MISSING DATA: what do I do?

Why and ***how*** to handle missing values in (longitudinal) research

Outline

1 The problem with **listwise / pairwise deletion**

- *Selection bias* and *missingness mechanisms*

2 More sophisticated statistical methodologies:

- ❑ **Multiple Imputation** (MI)
- ❑ **Full Information Maximum Likelihood** (FIML)
- ❑ **Propensity Score** (PS) **weighting**

3 Some more sins and mischiefs:

- How much missing is too much (or too little)?
- Imputing *only the covariates*…
- *Unethical* missing value handling

Practical examples

- Compare methods
- Illustrate benefits and challenges

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

$$\mathbf{Y} = \alpha + \beta_1 * \mathbf{X1} + \beta_2 * \mathbf{X2}$$

$$N = 6$$

$$\downarrow$$

$$N = \mathbf{3}$$

"listwise" / "pairwise" deletion

The *problem* with **listwise** (or **pairwise**) **deletion**

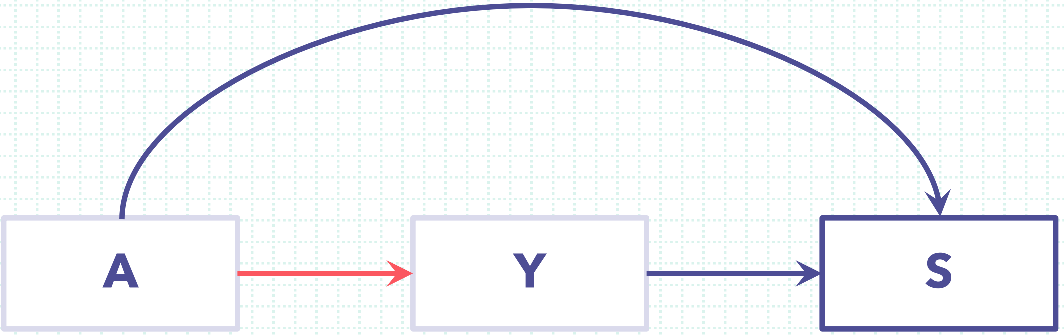
Step 1 : **admit that you have a problem**

😕 *Wasteful*

😱 May lead to **biased results !**

Listwise deletion can have serious implications for the **external** & **internal validity** of research findings.

Generalisability vs. *bias*



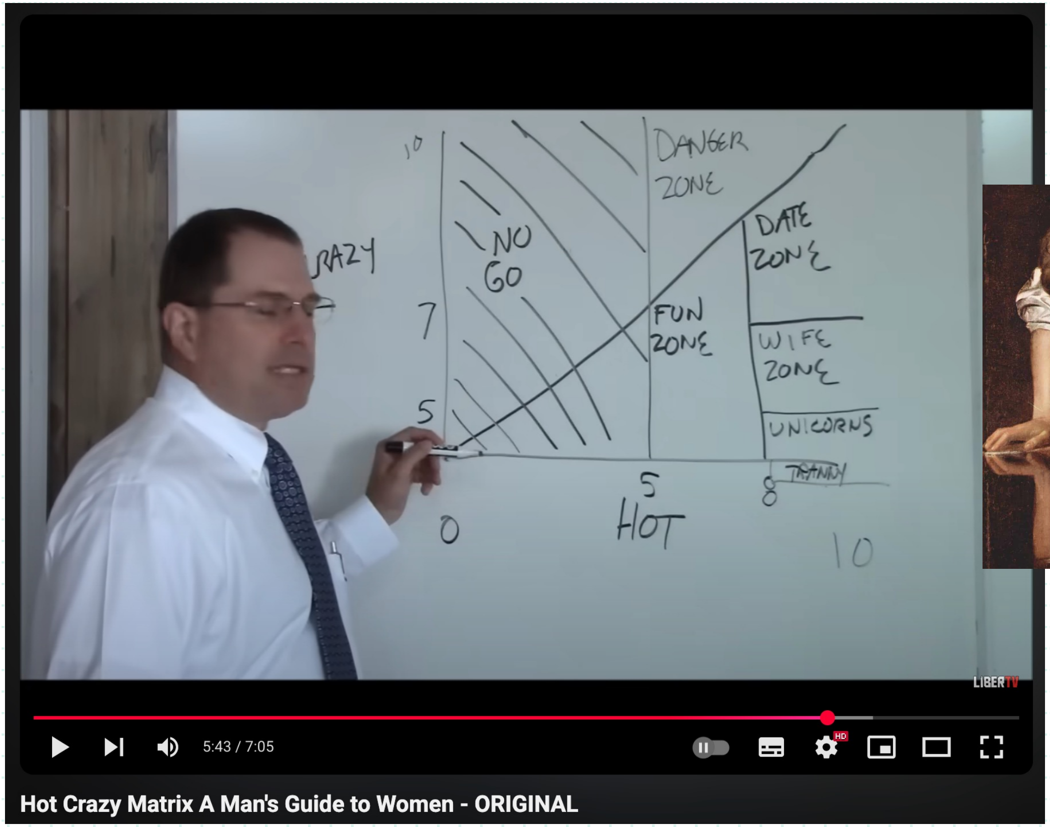
"collider bias"

Let's look at an example…

Research question:

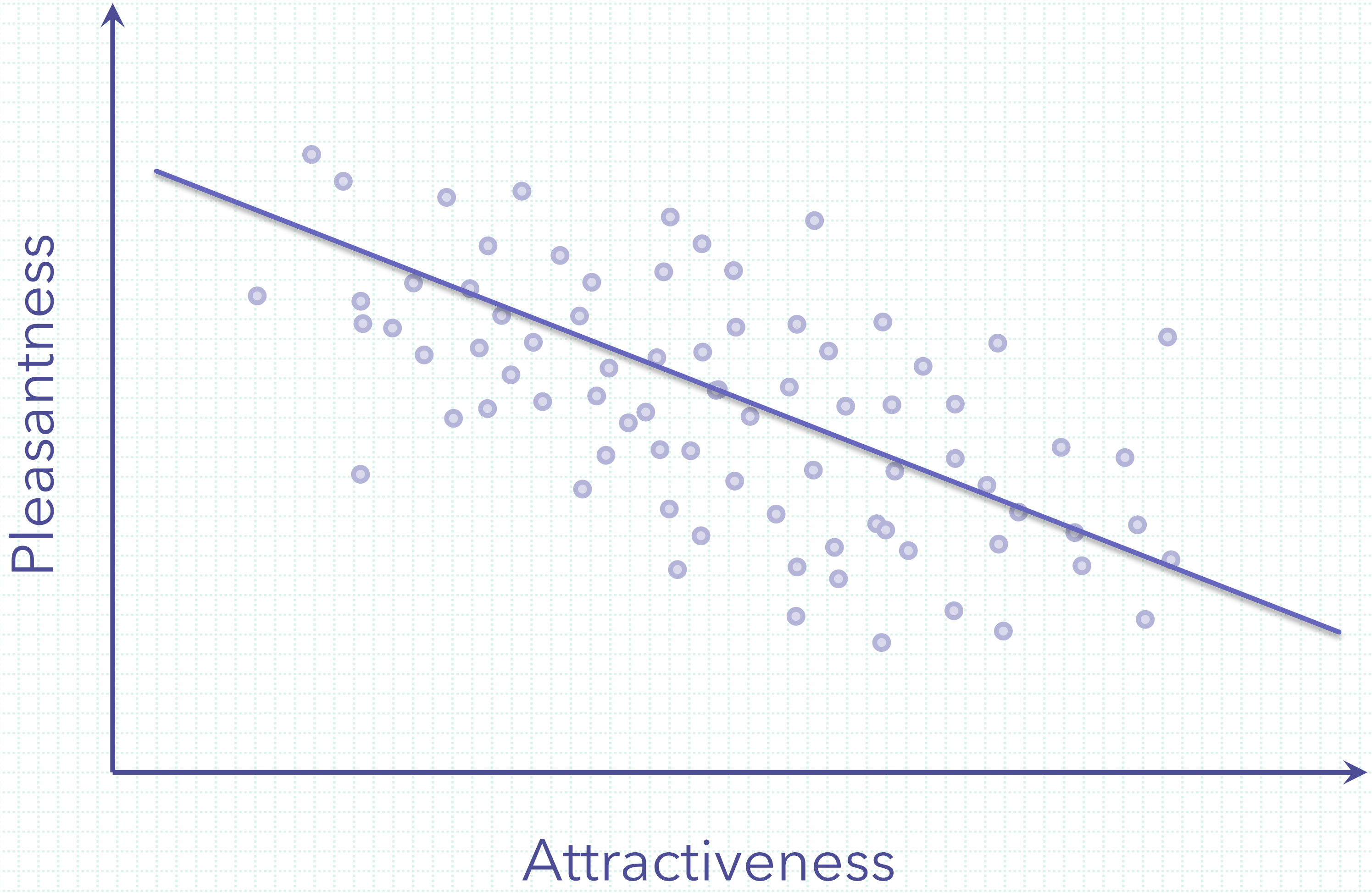
Are attractive people a***oles?

