DEPARTMENT OF
FAMILY AND
COMMUNITY
SERVICES

THE UNIVERSITY OF
MELBOURNE

HILDA PROJECT DISCUSSION PAPER SERIES
NO. 1/03, March 2003

Towards an Imputation Strategy for Wave 1 of the HILDA Survey

Nicole Watson and Mark Wooden

The HILDA Project was initiated, and is funded, by the Commonwealth Department of Family and Community Services

Contents

Introduction

Non-response is a problem that all surveys have to contend with. Some respondents do not provide complete answers to all questions, resulting in item non-response, and others do not provide an interview at all, resulting in unit non-response. The treatment of non-response requires the understanding of the characteristics of the non-respondents and the likely impact these non-respondents would have on the analysis of the survey data. In some cases, it is appropriate to treat non-response through weighting, and in others, through imputation. It may also be appropriate to not treat the non-respondents at all – effectively assuming that the non-respondents are like the respondents or that any bias introduced by excluding the non-respondents from the analysis is likely to be very small.

This paper discusses various issues surrounding the use of imputation in the Household, Income and Labour Dynamics in Australia (HILDA) Survey and seeks a way towards an imputation strategy. While much of the discussion centres on the imputation for Wave 1, some consideration is also given to the strategy for later waves where it might impact on Wave 1. The various options for imputation are discussed and the recommended approach for the HILDA Survey draws on the experiences of other organisations with large-scale surveys, especially longitudinal surveys.

Types of Non-Response

The treatment for non-response depends on the assumptions about the type of non-response that has occurred. Rubin (1976) was the first to formalise the response mechanism. There are essentially three types of non-response:

- *Missing completely at random* – this is where the event of a particular variable being missing is independent of its true value and of any observed variables. This type of 'missingness' occurs infrequently, except when the variables are missing by design (for example, you choose not to follow up a simple random sample of certain cases in an experiment).

- *Missing at random* – this is where the event of a particular variable being missing is independent of its true value, but is dependent on the variables that are not missing. (An example of this is where a respondent does not know their benefit income, but has reported sufficient information about their situation from which a reasonable estimate of their benefit income can be worked out.)

- *Not missing at random* – this is where the event of a particular variable being missing is dependent on its true value. (An example of this is where a respondent declines to report their income because it is very large.) This type of missingness is very difficult to adjust for.

Most approaches to non-response assume the data are missing at random (Dillman et al. 2002). When the missingness is not random, alternative approaches are required and this is an active area of research.

Weighting Versus Imputation

It is well recognised that unit non-response is treated through weighting and item non-response is treated through imputation (e.g., Lepkowski 1989, Nordholt 1998, Kalton and Brick 2000, Dillman et al. 2002). Essentially, weighting is an implicit form of imputation, where the weight of the missing unit is distributed to other like respondents.

For item non-response, however, it would be completely impractical to have a different weight for every variable. Weighting could be used to adjust for a string of item non-response for a unit, so that there would be a separate weight for various groups of variables.

Where weighting is not used, imputation can be considered to counter bias in the estimates when non-respondents are dissimilar to respondents. If the proportion of missing values for a variable is small, then the extent of the bias is likely to be also small and imputation should not be needed. Bennett (2001) suggests that when the proportion of missing data exceeds 10 per cent, then the issue of missingness needs to be addressed. Nordholt (1998) places an upper limit on the missingness at 40 per cent, after which he suggests imputation should not be done. For a variable with a very high rate of missing data, it is likely that the quality of the collected data is questionable and the ability to impute a reasonable value is limited.

To understand the approach that has been taken in weighting Wave 1 of the HILDA Survey, we must first understand the structure of the data. There are two primary units of analysis – the household and the individual. Non-responding households have been accounted for by adjusting the weights of the responding households. Non-responding persons (from responding households and non-responding households) have also been accounted for by adjusting the weights of responding persons. Within a responding household, non-responding persons can be seen as a string of non-response for which we have used weighting to make the non-response adjustments. (See Watson and Fry 2002 for more details on the weighting procedure used in Wave 1.) However, we have not, as yet, made any adjustments for item non-response occurring during the household or person interviews. This is where imputation may be beneficial, certainly for some key variables, such as income.

Candidates for Imputation

For the HILDA Survey we are limited by the amount of resources that can be spent on imputation and need to restrict our attention to a subset of variables. After discussion with FaCS, we agreed to initially consider several key variables, including:

- age;
- labour force status;
- household relationships;
- income and sources;
- housing costs;
- duration of most recent unemployment spell;
- education level;
- occupation and
- industry.

The details of the missingness for each of these variables are provided in Tables 1 and 2. Table 1 lists the person level variables and Table 2 shows the household level variables. The discussion below reflects the agreed position on the variables that will be imputed.

Person-Level Variables

Respondents

For respondents, there are no missing cases for age, labour force status and household relationships.

