Missing data

Wenda Zhou

June 20, 2017

Missing data

- Common occurence in many datasets, especially in surveys.
- Numerous possible causes: non-response bias, difficulty in collecting the data, etc.
- Important to understand the missingness as it can mislead our estimates.

Types of missingness

It is useful to distinguish three types of statistical missingness.

Missing completely at random (MCAR)

The missingness does not depend on any thing.

Types of missingness

It is useful to distinguish three types of statistical missingness.

Missing completely at random (MCAR)

The missingness does not depend on any thing.

Missing at random (MAR)

The missingness only depends on the value of what is observed (but not directly on the value of what is missing).
Example: men may be less likely to answer a question on depression, and also more likely to have depression. But men who have depression are just as likely to answer the question as those who don't.

Types of missingness

It is useful to distinguish three types of statistical missingness.

Missing completely at random (MCAR)

The missingness does not depend on any thing.

Missing at random (MAR)

The missingness only depends on the value of what is observed (but not directly on the value of what is missing).
Example: men may be less likely to answer a question on depression, and also more likely to have depression. But men who have depression are just as likely to answer the question as those who don't.

Missing not at random (MNAR)

The missingness may depend on the value of the missing data.

Survey biases

Non-response bias

Answers collected differs from potential answers that were not collected.

Survey biases

Non-response bias

Answers collected differs from potential answers that were not collected.

Response bias

Giving inaccurate answers to a question. Numerous possible causes:

- Phrasing of the question might push towards some answer
- Social desirability
- Demand characteristics: behaviour changes simply by being part of an experiment

Dealing with missing data

Complete case analysis

Ignore the observations where the data is missing, and only analyse complete cases.

- Simple strategy, good if very little missing data
- May discard too much data
- May cause bias if missing units systematically different

Dealing with missing data

Complete case analysis

Ignore the observations where the data is missing, and only analyse complete cases.

- Simple strategy, good if very little missing data
- May discard too much data
- May cause bias if missing units systematically different

Complete case analysis is correct when the data is missing completely at random.

MAR data

Data is rarely MCAR. However, MAR can be a reasonable assumption.

Complete case analysis

For MAR data, complete case analysis is correct as long as we control for every variable that may cause missingness.

MAR data

Data is rarely MCAR. However, MAR can be a reasonable assumption.

Complete case analysis

For MAR data, complete case analysis is correct as long as we control for every variable that may cause missingness.

Can still face problems with deleting too much data.

Imputation

Idea to handle missing data: guess the values that are missing!

Imputation

Guess the missing data from other observations.
For example: income not reported for a white college-educated man: can predict based on income of other white college-educated men.

Imputation

Idea to handle missing data: guess the values that are missing!

Imputation

Guess the missing data from other observations.
For example: income not reported for a white college-educated man: can predict based on income of other white college-educated men.

Can then "pretend" we have no missingness.

Imputation

- Often reduces bias in estimation
- Can handle cases with large amounts of missingness
- Can apply any method we desire