Understanding Society User Support - Support #2154

Scaling weights for pooled calendar year analysis

09/19/2024 04:57 PM - Emma Maun

Status:	Feedback	**Start date:**	09/19/2024
Priority:	Normal	**% Done:**	90%
Assignee:	Olena Kaminska		
Category:			

Description

Hello,

I have pooled data from waves 9 to 12 and am carrying out pooled cross-sectional analysis on people who subsequently died (just keeping their last wave of data). I would like to analyse the proportion of several variables across calendar years 2018-2020 and would like to check my strategy for creating and scaling the weights - I have used q14-16 in the weighting paper 2024. I think I need to both ensure yr1 and yr2 of each calendar year contributes equally to estimates, and that each calendar year as a whole contributes equally to the estimates.

I am calculating the weights for the full pooled dataset before selecting my sample (the final sample is very small). I first applied a scaling factor to balance yr 1 and yr 2 for one of the calendar years (2019). I then used the scaled yr1 and yr2 2019 to scale the other calendar years for each 12 month period. My syntax is pasted below in case helpful, please could you let me know if this is the right way to ensure pooled calendar year data is properly weighted?

With thanks,
Emma

```
**
*generate scaled self-completion weight for yearly analysis, using 2019 as the base for scaling.
ge wt_yr_scld=0  //scaled weight for analysis by calendar year

replace wt_yr_scld=indscui_xw if wave==10 & (month>=13 & month<=24)  //2019 for wave 10 - for scaling period months 13-24 in other waves

ge ind=1  //generate variable to scale the weights - first to balance the yr 1 and yr 2
sum ind [aw=indscui_xw] if wave==10 & (month>=1 & month<=12)
gen awtdtot=r(sum_w)   //weighted total for wave 10 mths 1-12
sum ind [aw=indscui_xw] if wave==11 & (month>=1 & month<=12)
gen bwtdtot=r(sum_w)   //weighted total for wave 11 mths 1-12

replace wt_yr_scld=indscui_xw*(awtdtot/bwtdtot)  if wave==11 & (month>=1 & month<=12)  //2019 in wave 11 - upweighted so yr2 weighted in same way as in wave 10, so it contributes as much to estimates.

*checking syntax
sum ind [aw=indscui_xw] if wave==10 & (month>=1 & month<=12)
sum ind [aw=wt_yr_scld] if wave==11 & (month>=1 & month<=12)

*use the scaled calendar year weights for 2019 to generate for the other calendar year weights so all contribute equally
drop ind awtdtot bwtdtot

ge ind=1
sum ind [aw=wt_yr_scld] if wave==11 & (month>=1 & month<=12)
gen awtdtot=r(sum_w)  //calendar year 2019 yr2 scaled
sum ind [aw=wt_yr_scld] if wave==10 & (month>=13 & month<=24)
gen dwtdtot=r(sum_w)   //calendar year 2019 yr1 scaled

*generate the remaining weighted totals for each 12 month period for scaling - first months 1-12 and then 13-24
sum ind [aw=indscui_xw] if wave==10 & (month>=1 & month<=12)
gen bwtdtot=r(sum_w)  //weighted total wave 10 mths 1-12 (2018)
sum ind [aw=indscui_xw] if wave==12 & (month>=1 & month<=12)
gen cwtdtot=r(sum_w)  //weighted total wave 12 mths 1-12 (2020)

sum ind [aw=indscui_xw] if wave==9 & (month>=13 & month<=24)
gen ewtdtot=r(sum_w)  //weighted total wave 9 mths 13-24 (2018)
sum ind [aw=indscui_xw] if wave==11 & (month>=13 & month<=24)
```

```
gen fwtdtot=r(sum_w)  //weighted total wave 11 mths 13-24 (2020)

scale each 12 month period to the same 12 month period in 2019, first for months 1-12 then 13-24
replace wt_yr_scld=indscui_xw(awtdtot/bwtdtot) if wave==10 & (month>=1 & month<=12)  //scaled weight wave 10 mths 1-12 = 2018
replace wt_yr_scld=indscui_xw*(awtdtot/cwtdtot) if wave==12 & (month>=1 & month<=12)  //scaled weight wave 12 mths 1-12 = 2020
replace wt_yr_scld=indscui_xw*(dwtdtot/ewtdtot) if wave==9 & (month>=13 & month<=24)  //scaled weight wave 9 mths 13-24 = 2018
replace wt_yr_scld=indscui_xw*(dwtdtot/fwtdtot) if wave==11 & (month>=13 & month<=24)  //scaled weight wave 11 mths 13-24 = 2020

*checking
sum ind [aw=indscui_xw] if (wave==9 & (month>=13 & month<=24)) | (wave==10 & (month>=1 & month<=12))  //2018 not scaled
sum ind [aw=wt_yr_scld] if (wave==9 & (month>=13 & month<=24)) | (wave==10 & (month>=1 & month<=12))  //2018

sum ind [aw=indscui_xw] if (wave==10 & (month>=13 & month<=24)) |  (wave==11 & (month>=1 & month<=12) )  //2019 not scaled
sum ind [aw=wt_yr_scld] if (wave==10 & (month>=13 & month<=24)) |  (wave==11 & (month>=1 & month<=12) )  //2019

sum ind [aw=indscui_xw] if (wave==11 & (month>=13 & month<=24)) | (wave==12 & (month>=1 & month<=12))  //2020 not scaled
sum ind [aw=wt_yr_scld] if (wave==11 & (month>=13 & month<=24)) | (wave==12 & (month>=1 & month<=12))  //2020
```

History

#1 - 09/20/2024 09:29 AM - Understanding Society User Support Team

- Status changed from New to In Progress

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

- % Done changed from 0 to 10

- Private changed from Yes to No

#2 - 09/20/2024 12:24 PM - Olena Kaminska

Emma,

Thank you for your question. If you use pooled analysis pooling all the waves between wave 9 and 12, why are you working with calendar year concept? If it is not important in your analysis, you don't need to involve it in weighting. For studying death you probably don't need to implement any scaling if you are using wave 9 to 12.

