.20
missing income	114 2.95 43.35 5.04	149 3.85 56.65 9.30	0 0.00 0.00 0.00	263 6.80
Total	2262 58.45	1603 41.42	5 0.13	3870 100.00

Statistics for Table of M by HEALTH

Statistic	DF	Value	Prob
Chi-Square	2	27.1847	<.0001
Likelihood Ratio Chi-Square	2	27.0159	<.0001
Mantel-Haenszel Chi-Square	1	16.5460	<.0001
Phi Coefficient		0.0838	
Contingency Coefficient		0.0835	
Cramer's V		0.0838	

WARNING: 33% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

Appendix 10: Cross Tabulations of Post-imputed Missing Data to Determine Missing Data Mechanism

Frequency Percent Row Pct Col Pct	Table of Z by NUMADULT						
Z	**NUMADULT(NUMBER OF ADULTS IN HOUSEHOLD)**						**Total**
	1	**2**	**3**	**4**	**5**	**6**	
income	1240 32.04 34.05 90.31	1834 47.39 50.36 95.62	416 10.75 11.42 98.58	128 3.31 3.51 96.24	18 0.47 0.49 100.00	6 0.16 0.16 100.00	3642 94.11
missing income	133 3.44 58.33 9.69	84 2.17 36.84 4.38	6 0.16 2.63 1.42	5 0.13 2.19 3.76	0 0.00 0.00 0.00	0 0.00 0.00 0.00	228 5.89
Total	1373 35.48	1918 49.56	422 10.90	133 3.44	18 0.47	6 0.16	3870 100.00

Statistics for Table of Z by NUMADULT

Statistic	DF	Value	Prob
Chi-Square	5	61.3778	<.0001
Likelihood Ratio Chi-Square	5	64.5243	<.0001
Mantel-Haenszel Chi-Square	1	47.7967	<.0001
Phi Coefficient		0.1259	
Contingency Coefficient		0.1249	
Cramer's V		0.1259	

Frequency Percent Row Pct Col Pct	Table of Z by NUMCHILD							
Z	**NUMCHILD(NUMBER OF CHILDREN IN HOUSEHOLD)**							**Total**
	0	**1**	**2**	**3**	**4**	**5**	**6**	
income	2636 68.11 72.38 92.95	465 12.02 12.77 96.88	401 10.36 11.01 98.28	116 3.00 3.19 96.67	20 0.52 0.55 95.24	4 0.10 0.11 100.00	0 0.00 0.00 0.00	3642 94.11
missing income	200 5.17 87.72 7.05	15 0.39 6.58 3.13	7 0.18 3.07 1.72	4 0.10 1.75 3.33	1 0.03 0.44 4.76	0 0.00 0.00 0.00	1 0.03 0.44 100.00	228 5.89
Total	2836 73.28	480 12.40	408 10.54	120 3.10	21 0.54	4 0.10	1 0.03	3870 100.00

Statistics for Table of Z by NUMCHILD

Statistic	DF	Value	Prob
Chi-Square	6	44.0377	<.0001
Likelihood Ratio Chi-Square	6	39.8416	<.0001
Mantel-Haenszel Chi-Square	1	17.7508	<.0001
Phi Coefficient		0.1067	
Contingency Coefficient		0.1061	
Cramer's V		0.1067	

WARNING: 36% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

Frequency Percent Row Pct Col Pct	Table of Z by TENURE				
Z	**TENURE**				**Total**
	owner-occupier	**LA/ other public**	**HA/co-op**	**private-rented**	
income	2473 63.90 67.90 96.68	636 16.43 17.46 89.70	287 7.42 7.88 83.92	246 6.36 6.75 94.25	3642 94.11
missing income	85 2.20 37.28 3.32	73 1.89 32.02 10.30	55 1.42 24.12 16.08	15 0.39 6.58 5.75	228 5.89
Total	2558 66.10	709 18.32	342 8.84	261 6.74	3870 100.00

