EP16: Missing Values in Clinical Research: Multiple Imputation

9. Imputation in Complex Settings

Nicole Erler

Department of Biostatistics, Erasmus Medical Center

n.erler@erasmusmc.nl

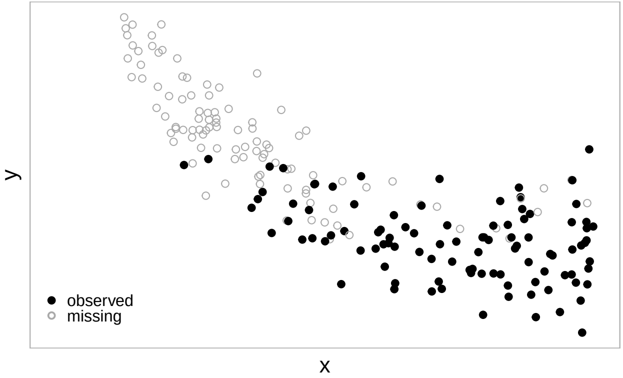
Erasmus MC
University Medical Center Rotterdam

Quadratic Effect

Consider the case where the **analysis model** (which we assume to be true) is

$$y = \beta_0 + \beta_1 x + \boldsymbol{\beta_2 x^2} + \ldots,$$

i.e., y has a **quadratic relationship** with x, and x is incomplete.



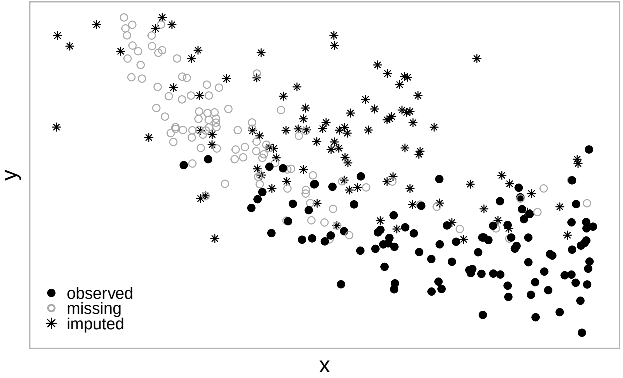
The original data show a curved pattern.

Quadratic Effect

The model used to **impute** x when using MICE (naively) is

$$x = \theta_{10} + \theta_{11}y + \dots,$$

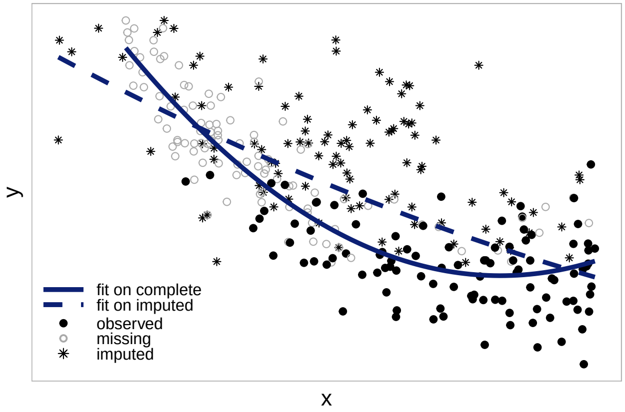
i.e., a **linear relation** between x and y is assumed.



The imputed values **distort the curved pattern** of the original data.

Quadratic Effect

The model fitted on the imputed data gives **severely biased results**; the non-linear shape of the curve has almost completely disappeared.

