
Subject: Missing values for anemia

Posted by [clarapd](#) on Fri, 19 Feb 2016 21:36:28 GMT

[View Forum Message](#) <> [Reply to Message](#)

Hi everyone, I'm new in this forum.

I'm trying to use the prevalence of anemia among women of reproductive age in several Sub-Saharan countries in a regression.

First, I merged the individual dataset with the household dataset with v001 and v002. Now, to match the anemia data (ha57_x, I understand that each one of these variables is linked to a woman in reproductive age of the household) with each woman I'm using the variable v003 (respondent's line number) and the ha0_x (Index to the household schedule):

gen anemia=ha57_1 if ha0_1=v003

replace anemia=ha57_2 if ha0_2=v003 ...

I'm doing something wrong because the number of missing values I get is always higher than in the report of the survey.

Can someone help me?

Many thanks,

Clara

Subject: Re: Missing values for anemia

Posted by [Liz-DHS](#) on Thu, 25 Feb 2016 23:15:51 GMT

[View Forum Message](#) <> [Reply to Message](#)

Dear User,

A response from one of our experts, Dr. Tom Pullum:

Quote:

You are apparently merging an IR file with an HR file. The HR file has one very long record per household. It is much easier to merge the IR file with the PR file, which has one record per household member, as follows. Open the IR file first, then gen hv000=v000; gen hv001=v001; gen hv002=v002; gen hvidx=v003; sort hv000 hv001 hv002 hvidx; save as temp.dta, replace. Then open the PR file; sort hv000 hv001 hv002 hvidx; merge hv000 hv001 hv002 hvidx using temp.dta.

However, you do not even have to do this. v457 in the IR file is the woman's anemia level. It's not necessary to do ANY merging. For each woman, you can also get the anemia levels of her children (up to 6 children born in the past five years) with hw57_1 through hw57_6. Here the subscripts refer to the woman's children. In the HR file, the subscripts refer to the line number in the household survey.

Subject: Re: Missing values for anemia
Posted by [clarapd](#) on Fri, 26 Feb 2016 14:26:18 GMT
[View Forum Message](#) <> [Reply to Message](#)

Thanks for you response, it's very easy now to generate this variable without merging.
However, the results I obtain now are the same I had and there is always a big number of missing values (.). Only a few are coded as 9.
I understand that all women in reproductive age (15-49) were tested for anemia.

Thanks for your help,

Clara

Subject: Re: Missing values for anemia
Posted by [Liz-DHS](#) on Mon, 29 Feb 2016 13:38:29 GMT
[View Forum Message](#) <> [Reply to Message](#)

Dear User,

Dr. Pullum asks:

Quote:

Can you give me a specific survey that has this problem? I just looked at Ghana 2008, more or less at random, and it does not have this problem. In general, "." should not be described as "missing" (although I sometimes do this myself!). It means "not applicable". Something in the skip pattern or in the eligibility for the question is responsible for this code. Tell me the survey and I should be able to reply right away.

Subject: Re: Missing values for anemia
Posted by [clarapd](#) on Mon, 29 Feb 2016 14:04:26 GMT
[View Forum Message](#) <> [Reply to Message](#)

For example, Benin 2011-2012. I see many (.) in the anemia variable in the individual recode and I understand that all women in reproductive age are eligible.

Thanks

Subject: Re: Missing values for anemia
Posted by [Liz-DHS](#) on Fri, 04 Mar 2016 21:10:44 GMT
[View Forum Message](#) <> [Reply to Message](#)

Dear User,

A response from Dr. Tom Pullum:

Quote:

I have talked with some people in the DHS office who are familiar with the Benin 2011-12 survey. They say that only a 1/3 subsample of women 15-49 were given the hemoglobin test. I see from the data file that there were 16,599 women and 5,513 had the test, which is almost exactly 1/3. Of the 5,513 women, 454 have code "9", which is not part of the label but would be the actual number "missing". The code ".", as I said before, means "not applicable". These numbers are unweighted.

I have looked very briefly at the report and the questionnaire. I cannot see anything about the subsampling or eligibility for this test. But maybe I am missing it because my French isn't good enough.

Apart from the subsampling, I have to warn you that DHS had problems with the implementing agency for this survey. Height and weight data were collected but are omitted from the report because of evidence of poor quality. That's very unusual. The anemia data should probably also have been suppressed. I think there should be some more explicit warnings, but I believe the report does describe, somewhere, the field problems and concerns about quality. Appendix C includes tables that indicate relatively high levels of genuinely missing responses (indicated by "9", "99", etc., rather than ".").

I believe that Benin is a special case. If you find other surveys with high levels of ".", I hope the report will provide documentation of subsampling. Let me know if you have other related questions.

Subject: Re: Missing values for anemia
Posted by [clarapd](#) on Wed, 09 Mar 2016 11:42:02 GMT
[View Forum Message](#) <> [Reply to Message](#)

Thanks for you help. I did not understand this because some of these questions are only asked to subsamples of the populations and this is country specific and it's not explained in the questionnaires.

Now, knowing this, I'm performing the analysis for 29 different countries. When I perform the analysis at a country level, I understand that I need to use the individual weights even if only a subsample answered the questions. My doubt is: when I append all datasets (29 countries) I need to rescale the weights to make each country representative of their population size, is that right? And for the rescale proces, I think that I need to drop (or to not take into account) the (.) values to make the subsample representative of the whole country population.

Thanks you very much.

Subject: Re: Missing values for anemia
Posted by [Liz-DHS](#) on Tue, 29 Mar 2016 20:14:29 GMT

[View Forum Message](#) <> [Reply to Message](#)

Dear User,
Do you still need assistance with this post?
Please let us know if you do.
Thanks!

Subject: Re: Missing values for anemia
Posted by [clarapd](#) on Wed, 30 Mar 2016 07:10:39 GMT
[View Forum Message](#) <> [Reply to Message](#)

Yes, my only doubt is if I need to rescale the weights of the countries I appended to make each country representative of their population size. And then, if I need to drop (or not to take into account) the (.) values to rescale.

Thanks you

Subject: Re: Missing values for anemia
Posted by [Liz-DHS](#) on Wed, 30 Mar 2016 14:53:05 GMT
[View Forum Message](#) <> [Reply to Message](#)

Dear User,
Response from Dr. Tom Pullum,
Quote:

Are you planning to pool multiple surveys? I would not recommend that. It will be best to calculate the prevalence of anemia separately for each survey, and not to calculate an overall prevalence for all surveys pooled. The combination of surveys does not correspond with a well-defined population and the dates of the surveys are different. But if you do want to pool, I recommend that you give equal weight to each survey. If, instead of that, you weight in proportion to the size of the country, your pooled estimate will be overwhelmed by the largest one or two countries.

If you want to weight in proportion to the country sizes (which I do not recommend) then you can use the population estimates for the number of children 0-4 (or other subpopulations) provided by the Population Division of the UN, for July 1 nearest to the mean or median date of the survey. The relevant number of cases in the survey would be the non-missing cases. That would also be the relevant number if you want to weight each survey equally.

The Stata commands for weighting in proportion to population size or for weighting each survey equally have already been posted on the forum, but let us know if you can't find them.
