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!!