Missing Data

Why Do Data Go Missing?

- We need to think about why data go missing, so we can make informed decisions about how to handle it.
- We can classify missing data as follows:
  - MCAR: missing completely at random
  - MAR: missing at random
  - MNAR: missing not at random
- This terminology can be confusing, and the topic can be challenging to think about.
- There will be no easy answers.

Missing Completely At Random (MCAR)

- Generally, this is the easiest situation to deal with, it means that the missing status of a value for a variable is not associated with its actual (unknown) value or the value of other variables. Its missingness is independent of everything!
- Unfortunately, we almost never can be sure of what class of missing data we are dealing with! We just make informed guesses.
- MCAR can happen for a variety of reasons, but it has to be something unrelated to the data you are collecting (e.g. an intermittent measurement failure).

Missing At Random (MAR)

- MAR means the data that are missing are associated with other variables in the dataset, but not with the actual values of the missing data.
- This is slightly more serious than the MCAR situation, in a sense.
- Suppose that men are less likely to answer questions about their salary than women (this is a made up example, I have no idea).
- Also assume that there is no relation between those that are missing and their actual salary (among men). Then the data are MAR.

Missing Not At Random (MNAR)

- MNAR means the data that are missing are associated with the value of the missing data.
- This is the most difficult scenario to deal with.
- It is also very difficult to know if it is the case. But sometimes, you can get a pretty good idea.
- Suppose someone collected data on household salary and that those with low household salary tended not to respond. What would this do to your measure of median household income?
- These data would be MNAR.
- Sometimes, if you could get more data, MNAR can become MAR.

Complete Case Analysis

- This is one of the approaches we have used so far for dealing with missing data.
- If an observation has missing data – get rid of it!
- This will only give you a valid result in your analysis if the data are MCAR, or possibly if there is a relatively small number of missing values.
- Data are seldom MCAR, so I would reserve this method for when there are few observations missing, or you have enough data to be fairly confident in your MCAR hypothesis.
- You can also try Little's test.

Pairwise

- You can analyze relationships between variables for which the data are complete.
- The data used will thus be different for different pairs of variables.

Dropping Variables

- This might seem like an obvious thing to do when you have a lot of missing data for a variable. But you should think about it.
- Do you really want to give up on what that variable could tell you?
- If the variable is mostly missing, it probably isn't much good.
- If the variable tends to be correlated with other variables you have, you might not need it.
- You might consider modeling methods that are tolerant of missing values, and you might consider visualizing what you do have available.

Mean Imputation

- I’m mentioning this method because you will undoubtedly hear about it. It is not the recommended way of doing things.
- This method involves replacing missing values with the mean for that variable.
- Shrinks the variance.
- Invalidates any linear models you try to build (at least for the response).
- Invalidates hypothesis tests.
- Invalidates confidence intervals.
- Affects correlations.
- Might be your only choice if your software is limited.

Regression Imputation

- Again, I'm going to mention this as something not to do, because it is done pretty often.
- Regress values of the variable that has missing data on those that don't.
- Use the model to predict the values that are missing.
- However, this will have the effect of emphasizing associations that may not really be that strong.

Multiple Imputation

- The big problem with regression imputation is that we treat the imputation as if it has no error.
- In multiple imputation, we build a number of models based on different samples of the data and then use all these models together.

Categorical

- Mode imputation. Yes, this is problematic.
- Treat missing data as a category – could be helpful, but often not terribly useful.
- Use models (could use regression imputation, logistic in this case, but why not use multiple imputation?)

KNN

- Works for both categorical and continuous variables!
- Find the k nearest “neighbors” to an observation by various distance metrics.
- Take the mean of the missing variable from the values in its neighbors (or mode for categorical).
- A simple and intuitive method that works pretty well. Probably not as good as Multiple Imputation, which has a firmer statistical basis.

Time Series

- LOCF – Last Observation Carried Forward
- NOCB – Next Observation Carried Back
- Linear Interpolation
- Seasonal Adjustment