Subject: Unique identifier in pooled DHS (countries and years)
Posted by marion on Tue, 10 Dec 2013 13:40:45 GMT

View Forum Message <> Reply to Message

Dear all,

I pooled the birth history datasets for all phases and all countries. My total sample consists of roughly 6,000,000 observations. I created a unique identifier by combining v000+v001+v002+v003+bidx, as it is recommended in the DHS manuals. However this combination does not uniquely identify each line (=each child). There are more than 1,500,000 observations which are not uniquely identified by this combination of variable. My guess is that is problem could be connected to the presents of different waves within the phases (e.g. for Uganda phase 5 there are the datasets UGBR5HDT and UGBR52DT). Do I have to assign a coding for the different waves within the phases? I did not read anything about this requirement in the manuals so I am not sure how to proceed best. Do you have any advises?

Many thanks and warm regards,
Marion

Subject: Re: Unique identifier in pooled DHS (countries and years)
Posted by Reduced-For(u)m on Wed, 11 Dec 2013 19:41:27 GMT

View Forum Message <> Reply to Message

Hi Marion,

   I don't have a great answer for how to use those variables to create a unique id number, but if all you want is a unique id (and don't want to use part of that ID to somehow group the obs by, say, country and round) then one easy way would just be:

gen id = _n

Then you'll get unique IDs of 1-N. If you have to match this to other data somehow, you could always sort it:

sort country wave something
gen id = _n

Or, if you want to check your intuition on what isn't unique, you could create a number of each "wave" within each "phase" and then add .01*wave (or 10000000*wave, however many 0s you need) to each identifier, and see if you have unique ID's.

Now, if there is some reason you need some more specific kind of ID number, then these might not work and you might need to wait for a DHS person to help.