9 User Guide

9.1 Einleitung

Wie in den vorangegangenen Kapiteln dargelegt wurde, zeichnen sich die HFCS-Daten durch einige Besonderheiten aus, die bei der Analyse berücksichtigt werden müssen. So sind die Daten multipel imputiert und verfügen über Survey- und Resampling-Gewichte. Darüber hinaus sind sie aufgrund der Struktur des Surveys in mehreren Files abgelegt. Diese Files unterscheiden sich hinsichtlich der Datenebene (Haushalts- oder Personenebene), der Nummer der Implicates (d. h., jedes Implicate ist eine eigene Datei) und darin, ob sie konstruiert oder erhoben wurden (Derived Variables, d. s. aggregierte Variablen, und Resampling-Gewichte versus Survey-Variablen). In diesem Kapitel[1] wird dem Benutzer ein Programmcode in Stata®[2] zur Verfügung gestellt, mit dem bei schrittweiser Anwendung all diesen Charakteristika Rechnung getragen werden kann.[3] Der Code wurde auszugsweise von Sébastien Pérez-Duarte[4] von der EZB bereitgestellt und liegt nun in einer leicht geänderten und erweiterten Fassung vor. Voraussichtlich wird auch die EZB im Frühjahr 2013 diverse Programmcodes zur Verfügung stellen. Der Programmcode findet sich jeweils in den blau-hinterlegten Passagen. Er kann direkt in das Stata®-Befehlsfenster[5] kopiert werden und muss gemäß der unten beschriebenen Sequenz ausgeführt werden (eine Änderung der Abfolge könnte den Code unbrauchbar machen). Zusätzlich dazu enthält der Online-Anhang ein Do-File „user_guide.do“ mit den im Folgenden beschriebenen Schritten.[6] Zunächst soll hier eine Möglichkeit, die einzelnen Files zusammenzuführen, erklärt werden. Anschließend wird ein Vorschlag zur Errichtung einer Struktur für die Imputationen und Survey-Informationen beschrieben. Beispiele für einfache Schätzbefehle sollen abschließend deren Verwendung exemplarisch darstellen.

9.2 Zusammenführung der Datenfiles

Die HFCS-Core-Daten, in denen alle international akkordierten Variablen enthalten sind, bestehen aus den fünf multipel imputierten Samples bzw. Implicates auf Haushaltsebene (Files H1–H5), den entsprechenden Samples auf Personenebene (Files P1–P5) und dem entsprechenden Set aggregierter Variablen[7] (Files D1–D5). Bevor mit der Erstellung eines neuen Datensatzes mit all diesen Files begonnen werden kann, muss der rechnerspezifische Pfad zu den Datensätzen und der Folder

[1] *Die Autoren nehmen von einer Beurteilung der in einem bestimmten Setup zu verwendenden Programme Abstand. So erfüllt insbesondere die Schätzung nichtlinearer Statistiken, sofern die Größe der Subsamples in den einzelnen Iterationen variiert, die Annahmen der in der Literatur bewiesenen Ergebnisse (siehe z. B. Little und Rubin, 2002) zu multiplen imputierten Daten nicht. Es liegt am Benutzer, die Validität und Eignung einzelner Schätzbefehle zu den jeweils gegebenen Bedingungen zu überprüfen.*

[2] *Die Codes wurden für die Stata®-Version 12.1 verfasst und sind nicht mit älteren Stata®-Versionen zu verwenden. Für die Stata®-Version 11.2 werden Kommentare zur Verfügung gestellt, die eine Verwendung ermöglichen. Aufgrund der klareren Gestaltung des Do-Files wird auf eine Übersetzung der Kommentare ins Deutsche im Programmcode verzichtet.*

[3] *Etwaige Änderungen bzw. Verbesserungen des Codes werden laufend im Online-Anhang aktualisiert werden.*

[4] *Principal Economist Statistician in der Statistics Development/Coordination Division der EZB.*

[5] *Aufgrund der Verarbeitung von Zeilenumbrüchen in Stata® müssen diese bei händischem Kopieren des Programmcodes eventuell gelöscht werden.*

