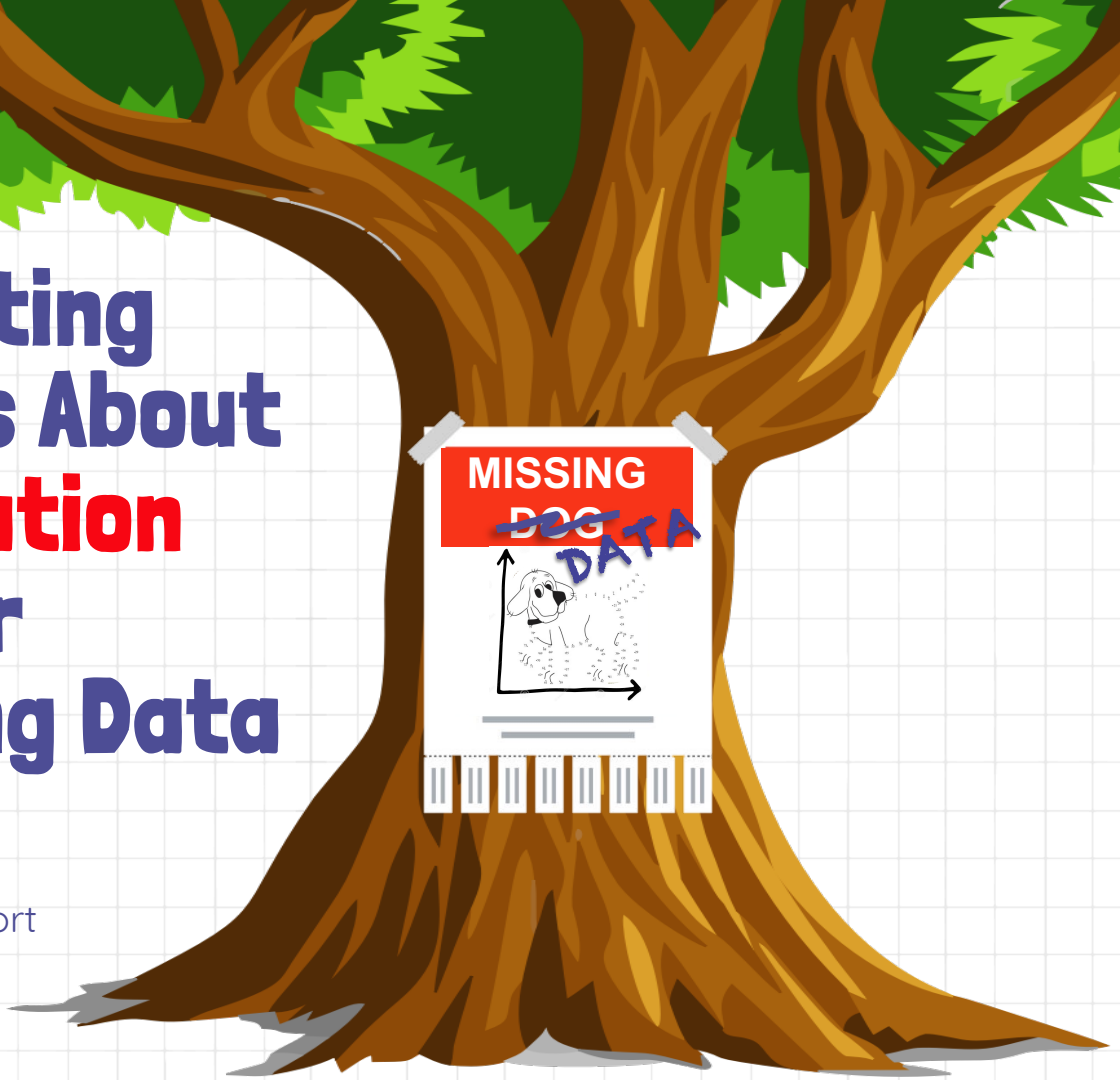
Rebutting Existing Misconceptions About Multiple Imputation as a Method for Handling Missing Data

Joost R. van Ginkel, Marielle Linting,
Ralph C. A. Rippe & Anja van der Voort



Summary

From: **statistical advisors**
To: **applied researchers**

So, you know how multiple imputation is great and all...
No how?

