Handling missing data in data science

Simulating the effects of missing data methods and how to present the results in an interactive plot with Github Pages

Rianne Schouten

1. PhD Candidate, Utrecht University
2. Developer Data & Analytics, Samen Veilig Midden-Nederland

March 20, 2018
Introduction

In this presentation:

1. Missing data methodology
   - What is missing data?
   - Missing data methods
   - Evaluation measures

2. Presentation with Github Pages
   - Why?
   - How?
Missing data methodology

```r
head(inc_data)
```

```r
## Y   X
## 1 3.608044 10.880361
## 2 4.959221 NA
## 3 5.284982 NA
## 4 3.415449 9.040452
## 5 5.147952 NA
## 6 5.872563 NA
```

```r
require(mice)
md.pattern(inc_data)
```

```r
## Y X
## 201 1 1 0
## 799 1 0 1
## 0 799 799
```
Missing data methodology
Missing data methodology

Observed Imputed on X
Missing data methodology
Missing data methodology

1. Drop incomplete rows/columns
2. Imputation
   - random imputation
   - mean/median imputation
   - regression imputation
   - random forest imputation
   - multiple imputation
   - and more...
3. Other methods such as
   - weighting procedures
   - likelihood based methods
   - and more...
Missing data methodology

Simulation study

1. Generate complete data (or use real dataset)
2. Generate missing values in complete data
3. Apply missing data method
4. Perform analysis and compare with complete data

Evaluation measures

- Statistical validity
- Imputation accuracy
- Prediction accuracy
Simulation study:

1. Split a given dataset into 60% training data and 40% testset.
2. Generate missing values.
3. Apply missing data method on the training data.
4. Fit a linear regression model on the completed training data.
5. Apply same missing data method on the test data.
6. Evaluate the performance of the regression model by calculating MSE.
Github Pages: Why?

Simulation with real dataset slump_test

MARZ

MSE vs. Proportion of incomplete rows

- drop
- mean
- median
- regression
- random
- stochastic
Github Pages: How?

https://rianneschouten.github.io/missing_data_science/
https://github.com/RianneSchouten
https://www.highcharts.com/demo
https://pages.github.com/
https://rianneschouten.github.io
Ask me questions or give me feedback:

Rianne Schouten: r.m.schouten@uu.nl

Follow my work: rianneschouten.github.io
What are missing data mechanisms?

- **MCAR**: Missingness is fixed, not related to any variable
- **MAR**: Missingness is related to an observed variable
- **MNAR**: Missingness is related to the missingness itself or to an unobserved variable

Example:

Consider outcome variable ‘income’ and feature ‘age’

- **MCAR**: Some age values are missing, both older and younger ages
- **MAR**: Age values are missing, especially for people with a high income
- **MNAR**: Age values are missing, especially for older people
What are missing data mechanisms: MCAR

Independent of value size, values on ‘feature X’ are missing

MCAR
What are missing data mechanisms: MAR and MNAR based on $Z$

Records with a large value on ‘$Z$’ are missing on ‘feature X’
What are missing data mechanisms: MAR and MNAR based on X

Records with a large value on ‘feature X’ are missing on ‘feature X’