Understanding Society User Support - Support #1128

How to match husbands and wives in USoc without dropping one or the other

01/14/2019 02:05 PM - Nico Ochmann

Status:	Feedback	**Start date:**	01/14/2019
Priority:	High	**% Done:**	80%
Assignee:	Nico Ochmann		

Description

Dear Stephanie, it is me again. I need your help with the following. I try to match husbands and wives (spouses) in USoc. This is what I am doing which is based on a previous suggestion from your team quite a while ago.

*manipulate data set to find age, UK arrival year etc. of spouse (sppno)
sort hidp pno
gen partnum=cond(pno < sppno, pno, sppno) if sppno>0

drop if sppno == 0 | sppno<0

bysort hidp partnum: egen numinpart = sum(sppno > 0)
tab numinpart

keep if numinpart 2

bysort hidp partnum: ge sp_age = cond(_n2,age(1),age(2),.) /// where age brackets 1 and age brackets 2, i.e.[] if I place a number within brackets, I get a goofy preview.
if partnum<.

bysort hidp partnum: ge sp_yr2uk4 = cond(_n==2,yr2uk4(1),yr2uk4(2),.) /// where yr2uk4 brackets 1 and yr2uk4 brackets 2, i.e.[]
if partnum<.

bysort hidp partnum: drop if (sex==1 & _n==2) | (sex==1 & _n==1) // drop females (2) or males (1), here I drop males so all variables defined are for wives and all sp_variables are for husbands.

Unfortunately, for my research question, I need to have husbands and wives matched, have variable characteristics for say wives and sp_characteristics for husbands WITHOUT having the dropping procedure of the previous line (i.e., bysort hidp partnum: drop if (sex==1 & _n==2) | (sex==1 & _n==1)). I hope I am making sense, I need to match wives and husbands and generate characteristics of both without dropping wives or husbands. The data should look like this:
hidp education_wife education_husband age_wife age_husband etc.
1 postgrad bachelor 50 60 etc.

I hope this is clear, if not please feel free to ask me.

Once again, I would very much appreciate your help and support.

Best wishes from Manchester.

Nico

History

#1 - 01/14/2019 03:54 PM - Stephanie Auty

- Status changed from New to In Progress

- % Done changed from 0 to 10

- Private changed from Yes to No

Many thanks for your enquiry. The Understanding Society team is looking into it and we will get back to you as soon as we can.

Best wishes,
Stephanie Auty - Understanding Society User Support Officer

#2 - 01/31/2019 03:44 PM - Stephanie Auty

- Status changed from In Progress to Feedback

- Assignee changed from Stephanie Auty to Nico Ochmann

- % Done changed from 10 to 70

Dear Nico,

Before that line of code (i.e., bysort hidp partnum: drop if (sex==1 & *n==2) | (sex==1 & _n==1)), you have two rows in the dataset for each couple, with one member of the couple defined in the sp* variables in one row, and the other member of the couple in the other row. You are dropping one of the rows so that you will be left with one row per couple. If all of the couples consisted of one man and one woman then I think that would be what you need. However, this code does not account for same sex couples. You will have no data for couples consisting of two men, and still have two rows for couples consisting of two women.

We have updated the worksheet for Example 7 in our course which deals with merging in this way, so you may find it helpful to look at that: https://moodlex.essex.ac.uk/course/view.php?id=76

Best wishes,
Stephanie

#3 - 02/03/2019 04:51 PM - Nico Ochmann

Dear Stephanie,

thanks for your help once again. I will have a look.

Best wishes.

Nico

#4 - 02/06/2019 04:15 PM - Nico Ochmann

Dear Stephanie,

I do have a follow-up question. An easy one I must admit, but I could not find anything in the user guide or elsewhere.

What is the difference between the _ppno and the _sppno. The latter refers to the spouse I see that, and the former refers to partner.

What does partner mean? Does it mean the spouse and the partners in unmarried couples? In sum, does _ppno refer to married and unmarried couples and _sppno to married couples only?

Cheerio and thank you very much.

Nico

#5 - 02/20/2019 03:12 PM - Stephanie Auty

- % Done changed from 70 to 80

Dear Nico,

That's right, partner includes spouse or cohabiting partner.

Best wishes,
Stephanie

#6 - 02/26/2019 01:45 PM - Nico Ochmann

Dear Stephanie,

I got one final, final question with regard to this open issue. I am confused about this following code:
bysort hidp partnum: drop if (sex==1 & _n==2) | (sex==1 & _n==1)
I did check the manual, but I am not sure what I am dropping here with this statement/command line: Sex==1 are men, but what do the _n==2 or _n==1 refer to?

Thank you very much.

Have a nice day.