Background:

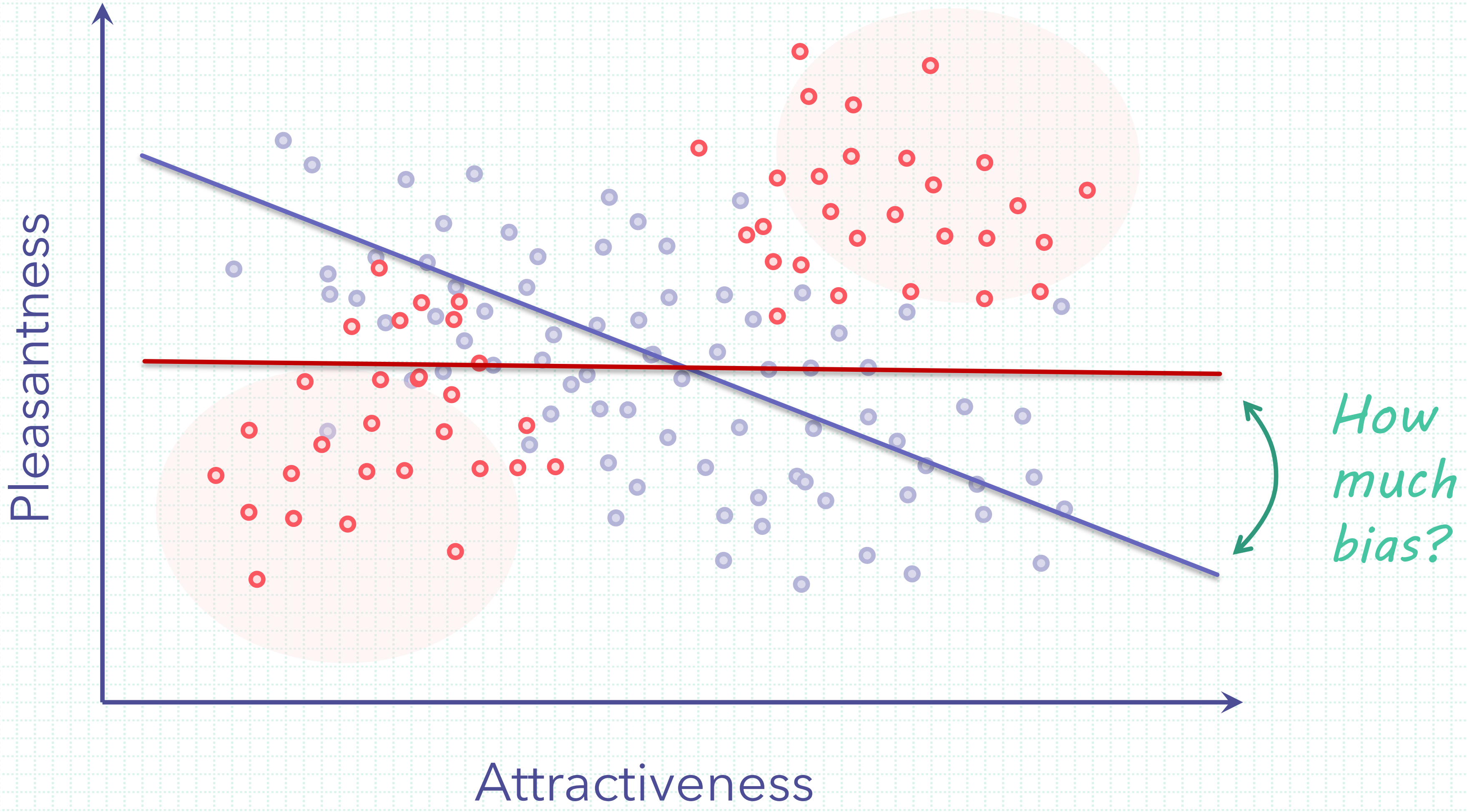


Hot Crazy Matrix A Man's Guide to Women - ORIGINAL

Experimental data



Experimental data



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

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*
… *but so is* NMAR

Under which missingness mechanisms does multiple imputation work?

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

2

3 (*better*) approaches to **addressing missing data**

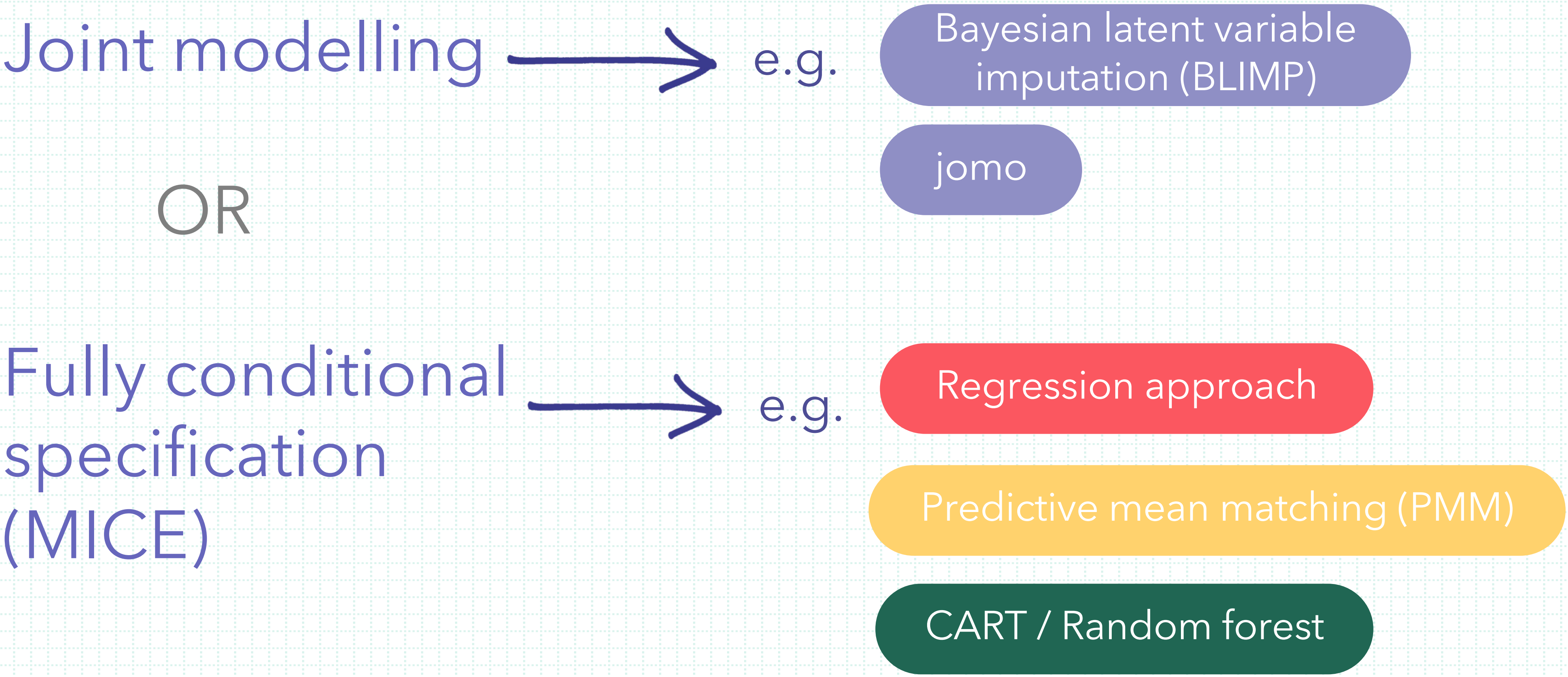
Multiple imputation (MI)