A sizeable proportion of cases are missing income. The Person Questionnaire (PQ) collected current wages and salaries, current benefits and five components for financial year income. The five financial year income components are:

- wages and salaries;
- benefits;
- business income;
- financial investment income; and
- other income.

Only these components will be imputed rather than all of the individual variables that make up each component. For the wage and salary components, any information about after tax amounts will be incorporated and used as real data (i.e. it won't be considered to be imputed). The income questions in the PQ are structured such that the respondent is asked whether they have a particular type of income and then how much they received from that source. Therefore, if the respondent has said they have income from a particular source, then we know that the amount cannot be zero. Total

income and after tax income can be readily worked out once the components are imputed.[1]

The duration of the most recent unemployment spell can be deduced for some respondents from the information provided in the calendar. For other respondents, the calendar provides information on the lower bound of the unemployment spell. It is only for the latter group that imputation would be required. The proportion of missing cases is relatively high, so this variable will be imputed.

In the calculation of the derived variable of highest level of education, inadequately described post-school courses were ignored and the highest level of education from the proper responses was used. Only the cases where all the post-school courses were inadequately described are classified in Table 1 as missing. However, the proportion of missing data is 2.4 per cent and the impact of this missingness is likely to be small on any analysis undertaken. This variable will not be imputed.

For occupation and industry, we restrict our attention to 2-digit level rather than the 4-digit level. At the 4-digit level, any imputation would be highly imprecise, especially when considering the longitudinal nature of the survey. At the 2-digit level, the occupation codes are broadly in order of the skill base, and a better prediction could be made. Two-digit industry would be difficult to impute well. The HILDA External Reference Group strongly argued against imputing these variables as any imputation would be very problematic in the longitudinal context of the survey. [2] The proportion of missing data is very small, being less than 1 per cent for occupation and less than 0.5 per cent for industry. These variables will not be imputed as the effect on analysis will be very small.

Non-Respondents in Partially Responding Households

It is the Melbourne Institute's view that imputation of missing variables for the non-respondents in responding households should only be considered if the focus is on creating household level information, such as total household income. It is not sensible to impute variables such as industry, occupation and unemployment duration for these people, when the analysis of these individuals is limited to the small number of valid or imputed variables. Therefore, only the following variables will be imputed for the non-respondents in responding households:

- age;
- labour force status;
- current wages and salary income;
- current benefit income;
- total last financial year individual income; and
- total last financial year after tax individual income.

[1] Note that after tax income is calculated from total income by applying the individual tax rates – it does not take into account the income of partners nor the number of dependents in the family.

[2] Imputation was discussed at the meeting held on 6 December 2002.

Financial year income by individual source, duration of most recent unemployment spell, highest level of education, occupation and industry will not be imputed for these individuals.

For age and labour force status, there are very few missing cases, so any imputation for these variables will be relatively trivial to implement and will have a minimal effect on any analysis undertaken. There are no missing cases for household relationships.

Household-Level variables

Total household income can be readily derived from the actual and imputed individual level incomes of people within the household.

With regard to housing costs, for people who are renting, this is the rental amount per month. For people who borrowed to purchase a house, this is the repayments paid per month. For the people who are living rent free (or have paid off their home), the amount is zero. However, several complexities arise. For the people who borrowed to purchase their home, we have collected the amount they pay off their original loan, together with any secondary loan secured against their home. We have not collected how much they pay back on loans from family or friends. Further, should we be considering the secondary loans against their home as a cost of housing, when they may have taken the loan out to pay for other big ticket items such as a car, boat, second home, etc? The proportion of missing cases for the various housing costs lie between 1.1 and 2.7 per cent. Any bias in the estimates or analysis due to this missing information is likely to be small. Therefore, no imputation will be undertaken on these variables.

Table 1: Scope of imputation required for person level variables in Wave 1

Item	Variables	Valid cases[a]	Missing cases[a]	% to impute[b]	Comments
Age					
At time of HF ivw	ahgage1-ahgage12 ahgage	R: 13969 E: 1153 C: 4790	R: 0 E: 5 C: 0	R: 0% E: 0.4% C: 0% A: <0.1%	
At 30th June 2001	ahhfag01-ahhfag12 ahhage	R: 13969 E: 1153 C: 4790	R: 0 E: 5 C: 0	R: 0% E: 0.4% C: 0% A:<0.1%	
Labour force status					
On HF	ahges1-ahges12	R: 13969 E: 1152	R: 0 E: 6	R: 0% E: 0.5% A:<0.1%	
Derived from PQ	aesdtl, aesbrd	R: 13969	R: 0	R: 0%	No imputation required – labour force status for enumerated non-responding individuals obtained from HF
Household relationships	arg02_01-arg12_11	R: 13969 E: 1158 C: 4790	R: 0 E: 0 C: 0	R: 0% E: 0% C: 0% A: 0%	No imputation required
Current income[c]					
Current wages and salaries – main job	awscme	R: 13613 E: 0	R: 356 E: 1158	R: 2.5% E: 100% A: 10.0%	Use variable that includes estimates from answers about after tax income. Assume positive wages and salaries for respondents.
Current wages and salaries – other jobs	awscoe	R: 13855 E: 0	R: 114 E: 1158	R: 0.8% E: 100% A: 8.4%	Use variable that includes estimates from answers about after tax income. Assume positive wages and salaries for respondents
Current wages and salaries – all jobs	awsce	R: 13507 E: 0	R: 462 E: 1158	R: 3.3% E: 100% A: 10.7%	Use variable that includes estimates from answers about after tax income. Assume positive wages and salaries for respondents.
Current benefits	abnc	R: 13833 E: 0	R: 136 E: 1158	R: 1.0% E: 100% A: 8.6%	Assume positive benefits for respondents

