

10 Generated Income Variables in SHARE Release 1

Agar Brugiavini, Enrica Croda, Omar Paccagnella, Roberta Rainato, and Guglielmo Weber

10.1 Introduction

This chapter documents the construction of *gross total annual income variables for 2003* in SHARE Release 1, at the individual as well as at the household level.¹

Let:

Y_{DIP}	gross individual income from employment
Y_{IND}	gross individual income from self-employment
Y_{PENS}	gross individual income from pension
Y_{REG}	gross individual income from private regular transfers (e.g. alimony...)
Y_L	gross individual income from long term care
Y_{BEN}	sum of the gross incomes of other household members and other benefits
Y_{AS}	capital assets income (income from bank accounts, from bonds, from stocks or shares and from mutual funds)
Y_{HO}	rent payments received, plus imputed rents

we define:

$$Y_{Ri} = Y_{DIP} + Y_{IND} + Y_{PENS} + Y_{REG} + Y_L$$

$$Y_{HH} = \sum_i Y_{Ri} + Y_{BEN} + Y_{AS} + Y_{HO}$$

where

Y_{Ri}	gross total individual income of respondent I
Y_{HH}	gross total household income.

The generated variables are provided in a Stata data set **sharerell_gv_inc.dta**, containing individual and household income information for each respondent. More precisely, in order to allow users to rely on multiple imputations, we provide 5 different final output data sets **sharerell_gv_incj.dta** ($j = 1, \dots, 5$). Section 10.6 for a brief discussion of multiple imputations.

10.2 Data availability and problems

10.2.1 General

We have chosen not to eliminate unusual values and possible outliers in the original data. The only exceptions to this rule are pension amounts in the Netherlands, where the Country Team deemed it necessary to transform the original data before using them in our programs (see Section 10.2.4 for details).

¹ Stata programs available upon request.

10.2.2 Availability

The SHARE questionnaire contains income-related questions in different modules. Income variables can be found in the employment and pensions module (EP), the household income module (HH), the housing module (HO), and the assets module (AS).

Questions may refer to different time frames. Employment and self-employment income amounts are asked directly as approximate yearly amounts. In contrast, the annual amount of income received from a specific pension or a specific regular payment needs to be calculated from 3 variables: average payment in 2003, the period covered by the payment, and the number of months in which the respondent has received that income payment in 2003. Lastly, long term care insurance income is asked as monthly amount. The income amount information available in each module is the following.

Module EP provides:

- gross annual income from employment in 2003
- gross annual earnings from self-employment in 2003
- gross income from pension, average amount of a typical payment in 2003
- gross income from regular transfers, average amount of a typical payment in 2003
- gross monthly income from long term care insurance

Module HH provides:

- gross annual income from other household members in 2003
- gross annual household payments (poverty relief, child benefits, ...) in 2003

Module HO provides:

- gross annual income or rent from secondary home
- amount still to pay on mortgage and loans, net of interest
- self-reported value of the house for home-owners

Module AS provides:

- gross annual interest from bank accounts, transaction accounts or saving accounts
- gross annual interest from government or corporate bonds
- gross annual dividend from stocks or shares
- gross annual interest or dividend from mutual funds or managed investment accounts

Note that some questions refer to the net rather than the gross value.

10.2.3 Euro and pre-Euro amounts

We express monetary amounts in Euro.

Non-Euro countries (Switzerland, Denmark and Sweden), however, report amounts in local currency. We convert these amounts into Euro by applying the exchange rates listed in Table 10.1.

