Subject: Re: Explanation of IPUMS-DHS
Posted by gange026 on Tue, 31 Jan 2017 19:51:14 GMT
View Forum Message <> Reply to Message

Merging between the original DHS files and IPUMS-DHS data extracts will become less necessary as we add more variables to IPUMS-DHS. We expect to double the number of IPUMS-DHS variables in our April 2017 data release. Below is some guidance on linking between an IPUMS-DHS extract file and the original DHS files, using the latest Nigerian sample as an example.

Merging IPUMS-DHS data with original DHS data requires all data files to have a linking key, a unique identifier that is identical in name and in character length. To merge individual-level data, the original DHS files need to have a variable called "IDHSPID" as the linking key. This is a unique identifier for each respondent. In IPUMS-DHS, IDHSPID is a concatenation of SAMPLE (a 4-digit number representing the country and year of the survey) and CASEID, which is a sample-specific unique identifier for the respondent.

To create IDHSPID, CASEID is assumed to be right-justified. This means that there are leading blanks in the data that cause CASEID to occupy the full variable width in the data, even if there are not numbers or letters to fill it. If you look at the data browser for the original 2003 or 2008 Nigerian DHS files, you should be able to see that there are a few spaces in every line of CASEID before the numbers begin. In the 2013 Nigeria children's recode, CASEID is not right-justified, and these spaces are not there.

An easy way to change this is to run the following command in Stata in your file for Nigeria 2013:
replace caseid=substr(" ", 1, 15 -length(caseid)) + caseid
There should be 15 spaces/blanks in the above quotation marks. If you then generate IDHSPID, there should be the appropriate number of spaces between the sample identifier and the CASEID.

A couple of other notes that may help you as you get further along in the merging process:

1. You will need to make sure IDHSPID is the same length in both the original DHS and IPUMS-DHS files. IPUMS-DHS stores IDHSPID in a 22-column string variable (str22), and you will want to make sure that when you create IDHSPID, that also has a width of 22. To do this you can add spaces at the beginning of the variable when you create it:
gen idhspid = " "+string(sample) + caseid
There should be 3 spaces/blanks in the quotation marks above to make up for differences in the width of CASEID in the original DHS data (15 columns) and IPUMS-DHS data (18 columns). The above code also assumes SAMPLE is a number rather than a string. If your version of SAMPLE is a string, you can simply run:
gen idhspid = " "+sample+caseid

2. If you are merging the original child recode dataset with children's data from IPUMS-DHS, you should keep in mind that IDHSPID is a unique identifier for the respondent, but not necessarily for children; children with the same mother will have the same value for IDHSPID. If you want to merge children's data with children's data, you should generate another variable that is a concatenation of IDHSPID and BIDX (Child's birth history index number) in both your original DHS

data and IPUMS-DHS data. This will be a unique identifier for each child in the data. For example:
gen idhspidk = idhspid + string(bidx)
When you merge, this is the variable you should use as your linking key:
merge 1:1 idhspidk using [filename]
If the IPUMS-DHS data you are using is from the women's file, you do not need to create this additional variable, and you should merge your files using IDHSPID as your linking key.

I have also attached a sample do-file for Stata, that goes through every step of the merging process with data from the children's recode files from Nigeria 2003, 2008, and 2013, and an IPUMS-DHS extract using children as the unit of analysis.

Please let us know if you have any more questions.

-IPUMS-DHS staff

File Attachments

1) merging_instructions_ng.do, downloaded 1058 times