# Methods to deal with missing data

K.G.M. Moons, PhD

k.g.m.moons@umcutrecht.nl

# Methods to handle missing data

Complete case analysis (CC)    analysis of the completely observed analysis

       X and Y are completely observed

=       !!when MAR less precise AND biased results!!

Available case analysis (AC)

Missing indicator method (See exercise)

Overall mean/median imputation

Subgroup mean/median imputation

Single (multivariable) regression based imputation

Multiple regression based imputation

# Complete Case

- All analyses (Y = X's models) on the same subset with completely observed X's and Y's

- See previous lecture and exercise:
  - If MCAR (though seldom):
    - Less precision
    - Unbiased/valid results

  - If MAR (as usual + more likely missing data become MAR with much documented/observed other data):
    - Less precision
    - (Very) biased/invalid results

# Example correlation/regression
## Clearly not MCAR - due to change in values/range → association changes
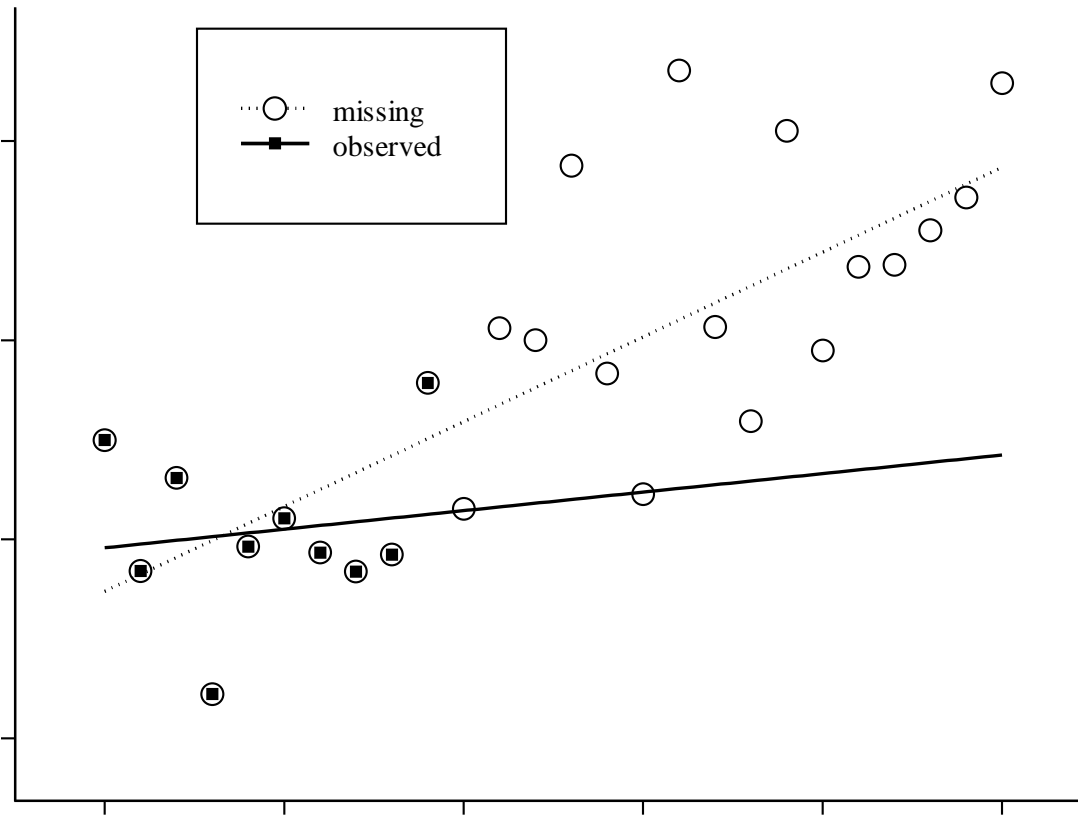


not missing completely at random because all in upper part and not everywhere

would affect the regression line! and therefore biased association

# Complete Case

- CC = in fact redefinition of study sample and thus of the source/target population (domain)
  - E.g. missings in specific hospital/country → study population becomes 'all subjects in completely observed hospitals/countries'

  - E.g. missing outcomes in RCTs → 'Results apply to patients who with complete f-up'

- **Redefinition of your study sample and thus source population is not a solution to a CC analysis if missing data are not MCAR?**

# Methods to handle missing data

arguable where to stop and how many NAs are acceptable- what percentage

Complete case analysis (CC)