- Missing mechanisms;
- Listwise, pairwise deletion and simple imputation;
- Multiple imputation (MI): how does it work.

Then **why is MI so rarely used**?

- SPSS is evil.
- Distrust:
    1) MI should only be used when the missingness is MAR.
    2) MI should only be used when too few cases are left after listwise deletion.
    3) If results differ from those of listwise deletion, MI must be wrong.
    4) Outcome variables must not be imputed.
    5) Predictor variables must not be imputed.
    6) You will end up with several different outcomes.

Something's missing ...

ID	Y	X1	X2
1	25	1	2
2	20	*NA*	7
3	*NA*	3	5
4	25	*NA*	*NA*
5	32	1	9
6	29	6	11

$$Y = a + b_1 * X1 + b_2 * X2$$

N = 6

N = 3

Something's missing ...

The easiest way to handle missing data is to exclude participants with missing data = **listwise deletion**.

→ *Default* in several statistical software's (SPSS 25.0)

BUT listwise deletion has 2 important **disadvantages**:

😕 **Wasteful**

😱 May lead to **biased results !**

→ Whether this will happen, depends on the underlying **missingness mechanism**.

Missingness mechanisms

MCAR

The probability of missingness does not depend on observed or unobserved data

MAR

The probability of missingness depends on observed data but not on unobserved data

NMAR

The probability of missingness depends on unobserved data

If data is missing at random, participants with missing data are a random subsample of the population of interest.

TRUE

You called 5 PhD students to help you send out some questionnaires, but one of them is hangovered and sends the letters to the wrong address.

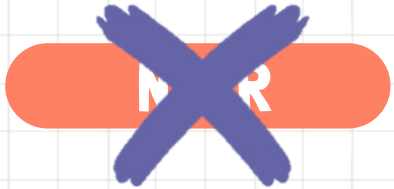
MAR

NMAR

The same PhD student made a mistake in the script and now all SBP values equal to 99 are encoded as missing.

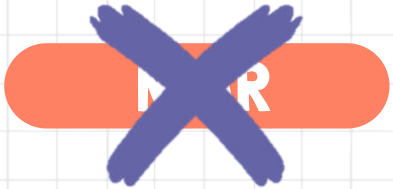
MCAR

MAR

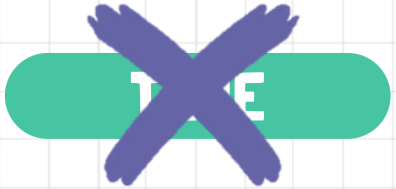


If missing values have the same frequency in all subgroups of people in the sample, the data is...

MAR



If data is missing completely at random, the best guess for the missing value is the sample median.



FALSE

I have the following model:

BMI ~ stress + diet score + sex

Richer participants are less likely to complete the diet questionnaire, but income is not a variable in my model, so I can cosider my data as MCAR.

TRUE

So how do I know which missingness mechanism applies ? YOU DON'T

Test the **MCAR** assumption:

- *T*- or χ^2-test
- Little's MCAR test (Little, 1988)

```
> naniar::mcar_test()
```

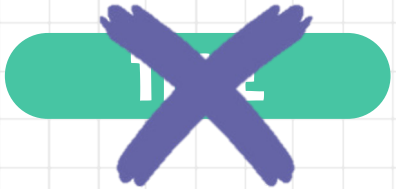
Significant

Data violate the MCAR assumption

Not significant

MCAR assumption is plausible

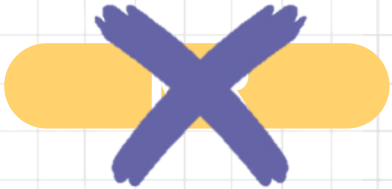
The null hypothesis of Little's MCAR test is that the missing values are randomly scattered across the data.

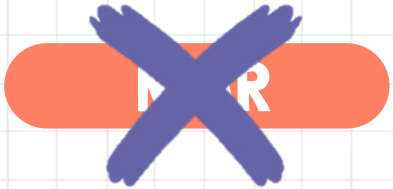


FALSE

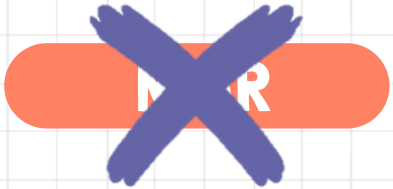
All t-test / Little tests are significant, which missingness mechanism likely underlies my data?

