MA8701 Advanced methods in statistical inference and learning

Week 3 (L5+L6): Missing data

Mette Langaas

1/7/23

Table of contents

Course homepage: https://wiki.math.ntnu.no/ma8701/2023v/start

(ELS 9.6, and van Buuren: "Flexible imputation of missing data") (**VanBuuren2008?**)

1 Missing data

When performing data analysis we often encounter data sets where some observations have missing values. It is important to understand the underlying mechanism for the observations

to be missing, so that we may treat the missing data appropriately.

If we do not treat missing data correctly, this may lead to * wrong conclusions (the observed data might not be representative for the population under study), * there is a loss of power for statistical tests (or length of CIs) so that conclusions are weaker than if the data were complete, and * the statistical analyses may be complicated since standard methods (assuming no missing data) can´t be used.

We first look at notation and then definitions of missing mechanisms.

1.1 Notation

- $\mathbf{y}$: response vector (no missing values)
- $\mathbf{X}$: the full covariate matrix
- $\mathbf{Z} = (\mathbf{X}, \mathbf{y})$: full responses and covariates
- $\mathbf{X}_{\text{obs}}$: the observed part of the covariate matrix
- $\mathbf{X}_{\text{mis}}$: the missing part of the covariate matrix
- $\mathbf{Z}_{\text{obs}} = (\mathbf{X}_{\text{obs}}, \mathbf{y})$: the observed responses and covariates
- $\mathbf{R}$: indicator matrix (0/1) for missingness/observability of $\mathbf{X}$, where 1 denotes observed and 0 denotes missing. (ELS 9.6 does the opposite, but that is not the common use.)
- θ: some parameter in the distribution of $\mathbf{R}$.

We may use the indicator matrix together with the missing data vector and observed data vector to write out the full covariate matrix.

The missing data mechanism is characterized by the conditional distribution of $\mathbf{R}$;

$$P(\mathbf{R} \mid \mathbf{Z}, \theta)$$

1.2 Missing mechanisms

1.2.1 Missing completely at random (MCAR)

$$P(\mathbf{R} \mid \mathbf{Z}, \theta) = P(\mathbf{R} \mid \theta)$$

- All observations have the same probability of being missing, and
- the missing mechanism is not related to the data (not to observed or missing data).

If observations with MCAR data are removed that should not bias the analyses (but the sample size will of cause be smaller), because the subset of complete observations should have the same distribution as the full data set.

Examples:

- measure weight, and the scales run out of battery
- similar mechanism to taking a random sample

1.2.2 Missing at random (MAR)

$$P(\mathbf{R} \mid \mathbf{Z}, \theta) = P(\mathbf{R} \mid \mathbf{Z}_{obs}, \theta)$$

- The probability of an entry to be missing depends (possibly) on observed data, but not on unobserved data.
- Thus, all information about the missing data that we need to perform valid statistical analysis is found in the observed data (but maybe difficult to construct a model for this).
- In a regression or classifiation setting this means that the missingness of a covariate may be dependent on the observed response.
- Remark: not dependent on what could have been observed (i.e. what is not observed).

Example:

- measure weight, and the scales have different missing proportions when being on a hard or soft surface

Most methods for handling missing data require the data to be MAR. If you know that the missingness is at least MAR, then there exists tests to check if the data also is MCAR.

1.2.3 Missing not at random (MNAR)

We have MNAR if we don´t have MAR or MCAR.

Then the missing mechanism could depend on what we could have measured (unobserved data) or other observed values (covariates or response that we are noe collecting). Statistical analyses can not be performed without modelling the underlying missing data mechanism.

Examples:

- the scales give more often missing values for heavier objects than for lighter objects
- a patient is too sick to perform some procedure that would show a high value of a measurement

---

1.2.4 Group discussion

Look at data to be used in the future - and say if MCAR, MNAR, MAR? Or come up with data where this is the case themselves? Suggest to create simulation experiments to check the invalidity or validity of complete case analysis under the three types of missingness?

2 Popular solutions to missing covariates

(**brandphd?**)

SE info fra SL i slides og evt annet for å få ryggdekning for når metodene er gyldige.

Use an analysis method that handles missing data: such as trees! More in Part 3.