Table 10.1 Exchange rates

Country	Currency	Old currency	Exchange rate (x to the Euro)
Austria	Euro	Austrian Schilling	13.7603
Germany	Euro	German Mark	1.95583
Sweden	Swedish Krona	-	9.1803
Netherlands	Euro	Dutch Guilder	2.20371
Spain	Euro	Spanish Peseta	166.386
Italy	Euro	Italian Lira	1936.27
France	Euro	French Franc	6.55957
Denmark	Danish Kroner	-	7.4388
Greece	Euro	Greek Dracma	340.75
Switzerland	Franc	-	1.5342

For Euro countries, if the answer to the Euro amount question is missing, but there is a non-missing value for the pre-Euro amount question, we use the latter (and convert the amount in Euro, see Chapter 3 for a description of the pre-Euro option in the survey instrument). For all countries, if the answer to the Euro amount question is “Don’t Know” or “Refuse”, we try to recover a value using the information available in the unfolding brackets.

10.2.4 Special procedures for particular variables

Some variables require special procedures

- Pension amounts
The annual amount of pension received is obtained using information from 3 variables: one amount variable (the average payment in 2003) and two frequency variables (the period covered by the payment, and number of months in which the respondent has received the payment in 2003)
To recover “invalid” (“Don’t Know”, “Refuse” or “.”) values, we use conditional hot-decking for amount variables, and linear regressions for frequency variables.
- Amount variables in module HO
We follow a strategy similar to the one described for pension amounts.
- Private Regular Payments
The annual amount of private regular payments is also obtained using information from 3 variables, one amount variable and two frequency variables. However, in this case, we follow a different strategy. First, we use hot-deck to recover “invalid” amount values. Next, we put to 0 the invalid values of the frequency variables.
- Amount variables in modules AS and HH
We follow a strategy similar to the one described for pension amounts.
- We have decided to impute rents for home-owners because they may represent a large fraction of resources at old age. We use information on self-reported house value and residual mortgage repayments derived from module HO. The interest rate of the imputed rents is fixed to 4% for all countries.

- Public pensions

In the Netherlands public pensions are received by all elderly individuals. In the case of couples in which both spouses don't work anymore, household-heads collect public pension both for themselves and for their spouses. Basically, public pensions in the Netherlands seem to represent household income rather than individual income. In contrast, Dutch occupational pensions are person-specific and are considered by the respondents as private pensions.

10.3 Imputations

We perform two types of imputations: imputations on amount variables, using the unfolding brackets (UBs) information and the hot-deck method, and imputations on frequency variables, using regression methods.

10.3.1 Unfolding brackets

The three bracket cut-off values (v_1, v_2, v_3) define 9 intervals (INT1,...INT9), depicted in Figure 10.1.

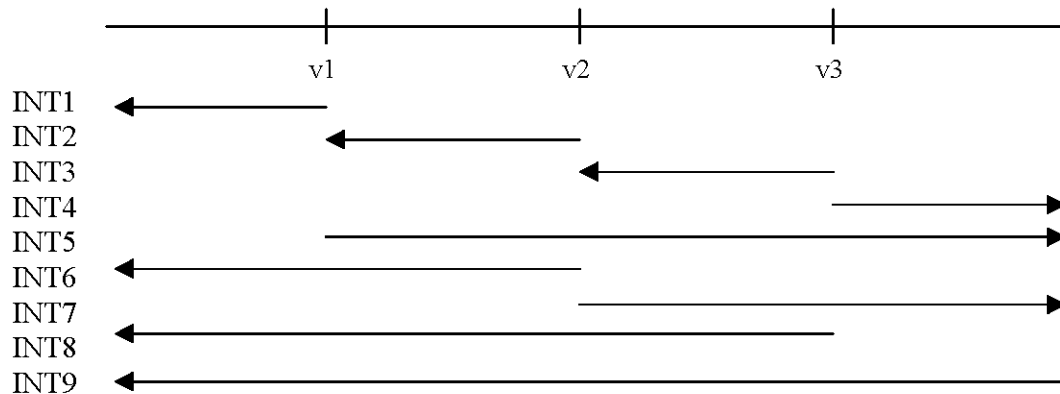


Figure 10.1: UB-defined intervals

The SHARE UB design is quite successful at recovering information for those respondents who were unwilling to answer or didn't know the exact answer.² Consider public old-age pension amount, for instance. Table 10.2 shows that 86.4% of the 8,602 respondents who report receiving income from public old-age pension, provide a 'valid/exact/continuous' answer to the amount question. More than half of the initial non-respondent complete the UB sequence, accounting for 8.3% of the recipients. Another 5.2% of the recipients enter the UB sequence and may give us some information so that we can narrow down the range in which their public old age income falls.

