Handling Data with Three Types of Missing Values:

A Simulation Study

Jennifer Boyko

Advisor: Ofer Harel
Department of Statistics
University of Connecticut
Storrs, CT

May 21, 2013

Outline

What is Missing Data?

- Unobserved values in a data set
- Unit nonresponse
- Item nonresponse
- Dropout
- Intermittent missed follow-up

The Missing Data Problem

- Present in many areas of research
- Small amounts can cause issues (Belin, 2009)
  - Simulation study
  - Removed 5% of the values
  - Complete case analysis: Significance
  - Completed data: No significance

The Missing Data Problem

- Most statistical package defaults use complete case analysis
- Problems include
  - bias
  - inefficiency
  - unrealistic standard errors

Pattern of Missingness

Maps which values are missing in a data set

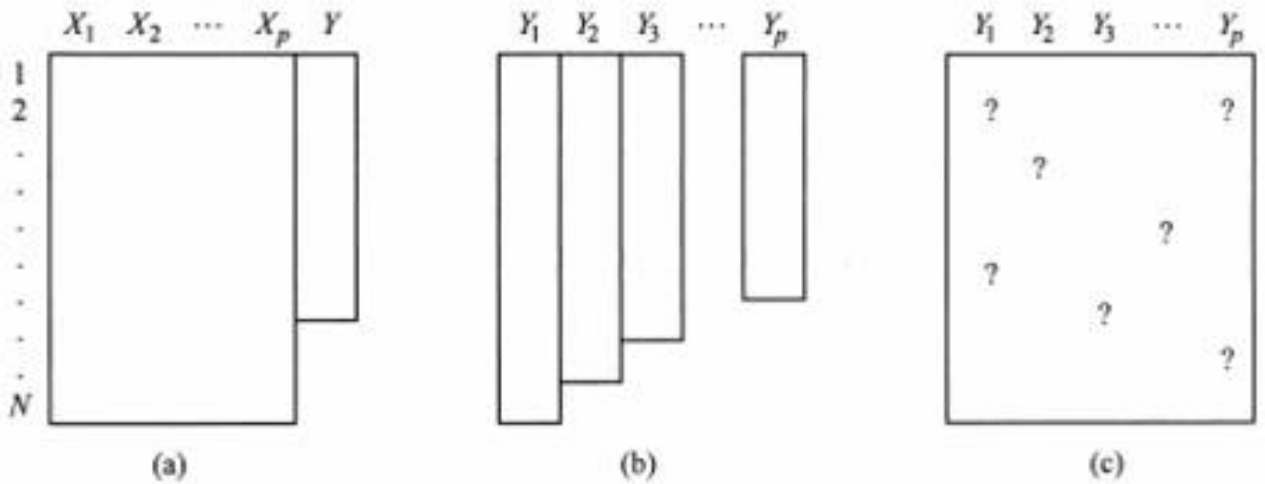


Figure 1. Patterns of nonresponse in rectangular data sets: (a) univariate pattern, (b) monotone pattern, and (c) arbitrary pattern. In each case, rows correspond to observational units and columns correspond to variables.

Figure: Schafer & Graham (2002)

Mechanisms of Missingness

Let

- Y be the complete data partitioned as (Y_{obs}, Y_{mis})
- R be an indicator variable indicating whether or not Y is observed or missing
- θ be the parameter of interest
- ϕ be the parameter of the missing data process
- M^+ be a matrix the same size as Y containing 0's and 1's corresponding to observed values of Y and missing values of Y, respectively

Mechanisms of Missingness

- Missing At Random (MAR)
  - $P(R|Y,\phi) = P(R|Y_{obs},\phi)$
  - Missingness depends on observed values of Y only
- Missing Completely At Random (MCAR)
  - $P(R|Y,\phi) = P(R,\phi)$
  - Missingness not dependent on observed or unobserved values of Y
  - Special case of MAR
- Missing Not At Random (MNAR)
  - Occurs when condition of MAR is violated
  - Missingness is dependent on Y_{mis} or some unobserved covariate

Ignorability

A missing data mechanism is classified as *ignorable* if two conditions are met:

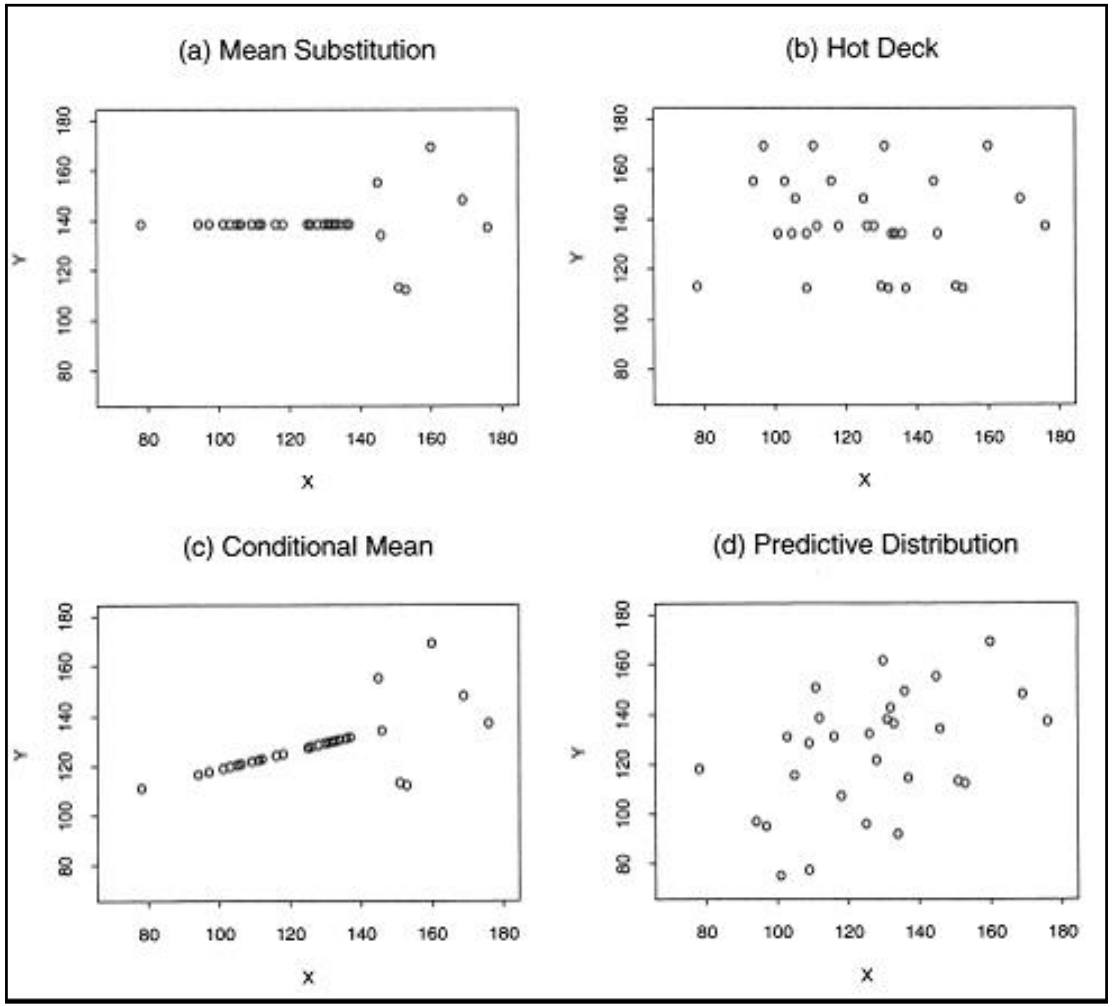
1. The data must be MAR or MCAR
2. θ and ϕ must be distinct
   - $P(\theta, \phi) = P(\theta)P(\phi)$
   - Joint parameter space is the Cartesian cross-product of the individual parameter spaces

Ignorability representes the weakest set of conditions under which the distribution of R does not need to be considered in Bayesian or likelihood-based inference of θ (Rubin, 1976)

