EP16: Missing Values in Clinical Research: Multiple Imputation

9. Imputation in Complex Settings

Nicole S. Erler

Department of Biostatistics, Erasmus Medical Center

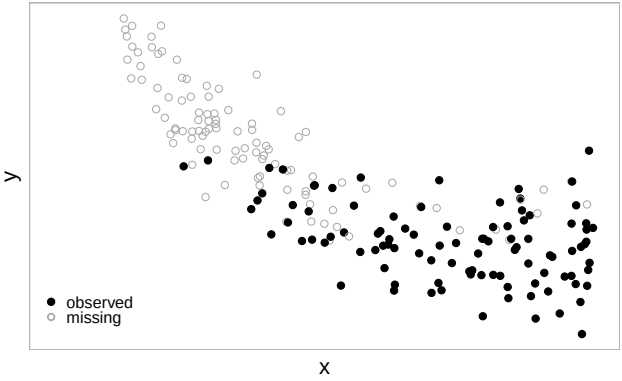
n.erler@erasmusmc.nl

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.



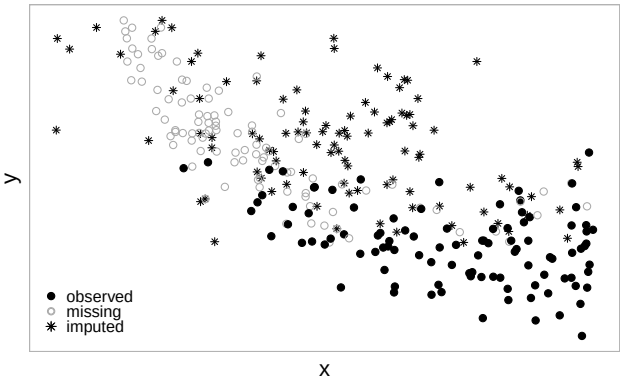
The original data show a curved pattern.

Quadratic Effect

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

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

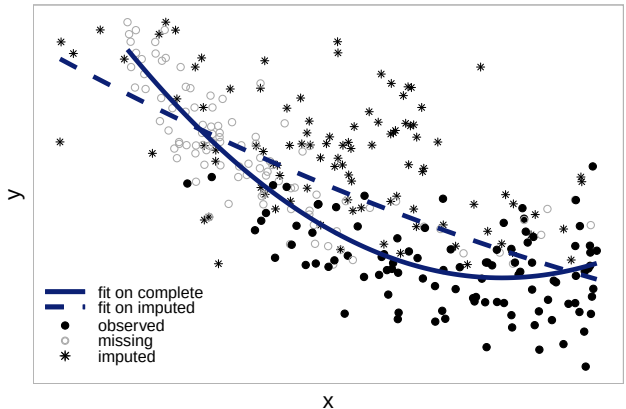
i.e., a **linear relation** between x and y is assumed.



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.

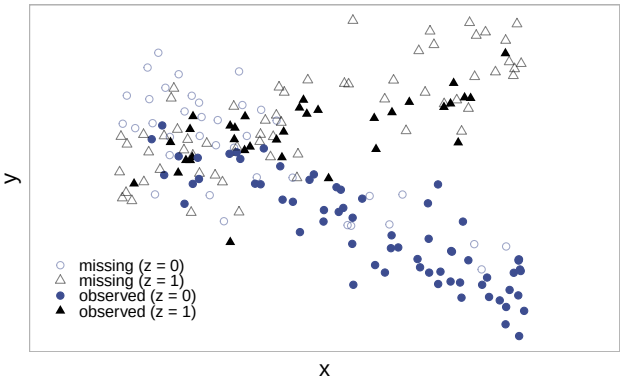


Interaction effect

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**.



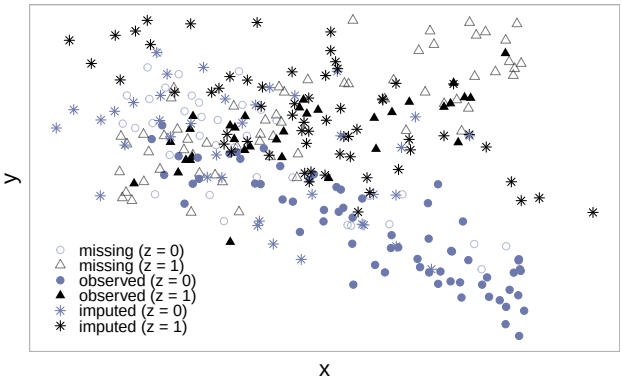
The original data shows a "<" shaped pattern.

