Subject: Re: Merging BR and PR data
Posted by Janet-DHS on Fri, 15 Jul 2022 12:49:49 GMT
View Forum Message <> Reply to Message

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

You have made a very interesting discovery. I ran the Stata lines that I sent you, but didn't even look for the warning "variables state cluster hh line do not uniquely identify observations in the using data" because I was so sure that the code was correct. I should have been more careful. That warning means that there are repeats or duplicates of some ID codes.

To find the duplicates, I ran these lines on IABR7Atemp.dta:

```
use e:\DHS\DHS_data\scratch\IABRtemp.dta, clear
gen n=1
collapse (sum) n, by(state cluster hh line)
tab n
```

Those lines give the number of cases with each ID code (for the merge). Here's what comes up:

```
. tab n

   (sum) n |      Freq.     Percent        Cum.
------------+-----------------------------------
         1 |  1,008,647       99.99       99.99
         2 |         58        0.01      100.00
------------+-----------------------------------
     Total |  1,008,705      100.00
```

Apparently there are 58 duplicates in a file of more than a million children. This is a DP error. Such duplicates can arise for different reasons but they should have been checked and resolved during data processing. I then enter the following line:

```
list state cluster hh line if n==2, table clean noobs nolabel
```

I won't reproduce the full list but here is the first duplicate:

```
  state   cluster   hh   line
    1       834   26      9
```

Using this example, let's look at the all the children in this household--that is, the cases in the BR file who have a code for b16. Here they are:

```
list v024 v001 v002 v003 b16 b3 b4 b8 if v024==1 & v001==834 & v002==26, table clean noobs

     v024   v001   v002   v003   b16     b3      b4   b8
```

jammu &	834	26	2	9	1301	male	11
jammu &	834	26	2	7	1191	male	20
jammu &	834	26	2	8	1174	female	21
jammu &	834	26	2	3	1132	male	25
jammu &	834	26	4	6	1428	female	0
jammu &	834	26	4	5	1363	female	6
jammu &	834	26	8	11	1336	female	8
jammu &	834	26	8	10	1311	male	10
jammu &	834	26	8	9	1287	male	12

There are 3 children age 18+ and 6 children under age 18. ("Child" just applies to someone in a birth history.) The problem is that in this list there are two children with b16=9: a boy age 11 whose mother has line number 2 and a boy age 12 (b8 is age) whose mother has line number 8. (There's something else wrong here, because the mother with v003 has a daughter age 21 who has b16=8. A 21-year old daughter of the woman on line 2 is not going to have a 12-year old son.) Next, let's look at the household listing in the PR file and find these same individuals.

```
use "C:\Users\26216\ICF\Analysis - Shared Resources\Data\DHSdata\IAPR7AFL.DTA", clear
. label list HV101
HV101:
           1 head
           2 wife or husband
           3 son/daughter
           4 son/daughter-in-law
           5 grandchild
           6 parent
           7 parent-in-law
           8 brother/sister
           9 co-spouse
          10 other relative
          11 adopted/foster child
          12 not related
          13 niece/nephew by blood
          14 niece/nephew by marriage
          15 brother-in-law or sister-in-law
          16 niece/nephew
          17 domestic servant
          98 don't know

. list hv024 hv001 hv002 hvidx hv101 hv104 hv105 hv112 if hv024==1 & hv001==834 &
hv002==26, table clean noobs nolabel
```

hv024	hv001	hv002	hvidx	hv101	hv104	hv105	hv112
1	834	26	1	1	1	50	.
1	834	26	2	2	2	45	.
1	834	26	3	3	1	25	.
1	834	26	4	4	2	20	.
1	834	26	5	5	2	6	4

1	834	26	6	5	2	0	4
1	834	26	7	3	1	24	.
1	834	26	8	3	2	30	.
1	834	26	9	5	1	12	8
1	834	26	10	5	1	10	8
1	834	26	11	5	2	8	8
1	834	26	12	10	2	14	0

Here are the relation to head codes for hv101:

```
. label list HV101
HV101:
           1 head
           2 wife or husband
           3 son/daughter
           4 son/daughter-in-law
           5 grandchild
           6 parent
           7 parent-in-law
           8 brother/sister
           9 co-spouse
          10 other relative
          11 adopted/foster child
          12 not related
          13 niece/nephew by blood
          14 niece/nephew by marriage
          15 brother-in-law or sister-in-law
          16 niece/nephew
          17 domestic servant
          98 don't know
```

hv112 is the line number of the mother if the child is under age 18. hv104 is sex, hv105 is age. It's pretty clear that this household has three mothers. First, the woman on line 2, age 45, has three grown children in the household: a son age 25 on line 3, whose wife age 20 is on line 4, a son on age 24 on line 7, and a daughter age 30 on line 8. I see that the ages for lines 7 and 8 in the PR file do not agree with the ages for lines 7 and 8 in the BR file, but this does not imply a DP error. Priority would be given to the ages in the BR file because they come from the individual interview with the mother but the values given in the household interview are retained, not over-written. You should give priority to b8 for age, rather than hv105, following this merge.

The second mother in the household is the woman on line 4; her children are on lines 5 (a girl age 6) and 6 (a girl age 0). The third mother is the woman on line 8, whose children are on lines 9 (a boy age 12), 10 (a boy age 10), and 11 (a girl age 8). The last person in the household is a 14 year old girl who is an "other relative" of the household head but whose mother is not in the household.

The children of the woman on line 8 have ages 9, 10, and 11 in the BR file, but ages 8, 10, and 12 in the PR file. The children of the woman on line 4 have ages 0 and 6 in both the BR and PR files.

The PR file clearly shows that the child on line 9 is the 12-year old son of the woman on line 8. This is the child at the bottom of the list from the BR file. The child at the TOP of that list, a boy age 11 who also is stated to have line 9, and whose mother is stated to be the woman on line 2, is incorrect. However, I don't have time to push further with this detective work. I just wanted to demonstrate that the merge command SHOULD work, and the only reason why Stata gives that warning is that the DP steps to reconcile such inconsistencies, which originate during data collection, did not extend quite far enough. 58 unresolved inconsistencies with such a large file is not serious. I recommend that you do the merge exactly as I first suggested. By saving cases with _merge==3, you will retain 58 children who probably genuinely were in the sample, and that's better than deleting them. Moreover, if you were to drop one child in each duplicate pair, which one would you drop? I will inform the DP staff of this issue.