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