Statistics for Table of Z by TENURE

Statistic	DF	Value	Prob
Chi-Square	3	119.3142	<.0001
Likelihood Ratio Chi-Square	3	101.0511	<.0001
Mantel-Haenszel Chi-Square	1	56.0817	<.0001
Phi Coefficient		0.1756	
Contingency Coefficient		0.1729	
Cramer's V		0.1756	

Frequency Percent Row Pct Col Pct	Table of Z by HIHSEX		
Z	**HIHSEX(sex of HIH)**		**Total**
	male	**female**	
income	2184 56.43 59.97 94.75	1458 37.67 40.03 93.16	3642 94.11
missing income	121 3.13 53.07 5.25	107 2.76 46.93 6.84	228 5.89
Total	2305 59.56	1565 40.44	3870 100.00

Statistics for Table of Z by HIHSEX

Statistic	DF	Value	Prob
Chi-Square	1	4.2375	0.0395
Likelihood Ratio Chi-Square	1	4.1820	0.0409
Continuity Adj. Chi-Square	1	3.9559	0.0467
Mantel-Haenszel Chi-Square	1	4.2364	0.0396
Phi Coefficient		0.0331	
Contingency Coefficient		0.0331	
Cramer's V		0.0331	

Fisher's Exact Test	
Cell (1,1) Frequency (F)	2184
Left-sided Pr <= F	0.9829
Right-sided Pr >= F	0.0239
Table Probability (P)	0.0067
Two-sided Pr <= P	0.0436

Sample Size = 3870

Frequency Percent Row Pct Col Pct	Table of Z by HIHAGEG								
Z	**HIHAGEG(age of HIH grouped)**								**Total**
	16-24	**25-34**	**35-44**	**45-54**	**55-64**	**65-74**	**75-84**	**85+**	
income	114 2.95 3.13 94.21	458 11.83 12.58 97.24	774 20.00 21.25 96.63	711 18.37 19.52 95.69	663 17.13 18.20 92.47	536 13.85 14.72 92.89	325 8.40 8.92 89.29	61 1.58 1.67 80.26	3642 94.11
missing income	7 0.18 3.07 5.79	13 0.34 5.70 2.76	27 0.70 11.84 3.37	32 0.83 14.04 4.31	54 1.40 23.68 7.53	41 1.06 17.98 7.11	39 1.01 17.11 10.71	15 0.39 6.58 19.74	228 5.89
Total	121 3.13	471 12.17	801 20.70	743 19.20	717 18.53	577 14.91	364 9.41	76 1.96	3870 100.00

Statistics for Table of Z by HIHAGEG

Statistic	DF	Value	Prob
Chi-Square	7	67.4360	<.0001
Likelihood Ratio Chi-Square	7	58.6966	<.0001
Mantel-Haenszel Chi-Square	1	47.9588	<.0001
Phi Coefficient		0.1320	
Contingency Coefficient		0.1309	
Cramer's V		0.1320	

Frequency Percent Row Pct Col Pct	Table of Z by PARTAGEG									
Z	**PARTAGEG(age of HIHs partner or spouse)**									**Total**
	16-24	**25-34**	**35-44**	**45-54**	**55-64**	**65-74**	**75-84**	**85+**	**no partner/spouse**	
income	49 1.27 1.35 100.00	287 7.42 7.88 98.63	524 13.54 14.39 98.50	476 12.30 13.07 96.95	429 11.09 11.78 95.55	268 6.93 7.36 92.10	101 2.61 2.77 92.66	5 0.13 0.14 83.33	1503 38.84 41.27 90.98	3642 94.11
missing income	0 0.00 0.00 0.00	4 0.10 1.75 1.37	8 0.21 3.51 1.50	15 0.39 6.58 3.05	20 0.52 8.77 4.45	23 0.59 10.09 7.90	8 0.21 3.51 7.34	1 0.03 0.44 16.67	149 3.85 65.35 9.02	228 5.89
Total	49 1.27	291 7.52	532 13.75	491 12.69	449 11.60	291 7.52	109 2.82	6 0.16	1652 42.69	3870 100.00

