An Approach to Probabilistic Modeling for Mixed Data Types with Missing Values

SCEA Luncheon, February 1, 2006

Outline

- Surveys and the problem of missing data
- Imputation
- Statistical modeling of imputed datasets
- Analysis of imputed datasets

Purpose

The aim of this presentation is to address one method used to analyze survey data. In particular, it attempts to address the following questions:

- Given a survey not designed for analysis, is there information that can be discerned?
- How can a dataset be developed from a survey when the responses to some questions are missing?
- What methods can be used to analyze this dataset to draw meaningful conclusions?

Problem

- A survey is created, distributed to, and returned from participants.
- Though the questions on the survey are important and gather high-quality input from the responders, the survey is poorly designed, and the questions are framed in a manner that does not lend them to be easily analyzed using traditional methods for extracting statistical information from survey results.
- What can be done to salvage the data?

Sample survey questions

- Which of the following are security measures you employ at the building entrance?

  [ ] Security Guards [ ] Thumbprint Reader [ ] PIN number entry pad

  [ ] External Cameras [ ] Metal Detectors [ ] Key

  [ ] Badge Reader [ ] Receptionist [ ] None

- At what type of area is the building located?

  [ ] Downtown (high-rise) [ ] Rural [ ] Outside the US

  [ ] Suburban [ ] Within 500 meters of a government building

- How many floors does the building have?

  [ ] 1 [ ] 6-10

  [ ] 2-5 [ ] 11+

- How many employees work inside the building?

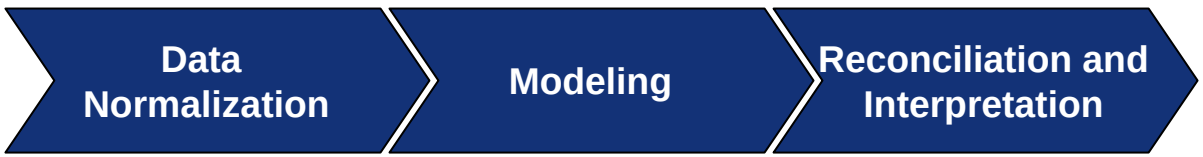
  [ ] 1-25 [ ] 51-100 [ ] 251-500 [ ] More than 1000

  [ ] 26-50 [ ] 101-250 [ ] 500-1000

How do we get from raw survey data to meaningful results?

This process can generally be split into three steps:

1. Data Normalization
2. Modeling
3. Reconciliation and Interpretation



Missing Data

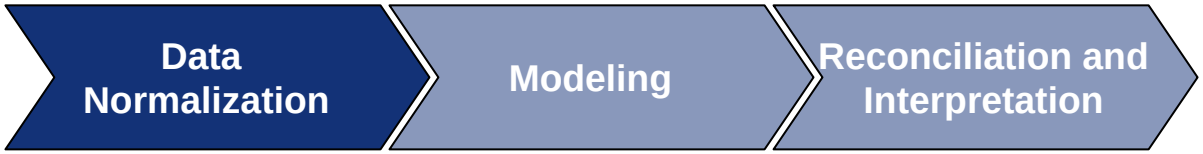
- In a perfect world, every participant in a survey would answer every question. In reality, this is rarely the case.

- Missing data often occurs due to:
  - Non-responders – participants who provide no answers to any questions
  - Partial responders – participants who respond to most questions but provide no response at all to others

- Missing data typically follows one of three patterns:
  - Monotone – Missing variables depend on prior variables; If the value for variable Y_i is missing, then for all $j>i$, Y_j is also missing
  - Missing at Random (MAR) – Missing variables are related to the values of the variables from which data is available, not the values of the missing variables
  - Missing Completely at Random (MCAR) – Missing variables do not depend on the values of any variable in the dataset

Missing Data (Continued)

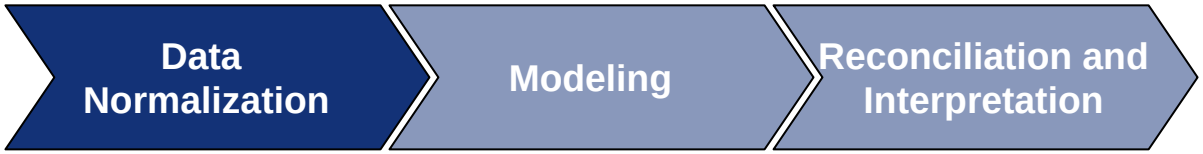
- Why is missing data a problem?
  - Missing data leads to bias, information loss, and a reduction in the number of observations
  - Datasets with missing values typically do not allow standard statistical approaches to be used on them unless the approaches are altered in some way or another
  - Altering a statistical method can be a difficult process, and once the method is tailored to the dataset, it may not be easily applied to other datasets
- How can the missing data be treated?
  - Ignore the missing values – Though a simple and easy fix to the problem, valuable data is lost, and bias can be introduced
  - Impute values for the missing data – Allows the dataset to be ‘repaired’ so that standard statistical models can be applied



