Understanding Society User Support - Support #2019

Tailored Weighting Guidance

12/27/2023 04:08 PM - Freya Buchanan

Status:	Resolved	**Start date:**	12/27/2023
Priority:	High	**% Done:**	100%
Assignee:	Understanding Society User Support Team		
Category:	Weights		

Description

I'm investigating the effect of mental health on labour market earnings. My dependent variable is fimnlabnet_dv and my independent variable is scghq2_dv. My population of interest is all individuals of working age and I'm using data from waves 1-13 of UKHLS. I will be using a fixed effects regression.

I'm looking to create a weight variable to be included in the regression. I've completed the Moodle course on creating your own tailored weights, have looked at the user guides, and have extensively scrolled through this user forum to see if anyone else has had similar issues but I'm still struggling.

My issues are as follows:

(1)
When attempting to predict attrition

code:

```
    xtreg fimnlabnet_dv scghq2_dv $controls if wave==2|3 [pw=b_indinus_lw], fe cluster(pidp)
    cap drop nonattrition
    gen nonattrition = e(sample)
    replace nonattrition = . if wave!=3
```

All variables are omitted from the regression due to collinearity when the weight is included, why may this be? I have checked extensively and as far as I can tell there is no collinearity present amongst my control variables (NB: $controls is a macro containing my control variables)

(2)
Through the moodle course I learned that I should use the longitudinal weight from the earliest wave of analysis as my base weight (b_indinus_lw, since there is no indinus_lw for wave a). I am then to run a logistic regression predicting response conditional on this base weight. However, when attempting to do this using Stata (code: logit resp2 $wave2predictors [pw=b_weight] if b_weight !=0 & b_weight !=.) I get the return code r(2000); outcome does not vary; remember: 0 = negative outcome, all other nonmissing values = positive outcome. I cannot work out why this might be. I have tried several varieties of the code and different predictors and combinations of predictors but I have had no success. Why may this be occurring and how can I remedy the issue?

(3)
Additionally, I would you please be able to tell me if my methodology for creating my own weights is correct? I'm using this methodology based off of the teachings in the Moodle page but I found it a little challenging to follow so I would hugely appreciate some clarification.
1. Select a base weight (in my case this would be the product of the design weight for the earliest wave in my analysis multiplied by the non-response
2. Predict response using logistic regression.

> When including the base weight in this regression, I get the return code r(2000); outcome does not vary; remember: 0 = negative outcome, all other nonmissing values = positive outcome. This is the issue I have noted above in (2)

3. Predict probabilities
4. Take the inverse of these probabilities (1/prob)
5. Multiply the ipw with the base weight (gen ipwXsampwgt = ipw*b_indinus_lw)

Please note that my panel is unbalanced but I wish for it to remain this way as creating a balanced panel would drop too many observations.

Thank you in advance for any assistance you may have to offer, I really appreciate it. If you need any further clarification please just let me know.

History

#1 - 12/27/2023 09:23 PM - Freya Buchanan

Apologies, the code I used in (1) is actually:

```
xtreg fimnlabnet_dv scghq2_dv $controls if wave==1|2 [pw=b_indinus_lw], fe cluster(pidp)
cap drop nonattrition
gen nonattrition = e(sample)
replace nonattrition = . if wave!=2
```

#2 - 01/02/2024 01:38 PM - Olena Kaminska

Freya,

Thank you for your question. I am not sure you need a tailored weight for your analysis. It sounds like you are planning to use a pooled analysis. Can you explain your data set up better? Are you pooling cross-sectional information from 13 waves? Or are you studying any longitudinal aspect, and if so, how many waves at a time are you looking at?

From what I can see in your description your tailored weight does not work, because your resp2 is wrong (all are respondents). Once you have a clear definition of your data structure, you will be able to define your resp2, and your nonresponse model will run.

Hope this helps,
Olena

#3 - 01/02/2024 08:38 PM - Freya Buchanan

Hi Olena,

Thank you for your swift response.

I have pooled data from all 13 waves using the 'append' command. So I think you are correct, it's pooled cross-sectional analysis as I'm using xtreg and xtlogit commands to look at how an individuals' mental health impacts their labour market outcomes but I'm not doing this for separate waves or over time. Apologies.

In this case, building off of other queries, I am led to believe the correct weighting approach would be to create a new weight that combines the `w'_indinus_lw weight from each wave, and then re-scale it to account for differences in sample sizes, is this correct?

I have attempted to do this using the following code:

```
gen weightscaled=0

gen ind=1
sum ind [aw=b_indinus_lw] if wave==2
gen bwtdtot=r(sum_w)

sum ind [aw=c_indinus_lw] if wave==3
gen cwtdtot=r(sum_w)

sum ind [aw=d_indinus_lw] if wave==4
gen dwtdtot=r(sum_w)

etc

replace weightscaled=c_indinus_lw*(bwtdtot/cwtdtot) if wave==3
replace weightscaled=d_indinus_lw*(bwtdtot/dwtdtot) if wave==4
replace weightscaled=e_indinus_lw*(bwtdtot/ewtdtot) if wave==5
replace weightscaled=f_indinus_lw*(bwtdtot/fwtdtot) if wave==6
replace weightscaled=g_indinus_lw*(bwtdtot/gwtdtot) if wave==7
replace weightscaled=h_indinus_lw*(bwtdtot/hwtdtot) if wave==8
replace weightscaled=i_indinus_lw*(bwtdtot/iwtdtot) if wave==9
replace weightscaled=j_indinus_lw*(bwtdtot/jwtdtot) if wave==10
replace weightscaled=k_indinus_lw*(bwtdtot/kwtdtot) if wave==11
replace weightscaled=l_indinus_lw*(bwtdtot/lwtdtot) if wave==12
replace weightscaled=m_indinus_lw*(bwtdtot/mwtdtot) if wave==13
```

Would you be able to advise if this code is correct please?