MCAR





All t-test / Little tests are not significant, which missingness mechanism likely underlies my data?



Alternatives to listwise deletion

Pairwise deletion

Reduce data waist

BUT: Limited applicability
Unbiased only under MCAR
Sample size unclear
Computational problems

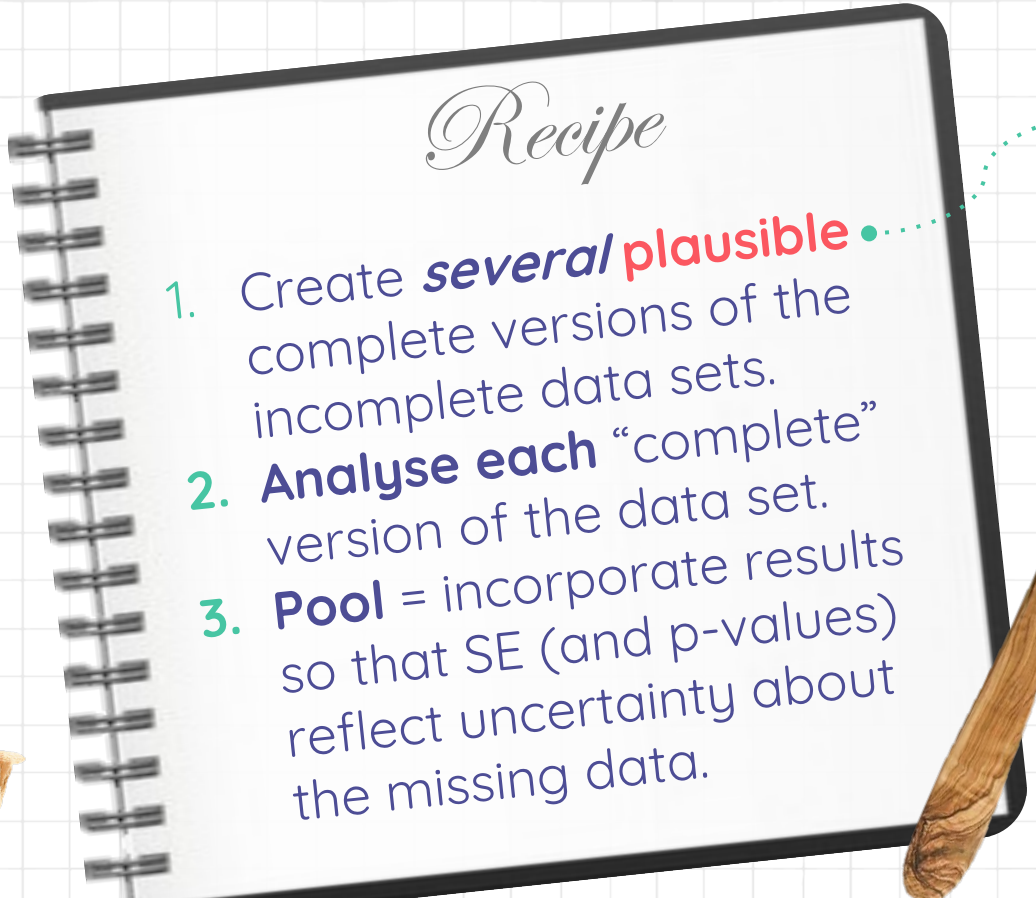
Single Imputation

Reduce data waist & computational problems

1. Variable mean
2. Regression
3. Stochastic regression

BUT: Biased variances and covariances (1 & 2)
Biased p-values and CIs (3)

Multiple imputation (MI)



K but how?

1. Create *plausible* complete data sets

Joint modelling

OR

Fully conditional specification

Regression approach

Predictive mean matching (PMM)

Procedure

1) Fill in starting values, based on the variables' marginal distributions.

For each variable X with missing values:

2) Fit a regression model for predicting its missing values.
3) Based in the model, replace the missing data with:

Random draws from the conditional distribution

The observed value of a "matching respondent"

4) Repeat until properties of the imputed values (i.e., means and SDs) stabilize.

