Subject: Missing values in NDHS 2003, 2008 and 2013
Posted by mertmert on Thu, 23 Jul 2015 14:04:20 GMT
View Forum Message <> Reply to Message

I am a Masters studentand I am currently working on my thesis about the changes in HIV/AIDS prevention knowledge in Nigeria. I have received and I am working with the Nigerian DHS data of 2003, 2008 and 2013 and I have a few questions concerning missing values and I was wondering if you could help me understand the data. Thank you very much in advance.

I have combined the men's and women's DHS data sets of 2003, 2008 and 2013 in Nigeria and find myself with a data set of 115,144 individuals. I have three questions concerning missing values in this data set:

1. Many of the demographic variables do not have any missing values: age, region, urban/rural residence, highest educational level and wealth. The variable "marital status" also has a very low number of missing values (out of 115,144 only 1 value of missing). Because the data set is so large, it is very surprising that those variables have no missing values at all. I was wondering if the data sets that were provided to me already dealt with missing values for the demographic variables and if yes how? (imputing the missing data with replacement values?)

2. My second question is about two HIV/AIDS knowledge variables. The variable "knowledge about condoms to reduce the chance of getting the AIDS virus" (mv752cp for men and v752cp for women) has 12,903 missing values and the variable "knowledge about limiting oneself to one faithful and uninfected sexual partner" (mv754dp for men and v754dp for women) has 12,883 missing values. However the missing values of these two variables come mostly from the same individuals. Once the 12,903 missing values of "knowledge about condoms to reduce the chance of getting the AIDS virus" are eliminated, the variable "knowledge about limiting oneself to one faithful and uninfected sexual partner" has only 154 missing values. Since the same individuals seem to not have answered these two questions, I was wondering if maybe the question was not asked to some specific groups/sub-groups for some reasons? or at random?

3. Finally, I am concerned with the variable "used condom during last intercourse" (mv761 for men and v761 for women). This variable has 30,277 missing values which is a lot. I was wondering if there was some kind of selection when this question was asked. Was it not asked to certain kind of respondents? Why are the missing values so high?

Again, thank you very much in advance.