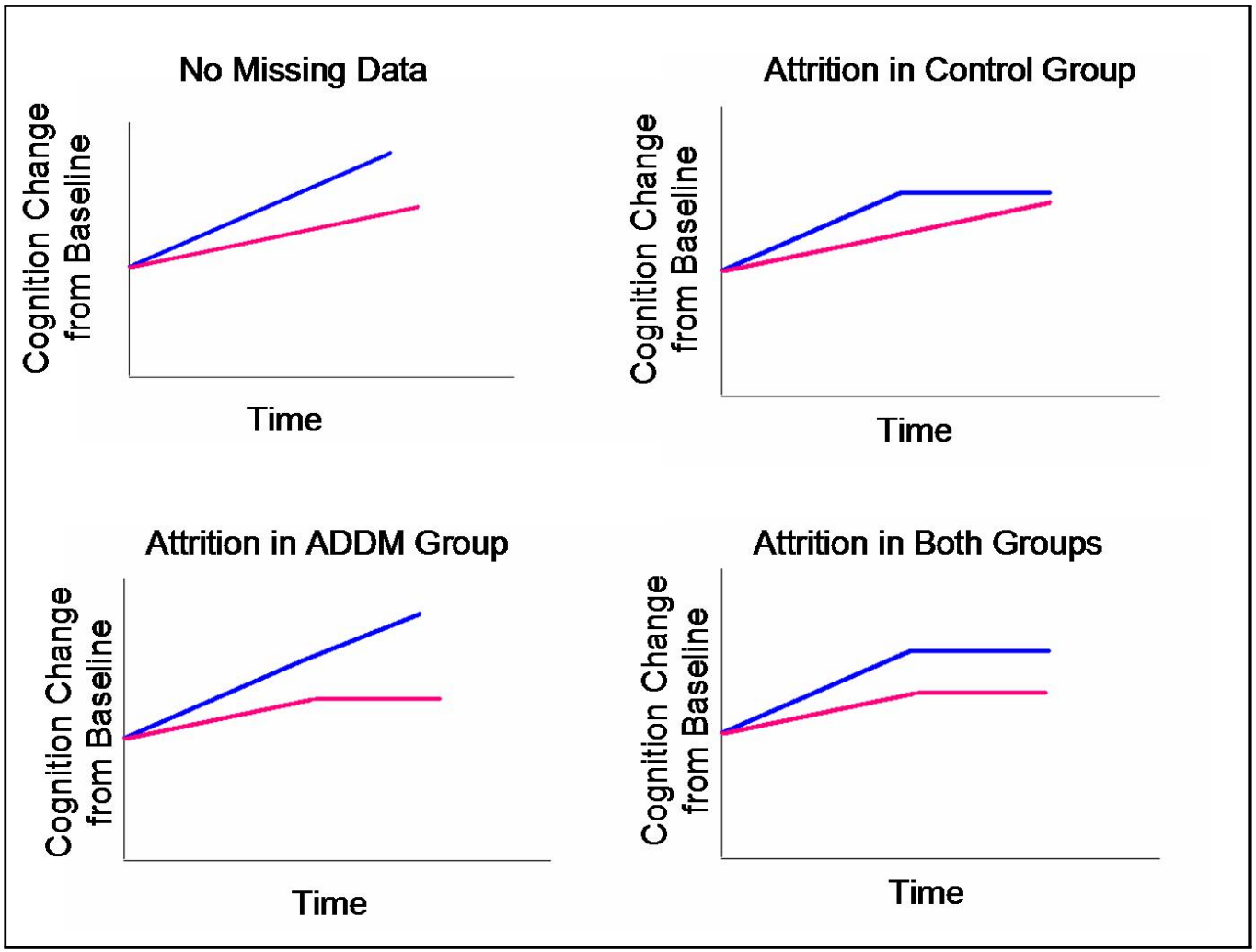
Older Methods

- Complete Case Analysis (CCA)
  - Can produce biased results
  - Default in many statistical packages
  - Loss of information
- Single Imputation
  - Fills in missing values with plausible values
  - Imputing unconditional means
  - Hot deck imputation
  - Conditional mean imputation
  - Last Observation Carried Forward (LOCF)