[6] *Die zwei Makros mit dem individuellen Pfad zu den Daten und zu den zusätzlich angeführten Do-Files müssen vor der Ausführung spezifiziert werden. Angesichts der Größe und Struktur der Daten und je nach Software- bzw. Hardwarespezifikationen kann die Ausführung des Do-Files längere Zeit in Anspruch nehmen.*

[7] *Die EZB wird voraussichtlich im Frühjahr 2013 die Definitionen der aggregierten Variablen zur Verfügung stellen.*

der anschließend zu verwendenden Do-Files definiert werden. Die zum Zusammenspielen verwendeten Variablen sind die Haushaltsidentifikation „sa0010", die Implicate-Nummer „im0100" und die Länderidentifikation „sa0100".

```
*********************************************************************
***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"
}
```

```
*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] *Die temp-Files werden für die Konfiguration der multipel imputierten Daten behalten und erst nach Abschluss dieses Schrittes gelöscht.*

Durch Umformung der P-Files (mit dem Befehl – `reshape` –), inklusive einer geeigneten Benennung und Beschriftung der P-File-Variablen, und durch Zusammenführung des daraus resultierenden Datensatzes mit den H-Files werden die sogenannten M-Files erstellt. Sie sind im „wide"-Format,[8] d. h., eine Zeile der Datenmatrix enthält die Informationen zu einem bestimmten Haushalt, während die Informationen zu jeder Person innerhalb eines Haushalts in einer eigenen Variablen festgehalten wird. Diese M-Files ergeben mit den D-Files zusammengeführt die gesamten Daten des HFCS im File „hfcs.dta".

9.3 Multiple Imputationen

Im nächsten Schritt werden sowohl die Originaldaten als auch die imputierten Samples in Statas® mi (d. h. `mi estimate`-Befehle zur geeigneten Anwendung der Struktur der multiplen Imputation) importiert. Da die Originaldaten nicht Teil der HFCS-Datenfiles sind, müssen sie aus den Informationen darüber, ob die Beobachtungen in den einzelnen Implicates variieren (was auf multiple Imputationen und daher fehlende Werte hindeuten würde), sowie den Informationen über fehlende Werte aus den Flags konstruiert werden.[9] Zuletzt müssen die originalen und imputierten Daten importiert und registriert werden. An dieser Stelle wird auf das Makro „`IMPUTEDVARS`" im unten stehenden Programmcode verwiesen, das einen String mit allen imputierten Variablen enthält, nachdem der entsprechende Loop ausgeführt wurde. Bei erfolgreicher Registrierung sollten nur einige wenige Variablen (z. B. die Implicate-Nummer „im0100") und die Flags als „unregistered varying" nach Eingabe von – `mi varying` – erscheinen.

[8] *Es besteht auch die Möglichkeit, die Datenfiles im „long"-Format zusammenzuführen, wobei ein fast identer Code verwendet wird und die Files auf Personenebene nicht umgewandelt werden müssen.*

[9] *Alle fehlenden Werte, also sowohl Missing Values wie auch „Weiß nicht", „Keine Angabe" und Filter-Missings, werden auf „." gesetzt. Eine Unterscheidung zwischen diesen Arten von fehlenden Werten ist auf Basis der Flags (Flag „0" weist ein Filter-Missing der Beobachtung aus) möglich. Flag-Variablen haben denselben Variablennamen, allerdings ist diesem jeweils ein „f" vorangestellt.*

```
*********************************************************************
***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-Variablen

