Using the RAND HRS Longitudinal File and RAND HRS Fat Files

Sample Programs for HRS Workshop

This document is intended to provide users with some examples of how to both set up and perform some simple descriptive analyses in SAS and Stata, using the RAND HRS data products. The sample programs assume that the data being used are set up in a folder called **C:\RandHRS**, and that any files you create are stored in a separate folder called **C:\MyPaper**. If you are storing your files and output a different folder, note that your paths/libnames will be different.

For space constraints, the RAND HRS Longitudinal File used in this lab exercise contains only the variables listed below. The full file is available for download from the University of Michigan at https://hrsdata.isr.umich.edu/data-products/rand (user registration required).

- HHIDPN: Individual respondent identifier, which is a combination of HHID and PN
- INW1 – INW15: These are flags that indicate whether an individual responded at a given wave, where 0 = "Non-response" and 1 = "Response, Alive."
- R13IWSTAT – R15IWSTAT: These variables provide another way to examine interview status. In addition to telling you whether the individual responded or not, these variables also indicate mortality status.
- HACOHORT: This variable identifies the entry cohort subsample. The responses include 0 = "HRS/AHEAD overlap," 1 = "AHEAD," 2 = "CODA," 3 = "HRS," 4 = "WB," 5 = "EBB," 6 = "MBB," and 7 = "LBB").
- RAGENDER: This variable identifies the gender of the respondent, where 1 = "Male" and 2 = "Female."
- RARACEM: This variable identifies the race of the respondent, where 1 = "White/Caucasian," 2 = "Black/African American," and 3 = "Other".
- R13SHLT – R15SHLT: These variables indicate, for Waves 13 (2016), 14 (2018), and 15 (2020), the respondent's self-report of health. The responses include 1 = "Excellent," 2 = "Very good," 3 = "Good," 4 = "Fair," and 5 = Poor." The corresponding variables for the spouse are S12SHLT – S14SHLT.
- R13OOPMD – R15OOPMD: These variables provide, for Waves 13 (2016), 14 (2018), and 15 (2020), the respondent's out of pocket medical expenditure for the last 2 years (or since the last interview). The corresponding variables for the spouse are S13OOPMD – S15OOPMD.
- R13AGEY_E – R15AGEY_E: These variables, for Waves 13 (2016), 14 (2018), and 15 (2020), indicate the age of the respondent (in years) at the final interview date. The corresponding variables for the spouse are S13AGEY_E – S15AGEY_E.

There are five parts to this lab, each of which should be run in the specified order:

Part #1: Initial Analysis

Run descriptive analyses using variables from the RAND HRS Longitudinal File (randhrs1992_2020v2).

- Run cross-tabulations (frequency tables) on variables that describe the interview response status during each wave and see how these variables interact with each other.
  - Hint: INW13 – INW15 and R13IWSTAT-R15IWSTAT are the two response status variables

Part #2: Subset the RAND HRS Longitudinal File

Create a file that contains a subset of variables from the RAND HRS Longitudinal File (randhrs1992_2020v2) and subset to those who responded in Waves 13, 14, or 15.

- Call the file wkshop1
- Hint: Use the INW13-INW15 vars to subset.

Part #3: Run Descriptive Analyses

Run descriptive analyses using the dataset created in **Part #2**.

- Create an indicator variable that the Respondent is age 80 years or older in Wave 15.
  - Name the indicator variable age80.

- Produce frequency tables that examine/compare the following:
  - Respondents' self-report of health by the created age indicator variable (age80) for Wave 15.
  - Respondents' self-report of health in Wave 15 by the Respondent's self-report of health in Wave 14.
  - Respondents' cohort by the created age indicator variables (age80) for Wave 15.

- Generate summary statistics for respondents' out of pocket medical expenditures:
  - For all Respondents in Waves 13 – 15.
  - By gender in Wave 15.
  - By the created age indicator variables (age80) in Wave 15.
  - By the Respondent's self-report of health in Wave 15.

