

Data Imputation

Imputation refers to the replacement of missing data with a substitute that allows data analysis to be conducted without being misleading. The major application possible for LSAC is the estimation of attributes where the interviewee refuses to answer the survey question or the attribute is missing. It is also possible to estimate the attributes of children not available for interview.

The following have been identified as key components of successful data imputation by the HILDA data management team. These will be adopted as the driving principles of the LSAC imputation policy:

- i. Imputation should not lead to biases or distributional changes in the data, or significant extra variance to estimators.
- ii. The imputation process should rely on data from the sample rather than making external assumptions about the likely nature of missing data.
- iii. Imputation should not lead to important sample estimates being based too heavily upon imputed values.

Non-response

Evidence from the LSAC Dress Rehearsal suggests that item non-response was minimal for responding units but that unit non-response was significant. The size of the study requires that a strategy be established to effectively deal with missing data and non-responding units.

The treatment of non-response requires an understanding of the characteristics of the non-respondents and the likely impact these non-respondents would have on the survey data. It is proposed that LSAC treats unit non-response through weighting and item non-response through imputation. At times it may also be appropriate to ignore non-respondents – effectively assuming that non-respondents are like respondents or that any bias introduced by excluding the non-respondents from the analysis will be very small.

Imputation in other Longitudinal Studies

The use of imputation in similar studies overseas has been minimal. The National Longitudinal Survey of Children and Youth performed limited imputation to their datasets, restricting it to missing values in variables used to calculate scales. The Millennium Cohort Study has not implemented any imputation within delivered datasets, preferring researchers to access an unmanipulated but ‘clean’ dataset.

HILDA’s investigation of imputation in other Australian longitudinal studies has also shown minimal useage to date. The General Customer Survey and Longitudinal Data Set, both run by FaCS, do not as yet include any imputation. For the Longitudinal Survey of Australian Youth, the treatment of missingness was left to 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 Survey of Employment



and Unemployment Patterns, undertaken by the ABS in 1994 to 1997, used imputation for wave non-response, but did not use 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.

LSAC – Whether to Impute

Based on other national and international longitudinal studies and the difficulties involved in longitudinal data imputation it would appear that either (a) no imputation should be conducted on the LSAC dataset and that researchers make allowance for missing data in their analyses; or (b) a minimalist imputation strategy should be adopted, primarily based on items exhibiting missingness levels of less than 10%.

In determining which option to adopt, the Data Management team will rely on input from FaCS, consortium members and future users of the data.

The following discussion focuses on possible strategies if option (b) is adopted.

Principles

It is proposed that the guiding principles for any LSAC data imputation strategy should encompass the following:

- Only items at the specified threshold of missingness (5-10%) will be considered for imputation.
- Any imputed variables should be clearly identified such that the users can use the imputed variable or original variable as they wish.
- The imputation should maintain, as far as possible, the underlying variability in the data.
- Imputation be used for scale construction where items are missing.

The LSAC imputation strategy will target specific variables with levels of missing data between five and ten percent. This proposed strategy is based on the following:

- Where there is a small amount of missing data the existence of item non-response is likely to be more nuisance value than substantially affecting analyses.
- By imputing these missing values, results of analyses are unlikely to change but it will ensure data is more ‘user friendly’ in that the ‘missing’ line can be eliminated from tables, thus avoiding any complication in table percentage calculations.
- Furthermore, the need to drop records when running regression models is avoided.

In essence, no harm can be done by imputing for small amounts of non-response but one is able to produce a more useable dataset.

However, where there is a significant amount of non-response (>10%) there is a potential for non-response bias beyond what can be corrected for in an imputation strategy. Imputation cannot create data where it doesn’t exist. It doesn’t increase the sample size, although imputing for a large amount of missing data can give analysts a



false impression of the sample size for an analysis. For any data item exhibiting missingness at such a level (i.e. >10%) it is proposed that imputation should not occur. Rather, end users of the dataset will be provided with reference information about such items. This information will seek to explain the low response rate and may investigate areas such as:

- whether the item was one that a particular group of respondents was unable to answer;
- whether the item was confusing;
- whether the information requested was too sensitive.

Candidates for Imputation

It is proposed that for those data items collected in Wave 1 that show a level of missingness between five and ten percent, that discussion will be entered into with management at FaCS to determine a priority listing of imputation candidates. It is acknowledged that LSAC is limited by the amount of resources that can be spent on imputation. Therefore, we propose to restrict our attention to a subset of variables exhibiting missingness levels between five and ten percent with priority given to those missing variables used in the calculation of scales.

Imputation over Time

As LSAC is a longitudinal study, the method of imputation needs to accommodate the impact of imputation over time. Many imputation techniques have been developed for cross-sectional surveys and the emphasis has been on population estimates and variances at a point in time.

When repeated observations are collected over time, the estimates of change between waves are very important. It is important to avoid any introduction of variability into estimates of change by imputing solely based on cross-sectional information, thereby suggesting there has been change when there has not been.

Furthermore, decreasing 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 may have been some, and thus should also be avoided.

It is possible that longitudinal information may be used in the imputation process. This would involve a recalculation of imputed values at every wave. However, whilst the use of longitudinal data may improve the ability of a predictive model to impute data, there would be no single, master dataset for any one wave. Longitudinal imputation options will be assessed in later LSAC technical papers.

Imputation Method

Adapting the experience of the HILDA survey for the purposes of LSAC, and acknowledging that the majority of variables within LSAC are categorical, the proposed imputation method is the ‘hot deck’ procedure. This technique appears to be



most appropriate due to its ease of use and ability to maintain the variability in the data.

The ‘hot deck’ method divides the complete cases into imputation classes based on key variables (e.g. family type). The incomplete case (i.e. the recipient) can then be matched to an imputation class from which a case with complete information (i.e. the donor) can be chosen. The missing data is then replaced with the valid data from the donor case.

If a significant number of continuous variables are found to require imputation, it is proposed that, for these variables, it may be more appropriate to implement the ‘nearest neighbour’ method of imputation. This method is an extension of the hot deck procedure where the distance between the non-respondent and respondents is calculated based on the observed variables and the closest respondent is chosen as the donor for the non-respondent.

Both the imputed and the original variables will be provided on the datasets. 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 researchers do not have to do any work in creating the original variable and the reason for the non-response is not lost.

