Maastricht UMC+

Welcome to the KEMTA Masterclasses

The department of clinical epidemiology and HTA (KEMTA) of the MUMC+ organizes monthly masterclasses for anyone interested in (methods of) scientific research. You can find our masterclasses (both the presentations and upcoming topics) on the KEMTA website:

https://www.mumc.nl/research/infrastructuur-en-ondersteuning/partners/kemta/masterclasses

https://www.linkedin.com/company/klinische-epidemiologie-medical-technology-assessment-kemta/

Maastricht UMC+

Maastricht UMC+

MISSING DATA

From Swiss cheese to valid outcomes

Lloyd Brandts, Sander M.J. van Kuijk
Department of Clinical Epidemiology and Medical Technology Assessment
lloyd.brandts@mumc.nl

Maastricht UMC+

SHORT REPORT Open Access

Common statistical and research design problems in manuscripts submitted to high-impact medical journals

Sara Fernandes-Taylor[1*], Jenny K Hyun[2], Rachelle N Reeder[1] and Alex HS Harris[1]

"Frequently, researchers fail to mention the missing data in their sample or fail to describe the extent of the missing data"

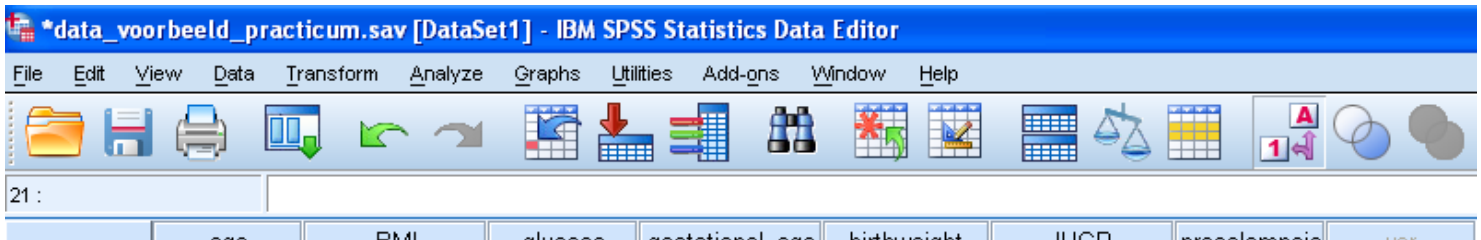
"… those researchers who do discuss missing data often do not describe their methods of data imputation or their evaluation of whether missing data are significantly related to any observed variables "

Theory

- Why bother?
- Prevent incomplete data
- Imputation methods
- Describe incomplete data

Why bother?

- SPSS only uses complete cases for analyses
- Result: loss of precision
- Potentially disturbed results
- Different samples for different analyses
- Ethics?



	age	BMI	glucose	gestational_age	birthweight	IUGR	preeclampsia	var
1	.	24,86	4,2	189	650	Yes	No	
2	.	20,57	6,6	217	1170	No	Yes	
3	26	.	3,4	196	600	Yes	No	
4	26	.	5,0	212	830	Yes	No	
5	25	25,14	.	214	1210	No	No	
6	32	20,44	.	189	650	No	No	
7	28	27,04	5,1	.	1150	No	No	
8	28	35,08	5,6	.	1600	No	Yes	
9	25	32,85	5,8	214	.	No	No	
10	29	20,72	5,3	221	.	Yes	No	
11	29	21,11	5,9	237	1920	.	Yes	
12	30	17,26	5,4	227	1660	.	Yes	
13	25	22,49	5,5	210	1160	No	No	
14	29	24,21	5,3	212	1070	No	No	
15	26	23,78	5,7	198	670	Yes	No	
16	39	25,39	5,6	203	925	No	No	
17	27	24,80	4,4	203	1250	No	No	
18	28	34,77	5,2	238	850	Yes	No	
19	30	26,22	5,4	203	1030	No	No	
20	26	33,46	5,7	201	730	No	Yes	
21	28	23,05	5,1	234	1910	No	Yes	
22	27	25,03	5,8	202	1080	No	Yes	
23	27	24,06	5,5	180	570	Yes	No	
24	29	25,71	5,3	217	1440	No	No	

Mechanisms of incomplete data

- Missing Completely At Random (MCAR)
- Missing At Random (MAR)
- Missing Not At Random (MNAR)

Missing Completely At Random

- Probability of missing is not associated with patient characteristics or outcomes

- Examples:
  - Lab technician drops blood sample
  - Questionnaire gets lost in the mail

- Simplest variant of incomplete data

Missing At Random

- Probability of missing depends on the value of other variables in the data file

- Example:
  - More missings based on age or sex

- Assumption of most missing imputation methods!

Missing Not At Random

- Probability of missing depends on the value itself, or on variables not in the data file

- Examples:
  - Telephone interview about alcohol consumption
  - Questions about income

- Most problematic type of missing data

Comments on these mechanisms

- No empirical methods to discriminate between MCAR, MAR, and MNAR
  - Think carefully about what may have happened during data collection
  - Are there differences between patients whose data is complete and those whose data is not?
- MAR is probably with (large) medical data files

What to do with incomplete data

- Preventing incomplete data
- Delete incomplete cases
- Delete incomplete variables
- Data imputation!

Preventing incomplete data – Design phase

Table 1. Eight Ideas for Limiting Missing Data in the Design of Clinical Trials.

Target a population that is not adequately served by current treatments and hence has an incentive to remain in the study.
Include a run-in period in which all patients are assigned to the active treatment, after which only those who tolerated and adhered to the therapy undergo randomization.
Allow a flexible treatment regimen that accommodates individual differences in efficacy and side effects in order to reduce the dropout rate because of a lack of efficacy or tolerability.
Consider add-on designs, in which a study treatment is added to an existing treatment, typically with a different mechanism of action known to be effective in previous studies.
Shorten the follow-up period for the primary outcome.
Allow the use of rescue medications that are designated as components of a treatment regimen in the study protocol.
For assessment of long-term efficacy (which is associated with an increased dropout rate), consider a randomized withdrawal design, in which only participants who have already received a study treatment without dropping out undergo randomization to continue to receive the treatment or switch to placebo.
Avoid outcome measures that are likely to lead to substantial missing data. In some cases, it may be appropriate to consider the time until the use of a rescue treatment as an outcome measure or the discontinuation of a study treatment as a form of treatment failure.

The Prevention and Treatment of Missing Data in Clinical Trials. RJ Little et al. NEJM. 2012.

Preventing incomplete data - Design fase

1. Include a motivated population
2. Include a run-in period
3. Adopt a flexible treatment plan
4. Consider "add-on" designs
5. Shorten the follow-up period for the primary outcome
6. Allow the use of "rescue medication"
7. Consider a "randomized withdrawal design"
8. Avoid using outcome measures with a high risk of missing data.