Item	Variables	Valid cases[a]	Missing cases[a]	% to impute[b]	Comments
Financial year income					
Wages and salaries	awsfe	R: 13380 E: 0	R: 589 E: 1158	R: 4.2% E: 100% A: 11.5%	Assume positive wages and salaries for respondents
Benefits	abnf	R: 13903 E: 0	R: 66 E: 1158	R: 0.5% E: 100% A: 8.1%	Assume positive benefits for respondents
Business income	abifp, abifn	R: 13457 E: 0	R: 512 E: 1158	R: 3.7% E: 100% A: 8.1%	Assume positive business income for respondents
Investments	aoifinvp, aoifinvn	R: 12831 E: 0	R: 1138 E: 1158	R: 8.1% E: 100% A 15.2%	Assume positive investment income for respondents
Other income	Aoifotht	R: 13859 E: 0	R: 110 E: 1158	R: 0.8% E: 100% A: 8.4%	Assume positive other income for respondents
Total income	atifep, atifen	R: 11913 E: 0	R: 2056 E: 1158	R: 14.7% E: 100% A: 21.2%	Use variable that includes estimates from answers about after tax income. Use sum of components of income.
Total after tax income	atiatp, atiatn	R: 11913 E: 0	R: 2056 E: 1158	R: 14.7% E: 100% A: 21.2%	Apply tax model to imputed total income
Duration of most recent unemployment spell	ajst	R: 638 E: 0	R: 90 E: ~50	R: 12.4% E: 100% A: 18.0%	Use calendar information from W1 and W2 to derive unemployment duration to nearest third of a month. Number of missing enumerated only cases may change slightly, depending on the imputed employment status for 6 cases.
Highest education level	aedhigh	R: 13632 E: 0	R: 337 E: 1158	R: 2.4% E: 100% A: 9.9%	Note that where a person had two qualifications, one described properly and another inadequately described, then the adequately described qualification was used to determine their highest qualification.
Occupation[d]					
Main job (for employed)	ajbmocc2	R: 8498 E: 0	R: 27 E: ~808	R: 0.3% E: 100% A: 8.8%	2 digit ASCO considered. Number of missing enumerated only cases may change slightly, depending on the imputed employment status for 6 cases.

Item	Variables	Valid cases[a]	Missing cases[a]	% to impute[b]	Comments
Last job (for unemployed and not in labour force)	aujljoc2	R: 4799 E: 0	R: 47 E: ~304	R: 1.0% E: 100% A: 6.5%	2 digit ASCO considered. Number of missing enumerated only cases may change slightly, depending on the imputed employment status for 6 cases. Note 641 PQ respondents did not get asked question on occupation (only asked of people without current job who had previously been in paid work).
Industry					
Main job (for employed)	ajbmind2	R: 8495 E: 0	R: 30 E: ~808	R: 0.4% E: 100% A: 9.0%	2 digit ANZSIC considered. Number of missing enumerated only cases may change slightly, depending on the imputed employment status for 6 cases.
Last job (for unemployed and not in labour force)[e]	aujljin2	R: 2217 E: 0	R: 7 E: ~140	R: 0.3% E: 100% A: 6.2%	2 digit ANZSIC considered. Number of missing enumerated only cases may change slightly, depending on the imputed employment status for 6 cases. Note 3257 PQ respondents did not get asked question on industry (only asked of people without current job and who had job in last 5 years).

Notes:

a. Persons completing a PQ are denoted by *R* (for respondent). Persons listed on the HF of responding households who were eligible to be interviewed, but did not complete a PQ are denoted by *E* (enumerated only). Persons not eligible to be interviewed (ie children) are denoted by *C*.

b. Percentage of all cases for which the variable is relevant that need to be imputed are denoted by *A*.

c. Wages and salaries for one year ago have been excluded from the list of income items to impute.

d. Father and mother occupations excluded from the list of occupation items considered.

e. This would require imputation of time since worked for pay for people who are not currently employed (aujtsjha, aujmtu, aujyru) to determine which people should answer the industry of previous employment. This would involve imputing responses for 3 respondents and approximately 304 enumerated persons who did not provide an interview.