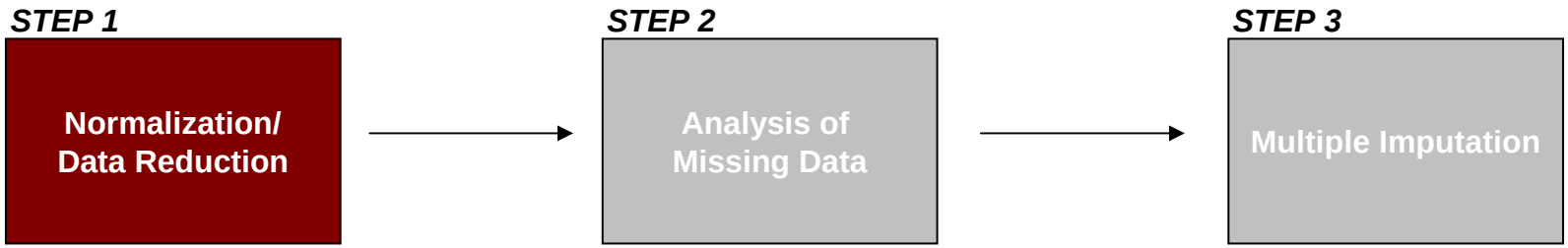
Data Normalization

This first step in processing the survey data prepares the dataset for modeling through:

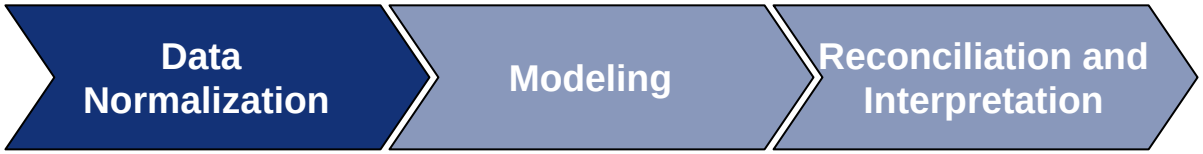
- Visualizing the data and analyzing the variables to be used in the model
- Choosing an imputation method
- Imputing data



Dataset Modification – Step 1

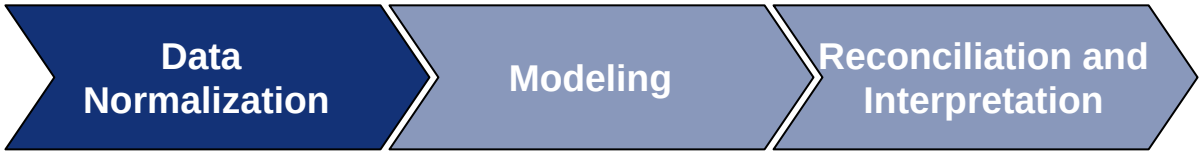


- The questions in the survey are analyzed to identify response and explanatory variables of interest
- Binary data is coded with 0(-)/1(+)
- The questions are further examined and classified to reduce the number of variables
- Survey records of non-responders are often deleted from the dataset at this step