Nachdem die Daten als multipel imputiert konfiguriert wurden, können sie nun als komplexe Survey-Daten designiert werden. Dabei werden jene Variablen, die Informationen über das Survey-Design enthalten, identifiziert; die Default-Methode für die Varianzschätzung wird festgelegt. Im vorliegenden Fall finden sich all diese Informationen in den finalen Survey-Gewichten (hw0010) und in den 1.000 Sets von Resampling-Gewichten (wr0001–wr1000), die sich in einem separaten File befinden und daher zuerst mit den Daten zusammengeführt werden müssen.

```
*********************************************************************
***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 Standardschätzverfahren

Die Daten sind nunmehr für die Analyse in Stata® aufbereitet. Nach der Eingabe von – `mi estimate: svy:` – gefolgt von dem betreffenden Schätzbefehl ermittelt Stata® unter Berücksichtigung der multiplen Imputationen und der Resampling-Gewichte korrekte Schätzungen und ihre Standardfehler.[10] Wenn die Samplegröße aufgrund der Imputationen je nach Implicate variiert, kann sich die Option – `esampvaryok` – als nützlich erweisen.[11] Die gleichzeitige Verwendung von Resampling-Gewichten und multipel imputierten Daten ist in den Stata®-Versionen vor Stata® 12 nicht möglich. Aus diesem Grund muss der zugrundeliegende Befehl[12] – `u_mi_estimate` – vor der Verwendung von Standardschätzbefehlen modifiziert werden. In Stata® 12 kann stattdessen – `vceok` – (nach dem Befehl – `mi estimate` –, z. B. „ `mi estimate, vceok:`… “) verwendet werden. Es wird darauf hingewiesen, dass Stata®, um die korrekte Varianz für Untergruppen (Subsamples) von Haushalten zu berechnen, die Definition einer Dummy-Variable für die jeweilige Untergruppe zusammen mit der Verwendung der Option für Subsamples (d. h. „ …`svy, subpop(dummy)`… “) benötigt (siehe zweites Beispiel im folgenden Programmcode).[13] Alternativ kann die Option – `over(variable)` – bei bestimmten Schätzbefehlen verwendet werden (siehe letztes Beispiel im folgenden Programmcode).

```
*********************************************************************
***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] *Eine korrekte Punktschätzung einer Statistik kann auf Basis der finalen Survey-Gewichte durchgeführt werden. Für die Berechnung der Varianz eines Schätzers werden die Resampling-Gewichte benötigt.*

[11] *Die Kombinationsregeln nach Rubin (siehe z. B. Little und Rubin, 2002) basieren auf der Annahme, dass in jedem Satz imputierter Daten dieselben Sets an Beobachtungen zur Anwendung kommen. Daher könnte es sein, dass die Regeln nicht gelten, wenn bei der Datenanalyse verschiedene Sets an Beobachtungen verwendet werden. Aus diesem Grund generiert –* `mi estimate` *– in diesem Fall eine Fehlermeldung. Wenn sich die Subsets in jeder fertigen Datenanalyse nicht zu sehr unterscheiden, könnten die herkömmlichen Formeln durchaus anwendbar sein. In diesem Fall kann sich der Benutzer entweder für die Option –* `esampvaryok` *– entscheiden oder eine andere Methode anwenden, um das Problem des Nichtzutreffens der oben genannten Annahme von Rubins Kombinationsregeln zu bewältigen. Den Autoren ist bis dato keine Abhandlung dieser Frage in der Literatur bekannt.*

[12] *Die notwendige Änderung stammt von der EZB und wird in einem separaten Do-File im Online-Anhang zur Verfügung gestellt.*

[13] *Die Verwendung einer if-Einschränkung beachtet die Unsicherheit der Größe des Subsamples nicht und liefert daher falsche Varianzschätzer.*

```
*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 Zusätzliche Schätzverfahren

Zur Berechnung eines Medians oder eines anderen Quantils muss ein anderes Stata®-Package verwendet werden. Das entsprechende Programm heißt – `medianize` – und stammt von der EZB (das Do-File befindet sich im Online-Anhang.). Bei der Verwendung des Programms ist allerdings Vorsicht geboten, da das Package bisher nur in einer eingeschränkten Umgebung getestet wurde. Darüber hinaus kommen der Befehl – `tabstat` – sowie die analytische Gewichtungsoption dieses Befehls in Stata® zur Anwendung. Allerdings gibt es nach dem Wissensstand der Autoren derzeit abgesehen von der Verwendung von solchen Ad-hoc-Verfahren keinen Befehl zur Schätzung nichtlinearer Statistiken (wie Medianen oder anderen Perzentilen).

```
**********************************************************************
***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-Anhang

Im Online-Anhang befindet sich der oben beschriebene Stata®-Code sowie die Do-Files, die für die Verwendung von Resampling-Gewichten in Kombination mit multipler Imputation bei älteren Stata®-Versionen bzw. für die Schätzung bestimmter Quantile notwendig sind. Sie soll laufend mit Programmcodes für verschiedene HFCS-relevante Themen aktualisiert werden. Jedem zusätzlichen Do-File wird eine entsprechende Dokumentation beigefügt.