Missing Data Analysis

May 2023

Click here to view the recording.

Outline

- Introduction
- What are missing data?
  - Sources of missingness
  - Mechanisms of missing data
- Methods to handle missing data
  - Single imputation, multiple imputation, inverse probability weighting, model based methods, and etc.
  - Relationship to types of missing data.
- General strategy to conduct data analysis with missing data.

Introduction

- Missing data happen frequently in research studies
- Ignoring missing data or using inappropriate methods to handle missing values can lead to misrepresentation of the target population, biased estimates and result in inaccurate conclusions.
- Discarding cases with missing values (complete case analysis) can lead to a loss of information, precision, and statistical power.
- Properly addressing missing data and documenting the chosen methods in the analysis can enhance the transparency and reproducibility of the research.

Sources of Missingness

- Data entry error
- A respondent chooses not to answer a question like `What is your salary?'
- Subjects missed visits, or lost to follow-up in a longitudinal study
- A new variable was created only partway through the data collection of a study.
- Measurements were outside of detectable range.

Types of Missingness

Missing data mechanism

- Missing Completely at Random (MCAR) - the probability of missingness is the same for all units. Like randomly poking holes in a data set.
- Missing at Random (MAR) - the probability of missingness in a variable depends only on available information (observed data).
- Missing Not at Random (MNAR) - the probability of missingness depends on information that has not been collected (unobserved data)

Types of Missingness

Type	Does missingness depend on observed data?	Does missingness depend on unobserved data?
MCAR	No	No
MAR	Yes	No
MNAR	Yes	Yes

Missing Completely at Random (MCAR)

- Examples:
  - A coin is flipped to determine whether an entry is removed.
  - Some pages are missing in the survey questionnaire
  - A random subsample of subjects were chosen to respond to questions.
  - A patient missed visits because of traffic jam
- Often missingness due to administrative reasons can be considered MCAR
- MCAR is the best case scenario, and the easiest to handle.
- Effect if you ignore: there is no effect on inferences (unbiased estimates), although may reduce statistical power

Missing at random (MAR)

- Examples:
  - men and women respond to the question "have you ever felt harassed at work?" at different rates.
  - Patients missed visits because of medication side effects (assuming medication side effects data are collected)
  - Patients lost to follow-up due to other medical conditions.
- Missing at Random can be handled properly.
- Effect if you ignore: inferences and predictions are biased.

Missing Not at Random (MNAR)

- Examples:
  - In a depression study, severe depressed subjects may be less likely to complete the questionnaire or answer specific questions about their depression
  - Patients drop out of a study because of good/bad outcomes, which we don't get to measure.
  - Patients missed medication may be less likely to respond to medication compliance questions.
- MNAR is difficult to handle properly
- if you ignore: inferences and predictions are biased.

Q: What is the missing mechanism of my data?

- Assess missing data mechanism: data inspection, summary statistics, domain knowledge.
- Compare summary statistics between cases with complete data and those with missing data.
  - Maybe to exclude the possibility of MCAR, but cannot confirm MCAR
  - Since we don't observe missing data, we can't confirm MAR vs MNAR.
- We generally think MCAR is unlikely in real life, other than from study design or administrative reasons (e.g., a random subsample is selected to respond questions).
- The true missing mechanism is probably mix of all 3.

Methods to handle missing data:

Complete case analysis

- analyze only the observations with complete information for all the variables of interest.
- Pros:
  1. Simple to do
  2. If MCAR, it provides unbiased estimates
- Cons:
  1. If MAR or MNAR, it can lead to biased estimates
  2. Inefficient use of available information => loss of statistical power

Single imputation: mean/median/mode imputation

- Replace missing values with the mean/median/mode of the observed values for the same variable.
- Pros:
  1. Simple to implement and understand.
- Cons:
  1. Doesn't preserve relationships between variables => generally leading to biased parameter estimates.
  2. Imputes the same value for all missing data => Underestimates the variance in the dataset
  3. Doesn't account for the uncertainty associated with imputing missing data => understates the variance of the parameter estimates

Single imputation: Last observation carried forward

- In longitudinal studies, replace missing values with the last observed values of the subject.
- Pros:
  1. Simple to implement and understand.
- Cons:
  1. Doesn't preserve relationships between variables => generally leading to biased parameter estimates.
  2. Assumes no change => Underestimates the variance in the datasets
  3. Fills in missing data with observed values, doesn't account for the uncertainty associated with imputing missing data => understate the variance of the parameter estimates

