Missing Data Project (optional extra credit)

Start Assignment

Due Monday by 11:59pm **Points** 15 **Submitting** a file upload
Available after Nov 12 at 12am

An extra credit optional (group project, with groups of any size from 1-5). Nearly all "real" datasets have columns/variables with missing values, so how does one deal with this? (This is discussed at end of Ch 9 on textbook).

- Missing Completely At Random (MCAR), you can actually just remove the rows of data that contain missing values for the columns/predictors
- Missing At Random (MAR), removing rows may bias results
- Missing Not At Random (MNAR), actually very hard to deal with

Review in notes/slides the definitions of each. Here we will deal with MAR dataset and solve via multiple imputation using Monte Carlo Markov Chain (MCMC) approach via the "mice" package in R.

Overview of the process:

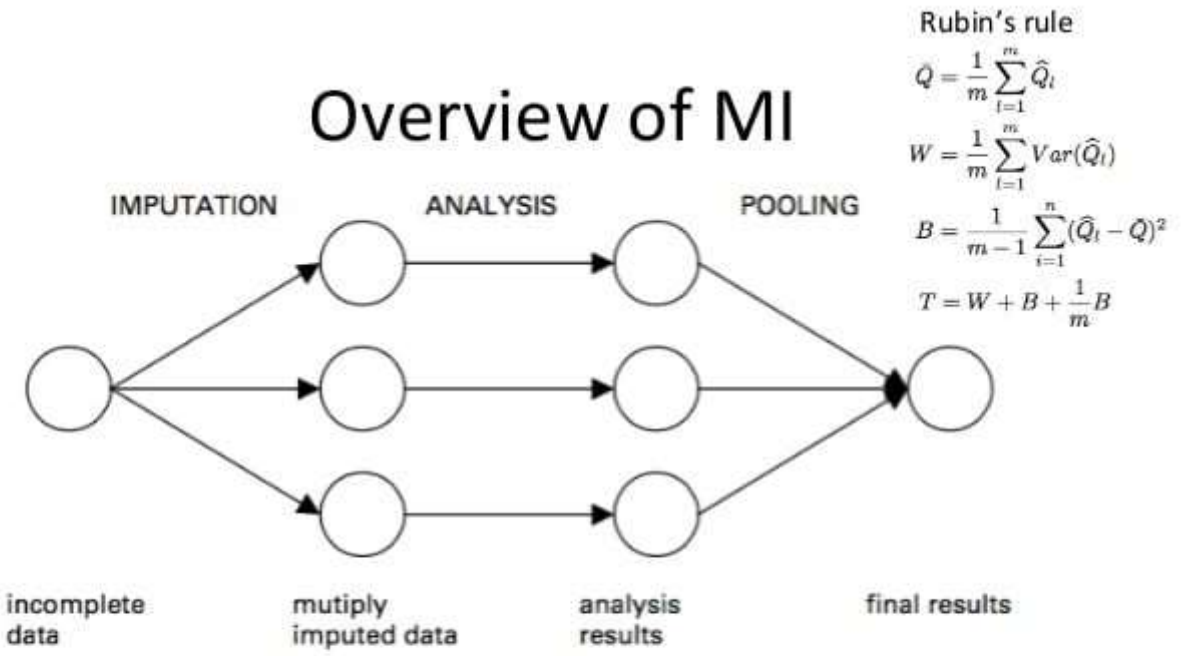

Figure : Schematic representation of the steps in multiple imputation. The process starts with an incomplete data set (on the left side), which is imputed m times (m=3 here) thus creating m completed data sets. Each complete data set is analyzed by using standard complete-data software, thus resulting in m analysis results. Finally, these m results are pooled into one final result that adequately reflects the amount of uncertainty in the estimates.

van Buuren 1999

First, a bit of reading:

- Appendices of **class notes** about missing data and multiple imputation
- **https://web.archive.org/web/20050212022244/http://www.stat.psu.edu/~jls/mifaq.html (https://web.archive.org/web/20050212022244/http://www.stat.psu.edu/~jls/mifaq.html)**
- **https://datascienceplus.com/handling-missing-data-with-mice-package-a-simple-approach/ (https://datascienceplus.com/handling-missing-data-with-mice-package-a-simple-approach/)**
- **https://datascienceplus.com/imputing-missing-data-with-r-mice-package/ (https://datascienceplus.com/imputing-missing-data-with-r-mice-package/)**

The data is the **(missing) crab data** ↓ **(https://ufl.instructure.com/courses/435845/files/64146337/download?download_frd=1)** we have covered in class but with missing values and the response is the count/number of male satellites (sat). Do the following, **justifying and explaining your choices.**

1. Describe missing data pattern, as shown in the datascience website links...and any other descriptive statistics you deem necessary. This is basically an "introduction" to the data and where the missing values are located.
2. Implement multiple imputation using an appropriate method for each variable to create multiple complete datasets and check whether the complete datasets are "similar" to the original data (that had the missing values). To view the list of imputation methods for each variable, read the **mice package documentation (https://cran.r-project.org/web/packages/mice/mice.pdf)** on the *mice.impute* methods and explain why and which you chose for each variable. The default is predictive mean matching "pmm" but explore others. For now, ignore the 2-level, "2l" methods.
3. Perform appropriate analysis on each full data set, using an appropriate GLM model for the specified problem.
4. Pool results from the multiple imputations
   A. to obtain model parameter estimates and standard errors to create 90% CIs on each parameter coefficient.
   B. You can perform full vs reduced model tests using D1() and D3() methods in the mice package for Wald and LRT tests respectively.
5. Interpret your model and state your conclusions.

p.s. Initially I was going to have you fit a zero inflated poisson (ZIP) model but the "pool" function in mice does not work on outputs from the function "zeroinfl" and the pooling would have to be done manually. And you would have to read up on ZIP literature and that is too much for an extra credit assignment.