K but how?

Recipe

1. Create ***several*** ***plausible*** complete versions of the incomplete data sets.
2. **Analyse each** "complete" version of the data set.
3. **Pool** = incorporate results so that SE (and p-values) reflect uncertainty about the missing data.

1. Create *plausible* complete data sets



Joint modelling → e.g.

- Bayesian latent variable imputation (BLIMP)
- jomo

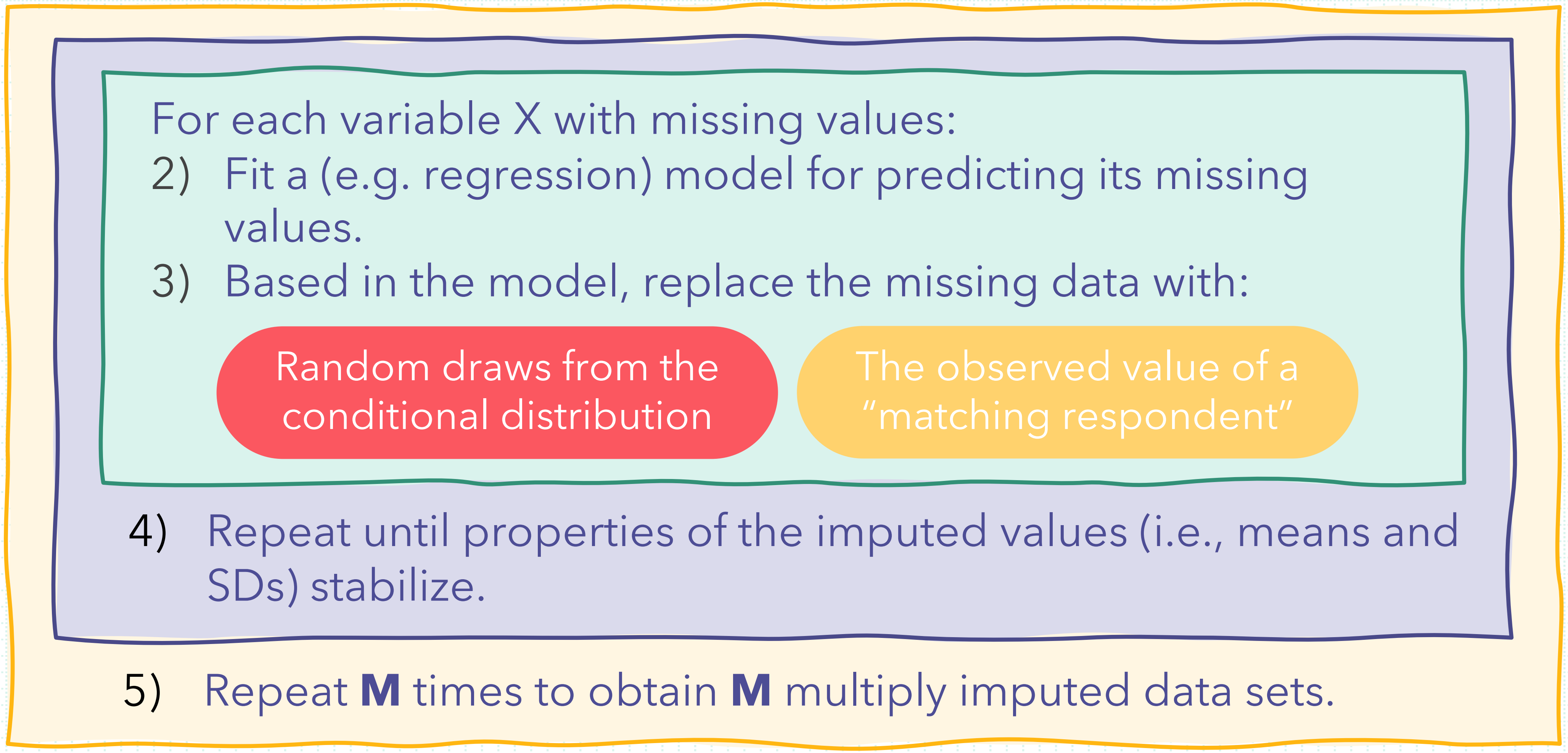
OR

Fully conditional specification (MICE) → e.g.

- Regression approach
- Predictive mean matching (PMM)
- CART / Random forest

Procedure

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



For each variable X with missing values:

2) Fit a (e.g. 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.

ID	Y	X1	X2
1	30	2	2
2	20	*NA*	7
3	*NA*	3	5
4	25	*NA*	*NA*
5	32	1	9
6	29	6	11

a) **Multiple imputation**

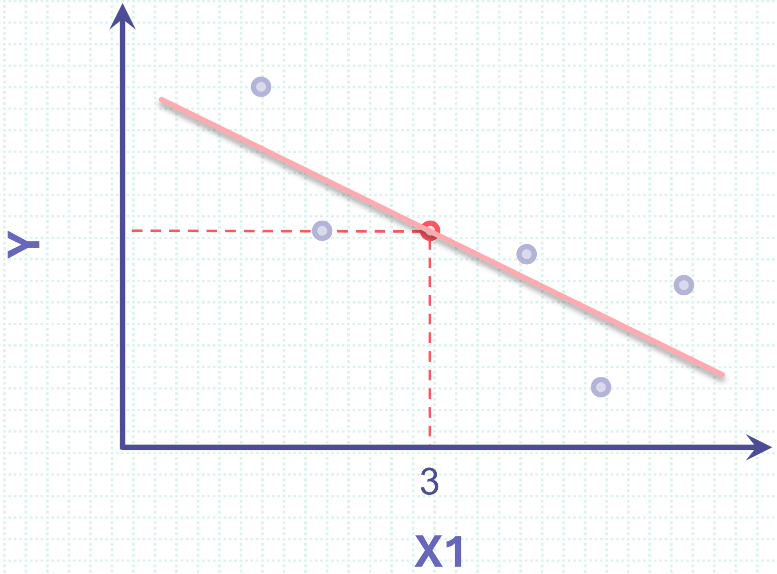
ID	Y	X1	X2
1	30	2	2
2	20	NA	7
3	NA	3	5
4	25	NA	NA
5	32	1	9
6	29	6	11

