Understanding Society User Support - Support #1257

Weight for unbalanced UKHLS panel data

10/09/2019 12:02 PM - Samir Sweida-Metwally

Status:	Resolved	**Start date:**	10/09/2019
Priority:	Normal	**% Done:**	100%
Assignee:			
Category:			

Description

Dear Olena,

I hope this message finds you well.

I am undertaking a longitudinal data analysis using wave 1 to wave 8 of UKHLS (I am using all waves - i.e., wave 1, 2, 3, 4, 5, 6, 7 & 8). The model is a two level logistic regression with observations (level 1) nested within individuals (level 2). My focus is on individuals (indresp file) who are aged between 16-64 and who completed a full interview (ivfio ==1). I am using Stata 15.1 to run my multilevel model. My dataset is unbalanced.

As I understand it I can only use the weights UKHLS provide to undertake analysis on a balanced panel dataset.

However, I don't want to run my analysis on a balanced panel dataset - **but instead I want to run my model on an unbalanced dataset.** This is because by using a balanced panel dataset I end up losing nearly 2/3rd of my observations and given the sub-groups I am interested in I am ending up with too few observations which is generating imprecise coefficients and too large standard errors/confidence intervals.

I have gone through the online discussion forum, notably the exchange you had with Ewan 4 years ago (https://iserswww.essex.ac.uk/support/issues/414), and I wanted to make sure I understood your advice correctly.

My questions are as follows:

1) Have I understood correctly that a possible remedy to my situation (**i.e. that I want to run a weighted model on an unbalanced dataset rather than on a balanced dataset**) is that I can use the cross-section weight of the earliest wave in my dataset (in this case wave 1) and apply that weight to all pidps across all waves?

2) If so, am I right that this weight would be **'a_indinus_xw' given that I am looking at wave 1 to wave 8?** That means that if I was looking at wave 2 to wave 8, the weight to be applied would be 'b_indinub_xw'. Is that correct?

3) If my understanding of point 2) (above) is correct, **am I right that 'a_indinus_xw' is the only weight I have to use in my model**? That is, that I don't have to use the cross-section weight for each and every single wave (i.e., waves 2, 3, 4, 5, 6, 7, and 8).

4) In the exchange you had with Ewan (see link above) you spoke of 'scaling the data'. **Am I correct in understanding that scaling does not apply to me** as I am not combining UKHLS with BHPS but only using UKHLS? That means that I would use the cross-section wave (e.g., 'a_indinus_xw' assuming my dataset is looking at wave 1 to wave 8) as is in my model without making any modifications to it, is that right?

5) Am I right in understanding that even if I did use the cross-section weight (e.g., 'a_indinus_xw' assuming my dataset is looking at wave 1 to wave 8) to run a weighted model on an unbalanced dataset, I would be increasing the number of observations in my model relative to running a weighted model on a balanced dataset using the UKHLS provided weight (i.e., 'h_indinus_lw' in this case), but I would still have observations being omitted from my model because any person who responded for the first time after wave 1 (i.e., in wave 2, 3, 4, 5, 6, 7 or 8) would have a weight of '0' as they would not have appeared in 'a_indinus_xw' ?

6) If my undersatidng of point 5) (above) is correct, does that mean that the only way for me to run a weighted model taking into account all the observations in my unbalanced dataset would be to create my own weights?

7) Finally, am I correct that the weight I should use **if I wanted to run a weighted version of my model on a balanced dataset (wave 1 to 8) I would use the weight called 'h_indinus_lw' provided by UKHLS?** Am I correct in understanding that by using the weight provided by UKHLS (here 'h_indinus_lw') I am only able to run a weighted version of my model on a balanced dataset because those pidps that do not have an observation in each and every wave (i.e., wave 1, 2, 3, 4, 5, 6, 7 & 8 in my case) would be given a weight of zero?

Apologies for the long message but I wanted to make sure I provided you with all the information you might need to be able to answer my questions. That said, if I have missed something and you do require further information please do not hesitate to let me know.

Thank you very much for your help Olena.

I look forward to hearing from you.