Interaction effect

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

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

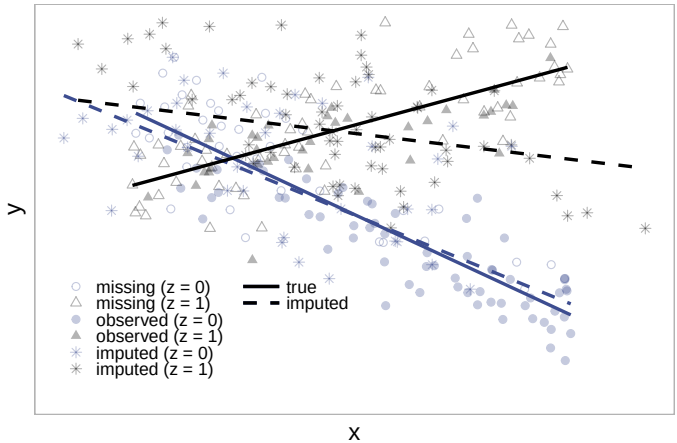
i.e., a linear relation between x and y is assumed.



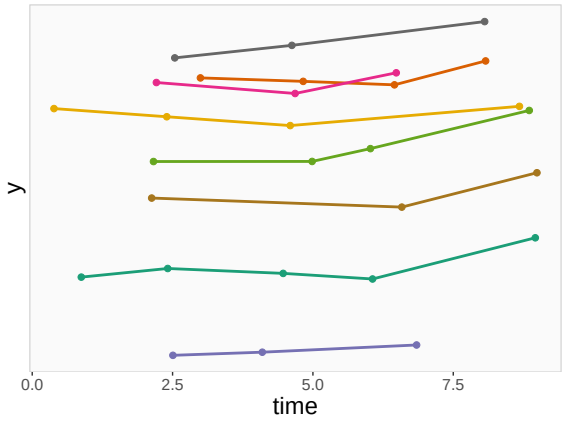
The "<" shaped pattern of the true data is **distorted by the imputed values**.

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
⋮	⋮	⋮	⋮	⋮	⋮	⋮

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

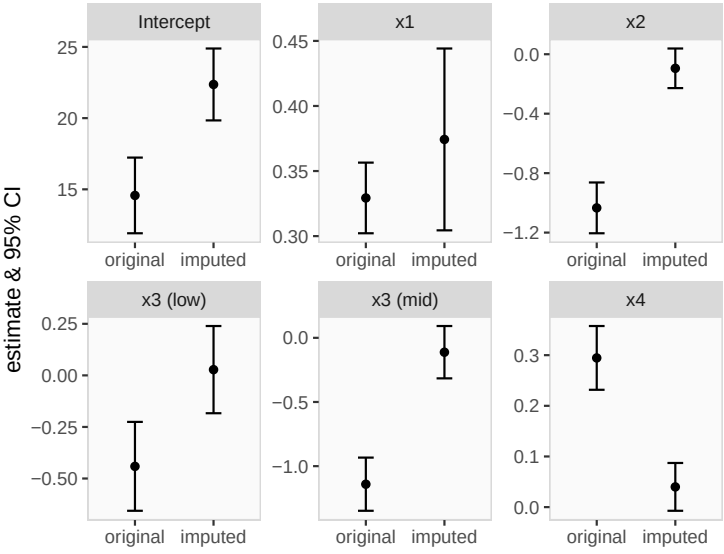
Longitudinal outcome

If we use MICE in the data in this (long) format, each row would be regarded as independent, which may cause 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
⋮	⋮	⋮	⋮	⋮	⋮	⋮

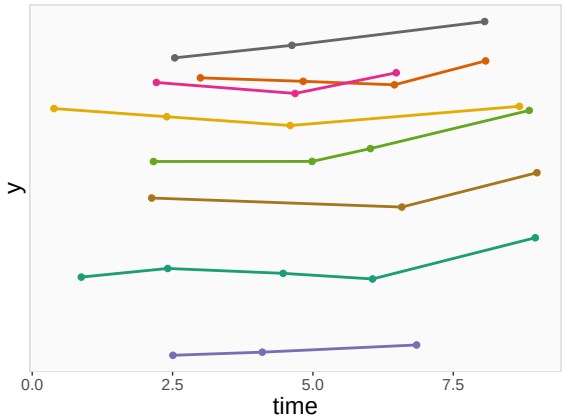
Longitudinal outcome



Estimates can be severely biased.

