

Appendix Estimation and Imputation

As far as possible, data are supplemented from external sources (CCO* or other). Where this is not possible, imputations are carried out for all variables required for calculating a household's disposable income. Other questionnaire variables are not imputed.

All imputation tasks are conducted using the IVEWare* macro and the multiple imputation method, which is based on a multivariate regression model.

Concerning individuals, a distinction was made between respondents (R) and item non-respondents (INR), who provided only some of the requisite information, and unit non-respondents (UNR), who refused to answer the individual questionnaire. There are different cases of non-response requiring imputation at household or individual level:

Individuals:

- Item non-response (INR) → imputation of missing income sub-components
- Unit non-response (UNR) → imputation of presence/absence of each income sub-component, followed by imputation of amount where appropriate

Households:

- Item non-response (INR) → imputation of missing income sub-components

Unlike individuals, a household can only correspond to item non-response, as households that are unit non-respondents are not valid.

1. Preliminary stages

1.1 Standardisation

Standardisation aims to calculate a standard annual full-time income per individual. For example, if a person has worked for six months part-time (50%), the resultant income will be multiplied by 4 ($\times 12/6 \times 100/50$) to arrive at a standardised value. This is the basis, which is similar for all individuals, on which imputations will be carried out. Imputed income amounts are then unstandardised on the basis of known information (duration and work-time quotient) for each individual and income sub-component.

For item non-response cases (INR), information used for income standardisation (work-time quotient, number of months worked, indication of period to which income corresponds) may be missing. Known information from respondents (R) is then used to impute these missing values. Income from this grouping can then be standardised as well.

Similarly, household variables stretching over a limited period are multiplied to arrive at standardised 12-month values.

1.2 Homogeneous imputation groups (HIG)

To impute missing values as accurately as possible, homogeneous imputation groups (HIG) are set up, within which imputations are conducted.

Groups of homogeneous individuals or households are created on the basis of specific shared features (explanatory variables) using a segmentation algorithm. Auxiliary variables used for segmentation purposes must be known in regard to both respondents and non-respondents for the income sub-

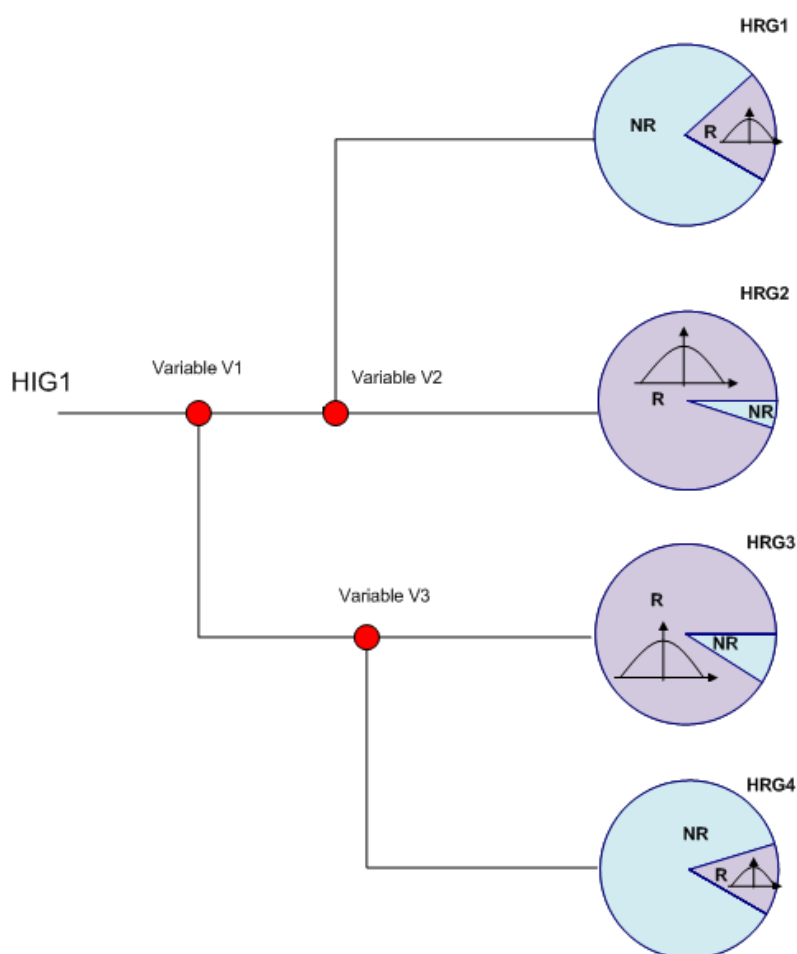
component in question. This segmentation tree is created by modelling the fact that one unit has a standardised amount for the income sub-component that is higher or lower than the median of the standardised incomes of that particular branch on the tree. HIGs are set up for each income variable.