Nico

#7 - 02/27/2019 10:55 AM - Nico Ochmann

Hi Stephanie,

something else came up in this context, my apologies to post another question in this regard. I am struggling with the following. Let's start out with this:

bysort hidp partnum: drop if (sex==1 & _n==2) | (sex==1 & _n==1) | sex==sp_sex // drop males
bysort female: sum pidp // number of females I get is 74,169
When I do this:
bysort hidp partnum: drop if (sex==2 & _n==2) | (sex==2 & _n==1) | sex==sp_sex // drop females
bysort female: sum pidp // number of males I get is 74,170
These two numbers are very close, and I conclude from this that the number of couples I have in my sample is about 74,169.
Here comes my problem and big question. Let's say I want to futher divide the sample into the subsample immigrant==1 or immigrant==0 (native).
Again I repeat this:
bysort hidp partnum: drop if (sex==1 & _n==2) | (sex==1 & _n==1) | sex==sp_sex // drop males
tab immigrant female // Now I get 42,264 natives and 13,688 immigrants for a total of 55,952
bysort hidp partnum: drop if (sex==2 & _n==2) | (sex==2 & _n==1) | sex==sp_sex // drop females
tab immigrant female // Here I get 41,863 natives and 12,958 immigrants for a total of 54,821.
Since immigrant status is missing more often than gender status, I see that I must lose observations, but what I do not understand is why there is such a huge difference between 55,952 and 54,821? My objective is to find the number of couples that are immigrants only, natives only, or mixed couples (immigrant wife/native husband or native wife/immigrant husband).
I really appreciate your help Stephanie.

Best wishes.

Nico

#8 - 03/01/2019 02:45 PM - Stephanie Auty

Dear Nico,

In reply to your first question, the _n refers to the number within the bysort group. You are using bysort hidp partnum:, so within each group of unique hidp and partnum, _n==1 for the first record, 2 for the second etc. _N refers to the last in the group. In this case it doesn't seem necessary, and just "bysort hidp partnum: drop if sex==1" would have the same effect, as there are two members of each couple.

I think the discrepancy in your most recent question is to do with the presence of same sex couples in the dataset. When you drop records where sex==1 you are dropping couples consisting of two men from the dataset, and if you drop where sex==2 you drop couples consisting of two women. Please do go back to the moodle course as I suggested above and look at the updated version of example 7, as this has some suggestions about using the data taking this into account.

Best wishes,
Stephanie

#9 - 03/01/2019 03:43 PM - Nico Ochmann

Dear Stephanie,

thank you very much for your kind reply and your help once again. Due to the discrepancy, my question was poorly stated/phrased. If you do not have an answer as to how to adjust for the discrepancies, no problem, I do not expect you to know it all. The following gives you my sample summary statistics for the variables I will be using. Let me give you an example of my concern. Lets take female== 0 and look at the sp_country variable with 52,020 observations. This number of observations should be close to equal to the number of observations under female==1 and country: 55,842. This is quite a discrepancy if I am correct in my thinking here. I admit for other variables the difference is not quite as pronounced. However, if you happen to know any way to adjust for this even if it entails dropping couples I would be very grateful.

Have a great day.

Cheers. Nico

-> female = 0

Variable	Obs	Mean	Std. Dev.	Min	Max
employed	49,173	.9411262	.2353907	0	1
hh_u7	74,895	.2067294	.4049624	0	1
kids	74,895	.6661459	1.034381	0	10
education	61,303	4.693359	2.284051	1	7
sp_education	60,769	4.720828	2.160083	1	7
yuk	55,260	6.11701	13.96495	0	87
spyuk	56,406	5.575666	12.73416	0	79
age	74,884	54.79936	14.97068	16	98
sp_age	74,890	52.16973	14.83276	17	99
first	67,268	.8544925	.3526144	0	1
sp_first	70,606	.8423788	.364388	0	1
region	74,874	6.596202	3.143511	1	12
year	74,895	2012.641	2.33733	2009	2018
cohort	55,260	.8848534	1.810811	0	8
sp_cohort	56,406	.8441123	1.696669	0	8

```
     ethn_dv |     74,212    3.090417    7.263106          1          97
  sp_ethn_dv |     74,631    3.215621    7.731485          1          97
     country |     54,751    788.4318    392.9279          1        1000
  sp_country |     52,020     95.0451    48.77885          5         997
   parentsuk |     60,985    .8113471    .3912359          0           1
```

-> female = 1

```
Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    employed |     50,696    .7460944    .4352485          0           1
       hh_u7 |     74,897    .2067373    .4049681          0           1
        kids |     74,897    .6666756    1.035046          0          10
   education |     60,846    4.719604    2.160695          1           7
sp_education |     61,237    4.693584    2.284055          1           7
-------------+---------------------------------------------------------
         yuk |     56,417    5.574047    12.73179          0          79
       spyuk |     55,266    6.116075    13.96254          0          87
         age |     74,892     52.1685    14.83495         17          99
      sp_age |     74,885    54.79927    14.97134         16          98
       first |     70,606    .8423788     .364388          0           1
-------------+---------------------------------------------------------
    sp_first |     67,274    .8545055    .3526014          0           1
      region |     74,876    6.597708    3.143389          1          12
        year |     74,897    2012.639     2.33632       2009        2018
      cohort |     56,417    .8439655    1.696301          0           8
   sp_cohort |     55,266    .8848478    1.810637          0           8
-------------+---------------------------------------------------------
     ethn_dv |     74,631    3.215835    7.731461          1          97
  sp_ethn_dv |     74,217    3.090518    7.262964          1          97
     country |     55,842    784.1448    393.2899          1        1000
  sp_country |     51,612    95.59897    51.16966          1         997
   parentsuk |     60,991       .8113    .3912733          0           1
```

#10 - 03/02/2019 03:51 PM - Nico Ochmann

Stephanie,

I should note that in coming up with the above summary stats, I did combine all eight waves in USoc.

Best wishes.

Nico

#11 - 03/04/2019 09:56 AM - Nico Ochmann

Dear Stephanie,

I found out a way to drop households that have a missing on any explanatory variable for both female or male.

Hence, the issue is resolved.

Thanks again for your help.

Have a lovely week.

Nico