Subject: What is the DHS position on the use of cluster-level data?
Posted by ld190 on Sun, 27 Mar 2016 16:27:36 GMT
View Forum Message <> Reply to Message

Dear All,

Apologies in advance for the long response and thank you for reading it. I've tried to be thorough in order to bypass old material and to get to what I see as the key ambiguities (for me) remaining in the question of how, and whether, cluster-level data from the DHS should be analyzed.

I have come a point in my research that am starting to use DHS data in earnest. I have downloaded and conducted some preliminary analysis of datasets from Senegal, and some of the things that I would like to do involve aggregating from individual to "cluster-level" characteristics. So for example, calculating that X percent of Women 15-49 sampled in cluster Y have attribute A and Z percent have B attribute. I am interested in clusters because my research is about residential communities - small areas where people are co resident and have some chance of knowing or being influenced by each other. I am able to do this using SPSS and aggregating cases by Cluster. This analysis produces some very interesting (and theoretically plausible) results. However, I am concerned about warnings that I've read, against the disaggregation of DHS data.

Despite attempting to work through the DHS's very helpful store of literature, and this forum, I remain unsure about the DHS' position on the direct use of cluster data. On the one hand, the official guides, and very enjoyable YouTube tutorials, seem to me to emphasise that the surveys are designed to be representative at the Regional and National levels only, meaning that further disaggregation is not possible. However, I'm not sure about the extent to which this applies to my research. I am not interested in estimating the prevalence of attribute A (which is very common) for any area except for the cluster (the Enumeration Area) itself. So I'm not interested in the surrounding administrative area or some other geographic area, for example. I'm just interested in the cluster of households from which the chosen households were randomly sampled. Considering this level of analysis, around 20 households sampled at random from a pool of, on average, 110 households, is the data so unrepresentative as to be useless? Do the observed attributes of the randomly sampled households (20) tell us nothing reliable about the attributes of the overall population (110)? What about if we average across a large number of clusters, to produce a distribution of values?

The guidance on this issue on the forum appears to me to provide a number of alternative possible answers and issues to consider.

One user on the forum seems to suggest that the use of cluster-level data is "noisy" (error prone) but basically OK as long is this is taken into account and that it is common to use this level of data for certain purposes:

http://userforum.dhsprogram.com/index.php?t=msg&goto=905
4&S=41b1f8e9c6ffff1e5ed1b91414054772&srch=aggregatin g+clusters#msg_9054

However, DHS staff member Trevor, on another post suggests that the use of cluster-level estimates are "impossible" because the sample sizes are too small. Although he is referring to

calculating Child Mortality rates, which is a very rare event, and so this measure might require an especially large sample size.

http://userforum.dhsprogram.com/index.php?t=msg&goto=852
4&S=41b1f8e9c6ffff1e5ed1b91414054772&srch=aggregatin g+clusters#msg_8524

On another post, Trevor says this:

Quote:You can and should still use hv005 as the sample weight, but doing your analysis with smaller geographic units is potentially problematic. The sample is designed to be representative at the region level, but not at the level of smaller units. As you disaggregate the data to smaller units the sample is less and less likely to be representative. The sample is also designed to provide a certain level of accuracy at the region level, and again as you disaggregate to smaller units the accuracy of those estimates gets worse and worse and the confidence intervals around the estimates quickly become very large and unreliable.

I found this advice slightly confusing. Presumably going from drawing inferences about the population at an officially representative level (region), to an intermediate level (like a small administrative unit) might reduce the representativeness of the data. This I because the size of the sample (N households) might be getting smaller relative to the size of target population (a whole administrative district). However, presumably at some point this trend will reverse? If we only tried to draw inferences about the Enumeration Area from which the sample is drawn, for example, then surely this is more representative than trying to use the cluster sample to draw inferences about, for example, a larger population within 5km2 of the Enumeration Area?