Just a note, you are using a interviewer identified death, that's roughly 50% of death overall, so please be aware of this.

Hope this helps,
Olena

#3 - 09/24/2024 04:04 PM - Emma Maun

Hi Olena,

Thank you for your response. The reason for pooling waves but using calendar years is that one of my research questions relates to changes over time and specifically looking at change/stability pre-covid to during the pandemic. Analysis by calendar wave is a more sensitive time period (1 year) than by wave (2 years). Can you say why scaling isn't needed in this case? I am looking at the circumstances of people who died in the 12 months before they died.

Thank you for your note on interviewer identified death. Can you tell me if there is another route to identifying people who died that I'm not aware of with U-Soc data? Also, if you are aware of % capture of people who died, has there been analysis of this you can point me to? That would be very useful for the limitations of my research.

With thanks,
Emma

#4 - 09/25/2024 01:45 PM - Olena Kaminska

Emma,

What you should do is create a dataset where you combined months 1-12 from wave x with months 13-24 from wave x-1. This will be one dataset for one calendar year X (wave x-1 would be a wave that started one year earlier, but months 13-24 correspond to year x). You then treat this as one wave and create one weight for this year, called x_weight. From your syntax I can see that you created separate weights for months 1-12 and 13-24. This won't represent the population. You need one weight for the combined data.
You then create 3 'waves' like these with 3 weights. You can then scale them, if you like. Make sure they are scaled to one number.

On death, I have worked on this topic and presented at ESRA 2021. This has not been published, but I am happy to share a presentation with you, if

you email me (email to usersupport with a request). I compared our estimates to official ONS death rates, but only by time, age and gender. We know that interviewers identify 50% of all death. Note, some became nonrespondents, sometimes long time ago, and subsequently died. So they are marked on the dataset as nonrespondents first, then often we have given up on contacting them, and later they died. Internally we linked our data to death registers, which identifies another 25% of death. But I noticed that these died with 2+ years since their last response. Out a couple of hundreds death that we identify additionally via registry only 1-2 have died in the last year (maybe this information would help you). So, most often they become nonrespondents first. Another 25% are not identified, but we know the proportions from the official statistics. Often these people move to a carehome, move between them, or spend long spells in hospital at which point we lose contact and our last address on the dataset does not correspond to the address of the death registry.
We only can check age by gender distribution of interviewer identified death and compare it to official statistics. But we have a guess that we identify many more death where there is a surviving partner or some other member of a household. This is because the household continues to respond and tells us about death. We tend to not have information on people living on their own before death. This may be important to your research, hence the information.

There is a little bit on death in UKHLS in this paper:
https://www.understandingsociety.ac.uk/wp-content/uploads/working-papers/2011-04.pdf

Hope this helps,
Olena

#5 - 09/25/2024 02:15 PM - Emma Maun

Hi Olena,

Thank you very much for the clarification of weighting for years and for the detailed information on capturing deaths. This is very important, as you say, for my research so thank you very much. I am emailing to request your presentation, many thanks for offering to share it.

With thanks,
Emma

#6 - 09/25/2024 05:43 PM - Emma Maun

Hi Olena,

I am just working on the weights for pooled analysis by years. When looking at the explanation in Q14 of the Weighting Questions Answered document, the document first notes "For example, if we pool sample months 1 to 12 from wave 3 with sample months 13 to 24 from wave 2, the former will be under-represented (as the responding sample size is smaller at wave 3 than at wave 2)1. To overcome this, we should scale the weights for these cases to give the same weighted total that this sample had at wave 2.". This is what I did first in the code above.

Further down, in the same doc and Q14 it notes "This rescaling becomes even more important when pooling data from more than one 12-month period (e.g. two calendar years). In that case, in addition to the imbalance between the 24 monthly samples, the relative contribution to the estimate (weighted sample size) will also tend to be less for the later year(s) unless rescaling is done, such that each year contributes equally to the estimate. This is achieved by scaling all of the weights to the relevant weighted totals from one common wave."

I had read these two paragraphs together but based on your response above, do I understand correctly that to correctly weight pooled calendar years of data, it is only necessary to create a weight for one of calendar years (with no rebalancing for smaller size of sample for months 1-12 in wave x+1) and then use that calendar year weight to scale all other calendar years of data so they each contribute equally to the analysis?

It is quite tricky to follow so thank you for any more help you can give.

All the best,
Emma

#7 - 09/27/2024 01:47 PM - Olena Kaminska

Emma,

Yes, your understanding is correct. In your situation first create a dataset for each year, and then rescale the weight for all the years to contribute the same amount to the analysis. To check, you can find a total of weights for each year (e.g. total weight if year==#). These should be the same for each year separately.

If so, you are good to go.
Best,
Olena

#8 - 10/02/2024 07:41 AM - Understanding Society User Support Team

- Status changed from In Progress to Feedback

- % Done changed from 10 to 90