Handling missing data in R

Workshop R-Ladies

Rianne Schouten

1. University Utrecht, Department of Methodology and Statistics
2. Samen Veilig Midden-Nederland

Download exercises at:
github.com/rianneschouten/workshops/R-Ladies Workshop or
github.com/rladiesamsterdam/2019_Oct_MissingData

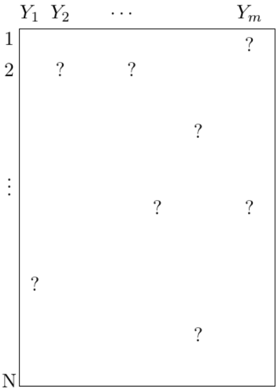
Welcome

What do you want to learn today?

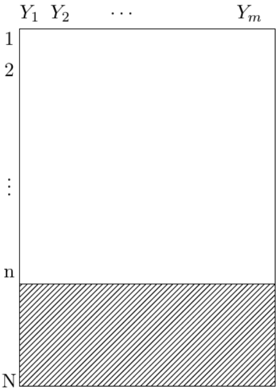
In this workshop:

1. Missing Values Analysis
   - Exercises
2. Implementing Missing Data Methods
   - **break**
   - Exercises
3. Evaluating Missing Data Methods
   - Exercises

1. Missing Values Analysis

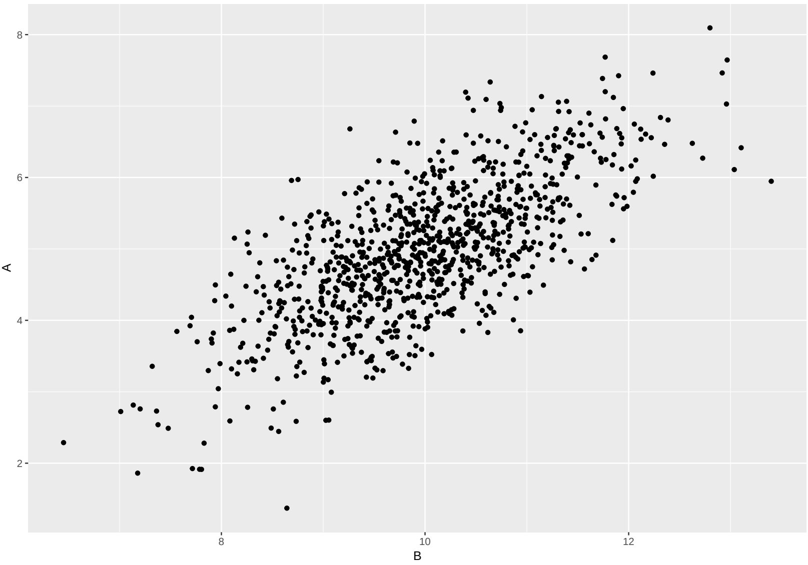


Item nonresponse



Unit nonresponse

1. Missing Values Analysis



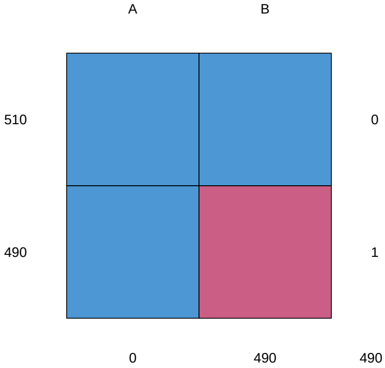
1. Missing Values Analysis

```r
head(inc_data)
```

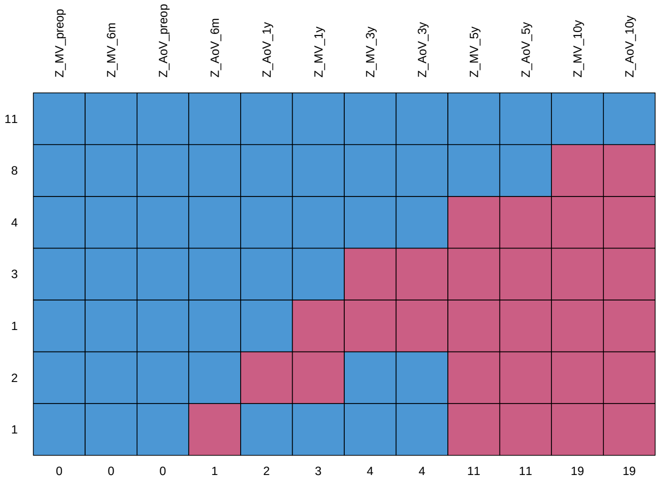
```
##          A         B
## 1 4.353964  8.420317
## 2 5.139020        NA
## 3 6.595914        NA
## 4 2.591296  8.083452
## 5 5.123381 10.667885
## 6 6.204706        NA
```

