9 User Guide

9.1 Introduction

As we have seen in the previous chapters, the HFCS data have several special features that must be taken into account when analyzing these data. The data are multiply imputed, they have survey weights and replicate weights. The HFCS data also come in several files, accounting for the structure of the survey. These files are distinguished by the level of data (household vs. personal), by the number of implicates (i.e. each implicate is a separate file), and by type, i.e. whether the data were collected or constructed (derived variables, i.e. aggregate variables, and replicate weights vs. survey variables). This chapter[1] provides Stata®[2] code that the user can employ step by step to account for all of these features.[3] Part of the code was provided by Sébastien Pérez-Duarte[4] of the ECB and is now available in a slightly altered and extended version. The ECB is expected to make available several program codes in spring 2013 as well. In this chapter, the program code for Stata® is given in the light blue-shaded area; it can be copied to the Stata® command window,[5] and has to be run according to the steps that are laid out below (an alteration of the sequence may invalidate the code). Additionally, the online appendix contains a do-file "user_guide.do" with all the steps that are laid out below.[6] This chapter first explains a way of merging the separate files, followed by the description of one possibility to set up the structure to use the imputations and survey information. Finally, some examples of simple estimation commands are provided in order to exemplify their use.

9.2 Merging the Data Files

The core HFCS data, which contain all internationally agreed variables, consist of the five multiple imputation samples or implicates at household level (files H1–H5), the corresponding sample at person level (files P1–P5) and the corresponding set of aggregate variables[7] (files D1–D5). Before we start creating a new data set containing all these files, the user has to provide the computer-specific path to the data sets and the folder of the do-files that are used later on. The variables used for merging are the household identifier "sa0010", the implicate number "im0100" and the country identifier "sa0100".

So-called M-files are created by reshaping the P-files (using the – `reshape` – command), including the appropriate naming of the P-file variables, and by

[1] *The authors do not make a judgement about which programs to use in a specific set-up. In particular if the size of the subsamples varies in each iteration, the estimation of discontinuous estimators does not comply with the assumptions underlying the findings on multiply imputed data in the literature (e.g. Little and Rubin, 2002). It is the responsibility of the user to check the validity and appropriateness of certain estimation commands in any given circumstances.*

[2] *The codes were written for Stata® version 12.1 and are not valid for previous Stata® versions. Comments are provided to facilitate the use of Stata® version 11.2.*

[3] *Any changes to and improvements of the code are made available on an ongoing basis in the online appendix.*

[4] *Principal Economist Statistician in the Statistics Development/Coordination Division of the ECB.*

[5] *Due to the way Stata handles line breaks, they may need to be deleted if the program code is copied by hand.*

[6] *The two macros containing the individual path to the data and the additionally provided do-files must be specified before execution. Due to the size and structure of the data and depending on the specification of the software and hardware, the execution of the do-file may take some time.*

[7] *The ECB is expected to make available the definitions of the derived variables in spring 2013.*

merging the resulting data set with the H-files. They are provided in *wide format,*[8] i.e. one line of the data matrix includes information on one particular household and the information on each person within a household is put in a separate variable. Merged with the D-files, these M-files yield the entire HFCS data set in the file "hfcs.dta".

```
*********************************************************************
***Merging the files of the HFCS data
*********************************************************************

*Set macro for the path to the data (must be specified by the user)
global hfcsdata="path to the appropriate folder where the data are stored"

*Set macro for the path to the do-files (must be specified by the user)
global hfcsdofile="path to the appropriate folder where the do-files are stored"

*Set working directory
cd "$hfcsdata"

*Merging the p and h files together (wide format)
forvalues i=1(1)5 {
  use "$hfcsdata\P`i'.dta", clear
  drop id hid survey
  foreach var of varlist sa0010-fra0500 {
    local `var'lab: variable label `var'
  }
  reshape wide ra0?0* fra0?0* p* fp*, i(sa0010 sa0100) j( ra0010)
  foreach j of varlist ra* fra* p* fp* {
  local last2car=substr("`j'", `=length("`j'")-1', 1)
    local last1car=substr("`j'", length("`j'"), 1)
    if "`last2car'"=="1" {
      local firstcar=substr("`j'",1, `=length("`j'")-2')
      rename `j' `firstcar'_`last2car'`last1car'
      label variable `firstcar'_`last2car'`last1car' "``firstcar'lab' - `last2car'`last1car'"
    }
    else {
      local firstcar=substr("`j'",1, `=length("`j'")-1')
      rename `j' `firstcar'_`last1car'
      label variable `firstcar'_`last1car' "``firstcar'lab' - `last1car'"
    }
  }
  save "$hfcsdata\P`i'_temp.dta", replace
  clear
  use "$hfcsdata\H`i'.dta", clear
  merge 1:1 sa0010 sa0100 im0100 using "$hfcsdata\P`i'_temp.dta",nogen
  save "$hfcsdata\M`i'.dta", replace
  erase "$hfcsdata\P`i'_temp.dta"
}
```

