Understanding Society User Support - Support #1859

sample size loss due to weighting

02/20/2023 02:26 PM - Caroline Kienast von Einem

Status:	Resolved		**Start date:**	02/20/2023
Priority:	Normal		**% Done:**	100%
Assignee:	Olena Kaminska			
Category:	Weights			

Description

Hi,

I am aware that weighting will affect and alter the sample size of the analysis, however, I am working with a pooled sample of participants from wave 3-6 and when I specify a weighted model my sample drops from ~45k to 27k. This seems quite significant, particularly once I start to investigate subgroup characterstics..

Would you be able to confirm whether a drop by ~20k is normal once weighting is applied (ai am using the longitudinal wave f weight "f_indinub_lw" / whether the below STATA code makes you think it is instead an error with my coding?

STATA CODE:

```
//Open wave 6:
use f_hidp f_psu f_strata pidp f_sex_dv f_age_dv f_indinub_lw using "$inpath\f_indresp", clear

save "test", replace

foreach w in c d e {

// Extract the variables needed
    use "$inpath/`w'_indresp", clear
    isvar pidp `w'_addrmov_dv `w'_adcts `w'_distmov_dv `w'_mvyr `w'_mvever `w'_plnowy4
    keep `r(varlist)'

// save each wave specific file
    save `w'junk.dta, replace
}

// Open the file for wave f and then add the rest of the wave specific files
use "test", clear
foreach w in c d e {
merge 1:1 pidp using `w'junk.dta
drop _merge
}

save "test", replace

// get rid of unwanted temporary files
foreach w in c d e  {
erase `w'junk.dta
}

mvdecode _all, mv(-9/-1)

//I only want those with data at wave 6
drop if f_hidp==.

tabulate f_sex_dv // -> n= 45,186

svyset f_psu , strata(f_strata) singleunit(scaled)|| pidp, weight(f_indinub_lw)
svy: tabulate f_sex_dv, count col // -> n=27,094
```

History

#1 - 02/21/2023 02:42 PM - Understanding Society User Support Team

- Status changed from New to In Progress

- Assignee changed from Understanding Society User Support Team to Olena Kaminska

- 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. We aim to respond to simple queries within 48 hours and more complex issues within 7 working days.

Best wishes,
Understanding Society User Support Team

#2 - 02/22/2023 02:13 PM - Olena Kaminska

Caroline,

Thank you for your question. When you mention pooled analysis, do you mean using data cross-sectionally from waves 3, 4, 5, and 6 separately in one analysis? If so, you should use cross-sectional weights. A longitudinal analysis using waves 3-6 would use f_indinub_lw, as you suggested, but this would exclude wave 6 boost - so the difference may be partially due to this, and partially because TSMs can't be used in a longitudinal analysis (by design).

Hope this helps,
Olena

#3 - 02/23/2023 10:37 AM - Understanding Society User Support Team

- Status changed from In Progress to Feedback

- % Done changed from 0 to 50

#4 - 02/23/2023 12:31 PM - Caroline Kienast von Einem

Dear Olena,

with pooled analysis I mean that my sample consists of individuals that have responded in wave 3, 4, 5 AND 6.

In each wave, I am dividing participants into either movers or stayers depending on whether they have relocated home or not, and then I am comparing the differences between movers and stayer in wave 6.

Does this clarify why the sample size drops so significantly once weights are applied?

Best wishes

#5 - 02/23/2023 01:08 PM - Olena Kaminska

Caroline,

As you describe it, you are using a longitudinal analysis with information from wave 6 and one or more of the waves 3-5. In such situation your weight selection is correct, and the sample size drops because not all samples are relevant to this longitudinal analysis (IEMB only started at wave 6), and TSMs are not eligible for your analysis. There is no problem with the sample size - it is just that some of the interviews are not relevant to your specific use of the data.

Hope this helps,
Olena

#6 - 11/30/2023 12:51 PM - Understanding Society User Support Team

- Status changed from Feedback to Resolved

- % Done changed from 50 to 100