Complete case analysis: discard all observations containing missing values. This is also called "listwise deletion". Wasteful, but will give valid inference for MCAR. If the missing is MAR a complete case analysis may lead to bias. In a regression setting if a missing covariate is dependent on the response, then parameter estimates may become biased.

Let each variable have a probability for missing values of 0.05, then for 20 variables the probability of an observation to be complete is $(1 - 0.05)^2 0 = 0.36$, for 50 variables 0.08. Not many observations left with complete case analysis. Of cause some variables may have more missing than others, and removing those variables first may of cause lead to less observations that are incomplete

Pairwise deletion: for example when calculating a correlation matrix only complete pairs may enter in the calculation. This is also called "available-case analysis". A covariance matrix constructed from pairwise deletion may not be positive definite.

Single imputation: here each missing value is imputed (filled in) once by some "estimate" and the data set is then complete and standard statistical methods may be used.

Versions of single imputation:

LOCF: Last observation carried forward. Time series etc. Not recommended, unless there is a reason to believe that nothing has changed.

Mean imputation: Replace the missing value with the mean of the covariate over all samples. Will decrease the variability in the data. "Common solution" within machine learning, but not so common in statistics. Will not give valid inference.

Problems with single imputation: Standard errors may be understimated giving too narrow CIs and too low p-values. Why: the imputed data are treated as fixed (known) values.

Likelihood approaches:

Bayesian approaches:

Multiple imputation: Devise a method to construct the distribution of each covariate (that can be missing) based on other covariates (often a regression method). Sample multiple observation for each missing value, and get m complete dataset. Analyse all m dataset and weigh the results together. R: package `mice`. Will give powerful and valid inference for MAR and MCAR

3 Multiple imputation

3.1 Imputation model

3.2 Analysis model

3.3 Rubin´s rules

From the statistics A, B, C then standard errors, confidence intervals and p-values may be found. The assumption used for statistical inference is that the distribution of A can be approximated by a multivariate normal distribution. Thus, the sample size needs to be sufficiently large.

3.4 R example

```
library(mice)
```

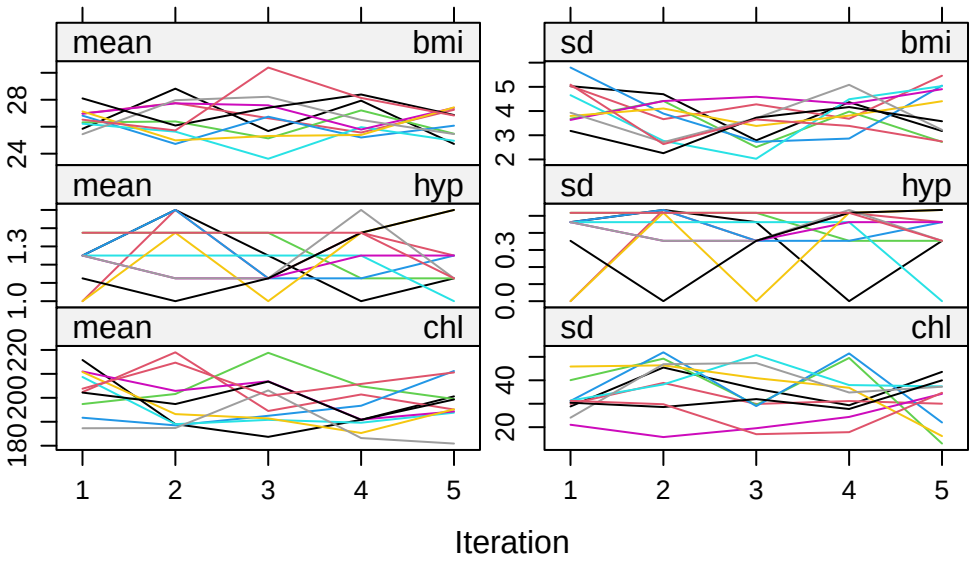
```
imp = mice(nhanes, print = FALSE, m = 10, seed = 24415)
fit = with(imp, lm(bmi ~ age))
# Number of missing observations for our variables
imp$nmis
```

```
age bmi hyp chl
  0   9   8  10
```