The Prevention and Treatment of Missing Data in Clinical Trials. RJ Little et al. NEJM. 2012.

Preventing incomplete data - Execution

Table 2. Eight Ideas for Limiting Missing Data in the Conduct of Clinical Trials.

Select investigators who have a good track record with respect to enrolling and following participants and collecting complete data in previous trials.
Set acceptable target rates for missing data and monitor the progress of the trial with respect to these targets.
Provide monetary and nonmonetary incentives to investigators and participants for completeness of data collection, as long as they meet rigorous ethical requirements.[15,16]
Limit the burden and inconvenience of data collection on the participants, and make the study experience as positive as possible.
Provide continued access to effective treatments after the trial, before treatment approval.
Train investigators and study staff that keeping participants in the trial until the end is important, regardless of whether they continue to receive the assigned treatment. Convey this information to study participants.
Collect information from participants regarding the likelihood that they will drop out, and use this information to attempt to reduce the incidence of dropout.
Keep contact information for participants up to date.

The Prevention and Treatment of Missing Data in Clinical Trials. RJ Little et al. NEJM. 2012.

Preventing incomplete data - Execution

1. Involve experienced researchers
2. Provide acceptable target rates and monitor
3. Use of incentives
4. Reduce burden and inconvenience of collection
5. Provide access to post-study treatment
6. Train investigators and study staff
7. Gather information about potential dropout risk among participants
8. Keep contact information up-to-date for participants.

The Prevention and Treatment of Missing Data in Clinical Trials. RJ Little et al. NEJM. 2012.

Preventing incomplete data

- Making patients realize that stopping study medication ≠ stopping participation study (RCT)
- Realizing that prevention is not always possible
- In that case: how do we deal with it?

Imputation methods

- Replacing the missing value with a (plausible) value
- After imputation, you have a complete data file
- After imputation, standard analysis techniques can be used
- **Does not add any new information!**

Imputation methods

- Last observation carried forward (LOCF)
- Hot-deck imputation
- Imputation with the mean
- Regression imputation
- Stochastic regression imputation
- Multiple imputation
- Expectation-maximization
- Inverse probability weighting
- Etc.

Imputation methods

- Last observation carried forward (LOCF)
- Hot-deck imputation
- **Imputation with the mean**
- **Regression imputation**
- **Stochastic regression imputation**
- **Multiple imputation**
- Expectation-maximization
- Inverse probability weighting
- Etc.

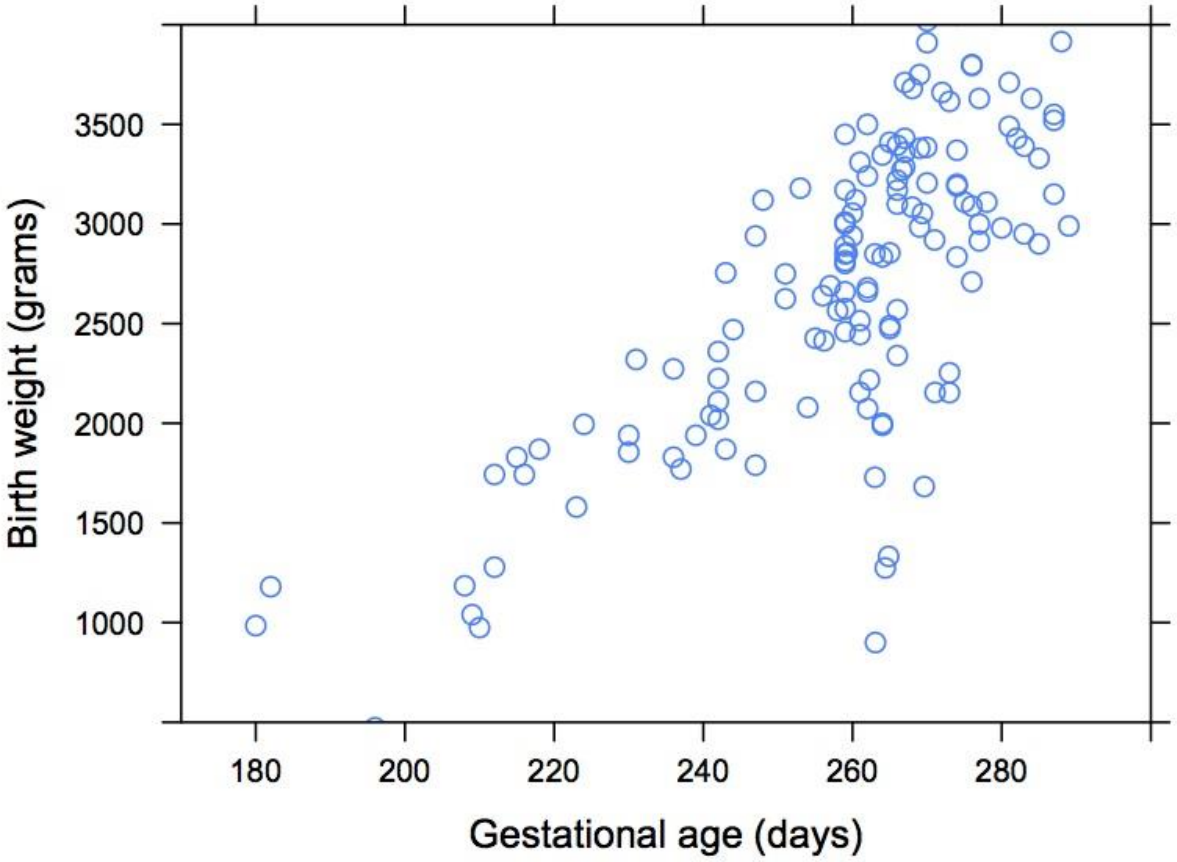
A good imputation method is one that...

