Understanding Society User Support - Support #414

Weights for unbalanced panel

09/10/2015 06:44 PM - Ewan Carr

Status:	Closed	**Start date:**	09/10/2015
Priority:	Normal	**% Done:**	100%
Assignee:	Olena Kaminska		
Category:	Weights		

Description

My question is very similar to this one, posted last month.

I am estimating a random intercept logistic regression model, drawing upon all available data in the BHPS and US samples. Since I am using repeated measures data, from multiple waves, my understanding was that I would need to apply longitudinal weights to this analysis (specifically, "the weights from the last wave of any longitudinal sequence"; A5-1).

However, I want to use all available data. I do not want to limit the analysis to individuals with full information (i.e. complete cases) between BHPS wave 1 and US wave 4.

In the response given here, it was suggested that the cross-sectional weights should be used in this situation:

> The last of these (unequal inclusion probabilities) is what weights are designed to deal with. **I suggest that for each observation you use the relevant cross-sectional weight.** That should correct for design probabilities and non-response.

However, this seems to go against the advice given in the User Guide (i.e., to use the longitudinal weights).

Can I confirm, therefore, that this approach (i.e. to apply the relevant cross-sectional weights) is a suitable strategy when treating the BHPS/US as an unbalanced panel?

Many thanks in advance.

History

#1 - 09/11/2015 04:50 PM - Olena Kaminska

I hope my answer helps to clarify yours and similar questions. There are potentially three scenarios:
1) if you are representing people over time, you should use longitudinal weights;
2) if you are looking at events within last number of years then you can pool the data and you should use cross-sectional weights;
3) some models allow for missing data and account for nonresponse (e.g. they are okay if person has only 1 observation, or if person missed some waves but answered other waves). In this situation make sure to use the cross-sectional weight from wave 1. Basically, your model 'corrects' for everything since wave 1, but has no idea what happened at wave 1 - thus wave 1 weight is super important.

The distinction between 2) and 3) is in the structure of your data (i.e. does the program think that you are looking at a person over time with some missing information or does it think that you are looking at many events clustered within a person).
The analysis you suggest is either 2) or 3). You will know it better.

One little note: if your scenario is 2) and you are using both BHPS and UKHLS data, make sure to scale your data - otherwise recent years are highly overrepresented.

Hope this helps,
Olena

#2 - 09/11/2015 05:16 PM - Ewan Carr

Many thanks, that's really helpful.

Weights

I'm pretty sure I want scenario (3):

> ...or does it think that you are looking at many events clustered within a person.

So, I need to do the following:

> ...In this situation make sure to use the cross-sectional weight from wave 1.

If I understand this correctly, this means **I want to weight my analyses with axrwght**. This also means, I think, that I can only include the 10,264 people who responded in BHPS wave 1 (i.e. who were assigned a non-missing value for axrwght).

Is this correct?

Scaling

> One little note: if your scenario is 2) and you are using both BHPS and UKHLS data, **make sure to scale your data** - otherwise recent years are highly overrepresented.

I'm not familiar with this procedure. Is this described in the documentation somewhere?

Many thanks again,

Ewan
--

Olena Kaminska wrote:

> I hope y answer helps to clarify yours and similar questions. There are potentially three scenarios:
> 1) if you are representing people over time, you should use longitudinal weights;
> 2) if you are looking at events within last number of years then you can pool the data and you should use cross-sectional weights;
> 3) some models allow for missing data and account for nonresponse (e.g. they are okay if person has only 1 observation, or if person missed some waves but answered other waves). In this situation make sure to use the cross-sectional weight from wave 1. Basically, your model 'corrects' for everything since wave 1, but has no idea what happened at wave 1 - thus wave 1 weight is super important.
>
> The distinction between 2) and 3) is in the structure of your data (i.e. does the program think that you are looking at a person over time with some missing information or does it think that you are looking at many events clustered within a person).
> The analysis you suggest is either 2) or 3). You will know it better.
>
> One little note: if your scenario is 2) and you are using both BHPS and UKHLS data, make sure to scale your data - otherwise recent years are highly overrepresented.
>
> Hope this helps,
> Olena

#3 - 09/22/2015 12:09 PM - Olena Kaminska

Ewan,

You can either use BHPS dataset with the first wave weight and have information from a longer period, or you could use UKHLS data from UKHLS w1 or combined UKHLS + BHPS data from w2. The latter two options will give you less time frame but more people.

