Assignments - Missing data workshop

dr. Iris Eekhout

Zurich | 27 & 28 June 2022

Missing value analyses

Assignment 1: amount of missing values

Use the builtin dataset `boys` *that is part of the* `mice` *package for this assignment about missing value analyses.*

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

Use the builtin dataset `boys` *that is part of the* `mice` *package for this assignment about missing value analyses.*

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 builtin `airquality` *data set for the assignments about evaluating the missing data mechanism.*

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

Multiple imputation

Use the builtin dataset `nhanes2` *from the package* `mice` *for the first assignment (5) on multiple imputation.*

Assignment 5: multiple imputation in `mice`

Perform multiple imputation on the builtin `nhanes2` data using all default options.

a. How many imputed datasets are generated?

Analyze each imputed dataset in a linear regression model using bmi as the dependent variable and age, hyp and chl as independent variables.

b. How many sets of results are generated?

Now combine the analysis results into the final results after multiple imputation.

c. What are the most relevant predictors for bmi?

Assignment 6: multiple imputation model amd convergence

Create the predictor matrix for the builtin `boys` data using the `make.predictorMatrix()` function in `mice`.

a. Adjust the predictor matrix so that the variables that have more than 50% missing values are excluded as predictors for the imputation.

Perform multiple imputation on the `boys` data with the predictor matrix designed at assignment 6a with 10 imputations and 10 iterations.

b. What methods are used for the imputation of each variable and explain why these are used.

c. Inspect the iteration plots. What are your observations?

d. Adjust the predictor matrix such that hgt and wgt are not imputed by bmi, or vice versa, and that hgt and wgt are not used as predictors together with bmi. Why do you think that these changes are needed?

e. Use the adjusted predictor matrix to impute the boys dataset again, with 10 imputations and 10 iterations. Inspect the iteration plots again, do you see improvements for hgt, wgt and bmi?

f. Another way to deal with the relation between hgt and wgt with bmi is to use "passive imputation" to impute the bmi variable. Adjust the method and predictor matrix in such a way that the bmi variable is passively imputed by hgt and wgt. Inspect the iteration plots again, do you see improvements for hgt, wgt and bmi?

Missing data in practice

Assignment 7: FIML

Use the builtin `airquality` *data for this assignment*

a. Repeat the example from the lecture on FIML and estimate the descriptive means for the `airquality` variables Ozone, Solar.R, Wind and Temp using FIML esimation.

b. Now estimate a regression between `Ozone` and `Temp` using FIML estimation (Ozone as dependent variable), what is the coefficient for Temp?

c. What is the Fraction of missing information in this regression analysis?

d. Add the other variables in the data as auxiliary variables to estimate the regression between Ozone and Temp. Did the coefficient for Temp change?

Assignment 8: Longitudinal missing data

Use the builtin `fdd` *data for this assignment and select only the following variables: "id", "trt", "age", "sex", "cbcl1", "yc1", "yc2", "yc3". See* `?fdd` *for more information about the dataset.*

a. Which variables have missing data and how many?

b. Analyze the data with a longitudinal model (yc1, yc2 and yc3 are the dependent repeated measurements for ptsd), using time, age, sex and cbcl1 as predictors. What can you conclude about the ptsd scores? How many study participants are used in the analysis?

c. Perform a multilevel multiple imputation with only a random intercept. Use 20 imputations and 10 iterations for the imputations. Repeat the longitudinal analysis and compare the parameter estimates with the estimates from the previous answer. Describe your findings.