... ensures unbiased results

... accurately reflects uncertainty

... in which the uncertainty due to missing data is included

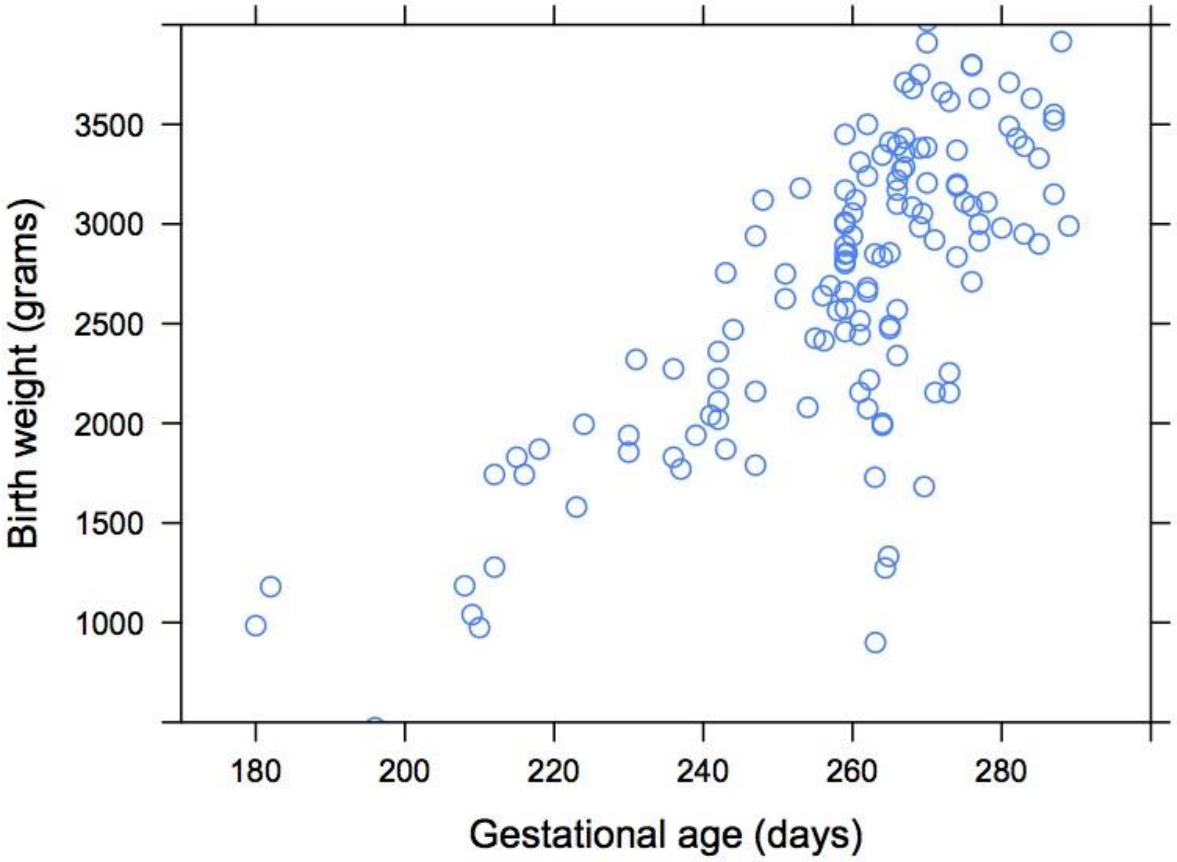
Example: Gestational age and birth weight



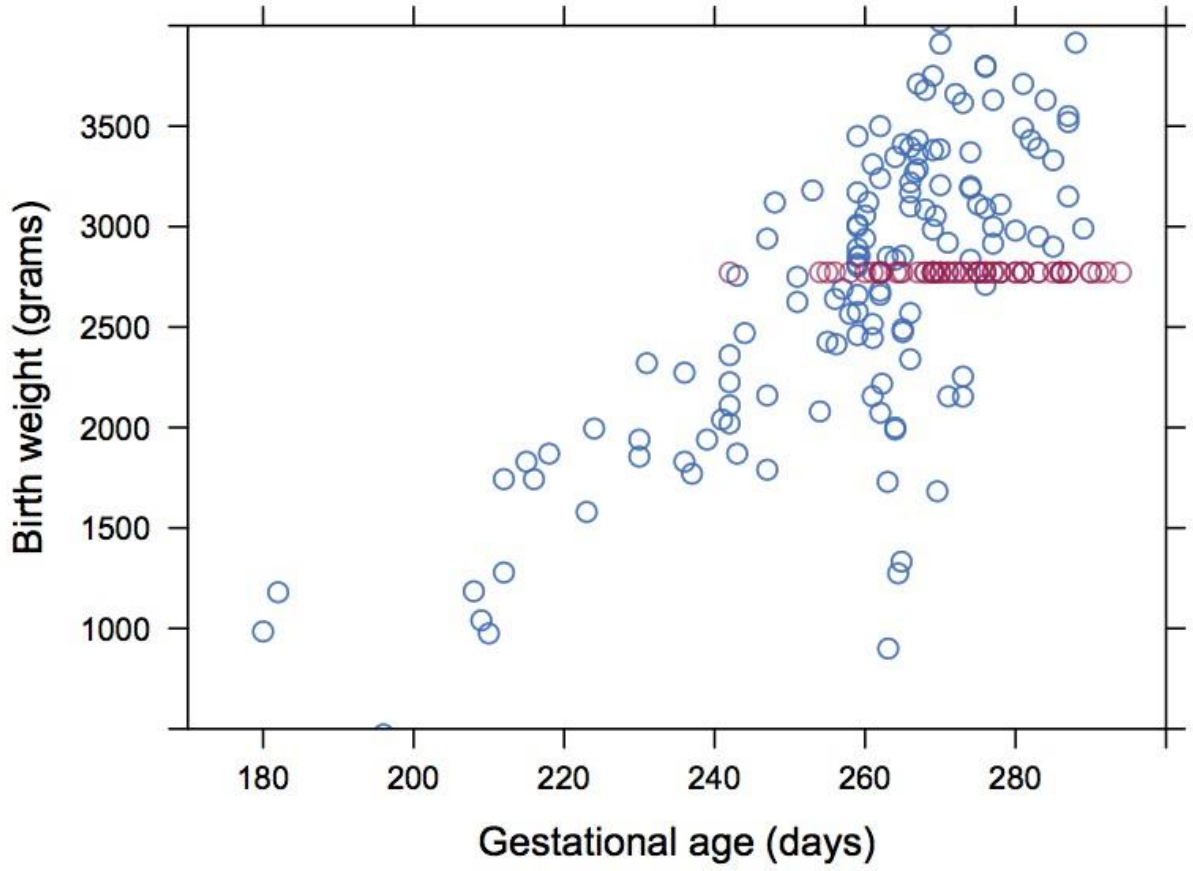
Imputation with the mean

- Easy
- Leads to bias towards the “null”
- Reduces variability within a variable

