Understanding Society User Support - Support #1185

Linking all waves of BHPS and UKHLS: Inconsistencies?

04/16/2019 11:38 PM - Nicole Schwitter

Status:	Resolved	**Start date:**	04/16/2019
Priority:	Normal	**% Done:**	100%
Assignee:			
Category:			

Description

Hello,

I've merged all the waves of the BHPS and Understanding Society into one master data file in the long-format (I have one row per person per wave). To check whether this has worked out correctly, I checked whether any respondents had changes in time-invariant variables like their sex. Doing that, I found quite a number of mismatches: Using the variable "sex" by "pidp", there was a change of sex in 15417 rows (and no change in 558476 rows). If I use "sex_dv", there is a change of sex in only 17 rows (no change in 279717 rows; sex_dv has a large number of missing values).
Is it possible that there are that many inconsistencies or is it more likely that I did anything wrong in the process of merging the datasets?

History

#1 - 04/17/2019 09:40 AM - Alita Nandi

- Status changed from New to Feedback

- Assignee changed from Alita Nandi to Nicole Schwitter

- % Done changed from 0 to 80

- Private changed from Yes to No

Hi Nicole,

Without looking at your code I cannot comment but I can say that when I appended 25 waves I too find similar inconsistencies. The reason you are getting these inconsistencies is because of proxy interviews. For these cases sex is coded as -7. We will look into why sex=-7 for proxy interviews as the missing value code of -7 is reserved for cases where the information is missing for proxy interviews - and that is not the case for sex.

Detailed answer:

After creating the long format file of 25 waves, I produced the mean of sex, sex_mean1 and counted mismatches of this mean with individual wave specific value of sex. I too found 15458 inconsistencies

bys pidp: egen sex_mean1=mean(sex)
cou if sex~=sex_mean1

Then I recoded proxies to missing and repeated the exercise and found 2545 mismatches.
recode sex -7=.
bys pidp: egen sex_mean2=mean(sex)
cou if sex~=sex_mean2

And when I restricted this to only those cases where sex is not missing the number of mismatches goes down to 492.

cou if sex~=sex_mean2 & sex<.

Best wishes,
Understanding Society User Support Team

#2 - 04/17/2019 10:02 AM - Alita Nandi

Also, sex_dv is only available for the 8 UKHLS waves and it does not have the same problem, that is, there is a valid value even for proxy respondents.

#3 - 04/17/2019 10:19 AM - Nicole Schwitter

Thank you Alita!
This is very helpful as I was mostly wondering whether I did something wrong in the matching process or whether it is something inherent in the data, but this explains the problem.

#4 - 03/02/2021 03:56 PM - Understanding Society User Support Team

- Status changed from Feedback to Resolved

- Assignee deleted (Nicole Schwitter)

- % Done changed from 80 to 100