Missing Data and Imputation

Barnali Das

NAACCR Webinar
May 2016

Outline

- Basic concepts
- Missing data mechanisms
- Methods used to handle missing data

What are missing data?

- General term: data we intended to collect but did not.
- More precise definition: data which we intended to collect to answer a specific question are missing on some variables for some observations.
- Missing data is a common problems in large data sets.
  - Study subjects fail to report to a clinic for monthly evaluations, or respondents refuse to answer certain questions.
  - In clinical trails and observation studies, complete covariate data is often not available for every subject due to the loss of hospital records or the unavailability of covariate measurements.
  - Missing data inevitable in cancer registry data. To limit bias from missing data NAACCR require data meet strict criteria.

Issues with Missing Data

- Missing data reduces representativeness of the sample.
- Depending on how missing data occurred, introduces bias into statistical estimates and leads to inefficient data analysis.

Three Goals:

- Minimize bias
- Maximize use of available information
- Get good estimates of uncertainty

* Not a goal: imputed values “close” to a real values.

Approaches to Handle Missing Data

Ideally

- Identify:
  - Plausible reasons for the data being missing (missingness mechanisms)
  - The sensitivity of the conclusions to different missingness mechanisms
- Then
  - Perform valid analysis under different plausible mechanisms, draw conclusions.
  - Discuss the implications and come to a valid interpretation of the study.

Practically

- Postulate a missingness mechanism and identify its class
- Perform a valid analysis for that class of missingness mechanism

Classifications of Missing Data

- Rubin(1976) defined three broad classes of mechanism, each with distinct implications for the analysis.
  1. MCAR- the missingness is independent of both the missing response and the observed response.
  2. MAR- the missingness is independent of the missing response given the observed values. The probability that Y is missing does not dependent on the value of Y, after controlling for observed variables.
  3. MNAR (Non-ignorable) - the missingness depends on both observed and missing responses. The MAR assumption is violated.
- Going beyond complete case analysis, we have to consider the missingness mechanism.
- These are the assumptions used in statistical methods for datasets with missing values, particular multiple imputation.

Missing Completely at Random (MCAR)

- MCAR- the missingness is independent of both the missing response and the observed response.
- MCAR is ideal situation.
- If data are MCAR, complete data subsample is a random sample from original target sample. Analysis is unbiased but less precise.
- It is not likely to be true in practice.
- MCAR assumptions can be examined but can not be fully confirmed

Missing at Random (MAR)

- MAR- the missingness is independent of the missing response given the observed values. The probability that Y is missing does not dependent on the value of Y, after controlling for observed variables.
  - E.g., the probability of missing income depends on marital status, but within each status, the probability of missing income doesn't depend on income.
- Considerably weaker assumption than MCAR.
- Impossible to test whether the MAR condition is satisfied.
- MAR is assumption we make for analysis, not a characteristic of the dataset.
- The reason for missingness may depend on the unobserved values, but conditional on data we observe they are independent.
- It makes the analysis much simpler!

Missing Not At Random (MNAR)

- MNAR - the missingness depends on both observed and missing responses.
  - E.g., People who weigh more may be less likely to report their weight on a questionnaire
- Under MNAR, both response of interest and the missingness mechanism need to be modeled.
- Effective estimation for MNAR missing data requires very good prior knowledge about missing data mechanism (pattern mixture model).
  - Data contain no information about what models would be appropriate
  - No way to test goodness of fit of missing data model
  - Results often very sensitive to choice model.

Examples: MCAR, MAR, MNAR

Survey 200 employees, 100 each of job type A and B. Income of type A average $60,000, Income of type B average $30,000. True average income $45,000.

