# GENERATED INCOME VARIABLES

# (SHORT) MEMO*

VENICE TEAM: AGAR BRUGIAVINI, ENRICA CRODA AND ROBERTA RAINATO§
PADUA TEAM: GUGLIELMO WEBER AND OMAR PACCAGNELLA§§

Revised: April 27, 2005

This document is composed of two parts:

Part I: Documentation of the Stata programs package to construct individual- and household-level income variables in SHARE

Part II: Multiple Imputations

A longer version of this document including Stata do-do files description and flow-chart us available upon request.

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

§ Agar Brugiavini: brugiavi@unive.it, Enrica Croda: enrica.croda@unive.it, Roberta Rainato: rob_2000@unive.it.
§§ Guglielmo Weber: guglielmo.weber@unipd.it, Omar Paccagnella: omar.paccagnella@unipd.it.

# CONTENTS

# PART I

## DOCUMENTATION OF THE STATA PROGRAMS PACKAGE TO CONSTRUCT INDIVIDUAL- AND HOUSEHOLD-LEVEL INCOME VARIABLES IN SHARE

### 1. Purpose

Our objective is to generate a measure of *gross total annual income for 2003*, at the individual as well as at the household level.

Let:

$Y_{Ri}$ gross total individual income of respondent i
$Y_{DIP}$ 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
$Y_{HH}$ gross total household income
$Y_{BEN}$ sum of the gross incomes of other household members ($Y_{HI}$) and other benefits ($Y_{OTH}$)
$Y_{AS}$ capital assets income (income from bank accounts - $Y_{BACC}$ -, from bonds - $Y_{BOND}$ -, from stocks or shares - $Y_{STOC}$ -, and from mutual funds - $Y_{FUND}$ -)
$Y_{HO}$ rent payments received ($Y_{RENT}$), plus imputed rents ($Y_{IRENT}$)

We define:

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

$Y_{HH} = ?_i Y_{Ri} + Y_{BEN} + Y_{AS} + Y_{HO}$

The Stata code described here generates a dataset INCOME.dta, containing individual and household income information for each respondent. This dataset is provided together with the programs.[1] More precisely, in order to allow users to rely on multiple imputations, we provide 5 different final output datasets INCOMEi.dta (i=1,…5).[2]

[1] The programs are provided on an "as-is" basis. They are distributed in the hope that they will be useful, but without warranties of any kind. All original material is provided under a Creative Commons Attribution-ShareAlike license.

[2] See below for details, and especially Part II for a brief discussion of multiple imputations.

## 2. Data availability and problems

### 2.1. General

We have chosen not to eliminate unusual values/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.[3]

### 2.2. Availability

Section EP provides:

- Gross annual income from employment in 2003 (ep205_)
- Gross annual earnings from self-employment in 2003 (ep207_)
- Gross income from pension, average amount of a typical payment in 2003 (ep078_1 … ep078_11)
- Gross income from regular transfers, average amount of a typical payment in 2003 (ep094_1 … ep094_5)
- Gross monthly income from long term care (ep086_)

Section HH provides:

- Gross annual income from other household members in 2003 (hh002_)
- Gross annual household payments (poverty relief, child benefits, …) in 2003 (hh011_)

Section HO provides:

- Gross annual income or rent from secondary home (ho030_)
- Multi-period payments for mortgage, net of interest (ho015_)
- Net multi-period income from imputed rents (ho024_)

Section AS provides:

- Gross annual interest from bank accounts, transaction accounts or saving accounts (as005_)
- Gross annual interest from government or corporate bonds (as009_)
- Gross annual dividend from stocks or shares (as015_)
- Gross annual interest or dividend from mutual funds or managed investment accounts (as058_)

Note: some variables refer to different time frames or report the net rather than the gross value.

### 2.3. Euro and pre-euro amounts

We express monetary amounts in euros.
For non-euro countries (Switzerland, Denmark and Sweden), which report amounts in local currency, we apply the exchange rate downloaded by the Amanda website: http://www.share-project.org/download/Amanda/convert%20to%20euro/convert_euro_130904.doc to convert amounts in euros.
For the 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 euros).
For all countries, if the answer to the euro amount question is DK or R, we try to recover a value using the information available in the unfolding brackets (UB).

