Subject: De-normalizing weights and svyset command in Stata
Posted by hannekeyserhegdahl on Mon, 12 Jan 2015 13:48:38 GMT
View Forum Message <> Reply to Message

Hi!

I have several questions regarding pooled datasets, weighting and svyset-command in Stata:

1) I am pooling datasets from multiple countries into regions to estimate HIV prevalence ratios between men and women in specific regions and to compare ratios from different regions. I have understood that I should de-normalize the weights in each of the countries before I pool them together, however I do not understand exactly how this is done. In the formula for de-normalizing weights, which value for v005 do I use/how do I find this value?

2) Should I do anything more to these de-normalized weights for Stata to understand that they represent different countries (like with the strata and cluster variables)?

3) I have already appended male and individual recode files and merged this "pooled" file with the HIV file for each of the countries. To asses HIV prevalence ratio, I am using the hiv05 (weight from HIV dataset). I have also done analyses on these datasets (each country separately) and did nothing to the hiv05-weigth, however, now I am a bit confused as to wether I should do something to this weight or not, since these datasets also are pooled in a way..

4) I have also understood that I should always use the svyset/svy command in analysing survey data in Stata, but I am using "GLM for the binomial family: binreg" (binreg is not supported by the svyset command), is it possible to work around this in some way?

5) When creating country-specific cluster and strata variables, to what value do I add the different 10000's (10000, 20000, etc), or do I just rename the variables?

Hope someone can clear up these questions for me, as I don't understand everything being discussed in previous threads.

Subject: Re: De-normalizing weights and svyset command in Stata
Posted by Reduced-For(u)m on Wed, 14 Jan 2015 20:53:35 GMT
View Forum Message <> Reply to Message

Not sure I can help on all of these, but I have a few suggestions:

1/2 - I don't know much about the HIV weights or whether they represent individuals or households, but in general I use the following type of procedure. For each survey (country-by-survey-round), I calculate the total sum of weights (Wc) and then take the individual weight (Wi) and divide it by the total weight (Wi/Wc). Now the sum of weights for that country/round equals 1, while preserving the relative probability across people. Then, if I want my weights to be population-level representative, I multiply those weights by the country's population (or, in the case of a sub-population (say, Men age 25-49 or something) by that sub-population).

That information comes from somewhere else (UN, World Bank, etc). Now you have weights that, within-country, distribute the population weight by probability of sampling, and across country factor in population differences. If the weights are representative of households and not individuals, the relevant population would be households in the country.

Note though: if you are using, say, West Africa as a region, Nigeria will basically swamp everything else. This may or may not be desirable, but you should look at your relative populations and decide whether or not you want one or two countries to dominate your estimates (you may want that, you may not).

3 - I know very little about the HIV testing, but if you are pooling people who were and were not tested, you'd get the wrong prevalence (because you don't know if the people in the other recodes are positive or not). Maybe I'm just worrying about nothing, but didn't immediately understand why you were pooling those to calculate HIV prevalence.

4/5 -you can do the svyset stuff mostly mechanically too. Binreg supports weights, so you can directly use those (pw=weight) and then it also supports clustering. You'd want something like:

```
binreg Y X [pw=weight], vce(cluster clustervar)
```

This won't get you the efficiency gains (read smaller standard errors) that accounting for stratification might get you, but that should be a small effect*. Note, when generating "clustervar" you want to do it so that each country has its own clusters (so cluster 14 in one country is different than cluster 14 from another, for instance). You might just generate a country/survey two-digit huber, and then gen clustervar = countrynum*1000 + clusternum - something like that.

Hope some of that helps.

*If you want to get fancy, you could probably bootstrap that using the stratification and drawing clusters with replacement, but my guess is the difference is going to be very small, and if you already have enough precision (are already statistically significant at acceptable levels) it might not be worth it, you can just say your SEs are "conservative".

Subject: Re: De-normalizing weights and svyset command in Stata
Posted by hannekeyserhegdahl on Fri, 16 Jan 2015 07:29:56 GMT

View Forum Message <> Reply to Message

Thank you, this was actually very helpful!

