Understanding Society User Support - Support #1609

Including/excluding zero weights changes SEs (but not coefficients)

11/22/2021 10:35 AM - Marie Mueller

Status:	Resolved	**Start date:**	11/22/2021
Priority:	Normal	**% Done:**	100%
Assignee:			
Category:			

Description

Hello,

In my project, I am analysing youth data. For my analyses, I define my analytic sample. For example, youth need to be from London and have a non-missing value on the outcome of interest. An additional condition for my analytic sample is that they have a non-zero study weight, as there are several youth with a weight of zero.

I see that you say that one does not need to worry about zero weights, as they will automatically be taken into account when you apply weights to your analysis. I understand that, if one has a weight of zero, data essentially do not contribute to the coefficient (as they 'weigh' zero). However, I realised that, while the **coefficient** does not change, the **SE** and therefore p value and CI do (due to the larger sample size when I include youth with zero weights).

Now I am not sure what would be more appropriate: excluding youth with zero weights or keeping them. I assumed that an individual with a zero weight means that this individual is not a 'valid case' for my analysis. Therefore, I excluded them from my analytic sample. Therefore, only individuals with a weight > zero were included in my analysis. To me this makes sense and I would assume that the associated SE is more valid than if I included all the youth with zero weights in my analysis (thereby increasing the sample size). In other words: if youth with zero weights should not be included in my analysis/contribute to the coefficient, why should I keep them, so they still affect the sample size and SEs?

I hope this makes sense – Thank you very much in advance!

Best wishes,
Marie

History

#1 - 11/22/2021 01:30 PM - Marie Mueller

PS In the FAQ on pooling cross-sectional data you say: "We strongly recommend that a non-zero value of the
weight variable is used to define the analysis base (see example below)." Removing individuals with a zero weight from my analytic sample produces the same result as specifying "if weight > 0" in the analysis itself. So, can I assume that removing individuals with a weight of zero from my analytic sample is okay?

#2 - 11/24/2021 09:06 AM - Understanding Society User Support Team

- Status changed from New to In Progress

- Assignee set to Olena Kaminska

- % Done changed from 0 to 10

- Private changed from Yes to No

#3 - 11/24/2021 09:08 AM - Understanding Society User Support Team

- Status changed from In Progress to Feedback

- % Done changed from 10 to 50

#4 - 11/24/2021 11:27 AM - Olena Kaminska

Marie,

I am not sure about your question. If you use weights correctly people with 0 weights won't contribute to your estimate. If you drop them first (I would discourage from this and would just rely on weights to do the correct job for you) - the weighted estimate should be identical, including std and CI.

If you want to avoid many zero weights you may want to create a tailored weight.

Does this help?
Olena

#5 - 11/24/2021 12:18 PM - Marie Mueller

Hi Olena,

Interesting! Let me give you an example that I just tested:

I have a sample of n = 2,671. Of these, 396 have a zero weight. Removing these from my analytic sample results in n = 2,275.

Let's say I want to test the effect of sex on SDQ total difficulties. I use:

```
svyset psu [pweight = weight], strata(strata)

svy: regress sdq_td sex
```

Three scenarios

1. I remove individuals with a zero weight from my analytic sample (n = 2,275) and run the above:
n = 2,275, df = 449, b = -0.024, SE = 0.360

2. I do **not** remove individuals with a zero weight from my analytic sample (n = 2,671) and run the above:
n = **2,671** , df = **496** , b = -0.024, SE = **0.352**

3. I do **not** remove individuals with a zero weight from my analytic sample (n = 2,671) and run the above **but** specify "if weight > 0" in the regression:
n = 2,275, df = 449, b = -0.024, SE = 0.360

Scenario 1 and 3 result in exactly the same results. Scenario 2 results in the same b, but in a different SE (which I assume is due to the larger n). In scenario 3, I do what you say in the FAQ: "We strongly recommend that a non-zero value of the weight variable is used to define the analysis base (see example below)." And this is the same as removing individuals with zero weights from my analytic sample beforehand.

