SAS® GLOBAL FORUM 2016

IMAGINE. CREATE. INNOVATE.

Simulation of Imputation Effects Under Different Assumptions

Danny Rithy

#SASGF

Simulation of Imputation Effects Under Different Assumptions

Danny Rithy (drithy@calpoly.edu)

California Polytechnic State University

ABSTRACT

Missing data is something that we cannot always prevent. Data can be missing due to subjects' refusing to answer a sensitive question or in fear of embarrassment. Researchers often assume that their data are “missing completely at random” or “missing at random." Unfortunately, we cannot test whether the mechanism condition is satisfied because missing values cannot be calculated. Alternatively, we can run simulation in SAS® to observe the behaviors of missing data under different assumptions: missing completely at random, missing at random, and being ignorable. We compare the effects from imputation methods if we assign a set of variable of interests to missing. The idea of imputation is to see how efficient substituted values in a data set affect further studies. This lets the audience decide which method(s) would be best to approach a data set when it comes to missing data.

MISSINGNESS MECHANISMS

When a response is missing,

- But does not depend on any variable of interest observed in dataset, it is said to be Missing Completely at Random (MCAR).
- But does not depend on the variable of interest but conditionally depends on some of the variable observed in the dataset, it is said to be Missing at Random (MAR).
- And depends on the variable of interest, it is said to be Not Missing at Random (NMAR).

METHODS

The main objective of imputation is to fill in missing values using variables with non-missing values prior to data analysis.

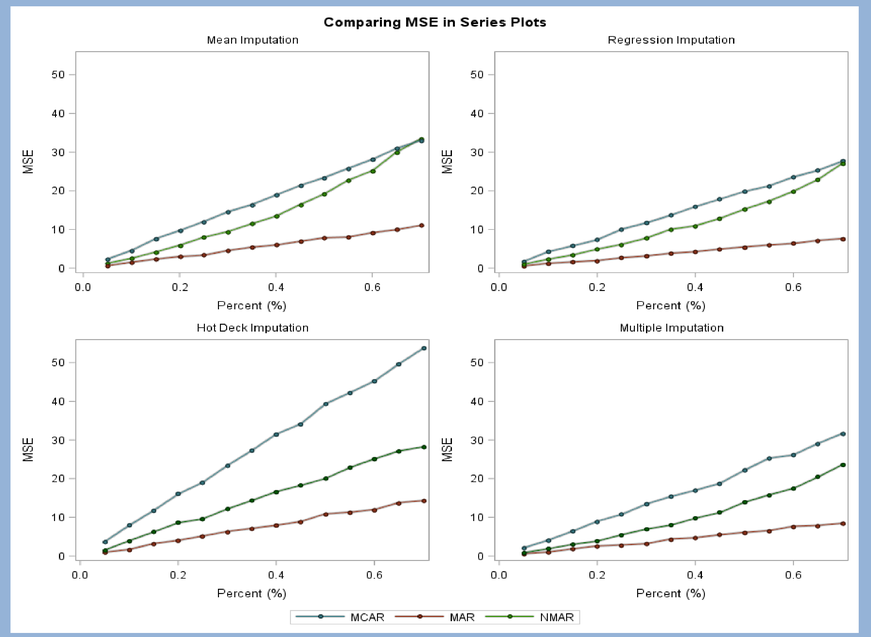
- **Likewise Deletion:** A method that simply removes any observations that contain at least one missing values. Note that this method does not require imputation. This can be done in a DATA step.
- **Mean Imputation:** A method that takes an average value of the variable of interest and imputes missing values. In SAS®, this can be done by using PROC MEANS or PROC SUMMARY.
- **Regression Imputation:** A method that estimates a regression model and predicts missing values given a number of independent variables. In SAS®, this can be done by using PROC REG.
- **Hot Deck Imputation:** A method for dealing with item nonresponse where every missing value is replaced with a value from a similar donor. In SAS®, this can be done by using PROC SURVEYIMPUTE.
- **Multiple Imputation:** A method for repeating the stochastic imputation *m* times where each imputation represents random samples. In SAS®, use the following procedures in order: PROC MI, PROC REG, and PROC MIANALYZE.

