Subject: weighting issues of multilevel modelling using the DHS survey data with multiple-stage sampling
Posted by YUJP on Mon, 03 Sep 2018 09:19:59 GMT
View Forum Message <> Reply to Message

Dear DHS expert,
I am reading extensively the historical and current forum discussion on the weighting issues of multilevel modelling using the DHS survey data with multiple-stage sampling. I would appreciate if you can help to enlighten me on the following question:

Basically, I am using a multilevel model t analysis dataset from DHS Cambodia 2014 with the outcome of the children under five diarrhoea and predictors at both the level of the children as well as the level of cluster (PSU). I am using the "melogit" command for this analysis (same results can be produced using the "meglm" command. I plan to use the scaling methods (methods A or B) as proposed by by Sophia Rabe-Hesketh (2006) (http://www.gllamm.org/JRSSAsurvey_06.pdf) and Adam C Carle (2009) (https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-9-49). One problem is that in the DHS database we only have the weight (v005 or hv005) that has taking the two stage sampling (cluster (PSU and women (or household) into consideration. As it was stated in the STATAMULTILEVEL MIXEDEFFECTS REFERENCEMANUAL RELEASE 15 (page 104) (https://www.stata.com/manuals/me.pdf), we don't have Wj or W i|j but only Wij:

"Now take these same data and fit a two-level model with meglm, it is not sufficient to use the single sampling weight wij , because weights enter the log likelihood at both the group level and the individual level. Instead, what is required for a two-level model under this sampling design is wj , the inverse of the probability that group j is selected in the first stage, and wi|j , the inverse of the probability that individual i from group j is selected at the second stage conditional on group j already being selected. You cannot use wij without making any assumptions about wj .

Given the rules of conditional probability, wij = wj wi|j . If your dataset has only wij , then you will need to either assume equal probability sampling at the first stage (wj = 1 for all j) or find some way to recover wj from other variables in your data; see Rabe-Hesketh and Skrondal (2006) and the references therein for some suggestions on how to do this, but realize that there is little yet known about how well these approximations perform in practice.

What you really need to fit your two-level model are data that contain wj in addition to either wij or wi|j . If you have wij--that is, the unconditional inclusion weight for observation i; j--then you need to divide wij by wj to obtain wi|j ."

However, when I re-read the DHS report of Cambodia, I found that there are actually information on the distribution of enumeration areas in the sampling by strata. (page 282 Appendix A Table A2, Cambodia Demographic and Health Survey 2014: https://dhsprogram.com/pubs/pdf/fr312/fr312.pdf) . If I call them Cj (j= strata 1, 2, ... 38)), as we can easily get the number of selected clusters per each strata, which I call them CSj (j= strata 1, 2, ... 38)), it seems that I would be able to calculate the probability that the clusters in each strata were selected (CSj/Cj) and thus the weight Wj = Cj/CSj). With Wj, when I can calculate the wi|j which is Wij/Wj.

I use the methods and the information in the Appendix of the report and recalculated the scaled weights and got a results which is a bit different from (but still very similar with) the results that was produced by using the wij (v005) and presume that the second level weight to be "1".

I would appreciate if you can guide me whether this is a valid solution to obtain the two level weights for the multilevel analysis using DHS data? Or at least this can provide a better (less biased) estimate of the parameters than the one using wij as the first level weight and presume the second level weight be "1"?
Many thanks in advance.

Subject: Re: weighting issues of multilevel modelling using the DHS survey data with multiple-stage sampling
Posted by Bridgette-DHS on Wed, 05 Sep 2018 12:13:52 GMT
View Forum Message <> Reply to Message

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

I believe there is a problem with your approach, because it does not take into account that a cluster's probability of selection is not a constant within each stratum, but is proportional to the size of the cluster. Without knowing the size of the cluster, you cannot calculate its probability of selection.

We at DHS hope to prepare guidelines for dealing with the requirement for separate weights for multi-level models. I don't think the results will be very sensitive to the allocation (as your comparison suggests) but we do need something better than just wij and 1. Your approach is worth pursuing but I think the best that can possibly come out of it will be a better approximation, because we simply don't know the size of the cluster.

Subject: Re: weighting issues of multilevel modelling using the DHS survey data with multiple-stage sampling
Posted by YUJP on Wed, 05 Sep 2018 18:32:47 GMT
View Forum Message <> Reply to Message

I really appreciate the attention to my question and this swift response! I will us the strategy to presume the children (women) level weight be "1" and use v005 as the cluster level weight. However, while I am also thinking I must have made it wrong, I still want to ask: on page 282 Appendix A Table A2, Cambodia Demographic and Health Survey 2014 (I am attaching it), in my (possibly wrong) understanding, they have listed the number of enumeration areaas (EAs) (which is used as the PSU in Cambodia) in the sampling framework, by domain and the residence type (denominator). As the number of selected clusters per each strata (numerator) can be easily calculated from the data sets, is numerator/denominator not the probability for each strata? Any further opinion to enlighten me would be greatly appreciated. Thanks!

File Attachments

1) Table A 2 DHS Cambodia 2014 report.pdf, downloaded 483 times

Subject: Re: weighting issues of multilevel modelling using the DHS survey data with multiple-stage sampling
Posted by Bridgette-DHS on Wed, 05 Sep 2018 23:38:14 GMT
View Forum Message <> Reply to Message

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

The probability of selection is not the same for every cluster in a stratum. If it were, then your calculation would be correct. The probability that a specific cluster will be selected is conditional on the size of the cluster. In sampling with probability proportional to size (pps), a cluster with twice as many households in the sampling frame will have twice the probability of being selected. That's the piece of information that we are missing.

Subject: Re: weighting issues of multilevel modelling using the DHS survey data with multiple-stage sampling
Posted by YUJP on Thu, 06 Sep 2018 13:00:42 GMT
View Forum Message <> Reply to Message

Dear Tom, Many thanks! Now I understand why it is not sufficient to based on the table to calculate the probability of the selection of the clusters. Best wishes Junping