With very best wishes,

Samir

History

#1 - 10/09/2019 02:34 PM - Olena Kaminska

Samir,

Thank you for your question. I think you are using pooled analysis putting together waves 1 to 8.
For this using just wave 1 weight would not be the best solution. Instead you would need to extract a cross-sectional weight from each of the waves and match it to the relevant wave. So, for example observations from wave 1 would have wave 1 weight, and observations from wave 2 would have wave 2 weight and so on.

It would be good to scale the weights as per discussion with Evan. But this is less important if you are using only UKHLS. One way to check is to note the total number of respondents in your models in each wave. If this differs substantially you would need this extra scaling. This is important only if you want to represent years evenly. If you are happy for the earlier years to contribute to your analysis more than later years you can skip this scaling factor.

Your points 5) and 6) are correct but are not relevant. You will have most observations if you use the method and weights above.

Your point 7) is correct. Yet, your sample size would go up if you start with wave 2 and you use weight ub instead of us; or from wave 6 you can use ui weight.

Thank you,
Olena

#2 - 10/09/2019 03:42 PM - Stephanie Auty

- Status changed from New to Feedback

- Assignee changed from Olena Kaminska to Samir Sweida-Metwally

- % Done changed from 0 to 70

- Private changed from Yes to No

#3 - 10/11/2019 12:39 PM - Samir Sweida-Metwally

Hello Olena,

Thanks a lot, this is v helpful.

As discussed, I applied the changes but **I wanted to check I used the correct weights**. Can you confirm the below is correct?

apply the 'a_indinus_xw' weight for all observations in wave 1
apply the 'b_indinub_xw' weight for all observations in wave 2
apply the 'c_indinub_xw' weight for all observations in wave 3
apply the 'd_indinub_xw' weight for all observations in wave 4
apply the 'e_indinub_xw' weight for all observations in wave 5
apply the 'f_indinui_xw' weight for all observations in wave 6
apply the 'g_indinui_xw' weight for all observations in wave 7
apply the 'h_indinui_xw' weight for all observations in wave 8

If the above is correct, I have a follow on question. Once I set svyset (see below code), **I get an error message in Stata (see below) when I run my multilevel model seemingly because the weights within pidps are not consistent.** I have run the model with the cross-section weights (as described above) without taking 'strata' and 'psu' into account so there is no issue there, but I would be keen to take the complex survey design into account to get more precise standard errors. Are you by any chance aware of how that could be done?

That said, **even if I did account for 'strata' and 'psu' I wouldn't expect the change in SE (which would be an increase, correct?) to be that large, and it would have no effect on the covariate coefficients (am I right?), meaning that the weighted model alone (i.e., without taking the complex survey design into account) is satisfactory. Is my reasoning correct?**

Code:
svyset, clear
svyset psu, strata(strata) weight(final_weight)

Stata Error Message:
'weights in variable final_weight not constant within groups defined by: pidp an error occurred when svy executed melogit'

As always, thanks for your ongoing support Olena.

I look forward to hearing from you.

With very best wishes,

Samir

#4 - 10/11/2019 05:37 PM - Samir Sweida-Metwally

Regarding my last question, I am essentially trying to understand the consequences on my model/results of not taking the complex survey design (i.e., strata and psu) into account.

Thank you.

#5 - 10/21/2019 10:48 AM - Olena Kaminska

Samir,

Thank you for your question. I am not very familiar with melogit command in Stata, but looking at the Stata help, I would suggest you try the following svyset commands:

1) make sure your data is in long format.

2) try the Stata suggested svyset command:

svyset psu [pw=final_weight], strata(strata)
By the way you should be able to get psu and strata variables from xwavedat that's ready for you, i.e. you wouldn't need to extract them from all the waves.

3) If the above doesn't work, try the following:

svyset psu || pidp, weight(final_weight)
svy: melogit y x || psu: || ssu:

My understanding is that psu should be as a psu in our dataset, and ssu is pidp.

This way you will not take into account stratification making your estimates conservative.

With regard to your last question not taking into account stratification is fine - your results are just more conservative. But you should take into account psu and weights - both, point estimates and relationships are likely to be influenced.

Hope this helps,
Olena

#6 - 10/22/2019 09:35 AM - Samir Sweida-Metwally

Hello Olena,

Thanks a lot for this.

