Understanding Society User Support - Support #1509

Real Household Income + Calendar Year vs Wave

02/15/2021 06:23 PM - Vikramsinh Patil

Status:	Resolved		**Start date:**	02/15/2021
Priority:	High		**% Done:**	100%
Assignee:				
Category:				

Description

Hello,

In my Metrics project, I am trying to explore the causal impact of income on subjective well-being by exploiting real wage fluctuations in the aftermath of the Brexit referendum. I am using data from Waves 8,9 and 10 of the UKHLS. I have merged net household income, month, and year of the household interview with individual-level files (through hidp) at the wave level and then merged these three waves through pidp to create a long format panel. Finally, using the month and year of household interview, I merged my combined panel with the CPI monthly indices.

My question is, that when it comes to analysis should my time variable be the wave or the calendar year? More specifically, if I were to try plotting average real household incomes over time, do I create calendar year-based cross-sections (so 2016, 2017, etc), compute the real wages for those year based cross-sections, and then plot the movement of these averages over the years? I realized that since in each wave, interviews are conducted over 2 years, averaging real wages for Wave 8, for example, would pick up incomes reported in 2016, 2017, and 2018 and by mixing these years, I was worried I would be misrepresenting the actual macro trend.

Closely related then, in my econometric analysis should my time dummies be the wave or the year of the interview?

Finally, if I should create calendar year based cross-sections and then merge to create a panel, what would be the appropriate weight variable to include? Going through the weighting FAQs, I think it should be j_indin_lw but I wanted to double-check if this is still correct if I have to adjust to using calendar year-based samples?

Apologies for these long-winded and perhaps basic queries but I just wanted to check any big and avoidable mistakes!

Thanking you,

Best,
Vikram

History

#1 - 02/17/2021 08:27 AM - Understanding Society User Support Team

- Status changed from New to In Progress

- Assignee set to Olena Kaminska

- % Done changed from 0 to 50

- Private changed from Yes to No

Hello Vikram,

The way you have set up the data is ok.

"More specifically, if I were to try plotting average real household incomes over time, do I create calendar year-based cross-sections (so 2016, 2017, etc), compute the real wages for those year based cross-sections, and then plot the movement of these averages over the years? I realized that since in each wave, interviews are conducted over 2 years, averaging real wages for Wave 8, for example, would pick up incomes reported in 2016, 2017, and 2018 and by mixing these years, I was worried I would be misrepresenting the actual macro trend."

- To this case, using year and not wave as the time variable is more appropriate

About weights: The weighting FAQ items 11 & 12 discuss this. I am assigning this issue to our Survey Statistician if she wants to add anything.

Best wishes,
Understanding Society User Support Team

#2 - 02/17/2021 07:27 PM - Vikramsinh Patil

Okay, thank you so much.

As a follow-up, if I xtset the data with pidp and wave as my time variable and add dummies for intdatey (2016, 2017, 2018, 2019), would that be

incorrect?

Further, for the purpose of graphing, instead of creating calendar year based cross sections and then computing real wage averages, would the same result be achieved simply by the following:-

by intdatey, sort: summarize rminc [where rminc is real monthly household income] and then taking the reported means?

#3 - 02/18/2021 02:42 PM - Understanding Society User Support Team

Hello Vikram,

Our remit is to answer questions related to the survey & its details and the datasets (incl data management). Due to staff time constraints we are not able to answer questions related to analysis and syntax (unless related to data management or weighting).

There are multiple ways of creating means in Stata, the method you have suggested will work but you have not specified the weights or the sample design. You can do that by using svyset:
svyset psu [pw=weight variable], strata(strata)
svy: mean rminc, over(year variable)

Best wishes,
Understanding Society User Support Team

#4 - 02/22/2021 02:56 PM - Olena Kaminska

Vikram,

Just to add, if you use a calendar year you will need to use three weights (see below an extract from FAW that explains this).

You can of course use wave as your time variable but I wouldn't include intdatey in your model. This variable is a processing variable and is related to the likelihood of response - I am not sure this has any meaning for your substantive model.

from FAQ, question 11

Let's say you are interested in studying December 2014. Your optimal option with the
largest sample size will be to combine all interviews carried out in December of 2014
from the following samples:
- Wave 5 sample months 21, 22, 23 and 24
- Wave 6 sample months 9, 10, 11 and 12
- Create a new variable that equals e_xxxxxus_zz weight for the wave 5
interviews and f_xxxxxus_zz weight for wave 6. No Northern Ireland
adjustment is needed. No extra nonresponse adjustment is needed as late
respondents in the month 24 sample are compensated for by bringing in the late
respondents from previous sample months. But you will need a scaling factor
(see Q12).
- Use psu and strata variables from xwave.dat to take into account clustering and
stratification.
Note if you want to study January 2014 for example, the information will come from 3
waves, because to compensate for missing of late respondents from wave 5, sample
month 1, you will need to include January respondents from wave 4, sample months 22-
24. The rest will follow the above example.
If you use respondents from calendar months / year just from one wave you will need an
extra adjustment for Northern Ireland and potentially also for late respondents (if your
period of interest includes sample months 1, 2 or 3).

#5 - 02/22/2021 07:52 PM - Vikramsinh Patil

Dear Olena,

Thank you so much for your response!

I apologize for asking something that did not fit with the remit of this user forum but thanks so much again for your response

If I may have just one more follow-up when I perform svy: mean (rminc), over(wave), STATA gives the output along with a message stating 'missing standard errors because of stratum with single sampling unit'.

Is there any way to address this?

Best,
Vikram

#6 - 02/24/2021 06:37 AM - Understanding Society User Support Team

- Status changed from In Progress to Feedback

#7 - 02/24/2021 06:37 AM - Understanding Society User Support Team

- Status changed from Feedback to In Progress

#8 - 02/25/2021 03:07 PM - Olena Kaminska

- Assignee changed from Olena Kaminska to Alita Nandi

#9 - 03/02/2021 02:34 PM - Understanding Society User Support Team

- Assignee changed from Alita Nandi to Olena Kaminska

#10 - 03/02/2021 04:16 PM - Understanding Society User Support Team

- Status changed from In Progress to Feedback

- Assignee changed from Olena Kaminska to Vikramsinh Patil

- % Done changed from 50 to 80

Hello Vikram,

This is an issue because in your final analysis sample, there are strata with only one PSU. When you specify the svyset use the singleunit option.

svyset psu [pw=weight], strata(strata) singleunit(scaled)

You can check Stata help to see the different ways in which singleunit can be specified, each results in a different way in which Stata computes the standard errors.

#11 - 10/12/2021 04:56 PM - Understanding Society User Support Team

- Status changed from Feedback to Resolved

- Assignee deleted (Vikramsinh Patil)

- % Done changed from 80 to 100