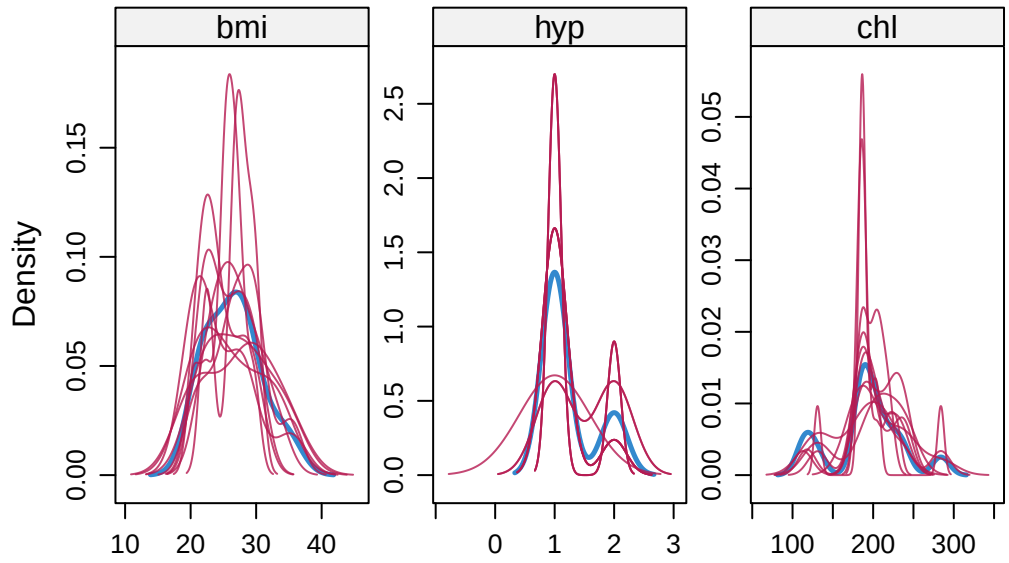
```r
# Summary of mice results
summary(pool(fit))
```

```
         term  estimate std.error statistic       df      p.value
1 (Intercept) 29.621111  2.248428 13.174145 12.62920 9.477666e-09
2         age -1.802222  1.108781 -1.625408 14.42517 1.257173e-01
```

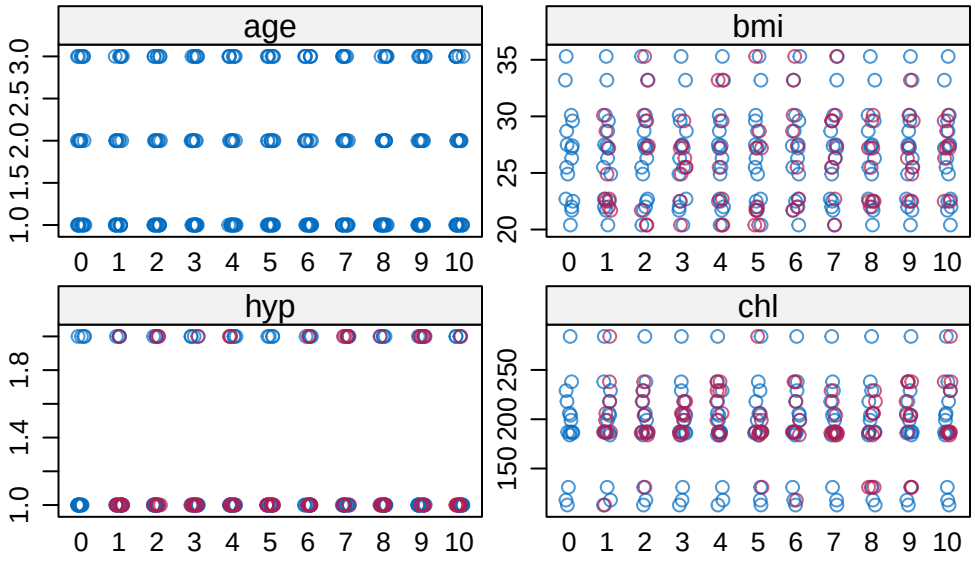
```r
# Trace line plot, can be used to check convergence
plot(imp)
```




```r
# Density of observed and imputed data, observed in blue
densityplot(imp)
```




```
# One dimensional scatter plots for observed and imputed data, observed in blue
stripplot(imp)
```



3.4.1 References

::: {#refs}