5) Repeat **M** times to obtain **M** multiply imputed data sets.

1. Use MI only for MAR
2. Use MI only if you have too many missing values
3. Trust complete-case analysis over MI
4. Do not impute outcomes
5. Do not impute predictors
6. What do I do with all those results anyway?

1

Multiple imputation should only be used when the missingness is MAR.

Under which missingness mechanisms does MI work?

	MCAR	MAR	NMAR
Multiple imputation	Unbiased	Can correct bias (but that depends on the model)	Cannot correct bias, but can reduce it!
Listwise deletion	Unbiased, but wasteful.	Biased!	Biased! (except very rare cases)

Preferring MI over listwise deletion does not depend on the acting missingness mechanism.

If it works under MAR, it also works under MCAR = MCAR is a *sufficient* but *not necessary* condition for MI.

Significant statistical tests of the MCAR assumption **do not** invalidate the use of MI. What they do invalidate is *listwise deletion*.

Neither the outcome of the MCAR tests, nor the actual underlying missingness mechanism are relevant for deciding whether or not to use MI.

2

Multiple imputation should only be used when too few cases are left after listwise deletion.

Missing data are not only a problem of power reduction

Under both MAR and NMAR, the dropout resulting from listwise deletion will be **systematic** = it will create bias.

MI will completely eliminate this bias under MAR, and partly eliminate it under NMAR.

Theoretically, MI is the preferred choice even with fewer missing values.

3

If results from statistical analyses obtained from multiple imputation differ from those of listwise deletion, the results of multiple imputation must be wrong.

Conclusions obtained from MI and listwise deletion differ: why?

Different results could stem from:

Increased power & correction for bias

Yey !

Incorrectly applied multiple imputation

Tip : check out van Buuren's **dos and don't's**
https://stefvanbuuren.name/fimd/sec-limitations.html

Conclusions obtained from MI and listwise deletion differ: what now?

To do:

- Check abnormal imputed values (e.g. outside data range)
- Compare patterns of observed and imputed values (e.g. in scatter plots or histograms)

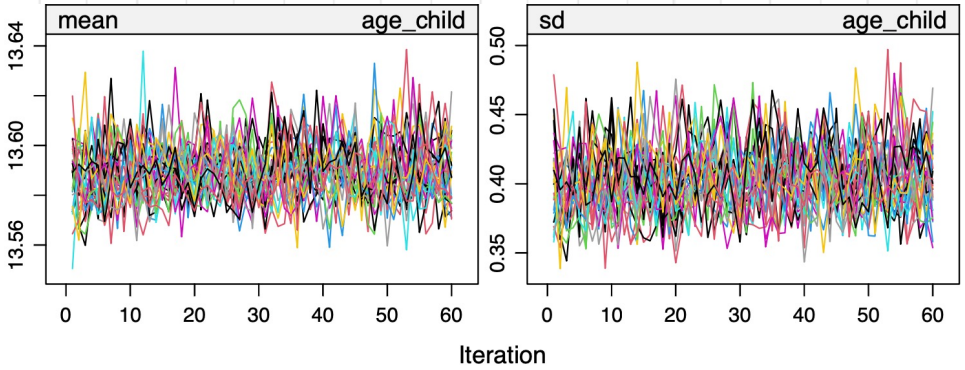
- If you spot anomalies, **adjust the imputation model** (don't just drop MI aside!)
- If there are no anomalies in the imputed values, **check the MCAR assumption** in the original sample.

Imputation quality control in mice

```
> imputation_list <- mice(original_dataset, m = 30)
```

Convergence plots: one or more parameters (e.g., mean and SD) against the iteration number, i.e., the sequence of imputed values (from starting value to final imputed value).
The different streams should be freely intermingled with one another, the variance between imputation chains (i.e., lines) should be equal to the variance within chains.

```
> plot(imputation_list)
```