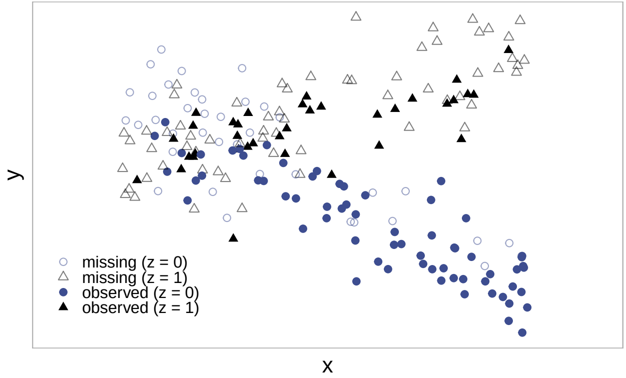


Interaction Effect

Another example: consider the analysis model (again, assumed to be true)

$$y = \beta_0 + \beta_x x + \beta_z z + \boldsymbol{\beta_{xz} xz} + \ldots,$$

i.e., y has a **non-linear relationship** with x due to the **interaction term**.



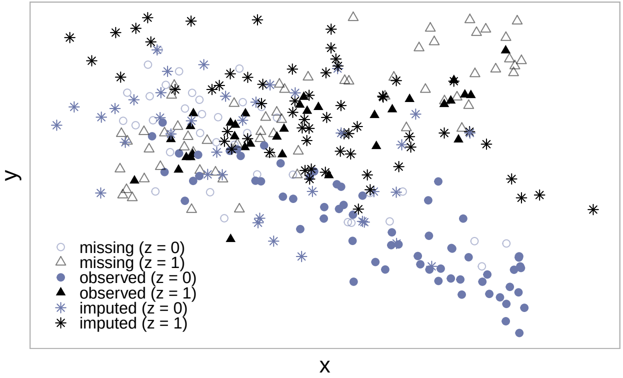
The original data shows a “<” shaped pattern.

Interaction Effect

The model used to impute x when using MICE (naively) is

$$x = \theta_{10} + \theta_{11}y + \theta_{12}z + \ldots,$$

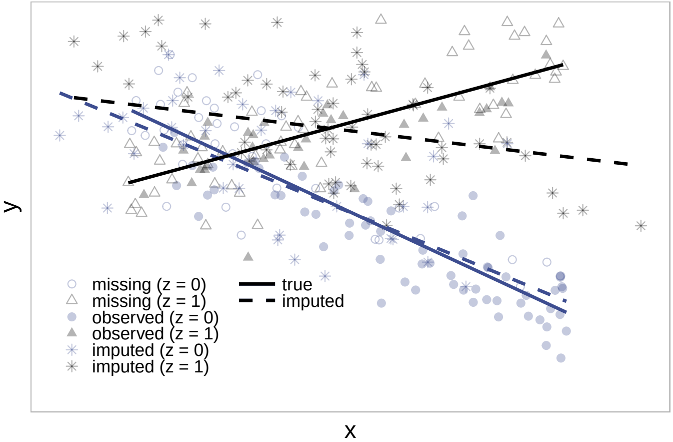
i.e., a linear relation between x and y is assumed.



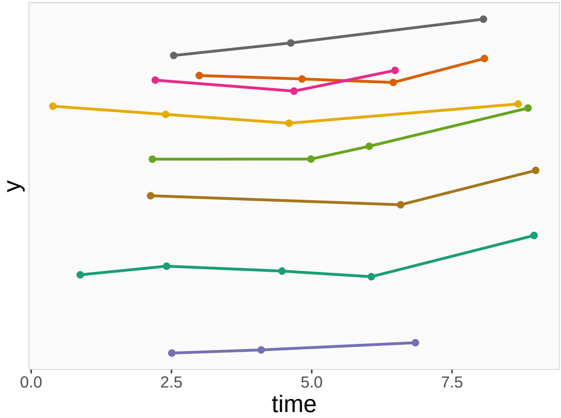
The "$<$" shaped pattern of the true data is **distorted by the imputed values**.

Interaction Effect

And the analysis on these naively imputed values leads to **severely biased estimates**.