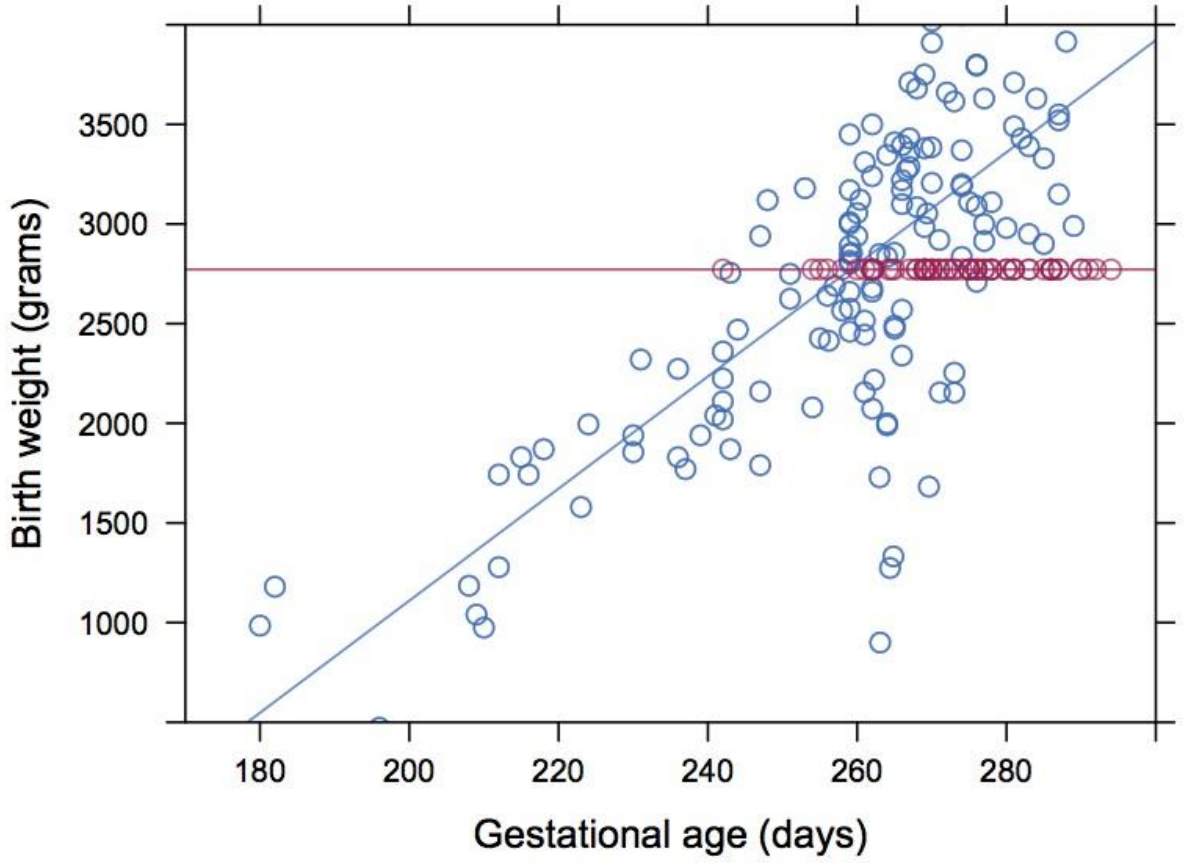
Imputation with the mean



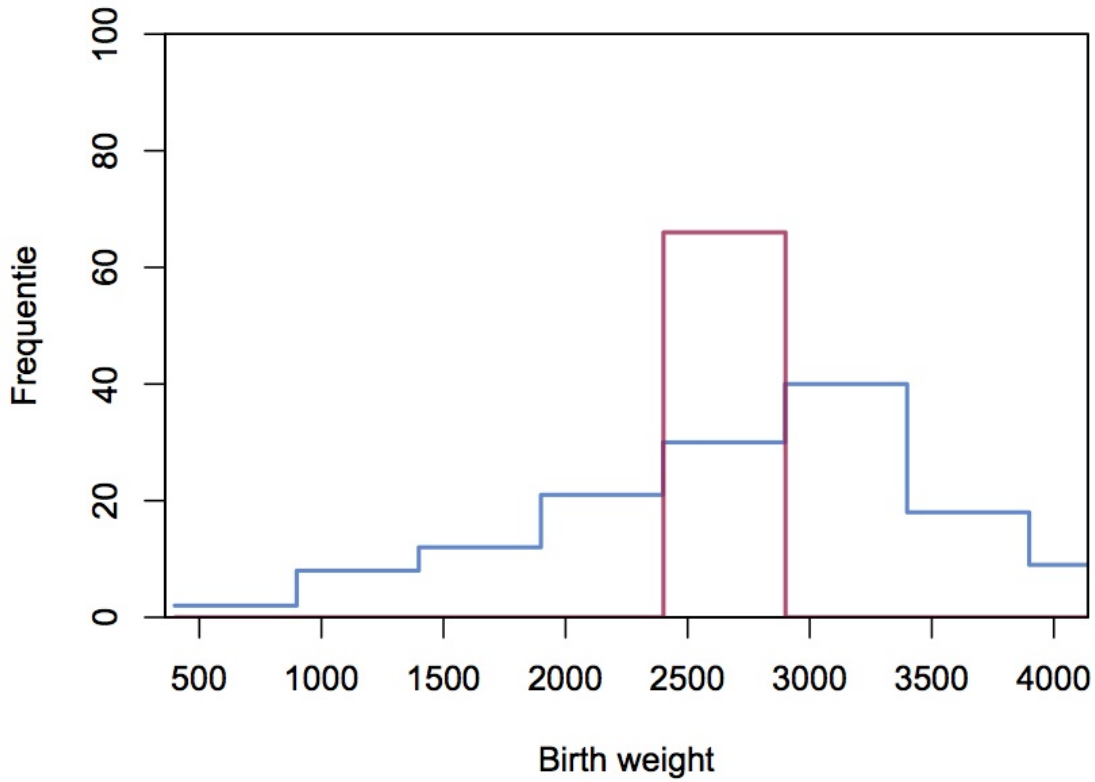
Imputation with the mean



Imputation with the mean



Imputation with the mean