What am I missing?

Best wishes,
Marie

#6 - 11/24/2021 03:07 PM - Olena Kaminska

Marie,

Are all of your 3 scenarios weighted? And do you use the same svyset for them?

Thanks,
Olena

#7 - 11/24/2021 03:23 PM - Marie Mueller

Hi Olena,

Yes, I use the same svyset (and therefore all 3 scenarios are weighted). I guess this is evident also in the coefficient which is exactly the same. The difference in scenario 2 is the n, df, SE, p, and CI. What I also noted: while the number of strata is the same in all three scenarios (because strata with no population members were omitted in scenario 2), this is not true for PSUs (there are more PSUs in scenario 2 than in scenarios 1 and 3).

Thank you!

Marie

#8 - 11/25/2021 11:43 AM - Olena Kaminska

Marie,

Yes, I see what you mean, and I was able to replicate your results. By expectation this should not happen. But Stata recognizes that this may happen in some circumstances:
"svy commands handle zero sampling weights properly. Standard commands ignore any observation with a weight of zero. Usually, this will yield the same standard errors, but sometimes
they will differ. Sampling weights of zero can arise from various postsampling adjustment
procedures. If the sum of weights for one or more PSUs is zero, svy and standard commands
will produce different standard errors, but usually this difference is very small."

from p. 104 https://www.stata.com/manuals/svy.pdf

Hope this helps,
Olena

#9 - 11/25/2021 12:00 PM - Marie Mueller

Hi Olena,

Very helpful - thank you! Now I am confident that I did use the weights correctly.

I guess it is on me to decide whether or not to include youth with zero weights in my analytic sample. As some of my p-values are very close to .05, I'd rather take the more conservative approach: remove youth with zero weights from my analytic sample, so SEs are based on youth that are actually considered in my analysis.

For a better understanding, I wanted to check: if you typically assume that including individuals with zero weights will result in the same estimates, why does the FAQ say: "We strongly recommend that a non-zero value of the weight variable is used to define the analysis base (see example below)."

https://www.understandingsociety.ac.uk/sites/default/files/downloads/general/weighting_faqs.pdf (page 9)

Thank you very much for your help!

Best wishes,
Marie

#10 - 11/25/2021 02:45 PM - Olena Kaminska

Marie,

The FAQ is not related to the question you have. It relates to a pooled dataset where if you pool over many years some years (like BHPS years) will have a smaller sample size than waves from UKHLS - in a pooled analysis this will skew it towards later years. The FAQ talks about how to correct for this - by scaling totals of non-zero weights (non-zero weight totals will be sample sizes for each wave used in your analysis, so that's what needs scaling in that situation).

Best,
Olena

#11 - 11/26/2021 08:08 AM - Marie Mueller

Hi Olena,

Thank you!

What confused me is this part:

> svyset psu2011 [pw=weight2011], strata(strata2011) singleunit(centered)
>
> svy: proportion jbstat2011 if weight2011>0

You seem to exclude individuals with a zero weight from the analysis. If I understand correctly, this part has nothing to do with the scaling itself, but is simply an example for a possible analysis. If weighting accounts for zero weights automatically (i.e., I do not need to remove them beforehand), why do you specify "if weight2011>0" here. Anyway, I may come back to the question about scaling in issue #1608 (after we clarified whether tailoring my own weight would be a good idea and what a suboptimal weight would be). [This issue can be marked as resolved, as the question about scaling does not really belong here.]

Best wishes,
Marie

#12 - 11/26/2021 12:29 PM - Olena Kaminska

Marie,

I can confirm that "if weight2011>0" is not necessary in this example and should result in the same estimates (by expectation) if this were not specified.

Thanks,
Olena

#13 - 11/26/2021 03:34 PM - Marie Mueller

Thanks!

#14 - 01/27/2022 02:05 PM - Understanding Society User Support Team

- Status changed from Feedback to Resolved

- Assignee deleted (Olena Kaminska)

- % Done changed from 50 to 100