Longitudinal Outcome



ID	y	x_1	x_2	x_3	x_4	time
1	✓	✓	NA	✓	✓	0.87
1	✓	✓	NA	✓	✓	2.41
1	✓	✓	NA	✓	✓	4.47
1	✓	✓	NA	✓	✓	6.06
1	✓	✓	NA	✓	✓	8.96
2	✓	✓	✓	NA	NA	3.00
2	✓	✓	✓	NA	NA	4.83
2	✓	✓	✓	NA	NA	6.45
2	✓	✓	✓	NA	NA	8.08
3	✓	✓	✓	NA	NA	2.51
3	✓	✓	✓	NA	NA	4.10
3	✓	✓	✓	NA	NA	6.85
4	✓	✓	NA	✓	✓	2.21
4	✓	✓	NA	✓	✓	4.68
4	✓	✓	NA	✓	✓	6.48
⋮	⋮	⋮	⋮	⋮	⋮	⋮

($x_1, \ldots, x_4$ are baseline covariates, i.e., not measured repeatedly, e.g. age at baseline, gender, education level, ...)

Longitudinal Outcome

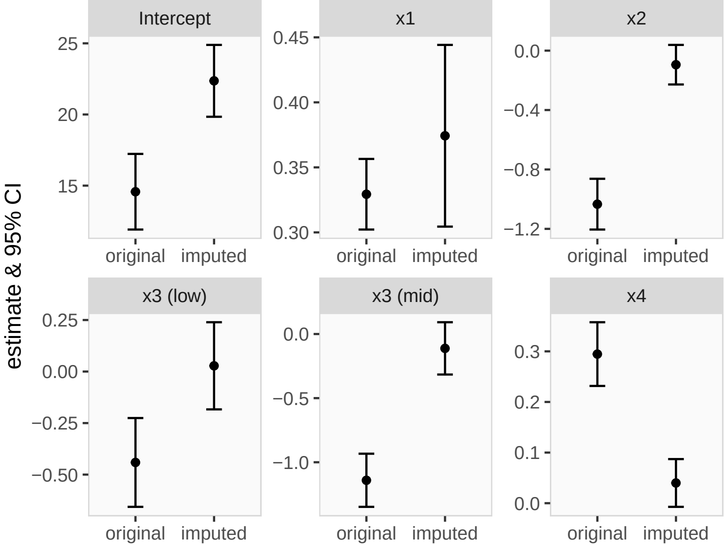
For data in long format:

- each row would be regarded as independent
- ➡ bias and **inconsistent imputations**

Imputed values of baseline covariates are imputed with different values, creating data that could not have been observed.

ID	y	x_1	x_2	x_3	x_4	time
1	✓	✓	girl	✓	✓	0.87
1	✓	✓	boy	✓	✓	2.41
1	✓	✓	girl	✓	✓	4.47
1	✓	✓	girl	✓	✓	6.06
1	✓	✓	girl	✓	✓	8.96
2	✓	✓	✓	mid	38.8	3.00
2	✓	✓	✓	high	39.9	4.83
2	✓	✓	✓	mid	40.1	6.45
2	✓	✓	✓	low	39.7	8.08
3	✓	✓	✓	high	40.7	2.51
3	✓	✓	✓	low	40.4	4.10
3	✓	✓	✓	mid	39.7	6.85
4	✓	✓	boy	✓	✓	2.21
4	✓	✓	boy	✓	✓	4.68
4	✓	✓	girl	✓	✓	6.48
⋮	⋮	⋮	⋮	⋮	⋮	⋮

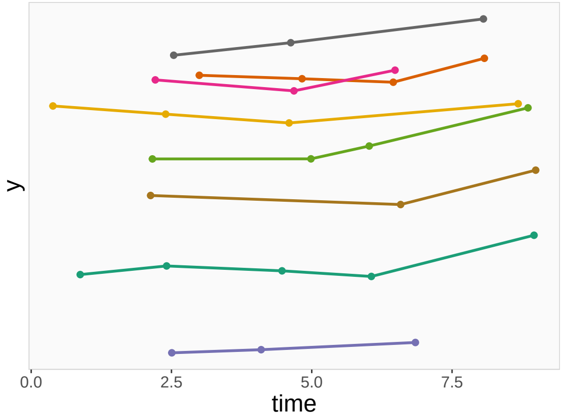
Longitudinal Outcome



Estimates can be severely biased.

