

Understanding Society User Support - Support #220

Problem with mergin files

11/06/2013 05:16 PM - Shilan Dargahi

Status:	Closed	Start date:	11/06/2013
Priority:	Urgent	% Done:	100%
Assignee:			
Category:			
Description			
I am using Current and Annual Net Household Income Variables of BHPS (waves 1–18) derived by Horacio Levy and Stephen P. Jenkins and I encountered the following problem: While browsing through the file and trying to merge the files to have a panel data set I found that two individuals have the same PID (I suppose they are the member of the same family). I attached a small Excel file to show you what I mean by this. For example some one with PID=10017992 in wave A is in a household of type: couple,1 ch,m lt 65 but in wave H this PID belongs to some one who is in a household of type: single woman,lt 20. I found this a bit confusing as it seems the person in wave H who is a man aged 20 has the same PID as the person who is woman and 65 years old at wave A.			

History

#1 - 11/06/2013 05:32 PM - Redmine Admin

- Target version set to BHPS

We don't allow any attachments. You can copy in tabulations or examples in the Notes field. It would be useful if you could give us an idea of the scale of the issue too. Thanks, Jakob

#2 - 11/06/2013 06:10 PM - Shilan Dargahi

I have set up a panel data using 18 waves of r_ecstat files. These files are part of the Derived Current and Annual Net Household Income data set, an unofficial supplement to the set of derived income variables in the official BHPS release.

I am not sure what exactly the scale of the issue is for what percentage of the constructed panel the problem exists but by quickly browsing through the data I spotted many cases like the following:

```
tab butype if pid==10017992
```

benefit unit type	Freq.	Percent	Cum.
couple,1 ch,m lt 65	6	35.29	35.29
f lone parent,lt 60	2	11.76	47.06
single woman,lt 20	2	11.76	58.82
single woman,20-39	7	41.18	100.00

Or another example:
tab butype if pid==10016848

benefit unit type	Freq.	Percent	Cum.
couple,no ch,m lt 65	8	53.33	53.33
couple,1 ch,m lt 65	5	33.33	86.67
single woman,20-39	2	13.33	100.00

I was not expecting to observe the same PID for these individuals but maybe there is a simple explanation here or I might be misunderstanding some thing.

Any help is greatly appreciated!

#3 - 11/06/2013 06:10 PM - Shilan Dargahi

Shilan Dargahi wrote:

I have set up a panel data using 18 waves of r_ecstat files. These files are part of the Derived Current and Annual Net Household Income data set, an unofficial supplement to the set of derived income variables in the official BHPS release.

I am not sure what exactly the scale of the issue is (for what percentage of the constructed panel the problem exists) but by quickly browsing through the data I spotted many cases like the following:

tab butype if pid==10017992

benefit unit type	Freq.	Percent	Cum.
-----+			
couple,1 ch,m lt 65	6	35.29	35.29
f lone parent,lt 60	2	11.76	47.06
single woman,lt 20	2	11.76	58.82
single woman,20-39	7	41.18	100.00
-----+			

Or another example:
tab butype if pid==10016848

benefit unit type	Freq.	Percent	Cum.
-----+			
couple,no ch,m lt 65	8	53.33	53.33
couple,1 ch,m lt 65	5	33.33	86.67
single woman,20-39	2	13.33	100.00
-----+			

I was not expecting to observe the same PID for these individuals but maybe there is a simple explanation here or I might be misunderstanding some thing.

Any help is greatly appreciated!

#4 - 11/19/2013 01:45 PM - Redmine Admin

- Status changed from New to In Progress
- % Done changed from 0 to 50

The benefit unit type relates to the composition of the household unit, which will vary naturally from wave to wave. The cross-wave file, XWLSTEN, can give you information such as last known sex of a respondent. This info may be useful in a small number of cases, where the data can be ambiguous.
Jakob

#5 - 12/03/2013 05:31 PM - Redmine Admin

- Status changed from In Progress to Closed
- % Done changed from 50 to 100