ClaraB, also a DHS staff member, offers this advice on the interpretation of cluster-samples:

http://userforum.dhsprogram.com/index.php?t=msg&goto=831
5&S=41b1f8e9c6ffff1e5ed1b91414054772&srch=aggregatin g+clusters#msg_8315

Quote:[inference about the] district location of the sampled clusters using a GIS software and the GPS dataset these data would not be statistically representative.

However, I'm unclear how to interpret this advice. Is the warning given because the user is trying to draw inferences about the district level (larger than the EA) from a single sample cluster?

Finally, a forum user posted this advice about the use of cluster-level measures:

Quote:cluster-level measurements are based on too few observations to be meaningful in and of themselves - as you say, there are wildly under-powered. A couple of things you could do: a) by averaging over many clusters, you can still get good estimates of community level variables, but each individual cluster-level point-estimate would be very, very noisy. But they may still mostly "agree" in some sense; b) so if in your hierarchical model you allow each cluster an unconstrained cluster-specific effect (like treating each cluster as a mini-experiment), you could look at those individual point-estimates on a scatter plot (say Beta across some variable you think would affect Beta); c) and then you could start restricting those Betas to have some particular distribution (a random slope model) and see how that changes your overall point estimate as you make your priors on the distribution of Beta more/less informative. I think this makes sense as a kind of

model-checking or informal/additional inference procedure. A leave-one-out cross-validation approach might make sense too, depending on how you end up thinking about each of these within-cluster estimates.

This user's scatter-plot suggestion is very close to what I have done in my own research.

I addition to searching the DHS forums, I've discovered that some published academic work has engaged with data at the cluster-level. Storey and Kaggwa from the Department of Population, Johns Hopkins University, have used cluster level data from the 1995, 2000 and 2005 Egypt Demographic and Health Surveys (EDHS).
This is a quote from the abstract for their paper:

Quote:Norms are defined at the cluster level, which serves as our community-level unit of analysis

The official site for the article is here:

http://muse.jhu.edu/journals/population_review/v048/48.1.sto rey.html

Also there has also been some research to actually estimate the error introduced from using cluster-level measures with DHS data. This was conducted by Øystein Kravdal, Professor of Demography at the University of Oslo.

Here is a quote from the abstract for his paper:

Quote:For example, researchers may consider including in their models the average education within the sample (cluster) of approximately 25 women interviewed in each primary sampling unit (PSU). However, this is only a proxy for the theoretically more interesting average among all women in the PSU, and, in principle, the estimated effect of the sample mean may differ markedly from the effect of the latter variable. Fortunately, simulation experiments show that the bias actually is fairly small - less than 14% - when education effects on first birth timing are estimated from DHS surveys in sub-Saharan Africa. If other data are used, or if the focus is turned to other independent variables than education, the bias may, of course, be very different. In some situations, it may be even smaller; in others, it may be unacceptably large. That depends on the size of the clusters, and on how the independent variables are distributed within and across communities. Some general advice is provided.

This paper is available to read, published in a Peer Reviewed Open Source Journal:

http://www.demographic-research.org/volumes/vol15/1/

Both of these papers, seem favorable to the use of cluster-level DHS data.

I wonder if the 'proof of the pudding is in the eating'? The results from my analysis of community level data are theoretically plausible, there is a clear pattern (agreement) in a scatter plot showing the relationship between two measures (the frequency of observations A and B in each cluster) across all the clusters and this pattern is consistent accross Senegalese DHS surveys in 2005, 2010 and 2014. Presumably, if the level of noise were so great that no meaningful information could be gained from cluster-level analysis, then a clear pattern of results like this would be quite

surprising?

Thank you again for reading through this long question. I am by no-means certain about any of this, I am new to this area of analysis and this kind of analysis. However I wanted to provide a detailed description of the problem that I am trying to grapple with.

If anyone can offer any further thoughts, clarification or advice on the use of cluster-level analysis with DHS data, I would be very grateful to hear it. Also, if there is some key DHS document (or other publication) that I have missed which elaborates on this issue would be grateful to receive a recommendation.

Many thanks in advance for your response.

Laurence.

P.s Thanks to UserRHS for the help in improving the formatting of this post.
