NC STATE UNIVERSITY

CSC-591: Foundations of Data Science
T/Th. 12:50-2:05pm. EBI-1005.

Ranga Raju Vatsavai
Chancellors Faculty Excellence Associate Professor in Geospatial Analytics
Department of Computer Science, North Carolina State University (NCSU)
Associate Director, Center for Geospatial Analytics, NCSU
&
Joint Faculty, Oak Ridge National Laboratory (ORNL)

W15: 11/24/15

NC STATE UNIVERSITY

Today

- Missing Data Analysis

NC STATE UNIVERSITY

Missing Data

- Missing data are ubiquitous in many fields
- Classical approaches
  - Discard incomplete cases
  - Fill the missing values
- Most of these techniques require strict assumption about the cause of missing data
- Modern approaches (1970's)
  - Maximum likelihood estimation
  - Multiple imputation

NC STATE UNIVERSITY

General Steps of Missing Data Analysis

- Identify patterns/reasons for missing data
- Understanding the distribution of missing data
- Decide on best methods of analysis

NC STATE UNIVERSITY

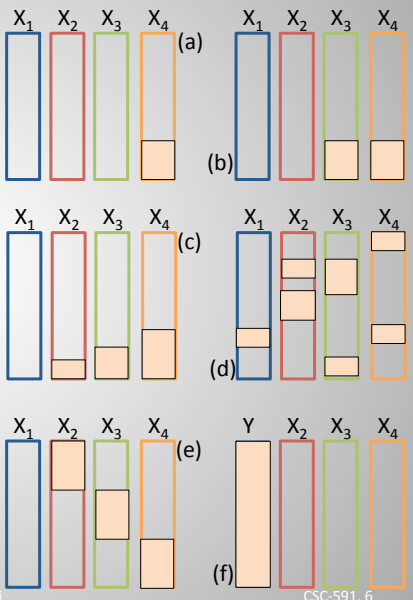
Missing Data Patterns

- Missing data pattern refers to the configuration of observed and missing values
  - Simply describes the location of the "holes" in the data and doesn't explain why the data are missing
- Missing data mechanisms describe possible relationship between measured variables and probability of missing data
  - Represent generic mathematical relationships between the data and missingness

NC STATE UNIVERSITY

Missing Data Patterns

- (a) Univariate pattern
- (b) Unit nonresponse pattern
- (c) Monotone pattern
- (d) General pattern
- (e) Planned missing pattern
- (f) Latent variable pattern



NC STATE UNIVERSITY

Missing Data Patterns

- (a) Univariate pattern
  - Missing values are isolated to a single variable
  - Relatively rare, but can arise experimental studies
- (b) Unit nonresponse pattern
  - Some characteristics available to all participants (e.g., social security), but some questions may not be answered by participants
- (c) Monotone pattern
  - Typically associated with longitudinal study where participants drop out and never return
- (d) General pattern
  - Most common
- (e) Planned missing pattern
  - three-form design is to distribute questionnaires across different forms and administer a subset of the forms to each respondent
- (f) Latent variable pattern
  - Value of latent variables are missing for entire sample

NC STATE UNIVERSITY

Missing Data Theory

- Rubin, et. al. introduced a classification system for missing data problems
  - Introduced three so-called missing data mechanisms that describe how the probability of a missing value relates to the data, if at all.
    - **MAR**: data are missing at random
    - **MCAR**: missing completely at random
    - **MNAR**: missing not at random

NC STATE UNIVERSITY

Missing at Random Data (MAR)

- Data are missing at random (MAR) when the probability of missing data on a variable *Y* is related to some other measured variable (or variables) in the analysis model but not to the values of *Y* itself
- Here "random" doesn't mean data is missing in haphazard fashion
- MAR actually means that a systematic relationship exists between one or more measured variables and the probability of missing data
  - But there is no way of confirming it

NC STATE UNIVERSITY

Example

Rating after 6-month probation

Employee Selection

IQ

Evaluate after 6-months

If company didn't hire applicants that scored in the lower quartile of IQ distribution

the probability of a missing job performance rating is solely a function of IQ scores and is unrelated to an individual's job performance