$$\mathbf{Y} = \alpha + \beta_1 * \mathbf{X1} + \beta_2 * \mathbf{X2}$$

a) **Multiple imputation**

ID	Y	X1	X2
1	30	2	2
2	20	*5*	7
3	*NA*	3	5
4	25	*4*	*10*
5	32	1	9
6	29	6	11

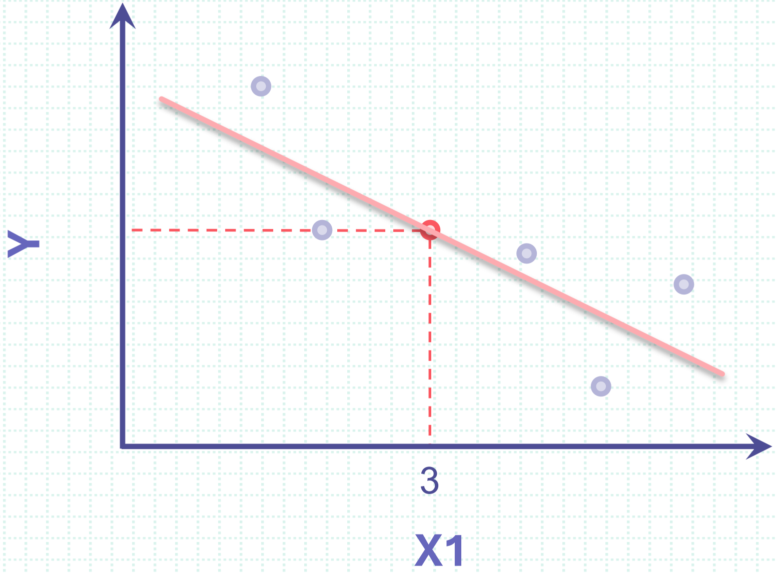
$$\mathbf{Y} = \alpha + \beta_1 * \mathbf{X1} + \beta_2 * \mathbf{X2}$$



a) **Multiple imputation**

ID	Y	X1	X2
1	30	2	2
2	20	*5*	7
3	*30*	3	5
4	25	*4*	*10*
5	32	1	9
6	29	6	11

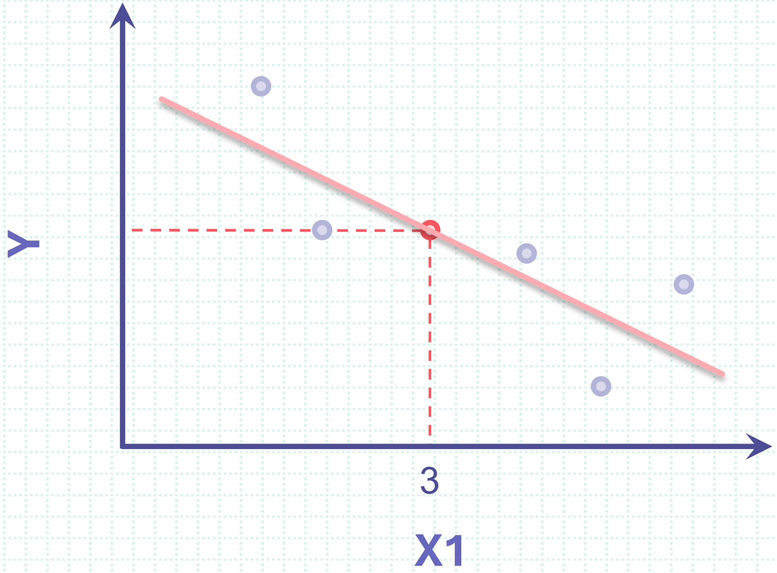
$$\mathbf{Y} = \alpha + \beta_1 * \mathbf{X1} + \beta_2 * \mathbf{X2}$$