Statistics for Table of Z by PARTAGEG

Statistic	DF	Value	Prob
Chi-Square	8	73.9920	<.0001
Likelihood Ratio Chi-Square	8	85.9906	<.0001
Mantel-Haenszel Chi-Square	1	69.8635	<.0001
Phi Coefficient		0.1383	
Contingency Coefficient		0.1370	
Cramer's V		0.1383	

Frequency Percent Row Pct Col Pct	Table of Z by PENSION				
Z	**PENSION(number of pensionable age householders)**				**Total**
	0.00	**1.00**	**2.00**	**3.00**	
income	2462 63.62 67.60 95.61	755 19.51 20.73 90.64	423 10.93 11.61 91.96	2 0.05 0.05 100.00	3642 94.11
missing income	113 2.92 49.56 4.39	78 2.02 34.21 9.36	37 0.96 16.23 8.04	0 0.00 0.00 0.00	228 5.89
Total	2575 66.54	833 21.52	460 11.89	2 0.05	3870 100.00

Statistics for Table of Z by PENSION

Statistic	DF	Value	Prob
Chi-Square	3	32.5751	<.0001
Likelihood Ratio Chi-Square	3	30.6832	<.0001
Mantel-Haenszel Chi-Square	1	22.3030	<.0001
Phi Coefficient		0.0917	
Contingency Coefficient		0.0914	
Cramer's V		0.0917	
WARNING: 25% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

Frequency Percent Row Pct Col Pct	Table of Z by HHTYPE								
Z	**HHTYPE(Household type)**								**Total**
	single adult	**small adult**	**single parent**	**small family**	**large family**	**large adult**	**older smaller**	**single pensioner run**	
income	509 13.15 13.98 90.57	651 16.82 17.87 95.88	214 5.53 5.88 93.04	516 13.33 14.17 98.66	276 7.13 7.58 98.22	393 10.16 10.79 97.52	566 14.63 15.54 92.64	517 13.36 14.20 88.98	3642 94.11
missing income	53 1.37 23.25 9.43	28 0.72 12.28 4.12	16 0.41 7.02 6.96	7 0.18 3.07 1.34	5 0.13 2.19 1.78	10 0.26 4.39 2.48	45 1.16 19.74 7.36	64 1.65 28.07 11.02	228 5.89
Total	562 14.52	679 17.55	230 5.94	523 13.51	281 7.26	403 10.41	611 15.79	581 15.01	3870 100.00

Statistics for Table of Z by HHTYPE

Statistic	DF	Value	Prob
Chi-Square	7	83.4773	<.0001
Likelihood Ratio Chi-Square	7	90.3119	<.0001
Mantel-Haenszel Chi-Square	1	3.6174	0.0572
Phi Coefficient		0.1469	
Contingency Coefficient		0.1453	
Cramer's V		0.1469	

Frequency Percent Row Pct Col Pct	Table of Z by UNDER5		
Z	**UNDER5(child under 5 in household)**		**Total**
	No	**Yes**	
income	3275 84.63 89.92 93.71	367 9.48 10.08 97.87	3642 94.11
missing income	220 5.68 96.49 6.29	8 0.21 3.51 2.13	228 5.89
Total	3495 90.31	375 9.69	3870 100.00

Statistics for Table of Z by UNDER5

Statistic	DF	Value	Prob
Chi-Square	1	10.5776	0.0011
Likelihood Ratio Chi-Square	1	13.4921	0.0002
Continuity Adj. Chi-Square	1	9.8404	0.0017
Mantel-Haenszel Chi-Square	1	10.5749	0.0011
Phi Coefficient		-0.0523	
Contingency Coefficient		0.0522	
Cramer's V		-0.0523	

