Subject: Guidance Needed on Weighting for Pooled DHS Data in Logistic Regression
Posted by ygkim127 on Sat, 02 Nov 2024 12:49:54 GMT
View Forum Message <> Reply to Message

Hello,

I am reaching out with a question regarding data weighting. I am conducting research on "The Effect of Girls' Empowerment on Adolescent Pregnancy in Sub-Saharan Africa," aiming to investigate whether increased aged 15-19 girls' empowerment has a positive effect on reducing adolescent pregnancy rates in this region.

I plan to pool data from 27 Sub-Saharan African countries and will be using DHS-7 and DHS-8 data from the IR datasets of these countries. The explanatory variable will be women's empowerment, while the dependent variable will be the pregnancy status of adolescents aged 15-19. I intend to perform logistic regression analysis using Stata.

Since I need the overall pooled set weights that can represent Sub-Saharan Africa, I want to ensure that I am correctly calculating and applying these weights. I have read previous posts on this users forum and the "Note on DHS standard weight de-normalization" file, but I would appreciate your guidance to confirm my understanding and approach.

I have conducted weight de-normalization for each country using the formula: V005×(total females age 15-49 in the country at the time of the survey)/(number of women age 15-49 interviewed in the survey)
I have extracted data only for married women aged 15-19 from each country.
I have used the "append" function to pool the data from the 27 countries into one dataset.
I want to apply weights when conducting logistic regression analysis, but I am unsure how to do so.
Please let me know if there are any mistakes in the sequence of these steps or in the weight de-normalization process. Additionally, I would greatly appreciate the exact Stata code for applying weights in the pooled dataset and conducting the logistic regression analysis.

I look forward to your response.

Thank you.

Subject: Re: Guidance Needed on Weighting for Pooled DHS Data in Logistic Regression
Posted by Bridgette-DHS on Mon, 04 Nov 2024 16:47:38 GMT
View Forum Message <> Reply to Message

Following is a response from Senior DHS staff member, Tom Pullum:

What you have done so far looks fine to me. For the last step, actually applying the weights, please check previous posts on the "svyset" command. This will adjust for the clusters and strata

in the surveys, as well as the weights. As you will see in the earlier posts, you need to construct a new variable for the cluster ID, to distinguish between v001 in different surveys. For example, you can enter "egen cluster_ID=group(survey v001)". You will also need a new variable for the statum ID. We recently posted (again) a file that specifies the stratification variable in all the surveys. In recent surveys it is v022=v023 and in most older surveys it is v024 x v025, but there are exceptions, and they are given in that file.

In the analysis I would include a fixed effect for survey with "i.survey". A multi-level model with a random effect for survey is not justified, in my opinion, and it would add complexity.

I expect that your outcome variable has huge variation from one survey to another, as well as variation within most surveys. The whole concept of pooling surveys with this kind of an outcome seems to me to be unnecessary, but of course you can do what you want.

Subject: Re: Guidance Needed on Weighting for Pooled DHS Data in Logistic Regression
Posted by ygkim127 on Thu, 07 Nov 2024 16:31:09 GMT
View Forum Message <> Reply to Message

Thank you for your answer. I am following your advice, but could you please check if I am doing it correctly?

gen preweight = (v005/1000000)
gen preweight_denormalized = preweight * (total females age 15-49 in the country at the time of the survey)/(number of women age 15-49 interviewed in the survey)
gen survey = "country name"
egen cluster_ID=group(survey v001)
egen stratum_ID = group(survey v023)
svyset cluster_ID [pw=preweight_denormalized], strata(stratum_ID)

Please correct me if there are any mistakes.
Thank you.

Subject: Re: Guidance Needed on Weighting for Pooled DHS Data in Logistic Regression
Posted by Bridgette-DHS on Thu, 07 Nov 2024 18:17:15 GMT
View Forum Message <> Reply to Message

Following is a response from Senior DHS staff member, Tom Pullum:

This looks fine to me. But the statement " gen survey = "country name" " will not work for creating a string variable. You need something like " gen str30 survey = "country name" ". For example, " gen str30 survey = "Kenya" " . 30 characters will be enough for most country names....