Longitudinal outcome

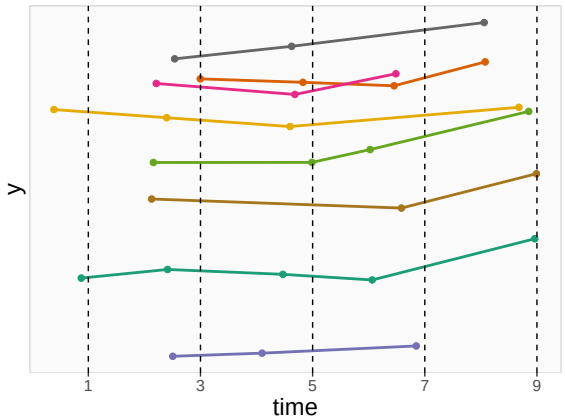
In some settings **imputation in wide format** may be possible.



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.



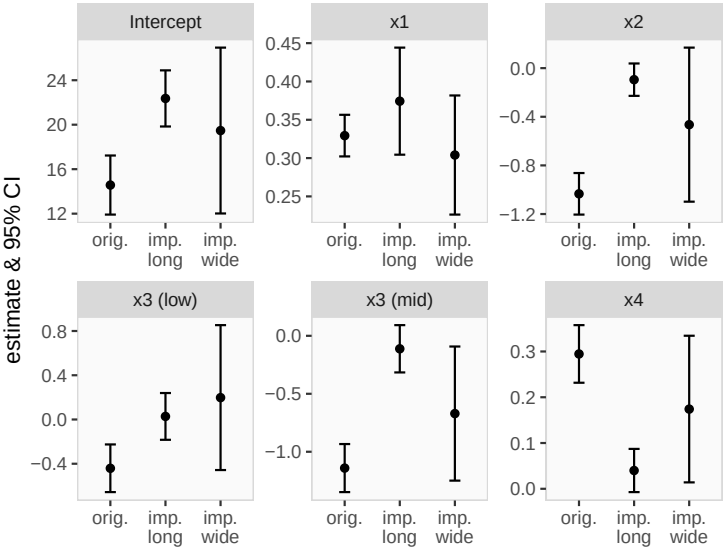
ID	y	x_1	x_2	x_3	x_4	time
1	✓	✓	NA	✓	✓	0.87
1	✓	✓	NA	✓	✓	2.41
1	✓	✓	NA	✓	✓	4.47
1	✓	✓	NA	✓	✓	6.06
1	✓	✓	NA	✓	✓	8.96
2	✓	✓	✓	NA	NA	3.00
2	✓	✓	✓	NA	NA	4.83
2	✓	✓	✓	NA	NA	6.45
2	✓	✓	✓	NA	NA	8.08
3	✓	✓	✓	NA	NA	2.51
3	✓	✓	✓	NA	NA	4.10
3	✓	✓	✓	NA	NA	6.85
4	✓	✓	NA	✓	✓	2.21
4	✓	✓	NA	✓	✓	4.68
4	✓	✓	NA	✓	✓	6.48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

Longitudinal outcome

id	y.1	time.1	y.3	time.3	y.5	time.5	y.7	time.7	y.9	time.9
1	31.59	0.87	31.79	2.41	31.67	4.47	31.54	6.06	32.5	8.96
2	NA	NA	36.23	3	36.15	4.83	36.07	6.45	36.63	8.08
3	NA	NA	29.76	2.51	29.84	4.1	30.01	6.85	NA	NA
4	NA	NA	36.12	2.21	35.87	4.68	36.35	6.48	NA	NA
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

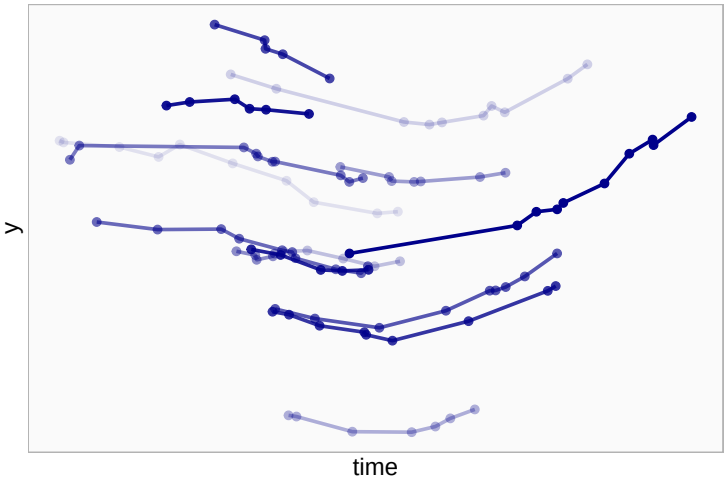
In this **wide format data** frame, missing values in the outcome and measurement times need to be imputed (to be able to use them as predictors to impute covariates), even though we would not need to impute them for the analysis (mixed model is valid when outcome measurements are M(C)AR).