Table 2: Scope of imputation required for household level variables in Wave 1

Item	Variables	Valid cases	Missing cases	% to impute	Comments
Total financial year income[a]	atifehp, atifehn	5442	2240	29.2%	Sum imputed income from household members.
Housing costs					
Rent payments	ahsrnt	2226	24	1.1%	
Mortgage repayments	ahsmg	5268	144	2.7%	
Second mortgage repayments	ahssl	5268	57	1.1%	

Notes:

a. Income components excluded at household level.

Imputation Methods

Having now decided on the variables that require imputation from Wave 1, we turn our attention to the myriad of techniques available for imputation. The following is not intended as a complete review of the techniques, but rather as a means of explaining the key differences between the different techniques.

Deterministic Imputation Techniques

Simple forms of imputation entail inserting the mean calculated from a group of like units in place of the missing value. This technique may be used when only means and totals are of interest. It does, however, underestimate the variability in the data.

Deterministic imputation techniques are usually discarded in favour of techniques that maintain the underlying distribution of the data. These preferred techniques are usually called 'stochastic' as the imputation allows for variability around the mean.

Stochastic Imputation Techniques

There are numerous stochastic imputation techniques, and the three main techniques are described below. The main premise behind these techniques is to divide the complete cases into like groups based on the observed characteristics of the incomplete cases and impute missing values based on the distribution of values from the complete cases. The quality of the imputation depends on how well the observed characteristics correlate with the true value of the missing variable. These techniques have been developed for situations where the missingness is considered to be random.

Regression Method

The regression method of imputation involves fitting a model using complete cases to predict the value of the missing variable for the incomplete cases. An error term, based on the complete cases, can be included in the prediction to maintain the underlying variability in the data. For example, you might build a regression model for income based on the observed explanatory variables for income, such as age, employment status, education attainment, household type, etc, and then you would use this regression model to predict the income for cases missing income.

Hot Deck

For the Hot Deck method, you divide the complete cases into imputation classes based on key explanatory variables. You then match the incomplete case (i.e. the recipient) to an imputation class and randomly choose a case with complete information (i.e. the donor). You would then replace the missing data with the valid data from the donor case. Where the imputation classes are too fine and you cannot match the recipient with a group of donors, you would collapse the classes. You would usually undertake regression modelling prior to defining the imputation classes to ensure that relevant variables are used to create these classes.

For example, if you wanted to impute income, you would build a regression model for income to identify key explanatory variables based on the complete cases. You would

then use these key explanatory variables (such as age, employment status, education attainment, household type, etc) to form imputation classes. You would then match your case with missing income to an imputation class and randomly choose a donor. The income for the donor is then inserted in place of the missing income for the non-respondent.

Nearest Neighbour

The nearest neighbour method can be seen as an extension of the hot deck method where you calculate the distance between the non-respondent and the respondents based on the observed variables and choose the closest respondent as the donor for the non-respondent. Alternatively, the nearest neighbour method can be seen as an extension of the regression method, where the regression model is the distance function and the nearest respondent is used as the donor.

Continuing the example of missing income, you would define a distance function based on key explanatory variables for income (such as age, employment status, education attainment, household type, etc) for both the respondents and non-respondents. You would then sort your dataset by the distance function and use the value of income of the respondent with the closest value of the distance measure to the non-respondent's to impute the missing income.

Single Versus Multiple Imputation

Where we only replace the missing variable with one imputed value, we run into the problem of users treating the imputed data as real data such that they underestimate the variances of the sample (that is, they will assume that the number of observations from which estimates are derived is larger than it actually is). Rubin and Schenker (1986) show that for simple situations in large samples that have 30 per cent of the data missing, single imputation results in the 90 per cent confidence intervials having below 80 per cent actual coverage.

To properly account for the multiple use of valid data, a technique such as multiple imputation is required (Rubin 1996, Graham and Hofer 2000).[3] Multiple imputation involves running the imputation procedure multiple times to produce typically between 5 and 10 different datasets (though 20 datasts are sometimes used) that have a range of imputed values for missing responses. Any analysis needs to be run seperately on each dataset. The results are then combined by averaging the estimates and variances from each dataset, with a small adjustment made to the variance to account for the number of datasets used.

Multiple imputation of the HILDA datasets would add an additional layer of complexity for users that would not be received favourably. This view is supported by the HILDA External Reference Group. When discussing the imputation undertaken for several lager scale US surveys, Marker et al. (2002) recognised that multiple imputation was a burden to users. He suggested that the naïve variances could be

[3] Alternatives to multiple imputation include replication with re-imputation and all-cases imputation (see Shao 2002 and Lee et al. 2002).

divided by the item response rates, but he recognised that this is an area needing further research. Another alternative would be to estimate the variances based on the non-imputed data (as suggested by the ABS).