[8] *It is also possible to merge the data files in "long" format using an almost identical code without needing to reshape the personal files.*

```
*Merging the core with the derived variables
forvalues i=1(1)5 {
  use "$hfcsdata\M`i'.dta", clear
  merge 1:1 sa0010 im0100 sa0100 using "$hfcsdata\D`i'.dta"
  save "$hfcsdata\temp`i'.dta", replace
}

*Merging the implicates together[1]
use "$hfcsdata\temp1.dta", clear
forvalues j=2(1)5 {
  append using "$hfcsdata\temp`j'.dta"
}

*Drop unnecessary variables and labels
drop _merge
label drop _merge

*Save the HFCS data
save "$hfcsdata\hfcs.dta", replace
```

[1] *The temporary files are kept for setting up the multiple imputed data and erased following this procedure.*

9.3 Multiple Imputation

The next step is to import into Stata's® `mi` (i.e. – `mi estimate` – commands for an appropriate use of the multiple imputation structure) both the original data and the imputed values. As the original data are not part of the HFCS data files, we have to construct them from the information about whether observations vary across implicates (indicating multiple imputation and, hence, missing values) and from the information about missing values taken from the flags.[9] Finally, original and imputed data must be imported and registered. Users should take note of the macro "`IMPUTEDVARS`" in the program code below, which contains a string listing all imputed variables after execution of the corresponding loop. If registration was successful only a few variables (e.g. the implicate number "im0100") and the flags should appear as "unregistered varying" when typing – `mi varying` – .

[9] *All missing values (including "Don't know", "No answer" and filter missings are set "."). The flags make it possible to differentiate between these types of missing values (filter missing observations were flagged with a "0"). Flag variables have the same variable name, but start with an "f" preceding the variable name.*

```
*********************************************************************
***Preparing the data for mi import
*********************************************************************

*Create the zero implicate to simulate the original data
*Use one implicate of the data
use "$hfcsdata\temp1.dta", clear
*Replace the implicate number by "0" to simulate the original data
replace im0100=0
*Append all other implicates
append using "$hfcsdata\hfcs.dta"

*For some reason string variables do not play well with mi-commands and need to
be encoded into numeric variables.
foreach var of varlist hb* hc* hd* hg* hh* hi* pa* pe* pf* pg* ra* sa0100 sb1000 {
  capture confirm numeric variable `var'
  if _rc {
    rename `var' `var'_string
    encode `var'_string, gen(`var')
    drop `var'_string
  }
}

*Set as soft missing (".") in im0100==0 all values varying, and also those whose
flags set them as imputed
global IMPUTEDVARS=""
foreach var of varlist hb* hc* hd* hg* hh* hi* pa* pe* pf* pg* ra* {
  capture confirm numeric variable `var'
  if !_rc {
    tempvar sd count
    quietly bysort sa0100 sa0010 : egen `sd'=sd(`var')
    quietly bysort sa0100 sa0010 : egen `count'=count(`var')
    quietly count if ((`sd'>0 & `sd' <.)|`count'<6|(f`var'>4000 & f`var'<5000))&im0100==0
    if r(N)>0 global IMPUTEDVARS "$IMPUTEDVARS `var'"
    quietly replace `var'=. if ((`sd'>0&`sd' <.)|`count'<6|(f`var'>4000&f`var'<5000))&im0100==0
    drop `sd' `count'
    disp ".", _continue
  }
}

*Drop unnecessary variables
drop id _merge

*Save the hfcs data
save "$hfcsdata\hfcs.dta", replace

*Erase temporary files that will not be needed anymore
forvalues i=1(1)5 {
  erase "$hfcsdata\temp`i'.dta"
}
```

```
*********************************************************************
***Import as multiply imputed data
*********************************************************************

*Import the imputation structure of the data into Stata
mi import flong, m(im0100) id(sa0100 sa0010) clear

*Register the variables that are imputed
mi register imputed $IMPUTEDVARS

*Check whether all imputed variables are registered
mi varying

*Save the hfcs-data with mi structure
save "$hfcsdata\hfcs.dta", replace
```

9.4 Survey Variables

Once having declared the data as multiply imputed, we can proceed by designating the data as complex survey data. To this end we identify the variables that contain information about the survey design and specify the default method of variance estimation. In our case, all this information is contained in the final survey weights (hw0010) and in the 1,000 sets of replicate weights (wr0001–wr1000), which are provided in a separate file and hence have to be merged with the data first.

```
*********************************************************************
***Setting up Complex Survey Design
*********************************************************************

*Encode country indicator
use "$hfcsdata\W.dta", clear
rename sa0100 sa0100_string
encode sa0100_string, gen(sa0100)
drop sa0100_string
save "$hfcsdata\Wtemp.dta", replace

*Using the hfcs-data with mi structure
use "$hfcsdata\hfcs.dta", clear

*Merging the data with replicate weights
merge m:1 sa0100 sa0010 using "$hfcsdata\Wtemp.dta"

*Drop unnecessary variables and files
drop _merge
erase "$hfcsdata\Wtemp.dta"

*Setting the appropriate survey structure using replicate weights
mi svyset [pw=hw0010], bsrweight(wr0001-wr1000) vce(bootstrap)

*Save the HFCS-data with mi svyset structure
save "$hfcsdata\hfcs.dta", replace
```