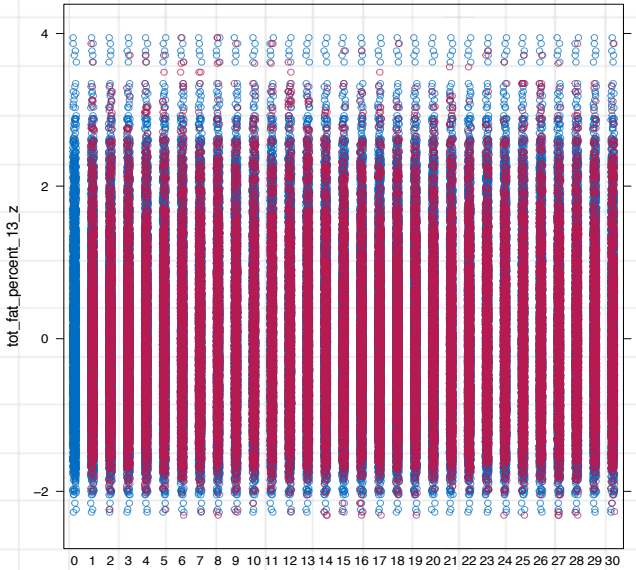
More about convergence?

https://stefvanbuuren.name/fimd/sec-algoptions.html

Imputation quality control in mice

Strip plots: one-dimensional scatterplot of the observed (in blue) and imputed values (in red) in each of the M datasets. Ideal to spot abnormal values, e.g., whether any imputed values fall outside the observed range.

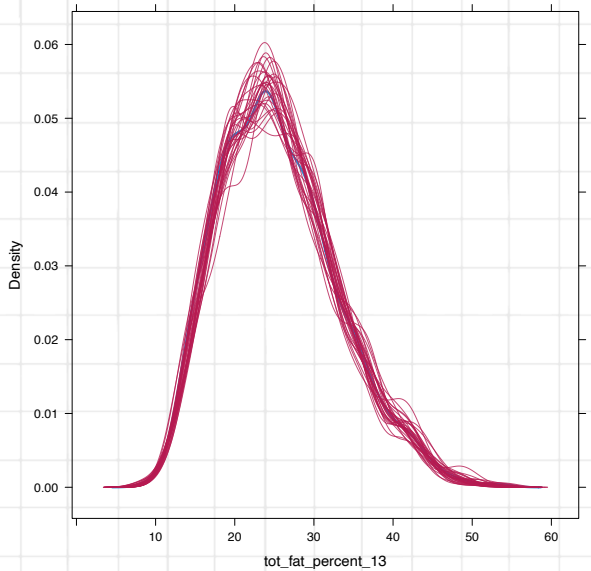
```
> stripplot(imputation_list, tot_fat_percent_13_z)
```

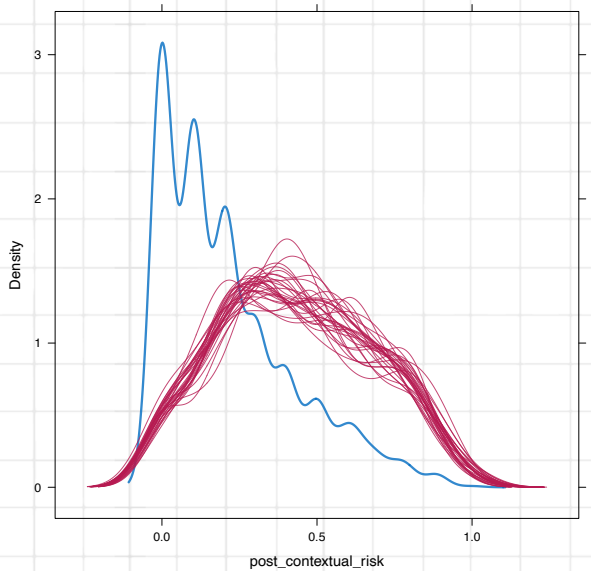


Imputation quality control in mice

Density plots: inspect and compare the density distribution of the observed values (thicker blue line) and the imputed data (M red lines).

```
> densityplot(imputation_list, ~ tot_fat_percent_13)
```

