Subject: Convert DHS (SPSS?) missing value codes to Stata codes in Stata dataset
Posted by econ_stata on Fri, 24 Jul 2015 14:42:13 GMT
View Forum Message <> Reply to Message

Greetings from a new member! I am quite new to the DHS survey data, hence please excuse the possibly naive question. Apologies also for a very long post- but I wanted to be quite clear & precise!

For the Indian NFHS-3 survey, I am trying to study the relations between maternal education & child health outcomes. Given the choice of dataset formats available, I chose the Stata data files, as I am quite familiar with Stata. I am looking mainly at the Child data file, KR.dta, but have also needed to merge in the individual & household files. But I find now that the data has many variables (e.g., child/parent height or weight for age percentiles hw4-hw11) which have special missing value codes (e.g., 999, 9999 etc.)   which are not recognized as missing codes by Stata.

E.g., from the NFHS Recode Manual (Description of the Demographic and Health Surveys Individual Recode Data File, 2012) :

Quote:"Special codes are used throughout the data file for certain responses. The general coding scheme is presented below. The codes given apply to 4 digit, 3 digit, 2 digit and 1 digit variables, respectively. If there are other special responses to questions, these are coded in decreasing order from these special codes, i.e., 9996, 996, 96, 6; 9995, 995, 95, 5; etc." 'Coding Standards', p3, Recode Manual) [I am not very familiar with SPSS coding conventions, but from searching around, these appear to be SPSS missing codes.]

Of course, if I know that certain specific variables have missing values coded as a particular set of numbers, it is quite simple to recode them, using, e.g., the mvdecode function in data, or even write a do-file to lop over each wrong code for each variable, & replace them with Stata missing codes, whether standard "." or extended ".a, .b etc. E.g., for some of the anthropometric variables , it is mentioned on p51-52 of the Recode Manual that missing values are coded 9999,9998, 99999, 99998. These I have already replaced with Stata's missing codes, via a do-file.

But is there any way to convert for sure & in one shot ALL the "wrong" (for Stata) codes? So far, the means I have discovered are:
1. Download the SPSS files & from within SPSS, save as Stata datasets.
2. From Stata, convert using a user-written module "-usespss-" but this works only on a windows system
3. Download the ascii file & Import into Stata
4. Use stat/transfer.
[from http://infoguides.gmu.edu/c.php?g=120989&p=1268167]

None of these are very convenient for me. e.g., I am now working with a Macbook, don't have SPSS or Stata/Trnsfer.

Has any other member dealt with this issue, or can anyone suggest what would be the best way for me tackle this? In particular, is there somewhere a program (Stata ado file) to convert all the missing value codes in the Stata files provided by DHS? Some time ago, when working with the SHARE (Survey European Health & Retirement) datasetI, found such a program (sharetom.ado) provided with it. I hope there is some such program available for DHS data users on Stata. It certainly seems strange that a file which has been distributed as a Stata dataset should have such wrong (for Stata) missing value codes.

Any help much appreciated!

Subject: Re: Convert DHS (SPSS?) missing value codes to Stata codes in Stata dataset
Posted by user-rhs on Fri, 24 Jul 2015 18:55:55 GMT

View Forum Message <> Reply to Message

Not sure what you mean by "wrong" codes. It is perhaps "inconvenient" that Stata do not recognise special codes as "missing," but they are certainly not "wrong." Sometimes it is useful for the user to distinguish missing values that were the result of a skip pattern (.) vs. missingness because of non-response or a response of "don't know." It is up to the user to manage the data in such a way that is useful to him or her. Why do you need to convert from SPSS to Stata? Are the India DHS data not available in Stata format?

Subject: Re: Convert DHS (SPSS?) missing value codes to Stata codes in Stata dataset
Posted by econ_stata on Sat, 25 Jul 2015 02:08:04 GMT

View Forum Message <> Reply to Message

Many thanks for your reply. I was perhaps not clear enough-though I did mention it in the message header and a few places in the message.   I am working with the Stata dataset downloaded from the DHS site, and yet encountering these codes!   And they affect quite a few observations for the variables I have looked at so far. Hence my question. Am I missing something?

Subject: Re: Convert DHS (SPSS?) missing value codes to Stata codes in Stata dataset
Posted by Bridgette-DHS on Fri, 07 Aug 2015 13:23:21 GMT

View Forum Message <> Reply to Message

Following is a response fron Senior DHS Stata Specialist, Tom Pullum:

I can suggest three different ways to deal with these kinds of missing value codes. I use them all the time.

As an example, take hw70, the height-for-age z-score. The anthropometry z-scores have several special codes in the vicinity of 9999. Sometimes you will find values in that vicinity that do not even have a label, but all such values must be excluded.