a) **Multiple imputation**

ID	Y	X1	X2
1	30	2	2
2	20	*5*	7
3	*30*	3	5
4	25	*4*	*10*
5	32	1	9
6	29	6	11

$$Y = \alpha + \beta_1 * X1 + \beta_2 * X2$$



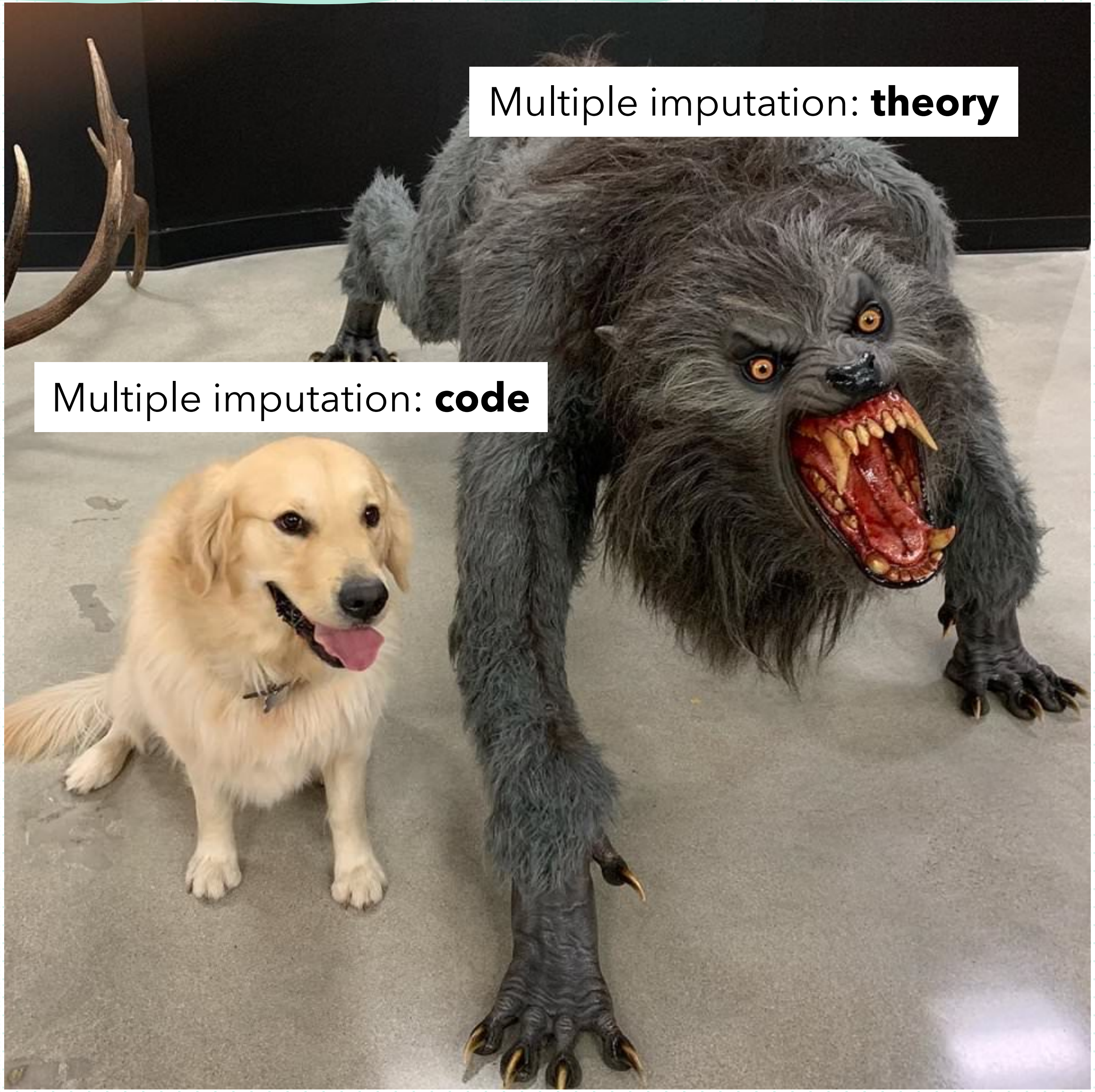
a) **Multiple imputation**

ID	Y	X1	X2
1	30	2	2
2	20	*NA*	7
3	*30*	3	5
4	25	*NA*	*10*
5	32	1	9
6	29	6	11

$$\mathbf{X1} = \alpha + \beta_1 * \mathbf{Y} + \beta_2 * \mathbf{X2}$$

… and so on

a) **Multiple imputation**


Multiple imputation: **theory**
Multiple imputation: **code**

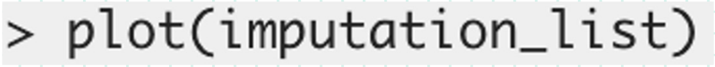
a) **Multiple imputation**

```r
meth <- make.method(data, defaultMethod = c("rf", "pmm", "polyreg"))

# Random forest imputation ran in parallel
imp_rf <- mice::futuremice(data,
                           method = meth,
                           m = 20,
                           maxit = 40,
                           ntree = 10,
                           rfPackage = "ranger",
                           n.core = 5,
                           n.imp.core = 4,
                           parallelseed = 3108,
                           print = TRUE)

# Or sequential: mice::mice()
```

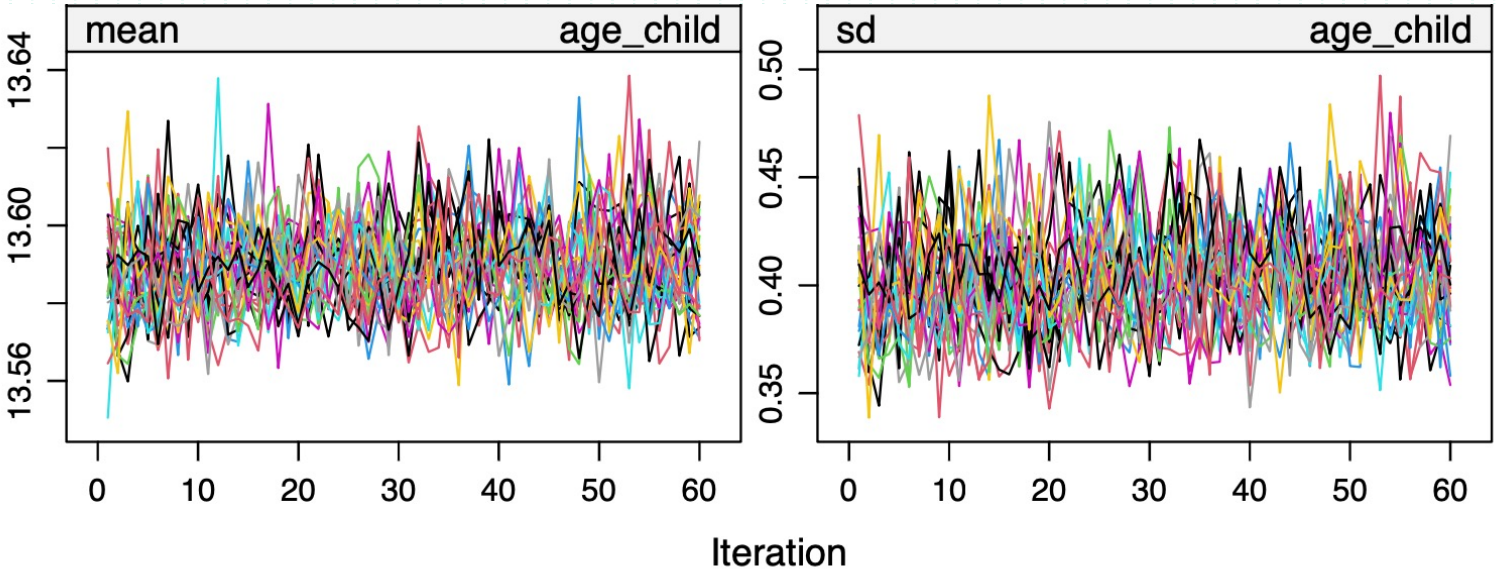
a) **Multiple imputation**

Assessing convergence *aka: how many iterations?*

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

Assessing convergence

aka: how many iterations?

Potential scale reduction factor (**PSR**) (Gelman and Rubin, 1992)
Measures the similarity of MCMC chains initiated from different random starting values.

$$\text{PSR} = \frac{\text{total variance}}{\text{average within−chain variance}}$$

```
mice::convergence(imp, diagnostic = "all", parameter = "mean")
```

As the *between-chain variation* (so: discrepancies across chains) ⇢ 0, the total variance ⇢ average within chain variance … PSR ⇢ 1

Cut-off: <1.05

3. Pool estimates: Rubin’s rules

Pooled **point estimate**: average of point estimates from each dataset

$$\bar{\theta} = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}_m$$

Where:

$\hat{\theta}_m$ is the parameter estimate from the mth imputed dataset

$\bar{\theta}$ is the pooled estimate across all M imputations.

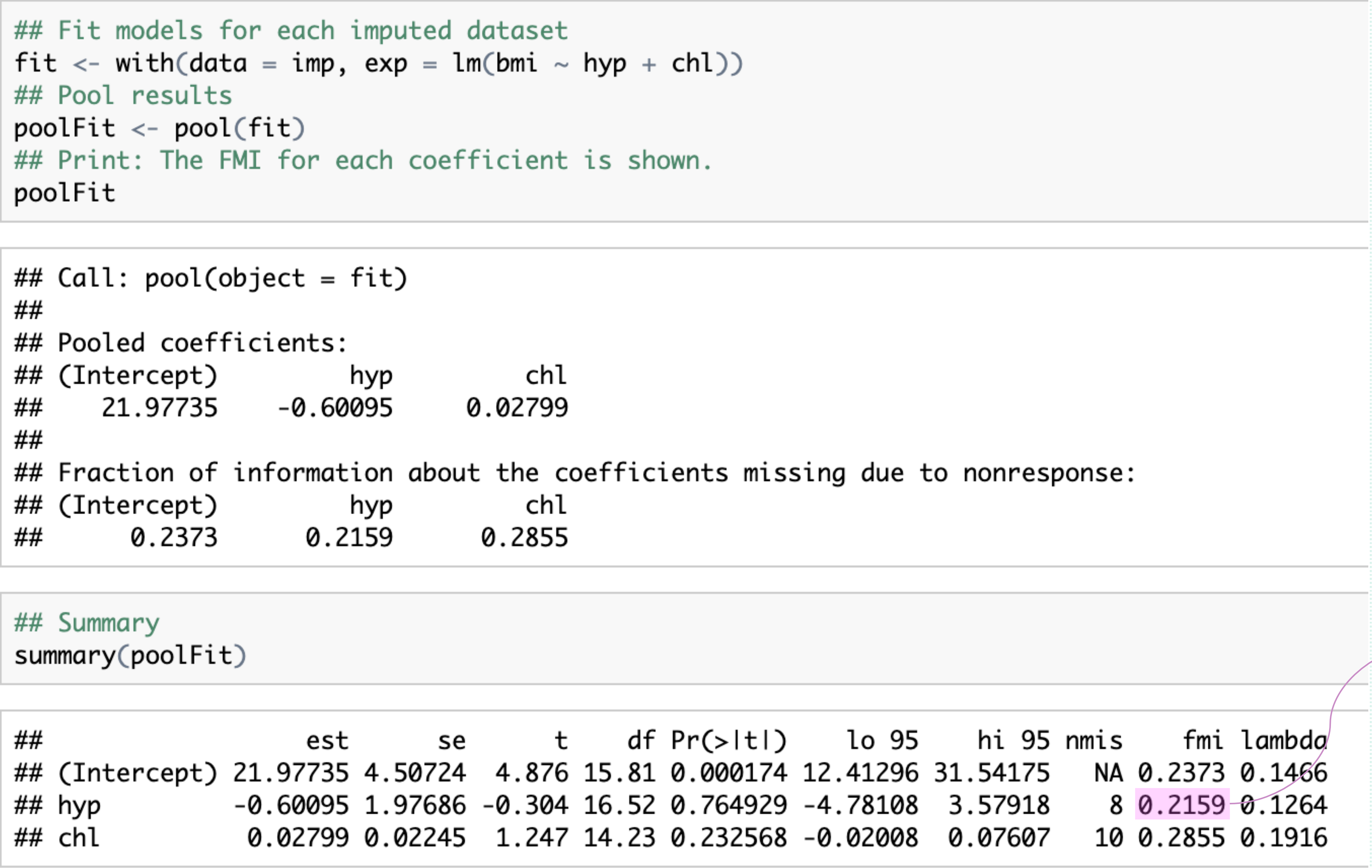
Pooled **standard error** estimate: incorporate corrections for the potential *inflation of the estimate due to sampling error.*