Single Imputation Methods



LOCF


No Missing Data
Cognition Change from Baseline
Time
Attrition in Control Group
Cognition Change from Baseline
Time
Attrition in ADDM Group
Cognition Change from Baseline
Time
Attrition in Both Groups
Cognition Change from Baseline
Time

Alternative Methods

- Maximum likelihood
- Bayesian
- **Multiple imputation**

Standard Multiple Imputation

Multiple imputation (Rubin, 1987) uses a three step process to analyze incomplete data sets:

1. Imputation
2. Analysis
3. Combination

Imputation Stage

- Idea: fill in $m > 1$ plausible values for the missing data to account for model uncertainty
- Create m complete data sets by drawing from the posterior predictive distribution of the missing values

Analysis Stage

- Analyze each of the m data sets using complete data methods
- Let Q denote the parameter of interest
- Let $\hat{Q}$ be the complete data estimate
- Let U be the variance of Q
- Assumption: $(\hat{Q} - Q)/\sqrt{U} \sim N(0, 1)$

Combination Stage

$$\bar{Q} = \frac{1}{m}\sum_{j=1}^{m}\hat{Q}^{(j)}$$

$$\bar{U} = \frac{1}{m}\sum_{j=1}^{m}U^{(j)}$$

$$B = \frac{1}{m-1}\sum_{j=1}^{m}\left(\hat{Q}^{(j)} - \bar{Q}\right)^2$$

$$T = \bar{U} + (1 + m^{-1})B$$

Combination Stage

$$\frac{(\bar{Q} - Q)}{\sqrt{T}} \sim t_{\nu}$$

$$\nu = (m-1)\left(1 + \frac{\bar{U}}{(1+m^{-1})B}\right)^2$$

Benefits of Multiple Imputation

- Adds variability to the imputed values
- Uses standard data analysis procedures after imputation
- Can be very efficient
- Can use the same set of imputations for several analyses

Two Stage Multiple Imputation

Two stage multiple imputation (Harel, 2009) considers a situation where we can have data missing for two different reasons

- Dropout in a longitudinal study vs. intermittent missing follow-up
- Refusal to answer a question vs. a "don't know" response
- Latent variable vs. missing planned observed values
- Death vs. dropout for other reasons
- Unit nonresponse vs. item nonresponse

Computational Efficiency

Originally developed by Shen (2000) with the intention of improving computational efficiency.

Y_1	Y_2	Y_3	Y_4	Y_5
?				
		?		
				?
?	?	?	?	?
?	?	?	?	?
?	?	?	?	?
?	?	?	?	?

Procedure

Imputation step is broken into two stages:

1. First draw m imputations of Y^A_{mis}
2. Conditioned on Y^A_{mis}, draw n imputations of Y^B_{mis}

Yields a total of mn completed data sets

Two Stage MI Combining Rules

$$\bar{Q} = \frac{1}{mn}\sum_{j=1}^{m}\sum_{k=1}^{n}\hat{Q}^{(j,k)}$$

$$\bar{U} = \frac{1}{mn}\sum_{j=1}^{m}\sum_{k=1}^{n}U^{(j,k)}$$

$$B = \frac{1}{m-1}\sum_{j=1}^{m}\left(\bar{Q}_{j.} - \bar{Q}_{..}\right)^2$$

$$W = \frac{1}{m(n-1)}\sum_{j=1}^{m}\sum_{k=1}^{n}\left(\hat{Q}^{(j,k)} - \bar{Q}_{j.}\right)^2$$

$$T = \bar{U} + (1 + m^{-1})B + (1 - n^{-1})W$$

Two Stage MI Combining Rules

$$\frac{\bar{Q} - Q}{\sqrt{T}} \sim t_{\nu_*}$$

$$\nu_*^{-1} = \frac{1}{m(n-1)} \left(\frac{(1 - 1/n)W}{T} \right)^2 + \frac{1}{m-1} \left(\frac{(1 + 1/m)B}{T} \right)^2$$

Benefits

- Can simplify imputation computationally
- Able to quantify how much missing information is due to each type of missing value which can help in planning future studies
- Allows for different mechanisms of missingness for each type of missing value (one ignorable and one nonignorable type of missing data)

Original Research

1. Multiple imputation in three stages including derivation of combining rules
2. Ignorability and rates of missing information
3. Application of methodology

Benefits

- Extend the benefits of two stage MI to allow for greater specificity regarding the data analysis
- Allows for missing data to be of three different types
- Allows for three different assumptions of the mechanisms of missingness
- Can quantify the variability and missing information due to each type of missing value

Example 1

Example of missing data due to dropout, intermittent missingness, and a missing covariate

X	Y_1	Y_2	Y_3	Y_4
?				
		?		
				?
	?			
?		?		
		?	?	?
		?		?
		?	?	?