	Job performance ratings			
IQ	Complete	MCAR	MAR	MNAR
78	9	—	—	9
84	13	13	—	13
84	10	—	—	10
85	8	8	—	—
87	7	7	—	—
91	7	7	7	—
92	9	9	9	9
94	9	9	9	9
94	11	11	11	11
96	7	—	7	—
99	7	7	7	—
105	10	10	10	10
105	11	11	11	11
106	15	15	15	15
108	10	10	10	10
112	10	—	10	10
113	12	12	12	12
115	14	14	14	14
118	16	16	16	16
134	12	—	12	12

NC STATE UNIVERSITY

Missing Completely At Random (MCAR)

- MCAR: the probability of missing data on a variable *Y* is unrelated to other measured variables and is unrelated to the values of *Y* itself
- In principle, it is possible to verify that missing data is MCAR
- The definition of MCAR requires that the observed data are a simple random sample of the hypothetically complete data set. This implies that the cases with observed job performance ratings should be no different from the cases that are missing their performance evaluations, on average.
  - To test this idea, you can separate the missing and complete cases and examine group mean differences on the IQ variable.
  - If the missing data patterns are randomly equivalent (i.e., the data are MCAR), then the IQ means should be the same, within sampling error.
  - From the example data, the complete cases have an IQ mean of 99.73, and the missing cases have a mean of 100.80.

NC STATE UNIVERSITY

Example

Rating after 6-month probation

Employee Selection

Hire All

Evaluate after 6-months

Ratings were deleted based on random number

Random numbers were not related to IQ or rating, consequently the probability of a missing job performance rating is not related to IQ scores or an individual's job performance

	Job performance ratings			
IQ	Complete	MCAR	MAR	MNAR
78	9	—	—	9
84	13	13	—	13
84	10	—	—	10
85	8	8	—	—
87	7	7	—	—
91	7	7	7	—
92	9	9	9	9
94	9	9	9	9
94	11	11	11	11
96	7	—	7	—
99	7	7	7	—
105	10	10	10	10
105	11	11	11	11
106	15	15	15	15
108	10	10	10	10
112	10	—	10	10
113	12	12	12	12
115	14	14	14	14
118	16	16	16	16
134	12	—	12	12

NC STATE UNIVERSITY

Missing Not At Random (MNAR)

- Data are missing not at random (MNAR) when the probability of missing data on a variable *Y* is related to the values of *Y* itself, even after controlling for other variables.
- Like the MAR mechanism, there is no way to verify that scores are MNAR without knowing the values of the missing variables.

NC STATE UNIVERSITY

Example

Rating after 6-month probation

Employee Selection

Hire All

Evaluate after 6-months

Suppose that the company hired all 20 applicants and subsequently terminated a number of individuals for poor performance prior to their 6-month evaluation.

probability of a missing job performance rating is dependent on one's job performance, even after controlling for IQ

	Job performance ratings			
IQ	Complete	MCAR	MAR	MNAR
78	9	—	—	9
84	13	13	—	13
84	10	—	—	10
85	8	8	—	—
87	7	7	—	—
91	7	7	7	—
92	9	9	9	9
94	9	9	9	9
94	11	11	11	11
96	7	—	7	—
99	7	7	7	—
105	10	10	10	10
105	11	11	11	11
106	15	15	15	15
108	10	10	10	10
112	10	—	10	10
113	12	12	12	12
115	14	14	14	14
118	16	16	16	16
134	12	—	12	12