$$SE_{\bar{\theta}} = \sqrt{V_W + \left(1 + \frac{1}{M}\right) V_B}$$

Where:

$V_W = \frac{1}{M}\sum_{m=1}^{M} SE_m^2$ is the within-imputation variance (i.e. mean of the squared standard errors from each imputed dataset)

$V_B = \mathrm{var}(\hat{\theta}_m)$ is the between-imputation variance.



```
## Fit models for each imputed dataset
fit <- with(data = imp, exp = lm(bmi ~ hyp + chl))
## Pool results
poolFit <- pool(fit)
## Print: The FMI for each coefficient is shown.
poolFit
```

```
## Call: pool(object = fit)
##
## Pooled coefficients:
## (Intercept)         hyp         chl
##    21.97735    -0.60095     0.02799
##
## Fraction of information about the coefficients missing due to nonresponse:
## (Intercept)         hyp         chl
##      0.2373      0.2159      0.2855
```

```
## Summary
summary(poolFit)
```

```
##                  est      se      t    df Pr(>|t|)    lo 95    hi 95 nmis    fmi lambda
## (Intercept) 21.97735 4.50724  4.876 15.81 0.000174 12.41296 31.54175   NA 0.2373 0.1466
## hyp         -0.60095 1.97686 -0.304 16.52 0.764929 -4.78108  3.57918    8 0.2159 0.1264
## chl          0.02799 0.02245  1.247 14.23 0.232568 -0.02008  0.07607   10 0.2855 0.1916
```

OR
`mitml`

Fraction of missing information

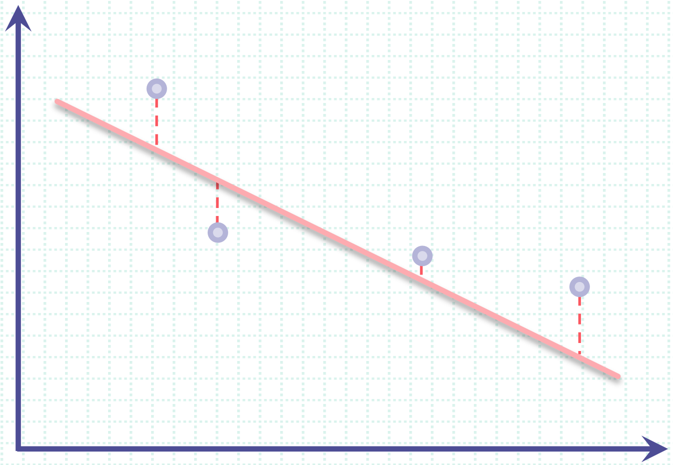
: variation due to missing data accounts for ~22% of the squared SE for the “hyp” predictor.

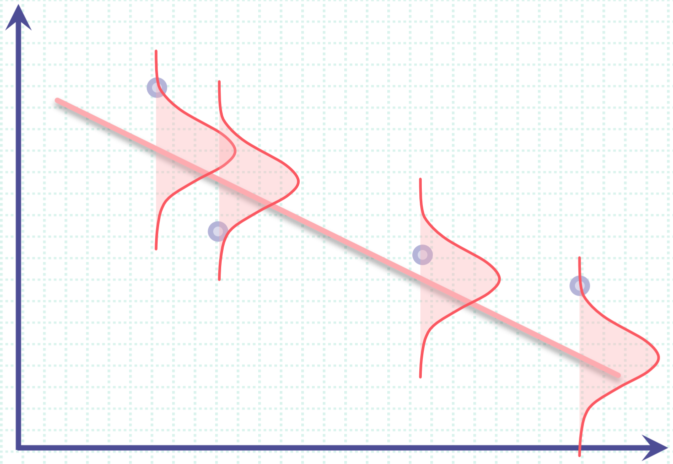
a) Multiple imputation

Full Information Maximum Likelihood (FIML)

Maximum Likelihood (ML) estimation

- ≠ least squares method, which stipulate that the estimates of the sample data (e.g. p or $\bar{x}$) should be close to the parameters of the model (π or μ).
- ML procedures take the opposite approach: find which values of the model parameters would make the data most likely.
  i.e., *maximize the probability of the* ***data*** by adjusting parameter estimates.



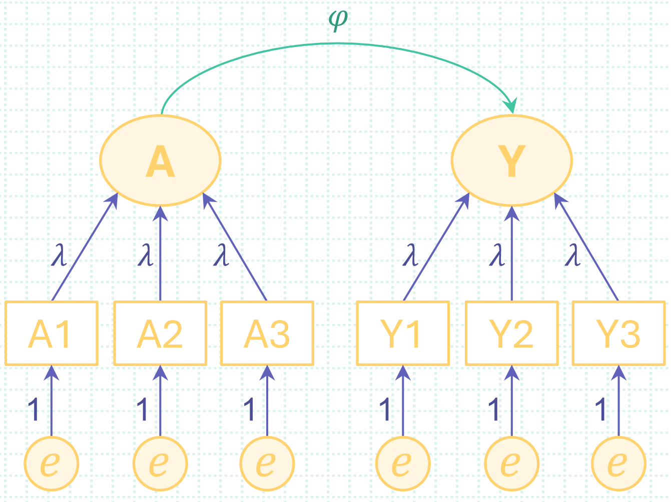


Observed correlation matrix

	A1	A2	A3	Y1	Y2	Y3
A1	1					
A2	0.4	1				
A3	0.3	0.2	1			
Y1	0.5	0.3	0.1	1		
Y2	0.3	0.4	0.2	0.6	1	
Y3	0.1	0.2	0.3	0.4	0.6	1

