Subject: Merging and appending data files
Posted by nora-dhs on Wed, 20 Jul 2022 08:48:07 GMT
View Forum Message <> Reply to Message

Hello, I am working with the DHS data and have some problems with merging different data files. This is what I want to do:

1) append data files of different survey waves and countries, e.g., append all HR data files for the African continent over the period 1999-2019 to one single data file called appended_HR. I then want to do the same with the other data types. I think I managed the first step and end up with four data files called appended_HR, appended_IR, appended_MR, appended_KR.

2) merge household characteristics and coordinates to the individuals. To this end, I want to merge appended_HR to the other appended files. To this end, I need unique identifiers. Here, I struggle. I noticed that some identifiers seem incorrectly coded (e.g., v001, v002, v003 are missing or do not correspond to mcaseid/caseid). I tried to solve these inconsistencies, but my approach does not work:

appended_HR:

duplicates tag v007 v000 v001 v002, gen(duple)

gen lhhid = strlen(hhid) // should be 12-character string

drop if duple != 0 & lhhid != 12 // drop if it's a duplicate and hhid is not of correct length

gen helpvar_v002 = substr(hhid,8,3) if duple != 0
destring helpvar_v002, gen(helpvar_v002num)
replace v002 = helpvar_v002num if duple != 0
drop helpvar_v002 helpvar_v002num duple

appended_IR, etc.:

duplicates tag v007 v000 v001 v002 v003, gen(duple)

gen lcaseid = strlen(casein) // should be 15-character string

drop if duple != 0 & lhhid != 15 // drop if it's a duplicate and caseid is not of correct length

gen helpvar_v002 = substr(caseid,8,3) // does not work, sometimes on another position
destring helpvar_v002, gen(helpvar_v002num) // does not work, Stata says: "contains nonnumeric characters; no generate"
replace v002 = helpvar_v002num if duple != 0
drop helpvar_v002 helpvar_v002num

gen helpvar_v003 = substr(caseid,11,2) // same here
destring helpvar_v003, gen(helpvar_v003num) // same here

replace v003 = helpvar_v003num if duple != 0
drop helpvar_v003 helpvar_v003num

Does anyone know how the correct approach would be?

Also, can you tell me what parts the caseid consists of in the below example? What does the "1" between "12" and "3" mean?

caseid .....12..1.3..4
v001 12
v002 3
v003 4

Thank you very much for your help!!

Subject: Re: Merging and appending data files
Posted by Janet-DHS on Thu, 21 Jul 2022 16:28:28 GMT
View Forum Message <> Reply to Message

Following is a response from DHS Research & Data Analysis Director, Tom Pullum:

The IR, MR, and KR files have individual women, men, or children as units. It would be virtually impossible to merge them with the HR file, which has households as units. You should use the PR file, which has individual household members as units, rather than the HR file.

You should do merges survey-by-survey and then append the merged files, rather than doing the appending first and then the merging.

The KR file includes children who are not in the PR file, and the PR file includes children who are not in the KR file. This is the trickiest merge and requires the use of b16 as well as v001 v002 v003.

I hardly ever use caseid or hhid in a merge. It is much easier to use the separate components of caseid, which are v001 v002 v003. There are survey-specific variations in the number of columns in caseid and hhid (which is made up of v001 v002) and in the number of columns assigned to the substrings for v001 v002 v003.

Some surveys in Francophone Africa have a sub-household code that must be used for merges.

Many surveys have survey-specific variables. Carrying them along will greatly increase the file size. Different surveys in the same country will have different coding for many variables, such as stratum, region, source of water, etc. I hope you are taking that into account and reducing the number of standard variables that you will keep.

My main advice would be that you do the merges for specific surveys and then append the surveys for each country. A massive file for all of the surveys done in Africa will be unwieldy.