Otherwise, looks good.

Subject: Re: Guidance Needed on Weighting for Pooled DHS Data in Logistic Regression
Posted by Hejie Wang on Sat, 09 Nov 2024 16:22:45 GMT
View Forum Message <> Reply to Message

gen survey = "country name"
egen cluster_ID=group(survey v001)
egen stratum_ID = group(survey v023)
Is this code correct because I see that it doesn't seem unique to see the same v001 from the same country in the data, for example, the BDKR7R file and the BDKR51FL file seem to be in different regions from the same v001. Would it make more sense to use gen survey="v000"

Subject: Re: Guidance Needed on Weighting for Pooled DHS Data in Logistic Regression
Posted by Bridgette-DHS on Mon, 11 Nov 2024 19:08:38 GMT
View Forum Message <> Reply to Message

All of your questions have been asked and answered earlier on the forum.

Subject: Re: Guidance Needed on Weighting for Pooled DHS Data in Logistic Regression
Posted by Hejie Wang on Tue, 12 Nov 2024 11:17:05 GMT
View Forum Message <> Reply to Message

Yes, it has been mentioned in the previous content, but I still have doubts. After merging documents from different countries, do I consider both first-level weights and second-level weights in the analysis? Also, would it not be better to randomize surveys (v000) and Psus conducted in different countries and at different times, since the association in the same country will also change at different times

---

Subject: Re: Guidance Needed on Weighting for Pooled DHS Data in Logistic Regression
Posted by Bridgette-DHS on Tue, 12 Nov 2024 17:10:16 GMT
View Forum Message <> Reply to Message

Following is a response from Senior DHS staff member, Tom Pullum:

You are jumping into a very complex analysis. I recommend that you pick ONE survey and develop a statistical model for just that one survey. See whether you find any significant relationships. You can then run that model repeatedly on other surveys.

Recently I worked with colleagues to analyze the anemia data in 9 countries, one country at a time. I will paste the citation and link below. We found virtually nothing. If we had pooled the surveys we would have found even less.

Users are attracted to multi-country pooled analyses because they seem sophisticated and state-of-the-art. The fact is that they are hardly ever justified. They tell you less, rather than more, about the outcome variables. DHS staff cannot provide further support for your unnecessarily complicated project.

Benedict, Rukundo K., Thomas W. Pullum, Sara Riese, and Erin Milner. 2024. Is child anemia associated with early childhood development? A cross-sectional analysis of nine Demographic and Health Surveys. PLOS ONE 19(2): e0298967. https://doi.org/10.1371/journal.pone.0298967. https://journals.plos.org/plosone/article?id=10.1371/journal .pone.0298967

Subject: Re: Guidance Needed on Weighting for Pooled DHS Data in Logistic Regression
Posted by Hejie Wang on Fri, 29 Nov 2024 18:43:58 GMT
View Forum Message <> Reply to Message

I have read previous posts on this users forum and the "Note on DHS standard weight de-normalization" file, and some questions have been solved. But some analysis questions still happen. Because I mainly use the statistical software R for analysis, I would like to know how to use the survey package for mixed-effect modeling after de-normalizing the weights of KR merge files for different countries, and what is recommended as a random term (country, v000, PSU, or region)

Subject: Re: Guidance Needed on Weighting for Pooled DHS Data in Logistic Regression
Posted by Bridgette-DHS on Mon, 02 Dec 2024 13:08:08 GMT
View Forum Message <> Reply to Message

Following is a response from Senior DHS staff member, Tom Pullum:

DHS staff have responded to dozens of forum posts on the topic of analyzing pooled surveys. We have nothing to add on this topic. Forum users are certainly free to post additional questions and responses but DHS staff (at least I, Tom Pullum, as of December 1, 2024) are finished with it.