Part #4: Merge the RAND HRS Longitudinal File to the RAND HRS Fat Files

Merge the dataset created in **Part #2** with the RAND HRS Fat Files for Waves 14 (HRS 2018) and 15 (HRS 2020). The raw variables that are selected from the fat files (described below) include questions that ask respondents how likely it is that they will spend $1,500 or more on out-of-pocket medical expenditures in the next year.

- Merge the dataset created in Part #2 with the fat files for Wave 14 (h18f2b) and Wave 15 (h20f1a).
  Important things to note:
  - For sorting and merging you can use HHIDPN.
  - All files are sorted by HHIDPN (HHID and PN).
    - Hint: Stata variable names are case-sensitive (all names are lower case) and sometimes the program requires that you re-sort files before merging. You can only merge two files at once.

- Keep HHIDPN, HHID and PN from the fat files, as well as the QP175 (Wave 14) and RP175 (Wave 15) variables described below.

- In this final merged dataset, keep only the individuals who responded to Waves 14 and 15.

- Call the file wkshop2

Part #5: Run analysis using the dataset created in Part #4.

- Create a 'lightly cleaned' version of the P175 variables and call them **p1500_14** (QP175) and **p1500_15** (RP175)
  - Things to note:
  - Recode the "Web non-response" (-8), "Don't Know" (998), and "Refuse" (999) answers to missing values.
  - For the purpose of this exercise, we're just recoding these values but in your real projects, when you are using variables from the RAND HRS Fat Files or the core HRS modules, you will usually have more cleaning to do. (This will be discussed more later this week). This exercise is to just show that the data will need to be cleaned in some way.
    - Use the HRS codebooks (https://hrsonline.isr.umich.edu/modules/meta/2018/core/codebook/h18p_r.htm#PQP175R, https://hrs.isr.umich.edu/sites/default/files/meta/2020/core/codebook/h20p_r.htm#PRP175R, and the clippings on the next page) as a reference.

- Create an indicator variable that =1 if the Respondent reported spending more than $3000 on out-of-pocket medical expenditure for the past two years in Wave 15.
  - Name the variable **spent3k_15**.
  - Make sure that if the Respondent has a missing value for R15OOPMD, the indicator variable is missing too.
    - Hint: SAS treats special missing values as low values, whereas Stata treats them as high values.
  - Note that R15OOPMD captures out of pocket medical expenditure for the last two years. For the purpose of this exercise, we're making an assumption that this expenditure was evenly distributed over the two years, but of course in actuality this may not always be the case.

- Produce some statistics that examine/compare the following:
  - Probability of spending $1,500 in the next year on out-of-pocket medical expenditures in Wave 14 (**p1500_14**) by actual reported out of pocket medical expenditure reported in Wave 15 (**spent3k_15**)
    - Also run using the raw variable (qp175) by spent3k_15 to see the importance of data cleaning when using variables from the RAND HRS Fat Files (or HRS data directly).
  - Probability of spending $1,500 in the next year on out-of-pocket medical expenditures in Wave 15 (**p1500_15**) by self-reported health in Wave 15 (R15SHLT)

QP175 PERCENT CHANCE R WILL PAY $1,500 MED EXP

Section: P Level: Respondent Type: Numeric Width: 3 Decimals: 0
Ref: SecP.EXPECTATIONS1.P175_

Please think about what you might spend out-of-pocket for your own medical expenses over the next year, including expenses such as doctor and dentist expenses, hospitals, nursing homes, prescription drugs and any others. Please include expenses that you would pay yourself (or a family member for you), but do not include what is covered by insurance.

On a scale from 0 to 100 (where 0 means absolutely no chance and 100 means absolutely certain), what are the chances that you will spend more than $1,500 during the coming year?