NC STATE UNIVERSITY

The Distribution of Missing Data

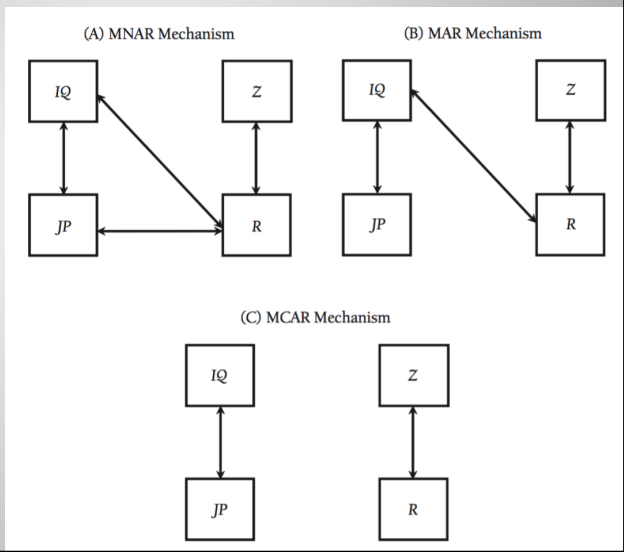
- Key idea behind Rubin's theory is that missingness is a variable that has a probability distribution. Specifically, Rubin defined a binary variable R to denote weather a score on a particular variable is missing (i.e., r=1 if score is observed, and r=0 if a value is missing)
- Defining the missing data as a variable implies that there is a probability distribution that governs whether *R* takes on a value of zero or one
- For MNAR, $p(R|Y_{obs}, Y_{mis}, \varphi)$
- For MAR, $p(R|Y_{obs}, \varphi)$
- For MCAR, $p(R|\varphi)$

Job performance		
Complete	MAR	Indicator
9	—	0
13	—	0
10	—	0
8	—	0
7	—	0
7	7	1
9	9	1
9	9	1
11	11	1
7	7	1
7	7	1
10	10	1
11	11	1
15	15	1
10	10	1
10	10	1
12	12	1
14	14	1
16	16	1
12	12	1

NC STATE UNIVERSITY

Graphical Representation of Rubin's Missing Data Mechanisms

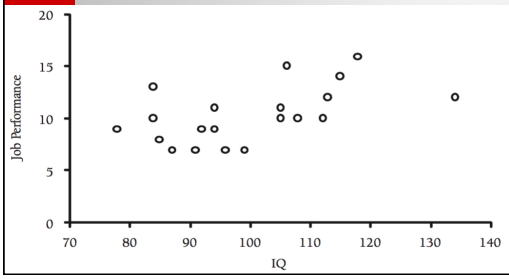
- Figure represents a bivariate scenario in which IQ scores are completely observed and JP scores are missing for some individuals
- Double headed arrows represent generic statistical associations and φ is a parameter that governs probability of a 0 or 1 on missing data indicator R.
- Z represents collection of unmeasured variables



NC STATE UNIVERSITY

Traditional Methods for Dealing with Missing Data

- Applicants take IQ and answers wellbeing q's during interview, and company subsequently hires upper half of IQ, a supervisor rates JP following a 6-month probationary period
- Note that the job performance scores are missing at random (MAR) because they are systematically missing as a function of IQ



Complete data		Missing data
IQ	Job performance	Job Performance
78	9	—
84	13	—
84	10	—
85	8	—
87	7	—
91	7	—
92	9	—
94	9	—
94	11	—
96	7	—
99	7	7
105	10	10
105	11	11
106	15	15
108	10	10
112	10	10
113	12	12
115	14	14
118	16	16
134	12	12

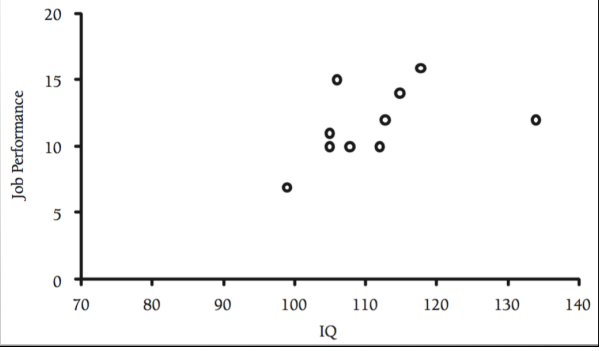
NC STATE UNIVERSITY

Listwise Deletion

- Also known as complete-case analysis discards the data for any case that has one or more missing values
  - Convenient and produces a common set of cases for all analysis
  - The primary problem with listwise deletion is that it requires MCAR data and can produce distorted parameter estimates when this assumption does not hold
  - Because IQ scores and job performance ratings are positively correlated, listwise deletion also excludes cases from the lower tail of the job performance distribution (i.e., the cases with low IQ scores). Not surprisingly, the remaining cases are unrepresentative of the hypothetically complete data set because they have systematically higher scores on both variables. Consequently, the listwise deletion mean estimates are too high.