[3] See below for more details.

### 2.4. Problems with particular variables

- Pension amounts.
  We generate a new variable, the annual amount of pension received, obtained using information from 3 variables:
  - average payment in 2003 (ep078_)
  - the period covered by the payment (ep074_)
  - number of months in which the respondent has received the payment in 2003 (ep208_)

  These three variables can take "invalid" values, for example "Don't Know", "Refuse" or ".". ep074_ can take also value "97.Other". If ep074_ is "97.Other", the respondent is asked to explain verbally what "Other" means.
  We currently implement the following procedure to recover "invalid" values:
  - for amount variables, use the hot-deck procedure
  - for frequency variables, use a linear regression
- Amount variables in section HO.
  We follow a strategy similar to the one described for pension amounts.
- Private Regular Payments.
  ep090_ and ep096_ present problems similar to the ones encountered respectively with ep074_ and ep208_ in the computation of pension amounts.
  In this case, we follow a different strategy. First, we use the hot-deck procedure to recover "invalid" (DK/REF/".") amount values. Next, we put to 0 the remaining invalid values (e.g. 97.Other) , both for amount and frequency variables.
- Amount variables in sections AS and HH
  We follow a strategy similar to the one described for pension amounts (see section 4 of this document for more details on section AS).
- 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 on residual mortgage repayments derived from section HO. The interest rate of the imputed rents is fixed to 4% for all countries. *Later on, we hope to receive a running mortgage rate from every country*.
- Public pensions (ep078_).
  Notice that in the Netherlands public pensions are received by all elderly individuals. In the case of a couple in which both spouses don't work anymore, the household head, usually the male spouse, collects public pension also for the spouse. 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.

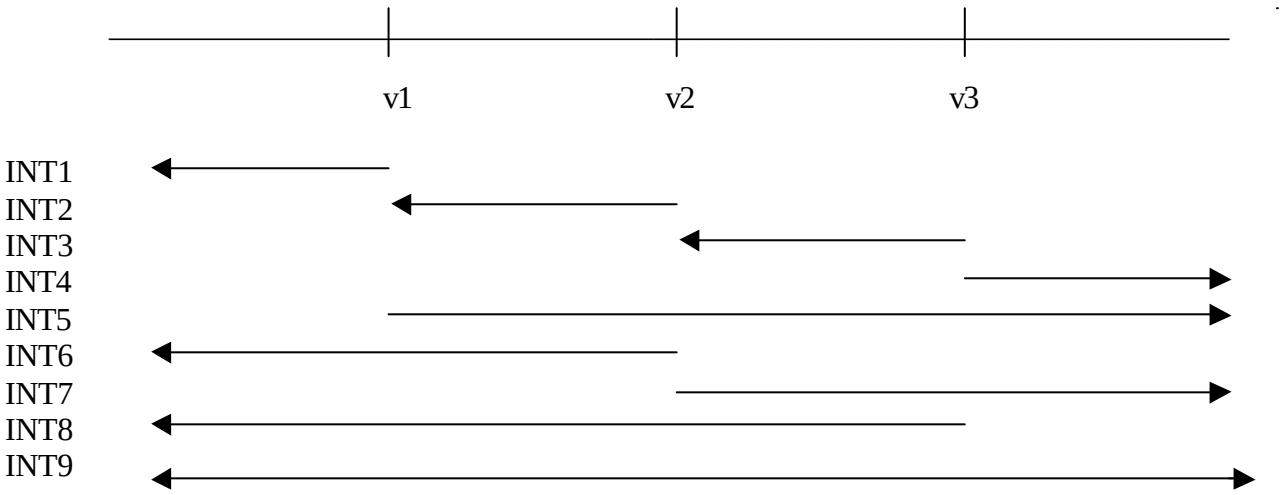
## 3. Imputations

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

### 3.1. Unfolding brackets and hot-deck

The three bracket cut-off values (v1,v2,v3) define 9 intervals (INT1,…INT9), depicted in Fig.1. We use the hot-deck procedure to produce imputations for those cases in which the amount is missing (see below) and the UBs provide enough information to identify an interval. For this purpose, the "Don't Know" or "Refuse" cases for the euro amounts (usually coded as 8e20 or 9e20), are considered as 'missing'.