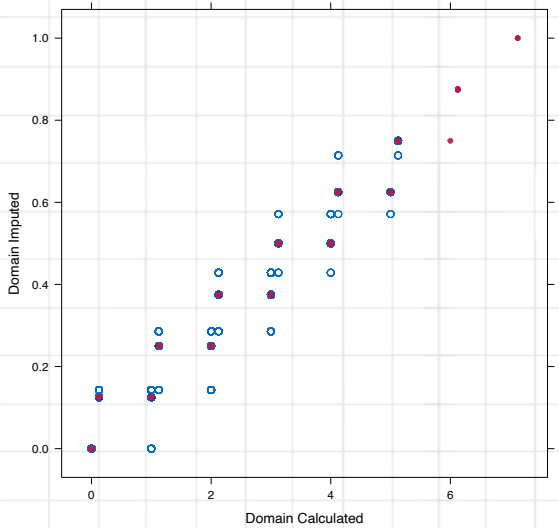




Imputation quality control in mice

Computed vs. imputed plots: sometimes it's best to explicitly compute a variable from the imputed components, rather than agnostically impute those values. You can use this plot to assess the quality of imputation in such cases.

```
> xyplot(imputation_list, post_contextual_risk ~I (material_deprivation + financial_problems + neiborhood_problems + income + unemployed + education / 7),
 na.groups = is.na(original_set[, "post_contextual_risk" ]))
```



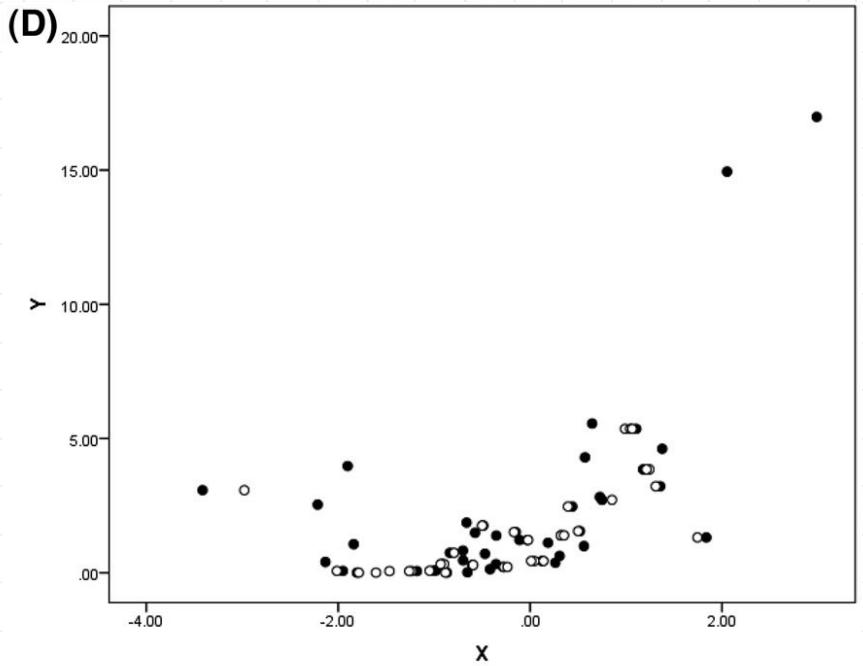
4 Outcome variables must not be imputed.

If an outcome variable were imputed using the same predictors as in the main analysis, wouldn't the imputed values incorrectly confirm the model?

Short answer: Not if the imputation model, the model used for analysis, and the model that generated the data are the same.

But, when that is *not* the case, well, let's see an example...

Simulated data example



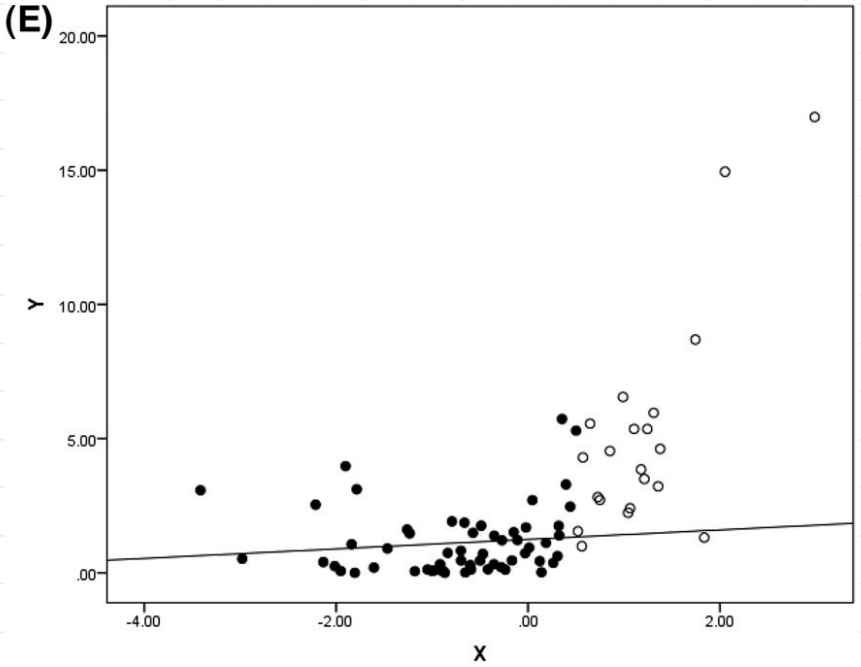
1. We simulate some bivariate data where X and Y are quadratically related.
2. We remove 40% of the data according to MCAR.
3. We *incorrectly* assume that X and Y are linearly related (and use a linear regression model for both multiple imputation and the analysis).

