Subject: merge IR and KR: duplicate line number in IR
Posted by bun_2019fall on Thu, 26 Aug 2021 17:52:27 GMT
View Forum Message <> Reply to Message

Hi DHS user forum,

I was trying to merge IR and KR (one observation in IR correspond to many obs in KR), based on v000, v001, and v002, v003, as suggested by the DHS website. However, I failed to do so because there are duplicate line numbers (v003) in IR files. Thus, my Stata code "1:m" does not work.

I tried to search the web and I did not find any information mentioning if there exists duplicate line number in the same IR files, and why? I wonder if I missed anything here? Any suggestion would be greatly appreciated!

Subject: Re: merge IR and KR: duplicate line number in IR
Posted by schoumaker on Thu, 26 Aug 2021 18:06:53 GMT
View Forum Message <> Reply to Message

Hello,
Are you sure you mentioned the KR data file as the using data file, if you use merge 1:m ?
Otherwise, try merge m:1.
Best,
Bruno

Subject: Re: merge IR and KR: duplicate line number in IR
Posted by bun_2019fall on Thu, 26 Aug 2021 18:49:29 GMT
View Forum Message <> Reply to Message

Hi Bruno,

Thanks so much for responding. Yes, I am sure that the using file is KR, while the master is IR.

Subject: Re: merge IR and KR: duplicate line number in IR
Posted by schoumaker on Fri, 27 Aug 2021 06:59:10 GMT
View Forum Message <> Reply to Message

Cand you send information on the files you trying to merge (their names) and the Stata command you use ?
best,
Bruno

Subject: Re: merge IR and KR: duplicate line number in IR
Posted by bun_2019fall on Fri, 27 Aug 2021 15:51:56 GMT
View Forum Message <> Reply to Message

Hi Bruno,

For example, I used "BOIR01FL", and below the dup variable (dup by v000 - country v001-cluster v002-household v003-line number) that I created have values > 0.

```
use "V:\DHS raw survey/BOIR01DT/BOIR01FL", clear
duplicates tag v000 v001 v002 v003, generate(dup)
tab dup
```

And below are all the IR files that I found the duplicates (note, I cut the "IR" string from the filename, so the orginal filename should be, for exxample, BR21FL=BRIR21FL). These are all raw DHS files and I did not modified anything in the raw data.

data_name	Freq.	Percent	Cum.
BO01FL.DTA	6,669	7.13	7.13
BR21FL.DTA	54	0.06	7.19
CI51FL.DTA	7,442	7.96	15.14
CM22FL.DTA	3,329	3.56	18.70
CM31FL.DTA	4,259	4.55	23.25
CO01FL.DTA	4,721	5.05	28.30
DR21FL.DTA	6,354	6.79	35.09
HN52FL.DTA	196	0.21	35.30
HT31FL.DTA	4,438	4.74	40.05
IA42FL.DTA	34,892	37.30	77.35
LK02FL.DTA	2	0.00	77.35
ML32FL.DTA	589	0.63	77.98
ML41FL.DTA	9,300	9.94	87.92
NG21FL.DTA	4	0.00	87.92
NI22FL.DTA	4,608	4.93	92.85
NI31FL.DTA	47	0.05	92.90
PE01FL.DTA	197	0.21	93.11
PE6IFL.DTA	272	0.29	93.40
SN02FL.DTA	237	0.25	93.65
SN4AFL.DTA	4	0.00	93.66
TD31FL.DTA	5,923	6.33	99.99
TN02FL.DTA	8	0.01	100.00
UG01FL.DTA	2	0.00	100.00
Total	93,547	100.00	

I wonder if I missed anything here, as to whether I correctly identified the unique women in each IR files, thus to link back to women who have children in the KR files? Thank you very much again!

Subject: Re: merge IR and KR: duplicate line number in IR
Posted by schoumaker on Fri, 27 Aug 2021 20:26:19 GMT
View Forum Message <> Reply to Message

Hello,
I looked at the Bolivia file. It seems the v002 variable is indeed always equal to 1. Maybe a dwelling variable was used for creating the caseid and not available in the data set.
I think you can just use the caseid variable (the woman's id) that is available in the IR and the KR file to merge these two files. It seems to work in Bolivia, but I did not check in the other countries.
Best,
Bruno

***
use BOIR01FL.DTA, clear
merge 1:m caseid using BOKR01FL.DTA
***

Subject: Re: merge IR and KR: duplicate line number in IR
Posted by bun_2019fall on Sat, 28 Aug 2021 02:45:37 GMT
View Forum Message <> Reply to Message

Hi Bruno,

Thank you for much for looking into the file, and I really appreciated it!

The "caseid" solution is almost perfect, except for one file "DRIR21FL":

use "V:\DHS raw survey/DRIR21DT/DRIR21FL.DTA", clear
*duplicates by caseid
duplicates tag caseid, generate(dup)
sort caseid
tab dup /*4 duplicates*/
*complete duplicates
duplicates tag, gen(dup1)
tab dup1   /*no duplicates*/

The "caseid" itself has 4 duplicates in this dataset, out of all DHS IR files that I have downloaded as of July 2021. I further checked, these observations are not complete duplicates. That said, to proceed with data analysis, I wonder if I should just drop the 4 observations? I wonder if DHS has any best practice regarding this issue? Thank you, again!