1. Missing Values Analysis: Where is my missing data?

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



1. Missing Values Analysis



1. Missing Values Analysis: Missingness Mechanisms

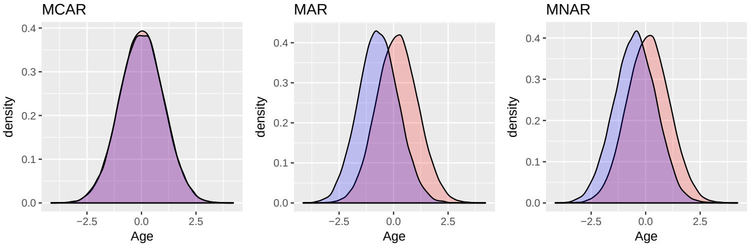
- MCAR: Missingness is not related to any variable
- MAR: Missingness is related to an observed variable
- MNAR: Missingness is related to the missingness itself or to an unobserved variable

For example: Consider variable ‘age’ (Y) and variable ‘length’ (X)

- MCAR: Length values are missing, both shorter and longer lengths
- MAR: Length values are missing for older childeren
- MNAR: Length values are missing for longer children

1. Missing Values Analysis

```r
R <- is.na(inc_data$length)
ggplot(data = inc_data, aes(age)) +
  geom_density(inc_data[R == 1, ], fill = "red") +
  geom_density(inc_data[R == 0, ], fill = "blue")
```



1. Missing Values Analysis: Exercises

https://github.com/RianneSchouten/workshops/tree/master/R-Ladies%20Workshop

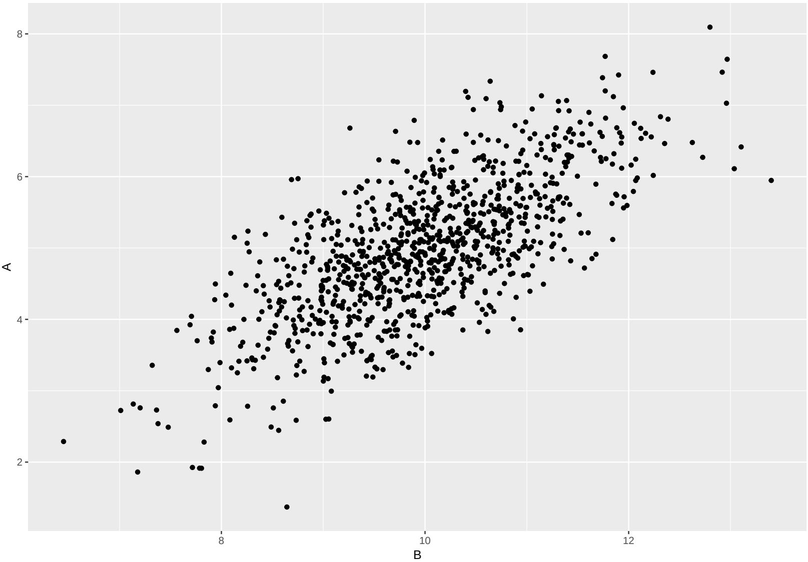
or

https://github.com/rladiesamsterdam/2019_Oct_MissingData

0. Download the folder
1. Open WorkshopRLadiesExercises.R
2. Install the two packages on lines 8 and 9 (wifi)
3. Start!

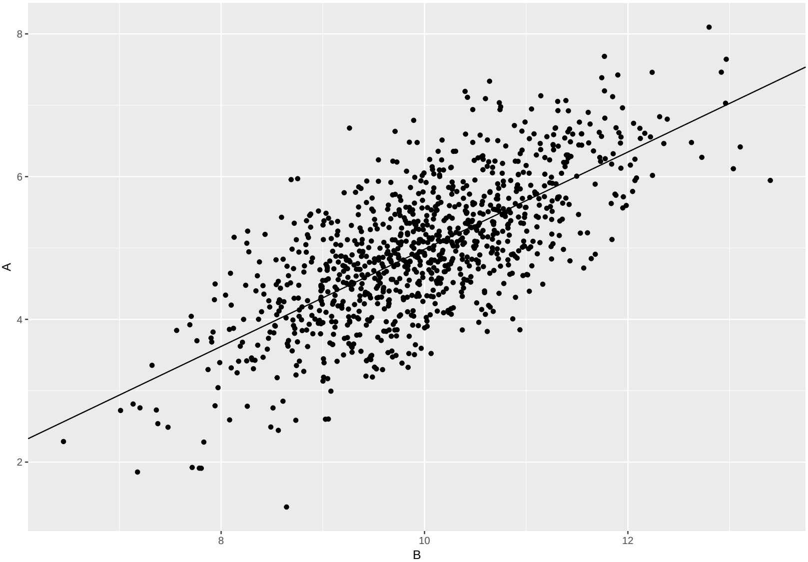
2. Implementing Missing Data Methods

Let's go back to this dataset:



2. Implementing Missing Data Methods

Let's go back to this dataset:



2. Implementing Missing Data Methods

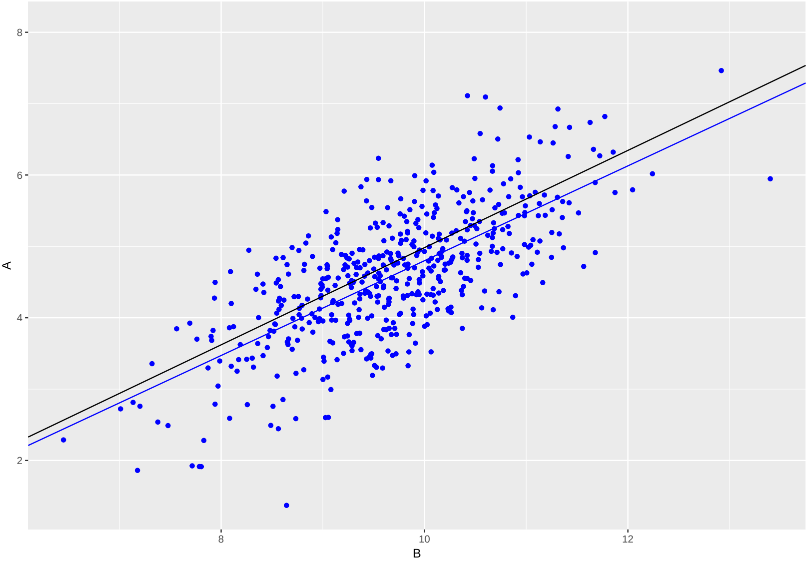
But we have:

2. Implementing Missing Data Methods

1. Methods without imputation
   - Listwise deletion (drop): `na.rm = TRUE` or `na.action = na.exclude`
   - Indicator method
   - Random forest analysis methods
   - And more . . .

2. Implementing Missing Data Methods

Listwise deletion (drop)



2. Implementing Missing Data Methods

2. Single imputation methods

- Mean imputation: `mice(data, method = "mean", m = 1, maxit = 1)`
- Regression imputation: `mice(data, method = "norm.predict", m = 1, maxit = 1)`
- Stochastic regression imputation: `mice(data, method = "norm.nob", m = 1, maxit = 1)`
- Last observation caried forwards: `tidyr::fill(data, variable)`
- And more...

imputation = filling in missing values

2. Implementing Missing Data Methods

Mean imputation

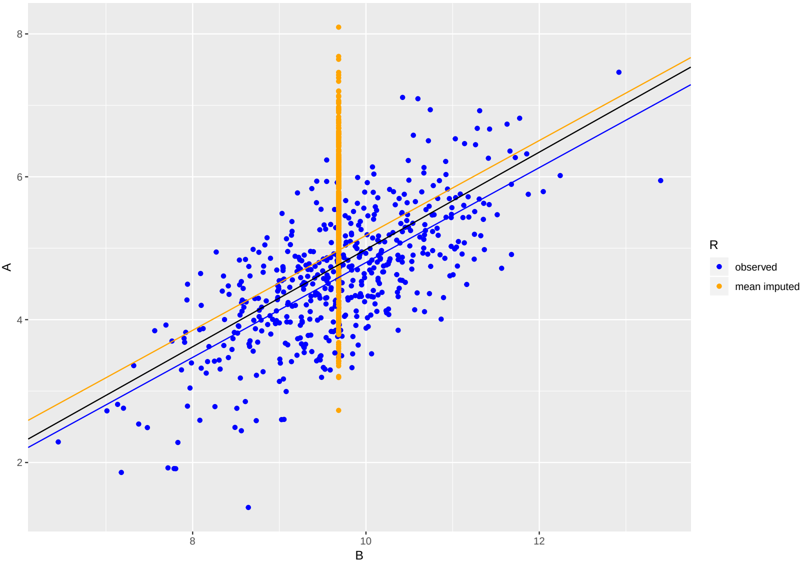
```r
mean_B <- mean(inc_data$B, na.rm = TRUE)
mean_B
```

```
## [1] 9.684854
```

```r
imp_data <- inc_data
imp_data[is.na(imp_data$B), 'B'] <- mean_B
```