Longitudinal outcome



Better, but large confidence intervals.

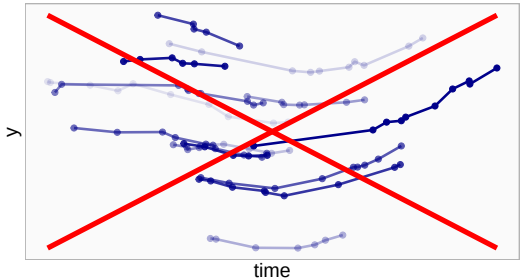
Longitudinal outcome



When the data is very **unbalanced**, transformation to wide format is not possible.

(Or at least transformation to wide format leads to variables with high proportions of missing values.)

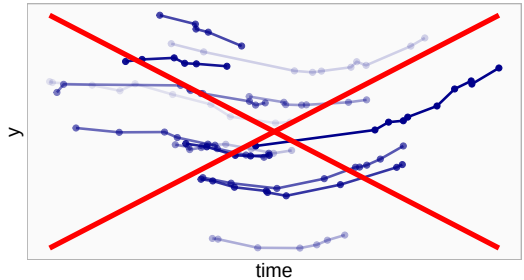
Longitudinal outcome

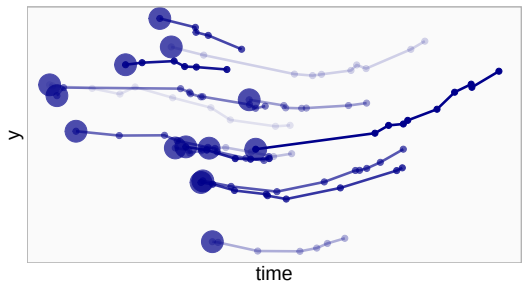


Naive approaches that are sometimes used are to

- **ignore the outcome** in the imputation

Longitudinal outcome

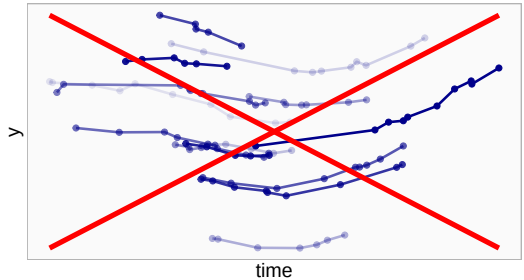


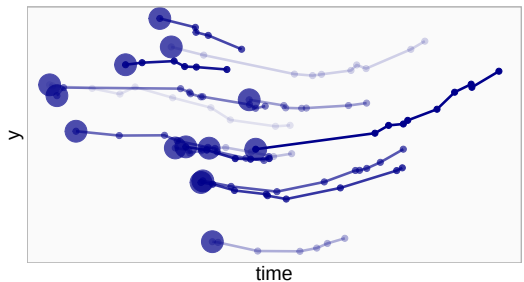


Naive approaches that are sometimes used are to

- **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**

However, **important information may be lost**, resulting in invalid imputations and biased results.

Survival Data

In **survival analysis**, the aim is to estimate the effect of covariates on the **time until an event** of interest happens.

Survival Data

In **survival analysis**, the aim is to estimate the effect of covariates on the **time until an event** of interest happens.

Commonly used method: **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

In **survival analysis**, the aim is to estimate the effect of covariates on the **time until an event** of interest happens.

Commonly used method: **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 outcomes are usually represented by the **observed event time** T and the **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

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

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

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

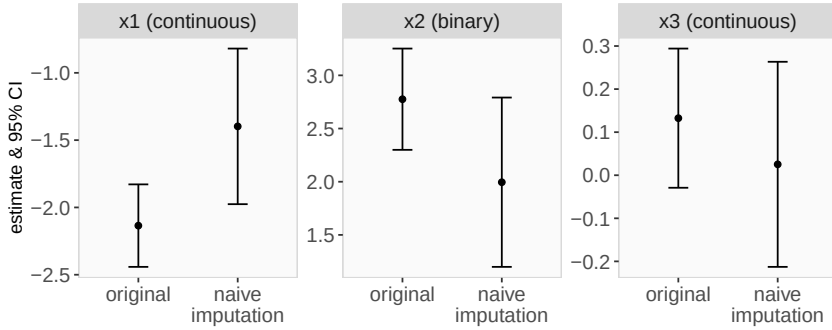
The **correct conditional distribution** of x given the other variables is, however,

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

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.