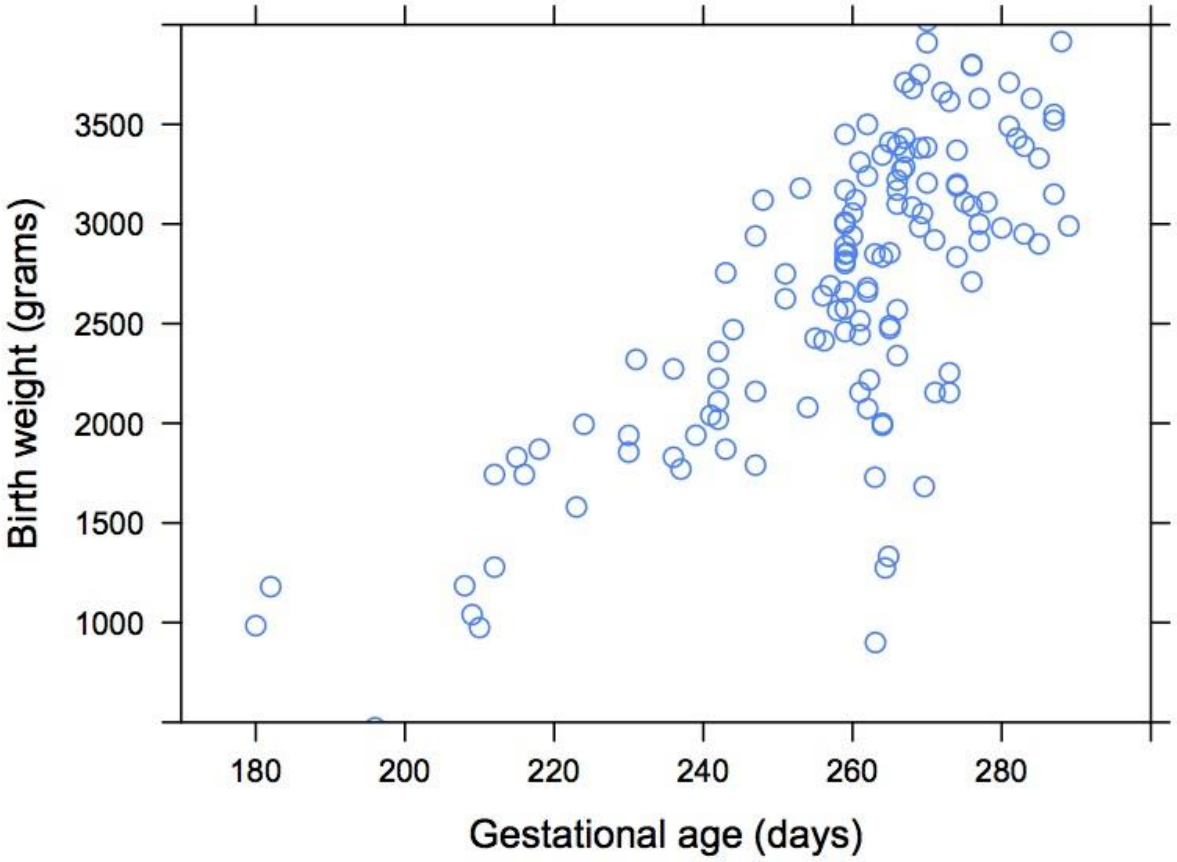
Regression imputation

- Predict each missing value using other variables in the database

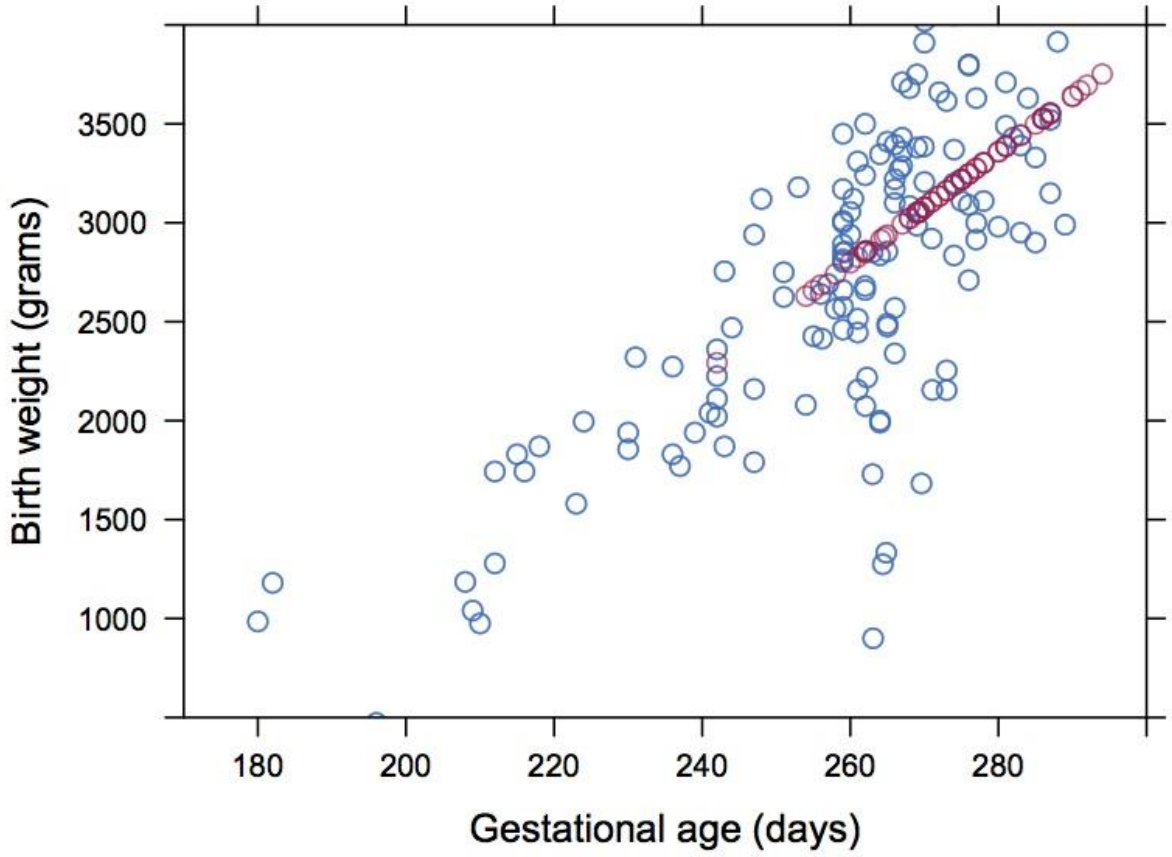
- Advantage:
  - Unbiased results at MAR or MCAR

