

Data Science Hackathon

Missing Values Treatment

Rianne Schouten

1. University Utrecht, Department of Methodology and Statistics
2. DPA Professionals, Data Science Excellence Program

January 18, 2017

Welcome

Introduction

- ▶ Rianne Schouten
- ▶ DPA Professionals
- ▶ Missing Data Specialist



Universiteit Utrecht



Welcome

Introduction

- ▶ Rianne Schouten
- ▶ DPA Professionals
- ▶ Missing Data Specialist

What do you expect of today?

Welcome

Introduction

- ▶ Rianne Schouten
- ▶ DPA Professionals
- ▶ Missing Data Specialist

What do you expect of today?

In this presentation:

- ▶ What is missing data?
- ▶ How to deal with missing data?
- ▶ Today's challenge

What is missing data?

	Y_1	Y_2	\dots	Y_m
1				?
2		?	?	
\vdots				
\vdots			?	?
\vdots	?			
N				?

Item nonresponse

Unit nonresponse

What is missing data?

```
head(inc_data)
```

```
##      outcome  feature
## 1 5.364150      NA
## 2 5.618218      NA
## 3 5.210518 10.73564
## 4 4.519554      NA
## 5 3.610414      NA
## 6 5.945725      NA
```

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

```
##      outcome feature
## 318          1       1    0
## 682          1       0    1
##           0      682 682
```

What is missing data?

- ▶ 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

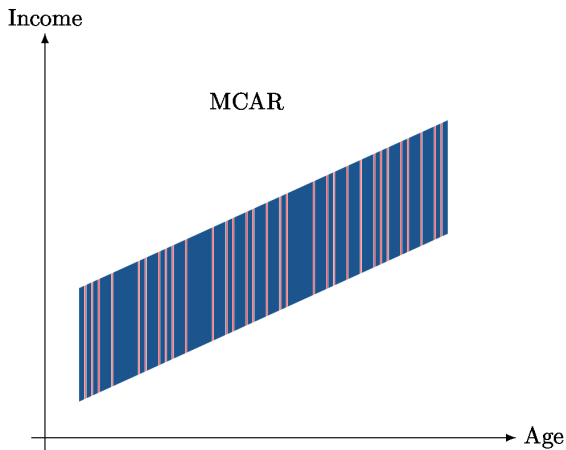
Example:

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

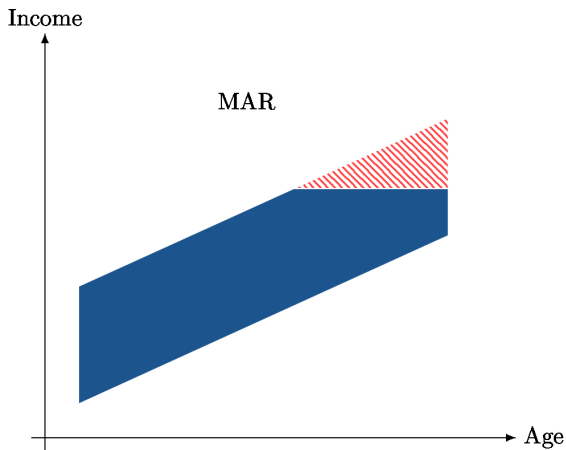
What is missing data: MCAR

Independent of value size, values on feature 'Age' are missing



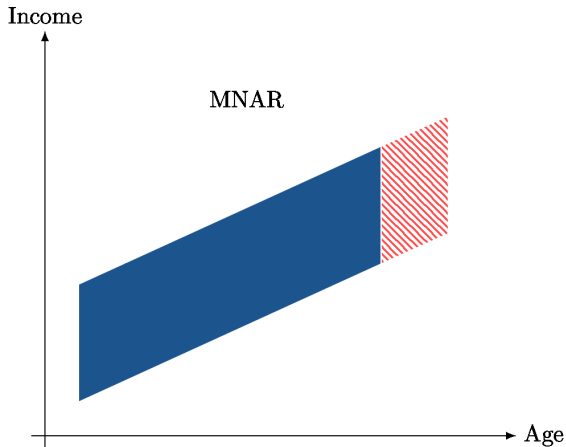
What is missing data: MAR

Records with a large value on 'Income' are missing on feature 'Age'

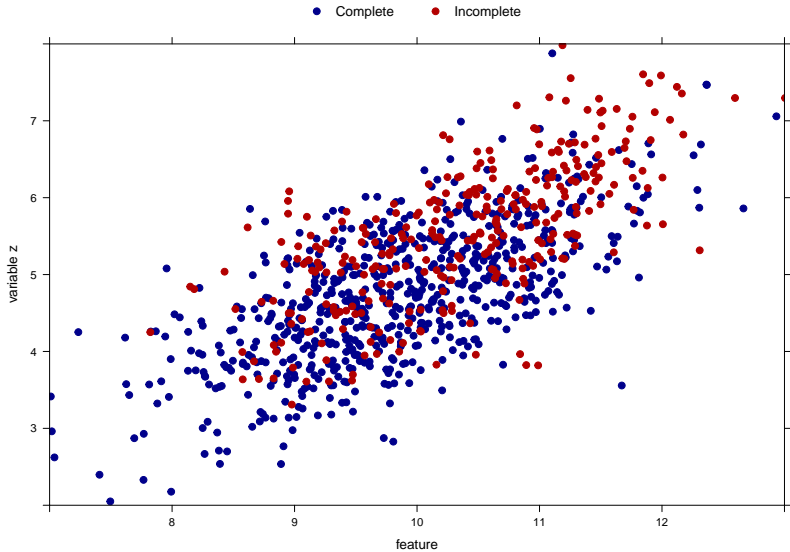


What is missing data: MNAR

Records with a large value on feature 'Age' are missing on feature 'Age'



What is missing data?