Subject: Re: Merging and appending data files
Posted by nora-dhs on Mon, 01 Aug 2022 12:40:40 GMT
View Forum Message <> Reply to Message

Hi Tom,

thank you very much for your detailed answer!
I have some questions to your recommendations:
- You said it is impossible to merge the IR, MR, and KR files with the HR files. This surprises me, as I followed the "Guide to DHS Statistics, Version May 2020" to do the merging, see 1.51 ff. In Example 1, they match the household characteristics from the HR files to individual children from the KR files, which is exactly what I want to do.
- Why do you recommend to do the merging first and then to append the files?
- I also prefer to use v001 v002 v003 to do the merging rather than the caseid. However, in some cases v001 etc. are missing or wrongly coded, so I need to create them manually from caseid. Do you have any advice on how to do that best? As you said, caseid and hhid have survey-specific variations and differ in length and structure between the surveys, so I can't find a straightforward code that works for all surveys.

Thanks a lot for your help!
All the best, Nora

Subject: Re: Merging and appending data files
Posted by Bridgette-DHS on Mon, 01 Aug 2022 16:03:41 GMT
View Forum Message <> Reply to Message

Following is another response from DHS Research & Data Analysis Director, Tom Pullum:

I'll add some suggestions but they may not answer all your questions.

In the HR file, there is one very wide line of data for each household, with household members identifiee with subscripts that range from 1 to 20. The HR file can be very efficient for a merge for strictly household-level variables, such as water, sanitation, or the length of the household interview. However, matching the line number in the IR file (v003=1, 2, 3, etc) with the line number in the HR file (subscripts _01, _02, _03, etc) is just too much work. Maybe someone can do it, but I have never even tried! The "long" format of the PR file is simply much easier.

Merging and then appending, in that sequence, is simpler. If you append and then merge, you will have to match on a survey ID code, and the data files do not include a unique survey ID code. You may think that v000 is a survey identifier, but it is not. Two surveys conducted within the same phase of DHS (for example the current phase is 8) will have the same value of v000. Also if

you append first you will have an extremely long file (lots of cases) and the data processing time will go way up. Merges in individual surveys are very fast.

There are a few old surveys in which v001 is missing but in those surveys it is given by v021. In almost all surveys, both v001 and v021 are included and are equal.

The following will tell you how to "unpack" the columns of caseid (or hhid).

* Open an IR file and enter this:

describe caseid

* this will tell you the string length, for example 12. Then:

```
forvalues li=1/12 {
gen col_`li'=substr(caseid,`li',1)
}
```

list col* v001 v002 v003 if _n<=50, table clean

Good luck!

Subject: Re: Merging and appending data files
Posted by kiran on Mon, 15 Jan 2024 18:09:58 GMT
View Forum Message <> Reply to Message

My question is that after merging some files from one particular year, how we are suppose to append files from different wave into one file, like I am looking at the impact of School Stipend Program on different outcomes and I want to append DHS 2018 women file with DHS 2013 women file and DHS 2006 women file altogether to create one pooled cross sectional data. Please if someone knows guide me on that.

Subject: Re: Merging and appending data files
Posted by Bridgette-DHS on Tue, 16 Jan 2024 17:12:18 GMT
View Forum Message <> Reply to Message

Following is a response from Senior DHS staff member, Tom Pullum:

The following lines show how to append the IR files from these three surveys. Note that v024 (region), which is an important variable, is coded differently in the three surveys. It is just one example of a variable that is defined differently in the three surveys. Such variables must be saved with survey-specific variable labels and then be recoded to a single variable that applies across the surveys.

```
* Specify a workspace
cd e:\DHS\DHS_data\scratch

use "...\PKIR52FL.DTA", clear
keep v*
rename v024 v024_1
gen survey=1
save PKIR.dta, replace

use "...\PKIR61FL.DTA", clear
keep v*
rename v024 v024_2
gen survey=2
quietly append using PKIR.dta
save PKIR.dta, replace

use "...\PKIR71FL.DTA", clear
keep v*
rename v024 v024_3
gen survey=3
quietly append using PKIR.dta
save PKIR.dta, replace

tab1 v024_*

* You must construct a new variable for v024 (region) because the codes were different
*   in the three surveys.  Many other variables are also different in the three surveys
```