2. Implementing Missing Data Methods

Mean imputation



2. Implementing Missing Data Methods

3. Multiple imputation methods

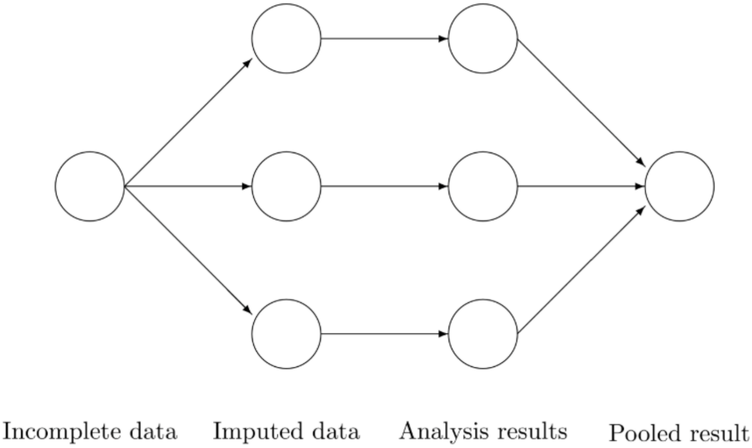


Figure 1.6: Scheme of main steps in multiple imputation.

2. Implementing Missing Data Methods

3. Multiple imputation methods

- Bayesian linear regression imputation
- Predictive mean matching
- And more...

```
imp <- mice(inc_data, method = "norm", m = 5, maxit = 5)
fit <- with(imp, lm(A ~ B))
summary(pool(fit))
```

2. Implementing Missing Data Methods: Exercises

3. Evaluating Missing Data Methods: Two Paradigms

1. Scientific Research
   - Statistical tests with p-values
   - Finding valid statistical estimates: unbiased
   - Comparison of estimates with the hypothesis: realistic standard error

3. Evaluating Missing Data Methods: Two Paradigms

1. Scientific Research

```
true_coefs
```

```
##                Estimate Std. Error   t value      Pr(>|t|)
## (Intercept) -1.8272914 0.21523608 -8.489708  7.424698e-17
## B            0.6810024 0.02145392 31.742567 1.988666e-153
```

```
inc_coefs
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -1.8431856 0.29202129 -6.311819 6.014993e-10
## B            0.6642674 0.03000525 22.138377 1.587365e-76
```

3. Evaluating Missing Data Methods

Use literature to know the best missing data method in your situation

Table 1.1: Overview of assumptions made by ad-hoc methods.

	Unbiased			Standard Error
	Mean	Reg Weight	Correlation	
Listwise	MCAR	MCAR	MCAR	Too large
Pairwise	MCAR	MCAR	MCAR	Complicated
Mean	MCAR	–	–	Too small
Regression	MAR	MAR	–	Too small
Stochastic	MAR	MAR	MAR	Too small
LOCF	–	–	–	Too small
Indicator	–	–	–	Too small

Table 1.1 provides a summary of the methods discussed in this section. The table addresses two topics: whether the method yields the correct results on average (unbiasedness), and whether it produces the correct standard error. Unbiasedness is evaluated with respect to three types of estimates: the mean, the regression weight (with the incomplete variable as dependent) and the correlation.

3. Evaluating Missing Data Methods: Two Paradigms

1. Scientific Research
2. Data Scientists
   - Prediction analysis without p-values
   - Finding good predictions of an outcome variable
   - Comparing trainingset with testset

3. Evaluating Missing Data Methods

A data science pipeline:

1. Split dataset into train/test
2. Data cleaning and feature selection in training set
3. Train prediction model (possibly with crossvalidation procedures)
4. Apply prediction model to test set and evaluate with f.e. mse

Where should the missing data be handled?

3. Evaluating Missing Data Methods

1. Split dataset into train/test
2. Data cleaning and feature selection in training set
   - Perform missing values analysis in training set
     - How similar will the missingness in the test set be?
   - Choose a missing data method
     - Can you save the parameters of the missing data method
     - Is it possible to apply the method in 1 row only?
     - Is it possible to apply the method if the missingness is in another variable?
     - How time consuming is the missing data method?
3. Train prediction model (possibly with crossvalidation procedures)
   - Train the model on the imputed dataset
4. Apply prediction model to test set and evaluate with f.e. mse
   - Make sure to use the parameters of the trained missing data model!

3. Evaluating Missing Data Methods: Exercises

Thank You

Contact information

Rianne Schouten, riannemargarethaschouten@gmail.com

Follow my work: rianneschouten.github.io

Literature

Flexible Imputation of Missing Data Free online version:
https://stefvanbuuren.name/fimd/