Fisher's Exact Test	
Cell (1,1) Frequency (F)	3275
Left-sided Pr <= F	2.434E-04
Right-sided Pr >= F	0.9999
Table Probability (P)	1.692E-04
Two-sided Pr <= P	4.677E-04

Frequency Percent Row Pct Col Pct	Table of Z by SIXTYPLS		
Z	**SIXTYPLS(person aged 60 or over in household)**		**Total**
	No	**Yes**	
income	2312 59.74 63.48 95.85	1330 34.37 36.52 91.22	3642 94.11
missing income	100 2.58 43.86 4.15	128 3.31 56.14 8.78	228 5.89
Total	2412 62.33	1458 37.67	3870 100.00

Statistics for Table of Z by SIXTYPLS

Statistic	DF	Value	Prob
Chi-Square	1	35.1832	<.0001
Likelihood Ratio Chi-Square	1	33.9176	<.0001
Continuity Adj. Chi-Square	1	34.3525	<.0001
Mantel-Haenszel Chi-Square	1	35.1741	<.0001
Phi Coefficient		0.0953	
Contingency Coefficient		0.0949	
Cramer's V		0.0953	

Fisher's Exact Test	
Cell (1,1) Frequency (F)	2312
Left-sided Pr <= F	1.0000
Right-sided Pr >= F	4.246E-09
Table Probability (P)	2.380E-09
Two-sided Pr <= P	7.852E-09

Frequency Percent Row Pct Col Pct	Table of Z by HEALTH			
Z	**HEALTH(Any long-term sick/disabled in household)**			**Total**
	No	**Yes**	**Unobtainable**	
income	2162 55.87 59.36 95.58	1475 38.11 40.50 92.01	5 0.13 0.14 100.00	3642 94.11
missing income	100 2.58 43.86 4.42	128 3.31 56.14 7.99	0 0.00 0.00 0.00	228 5.89
Total	2262 58.45	1603 41.42	5 0.13	3870 100.00

Statistics for Table of Z by HEALTH

Statistic	DF	Value	Prob
Chi-Square	2	21.8085	<.0001
Likelihood Ratio Chi-Square	2	21.6933	<.0001
Mantel-Haenszel Chi-Square	1	13.1879	0.0003
Phi Coefficient		0.0751	
Contingency Coefficient		0.0749	
Cramer's V		0.0751	

WARNING: 33% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

Appendix 11: 2002 Imputation Overview

Summary

- Carried out by National Centre for Social Research according to set of requirements specified by Communities Scotland.

- Hot deck imputation was used , described as dividing sample into imputation classes based of relevant characteristics of cases and these classes contain potential donor cases. A donor case is selected at random from the imputation class and the utem value for the case is assigned to the case with the missing value.

- The relevant characteristics were chosen using regression analysis. Very large and very small values were excluded from the imputation classes.

Earned income

Earned income was imputed for the Highest Income Householder and his or her spouse separately.

Original missing items of earned income shown below:

	Missing	Total	% Missing
Highest Income Householder	2313	9941	23.3%
Highest Income Householder other earnings	94	292	32.2%
Spouse Main Earnings	1258	5489	22.9%
Spouse Other Earnings	55	203	27.1%

The characteristics in the regression analysis which were significantly related to amounted earned by highest income householder (HIH) in their main job, and therefore used for imputation classes were:

- Age of HIH
- Sex of HIH
- Number of Rooms in Household
- Household type
- Whether HIH had more than one job

- Whether the HIH was in receipt of WFTC (Working F Tax Credit)
- Whether the HIH was in receipt of housing benefits
- Whether they owned or rented

For amount of second income for the highest income householder :

- Whether they owned or rented
- Whether the HIH was self employed

There were different relationships between characteristics and earned income of spouse/partner of the highest income householder. Spouses/partners were more likely to be working part-time and had different patterns of earning. The characteristics used to create imputation classes were:

-Number of rooms
-Spouse/partner age
-Spouse/partner sex
-Household type
-Whether the spouse/partner had more than one job
-Whether they owned or rented
-Whether the spouse/partner was in full time work
-Whether the spouse/partner was self employed

For the amount earned by spouse/partner from a second job, these were:

-Household type
-Whether spouse/partner in full time work

<u>Benefits and other regular income</u>

Amounts for these sources of income received for each benefit or income source, by both the highest income householder and their spouse. Totals received by both HIH and spouse were imputed for each benefit or other income item, where individual items were missing. In addition, there were households where a total for all benefits or for all other regular income was missing and this was imputed for these households.