If 50 employees refused to report their income,

1. Scenario under MCAR: average of 25 missing observations from each type and average income around $45,000 with slightly larger standard errors because of less observations.
2. Scenario under MAR : A has higher probability of missing income than B (i.e. 30 missing from A and 20 from B). However, within A and within B, probability of observing income doesn't depend on income. Average of observed income lower than true average ($45,000). i.e. (70*$60,000 + 80*$30,000)/150 = $44,000
3. Scenario under MNAR : probability of missing income associated with income – higher income less likely to be observed e.g., 40 missing A had above $80,000 salary and 10 missing B had above $40,000 salary). Average of observed income will be much lower than the true average

Imputation Goals

- Carefully complete the dataset
- Maintain the true underlying distribution of the data
- Maintain multivariate associations
  - Allow for low observed combinations of values to occur in the imputations - e.g., males with breast cancer
- Maintain point estimates
  - Reduce bias due to item missingness
  - Account for missing data mechanism

11

Imputation Goals (Continued)

- Maintain variances
  - Account for extra uncertainty due to imputation
- Preserve the shape of the data
  - Ranges
  - Spikes or rounded values
    - Income reported in multiples of $5,000 for most cases
    - Minutes traveled reported in multiples of 5 for most cases
- Preserve structured missingness
  - E.g., Questionnaires ~ 'skip patterns'
  - E.g., Treatment for females

12

Imputation Approaches (Continued)

Pre-imputation Steps

- Identify the variables to impute (target variables)
- Identify the missing data patterns
  - Structured (monotone)
    - What are the trigger items (the variables that the target variables depend on)?
    - E.g., Target variable = income; Trigger item = employment status
  - Non-structured (non-monotone)
    - Swiss cheese missing data

13

Imputation Approaches (Continued)

Pre-imputation Steps (continued)

- Compute item missing rates
  - More focus on items with higher missing rates
    - Model building
    - Diagnostics
- Identify item types
  - Continuous
  - Categorical (ordered) (e.g., health status: 1 = poor to 5= excellent)
  - Categorical (unordered) (e.g., race, marital status)
  - Cyclical (e.g., time of day)
- Review distributions of variables

14

Imputation Approaches (Continued)

Pre-imputation Steps (continued)

- Create a pool of predictors
  - From within the dataset
  - From external data
    - Tract-level percentages from the five-year American Community Survey data tables
      - Percent who rent
      - Percent who do not speak English well
      - Percent with less than high school education
    - Census data
    - Small area estimates for counties

15

Imputation Approaches (Continued)

Pre-imputation Steps (continued)

- Compute pairwise correlations
  - To understand the relationships between items, their predictability, expectations of the resulting imputations
- Identify key predictors
  - Use correlations
  - Use stepwise regression
  - Use research

16

Methods to Treat Missing Data

Available Case Methods

- Complete case method (listwise deletion).
  - Analyze only those subjects who are completely observed.
  - Easy to implement - works for any kind of statistical analysis
  - If data are MCAR, does not introduce any bias in parameter estimate and standard error estimates are appropriate.
  - May delete a large proportion of cases, resulting in loss of statistical power.
  - May introduce bias if MAR
- Pairwise deletion
  - Delete only the cases with complete responses for each calculation.
  - Different calculations in an analysis may be based on different sample sizes.
  - Approximately unbiased if MCAR
  - Biased estimates if MAR
  - Incorrect standard errors (no appropriate sample size)

Methods to Treat Missing Data

Single Imputation methods

- Mean substitution
  - Replace missing values with means
  - Causes bias in variance estimates
- Regression Imputation
  - Replace missing values with conditional means
- Last Observation Carried Forward
  - Replace missing values last observed value
- Hot Deck
  - Divide sample into homogeneous strata on observed variables. Within each stratum pick “donor” units with observed values to fill in missing values for other units.
- Often leads to biased parameter estimates (e.g. small variances)
- Leads to standard errors estimates that are biased downward
  - Treats imputed data as real data, ignore inherent uncertainty in imputed values.