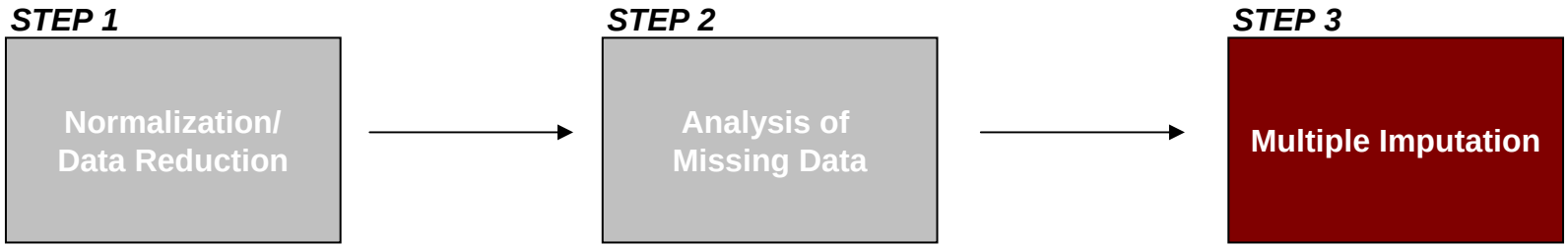


Dataset Modification – Step 2

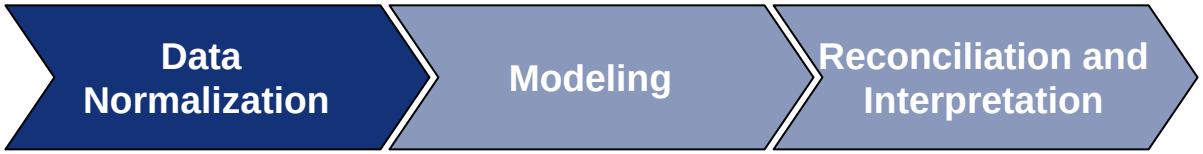
- The data is analyzed to determine the best approach for dealing with the missing data
- Possible methods of handling missing data:
  - Ignore missing values: Deleting the survey responses of those who failed to answer all questions
  - Single Imputation: Imputing one value for the variable (e.g., the mean for the variable) in each missing location
  - **Multiple Imputation (MI):** Multiple datasets are created by imputing several values for each missing value
- MI is most often the preferred method of imputation