$\rightarrow$

X	Y_1	Y_2	Y_3	Y_4
A				
		B		
				C
	B			
A		B		
		C	C	C
		B		C
		C	C	C

Example 2

Example with missing values due to item nonresponse, unit nonresponse, and latent class

Y_1	Y_2	Y_3	Y_4	Y_5
?	?			
?				
?		?		
?				
?				
?	?			
?				?
?	?	?	?	?
?	?	?	?	?
?	?	?	?	?
?	?	?	?	?

$\rightarrow$

Y_1	Y_2	Y_3	Y_4	Y_5
A	B			
A				
A		B		
A				
A				
A	B			
A				B
A	C	C	C	C
A	C	C	C	C
A	C	C	C	C
A	C	C	C	C

Process

Same as standard and two stage MI but with three stages in the imputation step and different combining rules

1. Impute L values of Y^A_{mis}
2. Conditioned on Y^A_{mis}, impute M values of Y^B_{mis}
3. Conditioned on Y^A_{mis} and Y^B_{mis}, impute N values of Y^C_{mis}

Yields a total of LMN completed data sets

A second, but equivalent, method draws simultaneously from the joint distribution of Y^A_{mis}, Y^B_{mis}, and Y^C_{mis}

Three Stage MI Combining Rules

$$\bar{Q} = \frac{1}{LMN} \sum_{l=1}^{L} \sum_{m=1}^{M} \sum_{n=1}^{N} \hat{Q}^{(l,m,n)}$$

$$\bar{U} = \frac{1}{LMN} \sum_{l=1}^{L} \sum_{m=1}^{M} \sum_{n=1}^{N} U^{(l,m,n)}$$

$$B = \frac{1}{L-1} \sum_{l=1}^{L} \left(\bar{Q}_{l..} - \bar{Q}_{...}\right)^2$$

$$W_1 = \frac{1}{L(M-1)} \sum_{l=1}^{L} \sum_{m=1}^{M} \left(\bar{Q}_{lm.} - \bar{Q}_{l..}\right)^2$$

$$W_2 = \frac{1}{LM(N-1)} \sum_{l=1}^{L} \sum_{m=1}^{M} \sum_{n=1}^{N} \left(\hat{Q}^{(l,m,n)} - \bar{Q}_{lm.}\right)^2$$