- Disadvantage:
  - Standard errors (SE's) too small

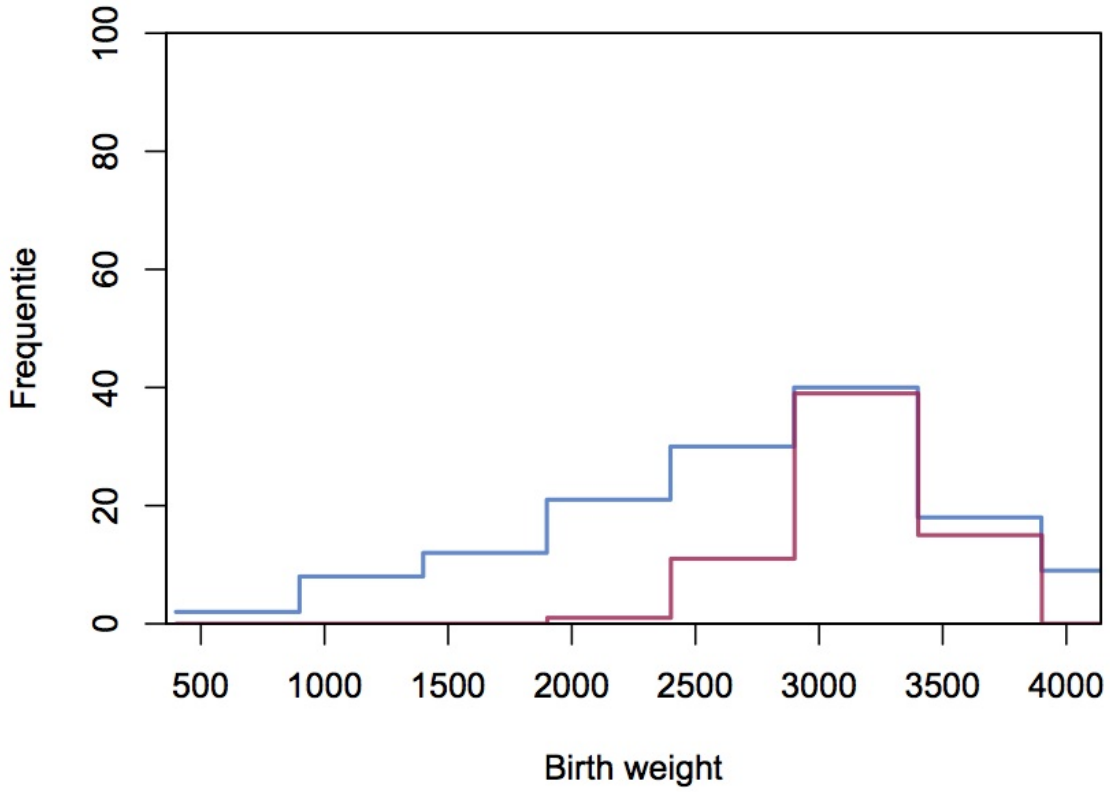
Regression imputation



Regression imputation



Regression imputation



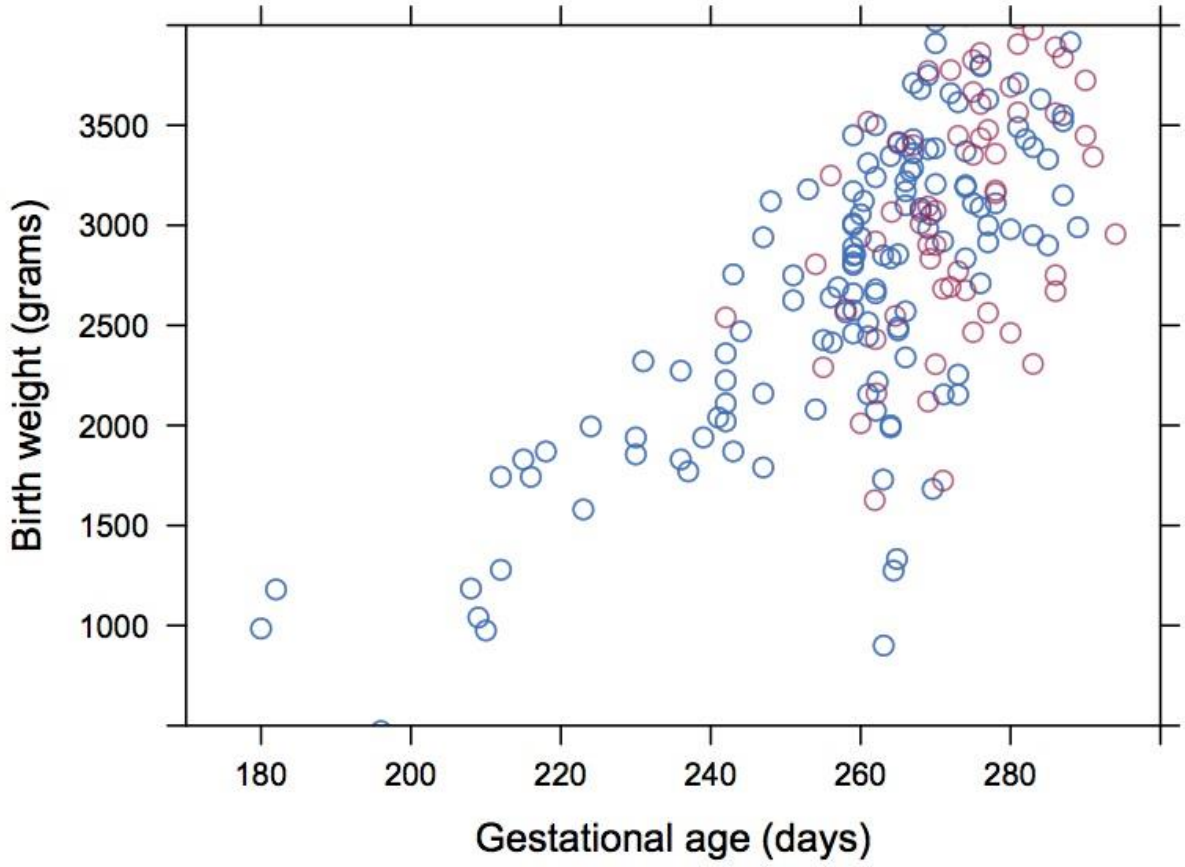
Stochastic regression imputation

- The same as with regression imputation, only now a random residue (error) is added

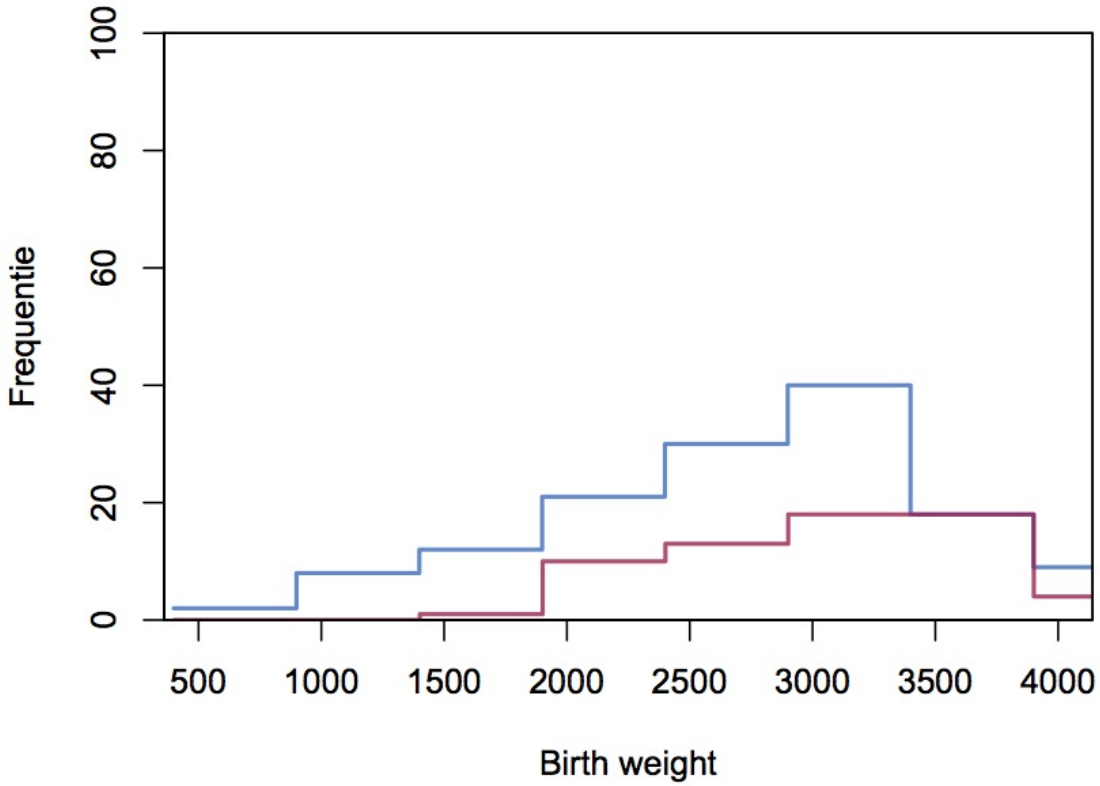
- Advantages:
  - Unbiased results at MAR or MCAR
  - Dispersion in the data is maintained