Methods to Treat Missing Data

Modern Approaches

- Maximum Likelihood (ML) method
  - Choose as parameter estimates those values would maximize the probability of observing what has, in fact, been observed.
  - Consistent (approximately unbiased in large samples)
  - Asymptotically efficient and normal
- Bayesian method
  - Specifying prior and distribution for the missing covariates
  - Missing values are sampled from fully conditional distribution via Gibbs sampler.
- Multiple Imputation (MI)
  - Utilized both ML and Bayesian approach
  - Impute missing values with several plausible values
  - Estimates are usually consistent, asymptotically efficient and normal.
  - Can be used in any kind of data and model
  - May get a different result every time you run it.

Some Slippery Slopes

What if you -

- Impute without best predictors available
  - Causes bias in point estimates
  - If MCAR, you don't need predictors
- Impute each item independently
  - Causes bias in correlations
- Impute without attention to missing value codes
  - Causes bias in everything for ordinal variables

<u>Should not treat imputed values as if they were observed</u>

20

References I : Basic Texts

Some good texts:

Little, RJ & Rubin, DB. “Statistical Analysis with Missing data”

Allison, PD. “Missing data”

Carpenter JR & Kenward MG. “Multiple Imputation and its Application”

Enders, C. “Applied Missing Data Analysis”

van Buuren, S. “Flexible Imputation of Missing Data”

References II

- Allison, P. (2012). Modern Methods for Missing Data. Webinar conducted for the American Statistical Association. May 2012
- Ault, K. (2012). Multiple imputation for ordinal variables: A comparison of SUDAAN Proc Impute and SAS Proc MI. Paper SD-12. SESUG 2012
- Creel, D. (2011). A comparison of the approximate bayesian bootstrap and the weighted sequential hot deck for multiple imputation. Proceedings of the Section on Survey Research Methods of the American Statistical Association, 4494-4500
- Judkins, D., Krenzke, T., Piesse, A., Fan, Z., and Haung, W.C. (2007). Preservation of skip patterns and covariate structure through semi-parametric whole questionnaire imputation. Proceedings of the Section on Survey Research Methods of the American Statistical Association, 3211-3218
- Kalton, G, and Kish, L. (1984). Some efficient random imputation methods. Comm. Statist. Theory Methods, A 13, 1919–1939

References III

- Kim, J., Brick, J.M., Fuller, W., and Kalton, G. (2006). On the bias of the multiple-imputation variance estimator in survey sampling. *Journal of the Royal Statistical Society*
- Kim, J.K. and Yu, C.Y. (2011). A semi-parametric estimation of mean functionals with non-ignorable missing data," *Journal of the American Statistical Association*, 106, 157-165
- Li, L., Lee, H., Lo, A., and Norman, G. (2008). Imputation of missing data for the Pre-Elementary Longitudinal Study. Proceedings of the Section on Survey Research Methods of the American Statistical Association
- Little, R.J. and Rubin, D.B. (2002) *Statistical Analysis with Missing Data*, 2nd Edition. J. Wiley & Sons, New York
- Little, R.J., Yosef, M., Cain, K.C., Nan, B., and Harlow, S.D. (2008). A hot-deck imputation procedure for gaps in longitudinal data on recurrent events. *Statistics in Medicine*, 27(1), 103-120

References IV

- Rao, J.N.K. and Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 811-822
- Rubin, D.B. (1996). Multiple imputation after 19+ years. *Journal of the American Statistical Association*, 91, 473-489
- Rubin, D.B. and Schenker, N. (1986). "Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse." *Journal of the American Statistical Association*. Vol. 81, 366-374.
- Schenker, N., Raghunathan, T., Chiu, P-L, Makuc, D, Zhang, G., and Cohen, A. (2008). Multiple imputation of family income and personal earnings in the National Health Interview Survey: methods and examples. Technical document written for the Centers for Disease Control, National Center for Health Statistics http://www.cdc.gov/nchs/data/nhis/tecdoc.pdf (accessed May 22, 2013)

References V

- Siddique, J. and Belin, T.R. (2008). Using an approximate Bayesian bootstrap to multiply impute nonignorable missing data. *Computational Statistics and Data Analysis* 53: 405-415
- Van Buuren, S. and Groothuis-Oudshoorn, K. (2011). MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, December 2011, Volume 45, Issue 3