	Cases with missing items	Total	% with any missing items
Benefits	5233	13530	38.7%
Other regular income	2766	6592	42.0%

Regression analysis was carried out for each type of benefit where there were a sufficient number of cases, to determine the relevant characteristics for imputation classes. However, where the number of cases was small, crosstabs and means were used.

Housing benefit was calculated in a number of stages, it was calculated directly where the information on the amount of rent before and after the housing benefit and the period for which the rent was paid. Otherwise housing benefit was imputed using the same categories as gross rent. Imputed housing benefit was restricted to an amount not exceeding the gross rent.

BENEFIT	CHARACTERISTICS
Income Support	- Collapsed version of household type - Whether the highest income householder is in ft work - Whether the highest income householder is retired
Working families tax credit	- Collapsed version of household type - Whether the highest income householder is in ft work - Banded total household income
Jobseeker's allowance	- Highest income householders marital status - Whether the highest income householder is in ft work
Housing benefit	HB calculated directly where rent questions allow. If rent after HB given, calculated from gross rent. Remaining cases hotdeck based on: - Tenure -Year moved in - Number of rooms - Age of highest income householder

	- Whether the highest income householder is in ft work - Whether the accommodation was tied
Council Tax Benefit	- Household type - Whether the highest income householder is in ft work
Earnings top-up	- Not carried out
Child benefit	- Total number of children - Household type
Child benefit at one parent rate	- Total number of children - Household type
Maternity allowance	- Whether the highest income householder is in ft work
State retirement pension	- Whether the highest income householder is retired - Whether the spouse is retired
Other benefit	- Household type
Widow's payment	- Household type
Widowed mother's allowance	- Household type
Widow's pension	- Whether the highest income householder is in ft work - Sex of the highest income householder
Incapacity benefit	- Whether highest income householder is long term ill/ disabled - Whether spouse is long term ill/disabled
Disabled persons tax credit	- Whether the highest income householder is in ft work - Collapsed household type
Disability living allowance care	- Whether highest income householder is long term ill/ disabled - Whether spouse is long term ill/disabled
Disability allowance mobility	- Whether highest income householder is long term ill/disabled - Whether spouse is long term ill/disabled
Industrial injury/disablement	- Whether the highest income householder is in ft work - Collapsed household type

Invalid care allowance	- Whether the highest income householder is in ft work - Whether the spouse is in ft work
Severe disablement benefit	- Household type
Statutory sick pay	- Whether the highest income householder is in ft work - Whether the spouse is in ft work
War disablement allowance	- Whether the highest income householder is long tem ill/disabled - Whether the spouse is in ft work
Disability premium	- Whether the highest income householder is long term ill/disabled - Whether spouse is long term ill/disabled
Attendance allowance	- whether spouse is retired

For other regular income sources, hotdeck groups were based on the following characteristics:

INCOME	CHARACTERISTICS
Non-state pension	- Household type - Whether HIH is single - Whether highest income householder is long term ill/disabled
Maintenance payments	- Household type - Age if highest income householder
Dig money	- Total number of adults - Total number of children - Whether HIH is single
Income from student loan	- Number of students in household

There were also households were no information had been given on any income sources. For some of those households a question on income band had been answered and the mid-point of the income band was imputed for their total income. Income was imputed in this way for 91 cases (0.5%) of social survey cases). Where there is no information on income band either, household income is unavailable. Household income is unavailable. Household income is unavailable for 107 cases (0.6% of social survey cases).

