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!