Now, if someone from DHS perhaps could give me some answers about the HIV-weight..?
Thanks in advance!

Subject: Re: De-normalizing weights and svyset command in Stata
Posted by Bridgette-DHS on Fri, 16 Jan 2015 16:24:34 GMT
View Forum Message <> Reply to Message

Following is a response from Senior DHS Stata Specialist, Tom PullumAny analysis using hiv03 (result of the HIV test) should be weighted with hiv05. Once you have merged with the AR file, you should ignore v005 or mv005 or hv005.

Your within-survey analyses are fine with the original hiv05 as the weight. Any adjustment to the weights related to pooling would be by a survey-specific multiplier and would have no effect on within-survey estimates.

My preferred way to handle the renumbering of clusters and strata in a pooled file is to use the "egen group" command. Within each survey, the cluster variable is always v001 (which is duplicated as v021). The stratum variable does not always have the same number and it is not always even named correctly. The strata are virtually always the combinations of region x v025 (v025 is urban/rural). I would find or construct that variable and then rename it as "strata", e.g. "gen strata=v022". You also need a unique identifier for "survey". You cannot rely on v000 for this, because v000 is a 3-character string such as "NG5", where "NG" is the country id and "5" is the phase of DHS. Sometimes there will be two surveys in the same phase, and v000 will be the same for both of them. (This is not an issue if you are using just one survey per country.)
Anyway, you will need a line such as

```
egen cluster_pooled=group(survey v001)
egen strata_pooled=group(survey strata)
```

and then you will have the unique identifiers.

To give equal weight to each survey, you need lines such as these FOR EACH SURVEY SEPARATELY:

```
scalar TOTWT=1000000
quietly summarize hiv05
scalar T=r(sum)
gen hiv05r=hiv05*TOTWT/T
```

You can do this adjustment before the pooling, or put those lines in a loop after the pooling, but just be sure that the recoding is survey-specific. These lines will remove the arbitrary factor of 1000000 from the original hiv05 and will give each survey an arbitrary TOTAL weight of 1000000. (That number could be anything you want.) This approach will give the same weight to every survey, regardless of the population of the country or the size of the sample. You need to make it very clear that your regional estimates were calculated that way. If, say, you wanted to weight each survey in proportion to its population size, you would replace "TOTWT" with the country's total population or the population age 15-49 or something like that.

I have not used "binreg" but have been using "glm..., family(binomial) link(log)" for many years. The two should be equivalent and I know glm works with svyset and svy. I'd be surprised if binreg doesn't, but if that's the case, you can switch to glm.

Let us know if any questions remain.

Subject: Re: De-normalizing weights and svyset command in Stata
Posted by Mercysh on Sat, 06 Jun 2015 11:33:43 GMT
View Forum Message <> Reply to Message

I am sorry I lost you in the 3rd paragraph, where Tom says that you can not rely on V000 because it is a string variable. What does one need to do if using two DHS waves/phases for three countries for example to have a unique identifier for "surveys" for each country and wave (one survey per phase). Secondly would a change in the regions need to be considered when pooling the data e.g. 2004 Malawi DHS was representative at selected districts (in addition to national) and the 2010 at all districts except that two were combined into one for survey purposes. I want to use svy prefix in Stata. Your help will be highly appreciated.

Subject: Re: De-normalizing weights and svyset command in Stata
Posted by Mercysh on Tue, 09 Jun 2015 14:10:42 GMT
View Forum Message <> Reply to Message

I think I got around part of the problem. I recoded shdist in Malawi 2010 data to be consistent with Malawi 2004 data but stuck with what to do with the "other districts" in Malawi 2004 data.

Subject: Re: De-normalizing weights and svyset command in Stata
Posted by DaniD on Mon, 09 Nov 2015 04:47:51 GMT
View Forum Message <> Reply to Message

Hi,

Thank you very much Tom for providing this code and explanation, it is extremely helpful! I am still uncertain as to whether the variables (cluster, strata, survey) need to have unique identifiers (cluster1, cluster2, etc.) for each survey included in the pooled (appended) data set. So whether it should be egen cluster_pooled= group(survey1 survey 2 cluster 1 cluster 2).

