Subject: Pooled data analysis from 25 countries
Posted by Mayank_Ag on Mon, 04 Feb 2019 12:10:48 GMT
View Forum Message <> Reply to Message

I am doing analysis after merging data from multiple waves of 25 countries. My regression specification includes variables that have been calculated at the psu level (employment, etc). My dependent variable is negative of HAZ. My unit of analysis are children below age of 5.

I have some questions regarding the sampling and weighting.

1. For certain countries cluster and PSU variables are not the same. In such cases which variable shall I use to create the otherwise PSU level variables.

2. Do I need to specify the strata variable while using svyset command in Stata. If yes, how do i deal with missing strata information.

3. I have already de-normalized the weights as suggested in earlier posts on this forum. Do I need to re-normalize the weights before using them? If yes, how shall I do it?

4. In one post on this forum I read that in multi country analysis data must be clustered at country level. Do I need to do that for this analysis? If yes, how do I cluster data at two different levels i.e., country level and then individual PSU level?

Subject: Re: Pooled data analysis from 25 countries
Posted by Bridgette-DHS on Wed, 06 Feb 2019 15:50:16 GMT
View Forum Message <> Reply to Message

Following is a response from Senior DHS Stata Specialist, Tom Pullum and Senior DHS Sampling Specialist, Mahmoud Elkasabi:

1. For certain countries cluster and PSU variables are not the same. In such cases which variable shall I use to create the otherwise PSU level variables.
In most of the countries, variables for cluster and PSU are the same. In cases where the two are different, you can use the cluster.

2. Do I need to specify the strata variable while using svyset command in Stata. If yes, how do i deal with missing strata information.
Yes, you can use HV022. In cases where HV022 is not consistent with the stratification described in sampling design in Appendix A of the final report, you can recode a new stratification variable. If you are using HV022, you should not expect to find missing strata information.

3. I have already de-normalized the weights as suggested in earlier posts on this forum. Do I need to re-normalize the weights before using them? If yes, how shall I do it?
I do not see a reason to re-normalize the de-normalized weights. This should not affect your regression.

4. In one post on this forum I read that in multi country analysis data must be clustered at country level. Do I need to do that for this analysis? If yes, how do I cluster data at two different levels i.e., country level and then individual PSU level?
I would recommend fixed effects for country or survey, rather than random effects. Just assign a survey code such as survey=1, 2, 3, etc. to each survey and include a term such as "i.survey" in the model specification.

Someone else could prefer random effects for surveys, especially if there are MANY surveys in the analysis. That would require a 3-level hierarchical model (for respondents / clusters / surveys).

When combining surveys, the strata and cluster codes but be unique. For example, you do not want cluster 1 in survey 1 to be confused with cluster 1 in survey 2. For example, you could have "egen cluster_id=group(survey hv021)" and "egen stratum_id=group(survey hv022)."

Subject: Re: Pooled data analysis from 25 countries
Posted by boyle014 on Wed, 06 Feb 2019 17:51:01 GMT
View Forum Message <> Reply to Message

Dear Mayank,

Thanks for the question. Tom and Mahmoud's answer works with the regular DHS. Since you also posted this query on the IPUMS DHS user forum as well, here's a similar response that uses the already-integrated data of IPUMS-DHS.

1. Can i directly use the idhspsu variable to create the requite variables for the surveys where cluster and PSU are not the same?

Yes.

2. Do i need to use idhsstrata variable while using svyset command?

Yes. svyset would still perform the weighted estimate you do not specify the strata, but the standard errors will be wrong.

To weight IPUMS-DHS data in Stata, the command is:

svyset [pw=perweight], psu(idhspsu) strata(idhsstrata)

This establishes the weights in Stata; they are then applied to relevant commands by putting

"svy:" at the beginning, such as:

svy: regress y x
svy: mean(y), over(x)

3. If yes, how do i deal with missing strata information?

This Forum has information on how to construct strata variables when they are missing. Fundamentally, it depends on the sampling design (which you can find in the appendices to the final reports). If the sample was stratified across urban/rural areas (typical), you can replace the strata variable (idhsstrata) with the urban/rural variable (urban).

4. Can i directly use the weight perweight for this analysis?

Yes.

5. In one post on this forum i read that in multi country analysis data must be clustered at country level. Do i need to do that for this analysis. If yes, how do i cluster data at two different levels i.e., country level and then individual psu level?

Whether it's necessary to cluster at the country level, the cluster level, or both depends on how much of the variation in your dependent variable is explained by these spatial variations. You can calculate this by running a null model, e.g.:

logit depvar [pweight = perweight] || idhspsu:
estat icc

If the rho is large (greater than 0.15 or so), then a mixed or multilevel model is appropriate. I've seen people cluster at the country, region, and psu level. These days, the psu level seems to be more common.

If the analysis combines only a few countries, then a dummy variable for each country except one is probably the best approach, and there would be no need to cluster at the country level. To cluster a multiple levels, here are the commands:

regress depvar [pweight = perweight] || idhspsu: || country:

Subject: Re: Pooled data analysis from 25 countries
Posted by Mayank_Ag on Wed, 06 Feb 2019 19:19:12 GMT
View Forum Message <> Reply to Message

Thanks a lot for the quick replies!!

I have some follow up questions. Sorry for the trouble but this is my first time with pooled data analysis.

1. My analysis has countries with widely varying no. of observations. For Ex:- Indian DHS 4 has around 200000 valid observations while some African country might only have around a few thousand observations. Is the weight perweight adjusted/applicable for such analysis? I read on some post in this forum that combining bigger surveys with smaller ones might give biased coefficients.

2. Regarding the clustering I got 2 different recommendations. As you can see Tom and Mahmoud suggested using a fixed effects model by adding a "i.survey" variable to my regression specification. On the other hand you have suggested a multilevel model. Can you please elaborate which model will be more suited for this analysis and how do i decide which one to use. I am sorry but i have never used multilevel models earlier.

Subject: Re: Pooled data analysis from 25 countries
Posted by boyle014 on Wed, 27 Feb 2019 14:06:00 GMT
View Forum Message <> Reply to Message

Dear Mayank,

DHS weights are not adjusted to take into account sample or population size in comparative analyses. There are separate discussions of when and how to do that in the User Forum though.

We are corresponding privately about your other questions, but I wanted to clarify my earlier post for other DHS users.

As to the specifics of your model, there are multiple ways to go. Including a survey fixed effect is the simplest approach and a good place to start. IPUMS DHS already includes a survey variable (SAMPLE), so you could use i.sample to do this with an IPUMS DHS data extract.

My post was providing the commands to do a multilevel model using IPUMS DHS variable names.

With respect to those commands, apologies--they included typos. The multilevel commands should have been "melogit" rather than "logit" and "mixed" rather than "regress". Further, one might try these without the weights in brackets first. Specifically:

melogit depvar || idhspsu:
estat icc

I hope your research is advancing well.