10.3.2 Imputation of amount variables: conditional hot-deck

We use the conditional hot-deck procedure to produce imputations for those cases in which respondents "Refuse" to answer or answer "Don't Know" to an amount question in modules EP, HO, AS and HH, and the associated UBs provide enough information to identify an interval. For this purpose, in the programming the "Refuse" or "Don't Know" cases for the Euro amounts (usually coded as 8e20 or 9e20), are considered as 'missing'.

² Chapter 12 provides a comparison of SHARE to ELSA and HRS for several variables.

Table 10.2 Public old age pension amount, percentage of recipients answering amount question or related UB sequence

Country	continuous amount	complete UB sequence	incomplete UB sequence	Total
AT	85.57	7.88	6.55	100
DE	84.25	8.38	7.37	100
SE	90.27	6.80	2.93	100
NL	82.89	11.18	5.92	100
ES	83.66	10.80	5.54	100
IT	87.71	10.63	1.67	100
FR	87.34	7.46	5.21	100
DK	85.93	9.09	4.97	100
GR	86.33	7.58	6.09	100
CH	94.8	2.26	2.94	100
All	86.43	8.32	5.24	100

The conditional hot-deck we implement is quite simple. We impute only the amount variable (and not the associated “ownership” variable that provides information regarding whether that income, pension or benefit was received), and we impute only one variable at a time. In the intervals 1 through 8, we stratify only by country. In contrast, in interval 9, we use a richer set of conditioning variables depending on the variable being imputed. Specifically, we stratify by country, gender and education for the imputation of employment incomes, by country, gender and age for pension incomes, and, lastly, by country and age for incomes from regular payments.³

For module HO, we use the imputed values for the two variables the self-reported house value and the residual mortgage provided by the Assets Working Group.

For module AS, we perform hot-deck imputations for the intervals INT1-INT8, and for INT9 we impute asset income as 2.5% of the associated imputed stock variable (using the imputations described in Chapter 11 of this Volume).

We perform one round of imputations for each variable. However, we do provide multiple imputations, obtained by running the whole income programs package multiple times.

10.3.3 Imputations for frequencies: regression method

For the imputation of relevant frequencies, we resort to linear regression techniques. In particular, we use the linear regression only for the frequencies of pensions received. The explanatory variables in these regressions are: age, gender, and indicators for whether the associated amount variable belong to the 1st, 2nd and 3rd quartile. The estimated coefficients for each frequency variable are produced separately by country.

³ Note that the hot-deck in a (conditioning variables, interval)-cell cannot be performed if there are no “donors” in that cell. In addition, the hot-deck is based on randomisation and repeating the procedure on exactly the same sample may give (slightly) different outcomes.

10.4 Naming conventions

10.4.1 General

Let X be an original variable and Y denote an aggregate variable derived from X . We use the following naming convention. YE denotes an amount variable possibly imputed and expressed in Euro. YP denotes the PPP-adjustment of YE , where we used the current OECD purchasing power parity, provided by Christelis, Jappelli and Padula (see Chapter 11 in this volume). Finally, YF denotes a flag variable indicating the nature of the imputations performed on the specific case.

10.4.2 Flag variables

We generate different types of flag variables, depending on the characteristics of the variables they are associated with.