1.3 Homogeneous response groups (HRG) within HIGs

Amongst individuals making up a HIG, a given variable will have respondents and non-respondents. Mirroring the principle used in creating HIGs, homogeneous response groups (HRG) are created by subdividing HIGs with the help of a segmentation algorithm, but this time based on the response (or lack thereof) to the question concerning the given variable. The aim is to create subsets of individuals with the highest probabilities of providing a homogeneous response to the given variable. To achieve this, some variables allowing for the best possible, though parsimonious, separation of groups of individuals that provided similar types of responses (response or non-response) are identified.

For each HRG, the aim is to obtain the median value from respondents. Medians taken from each HRG are used to model the missing values from the HIG's non-respondents. This procedure is supported by the hypothesis that probability of response is correlated with the amount of the income sub-component in question. Consequently, the median is used as the explanatory variable for modelling the missing income figures within the HIG. An example is shown in the following diagram (Figure 1).

Figure 1 Process for creating HRGs. In this example, a segmentation tree is created for HIG1. Three variables, V1, V2 and V3, are used to differentiate the four homogeneous response groups as accurately as possible. Medians calculated in regard to respondents in these HRGs will be used as explanatory variables for modelling missing values in HIG1.



2. Imputation of income variables

2.1 Individuals showing item non-response

Income variables are standardised using the aforementioned method. Within each HIG, missing amounts for each income sub-component are imputed by IVEWare*, then unstandardised to correspond once again to the duration and quotient of employment, whether actual or imputed.

2.2 Individuals showing unit non-response

Before imputing an income to unit non-respondents, the probability of an income sub-component being received must first be determined on the basis of the scant personal information available from grid and household questionnaires. A procedure is therefore implemented for each individual in order to impute the probability of a non-zero amount existing for each income sub-component.

Within the HIGs, for each individual, an iteration of 50 imputations is carried out on the basis of their grid profile, attributing a value of either 1 (amount>0) or 0 (amount=0). An average of these values indicates the probability that this person receives the income sub-component in question. Individuals who are unit non-respondents and who have the greatest probability of having a given income value will be imputed an amount by IVEWare*. To determine the number of them to whom an amount (>0) will be imputed, we take a percentage equivalent to that of individuals possessing the sub-component in question amongst the group of respondent individuals and individuals that have undergone imputation in regard to item non-response.

Selected individuals will then be imputed an income using a method similar to that used for item non-respondents. In contrast, where data is lacking for income standardisation, non-standardised income is imputed, determined on the basis of respondents' actual incomes. This is based on the assumption of a similar distribution of standardisation factors amongst respondents, individuals undergoing imputation for item non-response and those imputed for unit non-response within the HIGs that have been set up.

2.3 Households showing item non-response

The same process is applied as for individuals showing item non-response.

3. Imputation of other variables

Health insurance premiums

Annual health-insurance premiums are imputed deterministically on the basis of two core factors: city of residence and the person's age bracket (26 or over, 19-25 and 18 or less). For babies aged below 12 months, annual premiums are calculated depending on number of months as at 31 December 2016.¹ These sums feature under the variable *Tax on income and social contribution* (HY140G).

Total housing costs (HH070)

In computing the variable *Total housing costs* (HH070), the sum of ancillary costs is missing for some individuals. Amongst information providers, it was noted that rent after costs was tightly correlated with rent before costs. *Total housing costs* (HH070), was therefore imputed on the basis of a straightforward regression using the variable *Current rent* (HH060). Where both variables were missing, an amount was imputed using IVEWare*.

¹ Source: [Federal Public Health Office \(OFSP\)](#)

4. Checking imputations

Specific checks are carried out to verify the reliability of imputations. These include:

- Comparison of distribution curves for both observed and imputed values
- Chart representation of variation coefficients relating to the 50 imputation iterations for each value that is to be imputed
- Control of combinations of sub-components for a same individual between information providers and recipients of URN imputations

5. Imputed rent

Fictional or imputed rent (HY030G) corresponds to the savings made by a property owner or tenant living rent free or paying rent below the market price compared with a tenant paying market-price rent. It is calculated for people who have reported to CATI that they are property owners or tenants enjoying a preferential rent, a subsidy or who live rent-free.

For property owners, a “rent” is imputed in line with market prices and according to the characteristics of their dwelling. Actual expenses and mortgage interest paid by the owners (according to CATI) are deducted from this “rent”.

The Heckman model recommended by Eurostat is used to remove bias due to the part of the population that does not pay rent. This modelling combines a logistic and a linear regression. The dependent variable of the logistic model is the fact of being a market-price tenant or not (owners, tenants with a below market-price rent or rent free). In the next step, the linear regression’s dependent variable is the amount of rent. The linear regression uses *the inverse Mills ratio*, which was estimated in the first part of the model.

For tenants with a reduced rent, the rent plus expenses reported to the CATI are deducted as well as utility charges in relation to the dwelling (electricity, water, gas etc.) This calculation is different to the one recommended by Eurostat in order to be consistent with the one made for owners. This is because in the questionnaire it is too difficult for owners to separate expenses that are similar to those paid by tenants (electricity, water, gas) from those that are dissimilar (insurance, renovation, maintenance, etc.). We therefore deduct expenses and charges related to the dwelling for both tenants and owners.