Missing Data: Part II

Diagnosis of missing pattern before imputation

Saeid Shahraz

Heller School for Social Policy and Management
Brandeis University, Massachusetts, USA
shahraz@brandeis.edu
Saeid.Shahraz@saphirnetwork.org

September 14, 2012

Missing Completely At Random: MCAR

- Missing information are **unrelated** to both observed variables and the variable with missing values
- Example: Lab sample accidentally thrown out
- Missing information were independent of observed variables and variable 'blood sugar' (missing variable)

****

ANYONE WANTS TO TAKE A CRACK AT REPHRASING IT ALL?

****

Missing At Random: MAR

- Missing information are **related** to observed variables and **unrelated** to the variable with missing values
- Example: Missingness on household income in relationship between kidney donation and household income
- Missing information were dependent upon observed variables (e.g. gender as female) and unrelated to variable 'household income' (variable with missing values)

****

ANYONE WANTS TO TRY THIS EXAMPLE AGAIN?

****

Missing NOT At Random: MNAR

- Missing information are **related** to observed variables and **related** to the variable with missing values
- Example: Missingness on blood sugar reduction time in association between a synthetic insulin and blood sugar reduction time
- Missing information may depend on observed variables (e.g. age, gender) but more importantly are related to variable 'blood sugar level reduction time' (variable with missing values)

****

ANYONE LIKES TO GIVE IT A SHOT AND EXPLAIN IT AGAIN?

****

Diagnostics: External Information

- Try to get as much external information as you can on the reason for missing information
- Such information can be found from different sources but always check the formal documentations that come with the data set

Diagnostics: Associations in MCAR(1)

- In MCAR the distribution of non-missing values is not truncated or does not lack data on a spot
- In MCAR you cannot predict the probability of 'being in missing group' through observed information

Diagnostics: Associations in MCAR(2)

- In our example of MCAR " lab samples thrown out" if you look at the composition of the two populations (one with missing blood sugar variable and the other with non-missing blood sugar variable) you will not find any difference. This is a very useful test and you can run usually a Chi-Square test (or logistic regression) / T-test (or single regression) to find out about systematic differences in the composition of the two population. This is a popular check that you can see in many published articles.

Diagnostics: Associations in MAR

- In MAR the distribution of non-missing values is not truncated and does not lack data on a spot
- In MAR you can find association between probability of missingness and population characteristics
- In our example of MAR " kidney donation and household income" you may be able to predict the probability of a case to be in a missing population for example through gender.

Diagnostics: Associations in MNAR

- In MNAR the distribution of non-missing values is truncated or lacks data on a spot
- In MNAR it is not important to find out if the probability of missingness is predictable.
- In our example of MNAR " synthetic insulin and blood sugar reduction" you may be able to predict the probability of a case to be in a missing population but it is not probably useful because the reason for missingness is not random in the first place.

Take Home Points(1)

- Get external information
- Explore the distribution of missingness to rule out MNAR
- Run association tests to rule out MCAR
- If you get MAR you are lucky because you can use a well-known technique such as Multiple Imputations to fill in missing data

Take Home Points(2)

- if you end up with MCAR you are even luckier because complete case analysis does not give you biased estimates but maybe inefficient. If so, imputation techniques increase your efficiency
- If you wind up with MNAR you are not lucky because you cannot use well-known methods of imputations and in case you need to impute your data you need to use a rather arcane technique

Home Work

What do you think about non-response in population surveys? If you like you can email me your response.

THANK YOU and THANK DR. SYAMAK MOATTARI FOR COORDINATING THIS WEBINAR