**Available case analysis (AC)**

Missing indicator method (See exercise)

Overall mean/median imputation

Subgroup mean/median imputation

Single (multivariable) regression based imputation

Multiple regression based imputation

# Available case

little different from CCA

dangerous for prediction model research ( even more than complete case analysis)

- Use in each sub-analysis (in each model Y=X's) the records with complete data on the X's and Y in that model (sub-analysis)

- Most used method to deal with missing values

- Problems similar as CC
  - Less precision – although AC more efficient → the smaller models (with less X's) use more records/data
  - Unbiased if MCAR ; biased if MAR

- In prediction modeling AC is even a bigger problem than CC: Why?

  - Number of subjects is different for each model
  - Reduced + extended prediction models compared on different subjects
    - Interpretation or even estimation problems (e.g. added value)

less possibility of data being excluded '

# Methods to handle missing data
## Simple methods (ad hoc)

Complete case analysis (CC)

Available case analysis (AC)

**Missing indicator method (See exercise)**

Overall mean/median imputation

Subgroup mean/median imputation

Single (multivariable) regression based imputation

Multiple regression based imputation

# Methods to handle missing data

## The imputation methods!! <span>new ways to handle NAs</span>

Complete case analysis (CC)

Available case analysis (AC)

Missing indicator method (see exercise)

-------------------------------------------------------------------------

**Overall mean/median imputation**

Subgroup mean/median imputation

Single (multivariable) regression based imputation

Multiple regression based imputation

# Imputation methods

- **Imputation is replacement** –> preferred method

- Missing participant values on one or more study variables are replaced by 'predicted values/best guesses' which are based on the observed data

- Can be done if MAR – recall previous lecture and exercise:

    – If association between missingness (yes/no) with other observed variables...

    – ... These other observed variables thus convey information on missingess

    – ... And can thus be used to 'guess/predict' the missing variable value.

    best guessed as the observed values can predict the missing values

# Overall mean/median imputation

- For each missing on variable Z → <mark>overall mean</mark> of that variable Z from observed subject-values is <mark>imputed</mark>
  - Diseased (Outcome+) and Non-diseased (Outcome-) combined

- All imputations have same value for Z → Consequence?

  - <mark>Distributions of Z for Outcome+ and Outcome- merge (MORE overlap)</mark>
    the merge/overlap more which is why we get a biased estimate

  - Association of Z on Outcome dilutes → Bias

  - Also: distribution of Z becomes narrower <mark>(SD too low)</mark>   lower SD as reality
    - SE's of association underestimated → too often significant      p value too small

ALWAYS: ALSO IF MCAR!

# Methods to handle missing data

## The imputation methods!!

Complete case analysis (CC)

Available case analysis (AC)

Missing indicator method

----------------------------------------------------------------

Overall mean/median imputation

**Subgroup mean/median imputation**

Single (multivariable) regression based imputation

Multiple regression based imputation

# Subgroup mean/median imputation

- A priori relevant subgroups are defined → based on associations with variable (Z) with missings
  - E.g. per outcome category, sex, age groups, etc.

- Estimate mean/median for subgroup
  - For each missing on Z → subgroup mean is imputed

- More variations in imputed values than overall mean/median  standard error
  - Less bias though SE's still underestimated
  - Limited number of co-variates can a-priori be defined
  - Requires categorisation for continuous variables (loss of information)

# Methods to handle missing data

## The imputation methods!!

Complete case analysis (CC)

Available case analysis (AC)

Missing indicator method

----------------------------------------------------------------

Overall mean/median imputation

Subgroup mean/median imputation

**Single (multivariable) regression based imputation**

**Multiple regression based imputation**
**NEXT LECTURES**

# References

– A gentle introduction to imputation of missing values (Donders JCE 2006)

– Handling missing data in multivariable diagnostic research: a clinical example (van der Heijden JCE 2006).

– Using the outcome variable to impute missing values of predictor variables: a self fulfilling prophecy? (Moons JCE 2006)

– To Impute is better than to ignore (Janssen JCE 2010)

– Dealing with missing values when validating a prediction model (Janssen Clin Chem 2009)

– Imputation of missing outcomes in observational and randomised studies (Groenwold AJE 2012)

– Little et al; New Engl J Med 2012

– Randomized trials with missing outcomes: what to report and how to analyze (Groenwold et al, *CMAJ 2014)*

# Thank you for your attention