A. Labels of the flag variables of an amount variable (e.g. earnings or pensions) that follows an ownership question and for which unfolding bracket sequence is possible

- 1 *valid response*: The respondent provides a valid response (in Euro or non-Euro).
- 2 *complete bracket*: The respondent answers ‘refuse’ or ‘don’t know’ on the amount-question, enters the unfolding bracket sequence and follows it until the end. We include here answers of the ‘about’ category.
- 3 *incomplete bracket*: The respondent answers ‘refuse’ or ‘don’t know’ on the amount-question, enters the unfolding bracket sequence and at least provides a valid answer to the first question but does not finish this sequence for some reason. At some point in the sequence the respondent answers ‘refuse’ or ‘don’t know’.
- 5 *no value/bracket*: The respondent answers ‘refuse’ or ‘don’t know’ on the amount-question, enters the unfolding bracket sequence but does not provide a valid answer to the first question and does not finish this sequence for some reason.
- 6 *no ownership*: This respondent is not asked the amount question. The respondent answers in a previous question that he or she does not own this item or has no such source of income.
- 7 *rf/dk ownership*: This respondent is not asked the amount question. The respondent answers in a previous question on ownership ‘refuse’ or ‘don’t know’.
- 9 *no respondent for this module*: The questionnaire identifies the household, housing and financial respondent. If this household, housing and financial respondent does not answer the specific CAPI-module (e.g. a financial respondent does not answer the AS module), this flag is up.

B. Labels of the flag variables for an amount variable (e.g. long term care) without unfolding brackets and for frequency variables

- 1 *valid response*: The respondent provides a valid response (in Euro or non-Euro).
- 5 *rf/dk*: The respondent answers ‘refuse’ or ‘don’t know’ or no valid value “dot, missing”.
- 6 *no ownership*: This respondent is not asked the amount question. The respondent answers in a previous question that he or she does not own this item or has no such source of income.
- 7 *rf/dk ownership*: This respondent is not asked the amount question. The

respondent answers in a previous question on ownership ‘refuse’ or ‘don’t know’.

- 9 *no respondent for this module*: The questionnaire identifies the household, housing and financial respondent. If this household, housing and financial respondent does not answer the specific CAPI-module, this flag is up.
- 12 *does not apply to the country*: The specific question is not asked to respondents of that country (used only for long term care).

C. Labels of the flag variables of a composed amount variable (e.g. household income)

- 0 *does not apply*⁴
- 1 *no imputations*: The respondent provides valid responses to all questions on which this composed variable is based. Hence no imputations are needed.
- 5 *some imputations*: The respondent does not provide valid responses to all questions on which this composed variable is based and some imputations are needed to construct this variable.
- 11 *imputation failed*: The hot-deck procedure may fail – it happens very rarely - because there are no donors that can be used for that specific interval

10.5 Final output: list of variables in `sharerel1_gv_inc.dta`⁵

The names of the final variables provided are listed below. As mentioned above, the suffix e indicates that a variable is expressed in Euro (after conversion from original non-Euro values where applicable). The suffix p denotes a conversion of the Euro amount to an amount adjusted to reflect the differences in the price levels between countries. The suffix f denotes the flag variable associated to a specific variable.

The file `sharerel1_gv_inc.dta` contains individual and household income information for each respondent.

The gross annual individual income is delivered in variable `yre` (in Euro) and in variable `yrp` (in PPP-adjusted Euro). The gross annual household income is delivered in variable `yhhe` (in Euro) and in variable `yhhp` (in PPP-adjusted Euro).

We provide also relevant income components that were constructed and aggregated to obtain total income measures. Some of these income components are country-specific. Hence, we assign them generic names and labels. In particular, this is the case with `ypensk` ($k = 1, \dots, 11$) and `yreg_k` ($k = 1, \dots, 5$). The reader is referred to the SHARE web-site for further documentation on these variables.

IDs

<code>sampid2</code>	HOUSEHOLD ID
<code>cvid</code>	COVERSCREEN ID OF RESPONDENT
<code>country</code>	Country

Individual level variables

<code>yre</code>	gross annual individual income in Euro
<code>yinde</code>	gross annual self-employment income in Euro
<code>ydipe</code>	gross annual employment income in Euro
<code>yle</code>	gross annual long term care in Euro
<code>penske</code>	gross annual country specific pension income k in Euro, $k = 1$ to 11

⁴ The amount question is asked only if the respondent answer “yes” to the associated ownership question. “Does not apply” in this context means that the associated ownership variable is not “yes”.