One thing I noticed when using the commands:

```
    sum ind [aw=weightscaled] if wave==2
    sum ind [aw=weightscaled] if wave==3
    sum ind [aw=weightscaled] if wave==4
    etc
```

to check my weight variable, is that *sum ind [aw=weightscaled] if wave==2* returns 0 observations and the weight takes a value of 0. I understand why this has occurred however, I am wondering if it would be advisable to add the code *replace weightscaled=b_indinus_lw*(bwtdtot/bwtdtot) if wave==2*? I have not added it thus far as I was attempting to follow guidance on creating rescaled weights found in previous queries.

Additionally, my code excludes wave 1 as there is no a_indinus_lw. Is there an alternative weight I could use for wave 1 instead?

Moreover, when I attempted to use the weight with the command:

```
    xtreg logincome log_ghq_caseness [aweight=weightscaled], fe vce(cluster pidp)
```

I got this response from stata:

```
    weight must be constant within pidp
    r(199);
```

Do you have any advice on how to overcome this issue?

Additionally, would it be advisable to do a similar process (i.e., new variable creation and scaling) for psu and strata variables?

Thank you once again for taking the time to respond to my queries, I really appreciate your help.

Kind regards,
Freya

#4 - 01/03/2024 12:51 PM - Olena Kaminska

Freya,

Yes, this sounds correct. You would need _xw weights, not _lw weights if you are just pooling cross-sectional information. I will copy my general advice for rescaling below.

Strata and psu variables are constant, and you don't need to change them.

I will forward your message to another collegue to help you with xtreg.

Hope this helps,
Olena

Scaling weights

In pooled analysis and sometimes in other types of analysis you may need to apply an additional scaling to our weights. Our weights have a mean of 1 in each wave, which means that if combined in a pooled analysis the waves with smaller sample size will have a smaller contribution in your analysis. This includes BHPS waves and later waves (as sample size decreases with attrition). Ideally, when combining events / states over 30 years (for example) you want each year to have the same importance. To ensure this follow this example to calculate an additional scaling for your weights.

For example, you are looking at job quality and therefore are pooling information from wave 2, 4, 6 & 8 as these are the waves when the questions are asked. Here is how to create a scaled weight for this analysis.

```
ge weightscaled=0
replace weightscaled=b_indpxub_xw if wave=2

ge ind=1
sum ind [aw=b_indpxub_xw] if wave=2
gen bwtdtot=r(sum_w)
sum ind [aw=d_indpxub_xw] if if wave=4
gen dwtdtot=r(sum_w)
sum ind [aw=f_indpxub_xw] if if wave=6
gen fwtdtot=r(sum_w)
sum ind [aw=h_indpxub_xw] if if wave=8
gen hwtdtot=r(sum_w)

replace weightscaled=d_indpxub_xw*(bwtdtot/dwtdtot) if wave=4
replace weightscaled=f_indpxub_xw*(bwtdtot/fwtdtot) if wave=6
replace weightscaled=h_indpxub_xw*(bwtdtot/hwtdtot) if wave=8
```

You can double check by looking at the sum of ind with weightscaled for each wave – it should be the same.

```
sum ind [aw=weightscaled] if wave==2
sum ind [aw=weightscaled] if wave==4
sum ind [aw=weightscaled] if wave==6
sum ind [aw=weightscaled] if wave==8
```

#5 - 01/03/2024 12:51 PM - Olena Kaminska

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

#6 - 01/04/2024 09:21 AM - Understanding Society User Support Team

- Status changed from New to Feedback

- % Done changed from 0 to 50

- Private changed from Yes to No

Hello Freya

Here is the reply regarding xtreg

I can't tell for sure from the information provided, but this seems less to do with xtreg (after all, there are at most two waves per individual) to be something to do with the properties of the sample using the longitudinal weight b_indinus_lw **if** the model runs without the longitudinal weight specified. The zero weights have created a sample (possibly with only one observation per individual?). In combination with the 'fe' option, it could be that the implied design matrix for the model under the fixed-effects estimator for this sample displays collinearity and thus won't fit. Rerunning xtreg with the weights specified but using the 're' option will show whether it's the combination of the longitudinal weight **and** the fe option that's causing the problem, or simply the use of the weights.

In either case, the user will have to check the weight <> 0 sample to see whether it looks strange or not (number of non-missing cases per individual, things like that).

However, it does not seem that using the e(sample) option from the resulting fitted model is the best way to define the attrition variable and it would be best done manually using the mv commands.

I hope this information is helpful.

Best wishes,
Roberto Cavazos
Understanding Society User Support Team

#7 - 02/23/2024 02:35 PM - Understanding Society User Support Team

- Status changed from Feedback to Resolved

- % Done changed from 50 to 100