SIMULATION STUDY

For the simulation, MathAchieve dataset will be used where mathematics achievement score represents the response variable. The following steps summarizes on how to simulate missing data in SAS® Macro:

1. Assign some respondent’s answers to missing:
   - MCAR is induced by taking a simple random sample of the observations for the variable of interest and assign them to missing.
   - MAR takes a stratified sampling of the observations for the variable of interest and assign them to missing.
   - NMAR uses a condition to assign the variable of interest to missing.
2. Use an imputation method to impute the missing values
3. Calculate Mean Square Error (MSE)

SIMULATION RESULTS

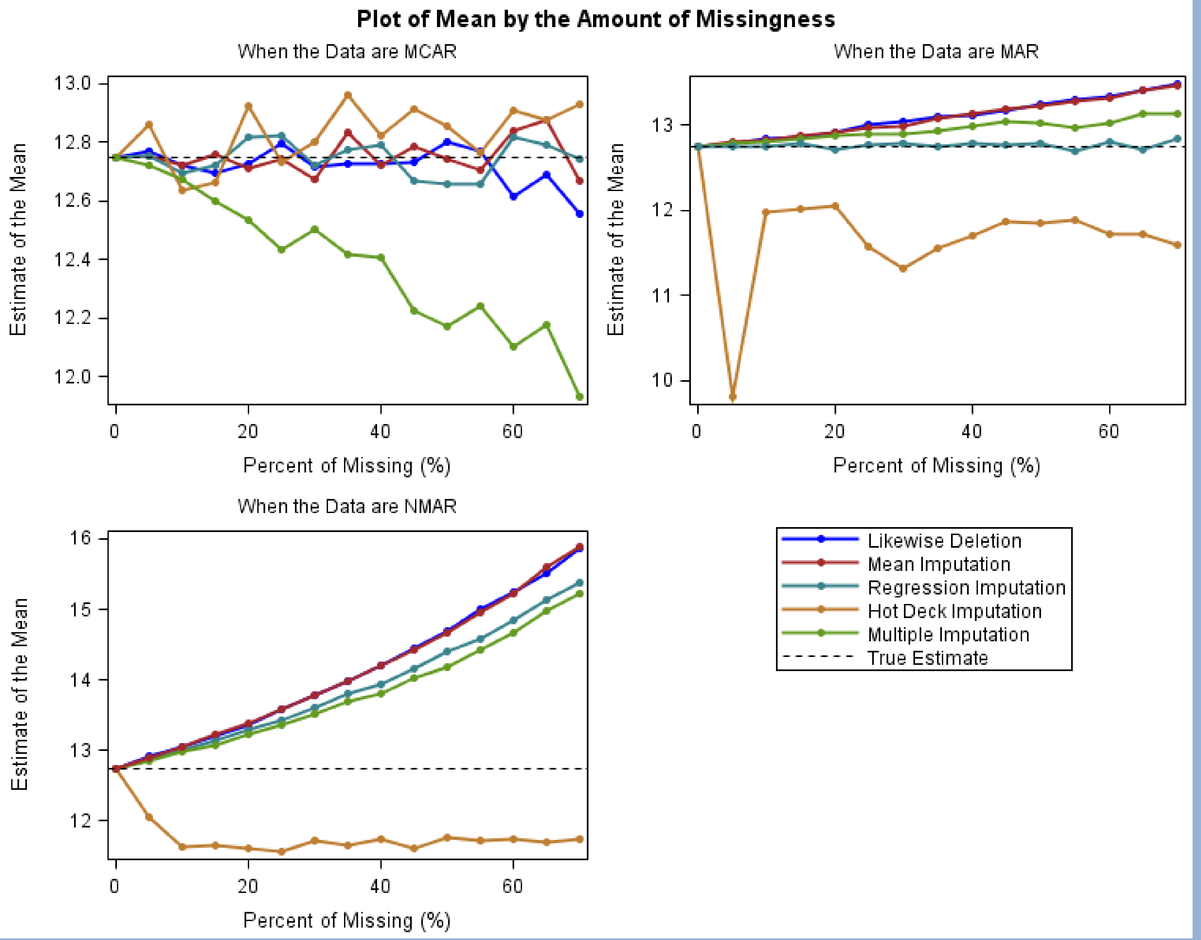


Simulation of Imputation Effects Under Different Assumptions

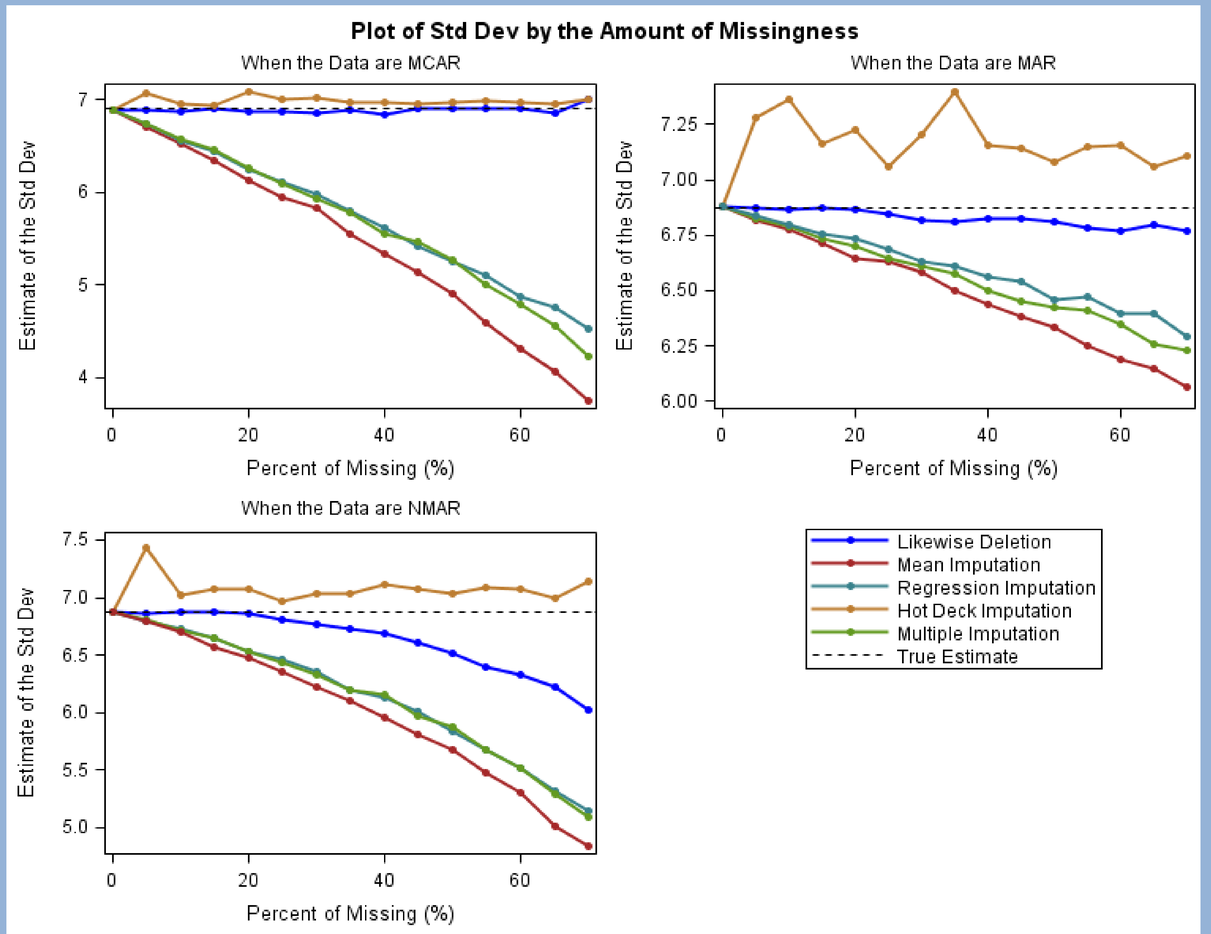
Danny Rithy (drithy@calpoly.edu)

California Polytechnic State University

SIMULATION RESULTS (CONT.)



