**How to Handle Missingness**

Missing Data Analyses Info from Multivariate Seminar:

* Basic Steps in Data Cleaning:
  + Check accuracy of data file (ex. if there are any impossible values)
  + Assess missing data patterns (and estimate missing data, if appropriate)
  + Examine normality of variables (and transform or deal with outliers, if necessary)
  + Assess multicollinearity
  + For some MV method that don’t make strict assumptions, so we can “violate at will”
* Data Accuracy
  + Run Frequencies
    - Are data within range?
    - Are means and SD reasonable?
    - Are missing data coded correctly?
    - Do some items have restricted ranges?
* Missing Data
  + Design consideration: well-designed surveys and experiments will have fewer missing data points
    - Not too many questions
    - Check surveys for completeness
    - Check response options for mutual exclusivity and exhaustiveness
    - Conduct research in quiet places
    - Pilot test computer programs
  + May or may not be a big deal, depending on:
    - Amount missing
    - Pattern of missing data
    - Reasons for missing data
  + Pattern of missingness is more important than amount?
  + Fewer than 5% of your observations missing won’t systematically bias your conclusions (may vary if you have a smaller sample, heuristic)
  + Data that are missing completely at random (MCAR) are the easiest to deal with
    - Not something that can ever been truly knowable
  + Data that are missing at random (MAR) are also pretty easy to deal with
  + Data that are missing not at random (MNAR) are serious because they compromise generalizability of results
  + Step 1: Test amount of missing data
    - Dummy code variable: missing data on variable x: Yes/no
  + Step 2: Test pattern of missing data
  + Step 3: Decide how to deal with missing data
    - Delete cases (Ana calls this the dumbass maneuver)
    - Estimate cases
    - Treat missing data as data
    - Repeat analyses with and without missing data
  + Missing Data Imputation (from most dumbass to least)
    - Prior knowledge (lol this is BS)
    - Mean substitution (reduces variability in your data which is not great)
    - Regression (uses other data in your data file to predict that person’s score, tends to slightly overfit so variance is slightly less)
    - Expectation maximization (assumes MAR data)
    - Multiple imputation
      * These last 3 ROCK!
* Missingness
  + May be candidates if:
    - Missing at greater than 5%
      * Unless there’s a logical reason there might be missing data from your theoretical or methodological perspective
        + Ex. Time 3 variables missing at greater than 5% might make sense because of attrition and other logical reasons
* When we impute via regressions, you’re predicted variables are not going to be extreme values, so you’re reducing variance (i.e. same mean but lower SD)
  + If you don’t have variability in your data, you don’t have anything to explain. AKA you need variability in order to explain variability.
  + So, we only want to replace the NAs with the predicted values, not all values 🡪 which is really good because we increase the variability back up close to the original
* A successful imputation will maintain variability and not distort the data!

Schlomer, G. L., Bauman, S., & Card, N. A. (2010). Best practices for missing data management

in counseling psychology. *Journal of Counseling psychology*, *57*(1), 1.

To determine whether a variable is missing at random:

* Dummy code the variable of interest (1 = missing, 0 = present)
* Examine whether this dummy-coded variable is associated with other variables in the model
* If this variable is unrelated to other variables in the model, then the pattern is considered missing completely at random (MCAR) OR not missing at random (NMAR)
* If this variable is related to other variables in the model, then it may be missing at random (MAR), but we cannot rule out that it is not missing at random (NMAR)
* When there is a pattern to missing data such that the likelihood of missingness is related to the score on that same variable had the participant responded, the data are NMAR
  + However, at this point, it is a conceptual question 🡪 is it likely that participants who are high (or low) on some variable more likely to have missing data?
    - Ex. Annual income question missing because those with high annual income might feel uncomfortable with the researcher knowing this

Deletion methods:

* Not typically recommended because it just ignores missingness
* Listwise deletion / case completion analysis = delete any cases that have missing values
  + May bias the data if the cases that have missing data are different in some way than the cases that don’t
  + Also results in a loss of statistical power
* Pairwise deletion / available case analysis = only cases that have data missing on a required variable are deleted from analysis
  + Makes bivariate correlations very complicated because you are using different cases for different comparisons

Nonstochastic methods

* Imputation = substituting a missing value with a plausible value
* Mean substitution = missing values are imputed with a mean value of that variable on the basis of the nonmissing values for that variable
  + Assumes that data is MCAR
    - If this is not true, it results in biased means
  + Also reduces the variance of that variable (underestimates variances and covariances)
  + Not recommended
* Regression substitution = using regression equations for based on nonmissing data to predict expected values for the missing data
  + Also produced biased means, not recommended
* Pattern-matching imputation
  + Still reduces the amount of variation within the data

Stochastic imputation methods

* Auxiliary variables = variables that are not included in the analysis model
  + Useful in imputation because the improve the precision of the imputation model by:
    - Including varialbes that account for the pattern of missing data
    - Improving the prediction of missing values by including variable that are correlated with the variable(s) that have missing data
* Stochastic regression = this method is a variant of the regression-based approach in which a stochastic, or random, value is added to the imputed predicted value
  + Stochastic variables are centered at zero, so they do not systemically change the mean
  + Provide unbiased means and variance estimates
* Expectation maximization (EM) = observed data are used to estimate parameters, which are then used to estimate the missing scores, type of maximum likelihood
  + All maximum likelihood methods are superior to the methods aforementioned
  + In the expectation step, the process is similar to the regression-based imputation. First, starting values for the parameters (e.g., means, covariances) are obtained with available data. Regression methods are used to impute, on the basis of these initial values, the values for the missing data.
  + When this step is completed, in the maximization step, new values for the parameters are calculated with the newly imputed data along with the original observed data.
  + Then the process starts over with the expectation step and continues until the estimates change very little from one iteration to the next (i.e., until the estimates converge)
  + Provides unbiased and efficient parameters
* Full Information Maximum Likelihood (FIML) = from a conceptual standpoint, FIML essentially “borrows” information about the probable values of y on the basis of the conditional expectation of y given x
  + Advantages:
    - The imputation procedure and the analysis are conducted within the same step
    - Unlike EM, FIML produces accurate standard errors by retaining the sample size
    - Unbiased results across a variety of parameter estimates, particularly at small sample sizes
    - Ability to manage missing data and conduct analyses in one step makes this approach much simpler than multiple imputation

#### Data Imputation ####

library(visdat)

vis\_miss(wediko.merge)

library("lme4")

library("plyr")

library(mice)

imputed\_Data <- mice(wediko.merge, m=5, maxit = 50, method = 'pmm', seed = 100)

summary(imputed\_Data)

# m=5 refers to the number of imputed datasets. Five is the default value.

# meth='pmm' refers to the imputation method. In this case we are using

# predictive mean matching as imputation method. Other imputation methods can be

# used, type methods(mice) for a list of the available imputation methods.

# Maxit is the # of iterations for each imputation

#### Anxious RS Model with imputed data ####

wedikofulltest <- with(data = imputed\_Data, exp = lm(obvqvic ~ anxrej \* mspss))

summary(wedikofulltest)

combine <- pool(wedikofulltest)

summary(combine)

#### Angry RS Model with imputed data ####

wedikofulltest.2 <- with(data = imputed\_Data, exp = lm(obvqvic ~ agrej \* mspss))

summary(wedikofulltest.2) ##summary with all 5 regressions

combine.2 <- pool(wedikofulltest.2) ##pooling the 5 different sets

summary(combine.2) ##summary of those

Links:

<https://www.analyticsvidhya.com/blog/2016/03/tutorial-powerful-packages-imputing-missing-values/>

<https://www.r-bloggers.com/imputing-missing-data-with-r-mice-package/>

<https://www.kdnuggets.com/2017/09/missing-data-imputation-using-r.html>