Therefore, the Melbourne Institute recommends that a single method of imputation be used, bearing in mind that estimates will appear more accurate than what they actually are. Guidance will need to be provided to users on the extent of such overstatement in the accurcy of the estimates.

Longitudinal Dimension to Imputation

As the HILDA Survey is longitudinal, consideration should be given to the impact of imputation over time. Many of the imputation techniques have been developed for surveys that are cross-sectional and the emphasis has been towards population estimates and variances at a point in time.

When we have repeated observations on an individual over time, the estimates of change between waves are very important. Therefore, we would not want to introduce variability into the estimates of change by imputing solely based on cross-sectional information, thereby suggesting there has been change when there has not been. Nor would we want to decrease the variability of the change estimates by imputing solely from information about an individual (such as carrying forward the last observed value), which would suggest there has been no change when there has been some.

Clearly, the imputation technique used in the second and subsequent waves of a longitudinal survey will need to incorporate information collected about that individual in previous waves. A combination of weighting and imputation may be required, depending on the response pattern overtime (Lepkowski 1989). Where there are a limited number of waves missing, imputation could be used to adjust for missingness of key variables.

Comparison of Imputation Methods

There are many imputation methods available and no one method stands out as being the best. Stochastic methods are clearly superior to deterministic methods when more than means and totals are of interest. The complexity of the imputation procedure and resources required to undertake the imputation will always be a consideration. The complexity of the datasets provided to users also needs to be kept in mind. The nearest neighbour and hot deck methods have the attraction that 'real' data are being used rather than a prediction.

The comparisons undertaken typically result in several methods that provide reasonable solutions to the missing data problem (see Nordholt 1998, Bennett 2001). Improvements are generally gained through adding disproportionately to the complexity of the imputation technique, so are often not justifiable.

This suggests that an imputation technique should be chosen to meet the following criteria:

(i) Must maintain the distribution of the underlying variables that need to be imputed.

(ii) Must not add significantly to the complexity of the dataset such that users will be discouraged from using the data.

(iii) Must be achievable within the resources available to undertake the imputation.

(iv) Must be readily understood by users.

Guidance on the appropriate imputation methods for the HILDA Survey can be taken from the experiences within Australia on both cross-sectional and longitudinal surveys. We also draw on the experiences of longitudinal studies conducted overseas that are similar in structure to the HILDA Survey. This is the focus of the next two sections.

Imputation in ABS Household Surveys

The Head of the Statistical Services Branch at the Australian Bureau of Statistics (ABS), Frank Yu, was consulted regarding the imputation undertaken in a number of household-based surveys. The key points from this meeting that might impact on the approach taken to imputation in the HILDA Survey are outlined below.

Extent of Imputation

The ABS does not do imputation in their frequently run household-based surveys, such as the Labour Force Survey. The main reason for this is that the proportion of missing responses is low, generally below 5 per cent.

For the Household Expenditure Survey and the Survey of Income and Housing Costs, imputation is undertaken. Both of these surveys have a structure similar to the HILDA Survey where information is collected at the household level as well as at the individual level. The main reason imputation is undertaken is that the proportion of missing responses for some key questions lies between 5 and 10 per cent.

Imputation Method Used

For the two household-based surveys where imputation has been undertaken, households where less than 50 per cent of the adults provided an interview were discarded from the dataset. For example, in a 3-adult household where only 1 person provided an interview, this household would not have been included in the dataset. The household and person weights were calculated based on the remaining households.

All variables that had missing values were imputed using a hot deck approach (including non-responding person in the responding households left in the dataset). The ABS developed a special SAS macro to undertake the imputation. The questions were divided up into eight blocks of questions about similar topics. Imputation classes were defined for each of these eight blocks depending on key variables. Different key variables could be used for different blocks. Cases with missing data were randomly matched to a donor within the same imputation class. Where a donor could not be identified, the imputation classes were collapsed until a donor could be found. The information from the donor was used to fill in the missing sections of the incomplete case for that block of the questionnaire. The imputed data was then checked against other information for the case to ensure the integrity of the data was maintained.

The editing and imputation steps were repeated until cases with missing data had been filled and all edits passed. When an edit failed, the imputed response was set to missing and was re-imputed. In a negligible number of cases, it was necessary to overwrite the collected data by imputed data to ensure the edits were satisfactorily met.

The ABS included the imputed values in the calculations of means, but excluded from the calculations of variances. That way, the variances were not artificially expanded by the apparent increase in the number of observed cases.

Resources Required

The Survey of Income and Housing Costs was the first household-based survey for which imputation was done. One Research Officer worked full-time for six to nine months on the imputation for this survey. The imputation for the Household Expenditure Survey expanded from the work undertaken for the Survey of Income and Housing Costs and involved one Research Officer working full-time for four to six months.

After the initial development work, the imputation task was given to the Survey area to undertake in future surveys, rather than being run by the Methodology Division.