Mortgages

Mortgage payments were imputed for households with endowment mortgages and for those with non-endowment mortgages. Mortgage items were missing in 17.7% of households with a mortgage. The same set of variables was significantly related to missing monthly payments for non-endowment mortgages and missing mortgage payments for endowment mortgages. The following variables were used to create imputation classes:

- Amount borrowed (grouped)
- Year moved in
- Number of Rooms
- Age of highest income earner
- Whether or not the highest income earner was in full time worker
- Whether or not the highest income earner was long term sick/disabled

The same variables, with the exception of whether the highest income earner was long term sick/disabled, were fitted for missing additional endowment payments. The classes for addition payments were created using the following variables:

- Amount borrowed (grouped)
- Year moved in
- Number of rooms
- Age of highest income earner
- Whether or not the highest income earner was in full time work

Rent

The amount of rent paid before deduction of housing benefit was imputed for households where this item was missing. This rent item was missing for 20.7% of households paying rent.

The variables significantly related to gross rent and used for imputation classes were:

- tenure
- year moved in
- number of rooms
- whether or not the spouse was in full time work
- age of highest income earner

- household type
- whether the accommodation was tied

Council tax payments

Council tax was imputed, where missing, for the purpose of determining whether or not a household was in Fuel Poverty (Fuel Poverty Statement 2002 definition). Annual council tax was imputed to calculate our best estimate of income according to the HBAI definition used in FPS 2002. This estimate of HBAI was used in the calculation of Fuel Poverty (2002 definition) only. Annual council tax after imputation is available for households where both a social survey and a full physical survey of the dwelling had been carried out.

Values were missing for 2154 (14%) cases. The distribution of council tax payments was substantially different for those house households that received council tax benefits and those that did not, and therefore missing values were imputed separately for these groups.

For households that did not receive council tax benefits, regression analysis showed that income, number of rooms and dwelling type were the best predictors of council tax payments. Cases were organised into groups based on all combinations of these three variables. Missing cases were then assigned the median value of the group to which they belong, after outliers had been removed.

For households that did receive council tax benefits, tenure and income were the best predictors of council tax payments. Again, cases were organised into groups based on all combinations of these three variables and missing cases were then assigned the value of the group to which they belong.

Loft Insulation

Values were missing for 952 (8%) cases. Regression analysis showed type of dwelling (house/flat), tenure and age of dwelling to be the best predictors of loft insulation. Cases were organised into groups based on all combinations of these three variables. Missing cases were then assigned according to the proportion of dwellings within each grouping that had each level of loft insulation.

Fuel expenditure

Values were missing for 146 (10%) cases. Regression analysis showed that the number of people in household, number of rooms in the dwelling and primary fuel source were the best predictors of fuel expenditure. Cases were organised into groups based on all combinations of

these three variables. Missing cases were then assigned the median value of the group to which they belong, after outliers had been removed.

The HBAI definition of income before housing costs is the definition of income used in the Scottish Fuel Poverty Statement published in August 2002.

Under this definition, income includes total income from all members of the household, including dependants, and includes the following components:

- usual net earnings from employment;

- profit or loss from self-employment;

- all Social Security benefits (including Housing benefit, but excluding Social Fund Loans) and Tax Credits;

- Income from occupational and private pensions;

- Investment income;

- maintenance payments, if a person received them directly;

- income from education grants and scholarships (including for students, top-up loans and parental contributions);

- the cash value of certain forms of income in kind (free school means, free welfare milk, and free school milk).

