Understanding Society User Support - Support #912

Duplicates of pid in the harmonized UKHLS/BHPS xwavedat

02/05/2018 08:13 PM - Michael Baumkautner

Status:	Resolved	**Start date:**	02/05/2018
Priority:	Immediate	**% Done:**	100%
Assignee:	Michael Baumkautner		
Category:	Data inconsistency		

Description

Dear User Support

I am trying to understand the identifier variables in the harmonized UKHLS/BHPS xwavedat.

1) Why are there records in xwavedat that have the same pid value?

2) And why do they have different pidp values?

See here:

```
use pid pidp using data\xwavedat, clear
recode pid (-8 = .)
keep if !missing(pid)
duplicates tag pid, gen(dup)
```

Thanks for your consideration

History

#1 - 02/12/2018 11:57 AM - Alita Nandi

- Category set to Data inconsistency

- Status changed from New to Feedback

- Assignee set to Michael Baumkautner

- Target version set to X M

- % Done changed from 0 to 50

- Private changed from Yes to No

Hello Michael,

Thank you for identifying these cases. You are correct there are 11 pids with 2 observations. We have identified the source of the problem and it will be resolved in the next release. These cases are all cases from the BHPS samples. Here are a couple of suggestions:
(1) Delete these 11 cases
(2) A quick fix that we can suggest is to use the best information from these duplicate cases (N=11) and then keep just one observation. Here is the code:

```
use "\\usocdist0\restricted$\working\crosswave\crosswave_wave07_1\datasets\xwavedat", clear
recode pid (-8 = .)
keep if !missing(pid)
bys pid: g dup=_N
keep if dup==2

mvdecode all, mv(-21/-1)
ds pidp pid, not
global vlist `r(varlist)'
foreach var of global vlist {
bys pidp: egen y`var'=mean(`var')
g prob_`var'=1 if y_`var'~=`var' & `var'<. & y_`var'<.
}
// Check that the values of the same variables are the same across the two rows.
su prob_*
// You will find a few variables where the values from the two rows are different. This is because the BHPS and UKHLS versions were not harmonised.
// (*scend_dv *generation *j1soc90 *1soc90_cc  *coh1m_dv coh1y_dv *lmar1y_dv). In these cases
```

```
// For the variables where either one of the rows is missing or both are the same do this:
foreach var of global vlist {
replace `var'=y_`var' if prob_`var'==.
}
// For the problem cases, decide on a rule and put that value for both rows.
// Next, keep one of the observations:
bys pid: keep if _n==1
```

Then append these cases to the original file, after first removing these 22 rows.

Please let me know if this does not work, or the response is not clear.

Best wishes,
Alita

#2 - 08/14/2018 04:07 PM - Stephanie Auty

- Status changed from Feedback to Resolved

- % Done changed from 50 to 100