Handling Missing Data: An Introduction

STA 210: Regression Analysis

Olanrewaju Michael Akande

Department of Statistical Science, Duke University

October 18, 2018

Outline

Motivation

- Most real world datasets often contain missing values.
- Ideally, analysts should first decide on how to deal with missing data before moving on to analysis.
- One needs to make assumptions and ask tons of questions, for example,
    - why are the values missing?
    - what is the pattern of missingness?
    - what is the proportion of missing values in the data?
- As a Bayesian, one could treat the missing values as parameters and estimate them simultaneously with the analysis, but even in that case, he/she must still first answer the same questions.

Motivation

- Simplest approach: complete/available case analyses – delete cases with missing data. Often problematic because:
  - it is sometimes infeasible (small n large p problem) – when we have a small number of observations but a large number of variables, we simply can not afford to throw away data, even when the proportion of missing data is small.
  - information loss – even when we do not have the small n, large p problem, we still lose information when we delete cases.
  - biased results – because the missing data mechanism is rarely random, features of the observed data can be completely different from the missing data.
- More principled approach: impute the missing data (in a statistically proper fashion) and analyze the imputed data.

Why should we care?

- Loss of power
  - can't regain lost power
- Any analysis must make an untestable assumption about the missing data
  - wrong assumption $\Rightarrow$ biased estimates
- Some popular analyses with missing data get biased standard errors
  - resulting in wrong p-values and confidence intervals
- Some popular analyses with missing data are inefficient
  - confidence intervals wider than they need be

What to do: loss of power

Approach by design:

- minimise amount of missing data
  - good communications with participants, for example, patients in clinical trial, participants in surveys and censuses, etc
  - follow up as much as possible; make repeated attempts using different methods
- Reduce the impact of missing data
  - collect reasons for missing data
  - collect information predictive of missing values

What to do: analysis

A suitable method of analysis would:

- Make the correct assumption about the missing data
- Give an unbiased estimate (under that assumption)
- Give an unbiased standard error (so that P-values and confidence intervals are correct)
- Be efficient (make best use of the available data)

BUT we can never be sure about what the correct assumption is
$\Rightarrow$ sensitivity analyses are essential!

How to approach the analysis

- Start by knowing:
  - extent of missing data
  - pattern of missing data (e.g. is X_1 always missing whenever X_2 is also missing?)
  - predictors of missing data and of outcome
- Principled approach to missing data:
  - identify a plausible assumption (needs discussion between statisticians and clinicians)
  - choose an analysis method that's valid under that assumption
- Some analysis methods are simple to describe but have complex and/or implausible assumptions

Types of missing data

- Unit nonresponse: the individual has no values recorded for any of the variables (we will not focus on this today!).

- Item nonresponse: the individual has values recorded for at least one variable, but not all variables

Table 1: Unit nonresponse vs item nonresponse

	X_1	X_2	Y
Complete cases →	✓	✓	✓
		✓	?
Item nonresponse	✓	?	?
		?	✓
Unit nonresponse →	?	?	?

Types of missing data mechanisms

- Missing completely at random (MCAR):
  - the reason for missingness does not depend on the values of the observed data or missing data
  - rarely plausible in practice
- Missing at random (MAR):
  - the reason for missingness may depend on the values of the observed data but not the missing data (conditional on the values of the observed data)
  - most commonly assumed in analysis.
- Missing not at random (MNAR or NMAR):
  - the reason for missingness depends on the actual values of the missing (unobserved) data
  - usually the case in real analysis, but analysis can be complex!

Mathematical formulation

- Consider the classical multiple regression setting with

$$Y_i = \boldsymbol{\beta}\mathbf{X}_i + \epsilon_i; \quad \epsilon_i \sim N(0, \sigma_\epsilon^2); \quad i = 1, \ldots, n$$

  where $\mathbf{X}_i = (1, X_{i1}, \ldots, X_{ip})$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$.

- Suppose for now, that $\mathbf{Y} = (Y_1, \ldots, Y_n)$ contains missing values but $\mathbf{X} = (\mathbf{X}_i, \ldots, \mathbf{X}_n)$ is fully observed.