Under this definition, income is net of the following items:

- income tax payments;

- National Insurance contributions;

- Council Tax;

- Contributions to occupational pension schemes (including additional voluntary contributions) and any contribution to person pensions;

- All maintenance and child support payments, which are deducted from the income of the person making the payment

- Parental contributions to students living away from home

Appendix 12: 2003/4 Onwards Imputation Overview

Software Used

Imputation of missing income for the Scottish House Condition Survey (SCHS) was applied by Methodology Directorate (MD) using the Canadian Census Edit and Imputation System (CANCEIS) which was developed to perform minimum change nearest neighbour imputation. CANCEIS is a generic system written in ANSI C and having the functionality to work with a variety of data types in which the user supplies their own data rules and requirements. Design of CANCEIS based on Nearest Neighbour Imputation Methodology (NIM) developed by Mike Bankier in 1992. CANCEIS is an integral part of the ONS Corporate Edit and Imputation Toolkit.

Method

For the 2004 SHCS survey cycle, SVS supplied a single dataset in the form of an SPSS data file, SENY1_all.sav, which contained some 3505 variables in respect of 3870 household level records.

The household income is defined to be the net income of the Respondent and their Partner. After discussion with Communities Scotland it was agreed that MD would impute for the following variables:

Variables imputed for Respondent	
Variable	Description
DIN1	Employment Status
DIN4	Amount of usual net pay
DIN5	Period of usual net pay
DIN6	Accuracy of usual net pay
DIN7	Hours worked per week
Variables imputed for Partner	
DIN35	Employment Status
DIN36	Number of Jobs
DIN38	For number of jobs = : Employed or self employed
DIN37	For number of jobs <1: Employed or self employed
DIN39	Amount of usual net pay
DIN40	Period of usual net pay

DIN41	Accuracy of usual net pay
DIN42	Hours worked per week
DIN56a	Self employed – usual pay or no usual pay
DIN56	Amount of self employed pay
DIN57	Period of self employed pay
DIN58i	Net or gross amount
DIN59	Number of hours worked per week

For each variable that contained missing items, a duplicate variable was created and imputed into and an associated binary variable was created to indicate which records contained imputed values. For example, for the original variableDIN4, the variable DIN4i was created and imputed into and the associated indicator variable DIN4ind was created.

Matching Variables

Compensation for missing items in surveys is generally based on equating, or specifying differences between respondents and non-respondents by controlling for other variables. Thus we can correct for non-response in the SHCS data by applying imputation in order to improve the quality of the statistical information. Methods of imputation vary depending on the type of data together with the extent and characteristics of the missing ness.

In order to identify matching variables, or predictors for the imputation process we consider each variable of interest in the form of a single nominal response variable reflecting whether we have an observed response or not. We can then fit a regression-like model to identify the optimum set of matching variables. This is a probabilistic model that defines a causal relationship in which change in one variable increases the probability of change in another but does not invariably produce the change.

By perturbing the SHCS 2003 data set and applying logistic regression we found a common set of matching variables which we then applied to the data to achieve an acceptable recovery. Thus, by fitting the logistic model to the SCHS 2003 data we have identified the following six variables to be used as baseline matching variables:

Variable	Description	Respondent	Partner
DVHSIZE	Household size	√	√
WTC	Whether receiving working tax credit or not	√	√

NROOMS	Number of rooms	√	√
HBENEFIT	Whether in receipt of housing benefit or not	√	√
TENURE	Tenure	√	√
RESPAGE	Age of Respondent (grouped)	√	
RESPSEX	Sex of Respondent	√	
RESPMAR	Marital Status of Respondent	√	
RESPWORK	Whether respondent works full or part time	√	
PARTAGE	Age of partner (grouped)		√
PARTSEX	Sex of partner		√
PARTMAR	Marital status of partner		√
PARTWORK	Whether partner works full time or part time		√

Other Income

Imputation process was described as above for the variables imputed

Variable	Description
DB15	Any other regular income or payment
Then for each income or payment:	
DB17?	Who receives
DB17B?	Amount
DB17C?	Period

Benefits to which this applies are:

A. Ocuupational Pension

B. Annuity

C. Maintenance

D. Rent from property

E. Dig Money

F. Accident Scheme – complete no imputation

G. Investment Income

H. Student Loan

I. Grant

J. Other Income