9.5 Standard Estimation Procedures

The data are now ready to be analyzed in Stata®. We write – `mi estimate: svy:` – followed by the estimation command in question. Stata® will then provide correctly calculated estimates and standard errors, taking into account both the multiple imputation framework and the replicate weights.[10] The – `esampvaryok` – option can be useful in case that the sample size varies across implicates due to imputations.[11] Stata® versions below Stata® 12 do not allow the use of replicate weights together with multiply imputed data. Thus an underlying command[12] called – `u_mi_estimate` – has to be modified before the standard estimation commands can be used. In Stata® 12 the option – `vceok` – (used after the – `mi estimate` – command, e.g. "`mi estimate, vceok:`…") can be used instead. It should be noted that in order to be able to calculate the correct variance for subsamples of households (see second example in the following program code), Stata® requires a dummy variable for each of these subsamples combined with the use of the option for subpopulations (e.g. "`…svy, subpop(dummy)…`").[13] Alternatively, it is possible to use the option – `over(variable)` – for certain estimation commands (see last example in the following program code).

```
*********************************************************************
***Using Standard Estimation Procedures
*********************************************************************

*Using the HFCS-data with mi svyset structure
use "$hfcsdata\hfcs.dta", clear

*Modified Stata command, which should be run before the estimation commands for
versions of Stata previous to Stata 12 (always update to the most recent Stata
version). Furthermore, exclude in this case the vceok option in the estimations
below.
*do "$hfcsdofile\modified u_mi_estimate 11.2.do"

*Mean of current value of primary housing unit
mi estimate, esampvaryok vceok: svy: mean hb0900
```

[10] *A correct point estimate of statistics can be carried out on the basis of the final survey weights. Replicate weights are needed to calculate a variance estimator.*

[11] *Rubin's combination rules (e.g. Little and Rubin, 2002) were derived assuming that the same set of observations is used in each imputed data set. They may thus not necessarily apply when the sets of observations used in the data analysis differ. This is why – `mi estimate` – errors out when this happens. In some cases, when the subsets used in each complete data analysis do not differ too much the conventional formulas may still be applicable. In this case, it is the choice of the user to use the – `esampvaryok` – option or find a better way to deal with the violation of the assumption of Rubin's combination rules described above. To our knowledge, this issue has not yet been addressed in the literature.*

[12] *The necessary alteration was written by the ECB and is provided as a separate do-file in the online appendix.*

[13] *The use of an if-condition does not account for the uncertainty of the subsample size and therefore yields wrong variance estimators.*

```
*Mean of current value of primary housing unit for part owner of the primary
housing unit
gen partowner=(hb0300==2)
mi estimate, esampvaryok vceok: svy, subpop(partowner): mean hb0900

*Proportions of owner/renter of primary housing unit
mi estimate, esampvaryok vceok: svy: proportion hb0300

*Ratio of current to acquisition value of primary housing unit
mi estimate, esampvaryok vceok: svy: ratio hb0900 hb0800

*Regression of current value of primary housing on acquisition value and year of
acquisition
mi estimate, esampvaryok vceok: svy: regress hb0900 hb0800 hb0700

*Average level of deposits according to gender of the first person
mi estimate, esampvaryok vceok: svy: mean da2101, over(ra0200_1)
```

9.6 Using Additional Estimation Procedures

Calculating a median or another quantile requires a different Stata® package. This program is called – `medianize` – and provided by the ECB (the respective do-file can be found in the online appendix). This program must be used with caution as it has so far been tested only in limited environments. Furthermore, it uses the command – `tabstat` – and the analytical weights option of this command in Stata®. To our knowledge, however, there is presently no command to estimate discontinuous estimators (like median or other percentiles) other than the use of these ad-hoc measures.

```
**********************************************************************
***Including Additional Estimation Procedures
**********************************************************************

*ECB-written command to calculate medians (and some other quantile statistics),
which should be run before the estimation command
capture program drop medianize
do "$hfcsdofile\medianize.do"

*Median of amount still owned in the first loan collateralized with primary
housing unit
mi estimate, esampvaryok vceok: svy: medianize hb1701

*Median of amount still owned in the first loan collateralized with primary
housing unit over gender of first person
mi estimate, esampvaryok vceok: svy: medianize hb1701, over(ra0200_1)

*10th percentile of amount still owned in the first loan collateralized with
primary housing unit over gender of first person
mi estimate, esampvaryok vceok: svy: medianize hb1701, over(ra0200_1) stat(p10)
```

9.7 Online Appendix

The online appendix contains the Stata® code described above and the do-files that are necessary to use replicate weights together with multiple imputation in older Stata® versions and to estimate certain quantiles. The online appendix is intended to be updated on an ongoing basis to include program codes for various other HFCS-relevant topics. Every additional do-file will be supplemented with appropriate documentation.