Interesting Observation: Listwise deletion can produce unbiased estimates of regression slopes un- der any missing data mechanism, provided that missingness is a function of a predictor variable and not the outcome variable

11/25/15



NC STATE UNIVERSITY

Pairwise Deletion

- Pairwise deletion (also known as available-case analysis) attempts to mitigate the loss of data by eliminating cases on an analysis-by-analysis basis.

Gender	8th grade math test score	12th grade math score
F	45	.
M	.	99
F	55	86
F	85	88
F	80	75
.	81	82
F	75	80
M	95	.
M	86	90
F	70	75
F	85	.

Gender	8th grade math test score	12th grade math score
F	45	.
M	.	99
F	55	86
F	85	88
F	80	75
.	81	82
F	75	80
M	95	.
M	86	90
F	70	75
F	85	.

NC STATE UNIVERSITY

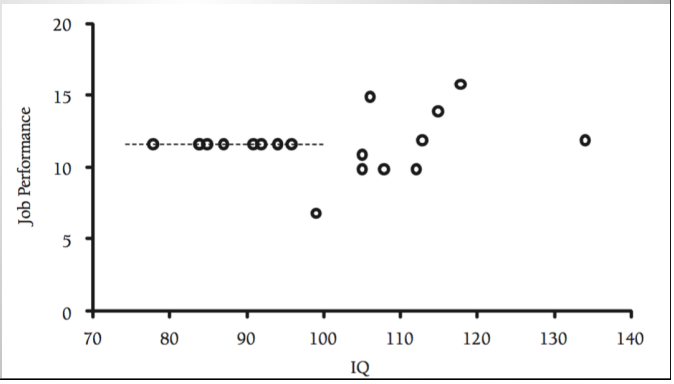
Single Imputation Methods

- Single imputation methods which impute (i.e., fill in) the data prior to analysis
  - Arithmetic mean imputation
  - Conditional mean imputation
  - Regression imputation
- Makes use of data that deletion approaches would otherwise discard
- However, produce biased parameter estimates
  - Stochastic regression imputation is the sole exception because it is the only approach that produces unbiased parameter estimates with MAR data

NC STATE UNIVERSITY

Single Imputation Methods

- **Arithmetic mean imputation** (also referred to as mean substitution and unconditional mean imputation) takes the seemingly appealing tack of filling in the missing values with the arithmetic mean of the available cases.

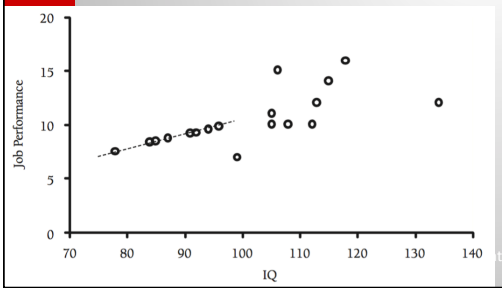


NC STATE UNIVERSITY

Regression Imputation

- Regression imputation (also known as conditional mean imputation) replaces missing values with predicted scores from a regression equation.
- Form employee selection data example, use 10 complete cases to estimate the regression of JP rating on IQ, the resulting regression eq: $JP_i = B_0 + B_1(IQ_i) = -2.065 + 0.123\ IQ_i$.

Overestimates correlations and R^2 static even when the data are MCAR



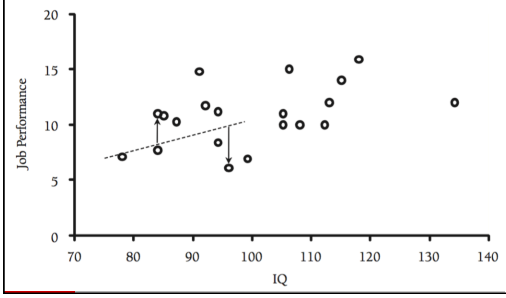
IQ	Job performance	Predicted score	Random residual	Stochastic imputation
78	—	7.53	−0.35	7.18
84	—	8.27	2.70	10.97
84	—	8.27	−0.59	7.68
85	—	8.39	2.39	10.78
87	—	8.64	1.64	10.28
91	—	9.13	5.77	14.90
92	—	9.25	2.47	11.72
94	—	9.50	−1.04	8.46
94	—	9.50	1.69	11.19
96	—	9.74	−3.58	6.16
99	7	—	—	—
105	10	—	—	—
105	11	—	—	—
106	15	—	—	—
108	10	—	—	—
112	10	—	—	—
113	12	—	—	—
115	14	—	—	—
118	16	—	—	—
134	12	—	—	—