Unfortunately both options suggested didn't work. More specifically, using "svyset psu [pw=final_weight], strata(strata)" generates an error message that reads:

"survey final weights not allowed with multilevel models; a final weight variable was svyset using the [pw=exp] syntax, but multilevel models require that each stage-level weight variable is svyset using the stage's corresponding weight() option an error occurred when svy executed melogit."

Meanwhile, running:

"svyset psu || pidp, weight(final_weight)
svy: melogit y x || psu: || ssu:"

generates an error message that reads:

"weights in variable final_weight not constant within groups defined by: psu pidp an error occurred when svy executed melogit."

I half expected that to be honest because the 'final_weight' has got cross-sectional waves so that the same pidp has different weights depending on the year/wave it relates to (i.e., 1 or 2, or 3 etc...)

That said, my thinking was that it should make sense to run the model as:

"svyset psu || pidp, weight(final_weight)
svy: melogit y x || pidp"

That is, signaling 'psu' through the svyset specification and letting the melogit command take care of the clustering of observations at the individual level through the '|| pidp' specification.

Would you be able to confirm if this understanding is correct or if it would be more appropriate to run the model as

"svyset psu || pidp, weight(final_weight)
svy: melogit y x || psu: "

As always, I appreciate your continued help with this.

With very best wishes,

Samir

#7 - 10/22/2019 10:58 AM - Alita Nandi

Hi Samir,

This issue is to do with how melogit works in Stata. My suggestion would be to look at Statalist and see if others have asked and answered this issue. If not, you can post an issue on Statalist and see what they have to say.

For example, here is an issue posted on Statalist about melogit and weights:
https://www.statalist.org/forums/forum/general-stata-discussion/general/1344286-problem-applying-sampling-weights-for-two-level-mixed-effects-logit-melogit

Having said that I had a look at melogit help in Stata. For pweights it said " pweight(varname) sampling weights at higher levels". This means that if your two levels are person (pidp) and time (wave), then pidp is the higher level and so Stata expects there to be a weight at the individual level only. Which means it should be the longitudinal weight from the last wave. But please ask about this on Statalist and confirm.

If you have questions about which weight to use, or need data managgement guidance in order to be able to use the weights, etc please post your question and we will get back to you. If you want to join the Helpdesk Hour tomorrow and ask Olena further questions on weights, please email usersupport@understandingsociety.ac.uk

Best wishes,
Alita

#8 - 10/22/2019 01:06 PM - Samir Sweida-Metwally

Dear Alita,

Thanks a lot for this.

This is v helpful and in line with my understanding from the readings. I had previously reached out via Statalist but hadn't had much luck - I'll try again and hopefully have more success this time.

Just one small point to clarify when you say 'which means it should be the longitudinal weight from the last wave', you mean in the case of if I was looking at undertaking a balanced panel data analysis? But in context of what I previously discussed above with Olena, such a weight could also be the cross-sectional one previously suggested for an unbalanced analysis. Correct?

Thanks a lot,

Samir

#9 - 10/22/2019 01:54 PM - Alita Nandi

The question you need to ask is why are you using weights. If your analysis is longitudinal (FE, RE, multi-level modelling where the 2 levels are person and time) then you should use longitudinal weights as it will account for non-response at wave 1 + attrition over the years. The only reason you have to use balanced panel is that we don't provide weights for any other combination of waves. So, you can either use the longitudinal weights we provide with balanced panel, or produce longitudinal weights which you can use with unbalanced panel data.

If on the other hand you are doing pooled cross-sectional analysis (e.g., pooled OLS) then you should use cross-sectional weights from all the waves as Olena suggested.

Hope this helps,
Alita

#10 - 10/22/2019 02:05 PM - Samir Sweida-Metwally

Thanks a lot Alita - v helpful. Please also forward my thanks to Olena for her help throughout.

Have a pleasant day.

Samir

#11 - 03/02/2021 03:33 PM - Understanding Society User Support Team

- Status changed from Feedback to Resolved

- Assignee deleted (Samir Sweida-Metwally)

- % Done changed from 70 to 100