(0---10---20---30---40---50---60---70---80---90---100)

0 Absolutely no chance
100 Absolutely certain

N	Min	Max	Mean	SD	Miss
15781	0	100	43.02	36.54	936

31	-8.	Web non-response
298	998.	DK (Don't Know); NA (Not Ascertained)
100	999.	RF (Refused)

RP175 PERCENT CHANCE R WILL PAY $1,500 MED EXP

Section: P Level: Respondent Type: Numeric Width: 3 Decimals: 0
Ref: SecP.EXPECTATIONS1.P175_

Please think about what you might spend out-of-pocket for your own medical expenses over the next year, including expenses such as doctor and dentist expenses, hospitals, nursing homes, prescription drugs and any others. Please include expenses that you would pay yourself (or a family member for you), but do not include what is covered by insurance.

On a scale from 0 to 100 (where 0 means absolutely no chance and 100 means absolutely certain), what are the chances that you will spend more than $1,500 during the coming year?

(0---10---20---30---40---50---60---70---80---90---100)

0 Absolutely no chance
100 Absolutely certain

WEB-ADMINISTERED TEXT: [*removed for space purposes but same as above text*]

N	Min	Max	Mean	SD	Miss
14250	0	100	43.65	36.15	1036

51	-8.	Web non-response
286	998.	DK (Don't Know); NA (Not Ascertained)
100	999.	RF (Refused)

SAS Code

PREP: Creating a formats catalog, naming libraries

```
/*CALL YOUR LIBNAME STATEMENTS*/
/* This assumes the sasfmts.sas7bat and randhrs1992_2020v2.sas7bdat files are stored in
  the same folder/directory you are working out of. If you are storing them
  in a different directory, you would name that directory here.
  Windows PC Example:
  libname library "c:\RandHRS";
  libname mypaper "C:\MyPaper";
  Linux/Server Example:
  libname library "//servername/folder/a/RandHRS/";
  libname mypaper "//servername/folder/a/MyPaper/;  */

libname library ".";
libname mypaper ".";


/*CREATE YOUR FORMAT CATALOG*/
proc format library=library CNTLIN=library.sasfmts;
run;
/* Note that your formats catalog needs to be in a folder called library.
   See https://www.rand.org/well-being/social-and-behavioral-
policy/centers/aging/dataprod/faq.html#q14
   for more details.;
```

Part #1:

```
proc freq data=library.randhrs1992_2020v2;
  tables inw13*r13iwstat
         inw14*r14iwstat
         inw15*r15iwstat / missing list;
run;
```

Part #2:

```
data mypaper.wkshop1;
    set library.randhrs1992_2020v2 (keep = hhidpn hacohort ragender raracem
                                           inw13 inw14 inw15
                                           r13iwstat r14iwstat r15iwstat
                                           r13shlt   r14shlt   r15shlt
                                           s13shlt   s14shlt   s15shlt
                                           r13oopmd  r14oopmd  r15oopmd
                                           s13oopmd  s14oopmd  s15oopmd
                                           r13agey_e r14agey_e r15agey_e
                                           s13agey_e s14agey_e s15agey_e);
   where inw13=1 or inw14=1 or inw15=1;
run;

proc contents data=mypaper.wkshop1;
run;
```

Part #3:

```
data mypaper.wkshop1;
    set mypaper.wkshop1;
    if missing(r15agey_e) then age80 = .;
    else if r15agey_e>=80 then age80 = 1;
    else if r15agey_e <80 then age80 = 0;
run;

proc freq data = mypaper.wkshop1;
  tables r15shlt*age80
         r15shlt*r14shlt
         hacohort*age80 ;
run;

proc means data = mypaper.wkshop1;
    var r13oopmd r14oopmd r15oopmd;
run;

proc means data = mypaper.wkshop1;
    class ragender;
    var r15oopmd;
run;

proc means data = mypaper.wkshop1;
    class  age80;
    var r15oopmd;
run;

proc means data = mypaper.wkshop1;
    class r15shlt;
    var r15oopmd;
run;
```

Part #4:

```
data mypaper.wkshop2;
  merge mypaper.wkshop1
        library.h18f2b (keep=hhidpn hhid pn qp175)
        library.h20f1a (keep=hhidpn hhid pn rp175);
    by hhidpn;
    if inw14=1 and inw15=1;
run;
```

