Imputation

-Applied Multivariate Analysis & Statistical Learning-

MMA chapter 3.12

Lecturer: Darren Homrighausen, PhD

Preamble:

- Define imputation and 'missingness'
- Give three types of missing values
- Outline some unsupervised techniques for imputation
- Give some practical considerations for dealing with missing data
- Look through some R examples

Missing data

Overview

It is very common to find data sets that have missing values, known as missingness

Ways of dealing with missingness are known as imputation methods

Note that missingness only refers to missing values in $\mathbb{X}$, not in $\mathbb{Y}$

(Filling in missing values in $\mathbb{Y}$ would be called prediction)

MMA chapter 3.12

Definitions

Imagine there is a hypothetical feature matrix $\mathbb{X}_c$

(The subscript 'c' means complete)

However, some missingness mechanism has conspired against you

This missingness mechanism will be denoted M and it will look like:

$$M_{ij} = \begin{cases} 1 & \text{if the value is missing} \\ \text{NA} & \text{if the value is observed} \end{cases}$$

(Here, we are using R's missing value indicator NA)

Hence, if we make a matrix $\mathbb{M} = [M_{ij}]$ then we observe

$$\mathbb{X} = \mathbb{X}_c * \mathbb{M}$$

(Here, we mean * to indicate element-wise multiplication as in R)

Thought experiment

Suppose we are running an experiment on a depression therapy

Some of the features we have collected are:

- Dep: depression score, standardized so that negative/positive means no depression/yes depression
- Prox: proximity to the testing center

Additionally, we have an automated phone call that goes out to every participant reminding them of the appointment

Thought experiment

Some scenarios:

- The automated phone call software has a bug and it only calls some participants and not others.
  - $\rightarrow$ Some of the participants forget and don't get evaluated
- Participants are much less likely to come to the testing center when values of Prox are large
  - $\rightarrow$ Some of the participants don't show up when Prox > 5 miles
- If a participant is too depressed, he/she is less likely to come into the testing center
  - $\rightarrow$ A participant doesn't show up if Dep > 2

Each of these scenarios results in missing data.

However, the types of missingness mechanism are quite different

Types of missing

MCAR: Missing Completely at Random

$$\mathbb{P}(\text{missing}|\text{complete data}) = \mathbb{P}(\text{missing})$$

(e.g. Some of the participants forget and don't come to get evaluated)

MAR: Missing at Random

$$\mathbb{P}(\text{missing}|\text{complete data}) = \mathbb{P}(\text{missing}|\text{observed data})$$

(e.g. Some of the participants don't show up when Prox $>$ 5 miles)

MNAR: Missing Not at Random

Everything else.

(e.g. If a participant is too depressed, he/she is less likely to come into the testing center)

LIST-WISE DELETION

An intuitive first take on dealing with missing values is to delete any observation with NAs

This is known formally as list-wise deletion

This is the default of most statistical computing packages/functions

EXAMPLE: If we wanted to do multiple linear regression:

```
summary(lm(Y~X))
...
##
## Residual standard error: 0.171 on 1 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.9746, Adjusted R-squared:  0.9238
## F-statistic: 19.18 on 2 and 1 DF,  p-value: 0.1594
```

Implications of list-wise deletion

MCAR

- Missingness is independent (of the data)

$$\mathbb{P}(\text{missing}|\text{complete data}) = \mathbb{P}(\text{missing})$$

- Only statistical implication is a loss of power

MAR

- Given the observed data, the missingness is independent

$$\mathbb{P}(\text{missing}|\text{complete data}) = \mathbb{P}(\text{missing}|\text{observed data})$$

- Statistical implications are a loss of power and bias

MNAR

- Everything else
- Statistical implications are a loss of power and bias. Needs outside information to address.

Alternatives to list-wise deletion

We will consider three alternatives to list-wise deletion

(there will be other, supervised imputation methods later e.g. RandomForest)

- Feature-wise deletion
- Mean/median/mode imputation
- Iterative imputation

Feature-wise deletion

Instead of removing rows of $\mathbb{X}$ (list-wise), we can remove columns of $\mathbb{X}$ (feature-wise)

A general guideline would be if a feature is missing more than 33%, use feature-wise deletion and then apply list-wise or another imputation scheme if any missing values remain

Note: I've found that including a new feature that is an indicator function for whether the removed feature was missing or not can be useful for predictions

Mean/median/mode imputation

For each feature, we can impute the missing values with some measure of centrality

The measure of centrality depends on the feature

- If the feature is quantitative with low skewness $\rightarrow$ mean
- If the feature is quantitative with high skewness $\rightarrow$ median
- If the feature is qualitative $\rightarrow$ mode

This method of imputation works okI if there is low/no correlation between the features

If there is some correlation between the features, then we can do better

Iterative imputation

We can instead do the following:

0. Use mean/median/mode imputation
1. for $j = 1, \ldots, p$:
   - treat x_j as the supervisor and run a multiple linear regression of x_j on $x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p$
   - impute the missing values of x_j with the fitted values of this regression
2. Repeat step 1. until the imputed values don't change much between iterations
3. Produce $\hat{\mathbb{X}}$, the imputed feature matrix

Note:

- If a feature x_j is qualitative, you would want to do logistic regression or some other classifier
- In general, you can use any supervised method instead of multiple linear regression

Multiple imputation

One general theme of these (single) imputation schemes we've discussed is that they all under estimate the variability in the complete data set

To better represent the omitted randomness, we can instead use multiple imputation

Multiple imputation: Instead of producing a single $\hat{\mathbb{X}}$, we

1. produce K draws $\hat{\mathbb{X}}_1, \ldots, \hat{\mathbb{X}}_K$ from a distribution
2. perform k different analyses using each of the $\hat{\mathbb{X}}_k$'s as the feature matrix
3. Summarize our results via some rule (e.g. averaging)

A classic source is D. Rubin (1987)

Missingness in practice

Possible considerations:

- Data size/complexity

  (Does it fit in RAM?)
- Business purpose

  (Is data precious? Development time?)
- Are any observations/features missing a large fraction of values?
- Type of features

  (Any sparsity? Is multivariate normality appropriate?)
- Any atypical missingness indicators?

  (e.g. using -1000 for income to indicate a missing value)

Postamble:

- Define imputation and 'missingness'

  (Missing values are quite common in practice and imputation is any procedure which estimates these missing values)
- Give three types of missing values

  (MCAR, MAR, and MNAR)
- Outline some unsupervised techniques for imputation

  (Mean/median/mode imputation and an iterative imputation scheme)
- Give some practical considerations for dealing with missing data

  (If development/computation time is the constraint, use a simple method. If data are precious, a more complex scheme would probably be worthwhile)
- Look through some R examples