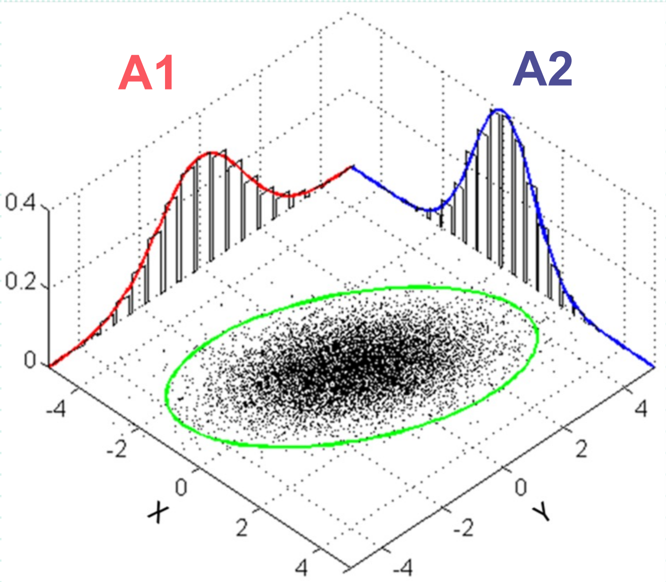
Model-implied correlation matrix

	A1	A2	A3	Y1	Y2	Y3
A1	$\mathrm{var}(e_{A1}) + \lambda_{A1}^2$					
A2	$\lambda_{A1}\lambda_{A2}$	$\mathrm{var}(e_{A2}) + \lambda_{A2}^2$				
A3	$\lambda_{A1}\lambda_{A3}$	$\lambda_{A2}\lambda_{A3}$	$\mathrm{var}(e_{A3}) + \lambda_{A3}^2$			
Y1	$\lambda_{A1}\lambda_{Y1}$	$\lambda_{A2}\lambda_{Y1}$	$\lambda_{A3}\lambda_{Y1}$	$\mathrm{var}(e_{Y1}) + \lambda_{Y1}^2$		
Y2	$\lambda_{A1}\lambda_{Y2}$	$\lambda_{A2}\lambda_{Y2}$	$\lambda_{A3}\lambda_{Y2}$	$\lambda_{Y1}\lambda_{Y2}$	$\mathrm{var}(e_{Y2}) + \lambda_{Y2}^2$	
Y3	$\lambda_{A1}\lambda_{Y3}$	$\lambda_{A2}\lambda_{Y3}$	$\lambda_{A3}\lambda_{Y3}$	$\lambda_{Y1}\lambda_{Y3}$	$\lambda_{Y2}\lambda_{Y3}$	$\mathrm{var}(e_{Y3}) + \lambda_{Y3}^2$



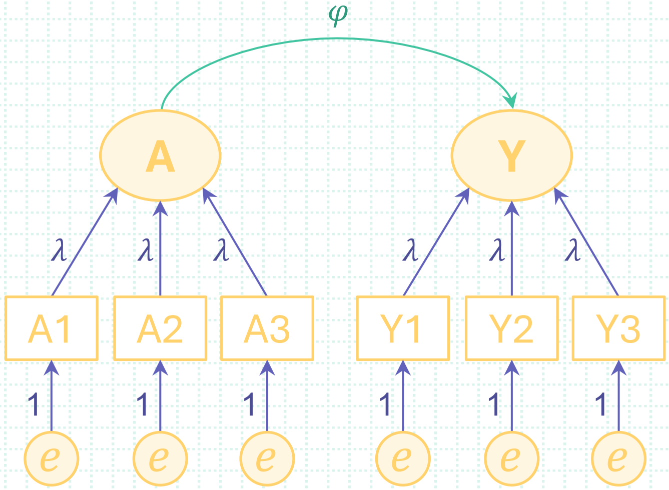
c) **Full information maximum likelihood (FIML)**

Observed correlation matrix



Model-implied correlation matrix

	A1	A2	A3	Y1	Y2	Y3
A1	var(e_{A1}) + λ_{A1}^2					
A2	$\lambda_{A1}\lambda_{A2}$	var(e_{A2}) + λ_{A2}^2				
A3	$\lambda_{A1}\lambda_{A3}$	$\lambda_{A2}\lambda_{A3}$	var(e_{A3}) + λ_{A3}^2			
Y1	$\lambda_{A1}\lambda_{Y1}$	$\lambda_{A2}\lambda_{Y1}$	$\lambda_{A3}\lambda_{Y1}$	var(e_{Y1}) + λ_{Y1}^2		
Y2	$\lambda_{A1}\lambda_{Y2}$	$\lambda_{A2}\lambda_{Y2}$	$\lambda_{A3}\lambda_{Y2}$	$\lambda_{Y1}\lambda_{Y2}$	var(e_{Y2}) + λ_{Y2}^2	
Y3	$\lambda_{A1}\lambda_{Y3}$	$\lambda_{A2}\lambda_{Y3}$	$\lambda_{A3}\lambda_{Y3}$	$\lambda_{Y1}\lambda_{Y3}$	$\lambda_{Y2}\lambda_{Y3}$	var(e_{Y3}) + λ_{Y3}^2



- **FIML** estimates the likelihood function *"row-by-row"*.
  Rows with missing data apply likelihood estimation *on the data that are available*.

- …assuming an appropriate distribution (e.g., *multivariate normal* for continuous data, or multinormal threshold models for binary or ordinal data).

c) **Full information maximum likelihood (FIML)**

“For missing Y’s, full information maximum likelihood (FIML) is preferred (Enders et al., 2020, von Hippel, 2007, Little, 1992) .”

FIML is the most straightforward (no preparation steps needed)

BUT:

- It is *not always available* in (open-source) packages for longitudinal data analysis

✗ `lme4`
✓ `lavaan`
✓ `OpenMx`
✓ `Mplus`

☝ (Like normal ML) this is based on multivariate normality ⇢ use robust standard errors

☝ It is based on the Y so exougenous variables are still deleted unless you set `fixed.x = FALSE` (in lavaan)

☝ Does not (natively) handle auxiliary variables (e.g. to change MNAR to MAR)

c) **Full information maximum likelihood (FIML)**

Propensity Scores (PS) Weighting

PS estimation can be used to *adjust for systematic differences between cases with complete and incomplete data*, including observational studies in which data is MNAR. ⇢ Not true

a) Estimate the **probability of missingness** *conditional on complete covariates*

b) Calculate *inverse probability* weights

$$\text{IPW} = {}^{1}/_{\text{probability of being a complete case}}$$

```
> twang::ps()
```

implements *gradient-boosted models* to balance the 2 groups on the set of covariates

c) ***Trim*** the weights (to reduce impact of outliers)

d) Assign "weights" to observations to *adjust their contribution: to* reflect their probability of being included in the analysis (Seaman and White, 2013).

2 critical decisions in PS estimation (applied to missing data)

1. *identify a* ***grouping variable*** [indicating whether cases are missing or complete]:
   - Typically, the *outcome* (e.g., missing neuroimaging data).
   - Note if you have multiple outcomes, PS estimates may *differ, e.g.* depending on imaging modality (e.g., functional vs. structural modalities).

2. *decide which* ***covariates*** will be used to balance groups.
   - Not trivial & likely to impact generalizability of study findings.
   - Common choice: *sociodemographic factors*, but ultimately depends on the research aims.
     - e.g., aim: comparing ADHD cases and controls ⇢ include e.g., medication history; comorbid anxiety…
   - Balancing covariates should be informed by both theory and missing data analysis, with caution: some covariates may have the potential to *introduce rather than reduce bias* (Seaman and White, 2013)

- If balancing covariates are also missing data, additional dummy variables indicating covariate missingness can be included as balancing covariates (D'Agostino and Rubin, 2000)

- Note: unlike other missing data approaches (e.g., FIML, MI) the analysis is only conducted on (re-weighted) observed cases. Thus, **the effective sample size does not differ** from a model that only includes complete cases.

Discussion point

is IPW ***ever*** preferable to MI or FIML?

3

We need to talk

…about a few more things

1… How much is too much missing data?

- There is no universal threshold. It depends *many* factors.

- **Be transparent** about the extent of missing data in the study:
  - Include an exploration of the missing data mechanism
  - Explain the approaches used for handling missing data
  - Analyse the potential impact of different approaches on the results.