SIMULATION RESULTS (CONT.)



Simulation of Imputation Effects Under Different Assumptions

Danny Rithy (drithy@calpoly.edu)

California Polytechnic State University

SIMULATION RESULTS (CONT.)

Comparing MSE in Series Plot

- As % missing increases, Mean Square Error (MSE) increases
- When the data are MCAR, MSE is the largest whereas MAR's MSE is the smallest
- MAR has MSE values that are lowest
- MAR is preferable to MCAR and NMAR

Plot of Mean and Standard Deviation by the Amount of Missingness

- MCAR's mean are unbiased estimates and SD are underestimated
- MAR's mean are slightly overestimated and SD are underestimated
- NMAR's mean are overestimated and SD are underestimated

LESSONS LEARNED

- Single variable imputation affects and decreases correlation and variability
- Multiple Imputation takes into account for variability due to stochastic error added per imputation
- When data are MCAR, Mean Imputation is appropriate
- When data are MAR, Regression Imputation & Multiple Imputation are appropriate
- When data are NMAR, advanced statistical models (e.g Structural Equations Modeling or MCMC) are preferable

CONCLUSIONS

It's difficult to determine whether each imputation method is the optimal choice because of unknown circumstances and uncertainty in missing data. The simulation results imply the more missingness in a dataset, the higher MSE will be for all imputation methods.

ACKNOWLEDGEMENTS

Statistics Department at Cal Poly | Senior Project Advisor: Soma Roy

REFERENCES

- Little, Roderick J. A., and Donald B. Rubin. *Statistical Analysis with Missing Data*. New York: Wiley, 1987. Print.
- Scheffer, Judi. "Dealing with Missing Data." *Research Letters in the Information and Mathematical Sciences* (2002): 153-60. Web.

MISSING DATA MECHANISMS IN SAS® MACRO

```
%macro MDprocedure(dat=, percent=, mechanism=, rv=, catevar=, grp1=, grp2=);
 %if &mechanism = "MCAR" %then
  %do;
   /* Use PROC SQL to get the number of samples from a specified dataset (dat) */
   /* Use PROC SURVEYSELECT to randomly select obs from the dataset*/
   /* Implement a DATA step from the outputted dataset in PROC SURVEYSELECT to assign selected obs's
      response variable (rv) to missing */
  %end;
 %else %if &mechanism = "MAR" %then
  %do;
   /* Begin the DATA step and create two SAS® data sets from a specified dataset (dat) to use the if-else statement based
      on categorical variable (catevar, grp1, and grp2) to output into one of two datasets*/
   /* Use PROC SQL to get the number of samples from the selected dataset above */
   /* Use PROC SURVEYSELECT to randomly select observations from the dataset*/
   /* Implement a DATA step from the outputted dataset in PROC SURVEYSELECT to assign selected obs's
      response variable (rv) to missing */
   /* Concatenate two datasets from the DATA step into one data set*/
  %end;
 %else
  %do;
   /* Begin the DATA step using a specified data set (dat) and set a condition in the if-else statement to output into one of
      the two datasets */
   /* Use PROC SQL to get the number of samples using the specified dataset */
   /* Use PROC SURVEYSELECT to randomly select observations from the dataset*/
   /* Implement a DATA step from the outputted dataset in PROC SURVEYSELECT to assign selected obs's
      response variable (rv) to missing */
   /* Concatenate two datasets from the DATA step into one data set*/
  %end;
%mend;
```

SAS® GLOBAL FORUM 2016

IMAGINE. CREATE. INNOVATE.

LAS VEGAS | APRIL 18-21

#SASGF