Subject: Re: What is the DHS position on the use of cluster-level data?
Posted by ld190 on Fri, 01 Apr 2016 10:47:13 GMT
View Forum Message <> Reply to Message

Dear user-rhs,

First, many thanks for your detailed and very helpful response. I really appreciate the time-taken.

To address the ambiguity you noticed in my post: As it happens I'm actually not using regression models. I am using an agent-based model, which is a kind of computer simulation. In this case the cluster-level measures that I am interested in are for purposes of model calibration and validation (rather than "fitting" as in regression models). Cluster level measure X will be used to "calibrate" the simulation - which means it will be used to set the value of a key parameter in the simulation. Then measure Y will be used to validate (i.e. test) the simulation by observing whether the simulated relationship between input variable X, and the output of the simulation, are the same as the relationships which exists in the real data between measures X and Y. Hence, scatter plots showing the relationship between simulation-input X and the subsequent simulated-output (which should correspond to Y) can be overlayed with the real cluster-level X and Y values - as an indication of the match between the simulation and the real clusters. I am interested in cluster-level measures because the model is a flexible model of social dynamics within a cluster (a village-sized residential community with a social network - etc.).

Having said that, the advice about regression is very useful for future reference - thank-you.

Having thought about what you've written, and about the recommendations of the Kravdal paper, I am now cautiously optimistic. It seems that the analysis is worth pursuing for the moment. Kravdal and you cite the importance of the absolute size of cluster samples (as well as the relative cluster population/ sample ratio). Having looked at the dataset from Senegal 2005, the cluster sizes are quite large (M = 38, SD = 11) and only 4.8% of clusters are below 20 cases. However, based on Kravdal's advice it will be important to check the within- and between- cluster variance of both measures. I'll also have to consider the average size of the cluster population itself. It may even be worth my creating a replication of something similar to Kravdal's simulation - in order to explore the viability of using this particular data-set in this way.

As and when I do a more in-depth investigation of the viability of using clusters for this purpose I will post about it here for the interest of future users of these particular measures.

Also thanks for the advice regarding a scientific justification for such choices. I agree that if a good justification can be found for the use of these measures and as long as one is open and honest about their limitations and drawbacks, there is no obligation to neglect their use.

Best,

Laurence.