Advice for the HILDA Survey

Drawing on the experiences that the ABS has had with imputation, Frank Yu made several suggestions for imputation in the HILDA Survey:

- Use hot deck imputation in preference to nearest neighbour imputation – it was felt that the semi-deterministic nature of the nearest neighbour technique would have adverse effects on the estimates.

- Consider using a package such as CART to identify the variables used to define the imputation classes – CART may provide advantages over regression techniques as it better handles a large number of variables.[4]

- Consider imputing all variables rather than just key variables – researchers will then be able to undertake analysis on the complete dataset, rather than restricting their analysis to responding persons for non-key variables.

- Imputation of all previous waves needs to be done each year – the main purpose of the HILDA Survey is to provide longitudinal data, so the accuracy of the imputation will be much more important than having a fixed dataset.

[4] CART is a decision tree tool developed by Salford Systems that identifies significant patterns and relationships.

Imputation in Panel Studies in Australia

The use of imputation has been relatively minimal to date in Australian longitudinal studies. The General Customer Survey and Longitudinal Data Set, both run by FaCS, do not, as yet, include any imputation. For the Longidutinal Survey of Australian Youth, the treatment of missingness was left to the researchers. In the Longitudinal Study of Immigrants to Australia, item imputation was not done and unit non-response was dealt with through weighting.

The Longitudinal Study of Women's Health does not routinely impute missing data. The proportion of missing data ranges from approximately 2 per cent in the younger aged sample to approximately 10 per cent in the older aged sample. The use of multiple imputation has recently been tested for a small group of people who, after not returning their self-completion questionnaire, did provided a short telephone interview. Five separate datasets were created using an expectation maximization algorithm in NORM to impute a select group of missing variables. The two main concerns resulting from this work were that some unrealistic values were generated by the imputation technique and the range of analyses supported by multiple imputation packages is limited.

The Survey of Employment and Unemployment Patterns, undertaken by the ABS in 1994 to 1997, had imputation for wave non-response, but did not have imputation for item non-response. Information from earlier waves was used to construct imputation classes and a donor from the same wave was identified using the hot deck procedure. No backcasting of the imputation was performed when information about the individual was obtained in later waves.

Imputation in Panel Studies Overseas

The experience of overseas panel studies that are like the HILDA Survey may also provide some guidance.

The German Socio-Economic Panel does not impute any data items for their public release files. They do, however, impute responses when providing a small number of variables to the Cross National Equivalent File or the European Community Household Panel. The method used for this imputation, based on Little and Su (1989), is a nearest neighbour technique that takes into account cross-sectional as well as longitudinal information in defining the nearest neighbour. The imputation relies only on the data collected for the same variable (such as total income) over time and for other units. It does not use other key indicators (such as age, employment status, etc). Note that households with non-responding individuals were excluded from the first wave. In subsequent waves, however, households with non-responding individuals were included in the data files.

The British Household Panel Study imputes item non-response and wave non-response (if the person has responded at another wave) for a small number of income and housing cost variables. They rely on weights to adjust for other forms of non-response. Both hot deck and regression imputation are used. The hot deck method is typically used to impute categorical variables and the regression method is used to impute continuous money amount variables. The regression model is used to find the nearest valid case to impute for the missing case, therefore maintaining the variability in the data. Buck (1997) found substantial gains from incorporating into the imputation methods as much information about the structure of the relationships in the data as possible. For cross-wave imputation, the same techniques are used and information from previous or subsequent waves are used in defining the imputation classes or regression model. Only single imputation is done, and it is left to the user to make allowances for the spurious increase in precision. For more details see Taylor et al. (2002).

The Canadian Survey of Labour and Income Dynamics imputes data from previous waves, which is then updated for current circumstances. Where the information from the previous year is not available, then the nearest neighbour technique is used to find a donor from the current year.

The US Panel Study of Income Dynamics undertakes hot deck imputation to replace missing values in most money value questions and questions on hours worked. Variables on the cost of food were imputed using sub-group means from the previous wave, which were updated by an inflation factor. (For further information, see Hofferth et al. 1998.)

The US Survey of Income and Program Participation imputes information on the core person questionnaire for non-responding persons in responding households using the hot-deck method. Item non-response is imputed using a partitioned sequential hot deck method which divides the variables to be imputed into five topic areas and uses different imputation classes for each. (See Pennell 1993 for more information.)

Proposed Imputation Strategy for the HILDA Survey

The preceding discussion of imputation methods and the experiences of various Australian and overseas organisations has contributed several ideas to the proposed strategy for imputing variables in the HILDA Survey. With this information in mind, it is recommended that the guiding principles for the imputation strategy should encompass the following:

(i) Longitudinal information should be used where possible. This will mean that the imputation is recalculated every wave.

(ii) The imputation strategy should be viable for the long term. The approach taken in Wave 1 should be sensibly replicated for future waves and be able to take on board the additional information for each wave.

(iii) Imputation is a time intensive process and has to be limited by the resources available to undertake the task. It is better to impute a small number of variables well rather than to impute a large number of variables poorly.

