Subject: Missing data
Posted by nwegbus on Tue, 08 Dec 2015 22:07:23 GMT
View Forum Message <> Reply to Message

Hi,
I'm a young researcher writing my first independent academic paper. I'm using the DHS 2008 Indidviual dataset in Stata and I have a lot of missingness (as much as 8505 on my outcome variable). When I try to do multiple imputation, the Stata 13 IC only allows for up to 1000 per variable, but I need to do at least 8505 to make it up to my original sample size which is 23,954.

Do you have any idea how I might get around this? Or what could I be doing wrong.

Many thanks for your help.

SN

Subject: Re: Missing data
Posted by Bridgette-DHS on Fri, 11 Dec 2015 12:32:30 GMT
View Forum Message <> Reply to Message

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

There was a similar question on the forum earlier this week. You are almost certainly interpreting the "." code in the Stata files. As DHS uses that code, it always means "Not Applicable". Those values are NOT MISSING. The question(s) on which the variable is based were not asked. For example, if a woman has not had any children in the past five years, then all the variables for such children will be coded ".". Values that are missing because of nonresponse or invalid codes will be assigned 9 or 99, etc. There are rarely enough such cases for you even to think about imputation.

Subject: Re: Missing data
Posted by nwegbus on Fri, 11 Dec 2015 13:52:38 GMT
View Forum Message <> Reply to Message

Thanks so much!

OK so I guess my follow-up question would be - so how do I deal with that because I can't drop those observations otherwise my variances will be biased right? But if I keep them, even my point estimates will be off. I'm thoroughly confused now. Is there a syntax/command for delineating my sample so as not to include those not-applicables in my model estimation?

I'm already working with the svy, sub pop: command since I'm actually only interested in a subpopulation of married women. So is there a command that would be supported by the above syntax which I can use to exclude the not-applicable? Or is it something I would have to take care of prior to weighting my data?

Many thanks.

SN

Subject: Re: Missing data
Posted by user-rhs on Sat, 12 Dec 2015 02:44:17 GMT
View Forum Message <> Reply to Message

Cross-posted from my response to your question here:
http://userforum.dhsprogram.com/index.php?t=msg&th=4728& amp; amp; amp; amp;goto=8743&S=890da8edd7880c05ebf2f52f2d8e9db3#msg_874 3

Edit: I also completely agree with Tom above that N/A is NOT the same as missing. See my post below for a discussion on skip patterns

When you run a regression model in Stata, Stata handles missing values with listwise deletion. This means that if even a single variable is missing from a list of covariates in your model, that observation will be excluded from analysis. The obvious problem when this happens is that your parameter estimates will usually be biased, unless the data are missing completely at random (MCAR). Data are rarely, if ever, MCAR.

Fortunately for you, before you go off and read Little and Rubin's rather excellent Statistical Analysis with Missing Data concurrently with Stata's Multiple Imputation Manual (recommended for the bold and adventurous types out there) to follow your chief evaluator's advice of doing "multiple imputation," there are things you should do to determine whether it is even necessary in the first place for you to do multiple imputation. (By the way, these are also the things Little and Rubin recommend doing in the first few chapters of their book). Many scientists have a tendency to go for the shiniest and fanciest new toy (and statistical models because we want to sound smart), but in many cases, the simple solutions may be sufficient.

Before I start, here's something that I think most seasoned statisticians will agree on:

The key to fitting good models is understanding your data and the data generation process. Therefore, you should familiarize yourself with the data (read the questionnaires, DHS recode manual, any data documentation that came with your dataset, DHS report for the country, run tabulations/cross-tabulations, etc.) before attempting to do any further analysis.

So, if you have not done so already:

1.) Examine each variable in the dataset to determine level of missingness. I like the user-written command -mdesc-, but this command will not give you the % missing if "missing" was coded as something other than (.) in the dataset. Doing a -tab, miss- for each variable will tell you exactly the numbers and proportions of system and non-system missing in those variables.