**Fig.1. UB-defined intervals**




### 3.2. Imputation of amount variables: hot-deck

Currently, we are using a simple hot-deck procedure for sections EP, HO, AS and HH

- we only impute the amount variable (and not the associated "ownership" variable that give us the information if that income, pension or benefit is received)
- we only impute one variable at a time
- the package of Stata income programs described here performs one round of imputations for each variable.[4]
- in the intervals 1 through 8, we stratify by country; in contrast, in interval 9, we use a richer set of conditioning variables depending on the variable being imputed.[5]

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.

[4] However, we do provide multiple imputations, obtained by running the whole Stata income programs package multiple times. See below for details, and especially Part II for a brief discussion of multiple imputations.
[5] For the imputation of employment incomes, we stratify by country, gender and education. For the imputation of pension incomes, we stratify by country, gender and age. Lastly, for the imputation of pension incomes, we stratify by country and age.

For section HO we use the imputed values for variables ho024 and ho015 provided by the WG-AS. For section 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 (imputed by the Salerno team).

### 3.3. Imputations for frequencies: regression method

For the imputations of some important frequencies, we resort to linear regression techniques. In particular, we use the linear regression only for the frequencies of pensions received (ep074_ and ep208_).
The regression use the following explanatory variables:

- age
- gender
- indicators for whether the associated amount variable belong to the intervals defined by the the 1st, 2nd and 3rd quartile.

We produce the estimated coefficients for each frequency variable within each country.

## 4. Naming conventions

### 4.1. General

Let X be an original variable, Y denotes an aggregate variable derived from X.
We use the following naming convention:

- YE denotes an amount variable possibly imputed and expressed in euros
- YP denotes the PPP-adjustment of YE, where we used the current OECD purchasing power parity, kindly provided by the Salerno team
- YF denotes a flag variable indicating the nature of the imputations performed on the specific case (see section 4.2. for more details)

### 4.2. Flag variables

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

#### 4.2.1. 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 'refusal' or 'don' 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 'refusal' 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 'refusal' or 'don' know'.

*4.not used*

*5.no value/bracket*
The respondent answers 'refusal' or 'don't know' on the amount-question, enters the unfolding bracket sequence but does no 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 'refusal' 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.

**4.2.2. 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)

*7. rf/dk.*
The respondent answers 'refusal' 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 'refusal' 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).

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

*0. does not apply*[6]

*1.no imputations*
The respondent provides valid responses to all questions on which this composed variable is based. Hence no imputations were needed.

*5.some imputations*
The respondent does not provide valid responses to all questions on which this composed variable is based and some imputations were 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)

## 5. Final output: list of variables in INCOME.dta[7]

The name of the final variables generated by our working group are listed below. As mentioned above, the suffix E indicates that a variable is expressed in euros (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 **INCOME.dta** contains individual and household income information for each respondent. The gross annual individual income is delivered in variable YrE (in euros) and in variable YrP (in ppp-adjusted euros). The gross annual household income is delivered in variable YhhE (in euros) and in variable YhhP (in ppp-adjusted euros).
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 Ypensi (i=1,…11) and Yreg_i (i=1,…5). The reader is referred to the Deviation files on the Amanda web-site for further details on these variables.

IDs

```
sampid2     HOUSEHOLD ID
cvid        COVERSCREEN ID OF RESPONDENT
gender      GENDER OF RESPONDENT
respid      RESPONDENT ID
Ctr         Country
```

Individual level variables

```
YrE         gross annual individual income in euros
```

[6] 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".
[7] Notice that we provide 5 different final output datasets INCOMEi.dta (i=1,…5). As motivated and described in Part II, they are obtained running 5 times the income programs package under different conditions. Each run produces as output INCOME.dta, which we renamed to INCOMEi.dta. See Part II for details.