(iv) Use of the imputed variables should be well within reach of all users.

(v) Any imputed variables should be clearly identified such that the users can use the imputed variable or original variable as they wish.

(vi) The imputation should maintain, as far as possible, the underlying variability in the data.

The following sections present the recommended imputation strategy for the HILDA Survey based on these principles.

Variables to Impute

As previously discussed, the variables that we will impute for respondents include:

- duration of most recent unemployment spell;
- two current summary income variables: wage and salaries, and benefits; and
- five financial year summary income variables: wages and salaries, benefits, business income, investment income, and other income.

The variables that we will impute for non-respondents in responding households include:

- age;
- labour force status;
- two current summary income variables: wage and salaries, and benefits; and
- total financial year income.

The main focus of the effort spent on imputation will be on the income variables. Income stands out as one topic where the rate of missing data is high and the variables will be often used. Note that it may also be necessary, depending on the quality of the imputed data, to combine some these income variables together.

Imputation for age and labour force status involves very few cases and can easily be provided as this was undertaken for the weighting (see Watson and Fry, 2002 for more details). Information on the duration of the current period of unemployment can draw on the experience recorded in the calendar to improve the current data and imputation could be used to fill the remaining gaps.

The other candidates identified by FaCS for imputation have very small amounts of missing data and as such it is believed that it would not adversely affect the analysis of the data if missing values are not imputed.

Non-Response Mechanism

For the variables with missing values in the HILDA Survey where we are not undertaking imputation, we will assume either that the data are missing completely at random, or, if it is missing at random, the bias introduced into the analysis is small because the rate of missingness is small.

For the variables that will be imputed, we will assume that the data are missing at random. Income, however, is sometimes given as an example of where the data may not be missing at random, and if this is the case, then we will attempt to quantify the problem. Non-random missingness is difficult to correct for, involves substantial assumptions on the reason for missingness and is usually done at the analysis stage. Therefore, the Melbourne Institute may have to provide a second-best option that is less controversial and will be suitable for most users.

Imputation Method

Multiple imputation techniques are not, as yet, supported by SPSS and Stata.[5] SAS has only recently provided an experimental multiple imputation procedure with Version 8 of SAS/STAT. Multiple imputation should be left to the advanced users if they should choose to pursue this.

Hot deck and nearest neighbour imputation techniques appear to be most popular for their ease of use and ability to maintain the variability in the data. Identifying the nearest neighbour through regression modelling seems a sensible approach. Therefore, the Melbourne Institute recommends that categorical variables are imputed using hot deck methods, and the continuous variables are imputed using the nearest neighbour-regression method.[6] In identifying the relevant variables to include in the imputation classes or regression models, a statistical package such as CART may be used.

---

[5] Note Stata does have 'regmsng' for linear regression with multiple imputation for missing variables, but has to be specially downloaded from the Stata web site. Multiple imputation tools in Stata do not, as yet, extend to other analysis procedures. Multiple imputation procedures are more readily available in S-Plus, along with dedicated packages such as NORM.

[6] For the very small number of cases requiring imputation of categorical variables (such as age and employment status) it is not worth the effort of building a multinomial logistic model, when a hot deck method has already been implemented for weighting purposes.

Imputation is a very resource intensive process, especially when done well. The relationships of observed variables to the variables with missing observations will need to be well understood and used to maximum advantage for the imputation to add value. Hirsch and Schumacher (2002) provide an example of a mis-specified model used to impute missing values for earnings in the US Current Population Survey. Unions status was not used in the hot deck imputation classes for earnings. As a result, the wage gap difference between union members and non-union members was biased towards zero. This highlights the need to carefully develop the imputation process.

Once information from Wave 2 is available, this can be used to help define the imputation classes used in the regression models, as appropriate, to impute missing cases in Wave 1.

Similarly, imputation for Wave 2 will use both Wave 1 and Wave 2 data. With the introduction of the second wave of information, the choices about the variables and cases to impute become more complex. It is recommended that:

- The imputation for Wave 2 should focus on a similar set of variables as for Wave 1, possibly extending to some wealth variables.

- For the cross-sectional file in Wave 2, only those households that fully or partially respond in Wave 2 should have imputed values and non-respondents in responding households should only have the smaller subset of variables imputed as done in Wave 1.

- For the longitudinal file, only those respondents that have not attrited should be imputed.

Presentation of Imputed Variables in Datasets

The main focus of imputation techniques is to improve population estimates obtained from survey data. Researchers undertaking analysis beyond the descriptive may prefer not to use the imputed values. If, for example, a researcher were to look at the predictors of income and include the imputed income values in their model, their results would be coloured by the relationships between income and other variables used to impute missing income. Instead, the researcher would use model-based procedures (to estimate the missing data model and substantive model simultaneously) or undertake their own imputation (Graham and Hofer 2002).