Three Stage MI Combining Rules

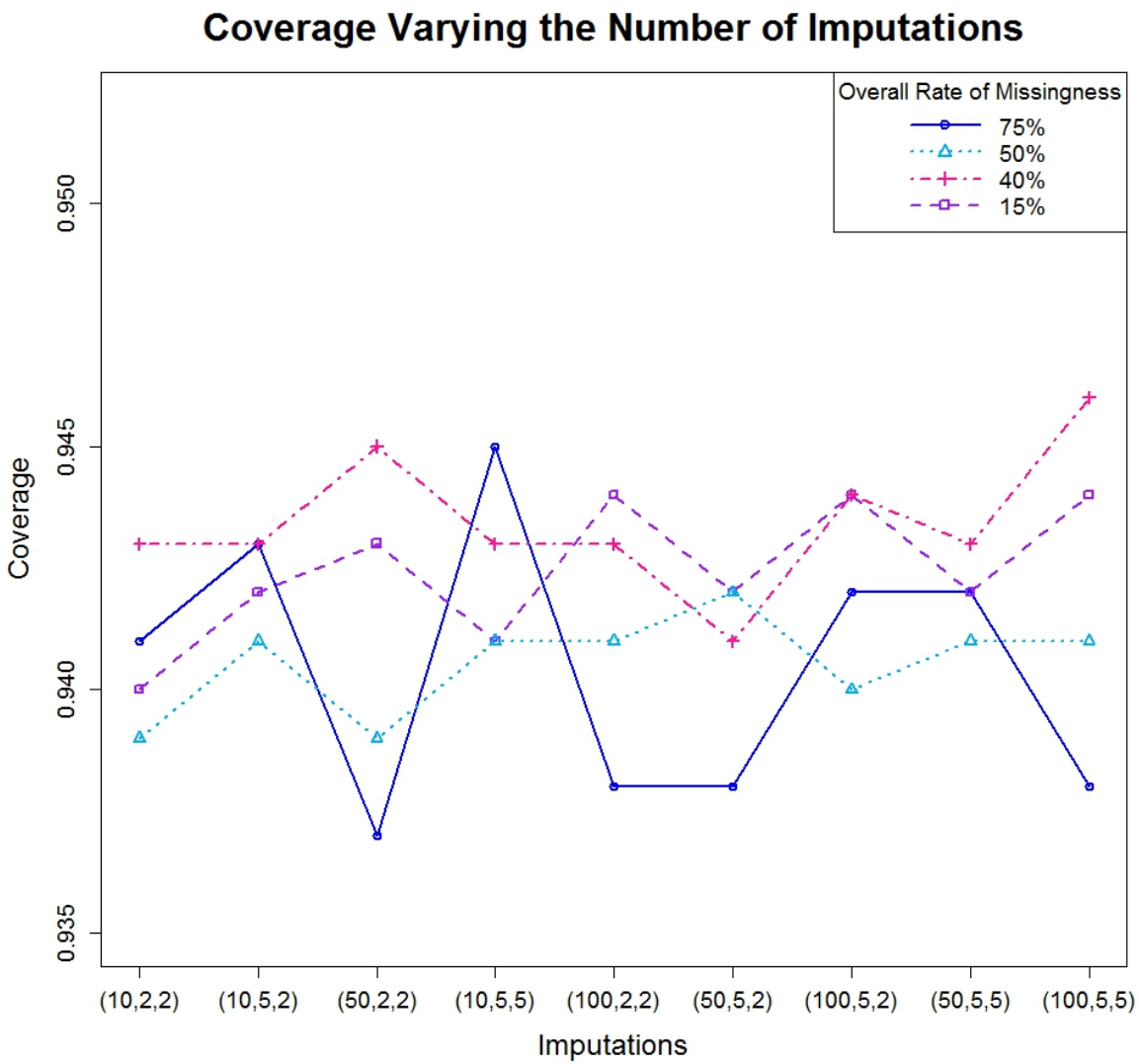
$$T = \bar{U} + (1 + L^{-1})B + (1 - M^{-1})W_1 + (1 - N^{-1})W_2$$

$$\nu^{-1} = \left[\frac{\left(1 + \frac{1}{L}\right) B}{T}\right]^2 (L-1)^{-1} + \left[\frac{\left(1 - \frac{1}{M}\right) W_1}{T}\right]^2 (L(M-1))^{-1} + \left[\frac{\left(1 - \frac{1}{N}\right) W_2}{T}\right]^2 (LM(N-1))^{-1}$$

A Simulation Study

- Let Q be the mean of Y
- Data generated with sample size 100
- Y is based on X so that Y has a known mean of 160
- Missingness:
  - Rate(A) corresponds to MCAR missingness in Y
  - Rate(B) corresponds to missingness in Y based on the upper percentile of X
  - Rate(C) corresponds to missingness in Y based on the lower percentile of X
- Data imputed and analyzed 1000 times for bias, MSE, and coverage

Results


Coverage Varying the Number of Imputations
Overall Rate of Missingness
75%
50%
40%
15%
Coverage
0.935
0.940
0.945
0.950
(10,2,2)
(10,5,2)
(50,2,2)
(10,5,5)
(100,2,2)
(50,5,2)
(100,5,2)
(50,5,5)
(100,5,5)
Imputations