- We can separate $\mathbf{Y}$ into the observed and missing parts, that is, $\mathbf{Y} = (\mathbf{Y}_{obs}, \mathbf{Y}_{mis})$. We can also do the same for $\mathbf{X}$ if it contains missing values.

Mathematical formulation

- Let $r_i = 1$ when Y_i is missing and $r_i = 0$ otherwise.
- Let $\mathbf{R} = (r_1, \ldots, r_n)$, and θ, the parameters associated with $\mathbf{R}$. This is the vector of missing indicators for $\mathbf{Y}$.
- When $\mathbf{X}$ contains missing values, we can also create a vector of missing indicators for each variable in $\mathbf{X}$ with missing entries.
- Assume θ and $(\boldsymbol{\beta}, \sigma_\epsilon^2)$ are distinct.

Mathematical formulation

- MCAR:

$$f(\mathbf{R}|\mathbf{Y}, \mathbf{X}, \theta, \boldsymbol{\beta}, \sigma_\epsilon^2) = f(\mathbf{R}|\theta)$$

- MAR:

$$f(\mathbf{R}|\mathbf{Y}, \mathbf{X}, \theta, \boldsymbol{\beta}, \sigma_\epsilon^2) = f(\mathbf{R}|\mathbf{Y}_{obs}, \mathbf{X}, \theta)$$

- MNAR:

$$f(\mathbf{R}|\mathbf{Y}, \mathbf{X}, \theta, \boldsymbol{\beta}, \sigma_\epsilon^2) = f(\mathbf{R}|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \mathbf{X}, \theta)$$

Each type of mechanism has a different implication on the likelihood of the observed data $\mathbf{Y}_{obs}$, and the missing data indicator $\mathbf{R}$. However, we will not go into that much detail today.

Types of missing data mechanisms: how to tell?

So how can we tell the type of mechanism we are dealing with?

In general, we don't know!!!

- Rare that data are MCAR (unless planned beforehand)
- Possible that data are MNAR
- Compromise: assume data are MAR if we include enough variables in model for the missing data indicator **R**.

Outline

Strategies for handling missing data

- Item nonresponse:
  - use complete/available cases analyses
  - single imputation methods
  - multiple imputation
  - model-based methods
- Unit nonresponse:
  - weighting adjustments
  - model-based methods (identifiability issues!).

Complete/available cases analyses

What can happen when using available case analyses with different types of missing data?

- MCAR: unbiased when disregarding missing data; variance increase (losing partially complete data)
- MAR: bias when missing data mechanism not modeled; variance increase (losing partially complete data)
- NMAR: generally biased!

Single imputation methods

- Marginal/conditional mean imputation
- Nearest neighbor imputation:
  - hot deck imputation
  - cold deck imputation
- Use observation from one of the previous time periods (for panel data)
  - LOCF – last observation carried forward
  - BOCF – baseline observation carried forward

Mean imputation

Plug in the variable mean for missing values.

- Point estimates of means OK under MCAR
- Variances and covariances underestimated.
- Distributional characteristics altered.
- Regression coefficients inaccurate.

Similar problems for plug-in conditional means.

Nearest neighbor imputation

Plug in donors' observed values.

- Hot deck: for each nonrespondent, find a respondent who "looks like" the nonrespondent in the same dataset
- Cold deck: find potential donors in an external but similar dataset. For example, respondents from a 2016 election poll survey might serve as potential donors for nonrespondents in the 2018 version of the same survey.
- Common metrics: Statistical distance, adjustment cells, propensity scores.

Nearest neighbor imputation

- Point estimates of means OK under MAR.
- Variances and covariances underestimated.
- Distributional characteristics OK.
- Regression coefficients OK under MAR.

Multiple imputation

- Fill in data sets m times with imputations.
- Analyze repeated data sets separately, then combine the estimates from each one.
- Imputations drawn from probability models for missing data.

Imputed Datasets

Table 2: Imputed datasets: missing values are replaced with plausible values.

y	x
✓	?
✓	✓
?	✓
✓	✓
?	✓
✓	?

