Assignments - Missing data workshop

dr. Iris Eekhout

2022-05-03

Missing value analyses

Use the built-in dataset `boys` *that is part of the* `mice` *package for the assignments about missing value analyses.*

Assignment 1: amount of missing values

a. How many variables have missing data?
b. How many rows in the data contain missing values?
c. How many overall matrix entries are missing? And how many observed?

Assignment 2: missing data patterns

a. How many different missing data patterns occur in the data?
b. What is the most frequently occurring pattern in the data?
c. Looking at patterns that occur more than incidental (once or twice), which variables happen to be missing together often?
d. Inspect the missing data pairs. With what other variable(s) is height observed together with in more than half of the cases?

Assignment 3: understanding missing data mechanisms

For the assignment about understanding the missing dat mechanisms, we will use a shiny application published on the website www.missingdata.nl/missing-data/missing-data-mechanisms/

The application button, links to a shiny application. In the application you can change the percentage of missing data, the sample size, and the missing data mechanism in the sidebar panel. With these settings data are generated for BMI ratings and knee pain scores with missing observations. The results are shown in a scatterplot, descriptives, histograms and a boxplot.

Set the percentage of missing data to 0 in order to see the complete data (no missings). Set the sample size at 100.

a. What is the mean and standard deviation of knee pain score? And the association between BMI and knee pain (coefficient, standard error and p-value)?

Set the missing data mechanism to "MCAR" and the percentage to 30%. Set the sample size at 100.

b. What are the mean and standard deviation of the knee pain score? What is association between BMI and knee pain?
c. How do these results compare to the complete data results?

Now change the sample size to 250, and set the percentage of missing data back to 0% (no missings). Look at the results and then change the percentage of missing data to 30%.

d. What happens to the association between BMI and knee pain? Explain differences with the previous answer (sample size 100).

Set the missing data mechanism to "MAR" and the percentage to 30%. Set the sample size at 250.

e. What is the association between BMI and knee pain? How does this compare to the association when the data were MCAR?
f. When there are 30% MAR missing data at sample size 250, at what BMI values do missing data on knee pain occur (inspect the scatterplot and the boxplots).

Set the missing data mechanism at "MNAR", the percentage of missing data at 30%% and the sample size at 1000.

g. Comparing the histograms, what knee pain values are mostly missing?

Set the sample size at 300 and the percentage of missing at 50%. Toggle between the three missing data mechanisms.

h. What happens with thee association between BMI and knee pain?
i. What happens with the mean and standard deviation of the knee pain score?

Assignment 4: Evaluating the missing data mechanism

Use the `boys` *dataset again for the assignments about evaluating the missing data mechanism.*

a. Evaluate the missing data mechanism for the boys data with univariate tests. What are your conclusions?
b. Evaluate the missing data mechanism for the boys data with a multivariate test. What are your conclusions?

Multiple imputation

Use the builtin dataset `nhanes` *from the package* `mice` *for the first assignment (5) on multiple imputation.*

Assignment 5: multiple imputation in mice

Assignment 6: multiple imputation predictor matrix

Assignment 7: multiple imputation methods