Results

- Coverage not below 93.5% in any of the cases
- No discernible pattern implies that increasing the number of imputations does not significantly improve coverage
- Percent bias was less than 1% in all cases
- MSE ranged from 15 to 28 with the MSE increasing as the percentage of missingness increases

Rates of Missing Information

- Refers to reduction in Fisher Information
- Helps with determination of number of imputations required at each stage
- Small numbers of imputations are required when the main concern is relative efficiency of point estimates
- Estimates for rates of missing information can be noisy for small numbers of imputations

Derivation

- Derived estimates for rates of missing information at each stage and associated variances
- Derived asymptotic distributions for the rates
- Conceptually: look at the reduction in missing information if each of the three types were known

Estimates

$$\hat{\lambda} = \frac{B + (1 - M^{-1})W_1 + (1 - N^{-1})W_2}{\bar{U} + B + (1 - M^{-1})W_1 + (1 - N^{-1})W_2}$$

$$\hat{\lambda}^{B,C|A} = \frac{W_1 + (1 - N^{-1})W_2}{\bar{U} + W_1 + (1 - N^{-1})W_2}$$

$$\hat{\lambda}^{C|A,B} = \frac{W_2}{\bar{U} + W_2}$$

$$\hat{\lambda}^{A} = \hat{\lambda} - \hat{\lambda}^{B,C|A}$$

$$\hat{\lambda}^{B|A} = \hat{\lambda}^{B,C|A} - \hat{\lambda}^{C|A,B}$$

A Simulation Study

- If data is MCAR, rate of missing information is equivalent to percentage of missing values
- All missingness in Y, all three types are MCAR
- Interested in if increasing number of imputations improves estimates

Results

Overall rate of missing information 50%, rate due to type A 25%, rate due to type $B|A$ 20%, rate due to type $C|A, B$ 5%

(L, M, N)	λ	λ^A	$\lambda^{B\|A}$	$\lambda^{C\|A,B}$
(10,2,2)	0.480 (0.894)	0.236 (0.894)	0.194 (0.882)	0.049 (0.907)
(10,5,2)	0.487 (0.892)	0.239 (0.877)	0.197 (0.930)	0.050 (0.939)
(10,5,5)	0.491 (0.892)	0.242 (0.881)	0.198 (0.923)	0.050 (0.917)
(50,2,2)	0.501 (0.933)	0.251 (0.933)	0.200 (0.937)	0.050 (0.923)
(50,5,2)	0.499 (0.921)	0.249 (0.920)	0.199 (0.921)	0.050 (0.943)
(50,5,5)	0.500 (0.910)	0.249 (0.922)	0.201 (0.919)	0.050 (0.894)
(100,2,2)	0.501 (0.911)	0.250 (0.933)	0.201 (0.944)	0.050 (0.931)
(100,5,2)	0.501 (0.930)	0.250 (0.921)	0.200 (0.917)	0.050 (0.926)
(100,5,5)	0.501 (0.923)	0.250 (0.942)	0.201 (0.925)	0.050 (0.869)

Table: Rates of Missing Information

Results

- Estimates stabilized as number of imputations increased
- Coverage mostly increased as number of imputations increased, with some anomalies
- Trend consistent for different percentages of missing values
- If rates of missing information are the main question of interest, need more imputations

Future Work

- Investigate if the order of imputation matters
- Ignorability assumptions
- Applying the work to a data set

Conclusion

- Common problem
- Improper procedures lead to incorrect inferences
- Under ignorability, MI is a good method
- Several types of missingness can be handled with two or three stage MI
- To achieve stable estimates for rates of missing information, number of imputations must increase

Selected Bibliography

Belin, T. (2009). Missing data: what a little can do and what researchers can do in response. *American Journal of Opthalmology* **148,** 820–822.

Harel, O. (2009). *Strategies for Data Analysis with Two Types of Missing Values: From Theory to Application*. Saarbrúcken, Germany: Lambert Academic Publishing.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **64,** 581–592.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Hoboken, New Jersey: John Wiley & Sons, Ltd, 1st ed.

Shen, Z. (2000). *Nested Multiple Imputation*. Ph.D. thesis, Department of Statistics, Harvard University, Cambridge, MA.