Yes, the scaling is not a common thing, and I am copying below my response to a similar question (support #228) below:
"2. Take the cross-sectional weight but scale it (and all other waves) such that each wave has the same weighted sample size. The explanation is below.

The aim of pooling data from different waves is often to represent events (e.g. number of events in the last 20 years). This works fine if each wave has the same number of people. As you know even BHPS does not have the same number of people (there is a boost in 1999 for example). While each single wave (once weighted using cross-sectional weight) represents the population in that year, the waves that have higher number of people will contribute to your estimates more than the waves with smaller number of people. Even before wave 3, if one uses pooled BHPS data to study events in GB over the last 20 years the years before 1999 would be underrepresented and therefore events after 1999 would make a larger contribution on your estimate.

It is easy to correct for this. 1) first calculate the weighted sample size for each wave (total of weight variable will give you this - note the weight variable should have mean of one); 2) take the average of weighted sample sizes across the waves you use; 3) divide the average by the weighted sample size for each year to get the scaling factor; 4) multiply the scaling factor for each wave by its cross-sectional weight. Use this product as a new weight for pooled data. This will ensure that each wave has the same weighted sample size and therefore each year has the same importance in your estimate. For example if one wave has weighted sample size of 1000 and another has 2000, then the average is 1500, the scaling factor for wave 1 is 1500/1000=1.5; for wave 2 is 1500/2000=0.75. The new weighted sample size (you could check this) will be the same in both waves (1500).

Treat the new BHPS + GPS + EMB sample in the same way - the scaling factor will be small for this wave and the scaling factor for BHPS waves will be over 1. But after correction your analysis will have higher precision (then if you were to not use GPS and EMB data) and will correctly and evenly represent all years. Finally, this method also corrects for differences in sample size due to non-response as well. In other words it should be used with pooled data even when there aren't sample boosts."

Best,
Olena

#4 - 09/24/2015 03:22 PM - Ewan Carr

Many thanks for your reply, this is very helpful.

> You can either use BHPS dataset with the first wave weight and have information from a longer period, or you could use UKHLS data from UKHLS w1 **or combined UKHLS + BHPS data from w2**. The latter two options will give you less time frame but more people.

I am aiming for the latter option: **combined UKHLS + BHPS data from w2**.

If I understand correctly, I want to use the **scaled**, **cross-sectional weight**, **from each wave** (and not the cross-sectional weight from BHPS wave 1, i.e. axrwght). This means I can include all BHPS respondents, not just the 10,264 responding in BHPS wave 1.

Many thanks again,

Ewan
--

#5 - 09/29/2015 03:42 PM - Olena Kaminska

Ewan,

You would need a weight called b_psnenub_xw or b_indpxub_xw (or another depending on the variables you are using). These weights are not currently on the dataset and will be released with the current release, but you can recreate them. To create these weight you need to assign the values of BHPS xw weight for BHPS sample and the values of UKHLS xw weight for UKHLS sample. For example, for enumeration use the following code:
gen b_psnenub_xw=0
replace b_psnenub_xw=b_psnenbh_xw if b_psnenbh_xw!=0
replace b_psnenub_xw=b_psnenus_xw if b_psnenus_xw!=0

Hope this helps,
Olena

#6 - 10/02/2015 08:38 PM - Ewan Carr

Olena,

Perfect, thanks. That makes sense. Many thanks for your help with all of this.

Ewan
--

#7 - 10/07/2015 09:10 AM - Redmine Admin

- Status changed from New to Closed

- % Done changed from 0 to 100

#8 - 11/10/2015 10:44 AM - Gundi Knies

- Category set to Weights

- Assignee set to Olena Kaminska

- Target version set to X M