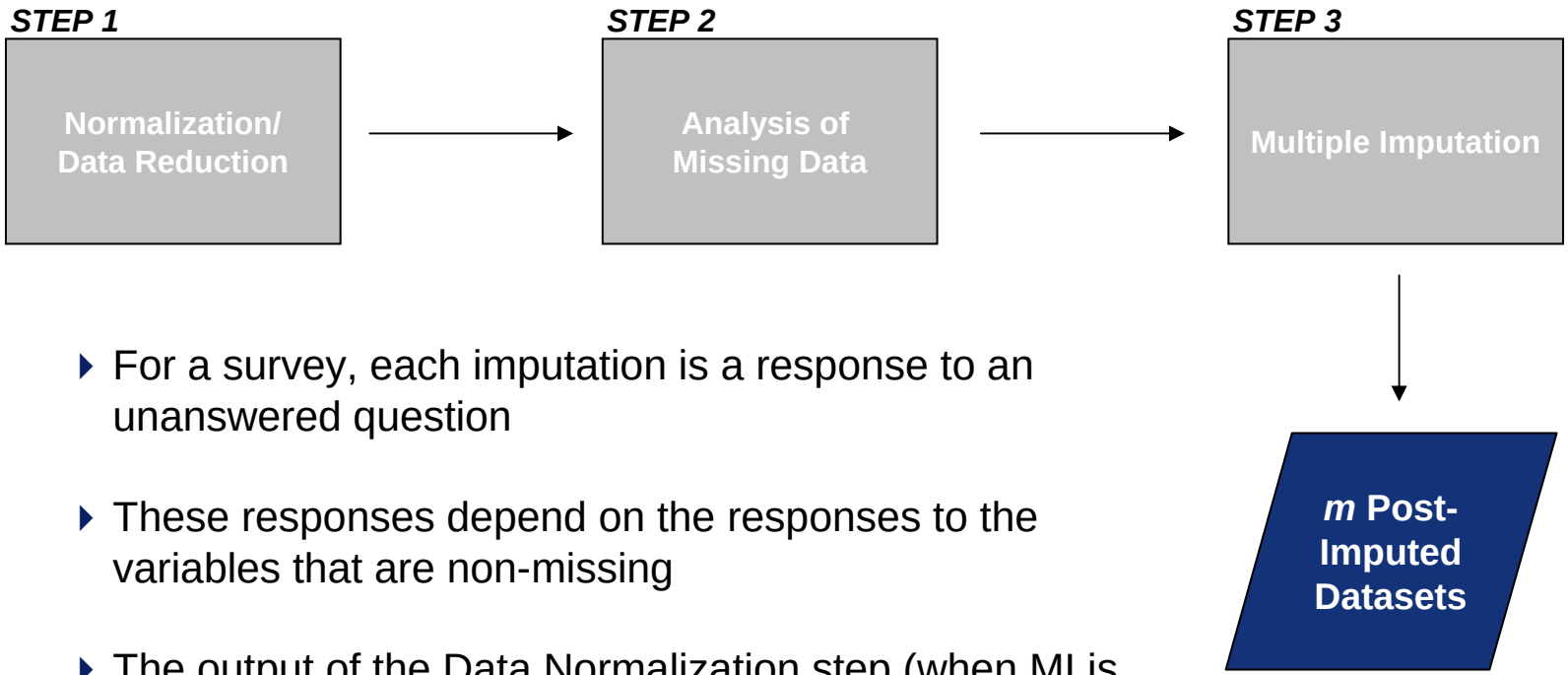
Dataset Modification – Step 3



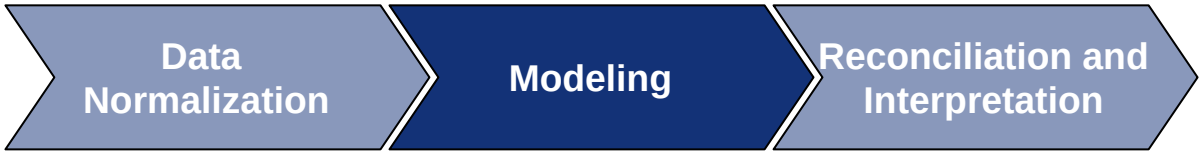
- Multiple Imputation fills in missing data *m* times thereby creating *m* different datasets
  - Imputed values vary across datasets; can be +/-; no numerical limit
  - *m* generally ranges between 2 and 10; *m = 5* is most commonly used in practice
- Three primary methods for generating imputed values:
  - Regression method
  - Propensity score method
  - **Monte Carlo Markov Chain (MCMC) method**
- MCMC method:
  - 'Markov Chains' are simulated to create the distribution of interest
  - Imputed values are drawn from this distribution



Data Normalization Output



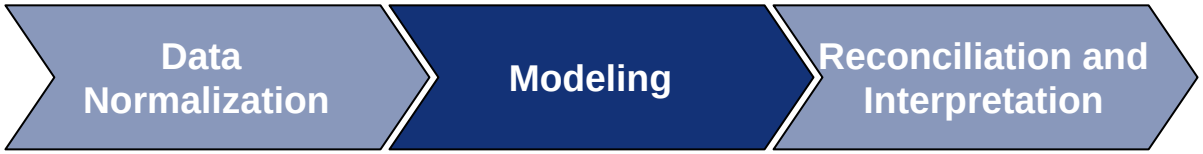
- For a survey, each imputation is a response to an unanswered question
- These responses depend on the responses to the variables that are non-missing
- The output of the Data Normalization step (when MI is used) is *m* complete post-imputation datasets across all variables of interest



Modeling

With complete, imputed datasets created, the next step is to employ standard modeling techniques on them to:

- Investigate interactions between pairs of response variables (log-linear analysis), and
- Identify relationships between the response variables and predictors



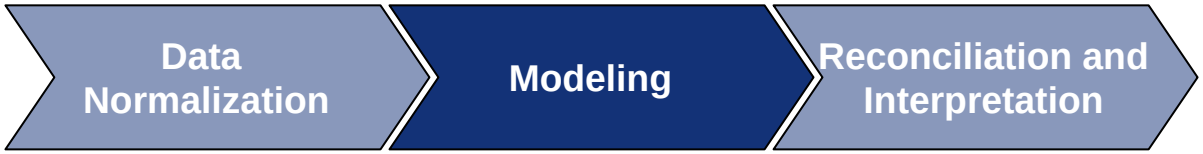
Modeling Process

- Log-linear modeling is first used to identify the interaction effects between the response variables within a single question in the survey.

$$y_{ij} = \lambda^{A} + \lambda^{B} + \ldots + \lambda^{AB} + \ldots + \lambda^{ABC}$$

- A log-linear analysis of these chosen response variables is then performed with respect to the variables in other questions to determine the most significant variables and interactions.