| | |
|---|---|
| YindE | gross annual self-employment income in euros |
| YdipE | gross annual employment income in euros |
| YlE | gross annual long term care in euros |
| Ypens1E | gross annual country specific pension income 1 in euros |
| Ypens2E | gross annual country specific pension income 2 in euros |
| Ypens3E | gross annual country specific pension income 3 in euros |
| Ypens4E | gross annual country specific pension income 4 in euros |
| Ypens5E | gross annual country specific pension income 5 in euros |
| Ypens6E | gross annual country specific pension income 6 in euros |
| Ypens7E | gross annual country specific pension income 7 in euros |
| Ypens8E | gross annual country specific pension income 8 in euros |
| Ypens9E | gross annual country specific pension income 9 in euros |
| Ypens10E | gross annual country specific pension income 10 in euros |
| Ypens11E | gross annual country specific pension income 11 in euros |
| Yreg_1E | gross annual country specific regular payment 1 in euros |
| Yreg_2E | gross annual country specific regular payment 2 in euros |
| Yreg_3E | gross annual country specific regular payment 3 in euros |
| Yreg_4E | gross annual country specific regular payment 4 in euros |
| Yreg_5E | gross annual country specific regular payment 5 in euros |

| | |
|---|---|
| YrP | 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) |
| Ypens1P | gross annual country specific pension income 1 ppp-adjusted (euro) |
| Ypens2P | gross annual country specific pension income 2 ppp-adjusted (euro) |
| Ypens3P | gross annual country specific pension income 3 ppp-adjusted (euro) |
| Ypens4P | gross annual country specific pension income 4 ppp-adjusted (euro) |
| Ypens5P | gross annual country specific pension income 5 ppp-adjusted (euro) |
| Ypens6P | gross annual country specific pension income 6 ppp-adjusted (euro) |
| Ypens7P | gross annual country specific pension income 7 ppp-adjusted (euro) |
| Ypens8P | gross annual country specific pension income 8 ppp-adjusted (euro) |
| Ypens9P | gross annual country specific pension income 9 ppp-adjusted (euro) |
| Ypens10P | gross annual country specific pension income 10 ppp-adjusted (euro) |
| Ypens11P | gross annual country specific pension income 11 ppp-adjusted (euro) |
| Yreg_1P | gross annual country specific regular payment 1 ppp-adjusted (euro) |
| Yreg_2P | gross annual country specific regular payment 2 ppp-adjusted (euro) |
| Yreg_3P | gross annual country specific regular payment 3 ppp-adjusted (euro) |
| Yreg_4P | gross annual country specific regular payment 4 ppp-adjusted (euro) |
| Yreg_5P | gross annual country specific regular payment 5 ppp-adjusted (euro) |

| | |
|---|---|
| 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 |
| Ip1F | flag for the gross annual country specific pension income 1 amount |
| Ip2F | flag for the gross annual country specific pension income 2 amount |
| Ip3F | flag for the gross annual country specific pension income 3 amount |
| Ip4F | flag for the gross annual country specific pension income 4 amount |
| Ip5F | flag for the gross annual country specific pension income 5 amount |
| Ip6F | flag for the gross annual country specific pension income 6 amount |
| Ip7F | flag for the gross annual country specific pension income 7 amount |
| Ip8F | flag for the gross annual country specific pension income 8 amount |
| Ip9F | flag for the gross annual country specific pension income 9 amount |
| Ip10F | flag for the gross annual country specific pension income 10 amount |
| Ip11F | flag for the gross annual country specific pension income 11 amount |
| Ireg1F | flag for the gross annual country specific regular payment 1 amount |

| | |
|---|---|
| Ireg2F | flag for the gross annual country specific regular payment 2 amount |
| Ireg3F | flag for the gross annual country specific regular payment 3 amount |
| Ireg4F | flag for the gross annual country specific regular payment 4 amount |
| Ireg5F | flag for the gross annual country specific regular payment 5 amount |

Household level Variables

| | |
|---|---|
| YhhE | gross annual household income in euros |
| YhiE | income from other household members in euros |
| YothE | other household benefits in euros |
| YrentE | rent value at household level in euros |
| YirentE | imputed rent value at household level in euros |
| YbaccE | bank account at household level in euros |
| YbondE | government or corporate bonds at household level in euros |
| YstocE | stocks or shares at household level in euros |
| YfundE | mutual funds at household level in euros |
| 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 |

