Subject: Sample weights and stratification - Nigeria 2008 and 2018
Posted by Goethe2014 on Mon, 11 May 2020 13:48:16 GMT
View Forum Message <> Reply to Message

Dear all,

Currently I am using DHS data in combination with Stata for the first time. I intend to estimate effects employing a Difference-in-Difference estimation on Nigerian DHS data from 2008 and 2018 (Individual/women recode). In this regard I would like to know more about the right way to weigh the data and account for the stratification process.

In literature I found that some scholars combine (append) two sets of data (DHS Year A and DHS Year B) and when running their regression account for the women's sampling weight by just including [pweight=v005]. As far as I understood from the DHS forum and manuals in this case we dont have to divide the sample weight by 1.000.000 as pweight can also handle it without doing so. My question now is whether it is that easy to just use the pweight command on the full/combined dataset as there are women from two distinct survey included whose sampling weight had been calculated for their original dataset (Year A OR Year B). Do I therefore have to reweigh the sample or is it really possible just to make use of [pweight=v005] as the data stems from different women and different year but the same country?

In addition I am also a bit confused whether I have to account for the stratification process which in the case of Nigeria was done by states and rural/urban. Some literature accounts for that fact, others ignores the stratification process.

Lastly, I struggle whether I have to make use of the svyset command at all when using DHS data. Again some literature just specifies the data as panel data using xtset command while others suggest svyset commands to account for the DHS survey characteristic.

In a paper which asks similar research questions, DHS data from two years from the same country is used and the authors also employ a Diff-in-Diff estimation. First, they define the data as panel data by using xtset command and then already run their regression model only including [pweight=v005] and vce(cluster v001) at the end.

I would really appreciate any help in order to generate the most robust results and understand DHS data better in general.
Greetings