2.) When you find one or more variables with huge chunks of missing data, think about the process that generated the missingness. Does it make sense that the information was missing on that person, or should there be a response there? Were the data missing because the respondent refused to answer it or didn't know the answer to it (e.g. 98, 99) or was it because the question was not asked of the respondent (for example the skip pattern in the questionnaire). Speaking of skip patterns, it is helpful to familiarize yourself with the questionnaire used to collect the data, because it will tell you why the person was not asked the question based on their responses to another question. If the person did not answer the question due to a skip pattern, it probably does NOT make sense to try to impute a response (it's missing for a reason--if you asked them about how many years they have lived with their current husband, and they have never been married, they probably will not be able to give you an answer). If the person was supposed to answer the question (e.g. 97, 98, 99 missing codes), and the data are missing in huge quantities based on those missing codes, then you probably should impute.

3.) Determine how you're going to handle missing data. For most variables, there should be little to no missing, but these can add up, especially if you have many model covariates. You have several options (each has its limitations, but what can you do):

 Do nothing and lose observations in listwise deletion--Some people may find this blasphemous, but if you lose 40 people out of a sample of 20,000, it's not a big deal If the variable that contains huge proportions missing is binary, consider changing it to 1-"Has the characteristic" and 0-"Otherwise" instead of 0-"Does not have the characteristic". That way, people with 99 and (.) can stay in the model If the variable that contains huge proportions missing is based on a skip pattern, consider recoding the missing to its own category and adding a "flag" (dummy) that takes on the value of 1 if the variable that determined whether the person got to answer the question was "Yes, eligible" and 0 otherwise. For example, if you have "number of miscarriages and stillbirths" as a model covariate, but this question was only asked of women who have had at least one pregnancy (the value will be . for women who have never been pregnant), then you can create a dummy variable called "ever had pregnancy" 0/1, and create a categorical variable based on "number of miscarriages and stillbirths" into something like "0 - Never had pregnancy; 1 - No miscarriages/stillbirths; 2 - 1 to 2 miscarriages/stillbirths; 3 - 3 to 4 miscarriages/stillbirths; 4 -5 or more miscarriages/stillbirths" This way, you do not lose the people who were never pregnant from the model.

Caveat: I had a prof. who handled all of her missingness in this way (creating a sort of "flag" variable for the missing generation process), but you have to be very careful when you do this because you make the assumption that ALL people who are missing share the same characteristic after controlling for all other model covariates, which may not be true.

4.) It is always a good idea to add variables into your model one by one (or chunk by chunk, if you prefer) just to see how the model responds to the addition of other variables. It is also always a good idea to run bivariate analysis before you fit your multivariate model so you get an idea of how things are supposed to be related and how they change once you control for other factors.

Good luck!

RHS

Subject: Concentration Index
Posted by danelnya on Fri, 01 Jul 2016 14:02:39 GMT
View Forum Message <> Reply to Message

Good Morning
I am a new user on this forum
I am actually do a research on poverty and health inequality in health using DHS data.
I want to know if anybody can show to me how to calculate concentration index using DHS data.
Thank's

Subject: Re: Concentration Index
Posted by Bridgette-DHS on Tue, 05 Jul 2016 16:44:47 GMT
View Forum Message <> Reply to Message

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

Open Stata and enter "help ginidesc". You will then see how to download this command and to access the help that goes with it.

Subject: Re: Missing data
Posted by filotempo on Fri, 09 Jun 2017 13:04:30 GMT
View Forum Message <> Reply to Message

Dear all,
a follow-up to the discussion about missing data.
I am analysing 2008 Bolivian DHS and I am dealing with a large amount of missings for the variables antenatal care, tetanus toxoid injections, place of delivery and assistance during delivery, which are important covariates in my analysis that has neonatal death as outcome.
About two thirds of the answers to those questions in the birth recode dataset were codede as ".", while very few observations were coded as 98 or 99.
Why would such information would be NOT APPLICABLE to children?
Any help would be highly appreciated.
Many thanks,
Filippo

Subject: Re: Missing data
Posted by Bridgette-DHS on Mon, 12 Jun 2017 11:47:13 GMT
View Forum Message <> Reply to Message

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

Many such questions are limited to the most recent birth. If you cross-tabulate one of those outcomes with bidx, you will probably see that all or virtually all of the "." codes are for children with bidx>1. (The most recent birth has bidx=1.) You can also look at the questionnaire in the back of the main report, and check the skip pattern.