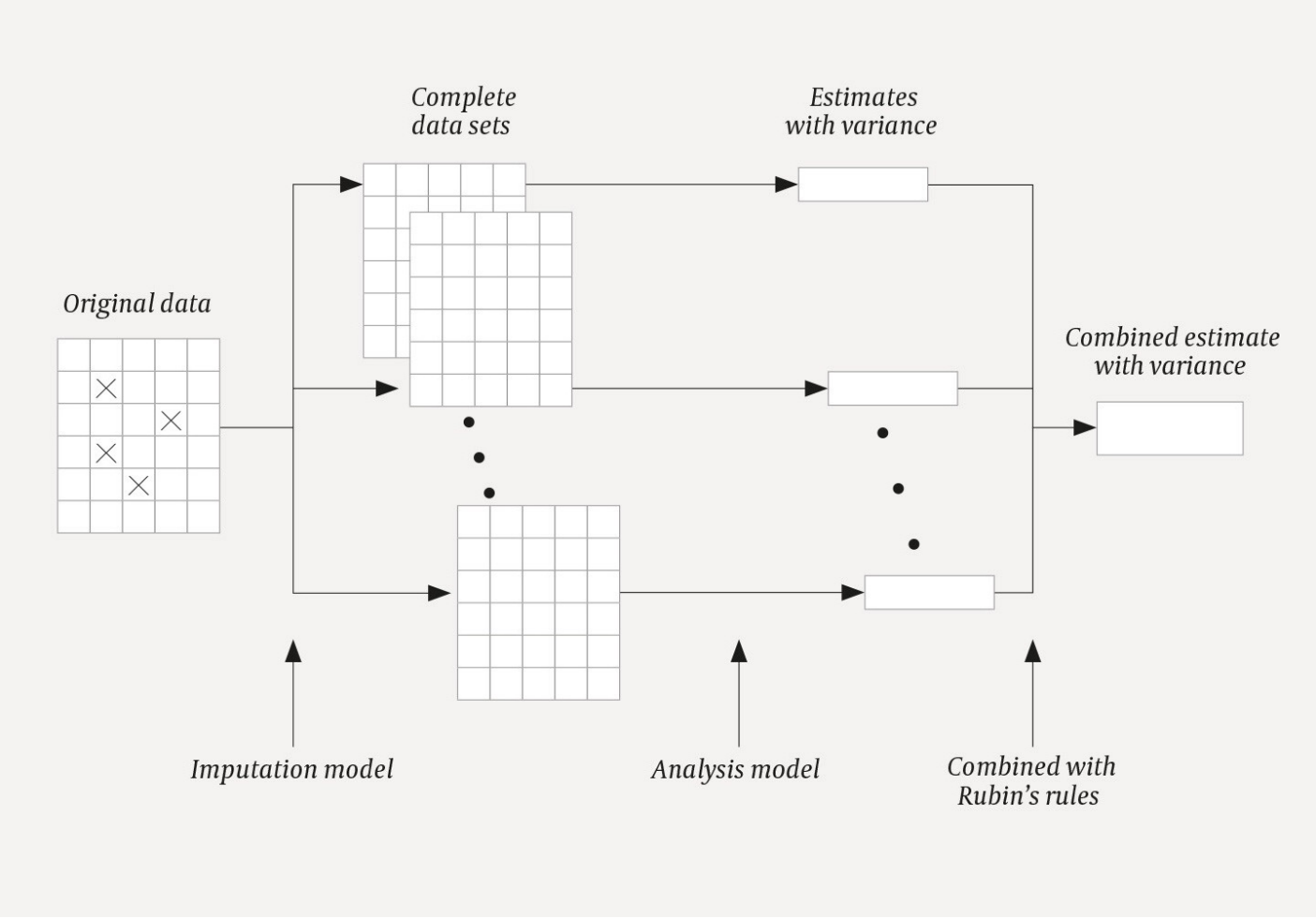
Single imputation: Regression imputation

- Estimates missing values using a regression model, where the variable with missing data is the dependent variable, and other variables in the dataset serve as independent variables.
- Pros:
  1. Makes use of the relationships between variables in the dataset. Can provide unbiased estimates if MAR.
  2. Provides more accurate imputations compared to mean/median/mode imputation or LOCF.
- Cons
  1. Assumes that the regression model is correctly specified
  2. Doesn't account for the uncertainty associated with imputing missing data => understates the variance of the parameter estimates

Single imputation: K-Nearest Neighbors

- Replaces missing values with the value(s) from the nearest neighbor(s) in the dataset, based on a distance metric calculated using other variables.
- Pros:
  1. Makes use of the relationships between variables in the dataset. Can provide unbiased estimates if MAR.
  2. Needs not assume parametric relationship between variables
- Cons
  1. Computationally more intensive
  2. Choice of distance metric and the number of neighbors (k) can impact the imputation results.
  3. Doesn't account for the uncertainty associated with imputing missing data => understates the variance of the parameter estimates