Longitudinal Outcome

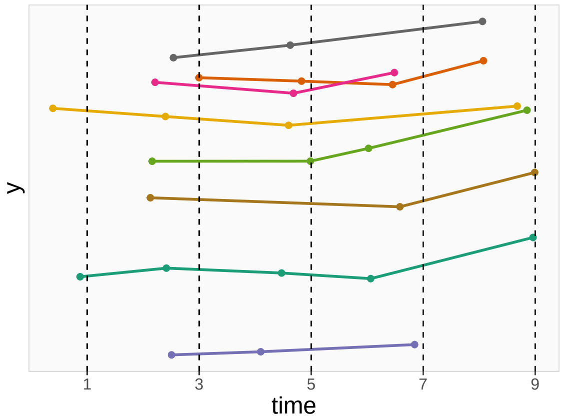
In some settings **imputation in wide format** may be possible.



ID	y	x_1	x_2	x_3	x_4	time
1	✓	✓	NA	✓	✓	0.87
1	✓	✓	NA	✓	✓	2.41
1	✓	✓	NA	✓	✓	4.47
1	✓	✓	NA	✓	✓	6.06
1	✓	✓	NA	✓	✓	8.96
2	✓	✓	✓	NA	NA	3.00
2	✓	✓	✓	NA	NA	4.83
2	✓	✓	✓	NA	NA	6.45
2	✓	✓	✓	NA	NA	8.08
3	✓	✓	✓	NA	NA	2.51
3	✓	✓	✓	NA	NA	4.10
3	✓	✓	✓	NA	NA	6.85
4	✓	✓	NA	✓	✓	2.21
4	✓	✓	NA	✓	✓	4.68
4	✓	✓	NA	✓	✓	6.48
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Longitudinal Outcome

In some settings **imputation in wide format** may be possible.



ID	y	x_1	x_2	x_3	x_4	time
1	✓	✓	NA	✓	✓	0.87
1	✓	✓	NA	✓	✓	2.41
1	✓	✓	NA	✓	✓	4.47
1	✓	✓	NA	✓	✓	6.06
1	✓	✓	NA	✓	✓	8.96
2	✓	✓	✓	NA	NA	3.00
2	✓	✓	✓	NA	NA	4.83
2	✓	✓	✓	NA	NA	6.45
2	✓	✓	✓	NA	NA	8.08
3	✓	✓	✓	NA	NA	2.51
3	✓	✓	✓	NA	NA	4.10
3	✓	✓	✓	NA	NA	6.85
4	✓	✓	NA	✓	✓	2.21
4	✓	✓	NA	✓	✓	4.68
4	✓	✓	NA	✓	✓	6.48
⋮	⋮	⋮	⋮	⋮	⋮	⋮