(a) Observed Data

y	x	y	x	y	x
✓	✓	✓	✓	✓	✓
✓	✓	✓	✓	✓	✓
✓	✓	✓	✓	✓	✓
✓	✓	✓	✓	✓	✓
✓	✓	✓	✓	✓	✓
✓	✓	✓	✓	✓	✓

(b) MI Datasets $(1, \ldots, m)$

MI: inferences from multiply-imputed datasets

Rubin (1987)

- Estimand: $Q = Q(X, Y)$
- Q can be estimated using q, with variance u.
- In a regression setting, suppose $Q = \beta_0$, then $q = \hat{\beta}_0$ and $u = Var(\hat{\beta}_0)$.
- In each imputed dataset d_i, where $i = 1, \ldots, m$

$$q_i = Q(d_i)$$

$$u_i = U(d_i)$$

MI: quantities needed for inference

- $\bar{q}_m = \sum_{i=1}^{m} \frac{q_i}{m}$
- $b_m = \sum_{i=1}^{m} \frac{(q_i - \bar{q}_m)^2}{m-1}$
- $\bar{u}_m = \sum_{i=1}^{m} \frac{u_i}{m}$

MI: inferences from multiply-imputed data

- MI estimate of Q: $\bar{q}_m$
- MI estimate of variance is:

$$T_m = (1 + 1/m)b_m + \bar{u}_m$$

- Use t-distribution inference for Q

$$\bar{q}_m \pm t_{1-\alpha/2}\sqrt{T_m}$$

Notice that the variance incorporates uncertainty both from within and between the m datasets.

MI: where should the imputations come from

So where should we get reasonable replacements for the missing values from? There are two general approaches:

- Sequential modeling
  - Estimate a sequence of conditional models (think separate regressions for each variable!)
  - Impute from each model
- Joint modeling
  - Choose a multivariate model for all the data
  - Estimate the model
  - Impute from the joint model

MI: sequential regression models

Suppose the data include Y_1, Y_2, Y_3

- Step 1: fill in missing values by simulating values from regressions based on complete cases
- Step 2: regress $Y_1|Y_2, Y_3$ using completed data
- Step 3: impute new values of Y1 from this model
- Step 4: repeat for $Y_2|Y_1, Y_3$ and $Y_3|Y_1, Y_2$
- Step 5: cycle through steps 2-4 times

MI: existing software for sequential regression approach

Free software packages

- MICE for R and Stata
- MI for R
- IVEWARE for SAS

Can specify many types of conditional models and include constraints on values.

MI: existing software for joint regression approach

A few examples of free software packages

- Multivariate normal data
  - R: NORM, Amelia II
  - SAS: proc MI
  - Stata: MI command
- Mixtures of multivariate normal distributions:
  - R: EditImpCont
- Multinomial data:
  - R: CAT – log-linear model

MI: pros and cons

- Advantages
  - Straightforward estimation of uncertainty
  - Flexible modeling of missing data
- Disadvantages (??)
  - Extra data sets to manage
  - Explicitly model-based

MI: a simple illustration

We will use a simple example created by Dr. Jerry Reiter

`https://www2.stat.duke.edu/~jerry/missingdata.txt`

Outline

Concluding remarks

- Ignoring missing data is risky.
- Single imputation procedures at best underestimate uncertainty and at worst fail to capture multivariate relationships.
- Multiple imputation recommended (or other model-based methods).
- We discussed MI for MAR data. When data are NMAR life much harder – get experts in missing data on your team.

Concluding remarks

- Incorporate all sources of uncertainty in imputations, including uncertainty in parameter estimates.
- Want models that accurately describe the distribution of missing values.
- Important to keep in mind that imputation model used only for cases with missing data.
  - Suppose you have 30% missing values
  - Also, suppose your model is "80% good" ("20% bad")
  - Then, completed data are only "6% bad"

Resources for learning more

- Little and Rubin (2002), *Statistical Analysis with Missing Data*, Wiley
- Schafer (1997), *Analysis of Incomplete Multivariate Data*, CRC Press
- Reiter and Raghunathan (2007), "The multiple adaptations of multiple imputation," *JASA*.

Acknowledgments

These slides contain materials adapted from courses taught by Dr. Jerry Reiter and Dr. Fan Li.

Questions?