Subject: Re: Missing data
Posted by Marejoha on Thu, 22 Mar 2018 12:43:35 GMT
View Forum Message <> Reply to Message

Hi,
I have a question relating to "." being not applicable. I am working with a few datasets, using among other variables the body mass index. I cannot find an explanation to why there are observations where bmi is noted as "." in so many datasets. Moeover why some datasets label the whole variable as not applicable
Many thanks,
Maren

Subject: Re: Missing data
Posted by Bridgette-DHS on Wed, 28 Mar 2018 11:58:50 GMT
View Forum Message <> Reply to Message

A variable is not applicable for a respondent either because the question was not asked in the survey or because the question was not asked of this respondent due to the flow or skip pattern of the questionnaire. Please take a look at the questionnaire for the country you are working with - all questionnaires are in the appendix of the final report.

Subject: Re: Missing data
Posted by boyle014 on Sun, 06 May 2018 22:33:19 GMT
View Forum Message <> Reply to Message

Dear Marejoha,

For samples in IPUMS-DHS, there is documentation for each variable. When you click on the name of a variable, you will see a series of tabs. One is "Universe," which explains who was asked the question. Another is "Survey Text," which provides the exact question wording (translated into English) and the ability to jump into the questionnaire to see the surrounding questions as well.

This can be very helpful for determining who has missing values and why.

Good luck!
Liz

Subject: Re: Missing data
Posted by Hassen on Wed, 09 May 2018 14:36:44 GMT
View Forum Message <> Reply to Message

Dear all,I have some questions based on 2016 Ethiopia DHS KR data set:-
1)Even if age (HW1for KR file) is necessary to analyze Childhood Nutritional status,There are alot of "." or not applicable symbol. So How can I deal with it? Can I clean these cases?
2)Another challenge to conduct my MPH thesis is,My target population are children aged 6-59 months old,So Can I delete/clean cases less than 6 months old children?
3) Regarding missing values(9999),flagged cases(9998),not present(9994),refused(9995)and 9996(others)in variables like WH70,WH71 and WH72,What are your recommendations? Can I delete/clean them? or Can I imputate them?
4) Hello my Heros,I am in many challenges regarding Complex samples,Imputation and creating Aggregated Community level factors like Multiple child deprivation index,Community women education rate etc using SPSS Version 24. So please kindly tell me you suggestion and recommendation on these issues.
Respectfully,Hassen

Subject: Re: Missing data
Posted by Bridgette-DHS on Fri, 11 May 2018 15:29:30 GMT
View Forum Message <> Reply to Message

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

It would be good if you (and other forum users) could always say what survey you are using. Like your question, my answer is generic. Cases with the "not applicable" code should be omitted from any analysis. Usually they will be omitted automatically, because a blank or a dot is not numeric and can't be included in calculations. Cases with missing value codes should also be omitted, and that must be done explicitly with an "if" or "select" statement. It is not a good idea to drop them from the file, because then you lose cases that are non-missing on other variables you may want to analyze in the same run. It is best to construct a new variable with a new name. For example, with hw70 you could call the recode "hw70r" or "HAZ". In Stata I would do something like this to go from hc70 (the name in the PR file) to stunted:

```
gen stunted=0
replace stunted=1 if hc70<-200
replace stunted=. if hc70<-600 | hc70>600
summarize stunted [iweight=hv005/1000000] if hv103==1
```

Subject: Re: Missing data
Posted by Hassen on Sun, 13 May 2018 18:31:41 GMT
View Forum Message <> Reply to Message

Thank you very much our Hero!!
With Best Wishes,Hassen