Missing Data

COMP3008

Lauren Ansell

Introduction

Today's topics:

- What do we mean by missing data?
- Why do we get missing data?
- Methods to deal with missingness.

Session learning outcomes - by the end of today's lecture you will be able to:

- know what the different types of missing data are.
- select an appropriate method for dealing with missingness.

Missing Data...AAARGH!

Missing data is where a value has not been recorded.

The majority of datasets will contain missing values to some degree.

Concluding why or how that data came to be missing will help inform how you deal with the missingness.

Missingness can affect the conclusions drawn from the data.

Why Do We Get Missing Data?

Missingness can arise from a number of reasons:

- equipment failure
- participant does not respond
- error during data entry
- data collection completed incorrectly

Types of Missing Data

Missing Completely At Random - MCAR

The missing information is independent of both observable variables and unobservable parameters of interest and entirely at random.

When data are MCAR, the analysis performed on the data is unbiased; however, data are rarely MCAR.

The missingness of data is unrelated to any study variable.

Missing At Random - MAR

Missing at random (MAR) occurs when the missingness is not random, but where missingness can be fully accounted for by variables where there is complete information.

Depending on the analysis method, these data can still induce parameter bias in analyses due to the contingent emptiness of cells.

However, if the parameter is estimated with Full Information Maximum Likelihood, MAR will provide asymptotically unbiased estimates.

Missing not at random

Missing not at random (MNAR) is data that is neither MAR nor MCAR.

Samuelson and Spirer[1] discussed how missing and/or distorted data about demographics, law enforcement, and health could be indicators of patterns of human rights violations.

[1] **jabine1992human**.

Structured missingness

An increasingly encountered problem arises in which data may not be MAR but missing values exhibit an association or structure, either explicitly or implicitly.

Structured missingness commonly arises when combining information from multiple studies, each of which may vary in its design and measurement set and therefore only contains a subset of variables from the union of measurement modalities.

The presence of structured missingness may be a hindrance to make effective use of data at scale, including through both classical statistical and current machine learning methods.

Planned missingness

Missing data can also be a deliberate part of study design.

Specifically, planned missingness is a research design strategy, employed in survey research, in which data are intentionally left uncollected from individual respondents.

The aim is to reduce burden while preserving the ability to estimate parameters for the full item set across the sample.

Solutions

There are three main methods for dealing with missing data:

- omission
- imputation
- analysis

The method we adopt will depend on the type and amount of missingness in the data.

Solutions - Omission

This is the simplest method of dealing with missingness.

All cases which contain missing data are removed from the dataset.

Solutions - Imputation

Some data analysis techniques are not robust to missingness, and require to "fill in", or impute the missing data.

Any multiply-imputed data analysis must be repeated for each of the imputed data sets and, in some cases, the relevant statistics must be combined in a relatively complicated way.

The expectation-maximization algorithm is an approach in which values of the statistics which would be computed if a complete dataset were available are estimated (imputed), taking into account the pattern of missing data.

Solutions - Analysis

Methods which take full account of all information available, without the distortion resulting from using imputed values as if they were actually observed.

Generative approaches:

- the expectation-maximisation algorithm
- full information maximum likelihood estimation

Discriminative approaches:

- Max-margin classification of data with absent features
- Partial identification methods may also be used.

Missing Data and Recommender Systems

We have seen that recommender systems require large datasets to create meaningful results.

However, not all customers will have rated all of the items.

We need a method to impute the missing values - matrix completion.

Principal components can be used to impute the missing values, through a process known as matrix completion.

We have a modified optimisation problem:

$$\underset{A\in\mathbb{R}^{n\times M},B\in\mathbb{R}^{p\times M}}{\textit{minimise}}\left\{\sum_{(i,j)\in\mathcal{O}}\left(x_{ij}-\sum_{m=1}^{M}a_{im}b_{jm}\right)^2\right\}$$

Algorithm

Step 1:

Create a complete data matrix $\tilde{\mathbf{X}}$ of dimension $n \times p$ of which the (i, j) element equals

$$\tilde{x}_{ij} = \begin{cases} x_{ij} & \text{if } (i,j) \in \mathcal{O} \\ \bar{x}_j & \text{if } (i,j) \notin \mathcal{O} \end{cases}$$

$\mathcal{O}$ denotes the observations that are observed in $\mathbf{X}$.

Step 2:

Solve:

$$\underset{A \in \mathbb{R}^{n \times M}, B \in \mathbb{R}^{p \times M}}{\textit{minimise}} \left\{ \sum_{j=1}^{p} \sum_{i=1}^{n} \left(\tilde{x}_{ij} - \sum_{m=1}^{M} a_{im} b_{jm} \right)^2 \right\}$$

by computing the principal components of $\tilde{\mathbf{X}}$.

For each element $(i, j) \notin \mathcal{O}$, set $\tilde{x}_{ij} \leftarrow \sum_{m=1}^{M} \hat{a}_{im} \hat{b}_{jm}$.

Compute the objective

$$\sum_{(i,j) \in \mathcal{O}} \left(x_{ij} - \sum_{m=1}^{M} \hat{a}_{im} \hat{b}_{jm} \right)^2$$

Step 3:

Return the estimated missing entries $\tilde{x}_{ij}$, $(i,j) \notin \mathcal{O}$.

Netflix

Netflix had its customers rate the content they watched on a scale of 1-5.

On average, each customer had only seen around 200 movies/shows - 99% of missing elements.

The key idea was that there will be an overlap in the content that users have watched and some of those will have similar content preferences.

Netflix can use similar customer ratings of content the next customer has not seen to predict what that customer would like.

Summary

Missing Data

- Missingness will be present in the majority of datasets you will deal with.
- Missing data can be an issue and we need to decide carefully how we deal with it.
- There are multiple methods to deal with missing data - the issue is selecting the most appropriate.
- When dealing with missingness, we must keep in mind that our conclusions could be affected by it.