- This analysis yields *m* different log-linear models for each ‘chosen’ response variable

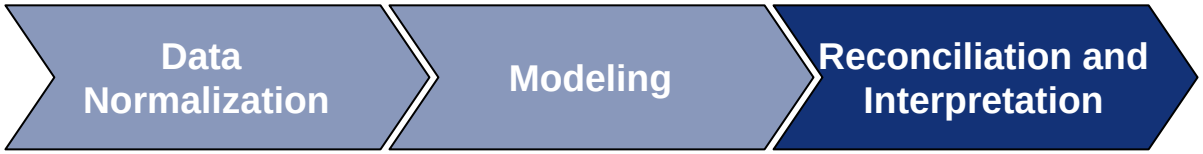


Modeling Process

- Logit models are used to analyze response problems between binary variables and multiple predictor variables to determine the probability of a particular response

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta^A + \beta^B + \beta^{AB}$$

- The chosen response variables and other significant variables identified in the log-linear analysis are pooled and included in a multivariate logistic regression model
- A logit model using these significant variables is run across the imputed datasets
- The same variables are used in the logistic regression for each imputed dataset, thereby generating *m* logit models



Reconciliation and Interpretation

- Data inferences set forth by J.L. Schafer and D.B. Rubin are used to consolidate the *m* logit models and determine a single final logit model for each response variable
- These logit models are then analyzed to gather insight into individual probabilities

Rubin's Rules

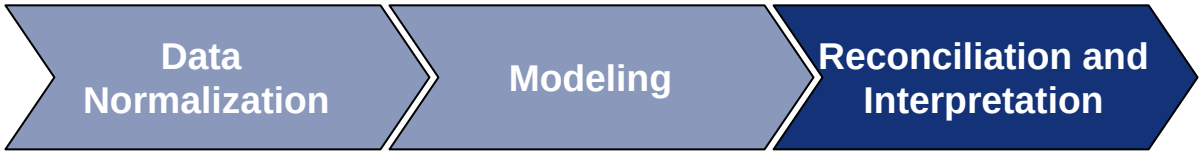
- Rubin's Rules are used to synthesize the results from the *m* logit models to generate a single logit model based on the survey data.

Given:

Parameter Q_i from logit model m_i	$\hat{Q}_i = \begin{pmatrix} \beta_0 \\ \beta_A \\ \beta_B \\ \beta_{AB} \end{pmatrix}$
Covariance matrix for each Q_i	$\hat{\mathbf{U}}_i = \text{cov}\,matrix$

These rules consist of the following:

Point Estimate for the Parameter Q_i	$\overline{Q} = \frac{1}{m}\sum_{i=1}^{m}\hat{Q}_i$
Within-Imputation Covariance Matrix	$\overline{\mathbf{U}} = \frac{1}{m}\sum_{i=1}^{m}\hat{\mathbf{U}}_i$
Between-Imputation Covariance Matrix	$\mathbf{B} = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{Q}_i - \overline{Q})(\hat{Q}_i - \overline{Q})^T$
Total Covariance Matrix	$\mathbf{T} = (1+r_1)\overline{\mathbf{U}}$, where $r_1 = (1+\frac{1}{m})tr(\mathbf{B}\mathbf{U}^{-1})/k$
F statistic with *k* degrees of freedom for testing $Q=Q_0$	$F = (\overline{\mathbf{Q}} - \mathbf{Q}_0)^T \mathbf{T}^{-1}(\overline{\mathbf{Q}} - \mathbf{Q}_0)/k$



Analysis

- After Rubin's Rules are used to generate one logit model for each of the chosen response variables, this model is evaluated
  - All significant variables (the predictors) are set to zero, and each predictor is varied through its possible states (for discrete-valued predictors) or ranges (for continuous-valued predictors) to determine the probability of each response occurring given the state or value of the predictor being varied
  - The relative contribution of each continuous-valued predictor is also calculated
- The process is repeated for each response variable of interest.

Sample Analysis

Model: logit(Badge Reader) = .464 - .892 (Suburban Area) + .186(Number of Employees)

Badge Reader: p-value = 3.323E-25, F=2976.9

Employees = 0, Suburban Area	Probability	Contribution
0	0.592	--
1	0.292	-0.21

Suburban Area = 0, Employees	Probability	Contribution	Relative Contribution
0	0.614	--	--
50	1.000	0.386	0.386
250	1.000	0.386	0.000
500	1.000	0.386	0.000

Summary

- This presentation summarizes one technique for extracting meaningful results from quality, but incomplete survey data
- It introduces the concept of Multiple Imputation and its use in addressing the problem of missing data
- It provides an overview of the statistical methods and analyses that can be used to generate traceable, statistically-sound results from surveys with missing data

- Thank you for attending! For more information, please contact
  - Deanna Ohwevwo (ohwevwo_deanna@bah.com)

Questions?