Subject: Pooled weights
Posted by denisshek on Thu, 20 Apr 2017 13:15:42 GMT
View Forum Message <> Reply to Message

I am doing an anlaysis using merged data set from different countries and different survey years. I understand that i have to denormalize the weights (v005) after having gone through previous post in the forum, however, i want to be sure i am doing the right thing.

Can i de-normalize such that "the weights in each survey sum up to 1" and then multiply these weights by the population of interest in this case women aged 15-49 years at the time of survey such that each country in total gets weight equal to the population size of women aged 15-49 years at the time of survey.

Thanks.

Subject: Re: Pooled weights
Posted by Bridgette-DHS on Fri, 21 Apr 2017 11:42:24 GMT
View Forum Message <> Reply to Message

Following is a response from Senior DHS Stata Specialist, Tom Pullum:

Yes, you can do that. Such weights are sometimes called "inflation" weights. The main problem with that approach, rather than, say, giving the same weight to each country, is that large countries will outweigh small countries and dominate the pooled estimates. I suggest that you try this approach and then see what happens to the estimates when you add or drop various countries--that is, do a sensitivity analysis. There is also a conceptual problem, that the surveys refer to different time points and the countries almost certainly do not comprise a standard geographic region or subregion. However, this is a judgment call by the researcher.

Subject: Re: Pooled weights
Posted by denisshek on Fri, 21 Apr 2017 18:42:44 GMT
View Forum Message <> Reply to Message

Thanks alot for the prompt reponse.

Subject: Weights in country-year level regressions
Posted by amil on Fri, 16 Nov 2018 18:07:44 GMT

View Forum Message <> Reply to Message

Dear Tom,

I have a related question but I am not sure your answer below applies to my case.

In brief, I am using all DHSs and created a country-year dataset (1 observation for each survey, for a total of 185 observations). It seems to me that, when running regressions, one should use weights that reflect the size of each country. For example India should weigh more in the regression than smaller countries, say Maldives.

What is the correct way to do regression analysis in this case?

I am thinking to run:

reg y x1 x2 [aweight=n]

analytic weights because I am using means calculated from each surveys (all these means have been calculated using DHS-provided sampling weights). n is the number of observations in the survey (or perhaps one should use population instead).

Your help is as always greatly appreciated!
thanks

Subject: Re: Weights in country-year level regressions
Posted by Bridgette-DHS on Mon, 19 Nov 2018 14:15:07 GMT
View Forum Message <> Reply to Message

Following is a response from Senior DHS Stata Specialist, Tom Pullum:

There have been several related postings on the forum. I agree that this is an instance in which aweights are appropriate. The options are (a) weight by sample size, (b) weight by population size, (c) weight equally. Your choice for a regression should be the same as your choice for an overall mean. The first question is this: why would you want to calculate an overall mean or regression from all these surveys? I can't think of any good justification for pooling 185 data files. Is there a meaningful population parameter that you are trying to estimate?

Many countries have had multiple surveys. Does it make sense to include one country once and another country six times, say? If you reduce to one survey per country, which survey would you select? Would you combine a 1990 survey from one country with a 2010 survey from another country?

If you weight by population size, and combine India and Maldives, then the impact of Maldives would be negligible. What's the value of pooling them?

I hope other forum users will add their viewpoints.

Subject: Re: Weights in country-year level regressions
Posted by amil on Mon, 19 Nov 2018 16:58:17 GMT
View Forum Message <> Reply to Message

Dear Tom,
many thanks for your reply. All your points are very useful.
I should have added that while the dependent variable is a calculated mean from the surveys, most (but not all) of the regressors are statistics (or variables created based on these) from other sources, provided at the country-year level.
The reason why I am including all the surveys, also multiple surveys for a country, is because I am also interested in exploring the time dimension.
You are totally right that by combining countries like India and Maldives (and weighting by sample size), then Maldives would have a negligible impact. So there would be no point in collapsing the data by survey. It would probably be more appropriate to simply estimate regressions directly from the data (without collapsing by country-year) even if I am using statistics from other sources at the country-year level.
Please let me know if you see anything wrong with this approach based on this new info.
Thanks so much.

Subject: Re: Weights in country-year level regressions
Posted by Bridgette-DHS on Mon, 19 Nov 2018 18:39:26 GMT
View Forum Message <> Reply to Message

Following is another response from Senior DHS Stata Specialist, Tom Pullum:

Even if you do not collapse the surveys, you still have to deal with the variation in sample sizes. Surveys with larger samples will tend to dominate. I prefer to revise v005, multiplying by a survey-specific factor. If, say, your combined data file with k surveys has N cases, you would revise the weights so that the weighted number of cases for each survey is the same, N/k. However, even that approach is vulnerable to criticism. Stata code to do this is posted.

Here at DHS, we usually do not pool surveys. When we combine successive surveys from one country into a single data file, that's usually just to make it easier to describe trends. There have been a few times when we pooled surveys because that was the only way to get enough cases, for some rare outcome. The main reason for not pooling is that the reference population, of which the data are supposed to be representative, is too difficult to define. But other users may have a different perspective.