- Disadvantage:
  - Standard errors still too small: uncertainty of imputation is not included

Stochastic regression imputation



Stochastic regression imputation

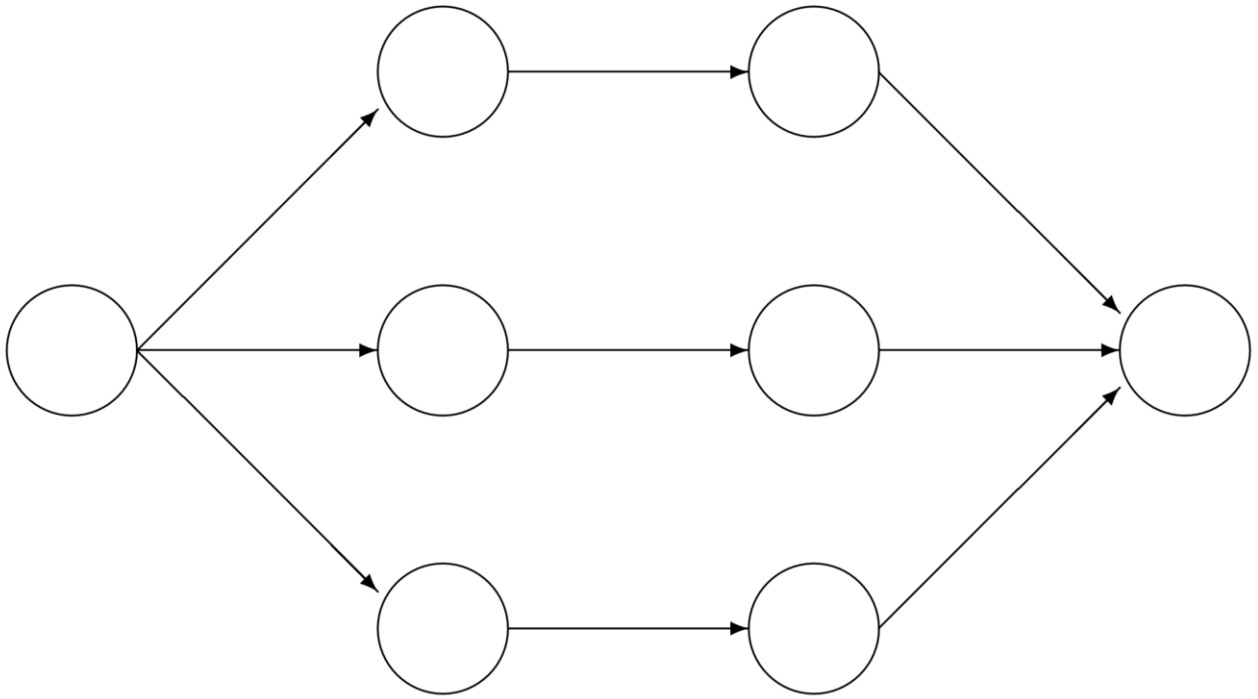


Multiple imputation

1. Create M datasets (M>1, usually 5, 10, or more)
2. Impute any dataset with stochastic regression imputation
3. Analyze all M datasets with standard techniques
4. Combine the M results into one pooled result
   - Point estimates are averaged
   - Standard errors are calculated using Rubin's rules

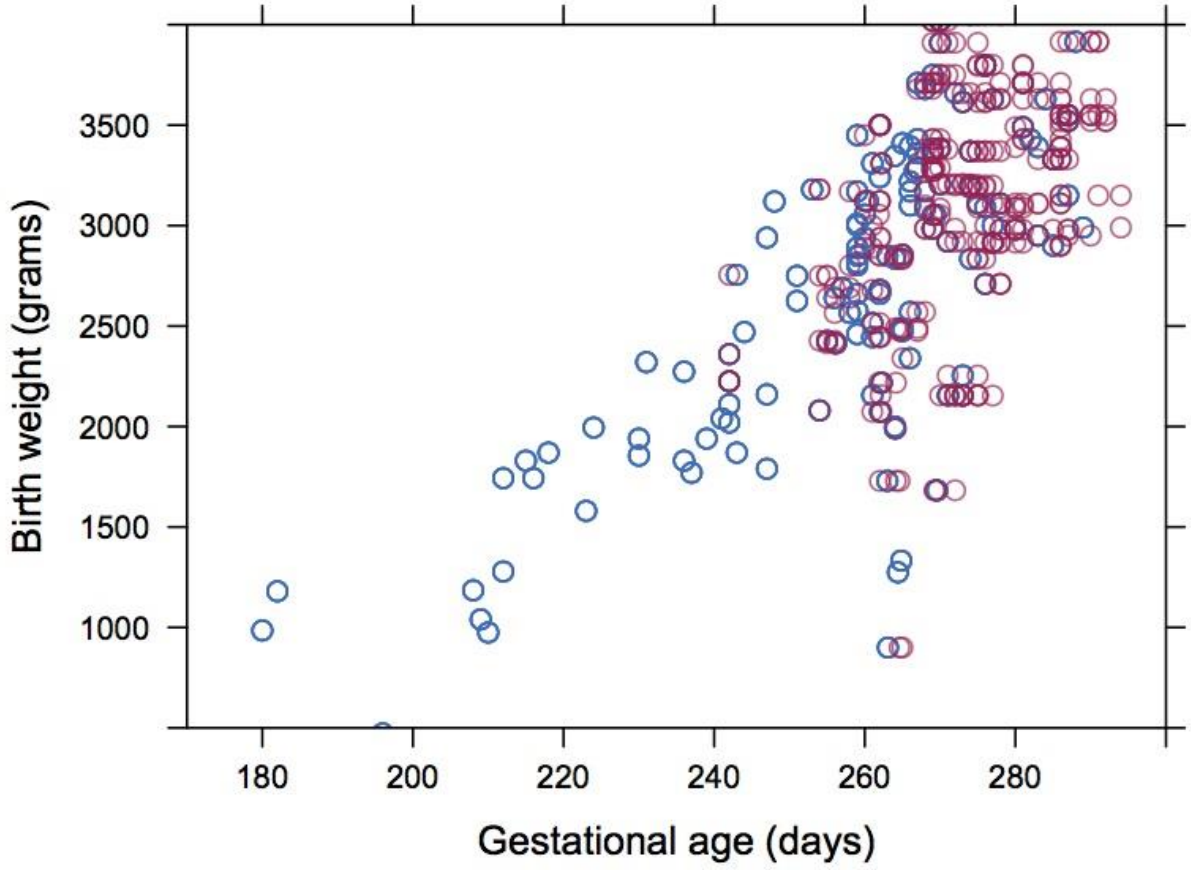
- Advantages
  - Same as with stochastic regression imputation
  - Finally the right standard errors