Multiple imputation



Multiple imputation:

1. Identify missing data:
2. Choose an imputation method (Imputation model) for example: Multiple Imputation by Chained Equations (MICE), Fully Conditional Specification (FCS), and Bayesian methods.
3. Create multiple imputed datasets: typically 5-10 or more datasets
4. Analyze each imputed dataset (Analysis model: linear regression, logistic regression, etc.)
5. Combine results: Use Rubin's rules
6. Report results

Multiple imputation:

- Pros:
  1. Can provide unbiased estimates if MAR.
  2. Improved efficiency: By using all available information in the data
  3. Flexibility: Multiple imputation can be applied to various types of missing data mechanisms (MCAR, MAR, MNAR) and is suitable for different types of variables (continuous, categorical, ordinal).
  4. Separation of imputation and analysis: One set of multiply imputed data can be used for different analyses.
  5. Accurate uncertainty estimation and confidence intervals.
  6. Available in popular statistical software packages.

Multiple imputation:

- Cons:
  1. Regular multiple imputation relies on certain assumptions, such as the missing data mechanism (e.g., MAR)
  2. Multiple imputation can't fully recover the information lost due to missing data, especially when a large proportion of data is missing or when the data are MNAR.
  3. Can be computationally intensive

Likelihood-based methods

- handle missing data by directly modeling the likelihood of the observed data, without requiring explicit imputation of missing values.
- E.g., Maximum Likelihood Estimation (MLE), Expectation-Maximization (EM) Algorithm, Bayesian methods, random effects models
- often rely on the assumption that the data are missing at random (MAR)

Inverse Probability Weighting (IPW)

Assign weights to the observed data based on the probability of being observed to adjust for the bias introduced by the missing values

1. Model the missing data mechanism, e.g., a logistic regression model with the outcome variable as an indicator of whether the data point is observed or missing
2. Calculate the inverse probability weights
3. Apply the weights in the analysis: perform a weighted analysis

Inverse Probability Weighting (IPW)

- Provide unbiased estimates under the MAR assumption
- Separation of outcome model and missing mechanism model. Allow incorporation of auxiliary variables that are related to the missingness but not included in the analysis model, which can help improve the MAR assumption's plausibility.
- Can be applied to a variety of statistical models and is relatively straightforward to implement in most statistical software packages.
- The method can be sensitive to extreme weights, which can result in high variability and inefficient estimates.

Methods for MNAR

- Missing Not at Random (MNAR)=> the missingness depends on the unobserved (missing) data itself.

1. Selection Models: jointly model the outcome variable and the missing data mechanism.
   - It has 2 parts: outcome model and missing mechanism model (which includes missing variable(s) as covariates.)
   - can provide unbiased estimates under the MNAR
   - require strong parametric assumptions and can be sensitive to model misspecification.

Methods for MNAR

2. Pattern Mixture Models (PMM):
   - separate the data into distinct patterns based on the missing data structure (e.g., lost to follow-up at 1 year, 2 years, 3 years.. etc)
   - model each pattern separately
   - Combine results to obtain an overall estimate
   - Choices of patterns can be arbitrary, and complex to implement.
   - Example: Impute missing data (using multiple imputation) in the intervention arm assuming they follow the same pattern as the observed data in the control arm.

Data	model
Treatment arm, no missing	With treatment effect
Treatment arm, with missing	No treatment effect
Control arm	No treatment effect

Methods for MNAR

3. Shared Parameter Models:
   - Useful in longitudinal or clustered data with missing values
   - These models jointly estimate the outcome model (e.g., linear mixed model) and a missing data model (e.g., logistic regression) that share latent random effects.
   - The shared random effects capture the correlation between the outcome and missingness => unbiased estimation under MNAR.
   - require strong parametric assumptions and can be computationally intensive

Analytic strategy

1. Explore the extent and patterns of missing data: Eg, summary tables

2. Assess missing data mechanism: data inspection, summary statistics, domain knowledge.

3. Choose an appropriate method: Most of the time, main analysis assumes MAR.
   - MCAR is unlikely; MAR or MNAR is more realistic.
   - Methods with MAR are more robust (multiple imputations, linear mixed effects model, IPW, etc)
   - Methods with MNAR rely heavily on the model assumptions of the relationship between outcome and unobserved.

4. Conduct sensitivity analysis, e.g., considering likely MNAR analyses to see if the conclusion still hold.

Thank you

Questions?