# PART II

## MULTIPLE IMPUTATIONS

### 6. Multiple Imputations

### 6.1. Background on multiple imputations[8]

Statistical inference with missing data is an important applied problem, because missing data (planned or unplanned) are commonly encountered in practice. Multiple imputations (MI) is one of the available procedures for analysing data sets with missing values entries [Little and Rubin ( 2002)].
MI is the technique that replaces each missing value with *M* (M=2 or more) acceptable values representing a distribution of possibilities. The *M* values are ordered in the sense that the first components of the vectors for the missing values are used to create one completed data set, the second components of the vectors for the missing values are used to create the second completed data set and so on. Thus, the *M* imputations for each missing datum create *M* complete data sets. Standard complete-data methods can be used to analyse each data set.

Advantages of MI:

- MI allows to analyse completed data sets (standard methods may be used).
- In many contexts the data collector is different from the data analyst, but data collector may have access to more and better information about non-respondents than the data analyst. This kind of information can often improve the imputed values.
- MI allows data collectors to reflect their uncertainty as to which values to impute: the resulting *M* complete-data analyses can be easily combined to create an inference that validly reflects sampling variability because of the missing values.
- If the method to create imputations is 'proper' [Rubin(1987)], then the resulting inferences will be statistically valid.
- Using MI, the missing data problem can be handled once, rather than many times, by the users. This implies consistency of the data-bases across users and a consequent consistency of answers from identical analyses.

The *M* complete-data analyses based on the *M* repeated imputations are then combined to create one repeated-imputation inference. Let $\hat{q}_l$, $U_l$, $l$=1,…, *M* be *M* complete-data estimates and their associated variances for a parameter $q$, calculated from the *M* repeated imputations under one model (e.g. OLS regression estimates). The final estimate of $q$ is

$$\bar{q} = \sum_{l=1}^{M} \frac{\hat{q}_l}{M}$$

The variability associated with this estimate has 2 components: the average within-imputation variance,

[8] This section has been contributed by the Padua Team.

$$\bar{U} = \sum_{l=1}^{M} \frac{U_l}{M}$$

and the between-imputation component

$$B = \frac{\sum_l \left(\hat{q}_l - \bar{q}\right)^2}{M-1}$$

The total variability associated with $q$ is

$$T = \bar{U} + \frac{M+1}{M} B$$

where $(M+1)/M$ is an adjustment for finite $M$. With scalar $q$, the approximate reference distribution for interval estimates and significance tests is a $t$ distribution

$$\left(q - \bar{q}\right) T^{-1/2} \sim t_v$$

where the degrees of freedom

$$v = (M-1)\left[1 + \frac{1}{M+1}\frac{\bar{U}}{B}\right]^2$$

is based on a Satterthwaite approximation [Rubin and Schenker (1986)].

### 6.2. Some references on multiple imputations


Little, J.A. and D.B. Rubin, 2002. *Statistical Analysis with Missing Data*, second edition. New York, John Wiley & Sons.

Rubin, D.B., 1987. *Multiple Imputation for Nonresponse in Surveys*. New York, John Wiley & Sons.

Rubin, D.B. and N. Schenker , 1986. Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse. *Journal of the American Statistical Association*, 81, pp. 366-374.


### 6.3. Multiple imputations in the generated income programs package

The Stata income programs package described here performs only one round of imputations for each variable using country-specific univariate conditional hotdeck 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 datasets, INCOMEi.dta (i=1,…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 Maintest2004 Share data,

- INCOME1.DTA uses Asset_R11 and seed = 123456789 (Stata's default)
- INCOME2.DTA uses Asset_R12 and seed = 1000
- INCOME3.DTA uses Asset_R13 and seed = 10000
- INCOME4.DTA uses Asset_R14 and seed = 100000
- INCOME5.DTA uses Asset_R15 and seed = 1000000

The assets datasets have been kindly provided by the Salerno team.[9]

[9] AssetRi needs to be renamed to Asset_R1 to run the income programs package, and that each run produces as output INCOME.dta, which we renamed to INCOMEi.dta.