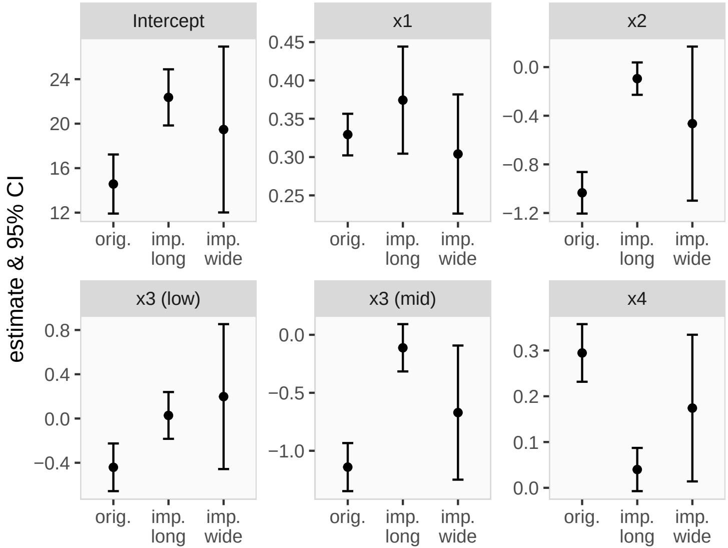
Longitudinal Outcome

id	y.1	time.1	y.3	time.3	y.5	time.5	y.7	time.7	y.9	time.9	...
1	31.6	0.9	31.8	2.4	31.7	4.5	31.5	6.1	32.5	9	...
2	NA	NA	36.2	3	36.1	4.8	36.1	6.5	36.6	8.1	...
3	NA	NA	29.8	2.5	29.8	4.1	30	6.8	NA	NA	...
4	NA	NA	36.1	2.2	35.9	4.7	36.3	6.5	NA	NA	...
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

In **wide format:**

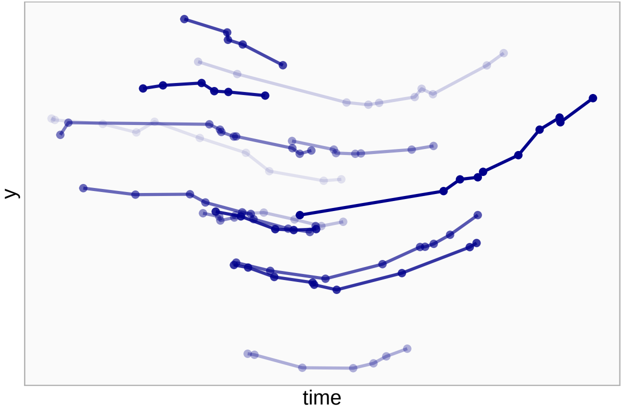
- missing values in outcome and measurement times need to be imputed (to be able to use them as predictors to impute covariates)
- **inefficient!** (we would not need to impute them for the analysis)

Longitudinal Outcome



Better, but large confidence intervals.

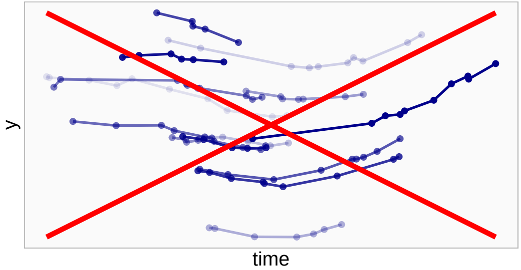
Longitudinal Outcome



Very **unbalanced** data: transformation to wide format not possible.

(Or requires summary measures)

