

Understanding Society User Support - Support #1810

Inconsistent income variable values between data releases

11/21/2022 09:51 AM - João Duro

Status:	Resolved	Start date:	11/21/2022
Priority:	Normal	% Done:	100%
Assignee:	Understanding Society User Support Team		
Category:	Data inconsistency		
Description			
I am referring to the Understanding Society (US) dataset in https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=6614			
I have started working with the US dataset when it contained waves 1-10, and I worked specifically with wave 9 (i.e. year=2018 or with files starting with i_*). Meanwhile a new wave has been released, and the most recent version also contains wave 11 (year=2020). My query relates to changes in the dataset for the "same" wave (i.e. wave 9) between the old release (which contains waves 1-10) and the new release (which contains waves 1-11). Just to clarify, I am not talking about changes over time (i.e. across waves), these are changes that I have noticed concerning only wave 9 (year=2018) between the old and new release of the US dataset. The following data is taken from wave 9 file i_indresp.			
Old dataset (waves 1-10):			
<pre>pidp i_dvage i_sex i_fimnnet_dv i_fimnlabnet_dv i_paynu_dv i_fimnpen_dv i_fimnsben_dv 1 68006127 47 2 1196.25 0 -8 0 1196.25 2 68006807 80 2 1872.07 0 -8 0 1872.07 3 68008847 59 2 934 0 -8 245 272 4 68009527 39 1 2102.17 1793 1793 0 80 5 68010887 53 2 1200 1200 1200 0 0 6 68011567 43 1 3163.81 3134.65 3134.65 0 0</pre>			
New dataset (waves 1-11):			
<pre>pidp i_dvage i_sex i_fimnnet_dv i_fimnlabnet_dv i_paynu_dv i_fimnpen_dv i_fimnsben_dv 1 68006127 47 2 1196.25 0 -8 0 1196.25 2 68006807 80 2 1854.66 0 -8 0 1854.66 3 68008847 59 2 934 0 -8 245 272 4 68009527 39 1 2102.17 1793 1793 0 80 5 68010887 53 2 1200 1200 1200 0 0 6 68011567 43 1 2713.83 2587.3 2587.3 0 0</pre>			
The above tables shows the pidp of 6 individuals together with some characteristics like age and sex, and 5 income net variables. Notice that between the two tables the second and last individuals (2 and 6) have different incomes values, and the remaining individuals (1, 3, 4, 5) have the same income values. I am aware that when new waves are added, the data of the old waves has to be recalculated. I understand that variables like i_fimnnet_dv and i_fimnlabnet_dv are derived from others, but this doesn't explain: 1) the changes in pension income reported for the individual 2 (was 1872.07 and now is 1854.66); and 2) the changes in net usual pay of the individual 6 (was 3134.65 and now is 2587.3). The last one is a very significant change in income. If I look further in the dataset I can find more individuals with different income values. The other non-income variables seem correct like age and sex and so on.			
I have also checked the file 6614_waves1_to_11_revisions_sep_2022.pdf that reports changes to the dataset, but does not mention any changes to these income variables. I have also looked into 6614_waves1_to_11_user_guide.pdf and there is a section on Top coding of income variables in page 44, where it is reported that variables like i_paynu_dv are top-coded at +- 8,333 per month in order to preserve the privacy of individuals with a very high income, but this does not affect the cases that I have reported above.			
Could please let me know if there is any other document that could explain these differences that I might have missed.			
Thanks, João			

History

#1 - 11/24/2022 08:22 AM - Understanding Society User Support Team

- Status changed from New to Feedback

- % Done changed from 0 to 80

Hello,

Sorry for the delay in getting back to you. If you check the imputation flags you will see that these changes are only observed for cases where the income (total or some of its components) have been imputed and in most cases the change is small. However, there are some cases, for which the imputed value changes a lot (as you have identified above). Note that the imputed flags are not 0-1 variables, rather they show the proportion of imputation involved. So, you could check if these large changes are found in cases where a large proportion of the income had to be imputed. For robustness check you could run their models excluding the cases with any imputed income or with a very proportion of income imputed.

Hope this helps.

Best wishes,
Understanding Society User Support Team

#2 - 11/24/2022 11:50 AM - João Duro

Thank you for your reply.

I can see that the individuals with inconsistency in income values have imputed flags on. For instance, the 6th individual has variable i_paynu_if set to 1, and all the other individuals with no inconsistency don't have the imputed flags on (i.e. they are set to zero).

This addresses my question.

Regards,
João

#3 - 11/25/2022 08:31 AM - Understanding Society User Support Team

- Status changed from Feedback to Resolved

- % Done changed from 80 to 100

#4 - 11/25/2022 08:32 AM - Understanding Society User Support Team

- Private changed from Yes to No