Understanding Society User Support - Support #2104

Dropped observations after specifying the complexity of the sample design

05/08/2024 05:09 PM - Yuliya Tavares

Status:	Feedback	**Start date:**	05/08/2024
Priority:	Normal	**% Done:**	80%
Assignee:	Understanding Society User Support Team		
Category:	Data management		

Description

Dear Team Member,

I am using 1-13 waves pooled together (undresp files), treating the merged dataset as cross-sectional. I have an issue with the following. When I run a linear regression with weights, the number of observations is 114,273. When I specify to STATA the complexity of the sample design:

svyset psu [pweight = indinui_xw], strata(strata) singleunit(centered)

And then run exactly the same regression (without weights), the number of observations drops significantly to 8,821. What would you recommend in this case?

I checked the forum but have not encountered similar issues posted. Thank you in advance.

History

#1 - 05/09/2024 05:13 PM - Understanding Society User Support Team

- Status changed from New to Feedback

- % Done changed from 0 to 80

The number of observations for estimations using weights and without are different as the number of cases with 0 weights are not included in the number of observations for weighted estimation. You can check this by excluding cases with zero weights from weighted and unweighted estimations - you will see that the number of observations remain the same.

But in your case, there is an additional issue - the weight variable you have specified, w_indinui_xw, is only available from onwards Wave 6. So, Stata drops all data from Waves 1-5 in weighted estimations using this weight. What you will need to do is create a new variable which will be the cross-sectional weights that are relevant for that wave, and use that weight variable in svyset. For example,
generat weight=indinus_xw if wave==1
replace weight=indinub_xw if wave>=2 & wave<=6
replace weight=indinui_xw if wave>=7 & wave<=13

I ran an example model of job satisfaction for those who are in paid employment using OLS with and without weights, and with svyset:
glo vlist jbsat i.sex_dv c.age_dv##c.age_dv i.ethn_dv i.mastat_dv i.hiqual_dv c.paygu_dv c.jbhrs
regress $vlist
regress $vlist [pw=weight]
svyset psu [pw=weight], strata(strata) singleunit(centered)
svy: $vlist

The number of observations were 237311, 232939, 253933. So, even though the observations did drop the extent was not as big as you found.

Please check and let us know if this resolves your issue.

Best wishes,
Understanding Society User Support Team

#2 - 05/16/2024 03:40 PM - Yuliya Tavares

Dear Team Member,

Thank you very much for a comprehensive reply! Generating new weights variable is an insight.

I have found the problem, when I merge waves using the following command (each file 'temp`w'' includes only relevant for my research variables):

use "$dir\tempa.dta", clear
foreach w in b c d e f g h i j k l m {
append using "$dir\temp`w'.dta"
}

All variables are successfully appended except for 'psu', 'strata' and 'ivfio'. Only wave 1 (2009-11) retains observations after appending, while waves 2-13 contain missing values. I read that these variables have identical values in each wave. How would you recommend solving this issue?

Thank you very much in advance.

#3 - 05/17/2024 04:57 PM - Understanding Society User Support Team

- Category set to Data management

- Private changed from Yes to No

Hi Yuliya,

It's hard to say what could have gone wrong (maybe you forgot to delete wave prefixes from the name of these variables?) In any case, you can always link psu and stata to your long file from the xwavedat file.

Best wishes,
Piotr Marzec
UKHLS User Support