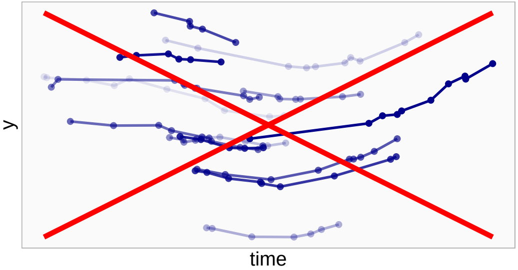
Longitudinal Outcome

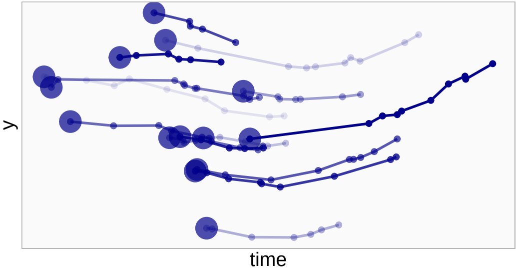


Naive approaches that are sometimes used are to

- **ignore the outcome** in the imputation

Longitudinal Outcome

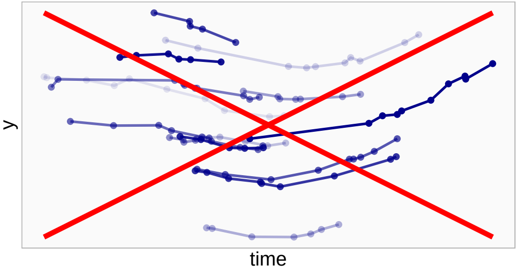


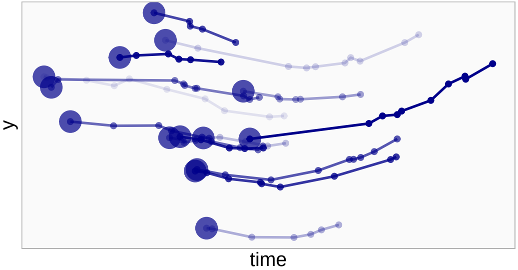


Naive approaches that are sometimes used are to

- **ignore the outcome** in the imputation, or to
- use only the **first/baseline outcome**

Longitudinal Outcome





Naive approaches that are sometimes used are to

- **ignore the outcome** in the imputation, or to
- use only the **first/baseline outcome**

➡ Important information may be lost!

➡ invalid imputations and biased results

Survival Data

Cox proportional hazards model

$$h(t) = h_0(t) \exp(x\beta_x + z\beta_z),$$

- $h(t)$: **hazard** = the instantaneous risk of an event at time t, given that the event has not occurred until time t
- $h_0(t)$: unspecified **baseline hazard**
- x and z: **incomplete** and **complete covariates**, respectively

Survival Data

Cox proportional hazards model

$$h(t) = h_0(t)\exp(x\beta_x + z\beta_z),$$

- $h(t)$: **hazard** = the instantaneous risk of an event at time t, given that the event has not occurred until time t
- $h_0(t)$: unspecified **baseline hazard**
- x and z: **incomplete** and **complete covariates**, respectively

Survival outcome representation:

- **observed event time** T
- **event indicator** D ($D = 1$: event, $D = 0$: censored).

Survival Data

Naive use of MICE

- T and D are treated just like any other variable.
- The resulting imputation model for X would have the form

$$p(x \mid T, D, \mathbf{z}) = \theta_0 + \theta_1 T + \theta_2 D + \theta_3 z + \ldots .$$

Survival Data

Naive use of MICE

- T and D are treated just like any other variable.
- The resulting imputation model for X would have the form

$$p(x \mid T, D, \mathbf{z}) = \theta_0 + \theta_1 T + \theta_2 D + \theta_3 z + \dots .$$

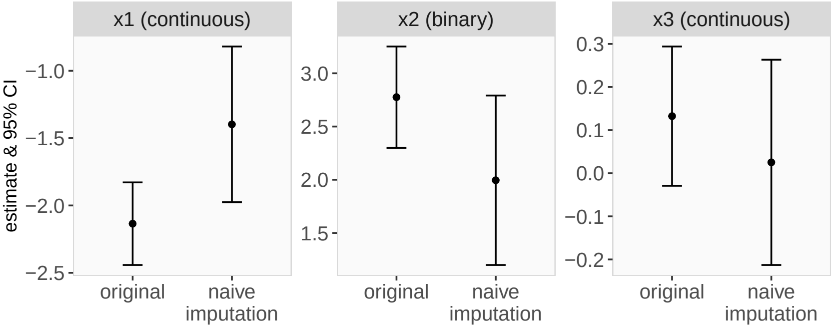
The **correct conditional distribution** of x given the other variables is, however,

$$\log p(x \mid T, D, z) = \log p(x \mid z) + D(\beta_x x + \beta_z z) - H_0(T) \exp(\beta_x x + \beta_z z) + const.,$$

where $H_0(T)$ is the cumulative baseline hazard. (White & Royston, 2009)

Survival Data

Using the naively assumed imputation model can lead to **severe bias**:



(Results from MICE imputation with two incomplete normal and one incomplete binary covariate.)

References

White, I. R., & Royston, P. (2009). Imputing missing covariate values for the cox model. *Statistics in Medicine*, *28*(15), 1982–1998.