⁵ Notice that we provide 5 different final output data sets `sharerel1_gv_incj.dta` ($j = 1, \dots, 5$). See Section 10.6 below for further details.

Generated Income Variables

yreg_ke	gross annual country specific regular payment k in Euro, k = 1 to 5
yip	gross annual individual income PPP-adjusted (Euro)
yindp	gross annual self-employment income PPP-adjusted (Euro)
ydipp	gross annual employment income PPP-adjusted (Euro)
ylp	gross annual long term care PPP-adjusted (Euro)
ypenskp	gross annual country specific pension income k PPP-adjusted (Euro), k = 1 to 11
yreg_kp	gross annual country specific regular payment k PPP-adjusted (Euro), k = 1 to 5
irf	flag for the gross annual individual income amount
iindf	flag for the gross annual self-employment income amount
idipf	flag for the gross annual employment income amount
ilf	flag for the gross annual long term care amount
ipkf	flag for the gross annual country specific pension income k amount, k = 1 to 11
Iregkf	flag for the gross annual country specific regular payment k amount, k = 1 to 5

Household level variables

yhhe	gross annual household income in Euro
yhie	income from other household members in Euro
yothe	other household benefits in Euro
yrente	rent value at household level in Euro
yirente	imputed rent value at household level in Euro
ybacce	bank account at household level in Euro
ybonde	government or corporate bonds at household level in Euro
ystoce	stocks or shares at household level in Euro
yfunde	mutual funds at household level in Euro
yhhp	gross annual household income PPP-adjusted (Euro)
yhip	Income from other household members PPP-adjusted (Euro)
yothp	other household benefits PPP-adjusted (Euro)
yrentp	rent value at household level PPP-adjusted (Euro)
yirentp	imputed rent value at household level PPP-adjusted (Euro)
ybaccp	bank account at household level PPP-adjusted (Euro)
ybondp	government or corporate bonds at household level PPP-adjusted (Euro)
ystocp	stocks or shares at household level PPP-adjusted (Euro)
yfundp	mutual funds at household level PPP-adjusted (Euro)
ihhf	flag for the gross annual household income
ihif	flag for the income from other household members
iothf	flag for other household benefits
irentf	flag for the rent value at household level
iirentf	flag for the imputed rent value at household level
ibaccf	flag for the bank account at household level
ibondf	flag for the government or corporate bonds at household level
istocf	flag for the stocks or shares at household level
ifundf	flag for the mutual funds at household level

10.6 Multiple imputations in the generated income programs package

The income programs package discussed here performs only one round of imputations for each variable using country-specific univariate conditional hot-deck as imputation method for amount variables and linear regressions as imputation method for frequency variables. However, we do provide multiple imputations, constructed as follows. We set the number of replications M to 5, and we provide 5 different final output data sets, **sharerell_gv_incj.dta** ($j = 1, \dots, 5$), each obtained running the income programs package using a different (imputed) assets data set as input and a different seed for the randomization in the hot-deck procedure. In particular, in addition to the original SHARE 2004 data,

- `sharerel1_gv_inc1.dta` uses `sharerel1_gv_as1` and `seed = 123456789` (Stata's default)
- `sharerel1_gv_inc2.dta` uses `sharerel1_gv_as2` and `seed = 1000`
- `sharerel1_gv_inc3.dta` uses `sharerel1_gv_as3` and `seed = 10000`
- `sharerel1_gv_inc4.dta` uses `sharerel1_gv_as4` and `seed = 100000`
- `sharerel1_gv_inc5.dta` uses `sharerel1_gv_as5` and `seed = 1000000`

The reader is referred to Chapter 11 for details on the assets datasets **`sharerel1_gv_asj.dta`** ($j = 1, \dots, 5$), and on the use of multiple imputations in estimation.

Acknowledgements

We are grateful to Dimitris Christelis and Adriaan Kalwij for helpful discussions.