Missing Data Imputation

Dr. Ratna Babu Chinnam

Industrial & Systems Engineering

Wayne State University

Dealing with Missing Data

- Use what you know about
  - Why data are missing
  - Distribution of missing data
- Decide on best analysis strategy to yield least biased estimates

Missing Data Mechanisms (1)

- Preliminaries:
  - Y_{obs}: Non-missing or observed data
  - Y_{miss}: Missing or unobserved data
- Missing Completely at Random (MCAR) - SAFE
  - Probability of missingness is unrelated to either observed (Y_{obs}) or unobserved (Y_{miss}) data
- Missing at Random (MAR)
  - Probability of missingness may be related to observed data (Y_{obs}) but is unrelated to unobserved data (Y_{miss})
- Missing Not at Random (MNAR) - RISKY
  - Probability of missingness is related to (unknown) value of unobserved data (Y_{miss}), even after conditioning on observed data (Y_{obs})

Missing Data Mechanisms (2)

- Appropriateness of different missing data treatments depends on underlying missing data mechanism
- "Real" missing data can seldom be classified into just one of the three (MCAR, MAR, MNAR)
- Because we don't have access to missing data (Y_{miss}), we cannot empirically test whether or not data is MNAR
- If we know (or can convincingly argue) that the data is *not* MNAR, tests for whether data is MCAR is available

Missing Data Treatments

- Deletion methods
  - Delete observations with incomplete data
  - Advantage: Simplicity; Disadvantage: Loss of data
- Mean/Median/Mode substitution
  - Replace missing value with variable mean/median/mode
  - Disadvantage: Reduces variability, weakens estimates; Option: add some random error
- Nearest neighbor imputation
  - Replace missing value with statistic derived from nearest neighbor methods (Matlab: Knnimpute)
- Regression imputation
  - Regression predicted value imputation; Option: add some error
- Multiple imputation