Subject: Clustering
Posted by ahmed89o on Wed, 25 May 2016 13:17:37 GMT
View Forum Message <> Reply to Message

Hello DHS user,
My key regressor is at the city level and my outcome at the individual level. I cluster my standard errors to the city level. Is it better to take into account the cluster effect of the DHS variable v001 made by dhs as well and how to do it stata? How to take into account the impact of the two clusters?

Subject: Re: Clustering
Posted by ahmed89o on Wed, 25 May 2016 16:15:50 GMT
View Forum Message <> Reply to Message

To clarify more the survey design is the following: strata based on urban or rural. the cluster based on district or village level and then household to women in the household. I study the impact of city level variable, which is something between the strata and cluster (bigger than cluster but smaller than strata), on individual level variable. I do not use multilevel modeling but the standard logistic model with standard errors clustered at the city level. So my question should I cluster for district and village (the original cluster) too. Ahmed Rashad

Subject: Re: Clustering
Posted by Reduced-For(u)m on Sun, 29 May 2016 21:53:06 GMT
View Forum Message <> Reply to Message

I think the general rule-of-thumb here is to cluster at whatever is larger: the level of your aggregate variable (city) or the level you need to account for the sampling design (PSU). It sounds like your "city" variable can include multiple PSUs, at which point I would suggest clustering at the city level. You do not need to cluster up to the level of the strata.

I don't know what you mean by "account for the impact of two clusters". Obviously, you need many clusters in order to "cluster" your standard error estimates (many like more than 40 or something, at least more than 15 or 20). Do you mean what if city includes more than one "cluster"? In that case, the answer is above: you want to cluster at the city level.

This way of thinking about it might help: if everyone in a city has the same value for your variable of interest, then are you getting N=# of observations amount of information, or C=# of cities worth of information? Probably we want to think "something in between" but it is probably closer to C than N... you are getting much less information from additional observations all with the same value of the right hand side variable than you would by getting a whole new city. Additional observations within a city (that already has many observations) gives you a little bit more information, but not a full new observation's worth. At the far extreme, you could imagine collapsing everything down to the city level and running your regression on those C

observations... each new piece of information within a city would just make each of those C observations slightly less variable (a slightly better estimate of the outcome for that city).

Subject: Re: Clustering
Posted by Faiza on Tue, 29 Oct 2019 11:01:33 GMT
View Forum Message <> Reply to Message

Dear DHS user
Can you please define the cluster number variable available in DHS.I used it in my multilevel logistic regression model.But my professor wasn't satisfied.He asked to define the basis of cluster formation.On what grounds clusters were generated?