R-Ladies Workshop

How to handle missing data in R

Date: 2 October 2019
Length: 1.5 hours

Author: Rianne Schouten
https://rianneschouten.github.io
https://www.linkedin.com/in/rianne-schouten-5bb29491/

Biography

My name is Rianne Schouten and I am a data enthusiast. When I studied Methodology and Statistics, I co-developed a procedure to generate missing data in complete datasets. I implemented the procedure in R-function `ampute`, and it is now part of imputation package `mice`.

Then, I worked as an external PhD candidate under supervision of dr. Gerko Vink and prof. dr. Stef van Buuren. My work focused on the effect of missing data on prediction oriented analyses. Unfortunately, I had to quit with my research when I changed jobs and no more funding was available.

Now, I work at Samen Veilig Midden-Nederland (i.e. a youth care organization) as Developer Data & Analytics. It is my job to structure and standardize the analyses of sensitive and highly classified data. I mainly work with databases and visualization software and I lead several projects.

Workshop description

When you work with data, it is most likely you will encounter missing data. Especially in application-oriented fields, available datasets are almost always incomplete. As most analyses require complete data, what to do with the missing values?

An often used and easy solution to handle missing data is deleting incomplete rows and/or columns from the dataset. Is this method the most appropriate one, though? What other methods are available and what is their influence on the outcome of the analysis?

The workshop has the following three parts.
1. **Missing Values Analysis**
What amount and what kind of missing data do I have in my dataset? What are possible reasons/mechanisms underlying the missingness?
2. **Evaluating Missing Data methods**
How do I know whether I choose the right missing data method? What should I take into account when deciding about a missing data method?
3. **Implementing Missing Data Methods**
How can I implement missing data methods in my analysis and/or in my pipeline?