IDS 702: Module 5.1

Introduction to missing data

Dr. Olanrewaju Michael Akande

IDS 702

Motivation

- Most real world datasets often suffer from nonresponse, that is, they 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, one must still ask the same questions.
- Ask as many questions as possible to help you figure out the most plausible assumptions!

Motivation

- Simplest approach: complete/available case analyses -- delete cases with missing data. Often problematic because:
  - it is just not feasible sometimes (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 due to the the smaller sample size
  - 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
  - so that confidence intervals are wider than they need be

What to do: loss of power

Approach by design:

- minimize 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 (or plausible) 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)
- However, 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 (through discussions between you as a data scientist and your clients)
  - choose an analysis method that's valid under that assumption
- Just because a method is simple to use does not make it plausible; some analysis methods are simple to describe but have complex and/or implausible assumptions.

Types of nonresponse (missing data)

- Unit nonresponse: the individual has no values recorded for any of the variables. For example, when participants do not complete a survey questionnaire at all.
- Item nonresponse: the individual has values recorded for at least one variable, but not all variables.

Unit nonresponse vs item nonresponse

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

Types of Missing Data Mechanism

- Data are said to be missing completely at random (MCAR) if the reason for missingness does not depend on the values of the observed data or missing data.
- For example, suppose
  - you handed out a double-sided survey questionnaire of 20 questions to a sample of participants;
  - questions 1-15 were on the first page but questions 16-20 were at the back; and
  - some of the participants did not respond to questions 16-20.
- Then, the values for questions 16-20 for those people who did not respond would be missing completely at random if they simply did not realize the pages were double-sided; they had no reason to ignore those questions.
- This is rarely plausible in practice!

IDS 702

Types of Missing Data Mechanism

- Data are said to be missing at random (MAR) if 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).
- Using our previous example, suppose
  - questions 1-15 include demographic information such as age and education;
  - questions 16-20 include income related questions; and
  - once again, some of the participants did not respond to questions 16-20.
- Then, the values for questions 16-20 for those people who did not respond would be missing at random if younger people are more likely not to respond to those income related questions than old people, where age is observed for all participants.
- This is the most commonly assumed mechanism in practice!

IDS 702

Types of Missing Data Mechanism

- Data are said to be missing not at random (MNAR or NMAR) if the reason for missingness depends on the actual values of the missing (unobserved) data.
- Continuing with our previous example, suppose again that
  - questions 1-15 include demographic information such as age and education;
  - questions 16-20 include income related questions; and
  - once again, some of the participants did not respond to questions 16-20.
- Then, the values for questions 16-20 for those people who did not respond would be missing not at random if people who earn more money are less likely to respond to those income related questions than old people.
- This is usually the case in real analysis, but analysis can be complex!

IDS 702

Types of missing data mechanisms: how to tell in practice?

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 $\boldsymbol{R}$.

IDS 702

WHY SHOULD WE CARE?

- Why should we care in practice? What does bias really mean here? How exactly does using only the complete cases affect our results for the three mechanisms?

- Let's attempt to answer these questions via simulations.

- Set $n = 10,000$. For $i = 1, \ldots, n$, generate
  - $x_i \overset{iid}{\sim} N(2, 1); \quad y_i | x_i \overset{iid}{\sim} N(-1 + 2x_i, \sigma^2 = 5^2)$
  - $r_i | y_i, x_i \sim \text{Bernoulli}(\pi_i); \quad \log\left(\dfrac{\pi_i}{1 - \pi_i}\right) = \theta_0 + \theta_1 y_i + \theta_2 x_i$

- Next, set y_i missing whenever $r_i = 1$.

- Set different values for $\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2)$ to reflect MCAR, MAR and MNAR.

- Let's use the R script here.

IDS 702

What's next?

Move on to the readings for the next module!

IDS 702