NC STATE UNIVERSITY

Stochastic Regression Imputation

- **Stochastic regression imputation** also uses regression equations to predict the incomplete variables from the complete variables, but it takes the extra step of augmenting each predicted score with a normally distributed residual term. Adding residuals to the imputed values restores lost variability to the data and effectively eliminates the biases associated with standard re-gression imputation schemes.
- Form employee selection data example, use 10 complete cases to estimate the regression of JP rating on IQ, the resulting regression eq: $JP_i = B_0 + B_1(IQ_i) = -2.065 + 0.123\ IQ_i + z_i$. Zi is a normal distribution with mean 0 variance = 6.65 (estimated from complete-case analysis). Use Monte Carlo simulation to generate 10 scores form normal distribution z_i.

Produces unbiased parameter estimates for MAR data



IQ	Job performance	Predicted score	Random residual	Stochastic imputation
78	—	7.53	–0.35	7.18
84	—	8.27	2.70	10.97
94	—	8.27	–0.59	7.68
35	—	8.39	2.39	10.78
37	—	8.64	1.64	10.28
91	—	9.13	5.77	14.90
92	—	9.25	2.47	11.72
94	—	9.50	–1.04	8.46
94	—	9.50	1.69	11.19
96	—	9.74	–3.58	6.16
99	7	—	—	—
05	10	—	—	—
05	11	—	—	—
06	15	—	—	—
08	10	—	—	—
12	10	—	—	—
13	12	—	—	—
15	14	—	—	—
18	16	—	—	—
34	12	—	—	—

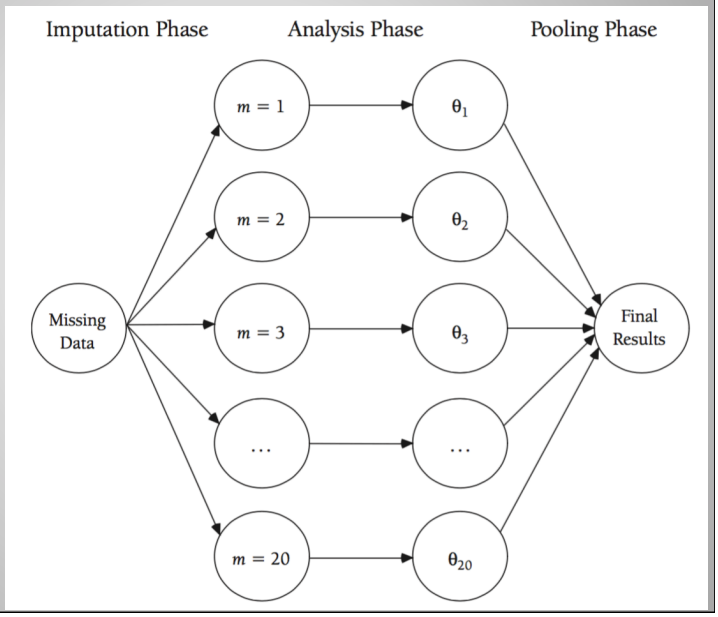
NC STATE UNIVERSITY

Hot-deck Imputation

- Hot-deck imputation is a collection of techniques that impute the missing values with scores from “similar” respondents.
- For example, consider a general population survey in which some respondents refuse to disclose their income. The hot-deck procedure classifies respondents into cells based on demographic characteristics such as gender, age, race, and marital status. It then replaces the missing val- ues with a random draw from the income distribution of respondents that shared the same constellation of demographic characteristics as the individual with missing data.

NC STATE UNIVERSITY

Multiple Imputation



NC STATE UNIVERSITY

Acknowledgements

- Schafer,JosephL.,JohnW.Graham. 2002.“Missing Data: Our View of the State of the Art.” Psychological Methods.
- Enders, Craig. 2010. Applied Missing Data Analysis. Guilford Press: New York.
- Little,RoderickJ.,DonaldRubin. 2002.StatisticalAnalysis with Missing Data. John Wiley & Sons, Inc: Hoboken.