Both the imputed and the original variables will be provided on the household and person files. The imputed variable will contain the original data for the non-missing cases and the imputed data for the missing cases. Researchers will then be able to choose which variable they wish to use. It is felt that this approach is better than having one variable with an imputation flag as the researchers do not have to do any work in creating the original variable and the reason for the non-response is not lost.

The presentation of the imputed information for non-responding persons in responding households needs some consideration. Following the current conventions for these people, this information would be appended to the household file along with the other information collected about these persons during the household interview.

This approach may become cumbersome for some users and it is suggested that in addition to the household and person file, a third file could be provided. This third file would be called the 'enumerated person file' and the current person file would be called the 'responding person file'. The enumerated person file would contain the basic person level information collected during the household interview together with the imputed person level variables. Researchers could then summarize this file to get household level characteristics, such as the number of household members employed, the number of household members earning more than $50,000 per year, etc.

References

Bennett, D.A. (2001) 'How can I deal with missing data in my study', *Australian and New Zealand Journal of Public Health*, vol. 25, no. 5, pp 464-469.

Buck, N. (1997) 'Imputation for Missing Income Data in a Panel Study', Paper presented at the IASS/IAOS Satellite Meeting on Longitudinal Studies, Jerusalem, 27-31 August, 1997 (Draft Paper, ESRC Research Centre on Micro-Social Change, University of Essex).

Dillman, D.A., Eltinge, J.L., Groves, R.M. and Little, R.J.A (2002) 'Survey Nonresponse in Design, Data Collection and Analysis', in *Survey Nonresponse*, edited by Groves, R.M., Dillman, D.A., Eltinge, J.L. and Little, R.J.A., Wiley, New York.

Graham, J.W., and Hofer, S.M. (2000) 'Multiple Imputation in Multivariate Research', in *Modeling Longitudinal and Multilevel Data: Practical Issues, Applied Approaches, and Specific Examples*, edited by Little, T.D., Schnabel, K.U. and Baumert, J., Lawrence Erlbaum Associates, New Jersey.

Hofferth, S., Stafford, F.P., Yeung, W.J., Duncan, G.J., Hill, M.S., Lepkowski, J., Morgan, J.N. (1998), 'A Panel Study of Income Dynamics: Procedures and Codebooks – Guide to the 1993 Interviewing Year', Institue for Social Research, The Unviersity of Michigan.

Hirsch, B.T., and Schumacher, E.J. (2002) 'Match Bias in Wage Gap Estimates Due to Earnings Imputation', Trinity University, June 2002, available at www.trinity.edu/bhirsch/.

Kalton, G., and Brick, M., (2000), 'Weighting in household panel surveys', *Researching Social and Economic Change: the uses of household panel studies*, ed Rose, D., Routledge, London.

Lepkowski, J.M., (1989), 'Treatment of Wave Nonresponse in Panel Surveys' in *Panel Surveys*, edited by Kasprzyk, D., Duncan, G.J., Kalton, G, Singh, M.P., Wiley, New York.

Little, R.J.A, and Su, H.L. (1989) 'Item Non-Response in Panel Surveys' in *Panel Surveys*, edited by Kasprzyk, D., Duncan, G., and Singh, M.P., Wiley, New York.

Lee, H., Rancour, E. and Särndal, C.E. (2002) 'Variance Estimation from Survey Data under Single Imputation', in *Survey Nonresponse*, edited by Groves, R.M., Dillman, D.A., Eltinge, J.L. and Little, R.J.A., Wiley, New York.

Nordholt, E.S., (1998), 'Imputation: Methods, Simulation Experiments and Practial Examples', *International Statistical Review*, Vol 66, p157-180.

Pennell, S.G. (1993) 'Cross-Sectional Imputation and Longitudinal Editing Procedures in the Survey of Income and Program Participation', Institute for Social Research, The Unviesity of Michigan.

Rubin, D.B. (1976) 'Inference and Missing Data', *Biometrika*, vol. 63, pp. 581-590.

Rubin, D.B. (1996) 'Multiple Imputation After 18+ Years', *Journal of the American Statistical Association*, vol. 91, pp.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*, vol. 81, pp. 336-374.

Shao, J. (2002) 'Replication Methods for Variance Estimation in Complex Surveys with Imputed Data', in *Survey Nonresponse*, edited by Groves, R.M., Dillman, D.A., Eltinge, J.L. and Little, R.J.A., Wiley, New York.

Taylor, M.F., Brice, J., Buck, N. and Prentice-Lane, E. (2002) British Household Panel Survey User Manual Volume A: Introcution, Technical Report and Appendices, Univesity of Essex, Colchester.

Watson, N, and Fry, T.R.L. (2002), 'The Household, Income and Labour Dynamics in Australia (HILDA) Survey: Wave 1 Weighting', HILDA Project Technical Paper Series No. 3/02, Melbourne Institute of Applied Economic and Social Research, University of Melbourne.