- "Researchers tend to be squeamish about analyses in which there are large amounts of missing data (≥50%; Sterne et al., 2009), but it is precisely in these instances when researchers *should* be dealing with it, and it is not inherently inappropriate to tackle a data set with large amounts of missing data."

- See Sterne et al. (2009) for reporting guidelines.

And… too little missing data?

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.

Missing data in the **exposure / predictor / X / independent variable**

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

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).

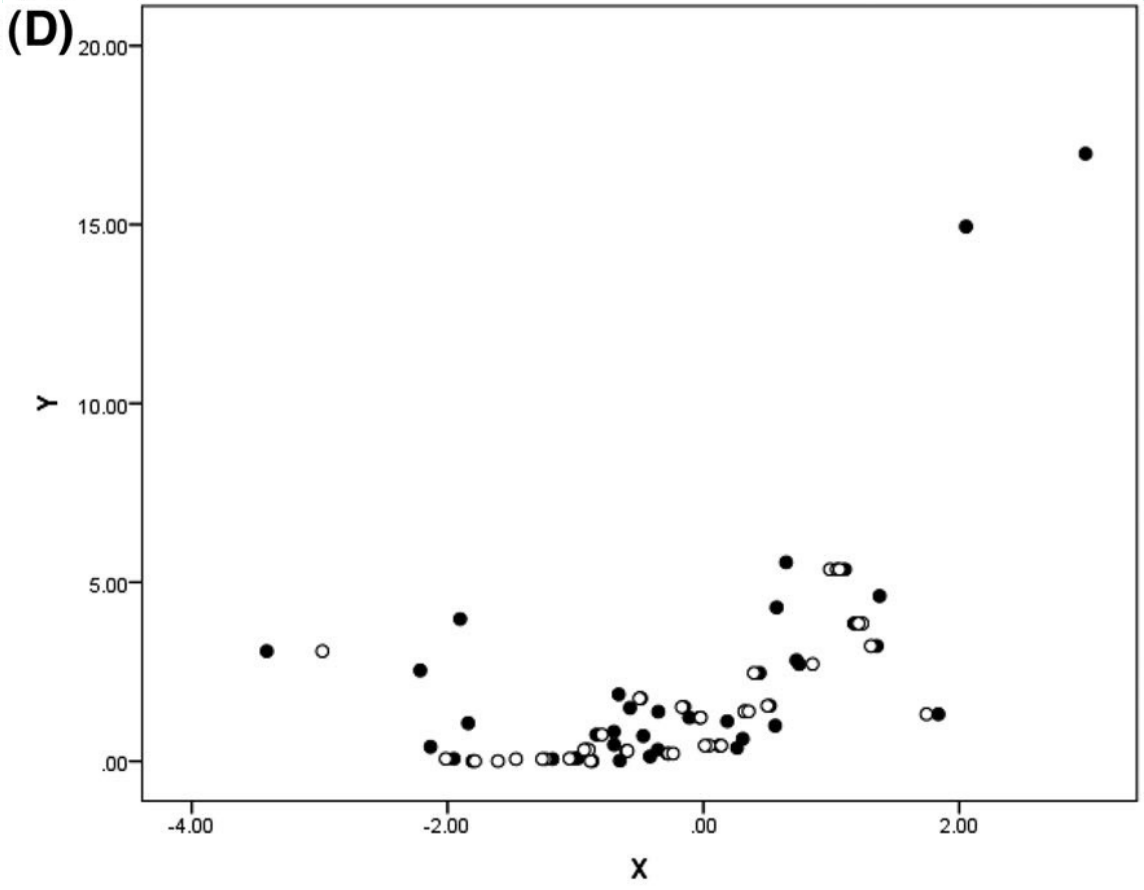
Missing data in the **outcome** / **dependent variable** / **Y**

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



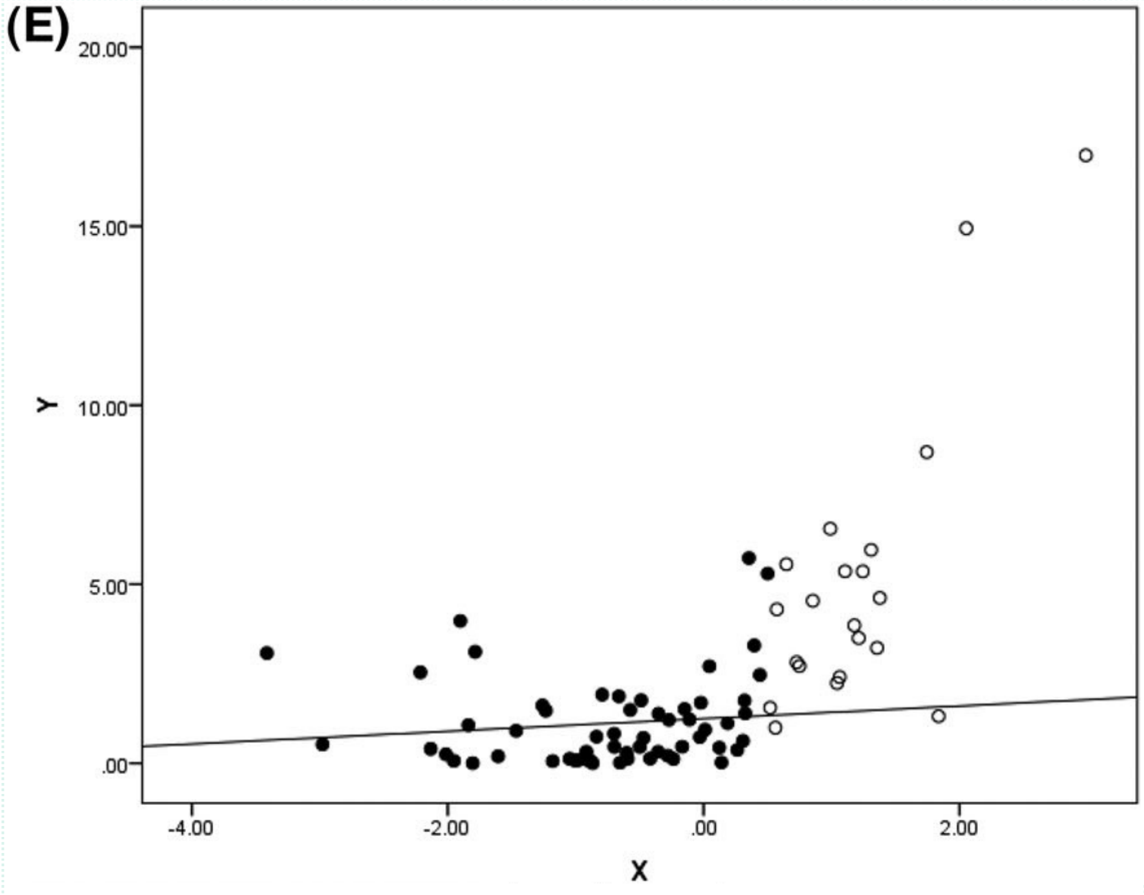
1. We simulate some bivariate data where X and Y are quadratically related.
2. We remove 40% of the data according to MCAR.
3. We *incorrectly* assume that X and Y are linearly related (and use a linear regression model for both multiple imputation and the analysis).

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.

3 points to enphasize

1. Missing data approaches (e.g., MI) should not be viewed as "*making up data*".

2. There is no one statistical procedure for all situations.

3. Missing data patterns by *sociodemographic groups* can be sourced, in part, to mistrust of historic and contemporary scientific practice.

Did I *miss* something ?