One approach would be simply to have a line such as "replace hw70=. If hw70>9000". Values with "." Are always considered by Stata to be missing and will be ignored from calculations. The problem with this is that you have now lost the original hw70. A second approach would be "gen hw70r=hw70" and "replace hw70r=. If hw70>9000". I add "r" for this kind of simple recode. Then any analysis would use hw70r in place of hw70, and you still have the original hw70. A third approach, when you have several related variables, could be something like the following. "gen hw7x_missing=0", "replace hw7x_missing=1 if hw70>9000 | hw71>9000 | hw72>9000". Then in your analysis, you could limit yourself to the cases that are non-missing on all variables by including "if hw7x_missing==0". I use this third approach if, say, I want to do a series of regression on exactly the same cases.

One more thing --in the DHS data files, the code "." Always means "not applicable". You should not confuse that meaning with what I have implied above, which is "please ignore in any calculations"!

Subject: Re: Convert DHS (SPSS?) missing value codes to Stata codes in Stata dataset
Posted by econ_stata on Fri, 04 Sep 2015 17:34:15 GMT
View Forum Message <> Reply to Message

Many thanks for your detailed & very helpful answer Dr. Pullum, and my apologies for the very late acknowledgement.

Subject: Re: Convert DHS (SPSS?) missing value codes to Stata codes in Stata dataset
Posted by phres110 on Thu, 06 Apr 2017 02:38:32 GMT
View Forum Message <> Reply to Message

Dear DHS experts
I am a regular user of dhs forum as i am working on my thesis using dhs data sets. Now my question is i am working on 5 dependent variables like Antenatal care, Postnatal care and delivery by c/s family planning and immunization of child. For independent variables profession, husband education, ethnicity and decision making i already deleted all missing values although my sample size is reduced but not as much. I am confused should i also deleted all missing values from dependent variables they are about 150 in ANC 86 in PNC and 49 in CS and about 500 in PNC? what should i do?
i am using SAS.
Regards

Subject: Re: Convert DHS (SPSS?) missing value codes to Stata codes in Stata

dataset
Posted by Bridgette-DHS on Thu, 06 Apr 2017 17:15:54 GMT
View Forum Message <> Reply to Message

Following is a response from Senior DHS Stata Specialist, Tom Pullum:

You are deleting the cases with missing values? This is not necessary and not desirable.

Your message refers to SPSS, Stata, and SAS. They have somewhat different ways of handling missing cases. In Stata, the only package I use, a blank or dot means not applicable. Those values are not exactly "missing". There's no value because the relevant question was not applicable. Stata will automatically exclude those cases from calculations. There will often be other codes such as 9, 99, etc., meaning that there should have been a response but there wasn't. You will want to exclude from calculations. You can do that by assigning those codes to a dot or doing some similar recode; just be sure that you never permanently over-write the original code. Excluding a case by dropping it from the file is not a good idea because it is quite possible that you will want those cases later for some other reason, and once they have been dropped you can only get them back by starting over.

Subject: Re: Convert DHS (SPSS?) missing value codes to Stata codes in Stata dataset
Posted by phres110 on Fri, 07 Apr 2017 01:01:36 GMT
View Forum Message <> Reply to Message

Dear Expert
Thank you very much for your suggestion.It is written in DHS guide:

"Other special responses and codes are: "inconsistent," "don't know," and "blank" (or "not applicable.") "Missing," "inconsistent," "don't know," and "blank" codes are excluded when calculating statistics such as means or medians; otherwise they are treated as real values. - See more at: http://www.dhsprogram.com/data/Data-Quality-and-Use.cfm#stha sh.4VQufOxh.dpuf"

i read in some papers that you may delete the cases if you are sure they are not specific for any reason. Although it may reduce your sample size but if there are not a high number of missing values its Ok. and some papers suggest not to delete the cases do imputation ....i am confused how to deal with missing values in my analysis. I am using SAS. Kindly guide me what to do in SAS that without deletion i can do my analysis correctly.
Many thanks and Regards.

Subject: Re: Convert DHS (SPSS?) missing value codes to Stata codes in Stata dataset
Posted by Bridgette-DHS on Fri, 07 Apr 2017 14:26:10 GMT

View Forum Message <> Reply to Message

Another response from Senior DHS Stata Specialist, Tom Pullum:

There is a large literature on imputation, by statisticians such as Roderick J.A. Little and Donald Rubin. "Multiple imputation" is perhaps the best approach (in Stata the relevant command is "mi impute"). You can use Heckman selection models ("heckman" in Stata) to counteract bias. I don't know the procedures in SAS. I personally would not consider either approach unless the level of missing was well over 10%, say, and I thought there was a systematic difference between the people with a response and the people without a response. You could construct a variable "missing" that is coded "." for NA, 0 if not missing, 1 if missing, and do logit regressions on the usual covariates, such as wealth, age, education, place of residence, etc., to see if there was a significant and systematic difference. (Basically a "selection model".) DHS cannot provide further guidance on the topic. You may need to talk with a statistician who knows the topic.