Part #5:

```
data mypaper.wkshop2;
  set mypaper.wkshop2;

  if qp175>=0 and qp175<=100 then p1500_14 = qp175;
  if rp175>=0 and rp175<=100 then p1500_15 = rp175;

    if missing(r15oopmd) then spent3k_15 = .;
  else if r15oopmd>=3000 then spent3k_15 = 1;
  else if r15oopmd <3000 then spent3k_15 = 0;

run;
```

Part #5: CONTINUED

```
proc means data = mypaper.wkshop2;
    class spent3k_15;
    var p1500_14;
run;

proc means data = mypaper.wkshop2;
    class spent3k_15;
    var qp175;
run;

proc means data = mypaper.wkshop2;
    class  r15shlt;
    var p1500_15;
run;
```

Stata Code

PREP: Naming Directories

```
clear all
version 14.1
set more off
set maxvar 20000
set matsize 800
set linesize 200
cap log close

/* I'm just using the local directory (i.e. programs, input datasets, and output
   datasets are all in the same location). But if your datasets are in a
   different location than your programs, you will need to declare your
   directories.

** WINDOWS PC EXAMPLE **
global library "c:\RandHRS"
global mypaper "C:\MyPaper"

** LINUX/SERVER EXAMPLE **
global library "//servername/folder/a/RandHRS/"
global mypaper "//servername/folder/a/MyPaper/"

** CALLING A FILE FROM A DIRECTORY EXAMPLE **
 use "$library/randhrs1992_2020v2.dta", clear
*/
```

Part #1:

```
use "randhrs1992_2020v2.dta", clear

tab inw13 r13iwstat, missing
tab inw14 r14iwstat, m
tab inw15 r15iwstat, m
```

Part #2:

```
#delimit ;
use hhidpn hacohort ragender raracem
    inw13 inw14 inw15
    r13iwstat r14iwstat r15iwstat
    r13shlt   r14shlt   r15shlt
    s13shlt   s14shlt   s15shlt
    r13oopmd  r14oopmd  r15oopmd
    s13oopmd  s14oopmd  s15oopmd
    r13agey_e r14agey_e r15agey_e
    s13agey_e s14agey_e s15agey_e
    using "randhrs1992_2020v2.dta";
#delimit cr

keep if inw13==1 | inw14==1 | inw15==1

save "wkshop1.dta", replace

describe
```

Part #3:

```
gen age80 = (r15agey_e>=80) if r15agey_e<.

save "wkshop1.dta", replace

tab r15shlt age80
tab r15shlt r14shlt
tab hacohort age80

sum r13oopmd r14oopmd r15oopmd

tabstat r15oopmd, by(ragender) stats(n mean min max)
tabstat r15oopmd, by(age80)    stats(n mean min max)
tabstat r15oopmd, by(r15shlt)  stats(n mean min max)
```

Part #4:

```
use hhidpn hhid pn qp175 using "h18f2b.dta"
sort hhidpn
save "h18x.dta", replace

use hhidpn hhid pn rp175 using "h20f1a.dta"
sort hhidpn
save "h20x.dta", replace

clear

use "wkshop1.dta"
sort hhidpn

merge 1:1 hhidpn using "h18x.dta", gen(mrg18)
merge 1:1 hhidpn using "h20x.dta", gen(mrg20)

keep if inw14==1 & inw15==1

tab mrg18 mrg20

save "wkshop2.dta", replace
```

Part #5:

```
gen p1500_14 = qp175 if qp175>=0 & qp175<=100
gen p1500_15 = rp175 if rp175>=0 & rp175<=100

gen spent3k_15 = (r15oopmd>=3000) if r15oopmd<.

save "wkshop2.dta", replace

tabstat p1500_14, by(spent3k_15) stats(n mean min max)
tabstat qp175,    by(spent3k_15) stats(n mean min max)
tabstat p1500_15, by(r15shlt)    stats(n mean min max)
```