Subject: Weighting for the Pakistan Special 2019 DHS on Stata
Posted by kgeorg7 on Sat, 13 Aug 2022 16:04:50 GMT
View Forum Message <> Reply to Message

Hi, I am working with the Pakistan Special 2019 DHS in Stata, I've combined the individual, household and services datasets with household and services data being applied to the individual level.

I am currently trying to do the weighting for the dataset. From what I understand from the final report, the Primary Sampling Unit (PSU) is the household (qhnumber) with the weight variable being qweight. For the strata variable, the final report states that 16 sampling strata were created with urban/rural for each of the eight regions. However, there is no specific strata variable, as such I was wondering if I should be creating a new strata variable to account for the 16 strata? Furthermore, would also appreciate feedback regarding my understanding of the PSU and weighting variables as well as whether my approach is appropriate considering I've combined the individual, household, and services dataset.

The code I'll be using: svyset qhnumber [pweight = qweight], strata(newstratavar) vce(linear) singleunit(center)

Any help will be appreciated, thank you!

Subject: Re: Weighting for the Pakistan Special 2019 DHS on Stata
Posted by Bridgette-DHS on Mon, 15 Aug 2022 20:20:18 GMT
View Forum Message <> Reply to Message

Following is a response from DHS Research & Data Analysis Director, Tom Pullum:

Your svyset command is correct (except that we usually omit "vce(linear)"). Unfortunately, the raw data files from this survey do not include an identifier for the strata. In standard recode files, the stratum variable is hv022 or v022 (which is usually repeated with hv023 or v023). Here is a comment from Ruilin Ren, the DHS sampling expert who worked on this survey. He refers to hv022/v022, which are NOT in the raw data (qregion and qtype are in the data), but his description of how the strata are constructed is valid. He is referring to a "recode data file" that was used for the preparation of the report but otherwise is not available.

"There is a stratification variable in the recode data file, HV022/V022 with 14 codes. Actually the sample was designed for 8 provinces stratified by urban rural which totals the number of strata to 16. Punjab and Islamabad were combined together, KPK and FATA were combined together to form survey domains, GB and AJK were stand alone. But the Pakistan NOS selected the sample and did not provide an identifier for FATA, which reduced the number of indefinable provinces in the data file to be 7, KPK and FATA were all together and labeled as KPK, so we can only code 14 strata. That is why HV022 and V002 have 14 codes. This is not a problem for any purpose of use because FATA was indeed a very small part of KPK before. A key point for Pakistan surveys is the fact that they made GB and AJK stand alone, any attempt to combine different surveys together, or to analysis the full data together may introduce serious bias (bias toward the two

small provinces GB and AJK) without a proper treatment of the sampling weights. This is an important advice for all data users using Pakistan survey data."

Subject: Re: Weighting for the Pakistan Special 2019 DHS on Stata
Posted by kgeorg7 on Sat, 27 Aug 2022 21:19:35 GMT
View Forum Message <> Reply to Message

Hi,

Thank you for your response, however, I just wanted to confirm two things.

First, the variable qregion has 6 provinces and these provinces were then stratified into urban/rural (qtype) which results in a total of 12 strata. Just wanted to confirm that there are only 12 strata and not 14 as previously stated since 2 of the total 8 provinces were collapsed into other provinces.

Second, regarding the bias from combining different surveys together, was this in reference to combining this 2019 DHS with those from other years or something else?

Thanks again!

Subject: Re: Weighting for the Pakistan Special 2019 DHS on Stata
Posted by Bridgette-DHS on Tue, 30 Aug 2022 15:20:12 GMT
View Forum Message <> Reply to Message

Following is a response from DHS Research & Data Analysis Director, Tom Pullum:

For your first item, the answer is "yes".

For the second item, there is a bias in the national estimate for 2019 if you include GB and AJK in a national estimate because they were over-sampled. You can get an unbiased estimate for Pakistan MINUS those areas, and an unbiased estimate for those areas, but you cannot combine them for a national estimate. (That is the position of the implementing agency in Pakistan and our samplers; I cannot explain it further.) If comparing successive surveys in Pakistan, you should work with the same geographic areas. However, as I understand it, the population of those areas is a very small part of the national population. There have been other countries, such as Mali, in which small areas (in terms of population, not necessarily in geographic size) have dropped or included in successive surveys based on the security situation at the time of the survey. It's unfortunate but unavoidable.

Subject: Re: Weighting for the Pakistan Special 2019 DHS on Stata
Posted by kgeorg7 on Fri, 23 Sep 2022 20:36:23 GMT
View Forum Message <> Reply to Message

Thank you for your help!

Subject: Re: Weighting for the Pakistan Special 2019 DHS on Stata
Posted by kgeorg7 on Wed, 23 Nov 2022 18:44:58 GMT
View Forum Message <> Reply to Message

Hi again,

I was conducting a final check on my code before finalizing my manuscript which is when I noticed that the primary sampling unit for the Pakistan Special 2019 DHS would be qhclust rather than qhnumber. Page 111 (Appendix A.1) of the Final Report states that 1396 primary sampling units were randomly selected which is the number that corresponds with the qhclust variable.

So, I just wanted to confirm that the survey set code would use qhclust instead of qhnumber, and thus be:
svyset qhclust [pweight = weight], strata(strata) singleunit(center)

Thank you!

Subject: Re: Weighting for the Pakistan Special 2019 DHS on Stata
Posted by Bridgette-DHS on Fri, 25 Nov 2022 12:50:06 GMT
View Forum Message <> Reply to Message

Following is a response from DHS Research & Data Analysis Director, Tom Pullum:

Yes, qhclust is the cluster ID code and is the primary sampling unit. qhnumber is the household number. Households are numbered within clusters. The variable for the sampling weight is qweight. The strata were the combinations of urban/rural and region. There does not appear to be a variable for strata in the file. You can construct svyset as follows:

egen qstrata=group(qregion qtype)
svyset qhclust [pweight = qweight], strata(qstrata) singleunit(centered)

There have been other postings on this survey, which was "Special" and does not have the usual variable names, although it had the usual kind of sampling design.

Subject: Re: Weighting for the Pakistan Special 2019 DHS on Stata
Posted by kgeorg7 on Fri, 25 Nov 2022 22:34:34 GMT
View Forum Message <> Reply to Message

Great, thank you very much for the confirmation and the code!