How to deal with missing data?

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...

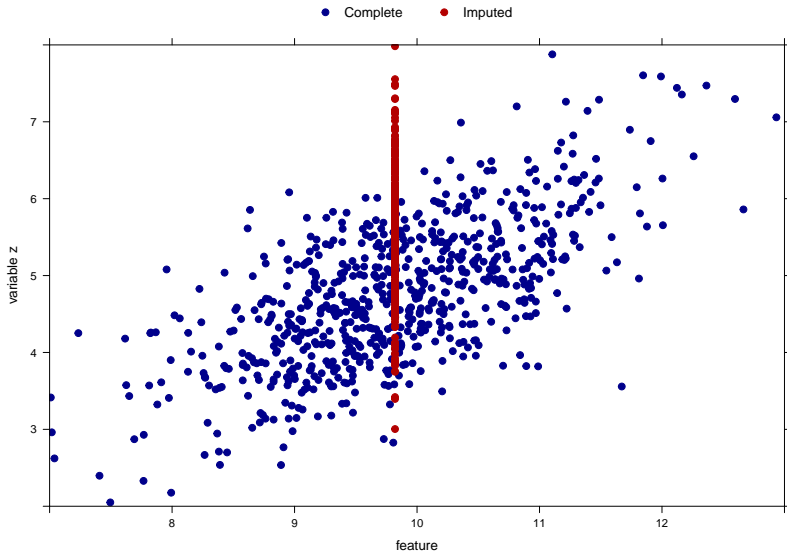
How to deal with missing data: mean imputation

```
head(inc_data)
```

```
##      outcome  feature
## 1 5.364150      NA
## 2 5.618218      NA
## 3 5.210518      NA
## 4 4.519554 9.939668
## 5 3.610414      NA
## 6 5.945725      NA
```

```
com_data <- inc_data
com_data[is.na(inc_data$feature), 'feature'] <-
  mean(inc_data$feature, na.rm = TRUE)
```

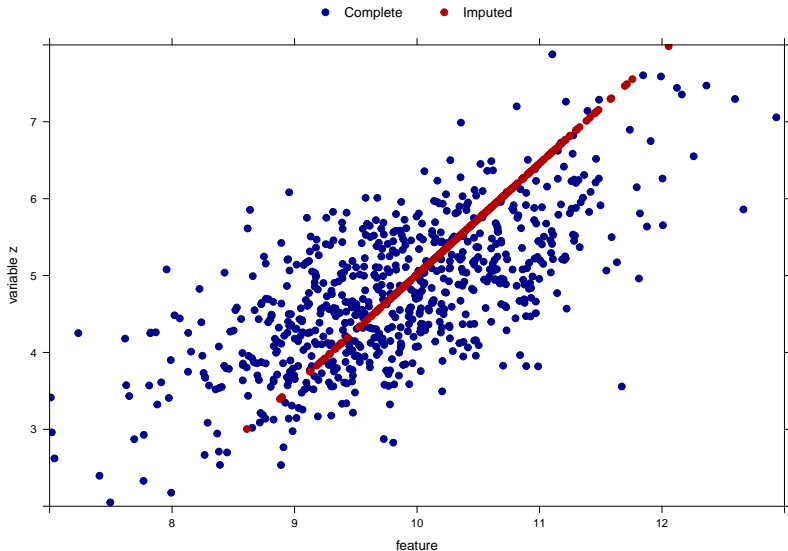
How to deal with missing data: mean imputation



How to deal with missing data: regression imputation

```
fit <- lm(feature ~ z, data = inc_data)
pred <- predict(fit, newdata = ic(inc_data))
com_data <- inc_data
com_data[is.na(inc_data$feature), 'feature'] <- pred
```

How to deal with missing data: regression imputation



Today's challenge

You receive:

- ▶ an incomplete training dataset
- ▶ outcome variable (dummy) is complete
- ▶ missingness in training and testset is comparable
- ▶ combination of MCAR, MAR and MNAR

Make sure:

- ▶ you fit your imputation method on your trainingset
- ▶ and transform/apply on testset

Question:

- ▶ How would you evaluate whether your imputation method is okay?

Contact information

Ask me anything, always:

Rianne Schouten, r.m.schouten@uu.nl, rianne.schouten@dpa.nl

Follow DPA Data Science: dpadatascience.nl

Follow my work: rianneschouten.github.io



Universiteit Utrecht



Work in progress

Simulation with real dataset slump_test