Will the imputed dataset confirm the incorrect (i.e., linear) statistical model ***more*** than when the outcome variable is not imputed?
No. They will give a similar (biased) regression coefficient and a similar (biased) standard error.
… and the rest is an old game of power …

4. We include a nonlinear term of X in the imputation model **or** use PMM.

Do you still think that imputation of Y will confirm the model of interest? Why?

Simulated data example



"Because the relationship between X and Y for the cases with missing data on Y is different than for the cases with observed values on Y."

e.g. the cases with the 40% highest values on X have missing data on Y.

In this case, you may indeed assume an incorrect statistical model and the imputed values can only confirm it.

But wait, what missingness mechanism is this?

So again, *both* MI and complete-case analysis will incorrectly estimate a similar regression line.

DISCUSSION POINT

Is a large *biased* sample better than a small sample?

Is the comparison between sampled and imputed data fair?

5 Predictor variables must not be imputed.

Conceptually, it makes no sense to predict missing data on a variable that is a predictor itself.

e.g., It is logically impossible that someone's age is (partly) influenced by someone's income.

Short answer: **it doesn't matter**. The model used for multiple imputation is not meant as a conceptually meaningful model.

MI is *only* used to accurately describe the relations and structures in the data (and impute data with similar properties).

6

Multiple imputation must not be used because you will end up with several different outcomes of your statistical analysis.

Which result to pick?

In MI **you are not supposed to pick one of the results**; it is the **pooled analysis** that you interpret as the final results.

BUT: the multiple-imputation framework is a work in progress… Pooling methods are not always readily available.

To do:

- ❑ Close SPSS and look for a statistical software package that has more options (e.g., mice in R)
- ❑ Use *ad hoc* methods for pooling the specific statistic.
- ❑ Always be transparent!

Take-home

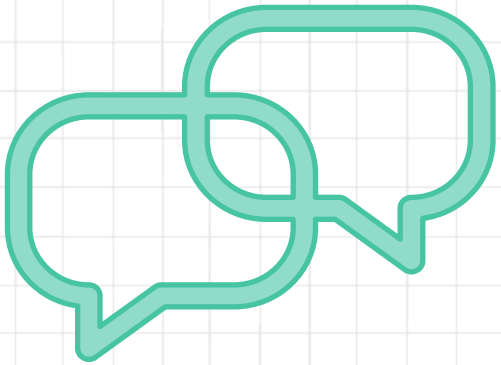
1. MI is (nearly) always better than listwise (and pairwise) deletion.
2. MCAR test should *not* be used to determine whether MI is justified.
3. MI does *not* confirm an incorrectly assumed model any more than complete-case analysis does.
4. Always perform quality control and adjust your MI models.
   e.g., always inspect the data for nonlinear patterns and include those in all models (or use PMM).
5. Any correlated variable is a good candidate predictor for an imputation model.

6. If you experience trouble with the multiple-imputation process, don't simply refrain from using it: contact a statistician!

When not to use MI?

- **Practical reasons**:
  e.g., there are very few missing values and the statistical analysis of interest is one for which pooled results are lacking.

- Some statistical analyses already have a **built-in method** for handling missing data:
  e.g., *full information maximum likelihood* (FIML) is used in latent class analysis, or SEM.
  e.g., *missing data passive* (MDP) cab be used for PCA



Did I *miss* something?