IFS Derived Variables User Guide

Aims

This document describes how to use the IFS Derived Variables data sets. The purpose of these data sets is to centralise the derivation of some of the more complex variables and reduce duplication of effort across the research community. Over time, new variables will be added and if researchers have derived variables that they would like to see added to the derived data set, they should contact zoe_o@ifs.org.uk. New releases of the data will be deposited periodically.

Data description

There is one derived dataset for each wave of data. The data do not contain the financial derived variables which are deposited separately. Users may also find the Pension Wealth derived variables useful which are also deposited separately.

Variables that are included in the data cross a number of topics. Variables may be included for a number of reasons. These include:

1) Variables that are complex to derive and may rely on information that is not in the public release dataset
2) Variables that are simpler to derive but are included for reasons of convenience. For example, variables which rely on feeding forward data from a previous wave may come under this category. The derived variable will have the data fed-forward from all previous ways so the user does not have to do this.
3) Variables that are direct copies of variables from the core data but are included for convenience. Examples include age or sex and all the identifying information. Sometimes these variables may have their name changed where we view this to be more convenient. This made clear in the documentation.

The data can be combined with all other ELSA datasets by using idauniq – the unique ELSA identifier.

Accompanying this document is a spreadsheet which contains two worksheets. The first (Variable List) is a simple list of the variables and their descriptions. In columns B and C the variables are assigned a category and (for economics and health only) a subcategory. The filters on these columns allow you to restrict the worksheet to show only variables in the category you are interested in.

Detailed variable information is available contained in the second worksheet (Detailed Variable Information). There is one box for each variable contains the following information:

- Variable name
- Variable description
- Waves in which the variable is available
- Source variables. These are the variables from the core data that were used to construct the derived variable
- Syntax. This is included where practical. There are cases where syntax is very complex, and it would not be possible to include it in the spreadsheet.
- Coding Frame – where applicable

Naming conventions

There are two naming conventions in the files.

1. Variables relating to partners (e.g. partner's age) are named with an _p suffix
2. Variables constructed using the pensions grid are named with a pp_ prefix

Coding conventions

Value codes for variables are used in a similar way to the core data. The following codes are used in the same way as in the core data:

-9 Refused
-8 Don't know
-1 inapplicable

In addition, there are two further "missing" codes. These are

-2 not asked
-3 not asked this wave

The -2 (not asked) code is similar to the -2 value in the core data which is used whenever there is some kind of interviewer or routing error. Here, -2 is used whenever someone has not been asked the question but ideally we would have asked them. This could be due to a known routing error in a particular wave of data (that may since have been fixed0 but is sometimes due to complex routing across waves where very uncommon patterns of data were not allowed for in routing and as a consequence the individual was not asked the question.

-3 (not asked this wave) is used whenever a variable has not been fielded in a particular wave.

Expenditure variables and unfolding brackets

A feature of ELSA is that when respondents are asked to report a monetary value, if they do not know the answer or they refuse to reveal the answer, further probing is carried out to see if the respondent is willing or able to give a range ("…more than £X or less than £Y"). These follow-up questions are referred to as "unfolding brackets". The questions are designed to elicit a minimum and maximum number within which the value lies. In the derived dataset, the variables relating to expenditure are derived using information on unfolding brackets. Answers from respondents who were able to provide an exact answer are combined with answers from unfolding brackets to create three variables that can be used consistently across different respondents. To take an example of food inside the home, we use the information from hofood and combine that with information from hofoodl, hofoodu, hofoode, and hofoodr to create three variables, foodinl, foodinu and foodint. Foodinl contains the lower bound for food expenditure and foodinu contains the upper bound taken either from the continuous question (hofood) or from hofoodl and hofoodu (for those who entered the unfolding brackets). Hofoodl and hofoodu will be identical for respondents who gave a continuous answer. The variable foodint tells you what "type" of answer was given by the respondent. This may be continuous (gave an exact answer), closed band (gave both an upper and

lower bound), open bound (gave a lower but not an upper bound), or completely missing (did not provide any information even after being asked the unfolding bracket questions).

Unlike in the financial derived variables, no imputation has been carried out on the expenditure variables. Users can either use the upper and lower bounds in their analysis or carry out their own imputation.