And then I am doing the weighting method you describe (scalar TOTWT=1000000 quietly summarize hiv05 scalar T=r(sum)
gen hiv05r=hiv05*TOTWT/T) for each survey before I append them. Is it correct that I can use the variable name hiv05r for each survey or does this need a unique identifier as well?

I know these are stupid questions but I am just having a hard time conceptualizing this process and it's difficult to check whether I am doing it correctly in stata.

Also, I asked in another thread and have not yet heard back but I am doing this with the couples data merged with the HIV data for multiple countries. I am assuming that I should be using all of the women's variables for this weighting process since the men were chosen on the basis of the women. Is this correct?

If someone could help with these lingering questions I would greatly appreciate it!

Cheers!

Dani

Subject: Re: De-normalizing weights and svyset command in Stata
Posted by DaniD on Mon, 09 Nov 2015 04:53:42 GMT
View Forum Message <> Reply to Message

Sorry my message got posted twice.

Subject: Re: De-normalizing weights and svyset command in Stata
Posted by Bridgette-DHS on Tue, 17 Nov 2015 15:20:32 GMT
View Forum Message <> Reply to Message

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

No problem--don't hesitate to ask questions. If you put several IR files (for example) into a single file, perhaps by appending them, you need to construct a variable called "survey" (for example) that is 1 for the first survey, 2 for the second survey, etc. Ideally you could just us v000 for this purpose, but there have been some instances of two successive surveys in the same country that have the same code for v000, so I always construct a new variable such as "survey". The command to construct the new cluster codes could be "egen cluster_all=group(survey v001)". Here v001 is the cluster variable (v021 is always exactly the same as v001) and there is no need to name it "cluster" but you can if you want. Both "survey" (after you construct it) and v001 are in all the surveys, and they do not need to have separate names within each survey--in fact that would defeat the purpose. You are right that hiv05r (or whatever you want to call it) would also be the same variable across the combined file. You do not need separate names within the individual surveys.

I can't tell what you mean by "should be using all of the women's variables for this weighting process" in your CR+AR merged file. The recommended weight in this file would be hiv05 for the men in the AR file. It would be applied to all runs using this file, including a run that, say, only happened to use variables from the women in the couples, because their probability of inclusion in this file is affected by the participation of the matched men in the HIV testing, and their participation (usually, at least) is more problematic than the participation of the women. No variables other than *v005 or hiv05 are ever involved in the calculation or selection of weights, so that phrase (in quotes) does not seem relevant.

Subject: Re: De-normalizing weights and svyset command in Stata
Posted by DaniD on Fri, 23 Jun 2017 18:25:23 GMT
View Forum Message <> Reply to Message

Thanks so much Bridgette and Tom for all your help!
Sorry, what I meant by "using all one gender's variables for the weighting process" was in

reference to creating the svyset. I am using the men's hiv weight for my data (CR merged with AR) so I was wondering if I should be using the men's variables for the cluster and strata identifiers as well or if it's fine to use the women's variables for that part.

To clarify, my understanding is that it shouldn't matter whether I use the men's or women's cluster and strata variables in the couple's file because the unit of analysis is the couple, is this correct?

Thanks!
Dani

Subject: Re: De-normalizing weights and svyset command in Stata
Posted by dnbhatta on Mon, 16 Oct 2017 21:36:03 GMT
View Forum Message <> Reply to Message

Hi,

Could you please send the reference document for de-normalizing weights (which you had mentioned, UN, World Bank, etc, my email: dnbhatta@yahoo.com), it would be a grate document for me. I have one question, after multiplied by the population. it will become a more than two or three digits, does it affect the results or shall we do some thing more to this.

Thank you.

Subject: Re: De-normalizing weights and svyset command in Stata
Posted by Bridgette-DHS on Tue, 17 Oct 2017 14:58:52 GMT
View Forum Message <> Reply to Message

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

There have been many postings about so-called de-normalization of the weights. I have nothing to add to them. Perhaps another user can help you. Regarding the last part of your question, substantive conclusions will not be affected by more than three significant digits, but if you want other people to be able to reproduce your estimates it will be best not to round the weights.