Multiple imputation



Incomplete data Imputed data Analysis results Pooled result

Multiple imputation



Multiple imputation in practice

- Assumes that data is MAR

- Prediction of values based on:
  - Linear regression (continuous), logistic regression (binary), multinomial log. Regression (categorical)
  - Assumes that continuous variables are normally distributed!

- If they are not normally distributed:
  - Transform before imputation, then transform back
  - Use *Predictive Mean Matching*

Epidemiology Biostatistics and Public Health - 2016, Volume 13, Number 1

Bias in regression coefficient estimates when assumptions for handling missing data are violated: a simulation study

Sander MJ van Kuijk [1,2], Wolfgang Viechtbauer [3], Louis L Peeters [4], Luc Smits [2]

(1) Clinical Epidemiology and Medical Technology Assessment, Maastricht University Medical Centre, Maastricht, The Netherlands
(2) Epidemiology, Maastricht University, PO Box 616, 6200 MD, Maastricht, The Netherlands
(3) Statistics and Methodology, Maastricht University, PO Box 616, 6200 MD, Maastricht, The Netherlands
(4) Obstetrics & Gynecology, Maastricht University Medical Centre, PO Box 5800, 6202 AZ, Maastricht, The Netherlands

Incomplete data reporting*

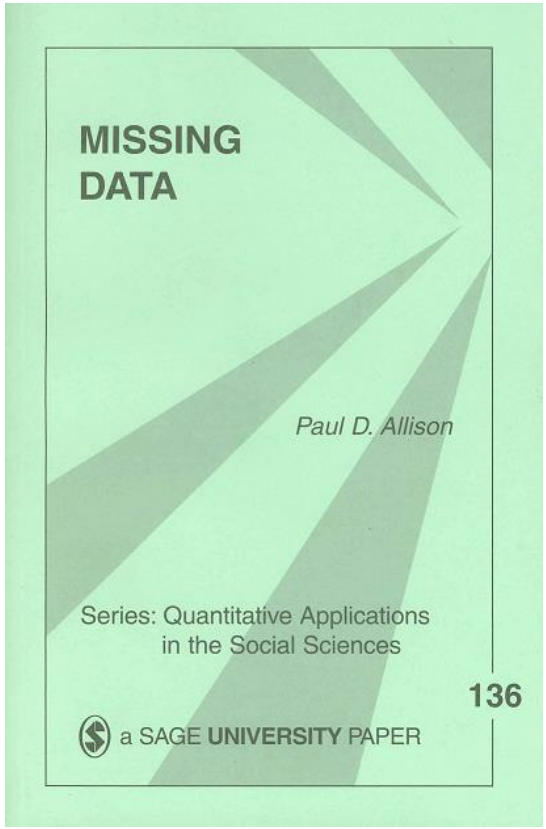
- Report the number (%) of missing values per variable, and the percentage of complete records
- Describe possible causes of incomplete data
- Describe any differences between complete and incomplete patients
- Describe the methods used for dealing with incomplete data!

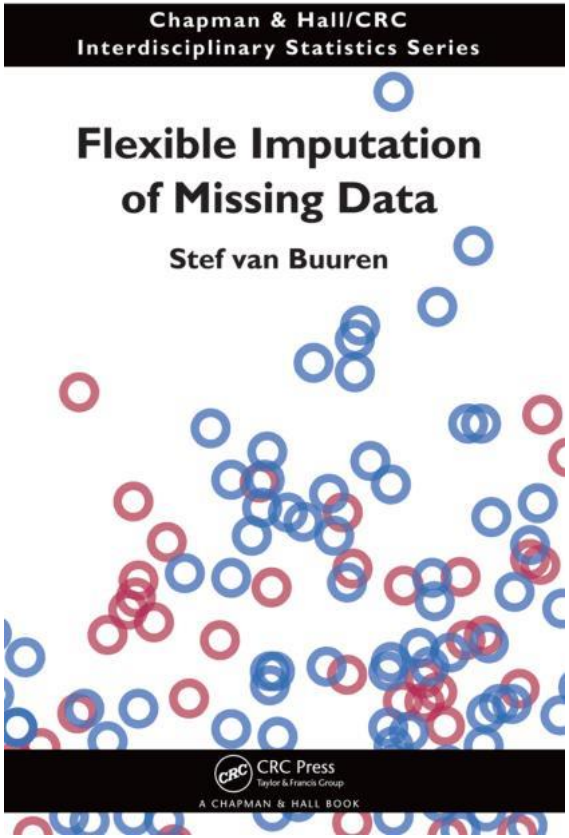
*Sterne et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009.

What did we talk about?

- If data is incomplete, a choice must always be made how to deal with it
- There are no foolproof ways to test which mechanism caused the incomplete data
- Multiple imputation gives (if the assumption is correct!) valid results

Want to know more?





KEMTA Masterclasses 2023

We have planned the following masterclasses this year:

17 January:	**Mixed methods, het beste van beide werelden (Daisy De Bruijn)**
2 February:	**Health Innovation Netherlands: a platform to support med tech innovation (Online)**
7 February:	**Early HTA (Bram Ramaekers en Sabine Grimm)**
4 April	**Het meten van patiënt voorkeuren in de klinische praktijk (Brigitte Essers)**
18 April:	**Missing data (Lloyd Brandts)**
11 May:	**Systematic reviews (Andrea Peeters)**
13 June:	**Patient-gerapporteerde uitkomsten en PROMs in klinisch onderzoek en dagelijkse zorg (Merel Kimman)**
September:	**Predictie (Sander van Kuijk)**
October:	**TBD**
November:	**TBD**
December	**TBD**

Time: van **16.30 tot 18.00u**

To register, for more information, or with suggestions/requests for topics, please contact Irene Vrancken (irene.vrancken@mumc.nl)

Maastricht UMC+

MISSING DATA

Thank you for your attention!

Lloyd Brandts, PhD
Department of Clinical Epidemiology and Medical Technology Assessment (KEMTA)
Contact: lloyd.brandts@mumc.nl

Maastricht UMC+